Manuel d'initiation à Stata (Version 8)

Kangni KPODAR

Janvier 2005

Centre d'Etudes et de Recherches sur le Développement International
65 Boulevard F. Mitterrand
63000 Clermont Ferrand
France

Ce document a été élaboré dans le cadre des Travaux Dirigés d'économétrie que j'assure au CERDI. Je remercie Christophe COTTET, Florent BRESSON et Kelly LABAR pour leurs commentaires. Toute erreur dans ce document reste de ma seule responsabilité. Merci de m'envoyer par correspondance les corrections ou les commentaires éventuels. Contact : roland.kpodar@u-clermont1.fr

Avertissements

Ce manuel est rédigé sur la base de la version 8 de Stata, les commandes et les syntaxes peuvent ne pas être les mêmes sur une version antérieure ou ultérieure.

Ce manuel n'aborde pas plusieurs aspects de Stata comme la programmation, la création de graphiques, l'estimation des équations simultanées, l'utilisation des techniques d'estimation en série temporelles, etc. Par ailleurs, pour les commandes présentées dans ce document, seule leur utilisation de base est décrite. Pour plus de détails sur une commande précise, les manuels officiels de Stata restent indispensables.

Tables des matières

PARTIE I : Présentation des logiciels Stat Transfer 7 et Stata 8 ..... **5**
1. Présentation de Stat Transfer 7 5
2. Présentation de Stata 8 9
2.1. Aspect général de l'interface Windows de Stata 8 9
2.2. La barre d'outils de Stata 8 11
PARTIE II : Travailler avec Stata 8 ..... **14**
1. Les fonctions et les expressions 14
1.1. Les opérateurs arithmétiques 14
1.2. Les expressions by, if et in 14
1.3. Les opérateurs de relation 16
1.4. Les fonctions 16
1.5. Les opérateurs logiques 17
2. Commandes de gestion des variables 17
2.1. Création de nouvelles variables : les commandes generate et egen 17
2.2. Autres commandes relatives à la gestion des variables 19
2.3. Abréviations des noms des variables et des commandes 20
2.4. Mettre des étiquettes pour les variables 21
3. Fusion de bases de données 22
4. Création d'un fichier do 24
5. Les statistiques descriptives 25
5.1. La commande summarize 25
5.2. La commande tabulate 26
5.3. Les coefficients de corrélation 27
6. Tests sur la moyenne, la variance et la distribution des variables 28
6.1. Test de comparaison de moyennes 28
6.3. Test sur la distribution de deux variables 30
7. Les régressions sur données transversales : 30
7.1. Les moindres carrés ordinaires (MCO) 30
7.2. La méthode des doubles moindres carrés 34
7.3. Les tests 36
7.3.1. Test de normalité des résidus 36
7.3.2. Test de Ramsey Reset 36
7.3.3. Test d'hétéroscédasticité 37

7.3.4. Test de Chow 37
7.3.5. Test d'endogénéité 39
7.3.6. Test de validité des instruments 40
7.3.7. Test sur les coefficients des variables 42
8. Les régressions sur données de panel 43
8.1. La commande collapse 43
8.2. Les statistiques descriptives 44
8.3. Le modèle à effets fixes 44
8.4. Le modèle à effets aléatoires 47
8.5. Le test de Hausman 49
8.6. L'estimateur de Hausman-Taylor 51
8.7. La méthode des moments généralisés (GMM) en panel dynamique 52
8.8. Les tests sur données de panel 57
8.8.1. Le test de normalité des résidus 57
8.8.2. Le test de Ramsey-Reset 57
8.8.3. Le test d'hétéroscédasticité 58
8.8.4. Le test d'autocorrélation des erreurs 58
8.8.5. Le test de Chow 60
8.8.6. Test d'endogénéité et tests sur les coefficients des variables 60
8.8.7. Le test de validité des instruments 60
9. Econométrie des variables qualitatives 60
9.1. Le modèle Logit 61
9.2. Le modèle Probit 64
9.3. Tableau de prédiction (qualité de la prédiction) 65
9.4. La commande mfx pour calculer les impacts marginaux et les élasticités 66
9.5. Quelques autres modèles Logit et Probit et leurs commandes Stata 67
10. Introduction aux séries temporelles 68
10.2.1. Test de stationnarité sur données temporelles 69
10.2.2. Test de stationnarité sur données de panel 69
10.2.3. Test d'autocorrélation d'erreurs en série temporelles 70
11. Exporter les tableaux des régressions 70
12. Ajout de nouveaux modules à Stata 73
Références 75
ANNEXE : La Méthode des Moments Généralisés en Panel Dynamique 76

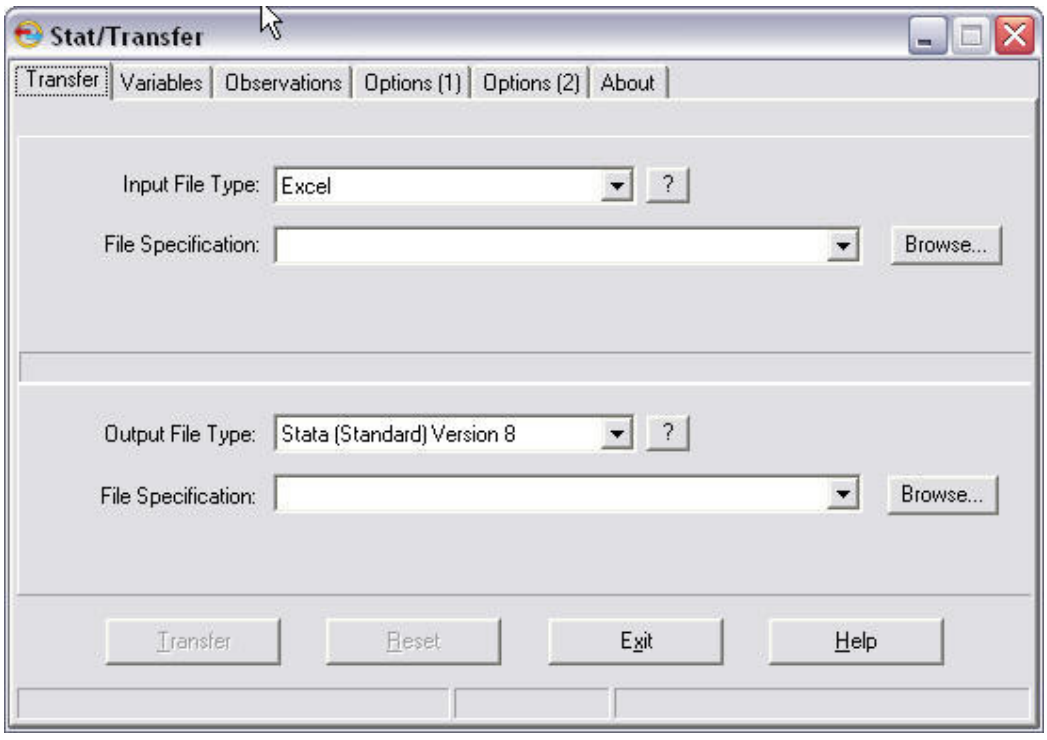
PARTIE I : Présentation des logiciels Stat Transfer 7 et Stata 8

Dans cette partie le logiciel Stat Transfer, qui permet de convertir les fichiers de bases de données dans un format compatible sous Stata, sera présenté dans une première section. Ensuite dans le seconde section, ce sera au tour du logiciel Stata d'être présenté, en particulier sa barre d'outils et ses différentes icônes.

1. Présentation de Stat Transfer 7

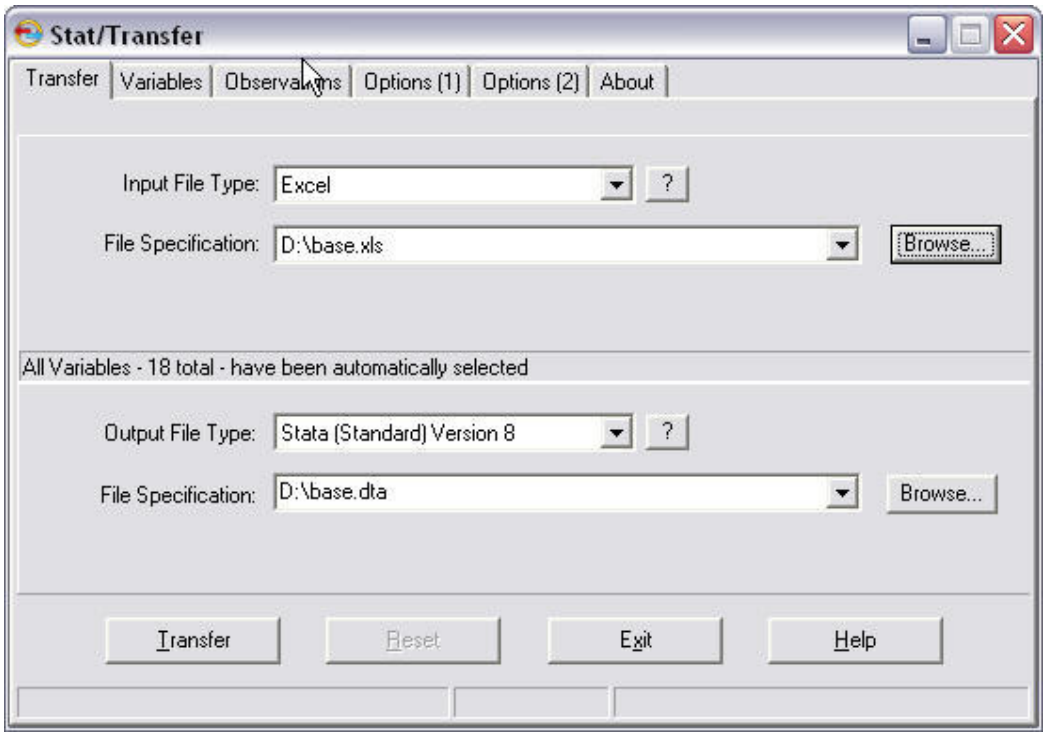
Les bases de données utilisables sous Stata doivent être dans un format spécifique (*.dta*). Le logiciel Stat Transfer permet d'obtenir ce format. En général les bases de données sont sous format Excel et peuvent être ensuite converties en format *.dta* par Stat Transfer. Notons que ce logiciel permet de faire également l'opération inverse (convertir un fichier *.dta* en fichier *.xls*). De manière générale, Stat Transfer permet de convertir tout fichier de bases de données en divers formats utilisables par des logiciels économétriques tels que Stata, SPSS, RATS, SAS etc. Les différentes étapes pour convertir un fichier en format *.dta* :

Etape 1 : Démarrer Stat Transfer, vous obtiendrez la capture d'écran ci-dessous :



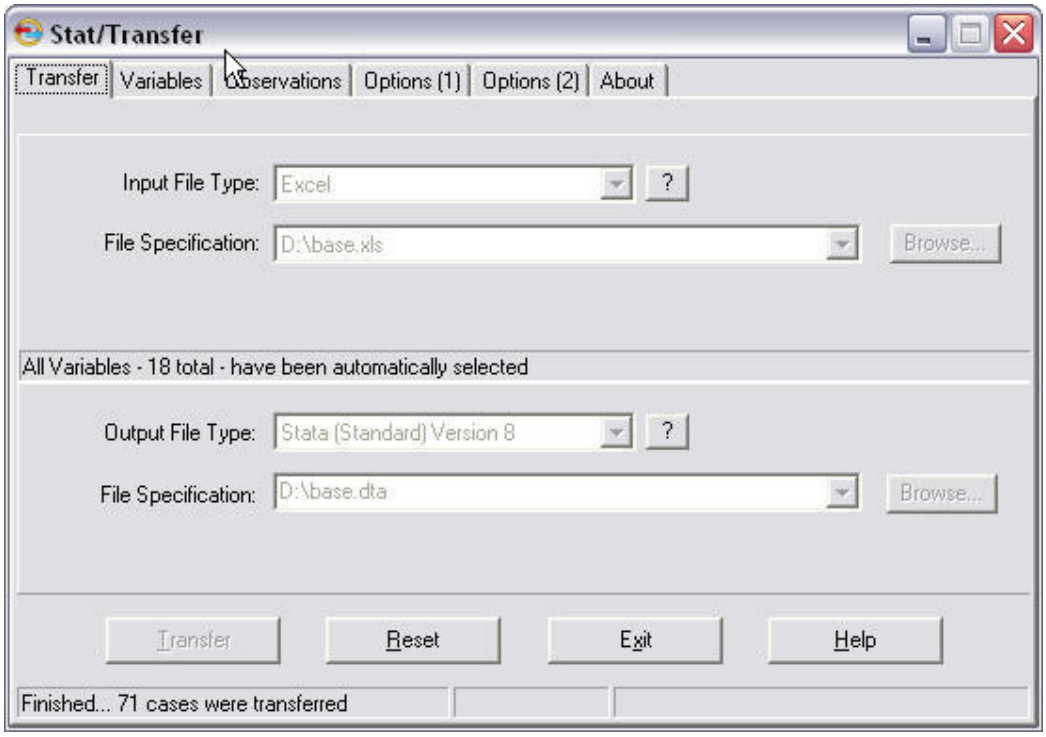
Input File Type permet de spécifier le format initial de la base de données, dans ce cas c'est le format Excel.
Output File Type permet de spécifier le format du fichier résultat.

Etape 2 : Spécifier le chemin d'accès du fichier *.xls*[1] à convertir en cliquant sur *Browse* (dans ce cas D:\Base.xls). Stat Transfer enregistrera par défaut le fichier *.dta* dans le même répertoire que le fichier source (D:\base.dta dans la capture d'écran ci-dessous)[2]



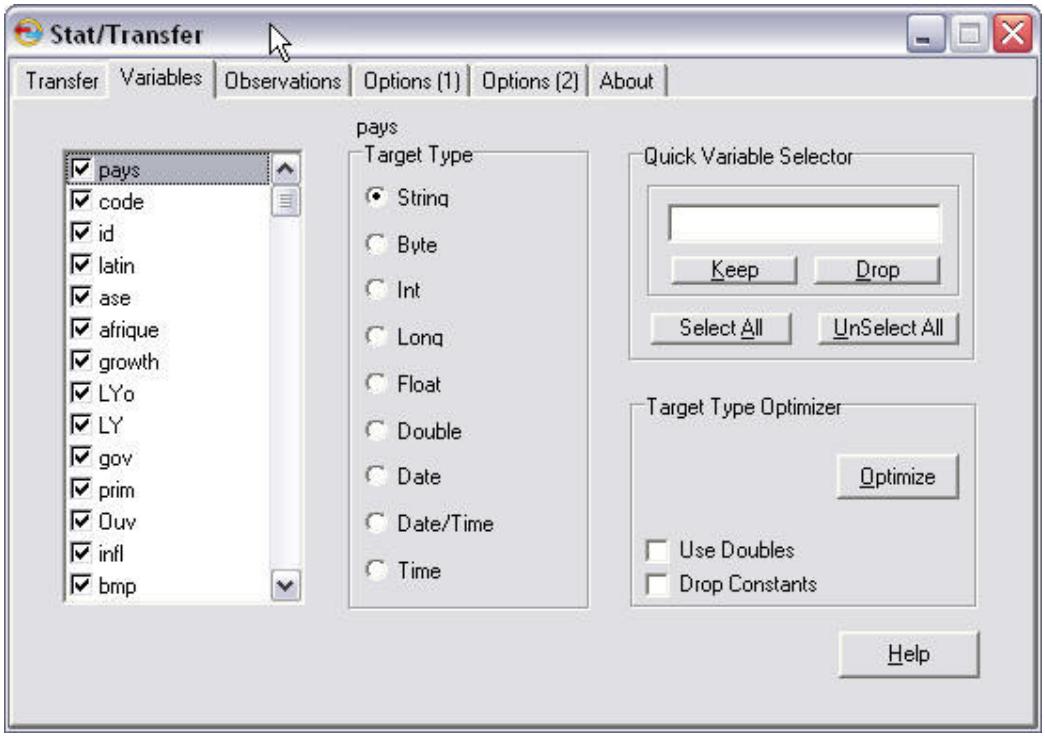
[1] Un fichier *.xls* peut contenir plusieurs feuilles, il est possible de spécifier à Stat Transfer la feuille contenant les données. Cette option apparaît automatiquement sous le chemin d'accès du fichier *.xls*. Mais Stat Transfer convertit mal les fichiers *.xls* s'il y a des données sur plusieurs feuilles du même classeur, il est donc préférable de n'avoir des données que sur une seule feuille du classeur.
[2] Vous pouvez modifier le nom et le répertoire par défaut du fichier *.dta*.

Etape 3 : Cliquer sur l'icône Transfer pour convertir le fichier. Dans la barre d'état inférieure, apparaît « Finished…71 cases were transfered », le chiffre 71 correspond au nombre d'observations dans la base de données. Cette indication permet de s'assurer que toutes les observations ont été bien prises en compte. Dans la barre d'état supérieure, le nombre de variables transférées est également indiqué « All Variables - 18 total - have been automatically selected ».



Vous avez certainement remarqué qu'il existe en haut de la fenêtre de Stat Transfer d'autres onglets à côté de celui de *Transfer* (*Variables*, *Observations*, *Options (1)*, *Options (2)* et *About*). Si généralement vous n'avez pas besoin d'utiliser ces onglets pour convertir vos fichiers, il est quand même nécessaire de savoir à quoi ils peuvent servir, en particulier les onglets *Variables* et *Observations*.

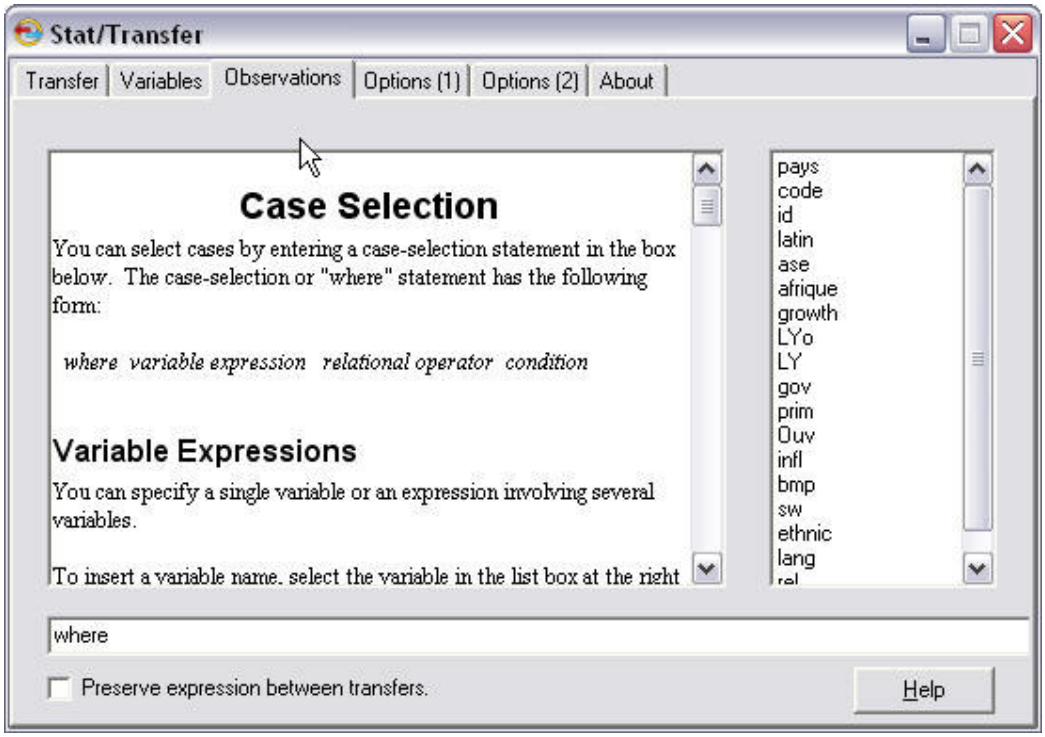
Dans la procédure de conversion ci-dessus, cliquer sur l'onglet *Variables* avant l'étape 3, on obtient la capture d'écran suivante :



Dans la colonne de gauche, toutes les variables de la base de données sont cochées par défaut. Si vous ne souhaitez pas transférer certaines variables, il suffit de les décocher. Dans la colonne du milieu figure le type de chaque variable. Dans Stata, deux types de variables sont couramment utilisés : le type *String* qui correspond au format alphanumérique[3] ou au format texte[4], et le type *Float* qui correspond au format numérique[5] (Exemple : dans la capture d'écran, la variable *pays* qui contient les noms des pays a été sélectionnée, Stat Transfer reconnaît automatiquement son format et lui attribut le type *String*)

L'onglet *Observations* permet de sélectionner les observations à inclure dans le fichier *.dta*, observations qui répondent à un critère précis spécifié dans la barre horizontale inférieure. Exemple : inclure **where** *growth>=0* dans la barre horizontale inférieure, a pour résultat de sélectionner dans la base de données uniquement les pays dont le taux de croissance économique (contenu dans la variable *growth*) est positif ou nul.

[3] Une suite de lettres et de chiffres.
[4] Une suite de lettres
[5] Une suite de chiffres

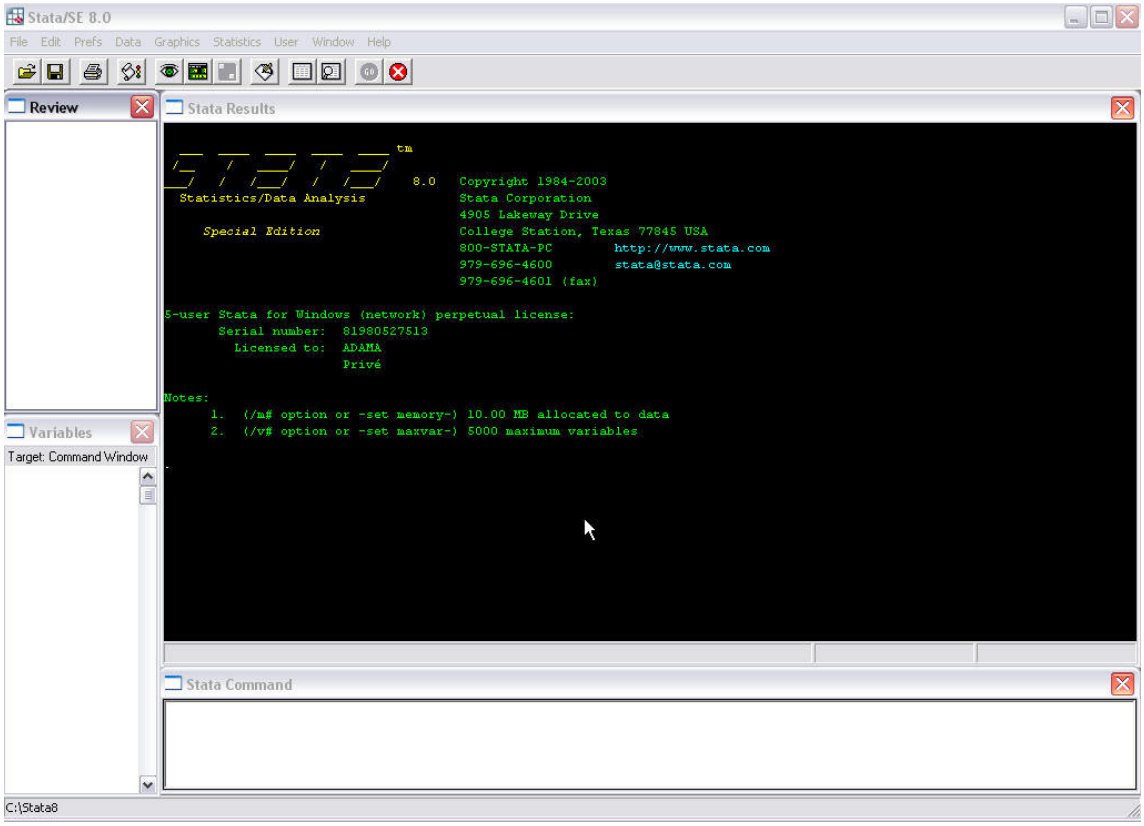


2. Présentation de Stata 8

En général, *Stata Corporation*, la société éditrice de Stata, lance une nouvelle version tous les deux ans. Actuellement, c'est la version 8. Cette version marque une réelle différence par rapport à la version 7 en ce sens que le logiciel a évolué vers plus de convivialité et de simplicité. Contrairement aux versions précédentes, Stata 8 dispose d'une barre de menu Windows pour les régressions économétriques (à l'exemple de Eviews). Cependant, il est préférable de se familiariser avec l'interface programme, qui offre plus de flexibilité dans la gestion des commandes et dans la résolution de calculs complexes. Dans ce manuel, seule l'utilisation de l'interface programme sera abordée.

2.1. Aspect général de l'interface Windows de Stata 8

Au démarrage du logiciel, on obtient la capture d'écran suivante :



La barre des menus rassemble les commandes pour gérer les bases de données, créer des graphiques et faire des régressions. La barre d'outils rassemble les raccourcis des commandes de bases pour ouvrir, créer ou enregistrer les fichiers gérés par Stata (fichiers de base de données, fichiers de résultats et les fichiers programmes). La description de la barre d'outils est détaillée plus bas. On observe également dans la capture ci-dessus que Stata dispose de quatre fenêtres indépendantes les unes des autres : (1) la fenêtre *Review* récapitule l'historique des commandes exécutées par Stata ; (2) les résultats de l'exécution des commandes sont affichés dans la fenêtre *Stata Results* ; (3) la liste des variables contenues dans la base de données est affichée dans la fenêtre *Variables* ; (4) la fenêtre *Stata Command* permet de saisir les commandes à exécuter par Stata. Chaque commande de Stata doit être validée en appuyant sur la touche Entrée[6]. Un clic sur une variable dans la fenêtre *Variables* permet d'afficher le nom de cette variable dans la fenêtre *Stata Command*.

[6] La touche Page up du clavier permet de revenir sur les commandes précédentes.

2.2. La barre d'outils de Stata 8

Dans l'ordre, l'icône ouvrir **(1)** ; enregistrer ; imprimer ; visualiser ou créer un fichier *log* **(2)** ; afficher l'aide et diverses options (mise à jour, lien vers le site Internet de Stata[7], etc.) **(3)** ; afficher les résultats **(4)** ; afficher un graphique ; ouvrir ou créer un fichier *do* **(5)** ; modifier la base de données **(6)** ; voir la base de données **(7)** ; faire défiler les résultats ; arrêter l'exécution d'une commande.

(1) A l'ouverture d'une base de données, Stata charge cette dernière dans la mémoire vive, le fichier ouvert n'est plus relié à sa source[8]. Par défaut 10 MB[9] sont alloués à la mémoire vive, mais cet espace peut s'avérer insuffisant pour de grosses bases de données. On peut l'augmenter par la commande suivante : **set memory #m** (remplacer # par la taille désirée, par exemple 40)[10]

(2) Un fichier *log* est un fichier (format texte) d'impression des commandes et des résultats de ses commandes au cours d'une session de Stata. Il permet de garder une trace des résultats des régressions économétriques. Un premier clic sur cet icône permet de créer un fichier *log* en spécifiant son chemin d'accès et le format *.log* (au lieu du format *.smcl* par défaut). Le fichier *log* est alors ouvert et prêt à enregistrer l'historique des commandes et des résultats. Avec un second clic sur cette icône, on a les options suivantes comme l'indique la capture d'écran ci-dessous : *(a)* la première permet de voir le contenu du fichier *log*, *(b)* la seconde permet de le fermer, *(c)* et la troisième permet de le suspendre pour un enregistrement ultérieur. Par la suite, vous verrez que la programmation rend plus simple la gestion des fichiers *log*.

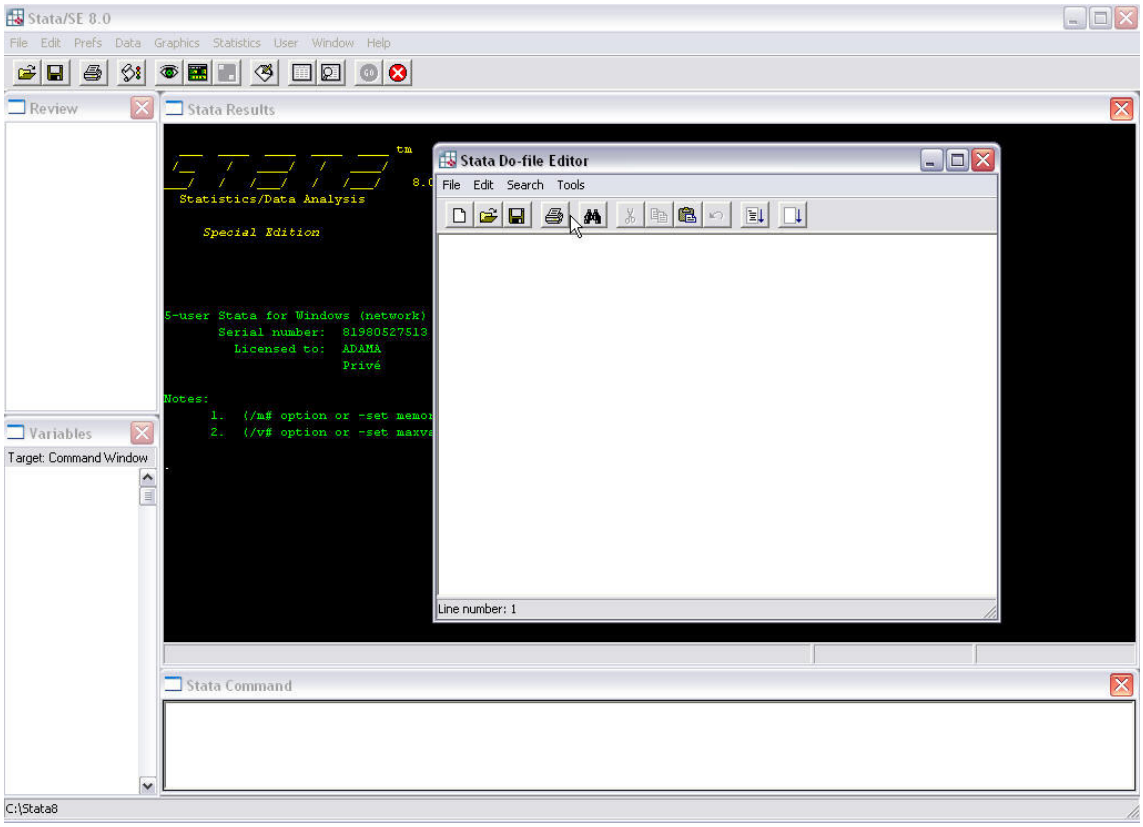
[7] www.stata.com
[8] Le fichier peut être supprimé dans son répertoire sans que cela n'affecte le fonctionnement de Stata.
[9] 10 MB pour *Stata Special Edition* et 1 MB pour *Intercooled Stata* (1 MB= 1024 kilobytes).
[10] Il faut quitter Stata, le relancer, puis appliquer la commande **set memory #m** avant d'ouvrir de nouveau la base de données. La ligne de commande **set memory #m, perm** spécifie de façon permanente la taille de la mémoire vive à allouer à Stata.

(3) L'aide de Stata est précieuse car on ne peut maîtriser qu'une infime partie des capacités du logiciel. De même, l'aide est utile pour maîtriser la syntaxe de chaque commande ainsi que ses subtilités. L'aide de Stata est facile d'utilisation si on connaît déjà la commande dont on a besoin, mais dans le cas contraire on a la possibilité de faire des recherches par mots clés. Il est également très facile de parcourir la table des matières de l'aide et de repérer la commande recherchée. L'aide est encore plus détaillée dans les manuels (officiels) d'utilisation de Stata, n'hésitez surtout pas à les consulter pour approfondir vos connaissances sur une commande précise. Pour afficher l'aide dans la fenêtre de résultats de Stata, il suffit de taper dans la fenêtre de commande : **help** suivi du nom de la commande (par exemple : **help graph** permet d'obtenir de l'aide sur les graphiques). Vous pouvez également utiliser le menu *Help* de la barre des menus.

(4) Cet icône permet de faire disparaître toute fenêtre qui vient masquer celle des résultats (par exemple : fenêtre d'aide, graphiques, etc.).

(5) Un fichier *do* (*do file*) est le fichier (format texte) dans lequel sont incluses les commandes de Stata sous forme de programme. Grâce à ce fichier, on garde une trace des commandes exécutées par Stata. Ce fichier contient donc un ensemble cohérent de commandes à exécuter par Stata dans l'ordre de leur apparition. Cliquer sur cette icône revient à ouvrir l'éditeur du fichier *do* comme le montre la capture ci-dessous. Dans l'éditeur du fichier *do*, on a encore une barre des menus et une barre d'outils. Dans la barre d'outils, il est assez intuitif de savoir ce à quoi correspondent les différentes icônes à l'exception des deux dernières probablement. Ces deux dernières icônes correspondent aux commandes du menu *Tools*. Les icônes *Do* et *Run* permettent d'exécuter les commandes du fichier *do* en entier, à la différence que l'icône *Run* ne permet pas de visualiser les résultats (l'intérêt est de voir si le fichier *do* s'exécute

correctement et qu'il ne comporte pas d'erreur qui en bloque l'exécution). La création des fichiers *do* sera abordée dans la section 4.



(6) Cette commande permet de visualiser la base de données avec la possibilité de modifier les données. On peut obtenir le même résultat avec la commande **edit**.

(7) Cette commande permet de visualiser la base de données sans possibilité de modification des données. La commande **browse** permet d'arriver au même résultat.

PARTIE II : Travailler avec Stata 8

Cette partie couvre l'utilisation des commandes de Stata pour créer et gérer des variables, fusionner des bases de données, créer des fichiers programmes, faire des statistiques descriptives, effectuer des régressions sur des données transversales et des données de panel, et faire des tests d'hypothèses. Pour faciliter la compréhension de la structure des syntaxes des différentes commandes, les commandes de Stata sont mises en caractères gras et les autres éléments (à l'exemple des noms des variables) sont mis en italiques pour signifier qu'ils relèvent du choix de l'utilisateur.

1. Les fonctions et les expressions

1.1. Les opérateurs arithmétiques

Addition	+
Soustraction	-
Multiplication	*
Division	/
Exposant	^

Exemple : **generate** *y = x^2* crée une nouvelle variable *y* telle que *y* soit le carré de *x*.

1.2. Les expressions *by*, *if* et *in*

(1) **by** permet de répéter une commande pour chaque valeur (ou modalité) d'une variable donnée. Syntaxe générale pour **by** :

by *variables* : *commande*

Avant d'utiliser **by**, il faut d'abord classer les observations en fonction des valeurs de la variable à laquelle la commande **by** va s'appliquer, la commande **sort** permet d'effectuer ce classement par ordre croissant.

Exemple 1 : soit une variable numérique nommée *continent* dont chaque valeur correspond à un continent donné :

sort *continent*

by *continent* : **list** *pays*

La commande **list** permet de faire une liste des modalités de la variable à laquelle elle s'applique. Ainsi, les deux lignes de commandes ci-dessus permettent de lister les noms des pays de la base de données pour chaque continent donné. On peut fusionner ces deux lignes de commandes par la ligne unique suivante :

bysort *continent* : **list** *pays*

Exemple 2 : soit *pri* une variable muette qui prend la valeur 1 pour les pays à revenu intermédiaire et 0 autrement, et soit la variable *pays* qui contient le nom des pays.

bysort *continent pri* : **list** *pays*

Cette ligne de commande classe d'abord les pays par continent, puis à l'intérieur des continents fait un classement en fonction de l'appartenance ou non du pays au groupe des pays à revenu intermédiaire, puis affiche le résultat.

Remarque : Pour certaines commandes, **by** se place non pas avant la commande, mais après celle-ci dans les options (exemple de la commande **collapse**, voir section 8.1)

(2) **if** permet de spécifier les conditions dans lesquelles une commande doit être exécutée. Syntaxe générale pour **if**

commande **if** *condition*

Exemple : **generate** *y* = *x*^(0.5) **if** *x*>=0 crée une variable *y* qui est égale à la racine carrée de la variable *x*, si *x* est positif.

(3) **in** permet de spécifier les observations auxquelles s'applique une commande. Syntaxe générale pour **in** :

commande **in** *intervalle*

Exemples :

- **list in** 1/9 affiche la première jusqu'à la neuvième observation de la base de données.
- **list in** *n* affiche la $n^{ième}$ observation, *n* peut être négatif, dans ce cas le décompte se fait à partir de la dernière observation.
- **list in** -1 affiche la dernière observation de la base de données.

Remarque générale : **by**, **if** et **in** peuvent être combinés à presque toutes les commandes de Stata, en particulier pour spécifier l'échantillon des régressions économétriques.

1.3. Les opérateurs de relation

Voici la liste des différents opérateurs de relation dans Stata :

Supérieur	>
Supérieur ou égal	>=
Inférieur	<
Inférieur ou égal	<=
Egal	=
N'est pas égal	~=
Différent de	!=

Remarque : Il existe une exception pour le signe d'égalité. En effet, lorsque la commande **if** précède une condition d'égalité, il faut utiliser le signe « = = » au lieu du signe « = » pour exprimer cette égalité.

Exemple : **list if** *x* = =10 liste les observations dont la valeur de *x* est égale à 10

list if *x*> • liste les observations dont les valeurs de *x* sont manquantes

1.4. Les fonctions

Voici une liste non exhaustive des fonctions mathématiques disponibles sur Stata :

Racine carrée	*sqrt*
Exponentielle	*exp*
Logarithme	*log*
Valeur absolue	*abs*
Partie entière	*int*

Exemple : **generate** *y* = **log(sqrt(abs(***x***)))** crée une variable *y* qui est égale au logarithme naturel de la racine carrée de la valeur absolue de la variable *x*.

Remarque 1 : Pour la fonction logarithme, Stata accepte l'expression **log** ou **ln**, mais les deux correspondent au logarithme naturel. Pour obtenir la fonction logarithme base 10, il faut utiliser l'expression **log10**.

Remarque 2 : Il existe bien d'autres fonctions telles que les fonctions de probabilité, les fonctions sur les matrices et les fonctions texte. Vous pouvez les consulter dans l'aide (menu *Help*) ou dans les Manuels de Stata.

1.5. Les opérateurs logiques

Ou | (combinaison de la touche *Altgr* et la touche « 6 » du pavé alphanumérique)
Et &

Exemple : **list if** *x*>3 **&** *x*<20 liste toutes les observations dont la valeur de *x* est comprise entre 3 et 20, bornes non comprises.

Remarque : l'opérateur **&** est prioritaire sur l'opérateur |
list if *x*>50 | (*x*>30 **&** *z*<2.5) équivaut à écrire **list if** *x*>50 | *x*>30 **&** *z*<2.5

2. Commandes de gestion des variables

2.1. Création de nouvelles variables : les commandes *generate* et *egen*

Pour créer une variable, deux commandes sont disponibles dans Stata : la commande **generate** et la commande **egen**. La commande **egen** est une extension de la commande **generate,** et elle est utilisée pour créer des variables à l'aide de fonctions spécifiques (voir l'aide pour la liste de fonctions utilisables avec la commande **egen**).

Exemples pour **generate** :

(1) **generate** *y* = *x* : crée une variable *y* dont les valeurs sont identiques à celles de la variable *x*.

(2) **generate** *y* = *x*+*z* : crée une variable *y* qui est à la somme des variables *x* et *z*.

(3) **generate** *y* = *x*>100 : crée une variable muette *y* qui est égale à 1 si la valeur de *x* est supérieure à 100 et 0 autrement. L'inconvénient avec cette syntaxe est que *y* sera égale à 1 même pour les valeurs manquantes de *x*, car pour Stata les valeurs manquantes sont remplacés par un point (•) considéré comme une valeur infinie. Pour éviter cela, il est plus approprié d'utiliser la syntaxe suivante : **generate** *y* = *x*>100 **if** *x*< •

(4) **generate** *y* = *x***[n]** : crée une constante *y* dont la valeur est égale à celle de la $n^{ième}$ observation de la variable *x*.

(5) **generate** *y* = *x***[_n-1]** : crée une variable *y* qui est égale à la valeur précédente de la variable *x*.[11]

(6) **generate** *y* = **"***ab***"** : crée une variable non numérique contenant le terme *ab* pour toutes les observations.

Exemples pour **egen** :

(1) **egen** *y* = **count(***x***)** : crée une variable *y* dont la valeur est constante et égale au nombre d'observations non manquantes de la variable *x*.

(2) **egen** *y* = **sd(***x***)** : crée une variable *y* dont la valeur constante est égale à l'écart-type de la variable *x*. A la place de **sd**, on peut utiliser la fonction **mad** pour calculer la médiane, la fonction **max** pour le maximum, la fonction **min** pour le minimum, la fonction **sum** pour la somme et la fonction **mean** pour la moyenne.

(3) **egen** *y* = **pctile(***x***)**, **p(***n***)** : permet de créer une variable *y* telle qu'elle soit égale au $n^{ième}$ percentile (*n*=50 correspond à la médiane).

(4) **egen** *idpays* = **group(***pays***)** : crée une variable *idpays* qui attribue un numéro par pays en les classant par ordre alphabétique (*pays* est une variable qui contient le nom des pays). Cette commande peut servir à créer des identifiants pays. De façon générale, la fonction **group** permet d'attribuer un numéro à chaque modalité de la variable à laquelle elle s'applique.

[11] Attention, cette commande ne permet pas du tout de créer une variable retardée, sauf dans le cas où toutes les observations de la base de données font référence à un seul pays ou une seule entité. Pour la création de variable retardée, se référer à la section 10.1.

(5) **egen** *y* = **rmax(***x,z***)** : pour chaque observation, la nouvelle variable *y* créée sera égale au maximum des valeurs des variables *x* et *z*. Lorsqu'on remplace **rmax** par la fonction **rsum**, la variable *y* créée sera égale à la somme en ligne des variables *x* et *z*. Il existe également les fonctions **rsd** et **rmean**.

Les commandes **generate** et **egen** peuvent être combinées avec **by**, **if** et **in**. Supposons par exemple que l'on dispose des données sur le PIB (contenues dans la variable nommée *pib*) pour un échantillon de pays, et que l'on souhaite calculer le PIB moyen par continent, la commande à appliquer est la suivante :

sort *continent*

by *continent* : **egen** *pib_m* = **mean(***pib***)**

La nouvelle variable *pib_m* créée va faire correspondre à chaque pays la valeur du PIB moyen du continent auquel il appartient.

Après **by**, on peut également mettre plus d'une variable comme spécifié dans l'exemple 2 de la section 1.2.

2.2. Autres commandes relatives à la gestion des variables

- Syntaxe pour renommer une variable : **rename** *ancien_nom nouveau_nom*
  Exemple : **rename** *gdp pib* : change le nom de variable *gdp* par *pib*
- La commande **replace** permet de modifier les valeurs d'une variable déjà existante : par exemple créer une variable muette *y* qui prend la valeur 1 si la variable *x* est positive et 0 autrement[12] :

  generate *y* = 0 : crée une variable *y* égale à 0 pour toutes les valeurs de *x*.

  replace *y* = 1 **if** *x*>0 **&** *x* <• : remplace les 0 par 1 uniquement pour les valeurs positives de *x* et les observations non manquantes.

  replace *y* = • **if** *x*>• : pour toutes les observations manquantes de *x*, la variable *y* aura également des observations manquantes.
- La commande **drop** permet de supprimer des variables, par exemple :
  drop *y* supprime la variable *y* de la base de données.
  La commande **drop** peut être combinée avec **by**, **if** et **in**.

[12] On peut utiliser également la syntaxe : **generate** *y* = *x*>0 **if** *x*< •

Exemples : **drop in** 1/9 supprime la première jusqu'à la neuvième observation.

drop if $x>0$ supprime toutes les observations dont la valeur de x est positive.

Dans ce dernier cas, n'oubliez pas que les valeurs manquantes de la variable x seront également supprimées parce qu'elles sont considérées par Stata comme des valeurs infinies. La commande **drop _all** ou **clear** supprime toutes les variables de la base de données.

- La commande **keep** (garder) marche de la même manière que la commande **drop** à la différence qu'elle produit le résultat inverse.

  Exemple : **keep in** 1/9 ne garde dans la base de données que la première jusqu'à la neuvième observation.

- La commande **sort** permet de classer les observations par ordre croissant d'une ou de plusieurs variables. Exemples : soit *pays* une variable contenant le nom des pays,

  sort *pays* : permet de classer les observations par ordre alphabétique des noms de pays

  sort *pays periode* : classe d'abord par ordre alphabétique les pays, puis pour chaque pays fait le classement des données par ordre chronologique (*periode*[13] étant la variable représentant le temps).

  Pour faire des classements, il existe une commande dans Stata nommée **gsort** qui offre plus de flexibilité en ce sens qu'elle permet de classer par ordre croissant ou par ordre décroissant ou les deux simultanément pour deux ou plusieurs variables différentes.

2.3. Abréviations des noms des variables et des commandes

Stata permet d'abréger les noms des commandes suivant des règles précises. Par exemple lorsqu'on fait appelle à l'aide de Stata sur la commande **generate**, on observe que la lettre **g** du mot **generate** est soulignée. Cela veut dire qu'on peut écrire ***g*** $y = x$ à la place de **generate** $y = x$. On peut rajouter une, deux ou plusieurs lettres de la commande **generate** à **g**, en d'autres termes on aura le même résultat si on écrit **ge** $y = x$ ou **gen** $y = x$ ou encore **gene** $y = x$. Le même principe s'applique aux autres commandes de Stata à l'exception de quelques unes comme **replace** qui n'autorisent pas d'abréviations.

[13] *periode* au lieu de *période* car les lettres accentuées ne sont pas reconnues par Stata.

Stata permet également quelques raccourcis pour les noms des variables. Exemples :

list *var** affiche toutes les variables dont les noms commencent par les lettres *var*

list **var* affiche toutes les variables dont les noms se terminent par les lettres *var*

list *x-y* affiche les variables *x* et *y* ainsi que toutes celles se trouvant entre ces deux variables dans l'ordre d'apparition dans la base de données.

On peut également abréger le nom des variables à moins que l'abréviation n'entre en conflit avec les noms des autres variables. Supposons qu'il existe dans une base de données deux variables nommées respectivement *pib* et *pib_ppa*. La limite d'abréviation du nom de la dernière variable sera *pib_*, on peut donc écrire par exemple **list** *pib_* à la place de **list** *pib_ppa*. Il est à noter que les noms des variables doivent respecter quelques principes de bases : *(a)* le nom d'une variable doit être le plus court possible, ne doit pas contenir de lettre accentuée, ni d'espace, ni de caractères spéciaux comme ?,& ou !. *(b)* les majuscules et les minuscules sont différenciées par Stata dans la gestion des noms de variables, pour exemple les variables *MAKE*, *Make* et *make* seront considérées par Stata comme trois variables distinctes.

2.4. Mettre des étiquettes pour les variables

Stata offre la possibilité d'attribuer une étiquette à chaque variable. Cette étiquette constitue une description de la variable puis qu'il n'est probablement pas évident à une tierce personne de deviner la signification d'une variable à partir de son nom. Par ailleurs les étiquettes apportent une meilleure lisibilité des résultats économétriques car dans les tableaux des régressions, on peut remplacer les noms des variables par leurs étiquettes.

Syntaxe générale : **label var** *nom_variable* **"***description***"**

Exemple : **label var** *pib* **"***Produit Interieur Brut***"**

Il est possible également d'étiqueter les valeurs d'une variable catégorielle. Par exemple : supposons qu'une variable muette nommée *pvd* prend la valeur 1 pour les pays en développement et 0 pour les autres. Les deux lignes de commandes suivantes permettent de mettre des étiquettes aux valeurs 1 et 0 :

label define *d_pvd* 1 "*pays en developpement*" 0 "*pays developpe*" (création de l'étiquette)

label values *pvd d_pvd* (association de l'étiquette *d_pvd* à la variable *pvd*)

La variable *pvd* qui était composée de 1 et de 0 sera désormais composée de "*pays en développement*" et "*pays developpe*". L'avantage est que la variable *pvd* garde sa nature *Float* (variable numérique) malgré qu'elle soit composée désormais de chaînes de caractères. Elle reste donc utilisable dans les régressions économétriques.

On peut également attribuer un nom à toute la base de données par la commande :

label data "*nom_base***"**

Pour supprimer les étiquettes pour les trois exemples ci-dessus :

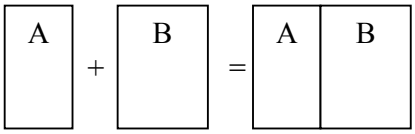
label var *pib*	supprime l'étiquette de la variable *pib*
label values *pvd*	dissocie l'étiquette *d_pvd* des valeurs de la variable *pvd*
label drop *d_pvd*	supprime l'étiquette *d_pvd*
label data	supprime le nom ou la description de la base de données

La commande **describe** permet de visualiser la liste des variables de la base de données, leurs formats et leurs étiquettes. Si la commande **describe** est suivie d'une ou deux plusieurs variables, seuls les informations les concernant seront affichées.

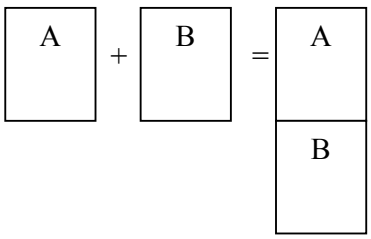
3. Fusion de bases de données

Les commandes **merge** et **append** permettent respectivement de fusionner horizontalement et verticalement deux bases de données. De façon générale, la commande **merge** vous permet d'ajouter de nouvelles variables à la base de données, et la commande **append** vous permet d'ajouter de nouvelles observations.

Commande **merge** :



Commande **append** :



Voici les différentes étapes pour fusionner deux bases de données. La procédure suivante concerne la commande **merge**. Tout d'abord il faut avoir une variable commune aux bases A et B qui permettra de faire la fusion, prenons par exemple les noms de pays contenus dans la variable nommée *pays*.

(1) Ouvrir la base A et classer les observations par nom des pays : **sort** *pays*

(2) Ouvrir la base B, classer les observations par nom des pays : **sort** *pays*
Enregistrer puis fermer la fenêtre Stata de la base B.

(3) Revenir dans la fenêtre de la base A pour appliquer la commande suivante :
merge *pays* **using "***Chemin d'accès de la base B***"**

(4) Enregistrer ensuite la nouvelle base obtenue.

Remarque 1 : La variable de fusion (nom des pays dans cet exemple) de la base A doit être rigoureusement identique à celle de la base B.

Remarque 2 : On peut également utiliser deux ou plusieurs variables pour faire la fusion (en particulier pour les panels) : par exemple fusionner par nom des pays et par période, on aura alors les commandes suivantes :

sort *pays periode*
merge *pays periode* **using "***Chemin d'accès de la base B***"**

Remarque 3 : Il se peut qu'il y ait des observations dans la base A qui ne se trouvent pas dans la base B et inversement. Dans ce cas, arrivé à l'étape 3 de la procédure, vous pouvez utiliser la commande **browse** pour afficher toute la base de données, puis faire un copier (*Ctrl+C*) coller de Stata vers Excel pour nettoyer la base des observations non désirées. Une autre possibilité est que Stata crée une variable indicatrice nommée *_merge* qui dans sa forme standard prend la valeur 1 lorsque les observations de la base de données résultante proviennent uniquement de la base A, la valeur 2 lorsque les observations proviennent uniquement de la base B et la valeur 3 lorsque les observations sont communes aux deux bases. Avec la commande **drop if** *_merge* = = 2, on peut supprimer les observations non désirées de la base B. Cependant on prend le risque de supprimer par exemple une observation de la base A qui a son correspondant dans la base B, mais dont l'identifiant est légèrement différent dans les deux bases par mégarde (Exemple : Côte d'Ivoire et Côte d'ivoire ne représentent pas le même pays).

Pour la commande **append**, la syntaxe est plus simple. Il suffit d'ouvrir la première base (base A) et dans la fenêtre de commande de Stata, taper la ligne de commande suivante :

append using "*Chemin d'accès de la base B***"**

4. Création d'un fichier do

Un fichier *do* comme je l'ai dit plus haut est un fichier programme qui contient un ensemble de lignes de commandes à exécuter par Stata. Il existe une série de lignes de commandes préliminaires standard qui permettent d'ouvrir la base et de créer un fichier *log* où seront stockés les résultats du programme. Voici la liste de commandes préliminaires :

version #	(1)
capture clear	(2)
capture log close	(3)
log using "*C:\resultat.log***", replace**	(4)
clear	(5)
set memory #**m**	(6)
set more off	(7)
use* "C:\base.dta***"**	(8)
preserve	(9)

- La ligne 1 permet de spécifier la version de Stata à laquelle se réfèrent les commandes. Certaines commandes peuvent changer d'une version à une autre. Par exemple, le fait de mettre version 7 en début du fichier *do* permet aux commandes de pouvoir être exécutées par Stata 8, même si celles-ci ont changé.
- La ligne 2 permet de vider la mémoire vive de Stata.
- La ligne 3 permet de fermer tout fichier *log* précédemment ouvert.
- La ligne 4 crée un fichier *log* nommé *resultat* sur le répertoire *C:*
- La ligne 5 exécute la même commande que la ligne 2.
- La ligne 6 définit la taille en MB allouée à la mémoire vive de Stata.
- La ligne 7 autorise Stata à faire défiler tous les résultats jusqu'à la fin du fichier *do*. Sans cette commande, les résultats seront présentés paquet par paquet.
- La ligne 8 spécifie le chemin d'accès de la base de données nommée *base.dta*

- La commande **preserve** de la ligne 9 fait une "photographie" d'origine de la base de données, puis la restitue à la fin de l'exécution du fichier *do*, même si entre temps elle a été modifie par les commandes du programme. Ainsi la base de donnée de départ reste intacte de toutes modifications.

Vous pouvez appliquer ces commandes préliminaires à d'autres bases de données en changeant uniquement les lignes 4 et 8. Après la ligne 9, vous pouvez mettre toutes vos commandes de gestions des variables, de statistiques descriptives et de régressions économétriques. Toutes les commandes déjà présentées et celles qui vont suivre peuvent être incluses dans un fichier *do*. Notons que vous pouvez inclure dans le fichier *do* des commentaires, pour cela il faut mettre le signe * avant le commentaire. Le signe * en début de ligne signifie à Stata de ne pas exécuter cette ligne.

L'icône *Do* de *Stata Do-file Editor* permet d'exécuter tout le fichier *do*. Vous pouvez également exécuter une seule ligne du fichier *do* en cliquant sur l'icône *Do* après avoir sélectionné cette ligne. Mais logiquement cela ne marche pas lorsque cette ligne comporte une variable définie en amont.

5. Les statistiques descriptives

5.1. La commande *summarize*

La commande **summarize** (ou **sum** en abrégé) calcule pour une variable ou une liste de variable la moyenne, l'écart-type, le minimum et le maximum de l'échantillon sélectionné.

Syntaxe générale : **sum** *noms_variables* **(if, in)**

Exemples :

sum *y* : la commande **sum** s'applique à la variable *y*. .

sum *y x* : la commande **sum** s'applique aux variables *x* et *y*.

bysort *continent* : **sum** *pib* : la commande **sum** s'applique séparément à chaque modalité de la variable *continent*.

sum *pib* **if** *continent* = **"***Afrique***"** la commande **sum** s'applique à la variable *pib* mais uniquement pour les observations dont la variable *continent* est égale à la suite de caractère *Afrique*.

Remarque 1 : lorsque aucune variable n'est spécifiée à la suite de la commande **sum**, alors les statistiques descriptives sont faites pour toutes les variables de la base de données.

Remarque 2 : **sum** *y*, **detail** (avec l'option **detail**, la commande **sum** donne en plus des statistiques standard, le *skewness* et le *kurtosis* de la variable *y*)

Remarque 3 : la commande **tabstat** permet également de faire des statistiques descriptives. Elle offre plus de flexibilités que la commande **sum** en ce sens qu'elle permet de faire un tableau unique de statistiques descriptives pour plusieurs variables, et elle offre une plus large panoplie de statistiques.

5.2. La commande tabulate

La commande **tabulate** (ou **tab** en abrégé) calcule les fréquences des observations d'une variable et permet de faire des tableaux croisés pour deux variables.

Exemples : **tab** *y* fait un tableau des valeurs de la variable *y* avec leur fréquence.

tab *y x* fait un tableau croisé des valeurs de *y* et de *x*.

tab *y*, **gen(***dy***)** Exécute la même commande que **tab** *y*, mais en plus une variable muette sera créée pour chaque modalité de la variable *y*. cette variable muette sera nommée *dy1* pour la première modalité de la variable *y*, *dy2* pour la seconde modalité, et ainsi de suite…

tab *y x*, **row** tableau croisé de *y* et *x* avec les fréquences en lignes.

tab *y x*, **col** tableau croisé de *y* et *x* avec les fréquences en colonnes.

Il existe d'autres variantes de la commande **tab**, il s'agit de **tab1** et **tab2**.

tab1 *y x* crée non pas un tableau croisé de *y* et *x*, mais un tableau séparé pour chacune de ces variables.

tab2 *y x z* crée un tableau croisé pour chaque combinaison possible de deux variables de cette liste de variables (*y x*, *y z*, et *x z*).

On peut combiner **tab** et ses variantes avec **by**, **if** et **in**.

5.3. Les coefficients de corrélation

Pour calculer la corrélation entre deux ou plusieurs variables : la syntaxe générale est la suivante :

pwcorr *variable1 variable2 ... variableN*, **obs sig star(**#**)**

L'option **obs** reporte le nombre d'observations utilisé pour calculer les coefficients de corrélation. L'option **sig** rajoute une ligne donnant la probabilité de rejet de l'hypothèse de non significativité du coefficient de corrélation. Pour l'option **star**, si # = 10, tous les coefficients de corrélation significatifs au seuil de 10% sont marqués d'une étoile (*). Exemple : **pwcorr** *x y z*, ***obs sig star*(**10**)**

Il existe une autre commande dénommée **correlate** (ou **corr**) qui produit la matrice des variances-covariances d'une liste de variables ou d'un ensemble de coefficients d'une régression.

corr *y x*, **cov**	produit la matrice des variances-covariances des variables *y* et *x*
corr *y x*, **cov mean**	produit des statistiques descriptives de *y* et *x* en plus de la matrice des variances-covariances.
corr, **cov _coef**	produit la matrice des variances-covariances des coefficients du modèle estimé le plus récent.

La commande **pcorr** quant à elle permet de calculer les coefficients de corrélations partielles :

pcorr *x y z*	calcule le coefficient de corrélation partielle entre la variable *x* et la variable *y* en maintenant la variable *z* constante, et le coefficient de corrélation partielle entre la variable *x* et la variable *z* en maintenant la variable *y* constante.

Les commandes **pwcorr**, **corr** et **pcorr** peuvent être combinées avec **by**, **if** et **in**.

6. Tests sur la moyenne, la variance et la distribution des variables

6.1. Test de comparaison de moyennes

La commande **ttest** de Stata permet de faire des comparaisons de moyennes.

Cas 1 : soit la variable *prim* le taux de scolarisation primaire moyen sur la période 1988-1997 pour un échantillon de pays en développement. Pour tester par exemple la différence de moyenne de taux de scolarisation entre le groupe des pays africains et les autres pays de l'échantillon. La commande et les résultats sont les suivants :

```
. ttest prim, by(afrique)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      43    1.026374    .0214091    .1403888     .983169     1.06958
       1 |      26    .7089842    .0568345    .2898005    .5919313    .8260371
---------+--------------------------------------------------------------------
combined |      69     .906778    .0311734    .2589461    .8445724    .9689837
---------+--------------------------------------------------------------------
    diff |            .3173901    .0519286                .2137402    .4210399
------------------------------------------------------------------------------
Degrees of freedom: 67

                     Ho: mean(0) - mean(1) = diff = 0

     Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
       t =   6.1121                 t =   6.1121                 t =   6.1121
   P < t =   1.0000             P > |t| =   0.0000             P > t =   0.0000
```

La variable *afrique* est une muette égale à 1 pour les pays africains et 0 autrement. La statistique du test se trouve en bas du tableau et montre qu'on ne peut pas rejeter l'hypothèse que le taux de scolarisation moyen des pays africains soit moins élevé que celui des autres pays en développement. Ce test suppose que la variance de la variable *prim* n'est pas différente d'un groupe à un autre, on peut relâcher cette hypothèse avec l'option **unequal** (**ttest** *prim*, **by(***afrique***) unequal**).

Cas 2 : on peut également tester l'égalité des moyennes de deux variables différentes. Par exemple, tester si en moyenne dans un échantillon de pays le taux de scolarisation primaire est plus élevé que le taux de scolarisation secondaire ou inversement. La syntaxe est :

ttest *prim* = *sec*, **unpaired unequal**

L'option **unpaired** signifie que les deux variables ne représentent pas la même chose. Cette option ne doit pas être utilisée lorsqu'on compare la moyenne d'une variable pour laquelle on

dispose de deux mesures obtenues de deux sources différentes. Dans ce dernier cas, on doit aussi enlever l'option **unequal**.

Cas 3 : la commande **ttest** sert également à tester l'égalité de la moyenne d'une variable à une valeur donnée. Exemple : **ttest** *prim* = 30, permet de tester si la moyenne de la variable *prim* est supérieure, inférieure ou égale à 30.

Remarque : pour effectuer des tests de comparaison des proportions, il faut utiliser la commande **prtest** de Stata dont le fonctionnement est similaire à celui de **ttest**.

6.2. Test de comparaison de variances

La commande **sdtest** de Stata permet de procéder à des comparaisons de variances. On retrouve les trois cas similaires à ceux de la commande **ttest**.

Cas 1 : comparaison de la variance du taux de scolarisation primaire des pays africains à celle des autres pays en développement.

```
. sdtest prim, by(afrique)

Variance ratio test

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      43    1.026374    .0214091    .1403888     .983169     1.06958
       1 |      26    .7089842    .0568345    .2898005    .5919313    .8260371
---------+--------------------------------------------------------------------
combined |      69     .906778    .0311734    .2589461    .8445724    .9689837
------------------------------------------------------------------------------

                         Ho: sd(0) = sd(1)

             F(42,25) observed   = F_obs           =    0.235
             F(42,25) lower tail = F_L   = F_obs   =    0.235
             F(42,25) upper tail = F_U   = 1/F_obs =    4.261

   Ha: sd(0) < sd(1)         Ha: sd(0) != sd(1)          Ha: sd(0) > sd(1)
  P < F_obs = 0.0000     P < F_L + P > F_U = 0.0002     P > F_obs = 1.0000
```

D'après les résultats du test, on peut conclure que la variance du taux de scolarisation primaire est plus élevée dans le groupe des pays africains que dans celui des autres pays en développement, donc sur la base du taux de scolarisation primaire le groupe des autres pays en développement est plus homogène.

Cas 2 : Test d'égalité de variances entre deux variables différentes :

sdtest *variable1 = variable2*

Cas 3 : **sdtest** *y = #* teste l'hypothèse Ho que la variance de la variable *y* est égale à la valeur #.

6.3. Test sur la distribution de deux variables

Ce test s'obtient à l'aide de la commande **tabulate** (ou **tab**) décrite plus haut. C'est le test de chi2 de Pearson dont l'hypothèse Ho est l'indépendance des lignes et des colonnes du tableau croisé. La syntaxe est la suivante :

tab *variable1 variable2*, **chi2**

7. Les régressions sur données transversales :

7.1. Les moindres carrés ordinaires (MCO)

Pour effectuer des régressions en MCO, il faut utiliser la commande **regress** (ou **reg**) suivi de la variable dépendante, des variables explicatives et éventuellement des options. La syntaxe générale est la suivante :

reg *var_dep var_explicatives* **(if, in)**, *options*

Exemple : Impact du tourisme sur la croissance économique

La base de données[14] est composée de 63 pays avec des données en moyenne sur la période 1988-1997. La variable *growth* est le taux de croissance du PIB par tête, *tourism* le nombre de touristes (en log), *lyo* le niveau du PIB par tête initial (1988), *prim* le taux de scolarisation primaire, *infl* le taux d'inflation et *sw* l'indicateur de politique d'ouverture de Sachs et Warner. Voici les résultats de l'estimation du modèle par les MCO.

[14] Disponible sur demande

```
. reg growth tourism lyo prim  infl sw, robust

Regression with robust standard errors                Number of obs =      58
                                                      F(  5,    52) =    9.56
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5703
                                                      Root MSE      =  2.0622

------------------------------------------------------------------------------
             |               Robust
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tourism |   2.533977   .5266953     4.81   0.000     1.477086    3.590869
         lyo |  -1.260772   .7039543    -1.79   0.079    -2.673359    .1518156
        prim |   .2660154   1.245608     0.21   0.832     -2.23348    2.765511
        infl |   -.140905   .0566845    -2.49   0.016    -.2546509   -.0271592
          sw |   1.284783   .7505155     1.71   0.093    -.2212367    2.790802
       _cons |   -10.2673   2.364197    -4.34   0.000     -15.0114   -5.523187
------------------------------------------------------------------------------
```

Sur la base de la lecture des *t* de *student* ou de leur probabilité, on note que la variable d'intérêt du modèle (*tourism*) a un coefficient positif et significatif à 1%, soutenant ainsi l'hypothèse assez évidente que le tourisme est corrélé positivement au taux de croissance économique. L'hypothèse de convergence est également soutenue par le coefficient négatif et significatif à 10% du niveau initial du PIB par tête. Les résultats montrent également que l'inflation est mauvaise pour la croissance économique tandis que l'ouverture commerciale lui est favorable. Parmi toutes les variables, seul le taux de scolarisation primaire n'est pas significatif au seuil conventionnel de 10%.

Le R^2 de l'estimation est égal à 0.57, la variabilité des variables explicatives du modèle expliquerait 57% de la variabilité du taux de croissance. Avec l'option **robust** (ou **ro**) associée à la commande **reg**, les *t* de *student* sont corrigés de l'hétéroscédasticté par la méthode de White.

Il existe d'autres options associées à la commande **reg**, par exemple l'option **noconstant** supprime la constante dans le modèle, l'option **cluster(***id_group***)** suppose que les observations sont indépendantes d'un groupe à un autre, mais pas nécessairement à l'intérieur d'un même groupe (*id_group* est la variable catégorielle des groupes).

On peut combiner la commande **reg** à **by**, **if** et **in**. Par exemple, soit *afrique* une variable muette égale à 1 pour les pays africains et 0 autrement :

reg *growth tourism lyo prim infl sw* **if** *afrique* = = 1, **robust** (le modèle sera estimé uniquement pour les pays africains)

bysort *afrique* : **reg** *growth tourism lyo prim infl sw,* **robust** (le modèle sera estimé séparément sur l'échantillon des pays africains et celui des autres pays)

La commande **predict** permet d'obtenir sur la base des coefficients estimés, entre autres, la valeur prédite de la variable dépendante ainsi que les résidus de la régression. La commande **predict** s'applique à la régression la plus récente et ses options diffèrent d'une technique économétrique à une autre[15]. La syntaxe générale est :

predict *variable* **(if, in)**, *options*

Pour les MCO et dans le cas de l'exemple ci-dessus :

reg *growth tourism lyo prim infl sw*, **robust**

predict *growth_hat*, **xb** les valeurs prédites du taux de croissance seront stockées dans la variable *growth_hat*

predict *growth_res*, **re** permet de prédire les résidus qui seront stockés dans la variable nommée *growth_res*

Remarque : Avec l'option **predict**, Stata calcule dans la mesure du possible les valeurs prédites de la variable dépendante ou des résidus pour toutes les observations de l'échantillon (pays, individu, etc.), même pour celles qui n'ont pas été prises en compte dans les régressions. Pour limiter la prédiction aux seules unités prises en compte dans les régressions, la condition **if e(sample)** est utilisée.

Exemples : **predict** *growth_hat* **if e(sample), xb**

predict *growth_res* **if e(sample), re**

On peut combiner **predict** avec **by**, **if** et **in**.

Que ce soit pour les MCO ou pour toute autre technique économétrique, chaque élément des résultats des régressions économétriques est stocké par Stata dans une macro, un scalaire ou une matrice, auquel on peut faire appel par la commande **ereturn list** qui s'applique à la régression économétrique la plus récente. Reprenons l'exemple de l'estimation de l'impact du tourisme sur la croissance économique.

[15] Pour connaître les options de la commande **predict**, consulter l'aide associée à la commande de régression concernée.

```
. quietly reg growth tourism lyo prim  infl sw, robust

. ereturn list

scalars:
                  e(N) =  58
               e(df_m) =  5
               e(df_r) =  52
                  e(F) =  9.556990649892306
                 e(r2) =  .5703164729191408
               e(rmse) =  2.062225131298739
                e(mss) =  293.5233837456215
                e(rss) =  221.1441695923252
               e(r2_a) =  .5290007491613659
                 e(ll) =  -121.111217977872
               e(ll_0) =  -145.6077013969634

macros:
            e(depvar) : "growth"
               e(cmd) : "regress"
           e(predict) : "regres_p"
             e(model) : "ols"
           e(vcetype) : "Robust"

matrices:
                e(b) :  1 x 6
                e(V) :  6 x 6

functions:
            e(sample)
```

Avec la commande **quietly** devant **reg**, Stata effectue la régression mais ne restitue pas le tableau des résultats. Avec la commande **ereturn list**, on note que le nombre d'observation est stocké dans le scalaire **e(N)**, le R^2 dans le scalaire **e(r2)**, le R^2 ajusté dans le scalaire **e(r2_a)**, la somme des carrés des résidus dans le scalaire **e(rss)**, etc. L'ensemble des coefficients est stocké dans la matrice **e(b)** et **e(V)** est la matrice de variances-covariances. Pour plus de détails, consulter les Manuels de Stata.

Ces données peuvent être récupérées à des fins de calculs ultérieurs. Exemples :

scalar *scr* = **e(rss)**	crée un scalaire nommé *scr* égal à la somme des carrés des résidus.
scalar *r_carre* = **e(r2)**	crée un scalaire *r_carre* égal au R^2

Pour chaque variable, la valeur du coefficient et son écart-type sont stockés respectivement dans les scalaires **_b[***nom_variable***]** et **_se[***nom_variable***]**. On peut les utiliser pour faire des calculs. Exemple : le *t* de *student* de la variable *tourism* (4.81) est le ratio de la valeur de son coefficient sur son écart-type, on peut le vérifier ci-dessous avec la commande **display** qui fait appel à la calculatrice de Stata:

```
. display _b[tourism]/_se[tourism]
4.811088
```

Stata offre la possibilité d'inclure automatiquement dans les régressions des variables muettes correspondant à chaque modalité d'une variable spécifiée. Supposons que dans le modèle présenté dans cette section, on souhaite inclure des variables muettes par continent, afin de contrôler pour les variations du taux de croissance économique d'un continent à un autre, en d'autres termes des facteurs géographiques. Soit la variable *conti* qui prend 1 pour les pays d'Afrique du Nord et du Moyen-Orient, 2 pour les pays d'Amérique latine, 3 pour les pays d'Asie et 4 pour les pays d'Afrique. Pour inclure des variables muettes pour chaque continent, il suffit de placer comme suit **xi :** devant **reg**, puis **i.***conti* dans la liste des variables explicatives[16].

```
. xi: reg growth tourism lyo prim  infl sw i.conti, robust
i.conti           _Iconti_1-4         (naturally coded; _Iconti_1 omitted)

Regression with robust standard errors                Number of obs =      58
                                                      F(  8,    49) =   14.69
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.6872
                                                      Root MSE      =  1.8126

------------------------------------------------------------------------------
             |               Robust
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tourism |   2.047687   .4836437     4.23   0.000     1.075769    3.019606
         lyo |  -.5377549   .7027413    -0.77   0.448    -1.949966    .8744566
        prim |  -1.121428   1.065158    -1.05   0.298    -3.261944    1.019087
        infl |  -.1389238   .0585121    -2.37   0.022    -.2565083   -.0213394
          sw |   .9713834   .8253392     1.18   0.245    -.6871977    2.629964
   _Iconti_2 |   1.367795   1.006714     1.36   0.180    -.6552722    3.390861
   _Iconti_3 |   3.410633   .8973515     3.80   0.000     1.607338    5.213929
   _Iconti_4 |   .4657152   1.078325     0.43   0.668    -1.701261    2.632691
       _cons |  -9.463751   3.118764    -3.03   0.004    -15.73114    -3.19636
------------------------------------------------------------------------------
```

Comme le montre le tableau ci-dessus la variable muette pour les pays d'Afrique du Nord et du Moyen-Orient (*conti*=1) a été exclue, donc la comparaison des taux de croissance se fait par rapport à ce groupe de pays de référence. On peut dire qu'en moyenne, les pays de l'Asie (*conti*=3) ont un taux de croissance plus élevé que les pays d'Afrique du Nord et du Moyen-Orient, alors que les performances économiques des pays d'Amérique latine et d'Afrique ne s'y diffèrent pas fondamentalement.

7.2. La méthode des doubles moindres carrés

La méthode des doubles moindres carrés (DMC) est utilisée lorsqu'une ou plusieurs variables sont endogènes au modèles. Il existe trois sources principales de l'endogénéité : *(a)* les erreurs

[16] Le même principe est valable pour les autres commandes de régressions.

de mesures sur les variables explicatives, *(b)* la double causalité : lorsque la variable explicative agit sur la variable dépendante et inversement, et *(c)* le biais de variable omise lorsqu'une variable non incluse dans le modèle est corrélée avec au moins une des variables explicatives. La commande **ivreg** de Stata permet de faire des régressions en DMC. La syntaxe générale est la suivante :

ivreg *var_dep var_exog* **(***var_endo = instrum***) (if, in)**, *options*

var_dep est la variable dépendante, *var_exog* est la liste des variables explicatives supposées exogènes, *var_endo* est la liste des variables explicatives supposées endogènes et *instrum* est la liste des instruments. Le nombre d'instruments utilisé doit être au moins égal au nombre de variables endogènes. Un instrument est une variable corrélée avec la variable supposée endogène mais non corrélée avec le résidu du modèle. Les options de la commande **ivreg** sont similaires à celles précitées pour la commande **reg**. *Idem* pour la commande **predict** pour obtenir les valeurs prédites de la variable dépendante et des résidus.

Reprenons l'exemple de l'impact du tourisme sur la croissance économique. On peut penser que dans ce modèle la variable *tourism* est endogène parce que mesurée avec erreurs, ou parce qu'il peut avoir des variables omises (l'instabilité politique par exemple) qui lui sont corrélées et qui ont également un effet direct sur la croissance économique. L'hypothèse d'orthogonalité des erreurs est donc violée pour les MCO. Pour effectuer la régression en DMC, j'ai choisi d'instrumenter la variable *tourism* par sa valeur retardée d'une décennie (1978-1987).

```
. ivreg growth lyo prim  infl sw (tourism = r_touris), robust

IV (2SLS) regression with robust standard errors       Number of obs =      58
                                                       F(  5,    52) =    8.24
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.5630
                                                       Root MSE      =  2.0798

------------------------------------------------------------------------------
             |               Robust
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tourism |   2.065948   .6358393     3.25   0.002     .7900433    3.341853
         lyo |  -1.059231   .7454959    -1.42   0.161    -2.555178    .4367156
        prim |   .7349815   1.346156     0.55   0.587    -1.966279    3.436242
        infl |  -.1474322   .0566298    -2.60   0.012    -.2610682   -.0337961
          sw |   1.501175   .7814268     1.92   0.060    -.0668728    3.069222
       _cons |  -8.761493   2.521005    -3.48   0.001    -13.82026   -3.702727
------------------------------------------------------------------------------
Instrumented:  tourism
Instruments:   lyo prim infl sw r_touris
------------------------------------------------------------------------------
```

Les résultats ci-dessus montrent que la partie exogène de la variable *tourism* exerce un effet positif et significatif sur la croissance économique.

7.3. Les tests

7.3.1. Test de normalité des résidus

La commande **sktest** permet de faire le test de normalité d'une variable donné. Dans l'exemple ci-dessous, les résidus sont prédits après l'estimation du modèle, puis la commande **sktest** est appliquée à la variable nommée *residu* qui contient les valeurs des résidus.

```
. quietly reg growth tourism lyo prim  infl sw, robust

. predict residu, resid
(5 missing values generated)

. sktest residu

                    Skewness/Kurtosis tests for Normality
                                                   ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
      residu |      0.544         0.411            1.08       0.5823
```

La probabilité du test est 0.58, on ne peut donc pas rejeter l'hypothèse Ho de normalité des erreurs.

7.3.2. Test de Ramsey Reset

Le test de Ramsey Reset permet de tester l'omission de variable explicative pertinente ou une mauvaise spécification du modèle. La statistique du test est obtenue avec la commande **ovtest**.

```
. quietly reg growth tourism lyo prim  infl sw, robust

. ovtest

Ramsey RESET test using powers of the fitted values of growth
       Ho:  model has no omitted variables
                  F(3, 49) =      1.24
                  Prob > F =      0.3051
```

La probabilité du test est 0.30, on ne peut donc pas rejeter l'hypothèse Ho au seuil de 10%.

Remarque : le test de Ramsey Reset (**ovtest**) ne se fait qu'après la commande **reg** uniquement.[17]

7.3.3. Test d'hétéroscédasticité

La commande **hettest** utilise le test de Breush-Pagan pour tester l'hypothèse d'homoscédasticité des résidus. Elle fonctionne sur le même principe que la commande **ovtest**, mais la différence est que la régression qui précède la commande **hettest** ne doit pas comporter d'option **robust**.

7.3.4. Test de Chow

Le test de Chow permet de tester la stabilité des coefficients de la régression sur deux sous échantillons différents. Malheureusement ce test n'est pas préprogrammé sur Stata. Cependant, on peut facilement réaliser ce test en suivant étape par étape la procédure suivante :

(1) On fait une estimation sur l'ensemble de l'échantillon, d'où on récupère la somme des carrés des résidus (SCR)

(2) On fait également une estimation sur chacun des deux sous échantillons, puis on extrait la somme des carrés des résidus respectifs (SCR1 et SCR2)

(3) On calcule la statistique du test qui suit une loi de Fisher :

$$\frac{SCR-(SCR1+SCR2)}{SCR1+SCR2} * \frac{n-2k}{k} \rightarrow F(k, n-2k)$$

k est le nombre de variables explicatives y compris la constante, et n est le nombre d'observations.

(4) Si la statistique calculée est inférieure à la statistique lue, on peut rejeter l'hypothèse de constance des coefficients.

Exemple : Test de la stabilité des coefficients du modèle de l'impact du tourisme sur la croissance économique sur deux sous échantillons différents : le groupe des pays africains et celui des autres pays de l'échantillon.

[17] Après la commande **anova** également, mais celle-ci n'est pas abordée dans ce manuel.

```
. *test de chow
. *regression sur tout l'echantillon
.
. reg growth tourism lyo prim infl sw, robust

Regression with robust standard errors                Number of obs =      58
                                                      F(  5,    52) =    9.56
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.5703
                                                      Root MSE      =  2.0622

------------------------------------------------------------------------------
             |               Robust
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tourism |   2.533977   .5266953     4.81   0.000     1.477086    3.590869
         lyo |  -1.260772   .7039543    -1.79   0.079    -2.673359    .1518156
        prim |   .2660154   1.245608     0.21   0.832     -2.23348    2.765511
        infl |   -.140905   .0566845    -2.49   0.016    -.2546509   -.0271592
          sw |   1.284783   .7505155     1.71   0.093    -.2212367    2.790802
       _cons |   -10.2673   2.364197    -4.34   0.000     -15.0114   -5.523187
------------------------------------------------------------------------------

. *recuperation de la somme des carres des residus
. scalar scr=e(rss)

. *recuperation du nombre d'observations
. scalar n=e(N)

. *regression sur l'echantillon des pays africains
. reg growth tourism lyo prim infl sw if afrique==1, robust

Regression with robust standard errors                Number of obs =      22
                                                      F(  5,    16) =  132.30
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.7468
                                                      Root MSE      =  1.5572

------------------------------------------------------------------------------
             |               Robust
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tourism |   2.193158   .5804482     3.78   0.002     .9626628    3.423653
         lyo |  -3.876517   1.338361    -2.90   0.011    -6.713715   -1.039319
        prim |  -.0937218   1.071642    -0.09   0.931    -2.365501    2.178058
        infl |  -.2236358   .0134453   -16.63   0.000    -.2521386    -.195133
          sw |   1.050236   1.045604     1.00   0.330    -1.166346    3.266817
       _cons |  -2.024132    4.35443    -0.46   0.648    -11.25511    7.206848
------------------------------------------------------------------------------

. *recuperation de la somme des carres des residus
. scalar scr1=e(rss)

. *regression sur l'echantillon des pays non africains
. reg growth tourism lyo prim infl sw if afrique==0, robust

Regression with robust standard errors                Number of obs =      36
                                                      F(  5,    30) =    7.98
                                                      Prob > F      =  0.0001
                                                      R-squared     =  0.3178
                                                      Root MSE      =   2.179

------------------------------------------------------------------------------
             |               Robust
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tourism |    1.96849   .8032893     2.45   0.020     .3279546    3.609026
         lyo |  -1.251909   .7560031    -1.66   0.108    -2.795874    .2920549
        prim |   .8802813   3.322838     0.26   0.793    -5.905859    7.666422
        infl |  -.0872288   .0282095    -3.09   0.004    -.1448402   -.0296173
          sw |   .5655836   1.268894     0.45   0.659    -2.025844    3.157011
       _cons |   -6.81276   4.953607    -1.38   0.179    -16.92938    3.303856
------------------------------------------------------------------------------

. *recuperation de la somme des carres des residus
. scalar scr2=e(rss)
```

```
. *calcul de la statistique du test
. scalar stat=((scr-(scr1+scr2))/(scr1+scr2))*((n-2*6)/6)

. *calcul de la probabilite du test
. display F(6,n-2*6,stat)
.85462085
```

La dernière ligne de commande permet de faire appel à la loi de Ficher F(*n1*,*n2*) pour calculer la probabilité pour une variable *x* (*n1* et *n2* étant les degrés de liberté). La syntaxe générale est : **display F(*n1*,*n2*,*x*)**. Pour le test de Chow ci-dessus la probabilité est de 0.85, on ne peut donc pas rejeter l'hypothèse Ho de stabilité des coefficients entre les deux sous échantillons.

7.3.5. Test d'endogénéité

Il n'existe pas de commande préprogrammée de test d'endogénéité sur Stata[18]. Cependant, on peut tester l'endogénéité grâce au test de Nakamura Nakamura qui se fait en deux étapes comme suit :

(1) Chaque variable endogène est régressée sur les variables exogènes du modèle et ses instruments.

(2) Les résidus de la première étape sont récupérés et inclus dans le modèle initial. Si les coefficients des résidus sont conjointement significatifs (Test de Fisher, section 7.3.7) alors on ne peut pas rejeter l'endogénéité des variables testées. Dans le cas d'une seule variable endogène (comme ci-dessous), la significativité du *t* de *student* du résidu permet de conclure ou non au rejet de l'hypothèse d'exogénéité.

Exemple : test de l'endogénéité de la variable *tourism* dans l'équation de croissance.

```
. *Etape 1
. reg tourism r_touris lyo prim  infl sw, ro

Regression with robust standard errors                Number of obs =      58
                                                      F(  5,    52) =  115.01
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.9284
                                                      Root MSE      =  .21384

------------------------------------------------------------------------------
             |               Robust
     tourism |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    r_touris |   1.015061   .0579243    17.52   0.000     .8988272    1.131295
         lyo |   -.120351   .0710574    -1.69   0.096    -.2629382    .0222361
        prim |   .0917429   .1559471     0.59   0.559    -.2211878    .4046737
        infl |    .002837   .0046517     0.61   0.545    -.0064973    .0121714
          sw |   .2427379   .0866992     2.80   0.007     .0687633    .4167125
       _cons |   .2924139   .2742528     1.07   0.291    -.2579147    .8427425
------------------------------------------------------------------------------
. *Etape 2
. predict res_tourism, re
```

[18] Le test d'Hausman programmé sur Stata peut-être utilisé comme test d'endogénéité (voir section 8.5)

```
(5 missing values generated)

. reg growth tourism res_tourism lyo prim infl sw , ro

Regression with robust standard errors                Number of obs =      58
                                                      F(  6,    51) =   11.55
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.6165
                                                      Root MSE      =  1.9673

------------------------------------------------------------------------------
             |               Robust
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tourism |   2.065948    .549608     3.76   0.000     .9625641    3.169332
 res_tourism |   3.403889   1.512939     2.25   0.029     .3665364    6.441241
         lyo |  -1.059231   .7055202    -1.50   0.139    -2.475622    .3571596
        prim |   .7349815   1.325386     0.55   0.582    -1.925842    3.395804
        infl |  -.1474322   .0452672    -3.26   0.002    -.2383099   -.0565544
          sw |   1.501175   .6666259     2.25   0.029     .1628675    2.839482
       _cons |  -8.761493   2.190219    -4.00   0.000    -13.15854   -4.364445
------------------------------------------------------------------------------
```

Dans le tableau ci-dessus, le résidu de l'équation de la première étape est significativement corrélé à la croissance économique, ce qui tend à soutenir l'hypothèse d'endogénéité de la variable *tourism*.

7.3.6. Test de validité des instruments

Le test de suridentification de Sargan permet de tester la validité des instruments utilisés dans les régressions en doubles moindres carrés. Ce test n'est pas préprogrammé sur Stata, il faut télécharger le module correspondant d'Internet (pour l'ajout des programmes additionnels à Stata, voir section 12). La commande est **overid** et fonctionne sous le même principe que la commande **ovtest**, mais la régression qui la précède ne doit pas contenir d'option **robust**.

Reprenons toujours le modèle de l'impact du tourisme sur la croissance économique. Une condition nécessaire pour réaliser le test de Sargan est que le modèle soit suridentifié, le nombre d'instruments doit être strictement supérieur au nombre de variables endogènes. Donc pour instrumenter la variable *tourism*, il faut trouver un autre instrument en plus de sa valeur retardée d'une décennie. Une variable exogène qui est susceptible d'être corrélée au tourisme et non corrélée à la croissance économique est par exemple le nombre de sites classés au patrimoine de l'UNESCO dans chaque pays. Cette variable est dénommée *unes*.

```
. ivreg growth lyo prim  infl sw (tourism = r_touris unes)

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =      58
-------------+------------------------------           F(  5,    52) =   11.36
       Model |  289.659861     5  57.9319721           Prob > F      =  0.0000
    Residual |  225.007693    52  4.32707101           R-squared     =  0.5628
-------------+------------------------------           Adj R-squared =  0.5208
       Total |  514.667553    57  9.02925532           Root MSE      =  2.0802

------------------------------------------------------------------------------
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tourism |   2.061324   .5385482     3.83   0.000     .9806483       3.142
         lyo |   -1.05724   .7911785    -1.34   0.187    -2.644856    .5303756
        prim |   .7396146   1.530201     0.48   0.631    -2.330958    3.810187
        infl |  -.1474966   .0456694    -3.23   0.002    -.2391389   -.0558544
          sw |   1.503313   .8358417     1.80   0.078    -.1739264    3.180552
       _cons |  -8.746617   2.572441    -3.40   0.001     -13.9086   -3.584637
------------------------------------------------------------------------------
Instrumented:  tourism
Instruments:   lyo prim infl sw r_touris unes
------------------------------------------------------------------------------

. overid

Tests of overidentifying restrictions:
Sargan N*R-sq test        0.415  Chi-sq(1)    P-value = 0.5196
Basmann test              0.367  Chi-sq(1)    P-value = 0.5445
```

La probabilité du test de Sargan est de 0.52, on ne peut donc pas rejeter l'hypothèse Ho de la validité des instruments. La commande **overid** donne également la statistique et la probabilité du test de suridentification de Basmann.

Remarque 1 : Le test de Sargan peut se faire étape par étape pour les cas où la commande **overid** ne s'applique pas[19] :

(a) on estime l'équation structurelle avec les doubles moindres carrés, puis on récupère les résidus.

(b) on régresse les résidus sur toutes les variables exogènes y compris les instruments, puis on récupère le R^2.

(c) sous l'hypothèse nulle de la validité des instruments, la statistique $n * R^2$ suit une loi de chi2 à q degrés de liberté, n étant le nombre d'observations et q est le nombre d'instruments extérieurs au modèle moins le nombre de variables explicatives endogènes.

Remarque 2 : Il existe une version étendue de la commande **ivreg** disponible sur Internet. Elle est dénommée **ivreg2**. Elle permet entre autres, d'obtenir la statistique du test de

[19] Voir section 8.8.6

suridentification de Hansen[20] avec l'option **robust** et de faire des estimations en DMC par la méthode des moments généralisés (GMM)

7.3.7. Test sur les coefficients des variables

a) Test linéaire

La commande pour effectuer les tests linéaires sur les coefficients des variables d'une régression est **test**. Cette commande s'applique à la régression la plus récente estimée. Soit *x*, *y* et *z* des variables explicatives d'un modèle, voici les commandes associées à un test linéaire, avec l'hypothèse Ho correspondant :

test *x*	Ho : le coefficient de la variable *x* n'est pas différent de zéro[21]
test *x* = #	Ho : le coefficient de la variable *x* n'est pas différent de #
test *x y z*	Ho : les coefficients des variables *x*, *y* et *z* ne sont pas différents de 0, ce test correspond au *F test* de Fisher
test *x* + *y* = *b*	Ho : la somme des coefficients des variables *x* et *y* n'est pas différente de 0
test *x* = *y* ou **test** *x*-*y* = 0	Ho : le coefficient de la variable *x* n'est pas différent de celui de *y*
test (*x* = *c***) (***y* = *d***)**	Ho : le coefficient de la variable *x* n'est pas différent de *c* et celui de la variable *y* n'est pas différent de *d*.

Remarque : on peut remplacer le nom de la variable par la macro contenant la valeur de son coefficient, en d'autres termes on peut écrire **test** *x* ou **test _b[***x***]**. Ce ne sera pas le cas des tests non linéaires, où seule l'expression **_b[***x***]** est permise pour désigner le coefficient de la variable *x*.

b) Test non linéaire

testnl _b[*x***]/_b[***y***]=_b[***z***]**	Ho: le ratio du coefficient de la variable *x* sur le coefficient de la variable *y* n'est pas différent de la valeur du coefficient de la variable *z*.

Pour tester plusieurs hypothèses de manière conjointe, il suffit de les mettre chacune entre parenthèses comme pour la commande **test**.

[20] Le test de Hansen est robuste à la présence d'hétéroscédasticité, alors que le test de Sargan ne l'est pas.
[21] La probabilité de ce test correspond à celle du *t* de *student*.

8. Les régressions sur données de panel

8.1. La commande *collapse*

La commande **collapse** permet de transformer une base de données à l'aide de fonctions comme la moyenne, l'écart-type, le maximum, le minimum, etc.

Syntaxe générale : **collapse** *(fonctions) variables*

Exemple : **collapse (mean)** *x y* **(sd)** *z*
Si la ligne de commande ci-dessus est appliquée à une base de donnée contenant les variables *x*, *y* et *z*, la base de données résultante sera composée d'une seule observation dont la valeur de la variable *x* (respectivement *y*) sera égale à la moyenne de cette même variable sur tout l'échantillon, tandis que la valeur de la variable *z* sera égale à son écart-type également calculé sur tout l'échantillon.

La combinaison de la commande **collapse** avec **by** peut-être assez utile pour calculer des moyennes par période nécessaires pour les données de panel. Supposons que vous disposiez d'une base de données avec des données annuelles par pays. Soit *pays*, la variable qui contient le nom des pays, et *periode* une variable indicatrice qui attribue à chaque année une valeur pour la période concernée. Les commandes qui suivent permettent de transformer la base de données annuelles en une base de données moyennes par pays et par période.

sort *pays periode*
collapse (mean) *variable 1 variable2 … variableN***, by(***pays periode***)**

Remarque 1 : la commande **collapse** modifie radicalement la base de données, pour pouvoir retrouver la base de données d'origine, il faut appliquer la commande **preserve** avant la commande **collapse**. Une copie de la base est ainsi faite et celle-ci peut-être restituée par la commande **restore**.

Remarque 2 : Il existe une autre version de la commande **collapse** qui se nomme **collapse2** et qui est disponible sur Internet. Cette commande dispose de fonctions supplémentaires comme

la fonction **first** pour générer les valeurs de début de période et la fonction **last** pour générer des valeurs de fin de période.

8.2. Les statistiques descriptives

Les commandes **xtsum** et **xttab** sont les versions panel des commandes **sum** et **tab** (le principe de fonctionnement et les syntaxes sont les mêmes). Elles donnent des statistiques intra et inter-individuelles des données.

Mais avant d'utiliser ces commandes, il faut indiquer à Stata la variable de dimension transversale et la variable de dimension temporelle. Cela se fait par la commande **tsset**.

Exemple : **tsset** *id tps* déclare des données de panel où *id* est la variable indicatrice de la dimension transversale et *tps* l'indicatif des périodes.

La commande **tsset** doit précéder les commandes sur données de panel, à défaut la variable de dimension transversale doit être spécifiée à chaque commande. Il est néanmoins préférable de mettre la commande **tsset** en début du fichier *do* juste après la commande **preserve**.

En somme, en amont des commandes sur des données de panel, il faut spécifier la variable de dimension transversale et la variable de dimension temporelle par **tsset**.

Pour les commandes qui suivent, il est sous-entendu que la commande **tsset** doit les précéder.

8.3. Le modèle à effets fixes

La commande **xtreg**…, **fe** de Stata permet d'estimer un modèle à effets fixes. La syntaxe générale est la suivante :

xtreg *var_dep var_explicatives* **(if, in)**, **fe i (***id***)**

L'option **fe** spécifie les effets fixes. La variable *id* est la variable d'identification de la dimension transversale des données. On peut se passer de cette option lorsque la commande **tsset** définit auparavant les variables de dimension transversale et de dimension temporelle.

Dans l'exemple qui suit, j'estime de nouveau l'impact du tourisme sur la croissance économique, mais avec des données en moyennes quinquennales sur la période 1968-1997. Chaque pays dispose théoriquement de six points d'observations.

```
. xtreg growth tourism lyo prim infl sw,fe

Fixed-effects (within) regression               Number of obs      =       114
Group variable (i): id                          Number of groups   =        58

R-sq:  within  = 0.6368                         Obs per group: min =         1
       between = 0.0018                                        avg =       2.0
       overall = 0.0036                                        max =         2

                                                F(5,51)            =     17.89
corr(u_i, Xb)  = -0.9431                        Prob > F           =    0.0000

------------------------------------------------------------------------------
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tourism |   .0270759   .0057554     4.70   0.000     .0155214    .0386304
         lyo |  -.0873546   .0135308    -6.46   0.000    -.1145189   -.0601903
        prim |   .0006829   .0002603     2.62   0.011     .0001604    .0012053
        infl |  -.0007259   .0008021    -0.90   0.370    -.0023363    .0008845
          sw |   .0107406   .0072223     1.49   0.143    -.0037587    .0252399
       _cons |   .1805307   .1039865     1.74   0.089    -.0282309    .3892923
-------------+----------------------------------------------------------------
     sigma_u |  .07662412
     sigma_e |  .01232982
         rho |  .97476053   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(57, 51) =     4.16                 Prob > F = 0.0000
```

La commande **xtreg** calcule trois statistiques de R^2. Pour le modèle à effets fixes, le R^2 le plus pertinent est le R^2 *within* car il donne une idée de la part de la variabilité intra-individuelle de la variable dépendante expliquée par celles des variables explicatives. Le R^2 *between* quant à lui donne une idée de la contribution des effets fixes au modèle.

Dans le tableau, il existe deux statistiques de test de Fisher. La première (en haut du tableau) teste la significativité conjointe des variables explicatives et la seconde (en bas du tableau) teste la significativité conjointe des effets fixes introduits.

La commande **xtreg**…, **fe** ne permet pas d'option **robust** pour la correction de l'hétéroscédasticité par la méthode de White. Pour le faire, il faut utiliser une autre commande nommée **areg** :

```
. areg growth tourism lyo prim infl sw, absorb(id) robust

Regression with robust standard errors                Number of obs =     114
                                                      F(  5,    51) =   13.27
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.9020
                                                      Adj R-squared =  0.7829
                                                      Root MSE      =  .01233

------------------------------------------------------------------------------
             |               Robust
      growth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tourism |   .0270759   .0055758     4.86   0.000      .015882    .0382698
         lyo |  -.0873546   .0138752    -6.30   0.000    -.1152102    -.059499
        prim |   .0006829   .0002554     2.67   0.010       .00017    .0011957
        infl |  -.0007259   .0009661    -0.75   0.456    -.0026654    .0012136
          sw |   .0107406   .0059401     1.81   0.076    -.0011846    .0226658
       _cons |   .1805307   .1096602     1.65   0.106    -.0396213    .4006827
-------------+----------------------------------------------------------------
          id |   absorbed                                     (58 categories)
```

La commande **areg** nécessite de spécifier la dimension transversale même si la commande **tsset** est utilisée avant. L'option **absorb(***id***)** spécifie la variable *id* comme représentant la dimension transversale. La différence entre **xtreg**…, **fe** et **areg** est qu'avec le premier, les données sont transformées en différences par rapport à la moyenne individuelle pour éliminer les effets fixes, alors que la commande **areg** revient à faire des MCO sur un modèle dans lequel on introduit une variable muette pour chaque individu ou pays.

Les options de *predict* pour **xtreg**…, **fe** sont :

predict *yhat*, **xb** — prédit la valeur estimée de la variable dépendante et la stocke dans la variable nommée *yhat*.

predict *yhat2*, **xbu** — prédit la somme de la valeur estimée de la variable dépendante et de l'effet fixe et la nomme *yhat2*.

predict *ef*, **u** — prédiction de l'effet fixe stocké dans la variable nommée *ef*.

predict *residu*, **e** — prédiction des résidus qui seront stockés dans la variable *residu*.

predict *ef_residu*, **ue** — la nouvelle variable *ef_residu* créée va contenir les prédictions de la somme du résidu et de l'effet fixe.

Les options de **predict** pour **areg** sont :

predict *yhat*, **xb** — prédit la valeur estimée de la variable dépendante et la stocke dans la variable nommée *yhat*.

predict *yhat2*, **xbd** — prédit la somme de la valeur estimée de la variable dépendante et de l'effet fixe et nomme cette prédiction *yhat2*.

predict *ef*, **d** — prédiction de l'effet fixe stocké dans la variable nommée *ef*.

predict *residu*, **r** — prédiction des résidus qui seront stockés dans la variable *residu*.

predict *ef_residu*, **dr** la nouvelle variable *ef_residu* créée va contenir les prédictions de la somme des résidus et de l'effet fixe.

Pour estimer un modèle à effets fixes avec des variables instrumentales, il faut utiliser la commande **xtivreg**…, **fe** dont la syntaxe générale est :

xtivreg *var_dep var_exgo* **(**var_endo = instrum**) (if, in)**, **fe i(***id***)**

Les options de **predict** sont les mêmes que pour **xtreg**…, **fe**

Remarque : comme **xtreg**…,**fe** , la commande **xtivreg**…, **fe** n'admet pas l'option **robust** pour corriger de l'hétéroscédasticité. Pour réaliser un telle estimation, il faut utiliser la commande **areg2** qui n'est pas intégrée par défaut dans Stata, et donc nécessite d'être téléchargée d'Internet.

8.4. Le modèle à effets aléatoires

Pour estimer un modèle à effets aléatoires, il suffit de remplacer l'option **fe** de **xtreg** par l'option **re**. La syntaxe générale est donc la suivante :

xtreg *var_dep var_explicatives* **(if, in)**, **re i (***id***)**

```
. xtreg growth tourism lyo prim infl sw, re

Random-effects GLS regression                   Number of obs      =       114
Group variable (i): id                          Number of groups   =        58

R-sq:  within  = 0.3317                         Obs per group: min =         1
       between = 0.5010                                        avg =       2.0
       overall = 0.4314                                        max =         2

Random effects u_i ~ Gaussian                   Wald chi2(5)       =     69.81
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      growth |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tourism |   .0091043   .0019265     4.73   0.000     .0053285    .0128802
         lyo |   -.007757   .0031227    -2.48   0.013    -.0138773   -.0016366
        prim |   .0001867   .0001291     1.45   0.148    -.0000663    .0004397
        infl |  -.0023485   .0007261    -3.23   0.001    -.0037716   -.0009254
          sw |   .0116479   .0057053     2.04   0.041     .0004657    .0228301
       _cons |  -.0785984   .0207657    -3.79   0.000    -.1192985   -.0378983
-------------+----------------------------------------------------------------
     sigma_u |  .01403291
     sigma_e |  .01232982
         rho |  .56433383   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Dans le cas de **xtreg**…, **re** , le R^2 le plus pertinent est le R^2 *between*, c'est la mesure de la part de la variabilité inter-individuelle de la variable dépendante expliquée par celles des variables explicatives. Le R^2 *within* quant à lui donne une idée de la contribution des effets aléatoires pays au modèle.

Les options de **predict** pour **xtreg**…, **re** sont les mêmes que pour **xtreg**…, **fe**

La commande **xttest0** après une estimation d'un modèle à effets aléatoires permet d'obtenir la statistique du test de Breusch-Pagan qui teste la significativité des effets aléatoires.

```
. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects:

        growth[id,t] = Xb + u[id] + e[id,t]

        Estimated results:
                         |       Var     sd = sqrt(Var)
                ---------+-----------------------------
                  growth |   .0007004       .0264648
                       e |    .000152       .0123298
                       u |   .0001969       .0140329

        Test:   Var(u) = 0
                             chi2(1) =     4.02
                          Prob > chi2 =     0.0450
```

Dans l'exemple ci-dessus, la probabilité de la statistique du test de Breush-Pagan montre que les effets aléatoires sont globalement significatifs à un seuil de 5%.

En ce qui concerne le modèle à effets aléatoires avec des variables instrumentales, il faut utiliser la commande **xtivreg**…, **re** dont la syntaxe générale est la suivante :

xtivreg *var_dep var_exgo* **(***var_endo* = *instrum***) (if, in)**, **re i(***id***)**

Les options de **predict** restent identiques à celles du modèle sans variables instrumentales.

Remarque : Il existe une version de **xtreg** nommée **xtregar** qui permet d'estimer un modèle à effets fixes ou un modèle à effets aléatoires avec des résidus autocorrélés d'ordre 1. En effet, lorsque la dimension temporelle du panel est relativement grande par rapport à la dimension transversale, on peut être confronté à des problèmes de séries temporelles.

8.5. Le test de Hausman

Les modèles à effets fixes et à effets aléatoires permettent de prendre en compte l'hétérogénéité des données mais les hypothèses sur la nature des effets spécifiques diffèrent d'un modèle à l'autre. Dans le premier cas, on suppose que les effets spécifiques peuvent être corrélés avec les variables explicatives du modèle, et dans le second cas on suppose que les effets spécifiques sont orthogonaux aux variables explicatives du modèle. Le test de spécification de Hausman permet de tester laquelle de ces deux hypothèses est appropriée aux données. En d'autres termes ce test permet de choisir entre le modèle à effets fixes et le modèle à effets aléatoires. La syntaxe est la suivante :

xtreg…, **fe**
est store *eq*1
xtreg…, **re**
hausman *eq*1

La première ligne de commande estime le modèle à effets fixes. La seconde ligne conserve les résultats du modèle précédent sous le nom *eq*1. La troisième ligne estime le modèle à effets aléatoire et la quatrième ligne exécute le test de Hausman proprement dit. Exemple :

```
. * test de Hausman
.
. quietly xtreg growth tourism lyo prim infl sw, fe

. est store eq1

. quietly xtreg growth tourism lyo prim infl sw, re

. hausman eq1

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |      eq1           .          Difference          S.E.
-------------+------------------------------------------------------------
     tourism |    .0270759     .0091043        .0179716        .0054235
         lyo |   -.0873546     -.007757       -.0795976        .0131656
        prim |    .0006829     .0001867        .0004962         .000226
        infl |   -.0007259    -.0023485        .0016226        .0003409
          sw |    .0107406     .0116479       -.0009073        .0044284
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(5) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       50.38
                Prob>chi2 =      0.0000
```

La probabilité du test dans l'exemple ci-dessus est inférieure à 10%, ce qui implique que le modèle à effets fixes est préférable au modèle à effets aléatoires. Lorsque la probabilité du test est supérieure au seuil de 10%, alors le test de Hausman ne permet de différencier le modèle à effet fixes du modèle à effets aléatoires. Dans ce cas, le choix de l'un ou l'autre modèle doit être justifié rigoureusement, et il dépend de la conviction de chaque auteur sur la pertinence d'un modèle par rapport à l'autre. On peut néanmoins se référer à quelques arguments d'ordre général qui facilitent le choix du modèle :

(a) Lorsque la variation intra individuelle des variables est plus forte que la variation inter individuelle[22], le modèle à effets fixes est plus approprié que le modèle à effets aléatoires et vice versa.

(b) Lorsque la dimension temporelle est très réduite, par exemple à deux périodes, le modèle à effets fixes donne de moins bons résultats que le modèle à effets aléatoires.

(c) Lorsqu'il existe dans le modèle une variable explicative invariante dans le temps dont on veut estimer l'impact marginal, on utilisera le modèle à effets aléatoires, mais sous l'hypothèse assez forte d'exogénéité des effets spécifiques.

Remarque 1 : On peut également utiliser le test de Hausman pour tester l'endogénéité d'une ou de plusieurs variables. La procédure est la suivante :

On estime le modèle en variables instrumentales (avec **ivreg**)
est store *eq*1
On estime ensuite le modèle en MCO (avec **reg**)
hausman *eq*1

Si la probabilité du test est inférieure à 10%, alors on rejette l'hypothèse d'exogénéité des variables explicatives instrumentées, la préférence va donc au modèle en DMC.

Remarque 2 : la procédure du test de Hausman diffère de la version 7 de Stata à la version 8. Dans la version 7, la procédure est la suivante :

[22] La commande **xtsum** permet d'obtenir les variances intra individuelles et les variances inter-individuelles.

xtreg…, **fe**		
hausman, save	ou	**xtreg**…, **re**
xtreg…, **re**		**xthausman**
hausman		

Remarque 4 : voir la commande **suest** pour une version généralisée du test de Hausman.

8.6. L'estimateur de Hausman-Taylor

La solution la plus simple pour pallier à la corrélation entre les effets spécifiques et les variables explicatives consiste à éliminer les effets spécifiques en recourant à l'estimateur *within* (**xtreg**…, **fe**) ou à estimer l'équation en première différence. Mais ces transformations ne permettent pas d'estimer l'impact d'une variable explicative invariante dans le temps. L'estimateur des variables instrumentales de Hausman-Taylor permet de lever cette limite.

Parmi les variables explicatives du modèle à estimer, il doit avoir deux groupes de variables : un premier groupe de variables qui varient dans le temps et entre les individus (G1), et un second groupe de variables invariantes dans le temps (G2). On suppose qu'il existe un nombre *n1* de variables appartenant à G1 et un nombre *n2* de variables appartenant à G2 qui sont doublement exogènes, c'est-à-dire corrélées ni avec l'effet spécifique, ni avec le terme d'erreur. Le nombre *n1* doit être supérieur ou égal au nombre de variables appartenant à G2 et qui sont corrélées avec l'effet spécifique, en d'autres termes les variables endogènes.

La syntaxe générale de la commande sur Stata est la suivante:

xthtaylor *var_dep var_explicatives* **(if, in)**, **endo(***liste1***) cons(***liste2***) small am**

liste1 est la liste des variables endogènes du modèle (séparées par un espace) y compris les variables endogènes invariantes dans le temps. Ici le terme endogène désigne une corrélation uniquement avec l'effet spécifique, mais pas avec le terme d'erreur.

liste2 est la liste de l'ensemble des variables invariantes dans le temps. Cette option est facultative car Stata détecte automatiquement les variables constantes dans le temps et celles qui varient selon les individus et le temps.

small est l'option qui permet de reporter les *t* de *student* au lieu des statistiques *z*, et le *F* de Fisher au lieu de la statistique de chi2.

am est une option exclusivement utilisée lorsqu'on souhaite estimer le modèle par la méthode d'Amemiya-MaCurdy au lieu de celle de Hausman-Taylor. Cette méthode alternative utilise des instruments additionnels pour améliorer l'efficience de l'estimateur[23]. Mais elle nécessite un panel cylindré et une période initiale identique pour chaque individu de la dimension transversale.

Les options de **predict** pour **xthtaylor** sont les mêmes que pour **xtreg**…, **fe**

8.7. La méthode des moments généralisés (GMM) en panel dynamique

C'est la méthode « magique » qui fait « fureur » chez les macro-économistes depuis quelques années. Selon ses défenseurs, cette méthode permet d'apporter des solutions aux problèmes de biais de simultanéité, de causalité inverse et de variables omises. La présentation détaillée de cette méthode pour les modèles de croissance se trouve en Annexe de ce document.

Un modèle dynamique est un modèle dans lequel un ou plusieurs retards de la variable dépendante figurent comme variables explicatives. A l'inverse des GMM en panel dynamique, les techniques économétriques standards comme les MCO ne permettent pas d'obtenir des estimations efficientes d'un tel modèle, à cause de la présence de la variable dépendante retardée à droite de l'équation (pour plus d'explications, voir Sevestre, 2002).[24]

Il existe deux variantes d'estimateur des GMM en panel dynamique : *(a)* l'estimateur GMM en première différence et *(b)* l'estimateur GMM en système.

L'estimateur GMM en première différence d'Arellano et Bond (1991) consiste à prendre pour chaque période la première différence de l'équation à estimer pour éliminer les effets spécifiques pays, et ensuite à instrumenter les variables explicatives de l'équation en première différence par leurs valeurs en niveau retardées d'une période ou plus. Quant à l'estimateur

[23] Voir Sevestre (2002)
[24] Il est à noter que la méthode GMM permet également d'estimer des modèles non dynamiques dont certaines variables explicatives sont endogènes.

GMM en système de Blundel et Bond (1998), il combine les équations en première différence avec les équations en niveau dans lesquelles les variables sont instrumentées par leurs premières différences. Blundel et Bond (1998) ont montré à l'aide des simulations de Monte Carlo que l'estimateur GMM en système est plus performant que celui en première différence, ce dernier donne des résultats biaisés dans des échantillons finis lorsque les instruments sont faibles.

Deux tests sont associés à l'estimateur des GMM en panel dynamique : le test de suridentification de Sargan/Hansen qui permet de tester la validité des variables retardées comme instruments, et le test d'autocorrélation d'Arellano et Bond où l'hypothèse nulle est l'absence d'autocorrélation de second ordre des erreurs de l'équation en différence.

Dans le modèle à estimer, l'utilisation des variables retardées comme instruments diffère selon la nature des variables explicatives :

(a) Pour les variables exogènes, leurs valeurs courantes sont utilisées comme instruments
(b) Pour les variables prédéterminées ou faiblement exogènes (des variables qui peuvent être influencées par les valeurs passées de la variable dépendante, mais qui restent non corrélées aux réalisations futures du terme d'erreur), leurs valeurs retardées d'au moins une période peuvent être utilisées comme instruments.
(c) Pour les variables endogènes, seules leurs valeurs retardées d'au moins deux périodes peuvent être des instruments valides.

Seul l'estimateur des GMM en première différence est préprogrammé sur Stata sous la commande **xtabond**. Mais en plus, cette commande ne permet d'estimer que les modèles dynamiques (modèles dans lesquels la variable dépendante retardée est incluse comme variable explicative). La commande **xtabond2** disponible sur Internet offre une alternative plus intéressante. Elle permet d'estimer des modèles dynamiques et non dynamiques aussi bien avec l'estimateur GMM en différence que l'estimateur GMM en système. La syntaxe générale est la suivante :

> **xtabond2** *var_dep var_explicatives* **(if, in)**, **noleveleq gmm(***liste1, options1***) iv(***list2, options2***) two robust small**

L'option **noleveleq** permet de spécifier l'estimateur GMM en différence. Lorsque cette option est omise, alors c'est l'estimateur GMM en système qui est utilisé.

gmm(*liste1, options1***)** : *liste1* est la liste des variables explicatives non exogènes du modèle ; *options1* regroupe les options suivantes qui doivent être séparées par des espaces : **lag(***a b***)**, **eq(diff)**, **eq(level)**, **eq(both)** et **collapse**.

- **lag(***a b***)** signifie que pour l'équation en différence, les variables retardées en niveau (de chaque variable de *list1*) datées de *t-a* à *t-b* seront utilisées comme instruments, alors que pour l'équation en niveau ce seront les différences premières datées de *t-a+1* qui seront utilisées comme instruments. Si $b = \bullet$, cela signifie que *b* est infini. Par défaut, $a = 1$ et $b = \bullet$. Exemple 1 : **gmm(***x y*, **lag(**2 .**))** $\Rightarrow$ toutes les valeurs retardées de *x* et *y* datées d'au moins deux périodes seront utilisées comme instruments. Exemple 2 : **gmm(***x*, **lag(**1 2**)) gmm(***y*, **lag(**2 3**))** $\Rightarrow$ pour la variable *x* les valeurs retardées d'une période et de deux périodes seront utilisées comme instruments alors que pour la variable *y* les valeurs retardées de deux et trois périodes seront utilisées comme instruments.
- Les options **eq(diff)**, **eq(level)** ou **eq(both)** signifient que les instruments doivent être utilisés respectivement pour l'équation en différence première, pour l'équation en niveau ou pour les deux équations. L'option par défaut est **eq(both)**.
- l'option **collapse** réduit la taille de la matrice des instruments et permet d'éviter le biais de sur-instrumentation dans des petits échantillons lorsque le nombre d'instruments s'approche du nombre d'observations, mais elle réduit l'efficience statistique de l'estimateur dans les grands échantillons.

iv(*list2*, *options2***)** : *list2* est la liste des variables strictement exogènes du modèle ou extérieures au modèle et *options2* regroupe les options **eq(diff)**, **eq(level)**, **eq(both)**, **pass** et **mz**.

- **eq(diff)**, **eq(level)** ou **eq(both)** ont les même fonctions que ci-dessus.
- Par défaut, les variables exogènes sont différenciées pour servir d'instruments dans les équations en différence première et sont utilisées telles quelles comme instruments dans les équations en niveau. L'option **pass** permet d'éviter que les variables exogènes soient différenciées pour instrumenter les équations en différence première (Exemple : **gmm(***x*, **eq(level)) gmm(**x, **eq(diff) pass)** permet d'utiliser la variable *x* en niveau

comme instrument dans l'équation en niveau aussi bien que dans l'équation en différence.

- L'option **mz** remplace les valeurs manquantes des variables exogènes par un zéro, permettant ainsi aux observations dont les données sur les variables exogènes sont manquantes d'être incluses dans la régression. Cette option n'a d'impact sur les coefficients que si les variables exogènes sont extérieures au modèle.

L'option **two** spécifie l'utilisation de l'estimateur GMM en deux étapes, mais cet estimateur à deux étapes, bien qu'il soit asymptotiquement plus efficace, entraîne des résultats biaisés. Pour pallier ce problème, la commande **xtabond2** procède à une correction de la matrice des covariances pour les échantillons finis. La question de savoir si l'estimateur GMM en une seule étape est meilleur que l'estimateur GMM corrigé en deux étapes (ou inversement) reste pour l'instant sans réponse.

L'option **robust** permet de corriger les *t* de *student* de l'hétéroscédasticité, alors que l'option **small** reporte les *t* de *student* à la place des statistiques *z*.

La commande **xtabond2** reporte par défaut les statistiques du test de Sargan/Hansen, et celles du test d'autocorrélation du premier et du second ordre.

Exemple : Impact du développement financier sur la croissance économique

Le modèle estimé est le suivant :

$$ly_{i,t} - ly_{i,t-1} = (\alpha - 1) * ly_{i,t-1} + \delta * df + \beta' X_{i,t} + u_i + v_t + e_{it}$$

Où ly_{it} représente le logarithme du PIB par tête réel, *df* le niveau de développement financier mesuré par le volume des crédits au secteur privé sur le PIB, *X* représente les variables de contrôle telles que le taux de scolarisation primaire (*prim*), l'ouverture commerciale (*sw*) et l'inflation (*infl*), *u* est l'effet spécifique pays, *v* est l'effet spécifique temporel et *e* le terme d'erreur, *i* et *t* représentent respectivement l'indice pays et l'indice temporel.

Avant d'estimer le modèle, on peut réécrire l'équation ci-dessus sous une forme dynamique :

$$ly_{i,t} = \alpha * ly_{i,t-1} + \delta * df + \beta' X_{i,t} + u_i + v_t + e_{it}$$

L'échantillon est composé de 63 pays, les données sont des moyennes quinquennales sur la période 1968-1997 (6 sous périodes de cinq ans).

```
. xtabond2 ly L.ly df prim sw infl tps3 tps4 tps5 tps6, robust small  iv(tps3 tps4
tps5 tps6) gmm(L.ly prim sw infl, lag(1 .) collapse) gmm(df, lag(2 .) collapse)

Building GMM instruments......
Estimating.
Performing specification tests.

Arellano-Bond dynamic panel-data estimation, one-step system GMM results
------------------------------------------------------------------------------
Group variable: id                              Number of obs      =       251
Time variable : tri5                            Number of groups   =        59
Number of instruments = 33                      Obs per group: min =         1
F(9, 58)      =    359.90                                      avg =      4.25
Prob > F      =     0.000                                      max =         5
------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ly           |
          L1 |   .9844256   .0303777    32.41   0.000      .923618    1.045233
df           |   .0037217    .001148     3.24   0.002     .0014237    .0060196
prim         |    .003489   .0016603     2.10   0.040     .0001656    .0068123
sw           |   .0839188   .0444522     1.89   0.064     -.005062    .1728996
infl         |  -.0387624   .0125066    -3.10   0.003    -.0637972   -.0137277
tps3         |  -.1064364   .0273141    -3.90   0.000    -.1611116   -.0517611
tps4         |  -.1272884   .0245494    -5.18   0.000    -.1764294   -.0781473
tps5         |  -.1212613   .0328504    -3.69   0.000    -.1870185   -.0555041
tps6         |  -.1371423   .0401844    -3.41   0.001    -.2175801   -.0567045
_cons        |   -.141556   .2123163    -0.67   0.508    -.5665531    .2834412
------------------------------------------------------------------------------
Hansen test of overid. restrictions: chi2(23) =   29.41   Prob > chi2 =  0.167

Arellano-Bond test for AR(1) in first differences: z =  -2.51  Pr > z =  0.012
Arellano-Bond test for AR(2) in first differences: z =  -1.60  Pr > z =  0.110
------------------------------------------------------------------------------
```

Note : La commande **xtabond2** et ses options sont saisies sur une même ligne dans le fichier *do*. La variable *L.ly* (qui représente le niveau du PIB par tête initial) est la valeur retardée de la variable *ly* (voir section 10.1 pour la création et l'utilisation de variables retardées). Les variables *tps3* à *tps6* sont les variables muettes temporelles[25]. Le développement financier est considéré comme une variable endogène, ce qui justifie l'utilisation de l'option **lag(**2 .**)** pour ses instruments. Les autres variables explicatives sont supposées prédéterminées, d'où l'option **lag(**1 .**)**. Compte tenu de la faiblesse de l'échantillon, l'option **collapse** a été utilisée pour limiter le biais de sur-instrumentation, étant donné que tous les retards des variables ont été utilisés comme instruments.

L'estimation du modèle en panel dynamique par la commande **xtabond2** donne la valeur α du coefficient du PIB par tête initial. Mais il faut calculer la valeur du coefficient de cette variable qui est $\alpha - 1$ dans le modèle de croissance. Il faut également calculer le *t* de *student*

[25] 4 variables muettes temporelles sont introduites dans le modèle parce que sur les six périodes, la première n'est pas prise en compte dans les régressions à cause de la présence de la variable dépendante retardée.

du coefficient $\alpha - 1$ qui est égale à $\frac{(\alpha - 1)}{\text{écart-type de } \alpha}$. La commande **lincom**[26] de Stata permet d'obtenir ces deux valeurs comme ci-dessous.

```
. lincom L.ly - 1

 ( 1)  L.ly = 1

------------------------------------------------------------------------------
          ly |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0155744   .0303777    -0.51   0.610     -.076382    .0452332
------------------------------------------------------------------------------
```

Le coefficient du PIB initial est donc égale à -0.0156, et est non significatif à 10% (*t*=-0.51). Pour le coefficient du développement financier, il est positif et significatif à 1%, ce qui soutient l'hypothèse que le développement financier exerce un effet favorable sur la croissance économique. Un accroissement du volume des crédits au secteur privé sur le PIB de 10% entraîne 3 points de pourcentage de taux de croissance additionnelle. Les résultats montrent également qu'un taux de scolarisation primaire élevé, une plus grande ouverture commerciale et un faible taux d'inflation sont bénéfiques pour la croissance économique.

Le test de Hansen ($p = 0.167$) et le test d'autocorrélation de second ordre d'Arellano et Bond ($p = 0.11$) ne permettent pas de rejeter l'hypothèse de validité des variables retardés en niveau et en différences comme instruments, et l'hypothèse d'absence d'autocorrélation de second ordre.

8.8. Les tests sur données de panel

8.8.1. Le test de normalité des résidus

Le test de normalité des résidus se fait de la même façon que pour les données transversales. La commande reste toujours **sktest**, mais le résidu prédit provient d'un modèle estimé par **xtreg**…, **fe** ou **xtreg**…, **re** selon l'estimateur choisi.

8.8.2. Le test de Ramsey-Reset

Le test de Ramsey Reset n'est pas disponible pour les régressions en panel après la commande **xtreg**, il faut donc le faire étape par étape comme suit :

[26] Cette commande permet de calculer la valeur, l'écart-type, le *t* de *student* et l'intervalle de confiance d'une combinaison linéaire des coefficients du modèle estimé le plus récent (voir l'aide de Stata pour plus de détails).

(1) Régresser le modèle à l'aide de la commande **xtreg**, puis récupérer les valeurs prédites de la variable dépendante.
(2) Estimer de nouveau le modèle structurel en introduisant comme variables explicatives additionnelles la variable dépendante prédite au carré, au cube et à la puissance 4.[27]
(3) Faire un test de Fisher (voir section 7.3.7) sur la significativité globale de ses nouvelles variables introduites. Si la probabilité du *F test* est supérieure à 10%, on ne peut donc pas rejeter l'hypothèse Ho d'une bonne spécification du modèle.

8.8.3. Le test d'hétéroscédasticité

Pour les régressions en panel, le test d'hétéroscédasticité de Breush-Pagan est donné par la commande **xttest0** (cf. section 8.4) après **xtreg**…, **re** (modèle à effets aléatoires). Pour tester l'hétéroscédasticité dans un modèle à effets fixes, il faut donc suivre étape par étape la procédure du test de Breush-Pagan :

(1) Régresser le modèle structurel par **xtreg**…, **fe**.
(2) Récupérer le résidu du modèle ci-dessus, puis l'élever au carré.
(3) Régresser le carré du résidu sur l'ensemble des variables explicatives du modèle structurel[28].
(4) La statistique du test est $n*R^2$, qui sous l'hypothèse Ho d'homoscédasticité suit une loi chi2 à *k-1* degré de liberté, *n* et R^2 sont respectivement le nombre d'observations et le coefficient de détermination du modèle de l'étape 3, *k* est le nombre de variables explicatives y compris la constante.

8.8.4. Le test d'autocorrélation des erreurs

Il n'existe pas de commande préprogrammé sur Stata pour faire un test d'autocorrélation de premier ordre AR(1) en panel. Mais cette limite a été comblée par la commande **xtserial** présentée dans le *Stata Journal* (2003), *volume 3*, *numéro 2*. C'est le test d'autocorrélation de Wooldrigde (2002) qui a été programmé sous le nom de **xtserial**. Les lignes de commandes suivantes permettent de télécharger ce package à partir du site Internet de Stata:

```
net from http://www.stata-journal.com/software/sj3-2/
net describe st0039
net install st0039
```

[27] Vous devez créer auparavant par la commande **gen** la variable dépendante prédite au carré, au cube et à la puissance 4.
[28] A cette étape, on peut utiliser la commande **reg** (MCO) à la place de **xtreg**…, **fe** car le résidu est déjà purgé des effets fixes.

Sans connexion Internet, il faut alors être abonné à *Stata Journal* avant d'avoir accès à ce package.

Exemple : **xtserial** *growth tourism lyo prim infl sw*

La ligne de commande ci-dessus teste l'autocorrélation des erreurs du modèle de l'impact du tourisme sur la croissance économique, modèle tel que spécifié dans la section 8.3. L'avantage de cette méthode est que vous n'avez pas à choisir à priori entre un modèle à effets fixes ou un modèle à effets aléatoires. L'hypothèse Ho de ce test est l'absence d'autocorrélation de premier ordre des résidus.

Une seconde façon de faire le test d'autocorrélation est de procéder de manière indirecte à l'aide de la commande **xtregar**. Comme cela a été spécifié plus haut (section 8.4), les commandes **xtregar**…, **fe** et **xtregar**…, **re** permettent d'estimer respectivement un modèle à effets fixes et un modèle à effets aléatoires avec des erreurs autocorrélés d'ordre 1.

xtregar *var_dep var_explicatives* **(if, in)**, **fe lbi**
xtregar *var_dep var_explicatives* **(if, in)**, **re lbi**

L'option **lbi** permet de reporter la statistique du test de Baltagi-Wu et celle du test de Durbin-Watson. Il faut utiliser les tables statistiques correspondantes pour interpréter les résultats des tests, l'hypothèse Ho étant l'absence d'autocorrélation des erreurs.

Une troisième façon de faire un test d'autocorrélation est la commande **pantest2**, mais celle-ci ne marche que pour les modèles à effets fixes, c'est-à-dire après **xtreg**…, **fe**. La commande **pantest2** est disponible sur Internet, il faut donc l'installer avant de l'exécuter. En plus du test d'autocorrélation, la commande **pantest2** donne également le *F test* de significativité des effets fixes et le test de normalité des résidus. Exemple :

xtreg *growth tourism lyo prim infl sw*, **fe**
pantest2 *tps*

tps est la variable de dimension temporelle du panel, il faut la spécifier après la commande **pantest2**.

8.8.5. Le test de Chow

Il se fait de la même façon que telle que présentée étape par étape pour les données transversales (section 7.3.4)

8.8.6. Test d'endogénéité et tests sur les coefficients des variables

Ils sont similaires sur données de panel que sur données transversales

8.8.7. Le test de validité des instruments

Le test de validité des instruments n'est pas préprogrammé sur Stata, il faut donc installer un module externe nommé **overidxt** disponible sur Internet. C'est la version panel de la commande **overid** présentée plus haut (section 7.3.5). De la même façon que la commande **overid** suit une régression avec la commande **ivreg**, la commande **overidxt** suit également une régression en variables instrumentales avec la commande **xtivreg**…, **fe**. Cependant **overidxt** ne marche pas après une régression à l'aide de **xtivreg**…, **re**. Il faut donc procéder au test de Sargan étape par étape comme présenté dans la section 7.3.5.

9. Econométrie des variables qualitatives

Dans les modèles ou la variable expliquée prend la valeur 0 ou 1, l'estimation linéaire n'est pas tout à fait appropriée car les valeurs prédites peuvent être dessous de 0 et au dessus de 1, ou compris entre les deux. De même, la faiblesse de la variance de la variable expliquée peut conduire à des estimations de mauvaise qualité lorsqu'on utilise les MCO.[29]

L'inadéquation du modèle linéaire conduit à modéliser, non pas la variable dépendante elle-même, mais la probabilité qu'elle prenne la valeur 1 ou 0. Pour modéliser cette probabilité, on suppose qu'il existe une variable latente $y*$ telle que : $y=1$ si $y* \geq 0$ et $y=0$ si $y* < 0$. Ensuite on suppose que cette variable $y*$ dépend linéairement d'un certain nombre de variables explicatives X :

$$y* = \beta' X + \varepsilon$$

Le modèle à estimer dépend de l'hypothèse faite sur la distribution du terme d'erreur ε. Les deux lois les plus utilisées sont la loi logistique et la loi normale.

[29] Cependant, il peut être toujours utile de comparer les résultats obtenus en MCO (modèle de probabilité linéaire) à ceux obtenus avec les techniques d'estimations en économétrie qualitatives comme le Logit ou le Probit.

- Lorsqu'on utilise la loi logistique, on parle de modèle Logit donc la fonction de répartition est : $f(x) = \frac{\exp(x)}{1+\exp(x)}$
- Lorsqu'on utilise la loi normale centrée réduite, on parle de modèle Probit dont la fonction de répartition est : $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

Il n'y a pas de test économétrique pour choisir entre les deux modèles. Dans la pratique, les résultats des deux méthodes sont similaires, sauf sur de très grands échantillons.

9.1. Le modèle Logit

La syntaxe pour estimer un modèle Logit sur Stata est la suivante :

logit *var_dep var_explicatives* **(if, in)**, **robust**

Comme la commande **reg** (MCO), la commande **logit** supporte l'option **cluster** et **noconstant**.

Les options standard de **predict** sont :

predict *yhat*, **p** crée une variable *yhat* dont les valeurs seront les probabilités prédites.

predict *yhat2*, **xb** crée une variable *yhat2* dont les valeurs seront les probabilités linéaires prédites équivalentes au modèle régressé en MCO.

predict *resid*, **r** crée une variable *resid* contenant les résidus estimés du modèle.

Exemple : Relation entre niveau de développement financier et crise financière.

logit *crise credit lgdp goveffec rulelaw*, **ro**

La variable dépendante (*crise*) prend la valeur 1 lorsque le pays a connu une crise financière sur la période 1993-1997, et 0 autrement. La variable d'intérêt (*credit*) représentant le niveau de développement financier est le volume des crédits au secteur privé rapporté au PIB. Les variables de contrôles sont le logarithme du PIB par tête (*lgdp*), un indicateur d'efficacité du gouvernement dans la gestion de la politique économique (*goveffec*) et un indicateur de l'état de droit (*rulelaw*). L'échantillon est composé de 78 pays, les variables *credit* et *lgdp* sont

mesurées en moyenne sur la période 1993-1997, les variables *goveffec* et *rulelaw* sont mesurées en 1997.

```
. logit crise lgdp credit goveffec rulelaw, ro

Iteration 0:   log pseudo-likelihood = -49.648105
Iteration 1:   log pseudo-likelihood = -44.381037
Iteration 2:   log pseudo-likelihood = -44.196006
Iteration 3:   log pseudo-likelihood = -44.194197
Iteration 4:   log pseudo-likelihood = -44.194196

Logit estimates                                   Number of obs   =         78
                                                  Wald chi2(4)    =       8.37
                                                  Prob > chi2     =     0.0788
Log pseudo-likelihood = -44.194196                Pseudo R2       =     0.1099

------------------------------------------------------------------------------
             |               Robust
       crise |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lgdp |  -.1316791   .7612493    -0.17   0.863      -1.6237    1.360342
      credit |   .0181833   .0088078     2.06   0.039      .0009204    .0354462
    goveffec |  -1.865142    .800122    -2.33   0.020     -3.433353   -.2969322
     rulelaw |   .4171538   .7922554     0.53   0.599     -1.135638    1.969946
       _cons |  -1.196863   2.431338    -0.49   0.623     -5.962199    3.568472
------------------------------------------------------------------------------
```

Les coefficients tels que présentés ci-dessus ne sont pas des impacts marginaux comme cela a été le cas pour les estimateurs étudiés jusqu'à présent, seuls leurs signes sont interprétables.

L'impact marginal dans un modèle Logit varie d'une observation à une autre (pays dans le cas ci-dessus), il dépend des valeurs des variables explicatives. L'impact marginal d'une variable explicative continue x_i est donné par la formule suivante :

$$\frac{\partial P}{\partial x_i} = \beta_i * \Pr * (1 - \Pr)$$

β_i est le coefficient de la variable x_i donné par Stata avec la commande **logit** et Pr est la probabilité prédite.

Dans l'exemple ci-dessus, puisque l'impact marginal de chaque variable varie d'un pays à un autre, il faut donc pour les besoins d'interprétation des résultats, calculer un impact marginal moyen qui sera celui d'un pays fictif. Il existe deux méthodes pour calculer l'impact marginal moyen d'une variable :

(1) calculer l'impact marginal pour un pays fictif qui a les caractéristiques moyennes de tout l'échantillon.

(2) calculer l'impact marginal en utilisant la moyenne sur tout l'échantillon de l'expression $\Pr * (1 - \Pr)$

Stata dispose de la commande **mfx**[30] (voir section 9.5) qui permet d'obtenir les effets marginaux d'un modèle Logit à l'aide de la première méthode. La commande **dlogit2** disponible sur Internet permet également d'estimer les effets marginaux à l'aide de la première méthode.

```
. dlogit2 crise lgdp credit goveffec rulelaw, ro

Marginal effects from logit                          Number of obs   =       78
                                                     chi2(4)         =    10.15
                                                     Prob > chi2     =   0.0379
Log Likelihood = -44.194196                          Pseudo R2       =   0.1099

------------------------------------------------------------------------------
       crise |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lgdp |  -.0280195   .1617127    -0.17   0.862    -.3449706    .2889316
      credit |   .0038692   .0017665     2.19   0.029     .0004069    .0073315
    goveffec |  -.3968766   .1603363    -2.48   0.013    -.7111299   -.0826233
     rulelaw |   .0887646   .1675018     0.53   0.596    -.2395329    .4170621
       _cons |   -.254676   .5161921    -0.49   0.622    -1.266394    .7570419
------------------------------------------------------------------------------
Marginal effects evaluated at

        lgdp     credit   goveffec    rulelaw      _cons
x    3.14963   33.37745  -.1423035  -.1785425          1
```

Le coefficient de la variable *credit* peut être interprété de la manière suivante : une augmentation du niveau de développement financier de 10% accroît de 3 points de pourcentage la probabilité qu'une crise financière survienne. La commande **dlogit2** donne en dessous du tableau les valeurs moyennes auxquelles les impacts marginaux ont été évalués[31].

La commande **dlogit2** permet également d'obtenir les résultats de l'estimation avec **logit**, il suffit d'y rajouter l'expression **log** comme options.

La seconde méthode de calcul des impacts marginaux moyens n'étant pas disponible sur Stata, il faut donc la programmer. On peut le faire pour une seule variable ou pour toutes les variables simultanément. Le programme pour calculer les impacts marginaux pour toutes les variables étant complexe, seul celui relatif à une seule variable sera présenté ici.

Pour la variable *credit* par exemple :

```
. quietly logit crise lgdp credit goveffec rulelaw, ro
```

[30] L'utilisation de cette commande ne se réduit pas non seulement au calcul des effets marginaux d'un modèle Logit, c'est pour cela qu'une section spécifique lui est consacrée.
[31] Les impacts marginaux peuvent être également évalués pour d'autres valeurs, veuillez consulter l'aide de la commande **dlogit2** pour plus de détails.

```
. predict pr if e(sample),p
(13 missing values generated)

. gen pr2=pr*(1-pr)
(13 missing values generated)

. egen pr2_m=mean(pr2)

. scalar margin_credit=_b[credit]*pr2_m

. display margin_credit
.00350171
```

On remarque que l'impact marginal calculé par la seconde méthode est de 0.00350171, ce qui est très proche de celui calculé par la première méthode : 0.0038692

9.2. Le modèle Probit

La commande **probit** sur Stata permet d'estimer les modèles Probit. Cette commande fonctionne de la même façon que la commande **logit** : la syntaxe est la même et les options sont quasiment identiques. La commande **probit** supporte les options de **predict** précitées pour la commande **logit** à l'exception de la prédiction des résidus.

Pour obtenir les effets marginaux dans un modèle Probit, il faut utiliser la commande **dprobit** préprogrammée sur Stata. Cette commande utilise la méthode de calcul des impacts marginaux aux valeurs moyennes des variables explicatives du modèle.

Exemple du modèle Logit estimé plus haut :

```
. probit crise lgdp credit goveffec rulelaw, ro

Iteration 0:   log pseudo-likelihood = -49.648105
Iteration 1:   log pseudo-likelihood = -44.425425
Iteration 2:   log pseudo-likelihood = -44.320815
Iteration 3:   log pseudo-likelihood = -44.320602

Probit estimates                                  Number of obs   =         78
                                                  Wald chi2(4)    =       8.94
                                                  Prob > chi2     =     0.0626
Log pseudo-likelihood = -44.320602                Pseudo R2       =     0.1073

------------------------------------------------------------------------------
             |               Robust
       crise |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lgdp |  -.0676586   .4436938    -0.15   0.879    -.9372824    .8019652
      credit |   .0105334   .0052631     2.00   0.045     .0002179    .0208489
    goveffec |  -1.084367   .4604022    -2.36   0.019    -1.986739   -.1819953
     rulelaw |   .2328497   .4527732     0.51   0.607    -.6545695    1.120269
       _cons |  -.7301225   1.418899    -0.51   0.607    -3.511114    2.050869
------------------------------------------------------------------------------

. dprobit crise lgdp credit goveffec rulelaw, ro

Iteration 0:   log pseudo-likelihood = -49.648105
```

```
Iteration 1:   log pseudo-likelihood = -44.425425
Iteration 2:   log pseudo-likelihood = -44.320815
Iteration 3:   log pseudo-likelihood = -44.320602

Probit estimates                                  Number of obs =     78
                                                  Wald chi2(4)  =   8.94
                                                  Prob > chi2   = 0.0626
Log pseudo-likelihood = -44.320602                Pseudo R2     = 0.1073

------------------------------------------------------------------------------
             |               Robust
      crise |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
-------------+----------------------------------------------------------------
        lgdp |  -.0240674   .1577395    -0.15   0.879   3.14963  -.333231  .285096
      credit |   .0037469   .0018233     2.00   0.045   33.3774   .000173   .00732
    goveffec |  -.3857296   .1596222    -2.36   0.019  -.142303  -.698583 -.072876
     rulelaw |    .082829   .1606634     0.51   0.607  -.178542  -.232065  .397723
-------------+----------------------------------------------------------------
      obs. P |   .3333333
     pred. P |   .3160018  (at x-bar)
------------------------------------------------------------------------------
    z and P>|z| are the test of the underlying coefficient being 0
```

9.3. Tableau de prédiction (qualité de la prédiction)

Que ce soit après un modèle Logit ou un modèle Probit, on peut faire un tableau de prédiction du modèle pour évaluer sa qualité à prédire les valeurs 0 et 1 de la variable dépendante. On fixe un seuil arbitraire de probabilité et on suppose que si la probabilité prédite est supérieure à ce seuil alors la variable dépendante est égale 1 (événement) et si la probabilité prédite est inférieure à ce seuil, alors la variable dépendante est égale à 0 (non événement). On compare ensuite ces prédictions aux vraies valeurs prises par la variable dépendante. Le seuil souvent utilisé est 0.5. On peut également utiliser comme seuil la moyenne de la variable dépendante. La commande **lstat** permet d'obtenir le tableau de prédiction après une estimation avec la commande **logit** ou **probit**. Exemple :

```
. quietly logit crise lgdp credit goveffec rulelaw, ro

. lstat

Logistic model for crise

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        10             3  |         13
     -     |        16            49  |         65
-----------+--------------------------+-----------
   Total   |        26            52  |         78

Classified + if predicted Pr(D) >= .5
True D defined as crise != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   38.46%
Specificity                     Pr( -|~D)   94.23%
Positive predictive value       Pr( D| +)   76.92%
Negative predictive value       Pr(~D| -)   75.38%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    5.77%
False - rate for true D         Pr( -| D)   61.54%
False + rate for classified +   Pr(~D| +)   23.08%
False - rate for classified -   Pr( D| -)   24.62%
```

```
--------------------------------------------------
Correctly classified                       75.64%
--------------------------------------------------
```

Le seuil utilisé ci-dessous est 0.5, c'est le seuil par défaut[32]. Le tableau de prédiction montre que pour les pays qui ont connu une crise financière (*crise*=1), 10 cas sur 13 ont été bien prédits (probabilité supérieure à 0.5%) et pour les pays n'ayant pas connu de crises (*crise*=0), 49 cas sur 65 ont été bien prédits. Le taux de prédiction du modèle est égale à la somme des cas correctement prédit rapportée au nombre totale d'observations, soit :

$$\frac{10+49}{78} * 100 = 75.64\%$$

La limite principale du tableau de prédiction est qu'une réalisation de la variable dépendante ayant une probabilité de 0.49 représente la même chose qu'avec une probabilité de 0.001.

9.4. La commande *mfx* pour calculer les impacts marginaux et les élasticités

La commande **mfx** permet de calculer les effets marginaux et les élasticités après une estimation. Elle s'applique à l'estimation la plus récente et fonctionne pour la plupart des techniques économétriques d'estimation, en l'occurrence les modèles Logit et Probit.

La syntaxe est la suivante : **mfx compute (if, in)**, *options*

Pour les options de base, on distingue[33] :

dydx l'effet marginal $\frac{\partial y}{\partial x}$ est calculé (option par défaut)

eyex l'élasticité $\frac{\partial \ln(y)}{\partial \ln(x)}$ est calculée

dyex la semi-élasticité $\frac{\partial y}{\partial \ln(x)}$ est calculée

eydx la semi-élasticité $\frac{\partial \ln(y)}{\partial x}$ est calculée

at(…**)** spécifie les valeurs auxquelles les impacts marginaux ou les élasticités doivent être calculées. Par défaut, les valeurs moyennes des variables

[32] On peut utiliser aussi comme seuil la moyenne de la variable dépendante. Soit *m* cette valeur, le tableau des prédictions s'obtient par la commande suivante : **lstat**, **cutoff**(*m*)
[33] Consulter l'aide de la commande **mfx** pour les options avancées

explicatives sont utilisées. Exemples : *(a)* **at(mean)** calcule les impacts marginaux à la valeur moyenne des variables explicatives. *(b)* **at(median)** calcule les impacts marginaux à la valeur médiane des variables explicatives. *(c)* **at(zero)** calcule les impacts marginaux à la valeur nulle des variables explicatives. *(d)* **at(mean** *x=a y=b***)** les impacts marginaux sont calculés pour les valeurs *a* de *x*, *b* de *y* et pour les valeurs moyennes des autres variables explicatives.

Exemple du modèle Logit de la section 9.1 :

```
. quietly logit crise lgdp credit goveffec rulelaw, ro

. mfx compute

Marginal effects after logit
      y  = Pr(crise) (predict)
         =  .30709123
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]       X
---------+--------------------------------------------------------------------
    lgdp |  -.0280195      .16171   -0.17   0.862  -.344971  .288932   3.14963
  credit |   .0038692      .00177    2.19   0.029   .000407  .007331   33.3774
goveffec |  -.3968766      .16034   -2.48   0.013   -.71113 -.082623  -.142303
 rulelaw |   .0887646       .1675    0.53   0.596  -.239533  .417062  -.178542
------------------------------------------------------------------------------
```

Le tableau produit par la commande **mfx** est rigoureusement identique à celle de la commande **dlogit2** plus haut (section 9.1). La dernière colonne du tableau ci-dessus donne les valeurs moyennes des variables explicatives, valeurs auxquelles sont calculés les impacts marginaux.

9.5. Quelques autres modèles Logit et Probit et leurs commandes Stata

Logit en panel avec effets fixes	**xtlogit**…, **fe**
Logit en panel avec effets aléatoires	**xtlogit**…, **re**
Logit conditionnel	**clogit**
Logit multinomial	**mlogit**
Logit ordonnée	**ologit**
Probit bivarié	**biprobit**
Probit en panel avec effets aléatoires	**xtprobit**…, **re**
Probit ordonnée	**oprobit**

10. Introduction aux séries temporelles

10.1. Création de variables retardées et de variables en différences

Avant d'utiliser les commandes de séries temporelles, il faut impérativement préciser la dimension temporelle par la syntaxe suivante :

tsset *tps* *tps* est le nom de la variable de dimension temporelle

Pour la création de variables, on utilise la commande **generate** (**gen** en abrégé)

gen *y* = **L.***x* crée une variable *y* égale à la valeur retardée d'une période de la variable *x*.

gen *y* = **L2.***x* crée une variable *y* égale à la valeur retardée de deux périodes de la variable *x*

gen *y* = **F.***x* crée une variable *y* égale à x_{t+1}

gen *y* = **F2.***x* crée une variable y égale à x_{t+2}

gen *y* = **D.***x* crée une variable *y* égale à la première différence de la variable *x*, soit $(x_t - x_{t-1})$

gen *y* = **D2.***x* crée une variable *y* égale à la différence de la différence de la variable *x*, soit $\left[(x_t - x_{t-1}) - (x_{t-1} - x_{t-2})\right]$

gen *y* = **S.***x* crée une variable *y* égale à la première différence de la variable *x*, soit $(x_t - x_{t-1})$

gen *y* = **S2.***x* crée une variable *y* égale à la différence seconde de la variable *x*, soit $(x_t - x_{t-2})$

gen *y* = **Sn.***x* crée une variable *y* égale à la différence $n^{ième}$ de la variable *x*, soit $(x_t - x_{t-n})$

Remarque 1 : les commandes ci-dessus fonctionnent également sur les données de panel.

Remarque 2 : plusieurs commandes de régressions ou de statistiques descriptives tolèrent l'utilisation des variables retardées ou en différence dans la liste des variables explicatives.

Exemple : **regress** *z x* **L.***x*, **ro** la variable dépendante *z* est régressée sur la variable *x* et sa valeur retardée d'une période.

sum L.*x* statistiques descriptives de la variable retardée de *x*.

10.2. Test de stationnarité sur données temporelles et sur données de panel

10.2.1. Test de stationnarité sur données temporelles

Le test de Dickey-Fuller augmenté

dfuller *variable* **(if, in)**, **noconstant lags(**#**) trend regress**

L'option **noconstant** exclut la constante de la régression. Pour l'option **lags(**#**)**, # spécifie le nombre de variables retardées en différences à inclure comme variables explicatives dans la régression associée au test de Dickey-Fuller. L'option **trend** inclut une tendance dans la régression et l'option **regress** affiche les résultats de la régression. L'hypothèse Ho du test est la présence d'une racine unitaire, donc une probabilité du test inférieure à 10% (ou une statistique calculée inférieure à la statistique lue à 10%) conduit au rejet de Ho.

Le test de Phillips-Perron

pperron *variable* **(if, in)**, **noconstant lags(**#**) trend regress**

Toutes les options de la commande **pperron** ont les mêmes significations que dans le cas de la commande **dfuller**, à l'exception de l'option **lags(**#**)** où # spécifie le nombre de retard de Newey-West à utiliser pour calculer la variance. L'hypothèse Ho du test de Phillips-Perron est la présence d'une racine unitaire, donc une probabilité du test inférieure à 10% (ou une statistique calculée inférieure à la statistique lue à 10%) conduit au rejet de Ho.

10.2.2. Test de stationnarité sur données de panel

Le test de Im-Pesaran-Shin et celui de Levin-Lin-Chu seront présentés dans cette section. Ces tests ne sont pas préprogrammés sur Stata, il convient donc de télécharger leurs modules à partir d'Internet. Pour pouvoir effectuer ces tests, il faut impérativement un panel cylindré (c'est-à-dire sans données manquantes).

Le test de Im-Pesaran-Shin (IPS)

ipshin *variable* **(if, in)**, **lags(**#**) trend nodemean**

L'option **lags(**#**)** spécifie le nombre de retards à inclure dans le modèle, l'option **trend** spécifie l'inclusion d'une tendance et l'option **nodemean** inclut des variables muettes temporelles dans le modèle. L'hypothèse Ho du test IPS est : toutes les séries sont non stationnaires contre l'hypothèse alternative : seule une fraction des séries individuelles est stationnaire. Une probabilité du test inférieure à 10% conduit au rejet de Ho.

Le test de Levin-Lin-Chu (LLC)

levinlin *variable* **(if, in)**, **lags(**#**) noconstant trend**

Pour les options, les définitions sont les mêmes que pour la commande **ipshn**. L'option **noconstant** supprime la constante du modèle. L'hypothèse Ho du test LLC est : toutes les séries sont non stationnaires contre l'hypothèse alternative : toutes les séries sont stationnaires. Une probabilité du test inférieure à 10% conduit au rejet de Ho.

10.2.3. Test d'autocorrélation d'erreurs en série temporelles

Test de Breush-Godfrey	commande **bgodfrey**
Test de Durbin-Watson	commande **durbina**

Les commandes ci-dessus fonctionnent sous le même principe que la commande **ovtest** après **regress** (section 7.3.2), mais uniquement en séries temporelles.

11. Exporter les tableaux des régressions

La commande **outreg** qui est disponible sur Internet permet d'exporter les résultats des régressions dans des tableaux. En voici un exemple :

```
reg growth tourism lyo, robust
outreg using output.doc, nolabel replace 3aster

reg growth tourism lyo prim , robust
```

```
outreg using output.doc, nolabel append 3aster

reg growth tourism lyo infl , robust
outreg using output.doc, nolabel append 3aster

reg growth tourism lyo sw, robust
outreg using output.doc, nolabel append 3aster

reg growth tourism lyo prim  infl sw, robust
outreg using output.doc, nolabel append 3aster
```

Tous les résultats des régressions en MCO ci-dessus seront inclus dans un tableau unique (en format texte) dans le fichier Word nommé output dont l'aspect est le suivant :

```
      (1)    (2)    (3)    (4)    (5)
      growth       growth       growth       growth       growth
tourism     2.791 2.862 2.609 2.655 2.534
      (5.68)***    (5.67)***    (5.76)***    (4.99)***    (4.81)***
lyo   -0.510       -0.765       -0.393       -1.024       -1.261
      (0.89)       (1.26)       (0.70)       (1.45)       (1.79)*
prim        0.380              0.266
            (0.32)                    (0.21)
infl              -0.148             -0.141
                  (2.47)**           (2.49)**
sw                      1.039 1.285
                        (1.33)       (1.71)*
Constant    -13.161      -13.197      -12.238      -11.459      -10.267
      (5.02)***    (4.99)***    (5.64)***    (4.27)***    (4.34)***
Observations       61     60     61     59     58
R-squared   0.45  0.46  0.54  0.47  0.57
Robust t statistics in parentheses
* significant at 10%; ** significant at 5%; *** significant at 1%
```

Notons que ce fichier est souvent créé dans le répertoire C:\Stata8[34]. Il faut donc l'ouvrir, ensuite convertir le texte en tableau dans Word en utilisant le menu *Tableau/Convertir/Texte en tableau*. Enfin, vous pouvez mettre en forme les bordures du tableau par le menu *Format/Bordure et trame* et éventuellement modifier le nom des variables, de mêmes que les notes en bas du tableau. Les valeurs absolues des *t* de *student* sont entre parenthèses.

[34] Ce répertoire peut changer d'une version à une autre de Stata, une recherche rapide sur le disque dur permet de retrouver le fichier.

	(1)	(2)	(3)	(4)	(5)
	growth	growth	growth	growth	growth
tourism	2.791	2.862	2.609	2.655	2.534
	(5.68)***	(5.67)***	(5.76)***	(4.99)***	(4.81)***
Lyo	-0.510	-0.765	-0.393	-1.024	-1.261
	(0.89)	(1.26)	(0.70)	(1.45)	(1.79)*
Prim		0.380			0.266
		(0.32)			(0.21)
Infl			-0.148		-0.141
			(2.47)**		(2.49)**
Sw				1.039	1.285
				(1.33)	(1.71)*
Constant	-13.161	-13.197	-12.238	-11.459	-10.267
	(5.02)***	(4.99)***	(5.64)***	(4.27)***	(4.34)***
Observations	61	60	61	59	58
R-squared	0.45	0.46	0.54	0.47	0.57

Robust t statistics in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

Revenons sur les options de la commande **outreg**. L'option **nolabel** évite que les noms des variables soient remplacés par leurs étiquettes (pour les étiquettes des variables, voir section 2.4). L'option **3aster** permet de mettre les étoiles de significativité des coefficients des variables explicatives. L'option **replace** se met uniquement pour la première régression (c'est à dire la première colonne du tableau), elle permet de remplacer le fichier *output.doc* existant si le fichier *do* a été déjà exécutée une fois. Pour les régressions autres que la première, il faut impérativement mettre l'option **append** pour **outreg** car c'est cette option qui permet d'ajouter les autres régressions à la première pour former le tableau.

Remarque 1 : la commande **outreg** peut s'appliquer après toutes les commandes d'estimations en Stata.

Remarque 2 : la commande **outreg** dispose d'une foule d'autres options que je vous invite à explorer avec l'aide de Stata (en tapant **help outreg** dans la fenêtre de commande de Stata). Par exemple, on peut remplacer le *t* de *student* par l'écart-type des coefficients. On peut également rajouter d'autres statistiques que le nombre d'observations et le R^2.

Remarque 3 : on peut changer le format du fichier dans lequel seront stockés les résultats. Par exemple on peut mettre *output.xls* à la place de *output.doc*, l'inconvénient est qu'Excel met automatiquement tous les nombres négatifs (à l'exemple des coefficients) entre parenthèses.

Pour ceux qui utilisent LaTeX pour le traitement de texte, je vous invite a consulter le manuel de Florent Bresson[35] qui recense un éventail assez large de modules de Stata qui permettent d'exporter les tableaux de Stata en version *.tex*.

12. Ajout de nouveaux modules à Stata

On a vu tout au long de ce manuel que plusieurs modules ne sont pas disponibles d'origine sur Stata et qu'il faut les télécharger à partir d'Internet. Bon nombre des modules externes de Stata sont disponibles sur le site Internet de Boston College[36], on peut les installer manuellement ou faire une installation automatique.

ssc install *nom_du_programme*, **all**

La syntaxe ci-dessus, lorsqu'elle est saisie dans la fenêtre de commandes de Stata, permet l'installation automatique des programmes à partir du site Internet de Boston College[37] (voir l'aide de la commande **ssc** pour plus de détails).

Exemples : **ssc install overid**, **all**
ssc install xtabond2, **all**

Pour ceux qui veulent en savoir plus sur l'installation automatique de nouveaux modules en général, je vous suggère de consulter l'aide de Stata sur la commande **net**.

Dans le cas de l'installation manuelle, il faut télécharger tous les fichiers indispensables au fonctionnement du module. Souvent, mais pas dans tous les cas, il s'agit d'un fichier *.ado* (fichier programme) et d'un fichier *.hlp* (fichier d'aide). Les fichiers doivent être enregistrés

[35] Disponible sur le réseau du CERDI pour les doctorants ou sur demande à florent.bresson@u-clermont1.fr
[36] http://ideas.repec.org/s/boc/bocode.html
[37] Bien évidement vous devez être connecté à Internet avant de lancer la commande **ssc**.

dans un dossier bien déterminé du répertoire C:\ado\plus ou C:\ado\personal[38] (répertoires à créer s'ils n'existent pas). Dans ce répertoire, il faut créer (s'il n'existe pas déjà) un dossier dont le nom sera la première lettre du nom de la commande, et ensuite y enregistrer les fichiers téléchargés. Par exemple, le module de la commande **overid** doit être enregistré dans le répertoire C:\ado\personal\o, celui de la commande **xtabond2** doit être enregistré dans le répertoire C:\ado\personal\x. Notons que dans certains cas le fichier *.ado* à télécharger apparaît plutôt sous forme de page Web, dans ce cas il faut l'enregistrer en choisissant le format *Page Web HTML uniquement (*.htm,*.html)*, puis ensuite renommer ce fichier en *.ado* en supprimant l'extension *.html*. Par exemple, l'enregistrement sous format *.html* du fichier programme de la commande **overid** donnera un fichier nommé par défaut *overid_ado.html*, il faut donc renommer ce fichier en *overid.ado* avant de l'installer dans le répertoire C:\ado\personal\o.

[38] Merci à Christopher F Baum pour avoir soulevé ce point.

Références

Arellano M. et S. Bond (1991) "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations", *Review of Economic Studies*, vol. 58, p. 277-297.

Arellano M. et O. Bover (1995) "Another Look at the Instrumental-Variable Estimation of Error-Components Models", *Journal of Econometrics*, vol. 68, n°1, p. 29-52.

Blundell R. et S. Bond (1998) "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models", *Journal of Econometrics*, vol. 87, n°1, p.115-143.

Bresson Florent (2004) *Présentation de commandes Stata pour LATEX*

Drukker D. (2003) "Testing for Serial Correlation in Linear Panel-Data Models", *Stata Journal*, vol. 3, n°2, p.168-177.

Sevestre P. (2002) *Econométrie des données de panel*, Paris, Dunod.

StataCorp. 2003. Stata Statistical Software: Release 8.0. College Station, TX: Stata Corporation

Wooldridge J. (2002) *Econometric Analysis of Cross Section and Panel Data*, Cambridge, The MIT Press.

ANNEXE : La Méthode des Moments Généralisés en Panel Dynamique

Considérons l'équation suivante :

$$y_{i,t} - y_{i,t-1} = (\alpha - 1)y_{i,t-1} + \beta' X_{i,t} + u_i + v_t + e_{it} \qquad (1)$$

Où y_{it} représente le logarithme du PIB par tête réel, X représente les variables explicatives du modèle, *u* l'effet spécifique pays, *v* l'effet spécifique temporel et *e* le terme d'erreur, *i* est l'indice pays, et *t* l'indice temporel.

L'équation (1) qui est équivalente à une équation de croissance, peut être réécrite de la façon suivante :

$$y_{i,t} = \alpha y_{i,t-1} + \beta' X_{i,t} + u_i + v_t + e_{it} \qquad (2)$$

Dans ce modèle, la présence de la variable dépendante retardée ne permet pas d'utiliser les techniques économétriques standard. On utilise la Méthode des Moments Généralisés en panel dynamique qui permet de contrôler pour les effets spécifiques individuels et temporels, et de pallier les biais d'endogénéité des variables. Il existe deux types d'estimateur : (a) l'estimateur d'Arellano et Bond (1991) ou GMM en différences et (b) l'estimateur des GMM en système. Notons que l'utilisation de ces deux estimateurs présuppose la quasi-stationnarité des variables de l'équation en niveau, et l'absence d'autocorrélation des résidus.

Dans l'estimateur d'Arellano et Bond (1991), la stratégie pour répondre à un éventuel biais de variable omise liés aux effets spécifiques est de différencier l'équation (2) en niveau. On obtient l'équation :

$$y_{i,t} - y_{i,t-1} = \alpha(y_{i,t-1} - y_{i,t-2}) + \beta'(X_{i,t} - X_{i,t-1}) + (v_t - v_{t-1}) + (e_{it} - e_{i,t-1}) \qquad (3)$$

La différence première élimine l'effet spécifique pays et par conséquent le biais de variables omises invariantes dans le temps. Par construction le terme d'erreur $(e_{it} - e_{i,t-1})$ est corrélé avec la variable retardée en différence $(y_{i,t-1} - y_{i,t-2})$. Les différences premières des variables explicatives du modèle sont instrumentés par les valeurs retardées (en niveau) de ces mêmes variables. Le but est de réduire les biais de simultanéité et le biais introduit par la présence de la variable dépendante retardée en différence dans le membre de gauche.

Sous l'hypothèse que les variables explicatives du modèle sont faiblement exogènes (elles peuvent être influencées par les valeurs passées du taux de croissance, mais restent non corrélées aux réalisations futures du terme d'erreur) et que les termes d'erreur ne soient pas

autocorrélés, les conditions de moments suivantes s'appliquent pour l'équation en première différence.

$$E\left[y_{i,t-s} \cdot \left(e_{i,t} - e_{i,t-1}\right)\right] = 0 \text{ pour } s \geq 2; t = 3,...,T \quad (4)$$

$$E\left[X_{i,t-s} \cdot \left(e_{i,t} - e_{i,t-1}\right)\right] = 0 \text{ pour } s \geq 2; t = 3,...,T \quad (5)$$

Le problème avec cet estimateur est qu'il souffre de la faiblesse des instruments, qui entraînent des biais considérables dans les échantillons finis, et sa précision est asymptotiquement faible. Plus précisément, les valeurs retardées des variables explicatives sont des faibles instruments de l'équation en différence première. Par ailleurs, la différentiation de l'équation en niveau élimine les variations inter-pays et ne prend en compte que les variations intra-pays.

L'estimateur GMM en système permet de lever cette limite. Il combine l'équation en différence avec celle en niveau. L'équation en différence première (Equation 3) est estimée simultanément avec l'équation en niveau (Equation 2) par les GMM. Dans l'équation en niveau, les variables sont instrumentés par leurs différences premières[39]. Blundell et Bond (1998) ont testé cette méthode à l'aide des simulations de Monte Carlo. Ces auteurs ont trouvé que l'estimateur GMM en système est plus efficient que l'estimateur des GMM en différences. Ce dernier produit des coefficients biaisés pour les petits échantillons. Le biais est d'autant plus important que les variables sont persistantes dans le temps, que les effets spécifiques sont importants et que la dimension temporelle du panel est faible.

Pour l'équation en niveau, on utilise des conditions additionnelles de moments en supposant que les variables explicatives sont stationnaires.

$$E\left[\left(y_{i,t-s} - y_{i,t-s-1}\right) \cdot \left(u_i + e_{i,t}\right)\right] = 0 \text{ pour } s = 1 \quad (6)$$

$$E\left[\left(X_{i,t-s} - X_{i,t-s-1}\right) \cdot \left(u_i + e_{i,t}\right)\right] = 0 \text{ pour } s = 1 \quad (7)$$

Les conditions de moments ci-dessus (4 à 7) combinées avec la Méthode des Moments Généralisées permettent d'estimer les coefficients du modèle. Pour tester la validité des variables retardées comme instruments, Arellano et Bond (1991), Arellano et Bover (1995), et Blundel et Bond (1998) suggèrent le test de suridentification de Sargan/Hansen. Par construction le terme d'erreur en différence première est corrélé au premier ordre, mais il ne doit pas l'être au second ordre. Pour tester cette hypothèse, ces mêmes auteurs suggèrent un test d'autocorrélation de second ordre.

[39] Seule la différence première la plus récente est utilisée, l'utilisation d'autres différences premières retardées entraînerait une redondance des conditions de moments (Arellano et Bover, 1995)