Département de Sciences Économique

Analyse économétrique Avec Stata 12

A. Bédard T. Boquet P. Lapierre

Université de Montréal

Sommaire

Préface 1

1 Introduction à STATA 2

1.1 Ouvrir STATA 2

1.2 La page d'ouverture 3

1.3 Fermer STATA 3

1.4 Utiliser un fichier contenant une base de données 4

1.4.1 La commande use 4

1.4.2 Chargement d'une base de données via la barre d'outils 5

1.4.3 La fenêtre data editor 5

1.4.4 La commande help 6

1.4.5 Le manuel de référence du logiciel 7

1.5 Importer des données à partir de CANSIM sur E-STAT 7

1.5.1 La méthode « copier-coller » 11

1.5.2 La méthode d'importation par STATA – Le fichier *.csv 12

1.6 Statistique descriptive 15

1.6.1 La commande list 16

1.6.2 La commande describe 16

1.6.3 La commande summarize 18

1.6.4 La commande tabulate 21

1.6.5 Coefficient de corrélation 23

1.7 Enregistrer son travail 24

1.8 La commande display 24

1.9 Créer un graphique 25

1.9.1 Histogramme 25

1.9.2 Nuage de points 26
1.9.3 Graphique en ligne 28
1.10 Générer de nouvelles variables 30
1.11 Générer des variables avec un lag ou un lead 33
1.12 Éliminer et renommer des variables 35
1.13 Le codage de variables 36
1.14 Les commandes sort 37
1.15 Organisation des données 39
1.16 Log file 41
1.17 Do file 43
1.18 Tests sur la moyenne, la variance et la distribution des variables 45
1.18.1 Inférence sur le paramètre d'une population suivant une distribution normale 46
1.18.2 Estimation de la différence de moyenne de deux populations suivant une loi normale 47
1.18.3 Estimation de la différence de moyenne de deux populations suivant une loi normale 49
2 Régression linéaire 52
2.1 Introduction 52
2.2 La structure des données en science économique 52
2.2.1 Les données en coupe 52
2.2.2 Séries temporelles 53
2.2.3 Les données mixtes 55
2.2.4 Les données longitudinales 55
2.3 Le modèle de régression linéaire simple 57
2.3.1 Rappel théorique 57
2.3.2 Le pouvoir explicatif de la régression linéaire 58

2.4 Tests d'hypothèse et intervalles de confiance 60
2.4.1 Introduction 60
2.4.2 Tests d'hypothèse 61
2.4.3 Intervalles de confiance 62
2.5 Comment rapporter les résultats d'une régression 62
2.6 Régression linéaire simple - Application sur STATA 64
2.6.1 Générer la régression linéaire simple 64
2.6.2 Interprétation de l'output STATA de la régression linéaire simple 65
2.6.3 Générer les valeurs prédites du modèle 68
2.6.4 Calculer l'élasticité des coefficients 69
2.6.5 Obtenir la matrice variance-covariance 71
2.6.6 Générer le graphique de la droite de régression 72
2.7 Le modèle de régression linéaire multiple 76
2.7.1 Rappel théorique 76
2.7.2 Test de Fisher (le test joint) 78
2.8 Modèle de régression linéaire multiple – Application sur STATA 79
2.8.1 Générer la régression linéaire multiple 79
2.8.2 Interprétation de l'output STATA de la régression linéaire multiple 80
2.8.3 Tester le niveau de significativité des variables 81
2.8.4 Générer un graphique avec les estimations du modèle de régression linéaire multiple 82
2.9 Hétéroscédasticité 86
2.9.1 Rappel théorique 86
2.9.2 Tester pour l'hétéroscédasticité 87
2.9.3 Estimateurs robustes à l'hétéroscédasticité 93
2.9.4 Estimateurs robustes à l'hétéroscédasticité – Application sur STATA 94

2.9.5 Les estimateurs moindres carrés pondérés (MCP) 97
2.9.6 Les estimateurs moindres carrés pondérés (MCP) – Application sur STATA 99
2.9.7 La méthode des moindres carrés généralisés faisables 101
2.9.8 La méthode des moindres carrés généralisés faisables – Application sur STATA 103
2.9.9 Observer l'hétéroscédasticité par graphique 106
2.10 Les différentes formes fonctionnelles de variables 109
2.10.1 Variables muettes 109
2.10.2 Variables logarithmiques 116
2.10.3 Les variables quadratiques 119
2.10.4 Les variables d'interaction 122
2.10.5 Générer une variable d'interaction sur STATA 123
2.11 Forme fonctionnelle d'un modèle 124
2.11.1 Spécificité du modèle 124
2.11.2 Tester la forme fonctionnelle d'un modèle sur STATA 125
2.12 Normalité dans le terme d'erreur 127
2.13 Tester la multicollinéarité des variables explicatives 131
2.13.1 Rappel théorique 131
2.13.2 Tester la multicolinéarité des variables explicatives – Application sur STATA .. 132
3 Variables dépendantes discrètes 136
3.1 Introduction 136
3.2 Modèles logit et probit binaires 136
3.2.1 Rappel théorique 136
3.2.2 La commande logit 140
3.2.3 Les commandes postestimation après la commande logit 141
3.2.4 Exemple 143

3.2.5 La commande probit 171
3.2.6 Les commandes postestimation après la commande Probit 171
3.2.7 Exemple 173
3.3 Modèles logit et probit multinomiaux 176
3.3.1 Rappel théorique 176
3.3.2 La commande mlogit 179
3.3.3 Les commandes postestimation après la commande mlogit 180
3.3.4 Exemple 182
3.3.5 La commande mprobit 216
3.3.6 Les commandes postestimation après la commande mprobit 217
3.3.7 Exemple 219
3.4 Modèles logit et probit ordonnés 220
3.4.1 Rappel théorique 220
3.4.2 La commande ologit 221
3.4.3 Les commandes postestimation après la commande ologit 222
3.4.4 Exemple 224
3.4.5 La commande oprobit 259
3.4.6 Les commandes postestimation après la commande oprobit 259
3.4.7 Exemple 261
4 Données de comptage 263
4.1 Introduction 263
4.2 Modèle de régression de Poisson 263
4.2.1 Rappel théorique 263
4.2.2 La commande poisson 264
4.2.3 Les commandes postestimation après la commande poisson 265

4.2.4 Exemple ... 267
4.3 Modèle de régression binomiale négative ... 298
4.3.1 Rappel théorique ... 298
4.3.2 La commande nbreg ... 299
5 Variables dépendantes limitées ... 305
5.1 Introduction ... 305
5.2 Modèle Tobit ... 305
5.2.1 Rappel théorique ... 305
5.2.2 La commande tobit (modèle Tobit type I) ... 307
5.2.3 Les commandes postestimation après la commande tobit ... 307
5.2.4 Exemple ... 310
5.3 Modèle de régression tronquée ... 341
5.3.1 Rappel théorique ... 341
5.3.2 La commande truncreg ... 343
5.3.3 Les commandes postestimation après la commande tobit ... 344
5.3.4 Exemple ... 346
6 Problèmes d'endogénéité ... 349
6.1 Introduction ... 349
6.2 Estimation par variables instrumentales ... 349
6.2.1 Rappel théorique ... 349
6.2.2 La commande ivregress ... 351
6.2.3 Les commandes postestimation après la commande ivregress ... 352
6.2.4 Exemple ... 354
7 Les Séries Temporelles univariées ... 368
7.1 Définition d'une série temporelle ... 368

7.1.1 Modèle Statique 368
7.1.2 Modèle à Retards Échelonnés Finis 368
7.1.3 Modélisation 368
7.2 Créer et nommer la série temporelle 369
7.3 Graphique de la tendance 370
7.3.1 Tendance linéaire 370
7.3.2 Tendance exponentielle 372
7.3.3 Tendance Quadratique 373
7.4 Utilisation de variables avec tendance dans une régression 374
7.5 Méthode partialling out 375
7.6 R^2 lorsqu'il y a une tendance 375
7.7 Tester pour des coupures dans la tendance 376
7.7.1 Coupure avec date inconnue 376
7.7.2 Coupure avec date connue 379
7.8 Saisonnalité 380
7.8.1 Commandes Stata 381
7.8.2 Enlever la saisonnalité 381
7.9 Stationnarité et test de racine unitaire 382
7.9.1 Test de Dickey-Fuller 382
7.10 Autocorrélation 384
7.10.1 Test pour l'autocorrélation 384
7.11 Test de Durbin-Watson sous les hypothèses classiques 385
7.11.1 Test AR(1) d'autocorrélation sans exogénéité stricte 386
7.11.2 Test d'autocorrélation de Breusch-Godfrey 387
7.11.3 Durbin Alternative Test 387

7.12 Modèle Autorégressif 388
7.12.1 Exemple 388
7.13 Corriger l'autocorrélation 390
7.13.1 MCQG sous AR(1) 390
7.13.2 Écart-types robustes à l'autocorrélation – Newey West 393
7.14 Hétéroscédasticité dans les séries temporelles 394
7.15 Autoregressive Conditional Heteroskedasticity 394
7.15.1 Exemple 395
7.16 Modèle de discrimination 396
7.16.1 Modèle Oaxaca-Blinder 396
7.16.2 Commande Stata 397

Préface

Ce guide est destiné aux étudiants désirant avoir une introduction aux outils de base du logiciel Stata afin de pouvoir estimer et analyser les modèles les plus communs en sciences sociales et en sciences de la gestion. Il est fortement recommandé d'avoir en parallèle la documentation officielle du logiciel, les livres de Wooldridge, (Introductory Econometrics : A Modern Approach, fourth edition, 2009) pour une explication intuitive des concepts théoriques, de Greene (Econometric Analysis, 2012) et Wooldridge (Econometric analysis of cross section and panel data, second edition, 2001) pour une explication plus technique. Les livres de Cameron & Trivedi (Microeconometrics: Methods and Applications, 2005), et Cameron & Trivedi (Microeconometrics using Stata, 2009) sont des références utiles pour les modèles non couverts par ce manuel. Nous incitons aussi les lecteurs à avoir un navigateur internet ouvert afin de pouvoir effectuer des recherches en ligne quand les explications dans le manuel officiel du logiciel, les ouvrages de référence et ce guide ne suffisent pas à la compréhension ou à l'avancement d'un projet de recherche. Ce guide a été rédigé pour la version 12 du logiciel Stata et utilise les nouvelles commandes disponibles.

1 Introduction à STATA

1.1 Ouvrir STATA

Il existe différentes avenues afin de démarrer le programme STATA. Dans notre cas, nous utiliserons Windows pour toutes les prochaines démonstrations. Tout d'abord, il est possible d'ouvrir le programme à partir du bureau sur l'icône de raccourcie de STATA. Il est également possible de passer par le **menu démarrer** → **Programmes** → **STATA** (selon la version installée). Finalement, il est simplement possible de localiser un fichier *.dta[1] sur le portable et de double-cliquer.

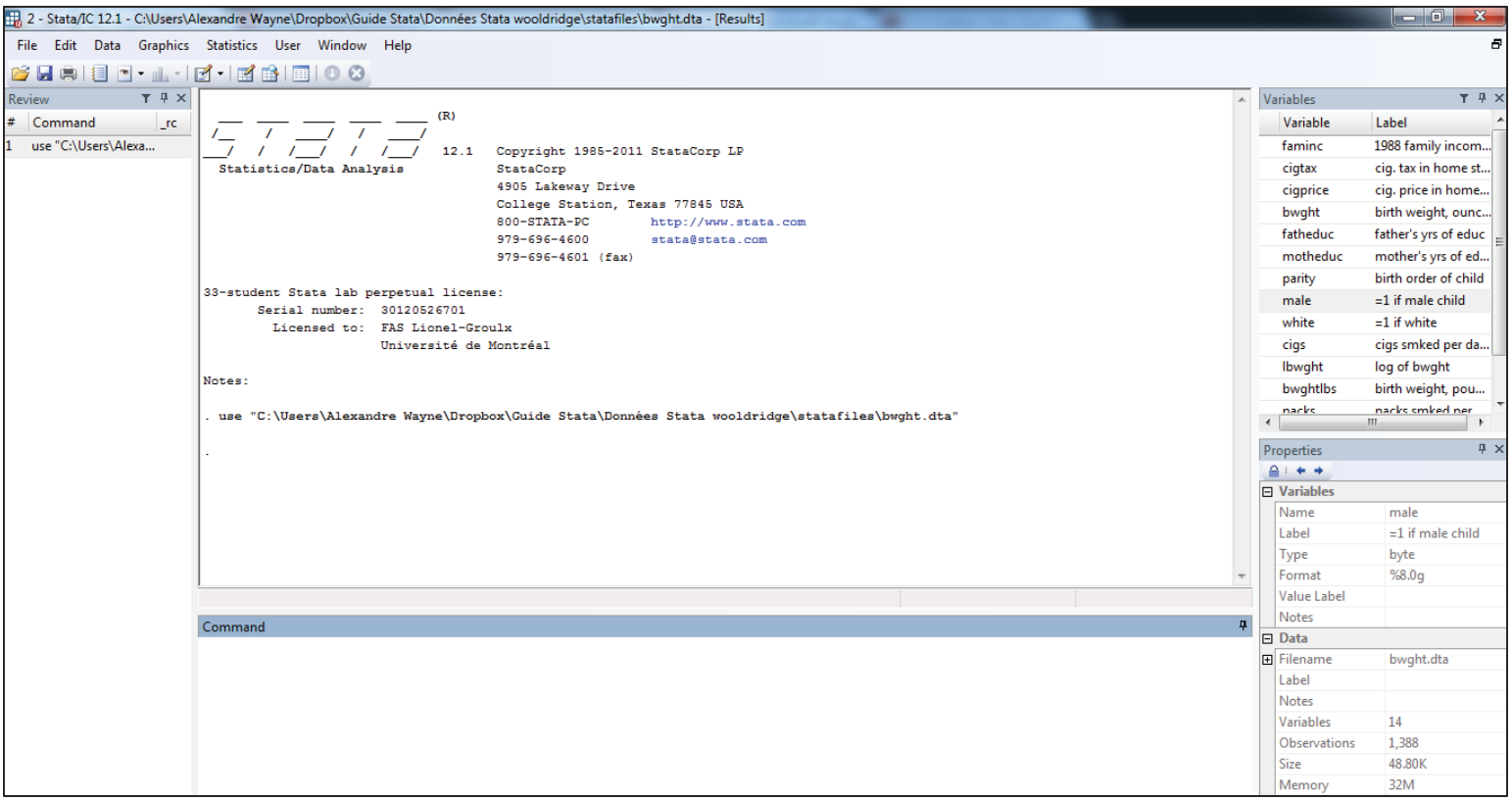
[1] À noter que tout fichier de base de données enregistré sur STATA 12 est libellé en dta. Il existe quelques exceptions que nous verrons ultérieurement.

1.2 La page d'ouverture

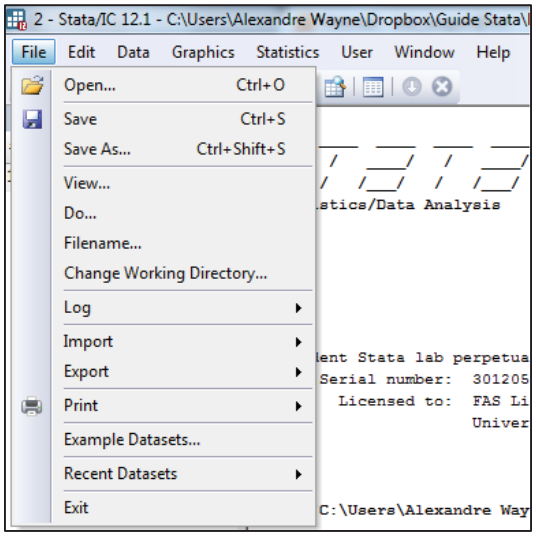
Cinq fenêtres importantes sont affichées une fois que la page d'ouverture de STATA est ouverte :

Commande : la fenêtre où l'on insère les différentes commandes du logiciel
Variables : la liste des variables qui composent la base de données de l'utilisateur STATA
Properties : la fenêtre qui décrit à la fois la base de données ainsi que les variables
Review : la liste des commandes récemment exécutées
Results : la fenêtre des résultats

1.3 Fermer STATA



Il existe deux façons de terminer une session sur STATA. La première consiste à cliquer sur l'onglet **File**



Et par la suite sélectionner l'onglet **Exit.**

La seconde façon de procéder à la fermeture de STATA consiste à taper `exit` dans la fenêtre de commande et d'appuyer sur la touche **enter** du clavier.

1.4 Utiliser un fichier contenant une base de données

Il existe plusieurs façons d'ouvrir une base de données à partir du logiciel STATA. La première consiste à sélectionner directement un fichier *.dta sur l'ordinateur soit par la commande `use` ou la barre d'outils présentée sur la page d'ouverture de STATA.

1.4.1 La commande use

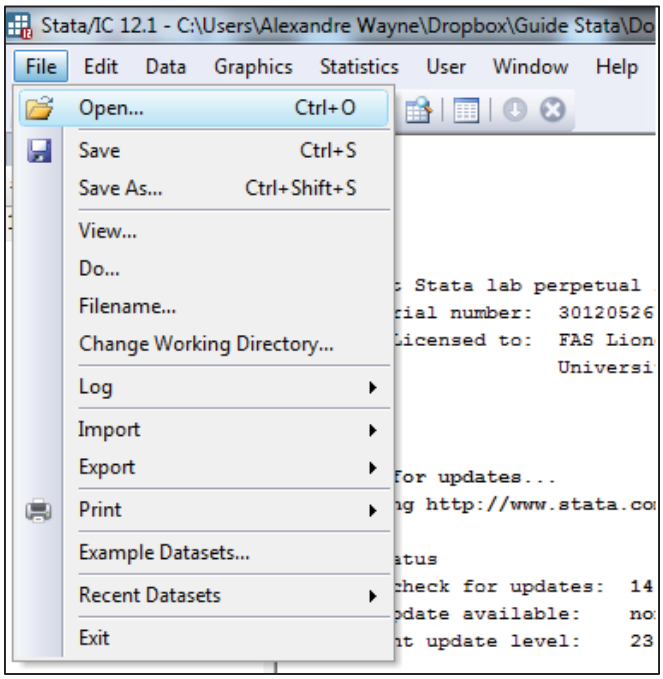
Afin de télécharger un fichier *.dta directement de l'ordinateur, nous pouvons taper la commande `use` ainsi que le lien définissant la localisation du dit fichier.

```
Command
use "C:\Users\Alexandre Wayne\Dropbox\Guide Stata\Données Stata wooldridge\statafiles\bwght.dta"
```

Notons qu'il ne faut pas oublier les guillemets entre le lien du fichier *.dta, sans quoi, la commande ne fonctionnera pas.

1.4.2 Chargement d'une base de données via la barre d'outils

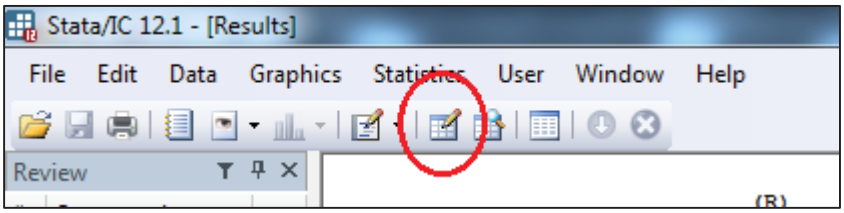
Afin d'ouvrir une base de données, il est également possible de procéder via la barre d'outils en cliquant sur **File > Open** et en sélectionnant le fichier *.dta que vous désirez télécharger.



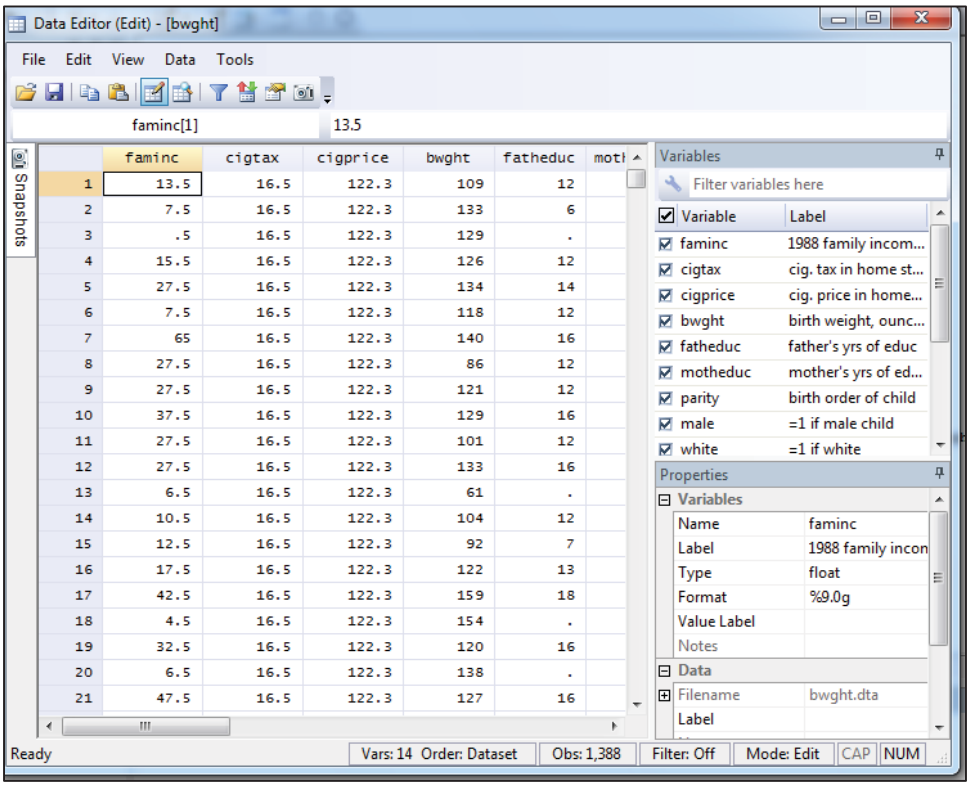
1.4.3 La fenêtre data editor

La fenêtre data editor s'avère être le classeur STATA contenant les données. Il existe deux alternatives pour y accéder.

1. Par la barre d'outils de la page d'ouverture, **data > data editor >data editor (edit)**.
2. Par l'icône encerclée en rouge dans l'image ci-dessous.



La fenêtre data editor ressemble au classeur ici-bas (à noter que le tableur est vide lorsqu'il n'y a pas de données téléchargées).



Dans le classeur, la première ligne représente les variables tandis que la première colonne représente les observations. Il est possible de voir le classeur en entier à l'aide des barres de défilement. De plus, nous pouvons modifier nos données directement en sélectionnant une entrée dans le classeur. Notons également que lorsqu'une entrée prend la valeur « . », cela signifie qu'il n'existe aucune valeur pour cette observation.

1.4.4 La commande help

La commande `help` de STATA propose un vaste éventail d'information sur l'ensemble des commandes qui composent le programme. Par exemple, advenant le cas où l'on désire obtenir de l'information sur la commande `use`, il est possible de procéder en exécutant la commande suivante :

```
help use
```

STATA fera apparaître une fenêtre d'aide portant sur la syntaxe générale, l'accès via la barre d'outils, la description des options, etc. propres à la commande `use`.

1.4.5 Le manuel de référence du logiciel

Dans le menu **Help>PDF documentation**, l'utilisateur pourra accéder à la documentation du logiciel. Les parties pertinentes de celle-ci avec la matière de ce guide sont : la partie Data management, la partie Base Reference, la partie Time Series et la partie Longitudinal Data/Panel Data. Tous les concepts ne sont évidemment pas traités, mais apprendre à naviguer dans cette documentation peut largement faciliter l'utilisation du logiciel.

1.5 Importer des données à partir de CANSIM sur E-STAT

La prochaine section vise à familiariser le lecteur avec l'extraction de données à partir de CANSIM sur E-STAT. Elle porte également sur une séance de travaux pratiques donnée dans le cadre du cours ECN 1160 – Analyse des données économiques. Nous énumérerons et expliciterons les étapes de cette section à l'aide d'un exemple où nous tenterons d'extraire une série temporelle sur le taux de chômage canadien.

1. Tapons l'adresse de la bibliothèque des lettres et sciences humaines de l'Université de Montréal - http://www.bib.umontreal.ca/SS/ - dans la barre de navigation.
2. Posons notre curseur sur l'onglet « Ressources électroniques » (partie supérieure de la page d'accueil) et cliquons sur « Information statistique ».

3. Cliquons sur l'onglet CANSIM sur E-STAT. Nous serons dès lors redirigés vers la page d'accueil de CANSIM sur E-STAT de Statistique Canada.

4. Sur cette page, nous aurons accès à une barre de recherche ainsi que différents thèmes de recherche classés par sujet (agriculture, aînés, etc.). Pour faire un lien avec notre exemple, nous pouvons taper « chômage » dans la barre de recherche. À la page suivante, vous aurez ainsi accès à différents tableaux détenant comme mot clé « chômage ». Libre à vous de choisir le tableau qui vous convient selon l'objectif de votre recherche. Dans notre cas, pour faciliter les choses, nous utiliserons le tableau 109-5324 (il suffit d'entrer le numéro de tableau dans la barre de recherche afin de procéder aux différentes caractéristiques de votre sous-ensemble de données – étape 5).

5. Ici, nous devons définir les caractéristiques de votre sous-ensemble, soit la géographie, le type de taux de chômage et la période de référence. Notons que ces caractéristiques sont propres au tableau que nous avons utilisé; celles-ci dépendent de la recherche que vous effectuez. Pour notre exemple, sélectionnons :

    a. Géographie → Canada.
    b. Taux de chômage → Taux de chômage, 15 ans et plus.
    c. Période de référence → 2006 à 2010.

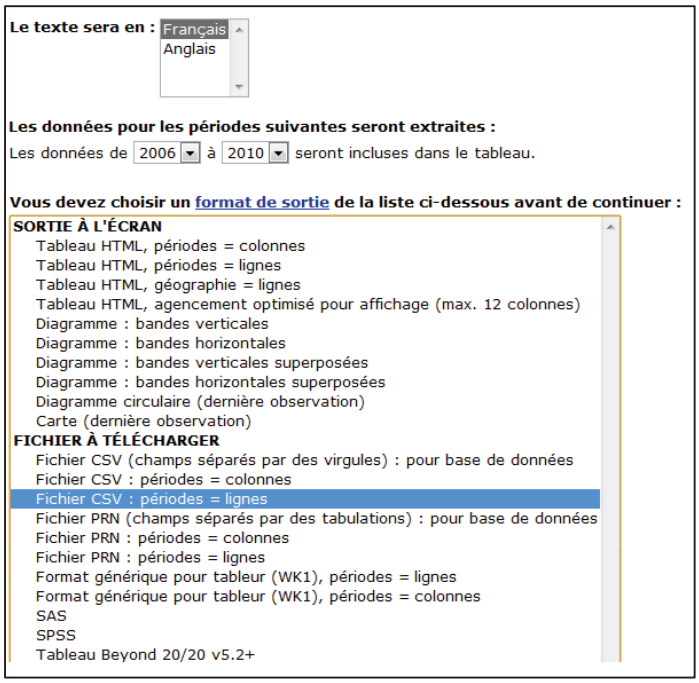
    Lorsque nous avons terminé de sélectionner nos caractéristiques, abaissons notre curseur jusqu'en bas de la page et cliquons sur « extraire tableau ».

6. La nouvelle page nous amène dans la spécification de la sortie où nous devons sélectionner le format de sortie. Pour notre exemple, cliquons sur :

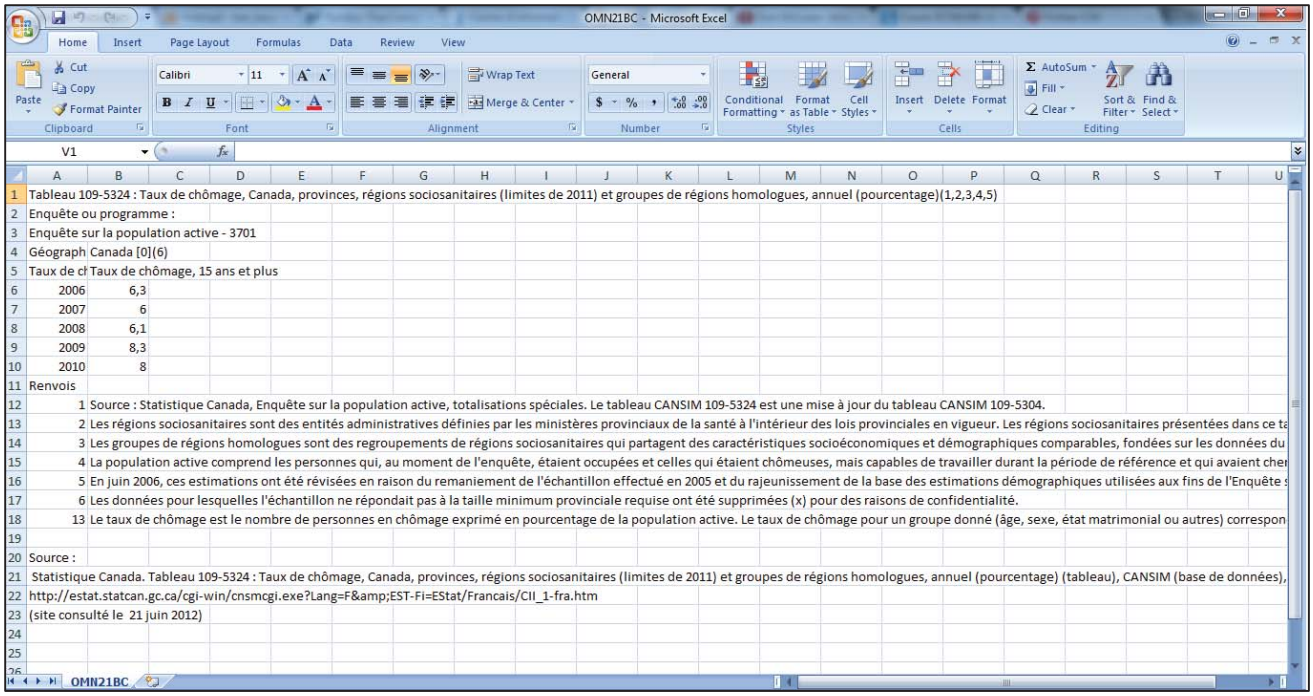
    a. Langue du texte → Français.
    b. Période d'extraction → 2006 à 2010.
    c. Format de sortie → Fichier CSV[2] : périodes = lignes

[2] Un fichier CSV est un fichier tableau.

Descendons notre curseur jusqu'en bas de la page et cliquons sur extraire maintenant. Notre fichier sera téléchargé avec Microsoft Excel.



7. Lorsque notre fichier est téléchargé, ouvrons-le. Normalement, si nous avons respecté toutes les étapes précédentes, nous devrions obtenir un tableau Excel qui ressemble à l'image ci-dessous. La prochaine étape consiste à importer les données qui nous intéressent de Excel à Stata. Nous verrons deux façons de procéder. Libre à au lecteur de choisir celle qui nous convient le mieux.



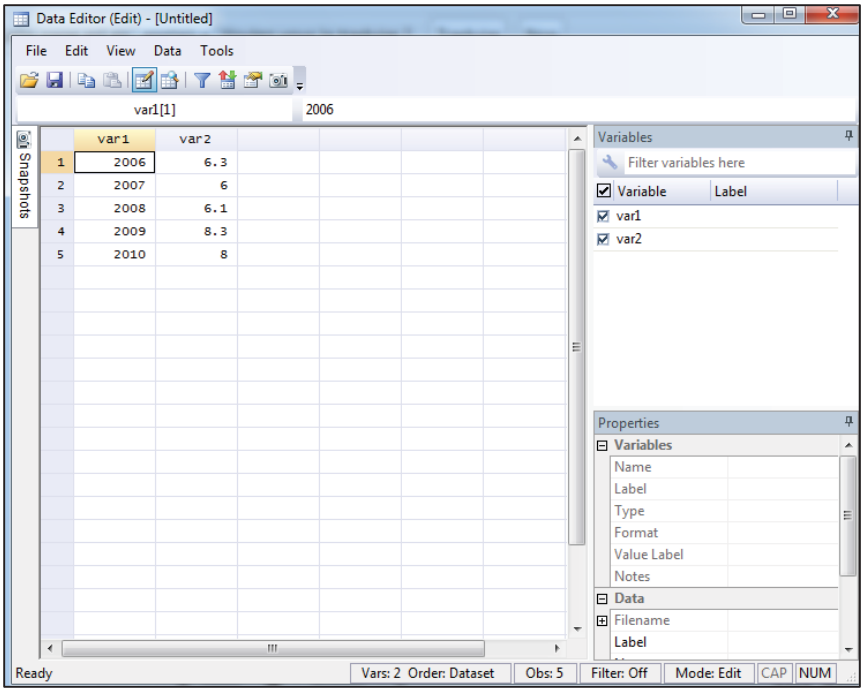
1	Tableau 109-5324 : Taux de chômage, Canada, provinces, régions sociosanitaires (limites de 2011) et groupes de régions homologues, annuel (pourcentage)(1,2,3,4,5)	
2	Enquête ou programme :	
3	Enquête sur la population active - 3701	
4	Géograph	Canada [0](6)
5	Taux de ch	Taux de chômage, 15 ans et plus
6	2006	6,3
7	2007	6
8	2008	6,1
9	2009	8,3
10	2010	8
11	Renvois	
12	1	Source : Statistique Canada, Enquête sur la population active, totalisations spéciales. Le tableau CANSIM 109-5324 est une mise à jour du tableau CANSIM 109-5304.
13	2	Les régions sociosanitaires sont des entités administratives définies par les ministères provinciaux de la santé à l'intérieur des lois provinciales en vigueur. Les régions sociosanitaires présentées dans ce ta
14	3	Les groupes de régions homologues sont des regroupements de régions sociosanitaires qui partagent des caractéristiques socioéconomiques et démographiques comparables, fondées sur les données du
15	4	La population active comprend les personnes qui, au moment de l'enquête, étaient occupées et celles qui étaient chômeuses, mais capables de travailler durant la période de référence et qui avaient che
16	5	En juin 2006, ces estimations ont été révisées en raison du remaniement de l'échantillon effectué en 2005 et du rajeunissement de la base des estimations démographiques utilisées aux fins de l'Enquête s
17	6	Les données pour lesquelles l'échantillon ne répondait pas à la taille minimum provinciale requise ont été supprimées (x) pour des raisons de confidentialité.
18	13	Le taux de chômage est le nombre de personnes en chômage exprimé en pourcentage de la population active. Le taux de chômage pour un groupe donné (âge, sexe, état matrimonial ou autres) correspon
19		
20	Source :	
21	Statistique Canada. Tableau 109-5324 : Taux de chômage, Canada, provinces, régions sociosanitaires (limites de 2011) et groupes de régions homologues, annuel (pourcentage) (tableau), CANSIM (base de données),	
22	http://estat.statcan.gc.ca/cgi-win/cnsmcgi.exe?Lang=F&EST-Fi=EStat/Francais/CII_1-fra.htm	
23	(site consulté le 21 juin 2012)	

1.5.1 La méthode « copier-coller »

La première méthode consiste à copier-coller votre série dans le classeur STATA (fenêtre data editor). Voici les étapes pour y parvenir :

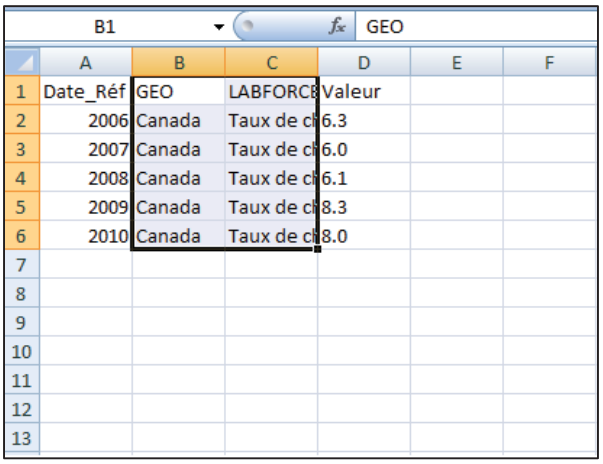
1. Tout d'abord, STATA n'identifie pas les virgules dans son tableau (il mettra en rouge toutes les données comportant des virgules collées dans le classeur, signe qu'il y a une erreur). Afin d'éviter cela, il faut changer les virgules de notre fichier Excel en points. Pour y arriver, de notre fichier Excel, sélectionnons les données qui nous intéressent (l'encadré dans l'image ci-dessous) et appuyons respectivement sur les touches Ctrl → F de notre clavier. Une fenêtre apparaîtra, cliquons sur l'onglet « remplacer » et tapons sur la touche virgule.

2. Lorsque les modifications sont apportées, sélectionnez vos données et copiez-les. Ensuite, ouvrez STATA et cliquez sur l'icône de data editor (edit). Lorsque la fenêtre est ouverte, sélectionnez la première entrée (coin supérieur gauche) de votre classeur et collez vos données. Vous avez maintenant un classeur STATA contenant deux variables (var1 et var2) et un total de cinq observations par variable.

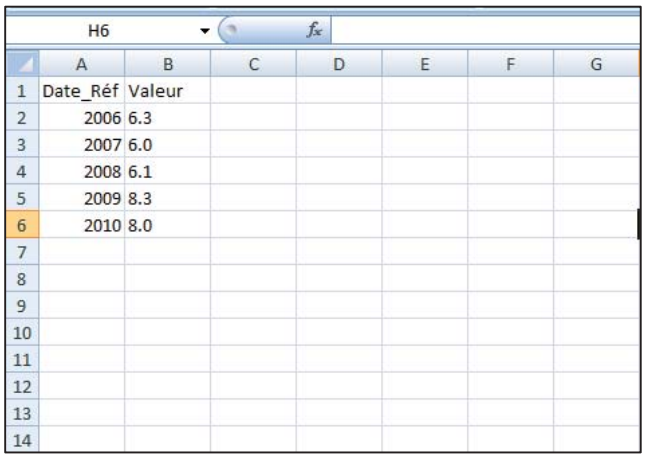


1.5.2 La méthode d'importation par STATA – Le fichier *.csv.

Lorsque le fichier à télécharger est en format *.csv, il est possible d'importer la base de données à l'aide de la commande `insheet`. En reprenant les mêmes données du tableau 109-5324, téléchargeons le fichier en cliquant sur « Fichier CSV (champs séparés par des virgules) : pour base de données » dans la section « Spécification de la sortie » sur CANSIM sur E-STAT. Lorsque le fichier est téléchargé, ouvrons le fichier OMNB2F5.csv. Le contenu devrait ressembler à l'image ci-dessous.

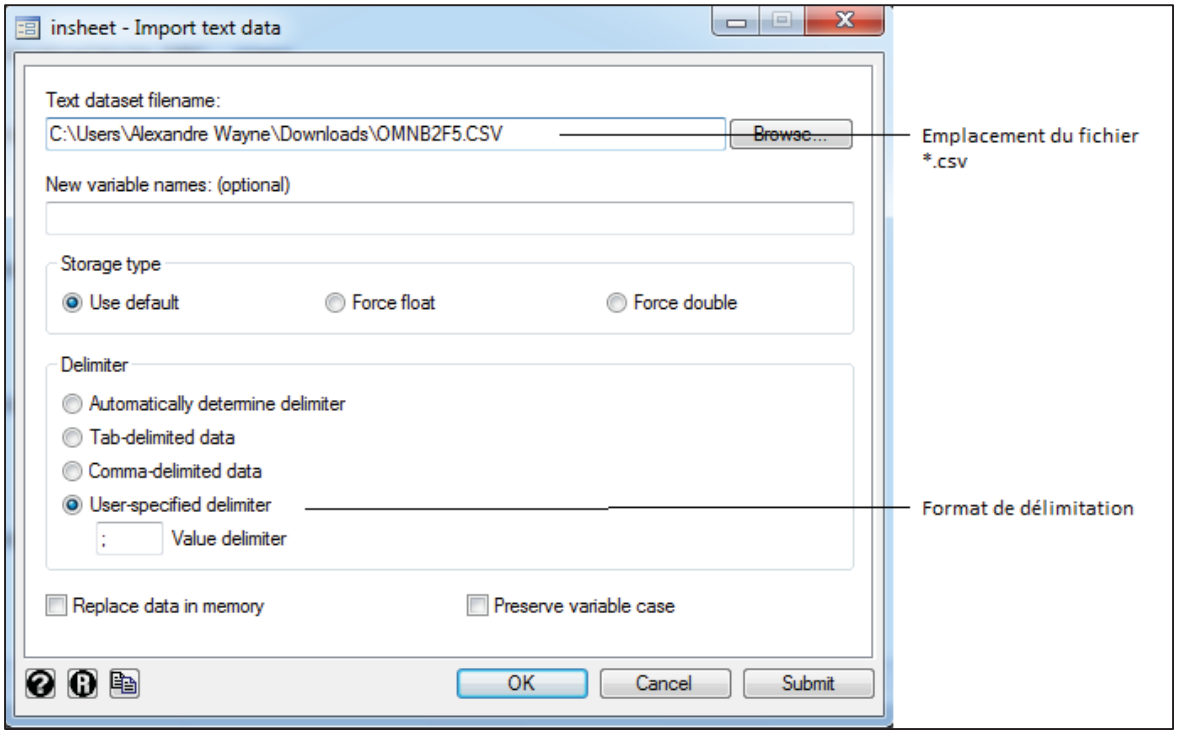


Ici, il est important de bien nettoyer le classeur ci-haut en sélectionnant seulement les variables d'intérêt tout en présentant le nom des variables sur la première ligne, sans quoi STATA risque d'importer notre base de données avec des erreurs. Advenant le cas où les séries comportent des virgules, nous devons interchanger les virgules pour des points. Le nettoyage final devrait ressembler à l'image ci-dessous.



	A	B	C	D	E	F	G
1	Date_Réf	Valeur					
2	2006	6.3					
3	2007	6.0					
4	2008	6.1					
5	2009	8.3					
6	2010	8.0					
7							
8							
9							
10							
11							
12							
13							
14							

Cette étape terminée, nous pouvons ouvrir STATA et sélectionner dans la barre d'outils les onglets suivant : **File > Import >Text data created by a spreasheet**. Dans la nouvelle fenêtre, nous devons insérer l'emplacement du fichier OMNB2F5.csv dans la case « Text dataset filename » ainsi que le format de délimitation dans l'espace « Delimiter ». Le fichier téléchargé utilise des données séparées par des points-virgules (notons que le format de délimitation dépend d'une source à l'autre). Ce faisant, insérons « ; » dans la case « user-specified delimiter ». Appuyons sur OK et le fichier sera importé correctement sur STATA.



La commande résumant les opérations sur STATA est la suivante :

```
insheet using "C:\Users\Alexandre Wayne\Downloads\OMNB2F5.CSV", delimiter(";")
```

1.6 Statistique descriptive

Lorsque nous importons de nouvelles données, il est utile et parfois même primordial d'avoir une vue d'ensemble sur celles-ci. Cette sous-section donne les outils nécessaires afin de dégager toutes statistiques descriptives (mesures de tendance centrale, mesures de dispersion, etc.) reliées aux données téléchargées sur STATA.

Tout d'abord, STATA utilise une syntaxe qui lui est propre afin de faire de la statistique descriptive. Cette syntaxe va comme suit :

```
command [varlist] [if] [in] [weight] [ , options]3
```

`command` est la commande principale utilisée
`[varlist]` est la liste des variables incluses dans la commande
`[if]` est une condition imposée dans la commande
`[in]` définit un intervalle spécifique dans les observations
`[weight]` pour des observations devant être pondérées différemment l'une de l'autre
`[ , options]` spécifie une option rattachée à votre commande

De plus, quelques spécifications en programmation pour STATA concernant les opérateurs logiques doivent être prises en compte afin de pouvoir utiliser optimalement les commandes ci-haut. Voici une liste exhaustive de ces opérateurs.

[3] Notons que les termes entre crochet s'avèrent des option qui peuvent être utilisées afin de dégager de l'information sur les variables. Cette syntaxe est également utilisée dans le cas de d'autres opérations : régression, création de graphiques, etc. Dans le cas des options spécifiques, le lecteur est inviter à se documenter via STATA 12 en insérant la commande help [la commande de statistique descriptive].

Opérateur	Signification
==	égal
>=	plus grand ou égal
<=	plus petit ou égal
>	plus grand
<	plus petit
&	et
|	ou
~	négation
!	négation

1.6.1 La commande list

La commande `list` reprend essentiellement les informations contenues dans le classeur STATA en les présentant en output dans la fenêtre des résultats. Bien entendu, celle-ci s'avère peu utile dans le cas d'un classeur comportant un grand nombre de variables et de données. Un truc utile afin d'éviter un débordement d'information dans la fenêtre des résultats est de délimiter un intervalle dans les observations avec l'option « `in` ». Par exemple, `list x in 1/20`, c'est-à-dire « lister les 20 premières observations de la variable x ».

1.6.2 La commande describe

La commande `describe` offre au lecteur de l'information générale, notamment le nombre d'observations et de variables ainsi que le nom des variables et leur « label », et ce, sur l'ensemble des variables qui composent le classeur STATA. En utilisant les données de Wooldridge (2009), BWGHT.dta, la commande `describe` offre l'output suivant :

```
use http://www.Stata.com/data/jwooldridge/eacsap/bwght.dta
describe

Contains data from C:\Users\Alexandre Wayne\Desktop\bwght.dta
  obs:         1,388
 vars:            14                          31 Dec 2001 11:49
 size:        49,968
-------------------------------------------------------------------------------------------------
---------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------------------------
---------------------------------------
faminc          float  %9.0g                  1988 family income, $1000s
cigtax          float  %9.0g                  cig. tax in home state, 1988
cigprice        float  %9.0g                  cig. price in home state, 1988
bwght           int    %8.0g                  birth weight, ounces
fatheduc        byte   %8.0g                  father's yrs of educ
motheduc        byte   %8.0g                  mother's yrs of educ
parity          byte   %8.0g                  birth order of child
male            byte   %8.0g                  =1 if male child
white           byte   %8.0g                  =1 if white
cigs            byte   %8.0g                  cigs smked per day while preg
lbwght          float  %9.0g                  log of bwght
bwghtlbs        float  %9.0g                  birth weight, pounds
packs           float  %9.0g                  packs smked per day while preg
lfaminc         float  %9.0g                  log(faminc)
-------------------------------------------------------------------------------------------------
---------------------------------------
Sorted by:
```

Élément intéressant à noter dans la deuxième colonne du tableau précédent, le format de stockage des données (`storage type`), définit selon le nombre de caractères des données afin de minimiser la mémoire utilisée. Dans les logiciels de statistiques et dans STATA, il existe deux types de données : les données en texte (string) et les données numériques. Parmi les données numériques, nous dénombrons cinq catégories de variables selon le nombre de caractères utilisés :

Byte : nombre entier dans l'intervalle [-127,100]

Int : nombre entier dans l'intervalle [-32 767, 32 740]

Long : nombre entier dans l'intervalle [-2 147 483 647, 2 147 483 620]

Float : nombre réel dans l'intervalle [-1.70141173319*10^38, 1.70141173319*10^38]

Double : nombre réel dans l'intervalle [-8.9884656743*10^307, 8.9884656743*10^307]

1.6.3 La commande summarize

La commande `summarize` (ou `sum`) offre des renseignements statistiques sur les variables qui composent le classeur STATA. Voici quelques exemples afin de familiariser le lecteur dans la procédure entourant la commande `summarize`. La base de données utilisée, BWGHT.dta, provient de Wooldridge (2009).

- `sum faminc`

```
. sum faminc

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      faminc |      1388    29.02666    18.73928         .5         65
```

Ce tableau présente dans l'ordre le nombre d'observations, la moyenne, l'écart type, la valeur minimale et la valeur maximale de la variable faminc dans l'échantillon.

```
sum faminc, detail

                 1988 family income, $1000s
-------------------------------------------------------------
      Percentiles      Smallest
 1%           .5             .5
 5%          3.5             .5
10%          6.5             .5       Obs                1388
25%         14.5             .5       Sum of Wgt.        1388

50%         27.5                      Mean           29.02666
                        Largest       Std. Dev.      18.73928
75%         37.5             65
90%           65             65       Variance       351.1608
95%           65             65       Skewness       .6176205
99%           65             65       Kurtosis       2.473396
```

Ce tableau, en plus des éléments énumérés ci-dessus, présente les valeurs attribuées pour chaque percentile, la variance, la symétrie de la distribution (*skewness*) et le coefficient d'aplatissement (*kurtosis*).

```
sum faminc if faminc<30

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      faminc |       814    16.04914    8.199656         .5       27.5
```

Ce tableau présente dans l'ordre le nombre d'observations, la moyenne, l'écart type, la valeur minimale et la valeur maximale pour toute observation dont la valeur est inférieure à 30.

```
sum faminc in 1/65

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      faminc |        65    28.63846     21.3961         .5         65
```

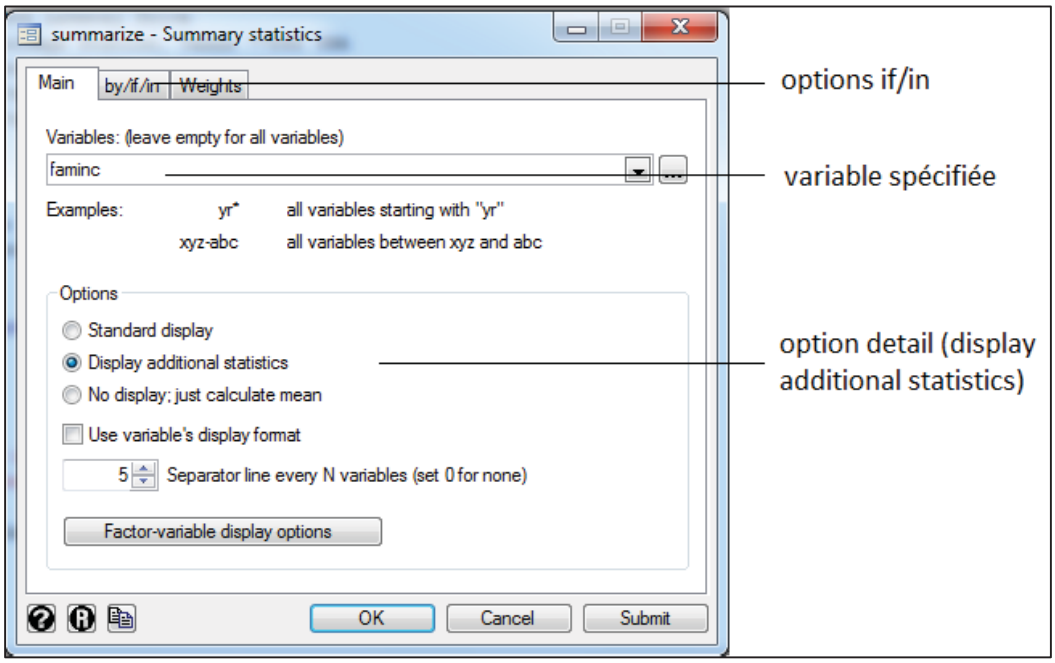
Encore une fois, ce tableau présente dans l'ordre le nombre d'observations, la moyenne, l'écart type, la valeur minimale et la valeur maximale pour les observations 1 à 65. Libre au lecteur de changer cet intervalle en modifiant le numérateur et le dénominateur de la fraction dans la commande (1 étant la première observation et 1388 étant la dernière recensée dans l'échantillon).

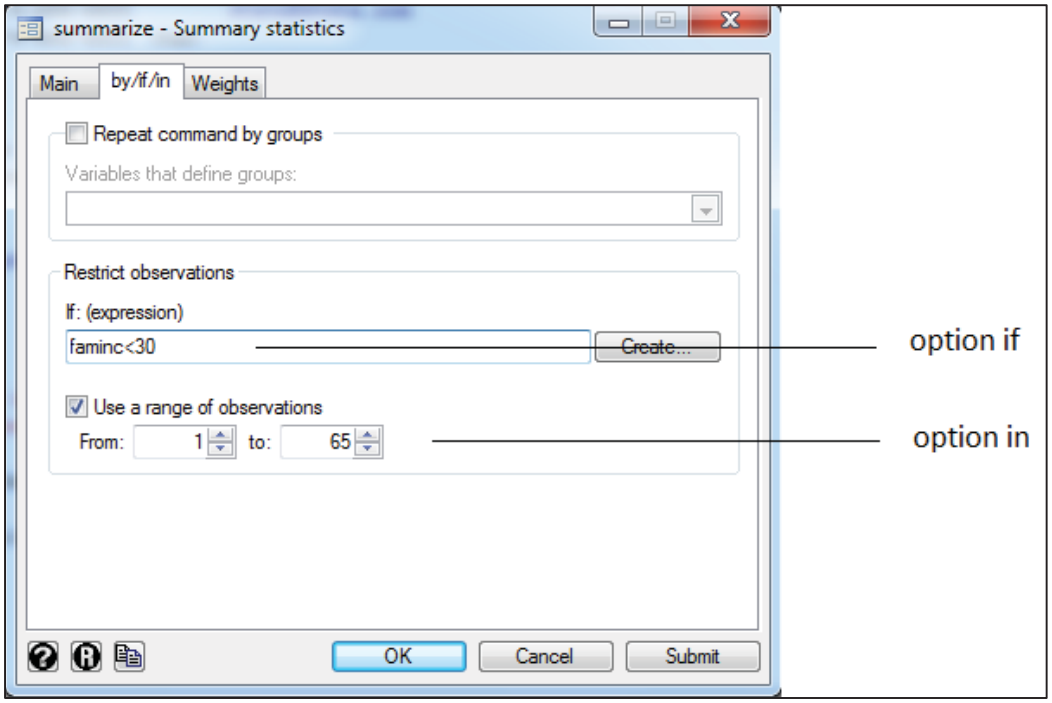
Pour obtenir une combinaison du troisième et quatrième exemple plus haut, et ce, en version détaillée, en reproduisant les tableaux ci-dessus, il est dès lors possible d'obtenir l'output suivant.

```
summarize faminc if faminc<30 in 1/65, detail

                 1988 family income, $1000s
-------------------------------------------------------------
      Percentiles      Smallest
 1%           .5             .5
 5%           .5             .5
10%           .5             .5       Obs                  39
25%          6.5             .5       Sum of Wgt.          39
```

```
50%        13.5                    Mean          14.01282
                       Largest     Std. Dev.     9.338296
75%        22.5           27.5
90%        27.5           27.5     Variance      87.20378
95%        27.5           27.5     Skewness       .125671
99%        27.5           27.5     Kurtosis      1.755575
```

Ce tableau présente les statistiques descriptives détaillées de la variable faminc, pour les observations 1 à 65 ayant des valeurs inférieures à 30. Bien entendu, il est également possible de procéder de façon similaire en passant par la barre d'outils de STATA, soit en sélectionnant Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Summary statistics. Voici la démarche démontrée en images.





1.6.4 La commande tabulate

La commande `tabulate` (ou `tab`) construit un tableau de distribution selon la ou les variables sélectionnées dans le classeur STATA. Encore une fois, les mêmes options explicitées précédemment peuvent être rattachées à cette commande selon les objectifs du lecteur. Nous présenterons ainsi les prémisses entourant la commande `tabulate`. Les prochains exemples utilisent la base de données BWGHT.dta de Wooldridge (2009).

```
tab white

=1 if white |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        299       21.54       21.54
          1 |      1,089       78.46      100.00
------------+-----------------------------------
      Total |      1,388      100.00
```

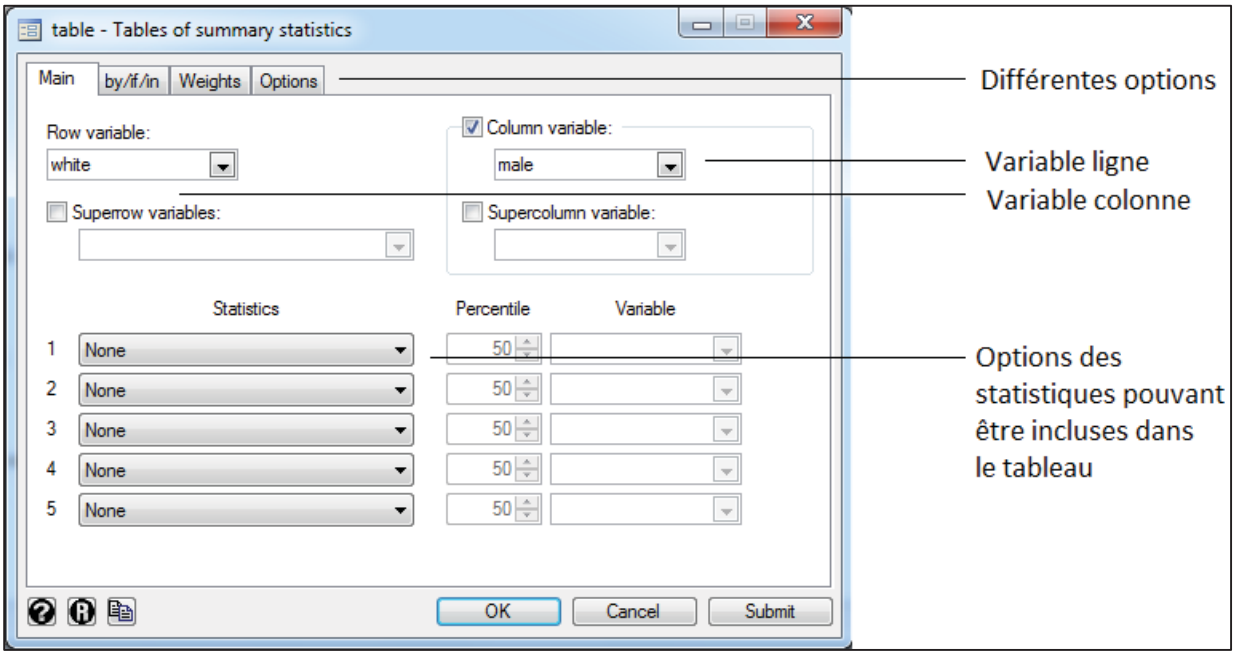
Ce tableau de distribution présente la fréquence (`freq.`), la fréquence relative (`percent`) et la fréquence relative cumulée (`cum`) pour la variable binaire white.

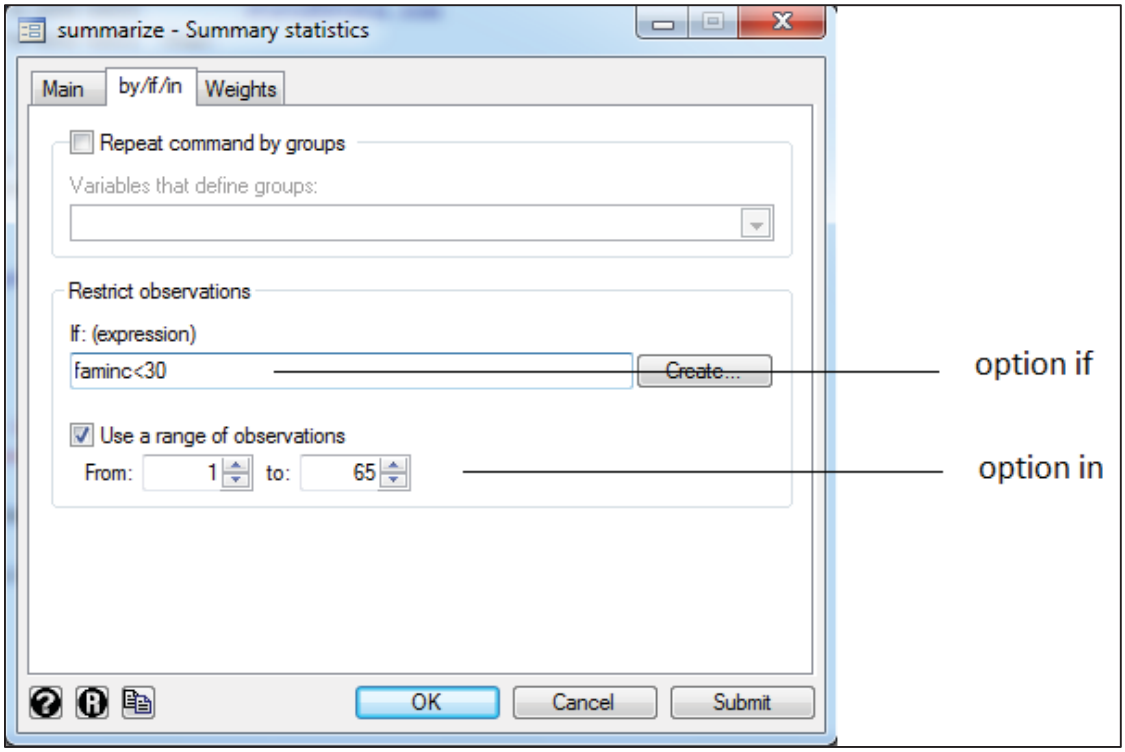
```
tab white male
```

```
     =1 if |   =1 if male child
     white |         0          1 |     Total
-----------+----------------------+----------
         0 |       137        162 |       299
         1 |       528        561 |     1,089
-----------+----------------------+----------
     Total |       665        723 |     1,388
```

Ce tableau présente la matrice de distribution pour les variables binaires white et male. Par exemple, nous dénombrons 137 observations dans la catégorie où white et male prennent conjointement la valeur zéro.

Il est également possible de recréer l'output précédent en passant par la barre d'outils, c'est-à-dire en sélectionnant respectivement Statistics > Tables > Table of summary statistics (table).





Comme dans l'exemple précédent portant sur la commande `summarize`, en répétant respectivement la même procédure que les deux images plus haut, il est possible de reproduire un tableau identique que celui obtenu dans l'output STATA via la commande tabulate.

1.6.5 Coefficient de corrélation

Il est possible de dégager de l'information sur la corrélation et la covariance entre deux ou plusieurs variables. En effet, le lecteur peut utiliser la commande `cor` afin de vérifier le degré de corrélation entre des variables. STATA affichera un output des coefficients de corrélation des variables d'intérêt (à noter qu'il est possible d'insérer plus de deux variables). L'output ci-dessous utilise les données BWGHT.dta de Wooldridge (2009).

```
cor motheduc fatheduc
(obs=1191)

             | motheduc fatheduc
-------------+------------------
    motheduc |   1.0000
    fatheduc |   0.6435   1.0000
```

Ici, le coefficient de corrélation entre les variables motheduc et fatheduc est de 0,6435. Nous pouvons également générer un tableau de covariances entre différentes variables en incorporant l'option `covariance` (ou `cov`) à la fin de la commande `cor`.

```
. cor faminc bwght fatheduc motheduc, cov
(obs=1191)

             |   faminc    bwght fatheduc motheduc
-------------+------------------------------------
      faminc |  322.425
       bwght |  28.9044  405.669
    fatheduc |  22.0362  4.58505  7.51458
    motheduc |  18.5389  2.19417  4.26427    5.844
```

1.7 Enregistrer son travail

Il existe entre autres trois façons d'enregistrer son travail sur STATA. La première consiste à procéder via la barre d'outils en sélectionnant les onglets suivants : **File** > **Save as**. Une fenêtre apparaîtra et le lecteur devra enregistrer son travail en format *.dta. Celui-ci peut d'autant plus procéder de la même manière en tapant la commande `save` selon la procédure suivante :

```
save [nom du fichier]
```

Finalement, il est également possible de copier-coller l'output présenté dans la fenêtre des résultats dans un document *.doc. Le lecteur devra par la suite modifier la police en « Courrier New » sur son document Word afin que l'output en *.doc soit identique à celui affiché sur STATA.

1.8 La commande display

La commande `display` commande à STATA d'effectuer une opération mathématique selon les directives données. Celle-ci donne ainsi la valeur de l'opération demandée. Par exemple, supposons que nous désirons calculer l'addition suivante : 2+4+5. Voyons comment résoudre cette opération triviale à partir de STATA.

```
display (2+4+5)
11
```

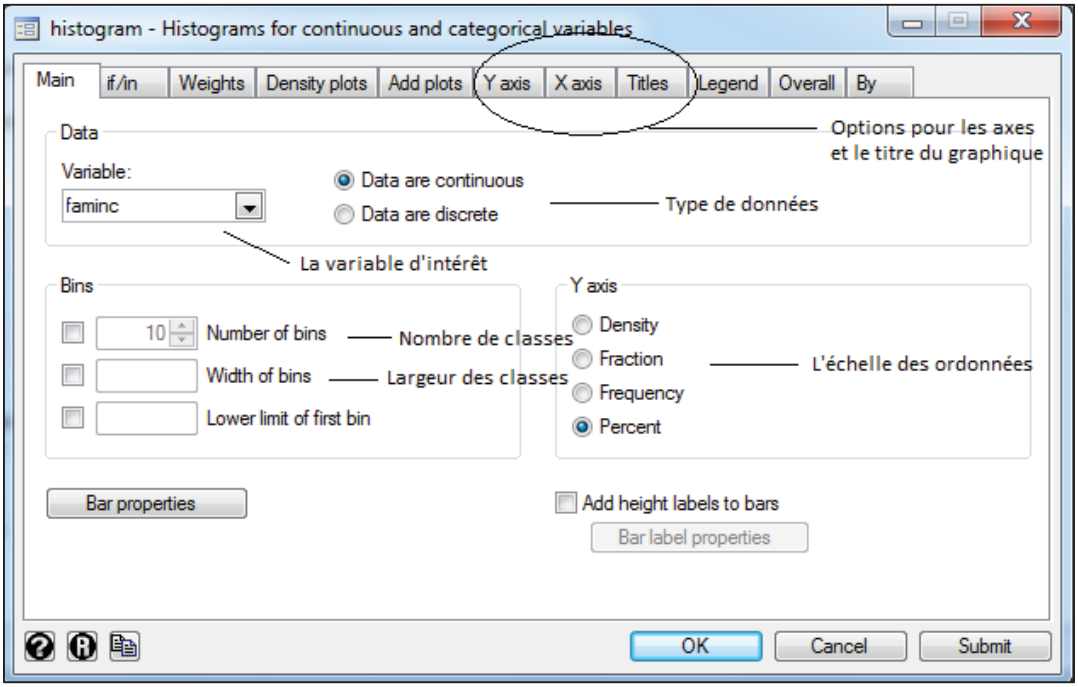
L'output affiche ainsi la valeur de l'opération. Dans notre cas, 2+4+5=11.

1.9 Créer un graphique

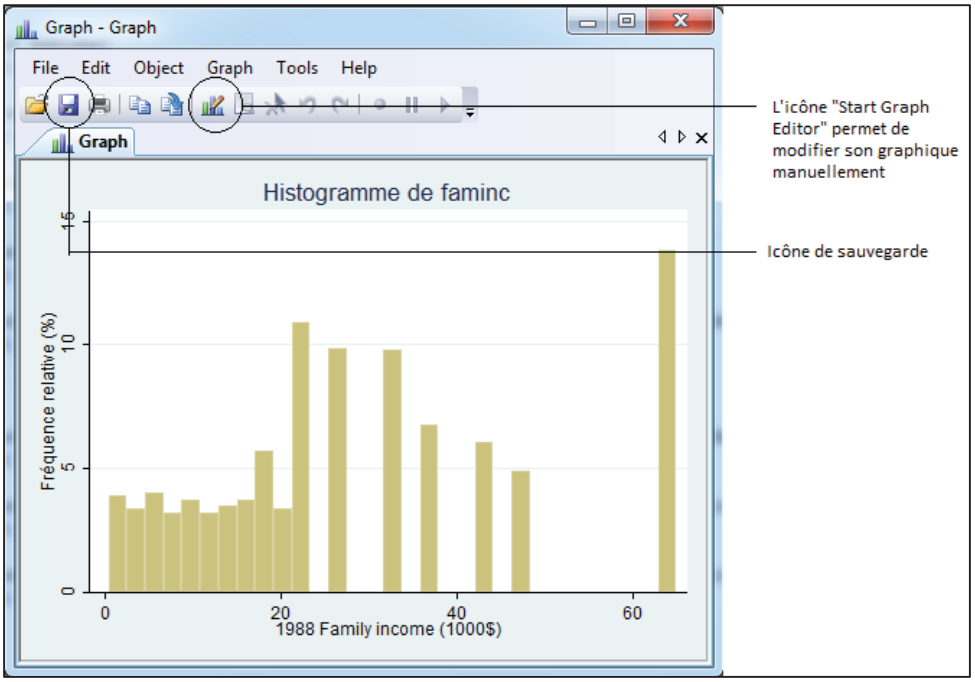
STATA offre la possibilité de construire un vaste éventail de graphiques de tout genre. Cette section présentera au lecteur la façon de construire un histogramme, un nuage de points et un graphique en ligne. Les deux prochains exemples utiliseront le fichier BWGHT.dta de Wooldridge (2009).

1.9.1 Histogramme

Pour construire un histogramme, sélectionnons dans la barre d'outils les onglets Graphics > Histogram. Un menu nommé « histogram – Histograms for continuous and categorical variables » apparaîtra et le lecteur sera amené à sélectionner les options désirées.



Une fois les options sélectionnées et les titres, c'est-à-dire le titre du graphique (Histogramme de faminc), de l'axe des X (1988 Family income (1000$)) et des Y (Fréquence relative (%)), sont complétés, le résultat devrait être celui de l'image ci-dessous.



Bien entendu, il est possible de sauvegarder son graphique en cliquant sur l'icône de sauvegarde. Le graphique sera enregistré en format *.gph. Copier-coller le graphique sur un fichier *.doc est également possible.

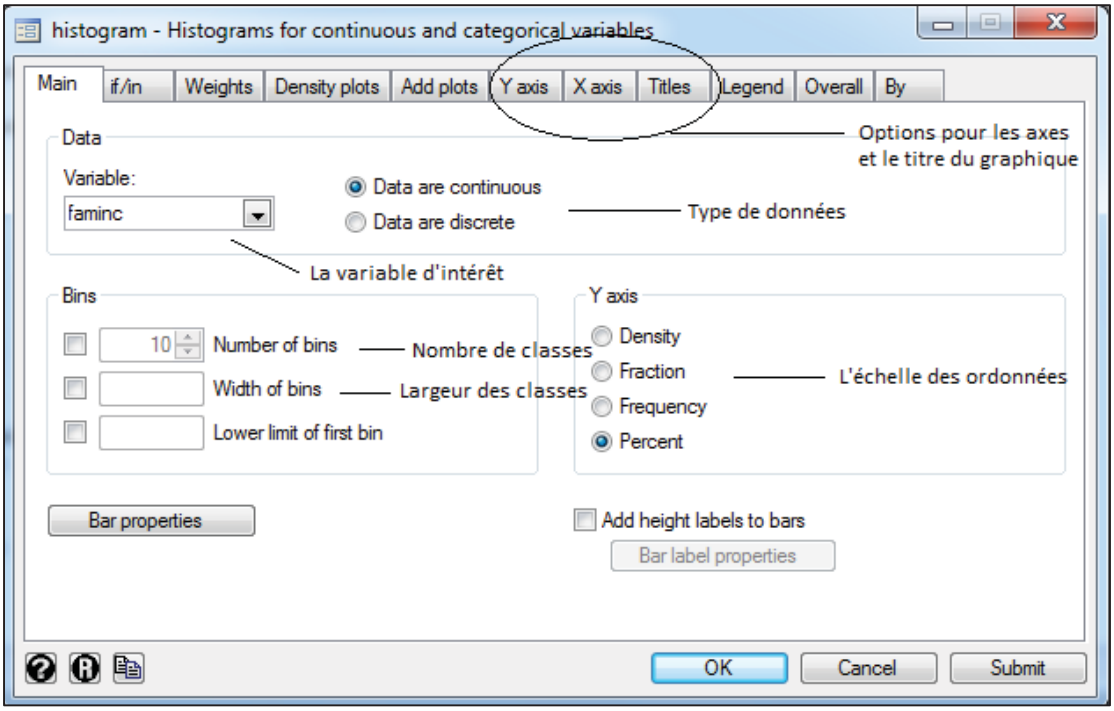
La commande résumant la procédure précédente est la suivante[4]:

```
histogram faminc, percent ytitle(Fréquence relative (%)) xtitle(1988 Family income (1000$))
title(Histogramme de faminc)
```

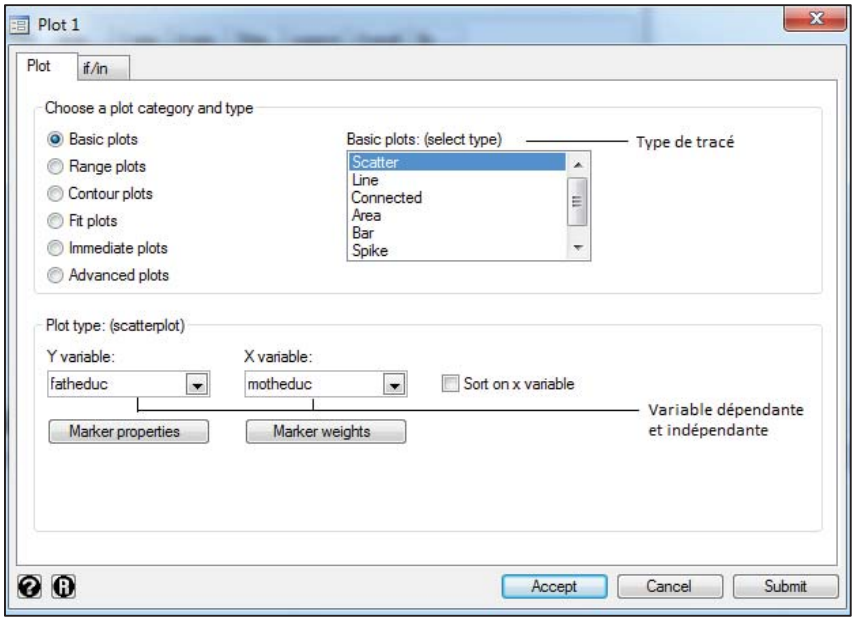
1.9.2 Nuage de points

Pour construire un nuage de points, passer par **Graphics > Twoway graph (scatter, line, etc.)**. Une fois la fenêtre apparue (image ci-dessous), cliquer sur « Create ».

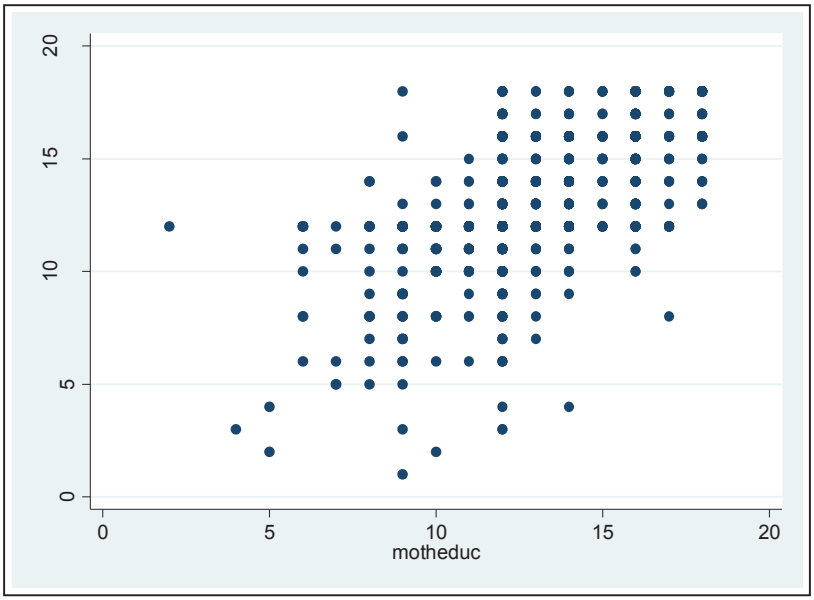
[4] Le lecteur peut toujours construire son graphique en utilisant les commandes appropriées. Ce guide offre les prémisses de base pour construire un graphique via la barre d'outils étant donné sa plus grande simplicité.



Nous pouvons maintenant sélectionner notre variable dépendante et indépendante ainsi que le type de dispersion. Cliquons sur « Scatter » dans l'onglet « Basic plots : (select type) » et sélectionnons les variables d'intérêt. Une fois terminé, appuyons sur « Submit ».



Le graphique devrait être celui ci-dessous

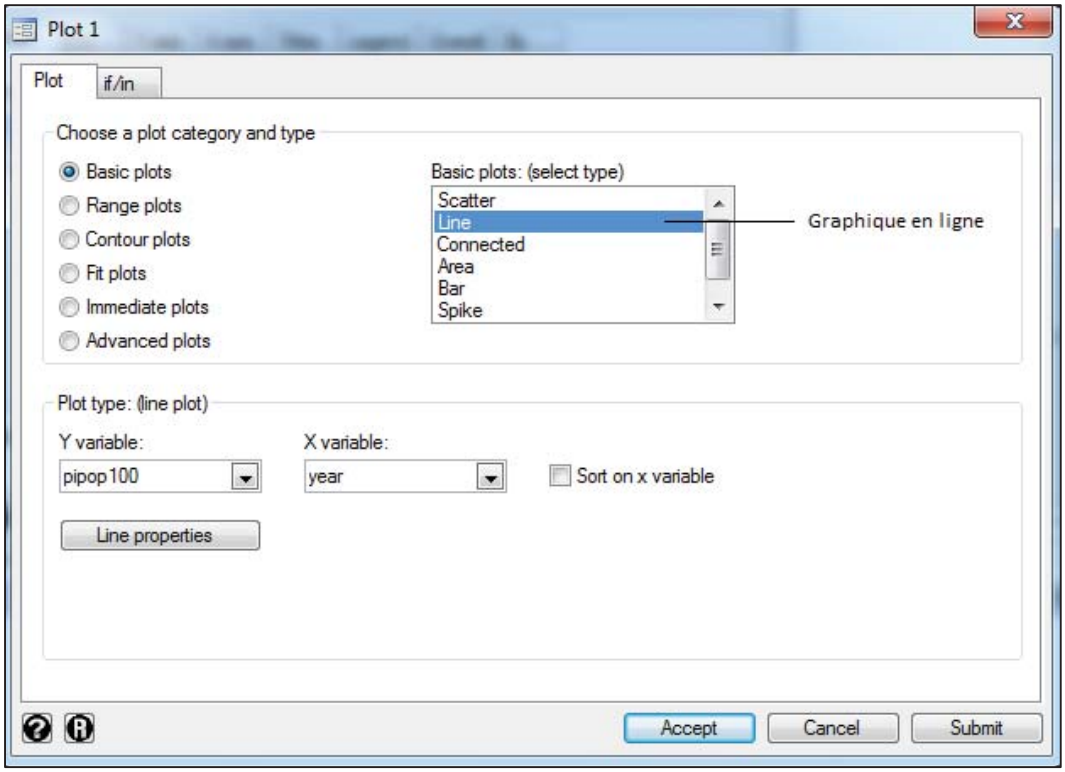


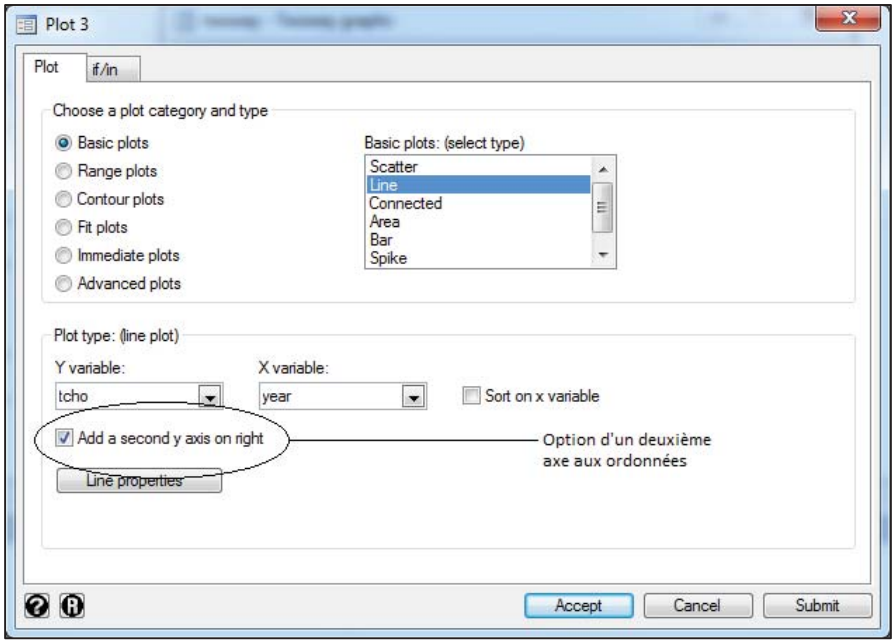
La commande résumant la procédure précédente est la suivante :

```
twoway (scatter fatheduc motheduc)
```

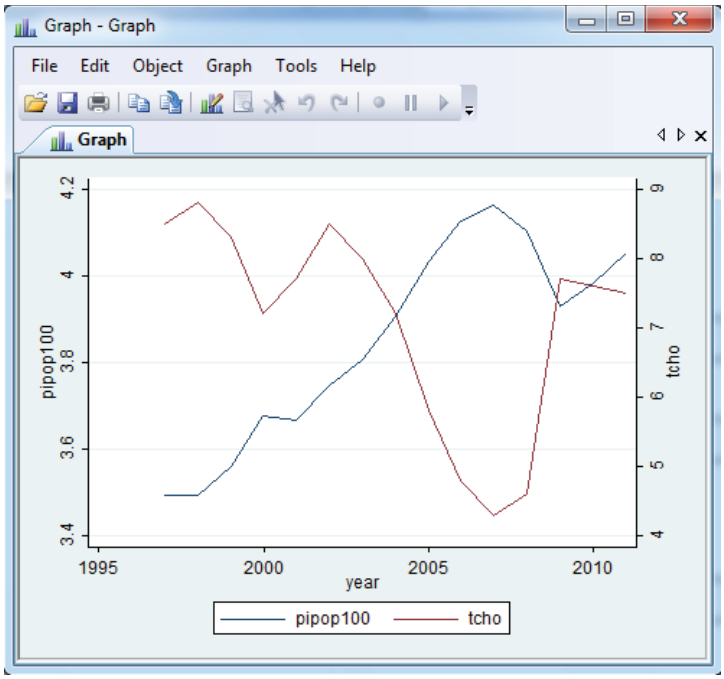
1.9.3 Graphique en ligne

L'utilisation d'un graphique en ligne est spécialement primée dans la manipulation de séries temporelles. La démarche à suivre dans la construction d'un tel graphique avec STATA est quasi similaire à celle utilisée pour un nuage de points. L'exemple suivant utilise les données des tableaux 379-0025, 282-0008 et 051-0001 de CANSIM sur E-STAT. Pour construire le graphique, il suffit de reprendre les mêmes étapes pour le nuage en points, mais en sélectionnant « line » dans la fenêtre « Basic plots : (select type) ». Il est également possible de construire un graphique à deux axes, fort utile lors d'une analyse de corrélation temporelle entre deux variables. Pour procéder, l'utilisateur STATA doit créer deux types de tracés. Dans le deuxième tracé, celui-ci doit cliquer sur la case « Add a second y axis on right ».





Avant d'appuyer sur « OK » dans la fenêtre « twoway – Twoway graphs », il faut s'assurer qu'aucun autre tracé, mis appart ceux que nous avons générés, ne soit activé. Advenant le cas, pour désactiver un tracé de trop, il suffit de sélectionner ledit tracé et de cliquer sur « Disable ». Le graphique obtenu devrait ressembler à l'image ci-dessous.



La commande résumant la procédure précédente est la suivante :

```
twoway (line pipop100 year) (line tcho year, yaxis(2))
```

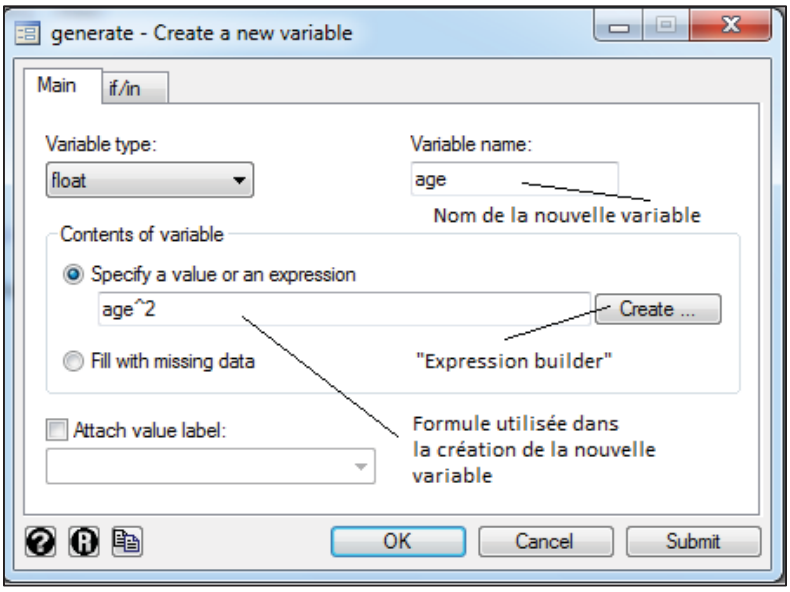
1.10 Générer de nouvelles variables

STATA offre la possibilité de créer de nouvelles variables à partir des variables composant déjà le classeur. Avant d'entamer le sujet, voici une liste d'opérateurs arithmétiques propres à STATA qui aideront l'utilisateur STATA dans la génération de nouvelles variables.

Opérateur	Signification
+	Addition
-	Soustraction
*	Multiplication
/	Division

^ Puissance

Ce guide présentera un exemple à partir des données FERTIL1.dta de Wooldridge (2009). Tout d'abord, dans la barre d'outils, sélectionnons **Data > Create or change data > Create new variable**. Une fenêtre similaire à l'image ci-dessous apparaîtra.



Le lecteur sera interpellé à nommer sa nouvelle variable dans l'encadré « New variable name » et à programmer la formule nécessaire dans la création de cette variable dans la case « Contents of new variable : (expression) ». Noter ici que l'expression **age^2** signifie « élever à la puissance deux toutes les valeurs des observations composant la variable age ». En appuyant sur « OK », la variable age2 sera ajoutée dans la fenêtre « Variables ». La commande raccourcie pour cet exercice s'avère être :

```
generate age2 = age^2
```

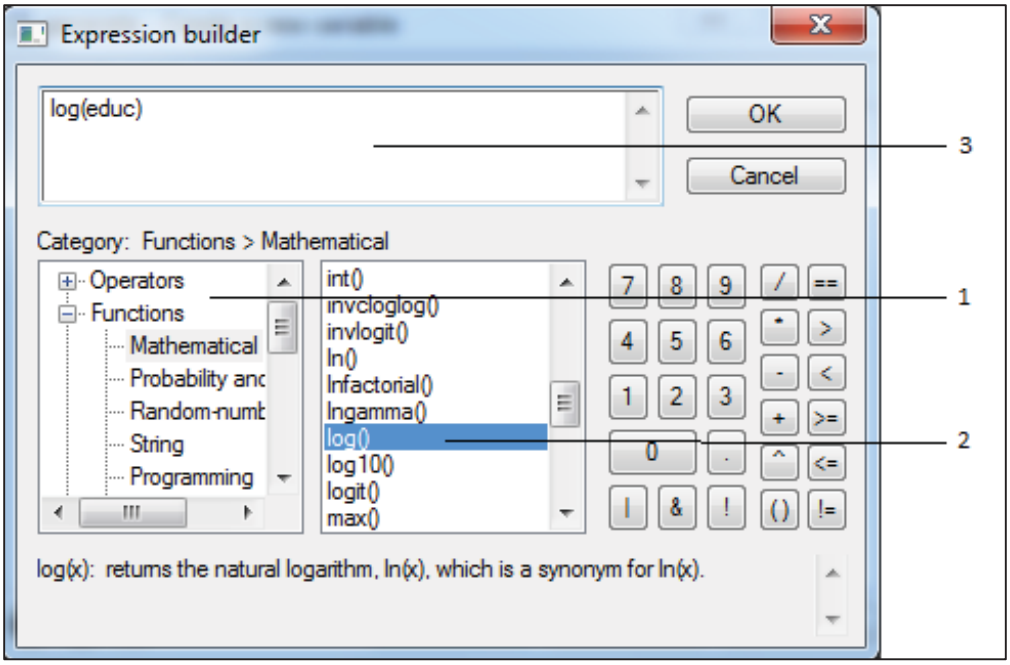
La logique dans cette commande est très simple. En effet, le premier terme « `generate` » (qui peut être remplacé par « `gen` ») est la commande maîtresse pour créer une nouvelle variable. Le second terme, « age2 », représente le nom donné à la nouvelle variable. Le troisième terme, age^2, est la formule utilisée dans la construction de la variable. En d'autres mots, nous demandons à STATA de générer une variable nommée age2 en élevant au carré la variable age

déjà existante dans le classeur. Voici quelques formules supplémentaires très utiles qui aideront le lecteur dans la création de nouvelles variables. On suppose ici que la variable utilisée est « x ».

Formule	**Signification**	**Exemple avec commande « gen »**[5]
log(x)	Logarithme naturel de x	`gen lx = log(x)`
exp(x)	Exponentiel de x	`gen expx = exp(x)`
x^y	Puissance y, y étant un scalaire	`gen xy = x^y`
sqrt(x)	Racine carrée de x	`gen sqrtx = sqrt(x)`
x+2	X plus 2	`gen x2 = x+2`
round(x)	Arrondie x à l'unité près	`gen roundx = round(x)`
x*y	Interaction entre les variables x et y	`gen xy = x*y`

Le lecteur doit se souvenir qu'il existe une multitude de formules dans le registre de STATA. Heureusement, STATA offre un menu complet portant sur toutes les formules disponibles dans le programme. Pour y accéder, nous sur la barre « create » du menu nommé « generate – Create a new variable » portant sur la création de nouvelles variables (dernière image). Un tableau, nommé « Expression builder », apparaîtra.

[5] Noter que le nom donné à la variable créée est subjective.



L'utilisateur STATA devra alors procéder ainsi :

1. Choisir la catégorie de formule. Dans l'exemple de ce guide, sélectionner la catégorie « Mathematical ».
2. Sélectionner la transformation (la formule en question). Dans l'exemple de ce guide, choisir log().
3. En double-cliquant sur la formule, l'expression log(x) apparaîtra dans la fenêtre supérieure. Réécrire le nom de la variable utilisée à la place de « x ». Dans l'exemple de ce guide, changer « x » pour « educ ». Appuyer sur OK.

Ce faisant, dans l'encadré « Contents of new variable : (expression) », la formule log(educ) sera ajoutée. Il ne faut pas oublier de donner un nom à cette nouvelle variable dans le cadrage « New variable name : » et le tour est joué.

1.11 Générer des variables avec un lag ou un lead

Lorsque nous travaillons avec des séries temporelles, il est très fréquent de vouloir générer des variables avec un lag ou un lead, c'est-à-dire avec un nombre de périodes n de retard ou d'avance. Voici les commandes à utiliser selon le contexte dans lequel sont ordonnées les variables d'intérêt. Les exemples suivants utilisent les données du tableau 379-0025 de CANSIM

sur E-STAT. Nous pouvons procéder en prenant en compte le nombre d'observations ou en créant une variable de temps.

En utilisant le nombre d'observations comme référence, voici les commandes d'intérêt.

Exemple	Signification
`gen piblag=pib[_n-1]`	Nous créons un retard d'une observation dans la variable pib, nouvelle variable que nous nommons piblag
`gen piblag2=pib[_n-2]`	Nous créons un retard de deux observations dans la variable pib, nouvelle variable que nous nommons piblag2
`gen piblea=pib[_n+1]`	Nous créons un lead d'une observation dans la variable pib, nouvelle variable que nous nommons piblea

En générant une variable de temps, nous pouvons suivre les étapes suivantes :

1. Générer une série chronologie de n observations :
   - `gen t=_n`
2. Commander à STATA de programmer la variable « t » comme une série temporelle :
   - `tsset t`
3. Générer une variable avec un retard avec la spécification « L. » [6]:
   - `gen piblag=L.pib`

[6] On peut également ordonner un retard de deux périodes en insérant L2.pib, de trois périodes avec L3.pib et ainsi de suite.

1.12 Éliminer et renommer des variables

Il existe des commandes raccourcies sur STATA afin d'éliminer, garder et renommer des variables. Voici un tableau résumant les fonctions de ces commandes (noter que x représente le nom d'une variable quelconque).

Commande	Signification	Exemple
`drop`	Éliminer une variable	drop x
`keep`	Garder une variable (et élimine toutes les autres)	keep x
`rename`	Renommer une variable	rename x nouveau x (où nouveau x est le nouveau nom de la variable x)
`label`	Modifier le label ou la description de la variable	label x "la variable x"

Évidemment, il est toujours possible d'utiliser ces commandes avec des options spéciales. Par exemple, si nous supposons que le classeur STATA possède les variables x et y, advenant que nous veuillons éliminer seulement les observations supérieures ou égales à p, p étant un nombre réel, nous pouvons insérer la commande suivante selon les opérateurs logiques vus précédemment :

```
drop if x>=p
```

Plusieurs spécificités peuvent dès lors être considérées selon les objectifs du lecteur. Pour de plus amples détails, ce dernier peut taper la commande `help drop`, `help keep`, `help rename` ou `help label` selon les besoins.

1.13 Le codage de variables

Le codage de variables est une étape forte importante dans la construction d'une base de données. La commande `replace` permet à l'utilisateur STATA de modifier des variables existantes dans son classeur. Par exemple, on suppose un classeur STATA qui ressemble au tableau ci-dessous :

Observations	**VariableA**	**VariableB**
1	1	9
2	1	9
3	9	1
4	9	1
5	1	9
6	7	7

Dans cet exemple, nous cherchons à remplacer les valeurs « 9 » et « 7 » des variables « variableA » et « variableB » par la valeur « 2 ». Pour apporter de telles modifications, nous utilisons la commande replace.

```
replace variableA=2 if variableA==9 & variableA==7
replace variableB=2 if variableB==9 & variableB==7
```

Ici, pour la première commande, l'interprétation se ferait comme suit : « remplacer par la valeur 2 toutes les valeurs 9 et 7 de la variable variableA ». L'interprétation est identique pour la deuxième commande. Encore une fois, selon l'objectif visé, libre à l'utilisateur STATA d'employer d'autres opérateurs logiques dans le codage de ses variables.

1.14 Les commandes sort

Il existe une commande utile afin d'économiser du temps dans la réorganisation de données, spécialement lorsque nous désirons classer en ordre croissant ou décroissant une série quelconque. Cette commande s'appelle `sort`. Voici deux exemples qui permettront au lecteur de mieux se familiariser avec cette commande. Nous supposons un classeur de la forme suivante. Male est une variable binaire qui prend la valeur 1 si l'observation est un homme et 0 autrement. educ est le nombre d'années de scolarité en années.

Observation	**male**	**educ**
1	1	12
2	0	17
3	1	15
4	0	13

Nous désirons ici classer en ordre croissant les valeurs composant la variable male (en veut classer en premier les femmes et en second les hommes). En utilisant la commande suivante, nous obtenons le nouveau classeur suivant :

```
sort male
```

Observation	**male**	**educ**
1	0	12
2	0	17
3	1	15
4	1	13

À l'inverse, on peut également classer en ordre décroissant les valeurs de la variable male en insérant la commande suivante :

```
gsort -male.
```

On peut également effectuer le même processus en conditionnant en ordre croissant les valeurs de educ au classement de male.

```
sort male educ
```

Observation	**male**	**educ**
1	0	13
2	0	17
3	1	12
4	1	15

Nous pouvons également refaire le même processus, mais cette fois-ci en conditionnant en ordre décroissant les valeurs de educ au classement de male en ajoutant un « - » avant educ.

```
sort male -educ
```

1.15 Organisation des données

Une base de données peut être organisée de façons différentes. Parmi celles-ci, nous dénotons deux formes particulières : le format « wide » et le format « long ». Voici deux tableaux caractérisant ces deux formats. Ici, market est une variable binaire.

Base de données : Format wide

observations	market	price2000	price2001	price2002
1	1	6.52	7	11.04
2	0	20.59	18.65	19.8
3	1	25.01	27.96	24.66

Base de données : Format long

observations	year	market	price
1	2000	1	6.52
1	2001	1	7
1	2002	1	11.04
2	2000	0	20.59
2	2001	0	18.65
2	2002	0	19.8
3	2000	1	25.01
3	2001	1	27.96
3	2002	1	24.66

STATA offre la possibilité de passer d'un format à l'autre grâce à la commande `reshape`. Avec cet exemple, afin de passer du format wide au format long, nous utilisons la commande suivante `reshape`.

```
reshape long price, i(observations) j(year)
```

Il est nécessaire que le suffixe de la variable longitudinale soit dans un format temporel approprié. Il est recommandé aux utilisateurs de se référer directement au guide du logiciel.

Règle générale, `i()` désigne la ou les variables que l'on souhaite garder dans le format actuel (dans le cas précédent, nous voulons garder « observations » en format long) et `j()` représente la ou les variables que nous désirons transformer dans l'autre format (dans le cas précédent, nous souhaitons créer la variable year en format long).

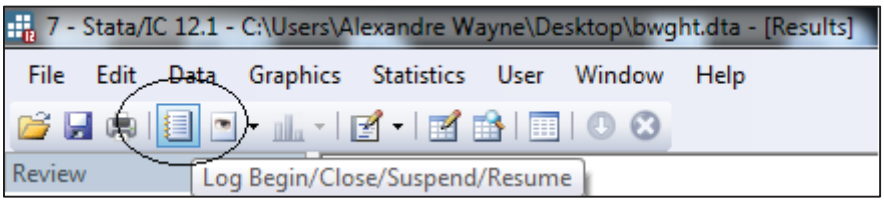
Nous procédons par l'inverse du format long au format wide via cette opération.

```
reshape wide price, i(observations) j(year)
```

1.16 Log file

Le Log file contient les outputs que présente STATA sous la forme d'un fichier de sortie, par exemple de format PDF. Cela peut s'avérer fort utile dans la remise d'un travail économétrique où le .do file et les outputs du travail exécuté sont demandés. La prochaine section présentera les étapes à suivre afin de préparer un Log file en utilisant la base de données BWGHT.dat de Wooldridge (2009).

1. Pour créer un Log file, appuyons sur l'icône encerclée présentée dans l'image ci-dessous. Il est également possible de procéder via la barre d'outils en sélectionnant : **File > Log > Begin.**



2. Donnons un nom au fichier *.smcl (Log file) et appuyons sur « enregistrer ».
3. La commande « log using "emplacement du fichier *.smcl" » ainsi que l'output ci-dessous apparaîtront dans la fenêtre des résultats de STATA.

```
. log using "C:\Users\Alexandre Wayne\Desktop\[nomdufichier].smcl"
-----------------------------------------------------------------------------------------------------
---------------------------------------
      name:  <unnamed>
       log:  C:\Users\Alexandre Wayne\Desktop\[nomdufichier].smcl
  log type:  smcl
 opened on:  20 Jul 2012, 15:04:01
```

4. Afin d'illustrer le concept du Log-File, insérons la commande suivante :

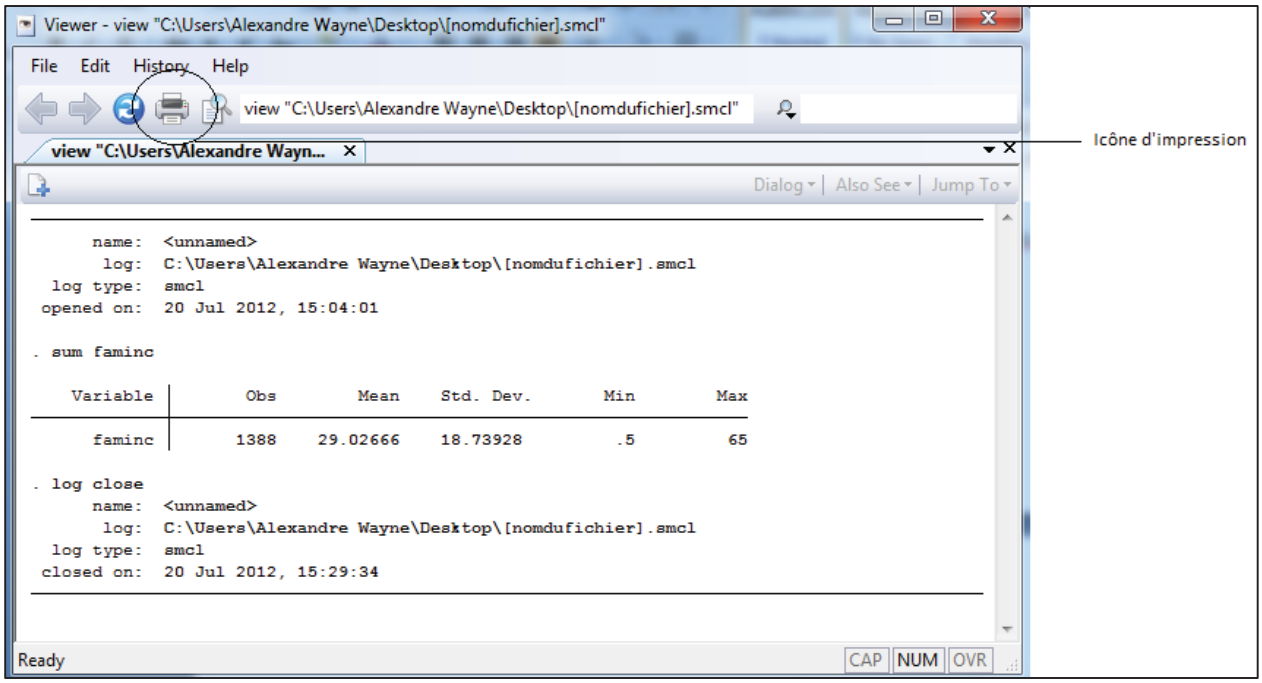
```
sum faminc
```

Un output STATA identique à l'exemple de la section 8.3 de ce guide devrait apparaître dans la fenêtre des résultats.

5. Cliquons sur l'icône de l'image précédente et sélectionnons la case « Close log file » afin d'empêcher STATA d'ajouter d'autres outputs sur le log file. STATA affichera l'output suivant dans la fenêtre des résultats ainsi que la commande « log close ».

```
      name:  <unnamed>
       log:  C:\Users\Alexandre Wayne\Desktop\[nomdufichier].smcl
  log type:  smcl
 closed on:  20 Jul 2012, 15:29:34
```

Il est par la suite possible de visionner son log file en passant par **File** > **Log** > **View** via la barre d'outils et en insérant l'emplacement du fichier *.smcl qui correspond au Log file crée dans la fenêtre « Choose File to View ». La fenêtre suivante devrait dès lors apparaître après ces étapes.



L'utilisateur STATA peut également imprimer son *log file* en cliquant sur l'icône encerclée de l'image ci-dessus.

Voici une liste exhaustive des commandes couramment utilisée afin de créer un log file.

Commande	Signification

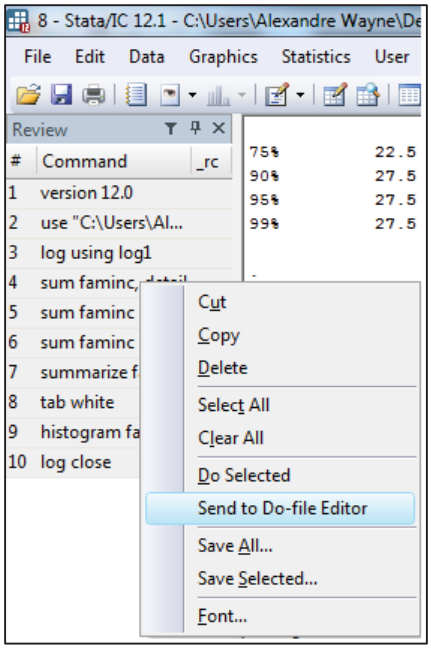
`log using [nomdufichier]`	Ouverture d'un log file
`log close`	Fermeture du log file [nomdufichier]
`print [nomdufichier].smcl`	Imprimer le log file [nomdufichier]
`log using [nomdufichier], replace`	Ouvrir un log file et le remplacer par un autre log file de même nom
`log using [nomdufichier], append`	Ouvrir un log file existant et ajouter de nouveaux résultats à la fin du fichier

Notons que les fichiers .do file incorporés dans ce guide peuvent servir d'exemple

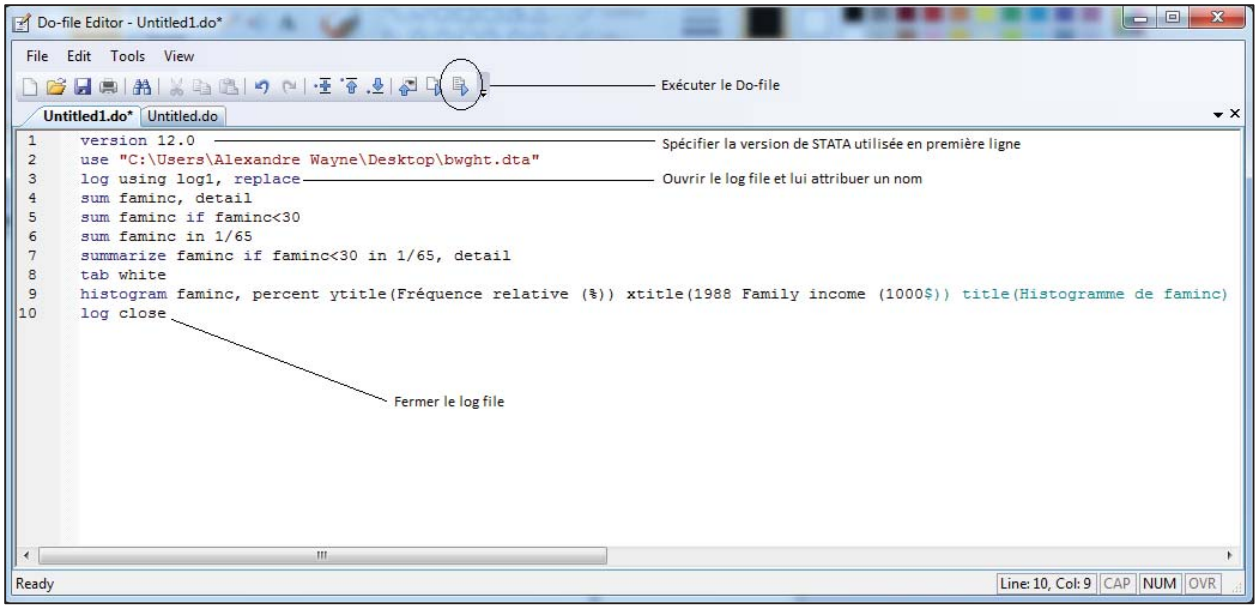
1.17 Do-file

La création d'un Do-file sur STATA s'avère fort utile lorsqu'un travail contient plusieurs commandes dans la fenêtre « Review ». Grâce au Do-file, STATA offre la possibilité d'enregistrer l'ensemble des commandes effectuées. Voici la démarche à suivre pour procéder.

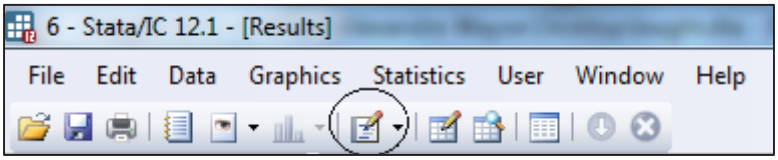
1. Dans la fenêtre « Review », avec le clique-droit de la souris du portable, sélectionnons « Select all » et refaire la même opération en cliquant sur « Send to Do-file Editor ». Nous transférons ainsi l'ensemble de nos commandes effectuées dans un menu spécial conçu pour la gestion de Do-file nommé « STATA Do-File Editor ». Notons également qu'il est possible de sélectionner individuellement des commandes spécifiques dans la fenêtre « Review ».



2. La fenêtre « Do-file Editor » apparaîtra. Celle-ci devrait contenir la liste de l'ensemble des commandes sélectionnées dans la fenêtre « Review ». Ajoutons la version de STATA utilisée dans la première ligne et enregistrons le Do-file (en format *.do) en cliquant sur l'icône de la disquette. La commande `about` donne également de l'information sur la version de STATA utilisée, advenant le cas où l'utilisateur STATA ne la connaît pas. Celui-ci peut également exécuter son Do-file en appuyant sur l'icône encerclée de l'image ci-dessous. Notons qu'un Do-file ne pourra être exécuté si aucune base de données propre au fichier n'est téléchargée préalablement. Ce faisant, il faut s'assurer que le Do-file contienne la commande de téléchargement de la base de données en question.



Pour télécharger un Do-file à partir de STATA, nous pouvons accéder à la fenêtre « STATA Do-File Editor » en cliquant sur l'icône encerclée de l'image ici-bas. Une fois la fenêtre apparue, cliquons sur l'icône de dossier et sélectionnons le fichier *.do désiré.



L'utilisateur STATA peut également copier-coller les commandes de son travail dans un fichier *.doc afin d'enregistrer son Do-file. Pour démarrer son Do-file, celui-ci n'aura qu'à copier-coller du fichier *.doc à la fenêtre « Command » le contenu de son Do-file.

1.18 Tests sur la moyenne, la variance et la distribution des variables

En statistique, nous sommes fréquemment amenés à tester des paramètres (moyennes et variances) de la population en inférant à partir des données qui composent notre échantillon. Le lecteur est amené à lire plus en détail la théorie couvrant ces tests d'hypothèse en lisant Newbold, Carlson et Thorne (2010).

1.18.1 Inférence sur le paramètre d'une population suivant une distribution normale

Voici la procédure à suivre sur STATA dans le cas où nous désirons inférer sur le paramètre d'une population.

Syntaxe:

```
ttest varname == # [if] [in] [, level(#)]
```

où `varname` est la variable à tester
`#` est la valeur à estimer
`[if] [in]` sont des contraintes dans l'échantillon
`[, level(#)]` est le niveau de confiance de l'intervalle de confiance

Menu :

Statistics > Summaries, tables, and tests > Classical tests of hypotheses > One-sample mean-comparison test

À partir des données BWGHT.dta de Wooldridge (2009), testons si la moyenne de faminc dans la population est égale à la moyenne de l'échantillon soustrait d'un demi-écart type.

1. Trouvons la moyenne et l'écart type de faminc dans l'échantillon et calculons $\overline{faminc} - \frac{1}{2}\sigma_{faminc}$.

```
sum faminc

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      faminc |      1388    29.02666    18.73928         .5         65

display (29.02666-0.5*18.73928)
19.65702
```

2. Testons $H_0: E(faminc) = 19{,}66$.

```
ttest faminc=19.65702


One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  faminc |    1388    29.02666    .5029888    18.73928    28.03996    30.01336
------------------------------------------------------------------------------
    mean = mean(faminc)                                           t =  18.6279
Ho: mean = 19.65702                              degrees of freedom =     1387
```

```
Ha: mean < 19.65702          Ha: mean != 19.65702          Ha: mean > 19.65702
Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

STATA affiche l'output des hypothèses alternatives des tests unilatéraux (Ha: moyenne de l'échantillon $<$ 19.65702 et Ha: moyenne de l'échantillon $>$ 19.65702) et du test bilatéral (Ha: la moyenne de l'échantillon $\neq$ 19.65702). Nous voyons que la statistique t est de 18,6279. Excepté pour l'hypothèse alternative Ha: moyenne $<$ 19.65702, nous avons suffisamment d'évidence statistique pour rejeter l'hypothèse nulle à un seuil de confiance de 1,00 % (voir les p-valeur, soit Pr($|T| > |t|$)). Nous pouvons donc conclure statistiquement que la moyenne de faminc dans la population est significativement plus grande que 19,65702.

1.18.2 Estimation de la différence de moyenne de deux populations suivant une loi normale

Échantillons indépendants

Voici la procédure à suivre sur STATA dans le cas où nous désirons inférer sur le la différence de moyennes de deux populations suivant une loi normale dans le cas d'échantillons indépendants.

Syntaxe

```
ttest varname1 == varname2 [if] [in], unpaired [unequal level(#)]
```

où `varname1` et `varname2` sont les variables d'intérêt
`[if] [in]` sont des contraintes dans l'échantillon
`unpaired` suppose une variance identique pour les deux populations
`[, level(#)]` est le niveau de confiance de l'intervalle de confiance

Menu :

Statistics $>$ Summaries, tables, and tests $>$ Classical tests of hypotheses $>$ Two-sample mean-comparison test

Voici un exemple tiré de Newbold, Carlson et Thorne (2010). Les résidents de Orange City se plaignent qu'ils paient des contraventions pour excès de vitesse sur la route plus importantes que leurs voisins de la ville de DeLand. Le maire de Orange City accepte d'étudier le problème

afin d'évaluer si les plaintes sont appuyées. 2 échantillons indépendants sont tirés de contraventions payées depuis les trois derniers mois par les résidents des deux villes. Voici les montants chiffrés.

Orange City (variable OC)	DeLand (variable DL)
100	95
125	87
135	100
128	75
140	110
142	105
128	85
137	95
156	
142	

On cherche à connaître s'il existe une différence dans le coût des contraventions entre les deux villes. On assume une variance égale dans les deux populations (les deux villes). Procédons sur STATA.

```
ttest OC=DL, unpaired

Two-sample t test with equal variances
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
      OC |      10       133.3    4.669166     14.7652    122.7376    143.8624
      DL |       8          94    4.022259    11.37667    84.48887    103.5111
---------+--------------------------------------------------------------------
combined |      18    115.8333    5.639641    23.92697    103.9347    127.7319
---------+--------------------------------------------------------------------
    diff |                39.3      6.3508                25.83691    52.76309
------------------------------------------------------------------------------
    diff = mean(OC) - mean(DL)                                    t =   6.1882
```

```
Ho: diff = 0                                     degrees of freedom =       16

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Sur la base de ces résultats, nous pouvons conclure avec suffisamment d'évidence statistique que les résidents de Orange City paient des contraventions plus chères que ceux de la ville de DeLand (intervalle de confiance positif et p-valeur=0,00 pour l'hypothèse alternative Ha : diff > 0).

1.18.3 Estimation de la différence de moyenne de deux populations suivant une loi normale

Échantillons dépendants.

Voici la procédure à suivre sur STATA dans le cas où nous désirons inférer sur le la différence de moyennes de deux populations suivant une loi normale dans le cas d'échantillons dépendants.

Syntaxe :

```
ttest varname1 == varname2 [if] [in] [, level(#)]
```

où `varname1` et `varname2` sont les variables d'intérêt
`[if] [in]` sont des contraintes dans l'échantillon
`[, level(#)]` est le niveau de confiance de l'intervalle de confiance

Menu :

Statistics > Summaries, tables, and tests > Classical tests of hypotheses > Mean-comparison test, paired data

Voici un exemple tiré de Newbold, Carlson et Thorne (2010). Nous cherchons à vérifier l'efficacité d'un programme de lecture dans le cadre de la commission scolaire STATA. Chaque élève a dû remplir un examen avant et après la mise en place du programme. Une note plus élevée au deuxième examen signifie que le programme est efficace. Nous tirons aléatoirement un échantillon de 7 enfants dans le bassin d'élèves composant la commission scolaire STATA.

Élève	**Note examen avant programme (variable**	**Note examen après programme (variable**

	pretest)	posttest)
1	**40**	**48**
2	**36**	**42**
3	**32**	**.**
4	**38**	**36**
5	**.**	**43**
6	**33**	**38**
7	**35**	**45**

L'élève numéro 3 a changé de commission scolaire lorsque le programme a été implanté. L'élève numéro 5 est arrivé dans la commission scolaire STATA lors de l'implantation du programme. Testons s'il existe une différence entre les moyennes des deux populations (avant et après la mise en place du programme).

```
ttest pretest=posttest


Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 pretest |       5        36.4    1.208305    2.701851    33.04521    39.75479
posttest |       5        41.8         2.2     4.91935    35.69182    47.90818
---------+--------------------------------------------------------------------
    diff |       5        -5.4    2.039608    4.560702   -11.06286    .2628591
------------------------------------------------------------------------------
     mean(diff) = mean(pretest - posttest)                         t =  -2.6476
 Ho: mean(diff) = 0                              degrees of freedom =        4

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.0286         Pr(|T| > |t|) = 0.0571          Pr(T > t) = 0.9714
```

L'interprétation de l'output STATA se fait comme dans la section 20.2. En regardant les résultats ci-dessus, nous avons suffisamment d'évidence statistique pour rejeter l'hypothèse nulle qu'il n'existe aucune différence entre les résultats de l'évaluation avant et après l'implantation du programme de lecture. Nous pouvons donc conclure statistiquement qu'il existe une différence significative au seuil de confiance de 5,0 % à l'avantage des notes de l'examen après la

mise en place du programme (p-valeur=0,0286 pour Ha : diff < 0). Le programme s'avérait donc efficace.

2 Régression linéaire

2.1 Introduction

Ce chapitre présente une introduction sur la base théorique nécessaire dans l'étude statistique de données quantitatives en économétrie ainsi que son application sur le logiciel STATA. Les modèles de régressions linéaires simples et multiples ainsi que les différents tests de validité seront développés théoriquement ainsi que concrètement sur STATA afin d'initier et former les étudiants aux cours d'économétrie du baccalauréat en sciences économiques de l'Université de Montréal. Des exemples issus des données de Wooldridge (2009) seront mis à la disposition des étudiants à la fin de chaque sous-section.

2.2 La structure des données en science économique

Les économistes travaillent habituellement avec quatre types de données en économétrie : les données en coupe, les séries temporelles, les données mixtes (« *pooled cross section* ») et les données longitudinales (de panel). La connaissance du type de données que nous manipulons est essentielle avant toute modélisation. L'utilisateur STATA peut utiliser la commande `describe` afin d'avoir une vue d'ensemble sur les données qui composent son classeur. Les prochaines sous-sections présenteront au lecteur un exemple de configuration de base de données propre à chaque type de données utilisées.

2.2.1 Les données en coupe

Les données en coupe sont fortement utilisées en science économique, spécialement en économie du travail, en organisation industrielle, en économie de la santé, etc. Celles-ci comprennent des unités observées dans un même instant de temps échantillonnées aléatoirement et indépendamment dans une population donnée : le salaire annuel en 2008 de travailleurs dans le secteur manufacturier au Canada, le nombre de cigarettes consommées par semaine par des étudiants, etc. Plusieurs précautions dans l'utilisation de ces données doivent cependant être prises (nous verrons plus loin dans cette partie comment vérifier ces anomalies sur STATA) :

- Biais de sélection issu d'un échantillonnage non aléatoire
- Biais d'erreur de mesure
- Hypothèse d'indépendance non respectée

Un classeur composé de données en coupes s'avère la forme la plus simple. Chaque ligne représente une unité observable. L'ordre de présentation, c'est-à-dire le classement d'une observation, n'a d'ailleurs pas d'importance. Par exemple, en jetant un coup d'œil au classeur ci-dessous, la première observation (la première ligne) représente un individu dont le revenu familial (faminc) est de 13500 $, de sexe masculin (male = 1), de race blanche (white = 1), etc.

Data Editor (Edit) - [bwght]

File Edit View Data Tools

faminc[1] 13.5

	faminc	cigtax	cigprice	bwght	fatheduc	motheduc	parity	male	white	cigs
1	13.5	16.5	122.3	109	12	12	1	1	1	0
2	7.5	16.5	122.3	133	6	12	2	1	0	0
3	.5	16.5	122.3	129	.	12	2	0	0	0
4	15.5	16.5	122.3	126	12	12	2	1	0	0
5	27.5	16.5	122.3	134	14	12	2	1	1	0
6	7.5	16.5	122.3	118	12	14	6	1	0	0
7	65	16.5	122.3	140	16	14	2	0	1	0
8	27.5	16.5	122.3	86	12	14	2	0	0	0
9	27.5	16.5	122.3	121	12	17	2	0	1	0
10	37.5	16.5	122.3	129	16	18	2	0	1	0
11	27.5	16.5	122.3	101	12	16	2	1	0	0
12	27.5	16.5	122.3	133	16	15	1	1	1	0
13	6.5	16.5	122.3	61	.	12	3	1	0	0
14	10.5	16.5	122.3	104	12	12	1	1	1	0
15	12.5	16.5	122.3	92	7	12	1	1	0	0
16	17.5	16.5	122.3	122	13	13	1	1	1	0
17	42.5	16.5	122.3	159	18	16	1	1	1	0
18	4.5	16.5	122.3	154	.	11	1	0	0	0
19	32.5	16.5	122.3	120	16	16	1	1	0	0
20	6.5	16.5	122.3	138	.	11	2	0	0	0
21	47.5	16.5	122.3	127	16	17	1	0	1	0

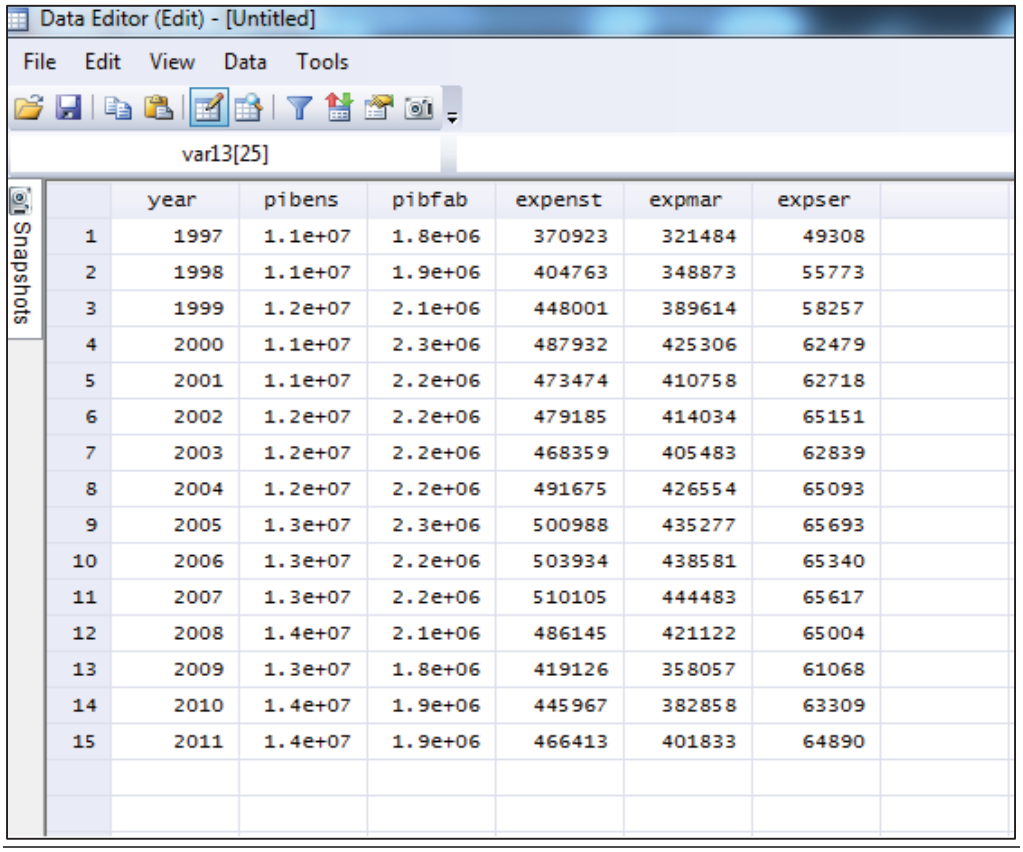
2.2.2 Séries temporelles

Les séries temporelles consistent à un ensemble d'observations d'une ou plusieurs variables recensées dans un intervalle de temps. Ce type de données est spécialement utilisé dans les domaines de la finance, de l'économie financière, de la conjoncture économique, etc. Par exemple, nous pouvons nous intéresser au taux de rendement quotidien d'une obligation 10 ans canadienne de 2009 à 2011, au cours de l'action ordinaire de la firme XY de 1990 à 2010, etc.

Encore une fois, il faut rester prudent sur l'utilisation de ces données notamment sur les points suivants (ces questions seront davantage développées dans la troisième partie de ce guide) :

- Est-ce que les observations sont indépendantes dans le temps?
- Est-ce que l'échantillon collecté est suffisamment grand?

La conception d'un classeur pour des séries temporelles reste relativement simple. Effectivement, chaque observation est classée en ordre chronologique où la variable de temps est placée dans la première colonne. L'exemple ci-dessous présente des données sur le PIB en dollars enchaînés de 2002 (en millions de $) et les exportations canadiennes en dollars enchaînés de 2002 (en millions de $) de 1997 à 2011. Celles-ci sont issues des tableaux 379-0027 et 380-0027 de CANSIM sur E-STAT. Par exemple, pour l'année 1997, le PIB canadien (pibens) était de 10657900 millions de dollars tandis que la valeur des exportations totales pour l'ensemble des industries (expenst) était de 370923 millions de dollars.



	year	pibens	pibfab	expenst	expmar	expser
1	1997	1.1e+07	1.8e+06	370923	321484	49308
2	1998	1.1e+07	1.9e+06	404763	348873	55773
3	1999	1.2e+07	2.1e+06	448001	389614	58257
4	2000	1.1e+07	2.3e+06	487932	425306	62479
5	2001	1.1e+07	2.2e+06	473474	410758	62718
6	2002	1.2e+07	2.2e+06	479185	414034	65151
7	2003	1.2e+07	2.2e+06	468359	405483	62839
8	2004	1.2e+07	2.2e+06	491675	426554	65093
9	2005	1.3e+07	2.3e+06	500988	435277	65693
10	2006	1.3e+07	2.2e+06	503934	438581	65340
11	2007	1.3e+07	2.2e+06	510105	444483	65617
12	2008	1.4e+07	2.1e+06	486145	421122	65004
13	2009	1.3e+07	1.8e+06	419126	358057	61068
14	2010	1.4e+07	1.9e+06	445967	382858	63309
15	2011	1.4e+07	1.9e+06	466413	401833	64890

2.2.3 Les données mixtes

Les données mixtes combinent à la fois les caractéristiques des données en coupes et temporelles. En effet, elles regroupent des observations recensées dans un même instant de temps à plusieurs fréquences dans le temps, les unités observables étant différentes d'une période de temps à l'autre. Celles-ci visent spécialement à analyser l'effet d'une politique quelconque à travers le temps, par exemple l'effet d'une augmentation du salaire minimum sur la consommation hebdomadaire d'alcool entre deux périodes.

2.2.4 Les données longitudinales

Les données longitudinales sont des données dont les mêmes unités sont observées sur une période de temps, par exemple une étude faisant cinq suivis médicaux échelonnés sur cinq ans sur 100 patients atteints du cancer du colon. Ce type de données devient de plus en plus primé par les économistes étant donné qu'il est possible de contrôler davantage certaines caractéristiques non observables de l'échantillon utilisé. La façon commune de construire un panel dans la littérature scientifique est l'insertion verticale de chaque unité observable selon l'année. En se basant sur les données LOWBRTH.dta de Wooldridge (2009), nous remarquons que les données sont classées respectivement en paires de deux périodes, soient 1987 et 1990. Chaque paire d'entrées représente la même unité observable, c'est-à-dire un État. Par exemple, les deux premières lignes du classeur correspondent à des données de l'Alabama pour les années 1987 et 1990.

Data Editor (Browse) - [lowbirth]

File Edit View Data Tools

lpcinc[16] 9.938324

	year	lowbrth	infmort	afdcprt	popul	pcinc	physic	afdcprc
1	1987	8	12.2	132	4084	12039	151	3.232125
2	1990	8.4	10.8	132	4041	14899	158	3.266518
3	1987	4.8	10.4	19	524	18461	138	3.625954
4	1990	4.8	10.5	24	550	20867	146	4.363636
5	1987	6.4	9.5	91	3400	14322	191	2.676471
6	1990	6.4	8.8	144	3665	16169	197	3.929059
7	1987	7.7	10.3	67	2388	11421	144	2.805695
8	1990	8.2	9.2	73	2351	14037	150	3.105062
9	1987	6	9	1708	27653	17770	242	6.176545
10	1990	5.8	7.9	2023	29760	20547	244	6.797715
11	1987	7.9	9.8	94	3293	15680	202	2.85454
12	1990	8	8.8	109	3294	18795	211	3.309047
13	1987	6.7	8.8	107	3212	21258	293	3.331258
14	1990	6.6	7.9	135	3287	25528	305	4.107089
15	1987	6.7	11.7	20	648	16305	191	3.08642
16	1990	7.6	10.1	22	666	20709	199	3.303303
17	1987	7.7	10.6	307	12022	15594	203	2.553652
18	1990	7.4	9.6	420	12938	18606	208	3.246251

2.3 Le modèle de régression linéaire simple

Dans le chapitre introductif de ce guide, nous avons vu qu'il est possible de quantifier le degré de corrélation entre deux variables ou plusieurs variables à l'aide de STATA. Cependant, ce niveau de dépendance peut prendre la forme d'une fonction telle que y=F(X), où y, la variable dépendante, est fonction du vecteur X, les variables indépendantes. Ce faisant, les modèles de régression visent à prédire les valeurs de y conditionnelles à d'autres variables, X. Le modèle de régression linéaire simple propose un modèle prédictif où une variable dépendante, y, dépend linéairement d'une seule variable explicative, x.

2.3.1 Rappel théorique

Le modèle de régression linéaire simple vise à expliquer y en terme de x, par exemple le résultat à un examen selon le nombre d'heures d'étude effectuée une semaine avant la date de l'examen en question. L'équation structurelle de cette relation prend la forme suivante :

$$y = \beta_0 + \beta_1 x + u$$

Où : y est la variable dépendante

x est la variable indépendante
u est le terme d'erreur, c.-à-d. les autres facteurs qui affectent y sauf x, mais non inclus dans le modèle
β_0 (l'intercepte) et β_1 (la pente) sont les paramètres du modèle à estimer

Ici, la linéarité du modèle se fait par rapport aux paramètres et non par rapport à la variable explicative. Ce faisant, un modèle du genre $y = \beta_0 + \beta_1 \log(x) + u$ est linéaire, mais $y = \beta_0 + {\beta_1}^2 x + u$ ne l'est pas. L'objectif de l'économétrie est donc d'inférer sur l'effet *ceteris paribus* de x sur y en gardant tous les autres facteurs constants ($\Delta u = 0$), soit :

$$\Delta y = \beta_1 \Delta x$$

$\frac{\Delta y}{\Delta x} = \beta_1$, l'effet marginal de x sur y

β_1 serait donc l'effet de x sur y en moyenne.

La méthode d'approximation communément utilisée pour inférer sur ce genre de relation est la méthode des moindres carrés ordinaires (MCO). Celle-ci repose sur 5 hypothèses importantes qui font des estimateurs MCO les meilleurs estimateurs sans biais (BLUE) :

1. Le modèle est linéaire par rapport aux paramètres.
2. L'échantillon est obtenu par un tirage aléatoire dans la population → y et x sont indépendamment et identiquement distribués (i.i.d.).
3. Il existe une variabilité dans les valeurs de x où $x_i \neq C$, une constante, pour certains i.
4. La valeur espérée de u conditionnelle à x est égale à zéro, $E(u|x) = 0$.

Sous les hypothèses 1 à 4, les estimateurs MCO sont non biaisés ($E(\hat{\beta}_j) = \beta_j$).

5. Homoscédasticité (constance de la variance) de u, $var\ (u|x) = \sigma^2,\ \sigma$ étant une constante.

Ces hypothèses vérifiées, nous pouvons dès lors dériver les estimateurs MCO et obtenir l'équation de la droite de régression suivante :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \text{ le modèle prédit}$$

$$\hat{\beta}_1 = \frac{\text{cov (x,y)}}{\text{var (x)}} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2} \text{ et } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

2.3.2 Le pouvoir explicatif de la régression linéaire

Il existe plusieurs indicateurs qui permettent d'expliquer le pouvoir explicatif d'une régression linéaire, c'est-à-dire l'efficacité de la prédiction de x sur y. Les méthodes de base couramment utilisées en économétrie portent principalement sur l'analyse de la variance de la régression ou la variabilité totale (SCT) de la régression autour de la moyenne des observations composant la variable dépendante. Celle-ci peut d'ailleurs être décomposée en deux parties, soient la variance expliquée (SCE) et la variance non expliquée (SCR), c'est-à-dire celle issue des résidus, e, de la régression linéaire) :

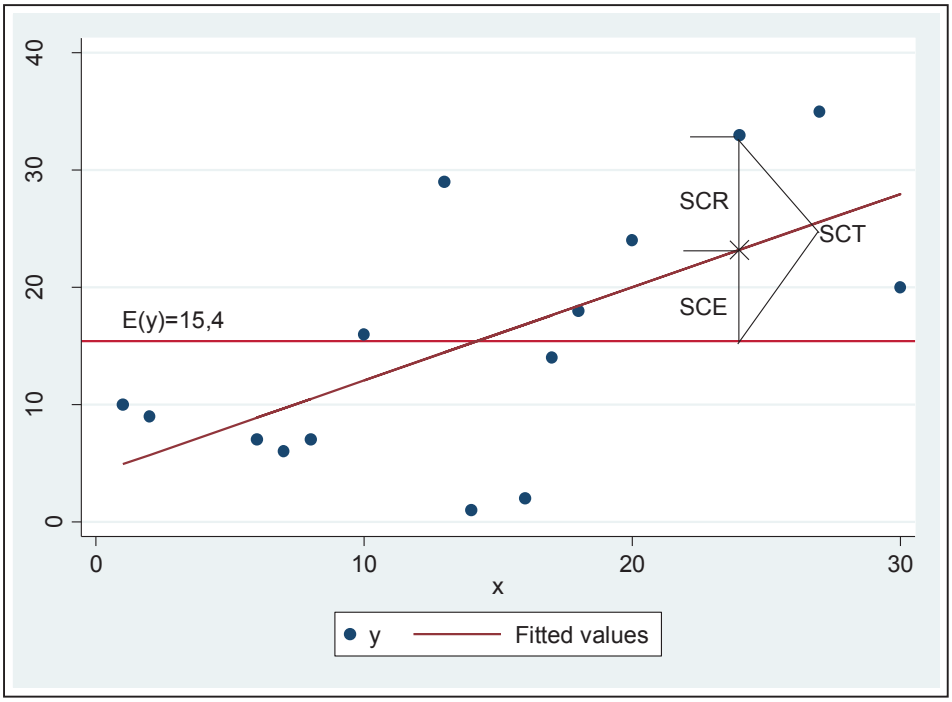
$$SCT = SCR + SCE$$

$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2$, la somme des carrés totaux – mesure la variation totale de y autour de la moyenne de l'échantillon.

$SCR = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$, la somme des carrés résiduels - mesure la partie de la variation totale de y autour de sa moyenne qui est non expliquée par la régression.

$SCE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \beta_1^2 \sum_{i=1}^{n}(x_i - \bar{x})$, la somme des carrés expliqués – mesure la partie de la variation totale de y autour de sa moyenne expliquée par la régression.

Un exemple graphique devrait permettre une meilleure compréhension de la variance totale de la régression.



Il est également possible de vérifier la qualité de l'ajustement du modèle à l'aide du coefficient de détermination (R^2) qui définit le pourcentage de la variation de y expliquée par la variation de x. Nous pouvons décomposer ce coefficient comme suit :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} = r^2, \text{ r étant le coefficient de corrélation entre x et y, } R^2 \in [0,1]$$

L'ajout de variables pertinentes au modèle (effet important) augmente la valeur du R^2 et ne la décroît pas. En effet, en ajoutant une variable pertinente, la variance résiduelle (SCR) du modèle baisse ce qui tend le R^2 vers 1. À noter qu'il est possible de comparer la qualité de deux modèles à l'aide de cette mesure si et seulement si la même variable dépendante est utilisée. Il faut toutefois rester prudent quant à l'utilisation de cette mesure. L'inclusion de variables provoquant un effet important, mais non significatif, sur la variable dépendante augmente quand même le coefficient de détermination. Dans ce cas, il est nécessaire de prendre en compte le coefficient de détermination ajusté ($\bar{R}^2$) qui pénalise pour l'ajout de variables explicatives

dont la statistique de student (t) est inférieure à 1,00 (la valeur du $\bar{R}^2$ diminuera). Il peut être également intéressant de comparer le $\bar{R}^2$ de deux modèles lorsque ceux-ci possèdent à la fois la même variable dépendante, mais également des variables indépendantes identiques, mais de formes fonctionnelles différentes. Voici la formule du $\bar{R}^2$.

$\bar{R}^2 = 1 - \frac{SCR}{SCT}\left(\frac{n-k-1}{n-1}\right)$, k + 1 étant le nombre de paramètres dans la régression et n le nombre d'observations,

En somme, la qualité des estimateurs ne repose pas uniquement sur le R^2, bien qu'il s'avère d'intérêt pour une analyse complète.

2.4 Tests d'hypothèse et intervalles de confiance

2.4.1 Introduction

La distribution de probabilité des estimateurs dépend de celle des erreurs, les estimateurs MCO étant des variables aléatoires. Ce faisant, afin d'inférer sur les paramètres d'une population à partir de notre échantillon, il faut connaître la distribution de probabilité du terme d'erreur, u. Sur la base du théorème central limite[7], on suppose généralement que u suit une distribution normale dans la population, d'où la 6e hypothèse du modèle des MCO.

6. Le terme d'erreur u est indépendant de la variable explicative, dans notre cas x, et suit une distribution normale avec moyenne zéro (E(u|x)=0) et variance σ^2 (var(u|x)= σ^2).

Sous les 6 hypothèses relatées précédemment (*the classical linear model assumptions*, CLM), il peut être démontré que les estimateurs MCO sont les meilleurs estimateurs sans biais à variance minimum parmi les estimateurs sans biais. Nous aurons donc :

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2)$$

$$\frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim N(0,1)$$

[7] Théorème central limite : la moyenne d'un grand nombre de variables aléatoires i.i.d suit approximativement une distribution normale.

Comme $\mathrm{sd}(\hat{\beta}_j)$ est inconnu, nous devons utiliser l'inférence statistique pour tester l'hypothèse sur β_j en choisissant $se(\hat{\beta}_j)$, l'erreur type, comme estimateur de $sd(\hat{\beta}_j)$. Sous les 6 hypothèses du *classical linear model*, nous pouvons démontrer que la variable aléatoire standardisée, $\frac{\hat{\beta}_j - \beta_j}{\mathrm{se}(\hat{\beta}_j)}$, suit une loi de Student (distribution t) avec $n-(k+1)$ degrés de liberté :

$\frac{\hat{\beta}_j - \beta_j}{\mathrm{se}(\hat{\beta}_j)} \sim \mathrm{t}_{n-k-1}$, k + 1 étant le nombre de paramètres du modèle structurel et n-k-1 le degré de liberté

Où $\mathrm{se}(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$ et $\hat{\sigma} = \frac{\mathrm{SCR}}{n-2}$, un estimateur consistent de σ.

2.4.2 Tests d'hypothèse

Dans la plupart des applications économétriques (STATA par exemple), nous testons à savoir si le paramètre de la population, β_j, est différent de zéro selon un seuil de confiance arbitraire (α), c'est-à-dire la probabilité de rejeter l'hypothèse nulle (H_0) sachant qu'elle est vraie. Ce faisant, pour toute hypothèse alternative (H_1), nous posons donc comme hypothèse nulle :

$$H_0: \beta_j = 0 \text{ avec}$$
$$H_1: \beta_j \neq 0 \text{ ou}$$
$$H_1: \beta_j > 0 \text{ ou}$$
$$H_1: \beta_j < 0$$

Ce faisant, nous pouvons rejeter l'hypothèse au seuil de significativité donné si et seulement si :

$$\left|\frac{\hat{\beta}_j - \beta_j}{\mathrm{se}(\hat{\beta}_j)}\right| > \mathrm{t}_{n-k-1,\ \alpha}, \text{ selon l'hypothèse alternative choisie}$$

En observant la valeur de la statistique t, il est possible de quantifier le niveau de significativité le plus petit possible permettant de rejeter l'hypothèse à l'aide de la p-valeur. En effet, la p-valeur est la probabilité d'observer une statistique t aussi petite que possible sachant que l'hypothèse nulle est vraie :

$p - valeur = P(|T| > |t|)$ dans le cas d'une hypothèse alternative unilatérale

$p - valeur = 2P(T > t)$ dans le cas d'une hypothèse alternative bilatérale

Si nous utilisons la p-valeur comme règle de décision, nous rejetons l'hypothèse nulle si :

$$p - valeur < \alpha$$

2.4.3 Intervalles de confiance

Sous les hypothèses du *classical linear model*, il est également possible de construire des intervalles de confiance pour le paramètre de la population à l'étude. En effet, si u suit une loi normale et si toutes les hypothèses du CLM tiennent, pour un intervalle de confiance bilatéral, β_j, le paramètre de la population peut être décrit par :

$$\hat{\beta}_j - t_{n-k-1,\ \alpha/2} se(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + t_{n-k-1,\ \alpha/2} se(\hat{\beta}_j)$$

L'interprétation ici va comme suit : pour un niveau de confiance $(1-\alpha)$, si un échantillon sélectionné aléatoirement était tiré à maintes reprises, la valeur du paramètre β_j serait située, en moyenne, autour de l'intervalle $\left[\hat{\beta}_j - t_{n-k-1,\frac{\alpha}{2}} se(\hat{\beta}_j),\ \hat{\beta}_j + t_{n-k-1,\ \alpha/2} se(\hat{\beta}_j)\right]$ dans 1-α des échantillons obtenus. Élément intéressant, lorsque le degré de liberté, n-k-1 est supérieur à 120, la loi de Student utilisée tend vers une loi normale. Il est donc possible de construire un intervalle de confiance pour β_j à l'aide d'une distribution normale standardisée (statistique z).

2.5 Comment rapporter les résultats d'une régression

Il existe une façon commune de présenter les résultats d'une régression. Voici la syntaxe générale à respecter dans le cas d'une régression linéaire simple.

$$\hat{y} = \underset{(se(\hat{\beta}_0))}{\hat{\beta}_0} + \underset{(se(\hat{\beta}_1))}{\hat{\beta}_1} x$$

$$n = n_0 \quad R^2 = p$$

où $n_0 \in \mathbb{Z}^+$ et $p \in [0,1]$

Les erreurs standards devraient toujours être placées entre parenthèses en dessous de leur coefficient respectif. De plus, il est important de présenter le nombre d'observations et la valeur du coefficient de détermination afin d'avoir une vue globale de la qualité de l'ajustement du modèle. Voici un exemple où nous cherchons à quantifier l'effet de l'éducation sur le revenu annuel en $ d'un individu. Nous effectuons la régression de annearn (le revenu annuel en $ d'un

individu) sur le niveau d'éducation d'un individu (en années). La régression utilise les données FRINGE.dta de Wooldridge (2009)

$$\widehat{annearn} = \underset{(1667,52)}{4313,08} + \underset{(130,18)}{685,17}\ educ$$

$$n = 616 \quad R^2 = 0{,}0432$$

Sur la base de ces résultats, nous voyons qu'une année d'éducation supplémentaire provoque en moyenne une augmentation de 685,17 $ du revenu annuel d'un individu. En poussant l'analyse plus loin, vérifions si les estimateurs MCO sont significatifs.

Testons β_0 au seuil de confiance de 0,05.

$$H_0: \beta_0 = 0$$

$$H_1: \beta_0 \neq 0$$

$$\left|\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)}\right| > t_{n-k-1,\ \alpha}$$

$$\left|\frac{4313,08}{1667,52}\right| > t_{614,\ 0,05}$$

$$2,59 > t_{614,\ 0,05}$$

Au seuil de 0,05, nous avons suffisamment d'évidence statistique pour rejeter l'hypothèse nulle.

Testons β_1 au seuil de confiance de 0,05.

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

$$\left|\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)}\right| > t_{n-k-1,\ \alpha}$$

$$\left|\frac{685,17}{130,18}\right| > t_{614,\ 0,05}$$

$$5,26 > t_{614,\ 0,05}$$

Au seuil de 0,05, nous avons suffisamment d'évidence statistique pour rejeter l'hypothèse nulle. Nous pouvons donc conclure qu'individuellement, nos estimateurs sont significatifs. Néanmoins, le coefficient de détermination reste relativement très bas. Nous pouvons donc conclure qu'il existe forcément d'autres facteurs, autres que l'éducation, qui peuvent influencer le revenu annuel en $ d'un individu.

2.6 Régression linéaire simple - Application sur STATA

2.6.1 Générer la régression linéaire simple

Pour effectuer une régression sur STATA avec la méthode des MCO, il faut utiliser la commande `regress` (ou `reg`).

Syntaxe :

```
regress depvar [indepvars] [if] [in] [weight] [, options]
```

où `depvar` est la variable dépendante
`[indepvars]` est la variable indépendante
`[if]` et `[in]` sont des contraintes dans l'échantillon
`[weight]` assigne des poids aux observations
`[, options]` présente des options supplémentaires dans la régression

Menu:

Statistics > Linear models and related > Linear régression

Options:

`Noconstant` : supprime l'intercepte dans le modèle prédit

`robust` : utilisation des estimateurs robustes à l'hétéroscédasticité[8]

`level(#)` : spécifie un intervalle de confiance aux tests d'hypothèse

`beta` : rapporte des coefficients beta standardisés

`noheader` : supprime l'entête de l'output STATA de la régression

[8] Voir section 6.4.

`notable` : supprime le tableau contenant la valeur des estimateurs

La section 3.1.2 donne un exemple introductif de la commande `reg`.

2.6.2 Interprétation de l'output STATA de la régression linéaire simple

Cette sous-section traite de l'interprétation de l'output STATA généré suite à l'opération d'une régression. Nous utiliserons les données FRINGE.dta de Wooldridge (2009) pour clarifier le tout à l'aide d'un exemple. Le modèle structurel (de la population) est le suivant :

$$annearn = \beta_0 + \beta_1 educ + u$$

Où annearn est le revenu annuel en $ et educ est le nombre d'années de scolarité

Nous cherchons à quantifier l'effet de l'éducation sur le revenu annuel en $ d'un individu. Vérifions cette relation avec STATA.

```
use http://fmwww.bc.edu/ec-p/data/wooldridge/FRINGE

reg annearn educ
```

1

Source	SS	df	MS
Model	2.1566e+09	1	2.1566e+09
Residual	4.7802e+10	614	77854167.8
Total	4.9959e+10	615	81234289.6

2

Number of obs =	616
F(1, 614) =	27.70
Prob > F =	0.0000
R-squared =	0.0432
Adj R-squared =	0.0416
Root MSE =	8823.5

3

annearn	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	685.1702	130.1822	5.26	0.000	429.5138	940.8266
_cons	4313.082	1667.517	2.59	0.010	1038.353	7587.81

En suivant la forme standard de présentation pour un modèle prédit (voir section 2.4), nous obtenons les résultats suivants :

$$\widehat{annearn} = \underset{(1667,52)}{4313,08} + \underset{(130,18)}{685,17}\ educ$$

$$n = 616 \quad R^2 = 0,0432$$

Voici un tableau descriptif détaillé des différentes statistiques composant l'output STATA de la régression de annearn sur educ.

Premier tiers – La variance de la régression

Composante	Valeur	Signification	Formule
Model SS - Somme des carrés expliqués	2.1566e+09	Mesure la partie de la variation totale de y autour de sa moyenne expliquée par la régression.	$\mathrm{SCE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
Residual SS – Somme des carrés résiduels	4.7802e+10	Mesure la partie de la variation de y autour de sa moyenne non expliquée par la régression.	$\mathrm{SCR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
Total SS – Sommes des carrés totaux	4.9959e+10	Mesure la variation totale de y autour de la moyenne de l'échantillon.	$\mathrm{SCT} = \sum_{i=1}^{n} (y_i - \bar{y})^2$
Model Df – Degré de liberté SCE	1		Nombre de variables indépendantes
Residual Df – Degré de liberté SCR	614		n-2
Total Df – Degré de liberté SCT	615		n-1
Model MS – Variance de la somme des carrés expliqués	2.1566e+09	Variance de la variation totale de y autour de sa moyenne expliquée par la régression.	$MSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
Residual MS – Variance de la somme des carrés résiduels	77854167.8	Variance de la variation totale de y autour de sa moyenne non expliquée dans le modèle	$\mathrm{MSE} = \hat{\sigma}^2 = \frac{\sum_i^n \hat{e}_i^2}{n-2}$ [9]
Total MS – Variance de la somme des carrés	81234289.6	Variance de la variable dépendante dans l'échantillon	$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$

[9] La variance du terme d'erreur est $var(e_i) = E[e_i - E(e_i]^2 = E(e_i^2)$. Cependant, le terme d'erreur est non-observé, sa vraie valeur reste donc inconnue. Ce paramètre est donc estimé par $\hat{\sigma}^2 = \frac{\sum_i^n \hat{e}_i^2}{n-2}$, un estimateur non biasé.

totaux

Deuxième tiers – Qualité de l'ajustement du modèle

Composante	Valeur	Signification	Formule
Number of observations – Nombre d'observations	616	Taille de l'échantillon	n
F(1, 614) = Test de Fisher[10]	27,70	F(q, n-k-1), q étant le nombre de paramètres contraints et n-k-1 le degré de liberté	$F = \left(\frac{SCR_C - SCR_{nc}}{SCR_{nc}}\right)\left(\frac{n-k-1}{q}\right)$
Prob > F	0,00	La p-valeur du test de Fisher	
R-squared – Coefficient de détermination	0,0432	Le pourcentage de la variation de y expliquée par la variation de x	$R^2 = \frac{SCE}{SCT}$
Adj R-squared – Coefficient de détermination ajusté	0,0416	définis le pourcentage de la variation de y expliquée par la variation de x corrigé selon le niveau de significativité de x	$\bar{R}^2 = 1 - \frac{SCR}{SCT}\left(\frac{n-k-1}{n-1}\right)$
Root MSE = $\hat{\sigma}$	8823,5	Erreur type de la somme des carrés résiduels	$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$

Troisième tiers – Coefficients et niveau significativité

Composante	Valeur	Signification	Formule
Coef. educ – Coefficients de educ	685,1702	Pente de la droite de régression	$\hat{\beta}_1 = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2}$
Coef. _cons – Coefficient de la constante ou l'intercepte	4313,082	Intercepte de la droite de régression	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

[10] Voir section 5.2 pour de plus amples détails

Std. Err. educ – Erreur type du coefficient de educ	130,1822	Écart type de l'estimateur $\hat{\beta}_1$	$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$
Std. Err._cons– Erreur type du coefficient de la constante	1667,517	Écart type de l'estimateur $\hat{\beta}_0$	$se(\hat{\beta}_0) = \frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})}}$
t educ – Statistique t de educ	5,26	Statistique t de la variable aléatoire standardisée $\hat{\beta}_1$	$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)}$
t _cons - Statistique t de la constante	2,59	Statistique t de la variable aléatoire standardisée $\hat{\beta}_0$	$t = \frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)}$
P>|t| educ – p-valeur de educ	0,000	La probabilité d'observer une statistique t aussi petite que possible sachant que l'hypothèse nulle, $\beta_1 = 0$, est vraie	
P>|t| _cons – p-valeur de la constante	0,010	La probabilité d'observer une statistique t aussi petite que possible sachant que l'hypothèse nulle, $\beta_0 = 0$, est vraie	
[95% Conf. Interval] educ – Intervalle de confiance de educ	[429,5138; 940,8266]	Intervalle de confiance au niveau de confiance 95 %	$\hat{\beta}_1 \pm t_{n-k-1,\ \alpha/2} se(\hat{\beta}_1)$
[95% Conf. Interval] _cons – Intervalle de confiance de la constante	[1038,353; 7587,81]	Intervalle de confiance au niveau de confiance 95 %	$\hat{\beta}_0 \pm t_{n-k-1,\ \alpha/2} se(\hat{\beta}_0)$

2.6.3 Générer les valeurs prédites du modèle

Il est possible de dériver les valeurs prédites et les résidus de la régression linéaire à l'aide de la commande `predict`.

Syntaxe :

`predict [type] newvar [if] [in] [, single_options]`

où `[type]` est le type de variable utilisée
`newvar` est le nom de la nouvelle variable générée
`[if]` et `[in]` sont des contraintes dans l'échantillon
`[, single_ options]` présente des options supplémentaires dans la commande postestimation `predict`

Menu:

Statistics > Postestimation > Prédictions, residuals, etc.

`Single_`options:

`xb` : calcule les valeurs prédites de la régression linéaire

`stdp` : calcule les erreurs types des valeurs prédites de la régression linéaire

`reside` : calcule les résidus de la régression linéaire

Le lecteur est invité à lire la section 6 sur l'hétéroscédasticité pour voir les applications de la commande `predict`.

2.6.4 Calculer l'élasticité des coefficients

Il est parfois intéressant de vérifier l'ampleur de l'effet de x sur y en calculant l'élasticité évaluée à leur moyenne. Tout d'abord, revoyons mathématiquement les prémisses de base du concept d'élasticité.

Dans le cas d'une régression linéaire, l'élasticité évaluée à la moyenne se définit comme suit :

$$\varepsilon = \frac{\Delta\ en\ \%\ de\ y}{\Delta\ en\ \%\ de\ x} = \frac{\Delta y}{\Delta x}\frac{\bar{x}}{\bar{y}}, \text{ ou}$$

$$\varepsilon = \frac{dy/\bar{y}}{dx/\bar{x}} = \frac{dy}{dx}\frac{\bar{x}}{\bar{y}} = \beta_j\frac{\bar{x}}{\bar{y}}$$

ε se définit ainsi comme étant l'effet en pourcentage d'une variation de 1,00 % de x sur y, évalué à leur moyenne. Nous pouvons calculer l'élasticité sur STATA à l'aide de la commande **margins**. La commande margins permet de calculer des statistiques à partir des prédictions d'un modèle préalablement construit sur STATA.

Syntaxe :

```
margins [marginlist] [if] [in] [weight] [, response_options options]
```

où `[marginlist]` est la liste de variable pour lesquelles nous discriminons
`[if] [in]` sont des contraintes dans l'échantillon
`[weight]` assigne des poids aux observations
`response_options` est l'option affiliée directement à la commande `margins`
`option` est l'option supplémentaire spécifiée sur `response_options`

Menu :

Statistics > Postestimation > Marginal means and predictive margins

Statistics > Postestimation > Marginal effects

Response_options:

`dydx(varlist)`: estime l'effet marginal (dans notre cas les coefficients, car l'effet marginal représente la pente de la droite de régression).

`eyex(varlist)` : estime les élasticités des variables

`dvex(varlist)` : estime les semi-élasticités (d(y)/d(ln (x))

`eydx(varlist)` : estime les semi-élasticités (d(ln (y))/d(x))

Options :

`at(#)`: estime les marges à des valeurs spécifiques des valeurs explicatives

`atmeans` : estime les marges à la moyenne des valeurs explicatives

Par exemple, en reprenant les résultats de la régression de annearn sur educ avec les données FRINGE.dta de Wooldridge (2009), calculons l'élasticité évaluée à la moyenne de annearn par rapport à educ. Ne pas oublier qu'il faut préalablement effectuer la régression avant d'utiliser cette commande postestimation. La commande `quietly` permet ici de régresser sur STATA, mais en cachant l'output des résultats.

```
quietly reg annearn educ
margins, eyex (educ) atmeans

Conditional marginal effects                      Number of obs   =        616
Model VCE    : OLS
Expression   : Linear prediction, predict()
ey/ex w.r.t. : educ
at           : educ            =    12.51461 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      ey/ex   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .665334   .1277387     5.21   0.000     .4149707    .9156973
```

L'output STATA ci-haut s'interprète sensiblement de la même façon que celui de la commande reg. L'élasticité de annearn par rapport à educ est 0,665334 : lorsque educ varie de 1,00 %, annearn varie de 0,665334 %. En regardant la p-valeur (P>|z|), nous pouvons confirmer que le coefficient est statistiquement significatif. À noter que nous aurions pu utiliser l'option `at(#)` ou l'abréviation `atmeans` en insérant la moyenne de educ et obtenir un output STATA identique à celui généré précédemment. Voici cette commande.

```
margins, eyex (educ) atmeans
```

2.6.5 Obtenir la matrice variance-covariance

Une commande postestimation forte importante sur STATA permet de calculer la matrice de variance-covariance des estimateurs MCO : la commande `estat vce`. En effet, les résultats obtenus suite à une régression sur STATA ne permettent pas d'obtenir la variance des estimateurs, mais seulement leurs écarts types. Aux fins de présentation, nous clarifierons dans cette sous-section l'ensemble des options importantes pour la commande `estat vce` étant donné les variantes intéressantes postestimation qu'elle offre.

Syntaxe :

```
estat vce [, estat_vce_options]
```

où `[, estat_vce_options]` représente les options reliées à `estat vce`

Menu :

Statistics > Postestimation > Reports and statistics

Estat_vce_options :

`covariance` : affiche la matrice des covariances des estimateurs

`correlation` : affiche la matrice des corrélations des estimateurs

En reprenant les résultats de la régression de annearn sur educ, utilisons la commande `estat vce` pour obtenir la matrice de variance-covariance.

```
quietly reg annearn educ
estat vce, cov
```

Covariance matrix of coefficients of regress model

```
        e(V) |       educ       _cons
-------------+------------------------
        educ |  16947.409
       _cons | -212090.22   2780613.1
```

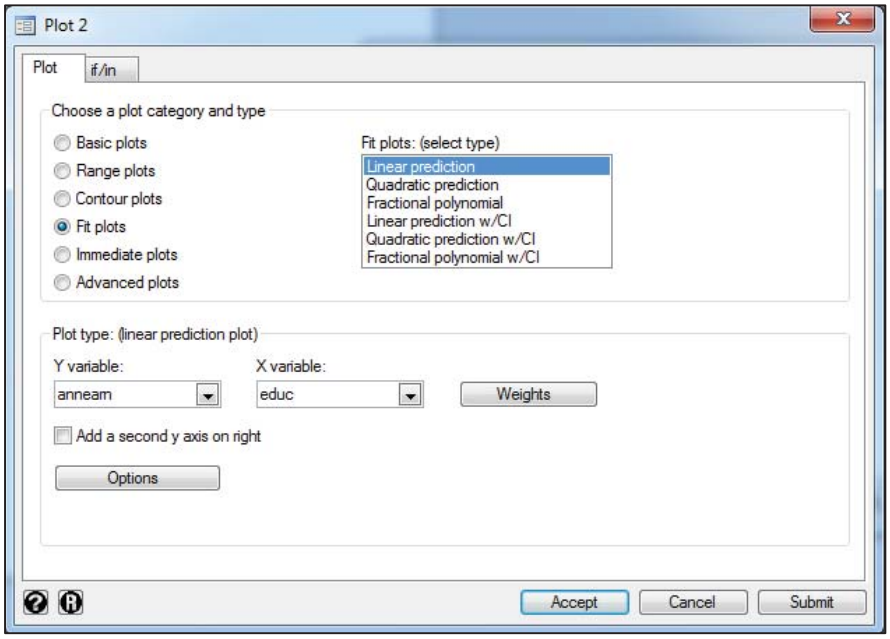
Le tableau ci-dessus peut également être retranscrit sous la forme générale suivante :

	$\hat{\beta}_1$	$\hat{\beta}_0$
$\hat{\beta}_1$	$\mathrm{cov}(\hat{\beta}_1, \hat{\beta}_1) = \mathrm{var}(\hat{\beta}_1)$	$\mathrm{cov}(\hat{\beta}_1, \hat{\beta}_0)$
$\hat{\beta}_0$	$\mathrm{cov}(\hat{\beta}_1, \hat{\beta}_0)$	$\mathrm{cov}(\hat{\beta}_0, \hat{\beta}_0) = \mathrm{var}(\hat{\beta}_0)$

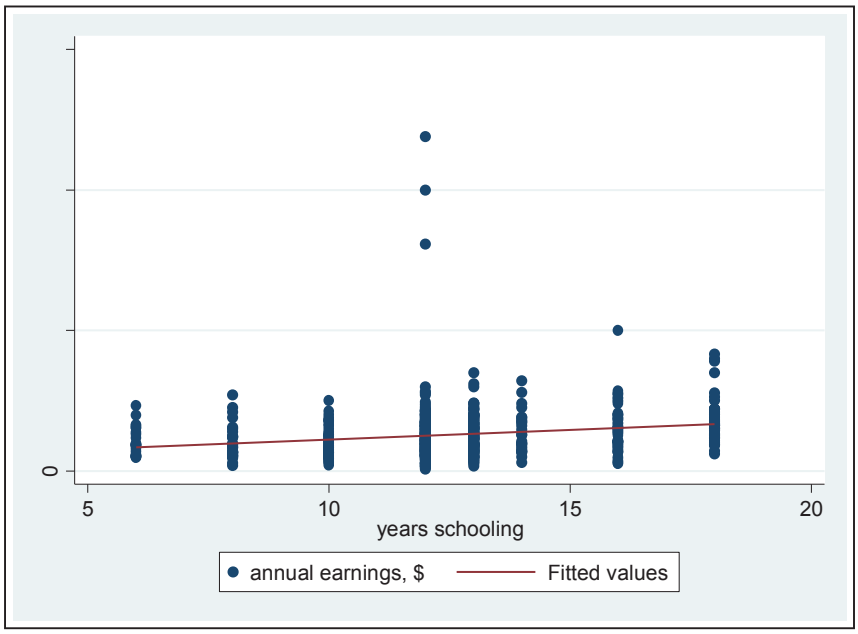
2.6.6 Générer le graphique de la droite de régression

Évidemment, la présentation des résultats d'une régression simple n'est complète que si nous y ajoutons le graphique de la droite de régression. Cette sous-section présente la méthode qui utilise la barre d'outils afin de faciliter la compréhension du lecteur, les options rattachées à la construction de graphiques étant relativement très nombreuses et offrant une infinité de possibilités. L'exemple ici consiste à utiliser les valeurs prédites du modèle linéaire annearn sur educ dans un graphique en ligne tout en utilisant les valeurs réelles de ces deux variables dans un nuage de points.

1. Dans la barre d'outils, sélectionner Graphics > Twoway graph (scatter, line, etc.). À partir de la fenêtre « twoway – Twoway graphs », créons un tracé en cliquant sur « Create ».
2. L'utilisateur STATA sera amené à construire son premier tracé (plot 1). Ici, nous allons créer un nuage de points avec comme variable dépendante annearn et variable indépendante educ. Pour y arriver, cliquons sur « Scatter » dans la fenêtre « Basic plots : (select type) » et sélectionnons la variable annearn dans la fenêtre « Y variable : » et la variable educ dans la fenêtre « X variable : ». Appuyons sur « Accept ».
3. Nous serons redirigés vers la fenêtre « twoway – Twoway graphs ». À partir d'ici, nous désirons tracer la droite de régression de annearn sur educ. Cliquons sur « Create ».
4. Dans la fenêtre « Plot 2 », cliquons sur « Fit plots » pour « tracé prédit » dans la sélection de cases « Choose a plot category and type » et par la suite sélectionnons « Linear prediction » dans la fenêtre « Basic plots : (select type). Choisissons finalement les mêmes variables d'intérêt qu'à l'étape 3 dans les mêmes fenêtres. Appuyons sur « Accept ». Nous serons encore redirigés vers la fenêtre « twoway – Twoway graphs ». Il est dès lors possible de générer le graphique désiré grâce à nos deux tracés en cliquant sur « OK ». L'image ci-dessous est une représentation des différentes étapes à faire sur la fenêtre « Plot 2 ».



Voici le graphique de la droite de régression de annearn sur educ.



La commande sommaire utilisée est la suivante :

```
twoway (scatter annearn educ) (lfit annearn educ)
```

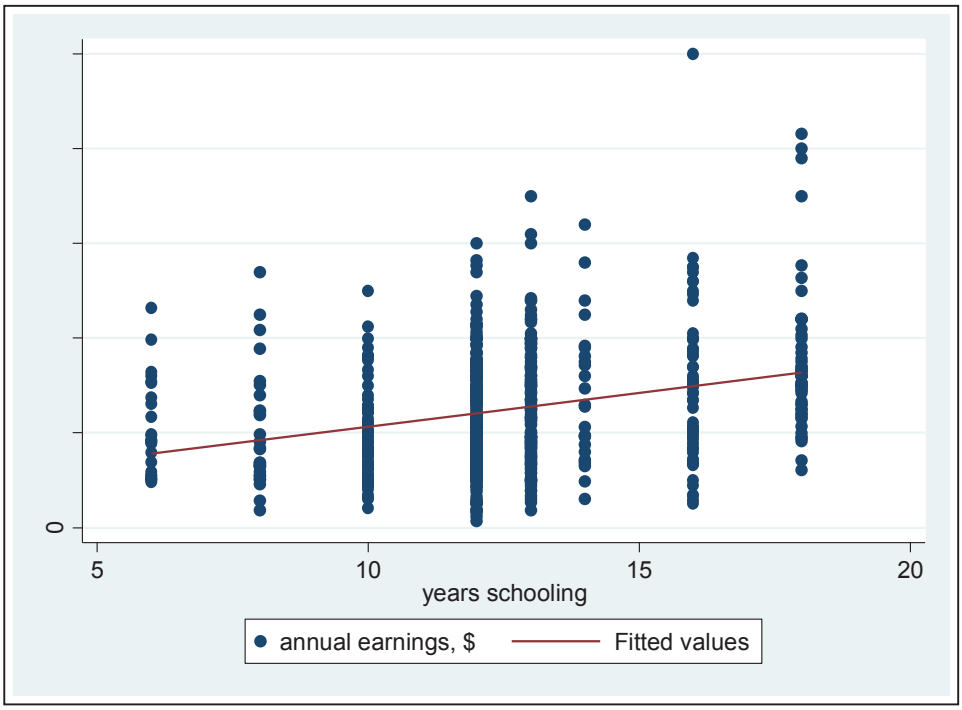
Nous pouvons voir ici que la qualité du graphique est passablement mauvaise étant donné que pratiquement toutes les valeurs que prennent les observations de annearn sont recensées à l'intérieur de l'intervalle [0, 50000]. L'intuition ici ne repose pas dans la qualité du graphique, mais dans la qualité de l'estimation. En effet, comme nous pouvons le voir dans le graphique précédent, il existe 3 observations de annearn ayant des valeurs extrêmes dans notre échantillon. Ces trois observations peuvent en quelque sorte fausser l'interprétation de nos résultats, étant donné que nous désirons obtenir des estimations à la moyenne. Effectuons à nouveau la régression de annearn sur educ pour toutes valeurs de annearn inférieures ou égales à 50000.

```
reg annearn educ if annearn<=50000

      Source |       SS       df       MS              Number of obs =     613
-------------+------------------------------           F(  1,   611) =   59.61
       Model |  2.3460e+09     1  2.3460e+09           Prob > F      =  0.0000
    Residual |  2.4046e+10   611  39354456.3           R-squared     =  0.0889
-------------+------------------------------           Adj R-squared =  0.0874
       Total |  2.6392e+10   612  43123544.4           Root MSE      =  6273.3
```

```
------------------------------------------------------------------------------
     annearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   714.6869    92.5647     7.72   0.000     532.9033    896.4705
       _cons |   3515.943   1186.025     2.96   0.003     1186.762    5845.124
------------------------------------------------------------------------------
```

Nous remarquons ici que les coefficients de educ et de _cons sont relativement différents de ceux obtenus en incluant les trois observations extrêmes (voir section 3.1.2.). Comme espérée, la suppression de ces observations fait également diminuer la variance de notre modèle.

Il est possible de spécifier à STATA un intervalle spécifique à respecter grâce à l'option if pour créer notre nouveau graphique. Dans la fenêtre « twoway – Twoway graphs », cliquons sur l'onglet « if/in » et insérons la commande « annearn<=50000» dans l'espace « If : expression ». Une fois la spécification faite, appuyons sur « OK ». Voici le graphique en question.



Voici la commande résumant les étapes précédentes qui inclue la spécification de l'intervalle.

```
twoway (scatter annearn educ) (lfit annearn educ) if annearn<=50000
```

2.7 Le modèle de régression linéaire multiple

Contrairement au modèle de régression linéaire simple, le modèle de régression multiple offre la possibilité de contrôler pour plusieurs variables permettant ainsi d'améliorer la prédiction de la variable dépendante. En effet, dans le cas d'une régression linéaire simple, l'hypothèse d'exogénéité stricte (E(u|x)=0) est peu réaliste étant donné qu'il existe une multitude de facteurs extérieurs qui peuvent influencer la variable dépendante. Le modèle de régression linéaire multiple permet ainsi de prendre en compte une plus grande proportion de variables indépendantes permettant ainsi de réduire le biais de variables omises.

2.7.1 Rappel théorique

La forme générale du modèle de régression multiple est la suivante :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u, \text{ pour k+1 paramètres inconnus}$$

Où : y est la variable dépendante

x_j sont les variables indépendantes, $j \in [1, k]$ pour k variables exogènes
u est le terme d'erreur contenant toutes les variables corrélées avec y, mais non incluses dans le modèle
β_0 est l'intercepte (ou la constante) du modèle
β_1 est le paramètre associé à x_1, β_2 est le paramètre associé à x_2, ... et β_k est le paramètre associé à x_k

Le concept *ceteris paribus* est très important ici étant donné l'inclusion de plusieurs variables indépendantes dans le modèle. Nous parlerons donc de l'effet partiel de x_j sur y, toute chose étant égale par ailleurs. Cet effet partiel est dans ce sens inclus dans β_j, la pente du régresseur x_j. Par exemple, supposons que nous désirons obtenir l'effet de x_2 sur y, soit l'équivalent de la dérivée première de y par rapport à x_2 ou l'effet marginal de x_2 sur y.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

$$\Delta y = \beta_2 \Delta x_2, \text{ où } \Delta u = 0 \text{ par hypothèse (nous y reviendrons)}$$

$$\frac{\Delta y}{\Delta x_2} = \frac{dy}{dx_2} = \beta_2, \text{ l'effet partiel de } x_2 \text{ sur y}$$

Encore une fois, il est possible de dériver les estimateurs MCO des paramètres β_k en minimisant la somme des carrés résiduels :

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik})^2$$

Afin de dériver le modèle prédit à l'aide de la méthode des moindres carrés ordinaires, nous devons assumer les hypothèses de Gauss-Markov suivantes :

1. Le modèle est linéaire par rapport aux paramètres.
2. L'échantillon est obtenu par un tirage aléatoire dans la population → $y_i, x_{i1}, x_{i2}, \ldots et\ x_{ik}$ sont indépendamment et identiquement distribués (i.i.d.).
3. Il n'existe pas de multicolinéarité parfaite entre les variables indépendantes. Aucune des variables explicatives ne peut s'avérer être une combinaison linéaire des autres variables explicatives.
4. La valeur espérée de u conditionnelle aux x_j est égale à zéro, $E(u|x_1, x_2, \ldots, x_k)=0$.

Sous les hypothèses 1 à 4, les estimateurs MCO sont sans non biaisés ($E(\hat{\beta}_j) = \beta_j$).

5. Homoscédasticité (constance de la variance) de u, $var(u_i|x_1, x_2, \ldots, x_k)=\sigma^2$, σ étant une constante et $i = 1, 2, \ldots, n$.

Sous les hypothèses 1 à 5, la variance des estimateurs MCO est donnée par la formule suivante :

$$var(\hat{\beta}_j) = \frac{\sigma^2}{SCT_j(1-R_j^2)},\ R_j^2 \neq 1, \text{ pas de multicolinéarité parfaite}$$

Où $SCT_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$ est la variation totale de la variable indépendante x_j
R_j^2 est le coefficient de détermination obtenu de la régression de x_j sur toutes les autres variables explicatives du modèle
σ^2 est la variance de u conditionnelle aux x_j

Comme σ^2 est un paramètre inconnu, nous pouvons l'estimer avec cet estimateur sans biais :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k-1} = \frac{SCR}{n-k-1}, \text{ où } E(\hat{\sigma}^2) = \sigma^2$$

Sous les hypothèses 1 à 5, les estimateurs sont BLUE où :

$\hat{\beta}_1 = \frac{\sum_i^n \hat{r}_{i1} y_i}{\sum_i^n \hat{r}_{i1}{}^2}$, où $\hat{r}_1$est le résidu de la régression de x_1 sur les autres régresseurs

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 - \cdots - \hat{\beta}_k \bar{x}_k$$

6. Le terme d'erreur u est indépendant des variables explicatives et suit une distribution normale avec moyenne zéro ($E(u|x_1, x_2, \ldots, x_k)=0$) et variance σ^2 ($var(u|\ x1, x2, \ldots, xk)= \sigma^2$). Cette hypothèse nous permet essentiellement d'inférer sur les paramètres de la population à partir de l'échantillon collecté (tests hypothèse – idem que dans une régression linéaire simple).

Dans l'ensemble, le modèle de régression linéaire multiple repose similairement sur les mêmes hypothèses que le modèle de régression linéaire simple. De plus, la décomposition de la variabilité totale de la régression, le R^2, les tests d'hypothèses et les intervalles de confiance se construisent intuitivement de la même façon. La différence majeure entre ces deux modèles se situe au niveau du nombre de variables indépendantes qui composent chacun des modèles.

2.7.2 Test de Fisher (le test joint)

Parfois, il est utile de tester si un ensemble de variables explicatives sont conjointement significatives. En effet, dans un modèle, il peut arriver que chaque variable explicative prise individuellement soit non significative, mais devienne conjointement significative une fois testée en groupe. Supposons le modèle général prédit suivant :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

Nous désirons tester la signification globale du modèle (toutes les variables prises conjointement). Soit l'hypothèse nulle suivante :

$$H_0: \beta_0 = 0, \beta_1 = 0, \beta_2 = 0, \ldots, \beta_k = 0$$

$H_1: \beta_0 \neq 0, \beta_1 \neq 0, \beta_2 \neq 0, \ldots, \beta_k \neq 0$, au moins un des paramètres est différent de zéro

Pour y parvenir, nous devons distinguer un modèle non contraint où il n'existe aucune contrainte sur les paramètres et un modèle contraint où on impose une ou plusieurs contraintes sur les paramètres (selon le nombre de paramètres testés). Le test se réalise à l'aide de la statistique F (de Fisher) :

$$F \equiv \left(\frac{SCR_C - SCR_{nc}}{SCR_{nc}}\right)\left(\frac{n-k-1}{q}\right),\ F > 0\ car\ SCR_C > SCR_{nc}$$

Où SCR_C est la somme des carrés résiduels du modèle contraint
SCR_{nc} est la somme des carrés résiduels du modèle non contraint
$n-k-1$ est le degré de liberté où k est le nombre de coefficients relatifs aux régresseurs
q est le nombre de restrictions linéaires imposées au modèle contraint

Le test F s'avère important dans l'optique où nous désirons tester si un groupe de variables explicatives corrélées ou non entre elles sont conjointement significatives de sorte à identifier quelles sont les variables pertinentes dans le modèle. Nous pouvons d'autant plus retranscrire le test F sous la forme de coefficients de détermination :

$$F \equiv \left(\frac{R^2_{nc} - R^2_c}{1 - R^2_{nc}}\right)\left(\frac{n-k-1}{q}\right),\ \text{où } R^2_{nc} > R^2_c$$

Notons également qu'il est possible de tester une seule variable avec la statistique F. Dans ce cas, le test F sera équivalent à un test bilatéral utilisant la statistique t où $q = 1$.

2.8 Modèle de régression linéaire multiple – Application sur STATA

2.8.1 Générer la régression linéaire multiple

La commande utilisée pour effectuer une régression linéaire multiple sur STATA est la même que dans le cas d'une régression linéaire simple, soit la commande `regress`. Toutes les options s'y rattachant restent donc les mêmes que celles explicitées dans la section 3.1.

2.8.2 Interprétation de l'output STATA de la régression linéaire multiple

L'output STATA présentant les résultats d'une régression linéaire multiple est quasi identique à celui de la régression linéaire simple. Pour identifier les quelques différences, reprenons l'exemple de la section 3.2.

Au départ, nous avions le modèle suivant :

$$annearn = \beta_0 + \beta_1 educ + u$$

Il y a fort à parier qu'il existe d'autres facteurs pouvant influencer le revenu annuel d'un individu. Nous n'avons qu'à penser à l'expérience par exemple. De plus, l'expérience et le niveau d'éducation sont probablement corrélés l'une et l'autre faisant en sorte que l'hypothèse d'exogénéité stricte reste peu réaliste ($E(u|x) \neq 0$). L'estimateur $\hat{\beta}_1$ est donc forcément biaisé. Pour palier le problème, supposons que annearn est fonction de educ et exper (expérience d'un individu) et que toutes les hypothèses de Gauss-Markov sont respectées. Nous avons donc l'équation structurelle suivante :

$$annearn = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

Par la méthode des MCO, trouvons les estimateurs de β_0, β_1 et β_2 à l'aide de STATA.

```
. reg annearn educ exper

      Source |       SS       df       MS              Number of obs =     616
-------------+------------------------------           F(  2,   613) =   53.98
       Model |  7.4811e+09     2  3.7406e+09           Prob > F      =  0.0000
    Residual |  4.2478e+10   613    69295171           R-squared     =  0.1497
-------------+------------------------------           Adj R-squared =  0.1470
       Total |  4.9959e+10   615  81234289.6           Root MSE      =  8324.4

------------------------------------------------------------------------------
     annearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    894.514   125.1184     7.15   0.000     648.8012    1140.227
       exper |   243.1261   27.73594     8.77   0.000     188.6571    297.5951
       _cons |  -2833.026    1771.87    -1.60   0.110    -6312.699    646.6466
------------------------------------------------------------------------------
```

La différence ici repose sur le test de Fisher (encadré 1) et l'ajout de la variable exper dans l'encadré 2. Le test de Fisher mesure ici la significativitée globale du modèle, car nous testons

pour $q = k$. La variable exper est également significative au seuil de 1,00 %, sa p-valeur étant de 0,000.

2.8.3 Tester le niveau de significativité des variables

La commande `test` permet de tester des hypothèses de linéarité après l'estimation du modèle. Celle-ci s'avère forte utile lorsque nous désirons effectuer un test de Fisher pour un groupe de variables d'intérêt. La commande `test` possède quatre syntaxes différentes selon les objectifs du test. Nous utiliserons la première syntaxe qui consiste à un test paramétrique sur des hypothèses linéaires ou composites de variables faisant parties du modèle économétrique. Ce test repose sur l'hypothèse nulle où les coefficients sont égaux à zéro. La commande `test` se base sur un test de Wald (voir Green (2012)).

Syntaxe :

```
test coeflist
```

où `coeflist` est la liste des coefficient à tester (la ou les variables à tester)

Menu :

Statistics > Postestimation > Tests > Test linear hypotheses

Illustrons à l'aide d'un exemple. Reprenons le modèle suivant:

$$annearn = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

Dont le modèle estimé à l'aide des MCO est :

$$\widehat{annearn} = \underset{(1771{,}87)}{-2833{,}026} + \underset{(125{,}1184)}{894{,}514}\ educ + \underset{(27{,}73594)}{243{,}1261}\ \text{exper}$$

$$n = 616 \quad R^2 = 0{,}1497$$

À l'aide de STATA, nous désirons tester si educ et exper sont conjointement significatifs, c'est-à-dire si leur coefficient est simultanément différent de zéro.

```
. test educ exper

 ( 1)  educ = 0
 ( 2)  exper = 0

       F(  2,   613) =   53.98
```

```
Prob > F =    0.0000
```

Ici, les points 1 et 2 représentent l'hypothèse nulle ($H_0: \beta_1 = \beta_2 = 0$). F(2, 613) est la statistique de Fisher avec $q = 2$ et $n - k - 1 = 613$ degrés de liberté. Finalement, Prob > F est la p-valeur. Sur la base de ses résultats, nous avons suffisamment d'évidence statistique pour rejeter l'hypothèse nulle, et ce, au seuil de confiance de 1,00 %.

2.8.4 Générer un graphique avec les estimations du modèle de régression linéaire multiple

Il est possible de générer un graphique à partir des valeurs prédites du modèle de régression linéaire multiple. En effet, le graphique construit ne peut contenir que deux variables, soit la variable dépendante et une variable indépendante parmi l'ensemble des régresseurs. Nous devons donc garder fixes les autres variables. Une façon commune de procéder est de fixer les autres variables à leur valeur moyenne. Nous pouvons générer un graphique du genre notamment en utilisant la commande `twoway function` ou en utilisant la commande post-estimation `marginsplot`. La méthode utilisant twoway function est beaucoup plus longue, mais décrit relativement bien l'intuition derrière les étapes pour réaliser un graphique d'une régression multiple. Commençons par celle-ci.

Syntaxe :

```
twoway function [[y]=] f(x) [if] [in] [, options]
```

où `[[y]=] f(x)` est la fonction représentée par le graphique
`[if] [in]` sont des contraintes dans l'échantillon
`[, options]` présente des options supplémentaires dans la création du graphique

Menu :

Graphics > Twoway graph (scatter, line, etc.)

`[, options]` :

`range(# #)` : tracer dans l'intervalle de x [#,#]

`range(varname)` : tracer dans l'intervalle de x [min x, max x]

`n(#)`: spécifie le nombre de points où f(x) est évalué. La valeur par défaut est n(300)

`droplines(numlist)` :

`base(#)` : spécifie la base pour les « droplines »

Par exemple, en estimant le modèle $annearn = \beta_0 + \beta_1 educ + \beta_2 exper + u$, nous désirons obtenir la représentation graphique de la relation entre les valeurs prédites de annearn selon educ et educ, en fixant exper à sa valeur moyenne. Voici les étapes pour y parvenir.

1. Trouver la moyenne de exper et la valeur minimale et maximale de educ.

```
. sum exper educ

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       exper |       616    18.61688    12.32906          0         60
        educ |       616    12.51461    2.733076          6         18
```

Nous avons donc E(exper) = 18,61688. Le minimum et le maximum de educ sont respectivement 6 et 18.

2. Estimer le modèle annearn sur educ et exper afin de trouver les estimateurs.

```
reg annearn educ exper

      Source |       SS       df       MS              Number of obs =     616
-------------+------------------------------           F(  2,   613) =   53.98
       Model |  7.4811e+09     2  3.7406e+09           Prob > F      =  0.0000
    Residual |  4.2478e+10   613    69295171           R-squared     =  0.1497
-------------+------------------------------           Adj R-squared =  0.1470
       Total |  4.9959e+10   615  81234289.6           Root MSE      =  8324.4

------------------------------------------------------------------------------
     annearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    894.514   125.1184     7.15   0.000     648.8012    1140.227
       exper |   243.1261   27.73594     8.77   0.000     188.6571    297.5951
       _cons |  -2833.026    1771.87    -1.60   0.110    -6312.699    646.6466
------------------------------------------------------------------------------
```

Nous avons $\hat{\beta}_0 = -2833{,}026; \hat{\beta}_1 = 894{,}514;\ \hat{\beta}_2 = 243{,}1261$

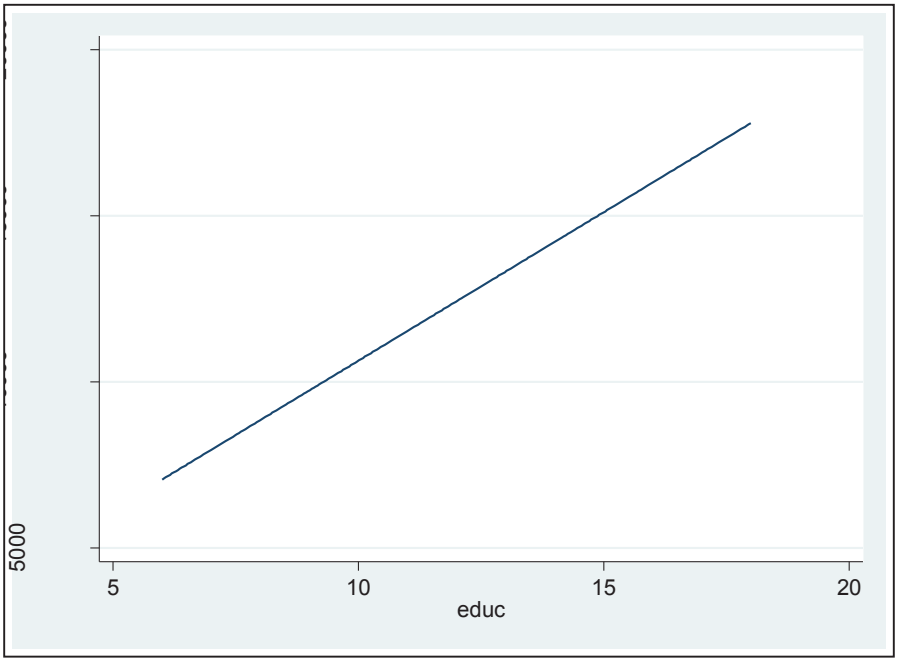
3. Calculer l'intercepte (α) (avec E(exper)) de la fonction $annearn = \alpha + \delta educ, \alpha = \hat{\beta}_0 + E(exper) * \hat{\beta}_2$.

```
display (-2833.026+243.1261*18.61688)
1693.2234
```

4. Générer le graphique en question avec la commande `twoway function`.

```
twoway function y=(1693.2234+894.514*x), range (6 18) ytitle(annearn) xtitle(educ)
```

Le graphique est le suivant :



En gardant exper à sa valeur moyenne, nous pouvons affirmer qu'une année d'éducation supplémentaire provoque en moyenne une hausse du salaire annuel de 894,51 $

La commande marginsplot présente graphiquement les statistiques préalablement calculées grâce à la commande margins que nous avons vue dans la section 3.1.4.

Syntaxe :

```
marginsplot [, options]11
```

Menu :

Statistics > Postestimation > Margins plots and profile plots

[11] Pour plus d'information concernant les différentes options se rattachant à la commande marginsplot, insérer la commande help marginsplot dans STATA 12.

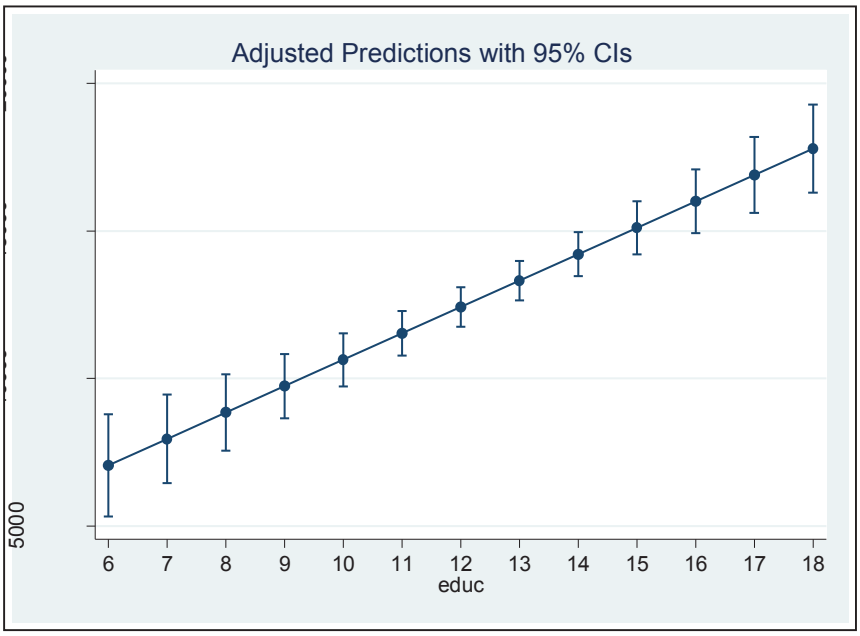
Essayons de reproduire le graphique précédent généré par STATA. Suite à la régression de annearn sur educ et exper, nous devons estimer les valeurs de annearn pour chaque valeur de educ (de 6 à 18) en assumant exper égale à sa moyenne. Nous devons donc utiliser la commande margins, commande déjà analysée dans la section 3.4. Voici la procédure sur STATA.

```
quietly reg annearn educ exper
margins , atmeans at(educ=(6(1)18))

Adjusted prédictions                            Number of obs   =        616
Model VCE    : OLS

Expression   : Linear prediction, predict()

1._at        : educ            =           6
               exper           =    18.61688 (mean)

2._at        : educ            =           7
               exper           =    18.61688 (mean)

3._at        : educ            =           8
               exper           =    18.61688 (mean)

4._at        : educ            =           9
               exper           =    18.61688 (mean)

5._at        : educ            =          10
               exper           =    18.61688 (mean)

6._at        : educ            =          11
               exper           =    18.61688 (mean)

7._at        : educ            =          12
               exper           =    18.61688 (mean)

8._at        : educ            =          13
               exper           =    18.61688 (mean)

9._at        : educ            =          14
               exper           =    18.61688 (mean)

10._at       : educ            =          15
               exper           =    18.61688 (mean)

11._at       : educ            =          16
               exper           =    18.61688 (mean)

12._at       : educ            =          17
               exper           =    18.61688 (mean)

13._at       : educ            =          18
               exper           =    18.61688 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   7060.309   881.4062     8.01   0.000     5332.785    8787.833
          2  |   7954.823   767.1792    10.37   0.000     6451.179    9458.466
          3  |   8849.337   656.9324    13.47   0.000     7561.773     10136.9
          4  |   9743.851   553.0513    17.62   0.000      8659.89    10827.81
          5  |   10638.36   459.8701    23.13   0.000     9737.036    11539.69
          6  |   11532.88   385.2331    29.94   0.000     10777.84    12287.92
          7  |   12427.39   341.5229    36.39   0.000     11758.02    13096.77
```

```
          8  |   13321.91   340.8525    39.08   0.000     12653.85    13989.97
          9  |   14216.42   383.4478    37.08   0.000     13464.88    14967.96
         10  |   15110.94   457.3766    33.04   0.000     14214.49    16007.38
         11  |   16005.45   550.1487    29.09   0.000     14927.18    17083.72
         12  |   16899.96   653.7914    25.85   0.000     15618.56    18181.37
         13  |   17794.48   763.8927    23.29   0.000     16297.27    19291.68
------------------------------------------------------------------------------
```

Nous avons donc un tableau des valeurs estimées de annearn pour chaque valeur de educ lorsque exper = 18,61688 (sa moyenne dans l'échantillon) ainsi que la p-valeur et l'intervalle de confiance. À partir des statistiques calculées à l'aide de margins, nous pouvons générer notre graphique avec la commande marginsplot.

```
marginsplot

Variables that uniquely identify margins: educ
```



`marginsplot` présente également les intervalles de confiance avec un seuil de confiance de 5,0 % pour chaque valeur de educ (de 6 à 18).

2.9 Hétéroscédasticité

2.9.1 Rappel théorique

Tel que relaté dans la section 4.1, les estimateurs MCO sont les meilleurs estimateurs sans biais et à variance minimum notamment si la variance du terme d'erreur est constante pour

toute p-valeur rise par les variables indépendantes (hypothèse numéro 5). Il arrive toutefois que cette hypothèse ne soit pas vérifiée. Nous sommes donc en présence d'hétéroscédasticité :

$$var(u|x_1, x_2, \dots, x_k) = \sigma_i^2$$

Bien que les estimateurs MCO restent sans biais et consistants malgré la présence d'hétéroscédasticité (en supposant que les hypothèses 1 à 4 tiennent), les écarts types obtenus pour les estimations des paramètres du modèle structurel sont biaisés. Ce faisant, nous ne pouvons inférer de façon fiable le niveau de significativité des estimateurs MCO à l'aide des tests de Student et de Fisher. En effet, en présence d'hétéroscédasticité, la statistique t ne suit plus une distribution de Student et ce problème ne peut être résolu en augmenter la taille de l'échantillon, pareillement pour la statistique F.

2.9.2 Tester pour l'hétéroscédasticité

Il existe plusieurs tests dans la littérature économétrique visant à vérifier la présence d'hétéroscédasticité dans le terme d'erreur. Nous verrons essentiellement deux de ces tests : le test de Breusch-Pagan et le test de White pour l'hétéroscédasticité.

Test de Breusch-Pagan pour l'hétéroscédasticité

Le test de Breusch-Pagan est un test visant à détecter toute forme linéaire d'hétéroscédasticité dans le terme d'erreur u. Voici l'intuition générale derrière ce test.

Soit le modèle de régression linéaire à deux variables explicatives suivant :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

Nous savons qu'il y a présence d'homoscédasticité si et seulement si :

$$var(u \,|\, x_1, x_2) = \sigma^2$$

Nous désirons tester l'hypothèse suivante :

$$H_0: var(u \mid x_1, x_2) = \sigma^2$$

Ou

$$H_0: E(u^2|x_1, x_2) = E(u^2) = \sigma^2$$

Ici, l'essentiel du test de Breusch-Pagan repose sur la possibilité que $E(u^2)$ soit corrélée avec x_1 et/ou x_2. Nous pouvons assumer la relation linéaire suivante :

$$u^2 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + v, \text{ avec } E(v|x_1, x_2) = 0$$

De cette équation, l'hypothèse nulle d'existence d'homoscédasticité est :

$$H_0: \alpha_1 = \alpha_2 = 0$$

Malheureusement, u est un paramètre inconnu. Nous devons donc utiliser les résidus, $\hat{u}$, de la régression de y sur x_1 et x_2.

$$\hat{u}^2 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + erreur$$

Avec les résultats de cette estimation, nous utilisons le R^2 (que nous dénoterons $R^2_{\hat{u}^2}$) pour construire la statistique F qui servira à vérifier s'il est possible de rejeter l'hypothèse nulle d'homoscédasticité dans le terme d'erreur u.

$$F = \left(\frac{R^2_{\hat{u}^2}}{1 - R^2_{\hat{u}^2}}\right)\left(\frac{n-k-1}{k}\right), \text{ k étant le nombre de variables indépendantes}$$

En rappel :

- Estimer $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ en assumant l'hypothèse 5 et extraire $\hat{u}$,
- Estimer $\hat{u}^2 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + erreur$.
- Tester $H_0: \alpha_1 = \alpha_2 = 0$ par un test de Fisher
- Si H_0 est rejetée, nous concluons que u est hétéroscédastique.

Test de Breusch-Pagan pour l'hétéroscédasticité – Application sur STATA (estat hettest)

Il est possible de reprendre les quatre étapes précédentes sur STATA afin de tester pour l'hétéroscédasticité du terme d'erreur d'un modèle. Procédons à l'aide d'un exemple en réutilisant le modèle :

$$annearn = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

Voici les étapes à suivre sur STATA:

1. Estimer le modèle annearn sur educ et exper.
   - `quietly reg annearn educ exper`
2. Extraire les résidus de la régression en (1).
   - `predict uhat, resid`
3. Élever au carré les résidus extraits en (2).
   - `gen uhat2=uhat^2`
4. Estimer le modèle $\text{uhat2} = \alpha_0 + \alpha_1 \text{educ} + \alpha_2 \text{exper} + erreur$
   - `quietly reg uhat2 educ exper`
5. Tester l'hypothèse : $H_0: \alpha_1 = \alpha_2 = 0$.
   - `test educ exper`

L'output STATA du test en (5) est la suivante :

```
( 1)  educ = 0
 ( 2)  exper = 0

       F(  2,   613) =    9.42
            Prob > F =    0.0001
```

Sur la base de ces résultats, nous avons suffisamment d'évidence statistique pour rejeter l'hypothèse nulle. Nous pouvons conclure que u est fort probablement hétéroscédastique. Heureusement, il existe une commande qui reprend exactement les prémisses du test de Breusch-Pagan, la commande postestimation `estat hettest`.

Syntaxe :

```
estat hettest
```

Menu :

Statistics > Postestimation > Reports and statistics

Voici un exemple illustrant la fonction de `estat hettest`. Nous reprendrons le modèle de annearn sur educ et exper.

```
quietly reg annearn educ exper
estat hettest


Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of annearn

         chi2(1)      =   357.30
         Prob > chi2  =   0.0000
```

Ici, la commande estat hettest utilise une statistique qui suit une distribution chi2 afin de tester pour l'hétéroscédasticité. Encore une fois, l'hypothèse nulle propose que la variance du terme d'erreur soit constante (homoscédasticité). Sur la base de cet output, Prob > chi2 tend vers zéro. Nous avons ainsi suffisamment d'évidence statistique pour rejeter l'hypothèse nulle. Il y a donc une forte évidence que le terme d'erreur soit hétéroscédastique.

Test de White pour l'hétéroscédasticité

Le concept du test de White pour l'hétéroscédasticité repose sur l'hypothèse (moins forte que l'homoscédasticité) que le terme d'erreur au carré, u^2, est non corrélé avec les x_j, les x_j^2 et les $x_j x_h$ pour $j \neq h$. Si nous utilisons un total de trois variables indépendantes, ce test repose sur l'estimation suivante :

$$\hat{u}^2 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_1^2 + \alpha_5 x_2^2 + \alpha_6 x_3^2 + \alpha_7 x_1 x_2 + \alpha_8 x_1 x_3 + \alpha_9 x_2 x_3 + v \qquad [1]$$

avec E(v|X)=0, X étant un vecteur de tous les régresseurs de l'équation [1]

Le test de White utilise une statistique F qui teste l'hypothèse que tous les coefficients de la régression ci-dessus, excepté pour α_0, sont égaux à zéro. Le problème dans le test de White

résulte dans son abondance de régresseurs dans l'équation [1] (27 régresseurs dans le cas d'une estimation originale avec 6 variables exogènes) jouant ainsi énormément sur l'arbitrage entre le degré de liberté et le nombre de variables indépendantes. L'ajout d'un nombre important de variables explicatives dans l'équation [1] peut rendre le test moins susceptible de produire des résultats significatifs. Il est possible de pallier ce problème en utilisant les valeurs prédites $\hat{y}$ et $\hat{y}^2$ du modèle original et de les remplacer par l'ensemble des régresseurs de l'équation [1] :

$$\hat{u}^2 = \alpha_0 + \alpha_1\hat{y} + \alpha_2\hat{y}^2 + v, \text{ avec } E(v|\hat{y}, \hat{y}^2)=0$$

En effet, $\hat{y}$ est en soit une combinaison linéaire des variables indépendantes, tout comme $\hat{y}^2$. Avec une statistique de Fisher, nous pouvons donc poser l'hypothèse nulle suivante :

$$H_0: \alpha_1 = \alpha_2 = 0, \text{ il y a homoscédasticité dans u}$$

Cette variante est en quelque sorte un cas spécial du test de White. Pour plus de détails, lire White (1990).

En rappel :

1. Estimer $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$ en assumant l'hypothèse 5 et extraire $\hat{u}$ et $\hat{y}$.
2. Estimer $\hat{u}^2 = \alpha_0 + \alpha_1\hat{y} + \alpha_2\hat{y}^2 + v$.
3. Tester $H_0: \alpha_1 = \alpha_2 = 0$ avec un test de Fisher
4. Si H_0 est rejetée, nous concluons que u est hétéroscédastique.

Test de White pour l'hétéroscédasticité – Application sur STATA

Il est possible de reprendre les étapes précédentes afin de générer un test de White en utilisant les valeurs prédites du modèle structurel. Reprenons le modèle annearn sur educ et exper.

Voici les étapes à suivre sur STATA 12:

1. Estimer le modèle $annearn = \beta_0 + \beta_1 educ + \beta_2 exper + u$.
    - `quietly reg annearn educ exper`
2. Extraire les résidus et les valeurs prédites du modèle estimé.
    - `predict uhat, resid`

- predict yhat

3. Générer les résidus et les valeurs prédites au carré.
   - gen uhat2=uhat^2
   - gen yhat2=yhat^2
4. Estimer le modèle $\text{uhat2} = \alpha_0 + \alpha_1 \text{yhat} + \alpha_2 \text{yhat2} + v$
   - reg uhat2 yhat yhat2
5. Tester l'hypothèse nulle suivante : $H_0: \alpha_1 = \alpha_2 = 0$.
   - test yhat yhat2

L'output STATA du test en (5) est la suivante :

```
. test yhat yhat2

 ( 1)  yhat = 0
 ( 2)  yhat2 = 0

       F(  2,   613) =   12.98
            Prob > F =    0.0000
```

Sur la base de ces résultats, nous avons suffisamment d'évidence statistique pour rejeter l'hypothèse nulle que la variance du terme d'erreur est constante (homoscédasticité).

Il existe également deux commandes sur STATA qui génèrent le test de White sur la base des régresseurs $(x_j, \mathrm{x}_\mathrm{j}^2 \text{ et } \mathrm{x_j x_h} \text{ pour } \mathrm{j} \neq \mathrm{h})$ et non des valeurs prédites sur modèle original: les commandes postestimation `whitetst` et `estat imtest`. La commande `whitetst` doit préalablement être installée avant son utilisation. Voici les étapes pour y parvenir :

1. Taper la commande findit whitetst.
2. Sur la fenêtre « Viewer – search whitetst, all », sélectionner l'hyperlien « sg137 from http://www.Stata.com/stb/stb55 ».
3. Cliquer sur « click here to install ». L'installation se fera automatiquement et la commande sera fonctionnelle sur STATA.

Le choix d'utilisation de `whitetst` et `estat imtest` repose essentiellement sur les préférences de l'utilisateur STATA. En effet, les deux commandes utilisent une statistique suivant une distribution chi2 et proposent ainsi exactement les mêmes résultats d'analyse. Le test de White utilisé dans STATA est un cas spécial du test de Breusch-Pagan où on relâche l'hypothèse que les erreurs suivent une distribution normale.

Éclaircissons l'étendue de ces commandes à l'aide d'un exemple. Nous reprendrons le modèle $annearn = \beta_0 + \beta_1 educ + \beta_2 exper + u$.

```
. quietly reg annearn educ exper

. estat imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      59.29      5    0.0000
            Skewness |      17.00      2    0.0002
            Kurtosis |  -2.96e+07      1    1.0000
---------------------+-----------------------------
               Total |  -2.96e+07      8    1.0000
---------------------------------------------------

. whitetst

White's general test statistic :  59.29029  Chi-sq( 5)  P-valeur  =  1.7e-11
```

En comparant les deux outputs de estat imtest et whitetst, nous pouvons conclure (sur la base du chi2 = 59,29), qu'il y a suffisamment d'évidence statistique pour rejeter l'hypothèse nulle d'homoscédasticité dans le terme d'erreur u. Notons que les statistiques portant sur la symétrie de la distribution (skewness) et le coefficient d'aplatissement (kurtosis) ne seront pas détaillées dans ce guide.

Il faut finalement rester prudent dans l'interprétation des tests de Breusch-Pagan et de White. Il n'est pas toujours juste de considérer le rejeter de H_0 comme étant une évidence statistique de l'hétéroscédasticité. En effet, une erreur dans la forme fonctionnelle du modèle peut rendre des tests significatifs alors qu'ils ne le sont pas en réalité. Il faut donc s'assurer que la spécification du modèle soit la bonne (voir section 8)

2.9.3 Estimateurs robustes à l'hétéroscédasticité

Une fois que nous avons détecté la présence d'hétéroscédasticité, il faut forcément corriger la situation. Une méthode de correction fortement utilisée en économétrie consiste à utiliser des écarts types qui tiennent compte de l'hétéroscédasticité, c'est-à-dire des écarts types robustes à l'hétéroscédasticité (HCSE). Les HCSE sont valides asymptotiques lorsque la taille de

l'échantillon est suffisamment large, et ce, en présence ou non d'hétéroscédasticité. Ces écarts types sont calculés à partir de la formule suivante :

$$se(\hat{\beta}_j) = \sqrt{\widehat{var(\hat{\beta}_j)}} = \sqrt{\frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{SCT_j^2}}$$

Où $\hat{r}_{ij}$ représente le résidu i de la régression de x_j sur les autres variables indépendantes.

$SCT_j \ = \sum_{i=1}^{n} (x_j - \bar{x}_j)^2$

Sous les hypothèses 1 à 4, il peut être démontré que $se(\hat{\beta}_j)$ est un estimateur valide de sd($\hat{\beta}_j$). Une fois les HCSE obtenus, nous pouvons construire des statistiques de Student robustes à l'hétéroscédasticité dont la forme générale est :

$$t = \frac{estimé - valeur\ hypothétique}{erreur\ standard}$$

L'unique différence entre la statistique t utilisant les MCO et celle utilisant les HSCE résulte dans le calcul des écarts types. En présence d'homoscédasticité et d'un échantillon petit, il est préférable d'utiliser les estimateurs MCO, car les HCSE ne sont consistants que si n est large. En effet, si le terme d'erreur u est homoscédastique et suit une loi normale, les statistiques t et F suivent exactement les distributions t et F et ce indépendamment de la taille de l'échantillon.

2.9.4 Estimateurs robustes à l'hétéroscédasticité – Application sur STATA

Sur STATA, nous pouvons obtenir les écarts types robustes à l'hétéroscédasticité en insérant l'option `[,robust]` à la fin de la commande `regress depvar [indepvars] [if] [in] [weight]` (voir section 3.1). Voici un exemple afin d'éclaircir le lecteur.

Soit le modèle structurel suivant :

$$annearn \ = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 age + \beta_4 age^2 + \beta_5 tenure + \beta_6 married + \beta_7 male + \beta_8 white + u$$

Où age est l'âge en années
age^2 est la variable age élevé au carré
tenure est le nombre d'années à l'emploi avec l'employeur actuel
married est une variable binaire où married = 1 si l'individu est marié
male est une variable binaire où male = 1 si l'individu est un homme
white est une variable binaire où white = 1 si l'individu est de race blanche

Voici les résultats du modèle prédits. Notons que les données utilisées proviennent encore une fois de FRINGE.dta de Wooldridge (2009).

```
reg annearn educ exper age age2 tenure married male white


      Source |       SS       df       MS              Number of obs =     616
-------------+------------------------------           F(  8,   607) =   26.32
       Model |  1.2866e+10     8  1.6082e+09           Prob > F      =  0.0000
    Residual |  3.7093e+10   607  61109507.1           R-squared     =  0.2575
-------------+------------------------------           Adj R-squared =  0.2477
       Total |  4.9959e+10   615  81234289.6           Root MSE      =  7817.3

------------------------------------------------------------------------------
     annearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   723.7702     120.54     6.00   0.000     487.0442    960.4963
       exper |   272.7116   63.70656     4.28   0.000     147.5996    397.8237
         age |   474.7455   174.6085     2.72   0.007     131.8353    817.6556
        age2 |  -6.968868   2.000022    -3.48   0.001    -10.89667   -3.041065
      tenure |  -33.86111   50.55024    -0.67   0.503    -133.1357    65.41349
     married |   953.6357   768.8835     1.24   0.215     -556.359     2463.63
        male |   4605.671   764.2142     6.03   0.000     3104.846    6106.495
       white |   2309.238   1093.141     2.11   0.035     162.4395    4456.036
       _cons |  -13558.19   3542.676    -3.83   0.000    -20515.58   -6600.801
------------------------------------------------------------------------------
```

Après analyse, nous remarquons que 7 coefficients sur 9 sont significatifs à un seuil d'au plus 5,00 %. Nous avons également procédé à un test d'ajustement du modèle que nous expliciterons dans la section 8. Après vérification, la forme fonctionnelle utilisée ci-dessus est valide ce qui n'aura pas d'impact sur la qualité des tests F pour vérifier l'hétéroscédasticité dans u. Vérifions, à l'aide du test de Breusch-Paggan, s'il existe ou non de l'hétéroscédasticité.

```
estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of annearn

         chi2(1)      =   278.05
         Prob > chi2  =   0.0000
```

Sur la base de ces résultats, nous avons suffisamment d'évidence statistique pour rejeter l'hypothèse d'homoscédasticité. Nous devons donc corriger les écarts types du modèle. Utilisons les écarts types robustes à l'hétéroscédasticité.

```
reg annearn educ exper age age2 tenure married male white, robust

Linear régression                                      Number of obs =     616
                                                       F(  8,   607) =   50.68
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2575
                                                       Root MSE      =  7817.3

------------------------------------------------------------------------------
             |               Robust
     annearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   723.7702      101.5     7.13   0.000     524.4365    923.1039
       exper |   272.7116   151.5382     1.80   0.072    -24.89125    570.3145
         age |   474.7455   260.9258     1.82   0.069    -37.68138    987.1723
        age2 |  -6.968868   2.533816    -2.75   0.006    -11.94498   -1.992757
      tenure |  -33.86111   97.94374    -0.35   0.730    -226.2108    158.4886
     married |   953.6357   624.8977     1.53   0.128    -273.5882     2180.86
        male |   4605.671   799.6281     5.76   0.000     3035.297    6176.044
       white |   2309.238   796.0305     2.90   0.004     745.9297    3872.546
       _cons |  -13558.19   4056.522    -3.34   0.001    -21524.71    -5591.67
```

Notons que les coefficients restent les mêmes, peu importe la méthode d'estimation utilisée (MCO ou robuste). Seuls les écarts types et les tests d'inférence changent. Cette fois-ci, 7 coefficients sur 9 sont significatifs, mais à un seuil d'au plus 10,0 %. Par exemple, pour les variables educ et exper, les statistiques de Student pour la régression MCO sont inférieures à celles obtenues par l'estimation robuste, et ce, en raison d'une surestimation des écarts types MCO. Conjointement, les variables sont statistiquement significatives (la statistique de Fisher étant égale à 50,68). La présentation standard pour représenter les résultats d'une régression utilisant des écarts types robustes à l'hétéroscédasticité est la suivante :

$$\begin{array}{lllllllllll}
\widehat{annearn} = & -13558{,}2 + & 723{,}8\ educ + & 272{,}76\ exper + & 474{,}7\ age - & 6{,}9\ age^2 - & 33{,}9\ tenure + & 953{,}6\ married + & 4605{,}7\ male + & 2309{,}2\ white \\
 & (3542{,}7) & (120{,}5) & (63{,}7) & (174{,}6) & (2{,}0) & (50{,}6) & (768{,}9) & (764{,}2) & (3542{,}7) \\
 & [4056{,}5] & [101{,}5] & [151{,}5] & [260{,}9] & [2{,}5] & [97{,}7] & [624{,}9] & [799{,}6] & [4056{,}5]
\end{array}$$

$$n = 616, \quad R^2 = 0{,}2575, \quad \bar{R}^2 = 0{,}2477$$

Les HCSE sont présentés entre crochets sous les écarts types MCO. Nous avons également ajouté le coefficient de détermination ajusté pour compléter la représentation. Il existe également une autre méthode pouvant en quelque sorte corriger l'hétéroscédasticité. Celle-ci consiste à transformer en logarithme naturel la variable dépendante de sorte à obtenir un modèle log-level (nous verrons plus en détail l'étendue de ce modèle dans la section 7). Dans la littérature scientifique, les variables de revenus ou de salaires sont généralement exprimées en logarithme naturel dans le processus de modélisation. Reprenons le modèle structurel de

l'exemple précédent en remplaçant annearn par log(annearn) comme variable dépendante. Celle-ci est d'ailleurs déjà disponible dans la base de données.

```
reg lannearn educ exper age age2 tenure married male white


      Source |       SS       df       MS              Number of obs =     616
-------------+------------------------------           F(  8,   607) =   58.79
       Model |  95.2846082     8  11.910576            Prob > F      =  0.0000
    Residual |  122.974306   607  .202593584           R-squared     =  0.4366
-------------+------------------------------           Adj R-squared =  0.4291
       Total |  218.258914   615  .354892543           Root MSE      =   .4501

------------------------------------------------------------------------------
    lannearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0550772   .0069405     7.94   0.000     .0414469    .0687074
       exper |   .0131631   .0036681     3.59   0.000     .0059594    .0203668
         age |   .0660754   .0100537     6.57   0.000     .0463312    .0858196
        age2 |  -.0008918   .0001152    -7.74   0.000    -.0011179   -.0006656
      tenure |   .0076887   .0029106     2.64   0.008     .0019726    .0134048
     married |   .0382133    .044271     0.86   0.388    -.0487295    .1251562
        male |   .4841225   .0440021    11.00   0.000     .3977076    .5705373
       white |   .1526568   .0629412     2.43   0.016     .0290479    .2762657
       _cons |   6.744754   .2039811    33.07   0.000     6.344159    7.145348
------------------------------------------------------------------------------
```

Ici, il faut être prudent dans la comparaison de ce modèle avec celui ayant comme variable dépendante annearn. En effet, bien que log(annearn) contienne implicitement annearn, les deux modèles ne possèdent pas la même variable expliquée. Les résultats ici obtenus, nous pouvons utiliser la commande postestimation `estat hettest` pour tester pour l'hétéroscédasticité.

```
    . estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
             Ho: Constant variance
             Variables: fitted values of lannearn

             chi2(1)      =     2.30
             Prob > chi2  =   0.1292
```

Sur la base de ces résultats, nous ne pouvons pas rejeter l'hypothèse nulle d'homoscédasticité. Ce faisant, la simple transformation logarithmique de annearn a permis de régler le problème d'hétéroscédasticité.

2.9.5 Les estimateurs moindres carrés pondérés (MCP)

La méthode des moindres carrés pondérés est une méthode d'estimation qui vise à corriger pour l'hétéroscédasticité lorsque la variance du terme d'erreur conditionnelle aux variables explicatives prend la forme suivante :

$var(u|X) = \sigma^2 h(X)$, X dénotant l'ensemble des variables explicatives

Ici, h(X) est une fonction des régresseurs déterminant l'hétéroscédasticité dans la variance de u. Comme une variance est positive, $h(X) > 0, \forall\, x_{ij}$. Évidemment, h(X) n'est pas toujours connu par le chercheur. Supposons ici que cette fonction est connue. Il est possible de transformer les estimateurs MCO contenant des erreurs hétéroscédastiques en estimateurs contenant des erreurs homoscédastiques. En effet, sachant cette forme,

$$var(u|X) = E(u^2|X) = \sigma^2 h(X)$$

Nous pouvons construire un poids pouvant éradiquer la corrélation entre la variance de u et les régresseurs.

$$var\left(\frac{u}{\sqrt{h(X)}}\right) = E\left(\left(u/\sqrt{h(X)}\right)^2\right) = \frac{\sigma^2 h(X)}{h(x)} = \sigma^2$$

$$\rightarrow var\left(u|X/\sqrt{h(X)}\right) = \frac{\sigma^2 h(X)}{h(X)} = \sigma^2$$

Ce faisant, nous pouvons obtenir des estimateurs plus efficients que ceux des MCO en pondérant par 1/h(X) le modèle structurel suivant :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u \quad [2]$$

$$\frac{1}{\sqrt{h}}(y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u)$$

$$y^* = \beta_0 + \beta_1 \mathrm{x}_1^* + \beta_2 \mathrm{x}_2^* + \cdots + \beta_k \mathrm{x}_\mathrm{k}^* + u^* \quad [3]$$

Ici, si l'équation originale [2] respecte les hypothèses 1 à 4, l'équation [3] les satisfait également, et ce, avec l'hypothèse d'homoscédasticité en plus. De plus, si u suit une distribution normale, u^* suit aussi la même distribution. Ce faisant, les estimateurs obtenus de [3], soient les estimateurs MCP (moindres carrés pondérés), sont BLUE et plus efficaces que les estimateurs MCO. La méthode MCP est un cas particulier de la méthode dite des moindres carrés généralisés (MCG). En moyenne, les estimateurs MCP sont estimations avec plus de précision (avec un écart type plus petit) par rapport aux estimateurs MCO. Cependant, il est possible que les écarts types associés soient plus grands advenant un problème dans la spécification de h(X)

ou du modèle économétrique. Effectivement, avec la méthode des MCP, nous attribuons un poids moins important sur les variables ayant une variance élevée comparativement à la méthode des MCO qui traite toutes les observations identiquement.

2.9.6 Les estimateurs moindres carrés pondérés (MCP) – Application sur STATA

La commande d'option `aweights` ou `aw` (avec syntaxe `regress depvar indepvars [aweight = 1/w]`) permet entre autres de réaliser une estimation MCP. Cette commande appose un poids proportionnel à la variance d'une observation pour toute observation, où nous assumons que la variance de l'observation j est σ^2/w_j, où w_j est le poids de l'observation j. STATA implémente également le poids w tel que $\sum_{j=1}^{n} \frac{1}{w_j}$.

Présentons cette commande à l'aide d'un exemple qui utilise les données 401KSUBS.dta de Wooldridge (2009).

Nous désirons estimer le modèle suivant :

$$nettfa = \beta_0 + \beta_1 \text{inc} + u$$

Où nettfa est le total des actifs net en 1000$ d'un individu
inc est le revenu d'un individu

Effectuons la première estimation MCO

```
. reg nettfa inc if fsize==1

      Source |       SS       df       MS              Number of obs =    2017
-------------+------------------------------           F(  1,  2015) =  181.60
       Model |  377482.064     1  377482.064           Prob > F      =  0.0000
    Residual |  4188482.98  2015   2078.6516           R-squared     =  0.0827
-------------+------------------------------           Adj R-squared =  0.0822
       Total |  4565965.05  2016  2264.86361           Root MSE      =  45.592

------------------------------------------------------------------------------
      nettfa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   .8206815      .0609    13.48   0.000     .7012479     .940115
       _cons |  -10.57095   2.060678    -5.13   0.000    -14.61223   -6.529671
------------------------------------------------------------------------------
```

L'estimation s'est réalisée sur l'échantillon des familles d'une personne seulement (`if fsize==1`). Nous notons ici que tous les coefficients sont significatifs au seuil de 1,00 %. Vérifions s'il y a présence d'hétéroscédasticité dans le modèle à l'aide du test de Breusch-Pagan pour l'hétéroscédasticité.

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of nettfa

         chi2(1)      =  1177.05
         Prob > chi2  =   0.0000
```

Nous voyons sur la base de ces résultats que nous pouvons rejeter l'hypothèse nulle d'homoscédasticité. Assumons que la variance du terme d'erreur u soit proportionnellement linéaire au revenu (inc) et suive la forme suivante :

$$var(u|X) = \sigma^2 inc$$

Voici la procédure pour obtenir les estimateurs MCP.

```
reg nettfa inc if fsize==1 [aw=1/inc]
(sum of wgt is   8.7566e+01)

      Source |       SS       df       MS              Number of obs =    2017
-------------+------------------------------           F(  1,  2015) =  153.71
       Model |  184535.798     1  184535.798           Prob > F      =  0.0000
    Residual |  2419034.45  2015  1200.51338           R-squared     =  0.0709
-------------+------------------------------           Adj R-squared =  0.0704
       Total |  2603570.25  2016   1291.4535           Root MSE      =  34.648

------------------------------------------------------------------------------
      nettfa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   .7870523   .0634814    12.40   0.000     .6625562    .9115484
       _cons |  -9.580702   1.653284    -5.79   0.000    -12.82303   -6.338378
```

Voici un tableau comparatif des résultats de l'estimation par MCO et MCP.

Variable dépendante : nettfa

Variable indépendante	**(1) MCO**	**(2) MCP**
inc	0,821	0,787

	(0,061)	(0,0635)
constante	-10,571	-9,581
	(2,061)	(1,653)
Observations	2017	2017
R^2	0,0827	0,0709

Nous remarquons que les estimateurs MCO et MCP restent relativement semblables; l'écart entre les bêtas des deux modèles pour la variable inc et la constante étant respectivement de 0,034 et -0,99. Les écarts types vont également dans le même sens. Notons que les deux R^2 ne sont pas comparables, la variable dépendante étant différente d'un modèle à l'autre. Tel que discuté dans la partie théorique de la sous-section précédente, nous pouvons noter que les écarts types de l'estimation MCP sont inférieurs à ceux de l'estimation MCO.

2.9.7 La méthode des moindres carrés généralisés faisables

Il s'avère souvent difficile de connaître la fonction h(X). Heureusement, nous pouvons modéliser la fonction h et utiliser les données que nous disposions afin d'estimer les paramètres inconnus du modèle d'intérêt. Ainsi, nous utilisons $\hat{h}(X)$ au lieu de h(X), car ce paramètre est inconnu. Cette méthode d'estimation se nomme la méthode des moindres carrés généralisés faisables (MCGF). Il existe plusieurs méthodes permettant d'estimer l'hétéroscédasticité. Supposons qu'elle prend la forme suivante :

$$var(u|X) = \sigma^2 h(X) = \sigma^2 e^{(\alpha_0+\alpha_1 x_1+\alpha_2 x_2+\cdots+\alpha_k x_k)}$$

où les x sont les variables indépendantes du modèle structurel original et les α des paramètres inconnus.

La raison qui pousse le choix de cette fonction exponentielle est de garantir que les variances soient positives. En effet, supposer une forme linéaire ne garantit pas cette nécessité. L'objectif ici est d'estimer h(X) (donc les α) afin de construire le bon poids qui corrigera pour

l'hétéroscédasticité. Transformons $var(u|X)$ afin qu'elle devienne linéaire en extrayant son logarithme naturel.

$$var(u|X) = \sigma^2 e^{(\alpha_0+\alpha_1 x_1+\alpha_2 x_2+\cdots+\alpha_k x_k)}$$

$$u^2 = \sigma^2 e^{(\alpha_0+\alpha_1 x_1+\alpha_2 x_2+\cdots+\alpha_k x_k)}$$

$$g = \log(u^2) = \delta_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k + v, \quad [4]$$

Où $\delta_0 = \log(\sigma^2) + \alpha_0$
v est le terme d'erreur et nous supposons que E(v|X)=0.

Nous assumons ici que [4] satisfait les hypothèses de Gauss-Markov. Nous pouvons, de ce fait, estimer les α avec la méthode des MCO. Comme u est inconnu, nous utilisons les résidus $\hat{u}$ afin d'estimer g. Nous pouvons dès lors obtenir le poids $1/\hat{h}$ en reprenant la valeur exponentielle de $\hat{g} \longrightarrow \hat{h} = e^{\hat{g}}$.

Fait important : les estimateurs MCGF sont biaisés du fait que h est estimé (ils ne sont plus BLUE). Ceux-ci restent cependant consistants et asymptotiquement plus efficients que les estimateurs MCO. S'il y a présence d'hétéroscédasticité, l'utilisation des MCGF est préférée aux MCO si la taille de l'échantillon est suffisamment grande.

En rappel :

1. Estimer avec les MCO le modèle $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$ et prédire $\hat{u}$.
2. Générer $log(\hat{u}^2)$.
3. Régresser $log(\hat{u}^2)$ sur tous les régresseurs.
4. Trouver les valeurs prédites $\hat{g}_i$ de la régression en (3).
5. Générer $\hat{h} = e^{\hat{g}}$.
6. Estimer la régression $\frac{y}{h} = \beta_0 + \beta_1 \frac{x_1}{h} + \beta_2 \frac{x_2}{h} + \cdots + \beta_k \frac{x_k}{h} + u$

2.9.8 La méthode des moindres carrés généralisés faisables – Application sur STATA

En supposant que h(X) prend une forme exponentielle exposée précédemment, nous pouvons réaliser une régression MCGF sur STATA en répétant le même processus théorique que dans la section 6.8. Nous baserons cet exemple sur celui de Wooldridge (2009) qui utilise les données SMOKE.dta. Nous cherchons à évaluer les déterminants qui influencent la consommation hebdomadaire de cigarettes. Soit le modèle suivant :

$$cigs = \beta_0 + \beta_1 \log(\text{income}) + \beta_2 \log(\text{cigprice}) + \beta_3 \text{educ} + \beta_4 \text{age} + \beta_5 \text{age}^2 + \beta_6 \text{restaurn} + u$$

Où cigs est le nombre de cigarettes fumées par jour
log(income) est le logarithme naturel du revenu annuel
log(cigprice) est le logarithme naturel du prix (en cennes) d'un paquet de cigarettes
educ est le nombre d'années d'éducation
age est l'âge en année et age^2 est la variable age élevée au carré
restaurn est une variable binaire où restaurn=1 si l'individu vit dans un État où il existe des interdictions de fumer dans un restaurant

1. Estimons le modèle de cigs sur log(income), log(cigprice), etc. avec les MCO. Notons qu'il faut générer la variable age^2 étant donné qu'elle n'existe pas dans les données. S'il y a hétéroscédasticité dans les erreurs, régressons ce modèle à l'aide des MCGF.

```
use http://fmwww.bc.edu/ec-p/data/wooldridge/SMOKE
gen age2=age^2
reg cigs lincome lcigpric educ age age2 restaurn
```

```
      Source |       SS       df       MS              Number of obs =     807
-------------+------------------------------           F(  6,   800) =    7.42
       Model |  8003.02506     6  1333.83751           Prob > F      =  0.0000
    Residual |  143750.658   800  179.688322           R-squared     =  0.0527
-------------+------------------------------           Adj R-squared =  0.0456
       Total |  151753.683   806  188.280003           Root MSE      =  13.405

------------------------------------------------------------------------------
        cigs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lincome |   .8802689   .7277838     1.21   0.227    -.5483223     2.30886
    lcigpric |  -.7508498   5.773343    -0.13   0.897    -12.08354    10.58184
        educ |  -.5014982   .1670772    -3.00   0.003    -.8294597   -.1735368
         age |   .7706936   .1601223     4.81   0.000      .456384    1.085003
        age2 |  -.0090228    .001743    -5.18   0.000    -.0124443   -.0056013
    restaurn |  -2.825085   1.111794    -2.54   0.011    -5.007462   -.6427079
       _cons |  -3.639884   24.07866    -0.15   0.880     -50.9047    43.62493
------------------------------------------------------------------------------
```

4 coefficients sur 7 sont significatifs: educ, age, age^2 et restaurn. Tous les effets prédits sont également cohérents intuitivement. Vérifions maintenant à l'aide d'un test de Breusch-Paggan s'il y a présence d'hétéroscédasticité dans le terme d'erreur.

```
estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of cigs

         chi2(1)      =    56.17
         Prob > chi2  =   0.0000
```

Nous voyons ici qu'il y a de fortes évidences qu'il existe de l'hétéroscédasticité dans u. Corrigeons la situation avec les estimateurs MCGF.

2. Obtenons les résidus, élevons-les au carré et transformons-les en logarithme naturel. Après ces étapes, nous aurons dérivé la variable constituant le logarithme naturel des résidus au carré.

```
predict uhat, resid

gen uhat2=uhat^2

gen luhat2=log(uhat2)
```

3. Estimons l'équation suivante :

$$g = \log(u^2) = \delta_0 + \alpha_1 \log(\text{income}) + \alpha_2 \log(\text{cigprice}) + \alpha_3 \text{educ} + \alpha_4 \text{age} + \alpha_5 \text{age}^2 + \alpha_6 \text{restaurn} + v.$$

```
quietly reg luhat2 cigs lincome lcigpric educ age age2 restaurn
```

4. Obtenons les valeurs prédites de la régression en (3) et construisons les poids finaux estimations ($\hat{h}$) de sorte que $\hat{h} = e^{\hat{g}}$.

```
predict fitted
(option xb assumed; fitted values)

gen hhat=exp(fitted)
```

5. Effectuons la régression MCGF.

```
reg cigs lincome lcigpric educ age age2 restaurn [aw=1/hhat]
```

```
(sum of wgt is   1.9977e+01)

      Source |       SS       df       MS              Number of obs =     807
-------------+------------------------------           F(  6,   800) =   17.06
       Model |  10302.6415     6  1717.10692           Prob > F      =  0.0000
    Residual |  80542.0684   800  100.677586           R-squared     =  0.1134
-------------+------------------------------           Adj R-squared =  0.1068
       Total |    90844.71   806  112.710558           Root MSE      =  10.034

------------------------------------------------------------------------------
        cigs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lincome |   1.295241   .4370118     2.96   0.003     .4374154    2.153066
    lcigpric |   -2.94028   4.460142    -0.66   0.510    -11.69524    5.814684
        educ |  -.4634462   .1201586    -3.86   0.000    -.6993095   -.2275829
         age |   .4819474   .0968082     4.98   0.000     .2919194    .6719755
        age2 |  -.0056272   .0009395    -5.99   0.000    -.0074713   -.0037831
    restaurn |  -3.461066   .7955047    -4.35   0.000    -5.022589   -1.899543
       _cons |    5.63533   17.80313     0.32   0.752    -29.31103    40.58169
```

Voici un tableau comparatif des résultats entre le modèle MCO et MCGF.

Variable dépendante : cigs

Variable indépendante	**(1) MCO**	**(2) MCGF**
lincome	0,88	1,295
	(0,728)	(0,437)
lcigpric	-0,751	-2,94
	(5,773)	(4,46)
educ	-0,501	-0,463
	(0,167)	(0,12)
age	0,771	0,481
	(0,16)	(0,097)
age^2	-0,009	-0,006
	(0,002)	(0,001)
restaurn	-2,825	-3,461

	(1,112)	(0,796)
constante	-3,64	5,634
	(24,079)	(17,803)
Observations	807	807
R^2	0,0527	0,1134

Encore une fois, nous remarquons que tous les écarts types des estimateurs MCGF sont inférieurs à ceux des MCO. L'effet revenu est également plus important et maintenant statistiquement significatif avec la méthode des MCGF. Tous les coefficients gardent également le même signe et restent sensiblement de même amplitude d'une estimation à l'autre. En ce sens, si les coefficients des MCO et des MCP/MCGF sont très différents (par exemple de signes opposés), il est fort possible qu'il y ait une erreur de spécification dans le modèle utilisé (construction des variables, manque de variables, absence de constante, etc.). Nous verrons dans la section 8 comment régler un problème de spécification. Qu'arrive-t-il si la forme estimée de h(X) est fausse, par exemple que $var(u|X) \neq \sigma^2 e^{(\alpha_0+\alpha_1 x_1+\alpha_2 x_2+\cdots+\alpha_k x_k)}$. Tout d'abord, les tests t et F obtenus sous l'hypothèse $var(u|X) = \sigma^2 h(X)$ ne sont plus valides et ce quelque soit la taille de l'échantillon collectée. Les estimateurs restent cependant sans biais si E(u|X)=0. Finalement, advenant une telle problématique, l'utilisation des écarts types robustes à l'hétéroscédasticité reste la solution la plus optimale.

2.9.9 Observer l'hétéroscédasticité par graphique

Une méthode visant à vérifier la présence d'hétéroscédasticité dans le terme d'erreur est de représenter graphiquement l'allure des résidus grâce à deux commandes postestimation.

Nous pouvons représenter dans un graphique à deux dimensions la distribution des résidus en fonction des valeurs prédites d'un modèle dans le cas où nous croyons que la variance de u est proportionnelle à tous les régresseurs.

Syntaxe :

```
rvfplot, yline(#)
```

où `yline(#)` est une option qui ajoute une droite horizontale au graphique.

Menu :

Graphics > Residual diagnostic plots > Residual versus fitted

Advenant le cas où nous soupçonnons que la variance de u est proportionnelle à une variable indépendante particulière, il est également possible de construire un graphique des résidus en fonction du régresseur en question.

Syntaxe :

```
rvpplot x1, yline(#)
```

où x1 est la variable indépendante corrélée avec la variance de u

Menu :

Graphics > Residual diagnostic plots > Residual versus predictor

Une dernière méthode consiste à prédire les résidus et de construire un graphique à partir des résidus générés.

```
reg y x1 x2 x3
predict uhat, resid
twoway (scatter uhat x1), yline(#)
```

où x1 est le régresseur susceptible d'être corrélé avec la variance de u

Il y a trois éléments importants à identifier lorsque nous analysons un graphique de résidus afin d'inspecter pour l'hétéroscédasticité :

1. Est-ce qu'il y une tendance observable dans les résidus? Si oui, il y a un risque d'hétéroscédasticité.
2. Est-ce que la variabilité dans les y est constante pour tout x. Sinon, il y a un risque d'hétéroscédasticité.
3. Est-ce qu'environ 95,0 % des résidus converge autour de deux écarts types à partir de l'axe y=0. Sinon, il y a risque d'hétéroscédasticité.

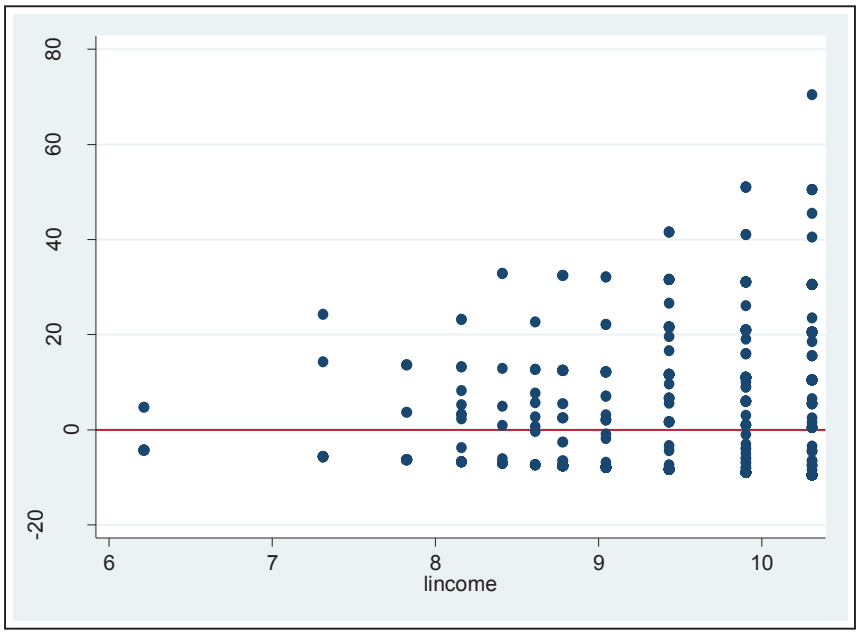
Voici un exemple qui nous aidera à identifier s'il y a présence d'hétéroscédasticité à partir du graphique des résidus. Avec les données SMOKE.dta de Wooldridge (2009), nous cherchons à

vérifier si le revenu à un effet sur la consommation hebdomadaire de cigarettes. Voici le modèle à estimer :

$$cigs = \beta_0 + \beta_1 \log(income) + u$$

Avec un graphique que nous construirons sur STATA, vérifions si la variance des résidus est proportionnelle à log(income).

```
quietly reg cigs lincome

rvpplot lincome, yline(0)
```



Par simple observation, nous voyons clairement que la disparité dans les résidus augmente avec le revenu (en logarithme naturel). La variabilité dans les résidus n'est pas constante pour toute valeur de lincome. Ce faisant, il y a de fortes évidences que la variance de u conditionnelle à lincome soit hétéroscédastique. L'analyse des résidus par graphique s'avère en soit un excellent complément aux tests que nous avons vus précédemment.

2.10 Les différentes formes fonctionnelles de variables

2.10.1 Variables muettes

Rappel théorique

Dans la plupart des travaux empiriques en économétrie, il arrive fréquemment d'inclure de l'information qualitative dans les modèles d'étude. Une façon commune de procéder est d'inclure cette information qualitative sous forme de variable binaire. Une variable binaire prend la valeur 0 ou 1 selon la définition donnée à la variable. Définissons ce concept.

Soit x une variable binaire, celle-ci prend les valeurs :

$x = 1$, s'il y a « succès »

$x = 0$, sinon

Le concept de « succès » dépend de la définition de la variable. L'exemple classique consiste à une variable qui prend la valeur 1 lorsque l'observation est un homme et 0 lorsque l'observation est une femme (le « succès » ici serait donc « homme »). Illustrons le concept de variable binaire dans le cas d'une régression multiple. Nous cherchons à vérifier s'il existe des disparités de revenus entre les hommes et les femmes avec le modèle suivant :

$$\log(annearn) = \beta_0 + \beta_1 male + \beta_2 educ + u$$

Où log(annearn) est le logarithme naturel du revenu annuel en 1000 de $ d'un individu
male une variable binaire où male=1 si l'observation est un homme
educ est le nombre d'années de scolarité

La variable d'intérêt ici est donc male et nous contrôlons pour educ. Estimons le modèle à l'aide de STATA.

```
reg lannearn educ male, robust

Linear régression                                      Number of obs =     616
                                                       F(  2,   613) =  144.64
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.3110
                                                       Root MSE      =  .49529

------------------------------------------------------------------------------
             |               Robust
    lannearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0547779   .0066946     8.18   0.000     .0416307    .0679252
        male |   .6141117   .0432875    14.19   0.000     .5291019    .6991215
       _cons |   8.214438   .0900295    91.24   0.000     8.037634    8.391242
------------------------------------------------------------------------------
```

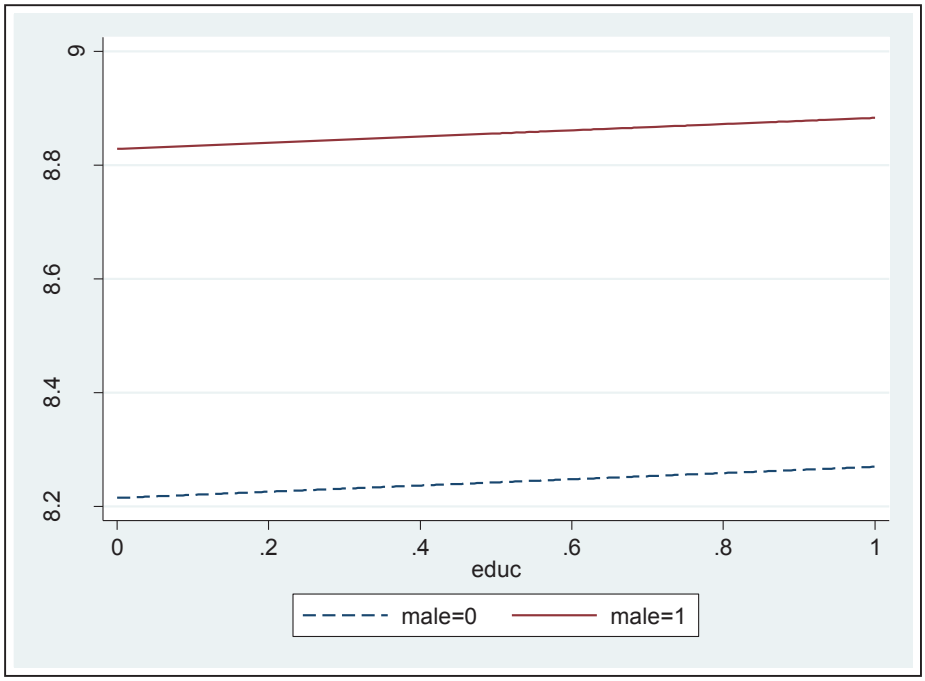
Notons que nous avons utilisé les estimateurs robustes à l'hétéroscédasticité suite à la conduite préalable d'un test de Breusch-Paggan significatif. Le coefficient d'intérêt est celui de male : 0,614. Cet effet partiel étant d'autant plus significatif, le modèle prédit qu'être un homme augmente le revenu annuel en millier de $ de 0,614 %. Il y donc une discrimination positive à l'égard des hommes par rapport aux femmes.

En termes d'espérance, si nous supposons que E(u|male, educ)=0, le paramètre β_0 s'interprète comme suit :

$$\beta_0 = E(\log(annearn)\,|male = 1, educ) - E(\log(annearn)|male = 0, educ)$$

Cette différence peut être dépeinte graphiquement par un changement d'intercepte entre les hommes et les femmes. Représentons cette différence graphiquement à l'aide de STATA.

```
twoway (function 8.214438+0.0547779*x) (function 8.8285497+0.0547779*x), ytitle(annearn) xtitle(educ)
```



Afin de construire ce graphique, nous avons utilisé les résultats de la régression de annearn sur male et educ. La droite en pointillés suit la forme suivante :

$$E(annearn|male = 0) = \hat{\beta}_0 + \hat{\beta}_1 E(male|male = 0) + \hat{\beta}_2 E(educ|male = 0)$$

$$E(annearn|male = 0) = \hat{\beta}_0 + \hat{\beta}_2 educ$$

$$E(annearn|male = 0) = 8{,}21 + 0{,}05 educ$$

La droite solide suit la forme suivante :

$$E(annearn|male = 1) = \hat{\beta}_0 + \hat{\beta}_1 E(male|male = 1) + \hat{\beta}_2 E(educ|male = 1)$$

$$E(annearn|male = 1) = \left(\hat{\beta}_0 + \hat{\beta}_1\right) + \hat{\beta}_2 educ$$

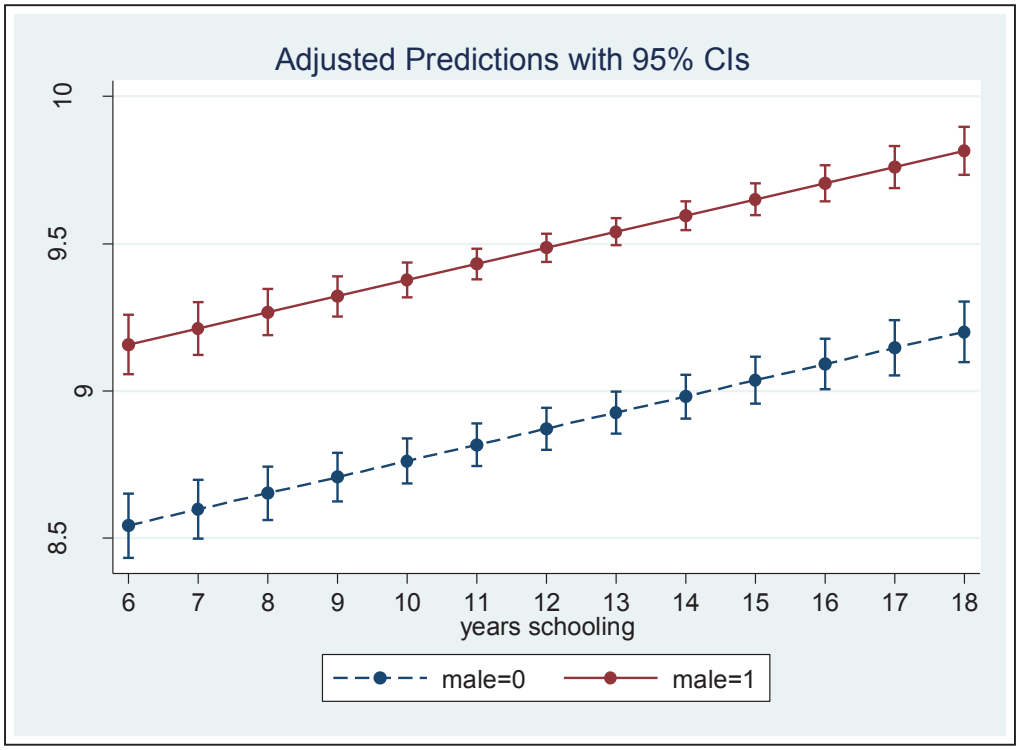
$$E(annearn|male = 1) = 8{,}83 + 0{,}05 educ$$

La différence d'intercepte dans le graphique représente ainsi la discrimination au niveau du revenu annuel des femmes par rapport aux hommes. Pour construire notre graphique, il est également plus rapide de procéder à l'aide de la commande postestimation margins. Voici la démarche.

```
quietly reg lannearn educ male,
margins , atmeans at(educ=(6(1)18) male=0 male=1)
```

```
Adjusted prédictions                            Number of obs   =        616
Model VCE    : Robust

Expression   : Linear prediction, predict()

1._at        : educ            =           6
               male            =           0

2._at        : educ            =           6
               male            =           1

3._at        : educ            =           7
               male            =           0

4._at        : educ            =           7
               male            =           1

5._at        : educ            =           8
               male            =           0

6._at        : educ            =           8
               male            =           1

7._at        : educ            =           9
               male            =           0

8._at        : educ            =           9
               male            =           1

9._at        : educ            =          10
               male            =           0

10._at       : educ            =          10
               male            =           1

11._at       : educ            =          11
               male            =           0

12._at       : educ            =          11
               male            =           1

13._at       : educ            =          12
               male            =           0

14._at       : educ            =          12
               male            =           1

15._at       : educ            =          13
               male            =           0

16._at       : educ            =          13
               male            =           1

17._at       : educ            =          14
               male            =           0

18._at       : educ            =          14
               male            =           1

19._at       : educ            =          15
               male            =           0

20._at       : educ            =          15
               male            =           1

21._at       : educ            =          16
               male            =           0

22._at       : educ            =          16
               male            =           1
```

```
23._at       : educ            =          17
               male            =           0

24._at       : educ            =          17
               male            =           1

25._at       : educ            =          18
               male            =           0

26._at       : educ            =          18
               male            =           1

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   8.543106   .0556551   153.50   0.000     8.434024    8.652188
          2  |   9.157217   .0514989   177.81   0.000     9.056281    9.258153
          3  |   8.597884     .05076   169.38   0.000     8.498396    8.697371
          4  |   9.211995   .0456458   201.81   0.000     9.122531    9.301459
          5  |   8.652662   .0463174   186.81   0.000     8.561881    8.743442
          6  |   9.266773   .0400572   231.34   0.000     9.188262    9.345284
          7  |   8.707439   .0424696   205.03   0.000     8.624201    8.790678
          8  |   9.321551   .0348606   267.40   0.000     9.253226    9.389877
          9  |   8.762217   .0393912   222.44   0.000     8.685012    8.839423
         10  |   9.376329   .0302584   309.87   0.000     9.317024    9.435635
         11  |   8.816995   .0372733   236.55   0.000     8.743941     8.89005
         12  |   9.431107   .0265617   355.06   0.000     9.379047    9.483167
         13  |   8.871773   .0362846   244.50   0.000     8.800657     8.94289
         14  |   9.485885   .0241891   392.16   0.000     9.438475    9.533295
         15  |   8.926551   .0365169   244.45   0.000     8.854979    8.998123
         16  |   9.540663   .0235443   405.22   0.000     9.494517    9.586809
         17  |   8.981329   .0379477   236.68   0.000     8.906953    9.055705
         18  |   9.595441   .0247628   387.49   0.000     9.546907    9.643975
         19  |   9.036107   .0404501   223.39   0.000     8.956826    9.115388
         20  |   9.650219   .0275988   349.66   0.000     9.596126    9.704311
         21  |   9.090885    .043841   207.36   0.000     9.004958    9.176812
         22  |   9.704997     .03162   306.93   0.000     9.643023    9.766971
         23  |   9.145663   .0479321   190.80   0.000     9.051718    9.239608
         24  |   9.759775   .0364362   267.86   0.000     9.688361    9.831188
         25  |   9.200441   .0525602   175.05   0.000     9.097425    9.303457
         26  |   9.814553   .0417732   234.95   0.000     9.732679    9.896427
------------------------------------------------------------------------------

marginsplot

Variables that uniquely identify margins: educ male
```

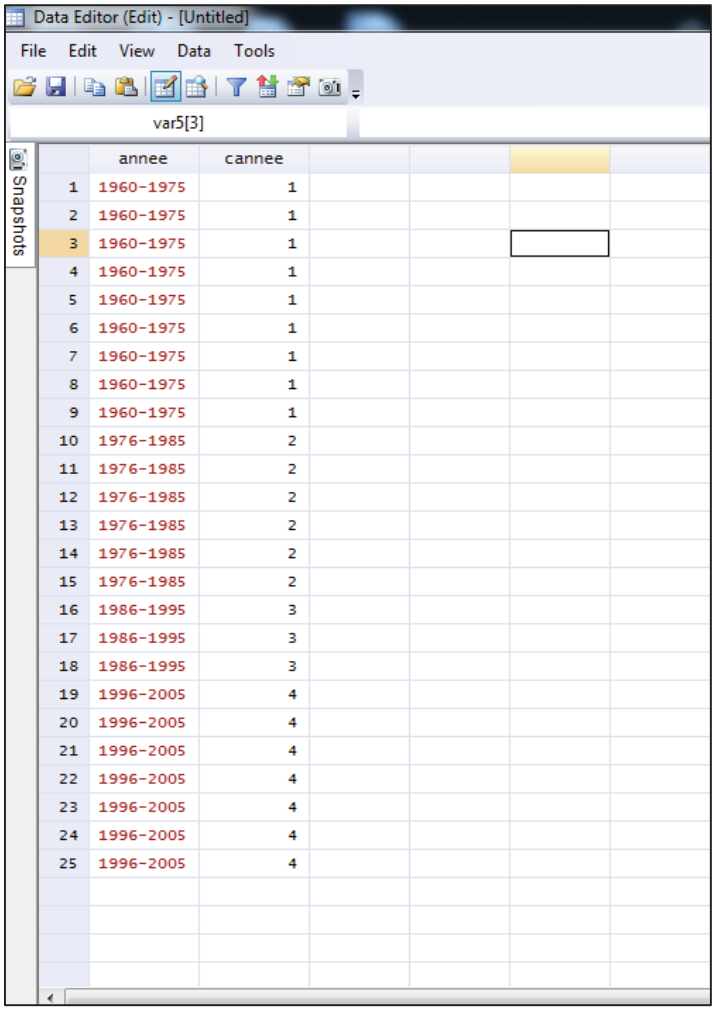


La droite supérieure (la ligne solide) représente en soit les hommes tandis que la droite inférieure (la ligne pointillée) représente les femmes.

Générer une variable muette sur STATA

Il est possible de générer des variables binaires sur STATA notamment à l'aide de la commande `replace`. Cette commande a déjà été présentée dans la partie introductive de ce guide. Nous ne ferons qu'illustrer un exemple utilisant cette commande dans cette section.

Voici un classeur fictif qui répartit un ensemble de logements selon l'année de construction. Un logement construit entre 1960-1975 prend la valeur 1; un logement construit entre 1976-1985 prend la valeur 2, etc.



	annee	cannee
1	1960-1975	1
2	1960-1975	1
3	1960-1975	1
4	1960-1975	1
5	1960-1975	1
6	1960-1975	1
7	1960-1975	1
8	1960-1975	1
9	1960-1975	1
10	1976-1985	2
11	1976-1985	2
12	1976-1985	2
13	1976-1985	2
14	1976-1985	2
15	1976-1985	2
16	1986-1995	3
17	1986-1995	3
18	1986-1995	3
19	1996-2005	4
20	1996-2005	4
21	1996-2005	4
22	1996-2005	4
23	1996-2005	4
24	1996-2005	4
25	1996-2005	4

Sans supprimer les variables déjà présentes dans le classeur, nous désirons ici construire une variable (construc) qui prendra les valeurs suivantes :

construc=1 si le logement est construit entre 1986 et 2005
construc=0 si autrement

Voici la procédure à suivre sur STATA.

1. Créer la variable construc qui s'avère la même que cannee.

```
gen construc=cannee
```

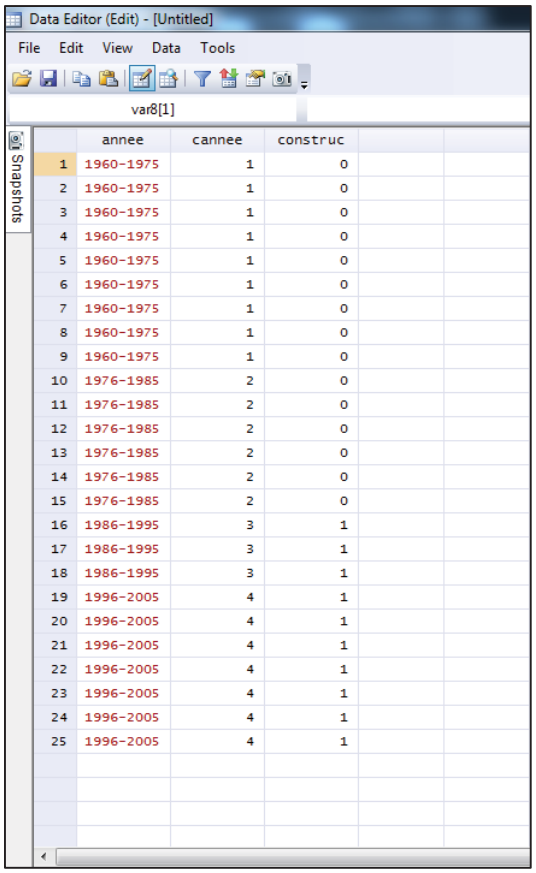
2. Donner la valeur 0 à construc pour toutes les valeurs de construct prenant la valeur 1 ou 2.

```
replace cannee=0 if (cannee==1 | cannee==2)
(15 real changes made)
```

3. Donner la valeur 1 à construc pour toutes les valeurs de construct prenant la valeur 3 ou 4.

```
replace cannee=1 if (cannee==3 | cannee==4)
(10 real changes made)
```

Nous avons maintenant un classeur qui inclut maintenant la variable binaire construc.



	annee	cannee	construc
1	1960-1975	1	0
2	1960-1975	1	0
3	1960-1975	1	0
4	1960-1975	1	0
5	1960-1975	1	0
6	1960-1975	1	0
7	1960-1975	1	0
8	1960-1975	1	0
9	1960-1975	1	0
10	1976-1985	2	0
11	1976-1985	2	0
12	1976-1985	2	0
13	1976-1985	2	0
14	1976-1985	2	0
15	1976-1985	2	0
16	1986-1995	3	1
17	1986-1995	3	1
18	1986-1995	3	1
19	1996-2005	4	1
20	1996-2005	4	1
21	1996-2005	4	1
22	1996-2005	4	1
23	1996-2005	4	1
24	1996-2005	4	1
25	1996-2005	4	1

2.10.2 Variables logarithmiques

Rappel théorique

À travers les exemples précédents, nous avons vu que plusieurs modèles allouaient une relation non linéaire et logarithmique entre deux variables (dépendante et indépendante). Cette spécification permet entre autres d'exprimer l'effet *ceteris paribus* de x sur y en terme d'élasticité. Selon la ou les variables d'intérêt transformées en logarithme naturel,

l'interprétation se fait différemment. Voici un tableau explicitant les formes fonctionnelles de variables logarithmiques.

Modèle	**Variable dépendante**	**Variable indépendante**	**Interprétation du coefficient**
Niveau-log	y	Log(x)	$\Delta y = (\beta_1/100)\%\Delta x$
Log-niveau	Log(y)	X	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	Log(y)	Log(x)	$\%\Delta y = \beta_1\%\Delta x$

Si y>0, un modèle avec log(y) comme variable dépendante satisfait parfois mieux les hypothèses du *Classical linear model* qu'un modèle avec y seulement comme variable dépendante. En effet, des variables strictement positives présentent quelquefois des distributions conditionnelles hétéroscédastiques ou asymétriques; prendre leur logarithme naturel peut aider à corriger l'hétéroscédasticité, entre autres. Prendre le logarithme permet également d'avoir des estimateurs moins sensibles aux observations extrêmes qui composent notre échantillon. Parmi les variables disposant de valeurs strictement supérieures à zéro, nous pouvons recenser des variables de salaires, de distances, de quantités, etc. À noter finalement qu'il est non approprié de comparer le R^2 de deux modèles dont l'un possède comme variable dépendante y et l'autre log(y), la variable explicative étant fondamentalement différente.

Générer une variable logarithmique sur STATA

Nous pouvons créer une variable logarithmique à partir d'une variable à niveau avec la commande `gen`. Illustrons à l'aide d'un exemple. Nous désirons estimer le modèle suivant :

$$\text{annearn} = \beta_0 + \beta_1 educ + \beta_2 tenure + u$$

où annearn est le revenu annuel en $

educ est le nombre d'années de scolarité
tenure est le nombre d'années de travail avec le même employeur

```
reg annearn educ tenure

      Source |       SS       df       MS              Number of obs =     616
-------------+------------------------------           F(  2,   613) =   23.60
       Model |  3.5714e+09     2  1.7857e+09           Prob > F      =  0.0000
    Residual |  4.6388e+10   613  75673228.6           R-squared     =  0.0715
-------------+------------------------------           Adj R-squared =  0.0685
       Total |  4.9959e+10   615  81234289.6           Root MSE      =    8699

------------------------------------------------------------------------------
     annearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   784.9712   130.4048     6.02   0.000     528.8768    1041.066
      tenure |    198.193   45.83703     4.32   0.000     108.1763    288.2096
       _cons |   1527.793   1765.693     0.87   0.387    -1939.747    4995.334
------------------------------------------------------------------------------
```

Le R^2 est de 0,0715, une valeur très faible. La variabilité de educ et tenure explique peu la variabilité de annearn. Educ et tenure sont statistiquement significatifs au seuil de 1,0 %. En analysant la variable tenure, nous notons qu'une année supplémentaire au sein du même employeur augmente en moyenne le revenu annuel de 198,193 $. Ces deux variables sont également conjointement significatives. Vérifions à l'aide d'un test de Breusch-Paggan s'il existe de l'hétéroscédasticité dans u.

```
estat hottest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of annearn

         chi2(1)      =    13.56
         Prob > chi2  =   0.0002
```

Sur la base de ce test, nous avons suffisamment d'évidence statistique pour rejeter l'hypothèse nulle d'homoscédasticité. Maintenant, effectuons la même régression, mais en utilisant log(annearn) à la place de annearn.

```
gen lannearn=log(annearn)
reg lannearn educ tenure

      Source |       SS       df       MS              Number of obs =     616
-------------+------------------------------           F(  2,   613) =   48.23
       Model |  29.6735558     2  14.8367779           Prob > F      =  0.0000
    Residual |  188.585358   613  .307643325           R-squared     =  0.1360
-------------+------------------------------           Adj R-squared =  0.1331
       Total |  218.258914   615  .354892543           Root MSE      =  .55466
```

```
------------------------------------------------------------------------------
    lannearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0664322   .0083147     7.99   0.000     .0501035    .0827609
      tenure |   .0205612   .0029226     7.04   0.000     .0148217    .0263008
       _cons |   8.302996   .1125817    73.75   0.000     8.081903    8.524088
------------------------------------------------------------------------------
```

Nous voyons que toutes les variables ainsi que la constante sont individuellement significatives. De plus, educ et tenure restent également conjointement significatives. Si nous reprenons l'interprétation du coefficient de tenure, nous notons qu'une année de travail supplémentaire au sein du même employeur provoque en moyenne une hausse du revenu annuel de 0,021 %. Vérifions finalement s'il y a présence d'hétéroscédasticité dans ce modèle.

```
    . hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lannearn

         chi2(1)      =     2.30
         Prob > chi2  =   0.1295
```

Sur la base de ces résultats, nous ne pouvons rejeter l'hypothèse nulle d'homoscédasticité au seuil de 10,0 %. Grâce à cet exemple, nous pouvoir voir que la transformation en logarithme d'une variable à niveau peut corriger en soit le problème d'hétéroscédasticité.

2.10.3 Les variables quadratiques

Rappel théorique

En économétrie, nous utilisons principalement des fonctions quadratiques afin de capturer le sens de l'effet marginal d'une variable explicative. Par exemple, supposons le modèle suivant :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + u$$

L'intuition dans l'interprétation de l'effet de x_1 sur y ne repose pas uniquement sur β_1, mais sur β_2 également. Nous pouvons calculer l'effet de x_1 sur y en dérivant y par rapport à x_1.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + u$$

$$\frac{dy}{dx_1} = \beta_1 + 2\beta_2 x_1$$

$$\Delta y = (\beta_1 + 2\beta_2 x_1)\Delta x_1 \qquad [5]$$

L'effet marginal de x_1 de y dépend de la valeur de x_1 dans l'équation [5]. Généralement, nous mesurons l'effet marginal de x_1 lorsque x_1 prend comme valeur sa moyenne dans l'échantillon. Dans [5], le signe de β_2 définit si l'effet marginal de x_1 est croissant ou décroissant. En effet,

Si $\beta_2 > 0$, l'effet marginal de x_1 est croissant

Si $\beta_2 < 0$, l'effet marginal de x_1 est décroissant

Générer une variable quadratique sur STATA

Encore une fois, nous pouvons générer une variable quadratique sur STATA à l'aide de la commande `gen`. Nous éclaircirons le concept de variable quadratique avec STATA à l'aide d'un exemple. À l'aide des données NBASAL.dta de Wooldridge (2009), nous désirons estimer un modèle expliquant les points par matchs (points) en fonction des années dans la ligue (exper), cette même variable au carré ($exper^2$), l'âge (age) et des années jouées au collège (educ).

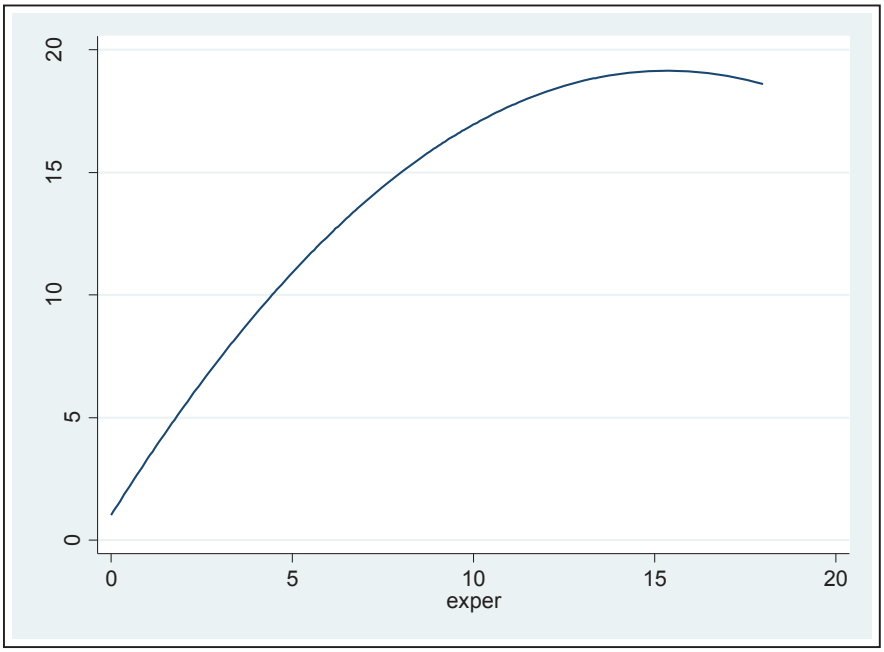
$$points = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 age + \beta_4 educ + u$$

```
. gen exper2=exper^2
. reg points exper exper2 age educ, robust

Linear régression                                      Number of obs =     269
                                                       F(  4,   264) =   13.32
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1412
                                                       Root MSE      =  5.5095

------------------------------------------------------------------------------
             |               Robust
      points |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   2.363631   .3989252     5.92   0.000     1.578151    3.149111
      exper2 |  -.0770269   .0257682    -2.99   0.003    -.1277642   -.0262895
         age |  -1.073958   .3138651    -3.42   0.001    -1.691956   -.4559609
        educ |  -1.286255   .4369337    -2.94   0.004    -2.146573   -.4259363
       _cons |   35.21831    7.43866     4.73   0.000     20.57166    49.86496
------------------------------------------------------------------------------
```

La variation du nombre de points par match suite à un changement dans l'expérience est expliquée à la fois par le coefficient de $exper$ et par celui de $exper2$. Plus précisément, en maintenant age et $coll$ constants, une année d'expérience supplémentaire dans la NBA entraine une augmentation du nombre de points par match de $2.363631 - 0.1540538exper$. La variable $exper$ a donc un effet marginal décroissant sur $points$, ce qui semble à priori normal puisqu'il est assez sensé d'imaginer un joueur qui a cessé d'apprendre et de s'améliorer après 15 ans d'expérience. On peut interpréter le coefficient sur $exper$ comme étant la variation dans $points$ quand $exper$ passe de 0 à 1.

Nous pouvons également créer un graphique des points estimations en fonction de l'expérience en gardant age et educ fixes afin de connaître à partir de combien d'années l'expérience réduit les points par match. Age et educ prendront comme valeur leur moyenne dans l'échantillon.

```
twoway function y=(2.363631*x-0.0770269*x^2)+1.016726512, range (0 18) ytitle(points)
xtitle(exper)
```



[12] $1{,}01675 = \hat{\beta}_o + \hat{\beta}_3 * \overline{age} + \hat{\beta}_{4l} * \overline{educ}$

Selon notre modèle, à partir du moment qu'un joueur a accumulé 15,34289 années d'expérience dans la NBA, chaque année d'expérience supplémentaire a pour effet de réduire le nombre de points par match, *ceteris paribus*. En effet,

$$\max_{exper} \; [2.363631exper - 0.0770269exper^2 + 1.0167265]$$

$$2.363631 - 0.1540538exper = 0$$

$$exper^* = 15{,}34289 \text{ est un maximum}$$

2.10.4 Les variables d'interaction

Rappel théorique

Il arrive parfois que l'effet partiel d'une variable indépendante soit fonction d'un autre régresseur. Par exemple, supposons le modèle suivant :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_4 + u$$

Nous désirons évaluer l'effet partiel de x_1 sur y.

$$\frac{\Delta y}{\Delta x_1} = \beta_1 + \beta_4 x_4$$

$\frac{\Delta y}{\Delta x_1}$ dépend linéairement de x_4. Le signe de β_4 implique si l'effet marginal de x_1 est croissant ou décroissant. Il existe donc un effet d'interaction entre x_1 et x_4. L'effet partiel de x_1 peut être évalué pour toute valeur de x_4. Nous utilisons généralement la moyenne de x_4 dans l'échantillon.

2.10.5 Générer une variable d'interaction sur STATA

Nous pouvons générer des variables d'interaction sur STATA à l'aide de la commande `gen`. Nous reprendrons l'exemple de la NBA de la section 7.3.2. Nous aimerions tester s'il existe une différence dans le salaire (wage) estimé entre un joueur de pointe (guard) étoile (allstar) et un joueur de centre (center) qui n'est pas un joueur étoile, en gardant les points par match (points), l'âge (age) et les années au collège (educ) fixes. Nous disposions de quatre variables binaires sur la position et le statut d'étoile des joueurs dans notre base de données.

- guard=1 si le joueur est joueur de pointe
- forward=1 si le joueur est un attaquant
- center=1 si le joueur est un centre
- allstar=1 si le joueur est une étoile

Le premier objectif ici est de générer des variables d'interaction qui définissent les catégories de joueurs suivants :

- guardstar: un joueur de pointe étoile
- guardnonstar : un joueur de pointe non étoile
- centerstar : un joueur de centre étoile
- forwardstar : un joueur d'attaque étoile
- forwardnonstar : un joueur d'attaque non-étoile

Nous assumerons comme groupe de base centernonstar, c'est-à-dire un joueur de centre non-étoile.

Voici la procédure à suivre pour générer ces variables sur STATA.

```
gen nonstar=(!allstar)
gen guardstar=guard*allstar
gen guardnonstar=guard*nonstar
gen centerstar=center*allstar
gen forwardstar=forward*allstar
gen forwardnonstar=forward*nonstar
```

Ensuite, vérifions s'il existe une différence dans le salaire estimé entre un joueur de pointe étoile et un joueur de centre non-étoile. Effectuons la régression suivante sur STATA.

```
reg wage points exper age educ guardstar guardnonstar centerstar forwardstar forwardnonstar,
robust


Linear régression                                      Number of obs =     269
                                                       F(  9,   259) =   35.42
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.5610
                                                       Root MSE      =  673.87

------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      points |   102.3945    10.3829     9.86   0.000     81.94888    122.8402
       exper |   86.30806   42.08734     2.05   0.041     3.431108     169.185
         age |  -2.333977   42.08881    -0.06   0.956     -85.2138    80.54585
        educ |  -64.50734   77.02455    -0.84   0.403    -216.1814    87.16675
   guardstar |  -569.0822   284.9177    -2.00   0.047    -1130.132   -8.032045
guardnonstar |  -473.2613   148.0945    -3.20   0.002    -764.8838   -181.6387
  centerstar |   480.0377   475.0664     1.01   0.313    -455.4467    1415.522
 forwardstar |  -256.8007   304.9973    -0.84   0.401    -857.3908    343.7894
forwardnon~r |  -224.8562   153.1877    -1.47   0.143    -526.5081    76.79577
       _cons |   528.5143   1095.214     0.48   0.630    -1628.143    2685.172
------------------------------------------------------------------------------
```

Le coefficient d'intérêt ici est celui de guardstar. *Ceteris paribus*, nous notons qu'un joueur de pointe étoile gagne en moyenne 569,08 milliers de $ de moins qu'un joueur de centre non étoile. La p-valeur de ce coefficient nous indique également que cette différence est statistiquement significative à un seuil de 5,0 %.

2.11 Forme fonctionnelle d'un modèle

2.11.1 Spécificité du modèle

Une spécificité du modèle erroné d'un modèle invalide l'hypothèse numéro 4 d'exogénéité stricte (E(u|X)=0) provoquant ainsi un biais d'endogénéité. En effet, dans le cas où un x_j est endogène, nous aurons $\mathrm{cov}(u, x_j) \neq 0$. Une erreur de spécification survient dans les cas suivants :

- Biais d'omission de variable corrélée avec la variable dépendante
- Biais d'erreur de mesure

- Forme fonctionnelle du modèle et/ou des variables incorrecte
    - Nous estimons $wage = \beta_0 + \beta_1 educ + u$ lorsque le vrai modèle est $\log(wage) = \beta_0 + \beta_1 educ + u$ ou nous omettons d'inclure une variable d'interaction lorsqu'elle est présente dans le vrai modèle.

Une façon pratique de vérifier si un modèle est correctement spécifié est de tester s'il existe des combinaisons linéaires statistiquement significatives de variables explicatives. Cette méthode se nomme « RESET » pour « *régression specification error test* ». Le concept derrière le test RESET est de tester si des combinaisons non linéaires des valeurs ajustées aident à valider ou invalider les variables indépendantes. Si ces combinaisons ont un pouvoir explicatif ou sont statistiquement significatives dans le nouveau modèle, le modèle original est mal spécifié. Voici les étapes à suivre pour réaliser un test RESET :

1. Estimer le modèle suivant $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$ et obtenir les valeurs prédites $\hat{y}$.
2. Estimer le modèle $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \alpha_0 \hat{y}^2 + \alpha_1 \hat{y}^2 + v$ en supposant E(v|X)=0.
3. Tester $H_0: \alpha_0 = \alpha_1 = 0$ avec un test de Fisher. Si nous pouvons rejeter l'hypothèse, il y a de fortes évidences de croire que le modèle soit mal spécifié.

Finalement, nous pouvons juger de la spécificité d'un modèle en effectuant un test de Fisher pour toutes les variables explicatives. Si nous pouvons rejeter l'hypothèse nulle à un niveau de significativité d'au plus 10,0 %, le modèle possède une forme fonctionnelle générale statistiquement adéquate.

2.11.2 Tester la forme fonctionnelle d'un modèle sur STATA

Il est possible de répéter les étapes du test RESET sur STATA. Illustrons à l'aide d'un exemple. Nous désirons estimer le modèle suivant et vérifier s'il est correctement spécifié à l'aide des données NBASAL.dta de Wooldridge (2009) :

$$points = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 age + \beta_4 age^2 + \beta_5 educ + u$$

Procédons sur STATA.

1. Générons age^2 et estimons $points = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 age + \beta_4 age^2 + \beta_5 educ + u$

```
gen age2=age^2
reg points exper2 exper age2 age educ

      Source |       SS       df       MS              Number of obs =     269
-------------+------------------------------           F(  5,   263) =    8.92
       Model |  1353.54692     5  270.709385           Prob > F      =  0.0000
    Residual |  7977.64396   263   30.333247           R-squared     =  0.1451
-------------+------------------------------           Adj R-squared =  0.1288
       Total |  9331.19088   268  34.8178764           Root MSE      =  5.5076

------------------------------------------------------------------------------
      points |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      exper2 |  -.1280723   .0524378    -2.44   0.015    -.2313237   -.0248209
       exper |   2.863828   .6127241     4.67   0.000     1.657359    4.070297
        age2 |   .0535514   .0491917     1.09   0.277    -.0433083    .1504112
         age |  -3.983695   2.689078    -1.48   0.140    -9.278557    1.311168
        educ |  -1.312604   .4510841    -2.91   0.004      -2.2008    -.424408
       _cons |   73.59034   35.93341     2.05   0.042     2.836553    144.344
```

2. Estimons $\widehat{points}$, $\widehat{points}^2$ et $\widehat{points}^3$, les valeurs prédites du modèle en (1).

```
predict yhat
(option xb assumed; fitted values)
gen yhat2=yhat^2
gen yhat3=yhat2*yhat
```

3. Estimons $points = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 age + \beta_4 age^2 + \beta_5 educ + \alpha_0 \widehat{points}^2 + \alpha_1 \widehat{points}^3$ et testons $H_0: \alpha_0 = \alpha_1 = 0$.

```
quietly reg points exper2 exper age2 age educ yhat2 yhat3
test yhat2 yhat3
( 1)  yhat2 = 0
( 2)  yhat3 = 0
F(  2,   261) =    0.42
Prob > F =    0.6547
```

Sur la base de ces résultats, avec une p-valeur de 0,6547, nous ne pouvons pas rejeter l'hypothèse nulle. Il n'existe donc pas d'erreur de spécification.

Heureusement, il existe une commande postestimation nommée estat ovtest qui s'avère en soit un test RESET. Cette commande test s'il existe une erreur de spécification causée par une variable importante omise dans le modèle régressé.

Syntaxe :

```
estat ovtest
```

Menu :

Statistics > Postestimation > Reports and statistics

Effectuons ce test suite au modèle estimé ci-dessus.

```
quietly reg points exper2 exper age2 age educ
estat ovtest
Ramsey RESET test using powers of the fitted values of points
       Ho:  model has no omitted variables
                 F(3, 260) =      0.54
                  Prob > F =      0.6571
```

L'hypothèse nulle de ce test stipule qu'il n'y a pas de variables importantes omises dans le modèle. Sur la base de la p-valeur (0,6571), nous ne pouvons rejeter cette hypothèse nulle. Notre modèle est donc correctement spécifié.

2.12 Normalité dans le terme d'erreur

Nous avons vu ultérieurement que les statistiques de Student et de Fisher ne sont valides pour tester les paramètres d'une régression si et seulement si l'hypothèse de normalité dans le terme d'erreur est admise. Grâce à Stata, nous pouvons juger de cette affirmation en comparant sur un graphique la distribution des résidus à celle d'une loi normale grâce à une méthode non paramétrique d'estimation de la densité de probabilité, KDE pour « Kernel density estimation ». La commande en question est **kdensity**[13].

Syntaxe :

```
kdensity varname [if] [in] [weight] [, options]
```

Menu:

Statistics > Nonparametric analysis > Kernel density estimation

[13] À noter que seules les options intéressantes pour cette sous-section sont incluses. Le lecteur est invité à lire davantage sur la liste des options à la commande kdensity en tapant la commande help kdensity sur STATA 12.

Options:

`normal`: ajoute une densité normale au graphique

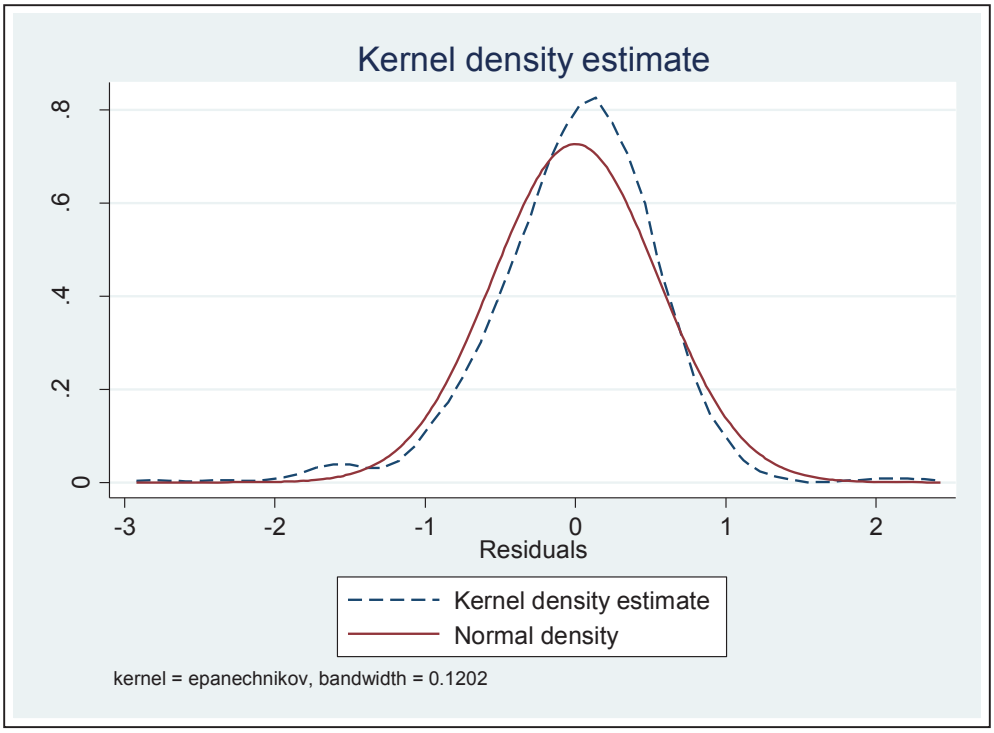
`student(#)` : ajoute une densité de Student selon le degré de liberté (#) au graphique

Voici un exemple concret qui nous permettra de mieux comprendre l'étendue de cette commande. À l'aide des données FRINGE.dta de Wooldridge (2009), nous cherchons à quantifier si les années de scolarité (educ) ont une influence sur le revenu annuel (log(annearn)) tout en contrôlant pour le mariage (married).

```
. reg lannearn educ married

      Source |       SS       df       MS              Number of obs =     616
-------------+------------------------------           F(  2,   613) =   53.73
       Model |  32.5545763     2  16.2772882           Prob > F      =  0.0000
    Residual |  185.704338   613  .302943455           R-squared     =  0.1492
-------------+------------------------------           Adj R-squared =  0.1464
       Total |  218.258914   615  .354892543           Root MSE      =   .5504

------------------------------------------------------------------------------
    lannearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0582265   .0081254     7.17   0.000     .0422695    .0741835
     married |   .3698522   .0478384     7.73   0.000     .2759053    .4637992
       _cons |   8.311096   .1101793    75.43   0.000     8.094722    8.527471
------------------------------------------------------------------------------
```

Nous voyons ici que l'effet d'une année supplémentaire de scolarité provoque une augmentation de 5,82 % du revenu annuel, en moyenne. Le R^2 de la régression est de 0,1492. Tous les paramètres sont également statistiquement significatifs. Après vérification par les tests RESET et de Breusch-Paggan, le modèle est correctement spécifié et ne présente pas de signes d'hétéroscédasticité. À l'aide de kdensity, vérifions maintenant la distribution des résidus pour connaître la validité de l'hypothèse 6.

```
predict uhat, resid
kdensity uhat, normal
```



Avec ce graphique, nous pouvons voir que la densité de Kernel ressemble quelque peu à une distribution normale avec une légère asymétrie vers la droite.

Il existe également une commande complémentaire à la densité de Kernel, la commande `pnorm`. La commande pnorm présente graphiquement les valeurs standardisées d'une distribution de probabilité suivant une loi normale. Ces valeurs sont par la suite réparties en fonction d'une droite représentant les vraies valeurs d'une distribution normale théorique. Les déviations par rapport à cette droite signifient qu'il y a peu de normalité dans la variable utilisée lors de la standardisation. Notons que pnorm est sensible aux observations présentant une non-normalité dans les quartiles médians.

Syntaxe :

```
pnorm varname [if] [in] [, options2]14
```

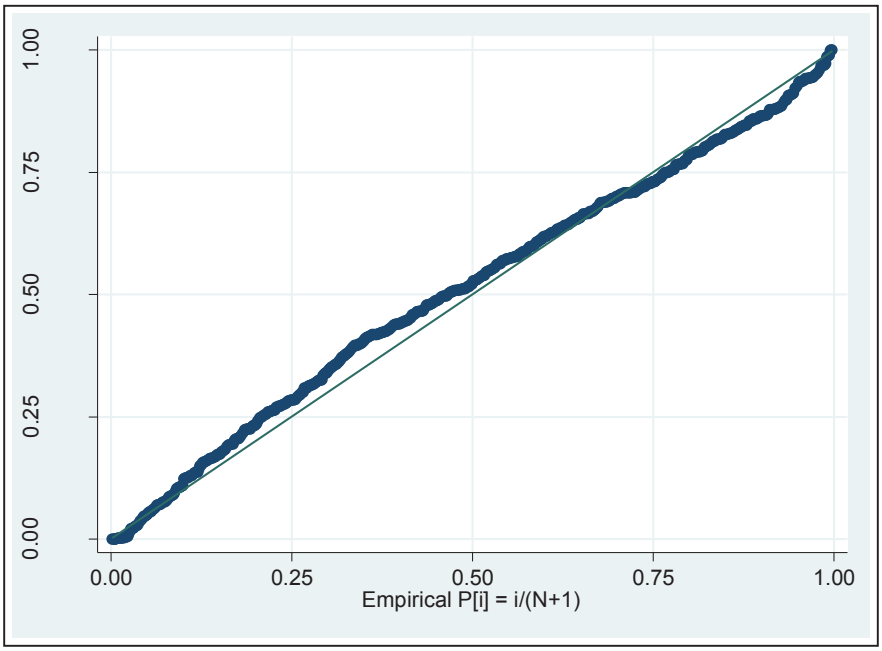
Menu :

Statistics > Summaries, tables, and tests > Distributional plots and tests > Normal probability plot, standardized

[14] Pour les options supplémentaires, taper help pnorm dans STATA 12.

Illustrons avec notre exemple.

```
pnorm uhat
```



Nous voyons clairement qu'il existe une déviation dans le deuxième et le dernier quartile entre les valeurs standardisées et la droite de normalité théorique.

Finalement, la commande `swilk` permet de tester si une variable aléatoire d'un échantillon provenant d'une population normalement distribué suit ou non une loi normale. L'hypothèse nulle de ce test assume que la variable aléatoire suit une distribution normale. Nous pouvons d'ailleurs utiliser cette commande pour un échantillon contenant 4 à 2000 observations. Ce test se base sur les travaux de Shapiro et Wilk (1965).

Syntaxe:

```
swilk varlist [if] [in] [, swilk_options]15
```

Menu:

Statistics > Summaries, tables, and tests > Distributional plots and tests > Shapiro-Wilk normality test

Testons avec les résidus de notre modèle.

[15] Pour les options supplémentaires, taper help pnorm dans STATA 12.

```
swilk uhat

                   Shapiro-Wilk W test for normal data

    Variable |    Obs       W           V         z       Prob>z
-------------+-------------------------------------------------
        uhat |    616    0.96463     14.365     6.465    0.00000
```

Selon la p-valeur (Prob>z), nous avons suffisamment d'évidence statistique pour rejeter l'hypothèse nulle de normalité dans les résidus de la régression que nous avons effectuée. Il faut donc rester vigilant avec l'interprétation de l'hypothèse 6 dans notre modèle.

2.13 Tester la multicolinéarité des variables explicatives

2.13.1 Rappel théorique

La multicolinéarité survient lorsque les variables indépendantes d'un modèle sont fortement corrélées entre elles. La multicolinéarité peut avoir plusieurs conséquences telles que :

- Erreurs standards des estimateurs élevées.
- Statistique t très faible.
- Changement non espéré de la magnitude ou du signe des coefficients.
- Coefficients non significatifs malgré un R^2 élevé dans la régression.

Ces conséquences proviennent essentiellement d'une sensibilité plus importante de la variance des estimateurs par rapport au degré corrélation qui existe entre les variables explicatives. En effet, si nous reprenons la formule de la variance d'un estimateur $\hat{\beta}_j$, $var(\hat{\beta}_j) = \frac{\sigma^2}{SCT_j(1-R_j^2)}$, ce degré de corrélation provient du facteur $(1 - R_j^2)$. Tel que défini ultérieurement, R_j^2 représente le R^2 de la régression de x_j sur les autres régresseurs. Plus R_j^2 est élevé, plus $var(\hat{\beta}_j)$ sera sensible aux autres régresseurs.

Ce faisant, dans un monde idéal, il est toujours préférable d'inclure des variables indépendantes corrélées avec la variable dépendante, mais non corrélées avec les autres variables indépendantes

du modèle. De cette façon, nous n'aggravons pas le problème de multicolinéarité et la sur spécification du modèle. Dans tous les cas, nous avons un arbitrage biais-variance devant nous.

2.13.2 Tester la multicolinéarité des variables explicatives – Application sur STATA

Nous pouvons utiliser la commande postestimation **estat vif** afin de vérifier le niveau de multicolinéarité entre les variables. Le test estat vif calcule les $\left(1 - R_j^2\right)$ pour chaque variable explicative du modèle préalablement estimé, nous donnant ainsi un aperçu du niveau de corrélation entre les variables indépendantes. Notons que cette commande ne règle pas le problème de multicolinéarité.

Syntaxe :

```
estat vif
```

Menu :

Statistics > Postestimation > Reports and statistics

Présentons un exemple afin de mieux comprendre cette commande. Nous aimerions savoir s'il existe un effet de revenu positif (log(income)) sur la consommation hebdomadaire de cigarettes (cigs). Nous disposons de données sur le prix d'un paquet de cigarettes et de plusieurs variables de contrôle telles l'âge (age), l'éducation (educ), la couleur de peau d'individu (white), s'il y a des restrictions de fumer dans les restaurants où l'individu vit (restaurn). Avant de trouver le modèle optimal, il est important de vérifier le niveau de corrélation entre les variables afin de détecter les risques de multicolinéarité entre celles-ci.

```
. cor cigs income cigpric educ age white restaurn
(obs=807)
             |     cigs   income  cigpric     educ      age    white restaurn
-------------+------------------------------------------------------------
        cigs |   1.0000
      income |   0.0532   1.0000
     cigpric |  -0.0114   0.0480   1.0000
        educ |  -0.0487   0.3344   0.0310   1.0000
         age |  -0.0415  -0.0640   0.0293  -0.1806   1.0000
       white |   0.0075  -0.0146  -0.0616  -0.0029  -0.0285   1.0000
    restaurn |  -0.0871   0.1060   0.1392   0.0605  -0.0389  -0.1042   1.0000
```

En analysant ce tableau de corrélation, nous voyons que la variable dépendante de notre modèle, cigs, est peu corrélée avec chacune des variables explicatives disponibles dans notre base de données. Il y a donc de fortes chances que le R^2 du modèle soit faible. De plus, excepté pour la corrélation entre income et educ (0,3344), toutes les variables indépendantes restent peu corrélées entre elles ce qui sous-entend un risque de multicolinéarité faible. Effectuons la première régression de cigs sur income, cigpric, educ, age, white et restaurn.

```
. reg cigs income cigpric educ age white restaurn

      Source |       SS       df       MS              Number of obs =     807
-------------+------------------------------           F(  6,   800) =    2.59
       Model |  2889.70751     6  481.617918           Prob > F      =  0.0172
    Residual |  148863.975   800  186.079969           R-squared     =  0.0190
-------------+------------------------------           Adj R-squared =  0.0117
       Total |  151753.683   806  188.280003           Root MSE      =  13.641

------------------------------------------------------------------------------
        cigs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0001306    .000056     2.33   0.020     .0000206    .0002405
     cigpric |    .004311   .1026571     0.04   0.967    -.1971981      .20582
        educ |  -.3680446   .1692771    -2.17   0.030    -.7003243   -.0357648
         age |  -.0439658   .0287385    -1.53   0.126    -.1003775    .0124459
       white |  -.1123855   1.480575    -0.08   0.940    -3.018657    2.793886
    restaurn |  -2.990576   1.137075    -2.63   0.009    -5.222579   -.7585731
       _cons |   13.14521    6.76139     1.94   0.052    -.1269484    26.41737
------------------------------------------------------------------------------
```

Comme attendu, le R^2 de la régression est très faible, voire nul. Toutes les variables sont conjointement significatives entre elles. De plus, excepté pour white dont l'effet ne repose essentiellement sur aucun fondement logique, tous les coefficients possèdent le signe espéré. La constante, income, educ et restaurn sont statistiquement significatifs. À la revue du tableau de corrélation ci-haut, nous voyons que white est peu corrélée avec cigs (elle est d'ailleurs fortement non significative dans le modèle). Afin d'identifier qu'elles sont les variables qui pourraient être à l'origine de la multicolinéarité dans notre modèle, nous pouvons également utiliser la commande estat vce, cor pour obtenir la matrice des corrélations de nos estimateurs.

```
. vce, cor

Correlation matrix of coefficients of regress model

        e(V) |   income   cigpric      educ       age     white  restaurn     _cons
-------------+------------------------------------------------------------------------
      income |   1.0000
     cigpric |  -0.0277    1.0000
        educ |  -0.3248   -0.0189    1.0000
```

```
         age |   0.0025   -0.0383    0.1690    1.0000
       white |   0.0037    0.0466   -0.0002    0.0308    1.0000
    restaurn |  -0.0856   -0.1305   -0.0187    0.0362    0.0971    1.0000
       _cons |  -0.0308   -0.9020   -0.2718    -0.2008   -0.2450   0.0725     1.0000
```

Effectivement, toutes les corrélations incluant la variable white sont très faibles. Par souci de parcimonie, effectuons à nouveau la régression précédente, mais en excluant la variable white.

```
. reg cigs income cigpric educ age restaurn

      Source |       SS       df       MS              Number of obs =     807
-------------+------------------------------           F(  5,   801) =    3.11
       Model |  2888.63535     5   577.72707           Prob > F      =  0.0087
    Residual |  148865.047   801  185.848998           R-squared     =  0.0190
-------------+------------------------------           Adj R-squared =  0.0129
       Total |  151753.683   806  188.280003           Root MSE      =  13.633

------------------------------------------------------------------------------
        cigs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0001306    .000056     2.33   0.020     .0000207    .0002405
     cigpric |   .0046741   .1024819     0.05   0.964    -.1964908    .2058389
        educ |  -.3680468    .169172    -2.18   0.030    -.7001196   -.0359739
         age |  -.0438986    .028707    -1.53   0.127    -.1002484    .0124512
    restaurn |  -2.982192   1.130995    -2.64   0.009    -5.202256   -.7621274
       _cons |   13.01946   6.551227     1.99   0.047     .1598616    25.87906
------------------------------------------------------------------------------
```

Ici, le R^2 de la régression reste inchangé. Aucun changement majeur n'est à noter; tous les résultats restent presque similaires à ceux de la première régression. Nous voyons qu'il existe un effet revenu faible sur la consommation hebdomadaire de cigarettes : en moyenne, un individu consomme une cigarette de plus par semaine lorsque son revenu annuel augmente de 1000 $. Cet effet est également statistiquement significatif. Notons que des tests supplémentaires sur la forme fonctionnelle et l'hétéroscédasticité ont également été effectués, ceux-ci étant non significatifs. Maintenant, voyons si la multicolinéarité a un impact important sur nos résultats.

```
. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
        educ |      1.16    0.862048
      income |      1.14    0.880131
         age |      1.04    0.965071
    restaurn |      1.03    0.969071
     cigpric |      1.02    0.977810
-------------+----------------------
    Mean VIF |      1.08
```

La colonne 1/VIF affiche les $\left(1 - R_j^2\right)$ de chaque variable explicative. Intuitivement, plus 1/VIF est élevé, plus le risque de multicolinéarité est faible (donc moins la variance des estimateurs sera élevée). Par exemple, pour la variable income, 0.12 % (1-0,862) de la variation de educ, restaurn, age, et cigpric est expliquée dans la variation de income, ce qui s'avère relativement faible. En moyenne, le $\left(1 - R_j^2\right)$ qui combine toutes les variables indique un très faible niveau de multicolinéarité (1/1,08=0,925).

3 Variables dépendantes discrètes

3.1 Introduction

Ce chapitre traite de l'estimation des paramètres des modèles ayant comme variable dépendante une variable discrète. Nous traiterons tout d'abord la situation où la variable dépendante ne prend que deux valeurs et par la suite des modèles où la variable dépendante peut prendre plusieurs valeurs discrètes. Notez que pour ce chapitre les modules Stata SPost, omodel seront utilisés.

3.2 Modèles logit et probit binaires

3.2.1 Rappel théorique

Modèle

Dans certains cas, le chercheur peut vouloir modéliser une situation ou une observation appartient ou non à un ensemble, une catégorie ou un groupe.

Pour simplifier notre rappel, nous prendrons comme exemple le fait qu'une personne fasse partie ou non de la portion de la population active en emploi. Notre variable dépendante peut donc être spécifiée comme suit :

$$y = \begin{cases} 1 & si\ la\ personne\ a\ un\ emploi \\ 0 & sinon \end{cases}$$

Nous nous intéressons donc à la probabilité que cette variable prenne la valeur 1 ($P(y = 1|x)$). Le chercheur peut donc vouloir connaître l'impact des variables exogènes sur la probabilité d'appartenir à une catégorie ou prédire si la variable y pour un nouvel individu prendra la valeur 1 (la personne à un emploi) ou 0 (la personne est sans emploi).

Dans le cas de variables continues définies sur un intervalle non borné, nous pouvons obtenir nos $\hat{y}_i$ (les valeurs prédites) et donner leur interprétation, et ce, pour chaque individu i. Dans le cas d'une variable qui doit refléter une probabilité, les valeurs de $\hat{y}_i$ doivent être strictement entre 0 et 1. À cet effet, nous allons modifier notre modèle linéaire afin de pouvoir prendre en compte

cette restriction. En effet, si le modèle estimé est linéaire, alors il y aura présence d'hétéroscédasticité et les valeurs prédites ne seront pas strictement entre 0 et 1.

Nous pouvons donc réécrire le modèle comme étant

$$P(y = 1|x) = G(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k) = G(\beta_0 + \boldsymbol{x\beta}),$$

Ou $G(.)$ est une fonction définie sur $\mathbb{R}$ et où nous voulons que

$$\lim_{\boldsymbol{x\beta} \to +\propto} Prob(Y = 1|x) = 1$$
$$\lim_{\boldsymbol{x\beta} \to -\propto} Prob(Y = 1|x) = 0$$

Le chercheur peut donc choisir toute fonction qui satisfait les deux dernières propriétés. De cet exercice, les deux fonctions les plus utilisées sont la fonction de densité cumulative normale

$$G(\beta_0 + \boldsymbol{x\beta}) = \Phi(\beta_0 + \boldsymbol{x\beta}) \equiv \int_{-\infty}^{\beta_0 + \boldsymbol{x\beta}} \left(\frac{1}{\sqrt{2\pi}}\right) e^{(-1/2)z^2} dz$$

et la fonction de densité cumulative logistique

$$G(\beta_0 + \boldsymbol{x\beta}) \ = \Lambda(\beta_0 + \boldsymbol{x\beta}) = \frac{exp^{\beta_0 + \boldsymbol{x\beta}}}{1 + exp^{\beta_0 + \boldsymbol{x\beta}}}$$

Nous laisserons le soin au lecteur de s'informer concernant les méthodes d'estimation.

Après avoir transformé la fonction, il est possible de pouvoir interpréter les effets marginaux et les rapports de cotes dans le cas d'un modèle logistique (« odds ratio ») comme les paramètres $(\widehat{\boldsymbol{\beta}})$ ne sont pas directement interprétables.

Rapports de cotes

Notons qu'une fois les estimateurs calculés, il est assez aisé de dériver les rapports de cotes aussi appelés ratios de chance et « odds ratios » (OR) associés à chaque coefficient pour le modèle logit:

$$OR = e^{\beta_i}$$

Pour chaque variable indépendante, ces ratios indiquent comment est affectée la probabilité de réalisation de l'évènement (la personne a un emploi) par rapport à la réalisation de l'évènement de base (la personne n'a pas d'emploi). Un OR supérieur à un signifie que ce risque relatif

augmente avec la variable explicative. Autrement dit, la réalisation de l'évènement donné devient plus probable par rapport à l'évènement de base. L'inverse est vrai pour un OR inférieur à un.

Effets marginaux

Les effets marginaux peuvent apparaître de plusieurs façons selon la nature de la variable explicative d'intérêt. En effet, pour connaître l'impact à la marge d'une variable explicative continue, il faut calculer

$$\frac{\partial E(y|\boldsymbol{x})}{\partial x} = G(\boldsymbol{x\beta})[1 - G(\boldsymbol{x\beta})]\boldsymbol{\beta}$$

Dans le cas d'une variable discrète binaire, x_1, nous devons simplement calculer

$$G(\beta_0 + \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k) - G(\beta_0 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

Dans le cas d'une variable discrète c_k prenant plus de deux modalités, nous calculerons la différence de probabilité estimée entre toutes les modalités que prend la variable étant donné que nous ne nous attendons pas à un effet linéaire dans toutes les situations. Ainsi, entre une modalité c_k et $c_k + 1$

$$G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k (c_k + 1)) - G(\beta_0 + \beta_2 x_2 + \cdots + \beta_k c_k)$$

Ces effets peuvent être calculés en moyenne, c'est-à-dire en prenant la moyenne de chaque effet marginal. Nous pouvons aussi les calculer à la moyenne ou à la médiane, c'est-à-dire en insérant la valeur moyenne ou médiane de chaque variable dans le calcul. Étant donné la nature non linéaire du modèle, il peut aussi être intéressant de calculer l'effet marginal à plusieurs valeurs de la variable afin de pouvoir le représenter graphiquement et ainsi l'analyser d'une manière plus étendue.

Mesures de qualité du modèle

Plusieurs mesures de qualité du modèle sont utilisées dans la littérature afin de répondre à la situation ou nous voulons prédire dans quelle catégorie sera un nouvel individu et de comparer les modèles entre eux. Nous énumérerons les principaux.

Pseudo-R^2 de McFadden :

$$R^2{}_{MF} = 1 - \frac{LV_M}{LV_0}$$

Où LV_M est la fonction de log-vraisemblance du modèle estimé et LV_0 est la fonction de log-vraisemblance du modèle n'ayant aucune variable explicative. Le pseudo-R^2 de McFadden varie entre 0 (le modèle n'est pas meilleur que le modèle trivial) et 1 (le modèle est « parfait ») (Rakotomalala, 2009).

Le critère d'information d'Akaike (AIC):

$$AIC = 2K - 2LV_M$$

Cette mesure est une évaluation de l'arbitrage biais-variance pour un modèle. Il ne permet pas d'effectuer un test au sens strict (car aucune hypothèse nulle n'est testée), mais demeure un outil comparatif intéressant pour la sélection d'un modèle. Pour un ensemble donné de candidats, le modèle préféré sera celui dont la valeur AIC est minimale. L'AIC possède la propriété intéressante de pénaliser pour la surspécification du modèle : à log-vraisemblance égale, le meilleur modèle est celui avec le moins de paramètres. Burnham et Anderson (2002) postulent arbitrairement qu'une différence de 10 unités entre les valeurs AIC de deux modèles permet de significativement discriminer le modèle qui se classe moins bien.

La courbe ROC

La courbe ROC met en relation le taux des valeurs « vraies » bien prédites (taux de vrais positifs) avec le taux de valeurs « vraies » mal prédites (taux de faux positif). Le premier taux, aussi appelé « sensibilité », est le ratio des valeurs « 1 » de la variable dépendante correctement prédites par le modèle sur le nombre total de valeurs « 1 » observées dans l'échantillon. Le second taux est le ratio des valeurs « 1 » de la variable dépendante incorrectement prédites, c'est-à-dire quand le modèle prédit la valeur 1 alors que la valeur observée est 0, sur le nombre total des valeurs « 0 » observées dans l'échantillon. Nous classons un individu comme ayant une valeur « 1 », par exemple lorsque la probabilité prédite de cette réalisation est supérieure au seuil s=0,5. La courbe ROC est une généralisation de la précédente méthode de classement qui fait varier s sur tout le continuum des valeurs possibles entre 0 et 1. Pour chaque valeur de s, le taux de vrais positifs (TVP) et le taux de faux positifs (TFP) sont calculés et représentés sur un plan cartésien avec le TFP en abscisse et le TVP en ordonnées. Cette façon de représenter

l'arbitrage entre le TVP et le TFP en fonction du seuil s permet d'évaluer globalement le pouvoir de discrimination du modèle logit. Si le modèle classe de façon complètement aléatoire (probabilité prédite systématique de 0,5 pour chaque réalisation) la courbe ROC aura l'allure d'une droite de 45 degrés entre 0 et 1. À l'inverse, si la classification du modèle est parfaite, la courbe ROC prendra l'allure d'un seul point en (0,1). Il est donc intéressant de considérer l'aire en dessous de la courbe ROC (AUC) pour obtenir un score du modèle en termes de classification. Nous pouvons définir la qualité de la classification comme suit :

AUC = 0.5	Aucune discrimination, classification aléatoire.
0.7 ≤ AUC < 0.8	Classification acceptable
0.8 ≤ AUC < 0.9	Classification excellente
AUC > 0.9	Classification exceptionnelle

(Rakotomalala, 2009)

3.2.2 La commande logit

Description

La commande logit dans Stata sert à estimer un modèle logistique et utilise la méthode du maximum de vraisemblance pour trouver les paramètres ($\hat{\beta}$) du modèle. Stata traitera les valeurs de la variable dépendante telles que

$$y = \begin{cases} 1 & y > 0 \\ 0 & y = 0 \end{cases}$$

Les variables négatives seront donc considérées comme manquantes. La variable dépendante doit donc être codée soit 0, soit 1.

Syntaxe

```
Logit [depvar] [indepvars [if] [in] [weight] [, options]
```

où `depvar` est la variable dépendante
`[indepvars]` est la variable indépendante
`[if]` et `[in]` sont des contraintes dans l'échantillon
`[weight]` assigne des poids aux observations
`[, options]` présente des options supplémentaires dans la régression

Menu

Statistics > Binary outcomes > Probit régression

Option

Erreur type/Robust

`vce(typevce)` : le type vce peut être `robust`, `cluster`(*variablecluster*). Pour plus d'informations, le lecteur peut se référer à la documentation Stata.

Résultats

`or` : les rapports de cotes seront affichés à la place des paramètres estimations.

3.2.3 Les commandes postestimation après la commande logit

`test [varexo]`: nous pouvons tester des restrictions linéaires comme dans un modèle de régression linéaire (par exemple : $\beta_1 = \beta_2$).

`estat classification` : nous donne accès à des statistiques de classement basées sur les $\hat{y}_i$.

`estat gof` : cette commande performe le test de Hosmer-Lemeshow sur la qualité du modèle. Ce test n'est pas très puissant, car il nous informe uniquement sur le fait que le modèle est acceptable.

`lroc` : cette commande calcule l'aire sous la courbe ROC et fournit un graphique celle-ci.

`lsens` : trace un graphique mettant en relation la sensitivité et la spécificité par rapport au point de coupure choisi.

`estat ic` : nous donne les valeurs des critères d'information (AIC et BIC).

`lincom [opérationlinéaire], or` : cette option nous permet de calculer les ratios relatifs de risque pour un type particulier d'individu en spécifiant les valeurs auxquelles ceux-ci veulent être évalués. Pour plus de renseignements sur cette commande, se référer à la page 911 du manuel du logiciel Stata.

`linktest` : cette commande teste globalement la bonne spécification du modèle. Une régression est effectuée sur les valeurs prédites $\hat{y}_i$ et les valeurs prédites au carré $\hat{y}_i{}^2$. Le modèle est bien

spécifié si le carré des valeurs prédites ne contient pas d'information pour expliquer y. Autrement dit, le coefficient devant la variable _hatsq doit être non significatif.

`lrtest [modèlecomplet] [modèlecontraint][, stats]` : cette commande performe un test de ratio de vraisemblance (likelihood ratio test) entre un modèle complet et un modèle contraint. Ce test nous renseigne sur la significativitée des coefficients non inclus dans le modèle contraint et est similaire au test-F joint qui est calculé avec la commande test. Le modèle contraint est donc un sous modèle du modèle complet. Ce test peut être utilisé pour un test de Chow.

`margins, [dydx(varexos)]`
`[at(varexo1=(valdébut(incrément)valfin)varexo2=(valdébut(incrément)valfin)]` :

cette commande nous permet de calculer les effets marginaux des variables spécifiées et les prédictions du modèle. L'option `dydx(.)` est nécessaire pour calculer les effets marginaux de variables continues, l'option `at` nous servira à pouvoir utiliser la commande `marginsplot` afin d'avoir un aperçu graphique de ces effets.

`marginsplot [varexos]` : cette commande doit être utilisée après `margins` pour avoir un aperçu graphique des effets marginaux sur les probabilités que la variable prenne la valeur 1.

`prchange [varlist], [help]` : calcule l'impact d'une variation discrète ou marginale des variables spécifiées pour les modèles catégoriels et de comptage. Cette commande ne fonctionne pas avec les variables considérées comme des facteurs (*factor variables*).

`contrast [varlist]` : permets de tester si la moyenne des effets est égale entre toutes les modalités d'une variable ordinale incluse dans la régression (x).

`contrast r.[varlist]` : permets de tester si les moyennes des effets sont égales entre les différentes modalités d'une variable ordinale quand elles sont comparées à la catégorie de référence.

`contrast ar.[varlist]` : permets de tester si l'estimateur d'une modalité est différent de celui qui le précède.

Syntaxe pour predict

`predict p if e(sample)` : prédit la probabilité que la variable dépendante prenne la valeur 1. La restriction if e(sample) est ajoutée afin de spécifier que les probabilités soient prédites pour les observations n'ayant aucune valeur manquante.

3.2.4 Exemple

Pour cette application, nous utiliserons le fichier BWGHT.dta disponible sur le site de Stata. Ce fichier est composé de données concernant le poids des bébés à la naissance et concernant la consommation de cigarettes.

```
use "http://www.Stata.com/data/jwooldridge/eacsap/bwght.dta", clear
```

Nous voulons donc estimer les paramètres d'un modèle ou la variable dépendante est le fait de fumer ou non pour un individu selon les caractéristiques auxquelles nous avons accès.

Après avoir redéfini une variable binaire qui égale 1 quand l'individu fume et 0 sinon, nous choisissons trois variables étant susceptibles de pouvoir la déterminer. Motheduc représente le niveau de scolarité de la mère de l'individu en année, white est une variable dichotomique qui égale 1 si l'individu est blanc et 0 sinon, lfaminc est le logarithme naturel du revenu du ménage auquel l'individu appartient. Il est plausible de penser que plus la mère est éduquée moins l'individu aura tendance à fumer. Il en est de même pour la variable concernant le revenu, son augmentation aura un effet négatif sur la probabilité que l'individu fume.

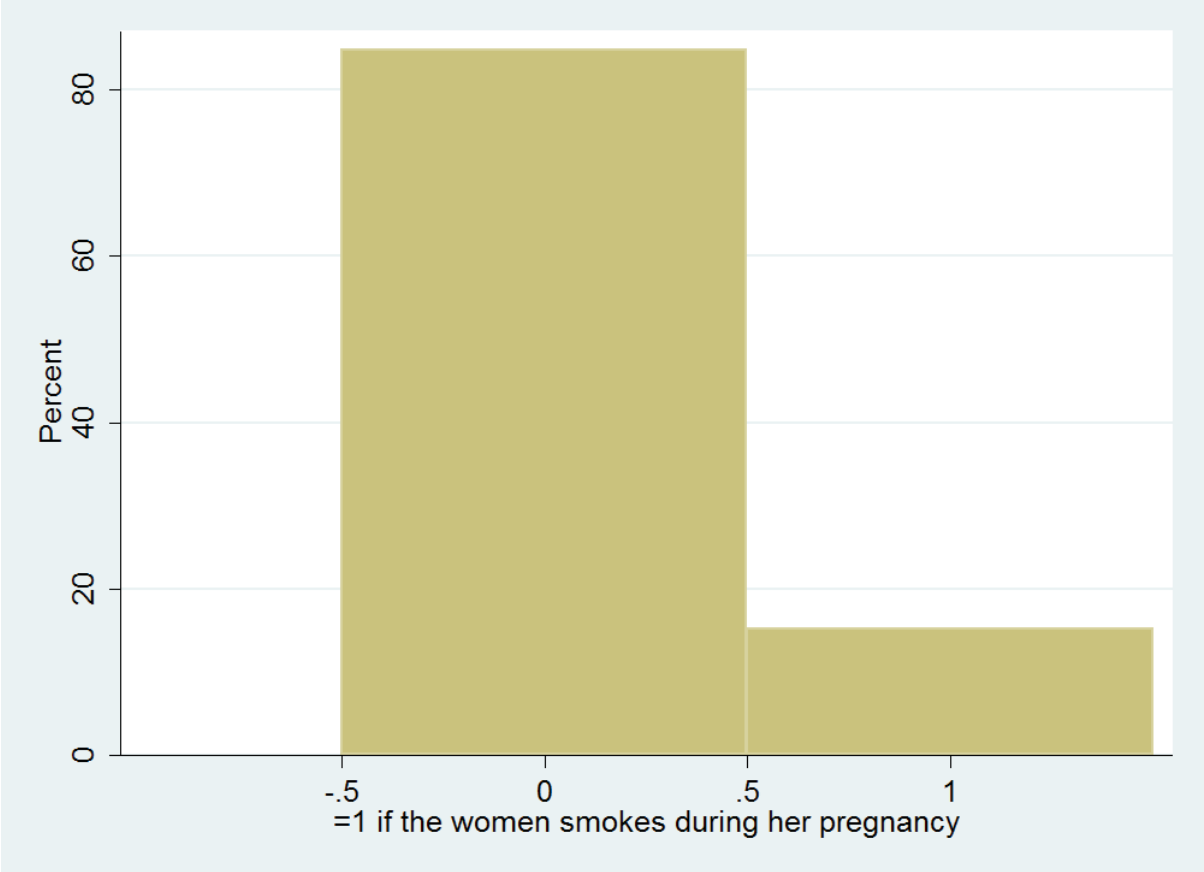
Description des données

Nous allons premièrement explorer les données disponibles.

```
describe smokes motheduc motheducc white lfaminc

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------------------
smokes          float  %9.0g                  =1 if the women smokes during her pregnancy
motheduc        byte   %8.0g                  mother's yrs of educ
motheducc        float  %9.0g                   Éducation de la mère: entre 0 et 8, 2: entre 8 et
                                               12, 3: plus de 12
white           byte   %8.0g                  =1 if white
lfaminc         float  %9.0g                  log(faminc)

summarize smokes motheduc motheducc white lfaminc

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      smokes |      1388    .1527378    .3598642          0          1
    motheduc |      1387    12.93583    2.376728          2         18
   motheducc |      1388    1.644092    .8047182          1          3
       white |      1388    .7845821    .4112601          0          1
     lfaminc |      1388    3.071271    .9180645  -.6931472   4.174387
```

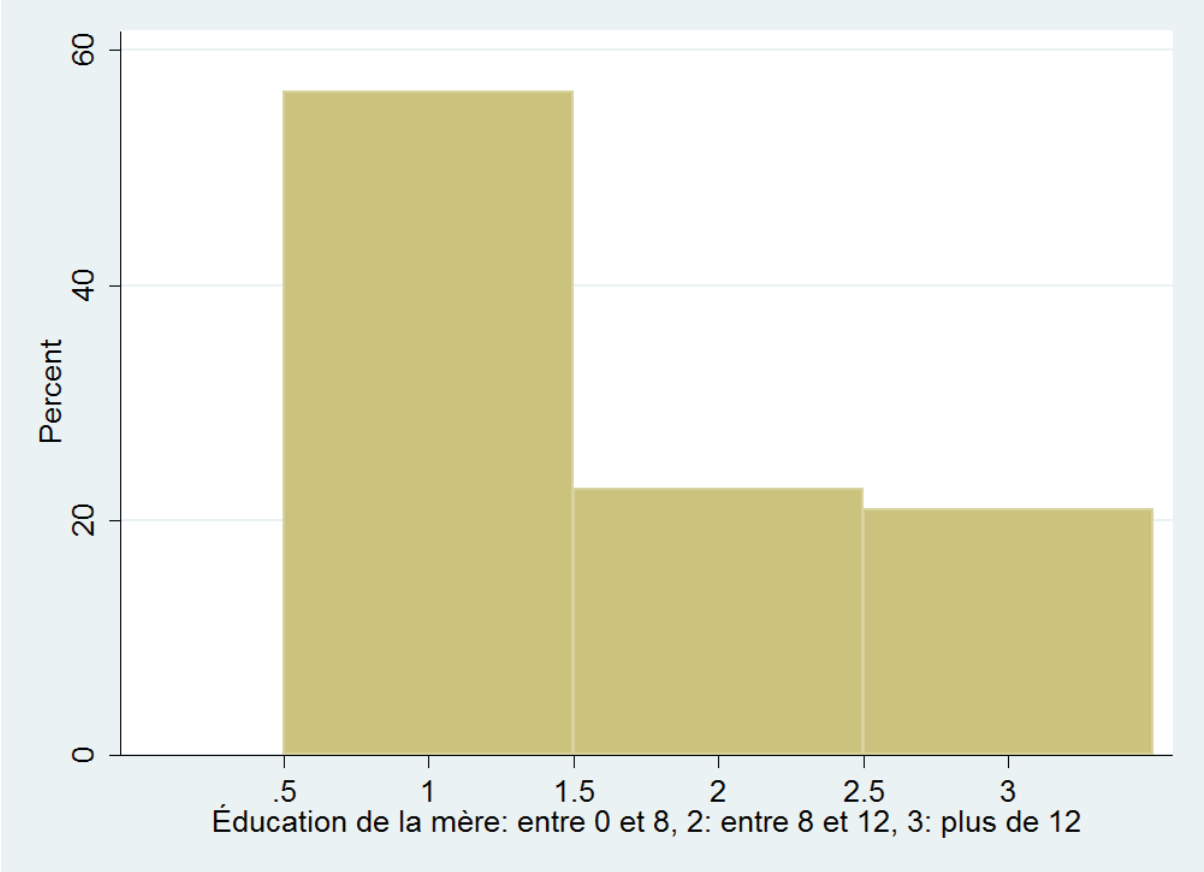
Nous voyons que les variables ont des valeurs minimum et maximum plausibles notament pour la variable dépendante et la variable dichotomique.

```
inspect smokes

smokes:  =1 if the women smokes during her      Number of Observations
------------------------------------------- ------------------------------
                                             Total   Integers   Nonintegers
|  #                          Negative          -         -          -
|  #                          Zero           1176      1176          -
|  #                          Positive        212       212          -
|  #                                         -----     -----      -----
|  #                          Total          1388      1388          -
|  #   #                      Missing           -
+----------------------                      -----
0                      1                     1388
   (2 unique values)

inspect motheducc

motheducc:  Éducation de la mère: entre 0 e     Number of Observations
------------------------------------------- ------------------------------
                                             Total   Integers   Nonintegers
|  #                          Negative          -         -          -
|  #                          Zero              -         -          -
|  #                          Positive       1388      1388          -
|  #                                         -----     -----      -----
|  #   #   #                  Total          1388      1388          -
|  #   #   #                  Missing           -
+----------------------                      -----
1                      3                     1388
   (3 unique values)


histogram smokes, discrete percent
(start=0, width=1)
```



Graphiquement, nous pouvons voir que peu de femmes de l'échantillon ont fumé pendant leur grossesse.

```
histogram motheducc, discrete percent
(start=1, width=1)
```




```
tab1 smokes motheduc motheducc lfaminc white

-> tabulation of smokes

  =1 if the |
      women |
     smokes |
 during her |
  pregnancy |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,176       84.73       84.73
          1 |        212       15.27      100.00
------------+-----------------------------------
      Total |      1,388      100.00

-> tabulation of motheduc

   mother's |
yrs of educ |      Freq.     Percent        Cum.
------------+-----------------------------------
          2 |          1        0.07        0.07
          3 |          1        0.07        0.14
          4 |          1        0.07        0.22
          5 |          2        0.14        0.36
          6 |          9        0.65        1.01
          7 |          7        0.50        1.51
          8 |         21        1.51        3.03
          9 |         43        3.10        6.13
         10 |         70        5.05       11.18
         11 |         67        4.83       16.01
         12 |        562       40.52       56.52
         13 |        122        8.80       65.32
         14 |        151       10.89       76.21
         15 |         41        2.96       79.16
```

```
         16 |        198       14.28       93.44
         17 |         37        2.67       96.11
         18 |         54        3.89      100.00
------------+-----------------------------------
      Total |      1,387      100.00
-> tabulation of motheducc

  Éducation |
de la mère: |
 entre 0 et |
8, 2: entre |
8 et 12, 3: |
 plus de 12 |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        784       56.48       56.48
          2 |        314       22.62       79.11
          3 |        290       20.89      100.00
------------+-----------------------------------
      Total |      1,388      100.00

-> tabulation of lfaminc

log(faminc) |      Freq.     Percent        Cum.
------------+-----------------------------------
  -.6931472 |         19        1.37        1.37
   .4054651 |         13        0.94        2.31
   .9162908 |         22        1.59        3.89
   1.252763 |         20        1.44        5.33
   1.504077 |         27        1.95        7.28
   1.704748 |         22        1.59        8.86
   1.871802 |         34        2.45       11.31
   2.014903 |         22        1.59       12.90
   2.140066 |         23        1.66       14.55
   2.251292 |         28        2.02       16.57
   2.351375 |         24        1.73       18.30
   2.442347 |         18        1.30       19.60
   2.525729 |         27        1.95       21.54
    2.60269 |         27        1.95       23.49
   2.674149 |         22        1.59       25.07
    2.74084 |         31        2.23       27.31
    2.80336 |         21        1.51       28.82
   2.862201 |         30        2.16       30.98
   2.917771 |         49        3.53       34.51
   2.970414 |         47        3.39       37.90
   3.113515 |        151       10.88       48.78
   3.314186 |        137        9.87       58.65
    3.48124 |        136        9.80       68.44
   3.624341 |         94        6.77       75.22
   3.749504 |         84        6.05       81.27
    3.86073 |         68        4.90       86.17
   4.174387 |        192       13.83      100.00
------------+-----------------------------------
      Total |      1,388      100.00

-> tabulation of white

=1 if white |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        299       21.54       21.54
          1 |      1,089       78.46      100.00
------------+-----------------------------------
      Total |      1,388      100.00
```

En observant les données dans les tables créées avec la commande tabulate, nous remarquons plus précisément la faible proportion de femmes ayant fumé pendant leur grossesse. Nous voyons aussi qu'il existe une accumulation dans la variable motheduc correspondant aux fins de cycles d'études. Une variable catégorielle est alors codée afin de prendre en compte leur effet.

Modélisation

Nous allons premièrement estimer un modèle de probabilité linéaire avec la commande `regress` tout en spécifiant l'option `robust` étant donné le problème connu d'hétéroscedasticité.

```
reg smokes c.motheduc i.white c.lfaminc, robust

Linear régression                                      Number of obs =    1387
                                                       F(  3,  1383) =   31.11
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0631
                                                       Root MSE      =  .34881

------------------------------------------------------------------------------
             |               Robust
      smokes |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    motheduc |  -.0293207   .0041666    -7.04   0.000    -.0374942   -.0211471
     1.white |    .048364     .02528     1.91   0.056    -.0012273    .0979552
     lfaminc |  -.0427425   .0132724    -3.22   0.001    -.0687787   -.0167062
       _cons |   .6255033   .0564588    11.08   0.000     .5147492    .7362573
------------------------------------------------------------------------------
```

Motheduc et lfaminc sont tous significatifs en considérant une erreur de première espèce de 5 % tandis que white est significatif en considérant une erreur de première espèce de 10 %. Les coefficients représentent les effets marginaux moyens. Une variation de motheduc d'une unité à un impact négatif de 2,9 % sur la probabilité de fumer, ceteris paribus.

Nous pouvons aussi calculer certaines mesures de qualité du modèle avec la commande `fitstat`. La plupart de ces statistiques ne sont pas absolues, mais relatives dans le sens où les statistiques de plusieurs modélisations doivent être comparées pour être utilisables. Nous indiquons donc à Stata de sauvegarder ces mesures sous le nom m1.

```
fitstat, saving(m1)

Measures of Fit for regress of smokes

Log-Lik Intercept Only:       -550.424   Log-Lik Full Model:           -505.227
D(1382):                      1010.454   LR(3):                          90.394
                                         Prob > LR:                       0.000
R2:                              0.063   Adjusted R2:                     0.061
AIC:                             0.736   AIC*n:                        1020.454
BIC:                         -8988.176   BIC':                          -68.690
BIC used by Stata:            1039.393   AIC used by Stata:            1018.454

(Indices saved in matrix fs_m1)
```

Nous allons maintenant observer les valeurs prédites que nous appellerons prhat avec la commande `predict`.

```
predict prhat
(option xb assumed; fitted values)
(1 missing value generated)
```

Nous savons qu'avec un modèle linéaire, les prédictions ne sont pas bornées entre 0 et 1. Ce qui s'avère un problème quand nous voulons interpréter la condition des individus ayant des valeurs prédites à l'extérieur de cet intervalle. Nous allons donc lister les individus étant dans cette situation.

```
list prhat if prhat<0

      +-----------+
      |     prhat |
      |-----------|
  10. | -.0088179 |
  46. | -.0220509 |
  50. | -.0323282 |
  99. | -.0030076 |
 104. | -.0323282 |
      |-----------|
 115. | -.0220509 |
 138. | -.0027014 |
 155. | -.0220509 |
 166. | -.0030076 |
 190. | -.0030076 |
      |-----------|
 196. | -.0323282 |
 227. | -.0806922 |
 232. | -.0323282 |
 237. | -.0030076 |
 247. | -.0323282 |
      |-----------|
        .
        .
        .
      |-----------|
1151. | -.0323282 |
1153. | -.0439251 |
1166. | -.0323282 |
1188. | -.0220509 |
1287. | -.0030076 |
      |-----------|
1303. | -.0323282 |
1307. | -.0806922 |
1315. | -.0323282 |
      +-----------+
```

```
list prhat if prhat>1

      +-------+
      | prhat |
      |-------|
 207. |     . |
      +-------+
```

De nombreuses valeurs prédites sont inférieures à 0 alors qu'aucune n'est supérieure à 1. Nous pouvons attribuer ce résultat à la faible fréquence de femmes ayant fumé pendant leur grossesse.

Afin de palier à ce problème de prédiction, nous allons maintenant utiliser un modèle de régression logistique et la commande logit.

```
logit smokes c.motheduc i.white c.lfaminc

Iteration 0:   log likelihood = -593.10529
Iteration 1:   log likelihood =  -551.1645
Iteration 2:   log likelihood = -548.74295
Iteration 3:   log likelihood = -548.73367
Iteration 4:   log likelihood = -548.73367

Logistic régression                               Number of obs   =       1387
                                                  LR chi2(3)      =      88.74
                                                  Prob > chi2     =     0.0000
Log likelihood = -548.73367                       Pseudo R2       =     0.0748

------------------------------------------------------------------------------
      smokes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    motheduc |  -.2518731   .0372045    -6.77   0.000    -.3247925   -.1789536
     1.white |   .3438842   .2002178     1.72   0.086    -.0485355    .7363038
     lfaminc |  -.2962299   .0866347    -3.42   0.001    -.4660308   -.1264289
       _cons |   2.012535   .4474219     4.50   0.000     1.135604    2.889466
------------------------------------------------------------------------------

estimates store m2
```

Motheduc et lfaminc sont tous les deux significatifs en considérant une erreur de première espèce de 5 % tandis que white est significatif en considérant une erreur de première espèce de 10 %. Les estimations ne sont pas directement interprétables, nous parlerons de l'impact des variables indépendantes dans la section destinée à leur interprétation.

Mesures d'adéquation

Nous pouvons par la suite vérifier la qualité du modèle en termes de classification grâce à plusieurs mesures que nous avons présentées dans la partie théorique (taux de bonnes classifications, courbe ROC). Nous allons premièrement afficher des statistiques simples de classement des individus par notre modèle avec la commande `estat classification`. Cette commande met en comparaison le classement de notre modèle par rapport à son statut véritable. Quand la variable dépendante est binaire, nous disposons de quatre cas résumés dans le tableau ci-dessous

	La femme a fumé pendant sa grossesse	La femme n'a pas fumé pendant sa grossesse
Le modèle prédit que la femme a fumé pendant sa grossesse	n_{11}	n_{12}
Le modèle prédit que la femme n'a pas fumé pendant sa grossesse	n_{21}	n_{22}

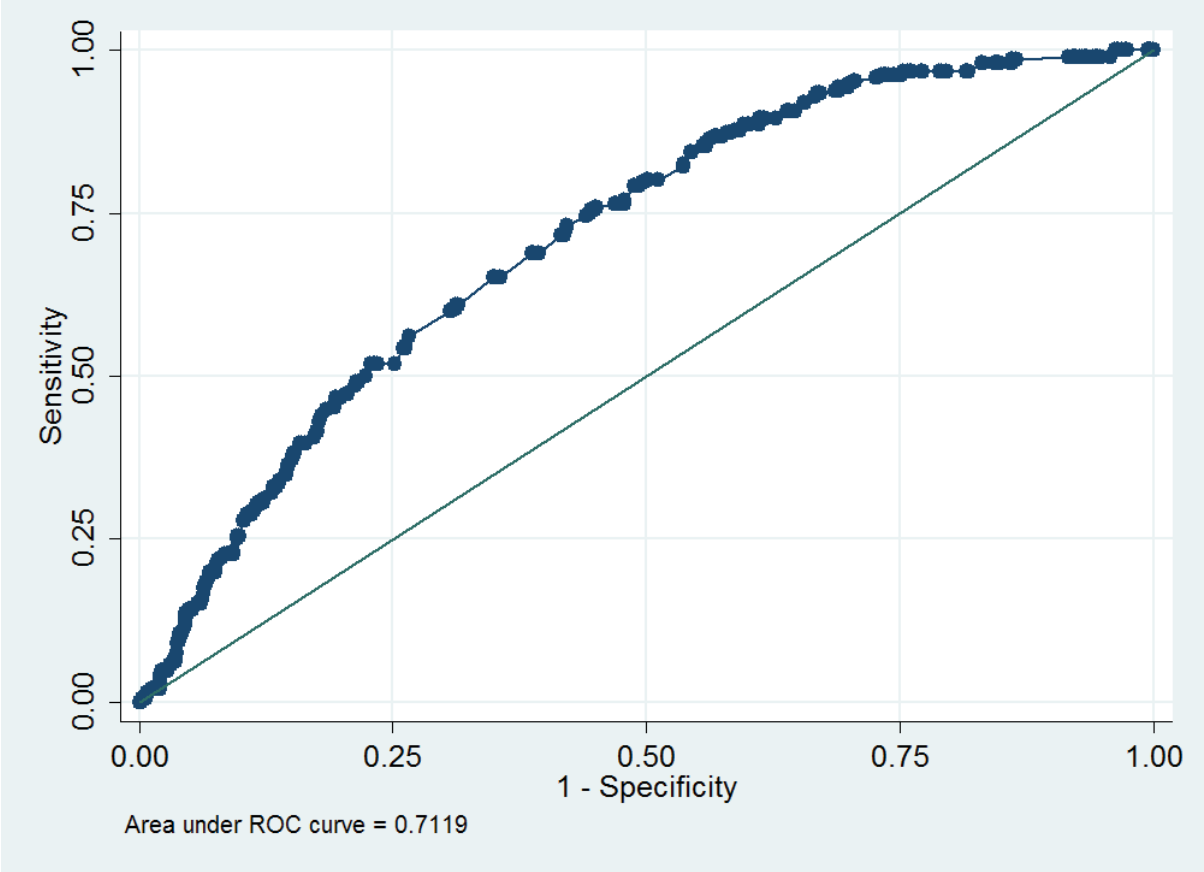
```
quietly logit smokes c.motheduc i.white c.lfaminc
```

```
estat classification

Logistic model for smokes

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         3             9  |         12
     -     |       209          1166  |       1375
-----------+--------------------------+-----------
   Total   |       212          1175  |       1387

Classified + if predicted Pr(D) >= .5
True D defined as smokes != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)    1.42%
Specificity                     Pr( -|~D)   99.23%
Positive predictive value       Pr( D| +)   25.00%
Negative predictive value       Pr(~D| -)   84.80%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    0.77%
False - rate for true D         Pr( -| D)   98.58%
False + rate for classified +   Pr(~D| +)   75.00%
False - rate for classified -   Pr( D| -)   15.20%
--------------------------------------------------
Correctly classified                        84.28%
--------------------------------------------------
```

Nous pouvons observer un classement global du modèle très satisfaisant comme il classe correctement les individus à 84,28 %. Cependant, si nous regardons de plus près la façon dont il classe les individus, nous remarquons que 0,77 % des femmes ayant fumé sont classées correctement.

Nous allons vérifier si notre modèle parvient à discriminer adéquatement les individus grâce à la courbe ROC avec la commande `lroc`.

```
lroc
```

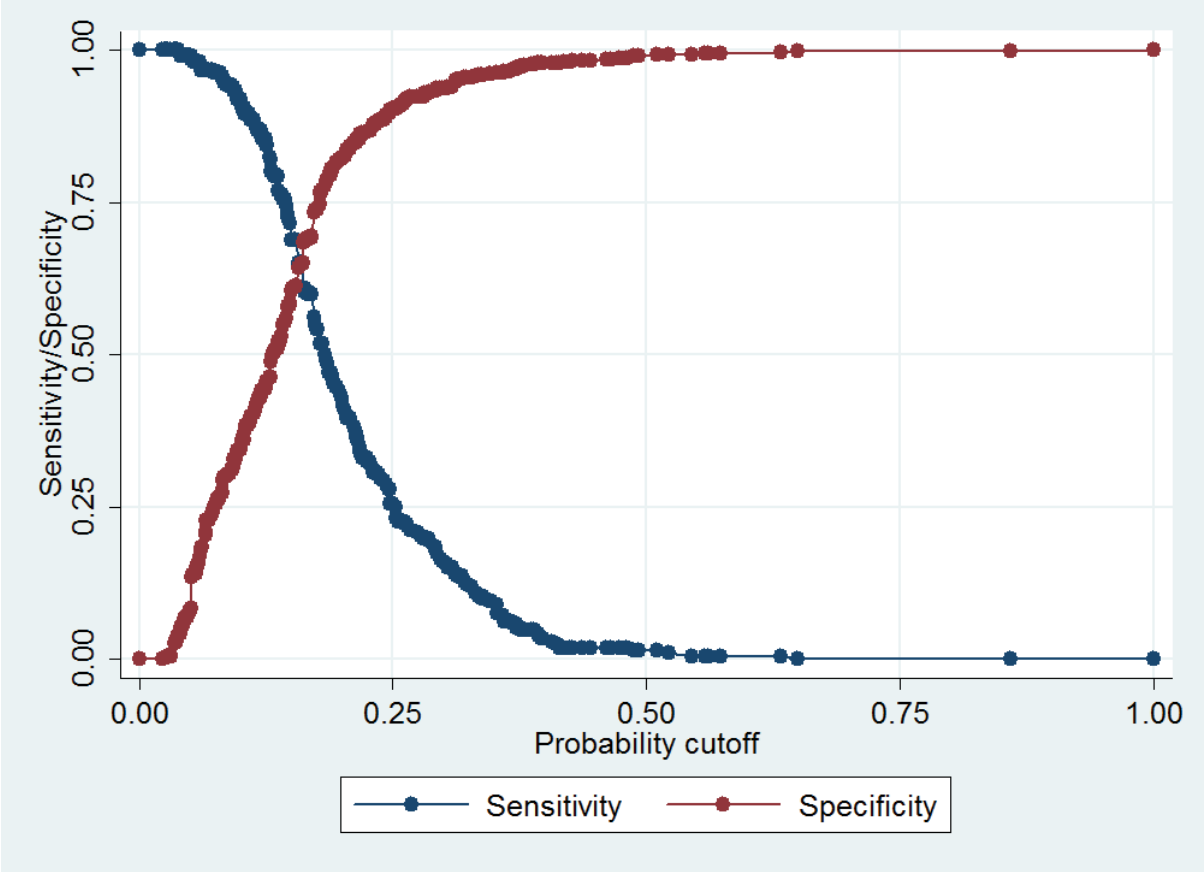



```
Logistic model for smokes

number of observations =     1387
area under ROC curve   =   0.7119
```

Nous pouvons graphiquement observer le modèle en termes de classification. La commande nous rapporte également l'aire sous la courbe ROC (*Area Under roc Curve*) de 0,7119 qui reflète une discrimination acceptable des individus (voir partie théorique pour les mesures).

Dans la même catégorie de mesures au niveau de la classification, nous pouvons confronter la spécificité à la sensitivité par rapport au point de coupure choisi pour classer les observations.

```
lsens
```



Si nous voulons que le modèle puisse prédire chaque modalité plus équitablement, nous pouvons spécifier manuellement le point de coupure (*probability cutoff*). Nous pouvons prendre comme base l'intersection des deux courbes qui donne une spécificité égale à la sensitivité. Cette approche est intéressante si le but de la modélisation est de prédire quelles femmes sont plus à risque de fumer pendant leur grossesse.

Afin de pouvoir valider le modèle au niveau de sa forme et des variables choisies, il est possible de générer d'autres tests comme le test de Hosmer-Lemeshow. Ce test suit une loi de χ^2 avec $g-2$ degrés de liberté ou g est le nombre de groupes. Les hypothèses testées sont

H_0: le modèle est approprié aux données

H_1: le modèle n'est pas approprié aux données

Le lecteur intéressé par la théorie derrière ce test peut se référer à Lemeshow & Hosmer (1982).

```
estat gof
Logistic model for smokes, goodness-of-fit test
```

```
       number of observations =      1387
 number of covariate patterns =       320
            Pearson chi2(316) =       322.60
                  Prob > chi2 =         0.3870
```

Nous constatons que le modèle est adéquat, car la p-valeur est supérieure à 10 %.

Afin de nous assurer que la statistique ne soit pas calculée avec un trop grand nombre de groupes, nous pouvons former manuellement 10 groupes pour la calculer (cette manipulation est souvent nécessaire pour les échantillons de petite taille). Nous ajoutons l'option `group(10)` et obtenons les résultats suivants.

```
estat gof, group(10)

Logistic model for smokes, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)

       number of observations =      1387
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =         7.31
                  Prob > chi2 =         0.5039
```

Ici, nous constatons que le modèle est adéquat, car la p-valeur est supérieure à 10 %.

Nous pouvons tester avec `linktest` si le modèle est bien spécifié en régressant la variable dépendante sur les valeurs prédites et les valeurs prédites au carré pour s'assurer que ceux-ci n'ont aucun pouvoir explicatif.

```
linktest

Iteration 0:   log likelihood = -593.10529
Iteration 1:   log likelihood = -547.94911
Iteration 2:   log likelihood = -539.04634
Iteration 3:   log likelihood = -538.91661
Iteration 4:   log likelihood = -538.91656

Logistic régression                               Number of obs   =       1387
                                                  LR chi2(2)      =     108.38
                                                  Prob > chi2     =     0.0000
Log likelihood = -538.91656                       Pseudo R2       =     0.0914

------------------------------------------------------------------------------
      smokes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   -.414068   .3896613    -1.06   0.288     -1.17779    .3496541
      _hatsq |  -.4940154   .1302451    -3.79   0.000    -.7492911   -.2387397
       _cons |   -.803427   .2997056    -2.68   0.007    -1.390839   -.2160147
------------------------------------------------------------------------------
```

Nous remarquons que les valeurs prédites ne sont pas significatives tandis que les valeurs prédites au carré le sont. Les variables choisies n'ont donc pas un pouvoir de prédiction assez fort et le modèle peut être mal spécifié.

Il faudrait donc trouver d'autres variables à ajouter au modèle. Les autres variables de la base de données ne nous apportant pas plus de significativité, nous continuerons donc à utiliser cette modélisation pour exécuter les exemples des commandes.

Nous sauvegardons par la suite les mesures d'adéquation du modèle données par la commande `fitstat` sous le nom m2.

```
fitstat, saving(m2)

Measures of Fit for logit of smokes

Log-Lik Intercept Only:       -593.105   Log-Lik Full Model:           -548.734
D(1382):                      1097.467   LR(3):                          88.743
                                         Prob > LR:                       0.000
McFadden's R2:                   0.075   McFadden's Adj R2:               0.066
ML (Cox-Snell) R2:               0.062   Cragg-Uhler(Nagelkerke) R2:      0.108
McKelvey & Zavoina's R2:         0.141   Efron's R2:                      0.053
Variance of y*:                  3.830   Variance of error:               3.290
Count R2:                        0.843   Adj Count R2:                   -0.028
AIC:                             0.798   AIC*n:                        1107.467
BIC:                         -8901.162   BIC':                          -67.039
BIC used by Stata:            1126.407   AIC used by Stata:            1105.467

(Indices saved in matrix fs_m2)
```

Tests d'inférence

Nous utilisons ensuite un test de Wald pour savoir si white et motheduc sont conjointement égales à 0 avec la commande `test`.

```
test 1.white c.motheduc

 ( 1)  [smokes]1.white = 0
 ( 2)  [smokes]motheduc = 0

           chi2(  2) =   49.73
         Prob > chi2 =    0.0000
```

Nous rejetons fortement l'hypothèse nulle de nullité simultanée des coefficients.

Un autre moyen de tester leur significativité est d'utiliser un test du rapport de vraisemblance en comparant le modèle contraint au modèle complet. Nous spécifions que la variable motheduc ne doit pas être manquante pour avoir les mêmes observations que dans la régression précédente. Nous commençons par estimer le modèle contraint et nous sauvegardons les estimations sous le nom mtest. Nous utilisons ensuite la commande `lrtest` en prenant soins de placer le modèle complet avant le modèle contraint.

```
quietly logit smokes c.lfaminc if motheduc~=.

estimates store mtest
```

```
lrtest m2 mtest

Likelihood-ratio test                                 LR chi2(2)  =     53.24
(Assumption: mtest nested in m2)                      Prob > chi2 =    0.0000
```

Nous tirons la ici la même conclusion qu'avec le test de Wald. Nous rejetons fortement l'hypothèse nulle.

Interprétation des relations entre variables dépendantes et variables indépendantes

Comme les coefficients β ne sont pas directement interprétables, nous devons utiliser une transformation ou d'autres calculs pour vérifier l'impact des variables. Nous commencerons avec la transformation en termes d'*odds ratios* en spécifiant l'option `or` après la commande `logit`.

```
logit, or

Logistic régression                               Number of obs   =       1387
                                                  LR chi2(3)      =      88.74
                                                  Prob > chi2     =     0.0000
Log likelihood = -548.73367                       Pseudo R2       =     0.0748

------------------------------------------------------------------------------
      smokes | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    motheduc |   .7773434   .0289207    -6.77   0.000     .7226773    .8361447
     1.white |   1.410415   .2823902     1.72   0.086     .9526236    2.088203
     lfaminc |   .7436165    .064423    -3.42   0.001     .6274879    .8812368
       _cons |   7.482259   3.347727     4.50   0.000     3.113053     17.9837
------------------------------------------------------------------------------
```

Nous voyons donc que si l'éducation de la mère varie d'une année, l'individu aura 0,777 fois moins de chance de fumer, toute chose étant égale par ailleurs. Nous pouvons aussi affirmer que l'individu aura 1,287 fois plus de chances (1/0,777) de ne pas fumer, en moyenne. Si la personne est blanche, elle aura 1,41 fois plus de chances de fumer, *ceteris paribus*.

Grâce à la commande `margins` et sa panoplie d'options, nous pouvons aussi calculer l'impact marginal des variables et les représenter graphiquement en exécutant la commande `marginsplot` après l'estimation. Nous pouvons calculer ces effets marginaux à différentes valeurs des variables nous permettant d'explorer les effets en détail comme ils sont non linéaires.

Nous estimons premièrement l'effet marginal moyen de toutes les variables

```
margins, dydx(*)

Average marginal effects                          Number of obs   =       1387
Model VCE    : OIM

Expression   : Pr(smokes), predict()
dy/dx w.r.t. : motheduc 1.white lfaminc
```

```
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    motheduc |  -.0303406   .0043807    -6.93   0.000    -.0389266   -.0217545
     1.white |    .039117   .0214193     1.83   0.068     -.002864    .0810981
     lfaminc |  -.0356838   .0103641    -3.44   0.001     -.055997   -.0153706
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Une variation unitaire de la variable motheduc lorsque celle-ci est à sa moyenne amène une variation négative de la probabilité qu'une femme ait fumé pendant sa grossesse de 3 %, toute chose étant égale par ailleurs. *Ceteris paribus*, une femme blanche a une probabilité plus grande de 3,9 % d'avoir fumé pendant sa grossesse. De plus, en moyenne, une variation du revenu familial de 1 pour cent fait baisser la probabilité qu'une femme ait fumé pendant sa grossesse de 3,6 %. Nous voyons encore que les variables motheduc et lfaminc sont largement significatives tandis que la variable white n'est significative qu'à un niveau de 10 %.

Nous pouvons aussi estimer les effets marginaux lorsque les variables indépendantes sont à leur moyenne. Il faut aussi noter que si nous spécifions `atmeans` et qu'il existe des variables catégorielles et binaires, il faut alors spécifier des valeurs d'intérêt pour celles-ci étant donné l'interprétation ambiguë d'une moyenne dans ce cas.

Nous allons cependant utiliser l'option `atmeans` pour calculer l'effet moyen à la moyenne étant donné que celle-ci ne change pas les résultats alors que la variable white est binaire.

```
margins, dydx(*) atmeans

Conditional marginal effects                      Number of obs   =       1387
Model VCE    : OIM

Expression   : Pr(smokes), predict()
dy/dx w.r.t. : motheduc 1.white lfaminc
at           : motheduc        =    12.93583 (mean)
               0.white         =    .2155732 (mean)
               1.white         =    .7844268 (mean)
               lfaminc         =    3.072033 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    motheduc |  -.0288047   .0040128    -7.18   0.000    -.0366697   -.0209397
     1.white |   .0366196    .019809     1.85   0.065    -.0022054    .0754445
     lfaminc |  -.0338774   .0098448    -3.44   0.001    -.0531728    -.014582
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Une différence est visible entre les deux calculs comme le niveau de chaque variable est prise en compte dans le calcul des effets marginaux.

Nous pouvons également calculer l'effet marginal à la médiane des variables motheduc et lfaminc en spécifiant l'option `at((median)motheduc lfaminc)`.

```
margins, dydx(*) at((median)motheduc lfaminc)

Conditional marginal effects                      Number of obs   =       1387
Model VCE    : OIM

Expression   : Pr(smokes), predict()
dy/dx w.r.t. : motheduc 1.white lfaminc
at           : motheduc        =          12 (median)
               lfaminc         =    3.314186 (median)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    motheduc |  -.0324794   .0054252    -5.99   0.000    -.0431126   -.0218461
     1.white |   .0413272   .0221581     1.87   0.062    -.0021019    .0847562
     lfaminc |  -.0381993   .0102806    -3.72   0.000    -.0583488   -.0180497
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Encore une fois, nous observons une différence dans les effets marginaux.

Étant donné les résultats obtenus précédemment, nous allons premièrement calculer les effets marginaux moyens et les effets marginaux à la moyenne à plusieurs valeurs de la variable lfaminc en différentiant par ethnie. Deuxièmement, nous allons calculer les effets marginaux moyens et les effets marginaux à la moyenne à plusieurs valeurs de la variable motheduc en différentiant par ethnie. Nous combinerons donc les options `at()` et `atmeans`.

```
margins white, dydx(motheduc) at(lfaminc=(-1(0.2)5)) atmeans

Conditional marginal effects                      Number of obs   =       1387
Model VCE    : OIM

Expression   : Pr(smokes), predict()
dy/dx w.r.t. : motheduc

1._at        : motheduc        =    12.93583 (mean)
               0.white         =    .2155732 (mean)
               1.white         =    .7844268 (mean)
               lfaminc         =          -1

2._at        : motheduc        =    12.93583 (mean)
               0.white         =    .2155732 (mean)
               1.white         =    .7844268 (mean)
               lfaminc         =         -.8

                              .
                              .
                              .

30._at       : motheduc        =    12.93583 (mean)
               0.white         =    .2155732 (mean)
               1.white         =    .7844268 (mean)
               lfaminc         =         4.8

31._at       : motheduc        =    12.93583 (mean)
               0.white         =    .2155732 (mean)
               1.white         =    .7844268 (mean)
```

```
              lfaminc         =           5

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
motheduc     |
   _at#white |
        1 0  |  -.0506666   .0082234    -6.16   0.000    -.0667842    -.034549
        1 1  |  -.0575313   .0084781    -6.79   0.000    -.0741481   -.0409145
        2 0  |  -.0493225   .0080184    -6.15   0.000    -.0650382   -.0336068
        2 1  |  -.0564933   .0084036    -6.72   0.000    -.0729641   -.0400226
                                             .
                                             .
                                             .

       30 0  |   -.015289   .0044709    -3.42   0.001    -.0240518   -.0065262
       30 1  |  -.0204591   .0043571    -4.70   0.000     -.028999   -.0119193
       31 0  |  -.0145177   .0044489    -3.26   0.001    -.0232375    -.005798
       31 1  |  -.0194816   .0043919    -4.44   0.000    -.0280896   -.0108736
------------------------------------------------------------------------------
```

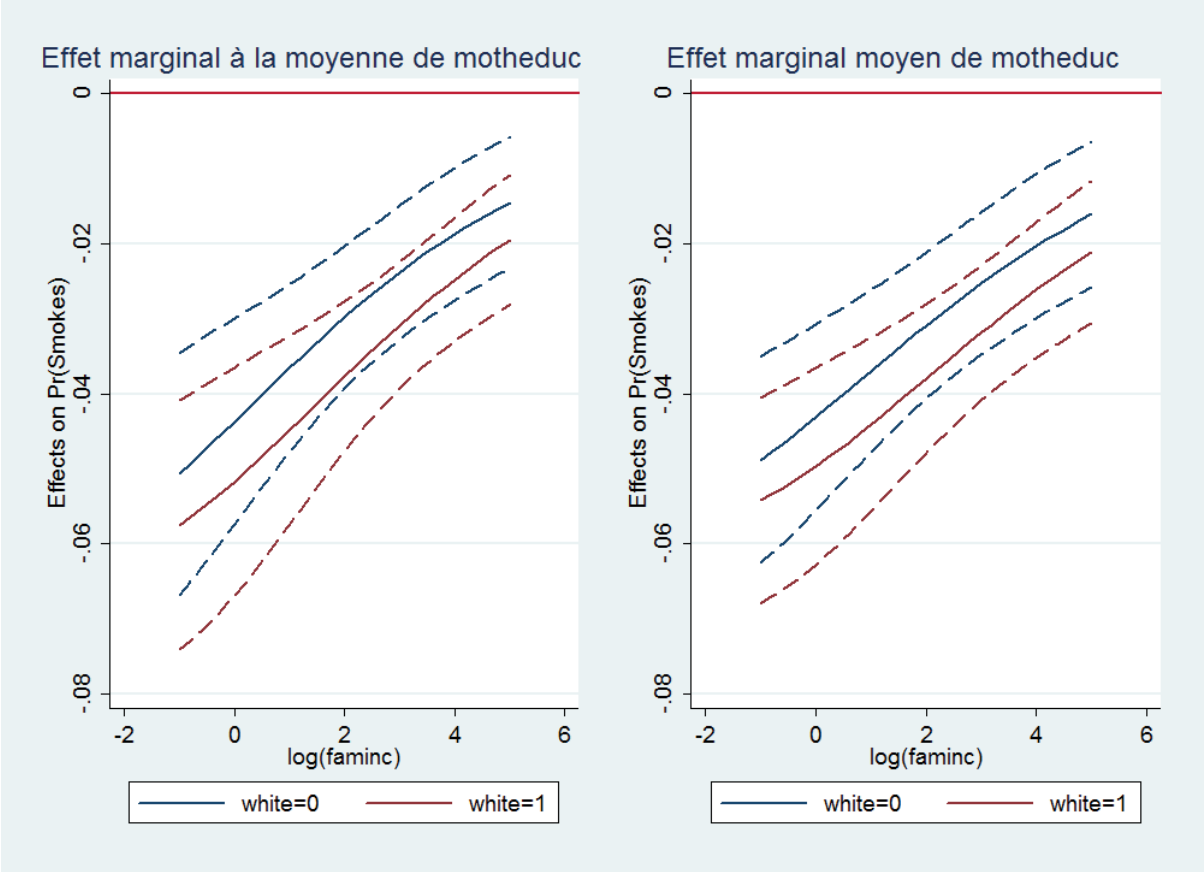
La première chose à observer est que l'effet marginal de motheduc est significatif pour toutes les valeurs de lfaminc et de white. Nous pouvons donc interpréter cet effet marginal sans ambiguïté et le représenter graphiquement avec la commande marginsplot. Nous ajoutons les options `recast(line)` pour spécifier que les effets doivent être représentés par une ligne, `recastci(rline)` `ciopts(lpattern(dash))` pour que les intervalles de confiance soient représentés par des lignes discontinues et nous ajoutons un titre avec la commande `title()`. Une ligne horizontale est aussi ajoutée pour y=0 pour avoir un repère au niveau de la significativité de l'effet marginal. Nous enregistrons le graphique avec la commande `graph save` pour pouvoir le combiner avec les graphiques suivant par la suite.

```
marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) title(Effet marginal à la
moyenne de motheduc) yline(0)

  Variables that uniquely identify margins: lfaminc white

graph save margins9.gph, replace
```

Nous effectuons la même opération pour obtenir les effets marginaux à la moyenne des variables afin de les représenter graphiquement. Nous enlevons donc l'option `atmeans` de la commande `margins`.

```
margins white, dydx(motheduc) at(lfaminc=(-1(0.2)5))

Average marginal effects                        Number of obs   =       1387
Model VCE    : OIM

Expression   : Pr(smokes), predict()
dy/dx w.r.t. : motheduc

1._at        : lfaminc         =          -1

2._at        : lfaminc         =         -.8

               .
```

```
                .
                .

30._at       : lfaminc         =         4.8

31._at       : lfaminc         =           5

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
motheduc     |
   _at#white |
        1 0  |  -.0488283   .0070219    -6.95   0.000     -.062591   -.0350655
        1 1  |  -.0542235   .0069934    -7.75   0.000    -.0679304   -.0405166
        2 0  |   -.047743   .0069152    -6.90   0.000    -.0612966   -.0341895
        2 1  |  -.0534241   .0069774    -7.66   0.000    -.0670996   -.0397487

                                                  .
                                                  .
                                                  .

       30 0  |  -.0168555   .0049286    -3.42   0.001    -.0265153   -.0071957
       30 1  |  -.0220345   .0047967    -4.59   0.000    -.0314358   -.0126332
       31 0  |  -.0160641   .0049198    -3.27   0.001    -.0257068   -.0064214
       31 1  |  -.0210714   .0048396    -4.35   0.000    -.0305568   -.0115859
------------------------------------------------------------------------------

marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) title(Effet marginal moyen de
motheduc) yline(0)

  Variables that uniquely identify margins: lfaminc white

graph save margins2.gph, replace

graph combine margins1.gph margins2.gph, ycommon
```



Graphiquement, nous distinguons mieux les effets de motheduc quand lfaminc varie. Nous observons que plus le revenu familial augmente moins l'effet de l'éducation de la mère a un impact négatif sur la probabilité de fumer, toute chose égale par ailleurs. Nous notons aussi que l'effet négatif de motheduc est relativement plus marqué pour les femmes blanches, ceteris paribus. L'effet marginal moyen n'est pas très différent de l'effet marginal à la moyenne.

Nous obtenons maintenant les effets marginaux de lfaminc par rapport à une variation selon l'ethnie et reproduisons un autre graphique.

Premièrement, calculons les effets marginaux à la moyenne

```
margins white, dydx(lfaminc) at(lfaminc=(-1(0.2)5)) atmeans

Conditional marginal effects                    Number of obs   =       1387
Model VCE    : OIM

Expression   : Pr(smokes), predict()
dy/dx w.r.t. : lfaminc

1._at        : motheduc        =    12.93583 (mean)
               0.white         =    .2155732 (mean)
               1.white         =    .7844268 (mean)
               lfaminc         =          -1
```

```
2._at        : motheduc        =    12.93583 (mean)
               0.white         =    .2155732 (mean)
               1.white         =    .7844268 (mean)
               lfaminc         =         -.8

                      .
                      .
                      .

30._at       : motheduc        =    12.93583 (mean)
               0.white         =    .2155732 (mean)
               1.white         =    .7844268 (mean)
               lfaminc         =         4.8

31._at       : motheduc        =    12.93583 (mean)
               0.white         =    .2155732 (mean)
               1.white         =    .7844268 (mean)
               lfaminc         =           5

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lfaminc      |
   _at#white |
        1 0  |  -.0595894   .0252953    -2.36   0.018    -.1091673   -.0100116
        1 1  |   -.067663   .0269477    -2.51   0.012    -.1204795   -.0148464
        2 0  |  -.0580086   .0245868    -2.36   0.018    -.1061978   -.0098194
        2 1  |  -.0664423   .0267396    -2.48   0.013    -.1188509   -.0140336

                                      .
                                      .
                                      .

       30 0  |  -.0179815   .0032309    -5.57   0.000    -.0243139   -.0116491
       30 1  |  -.0240622   .0045479    -5.29   0.000    -.0329759   -.0151484
       31 0  |  -.0170744    .002933    -5.82   0.000    -.0228229   -.0113259
       31 1  |  -.0229125   .0040118    -5.71   0.000    -.0307755   -.0150495
------------------------------------------------------------------------------
```

Graphique des effets marginaux et enregistrement de ceux-ci.

```
marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) title(Effet marginal à la
moyenne de lfaminc) yline(0)

  Variables that uniquely identify margins: lfaminc white

graph save margins3.gph, replace
(file margins3.gph saved)
```

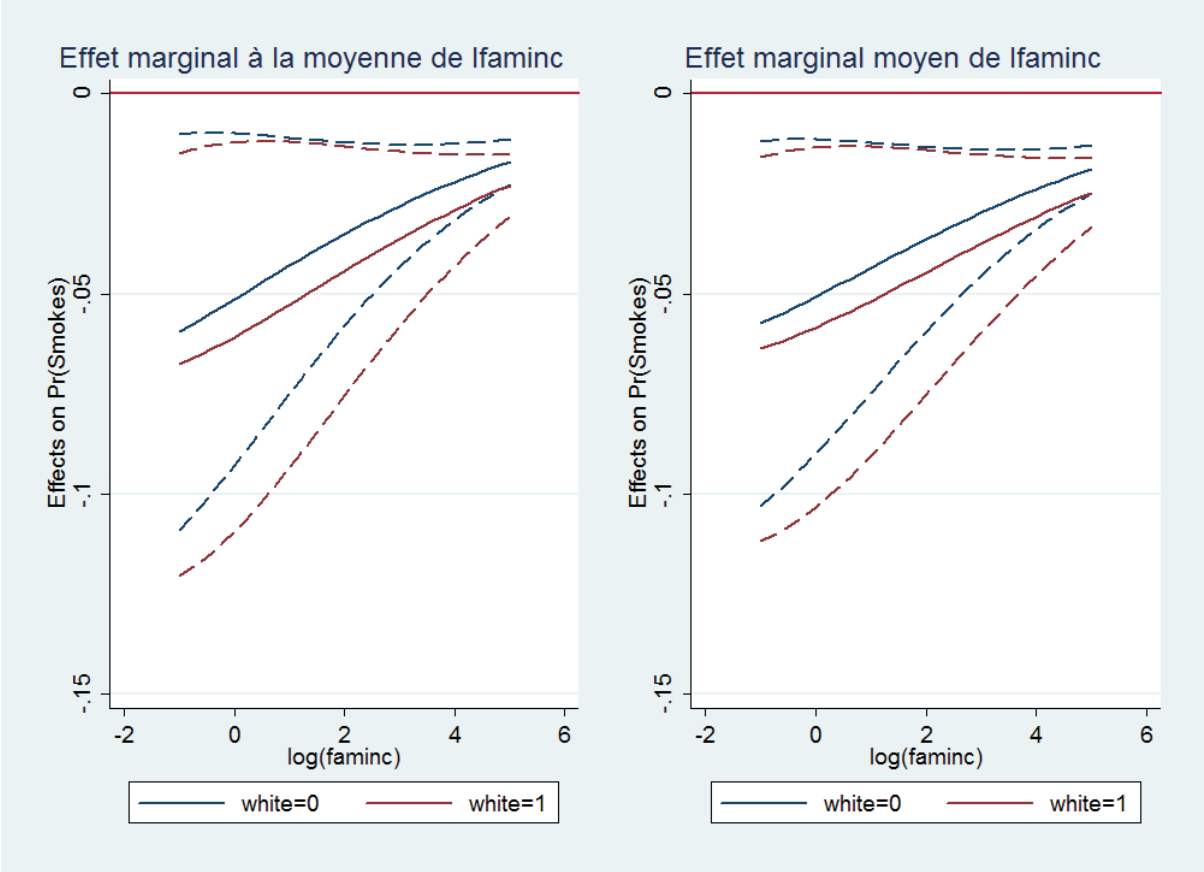
Deuxièmement, calculons les effets marginaux moyens

```
margins white, dydx(lfaminc) at(lfaminc=(-1(0.2)5))

Average marginal effects                        Number of obs   =       1387
Model VCE    : OIM

Expression   : Pr(smokes), predict()
dy/dx w.r.t. : lfaminc

1._at        : lfaminc         =          -1

2._at        : lfaminc         =         -.8

                      .
                      .
                      .

30._at       : lfaminc         =         4.8
```

```
31._at       : lfaminc         =           5

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
lfaminc      |
   _at#white |
        1 0  |  -.0574273   .0232873    -2.47   0.014    -.1030695   -.0117851
        1 1  |  -.0637727   .0244669    -2.61   0.009    -.1117269   -.0158185
        2 0  |   -.056151   .0227419    -2.47   0.014    -.1007242   -.0115777
        2 1  |  -.0628326   .0243279    -2.58   0.010    -.1105144   -.0151507

                                      .
                                      .
                                      .

       30 0  |  -.0198239   .0034014    -5.83   0.000    -.0264905   -.0131573
       30 1  |   -.025915    .004927    -5.26   0.000    -.0355717   -.0162583
       31 0  |  -.0188931   .0030763    -6.14   0.000    -.0249227   -.0128636
       31 1  |  -.0247822   .0043704    -5.67   0.000     -.033348   -.0162164
------------------------------------------------------------------------------
```

Nous reproduisons un graphique de ces effets et produisons un graphique combiné.

```
marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) title(Effet marginal moyen de
lfaminc) yline(0)

  Variables that uniquely identify margins: lfaminc white

graph save margins4.gph, replace
(file margins4.gph saved)

graph combine margins3.gph margins4.gph, ycommon
```



Nous pouvons remarquer que le logarithme du revenu familial a un effet marginal de moins en moins négatif sur la probabilité de fumer en moyenne et lorsque les variables sont à la moyenne.

Il est aussi possible de calculer les prédictions du modèle en enlevant l'option `dydx()`.

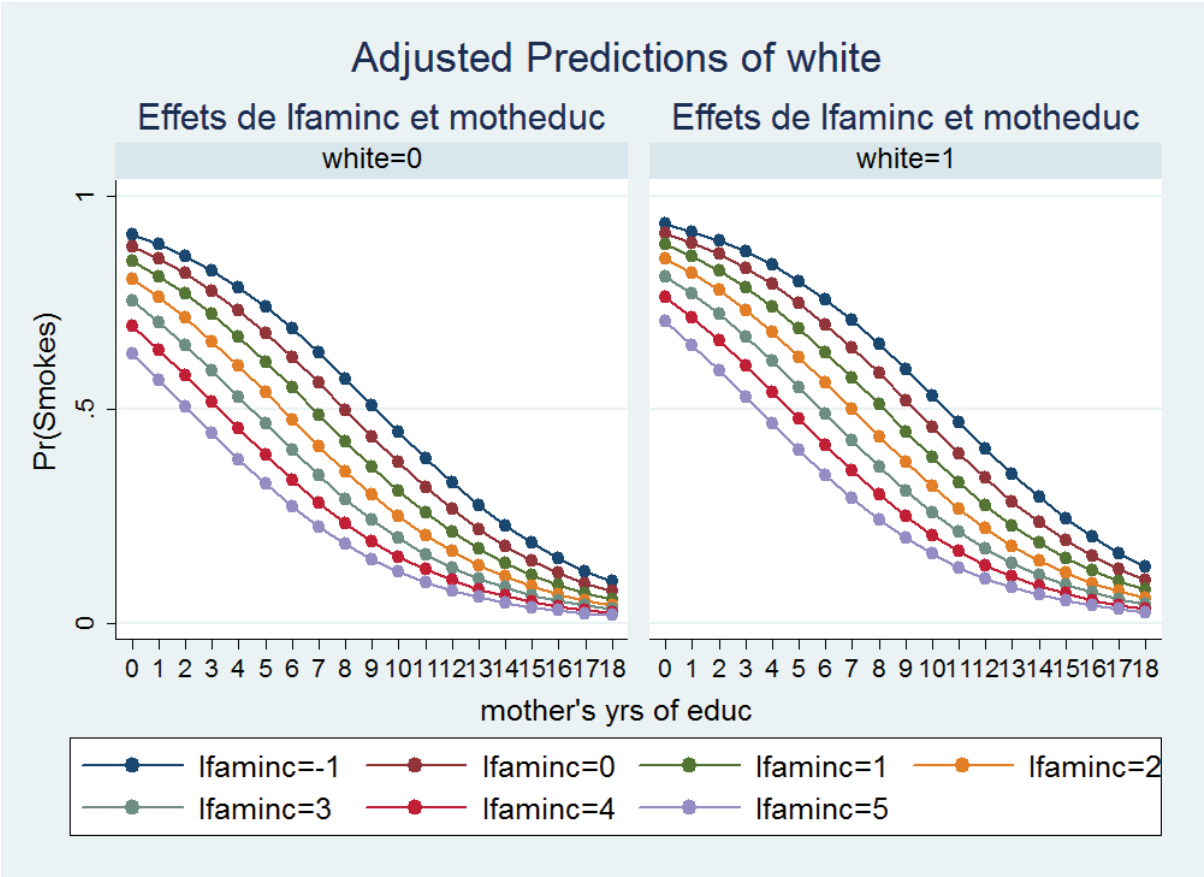
```
margins white, at(motheduc=(0(1)18)lfaminc=(-1(1)5))

Adjusted prédictions                            Number of obs   =       1387
Model VCE    : OIM

Expression   : Pr(smokes), predict()

1._at        : motheduc        =           0
               lfaminc         =          -1

2._at        : motheduc        =           0
               lfaminc         =           0

3._at        : motheduc        =           0
               lfaminc         =           1

4._at        : motheduc        =           0
               lfaminc         =           2

5._at        : motheduc        =           0
               lfaminc         =           3

6._at        : motheduc        =           0
               lfaminc         =           4
```

```
7._at        : motheduc        =           0
               lfaminc         =           5

8._at        : motheduc        =           1
               lfaminc         =          -1

9._at        : motheduc        =           1
               lfaminc         =           0

10._at       : motheduc        =           1
               lfaminc         =           1

11._at       : motheduc        =           1
               lfaminc         =           2

                      .
                      .
                      .


127._at      : motheduc        =          18
               lfaminc         =          -1

128._at      : motheduc        =          18
               lfaminc         =           0

129._at      : motheduc        =          18
               lfaminc         =           1

130._at      : motheduc        =          18
               lfaminc         =           2

131._at      : motheduc        =          18
               lfaminc         =           3

132._at      : motheduc        =          18
               lfaminc         =           4

133._at      : motheduc        =          18
               lfaminc         =           5


------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _at#white |
        1 0  |   .9096003   .0382921    23.75   0.000     .8345493    .9846514
        1 1  |   .9341741   .0298912    31.25   0.000     .8755884    .9927597
        2 0  |   .8821069   .0465293    18.96   0.000     .7909111    .9733027
        2 1  |   .9134431   .0359642    25.40   0.000     .8429546    .9839316
        3 0  |   .8476523   .0575385    14.73   0.000     .7348789    .9604256
        3 1  |   .8869731   .0439897    20.16   0.000     .8007548    .9731914
        4 0  |   .8053502   .0721561    11.16   0.000     .6639268    .9467735
        4 1  |   .8537048   .0548936    15.55   0.000     .7461153    .9612943
        5 0  |   .7547013   .0907365     8.32   0.000      .576861    .9325415
        5 1  |    .812712   .0695586    11.68   0.000     .6763797    .9490443
        6 0  |   .6958504   .1127078     6.17   0.000     .4749471    .9167538
        6 1  |   .7634159   .0883899     8.64   0.000      .590175    .9366569
        7 0  |   .6298062   .1363437     4.62   0.000     .3625774    .8970349
        7 1  |    .705841   .1108789     6.37   0.000     .4885223    .9231597
        8 0  |   .8866421   .0440433    20.13   0.000     .8003188    .9729655
        8 1  |   .9168864   .0350889    26.13   0.000     .8481135    .9856594

                                          .
                                          .
                                          .

      127 0  |   .0975322   .0387069     2.52   0.012     .0216681    .1733963
      127 1  |   .1322664   .0541938     2.44   0.015     .0260485    .2384844
      128 0  |   .0743866   .0259449     2.87   0.004     .0235355    .1252377
```

```
     128 1  |   .1018079   .0364471     2.79   0.005     .0303729    .1732429
     129 0  |   .0563906    .017292     3.26   0.001     .0224988    .0902823
     129 1  |   .0777351   .0237537     3.27   0.001     .0311787    .1242914
     130 0  |   .0425481   .0118235     3.60   0.000     .0193744    .0657218
     130 1  |   .0589806    .015271     3.86   0.000       .02905    .0889111
     131 0  |   .0319884   .0086413     3.70   0.000     .0150518     .048925
     131 1  |   .0445323   .0101264     4.40   0.000     .0246849    .0643798
     132 0  |   .0239838   .0068858     3.48   0.000     .0104879    .0374797
     132 1  |   .0334974   .0074244     4.51   0.000     .0189458     .048049
     133 0  |   .0179451     .00585     3.07   0.002     .0064793    .0294109
     133 1  |    .025125   .0061746     4.07   0.000     .0130231    .0372269
------------------------------------------------------------------------------
```

Nous pouvons représenter graphiquement ces prédictions grâce à la commande marginsplot.

```
marginsplot, noci by(white) legend(cols(4)) title(Effets de lfaminc et motheduc)
```



Nous pouvons globalement observer l'effet des variables sur les probabilités prédites. Clairement, nous observons un effet négatif de la variable motheduc sur la probabilité de fumer. Le revenu a aussi un effet négatif sur la probabilité de fumer durant la grossesse tandis qu'à travers les effets des deux autres variables les femmes blanches ont plus de probabilité de fumer.

Tests sur les variables catégorielles

La variable motheducc correspond à certaines valeurs de l'éducation et est codée comme suit 1 : diplôme du secondaire (12th grade).

2 : diplôme de premier cycle universitaire.

3 : diplôme de second cycle universitaire (maîtrise).

Nous estimons donc une régression logistique en remplaçant motheduc par motheducc et nous indiquons à Stata que la variable est catégorielle avec le préfixe i.

```
logit smokes i.motheducc i.white c.lfaminc

Iteration 0:   log likelihood =  -593.2711
Iteration 1:   log likelihood =  -551.3655
Iteration 2:   log likelihood = -545.62805
Iteration 3:   log likelihood = -545.49329
Iteration 4:   log likelihood = -545.49321
Iteration 5:   log likelihood = -545.49321

Logistic régression                               Number of obs   =       1388
                                                  LR chi2(4)      =      95.56
                                                  Prob > chi2     =     0.0000
Log likelihood = -545.49321                       Pseudo R2       =     0.0805

------------------------------------------------------------------------------
      smokes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   motheducc |
          2  |  -.6684841   .2082536    -3.21   0.001    -1.076654   -.2603146
          3  |  -2.206939   .3991534    -5.53   0.000    -2.989265   -1.424613
             |
     1.white |   .3865328   .1991701     1.94   0.052    -.0038333     .776899
     lfaminc |  -.2996932   .0851848    -3.52   0.000    -.4666524   -.1327339
       _cons |  -.7610281   .2370093    -3.21   0.001    -1.225558   -.2964984
------------------------------------------------------------------------------
```

Nous voyons que les modalités de la variable sont significatives et nous aimerions tester celles-ci de différentes façons. En effet, nous pourrions imaginer une hypothèse de recherche nous amenant à vérifier que l'impact des diplômes est bien différent.

Nous aimerions savoir si les coefficients de chaque modalité sont tous égaux avec la commande

```
contrast

contrast motheducc

Contrasts of marginal linear prédictions

Margins      : asbalanced

-------------------------------------------------
             |         df        chi2     P>chi2
-------------+-----------------------------------
   motheducc |          2       36.74     0.0000
-------------------------------------------------
```

Nous avons suffisamment d'évidence statistique pour rejeter l'hypothèse nulle d'égalité des coefficients.

Nous pouvons maintenant tester s'il existe une différence significative entre un diplôme de premier cycle et un diplôme du secondaire ainsi qu'un diplôme de deuxième cycle et un diplôme du secondaire.

```
contrast r.motheducc

Contrasts of marginal linear prédictions

Margins      : asbalanced

-------------------------------------------------
             |         df        chi2     P>chi2
-------------+-----------------------------------
   motheducc |
   (2 vs 1)  |          1       10.30     0.0013
   (3 vs 1)  |          1       30.57     0.0000
      Joint  |          2       36.74     0.0000
-------------------------------------------------

------------------------------------------------------------
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
   motheducc |
   (2 vs 1)  |  -.6684841   .2082536     -1.076654   -.2603146
   (3 vs 1)  |  -2.206939   .3991534     -2.989265   -1.424613
------------------------------------------------------------
```

Nous pouvons ici rejeter l'hypothèse que l'effet d'un diplôme du secondaire est égal à l'effet d'un diplôme de premier cycle ainsi que que l'effet d'un diplôme du secondaire est égal à l'effet d'un diplôme de deuxième cycle pour la mère.

Nous pouvons aussi tester la différence entre les différentes modalités adjacentes de la catégorie avec le préfixe `ar.` devant la variable.

```
contrast ar.motheducc

Contrasts of marginal linear prédictions

Margins      : asbalanced

-------------------------------------------------
             |         df        chi2     P>chi2
-------------+-----------------------------------
   motheducc |
   (2 vs 1)  |          1       10.30     0.0013
   (3 vs 2)  |          1       13.14     0.0003
      Joint  |          2       36.74     0.0000
-------------------------------------------------

------------------------------------------------------------
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
   motheducc |
   (2 vs 1)  |  -.6684841   .2082536     -1.076654   -.2603146
   (3 vs 2)  |  -1.538455   .4244366     -2.370335   -.7065742
```

```
------------------------------------------------------------
```

Dans ce tableau, nous n'avons qu'une valeur d'intérêt comme les deux autres ont déjà été reportées dans les autres tableaux. Nous pouvons rejeter l'hypothèse que l'effet d'avoir un diplôme de premier cycle est le même que l'effet d'avoir un diplôme de deuxième cycle pour la mère.

Une autre manière de tester les hypothèses précédentes en un seul test est d'utiliser la commande `pwcompare`. Il faut par contre spécifier à Stata plus d'options pour que les statistiques et les estimations voulues soient reportées dans le tableau. Nous ajoutons donc `cimargins` pour avoir les intervalles de confiances, `effects` pour avoir les estimations et nous demandons à Stata de classer les variables par ordre d'importance. L'option `mcompare` nous permet d'ajouter une option pour que les niveaux de significativité soient ajustés avec la méthode de Bonferonni qui prend en compte le nombre de comparaisons effectuées simultanément sur le même échantillon.

```
pwcompare i.motheduc, mcompare(bonferroni) cimargins effects sort

Pairwise comparisons of marginal linear prédictions

Margins      : asbalanced

------------------------------------------------------------
             |                                Unadjusted
             |     Margin   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
   motheducc |
          3  |  -2.774701   .4758266     -3.707304   -1.842098
          2  |  -1.236246   .3195343     -1.862522   -.6099702
          1  |  -.5677617    .228381      -1.01538   -.1201432
------------------------------------------------------------

---------------------------
             |    Number of
             |  Comparisons
-------------+-------------
smokes       |
   motheducc |            3
---------------------------

------------------------------------------------------------------------------
             |                            Bonferroni           Bonferroni
             |   Contrast   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
smokes       |
   motheducc |
     3 vs 1  |  -2.206939   .3991534    -5.53   0.000    -3.162504   -1.251374
     3 vs 2  |  -1.538455   .4244366    -3.62   0.001    -2.554547    -.522362
     2 vs 1  |  -.6684841   .2082536    -3.21   0.004    -1.167039   -.1699292
------------------------------------------------------------------------------
```

Avec les ajustements de Bonferonni, nous avons les mêmes conclusions que précédemment, c'est-à-dire que l'effet de chaque modalité est statistiquement différent lorsque celles-ci sont comparées deux à deux.

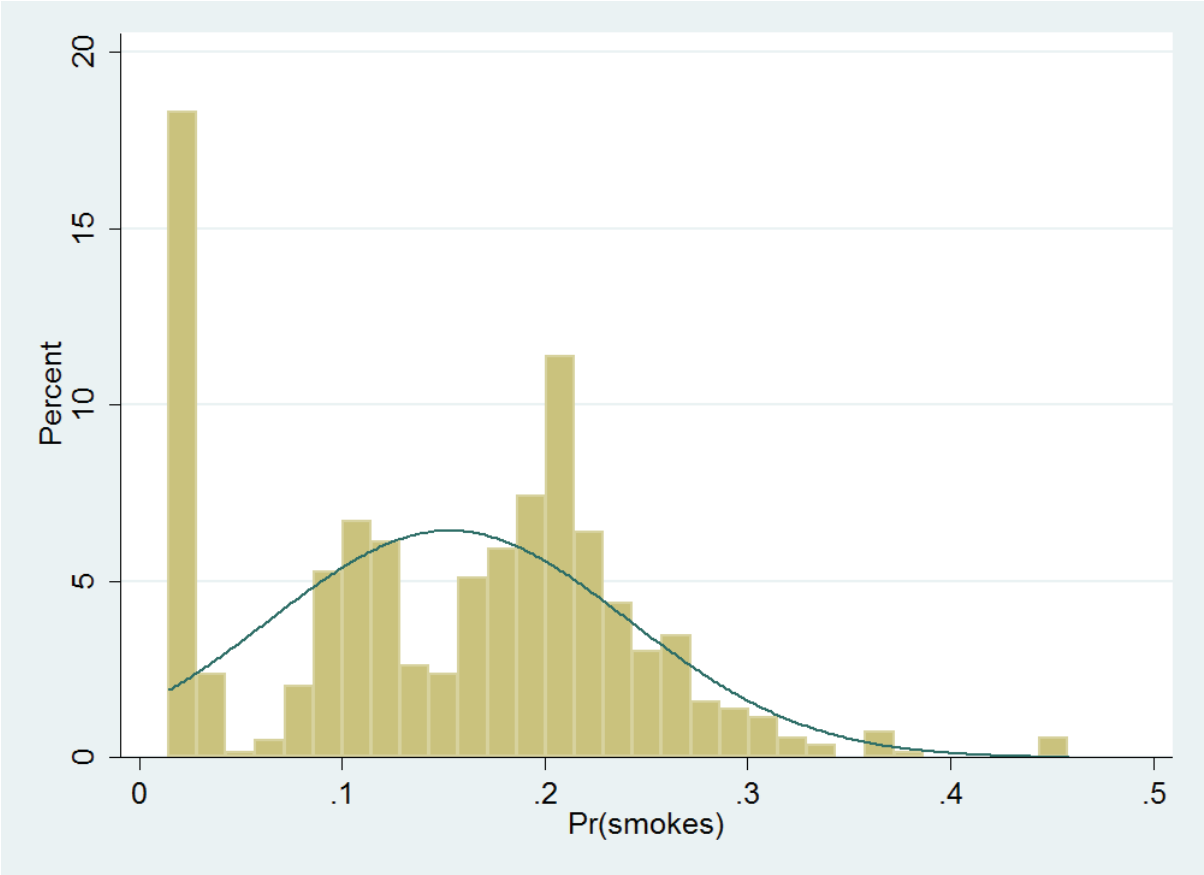
Prédictions

Sans spécifier d'option, nous pouvons prédire les probabilités que l'individu ait fumé pendant sa grossesse grâce à la commande `predict`.

```
predict plogit if e(sample)
(option pr assumed; Pr(smokes))
```

Nous pouvons représenter graphiquement les probabilités prédites.

```
histogram plogit, percent normal
(bin=31, start=.01449997, width=.01431953)
```



Nous remarquons que la plupart des probabilités prédites sont inférieures à 0.5, expliquant ainsi le taux de mauvaise prédiction concernant les femmes ayant fumé pendant leur grossesse.

3.2.5 La commande probit

Description

La commande `probit` dans Stata sert à estimer un modèle probit et utilise la méthode du maximum de vraisemblance pour estimer les paramètres ($\hat{\beta}$) du modèle. Stata traite les valeurs de la variable dépendante telles que

$$y = \begin{cases} 1 & y > 0 \\ 0 & y = 0 \end{cases}$$

Les variables négatives seront considérées comme manquantes. Il faut donc faire attention à avoir une variable codée qui prend strictement la valeur 0 ou 1.

Syntaxe

`probit [depvar] [indepvars] [if] [in] [, options]`

où `depvar` est la variable dépendante
`[indepvars]` est la variable indépendante
`[if]` et `[in]` sont des contraintes dans l'échantillon
`[weight]` assigne des poids aux observations
`[, options]` présente des options supplémentaires dans la régression

Menu

Statistics > Binary outcomes > Probit régression

Option

Erreur type/Robust

`vce(typevce)` : le type vce peut être `robust`, `cluster`(*variablecluster*). Pour plus d'informations, le lecteur peut se référer à la documentation de Stata.

3.2.6 Les commandes postestimation après la commande Probit

`test [varexo]`: nous pouvons tester des restrictions linéaires comme dans un modèle de régression linéaire (par exemple : $\beta_1 = \beta_2$).

`estat classification` : nous donne accès à des statistiques de classement basées sur les $\hat{y}_i$.

`estat gof` : cette commande performe le test de Hosmer-Lemeshow sur la qualité du modèle. Il n'est pas très puissant, car il nous informe seulement sur le fait que le modèle est acceptable.

`lroc` : cette commande calcule l'aire sous la courbe ROC et fournit un graphique de la courbe

`lsens` : trace un graphique mettant en relation la sensitivité et la spécificité par rapport au point de coupure choisi.

`estat ic` : nous donne les valeurs des critères d'information (AIC et BIC).

`lincom [opérationlinéaire], or` : cette option nous permet de calculer les ratios relatifs de risque pour un type particulier d'individu en spécifiant les valeurs auxquelles ceux-ci veulent être évalués. Pour plus de renseignements sur cette commande, se référer à la page 911 du manuel du logiciel Stata.

`linktest` : cette commande teste globalement la bonne spécification du modèle. Une régression est effectuée sur les valeurs prédites $\hat{y}_i$ et les valeurs prédites au carré $\hat{y}_i{}^2$. Le modèle est bien spécifié si le carré des valeurs prédites ne contient pas d'information pour expliquer y. Autrement dit, le coefficient devant la variable _hatsq doit être non significatif.

`lrtest [modèlecomplet] [modèlecontraint][, stats]` : cette commande performe un test de ratio de vraisemblance (likelihood ratio test) entre un modèle complet et un modèle contraint. Ce test nous renseigne sur la significativité des coefficients non inclus dans le modèle contraint et est similaire au test-F joint qui est calculé avec la commande test. Le modèle contraint est donc un sous modèle du modèle complet. Ce test peut être utilisé pour un test de Chow.

`margins, [dydx(varexos)] [at(varexo1=(valdébut(incrément)valfin)varexo2=(valdébut(incrément)valfin)]` :

cette commande nous permet de calculer les effets marginaux des variables spécifiées et les prédictions du modèle. L'option dydx(.) est nécessaire pour calculer les effets marginaux de variables continues, l'option at nous servira pour pouvoir utiliser la commande `marginsplot` afin d'avoir un aperçu graphique de ces effets.

`marginsplot [varexos]` : cette commande doit être utilisée après marginsplot pour avoir un aperçu graphique des effets marginaux sur les probabilités que la variable prenne la valeur 1.

`prchange [varlist], [help]` : calcule l'impact d'une variation discrète ou marginale des variables spécifiées pour les modèles catégoriels et de comptage. Cette commande ne fonctionne pas avec les variables considérées comme des facteurs (*factor variables*).

`contrast [varlist]` : permets de tester si la moyenne des effets est égale entre toutes les modalités d'une variable ordinale incluse dans la régression (x).

`contrast r.[varlist]` : permets de tester si les moyennes des effets sont égales entre les différentes modalités d'une variable ordinale quand elles sont comparées à la catégorie de référence.

`contrast ar.[varlist]` : permets de tester si la moyenne d'un groupe est différente du groupe qui le précède comme la variable est ordinale.

Pour tester des différences plus complexes, le lecteur peut se référer à la page 314 du manuel du logiciel Stata.

Syntaxe pour predict

`predict p if e(sample)` : prédit la probabilité que la variable dépendante prenne la valeur 1. La restriction if e(sample) est ajoutée afin de spécifier que les probabilités soient prédites pour les observations n'ayant aucune valeur manquante.

3.2.7 Exemple

Nous allons maintenant présenter un exemple pour le modèle probit. Les résultats ne seront pas interprétés comme le choix entre l'une des deux distributions d'erreur ne change quasiment pas les résultats obtenus par rapport au modèle logit. Les lecteurs peuvent se référer à l'exemple précédent pour l'interprétation des résultats.

La commande `probit` est utilisée avec les mêmes variables que le premier modèle estimé avec la commande `logit`. La seule option non utilisée dans cet exemple est l'option `or` qui n'est permise qu'avec une forme logistique étant donné la transformation effectuée et sa signification spécifique au modèle.

```
probit smokes c.motheduc i.white c.lfaminc

Iteration 0:   log likelihood = -593.10529
Iteration 1:   log likelihood = -547.59879
Iteration 2:   log likelihood = -546.77099
Iteration 3:   log likelihood = -546.76991
Iteration 4:   log likelihood = -546.76991
```

```
Probit régression                                 Number of obs   =       1387
                                                  LR chi2(3)      =      92.67
                                                  Prob > chi2     =     0.0000
Log likelihood = -546.76991                       Pseudo R2       =     0.0781

------------------------------------------------------------------------------
      smokes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    motheduc |  -.1450599   .0207899    -6.98   0.000    -.1858074   -.1043124
     1.white |   .1896765   .1098805     1.73   0.084    -.0256853    .4050383
     lfaminc |  -.1669109   .0498894    -3.35   0.001    -.2646923   -.0691296
       _cons |   1.126276   .2504611     4.50   0.000     .6353817    1.617171
------------------------------------------------------------------------------
```

Le signe des coefficients ne diffère pas du signe des coefficients estimations avec le modèle logit. Les p-valeur ne diffèrent que très peu de celles calculées avec le modèle logit.

Nous sauvegardons les estimations du modèle sous le nom m3

```
estimates store m3
```

Nous pouvons aussi comparer les mesures d'adéquation des deux modèles avec la commande `fistat` en spécifiant l'option `using(nomdumodèle)` après les avoir sauvegardés. Nous ajoutons l'option `force` pour pouvoir comparer deux types d'estimation différentes (logit vs probit).

```
fitstat, using(m2) force

Measures of Fit for probit of smokes

Warning: Current model estimated by probit, but saved model estimated by logit

Warning: N's do not match.

                              Current             Saved        Difference
Model:                         probit             logit
N:                               1388              1387                 1
Log-Lik Intercept Only        -593.271          -593.105            -0.166
Log-Lik Full Model            -545.349          -548.734             3.385
D                             1090.698(1381)    1097.467(1382)       6.769(1)
LR                              95.844(4)         88.743(3)          7.101(1)
Prob > LR                        0.000             0.000             0.008
McFadden's R2                    0.081             0.075             0.006
McFadden's Adj R2                0.069             0.066             0.003
ML (Cox-Snell) R2                0.067             0.062             0.005
Cragg-Uhler(Nagelkerke) R2       0.116             0.108             0.008
McKelvey & Zavoina's R2          0.196             0.141             0.055
Efron's R2                       0.060             0.053             0.007
Variance of y*                   1.243             3.830            -2.587
Variance of error                1.000             3.290            -2.290
Count R2                         0.847             0.843             0.004
Adj Count R2                     0.000            -0.028             0.028
AIC                              0.796             0.798            -0.003
AIC*n                         1104.698          1107.467            -2.769
BIC                          -8901.692         -8901.162            -0.530
BIC'                            -66.902           -67.039             0.137
BIC used by Stata             1126.876          1126.407             0.469
AIC used by Stata             1100.698          1105.467            -4.769

Note: p-valeur  for difference in LR is only valid if models are nested.
```

Comme les mesures de R^2 sont plus élevées pour le modèle probit, nous pouvons remarquer que le modèle probit performe un peu mieux que le modèle logit. Nous voyons aussi que les critères d'information, le AIC et le BIC sont moins élevés.

3.3 Modèles logit et probit multinomiaux

3.3.1 Rappel théorique

Modélisation

Le modèle logit multinomial est nécessaire quand le nombre de modalités que prend la variable dépendante est supérieur à deux. Nous sommes ici dans le cadre d'une régression logistique polytomique qui est une généralisation de la régression logistique binaire.

En effet, si nous voulons modéliser une variable qui prend plusieurs modalités et que celles-ci sont non ordonnées, c'est-à-dire que nous ne pouvons pas les classer en ordre de grandeur, nous pouvons considérer les modèles logit et probit multinomiaux.

Nous pouvons construire notre régression logistique multinomiale en modélisant $(K-1)$ rapports de probabilités comme nous prenons une modalité comme référence. Nous exprimerons les estimateurs logit par rapport à celle-ci. Nous pouvons constater que nous avons 2-1=1 rapport de probabilités à modéliser pour le modèle logit comme la catégorie de base est 0.

Nous voulons donc savoir quelle est la probabilité qu'une observation soit dans le groupe j.

Pour un individu α, les probabilités vérifient

$$\sum_{j=1}^{J} P(y_\alpha = j) = 1$$

Autrement dit, la somme des probabilités d'appartenance à chaque catégorie estimées doit être égale à 1.

Le modèle logistique peut être écrit:

$$P(y_i = j) = \frac{e^{\beta'_j x_i}}{\sum_{k=0}^{K} e^{\beta'_k x_i}}, \quad j = 0,1,\dots,K$$

Où les x_i sont nos variables indépendantes et les $j = 0,1,\dots,J$ représentent l'ensemble de catégories que prend la variable dépendante.

Étant donné la forme du modèle, il faut ajouter une restriction afin que le modèle soit identifié, car il y a plus qu'une solution pour les paramètres du système d'équations engendré par la

modélisation. Il faut donc contraindre les paramètres associés à une modalité égale à 0 afin de pouvoir exprimer les paramètres des autres modalités par rapport à celle-ci. Ainsi, les paramètres pourront numériquement être estimations.

C'est par la méthode du maximum de vraisemblance que les estimateurs sont calculés. Le lecteur intéressé peut aisément trouver de l'information dans les ouvrages référencés à la fin de ce manuel.

Nous ne dériverons pas explicitement le modèle probit multinomial qui est plus fastidieux à reproduire que le modèle logit multinomial. Globalement, les deux distributions sont très semblables et les effets marginaux diffèrent peu. Le choix de la distribution des erreurs est laissé à la discrétion du chercheur selon la nature des variables et selon la littérature disponible sur le sujet. De plus, il est important de noter que modéliser correctement plus de quatre catégories avec un modèle probit multinomial peut s'avérer complexe étant donné les méthodes d'estimation des paramètres. Il est donc conseillé d'utiliser un modèle logistique multinomial pour modéliser une variable avec de nombreuses modalités.

Rapports de cotes

Dans les modèles logistiques multinomiaux, les rapports de cotes (aussi appelés ratios relatifs de risque) sont aussi des manières de pouvoir décrire les relations entre la variable dépendante et les variables de contrôle.

$$RRR = e^{\beta_i}$$

Essentiellement, pour chaque variable indépendante, ces ratios indiquent de quelle façon le risque (ou la probabilité) de réalisation de l'un des évènements donnés de la variable dépendante est affectée par rapport au risque de réalisation de l'évènement de base. Un RRR supérieur à 1 signifie que ce risque relatif augmente avec une variation de la variable explicative. Autrement dit, la réalisation de l'évènement donné devient plus probable par rapport à l'évènement de base. L'inverse est vrai pour un RRR inférieur à un.

Il est important de noter que le modèle logit multinomial s'appuie sur l'hypothèse que les alternatives sont indépendantes. Cette hypothèse est plus connue sous le nom d'*Independance of Irrelevant Alternatives* et se traduit par une constance des ratios relatifs de risque si une nouvelle alternative est ajoutée à l'ensemble de choix. Pour de plus amples explications, le

lecteur peut se rapporter à l'article de McFadden (1974) et son exemple du choix d'autobus bleu et d'autobus rouge.

Taux de classification (*Count* $\mathbf{R^2}$) :

Une autre mesure de qualité consiste à classer chaque observation selon sa probabilité d'appartenir à une modalité versus une autre. Le classement prédit pour un individu α est alors

$$\hat{y}_\alpha = y_{\alpha j} \quad \Leftrightarrow \quad y_{\alpha j} = arg\,\max_j \left[\hat{\pi}_j(\alpha)\right]$$

Où $\hat{\pi}_j$ est la probabilité prédite du modèle pour la modalité j de la variable dépendante. Nous allons donc classer dans une modalité j l'individu qui aura la probabilité d'appartenir à la modalité j comme probabilité maximale parmi les J possibilités disponibles. À partir de ce classement, il est possible de générer une matrice de confusion(Kohavi & Provost, 1998). Il nous faudra premièrement effectuer le comptage du nombre de personnes appartenant à la catégorie y_j et qui on été affectés à $\hat{y}_k$ que l'on appellera n_{kj} où $|n_{kj}|$ est son cardinal

$$\left|n_{kj}\right| = \{z, y(z) = y_j \; et \; \hat{y}(z) = \hat{y}_k, \forall k = j\}$$

Le taux d'erreur est obtenu en effectuant le rapport des observations « mal classées » sur le nombre n d'observations totales :

$$\varepsilon = \frac{\sum_k \sum_{j \neq k} n_{kj}}{n} = 1 - \frac{\sum_k n_{kk}}{n}$$

La matrice de confusion compare les fréquences des valeurs de la variable dépendante dans l'échantillon avec les fréquences des valeurs prédites par le modèle. Similairement au test du ratio de maximum de vraisemblance, nous allons pouvoir comparer le taux d'erreur du modèle trivial au taux d'erreur du modèle alternatif afin d'avoir une base de comparaison. Il est possible de dériver un pseudo-R^2 basé sur ces taux:

$$R_\varepsilon^2 = 1 - \frac{\varepsilon_M}{\varepsilon_T}$$

Ici, $R_\varepsilon^2 = 1$ si le modèle est « parfait » et $R_\varepsilon^2 = 0$ si le modèle prédit aussi bien que le modèle trivial.

3.3.2 La commande mlogit

Description

La commande mlogit dans Stata sert à estimer un modèle logistique multinomial ou régression logistique polytomique et utilise la méthode du maximum de vraisemblance pour trouver les paramètres ($\hat{\beta}$) du modèle. Ce type de modèle doit être choisi lorsque la variable dépendante est discrète et composée de modalités qui ne peuvent pas être naturellement ordonnées. Stata traite les valeurs de la variable dépendante telles qu'elles seront spécifiées par l'utilisateur. Ainsi, il est important de vérifier le codage de la variable dépendante en utilisant la commande `summarize`. La première catégorie est généralement codée 0. Les variables négatives seront considérées comme manquantes. Il faut donc s'assurer d'avoir une variable codée strictement positivement 0, 1, 2, 3, etc. Lorsque la variable dépendante ne prend que deux modalités, les paramètres estimations sont les mêmes qu'en utilisant la commande logistic ou logit.

Syntaxe

```
mlogit [depvar] [indepvars] [if] [in] [weight] [, options]
```

où `depvar` est la variable dépendante
`[indepvars]` est la variable indépendante
`[if]` et `[in]` sont des contraintes dans l'échantillon
`[weight]` assigne des poids aux observations
`[, options]` présente des options supplémentaires dans la régression

Menu

Statistics > Categorical outcomes > Multinomial logistic régression

Option

Erreur type/Robust

`vce(typevce)` : le type vce peut être `robust`, `cluster`(*variablecluster*). Pour plus d'informations, le lecteur peut se référer à la documentation Stata.

Résultats

`baseoutcome(#)` : permets d'indiquer quelle est la catégorie qui doit être prise comme catégorie de référence (catégorie de base). Par défaut, c'est la catégorie qui a la fréquence la plus élevée qui est choisie par le logiciel comme catégorie de référence.

`rrr` : rapporte les ratios relatifs de risque (rapport de cotes), c'est-à-dire les e^{β}. Les erreurs standards et les intervalles de confiance sont similairement transformés. rrr peut être performé lors de l'estimation ou simplement après l'estimation pour obtenir un nouveau tableau.

3.3.3 Les commandes postestimation après la commande mlogit

`test [[label ou valeur de la modalité]] [varexo]`: test de Wald avec lequel nous pouvons tester des restrictions linéaires (par exemple $\beta_1 = \beta_2$ ou $\beta_1 = 0$).

`lrtest [modèlecomplet] [modelcontraint][, stats]` : cette commande performe un test de ratio de vraisemblance (*likelihood ratio test*) entre un modèle complet et un modèle contraint. Ce test nous renseigne sur la significativité des coefficients non inclus dans le modèle contraint et est similaire au test-F joint qui est calculé avec la commande test. Le modèle contraint est donc un sous modèle du modèle complet. Ce test peut être utilisé pour un test de Chow advenant où nous voulons savoir si les coefficients du modèle sont différents entre plusieurs tranches de l'échantillon.

`mlogtest, [all]` : calcule une grande variété de tests pour les modèles logistiques multinomiaux. L'option all permet de calculer tous les tests possibles soit un *likelihood ratio test* où nous testons l'hypothèse nulle que les coefficients associés avec les variables testées sont 0 dans toutes les équations; un test de Wald testant la même hypothèse nulle; un test de Hausman vérifiant l'hypothèse IIA; un test de Small-Hsiao vérifiant l'hypothèse IIA; un test de Wald où nous testons l'hypothèse nulle que deux alternatives peuvent être combinées et un *likelihood ratio test* où nous testons la même hypothèse nulle.

`listcoef [, const help]`: donne une liste des coefficients pour le dernier modèle estimé. L'option const inclut la constante dans le tableau et l'option help ajoute une légende des variables diponibles dans le tableau. Pour plus de renseignement concernant les options, l'utilisateur peut se référer à l'aide fournie par le module en utilisant la commande help listcoef.

`fitstat` : indices de qualité du modèle (fit ou adéquation).

```
margins,[dydx(varexos)][at(varexo1=(valdébut(incrément)valfin)varexo2=(valdébut(incrément)valfin)
] [predict(outcome(label ou valeur de la modalité))]:
```

cette commande nous permet de calculer les effets marginaux des variables spécifiées et les prédictions du modèle. L'option `dydx(.)` est nécessaire pour calculer les effets marginaux de variables continues, l'option at nous servira à utiliser la commande `marginsplot` afin d'avoir un aperçu graphique de ces effets. Dans le cas d'un modèle logistique multinomial, il est aussi nécessaire de spécifier pour quelle catégorie nous voulons les effets marginaux.

`marginsplot [varexos]` : cette commande doit être utilisée après margins pour avoir un aperçu graphique des effets marginaux sur les probabilités.

`prchange [varlist], [outcome(label ou valeur de la modalité)] [help]`: calcule l'impact d'une variation discrète ou marginale des variables spécifiées pour les modèles catégoriels et de comptage.

`pwcompare [varexo], [cimargins] [eq(label ou valeur de l'équation)]` : permets de comparer l'impact de chaque modalité d'une variable catégorielle entre elles et de tester si elles ont un effet significativement différent.

`contrast [varexo], [eq(label ou valeur de l'équation)]` : permets de tester si la moyenne des effets est égale entre toutes les modalités d'une variable catégorielle incluse dans la régression (x).

`contrast r.[varexo], [eq(label ou valeur de l'équation)]` : permets de tester si la moyenne des effets est égale entre les différentes modalités d'une variable catégorielle lorsqu'elles sont comparées à la catégorie de référence.

`contrast ar.[varexo] , [eq(label ou valeur de l'équation)]`: permets de tester si le coefficient d'une modalité est différent du coefficient de la modalité qui le précède.

`predict p if e(sample) [, outcome(label ou valeur de la modalité)]` : prédit la probabilité que la variable dépendante prenne la valeur 1. La restriction `if e(sample)` est ajoutée afin de spécifier que les probabilités soient prédites pour les observations n'ayant aucune valeur manquante. Dans ce cas-ci, il faut aussi spécifier la valeur de la modalité pour laquelle les probabilités sont prédites.

3.3.4 Exemple

Pour cette application, nous utiliserons le fichier KEANE.dta disponible sur le site de Stata. Ce fichier est composé de données de panel concernant l'éducation et l'emploi d'un échantillon d'homme de 1981 à 1987. Nous n'utiliserons que les données pour l'année 1987.

```
use "http://www.Stata.com/data/jwooldridge/eacsap/keane.dta", clear
```

Nous nous intéressons à estimer un modèle où la variable dépendante prend trois valeurs représentant l'activité d'un individu de l'échantillon. Nous voulons modéliser la probabilité que celui-ci soit 1) sans emploi et pas aux études, 2) un étudiant et sans emploi ou 3) un travailleur pas aux études.

Parmi toutes les variables, nous sélectionnons une variable concernant l'éducation (educ), une variable concernant l'expérience (exper) et une variable binaire concernant l'ethnie de l'individu (black). Nous nous attendons à ce que l'éducation ait un effet négatif sur la probabilité de ne pas travailler et de ne pas étudier. De plus, l'expérience devrait avoir tendance à baisser la probabilité de ne pas travailler et de ne pas étudier. En effet, une variation positive de ces deux variables devrait amener une augmentation du salaire. Celui-ci devrait également avoir un impact sur la décision d'un individu d'entrer sur le marché du travail.

Description des données

Nous allons premièrement explorer les données disponibles et celles qui nous intéressent

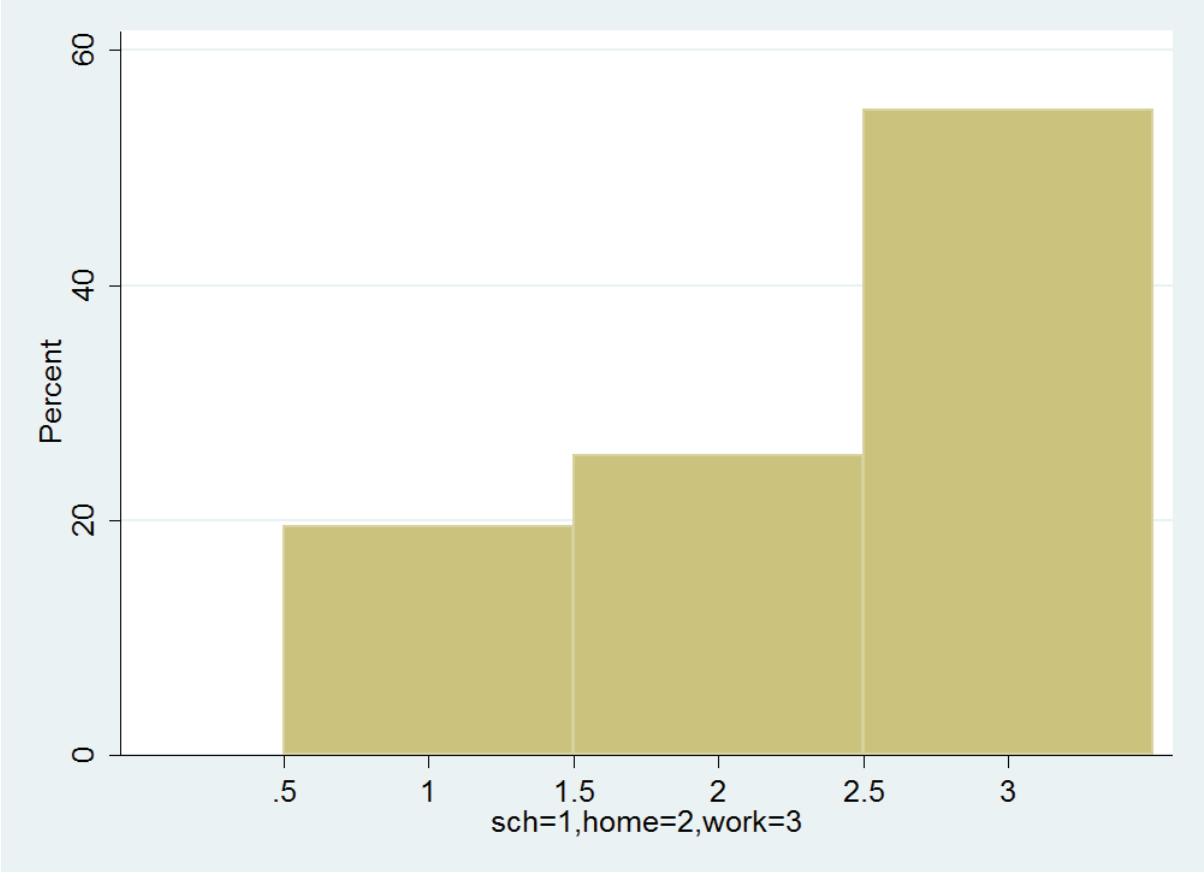
```
describe status educ educc exper expersq black

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------------------
status          byte   %9.0g                  sch=1,home=2,work=3
educ            byte   %9.0g                  years of schooling
educc            float  %9.0g                   Éducation, 1: entre 7 et 11, 2: 11 et 12, 3: 12 et
                                               13, 4, supérieur à 13
exper           byte   %9.0g                  expwc + expbc + expser
expersq         byte   %9.0g                  exper^2
black           byte   %9.0g                  =1 if black
```

```
summarize status educ educc exper expersq black

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      status |     12665    2.354205    .7869674          1          3
        educ |     12723    12.06437    2.092468          7         19
```

```
       educc |     12723    2.672876    1.028845          1          4
       exper |     12723    1.651969    1.930829          0         10
     expersq |     12723    6.456811    11.48361          0        100
-------------+---------------------------------------------------------
       black |     12723    .3767193    .4845826          0          1

inspect status

status:  sch=1,home=2,work=3                    Number of Observations
----------------------------                 ------------------------------
                                             Total   Integers   Nonintegers
|          #                 Negative           -          -          -
|          #                 Zero               -          -          -
|          #                 Positive       12665      12665          -
|          #                                -----      -----      -----
|  #   #   #                 Total          12665      12665          -
|  #   #   #                 Missing           58
+----------------------                      -----
1                     3                      12723
   (3 unique values)

inspect educc

educc:  Éducation, 1: entre 7 et 11, 2: 11      Number of Observations
------------------------------------------- ------------------------------
                                             Total   Integers   Nonintegers
|          #                 Negative           -          -          -
|          #                 Zero               -          -          -
|          #                 Positive       12723      12723          -
|          #                                -----      -----      -----
|  #       #   #             Total          12723      12723          -
|  #   #   #   #             Missing            -
+----------------------                      -----
1                     4                      12723
   (4 unique values)
```

Les variables ont des valeurs minimum et maximum plausibles. Nous avons aussi créé une variable catégorielle pour l'éducation afin de pouvoir effectuer une certaine partie de l'exemple.

```
histogram status, discrete percent
(start=1, width=1)
```




```
tab1 status educ educc exper expersq black

-> tabulation of status

sch=1,home= |
   2,work=3 |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      2,473       19.53       19.53
          2 |      3,233       25.53       45.05
          3 |      6,959       54.95      100.00
------------+-----------------------------------
      Total |     12,665      100.00

-> tabulation of educ

   years of |
  schooling |      Freq.     Percent        Cum.
------------+-----------------------------------
          7 |        135        1.06        1.06
          8 |        452        3.55        4.61
          9 |        803        6.31       10.93
         10 |      1,167        9.17       20.10
         11 |      1,824       14.34       34.43
         12 |      4,392       34.52       68.95
         13 |      1,174        9.23       78.18
         14 |        993        7.80       85.99
         15 |        699        5.49       91.48
         16 |        880        6.92       98.40
         17 |        142        1.12       99.51
         18 |         52        0.41       99.92
         19 |         10        0.08      100.00
------------+-----------------------------------
      Total |     12,723      100.00
```

```
-> tabulation of educc

 Éducation, |
 1: entre 7 |
  et 11, 2: |
  11 et 12, |
   3: 12 et |
     13, 4, |
supérieur à |
         13 |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      2,557       20.10       20.10
          2 |      1,824       14.34       34.43
          3 |      5,566       43.75       78.18
          4 |      2,776       21.82      100.00
------------+-----------------------------------
      Total |     12,723      100.00

-> tabulation of exper

    expwc + |
    expbc + |
     expser |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      5,248       41.25       41.25
          1 |      2,184       17.17       58.41
          2 |      1,737       13.65       72.07
          3 |      1,282       10.08       82.14
          4 |        947        7.44       89.59
          5 |        628        4.94       94.52
          6 |        407        3.20       97.72
          7 |        200        1.57       99.29
          8 |         70        0.55       99.84
          9 |         16        0.13       99.97
         10 |          4        0.03      100.00
------------+-----------------------------------
      Total |     12,723      100.00

-> tabulation of expersq

    exper^2 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      5,248       41.25       41.25
          1 |      2,184       17.17       58.41
          4 |      1,737       13.65       72.07
          9 |      1,282       10.08       82.14
         16 |        947        7.44       89.59
         25 |        628        4.94       94.52
         36 |        407        3.20       97.72
         49 |        200        1.57       99.29
         64 |         70        0.55       99.84
         81 |         16        0.13       99.97
        100 |          4        0.03      100.00
------------+-----------------------------------
      Total |     12,723      100.00

-> tabulation of black

=1 if black |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      7,930       62.33       62.33
          1 |      4,793       37.67      100.00
------------+-----------------------------------
      Total |     12,723      100.00
```

Dans les tableaux de fréquences et les graphiques, nous pouvons remarquer que toutes les variables possèdent des distributions qui ne posant aucun problème si nous les incluons dans le modèle ou au niveau de leur interprétation.

Modélisation

Nous pouvons utiliser la commande `mlogit` pour ajuster un modèle logistique multinomial en faisant dépendre le statut de l'éducation{Wooldridge, 2009 #149}, de l'expérience et de l'ethnie avec une interaction concernant l'expérience pour que celle-ci ait un effet quadratique (équivalent à inclure l'expérience au carré). Nous demandons à Stata de considérer seulement l'année 1987 avec l'option `if year==87` pour ne pas avoir à traiter des effets à travers le temps.

```
mlogit status educ c.exper##c.exper black if year==87

Multinomial logistic régression                   Number of obs   =       1717
                                                  LR chi2(8)      =     583.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -907.85723                       Pseudo R2       =     0.2433

------------------------------------------------------------------------------
         status |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
1               |
           educ |   .3146573   .0651096     4.83   0.000     .1870448    .4422699
          exper |  -.8487367   .1569856    -5.41   0.000    -1.156423   -.5410507
                |
c.exper#c.exper |   .0773003   .0229217     3.37   0.001     .0323746    .1222261
                |
          black |  -.3113612   .2815339    -1.11   0.269    -.8631574     .240435
          _cons |  -5.543798   1.086409    -5.10   0.000    -7.673121   -3.414475
----------------+-------------------------------------------------------------
2               |
           educ |  -.3589739   .0319425   -11.24   0.000    -.4215801   -.2963677
          exper |  -.9549516   .0966353    -9.88   0.000    -1.144353   -.7655499
                |
c.exper#c.exper |   .0647852   .0127634     5.08   0.000     .0397693    .0898011
                |
          black |   .5016554   .1484078     3.38   0.001     .2107814    .7925295
          _cons |   4.734077    .461716    10.25   0.000      3.82913    5.639024
----------------+-------------------------------------------------------------
3               |  (base outcome)
------------------------------------------------------------------------------
```

Nous voyons que toutes les catégories de la variable dépendante sont comparées par rapport à la catégorie 3 (les personnes qui sont en emploi) comme nous n'avons pas spécifié l'option `outcome()`. C'est donc la catégorie avec la plus grande fréquence qui est choisie comme groupe de base.

Comme les coefficients ne sont pas directement interprétables, nous ne nous attarderons que très peu sur cette sortie. Les signes des coefficients nous donnent néanmoins une indication sur la direction de l'effet de chaque variable. Un niveau d'éducation plus élevé aurait un effet positif sur le fait d'appartenir à la catégorie actuellement aux études tandis que plus d'expérience aurait un effet de moins en moins négatif sur celle-ci, toute chose étant égale par ailleurs. Le fait d'être d'ethnie noire a un effet négatif sur la probabilité d'appartenir à la catégorie des

personnes étant actuellement aux études. Nous nous passerons l'interprétation de l'équation 2 comme il est plus intéressant d'interpréter les effets marginaux et les ratios relatifs de risque.

Excepté le coefficient de la variable black dans la première équation, nous pouvons aussi remarquer que les coefficients sont pratiquement tous significatifs.

Afin d'avoir un aperçu de l'estimation en changeant la catégorie de base, nous pouvons spécifier `baseoutcome(2)` qui nous donne de l'information sur l'effet des variables sur la probabilité de passer de la catégorie 2 à la catégorie 1 et de la catégorie 2 à la catégorie 3.

```
mlogit status educ c.exper##c.exper black if year==87, baseoutcome(2)

Multinomial logistic régression                   Number of obs   =       1717
                                                  LR chi2(8)      =     583.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -907.85723                       Pseudo R2       =     0.2433

------------------------------------------------------------------------------
         status |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
1               |
           educ |   .6736313   .0698999     9.64   0.000       .53663    .8106325
          exper |   .1062149    .173282     0.61   0.540    -.2334116    .4458414
                |
c.exper#c.exper |   .0125152   .0252291     0.50   0.620     -.036933    .0619633
                |
          black |  -.8130166   .3027231    -2.69   0.007    -1.406343   -.2196902
          _cons |  -10.27787   1.133336    -9.07   0.000    -12.49917   -8.056578
----------------+-------------------------------------------------------------
2               |  (base outcome)
----------------+-------------------------------------------------------------
3               |
           educ |   .3589739   .0319425    11.24   0.000     .2963677    .4215801
          exper |   .9549516   .0966353     9.88   0.000     .7655499    1.144353
                |
c.exper#c.exper |  -.0647852   .0127634    -5.08   0.000    -.0898011   -.0397693
                |
          black |  -.5016554   .1484078    -3.38   0.001    -.7925295   -.2107814
          _cons |  -4.734077    .461716   -10.25   0.000    -5.639024    -3.82913
```

Nous sauvegardons les estimations comme nous en aurons besoins pour différents tests.

```
estimates store m2
```

Tests des hypothèses de modélisation

Le modèle logistique multinomial repose sur l'hypothèse d'*Independance of Irrelevant Alternatives* et sur le fait que les catégories sont bien distinctes. Afin de tester ces hypothèses, nous pouvons utiliser la commande `mlogtest` et l'option `all` pour demander à Stata de performer tous ces tests.

```
mlogtest, all
```

Les deux premiers tests testent la nullité du coefficient de chaque variable simultanément dans les deux équations précédemment estimées premièrement avec un test du rapport de vraisemblance et deuxièmement avec un test de Wald.

```
**** Likelihood-ratio tests for independent variables (N=1717)

 Ho: All coefficients associated with given variable(s) are 0.

             |       chi2   df   P>chi2
-------------+------------------------
        educ |    207.970    2    0.000
       exper |    111.243    2    0.000
     expersq |     29.820    2    0.000
       black |     13.671    2    0.001
---------------------------------------

**** Wald tests for independent variables (N=1717)

 Ho: All coefficients associated with given variable(s) are 0.

             |       chi2   df   P>chi2
-------------+------------------------
        educ |    160.703    2    0.000
       exper |    114.905    2    0.000
     expersq |     34.376    2    0.000
       black |     13.695    2    0.001
---------------------------------------
```

Pour notre échantillon, nous voyons que les conclusions des deux tests ne diffèrent pas, les variables sont toutes statistiquement significatives.

Le deuxième test effectué est le test pour l'hypothèse IIA. Ici, nous testons si les modalités de la variable indépendante sont indépendantes les unes des autres en supprimant successivement une catégorie et en comparant pour le premier test les coefficients et le deuxième test la vraisemblance du modèle complet et du modèle ou une modalité a été omise. L'hypothèse testée est H_0: les rapports de chance d'une alternative sont indépendants des rapports de chance des autres alternatives.

```
**** Hausman tests of IIA assumption (N=1717)

 Ho: Odds(Outcome-J vs Outcome-K) are independent of other alternatives.

 Omitted |      chi2   df   P>chi2   evidence
---------+-----------------------------------
       1 |     1.031    5    0.960   for Ho
       2 |     1.609    5    0.900   for Ho
---------------------------------------------

**** Small-Hsiao tests of IIA assumption (N=1717)

 Ho: Odds(Outcome-J vs Outcome-K) are independent of other alternatives.

 Omitted |  lnL(full)  lnL(omit)    chi2   df   P>chi2   evidence
---------+-------------------------------------------------------
       1 |   -288.759   -286.424   4.670    5    0.457   for Ho
       2 |   -155.431   -150.790   9.283    5    0.098   for Ho
-----------------------------------------------------------------
```

Nous n'avons pas assez d'évidence statistique pour rejeter l'hypothèse nulle d'indépendance entre les alternatives. Celle-ci est donc respectée.

Le prochain test peut être intéressant lorsque nous voyons que de nombreux coefficients dans une équation ne sont pas significatifs. Il teste si nous pouvons combiner deux modalités de la variable dépendante ce qui est équivalent à tester si les coefficients d'une équation sont tous 0, autrement dit si seule la constante est significative.

```
**** Wald tests for combining alternatives (N=1717)

 Ho: All coefficients except intercepts associated with a given pair
     of alternatives are 0 (i.e., alternatives can be combined).

Alternatives tested|     chi2   df   P>chi2
-------------------+------------------------
       1-       2  |  130.257    4    0.000
       1-       3  |  136.281    4    0.000
       2-       3  |  300.727    4    0.000
--------------------------------------------

**** LR tests for combining alternatives (N=1717)

 Ho: All coefficients except intercepts associated with a given pair
     of alternatives are 0 (i.e., alternatives can be collapsed).

Alternatives tested|     chi2   df   P>chi2
-------------------+------------------------
       1-       2  |  199.865    4    0.000
       1-       3  |  167.411    4    0.000
       2-       3  |  412.042    4    0.000
--------------------------------------------
```

Ici, nous ne pouvons pas combiner de modalités, car nous rejetons l'hypothèse nulle dans tous les cas, et ce, avec les deux méthodes.

Nous estimons à nouveau maintenant le modèle en utilisant les opérateurs et les variables factorielles.

```
quietly mlogit status educ c.exper##c.exper i.black if year==87
```

Nous utilisons la commande `fitstat` pour avoir accès aux mesures d'adéquation du modèle et nous sauvegardons les résultats sous l'appellation m2.

Mesures d'adéquation

```
fitstat, saving(m2)

Measures of Fit for mlogit of status

Log-Lik Intercept Only:      -1199.718   Log-Lik Full Model:           -907.857
D(1699):                      1815.714   LR(8):                         583.722
                                         Prob > LR:                       0.000
McFadden's R2:                   0.243   McFadden's Adj R2:               0.228
ML (Cox-Snell) R2:               0.288   Cragg-Uhler(Nagelkerke) R2:      0.383
```

```
Count R2:                        0.796   Adj Count R2:                   0.186
AIC:                             1.078   AIC*n:                       1851.714
BIC:                        -10839.005   BIC':                        -524.135
BIC used by Stata:            1890.198   AIC used by Stata:           1835.714

(Indices saved in matrix fs_m2)
```

Ces mesures doivent être comparées lors de la modélisation et le choix des variables pour avoir des indications sur la qualité du modèle en termes d'explicativité de la variable dépendante.

Tests d'inférence

Nous testons des contraintes linaires ou nous restreignons black et educ à 0 dans les deux équations

```
test black educ

 ( 1)  [1]black = 0
 ( 2)  [2]o.black = 0
 ( 3)  [3]black = 0
 ( 4)  [1]educ = 0
 ( 5)  [2]o.educ = 0
 ( 6)  [3]educ = 0
       Constraint 2 dropped
       Constraint 5 dropped

           chi2(  4) =  198.48
         Prob > chi2 =    0.0000
```

Les coefficients des deux variables sont clairement statistiquement différents de 0.

Nous pouvons aussi effectuer un test du rapport de vraisemblance concernant les mêmes variables. La première étape est d'estimer le modèle contraint.

```
mlogit status c.exper##c.exper if year==87

Multinomial logistic régression                   Number of obs   =       1717
                                                  LR chi2(4)      =     305.71
                                                  Prob > chi2     =     0.0000
Log likelihood = -1046.8633                       Pseudo R2       =     0.1274

------------------------------------------------------------------------------
         status |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
1               |
          exper |   -1.04547   .1480985    -7.06   0.000    -1.335737   -.7552018
                |
c.exper#c.exper |   .0771984   .0225993     3.42   0.001     .0329047    .1214922
                |
          _cons |  -.6668683   .1741553    -3.83   0.000    -1.008206   -.3255301
----------------+-------------------------------------------------------------
2               |
          exper |  -.8288785   .0885893    -9.36   0.000     -1.00251   -.6552467
                |
c.exper#c.exper |   .0618822   .0122411     5.06   0.000     .0378901    .0858743
                |
          _cons |   .2651945   .1277839     2.08   0.038     .0147427    .5156463
----------------+-------------------------------------------------------------
3               |  (base outcome)
------------------------------------------------------------------------------
```

Nous sauvegardons les estimations sous le nom mtest

```
estimates store mtest
```

Nous effectuons le test du rapport de vraisemblance avec la commande `lrtest` suivi du modèle complet et du modèle contraint.

```
lrtest m2 mtest

Likelihood-ratio test                                 LR chi2(4)  =    278.01
(Assumption: mtest nested in m2)                      Prob > chi2 =    0.0000
```

Encore une fois, nous voyons que les coefficients des variables sont statistiquement significatifs dans les deux équations.

Prédictions

Cette partie concerne la prédiction des probabilités et le classement des individus avec l'aide du modèle. Nous commençons par prédire la probabilité d'appartenance de chaque individu aux catégories 1, 2 ou 3.

```
predict p1 if e(sample), outcome(1)
(option pr assumed; predicted probability)
(11006 missing values generated)

predict p2 if e(sample), outcome(2)
(option pr assumed; predicted probability)
(11006 missing values generated)

predict p3 if e(sample), outcome(3)
(option pr assumed; predicted probability)
(11006 missing values generated)
```

Nous créons ensuite des variables dichotomiques définies comme suit

$$p1best = \begin{cases} 1 & si\, p1 > p2\ et\ p1 > p3 \\ 0 & sinon \end{cases}$$

$$p2best = \begin{cases} 1 & si\, p2 > p1\ et\ p2 > p3 \\ 0 & sinon \end{cases}$$

$$p3best = \begin{cases} 1 & si\, p3 > p2\ et\ p3 > p1 \\ 0 & sinon \end{cases}$$

```
gen p1best if year==87 = ( p1>p2 & p1>p3 )
(10985 missing values generated)

replace  p1best=. if p1==.
(21 real changes made, 21 to missing)

gen p2best= ( p2>p1 & p2>p3 )

replace  p2best=. if p2==.
(11006 real changes made, 11006 to missing)

gen p3best= ( p3>p1 & p3>p2 )
```

```
replace  p3best=. if p3==.
(11006 real changes made, 11006 to missing)
```

Nous générons ensuite la variable status prédite par la modèle où

$$statushat = \begin{cases} 1 \; si \; p1best = 1 \\ 2 \; si \; p2best = 1 \\ 3 \; si \; p3best = 1 \end{cases}$$

```
gen statushat  = 1 if(p1best== 1 & p2best== 0 & p3best== 0)
(12708 missing values generated)

replace statushat = 2 if(p1best== 0 & p2best== 1 & p3best== 0)
(193 real changes made)

replace statushat = 3 if(p1best== 0 & p2best== 0 & p3best== 1)
(1509 real changes made)
```

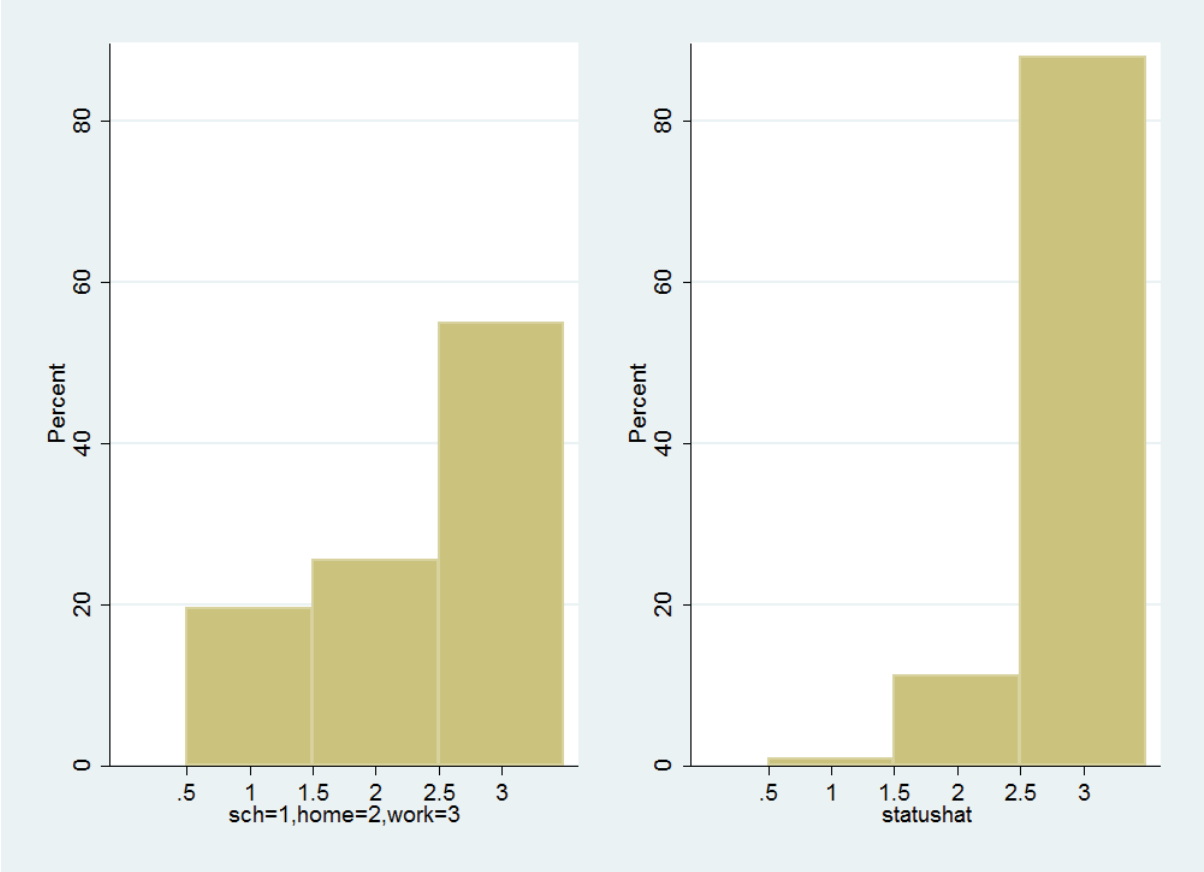
Nous créons enfin un tableau croisé des variables status et statushat en précisant l'option row pour avoir la fréquence de la cellule par rapport au total de la ligne en pourcentage et `nofreq` pour ne pas afficher la fréquence de la cellule. Nous affichons ainsi la proportion d'individus classés par le modèle par rapport à la vraie catégorie.

```
tab status statushat if year==87, row nofreq

sch=1,home |            statushat
 =2,work=3 |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |     12.12       4.04      83.84 |    100.00
         2 |      0.00      39.16      60.84 |    100.00
         3 |      0.23       4.59      95.18 |    100.00
-----------+---------------------------------+----------
     Total |      0.87      11.24      87.89 |    100.00
```

Nous voyons que le modèle classe correctement 12,12 % des individus étant scolarisés, 39,16 % des individus étant sans emploi et 95,18 % des individus étant employés.

Afin de nous apercevoir graphiquement de la différence entre les distributions des deux variables, nous allons combiner les histogrammes de celles-ci.

```
histogram status, discrete percent
(start=1, width=1)

histogram statushat, discrete percent
(start=1, width=1)
```



Interprétation des relations entre variable dépendante et variables indépendantes

Une autre manière d'avoir les résultats de l'estimation est de spécifier `rrr` comme option pour avoir accès aux ratios relatifs de risque.

```
mlogit, rrr

Multinomial logistic régression                   Number of obs   =       1717
                                                  LR chi2(8)      =     583.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -907.85723                       Pseudo R2       =     0.2433

------------------------------------------------------------------------------
         status |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
1               |
           educ |   1.961347   .1370979     9.64   0.000     1.710234     2.24933
          exper |   1.112061   .1927001     0.61   0.540     .7918276    1.561804
                |
c.exper#c.exper |   1.012594   .0255469     0.50   0.620     .9637407    1.063923
                |
          black |   .4435181   .1342632    -2.69   0.007     .2450377    .8027674
          _cons |   .0000344    .000039    -9.07   0.000     3.73e-06     .000317
----------------+-------------------------------------------------------------
2               |  (base outcome)
----------------+-------------------------------------------------------------
3               |
```

```
           educ |   1.431859   .0457372    11.24   0.000     1.344965    1.524368
          exper |   2.598545   .2511112     9.88   0.000     2.150176     3.14041
                |
c.exper#c.exper |   .9372688   .0119628    -5.08   0.000      .914113    .9610111
                |
          black |   .6055274    .089865    -3.38   0.001     .4526983    .8099511
          _cons |   .0087906   .0040587   -10.25   0.000     .0035563    .0217285
------------------------------------------------------------------------------
```

La significativité des coefficients ne change pas, mais nous pouvons cette fois-ci les interpréter. *Ceteris paribus*, si la variable educ varie d'une unité, nous pouvons observer qu'un individu a 1.96 fois plus de chance d'appartenir aux personnes actuellement scolarisées qu'aux personnes qui sont sans emploi et non scolarisées. La variable exper ne peut cependant pas être interprétée facilement. Nous aurons besoin de la commande `margins`.

Afin d'utiliser les commandes du module SPost, nous estimons à nouveau le modèle sans les variables factorielles.

```
quietly mlogit status educ exper expersq black if year==87
```

La première commande est la commande `listcoef` qui nous permet de comparer l'effet des variables à travers toutes les catégories de la variable status.

```
listcoef, help

mlogit (N=1717): Factor Change in the Odds of status

Variable: educ (sd=2.4812926)

Odds comparing    |
Alternative 1     |
to Alternative 2  |      b         z     P>|z|     e^b   e^bStdX
------------------+-----------------------------------------------
1       -2        |   0.67363    9.637   0.000   1.9613   5.3200
1       -3        |   0.31466    4.833   0.000   1.3698   2.1831
2       -1        |  -0.67363   -9.637   0.000   0.5099   0.1880
2       -3        |  -0.35897  -11.238   0.000   0.6984   0.4104
3       -1        |  -0.31466   -4.833   0.000   0.7300   0.4581
3       -2        |   0.35897   11.238   0.000   1.4319   2.4369
-----------------------------------------------------------------
```

Nous voyons que la variable d'éducation a un effet significatif pour des effets à travers toutes les catégories. *Ceteris paribus*, nous remarquons qu'un individu a plus de chance d'être scolarisé si son éducation varie d'une année par rapport à être sans emploi ou en emploi. Pareillement, nous pouvons aussi remarquer qu'un individu plus éduqué a plus de chance de se trouver en emploi que d'être sans emploi.

Nous laissons l'output pour la variable exper et expersq sans l'interpréter, car nous ne pouvons pas capter l'effet d'interaction correctement avec l'estimation sous cette forme et avec la commande `listcoef`. Nous allons développer cette interprétation avec la commande `margins`.

```
Variable: exper (sd=2.316598)

Odds comparing    |
Alternative 1     |
to Alternative 2  |      b         z     P>|z|     e^b   e^bStdX
------------------+-------------------------------------------
1       -2        |   0.10621    0.613   0.540   1.1121   1.2790
1       -3        |  -0.84874   -5.406   0.000   0.4280   0.1400
2       -1        |  -0.10621   -0.613   0.540   0.8992   0.7819
2       -3        |  -0.95495   -9.882   0.000   0.3848   0.1095
3       -1        |   0.84874    5.406   0.000   2.3367   7.1433
3       -2        |   0.95495    9.882   0.000   2.5985   9.1361
---------------------------------------------------------------

Variable: expersq (sd=18.138387)

Odds comparing    |
Alternative 1     |
to Alternative 2  |      b         z     P>|z|     e^b   e^bStdX
------------------+-------------------------------------------
1       -2        |   0.01252    0.496   0.620   1.0126   1.2548
1       -3        |   0.07730    3.372   0.001   1.0804   4.0637
2       -1        |  -0.01252   -0.496   0.620   0.9876   0.7969
2       -3        |   0.06479    5.076   0.000   1.0669   3.2385
3       -1        |  -0.07730   -3.372   0.001   0.9256   0.2461
3       -2        |  -0.06479   -5.076   0.000   0.9373   0.3088
---------------------------------------------------------------

Variable: black (sd=.48546152)

Odds comparing    |
Alternative 1     |
to Alternative 2  |      b         z     P>|z|     e^b   e^bStdX
------------------+-------------------------------------------
1       -2        |  -0.81302   -2.686   0.007   0.4435   0.6739
1       -3        |  -0.31136   -1.106   0.269   0.7324   0.8597
2       -1        |   0.81302    2.686   0.007   2.2547   1.4839
2       -3        |   0.50166    3.380   0.001   1.6515   1.2758
3       -1        |   0.31136    1.106   0.269   1.3653   1.1632
3       -2        |  -0.50166   -3.380   0.001   0.6055   0.7839
---------------------------------------------------------------
```

De ces tableaux, nous pouvons noter qu'un individu noir de notre échantillon a plus de chance d'être sans emploi qu'en emploi, *ceteris paribus*. Il a aussi plus de chance d'être sans emploi que scolarisé et plus de chance de travailler que d'être scolarisé.

L'option `help` nous permet d'avoir accès à la signification de chaque colonne du tableau.

```
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-valeur  for z-test
     e^b = exp(b) = factor change in odds for unit increase in X
 e^bStdX = exp(b*SD of X) = change in odds for SD increase in X
```

Une autre commande d'intérêt du module SPost est la commande `prchange` qui nous permet d'avoir accès aux effets marginaux à la moyenne et à l'effet d'autres variations concernant les

probabilités pour un individu de se trouver dans l'une ou l'autre des catégories de la variable dépendante.

```
prchange, help

mlogit: Changes in Probabilities for status

educ
            Avg|Chg|           1           2           3
Min->Max   .36339941   .21288132  -.54509914   .33221778
   -+1/2    .0303475   .01086067  -.04552126   .03466058
  -+sd/2   .07578843   .02739228  -.11368266   .08629036
MargEfct     .030308   .01082604    -.045462   .03463597

exper
            Avg|Chg|           1           2           3
Min->Max   .56083208  -.10768881  -.73355932   .84124811
   -+1/2   .09005499  -.02080508  -.11427741   .13508248
  -+sd/2   .21155314  -.04768537  -.26964433    .3173297
MargEfct   .08961937  -.02082885   -.1136002   .13442905

expersq
            Avg|Chg|           1           2           3
Min->Max   .61392254   .35843275   .56245105  -.92088382
   -+1/2   .00640988   .00199379   .00762103  -.00961483
  -+sd/2    .1172177    .0370772   .13874936  -.17582655
MargEfct   .00640971   .00199363   .00762094  -.00961456

black
         Avg|Chg|           1           2           3
0->1   .04391455  -.01100145   .06587184  -.05487037

                 1           2           3
Pr(y|x)  .03036102  .14338864  .82625031

          educ    exper  expersq    black
   x=  12.5492   3.4403  17.1992  .379732
sd_x=  2.48129   2.3166  18.1384  .485462

 Pr(y|x): probability of observing each y for specified x values
Avg|Chg|: average of absolute value of the change across categories
Min->Max: change in predicted probability as x changes from its minimum to
          its maximum
    0->1: change in predicted probability as x changes from 0 to 1
   -+1/2: change in predicted probability as x changes from 1/2 unit below
          base value to 1/2 unit above
  -+sd/2: change in predicted probability as x changes from 1/2 standard
          dev below base to 1/2 standard dev above
MargEfct: the partial derivative of the predicted probability/rate with
          respect to a given independent variable
```

Ici, nous n'interpréterons pas l'effet de exper et expersq. Ce tableau est intéressant, car il nous donne les grandes lignes des effets des variables pour différentes variations, soit de la valeur minimale à la valeur maximale de la variable pour une variation de $\frac{1}{2}$ de la variable à partir de la valeur de base, pour une variation de $\frac{1}{2}$ déviation standard et enfin pour une variation de 0 à 1 pour une variable binaire. Nous avons aussi une idée de la valeur moyenne du changement à travers les catégories et de la fréquence observée. En prenant pour exemple l'éducation, nous voyons que l'effet est négatif sur la probabilité d'être sans emploi et positif sur la probabilité d'être scolarisé ou en emploi.

La partie interprétation des résultats peut s'avérer plus précise grâce à la commande `margins`, la commande `marginsplot` et leurs options. Nous estimons à nouveau silencieusement le modèle en incluant les variables factorielles et les opérateurs. Il est nécessaire d'estimer le modèle ainsi afin que Stata puisse calculer les bons effets marginaux. Autrement, les effets marginaux seront erronés et l'interprétation des résultats sera fausse.

```
quietly mlogit status educ c.exper##c.exper i.black if year==87
```

Nous estimons premièrement les effets marginaux moyens pour la modalité 1. Nous spécifions donc `predict(outcome(1))` et `dydx(*)` pour obtenir les effets marginaux de toutes les variables.

```
margins, dydx(*) predict(outcome(1))

Average marginal effects                        Number of obs   =       1717
Model VCE    : OIM

Expression   : Pr(status==1), predict(outcome(1))
dy/dx w.r.t. : educ exper 1.black

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0173786   .0029057     5.98   0.000     .0116835    .0230737
       exper |  -.0194377   .0042472    -4.58   0.000    -.0277621   -.0111133
     1.black |   -.017991   .0116456    -1.54   0.122     -.040816     .004834
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

L'éducation a un effet marginal moyen positif et significatif de 1,7 % par année supplémentaire sur la probabilité d'être scolarisé toutes choses égales par ailleurs. L'expérience a quant à elle un effet marginal moyen négatif et significatif de 1,9 % par année supplémentaire sur la probabilité d'être scolarisé, *ceteris paribus*. L'effet de la variable dichotomique black n'est pas statistiquement différent de 0.

Nous demandons le même type d'effet marginal concernant l'appartenance à la catégorie d'individus n'ayant pas d'emploi.

```
margins, dydx(*) predict(outcome(2))

Average marginal effects                        Number of obs   =       1717
Model VCE    : OIM

Expression   : Pr(status==2), predict(outcome(2))
dy/dx w.r.t. : educ exper 1.black

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0429471   .0030007   -14.31   0.000    -.0488283   -.0370659
       exper |  -.0619329   .0039382   -15.73   0.000    -.0696516   -.0542143
     1.black |   .0614621   .0178146     3.45   0.001     .0265461     .096378
```

```
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Dans ce cas-ci, les variables educ et exper ont toutes les deux un impact négatif et significatif de respectivement 4,3 % et 6,2 % sur la probabilité de ne pas travailler par année supplémentaire. Un individu d'origine noire aura quant à lui 6,1 % plus de probabilité de ne pas travailler, en moyenne.

De même, pour la dernière modalité nous avons :

```
margins, dydx(*) predict(outcome(3))

Average marginal effects                        Number of obs   =       1717
Model VCE    : OIM

Expression   : Pr(status==3), predict(outcome(3))
dy/dx w.r.t. : educ exper 1.black

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0255685   .0040739     6.28   0.000     .0175837    .0335533
       exper |   .0813706   .0044553    18.26   0.000     .0726383    .0901029
     1.black |   -.043471   .0199503    -2.18   0.029     -.082573   -.0043691
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Ceteris paribus, nous voyons que l'expérience et l'éducation ont toutes deux un effet positif et significatif de respectivement 2,6 % et 8,1 % sur la probabilité de travailler. Nous confirmons donc nos hypothèses posées en introduction. Un individu d'origine noire aura quant à lui 4,3 % moins de probabilité de travailler, toutes choses étant égales par ailleurs.

Nous pouvons maintenant calculer les effets marginaux conditionnellement à ce que les variables educ, exper et black soient à leur moyenne (les mêmes résultats sont donnés par Stata en excluant la condition sur black).

```
margins, dydx(*) atmeans  predict(outcome(1))

Conditional marginal effects                    Number of obs   =       1717
Model VCE    : OIM

Expression   : Pr(status==1), predict(outcome(1))
dy/dx w.r.t. : educ exper 1.black
at           : educ            =    12.54921 (mean)
               exper           =    3.440303 (mean)
               0.black         =    .6202679 (mean)
               1.black         =    .3797321 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0073312   .0015143     4.84   0.000     .0043631    .0102992
       exper |  -.0054125   .0019287    -2.81   0.005    -.0091928   -.0016323
```

```
     1.black |  -.0073531   .0055632    -1.32   0.186    -.0182567    .0035505
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Quand les variables sont à leur moyenne, les effets marginaux ne changent pas de signe, mais sont un peu plus faibles que les effets marginaux à la moyenne. L'éducation a un impact de 0,7 % (comparé à 1,7 % en moyenne) par année supplémentaire tandis que l'expérience a un impact de 0,5 % (comparé à 1,9 % en moyenne) concernant l'appartenance au groupe d'individus étant scolarisés, toutes choses égales par ailleurs. La variable black est encore une fois non significative.

Pour la modalité 2, nous avons :

```
margins, dydx(*) atmeans  predict(outcome(2))

Conditional marginal effects                      Number of obs   =       1717
Model VCE    : OIM

Expression   : Pr(status==2), predict(outcome(2))
dy/dx w.r.t. : educ exper 1.black
at           : educ            =    12.54921 (mean)
               exper           =    3.440303 (mean)
               0.black         =    .6202679 (mean)
               1.black         =    .3797321 (mean)
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0349842    .003619    -9.67   0.000    -.0420773   -.0278911
       exper |  -.0478969   .0044524   -10.76   0.000    -.0566234   -.0391704
     1.black |   .0511596   .0155613     3.29   0.001     .0206599    .0816592
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Cette fois-ci, les effets marginaux à la moyenne ont un impact assez semblable aux effets marginaux moyens. Nous ne nous attarderons pas à les interpréter.

La médiane peut également être d'intérêt pour calculer les effets marginaux. Nous spécifions donc l'option `at((median)…)` suivie du nom des variables pour obtenir les effets marginaux à la médiane pour l'appartenance à la modalité 1.

```
margins, dydx(*) at((median)exper educ) predict(outcome(1))

Conditional marginal effects                      Number of obs   =       1717
Model VCE    : OIM

Expression   : Pr(status==1), predict(outcome(1))
dy/dx w.r.t. : educ exper 1.black
at           : educ            =          12 (median)
               exper           =           3 (median)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

```
        educ |   .0072241   .0012823     5.63   0.000     .0047109    .0097374
       exper |  -.0057826   .0020379    -2.84   0.005    -.0097768   -.0017884
     1.black |  -.0072963   .0053115    -1.37   0.170    -.0177068    .0031141
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Si nous comparons l'effet d'educ à la médiane de 0,72 % et celui à la moyenne de 0,73 %, nous pouvons nous apercevoir qu'il n'y a pas une grande différence dans les données que nous analysons. Nous pouvons essayer l'exercice sur les autres catégories et pour d'autres valeurs fixes.

Nous nous intéresserons plutôt à calculer les effets marginaux estimations conditionnellement à plusieurs valeurs d'une variable sur une plage de valeurs afin de pouvoir les représenter graphiquement. Encore une fois, nous pouvons spécifier d'autres conditions et demander à Stata de calculer les effets marginaux moyens ou les effets marginaux à la moyenne.

Nous commençons par calculer les effets marginaux sous plusieurs angles. Tout d'abord, nous voulons visualiser la différence black et non-black et nous spécifions donc la variable black directement après `margins`. Nous aimerions connaître l'effet marginal de la variable exper et nous la plaçons dans les parenthèses de l'option `dydx()`. Afin d'avoir plusieurs points sur notre graphique, nous utilisons l'option `at` de sorte que l'éducation soit calculée à 13 niveaux soit de 7 à 19. Nous ajoutons l'option `atmeans` pour que les autres variables soient à la moyenne et `predict(outcome(1))` pour que les effets marginaux soient relatifs à la probabilité pour un individu d'appartenir à la catégorie 1. Nous n'incluons pas toute la sortie Stata dans l'exemple ci-dessous.

```
margins black, dydx(exper) at(educ=(7(1)19)) atmeans predict(outcome(1))

Conditional marginal effects                      Number of obs   =       1717
Model VCE    : OIM

Expression   : Pr(status==1), predict(outcome(1))
dy/dx w.r.t. : exper

1._at        : educ            =           7
               exper           =    3.440303 (mean)
               0.black         =    .6202679 (mean)
               1.black         =    .3797321 (mean)

2._at        : educ            =           8
               exper           =    3.440303 (mean)
               0.black         =    .6202679 (mean)
               1.black         =    .3797321 (mean)

                              .
                              .
                              .

12._at       : educ            =          18
               exper           =    3.440303 (mean)
               0.black         =    .6202679 (mean)
```

```
                1.black         =    .3797321 (mean)

13._at        : educ            =          19
                exper           =    3.440303 (mean)
                0.black         =    .6202679 (mean)
                1.black         =    .3797321 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
exper        |
   _at#black |
        1 0  |  -.0002715   .0003023    -0.90   0.369    -.0008641     .000321
        1 1  |  -.0000568   .0001345    -0.42   0.673    -.0003203    .0002067
        2 0  |  -.0006092   .0005098    -1.19   0.232    -.0016083      .00039
        2 1  |  -.0002099    .000243    -0.86   0.388    -.0006861    .0002663

                                                .
                                                .
                                                .

       12 0  |  -.0350697    .007326    -4.79   0.000    -.0494283    -.020711
       12 1  |   -.026968   .0076145    -3.54   0.000    -.0418922   -.0120439
       13 0  |  -.0441532   .0091083    -4.85   0.000     -.062005   -.0263013
       13 1  |  -.0349762   .0098936    -3.54   0.000    -.0543673    -.015585
------------------------------------------------------------------------------
```

Pour avoir un aperçu graphique de toutes les valeurs calculées, nous faisons appel à la commande `marginsplot` et spécifions quelques options afin de rendre le graphique plus compréhensible. Nous voulons que les effets soient représentés par une ligne avec l'option `recast(line)` et les intervalles de confiance par des lignes discontinues avec l'option `recastci(rline) ciopts(lpattern(dash))`. De plus, nous ajoutons un titre et une ligne horizontale à y=0.

```
marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) title(Effet marginal à la
moyenne de exper) yline(0)

  Variables that uniquely identify margins: educ black

graph save marginsmlogit1.gph, replace
(file marginsmlogit1.gph saved)
```

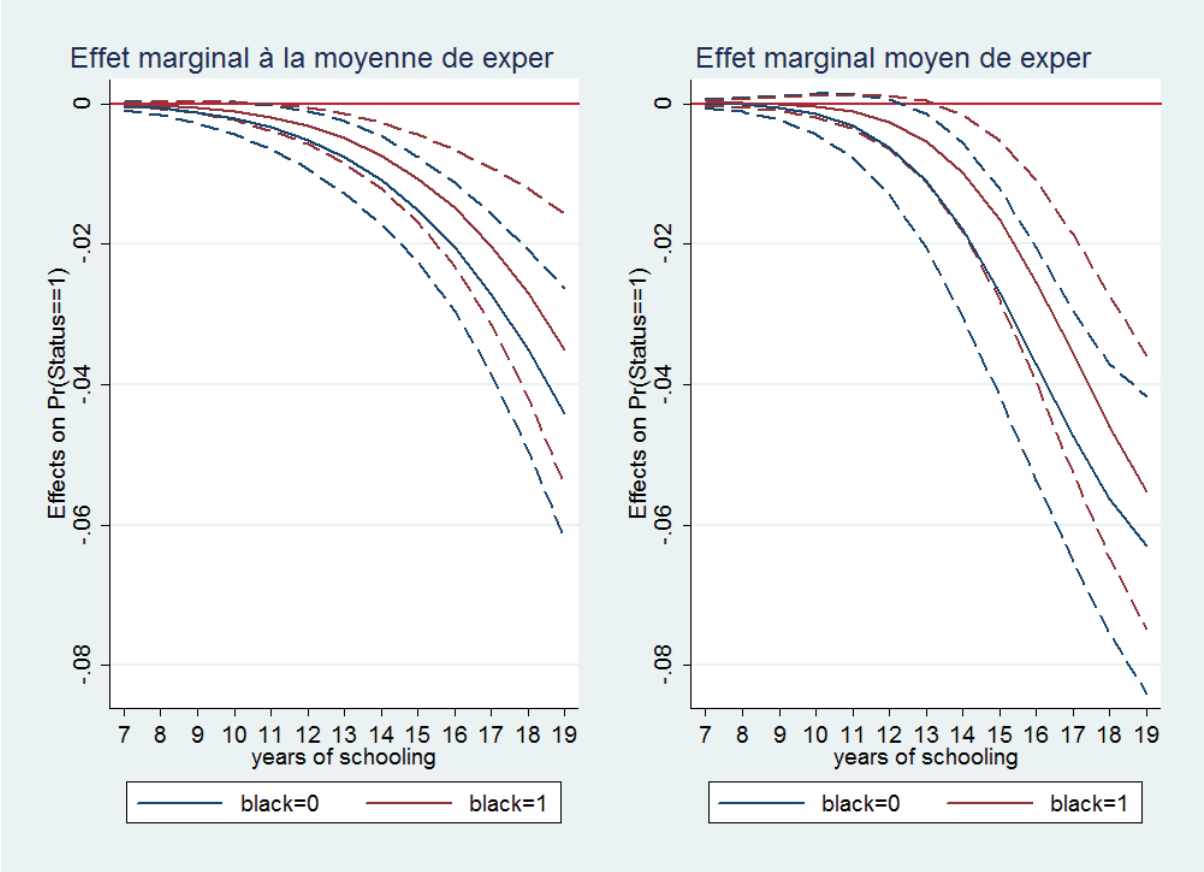
Nous allons combiner les graphiques afin de pouvoir comparer les effets marginaux moyens et les effets marginaux à la moyenne. Nous estimons donc maintenant l'effet marginal moyen de la variable exper sans spécifier l'option `atmeans`.

```
margins black, dydx(exper) at(educ=(7(1)19)) predict(outcome(1))

Average marginal effects                        Number of obs   =       1717
Model VCE    : OIM

Expression   : Pr(status==1), predict(outcome(1))
dy/dx w.r.t. : exper

1._at        : educ            =           7

2._at        : educ            =           8
```

```
                                .
                                .
                                .

12._at       : educ            =          18

13._at       : educ            =          19

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
exper        |
   _at#black |
        1 0  |   .0000559   .0003054     0.18   0.855    -.0005426    .0006544
        1 1  |   .0001178   .0001748     0.67   0.501    -.0002248    .0004604
        2 0  |  -.0001121   .0005213    -0.21   0.830    -.0011339    .0009097
        2 1  |   .0000954   .0002855     0.33   0.738    -.0004641    .0006548

                                                .
                                                .
                                                .

       12 0  |  -.0562784   .0097988    -5.74   0.000    -.0754837   -.0370731
       12 1  |  -.0460697   .0095379    -4.83   0.000    -.0647636   -.0273759
       13 0  |  -.0629091   .0107873    -5.83   0.000    -.0840518   -.0417663
       13 1  |  -.0553135   .0099557    -5.56   0.000    -.0748264   -.0358007
------------------------------------------------------------------------------
```

Nous demandons le même type de graphique que précédemment.

```
marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) title(Effet marginal moyen de
exper) yline(0)

  Variables that uniquely identify margins: educ black

graph save marginsmlogit2.gph, replace
(file marginsmlogit2.gph saved)

graph combine marginsmlogit3.gph marginsmlogit4.gph, ycommon
```



Nous pouvons voir que l'effet marginal moyen et l'effet marginal à la moyenne sont différents pour des valeurs élevées de la variable éducation. Nous notons aussi que la différence s'accentue entre les individus d'origine noire et les individus n'étant pas d'origine noire au fur et à mesure que la variable d'éducation varie, et ce, pour les deux types de calculs.

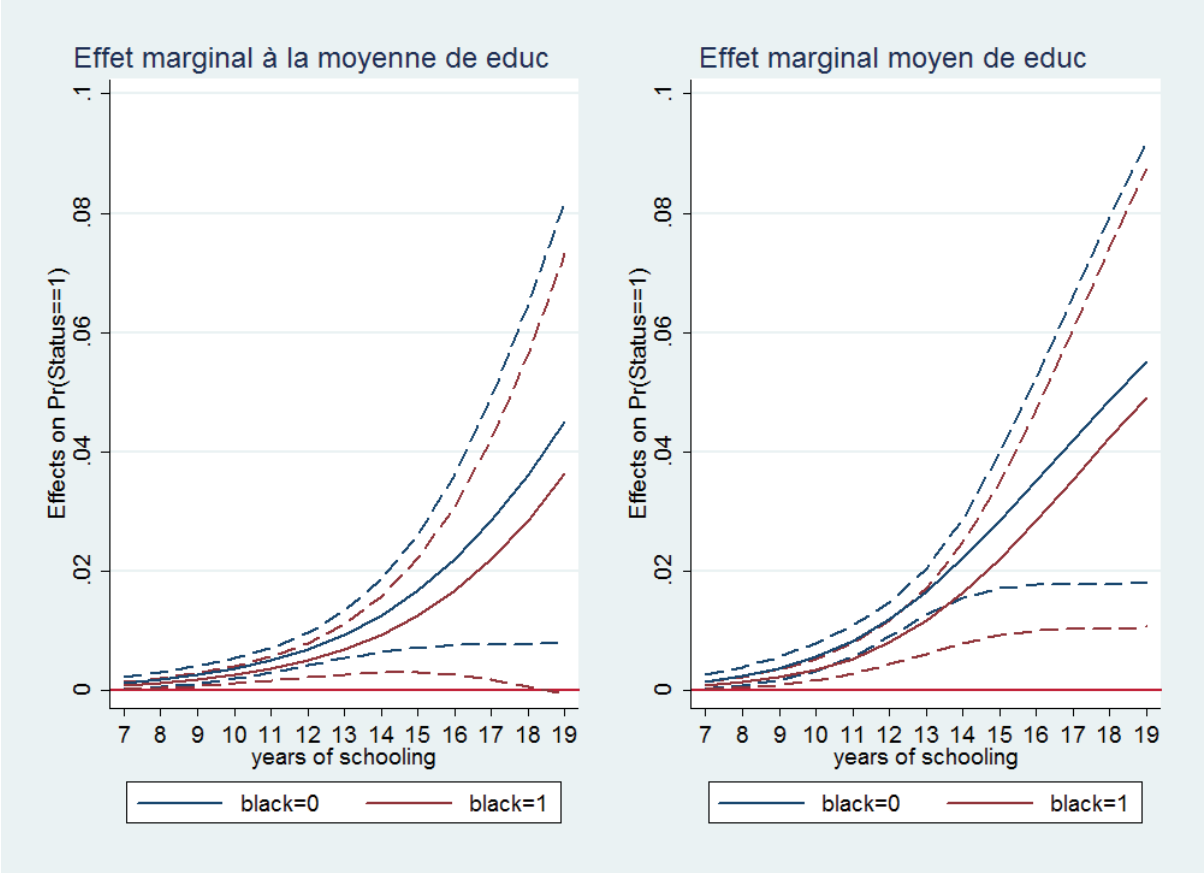
Nous voulons maintenant connaître l'impact marginal de l'éducation lorsque celle-ci varie et effectuer la même comparaison graphique entre les deux méthodes de calcul pour la même modalité de la variable dépendante. Nous suivons donc les mêmes étapes que précédemment en échangeant exper pour educ dans la parenthèse de l'option `dydx()`.

```
margins black, dydx(educ) at(educ=(7(1)19)) atmeans predict(outcome(1))

Conditional marginal effects                      Number of obs   =       1717
Model VCE    : OIM

Expression   : Pr(status==1), predict(outcome(1))
dy/dx w.r.t. : educ

1._at        : educ            =           7
               exper           =    3.440303 (mean)
               0.black         =    .6202679 (mean)
               1.black         =    .3797321 (mean)

2._at        : educ            =           8
```

```
               exper           =    3.440303 (mean)
               0.black         =    .6202679 (mean)
               1.black         =    .3797321 (mean)

                              .
                              .
                              .

12._at       : educ            =          18
               exper           =    3.440303 (mean)
               0.black         =    .6202679 (mean)
               1.black         =    .3797321 (mean)

13._at       : educ            =          19
               exper           =    3.440303 (mean)
               0.black         =    .6202679 (mean)
               1.black         =    .3797321 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
educ         |
   _at#black |
        1 0  |    .001272   .0005107     2.49   0.013      .000271     .002273
        1 1  |   .0008008   .0003191     2.51   0.012     .0001754    .0014263
        2 0  |   .0018636   .0006312     2.95   0.003     .0006266    .0031007
        2 1  |   .0012306   .0004294     2.87   0.004     .0003891    .0020721

                                                 .
                                                 .
                                                 .

       11 0  |   .0284457   .0105431     2.70   0.007     .0077816    .0491097
       11 1  |   .0220361   .0102924     2.14   0.032     .0018632    .0422089
       12 0  |   .0362084   .0145183     2.49   0.013      .007753    .0646637
       12 1  |   .0285462   .0142082     2.01   0.045     .0006986    .0563938
       13 0  |   .0450589   .0188829     2.39   0.017     .0080492    .0820686
       13 1  |   .0363254   .0187956     1.93   0.053    -.0005134    .0731641
------------------------------------------------------------------------------

marginsplot,  recast(line)  recastci(rline)  ciopts(lpattern(dash))  title(Effet  marginal  à  la
moyenne de educ) yline(0)

  Variables that uniquely identify margins: educ black

graph save marginsmlogit3.gph, replace
(file marginsmlogit3.gph saved)

margins black, dydx(educ) at(educ=(7(1)19)) predict(outcome(1))

Average marginal effects                        Number of obs   =       1717
Model VCE    : OIM

Expression   : Pr(status==1), predict(outcome(1))
dy/dx w.r.t. : educ

1._at        : educ            =           7

2._at        : educ            =           8

                              .
                              .
                              .

12._at       : educ            =          18

13._at       : educ            =          19

------------------------------------------------------------------------------
             |            Delta-method
```

```
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
educ         |
   _at#black |
        1 0  |   .0015206   .0005898     2.58   0.010     .0003646    .0026766
        1 1  |   .0008917   .0003358     2.66   0.008     .0002335    .0015499
        2 0  |   .0023928   .0007871     3.04   0.002     .0008501    .0039355
        2 1  |   .0014361   .0004742     3.03   0.002     .0005067    .0023655

                                                  .
                                                  .
                                                  .

       12 0  |   .0486865   .0157745     3.09   0.002     .0177689     .079604
       12 1  |   .0423149   .0162999     2.60   0.009     .0103677    .0742621
       13 0  |   .0550641   .0187871     2.93   0.003      .018242    .0918862
       13 1  |   .0490613   .0195833     2.51   0.012     .0106788    .0874438
------------------------------------------------------------------------------

marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) title(Effet marginal moyen de
educ) yline(0)

  Variables that uniquely identify margins: educ black

graph save marginsmlogit4.gph, replace

graph combine marginsmlogit3.gph marginsmlogit4.gph, ycommon
```

Après combinaison, nous obtenons le graphique ci-dessous



Nous pouvons observer un effet positif et des intervalles de confiances qui deviennent rapidement très larges. Il existe aussi une différence entre un individu d'origine noire et d'origine non noire qui s'accentue avec le niveau d'éducation atteint. Nous pourrions penser que faire des études accroit la capacité de continuer ses études à un niveau supérieur, mais accroit le salaire espéré et donc la possibilité de rentrer sur le marché du travail.

Ces analyses peuvent être obtenues pour chaque variable en faisant varier une variable continue ou catégorielle sur une plage de données voulue. Il existe de nombreuses possibilités pour analyser les relations entre les variables indépendantes et la variable dépendante sachant que les effets marginaux dépendent du niveau où les variables sont observées. Il faut donc faire un choix judicieux pour explorer en détail ces effets marginaux.

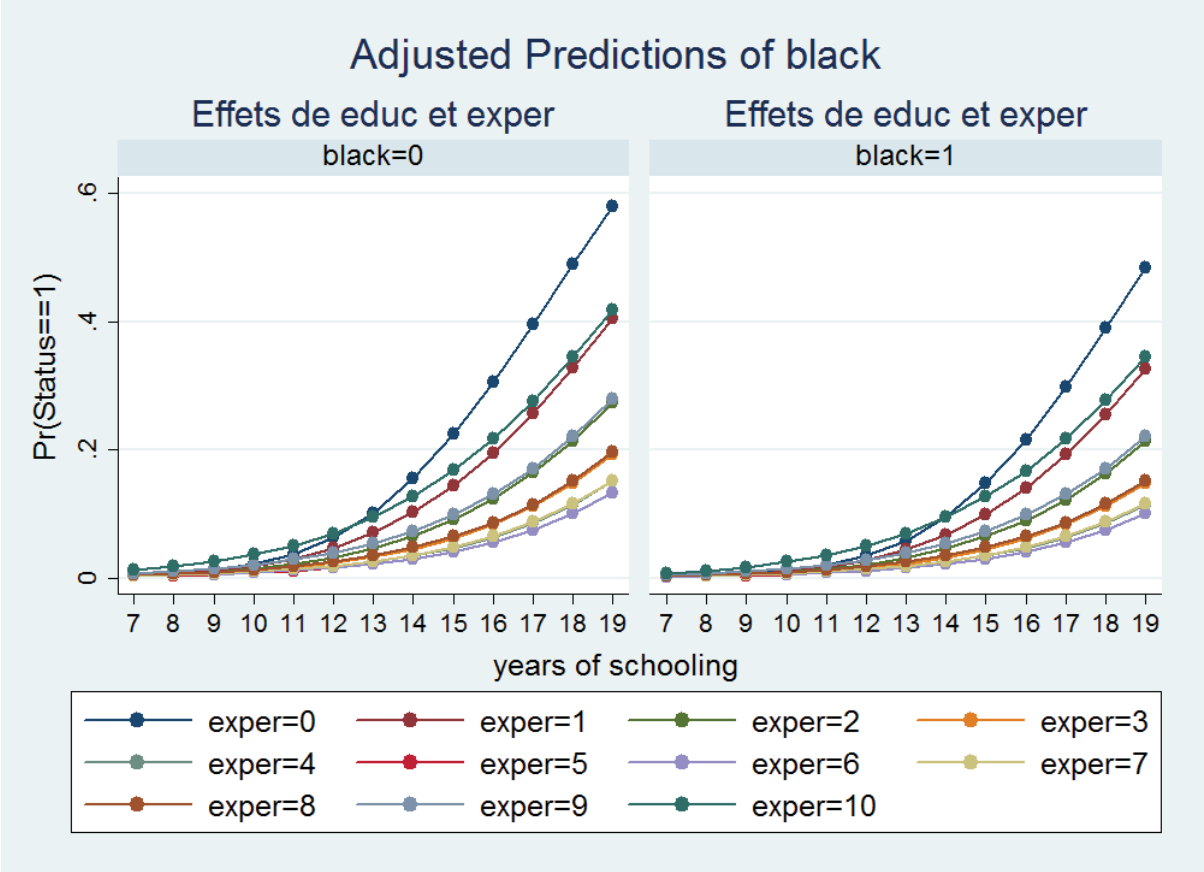
Nous pouvons avoir un portrait plus global des relations entre les variables en enlevant l'option `dydx()` afin de calculer les prédictions ajustées à plusieurs niveaux des variables grâce à l'option `at`. Celle-ci fonctionne exactement comme dans les commandes précédentes.

```
margins black, at(educ=(7(1)19)exper=(0(1)10)) predict(outcome(1))

Adjusted prédictions                            Number of obs   =       1717
Model VCE    : OIM

Expression   : Pr(status==1), predict(outcome(1))

1._at        : educ            =           7
               exper           =           0

2._at        : educ            =           7
               exper           =           1

                              .
                              .
                              .

10._at       : educ            =           7
               exper           =           9

11._at       : educ            =           7
               exper           =          10

12._at       : educ            =           8
               exper           =           0

13._at       : educ            =           8
               exper           =           1

                              .
                              .
                              .

21._at       : educ            =           8
               exper           =           9

22._at       : educ            =           8
               exper           =          10
```

```
                              .
                              .
                              .

142._at      : educ            =          19
               exper           =           9

143._at      : educ            =          19
               exper           =          10

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _at#black |
        1 0  |   .0034517   .0022558     1.53   0.126    -.0009695    .0078729
        1 1  |   .0015953   .0009903     1.61   0.107    -.0003456    .0035362
        2 0  |   .0034082   .0020551     1.66   0.097    -.0006198    .0074362
        2 1  |   .0016503   .0009392     1.76   0.079    -.0001906    .0034912

                                                 .
                                                 .
                                                 .

      142 0  |   .2794946   .1973425     1.42   0.157    -.1072895    .6662788
      142 1  |   .2207742   .1813979     1.22   0.224    -.1347591    .5763075
      143 0  |   .4186346   .3038009     1.38   0.168    -.1768042    1.014073
      143 1  |   .3444639   .2968089     1.16   0.246    -.2372709    .9261987
------------------------------------------------------------------------------
```

Pour représenter graphiquement ces nombreux calculs, nous allons supprimer les intervalles de confiance avec l'option `noci` et agrandir le nombre de colonnes dans la légende avec l'option `legend(cols(4))`.

```
marginsplot, noci by(black) legend(cols(4)) title(Effets de educ et exper)

  Variables that uniquely identify margins: educ exper black
```

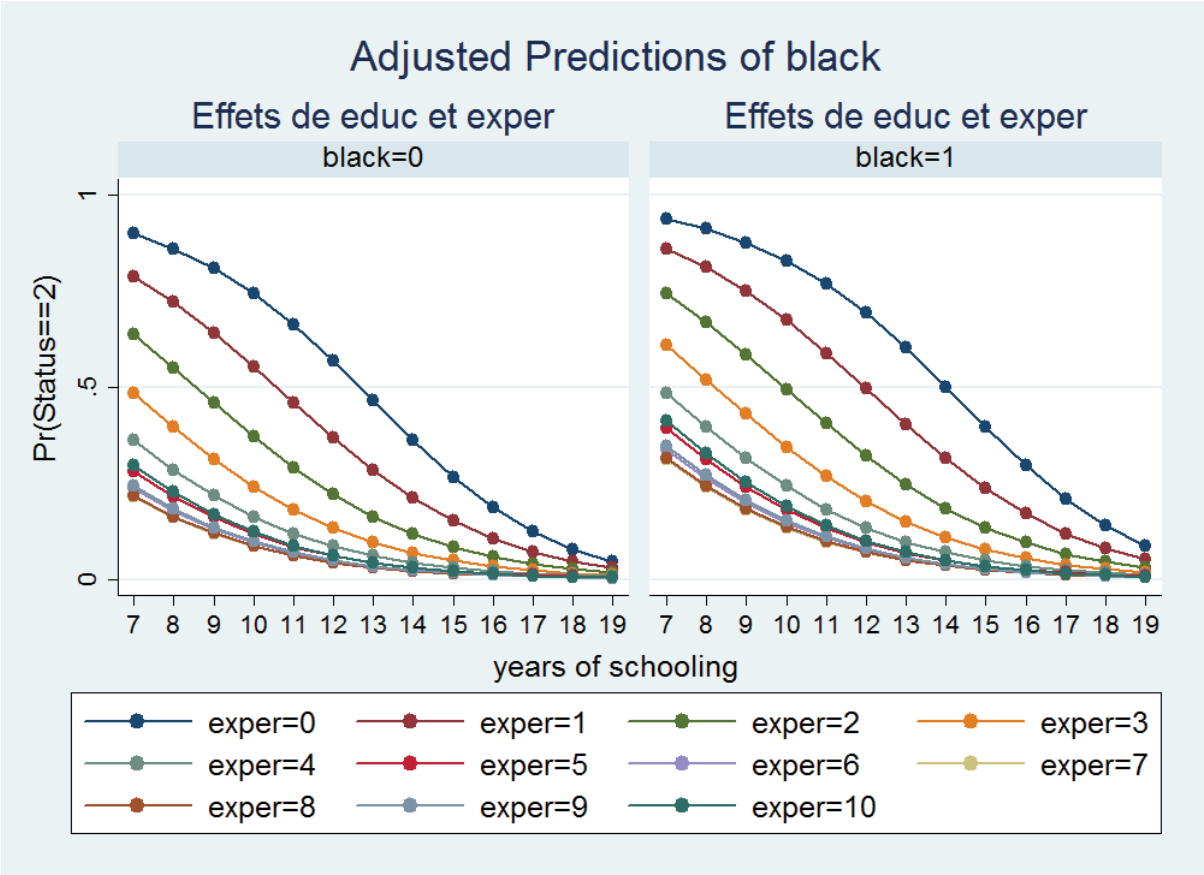


Le graphique que nous obtenons nous permet d'explorer les prédictions ajustées du modèle pour des valeurs différentes des variables educ et exper. Nous comprenons que davantage d'éducation augmente la probabilité pour un individu d'être scolarisé à travers n'importe quel niveau d'expérience et indépendamment de la variable ethnique. Par contre, au niveau de l'expérience, nous constatons qu'avoir peu d'expérience et beaucoup d'expérience augmente la probabilité d'être scolarisé tandis que les individus ayant une expérience moyenne sont ceux qui ont le moins de probabilité d'être scolarisés et cela indépendamment de l'ethnie et de l'éducation. Comme les tendances se ressemblent fortement entre ethnies, nous pouvons affirmer que les individus ayant des origines noires ont généralement moins de probabilité d'être scolarisés.

Nous répétons maintenant les mêmes commandes pour analyser l'appartenance au groupe d'individus ne travaillant pas (modalité 2).

```
margins black, at(educ=(7(1)19)exper=(0(1)10)) predict(outcome(2))

Adjusted prédictions                            Number of obs   =       1717
Model VCE    : OIM

Expression   : Pr(status==2), predict(outcome(2))
```

```
1._at        : educ            =           7
               exper           =           0

2._at        : educ            =           7
               exper           =           1

                               .
                               .
                               .

10._at       : educ            =           7
               exper           =           9

11._at       : educ            =           7
               exper           =          10

12._at       : educ            =           8
               exper           =           0

13._at       : educ            =           8
               exper           =           1

                               .
                               .
                               .

21._at       : educ            =           8
               exper           =           9

22._at       : educ            =           8
               exper           =          10

23._at       : educ            =           9
               exper           =           0

24._at       : educ            =           9
               exper           =           1

                               .
                               .
                               .

142._at      : educ            =          19
               exper           =           9

143._at      : educ            =          19
               exper           =          10

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _at#black |
        1 0  |   .8990285   .0238603    37.68   0.000     .8522631    .9457939
        1 1  |   .9368684   .0143968    65.07   0.000     .9086511    .9650856
        2 0  |   .7883254    .037748    20.88   0.000     .7143408    .8623101
        2 1  |   .8606661   .0241718    35.61   0.000     .8132902     .908042

                                                  .
                                                  .
                                                  .

      142 0  |   .0031337   .0016592     1.89   0.059    -.0001183    .0063856
      142 1  |    .005581   .0029852     1.87   0.062    -.0002698    .0114319
      143 0  |   .0033275   .0025571     1.30   0.193    -.0016843    .0083393
      143 1  |   .0061732    .004594     1.34   0.179    -.0028309    .0151774
------------------------------------------------------------------------------

marginsplot, noci by(black) legend(cols(4)) title(Effets de educ et exper)

  Variables that uniquely identify margins: educ exper black
```

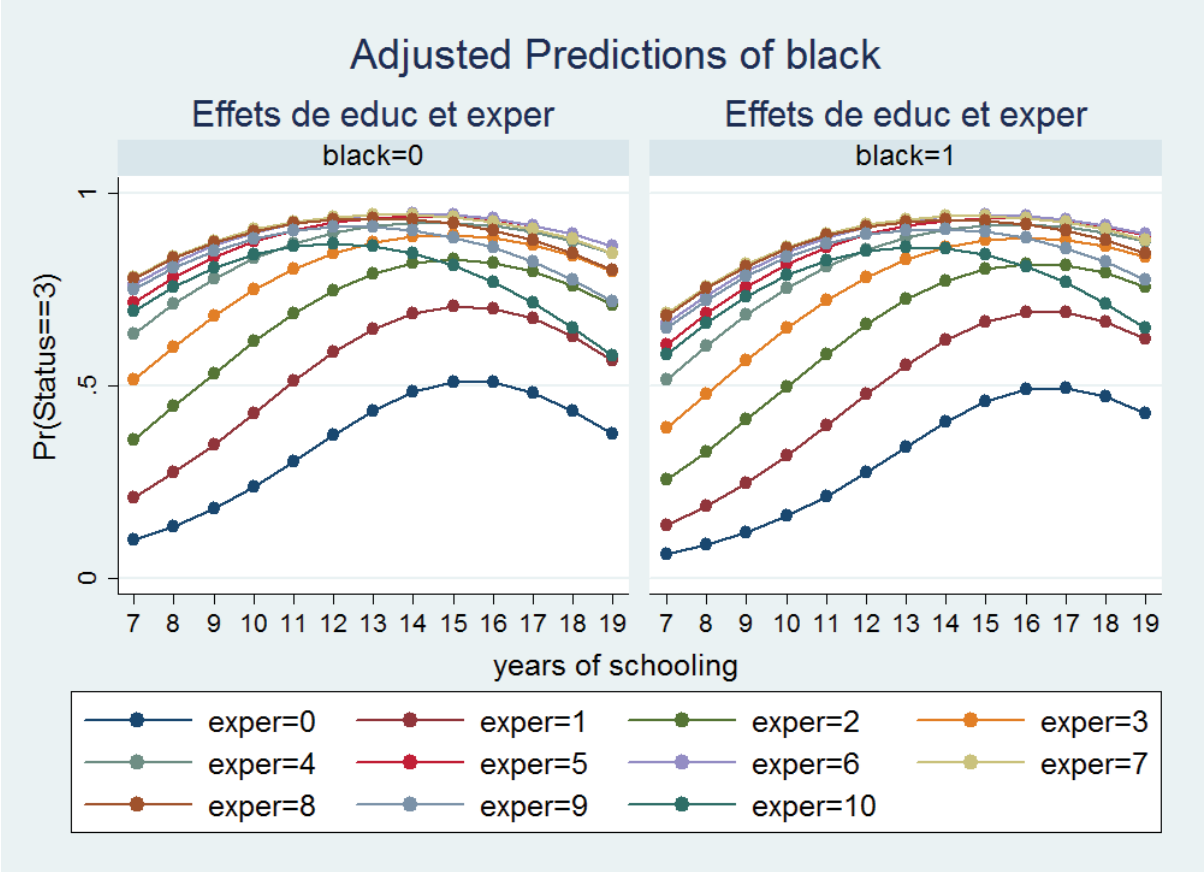


Cette fois-ci, nous remarquons que l'éducation a un effet négatif sur la probabilité d'être sans emploi. Cet effet change de forme dépendamment du niveau d'expérience, mais s'avère clairement décroissant et devient totalement convexe à partir de deux années d'expérience. La probabilité d'être sans emploi diminue si l'expérience augmente jusqu'à huit années d'expérience. De plus, une augmentation de l'expérience augmente la probabilité pour un individu d'être sans emploi.

Nous répétons la même analyse pour la dernière catégorie de la variable dépendante avec les mêmes commandes.

```
margins black, at(educ=(7(1)19)exper=(0(1)10)) predict(outcome(3))


Adjusted prédictions                            Number of obs   =       1717
Model VCE    : OIM

Expression   : Pr(status==3), predict(outcome(3))

1._at        : educ            =           7
               exper           =           0
```

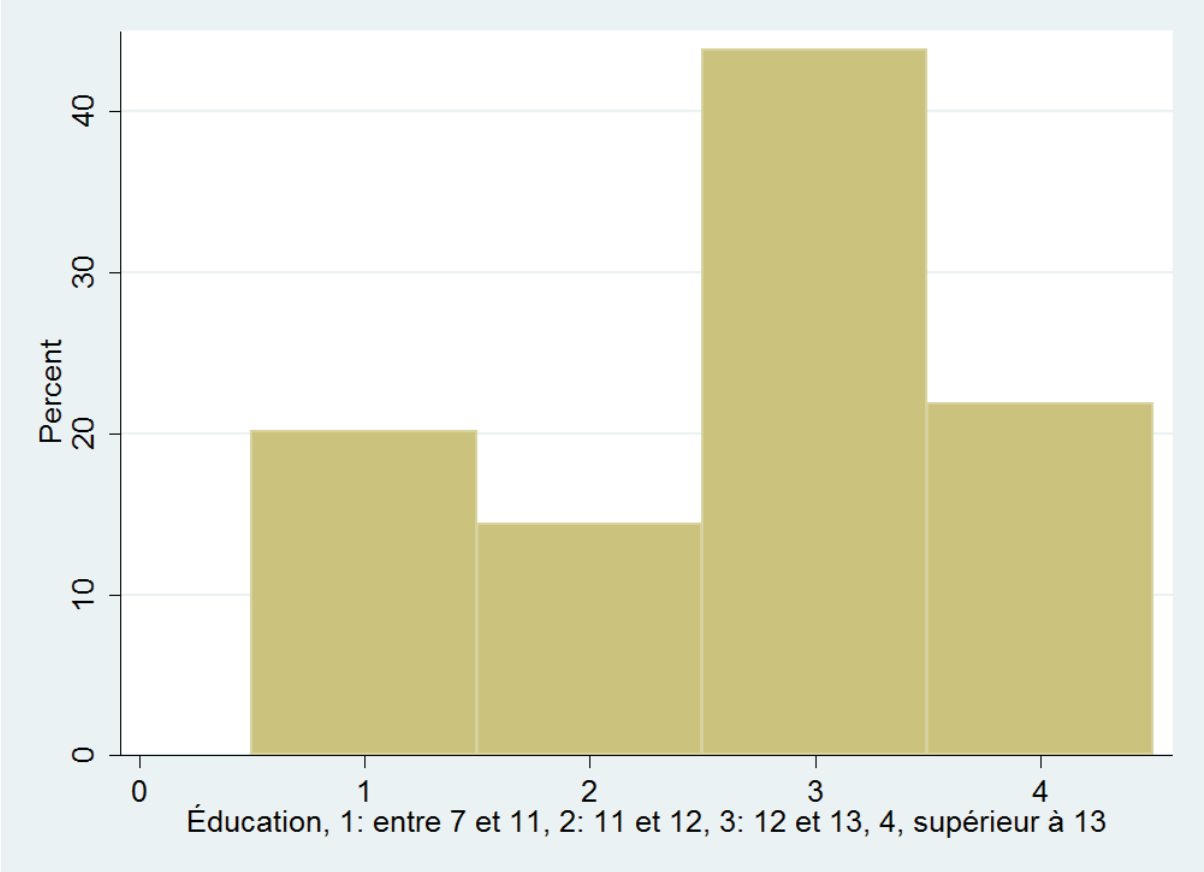
```
2._at        : educ            =           7
               exper           =           1

                              .
                              .
                              .

10._at       : educ            =           7
               exper           =           9

11._at       : educ            =           7
               exper           =          10

12._at       : educ            =           8
               exper           =           0

13._at       : educ            =           8
               exper           =           1

                              .
                              .
                              .

142._at      : educ            =          19
               exper           =           9

143._at      : educ            =          19
               exper           =          10

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _at#black |
        1 0  |   .0975198   .0232474     4.19   0.000     .0519557    .1430839
        1 1  |   .0615363   .0141319     4.35   0.000     .0338383    .0892343
        2 0  |   .2082664   .0373401     5.58   0.000     .1350811    .2814517
        2 1  |   .1376836   .0239783     5.74   0.000      .090687    .1846802

                                                  .
                                                  .
                                                  .

      142 0  |   .7173717   .1965764     3.65   0.000      .332089    1.102654
      142 1  |   .7736447   .1803048     4.29   0.000     .4202537    1.127036
      143 0  |    .578038   .3021831     1.91   0.056    -.0142299    1.170306
      143 1  |   .6493629   .2942788     2.21   0.027     .0725869    1.226139
------------------------------------------------------------------------------

marginsplot, noci by(black) legend(cols(4)) title(Effets de educ et exper)

  Variables that uniquely identify margins: educ exper black
```



L'effet de l'éducation est clairement quadratique concernant la probabilité de travailler et plus d'éducation augmente en moyenne la probabilité de travailler. Le maximum de l'effet varie selon l'expérience et il faut l'explorer plus en détail afin de pouvoir l'identifier pour chaque année d'expérience supplémentaire. L'expérience a également un effet quadratique. En effet, celle-ci fait augmenter la probabilité de travailler et atteint son maximum à la sixième année. La différence entre les ethnies n'est pas flagrante, mais existe néanmoins. Encore une fois, l'ethnie ne change pas les effets en termes de rang, mais en termes d'amplitude. Notons néanmoins que les individus ayant à la fois des origines noires, un niveau d'éducation très élevé et plus de six ans d'expérience ont plus de probabilité de travailler que les individus d'origine non noire dans la même situation.

Afin d'utiliser la commande `pwcompare` et `contrast`, nous avons créé une variable catégorielle concernant l'éducation avec des points de coupure arbitraires.

```
histogram educc, discrete percent
(start=1, width=1)
```



Nous estimons un modèle multinomial logistique avec la variable catégorielle créée pour l'éducation.

```
mlogit status i.educc c.exper##c.exper i.black if year==87

Iteration 0:   log likelihood = -1199.7182
Iteration 1:   log likelihood = -960.17767
Iteration 2:   log likelihood = -912.27106
Iteration 3:   log likelihood = -911.35531
Iteration 4:   log likelihood = -911.35398
Iteration 5:   log likelihood = -911.35398

Multinomial logistic régression                   Number of obs   =       1717
                                                  LR chi2(12)     =     576.73
                                                  Prob > chi2     =     0.0000
Log likelihood = -911.35398                       Pseudo R2       =     0.2404

------------------------------------------------------------------------------
         status |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
1               |
          educc |
             2  |  -.0063821   .9232708    -0.01   0.994     -1.81596    1.803196
             3  |   .4832895   .6630833     0.73   0.466    -.8163298    1.782909
             4  |   1.960897    .633631     3.09   0.002     .7190029    3.202791
                |
          exper |  -.9461064   .1506675    -6.28   0.000    -1.241409   -.6508035
                |
c.exper#c.exper |   .0900361   .0224182     4.02   0.000     .0460972     .133975
                |
        1.black |  -.4020934   .2763706    -1.45   0.146    -.9437699    .1395831
```

```
          _cons |  -2.288546   .6479888    -3.53   0.000    -3.558581   -1.018512
----------------+----------------------------------------------------------------
2               |
          educc |
             2  |  -.5414335   .2332924    -2.32   0.020    -.9986782   -.0841889
             3  |  -.9504979    .180711    -5.26   0.000    -1.304685   -.5963108
             4  |  -2.431593    .227032   -10.71   0.000    -2.876567   -1.986618
                |
          exper |  -.9614668   .1002418    -9.59   0.000    -1.157937   -.7649965
                |
c.exper#c.exper |   .0659904   .0130813     5.04   0.000     .0403515    .0916294
                |
        1.black |   .4910117   .1489629     3.30   0.001     .1990497    .7829737
          _cons |   1.406806   .2202981     6.39   0.000     .9750301    1.838583
----------------+----------------------------------------------------------------
3               |  (base outcome)
------------------------------------------------------------------------------
```

Nous allons avant tout tester dans l'équation 1 avec l'option `eq(1)` si tous les coefficients associés à chaque modalité sont égaux avec la commande `contrast`.

```
contrast educc, eq(1)

Contrasts of marginal linear prédictions

Margins      : asbalanced

-------------------------------------------------
             |         df        chi2     P>chi2
-------------+-----------------------------------
1            |
       educc |          3       22.96     0.0000
-------------------------------------------------
```

Nous pouvons rejeter l'hypothèse nulle d'égalité des coefficients.

Nous testons toujours dans l'équation 1 si chaque catégorie est différente de la catégorie de base grâce au préfixe `r`.

```
contrast r.educc, eq(1)

Contrasts of marginal linear prédictions

Margins      : asbalanced

-------------------------------------------------
             |         df        chi2     P>chi2
-------------+-----------------------------------
1            |
       educc |
   (2 vs 1)  |          1        0.00     0.9945
   (3 vs 1)  |          1        0.53     0.4661
   (4 vs 1)  |          1        9.58     0.0020
      Joint  |          3       22.96     0.0000
-------------------------------------------------

--------------------------------------------------------------
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
1            |
       educc |
   (2 vs 1)  |  -.0063821   .9232708      -1.81596    1.803196
   (3 vs 1)  |   .4832895   .6630833     -.8163298    1.782909
   (4 vs 1)  |   1.960897    .633631      .7190029    3.202791
```

```
------------------------------------------------------------
```

Comme nous ne rejetons pas l'hypothèse nulle dans ces cas, il n'existe pas de différence entre le coefficient de la catégorie 2 et le coefficient de la catégorie 1 ainsi que le coefficient de la catégorie 3 et le coefficient de la catégorie 1. La quatrième catégorie est statistiquement différente de la première.

Nous testons maintenant le coefficient d'une catégorie par rapport au coefficient de la catégorie précédente avec le préfixe ar. toujours dans l'équation 1.

```
contrast ar.educc, eq(1)

Contrasts of marginal linear prédictions

Margins      : asbalanced

-------------------------------------------------
             |         df        chi2     P>chi2
-------------+-----------------------------------
1            |
       educc |
   (2 vs 1)  |          1        0.00     0.9945
   (3 vs 2)  |          1        0.39     0.5303
   (4 vs 3)  |          1       14.68     0.0001
      Joint  |          3       22.96     0.0000
-------------------------------------------------

------------------------------------------------------------
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
1            |
       educc |
   (2 vs 1)  |  -.0063821   .9232708      -1.81596    1.803196
   (3 vs 2)  |   .4896716   .7802512     -1.039593    2.018936
   (4 vs 3)  |   1.477607   .3856632      .7217214    2.233493
------------------------------------------------------------
```

Dans ce cas, seul le coefficient de la catégorie 4 est statistiquement différent du coefficient de la catégorie 3. Il semble que nous pourrions regrouper les catégories 1, 2 et 3 en une seule catégorie. Nous pouvons aussi utiliser la commande `pwcompare` qui nous donnera directement les tests pour des différences entre paires de coefficients en ajustant les erreurs standards par la méthode de Bonferroni.

```
pwcompare i.educc, mcompare(bonferroni) cimargins effects sort eq(1)

Pairwise comparisons of marginal linear prédictions

Margins      : asbalanced

------------------------------------------------------------
             |                                Unadjusted
             |     Margin   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
1            |
       educc |
          2  |  -2.495975   .7554331     -3.976597   -1.015353
          1  |  -2.489593   .6350248     -3.734219   -1.244967
```

```
          3  |  -2.006303   .4134356     -2.816622   -1.195985
          4  |  -.5286961   .2015509     -.9237285   -.1336637
-------------------------------------------------------------

---------------------------
             |    Number of
             |  Comparisons
-------------+-------------
1            |
       educc |            6
---------------------------

------------------------------------------------------------------------------
             |                            Bonferroni           Bonferroni
             |   Contrast   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
       educc |
     2 vs 1  |  -.0063821   .9232708    -0.01   1.000    -2.442208    2.429444
     3 vs 1  |   .4832895   .6630833     0.73   1.000    -1.266095    2.232674
     3 vs 2  |   .4896716   .7802512     0.63   1.000    -1.568832    2.548175
     4 vs 3  |   1.477607   .3856632     3.83   0.001     .4601287    2.495086
     4 vs 1  |   1.960897    .633631     3.09   0.012     .2892153    3.632578
     4 vs 2  |   1.967279   .7535801     2.61   0.054    -.0208593    3.955417
------------------------------------------------------------------------------
```

Il semble bien que nous puissions regrouper les catégories 1, 2 et 3 selon ces tests. Nous pourrions procéder par étape et regrouper la catégorie 1 et 2, tester pour une différence par paire et encore une fois regrouper si les coefficients ne sont pas statistiquement différents.

3.3.5 La commande mprobit

Description

La commande mprobit dans Stata sert à estimer un modèle probit multinomial ou régression probit polytomique et utilise la méthode du maximum de vraisemblance pour estimer les paramètres ($\hat{\beta}$) du modèle. Ce type de modèle doit être choisi quand la variable dépendante est discrète et composée de modalités qui ne peuvent pas être naturellement ordonnées. Stata traitera les valeurs de la variable dépendante telles qu'elles seront spécifiées par l'utilisateur. Ainsi, il est important de vérifier le codage de la variable dépendante en utilisant la commande `summarize`. La première catégorie est généralement codée 0. Les variables négatives seront considérées comme manquantes. Il faut donc faire attention à avoir une variable codée strictement positivement 0, 1, 2, 3, etc. Quand la variable dépendante ne prend que deux modalités, les paramètres estimations seront les mêmes qu'en utilisant la commande `probit`. Ce modèle est basé sur l'hypothèse que les termes d'erreur sont indépendants, aléatoires et suivent une loi normale standard.

Syntaxe

`mprobit [depvar] [indepvars] [if] [in] [weight] [, options]`

où `depvar` est la variable dépendante
`[indepvars]` est la variable indépendante
`[if]` et `[in]` sont des contraintes dans l'échantillon
`[weight]` assigne des poids aux observations
`[, options]` présente des options supplémentaires dans la régression

Menu

Statistics > Categorical outcomes > Independent multinomial probit

Option

Erreur type/Robust

`vce(typevce)` : le type vce peut être `robust`, `cluster`(*variablecluster*). Pour plus d'informations, le lecteur peut se référer à la documentation Stata.

Résultats

`baseoutcome(#)` : permets d'indiquer quelle catégorie doit être prise comme catégorie de référence (catégorie de base). Par défaut, c'est la catégorie qui a la fréquence la plus élevée qui est choisie par le logiciel comme catégorie de référence.

3.3.6 Les commandes postestimation après la commande mprobit

`test [[label ou valeur de la modalité]] [varexo]`: test de Wald avec lequel nous pouvons tester des restrictions linéaires (par exemple $\beta_1 = \beta_2$ ou $\beta_1 = 0$).

`listcoef [, const help]`: donne une liste des coefficients pour le dernier modèle estimé. L'option `const` inclut la constante dans le tableau et l'option `help` ajoute une légende des variables disponibles dans le tableau. Pour plus de renseignements concernant les options, l'utilisateur peut se référer à l'aide fournie avec le module en utilisant la commande `help listcoef`.

`contrast [varexo], [eq(label ou valeur de l'équation)]` : permets de tester si les moyennes des effets sont égales entre toutes les modalités d'une variable ordinale incluse dans la régression (x).

`contrast r.[varexo], [eq(label ou valeur de l'équation)]` : permet de tester si la moyenne des effets est égale entre les différentes modalités d'une variable catégorielle lorsqu'elles sont comparées à la catégorie de référence.

`contrast ar.[varexo], [eq(label ou valeur de l'équation)]` : permets de tester si le coefficient d'une modalité est différent du coefficient de la modalité qui le précède.

`fitstat` : indices de qualité du modèle (fit ou adéquation).

`lrtest [modèlecomplet] [modelcontraint][, stats]` : cette commande performe un test de ratio de vraisemblance (*likelihood ratio test*) entre un modèle complet et un modèle contraint. Ce test nous renseigne sur la significativité des coefficients non inclus dans le modèle contraint et est similaire au test-F joint qui est calculé avec la commande test. Le modèle contraint est donc un sous modèle du modèle complet. Ce test peut être utilisé pour un test de Chow si nous voulons savoir si les coefficients du modèle sont différents entre plusieurs tranches de l'échantillon.

```
margins,                                                                    [dydx(varexos)]
[at(varexo1=(valdébut(incrément)valfin)varexo2=(valdébut(incrément)valfin)] [predict(outcome(labe
l ou valeur de la modalité))]:
```

cette commande nous permet de calculer les effets marginaux des variables spécifiées et les prédictions du modèle. L'option `dydx(.)` est nécessaire pour calculer les effets marginaux de variables continues, l'option at nous sera utile pour utiliser la commande marginsplot afin d'avoir un aperçu graphique de ces effets. Dans le cas d'un modèle logistique multinomial, il est aussi nécessaire de spécifier pour quelle catégorie nous désirons les effets marginaux.

`marginsplot [varexos]` : cette commande doit être utilisée après margins pour avoir un aperçu graphique des effets marginaux sur les probabilités.

`prchange [varlist], [outcome(label ou valeur de la modalité)] [help]`: calcule l'impact d'une variation discrète ou marginale des variables spécifiées pour les modèles catégoriels et de comptage.

`predict p if e(sample) [, outcome(label ou valeur de la modalité)]` : prédit la probabilité que la variable dépendante prenne la valeur 1. La restriction if e(sample) est ajoutée afin de spécifier que les probabilités soient prédites pour les observations n'ayant aucune valeur manquante.

Dans ce cas-ci, il faut aussi spécifier la valeur de la modalité pour laquelle les probabilités sont prédites.

`pwcompare [varexo], [cimargins] [eq(label ou valeur de l'équation)]` : permets de comparer l'impact de chaque modalité d'une variable catégorielle entre elles et de tester si ces modalités ont un effet significativement différent.

3.3.7 Exemple

Nous allons estimer un modèle probit multinomial avec la commande `mlogit`.

```
mprobit status educ c.exper##c.exper i.black if year==87

Iteration 0:   log likelihood = -912.50781
Iteration 1:   log likelihood = -908.64535
Iteration 2:   log likelihood = -908.41824
Iteration 3:   log likelihood = -908.41773
Iteration 4:   log likelihood = -908.41773

Multinomial probit régression                     Number of obs   =       1717
                                                  Wald chi2(8)    =     480.82
Log likelihood = -908.41773                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
         status |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
1               |
           educ |    .162221   .0385226     4.21   0.000     .0867181    .2377239
          exper |  -.6721982   .1027514    -6.54   0.000    -.8735872   -.4708091
                |
c.exper#c.exper |   .0592846   .0142553     4.16   0.000     .0313448    .0872244
                |
        1.black |  -.2244359   .1795638    -1.25   0.211    -.5763746    .1275028
          _cons |  -2.985019   .6285942    -4.75   0.000     -4.21704   -1.752997
----------------+-------------------------------------------------------------
2               |
           educ |  -.2788583    .024389   -11.43   0.000    -.3266598   -.2310567
          exper |  -.7859898    .073674   -10.67   0.000    -.9303883   -.6415914
                |
c.exper#c.exper |     .05491    .009636     5.70   0.000     .0360238    .0737963
                |
        1.black |    .380267   .1145264     3.32   0.001     .1557994    .6047345
          _cons |   3.721129   .3604036    10.32   0.000     3.014751    4.427507
----------------+-------------------------------------------------------------
3               |  (base outcome)
------------------------------------------------------------------------------
```

Nous ne pouvons pas utiliser les ratios de risques relatifs par construction du modèle et il faut donc dériver les effets marginaux avec la commande margins. Nous ne nous attarderons pas à le faire dans ce cas la et nous référons le lecteur à la partie sur le modèle logistique multinomial.

3.4 Modèles logit et probit ordonnés

3.4.1 Rappel théorique

Un modèle logistique ordonné peut être nécessaire si le but de l'étude est de mesurer l'impact de variables exogènes sur une variable dépendante discrète ayant des modalités qui peuvent être naturellement ordonnées. Nous nous concentrerons sur le *proportional odds model*.

Nous avons une variable latente (non observée) $\tilde{y}$ répondant à cette relation

$$\tilde{y}_i = X_i\beta + u_i$$

Ou u_i peut suivre une loi logistique ou normale.

Nous observons

$$y_i = 0 \text{ si } \tilde{y}_i \leq 0$$

$$y_i = 1 \text{ si } 0 < \tilde{y}_i \leq \mu_1$$

$$y_i = 2 \text{ si } \mu_1 < \tilde{y}_i \leq \mu_2$$

$$\vdots$$

$$y_i = J \text{ si } \mu_{J-1} \leq \tilde{y}_i$$

Ou les μ sont des paramètres inconnus à estimer au même titre que les β.

Nous avons donc

$$\Pr(y = 0|x) = G(-x\beta),$$

$$\Pr(y = 1|x) = G(\mu_1 - x\beta) - (G(-x\beta)),$$

$$\Pr(y = 2|x) = G(\mu_2 - x\beta) - (G(\mu_1 - x\beta)),$$

$$\vdots$$

$$\Pr(y = J|x) = 1 - \left(G\left(\mu_{J-1} - x\beta\right)\right),$$

Ou $G(.)$ est la fonction de densité normale ou logistique.

Pour que les probabilités soient positives il faut que

$$0 < \mu_1 < \mu_2 < \cdots < \mu_{J-1}$$

Les effets marginaux des variables exogènes sur les probabilités d'appartenance à chaque modalité de la variable ne sont pas égaux aux coefficients. En prenant l'exemple de trois catégories et en supposant que les erreurs suivent une loi normale de moyenne 0 et de variance σ^2 nous pouvons réécrire

$$\Pr(y = 0|x) = 1 - \Phi(x\beta),$$

$$\Pr(y = 1|x) = \Phi(\mu_1 - x\beta) - (\Phi(-x\beta)),$$

$$\Pr(y = 2|x) = 1 - \big(\Phi(\mu_1 - x\beta)\big).$$

Les effets marginaux sont donc

$$\frac{\partial \Pr(y = 0|x)}{\partial x} = -\phi(x\beta)\beta,$$

$$\frac{\partial \Pr(y = 1|x)}{\partial x} = [\phi(-x\beta) - \phi(\mu_1 - x\beta)]\beta,$$

$$\frac{\partial \Pr(y = 2|x)}{\partial x} = \phi(\mu_1 - x\beta)\beta.$$

Ce modèle suppose que les relations sont les mêmes entre chaque paire de modalités de la variable dépendante d'où sa dénomination « proportionnal odds model ». Il faut donc comprendre que les coefficients qui décrivent la relation entre les deux premières modalités sont les mêmes qui décrivent la relation entre les deux dernières modalités ou encore les deux modalités extrêmes. Il n'y a donc qu'une équation estimée pour le modèle et l'hypothèse doit être testée afin de pouvoir valider l'interprétation des effets marginaux.

3.4.2 La commande ologit

Description

La commande ologit dans Stata sert à estimer un modèle logistique multinomial ordonné ou régression logistique polytomique ordonnée, et utilise la méthode du maximum de vraisemblance pour estimer les paramètres ($\hat{\beta}$) du modèle. Ce type de modèle doit être choisi

quand la variable dépendante est discrète et composée de modalités qui peuvent être naturellement ordonnées. Stata traitera les valeurs de la variable dépendante telles qu'elles seront spécifiées par l'utilisateur. Ainsi, il est important de vérifier le codage de la variable dépendante en utilisant la commande summarize (lien vers commande). La première catégorie est généralement codée 0. Les variables négatives seront considérées comme manquantes. Il faut donc faire attention à avoir une variable codée strictement positivement 0, 1, 2, 3, etc. Quand la variable dépendante ne prend que deux modalités, les paramètres estimations seront les même qu'en utilisant la commande logit.

Syntaxe

`ologit [depvar] [indepvars] [if] [in] [weight] [, options]`

où `depvar` est la variable dépendante
`[indepvars]` est la variable indépendante
`[if]` et `[in]` sont des contraintes dans l'échantillon
`[weight]` assigne des poids aux observations
`[, options]` présente des options supplémentaires dans la régression

Menu

Statistics > Ordinal outcomes > Ordered logistic régression

Option

Erreur type/Robust

`vce(typevce)` : le type vce peut être `robust`, `cluster`(*variablecluster*). Pour plus d'informations, le lecteur peut se référer à la documentation Stata.

Résultats

`or` : calcule les ratios de chances. Les erreurs types et les intervalles de confiance sont similairement transformés. `or` peut être spécifié lors de l'estimation ou simplement après l'estimation pour obtenir un nouveau tableau.

3.4.3 Les commandes postestimation après la commande ologit

`test [[label ou valeur de la modalité]] [varexo]`: test de Wald avec lequel nous pouvons tester des restrictions linéaires (par exemple $\beta_1 = \beta_2$ ou $\beta_1 = 0$).

`listcoef [, const help]` : donne une liste des coefficients pour le dernier modèle estimé. L'option const inclut la constante dans le tableau et l'option help rajoute une légende des variables disponibles dans le tableau. Pour plus de renseignements concernant les options, l'utilisateur peut se référer à l'aide fournie avec le module en utilisant la commande help listcoef.

`contrast [varexo]` : permet de tester si les moyennes des effets sont égales entre toutes les modalités d'une variable ordinale incluse dans la régression (x).

`contrast r.[varexo]` : permet de tester si les moyennes des effets sont égales entre les différentes modalités d'une variable ordinale quand elles sont comparées à la catégorie de référence.

`contrast ar.[varexo]` : permet de tester si le coefficient d'une modalité est différent du coefficient de la modalité qui le précède.

`fitstat` : indices de qualité du modèle (fit ou adéquation).

`lrtest [modèlecomplet] [modelcontraint][, stats]` : cette commande performe un test de ratio de vraisemblance (likelihood ratio test) entre un modèle complet et un modèle contraint. Ce test nous renseigne sur la significativité des coefficients non inclus dans le modèle contraint et est similaire au test-F joint qui est calculé avec la commande test. Le modèle contraint est donc un sous modèle du modèle complet. Ce test peut être utilisé pour un test de Chow si nous voulons savoir si nous voulons savoir si les coefficients du modèle sont différents entre plusieurs tranches de l'échantillon.

```
margins, [dydx(varexos)]
[at(varexo1=(valdébut(incrément)valfin)varexo2=(valdébut(incrément)valfin)] [predict(outcome(labe
l ou valeur de la modalité))]:
```

cette commande nous permet de calculer les effets marginaux des variables spécifiées et les prédictions du modèle. L'option dydx(.) est nécessaire pour calculer les effets marginaux de variables continues, l'option at nous servira pour pouvoir utiliser la commande marginsplot afin d'avoir un aperçu graphique de ces effets. Dans le cas d'un modèle logistique multinomial, il est aussi nécessaire de spécifier pour quelle catégorie nous voulons les effets marginaux.

`marginsplot [varexos]` : cette commande doit être utilisée après margins pour avoir un aperçu graphique des effets marginaux sur les probabilités.

`prchange [varlist], [outcome(label ou valeur de la modalité)] [help]`: calcul l'impact d'une variation discrète ou marginale des variables spécifiées pour les modèles catégoriels et de comptage.

`pwcompare [varexo], [cimargins]` : permet de comparer l'impact de chaque modalité d'une variable catégorielle entre elles et de tester si elles ont un effet significativement différent. Voir dans l'application pour avoir les différentes options.

Syntaxe pour predict

`predict p if e(sample), [outcome(label ou valeur de la modalité)]` : prédit la probabilité que la variable dépendante prenne la valeur 1. La restriction `if e(sample)` est ajoutée afin de spécifier que les probabilités soient prédites pour les observations n'ayant aucune valeur manquante. Dans ce cas-ci, il faut aussi spécifier la valeur de la modalité pour laquelle les probabilités sont prédites.

3.4.4 Exemple

Nous utiliserons pour cette application un fichier comportant des données sur la présence aux cours et la performance scolaire.

```
use "http://www.Stata.com/data/jwooldridge/eacsap/attend.dta", clear
```

Nous imaginerons ici que nous avons accès à une variable catégorielle concernant le GPA au terme d'une session d'un étudiant que nous voulons expliquer par le GPA de la session passée, le pourcentage de cours où l'étudiant était présent, le pourcentage de travaux rendus durant la session, et le fait ou non d'être un étudiant de première année.
Nous créons premièrement la variable catégorielle ordonnée concernant la mesure du GPA de la session.

```
gen termgpac=1 if termgpa>=0 & termgpa<=1
(657 missing values generated)

replace termgpac=2 if termgpa>1 & termgpa<=2
(110 real changes made)

replace termgpac=3 if termgpa>2 & termgpa<=3
(346 real changes made)

replace termgpac=4 if termgpa>3 & termgpa<=4
(201 real changes made)

replace termgpac=. if termgpa==.
(0 real changes made)
```

```
label variable termgpac "GPA , 1: entre 0 et 1, 2: entre 1 et 2, 3: entre 2 et 3, 4: entre 3 et
4"
```

Nous créons ensuite une variable catégorielle concernant le pourcentage de travaux rendus.

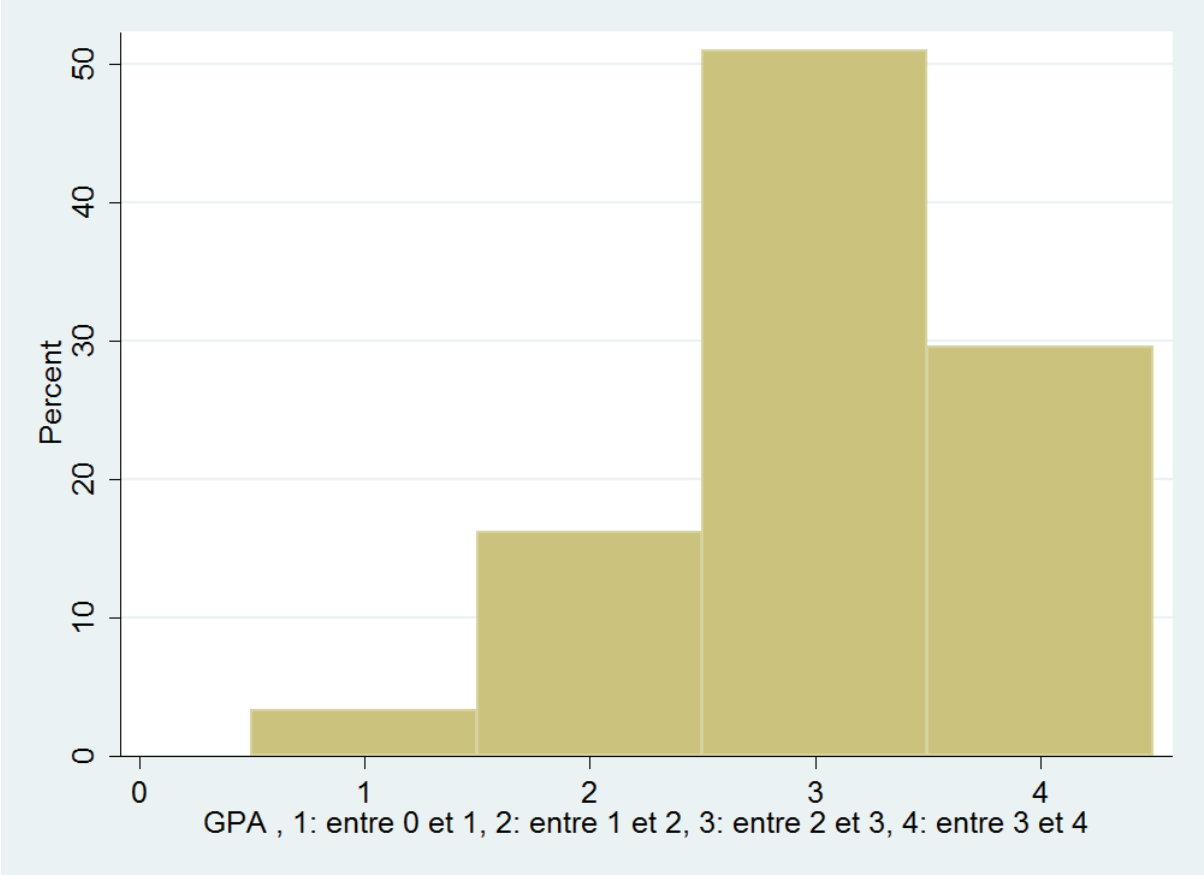
```
gen hwrtec=1 if hwrte>=0 & hwrte<=25
(664 missing values generated)

replace hwrtec=2 if hwrte>25 & hwrte<=50
(39 real changes made)

replace hwrtec=3 if hwrte>50 & hwrte<=75
(103 real changes made)

replace hwrtec=4 if hwrte>75 & hwrte<=100
(516 real changes made)

replace hwrtec=. if hwrte==.
(0 real changes made)
```

Description des données

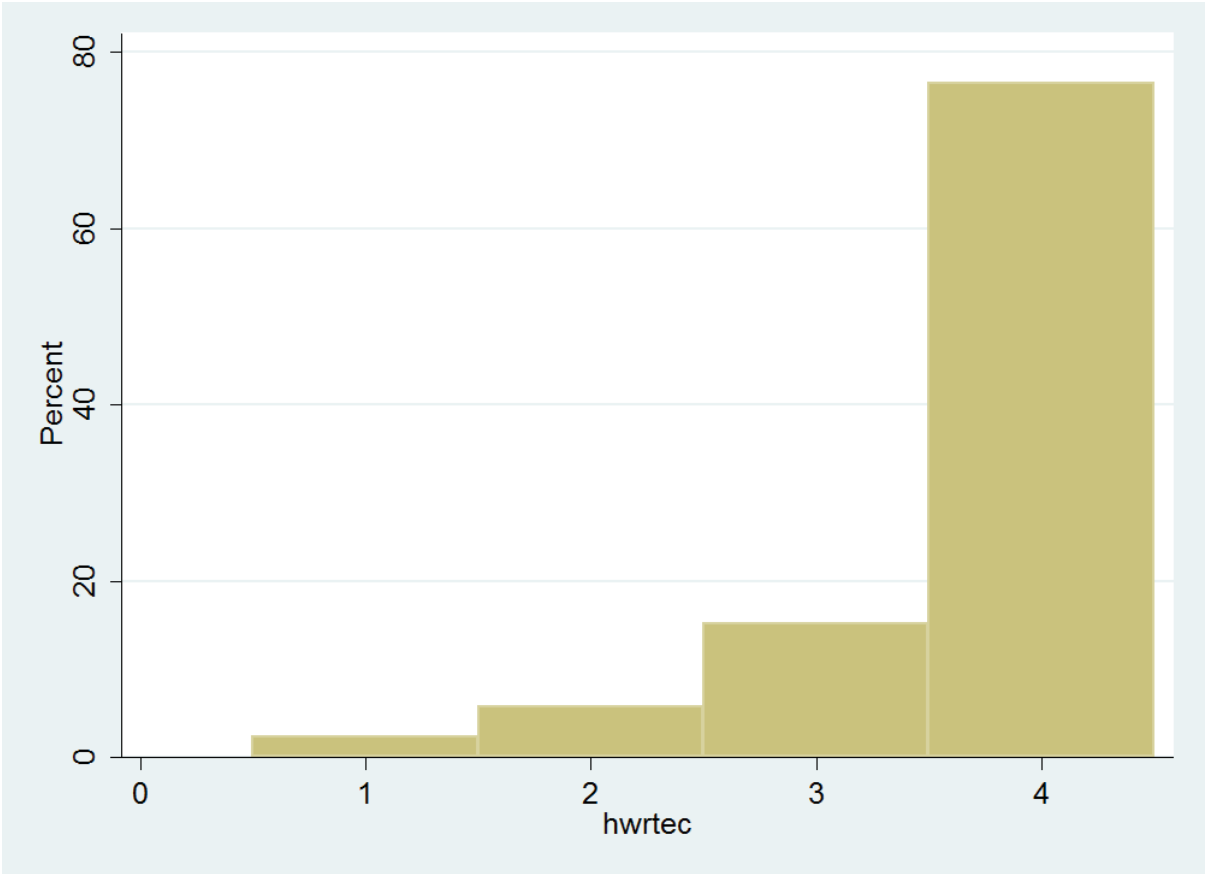
Nous explorons par la suite les données que nous avons sélectionnées

```
describe termgpac priGPA atndrte hwrtec frosh

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------------------
termgpac        float  %9.0g                  GPA , 1: entre 0 et 1, 2: entre 1 et 2, 3: entre 2
                                              et 3, 4: entre 3 et 4
priGPA          float  %9.0g                  cumulative GPA prior to term
atndrte         float  %9.0g                  percent classes attended
hwrtec          float  %9.0g
frosh           byte   %8.0g                  =1 if freshman

summarize termgpac priGPA atndrte hwrtec frosh

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    termgpac |       680    3.066176    .7675507          1          4
      priGPA |       680    2.586775    .5447141       .857       3.93
     atndrte |       680    81.70956    17.04699       6.25        100
      hwrtec |       674    3.660237    .6951267          1          4
       frosh |       680    .2323529    .4226438          0          1

inspect termgpac

termgpac:  GPA , 1: entre 0 et 1, 2: entre      Number of Observations
------------------------------------------- ------------------------------
                                            Total   Integers   Nonintegers
|          #                 Negative          -          -          -
|          #                 Zero              -          -          -
|          #                 Positive        680        680          -
|          #   #                           -----      -----      -----
|          #   #             Total           680        680          -
|  .   #   #   #             Missing           -
+----------------------                     -----
1                      4                      680
   (4 unique values)

. inspect hwrtec

hwrtec:                                         Number of Observations
---------                                   ------------------------------
                                            Total   Integers   Nonintegers
|              #             Negative          -          -          -
```

```
|              #             Zero               -         -          -
|              #             Positive         674       674          -
|              #                            -----     -----      -----
|              #             Total            674       674          -
|  .   .   #   #             Missing            6
+----------------------                     -----
1                      4                      680
   (4 unique values)

histogram termgpac, discrete percent
(start=1, width=1)
```




```
histogram hwrtec, discrete percent
(start=1, width=1)
```




```
tab1 termgpac priGPA atndrte hwrtec frosh

-> tabulation of termgpac

   GPA , 1: |
 entre 0 et |
1, 2: entre |
 1 et 2, 3: |
 entre 2 et |
3, 4: entre |
     3 et 4 |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         23        3.38        3.38
          2 |        110       16.18       19.56
          3 |        346       50.88       70.44
          4 |        201       29.56      100.00
------------+-----------------------------------
      Total |        680      100.00

-> tabulation of priGPA

 cumulative |
  GPA prior |
    to term |      Freq.     Percent        Cum.
------------+-----------------------------------
       .857 |          1        0.15        0.15
       1.19 |          1        0.15        0.29

                                .
                                .
                                .

       3.91 |          1        0.15       99.56
       3.92 |          1        0.15       99.71
```

```
       3.93 |          2        0.29      100.00
------------+-----------------------------------
      Total |        680      100.00

-> tabulation of atndrte

    percent |
    classes |
   attended |      Freq.     Percent        Cum.
------------+-----------------------------------
       6.25 |          1        0.15        0.15
      9.375 |          1        0.15        0.29
       12.5 |          1        0.15        0.44
     15.625 |          2        0.29        0.74
      18.75 |          2        0.29        1.03
     28.125 |          4        0.59        1.62
      31.25 |          7        1.03        2.65
     34.375 |          2        0.29        2.94
       37.5 |          5        0.74        3.68
     40.625 |          2        0.29        3.97
      43.75 |          4        0.59        4.56
     46.875 |          8        1.18        5.74
         50 |          8        1.18        6.91
     53.125 |         12        1.76        8.68
      56.25 |         12        1.76       10.44
     59.375 |         13        1.91       12.35
       62.5 |         10        1.47       13.82
     65.625 |         15        2.21       16.03
      68.75 |          9        1.32       17.35
     71.875 |         27        3.97       21.32
         75 |         34        5.00       26.32
     78.125 |         49        7.21       33.53
      81.25 |         42        6.18       39.71
     84.375 |         64        9.41       49.12
       87.5 |         72       10.59       59.71
     90.625 |         70       10.29       70.00
      93.75 |         67        9.85       79.85
     96.875 |         71       10.44       90.29
        100 |         66        9.71      100.00
------------+-----------------------------------
      Total |        680      100.00

-> tabulation of hwrtec

     hwrtec |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         16        2.37        2.37
          2 |         39        5.79        8.16
          3 |        103       15.28       23.44
          4 |        516       76.56      100.00
------------+-----------------------------------
      Total |        674      100.00

-> tabulation of frosh

      =1 if |
   freshman |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        522       76.76       76.76
          1 |        158       23.24      100.00
------------+-----------------------------------
      Total |        680      100.00
```

Les données disponibles dans l'échantillon ont l'air de bien se prêter au type d'analyse que nous voulons effectuer. Il y n'y a pas de données manquantes, les variables sont codées correctement, les fréquences sont acceptables concernant les variables catégorielles. Nous allons tout de même

effectuer des tests sur la variable hwrtec étant donné le faible nombre d'étudiants ayant rendu entre 0 % et 25 % de leurs devoirs.

Modélisation

Nous estimons un modèle logistique ordonné avec les variables sélectionnées auparavant avec la commande ologit.

```
ologit termgpac c.priGPA c.atndrte i.hwrtec i.frosh

Iteration 0:   log likelihood = -742.82548
Iteration 1:   log likelihood = -560.86181
Iteration 2:   log likelihood = -541.69528
Iteration 3:   log likelihood = -540.47955
Iteration 4:   log likelihood = -540.47575
Iteration 5:   log likelihood = -540.47575

Ordered logistic régression                     Number of obs   =        674
                                                LR chi2(6)      =     404.70
                                                Prob > chi2     =     0.0000
Log likelihood = -540.47575                     Pseudo R2       =     0.2724

------------------------------------------------------------------------------
    termgpac |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      priGPA |    2.69323   .2126047    12.67   0.000     2.276532    3.109927
     atndrte |   .0213841   .0066941     3.19   0.001     .0082639    .0345043
             |
      hwrtec |
          2  |   1.027423   .6174534     1.66   0.096    -.1827632     2.23761
          3  |   1.371838   .5838661     2.35   0.019     .2274818    2.516195
          4  |   2.441417   .5835358     4.18   0.000     1.297708    3.585126
             |
     1.frosh |   .3122286   .2033122     1.54   0.125    -.0862559    .7107131
-------------+----------------------------------------------------------------
       /cut1 |   5.879737   .7269155                      4.455009    7.304465
       /cut2 |   8.753143   .7835211                       7.21747    10.28882
       /cut3 |   12.33687   .8569134                      10.65735    14.01639
------------------------------------------------------------------------------
```

Nous remarquons que le modèle semble bien performer avec un pseudo R^2 plutôt élevé et la plupart des coefficients significatifs. Nous ne pouvons pas directement interpréter les coefficients et nous nous contenterons donc de sauvegarder les estimations avec la commande estimates store sous le nom m1.

```
estimates store m1
```

Tests des hypothèses de modélisation

Avant d'interpréter les coefficients, il nous faut vérifier une des hypothèses de notre modèle qui est la proportionnalité des coefficients entre chaque paire de modalités. Nous utilisons premièrement la commande omodel pour tester cette hypothèse.

```
omodel logit termgpac priGPA atndrte hwrtec frosh
```

```
Iteration 0:   log likelihood = -742.82548
Iteration 1:   log likelihood = -559.69789
Iteration 2:   log likelihood = -542.13417
Iteration 3:   log likelihood = -541.40859
Iteration 4:   log likelihood =  -541.4052

Ordered logit estimates                           Number of obs   =        674
                                                  LR chi2(4)      =     402.84
                                                  Prob > chi2     =     0.0000
Log likelihood =  -541.4052                       Pseudo R2       =     0.2712

------------------------------------------------------------------------------
    termgpac |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      priGPA |   2.698716   .2123894    12.71   0.000     2.282441    3.114992
     atndrte |   .0205193   .0065969     3.11   0.002     .0075897    .0334489
      hwrtec |   .8286642   .1487421     5.57   0.000      .537135    1.120193
       frosh |   .2949104   .2027187     1.45   0.146     -.102411    .6922318
-------------+----------------------------------------------------------------
       _cut1 |   6.703632   .6226882          (Ancillary parameters)
       _cut2 |   9.598665   .6841287
       _cut3 |   13.16389   .7800429
------------------------------------------------------------------------------

Approximate likelihood-ratio test of proportionality of odds
across response categories:
         chi2(8) =      5.38
       Prob > chi2 =    0.7167
```

L'hypothèse nulle est ici il y égalité des coefficients entre toutes les catégories. Deux modèles sont estimations, premièrement un modèle à proportional odds et deuxièmement un modèle généralisé permettant des différences de coefficients pour chaque catégorie. Si les coefficients ne sont pas significativement différents entre les deux estimations, alors il convient d'utiliser un modèle à proportional odds, sinon. Nous n'avons pas assez d'évidences statistiques pour rejeter l'hypothèse nulle et concluons que les coefficients sont égaux entre toutes les catégories.

Le deuxième test est disponible avec la commande brant du module SPost. Nous spécifions l'option detail pour avoir accès aux coefficients des deux équations utilisées dans le test de Brant. Celui-ci consiste à obtenir les estimations de modèles logit séparés pour chaque modalité de la variable dépendante et à examiner la restriction que les coefficients de toutes les équations sont égaux.

```
brant, detail

Estimated coefficients from j-1 binary régressions

                y>1         y>2         y>3
 priGPA   2.6891301   2.3734193   2.9124478
atndrte   .02410456   .01642564   .02610233
 hwrtec   1.0815843    .8196268   .71579972
  frosh  -.12007495   .50860572   .22397263
  _cons  -7.3269602  -8.5398172  -13.796245

Brant Test of Parallel Régression Assumption

    Variable |      chi2   p>chi2    df
-------------+--------------------------
         All |      6.86    0.552     8
```

```
-------------+--------------------------
      priGPA |      1.83    0.400     2
     atndrte |      0.69    0.707     2
      hwrtec |      0.89    0.642     2
       frosh |      1.38    0.502     2
---------------------------------------

A significant test statistic provides evidence that the parallel
régression assumption has been violated.
```

Nous tirons la même conclusion que pour le test précédent ou l'hypothèse présupposée est vérifiée. Au cas où celle-ci ne l'est pas, il est possible d'utiliser la commande mlogit pour estimer le modèle. L'utilisateur expérimenté peut aussi utiliser un modèle ordonné généralisé et installer le module Stata gologit2.

Mesures d'adéquation

Nous allons maintenant obtenir les mesures d'adéquations du modèle grâce à la commande fitstat et sauvegarder ces mesures sous le nom m2.

```
fitstat, saving(m2)

Measures of Fit for ologit of termgpac

Log-Lik Intercept Only:       -742.825   Log-Lik Full Model:           -541.405
D(667):                       1082.810   LR(4):                         402.841
                                         Prob > LR:                       0.000
McFadden's R2:                   0.271   McFadden's Adj R2:               0.262
ML (Cox-Snell) R2:               0.450   Cragg-Uhler(Nagelkerke) R2:      0.506
McKelvey & Zavoina's R2:         0.522
Variance of y*:                  6.885   Variance of error:               3.290
Count R2:                        0.662   Adj Count R2:                    0.305
AIC:                             1.627   AIC*n:                        1096.810
BIC:                         -3261.514   BIC':                         -376.788
BIC used by Stata:            1128.403   AIC used by Stata:            1096.810

(Indices saved in matrix fs_m2)
```

Ces indices sont le plus souvent utilisés afin de sélectionner un meilleur modèle (nouvelles variables ou restrictions linéaires).

Tests d'inférence

Nous pouvons aussi tester des restrictions linéaires avec la commande test. Nous voulons ici savoir si les coefficients de la variable hwrtec et de la variable frosh sont conjointement 0.

```
test 2.hwrtec 3.hwrtec 4.hwrtec 1.frosh

 ( 1)  [termgpac]2.hwrtec = 0
 ( 2)  [termgpac]3.hwrtec = 0
 ( 3)  [termgpac]4.hwrtec = 0
 ( 4)  [termgpac]1.frosh = 0

           chi2(  4) =   34.15
         Prob > chi2 =    0.0000
```

Nous pouvons rejeter l'hypothèse nulle de nullité jointe des coefficients des variables hwrtec et frosh.

Nous pouvons aussi tester ces restrictions à l'aide de la commande `lrtest` avec un test du rapport de vraisemblance. Nous estimons donc le modèle contraint.

```
ologit termgpac c.priGPA c.atndrte  if hwrtec~=. & frosh~=.

Iteration 0:   log likelihood = -742.82548
Iteration 1:   log likelihood = -570.93953
Iteration 2:   log likelihood = -558.39117
Iteration 3:   log likelihood = -558.24512
Iteration 4:   log likelihood = -558.24499
Iteration 5:   log likelihood = -558.24499

Ordered logistic régression                       Number of obs   =        674
                                                  LR chi2(2)      =     369.16
                                                  Prob > chi2     =     0.0000
Log likelihood = -558.24499                       Pseudo R2       =     0.2485

------------------------------------------------------------------------------
    termgpac |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      priGPA |   2.557013    .197955    12.92   0.000     2.169029    2.944998
     atndrte |    .040732   .0056017     7.27   0.000     .0297528    .0517112
-------------+----------------------------------------------------------------
       /cut1 |   5.184619   .5435904                      4.119202    6.250037
       /cut2 |   7.839547   .5758625                      6.710877    8.968216
       /cut3 |   11.30293   .6710185                      9.987759     12.6181
------------------------------------------------------------------------------
```

Nous sauvegardons les coefficients du modèle afin de pouvoir tester la significativité jointe des coefficients.

```
estimates store mtest
```

Nous performons enfin le test du rapport de vraisemblance qui nous donne le même résultat que précédemment. Il faut faire attention de spécifier les modèles dans (modèle complet ; modèle contraint) le bon ordre afin que le test puisse être effectué correctement,

```
lrtest m1 mtest

Likelihood-ratio test                                 LR chi2(4)  =     35.54
(Assumption: mtest nested in m1)                      Prob > chi2 =    0.0000
```

Interprétation des relations entre variable dépendante et variables indépendantes

Nous estimons à nouveau donc le modèle en termes d'odds ratio pour pouvoir donner une interprétation aux coefficients.

```
ologit, or
```

```
Ordered logistic régression                     Number of obs   =        674
                                                LR chi2(6)      =     404.70
                                                Prob > chi2     =     0.0000
Log likelihood = -540.47575                     Pseudo R2       =     0.2724

------------------------------------------------------------------------------
    termgpac | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      priGPA |   14.77933   3.142155    12.67   0.000     9.742835    22.41941
     atndrte |   1.021614   .0068388     3.19   0.001     1.008298    1.035106
             |
      hwrtec |
          2  |   2.793858   1.725077     1.66   0.096     .8329654    9.370906
          3  |   3.942592   2.301946     2.35   0.019     1.255435    12.38139
          4  |   11.48931   6.704425     4.18   0.000     3.660896    36.05791
             |
     1.frosh |   1.366467   .2778194     1.54   0.125     .9173594    2.035442
-------------+----------------------------------------------------------------
       /cut1 |   5.879737   .7269155                      4.455009    7.304465
       /cut2 |   8.753143   .7835211                       7.21747    10.28882
       /cut3 |   12.33687   .8569134                      10.65735    14.01639
------------------------------------------------------------------------------
```

Nous voyons que pour une variation d'un point du GPA passé, un étudiant a 14,77 fois plus de chance d'avoir un GPA de session entre 3 et 4 contre toutes les autres alternatives combinées c'est-à-dire d'avoir un GPA passé entre 0 et 3. De même, si un étudiant rend entre 75 % et 100 % de ses travaux, il a 11,75 fois plus de chances d'avoir un GPA de session entre 2 et 4 plutôt que d'avoir un GPA de session entre 0 et 2. Il a aussi 11,75 fois plus de chances d'avoir un GPA entre 1 et 4 qu'un GPA entre 0 et 1.

Par la suite, nous estimons de nouveau le modèle sans les opérateurs afin de pouvoir utiliser les commandes du module SPost

```
quietly ologit termgpac priGPA atndrte hwrtec frosh
```

Nous regardons par la suite une liste des coefficients avec la commande listcoef.

```
listcoef, help

ologit (N=674): Factor Change in Odds

  Odds of: >m vs <=m

----------------------------------------------------------------------
    termgpac |      b         z     P>|z|    e^b    e^bStdX      SDofX
-------------+--------------------------------------------------------
      priGPA |   2.69872   12.706   0.000  14.8606   4.3392     0.5438
     atndrte |   0.02052    3.110   0.002   1.0207   1.4029    16.4997
      hwrtec |   0.82866    5.571   0.000   2.2903   1.7790     0.6951
       frosh |   0.29491    1.455   0.146   1.3430   1.1322     0.4211
----------------------------------------------------------------------
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-valeur  for z-test
     e^b = exp(b) = factor change in odds for unit increase in X
 e^bStdX = exp(b*SD of X) = change in odds for SD increase in X
   SDofX = standard deviation of X
```

Nous pouvons voir grâce à la colonne e^bStdX une variation normalisée de chaque variable, nous permettant d'avoir un ordre d'idée de l'effet de passer d'une tranche de la distribution de chaque variable à une autre.

Nous regardons aussi une liste des prédictions des probabilités à différentes valeurs des variables indépendantes.

```
prchange, help

ologit: Changes in Probabilities for termgpac

priGPA
            Avg|Chg|           1           2           3           4
Min->Max   .46054661  -.40230357  -.51878965   .02066833   .90042491
   -+1/2   .21522124  -.02186975  -.25279684  -.15577587   .43044248
  -+sd/2   .11783493  -.00984308  -.12907328  -.09675348   .23566987
MargEfct   .21628571  -.01662868  -.22924072  -.18670201   .43257142

atndrte
            Avg|Chg|           1           2           3           4
Min->Max   .13805744  -.02444052  -.25167438   .06059951   .21551535
   -+1/2    .0016445  -.00012644  -.00174301  -.00141954     .003289
  -+sd/2   .02713859  -.00209573  -.02881604  -.02336538    .0542772
MargEfct    .0016445  -.00012643    -.001743  -.00141956     .003289

hwrtec
            Avg|Chg|           1           2           3           4
Min->Max   .21358488    -.048843  -.37832678   .20467332   .22249643
   -+1/2    .0664735   -.0052477  -.07121426    -.056485   .13294703
  -+sd/2   .04618754  -.00359673  -.04920951  -.03956884   .09237508
MargEfct    .0664124  -.00510598  -.07039036  -.05732847   .13282481

frosh
        Avg|Chg|           1           2           3           4
0->1   .02476376  -.00168553  -.02356001  -.02428198   .04952753

                 1          2          3          4
Pr(y|x)  .00620014  .09518485   .6981349  .20048013

        priGPA  atndrte   hwrtec    frosh
   x=  2.59182  82.1217  3.66024   .22997
sd_x=  .543844  16.4997  .695127  .421126

 Pr(y|x): probability of observing each y for specified x values
Avg|Chg|: average of absolute value of the change across categories
Min->Max: change in predicted probability as x changes from its minimum to
          its maximum
    0->1: change in predicted probability as x changes from 0 to 1
   -+1/2: change in predicted probability as x changes from 1/2 unit below
          base value to 1/2 unit above
  -+sd/2: change in predicted probability as x changes from 1/2 standard
          dev below base to 1/2 standard dev above
MargEfct: the partial derivative of the predicted probability/rate with
          respect to a given independent variable
```

Encore une fois, cette commande nous donne accès à des prédictions de probabilité d'appartenance aux catégories pour des valeurs spécifiques des variables. Ainsi, si un étudiant rend tous ses devoirs, il peut s'attendre à une variation moyenne de 22,25 % de probabilité supplémentaire d'avoir un GPA de session entre 3 et 4.

Nous allons maintenant estimer de nouveau le modèle avec les opérateurs afin de pouvoir dériver les effets marginaux.

```
quietly ologit termgpac c.priGPA c.atndrte i.hwrtec i.frosh
```

Nous commençons par les effets marginaux moyens des variables pour l'appartenance au groupe d'étudiants ayant un GPA entre 0 et 1. Nous utilisons donc la commande `margins` et spécifions les options `dydx` pour avoir la dérivée de notre fonction par rapport à toutes nos variables `(*)` et `predict(outcome(1))` pour la première catégorie.

```
margins, dydx(*) predict(outcome(1))

Average marginal effects                        Number of obs   =        674
Model VCE    : OIM

Expression   : Pr(termgpac==1), predict(outcome(1))
dy/dx w.r.t. : priGPA atndrte 2.hwrtec 3.hwrtec 4.hwrtec 1.frosh

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      priGPA |  -.0648998   .0113075    -5.74   0.000     -.087062   -.0427375
     atndrte |  -.0005153   .0001739    -2.96   0.003    -.0008562   -.0001744
             |
      hwrtec |
          2  |  -.0583982   .0418552    -1.40   0.163     -.140433    .0236365
          3  |  -.0707993   .0413905    -1.71   0.087    -.1519232    .0103247
          4  |  -.0941023   .0415101    -2.27   0.023    -.1754607   -.0127439
             |
     1.frosh |  -.0072674   .0047004    -1.55   0.122      -.01648    .0019451
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Nous voyons que les effets sont faibles, mais significatifs en tolérant une erreur de première espèce de 10 % (excepté pour la 2e modalité de la variable hwrtec) concernant la première modalité de la variable dépendante. Une augmentation du GPA passé réduit la probabilité d'appartenir à cette catégorie de 6,5 % par point de GPA tandis que rendre tous ses travaux réduit la probabilité d'avoir un GPA de session entre 0 et 1 de 9,4 %. Nous pouvons noter qu'à la base, aucun individu n'a trop de chance d'appartenir à cette modalité ce qui explique les effets faibles des variables.

Nous pouvons obtenir ces effets marginaux moyens pour toutes les modalités de la variable dépendante et nous laissons le soin au lecteur d'essayer de les calculer et de les interpréter lui-même.

Nous interprétons donc directement l'effet marginal à la moyenne pour la modalité 4 que nous obtenons en spécifiant l'option `atmeans` en plus à la commande précédente.

```
margins, dydx(*) atmeans  predict(outcome(4))

Conditional marginal effects                      Number of obs   =        674
Model VCE    : OIM

Expression   : Pr(termgpac==4), predict(outcome(4))
dy/dx w.r.t. : priGPA atndrte 2.hwrtec 3.hwrtec 4.hwrtec 1.frosh
at           : priGPA          =     2.59182 (mean)
               atndrte         =    82.12166 (mean)
               1.hwrtec        =    .0237389 (mean)
               2.hwrtec        =    .0578635 (mean)
               3.hwrtec        =     .152819 (mean)
               4.hwrtec        =    .7655786 (mean)
               0.frosh         =    .7700297 (mean)
               1.frosh         =    .2299703 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      priGPA |   .4298531   .0379372    11.33   0.000     .3554976    .5042086
     atndrte |    .003413   .0010727     3.18   0.001     .0013105    .0055155
             |
      hwrtec |
          2  |   .0472665   .0283118     1.67   0.095    -.0082237    .1027566
          3  |   .0751918   .0250493     3.00   0.003     .0260961    .1242876
          4  |     .22364   .0266078     8.41   0.000     .1714897    .2757904
             |
     1.frosh |   .0523629   .0357637     1.46   0.143    -.0177328    .1224585
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Quand toutes les variables sont à leur moyenne, tous les coefficients sont significatifs à 10 %. L'effet d'avoir un GPA passé d'un point supérieur a une influence positive de 42,9 % sur la probabilité d'avoir une GPA de session entre 3 et 4. Un étudiant ayant rendu tous ses travaux aura une probabilité supérieure de 22,4 % d'avoir un GPA entre 3 et 4.

Nous allons maintenant comparer l'effet à la moyenne à l'effet à la médiane du GPA passé et du pourcentage de cours pour lesquels l'étudiant était présent.

```
margins, dydx(*) at((median)c.priGPA c.atndrte) predict(outcome(4))

Average marginal effects                          Number of obs   =        674
Model VCE    : OIM

Expression   : Pr(termgpac==4), predict(outcome(4))
dy/dx w.r.t. : priGPA atndrte 2.hwrtec 3.hwrtec 4.hwrtec 1.frosh
at           : priGPA          =        2.56 (median)
               atndrte         =        87.5 (median)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      priGPA |   .4459039   .0344244    12.95   0.000     .3784333    .5133745
     atndrte |   .0035405   .0011653     3.04   0.002     .0012565    .0058244
             |
      hwrtec |
          2  |   .0488216   .0296825     1.64   0.100    -.0093549    .1069982
          3  |   .0775471   .0258678     3.00   0.003     .0268472     .128247
          4  |   .2289231    .025708     8.90   0.000     .1785363    .2793098
             |
     1.frosh |   .0540097   .0364346     1.48   0.138    -.0174009    .1254202
```

```
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Nous pouvons remarquer que l'effet à la médiane observé pour les variables priGPA et la modalité 4 de hwrtec a un effet un peu plus grand sur la probabilité d'avoir un GPA de session entre 3 et 4.

Afin de pouvoir mieux observer ces effets, nous allons calculer les effets marginaux de la variable atndre(pourcentage de cours où l'étudiant était présent) quand priGPA (le GPA de passé) varie en différenciant par rapport à la variable hwrtec(pourcentage de travaux rendus). Nous voulons faire varier priGPA de 1 à 4 par tranche de 0.2 et utilisons donc l'option `at(priGPA=(1(0.2)4))`.

Premièrement quand toutes les variables sont à leur moyenne avec l'option `atmeans`.

```
margins i.hwrtec, dydx(atndrte) at(priGPA=(1(0.2)4)) atmeans predict(outcome(1))

Conditional marginal effects                      Number of obs   =        674
Model VCE    : OIM

Expression   : Pr(termgpac==1), predict(outcome(1))
dy/dx w.r.t. : atndrte

1._at        : priGPA          =           1
               atndrte         =    82.12166 (mean)
               1.hwrtec        =    .0237389 (mean)
               2.hwrtec        =    .0578635 (mean)
               3.hwrtec        =     .152819 (mean)
               4.hwrtec        =    .7655786 (mean)
               0.frosh         =    .7700297 (mean)
               1.frosh         =    .2299703 (mean)

2._at        : priGPA          =         1.2
               atndrte         =    82.12166 (mean)
               1.hwrtec        =    .0237389 (mean)
               2.hwrtec        =    .0578635 (mean)
               3.hwrtec        =     .152819 (mean)
               4.hwrtec        =    .7655786 (mean)
               0.frosh         =    .7700297 (mean)
               1.frosh         =    .2299703 (mean)

                              .
                              .
                              .

15._at       : priGPA          =         3.8
               atndrte         =    82.12166 (mean)
               1.hwrtec        =    .0237389 (mean)
               2.hwrtec        =    .0578635 (mean)
               3.hwrtec        =     .152819 (mean)
               4.hwrtec        =    .7655786 (mean)
               0.frosh         =    .7700297 (mean)
               1.frosh         =    .2299703 (mean)

16._at       : priGPA          =           4
               atndrte         =    82.12166 (mean)
               1.hwrtec        =    .0237389 (mean)
               2.hwrtec        =    .0578635 (mean)
               3.hwrtec        =     .152819 (mean)
               4.hwrtec        =    .7655786 (mean)
               0.frosh         =    .7700297 (mean)
               1.frosh         =    .2299703 (mean)
```

```
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
atndrte      |
  _at#hwrtec |
        1 1  |  -.0034783   .0020224    -1.72   0.085    -.0074421    .0004855
        1 2  |  -.0052021   .0019156    -2.72   0.007    -.0089566   -.0014475
        1 3  |  -.0053458   .0016675    -3.21   0.001     -.008614   -.0020776
        1 4  |  -.0040411   .0012893    -3.13   0.002    -.0065681   -.0015142
        2 1  |  -.0045393   .0021211    -2.14   0.032    -.0086965   -.0003821
        2 2  |  -.0052889   .0015157    -3.49   0.000    -.0082596   -.0023183
        2 3  |  -.0049588   .0014228    -3.49   0.000    -.0077474   -.0021701
        2 4  |  -.0029463   .0010056    -2.93   0.003    -.0049171   -.0009754

                                              .
                                              .
                                              .

       15 1  |   -.000044   .0000294    -1.49   0.135    -.0001017    .0000137
       15 2  |  -.0000158   8.58e-06    -1.84   0.066    -.0000326    1.04e-06
       15 3  |  -.0000112   6.23e-06    -1.80   0.072    -.0000234    1.01e-06
       15 4  |  -3.84e-06   2.25e-06    -1.71   0.087    -8.24e-06    5.58e-07
       16 1  |  -.0000257   .0000179    -1.44   0.151    -.0000608    9.38e-06
       16 2  |  -9.22e-06   5.30e-06    -1.74   0.082    -.0000196    1.18e-06
       16 3  |  -6.53e-06   3.85e-06    -1.70   0.090    -.0000141    1.01e-06
       16 4  |  -2.24e-06   1.38e-06    -1.62   0.105    -4.95e-06    4.68e-07
------------------------------------------------------------------------------
```

Nous demandons le graphique des effets marginaux et utilisons des options graphiques afin que celui-ci soit plus facilement interprétable. L'option recast(line) nous permet de représenter l'effet marginal par une ligne. L'option recastci(rline) ciopts(lpattern(dash)) nous permet de représenter l'intervalle de confiance par une ligne discontinue. Nous spécifions un titre avec l'option `title()` une ligne horizontale pour y=0 avec l'option `yline(0)` des labels pour les x placés à la verticale pour une meilleure lisibilité avec l'option `xlabel(, angle(90))`, des labels pour les y placés à l'horizontale avec l'option `ylabel(, angle(0))` et nous supprimons le titre sur l'axe des y avec l'option `title("")`.

```
marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) title(Effet marginal à la
moyenne de atndrte) yline(0) xlabel(, angle(90)) ylabel(, angle(0)) ytitle("")

  Variables that uniquely identify margins: priGPA hwrtec
```

Nous sauvegardons le graphique sur le disque avec la commande `graph save` afin de pouvoir le combiner avec celui des effets marginaux moyens estimations ci-dessous.

```
graph save marginsologit1.gph, replace
(file marginsologit1.gph saved)
```

Nous estimons maintenant les effets marginaux moyens.

```
margins i.hwrtec, dydx(atndrte) at(priGPA=(1(0.2)4)) predict(outcome(1))

Average marginal effects                        Number of obs   =        674
Model VCE    : OIM

Expression   : Pr(termgpac==1), predict(outcome(1))
```

```
dy/dx w.r.t. : atndrte

1._at        : priGPA          =           1

2._at        : priGPA          =         1.2

                               .
                               .
                               .

15._at       : priGPA          =         3.8

16._at       : priGPA          =           4

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
atndrte      |
  _at#hwrtec |
        1 1  |  -.0034967   .0019779    -1.77   0.077    -.0073734    .0003801
        1 2  |  -.0050634   .0017652    -2.87   0.004    -.0085231   -.0016037
        1 3  |  -.0051761   .0015273    -3.39   0.001    -.0081696   -.0021826
        1 4  |  -.0039733   .0012124    -3.28   0.001    -.0063495   -.0015971
        2 1  |  -.0044792   .0020029    -2.24   0.025    -.0084049   -.0005535
        2 2  |   -.005117   .0013827    -3.70   0.000     -.007827    -.002407
        2 3  |  -.0048105   .0013064    -3.68   0.000    -.0073709   -.0022501
        2 4  |  -.0029574   .0009862    -3.00   0.003    -.0048904   -.0010245

                                             .
                                             .
                                             .

       15 1  |  -.0000478   .0000326    -1.46   0.143    -.0001117    .0000162
       15 2  |  -.0000172   9.58e-06    -1.79   0.073    -.0000359    1.62e-06
       15 3  |  -.0000122   7.09e-06    -1.72   0.086    -.0000261    1.72e-06
       15 4  |  -4.18e-06   2.57e-06    -1.63   0.104    -9.22e-06    8.57e-07
       16 1  |  -.0000279   .0000199    -1.41   0.159    -.0000669     .000011
       16 2  |    -.00001   5.92e-06    -1.69   0.090    -.0000216    1.57e-06
       16 3  |  -7.11e-06   4.37e-06    -1.63   0.104    -.0000157    1.45e-06
       16 4  |  -2.44e-06   1.58e-06    -1.55   0.122    -5.53e-06    6.50e-07
------------------------------------------------------------------------------
```

Les options du graphique précédent seront aussi utilisées pour le prochain.

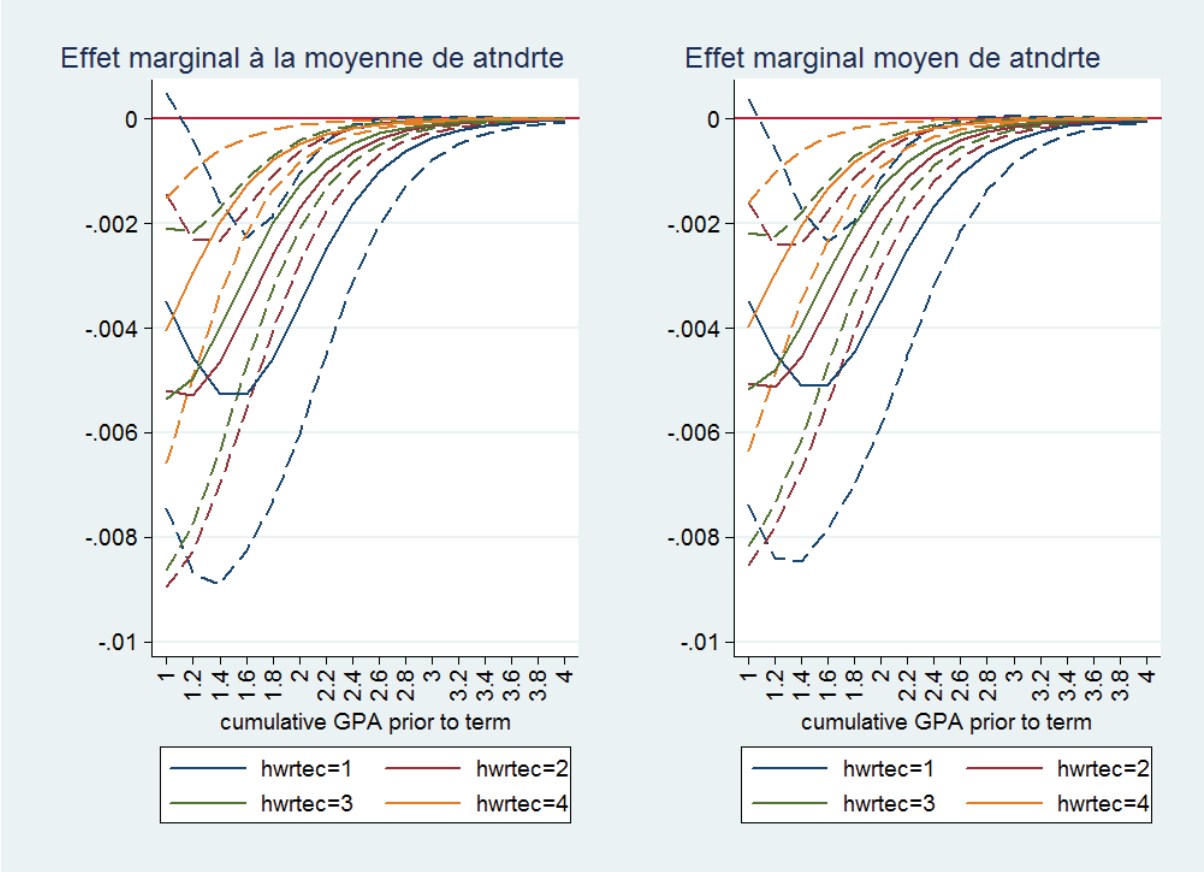
```
marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) title(Effet marginal moyen de
atndrte) yline(0) xlabel(, angle(90)) ylabel(, angle(0)) ytitle("")

  Variables that uniquely identify margins: priGPA hwrtec
```

Nous le sauvegardons de la même manière.

```
graph save marginsologit2.gph, replace
(file marginsologit2.gph saved)
```

Nous combinons les deux graphiques grâce à la commande `graph combine`.

```
graph combine marginsologit1.gph marginsologit2.gph, ycommon
```



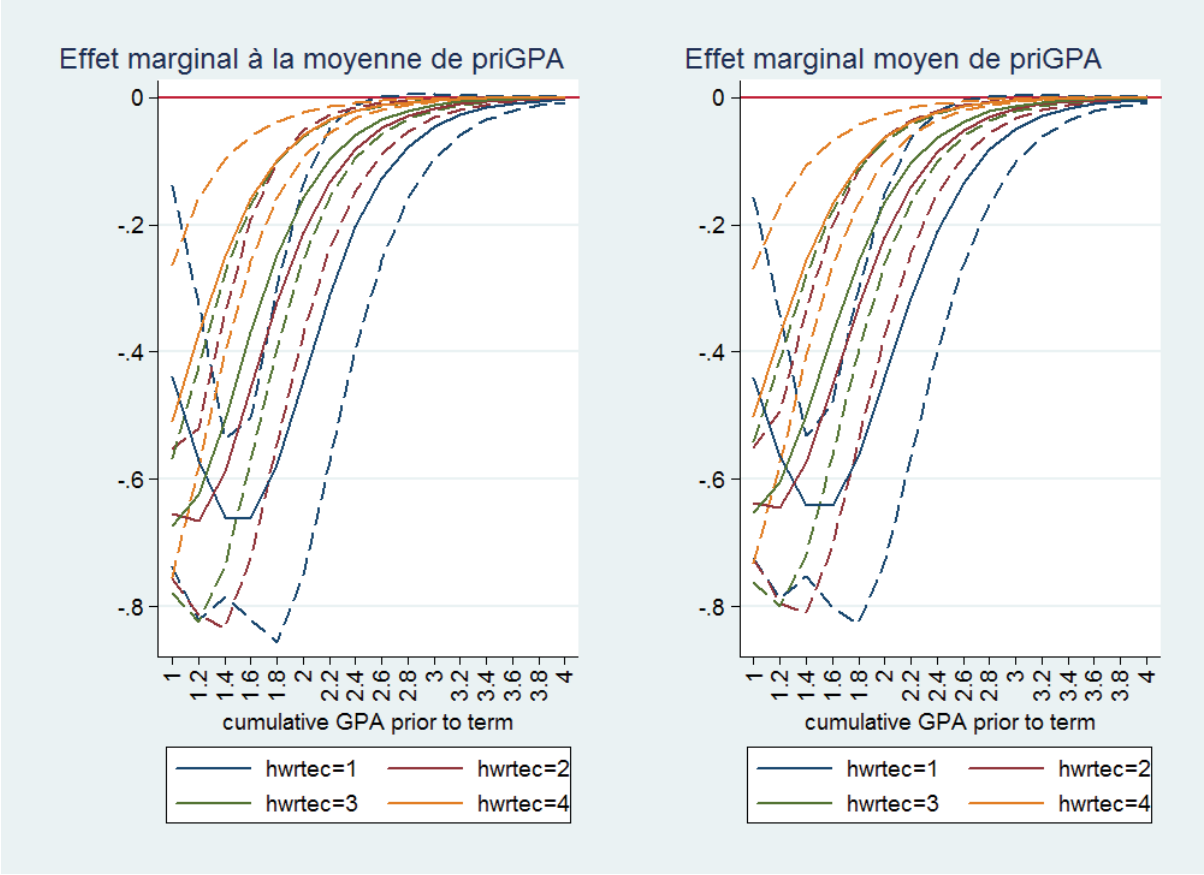
Ce graphique est un peu plus intéressant que le tableau complet des effets marginaux, car il nous donne une bonne idée globale de la taille de l'effet marginal de atndrte quand les deux autres variables varient. Nous voyons donc que l'effet minimum de la variable atndrte est atteint quand le GPA passé est entre 1,4 et 1,6 quand l'étudiant ne rend qu'entre 0 % et 25 % de travaux. Nous voyons que pour une variation de 10 % de la présence aux cours la probabilité d'avoir un GPA de session compris entre 0 et 1 diminue d'un peu plus de 5 %, que ce soit pour l'effet moyen ou pour l'effet à la moyenne.

Nous regardons maintenant l'effet marginal du GPA passé quand lui-même varie selon les niveaux de la variable hwrtec. Nous procédons exactement comme l'exemple précédent de margins et allons directement interpréter le graphique. Les sorties Stata sont donc mises à disposition pour information.

```
margins i.hwrtec, dydx(priGPA) at(priGPA=(1(0.2)4)) atmeans predict(outcome(1))

Conditional marginal effects                      Number of obs   =        674
Model VCE    : OIM
Expression   : Pr(termgpac==1), predict(outcome(1))
dy/dx w.r.t. : priGPA
```

```
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
priGPA       |
  _at#hwrtec |
        1 1  |  -.4380809   .1527185    -2.87   0.004    -.7374036   -.1387582
        1 2  |  -.6551771   .0520357   -12.59   0.000    -.7571653    -.553189
        1 3  |  -.6732779    .054296   -12.40   0.000    -.7796962   -.5668596
        1 4  |  -.5089625   .1252346    -4.06   0.000    -.7544179   -.2635072
        2 1  |  -.5717064   .1274099    -4.49   0.000    -.8214252   -.3219877
        2 2  |  -.6661142   .0749161    -8.89   0.000    -.8129471   -.5192813
        2 3  |  -.6245348    .101686    -6.14   0.000    -.8238358   -.4252338
        2 4  |  -.3710674   .1078465    -3.44   0.001    -.5824426   -.1596921

                                                .
                                                .
                                                .

       15 1  |  -.0055397   .0034338    -1.61   0.107    -.0122698    .0011904
       15 2  |  -.0019881   .0009926    -2.00   0.045    -.0039334   -.0000427
       15 3  |  -.0014094   .0005999    -2.35   0.019    -.0025851   -.0002337
       15 4  |   -.000484   .0001988    -2.43   0.015    -.0008737   -.0000943
       16 1  |  -.0032382   .0020683    -1.57   0.117     -.007292    .0008157
       16 2  |  -.0011608   .0006053    -1.92   0.055    -.0023472    .0000255
       16 3  |  -.0008228    .000373    -2.21   0.027    -.0015539   -.0000918
       16 4  |  -.0002825   .0001245    -2.27   0.023    -.0005266   -.0000384
------------------------------------------------------------------------------

marginsplot, recast(line) nolabel recastci(rline) ciopts(lpattern(dash)) title(Effet marginal à
la moyenne de priGPA) yline(0) xlabel(, angle(90)) ylabel(, angle(0)) ytitle("")

  Variables that uniquely identify margins: priGPA hwrtec

graph save marginsologit3.gph, replace
(file marginsologit3.gph saved)

margins i.hwrtec, dydx(priGPA) at(priGPA=(1(0.2)4)) predict(outcome(1))

Average marginal effects                        Number of obs   =        674
Model VCE    : OIM

Expression   : Pr(termgpac==1), predict(outcome(1))
dy/dx w.r.t. : priGPA

1._at        : priGPA          =           1

2._at        : priGPA          =         1.2

                               .
                               .
                               .

15._at       : priGPA          =         3.8

16._at       : priGPA          =           4

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
priGPA       |
  _at#hwrtec |
        1 1  |  -.4403881   .1448071    -3.04   0.002    -.7242047   -.1565715
        1 2  |  -.6377109   .0448213   -14.23   0.000    -.7255591   -.5498627
        1 3  |    -.65191   .0566711   -11.50   0.000    -.7629833   -.5408367
        1 4  |  -.5004204   .1185401    -4.22   0.000    -.7327546   -.2680861
        2 1  |  -.5641314   .1140572    -4.95   0.000    -.7876795   -.3405833
        2 2  |  -.6444594   .0769968    -8.37   0.000    -.7953705   -.4935484
        2 3  |  -.6058668   .0989862    -6.12   0.000    -.7998762   -.4118574
        2 4  |  -.3724763    .103227    -3.61   0.000    -.5747975   -.1701551
```

```
                                                  .
                                                  .
                                                  .
       15 1  |  -.0060171   .0036594    -1.64   0.100    -.0131894    .0011552
       15 2  |  -.0021615   .0010417    -2.08   0.038    -.0042032   -.0001199
       15 3  |  -.0015327    .000646    -2.37   0.018    -.0027987   -.0002666
       15 4  |  -.0005264   .0002183    -2.41   0.016    -.0009543   -.0000986
       16 1  |  -.0035195   .0022112    -1.59   0.111    -.0078534    .0008144
       16 2  |  -.0012624   .0006383    -1.98   0.048    -.0025135   -.0000113
       16 3  |  -.0008949    .000403    -2.22   0.026    -.0016848    -.000105
       16 4  |  -.0003073   .0001369    -2.25   0.025    -.0005755    -.000039
------------------------------------------------------------------------------

marginsplot, recast(line) nolabel recastci(rline) ciopts(lpattern(dash)) title(Effet marginal
moyen de priGPA) yline(0) xlabel(, angle(90)) ylabel(, angle(0)) ytitle("")

  Variables that uniquely identify margins: priGPA hwrtec

graph save marginsologit4.gph, replace
(file marginsologit4.gph saved)

graph combine marginsologit3.gph marginsologit4.gph, ycommon
```



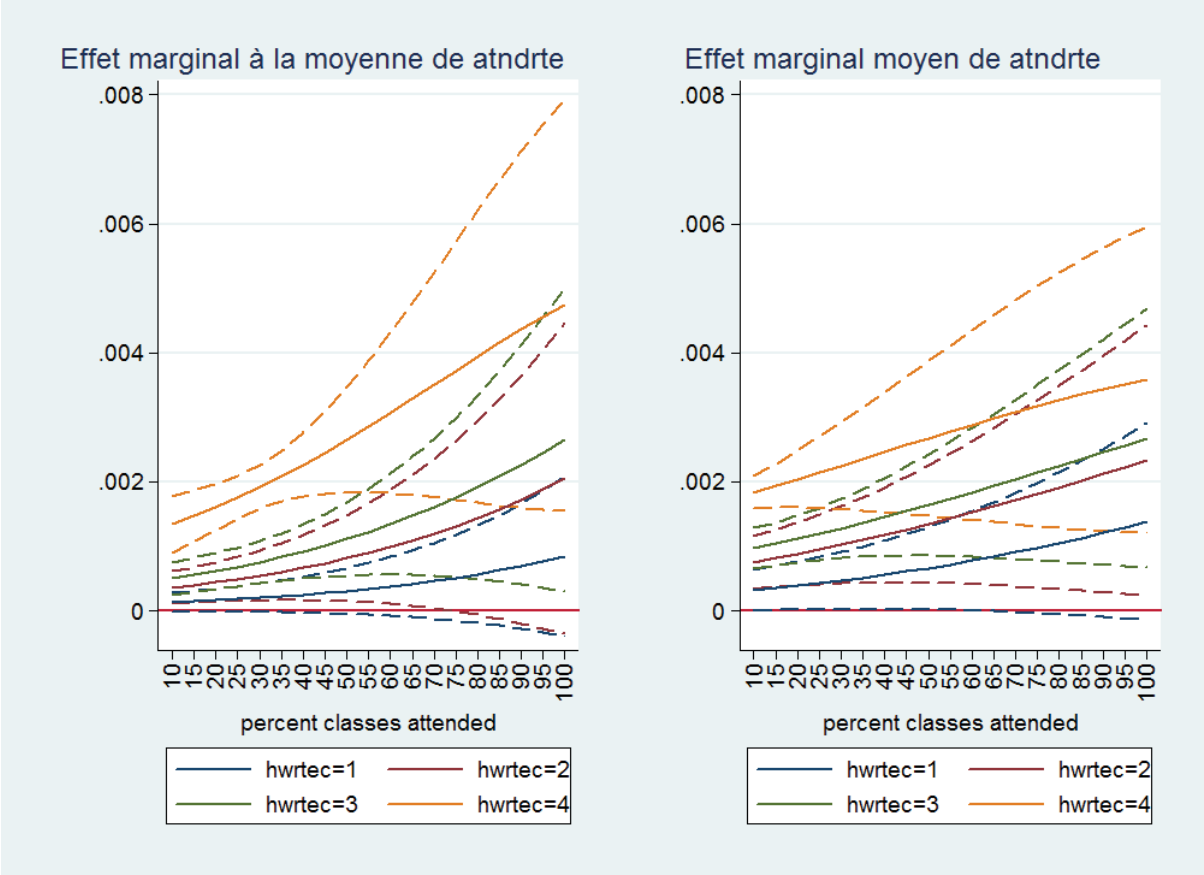
Nous voyons que l'effet marginal moyen de GPA quand l'étudiant rend entre 0 % et 25 % de ses devoirs, atteint son minimum (entre -60 % et -70 %) quand l'étudiant a un GPA entre environ 1,4 et 1,6. Il est intéressant de voir que l'effet converge vers 0 quand la variable priGPA tend vers son maximum. Les effets moyens et à la moyenne sont sensiblement les mêmes.

Nous observons maintnenant l'effet pour l'appartenance à la catégorie 4.

```
margins i.hwrtec, dydx(atndrte) at(atndrte=(10(5)100)) atmeans predict(outcome(4))

Conditional marginal effects                      Number of obs   =        674
Model VCE    : OIM

Expression   : Pr(termgpac==4), predict(outcome(4))
dy/dx w.r.t. : atndrte

1._at        : priGPA          =     2.59182 (mean)
               atndrte         =          10
               1.hwrtec        =    .0237389 (mean)
               2.hwrtec        =    .0578635 (mean)
               3.hwrtec        =     .152819 (mean)
               4.hwrtec        =    .7655786 (mean)
               0.frosh         =    .7700297 (mean)
               1.frosh         =    .2299703 (mean)

2._at        : priGPA          =     2.59182 (mean)
               atndrte         =          15
               1.hwrtec        =    .0237389 (mean)
               2.hwrtec        =    .0578635 (mean)
               3.hwrtec        =     .152819 (mean)
               4.hwrtec        =    .7655786 (mean)
               0.frosh         =    .7700297 (mean)
               1.frosh         =    .2299703 (mean)

                             .
                             .
                             .

18._at       : priGPA          =     2.59182 (mean)
               atndrte         =          95
               1.hwrtec        =    .0237389 (mean)
               2.hwrtec        =    .0578635 (mean)
               3.hwrtec        =     .152819 (mean)
               4.hwrtec        =    .7655786 (mean)
               0.frosh         =    .7700297 (mean)
               1.frosh         =    .2299703 (mean)

19._at       : priGPA          =     2.59182 (mean)
               atndrte         =         100
               1.hwrtec        =    .0237389 (mean)
               2.hwrtec        =    .0578635 (mean)
               3.hwrtec        =     .152819 (mean)
               4.hwrtec        =    .7655786 (mean)
               0.frosh         =    .7700297 (mean)
               1.frosh         =    .2299703 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
atndrte      |
  _at#hwrtec |
        1 1  |   .0001326   .0000738     1.80   0.072    -.0000121    .0002772
        1 2  |   .0003622   .0001251     2.90   0.004      .000117    .0006074
        1 3  |    .000504   .0001287     3.92   0.000     .0002518    .0007561
        1 4  |   .0013417   .0002228     6.02   0.000      .000905    .0017784
        2 1  |   .0001473   .0000817     1.80   0.071    -.0000128    .0003075
        2 2  |   .0004015   .0001387     2.89   0.004     .0001296    .0006734
        2 3  |   .0005578   .0001342     4.16   0.000     .0002948    .0008207
        2 4  |   .0014707   .0002021     7.28   0.000     .0010746    .0018668

                                                  .
                                                  .
                                                  .

       18 1  |   .0007661   .0005598     1.37   0.171    -.0003311    .0018634
```

```
       18 2  |    .001881   .0011001     1.71   0.087    -.0002752    .0040371
       18 3  |   .0024537   .0010714     2.29   0.022     .0003538    .0045537
       18 4  |   .0045537   .0015276     2.98   0.003     .0015597    .0075477
       19 1  |   .0008455   .0006326     1.34   0.181    -.0003944    .0020853
       19 2  |   .0020479   .0012295     1.67   0.096    -.0003618    .0044577
       19 3  |   .0026509    .001202     2.21   0.027     .0002951    .0050068
       19 4  |   .0047333    .001627     2.91   0.004     .0015444    .0079223
------------------------------------------------------------------------------

marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) title(Effet marginal à la
moyenne de atndrte) yline(0) xlabel(, angle(90)) ylabel(, angle(0)) ytitle("")

  Variables that uniquely identify margins: atndrte hwrtec

graph save marginsologit21.gph, replace
(file marginsologit21.gph saved)

margins i.hwrtec, dydx(atndrte) at(atndrte=(10(5)100)) predict(outcome(4))

Average marginal effects                        Number of obs   =        674
Model VCE    : OIM

Expression   : Pr(termgpac==4), predict(outcome(4))
dy/dx w.r.t. : atndrte

1._at        : atndrte         =          10

2._at        : atndrte         =          15

                              .
                              .
                              .

18._at       : atndrte         =          95

19._at       : atndrte         =         100

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
atndrte      |
  _at#hwrtec |
        1 1  |   .0003257   .0001628     2.00   0.046     6.50e-06    .0006448
        1 2  |    .000753   .0002077     3.62   0.000     .0003458    .0011602
        1 3  |   .0009645   .0001619     5.96   0.000     .0006471    .0012819
        1 4  |   .0018352   .0001295    14.17   0.000     .0015813     .002089
        2 1  |   .0003575   .0001763     2.03   0.043     .0000119     .000703
        2 2  |   .0008148   .0002258     3.61   0.000     .0003721    .0012574
        2 3  |   .0010375   .0001698     6.11   0.000     .0007048    .0013702
        2 4  |   .0019364   .0001727    11.21   0.000     .0015979    .0022748

                                                  .
                                                  .
                                                  .

       18 1  |   .0012911   .0007204     1.79   0.073    -.0001209     .002703
       18 2  |    .002225   .0010014     2.22   0.026     .0002622    .0041878
       18 3  |   .0025667   .0009578     2.68   0.007     .0006894     .004444
       18 4  |   .0035091   .0011685     3.00   0.003     .0012188    .0057993
       19 1  |   .0013777   .0007788     1.77   0.077    -.0001488    .0029042
       19 2  |   .0023308   .0010676     2.18   0.029     .0002383    .0044234
       19 3  |   .0026722    .001023     2.61   0.009     .0006672    .0046772
       19 4  |   .0035786    .001209     2.96   0.003      .001209    .0059482
------------------------------------------------------------------------------

marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) title(Effet marginal moyen de
atndrte) yline(0) xlabel(, angle(90)) ylabel(, angle(0)) ytitle("")

  Variables that uniquely identify margins: atndrte hwrtec
```

```
graph save marginsologit22.gph, replace
(file marginsologit22.gph saved)

graph combine marginsologit21.gph marginsologit22.gph, ycommon
```



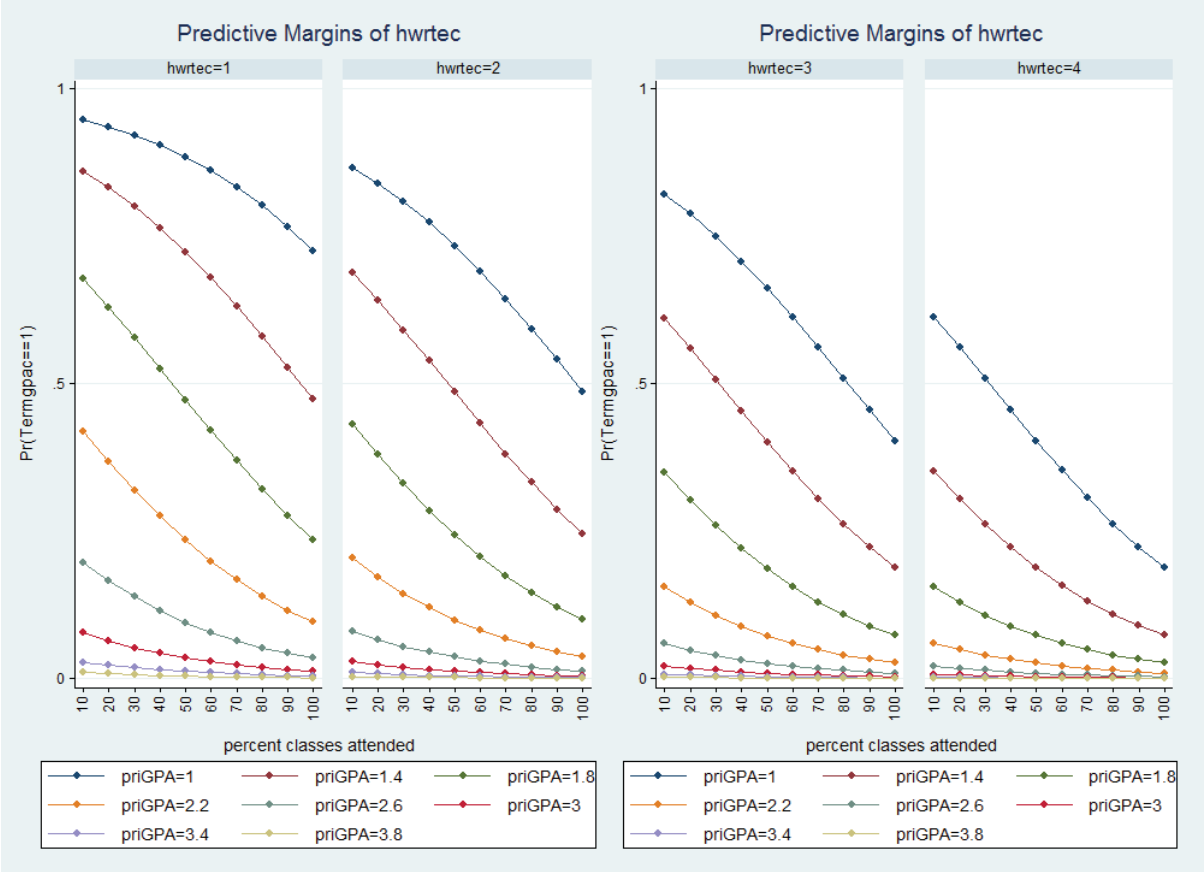
Nous voyons que l'effet marginal de la variable atndrte quand celle-ci varie est croissant pour tous les niveaux de hwrtec. Plus un étudiant rend de devoirs plus l'effet marginal de atndrte sera élevé.

Nous allons maintenant représenter graphiquement les prédictions du modèle par rapport à des valeurs différentes de atndrte et priGPA en différenciant par la variable hwrtec. Nous allons utiliser une astuce pour avoir des graphiques plus lisibles en séparant les catégories de la variable hwrtec en deux groupes. Nous allons donc représenter graphiquement les deux premières catégories de la variable hwrtec dans un graphique et les deux dernières dans un autre graphique. Nous allons combiner ces deux graphiques pour avoir les 4 catégories réunies dans un seul graphique.

```
margins 1.hwrtec 2.hwrtec, at(atndrte=(10(10)100)priGPA=(1(0.4)4)) predict(outcome(1))

Predictive margins                                Number of obs   =        674
```

```
Model VCE    : OIM

Expression   : Pr(termgpac==1), predict(outcome(1))

1._at        : priGPA          =           1
               atndrte         =          10

2._at        : priGPA          =           1
               atndrte         =          20

                              .
                              .
                              .

9._at        : priGPA          =           1
               atndrte         =          90

10._at       : priGPA          =           1
               atndrte         =         100

11._at       : priGPA          =         1.4
               atndrte         =          10

12._at       : priGPA          =         1.4
               atndrte         =          20

                              .
                              .
                              .

19._at       : priGPA          =         1.4
               atndrte         =          90

20._at       : priGPA          =         1.4
               atndrte         =         100

                              .
                              .
                              .

79._at       : priGPA          =         3.8
               atndrte         =          90

80._at       : priGPA          =         3.8
               atndrte         =         100

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _at#hwrtec |
        1 1  |   .9474897   .0304758    31.09   0.000     .8877582    1.007221
        1 2  |   .8661088   .0535454    16.18   0.000     .7611617    .9710559
        2 1  |    .935788   .0355857    26.30   0.000     .8660414    1.005535
        2 2  |    .839385   .0591556    14.19   0.000      .723442    .9553279

                                                  .
                                                  .
                                                  .

       79 1  |   .0017563   .0011722     1.50   0.134    -.0005412    .0040538
       79 2  |   .0006293   .0003511     1.79   0.073    -.0000588    .0013175
       80 1  |   .0014187    .000977     1.45   0.146    -.0004962    .0033335
       80 2  |   .0005082   .0002979     1.71   0.088    -.0000757    .0010921
------------------------------------------------------------------------------

marginsplot, noci by(hwrtec) legend(cols(3)) xlabel(, angle(90)) ylabel(0(0.5)1, angle(0))

  Variables that uniquely identify margins: atndrte priGPA hwrtec

graph save marginsglobologit1, replace
```

```
(file marginsglobologit1.gph saved)

margins 3.hwrtec 4.hwrtec, at(atndrte=(10(10)100)priGPA=(1(0.4)4)) predict(outcome(1))

Predictive margins                                 Number of obs   =        674
Model VCE    : OIM

Expression   : Pr(termgpac==1), predict(outcome(1))

1._at        : priGPA          =           1
               atndrte         =          10

2._at        : priGPA          =           1
               atndrte         =          20

                              .
                              .
                              .

79._at       : priGPA          =         3.8
               atndrte         =          90

80._at       : priGPA          =         3.8
               atndrte         =         100

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _at#hwrtec |
        1 3  |    .821044   .0723926    11.34   0.000     .6791571    .9629309
        1 4  |   .6124721   .1250142     4.90   0.000     .3674487    .8574955
        2 3  |   .7875521   .0758338    10.39   0.000     .6389206    .9361837
        2 4  |   .5608934   .1185098     4.73   0.000     .3286184    .7931684

                                                  .
                                                  .
                                                  .

       79 3  |   .0004461   .0002155     2.07   0.039     .0000236    .0008685
       79 4  |   .0001531   .0000712     2.15   0.031     .0000136    .0002926
       80 3  |   .0003602   .0001797     2.00   0.045     7.92e-06    .0007125
       80 4  |   .0001236   .0000583     2.12   0.034     9.30e-06     .000238
------------------------------------------------------------------------------

marginsplot, noci by(hwrtec) legend(cols(3)) xlabel(, angle(90)) ylabel(0(0.5)1, angle(0))

  Variables that uniquely identify margins: atndrte priGPA hwrtec

graph save marginsglobologit2, replace
(file marginsglobologit2.gph saved)

graph combine marginsglobologit1.gph marginsglobologit2.gph, ycommon
```



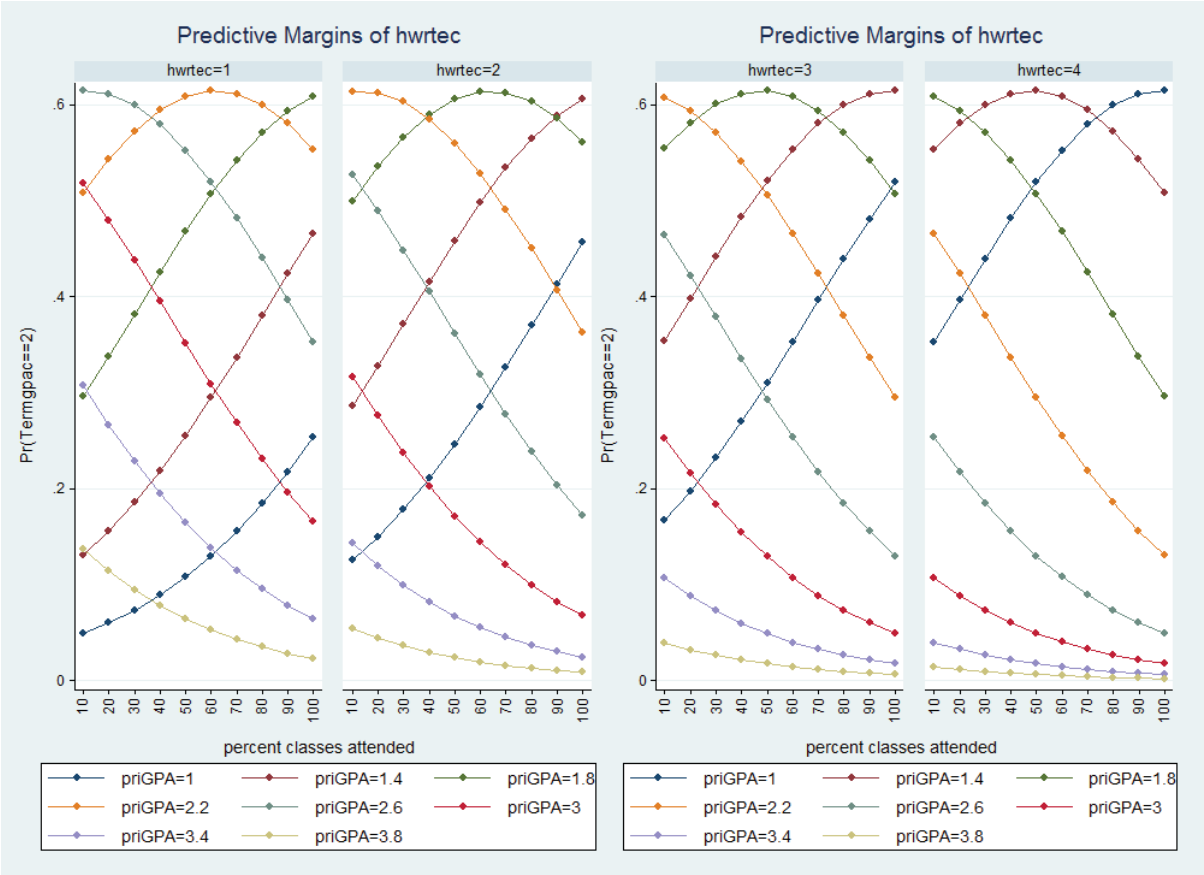
La prédiction linéaire nous donne un aperçu global des effets de chaque variable. Nous voyons clairement que plus les étudiants rendent leurs devoirs moins ils ont de probabilité d'avoir un GPA de session entre 0 et 1. Une augmentation de la présence aux cours réduit aussi la probabilité de finir la session avec un mauvais GPA. Nous remarquons que le GPA passé a une grande influence et c'est cette mesure qui détermine le plus l'appartenance aux catégories.

Nous regardons maintenant l'effet des variables sur la probabilité d'avoir un GPA de session entre 1 et 2. Nous réalisons donc les mêmes calculs avec la commande margins mais en changeant simplement l'outcome prédit.

```
margins 1.hwrtec 2.hwrtec, at(atndrte=(10(10)100)priGPA=(1(0.4)4)) predict(outcome(2))

Predictive margins                                  Number of obs   =        674
Model VCE    : OIM

Expression   : Pr(termgpac==2), predict(outcome(2))

1._at        : priGPA          =           1
               atndrte         =          10

2._at        : priGPA          =           1
               atndrte         =          20
```

```
                              .
                              .
                              .

79._at       : priGPA          =         3.8
               atndrte         =          90

80._at       : priGPA          =         3.8
               atndrte         =         100

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _at#hwrtec |
        1 1  |   .0493854    .028626     1.73   0.084    -.0067206    .1054914
        1 2  |   .1252103   .0502084     2.49   0.013     .0268036    .2236169
        2 1  |    .060345   .0333966     1.81   0.071    -.0051112    .1258011
        2 2  |    .149887   .0553925     2.71   0.007     .0413197    .2584544

                                              .
                                              .
                                              .

       79 1  |   .0284277   .0177693     1.60   0.110    -.0063995    .0632549
       79 2  |   .0103909   .0049664     2.09   0.036     .0006569    .0201249
       80 1  |   .0230987   .0149215     1.55   0.122    -.0061469    .0523442
       80 2  |   .0084096   .0042516     1.98   0.048     .0000766    .0167425
------------------------------------------------------------------------------

marginsplot, noci by(hwrtec) legend(cols(3)) xlabel(, angle(90)) ylabel(0(0.2)0.6, angle(0))

  Variables that uniquely identify margins: atndrte priGPA hwrtec

graph save marginsglobologit3, replace
(file marginsglobologit3.gph saved)

margins 3.hwrtec 4.hwrtec, at(atndrte=(10(10)100)priGPA=(1(0.4)4)) predict(outcome(2))

Predictive margins                                 Number of obs   =        674
Model VCE    : OIM

Expression   : Pr(termgpac==2), predict(outcome(2))

1._at        : priGPA          =           1
               atndrte         =          10

2._at        : priGPA          =           1
               atndrte         =          20

                              .
                              .
                              .

79._at       : priGPA          =         3.8
               atndrte         =          90

80._at       : priGPA          =         3.8
               atndrte         =         100

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _at#hwrtec |
        1 3  |   .1667503   .0674977     2.47   0.013     .0344572    .2990434
        1 4  |   .3527857   .1104589     3.19   0.001     .1362903    .5692812
        2 3  |   .1973767      .0706     2.80   0.005     .0590033    .3357501
        2 4  |    .396441   .1033213     3.84   0.000     .1939349    .5989471

                                              .
```

```
                                                  .
                                                  .

        79 3  |   .0073888   .0027218     2.71   0.007     .0020542    .0127234
        79 4  |   .0025496   .0008508     3.00   0.003      .000882    .0042171
        80 3  |   .0059759    .002282     2.62   0.009     .0015034    .0104485
        80 4  |   .0020599   .0006904     2.98   0.003     .0007066    .0034131
------------------------------------------------------------------------------

marginsplot, noci by(hwrtec) legend(cols(3)) xlabel(, angle(90)) ylabel(0(0.2)0.6, angle(0))
  Variables that uniquely identify margins: atndrte priGPA hwrtec

graph save marginsglobologit4, replace
(file marginsglobologit4.gph saved)

graph combine marginsglobologit3.gph marginsglobologit4.gph, ycommon
```



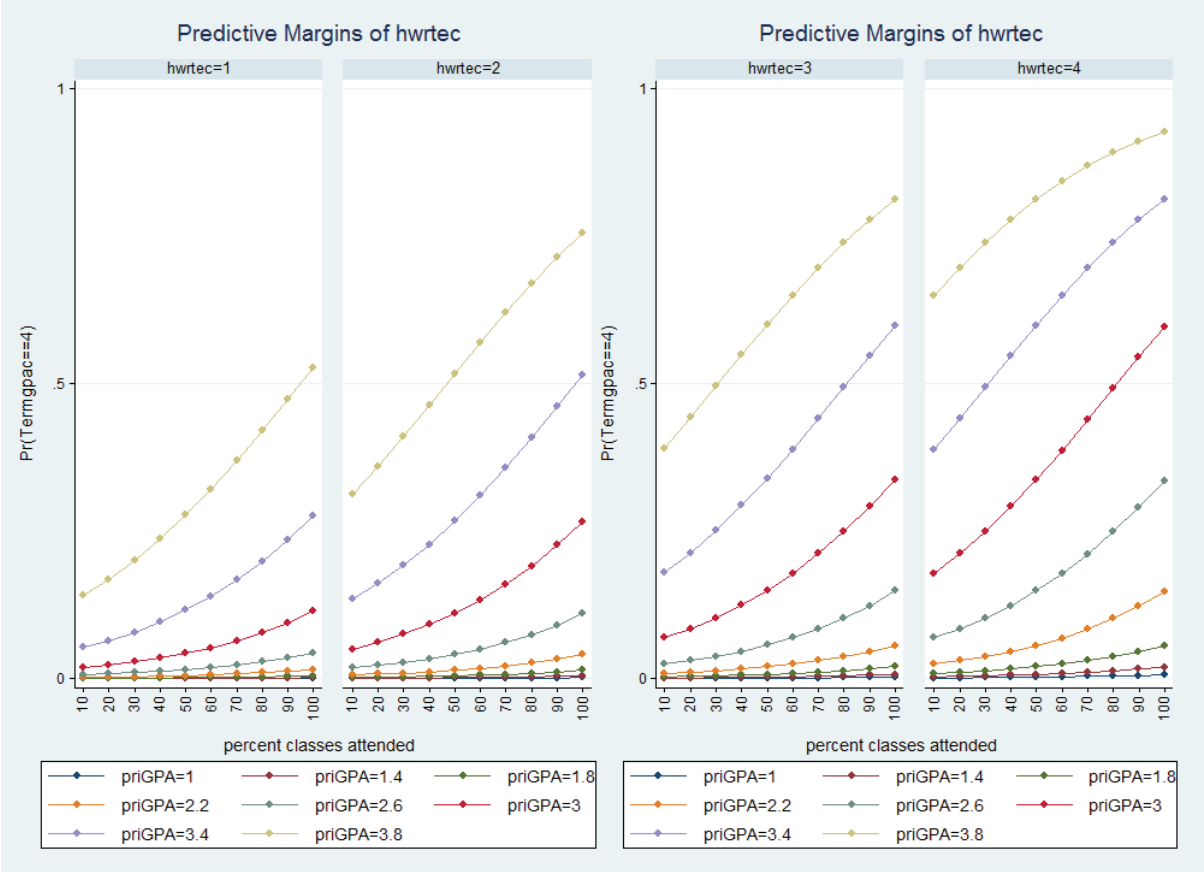
Nous voyons que cette fois-ci, l'effet de rendre plus ou moins de devoirs pour les étudiants change aussi la forme de l'effet de la variable atndrte. Par exemple, si un étudiant ayant un GPA passé de 2,2 qui rendait entre 0 et 25 % de ses devoirs décide de tout rendre, l'effet de atndrte passera de concave à convexe. Quand un étudiant à un GPA passé aux alentours de 1, le fait de rendre plus de travaux lui permettra d'augmenter strictement sa probabilité d'avoir un GPA de session entre 1 et 2. Si nous prenons un étudiant ayant un GPA passé de 1,4 il aura strictement plus de chance d'avoir un GPA de session entre 1 et 2 jusqu'à 75 % de travaux

rendu. S'il rend entre 75 % et 100 % de ses travaux, nous pouvons voir que cette probabilité commence à réduire au profit d'une plus grande chance d'avoir un GPA de session plus élevé que 2.

Nous pouvons répéter l'opération pour les autres modalités de la variable termgpac. Nous laissons au lecteur le soin de consulter le fichier discretordon.do pour avoir accès aux sorties Stata et graphiques concernant la modalité 3. Nous passons donc directement aux prédictions concernant la modalité 4.

```
margins 1.hwrtec 2.hwrtec, at(atndrte=(10(10)100)priGPA=(1(0.4)4)) predict(outcome(4))

Predictive margins                                      Number of obs   =        674
Model VCE    : OIM

Expression   : Pr(termgpac==4), predict(outcome(4))

1._at        : priGPA          =           1
               atndrte         =          10

2._at        : priGPA          =           1
               atndrte         =          20

                                .
                                .
                                .

79._at       : priGPA          =         3.8
               atndrte         =          90

80._at       : priGPA          =         3.8
               atndrte         =         100

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _at#hwrtec |
        1 1  |   .0000871   .0000626     1.39   0.165    -.0000357    .0002098
        1 2  |   .0002432   .0001359     1.79   0.073    -.0000231    .0005095
        2 1  |   .0001078   .0000751     1.44   0.151    -.0000394     .000255
        2 2  |   .0003011   .0001604     1.88   0.060    -.0000132    .0006155

                                                .
                                                .
                                                .

       79 1  |    .473523   .1524886     3.11   0.002     .1746508    .7723951
       79 2  |   .7145119   .0905183     7.89   0.000     .5370993    .8919246
       80 1  |   .5267042   .1565929     3.36   0.001     .2197878    .8336206
       80 2  |   .7559464   .0867841     8.71   0.000     .5858527    .9260402
------------------------------------------------------------------------------

marginsplot, noci by(hwrtec) legend(cols(3)) xlabel(, angle(90)) ylabel(0(0.5)1, angle(0))

  Variables that uniquely identify margins: atndrte priGPA hwrtec

graph save marginsglobologit7, replace
(file marginsglobologit7.gph saved)


margins 3.hwrtec 4.hwrtec, at(atndrte=(10(10)100)priGPA=(1(0.4)4)) predict(outcome(4))
```

```
Predictive margins                                 Number of obs   =        674
Model VCE    : OIM

Expression   : Pr(termgpac==4), predict(outcome(4))

1._at        : priGPA          =           1
               atndrte         =          10

2._at        : priGPA          =           1
               atndrte         =          20

                               .
                               .
                               .

79._at       : priGPA          =         3.8
               atndrte         =          90

80._at       : priGPA          =         3.8
               atndrte         =         100

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _at#hwrtec |
        1 3  |   .0003431   .0001951     1.76   0.079    -.0000393    .0007256
        1 4  |   .0009993   .0005763     1.73   0.083    -.0001303    .0021289
        2 3  |   .0004249   .0002241     1.90   0.058    -.0000143    .0008641
        2 4  |   .0012372   .0006528     1.90   0.058    -.0000422    .0025167

                                                  .
                                                  .
                                                  .

       79 3  |   .7791625   .0546972    14.25   0.000      .671958    .8863671
       79 4  |   .9111983    .019444    46.86   0.000     .8730888    .9493078
       80 3  |   .8136793   .0499825    16.28   0.000     .7157155    .9116432
       80 4  |   .9270315   .0160787    57.66   0.000     .8955178    .9585452
------------------------------------------------------------------------------

marginsplot, noci by(hwrtec) legend(cols(3)) xlabel(, angle(90)) ylabel(0(0.5)1, angle(0))

  Variables that uniquely identify margins: atndrte priGPA hwrtec

graph save marginsglobologit8, replace
(file marginsglobologit8.gph saved)

graph combine marginsglobologit7.gph marginsglobologit8.gph, ycommon
```



Nous pouvons déduire que, plus un étudiant a un GPA passé élevé, plus il travaille et plus il assistera aux cours et moins il aura de chance de se retrouver avec un GPA de session faible. Cela implique aussi qu'il aura plus de chance d'avoir un GPA de session plus élevé mais que son GPA passé qui détermine la plus grosse part des probabilités d'appartenance à une modalité de la variable dépendante détermine aussi sa marge d'évolution à travers les catégories.

Nous allons maintenant prédire les probabilités d'appartenance à chaque classe pour chaque individu et procéder au classement des individus.

```
predict p1 if e(sample), outcome(1)


(option pr assumed; predicted probability)
(6 missing values generated)

predict p2 if e(sample), outcome(2)
(option pr assumed; predicted probability)
(6 missing values generated)

predict p3 if e(sample), outcome(3)
(option pr assumed; predicted probability)
(6 missing values generated)

predict p4 if e(sample), outcome(4)
(option pr assumed; predicted probability)
(6 missing values generated)
```

Nous créons ensuite des variables dichotomiques définies comme suit

$$p1best = \begin{cases} 1 & si\ p1 > p2\ et\ p1 > p3\ et\ p1 > p4 \\ 0 & sinon \end{cases}$$

$$p2best = \begin{cases} 1 & si\ p2 > p1\ et\ p2 > p3\ et\ p2 > p4 \\ 0 & sinon \end{cases}$$

$$p3best = \begin{cases} 1 & si\ p3 > p2\ et\ p3 > p1\ et\ p3 > p4 \\ 0 & sinon \end{cases}$$

$$p4best = \begin{cases} 1 & si\ p4 > p2\ et\ p4 > p1\ et\ p4 > p3 \\ 0 & sinon \end{cases}$$

```
gen p1best = ( p1>p2 & p1>p3 & p1>p4)

replace  p1best=. if p1==.
(6 real changes made, 6 to missing)
gen p2best= ( p2>p1 & p2>p3 & p2>p4)

replace  p2best=. if p2==.
(6 real changes made, 6 to missing)
gen p3best= ( p3>p1 & p3>p2 & p3>p4)

replace  p3best=. if p3==.
(6 real changes made, 6 to missing)
gen p4best= ( p4>p1 & p4>p2 & p4>p3)

replace  p4best=. if p4==.
(6 real changes made, 6 to missing)
```

Nous générons ensuite la variable status prédite par la modèle ou

$$termgpahat = \begin{cases} 1\ si\ p1best = 1 \\ 2\ si\ p2best = 1 \\ 3\ si\ p3best = 1 \\ 4\ si\ p4best = 1 \end{cases}$$

```
gen termgpachat  = 1 if(p1best== 1 & p2best== 0 & p3best== 0 & p4best==0)
(674 missing values generated)

replace termgpachat = 2 if(p1best== 0 & p2best== 1 & p3best== 0 & p4best==0)
(68 real changes made)

replace termgpachat = 3 if(p1best== 0 & p2best== 0 & p3best== 1 & p4best==0)
(442 real changes made)

replace termgpachat = 4 if(p1best== 0 & p2best== 0 & p3best==0 & p4best== 1)
(158 real changes made)
```

Il convient enfin de créer un tableau croisé de la variable observée et de la variable prédite en précisant l'option row pour avoir la fréquence de la cellule par rapport au total de la ligne en pourcentage et nofreq pour ne pas afficher la fréquence de la cellule. Nous affichons ainsi la proportion d'individus classés par le modèle par rapport à la vraie catégorie.

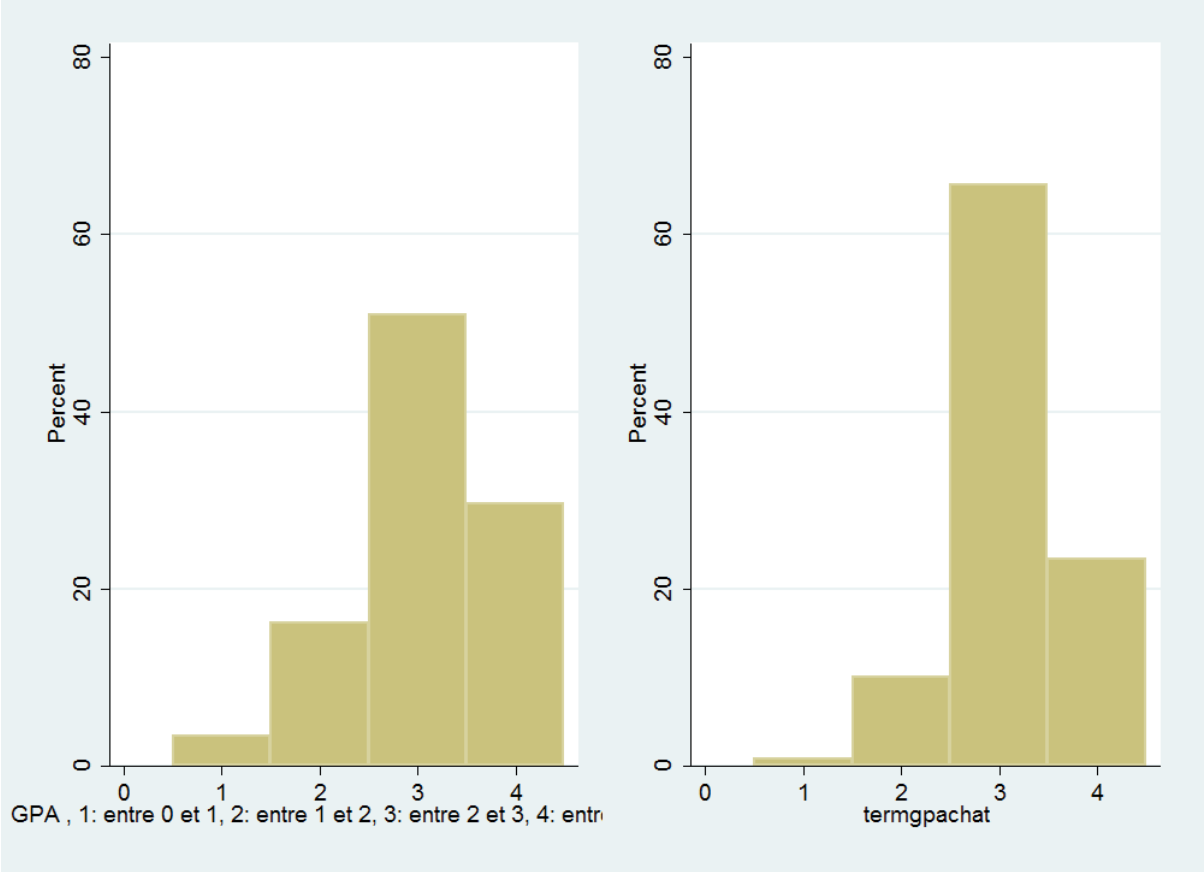
```
tab termgpac termgpachat, row nofreq
```

```
  GPA , 1: |
entre 0 et |
     1, 2: |
entre 1 et |
     2, 3: |
entre 2 et |
     3, 4: |
entre 3 et |                  termgpachat
         4 |         1          2          3          4 |     Total
-----------+--------------------------------------------+----------
         1 |     23.81      57.14      19.05       0.00 |    100.00
         2 |      0.94      32.08      66.98       0.00 |    100.00
         3 |      0.00       6.07      82.95      10.98 |    100.00
         4 |      0.00       0.50      39.80      59.70 |    100.00
-----------+--------------------------------------------+----------
     Total |      0.89      10.09      65.58      23.44 |    100.00
```

Nous voyons que la catégorie la mieux classée est la catégorie 3 avec 82,95 % de bonnes prédictions. Suit ensuite la catégorie 4 avec 59,7 % de bonnes prédictions, la catégorie 2 avec 32.08 % de bonnes prédictions et la catégorie 1 avec 23.81 % de bonnes prédictions.

Nous traçons ensuite l'histogramme des deux variables côtes à côtes afin de pouvoir les comparer graphiquement. Nous utilisons les commandes graph save et graph combine afin de créer un graphique plus facile à interpréter.

```
histogram termgpac, discrete percent
(start=1, width=1)

graph save termgpac.gph, replace
(file termgpac.gph saved)

histogram termgpachat, discrete percent
(start=1, width=1)

graph save termgpachat.gph, replace
(file termgpachat.gph saved)

graph combine termgpac.gph termgpachat.gph, ycommon
```



Notre modèle a tendance à surclasser les individus dans la troisième catégorie et sous-classer les individus dans les autres catégories. Cependant, il a une efficacité globale satisfaisante si nous nous référons au tableau donné par la commande `fitstat` un peu plus haut avec une bonne prédiction globale (toutes catégories confondues de 66,2 %.

Tests sur les variables catégorielles

Étant donné que nous avons une variable catégorielle concernant le nombre de devoirs rendus, nous allons tester si les modalités de la variable hwrtec sont significativement différentes. Premièrement, nous testons l'égalité jointe de toutes les modalités avec la commande `contrast`.

```
contrast hwrtec

Contrasts of marginal linear prédictions

Margins      : asbalanced

-------------------------------------------------
             |          df        chi2     P>chi2
-------------+-----------------------------------
termgpac     |
      hwrtec |           3       32.91     0.0000
-------------------------------------------------
```

Nous pouvons rejeter l'hypothèse nulle d'égalité des coefficients. Nous regardons donc s'il existe une différence entre la modalité de référence et chaque coefficient en plaçant le préfixe `r.` devant la variable à tester.

```
contrast r.hwrtec

Contrasts of marginal linear prédictions

Margins      : asbalanced

-------------------------------------------------
             |         df        chi2     P>chi2
-------------+-----------------------------------
termgpac     |
      hwrtec |
   (2 vs 1)  |          1        2.77     0.0961
   (3 vs 1)  |          1        5.52     0.0188
   (4 vs 1)  |          1       17.50     0.0000
      Joint  |          3       32.91     0.0000
-------------------------------------------------

------------------------------------------------------------------
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
termgpac     |
      hwrtec |
   (2 vs 1)  |   1.027423   .6174534     -.1827632     2.23761
   (3 vs 1)  |   1.371838   .5838661      .2274818    2.516195
   (4 vs 1)  |   2.441417   .5835358      1.297708    3.585126
------------------------------------------------------------------
```

Il y a une différence significative si nous tolérons une erreur de type I de 10 %. Pour chaque paire comparée. Nous testons donc s'il existe une différence entre le coefficient d'une modalité et celui de la modalité précédente en spécifiant le préfixe ar. devant la variable testée.

```
contrast ar.hwrtec

Contrasts of marginal linear prédictions

Margins      : asbalanced

-------------------------------------------------
             |         df        chi2     P>chi2
-------------+-----------------------------------
termgpac     |
      hwrtec |
   (2 vs 1)  |          1        2.77     0.0961
   (3 vs 2)  |          1        0.73     0.3932
   (4 vs 3)  |          1       18.98     0.0000
      Joint  |          3       32.91     0.0000
-------------------------------------------------

------------------------------------------------------------------
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
termgpac     |
      hwrtec |
   (2 vs 1)  |   1.027423   .6174534     -.1827632     2.23761
   (3 vs 2)  |   .3444151   .4033975     -.4462295     1.13506
   (4 vs 3)  |   1.069579   .2455172      .5883738    1.550784
------------------------------------------------------------------
```

Cette fois-ci, nous pouvons observer qu'il n'y a pas de différence statistiquement significative entre le coefficient de la modalité 2 et de la modalité 3.

Nous comparerons pour finir toutes les paires de catégories que nous pouvons former grâce à la commande pwcompare. Nous pouvons aussi tenir compte du nombre de comparaisons simultanément effectuées pour nos tests en utilisant la méthode de Bonferroni avec l'option `mcompare(bonferroni)`.

```
pwcompare i.hwrtec, mcompare(bonferroni) cimargins effects sort

Pairwise comparisons of marginal linear prédictions

Margins      : asbalanced

------------------------------------------------------------
             |                                Unadjusted
             |     Margin   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
termgpac     |
      hwrtec |
          1  |   .1561143   .1016561      -.043128    .3553566
          2  |   1.183538   .6297983     -.0508444     2.41792
          3  |   1.527953   .5943138       .363119    2.692786
          4  |   2.597531   .5973592      1.426729    3.768334
------------------------------------------------------------

---------------------------
             |    Number of
             |  Comparisons
-------------+-------------
termgpac     |
      hwrtec |            6
---------------------------

------------------------------------------------------------------------------
             |                            Bonferroni           Bonferroni
             |   Contrast   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
termgpac     |
      hwrtec |
     3 vs 2  |   .3444151   .4033975     0.85   1.000    -.7198513    1.408681
     2 vs 1  |   1.027423   .6174534     1.66   0.577    -.6015777    2.656424
     4 vs 3  |   1.069579   .2455172     4.36   0.000     .4218411    1.717316
     3 vs 1  |   1.371838   .5838661     2.35   0.113    -.1685506    2.912227
     4 vs 2  |   1.413994   .4020975     3.52   0.003     .3531573     2.47483
     4 vs 1  |   2.441417   .5835358     4.18   0.000     .9018995    3.980935
------------------------------------------------------------------------------
```

En corrigeant pour le nombre de catégories, nous voyons que les modalités 1, 2 et 3 ne sont pas statistiquement significatives. Nous pourrions donc essayer de créer une variable binaire pour les nombres de travaux rendus qui serait égale à 1 si le nombre de travaux rendus est supérieur à 75 % et 0 sinon.

3.4.5 La commande oprobit

Description

La commande oprobit dans Stata sert à estimer un modèle probit multinomial ordonné ou régression probit polytomique ordonnée et utilise la méthode du maximum de vraisemblance pour estimer les paramètres ($\hat{\beta}$) du modèle. Ce type de modèle doit être choisi quand la variable dépendante est discrète et composée de modalités qui peuvent être naturellement ordonnées. Stata traitera les valeurs de la variable dépendante telles qu'elles seront spécifiées par l'utilisateur. Ainsi, il est important de vérifier le codage de la variable dépendante en utilisant la commande summarize. La première catégorie est généralement codée 0. Les variables négatives seront considérées comme manquantes. Il faut donc faire attention à avoir une variable codée strictement positivement 0, 1, 2, 3, etc. Quand la variable dépendante ne prend que deux modalités, les paramètres estimations seront les même qu'en utilisant la commande probit.

Syntaxe

`oprobit [depvar] [indepvars] [if] [in] [weight] [, options]`

où `depvar` est la variable dépendante
`[indepvars]` est la variable indépendante
`[if]` et `[in]` sont des contraintes dans l'échantillon
`[weight]` assigne des poids aux observations
`[, options]` présente des options supplémentaires dans la régression

Menu

Statistics > Ordinal outcomes > Ordered probit régression

Option

Erreur type/Robust

`vce(typevce)` : le type vce peut être `robust`, `cluster`(*variablecluster*). Pour plus d'informations, le lecteur peut se référer à la documentation Stata.

3.4.6 Les commandes postestimation après la commande oprobit

`test [[label ou valeur de la modalité]] [varexo]`: test de Wald avec lequel nous pouvons tester des restrictions linéaires (par exemple $\beta_1 = \beta_2$ ou $\beta_1 = 0$).

`listcoef [, const help]` : donne une liste des coefficients pour le dernier modèle estimé. L'option const inclut la constante dans le tableau et l'option help rajoute une légende des variables diponibles dans le tableau. Pour plus de renseignement concernant les options, l'utilisateur peut se référer à l'aide fournie avec le module en utilisant la commande help listcoef.

`contrast [varexo]` : permet de tester si les moyennes des effets sont égales entre toutes les modalités d'une variable ordinale incluse dans la régression (x).

`contrast r.[varexo]` : permet de tester si les moyennes des effets sont égales entre les différentes modalités d'une variable ordinale quand elles sont comparées à la catégorie de référence.

`contrast ar.[varexo]` : permet de tester si le coefficient d'une modalité est différent du coefficient de la modalité qui le précède.

`fitstat` : indices de qualité du modèle (fit ou adéquation).

`lrtest [modèlecomplet] [modelcontraint][, stats]` : cette commande performe un test de ratio de vraisemblance (likelihood ratio test) entre un modèle complet et un modèle contraint. Ce test nous renseigne sur la significativité des coefficients non inclus dans le modèle contraint et est similaire au test-F joint qui est calculé avec la commande test. Le modèle contraint est donc un sous modèle du modèle complet. Ce test peut être utilisé pour un test de Chow si nous voulons savoir si nous voulons savoir si les coefficients du modèle sont différents entre plusieurs tranches de l'échantillon.

```
margins,                                                                    [dydx(varexos)]
[at(varexo1=(valdébut(incrément)valfin)varexo2=(valdébut(incrément)valfin)] [predict(outcome(labe
l ou valeur de la modalité))]:
```

cette commande nous permet de calculer les effets marginaux des variables spécifiées et les prédictions du modèle. L'option dydx(.) est nécessaire pour calculer les effets marginaux de variables continues, l'option at nous servira pour pouvoir utiliser la commande marginsplot afin d'avoir un aperçu graphique de ces effets. Dans le cas d'un modèle logistique multinomial, il est aussi nécessaire de spécifier pour quelle catégorie nous voulons les effets marginaux.

`marginsplot [varexos]` : cette commande doit être utilisée après margins pour avoir un aperçu graphique des effets marginaux sur les probabilités.

`prchange [varlist], [outcome(label ou valeur de la modalité)] [help]`: calcul l'impact d'une variation discrète ou marginale des variables spécifiées pour les modèles catégoriels et de comptage.

`pwcompare [varexo], [cimargins]` : permet de comparer l'impact de chaque modalité d'une variable catégorielle entre elles et de tester si elles ont un effet significativement différent. Voir dans l'application pour avoir les différentes options.

Syntaxe pour predict

`predict p if e(sample), [outcome(label ou valeur de la modalité)]` : prédit la probabilité que la variable dépendante prenne la valeur 1. La restriction `if e(sample)` est ajoutée afin de spécifier que les probabilités soient prédites pour les observations n'ayant aucune valeur manquante. Dans ce cas-ci, il faut aussi spécifier la valeur de la modalité pour laquelle les probabilités sont prédites.

3.4.7 Exemple

Nous allons utiliser la commande `oprobit` pour estimer un modèle probit ordonné dans cette partie en reprenant la même base de données que pour le modèle ordonné.

```
oprobit termgpac c.priGPA c.atndrte i.hwrtec i.frosh

Iteration 0:   log likelihood = -742.82548
Iteration 1:   log likelihood = -544.16931
Iteration 2:   log likelihood = -540.38331
Iteration 3:   log likelihood = -540.38027
Iteration 4:   log likelihood = -540.38027

Ordered probit régression                       Number of obs   =        674
                                                LR chi2(6)      =     404.89
                                                Prob > chi2     =     0.0000
Log likelihood = -540.38027                     Pseudo R2       =     0.2725

------------------------------------------------------------------------------
    termgpac |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      priGPA |    1.52426   .1138352    13.39   0.000     1.301147    1.747373
     atndrte |   .0117412   .0037271     3.15   0.002     .0044363    .0190462
             |
      hwrtec |
          2  |    .580577   .3508759     1.65   0.098    -.1071271    1.268281
          3  |   .7711446   .3277328     2.35   0.019     .1288002    1.413489
          4  |    1.36417   .3264752     4.18   0.000     .7242904     2.00405
             |
     1.frosh |   .1862007   .1146199     1.62   0.104    -.0384502    .4108517
-------------+----------------------------------------------------------------
       /cut1 |   3.340067   .4070045                      2.542353    4.137781
       /cut2 |    4.88707   .4280097                      4.048186    5.725953
       /cut3 |   6.929221   .4577432                       6.03206    7.826381
------------------------------------------------------------------------------
```

Dans ce cas, nous ne pourrons pas calculer de ratio de taux d'incidence par construction du modèle et il faut donc calculer les effets marginaux avec la commande margins. Le lecteur est invité à se référer à l'exemple pour le modèle logistique ordonné comme les commandes et les interprétations sont les mêmes.

Nous allons comparer les mesures d'adéquations des deux modèles avec la commande fitstat en utilisant l'option force comme deux commandes différentes ont été utilisées pour estimer les modèles.

```
fitstat, using(m2) force

Measures of Fit for oprobit of termgpac

Warning: Current model estimated by oprobit, but saved model estimated by ologit

                              Current             Saved        Difference
Model:                        oprobit            ologit
N:                                674               674                 0
Log-Lik Intercept Only        -742.825          -742.825             0.000
Log-Lik Full Model            -540.380          -541.405             1.025
D                            1080.761(663)     1082.810(667)        2.050(4)
LR                            404.890(6)        402.841(4)          2.050(2)
Prob > LR                       0.000             0.000             0.359
McFadden's R2                   0.273             0.271             0.001
McFadden's Adj R2               0.258             0.262            -0.004
ML (Cox-Snell) R2               0.452             0.450             0.002
Cragg-Uhler(Nagelkerke) R2      0.508             0.506             0.002
McKelvey & Zavoina's R2         0.534             0.522             0.011
Variance of y*                  2.144             6.885            -4.741
Variance of error               1.000             3.290            -2.290
Count R2                        0.662             0.662             0.000
Adj Count R2                    0.305             0.305             0.000
AIC                             1.636             1.627             0.009
AIC*n                        1102.761          1096.810             5.950
BIC                         -3237.511         -3261.514            24.003
BIC'                         -365.811          -376.788            10.977
BIC used by Stata            1139.380          1128.403            10.977
AIC used by Stata            1098.761          1096.810             1.950

Difference of   10.977 in BIC' provides very strong support for saved model.

Note: p-valeur  for difference in LR is only valid if models are nested.
```

Nous pouvons voir que le modèle logistique ordonné convient mieux à nos données que le modèle probit ordonné selon le critère d'information bayésien (BIC).

4 Données de comptage

4.1 Introduction

Dans certains cas il peut arriver que la variable dépendante ait une distribution asymétrique. Les données pourraient être analysées en utilisant la régression linéaire multiple, mais la prépondérance de zéros et de petites valeurs ainsi que la nature discrète de la variable dépendante mènent à considérer un modèle permettant une estimation prenant en compte ces caractéristiques.

4.2 Modèle de régression de Poisson

4.2.1 Rappel théorique

Le modèle de régression de Poisson est l'un des modèles le plus utilisé pour ce type de données. Chaque y_i est ainsi tiré d'une distribution de Poisson ayant comme paramètre λ_i, qui est en réalité une fonction des variables exogènes x_i.

La première équation du modèle est donc

$$\Pr(Y_i = y_i|x_i) = \frac{e^{-\lambda_i}\lambda_i{}^{y_i}}{y_i!}, y_i = 0,1,2, \ldots.$$

La modélisation la plus commune pour λ_i est le modèle log-linéaire,

$$\ln(\lambda_i) = x_i\beta$$

Où le nombre de fois où l'évènement se produit par période équivaut à

$$E[y_i|x_i] = Var[y_i|x_i] = \lambda_i = e^{x_i\beta}$$

Et les effets marginaux sont donnés par

$$\frac{\partial E[y_i|x_i]}{\partial x_i} = \lambda_i\beta$$

Le modèle de poisson présuppose une distribution particulière de la variable dépendante. Il existe donc d'autres modèles pour les données de comptage tel que le modèle de régression binomiale négative couvert par la suite.

Ratio de taux d'incidence

Une autre façon d'interpréter les résultats est de calculer les ratios de taux d'incidence (incidence-rate ratios) calculés comme étant e^{β}. Il peut s'interpréter comme étant un facteur multiplicatif du comptage espéré de la variable. Un ratio de taux d'incidence supérieur à 1(inférieur à 1) signifie que pour une variation d'une unité de la variable indépendante, l'espérance de la variable dépendante aura une variation positive (négative).

Mesures de qualité du modèle

Nous pouvons utiliser le critère d'information d'Akaike (AIC) présenté dans le rappel théorique des modèles logistiques et probit binaire.

4.2.2 La commande poisson

Description

La commande poisson dans Stata sert à estimer un modèle de poisson ou régression de poisson, et utilise la méthode du maximum de vraisemblance pour estimer les paramètres ($\hat{\beta}$) du modèle.

Syntaxe

```
poisson [depvar] [indepvars] [if] [in] [weight] [, options]
```

où `depvar` est la variable dépendante
`[indepvars]` est la variable indépendante
`[if]` et `[in]` sont des contraintes dans l'échantillon
`[weight]` assigne des poids aux observations
`[, options]` présente des options supplémentaires dans la régression

Menu

Statistics > Count outcomes > Poisson régression

Option

`exposure(variable)` : inclure $\ln(variable)$ dans le modèle avec un coefficient contraint à 1.

`offset(variable)` : inclure la variable dans le modèle avec un coefficient contraint à 1

Erreur type/Robust

`vce(typevce)` : `typevce` peut être `robust`, `cluster(variable de cluster)`.

Résultats

`irr` : rapporte le ratio taux-incidence. Les erreurs standards et les intervalles de confiance sont similairement transformés. Les paramètres sont cependant estimations de la même manière, seule la forme de présentation change.

4.2.3 Les commandes postestimation après la commande poisson

`test [[label ou valeur de la modalité]] [varexo]`: test de Wald avec lequel nous pouvons tester des restrictions linéaires (par exemple $\beta_1 = \beta_2$ ou $\beta_1 = 0$).

`listcoef [, const help]` : donne une liste des coefficients pour le dernier modèle estimé. L'option `const` inclut la constante dans le tableau et l'option `help` rajoute une légende des variables disponibles dans le tableau. Pour plus de renseignements concernant les options, l'utilisateur peut se référer à l'aide fournie avec le module en utilisant la commande `help listcoef`.

`contrast [varexo]` : permet de tester si les moyennes des effets sont égales entre toutes les modalités d'une variable ordinale incluse dans la régression (x).

`contrast r.[varexo]` : permet de tester si les moyennes des effets sont égales entre les différentes modalités d'une variable ordinale quand elles sont comparées à la catégorie de référence.

`contrast ar.[varexo]` : permet de tester si le coefficient d'une modalité est différent du coefficient de la modalité qui le précède.

`fitstat` : indices de qualité du modèle (fit ou adéquation).

`lrtest [modèlecomplet] [modelcontraint][, stats]` : cette commande performe un test de ratio de vraisemblance (likelihood ratio test) entre un modèle complet et un modèle contraint. Ce test nous renseigne sur la significativité des coefficients non inclus dans le modèle contraint et est similaire au test-F joint qui est calculé avec la commande test. Le modèle contraint est donc un sous modèle du modèle complet. Ce test peut être utilisé pour un test de Chow si nous voulons savoir si nous voulons savoir si les coefficients du modèle sont différents entre plusieurs tranches de l'échantillon.

```
margins, [dydx(varexos)]
[at(varexo1=(valdébut(incrément)valfin)varexo2=(valdébut(incrément)valfin)] [predict(outcome(labe
l ou valeur de la modalité))]:
```

cette commande nous permet de calculer les effets marginaux des variables spécifiées et les prédictions du modèle. L'option `dydx(.)` est nécessaire pour calculer les effets marginaux de variables continues, l'option `at` nous servira pour pouvoir utiliser la commande `marginsplot` afin d'avoir un aperçu graphique de ces effets. Dans le cas d'un modèle logistique multinomial, il est aussi nécessaire de spécifier pour quelle catégorie nous voulons les effets marginaux.

`marginsplot [varexos]` : cette commande doit être utilisée après `margins` pour avoir un aperçu graphique des effets marginaux sur les probabilités.

`prchange [varlist], [help]`: calcule l'impact d'une variation discrète ou marginale des variables spécifiées pour les modèles catégoriels et de comptage.

`pwcompare [varexo], [cimargins]` : permet de comparer l'impact de chaque modalité d'une variable catégorielle entre elles et de tester si elles ont un effet significativement différent. Voir dans l'application pour avoir les différentes options.

`estat gof` : est une mesure d'adéquation du modèle, une statistique de déviance et la statistique de Pearson sont testées.

Syntaxe pour predict

`predict [type] [if] [, statistique nooffset]` : il existe plusieurs prédictions intéressantes selon la statistique demandée.

Si statistique=`n`, c'est le nombre d'évènements qui est prédit (n est considéré par défaut si rien n'est spécifié).

Si statistique=`ir`, c'est le ratio d'incidence exp(xβ) qui est calculé, qui est le nombre d'évènements quand l'exposition est 1. Spécifier ir est équivalent à spécifier n si l'utilisateur n'a utilisé ni `offset()` ni `exposure()` au moment de l'estimation du modèle.

Si statistique=`pr`(n), c'est la probabilité $\Pr(y_j = n)$ qui est calculée, où n est un entier non négatif spécifié comme un nombre ou une variable.

Si statistique=`pr`(a,b), c'est la probabilité $\Pr(a \leq y_j \leq b)$ qui est calculée, où a and b sont des entiers non négatifs spécifiés comme des nombres ou des variables.

b manquant ($b \geq .$) signifie $+\infty$;

`pr(20,.)` calcule $\Pr(y_j \geq 20)$;

`pr(20,b)` calcule $\Pr(y_j \geq 20)$ pour les observations pour lesquelles $b \geq .$ et calcule $\Pr(20 \leq y_j \leq b)$ autrement;

4.2.4 Exemple

Pour cette application, nous allons utiliser un fichier comportant des données sur le nombre d'enfants par femme au Botswana. Nous avons aussi accès à des variables concernant l'âge, le niveau d'éducation, l'état matrimonial, le fait ou non de vivre dans une région urbaine, d'avoir l'électricité et enfin d'avoir la télévision.

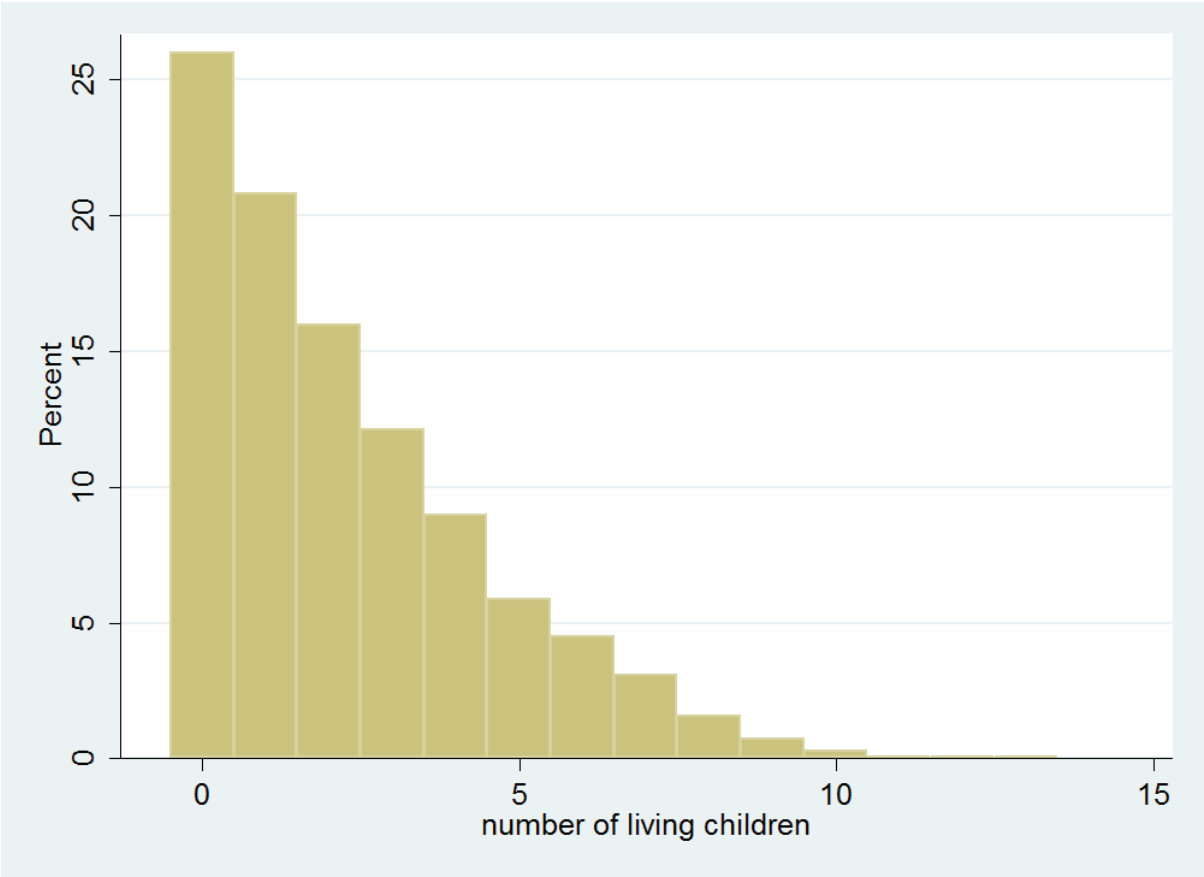
```
use "http://www.Stata.com/data/jwooldridge/eacsap/fertil2.dta", clear
```

Description des données

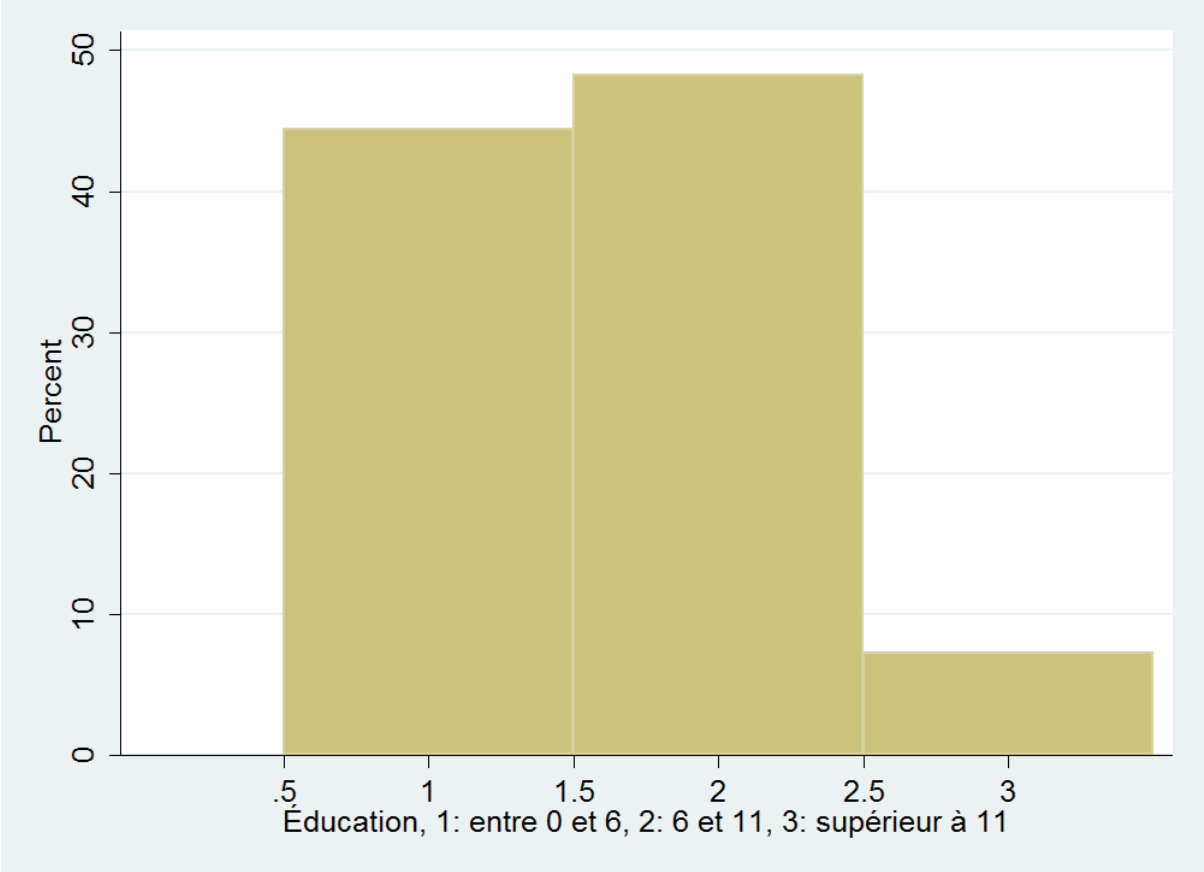
Nous allons premièrement explorer les données d'intérêt.

```
describe children age educc educ7 evermarr urban electric tv

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------------------
children        byte   %8.0g                  number of living children
age             byte   %8.0g                  age in years
educc           float  %9.0g                  Éducation, 1: entre 0 et 6, 2: 6 et 11, 3:
                                              supérieur à 11
educ7           byte   %9.0g                  =1 if educ >= 7
evermarr        byte   %9.0g                  =1 if ever married
urban           byte   %8.0g                  =1 if live in urban area
electric        byte   %8.0g                  =1 if has electricity
tv              byte   %8.0g                  =1 if has tv

summarize children age educc educ7 evermarr urban electric tv

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    children |      4361    2.267828    2.222032          0         13
         age |      4361    27.40518    8.685233         15         49
       educc |      4361    1.628526    .6159596          1          3
       educ7 |      4361    .5556065    .4969553          0          1
    evermarr |      4361    .4767255    .4995153          0          1
-------------+--------------------------------------------------------
       urban |      4361    .5166246    .4997808          0          1
    electric |      4358    .1402019    .3472363          0          1
          tv |      4359    .0929112    .2903413          0          1

inspect children

children:  number of living children            Number of Observations
-------------------------------------         ------------------------------
                                              Total   Integers   Nonintegers
|  #                          Negative           -          -          -
|  #                          Zero            1132       1132          -
|  #                          Positive        3229       3229          -
|  #                                         -----      -----      -----
|  #   #                      Total           4361       4361          -
```

```
|  #   #   .   .   .         Missing            -
+----------------------                      -----
0                     13                      4361
  (14 unique values)

inspect educc

educc:  Éducation, 1: entre 0 et 6, 2: 6 et     Number of Observations
------------------------------------------- -------------------------------
                                             Total   Integers   Nonintegers
|      #                     Negative           -          -          -
|  #   #                     Zero               -          -          -
|  #   #                     Positive        4361       4361          -
|  #   #                                     -----      -----      -----
|  #   #                     Total           4361       4361          -
|  #   #   .                 Missing            -
+----------------------                      -----
1                      3                      4361
   (3 unique values)

histogram children, discrete percent
(start=0, width=1)
```




```
histogram educc, discrete percent
(start=1, width=1)
```




```
tab1 children age educc educ7 evermarr urban electric tv

-> tabulation of children

  number of |
     living |
   children |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,132       25.96       25.96
          1 |        907       20.80       46.76
          2 |        696       15.96       62.71
          3 |        528       12.11       74.82
          4 |        392        8.99       83.81
          5 |        255        5.85       89.66
          6 |        197        4.52       94.18
          7 |        134        3.07       97.25
          8 |         68        1.56       98.81
          9 |         32        0.73       99.54
         10 |         13        0.30       99.84
         11 |          3        0.07       99.91
         12 |          3        0.07       99.98
         13 |          1        0.02      100.00
------------+-----------------------------------
      Total |      4,361      100.00

-> tabulation of age

     age in |
      years |      Freq.     Percent        Cum.
------------+-----------------------------------
         15 |        151        3.46        3.46
         16 |        205        4.70        8.16
         17 |        141        3.23       11.40
         18 |        243        5.57       16.97
```

```
         19 |        205        4.70       21.67
         20 |        219        5.02       26.69
         21 |        172        3.94       30.64
         22 |        206        4.72       35.36
         23 |        156        3.58       38.94
         24 |        196        4.49       43.43
         25 |        184        4.22       47.65
         26 |        201        4.61       52.26
         27 |        149        3.42       55.68
         28 |        192        4.40       60.08
         29 |        153        3.51       63.59
         30 |        163        3.74       67.32
         31 |        118        2.71       70.03
         32 |        138        3.16       73.19
         33 |        106        2.43       75.62
         34 |        118        2.71       78.33
         35 |        105        2.41       80.74
         36 |        103        2.36       83.10
         37 |         70        1.61       84.71
         38 |        102        2.34       87.04
         39 |         73        1.67       88.72
         40 |         75        1.72       90.44
         41 |         72        1.65       92.09
         42 |         42        0.96       93.05
         43 |         35        0.80       93.85
         44 |         48        1.10       94.96
         45 |         44        1.01       95.96
         46 |         48        1.10       97.06
         47 |         44        1.01       98.07
         48 |         50        1.15       99.22
         49 |         34        0.78      100.00
------------+-----------------------------------
      Total |      4,361      100.00

-> tabulation of educc

 Éducation, |
 1: entre 0 |
 et 6, 2: 6 |
  et 11, 3: |
supérieur à |
         11 |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      1,938       44.44       44.44
          2 |      2,105       48.27       92.71
          3 |        318        7.29      100.00
------------+-----------------------------------
      Total |      4,361      100.00

-> tabulation of educ7

 =1 if educ |
       >= 7 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,938       44.44       44.44
          1 |      2,423       55.56      100.00
------------+-----------------------------------
      Total |      4,361      100.00

-> tabulation of evermarr

 =1 if ever |
    married |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,282       52.33       52.33
          1 |      2,079       47.67      100.00
------------+-----------------------------------
      Total |      4,361      100.00

-> tabulation of urban
```

```
 =1 if live |
   in urban |
       area |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,108       48.34       48.34
          1 |      2,253       51.66      100.00
------------+-----------------------------------
      Total |      4,361      100.00


-> tabulation of electric

  =1 if has |
electricity |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      3,747       85.98       85.98
          1 |        611       14.02      100.00
------------+-----------------------------------
      Total |      4,358      100.00

-> tabulation of tv

  =1 if has |
         tv |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      3,954       90.71       90.71
          1 |        405        9.29      100.00
------------+-----------------------------------
      Total |      4,359      100.00
```

Nous voyons que la variable concernant le nombre d'enfants a une distribution comportant un grand nombre de zéros. De plus, cette variable est de nature discrète, c'est le cas où nous pouvons utiliser un modèle de régression de Poisson.

Modélisation

Nous utilisons donc la commande Poisson pour estimer le modèle. Nous spécifions donc $\ln(\lambda) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 educc + \beta_4 evermarr + \beta_5 urban + \beta_6 electric + \beta_7 tv$.

```
poisson children c.age##c.age i.educc i.evermarr i.urban i.electric i.tv

Iteration 0:   log likelihood = -6495.4958
Iteration 1:   log likelihood = -6487.6302
Iteration 2:   log likelihood = -6487.6175
Iteration 3:   log likelihood = -6487.6175

Poisson régression                              Number of obs   =       4358
                                                LR chi2(8)      =    6186.23
                                                Prob > chi2     =     0.0000
Log likelihood = -6487.6175                     Pseudo R2       =     0.3228

------------------------------------------------------------------------------
    children |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .3406061   .0099491    34.23   0.000     .3211063     .360106
             |
 c.age#c.age |  -.0041585   .0001454   -28.61   0.000    -.0044434   -.0038736
             |
       educc |
          2  |  -.1103577   .0244803    -4.51   0.000    -.1583383   -.0623772
          3  |   -.419865   .0525214    -7.99   0.000     -.522805   -.3169251
             |
  1.evermarr |   .3155895   .0244743    12.89   0.000     .2676208    .3635582
```

```
     1.urban |  -.0975435    .021596    -4.52   0.000    -.1398708   -.0552162
  1.electric |  -.1084481   .0386483    -2.81   0.005    -.1841974   -.0326987
        1.tv |  -.0965831   .0481365    -2.01   0.045    -.1909289   -.0022373
       _cons |  -5.473329   .1629479   -33.59   0.000    -5.792701   -5.153957
------------------------------------------------------------------------------
```

Les estimations brutes dans le tableau ci-dessus ne sont pas facilement interprétables et nous allons donc les sauvegarder et l'interprétation sera développée pour le tableau des ratios d'incidences. Il est cependant intéressant de remarquer que tous les coefficients sont significatifs à un niveau de 5%.

Nous utilisons la commande `estimates store` pour sauvegarder les résultats sous le nom m1.

```
estimates store m1
```

Tests d'inférence

Nous pouvons tester des restrictions linéaires sur les coefficients grâce à la commande `test`. Ici, nous voulons tester si les coefficients devant evermarr, la deuxième modalité de la variable educc et la troisième modalité de la variable educc sont simultanément statistiquement différents de 0.

```
test 1.evermarr 2.educc 3.educc

 ( 1)  [children]1.evermarr = 0
 ( 2)  [children]2.educc = 0
 ( 3)  [children]3.educc = 0

           chi2(  3) =  250.63
         Prob > chi2 =    0.0000
```

Nous pouvons clairement rejeter l'hypothèse nulle de nullité simultanée des coefficients dans l'équation précédement estimée.

Nous pouvons aussi effectuer les mêmes restrictions avec un test du rapport de vraisemblance. Nous estimons donc le modèle contraint et sauvegardons les estimations sous le nom mtest. Nous spécifions `if educ7~=. & evermarr ~=.` pour estimer le modèle avec les mêmes observations que le modèle précédent.

```
poisson children c.age##c.age i.urban i.electric i.tv  if educ7~=. & evermarr ~=.

Iteration 0:   log likelihood = -6627.4294
Iteration 1:   log likelihood = -6619.0473
Iteration 2:   log likelihood = -6619.0332
Iteration 3:   log likelihood = -6619.0332

Poisson régression                              Number of obs   =       4358
                                                LR chi2(5)      =    5923.40
                                                Prob > chi2     =     0.0000
Log likelihood = -6619.0332                     Pseudo R2       =     0.3091

------------------------------------------------------------------------------
```

```
    children |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .3752082    .009602    39.08   0.000     .3563886    .3940277
             |
 c.age#c.age |  -.0045424   .0001423   -31.91   0.000    -.0048214   -.0042634
             |
     1.urban |   -.093579   .0214277    -4.37   0.000    -.1355765   -.0515815
  1.electric |  -.1705386   .0381613    -4.47   0.000    -.2453333   -.0957438
        1.tv |  -.2014624   .0459295    -4.39   0.000    -.2914826   -.1114422
       _cons |  -6.007818   .1569878   -38.27   0.000    -6.315509   -5.700128
------------------------------------------------------------------------------
```

```
estimates store mtest
```

Nous testons maintenant si les coefficients sont significatifs en performant le test avec la commande `lrtest`.

```
lrtest m1 mtest

Likelihood-ratio test                                 LR chi2(3)  =    262.83
(Assumption: mtest nested in m1)                      Prob > chi2 =    0.0000
```

Afin de pouvoir savoir si le modèle utilisé est une bonne façon de modéliser les liens supposés entre les variables, nous nous servons de la commande `estat gof`.

```
estat gof

         Déviance  goodness-of-fit =  4152.709
         Prob > chi2(4352)        =    0.9847

         Pearson goodness-of-fit  =  3518.389
         Prob > chi2(4352)        =    1.0000
```

Les deux statistiques sont significatives nous permettant de confirmer l'utilisation du modèle de régression de Poisson pour nos données.

Nous utilisons ensuite la commande `fitstat` pour avoir les mesures d'adéquation du modèle estimé et nous sauvegardons ces mesures sous le nom m2.

```
fitstat, saving(m2)

Measures of Fit for poisson of children

Log-Lik Intercept Only:      -9580.731   Log-Lik Full Model:          -6492.802
D(4349):                     12985.604   LR(7):                        6175.859
                                         Prob > LR:                       0.000
McFadden's R2:                   0.322   McFadden's Adj R2:               0.321
ML (Cox-Snell) R2:               0.758   Cragg-Uhler(Nagelkerke) R2:      0.767
AIC:                             2.984   AIC*n:                       13003.604
BIC:                        -23458.010   BIC':                        -6117.201
BIC used by Stata:           13052.642   AIC used by Stata:           13001.604
```

Nous générons ensuite une variable discrète prédite en utilisant la règle d'affectation suivante

$$childrenhat = 0 \; if \; yhat < 0{,}5$$

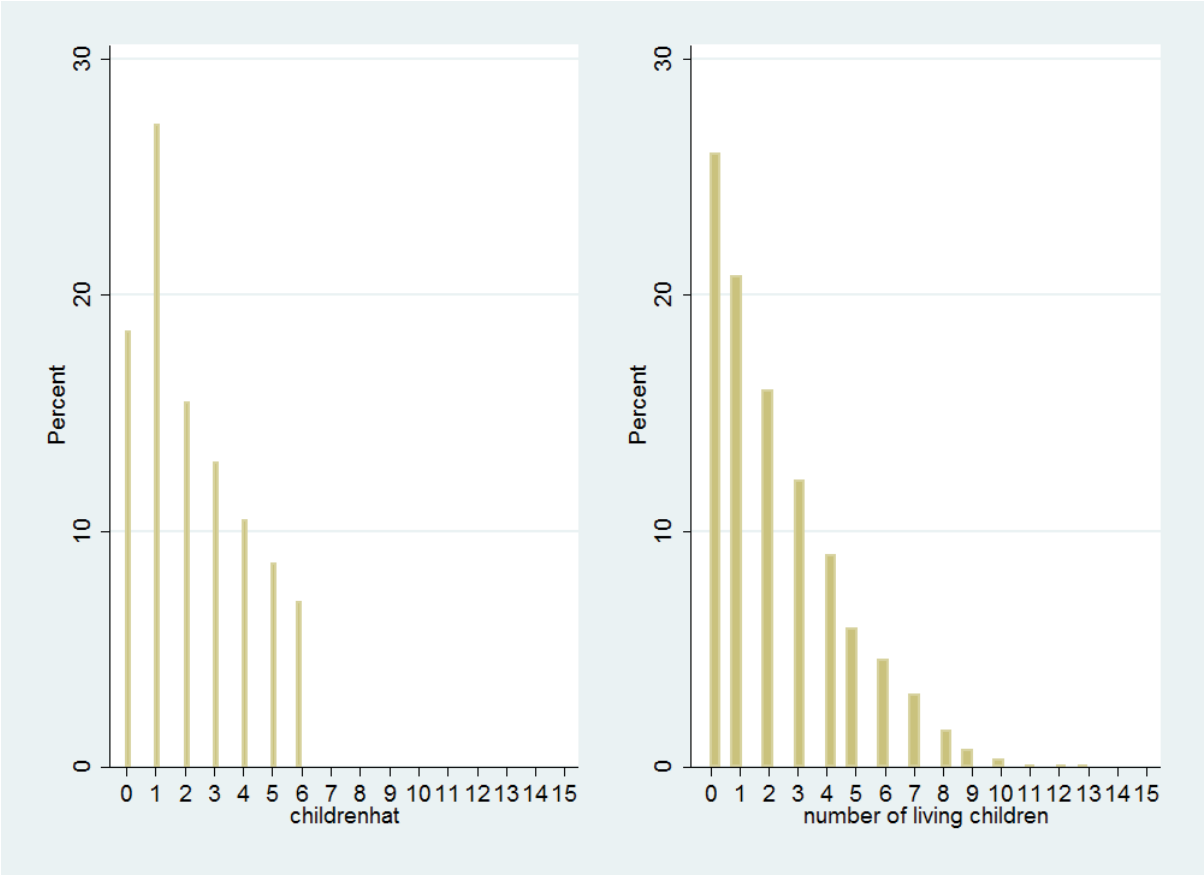
$$childrenhat = 1 \; if \; 0{,}5 < yhat < 1{,}5$$

$$childrenhat = 2\ if\ 1{,}5 < yhat < 2{,}5$$

$$\vdots$$

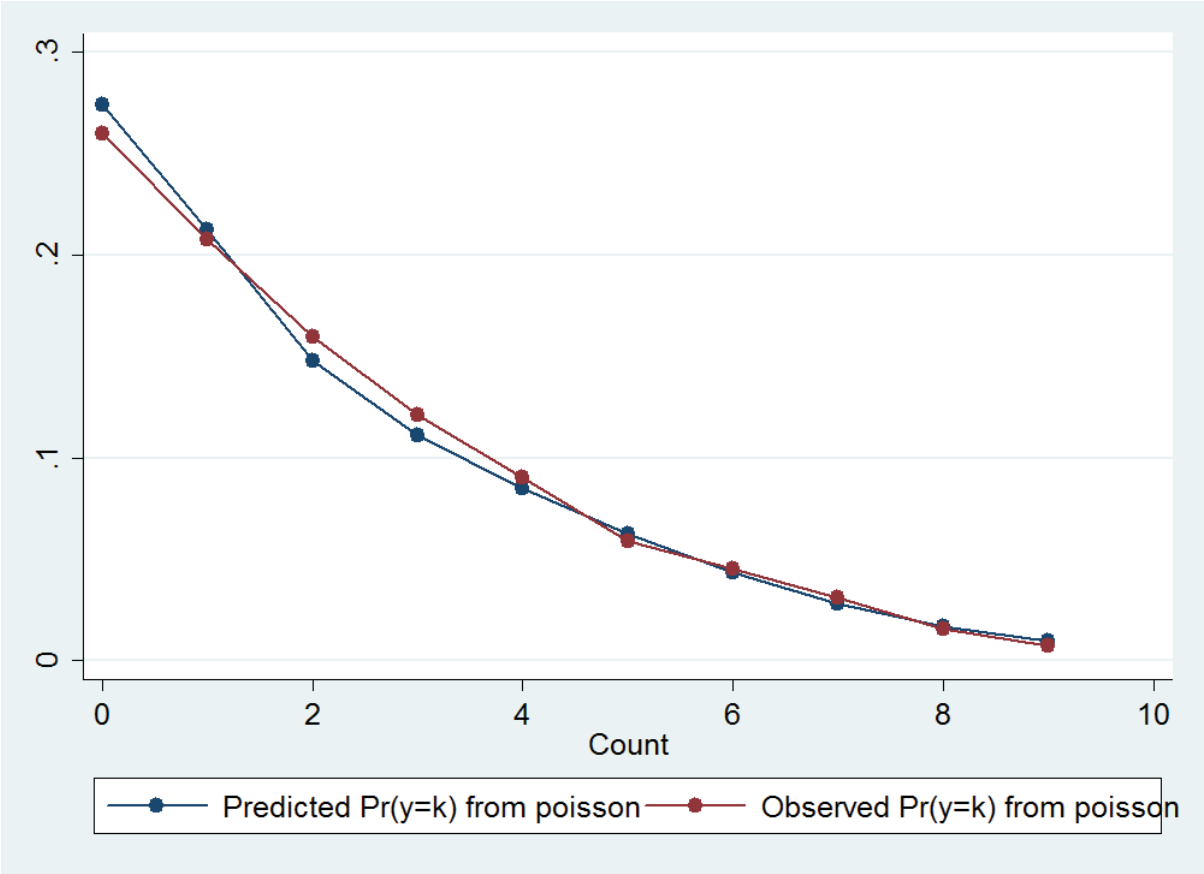
$$childrenhat = 9\ if\ 8{,}5 < yhat < 9{,}5$$

```
predict yhat

gen childrenhat=0 if yhat<0.5
replace childrenhat=1 if yhat>=0.5 & yhat<1.5
replace childrenhat=2 if yhat>=1.5 & yhat<2.5
replace childrenhat=3 if yhat>=2.5 & yhat<3.5
replace childrenhat=4 if yhat>=3.5 & yhat<4.5
replace childrenhat=5 if yhat>=4.5 & yhat<5.5
replace childrenhat=6 if yhat>=5.5 & yhat<6.5
replace childrenhat=7 if yhat>=6.5 & yhat<7.5
replace childrenhat=8 if yhat>=7.5 & yhat<8.5
replace childrenhat=9 if yhat>=8.5 & yhat<9.5
replace childrenhat=. if yhat==.

histogram childrenhat, percent
graph save childrenhatpoisson.gph, replace
histogram children, percent
graph save children2.gph, replace
graph combine yhatpoisson.gph children2.gph, ycommon
```



Nous utilisons la commande `prcounts` du module SPost afin de pouvoir représenter graphiquement les probabilités d'observer les différentes valeurs de la variables dépendante vs les probabilités d'observer les valeurs prédites par le modèle.

```
prcounts po, plot
twoway scatter popreq poobeq poval, c(l l)
```



Nous remarquons que le modèle performe bien si nous regardons le dernier graphique. La probabilité que la variable prédite soit égale un chiffre k est très proche de la probabilité que la variable observé soit égale à ce même chiffre. Nous remarquons tout de même que le modèle prédit plus de femmes avec 1 enfant que de femmes sans enfants si nous reclassons les valeurs prédites avec notre règle d'affectation.

Interprétation des relations entre variable dépendante et variables indépendantes

Nous estimons le modèle avec l'option `irr` qui nous permet d'avoir accès aux ratios de taux d'incidence afin de mieux pouvoir interpréter les coefficients du modèle.

```
poisson children c.age##c.age i.educc i.evermarr i.urban i.electric i.tv , irr

Iteration 0:   log likelihood = -6495.4958
Iteration 1:   log likelihood = -6487.6302
Iteration 2:   log likelihood = -6487.6175
Iteration 3:   log likelihood = -6487.6175

Poisson régression                              Number of obs   =       4358
                                                LR chi2(8)      =    6186.23
                                                Prob > chi2     =     0.0000
Log likelihood = -6487.6175                     Pseudo R2       =     0.3228

------------------------------------------------------------------------------
    children |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.405799   .0139864    34.23   0.000     1.378652    1.433481
             |
 c.age#c.age |   .9958501   .0001448   -28.61   0.000     .9955665    .9961339
             |
       educc |
          2  |   .8955137   .0219225    -4.51   0.000      .853561    .9395285
          3  |   .6571355   .0345137    -7.99   0.000     .5928552    .7283853
             |
  1.evermarr |   1.371067   .0335559    12.89   0.000     1.306852    1.438439
     1.urban |   .9070629   .0195889    -4.52   0.000     .8694706    .9462806
  1.electric |   .8972255   .0346763    -2.81   0.005     .8317716    .9678301
        1.tv |   .9079344   .0437048    -2.01   0.045     .8261913    .9977652
       _cons |   .0041972   .0006839   -33.59   0.000     .0030497    .0057765
------------------------------------------------------------------------------
```

Nous pouvons maintenant interpréter les coefficients comme des facteurs. Pour une femme de notre échantillon qui a l'électricité quand toutes les autres variables restent constantes, nous nous attendons à observer un nombre d'enfants qui baisse par un facteur de 0,90 par rapport à une femme n'ayant pas l'électricité. Si une femme a déjà été mariée (evermarr) et en supposant toutes les autres variables constantes, nous nous attendons à observer un nombre d'enfants 1,37 fois plus grand que pour une femme qui n'a jamais été mariée.

Nous allons maintenant estimer de nouveau le modèle pour pouvoir utiliser les commandes du module SPost.

```
quietly poisson children age agesq educc educc evermarr urban electric tv
```

Premièrement, nous allons opposer directement les coefficients et les ratios de taux d'incidence en mesurant ceux-ci à différentes valeurs avec la commande `listcoef`.

```
Listcoef, help

poisson (N=4358): Factor Change in Expected Count

 Observed SD: 2.2219115

----------------------------------------------------------------------
    children |      b         z     P>|z|     e^b    e^bStdX      SDofX
-------------+--------------------------------------------------------
         age |   0.33818   34.078   0.000   1.4024  18.8669     8.6859
       agesq |  -0.00414  -28.478   0.000   0.9959   0.1131   526.9622
       educc |  -0.15695   -7.929   0.000   0.8547   0.9078     0.6160
       urban |   0.31304   12.796   0.000   1.3676   1.1694     0.4998
```

```
    electric |  -0.09291   -4.314   0.000   0.9113   0.9683     0.3472
          tv |  -0.12196   -3.159   0.002   0.8852   0.9652     0.2904
------------------------------------------------------------------------
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-valeur  for z-test
     e^b = exp(b) = factor change in expected count for unit increase in X
 e^bStdX = exp(b*SD of X) = change in expected count for SD increase in X
   SDofX = standard deviation of X
```

Nous remarquons qu'encore une fois, il est difficile de pouvoir interpréter les coefficients d'âge et âge au carré, il nous faudra donc explorer les relations avec les effets marginaux dérivés plus tard.

Nous allons maintenant estimer de nouveau l'équation en utilisant les opérateurs afin de pouvoir calculer les effets marginaux.

```
quietly poisson children c.age##c.age i.educc i.evermarr i.urban i.electric i.tv
```

Nous commençons premièrement par calculer les effets marginaux moyens des variables avec la commande `margins` et l'option `dydx(*)`.

```
margins, dydx(*)

Average marginal effects                        Number of obs   =       4358
Model VCE    : OIM

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : age 2.educc 3.educc 1.evermarr 1.urban 1.electric 1.tv

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1379834   .0025351    54.43   0.000     .1330148     .142952
             |
       educc |
          2  |  -.2520032   .0550817    -4.58   0.000    -.3599612   -.1440451
          3  |  -.8269311   .0879851    -9.40   0.000    -.9993787   -.6544835
             |
  1.evermarr |   .6770647   .0499144    13.56   0.000     .5792342    .7748953
     1.urban |  -.2205222   .0487033    -4.53   0.000    -.3159788   -.1250655
  1.electric |  -.2361803    .080831    -2.92   0.003    -.3946062   -.0777545
        1.tv |  -.2103574   .1006706    -2.09   0.037    -.4076682   -.0130466
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

En moyenne, une variation d'une année de l'âge d'une femme fait augmenter le nombre d'enfants qu'elle aura de 0,14, *ceteris paribus*. Plus les femmes sont éduquées et moins elles ont tendance à avoir des enfants. Ainsi, toutes choses égales par ailleurs, une femme ayant plus de 11 ans d'éducation a en moyenne 0,83 enfant en moins.

Nous calculons ensuite les effets marginaux des variables à la moyenne en ajoutant l'option `atmeans`.

```
margins, dydx(*) atmeans

Conditional marginal effects                       Number of obs   =       4358
Model VCE    : OIM

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : age 2.educc 3.educc 1.evermarr 1.urban 1.electric 1.tv
at           : age             =    27.40431 (mean)
               1.educc         =    .4444699 (mean)
               2.educc         =    .4825608 (mean)
               3.educc         =    .0729693 (mean)
               0.evermarr      =    .5236347 (mean)
               1.evermarr      =    .4763653 (mean)
               0.urban         =    .4834787 (mean)
               1.urban         =    .5165213 (mean)
               0.electric      =    .8597981 (mean)
               1.electric      =    .1402019 (mean)
               0.tv            =    .9070675 (mean)
               1.tv            =    .0929325 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .2337415   .0058444    39.99   0.000     .2222867    .2451964
             |
       educc |
          2  |  -.2357015   .0523718    -4.50   0.000    -.3383483   -.1330546
          3  |  -.7734381   .0837614    -9.23   0.000    -.9376075   -.6092688
             |
  1.evermarr |   .6622654   .0504512    13.13   0.000     .5633828     .761148
     1.urban |  -.2027401   .0450713    -4.50   0.000    -.2910782   -.1144021
  1.electric |  -.2164503   .0742117    -2.92   0.004    -.3619025   -.0709981
        1.tv |  -.1926925   .0923137    -2.09   0.037    -.3736241   -.0117609
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Le signe des effets marginaux ne change pas, mais nous voyons que pour la variable âge la taille de l'effet change. En effet, pour une variation d'une année supplémentaire de la variable âge, une femme aura 0,23 enfant supplémentaire quand toutes les variables sont à leur moyenne. Nous savons que cet effet est quadratique et nous allons plus tard calculer les effets marginaux de la variable âge à plusieurs valeurs de celle-ci.

Nous calculons maintenant les effets marginaux à la médiane de la variable âge en spécifiant l'option `at((median)c.age)`.

```
margins, dydx(*) at((median)c.age)

Average marginal effects                           Number of obs   =       4358
Model VCE    : OIM

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : age 2.educc 3.educc 1.evermarr 1.urban 1.electric 1.tv
at           : age             =          26 (median)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .2240669   .0049941    44.87   0.000     .2142786    .2338552
             |
       educc |
```

```
          2  |  -.2026057   .0451149    -4.49   0.000    -.2910293   -.1141821
          3  |  -.6648367   .0718866    -9.25   0.000    -.8057318   -.5239416
             |
  1.evermarr |   .5674531   .0439619    12.91   0.000     .4812894    .6536169
     1.urban |   -.175631   .0390159    -4.50   0.000    -.2521007   -.0991612
  1.electric |  -.1875242   .0641318    -2.92   0.003    -.3132203   -.0618281
        1.tv |  -.1670544   .0798657    -2.09   0.036    -.3235884   -.0105205
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

L'interprétation change comme c'est l'effet moyen des variables calculées à la médiane de la variable age qui est calculé. L'effet moyen d'avoir été marié pour une femme est une augmentation de 0,57 enfant quand la variable âge est à sa moyenne, toutes choses égales par ailleurs.

L'effet marginal de la variable âge devrait avoir une magnitude différente dépendamment du niveau de celle-ci. Nous allons observer cet effet graphiquement en calculant l'effet marginal de la variable âge à des niveaux différents. Nous utilisons donc la commande `margins` avec l'option `at(age=(15(1)49))` pour calculer les effets marginaux à la moyenne dans un premier temps et les effets marginaux moyens dans un deuxième temps.

```
margins i.educc, dydx(age) at(age=(15(1)49)) atmeans

Conditional marginal effects                      Number of obs   =       4358
Model VCE    : OIM

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : age

1._at        : age             =          15
               1.educc         =    .4444699 (mean)
               2.educc         =    .4825608 (mean)
               3.educc         =    .0729693 (mean)
               0.evermarr      =    .5236347 (mean)
               1.evermarr      =    .4763653 (mean)
               0.urban         =    .4834787 (mean)
               1.urban         =    .5165213 (mean)
               0.electric      =    .8597981 (mean)
               1.electric      =    .1402019 (mean)
               0.tv            =    .9070675 (mean)
               1.tv            =    .0929325 (mean)

2._at        : age             =          16
               1.educc         =    .4444699 (mean)
               2.educc         =    .4825608 (mean)
               3.educc         =    .0729693 (mean)
               0.evermarr      =    .5236347 (mean)
               1.evermarr      =    .4763653 (mean)
               0.urban         =    .4834787 (mean)
               1.urban         =    .5165213 (mean)
               0.electric      =    .8597981 (mean)
               1.electric      =    .1402019 (mean)
               0.tv            =    .9070675 (mean)
               1.tv            =    .0929325 (mean)

                                  .
                                  .
                                  .
```

```
35._at       : age             =          49
               1.educc         =    .4444699 (mean)
               2.educc         =    .4825608 (mean)
               3.educc         =    .0729693 (mean)
               0.evermarr      =    .5236347 (mean)
               1.evermarr      =    .4763653 (mean)
               0.urban         =    .4834787 (mean)
               1.urban         =    .5165213 (mean)
               0.electric      =    .8597981 (mean)
               1.electric      =    .1402019 (mean)
               0.tv            =    .9070675 (mean)
               1.tv            =    .0929325 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
   _at#educc |
        1 1  |   .0634635    .001971    32.20   0.000     .0596004    .0673267
        1 2  |   .0568325   .0015576    36.49   0.000     .0537795    .0598854
        1 3  |   .0417042   .0022799    18.29   0.000     .0372356    .0461727
        2 1  |   .0754044   .0020706    36.42   0.000     .0713461    .0794627
        2 2  |   .0675257    .001626    41.53   0.000     .0643388    .0707126
        2 3  |   .0495509   .0026164    18.94   0.000     .0444228    .0546791

                                                .
                                                .
                                                .

       34 1  |  -.2307337   .0122892   -18.78   0.000    -.2548201   -.2066472
       34 2  |  -.2066252   .0113987   -18.13   0.000    -.2289663    -.184284
       34 3  |  -.1516233   .0106457   -14.24   0.000    -.1724884   -.1307582
       35 1  |   -.247447   .0110915   -22.31   0.000    -.2691859    -.225708
       35 2  |  -.2215921   .0105558   -20.99   0.000    -.2422812   -.2009031
       35 3  |  -.1626062   .0104854   -15.51   0.000    -.1831572   -.1420552
------------------------------------------------------------------------------
```

Nous utilisons la commande `margins` pour l'aperçu graphique de ces effets en ajoutant des options graphiques afin d'avoir un graphique plus lisible. Les effets marginaux seront représentés par une ligne [`recast(line)`] et les intervalles de confiance par une ligne en pointillés [`recastci(rline) ciopts(lpattern(dash))`].

```
marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) title(Effet marginal à la
moyenne de age) yline(0) xlabel(, angle(90)) ylabel(, angle(0))

  Variables that uniquely identify margins: age educc
```

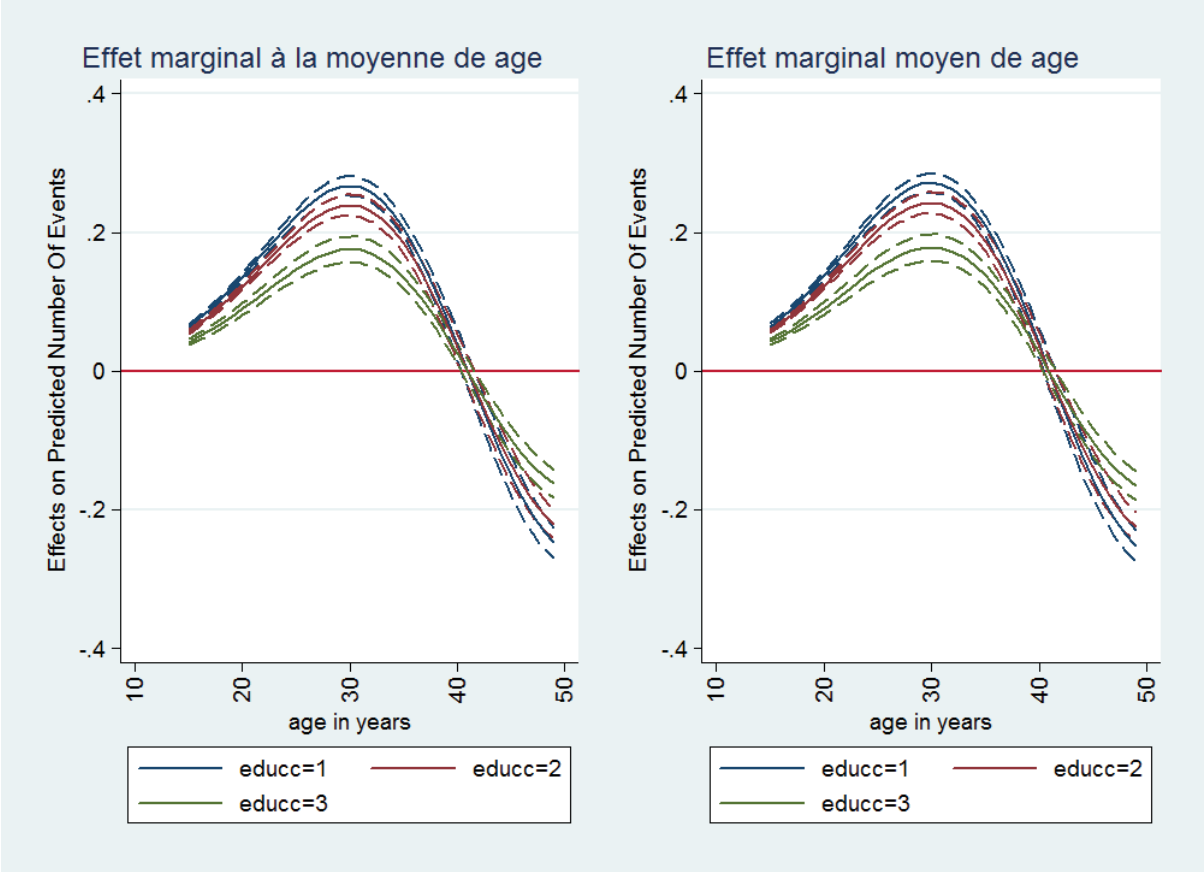
Nous sauvegardons le graphique afin de pouvoir le combiner avec le graphique des effets marginaux moyens.

```
graph save marginspoisson1.gph, replace
(file marginspoisson1.gph saved)
```

Nous répétons les mêmes commandes pour les effets marginaux moyens.

```
margins i.educc, dydx(age) at(age=(15(1)49))

Average marginal effects                        Number of obs   =       4358
Model VCE    : OIM

Expression   : Predicted number of events, predict()
```

```
dy/dx w.r.t. : age

1._at        : age             =          15

2._at        : age             =          16

                              .
                              .
                              .

34._at       : age             =          48

35._at       : age             =          49

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
   _at#educc |
        1 1  |   .0644318   .0020054    32.13   0.000     .0605013    .0683624
        1 2  |   .0576996   .0016032    35.99   0.000     .0545574    .0608417
        1 3  |   .0423404    .002333    18.15   0.000     .0377679    .0469129
        2 1  |   .0765549   .0021007    36.44   0.000     .0724376    .0806721
        2 2  |   .0685559   .0016728    40.98   0.000     .0652772    .0718346
        2 3  |   .0503069   .0026764    18.80   0.000     .0450614    .0555525

                                                   .
                                                   .
                                                   .

       34 1  |   -.234254   .0123443   -18.98   0.000    -.2584484   -.2100596
       34 2  |  -.2097777   .0114901   -18.26   0.000    -.2322979   -.1872574
       34 3  |  -.1539366   .0107898   -14.27   0.000    -.1750842   -.1327891
       35 1  |  -.2512223   .0110955   -22.64   0.000    -.2729691   -.2294755
       35 2  |   -.224973   .0106175   -21.19   0.000     -.245783    -.204163
       35 3  |  -.1650871   .0106257   -15.54   0.000    -.1859131    -.144261
------------------------------------------------------------------------------

marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) title(Effet marginal moyen de
age) yline(0) xlabel(, angle(90)) ylabel(, angle(0))

  Variables that uniquely identify margins: age educc

graph save marginspoisson2.gph, replace
(file marginspoisson2.gph saved)

graph combine marginspoisson1.gph marginspoisson2.gph, ycommon
```



L'interprétation qui va suivre est à considérer toutes choses égales par ailleurs. L'effet marginal moyen et l'effet marginal à la moyenne de la variable âge sont positifs et concaves sur la portion de 15 à 40 ans. Il atteint son maximum aux alentours de 30 ans (nous pouvons vérifier les valeurs exactes dans le tableau de résultats de la commande `margins`). Le point d'inflexion se situe aux alentours de 40 ans et l'effet devient négatif et est convexe de 40 ans à la limite d'âge dans notre échantillon. Nous pouvons aussi observer une différence d'effet marginal entre les niveaux d'éducation. Généralement, plus une femme a un niveau d'éducation élevé moins l'effet marginal de l'âge sur le nombre d'enfants est élevé quand l'âge varie. L'analyse graphique présentée dans cet exemple est loin d'être exhaustive et les tendances et informations repérées sur les graphiques des effets marginaux

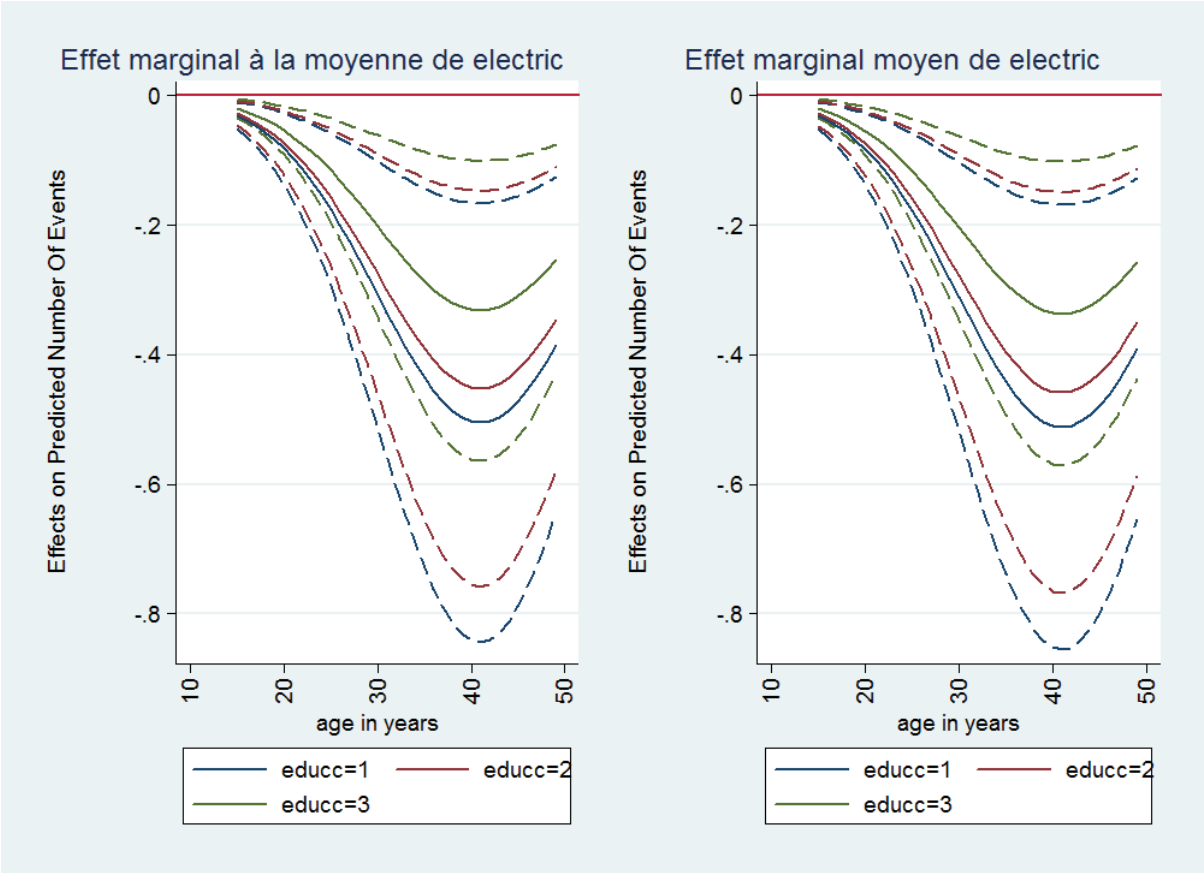
Nous recommençons l'exercice pour observer l'effet marginal de la variable dichotomique electric.

```
margins i.educc, dydx(electric) at(age=(15(1)49)) atmeans

Conditional marginal effects                    Number of obs   =       4358
Model VCE    : OIM

Expression   : Predicted number of events, predict()
```

```
dy/dx w.r.t. : 1.electric




1._at        : age             =          15
               1.educc         =    .4444699 (mean)
               2.educc         =    .4825608 (mean)
               3.educc         =    .0729693 (mean)
               0.evermarr      =    .5236347 (mean)
               1.evermarr      =    .4763653 (mean)
               0.urban         =    .4834787 (mean)
               1.urban         =    .5165213 (mean)
               0.electric      =    .8597981 (mean)
               1.electric      =    .1402019 (mean)
               0.tv            =    .9070675 (mean)
               1.tv            =    .0929325 (mean)

2._at        : age             =          16
               1.educc         =    .4444699 (mean)
               2.educc         =    .4825608 (mean)
               3.educc         =    .0729693 (mean)
               0.evermarr      =    .5236347 (mean)
               1.evermarr      =    .4763653 (mean)
               0.urban         =    .4834787 (mean)
               1.urban         =    .5165213 (mean)
               0.electric      =    .8597981 (mean)
               1.electric      =    .1402019 (mean)
               0.tv            =    .9070675 (mean)
               1.tv            =    .0929325 (mean)

                             .
                             .
                             .

34._at       : age             =          48
               1.educc         =    .4444699 (mean)
               2.educc         =    .4825608 (mean)
               3.educc         =    .0729693 (mean)
               0.evermarr      =    .5236347 (mean)
               1.evermarr      =    .4763653 (mean)
               0.urban         =    .4834787 (mean)
               1.urban         =    .5165213 (mean)
               0.electric      =    .8597981 (mean)
               1.electric      =    .1402019 (mean)
               0.tv            =    .9070675 (mean)
               1.tv            =    .0929325 (mean)

35._at       : age             =          49
               1.educc         =    .4444699 (mean)
               2.educc         =    .4825608 (mean)
               3.educc         =    .0729693 (mean)
               0.evermarr      =    .5236347 (mean)
               1.evermarr      =    .4763653 (mean)
               0.urban         =    .4834787 (mean)
               1.urban         =    .5165213 (mean)
               0.electric      =    .8597981 (mean)
               1.electric      =    .1402019 (mean)
               0.tv            =    .9070675 (mean)
               1.tv            =    .0929325 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.electric   |
   _at#educc |
        1 1  |  -.0306802   .0105555    -2.91   0.004    -.0513686   -.0099918
        1 2  |  -.0274745   .0094825    -2.90   0.004    -.0460599   -.0088891
        1 3  |   -.020161   .0071903    -2.80   0.005    -.0342538   -.0060683
```

```
        2 1  |  -.0379136   .0130202    -2.91   0.004    -.0634328   -.0123944
        2 2  |  -.0339522   .0116986    -2.90   0.004     -.056881   -.0110233
        2 3  |  -.0249144   .0088707    -2.81   0.005    -.0423005   -.0075282

                                                .
                                                .
                                                .

       34 1  |  -.4108062   .1411778    -2.91   0.004    -.6875097   -.1341028
       34 2  |  -.3678826   .1272684    -2.89   0.004    -.6173241   -.1184411
       34 3  |  -.2699554   .0963425    -2.80   0.005    -.4587833   -.0811275
       35 1  |  -.3858136   .1327393    -2.91   0.004    -.6459778   -.1256494
       35 2  |  -.3455014   .1196664    -2.89   0.004    -.5800431   -.1109596
       35 3  |  -.2535318   .0905836    -2.80   0.005    -.4310723   -.0759913
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.

marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) title(Effet marginal à la
moyenne de electric) yline(0) xlabel(, angle(90)) ylabel(, angle(0))

  Variables that uniquely identify margins: age educc

graph save marginspoisson3.gph, replace
(file marginspoisson3.gph saved)

margins i.educc, dydx(electric) at(age=(15(1)49))

Average marginal effects                        Number of obs   =       4358
Model VCE    : OIM

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : 1.electric

1._at        : age             =          15

2._at        : age             =          16

                              .
                              .
                              .

34._at       : age             =          48

35._at       : age             =          49

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.electric   |
   _at#educc |
        1 1  |   -.031106   .0106963    -2.91   0.004    -.0520704   -.0101415
        1 2  |  -.0278558   .0096096    -2.90   0.004    -.0466903   -.0090214
        1 3  |  -.0204408   .0072871    -2.81   0.005    -.0347233   -.0061584
        2 1  |  -.0384398   .0131936    -2.91   0.004    -.0642989   -.0125807
        2 2  |  -.0344234   .0118551    -2.90   0.004    -.0576589   -.0111878
        2 3  |  -.0252602   .0089899    -2.81   0.005      -.04288   -.0076403

                                                .
                                                .
                                                .

       34 1  |  -.4165076   .1430191    -2.91   0.004    -.6968199   -.1361953
       34 2  |  -.3729883   .1289363    -2.89   0.004    -.6256989   -.1202777
       34 3  |  -.2737019   .0976128    -2.80   0.005    -.4650195   -.0823844
       35 1  |  -.3911681   .1344716    -2.91   0.004    -.6547277   -.1276086
       35 2  |  -.3502964   .1212356    -2.89   0.004    -.5879139    -.112679
       35 3  |  -.2570505   .0917786    -2.80   0.005    -.4369332   -.0771678
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

```
marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) title(Effet marginal moyen de
electric) yline(0) xlabel(, angle(90)) ylabel(, angle(0))

  Variables that uniquely identify margins: age educc

graph save marginspoisson4.gph, replace
(file marginspoisson4.gph saved)

graph combine marginspoisson3.gph marginspoisson4.gph, ycommon
```



Nous remarquons clairement une différence dans l'effet marginal moyen et à la moyenne de electric à travers les différentes catégories de la variable educc. Dans les deux cas, l'effet de la variable âge sur l'effet marginal de la variable electric a une allure convexe en plus grande partie avec un minimum aux alentours de 41 ans. Cette fois-ci, nous remarquons qu'un niveau d'éducation plus élevé a un effet moins négatif sur l'effet marginal de electric concernant le nombre d'enfants.

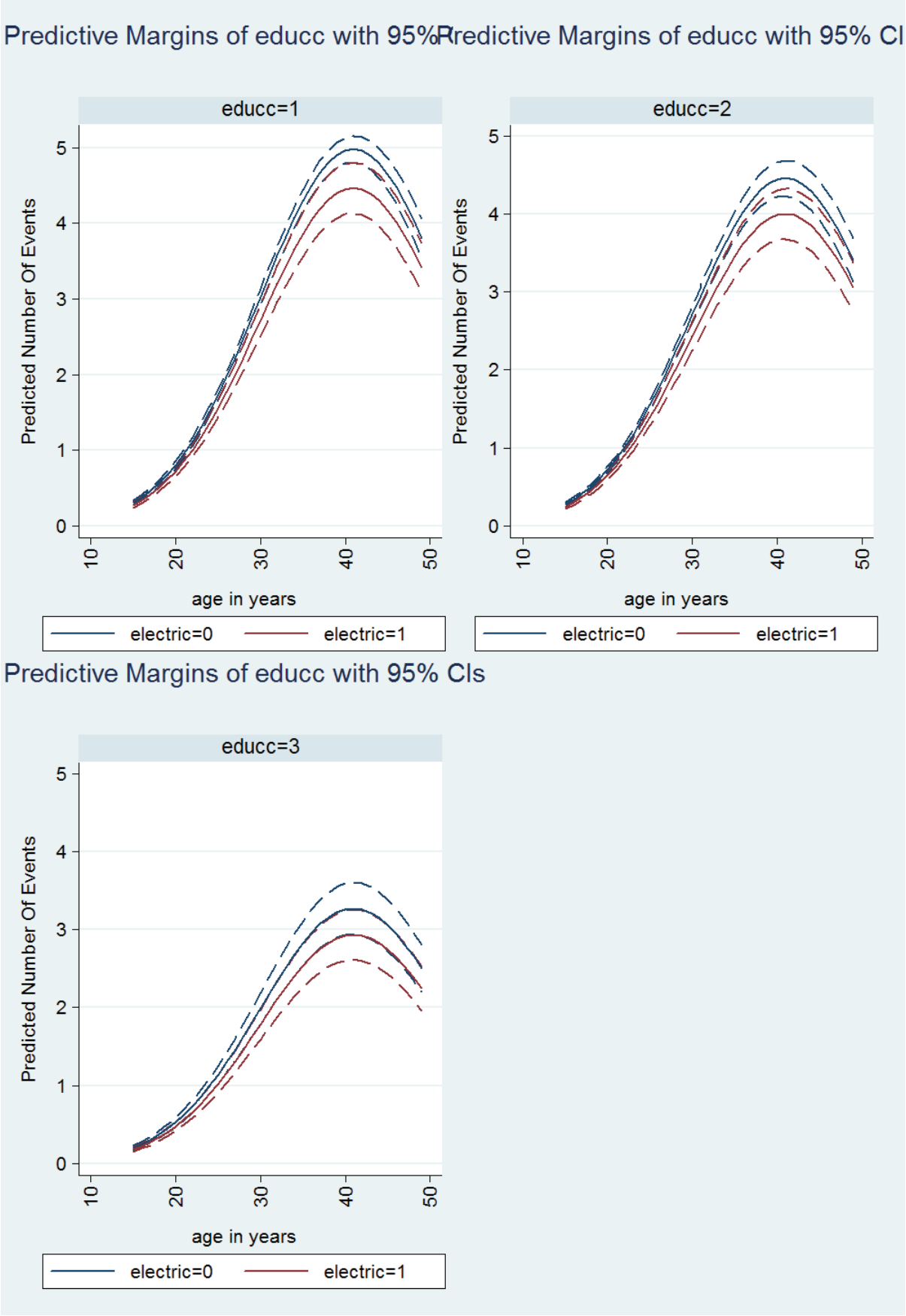
Nous allons maintenant regarder les prédictions de notre modèle à différentes valeurs des variables explicatives. Nous enlevons alors l'option `dydx` qui nous donnait accès aux effets marginaux et rajoutons l'option at pour obtenir une variation de la variable electric en plus de la variation de la variable age [`at(age=(15(1)49)electric=(0 1))`]. Nous séparons aussi le calcul des

effets marginaux selon les catégories de la variable educc. Nous allons donc estimer trois tableaux d'effets marginaux différents et tracer trois graphiques pour les combiner par la suite.

```
margins 1.educc, at(age=(15(1)49)electric=(0 1))

Predictive margins                                 Number of obs   =       4358
Model VCE    : OIM

Expression   : Predicted number of events, predict()

1._at        : age             =          15
               electric        =           0

2._at        : age             =          15
               electric        =           1

                              .
                              .
                              .

69._at       : age             =          49
               electric        =           0

70._at       : age             =          49
               electric        =           1

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _at#educc |
        1 1  |   .3026624   .0160624    18.84   0.000     .2711808    .3341441
        2 1  |   .2715565   .0174477    15.56   0.000     .2373595    .3057534

                                                  .
                                                  .
                                                  .

       69 1  |   3.806081   .1336326    28.48   0.000     3.544166    4.067996
       70 1  |   3.414913   .1662738    20.54   0.000     3.089022    3.740804
------------------------------------------------------------------------------

marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) by(educc) legend(cols(3))
xlabel(, angle(90)) ylabel(0(1)5, angle(0)) title(" ")

  Variables that uniquely identify margins: age electric

graph save marginsglobpoisson1, replace
(file marginsglobpoisson1.gph saved)

margins 2.educc, at(age=(15(1)49)electric=(0 1))

Predictive margins                                 Number of obs   =       4358
Model VCE    : OIM

Expression   : Predicted number of events, predict()

1._at        : age             =          15
               electric        =           0

2._at        : age             =          15
               electric        =           1

                              .
                              .
                              .

69._at       : age             =          49
```

```
               electric        =           0

70._at       : age             =          49
               electric        =           1

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _at#educc |
        1 2  |   .2710384     .01346    20.14   0.000     .2446573    .2974194
        2 2  |   .2431825   .0146927    16.55   0.000     .2143854    .2719796

                                                   .
                                                   .
                                                   .

       69 2  |   3.408398   .1414881    24.09   0.000     3.131086    3.685709
       70 2  |   3.058101   .1598923    19.13   0.000     2.744718    3.371485
------------------------------------------------------------------------------

marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) by(educc) legend(cols(3))
xlabel(, angle(90)) ylabel(0(1)5, angle(0)) title(" ")

  Variables that uniquely identify margins: age electric

graph save marginsglobpoisson2, replace
(file marginsglobpoisson2.gph saved)

margins 3.educc, at(age=(15(1)49)electric=(0 1))

Predictive margins                                 Number of obs   =       4358
Model VCE    : OIM

Expression   : Predicted number of events, predict()

1._at        : age             =          15
               electric        =           0

2._at        : age             =          15
               electric        =           1

                                .
                                .
                                .

69._at       : age             =          49
               electric        =           0

70._at       : age             =          49
               electric        =           1

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _at#educc |
        1 3  |   .1988902   .0139759    14.23   0.000      .171498    .2262825
        2 3  |   .1784494   .0132092    13.51   0.000     .1525599    .2043389

                                                   .
                                                   .
                                                   .

       69 3  |   2.501111   .1544244    16.20   0.000     2.198445    2.803777
       70 3  |   2.244061   .1450833    15.47   0.000     1.959703    2.528419
------------------------------------------------------------------------------

marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) by(educc) legend(cols(3))
xlabel(, angle(90)) ylabel(0(1)5, angle(0)) title(" ")
```

```
  Variables that uniquely identify margins: age electric

graph save marginsglobpoisson3, replace
(note: file marginsglobpoisson3.gph not found)
(file marginsglobpoisson3.gph saved)

graph combine marginsglobpoisson1.gph marginsglobpoisson2.gph marginsglobpoisson3.gph, ycommon
```


Predictive Margins of educc with 95% CI
educc=1
educc=2
educc=3
Predicted Number Of Events
age in years
electric=0
electric=1

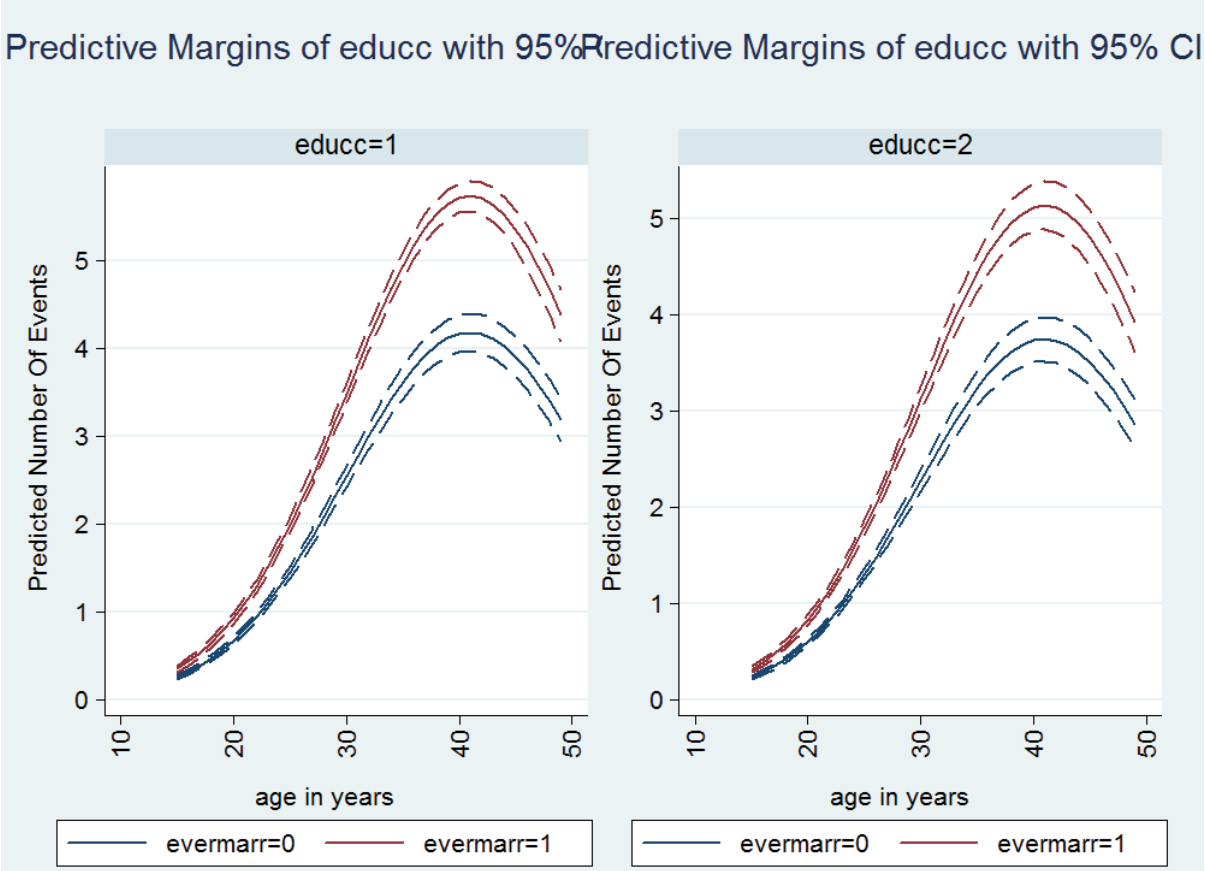
La variable electric a un effet différent à travers les variables d'éducation et la variable age, nous l'avons vu au niveau des effets marginaux. Ici, nous remarquons aussi les différences au niveau des prédictions moyennes qui sont maximales pour les femmes de notre échantillon à 40 ans. Une femme ayant le plus faible niveau d'éducation a en moyenne une différence proche d'un enfant si elle a l'électricité, passant d'environ 5 à environ 4 enfants. Plus le niveau d'éducation est élevé et plus cette différence s'amenuise.

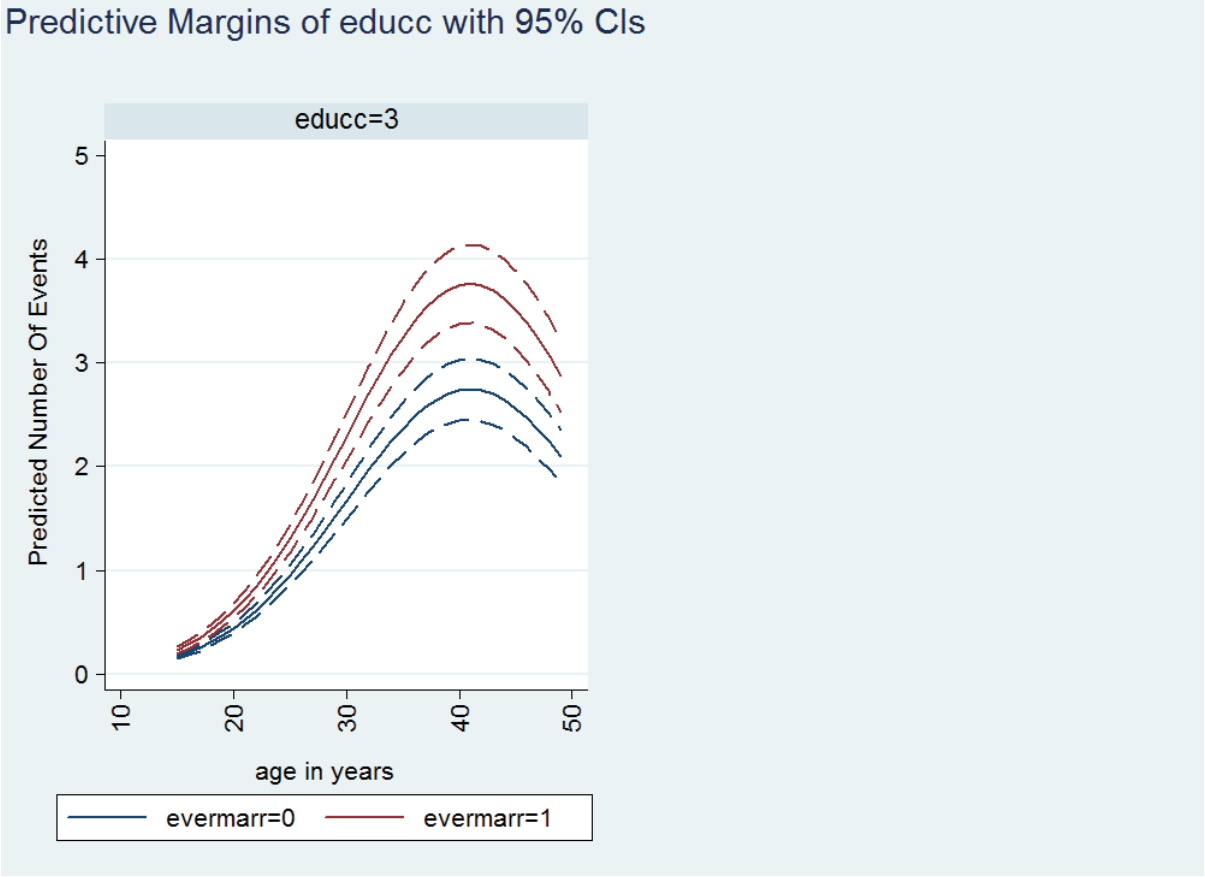
Nous allons maintenant observer s'il existe des différences dans les prédictions moyennes à travers les valeurs de la variable evermarr. Nous utilisons la même méthode que dans l'exemple précédent.

```
margins 1.educc, at(age=(15(1)49)evermarr=(0 1))

Predictive margins                                 Number of obs   =       4358
Model VCE    : OIM

Expression   : Predicted number of events, predict()

1._at        : age             =          15
               evermarr        =           0

2._at        : age             =          15
               evermarr        =           1

                               .
                               .
                               .

69._at       : age             =          49
               evermarr        =           0

70._at       : age             =          49
               evermarr        =           1

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _at#educc |
        1 1  |   .2538803     .01317    19.28   0.000     .2280675    .2796931
        2 1  |    .348087   .0196257    17.74   0.000     .3096214    .3865526

                                                  .
                                                  .
                                                  .

       69 1  |    3.19263   .1261605    25.31   0.000     2.945359      3.4399
       70 1  |    4.37731   .1511821    28.95   0.000     4.080999    4.673621
------------------------------------------------------------------------------

marginsplot, noci by(educc) legend(cols(3)) xlabel(, angle(90)) ylabel(0(1)5, angle(0))

  Variables that uniquely identify margins: age evermarr

graph save marginsglobpoisson4, replace
(note: file marginsglobpoisson4.gph not found)
(file marginsglobpoisson4.gph saved)
```

```
margins 2.educc, at(age=(15(1)49)evermarr=(0 1))

Predictive margins                                     Number of obs   =       4358
Model VCE    : OIM

Expression   : Predicted number of events, predict()

1._at        : age             =          15
               evermarr        =           0

2._at        : age             =          15
               evermarr        =           1

                              .
                              .
                              .

69._at       : age             =          49
               evermarr        =           0

70._at       : age             =          49
               evermarr        =           1

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _at#educc |
        1 2  |   .2273533   .0108464    20.96   0.000     .2060948    .2486118
        2 2  |   .3117167   .0166678    18.70   0.000     .2790484     .344385

                                                  .
                                                  .
                                                  .

       69 2  |   2.859044   .1274518    22.43   0.000     2.609243    3.108845
       70 2  |   3.919941   .1622814    24.16   0.000     3.601876    4.238007
------------------------------------------------------------------------------

marginsplot, noci by(educc) legend(cols(3)) xlabel(, angle(90)) ylabel(0(1)5, angle(0))

  Variables that uniquely identify margins: age evermarr

graph save marginsglobpoisson5, replace
(note: file marginsglobpoisson5.gph not found)
(file marginsglobpoisson5.gph saved)

margins 3.educc, at(age=(15(1)49)evermarr=(0 1))

Predictive margins                                     Number of obs   =       4358
Model VCE    : OIM

Expression   : Predicted number of events, predict()

1._at        : age             =          15
               evermarr        =           0

2._at        : age             =          15
               evermarr        =           1

                              .
                              .
                              .

69._at       : age             =          49
               evermarr        =           0

70._at       : age             =          49
               evermarr        =           1

------------------------------------------------------------------------------
```

```
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _at#educc |
        1 3  |   .1668338   .0114164    14.61   0.000     .1444581    .1892094
        2 3  |   .2287403   .0165944    13.78   0.000      .196216    .2612647

                                                   .
                                                   .
                                                   .

       69 3  |    2.09799   .1328738    15.79   0.000     1.837562    2.358418
       70 3  |   2.876486   .1757968    16.36   0.000      2.53193    3.221041
------------------------------------------------------------------------------

marginsplot, noci by(educc) legend(cols(3)) xlabel(, angle(90)) ylabel(0(1)5, angle(0))

  Variables that uniquely identify margins: age evermarr

graph save marginsglobpoisson6, replace
(note: file marginsglobpoisson6.gph not found)
(file marginsglobpoisson6.gph saved)

graph combine marginsglobpoisson4.gph marginsglobpoisson5.gph marginsglobpoisson6.gph, ycommon
```


Predictive Margins of educc with 95%
Predictive Margins of educc with 95% CI
educc=1
educc=2
Predicted Number Of Events
0
1
2
3
4
5
10
20
30
40
50
age in years
evermarr=0
evermarr=1


Predictive Margins of educc with 95% CIs
educc=3
Predicted Number Of Events
0
1
2
3
4
5
10
20
30
40
50
age in years
evermarr=0
evermarr=1

Cette fois-ci, la différence entre les deux modalités de la variable dichotomique est plus marquée que dans l'exemple précédent. Nous remarquons que l'effet de la variable age atteint encore un maximum aux alentours de 40 ans. La différence de prédiction moyenne dépasse maintenant un enfant si une femme s'est a déjà été mariée passant d'un peu moins de 4 enfants à un peu plus de 5 enfants pour un âge de 40 ans et pour le plus faible niveau d'éducation. Il est aussi intéressant de voir qu'il y a plus de variabilité dans les prédictions quand le niveau d'éducation augmente. Une femme ayant un niveau d'éducation faible aura une plus forte propension à agir de la même façon que la moyenne des femmes ayant le même niveau d'éducation qu'une femme ayant un niveau d'éducation plus élevé.

Nous voulons maintenant tester la différence de coefficients à travers les catégories de la variable educc. Autrement dit, nous voulons savoir si tous les niveaux d'éducation n'ont pas le même effet sur la variable dépendante. Nous utilisons donc la commande `contrast` pour tester l'égalité des coefficients de la variable educc.

```
contrast educc

Contrasts of marginal linear prédictions

Margins      : asbalanced

--------------------------------------------------
             |         df        chi2     P>chi2
-------------+------------------------------------
       educc |          2       70.64     0.0000
--------------------------------------------------
```

Nous pouvons clairement rejeter l'hypothèse nulle d'égalité des coefficients de la variable catégorielle testée. Le test qui suit est un peu plus précis et nous renseigne sur la différence entre un coefficient et la catégorie de base. Nous utilisons simplement le préfixe `r.` devant notre variable à tester.

```
contrast r.educc

Contrasts of marginal linear prédictions

Margins      : asbalanced

--------------------------------------------------
             |         df        chi2     P>chi2
-------------+------------------------------------
       educc |
   (2 vs 1)  |          1       20.32     0.0000
   (3 vs 1)  |          1       63.91     0.0000
      Joint  |          2       70.64     0.0000
--------------------------------------------------

----------------------------------------------------------------
```

```
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       educc |
   (2 vs 1)  |  -.1103577   .0244803     -.1583383   -.0623772
   (3 vs 1)  |   -.419865   .0525214      -.522805   -.3169251
--------------------------------------------------------------
```

Nous pouvons conclure qu'il existe bien une différence entre chaque coefficient et notre catégorie de base pour la variable d'éducation. En effet, toutes nos p-valeurs sont significatives, nous avons donc des évidences statistiques pour rejeter H_0 dans les deux tests.

Nous aimerions aussi savoir s'il existe une différence entre les coefficients de la modalité 3 et de la modalité 2. Nous utilisons donc le préfixe `ar.` pour tester si un coefficient est statistiquement différent du coefficient qui le précède.

```
contrast ar.educc

Contrasts of marginal linear prédictions

Margins      : asbalanced

-------------------------------------------------
             |         df        chi2     P>chi2
-------------+-----------------------------------
       educc |
   (2 vs 1)  |          1       20.32     0.0000
   (3 vs 2)  |          1       35.27     0.0000
      Joint  |          2       70.64     0.0000
-------------------------------------------------

--------------------------------------------------------------
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       educc |
   (2 vs 1)  |  -.1103577   .0244803     -.1583383   -.0623772
   (3 vs 2)  |  -.3095073    .052116     -.4116528   -.2073619
--------------------------------------------------------------
```

Encore une fois, le test qui nous intéresse a une p-valeur très significative et nous pouvons rejeter l'hypothèse nulle d'absence de différence entre les coefficients de la modalité 2 et de la modalité 3.

Nous utilisons enfin la commande `pwcompare` afin de tester une différence entre toutes les paires de coefficients qu'il est possible de former parmi les coefficients de la variable educc. Nous spécifions ici l'option `mcompare(bonferroni)` afin de prendre en compte le nombre de comparaisons effectuées dans le calcul de la p-valeur et les options `cimargins`, `effetc` et `sort` afin d'avoir accès aux différences et leurs intervalles de confiance en ordre de grandeur.

```
pwcompare i.educc, mcompare(bonferroni) cimargins effects sort

Pairwise comparisons of marginal linear prédictions

Margins      : asbalanced
```

```
------------------------------------------------------------
             |                                Unadjusted
             |     Margin   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------
       educc |
          3  |  -5.886687   .1705496     -6.220958   -5.552416
          2  |  -5.577179   .1640658     -5.898742   -5.255616
          1  |  -5.466822   .1675656     -5.795244   -5.138399
------------------------------------------------------------

---------------------------
             |    Number of
             |  Comparisons
-------------+-------------
children     |
       educc |            3
---------------------------

------------------------------------------------------------------------------
             |                            Bonferroni           Bonferroni
             |   Contrast   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
children     |
       educc |
     3 vs 1  |   -.419865   .0525214    -7.99   0.000    -.5456001     -.29413
     3 vs 2  |  -.3095073    .052116    -5.94   0.000     -.434272   -.1847427
     2 vs 1  |  -.1103577   .0244803    -4.51   0.000    -.1689631   -.0517523
------------------------------------------------------------------------------
```

Nous pouvons encore une fois observer la significativité des tests concernant les différences entre chaque paire de coefficients comme toutes les p-valeurs sont significative.

Tests des hypothèses de modélisation

Nous savons que notre modèle arrive bien à cerner les relations qui existent entre les variables que nous avons choisies pour l'analyse. Nous allons aussi vérifier s'il n'y a pas présence de surdispersion dans nos données. Nous voulons donc tester l'hypothèse nulle d'équidispersion des données, c'est-à-dire que $Var(y|x) = E(y|x)$ contre l'hypothèse alternative de sur-dispersion qui est basée sur l'équation $Var(y|x) = E(y|x) + \alpha^2 E(y|x)$ qui est la variance pour le modèle de régression binomiale négative 2 (NB2). Nous testons donc $H_0: \alpha = 0$ contre $H_1: \alpha > 0$. Cette hypothèse peut être testée en construisant une régression d'une transformation des résidus de la régression $\left(\frac{\{(y-\hat{\mu})^2 - y\}}{\hat{\mu}}\right)$ avec la variable dépendante sur les résidus sans constante. Nous devons regarder si le coefficient devant la variable des résidus que nous appellerons muhat est significatif ou non. Un coefficient significatif indique la présence de sur-dispersion et un coefficient non significatif indique le respect de l'hypothèse d'équi-dispersion.

```
quietly poisson children c.age##c.age i.educc i.evermarr i.urban i.electric i.tv

predict muhat, n
(3 missing values generated)

quietly generate ystar=((children-muhat)^2-children)/muhat
```

```
regress ystar muhat, noconstant noheader
------------------------------------------------------------------------------
       ystar |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       muhat |   -.034961   .0060031    -5.82   0.000    -.0467302   -.0231919
------------------------------------------------------------------------------
```

Nous voyons qu'il y a présence de sur-dispersion selon ce test et nous devons trouver un moyen de la corriger.

La première façon est d'utiliser la commande `poisson` avec l'option `robust`. Nous voyons que les variables sont encore toutes significatives même en calculant les erreurs standards robustes.

```
poisson children c.age##c.age i.educc i.evermarr i.urban i.electric i.tv, robust


Iteration 0:   log pseudolikelihood = -6495.4958
Iteration 1:   log pseudolikelihood = -6487.6302
Iteration 2:   log pseudolikelihood = -6487.6175
Iteration 3:   log pseudolikelihood = -6487.6175

Poisson régression                              Number of obs   =       4358
                                                Wald chi2(8)    =    6381.08
                                                Prob > chi2     =     0.0000
Log pseudolikelihood = -6487.6175               Pseudo R2       =     0.3228

------------------------------------------------------------------------------
             |               Robust
    children |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .3406061   .0094572    36.02   0.000     .3220703     .359142
             |
 c.age#c.age |  -.0041585   .0001443   -28.81   0.000    -.0044414   -.0038756
             |
       educc |
          2  |  -.1103577   .0212859    -5.18   0.000    -.1520773   -.0686381
          3  |   -.419865   .0433586    -9.68   0.000    -.5048463   -.3348838
             |
  1.evermarr |   .3155895   .0231645    13.62   0.000      .270188     .360991
     1.urban |  -.0975435   .0199149    -4.90   0.000    -.1365759   -.0585111
  1.electric |  -.1084481   .0366029    -2.96   0.003    -.1801884   -.0367077
        1.tv |  -.0965831   .0447863    -2.16   0.031    -.1843627   -.0088035
       _cons |  -5.473329   .1476551   -37.07   0.000    -5.762728   -5.183931
------------------------------------------------------------------------------
```

La deuxième façon de prendre en compte la sur-dispersion est d'utiliser une forme plus générale pour les données de comptage, la régression négative binomiale, nous posons les bases du modèle dans la prochaine section.

4.3 Modèle de régression binomiale négative

4.3.1 Rappel théorique

Le modèle de régression de Poisson a une hypothèse forte sur la distribution des données. En effet étant donné que la moyenne est égale à la variance, on parle d'équi-dispersion de celles-ci. Si nous soupçonnons une sur-dispersion des données, c'est-à-dire que la variance est plus grande que la moyenne, il est alors possible d'utiliser un modèle de régression binomiale négative comme décrit dans Cameron & Trivedi (2005). Supposons une variable tirée d'une distribution de Poisson

$$f(y|\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}$$

Le paramètre λ est maintenant supposé aléatoire plutôt que déterminé par les variables indépendantes du modèle, les $\boldsymbol{x}$. Nous spécifions $\lambda = \mu v$ ou μ est une fonction des $\boldsymbol{x}$, dans le cas de Stata $\exp(\boldsymbol{x}'\beta)$, et $v > 0$ est i.i.d avec une densité $g(v|\alpha)$. Nous permettons aux observations d'avoir un λ différent ou cette hétérogénéité non observée est due au composant aléatoire non observé v. L'interprétation des paramètres du modèle de régression binomiale négative reste la même que pour le modèle de régression de Poisson si $E[v] = 1$ car cela entraine $E[\lambda|\mu] = \mu$.

Nous allons présenter la *mixture density* (densité mélange) de la fonction obtenue (binomiale négative) après transformation algébrique quand $f(y|\lambda)$ est une densité de Poisson et $g(v|\alpha) = \frac{v^{\delta-1}e^{(-v\delta)}\delta^{\delta}}{\Gamma(\delta)}$

$$h(y|\mu,\delta) = \left(\frac{\Gamma(\alpha^{-1}+y)}{\Gamma(\alpha^{-1})\Gamma(y+1)}\right)\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\mu+\alpha^{-1}}\right)^{y}$$

Il peut alors être montré que la variance est toujours supérieure à la moyenne, pour cette distribution quand $E[v] = 1$, ce qui est notre cas. Dans Stata, la commande `nbreg` utilise une modélisation de la variance quadratique qui correspond au modèle NB2 (negative binomial model type 2). Le modèle est estimé par maximum de vraisemblance.

Effets marginaux

Les effets marginaux sont interprétables exactement comme les effets marginaux du modèle de régression de Poisson.

Mesures de qualité du modèle

Nous pouvons utiliser le critère d'information d'Akaike (AIC) présenté dans le rappel théorique des modèles logistiques et probit binaires.

4.3.2 La commande nbreg

Description

La commande `nbreg` dans Stata sert à estimer un modèle de régression binomiale négative et utilise la méthode du maximum de vraisemblance pour estimer les paramètres ($\hat{\beta}$) du modèle NB2 (negative binomial model type 2).

Syntaxe

```
nbreg [depvar] [indepvars] [if] [in] [weight] [, options]
```

où `depvar` est la variable dépendante
`[indepvars]` est la variable indépendante
`[if]` et `[in]` sont des contraintes dans l'échantillon
`[weight]` assigne des poids aux observations
`[, options]` présente des options supplémentaires dans la régression

Menu

Statistics > Count outcomes > Negative binomial régression

Option

Erreur type/Robust

`vce(typevce)` : `typevce` peut être robust, `cluster (variable de cluster)`.

Résultats

`irr` : rapporte le ratio taux-incidence. Les erreurs standards et les intervalles de confiance sont similairement transformés. Les paramètres sont cependant estimations de la même manière, seule la forme de présentation change.

Les commandes postestimation après la commande nbreg

`test [[label ou valeur de la modalité]] [varexo]`: test de Wald avec lequel nous pouvons tester des restrictions linéaires (par exemple $\beta_1 = \beta_2$ ou $\beta_1 = 0$).

`listcoef [, const help]` : donne une liste des coefficients pour le dernier modèle estimé. L'option `const` inclut la constante dans le tableau et l'option `help` rajoute une légende des variables disponibles dans le tableau. Pour plus de renseignements concernant les options, l'utilisateur peut se référer à l'aide fournie avec le module en utilisant la commande `help listcoef`.

`contrast [varexo]` : permet de tester si les moyennes des effets sont égales entre toutes les modalités d'une variable ordinale incluse dans la régression (x).

`contrast r.[varexo]` : permet de tester si les moyennes des effets sont égales entre les différentes modalités d'une variable ordinale quand elles sont comparées à la catégorie de référence.

`contrast ar.[varexo]` : permet de tester si la moyenne d'un groupe est différente du groupe qui le précède comme la variable est ordinale.

`fitstat` : indices de qualité du modèle (fit ou adéquation).

`lrtest [modèlecomplet] [modelcontraint][, stats]` : cette commande performe un test de ratio de vraisemblance (likelihood ratio test) entre un modèle complet et un modèle contraint. Ce test nous renseigne sur la significativité des coefficients non inclus dans le modèle contraint et est similaire au test-F joint qui est calculé avec la commande test. Le modèle contraint est donc un sous modèle du modèle complet. Ce test peut être utilisé pour un test de Chow si nous voulons savoir si nous voulons savoir si les coefficients du modèle sont différents entre plusieurs tranches de l'échantillon.

```
margins, [dydx(varexos)]
[at(varexo1=(valdébut(incrément)valfin)varexo2=(valdébut(incrément)valfin)] [predict(outcome(labe
l ou valeur de la modalité))]:
```

cette commande nous permet de calculer les effets marginaux des variables spécifiées et les prédictions du modèle. L'option `dydx(.)` est nécessaire pour calculer les effets marginaux de variables continues, l'option `at` nous servira pour pouvoir utiliser la commande `marginsplot` afin d'avoir un aperçu graphique de ces effets. Dans le cas d'un modèle logistique multinomial, il est aussi nécessaire de spécifier pour quelle catégorie nous voulons les effets marginaux.

`marginsplot [varexos]` : cette commande doit être utilisée après `margins` pour avoir un aperçu graphique des effets marginaux sur les probabilités.

`prchange [varlist], [help]`: calcule l'impact d'une variation discrète ou marginale des variables spécifiées pour les modèles catégoriels et de comptage.

`pwcompare [varexo], [cimargins]`, permet de comparer l'impact de chaque modalité d'une variable catégorielle entre elles et de tester si elles ont un effet significativement différent. Voir dans l'application pour avoir les différentes options.

`estat gof` : est une mesure d'adéquation du modèle, une statistique de déviance et la statistique de Pearson sont testées.

Syntaxe pour predict

`predict [type] [if] [, statistique nooffset]` : il existe plusieurs prédictions intéressantes selon la statistique demandée.

Si statistique=n, c'est le nombre d'évènements qui est prédit (n est considéré par défaut si rien n'est spécifié).

Si statistique=ir, c'est le ratio d'incidence exp(xβ) qui est calculé, qui est le nombre d'évènements quand l'exposition est 1. Spécifier ir est équivalent à spécifier n si l'utilisateur n'a utilisé ni `offset()` ni `exposure()` au moment de l'estimation du modèle.

Si statistique=pr(n), c'est la probabilité $\Pr(y_j = n)$ qui est calculée, où n est un entier non négatif spécifié comme un nombre ou une variable.

Si statistique=pr(a,b), c'est la probabilité $\Pr(a \leq y_j \leq b)$ qui est calculée, où a and b sont des entiers non négatifs spécifiés comme des nombres ou des variables.

b manquant ($b \geq .$) signifie $+\infty$;

pr(20,.) calcule $\Pr(y_j \geq 20)$;

pr(20,b) calcule $\Pr(y_j \geq 20)$ pour les observations pour lesquelles $b \geq .$ et calcule $\Pr(20 \leq y_j \leq b)$ autrement.

Exemple

Nous reprenons l'exemple de la section sur le modèle de régression de poisson ou nous avions vu qu'il existait un problème de sur-dispersion des données. Le test est reporté ci-dessous

```
quietly poisson children c.age##c.age i.educc i.evermarr i.urban i.electric i.tv

predict muhat, n
(3 missing values generated)

quietly generate ystar=((children-muhat)^2-children)/muhat

regress ystar muhat, noconstant noheader
------------------------------------------------------------------------------
       ystar |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       muhat |   -.034961   .0060031    -5.82   0.000    -.0467302   -.0231919
------------------------------------------------------------------------------
```

Si nous voulons utiliser le modèle de régression binomiale négative, nous utilisons la commande

```
nbreg.


nbreg children c.age##c.age i.educc i.evermarr i.urban i.electric i.tv

Fitting Poisson model:

Iteration 0:   log likelihood = -6495.4958
Iteration 1:   log likelihood = -6487.6302
Iteration 2:   log likelihood = -6487.6175
Iteration 3:   log likelihood = -6487.6175

Fitting constant-only model:

Iteration 0:   log likelihood = -8770.3433
Iteration 1:   log likelihood = -8698.2669
Iteration 2:   log likelihood = -8698.2589
Iteration 3:   log likelihood = -8698.2589

Fitting full model:

Iteration 0:   log likelihood = -7528.5744
Iteration 1:   log likelihood = -6891.8034
Iteration 2:   log likelihood = -6677.2809
Iteration 3:   log likelihood = -6512.1038
Iteration 4:   log likelihood = -6487.6628
Iteration 5:   log likelihood = -6487.6185
Iteration 6:   log likelihood = -6487.6176
```

```
Iteration 7:   log likelihood = -6487.6175
Iteration 8:   log likelihood = -6487.6174
Iteration 9:   log likelihood = -6487.6173  (not concave)
Iteration 10:  log likelihood = -6487.6173  (backed up)

Negative binomial régression                      Number of obs   =       4358
                                                  LR chi2(8)      =    4421.28
Dispersion     = mean                             Prob > chi2     =     0.0000
Log likelihood = -6487.6173                       Pseudo R2       =     0.2541

------------------------------------------------------------------------------
    children |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .3406077   .0099491    34.24   0.000     .3211078    .3601076
             |
 c.age#c.age |  -.0041585   .0001454   -28.61   0.000    -.0044434   -.0038736
             |
       educc |
          2  |  -.1103566   .0244804    -4.51   0.000    -.1583372   -.0623759
          3  |  -.4198635   .0525214    -7.99   0.000    -.5228035   -.3169234
             |
  1.evermarr |   .3155881   .0244743    12.89   0.000     .2676194    .3635568
     1.urban |  -.0975422    .021596    -4.52   0.000    -.1398695   -.0552148
  1.electric |  -.1084474   .0386484    -2.81   0.005    -.1841968    -.032698
        1.tv |  -.0965825   .0481365    -2.01   0.045    -.1909284   -.0022366
       _cons |  -5.473353   .1629483   -33.59   0.000    -5.792726    -5.15398
-------------+----------------------------------------------------------------
    /lnalpha |  -17.33736    155.205                     -321.5336    286.8588
-------------+----------------------------------------------------------------
       alpha |   2.95e-08   4.59e-06                      2.3e-140    3.8e+124
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) = 3.8e-04 Prob>=chibar2 = 0.492
```

Les coefficients dans le tableau ci-dessus ne sont pas facilement interprétables et il faut donc utiliser les mêmes commandes postestimations que pour la commande `poisson` utilisée dans la section précédente. Le test du rapport de vraisemblance disponible au bas du tableau nous permet de tester à nouveau la présence de sur-dispersion dans les données. Nous voyons que nous n'avons pas assez d'évidences statistiques pour rejeter l'hypothèse de nullité du coefficient alpha dans la régression selon ce test. Il n'y aurait donc pas présence de sur-dispersion. Le modèle de régression de Poisson et le modèle de régression binomiale négative devraient donc être équivalents. Nous allons sauvegarder les estimations et effectuer un test de hausman pour voir s'il existe une différence entre les coefficients des deux modèles.

```
hausman m2 m1

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       m2           m1         Difference          S.E.
-------------+----------------------------------------------------------------
         age |    .3406077     .3406061        1.55e-06        .0000211
 c.age#c.age |   -.0041585    -.0041585       -2.63e-08        3.29e-07
   2bn.educc |   -.1103566    -.1103577        1.17e-06        .0000347
     3.educc |   -.4198635     -.419865        1.57e-06        .0000559
  1.evermarr |    .3155881     .3155895       -1.40e-06        .0000358
     1.urban |   -.0975422    -.0975435        1.33e-06        .0000366
  1.electric |   -.1084474    -.1084481        6.61e-07        .0000496
        1.tv |   -.0965825    -.0965831        6.18e-07        .0000551
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from nbreg
```

```
          B = inconsistent under Ha, efficient under Ho; obtained from poisson

    Test:  Ho:  difference in coefficients not systematic

                  chi2(8) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        0.02
                Prob>chi2 =      1.0000
```

Nous pouvons conclure que les coefficients ne sont pas statistiquement différents et que nous pourrions donc utiliser l'un ou l'autre des modèles afin de pouvoir analyser les données.

```
fitstat, using(m2) force

Measures of Fit for nbreg of children

Warning: Current model estimated by nbreg, but saved model estimated by poisson

                               Current             Saved        Difference
Model:                           nbreg           poisson
N:                                4358              4358                 0
Log-Lik Intercept Only       -8698.259         -9580.731           882.472
Log-Lik Full Model           -6487.617         -6950.697           463.080
D                           12975.235(4343)   13901.394(4350)     926.159(7)
LR                           4421.283(8)       5260.069(6)        838.786(2)
Prob > LR                       0.000             0.000             0.000
McFadden's R2                   0.254             0.275            -0.020
McFadden's Adj R2               0.252             0.274            -0.021
ML (Cox-Snell) R2               0.637             0.701            -0.063
Cragg-Uhler(Nagelkerke) R2      0.649             0.710            -0.060
AIC                             2.984             3.194            -0.209
AIC*n                       13005.235         13917.394          -912.159
BIC                        -23418.100        -22550.599          -867.501
BIC'                        -4354.245         -5209.790           855.545
BIC used by Stata           13059.032         13960.052          -901.020
AIC used by Stata           12995.235         13915.394          -920.159

Note: p-valeur  for difference in LR is only valid if models are nested.
```

Selon les résultats sauvegardés, nous voyons que les différentes mesures d'adéquation n'indiquent pas les mêmes conclusions. Les différents R^2 avantagent le modèle de régression de Poisson alors que le BIC et le AIC avantagent le modèle de régression négative binomiale.

5 Variables dépendantes limitées

5.1 Introduction

Dans certains cas, le chercheur peut vouloir inclure dans sa modélisation une variable dépendante ayant un comportement particulier pour certaines valeurs. En effet, la mesure de certaines variables ne peut être effectuée dans le sens où elle ne peut être observée. La variable y^* a donc une interprétation continue et l'intérêt principal est d'estimer $E(y^*|x)$. Si y^* et x sont observés pour toute la population, alors il est possible d'estimer les effets marginaux des variables exogènes par moindres carrés ordinaires. Dans le cas ou y^* est censuré par le bas (il existe une valeur minimale de y^*) ou par le haut (il existe une valeur maximale de y^*) ou les deux, les estimations obtenues par moindres carrés ordinaires seront biaisées, car certaines observations ne sont pas observables pour toute la population. Par exemple, il existe des sondages où les ménages reportent leurs revenus de façon continue quand celui-ci est inférieur à 100,000 $ et reportent que leurs revenus sont supérieurs à 100,000 $ sinon.

Il existe aussi des situations où nous voulons essayer de modéliser des variables où les décisions peuvent être représentées par des solutions de coin. La variable est alors accumulé à un coin des valeurs possibles de la variable dépendante. Nous pouvons prendre comme exemple le nombre annuel d'heure travaillées qui est la situation couverte par l'exemple.

5.2 Modèle Tobit

5.2.1 Rappel théorique

Quand nous sommes en présence d'une variable dépendante censurée, il convient de définir un modèle comme suit

$$y^*_i = x_i\beta + \varepsilon_i$$

$$y_i = y^*_i \text{ si } a < y^*_i < b$$

$$y_i = a \text{ si } y^*_i \leq a$$

$$y_i = b \text{ si } y^*{}_i \geq b$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

Où $y^*{}_i$ est une variable latente (non observée), a correspond à la valeur minimale observée de la variable $y^*{}_i$ et b correspond à la valeur maximale observée de la variable $y^*{}_i$. Les paramètres du modèle sont estimations par maximum de vraisemblance et ils ne représentent qu'une partie des informations disponibles de l'estimation. Les β mesurent comment la variable non observée $y^*{}_i$ varie par rapport à une variation des variables explicatives.

Les autres effets d'intérêt sont donc :

Les effets marginaux des variables explicatives sur l'espérance de la variable tronquée $E(y_i|a < y_i < b)$ mesurent la variation de la variable $y^*{}_i$ (non observée) pour une variation d'une variable explicative pour la sous-population ou la variable y_i n'est pas observée à a ou à b. C'est la partie de la marge intensive qui est affectée par les variables explicatives.

Effets marginaux

Les effets marginaux des variables explicatives sur l'espérance $E(y_i)$ mesurent la variation de la variable observée y_i pour une variation d'une variable explicative. C'est l'effet total qui est observé.

Les effets marginaux des variables explicatives sur la probabilité $\Pr(a < y_i < b)$ décrivent la variation de la probabilité que les individus aient une observation y_i non censurée par rapport à une variation d'une variable explicative. C'est la partie de l'effet à la marge intensive qui est affectée par les variables explicatives.

Ces effets marginaux peuvent être réécrits

$$\frac{\partial E(y)}{\partial x} = \left[\frac{\partial \Pr(a < y < b)}{\partial x} E(y_i|a < y < b)\right] + \Pr(a < y < b)\frac{\partial E(y|a < y < b)}{\partial x}$$

Mesures de qualité du modèle

Nous pouvons utiliser le critère d'information d'Akaike (AIC) présenté dans le rappel théorique des modèles logistiques et probit binaires.

5.2.2 La commande tobit (modèle Tobit type I)

Description

La commande `tobit` sert à estimer un modèle de régression ou les valeurs de censures sont fixées.

Syntaxe

`tobit [depvar] [indepvars] [if] [in] [weight] , ll[(#)] ul[(#)] [options]`

où `depvar` est la variable dépendante
`[indepvars]` est la variable indépendante
`[if]` et `[in]` sont des contraintes dans l'échantillon
`[weight]` assigne des poids aux observations
`[, options]` présente des options supplémentaires dans la régression

Menu

Statistics > Linear models and related > Censored régression > Tobit régression

Option

Modèle

`ll[(#)]` : censure à gauche, limite inférieure, équivalente à la valeur où les observations sont accumulées alors qu'elles devraient prendre une valeur inférieure à celle-ci.

`ul[(#)]` : censure à droite, limite supérieure, équivalente à la valeur où les observations sont accumulées alors qu'elles devraient prendre une valeur supérieure à celle-ci.

Erreur type/Robust

`vce(typevce)` : `typevce` peut être `robust`, `cluster(variable de cluster)`.

5.2.3 Les commandes postestimation après la commande tobit

`test [[label ou valeur de la modalité].] [varexo]`: test de Wald avec lequel nous pouvons tester des restrictions linéaires (par exemple $\beta_1 = \beta_2$ ou $\beta_1 = 0$).

`listcoef [, const help]` : donne une liste des coefficients pour le dernier modèle estimé. L'option `const` inclut la constante dans le tableau et l'option `help` rajoute une légende des variables disponibles dans le tableau. Pour plus de renseignements concernant les options, l'utilisateur peut se référer à l'aide fournie avec le module en utilisant la commande `help listcoef`.

`contrast [varexo]` : permet de tester si les moyennes des effets sont égales entre toutes les modalités d'une variable ordinale incluse dans la régression (x).

`contrast r.[varexo]` : permet de tester si les moyennes des effets sont égales entre les différentes modalités d'une variable ordinale quand elles sont comparées à la catégorie de référence.

`contrast ar.[varexo]` : permet de tester si la moyenne d'un groupe est différente du groupe qui le précède comme la variable est ordinale.

`fitstat` : indices de qualité du modèle (fit ou adéquation).

`lrtest [modèlecomplet] [modelcontraint][, stats]` : cette commande performe un test de ratio de vraisemblance (likelihood ratio test) entre un modèle complet et un modèle contraint. Ce test nous renseigne sur la significativité des coefficients non inclus dans le modèle contraint et est similaire au test-F joint qui est calculé avec la commande test. Le modèle contraint est donc un sous modèle du modèle complet. Ce test peut être utilisé pour un test de Chow si nous voulons savoir si les coefficients du modèle sont différents entre plusieurs tranches de l'échantillon.

```
margins,
[dydx(varexos)][at(varexo1=(valdébut(incrément)valfin)varexo2=(valdébut(incrément)valfin)][predic
t(e(ll,ul))]:
```

cette commande nous permet de calculer les effets marginaux des variables explicatives sur l'espérance $E(y|a < y < b)$ ($\left(\frac{\partial E(y|a < y < b)}{\partial x}\right)$=marge extensive). L'option `dydx(.)` est nécessaire pour calculer les effets marginaux de variables continues, l'option `at` nous servira pour pouvoir utiliser la commande `marginsplot` afin d'avoir un aperçu graphique de ces effets. Quelques raccourcis comme `dydx(*)` pour sélectionner toutes les variables et `atmeans` pour évaluer les effets marginaux à la moyenne des variables dans l'échantillon peuvent s'avérer très utiles. Si la variable ne possède qu'une borne inférieure alors la syntaxe est `[predict(e(ll,.))]`. Si la variable ne possède qu'une borne supérieure alors la syntaxe est `[predict(e(.,ul))]`.

```
margins,
[dydx(varexos)][at(varexo1=(valdébut(incrément)valfin)varexo2=(valdébut(incrément)valfin)][predic
t(ystar(ll,ul))]:
```

cette commande nous permet de calculer les effets marginaux des variables explicatives sur l'espérance $E(y)$ ($\left(\frac{\partial E(y|a<y<b)}{\partial x}\right)$=marge extensive). L'option `dydx(.)` est nécessaire pour calculer les effets marginaux de variables continues, l'option at nous servira pour pouvoir utiliser la commande `marginsplot` afin d'avoir un aperçu graphique de ces effets. Quelques raccourcis comme `dydx(*)` pour sélectionner toutes les variables et `atmeans` pour évaluer les effets marginaux à la moyenne des variables dans l'échantillon peuvent s'avérer très utiles. Si la variable ne possède qu'une borne inférieure alors la syntaxe est `[predict(e(ll,.))]`. Si la variable ne possède qu'une borne supérieure alors la syntaxe est `[predict(e(.,ul))]`.

`marginsplot [varexos]` : cette commande doit être utilisée après margins pour avoir un aperçu graphique des effets marginaux sur les probabilités.

`pwcompare [varexo], [cimargins]` : permet de comparer l'impact de chaque modalité d'une variable catégorielle entre elles et de tester si elles ont un effet significativement différent. Voir dans l'application pour avoir les différentes options.

Syntaxe pour predict

`predict [type] newvar [if] [in] [, statistic nooffset]` : il existe plusieurs prédictions intéressantes selon la statistique demandée.

Si statistique=xb, c'est la prédiction linéaire qui est prédite (la prédiction par défaut).

Si statistique=pr(a,b), c'est la probabilité $\Pr(a \leq x_j b + u_j \leq b)$ qui est calculée, la probabilité que $y_j|x_j$ soit observé dans l'intervalle (a,b), où a and b sont des entiers non négatifs spécifiés comme des nombres ou des variables. Si lb et ub sont des noms de variables alors

Pr(20,30) calcule $\Pr(20 \leq x_j b + u_j \leq 30)$;

Pr(lb,ub) calcule $\Pr\left(lb \leq x_j b + u_j \leq ub\right)$; et

Pr(20,ub) calcule $\Pr(20 \leq x_j b + u_j \leq ub)$.

a manquant $(a \geq .)$ signifie $-\infty$; pr(.,30) calcule $\Pr(-\infty \leq x_j b + u_j \leq 30)$ et pr(lb,30) calcule $\Pr(-\infty \leq x_j b + u_j \leq 30)$ pour les observations pour lesquelles lb est manquante et $\Pr(lb \leq x_j b + u_j \leq 30)$ autrement.

b manquant ($b \geq .$) signifie $+\infty$; pr(.,30) calcule $\Pr(20 \leq x_j b + u_j \leq +\infty)$ et pr(20,ub) calcule $\Pr\Pr(20 \leq x_j b + u_j \leq +\infty)$ pour les observations pour lesquelles ub est manquante et $\Pr(20 \leq x_j b + u_j \leq ub)$ autrement.

Si statistique=residuals, ce sont les résidus qui seront générés pour chaque individu i.

5.2.4 Exemple

Nous utiliserons la base de données mroz.dta disponible sur le site de Stata qui est une référence utilisée dans Wooldridge (2010) pour l'exemple de cette partie.

```
use "http://www.Stata.com/data/jwooldridge/eacsap/mroz.dta", clear
```

Nous créons par la suite une variable catégorielle pour l'éducation.

```
gen educc=1 if 0<=educ & educ<13
(212 missing values generated)

replace educc=2 if 13<=educ & educ<16
(109 real changes made)

replace educc=3 if educ>=16
(103 real changes made)

replace educc=. if educ==.
(0 real changes made)

label variable educc "Éducation : entre 0 et 8, 2: entre 8 et 12, 3: plus de 12"
```

Nous créons un clone de la variable concernant le revenu des femmes pour que les graphiques que nous allons effectuer par la suite aient une présentation plus compacte.

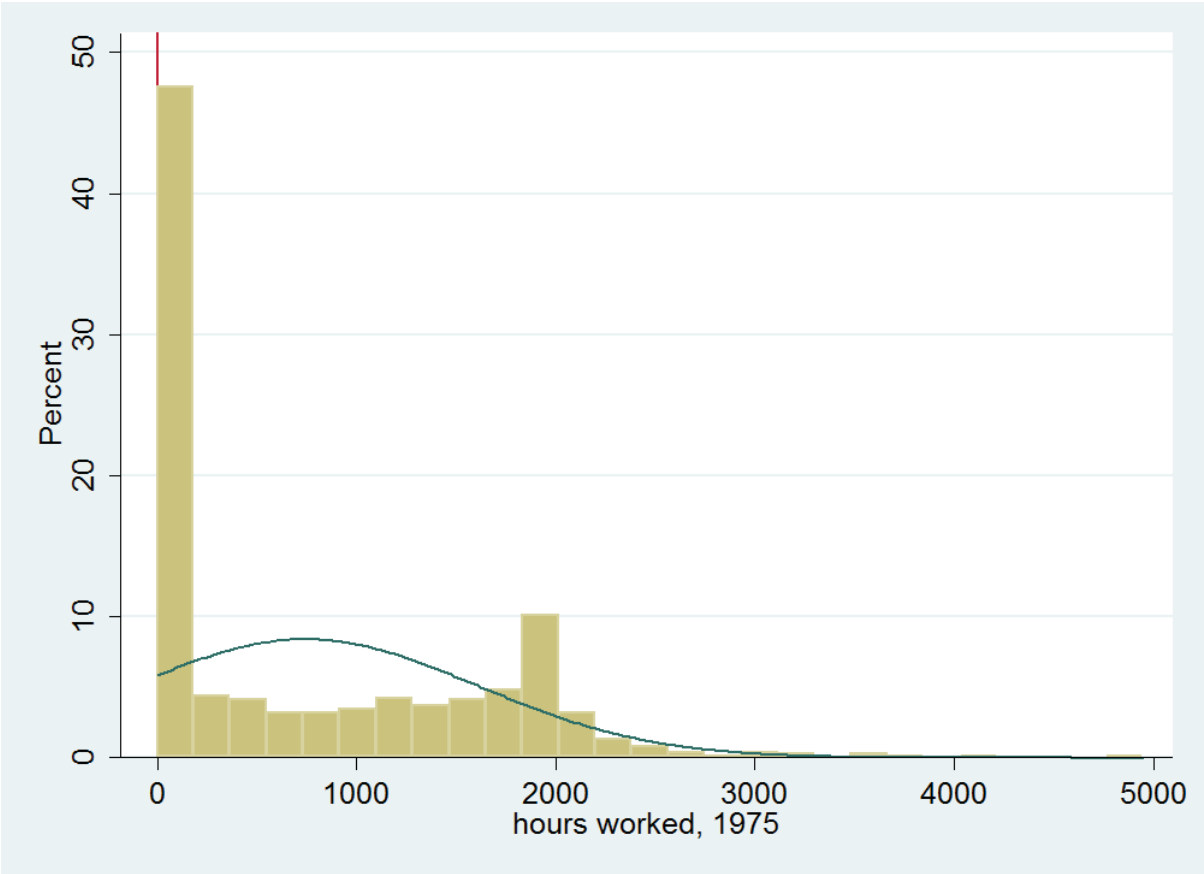
```
gen nwi=nwifeinc
```

Description des données

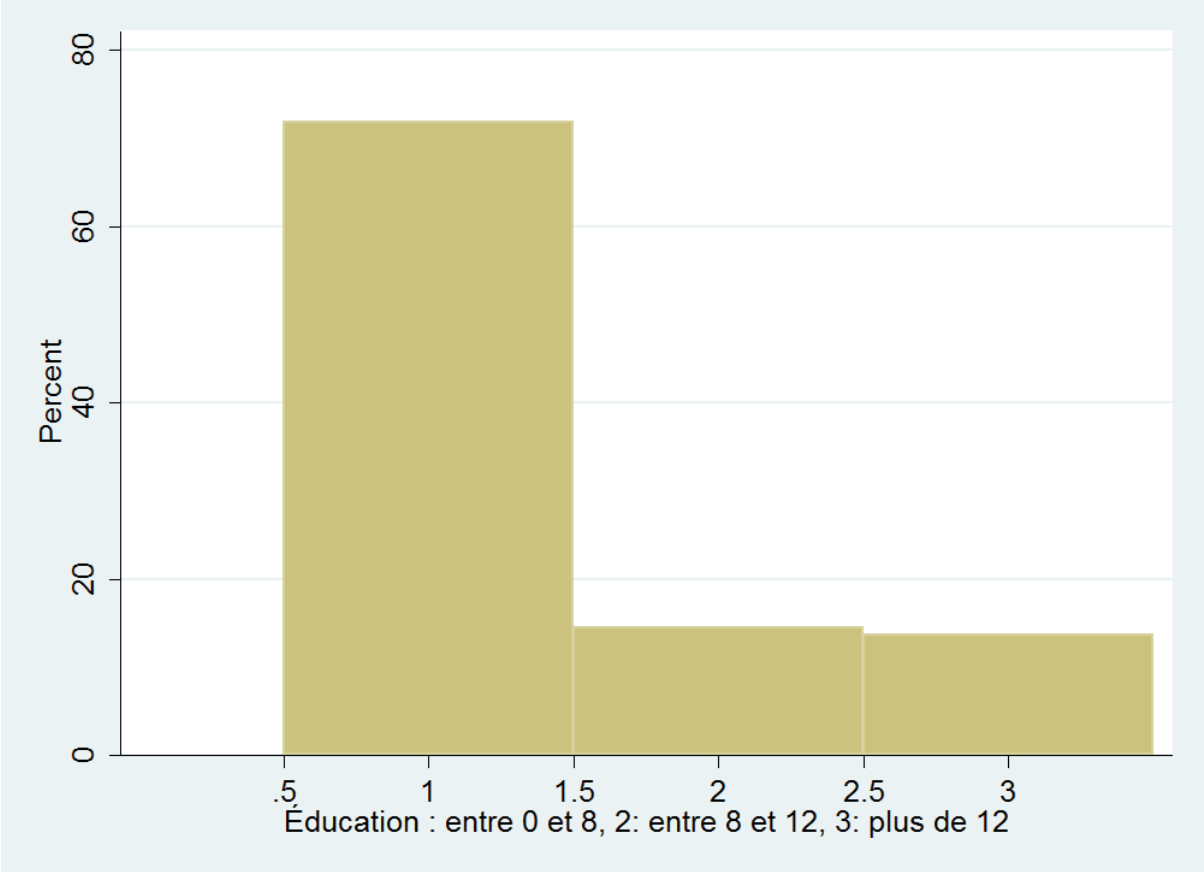
Pour commencer l'analyse, nous explorons les données intéressantes pour cette application dans la base de données.

```
describe hours nwi educc exper age kidslt6 kidsge6

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------------------
hours           int    %9.0g                  hours worked, 1975
nwi             float  %9.0g
educc           float  %9.0g                  Éducation : entre 0 et 8, 2: entre 8 et 12, 3: plus
de 12
exper           byte   %9.0g                  actual labor mkt exper
age             byte   %9.0g                  woman's age in yrs
kidslt6         byte   %9.0g                  # kids < 6 years
kidsge6         byte   %9.0g                  # kids 6-18

summarize hours nwi educc exper age kidslt6 kidsge6

    Variable |       Obs        Mean    Std. Dev.       Min        Max
```

```
-------------+--------------------------------------------------------
       hours |       753    740.5764    871.3142          0       4950
         nwi |       753    20.12896     11.6348  -.0290575         96
       educc |       753    1.418327    .7194367          1          3
       exper |       753    10.63081     8.06913          0         45
         age |       753    42.53785    8.072574         30         60
-------------+--------------------------------------------------------
     kidslt6 |       753    .2377158     .523959          0          3
     kidsge6 |       753    1.353254    1.319874          0          8

inspect hours

hours:  hours worked, 1975                      Number of Observations
---------------------------                   ------------------------------
                                              Total   Integers   Nonintegers
|  #                          Negative            -          -          -
|  #                          Zero              325        325          -
|  #                          Positive          428        428          -
|  #                                          -----      -----      -----
|  #   #                      Total             753        753          -
|  #   #   .   .   .          Missing             -
+----------------------                       -----
0                   4950                        753
(More than 99 unique values)

inspect educc

educc:  Éducation : entre 0 et 8, 2: entre      Number of Observations
------------------------------------------- ------------------------------
                                              Total   Integers   Nonintegers
|  #                          Negative            -          -          -
|  #                          Zero                -          -          -
|  #                          Positive          753        753          -
|  #                                          -----      -----      -----
|  #                          Total             753        753          -
|  #   #   #                  Missing             -
+----------------------                       -----
1                     3                         753
   (3 unique values)

histogram hours, percent normal xline(0)
(bin=27, start=0, width=183.33333)
```




```
histogram educc, discrete percent
(start=1, width=1)
```



Nous remarquons que la distribution de la variable dépendante est asymétrique à gauche avec un grand nombre de 0. Il existe surement une relation latente que nous devrions prendre en compte, car cette variable semble avoir une partie discrète et une partie continue.

Modélisation

Nous allons premièrement commencer par utiliser la régression linéaire et estimer les paramètres par moindres carrés ordinaires pour estimer les effets des variables sélectionnées sur la portion de la variable dépendante supérieure à 0.

```
reg hours c.nwi i.educc c.exper##c.exper c.age if hours>0, robust

Linear régression                                      Number of obs =     428
                                                       F(  6,   421) =    8.96
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1112
                                                       Root MSE      =  737.05

------------------------------------------------------------------------------
                 |               Robust
           hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------
             nwi |  -.3858206   3.305439    -0.12   0.907     -6.88304    6.111399
                 |
           educc |
```

```
             2  |   73.96514    115.075     0.64   0.521     -152.228    300.1583
             3  |  -132.6496   99.45736    -1.33   0.183    -328.1445    62.84522
                |
          exper |   56.04848   14.52667     3.86   0.000     27.49465    84.60231
                |
c.exper#c.exper |  -.7182604   .4042438    -1.78   0.076    -1.512848    .0763272
                |
            age |  -10.05192   5.475659    -1.84   0.067    -20.81495    .7111194
          _cons |   1182.768   218.7514     5.41   0.000      752.787    1612.749
------------------------------------------------------------------------------
```

Les estimations par moindres carrés ordinaires sont biaisées si nous ne prenons pas en compte la nature de la distribution.

Nous recommençons donc l'estimation afin de tenter de corriger le biais avec la commande `tobit` ou nous spécifions la censure à `hours=0` avec l'option `ll(0)` (qui est nécessaire).

```
tobit hours c.nwi i.educc c.exper##c.exper c.age, ll(0)

Tobit régression                                  Number of obs   =        753
                                                  LR chi2(6)      =     202.62
                                                  Prob > chi2     =     0.0000
Log likelihood = -3853.5798                       Pseudo R2       =     0.0256

------------------------------------------------------------------------------
          hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
            nwi |  -9.084071   4.565897    -1.99   0.047    -18.04759   -.1205529
                |
          educc |
             2  |   307.6622   137.5739     2.24   0.026     37.58469    577.7397
             3  |   363.2303   139.6166     2.60   0.009     89.14282    637.3178
                |
          exper |    151.751   17.99108     8.43   0.000     116.4319    187.0701
                |
c.exper#c.exper |  -2.377744   .5636933    -4.22   0.000    -3.484356   -1.271133
                |
            age |  -33.77527   6.755802    -5.00   0.000    -47.03788   -20.51265
          _cons |   640.0859   294.6218     2.17   0.030     61.70069    1218.471
----------------+-------------------------------------------------------------
         /sigma |   1180.015   43.98574                      1093.665    1266.365
------------------------------------------------------------------------------
  Obs. summary:        325  left-censored observations at hours<=0
                       428     uncensored observations
                         0 right-censored observations
```

Comme les estimations ne sont pas directement interprétables (les β sont les effets marginaux sur la variable latente) nous allons pour l'instant comparer le tableau d'estimation par MCO et le tableau d'estimation avec le modèle Tobit. Premièrement, nous pouvons remarquer que les coefficients devant chaque variable ont le même signe. Nous voyons qu'ils sont significatifs pour toutes les variables quand l'estimation est réalisée avec la commande `tobit`, les coefficients des variables nwi la deuxième et la troisième modalité de la variable educc et le terme d'interaction pour la variable exper ne sont pas significatifs dans la régression par MCO. Nous sauvegardons les estimations sous le nom m1 avec la commande `estimates store`.

```
estimates store m1
```

Tests d'inférence

Nous pouvons aussi tester des restrictions linéaires avec la commande test. Nous testons ici si les coefficients de nwi, de la deuxième et de la troisième modalité de la variable educc sont simultanément 0.

```
test nwi 2.educc 3.educc

 ( 1)  [model]nwi = 0
 ( 2)  [model]1b.educc = 0
 ( 3)  [model]2.educc = 0
 ( 4)  [model]3.educc = 0
       Constraint 2 dropped

       F(  3,   747) =    3.62
            Prob > F =    0.0128
```

Nous rejetons l'hypothèse nulle de nullité simultanée des coefficients à un niveau de 5%. Nous pouvons aussi effectuer un test du rapport de vraisemblance pour tester la significativité des coefficients. Nous estimons donc le modèle contraint

```
tobit hours c.exper##c.exper c.age if educc~=. & nwi~=., ll(0)

Tobit régression                                  Number of obs   =        753
                                                  LR chi2(3)      =     191.67
                                                  Prob > chi2     =     0.0000
Log likelihood = -3859.0584                       Pseudo R2       =     0.0242

------------------------------------------------------------------------------
          hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
          exper |   155.4389    17.9682     8.65   0.000      120.165    190.7128
                |
c.exper#c.exper |  -2.369404   .5633412    -4.21   0.000    -3.475317   -1.263491
                |
            age |  -37.19969   6.681946    -5.57   0.000    -50.31724   -24.08215
          _cons |   658.8876   289.3435     2.28   0.023     90.86801    1226.907
----------------+-------------------------------------------------------------
         /sigma |   1187.196   44.28298                      1100.263     1274.13
------------------------------------------------------------------------------
  Obs. summary:        325  left-censored observations at hours<=0
                       428     uncensored observations
                         0 right-censored observations
```

Nous sauvegardons les estimations sous le nom mtest.

```
estimates store mtest
```

Nous effectuons maintenant le test avec la commande `lrtest`

```
lrtest m1 mtest

Likelihood-ratio test                                 LR chi2(3)  =     10.96
(Assumption: mtest nested in m1)                      Prob > chi2 =    0.0120
```

Encore une fois nous ne pouvons pas rejeter l'hypothèse nulle de nullité des coefficients des variables nwi et des modalités de educc.

Nous estimons à nouveau le modèle sans les opérateurs afin de pouvoir utiliser les commandes du module SPost.

```
quietly tobit hours nwi educc exper expersq age kidslt6, ll(0)
```

Nous utilisons la commande `fitstat` pour avoir les mesures d'adéquation du modèle.

```
fitstat

Measures of Fit for tobit of hours

Log-Lik Intercept Only:      -3954.892   Log-Lik Full Model:          -3853.984
D(746):                       7707.969   LR(5):                         201.815
                                         Prob > LR:                       0.000
McFadden's R2:                   0.026   McFadden's Adj R2:               0.024
ML (Cox-Snell) R2:               0.235   Cragg-Uhler(Nagelkerke) R2:      0.235
McKelvey & Zavoina's R2:         0.275
Variance of y*:            1923947.656   Variance of error:         1395535.453
AIC:                            10.255   AIC*n:                        7721.969
BIC:                          2766.416   BIC':                         -168.694
BIC used by Stata:            7754.337   AIC used by Stata:            7721.969
```

Nous générons ensuite plusieurs prédictions du modèle Tobit. Premièrement, des prédictions linéaires avec la commande `predict` sans spécifier d'option (par défaut l'option `xb`).

Prédictions

```
predict yhat
(option xb assumed; fitted values)
```

Deuxièmement les prédictions de la variable dépendante non conditionnellement à ce que celle-ci soit supérieure à 0 avec l'option `ystar(0,.)`.

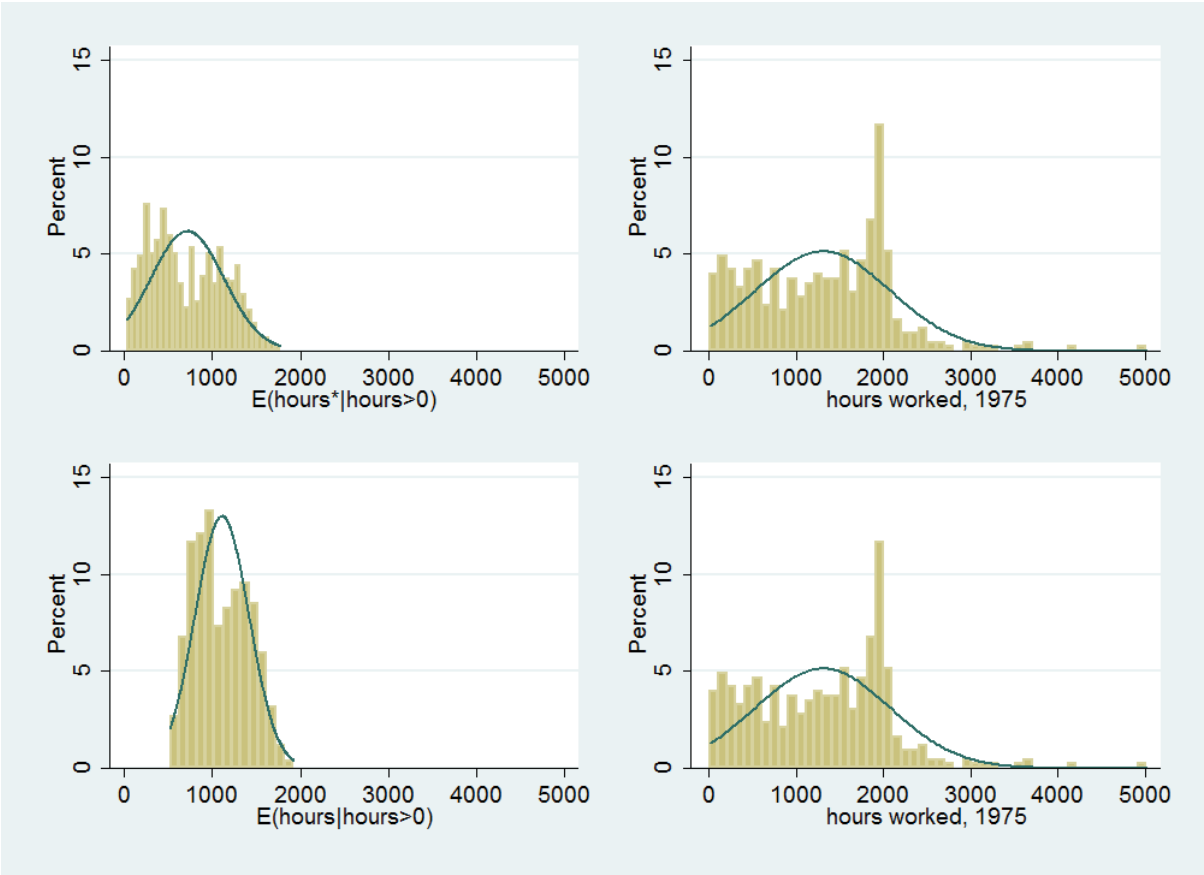
```
predict yhatt, ystar(0,.)
```

Nous continuons avec les prédictions de la variable dépendante conditionnellement à ce que celle-ci soit supérieure à 0 avec l'option `e(0,.)`.

```
predict e, e(0,.)
```

Nous allons à présent comparer les deux dernières prédictions avec la variable dépendante observée à l'aide d'histogrammes. Les histogrammes seront générés séparément et combinés par la suite aux fins de comparaison. Nous utilisons les options `percent` pour que notre histogramme soit considéré en pourcentage et `normal` pour avoir une courbe normale superposée sur les graphiques.

```
histogram yhatt, percent normal xlabel(0(1000)5000) width(100)
graph save yhatttobit.gph, replace
histogram hours if hours>0, percent normal xlabel(0(1000)5000) width(100)
graph save hours2.gph, replace
histogram e, percent normal xlabel(0(1000)5000) width(100)
graph save yhatttobit2.gph, replace
graph combine yhattobit.gph hours2.gph yhatttobit2.gph hours2.gph, ycommon
graph export yhatttobit_hours.tif, replace width(1000)
```



Nous voyons effectivement que la distribution non conditionnelle est proche de la distribution de la variable observée quand nous observons seulement les variables supérieures à 0, mais que la moyenne est plus basse que ce que nous pouvons observer. La distribution conditionnelle à ce que les heures soient positives est beaucoup plus centrée sur la moyenne qui semble correspondre à la moyenne des heures travaillées observées.

Interprétation des relations entre variable dépendante et variables discrètes

Nous estimons à nouveau maintenant le modèle avec les opérateurs afin de calculer les effets marginaux.

```
quietly tobit hours c.nwi i.educc c.exper##c.exper c.age i.kidslt6, ll(0)
```

La commande `margins` et l'option `dydx(*)` nous permettent d'estimer les effets marginaux moyens des variables de contrôle.

```
margins, dydx(*)

Average marginal effects                        Number of obs   =        753
Model VCE    : OIM

Expression   : Linear prediction, predict()
dy/dx w.r.t. : nwi 2.educc 3.educc exper age

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         nwi |  -9.084071   4.565897    -1.99   0.047    -18.03306   -.1350761
             |
       educc |
          2  |   307.6622   137.5739     2.24   0.025     38.02228    577.3021
          3  |   363.2303   139.6166     2.60   0.009     89.58691    636.8738
             |
       exper |   101.1963   8.221971    12.31   0.000     85.08155    117.3111
         age |  -33.77527   6.755802    -5.00   0.000     -47.0164   -20.53414
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Ces effets marginaux se répercutent sur la variable latente (non observable) et ne sont pas d'un grand intérêt.

Nous utilisons donc l'option `predict(e(0,.))` pour avoir les effets marginaux des variables sur le nombre d'heures travaillées pour les femmes qui travaillaient déjà. C'est une composante de la marge intensive. Nous pouvons l'appeler plus généralement l'effet marginal moyen sur la variable dépendante sachant qu'elle n'est pas censurée.

```
margins, dydx(*) predict(e(0,.))

Average marginal effects                        Number of obs   =        753
Model VCE    : OIM

Expression   : E(hours|hours>0), predict(e(0,.))
dy/dx w.r.t. : nwi 2.educc 3.educc exper age

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         nwi |   -4.01539   2.018921    -1.99   0.047    -7.972404   -.0583772
             |
       educc |
          2  |    139.496    65.2165     2.14   0.032     11.67403     267.318
          3  |   166.7831   67.43752     2.47   0.013     34.60799    298.9582
             |
       exper |   40.07545   2.965079    13.52   0.000       34.264     45.8869
         age |  -14.92953   2.995719    -4.98   0.000    -20.80103   -9.058026
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

En moyenne et toutes choses égales par ailleurs, une variation d'une unité de nwi va amener une baisse de 4 heures travaillées par an pour les femmes travaillant déjà. Nous remarquons que toutes choses égales par ailleurs, plus une femme qui travaillait déjà a un diplôme d'études élevé, plus son nombre d'heures travaillé sera grand. Une année d'expérience supplémentaire fera augmenter le nombre d'heures travaillées par année pour les femmes qui travaillaient déjà, ceteris paribus.

Nous allons maintenant observer l'effet marginal moyen sur le nombre d'heures travaillées non conditionnellement à ce que l'individu de l'échantillon travaille. Cet effet est l'effet total, ou l'effet marginal moyen sur la variable dépendante, censurée ou non. L'option utilisée est donc `predict(ystar(0,.))`.

```
margins, dydx(*) predict(ystar(0,.))

Average marginal effects                        Number of obs   =        753
Model VCE    : OIM

Expression   : E(hours*|hours>0), predict(ystar(0,.))
dy/dx w.r.t. : nwi 2.educc 3.educc exper age

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         nwi |  -5.353268   2.686738    -1.99   0.046    -10.61918   -.0873582
             |
       educc |
          2  |   186.4009   87.01701     2.14   0.032     15.85066    356.9511
          3  |   222.8083   89.96426     2.48   0.013     46.48163    399.1351
             |
       exper |   52.89413   3.605872    14.67   0.000     45.82675    59.96151
         age |  -19.90386   3.949572    -5.04   0.000    -27.64488   -12.16284
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

L'effet total sur les heures travaillées des individus de l'échantillon quand la variable nwi varie d'une unité est une diminution des heures travaillées de 5,4 toutes choses égales par ailleurs.

Nous pouvons maintenant nous intéresser aux effets marginaux des variables sur la probabilité qu'une femme ne travaillant pas, un individu avec `hours=0` rentre sur le marché du travail, c'est-à-dire que `hours>0`. C'est un calcul des effets marginaux des variables indépendantes sur la probabilité que la variable dépendante ne soit pas censurée qui est une composante de l'effet à la marge intensive. Nous utilisons l'option `predict(pr(0,.))`.

```
margins, dydx(*) predict(pr(0,.))

Average marginal effects                        Number of obs   =        753
Model VCE    : OIM
```

```
Expression   : Pr(hours>0), predict(pr(0,.))
dy/dx w.r.t. : nwi 2.educc 3.educc exper age

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         nwi |  -.0025161   .0012587    -2.00   0.046     -.004983   -.0000491
             |
       educc |
          2  |   .0848054       .037     2.29   0.022     .0122868    .1573241
          3  |   .0994791   .0371252     2.68   0.007     .0267152    .1722431
             |
       exper |   .0291979   .0019213    15.20   0.000     .0254323    .0329635
         age |  -.0093549   .0018075    -5.18   0.000    -.0128975   -.0058123
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Nous observons que pour une variation d'une unité de la variable nwi une femme ayant des heures travaillées égales à 0 a 0,3 % plus de probabilité d'avoir des heures travaillées supérieures à 0.

Nous pouvons aussi calculer les effets marginaux à la moyenne, nous interpréterons donc l'effet conditionnellement à ce que toutes les variables soient à la moyenne impliquant donc toutes choses égales par ailleurs.

```
margins, dydx(*) predict(e(0,.)) atmeans

Conditional marginal effects                      Number of obs   =        753
Model VCE    : OIM

Expression   : E(hours|hours>0), predict(e(0,.))
dy/dx w.r.t. : nwi 2.educc 3.educc exper age
at           : nwi             =    20.12896 (mean)
               1.educc         =    .7184595 (mean)
               2.educc         =    .1447543 (mean)
               3.educc         =    .1367862 (mean)
               exper           =    10.63081 (mean)
               age             =    42.53785 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         nwi |  -4.148791   2.087055    -1.99   0.047    -8.239344   -.0582378
             |
       educc |
          2  |   144.6831   67.99841     2.13   0.033     11.40868    277.9576
          3  |   173.1296   70.19351     2.47   0.014     35.55283    310.7063
             |
       exper |   46.21743   4.152704    11.13   0.000     38.07828    54.35658
         age |  -15.42552   3.026612    -5.10   0.000    -21.35757   -9.493468
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Pour une femme moyenne travaillant déjà et toutes choses égales par ailleurs, une augmentation d'une unité de la variable nwi fera diminuer le nombre d'heures travaillées annuel de 4,2 heures. Il n'y a pas de grosse différence avec l'effet marginal moyen.

Nous pouvons aussi calculer l'effet marginal des variables sur la variable indépendante non conditionnellement à ce que celle-ci soit supérieure à 0 et l'effet marginal des variables sur la probabilité que la variable indépendante soit non censurée à la moyenne de toutes les autres variables. Nous reportons les sorties Stata ci-dessous, mais nous ne les interprèterons pas. Le lecteur peut se reporter aux tableaux et interprétations des effets marginaux moyens et être prudent afin de prendre en compte la conditionnalité (effets à la moyenne plutôt qu'effets moyens) des prochaines commandes.

```
margins, dydx(*) predict(ystar(0,.)) atmeans

Conditional marginal effects                      Number of obs   =        753
Model VCE    : OIM

Expression   : E(hours*|hours>0), predict(ystar(0,.))
dy/dx w.r.t. : nwi 2.educc 3.educc exper age
at           : nwi              =    20.12896 (mean)
               1.educc          =    .7184595 (mean)
               2.educc          =    .1447543 (mean)
               3.educc          =    .1367862 (mean)
               exper            =    10.63081 (mean)
               age              =    42.53785 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         nwi |  -5.917638   2.976833    -1.99   0.047    -11.75212    -.083152
             |
       educc |
          2  |   205.9716   96.36831     2.14   0.033     17.09322    394.8501
          3  |   246.1805   99.16695     2.48   0.013     51.81683    440.5441
             |
       exper |   65.92234   5.918619    11.14   0.000     54.32206    77.52262
         age |  -22.00223   4.317148    -5.10   0.000    -30.46368   -13.54077
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.

margins, dydx(*) predict(pr(0,.)) atmeans

Conditional marginal effects                      Number of obs   =        753
Model VCE    : OIM

Expression   : Pr(hours>0), predict(pr(0,.))
dy/dx w.r.t. : nwi 2.educc 3.educc exper age
at           : nwi              =    20.12896 (mean)
               1.educc          =    .7184595 (mean)
               2.educc          =    .1447543 (mean)
               3.educc          =    .1367862 (mean)
               exper            =    10.63081 (mean)
               age              =    42.53785 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         nwi |  -.0028472   .0014308    -1.99   0.047    -.0056514   -.0000429
             |
       educc |
          2  |   .0942155    .040383     2.33   0.020     .0150663    .1733648
          3  |   .1099664   .0401473     2.74   0.006     .0312791    .1886537
             |
       exper |   .0317174   .0024672    12.86   0.000     .0268817    .0365531
```

```
         age |   -.010586   .0021532    -4.92   0.000    -.0148062   -.0063658
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Nous pouvons aussi essayer de visualiser les effets marginaux en les calculant à plusieurs valeurs d'une ou plusieurs variables. Nous utiliserons donc l'option `at(age=(30(1)60))` pour faire varier l'âge de 30 à 60 ans par un incrément d'une unité. Nous allons par la suite représenter graphiquement les sorties Stata avec la commande `marginsplot`. Nous calculerons l'effet marginal moyen pour les trois types d'effets marginaux.

Nous commençons par calculer les effets marginaux pour la variable dépendante conditionnellement à ce qu'elle soit censurée.

```
margins i.educc, dydx(nwi) at(age=(30(1)60)) predict(e(0,.))

Average marginal effects                        Number of obs   =        753
Model VCE    : OIM

Expression   : E(hours|hours>0), predict(e(0,.))
dy/dx w.r.t. : nwi

1._at        : age             =          30

2._at        : age             =          31

                               .
                               .
                               .

30._at       : age             =          59

31._at       : age             =          60

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
nwi          |
   _at#educc |
        1 1  |  -4.624814   2.283835    -2.03   0.043    -9.101048   -.1485792
        1 2  |  -5.230011   2.646776    -1.98   0.048     -10.4176   -.0424243
        1 3  |  -5.340713   2.723481    -1.96   0.050    -10.67864   -.0027888
        2 1  |  -4.559701    2.25292    -2.02   0.043    -8.975344    -.144059
        2 2  |  -5.162833   2.614899    -1.97   0.048    -10.28794   -.0377256
        2 3  |  -5.273405   2.691574    -1.96   0.050    -10.54879    .0019837

                                                   .
                                                   .
                                                   .

       30 1  |  -2.945014    1.50497    -1.96   0.050    -5.894701    .0046737
       30 2  |  -3.418148   1.798171    -1.90   0.057    -6.942499    .1062024
       30 3  |  -3.509488   1.863722    -1.88   0.060    -7.162317    .1433409
       31 1  |  -2.896526   1.482968    -1.95   0.051    -5.803089    .0100375
       31 2  |  -3.363488   1.772842    -1.90   0.058    -6.838194    .1112186
       31 3  |  -3.453762   1.837743    -1.88   0.060    -7.055673    .1481485
------------------------------------------------------------------------------
```

La commande `marginsplot` nous permet par la suite de représenter graphiquement ces effets marginaux. L'option `recast(line)` nous permet de représenter l'effet marginal par une ligne et

l'option `recastci(rline) ciopts(lpattern(dash))` nous permet de représenter les intervalles de confiance par des lignes discontinues.

```
marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) title(Effet marginal moyen
conditionnel de nwi) yline(0) ylabel(, angle(0))

  Variables that uniquely identify margins: age educc
```

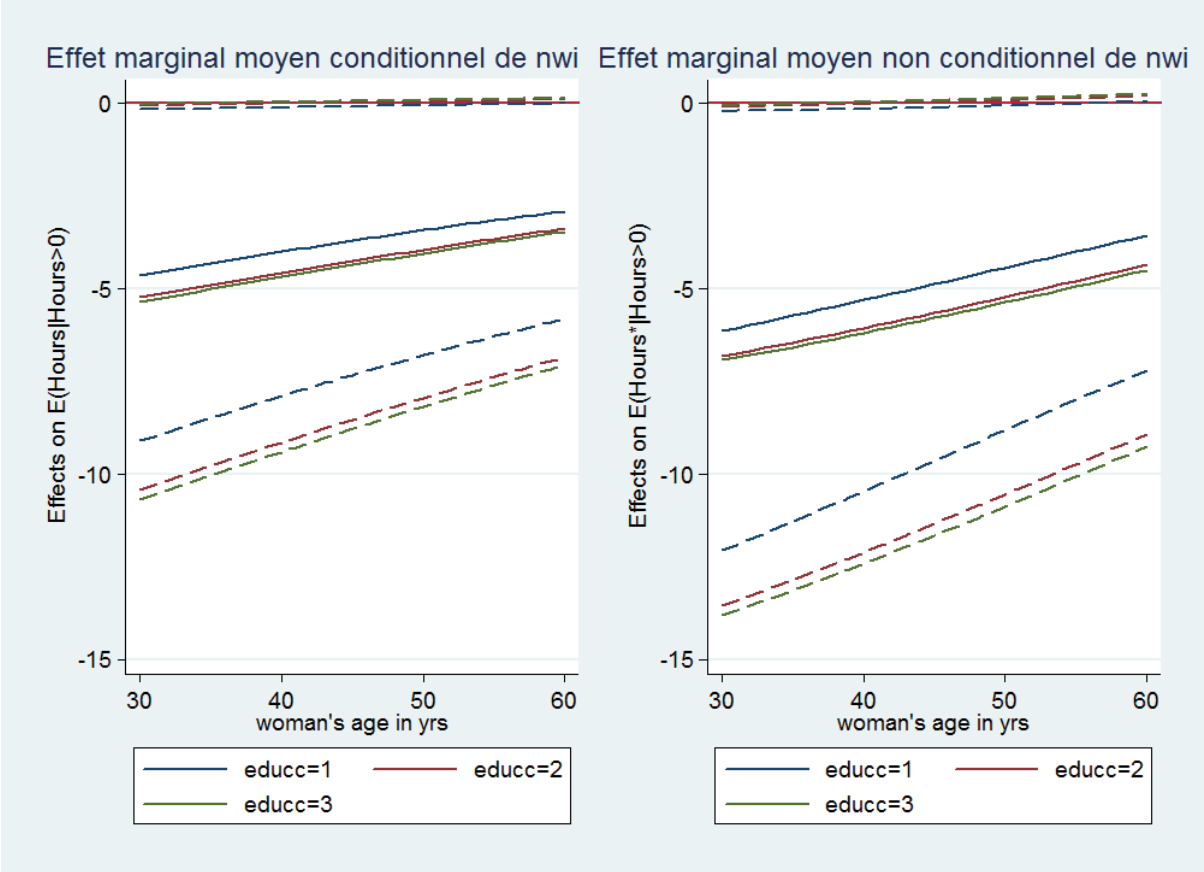
Le graphique est ensuite sauvegardé avec la commande `graph save`.

```
graph save marginstobit1.gph, replace
(file marginstobit1.gph saved)
```

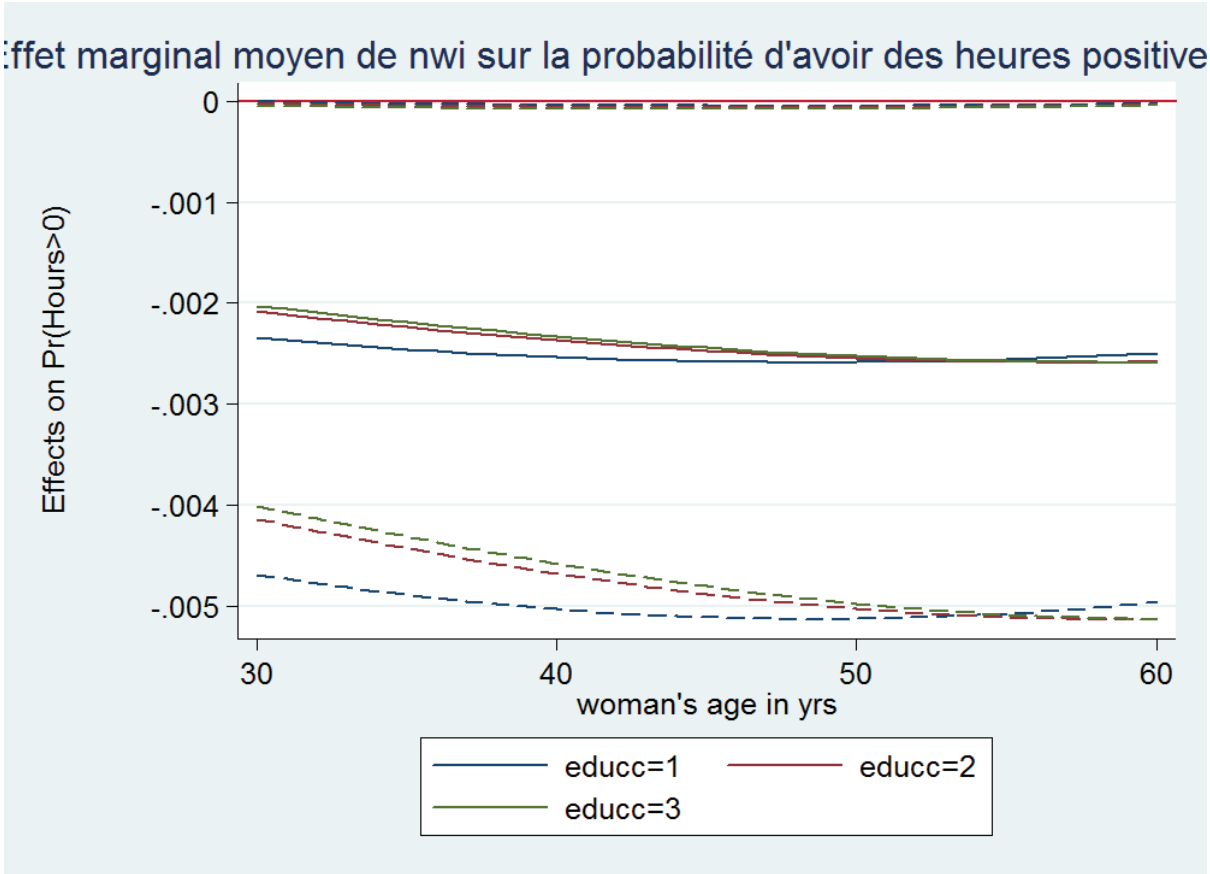
Nous recommençons la même opération pour obtenir les effets marginaux sur la variable indépendante sachant qu'elle est non censurée.

```
margins i.educc, dydx(nwi) at(age=(30(1)60)) predict(ystar(0,.))

Average marginal effects                        Number of obs   =        753
Model VCE    : OIM

Expression   : E(hours*|hours>0), predict(ystar(0,.))
dy/dx w.r.t. : nwi

1._at        : age             =          30

2._at        : age             =          31

                              .
                              .
                              .

30._at       : age             =          59

31._at       : age             =          60

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
nwi          |
   _at#educc |
        1 1  |  -6.126344   3.025933    -2.02   0.043    -12.05706   -.1956237
        1 2  |  -6.809407   3.434719    -1.98   0.047    -13.54133   -.0774817
        1 3  |  -6.923518    3.51326    -1.97   0.049    -13.80938   -.0376547
        2 1  |  -6.046667   2.987675    -2.02   0.043     -11.9024    -.190932
        2 2  |  -6.738568   3.401795    -1.98   0.048    -13.40596   -.0711712
        2 3  |  -6.854525   3.481593    -1.97   0.049    -13.67832   -.0307279

                                                  .
                                                  .
                                                  .

       30 1  |  -3.652488   1.887584    -1.94   0.053    -7.352084    .0471076
       30 2  |  -4.439287   2.366884    -1.88   0.061    -9.078294     .199721
       30 3  |  -4.583041   2.466485    -1.86   0.063    -9.417262    .2511808
       31 1  |  -3.567998   1.850724    -1.93   0.054    -7.195351    .0593547
       31 2  |  -4.351983   2.328965    -1.87   0.062    -8.916671    .2127052
       31 3  |  -4.495655   2.428588    -1.85   0.064    -9.255601    .2642902
------------------------------------------------------------------------------

marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) title(Effet marginal moyen non
conditionnel de nwi) yline(0) ylabel(, angle(0))

  Variables that uniquely identify margins: age educc
```

```
graph save marginstobit2.gph, replace
(file marginstobit2.gph saved)

margins i.educc, dydx(nwi) at(age=(30(1)60)) predict(pr(0,.))

Average marginal effects                        Number of obs   =        753
Model VCE    : OIM

Expression   : Pr(hours>0), predict(pr(0,.))
dy/dx w.r.t. : nwi

1._at        : age             =          30

2._at        : age             =          31

                              .
                              .
                              .

30._at       : age             =          59

31._at       : age             =          60

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
nwi          |
   _at#educc |
        1 1  |  -.0023468   .0011957    -1.96   0.050    -.0046903   -3.32e-06
        1 2  |   -.002081   .0010495    -1.98   0.047     -.004138    -.000024
        1 3  |  -.0020258   .0010111    -2.00   0.045    -.0040074   -.0000442
        2 1  |   -.002371   .0012058    -1.97   0.049    -.0047344   -7.74e-06
        2 2  |  -.0021136   .0010638    -1.99   0.047    -.0041986   -.0000287
        2 3  |  -.0020596    .001026    -2.01   0.045    -.0040705   -.0000486

                                                  .
                                                  .
                                                  .

       30 1  |  -.0025091   .0012658    -1.98   0.047      -.00499   -.0000283
       30 2  |   -.002586   .0012992    -1.99   0.047    -.0051324   -.0000397
       30 3  |  -.0025873   .0012967    -2.00   0.046    -.0051287   -.0000458
       31 1  |  -.0024937   .0012596    -1.98   0.048    -.0049625    -.000025
       31 2  |  -.0025834   .0012998    -1.99   0.047     -.005131   -.0000357
       31 3  |   -.002587   .0012989    -1.99   0.046    -.0051327   -.0000413
------------------------------------------------------------------------------

marginsplot, recast(line) recastci(rline) ciopts(lpattern(dash)) title(Effet marginal moyen de
nwi sur la probabilité d'avoir un revenu) yline(0) ylabel(, angle(0))

  Variables that uniquely identify margins: age educc

graph combine marginstobit1.gph marginstobit2.gph, ycommon
```



Nous pouvons remarquer que l'effet marginal moyen conditionnel et non conditionnel est à la limite d'être significatif à un niveau de 5 % pour les différentes valeurs des calculs. Ils sont par contre significatifs si l'on tolère une erreur de première espèce de 10 %. Les deux effets marginaux moyens de la variable nwi sur la variable dépendante, qu'elle soit censurée ou non ont un impact de moins en moins négatifs quand l'âge de l'individu augmente. Nous voyons qu'il n'existe pas de grande différence entre les deux dernières modalités de la variable educc sur ce graphique.



L'effet marginal sur les probabilités que la variable dépendante soit censurée ou non est significatif pour tous les calculs sauf pour le premier à un niveau de 5 %. Nous pouvons aussi remarquer que les effets sont petits, soit moins de 1 % par variation d'une unité de la variable nwi.

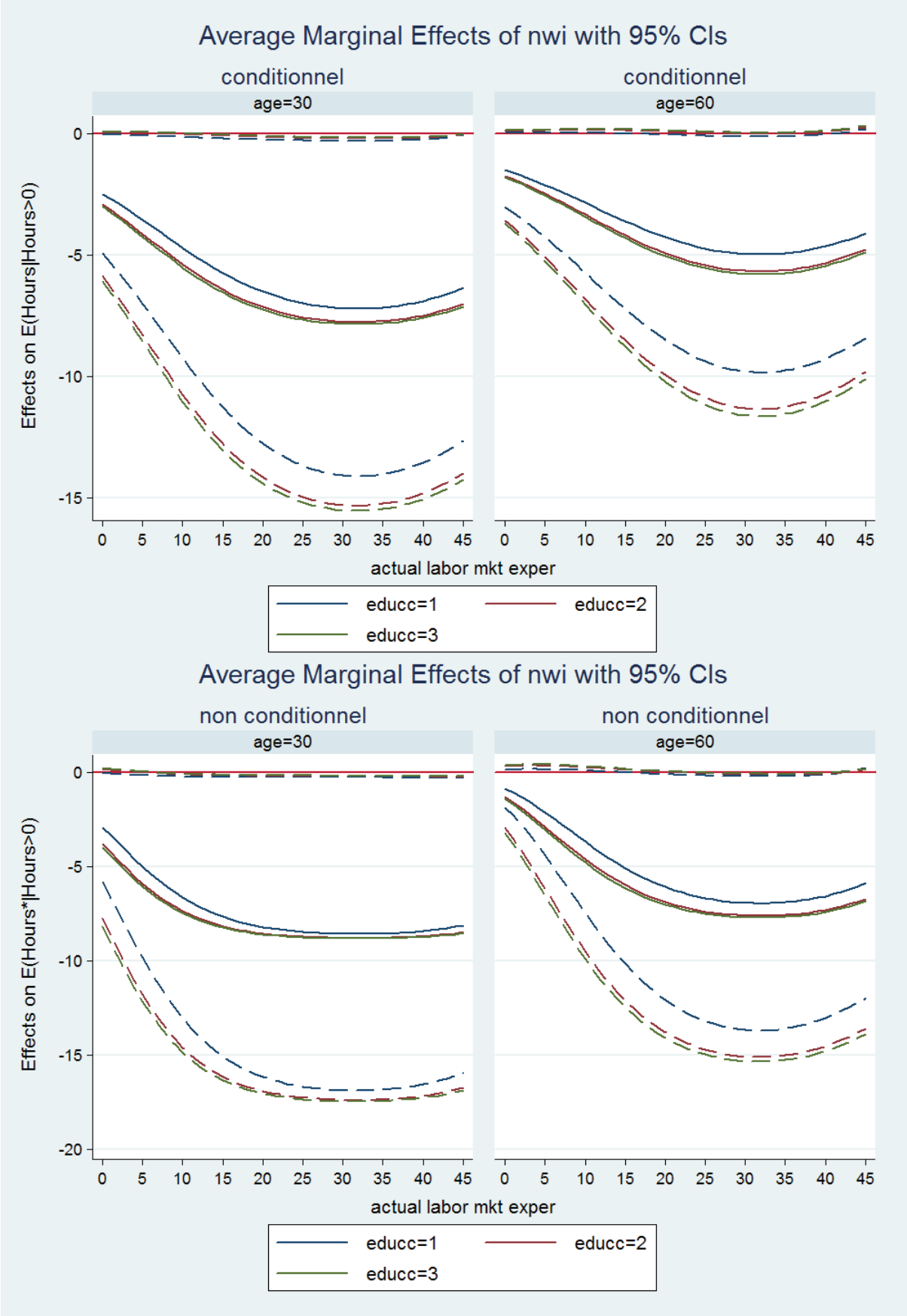
Nous allons ensuite calculer les effets marginaux à différentes valeurs de la variable exper au minimum et au maximum de la variable age en différenciant pour le niveau d'éducation. Nous insérons donc une option `at` et deux variables. Nous allons ensuite représenter graphiquement les effets marginaux toujours avec la commande `marginsplot` et combiner les graphiques des effets marginaux calculés conditionnellement à ce que la variable soit censurée et non conditionnellement à ce que la variable soit censurée.

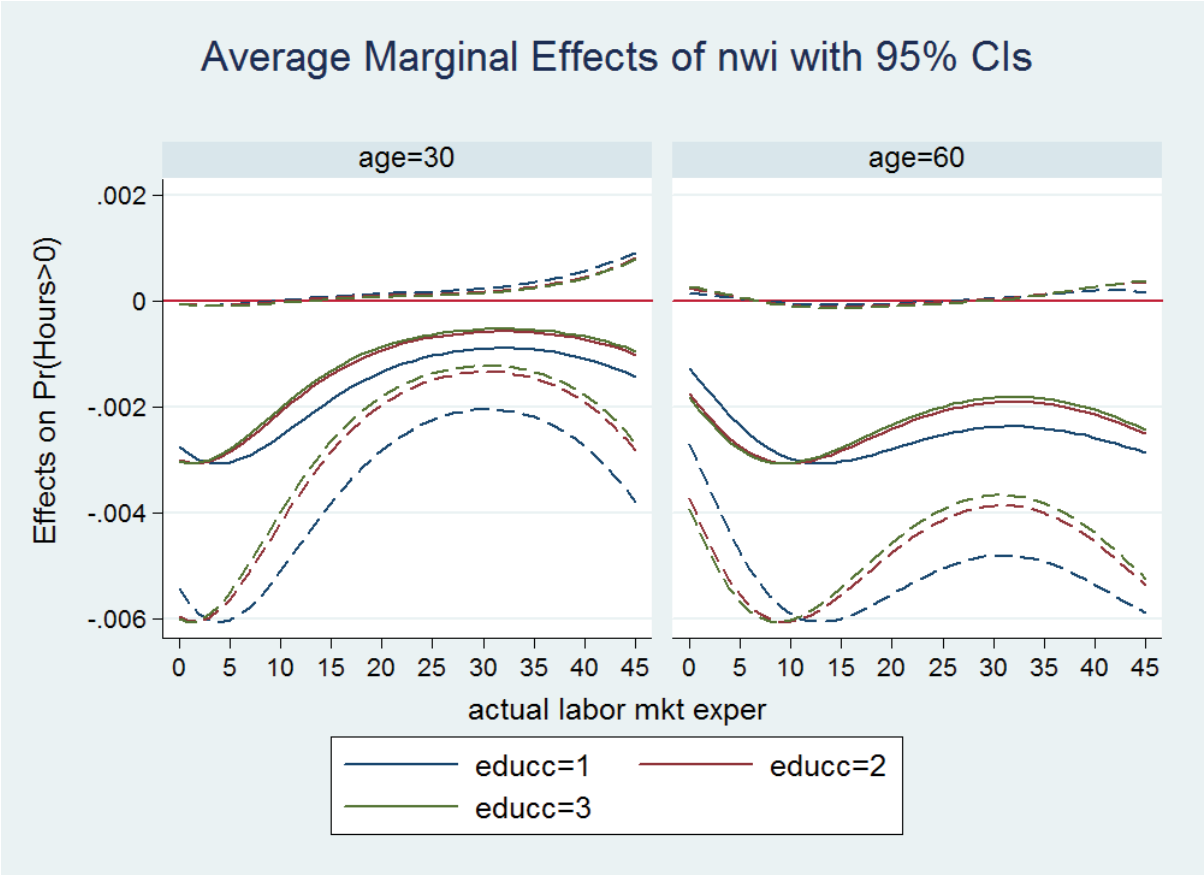
```
/*Effet marginal de nwi quand exper varie et age est à son minimum et son maximum en
différençiant par la variable educc*/
margins i.educc, dydx(nwi) at(age=(30 60)exper=(0(1)45)) predict(e(0,.))

Average marginal effects                        Number of obs   =        753
Model VCE    : OIM

Expression   : E(hours|hours>0), predict(e(0,.))
dy/dx w.r.t. : nwi
```

```
1._at        : exper           =           0
               age             =          30

2._at        : exper           =           0
               age             =          60

3._at        : exper           =           1
               age             =          30

4._at        : exper           =           1
               age             =          60

                              .
                              .
                              .

89._at       : exper           =          44
               age             =          30

90._at       : exper           =          44
               age             =          60

91._at       : exper           =          45
               age             =          30

92._at       : exper           =          45
               age             =          60


------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
nwi          |
   _at#educc |
        1 1  |  -2.485507   1.250644    -1.99   0.047    -4.936723   -.0342905
        1 2  |  -2.911214    1.51147    -1.93   0.054    -5.873641    .0512119
        1 3  |  -2.995374   1.573606    -1.90   0.057    -6.079586    .0888377
        2 1  |   -1.49697   .7834059    -1.91   0.056    -3.032417    .0384775
        2 2  |  -1.739502   .9387804    -1.85   0.064    -3.579478    .1004737
        2 3  |  -1.788083   .9764085    -1.83   0.067    -3.701808    .1256431

                                                  .
                                                  .
                                                  .

       91 1  |  -6.346517   3.219229    -1.97   0.049    -12.65609   -.0369451
       91 2  |  -7.004889   3.562399    -1.97   0.049    -13.98706    -.022716
       91 3  |  -7.117043   3.637343    -1.96   0.050     -14.2461    .0120188
       92 1  |  -4.123212   2.193151    -1.88   0.060    -8.421708    .1752853
       92 2  |  -4.764838   2.568664    -1.85   0.064    -9.799327    .2696507
       92 3  |  -4.885898   2.660344    -1.84   0.066    -10.10008    .3282806
------------------------------------------------------------------------------

marginsplot, by(age) recast(line) recastci(rline) ciopts(lpattern(dash)) title(conditionnel)
yline(0) ylabel(, angle(0)) xlabel(0(5)45) xdimension(exper)

  Variables that uniquely identify margins: age exper educc

graph save marginstobit4.gph, replace
(file marginstobit4.gph saved)

margins i.educc, dydx(nwi) at(age=(30 60)exper=(0(1)45)) predict(ystar(0,.))

Average marginal effects                        Number of obs   =        753
Model VCE    : OIM

Expression   : E(hours*|hours>0), predict(ystar(0,.))
dy/dx w.r.t. : nwi
```

```
1._at        : exper           =           0
               age             =          30

2._at        : exper           =           0
               age             =          60

3._at        : exper           =           1
               age             =          30

4._at        : exper           =           1
               age             =          60

                              .
                              .
                              .

91._at       : exper           =          45
               age             =          30

92._at       : exper           =          45
               age             =          60


------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
nwi          |
   _at#educc |
        1 1  |  -2.902181   1.483292    -1.96   0.050     -5.80938    .0050175
        1 2  |   -3.78866   2.019717    -1.88   0.061    -7.747232    .1699117
        1 3  |  -3.955737    2.13957    -1.85   0.064    -8.149218    .2377431
        2 1  |  -.8413331   .5157743    -1.63   0.103    -1.852232     .169566
        2 2  |  -1.303055   .8363889    -1.56   0.119    -2.942347     .336237
        2 3  |  -1.401928   .9176583    -1.53   0.127    -3.200506    .3966488
                                              .
                                              .
                                              .

       91 1  |  -8.091477    3.99935    -2.02   0.043    -15.93006    -.252895
       91 2  |  -8.465247   4.214086    -2.01   0.045     -16.7247   -.2057914
       91 3  |  -8.519292   4.252528    -2.00   0.045    -16.85409    -.184489
       92 1  |  -5.871548   3.116575    -1.88   0.060    -11.97992    .2368268
       92 2  |  -6.697964   3.509513    -1.91   0.056    -13.57648    .1805546
       92 3  |  -6.834756   3.597056    -1.90   0.057    -13.88486    .2153444
------------------------------------------------------------------------------

marginsplot, by(age) recast(line) recastci(rline) ciopts(lpattern(dash)) title(non conditionnel)
yline(0) ylabel(, angle(0)) xlabel(0(5)45) xdimension(exper)

  Variables that uniquely identify margins: age exper educc

graph save marginstobit5.gph, replace
(file marginstobit5.gph saved)

margins i.educc, dydx(nwi) at(age=(30 60)exper=(0(1)45)) predict(pr(0,.))

Average marginal effects                        Number of obs   =        753
Model VCE    : OIM

Expression   : Pr(hours>0), predict(pr(0,.))
dy/dx w.r.t. : nwi

1._at        : exper           =           0
               age             =          30

2._at        : exper           =           0
               age             =          60

3._at        : exper           =           1
               age             =          30
```

```
4._at        : exper           =           1
               age             =          60

                              .
                              .
                              .

91._at       : exper           =          45
               age             =          30

92._at       : exper           =          45
               age             =          60

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
nwi          |
   _at#educc |
        1 1  |  -.0027409    .001367    -2.01   0.045    -.0054201   -.0000617
        1 2  |  -.0029931   .0015048    -1.99   0.047    -.0059425   -.0000437
        1 3  |  -.0030193   .0015192    -1.99   0.047    -.0059968   -.0000417
        2 1  |  -.0012727   .0007258    -1.75   0.080    -.0026953    .0001499
        2 2  |  -.0017359   .0010095    -1.72   0.086    -.0037145    .0002428
        2 3  |  -.0018228   .0010712    -1.70   0.089    -.0039223    .0002767
                                                  .
                                                  .
                                                  .

       91 1  |  -.0014339   .0011957    -1.20   0.230    -.0037775    .0009097
       91 2  |  -.0010073   .0009274    -1.09   0.277     -.002825    .0008104
       91 3  |  -.0009383   .0008823    -1.06   0.288    -.0026676    .0007909
       92 1  |  -.0028503   .0015407    -1.85   0.064    -.0058701    .0001695
       92 2  |  -.0024995   .0014563    -1.72   0.086    -.0053539    .0003549
       92 3  |  -.0024234   .0014326    -1.69   0.091    -.0052313    .0003845
------------------------------------------------------------------------------

marginsplot, by(age) recast(line) recastci(rline) ciopts(lpattern(dash)) title(" ") yline(0)
ylabel(, angle(0)) xlabel(0(5)45) xdimension(exper)

  Variables that uniquely identify margins: age exper educc

graph export marginstobit6.tif, replace width(1000)
(file marginstobit6.tif written in TIFF format)

graph combine marginstobit4.gph marginstobit5.gph, ycommon ysize(8) imargin(zero) iscale(0.6)
rows(2)
```


Average Marginal Effects of nwi with 95% CIs
conditionnel
age=30
conditionnel
age=60
Effects on E(Hours|Hours>0)
0
-5
-10
-15
0 5 10 15 20 25 30 35 40 45
actual labor mkt exper
educc=1
educc=2
educc=3
Average Marginal Effects of nwi with 95% CIs
non conditionnel
age=30
non conditionnel
age=60
Effects on E(Hours*|Hours>0)
0
-5
-10
-15
-20
0 5 10 15 20 25 30 35 40 45
actual labor mkt exper
educc=1
educc=2
educc=3



Dans le cas des effets marginaux sur les probabilités, nous remarquons que ceux-ci ont un impact très faible. De plus, ils sont non significatifs pour la plupart des calculs.

Nous pouvons par la suite calculer les prédictions du modèle à des valeurs différentes de plusieurs variables. Nous utiliserons donc l'option `at` pour les variables exper et nwi en différenciant pour la variable d'éducation.

```
/*Effets globaux sur les prédictions conditionnelles des variables educc et exper selon le
salaire*/
. margins 1.educc, at(exper=(0(1)45)nwi=(0(20)100)) predict(e(0,.))

Predictive margins                                  Number of obs   =        753
Model VCE    : OIM

Expression   : E(hours|hours>0), predict(e(0,.))

1._at        : nwi             =           0
               exper           =           0

2._at        : nwi             =           0
               exper           =           1

                                .
                                .
                                .
```

```
45._at       : nwi             =           0
               exper           =          44

46._at       : nwi             =           0
               exper           =          45

47._at       : nwi             =          20
               exper           =           0

48._at       : nwi             =          20
               exper           =           1

49._at       : nwi             =          20
               exper           =           2

                              .
                              .
                              .

275._at      : nwi             =         100
               exper           =          44

276._at      : nwi             =         100
               exper           =          45

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _at#educc |
        1 1  |   708.3869   44.57945    15.89   0.000     621.0128     795.761
        2 1  |   746.2666   44.82289    16.65   0.000     658.4154    834.1178

                                                  .
                                                  .
                                                  .

      275 1  |   1096.623   291.3498     3.76   0.000     525.5879    1667.658
      276 1  |   1070.674   294.2901     3.64   0.000     493.8758    1647.472
------------------------------------------------------------------------------
```

Nous demandons ensuite une représentation graphique des prédictions avec la commande `marginsplot`. Nous ajoutons l'option `recast(line)` pour représenter les effets marginaux par une ligne, `noci` pour ne pas afficher les intervalles de confiance, `legend(cols(3))` pour restreindre la légende à trois colonnes et spécifions la forme et les valeurs de l'axe des y avec l'option `ylabel(500 1000 2000, angle(0))`.

```
marginsplot, recast(line) noci legend(cols(3)) ylabel(500 1000 2000, angle(0)) title(" ")

  Variables that uniquely identify margins: exper nwi
```

Nous enregistrons le graphique avec la commande `graph save`.

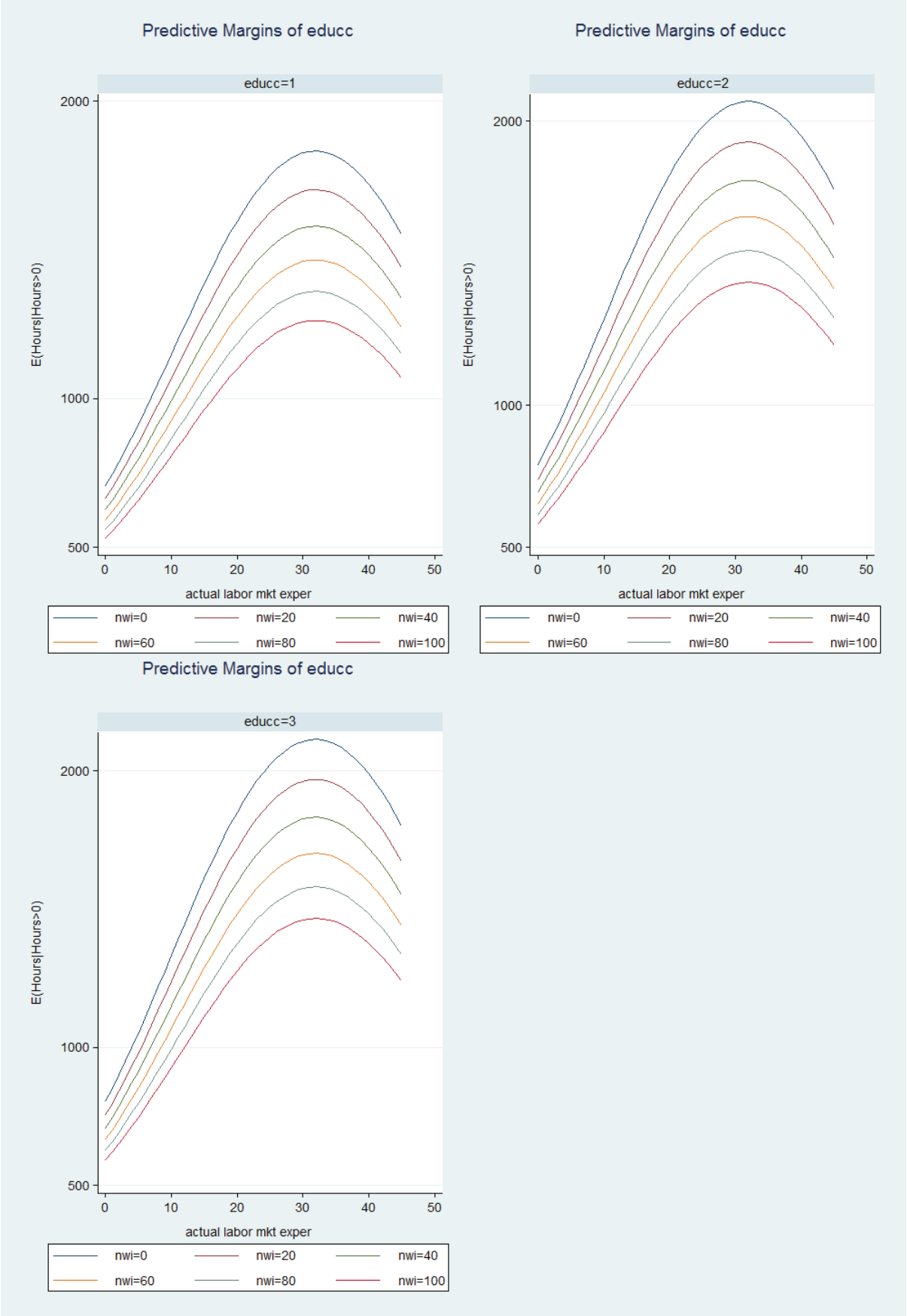
```
graph save marginsglobtobit1, replace
(file marginsglobtobit1.gph saved)
```

Nous recommençons la même opération pour la deuxième et la troisième modalité de la variable educc.

```
/*Effets globaux sur les prédictions conditionnelles des variables educc et exper selon le
salaire*/
```

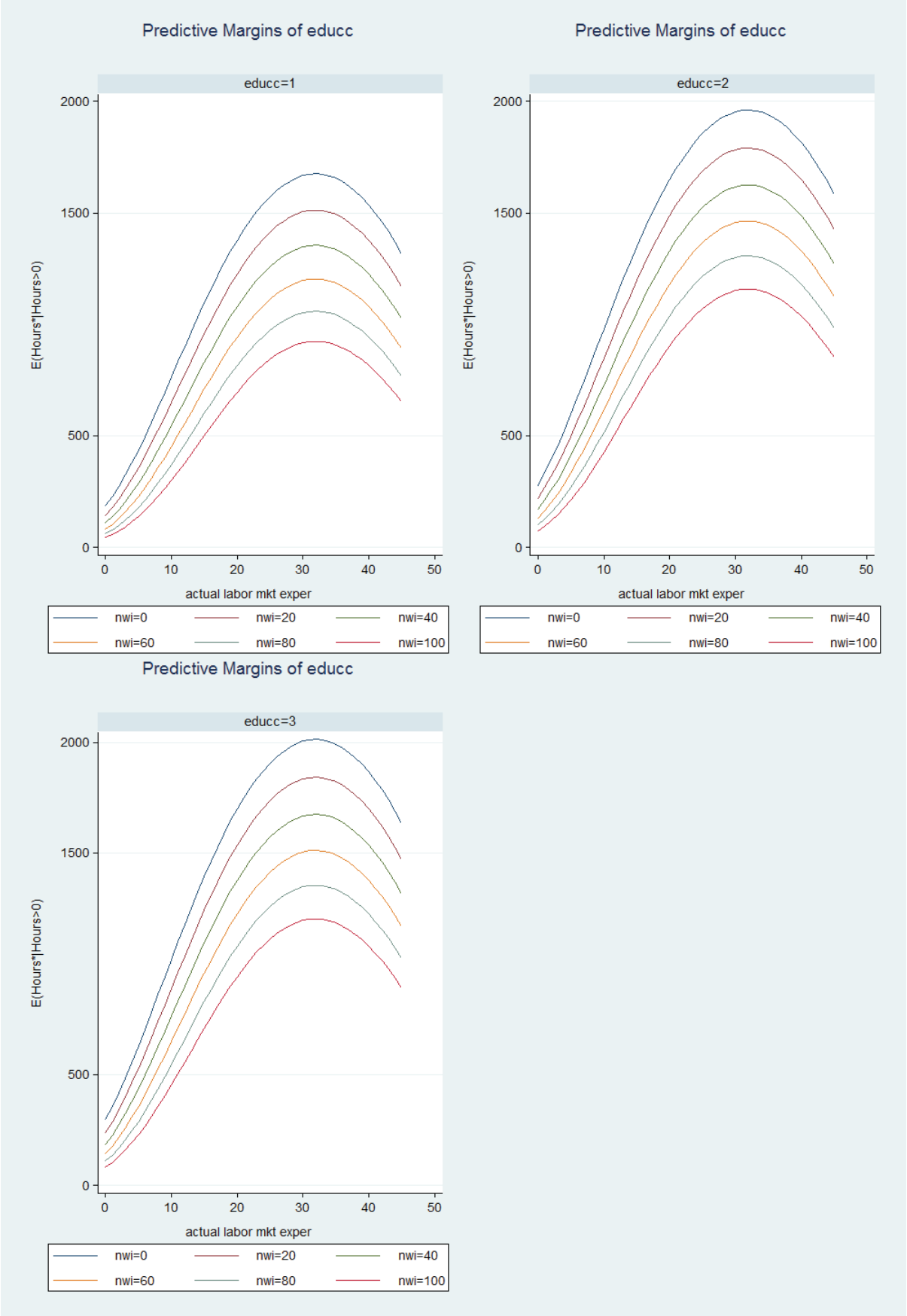
```
margins 2.educc, at(exper=(0(1)45)nwi=(0(20)100)) predict(e(0,.))

Predictive margins                                  Number of obs   =        753
Model VCE    : OIM

Expression   : E(hours|hours>0), predict(e(0,.))

1._at        : nwi             =           0
               exper           =           0

2._at        : nwi             =           0
               exper           =           1

                              .
                              .
                              .

275._at      : nwi             =         100
               exper           =          44

276._at      : nwi             =         100
               exper           =          45

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _at#educc |
        1 2  |   789.6907   62.40776    12.65   0.000     667.3737    912.0077
        2 2  |    834.025   64.43715    12.94   0.000     707.7305    960.3195

                                                  .
                                                  .
                                                  .

      275 2  |   1242.317   329.7987     3.77   0.000     595.9231     1888.71
      276 2  |   1212.298   333.9669     3.63   0.000     557.7349    1866.861
------------------------------------------------------------------------------

marginsplot, recast(line) noci legend(cols(3)) ylabel(500 1000 2000, angle(0)) title(" ")

  Variables that uniquely identify margins: exper nwi

graph save marginsglobtobit2, replace
(file marginsglobtobit2.gph saved)

/*Effets globaux sur les prédictions conditionnelles des variables educc et exper selon le
salaire*/
margins 3.educc, at(exper=(0(1)45)nwi=(0(20)100)) predict(e(0,.))

Predictive margins                                  Number of obs   =        753
Model VCE    : OIM

Expression   : E(hours|hours>0), predict(e(0,.))

1._at        : nwi             =           0
               exper           =           0

2._at        : nwi             =           0
               exper           =           1

                              .
                              .
                              .

275._at      : nwi             =         100
               exper           =          44

276._at      : nwi             =         100
               exper           =          45
```

```
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _at#educc |
        1 3  |   805.7887   68.10559    11.83   0.000     672.3043    939.2732
        2 3  |   851.4019   70.28101    12.11   0.000     713.6537    989.1502

                                                   .
                                                   .
                                                   .

      275 3  |   1270.923   338.7495     3.75   0.000     606.9862     1934.86
      276 3  |   1240.133   343.4232     3.61   0.000      567.036     1913.23
------------------------------------------------------------------------------

marginsplot, recast(line) noci legend(cols(3)) ylabel(500 1000 2000, angle(0)) title(" ")

  Variables that uniquely identify margins: exper nwi

graph save marginsglobtobit3, replace
(file marginsglobtobit3.gph saved)

graph combine marginsglobtobit1.gph marginsglobtobit2.gph marginsglobtobit3.gph, ycommon
iscale(0.4) imargin(left) ysize(8)
```

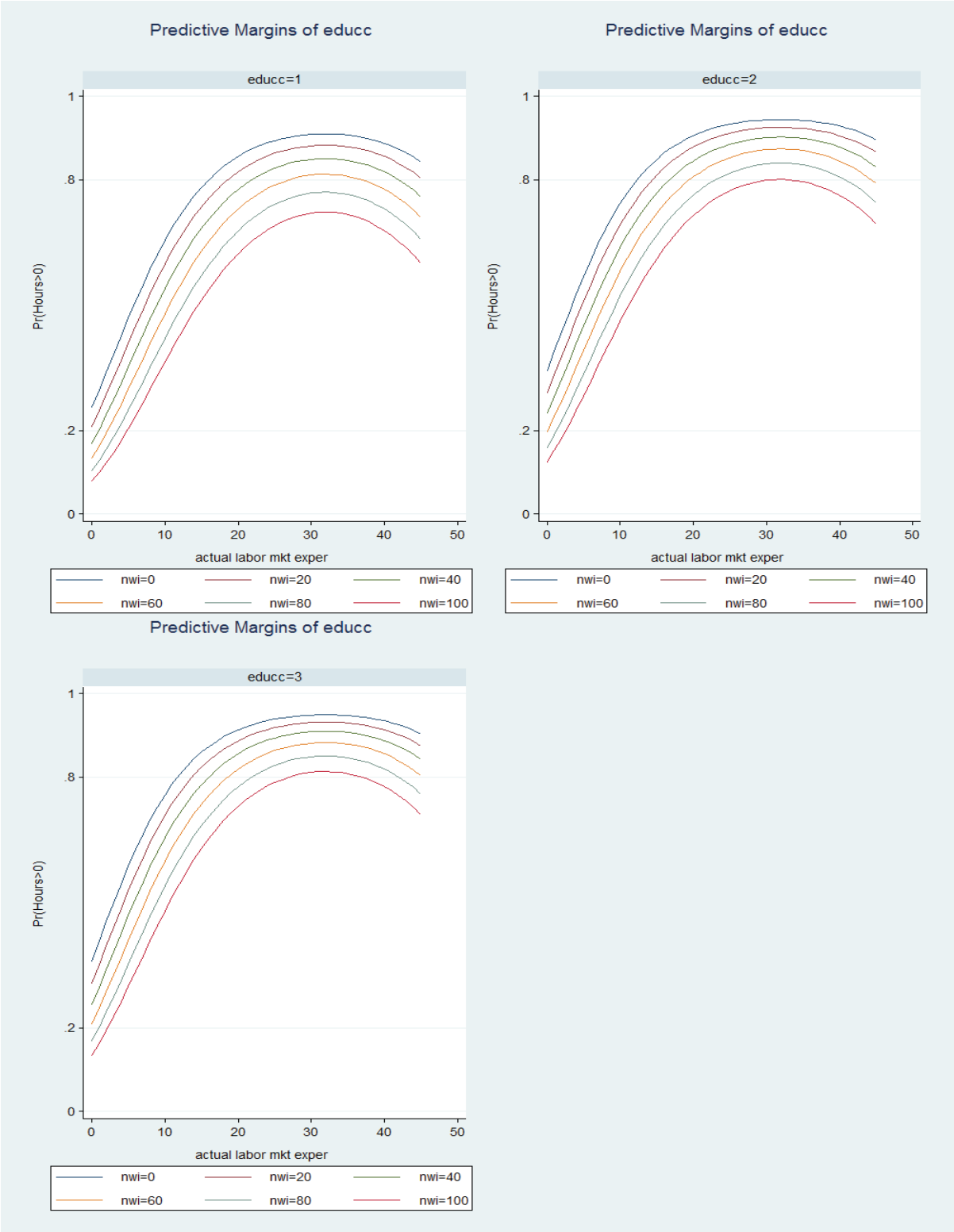

Predictive Margins of educc
educc=1
E(Hours|Hours>0)
2000
1000
500
0
10
20
30
40
50
actual labor mkt exper
nwi=0
nwi=20
nwi=40
nwi=60
nwi=80
nwi=100
Predictive Margins of educc
educc=2
E(Hours|Hours>0)
2000
1000
500
0
10
20
30
40
50
actual labor mkt exper
nwi=0
nwi=20
nwi=40
nwi=60
nwi=80
nwi=100
Predictive Margins of educc
educc=3
E(Hours|Hours>0)
2000
1000
500
0
10
20
30
40
50
actual labor mkt exper
nwi=0
nwi=20
nwi=40
nwi=60
nwi=80
nwi=100

Nous pouvons observer les effets globaux des variables quand plusieurs variables varient. L'important est de trouver une manière d'arranger les graphiques pour que ceux-ci soient facilement interprétables et compréhensibles. C'est une seconde manière de décrire les effets marginaux que nous avons représentés graphiquement. Nous voyons ainsi que l'expérience a un effet concave avec un maximum aux alentours de 30 ans, le revenu a un effet négatif sur le nombre d'heures travaillées et l'éducation un effet positif sur le nombre d'heures travaillées.

Nous pouvons recommencer l'exercice pour la prédiction de la variable dépendante non conditionnellement au fait qu'elle soit supérieure à 0. La même démarche est adoptée que précédemment nous ne présenterons pas les tableaux de résultats et les commandes et seulement le graphique.


Predictive Margins of educc
educc=1
E(Hours*|Hours>0)
2000
1500
500
0
0
10
20
30
40
50
actual labor mkt exper
nwi=0
nwi=20
nwi=40
nwi=60
nwi=80
nwi=100
Predictive Margins of educc
educc=2
E(Hours*|Hours>0)
2000
1500
500
0
0
10
20
30
40
50
actual labor mkt exper
nwi=0
nwi=20
nwi=40
nwi=60
nwi=80
nwi=100
Predictive Margins of educc
educc=3
E(Hours*|Hours>0)
2000
1500
500
0
0
10
20
30
40
50
actual labor mkt exper
nwi=0
nwi=20
nwi=40
nwi=60
nwi=80
nwi=100

Après avoir estimé à nouveau et représenté graphiquement les prédictions concernant la probabilité pour un individu d'être censuré ou non, nous obtenons le graphique ci-dessous.



Nous pouvons observer un effet encore une fois globalement concave en expérience avec un maximum aux alentours de 30 ans. La probabilité pour une femme d'avoir des heures de travail supérieures à 0 dépend négativement du salaire et ne semble pas varier beaucoup selon le niveau d'éducation.

Nous allons maintenant tester si le coefficient des deux modalités de la variable d'éducation est significativement différent. Premièrement, en testant l'hypothèse nulle d'égalité des coefficients de chaque modalité de la variable avec la commande `contrast`.

```
contrast educc

Contrasts of marginal linear prédictions

Margins      : asbalanced

-------------------------------------------------
             |         df           F        P>F
-------------+-----------------------------------
model        |
       educc |          2        4.84     0.0082
             |
    Residual |        747
-------------------------------------------------
```

Il est clair que nous rejetons l'hypothèse nulle d'égalité des coefficients à un niveau de 1 %.

Nous testons donc maintenant si la deuxième et la troisième modalité sont significativement différentes de la première en ajoutant le préfixe `r.` devant la variable.

```
contrast r.educc

Contrasts of marginal linear prédictions

Margins      : asbalanced

-------------------------------------------------
             |         df           F        P>F
-------------+-----------------------------------
model        |
       educc |
   (2 vs 1)  |          1        5.00     0.0256
   (3 vs 1)  |          1        6.77     0.0095
      Joint  |          2        4.84     0.0082
             |
    Residual |        747
-------------------------------------------------

------------------------------------------------------------------
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
model        |
       educc |
   (2 vs 1)  |   307.6622   137.5739      37.58469    577.7397
   (3 vs 1)  |   363.2303   139.6166      89.14282    637.3178
------------------------------------------------------------------
```

Nous pouvons rejeter l'hypothèse d'égalité des coefficients entre la modalité 1 et la modalité 2 à un niveau de 5 % tandis que nous rejetons l'hypothèse d'égalité des coefficients entre la modalité 1 et la modalité 3 à un niveau de 1 %.

Nous allons maintenant tester s'il existe une différence significative entre la modalité 2 et la modalité 3 avec le préfixe `ar`.

```
contrast ar.educc

Contrasts of marginal linear prédictions

Margins      : asbalanced

-------------------------------------------------
             |         df           F        P>F
-------------+-----------------------------------
model        |
       educc |
   (2 vs 1)  |          1        5.00     0.0256
   (3 vs 2)  |          1        0.10     0.7483
      Joint  |          2        4.84     0.0082
             |
    Residual |        747
-------------------------------------------------

------------------------------------------------------------------
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
model        |
       educc |
   (2 vs 1)  |   307.6622   137.5739      37.58469    577.7397
   (3 vs 2)  |   55.56813    173.136     -284.3229    395.4592
------------------------------------------------------------------
```

Nous ne pouvons pas rejeter l'hypothèse nulle d'absence de différence entre le coefficient de la modalité 2 et de la modalité 3.

Nous pouvons aussi effectuer tous ces tests avec la commande `pwcompare` et les options `mcompare(bonferonni)` afin de prendre en compte les comparaisons multiples pour le calcul de la p-valeur , `cimargins effets sort` pour que les effets marginaux et les intervalles de confiance soient classés par ordre de grandeur.

```
pwcompare i.educc, mcompare(bonferroni) cimargins effects sort

Pairwise comparisons of marginal linear prédictions

Margins      : asbalanced

------------------------------------------------------------------
             |                                  Unadjusted
             |     Margin   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
model        |
       educc |
          1  |   640.0859   294.6218      61.70069    1218.471
          2  |   947.7481   313.6135      332.0794    1563.417
          3  |   1003.316   318.7638      377.5367    1629.096
```

```
-------------------------------------------------------------

----------------------------
             |    Number of
             |  Comparisons
-------------+--------------
model        |
       educc |            3
----------------------------

------------------------------------------------------------------------------
             |                            Bonferroni           Bonferroni
             |   Contrast   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
model        |
       educc |
     3 vs 2  |   55.56813    173.136     0.32   1.000    -359.8517     470.988
     2 vs 1  |   307.6622   137.5739     2.24   0.077     -22.4305    637.7549
     3 vs 1  |   363.2303   139.6166     2.60   0.028     28.23655    698.2241
------------------------------------------------------------------------------
```

Avec l'ajustement de la p-valeur , nous voyons maintenant qu'il n'y a pas de différence entre la troisième et la deuxième modalité et pas de différence entre la deuxième et la première modalité en considérant un seuil de 5 %. Il serait possible d'essayer de regrouper la troisième et la deuxième modalité et d'effectuer de nouveau l'analyse.

5.3 Modèle de régression tronquée

5.3.1 Rappel théorique

Il existe des cas ou nous ne pouvons observer une variable pour un certain type d'individu dans de la population. Par exemple, si nous voulons estimer un modèle explicatif du salaire, il nous manquera des données sur celui-ci quand l'individu ne travaille pas. L'échantillon que nous observons est donc en partie choisi par rapport aux valeurs de y_i. Nous pouvons penser au problème de sélection pour un individu étant

$$y_i = y_i^* \text{ si } a_1 \leq y_i^* \leq a_2$$

$$y_i \text{ n'est pas observé si } y_i^* < a_1$$

$$y_i \text{ n'est pas observé si } y_i^* > a_2$$

Ou a_1 et a_2 sont deux constantes vérifiant $a_1 < a_2$. Quand y_i se trouve dans l'intervalle (a_1, a_2), nous pouvons observer y_i et $\boldsymbol{x_i}$. Si y_i est en dehors de l'intervalle, nous n'observons ni l'un ni l'autre. Il existe donc une partie de notre population qui ne peut pas être observée.

Nous devons donc spécifier une distribution conditionnelle de y_i sachant $\boldsymbol{x_i}$. Si nous assumons que la distribution conditionnelle est une Normale($\boldsymbol{x\beta}, \sigma^2$) et qu'il n'y a qu'une troncation par le bas à la valeur 0 nous avons

$$E[y_i|x_i] = E[y_i^*|x_i, y_i^* \geq 0] = x_i'\beta + [\varepsilon_i|\varepsilon_i \geq -x_i'\beta]$$

$$= x_i'\beta + \sigma \frac{\phi\left(-\frac{x_i'\beta}{\sigma}\right)}{1 - \Phi\left(-\frac{x_i'\beta}{\sigma}\right)}$$

Où $\phi(.)$est la densité normale standard et $\Phi(.)$ est la fonction de densité normale cumulée.

Nous pouvons réécrire

$$E[y_i|x_i] = x_i'\beta + \sigma\lambda_i$$

Où

$$\lambda_i = \frac{\phi\left(-\frac{x_i'\beta}{\sigma}\right)}{1 - \Phi\left(-\frac{x_i'\beta}{\sigma}\right)} = \frac{\phi\left(\frac{x_i'\beta}{\sigma}\right)}{\Phi\left(\frac{x_i'\beta}{\sigma}\right)}$$

Nous voyons que ne pas prendre en compte le dernier terme de l'expression d'espérance conditionnelle entrainera un biais de variable omise comme λ_i est surement corrélée avec x_i.

Nous pouvons estimer les paramètres du modèle par maximum de vraisemblance.

Effets marginaux

Les effets marginaux dérivés de la régression tronquée quand nous considérons seulement une troncation par le bas à la valeur 0 sont alors

$$\frac{\partial E[y_i|x_i]}{\partial x_i} = \beta\delta_i$$

Ou δ_i est une composante de la variance de $[\varepsilon_i|\varepsilon_i \geq -x_i'\beta]$

$$Var[\varepsilon_i|\varepsilon_i \geq -x_i'\beta] = \sigma^2\delta_i$$

$$\delta_i = 1 - a_i\lambda_i - \lambda_i^2$$

Sachant que $0 < \delta_i < 1$ l'effet de troncation réduit la variation dans le modèle de régression tronquée et a un réducteur sur les effets marginaux.

Mesures de qualité du modèle

Nous pouvons utiliser le critère d'information d'Akaike (AIC) présenté dans le rappel théorique des modèles logistiques et probit binaires.

5.3.2 La commande truncreg

Description

La commande `truncreg` sert à estimer un modèle de régression ou les valeurs à partir desquelles l'échantillon est tronqué sont fixées.

Syntaxe

```
truncreg [variable dépendante] [variable(s) indépendante(s)] [if] [in] [weight] , ll[(#)] ul[(#)]
[options]
```

où `depvar` est la variable dépendante
`[indepvars]` est la variable indépendante
`[if]` et `[in]` sont des contraintes dans l'échantillon
`[weight]` assigne des poids aux observations
`[, options]` présente des options supplémentaires dans la régression

Menu

Statistics > Linear models and related > Censored régression > Truncated régression

Option

`ll[(#)]` : censure à gauche, limite inférieure, équivalent à la valeur où les observations sont accumulées alors qu'elles devraient prendre une valeur inférieure à celle-ci.

`ul[(#)]` : censure à droite, limite supérieure, équivalent à la valeur où les observations sont accumulées alors qu'elles devraient prendre une valeur supérieure à celle-ci.

Erreur type/Robust

`vce(typevce)` : `typevce` peut être `robust`, `cluster` (variable de cluster).

5.3.3 Les commandes postestimation après la commande tobit

`test [[label ou valeur de la modalité].] [varexo]`: test de Wald avec lequel nous pouvons tester des restrictions linéaires (par exemple $\beta_1 = \beta_2$ ou $\beta_1 = 0$).

`listcoef [, const help]` : donne une liste des coefficients pour le dernier modèle estimé. L'option `const` inclut la constante dans le tableau et l'option `help` rajoute une légende des variables disponibles dans le tableau. Pour plus de renseignements concernant les options, l'utilisateur peut se référer à l'aide fournie avec le module en utilisant la commande `help listcoef`.

`contrast [varexo]` : permet de tester si les moyennes des effets sont égales entre toutes les modalités d'une variable ordinale incluse dans la régression (x).

`contrast r.[varexo]` : permet de tester si les moyennes des effets sont égales entre les différentes modalités d'une variable ordinale quand elles sont comparées à la catégorie de référence.

contrast ar.[varexo] : permet de tester si la moyenne d'un groupe est différente du groupe qui le précède comme la variable est ordinale.

`fitstat` : indices de qualité du modèle (fit ou adéquation).

`lrtest [modèlecomplet] [modelcontraint][, stats]` : cette commande performe un test de ratio de vraisemblance (*likelihood ratio test*) entre un modèle complet et un modèle contraint. Ce test nous renseigne sur la significativité des coefficients non inclus dans le modèle contraint et est similaire au test-F joint qui est calculé avec la commande `test`. Le modèle contraint est donc un sous modèle du modèle complet. Ce test peut être utilisé pour un test de Chow si nous voulons savoir si les coefficients du modèle sont différents entre plusieurs tranches de l'échantillon.

```
margins,
[dydx(varexos)][at(varexo1=(valdébut(incrément)valfin)varexo2=(valdébut(incrément)valfin)]
[predict(e(ll,ul))]:
```

cette commande nous permet de calculer les effets marginaux des variables explicatives sur l'espérance $E(y|a < y < b)$ ($\left(\frac{\partial E(y|a < y < b)}{\partial x}\right)$=marge extensive). L'option `dydx(.)` est nécessaire pour calculer les effets marginaux de variables continues, l'option `at` nous servira pour pouvoir utiliser la commande `marginsplot` afin d'avoir un aperçu graphique de ces effets. Quelques raccourcis comme `dydx(*)` pour sélectionner toutes les variables et `atmeans` pour évaluer

les effets marginaux à la moyenne des variables dans l'échantillon peuvent s'avérer très utiles. Si la variable ne possède qu'une borne inférieure alors la syntaxe est `[predict(e(ll,.))]`. Si la variable ne possède qu'une borne inférieure alors la syntaxe est `[predict(e(.,ul))]`.

```
margins, [dydx(varexos)]
[at(varexo1=(valdébut(incrément)valfin)varexo2=(valdébut(incrément)valfin)]
[predict(ystar(ll,ul))]:
```

cette commande nous permet de calculer les effets marginaux des variables explicatives sur l'espérance $E(y)$ ($\left(\frac{\partial E(y|a<y<b)}{\partial x}\right)$=marge extensive). L'option `dydx(.)` est nécessaire pour calculer les effets marginaux de variables continues, l'option `at` nous servira pour pouvoir utiliser la commande `marginsplot` afin d'avoir un aperçu graphique de ces effets. Quelques raccourcis comme `dydx(*)` pour sélectionner toutes les variables et `atmeans` pour évaluer les effets marginaux à la moyenne des variables dans l'échantillon peuvent s'avérer très utiles. Si la variable ne possède qu'une borne inférieure alors la syntaxe est `[predict(e(ll,.))]`. Si la variable ne possède qu'une borne inférieure alors la syntaxe est `[predict(e(.,ul))]`.

`marginsplot [varexos]` : cette commande doit être utilisée après `margins` pour avoir un aperçu graphique des effets marginaux sur les probabilités.

`pwcompare [varexo], [cimargins]` : permet de comparer l'impact de chaque modalité d'une variable catégorielle entre elles et de tester si elles ont un effet significativement différent. Voir dans l'application pour avoir les différentes options.

Syntaxe pour predict

`predict [type] newvar [if] [in] [, statistic nooffset]` : il existe plusieurs prédictions intéressantes selon la statistique demandée.

Si statistique=xb, c'est la prédiction linéaire qui est prédite (la prédiction par défaut).

Si statistique=pr(a,b), c'est la probabilité $\Pr(a \leq x_j b + u_j \leq b)$ qui est calculée, la probabilité que $y_j|x_j$ soit observé dans l'intervalle (a,b), où a and b sont des entiers non négatifs spécifiés comme des nombres ou des variables. Si lb et ub sont des noms de variables alors

Pr(20,30) calcule $\Pr(20 \leq x_j b + u_j \leq 30)$;

Pr(lb,ub) calcule $\Pr\big(lb \leq x_j b + u_j \leq ub\big)$; et

Pr(20,ub) calcule $\Pr(20 \leq x_j b + u_j \leq ub)$.

a manquant ($a \geq .$) signifie $-\infty$; pr(.,30) calcule $\Pr(-\infty \leq x_j b + u_j \leq 30)$ et pr(lb,30) calcule $\Pr(-\infty \leq x_j b + u_j \leq 30)$ pour les observations pour lesquelles lb est manquante et $\Pr(lb \leq x_j b + u_j \leq 30)$ autrement.

b manquant ($b \geq .$) signifie $+\infty$; pr(.,30) calcule $\Pr(20 \leq x_j b + u_j \leq +\infty)$ et pr(20,ub) calcule $\Pr\Pr(20 \leq x_j b + u_j \leq +\infty)$ pour les observations pour lesquelles ub est manquante et $\Pr(20 \leq x_j b + u_j \leq ub)$ autrement.

Si statistique=residuals, ce sont les résidus qui seront générés pour chaque individu i.

5.3.4 Exemple

Nous utilisons la même base de données que pour l'exemple du modèle Tobit. Nous allons cette fois-ci supposer que l'effet des variables dépendantes doit être calculé en prenant en compte que l'échantillon qui génère les données n'est disponible que pour les femmes étant sur le marché du travail.

Nous commençons par estimer de nouveau notre modèle par régression linéaire avec la méthode des MCO sur la partie de l'échantillon où les heures sont strictement positives.

```
reg hours c.nwi i.educc c.exper##c.exper c.age if hours>0

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  6,   421) =    8.78
       Model |  28604825.8     6  4767470.96           Prob > F      =  0.0000
    Residual |   228706194   421  543245.117           R-squared     =  0.1112
-------------+------------------------------           Adj R-squared =  0.0985
       Total |   257311020   427   602601.92           Root MSE      =  737.05

------------------------------------------------------------------------------
          hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
            nwi |  -.3858206   3.654246    -0.11   0.916     -7.56866    6.797019
                |
          educc |
              2 |   73.96514   102.4462     0.72   0.471    -127.4047    275.3349
              3 |  -132.6496   98.17676    -1.35   0.177    -325.6273    60.32806
                |
          exper |   56.04848   14.59653     3.84   0.000     27.35733    84.73963
                |
c.exper#c.exper |  -.7182604   .4436203    -1.62   0.106    -1.590247    .1537263
                |
            age |  -10.05192   5.494241    -1.83   0.068    -20.85148    .7476446
          _cons |   1182.768   234.1734     5.05   0.000     722.4734    1643.063
------------------------------------------------------------------------------
```

Nous avons donc dans le tableau ci-dessus les effets marginaux moyens des variables qui ne prennent pas en compte l'aspect tronqué de la variable hours. Ces estimations peuvent être biaisées et c'est pourquoi nous estimons la régression avec la commande `truncreg`.

```
truncreg hours c.nwi i.educc c.exper##c.exper c.age, ll(0)
(note: 325 obs. truncated)

Fitting full model:

Iteration 0:   log likelihood = -3408.4624
Iteration 1:   log likelihood = -3398.2356
Iteration 2:   log likelihood = -3398.1402
Iteration 3:   log likelihood = -3398.1401

Truncated régression
Limit:   lower =          0                             Number of obs =    428
         upper =       +inf                             Wald chi2(6)  =  46.63
Log likelihood = -3398.1401                             Prob > chi2   = 0.0000

------------------------------------------------------------------------------
          hours |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
            nwi |  -.9122765   5.329034    -0.17   0.864    -11.35699    9.532437
                |
          educc |
              2 |   106.6618   141.4496     0.75   0.451    -170.5743    383.8979
              3 |  -185.8877   142.3969    -1.31   0.192    -464.9805    93.20509
                |
          exper |   87.64866    21.7497     4.03   0.000     45.02002    130.2773
                |
c.exper#c.exper |  -1.267209   .6272191    -2.02   0.043    -2.496536   -.0378822
                |
            age |  -14.59732   7.935355    -1.84   0.066    -30.15033    .9556892
          _cons |   910.3709   344.2213     2.64   0.008     235.7095    1585.032
----------------+-------------------------------------------------------------
         /sigma |   870.7478   45.66682    19.07   0.000     781.2425    960.2531
------------------------------------------------------------------------------
```

Les coefficients ci-dessus représentent donc les effets marginaux en incluant la variable sigma ce qui corrige pour le biais d'omission de variable. Si nous utilisons la commande `margins`, nous obtiendrons les mêmes résultats.

```
margins, dydx(*)

Average marginal effects                        Number of obs   =        428
Model VCE    : OIM

Expression   : Linear prediction, predict()
dy/dx w.r.t. : nwi 2.educc 3.educc exper age

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         nwi |  -.9122765   5.329034    -0.17   0.864    -11.35699    9.532437
             |
       educc |
          2  |   106.6618   141.4496     0.75   0.451    -170.5743    383.8979
          3  |  -185.8877   142.3969    -1.31   0.192    -464.9805    93.20509
             |
       exper |   54.60648   8.713272     6.27   0.000     37.52878    71.68418
         age |  -14.59732   7.935355    -1.84   0.066    -30.15033    .9556892
```

```
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Nous calculons maintenant les effets marginaux conditionnellement à ce que la variable soit supérieure à 0 avec l'option `predict(e(o,.))`.

```
margins, dydx(*) predict(e(0,.))

Average marginal effects                        Number of obs   =        428
Model VCE    : OIM

Expression   : E(hours|hours>0), predict(e(0,.))
dy/dx w.r.t. : nwi 2.educc 3.educc exper age

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         nwi |  -.6425333   3.753126    -0.17   0.864    -7.998525    6.713458
             |
       educc |
          2  |   77.46659   104.0283     0.74   0.456    -126.4252    281.3584
          3  |   -126.661     94.504    -1.34   0.180    -311.8854    58.56343
             |
       exper |   36.34818   5.128502     7.09   0.000      26.2965    46.39986
         age |  -10.28116   5.560504    -1.85   0.064    -21.17955    .6172229
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

6 Problèmes d'endogénéité

6.1 Introduction

Dans les hypothèses de base du modèle linéaire, il faut que les variables explicatives ne soient pas corrélées avec le terme d'erreur pour que les estimateurs soient consistants. Il existe cependant des cas ou nous savons qu'une ou plusieurs variables violent cette hypothèse. Cette source d'endogénéité peut être causée par :

- Une variable omise contenue dans le terme d'erreur et corrélée avec une variable explicative incluse dans le modèle
- Deux variables peuvent se déterminent simultanément, autrement dit, une variable présumée indépendante du modèle détermine une variable dépendante ET la même variable présumée indépendante est déterminée par la variable dépendante.

$$x = \text{activité sportive}$$

$$y = \text{poids}$$

$$x \leftrightarrows y$$

- Il peut exister des erreurs de mesures qui sont corrélées avec au moins une caractéristique de l'individu.

6.2 Estimation par variables instrumentales

6.2.1 Rappel théorique

Pour commencer, nous avons le modèle suivant

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

et sachant que

$$E[\varepsilon] = 0$$

$$Cov(x_j, \varepsilon) = 0 \qquad j = 1, \dots, k-1$$

$$Cov(x_k, \varepsilon) \neq 0$$

La variable x_k est corrélée avec le terme d'erreur. Nous pouvons voir que les variables explicatives de x_1 à x_{k-1} sont exogènes, mais que x_k est endogène. Nous prendrons comme exemple le cas ou le terme d'erreur contient une variable omise corrélée avec x_k. Ainsi, tous les β sont inconsistants. Il existe une méthode pour corriger le biais occasionné par ce problème s'il existe une variable disponible z_1 non incluse dans le précédent modèle qui satisfait aux deux conditions suivantes

1. La covariance entre z_1 et u dois être égale à 0 (les deux variables doivent être indépendantes).

$$\mathrm{Cov}(z_1, u) = 0$$

2. Il faut aussi que le coefficient θ_1 dans la projection linéaire de x_k sur toutes les variables incluses dans le modèle soit statistiquement différent de 0.

$$x_k = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_{k-1} x_{k-1} + \theta_1 z_1 + r_k$$

Où $E(r_k) = 0$ et r_k n'est pas corrélé avec $x_1, x_2, \dots, x_{k-1}, z_1$

Si les deux conditions sont satisfaites, alors z_1 est une variable instrumentale candidate pour x_k ou plus simplement un instrument de x_k.

Il est à noter qu'il est possible d'inclure plusieurs instruments qui respectent les deux conditions énoncées ci-dessus. Il faut alors que $z_1, z_2, \dots, z_m$ soient orthogonales au terme d'erreur, c'est-à-dire que la covariance entre ces variables et le terme d'erreur soit nulle et que leur estimateur linéaire respectif $\theta_1, \theta_2, \dots, \theta_m$ soit non nuls dans la projection linéaire

$$x_k = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_{k-1} x_{k-1} + \theta_1 z_1 + \theta_2 z_2 + \cdots + \theta_m z_m + r_k$$

Les estimateurs sont le plus souvent obtenus par moindres carrés en deux étapes (*two stages least squares*).

Premièrement, il faut estimer les paramètres de la projection linéaire de x_k sur les autres variables exogènes et les instruments par moindres carrés ordinaires et calculer les valeurs prédites $\hat{x}_k$ pour chaque individu i.

Deuxièmement, il faut effectuer la régression de y sur un nouveau vecteur de variables explicatives composé de $x_1, x_2, \dots, \hat{x}_k$ pour obtenir les estimateurs $\hat{\beta}_{MC2E}$.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k \hat{x}_k + \varepsilon$$

La qualité des estimateurs dépend fortement des deux hypothèses posées sur les instruments, car la taille du biais dépend de la force de la relation entre les instruments et la variable endogène (que nous voulons forte) et la force de la relation entre les instruments et le terme d'erreur ε (que nous voulons égale à 0).

6.2.2 La commande ivregress

Description

La commande `ivregress` permet d'obtenir une estimation des paramètres du modèle $\hat{\beta}_{MC2E}$ directement sans effectuer manuellement les deux étapes pour estimer le modèle. Cette commande supporte aussi l'estimation par maximum de vraisemblance avec information limitée et la méthode généralisée des moments. Pour plus d'informations, l'utilisateur peut se référer au guide Stata.

Syntaxe

```
ivregress estimator [depvar] [indepvars] (endvar=instru) [if] [in] [weight] [, option]
```

où `depvar` est la variable dépendante
`[indepvars]` est la variable indépendante
`(endvar=instru)` est l'équation $variableendogène = \delta_0 + \delta_1 instrument_1 + \cdots + \delta_g instrument_g + \epsilon$
`[if]` et `[in]` sont des contraintes dans l'échantillon
`[weight]` assigne des poids aux observations
`[, options]` présente des options supplémentaires dans la régression

Menu

Statistics > Endogenous covariates > Single-equation instrumental-variables régression

Estimator

`2sls` : l'estimation des paramètres est effectuée par moindres carrés en deux étapes ou *two-stages least squares* (2SLS).

`liml` : l'estimation des paramètres est effectuée par maximum de vraisemblance à information limitée (LIML)

`gmm` : l'estimation des paramètres est effectuée par méthode des moments généralisée, *general method of moment* (GMM).

Option

Erreur type/Robust

`vce(typevce)` : le type vce peut être `robust`, `cluster(variablecluster)`.

Résultats

`first` : rapporte la régression de la première étape des moindres carrés en deux étapes

6.2.3 Les commandes postestimation après la commande ivregress

`estat endogenous` : calcule une statistique de test pour savoir si les variables considérées comme endogènes dans le modèle ne sont pas exogènes. Après une estimation par moindres carrés en deux étapes, les statistiques de Durbin et de Wu-Hausman sont calculées. Si la statistique des tests est significative, alors les variables sont bien endogènes.

`estat firststage` : calcule différentes statistiques qui mesurent la validité des instruments choisis.

`estat overid` : calcule une statistique pour vérifier si l'hypothèse d'indépendance entre l'instrument et le terme d'erreur structurel est vérifiée.

`test [[label ou valeur de la modalité]] [varexo]`: test de Wald avec lequel nous pouvons tester des restrictions linéaires (par exemple $\beta_1 = \beta_2$ ou $\beta_1 = 0$).

`contrast [varexo]` : permet de tester si les moyennes des effets sont égales entre toutes les modalités d'une variable ordinale incluse dans la régression (x).

`contrast r.[varexo]` : permet de tester si les moyennes des effets sont égales entre les différentes modalités d'une variable ordinale quand elles sont comparées à la catégorie de référence.

`contrast ar.[varexo]` : permet de tester si la moyenne d'un groupe est différente du groupe qui le précède comme la variable est ordinale.

`margins, [dydx(varexos)],`
`[at(varexo1=(valdébut(incrément)valfin)varexo2=(valdébut(incrément)valfin)]` :

cette commande nous permet de calculer les effets marginaux des variables spécifiées et les prédictions du modèle. L'option `dydx(.)` est nécessaire pour calculer les effets marginaux de variables continues, l'option `at` nous servira pour pouvoir utiliser la commande `marginsplot` afin d'avoir un aperçu graphique de ces effets. Quelques raccourcis comme `dydx(*)` pour sélectionner toutes les variables et atmeans pour évaluer les effets marginaux à la moyenne des variables dans l'échantillon peuvent s'avérer très utiles.

`marginsplot [varexos]` : cette commande doit être utilisée après `margins` pour avoir un aperçu graphique des effets marginaux sur les probabilités.

`pwcompare [varexo], [cimargins]` : permet de comparer l'impact de chaque modalité d'une variable catégorielle entre elles et de tester si elles ont un effet significativement différent. Voir dans l'application pour avoir les différentes options.

Syntaxe pour predict

`predict [type] newvar [if] [in] [, statistic nooffset]` : il existe plusieurs prédictions intéressantes selon la statistique demandée.

Si statistique=`xb`, c'est la prédiction linéaire qui est prédite (la prédiction par défaut).

Si statistique=`pr(a,b)`, c'est la probabilité $\Pr(a \leq x_j b + u_j \leq b)$ qui est calculée, la probabilité que $y_j|x_j$ soit observé dans l'intervalle (a, b), où a and b sont des entiers non négatifs spécifiés comme des nombres ou des variables. Si lb et ub sont des noms de variables alors

Pr(20,30) calcule $\Pr(20 \leq x_j b + u_j \leq 30)$;

Pr(lb,ub) calcule $\Pr\left(lb \leq x_j b + u_j \leq ub\right)$; et

Pr(20,ub) calcule $\Pr(20 \leq x_j b + u_j \leq ub)$.

a manquant ($a \geq .$) signifie $-\infty$; pr(.,30) calcule $\Pr(-\infty \leq x_j b + u_j \leq 30)$ et pr(lb,30) calcule $\Pr(-\infty \leq x_j b + u_j \leq 30)$ pour les observations pour lesquelles lb est manquante et $\Pr(lb \leq x_j b + u_j \leq 30)$ autrement.

b manquant ($b \geq .$) signifie $+\infty$; pr(.,30) calcule $\Pr(20 \leq x_j b + u_j \leq +\infty)$ et pr(20,ub) calcule $\Pr\Pr(20 \leq x_j b + u_j \leq +\infty)$ pour les observations pour lesquelles ub est manquante et $\Pr(20 \leq x_j b + u_j \leq ub)$ autrement.

6.2.4 Exemple

Nous utiliserons la base de données mroz.dta disponible sur le site de Stata qui est utilisée dans Wooldridge (2010) (utilisées dans Mroz, T. A. 1987. The *sensitivity of an empirical model of married women's hours to work economic and statistical assumptions*. Econometrica 55: 765–799.) pour l'exemple de cette partie.

```
use "http://www.Stata.com/data/jwooldridge/eacsap/mroz.dta", clear
```

Description des données

Les données intéressantes pour cette application dans la base de données peuvent être résumées ci-dessous

```
describe lwage educ exper expersq kidsge6n kidslt6n motheduc fatheduc huseduc

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------------------
lwage           float  %9.0g                  log(wage)
educ            byte   %9.0g                  years of schooling
exper           byte   %9.0g                  actual labor mkt exper
expersq         int    %9.0g                  exper^2
kidsge6n        float  %9.0g                   =1 if # kids between 6-18 is between 0 and 2, =2 if
                                              # kids between 6-18 is great
kidslt6n        float  %9.0g                  =1 if # kids under 6 is greater or equal to 1
motheduc        byte   %9.0g                  mother's years of schooling
fatheduc        byte   %9.0g                  father's years of schooling
huseduc         byte   %9.0g                  husband's years of schooling
```

Nous chercherons à expliquer les effets des déterminants des salaires représentés par la variable lwage. Selon la littérature, nous savons que l'expérience et un effet positif décroissant sur le salaire et l'éducation a un effet positif sur le salaire. Le modèle linéaire associé à nos hypothèses peut être réécrit

$$lwage = \beta_0 + \beta_1 exper + \beta_2 expersq + \beta_5 educ + \varepsilon$$

Modélisation

Nous pouvons estimer la régression sur Stata

```
reg lwage c.exper##c.exper c.educ

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  3,   424) =   26.29
       Model |  35.0222967     3  11.6740989           Prob > F      =  0.0000
    Residual |  188.305144   424  .444115906           R-squared     =  0.1568
-------------+------------------------------           Adj R-squared =  0.1509
       Total |  223.327441   427  .523015084           Root MSE      =  .66642

------------------------------------------------------------------------------
          lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
          exper |   .0415665   .0131752     3.15   0.002     .0156697    .0674633
                |
c.exper#c.exper |  -.0008112   .0003932    -2.06   0.040    -.0015841   -.0000382
                |
           educ |   .1074896   .0141465     7.60   0.000     .0796837    .1352956
          _cons |  -.5220406   .1986321    -2.63   0.009    -.9124667   -.1316144
------------------------------------------------------------------------------
```

Malheureusement, il faut soupçonner que la variable d'éducation ($educ$) est endogène. Nous pouvons attribuer cette endogénéité à l'omission d'une variable mesurant le talent. Cette variable peut difficilement être obtenue et nous n'y avons pas accès dans la base de données. Les estimations ne sont donc pas consistantes étant donné une violation d'une des hypothèses de base du modèle linéaire et nous devons trouver un autre moyen d'estimer les paramètres. Nous pouvons estimer ce modèle en utilisant les moindres carrés en deux étapes pour pallier au biais d'endogénéité si nous pouvons trouver une ou plusieurs variables qui répondent aux hypothèses énoncées dans la partie théorique. Dans la base de données, nous pouvons observer trois variables qui semblent être éligibles pour être des instruments de la variable $educ$. L'éducation de la mère, l'éducation du père et l'éducation du mari semblent être corrélées avec l'éducation de l'individu, mais pas corrélées avec son salaire. Nous estimons donc le modèle

$$educ = \delta_0 + \delta_1 exper + \delta_2 expersq + \theta_1 motheduc + \theta_2 fatheduc + \theta_3 huseduc + r$$

Où les instruments de la variable $educ$ sont $motheduc$, $fatheduc$ et $huseduc$. Nous pouvons produire la régression par moindres carrés ordinaires et trouver

```
reg educ exper expersq motheduc fatheduc huseduc if lwage ~=.

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  5,   422) =   63.30
       Model |  955.830608     5  191.166122           Prob > F      =  0.0000
    Residual |  1274.36565   422  3.01982382           R-squared     =  0.4286
-------------+------------------------------           Adj R-squared =  0.4218
       Total |  2230.19626   427  5.22294206           Root MSE      =  1.7378

------------------------------------------------------------------------------
```

```
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0374977   .0343102     1.09   0.275    -.0299424    .1049379
     expersq |  -.0006002   .0010261    -0.58   0.559    -.0026171    .0014167
    motheduc |   .1141532   .0307835     3.71   0.000     .0536452    .1746613
    fatheduc |   .1060801   .0295153     3.59   0.000     .0480648    .1640955
     huseduc |   .3752548   .0296347    12.66   0.000     .3170049    .4335048
       _cons |   5.538311   .4597824    12.05   0.000     4.634562     6.44206
------------------------------------------------------------------------------
```

Les instruments ont des paramètres significativement différents de 0, la p-valeur étant de 0,000 pour la variable *motheduc*, 0,000 pour la variable *fatheduc* et 0,000 pour la variable *huseduc*. Ces variables ont bien une relation linéaire avec l'éducation de l'individu.

Nous pouvons aussi effectuer un test de Wald pour nous assurer qu'au moins un des coefficients des instruments est significativement différent de 0.

```
test motheduc fatheduc huseduc

 ( 1)  motheduc = 0
 ( 2)  fatheduc = 0
 ( 3)  huseduc = 0

       F(  3,   422) =  104.29
            Prob > F =    0.0000
```

Nous trouvons aussi que nous pouvons fortement rejeter l'hypothèse nulle que tous les coefficients relatifs aux instruments sont simultanément égaux à 0.

À partir de ces résultats, il est possible d'essayer d'estimer le modèle par moindres carrés en deux étapes en utilisant la commande `ivregress`. Spécifier `first` nous donne accès aux résultats de la première étape de la régression.

```
ivregress 2sls lwage c.exper##c.exper (educ=fatheduc huseduc motheduc ), first

First-stage régressions
-----------------------

                                                  Number of obs   =        428
                                                  F(   5,    422) =      63.30
                                                  Prob > F        =     0.0000
                                                  R-squared       =     0.4286
                                                  Adj R-squared   =     0.4218
                                                  Root MSE        =     1.7378

------------------------------------------------------------------------------
           educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
          exper |   .0374977   .0343102     1.09   0.275    -.0299424    .1049379
                |
c.exper#c.exper |  -.0006002   .0010261    -0.58   0.559    -.0026171    .0014167
                |
       fatheduc |   .1060801   .0295153     3.59   0.000     .0480648    .1640955
        huseduc |   .3752548   .0296347    12.66   0.000     .3170049    .4335048
       motheduc |   .1141532   .0307835     3.71   0.000     .0536452    .1746613
          _cons |   5.538311   .4597824    12.05   0.000     4.634562     6.44206
------------------------------------------------------------------------------
```

```
Instrumental variables (2SLS) régression               Number of obs =     428
                                                       Wald chi2(3)  =   34.90
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.1495
                                                       Root MSE      =  .66616

------------------------------------------------------------------------------
          lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
           educ |   .0803918    .021672     3.71   0.000     .0379155    .1228681
          exper |   .0430973   .0132027     3.26   0.001     .0172204    .0689742
                |
c.exper#c.exper |  -.0008628   .0003943    -2.19   0.029    -.0016357   -.0000899
                |
          _cons |  -.1868572   .2840591    -0.66   0.511    -.7436029    .3698885
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper c.exper#c.exper fatheduc
               huseduc motheduc
```

Tests des hypothèses de modélisation

Afin de tester l'endogénéité de la variable *educ*, il est possible d'effectuer un test d'Hausman entre les paramètres estimations par une régression par moindres carrés ordinaires et une régression par moindres carrés en deux étapes après avoir sauvegardé ceux-ci.

```
hausman m2 m1

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       m2           m1         Difference          S.E.
-------------+----------------------------------------------------------------
        educ |    .0803918     .1074896       -.0270979         .016418
       exper |    .0430973     .0415665        .0015308        .0008524
c.exper#c.~r |   -.0008628    -.0008112       -.0000516        .0000293
------------------------------------------------------------------------------
                       b = consistent under Ho and Ha; obtained from ivregress
          B = inconsistent under Ha, efficient under Ho; obtained from regress

    Test:  Ho:  difference in coefficients not systematic

                  chi2(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        2.72
                Prob>chi2 =      0.4361
                (V_b-V_B is not positive definite)
```

Nous pouvons observer qu'il existe une différence significative entre les coefficients des deux régressions et que la variable *educ* peut être considérée comme endogène selon ce test.

Il est aussi possible d'effectuer un autre test pour la présence d'endogénéité grâce à la commande `estat endogenous`

```
estat endogenous

  Tests of endogeneity
  Ho: variables are exogenous

  Durbin (score) chi2(1)          =  2.74613  (p = 0.0975)
  Wu-Hausman F(1,423)             =  2.73158  (p = 0.0991)
```

Nous pouvons voir que la variable *educ* est considérée comme endogène par ce test si nous tolérons une erreur de première espèce de 5%.

Nous pouvons tester la corrélation entre les instruments et la variable endogène. La commande `estat firststage` nous donne accès à plusieurs statistiques nous permettant de juger la validité des instruments choisis.

```
estat firststage

First-stage régression summary statistics
  ---------------------------------------------------------------------------
                |            Adjusted      Partial
      Variable |   R-sq.       R-sq.        R-sq.      F(3,422)   Prob > F
  -------------+-------------------------------------------------------------
          educ |  0.4286      0.4218       0.4258       104.294    0.0000
  ---------------------------------------------------------------------------


  Minimum eigenvalue statistic = 104.294

  Critical Values                       # of endogenous regressors:    1
  Ho: Instruments are weak              # of excluded instruments:     3
  ---------------------------------------------------------------------
                                      |    5%     10%     20%     30%
  2SLS relative bias                  |  13.91    9.08    6.46    5.39
  ------------------------------------+--------------------------------
                                      |   10%     15%     20%     25%
  2SLS Size of nominal 5% Wald test   |  22.30   12.83    9.54    7.80
  LIML Size of nominal 5% Wald test   |   6.46    4.36    3.69    3.32
  ---------------------------------------------------------------------
```

Afin d'avoir un premier aperçu de la force de la relation entre les instruments et la variable endogène nous pouvons regarder le « *partial R-squared* » qui est une mesure similaire au R^2 mais qui ne prend en compte que les variations de la variable endogène causées par les instruments.

Stock & Yogo (2005) proposent un test plus précis concernant le biais toléré au niveau des estimateurs. Nous pouvons prendre en note la statistique F fournie par le tableau (F(3,422)=104.294) qui est une mesure de la significativité jointe des paramètres des instruments. La p-valeur nous fournissant que peu d'évidence sur la qualité des instruments, il faut plutôt se tourner vers le tableau nous présentant le biais relatif des estimateurs par moindres carrés en deux étapes par rapport aux estimateurs MCO. Nous voudrions que la statistique F soit supérieure aux valeurs critiques pour un biais relatif des estimateurs indiqués en pourcentage. Dans notre exemple nous pouvons voir que la statistique F est supérieure à la valeur critique pour un biais relatif de 5 % (104.294>13.91) sur la ligne « *2SLS relative bias* » . La deuxième façon de pouvoir prendre en compte la qualité des instruments est de considérer le

seuil auquel nous voulons rejeter l'hypothèse nulle d'un test de Wald à un niveau de 5 %. Encore une fois, il nous faut comparer la statistique F au seuil indiqué au niveau choisi. Ici, dans tous les cas du tableau, nous pouvons conclure que les instruments ont un lien assez fort avec la variable endogène pour être utilisés.

La deuxième condition pour que les instruments soient valides est que ceux-ci ne doivent pas être corrélés avec le terme d'erreur. Quand nous spécifions plusieurs instruments, il est possible avec la commande `estat overid` de tester simultanément si la forme structurelle du modèle est correcte et si les instruments ne sont pas corrélés avec le terme d'erreur. Il faut donc interpréter les résultats de celui-ci avec précaution.

```
estat overid

  Tests of overidentifying restrictions:

  Sargan (score) chi2(2) =  1.11504  (p = 0.5726)
  Basmann chi2(2)        =  1.10228  (p = 0.5763)
```

Dans le cas de notre estimation précédente, il existe ou alors une corrélation faible entre les instruments et le terme d'erreur ou alors la forme structurelle est mal spécifiée.

Tests d'inférence

Il est aussi possible de tester la significativité jointe de plusieurs coefficients avec la commande `test`. Ici, nous testons si les coefficients devant les variables exper et expersq sont simultanément 0.

```
test c.exper  c.exper#c.exper

 ( 1)  exper = 0
 ( 2)  c.exper#c.exper = 0

           chi2(  2) =   19.92
         Prob > chi2 =    0.0000
```

Nous pouvons rejeter l'hypothèse nulle de nullité des coefficients testés.

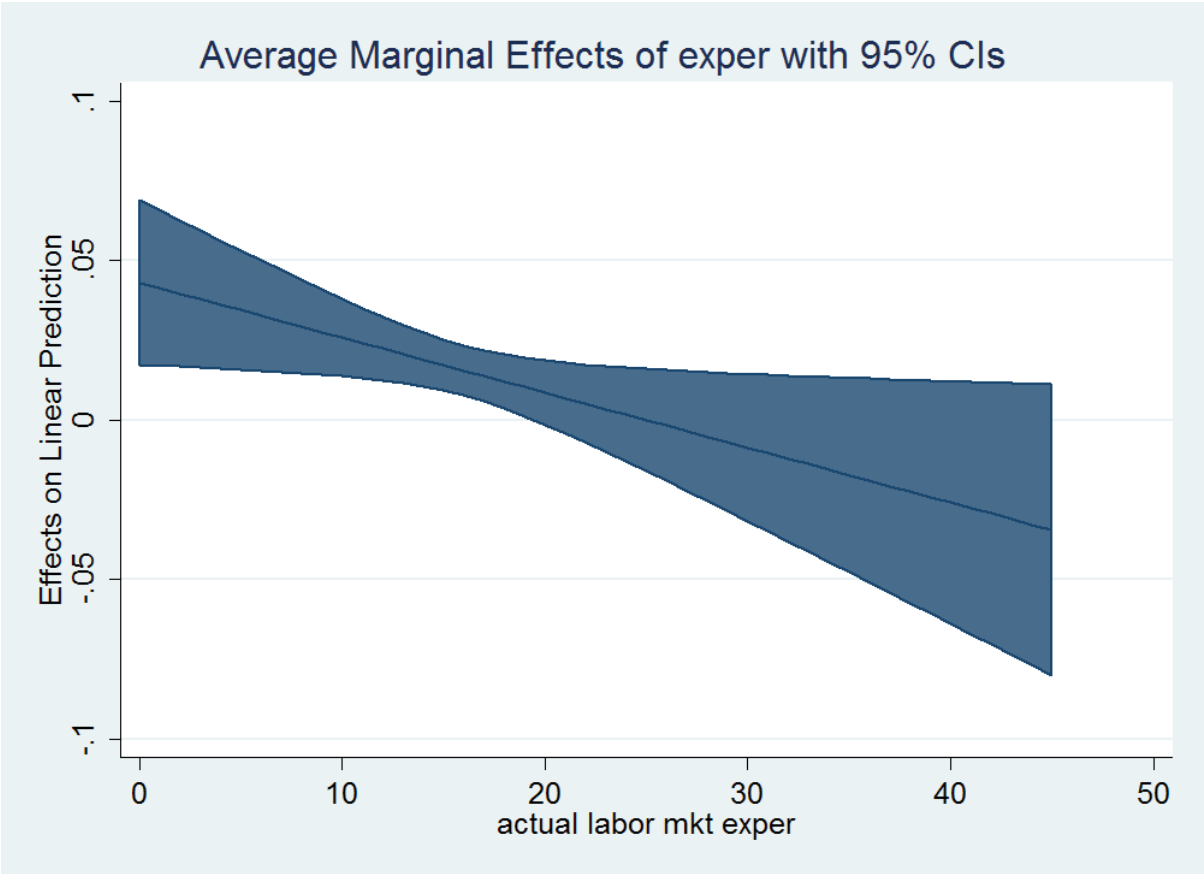
Interprétation des relations entre variable dépendante et variables indépendantes

Si nous voulons explorer les effets d'interaction d'une façon plus précise, la commande `margins` peut être d'une grande utilité. Nous regardons ici comment varie l'effet marginal de la variable exper à différents niveaux de celle-ci. Nous utilisons les options `dydx(exper)` et `at(exper=(0(1)45))` pour faire varier l'âge de 0 à 45 ans par incrément d'une année.

```
margins, dydx(exper) at(exper=(0(1)45))

Average marginal effects                        Number of obs   =        428
Model VCE    : Unadjusted

Expression   : Linear prediction, predict()
dy/dx w.r.t. : exper

1._at        : exper           =           0

2._at        : exper           =           1

                               .
                               .
                               .

45._at       : exper           =          44

46._at       : exper           =          45

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
exper        |
         _at |
          1  |   .0430973   .0132027     3.26   0.001     .0172204    .0689742
          2  |   .0413717   .0124535     3.32   0.001     .0169633    .0657801

                                                  .
                                                  .
                                                  .

         45  |  -.0328288   .0224792    -1.46   0.144    -.0768872    .0112296
         46  |  -.0345544   .0232557    -1.49   0.137    -.0801346    .0110259
------------------------------------------------------------------------------
```

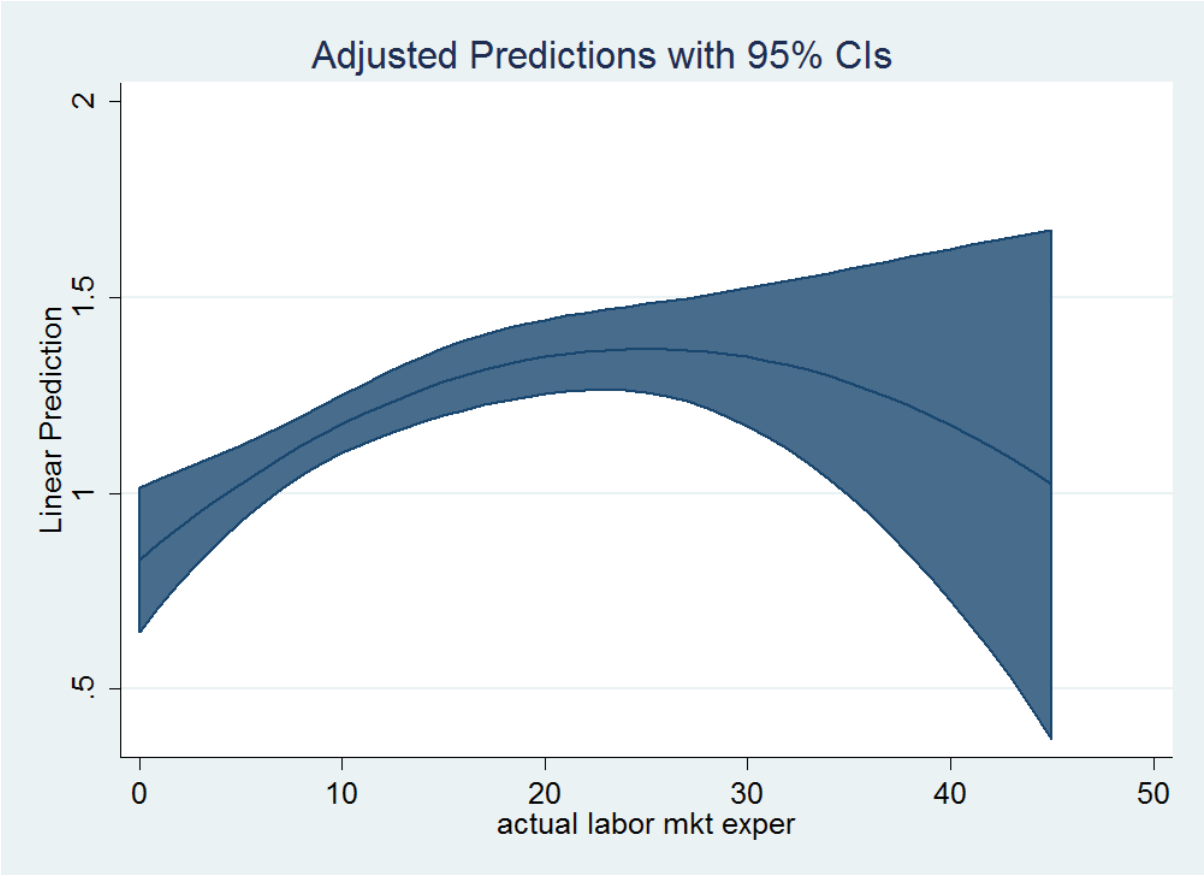
Nous utilisons par la suite la commande `marginsplot` pour représenter graphiquement l'effet. Les intervalles de confiance sont représentés par les aires bleues grâce à l'option `recastci(rarea)`.

```
marginsplot, recast(line) recastci(rarea)

  Variables that uniquely identify margins: exper
```




```
margins, at(exper=(0(1)45)) atmeans

Adjusted prédictions                            Number of obs   =        428
Model VCE    : Unadjusted

Expression   : Linear prediction, predict()

1._at        : educ            =    12.65888 (mean)
               exper           =           0
               0.kidsge6n      =    .3481308 (mean)
               1.kidsge6n      =    .4579439 (mean)
               2.kidsge6n      =    .1939252 (mean)
               0.kidslt6n      =    .8761682 (mean)
               1.kidslt6n      =    .1238318 (mean)

2._at        : educ            =    12.65888 (mean)
               exper           =           1
               0.kidsge6n      =    .3481308 (mean)
               1.kidsge6n      =    .4579439 (mean)
               2.kidsge6n      =    .1939252 (mean)
               0.kidslt6n      =    .8761682 (mean)
               1.kidslt6n      =    .1238318 (mean)

                              .
                              .
                              .

45._at       : educ            =    12.65888 (mean)
               exper           =          44
               0.kidsge6n      =    .3481308 (mean)
               1.kidsge6n      =    .4579439 (mean)
               2.kidsge6n      =    .1939252 (mean)
               0.kidslt6n      =    .8761682 (mean)
               1.kidslt6n      =    .1238318 (mean)
```

```
46._at       : educ            =    12.65888 (mean)
               exper           =          45
               0.kidsge6n      =    .3481308 (mean)
               1.kidsge6n      =    .4579439 (mean)
               2.kidsge6n      =    .1939252 (mean)
               0.kidslt6n      =    .8761682 (mean)
               1.kidslt6n      =    .1238318 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   .8308123   .0941714     8.82   0.000     .6462398    1.015385
          2  |   .8730468   .0829035    10.53   0.000     .7105588    1.035535

                                                  .
                                                  .
                                                  .

         45  |    1.05672   .3090923     3.42   0.001     .4509107     1.66253
         46  |   1.023029   .3312409     3.09   0.002     .3738086    1.672249
------------------------------------------------------------------------------

marginsplot, recast(line) recastci(rarea)

  Variables that uniquely identify margins: exper
```



Nous pouvons clairement visualiser l'effet quadratique de l'expérience avec un intervalle de confiance qui devient très grand pour de nombreuses années d'expérience, *ceteris paribus*.

Tests sur des variables catégorielles

Si nous n'avions accès qu'à une variable catégorielle concernant l'expérience de chaque individu, il nous faudrait tester si les coefficients de chaque modalité ont un effet distinct les uns des autres. Nous estimons à nouveau donc notre modèle à l'aide de la commande `ivregress` avec notre variable catégorielle pour l'expérience sans interaction.

Si nous voulions tester si les différentes modalités de la variable ordinale experc sont significativement différentes, nous pouvons utiliser la commande `contrast` de différentes façons.

```
ivregress 2sls lwage i.experc (educ=fatheduc huseduc motheduc ), first
First-stage régressions
-----------------------

                                                  Number of obs   =        428
                                                  F(   6,    421) =      52.75
                                                  Prob > F        =     0.0000
                                                  R-squared       =     0.4291
                                                  Adj R-squared   =     0.4210
                                                  Root MSE        =     1.7390

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      experc |
          0  |          0  (empty)
          1  |   .3335239   .2830455     1.18   0.239    -.2228344    .8898823
          2  |    .232027   .2649993     0.88   0.382    -.2888595    .7529136
          3  |   .4921562   .2631098     1.87   0.062    -.0250164    1.009329
             |
    fatheduc |   .1012286   .0299022     3.39   0.001     .0424525    .1600048
     huseduc |   .3737594   .0297234    12.57   0.000     .3153346    .4321843
    motheduc |   .1154246   .0308411     3.74   0.000      .054803    .1760463
       _cons |   5.634963   .4329003    13.02   0.000     4.784048    6.485879
------------------------------------------------------------------------------


Instrumental variables (2SLS) régression               Number of obs =     428
                                                       Wald chi2(4)  =   29.95
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.1384
                                                       Root MSE      =  .67052

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0750269    .022271     3.37   0.001     .0313764    .1186773
             |
      experc |
          1  |   .2330221   .1098355     2.12   0.034     .0177484    .4482957
          2  |   .3108953   .1026884     3.03   0.002     .1096298    .5121608
          3  |   .3922503   .1019029     3.85   0.000     .1925243    .5919764
             |
       _cons |  -.0307208   .2815304    -0.11   0.913    -.5825102    .5210687
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   1.experc 2.experc 3.experc fatheduc huseduc motheduc
```

Premièrement, nous pouvons tester si les effets de chaque modalité sont égaux.

```
contrast experc
```

```
Contrasts of marginal linear prédictions

Margins      : asbalanced

-------------------------------------------------
             |         df        chi2     P>chi2
-------------+-----------------------------------
      experc |          3       15.68     0.0013
-------------------------------------------------
```

Nous avons assez d'évidences pour rejeter l'hypothèse que tous les coefficients sont égaux.

Nous pouvons essayer de connaître plus précisément les différences existantes entre les catégories de la variable experc en ajoutant `r.` devant la variable. Ceci nous permet de tester les différences entre une catégorie *j* et la catégorie de base.

```
contrast r.experc

Contrasts of marginal linear prédictions

Margins      : asbalanced

-------------------------------------------------
             |         df        chi2     P>chi2
-------------+-----------------------------------
      experc |
   (1 vs 0)  |          1        4.50     0.0339
   (2 vs 0)  |          1        9.17     0.0025
   (3 vs 0)  |          1       14.82     0.0001
      Joint  |          3       15.68     0.0013
-------------------------------------------------

------------------------------------------------------------------
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
      experc |
   (1 vs 0)  |   .2330221   .1098355      .0177484    .4482957
   (2 vs 0)  |   .3108953   .1026884      .1096298    .5121608
   (3 vs 0)  |   .3922503   .1019029      .1925243    .5919764
------------------------------------------------------------------
```

Nous voyons que par rapport à la catégorie de base qui représente le fait pour un individu de la population d'avoir entre 0 et 4 ans d'expérience, que les autres catégories sont significativement différentes.

Nous regardons maintenant s'il existe une différence significative entre la deuxième catégorie et la première catégorie et une différence significative entre la troisième catégorie et la deuxième catégorie. Nous utilisons le préfixe `ar.` pour comparer le coefficient de la catégorie *j* au coefficient de la catégorie *j-1*.

```
contrast ar.experc

Contrasts of marginal linear prédictions

Margins      : asbalanced
```

```
-----------------------------------------------
             |         df        chi2     P>chi2
-------------+---------------------------------
      experc |
   (1 vs 0)  |          1        4.50     0.0339
   (2 vs 1)  |          1        0.75     0.3879
   (3 vs 2)  |          1        0.97     0.3245
      Joint  |          3       15.68     0.0013
-----------------------------------------------

-------------------------------------------------------------
             |   Contrast   Std. Err.     [95% Conf. Interval]
-------------+-----------------------------------------------
      experc |
   (1 vs 0)  |   .2330221   .1098355      .0177484    .4482957
   (2 vs 1)  |   .0778732   .0901894     -.0988948    .2546413
   (3 vs 2)  |   .0813551   .0825809     -.0805006    .2432107
-------------------------------------------------------------
```

Nous remarquons qu'il n'y a pas de différence significative entre l'estimateur de la catégorie 2 et l'estimateur de la catégorie 1. Il en est de même pour les modalités 2 et 3.

Une autre façon de comparer les coefficients estimations d'une variable catégorielle est d'utiliser la commande `pwcompare`, qui nous permet de comparer par paire des niveaux de variables catégorielles. Nous garderons la même variable que précédemment et nous ajoutons les options `cimargins`, `effect` et `sort` pour que les estimations, les écarts types et les valeurs p soient disponibles en ordre de grandeur.

```
pwcompare i.experc, mcompare(bonferroni) cimargins effects sort

Pairwise comparisons of marginal linear prédictions

Margins      : asbalanced

-------------------------------------------------------------
             |                                Unadjusted
             |     Margin   Std. Err.     [95% Conf. Interval]
-------------+-----------------------------------------------
      experc |
          0  |  -.0307208   .2815304     -.5825102    .5210687
          1  |   .2023013   .3012081     -.3880557    .7926583
          2  |   .2801745   .2880526     -.2843981    .8447472
          3  |   .3615296   .2854893     -.1980192    .9210783
-------------------------------------------------------------

----------------------------
             |    Number of
             |  Comparisons
-------------+--------------
      experc |            6
----------------------------

-----------------------------------------------------------------------------
             |                            Bonferroni           Bonferroni
             |   Contrast   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
      experc |
     2 vs 1  |   .0778732   .0901894     0.86   1.000    -.1600697    .3158162
     3 vs 2  |   .0813551   .0825809     0.99   1.000    -.1365146    .2992248
     3 vs 1  |   .1592283   .0899624     1.77   0.460    -.0781157    .3965722
     1 vs 0  |   .2330221   .1098355     2.12   0.203    -.0567523    .5227964
     2 vs 0  |   .3108953   .1026884     3.03   0.015      .039977    .5818136
```

```
     3 vs 0  |   .3922503   .1019029     3.85   0.001     .1234043    .6610964
------------------------------------------------------------------------------
```

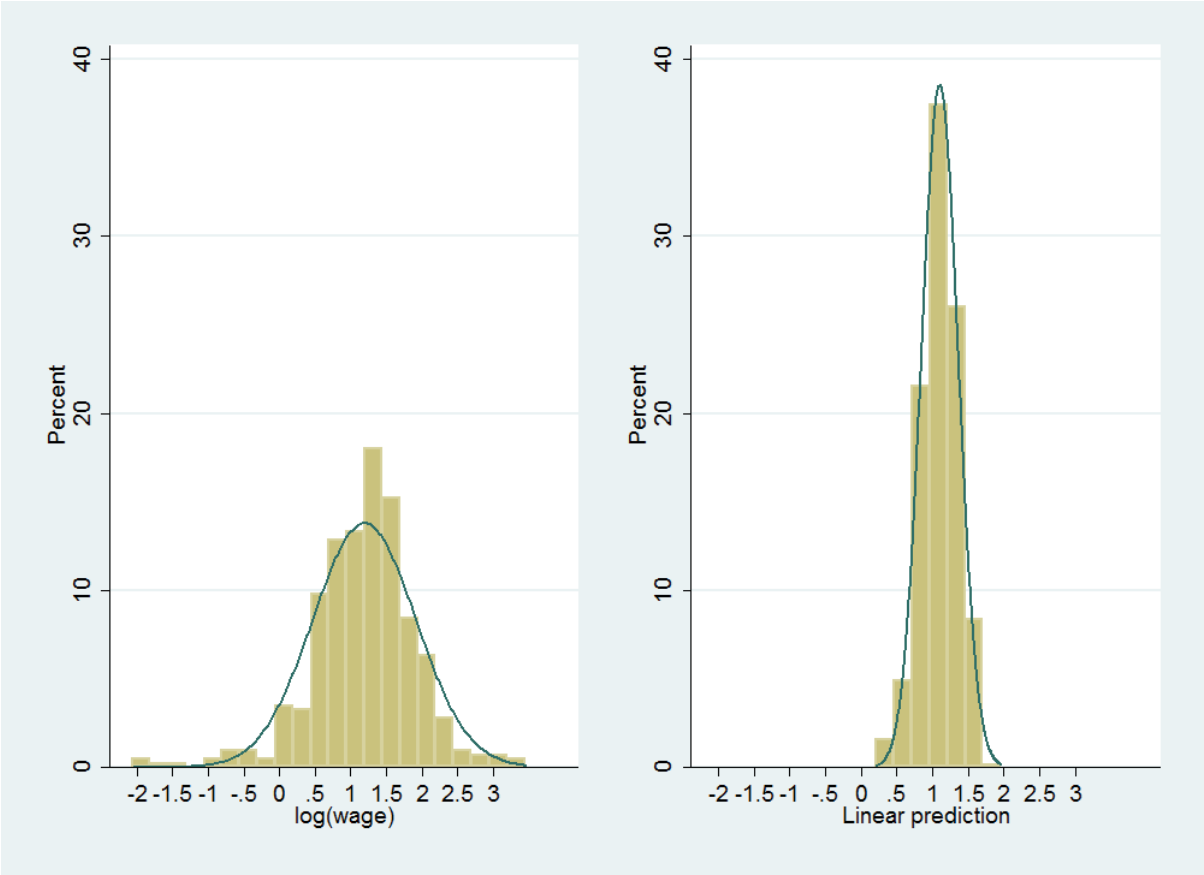
Nous voyons lorsque nous ajustons pour le nombre de comparaisons par la méthode de Bonferroni avec l'option `mcompare(bonferroni)`, nous avons des résultats qui diffèrent pour la comparaison entre la modalité 0 et la modalité 1.

Pour conclure à ce niveau, nous pourrions regrouper des modalités et tester à nouveau la différence des coefficients jusqu'à ce que nous arrivions à une différence significative entre toutes les paires de coefficients comparées.

Prédiction

Nous allons pour finir calculer les valeurs prédites des résidus avec la commande `predict`. Nous allons aussi prédire les résidus et les représenter graphiquement

```
predict yhat
(option xb assumed; fitted values)

histogram lwage, percent normal saving(lwage)
histogram yhat,  percent normal saving(yhat)
graph combine lwage.gph yhat.gph, ycommon
```



Alors que la moyenne semble être assez justement estimée, nous voyons que les prédictions souffrent d'un manque de dispersion. Nous avons pu constater que les intervalles de confiance étaient larges aux valeurs extrêmes. Notre analyse, en moyenne, ne nous apporte donc pas beaucoup d'informations sur l'impact des variables explicatives à leurs valeurs extrêmes.

7 Les Séries Temporelles univariées

7.1 Définition d'une série temporelle

Une séquence de variables aléatoires indexées par le temps est un processus stochastique ou processus temporel. Lorsque nous collectons un ensemble de données dans le temps, nous obtenons un résultat possible du processus stochastique.

Modèle Statique

Nous avons des données sur 2 variables soit y et z à travers le temps

$$Yt = \beta_0 + \beta_1 z_{t1} + u_t, t = 1, 2, 3, \dots, n$$

- Nous modélisons une relation contemporaine entre y et z
- Nous postulons que dans un modèle statique, un changement dans z à temps t a un effet immédiat sur y
- $\Delta y_t = \beta_1 \Delta z_t$ lorsque $\Delta u_t = 0$

Modèle à Retards Échelonnés Finis

Modèle où au moins une variable influence y avec un retard. Exemple de modèle d'ordre 2

$$Y_t = \beta_0 + \theta_0 z_t + \theta_1 z_{t-1} + \theta_2 z_{t-2} + u_t$$

Pour interpréter les coefficients, supposez que « z » = « c » où « c » est une constante pour toutes les périodes. Au temps t, $z \uparrow$ d'une unité,

Modélisation

Étapes à suivre quand nous avons à modéliser une variable temporelle sont les suivantes:

1. Nommer la série temporelle
2. Représenter graphiquement les données dans le temps afin de visualiser les tendances
3. Tester les coupures dans la tendance avec le *Quandt Likelihood Ratio* et le test de Chow
4. Tester la saisonnalité
5. Effectuer un test de racine unitaire pour l'hypothèse de stationnarité
6. Tester l'autocorrélation et la corriger

7. Déterminer l'ordre du modèle autorégressif
8. Tester l'hétéroscédasticité et la corriger
9. Tester le modèle ARCH
10. Effectuer la régression finale

7.2 Créer et nommer la série temporelle

La première étape pour une série temporelle sur Stata est d'annoncer au logiciel que nous avons une série temporelle. Pour ce faire, il faut avoir une variable de temps. Nous verrons deux méthodes pour créer cette variable :

Créer une variable *time* avec la fonction :

```
gen time = _n
```

Ensuite, il nous faudra indiquer que *time* est votre variable temporelle.

```
tsset time
```

L'autre méthode serait de télécharger le module Stata tsmktim (`findit tsmktim` et installer le fichier dm81) créer une variable avec cette fonction :

```
tsmktim time, start(1977m1)
```

Alors Stata créera une variable de temps mensuelle *time* à partir de 1977 au mois 1 jusqu'à la fin de vos données.

Si par exemple vous avez des données trimestrielles qui commencent en 1977 2e trimestre, il faudrait faire la commande :

Pour les besoins de la cause, nous allons créer une série de données

```
clear
set obs 100
obs was 0, now 100
set seed 10101
generate x1var = runiform()
gen t=(_n-64)/15
gen r=_n+45
gen x1var2=(x1var-0.5)*100
gen t2=t-10*t^2
gen x1var3=x1var2+t2+r
```

Ensuite nous utilisons la commande `tsmktim` dans Stata

```
tsmktim time, start(1977q2)
```

```
list time in 1/10

     +--------+
     |   time |
     |--------|
  1. | 1977q2 |
  2. | 1977q3 |
  3. | 1977q4 |
  4. | 1978q1 |
  5. | 1978q2 |
     |--------|
  6. | 1978q3 |
  7. | 1978q4 |
  8. | 1979q1 |
  9. | 1979q2 |
 10. | 1979q3 |
```

7.3 Graphique de la tendance

Une fois que nous avons indiqué à Stata que nous allons travailler avec des données temporelles, il faut faire un graphique de nos données afin de voir s'il y a une tendance linéaire, exponentielle ou quadratique.

$$Yt = \theta_0 + \theta_1 t + e_t, t = 1, 2, \ldots$$

Où e_t est i.i.d. $E(e_t) = 0$, $Var(e_t) = \sigma_t^2$

θ_1 : *ceteris paribus*, θ_1 mesure $\Delta\, y_t$d'une période à l'autre lorsque $\Delta\, e_t = 0$

$$\Delta\, y_t = y_t - y_{t-1} = \theta_1$$

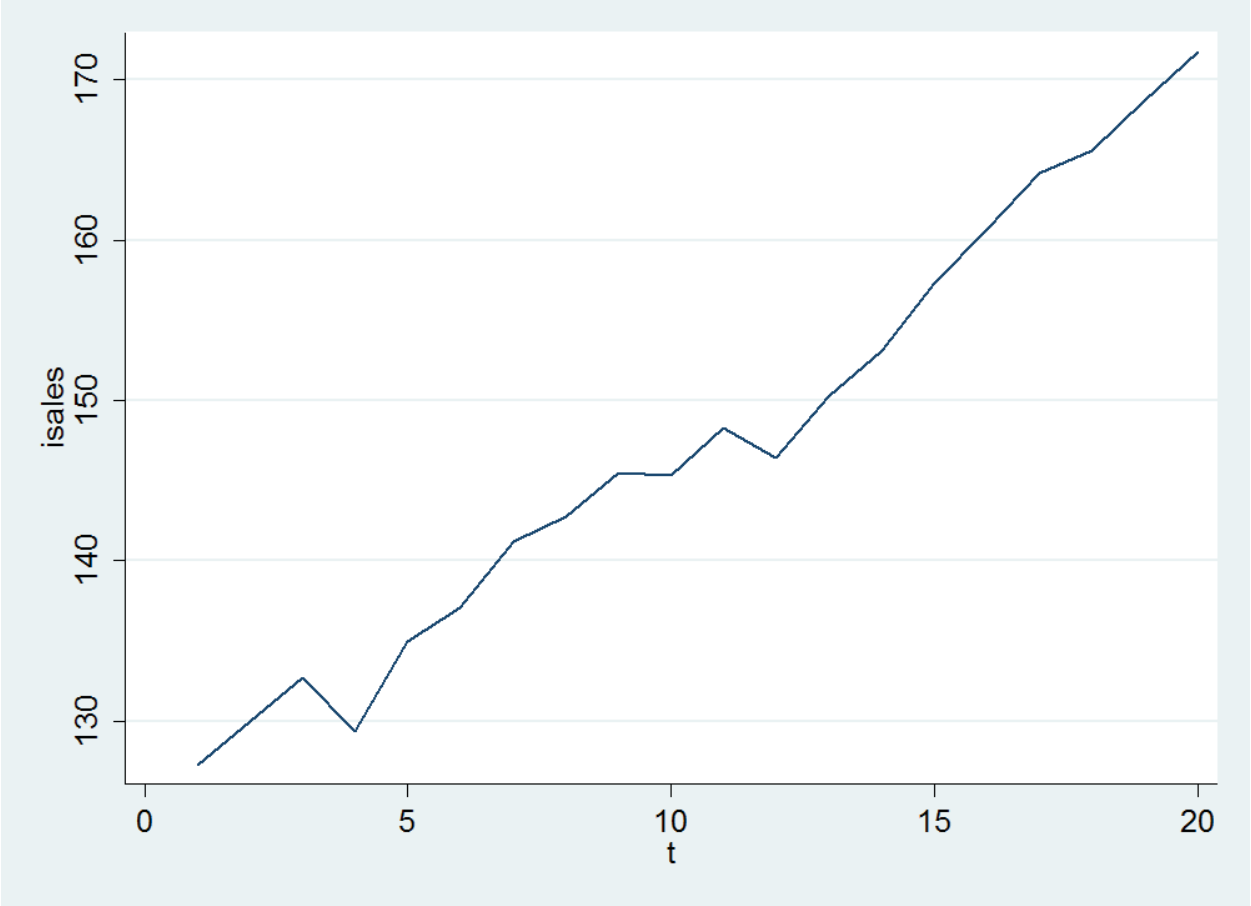
$$E(y_t) = \theta_0 + \theta_1 t$$

si (e_t) est une séquence i.i.d., alors (y_t) est une séquence indépendante, mais (e_t) sera sûrement corrélé dans le temps.

Tendance linéaire

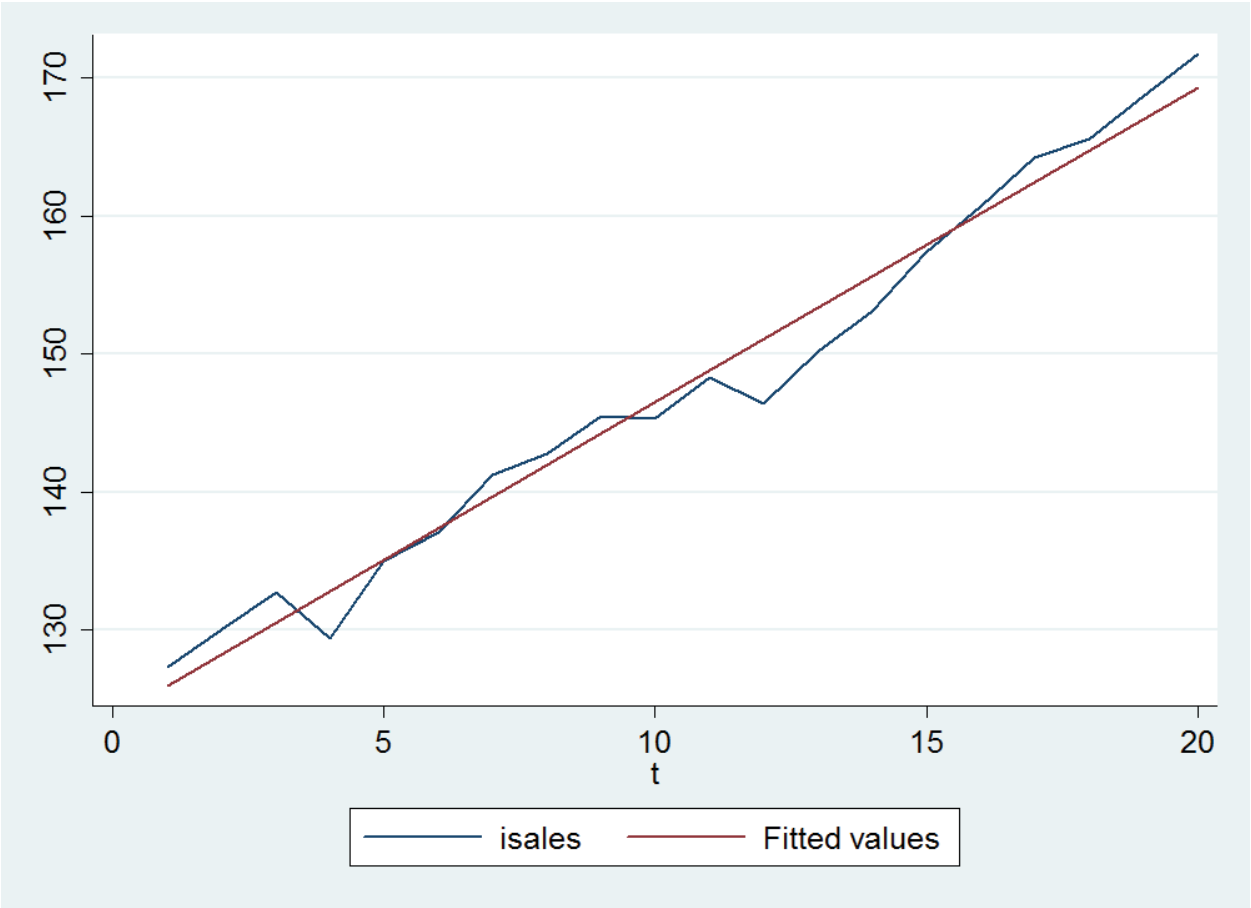
Nous allons maintenant représenter graphiquement les données pour faire ressortir les tendances.

```
webuse qsales

tsset t

tsline isales
```



Pour inclure graphiquement la tendance linéaire, nous utilisons l'option `lfit` dans la commande `twoway`.

```
twoway (line isales t) (lfit isales t)
```



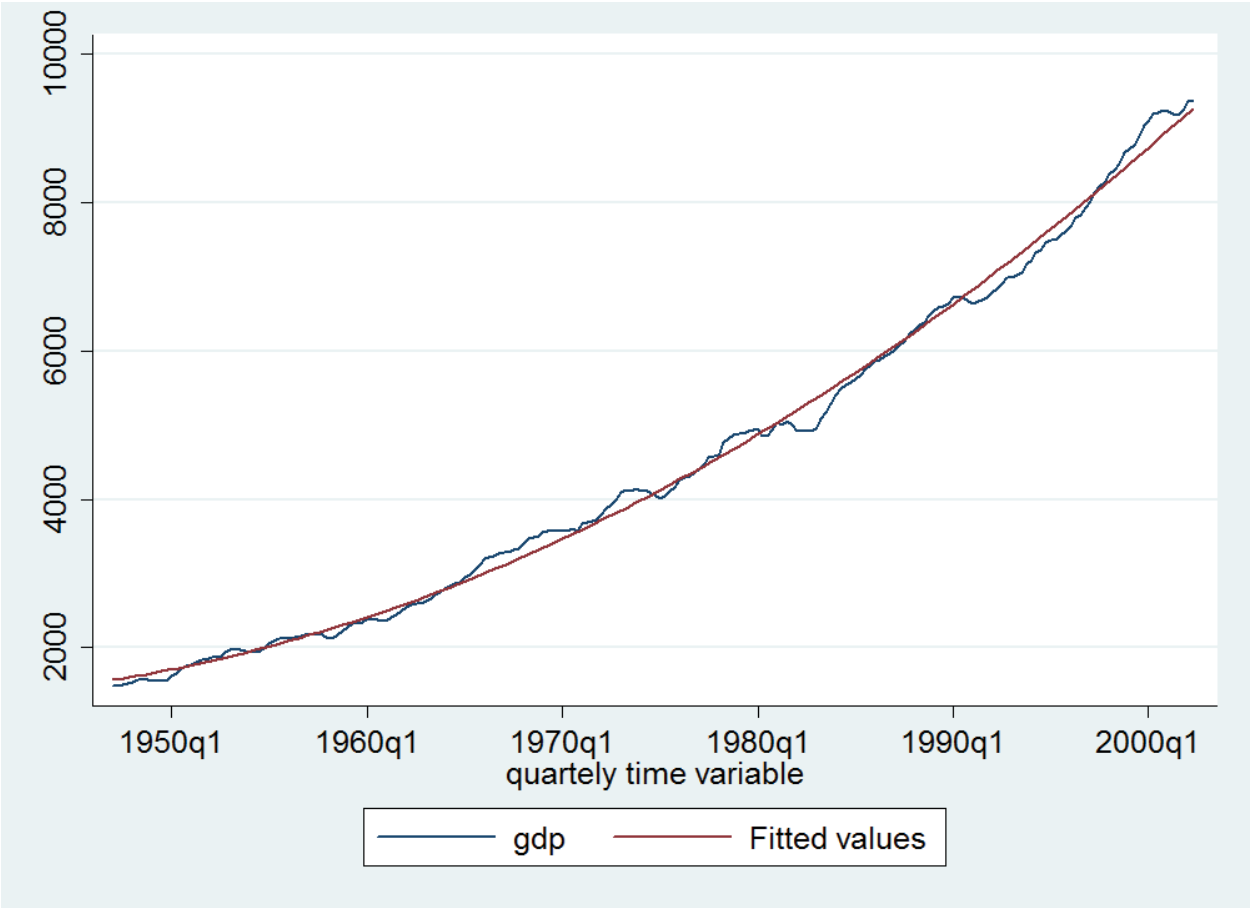
À partir du graphique, nous pouvons constater que ces données ont une tendance linéaire.

Tendance exponentielle

Nous pouvons modéliser la tendance avec un log, $\log(y_t) = \beta_0 + \beta_1 t + e_t$, t = 1,2, … n. Le coefficient $\beta_1 = \Delta \log(y_t) = \log(y_t) - \log(y_{t-1}) \approx \frac{(y_t - y_{t-1})}{y_{t-1}}$ qui est le taux de croissance.

Nous allons utiliser le fichier m1gdp de Stata et générer le logarithme naturel du *gdp*. Nous utiliserons l'option `qfit` dans la commande `twoway`.

```
webuse m1gdp

gen gdp=exp(ln_gdp)

twoway (line gdp t) (qfit gdp t)
```



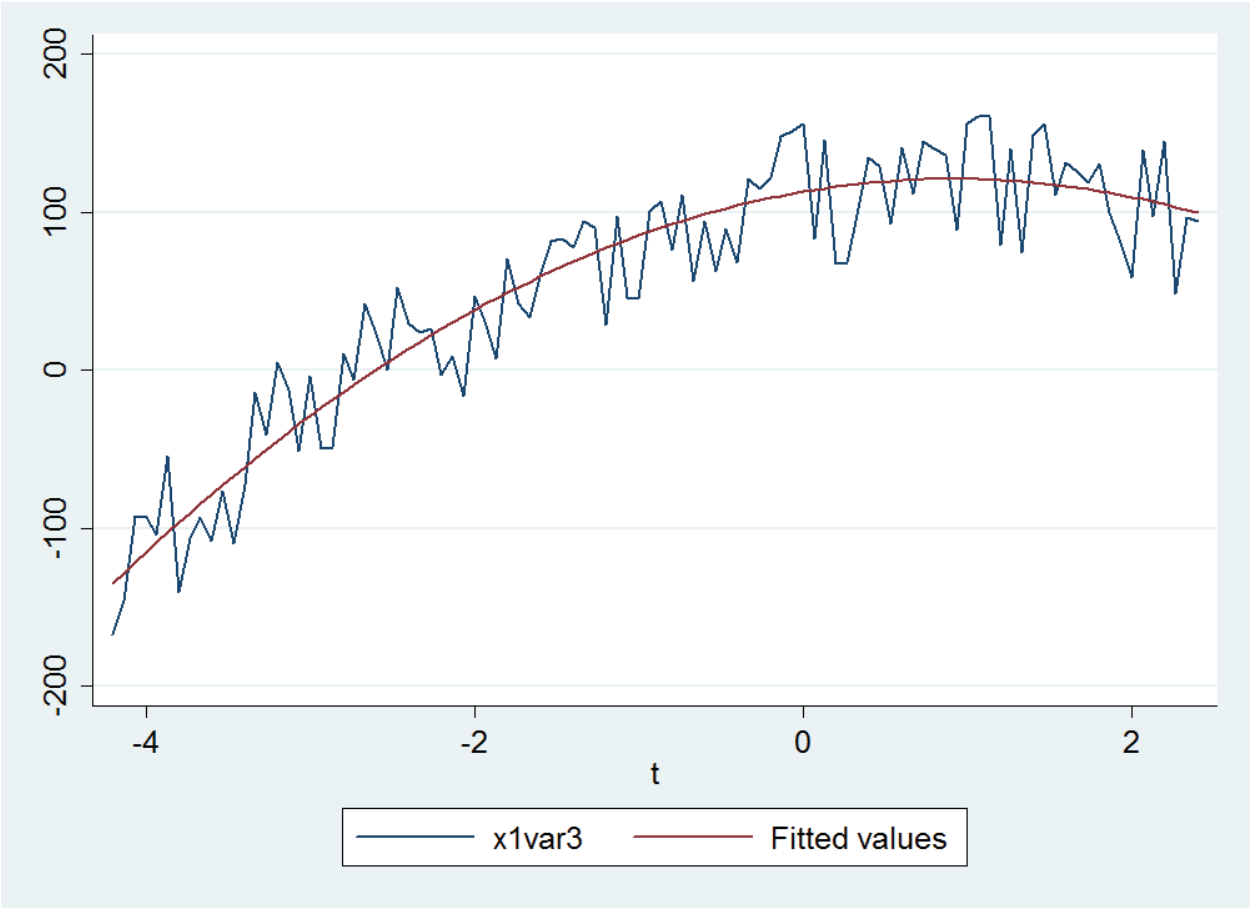
Pour enlever la tendance exponentielle, les données doivent être transformées en utilisant la fonction logarithme naturel.

Tendance Quadratique

Nous pouvons aussi modéliser la tendance avec une forme quadratique, $y_t = \vartheta_0 + \vartheta_1 t + \vartheta_2 t^2 + e_t$. Dans le graphique, $\vartheta_1 > 0 \; et \; \vartheta_2 > 0$. La pente approximative est $\frac{\Delta y_t}{\Delta t} = \vartheta_1 + 2\vartheta_2 t$

Nous générons une séquence de données pour 100 observations en simulant une tendance quadratique

```
clear
set obs 100
set seed 10101
generate x1var = runiform()
gen t=(_n-64)/15
gen r=_n+45
gen x1var2=(x1var-0.5)*100
gen t2=t-10*t^2
gen x1var3=x1var2+t2+r
twoway (line x1var3 t) (qfit x1var3 t)
```



Pour nous assurer de ne pas avoir une mauvaise forme fonctionnelle, nous devons inclure la variable de temps ainsi que la variable de temps au carré, t^2.

7.4 Utilisation de variables avec tendance dans une régression.

Les variables ayant une tendance ne violent pas nécessairement les hypothèses TS.1 à TS.6, mais il est possible que des facteurs non observés ayant une tendance influencent y et pourrait être corrélé avec les variables explicatives. Si nous ignorons les tendances, il est possible d'avoir une régression fallacieuse.

$$yt = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 t + \mu_t$$

La variable explicative t permet à $y \uparrow$ ou $\downarrow$ pour des raisons non reliées à x_1 et x_2. Si la régression respecte TS.1 à TS.3, alors omettre t implique un biais, surtout si x_1 et x_2 ont une tendance.

Parfois ajouter t rend les autres variables explicatives significatives surtout lorsque y et les X ont des tendances.

Une interprétation éliminant la tendance, inclure t permet une interprétation des coefficients sans tendance.

7.5 Méthode *partialling out*

Premièrement, nous estimons les régressions

$$y_t = \delta_0 + \delta_1 t + \varepsilon$$

$$x_{t1} = \theta_0 + \theta_1 t + \varepsilon$$

$$x_{t2} = \gamma_0 + \gamma_1 t + \varepsilon$$

Ensuite, nous générons les résidus de chaque régression, soit $\ddot{y}_t, \ddot{x}_{t1}, \ddot{x}_{t2}$. Nous avons ainsi des variables où la tendance a été éliminée linéairement.

Deuxièmement, nous estimons la régression suivante pour obtenir $\hat{\beta}_1$ *et* $\hat{\beta}_2$.

$$\ddot{y}_t = \beta_0 + \beta_1 \ddot{x}_{t1} + \beta_2 \ddot{x}_{t2} + \varepsilon$$

Si t est omis, alors la tendance n'est pas éliminée. Si y_t contient une tendance et que celle-ci est corrélée à un ou plusieurs des x_{tj}, alors les estimateurs sont biaisés, l'hypothèse d'indépendance entre le terme d'erreur et les variables explicatives est violée.

7.6 R^2 lorsqu'il y a une tendance

Le R^2 est souvent très élevé avec les séries temporelles.

$$R^2 = 1 - \frac{\hat{\sigma}_y^2}{\hat{\sigma}_y^2}$$

Pour éliminer l'effet de tendance dans le R^2, nous devons la calculer à partir de la régression

$$\ddot{y}_t \text{ sur } \ddot{x}_{t1} \; et \; \ddot{x}_{t2}$$

qui donnera

$$R^2 = 1 - \frac{SRC}{\sum \hat{y}^2}$$

Comme $\Sigma\ddot{y}^2 \leq \Sigma(y_t - \bar{y})^2$, le R^2 est plus petit et explique mieux les effets des variables indépendantes sur la variable dépendante.

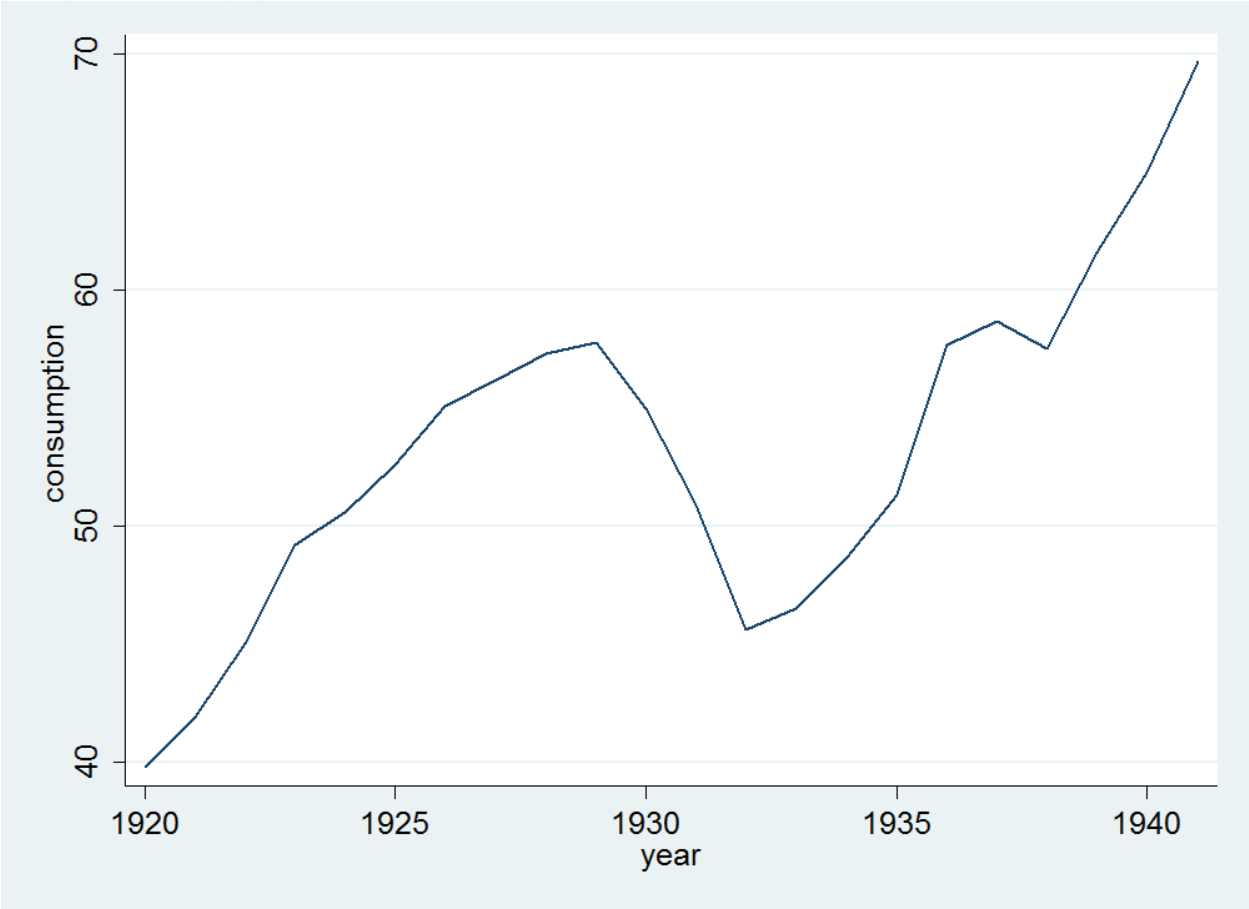
7.7 Tester pour des coupures dans la tendance

Coupure avec date inconnue

Test Quandt likelihood

Il est possible qu'une série temporelle ait des coupures. Comme nous voyons dans le graphique ci-dessous, la tendance de la consommation n'est pas constante, alors les coefficients ne sont pas constants à travers le temps :

```
webuse klein, clear
tsset yr
twoway line consump yr
```



Cependant, étant donné que nous ne pouvons généralement pas tester précisément quand les coupures ont lieu, le test « Quandt likelihood Ratio » nous permet de situer les possibles coupures. Selon le QLR, nous devons éliminer les premiers et derniers 15% des séries temporelles.

Programme Stata[16]

Nous générons premièrement de nouvelles variables clones qui seront utilisées dans le programme. Il faut donc générer la variable vt qui est un clone de notre variable de temps et la variable vardep qui est un clone de notre variable dépendante. Pour notre application nous avons donc

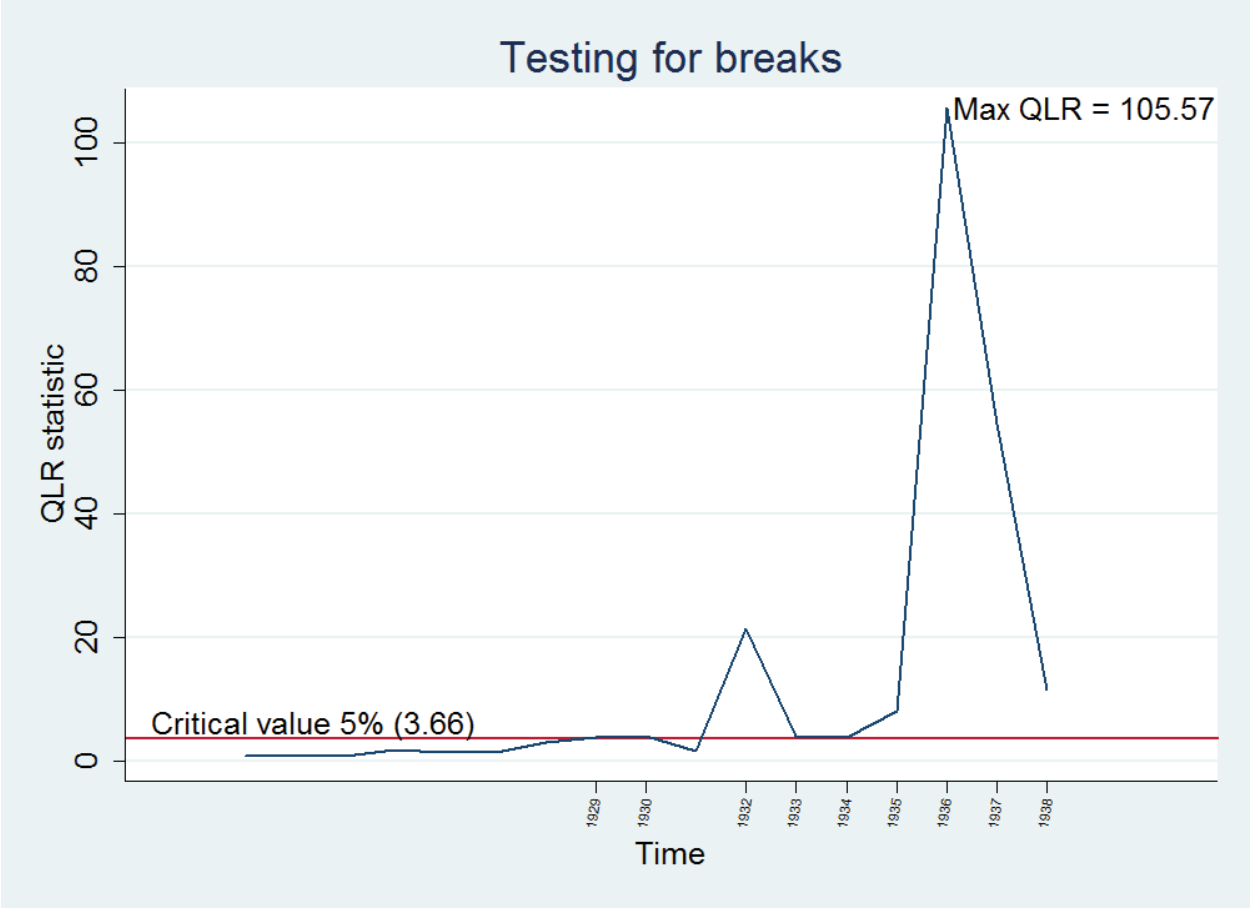
```
gen vt=yr
gen vardep=consump
sum vt
local time=r(max)-r(min)+1
local i = round(`time'*.15)
local f = round(`time'*.85)
local var = "vardep"
gen diff`var' = d.`var'
gen chow`var' = .
gen qlr`var' = .
set more off
while `i'<=(`f') {
gen di = (_n > `i')
cap gen d_`var'1 = di*l1.`var'
cap gen d_`var'2 = di*l2.`var'
cap gen d_`var'3 = di*l3.`var'
cap gen d_`var'4 = di*l4.`var'
qui reg diff`var' L(1/4).diff`var' di, r
qui test di
sca chow = r(F)
cap replace chow`var' = r(F) in `i'
qui reg diff`var' L(1/4).diff`var' di d_`var'1 d_`var'2 d_`var'3 d_`var'4, r
qui test di d_`var'1 d_`var'2 d_`var'3 d_`var'4
sca qlr = r(F)
cap replace qlr`var' = r(F) in `i'
dis "`i' " %ty vt[`i'] " " %8.3f chow " " %8.3f qlr
drop di d_`var'1 d_`var'2 d_`var'3 d_`var'4
local i = `i' + 1
}
sum qlr`var'
local maxvalue=r(max)
gen maxdate=vt if qlr`var'==`maxvalue'
local maxvalue1=round(`maxvalue',0.01)
local critical=3.66
*Replace with the appropriate critical value (see Stock & Watson)*
sum vt
local mindate=r(min)
sum maxdate
local maxdate=r(max)
gen break=vt if qlr`var'>=`critical' & qlr`var'!=.

levelsof break, local(break1)
1929 1930 1932 1933 1934 1935 1936 1937 1938
```

[16] Le programme apparaissant sur cette page est une modification du programme disponible dans, Torres-Reyna, Oscar, Time Series (ver. 1.5), présentation en ligne consultée le 28/08/2012 http://dss.princeton.edu/training/TS101.pdf

```
list vt qlr`var' if break!=.
     +-----------------+
     |   yr   qlrcon~p |
     |-----------------|
 10. | 1929   3.771517 |
 11. | 1930   4.046748 |
 13. | 1932   21.35588 |
 14. | 1933   4.062252 |
 15. | 1934   3.813575 |
     |-----------------|
 16. | 1935   8.038505 |
 17. | 1936    105.567 |
 18. | 1937   54.44305 |
 19. | 1938   11.48017 |
     +-----------------+
```

Comme nous pouvons voir, le test QLR nous donne des années où il y aurait possiblement des coupures. Cependant, nous ne pouvons pas considérer les années 1935 à 1938, car elles font parties du dernier 15 % de la série temporelle.

```
twoway tsline qlr`var', title(Testing for breaks) xlabel(`break1', angle(80) labsize(1.5)) yline(`critical') ytitle(QLR statistic) xtitle(Time) ttext(`critical' `mindate' "Critical value 5% (`critical')", placement(ne)) ttext(`maxvalue' `maxdate' "Max QLR = `maxvalue1'", placement(e))
```



Ce graphique nous permet de voir que les années 1929, 1930, 1933 et 1934 dépassent très légèrement le seuil critique de 5 % de la statistique *sup-Wald* tandis que l'année 1932 le dépasse largement. Les dernières années ne sont pas représentatives et la statistique de *sup-Wald* ne doit pas être prise en compte étant donné qu'elles font parties du dernier 15 % des données.

Coupure avec date connue

Test de Chow

Le test de Chow permet de vérifier s'il existe une coupure pour les dates proposées par le test *Quandt likelihood*. La procédure de Chow teste si les coefficients sont différents à travers le temps.

Supposons que sous H_0 le modèle est

$$y_i = X'_i\beta + e_i \, pour \, i = 1, \dots, n$$

et sous H_1

$$y_i = X'_i\beta_1 + e_i \, pour \, i = 1, \dots, n$$

$$y_i = X'_i\beta_2 + e_i \, pour \, i = n+1, \dots, m.$$

Nous voulons tester si $\beta_1 = \beta_2$.

Ce modèle est égal à :

$$y = X\beta_1 + Z\delta + e,$$

où $\delta = \beta_2 - \beta_1$. Alors, un test F est approprié et correspond à tester l'hypothèse nulle $\delta = 0$.

Soit SSR_{nc}, le SSR du modèle $y = X\beta_1 + Z\delta + e$, nous pouvons démontrer que

$SSR_{nc} = SSR_1 + SSR_2$,

Où SSR_1 et SSR_2 sont les SSR des modèles ($y_i = x'_i\beta_1 + e_i$) et ($y_i = x'_i\beta_2 + e_i$) respectivement.

Donc,

$$F = \frac{(SSR_c - SSr_1 - SSR_2)/k}{SSR_1 + SSR_2/(n-2k)}$$

Suit une distribution F(k, n-2k) sous H_0, car il y a 2k paramètres dans le modèle non contraint.

Nous voulons tester si les coefficients sont différents entre deux groupes d'observations. Pour faire un test de Chow il faut créer une variable binaire qui différencie les deux groupes. On fait ensuite la régression de départ à laquelle on ajoute la variable binaire ainsi que les termes d'interactions de toutes les variables avec la variable binaire de groupe

Commande Stata

```
gen break1 = (yr>(1932))

generate break1_invest = break1*invest

reg consump invest break1 break1_invest
      Source |       SS       df       MS              Number of obs =      22
-------------+------------------------------           F(  3,    18) =   10.56
       Model |    722.9735     3  240.991167           Prob > F      =  0.0003
    Residual |  410.801378    18  22.8222988           R-squared     =  0.6377
-------------+------------------------------           Adj R-squared =  0.5773
       Total |  1133.77488    21  53.9892799           Root MSE      =  4.7773

------------------------------------------------------------------------------
      consump |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
       invest |   .7006278   .3875858     1.81   0.087    -.1136597    1.514915
       break1 |   7.774459   2.226425     3.49   0.003     3.096914      12.452
break1_invest |   1.464618   .6477659     2.26   0.036     .1037123    2.825524
        _cons |   49.08331   1.550343    31.66   0.000     45.82616    52.34046
------------------------------------------------------------------------------

test break break_invest

 ( 1)  break = 0
 ( 2)  break_invest = 0
       F(  2,    18) =   11.54
            Prob > F =    0.0006
```

Nous pouvons rejeter l'hypothèse nulle qu'il n'y a pas de différence dans les coefficients. Dans une telle situation, nous pouvons sélectionner une seule partie de la série temporelle pour l'étude si nous avons suffisamment de données, sinon les coefficients estimés seront biaisés et faux, ou il faudra créer une variable binaire afin d'indiquer qu'il y a eu une brisure dans la série temporelle.

7.8 Saisonnalité

Si les données sont mensuelles ou trimestrielles, il y a possibilité de saisonnalité. Nous devons modéliser y_t en fonction des mois ou trimestres. Pour désaisonnaliser notre série, nous pouvons inclure des variables muettes saisonnières, soit n-1 variables muettes. Si nos données sont mensuelles, nous devons en inclure 11, afin de garder une variable de référence et ainsi éliminer la multicolinéarité.

$$yt = \beta_0 + \beta_1 x_{t1} + \dots + \beta_k x_{tk} + \delta_1 m_1 + \dots + \delta_{11} m_{11} + u_t$$

Pour tester la saisonnalité, nous devons faire un test de Fisher suivant les hypothèses suivantes :

H_0 : $\delta_1 = 0, \dots, \delta_{11} = 0$ (aucune saisonnalité) vs

H_1 : $\delta_1 \geq 0, \dots, \delta_{11} \geq 0$ (il y a saisonnalité si au moins un des termes n'égale pas à zéro)

$$F = \frac{(R_r^2 - R_{ur}^2)/q}{(1 - R_{ur}^2)/(n-k-1)}$$

Commandes Stata

Pour tester s'il y a présence de saisonnalité, il faut créer des variables muettes mensuelles ou trimestrielles selon la nature des données.

```
gen q = quarter(dofq(t))
gen q1 = (q==1)
gen q2 = (q==2)
gen q3 = (q==3)
gen q4 = (q==4)
```

Faites la régression en incluant toutes les variables explicatives et muettes sauf la variable muette de référence q4 dans ce cas-ci.

```
reg y x q1 q2 q3
```

Faites le test de significativité jointe pour tester s'il y a présence de saisonnalité (test Fisher)

```
test q1 q2 q3
```

Si vous ne trouvez pas de présence de saisonnalité, probablement qu'une tendance linéaire ou exponentielle sera suffisante (selon ce que vous avez obtenu lorsque vous avez fait le graphique de vos données) pour votre régression.

S'il y a présence de saisonnalité, vous pouvez soit :

1- inclure les variables muettes dans la régression (sauf la variable de référence); ou
2- enlever les effets saisonniers dans les variables par la méthode suivante :

Enlever la saisonnalité

Premièrement, nous effectuons les régressions suivantes pour enlever l'effet saisonnier de nos variables

Régresser y_t, x_{t1}, x_{t2} sur une constante et sur les variables muettes. Ensuite, générer les résidus de chacune des régressions, soit :

$\ddot{y}_t$, $\ddot{x}_{t1}$, $\ddot{x}_{t2}$, ex : $\ddot{y}_t = y_t - \hat{\alpha}_0 - \hat{\alpha}_1 m_1 - \ldots - \hat{\alpha}_{11} m_{11}$

alors $\ddot{y}_t$ est désaisonnalisé linéairement.

1- Régresser $\ddot{y}_t$ sur $\ddot{x}_{t1}$ et $\ddot{x}_{t2}$ ce qui donne l'estimation de β_1 et β_2

Étapes sur Stata

```
webuse air2, clear
reg air q1 q2 q3
predict ry, resid
reg x1 q1 q2 q3
predict rx1, resid
reg x2 q1 q2 q3
predict rx2, resid
reg ry rx1 rx2
```

7.9 Stationnarité et test de racine unitaire

Si nous trouvons de l'autocorrélation, la prochaine étape avant de déterminer l'ordre du modèle autorégressif est de déterminer si les séries sont stationnaires, faiblement dépendantes.

Une série temporelle est stationnaire lorsque son processus stochastique a une distribution des probabilités stable à travers le temps

Test de Dickey-Fuller

La commande `dfuller` permet de faire le test Dickey-Fuller Augmenté. Elle teste si une variable a une présence de racine unitaire. Nous devons donc tester toutes les variables incluses dans le modèle.

$$y_t = \delta + \rho y_{t-1} + e_t, t = 1, 2, \ldots, n$$

$H_0 : \rho = 1$ (racine unitaire)

$H_1 : \rho \neq 1$

Lorsque $|\rho| < 1$, $\{y_t\}$ est un processus stable, alors faiblement dépendant.

La fonction permet d'inclure une tendance, un drift et des lags

Exemple

```
webuse air2, clear
(TIMESLAB: Airline passengers)

. tsset t
        time variable:  t, 1 to 144
                delta:  1 unit

. dfuller air

Dickey-Fuller test for unit root                   Number of obs   =       143

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -1.748            -3.496            -2.887            -2.577
------------------------------------------------------------------------------
MacKinnon approximate p-valeur  for Z(t) = 0.4065
```

À partir de l'exemple, nous ne pouvons pas rejeter l'hypothèse de racine unitaire.

Nous pouvons aussi tester la racine unitaire en incluant une tendance.

```
dfuller air, trend

Dickey-Fuller test for unit root                   Number of obs   =       143

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -4.639            -4.026            -3.444            -3.144
------------------------------------------------------------------------------
MacKinnon approximate p-valeur  for Z(t) = 0.0009
```

Nous rejetons l'hypothèse de racine unitaire lorsque nous ajoutons l'option d'une tendance

- Si la série temporelle a une tendance linéaire, alors prenez la première différence
- Si la série temporelle a une tendance exponentielle, prenez le logarithme naturel et ensuite prenez la première différence.
- Si la série temporelle a une saisonnalité significative, il faut inclure des variables muettes pour éliminer la saisonnalité

Pour rendre la série stationnaire, il faut prendre la première différence et ensuite nous allons de nouveau tester s'il y a présence de racine unitaire.

```
dfuller d.air

Dickey-Fuller test for unit root                   Number of obs   =       142

                               ---------- Interpolated Dickey-Fuller ---------
                  Test         1% Critical       5% Critical      10% Critical
               Statistic           Value             Value             Value
------------------------------------------------------------------------------
 Z(t)             -8.580            -3.496            -2.887            -2.577
```

```
------------------------------------------------------------------------------
MacKinnon approximate p-valeur  for Z(t) = 0.0000
```

En prenant la première différence, nous avons transformé une série avec une racine unitaire en une série covariante-stationnaire.

Il faut donc estimer le modèle en transformant la variable air dans notre cas en créant une nouvelle variable qui est la différence de air et du lag de air.

7.10 Autocorrélation

Il faut aussi tester l'autocorrélation pour détecter les mauvaises spécifications dynamiques. Les modèles statiques et à retard fini ont souvent des erreurs autocorrélées même s'il n'y a aucune mauvaise spécification.

Test pour l'autocorrélation

Test t pour AR(1) avec exogénéité stricte

$u_t = \rho u_{t-1} + e_t$ où $t = 2, \ldots, n$, $\{u_t\}$ est faiblement dépendant

Nous supposons $E(e_t | u_{t-1}, u_{t-2}, \ldots) = 0$ et Var $(e_t | u_{t-1})$ = Var$(e_t) = \sigma_e^2$

Sous H_0 : $\rho = 0$ on suppose l'absence d'autocorrélation

(i) Régresser par MCO $y_t \; sur \; x_{t-1}, \ldots, x_{t-k}$ et obtenons les résidus MCO $\hat{u}_t$ $\forall t$

(ii) Régresser $\hat{u}_t$ sur $\hat{u}_{t-1}$

(iii) Tester $\hat{\rho}$ avec la statistique t

Exemple

En prenant le même fichier que précédemment, nous estimons le modèle suivant

```
reg air t

      Source |       SS       df       MS              Number of obs =     144
-------------+------------------------------           F(  1,   142) =  828.20
       Model |  1756825.04     1  1756825.04           Prob > F      =  0.0000
    Residual |  301219.118   142   2121.2614           R-squared     =  0.8536
-------------+------------------------------           Adj R-squared =  0.8526
       Total |  2058044.16   143  14391.9172           Root MSE      =  46.057

------------------------------------------------------------------------------
         air |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |   2.657184   .0923325    28.78   0.000      2.47466    2.839708
```

```
       _cons |   87.65278   7.716347    11.36   0.000     72.39902    102.9065
------------------------------------------------------------------------------

predict r, resid

reg r l.r

      Source |       SS       df       MS              Number of obs =     143
-------------+------------------------------           F(  1,   141) =  161.41
       Model |  160519.244     1  160519.244           Prob > F      =  0.0000
    Residual |  140226.127   141  994.511537           R-squared     =  0.5337
-------------+------------------------------           Adj R-squared =  0.5304
       Total |  300745.371   142  2117.92515           Root MSE      =  31.536

------------------------------------------------------------------------------
           r |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           r |
         L1. |   .7317944   .0576011    12.70   0.000      .617921    .8456677
             |
       _cons |  -.3476115   2.637208    -0.13   0.895     -5.56119    4.865967
------------------------------------------------------------------------------
```

Nous voyons, à partir de la statistique t, que nous sommes en présence d'autocorrélation puisque la période de retard des résidus est significative.

7.11 Test de Durbin-Watson sous les hypothèses classiques

C'est un autre test pour AR(1)

La statistique DW est basée sur les résidus de MCO

$$DW = \frac{\Sigma(\hat{u}_t - \hat{u}_{t-1})^2}{\Sigma \hat{u}_t^2} \approx 2(1 - \hat{\rho})$$

La distribution de la statistique DW nécessite les hypothèses classiques

$$H_0: \rho = 0 \; vs \; H_1: \rho > 0$$

$$si \; \hat{\rho} \approx 0 \implies DW \approx 2$$

$$si \; \hat{\rho} > 0 \implies DW < 2$$

$$si \; \hat{\rho} < 0 \implies DW > 2$$

Nous recherchons une valeur qui est significativement inférieure ou supérieure à 2 pour rejeter H_0. Cependant, nous devons comparer le test de DW avec 2 valeurs critiques d_U (*upper limit*) et d_L (*lower limit*).

$$si \; DW < d_L \implies rejeter \; H_0$$

$si\ DW > d_U \Longrightarrow rejeter\ H_0$

$si\ d_L \leq DW \leq d_U \Longrightarrow pas\ de\ conclusion$

Pour avoir les 2 valeurs critiques, le lecteur est invité à consulter le site http://www.stanford.edu/~clint/bench/dwcrit.htm

Exemple

```
webuse klein, clear

. tsset yr
        time variable:  year, -11 to 10
                delta:  1 unit

reg consump wagegovt if yr>1932

      Source |       SS       df       MS              Number of obs =      22
-------------+------------------------------           F(  1,    20) =   17.72
       Model |  532.567711     1  532.567711           Prob > F      =  0.0004
    Residual |  601.207167    20  30.0603584           R-squared     =  0.4697
-------------+------------------------------           Adj R-squared =  0.4432
       Total |  1133.77488    21  53.9892799           Root MSE      =  5.4827

------------------------------------------------------------------------------
     consump |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    wagegovt |    2.50744   .5957173     4.21   0.000     1.264796    3.750085
       _cons |   40.84699   3.192183    12.80   0.000     34.18821    47.50577
------------------------------------------------------------------------------

estat dwatson

Durbin-Watson d-statistic(  2,    22) =  .3217998
```

Selon la statistique Durbin-Watson, nous testons s'il y a présence d'autocorrélation d'ordre 1 et nous rejetons l'hypothèse nulle qu'il n'y a pas présence d'autocorrélation.

Test AR(1) d'autocorrélation sans exogénéité stricte

Ni la statistique-t, ni DW ne sont valides lorsque n est grand. Pour contrer le problème, Durbin's h stat.

(i) Régresser par MCO y_t sur $x_{t1}, \ldots, x_{tk}$ et obtenir les résidus de MCO $\hat{u}_t$

(ii) Régressez $\hat{u}_t$ sur $x_{t1}, \ldots, x_{tk}, \hat{u}_{t-1}$ pour $t = 2, \ldots n$ et obtenir $\hat{\rho}$ qui est le coefficient de $\hat{u}_{t-1}$

(iii) Tester $H_0 : \hat{\rho} = 0$

La t calculée du résidu à retard est un test valide de $H_0 : \hat{\rho} = 0$ dans un modèle AR(1) sous $Var(u_t \mid x_t, u_{t-1}) = \sigma^2$. L'inclusion de $x_{t1}, \ldots, x_{tk}$ permet à x_{tj} d'être corrélé avec u_{t-1} et assure que $t_{\hat{\rho}}$ soit une distribution approximative.

Test d'autocorrélation de Breusch-Godfrey

Nous testons $H_0 : \rho_1 = 0, \rho_2 = 0, \ldots, \rho_q = 0$ du modèle $u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \ldots + \rho_q u_{t-q} + e_t$

(i) Régresser par MCO y_t sur $x_{t1}, \ldots, x_{tk}$ et obtenir les résidus de MCO $\hat{u}_t$

(ii) Régressez $\hat{u}_t$ sur $x_{t1}, \ldots, x_{tk}, \hat{u}_{t-1}, \ldots, \hat{u}_{t-q}$ $\forall t = (q+1), \ldots, n$

(iii) Test joint de Fisher (F-test) sur $\hat{u}_{t-1}, \ldots, \hat{u}_{t-q}$

Avec des données trimestrielles ou mensuelles non désaisonnalisées, nous voudrions tester pour la forme saisonnière d'autocorrélation. Dans le cas trimestriel, les erreurs suivent par exemple $u_t = \rho_4 u_{t-4} + e_t$.

Exemple

```
webuse klein
tsset yr
quietly regress consump wagegovt L.consump L2.consump

estat bgodfrey

Breusch-Godfrey LM test for autocorrelation
---------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-------------------------------------------------------------
       1     |         14.264               1                   0.0002
---------------------------------------------------------------------------
                        H0: no serial correlation

. estat bgodfrey, small lags(1/2)

Breusch-Godfrey LM test for autocorrelation
---------------------------------------------------------------------------
    lags(p)  |          F                  df                 Prob > F
-------------+-------------------------------------------------------------
       1     |       14.264           (  1,   19 )              0.0013
       2     |        8.078           (  2,   18 )              0.0031
---------------------------------------------------------------------------
                        H0: no serial correlation
```

Nous pouvons remarquer que selon les deux tests, il y a de l'autocorrélation dans la variable dépendante.

Durbin Alternative Test

Ce test diffère du test classique de Durbin Watson parce qu'il n'impose pas l'exogénéité stricte des régresseurs. Ce test vérifie aussi l'autocorrélation, mais pas pour un ordre particulier. Ce test est un cas particulier du test Breusch-Godfrey, qui considère plus qu'un 1 ordre, et donc est plus général.

H_0 : aucune présence d'autocorrélation

H_1 : présence d'autocorrélation

Exemple

```
estat durbinalt, small

Durbin's alternative test for autocorrelation
---------------------------------------------------------------------------
    lags(p)  |          F                  df                 Prob > F
-------------+-------------------------------------------------------------
       1     |       35.035           (  1,   19 )              0.0000
---------------------------------------------------------------------------
                        H0: no serial correlation
```

Nous pouvons encore une fois rejeter l'hypothèse nulle qu'il n'a pas de présence d'autocorrélation.

7.12 Modèle Autorégressif

La prochaine étape lorsque nous avons transformé nos séries en un processus stationnaire, est de déterminer l'ordre du modèle autorégressif. Pour déterminer l'ordre, nous pouvons soit :

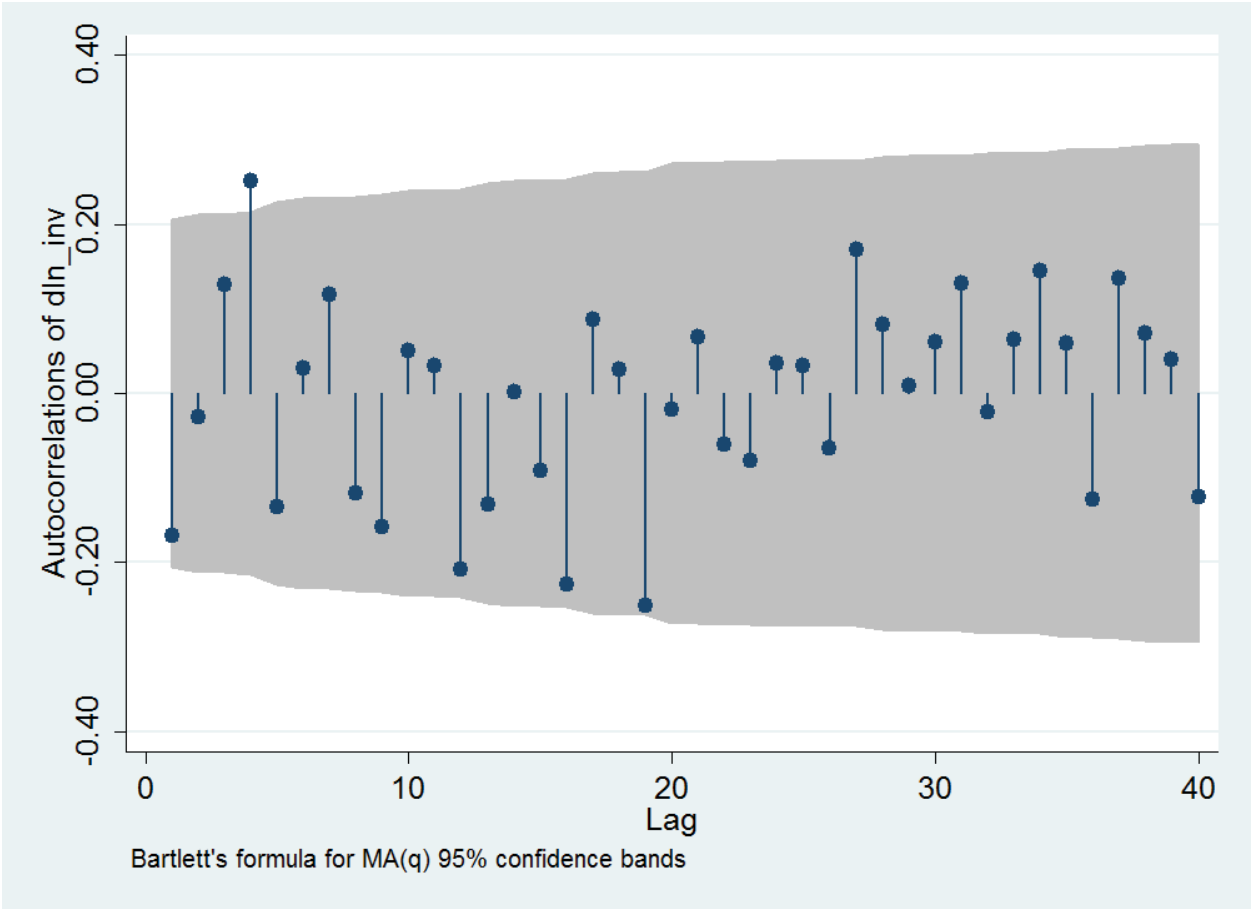
1. Tester l'autocorrélation pour un AR(q) jusqu'au moment qu'il n'y a plus d'autocorrélation
2. Nous demandons la statistique AIC (Akaike's Information Criteria) avec la fonction `varsoc` dans Stata
3. Faites l'autocorrélogramme avec la fonction `corrgram` et `ac` pour faire le graphique

Exemple

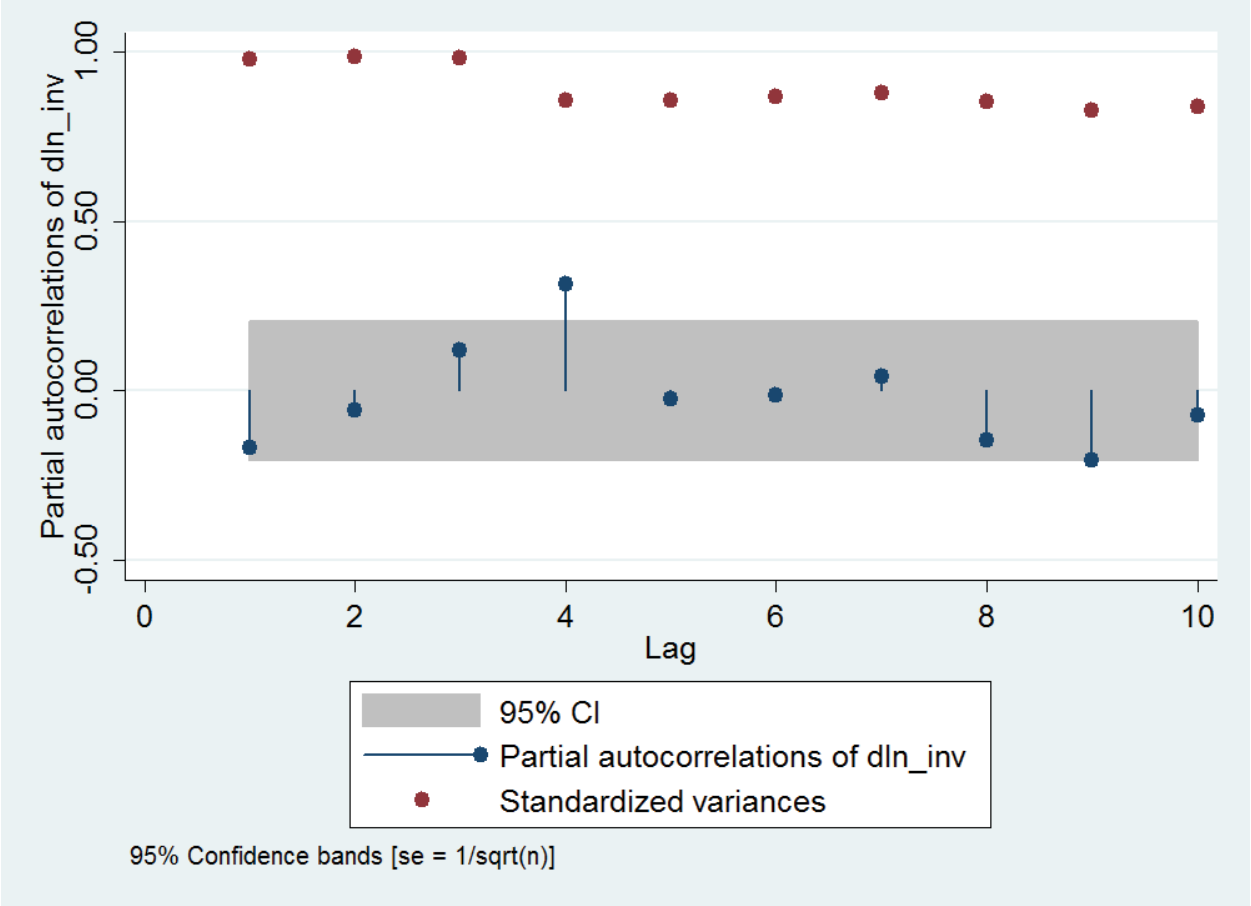
```
webuse lutkepohl2
(Quarterly SA West German macro data, Bil DM, from Lutkepohl 1993 Table E.1)

varsoc dln_inv

   Selection-order criteria
   Sample:  1961q2 - 1982q4                     Number of obs      =        87
  +---------------------------------------------------------------------------+
  |lag |    LL      LR      df    p      FPE       AIC      HQIC      SBIC    |
  |----+----------------------------------------------------------------------|
  |  0 |  147.073                      .002038  -3.35799  -3.34658  -3.32965* |
  |  1 |  148.538  2.9308    1  0.087  .002016  -3.36869  -3.34587    -3.312  |
  |  2 |  148.734   .3913    1  0.532  .002054   -3.3502  -3.31596  -3.26517  |
  |  3 |  149.472   1.476    1  0.224  .002066  -3.34418  -3.29853   -3.2308  |
  |  4 |  154.123  9.3022*   1  0.002    .0019* -3.42811* -3.37105* -3.28639  |
  +---------------------------------------------------------------------------+
   Endogenous:  dln_inv
    Exogenous:  _cons
```

Pour faire le graphique nous utilisons la commande

```
ac dln_inv
```




```
corrgram dln_inv, lags(10)

                                          -1       0       1 -1       0       1
 LAG       AC       PAC      Q     Prob>Q  [Autocorrelation]  [Partial Autocor]
-------------------------------------------------------------------------------
1       -0.1690  -0.1693   2.6862  0.1012          -|                  -|
2       -0.0288  -0.0591   2.7649  0.2510           |                   |
3        0.1283   0.1179   4.3473  0.2263           |-                  |
4        0.2512   0.3133   10.485  0.0330           |--                 |--
5       -0.1349  -0.0264   12.277  0.0312          -|                   |
6        0.0300  -0.0145   12.366  0.0543           |                   |
7        0.1170   0.0430   13.745  0.0559           |                   |
8       -0.1176  -0.1463   15.156  0.0562           |                  -|
9       -0.1582  -0.2046   17.739  0.0383          -|                  -|
10       0.0498  -0.0731   17.998  0.0550           |                   |

pac dln_inv, srv lag(10)
```



Nous pouvons voir que dans chacun des tableaux ci-dessus, le nombre de périodes où le retard est autocorrélé est de quatre. Dans le premier, `varsoc` l'astérisme nous indique que la dernière période où il y a autocorrélation est la 4e. De même, avec le graphique `ac` nous constatons que la 4e période sort de du nuage gris, la zone de non-significativité à 95 %. Ce qui confirme le `varsoc`. La dernière colonne du `corrgram`, la corrélation partielle, nous indique qu'avec le trait dominant de la 4e période de retard que les autocorrélations suivantes sont mineures.

7.13 Corriger l'autocorrélation

MCQG sous AR(1)

ρ est rarement connu, alors la méthode des MCG est problématique

(i) Régresser y_t sur $x_{t1}, \ldots, x_{tk}$ et obtenir les résidus MCO $\hat{u}_t$

(ii) Régresser $\hat{u}_t$ sur $\hat{u}_{t-1}$ et obtenir ρ

(iii) Régresser $\tilde{y}_t = \beta_0\tilde{x}_{t0} + \beta_1\tilde{x}_{t1} + \cdots + \beta_k\tilde{x}_{tk} + e_t$

Nous pouvons utiliser 2 méthodes d'estimation; la méthode Cochrane-Orcutt et la méthode Prais-Winsten. La première méthode omet la première observation et la seconde méthode l'utilise.

Les méthodes CO et PW utilisent une méthodologie par itération afin de calculer ρ.

1. Une fois que l'estimateur MCQG est trouvé, nous calculons un ensemble de résidus
2. Nous obtenons un nouvel estimateur de ρ de $\hat{u}_t$ sur $\hat{u}_{t-1}$
3. Nous transformons les données utilisant le nouveau ρ
4. Nous estimons par MCO $\tilde{y}_t \; sur \; \tilde{x}_t$
5. Nous répétons plusieurs fois

Exemple Prais-Winsten

```
webuse idle, clear

tsset t
        time variable:  t, 1 to 30
                delta:  1 unit

varsoc usr

   Selection-order criteria
   Sample:  5 - 30                              Number of obs      =        26
  +---------------------------------------------------------------------------+
  |lag |    LL      LR      df    p      FPE       AIC      HQIC      SBIC    |
  |----+----------------------------------------------------------------------|
  |  0 | -71.0621                      14.9602   5.54323   5.55717   5.59162  |
  |  1 | -63.1521   15.82    1  0.000  8.79448    5.0117   5.03956   5.10847  |
  |  2 | -60.1347  6.0347*   1  0.014  7.53584   4.85652   4.89832   5.00168  |
  |  3 | -58.2722  3.7251    1  0.054  7.06217*  4.79017*   4.8459*  4.98372* |
  |  4 | -57.7302  1.0839    1  0.298  7.33291    4.8254   4.89507   5.06734  |
  +---------------------------------------------------------------------------+
```

Nous détectons une autocorrélation d'ordre 3 (AIC). Nous allons donc utiliser la commande `prais` pour la corriger.

```
prais usr idle

Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.3518
Iteration 2:  rho = 0.4731
Iteration 3:  rho = 0.5232
Iteration 4:  rho = 0.5425
Iteration 5:  rho = 0.5496
Iteration 6:  rho = 0.5522
Iteration 7:  rho = 0.5531
Iteration 8:  rho = 0.5534
Iteration 9:  rho = 0.5535
Iteration 10:  rho = 0.5535
Iteration 11:  rho = 0.5535
Iteration 12:  rho = 0.5535
Iteration 13:  rho = 0.5535
Iteration 14:  rho = 0.5535

Prais-Winsten AR(1) régression -- iterated estimates
```

```
      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =    7.12
       Model |  43.0076941     1  43.0076941           Prob > F      =  0.0125
    Residual |  169.165739    28  6.04163354           R-squared     =  0.2027
-------------+------------------------------           Adj R-squared =  0.1742
       Total |  212.173433    29  7.31632528           Root MSE      =   2.458

------------------------------------------------------------------------------
         usr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        idle |  -.1356522   .0472195    -2.87   0.008    -.2323769   -.0389275
       _cons |   15.20415   4.160391     3.65   0.001     6.681978    23.72633
-------------+----------------------------------------------------------------
         rho |   .5535476
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.295766
Durbin-Watson statistic (transformed) 1.476004
```

Le paramètre rho est estimé par itération et il est ensuite pris en compte dans la régression. Nous avons ainsi des estimateurs qui ont été corrigés pour l'autocorrélation d'ordre (q), dans notre cas 3.

Exemple Cochrane-Orcutt

```
prais usr idle, corc

Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.3518
Iteration 2:  rho = 0.4854
Iteration 3:  rho = 0.5411
Iteration 4:  rho = 0.5612
Iteration 5:  rho = 0.5677
Iteration 6:  rho = 0.5698
Iteration 7:  rho = 0.5705
Iteration 8:  rho = 0.5707
Iteration 9:  rho = 0.5708
Iteration 10:  rho = 0.5708
Iteration 11:  rho = 0.5708
Iteration 12:  rho = 0.5708
Iteration 13:  rho = 0.5708

Cochrane-Orcutt AR(1) régression -- iterated estimates

      Source |       SS       df       MS              Number of obs =      29
-------------+------------------------------           F(  1,    27) =    6.49
       Model |  40.1309584     1  40.1309584           Prob > F      =  0.0168
    Residual |  166.898474    27  6.18142498           R-squared     =  0.1938
-------------+------------------------------           Adj R-squared =  0.1640
       Total |  207.029433    28  7.39390831           Root MSE      =  2.4862

------------------------------------------------------------------------------
         usr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        idle |  -.1254511   .0492356    -2.55   0.017    -.2264742    -.024428
       _cons |   14.54641   4.272299     3.40   0.002      5.78038    23.31245
-------------+----------------------------------------------------------------
         rho |   .5707918
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.295766
Durbin-Watson statistic (transformed) 1.466222
```

En ajoutant l'option `corc`, nous entreprenons le test CO.

Nous pouvons voir une différence dans les résultats des 2 régressions. Étant donné que la méthode CO omet la première observation et que nous avons peu d'observations, les régresseurs estimations diffèrent dans les deux modèles.

La fonction `prais` permet aussi d'inclure l'option `robust` pour contrer l'hétéroscédasticité.

Écarts-types robustes à l'autocorrélation – Newey West

Estimation par MCO et corriger les écarts-types

- par exogénéité stricte alors MCQG n'est pas consistent
- avec MCQG les erreurs sont supposées suivre AR(1)

$$yt = \beta_0 + \beta_1 x_{t1} + \dots + \beta_k x_{tk} + e_t$$

x_{t1} est une fonction linéaire des autres variables explicatives et erreurs soit;

$$x_{t1} = \delta_0 + \delta_2 x_{t2} + \dots + \delta_k x_{tk} + r_t$$

$$Var(\hat{\beta}) = \frac{Var\Sigma r_t u_t}{(\Sigma \mathrm{E}(r_i^2))^2}$$

(i) Régresser par MCO ce qui donne se($\ddot{\beta}$), σ, résidus

(ii) Calculer $\hat{r}_t$ de la régression x_{t1} sur $x_{t2}, \dots, x_{tk}$

(iii) Former $\hat{a}_t = \hat{r}_t \hat{u}_t$

(iv) Calculer se($\hat{\beta}$)

Empiriquement, l'écart-type robuste à l'hétéroscédasticité est plus grand.

Exemple

La commande dans Stata est `newey` :

```
newey y [x] [if] [in] [weight] , lag(#)
```

L'option `lag` permet d'inclure le nombre d'ordres autorégressifs que nous avons trouvé autocorrélés précédemment.

```
newey usr idle, lag(3)

Régression with Newey-West standard errors          Number of obs  =        30
maximum lag: 3                                      F(  1,    28)  =     10.90
                                                    Prob > F       =    0.0026
```

```
------------------------------------------------------------------------------
             |             Newey-West
         usr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        idle |  -.2281501   .0690927    -3.30   0.003    -.3696801     -.08662
       _cons |   23.13483   6.327031     3.66   0.001     10.17449    36.09516
------------------------------------------------------------------------------
```

La méthode Newey-West permet une interprétation des coefficients estimations tandis que les méthodes des MCQG, soit la méthode Prais-Winsten et celle de Cochrane-Orcutt, ne le permet pas

7.14 Hétéroscédasticité dans les séries temporelles

Tout comme dans la section concernant les données en coupe transversale (voir exemple détaillé dans la partie régression linéaire), nous pouvons tester l'homoscédasticité avec le test Breusch-Pagan, soit la fonction `bpagan` ou manuellement par les étapes suivantes :

1. Récupérer les résidus de la régression qu'on désire tester;
2. Générer le carré des résidus;
3. Régresser le carré des résidus sur les variables indépendantes de la régression originale;
4. Tester si les coefficients sont conjointement significatifs (test F ou test LM).

Vérifier si la statistique de Fisher dans la sortie de Stata est significative ou non. Si nous rejetons l'hypothèse nulle, il y a hétéroscédasticité.

Pour corriger l'hétéroscédasticité, nous pouvons utiliser l'option `robust` dans Stata ou le modèle MCQG AR(1) avec la fonction prais ou les écarts-types Newey-West.

7.15 Autoregressive Conditional Heteroskedasticity

Même si la variance des résidus est constante en fonction des résidus, il est possible d'avoir d'autre forme d'hétéroscédasticité. Le modèle ARCH conditionne la variance des résidus sur ses valeurs passées et sur les régresseurs.

Étape pour tester un ARCH (1)

1. Estimer la régression avec la commande reg
2. Prédire les résidus et les élever au carré
3. Régressez les résidus au carré sur les résidus avec un retard d'une période (`l.variable`)
4. Testez si le coefficient est significatif

5. Si le coefficient est significatif, il y a présence d'hétéroscedasticité conditionnelle

Vous pouvez aussi utiliser la fonction `estat archlm` dans Stata

```
estat archlm, archlm_options
```

Exemple

```
webuse klein, clear

.
.  tsset yr
        time variable:  yr, 1920 to 1941
                delta:  1 unit

.
.  regress consump wagegovt

      Source |       SS       df       MS              Number of obs =      22
-------------+------------------------------           F(  1,    20) =   17.72
       Model |  532.567711     1  532.567711           Prob > F      =  0.0004
    Residual |  601.207167    20  30.0603584           R-squared     =  0.4697
-------------+------------------------------           Adj R-squared =  0.4432
       Total |  1133.77488    21  53.9892799           Root MSE      =  5.4827

------------------------------------------------------------------------------
     consump |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    wagegovt |    2.50744   .5957173     4.21   0.000     1.264796    3.750085
       _cons |   40.84699   3.192183    12.80   0.000     34.18821    47.50577
------------------------------------------------------------------------------

.
.  estat archlm, lags(1 2 3)
LM test for autoregressive conditional heteroskedasticity (ARCH)
-----------------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+---------------------------------------------------------------
       1     |          5.543               1                   0.0186
       2     |          9.431               2                   0.0090
       3     |          9.039               3                   0.0288
-----------------------------------------------------------------------------
         H0: no ARCH effects      vs.  H1: ARCH(p) disturbance
```

Le test nous permet de voir que nous pouvons rejeter l'absence d'effet ARCH d'ordre 1, 2, 3.

Si vous êtes en présence d'ARCH utilisez les méthodes MCQG ou les écarts types Newey-West.

8 Modèle de discrimination

8.1 Modèle Oaxaca-Blinder

Le modèle Oaxaca-Blinder est un modèle de décomposition des différences moyennes entre deux différents groupes. La méthode est utilisée afin d'identifier et quantifier les contributions des différences de groupes dans les régresseurs. L'idée du modèle est d'expliquer la distribution de la variable réalisée par des facteurs socio-économiques. Le modèle de décomposition de Oaxaca-Blinder est souvent utilisé en économie du travail pour déterminer la discrimination salariale.

Le modèle divise la différence de la variable endogène entre 2 groupes, la partie expliquée et non expliquée qui est souvent prise pour une mesure de discrimination. Le modèle repose entre autres sur l'hypothèse de groupe mutuellement exclusif.

Le modèle de décomposition Oaxaca-Blinder a quelques limites, entre autres il découle d'une approche d'équilibre partiel. Le modèle est aussi limité par le fait que nous ne pouvons pas faire d'interprétation de causalité.

Étant donné 2 groupes, soit A et B, le modèle Oaxaca-Blinder cherche la différence moyenne de la variable endogène entre ces groupes.

$$R = E(Y_A) - E(Y_B)$$

$$R = E(Y_A) - E(Y_B) = E(X_A)'\beta_A - E(X_B)'\beta_B$$

Le modèle prend la différence de l'espérance des variables endogènes et exogènes. La fonction d'Oaxaca-Blinder de Stata estime par défaut le modèle suivant, le modèle *three-fold* :

$$R = [E(X_A) - E(X_B)]'\beta_B + E(X_B)'(\beta_A - \beta_B) + [E(X_A) - E(X_B)]'(\beta_A - \beta_B)$$

$$R = E + C + I$$

où la différence de Y dans les groupes est décomposée par l'effet de dotation, E, la contribution des différences dans les coefficients, C, et un terme d'interaction qui tente de prendre en compte que E et C existe simultanément dans les deux groupes. La décomposition est prise en compte par rapport au groupe B mais pourrait aussi être par rapport à A.

Le modèle d'Oaxaca-Blinder permet aussi une décomposition *two-fold*, soit

$$R = [E(X_A) - E(X_B)]' \beta^* + [E(X_B)' (\beta_A - \beta^*) + E(X_B)' (\beta^* - \beta_B)]$$

$$R = Q + U$$

où Q est la partie de la différence de Y qui est expliquée par la différence des coefficients entre les groupes, l'effet de quantité, et U est la partie non expliquée, qui comprend toutes les variables non observées. Cette partie est souvent attachée à la discrimination.

Il peut être important de pondérer les coefficients en fonction du nombre d'observations dans chacun des groupes.

Commande Stata

```
oaxaca y X if in weight, by (groupvar) options
```

Il est important de toujours bien spécifier par quel groupe nous voulons décomposer.

```
findit oaxaca
```

Choisir le module SJ8-4 st0151

Installer le fichier st0151/oaxaca.dta

```
use oaxaca.dta
(Excerpt from the Swiss Labor Market Survey 1998)

oaxaca lnwage educ exper tenure, by(female)

Blinder-Oaxaca decomposition                      Number of obs   =       1434

           1: female = 0
           2: female = 1

------------------------------------------------------------------------------
       lnwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
Differential  |
 Prediction_1 |   3.440222   .0174874   196.73   0.000     3.405947    3.474497
 Prediction_2 |   3.266761   .0218522   149.49   0.000     3.223932    3.309591
   Difference |   .1734607    .027988     6.20   0.000     .1186052    .2283163
--------------+---------------------------------------------------------------
Decomposition |
   Endowments |   .0852798    .015693     5.43   0.000     .0545222    .1160375
 Coefficients |    .082563   .0255804     3.23   0.001     .0324263    .1326996
  Interaction |    .005618    .010966     0.51   0.608    -.0158749    .0271109
------------------------------------------------------------------------------
```

Dans l'exemple ci-dessus, nous avons entrepris l'estimation du modèle de décomposition de la différence moyenne du taux horaire des hommes et des femmes en fonction des paramètres : éducation (educ), expérience (exper) et le nombre d'année de permanence (tenure). Le coefficient de « Prediction_1 » est la moyenne du taux horaire en log naturel des hommes et « Prediction_2 » est la moyenne des femmes tandis que le coefficient .1734607 est la différence entre les deux. La deuxième partie du tableau est la décomposition de la différence en fonction de l'effet de dotation, de la différence des coefficients et de l'interaction.

Mise en garde

Comme nous avons mentionné, le terme d'interaction est souvent interprété comme valeur ou preuve de la discrimination. Cependant, si le modèle sous-jacent n'inclut pas toutes les variables explicatives, conclure présence de discrimination serait incorrect. Ceci veut dire que si vous n'incluez pas toutes les variables pouvant expliquer le taux horaire, l'effet de ces variables serait dans le terme d'interaction et le terme d'interaction ne représenterait pas uniquement la discrimination.

Option

`noisily` : l'option `noisily` peut être rajoutée afin de voir les régressions faites séparément pour les hommes et les femmes

```
oaxaca lnwage educ exper tenure, by(female) noisily

Model for group 1

      Source |       SS       df       MS              Number of obs =     751
-------------+------------------------------           F(  3,   747) =  101.14
       Model |   49.613308     3  16.5377693           Prob > F      =  0.0000
    Residual |  122.143834   747  .163512495           R-squared     =  0.2889
-------------+------------------------------           Adj R-squared =  0.2860
       Total |  171.757142   750  .229009522           Root MSE      =  .40437

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0820549   .0060851    13.48   0.000      .070109    .0940008
       exper |   .0098347   .0016665     5.90   0.000     .0065632    .0131062
      tenure |   .0100314   .0020397     4.92   0.000     .0060272    .0140356
       _cons |    2.24205   .0778703    28.79   0.000      2.08918    2.394921
------------------------------------------------------------------------------

Model for group 2

      Source |       SS       df       MS              Number of obs =     683
-------------+------------------------------           F(  3,   679) =   40.34
       Model |  33.5197344     3  11.1732448           Prob > F      =  0.0000
    Residual |   188.08041   679  .276996185           R-squared     =  0.1513
-------------+------------------------------           Adj R-squared =  0.1475
       Total |  221.600144   682  .324926897           Root MSE      =   .5263
```

```
------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0877579   .0087108    10.07   0.000     .0706546    .1048611
       exper |   .0131074   .0028971     4.52   0.000     .0074191    .0187958
      tenure |   .0036577   .0035374     1.03   0.301    -.0032878    .0106032
       _cons |   2.097806   .1091691    19.22   0.000     1.883457    2.312156
------------------------------------------------------------------------------

Blinder-Oaxaca decomposition                      Number of obs   =       1434

           1: female = 0
           2: female = 1

------------------------------------------------------------------------------
       lnwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
Differential  |
 Prediction_1 |   3.440222   .0174874   196.73   0.000     3.405947    3.474497
 Prediction_2 |   3.266761   .0218522   149.49   0.000     3.223932    3.309591
   Difference |   .1734607    .027988     6.20   0.000     .1186052    .2283163
--------------+---------------------------------------------------------------
Decomposition |
   Endowments |   .0852798    .015693     5.43   0.000     .0545222    .1160375
 Coefficients |    .082563   .0255804     3.23   0.001     .0324263    .1326996
  Interaction |    .005618    .010966     0.51   0.608    -.0158749    .0271109
------------------------------------------------------------------------------
```

`detail` : l'option `détail` subdivise les effets de dotation, de coefficient et d'interaction en fonction des paramètres

```
oaxaca lnwage educ exper tenure, by(female) detail

Blinder-Oaxaca decomposition                      Number of obs   =       1434

           1: female = 0
           2: female = 1

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Differential |
Prediction_1 |   3.440222   .0174874   196.73   0.000     3.405947    3.474497
Prediction_2 |   3.266761   .0218522   149.49   0.000     3.223932    3.309591
  Difference |   .1734607    .027988     6.20   0.000     .1186052    .2283163
-------------+----------------------------------------------------------------
Endowments   |
        educ |   .0510912    .012239     4.17   0.000     .0271031    .0750792
       exper |   .0254173   .0088089     2.89   0.004     .0081522    .0426824
      tenure |   .0087714   .0086201     1.02   0.309    -.0081238    .0256665
       Total |   .0852798    .015693     5.43   0.000     .0545222    .1160375
-------------+----------------------------------------------------------------
Coefficients |
        educ |  -.0640559   .1193498    -0.54   0.591    -.2979772    .1698654
       exper |  -.0397237     .04058    -0.98   0.328    -.1192591    .0398117
      tenure |   .0420986   .0270201     1.56   0.119    -.0108598    .0950571
       _cons |   .1442439   .1340957     1.08   0.282    -.1185788    .4070667
       Total |    .082563   .0255804     3.23   0.001     .0324263    .1326996
-------------+----------------------------------------------------------------
Interaction  |
        educ |  -.0033202   .0062283    -0.53   0.594    -.0155274    .0088871
       exper |  -.0063464   .0066988    -0.95   0.343    -.0194758     .006783
      tenure |   .0152845   .0101496     1.51   0.132    -.0046084    .0351774
       Total |    .005618    .010966     0.51   0.608    -.0158749    .0271109
------------------------------------------------------------------------------
```

`threefold(reverse)` : l'option `threefold(reverse)` substitue les groupes. La décomposition est exprimée à partir du groupe B à la place du groupe A

```
oaxaca lnwage educ exper tenure, by(female) threefold(reverse)

Blinder-Oaxaca decomposition                      Number of obs   =       1434

           1: female = 0
           2: female = 1

------------------------------------------------------------------------------
       lnwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
Differential  |
 Prediction_1 |   3.440222   .0174874   196.73   0.000     3.405947    3.474497
 Prediction_2 |   3.266761   .0218522   149.49   0.000     3.223932    3.309591
   Difference |   .1734607    .027988     6.20   0.000     .1186052    .2283163
--------------+---------------------------------------------------------------
Decomposition |
   Endowments |   .0908978   .0136196     6.67   0.000     .0642038    .1175917
 Coefficients |   .0881809    .026692     3.30   0.001     .0358655    .1404963
  Interaction |   -.005618    .010966    -0.51   0.608    -.0271109    .0158749
------------------------------------------------------------------------------
```

`weight` : l'option de poids est utilisée pour pondérer en fonction du nombre d'observation. L'option génère la décomposition *two-fold* et pondère de la façon suivante

$$\hat{\beta}^* = \frac{n_A}{n_A + n_B}\hat{\beta}_A + \frac{n_B}{n_A + n_B}\hat{\beta}_B$$

Cette option est importante lorsque le nombre d'observations dans les groupes est grandement différent.

```
oaxaca lnwage educ exper tenure, by(female) weight(1)

Blinder-Oaxaca decomposition                      Number of obs   =       1434

           1: female = 0
           2: female = 1

------------------------------------------------------------------------------
       lnwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
Differential  |
 Prediction_1 |   3.440222   .0174874   196.73   0.000     3.405947    3.474497
 Prediction_2 |   3.266761   .0218522   149.49   0.000     3.223932    3.309591
   Difference |   .1734607    .027988     6.20   0.000     .1186052    .2283163
--------------+---------------------------------------------------------------
Decomposition |
    Explained |   .0908978   .0136196     6.67   0.000     .0642038    .1175917
  Unexplained |    .082563   .0255804     3.23   0.001     .0324263    .1326996
------------------------------------------------------------------------------
```

`pooled` : l'option `pooled` génère la décomposition *two-fold* sans pondération. Cette option décompose seulement entre expliqué et non expliqué et ne différentie pas entre les divers effets de la décomposition *three-fold*. Le *two-fold* inclut la variable groupe comme variable de contrôle

```
oaxaca lnwage educ exper tenure, by(female) pooled

Blinder-Oaxaca decomposition                      Number of obs   =       1434

          1: female = 0
          2: female = 1

------------------------------------------------------------------------------
              |               Robust
       lnwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
Differential  |
 Prediction_1 |   3.440222   .0174586   197.05   0.000     3.406004     3.47444
 Prediction_2 |   3.266761   .0218042   149.82   0.000     3.224026    3.309497
   Difference |   .1734607   .0279325     6.21   0.000      .118714    .2282075
--------------+---------------------------------------------------------------
Decomposition |
    Explained |    .089347   .0137531     6.50   0.000     .0623915    .1163026
  Unexplained |   .0841137    .025333     3.32   0.001      .034462    .1337654
------------------------------------------------------------------------------
```

Comme vous pouvez voir, les coefficients de la décomposition sont différents de ceux utilisant l'option `weight`.

`omega` : l'option `omega` diffère seulement de l'option `pooled` par le fait qu'elle exclut la variable de groupe comme variable de contrôle et donc a un effet sur les coefficients décomposés.

```
oaxaca lnwage educ exper tenure, by(female) omega

Blinder-Oaxaca decomposition                      Number of obs   =       1434

          1: female = 0
          2: female = 1

------------------------------------------------------------------------------
              |               Robust
       lnwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
Differential  |
 Prediction_1 |   3.440222   .0174586   197.05   0.000     3.406004     3.47444
 Prediction_2 |   3.266761   .0218042   149.82   0.000     3.224026    3.309497
   Difference |   .1734607   .0279325     6.21   0.000      .118714    .2282075
--------------+---------------------------------------------------------------
Decomposition |
    Explained |   .0925597   .0140752     6.58   0.000     .0649728    .1201466
  Unexplained |    .080901   .0243589     3.32   0.001     .0331585    .1286435
------------------------------------------------------------------------------
```

Autres options importantes : les options `vce` pour l'hétéroscédasticité et `cluster` peuvent être ajoutées au modèle.

Mise en garde

`svy` : l'option `svy` est utilisée lorsque nous avons des données d'enquête. Il faut en premier indiquer le poids et ensuite inclure `svy` dans la fonction.

Lorsque nous travaillons avec des données d'enquête, il est suggéré d'utiliser l'option qui indique le poids d'échantillonnage.

```
svyset [pw=wt]

      pweight: wt
          VCE: linearized
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>

. oaxaca lnwage educ exper tenure, by(female) pooled svy

Blinder-Oaxaca decomposition

Number of strata   =         1                  Number of obs      =      1647
Number of PSUs     =      1647                  Population size    = 1657.1804
                                                 Design df          =      1646

           1: female = 0
           2: female = 1

------------------------------------------------------------------------------
              |             Linearized
       lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
Differential  |
 Prediction_1 |   3.405696   .0226311   150.49   0.000     3.361307    3.450085
 Prediction_2 |   3.193847   .0276463   115.53   0.000     3.139622    3.248073
   Difference |   .2118488    .035728     5.93   0.000     .1417718    .2819259
--------------+---------------------------------------------------------------
Decomposition |
    Explained |   .1107614   .0189967     5.83   0.000     .0735011    .1480216
  Unexplained |   .1010875   .0315911     3.20   0.001     .0391246    .1630504
------------------------------------------------------------------------------
```

Bibliographie

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach* (éd. 2). Springer-Verlag.

Cameron, A. C., & Trivedi, P. K. (2009). *Microeconometrics using Stata.* College Station: Stata Press.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications.* New York: Cambridge University Press.

Greene, W. (2012). *Econometric Analysis* (éd. Seventh edition). Upper Saddle River: Pearson Education.

Kohavi, R., & Provost, F. (1998). On Applied Research in Machine Learning. *Special Issue on Applications of Machine Learning and the Knowledge Discovery Process* .

Lemeshow, S., & Hosmer, D. W. (1982). A review of Goodness of fit statistics for use une the developpement of logistic regression model. *American Journal of Epidemiology* , 92-106.

McFadden, D. (1974). The Measurement of Urban Travel Demand. *Journal of Public Economics* , 303-328.

Rakotomalala, R. (2009). *Pratique de la Régression Logistique, Regression Logistique Binaire et Polytomique.* non publié.

Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. Dans *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg* (pp. 80-108). New-York: Cambridge University Press.

Wooldridge, J. M. (2001). *Econometric analysis of cross section and panel data, second edition.* Cambridge: MIT press.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data, Second Edition.* Cambridge: MIT Press .

Wooldridge, J. M. (2009). *Introductory Econometrics : A Modern Approach, fourth edition.* Mason: South-Western.