1 LA RÉGRESSION MULTIPLE

1.1 Le modèle linéaire classique.

(Johnston et Dinardo, ch. 3)

$$y_n = \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_k x_{kn} + \epsilon_n, \quad n = 1, \cdots, N \tag{1}$$

où y_n représente le vecteur des variables endogènes, x_{in} sont les variables exogènes et ϵ_n est un vecteur de termes d'erreur. La présence du terme d'erreur signifie que la relation n'est pas exacte. En particulier, ce terme peut contenir les variables manquantes (mais peu pertinentes) ou les erreurs de mesure.

Exemples:

1. Équation de demande
2. Fonction de production
3. Modèles macroéconomiques

On réecrit le modèle 1 sous une forme matricielle

$$\underbrace{Y}_{N\times 1} = \underbrace{X\beta}_{N\times K, K\times 1} + \underbrace{\epsilon}_{N\times 1} \tag{2}$$

où

$$Y = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{N\times 1}, \quad X = \underbrace{\begin{bmatrix} x_{11} & x_{21} & \cdots & x_{k1} \\ x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1N} & x_{2N} & \cdots & x_{kN} \end{bmatrix}}_{N\times K}, \quad \beta = \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}}_{K\times 1}, \quad \varepsilon = \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}}_{N\times 1}$$

1.1.1 Hypothèses du modèle linéaire classique

1-La forme fonctionnelle est linéaire. Le modèle linéaire est plus général qu'on peut le croire. Par exemple, prenons le modèle linéaire suivant:

$$y_n = Ax_n^{\beta} \exp(\epsilon). \tag{3}$$

On peut appliquer une transformation logarithmique à ce modèle pour obtenir une forme fonctionnelle linéaire. Ainsi,

$$\ln y_n \quad = \quad \ln A + \beta \ln x_n + \epsilon. \tag{4}$$

On peut également considérer l'exemple suivant:

$$y_n \quad = \quad \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n}^2 + \epsilon, \tag{5}$$

2- X est fixe (non aléatoire). Pour des échantillons répétés, X prend toujours la même valeur.

3- La matrice X est de $Rang = K$. Ceci implique que:

- Le nombre d'observations $\geq$ le nombre de variables explicatives.

- Il n'y a pas de relation linéaire parfaite entre les variables explicatives.

Donc, X est de plein rang.

Examinons, l'exemple suivant:

$$\underbrace{y_n}_{N\times 1} = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \epsilon_n. \tag{6}$$

On suppose que

$$x_{2n} = cx_{1n}$$

et on suppose que cette dernière relation est non observable. Alors,

$$\begin{aligned} y_n &= \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \epsilon_n, \\ y_n &= \beta_0 + \underbrace{(\beta_1 + c\beta_2)}_{\beta_1^*} x_{1n} + \epsilon_n, \\ y_n &= \beta_0 + \beta_1^* x_{1n} + \epsilon_n, \end{aligned}$$

avec $\beta_1^* = (\beta_1 + c\beta_2)$

4- $E(\epsilon) = 0$ où E est l'espérance mathématique. On aura alors,

$$\begin{aligned} E(Y) &= E(X\beta) + E(\epsilon) \quad (7) \\ &= X\beta. \quad (8) \end{aligned}$$

On a donc,

$$E(\epsilon) = \begin{bmatrix} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_N) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

5-La matrice de variance-covariance des termes d'erreurs est donnée par:

$$E(\underbrace{\epsilon}_{N\times 1}\underbrace{\epsilon'}_{1\times N}) = \sigma^2 \underbrace{I}_{N\times N}, \quad \text{(sphérique)} \quad (9)$$

puisque

$$E(\epsilon) = 0$$

ceci implique que

$$E(\epsilon\epsilon') = Var(\epsilon)$$

On a donc,

$$\begin{bmatrix} var(\epsilon_1) & cov(\epsilon_1, \epsilon_2) & \cdots & cov(\epsilon_1, \epsilon_N) \\ cov(\epsilon_2\epsilon_1) & var(\epsilon_2) & \cdots & cov(\epsilon_2, \epsilon_N) \\ \vdots & \vdots & \vdots & \vdots \\ cov(\epsilon_N\epsilon_1) & cov(\epsilon_N, \epsilon_2) & \cdots & var(\epsilon_N) \end{bmatrix}$$

$\Rightarrow var\left[\epsilon\right] = \sigma^2 I$

- Chaque terme d'erreur ϵ à la même variance, σ^2 (homoscédasticité vs. hétéroscédasticité).

- Les termes d'erreurs ne sont pas corrélés entre eux.

Remarque: On ne spécifie pas la loi pour le terme d'erreur

1.2 Les moindres carrés ordinaires

On cherche à minimiser la somme des carrés des termes d'erreur. Ainsi,

$$\begin{aligned} S &= \sum_{n=1}^{N}(Y_n - \beta_1 X_{1n} - \beta_2 X_{2n} - ... - \beta_k X_{kn})^2, \\ &= \underbrace{(Y - X\beta)'}_{1\times N}\underbrace{(Y - X\beta)}_{N\times 1} \end{aligned} \tag{10}$$

L'estimateur de β est obtenu comme étant la solution du problème suivant:

$$\hat{\beta} = argmin\ (Y - X\beta)'(Y - X\beta) \tag{11}$$

$$= argmin\ S \tag{12}$$

On réécrit la fonction S:

$$Y'Y - 2Y'X\beta + \beta'X'X\beta \tag{13}$$

Remarque:

$$\frac{\partial Z'AZ}{\partial Z} = 2AZ,$$

et

$$\frac{\partial AZ}{\partial Z} = A'.$$

Les conditions du premier ordre (C.P.O.) sont par les équations suivantes:

$$\frac{\partial S}{\partial \beta} = -2X'Y + 2X'X\beta = 0, \tag{14}$$

donc,

$$(X'X)\hat{\beta} = X'Y, \tag{15}$$

mais comme $(X'X)$ est de rang K, on peut donc l'inverser. On obtient,

$$\hat{\beta} = (X'X)^{-1}X'Y, \tag{16}$$

De plus, la dérivé seconde est donnée par

$$\frac{\partial^2 S}{\partial \beta \partial \beta'} = 2X'X, \tag{17}$$

qui est une matrice définie positive. On a donc un minimum.

Remarque: M.C.O. vs. M.C.G.

Pour les moindres carrés généralisés, on a

$$S = (Y - X\beta)'W(Y - X\beta) \tag{18}$$

où W est positive définie. Les moindres carrés ordinaires correspondent au cas où $W = I$.

On va maintenant étudier les propriétés de l'estimateur des moindres carrés ordinaires. En particulier, on va montrer que cet estimateur est sans biais et qu'il est l'estimateur optimal parmi les estimateurs linéaires sans biais.

Definition 1 *Un estimateur $\hat{\beta}$ du vecteur de paramètres β est sans biais si et seulement si*

$$E(\hat{\beta}) = \beta$$

.

On va montrer que l'estimateur des M.C.O. est un estimateur sans biais de β

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}X'(X\beta + \epsilon) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\epsilon \end{aligned}$$

et en introduisant l'opérateur espérance,

$$\begin{aligned} E(\hat{\beta}) &= \beta + E((X'X)^{-1}X'\epsilon) \\ &= \beta + (X'X)^{-1}X'E(\epsilon) \\ &= \beta \end{aligned}$$

puisque $E(\epsilon) = 0$. Donc, l'estimateur M.C.O. est sans biais.

Definition 2 *Un estimateur est optimal parmi la classe des estimateurs sans biais si sa variance est la plus petite parmi cette classe.*

Théorème 1 (Théorème de Gauss-Markov) *L'estimateur des M.C.O. est optimal parmi les estimateurs linéaires sans biais et il a pour variance*

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1} \tag{19}$$

Preuve

$\hat{\beta}$ est bien un estimateur linéaire, en effet

$$\hat{\beta} = (X'X)^{-1}X'Y = AY. \tag{20}$$

La matrice de variance-covariance de $\hat{\beta}$ est

$$\begin{aligned} E(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))' &= E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] \\ &= E\left[(X'X)^{-1}X'(\epsilon\epsilon')X(X'X)^{-1}\right] \\ &= (X'X)^{-1}X'E(\epsilon\epsilon')X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

On va maintenant montrer que l'estimateur M.C.O. est de variance minimale. Prenons un autre estimateur linéaire de forme générale.

$$\beta^* = CY \tag{21}$$

ainsi,

$$\beta^* = CY = C(X\beta + \epsilon) \tag{22}$$

et

$$E(\beta^*) = CX\beta \tag{23}$$

On suppose que cet estimateur est sans biais, on aura alors que $CX = I$. Comparons maintenant la variance de β^* avec la variance de $\hat{\beta}$;

$$\begin{aligned} var(\beta^*) &= var(\beta^* - \hat{\beta} + \hat{\beta}) \\ &= var(\beta^* - \hat{\beta}) + 2cov(\beta^* - \hat{\beta}, \hat{\beta}) + var(\hat{\beta}). \end{aligned}$$

On cherche maintenant l'expression pour le terme de covariance:

$$\begin{aligned} cov(\beta^* - \hat{\beta}, \hat{\beta}) &= E\left[\left(\beta^* - \beta - (\hat{\beta} - \beta)\right)\left(\hat{\beta} - \beta\right)'\right] \\ &= E\left(C\epsilon - (X'X)^{-1}X'\epsilon, ((X'X)^{-1}X'\epsilon)'\right) \\ &= \sigma^2 CX(X'X)^{-1} - \sigma^2(X'X)^{-1} = 0 \end{aligned}$$

puisque $CX = I$.

On a donc que:

$$var(\beta^*) = var(\beta^* - \hat{\beta}) + var(\hat{\beta}).$$

La matrice $var(\beta^* - \hat{\beta})$ étant semi définie positive, la matrice de variance-covariance de β^* est donc plus grande ou égale à la matrice de variance-covariance de $\hat{\beta}$. Ce résultat est valide pour toute matrice C (CQFD).

En anglais, on dira Best Linear Unbiased Estimator (BLUE).

Commentaires

- Résultats de petit ćhantillon
- Pas besoin de spécifier la densité du terme d'erreur.

1.3 Estimateur de σ^2

Ce terme n'apparait pas dans S. On a fait l'hypothèse suivante:

$$E(\epsilon\epsilon') = \sigma^2 I.$$

On peut utiliser les résidus des M.C.O. pour calculer cet estimateur. L'estimateur sans biais de σ est:

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{N-K}$$

où $\hat{\epsilon} = Y - X\hat{\beta}$.

Preuve

$$\begin{aligned}
\hat{\epsilon} &= Y - X\hat{\beta} \\
&= Y - X(X'X)^{-1}X'Y \\
&= \left[I - X(X'X)^{-1}X'\right] Y = MY \\
&= \left[I - X(X'X)^{-1}X'\right] [X\beta + \epsilon] \\
&= X\beta - X(X'X)^{-1}X'X\beta + M\epsilon \\
&= M\epsilon \\
\hat{\epsilon} &= M\epsilon
\end{aligned}$$

La matrice M est orthogonale à la matrice X, c.a.d. $MX = 0$. De plus, la matrice M est symétrique ($M = M'$) et idempotente ($MM = M$). Donc,

$$\begin{aligned} E(\hat{\epsilon}'\hat{\epsilon}) &= E(\epsilon' M' M \epsilon) \\ &= E(\epsilon' M \epsilon) \\ &= E\left[tr(\epsilon' M \epsilon)\right] \\ &= E\left[tr(M \epsilon \epsilon')\right] \\ &= tr\left[M E(\epsilon\epsilon')\right] \\ &= tr\ M \left[E(\underbrace{\epsilon\epsilon'}_{\sigma^2\ I})\right] \\ &= \sigma^2 tr\ M \end{aligned}$$

en utilisant le fait que $tr(AB) = tr(BA)$.

On cherche maintenant la valeur de trM,

$$\begin{aligned} tr\ M &= tr\left[I_N - X(X'X)^{-1}X'\right] \\ &= tr\ I_N - tr\ (X(X'X)^{-1}X') \\ &= tr\ I_N - tr\ ((X'X)^{-1}X'X) \\ &= tr\ I_N - tr\ I_K = N - K \end{aligned}$$

Ainsi

$$E\hat{\epsilon}'\hat{\epsilon} = \sigma^2(N - K)$$

ce qui implique que

$$E\left(\frac{\hat{\epsilon}'\hat{\epsilon}}{N - K}\right) = \sigma^2$$

qui est donc un estimateur sans biais de σ^2. On a les expressions équivalentes suivantes:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{N - K} \\ &= \frac{Y'\left[I - X(X'X)^{-1}X'\right]Y}{N - K}. \end{aligned}$$

En résumé, $\hat{\beta}$ est une fonction linéaire de Y,
$\Rightarrow \hat{\beta} = AY$.
$\hat{\sigma}^2$ est une fonction quadratique de Y,
$\Rightarrow \hat{\sigma}^2 = \frac{Y'AY}{N-K}$.

Il est important de noter qu'il n'existe pas de propriété d'optimalité pour $\hat{\sigma}^2$.

1.4 Aspects algébriques des M.C.O

On avait comme C.P.O.

$$-2X'Y + 2X'X\hat{\beta} = -X'(Y - X\hat{\beta}) = -X'\hat{\epsilon} \tag{24}$$

Ce qui veut dire que chaque colonne des X est orthogonale aux résidus, c.à.d. $X_k'\hat{\epsilon} = 0$.

Puisque la première colonne de la matrice X est une colonne de 1, on a les implications suivantes:

- La somme des résidus est égale à zéro. En effet, $X_1'\hat{\epsilon} = i'\hat{\epsilon} = \sum_{i=1}^{n} \hat{\epsilon}_i = 0$.

- L'hyperplan de la régression passe par le point moyen des données, c.à.d. $\bar{Y} = \bar{X}\hat{\beta}$.

- $E(y) = X\hat{\beta}$.

Il est important de savoir qu'aucune de ces implications tient si la régression ne contient pas un vecteur de 1.

1.5 Interprétation géométrique des M.C.O.

(Chapitre 1. Davidson et MacKinnon (1993))

Concept de projection:

$$Y = \underbrace{X\hat{\beta}}_{\hat{Y}} + \hat{\epsilon}$$

$$Y = \underbrace{X(X'X)^{-1}X'}_{P_x} Y + \hat{\epsilon}.$$

P_x est la matrice de projection orthogonale dans l'espace engendré par les X.

$$\begin{aligned} \hat{\epsilon} &= Y - X\hat{\beta} \\ &= Y - X(X'X)^{-1}X'Y \\ &= \left[I - X(X'X)^{-1}X'\right] Y \\ &= M_x Y. \end{aligned}$$

M_x est matrice de projection de Y sur l'espace orthogonale aux X. On dit que deux vecteurs sont orthogonaux si $A'B = 0$. Ainsi, $[I - X(X'X)^{-1}X']\, X = 0$. Par les CPO

$$\begin{aligned} -2X'Y + 2X'X\hat{\beta} &= 0 \\ X'(Y - X\hat{\beta}) &= 0 \\ X'\hat{\epsilon} &= 0 \end{aligned}$$

Les résidus sont orthogonaux aux vecteurs des variables explicatives. De plus,

$$\begin{aligned} Y &= \hat{Y} + \hat{\epsilon} \\ &= X(X'X)^{-1}X'Y + \left[I - X(X'X)^{-1}X'\right] Y \\ &= P_x Y + M_x Y \end{aligned}$$

où $P_x Y$ représente ce qui est expliqué dans l'espace engendré par les X et $M_x Y$ représente ce qui n'est pas expliqué par l'espace engendré par les X. On a donc une décomposition orthogonale de l'espace.

1.6 Projection partielle:(Théorème de Frisch-Waugh)

On suppose le modèle linéaire classique avec deux groupes (vecteurs) de variables explicatives.

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

On cherche l'expression analytique du vecteur β_2. Pour les M.C.O. on a que:

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

où $X = (X_1X_2)$ et $K_1 + K_2 = K$. On réécrit:

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_1'Y \\ X_2'Y \end{bmatrix} \begin{matrix} (1) \\ (2) \end{matrix}$$

En utilisant l'équation 1, on obtient:

$$X_1'X_1\hat{\beta}_1 + X_1'X_2\hat{\beta}_2 = X_1'Y$$

puisque $X_1'X_1$ est de rang complet, on peut inverser. Alors

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'[Y - X_2\hat{\beta}_2]$$

En substituant ce résultat dans l'équation 2

$$X_2'X_1\hat{\beta}_1 + X_2'X_2\hat{\beta}_2 = X_2'Y$$

$$X_2'X_1[(X_1'X_1)^{-1}X_1'(Y - X_2\hat{\beta}_2)] + X_2'X_2\hat{\beta}_2 = X_2'Y$$

En manipulant cette expression, on obtient:

$$\hat{\beta}_2 = [X_2'[I - X_1(X_1'X_1)^{-1}X_1']X_2]^{-1}X_2'[I - X_1(X_1'X_1)^{-1}X_1']Y.$$

On définit $M_1 = [I - X_1(X_1'X_1)^{-1}X_1']$ et on a que,

$$M_1X_1 = 0$$

et

$$M_1'M_1 = [I - X_1(X_1'X_1)^{-1}X_1']'[I - X_1(X_1'X_1)^{-1}X_1']$$

$$M_1'M_1 = [I - X_1(X_1'X_1)^{-1}X_1'] \Rightarrow \text{idempotente}$$

Donc,

$$\hat{\beta}_2 = (X_2'M_1X_2)^{-1}X_2'M_1Y$$

$$\hat{\beta}_2 = (X_2'M_1'M_1X_2)^{-1}X_2'M_1'M_1Y$$

En définisssant X_2^* et Y^* comme étant:

$$X_2^* = M_1X_2 \text{ et } Y^* = M_1Y.$$

Alors,

$$\hat{\beta}_2 = (X_2^{*\prime}X_2^*)^{-1}X_2^{*\prime}Y^*.$$

X_2^* correspondent aux résidus la régression de X_2 sur X_1 et Y^* aux résidus de la régression de Y sur X_1. En effet:

$$X_2 = X_1\theta + u$$

$$\hat{\theta} = (X_1'X_1)^{-1}X_1'X_2$$

$$\Rightarrow X_2 = X_1(X_1'X_1)^{-1}X_1'X_2 + \hat{u}$$

$$\Rightarrow X_2 - X_1(X_1'X_1)^{-1}X_1'X_2 = \hat{u}$$

$$[I - X_1(X_1'X_1)^{-1}X_1']X_2 = \hat{u}$$

$$M_1X_2 = \hat{u}$$

.

$\hat{\beta}_2$ correspond donc à l'estimation une fois que l'on a purgé X_2 et Y de X_1.

$\Rightarrow$ Donc, c'est ce qui provient exclusivement de X_2 (information orthogonale à X_1).

APPLICATIONS

1. **Omission de variables explicatives**

   On a le même modèle

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

   On obtenait que

$$\begin{aligned} X_2'X_1\hat{\beta}_1 + X_2'X_2\hat{\beta}_2 &= X_2'Y \\ \Rightarrow X_2'X_2\hat{\beta}_2 &= -X_2'X_1\hat{\beta}_1 + X_2'Y \\ \hat{\beta}_2 &= -(X_2'X_2)^{-1}X_2'X_1\hat{\beta}_1 + (X_2'X_2)^{-1}X_2'Y \end{aligned}$$

   Implications de l'omission du vecteur variables explicatives X_1:

   - Si X_1 et X_2 sont corrélées, alors $\hat{\beta}_2$ est biaisé.
   - Si X_1 et X_2 ne sont pas corrélés alors $\hat{\beta}_2$ est sans biais.

   On peut déduire ces deux implications directement de la formule du théorème Frisch-Waugh pour $\hat{\beta}_2$ donnée par:

$$\hat{\beta}_2 = \left[X_2'[I - X_1(X_1'X_1)^{-1}X_1']X_2\right]^{-1} X_2'[I - X_1(X_1'X_1)^{-1}X_1']Y$$

   .

2. **Déviation par rapport à la moyenne**

   On a toujours le même modèle

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

mais

$$X_1 = i = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix}$$

Si on régresse une variable Z (par exemple) sur i, on obtient la moyenne de Z. En effet

$$m_z = (i'i)^{-1}i'Z = \frac{1}{N}\sum_{n=1}^{N} Z_n$$

Dans le cas où X_1 est égale à i, $\hat{\beta}_2$ correspond à la régression de $(Y - i\bar{Y})$ sur $(X_2 - i\bar{X}_2)$ par le théorème de théorème Frisch-Waugh où $\bar{Y}$ et $\bar{X}_2$ sont les moyennes de Y et X_2.

On a

$$\hat{\beta}_2 = \left[X_2'[I - X_1(X_1'X_1)^{-1}X_1'\right] X_2]^{-1}X_2'[I - X_1(X_1'X_1)^{-1}X_1']Y$$

$$X_2^* = M_1X_2 = [I - i(i'i)^{-1}i']X_2 = (X_2 - i\bar{X}_2)$$

$$Y^* = M_1Y = [I - i(i'i)^{-1}i']Y = (Y - i\bar{Y})$$

$$\hat{\beta}^* = (X_2^{*\prime}X_2^*)^{-1}X_2^{*\prime}Y^*$$

3. **Coefficient de détermination** $(R^2, \bar{R}^2)$

   -Mesure de la performance d'un modèle.

   -Représente la partie expliquée de la variable dépendante par le modèle.

   On a toujours le même modèle:

$$Y = X\beta + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon$$

où ϵ est la partie qui n'est pas expliquée par le modèle et on suppose que $X_1 = i$

$$Y = \hat{Y} + \hat{\epsilon}$$

On a que

$$Y'Y = \hat{Y}'\hat{Y} + \hat{\epsilon}'\hat{\epsilon}$$

puisque $X'\hat{\epsilon} = 0$,

$$\Rightarrow Y'Y = \beta' X' X\hat{\beta} + \hat{\epsilon}'\hat{\epsilon}$$

On définit $M_i = [I - i(i'i)^{-1}i']$. Alors $M_iY = Y - i\bar{Y}$. On prémultiplie notre modèle par M_i.

$$M_iY = M_iX_1\hat{\beta}_1 + M_iX_2\hat{\beta}_2 + M_i\hat{\epsilon}.$$

On a que $MiX_1 = 0$ puisque $X_1 = i$ et $M_i\hat{\epsilon} = \hat{\epsilon}$ puisque $E(\epsilon) = 0$.
Alors,

$$M_iY = M_iX_2\hat{\beta}_2 + \hat{\epsilon}$$

$$Y'M_i'M_iY = \hat{\beta}_2'X_2'M_i'M_iX_2\hat{\beta}_2 + \hat{\epsilon}'\hat{\epsilon}$$

$$R^2 = \frac{\hat{\beta}_2'X_2'M_i'M_iX_2\hat{\beta}_2}{Y'M_iY} = 1 - \frac{\hat{\epsilon}'\hat{\epsilon}}{Y'M_iY}$$

où $Y'M_iY$ est la variance de Y ($\frac{(Y-i\bar{Y})'(Y-i\bar{Y})}{N}$). L'expression pour le R^2 implique que

$$0 \leq R^2 \leq 1$$

Propriété non souhaitable:

Si on augmente le nombre de régresseurs alors la statistique R^2 augmente.

On va définir le R^2 ajusté qui introduit une pénalité pour l'augmentation du nombre de régresseur

$$\bar{R}^2 = 1 - \left[\frac{\left(\frac{\hat{\epsilon}'\hat{\epsilon}}{N-K_2} \right)}{\left(\frac{Y'M_iY}{N-1} \right)} \right]$$

$$\bar{R}^2 = 1 - \frac{(\hat{\epsilon}'\hat{\epsilon})/(N-K_2)}{(Y'M_iY)/(N-1)}$$

où K_2 est la dimension de X_2.

$$\bar{R}^2 = 1 - \frac{N-1}{N-K_2}(1-R^2) \Rightarrow \text{peut être négatif}$$

1.7 Tests d'hypothèses: tests de restrictions linéaires et tests de changement structurel

Greene Chapitre 7

Johnston Chapitre 5 et 6

Rappel

Definition 3 *On appelle loi du khi-deux à K degrés de liberté la loi de $Y = X_1^2 + ... + X_K^2 = ||X||^2$, où les variables X_K sont indépendantes, de lois respectives $N(m_K, 1)$. Lorsque $m_K = 0$, pour tout K, on parle de khi-deux centrée, de Khi-deux décentrée dans le cas contraire.*

Propriété 1 *La loi de $Y = X_1^2 + ... + X_K^2$, les X_k suivant indépendamment $N(m_k, 1)$, ne dépend que de K et de $\lambda = \sum_{k=1}^{K} m_k^2 = \|m\|^2$. Elle peut être notée $\chi^2(K, \lambda)$. λ est appelé paramètre de décentrage. Une loi du Khi-deux centrée est notée $\chi^2(K) = \chi^2(K, 0)$.*

Propriété 2 *Si $Y \sim \chi^2(K, \lambda)$, alors $EY = K + \lambda$, $VAR(y) = 2(K + 2\lambda)$. En particulier si $EY \sim \chi^2(k)$, $E(Y) = K$, $VAR(y) = 2K$.*

Definition 4 *: On appelle loi de Student à K degrés de liberté, la loi de $Z = \frac{X}{\sqrt{\frac{Y}{K}}}$, où X et Y sont deux variables indépendantes suivant respectivement $N(m, 1)$ et $\chi^2(K)$. Le paramètre m est la paramètre de décentrage. Les lois de Student centrées $(m = 0)$ sont notées $T(K)$.*

La loi de Student contrées $T(K)$ admet pour densité:

$$f(Z) = \frac{\Gamma(\frac{K+1}{2})}{\Gamma(\frac{K}{2})\Gamma(\frac{1}{2})} \frac{1}{\sqrt{K}} \frac{1}{(1 + \frac{Z^2}{K})^{\frac{K+1}{2}}}$$

où Γ est la loi de Gamma.

Definition 5 *: La loi de Fisher de degrés de liberté K_1 et K_2, notées $F(K_1, K_2)$ est la loi de $Z = \frac{Y_1/K_1}{Y_2/K_2}$, où les variables Y_1 et Y_2 sont indépendantes et suivent respectivement $\chi^2(K_1)$ et $\chi^2(K_2)$.*

1.7.1 Tests d'hytpothèses pour le modèle M.C.O.

Pour l'estimateur des M.C.O, on a que

$$\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon$$

Jusqu'à maintenant, on a fait aucune hypothèse sur la loi que suit le terme d'erreur ϵ. Si on veut faire de l'inférence sur β, une hypothèse devient nécessaire pour ϵ afin d'obtenir des propriétés de petit échantillon. En effet, par l'expresseion suivante:

$$\hat{\beta} = constante + A\epsilon$$

où $A = (X'X)^{-1}X'$, l'estimateur des m.c.o. est une fonction linéaire des termes d'erreur. En plus des hypothèses du modèle linéaire classique, on va supposer que

$$\epsilon \sim N(0, \sigma^2 I).$$

$\hat{\beta}$ suit alors une loi normale. On aura alors des propriétés de petit échantillon.

Lemme 1 *Si $Z \sim N(\mu, \Sigma)$. Alors*

$$AZ + b \sim N[A\mu + b, A\Sigma A']$$

On aura pour $\hat{\beta}$ que

$$\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon.$$

Par le Lemme plus haut, on a l'implication suivante:

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$

Pour un élement du vecteur β, on a

$$\hat{\beta}_k \sim N(\beta_k, \sigma^2 (X'X)^{-1}_{kk})$$

Si on définit S^{kk} comme étant le k ième élément diagonale de $(X'X)$, on aura la statistique centrée réduite suivante:

$$Z_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 S^{kk}}}$$

On va pouvoir faire des tests sur différentes hypothèses par rapport à β.

Exemple:

$$H_o : \beta_k = 0;$$

$$H_1 : \beta_k \neq 0$$

Problème: on ne connaît pas σ^2 pour construire Z_k. Si on connaissait σ^2 on pourrait effectuer un test directement et obtenir un intervalle de confiance. Par exemple, on aurait l'intervalle de confiance suivant où il existe 95% des chances que $\hat{\beta} - \beta$ soit entre -1.96 et 1.96.

Lorsqu'on effectue un test, on doit choisir entre deux possibilités: rejeter ou de pas rejeter H_0 face à H_1. On peut alors commettre deux types d'erreurs:

- **Erreur de Type 1:**
  Rejet de H_0 lorsque H_0 est vraie.

- **Erreur de Type 2:**
  Non rejet de H_0 lorsque H_0 est fausse.

Pour l'approche classique en statistique (l'approche de Neyman), H_0 et H_1 ne sont pas considérées de façon symétrique. On va contrôler le risque d'erreur de type 1. On choisira donc une valeur qui correspond à une certaine

probabilité de commettre une erreur de type 1. On définit le niveau d'un test comme étant cette probabilité d'erreur de type 1. La puissance d'un test, pour sa part, est donnée par la probabilité de rejeter H_0 lorsque H_0 est fausse. L'estimateur des m.c.o. va maximiser la puissance des tests puisque l'estimateur a la variance minimale.

On avait donc la statistique centrée réduite

$$Z_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 S^{kk}}} \sim N(0,1)$$

On ne connaît pas σ^2. On va donc utiliser son estimateur:

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{N-K} \text{ où } \hat{\epsilon} = M\epsilon$$

On construit l'expression suivante:

$$(N-K)\frac{\hat{\sigma}^2}{\sigma^2} = \frac{(N-K)\hat{\epsilon}'\hat{\epsilon}}{(N-K)\sigma^2} = (\frac{\epsilon'}{\sigma})M(\frac{\epsilon}{\sigma})$$

Puisque M est une matrice idempotente, on a donc une forme quadratique idempotente symétrique d'un vecteur suivant une loi normale standard.

Lemme 2 *Si $Z \sim N(0, \sigma^2 I)$ et A est idempotente de rang r. Alors*

$$\frac{1}{\sigma^2} Z'AZ \sim \chi^2(r)$$

On sait que le rang de M est $N-K$. Par le Lemme énoncé plus haut,

$$(\frac{\epsilon'}{\sigma})'M(\frac{\epsilon}{\sigma}) \sim \chi^2(N-K)$$

On a donc:

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 S^{kk}}} \sim N(0,1)$$

et

$$(N-K)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(N-K)$$

On va donc construire la statistique t

$$t = \frac{\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 S^{kk}}}}{\sqrt{\frac{(N-K)\frac{\hat{\sigma}^2}{\sigma^2}}{N-K}}} = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2 S^{kk}}}.$$

Par la DÉFINITION 4, cette statistique suivra une loi de student à $N - K$ degré de liberté si le numérateur et le dénominateur sont deux variables indépendantes.

Pour montrer l'indépendance entre

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 S^{kk}}} \quad \text{et } \sqrt{(N-K)\frac{\hat{\sigma}^2}{\sigma^2}},$$

il est suffisant de montrer l'indépendance entre

$$\frac{\hat{\beta} - \beta}{\sigma} = (X'X)^{-1}X'(\frac{\epsilon}{\sigma}) \quad \text{et} \quad M(\frac{\epsilon}{\sigma}).$$

Lemme 3 *Supposons $Z \sim N(0, \sigma^2 I)$, $Z'AZ$ une forme quadratique où A est une matrice idempotente d'ordre N et LZ est un vecteur de m éléments, ceux-ci étant une combinaison linéaire de Z, alors les variables AZ et LZ sont indépendantes si $LA' = 0$.*

Preuve

$$E(AZZ'L') = 0$$

$$E(AZZ'L') = AE(Z'Z)L' = \sigma^2 AL' = 0$$

$$\Rightarrow AL' = 0 \text{ et } LA' = 0$$

On définit

$$L = (X'X)^{-1}X' \quad \text{et} \quad A = M = [I - X(X'X)^{-1}X'].$$

On doit donc avoir

$$AL' = ML' = 0$$

$$[I - X(X'X)^{-1}X']X(X'X)^{-1} = X(X'X)^{-1} - X(X'X)^{-1}X`X(X`X)^{-1} = 0$$

On a donc

$$t = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2 S^{kk}}} \sim T(N-K).$$

On construit un intervalle de confiance avec la valeur critique de la loi de student pour le niveau désiré. Ainsi, l'intervalle de confiance est donné par:

$$\hat{\beta}_k \pm C_\alpha \sqrt{\sigma^2 S^{kk}}$$

où C_α est la valeur critique.

1.7.2 Tests sur des restrictions linéaires du vecteur de paramètres β

On s'intéresse à un test pour plus d'un coefficient. On utilise la formulation générale suivante:

$$H_0 : R\beta = q$$

où R est dimension $J \times K$ et de rang $J < K$ et q est de dimension $J \times 1$. Cette condition de rang implique qu'il n'existe pas de combinaison linéaire entre les colonnes de R.

Exemples:

1. $H_0 : \beta_k = 0$
   $R = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & . & . & . & 0 \end{bmatrix}$, et $q = 0$
2. $H_0 : \beta_j = \beta_i \Rightarrow \beta_j - \beta_i = 0$
   $R = \begin{bmatrix} 0 & 0 & 1 & 0 & . & . & . & 0 & -1 & 0 & 0 \end{bmatrix}$, et $q = 0$
3. $H_0 : \beta_2 + \beta_3 + \beta_4 = 1$
   $R = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & . & . & . & 0 \end{bmatrix}$, et $q = 1$
4. $H_0 : \beta_1 = 0,\ \beta_2 = 0\ ,\ \beta_3 = 0$
   $R = \begin{bmatrix} 1 & 0 & 0 & 0 & . & . & . & 0 \\ 0 & 1 & 0 & 0 & . & . & . & 0 \\ 0 & 0 & 1 & 0 & . & . & . & 0 \end{bmatrix}$, et $q = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

On remplace β par son estimateur et on évalue si $R\hat{\beta} - q$ est statistiquement différent de zéro. On a que

$$E(R\hat{\beta}) = RE(\hat{\beta}) = R\beta$$

et

$$VAR(R\hat{\beta}) = E(R(\hat{\beta} - \beta))(\hat{\beta} - \beta)'R')$$
$$= \sigma^2 R(X'X)^{-1}R'$$

Puisque $\hat{\beta}$ suit une loi normale multivariée,

$$R\hat{\beta} \sim N(R\beta, \sigma^2 R(X'X)^{-1}R')$$

alors

$$(R\hat{\beta} - R\beta) \sim N(0, \sigma^2 R(X'X)^{-1}R')$$

On veut faire le test suivant:

$$H_0 : R\beta - q = 0$$

On va utiliser le lemme suivant:

Lemme 4 *Si* $Z \sim N(\mu, \Sigma)$ *alors* $\Sigma^{-\frac{1}{2}}(Z-\mu) \sim N(0, I)$. *Si* $Z \sim N(\mu, \Sigma)$ *alors* $(Z - \mu)'\Sigma^{-1}(Z - \mu) \sim \chi^2(n)$ *où* n *est la dimension de* Z *et de* Σ *est de rang complet (de rang* n*).*

On va faire une forme quadratique pondérée par la variance (Test de Type Wald) et en utilisant le Lemme plus haut, on connaît la loi de cette forme. Ainsi,

$$(R\hat{\beta} - q)'[\sigma^2 R(X'X)^{-1}R']^{-1}(R\hat{\beta} - q) \sim \chi^2(J)$$

On ne peut utiliser cette expression puisque σ^2 n'est pas connu. On utilise plutôt son estimateur $\hat{\sigma}^2$. On va former la statistique F suivante:

$$F = \frac{(R\hat{\beta} - q)'[\sigma^2 R(X'X)^{-1}R']^{-1}(R\hat{\beta} - q)/J}{((N-K)\frac{\hat{\sigma}^2}{\sigma^2})/(N-K)}.$$

Réécrivons cette statistique comme étant $F = \frac{X/J}{Y/(N-K)}$. Cette statistique suivra une loi de Fisher $F(J, N-K)$ si X et Y sont indépendants, puisque $X \sim \chi^2(J)$ et que $Y \sim \chi^2(N-K)$. On va utiliser le Lemme suivant pour montrer l'indépendance.

Lemme 5 *Si $Z \sim N(0, \sigma^2 I)$, $Z'AZ$ et $Z'BZ$ sont deux formes quadratiques et que A et B sont des matrices idempotentes symétriques, $Z'AZ$ et Z'BZ sont indépendantes si $AB = 0$.*

Preuve: On a $A = A'A$ et $B = B'B$

$$Z'AZ = Z'A'AZ,\ Z_1 = AZ$$
$$Z'BZ = Z'B'BZ,\ Z_2 = BZ$$
$$E(Z_1 Z_2') = AE(ZZ')B' = \sigma^2 AB' = 0 \Rightarrow AB' = 0$$

On réécrit la statistique F:

$$F = \frac{(R(\hat{\beta} - \beta)/\sigma)'[R(X'X)^{-1}R']^{-1}(R(\hat{\beta} - \beta)/\sigma)/J}{(M\frac{\epsilon}{\sigma})'(M\frac{\epsilon}{\sigma})/N - K}$$

puisque $\frac{R(\hat{\beta}-\beta)}{\sigma} = R(X'X)^{-1}X'(\frac{\epsilon}{\sigma})$.

Alors le numérateur de F est égal à

$$(\frac{\epsilon}{\sigma})'X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'(\frac{\epsilon}{\sigma})$$

où $X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'$. Prenons cette dernière comme étant A et la matrice M comme étant B. On peut maintenant appliquer le Lemme. On doit donc montrer $AB' = 0$ ou de façon équivalente que $AM = 0$

$$[X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'][I - X(X'X)^{-1}X'] =$$

$$X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'$$

$$-X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'X(X'X)^{-1}X'$$

et cette dernière expression est bien égale à 0. Alors, $F \sim F(J, N-K)$

On peut réécrire la statistique F de la façon suivante:

$$F = (R\hat{\beta} - q)'[\hat{\sigma}^2 R(X'X)^{-1}R']^{-1}(R\hat{\beta} - q)/J$$

Commentaires:

Pour un test portant sur un seul paramètre.

$$T^2(N-K) = F(1, N-K)$$

Pour un test conjoint, pourquoi ne pas utiliser une statistique t sur deux tests séparés?

$$H_0 : \beta_1 = 0 \text{ et } H_0 : \beta_2 = 2$$

Raison: $\hat{\beta}_1$ et $\hat{\beta}_2$ sont corrélées, la statistique F en tient compte.

1.7.3 DEUX APPROCHES POUR EFFECTUER DE L'INFÉRENCE

(Greene, Ch. 7, Johnston, ch. 6.)

1.7.4 Inférence basée sur un modèle sans contrainte

On a le modèle non contraint suivant:

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

On effectue alors un test sur une ou plusieurs restrictions: $R\beta = q$. Prenons l'exemple suivant:

$$H_0 : \quad \beta_2 = 0$$

On se pose la question suivante: est-ce que le vecteur β_2 est significativement différent de zéro en tenant compte de l'incertitude entourant cet estimateur?

1.7.5 Inférence basée sur un modèle contraint

Pour l'hypothèse nulle définie plus haut, on a le modèle contraint correspondant qui est:

$$Y = X_1\beta_1 + \epsilon^*$$

Ce modèle aura une moins bonne performance que le modèle non contraint à moins que la contrainte soit respectée (en tenant compte de l'incertitude).
On peut effectuer un test sur les restrictions en comparant la performance du modèle contraint vs. non contraint.

On va utiliser la forme générale pour les restrictions:

$$R\beta = q.$$

On incorpore ces restrictions au problème des M.C.O. On minimise

$$S(\beta) = (Y - X\beta)'(Y - X\beta)$$

sous la contrainte que

$$R\beta - q = 0$$

On écrit le $\mathcal{L}$agrangien

$$\mathcal{L}(\beta) = (Y - X\beta)'(Y - X\beta) + 2\lambda'(R\beta - q)$$

où λ est un vecteur $(J \times 1)$.

Les conditions du premier ordre sont données par les expressions suivantes:

$$\begin{aligned} \frac{\partial \mathcal{L}(\beta)}{\partial \beta} &= -2X'Y + 2X'X\beta + 2R'\lambda = 0 \\ \frac{\partial \mathcal{L}(\beta)}{\partial \lambda} &= 2(R\beta - q) = 0 \end{aligned}$$

En utilisant le premier groupe de C.P.O. et en posant l'égalité à zéro, on obtient

$$-X'Y + X'X\beta^* = -R'\lambda^*$$

On prémultiplie par $(X'X)^{-1}$,

$$\beta^* = \underbrace{(X'X)^{-1}X'Y}_{\hat{\beta}} - (X'X)^{-1}R'\lambda^*$$

Par le deuxième groupe de C.P.O., on a:

$$R\beta^* = q.$$

Alors,

$$R\beta^* = R\hat{\beta} - R(X'X)^{-1}R'\lambda^* = q.$$

Ce qui implique,

$$\lambda^* = \left[R(X'X)^{-1}R'\right]^{-1}(R\hat{\beta} - q)$$

On substitue ce résultat dans l'expression pour β^* pour obtenir

$$\beta^* = \hat{\beta} - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}(R\hat{\beta} - q)$$

Interprétation:

Si la restriction est respectée alors

$$E(\beta^*) = E(\hat{\beta}) = \beta.$$

Sinon

$$E(\beta^*) \neq E(\hat{\beta}) = \beta \quad \Rightarrow \quad \text{biais pour} \quad \beta^*$$

La matrice de variance-covariance de β^* est égale à

$$Var(\beta^*) = \sigma^2(X'X)^{-1} - \sigma^2 \underbrace{(X'X)^{-1}R' \left[R(X'X)^{-1}R'\right]^{-1} R(X'X)^{-1}}_{\text{matrice semi-définie}}$$

$$\Rightarrow \quad var(\beta^*) < var(\hat{\beta})$$

Donc, si la contrainte est respectée, β^* est

1. un estimateur sans biais,

2. un estimateur plus précis (variance plus petite).

On peut maintenant construire un test pour les restrictions $R\beta = q$ basé sur la différence de la "performance" entre le modèle contraint et le modèle non contraint. Ce test est basé sur la somme des carrés des résidus. On a que

$$\begin{aligned}
\epsilon^* &= Y - X\beta^* \\
&= Y - X\hat{\beta} - X\beta^* + X\hat{\beta} \\
&= Y - X\hat{\beta} - X(\beta^* - \hat{\beta}) \\
&= \hat{\varepsilon} - X(\beta^* - \hat{\beta}) \\
\epsilon^{*'}\epsilon^* &= \hat{\epsilon}'\hat{\epsilon} + \underbrace{(\beta^* - \hat{\beta})'X'X(\beta^* - \hat{\beta})}_{\text{matrice non négative définie}} \geq \hat{\epsilon}'\hat{\epsilon}
\end{aligned}$$

La somme des carrés des résidus du modèle non contraint est plus petite que la somme des carrés des résidus pour le modèle contraint.

$$
\begin{aligned}
R^2 &= 1 - \frac{\hat{\epsilon}'\hat{\epsilon}}{Y'M_\iota Y} \quad \text{où} \quad M_\iota = I - \iota(\iota'\iota)^{-1}\iota' \\
R^{*2} &= 1 - \frac{\epsilon^{*'}\epsilon}{Y'M_\iota Y} \quad \leq R^2
\end{aligned}
$$

On a donc que

$$\epsilon^{*'}\epsilon^* - \hat{\epsilon}'\hat{\epsilon} = (\beta^* - \hat{\beta})'X'X(\beta^* - \hat{\beta})$$

et on sait que

$$\beta^* - \hat{\beta} = -(X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}(R\hat{\beta} - q).$$

On peut maintenant facilement montrer que l'expression $\epsilon^{*'}\epsilon^* - \hat{\epsilon}'\hat{\epsilon}$ est égale au numérateur de la statistique "F" divisé par J. Alors

$$\frac{(\epsilon^{*'}\epsilon^* - \hat{\epsilon}'\hat{\epsilon})/J}{\hat{\epsilon}'\hat{\epsilon}/N - K} \sim F(J, N - K)$$

(comparaison avec la statistique F).

Donc, la statistique F peut être considérée comme une statistique basée sur un test comparant le modèle contraint vs. non contraint.

Si on divise le numérateur et le dénominateur par $Y'M_\iota Y$, on obtient

$$\frac{(R^2 - R^{*2})/J}{(1 - R^2)/N - K} \sim F(J, N - K)$$

Cas particulier: Un test sur tous les coeficients sauf la constante, $\Rightarrow R^{*2} = 0$

$$\frac{R^2/J}{(1 - R^2)/N - K} \sim F(J, N - K)$$

En résumé, il y a deux approches pour effectuer un test sur des restrictions linéaires:

1. Modèle non contraint

$$F = (R\hat{\beta} - q)'\left[\hat{\sigma}^2 R(X'X)^{-1}R'\right]^{-1}(R\hat{\beta} - q)/J \quad \sim F(J, N - k)$$

2. Modèle contraint vs. non contraint

$$\frac{(\varepsilon^{*'}\varepsilon^* - \hat{\varepsilon}'\hat{\varepsilon})/J}{\hat{\varepsilon}'\hat{\varepsilon}/N - k} \sim F(J, N - k)$$

1.8 Tests de changement structurel

On cherche à évaluer si la relation entre la variable endogène et les variables exogènes est stable pour notre échantillon (test de Chow). Pour une date fixée, on a donc,

$$\begin{aligned} \underbrace{Y^1}_{n_1\times 1} &= \underbrace{X^1}_{n_1\times K} \quad \underbrace{\beta^1}_{K\times 1} + \underbrace{\epsilon^1}_{n_1\times 1} \\ \underbrace{Y^2}_{n_2\times 1} &= \underbrace{X^2}_{n_2\times K} \quad \underbrace{\beta^2}_{K\times 1} + \underbrace{\epsilon^2}_{n_2\times 1} \\ n_1 + n_2 &= N \end{aligned}$$

Modèle non contraint:

$$\begin{bmatrix} Y^1 \\ Y^2 \end{bmatrix} = \begin{bmatrix} X^1 & 0 \\ 0 & X^2 \end{bmatrix} \quad \begin{bmatrix} \beta^1 \\ \beta^2 \end{bmatrix} + \begin{bmatrix} \epsilon^1 \\ \epsilon^2 \end{bmatrix}$$

L'hypothèse nulle correspondant à l'absence de changement structurel est:

$$H_0 : \quad \beta^1 = \beta^2$$

Modèle contraint:

$$\underbrace{\begin{bmatrix} Y^1 \\ Y^2 \end{bmatrix}}_{N\times 1} = \underbrace{\begin{bmatrix} X^1 \\ X^2 \end{bmatrix}}_{N\times K} \quad \underbrace{\beta}_{K\times 1} + \underbrace{\begin{bmatrix} \epsilon^1 \\ \epsilon^2 \end{bmatrix}}_{N\times 1}$$

On peut effectuer un test basé sur les deux approches présentées précédemment:

1. Avec le modèle non contraint, l'hypothèse nulle s'écrit:

   $$R\hat{\beta} \Rightarrow \hat{\beta}^1 - \hat{\beta}^2 = 0.$$

   On construit la statistique F pour cette hypothèse.

2. Avec le modèle contraint vs. non contraint, on construit la statistique suivante:

   $$\frac{\epsilon^{*'}\epsilon^* - \hat{\epsilon}'\hat{\epsilon}}{\hat{\epsilon}'\hat{\epsilon}} \frac{J}{N-k} \quad \sim F(J, N-k)$$

et

$$\hat{\epsilon}'\hat{\epsilon} = \left[\hat{\epsilon}^{1'}\hat{\epsilon}^{2'}\right] \begin{bmatrix} \hat{\epsilon}^1 \\ \hat{\epsilon}^2 \end{bmatrix} = \hat{\epsilon}^{1'}\hat{\epsilon}^1 + \hat{\epsilon}^{2'}\hat{\epsilon}^2$$

$$\epsilon^{*'}\epsilon^* = \left[\epsilon^{*1'}\epsilon^{*2'}\right] \begin{bmatrix} \epsilon^{*1'} \\ \epsilon^{*2'} \end{bmatrix} = \epsilon^{*1'}\epsilon^{*1} + \epsilon^{*2'}\epsilon^{*2}$$

Cas particulier:

Examinons le cas particulier d'un changement structurel pour un sous-vecteur de paramètres. On a le modèle non contraint suivant:

$$\begin{aligned} Y^1 &= X_1^1\beta_1^1 + X_2^1\beta_2 + \epsilon^1 \\ Y^2 &= X_1^2\beta_1^2 + X_2^2\beta_2 + \epsilon^2. \end{aligned}$$

Le vecteur de paramètres β_1 peut varier pour les deux sous-échantillons

L'hypothèse nulle est:

$$H_0 : \quad \beta_1^1 = \beta_1^2.$$

Le modèle contraint est

$$\begin{aligned} Y^1 &= X_1^1\beta_1 + X_2^1\beta_2 + \epsilon^1 \\ Y^2 &= X_1^2\beta_1 + X_2^2\beta_2 + \epsilon^2 \end{aligned}$$

On peut effectuer un test du modèle contraint vs. non contraint comme nous l'avons vu précédemment.

Commentaires: Ici, la date du changement structurel a été fixé a priori. Habituellement, on ne connaît la date du changement structurel. On peut considérer une date inconnue. Ceci aura un impact sur la valeur critique (Andrews 1993).

1.8.1 Test de restrictions non linéaires

On va considérer des restrictions non linéaires entre les paramètres. L'hypothèse nulle s'écrit:

$$H_0 : \quad f(\beta) = q$$

où $f(\cdot)$ est une fonction non linéaire de dimension $J \times 1$. La statistique du test est donnée par l'expression suivante:

$$(f(\hat{\beta}) - q)' \left[Var f(\hat{\beta}) \right]^{-1} (f(\hat{\beta}) - q) \quad \sim \quad \chi^2(J)$$

On perd ici les propriétés de petit échantillon. Pour calculer la variance d'une fonction non linéaire, on fait alors une approximation par une expansion de Taylor

$$f(\hat{\beta}) = f(\beta) + \left(\frac{\delta f}{\delta \beta'} \right) (\hat{\beta} - \beta) + ...$$

$$Var(f(\hat{\beta})) = \underbrace{var(f(\beta))}_{fixe} + Var \left[\left(\frac{\delta f'}{\delta \beta} \right) (\hat{\beta} - \beta) \right]$$

$$var(f(\hat{\beta})) = \left(\frac{\delta f}{\delta \beta'} \right) var(\hat{\beta}) \left(\frac{\delta f}{\delta \beta'} \right)'$$

On peut donc construire la statistique présentée plus haut.

2 Théorie en grand échantillon et modèles à régresseurs aléatoires.

2.1 Convergence en probabilité et convergence en loi

Sous l'hypothèse que

1. X est une matrice fixe et de rang complet (K)
2. $E(\epsilon) = 0$, $E(\epsilon\epsilon') = \sigma^2 I$ et $\epsilon \sim N(0, \sigma^2 I)$

on a des propriétés de petit échantillon. En effet, on connaît la loi exacte des estimateurs et des tests (T, F).

Si on abandonne l'hypothèse que le vecteur ϵ suit une loi normale multivariée alors on ne connaît pas la loi des tests en petit échantillon. On doit alors faire appel à la théorie en grand échantillon. De plus, on va également examiner l'impact du relâchement de l'hypothèse que X est fixe.

2.2 Théorie en grand échantillon

(Johnston, 2.4.1 à 2.4.3, Greene 4.4)

On s'intéresse au comportement d'une variable aléatoire lorsque le nombre d'observations (N) tend vers l'infini.

Convergence en probabilité

Nous allons introduire et définir le concept de convergence en probabilité.

Prenons l'exemple suivant: On a une variable aléatoire $X_1, \cdots, X_N$ d'espérance μ et de variance σ^2 où les X_i ne sont pas corrélées, alors

$$E(\bar{X}_N) = \mu \quad \text{et} \quad VAR(\bar{X}_N) = \frac{\sigma^2}{N}$$

où

$$\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$$

On voit que la variance tend vers zéro lorsque N tend vers l'infini. Ce qui implique que la loi empirique de $\bar{X}_N$ est de plus en plus concentrée autour de μ lorsque N augmente.

Prenons maintenant un intervalle centré autour de μ, soit $\mu \pm \epsilon$. On va définir la probabilité que la variable $\bar{X}_N$ soit comprise dans cet intervalle. Ainsi,

$$Pr\{\mu - \epsilon < \bar{X}_N < \mu + \epsilon\} = Pr\{|\bar{X}_N - \mu| < \epsilon\}$$

Cette probabilité varie avec ϵ. Puisque la variance de $\bar{X}_N$ décroît de façon monotone lorsque N augmente, il existe un certain $N^\star$ et δ $(0 < \delta < 1)$ tels que un ϵ donné

$$Pr\{|\bar{X}_N - \mu| < \epsilon\} = 1 - \delta.$$

Lorsque $N \to \infty$, la probabilité que $\bar{X}_N$ appartienne à un intervalle bien précis devient plus élevée, et donc δ devient plus petit. On a alors,

$$\lim_{N\to\infty} Pr\{|\bar{X}_N - \mu| < \epsilon\} = 1$$

pour tout $\epsilon > 0$. Ce qui veut dire que la probabilité que $\bar{X}_N$ appartienne à un intervalle centré sur μ arbitrairement petit peut être rendu aussi voisine de 1 qu'on le désire, en prenant N suffisamment grand.

On réecrit la probabilité limite de la façon suivante.

$$plim\bar{X}_N = \mu.$$

Ce qui veut dire que la moyenne empirique est un estimateur convergent de l'espérance mathématique μ. De plus, on sait que la moyenne empirique est un estimateur sans biais peu importe la dimension de l'échantillon.

Prenons maintenant un autre estimateur m_N où

$$E(m_N) = \mu + \frac{c}{N}$$

où c est une constante quelconque. Cet estimateur m_N n'est pas sans biais en petit échantillon, cependant

$$\lim_{N\to\infty} E(m_N) = \mu$$

Alors m_N est asymptotiquement sans biais. Si la variance de m_N tend vers zéro alors M_N converge en probabilité vers μ. Donc, m_N est un estimateur convergent de μ.

Prenons un autre exemple: X_N prend les valeurs 0 et N avec des probabilités respectives de $(1 - \frac{1}{N})$ et $(\frac{1}{N})$. Alors $X_N \xrightarrow{p} 0$.

Definition 6 *On dit que X_N converge en probabilité vers une constante X si et seulement si*

$$\forall \epsilon > 0 \quad Pr[|X_N - X| > \epsilon] \to 0$$

lorsque N tend vers l'infini. On note $X_N \xrightarrow{p} X$.

Definition 7 *Une suite d'estimateur $\hat{\theta}_n$ est convergente vers θ si et seulement si*

$$plim(\hat{\theta}_n) = \theta$$

Une condition suffisante pour qu'un estimateur soit convergent en probabilité est qu'il soit asymptotiquement sans biais et que sa variance tend vers zéro. Ceci correspond à la convergence en moyenne quadratique. Donc, la convergence en moyenne quadratique est une condition suffisante pour avoir la convergence en probabilité.

Convergence en moyenne quadratique

Definition 8 (Convergence en moyenne quadratique) *Supposons que X_N a pour moyenne μ_N et variance σ_N^2 et que la limite de μ_N et σ_N^2 est c et 0 respectivement. On dira que X_n converge en moyenne quadratique vers c et*

$$PlimX_n = c$$

Preuve: voir Greene, p. 116.

Implication:

La convergence en moyenne quadratique implique la convergence en probabilité. Cependant, l'inverse n'est pas vrai.

Considérons l'exemple suivant:

$$\begin{aligned} X_N &= 0 \quad \text{avec une probabilté de} \quad 1 - \frac{1}{N} \\ &= N \quad \text{avec une probabilité de} \quad \frac{1}{N} \end{aligned}$$

L'espérance de X_N est égale à 1 pour $\forall N$. Mais ce n'est pas sa probabilité limite. De plus, la variance de X_N est égale à $(N-1)$.

Théorème 2 (Théorème de Slustsky) *Pour une fonction continue $g(X_N)$ qui ne dépend pas de N,*

$$plim\ g(X_N) = g(plimX_N)$$

Exemples:

$$\begin{aligned} plim(X_N)^2 &= (plimX_N)^2 \\ plim(X_N^{-1}) &= (plimX_N)^{-1} \\ plim\left(\frac{X_N}{Y_N}\right) &= \frac{plimX_N}{plimY_N} \end{aligned}$$

Pour les vecteurs, on a

$$plim AB = plim A \; plim B$$

et

$$plim(A^{-1}) = (plim A)^{-1}$$

Est-ce que l'estimateur des M.C.O converge en probabilité vers β?

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y \\ &= \beta + (X'X)^{-1}X'\epsilon = \beta + \left(\frac{1}{N}X'X\right)^{-1}\left(\frac{1}{N}X'\epsilon\right) \end{aligned}$$

On suppose que

$$\lim_{N\to\infty} \frac{1}{N}X'X = Q,$$

où Q est positive définie.

Alors

$$\lim_{N\to\infty} \left(\frac{X'X}{N}\right)^{-1} = Q^{-1}$$

$$plim\hat{\beta} = \beta + plim\left(\frac{1}{N}X'X\right)^{-1}\left(\frac{1}{N}X'\epsilon\right)$$

puisque

$$plim AB = plim A \; plim B$$

$$\begin{aligned} plim\hat{\beta} &= \beta + plim\left(\frac{1}{N}X'X\right)^{-1} plim\left(\frac{1}{N}X'\epsilon\right) \\ plim\hat{\beta} &= \beta + Q^{-1} plim\left(\frac{1}{N}X'\epsilon\right) \end{aligned}$$

On cherche la probabilité limite du dernier terme. Puisque X est fixe (non aléatoire)

$$\begin{aligned} E\left(\frac{1}{N}X'\epsilon\right) &= \frac{1}{N}X'E(\epsilon) = 0 \\ Var\left(\frac{1}{N}X'\epsilon\right) &= E\left(\frac{1}{N}X'\epsilon\epsilon'X\frac{1}{N}\right) \end{aligned}$$

$$\begin{aligned} &= \frac{1}{N} X' E(\epsilon\epsilon') X \frac{1}{N} \\ &= \frac{\sigma^2}{N} \left(\frac{X'X}{N} \right) \\ \lim_{N\to\infty} Var\left(\frac{1}{N} X'\epsilon\right) &= 0 \times Q = 0 \end{aligned}$$

Puisque $E\left(\frac{1}{N} X'\epsilon\right) = 0$ et que la variance tend vers zéro, ce terme converge en moyenne quadratique et donc on a la convergence en probabilité:

$$plim\left(\frac{1}{N} X'\epsilon\right) = 0.$$

On a donc

$$plim\hat{\beta} = \beta + Q^{-1} plim\left(\frac{1}{N} X'\epsilon\right)$$

où

$$plim\left(\frac{1}{N} X'\epsilon\right) = 0$$

donc

$$plim\hat{\beta} = \beta$$

Alors $\hat{\beta}$ est un estimateur convergent de β.

Convergence en loi

On utilise le même exemple. Puisque $Var(\bar{X}_N) \to 0$, on a un point de masse à μ, on a une loi dégénérée. On appelle $f(\bar{X}_N)$ la densité de la loi de $\bar{X}_N$ peu importe N. On va étudier ce que devient $f(\bar{X}_N)$ lorsque N tend vers l'infini. Une transformation de $\bar{X}_N$ permet d'obtenir un loi limite qui n'est pas dégénérée. Prenons,

$$Z_N = \sqrt{N}(\bar{X}_N - \mu)$$

On a que

$$E(Z_N) = 0 \quad \text{et} \quad Var(Z_N) = \sigma^2$$

Lorsqu'on connaît la loi de X_N, on peut pondérer $\bar{X}_N$ afin d'obtenir une loi qui n'est pas dégénérée.

On étudie la convergence en loi lorsque la loi en échantillon fini ne peut être obtenue. On peut alors considérer la loi limite comme une approximation de la loi inconnue en échantillon de taille finie.

Exemple: On suppose que

$$E(X) = \mu \text{ et } Var(X) = \sigma^2$$

mais on ne connaît pas la loi de X. X ne suit pas une loi normale.

Le théorème centrale limite nous dit que la loi limite de

$$Z_N = \sqrt{N}(\bar{X}_N - \mu) \text{ est une } N(0, \sigma^2).$$

On dira alors que $Z_N = \sqrt{N}(\bar{X}_N - \mu)$ converge en loi vers une $N(0, \sigma^2)$

On peut écrire également

$$Z_N = \sqrt{N}(\bar{X}_N - \mu) \overset{loi}{\rightarrow} N(0, \sigma^2)$$

Definition 9 *X_N converge en loi vers une variable aléatoire X avec une fonction de répartition $F(X)$ si*

$$\lim_{N \to \infty} |F(X_N) - F(X)| = 0$$

pour tout point de continuité de $F(X)$.

Remarque:

C'est un concept qui s'applique sur la loi de X_N et non sur X_N. Ainsi, on ne peut dire que X_N converge vers X. Examinons l'exemple suivant:

Exemple:

$$Prob(X_N = 1) = \frac{1}{2} + \frac{1}{N}$$

$$Prob(X_N = 2) = \frac{1}{2} - \frac{1}{N}$$

Lorsque $N \to \infty$, les deux probabilités convergent vers $\frac{1}{2}$, mais X_N ne converge pas vers une seule constante (donc pas de convergence en probabilité). La convergence en probabilité implique la convergence en loi et non l'inverse.

Théorème 3 (Théorème central limite) *Si $X_1, ..., X_N$ est une suite de variables aléatoires avec une certaine densité, une moyenne bornée $E(\bar{X}_N < \infty)$ et une variance finie σ^2, alors*

$$\sqrt{N}(\bar{X}_N - \mu) \overset{loi}{\to} N(0, \sigma^2)$$

Normalité asymptotique de l'estimateur des M.C.O.

Greene 6.7.3

On a que

$$\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{X'X}{N}\right)^{-1} \frac{X'\epsilon}{\sqrt{N}}$$

On va étudier la loi limite de $\frac{1}{\sqrt{N}}X'\epsilon$.

On a les résultats suivant:

$$\begin{aligned} Var\left(\frac{X'\epsilon}{\sqrt{N}}\right) &= \sigma^2 \frac{X'X}{N} \\ \lim_{N\to\infty} \sigma^2 \left(\frac{X'X}{N}\right) &= \sigma^2 Q \quad \text{et} \quad Q < \infty, \text{bornée,} \end{aligned}$$

on peut appliquer théorème central limite à l'expression $\frac{1}{\sqrt{N}}X'\epsilon$.

On applique donc le théorème central limite, ainsi,

$$\frac{1}{\sqrt{N}}X'\epsilon \overset{loi}{\to} N(0, \sigma^2 Q).$$

$$\sqrt{N}(\hat{\beta} - \beta) = \frac{X'X}{N}^{-1} \frac{X'\varepsilon}{\sqrt{N}} \overset{loi}{\to} N(0, Q^{-1}\sigma^2 Q Q^{-1})$$

Alors,

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{loi} N(0, \sigma^2 Q^{-1}).$$

On peut construire les statistiques de tests t et F

$$t \overset{loi}{\to} N(0,1)$$

$$F \overset{loi}{\to} \frac{X^2(J)}{J}$$

On espère que la loi limite soit une bonne approximation de la loi de notre échantillon.

3 Estimation par maximum de vraisemblance

Johnston-Dinardi, 5.1-5.4

On va étudier les implications de l'hypothèse de normalité sur les propriétés asymptotiques de l'estimateur des M.C.O.

Definition 10 *Un estimateur est asymptotiquement optimal si il est convergent, si il suit une loi limite normale et si sa matrice de variance-covariance est plus petite que tout autre estimateur convergent ayant pour loi limite une normale. (pendant en grand échantillon de Gauss-Markov)*

Cette définition est le pendant en grand échantillon de Gauss-Markov.

On va montrer que l'estimateur du maximum de vraisemblance est asymptotiquement optimal et on va le comparer à l'estimateur des M.C.O.

3.1 Présentation de l'estimateur du maximum de vraisemblance

On a une suite de variables aléatoires et on veut savoir quelle est la densité (ou la fonction de répartition) qui a pu générer cette suite. Donc, quelle densité paramétrique a pu produire la suite observée.

$$f(X_N, \theta).$$

Exemple:

$$\begin{aligned} X &= 0 \quad \text{avec une probabilité égale à} \quad p \\ X &= 1 \quad \text{avec une probabilité égale à} \quad 1-p \end{aligned}$$

On veut connaître le paramètre p. On observe la suite de réalisations suivantes

$$X_N = \{1\ 0\ 0\ 0\ 1\ 0\ 0\ 1\ 0\ 0\}$$

On va chercher p qui maximise la probabilité d'observer un tel échantillon. On aura la densité conjointe suivante:

$$L(X,p) = (1-p).\ p.\ p.\ p.\ (1-p)\ .p\ .p\ .(1-p)\ .p\ .p$$

où $L(\cdot)$ est une Bernouilli. On peut écrire la densité conjointe de cette façon puisque ce sont des événements indépendants. On réécrit la densité conjointe,

$$L(X,p) = p^{N_1}(1-p)^{N_2}$$

où N_1 est la nombre de fois que 0 est observée et N_2 est le nombre de fois que la valeur 1 est observée.

On cherche donc la valeur de p qui rend le plus probable cet échantillon. On va maximiser $L(X,p)$ par rapport à p, ce qui revient à

$$\max_p\ lnL(X,P)$$

$$\max_p\ N_1\ ln\ p + N_2\ ln\ (1-p)$$

C.P.O

$$\frac{\partial lnL}{\partial p} : \frac{N_1}{p} - \frac{N_2}{(1-p)} = 0$$

$$\Rightarrow \hat{p} = \frac{N_1}{N_1 + N_2} = 0.70.$$

Examinons maintenant le modéle linéaire classique,

$$Y = X\beta + \epsilon \quad \text{et} \quad \epsilon \sim N(0, \sigma^2).$$

En connaissant la densité du vecteur Y, on peut chercher le vecteur de paramètres β qui a le plus vraisemblablement engendré les observations y conditionnellement aux observations X.

On doit connaître la densité de Y. Puisque X est fixe et que le modèle est linéaire, la densité de Y est directement fonction de la densité de ϵ.

Quelle est la densité de Y? On va utiliser le résultat suivant: si ϵ suit une loi normale à plusieurs dimensions, il en est de même pour Y et, en particulier

$$f(Y) = f(\epsilon)\left|\frac{\partial \epsilon}{\partial Y'}\right|$$

où $\left|\frac{\partial \epsilon}{\partial Y'}\right|$ est la valeur absolue du déterminant de la matrice des dérivées partielles et $\frac{\partial \epsilon}{\partial Y'}$.

$$\begin{vmatrix} \frac{\partial \epsilon_1}{\partial y_1} & \frac{\partial \epsilon_1}{\partial y_2} & . & . & . & \frac{\partial \epsilon_1}{\partial y_N} \\ . & & & & & \\ . & & & & & \\ . & & & & & \\ \frac{\partial \epsilon_N}{\partial y_1} & \frac{\partial \epsilon_N}{\partial y_2} & . & . & . & \frac{\partial \epsilon_N}{\partial y_N} \end{vmatrix} = 1$$

On a donc $f(y) = f(\epsilon)$.

On a supposé que $\epsilon \sim N(0, \sigma^2 I)$. La densité de ϵ_n est donnée par la densité de la loi normale suivante:

$$f(\epsilon_n) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(y_n - x_n'\beta)^2\right).$$

Puisque les ϵ_t sont indépendants, alors

$$f(\epsilon) = \prod_{n=1}^{N} f(\epsilon_n).$$

De façon matricielle, on aura

$$f(\epsilon) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2}\epsilon'\epsilon\right).$$

La vraisemblance de Y est donc

$$L(\theta; Y, X) = f(Y) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta)\right)$$

où $\theta = (\beta', \sigma^2)'$.

Il est équivalent de maximiser la vraisemblance que de maximiser le logarithme de la vraisemblance. Le problème de maximisation à résoudre est donc le suivant:

$$\max_{\theta} \ln \; L(\theta; Y, X) = -\frac{N}{2} \ln 2\pi - \frac{N}{2} \ln \sigma^2 - \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta).$$

Les C.P.O sont données par les équations suivantes:

$$\begin{aligned} \frac{\partial lnL}{\partial \beta} &= -\frac{1}{2\sigma^2}(-2X'Y + 2X'X\beta) = \frac{1}{\sigma^2}(X'Y - X'X\beta) = 0 \\ \frac{\partial lnL}{\partial \sigma^2} &= \frac{-N}{2\sigma^2} + \frac{1}{2\sigma^4}(Y - X\beta)'(Y - X\beta) = 0 \end{aligned}$$

On obtient donc

$$\hat{\beta} = (X'X)^{-1}X'Y$$

et

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{N}.$$

Ainsi,

$$\hat{\beta}_{m.v.} = \hat{\beta}_{m.c.o.} \quad \text{mais} \quad \hat{\sigma}^2_{m.v.} \neq \hat{\sigma}^2_{m.c.o.}.$$

Ce qui implique que $\hat{\sigma}^2_{M.V.}$ est un estimateur biaisé de σ^2.

3.2 Propriétés de l'estimateur du M.V.

On examine les propriétés

1. à distance finie (petit échantillon)

2. asymptotique

3.2.1 Propriétés en petit échantillon

S'il existe un estimateur dont la variance est égale à la borne inférieure de la variance, alors il est donné par la méthode du maximum de vraisemblance.
La borne inférieure de la variance est donnée par le théorème de Cramer-Rao. C'est un seuil inférieur pour n'importe quel estimateur sans biais (pas seulement linéaire).

On ne peut pas toujours atteindre la borne de Cramer-Rao pour un estimateur sans biais.

Théorème 4 (Le théorème de Cramer-Rao) *La matrice suivante:*

$$Var(\hat{\theta}) - I^{-1}(\theta)$$

est une matrice semi-définie positive où

$$I(\theta) = -E\left(\frac{\partial^2\ lnL}{\partial\theta\partial\theta'}\right) = E\left[\frac{\partial lnL}{\partial\theta}\frac{\partial \ln L}{\partial\theta'}\right]$$

qui est une matrice positive définie. La matrice $I(\theta)$ est appelée matrice d'information de Fisher.

La matrice $I(\theta)$ est donc définie comme étant:

$$I(\theta) = -E\begin{bmatrix} \frac{\partial^2 lnL}{\partial\theta_1^2} & \frac{\partial^2 lnL}{\partial\theta_1\partial\theta_2} & \cdot & \cdot & \cdot & \frac{\partial^2 lnL}{\partial\theta_1\partial\theta_K} \\ \frac{\partial^2 lnL}{\partial\theta_2\partial\theta_1} & \frac{\partial^2 lnL}{\partial\theta_2^2} & \cdot & \cdot & \cdot & \frac{\partial^2 lnL}{\partial\theta_2\partial\theta_K} \\ \cdot & \cdot & & & & \cdot \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & & \cdot & \cdot \\ \frac{\partial^2 lnL}{\partial\theta_K\partial\theta_1} & \cdot & \cdot & \cdot & \cdot & \frac{\partial^2 lnL}{\partial\theta_K^2} \end{bmatrix}$$

Pour l'estimateur du maximum de vraisemblance, on a que

$$VAR(\hat{\theta}_{m.v.}) = I^{-1}(\theta).$$

Examinons ceci pour le modèle linéraire: $Y = X\beta + \epsilon$

Les dérivées secondes sont

$$\begin{aligned}
\frac{\partial^2 \ln L}{\partial\beta\partial\beta'} &= -\frac{1}{\sigma^2}X'X \\
\frac{\partial^2\ lnL}{\partial^2(\sigma^2)} &= \frac{N}{2\sigma^4} - \frac{(Y-X\beta)'(Y-X\beta)}{\sigma^6} \\
\frac{\partial^2\ lnL}{\partial\beta\partial\sigma^2} &= \frac{1}{\sigma^4}(X'Y - X'X\beta) = \frac{1}{\sigma^4}X'\epsilon
\end{aligned}$$

On prend l'espérance de chaque expression

$$\begin{aligned}
E\frac{\partial^2 \ln L}{\partial\beta\partial\beta'} &= -\frac{1}{\sigma^2}X'X \\
E\frac{\partial^2\ lnL}{\partial^2(\sigma^2)} &= \frac{N}{2\sigma^4} - N\frac{\sigma^2}{\sigma^6} = -\frac{N}{2\sigma^4} \\
E\frac{\partial^2\ lnL}{\partial\beta\partial\sigma^2} &= -\frac{1}{\sigma^4}E(X'\epsilon) = 0
\end{aligned}$$

La matrice de variance-covariance du maximum de vraisemblance est donc

$$I(\theta)^{-1} = \begin{bmatrix} \frac{1}{\sigma^2}(X'X) & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix}^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{bmatrix}.$$

Pour le modèle linéaire classique, la borne de Cramer-Rao est donc,

$$I(\theta)^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{bmatrix}$$

L'estimateur $\hat{\beta}_{m.c.o.}$ atteint la borne de Cramer-Rao puisque sa variance est la même que l'estimateur du maximum de vraisemblance. Qu'en est-il de $\sigma^2_{m.c.o.}$? Cet estimateur est donné par l'expression suivante:

$$\hat{\sigma}^2_{M.C.O} = \frac{\hat{\epsilon}'\hat{\epsilon}}{N-K} = \frac{\epsilon' M \epsilon}{N-K}.$$

Examinons cet estimateur à l'aide du résultat suivant: si $X \sim N(0, \sigma^2)$ et A est une matrice idempotente de rang r, alors $\frac{1}{\sigma^2}X'AX \sim \chi^2(r)$.

On a donc,

$$\frac{1}{\sigma^2}\epsilon' M\epsilon \sim \chi^2(N-K)$$
$$\Rightarrow \frac{1}{\sigma^2}\frac{\varepsilon' M\varepsilon}{N-K} \sim \frac{\chi^2(N-K)}{N-K}$$
$$\Rightarrow \hat{\sigma}^2 \sim \frac{\sigma^2}{(N-K)}\chi^2(N-K).$$

On sait que la variance d'une khi-deux centrée est égale à deux fois le nombre de degrés de liberté. Alors,

$$Var(\hat{\sigma}^2) = \frac{\sigma^4}{(N-K)^2}2(N-K) = \frac{2\sigma^4}{N-K} > \frac{2\sigma^4}{N}$$

-Donc l'estimateur de σ^2 des moindres carrés ordinaires n'atteint pas la borne de Cramer-Rao.

-En petit échantillon, aucun estimateur sans biais ne peut atteindre la borne de Cramer-Rao.

En résumé, pour des échantillon finis;

-$\hat{\sigma}^2_{M.V}$ atteint la borne de Cramer-Rao mais il n'est pas sans biais.

-$\hat{\sigma}^2_{M.C.O}$ est sans biais, mais il n'atteint pas la borne de Cramer-Rao.

Comment peut-on obtenir un estimateur de la matrice de variance-covariance de l'estimateur du maximum de vraisemblance de θ. On a l'égalité suivante $Var(\hat{\theta}_{M.V}) = I(\hat{\theta})^{-1}$ où

$$I(\theta)^{-1} = \left[-E\left[\frac{\partial^2 lnL(\theta)}{\partial\theta\partial\theta'}\right]\right]^{-1}.$$

On obtient un estimateur de cette matrice en l'évaluant à l'estimateur du maximum de vraisemblance. Ainsi,

$$\hat{I}(\hat{\theta})^{-1} = \left(\frac{-\partial^2 lnL(\hat{\theta})}{\partial\hat{\theta}\partial\hat{\theta}'}\right)^{-1}.$$

Cependant, $I(\hat{\theta})^{-1}$ est souvent compliqué à obtenir. Plus simplement, on peut calculer

$$I(\hat{\theta})^{-1} = \left[\sum_{n=1}^{N} \hat{g}_n \hat{g}_n'\right]^{-1}$$

où

$$\hat{g}_n = \frac{\partial ln f(x_i, \hat{\theta})}{\partial \hat{\theta}}$$

Propriété 3 (Propriété d'invariance) *L'estimateur du maximum de vraisemblance de $g(\theta)$ est $g(\hat{\theta})$ ou $\hat{\theta}$ est l'estimateur du maximum de vraisemblance de θ.*

Remarques:

- En connaissant $\hat{\theta}$, on obtient $g(\hat{\theta}) \Rightarrow g$ doit être une fonction continue.
- On peut changer la paramétrisation de la fonction de vraisemblance pour simplifier l'estimation.

3.2.2 Propriétés de grand échantillon de l'estimateur de maximum de vraisemblance

Sous certainces conditions de régularité, l'estimateur du maximum de vraisemblance est convergent, asymptotiquement normal et asymptotiquement optimal (même pour σ^2). En effet, l'estimateur du maximum de vraisemblance de σ^2 est donné par

$$\hat{\sigma}^2_{m.v.} = \frac{\hat{\epsilon}'\hat{\epsilon}}{N}$$

et donc

$$E(\hat{\sigma}^2_{m.v.}) = \frac{N-K}{N}\sigma^2.$$

Lorsque N tend vers l'infini, cette expression tend vers σ. Puisque que sa variance tend vers zéro, on a convergence en moyenne quadratique. On sait que la convergence en moyenne quadratique implique la convergence en probabilité. Alors,

$$plim\hat{\theta} = \theta$$

De plus, en employant le théorème central limite, on peut montrer que

$$\hat{\theta} \sim AN(\theta, I(\theta)^{-1})$$

où $I(\theta)^{-1}$ est la borne de Cramer-Rao.

3.3 Non-normalité des erreurs et régresseurs aléatoires

1. X est fixe (non aléatoire) et de rang complet.
2. $E(\varepsilon) = 0$, $VAR(\varepsilon) = \sigma^2 I$ mais on ne suppose pas que $\varepsilon \sim N(0, \sigma^2 I)$

On a obtenu que l'estimateur des M.C.O. de β est le meilleur estimateur linéaire sans biais.

Questions: Si on n'a pas la normalité, qu'est-ce qui arrive avec les tests?
Réponses: On aura une justification en grand échantillon.

On va examiner la loi asymptotique de $\hat{\beta}_{m.c.o.}$ lorsqu'on ne suppose pas que les termes d'erreurs suivent une loi normale multivariée. On considère que X est fixe.

On a l'hypothèse que

$$\lim_{n\to\infty} \frac{1}{N} X'X = Q < \infty.$$

On a également montré que

$$\frac{1}{\sqrt{N}} X'\epsilon \sim AN(0, \sigma^2 Q).$$

Ceci implique le résultat suivant:

$$\sqrt{N}(\hat{\beta} - \beta) \stackrel{loi}{\rightarrow} N(0, \sigma^2(\frac{X'X}{N})^{-1})$$

On peut donc établir la loi asymptotique de l'estimateur des moindres carrés ordinaires de β. Cependant, on ne peut rien dire sur la loi en petit échantillon. On doit espérer que la loi asymptotique est une bonne approximation de la loi en petit échantillon.

Examinons le cas où la matrice des variables explicatives X est aléatoires.
Exemple: le revenu des ménages varie selon l'échantillon.
On va adopter la stratégie suivante:

1. On cherche les propriétés des estimateurs conditionnels à X.
2. On cherche ensuite les propriétés marginales par moyennage de la loi conditionnelle.

On doit faire deux hypothèses importantes:

1. $g(X)$ (densité de X) ne dépend pas de β et σ^2
2. X et ϵ sont indépendants.

Dans un contexte de série temporelle, la deuxième hypothèse implique ϵ_t est indépendant des valeurs passées, présente et futures. Il est important de souligner que l'hypothèse 2 peut souvent être violé en pratique.
Exemple:

$$y_t = \alpha + \beta y_{t-1} + \epsilon_t$$

alors y_t est correlée avec ϵ_{t-1}.

Si les deux hypothèses sont respectées, alors

$$\begin{aligned} f(\epsilon/X) &= f(\epsilon) \\ E(\epsilon/X) &= E(\epsilon) \\ E(Y/X) &= E((X\beta + \epsilon)/X) = X\beta + E(\epsilon) = X\beta \\ E(\epsilon'\epsilon/X) &= E(\epsilon'\epsilon) \end{aligned}$$

L'espérance marginale de Y est

$$E(Y) = E(X\beta) + E(\varepsilon) = E(X)\beta.$$

Écrivons maintenant la densité conjointe de nos observations X et Y. La densité conjointe est donnée par:

$$\begin{aligned} f(Y, X; \beta\ \sigma^2) &= f(Y/X; \beta\ \sigma^2)\ g(X) \\ f(Y, X; \beta\ \sigma^2) &= f(\epsilon/X; \beta\ \sigma^2)\ g(X) \\ f(Y, X; \beta\ \sigma^2) &= f(\epsilon; \beta\ \sigma^2)\ g(X) \end{aligned}$$

On ajoute l'hypothèse de normalité pour les erreurs

$$\varepsilon \sim N(0, \sigma^2 I).$$

On peut maintenant estimer par maximum de vraisemblance. Le log de la vraisemblance est:

$$\ln L = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) + \ln\ g(X)$$

Puisque $g(X)$ ne dépend pas de β et σ^2, on obtient les mêmes C.P.O.. Cependant la matrice $I(\theta)^{-1}$ sera

$$I(\theta)^{-1} = \begin{bmatrix} \sigma^2 E(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{bmatrix}$$

Est-ce que $\hat{\beta}$ (marginal) est encore sans biais ? Pour répondre à cette question, on va introduire le résultat suivant. Supposons une densité conjointe de deux variables aléatoires $f(W, Z)$ et $g(W, Z)$ une fonction de ces deux variables. On va chercher à évaluer $E\left(g(W, Z)\right)$.

$$\begin{aligned} Eg(W, Z) &= \int_Z \int_W g(W, Z) f(W, Z) dW\ dZ \\ Eg(W, Z) &= \int_Z \int_W g(W, Z) f(W/Z) f(Z) dW\ dZ \\ Eg(W, Z) &= \int_Z \left[\int_W g(W, Z) f(W/Z) dW\right] f(Z) dZ \\ Eg(W, Z) &= E_Z E_{W/Z}(g(W, Z)) \end{aligned}$$

Appliquons maintenant ce résultat. On a que

$$\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon$$

$$E\hat{\beta} = \beta + E\left[(X'X)^{-1}X'\epsilon\right]$$
$$E\hat{\beta} = \beta + E_X E_{\epsilon/X}\left[(X'X)^{-1}X\epsilon\right]$$
$$E\hat{\beta} = \beta + E_X\left[(X'X)^{-1}XE(\epsilon/X)\right].$$

Ceci implique que $E\hat{\beta} = \beta$ puisque $E(\epsilon/X) = 0$. $\hat{\beta}$ est toujours un estimateurs sans biais de β.

Pour la matrice de variance-covariance de $\hat{\beta}$, on applique le même résultat. On a

$$\begin{aligned}
Var(\hat{\beta}) &= E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \\
Var(\hat{\beta}) &= E\left[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}\right] \\
Var(\hat{\beta}) &= E_X[E_{\epsilon/X}(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}] \\
Var(\hat{\beta}) &= E_X[\sigma^2(X'X)^{-1}] \\
Var(\hat{\beta}) &= \sigma^2 E(X'X)^{-1}
\end{aligned}$$

qui atteint la borne de Cramer-Rao.

En résumé, si X est aléatoire et indépendant de ϵ

- En supposant la normalité des erreurs ϵ, l'estimateur des M.C.O. est le même que l'estimateur maximum du vraisemblance et c'est le même que dans le cas où X est fixe.
- En échantillon fini, l'estimateur M.C.O.(M.V) est sans biais et il atteint la borne de Cramer-Rao.

- L'estimateur habituel $\hat{\sigma}^2(X'X)^{-1}$ est un estimateur sans biais de cette borne.
- Les test T et F préservent les propriétes de petit échantillon.

Examinons le cas où la matrice X est correlée avec ϵ. On a alors

$$E(\epsilon/X) \neq 0$$

Conséquence:

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'Y \\
\hat{\beta} &= (X'X)^{-1}X'(X\beta+\epsilon) \\
\hat{\beta} &= \beta + (X'X)^{-1}X'\epsilon \\
E(\hat{\beta}) &= \beta + E\left[(X'X)^{-1}X'\epsilon\right] \neq \beta
\end{aligned}$$

puisque $E(\epsilon/X) \neq 0$, alors l'estimateur des M.C.O. n'est pas sans biais. On aura le même résultat asymptotique si la corrélation entre X et ϵ ne disparaît pas lorsque $N \to \infty$.

Solution: estimateur à variables instrumentales

Une variable instrumentale est une variable qui n'est pas corrélée avec le terme d'erreur et qui est corrélée avec les variables explicatives.

On a le même modèle:

$$Y = X\beta + \epsilon$$

mais

$$E(\epsilon/X) \neq 0$$

On va utiliser une matrice de variables instrumentales telle que

$$E(Z'\epsilon) = 0 \quad \text{et} \quad E(Z'X) \neq 0$$

Supposons que Z a la même dimension que la matrice X. On prémultiplie le modèle par Z'. On a donc

$$Z'Y = Z'X + Z'\epsilon$$

On suppose de plus que

$$\begin{aligned} plim\frac{1}{N}Z'Z &= Q_{ZZ} < \infty, \quad \text{définie positive} \\ plim\frac{1}{N}Z'X &= Q_{ZX} < \infty, \quad \text{définie positive} \\ plim\frac{1}{N}Z'\epsilon &= Q_{Z\epsilon} = 0; \end{aligned}$$

On aura

$$plim\left(\frac{1}{N}Z'Y\right) = plim\left(\frac{1}{N}Z'X\right)\beta + plim\left(\frac{1}{N}Z'\varepsilon\right)$$

alors,

$$plim\hat{\beta} = \left[plim\left(\frac{1}{N}Z'X\right)\right]^{-1} plim\left(\frac{1}{N}Z'Y\right)$$

L'estimateur à variables instrumentales sera donc

$$\hat{\beta}_{v.i.} = (Z'X)^{-1}Z'Y.$$

On peut montrer que cet estimateur converge en probabilité.
On cherche maintenant à obtenir sa loi asymptotique. On procède de la même façon que pour les M.C.O. On travaille avec $\frac{1}{\sqrt{N}}(Z'\epsilon)$ au lieu de $\frac{1}{\sqrt{N}}(X'\epsilon)$.
Par le théorème central limite,

$$\left(\frac{1}{\sqrt{N}}(Z'\epsilon)\right) \quad \xrightarrow{loi} N(0, \sigma^2 Q_{ZZ})$$

alors

$$\sqrt{N}(\beta_{v.i.} - \beta) = \frac{1}{N}(Z'X)^{-1}\frac{1}{\sqrt{N}}(Z'\epsilon)$$

On aura

$$\beta_{v.i.} \quad \xrightarrow{loi} N(\beta, \sigma^2 Q_{ZX}^{-1} Q_{ZZ} Q_{XZ}^{-1})$$

La matrice Z peut contenir plus de variables instrumentales que de variables explicatives dans la matrice X. On va supposer que la matrice Z est de

dimension $N \times L$ où $L > k$. On pourrait choisir k combinaisons linéaires des L variables instrumentales. Le meilleur choix sera la projection de la matrice X sur l'espace engendré par les Z. On aura ainsi

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

et l'estimateur à variables instrumentales sera

$$\begin{aligned} b_{v.i.} &= (\hat{X}'X)^{-1}\hat{X}'Y \\ &= [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'Y \end{aligned}$$

3.4 Tests d'hypothèses

Greene 4.9
La trilogie des tests;

1. Wald
2. Rapport de vraisemblance
3. Multiplicateur de Lagrange

On a vu que si $X \sim N(\mu, \Sigma)$, alors

$$(X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi^2(J)$$

On considère un estimateur du maximum de vraissemblance $\hat{\theta}$. Il existe trois approches pour effectuer des tests sur ce vecteur.

On va supposer l'hypothèse nulle suivante:

$$H_0 : \ C(\theta) = q, \ \text{ où } \ dim(q) = J$$

1. **Test de type Wald**

Cette statistique de test est basée sur l'estimateur non contraint. La staistique de Wald est:

$$W = \left(C(\hat{\theta}) - q\right)' \left[Var\left(C(\hat{\theta}) - q\right)\right]^{-1} \left(C(\hat{\theta}) - q\right) \overset{loi}{\to} \chi^2(J)$$

et

$$Var\left(C(\hat{\theta}) - q\right) \simeq \left(\frac{\partial C(\hat{\theta})}{\partial \theta'}\right) Var(\hat{\theta}) \left(\frac{\partial C(\hat{\theta})}{\partial \theta'}\right)'$$

On estime donc le modèle non contraint et on construit la statistique W.

2. **Test du multiplicateur Lagrange**

Cette statistique est calculée à partir de l'estimation contrainte. On estime le modèle contraint et on fait un test sur la dérivée par rapport aux paramètres.

On a les C.P.O. suivantes:

$$\frac{\partial lnL(\theta)}{\partial \theta^c} + \left(\frac{\partial C}{\partial \theta'}\right)' \lambda = 0.$$

Si la contrainte n'est pas mordante, alors

$$\frac{\partial lnL(\hat{\theta}^c)}{\partial \theta^c} \approx 0 \quad \text{et} \quad \hat{\lambda} \approx 0$$

On peut montrer que sous l'hypothèse nulle que

$$\frac{1}{\sqrt{N}} \frac{\partial lnL(\hat{\theta}_c)}{\partial \theta^c} \overset{loi}{\to} N\left[0, \frac{1}{N} I(\theta)\right]$$

en appliquant le théorème central limite.

La statistique du multiplicateur de Lagrange est définie comme étant:

$$LM = \left(\frac{\partial lnL(\hat{\theta}_c)}{\partial \theta^c}\right) \left[I(\hat{\theta}_c)\right]^{-1} \left(\frac{\partial lnL(\hat{\theta}_c)}{\partial \theta^c}\right) \overset{loi}{\to} \chi^2(J).$$

La statistique du mul;tiplicatuer de Lagrange est également appelée statistique du *score*.

3. **Test du ratio de vraisemblance**

$$-2(\ln L(\hat{\theta}_c) - \ln L(\hat{\theta}) \quad \stackrel{loi}{\rightarrow} \chi^2(J)$$

Ces trois tests sont asymptotiquement équivalents, mais ils peuvent se comporter différemment à distance finie (petit échantillon).

3.4.1 Les tests Wald, LR, et LM pour le modèle linéaire classique

On a donc le modèle suivant:

$$Y = X\beta + \epsilon \qquad \epsilon \sim N(0, \sigma^2 I),$$

et l'hypothèse nulle suivante:

$$H_0 : \quad R\beta = q$$

On a un estimateur non contraint $\hat{\beta}_{nc}$ et un estimateur contraint $\hat{\beta}_c$. On va montrer qu'on a le préordre suivant en petit échantillon;

$$LM \leq LR \leq WALD.$$

On définit

$$\begin{aligned} \hat{\sigma}^2_{nc} &= \frac{1}{N}(Y - X\hat{\beta}_{nc})'(Y - X\hat{\beta}_{nc}) \\ \hat{\sigma}^2_{c} &= \frac{1}{N}(Y - X\hat{\beta}_{c})'(Y - X\hat{\beta}_{c}) \end{aligned}$$

1. Test de Wald

$$W = (R\hat{\beta} - q)' \left[\hat{\sigma}^2_{nc} R(X'X)^{-1}R'\right]^{-1} (R\hat{\beta} - q)$$

puisque

$$\hat{\sigma}^2_c - \hat{\sigma}^2_{nc} = \frac{1}{N}(X\hat{\beta}_c - X\hat{\beta}_{nc})'(X\hat{\beta}_c - X\hat{\beta}_{nc})$$

et par la relation

$$\hat{\beta}_c = \hat{\beta}_{nc} - (X'X)^{-1}R' \left[R(X'X)^{-1}R'\right]^{-1} (R\hat{\beta} - q)$$

Alors

$$W = N\left(\frac{\hat{\sigma}_c^2 - \hat{\sigma}_{nc}^2}{\hat{\sigma}_{nc}^2}\right)$$

2. Test du multiplicateur de $\mathcal{L}$agrange (Score)

$$LM = \hat{\sigma}_c^2 \hat{\lambda}' R(X'X)^{-1}R'\hat{\lambda}$$

et puisque

$$\hat{\lambda} = -\frac{1}{\hat{\sigma_c^2}} \left[R(X'X)^{-1}R'\right]^{-1} (R\hat{\beta} - q).$$

Alors, la statistique LM peut être réécrite comme étant

$$LM = \frac{1}{\hat{\sigma_c^2}}(R\hat{\beta} - q)' \left[R(X'X)^{-1}R'\right]^{-1} (R\hat{\beta} - q).$$

Ce qui donne,

$$LM = N\left(\frac{\hat{\sigma}_c^2 - \hat{\sigma}_{nc}^2}{\hat{\sigma}_c^2}\right)$$

3. Test du ratio de vraissemblance

$$-2\left[\ln L^c - \ln L^{nc}\right] = N \ln \frac{\hat{\sigma}_c^2}{\hat{\sigma}_{nc}^2}$$

Comparaison des trois statistiques de test:

On va utiliser la relation suivante:

$$\frac{X}{1+X} \leq \log(1+X) \leq X \quad \forall X > -1$$

On pose que

$$X = \frac{\hat{\sigma}_c^2 - \hat{\sigma}_{nc}^2}{\hat{\sigma}_{nc}^2}$$

Ce qui implique que

$$LM \leq LR \leq Wald$$

Ce préordre est vrai en petit échantillon. De façon asymptotique, les trois statistiques sont équivalentes.

4 MOINDRES CARRÉS GÉNÉRALISÉS, M.C.G

On a toujours le même modèle

$$Y = X\beta + \epsilon$$

On avait fait l'hypothèse suivante pour la matrice de variance-covariance des termes d'erreurs,

$$E(\epsilon\epsilon') = \sigma^2 I$$

On dit que les erreurs sont sphériques.

Implications:

1. La variance est constante (homoscédasticité).
2. Les covariances sont nulles.

On va maintenant relâcher ces deux hypothèses.

Hypothèse générale:

On considère le cas général où la matrice de variance-covariance est donnée par

$$E(\epsilon\epsilon') = \sigma^2\Omega$$

Hétéroscédasticité:

En présence d'hétéroscédasticité la matrice de variance-covariance des termes d'erreurs aura la forme générale suivante:

$$\sigma^2\Omega = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & & 0 & \sigma_n^2 \end{bmatrix}$$

La variance est différente selon les observations. On retrouve l'hétéroscédasticité surtout en microéconomie et en série temporelle.

Autocorrélation (séries temporelles)

$$\sigma^2\Omega = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{N-1} \\ \rho_1 & 1 & & & \\ \rho_2 & & 1 & & \\ \vdots & & & \ddots & \\ \rho_{N-1} & & & & 1 \end{bmatrix}$$

où

$$E(\epsilon_n \epsilon_{n-1}) = \rho_1 \neq 0$$

Il y a donc un lien entre les termes d'erreurs pour différentes observations.

4.1 Comportement de l'estimateur des M.C.O à distance finie

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y \\ E(\hat{\beta}) &= E_X E(\hat{\beta}/X) \\ &= \beta + E_X E\left[(X'X)^{-1}X'\epsilon)/X\right] = \beta. \end{aligned}$$

L'estimateur des M.C.O. est donc sans biais.

Cherchons maintenant sa variance,

$$\begin{aligned} Var(\hat{\beta}) &= E(\hat{\beta}-\beta)(\hat{\beta}-\beta)' \\ &= E\left[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}\right] \\ &= E_X E_{\epsilon/X}\left[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}\right] \\ &= \sigma^2 E_X (X'X)^{-1}X'\Omega X(X'X)^{-1} \end{aligned}$$

Si $\epsilon \sim N(0,\sigma^2)$, $\hat{\beta}$ est alors une fonction linéaire de ϵ. On a alors,

$$\hat{\beta} \quad \sim \quad N(\beta, \sigma^2 E_X (X'X)^{-1}X'\Omega X(X'X)^{-1})$$

Donc, on ne peut utiliser la variance des m.c.o. pour faire de l'inférence. On doit cependant utiliser la matrice plus haut et non la matrice correspondant au modèle sans hétéroscédasticité et autocorrélation, c.a.d. $\sigma^2(X'X)^{-1}$

4.2 Propriétés asymptotiques de l'estimateur des M.C.O.

La matrice de $\hat{\beta}$ tend vers zéro lorsque le nombre d'observations tend vers l'infini. En effet,

$$Var(\hat{\beta}) = \frac{\sigma^2}{N} E_X \left[\left(\frac{X'X}{N}\right)^{-1} \frac{X'\Omega X}{N} \left(\frac{X'X}{N}\right)^{-1} \right] \overset{N\to\infty}{\to} 0$$

si

$$plim \left(\frac{X'X}{N}\right)^{-1} = Q < \infty$$

et

$$plim \left(\frac{X'\Omega X}{N}\right) = Q^* < \infty.$$

L'estimateur est sans biais et sa variance tend vers zéro, alors on a convergence en moyenne quadratique et donc convergence en probabilité,

$$\hat{\beta} \xrightarrow{p} \beta$$

Exemple: Examinons un cas où on n'a pas la convergence en moyenne quadratique (et donc en probabilité). On suppose le modèle suivant:

$$Y = m_Y + \epsilon$$

avec la matrice de variance-covariance ϵ égale à $\sigma^2\Omega$ où

$$\Omega = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & & & \\ \rho & & 1 & & \\ \vdots & & & \ddots & \\ \rho & & & & 1 \end{bmatrix}$$

On est dans une situation où la dépendance temporelle ne diminue pas dans le temps. On peut montrer que

$$\begin{aligned} Var(\bar{Y}) &= \frac{\sigma^2}{N}(1-\rho+N\rho) \quad \rightarrow \sigma^2\rho \neq 0 \\ \left(\frac{X'\Omega X}{N}\right) &= 1+\rho(N-1) \quad \rightarrow \quad \infty \end{aligned}$$

Pour avoir convergence de l'estimateur dans le cas avec autocorrélation, la dépendance temporelle doit diminuer dans le temps.

On peut obtenir la loi asymptotique de $\hat{\beta}$ des moindres carrés ordinaires en présence d'hétéroscédasticité et d'autocorrélation. Ainsi,

$$\sqrt{N}(\hat{\beta}-\beta) = \left(\frac{X'X}{N}\right)^{-1} \frac{1}{\sqrt{N}} X'\epsilon$$

et si

$$plim\left(\frac{X'X}{N}\right) = Q$$

$$plim\sqrt{N}(\hat{\beta}-\beta) = Q^{-1} plim\left(\frac{1}{\sqrt{N}}X'\epsilon\right)$$

On applique le théorème cental limite et on obtient

$$\sqrt{N}(\hat{\beta}-\beta) \stackrel{loi}{\rightarrow} N\left(0, \sigma^2 Q^{-1}Q^*Q^{-1}\right)$$

4.3 Estimateur des moindres carrés généralisés

Si Ω est une matrice symétrique définie positive, alors elle peut s'écrire

$$\Omega = C\Lambda C'$$

où C est une matrice contenant les vecteurs propres de Ω et Λ est une matrice diagonale avec les valeurs propres sur sa diagonale et $C'C = I$.

On peut réecrire

$$C\Lambda C' = C\Lambda^{1/2}\Lambda^{1/2}C'$$

et on a que

$$C' = C^{-1}$$

puisque que C est une matrice contenant les vecteurs propres de Ω.

On définit

$$P' = C\Lambda^{-1/2}.$$

On a alors,

$$\Omega^{-1} = P'P = \underbrace{C\Lambda^{-1/2}}_{P'}\underbrace{\Lambda^{-1/2}C'}_{P}$$

On prémultiplie le modèle par P

$$\begin{aligned} PY &= PX\beta + P\epsilon \\ Y^* &= X^*\beta + \epsilon^* \end{aligned}$$

et

$$\begin{aligned} Var(\epsilon^*) &= P\sigma^2\Omega P' = \sigma^2 PC\Lambda^{1/2}\Lambda^{1/2}C'P' \\ &= \sigma^2\Lambda^{-1/2}C'C\Lambda^{1/2}\Lambda^{1/2}C'C\Lambda^{-1/2} = \sigma^2 I \end{aligned}$$

puisque $C'C = I$. On retombe sur le modèle standard. L'estimateur des moindres carrés généralisés est donnée par

$$\begin{aligned} \hat{\beta}_{m.c.g.} &= (X^{*'}X^*)^{-1}X^{*'}Y^* \\ &= (X'P'PX)^{-1}X'P'PY \\ &= \left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}Y. \end{aligned}$$

On peut montrer que cet estimateur est sans biais puisque,

$$E(\epsilon^*/X^*) = 0$$

étant donné que

$$E(P\varepsilon/PX) = 0.$$

De plus,

$$Var(\hat{\beta}_{m.c.g.}/X) = \sigma^2(X^{*'}X^*)^{-1} = \frac{\sigma^2}{N}\left(\frac{X'\Omega^{-1}X}{N}\right)^{-1} \quad \overset{N\to\infty}{\to} \quad 0.$$

Alors $\beta_{m.c.g.}$ est un estimateur convergent de β.

On a donc que

$$\begin{aligned}\hat{\beta}_{m.c.g.} &= (X^{*'}X^*)^{-1}X^{*'}Y^* \\ &= [X'P'PX]^{-1}X'P'PY \\ &= \left[X'\Omega^{-1}X\right]^{-1}X'\Omega^{-1}Y\end{aligned}$$

On va considérer dans un premier temps que la matrice Ω est connu.

En présence d'hétéroscédasticité et d'autocorrélation, l'estimateur $\beta_{m.c.g.}$ est l'estimateur linéaire sans biais à variance minimale. Pour obtenir ce résultat, on applique le théorème de Gauss-Markov sur

$$Y^* = X^*\beta + \epsilon*$$

Ceci correspond au cas général: le théorème de Aitken (1935) et Gauss-Markov est un cas particulier pour $\Omega = I$.

Pour les tests, on modifie les statistiques de la façon suivante:

$$F = \frac{(R\hat{\beta} - q)'\left[R\hat{\sigma}^2(X^{*'}X^*)^{-1}R'\right]^{-1}(R\hat{\beta} - q)}{J} \quad \sim \quad F(J, N-k)$$

où

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\hat{\epsilon}^{*'}\hat{\epsilon}^*}{N-K} = \frac{\hat{\epsilon}'P'P\hat{\epsilon}}{N-K} = \frac{\hat{\epsilon}'\Omega^{-1}\hat{\epsilon}}{N-K} \\ &= \frac{(Y - X\hat{\beta})'\Omega^{-1}(Y - X\hat{\beta})}{N-K}\end{aligned}$$

et de la même façon que pour les m.c.o., l'estimateur contraint β^*_{mcg} est égal à

$$\beta^*_{mcg} = \hat{\beta}_{mcg} - (X^{*'}X^*)^{-1}R' \left[R(X^{*'}X^*)^{-1}R' \right]^{-1} (R\hat{\beta} - q)$$

$$\beta^*_{mcg} = \hat{\beta}_{mcg} - (X'\Omega^{-1}X)^{-1}R' \left[R(X'\Omega^{-1}X^*)^{-1}R' \right]^{-1} (R\hat{\beta} - q)$$

Tous les résultats pour les tests obtenus pour les m.c.o. s'appliquent à l'estimateur des m.c.g..

Le problème général en présence d'hétéroscédasticité et d'autocorrélation consiste à minimiser la somme des carré des résidus pondérés par Ω^{-1}. Ainsi,

$$\hat{\beta}_{mcg} = \arg\min \quad (Y - X\beta)'\Omega^{-1}(Y - X\beta)$$

Par les CPO, on obtient

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

Pour les M.C.O., la pondération est égale à I.

4.4 Estimateur du maximum de vraisemblance

Si on a

$$Z \quad \sim \quad N(\mu, \Sigma),$$

alors, la densité s'écrit

$$f(Z) = (2\pi)^{-\frac{N}{2}} |\Sigma|^{-\frac{1}{2}} exp\left(-\frac{1}{2}(Z - \mu)'\Sigma^{-1}(Z - \mu) \right)$$

où Z est un vecteur $N \times 1$.

Pour notre modèle avec $E(\epsilon\epsilon') = \sigma^2\Omega$, la log vraisemblance sera alors donnée par l'expression suivante:

$$\begin{aligned} \ln L &= -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\sigma^2\Omega| - \frac{1}{2}\epsilon'(\sigma^2\Omega)^{-1}\epsilon \\ \ln L &= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln|\sigma^2| - \frac{1}{2}\ln|\Omega| - \frac{1}{2\sigma^2}(Y - X\beta)'\Omega^{-1}(Y - X\beta) \end{aligned}$$

puisque

$$|\sigma^2\Omega| = (\sigma^2)^N|\Omega|.$$

Les C.P.O. sont donnée par les équations suivantes:

$$\begin{aligned}
\frac{\delta \ln L}{\delta\beta} &= \frac{1}{\sigma^2}X'\Omega^{-1}(Y - X\beta) = \frac{1}{\sigma^2}X^{*'}(Y^* - X^*\beta) = 0 \\
\frac{\delta \ln L}{\delta\sigma^2} &= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}(Y - X\beta)'\Omega^{-1}(Y - X\beta) = 0 \\
&= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}(Y^* - X^*\beta)'(Y^* - X^*\beta) = 0
\end{aligned}$$

Par les C.P.O., l'estimateur du maximum de vraisemblance de β est:

$$\hat{\beta}_{mv} = (X^{*'}X^*)^{-1}X^{*'}Y^* = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$$

et celui de σ est:

$$\hat{\sigma}^2_{mv} = \frac{(Y - X\hat{\beta})'\Omega^{-1}(Y - X\hat{\beta})}{N}$$

L'estimateur des M.C.G. de β est aussi l'estimateur du maximum de vraisemblance. De plus, σ^2_{mv} n'est pas sans biais. On a donc les mêmes conclusions que pour le modèle sans hétéroscédasticité et sans autocorrélation.

On peut montrer également

$$\begin{bmatrix} \hat{\beta}_{MV} \\ \sigma^2_{MV} \end{bmatrix} \overset{N\to\infty}{\to} N\left(\begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix}, \begin{pmatrix} \sigma^2 E_X(X'\Omega^{-1}X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{pmatrix}\right)$$

On peut effectuer les tests de type Wald, LM et LR de la même façon.

Problème: Ω n'est pas connu.

On voudrait donc estimer Ω, cependant Ω est une matrice symétrique et elle contient $\frac{N(N+1)}{2}$ éléments différents. On a seulement N observations pour estimer $\frac{N(N+1)}{2}$ éléments.

Stratégie:

On fera dépendre Ω d'un nombre restreint de paramètres.

Exemple: Autocorrélation

$$\Omega = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{N-2} & \rho^N \\ \rho & 1 & \rho & & & \vdots \\ \rho^2 & & \ddots & & & \vdots \\ \rho^3 & & & \ddots & & \vdots \\ \cdots & & & & \ddots & \rho \\ \rho^{N-1} & \rho N-2 & \ddots & \rho^2 & \rho & 1 \end{bmatrix}$$

On a donc seulement σ^2 et ρ à estimer. L'estimateur M.C.G. sera

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

Remarques:

1. On a besoin seulement d'un estimateur convergent de Ω (et non pas efficace)
2. On perd les propriétés à distance finie (sauf pour des cas très simple)
3. L'estimateur M.C.G. sera alors optimal seulement assymptotiquement.

4.5 Hétéroscédasticité des erreurs

On a la matrice de variance-covariance pour les termes d'erreurs suivante:

$$E(\epsilon\epsilon') = \sigma^2\Omega = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ \vdots & & & \sigma_N^2 \end{bmatrix}$$

On peut récrire cette matrice de la façon suivante,

$$\sigma^2\Omega = \sigma^2 \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ \vdots & & & \omega_N \end{bmatrix}$$

Ainsi,

$$\sigma_n^2 = \sigma^2\omega_n \quad \text{pour tout } n$$

4.6 L'estimateur des M.C.O, en présence d'hétéroscédasticité

On a montré précédemment que pour l'estimateur des M.C.O. en présence de la matrice de variance-covariance générale, on a

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y \\ var(\hat{\beta}) &= \sigma^2 E_X(X'X)^{-1}X'\Omega X(X'X)^{-1} \end{aligned}$$

On peut récrire la partie du centre de la façon suivante:

$$\frac{X'\Omega X}{N} = \frac{1}{N}\Sigma_{n=1}^{N}\omega_n x_n x_n'$$

où x_n est un vecteur colonne de dimensions $K \times 1$ contenant l'observation n de chaque variable explicative.

Si $\frac{X'\Omega X}{N}$ est une matrice définie positive, alors $\hat{\beta} \xrightarrow{p} \beta$. En utilisant l'estimateur des M.C.O., la différence entre la matrice de variance-covariance (conditionnelle à X) du cas sans hétéroscédasticité et avec hétéroscédasticité est:

1. Sans hétéroscédasticité: $\sigma^2(X'X)^{-1}$

2. Avec hétéroscédasticité: $\sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$

La différence est donc:

$$\frac{\sigma^2}{N}\left(\frac{X'X}{N}\right)^{-1}\left[\frac{X'X}{N}-\frac{X'\Omega X}{N}\right]\left(\frac{X'X}{N}\right)^{-1}$$

La différence dépend donc de

$$\left[\frac{(X'X)}{N}-\frac{(X'\Omega X)}{N}\right]=\frac{1}{N}\sum_{n=1}^{N}x_nx_n'-\frac{1}{N}\sum_{n=1}^{N}\omega_nx_nx_n'$$

4.6.1 Estimateur de $\frac{X'\Omega X}{N}$ proposé par White (1980)

Cet estimateur a deux caractéristiques,

1. Estimateur non paramétrique.

2. L'hétéroscédasticité est reliée à X.

On doit évaluer

$$\Sigma=\sigma^2\frac{X'\Omega X}{N}=\frac{1}{N}\Sigma_{n=1}^{N}\sigma_n^2x_nx_n'$$

On estime Σ par

$$\hat{\Sigma}=\frac{1}{N}\Sigma_{n=1}^{N}\hat{\epsilon}_n^2x_nx_n'$$

où $\hat{\epsilon}$ est le résidu obtenu en applicant les M.C.O. Ainsi,

$$\hat{\epsilon}_n=y_n-\hat{\beta}_{MCO}x_n$$

White (1980) démontre que

$$\hat{\Sigma}\quad\xrightarrow{p}\quad\Sigma.$$

On peut donc obtenir un estimateur de la matrice de variance-covariance de l'estimateur des moindres carrés ordinaires présence d'hétéroscédasticité si on estime par M.C.O.

$$var(\hat{\beta}_{m.c.o.}/X)=N(X'X)^{-1}\underbrace{\hat{\Sigma}}_{White}(X'X)^{-1}$$

Pour cet estimateur, on a les caractéristiques suivantes:

1. On ne précise pas le type d'hétéroscédasticité.

2. Estimateur non paramétrique.

4.7 Tests pour détecter de l'hétéroscédasticité des erreurs

4.7.1 Test de White (1980)

Pour le test de White, on a l'hypothèse nulle suivante:

$$\begin{aligned} H_0 \quad &: \quad \sigma_n^2 = \sigma^2 \quad \text{pour tout } n \\ H_1 \quad &: \quad \sigma_n^2 \neq \sigma^2 \quad \text{pour au moins un } n \end{aligned}$$

Ce test est général, donc moins puissant. Nous allons voir un peu plus tard un test plus puissant mais spécifique à certaines alternatives.

Le test est basé sur la différence entre

$$\sigma^2 \left(\frac{X'X}{N} \right) \quad \text{et} \quad \sigma^2 \left(\frac{X'\Omega X}{N} \right)$$

Nous avons respectivement,

1. L'estimateur M.C.O.

$$\hat{\sigma}^2 \left(\frac{X'X}{N} \right) = \hat{\sigma}^2 \frac{1}{N} \Sigma_{n=1}^{N} x_n x_n'$$

2. L'estimateur de White

$$\hat{\Sigma} = \frac{1}{N} \Sigma_{n=1}^{N} \hat{\epsilon}_n^2 x_n x_n'$$

Le test cherche à évaluer si la différence entre les deux estimateurs est significative, donc si

$$\frac{1}{N} \sum_{n=1}^{N} (\hat{\epsilon}_n^2 - \hat{\sigma}^2) x_n x_n'$$

est significativement différente de zéro.

La matrice $x_n x_n'$ est symétrique, elle comporte $\frac{K(K+1)}{2}$ éléments différents. On utilise seulement les éléments différents pour effectuer le test. Introduire des éléments semblables ajoute aucune information supplémentaire au prix d'une puissance plus faible.

On définit le vecteur ψ_n comme étant

$$\psi_n = (\psi_{1n}, \psi_{2n}, \cdots, \psi_{mn})'$$

où $\psi_{ln} = x_{in}x_{jn}$, pour $i \geq j$ et $i = 2, \cdots, k$ et $j = 1, \cdots, k$, et $l = 1, \cdots, m$ et $m = \frac{K(K+1)}{2} - 1$. On a enlevé la constante, c'est la raison pour laquelle nous avons le terme -1. Ainsi ψ_n est un vecteur colonne de dimension $\left(\frac{K(K+1)}{2} - 1\right)$.

On définit

$$D_N = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} (\hat{\epsilon}_n - \hat{\sigma}^2)\psi_n$$

et la variance de D_N est

$$Var(D_N) = \frac{1}{N} \sum_{n=1}^{N} (\hat{\epsilon_n} - \hat{\sigma}^2)^2 (\psi_n - \bar{\psi})(\psi_n - \bar{\psi})'$$

où $\bar{\psi}$ est le vecteur contenant la moyenne de chaque ψ_l où $l = 1, \cdots, m$.

Le test de White (1980) est donc égal à

$$D_N' \left(Var(D_N)^{-1}\right) D_N \quad \stackrel{loi}{\rightarrow} \quad \chi^2 \underbrace{\left(\frac{K(K+1)}{2} - 1\right)}_{\text{degrés de liberté}}$$

Asymptotiquement, cette statistique est équivalente à effectuer la régression suivante

$$\hat{\epsilon}_n^2 = \alpha_0 + \alpha_1 \psi_{1n} + \alpha_2 \psi_{2n} + \cdots + + \alpha_m \psi_{mn} + u_n$$

et à calculer la statistique

$$NR^2 \quad \stackrel{loi}{\rightarrow} \quad \chi^2 \underbrace{\left(\frac{K(K+1)}{2} - 1\right)}_{\text{degrés de liberté}}$$

Ceci est donc un test conjoint de l'hypothèse nulle suivante:

$$\alpha_1 = \alpha_2 = \cdots = \alpha_m = 0,$$

c.à.d. que

$$E(\epsilon_n^2) = \alpha_0 = \sigma^2$$

donc que les erreurs sont homoscédastiques.

Question: pourquoi NR^2 correspond au test conjoint suivant?

$$H_0 : \quad \alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$$

On a vu dans le modèle standard $Y = X\beta + \epsilon$ que le R^2 était égal à

$$R^2 = \frac{\hat{\beta}_2' X_2' M_\iota X_2 \hat{\beta}_2}{Y' M_\iota Y}$$

où $M_\iota = [I - \iota(\iota'\iota)^{-1}\iota']$ et ι est un vecteur colonne de dimension N. $\hat{\beta}_2$ est le vecteur contenant les coefficients correspondants à X_2 qui est la matrice des variables explicatives autres que la constante.

Par le théorème de Frisch-Waugh, on obtenait que

$$\hat{\beta}_2 = \left[X_2' M_\iota X_2\right]^{-1} X_2' M_\iota Y$$

On réecrit le R^2 comme étant

$$\begin{aligned} R^2 &= \frac{Y' M_\iota X_2 \left[X_2' M_\iota X_2\right]^{-1} X_2' M_\iota X_2 \left[X_2' M_\iota X_2\right]^{-1} X_2' M_\iota Y}{Y' M_\iota Y} \\ &= \frac{Y' M_\iota X_2 \left[X_2' M_\iota X_2\right]^{-1} X_2' M_\iota Y}{Y' M_\iota Y} \end{aligned}$$

Appliquons ceci à la régression suivante:

$$\hat{\epsilon}^2 = \alpha_0 \iota + \psi\alpha + \mu$$

où $\psi = (\psi_1 \psi_2 \cdots \psi_N)$ est une matrice de dimension $N \times \left(\frac{K(K+1)}{2} - 1\right)$ et $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_m)'$. On a alors que $Y = \hat{\epsilon}^2$ et $X_2 = \psi$. Le R^2 de cette régression est donnée par

$$R^2 = \frac{\hat{\epsilon}^{2'} M_\iota \psi \left[\psi' M_\iota \psi\right]^{-1} \psi' M_\iota \hat{\epsilon}^2}{\hat{\epsilon}^{2'} M_\iota \hat{\epsilon}^2}$$

De plus, on a que:

$$M_\iota \hat{\epsilon}^2 = \hat{\epsilon}^2 - \hat{\sigma}^2 \iota$$

puisque

$$\frac{1}{N} \sum_{n=1}^{N} \hat{\epsilon}_n^2 = \hat{\sigma}^2$$

et ι est un vecteur colonne de dimension N contenant la valeur 1 pour chaque élément.

On réecrit le R^2,

$$\begin{aligned} R^2 &= \frac{(\hat{\epsilon}^2 - \hat{\sigma}^2 \iota)' \psi \left[\psi' M_\iota \psi\right]^{-1} \psi' (\hat{\epsilon}^2 - \hat{\sigma}^2 \iota)}{(\hat{\epsilon}^2 - \hat{\sigma}^2 \iota)' (\hat{\epsilon}^2 - \hat{\sigma}^2 \iota)} \\ &= \sum_{n=1}^{N} (\hat{\epsilon}_n^2 - \hat{\sigma}^2) \psi_n' \left[\sum_{n=1}^{N} (\hat{\epsilon}_n^2 - \hat{\sigma}^2)^2 \sum_{n=1}^{N} (\psi_n - \bar{\psi})(\psi_n - \bar{\psi})'\right]^{-1} \sum_{n=1}^{N} (\hat{\epsilon}_n^2 - \hat{\sigma}^2) \psi_n. \end{aligned}$$

Sous l'hypothèse nulle, cette expression est égale à

$$\left(D_N' \left(var(D_N)^{-1}\right) D_N\right) / N.$$

Ce qui nous donne,

$$NR^2 = D_N (var(D_N))^{-1} D_N'$$

C.Q.F.D.

1. Plus le R^2 est grand, plus il y a possibilité d'hétéroscédasticité.

2. Test général, donc peu puissant

4.7.2 Test de Goldfeld - Quandt

L'hétéroscédasticité dépend d'une variable explicative X_i et on sait laquelle. On aura alors un test plus puissant si on choisit bien X.

Exemple

$$\sigma_n^2 = \sigma^2 x_{i,n}$$

Procédure du test:

1. On ordonne les observations selon la taille de $X_i \longrightarrow X_i^*$
2. $\hat{\epsilon}^* = Y - X^* \hat{\beta}$
3. On sépare l'échantillon en trois groupes
   (a) X_i élevées
   (b) X_i moyennes
   (c) X_i faibles
4. On utilise seulement les groupes 1 et 3.
5. (a) $\hat{\epsilon}_1$: vecteur des résidus du groupe (1)
   (b) $\hat{\epsilon}_3$: vecteur des résidus du groupe (3)

La statistique du test est

$$\frac{\hat{\epsilon}_1' \hat{\epsilon}_1}{\hat{\epsilon}_3' \hat{\epsilon}_3} \quad \sim \quad F(n_1 - k, n_3 - k)$$

où n_1 est le nombre d'observations dans le groupe (1) et n_3 est le nombre d'observations dans le groupe (3)

4.7.3 Test de Breusch-Pagan (1979)

1. Test plus général

2. Test du multiplicateur de $\mathcal{L}$agrange (ou du score)

On considère une forme générale d'hétéroscédasticité,

$$\sigma_n^2 = f(\alpha' z_n)$$

où z_n est un vecteur dont le premier élément est 1 et les autres éléments peuvent contenir les observations x_n ou des transformations de ces observations. On décompose le vecteur α de la façon suivante:

$$\alpha = (\alpha_0, \alpha_1, \alpha_2, \cdots, \alpha_p)'.$$

si $\alpha_1, \alpha_2, \cdots, \alpha_p = 0$, alors $\alpha' z_t = \alpha_0$ et

$$\sigma^2 = f(\alpha_0) = \sigma^2$$

qui est une constante. Les résidus sont donc homoscédastiques.

L'hypothèse nulle est donc:

$$H_0 : \quad \alpha_1, \alpha_2, \cdots, \alpha_p = 0$$

Dérivons maintenant le test du multiplicateur de $\mathcal{L}$agrange pour une telle hypothèse nulle. On a donc le modèle suivant:

$$\begin{aligned} y_n &= \beta' x_n + \epsilon_n \\ &\text{où} \\ \epsilon_n &\sim N(0, \sigma_n^2) \\ &\text{et} \\ \sigma_n^2 &= f(\alpha' z_n) \end{aligned}$$

La log-vraissemblance est

$$\ln L(\beta', \alpha', x_n) = -\frac{N}{2}\ln 2\pi - \frac{1}{2}\sum_{n=1}^{N}\ln\sigma_n^2 - \frac{1}{2}\sum_{n=1}^{N}\left(\frac{(y_n - \beta' x_n)^2}{\sigma_n^2}\right)$$

Le test du multiplicateur de $\mathcal{L}$agrange consiste à évaluer sous H_0 si les C.P.O. sont significativement différentes de zéro. On définit $\theta = (\beta', \alpha')'$, alors le test LM est

$$LM = \left(\frac{\partial \ln L(\tilde{\theta})}{\partial \theta}\right)' I(\tilde{\theta})^{-1}\left(\frac{\partial \ln L(\tilde{\theta})}{\partial \theta}\right)$$

où θ est l'estimateur sous H_0. Puisque la matrice d'information est diagonale par morceaux (c. à. d. $I_{\beta\alpha} = I_{\alpha\beta} = 0$), la statistique du test est:

$$LM = \left(\frac{\partial \ln L(\tilde{\theta})}{\partial \alpha}\right)' I_{\alpha\alpha}(\tilde{\theta})^{-1}\left(\frac{\partial \ln L(\tilde{\theta})}{\partial \alpha}\right).$$

Évaluons maintenant les C.P.O. par rapport à α sous l'hypothèse nulle

$$\begin{aligned}
\frac{\partial \ln L(\tilde{\theta})}{\partial \alpha} &= \frac{1}{2}\left[\hat{\sigma}^{-2}\frac{\partial f(\hat{\alpha_0})}{\partial h_n}\right]\sum_{n=1}^{N} z_n\left(\hat{\sigma}^{-2}\hat{\epsilon}_n^2 - 1\right) \\
\frac{\partial \ln L(\tilde{\theta})}{\partial \alpha \partial \alpha} &= I_{\alpha\alpha'}(\hat{\theta}) = \frac{1}{2}\left[\hat{\sigma}^{-2}\frac{\partial f(\hat{\alpha_0})}{\partial h_n}\right]^2\sum_{n=1}^{N} z_n z_n
\end{aligned}$$

où $h_n = \alpha' z_n$.

Alors, la statistique du test est

$$LM = \frac{1}{2}\left(\sum_{n=1}^{N} z_n(\hat{\sigma}^{-2}\hat{\epsilon}_n^2 - 1)\right)'\left[\sum_{n=1}^{N} z_n z_n'\right]^{-1}\sum_{n=1}^{N} z_n\left(\hat{\sigma}^{-2}\hat{\epsilon}_n^2 - 1\right)$$

La statistique ne dépend pas de la forme de la fonction $f(.)$. On peut réécrire sous forme vectorielle

$$LM_N = \frac{1}{2}(\hat{\sigma^{-2}}\hat{\epsilon}^2 - \iota)' Z(Z'Z)^{-1}Z'(\hat{\sigma^{-2}}\hat{\epsilon}^2 - \iota) \quad \overset{loi}{\rightarrow} \quad \chi^2(p-1)$$

où ι est un vecteur de dimension N contenant des 1, et $Z = (z_1, z_2, \cdots, z_N)$

Cette statistique correspond à la moitié de la somme des carrées de la partie expliquée de la régression de $\frac{\hat{\epsilon}^2}{\hat{\sigma}^2}$ sur Z.

On peut effectuer également un test de type LR

$$LR = -2(\ln L_c - \ln L_{nc}) \quad \stackrel{loi}{\rightarrow} \quad \chi^2(p)$$

et un test de type Wald (Glesjeris test)

$$Wald = \hat{\alpha}' var(\hat{\alpha}) \hat{\alpha}$$

où $\alpha' = (\alpha_1, \alpha_2, \cdots, \alpha_p)'$

4.8 Estimation efficace des modèles avec erreurs hétéroscédastiques

1. Ω a trop de paramètres à estimer.

2. On choisit une forme paramétrique avec un nombre limité de paramètres.

Exemples:

$$\begin{aligned} \sigma_n^2 &= \sigma^2 x_n \\ \sigma_n^2 &= f(\alpha' x_n) \end{aligned}$$

4.8.1 Procédure en 2 étapes par M.C.G.

1. On effectue un M.C.O. (estimateur sans biais)

$$\begin{aligned} \hat{\epsilon} &= Y - X\hat{\beta}_{MCO} \\ &= Y - X\beta - X \underbrace{\hat{\beta} - \beta}_{\stackrel{p}{\longrightarrow} 0} \end{aligned}$$

$$\hat{\epsilon}^2 = f(Z\alpha) + u$$

On estime $\hat{\alpha}$ (par M.C.O. ou moindres carrées non linéaires) avec un estimé de $\hat{\alpha}$, on obtient un estimé de $\hat{\Omega}$.

2.

$$\begin{aligned}\hat{\beta}_{mcg} &= (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}Y \\ \hat{\sigma}^2_{mcg} &= \frac{(Y - X\hat{\beta})'\hat{\Omega}^{-1}(Y - X\hat{\beta})'}{N - K}\end{aligned}$$

On perd les propriétés à distance finie (petit échantillon).

4.8.2 Estimation par Maximum de vraisemblance

On écrit la vraisemblance avec $\sigma_n^2 = f(\alpha' z_n)$

$$\ln L(\beta, \alpha; x) = -\frac{N}{2}\ln 2\pi - \frac{1}{2}\sum_{n=1}^{N} lnf(\alpha' z_n) - \frac{1}{2}\sum_{n=1}^{N}\frac{(y_n - \beta' x_n)^2}{f(\alpha' z_n)}$$

5 Autocorrélation des erreurs

Notion de séries temporelles. On cherche à exprimer la dépendance temporelle des résidus de façon paramétriques.

5.1 Concepts de séries temporelles

Definition 11 *Un processus X_t est stationnaire du second ordre si*

1. $EX_t = m$ *(indépendant de t),* $\quad \forall t$ *et,*
2. $EX_t^2 < \infty, \quad \forall t,$
3. $cov(X_t, X_{t-h}) = \gamma(h)$ *est indépendant de t, pour* $\forall t$ *et dépend seulement de h.*

On va examiner trois types de processus paramétriques.

5.1.1 Processus autorégressifs (AR)

Definition 12 *On appelle processus autorégressif d'ordre p, un processus stationnaire X_t vérifiant une relation du type,*

$$X_t = \mu + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \epsilon_t$$

où ϵ est un bruit blanc.

Qu'est-ce qu'un bruit blanc

1. $E(\epsilon) = 0$
2.

$$\begin{aligned} E(\epsilon_t, \epsilon_{t-k}) &= \sigma^2 \text{ si } k = 0 \\ &= 0 \text{ autrement} \end{aligned}$$

Exemple: processus AR(1) sans constante

$$X_t = \phi_1 X_{t-1} + \epsilon_t$$

Question: est-ce que ce processus est stationnaire du second ordre?

On vérifie les conditions 2) et 3), on reviendra à la condition 1) plus tard. Examinons la deuxième condition:

$$var(X_t) = E(X_t)^2 - E(X_t)E(X_t) = \gamma(0).$$

Puisque la constante est égale à zéro, alors,

$$\begin{aligned} var(X_t) = E(X_t X_t) &= \phi E(X_t X_{t-1}) + E(X_t \epsilon_t) \\ \gamma(0) &= \phi\gamma(1) + E(X_t \epsilon_t) \end{aligned}$$

On cherche dans un premier temps la valeur de $E(X\epsilon)$

$$E(X_t\epsilon_t) = \phi E(X_{t-1}\epsilon_t) + E(\epsilon_t\epsilon_t)$$

$$E(X_t\epsilon_t) = \sigma^2$$

puisque $E(X_{t-1}\epsilon_t) = 0$.

Ainsi,

$$E(X_tX_{t-1}) = \phi E(X_{t-1}X_{t-1}) + E(X_{t-1}\epsilon_t)$$

$$\gamma(1) = \phi\gamma(0).$$

On a donc que

$$\begin{aligned}
\gamma(0) &= \phi\left(\phi\gamma(0)\right) + \sigma^2 \\
\gamma(0) &= \frac{\sigma^2}{(1-\phi^2)} < \infty \text{ si } |\phi| < 1 \\
&et \\
\gamma(1) &= \phi\frac{\sigma^2}{(1-\phi^2)} \text{ ne dépend pas de t} \\
&et \\
\gamma(2) &= E(X_tX_{t-2}) = \phi E(X_{t-1}X_{t-2}) + E(X_{t-2}\epsilon_t)
\end{aligned}$$

puisque $E(X_{t-2}\epsilon_t) = 0$ alors,

$$\begin{aligned}
\gamma(2) &= \phi\gamma(1) \\
\gamma(2) &= \phi^2\frac{\sigma^2}{(1-\phi^2)}
\end{aligned}$$

De façon générale,

$$\gamma(h) = \phi^h\frac{\sigma^h}{(1-\phi^2)} \quad \forall t$$

Donc, stationnaire du second ordre si $|\phi| < 1$. Si $|\phi| = 1$, on aura ce qu'on appelle une racine unité. Dans ce cas la variable est non stationnaire.

5.1.2 Processus moyennes mobiles (MA)

Definition 13 *On appelle processus moyenne mobile d'ordre q, un processus* X_t *défini par*

$$X_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q}$$

où ϵ_t *est un bruit blanc.*

Exemple: processus MA(1) avec une constante nulle

$$X_t = \epsilon_t + \theta\epsilon_{t-1}$$

Question: est-ce que ce processus est stationnaire du second ordre?

$$\begin{aligned}
E(X_t) &= E(\epsilon_t) + \theta E(\epsilon_{t-1}) = 0 \\
var(X_t) &= E\left[(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_t + \theta\epsilon_{t-1})\right] \\
&= \sigma^2 + \theta^2\sigma^2 = (1+\theta^2)\sigma^2 < \infty, \quad \forall\theta \\
\gamma(1) &= cov(X_tX_{t-1}) = E\left[(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_{t-1} + \theta\epsilon_{t-2})\right] \\
&= \theta\sigma^2 \\
\gamma(2) &= cov(X_tX_{t-2}) = E\left[(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_{t-2} + \theta\epsilon_{t-3})\right] \\
&= 0 \\
\gamma(3) &= 0 \\
&\vdots \\
\gamma(k) &= 0, \quad \text{pour } k > 1
\end{aligned}$$

Remarque:

Pas de condition sur le paramètre θ pour avoir un processus stationnaire.

5.1.3 Processus ARMA

Definition 14 *Un processus stationnaire X_t admet une représenation ARMA(p,q) minimale s'il satisfait*

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} + \cdots - \phi_p X_{t-p} = \mu\epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q}$$

où ϵ est un bruit blanc.

Exemple: processus ARMA(1,1) avec une constante nulle

$$X_t = \phi X_{t-1} + \epsilon_t + \theta\epsilon_{t-1}$$

On définit un opérateur de retard "L" tel que

$$L\ X_t = X_{t-1}, \quad L^n\ X_t = X_{t-n}$$

Si on inverse un processus autorégressif, on obtient un processus MA(∞).

Exemple: processus AR(1)

$$\begin{aligned} X_t &= \phi X_{t-1} + \epsilon_t \\ (1-\phi L)X_t &= \varepsilon \\ \Rightarrow \quad X_t &= \frac{\epsilon}{(1-\phi L)} = \sum_{i=0}^{\infty} \phi^i L^i \epsilon_t \end{aligned}$$

donc,

$$X_t = \epsilon_t + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + \cdots + \phi^\infty\epsilon_{t-\infty}.$$

Si on inverse un processus MA(q), on obtient un processus AR(∞).

Exemple: MA(1)

$$X_t = \epsilon_t + \theta\epsilon_{t-1} = (1+\theta L)\epsilon_t$$

$$\frac{X_t}{(1+\theta L)} \quad = \quad \frac{X_t}{(1-(-\theta L))}.$$

Ce qui donne

$$\sum_{i=0}^{\infty}(-\theta)^i L^i X_t \quad = \quad \epsilon_t$$

Prenons notre modèle,

$$Y = X\beta + \epsilon$$

et on suppose que les termes d'erreurs de la régression suivent un processus AR(1), alors sa matrice de variance-covariance sera égale à

$$\sigma^2\Omega = \frac{\sigma_u^2}{(1-\phi^2)} \begin{bmatrix} 1 & \phi & \phi^2 & \phi^3 & \cdots & \phi^{T-1} \\ \phi & 1 & \phi & \phi^2 & \cdots & \vdots \\ \phi^2 & \phi & 1 & \phi & \cdots & \vdots \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & & \ddots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \cdots & \phi & \cdots & 1 \end{bmatrix}$$

Cette matrice est seulement fonction des paramètres σ_u^2 et ϕ.

5.2 Conséquences pour l'estimateur des M.C.O.

Dans la cas général avec autocorrélation des erreurs, l'estimateur M.C.O. est

1. Sans biais,
2. Converge si $\frac{X'\Omega X}{T}$ est finie, donc X_t doit bien se comporter et la corrélation entre les erreurs doit s'estomper dans le temps (exemple: $\phi < 1$).
3. Normale de façon asymptotique, mais elle est très difficile à établir.

Donc, l'estimateur M.C.O. est sans biais, convergent et asymptotiquement normal. Sa matrice de variance-covariance conditionnelle est:

$$var(\hat{\beta}^{MCO}/X) = \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$$

5.2.1 Estimation de Ω

Si la forme paramétrique est connue, (ex: AR, MA ou ARMA), on estime cette représentation et on obtient $\hat{\Omega} = \Omega(\hat{\theta})$ où $\hat{\theta}$ sont les estimés de la représentation.

Il existe également un estimateur non paramétrique (comme celui de White pour l'hétéroscédasticité). On cherche donc à estimer $\sigma^2 \frac{X'\Omega X}{T}$. Cette matrice est égale à

$$\hat{\Sigma} = \hat{\sigma}^2 \frac{X'\hat{\Omega}X}{T} = S_T = \underbrace{S_0}_{White} + \frac{1}{T}\sum_{j=1}^{L}\sum_{t=j+1}^{T} w_j \hat{\epsilon}_t \hat{\epsilon}_{t-j}(x_t x'_{t-j} + x_{t-j}x'_t)$$

où $S_0 = \frac{1}{T}\sum_{t=1}^{T} \hat{\epsilon_t^2} x_t x'_t$, et w_j est une de pondération qui dépend de j pour assurer que la matrice S_t soit positif définie.

exemple:

$$w_j = 1 - \frac{j}{L+1}$$

C'est la fenêtre de Bartlett proposée par Newey et West (1987). Le problème est le choix de "L": Newey et West (1994) propose une méthode de sélection automatique selon les données.

5.3 Tests de l'autocorrélation des erreurs

5.3.1 Test de Durbin-Watson

On a le modèle

$$y_t = \beta' x_t + \varepsilon_t \quad \text{et} \quad \varepsilon_t = \rho\varepsilon_{t-1} + \mu_t$$

$$H_0 \quad : \quad \rho = 0$$

$$H_1 \quad : \quad \rho \neq 0$$

Le test de Durbin-Watson est basé sur la statistique

$$d = \frac{\sum_{t=2}^{T}(\hat{\epsilon}_t - \hat{\epsilon}_{t-1})^2}{\sum_{t=1}^{T}\hat{\epsilon}_t^2} \quad \approx \quad 2(1-r)$$

où $r = \hat{\rho} = \left(\sum_{t=2}^{T}\hat{\epsilon}_{t-1}\hat{\epsilon}_{t-1}\right)^{-1}\sum_{t=2}^{T}\hat{\epsilon}_{t-1}\hat{\epsilon}_t$.

Si $r \approx 1 \quad \Rightarrow \quad d = 0 \quad \Rightarrow$ autocorrélation fortement positive.

Si $r \approx 0 \quad \Rightarrow \quad d = 2 \quad \Rightarrow$ pas d'autocorrélation.

Si $r \approx -1 \quad \Rightarrow \quad d = 4 \quad \Rightarrow$ autocorrélation fortement négative.

Problème: La loi du test de Durbin-Watson dépend des observations x_t,

En effet, si on écrit la statistique "d" sous la forme vectorielle, on obtient

$$d = \frac{\hat{\epsilon}' A \hat{\epsilon}}{\hat{\epsilon}'\hat{\epsilon}}$$

où

$$A = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & 0 & -1 & 2 & -1 & \cdots & 0 \\ \vdots & & & & \ddots & & \vdots \\ \vdots & & & & -1 & 2 & -1 \\ \vdots & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

et

$$\hat{\epsilon} = M\epsilon = [I - X(X'X)^{-1}X]$$

$$d = \frac{\epsilon' M' A M \epsilon}{\epsilon' M \epsilon}.$$

On voit bien que d dépend des observations. Durbin et Watson ont réussi à borner la loi de la statistique d, mais il y a une zone d'indétermination qui

dépend du nombre d'observations. En particulier, si T augmente, cette zone diminue.

On aura deux valeurs critiques (d_u et d_l) qui détermine les zones de rejet. On ne rejette pas H_0 si $d > d_u$ et on rejette H_0 si $d < d_l$. Si $d_l < d < d_u$, on ne peut décider.

Dans le cas où, la valeur de d excède 2, alors l'hypothèse alternative est une autocorrélation négative. On utilise $4 - d$ pour effectuer le test.

Il y a deux conditions importantes pour utiliser le test de Durbin-Watson,

1. On doit absolument inclure une constante.

2. X doit être fixe. Par exemple, on ne peut inclure des variables retardés comme régresseurs.

De plus, si les erreurs sont caractérisées par un processus AR(p), le paramètre AR(1), $\hat{\rho}$, ne contient pas toute les informations sur la dépendance temporelle.

5.3.2 Test de Breusch (1978) et Godfrey (1978)

On est en présence d'un test de type LM. Les hypothèses nulles et alternatives sont les suivantes:

$$\begin{aligned} H_0 &: \text{ pas d'autocorrélation} \\ H_1 &: \epsilon_t \sim AR(p) \text{ ou } \epsilon_t \sim MA(p) \end{aligned}$$

Le test consiste à effectuer une régression de $\hat{\epsilon}_t$ sur les x_t et $\hat{\epsilon}_{t-1}, \hat{\epsilon}_{t-2}, \cdots, \hat{\epsilon}_{t-p}$ et à calculer la statistique suivante:

$$TR^2 \sim \chi^2(p)$$

Puisque $X'\epsilon = 0$ (par hypothèse), le test est équivalent à regresser $\hat{\epsilon}_t$ sur la partie des $\hat{\epsilon}_{t-1}, \cdots, \hat{\epsilon}_{t-p}$ qui n'est pas expliqué par les X_t (application du théorème de F.W.).
Si R^2 est significativement différent de zéro, il y a autocorrélation. C'est un

test conjoint de p coeficients. Bien sûr, le choix de p est important pour la puissance du test. Ce test est valide avec variable retardés comme régresseurs pour l'équation de Y

5.3.3 Test de Box et Pierce

Le test de Box et Pierce (appelé également test du "portemanteau") est basé sur la statistique suivante:

$$Q_T = T \sum_{j=1}^{L} \hat{r}_j^2 \sim \chi^2(L)$$

où

$$\hat{r}_j = \left(\sum_{t=j+1}^{T} \hat{\epsilon}_{t-j} \hat{\epsilon}_{t-j} \right)^{-1} \sum_{t=j+1}^{T} \hat{\epsilon}_{t-j} \hat{\epsilon}_t$$

Lung et Box ont proposé un ajustement en petit échantillon de cette statistique,

$$Q_t^{LB} = T(T+2) \sum_{j=1}^{L} \frac{\hat{r}_j^2}{T-j}$$

1. La puissance du test dépend du choix de "L".

2. Le test de Breusch et Godfrey semble plus puissant que les test de Box-Pierce et Ljung-Box.

5.4 Estimation efficace des modèles avec erreurs autocorrelées

Examinons le cas où les erreurs suivent un processus $AR(1)$. On a le modèle suivant:

$$Y = X\beta + \epsilon \quad \text{où} \quad \epsilon = \epsilon_{t-1}\rho + \mu$$

Alors,

$$Var(\epsilon) = \sigma_\epsilon^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{T-3} \\ \vdots & & & & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-1} & \cdots & \cdots & \cdots & 1 \end{bmatrix}$$

$$= \frac{\sigma_\mu^2}{(1-\rho^2)} \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{T-3} \\ \vdots & & & & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-1} & \cdots & \cdots & \cdots & 1 \end{bmatrix}$$

$$= \sigma_\mu^2 \Omega$$

où

$$\Omega = \frac{1}{(1-\rho^2)} \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{T-3} \\ \vdots & & & & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-1} & \cdots & \cdots & \cdots & 1 \end{bmatrix}.$$

La matrice inverse est donnée par:

$$\Omega^{-1} = \begin{bmatrix} 1 & -\rho & 0 & 0 & 0 & & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & & & \vdots \\ 0 & -\rho & 1+\rho^2 & -\rho & & & \vdots \\ \vdots & & & & \ddots & & \vdots \\ \vdots & & & & -\rho & 1+\rho^2 & -\rho \\ 0 & \cdots & \cdots & \cdots & \cdots & -\rho & 1 \end{bmatrix}$$

La matrice de transformation P est tel que $\Omega^{-1} = P'P$ et

$$P = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & 0 & 0 & 0 \\ -\rho & 1 & 0 & 0 & 0 & \vdots \\ 0 & -\rho & 1 & 0 & & \vdots \\ 0 & 0 & -\rho & 1 & 0 & \vdots \\ \vdots & & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & -\rho & 1 \end{bmatrix}$$

On fait donc des M.C.O. sur

$$\begin{aligned} PY &= PX\beta + P\epsilon \\ Y^* &= X^*\beta + \epsilon^* \end{aligned}$$

$$Y^* = \begin{bmatrix} \sqrt{1-\rho^2}y_1 \\ y_2 - \rho y_1 \\ y_3 - \rho y_2 \\ \vdots \\ y_T - \rho y_{T-1} \end{bmatrix} \qquad X^* = \begin{bmatrix} \sqrt{1-\rho^2}x_1 \\ x_2 - \rho x_1 \\ x_3 - \rho x_2 \\ \vdots \\ x_T - \rho x_{T-1} \end{bmatrix}$$

On remarque que la première observation de Y^* et X^* est différente. On peut récrire pour $t = 2, \cdots, T$

$$y_t = \beta' x_t + \epsilon_t, \quad \text{où} \quad \epsilon_t = \rho\epsilon_{t-1} + u_t.$$

Ce qui implique donc,

$$\begin{aligned} y_t - \rho y_{t-1} &= \beta' x_t - \rho\beta' x_{t-1} + u_t \\ y_t &= \rho y_{t-1} + \beta' x_t - \rho\beta' x_{t-1} + u_t \end{aligned}$$

et u_t est homoscédastique.

Si ρ est inconnue, on doit obtenir un estimé. On peut utiliser l'estimateur des M.C.O. pour obtenir $\hat{\rho}$, alors

$$\hat{\rho} = (\hat{\epsilon}'_{-1}\hat{\epsilon}_{-1})^{-1}\hat{\epsilon}'_{-1}\hat{\epsilon}$$

où $\hat{\epsilon}_{-1}$ est le vecteur retardé d'une période.

On peut effectuer directement les M.C.O. sur

$$y_t = \rho y_{t-1} + \beta' x_t - \rho\beta' x_{t-1} + \mu_t$$

pour $t = 2, \cdots, T$.

5.4.1 Maximum de vraisemblance

On sait que

$$f(x_1, x_2) = f(x_1/x_2)f(x_2)$$

Dans le cas qui nous intéresse, la densité conjointe sera donnée par:

$$f(y_1, y_2, \cdots, y_T) = f(y_1)f(y_2/y_1)f(y_3/y_2)\cdots f(y_T/y_{T-1})$$

La première observation du modèle transformé est

$$\sqrt{1-\rho^2}y_1 = \sqrt{1-\rho^2}\beta' x_t + u_1$$

et pour $t = 2, \cdots, T$

$$y_t = \rho y_{t-1} + \beta' x_t - \rho\beta' x_{t-1} + u_t$$

On cherche $f(y_t)$, on a vu que

$$f(y_1) = f(u_1) \underbrace{\left|\frac{\partial u_1}{\partial y_1}\right|}_{Jacobien}$$

alors $\left|\frac{\partial u_1}{\partial y_1}\right| = \sqrt{1-\rho^2}$ et $f(u_1) = f(\epsilon_1\sqrt{1-\rho^2})$ puisque $var(u_1) = (1-\rho^2)var(\epsilon_1)$. Donc

$$f(y_1) = \sqrt{1-\rho^2} f\left(\sqrt{(1-\rho^2)}(y_1 - \beta' x_1)\right).$$

On peut réécrire comme étant

$$f(y_1) = \sqrt{1-\rho^2}\left[2\pi\sigma_u\right]^{\frac{-1}{2}} \exp\left(\frac{-1}{2}\frac{(1-\rho^2)}{\sigma_u^2}(y_1 - \beta' x_1)^2\right).$$

La log-vraisemblance est alors donnée par

$$\ln L \;=\; \ln f(y_1) + \sum_{t=2}^{T} \ln f(y_t/y_{t-1})$$

On maximise par rapport à β, σ^2, ρ pour obtenir les estimateurs.

5.5 ARCH- Hétéroscédasticité conditionnelle de forme autorégressive

On est en présence de persistance de la variance (finance, macroéconomie), ex: inflation, Bon du trésor.

La variance du terme d'erreur au temps t dépend de la variance des termes d'erreurs retardés.

Une version simple du modèle ARCH est

$$y_t = \beta' x_t + \epsilon_t$$

où

$$\epsilon_t = u_t(\alpha_0 + \alpha_1 \epsilon_{t-1}^2)^{\frac{1}{2}} \quad \text{et} \quad u_t \sim N(0,1).$$

Ceci est un preocessus ARCH(1). On a pour ce processus que

$$\begin{aligned} E(\epsilon_t/\epsilon_{t-1}) &= 0 \\ var(\epsilon_t/\epsilon_{t-1}) &= E(\epsilon_t^2/\epsilon_{t-1}^2) \\ &= E(\mu_t^2)(\alpha_0 + \alpha_1 \epsilon_{t-1}^2) \\ &= \alpha_0 + \alpha_1 \epsilon_{t-1}^2 \end{aligned}$$

Donc, ϵ_t est hétéroscédastique conditionnellement à ϵ_{t-1}, c'est donc une forme autorégressive.

La variance marginale est donnée par

$$\begin{aligned} var(\epsilon_t) &= E\left(u_t^2(\alpha_0 + \alpha_1 \epsilon_{t-1}^2)\right) \\ &= \alpha_0 + \alpha_1 E(\epsilon_{t-1}^2) \\ &= \alpha_0 + \alpha_1 var(\epsilon_{t-1}) \end{aligned}$$

Si le processus est stationnaire du second ordre, alors

$$\begin{aligned} var(\epsilon_t) &= var(\epsilon_{t-1}) \\ \Rightarrow \quad var(\epsilon_t) &= \frac{\alpha_0}{1-\alpha_1} \end{aligned}$$

Les hypothèses du modèle linéaire sont respectées, donc l'estimateur des M.C.O. est l'estimateur linéaire optimal de β. Cependant, il existe un estimateur plus efficace non linéaire.

La fonction de vraissemblance pour ce modèle est conditionnelle aux valeurs de départ y_0 et X_0.

$$\ln L = \text{constante} - \frac{1}{2}\sum_{t=1}^{T}\ln(\alpha_0 + \alpha_1\epsilon_{t-1}^2) - \frac{1}{2}\sum_{t=1}^{T}\frac{\epsilon_t^2}{\alpha_0 + \alpha_1\epsilon_{t-1}^2}$$

où $\epsilon_t = y_t - \beta' x_t$.
On maximise par rapport à $\beta, \alpha_0, \alpha_2$, Il existe également une méthode en 4 étapes des M.C.G. (pp. 798, Greene).

5.5.1 Test pour les ARCH

Test de type LM, Engle 1982:
On estime par les M.C.O., on obtient $\hat{\epsilon}_t$, on effectue la régression suivante

$$\hat{\epsilon}_t^2 = \alpha_0 + \alpha_1\hat{\epsilon}_{t-1}^2 + \alpha_1\hat{\epsilon}_{t-2}^2 + \cdots + \alpha_p\hat{\epsilon}_{t-p}^2 + u_t$$

Le test consiste à calculer la statistique

$$TR^2 \quad \sim \quad \chi^2(p)$$

pour cette régression.

5.5.2 GARCH-Generalized Autoregressive Conditional Heteroscedasticity

On a encore le modèle suivant

$$y_t = \beta' x_t + \epsilon_t$$

et

$$\epsilon_t = \sqrt{h_t}u_t$$

On avait pour le modèle ARCH (p)

$$h_t = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \alpha \epsilon_{t-2}^2 + \cdots + \alpha_p \epsilon_{t-p}^2$$

Le GARCH est une généralisation avec une composante moyenne mobile. On aura

$$h_t = \alpha_0 + \delta_1 h_{t-1} + \delta_2 h_{t-2} + \cdots + \delta_r h_{t-r} + \alpha_1 \epsilon_{t-1}^2 + \alpha \epsilon_{t-2}^2 + \cdots + \alpha \epsilon_{t-p}^2$$

6 Représentation univariée de séries temporelles

6.1 Modèle de régression dynamique

Un modèle dynamique est un modèle dont la variable dépendante est fonction de ses retards et des retards des autres variables prédéterminées ou exogènes.

Exemple: Représentations AR, MA et ARMA.

1. AR: $y_t = \rho y_{t-1} + \varepsilon_t$
2. MA: $y_t = \varepsilon_t + \theta \varepsilon_{t-1}$
3. ARMA: $y_t = \rho y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$.

Exemple: Modèle d'ajustement partiel.

Il existe un niveau désiré de la variable y_t noté y^*, tel que

$$y_t^* = x_t' \beta^* + u_t$$

L'agent ne peut atteindre y_t^* à chaque période parce qu'il y a des coûts d'ajustements. Au lieu y_t s'ajuste vers y_t^* de la façon suivante:

$$y_t - y_{t-1} = (1 - \delta)(y_t^* - y_{t-1}) + v_t.$$

En combinant les deux équations, on obtient

$$\begin{aligned} y_t &= y_{t-1} - (1-\delta)y_{t-1} + (1-\delta)x_t'\beta^* + (1-\delta)u_t + v_t \\ &= x_t\beta + \delta y_{t-1} + e_t \ \text{ et } \ |\delta| < 1 \end{aligned}$$

où $\beta = (1-\delta)\beta$ et $e_t = (1-\delta)u_t + v_t$.
On effectue un M.C.O. sur cette équation et on obtient $\hat{\beta}*$ et $\hat{\delta}$ à partir de l'estimateur de $\hat{\beta}$ et $\hat{\delta}$.

Problème

Lorsque des variables dépendantes retardées sont incluses dans la régression, l'estimateur des M.C.O. est biaisé en petit échantillon. Donc la matrice des variables explicatives X contient des y retardées. On aura que

$$E\left[(X'X)^{-1}X'u\right] \neq 0$$

On a bien que $E(X'u) = 0$ mais $E\left[(X'X)^{-1}X'u\right] \neq 0$ car $X'X$ contient des valeurs pour $\forall t$ qui sont correlées avec $X'\mu$. En effet, les y_t sont corrélées avec les u_{t-j} pour $j \geq 1$.

Exemple:

$$y_t = \rho y_{t-1} + u_t \ \text{ où } \ |\rho| < 1.$$

On peut réecrire

$$\begin{aligned} y_t &= \sum_{j=0}^{t} \rho^j u_{t-j} \\ \Rightarrow \quad E(y_t u_{t+j}) &= 0 \ \text{ pour } \forall j > 0 \\ \text{mais } \ E(y_t u_{t-j}) &\neq \text{ pour } \forall i \geq 0 \end{aligned}$$

Ainsi,

$$\hat{\rho_T} = (\sum_{t=2}^{T} y_{t-1}y_{t-1})^{-1} \sum_{t=2}^{T} y_{t-1}y_t$$

$$= \rho + \underbrace{(\sum_{t=2}^{T} y_{t-1}y_{t-1})^{-1}}_{(1)} \underbrace{\sum_{t=2}^{T} y_{t-1}u_t}_{(2)}$$

L'espérance de ceci n'égale pas zéro, parce que le terme (1) est corrélé avec le terme (2).

Calculer l'espérance de cette expression est assez compliquée. Donc, en petit échantillon, l'estimateur des M.C.O. d'un modèle dynamique incluant une variable endogène retardée est biaisé.
Cependant l'estimateur des M.C.O. est convergent . Asymptotiquement, on peut écrire

$$plim\hat{\rho} = \rho + plim(Y'_{-1}Y_{-1})^{-1}plim(Y_{-1}u) = \rho$$

par le théorème de Slutsky où Y_{-1} est le vecteur contenant la variable Y retardée d'une période. $Plim(Y'_{-1}Y_{-1})$ est finie et non singulière et $plim(Y_{-1}u) = 0$, alors l'estimateur M.C.O. est convergent.

6.2 Processus non stationnaires, régression fictive et cointégration

La plupart des séries macroéconomiques et financières ne sont pas stationnaires. La moyenne et la variance ne sont pas indépendantes de t.

6.2.1 Processus particuliers

Marche aléatoire

$$y_t = y_{t-1} + \varepsilon_t$$

Ce processus correspond à un processus AR(1) avec un coefficient égal à 1.
On ne peut prédire les variations de y_t, c.à.d. $\Delta y_t = y_t - y_{t-1}$.

On va étudier ce processus. Réecrivons ce processus:

$$y_t = y_{t-1} + \varepsilon_t$$

$$
\begin{aligned}
&= y_{t-2} + \varepsilon_{t-1} + \varepsilon_t \\
&= y_{t-3} + \varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t \\
&\vdots \\
y_t &= y_0 + \sum_{i=1}^{t} \varepsilon_i
\end{aligned}
$$

Les chocs ne s'estompent pas dans le temps. On dit que les chocs sont permanents. Pour un processus AR(1) stationnaire ($|\phi| < 1$), on a que

$$y_t = \rho^t t_0 + \sum_{j=0}^{t-1} \phi^j \varepsilon_{t-j}.$$

Contrairement à la marche aléatoire, l'effet des chocs s'estompent dans le temps puisque ϕ^i tend vers zéro, pour $|\phi| < 1$, lorsque i tend vers l'infini. On dira que les chocs ont un effet transitoire (mean reverting).

Calculons les moments d'ordre 1 et 2 de la marche aléatoire:

$$
\begin{aligned}
E(y_t) &= y_0 \ \ \forall t \\
var(y_t) &= var\left[y_0 + \sum_{i=1}^{t} \varepsilon_i\right] = t\sigma^2
\end{aligned}
$$

Lorsque le nombre d'observations (t) tend vers l'infini, la variance tend vers l'infini. La variance n'est pas donc bornée. Ce processus est donc non stationnaire.

La fonction d'autocovariance est donnée par:

$$
\begin{aligned}
cov(y_t y_{t-k}) &= E\left[(y_t - y_0)(y_{t-k} - y_0)\right] \\
&= E\left[(\varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2} + \cdots)(\varepsilon_{t-k} + \varepsilon_{t-k-1} + \varepsilon_{t-k-2} + \cdots)\right] \\
&= E\left[(\varepsilon_{t-k}^2 + \varepsilon_{t-k-1}^2 + \varepsilon_{t-k-2}^2 + \cdots)\right] \\
&= (t-k)\sigma^2.
\end{aligned}
$$

Elle dépend donc de t.

La fonction d'autocorrélation est:

$$\rho_k = \frac{cov(y_t y_{t-k})}{[var(y_t) var(y_{t-k})]^{\frac{1}{2}}} = \frac{(t-k)\sigma^2}{[t\sigma^2(t-k)\sigma^2]^{\frac{1}{2}}}$$

$$= \left[\frac{t-k}{t}\right]^{\frac{1}{2}}.$$

La fonction d'autocorrélation décroît donc lentement en k. Si on différencie la marche aléatoire, on obtient un bruit blanc

$$\Delta y_t = y_t - y_{t-1} = \varepsilon_t$$

Exemples: Rendement boursier et taux de change.

Marche aléatoire avec dérive (Random Walk)

La marche aléatoire avec dérive est définie comme etant:

$$y_t = \mu + y_{t-1} + \varepsilon_t$$

où ε_t est un bruit blanc.

Réecrivons ce processus,

$$\begin{aligned} y_t &= \mu + y_{t-1} + \varepsilon_t \\ &= \mu + \mu + y_{t-2} + \varepsilon_{t-1} + \varepsilon_t \\ &= \mu + \mu + \mu + y_{t-3} + \varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t \\ &\vdots \\ y_t &= y_0 + \mu t + \sum_{i=1}^{t} \varepsilon_i. \end{aligned}$$

C'est donc un processus caractérisé par une tendance linéaire et dont les chocs ont un effet permanent sur y_t.

Les fonctions d'autocovariance et d'autocorrélation sont données par les expressions suivantes:

$$\begin{aligned} var(y_t) &= t\sigma^2 \\ cov(y_t y_{t-k}) &= (t-k)\sigma^2 \quad \Rightarrow \quad \text{dépend de } t \\ \rho_k &= \left[\frac{t-k}{t}\right]^{\frac{1}{2}}. \end{aligned}$$

Ces expressions sont les mêmes que pour la marche aléatoire sans dérive.

De façon générale, on dira que y_t est intégrée d'ordre d si cette variable doit être différenciée d fois pour être stationnaire. On notera

$$y \sim I(d)$$

En particulier,

$$\begin{aligned} y &\sim I(1) \Rightarrow \Delta y_t \text{ est stationnaire,} \\ y &\sim I(0) \Rightarrow y_t \text{ est stationnaire.} \end{aligned}$$

Exemple: ARIMA(p,d,q)

$$\phi_p(L)(1-L)^d y_t = \theta_q(L)\varepsilon_t$$

On a vu que pour une série $I(1)$, la fonction d'autocorrélation décroît lentement. On doit cependant faire un test formel pour savoir si la série est stationnaire ou pas. On appelle ces tests, des tests de racine unité ou unitaire. Le test consiste à savoir si la plus grande racine est égale à 1.

Exemple: Processus AR(1)

$$y_t = \mu + \phi_1 y_{t-1} + \varepsilon_t$$

$$\begin{aligned} H_0 &: \quad \phi_1 = 1 \\ H_1 &: \quad \phi_1 < 1 \end{aligned}$$

On peut réecrir le processus AR(1) tel que

$$\begin{aligned} \Delta y_t &= \mu + (\phi - 1) y_{t-1} + \varepsilon_t \\ \Delta y_t &= \mu + \alpha y_{t-1} + \varepsilon_t \end{aligned}$$

$$\begin{aligned} H_0 &: \quad \alpha = 0 \quad \Rightarrow \text{non stationnaire} \\ H_1 &: \quad \alpha < 0 \quad \Rightarrow \text{stationnaire} \end{aligned}$$

On appelle ce test, le test de Dickey-Fuller. La loi asymptotique de la statistique t est non standard. La loi dépend du nombre de termes déterministes dans les équations suivantes:

	5%	1%
$\Delta y_t = \alpha y_{t-1} + \varepsilon_t$	-1.95	-2.6
$\Delta y_t = \mu + \alpha y_{t-1} + \varepsilon_t$	-2.89	-3.51
$\Delta y_t = \mu + \beta t + \alpha y_{t-1} + \varepsilon_t$	-3.45	-4.04

Si on utilise les valeurs critiques standards, on rejette H_0, donc la racine unité, trop souvent.

Pour un processus AR(p), on aura le test augmenté par les retards de la variable en différence

$$\Delta y_t = \mu + \alpha y_{t-1} + \sum_{i=1}^{p} \phi_i \Delta y_{t-i} + \varepsilon_t$$

$$\begin{aligned} H_0 &: \quad \alpha = 0 \\ H_1 &: \quad \alpha < 0 \end{aligned}$$

Les valeurs critiques sont le mêmes que pour le processus AR(1).

Pour effectuer le test de Dickey-Fuller augmenté, on doit choisir le nombre de retards. Ce choix est très important. Si notre équation ne comporte pas assez de retards, alors le test n'aura pas le bon niveau puisque que $\hat{\alpha}$ sera biaisé. Si l'équation comporte trop de retards, alors le tests sera moins puissants (perte de degrés de liberté).

Choix du nombre de retards

Campbell et Perron ont proposé une procédure pour fixer le nombre de retards. On fixe un nombre de retards maximal et on effectue un test sur la signification du dernier retard de l'équation. Si le dernier retard n'est pas significatif, on le retranche. On estime à nouveau et on effectue un test sur la signification du dernier retard. On continue cette procédure jusqu'à trouver un retard significatif. On peut ensuite effectuer le test de racine unité.

6.2.2 Régression fictive (Spurious Regression)

On a le modèle suivant:

$$y_t = \alpha_0 + \beta x_t + \varepsilon_t$$

Les hypothèses habituelles sont:

- y_t et x_t sont stationnaires
- $E(\varepsilon_t) = 0$ et $var(\varepsilon_t) < \infty$

Si y_t et x_t sont des processus non stationnaires ($I(1)$), et que $\beta = 0$, on aura une régression fictive (Granger et Newbold (1974)).

Une régression fictive est caractérisée par (Phillips (1986)):

- L'estimateur $\hat{\beta}$ ne converge pas en probabilité vers une constante (zéro dans ce cas-ci) mais vers une variable aléatoire. L'incertitude ne s'estompe pas asymptotiquement.
- R^2 tend vers une variable aléatoire lorsque T tend vers l'infini.
- La statistique t_β diverge lorsque t tend vers l'infini
- $DW = \frac{\sum_{t=0}^{T}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})}{\sum_{t=0}^{T}\hat{\varepsilon}_t} \to 0$ en probabilité lorsque t tend vers l'infini.

Ainsi, ces résultats apparaissent intéressants selon les critères du R^2 et de la statistique t. Cependant, ils sont dénués de sens. L'estimateur des M.C.O. est biaisé et la loi assymptotique de $\hat{\beta}$ est non standard. La raison est que le terme d'erreur est non stationnaire. En effet, si $\beta = 0$, on a le modèle suivant:

$$y_t = \alpha_0 + \varepsilon_t.$$

Puisque y_t est $I(1)$, alors ε_t est $I(1)$. On aura également le même problème si y_t est $I(1)$ et x_t est $I(0)$. Le terme d'erreur sera alors non stationnaire.

Solution

On différencie les séries $I(1)$ pour les rendre stationnaires.

6.3 Cointégration

On a deux variables y_t et x_t qui sont intégrées d'ordre 1. Donc, la variancce de ces variables est infinie. Cependant, une combinaisons linéaire de ces deux variables peut être stationnaire.

Exemples:

- Revenu disponible et la consommation,
- Taux d'intérêt de long et de court terme
- Parité du pouvoir d'achat.

On a donc que y_t est $I(1)$ et x_t est $I(1)$. La cointégration implique la relation suivante de long terme

$$y_t = \mu + \beta x_t + \varepsilon_t$$

et le terme d'erreur est stationnaire $I(0)$). On peut réecrire de la façon suivante

$$y_t - \beta x_t = \mu + \varepsilon_t$$

On dira que $(1, -\beta)$ est le vecteur de cointégration. On peut également avoir de la cointégration entre plusieurs variables.

Definition 15 *Les séries $X_{jt}, j = 1, \cdots, m$ où X_t est intégré d'ordre d sont dites cointégrées si et seulement s'il existe une combinaison linéaire non nulle des séries qui est intégrée d'ordre strictement inférieur à d. On dira que*

$$X \sim CI(d, b)$$

où b est le degré de cointégration.

Cette combinaison linéaire est appelée vecteur de cointégration. On a le vecteur des séries X_t alors

$$\alpha' X_t = Z_t$$

sera stationnaire si les séries X_t sont $I(1)$ et cointégrées. Z_t peut contenir une constante ou une tendance déterministe . Cependant la variance de Z_t est finie.

Exemple

$$\begin{aligned} y_{1t} &= \lambda_1 y_{2t} + \mu_{1t} \text{ et } (1-\rho_1 L)\mu_{1t} = \varepsilon_{1t} \\ y_{2t} &= \lambda_2 y_{1t} + \mu_{2t} \text{ et } (1-\rho_2 L)\mu_{2t} = \varepsilon_{2t} \\ \text{et } \varepsilon_t &= (\varepsilon_{1t}, \varepsilon_{2t}) \text{ où } \varepsilon_t \sim BB(0, \Sigma) \end{aligned}$$

- Si $|\rho| < 1$ et $|\rho_2| < 2$, alors y_{1t} et y_{2t} sont $I(0)$.
- Si $\rho_i = 1$ et $\rho_j = 1$, alors y_{1t} et y_{2t} sont I(1).
- Si $\rho_i = 1$ $|\rho_j| < 1$, les deux variables sont $I(1)$, mais elles sont cointégrées.

par exemple, si $\rho_2 < 1$ et $\rho_1 = 1$ alors,

$$y_{2t} - \lambda_2 y_{1t} = \mu_{2t} = \rho_2 \mu_{2,t-1} + \varepsilon_{2t}$$

puisque $\rho_2 < 1$, alors la combinaison linéaire $(1, -\lambda_2)$ est stationnaire.

On doit estimer ces vecteurs de cointégration et tester s'il y a cointégration.

La cointégration implique un lien à long terme entre les séries cointégrées. Pour bien comprendre ceci, on va examiner l'exemple suivant:

$$\begin{aligned} y_t &= \mu_{yt} + \varepsilon_{yt} \text{ où } \varepsilon_{yt} \text{ est un bruit blanc} \\ x_t &= \mu_{xt} + \varepsilon_{xt} \text{ où } \varepsilon_{xt} \text{ est un bruit blanc} \end{aligned}$$

et

$$\begin{aligned} \mu_{yt} &= \mu_{yt-1} + \eta_{yt} \text{ où } \eta_{yt} \text{ est un bruit blanc} \\ \mu_{xt} &= \mu_{xt-1} + \eta_{xt} \text{ où } \eta_{xt} \text{ est un bruit blanc} \end{aligned}$$

alors,

$$\begin{aligned}\mu_{yt} &= \mu_{y0} + \sum_{i=1}^{t} \eta_{yi} \\ \mu_{xt} &= \mu_{x0} + \sum_{i=1}^{t} \eta_{xi}\end{aligned}$$

puisque μ_{yt} et μ_{xt} sont $I(1)$, y_t et x_t sont également $I(1)$.

On va maintenant supposer que les chocs ayant un effet permanent sur y_t et x_t sont liés par la relation suivante:

$$\eta_{y_t} = a\eta_{x_t}.$$

On va montrer que y_t et x_t sont alors cointégrées. En effet, on a que

$$\begin{aligned}y_t &= \mu_{y0} + \sum_{i=1}^{t} \eta_{yi} + \varepsilon_{y_t} \\ x_t &= \mu_{x0} + \sum_{i=1}^{t} \eta_{xi} + \varepsilon_{x_t}\end{aligned}$$

Si on prend comme vecteur de cointégration $(1, -a)$, on obtient

$$\begin{aligned}y_t - ax_t &= \mu_{y0} - a\mu_{x0} + \sum_{i=1}^{t} \eta_{yi} - a\sum_{i=1}^{t} \eta_{xi} + \varepsilon_{yt} - a\varepsilon_{xt} \\ y_t - ax_t &= \mu_{y0} - a\mu_{x0} + \varepsilon_{yt} - a\varepsilon_{xt} \\ y_t - ax_t &= \mu^* + \varepsilon_t^*\end{aligned}$$

et ε_t^* est $I(0)$. La combinaison linéaire $y_t - ax_t$ est donc stationnaire. Puisque le lien entre les chocs ayant un effet permanent sur y_t et x_t impique la cointgration pour ces deux séries, elles sont donc liées à long terme.

6.4 Test de cointégration (Engel et Granger)

On suppose que nous avons effectué un test de racine unité sur y_t et x_t et que nous ne pouvons rejeter l'hypothèse de racine unité (non stationnaire). On pense que y_t et x_t sont reliées à long terme. On va donc effectuer un test de cointégration.

On effectue un M.C.O. sur

$$y_t = \mu + \beta x_t + \varepsilon_t$$

et on fait un test de racine unité sur ε_t. Donc,

$$\Delta\varepsilon_t = \rho\varepsilon_{t-1} + \sum_{i=1}^{p} \delta_i \Delta\varepsilon_{t-i} + \mu_t$$

$$\begin{aligned} H_0 &: \quad \rho = 0 \\ H_1 &: \quad \rho < 0 \end{aligned}$$

La loi asymptotique est non standard et dépend du nombre de variables dans la première estimation par M.C.O. et de la présence d'une constante et une tendance dans cette régression.

À la première étape, on doit choisir la variable à gauche pour la régression. C'est une question de normalisation. De façon asymptotique, ce choix ne fait aucune différence. En petit échantillon, ce choix influence le résultat du test de cointégration. Ce test souffre du même problème que les tests de racine unité, il est peu puissant

Tests multivariés

- Test de Johansen (1991): Extension multivariées du test de Dickey-Fuller (dans un contexte de maximum de vraissemblance),
- Stock-Watson (1988),
- Phillips-Ouliaris (1990).

6.4.1 Estimation

Supposons un matrice Y de m variables cointégrées. On décompose Y comme étant

$$Y = (y_1 \;\; Y^*)$$

On peut estimer la relation de cointégration par une méthode simple d'estimation; M.C.O. (Engle et Granger 1987). On a donc

$$y_1 = X\beta + Y^*\alpha^* + u$$

où $X\beta$ contient les composantes déterministes (constante, tendance, $\cdots$)

Problèmes:

On va voir que l'estimateur des M.C.O. est un estimateur convergent (et même "superconvergent"), mais que cet estimateur peut avoir un biais important en petit chantillon. Ce biais provient de deux sources. De façon générale, si y_1 et Y^* sont cointégrées, ceci implique que ces séries se déterminent conjointement, alors on a

$$E(Y^{*'}u) \neq 0.$$

L'estimateur n'est donc pas sans biais en petit échantillon.

De plus, si le terme d'erreur est autocorrélé, ceci introduira également un biais en petit échantillon. Examinons, ce problème avec l'exemple suivant: On a la relation de cointégration suivante entre y_{1t} et y_{2t}

$$\begin{aligned} y_{1t} &= \alpha_1 y_{2t} + u_{1t} \;\; \text{où } u_{1t} = \rho_1 u_{1,t-1} + \varepsilon_{1t} \;\; \text{et} \;\; \rho_1 < 1 \\ y_{1t} &= \alpha_1 y_{2t} + \underbrace{\rho_1(y_{1,t-1} - \alpha_1 y_{2,t-1}) + \varepsilon_{1t}}_{v_t} \end{aligned}$$

Si on estime α_1 par les moindres carrés ordinaires, alors le terme d'erreur v_t est corrélés avec y_{2t}. En effet

$$E(y_{2t}v_t) = E\left(y_{2t}\left(\rho_1(y_{1,t-1} - \alpha_1 y_{2,t-1}) + \varepsilon_{1t}\right)\right) \neq 0$$

Cependant, de façon asymptotique, l'estimateur des moindres carrées ordinaires est convergent (et même "super convergent"). Pour le cas général, en supposant que $X\beta = 0$, on a l'estimateur des M.C.O.

$$\begin{aligned} \hat{\alpha^*} &= (Y^{*'}Y^*)^{-1}Y^{*'}y_1 \\ \rightarrow \hat{\alpha^*} - \alpha^* &= \underbrace{(Y^{*'}Y^*)^{-1}}_{Op(T^2)} \underbrace{Y^*u}_{Op(T)} \end{aligned}$$

Puisque le terme $(Y^{*'}Y^*)^{-1}$ augmente plus rapidement que Y^*u, de façon asymptotique $(\hat{\alpha^*} - \alpha^*)$ tend vers zéro même si Y^* est corrélé avec u. La vitesse de convergence est égale à T.

La loi asymptotique de $\hat{\alpha^*}$ est donnée par ce qui suit. On définit un vecteur de mouvement Brownien

$$B(r) = (B_1(r) \quad B_*(r)')' \text{ et } E(B(r)B(r)') = \Sigma$$

où $B_1(r)$ est un scalaire et $B_*(r)$ est un vecteur de même dimension que Y^*. On peut montrer que

$$T(\hat{\alpha^*} - \alpha^*) \Longrightarrow \left[\int_0^1 B_*(r)'B_r(r)dr\right]^{-1} \int_0^1 B_*(r)'dB_1(r)$$

Donc, de façon asymptotique, il n'y a pas de biais, cependant, en petit échantillon le biais peut être important.

Reprenons notre exemple avec autocorrélation du terme d'erreur.

$$y_{1t} = \alpha_1 y_{2t} + \underbrace{\rho_1(y_{1,t-1} - \alpha_1 y_{2,t-1}) + \varepsilon_{1t}}_{v_t}$$

En regressant y_{1t} sur y_{2t}, on ignore le terme $\rho_1(y_{1,t-1} - \alpha_1 y_{2,t-1})$. Ce terme est stationnaire $I(0)$ et y_{2t} est $I(1)$. On voit que ce terme n'influencera pas de façon asymptotique l'estimateur des M.C.O.. En petit échantillon, si ρ_1 est grand, alors l'estimateur M.C.O. aura un biais important.

On peut corriger les deux sources de biais en petit échantillon avec l' equation suivante:

$$y_{1t} = \beta' X_t + \alpha^{*\prime} Y_t^* + \sum_{j=-p}^{p} \gamma_j \Delta Y_{t-j}^* + \sum_{i=1}^{q} \rho_i (y_{1,t-i} - \alpha^{*\prime} Y_{t-i}^*) + \varepsilon_t.$$

Les termes additionnels de valeurs retardées et avancées de ΔY_t sont inclus pour corriger le biais d'endogénéité de Y_t et les termes de retards de la relation de cointégration sont introduit pour corriger le biais induit par l'autocorrélation du terme d'erreur. Ces deux corrections ont été proposées par Phillips et Loretan (1991) (voir aussi Stock-Watson 1993). En introduisant ces deux corrections, on peut utiliser les valeurs critiques standards pour effectuer des tests sur le vecteur α^*.

6.4.2 Modélisation avec des variables cointégrées

Modèle à correction d'erreurs: basé sur une équation

On cherche à modéliser la dynamique d'une variable en tenant compte de la relation de cointégration. On comprend de façon intuitive que la dynamique de la variable sera influencée par la relation de cointégration. La stratégie consiste à modéliser la variable I(1) en différence de telle sorte qu'elle soit stationnaire et à introduire la relation de cointégration comme variable explicative. On a ainsi ce qu'on appelle un MODÈLE À CORRECTION D'ERREURS. Ce modèle est donnée par l'équation suivante pour le vecteur $Y_t = (y_{1t} \quad Y_t^{*\prime})$,

$$\Delta y_{1t} = \mu + \beta\alpha Y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta Y_{t-j} + \varepsilon_t$$

où $\alpha = (1 \quad -\alpha^{*\prime})'$. Cette représentation inclut seulement des variables stationnaires si le vecteur de cointégration est connue. On peut donc effectuer de l'inférence de façon habituelle. Si le vecteur de cointégration n'est pas connue, on peut procéder en deux étapes:

1. On effectue une régression pour obtenir un estimé de $\hat{\alpha}$.

2. On estime ensuite la représentation à correction d'erreurs en remplaçant α par $\hat{\alpha}$.

6.5 Contexte multivarié

6.5.1 Représentation vectorielle autorégressive VAR

On suppose un vecteur Y_t de dimension $(m \times 1)$. Alors, un VAR(P) sera

$$Y_t = \alpha + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + u_t.$$

On peut réecrire ce système d'équations de la façon suivante en utilisant l'opérateur de retard L:

$$Y_t = \alpha + \phi(L) Y_{t-1} + U_t$$

C'est un système ayant la forme S.U.R.E. avec les mêmes variables explicatives à droite. On peut donc estimer équation par équation à l'aide des moindres carrés ordinaires.

- **Inconvénient d'un VAR:** beaucoup de paramètres à estimer. Il y a $m + Pm^2$ paramètres à estimer.
- **Avantage d'un VAR:**
  - facile à estimer (M.C.O. sur chaque équation)
  - on n'a pas à choisir quelles variables sont endogènes ou exogènes

On peut effectuer des tests de causalité au sens de "Granger". On se pose la question suivante: est-ce que les retards de la variable Y_{jt} aident à prédire la variable Y_{it}?

Definition 16 *On dit que Y_{jt} " cause au sens de Granger " Y_{it} si et seulement si*

$$E(Y_{it}/Y_{i,-1}, Y_{j,-1}) \neq E(Y_{it}/Y_{i,-1})$$

ou Y_{jt} ne "cause pas au sens de Granger " Y_{it} si et seulement si

$$E(Y_{it}/Y_{i,-1}, Y_{j,-1}) = E(Y_{it}/Y_{i,-1})$$

où $Y_{i,-1}, Y_{j,-1}$ sont des vecteurs contenant les variables respectives retardées.

On effectue un test conjoint de type F sur les coefficients des retards de Y_{jt} dans l'équation de Y_{it}. On connaît seulement la loi asymptotique du test. De plus, le résultat dépend des variables incluses dans le VAR.

Représentation autorégressive à correction d'erreurs (VECM)

On veut modéliser un vecteur Y_t de m variables. On suppose le VAR suivant:

$$\begin{aligned} Y_t &= \Pi_1 Y_{t-1} + \Pi_2 Y_{t-2} + \cdots + \Pi_p Y_{t-p} + u_t \\ \Pi(L) Y_t &= u_t \end{aligned}$$

On peut réecrire ce VAR de la façon suivante:

$$\begin{aligned} \Delta Y_t &= \Pi Y_{t-1} + \Gamma_1 \Delta Y_{t-1} + \Gamma_2 \Delta Y_{t-2} \cdots + \Gamma_{p-1} \Delta Y_{t-p-1} + u_t \\ \text{où } \Pi &= \Pi_1 + \Pi_2 + \cdots + \Pi_p - 1 \\ \text{et } \Gamma_i &= -\sum_{j=i+1}^{p} \Pi_i \end{aligned}$$

où ΠY_{t-1} est stationnaire.

Il y a trois cas possibles:

1. Π a un rang égal à zéro; les variables sont $I(1)$ et il n'existe pas de cointégration.

2. rang$(\Pi) = r < m$; les variables sont I(1) mais il existe r relations de cointégration entre les variables.

3. Π est de rang complet, les variables sont alors stationnaires $I(0)$.

Si rang(Π) est r et que $0 < r < m$, alors on peut réecrire Π de la façon suivante:

$$\Pi = \beta\alpha'$$

où

- α contient les vecteurs de cointégration, c'est une matrice de dimension $m \times r$.
- β est la matrice des paramètres d'ajustement vers l'équilibre de long terme. β est une matrice de dimension $m \times r$.

β, α ne sont pas identifiables à partir de l'estimé de la matrice Π, en effet

$$\Pi = \beta\alpha' = \beta\Gamma\Gamma^{-1}\alpha^* = \beta^*\alpha^{*'}$$

On peut maintenant énoncer le théorème de représentation d'Enger et Granger (1987)

Théorème 5

Les séries $I(1)$ $\{Y_t\}$ ne sont pas cointégrées si, et seulement si, le modèle s'écrit

$$\Pi(L)\Delta Y_t = u_t$$

où $\Pi(L)$ est un polynôme de degré $p-1$ tel que les racines de $|\Pi(L)|$ sont à l'extérieur du disque unité.

Si les séries Y_t sont cointégrées et si α' est une matrice $(r \times m)$ dont les lignes sont des vecteurs de cointégration indépendants, le modèle admet une représentation à correction d'erreurs du type

$$\Delta Y_t = \Pi Y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta Y_{t-i} + u_t$$

On doit donc inlure le terme de correction d'erreurs ΠY_{t-1} si il y a cointégration, sinon on est dans un cas d'omission de variables explicatives, ce qui entraîne un biais des estimateurs.

En présence de variables $I(1)$ et de relations de cointégration, on peut également estimer le VAR en niveau

$$Y_t = \Pi_1 Y_{t-1} + \Pi_2 Y_{t-2} + \cdots + \Pi_p Y_{t-p} + u_t$$

Implications:

- Les estimateurs sont convergents ou super-convegents ($\sqrt{T}$ ou T),
- La loi asymptotique des estimateurs n'est pas toujours normale (voir Sims, Stock et Watson, 1990)

En particulier, certains coefficients estimés ont une loi asymptotique fonction de mouvements Browniens.

7 La méthode des moments généralisés

La méthode des moments généralisés (GMM) consiste à estimer des paramètres d'intérêt à l'aide de conditions de moments appelées également conditions d'orthogonalité. L'estimateur est obtenu à l'aide des conditions de moments empiriques correspondantes aux conditions de moments théoriques.

Prenons, par exemple, une variable y_t où $t = 1, \cdots, T$. Le premier moment est donnée par:

$$E(y_t) = \mu.$$

Le scalaire μ est donc la moyenne non conditionnelle de la loi des y_t. On obtient un estimateur de μ par l' equivalent empirique de la condition de moment théorique. Ainsi,

$$E(y_t - \mu) = 0$$

et l'estimateur de la méthode des moments est:

$$\frac{1}{T}\sum_{T=1}^{T}(y_t - \hat{\mu}) = 0 \Rightarrow \hat{\mu} = \frac{1}{T}\sum_{T=1}^{T} y_t.$$

L'estimateur $\hat{\mu}$ satisfait la condition de moment empirique.

De la même façon, la variance théorique de y_t est donnée par

$$E(y_t - \mu)^2 = \sigma^2.$$

Cette condition de moment peut naturellement être réécrite comme étant:

$$E((y_t - \mu)^2 - \sigma^2) = 0.$$

L'estimateur de la méthode des moments est obtenu en égalisant à zéro la condition de moment empirique correspondante,

$$\frac{1}{T}\sum_{T=1}^{T}\left[(y_t-\hat{\mu})^2-\hat{\sigma}^2\right]=0$$

$$\Rightarrow \hat{\sigma}^2=\frac{1}{T}\sum_{T=1}^{T}(y_t-\hat{\mu})^2$$

Examinons maintenant le modèle linéaire suivant:

$$y_t=\beta' x_t+\varepsilon_t$$

où β est un vecteur de paramètres de dimension $K\times 1$, x_t est un vecteur de variables explicatives aléatoires de dimension $K\times 1$, $E(\varepsilon_t)=0$, $E(\varepsilon_t\varepsilon_{t-j})=\sigma^2$ pour $j=0$ et 0 autrement (bruit blanc faible).

Pour que l'estimateur des moindres carrés ordinaires soit sans biais, il doit respecter la condition suivante:

$$E(\varepsilon_t/x_t)=0,$$

qui correspond aux conditions de moments suivantes:

$$E(x_t\varepsilon_t)=0.$$

On a donc ici K conditions de moments.

L'estimateur de la méthode des moments généralisés est celui qui égalise les conditions de moments empiriques à zéro,

$$\frac{1}{T}\sum_{t=1}^{T}x_t\hat{\varepsilon}_t=\frac{1}{T}\sum_{t=1}^{T}x_t(y_t-\hat{\beta}'x_t)=0.$$

On obtient ainsi l'estimateur de la méthodes des moments généralisés

$$\hat{\beta}=(X'X)^{-1}X'Y$$

écrit de façon matricielle où X est une matrice $T \times K$ contenant les variables explicatives et Y est le vecteur contenant les observations de la variables dépendantes. Cet estimateur correspond à l'estimateur des moindres carrés ordinaires.

Si $E(x_t\varepsilon_t) \neq 0$, on peut utiliser un vecteur de variables instrumentales z_t de dimension égal ou plus grande que K, tel que $E(x_t z_t') \neq 0$ et $E(z_t\varepsilon_t) = 0$.

L'estimateur à variables instrumentales sera sans biais s'il respecte les conditions d'orthogonalité suivantes:

$$E(z_t\varepsilon_t) = 0.$$

L'estimateur est donné par le vecteur de paramètres qui égalise les conditions de moments empiriques à zéro. Ainsi,

$$\frac{1}{T}\sum_{t=1}^{T} z_t\hat{\varepsilon}_t = \frac{1}{T}\sum_{t=1}^{T} z_t(y_t - \hat{\beta}' x_t) = 0.$$

On peut réécrire les conditions de moments de façon matricielle:

$$\frac{1}{T}Z'\hat{\varepsilon} = \frac{1}{T}Z'(Y - X\hat{\beta}) = 0$$

où Z est la matrice contenant les variables instrumentales de dimension $T \times Q$.

Si le vecteur de variables instrumentales z_t a la même dimension que le vecteur x_t, alors l'estimateur de la méthode des moments généralisés (estimateur à variables instrumentales) est donné par

$$\hat{\beta}_{V.I} = (Z'X)^{-1}Z'Y.$$

Cet estimateur est également connu comme étant l'estimateur des moindres carrés indirects.

Lorsque le nombre d'instruments est plus grand que le nombre de paramètres d'intérêt, on peut utiliser une combinaison des instruments telle que

$$E(AZ'\varepsilon) = 0$$

où

$$E(Z^{*\prime}\varepsilon)$$

et $Z^* = ZA'$, A est de dimension $K \times Q$ et de rang égal à K.

L'estimateur est alors donnée par

$$\frac{1}{T}AZ'\hat{\varepsilon} = \frac{1}{T}AZ'(Y - X\hat{\beta}) = 0,$$

alors

$$\hat{\beta}_{V.I} = (AZ'X)^{-1} AZ'Y.$$

On a autant d'estimateur qu'il existe de matrice A. Il y a donc une infinité d'estimateur. On peut montrer que l'estimateur optimal est obtenu avec

$$A = X'Z(Z'Z)^{-1},$$

alors l'estimateur de la méthode des moments généralisés est:

$$\hat{\beta}_{V.I} = \left[X'Z(Z'Z)^{-1}Z'X\right]^{-1} X'Z(Z'Z)^{-1}Z'Y,$$

qui correspond à l'estimateur des doubles moindes carrés (Two stage least squares).

Soit un modèle général non linéaire:

$$Y = h(X, \beta) + \varepsilon$$

où $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_t)'$ et $E(\varepsilon) = 0, E(\varepsilon\varepsilon^1) = \Omega$ où Ω est une matrice definie positive.

On peut donc avoir de l'autocorrélation et/ou de l'hétéroscédasticité. Il est de plus possible que $E(X'\varepsilon) \neq 0$. Cependant, il existe un vecteur de variables instrumentales z_t de dimension q tel que $E(Z'\varepsilon) = 0$.

On peut donc utiliser les conditions de moments empiriques pour obtenir un estimateur de β, ainsi

$$\frac{1}{T}\sum_{t=1}^{T} z_t\hat{\varepsilon}_t = \frac{1}{T}\sum_{t=1}^{T} z_t(y_y - h(x_t\hat{\beta})) = 0.$$

Si $Q = K$, c.a.d., si la dimension du vecteur de variables instrumentales est égale à la dimension de β, alors on aura l'égalité à zéro et l'estimateur est unique. Si $Q > K$, on aura besoin d'une mesure de distance par rapport à zéro. On prendra une mesure quadratique. L'estimateur de la méthode des moments généralisés sera donné comme la solution du problème suivant:

$$\hat{\beta} = \arg\min \left(\frac{1}{T}\sum_{t=1}^{T} z_t\varepsilon_t\right)' W_T \left(\frac{1}{T}\sum_{t=1}^{T} z_t\varepsilon_t\right)$$

où W_T est une matrice définie positive qui peut dépendre des observations.

Il existe autant d'estimateur qu'il existe de matrice de pondération W. Hansen (1982) a montré que l'estimateur optimal est obtenu pour $W = S^{-1}$ où

$$S = \lim_{T\to\infty} Var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} z_t\varepsilon_t\right).$$

Pour le cas avec hétéroscédasticité seulement, on a que

$$S = \lim_{T\to\infty} \frac{1}{T}\sum_{i=1}^{T} \sigma_{ii} z_i z_i'.$$

White (1982) a proposé l'estimateur convergent suivant:

$$S_T = \frac{1}{T}\sum_{i=1}^{T} \hat{\varepsilon}_i^2 z_i z_i',$$

et il a montré que $S_T \xrightarrow{P} S$.

Lorsqu'il y a autocorrélation des erreurs, Newey-West (1987) ont démontré que l'estimateur suivant était convergent

$$S_T = \frac{1}{T}\sum_{l=0}^{r}\omega(l)\sum_{i=l}^{T}\hat{\varepsilon}_i\hat{\varepsilon}_{i-l}(z_i z'_{i-l} + z_{i-l}z'_i)$$

où $\omega(l) = 1 - \frac{l}{r+1}$ (fenêtre de Bartlett). On peut choisir r de façon endogène (Newy-West (1994)).

On peut maintenant examiner la méthode des moments généralisés dans un contexte général. De façon générale, on a les conditions de moments théoriques suivantes:

$$E\left[f(x_t, \theta_0)\right] = 0$$

où θ_0 est un vecteur de paramètres d'intérêt de dimension p. x_t est un vecteur de séries observées stationnaires et f est une fonction continue de dimension q où $q \geq p$.

$\hat{\theta}_T$ sera choisi tel que les conditions de moments empiriques sont le plus proche de zéro, c.a.d.

$$\frac{1}{T}\sum_{t=1}^{T} f(x_t, \theta).$$

Si $q = p$, alors on aura l'égalité à zéro, l'estimateur sera unique. Si $q > p$, on minimise une certaine mesure de distance par rapport à zéro.

Definition 17 *Étant donné une matrice symétrique définie positive W_T de dimension $q \times q$ dépendant éventuellement des observations, on appelle l'estimateur de la méthode des moments généralisés associé à W_T, une solution $\hat{\theta}_T(W_T)$ du problème*

$$\min_{\theta\in\Theta}\left(\frac{1}{T}\sum_{t=1}^{T} f(x_t, \theta)\right)' W_T \left(\frac{1}{T}\sum_{t=1}^{T} f(x_t, \theta)\right).$$

La matrice W_T mesure l'importance relative donnée aux conditions de moments.

On peut montrer que:

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \overset{loi}{\to} N\left(0, (D'WD)^{-1}D'WSWD(D'WD)^{-1}\right)$$

où

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\partial f(x_t, \hat{\theta}_T)}{\partial \theta'} \overset{p}{\to} D = E\left[\frac{\partial f(x_t, \theta_0)}{\partial \theta'}\right]$$

et $W_T \overset{p}{\to} W$, et

$$S = \lim_{T\to\infty} Var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f(x_t, \theta)\right).$$

Il existe autant d'estimateur qu'il existe de matrice W_T. Un estimateur optimal est obtenu avec la matrice $W_T = S_T^{-1}$. On a alors

$$\sqrt{T}(\hat{\theta_T} - \theta_0) \overset{loi}{\to} N\left(0, \left(D'S^{-1}D\right)^{-1}\right).$$

Pour ce choix de W_T, la matrice de variance-covariance est la plus petite possible.

Lorsque $q > p$, on obtient un test de spécification basé sur les conditions de suridentification. Ce test est donné par la statistique suivante:

$$J_T = T\left(\frac{1}{T}\sum_{t=1}^{T} f(x_t, \hat{\theta}_T)\right)' W_T \left(\frac{1}{T}\sum_{t=1}^{T} f(x_t, \hat{\theta})\right)$$

et suit une loi du $\chi^2(q-p)$.

La méthode des moments généralisés est une méthode en deux étapes.

1. À la première étape, on estime θ pour une matrice W quelconque. Habituellement, on utilise la matrice identité. Puisque l'estimateur obtenu est convergent, on peut l'utiliser pour construire un estimateur convergent de S.

2. Ayant obtenu un estimateur convergent de S, on réestime θ avec $W_T = S_T^{-1}$. Ainsi, on obtient un estimateur optimal.

En petit échantillon, il est plutôt préférable d'itérer plusieurs fois. De façon asymptotique, la procédure en deux étapes est suffisante pour obtenir un estimateur optimal.

8 Modèles à variables dépendantes qualitatives

Ref: Johnston et Dinardo, chap 13.

Les données statistiques ont souvent un caractère qualitatif.

Exemples:

- catégorie socio-professionnelle
- le fait de travailler ou non
- acheter ou ne pas acheter un produit
- choisir un mode de transport

Problèmes

Absence de continuité et souvent on a absence d'ordre naturel entre les modalités que peut prendre le caractère qualitatif.

De façon générale, la variable dépendante y peut prendre $K+1$ modalités tel que $K = 0, 1, \cdots, K$. On dira que si

$K + 1 = 2$: la variable est dichotomique

$K + 1 = 3$: la variable est trichotomique

$K + 1 > 3$: la variable est polytomique

8.1 Le modèle dichotomique simple

On suppose que la variable dépendante y est dichotomique. Les deux modalités qu'elle peut prendre sont par convention codées 0 et 1.

Dans un premier temps, on va chercher à bien comprendre les différences entre modèles qualitatifs et modèles quantitatifs.

On suppose N observations y_i et un vecteur de K variables exogènes $x_i = (x_{ii}, x_{2i}, \cdots, x_{ki})'$

Question: Peut-on utiliser le modèle linéaire?

Le modèle linéaire s'écrirait:

$$y_i = \beta' x_i + u_i$$

pour $i = 1, \cdots, N$.

Inadéquation d'une telle formulation peut être facilement comprise par des arguments intuitifs et par des arguments mathématiques.

Arguments intuitifs

- Les deux membres de l'égalité sont de nature différente: y_i est une variable qualitative et $\beta' xi + u_i$ est une variable quantitative.
- La valeur y_i est codée de façon arbitraire comme étant 0 et 1. La valeur de β dépendra de ce codage. Par exemple, si le codage était (0,2), alors on aurait 2β. La valeur de paramètre $\hat{\beta}$ est donc non interprétable.
- Le graphique des observations montre bien que l'approximation linéaire est peu adaptée au problème. En effet, on peut difficilement bien approximer ce nuage par une droite.

Arguments mathématiques:

- Comme y_i ne peut prendre que deux valeurs (0 et 1), il en est de même du terme d'erreur u_i. Ainsi, u_i prend la valeur $1-\beta' x_i$ avec une probabilité p_i et prend la valeur $-\beta' x_i$ avec une probabilité $1-p_i$. Le terme d'erreur u_i admet obligatoirement une loi discrète, ce qui interdit de faire l'hypothèse de normalité.

- Si nous imposons aux termes d'erreur d'être de moyenne nulle, p_i est déterminée de manière unique

$$E(u_i) = p_i(1 - \beta' x_i) - (1 - p_i)\beta' x_i = 0.$$

  Ceci implique que $p_i = \beta' x_i$. β sera alors contraint puisque p_i est une probabilité donc comprise entre 0 et 1. On aura donc que $0 \leq \beta' x_i \leq 1$. Ces contraintes peuvent être non compatibles.

- La variance de u_i vaut:

$$V(u_i) = \beta' x_i * (1 - \beta' x_i).$$

  On a donc hétéroscédasticité des erreurs. Cependant, on ne peut appliquer les M.C.O. ou M.C.G. car la variance dépend de β.

Spécification du modèle dichotomique

Prenons un exemple: un individu choisit de travailler (y=1) ou de ne pas travailler (y=0). On dénote l'utilité retirée de son choix par U_{i1}et U_{i0}. L'utilité dépend de certaines variables comme par exemple: l'âge, l'éducation, le nombre d'enfants, sa richesse etc. On aura donc:

$$U_{i0} = \alpha_0 + \gamma_0' W_i + \varepsilon_{i0}$$
$$U_{i1} = \alpha_1 + \gamma_1' W_i + \varepsilon_{i1}$$

Ainsi,

$$y_i = 1 \qquad si \qquad U_{i1} > U_{i0}$$

et

$$y_i = 0 \qquad si \qquad U_{i0} > U_{i1}$$

Examinons maintenant les probabilités des évènements $y_i = 1$ et $y_i = 2$,

$$\begin{aligned} Prob(y_i = 1) &= Prob(U_{i1} > U_{i0}) \\ &= Prob\left[(\varepsilon_{i0} - \varepsilon_{i1} < (\alpha_1 - \alpha_0) + (\gamma_1 - \gamma_0)' W_i\right] \\ &= F(\beta' x_i) = Prob(\varepsilon_{i0} - \varepsilon_{i1} < \beta' x_i) \end{aligned}$$

où $x_i = (1, W_i')'$ et $\beta = ((\alpha_1 - \alpha_0), (\gamma_1 - \gamma_0)')'$ et $Prob(y_i = 0) = 1 - F(\beta' x_i)$

Dans le cas du modèle linéaire, on aurait

$$F(\beta' x_i) = \beta' x_i$$

En effet,

$$\begin{aligned} E(y_i/x_i) &= 0\left[1 - F(\beta' x_i)\right] + 1\left[F(\beta' x_i)\right] \\ &= F(\beta' x_i), \end{aligned}$$

ce qui nous donne que $F(\beta' x_i) = \beta' x_i$ puisque

$$\begin{aligned} y_i &= E(y_i/x_i) + y_i - E(y_i/x_i) \\ &= \beta' x_i + u_i \end{aligned}$$

On a vu que ce modèle est insatisfaisant. En effet, $F(\beta' x_i)$ est une fonction de répartition et doit donc être comprise entre 0 et 1.

On peut utiliser la fonction de répartition de la loi normale. Ainsi,

$$Prob(y_i = 1) = \int_{-\infty}^{\beta' xi} \phi(t)dt = \Phi(\beta' x_i) = p_i$$

où $\phi(t)$ est la densité d'une loi normale. On appelle cette spécification le modèle Probit.

On utilise également la fonction de répartition de la loi logistique pour des raisons de manipulation mathématique. Ainsi,

$$Prob(y_i = 1) = \frac{1}{1 + e^{-\beta' x_i}} = p_i$$

On va souvent modéliser la variable qualitative à l'aide d'un seul inobservable y_i^*. Ce seuil sera fonction de variables explicatives. La variable inobservable pourra alors être modélisée par une relation linéaire telle que,

$$y_i^* = \beta' x_i + u_i$$

La variable qualitative observée est définie à partir de cette variable inobservable. On aura

$$y_i = 0 \qquad si \qquad y_i^* > l_i$$

et

$$y_i = 1 \qquad si \qquad y_i^* < l_i$$

Cherchons maintenant la loi de y_i. On suppose que les termes d'erreur sont indépendants de moyenne nulle et que $\frac{\mu_i}{\sigma}$ suit une certaine loi de répartition F. Alors,

$$\begin{aligned} Prob\left[y_i = 1\right] &= Prob\left[y_i^* < l_i\right] = Prob\left[\beta' x_i + u_i < l_i\right] \\ &= Prob\left[\frac{\mu_i}{\sigma} < \frac{l_i}{\sigma} - \frac{\beta' x_i}{\sigma}\right] = F\left[\frac{l_i}{\sigma} - \frac{\beta' x_i}{\sigma}\right] = p_i. \end{aligned}$$

Par l'hypothèse d'indépendance, on peut écrire la vraisemblance. Ainsi,

$$\begin{aligned} L(Y;\beta,\sigma) &= \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \\ &= \prod_{i=1}^{n}\left[F\left(\frac{l_i}{\sigma} - \frac{\beta' x_i}{\sigma}\right)^{y_i}\left(1 - F\left(\frac{l_i}{\sigma} - \frac{\beta' x_i}{\sigma}\right)\right)^{1-y_i}\right]. \end{aligned}$$

On réécrit

$$L(Y;\theta) = \prod_{i=1}^{n}\left[F(z_i,\theta)^{y_i}(1 - F(z_i,\theta))^{1-y_i}\right]$$

où $\theta = (\frac{1}{\sigma}, \frac{\beta'}{\sigma})'$

On suppose par la suite une loi de répartition, et on maximise la log-vraisemblance pour obtenir les estimateurs.