SEMINAIRE D'ECONOMETRIE

Rafik Bouklia-Hassane

Janvier 2020

PLAN DU SEMINAIRE

Leçon 1: Le modèle linéaire

Leçon 2: Introduction à Stata

Leçon 3: Econométrie des variables qualitatives

Leçon 4: Séries temporelles et méthodologie de Box et Jenkins

Leçon 5: Modélisation VAR et modèles à correction d'erreurs

Leçon 6: Econométrie des panels

Leçon 7: Introduction aux panels dynamiques

BIBLIOGRAPHIE

Bourbonnais Régis: Econométrie, Dunod 2015

Meddahi Nour: Statistiques pour économistes – ISGP – Alger 2015

Woolridge Jeffrey M.: Introductory Econometrics: A Modern Approach

Leçon 1: La démarche économétrique et le modèle linéaire

1- SPECIFICATION DU MODELE

L'objet est d'expliquer une variable y_t traduisant un phénomène économique (variable dépendante) à l'aide d'autres variables économiques $\boldsymbol{x_t}$ (variables explicatives). L'étape de spécification d'un modèle consiste à (i) identifier les variables explicatives et (ii) à se donner une forme fonctionnelle reliant les variables explicatives à la variable dépendante:

$$y_t = f(x_t, u_t)$$

Les termes u_t sont les termes d'erreur. Ils sont inconnus mais certaines hypothèses sur ces termes vont être faites par la suite.

Différentes formes fonctionnelles peuvent être retenues:

Modèle	Forme fonctionnelle	Interprétation
Linéaire	$y = b_1 + b_2 x + u$	La variation y est proportionnelle à x
Log-linéaire	$y = b_1 x^{b_2} u$	Le taux de variation de y est proportionnel au taux de variation de X (b_2 est une élasticité)
Exponentiel	$y = \exp(b_1 + b_2 x + u)$	Le taux de variation de y est proportionnel à la variation de X (b_2 est une semi-élasticité)
Logarithmique	$y = b_1 + b_2 Log x + u$	La variation de Y est proportionnelle au taux de variation de X

Dans le cas d'une spécification linéaire, le modèle peut s'écrire sous forme matricielle:

$$\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{u} \qquad (1)$$

2- CHOIX DU TYPE DE DONNEES

Les données utilisées peuvent être de plusieurs types:

1) *Données structurées en séries temporelles (time serie).* Dans ce cas, les données portent sur un individu (une entreprise, un pays,…) observé à plusieurs dates.

2) *Données en coupe transversale (cross section).* Dans ce cas, les données concernent plusieurs individus (ensemble d'entreprises observées à un instant donné, ensemble de pays, ensemble des localités d'un pays, etc) observés à date fixée.

3) *Données en panel*. Dans ce cas, les données concernent plusieurs individus observés à plusieurs dates.

CHOIX DU TYPE DE DONNEES

Deux structurations en panel des données sur le nombre d'homicides par 100,000 habitants pour quatre pays :

Structure longue

countryname	year	Nbr_homicides
Afghanistan	2010	3.4
Afghanistan	2011	4.1
Afghanistan	2012	6.3
Afghanistan	2013	NA
Afghanistan	2014	NA
Afghanistan	2015	10
Algeria	2010	0.7
Algeria	2011	0.8
Algeria	2012	1.4
Algeria	2013	1.3
Algeria	2014	1.5
Algeria	2015	1.4
France	2010	1.3
France	2011	1.4
France	2012	1.2
France	2013	1.2
France	2014	1.2
France	2015	1.6
United States	2010	4.8
United States	2011	4.7
United States	2012	4.7
United States	2013	4.5
United States	2014	4.5
United States	2015	5

Structure large

countryname	Homicide 2010	homicide 2011	homicide 2012	homicide 2013	homicide 2014	homicide 2015
Afghanistan	3.4	4.1	6.3	NA	NA	10
Algeria	0.7	0.8	1.4	1.3	1.5	1.4
France	1.3	1.4	1.2	1.2	1.2	1.6
United States	4.8	4.7	4.7	4.5	4.5	5

3- ESTIMATION DU MODELE

Cette étape consiste à estimer les coefficients intervenant dans la relation entre y et les variables x, soit, dans le cas linéaire

$$y_t = b_1 x_{1t} + b_2 x_{2t} + \cdots + b_K x_{Kt} + u_t \qquad (1')$$

Les coefficients $\boldsymbol{\beta} = (b_1, b_2, \ldots, b_K)'$ sont fixes mais inconnues. Il s'agit alors d'en trouver des estimateurs à partir d'un échantillon de N observations de y et des K variables explicatives:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \text{ et } \boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{K1} \\ \vdots & \ddots & \vdots \\ x_{1N} & \cdots & x_{KN} \end{bmatrix}$$

où la matrice $\boldsymbol{y}$ est (N, 1), et $\boldsymbol{X}$ est (N,K),

Si on se base uniquement sur la relation (1), alors la solution est indéterminée. Pour cette raison, pour déterminer les estimateurs des coefficients $\boldsymbol{\beta}$, on se donne un critère d'optimisation. Le plus souvent, ce critère est la minimisation de l'erreur moyenne quadratique (EMQ). Celle-ci est donnée par:

$$EMQ(\beta) = \sum_{t=1}^{T} [y_t - (b_1 x_{1t} + b_2 x_{2t} + \cdots + b_K x_{Kt})]^2$$

Noter que la parenthèse du membre de droite représente les valeurs de y_t prédites par le modèle (1')

4- VALIDATION DU MODELE

Cette étape consiste à s'interroger pour savoir si le modèle spécifié 's'adapte' aux observations empiriques. Pour cela, on met en œuvre des tests qui permettent ou non de valider empiriquement le modèle spécifié, c'est-à-dire de valider ou non la théorie économique qui est représentée par ce modèle. Parmi ces tests :

1- *tests de significativité des coefficients estimés*. Si dans la relation (1') un des coefficient b_i est statistiquement égal à 0, alors on peut considérer que la variable x_i qui lui est associée n'a pas d'impact sur la variable dépendante. Il y a lieu alors de 're-spécifier' le modèle.

2- *tests de bruit blanc*. Ce test consiste à vérifier que les résidus d'estimation forment un bruit blanc. En séries temporelles, un bruit blanc est un processus ε_t vérifiant

a) $E(u_t) = 0$ (processus est centré) ;

b) $V(u_t) = \sigma^2$ (la variance est indépendante du temps et

c) $Cov(u_t, u_{t+k}) = 0$ pour $k \neq 0$ (l'autocovariance est nulle sauf à l'origine quand $k = 0$ auquel cas elle est égale à la variance σ^2).

Le processus est *iid* (indépendant et identiquement distribué) si la condition c) est remplacée par la condition c'): les résidus d'estimation sont de même distribution et mutuellement indépendants,

Si $\hat{y}_t$ est la valeur estimée de y_t, alors le résidu d'estimation est $\hat{u}_t = y_t - \hat{y}_t$.

Un test de bruit blanc revient à tester que les résidus d'estimation $\hat{u}_t$ suivent un bruit blanc,

VALIDATION DU MODELE

Des tests de robustesse peuvent également être effectués comme:

- La robustesse par rapport à la période d'estimation: le modèle reste-t-il valide si la période d'estimation est modifiée ?
- La robustesse par rapport au choix des variables proxy retenues
- Les coefficients estimés sont-ils stables dans le temps?
- Etc.

5- L'ANALYSE POST ESTIMATION

Des investigations peuvent être réalisées à partir des résultats d'estimation du modèle comme:

- Les simulations de chocs: si telle variable explicative augmente, quel sera son impact sur la variable expliquée (discuter le signe de l'impact et également son intensité);
- La prévisions de la variable dépendante étant donné les évolutions attendues des variables explicatives.

Cette analyse permet d'aider les décideurs (pouvoirs publics, chefs d'entreprise,…) à rationnaliser leurs actions en anticipant les conséquences de leurs politiques publiques pour les premiers et l'impact des changements de l'environnement économique par exemple pour les seconds

1- LE MODELE LINEAIRE DE REGRESSION: L'ESTIMATEUR MCO

Notations:

y_t: variable dépendante au temps t (t sera l'identifiant de l'individu si on est en coupe transversale);

x_{kt}: variable explicative k à la date t; $x_k = (x_{k1}, x_{k2}, \dots, x_{kN})'$

b_k:coefficient de la variable explicative x_k ;

u_t: terme d'erreur à la date t.

K: nombre de variables explicatives et N nombre de périodes (d'individus en cross section)

X: matrice de format (N, K) des variables explicatives)

$\boldsymbol{\beta}$: vecteur colonne à K composantes des coefficients b_k

y: vecteur colonne à N composantes représentant la variable dépendante $(y_1, y_2, \dots, y_N)'$

LE MODELE LINEAIRE DE REGRESSION: L'ESTIMATEUR MCO

On se donne donc le modèle linéaire suivant:

$$y_t = b_1 x_{1t} + b_2 x_{2t} + \cdots + b_K x_{Kt} + u_t \text{ avec t=1,…N}$$

Sous forme matricielle, le modèle se réécrit comme dans (1):

$$\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{u}$$

Dans le modèle avec constante, le plus souvent utilisé, $x_{1t} = 1 \; \forall t = 1, \ldots, N$ de sorte que:

$$y_t = b_1 + b_2 x_{2t} + \cdots + b_K x_{Kt} + u_t \quad \text{avec } t = 1, \ldots N$$

Sous forme matricielle, cela signifie que la première colonne de X est composée uniquement que de 1. En général, on n'a pas besoin d'écrire cette variable constante. Elle est automatiquement générée par le logiciel à la demande.

LE MODELE LINEAIRE DE REGRESSION: L'ESTIMATEUR MCO

Critère de détermination de l'estimateur MCO

Soit $\widehat{\boldsymbol{\beta}}$ une approximation des coefficients inconnus β.L'erreur moyenne quadratique qu'on commet du fait de cette approximation est

$$EMQ = \frac{1}{N}\sum_{t=1}^{N}\left[y_t - \left(\hat{\beta}_1 x_{1t} + \hat{\beta}_2 x_{2t} + \cdots + \hat{\beta}_K x_{Kt}\right)\right]^2$$

EMQ peut également s'écrire matriciellement:

$$EMQ = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$$

En effet, étant donné un vecteur $V = \begin{pmatrix} v_1 \\ \vdots \\ v_N \end{pmatrix}$, alors $V'V = {v_1}^2 + {v_2}^2 + \cdots + {v_N}^2$ (voir document sur le calcul matriciel).

Par définition, l'estimateur MCO $\widehat{\boldsymbol{\beta}}_{\boldsymbol{MCO}}$ est celui qui minimise l'erreur moyenne quadratique EMQ.

$\widehat{\boldsymbol{\beta}}_{\boldsymbol{MCO}}$ est donc solution du problème:

$$\underbrace{Min}_{\beta}\ (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) \qquad \textbf{(P)}$$

LE MODELE LINEAIRE DE REGRESSION: L'ESTIMATEUR MCO

Expression de l'estimateur $\widehat{\boldsymbol{\beta}}_{\boldsymbol{MCO}}$

Le problème (P) se réécrit:

$$\underbrace{Min}_{\beta} \boldsymbol{y'y + 2\beta' X' y + \beta' X' X\beta}$$

(à démontrer comme exercice en utilisant les propriétés de la transposition des matrices. Voir document remis)

En dérivant matriciellement cette expression par rapport à β et en égalant cette dérivée à zéro, il vient:

$$\boldsymbol{-2X'y + 2X'X\beta} = 0 \qquad (2)$$

(A démontrer comme exercice en utilisant les propriétés sur la dérivation matricielle. Pour rappel :

Si $b'Ab$ est une forme quadratique (A est une matrice carrée symétrique et b un vecteur colonne), alors:

$$\frac{\partial b'Ab}{\partial b} = 2Ab$$

LE MODELE LINEAIRE DE REGRESSION: L'ESTIMATEUR MCO

En égalisant cette expression à 0, il vient alors :

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{MCO}} = (\boldsymbol{X'X})^{-1}(\boldsymbol{X'y})$$

Remarque 1: pour que la solution ci-dessus existe, il est nécessaire de faire l'hypothèse que la matrice $(\boldsymbol{X'X})$ est inversible. Cela se produit si les variables x_k ne sont pas parfaitement colinéaires, c'est-à-dire si:

la matrice **X** est de rang K (plein rang colonne) (H1)

Pourquoi faire cette hypothèse? Supposons à l'inverse que les variables $\boldsymbol{x_1}, \boldsymbol{x_2}$ *et* $\boldsymbol{x_3}$ par exemple soient colinéaires telles que $\boldsymbol{x_3} = \boldsymbol{x_1} + \boldsymbol{x_2}$. Cette relation montre que si on connaît $\boldsymbol{x_1}$ et $\boldsymbol{x_2}$ alors on peut connaître parfaitement x_3. Par conséquent, la variable $\boldsymbol{x_3}$ n'apporte aucune information supplémentaire. Elle est redondante et le modèle est mal spécifié.

Cette hypothèse *[(X'X) est inversible]* est connue sous le nom d'absence de multicolinérarité entre les variables explicatives.

LE MODELE LINEAIRE DE REGRESSION: L'ESTIMATEUR MCO

Remarque 2 : la condition d'optimisation (2) est nécessaire mais non suffisante. Il faut s'assurer que celle-ci définit bien un minimum. Celle condition de deuxième ordre est assurée car la dérivée seconde, obtenue en dérivant de nouveau le membre de gauche de (2) par rapport à β, est $2X'X$. On montre que cette matrice est définie positive (A démontrer en utilisant les propriétés des formes quadratiques: voir document de rappel de mathématiques remis). Par conséquent, on est bien en présence d'un minimum.

Remarque 3: On peut exprimer $\widehat{\boldsymbol{\beta}}_{\boldsymbol{MCO}}$ en fonction de *u:*

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{MCO}} = (\boldsymbol{X'X})^{-1}(\boldsymbol{X'y}) = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}(\boldsymbol{X\beta} + \boldsymbol{u}) = (\boldsymbol{X'X})^{-1}(\boldsymbol{X'X})\boldsymbol{\beta} + (\boldsymbol{X'X})^{-1}\boldsymbol{X'u}.$$

D'où la relation (qui sera utilisée par la suite) :

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X'X})^{-1}\boldsymbol{X'u} \qquad (3)$$

2- TP1 SUR STATA: REGRESSION LINEAIRE

Exercice: Etude des déterminants du phénomène de criminalité dans le monde.

On se propose d'étudier le phénomène de criminalité, plus précisément d'analyser les facteurs qui sont a l'origine du nombre d'homicides observés dans les différents pays à travers le monde. Pour cela, on construit puis on teste un modéle qui expliquerait ces homicides par

- des facteurs démographiques (taille de la population, pourcentage de jeunes dans la population, urbanisation),

- et par des facteurs économiques (niveau plus ou moins élevé de développement des pays en question)

Que peut-on conclure?

[*Remarque*: on assimilera les données sur le nombre d'homicide à une variable continue (on reviendra sur cette hypothèse plus tard). Comme par ailleurs, le nombre d'homicides est toujours positif, on spécifiera le modèle sous une forme exponentielle qu'on log-linéarisera du fait que le nombre d'homicides n'est jamais égal à zéro. Au total, on retiendra une forme fonctionnelle de type: $\log(nbr\ homicides_t) = \beta_1 + \beta_1 x_{t1} + \cdots + \beta_K x_{tK} + u_t$]

2- TP1 SUR STATA: REGRESSION LINEAIRE

Les commandes qui seront mobilisées sont:

- L'utilisation de Fichier 'do';
- L'importation des données à partir de Excel;
- L'importation des données à partir de la base de données WDI (World Development Indicators) de la Banque mondiale;
- **rename**: renommer une variable (exple: rename oldvariable newvariable)
- **sort**: classer les données suivant une variable donnée (exple: sort var1)
- **keep**: ne garder que les variables indiquées (keep var1 var2 var3 etc)
- **keep if**: (condition) : ne garder que les enregistrements qui vérifient la condition spécifiée (exemple: keep if var1==100). Remarquez le symbole '==': après la commande *if*, on écrit toujours deux fois le signe d'égalité == car il ne s'agit d'une affectation d'une valeur à la variable considérée.
- **drop**: exclure les variables indiquées (drop var1 var2 var3 etc)
- **drop if** (condition): exclure tous les enregistrements qui vérifient la condition spécifiées (exple: drop if var2==200)

2- TP1 SUR STATA: REGRESSION LINEAIRE

Premières sorties de la commande *reg*

- Exécuter une régression linéaire (commande : reg)

  reg vardep varindep1 varindep2 varindep3

- Générer la variable calculée qu'on va appeler 'ychapeau' :

  predict ychapeau, **xb**

- Générer la variable représentant les résidus d'estimation qu'on va appeler 'uchapeau'

  predict uchapeau , **residual**

- Construire des graphes

- On va porter les pays sur l'axe des x et le taux d'homocides observé, le taux d'homicides calculé et le résidus d'estimation. Cependant, on ne peut porter sur l'axe des x que des nombre et non pas des caractères comme le nom de pays. Pour cela, on va classer les pays par ordre alphabétique et leur affecter un numéro selon cet ordre:

  sort countryname
  generate id=**_n**

2- TP1 SUR STATA: REGRESSION LINEAIRE

Sortie des graphiques

twoway line vardep id

twoway line ychapeau id

twoway line vardep ychapeau id (les deux variables vardep et ychapeau sur le même graphe

twoway (line vardep id) **(line** ychapeau id) (donne les mêmes sorties que la commande ci-dessus.

twoway (line vardep id) **(line** ychapeau id, **mlcolor**(red)) (donne les mêmes sorties que la commande ci-dessus mais avec le deuxième graphe en rouge).

Scatter graph

L'ESTIMATEUR MCO: QUALITE DE L'AJUSTEMENT

On pose avec les notations ci-dessus :

$$SST = \sum_{t=1}^{T} (y_t - \bar{y}_t)^2$$

$$SSE = \sum_{t=1}^{T} (\hat{y}_t - \bar{y}_t)^2$$

$$SSR = \sum_{t=1}^{T} (y_t - \hat{y}_t)^2 = \sum_{t=1}^{T} \hat{u}_t{}^2$$

Alors, le coefficient de détermination R^2 est donné par:

$$R^2 = \frac{SSE}{SST}$$

R^2 s'interpréte comme la part de la variance de y expliquée par la régression. C'est un indicateur du 'goodness of fit'.

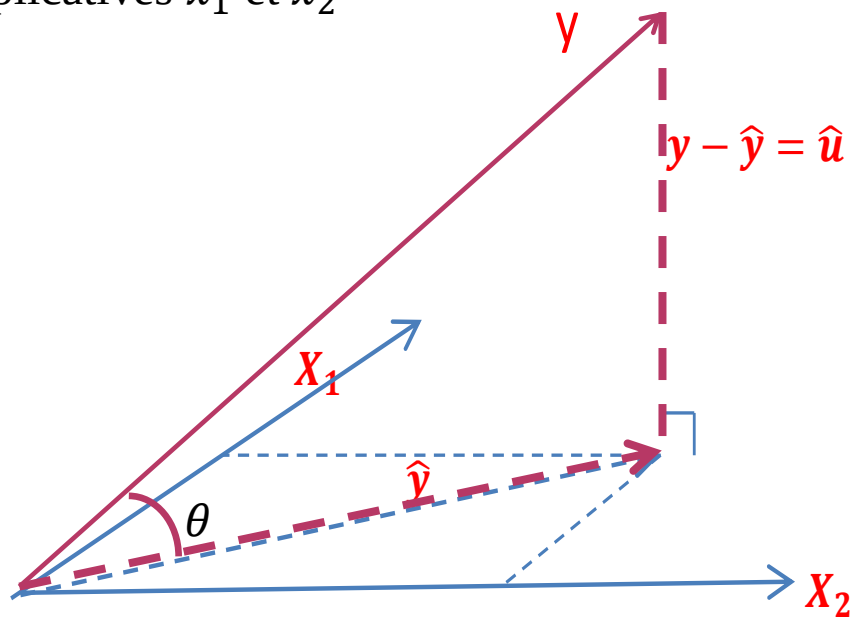
R^2 est également le carré du coefficient de corrélation entre les variables y et $\hat{y}$. C'est donc un indicateur qui est toujours compris entre 0 et 1.

Ce coefficient croit automatiquement avec le nombre de variables explicatives. C'est une limite de cette indicateur. D'où l'utilisation du R^2 ajusté $(\overline{R^2})$ qui introduit une pénalité lorsqu'on introduit des variables supplémentaires:

$$\overline{R^2} = 1 - [SSR/(N-K)]/[SST/(n-1)]$$

REPRESENTATION GEOMETRIQUE DE LA REGRESSION PAR MCO

On considère le modèle linéaire :$y_t = b_1 x_{1t} + b_2 x_{2t} + u_t$
Alors $\hat{y}$ obtenue par MCO est la projection de *y* sur l'espace engendré par les variables explicatives x_1 et x_2



Si les variables sont centrées, alors R^2 est le cosinus de l'angle θ. On voit que la régression est d'autant plus précise ($\|\hat{u}\|$ petit) que θ est petit et donc que $cos\theta$ est proche de 1.
On remarque également que le R^2 peut se réécrire:
$R^2 = 1 - SSR/SST$

On observe également que:

- $\hat{u}$ est orthogonal aux variables explicatives $x_1$et x_2: leur produit scalaire est nul: $\sum_{t=1}^{N} x_{it}\hat{u}_t = 0$ pour i=1,2
- $\hat{u}$ est orthogonal à $\hat{y}$. D'où; $\hat{y}'\,\hat{u} = \sum_{t=1}^{N} y_t\hat{u}_t = 0$)
- SST=SSE+SSR (théorème de Pythagore) – (retrouvez tous ces résultats analytiquement

L'ESTIMATEUR MCO: PROPRIETES DANS LE CAS D'HOMOSCEDASTICITE DES ERREURS

Espérance mathématique de l'estimateur MCO :

Avec l'hypothèse d'absence de multicolinéarité (H1), on fait de plus l'hypothèse que :

$$E(\boldsymbol{u}) = 0 \qquad \text{(H2)}$$

Propriété 1: Sous les hypothèses (H1) et (H2), l'estimateur MCO $\hat{\beta}$ est sans biais $\left(E(\hat{\beta}) = \beta\right)$.

En effet, d'après (3), $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}$. Donc: $\boldsymbol{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{E}(\boldsymbol{u}.)$. D'où:

$$\boldsymbol{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

Variance de l'estimateur MCO:

— Expression de la variance des coefficients estimés.

D'après (3): $Var(\widehat{\boldsymbol{\beta}}) = Var\left(\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\right)$. Comme $Var(\boldsymbol{A} + \boldsymbol{X}) = \boldsymbol{X}$, alors $Var(\widehat{\boldsymbol{\beta}}) = \mathrm{Var}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}$. Comme de plus $Var(\boldsymbol{AX}) = \boldsymbol{A}Var(\boldsymbol{X})\boldsymbol{A}'$, alors:

$$Var(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathrm{Var}(\mathbf{u})\mathbf{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \qquad (4)$$

L'ESTIMATEUR MCO: PROPRIETES DANS LE CAS D'HOMOSCEDASTICITE DES ERREURS

Ainsi, $Var(\widehat{\boldsymbol{\beta}})$ dépend de la valeur de $Var(\boldsymbol{u})$. Dans ce cadre, on fait l'hypothèse l'homoscédasticité des erreurs:

Le processus des erreurs $u = (u_t)$ vérifie:

$$Var(\boldsymbol{u}) = \sigma_u^2 \, \mathrm{I} \qquad \text{(H3)}$$

Alors: $Var(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\sigma_u^2\boldsymbol{I}\mathbf{X}(\boldsymbol{X'X})^{-1} = \sigma^2(\boldsymbol{X'X})^{-1}$. D'où:

Propriété 2:

Sous les hypothèses (H1)et (H3), on a:

$$Var(\widehat{\boldsymbol{\beta}}) = \sigma_u^2 \, (\boldsymbol{X'X})^{-1}.$$

Cependant, la variance de l'erreur σ_u^2 est inconnue. Un estimateur sans biais de σ_u^2 est :

$$\hat{\sigma}^2 = \widehat{\boldsymbol{u}}'\widehat{\boldsymbol{u}}/(N-K) \qquad (5)$$

L'ESTIMATEUR MCO: PROPRIETES DANS LE CAS D'HOMOSCEDASTICITE DES ERREURS

— Efficience de l'estimateur MCO: théorème de Gauss-Markov.

Théorème de Gauss Markov:

Sous les hypothèses (H1), (H2) et (H3), l'estimateur MCO est BLUE (Best Linear Unbiaised Estimator).

Cela signifie que parmi les estimateurs $\tilde{\boldsymbol{\beta}}$ qui sont:

- linéaires en **y** ($c\grave{a}d. \tilde{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{y}$),

- sans biais ($c\grave{a}d\ E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$),

alors l'estimateur MCO $\hat{\boldsymbol{\beta}}$ est celui dont la variance est minimale.

Remarque: Variance minimale est prise dans le sens où la matrice $V(\tilde{\beta}) - V(\tilde{\beta})$ est semi définie positive.

Pour une démonstration, voir Woolridge page 760.

C- INFERENCE STATISTIQUE DANS LE MODELE DE REGRESSION LINEAIRE GAUSSIEN

Distribution de probabilité des coefficients

Afin de procéder aux inférences statistiques, il est nécessaire de se donner une distribution de probabilité des erreurs du modèle. Pour cela, on suppose que les erreurs u_t sont indépendants et identiquement distribuées (iid) suivant la loi normale $N(0, \sigma_u^2)$.

D'où l'hypothèse suivante:

Les erreurs (u_t) sont indé pendants et identiquement distribuées suivant la loi normale:

$$\boldsymbol{u} \rightsquigarrow \mathcal{N}(0, \sigma_u^2 \boldsymbol{I}) \qquad \text{(H4)}$$

Alors:

Propriété 3:

Sous (H1)-(H4), les coefficients estimés $\widehat{\boldsymbol{\beta}}$ suivent la loi normale multivariée d'espérance $\boldsymbol{\beta}$ et de matrice de variance-covariance $\sigma_u^2\ (\boldsymbol{X'X})^{-1}$ **(**stabilité par linéarité de la loi normale) :

$$\widehat{\boldsymbol{\beta}} \rightsquigarrow \mathcal{N}\left(\beta, \sigma_u^2\ (\boldsymbol{X'X})^{-1}\right)$$

Pour un coefficient donné $\hat{\beta}_j$, sa variance est donnée par $\text{var}(\hat{\beta}_j) = \sigma_{\hat{\beta}_j}^2 = \sigma_u^2 c_{jj}$ où c_{jj} est l'élément d'ordre j situé sur la diagonale de $(\boldsymbol{X'X})^{-1}$ de sorte que:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}} \rightsquigarrow \mathcal{N}\left(\beta_j, \sigma_{\hat{\beta}_j}^2\right) \text{ avec } \sigma_{\hat{\beta}_j}^2 = \sigma_u^2\, c_{jj}$$

INFERENCE STATISTIQUE DANS LE MODELE DE REGRESSION LINEAIRE GAUSSIEN

En passant aux variables centrées et réduites, on voit que :

$$\frac{\hat{\beta}_j - \beta_j}{\sigma_u \sqrt{c_{jj}}} \rightsquigarrow \mathcal{N}(0,1) \qquad (6)$$

où $\sigma_u \sqrt{c_{jj}}$ est l'écart-type $\sigma_{\hat{\beta}_j}$ de $\hat{\beta}_j$

On voit que la normalité de $\boldsymbol{u}$ se transmet en quelque sorte à $\widehat{\boldsymbol{\beta}}$

Le problème avec cette formalisation est que l'écart-type σ_u est inconnu et donc l'écart type . $\sigma_{\hat{\beta}_j}$ est lui aussi inconnu. Cependant, on dispose avec (5) d'un estimateur $\hat{\sigma}_u$ sans biais de σ_u. Par ailleurs, on montre qu'en substituant $\hat{\sigma}_u$ à σ_u dans (6), on parvient à la propriété suivante:

Propriété 4:

Sous (H1)-(H4), l'estimateur standardisé $\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_u \sqrt{c_{jj}}}$ suit une loi de Student à N-K degrés de liberté où $\hat{\sigma}_u \sqrt{c_{jj}}$ est l'écart-type estimé de $\hat{\beta}_j$ et dans lequel $\hat{\sigma}_u$ est donné par (5). En notant cet écart-type par $\hat{\sigma}_{\beta_j}$, alors*:*

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\beta_j}} \rightsquigarrow t_{N-K} \qquad (7)$$

Il faut noter que plus le degré de liberté est grand et plus la distribution de Student s'approche de la distribution normale. Cette propriété est largement exploitée dans les études économétriques.

INFERENCE STATISTIQUE DANS LE MODELE DE REGRESSION LINEAIRE GAUSSIEN

Un paramètre β_j du modèle représente une caractéristique inconnue de la population et ne peut être connu avec certitude. On peut toutefois faire des hypothèses sur sa valeur (particulièrement $\beta_j = 0$) et tester cette hypothèse.

Test de l'hypothèse $\beta_j = 0$:

On considère l'hypothèse nulle H_0: $\beta_j = 0$ contre l'hypothèse alternative H_1: $\beta_j \neq 0$.

D'après (7), $\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\beta_j}} \rightsquigarrow t_{N-K}$

Donc sous l'hypothèse H_0, $\frac{\hat{\beta}_j}{\hat{\sigma}_{\beta_j}} \rightsquigarrow t_{N-K}$. Si on pose $t_{calc} = \frac{\hat{\beta}_j}{\hat{\sigma}_{\beta_j}}$, alors la règle de décision est:

- Si $t_{calc} > t_{N-K}^{\alpha/2}$ alors on rejette H_0 et on considère que β_j est significativement différent de 0 au seuil de α.
- Si $t_{calc} < t_{N-K}^{\alpha/2}$ alors on accepte H_0 et on considère que β_j est non significativement différent de 0 au seuil de α.

Pour rappel, pour $\alpha = 5\%$, $t_{N-K}^{\alpha/2}$=2 pour N-K=60 et $t_{N-K}^{\alpha/2}$=1.98 pour N-K=120

INFERENCE STATISTIQUE DANS LE MODELE DE REGRESSION LINEAIRE GAUSSIEN

Test de l'hypothèse $\beta_j = b \neq 0$:

L'hypothèse nulle est H_0: $\beta_j = b$ contre l'hypothèse alternative H_1: $\beta_j \neq b$.

La conduite du test est similaire au cas où $b = 0$ sauf que la statistique t_{calc} sera:

$$t_{calc} = \frac{\hat{\beta}_j - b}{\hat{\sigma}_{\beta_j}}$$

Test de signification globale de la régression :

Il s'agit de tester l'hypothèse que tous les coefficients de la régression (hormis la constante) sont nuls:

L'hypothèse nulle est H_0: $\beta_2 = \beta_3 = \cdots = \beta_K = 0$ contre l'hypothèse alternative H_1: $\exists j$ tel que $\beta_j \neq 0$.

La statistique du test est :

$$F^* = \frac{\sum_{t=1}^{N}(\hat{y}_t - \bar{y})^2/(K-1)}{\sum_{t=1}^{T}\hat{u}_t{}^2/(N-K)} = \frac{SSE/(K-1)}{SSR/(N-K)}$$

Sous l'hypothèse H_0, la statistique F^* suit une loi de Fisher:

$$F^* \rightsquigarrow F(K-1, N-K)$$

Si $F^* > F\ théorique$, alors, on rejette l'hypothèse H_0 selon laquelle la régression n'est pas globalement significative

2- TP2 SUR STATA: MISE EN ŒUVRE DE LA REGRESSION MCO SOUS STATA

Commande stata: reg homicides popjeune pibcap (interface)

SSR SSE N-K K-1 $SSE/(K-1)$ $SSR/(N-K)$

Source	SS	df	MS		
Model	2631.43585	2	1315.71793	Number of obs =	116
Residual	19059.5549	113	168.668627	F(2, 113) =	7.80
Total	21690.9907	115	188.617311	Prob > F =	0.0007
				R-squared =	0.1213
				Adj R-squared =	0.1058
				Root MSE =	12.987

homicides	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
popjeune	.885368	.3303054	2.68	0.008	.2309734	1.539763
pibcap	-.0577112	.0763017	-0.76	0.451	-.2088787	.0934563
_cons	-12.16653	8.994334	-1.35	0.179	-29.98593	5.652866

$$F_{calc} = \frac{SSE/(K-1)}{SSR/(N-K)}$$

p-value: seuil pour lequel le F calculé est égal au F théorique (test de significativité globale de la régression

Coefficient estimé $\hat{\beta}_j$

Ecart-type estimé $\hat{\sigma}_j$

$$t_{calc} = \frac{\hat{\beta}_j}{\hat{\sigma}_j}$$

p-value: seuil pour lequel le t calculé est égal au t théorique

$$[\hat{\beta}_j - t_{N-K}^{2.5\%}\hat{\sigma}_j,\ \hat{\beta}_j + t_{N-K}^{2.5\%}\hat{\sigma}_j]$$