Guide d'économétrie appliquée à l'intention des étudiants du cours ECN 3950

(guide d'application informel de ce que vous avez vu pendant le bac)

Simon Leblond
Isabelle Belley-Ferris
Département de sciences économiques
Université de Montréal

Octobre 2004

Table des matières

Introduction

Maintenant le cours d'économétrie appliquée terminé, il s'agit de mettre tous vos acquis économétriques en application dans le cadre de ECN 3950. Ce guide a donc pour but de vous accompagner dans la partie économétrique de votre travail en vous aidant à répondre aux questions Quoi? Comment? et Pourquoi? dans un contexte appliqué (et parfois simplifié...).

En entreprenant un travail d'économie faisant appel à l'économétrie, vous devrez vous demander: quelle est la question à laquelle je veux répondre? quel modèle dois-je utiliser pour y répondre? quels tests dois-je effectuer pour m'assurer que mon modèle est bon? quelles corrections dois-je apporter au modèle? et, finalement, qu'est-ce que mes résultats veulent dire? Ces questions seront adressées dans le cadre de trois types de modèles: les micro-données, les séries chronologiques et les données en panel.

Les sujets discutés dans ce document ont été déterminés en concertation avec M. Vaillancourt d'après les problèmes rencontrés par les étudiants du cours à l'hiver 2004. Les explications données ici sont inspirées surtout de mes notes de cours d'économétrie du bac et de la maîtrise. Je remercie donc les professeurs de ces cours, MM. McCausland, Meddahi, Montmarquette et Perron, ainsi que Mme Gonçalves. Je demeure par ailleurs responsable de toute erreur dans ce document.

1. Micro-données

1.1 Hétéroscédasticité

L'hétéroscédasticité qualifie des données qui n'ont pas une variance constante, i.e. $\mathrm{Var}(e) \neq \sigma_e^2$. L'hétéroscédasticité ne biaise pas l'estimation des coefficients, mais l'inférence habituelle n'est plus valide puisque les écarts-types trouvés ne sont pas les bons. L'hétéroscédasticité est une situation rencontrée fréquemment dans les données, il est donc important de savoir la détecter et la corriger.

Détection de l'hétéroscédasticité Plusieurs tests se ressemblant existent pour détecter l'hétéroscédasticité. Deux de ces tests, le test de Breusch-Pagen et le test de White sont décrits dans le guide d'économétrie appliqué pour Stata à la section 5.1. L'idée générale de ces tests est de vérifier si le carré des résidus peut être expliqué par les variables du modèle. Si c'est le cas, il y a hétéroscédasticité.

Interprétation des résultats des tests d'hétéroscédasticité Les deux tests mentionnés plus haut utilisent un test *F*. Ce test et son interprétation sont décrits ici de façon générale. L'étudiant pourra donc s'y référer chaque fois qu'un test *F* est utilisé.

Comme c'est le cas pour tous les tests statistiques, on doit comparer la valeur obtenue aux valeurs critiques de la statistique concernée. La statistique *F* 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.[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. $Prob[F_{q,n-k} \leq f]=0.95$. 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 (implicitement, on l'accepte).

Note, les résultats de statistiques sont souvent donnés sous la forme de « p-value », un nombre entre 0 et 1 qui indique la probabilité sous H_0 d'obtenir la valeur trouvée.
Ainsi, si le « p-value » est sous le α désiré, 5% par exemple, on rejette l'hypothèse nulle. Un « p-value » de 0.0000 rejette très fortement l'hypothèse nulle.

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, i.e. les variables du modèle n'expliquent pas la variance observée donc il y a homoscédasticité. L'hypothèse alternative est l'hypothèse d'hétéroscédasticité. Ainsi, si on rejette l'hypothèse nulle (« p-value » < alpha), on peut conclure à la présence d'hétéroscédasticité.

Stata affiche toujours le « 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.

Correction de l'hétéroscédasticité Il existe deux solutions au problème d'hétéroscédasticité :

1. paramétriser la matrice de variance-covariance des erreurs (MCG);
2. utiliser les MCO et corriger les écarts-types par la méthode d'Eicker-White.

Seule la deuxième solution est discutée ici en raison de sa simplicité. En fait, dans Stata, il suffit d'ajouter l'option 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.

[1] Le test F est expliqué en détails dans Wooldridge à la page 140

1.2 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:

$$y_i = X_i\beta + e_i$$

$$z*_i = \alpha_i\gamma + \mu_i$$

$z*_i$ est une variable latente et y_i est observable si et seulement si $z*_i > 0$, i.e. y_i 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 (!), elle est décrite à la section 9.4 du guide d'économétrie appliqué pour Stata. 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.

2. Séries chronologiques

2.1 Qu'est-ce qu'une série chronologique?

Les séries chronologiques se distinguent des données en coupe transversale par le fait qu'elles possèdent un ordre...chronologique! Une série chronologique est en fait 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 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.

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

- $E[y_t] = \mu$ (l'espérance ne dépend pas de t)
- $var[y_t] = \sigma^2$ (la variance ne dépend pas de t)
- $cov[y_t, y_s] = \gamma_k, k = t - s$ (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:

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.

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. La section suivante traite de cette possibilité.

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: $var(y_t) = t\sigma_\mu^2$. 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.

2.3 Procédure pour stationnariser une série chronologique

2.3.1 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 États-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.

Test de Chow Ce test est décrit à la section 5.2 du guide d'économétrie appliqué pour Stata. 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*.

Interprétation du test de Chow Le résultat du test de Chow est un test *F*. Ce test est expliqué plus tôt dans ce guide pour l'hétéroscédasticité. 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.

2.3.2 Racine unitaire

On désire s'assurer que la série n'est pas parfaitement autocorrélée, i.e. $\rho \neq 1$ dans $y_t = \alpha + \rho y_{t-1} + e_t$ ou, de façon équivalente, $\theta \neq 0$ dans $\Delta y_t = \alpha + \theta y_t + \in_t$. 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.

Test de Dickey-Fuller[2] Ce test est décrit à la section 11.4.1 du guide d'économétrie appliqué pour Stata. 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:

$$\Delta y_t = \mu + \beta t + \theta y_{t-1} + \epsilon_t, \epsilon_t \sim iid(0, \sigma^2)$$

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)[3]. 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. 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).

Interpréter les tests de racine unitaire Vous avez finalement réussi à vous décider sur un model à tester et votre logiciel économétrique 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. $y_t = \alpha + \rho y_{t-1} + e_t$ devient donc $\Delta y_t = \alpha + \theta y_t + \epsilon_t$. On voit bien que si l'hypothèse nulle tient, $\theta = 0$ et le terme disparaît du modèle.

[2] Cette section s'appuie beaucoup sur les notes de cours de ECN 6228 donné par Benoit Perron à l'hiver 2004. Toute erreur m'est par ailleurs entièrement imputable.
[3] On peut également utilisé un test de Phillips-Perron dans ce cas, mais ce test n'est pas discuté ici. La commande pour ce test est discutée à la section 11.4.1 du guide d'économétrie appliqué pour Stata.

Deux mises en garde:

- 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. Ainsi, si notre modèle cherche à trouver les déterminants du chômage et qu'on a dû le différencier, on pourrait interpréter le résultat comme « une hausse de croissance du PIB a un impact négatif sur la croissance du taux de chômage ». Si nos variables sont en log, la variation peut s'interpréter comme une variation en pourcentage (pour un coefficient arbitrairement près de 0).

2.3.3 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), si pour un θ donné $y_t - \theta x_t$ est I(0), 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 ADF 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. Données en panel

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.

3.1 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 : $Y_{it} = \gamma + X_{it}\beta + u_i + e_{it}$ .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 : $Y_{it} = \gamma + X_{it}\beta + \delta_t + u_i + e_{it}$. 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 $Y_{it} = \gamma + X_{it}\beta + u_i + e_{it}, e_{it} \sim iid$. En Stata, la commande **xtreg** effectue directement cette analyse.

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

tsset *variable de panel variable de temps*
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

Effets fixes :
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*

Effets aléatoires :
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 : $u_i \sim N(0,\sigma^2)$. 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.

$$Y_{it} = X_{it}\beta + \varepsilon_{it}$$

$$\varepsilon_{it} = e_{it} + u_i$$

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_i) 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 :

$$W = (\beta_f - \beta_a)'\mathrm{var}(\beta_f - \beta_a)^{-1}(\beta_f - \beta_a)$$

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)*

3.2 Corrélation et hétéroscédasticité

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

$$\Omega = \begin{bmatrix} \sigma^2 \mathrm{I}_{TxT} & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \sigma^2 \mathrm{I}_{TxT} \end{bmatrix}_{NTxNT}$$

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 $\sigma_{it}^2 = \sigma_i^2$ pour tout t. La dimension nouvelle des données de panels consiste à s'assurer que la variance est la même pour tous les individus : $\sigma_i^2 = \sigma^2$ 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).

Test

1. **Test d'hétéroscédasticité** Pour détecter l'hétéroscédasticité, le raisonnement est le même que celui décrit à la section 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 $\sigma_{it}^2 = \sigma^2$ pour tout i,t ce qui implique nécessairement que $\sigma_{it}^2 = \sigma_i^2$ pour tout t et $\sigma_i^2 = \sigma^2$ 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).

   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*** .Tandis qu'avec un modèle à effets aléatoires, STATA ne permet pas de tester la corrélation (test3), bien que cela soit théoriquement possible (on continue donc au test 4).

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

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 $\sigma_{it}^2 \neq \sigma^2$ 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. : $E(e_{it}e_{jt}) \neq 0$ pour $i \neq j$, 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 $E(e_{it}e_{is}) \neq 0$ pour $t \neq s$ 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 **xtreg*y 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 variance-covariance 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)