Université de Montréal

Guide d'économétrie appliquée pour Stata
Pour
ECN 3950 et FAS 3900

août 2005

par
Estelle Ouellet

avec l'apport de
Isabelle Belley-Ferris
et
Simon Leblond

Université de Montréal

Merci à Ghislaine Geoffrion, Linda Lee et François Vaillancourt pour leurs précieux conseils économétriques.

Table des matières

Préface

Ceci est la troisième version d'un guide d'économétrie appliqué à Stata créé pour aider les étudiants dans leur cours de FAS 3900 (séminaire d'économie politique) ou d'ECN 3950 (Atelier d'économie appliquée). La première version a été élaborée par Simon Leblond en décembre 2003. Les renseignements contenus dans cette première ébauche correspondent en partie à ce qui est inscrit dans la troisième section du présent manuel. Un an plus tard, Isabelle Belley-Ferris a ajouté une section descriptive pour compléter le travail fait par Simon. Cette section a aussi été intégrée dans la présente version du manuel, à la section 3.

En somme, les sections sur l'extraction des données (section 1) et sur le traitement des données (section 2) du présent manuel sont inédites, alors que l'on retrouve quelques ajouts au travail fait précédemment par Simon Leblond et Isabelle Belley-Ferris, correspondant à la dernière section du guide.

Introduction

Ce guide vous servira d'outil de référence tout au long de la session. Nous avons tenté de rassembler toute la matière essentielle à la réussite de votre cours (FAS 3900 ou ECN 3950) dans ce manuel. Cela dit, il se peut que pour certains, des sections de ce manuel contiennent des notions triviales ou trop avancées en fonction de vos objectifs de recherche. Il vous suffira donc de sauter à la prochaine section plus rapidement.

Chaque section présente le but de l'opération qui y est traitée. Les commandes appropriées sont ensuite présentées, d'abord individuellement, puis dans le cadre d'un exemple concret. Prenez note que ce texte décrit seulement les fonctions ainsi que leurs options les plus souvent utilisées pour votre cours, il n'est donc pas du tout exhaustif. Un conseil : apprenez à utiliser l'aide de Stata. Il s'agit d'un outil fort utile pour découvrir de nouvelles fonctions ou pour connaître l'ensemble des options disponibles pour les fonctions décrites dans ce guide.

Le chapitre 1 vous indiquera comment trouver les données d'enquête dont vous aurez besoin pour réaliser votre recherche. Le chapitre 2 s'adresse plus particulièrement aux élèves de FAS 3900, décrivant la base des manipulations économétriques. Le chapitre 3 s'adresse a priori aux étudiants de ECN 3950, exposant des manipulations plus poussées. Il abordera des sujets spécifiques de l'économétrie. Il introduit peu de nouvelles fonctions, se concentrant plutôt sur la démarche à adopter pour effectuer l'opération en question.

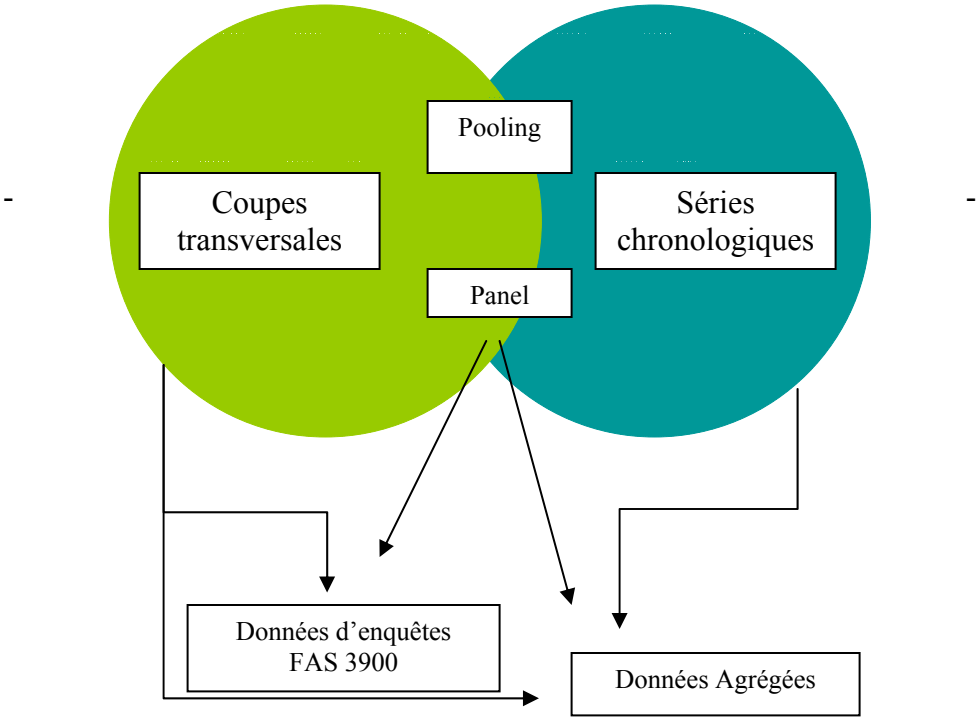
1 Extraction de données

1.1 Les données

Il existe deux types de fichiers de données (d'enquête et agrégées) à partir desquels sont faites les manipulations économétriques qui permettront d'estimer la valeur du lien entre deux variables. Les données d'enquête sont des données brutes, pratiquement illisibles sans le fichier de documentation (ou cliché d'enregistrement). Le fichier contient les réponses du répondant qui sont codés numériquement (ex. : Le recensement). Par exemple, voici un extrait d'un cliché d'enregistrement :

SEX	Sex of respondent		Sexe du répondant
	Male	1	Hommes
	Female	2	Femmes
MARSTAT	Marital status of respondent		État matrimonial du répondant
	Married or common-law	1	Mariés ou vivant en union libre
	Single, never married	2	Célibataires, n'ont jamais été mariés
	Widowed	3	Veufs ou veuves
	Separated/divorced	4	Séparés/divorcés

La première colonne nous donne le nom de la variable, les troisième et quatrième donnent le numéro correspondant à la variable. Donc, si un répondant répond à la question « sexe » qu'il est un homme, le numéro 1 apparaîtra dans le fichier de données. De la même façon, à la question sur le statut matrimonial, si le répondant affirme qu'il est veuf, c'est le numéro 3 qui apparaîtra dans le fichier de données. C'est ce type de données qui devra être utilisé pour le cours FAS 3900. Les données agrégées sont des données qui ont subit un traitement statistique. En examinant ce type de fichier, on comprend ce que les données signifient parce qu'elles ont été manipulées (ex. : Les PIB provinciaux). Ces deux types de fichier de données peuvent être structurés de façons différentes, et permettent de réaliser différents types d'analyses économétriques (voir schéma ci-dessous).



On associe souvent les données structurées en **coupe transversale** à l'analyse statique dans les domaines de la microéconomie (i.e. économie du travail, finance publiques provinciales ou municipales, organisation industrielle, etc.). Les données sur les individus, ménages, compagnies, villes, etc. à un point donnée dans le temps sont les plus utilisées pour les études microéconométriques. La structure en coupe transversale devra être utilisé par les étudiants de FAS 3900, puisque les autres types de structures de

données privilégient les données agrégées (séries chronologiques) ou nécessitent des connaissances plus poussées en économétrie (pooling et panel).

Les séries chronologiques sont généralement utilisées lors d'études relevant du domaine de la macroéconomie (i.e. indice des prix à la consommation, produit intérieur brut, vente annuelle de voiture dans l'industrie automobile, etc.). Ce type de bases de données est donc composé de données agrégées et est privilégiée pour faire des études macroéconométriques (généralement des prévisions).

Finalement, les structures appelées **pooling** et **panel**, comportent les caractéristiques des structures de données précédentes. Le pooling a pour but de comparer l'évolution de la relation entre un échantillon et une caractéristique clé à travers le temps (ce type de base de données est très fréquemment utilisé pour évaluer l'impact d'une politique publique sur un échantillon). Le panel est très semblable au pooling, mais la différence réside dans le fait que les unités de l'échantillon restent les mêmes à travers le temps.

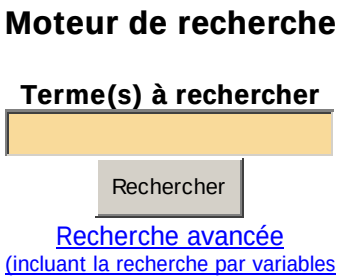
1.2 Les divers systèmes d'extraction de données

Pour débuter votre régression, vous aurez besoin de trouver vos données. Il existe plusieurs systèmes d'extraction de données, nous vous en présentons trois. Le premier, Sherlock, regroupe des fichiers de micro-données provenant d'enquêtes faites au Canada et au Québec. Dans le second, Cansim, on retrouve les données statistiques agrégées recensées par Statistiques Canada. Finalement, le troisième, ICSPR, couvre des enquêtes provenant des Etats-Unis. Dans chacune des sections du présent chapitre, vous retrouverez un lien avec le site, une image du moteur de recherche provenant du site ainsi que la façons de l'utiliser.

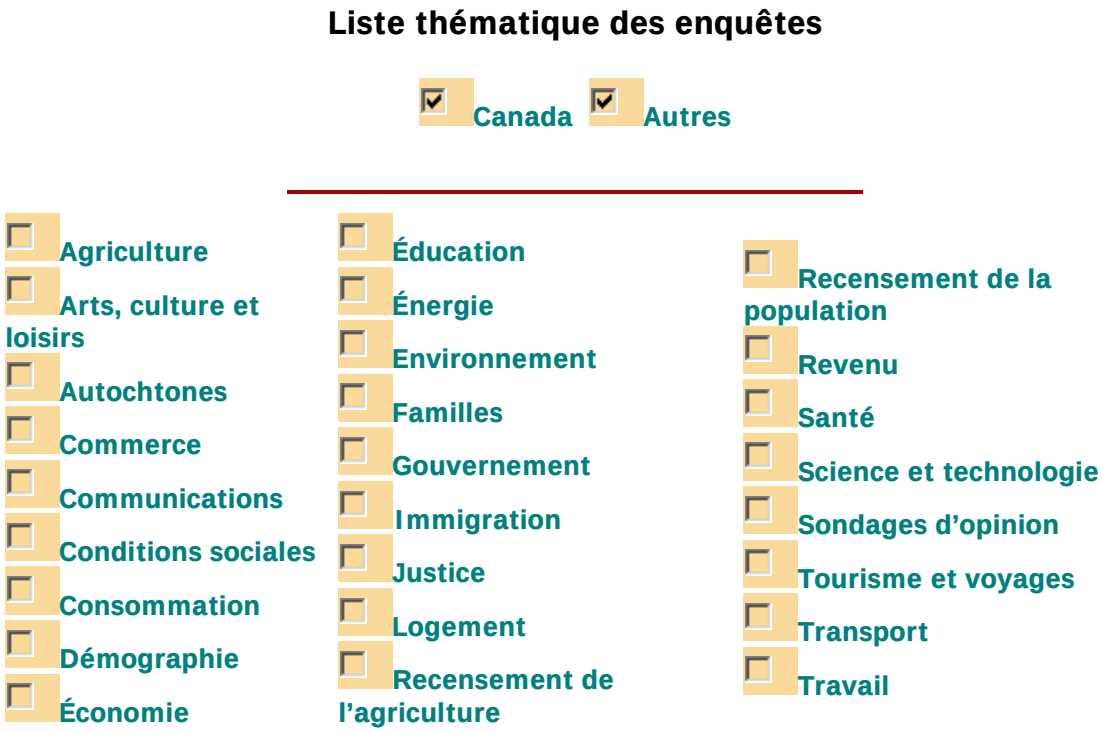
1.2.1 Sherlock

http://sherlock.crepuq.qc.ca/

Avec Sherlock, trois façons de faire votre recherche se présentent à vous. Vous pouvez soit faire une recherche à l'aide de la « liste d'enquête » dont l'hyperlien se trouve en haut de la page d'accueil, soit faire une recherche simple ou avancée avec le moteur de recherche ci-dessous :

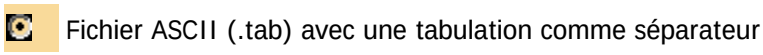


Ou chercher selon des thèmes en cochant la case approprié :



Lorsque que vous avez trouvé l'enquête désirée, deux documents sont à sélectionner. Le premier est le cliché d'enregistrement. Celui-ci vous donne le nom de chaque variable ou groupe de variable qui compose votre fichier de micro-données. Cela vous permettra d'écrire vos commandes dans Stata. Le second document est le fichier de micro-données en tant que tel. Pour obtenir ces documents, vous cliquez sur « cliché d'enregistrement » dans la section « documentation sur les données d'enquête » et sur « extraction » dans la section « accès aux données.

La deuxième manœuvre vous mène aux procédures d'extraction. Dans la première page qui apparaît après avoir cliqué sur « extraction », cliquez sur :



Ensuite, vous continuez la procédure d'extraction. Une liste de variables apparaîtra.[1] Voilà donc la première occasion d'utiliser le cliché d'enregistrement. Il est extrêmement important de bien le lire et de devenir familier avec les diverses variables qui composent le fichier, cela vous évitera des démarches inutiles et bien des erreurs puisque vous devez ensuite sélectionner vos variables. Il est important de bien réfléchir lors de la sélection, parce que si vous en oubliez, vous devrez tout recommencer du début. Vaut mieux donc en choisir plus que moins. Lorsque les variables désirées sont cochées, il ne vous reste qu'à écrire votre courriel au bas de la page. Dans les minutes qui

[1] Il se peut que la liste des variables n'apparaisse pas. Vous devrez alors télécharger le fichier de micro-données au complet. C'est en utilisant SPSS et statTransfert que vous pourrez sélectionner vos variables. Si vous avez des questions à ce sujet, Pascal Martinolli est familier avec cette procédure.

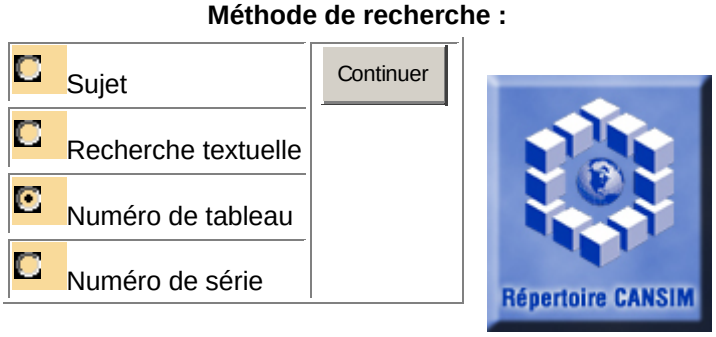
suivent, si tout va bien (il arrive que vous deviez refaire une procédure d'extraction parce que Sherlock n'est pas parfait !) vous recevrez un courriel dans votre boîte de réception vous donnant le l'hyperlien à vos données.

Note : Il y a un maximum des 30 variables pour une extraction. Si vous pensez avoir besoin de plus de variable pour votre régression, il faudra refaire plus d'une extraction. Ensuite, il est important de s'assurer que les données de chaque extraction concordent (que les réponses soient attribuées au bon répondant), et utiliser la fonction *merge* (*merge nomdesvariables*) dans Stata pour fusionner les diverses extractions pour avoir une base de données complète.

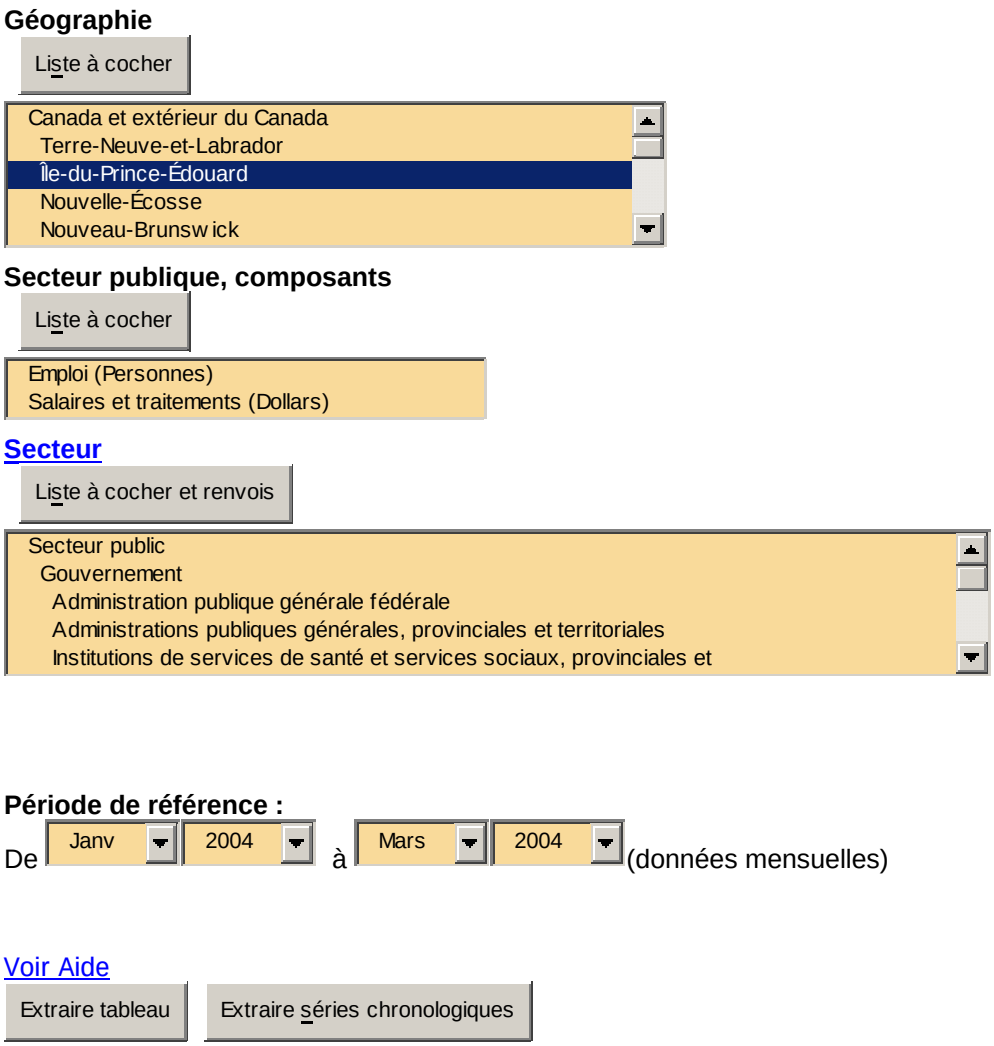
1.2.2 Cansim

À partir du site de la bibliothèque des sciences humaines de l'Université de Montréal, cliquez sur **Cansim II dans E-Stat** ou utilisez http://estat.statcan.ca/cgi-win/CNSMCGI.exe?Lang=F&CANSIMFile=EStat/Francais/CII_1_F.htm

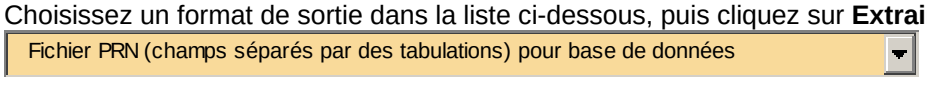
Pour faire une recherche dans Cansim, vous pouvez soit consulter le répertoire et ensuite entrer le numéro de tableau tel qu'indiqué ci-dessous :



Vous pouvez aussi, comme avec Sherlock, faire une recherche par thème (sujet) ou faire une recherche textuelle. Lorsque l'enquête est sélectionnée, un page comme celle-ci apparaît :



Faites les sélections nécessaires, et ensuite, cliquez sur Extraire tableau. La page suivante vous demandera de choisir un format de sortie. Sélectionnez :



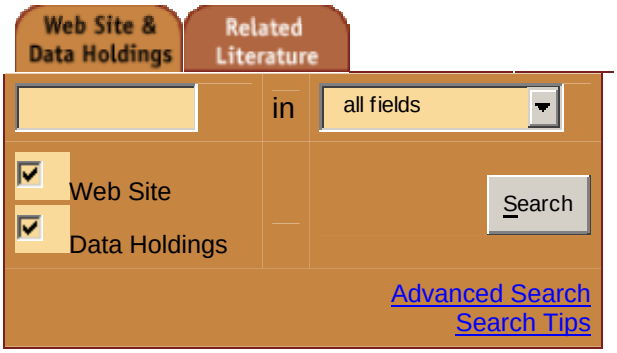
Un fichier bloc-notes (notepad) apparaîtra.

Vous pouvez aussi utiliser la ressource de l'Université de Toronto, qui contiens plus de bases de données et qui est mise à jour plus fréquemment que celle de E-Stat. Pour ce faire, utilisez http://dc2.chass.utoronto.ca/cansim2/French/. Vous pouvez aussi y accéder par l'entremise du site de la bibliothèque des sciences humaines de l'université de Montréal.

1.2.3 ICSPR

http://www.icpsr.umich.edu/

Le site de l'Inter-University Consortium for Political and Social Reaserch est une autre source très pratique. Lorsque vous arrivez à la page principale du site, vous pouvez soit cliquer le lien « Data Use Tutorial » pour vous familiariser avec le site ou cliquer sur « Data Access and Analysis » pour débuter la recherche de données directement. Une fois rendu à cette étape, comme pour Sherlock, deux façons de faire votre recherche se présentent à vous. Vous pouvez soit faire une recherche simple ou avancée avec le moteur de recherche ci-dessous :



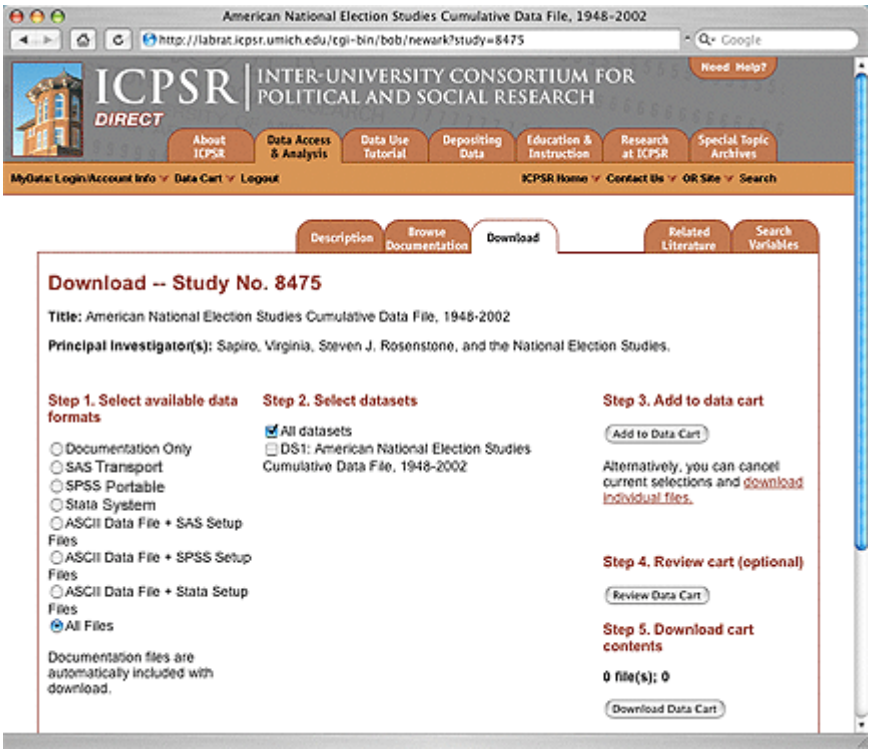
Ou chercher selon des thèmes en sélectionnant l'hyperlien approprié :

Other Methods of Finding Data

Bibliography of Data-Related Literature
Browse/Search Data by Subject
Data Available on CD-ROM
Popular Data (top 25 studies)
Publication-Related Archive
Recent Updates & Additions
Series Data
Social Science Variables Database
Special Topic Archives

Search Tips

Lisez la description de l'enquête (un conseil : dans la description, allez voir les « subject terms », peut-être que ces hyperliens vous mèneront à une enquête cernant mieux votre sujet de recherche). Si l'enquête semble convenir, cliquez sur « download » et une fenêtre comme celle-ci apparaîtra :



L'étape 1 requiert que vous choisissiez le format dans lequel le fichier de données apparaîtra. Si Stata n'apparaît pas dans la sélection, utilisez SPSS et la section 2.2 du présent chapitre vous expliquera comment formater le fichier pour que Stata puisse le lire. L'étape 2 consiste en la sélection des variables dont vous avez besoin pour votre enquête. À l'étape 3, vous ajoutez le fichier à votre « panier ». La quatrième étape est téléchargement du fichier. Le fichier qui apparaîtra est un fichier zip qu'il faudra décompresser pour que Stata puisse le lire.

Note : Tout comme l'ICSPR, l'American Community Survey est un bon système d'extraction pour les fichiers de données d'enquêtes. Pour sa part, le U.S. Census Bureau est une bonne source pour obtenir des données agrégées. Vous pouvez par ailleurs trouver d'autres systèmes d'extraction sur la page des données numériques de la bibliothèque des lettres et sciences humaines de l'Université de Montréal (http://www.bib.umontreal.ca/SS/num/). Si vous avez des problèmes reliés à ces systèmes d'extraction, adressez-vous à Maryna Beaulieu du Centre de données numériques et géospatiales de la bibliothèque des lettres et sciences humaines de l'Université de Montréal. Maryna.beaulieu@umontréal.ca ou (514) 343-6111, ext. 0994.

1.3 Convertir des fichiers pour Stata

Dans le cadre de cette section, nous vous expliquerons les solutions aux problèmes les plus courants. C'est-à-dire, le cas des fichiers extraits sans sélection parmi les variables (fichiers complets), les fichier formatés pour SPSS, et les fichiers exportés à partir de Cansim.

En ce qui concerne les deux premier cas, il suffit d'utiliser le programme StatTransfert qui se trouve sur les ordinateurs du local C-3001 (où vous faites vos TP) du pavillon Lionel-Groulx.

Dans le dernier cas, la démarche est un peu plus longue. Tout d'abord, tel que mentionné précédemment, vous avec sélectionné un format de fichier PRN pour base de données. Cela sort en bloc note (format .txt.). Afin de rendre ce fichier bloc note lisible pour Stata (puisque le fichier contient du texte et que Stata ne peut lire les textes), vous devez importer ce fichier .txt dans Excel. Pour ce faire, ouvrez Excel et allez à *données*, *données externes* puis *importer des données*. Suivez les indications, et quand le processus est terminé, vous aurez un fichier de données manipulable grâce à Excel. Enlevez les colonnes de texte et remplacez les virgules par des points s'il y a lieu, puisque Stata ne peut les lire. Lorsque le fichier de données est adéquatement modifié pour Stata, enregistré le tout en format txt. Vous avez maintenant un fichier lisible pour Stata à partir d'un fichier provenant de Cansim.

Note : Il arrive qu'il y ait des données manquantes dans les fichiers de données. Dans le cadre de **données d'enquête**, un répondant peut refuser de répondre à une question. Dans ce cas, on retrouve un espace vide dans la base de donnée, ce qui peut venir fausser les résultats de manipulations. Il faut remplacer ces espaces par des points, Stata ne pouvant pas lire les espaces vides. Allez dans le bloc note et faites Ctrl-H. En ce les fichiers de **données agrégées,** deux options s'offrent à vous pour combler les espaces vides. Vous pouvez soit faire des moyennes mobiles, en utilisant la valeur avant et la valeur après la donnée manquante (par exemple, s'il manque le PIB pour l'année 2003, additionnez celui de 2002 à celui de 2004, et divisez le en deux. Cela donne la moyenne

mobile pour 2003.) Une idée de la valeur manquante peut aussi vous être donnée en faisant un graphique. Or, puisqu'il manque des données, Stata ne pourra produire le graphique, vous devrez donc aller dans Excel…

2 Le traitement des données

2.1 Rappel de notions théoriques d'économétrie

2.1.1 Qu'est-ce que l'économétrie ?

Les régressions sont des outils qui permettent, entre autre, d'estimer l'effet marginal de la variation d'une unité de la variable indépendante sur la variable dépendante. On peut, par exemple, tester des théories économiques, évaluer l'impact d'une politique publique sur un échantillon de population ou même de faire des prévisions…

Pour faire une régression, il faut que tous les autres facteurs (d'autres variables indépendantes) pouvant influencer la variable dépendante soient maintenus constants. Leur effet potentiel sur la variable dépendante pourrait être capté par la variable indépendante d'intérêt et ainsi être à la source d'une augmentation (ou diminution) marginale sur la variable dépendante. Même si cela est quasi-impossible, il faut tenter de contrôler pour le maximum de variables indépendantes pertinentes (l'ajout de variables réduit le nombre de degrés de liberté[2]) afin d'être certain de la validité du lien de causalité prédit entre la variable dépendante et la variable indépendante d'intérêt.

2.1.2 La différence entre un estimateur non-biaisé et efficace, et une variable significative

Pour pouvoir assumer que les coefficients de la MCO sont non-biaisés, c'est-à-dire que la valeur prédite par l'estimateur converge vers la valeur dans la population, on

[2] Voir section 2.1.3 pour une définition des degrés de liberté

doit faire l'hypothèse que les 4 conditions suivantes sont respectées dans notre échantillon.

1. Les paramètres suivent une fonction linéaires : $y = \beta_0 + \beta_1 x + u$
2. L'échantillon est identiquement et indépendamment distribué (iid).
3. L'espérance du terme d'erreur sachant x est égale à zéro. E (u/x) = 0
4. Pas de multicolinéarité exacte.

Si l'échantillon est homoscédastique[3] et qu'il n'y a pas d'autocorrélation, on peut aussi assumer qu'on a des estimateurs efficace ou BLUE (Best linear Unbiased Estimators).

Maintenant, on peut déterminer si une variable est significative ou non en utilisant un test d'hypothèse. Une variable est significative lorsque la statistique du test (t, f, etc.) calculée par Stata se trouve dans la zone de rejet de l'hypothèse nulle, on suppose donc que $\beta>0$ ou $\beta<0$ ou $\beta\neq0$. On peut aussi utiliser la « p-value » pour déterminer si le coefficient passe le test de signification. La section ci-dessous fait un rappel de ce que sont les tests d'hypothèses.

Pour pouvoir assumer que les coefficients du modèle probit sont non-biaisés, le principe reste le même que celui de la MCO à cause de la variable latente ($y^*_i = \beta_0 + \beta' x_i + u_i$)[4]. Pour déterminer si la variable indépendante du probit est significative, on doit faire comme pour la MCO, c'est-à-dire, faire passer un test de signification.

[3] Voir section 2.1.4 pour un définition de l'homoscédasticité
[4] voir section 3.5.2

2.1.3 Les tests d'hypothèses

Faire un test d'hypothèse consiste vérifier si l'effet marginal de β sur la variable dépendante est nul ou non nul en comparant une statistique de test calculée à l'aide de paramètres estimés (β et σ) à une statistique critique. Dans cette section, nous vous parlerons des quatre statistiques de test les plus souvent utilisées dans le cadre de votre cours, soit la t de Student, la f de Fisher, la z de la distribution normal standard et la « p-value ». Avant de parler plus précisément des quatre statistiques de test, nous ferons un bref rappel des principes fondamentaux du test d'hypothèse.

La première chose à faire est de formuler l'hypothèse que l'on veut tester. On doit donc définir notre hypothèse nulle (H_0) et l'hypothèse alternative. Dans le cadre de régression, H_0 consiste, la plupart du temps, en un coefficient égal à zéro ($H_0 : \beta=0$). En termes économiques, cela veut dire que l'effet marginal des coefficients sur la variable dépendante est nul. L'hypothèse alternative peut aussi prendre diverses formes selon le cas : $H_1 : \beta\neq0$, $H_1 : \beta>0$ ou $H_1 : \beta<0$. La formulation de l'hypothèse alternative est très importante puisqu'elle vient influencer la zone de rejet du test. Cette zone est déterminée en fonction du niveau de confiance choisi (α) et si de si on fait un test à une ou deux queues. Plus le niveau de confiance est élevé, plus le test est précis. En sciences humaines, on choisi généralement un niveau de 5%. Dans ce cas, il y a 5% des chances que l'on rejette l'hypothèse nulle alors qu'elle est vrai. De plus, quand la situation le permet, il est préférable de privilégier un test bilatéral pour avoir un test plus précis.

La **statistique t** se calcule ainsi : ^β - β / σ^β. Stata donne cette statistique dans le tableau des résultats d'un régression par MCO. Alors, avec un niveau de confiance de 95% et un nombre infini de degrés de libertés, si H_0 : β=0 et H_1 : β≠0, la zone de non-rejet sera de -1.96 à 1.96. Ceci est un test bilatéral. Si H_0 : β=0 et H_1 : β>0, la zone de non-rejet sera de 0 à 1.64. Si H_0 : β=0 et H_1 : β<0, la zone de non-rejet sera de -1.64 à 0. Ceux-ci sont des tests unilatéraux. Donc, on rejette H_0 : β=0 si la statistique t donnée se trouve à l'extérieure de l'intervalle de confiance. Si t est rejeté, cela veut dire que notre coefficient a un impact sur notre variable indépendante, donc qu'elle est statistiquement significative.

En ce qui concerne la **statistique z**, avec un niveau de confiance de 95%, si H_0 : β=0 et H_1 : β≠0, la zone de non-rejet sera aussi de -1.96 à 1.96. Si H_0 : β=0 et H_1 : β>0, la zone de non-rejet sera de 0 à 1.645. Si H_0 : β=0 et H_1 : β<0, la zone de non-rejet sera de -1.645 à 0. Donc, on rejette H_0 : β=0 si la statistique z donnée se trouve à l'extérieure de la zone de non-rejet. Si z est rejeté, cela veut dire que notre coefficient a un impact sur notre variable indépendante, donc qu'elle est statistiquement significative.

La **statistique f** (test de signification conjointe de Fisher) est caractérisée par deux valeurs: *q*, le nombre de contraintes, i.e. le nombre de degrés de libertés du numérateur et *k*, le nombre de coefficients du modèle non-contraint, (*n* – *k*) est le nombre de degrés de libertés du dénominateur.

$$f = (R^2 / k) / (1- R^2) (n-k-1)$$

Dans le cas où on a deux contraintes et où (*n* – *k*) peut être considéré infini (>100), la valeur critique de la statistique f à 95% est 3.00, i.e. *Prob* [q,n k F . . f] = 0.95. Ainsi, si la valeur de la statistique f obtenue est supérieure à la valeur critique, on rejette

l'hypothèse nulle. Dans le cas contraire, on ne peut rejeter l'hypothèse nulle. Un exemple d'hypothèse à tester est donné à la section 3.1 du guide.

La **« p-value »** est une probabilité (entre 0 et 1) qui indique la probabilité sous $H_0 : \beta=0$ d'obtenir la valeur trouvée. Ainsi, si le « p-value » est sous le α désiré (5%), on rejette l'hypothèse nulle. Une « p-value » de 0.0000 rejette très fortement l'hypothèse nulle.

2.1.4 Homoscédasticité vs Hétéroscédasticité

Si, par hypothèse, on assume que le terme d'erreur de notre modèle est homoscédastique, on peut dire que l'on a des coefficients efficaces. L'homoscédasticité qualifie une variance constante des résidus de données composant l'échantillon. À l'inverse, on dit qu'il y a hétéroscédasticité lorsque la variance des résidus du modèle n'est pas constante. L'hétéroscédasticité ne biaise pas l'estimation par MCO des coefficients, mais révèle l'inefficacité des coefficients. En effet, puisque les écarts-types trouvés sont surestimés ou sous-estimés, on ne peut se référer à une table afin de comparer la valeur obtenue aux valeurs critiques de la statistique concernée puisque la valeur obtenue n'est pas la bonne. L'hétéroscédasticité est une situation rencontrée fréquemment dans les données, il est donc important de savoir la détecter[5] et la corriger[6].

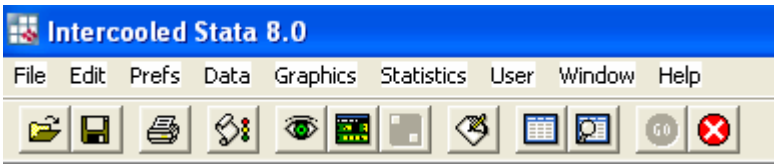
2.2 Commandes de base sur Stata

Le texte qui se trouve en *italique* désigne le texte tel qu'il serait entré à l'ordinateur, dans Stata, pour obtenir les manipulation souhaités.

[5] Voir section 3.1
[6] Voir section 2.5.1

2.2.1Pour débuter l'analyse

Lorsqu'on utilise Stata, il est préférable d'utiliser un fichier **Do-file**. Ce faisant, il est plus facile de sauver les commandes de programmation. Pour ne pas perdre ses commandes après une séance de TP, la façon la plus simple est de copier-coller son Do.file dans un document Word, de sauver ce fichier Word sur le bureau de l'ordinateur, et de se l'envoyer, en attachement, sur son compte de courriel (umontreal.ca, hotmail, etc.). Pour travailler avec le Do.file, on clique sur l'icône  se trouvant sur la barre d'outils de Stata.



Un page blanche apparaît, et c'est sur cette page blanche que l'on inscrit toutes les commandes que le veut faire exécuter sur nos données. Pour exécuter les commandes désirées, il suffit d'appuyer sur l'icône de la  se trouvant sur la barre d'outil du fichier Do-file.

On débute le programme avec la commande *clear* afin de vider la mémoire de Stata. Ensuite, on fixe un seuil de mémoire disponible pour Stata avec *set memory 800m* (ici, le chiffre peut varier entre 32m et 2000m, essayez en plusieurs jusqu'à ce que Stata accepte le seuil). Si vous travaillez avec beaucoup de variables, il est préférable de fixer une grandeur de matrice. Selon la version de Stata employée, le seuil maximal varie

(version 5 = 400, version 8 = 11 000) : *set matsize 400* ou *set matsize 11000.* Par la suite, on ouvre le fichier log et on renomme notre fichier de données : *log using nouveaunom.log, replace.* Puisqu'en faisant cette commande, on ouvre notre fichier log, on devra fermer le fichier log à la fin du programme ou à chaque fois que Stata détecte une erreur dans notre programmation (on doit, dans ce cas corrige l'erreur de programmation et on recommence du début). Pour fermer le fichier log, on doit inscrire *log close.*

Importation de données. Avant d'aller chercher votre fichier de données pour le lire sur Stata, assurez vous que le format de celles-ci est compatible avec Stata. Outre les séparateurs de données qui doivent correspondre à la commande choisie, il faut aussi s'assurer que le séparateur de décimales soit un point (.) et que les milliers ne soient pas séparés par un espace.

Il existe deux façons d'aller chercher votre fichier. L'utilisation de l'une ou l'autre dépend de la manière dont vos données sont disposées dans votre fichier. Rapide et efficace, *insheet* permet d'importer les données d'un fichier texte possédant une observation par ligne et dont les données sont séparées par des tabulations ou des virgules. Si le nom des variables est sur la première ligne:

insheet using "*nomdefichier*"[7]

Si le fichier ne contient pas le nom des variables:

insheet [nom des variables] using "*nomdefichier*"

Infile[8] permet plus de flexibilité que *insheet* en permettant que les observations soient sur plusieurs lignes ou que les données soient séparées par des espaces.

[7] Nomdefichier indique le nom complet, donc avec le chemin d'accès et l'extension. Par exemple, si votre fichier provient du bloc note : `C:\Documents and Settings\p0678264\Bureau\EXTRACT.TAB`

Données pondérées. Il est fortement possible que les données votre base de données soient pondérées. Par exemple, dans un recensement, les répondants n'ont pas tous le même poids dans le sondage. En effet, un répondant de l'Ile du Prince Édouard n'a pas le même poids qu'un répondant provenant de l'Ontario, la population de l'Ontario représentant 39% de la population canadienne alors que la population de l'Ile du Prince Édouard ne représente que 0.4% de celle-ci. Si vos données sont pondérés, cela sera généralement indiqué sur à la fin de la base de données ou dans le cliché d'enregistrement. Si tel est le cas, vous devrez l'indiquer à Stata à l'aide de la fonction suivante : *svyset [pweight = nomdevariabledepondération]* Lorsque Stata est avisé que vos données sont pondérées, il suffit d'ajouter *svy* avant chaque fonction (exemple : *svymean, svyregress, svyprobit…).* Ce faisant, vous n'avez pas à toujours ajouter *[pweight = fweight]* comme option à la fin de vos fonction. Une exception à cette règle existe toutefois. En faisant une régression par MCO, vous ne pouvez corriger pour l'hétéroscédasticité en utilisant l'option *robust* si vous utilisez la fonction *svyregress*. Donc, il faut utiliser *regress* et ajouter *[pweight = fweight]* en option à la fonction.[9]

Voici un exemple de ce à quoi devrait ressembler un début de fichier Do-file :

```
clear
set memory 800m
log using nouveaunom.log, replace
insheet using "C:\Documents and Settings\p0678264\Bureau\EXTRACT.TAB"
svyset [pweight = fweight]
```

[8] Pour plus de détails sur cette fonction, voir le Guide d'économétrie appliquée pour Stata de Simon Leblond, page 8. http://www.sceco.umontreal.ca/bibliotheque/guides/GuideEconometrieStata.pdf

[9] Voir exemple à la section 2.5.1

2.2.2Création de nouvelles variables.

C'est la commande *generate* ou *g* qui permet de créer de nouvelles variables. *g nomnouvellevariable = nomvariable.* Le nom de la variable indiqué du côté gauche du signe d'égalité est le nom de la variable que l'on veut créer par l'entremise de l'opération, et le nom du côté droit est le nom de la variable tel qu'attribué ou correspondant à une variable dans le fichier de données. On peut désirer créer des nouvelles variables parce que celles que l'on retrouve dans notre base de données ne nous satisfait pas. Par exemple, si on veut estimer l'impact d'être une femme sur le niveau d'éducation, la variable « sex » qui englobe les hommes et les femmes n'est pas assez précise. On créera donc une variable binaire ou dummy (surtout pour les étudiants de FAS 3900). Par exemple : *g femme = (sex==2)* et *g homme = (sex==1).* En terme littéraire cela veut dire : générer la nouvelle variable appelée *femme* lorsque la variable d'origine appelé *sex* prend la valeur 2 (lorsque le répondant est une femme, 2 est inscrit dans la base de donnée). Pour les étudiants de ECN 3950, créer une nouvelle variable voudra probablement dire la modifier mathématiquement. Par exemple : *g salaire1 = log(salaire)* ou *g salaire2 = salaire^2.*

Dans le tableau suivant, vous trouverez les opérateurs logiques et de comparaisons les plus fréquemment utilisés.

Soustraction	-	Addition	+
Division	/	Multiplication	*
Non (¬)	~	Puissance	^
Ou	|	Et	&
Renvoie l'argument	Min	Renvoie l'argument possédant la	Max

possédant la valeur la moins élevée	$(x_1 \ldots x_n)$	valeur la plus élevée	$(x_1 \ldots x_n)$
Différent	~=	Égal	==
Racine carrée de x	Sqrt(x)	e^x	Exp(x)
Plus petit	<	Plus grand	>
Logarithme de x	Log(x)	Σ x	Sum(x)
Plus petit ou égal	<=	Plus grand ou égal	>=

2.2.3 Divers

Il est possible d'insérer des commentaires dans son programme en prenant soin de débuter la ligne de commentaire par le symbole '*'. Par exemple: * Ceci est un commentaire.

La majorité des fonctions peuvent être suivies de *if* qui permet de spécifier une condition pour que la commande soit exécutée. *if* est placé après la fonction, mais avant les options. Par exemple: *regress y x1 x2 x3 if sex==1*

La majorité des fonctions peuvent être suivies de *in* qui permet de spécifier l'étendue des données affectées par la fonction. *in* est placé après la fonction, mais avant les options. L'étendue peut prendre la forme # ou #/#, et # peut-être un nombre positif, l (dernière observation), f (première observation) ou un nombre négatif (distance par rapport à la dernière observation). Par exemple : *regress y x1 x2 x3 in f/60* (les 60 premières observations) ou *regress y x1 x2 x3 in -10/l* (les 10 dernières observations).

Si vous voulez afficher à l'écran la valeur de certaines variables, faite *list nomsdesvariables.* Par exemple : *list sex in -10/l* (Stata affichera la valeur des 10 dernières observation de la variable sexe).

Si vous ne voulez pas retenir certaines catégories d'une variable binaire, vous pouvez utiliser la fonction *drop.* Par exemple : *drop if sex==2*

2.3 Statistiques de l'échantillon

Il est toujours recommandé d'examiner le portrait de notre échantillon avant de commencer à faire des manipulations. En effet, cela permet de vérifier s'il y a des anomalies dans l'échantillon qui pourraient venir biaiser les estimateurs. Un exemple d'anomalie pourrait être de retrouver quelques données très éloignées de la moyenne (le revenu de Bill Gates dans un échantillon du revenu d'infirmières). Pour ce faire, deux options s'offrent à vous. Si vous êtes plus visuel, faire un graphique (voir section 2.5) des données s'avère la meilleure option. Sinon, il suffit d'utiliser l'option *summarize* (ou *mean*). Vous n'avez qu'à inscrire la fonction suivit du nom de vos variables. Par exemple :

svymean homme femme age1519 age2024 age2529 age3034 age3539 age4044 age4549 age5054 age5559 age6069 celibataire marieunionlibre veufs separedivorce eduprimaire educsecondpartielles diplomesecondaire etudespostsec diplomepostsec bacc diplomedeuxiemecycle salaire lnsalaire Un tableau comme celui-ci apparaît.

```
pweight:  fweight                                 Number of obs    =     42335
Strata:   <one>                                    Number of strata =         1
PSU:      <observations>                          Number of PSUs   =     42335
                                                  Population size  =  13939734

------------------------------------------------------------------------------
    Mean |   Estimate    Std. Err.   [95% Conf. Interval]        Deff
---------+--------------------------------------------------------------------
   homme |   .3930934    .0030632    .3870895    .3990974    1.665059
   femme |   .6069066    .0030632    .6009026    .6129105    1.665059
 age1519 |   .0488869    .0013093    .0463206    .0514532    1.560858
 age2024 |   .1142093    .0019988    .1102915     .118127    1.671908
 age2529 |   .1186154    .0021007     .114498    .1227328    1.786939
 age3034 |   .1150967    .0020376    .1111029    .1190904    1.725733
 age3539 |   .1258951    .0020688    .1218401    .1299501    1.646534
 age4044 |   .1452781    .0021569    .1410505    .1495057    1.586109
```

```
 age4549 |   .1322608    .0020378    .1282667    .1362549    1.531739
 age5054 |   .1100464     .001911    .1063007     .113792    1.578657
 age5559 |   .0636861    .0015086    .0607292    .0666429    1.615699
 age6069 |   .0260253    .0009853    .0240942    .0279565    1.621276
celiba~e |   .2914449    .0028989     .285763    .2971269    1.722782
marieu~e |   .6256523    .0030379     .619698    .6316067    1.668121
   veufs |   .0071991    .0004553    .0063067    .0080914    1.227814
separe~e |   .0757037      .00158    .0726068    .0788005    1.510388
edupri~e |   .0300358    .0010069    .0280622    .0320094    1.473328
educse~s |   .0841567    .0016777    .0808683    .0874451    1.546026
diplo~re |   .1698306    .0023563    .1652123    .1744489    1.667056
etudes~c |   .0828385    .0017234    .0794606    .0862164    1.654926
diplom~c |   .3998383    .0030275    .3939044    .4057723    1.616983
    bacc |    .172731     .002412    .1680034    .1774585    1.723557
diplo~le |    .060569    .0015532    .0575247    .0636133    1.794879
 salaire |   17.10388    .0541942    16.99766     17.2101    1.684761
lnsala~e |    2.72358    .0029882    2.717723    2.729437    1.641865
------------------------------------------------------------------------------
```

On retrouve le nom des variables dans la première colonne. La proportion de l'échantillon qui se retrouve dans les catégories de chaque variable (en bref, la moyenne) se trouve dans la seconde colonne. On ne doit pas oublier que la somme doit être égale à 1 pour chaque groupe de variables dichotomiques. Nous ne nous attarderons pas aux autres colonnes, car elles sont moins utiles pour l'analyse.

Finalement, en utilisant ces données, vous pourrez faire des tableaux croisés qui vous donneront une intuition des résultats de votre régression (i.e. les moyennes donnent une bonne idée du signe (+/-) du coefficient. Voici un exemple de tableau croisé.

Origine du vote, élection présidentielle, USA, 2004, %		
	Républicain	Démocrates
Catholique	1	40
Protestants	54	5

Source : Résultats fictifs

La fonction *table* ou *tabulate* dans Stata s'avère une autre façon de créer un tableau croisé. Il suffit d'inscrire la fonction suivit des variables que l'on désire voir dans

le tableau, en prenant soins de toujours inscrire le nom de la variable dépendante en premier.

2.4 Graphiques et tableaux[10]

Pour tracer des graphiques, il suffit d'inscrire *graph* suivit du type de graphique ainsi que le nom des variables que l'on veut sur le graphique. Le type de graphique peut prendre les valeurs suivantes:

• twoway (t): nuage de points à deux axes; valeur par défaut si plusieurs variables sont affichées. La première variable spécifiée est toujours la variable dépendante.

• bar (b): graphique à barres

• pie (p): graphique en pointe de tartes

Par exemple : *graph bar education femme homme*

Si on souhaite donner un titre au graphique :

graph bar education femme homme, title(inscrire le titre souhaiter)

2.5 Régressions

2.5.1Régression par les moindres carrés ordinaires (MCO)

La régression par les MCO est une méthode qui permet d'estimer les coefficients d'une régression linéaire multivariée en minimisant la somme des carrés des résidus. La régression par MCO permet d'obtenir des estimateurs BLUE.

$$\hat{y} = \beta 0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x + {}_3 \ldots \beta_n x_n$$

[10] Pour plus de détails sur cette fonction, voir le Guide d'économétrie appliquée pour Stata de Simon Leblond, page 8. http://www.sceco.umontreal.ca/bibliotheque/guides/GuideEconometrieStata.pdf

En faisant une régression à l'aide de Stata, vous obtiendrez donc une liste de coefficients (β1, β2, β3, …βn). Ici, il est préférable de ne pas inscrire les coefficients sous la forme d'une équation (par exemple : $\hat{y} = 0.33 + 0.25x_1 + 0.25x_2 + 0.25x_3$). Si vous décidez de le faire, il est important d'inscrire la valeur de la statistique t en dessous de chaque coefficient afin que le lecteur sache si les variables sont significatives ou non. Une façon plus élégante de présenter les résultats est d'inscrire le nom des variables ainsi que leur coefficient et leur statistique t dans un tableau, et souligner les variables qui passent le test de « signification » en gras.

Pour programmer une régression sur Stata, il suffit d'inscrire *regress* suivit de la variable dépendante et des variables indépendantes. Dans notre programmation, on doit laisser tomber une catégorie pour chaque variable binaire (pour éviter le problème de multicolinéarité parfaite). L'exemple suivant inclus aussi une condition (*if prive==1)*, la pondération et la correction de l'hétéroscédasticité. La régression par MCO de la section 3.7.1 a été faite à partir de cette programmation.

regress lnsalaire homme age1519 age2529 age3034 age3539 age4044 age4549 age5054 age5559 age6069 marieunionlibre veufs separedivorce educsecondpartielles diplomesecondaire etudespostsec diplomepostsec bacc diplomedeuxiemecycle dmois pans dans tans tempsplein couverturesyndicale professionel personneldebureau sante education hotellerierestauration protection saisonnier temporairecontractuel occasionnelautre entre20et99employes entre100et500employes plusde500employes if prive==1 [pweight = fweight] , robust

Correction de l'hétéroscédasticité. Effectuer une régression par MCO et calculer les variances robustes d'Eicker-White. Dans Stata, il suffit d'ajouter l'option *robust* (exemple : *regress y x1 x2 x3, robust*) à sa régression pour corriger les écarts-types. Toutes les interprétations et les tests s'effectuent comme auparavant avec les nouveaux écarts-types. Il peut être tentant d'utiliser systématiquement les écarts-types robustes,

mais il faut savoir que cette méthode gonfle les écarts-types inutilement et réduit la puissance des tests lorsque ceci n'est pas nécessaire. Il faut donc s'abstenir de l'utiliser lorsqu'elle ne s'avère pas nécessaire.

2.5.2 Probit/Dprobit

Le probit fait partie de la famille des modèles de régression pour variables dépendantes prenant des valeurs dichotomiques. On parle ici des probit, logit, etc. Dans le cadre de ce guide, nous nous attarderons uniquement au plus simple de ces modèles, c'est-à-dire le probit.

Le probit permet de comprendre l'effet d'une variable indépendante sur la probabilité de se retrouver dans un état. On arrive essentiellement au même but que celui de la MCO, c'est-à-dire de « prédire » la valeur d'une variable dépendante à l'aide de variables indépendantes (ou explicatives). Néanmoins, dans le cas d'un probit, la variable dépendante est qualitative. Le modèle ressemble à ceci :

$$y^*_i = \beta_0 + \beta' x_i + u_i$$

y^*_i est une variable latente, c'est-à-dire qu'elle est inobservable (ex. : propension à acheter, préférence d'avoir des enfants, préférence pour un parti politique…). Néanmoins, on peut observer le comportement de l'individu. Par exemple : l'individu a acheté une voiture ou l'individu vote pour tel parti politique, etc. Dans le modèle probit, la variable dépendante est une variable binaire (dummy) dont la valeur est 1 quand l'événement se produit, et 0 quand il ne se produit pas (note : le zéro est un seuil choisi arbitrairement).

$$Y=1 \text{ si } y^*_i > 0 \text{ ou } Y=1 \text{ si } (\beta_0 + \beta' x_i + u_i) > 0$$
$$Y=0 \text{ autrement}$$

Le modèle probit donne l'effet de la variation d'une unité de la variable indépendante sur la probabilité que l'évènement se produise. Sa distribution normale standard cumulative ($\Phi(z)$) permet de restreindre la distribution des valeurs que le paramètre β_i peut prendre à des valeurs entre 0 et 1.

Prob (y=1) = prob ($\beta_0 + \beta_i x_i + u_i > 0$)
Prob (y=1) = prob ($u_i > - (\beta_0 + \beta_i x_i)$)
Prob (y=1) = 1- Φ ($-(\beta_0 + \beta_i x_i)$)
Parce que $u_i \sim N(0,\sigma^2)$
Prob (y=1) = Φ ($\beta_0 + \beta_i x_i$)

En somme, ce que l'on cherche à connaître ici est l'effet de x_i sur la probabilité de voir l'évènement se produire. Or, le probit, tel que formulé ci-dessus, nous donne la probabilité associée à une valeur donnée de la valeur latente (y^*_i) exprimé par la combinaison linaire des variables indépendantes. La façon d'obtenir l'effet de x sur la probabilité que l'évènement se produise est de faire un dprobit. Le dprobit dérive la fonction sur x_i.

δ Prob (y=1) / δ x_i
δ (Φ ($\beta 0 + \beta_i x_i$)) / δ x_i)
($\delta \Phi / \delta$ f) * (δ f/ δ x_i)
Prob (y=1) = fi ($\beta_0 + \beta_i x_i$) * (β_i)
où fi est la fonction de densité de probabilité

En langage très vulgarisé, le dprobit remplace par le x_i par la valeur pour l'individu moyen, ce qui permet de calculer la probabilité qu'un individu moyen se retrouve dans un certain état.

Pour faire un probit, il suffit d'inscrire *probit* suivit de la variable dépendante et des variables indépendantes (comme pour la régression…). *probit variabledépendante variable indépendante* On inscrit *dprobit* suivit de la variable dépendante et des variables

indépendantes pour faire un dprobit. Les probit/dprobit de la section 3.7.2 a été fait à partir de cette programmation.

probit ref rural panglo km2 ymedian pgv pimm tax5ans pop ratiopop mois quebec age
dprobit ref rural panglo km2 ymedian pgv pimm tax5ans pop ratiopop mois quebec age

Options: probit possède en grande partie les mêmes options que *regress*.

Note: Ici *predict* donne par défaut la probabilité. Pour avoir l'estimation linéaire, il faut préciser *xt* dans les options de *predict*.
Note : Pour le Probit, on n'a pas besoin de corriger pour l'hétéroscédasticité puisque que l'échantillon est homoscédastique par hypothèse.

2.6 L'interprétation des résultats

Dans le cadre de cette section, nous allons décrire et expliquer la signification des résultats des régressions (MCO et probit).

2.6.1 Régression par MCO

regress lnsalaire homme age2024 age2529 age3034 age3539 age4044 age4549 age5054 age5559 age6069 marieunionlibre veufs separedivorce educsecondpartielles diplomesecondaire etudespostsec diplomepostsec bacc diplomedeuxiemecycle dmois pans dans tans tempsplein couverturesyndicale professionel personneldebureau sante education hotellerierestauration protection saisonnier temporairecontractuel occasionnelautre entre20et99employes entre100et500employes plusde500employes if prive==1 [pweight = fweight] , robust

```
prive==1 [pweight = fweight] , robust
(sum of wgt is   7.6542e+06)

Regression with robust standard errors                Number of obs =   22265
                                                      F(37, 22227) =  541.10
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4913
                                                      Root MSE      =  .30906
------------------------------------------------------------------------------
             |               Robust
   lnsalaire |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       homme |    .138624   .0059218    23.41   0.000     .1270169    .1502311
     age2024 |   .0612328    .008348     7.34   0.000     .0448702    .0775954
     age2529 |   .1952287   .0118512    16.47   0.000     .1719996    .2184578
     age3034 |   .2519569   .0130786    19.26   0.000      .226322    .2775918
     age3539 |   .2882838   .0129534    22.26   0.000     .2628942    .3136734
     age4044 |   .2918056   .0125381    23.27   0.000       .26723    .3163811
```

```
     age4549 |   .3182847    .013374    23.80   0.000     .2920707    .3444987
     age5054 |   .3130712     .01386    22.59   0.000     .2859045    .3402379
     age5559 |    .314402   .0158076    19.89   0.000     .2834179    .3453861
     age6069 |   .2059092   .0216126     9.53   0.000      .163547    .2482715
marieunion~e |   .0454065   .0075136     6.04   0.000     .0306792    .0601337
       veufs |   .0622398   .0363485     1.71   0.087    -.0090058    .1334854
separedivo~e |   .0009371   .0127897     0.07   0.942    -.0241317    .0260059
educsecond~s |   .0584898   .0118429     4.94   0.000     .0352768    .0817027
diplomesec~e |   .1807878   .0119777    15.09   0.000     .1573107    .2042649
etudespost~c |   .1701347   .0133757    12.72   0.000     .1439173    .1963521
diplomepos~c |   .2166452   .0116543    18.59   0.000     .1938018    .2394885
        bacc |    .338646    .017189    19.70   0.000     .3049543    .3723376
diplomedeu~e |   .3521967   .0293534    12.00   0.000      .294662    .4097313
       dmois |  -.0098113   .0108557    -0.90   0.366    -.0310892    .0114665
        pans |   .0411321   .0083154     4.95   0.000     .0248334    .0574308
        dans |   .0890161   .0108985     8.17   0.000     .0676543    .1103779
        tans |   .1918873   .0101919    18.83   0.000     .1719105    .2118641
  tempsplein |   .1192996   .0071305    16.73   0.000     .1053234    .1332758
couverture~e |   .0400635   .0066268     6.05   0.000     .0270744    .0530526
professionel |  -.2397607   .0393663    -6.09   0.000    -.3169214      -.1626
personneld~u |  -.5565081   .0386123   -14.41   0.000    -.6321908   -.4808253
       sante |  -.3871212   .0427172    -9.06   0.000      -.47085   -.3033924
   education |   -.345052   .0653186    -5.28   0.000     -.473081    -.217023
hotellerie~n |  -.7273983   .0395429   -18.40   0.000    -.8049052   -.6498914
  protection |  -.6459752   .0392655   -16.45   0.000    -.7229383   -.5690122
  saisonnier |  -.0775936   .0157025    -4.94   0.000    -.1083715   -.0468156
temporaire~l |  -.0256905   .0112985    -2.27   0.023    -.0478363   -.0035446
occasionne~e |  -.0435002   .0118475    -3.67   0.000     -.066722   -.0202783
entre20et9~s |    .043604   .0065027     6.71   0.000     .0308582    .0563498
entre100et~s |   .1171781   .0078496    14.93   0.000     .1017923    .1325639
plusde500e~s |   .1694895   .0127836    13.26   0.000     .1444328    .1945463
       _cons |   2.354517   .0414752    56.77   0.000     2.273223    2.435811
------------------------------------------------------------------------------
```

La R-squared (R-carré) est la proportion de la variation de la variable indépendante qui est expliquée par les variables indépendantes. Il est préférable d'utiliser un R-carré ajusté puisque le R-carré est affecté par le nombre de variables indépendantes. Le R-carré est biaisé à la hausse lorsque le nombre de variables indépendantes est élevé. La plupart du temps, Stata donne le R-carré et le R-carré ajusté. Sinon, vous devez le calculer vous-même. Si vous obtenez un R-carré qui semble petit, il ne faut pas nécessairement rejeter votre modèle. Un faible R-carré vous donne plutôt l'indice qu'il manquerait des variables explicatives pertinentes à votre modèle.

Dans la première colonne, on retrouve nos variables indépendantes. Les coefficients (β) sont dans la seconde colonne (note: le coefficients sont toujours exprimés dans les unités de la variable dépendante). Dans la troisième colonne, on retrouve les

écarts types estimés. La quatrième colonne donne la statistique t. La statistique t est essentielle afin de déterminer si les coefficients **sont significatifs**. La statistique t doit être interprétée à l'aide de la table de Student. Finalement, les dernières colonnes donnent l'intervalle de confiance à un niveau de 5%.

2.6.2 Probit/Dprobit

probit ref rural panglo km2 ymedian pgv pimm tax5ans pop ratiopop mois quebec age

```
Probit estimates                                  Number of obs   =        207
                                                  LR chi2(12)     =     118.00
                                                  Prob > chi2     =     0.0000
Log likelihood = -81.509878                       Pseudo R2       =     0.4199

------------------------------------------------------------------------------
         ref |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       rural |   -.036977   .2883437    -0.13   0.898    -.6021203    .5281663
      panglo |   3.687492   1.020701     3.61   0.000     1.686954     5.68803
         km2 |  -.0015299   .0008365    -1.83   0.067    -.0031693    .0001096
     ymedian |   .0000874   .0000324     2.70   0.007      .000024    .0001508
         pgv |  -3.354991   .9138396    -3.67   0.000    -5.146084   -1.563898
        pimm |   1.599559    1.47169     1.09   0.277      -1.2849    4.484018
     tax5ans |  -1.116375      .7738    -1.44   0.149    -2.632995    .4002455
         pop |  -.0000153   7.91e-06    -1.94   0.052    -.0000308    1.51e-07
    ratiopop |   .0207279    .271442     0.08   0.939    -.5112887    .5527445
        mois |  -.0429848   .0265225    -1.62   0.105     -.094968    .0089983
      quebec |   3.000821   1.156023     2.60   0.009     .7350584    5.266584
         age |   .0725823   .0370994     1.96   0.050    -.0001311    .1452958
       _cons |  -4.404904   2.140513    -2.06   0.040    -8.600233   -.2095753
```

Le tableau des résultats du *probit* est très similaire à celui de la régression par MCO : on retrouve la statistique z au lieu de la statistique t. Le coefficient s'interprète toutefois plus ou moins bien, comparé à celui résultant d'une MCO. Le coefficient donne l'effet marginal d'une variation d'une unité de la variable indépendante x_i sur la valeur de la variable latente y^*_i. Un *dprobit* s'avère donc nécessaire.

dprobit ref rural panglo km2 ymedian pgv pimm tax5ans pop ratiopop mois quebec age

```
Probit estimates                                  Number of obs =    207
                                                  LR chi2(12)   = 118.00
                                                  Prob > chi2   = 0.0000
Log likelihood = -81.509878                       Pseudo R2     = 0.4199
```

```
------------------------------------------------------------------------------
     ref |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
   rural*|   -.013911   .1080834    -0.13   0.898   .188406  -.225751  .197929
  panglo |   1.392722   .4030385     3.61   0.000   .129188   .602781  2.18266
     km2 |  -.0005778   .0003107    -1.83   0.067   117.678  -.001187  .000031
 ymedian |    .000033   .0000125     2.70   0.007   21781.9   8.4e-06  .000058
     pgv |   -1.26714   .3232523    -3.67   0.000    .19797   -1.9007 -.633577
    pimm |   .6041343   .5583417     1.09   0.277   .654607  -.490195  1.69846
 tax5ans |  -.4216414   .2911896    -1.44   0.149   .073657  -.992362   .14908
     pop |  -5.79e-06   2.99e-06    -1.94   0.052     20583  -.000012  6.2e-08
ratiopop |   .0078287   .1025197     0.08   0.939   .892683  -.193106  .208764
    mois |  -.0162348   .0099063    -1.62   0.105   28.1256  -.035651  .003181
  quebec*|   .6917028   .0583106     2.60   0.009   .062802   .577416  .805989
     age |   .0274135   .0140448     1.96   0.050   39.9246  -.000114  .054941
---------+--------------------------------------------------------------------
  obs. P |   .4154589
 pred. P |   .3703592  (at x-bar)
```

Dans le tableau des resultats du *dprobit*, on retrouve deux nouvelles colonnes : celle du dF/dx et celled u x-bar. Le dF/dx donne l'effet marginal d'une variation d'une unité de la variable indépendante sur la probabilité de se retrouve dans un certain état (prob(y=1)). Le x-bar correspond à la probabilité d'obtenir un individu dans la catégorie moyenne.

2.6.3 Interprétation économique

Dans le cas du probit comme de la régression par les MCO, des erreurs d'interprétation peuvent survenir. Tout d'abord, dans votre modèle on retrouve une variable dépendante et des variables indépendantes. Dans le cadre de votre recherche, vous vous intéressez à la relation entre une variable indépendante en particulier et la variable dépendante. Donc, les autres représentent des variables de contrôle (i.e. *ceteris paribus*), que vous devrez par ailleurs interpréter. Ainsi, nous voulons souligner que le lien de causalité prévaut entre la variable dépendante et la variable indépendante d'intérêt, et non pas entre les variables indépendantes. Par exemple, l'objectif d'interprétation n'est pas de comparer les coefficients de deux variables indépendantes du modèle afin de vérifier laquelle a une influence plus grande sur la variable dépendante.

Ensuite, si vous avez une variable indépendante qui cause indirectement une autre variable indépendante qui explique la variable dépendante, il est préférable de l'enlever du modèle. Par exemple, l'intelligence (variable indépendante) cause la scolarité (variable indépendante) qui explique le revenu (variable dépendante). Dans ce cas-ci, il faudrait enlever la variable scolarité parce qu'il devient difficile de déterminer si elle cause directement le revenu ou si elle a un effet sur le salaire par l'intermédiaire de la variable intelligence. Ici, l'intelligence serait une variable proxy.[11]

3 Manipulations plus poussées

3.1 Hétéroscédasticité

Détecter l'hétéroscédasticité. Plusieurs tests se ressemblant existent pour détecter l'hétéroscédasticité. On aborde dans ce chapitre deux de ces tests, le test de Breusch-Pagen et le test de White. L'idée générale de ces tests est de vérifier si le carré des résidus peut être expliqué par les variables explicatives du modèle. Si c'est le cas, il y a hétéroscédasticité.

La plus simple est le test de Breusch-Pagen:

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).

```
reg y x1 x2
predict u, r
g u2 = u^2
reg u2 x1 x2
```

[11] Faire attention de ne pas mélanger les variables instrumentales (que l'ont doit utiliser lorsqu'il y a corrélation entre les résidus et les x_i (corr (x_i ,u)≠0) et les variables proxy.

Il suffit alors de regarder la statistique F donnée par Stata.

La faiblesse du test de Breusch-Pagan est qu'il suppose les erreurs normalement distribuées. Afin de laisser tomber cette hypothèse, il suffit d'ajouter le carré des variables indépendantes et leurs produits croisés dans la régression de l'étape 3, il s'agît là du test de White. Afin de limiter le nombre de paramètre, on peut utiliser un test de White légèrement modifié:

$u^2 = \beta 0 + \beta 1\hat{y} + \beta 2\hat{y}^2 + e$

On procède pour le reste exactement de la même façon que pour le test de Breusch-Pagan.

Interprétation des résultats des tests d'hétéroscédasticité. Les deux tests mentionnés plus haut utilisent un test *F*. Dans le contexte d'un test d'hétéroscédasticité, l'hypothèse nulle est que tous les coefficients de la régression des résidus au carré sont nuls, bref il y a homoscédasticité. L'hypothèse alternative est qu'il y a hétéroscédasticité. Ainsi, si on rejette l'hypothèse nulle (« p-value » < alpha), on peut conclure à la présence d'hétéroscédasticité. Stata affiche toujours la «p-value» du test *F* de «overall significance» lorsqu'il effectue une régression. C'est exactement le test qui nous intéresse dans le cas de l'hétéroscédasticité. Il n'est donc pas nécessaire d'effectuer un test supplémentaire après la régression.

3.2 Séries chronologiques

Une série chronologique est le résultat d'un processus stochastique (aléatoire) indexé en fonction du temps. Plusieurs problèmes sont propres aux séries chronologiques, notamment en raison de la corrélation du terme d'erreur entre les observations

(autocorrélation) et de la possibilité de changement du processus générateur de données d'une époque à l'autre. Les sections qui suivent adressent la question de comment s'assurer que l'on peut travailler avec nos données chronologiques.

Il est tout d'abord important de modéliser les données, notamment les données financières parce qu'elles contiennent beaucoup de bruit, pour rendre le terme d'erreur blanc. Pour ce faire, il suffit d'inscrire :

arima variabledependante variableindependante, arima(p,d,q)

où p est le nombre de AR, d le nombre de différenciation et q le nombre de MA.

Il n'est pas nécessaire de préciser de variables indépendantes.

ex: AR(1)
arima t, arima(1,0,0)
ex: MA(1)
arima t, arima(0,0,1)
ex: ARIMA(1,1,2)
arima t, arima(1,1,2)

Pour choisir p et q, il est bon de regarder l'autocorrélogramme partiel (nombre de AR) et l'autocorrélogramme (nombre de MA) de la variable qui nous intéresse.

Lorsqu'on travaille avec des séries chronologiques dans Stata, il est nécessaire de l'en aviser par la commande *tsset.* On commence donc par générer la variable de temps (*t*). Ensuite, on écrit *tsset* suivit du nom de la variable de temps. Ex. : *tsset = t*

Voici comment reproduire l'équivalent des opérateurs Avance et Retard dans Stata pour travailler sur les séries chronologiques. L'opérateur *l* est l'opérateur Retard de stata. Il peut être utilisé avec toutes les fonctions qui acceptent les séries temporelles une fois que la déclaration de séries temporelles à été faite.

l#.variable, où *variable* est la variable sur laquelle l'opérateur doit agir et # est le nombre de retards à appliquer. Si # est omis, un seul retard est appliqué (équivalent à *l1.variable*).

```
tsset t
* un modèle AR2
regress y l.y l2.y
```

L'opérateur *f* est l'opérateur Avance de stata. Il peut être utilisé avec toutes les fonctions qui acceptent les séries temporelles une fois que la déclaration de séries temporelles à été faite.

f#.variable, où variable est la variable sur laquelle l'opérateur doit agir et # est le nombre d'avance à appliquer. Si # est omis, une seule avance est appliquée (équivalent à *f1.variable*).

```
tsset t
* une autre formulation pour un modèle AR2
regress f.y y l.y
```

3.2.1 Test d'autocorrélation

Inutile de mentionner que l'autocorrélation est un problème qui n'est pertinent que dans le cas des séries temporelles. . . Le test ρ est le test le plus simple à effectuer pour tester la présence d'autocorrélation:

1. récupérer les résidus de la régression qu'on désire tester;
2. régresser $\hat{u}_t$ sur $\hat{u}_{t-1}$ à $\hat{u}_{t-n}$ et X
3. Tester la signification conjointe des coefficients de cette régression par un test F.

Choisissons n périodes égal à 3.

```
reg y x1 x2
predict u, r
reg u l.u l2.u l3.u
```

Il suffit alors de regarder la statistique F donnée par Stata.

3.2.2 Stationnarité

Pour travailler avec des données temporelles, elles doivent conserver une distribution constante dans le temps. C'est le concept de stationnarité.

Série chronologique stationnaire : la distribution des variables chronologiques ne varie pas dans le temps. Un concept moins fort de stationnarité est généralement utilisé, la covariance-stationnarité ou stationnarité au second degré.

Série chronologique covariance-stationnaire:

• E [y t] = μ (l'espérance ne dépend pas de t)

• var [y t] = σ^2 (la variance ne dépend pas de t)

• cov [yt,ys] = γ k = t - s (la covariance ne dépend que de t-s)

Ainsi, si nos variables passées sont semblables à nos variables futures, on peut utiliser le passé pour tenter de prédire (sic) le futur.

Si nos données ne sont pas stationnaires, on se retrouve avec:

• biais de prévision

• prévision inefficace

• mauvaise inférence

Il existe trois sources principales de non-stationnarité. 1- **Changement structurel (break)** La fonction de régression change dans le temps, soit de façon discrète, soit de façon graduelle. Par exemple, dans le cas d'un changement politique. La démarche à suivre est détaillée dans la sous-section ci-dessous. 2- **Tendance déterministe** Les données suivent une tendance qui a une fonction définie: t, t^2, etc. Afin de résoudre le problème, il suffit d'inclure une variable de tendance dans le modèle de régression: $y = \beta_0 + \beta_1 t + \beta_2 x$ Malheureusement, tout n'est pas aussi simple que ça en a l'air: très souvent,

ce qu'on pense être une tendance déterministe est en fait une tendance stochastique. 3- **Tendance stochastique (racine unitaire)** Les données suivent une marche aléatoire avec ou sans dérive avec un coefficient de 1 pour le terme autorégressé : $y_t = y_{t-1} + \mu_t$. Il y a non-stationnarité car la variance n'est pas constante: $var(y_t) = t\sigma^2_\mu$. Les tests à effectuer pour détecter la présence d'une racine unitaire et les corrections à apporter dans ce cas sont décrits à la prochaine section.

Procédure pour stationnariser une série chronologique

Changement structurel On peut corriger cette situation en ajoutant une variable binaire ou une variable d'interaction qui modélise le changement structurel. Il n'existe pas de test à proprement parler pour identifier un changement structurel. L'identification se fait plutôt par analyse graphique et par analyse historique: Observe-t-on une variation importante dans les variables? Connaît-on un événement important qui aurait pu changer la distribution des variables dans le temps? Exemple: on étudie les exportations du Québec aux Etats-Unis de 1980 à aujourd'hui. Ne pas prendre en considération que l'ALE serait une erreur, puisque ce dernier change les règles du jeu à compter de 1991 (année d'entrée en vigueur de l'accord). Il faut donc inclure une variable binaire, « y1991 » par exemple, qui sera égale à zéro de 1980 à 1990, puis égale à un pour les années subséquentes. Nous posons donc implicitement l'hypothèse que la droite de régression se déplace parallèlement vers le haut à compter de 1991 (l'ordonnée à l'origine n'est plus la même). Si on avait plutôt supposé que c'était la pente qui avait été affecté, il aurait fallu ajouter une variable d'interaction. Bien qu'il n'existe pas de test pour identifier un changement structurel, il en existe tout de même un pour vérifier si le changement structurel soupçonné est réel ou non. : le **Test de Chow.** Ce que ce test vérifie dans les

faits, c'est si le coefficient d'une variable est différent pour deux groupes de données. Dans l'exemple donné plus tôt, le test de Chow vérifierait si la constante est statistiquement différente avant et après l'ALE. Le résultat du test est une statistique *F*. Dans le contexte du test de Chow, l'hypothèse nulle est qu'il n'y a pas de changement structurel, i.e. les coefficients sont égaux pour les deux groupes de données. Donc, si on rejette l'hypothèse nulle (« p-value »< alpha), il y a bel et bien changement structurel et on est justifié de le modéliser.

Considérez le modèle suivant:$y = \beta 0 + \beta_1 x_1 + \beta_2 x_2 + u$

La façon "classique" d'effectuer le test de Chow est d'effectuer la régression du modèle pour les deux groupes de façon indépendante et pour les deux groupes ensemble:

$\hat{Y}_1 = \beta_{10} + \beta_{11} x_{11} + \beta_{12} x_{12}$
$\hat{Y}_2 = \beta_{20} + \beta_{21} x_{21} + \beta_{22} x_{22}$
$\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

puis de tester si les coefficients sont statistiquement différents par un test F: $H_0 : \beta_1 - \beta_2 = 0$, $H_1 : \beta_1 - \beta_2 \neq 0$.

$$F = ((S\hat{S}R_y - S\hat{S}R_{y1} - S\hat{S}R_{y2})/q) / ((S\hat{S}R_{y1} - S\hat{S}R_{y2}) /n_1 + n_2 - 2k))$$

Rappel: $S\hat{S}R_y$ est la somme au carré des résidus $(\Sigma \hat{u}_i^2) = \Sigma (y_i - \hat{y}_i)^2$ et q est le nombre de contraintes et k le nombre de coefficients, ici q = k = 3

Une autre façon plus rapide d'effectuer ce test est de construire une variable binaire égale à un pour les observations du deuxième groupe et de faire une seule régression sur les variables originales et sur les termes d'interaction avec la variable binaire[2]:

Soit δ la variable binaire: $\hat{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \delta + \beta_4 x_1 \delta + \beta_5 x_2 \delta$

On désire maintenant tester si $\beta_0 = (\beta_0 + \beta_3)$, si $\beta_1 = (\beta_1 + \beta_4)$ et si $\beta_2 = (\beta_2 + \beta_5)$. Ce qui revient à tester si β_3, β_4 et β_5 sont conjointement différent de 0. Ceci peut être facilement effectué par un test de F.

ex:
g g2 = (groupe == 2)
*g g2x1 = g2*x1*
*g g2x2 = g2*x2*
reg y x1 x2 g2 g2x1 g2x2
test g2 g2x1 g2x2

Tendance déterministe Afin de régler le problème de la présence d'une tendance temporelle, il suffit de la modéliser. Il faut faire attention de bien choisir la tendance la mieux adaptée à nos données: linéaire, quadratique, logarithmique, etc.

ex: tendance quadratique
t = n
t2 = t^2
tsset t
regress y t t2

Racine unitaire On désire s'assurer que la série n'est pas parfaitement autocorrélée, i.e. $p \neq 1$ dans $y_1 = \alpha + \rho y_{t-1} + e_t$ ou, de façon équivalente, $\theta \neq 0$ dans $\Delta y_t = \alpha + \theta\ y_{t-1} + \varepsilon_t$. La seconde forme est généralement utilisée pour effectuer des tests. L'hypothèse nulle est donc $H_0 : \theta = 0$. Le test *t* ne tient malheureusement pas dans ce cas, car les données sont... non stationnaires sous H_0! Il faut donc utiliser une loi de Dickey-Fuller. Le test de **Dickey-Fuller** (DF) teste s'il y a une racine unitaire dans le processus générateur de données. La loi de DF sur laquelle le test se base diffère en fait selon l'hypothèse alternative qu'elle teste. Le choix de l'hypothèse alternative est donc

primordial pour la validité du test. Ce choix doit se baser sur l'analyse de l'économètre. Soit le modèle suivant:

$\Delta y_t = u + \beta t + \theta\, y_{t-1} + \varepsilon_{t,,}\ \varepsilon_{t,,}$ iid $(0,\alpha^2)$

Les hypothèses nulles et alternatives possibles sont:

- H_0 : $\theta = 1$(il y a une racine unitaire)
- H_{1A}: $\theta < 1$, $\mu = 0$, $\beta = 0$ (pas de constante ni de tendance)
- H_{1b}: $\theta < 1$, $\mu \neq 0$, $\beta = 0$ (une constante, mais pas de tendance)
- H_{1c}: $\theta < 1$, $\mu \neq 0$, $\beta \neq 0$. (une constante et une tendance)

Il faut spécifier dans Stata l'hypothèse alternative qu'on désire tester à l'aide des options *trend* et *constant*. Enfin, s'il y a de l'autocorrélation dans les données, il faut utiliser un test de Dickey-Fuller augmenté (ADF) (ou Phillips-Perron). Ce test ajoute des retards au modèle testé afin de contrôler pour l'autocorrélation. Par défaut, Stata effectue un test ADF avec un nombre prédéterminé de retards. Il faut par ailleurs faire attention car si on a trop peu de retards, le résidu est autocorrélé et le test incorrect, alors que s'il y en a trop, la puissance du test est diminuée.Il peut-être pertinent de faire un autocorrélogramme avant de faire ce test. Le nombre de retards à inclure peut être contrôlé grâce à l'option *lags*. Un test de DF standard est obtenu en fixant *lags(0)*.

Donc, pour travailler avec sur le modèle $\Delta y_t = \alpha + \theta\, y_{t-1} + \varepsilon$ plutôt que sur $y_1 = \alpha + \rho y_{t-1} + e_t$, il faut utiliser la fonction *arima* dans Stata. Pour effectuer un test DF augmenté sur une variable, on écrit *dfuller nomdevariable, option*. Les options sont *lags* qui spécifie le nombre de retards, et *trends* et *constant* qui permet d'inclure une variable de tendance et un constante dans la régression selon l'hypothèse nulle à tester choisie. Le test *pperron* possède exactement la même structure et les mêmes options que *dfuller*, mais effectue un test Phillips-Perron plutôt qu'un test Dickey-Fuller augmenté.

Interpréter les tests de racine unitaire Vous avez finalement réussi à vous décider sur un model à tester et votre logiciel statistique vient de vous donner un résultat? Maintenant, que devez-vous en conclure? Généralement, comme c'est le cas pour tous les tests, vous obtiendrez deux valeurs: la statistique de test et le « p-value » associé à cette statistique. Vous pouvez comparer la statistique de test aux valeurs critiques de la loi correspondante, mais il est plus simple, surtout dans ce cas, de regarder le « p-value ». Si celui-ci est inférieur au niveau de confiance que vous avez fixé, 5% par exemple, vous rejetez l'hypothèse nulle: ouf! Tout va bien, il n'y a pas de racine unitaire. Dans le cas contraire, on doit corriger le modèle tel qu'exposé ci-dessous.

Corrections à apporter au modèle La façon de corriger un modèle est de le différencier, i.e. soustraire à chaque observation la valeur de la période précédente. $y_1 = \alpha + \rho y_{t-1} + e_t$ devient donc $\Delta y_t = \alpha + \theta\, y_{t-1} + \varepsilon_t$ On voit bien que si l'hypothèse nulle tient, $\theta = 0$ et le terme disparaît du modèle. En d'autres termes, le fait de différencier au premier degré permet de retrouver la forme AR, MA ou ARMA, qui sont stationnaires. Deux mises en gardes :

• Il ne faut pas différencier un modèle avec tendance déterministe.

• Ne devenez pas fou avec la différenciation! De un, surdifférencier « au cas où » est néfaste et, de deux, la puissance de ces tests n'est pas énorme et, donc, le risque d'erreur est grand. Dans le doute, puisque de toutes façons vous risquez d'avoir un biais, ne différenciez pas. Aussi, différencier plusieurs fois enlève tout potentiel d'interprétation au modèle. Vous aurez beau dire que votre modèle est désormais stationnaire, mais si vous ne pouvez pas l'interpréter, vous n'êtes pas avancé.

Interpréter le modèle après les corrections Un modèle différencié s'interprète comme l'impact d'une variation de la variable indépendante sur la variation de la variable dépendante. Si nos variables sont en log, la variation peut s'interpréter comme une variation en pourcentage (pour un coefficient arbitrairement près de 0). Par ailleurs, il est parfais intéressant d'utiliser les taux de croissance plutôt qu'une première différenciation.

4.2.3 Co-intégration

La co-intégration est une situation rencontrée lorsque deux séries possédant une racine unitaire ont une même tendance stochastique. Par exemple, les taux d'intérêts pour deux obligations de termes différents sont généralement considérés co-intégrés: ils suivent une tendance similaire avec une différence constante (la prime de risque). Soit $\{x_t\}$ et $\{y_t\}$ I(1) (= racine unitaire), si pour un θ donné $y_t - \theta x_t$ est I(0) (=absence de racine unitaire), alors on dit que $\{x_t\}$ et $\{y_t\}$ sont co-intégrés avec le paramètre d'intégration θ .

Pourquoi un test de co-intégration Si $\{x_t\}$ et $\{y_t\}$ sont bel et bien co-intégrés, alors $\hat{\beta}$ de la régression $y_t = \alpha + \beta x_t + e_t$ est convergent et il n'y a pas de correction à apporter. Dans le cas contraire, il faut suivre la démarche donnée pour une racine unitaire et estimer le modèle en différences.

Test de co-intégration On construit $\hat{e}_t = y_t - \hat{\alpha} - \hat{\beta} x_t$ et on teste $\hat{e}_t$ pour une racine unitaire. Il faut utiliser le test Dickey-Fuller Augmenté car, sous H_0 ($\hat{e}_t$ a une racine unitaire) la régression est illusoire et la statistique ne suit pas la loi de DF. Sinon, la démarche et l'interprétation sont identiques à celles pour une racine unitaire.

3.3 Données en panel

Une base de données d'un panel pourrait ressembler à ceci :

Panel	Année	Variable revenu	Variable éducation	Variable n
1	2000	50 000	18	…
1	2001	55 000	20	…
2	2000	45 400	18	…
2	2001	100 000	25	…
…	…	…	…	…

Pour indiquer à Stata que l'on travaille avec des données en panel, il suffit de reprendre la fonction vue en section 4.3 (*tsset)* et d'ajouter la variable de panel avant la variable de temps. Par exemple :

g année
Tsset panel année

Une fois *tsset* déclaré pour des données en panel, il est possible de travailler avec la famille *xt* de Stata. Il existe une telle fonction pour chacun des types de régression : *xtreg, xtlogit, xtprobit, xttobit, xtgls, etc.*

Les données en panel possèdent deux dimensions : une pour les individus (ou une quelconque unité d'observation) et une pour le temps. Elles sont généralement indiquées par l'indice *i* et *t* respectivement. Il est souvent intéressant d'identifier l'effet associé à chaque individu, i.e. un effet qui ne varie pas dans le temps, mais qui varie d'un individu à l'autre. Cet effet peut être fixe ou aléatoire. En plus de la question des effets individuels, la question de la corrélation et de l'hétéroscédasticité dans le cadre des données de panels est adressée. Bien qu'elle ne soit pas adressée ici, la question du biais de sélection doit également être considérée pour les données de panels.

3.3.1 Effets fixes vs. Effets aléatoires

La discussion suivante se concentrera sur la modélisation des effets individuels u_i pour des données en panel de la forme suivante : $Y_{it} = X_{it}\,\beta + u_i + e_{it}$. Cependant, il peut aussi s'avérer intéressant d'identifier l'effet associé à chaque période *t* .On peut inclure des effets temporels δ_t afin de tenir compte des changements dans l'environnement comme, par exemple, de cycles économiques. L'idée est la même que pour les effets individuels, c'est pourquoi nous ne nous y attarderons pas. On peut bien évidemment combiner les deux types d'effets : $Y_u = \gamma + X_u\,\beta + \delta_t + u_i + e_u$. Ces effets, individuels ou temporels, peuvent être captés en ajoutant une variable dichotomique pour chaque individu.

Test de présence d'effets individuels La première étape consiste à vérifier s'il y a bel et bien présence d'effets individuels dans nos données. On peut représenter ces effets par une intercepte propre à chaque individu, u_i. On cherche donc à tester l'hypothèse nulle $H_0 : u_i = 0$ dans la régression $Y_{it} = \gamma + X_{it}\,\beta + u_i + e_{it,}\ e_u \sim$ iid. En Stata, la commande *xtreg* effectue directement cette analyse.

Rappelons qu'au début de l'analyse, on déclare nos données en panel :

tsset variabledepanel variabledetemps
xtreg y x1 x2 ...,fe

Interprétation du Test L'hypothèse nulle de ce test est qu'il y a seulement une intercepte commune, aucun effet individuel. Le résultat est une statistique *F* avec (N-1,NT-N-K-1) degré de liberté. Si on rejette l'hypothèse nulle, alors on doit inclure des effets individuels dans le modèle.

Modélisation du modèle en présence d'effets individuels :

Dans le cas d'un effet fixe, la méthode la plus simple de capter cet effet est de supposer qu'il existe pour chacun de nos groupes et, ainsi, d'ajouter une variable binaire par groupe (sans oublier, comme d'habitude, d'en laisser tomber une). Donc si nous avons cinq groupes et quatre périodes de temps, nous aurons un total de sept variables binaires. Il peut être préférable dans certains cas de ne pas inclure de constante pour comparer tous les groupes entre eux. Dans le dernier exemple, on pourrait ainsi laisser tomber la constante et inclure cinq variables binaires pour les groupes et trois variables binaires pour les années. Ajout manuellement de variables binaires pour chaque groupe et chaque année. Par exemple: régression sur cinq échantillons tirés de 1980, 81, 82 et 83.

** création des variables binaires*
a81 = (annee == 1981)
a82 = (annee == 1982)
a83 = (annee == 1983)
g2 = (groupe == 2)
g3 = (groupe == 3)
g4 = (groupe == 4)
g5 = (groupe == 5)
** régression*
regress y x1 x2 a81 a82 a83 g2 g3 g4 g5

Une autre manière de capter les effets individuels, qui est équivalente à l'ajout de variables dichotomiques, est d'utiliser un estimateur «within», qui s'implémente facilement en STATA. Cet estimateur mesure la variation de chaque observation par rapport à la moyenne de l'individu auquel appartient cette observation :

$Y_{it} - \overline{Y}_i = \beta(X_{it} - \overline{X}_i) + e_{it} - \overline{e}_i$. Les effets individuels sont donc éliminés et l'estimateur de MCO peut être utilisé sur les nouvelles variables.

xtreg y x1 x2 ..., fe

On peut aussi modéliser les effets individuels de façon aléatoire: variant autour d'une moyenne. On suppose le plus souvent qu'ils suivent une loi normale : $u_t \sim N(0,\sigma^2)$. On considère alors que l'erreur du modèle est composée de l'erreur usuelle spécifique à l'observation *i, t* et de l'erreur provenant de l'intercepte aléatoire.

$Y_u = X_u\beta + \varepsilon_u$
$\varepsilon_{tt} = e_{tt} + u_t$
xtreg y x1 x2 ..., re

On doit maintenant choisir quelle modélisation se prête le mieux à nos données. Notons que les effets fixes sont plus généraux que les effets aléatoires puisqu'ils n'imposent pas de structure aux effets individuels. Cependant, on perd N-1 degrés de liberté en modélisant les effets individuels de manière fixe (inclusion implicite de N variables dummies moins l'intercepte générale), ce qui rend l'estimation des coefficients des variables explicatives moins efficientes. Par ailleurs, le coefficient de toute variable explicative qui ne varie pas dans le temps pour un même individu (la race, le sexe…) n'est pas estimable puisque l'estimateur «whitin» l'élimine$(X_{it} - \overline{X}_i = 0)$. On peut donc être tenté de se tourner vers une modélisation aléatoire des effets individuels.

Malheureusement, leur efficacité repose sur une hypothèse cruciale à savoir que, pour que les estimateurs d'effet aléatoires soient non biaisés, il ne doit pas y avoir de corrélation entre les effets aléatoires (u_t) et les variables explicatives.

Le test d'Hausman Le test d'Hausman est un test de spécification qui permet de déterminer si les coefficients des deux estimations (fixe et aléatoire) sont statistiquement différents. L'idée de ce test est que, sous l'hypothèse nulle d'indépendance entre les erreurs et les variables explicatives, les deux estimateurs sont non biaisés, donc les

coefficients estimés devraient peu différer. Le test d'Hausman compare la matrice de variance-covariance des deux estimateurs : $W = (\beta_f-\beta_a)'var(\beta_f-\beta_a)^{-1}(\beta_f-\beta_a)$.

Le résultat suit une loi χ^2 avec K-1 degré de liberté. Si on ne peut rejeter la nulle, i.e. si la p-value est supérieure au niveau de confiance, on utilisera les effets aléatoires qui sont efficaces s'il n'y a pas de corrélation entre les erreurs et les variables explicatives.

xtreg y x1 x2 ..., fe (réalise la régression en supposant des effets fixes)
estimates store fixe (conserve les coefficients)
xtreg y x1 x2 ..., re (réalise la régression en supposant des effets aléatoires)
hausman fixe (calcule W)

3.3.2 Corrélation et hétéroscédasticité

Soit la matrice de la variance-covariance des erreurs. Pour pouvoir utiliser les estimateurs MCO, cette matrice doit respecter la forme suivante :

$$\Omega = \begin{bmatrix} \sigma^2 I_{TxT} & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \sigma^2 I_{TxT} \end{bmatrix}_{NTxNT}$$

On doit donc vérifier les hypothèses d'homoscédasticité et de corrélation. Quatre tests permettent de vérifier si nos données respectent ces hypothèses dans le contexte de données en panels.

En ce qui concerne l'hypothèse d'homoscédasticité (test1 et test2), on doit vérifier si la variance des erreurs de chaque individu est constante : pour tout individu *i*, on doit donc avoir $\sigma^2 = \sigma^2$ pour tout t. La dimension nouvelle des données de panels consiste à s'assurer que la variance est la même pour tous les individus : $\sigma^2 = \sigma^2$ pour tout i.

Pour la corrélation, l'aspect nouveau auquel on doit porter attention concerne la possibilité de corrélation des erreurs entre les individus (test3).On doit aussi vérifier que les erreurs ne sont pas autocorrélées et ce, pour chaque individu (test4).

1. Test d'hétéroscédasticité Pour détecter l'hétéroscédasticité, le raisonnement est le même que celui décrit à la section 4.1 et on utilise sensiblement la même procédure. On peut aussi, comme mentionné dans cette même section, utiliser le test de White. Pour le Test de Breusch-Pagan :

xtreg y x1 x2 ..., fe/re (régression)
predict résidus (récupère les résidus)
gen résidus2 = résidus^2 (génère les résidus carrés)
reg résidus2 x1 x2 ... (régression des résidus sur les variables explicatives)

Si on ne peut rejeter l'hypothèse nulle d'homoscédasticité, alors on a $\sigma^2_{tt}= \sigma^2$, pour tout i, t ce qui implique nécessairement que $\sigma^2_{tt}= \sigma^2_t$ pour tout t et $\sigma^2_t= \sigma^2$ pour tout i. Il n'est alors pas nécessaire de faire le test 2. Si notre modèle ne contient pas d'effets individuels ou s'il contient des effets fixes, on continu l'analyse au test de corrélation (test 3). Cependant, bien que cela soit théoriquement possible, STATA ne permet pas de tester la corrélation si notre modèle inclut des effets aléatoires (on continue donc au test 4). Si on fait l'hypothèse qu'il y a corrélation, il est préférable d'utiliser des effets fixes.

Ayant conclu à l'hypothèse d'homoscédasticité avec un modèle à effets fixe, on continue l'analyse (au test 3) avec la commande : *xtreg y x1 x2 ..., fe* . Par contre, si on conclut à la présence d'hétéroscédasticité, on effectue le test 2, que ce soit avec un modèle à effets fixes ou aléatoires, pour tenter d'obtenir plus d'informations sur la forme de l'hétéroscédasticité. On utilise alors les MCG (GLS en anglais) où

$\hat{\beta}_{MCG} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y$ et $Var(\hat{\beta}_{MCG}) = (X'\hat{\Omega}^{-1}X)^{-1}$

2. Test d'hétéroscédasticité inter-individus Ce test-ci est conçu pour tester l'hypothèse spécifique d'homoscédasticité inter-individus. STATA utilise un test Wald modifié, qui est essentiellement un test F. Sous l'hypothèse nulle, le test suppose que la variance des erreurs est la même pour tous les individus : $\sigma_i^2 = \sigma^2 \forall i = 1,...,N$ et la statistique suit une loi χ^2 de degré de liberté N.

xtgls y x1 x2...,
xttest3

Si la valeur obtenue est inférieure à la valeur critique, on ne peut rejeter l'hypothèse nulle : la variance des erreurs est la même pour tous les individus. Étant donné que nous avions déjà conclue à la présence d'hétéroscédasticité sous une forme quelconque au test 1, on en déduit que nos données ont la structure suivante : homoscédasticité intra-individus $\sigma_{it}^2 = \sigma_i^2 \forall t$

et hétéroscédasticité inter-individus $\sigma_i^2 \neq \sigma^2 \forall i = 1,...,N$

Le rejet de l'hypothèse nulle ne nous permet cependant pas de spécifier d'avantage la structure de l'hétéroscédasticité. On demeure avec la conclusion précédente d'hétéroscédasticité $\sigma^2_{it} \neq \sigma^2$, pour tout i, t, sans pouvoir en dire plus.

3. Corrélation contemporaine entre individus Pour tester la présence de corrélation des erreurs inter-individus pour une même période, i.e. : E (e_{it} e_{jt}) ≠ 0 pour i ≠ j**,** on utilise un test Breusch-Pagan. L'hypothèse nulle de ce test est l'indépendance des résidus entre les individus. Ce test vérifie que la somme des carrés des coefficients de corrélation entre les erreurs contemporaines est approximativement zéro. Puisqu'il est seulement nécessaire de tester ceux sous la diagonale, la statistique résultante suit une χ^2 de degré de liberté N(N-1)/2, équivalent au nombre de restrictions testées.

xtreg y x1 x2 ..., fe /ou xtgls y x1 x2...,

xttest2

Si la valeur obtenue est supérieure à la valeur critique, on rejette l'hypothèse nulle: les erreurs sont corrélées de manière contemporaine. On corrige pour la corrélation en utilisant la fonction :

xtgls y x1 x2 ...,panel(corr)

4. Autocorrélation intra-individus On cherche à vérifier si les erreurs sont autocorrélées E (e_{it} e_{is}) ≠ 0 pour t ≠ s de forme autorégressive (AR1) : $e_{it} = \rho e_{it-1} + z_{it} \forall i = 1,...N$. S'il y a de l'autocorrélation, les matrices identités le long de la diagonale sont remplacées par des matrices de la forme suivante :

$$\Delta = \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}_{TxT, pour T=3}$$

STATA réalise un test Wald dont l'hypothèse nulle est celle d'absence d'autocorrélation des erreurs. Si on rejette cette hypothèse, i.e. si la valeur obtenue est supérieure à la valeur critique, les erreurs des individus sont autocorrélées.

xtserial y x1 x2 ...

On ajuste la forme de la matrice Ω afin de tenir compte de l'autocorrélation dans les erreurs des individus en utilisant soit :

xtgls y x1 x2 ...,panel(...)corr(ar1)

soit :

xtregar y x1 x2 ...,re/fe

Correction : résumé Donc en résumé, s'il n'y a aucun effet individuel, pas d'hétéroscédasticité ni de corrélation, les estimateurs MCO usuels sont valides. On effectue alors du « pooling », c'est-à-dire qu'on considère les données comme N*T observations non-panélisées et on effectue une régression standard :

reg y x1 x2 ...

S'il y a des effets individuels mais pas d'hétéroscédasticité ni de corrélation, on utilise la commande *xtregy x1 x2 ...,re/fe* qu'on corrige si nécessaire pour l'autocorrélation :

xtregar y x1 x2 ...,re/fe

Finalement, dans les autres cas, on utilise des variantes de la fonction *xtgls***.** Cette fonction estime le modèle par MCG et permet de combiner les diverses conclusions aux tests précédents. Les estimateurs β sont estimés en ajustant la matrice de variancecovariance des erreurs Ω afin de tenir compte de la présence d'hétéroscédasticité intra et inter individus et/ou autocorrélation inter-individus de type autorégressif de premier ordre et/ou corrélation inter-individus.

Il suffit de spécifier un des trois choix de structure de variance de panel : (iid | heteroskedastic | correlated) combiné avec un des trois choix de structure de corrélation intra-individu : (independent | ar1 | psar1). Le choix de ar1 signifie qu'on suppose un coefficient d'autorégression ρ commun pour tous les individus tandis que le choix de psar1 permet aux individus d'avoir des coefficients différents $\rho_i \neq \rho_j \forall i \neq j$. Cependant, le choix d'un ρ commun permet une meilleure estimation des β, si cette restriction est correcte, ce qui est le but de l'analyse.

xtgls y x1 x2 ...,panels(iid ou hetero ou corr) corr(independent ou ar1 ou psar1)

3.4 Variables instrumentales, doubles moindres carrés et test d'endogénéité

Lorsqu'une variable "indépendante" est corrélée avec le terme d'erreur, les hypothèses classiques du modèle linéaire sont violées et on se retrouve face à un problème d'endogénéité. Dans ces cas, on peut faire appel à l'estimateur de variables instrumentales (VI) ou aux doubles moindres carrés ordinaires (DMCO).

3.4.1 Estimateur Variables Instrumentales

Soit Z, une matrice de VI et X, la matrice originale. L'estimateur VI est donné par:

$$\hat{\beta}_{(VI)} = (Z'X)^{-1}Z'y$$

et l'estimateur VI de la covariance par:

$$\hat{\sigma}^2(Z'X)^{-1}(Z'Z)(X'Z)^{-1}$$

où

$$\hat{\sigma}^2 = 1/T\,(y - X\beta_{(IV)})'(y - X\beta_{(IV)}).$$

ou, lorsque J > K (J étant le nombre de VI et K le nombre de variables indépendantes), par:

$$\hat{\beta}_{(IV)} = [X'Z(Z.'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'y.$$

$$\hat{\sigma}^2\,[X'Z(Z'Z)^{-1}Z'X]^{-1}.$$

ivreg permet de faire directement une régression par DMCO. On inscrit donc :

ivreg variabledependante variablesindependantes (variabledependante = variable(s)intrumentale options

où options peut prendre les mêmes valeurs que pour *regress*, ainsi que *first* qui affiche les résultats de la première régression.

ex:

ivreg y1 z1 z2 (y2=x1), r first

predict peut être utilisé après *ivreg*

3.4.2 DMCO

Les trois hypothèses soutenant les DMCO :

1-Le terme d'erreur ne doit pas être corrélé avec la variable instrumentale.

2-La variable dont on suppose souffrir d'endogénéité doit être fortement corrélée avec la variable instrumentale, mais pas corrélée avec le terme d'erreur.

3-La variable instrument doit être différente de la variable qui souffre d'endogénéité, même à un multiple près.

Les doubles moindres carrés ordinaires permettent d'effectuer une régression en substituant la variable qui potentiellement souffre d'endogénéité par une variable instrumentale. Voici un exemple :

Les subventions aux entreprises (x) on un impact sur la croissance du PIB (y). Dû à des contraintes de disponibilité, nous n'arrivons pas à trouver les données sur les subventions. Donc, supposant qu'il existe un lien positif entre la variable subvention et efficacité, la variable efficacité serait lier au terme d'erreur, dû à l'omission d'une variable pertinente. Dans ce cas-ci, la variable instrumenté serait efficacité, car elle souffre d'endogénéité. Une variable instrumentale possible serait la taille des entreprises. On choisi cette dernière parce qu'intuitivement, on suppose que le nombre d'employé n'est pas lié à la variable subvention. De plus, la variable taille d'entreprise est liée à la variable efficacité (plus l'index efficacité est élevé, le nombre d'employé nécessaire diminue).

Soit le modèle suivant:

$y1 = \beta0 + \beta1x1 + \beta2x2 + \beta3y2 + u$

et soit z une VI de y2.

Comme leur nom l'indique, les DMCO se font en deux étapes.

1. Estimation de la variable endogène: Régression de y2 sur toutes les variables indépendantes (x1 et x2 ici) et la/les VI pour y2 (z ici).

   On récupère $\hat{y}2$, l'estimation linéaire de y2.

2. Régression du modèle avec $\hat{y}2$: Régression de y1 sur une constante, x1, x2 et $\hat{y}2$. Cette dernière régression ne souffrant plus d'endogénéité, les $\hat{\ }\beta$ ainsi obtenus sont non-biaisés.

Vous pouvez effectuer les deux régressions par vous même ou utiliser la fonction *ivreg* à la section précédente.

3.4.3 Test d'endogénéité

Le test de Hausman permet de vérifier s'il existe bel et bien une différence entre l'estimateur de variable instrumentale et l'estimateur MCO, vérifiant ainsi s'il y a bel et bien endogénéité des variables (si les deux estimateurs sont consistants, ils seront asymptotiquement égaux). Sous H_0, la statistique de Hausman est:

$$H = [\beta_{(V I)} - b]'[\sigma^2 [(X'Z(Z'Z)^{-1}Z'X]^{-1} - \sigma^2 (X'X)^{-1}]^{-1}[\beta_{(V I)} - b] \sim \chi^2(J)$$

La fonction *hausman* effectue le test de spécification d'Hausman. Estimation du modèle moins efficient, mais convergent (VI ici) :

hausman, save

Estimation du modèle efficient, mais peut-être pas convergent (MCO ici) :

hausman

Options: constant, indique que la constante doit être inclue dans la comparaison des deux modèles.

ex:

ivreg y1 z1 z2 (y2=x1)
hausman, save
reg y1 z1 z2 y2
hausman, constant

3.5 Estimateurs du maximums de vraissemblance (EMV)

La fonction de vraisemblance est la probabilité jointe des observations étant donné les paramètres d'intérêts, i.e.:

$$L(\theta|y) = f(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

L'estimateur du maximum de vraisemblance (EMV) a pour but de choisir le vecteur de paramètres θ qui maximise la fonction de vraisemblance, i.e. pour lequel les données observées sont les plus probables. Pour simplifier les choses, la fonction de log-vraisemblance, $L(\theta\ |y)$, est généralement utilisée. Prenons l'exemple d'un échantillon normalement distribué, de moyenne 0 de variance σ^2:

$$\begin{aligned} f(y|X, \beta, \sigma^2) &= \prod_{t=1}^{T} (2\pi\sigma^2)^{-1/2} \exp[(y_t - x_t'\beta)^2] \\ &= (2\pi\sigma^2)^{-T/2} \exp\left[-\frac{(y - X\beta)'(y - X\beta)}{2\sigma^2}\right]. \end{aligned}$$

La log-vraisemblance est

$$\mathcal{L}(\beta, \sigma^2) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log\sigma^2 - \frac{(y - X\beta)'(y - X\beta)}{2\sigma^2}.$$

Les CPO sont:

$$\frac{\delta lnL}{\delta \beta} = \frac{(y - X\beta)(y - X\beta)}{2\sigma^2}$$

$$\frac{\delta lnL}{\delta \sigma^2} = -\frac{T}{2\sigma^2} + \frac{(y - X\beta)'(y - X\beta)}{2\sigma^4}$$

Ce qui nous permet de trouver

$$\hat{\beta} = (X'X)X'y$$

$$\hat{\sigma^2} = \frac{(y - X\beta)'(y - X\beta)}{T} = \frac{\hat{e}'\hat{e}}{T}$$

*ml p*ermet de faire une estimation par maximum de vraisemblance pour une équation donnée. Cette fonction étant fort complexe et très peu utilisée dans le cadre des problèmes abordés dans ce guide, il est laissé à la discrétion du lecteur le soin de consulter l'aide de Stata à son sujet.

Stata estime automatiquement par maximum de vraisemblance les régressions qui doivent être traitées par EMV. Les modèles à variable dépendante qualitative, les modèles de durée et les ARIMA sont des exemples de tels cas.

3.6 Moindres carrés généralisés

La méthode des moindres carrés généralisés (MCG) cherche à modéliser la fonction de la variance. Nous obtenons alors l'estimateur MCG

$\hat{\beta}^{MCG} = (X'V^{-1}X)^{-1}X'V^{-1}y$

ou encore

$\hat{\beta}^{MCG} = (X'W^{-1}X)^{-1}W'V^{-1}y$

et sa variance est

$var[\hat{\beta}] = \sigma^2 (X'V^{-1}X)^{-1}$.

Où V et W sont égaux à

$$W = \sigma^2 \begin{bmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_n \end{bmatrix} \equiv \sigma^2 V$$

La fonction *vwls* permet de faire une régression linéaire pondérée par la variance. On inscrit : *vwls variabledependante variablesindependantes [poids], options*

Ici, l'option serait *sd (nomvariabl)*. Elle fournit une estimation de l'écart-type de la variable dépendante. Par exemple:

vwls y x1 x2, sd(sigma2ch)

où *sigma2ch* est une estimation de l'écart-type de y.

predict peut être utilisé après *vwls*

3.7 Le logit et le tobit

Le logit est un modèle a essentiellement la même fonction que le probit et repose sur les mêmes principes, mais a la différence du probit, il utilise un fonction de répartition logistic pour calculer l'effet de xi sur la probabilité associée à une valeur donnée de la valeur latente (y^*_i). Les économistes préfèrent généralement utiliser le modèle probit car le logit n'est généralement pas problématique avec les modèles univariés.

logit variabledépendante variableindépendante

Options: *logit* possède en grande partie les mêmes options que *regress*.

Un *tobit* est essentiellement un modèle dont les données sont censurées. Comme le *probit*, le *tobit* suit une loi normale.

tobit variabledépendante variableindépendante

Options: ll(#), ul(#): indiquent respectivement que les données sont tronquées à gauche ou à droite. Une ou les deux de ces options doivent être spécifiées. # indique le point de troncation. Si # n'est pas précisé, Stata suppose qu'il s'agit respectivement de la valeur minimum et de la valeur maximum. Les autres options de *tobit* sont en grande partie commune avec *regress*. Par exemple:

tobit y x1 x2 x3 x4, ll(0)

3.8 Biais de sélection

Un biais de sélection existe si la présence d'une observation dans l'échantillon est déterminée par un ou des facteurs extérieurs. Si c'est le cas, il faut utiliser la méthode de Heckman pour corriger le biais de sélection. Mathématiquement, ce biais peut être exprimé comme:

$Yi=xi\beta + ei$

$Z^*i = \alpha i\gamma+ui$

Z*i est une variable latente et i y est observable si et seulement si i Z*i > 0, i.e. yi sera observé si un niveau « d'utilité » arbitraire expliqué par un ou des facteurs extérieurs est atteint.

La détection d'un biais de sélection est intuitive: existe-t-il des facteurs qui pourraient influencer la nature aléatoire de l'échantillon? Peut-on les caractériser? Si on détermine qu'il y a biais de sélection, il faut le corriger par la méthode de Heckman. La commande de Stata pour la méthode de Heckman est *Heckman*.

L'idée est de modéliser l'équation de sélection (Z*i) qui agît comme un probit: si Z*i > 0, alors z = 1 (sinon z = 0) et on observe la donnée. On corrige alors l'estimation de l'espérance conditionnelle de y par un « facteur de biais » (l'inverse du ratio de Mills).

Attention, la sélection doit être expliquée par un ou des facteurs extérieurs, ils ne doivent pas se retrouver dans le modèle original, sinon le tout se simplifie et revient à faire un MCO.

Annexe A : Résumé des fonctions dans Stata

Fonction	Abréviation	Description	Forme
Importation de Données			
infile	inf	Importe les données d'un fichier.	`infile` *nom_des_variables* `using` *nomdefichier*
insheet		Importe les données d'un fichier (séparateurs: tabulations ou virgules).	`insheet using` *nom_de_fichier*
Transformation de Variables			
generate	g	Crée une nouvelle variable.	`generate` *nouvelle_variable* = *expression*
replace		Remplace une variable existante.	`replace` *variable_existante* = *expression*
abs		Valeur absolue.	`abs(`x`)`
exp		Exponentiel.	`exp(`x`)`
log		Logarithme naturel.	`log(`x`)`
max		Renvoie l'argument possédant la valeur la plus élevée.	`max(`$x_1,\ldots,x_n$`)`
min		Renvoie l'argument possédant la valeur la moins élevée.	`min(`$x_1,\ldots,x_n$`)`
mod		Modulo de x par rapport à y.	`mod(`x`,`y`)`
sqrt		Racine carrée.	`sqrt(`x`)`
sum		Somme de tous les éléments de x. 46	`sum(`x`)`

Fonctions Matricielles

matrix	mat	Crée ou modifie une matrice.	`matrix` *nom_de_la_matrice* = *expression*
matrix get		Permet d'obtenir copie d'une matrice système.	`matrix` *variable* = `get(`*matrice_système*`)`
mkmat		Transforme des variables en vecteurs/matrice.	`mkmat` *nom(s)_de_variable(s)*, `matrix` [(*nomdelanouvellematrice*)]
svmat		Transforme les colonnes d'une matrice en variables.	`svmat` *matrice*, [`names(`*nom_col*1, *nom_col*2,...)]
colsof		nombre de colonnes d'une matrice.	`colsof(`A`)`
det		Déterminant d'une matrice.	`det(`A`)`
diag		Matrice diagonale $n \times n$, avec pour diagonale les éléments de V.	`diag(`V`)`
el		Élément a_{ij} d'une matrice.	`el(`A,i,j`)`
I		Matrice identité $n \times n$.	`I(`n`)`:
inv		Inverse d'une matrice carrée.	`inv(`A`)`
rowsof		Nombre de rangées d'une matrice.	`rowsof(`A`)`
vecdiag		Extrait la diagonale d'une matrice carrée sous forme de vecteur.	`vecdiag(`A`)`

Fonctions Diverses

graph	gr	Trace un graphique.	`graph` *nomdesvariables*, [*typedegraphique*, *autresoptions*]
list		Affiche à l'écran les variables spécifiées.	`list` [*nom(s)_de_variable(s)*]
log		Enregistre la session.	`log using` *nom_de_fichier*
more		Active ou désactive l'affichage de `--more--`.	`more on/off`
set matsize	set mat	Fixe la taille maximale des matrices.	`set matsize` #

Fonctions Diverses (suite)

uniform		Donne une valeur aléatoire entre 0 et 1 (ditribution uniforme sur [0,1)).	`uniform()`
tsset		Déclaration de séries temporelles/Données pannel.	`tsset` *variable_de_temps*
l		Opérateur retard.	`l#.`*variable*
f		Opérateur avance.	`f#.`*variable*

Fonctions Économétriques

regress	reg	Effectue une régression linéaire par MCO.	`regress` *var_dep* [*vars_inds*]
predict		Calcule les valeurs prédites, les résidus, etc.	`predict` *nouvelle_variable*, options
test	t	Effectue des tests d'hypothèse.	`test` [*expression1* = *expression2*]
ivreg		Effectue une régression par DMCO.	`ivreg` *var_dep vars_inds* (*var_dep* = *VI*), options
hausman		Effectue le test de spécification d'Hausman.	`hausman` / `hausman, save`
vwls		Effectue une régression pondérée par la variance (FGLS).	`vwls` *var_dep vars_inds* [*poids*], options
probit	prob	Estime un modèle probit.	`probit` *var_dep vars_inds*
logit		Estime un modèle logit.	`logit` *var_dep vars_inds*
tobit		Estime un modèle tobit.	`tobit` *var_dep vars_inds*, [ll(#)] [ul(#)]
dfuller		Effectue le test de Dickey-Fuller augmenté.	`dfuller` *nom_de_variable*, options
pperron		Effectue le test de Phillips-Perron.	`pperron` *nom_de_variable*, options
corrgram		Produit une table des autocorrélations et des autocorrélations partielles.	`corrgram` *nom_de_variable*, *option*
xtreg		Effectue une régression sur des données panel.	`xtreg` *var_dep vars_inds*, [fe] [re] [mle]

Annexe B: Exemple d'un programme Stata complet

Exemple d'un programme Stata complet pour faire une régression linéaire par les moindres carrés ordinaires à partir de l'Enquête sur la population active 2002 de statistiques canada. Pour trouver les données, voir :
http://sherlock.crepuq.qc.ca/cgi-bin/sherlock.pl?langue=F&action=information_enquete&id_enquete=ENQ-10310
.

```
clear

set memory 800m

log using wage.log, replace

insheet using "C:\Documents and Settings\p0678264\Bureau\EXTRACT.TAB"

destring ftptmain tenure hrlyearn union permtemp estsize, replace

saveold "C:\Documents and Settings\p0678264\Bureau\donnee.dta", replace

svyset [pweight = fweight]

g age1519 = (age_12==1)
g age2024 = (age_12==2)
g age2529 = (age_12==3)
g age3034 = (age_12==4)
g age3539 = (age_12==5)
g age4044 = (age_12==6)
g age4549 = (age_12==7)
g age5054 = (age_12==8)
g age5559 = (age_12==9)
g age6069 = (age_12==10 | age_12==11)

g femme = (sex==2)
g homme = (sex==1)

g marieunionlibre = (marstat==1 | marstat==2)
g celibataire = (marstat==6)
g veufs = (marstat==3)
g separedivorce = (marstat==4 | marstat==5)

g eduprimaire = (educ90==0)
g educsecondpartielles = (educ90==1)
g diplomesecondaire = (educ90==2)
g etudespostsec = (educ90==3)
g diplomepostsec = (educ90==4)
g bacc = (educ90==5)
```

```
g diplomedeuxiemecycle = (educ90==6)

g pmois = (tenure>=1 & tenure<=6)
g dmois = (tenure>6 & tenure<=12)
g pans = (tenure>12 & tenure<=60)
g dans = (tenure>60 & tenure<=120)
g tans = (tenure>120 & tenure<=240)

g tempsplein = (ftptmain==1)
g tempspart = (ftptmain==2)

g secteurpublic = (cowmain==1)
g secteurprive = (cowmain==2)

g administrationprovplus = .
replace administrationprovplus = 1 if (naics_43==42 | naics_43==36 |
naics_43==37)
replace administrationprovplus = 0 if (naics_43==1 | naics_43==2 |
naics_43==3 | naics_43==4 | naics_43==6 | naics_43==7 | naics_43==8 |
naics_43==9 | naics_43==10 | naics_43==11 | naics_43==12 | naics_43==13
| naics_43==14 | naics_43==15 | naics_43==16 | naics_43==17 |
naics_43==18 | naics_43==19 | naics_43==20 | naics_43==21 |
naics_43==22 | naics_43==23 | naics_43==24 | naics_43==25 |
naics_43==26 | naics_43==27 | naics_43==29 | naics_43==30 |
naics_43==31 | naics_43==32 | naics_43==33 | naics_43==34 |
naics_43==35 | naics_43==39 | naics_43==40 | naics_43==28 |
naics_43==38| naics_43==41 | naics_43==43 | naics_43==5)

g autrepub = .
replace autrepub = 1 if (naics_43==41 | naics_43==43 | naics_43==5 |
naics_43==38 | naics_43==28)
replace autrepub = 0 if (naics_43==1 | naics_43==2 |  naics_43==3 |
naics_43==4 | naics_43==6 | naics_43==7 | naics_43==8 | naics_43==9 |
naics_43==10 | naics_43==11 | naics_43==12 | naics_43==13 |
naics_43==14 | naics_43==15 | naics_43==16 | naics_43==17 |
naics_43==18 | naics_43==19 | naics_43==20 | naics_43==21 |
naics_43==22 | naics_43==23 | naics_43==24 | naics_43==25 |
naics_43==26 | naics_43==27 | naics_43==29 | naics_43==30 |
naics_43==31 | naics_43==32 | naics_43==33 | naics_43==34 |
naics_43==35 | naics_43==39 | naics_43==40 | naics_43==36 |
naics_43==37 | naics_43==42)

g prive = .
replace prive = 1 if (naics_43==1 | naics_43==2 | naics_43==3 |
naics_43==4 | naics_43==6 | naics_43==7 | naics_43==8 | naics_43==9 |
naics_43==10 | naics_43==11 | naics_43==12 | naics_43==13 |
naics_43==14 | naics_43==15 | naics_43==16 | naics_43==17 |
naics_43==18 | naics_43==19 | naics_43==20 | naics_43==21 |
naics_43==22 | naics_43==23 | naics_43==24 | naics_43==25 |
naics_43==26 | naics_43==27 | naics_43==29 | naics_43==30 |
naics_43==31 | naics_43==32 | naics_43==33 | naics_43==34 |
naics_43==35 | naics_43==39 | naics_43==40)
replace prive = 0 if (naics_43==36 | naics_43==37 | naics_43==28 |
naics_43==38| naics_43==41 | naics_43==42 | naics_43==5 | naics_43==43)

g couverturesyndicale = (union==1 | union==2)
g noncouvert = (union==3)
```

```
g cadres = (soc91_47==1)
g professionel = (soc91_47==5 | soc91_47==22 | soc91_47==23 |
soc91_47==24)
g personneldebureau = (soc91_47==10)
g sante = (soc91_47==13 | soc91_47==14 | soc91_47==15 | soc91_47==16)
g education = (soc91_47==18)
g hotellerierestauration = (soc91_47==26 | soc91_47==29)
g protection = (soc91_47==28 | soc91_47==47)

g permanent = (permtemp==1)
g saisonnier = (permtemp==2)
g temporairecontractuel = (permtemp==3)
g occasionnelautre = (permtemp==4)

g moinsde20employes = (estsize==1)
g entre20et99employes = (estsize==2)
g entre100et500employes = (estsize==3)
g plusde500employes = (estsize==4)

g salaire = hrlyearn

g lnsalaire = log(hrlyearn)


regress lnsalaire administrationprovplus autrepub homme age2024 age2529
age3034 age3539 age4044 age4549 age5054 age5559 age6069 marieunionlibre
veufs separedivorce educsecondpartielles diplomesecondaire
etudespostsec diplomepostsec bacc diplomedeuxiemecycle dmois pans dans
tans tempsplein couverturesyndicale professionel personneldebureau
sante education hotellerierestauration protection saisonnier
temporairecontractuel occasionnelautre entre20et99employes
entre100et500employes plusde500employes [pweight = fweight] , robust


log close
```