1 LA RÉGRESSION MULTIPLE

1.1 Le modèle linéaire

Le modèle linéaire est défini comme étant :

$$y_n = \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_K x_{nK} + \epsilon_n, \quad n = 1, \cdots, N \tag{1}$$

où y_n représente le vecteur des variables endogènes, x_{ni} sont les variables exogènes i pour $i = 1, \cdots, K$ 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
4. Analyse de politiques économiques
5. Analyse de politiques sociales

Le modèle (1) peu être récrit sous la forme suivante :

$$Y = x_1\beta_1 + x_2\beta_2 + \cdots + x_K\beta_K + \epsilon,$$

où le vecteur Y contient les observations y_n, les vecteurs x_i les observations des variables explicatives x_i tel que $x_i = (x_{1i}, x_{2i}, x_{3i}, \cdots, x_{ni})'$ et ϵ le vecteur contenant les termes d'erreur pour toutes les observations.
On peut réecrire 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

$$Y = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{N\times 1}, \quad X = \underbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{iK} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{bmatrix}}_{N\times K}, \quad \beta = \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}}_{K\times 1}, \quad \varepsilon = \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}}_{N\times 1}$$

1.1.1 Hypothèses du modèle linéaire

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

$$y_n = A x_n^{\beta} \exp(\epsilon_n). \tag{3}$$

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

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

On peut également considérer l'exemple suivant :

$$y_n \quad = \quad \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2}^2 + \epsilon_n, \tag{5}$$

2- X est contient des variables aléatoires ou fixes. Lorsque les variables sont fixes, pour des échantillons répétés, X prend toujours la même valeur.
Exemple pour X **aléatoire** : le revenu des ménages varie selon l'échantillon

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 :

$$\underbrace{y_n}_{N\times 1} = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \epsilon_n. \tag{6}$$

On suppose que

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

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

$$\begin{aligned} y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \epsilon_n, \\ y_n &= \beta_0 + \underbrace{(\beta_1 + c\beta_2)}_{\beta_1^*} x_{n1} + \epsilon_n, \\ y_n &= \beta_0 + \beta_1^* x_{n2} + \epsilon_n, \end{aligned}$$

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

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

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

On a donc,

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

La nullité de l'espérance conditionnelle entraîne celle de l'espérance non conditionnelle puisque

$$E(\epsilon) = E_X\left[E(\epsilon/X)\right] = E_X(0) = 0$$

par la loi des projections itérées. De plus, l'hypothèse de la nullité de l'espérance conditionnelle implique également que $E(X'\epsilon) = 0$ mais l'inverse n'est pas vrai.

Loi des projections itérées :

Supposons deux variables aléatoires z et w dont la loi est sur un support continu et celle-ci est bornée. On aura alors

$$E_w\left[E(z/w)\right] = E(z).$$

Preuve : Par définition,

$$E_w\left[E(z/w)\right] = \int_w \left[\int_z z f(z/w) dz\right] f(w) dw.$$

On sait que $f(z/w) = \frac{f(z,w)}{f(w)}$. On peut donc réécrire de la façon suivante :

$$\begin{aligned}
\int_w \left[\int_z z f(z/w) dz\right] f(w) dw &= \int_w \int_z z f(z/w) f(w) dz dw \\
&= \int_w \int_z z f(z,w) dz dw \\
&= \int_z z \left[\int_w f(z,w) dw\right] dz \\
&= \int_z z f(z) dz = E(z),
\end{aligned}$$

puisque $\int_w f(z,w) dw = f(z)$.

Dans un contexte de série temporelle, la deuxième hypothèse implique que l'espérance de chaque ϵ_i conditionnellement à toutes ls observations de la matrice X est nulle. 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}.

Sous l'hypothèse que $E(\epsilon/X) = 0$, on a donc que $E(\epsilon) = 0$ ce qui implique que l'espérance marginale de Y est

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

5-La matrice de variance-covariance conditionnnelle des termes d'erreurs a la forme générale suivante :

$$E(\underbrace{\epsilon}_{N\times 1}\underbrace{\epsilon'}_{1\times N}/X) = \sigma^2 \underbrace{\Omega}_{N\times N}, \tag{9}$$

puisque

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

ceci implique que

$$E(\epsilon\epsilon'/X) = var(\epsilon/X).$$

Par la loi des projections itérées

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

On a donc,

$$E(\epsilon\epsilon'/X)\begin{bmatrix} var(\epsilon_1/X) & cov(\epsilon_1,\epsilon_2/X) & \cdots & cov(\epsilon_1,\epsilon_N/X) \\ cov(\epsilon_2\epsilon_1/X) & var(\epsilon_2/X) & \cdots & cov(\epsilon_2,\epsilon_N/X) \\ \vdots & \vdots & \vdots & \vdots \\ cov(\epsilon_N\epsilon_1/X) & cov(\epsilon_N,\epsilon_2/X) & \cdots & var(\epsilon_N/X) \end{bmatrix}$$

En présence d'homoscédasticité, la matrice de variance-covariance conditionnelle et non conditionnelle est donnée par $\sigma^2 I_N$ où I_N est une matrice identité de dimension $N \times N$. Cela implique que

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

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

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

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

on a alors hétéroscédasticité et/ou autocorrélation des erreurs.

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.

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

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

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

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,

$$\begin{aligned} S &= \sum_{n=1}^{N}(y_n - \beta_1 x_{n1} - \beta_2 x_{n2} - ... - \beta_K x_{nK})^2, \\ &= \underbrace{(Y - X\beta)'}_{1\times N}\underbrace{(Y - X\beta)}_{N\times 1} \end{aligned} \tag{10}$$

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 en présence d'homoscédasticité.

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.

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. Pour ce faire 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 $E\left(g(w,z)\right)$.

$$\begin{aligned} Eg(w,z) &= \int_z \int_w g(w,z) f(w,z) dw\, dw \\ Eg(w,z) &= \int_z \int_w g(w,z) f(w/z) f(z) dw\, dz \\ Eg(w,z) &= \int_z \left[\int_w g(w,z) f(w/z) dw \right] f(z) dz \\ Eg(w,z) &= E_z E_{w/z}(g(w,z)) \end{aligned}$$

Appliquons maintenant ce résultat. On a que

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

On substitue Y dans cette expression,

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

Ainsi,

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

On prend maintenant l'espérance,

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

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

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

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

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

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

et de façon conditionnelle

$$Var(\hat{\beta}/X) \quad = \quad \sigma^2 \left[(X'X)^{-1}X'\Omega X(X'X)^{-1}\right]$$

On va maintenant s'intéresser au cas avec erreurs homoscédastiques. Cela nous permettra de comparer la variance de différents estimateurs. On a donc que $E(\epsilon\epsilon') = \sigma^2 I_N$.

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 en présence d'homoscédastiscité parmi les estimateurs linéaires sans biais et il a pour variance*

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

Preuve

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

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

On a déjà montré que la matrice de variance-covariance conditionnelle de $\hat{\beta}$ est, de façon générale, donnée par l'expression suivante :

$$Var(\hat{\beta}/X) \quad = \quad \sigma^2 \left[(X'X)^{-1}X'\Omega X(X'X)^{-1}\right].$$

Pour le cas avec homoscédasticité, $\Omega = I_N$, ceci donne

$$Var\left(\hat{\beta}/X\right) = \sigma^2(X'X)^{-1}$$

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

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

ainsi,

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

et

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

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

$$\begin{aligned} var(\beta^*/X) &= var\left[(\beta^* - \hat{\beta} + \hat{\beta})/X\right] \\ &= var\left[(\beta^* - \hat{\beta})/X\right] + 2cov\left[(\beta^* - \hat{\beta}, \hat{\beta})/X\right] + var(\hat{\beta}/X). \end{aligned}$$

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

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

puisque $CX = I$.

On a donc que :

$$var(\beta^*/X) = var\left[(\beta^* - \hat{\beta})/X\right] + var(\hat{\beta}/X).$$

La matrice $var\left[(\beta^* - \hat{\beta})/X\right]$ étant semi définie positive, la matrice de variance-covariance conditionnelle de β^* est donc plus grande ou égale à la matrice de variance-covariance conditionnelle 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. Sous l'hypothèse d'homoscédasticité :

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

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,

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

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

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

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. $\bar{Y} = \bar{X}\hat{\beta}$ (puisque $i'Y = i'X\hat{\beta} + i'\hat{\epsilon}$).

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

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,

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

où P_xY représente ce qui est expliqué dans l'espace engendré par les X et M_xY 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 $\hat{\beta}_2$. 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,

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

La matrice X_2^* correspond aux résidus de la régression des vecteurs colonnes de X_2 sur X_1 et Y^* aux résidus de la régression de Y sur X_1. En effet, prenons la colonne i de la matrice X_2 et posons :

$$X_{2i} = X_1\theta + u$$

$$\hat{\theta} = (X_1'X_1)^{-1}X_1'X_{2i}$$

$$\Rightarrow X_{2i} = X_1(X_1'X_1)^{-1}X_1'X_{2i} + \hat{u}$$

$$\Rightarrow X_{2i} - X_1(X_1'X_1)^{-1}X_1'X_{2i} = \hat{u}$$

$$[I - X_1(X_1'X_1)^{-1}X_1']X_{2i} = \hat{u}$$

$$M_1X_{2i} = \hat{u}$$

et ceci est valide pour chaque colonne i de la matrice X_2. On peut maintenant énoncer le théorème de Frisch-Waugh.

Théorème 2 (Théorème de Frisch-Waugh) *L'estimateur du sous-vecteur β_2 obtenu à l'aide de la régression par moindres carrés ordinaires du vecteur Y sur les ensembles de variables explicatives X_1 et X_2 correspond à la regression des résidus de la régression de Y sur X_1 sur les résidus de la régression des vecteurs colonnes de X_2 sur X_1.*

Le vecteur de paramètres $\hat{\beta}_2$ mesure donc ce qui provient exclusivement de X_2 (information orthogonale à X_1).

1.7 APPLICATIONS

1.7.1 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 :

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

.

1.7.2 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

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

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

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

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

1.7.3 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 $M_iX_1 = 0$ puisque $X_1 = i$ et $M_i\hat{\epsilon} = \hat{\epsilon}$ puisque $i'\hat{\epsilon} = 0$.
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 $Y'M_iY$ correspond a la variance empirique de Y lorsque ce terme est divisé par $N-1$, c.a.d. $(\frac{(Y-i\bar{Y})'(Y-i\bar{Y})}{N-1})$. L'expression pour le R^2 implique que

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

Pour que ce résultat tienne il faut absolument qu'un vecteur de 1 soit inclut dans la régression.

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

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

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

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

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

Quelques définitions et propriétés utiles

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 centrées $T(K)$ 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* $\chi^2(K_2)$*.*

1.8.1 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'erreurs. En plus des hypothèses du modèle linéaire classique, on va supposer que

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

$\hat{\beta}$ 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 :

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

Pour un élement du vecteur β, on a

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

Si on définit S^{kk} comme étant le k ième élément diagonale de $(X'X)^{-1}$, 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 ne 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

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

qui suit conditionnellement aux X une loi normale centrée réduite. 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.

1.8.2 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 $J \leq K$ 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

$$VAR(R\hat{\beta}/X) = E(R(\hat{\beta} - \beta))(\hat{\beta} - \beta)'R'/X)$$

$$= \sigma^2 R(X'X)^{-1}R'$$

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

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

alors

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

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

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

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

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

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.

1.8.3 DEUX APPROCHES POUR EFFECTUER DE L'INFÉRENCE

1.8.4 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 ?

1.8.5 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

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

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

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

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

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

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

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

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

(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

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

1.9 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

$$\hat{\epsilon}'\hat{\epsilon} = \begin{bmatrix} \hat{\epsilon}^{1\prime}\hat{\epsilon}^{2\prime} \end{bmatrix} \begin{bmatrix} \hat{\epsilon}^1 \\ \hat{\epsilon}^2 \end{bmatrix} = \hat{\epsilon}^{1\prime}\hat{\epsilon}^1 + \hat{\epsilon}^{2\prime}\hat{\epsilon}^2$$

$$\epsilon^{*\prime}\epsilon^* = \begin{bmatrix} \epsilon^{*1\prime}\epsilon^{*2\prime} \end{bmatrix} \begin{bmatrix} \epsilon^{*1\prime} \\ \epsilon^{*2\prime} \end{bmatrix} = \epsilon^{*1\prime}\epsilon^{*1} + \epsilon^{*2\prime}\epsilon^{*2}$$

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).

1.9.1 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 continue de dimension $J \times 1$ et qui ne dépend pas du nombre d'observations. La statistique du test est donnée par l'expression suivante :

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

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

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

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

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

On peut donc construire la statistique présentée plus haut.

2 Théorie en grand échantillon

2.1 Convergence en probabilité et convergence en loi

Sous l'hypothèse que $E(\epsilon/X) = 0$, $E(\epsilon\epsilon'/X = \sigma^2 I$ et $\epsilon/X \sim N(0, \sigma^2 I)$ on a des propriétés de petit échantillon. En effet, on connaît la loi conditionnelle exacte des estimateurs et des tests (T, F).

Si on abandonne l'hypothèse que le vecteur ϵ suit une loi conditionnelle 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

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 pour 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*

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

avec $n = 1, \ldots, N$.

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*

$$plim X_N = c$$

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

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

Exemples :

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

Pour les vecteurs, on a

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

et

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

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

$$plim_{N\to\infty}\frac{1}{N}X'X = E(\frac{X'X}{N}) = Q,$$

où Q est positive définie.

Alors

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

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

puisque

$$plimAB = plimA\, plimB$$

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

On cherche la probabilité limite du dernier terme.

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

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 une 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.

Reprenons notre exemple avec $\bar{X}_N$. 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

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

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

$$\lim_{N \to \infty} |F(Z_N) - F(Z)| = 0$$

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

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 4 (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.

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 :

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

on peut appliquer le théorème central limite à l'expression $\frac{1}{\sqrt{N}}X'\epsilon$. En effet, $x_{k1}\epsilon_1,\ldots,x_{kN}\epsilon_N$ est une suite de variables aléatoires pour toutes les variables explicatives $k=1,\ldots,K$ avec une certaine densité, une moyenne bornée et une variance finie.

On applique donc le théorème central limite, ainsi,

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

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

Alors,

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

On peut construire les statistiques de tests t et F

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

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

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

3 Estimation par maximum de vraisemblance

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.*

Cette définition est le pendant en grand échantillon de Gauss-Markov à la différence qu'il n'est pas restreint aux estimateurs linéaires.

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 où $X_N = \{x_1, x_2, \ldots, x_N\}$, c.a.d.

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

Exemple : Prenons une suite de variables aléatoires indépendantes x_i qui suivent une loi de Bernoulli :

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

pour $i = 1, \ldots, N$. On veut connaître le paramètre p. On observe la suite de réalisations suivantes

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

où $N = 10$. On va chercher p qui maximise la probabilité d'observer un tel échantillon. On aura la densité conjointe suivante :

$$f(X_N; p) = (1-p).\ p.\ p.\ p.\ (1-p)\ .p\ .p\ .(1-p)\ .p\ .p.$$

On peut écrire la densité conjointe de cette façon puisque ce sont des événements indépendants. On réécrit la densité conjointe,

$$L(p; X_N) = f(X_N; p) = p^{N_1}(1-p)^{N_2}$$

où N_1 est la nombre de fois que 0 est observée, N_2 est le nombre de fois que la valeur 1 est observée et $L(p; X_N)$ est appelé la vraisemblance.

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

$$\max_p\ lnL(p; X_N)$$

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

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,

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

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. En conditionnant sur X et par le fait que le modèle est linéaire, la densité de Y est directement fonction de la densité de ϵ.

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

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

Puisque les ϵ_t sont indépendants, alors

$$f(\epsilon; \sigma^2) = \prod_{n=1}^{N} f(\epsilon_n; \sigma^2).$$

De façon matricielle, on aura

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

Lorsque une variable aléatoire z suit une $N(\mu, \sigma^2)$, sa densité de probabilité est donnée par :

$$f(z; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(z-\mu)^2\right).$$

Puisque $Y = X\beta$, on a alors que $Y/X \sim N(X\beta, \sigma^2)$. Pour chaque observation y_n, la densité de probabilité conditionnelle est :

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

Puisque les densités de probabilité conditionnelles sont indépendantes, la densité de probabilité conjointe conditionnelle est alors la multiplication des densité de probabilité conditionnelle pour chaque y_n. Ainsi,

$$f(Y/X; \beta, \sigma^2) = \prod_{n=1}^{N} f(y_n/x_n; \beta, \sigma^2).$$

De façon matricielle, on aura

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

La vraisemblance de Y est donc

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

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

$$\hat{\beta}_{M.V.} = (X'X)^{-1}X'Y$$

et

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

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 5 (Le théorème de Cramer-Rao) *Pour tout estimateur $\hat{\theta}$, 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

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

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

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

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

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

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 $X \sim N(0, \sigma^2 I)$ 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 échantillons 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,

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

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 si $g(\theta)$ est continue et continuellement différentiable*

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^2. 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

$$\sqrt{N}\left(\hat{\theta} - \theta)\right) \overset{loi}{\to} N\left(0, \left(\frac{I(\theta)}{N}\right)^{-1}\right)$$

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

3.3 Tests d'hypothèses

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 $\hat{\theta}_{nc}$. La statistique de Wald est :

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

et

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

On estime donc le modèle non contraint et on construit la statistique W. La loi asymptotique est obtenue de la façon suivante : on effectue un développement limité de $C(\hat{\theta})$ autour de la vraie valeur θ. Ainsi :

$$C(\hat{\theta}_{nc}) = C(\theta) + \left(\frac{\partial C(\theta)}{\partial \theta'}\right) \left(\hat{\theta}_{nc} - \theta\right) + ...$$

ce qui donne sous la nulle et en prémultipliant par $\sqrt{N}$ que :

$$\sqrt{N}\left(C(\hat{\theta}_{nc}) - q\right) = \left(\frac{\partial C(\hat{\theta})}{\partial \theta'}\right) \sqrt{N}\left(\hat{\theta}_{nc} - \theta\right) + ...$$

Par la normalité asymptotique de $\sqrt{N}\left(\hat{\theta}_{nc} - \theta\right)$, on obtient le résultat désiré.

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. Le Lagrangien s'écrit :

$$\ln L^*(\theta) = \ln L(\theta) + \lambda'(C(\theta) - q).$$

On a les C.P.O. suivantes :

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

Si la contrainte n'est pas mordante, alors

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

On peut montrer que sous l'hypothèse nulle que

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

en appliquant le théorème central limite.

La statistique du multiplicateur de Lagrange est définie comme étant :

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

Cette statistique peut également s'écrire

$$LM = \hat{\lambda}' \frac{\partial C(\hat{\theta}_c)}{\partial \theta'} \left[I(\hat{\theta}_c)\right]^{-1} \left(\frac{\partial C(\hat{\theta}_c)}{\partial \theta'}\right)' \hat{\lambda} \xrightarrow{loi} \chi^2(J).$$

La statistique du multiplicatuer de Lagrange est également appelée statistique du *score*.

3. **Test du ratio de vraisemblance**

$$-2(\ln L(\hat{\theta}_c) - \ln L(\hat{\theta}_{nc}) \xrightarrow{loi} \chi^2(J)$$

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

3.3.1 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

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

   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

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

   Alors

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

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

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

   et puisque

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

   Alors, la statistique LM peut être réécrite comme étant

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

Ce qui donne,

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

3. Test du rapport de vraissemblance

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

Comparaison des trois statistiques de test :

On va utiliser la relation suivante :

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

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

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

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 variance-covariance 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

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

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

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

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, ce qui implique que $CC' = CC^{-1} = I$. De plus, on aura que

$$\Omega^{-1} = (C\Lambda C')^{-1} = (C')^{-1}\Lambda^{-1}C^{-1} = C\Lambda^{-1}C'.$$

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

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

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,

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

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

On a donc que

$$\hat{\beta}_{m.c.g.} = \left[X'\Omega^{-1}X\right]^{-1} X'\Omega^{-1}Y.$$

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

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

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^2 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

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

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 asymptotiquement.

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 :

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

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

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

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

$$\Sigma = \sigma^2 \frac{X'\Omega X}{N} = \frac{1}{N}\sum_{n=1}^{N} \sigma_n^2 x_n x_n'.$$

On estime Σ par

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

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 en 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. C'est ce que l'on appelle un estimateur robuste à l'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.

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

2. L'estimateur de White

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

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

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

et la variance de D_N est

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

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

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

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

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

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

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

On réecrit le R^2 comme étant

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

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

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

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,

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

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,

$$NR^2 \quad = \quad D_N(var(D_N))^{-1}D'_N$$

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

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

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 :

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

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-vraisemblance est

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

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

$$
\begin{aligned}
\frac{\partial \ln L(\tilde{\theta})}{\partial \alpha} &= \frac{1}{2} \left[\hat{\sigma}^{-2} \frac{\partial f(\hat{\alpha}_0)}{\partial h_n} \right] \sum_{n=1}^{N} z_n \left(\hat{\sigma}^{-2} \hat{\epsilon}_n^2 - 1 \right) \\
\frac{\partial \ln L(\tilde{\theta})}{\partial \alpha \partial \alpha'} &= I_{\alpha\alpha'}(\tilde{\theta}) = \frac{1}{2} \left[\hat{\sigma}^{-2} \frac{\partial f(\hat{\alpha}_0)}{\partial h_n} \right]^2 \sum_{n=1}^{N} z_n z_n
\end{aligned}
$$

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 $f(\cdot)$. On peut réécrire sous forme vectorielle

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

où ι est un vecteur de dimension N contenant des 1, et $Z = (z_1, z_2, \cdots, z_N)'$ est la matrice contenant les variable z_n pour toutes les observations et autres que la constante.

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 également effectuer le test en régressant $\hat{\epsilon}^2$ sur une constante et les variables Z et calculer la statistique NR^2 qui suit également une $\chi^2(p)$.

On peut effectuer également un test de type LR

$$LR = -2\left(\ln L(\tilde{\theta}/Y, X) - \ln L(\hat{\theta}/Y, X)\right) \overset{loi}{\rightarrow} \chi^2(p)$$

où $\tilde{\theta}$ est l'estimateur contraint (donc avec $\alpha_1 = \alpha_2 = \ldots = 0$ et $\hat{\theta}$ l'estimateur non contraint. 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 : l'hétéroscédasticité dépend de la variable X_i

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

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

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

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

$$\hat{\epsilon}^2 = X_i\alpha + u$$

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 : sous quelle condition 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,

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

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

On a donc que

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

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,

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

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 ?

$$
\begin{aligned}
E(X_t) &= E(\epsilon_t) + \theta E(\epsilon_{t-1}) = 0 \\
var(X_t) &= E\left[(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_t + \theta\epsilon_{t-1})\right] \\
&= \sigma^2 + \theta^2\sigma^2 = (1+\theta^2)\sigma^2 < \infty, \ \ \forall\theta \\
\gamma(1) &= cov(X_t X_{t-1}) = E\left[(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_{t-1} + \theta\epsilon_{t-2})\right] \\
&= \theta\sigma^2 \\
\gamma(2) &= cov(X_t X_{t-2}) = E\left[(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_{t-2} + \theta\epsilon_{t-3})\right] \\
&= 0 \\
\gamma(3) &= 0 \\
&\vdots \\
\gamma(k) &= 0, \ \text{pour } k > 1
\end{aligned}
$$

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

$$L\ X_t = X_{t-1}, \ \ L^n\ X_t = X_{t-n}$$

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)

$$X_t = \epsilon_t + \theta\epsilon_{t-1} = (1+\theta L)\epsilon_t$$

$$\frac{X_t}{(1+\theta L)} = \frac{X_t}{(1-(-\theta L))}.$$

Ce qui donne

$$\sum_{i=0}^{\infty}(-\theta)^i L^i X_t = \epsilon_t$$

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

$$\sigma^2\Omega = \frac{\sigma_\epsilon^2}{(1-\phi^2)} \begin{bmatrix} 1 & \phi & \phi^2 & \phi^3 & \cdots & \phi^{T-1} \\ \phi & 1 & \phi & \phi^2 & \cdots & \vdots \\ \phi^2 & \phi & 1 & \phi & \cdots & \vdots \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & & \ddots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \cdots & \phi & \cdots & 1 \end{bmatrix}$$

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 positive 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

$$y_t = \beta' x_t + \varepsilon_t \text{ et } \varepsilon_t = \rho\varepsilon_{t-1} + \mu_t$$

$$\begin{aligned} H_0 &: \quad \rho = 0 \\ H_1 &: \quad \rho \neq 0 \end{aligned}$$

Le test de Durbin-Watson est basé sur la statistique

$$d = \frac{\sum_{t=2}^{T}(\hat{\epsilon}_t - \hat{\epsilon}_{t-1})^2}{\sum_{t=1}^{T}\hat{\epsilon}_t^2} \approx 2(1-r)$$

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 $r \approx 1 \Rightarrow d = 0 \Rightarrow$ autocorrélation fortement positive.
Si $r \approx 0 \Rightarrow d = 2 \Rightarrow$ pas d'autocorrélation.
Si $r \approx -1 \Rightarrow d = 4 \Rightarrow$ 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. L'hypothèse $E(\epsilon/X) = 0$ doit être respectée. 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 :

$$TR^2 \overset{loi}{\rightarrow} \chi^2(p)$$

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 :

$$Q_T = T \sum_{j=1}^{L} \hat{r}_j^2 \overset{loi}{\rightarrow} \chi^2(L)$$

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 :

$$Y = X\beta + \epsilon \text{ où } \epsilon = \epsilon_{t-1}\rho + \mu$$

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 $t = 2, \cdots, T$

$$y_t = \beta' x_t + \epsilon_t, \text{ où } \epsilon_t = \rho\epsilon_{t-1} + u_t.$$

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

$$\ln L \quad = \quad \ln f(y_1) + \sum_{t=2}^{T} \ln f(y_t/y_{t-1})$$

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

$$\epsilon_t = u_t(\alpha_0 + \alpha_1\epsilon_{t-1}^2)^{\frac{1}{2}} \text{ et } u_t \sim N(0,1).$$

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

$$TR^2 \overset{loi}{\rightarrow} \chi^2(p)$$

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 Système d'équations et variables instrumentales

6.1 Systèmes d'équations non simultanées

On s'intéresse à un groupe d'équations :

$$\begin{aligned} y_1 &= X_1 \beta_1 + u_1 \\ y_2 &= X_2 \beta_2 + u_2 \\ &\vdots \\ y_M &= X_M \beta_M + u_M \end{aligned}$$

On aura donc M équations de T observations

6.1.1 Modèles de régressions apparamment non reliées

En anglais : seemingly unrelated regression model (SURE))

On réécrit les équations plus haut :

$$y_m = X_m\beta_m + u_m \quad \text{pour} \quad m = 1, \cdots, M$$

et

$$u = (u_1', u_2', \cdots, u_m')'$$

est un vecteur de dimension $(TM \times 1)$ avec $E(u) = 0$ et $E(uu') = V$.

On suppose qu'il n'y a pas de corrélation entre les erreurs à des périodes différentes. Ainsi,

$$\begin{aligned} E(u_{mt}u_{ns}) &= \sigma_{mn} \quad \text{si} \quad t = s \\ &= 0 \text{ autrement,} \end{aligned}$$

$$E(u_m u_n') = \sigma_{mn} I_T$$

$$E(uu') = V = \begin{bmatrix} \sigma_{11}I & \sigma_{12}I & \cdots & \sigma_{1M}I \\ \sigma_{21}I & \sigma_{22}I & \cdots & \sigma_{2M}I \\ \vdots & & & \\ \sigma_{M1}I & \sigma_{M2}I & \cdots & \sigma_{MM}I \end{bmatrix}$$

Chaque équation du système est un modèle linéaire classique. Cependant, les termes d'erreurs entre les équations sont reliées entre eux de façon contemporaine. L'estimateur des MCO est encore sans biais mais il n'est pas optimal. La méthode optimale est les moindres carrés généralisés. Quel est l'estimateur MCG ?

On réécrit le système :

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} X_1 & 0 & 0 & \cdots & 0 \\ 0 & X_2 & 0 & \cdots & 0 \\ 0 & 0 & X_3 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & X_M \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix}$$

Ce qui donne sous forme matricielle :

$$Y = X\beta + u$$

Pour une observation t, la matrice de variance-covariance (M $\times$ M) est :

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2M} \\ \vdots & & & \\ \sigma_{M1} & \sigma_{M2} & \cdots & \sigma_{MM} \end{bmatrix}$$

Alors $E(uu') = V = \Sigma \otimes I$ où $\otimes$ représente le produit kronecker et $V^{-1} = \Sigma^{-1} \otimes I$. L'estimateur des moindres carrés généralisés est :

$$\begin{aligned} \hat{\beta} &= (X'V^{-1}X)^{-1}X'V^{-1}Y \\ &= (X'(\Sigma^{-1} \otimes I)X)^{-1}X'(\Sigma^{-1} \otimes I)Y \\ &= \begin{bmatrix} \sigma^{11}X_1'X_1 & \sigma^{12}X_1'X_2 & \cdots & \sigma^{1M}X_1'X_M \\ \sigma^{21}X_2'X_1 & \sigma^{22}X_2'X_2 & & \sigma^{2M}X_2'X_M \\ \vdots & \vdots & & \vdots \\ \sigma^{M1}X_M'X_2 & \sigma^{M2}X_M'X_2 & & \sigma^{MM}X_M'X_M \end{bmatrix}^{-1} \begin{bmatrix} \sum_{m=1}^{M}\sigma^{1m} & X_1'Y_m \\ \sum_{m=1}^{M}\sigma^{2m} & X_2'Y_m \\ \vdots & \\ \sum_{m=1}^{M}\sigma^{Mm} & X_M'Y_m \end{bmatrix} \end{aligned}$$

où le ij élément de Σ^{-1} est σ^{ij}.

Comparaison entre l'estimateur des MCO et des MCG.

1- Si $\sigma_{mn} = 0$, aucun gain à utiliser l'estimateur MCG.

2- Si les variables explicatives sont les mêmes dans chaque équation, c'est-à-dire $X_m = X_n$, l'estimateur MCO est le même que l'estimateur MCG (à faire en exercice).

On peut dire de façon générale :

1- Plus la corrélation est grande entre les erreurs des différentes équations, plus le gain à utiliser l'estimateur des MCG est grand.

2- Moins il y a de corrélation entre les X_m des différentes équations, plus le gain a utiliser l'estimateur des MCG est grand.

Jusqu'ici, on a supposé que Σ est connu, ce qui est rarement le cas. Quoi faire si Σ est inconnu ?

On utilise les résidus de l'estimateur des MCO sur chaque équation (estimateur sans biais) et on estime de la façon suivante :

$$\hat{\sigma}_{mn} = \frac{\hat{u}'_m \hat{u}_n}{T}$$

Ceci est un estimateur convergent de σ_{mn}. On peut aussi construire un estimateur de la matrice Σ et effectuer les moindres carrés généralisés.

L'inférence s'effectue de façon habituelle (pas de propriété de petit échantillon). Pourquoi ?

Estimateur du maximum de vraisemblance

On suppose ici la normalité des erreurs. On maximise la log-vraisemblance suivante par rapport à β et Σ. Le log de la vraisemblance s'écrit alors

$$\begin{aligned} lnL &= -\frac{MT}{2} ln 2\pi - \frac{1}{2} ln|V| - \frac{1}{2} u' V^{-1} u \\ &= -\frac{MT}{2} ln 2\pi - \frac{1}{2} ln|\Sigma \otimes I| - \frac{1}{2} u' (\Sigma^{-1} \otimes I) u \\ &= -\frac{MT}{2} ln 2\pi - \frac{T}{2} ln|\Sigma| - \frac{1}{2} u' (\Sigma^{-1} \otimes I) u \end{aligned}$$

6.2 Variables instrumentales

Nous allons examiner deux situations très fréquentes avec lesquelles l'estimateur des moindres carrés ordinaires ne peut être utilisé.

6.2.1 Système d'équations simultanées

Il existe ici une relation entre les différentes équations qui peut être donnée par la théorie économique. Les variables se déterminent donc de façon conjointe.

Exemple : On a le modèle structurel suivant :

équation de demande :	$q_d = \alpha_1 p + \alpha_2 y + \epsilon_d$
équation d'offre :	$q_s = \beta_1 p + \epsilon_s$
condition d'équilibre :	$q_d = q_s = q$

Ces trois équations déterminent conjointement q et p. Ici, y est une variable exogène. On suppose que : $E(\epsilon_{dt}) = 0$, $E(\epsilon_{st}) = 0$, $E(\epsilon_{dt}^2) = \sigma_d^2$, $E(\epsilon_{st}^2) = \sigma_s^2$, $E(\epsilon_{dt}\epsilon_{st}) = 0$ et

$$E[\epsilon_{dt} y_t] = E[\epsilon_{st} y_t] = 0$$

et les variables sont mesurées en déviations par rapport à la moyenne.

On résout en fonction de p et q pour obtenir la forme réduite :

$$\begin{aligned} p &= \frac{\alpha_2}{\beta_1 - \alpha_1} y + \frac{\epsilon_d - \epsilon_s}{\beta_1 - \alpha_1} = \pi_1 y + \nu_1 \\ q &= \frac{\beta_1 \alpha_2}{\beta_1 - \alpha_1} y + \frac{\beta_1 \epsilon_d - \alpha_1 \epsilon_s}{\beta_1 - \alpha_1} = \pi_2 y + \nu_2 \end{aligned}$$

et on suppose que $\beta_1 \neq \alpha_1$. Cette hypothèse veut tout simplement dire que la pente de la demande n'est pas la même que la pente de l'offre.

Qu'est-ce qui arrive si on effectue des MCO sur l'équation de demande et d'offre ? Prenons l'équation de l'offre. L'estimateur des MCO est donné par :

$$\hat{\beta}_{1M.C.O.} = (p'p)^{-1} p'q.$$

Prenons l'espérance de cet estimateur :

$$E\hat{\beta}_{1.M.C.O.} = E[(p'p)^{-1}p'(p\beta_1 + \epsilon_s] = E\left[(p'p)^{-1}p'p\beta\right] + E\left[(p'p)^{-1}p'\epsilon_s\right]$$

$$= \beta_1 + E\left[(p'p)^{-1}p'\epsilon_s\right]$$

Mais, $E(p'\epsilon_s) \neq 0$ car

$$E(p'\epsilon_s) = E\left(\frac{\alpha_2}{\beta_2 - \alpha_1}y'\epsilon_s\right) + E\left(\frac{1}{\beta_1 - \alpha_1}(\epsilon_d - \epsilon_s)'\epsilon_s\right) \neq 0.$$

L'estimateur des MCO est biaisé parce que p est une variable endogène.

De façon générale, la convergence en probabilité de l'estimateur des M.C.O. nécessite que $E(X'\epsilon) = 0$ pour le modèle linéaire $Y = X\beta + \epsilon$. En effet, la convergence en probabilité de l'estimateur des M.C.O requiert que

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

L'expression $\frac{1}{N}X'\epsilon$ est un estimateur convergent de $E(X'\epsilon)$ qui doit donc être égale à zéro. La matrice de variables explicatives doit donc être orthogonale aux terme d'erreurs. On peut réécrire cette condition sous la forme $E(x_n\epsilon_n) = 0$ pour $n = 1, \ldots, N$. Remarquez que cette condition est moins restrictive que la condition $E(\epsilon/X) = 0$ qui garantit un estimateur sans biais en petit échantillon. De plus, l'estimateur des M.C.O est obtenu à l'aide de la conditions d'orthogonalité $X'\hat{\epsilon} = 0$ (voir les équations (14) et (24) de la première section).
Solution : estimateur avec des variables instrumentales.

6.2.2 Erreurs de mesure

Supposons le modèle suivant :

$$Y = W\beta + \epsilon$$

Malheureusement on observe pas directement exactement W, on observe plutôt X où

$$\begin{aligned} X &= W + U \\ E(U/W) &= 0. \end{aligned}$$

La matrice W est donc observée augmentée d'une erreur de mesure. On veut faire une régression avec les observations X. Puisque $W = X - U$, le modèle en fonction de X devient

$$\begin{aligned} Y &= (X - U)\beta + \epsilon \\ Y &= X\beta - U\beta + \epsilon \\ Y &= X\beta + v \end{aligned}$$

où $v = -U\beta + \epsilon$. L'estimateur MCO n'est pas convergent, car X et v dépendent de U. Ainsi,

$$\begin{aligned} \text{plim } (X'v) &\neq 0 \\ \text{plim } X'(-U\beta + \epsilon) &= \text{plim } (-X'U\beta + X'\epsilon) \neq 0 \end{aligned}$$

puisque plim $(X'U\beta) \neq 0$.
Solution : estimateur avec des variables instrumentales.

6.2.3 Estimateurs à variables instrumentales

On suppose le modèle suivant :

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

mais $E(X'\epsilon) \neq 0$. L'estimateur des MCO ne converge donc pas en probabilité.

On utilise alors ce qu'on appelle des variables instrumentales Z qui sont corrélées avec X mais non corrélées avec ϵ. Dans le cas où Z a la même dimension que X, on prémultiplie le modèle par Z'. Ainsi

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

et en prenant l'espérance

$$E(Z'Y) = E(Z'X)\beta + E(Z'\epsilon).$$

Sous l'hypothèse que $E(Z'\epsilon) = 0$, on obtient alors :

$$\beta = (E(Z'X))^{-1}E(Z'Y).$$

L'inverse de $E(Z'X)$ doit donc exister, ce qui implique que le rang de cette matrice doit être égal à K, le rang de la matrice X, ce qui correspond au nombre de variables explicatives. Cette **condition de rang** correspond à la condition suffisante pour identifier et donc estimer β.

Pour obtenir l'estimateur sur données, on remplace alors les espérances par les valeurs observées, ce qui nous donne l'estimateur à variables instrumentales :

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

Par la loi des grands nombres, cet estimateur converge en probabilité sous les hypothèses suivantes :

$$\begin{aligned} plim\frac{Z'X}{N} &= E\left(\frac{Z'X}{N}\right) = Q_{ZX} < \infty \\ plim\frac{Z'\epsilon}{N} &= E\left(\frac{Z'\epsilon}{N}\right) = 0. \end{aligned}$$

On a également par le théorème centrale limite

$$\frac{1}{\sqrt{N}}Z'\epsilon \overset{loi}{\rightarrow} N\left(0, \sigma^2 E\left(\frac{Z'Z}{N}\right)\right)$$

sous l'hypothèse que $E(\epsilon\epsilon'/Z) = \sigma^2 I$.

La matrice de variance-covariance asymptotique est alors donnée par :

$$\begin{aligned} \lim_{T\rightarrow\infty} var\hat{\beta}_{VI} &= [E(Z'X)]^{-1}E[Z'\epsilon\,\epsilon'Z][E(X'Z)]^{-1} \\ &= \sigma^2[E(Z'X)]^{-1}E[Z'Z][E(X'Z)]^{-1}. \end{aligned}$$

Avec les hypothèses suivantes : $plim\frac{Z'Z}{N} = E(\frac{Z'Z}{N}) = Q_{ZZ}$ et $plim\frac{X'Z}{N} = E(\frac{X'Z}{N}) = Q_{XZ}$ et en appliquant le théorème central limite sur $\frac{1}{\sqrt{N}}Z'\epsilon$, on peut

alors montrer que :

$$\sqrt{N}(\hat{\beta}_{VI} - \beta) \overset{Loi}{\to} N\left(0, \sigma^2 Q_{ZX}^{-1} Q_{ZZ} Q_{XZ}^{-1}\right).$$

L'estimateur de la matrice de variance covariance sera donné par

$$\widehat{Var}\left(\hat{\beta}_{IV}\right) = \hat{\sigma}^2 \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1},$$

où $\hat{\sigma}^2$ est un estimateur convergent de σ^2 tel que

$$\hat{\sigma}^2 = \frac{1}{N-K}\hat{\epsilon}'\hat{\epsilon} = \frac{1}{N-K}(Y - X\hat{\beta}_{VI})'(Y - X\hat{\beta}_{VI}).$$

Plus grande est la corrélation entre Z et X qui est fonction de $E(Z'X)$, meilleur sera l'estimateur en petit échantillon et de façon asymptotique.

Revenons à notre exemple. Puisque y est exogène, (non corrélée avec les termes d'erreurs) et que p est corrélée avec ϵ_s (voir la forme réduite), on utilise alors y comme variable instrumentale. On a alors l'estimateur suivant :

$$\hat{\beta}_{1,VI} = (y'p)^{-1}y'q.$$

On peut aussi utiliser la forme réduite pour obtenir un estimateur des moindres carrés indirects. On effectue un MCO sur les deux équations de la forme réduite, on obtient $\hat{\pi}_1$ et $\hat{\pi}_2$, alors

$$\hat{\beta}_{1,MCI} = \frac{\hat{\pi}_2}{\hat{\pi}_1} = \frac{(y'y)^{-1}y'q}{(y'y)^{-1}y'p} = (y'p)^{-1}y'q.$$

qui est donc égal à l'estimateur à variables instrumentales. On pourrait vouloir effectuer le même chose avec l'équation de demande. Cette équation n'est cependant pas identifiée. On a donc un problème d'identification.

Le problème provient du fait que l'on cherche à estimer $\alpha_1, \alpha_2, \beta_1$ et on a seulement à partir de la forme réduite : π_1, π_2. On ne peut identifier 3 paramètres à l'aide de 2 paramètres de la forme réduite.

Démonstration graphique du problème d'identification.

Examinons à nouveau à l'aide de l'estimateur à variables intrumentales l'estimation des équations de demande et d'offre présentées plus haut. Prenons, dans un premier temps, l'estimation du paramètre β_1 de l'équation d'offre. Nous avons vu que ce paramètre pouvait être identifié. Pour ce cas, l'estimateur à variables instrumentales est donnée par :

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

et la variable instrumentale est y. On a donc que $Y = q$, $X = p$, $Z = y$ et $\beta = \beta_1$. La matrice $Z'X$ est de plein rang égal ici à 1, l'identification du paramètre β_1 est donc possible puisque la condition suffisante est satisfaite. Examinons maintenant l'équation de demande à la lumière de l'estimateur à variables instrumentales. Peux-t-on identifier les paramètres α_1 et α_2 ? Nous avons vu précédemment que nous ne pouvions identifier ces deux paramètres. Pour cette équation de la demande, $\beta = (\alpha_1, \alpha_2)'$, $X = (p, y)$ et $Z = y$. La matrice $Z'X$ n'est pas de plein rang, la condition suffisante d'identification n'est donc pas respectée.

Prenons maintenant le système d'équations suivant :

$$\begin{aligned} y_1 &= \beta_1 y_2 + \delta_{11} z_1 + \epsilon_1 \\ y_2 &= \beta_2 y_1 + \delta_{22} z_2 + \epsilon_2 \end{aligned}$$

où z_1 et z_2 sont des variables exogènes. Est-ce que l'on peut estimer β_1, β_2, δ_{11} et δ_{11} par variables instrumentales ?

On peut réécrire ce système sous une forme réduite. On obtiendra alors :

$$\begin{aligned} y_1 &= \pi_{11} z_1 + \pi_{12} z_2 + \nu_1 \\ y_2 &= \pi_{21} z_1 + \pi_{22} z_2 + \nu_2. \end{aligned}$$

On peut estimer ces équations de forme réduite par les moindres carrés ordinaires et ainsi obtenir : $\hat{\pi}_{11}$, $\hat{\pi}_{12}$, $\hat{\pi}_{21}$ et $\hat{\pi}_{22}$. On aura alors 4 paramètres estimés pour identifier 4 paramètres structurels. On a donc à résoudre un système d'équations avec autant d'équations que d'inconnues. La solution est alors unique. On peut donc identifier les paramètres structurels.

Pour l'estimateur à variables instrumentales dans le cas de la première équation structurelle, $\beta = (\beta_1, \delta_{11})'$, $X = (y_2, z_1)$ et $Z = (z_1, z_2)$. La matrice $Z'X$ est de rang plein si et seulement si $\pi_{22} \neq 0$. Si c'est le cas, la condition suffisante d'identification est satisfaite. Pour la deuxième équation structurelle, on a alors : $\beta = (\beta_2, \delta_{22})'$, $X = (y_1, z_1)$ et $Z = (z_1, z_2)$. La matrice $Z'X$ est de rang plein si et seulement si $\pi_{11} \neq 0$. Si c'est le cas, la condition suffisante d'identification est satisfaite. On remarque que la condition suffisante de rang implique donc que la variable endogène de l'équation doit dépendre d'une variable exogène autre que celle déjà comprise dans l'équation structurelle à estimer. Examinons un peu plus en détail pourquoi. Pour la première équation on aura que

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

pour $\beta = (\beta_1, \delta_{11})'$, $X = (y_2, z_1)$ et $Z = (z_1, z_2)$. En prenant les probabilités limites

$$\begin{aligned}
\text{plim } \hat{\beta}_{VI} &= \text{plim } (Z'X)^{-1}Z'Y \\
&= \text{plim } (Z'X)^{-1}Z'(X\beta + \epsilon_1) \\
&= \beta + \text{plim } \left(\frac{Z'X}{N}\right)^{-1} \text{plim } \left(\frac{Z'\epsilon_1}{N}\right) \\
&= \beta
\end{aligned}$$

si $\left(\frac{Z'X}{N}\right)$ est de rang complet. Généralement, on a

$$\begin{aligned}
\text{plim } \left(\frac{Z'X}{N}\right) &= \text{plim } \frac{1}{N} \left(\begin{pmatrix} z_1' \\ z_2' \end{pmatrix} \begin{pmatrix} y_2 & z_1 \end{pmatrix} \right) \\
&= \text{plim } \frac{1}{N} \begin{pmatrix} z_1'y_2 & z_1'z_1 \\ z_2'y_2 & z_2'z_1 \end{pmatrix}
\end{aligned}$$

Regardons chaque terme de cette matrice à partir de la forme réduite

$$\begin{aligned}
\text{plim } \frac{1}{N}\left(z_1' y_2\right) &= \text{plim } \frac{z_1'}{N}(\pi_{21} z_1 + \pi_{22} z_2 + v_2) \\
&= \text{plim } \frac{1}{N}(\pi_{21} z_1' z_1 + \pi_{22} z_1' z_2 + z_1' v_2) \\
&= \text{plim } \frac{1}{N}(\pi_{21} z_1' z_1 + \pi_{22} z_1' z_2) \\
\text{plim } \frac{1}{N}\left(z_2' y_2\right) &= \text{plim } \frac{z_2\prime}{N}(\pi_{21} z_1 + \pi_{22} z_2 + v_2) \\
&= \text{plim } \frac{1}{N}(\pi_{21} z_2\prime z_1 + \pi_{22} z_2\prime z_2)
\end{aligned}$$

On a alors la matrice

$$\begin{aligned}
\text{plim } \left(\frac{Z'X}{N}\right) &= \text{plim } \frac{1}{N}\begin{pmatrix} z_1' y_2 & z_1' z_1 \\ z_2' y_2 & z_2' z_1 \end{pmatrix} \\
&= \text{plim } \frac{1}{N}\begin{pmatrix} \pi_{21} z_1' z_1 + \pi_{22} z_1' z_2 & z_1' z_1 \\ \pi_{21} z_2' z_1 + \pi_{22} z_2' z_2 & z_2' z_1 \end{pmatrix}
\end{aligned}$$

Si $\pi_{22} = 0$, on a multicollinéarité entre les deux colonnes. Cette conclusion est due au fait que y_1 dépend de y_2 et z_1 et que y_2 dépend seulement de z_1 selon la forme réduite lorsque $\pi_{22} = 0$.

De même pour l'autre équation, si $\pi_{11} = 0$. Il y a aussi multicollinéarité. On se rappelle en effet que

$$\begin{aligned}
y_1 &= \beta_1 y_2 + \delta_{11} z_1 + \epsilon_1 \\
y_2 &= \beta_2 y_1 + \delta_{22} z_2 + \epsilon_2.
\end{aligned}$$

Pour que la condition de rang soit valide, il faut que la variable explicative endogène dépende de variables exogènes autre que les variables exogènes inclues dans la régression.

On va maintenant examiner le cas où on a plus d'instruments que de variables endogènes. Prenons le système d'équations suivant :

$$\begin{aligned}
y_1 &= \beta_1 y_2 + \delta_{11} z_1 + \epsilon_1 \\
y_2 &= \delta_{21} z_1 + \delta_{22} z_2 + \delta_{23} z_3 + \epsilon_2.
\end{aligned}$$

avec $E(\epsilon_1\epsilon_2') \neq 0$ et z_1, z_2 et z_3 sont des variables exogènes. Puisque y_2 dépend de z_1, z_2 et z_3, on peut utiliser ces variables comme instruments pour estimer les paramètres structurels de la première équation. On aura alors pour notre estimateur à variables instrumentales que $\beta = (\beta_1, \delta_{11})'$, $X = (y_2, z_1)$ et $Z = (z_1, z_2, z_3)$. On remarque que toutes les combinaisons linéaires des instruments Z, c.a.d. ZA pour une matrice A compatible, ne sont pas corrélées avec le terme d'erreur ϵ_1. En effet $E(A'Z'\epsilon_1) = 0$, puisque z_1, z_2 et z_3 sont des variables exogènes. L'objectif serait donc de choisir une combinaison linéaire ZA de dimension $N \times 2$ telle que $A'Z'X$ est inversible. L'estimateur à variables instrumentales sera alors donné par :

$$\hat{\beta}_{VI} = (A'Z'X)^{-1}A'Z'y_1.$$

La question est donc de savoir s'il existe une combinaison linéaire optimale dans le sens où cette combinaison linéaire nous donne l'estimateur à variables instrumentales avec une variance minimale. Il en existe une et cette combinaison linéaire optimale nous est donnée par la projection de la ou des variables endogènes (y_2 dans ce cas-ci) sur les variables exogènes.

Réécrivons la deuxième équation structurelle de la façon suivante :

$$y_2 = Z\Pi + \epsilon_2.$$

On peut obtenir la projection de y_2 sur les variables exogènes Z par l'estimateur des moindres carrés sur cette équation. On aura alors $\hat{y}_2 = Z\hat{\Pi} = Z(Z'Z)^{-1}Z'y_2$. Utilisons maintenant cette projection comme instruments. On définit $\hat{X} = (\hat{y}_2, z_1)$ et $X = (y_2, z_1)$, l'estimateur VI sera alors :

$$\hat{\beta}_{VI} = (\hat{X}'X)^{-1}\hat{X}'y_1.$$

et $\hat{X} = Z(Z'Z)^{-1}Z'X$. La combinaison linéaire est donc donnée ici par la matrice de projection des X sur Z, c.a.d. $ZA = Z(Z'Z)^{-1}Z'X$. L'estimateur s'écrit donc :

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

Problème à faire : démontrez que cet estimateur est convergent et asymptotiquement normal.

On appelle cet estimateur, l'estimateur des doubles moindes carrés (Two-Stage Least Squares, TSLS). Ce nom provient du fait que la projection en première étape consiste à effectuer des moindres carrés de la ou des variables endogènes (y_2 pour notre exemple) sur les variables exogènes Z pour obtenir $\hat{X}$ et à effectuer un autre moindres carrés de la variable à expliquer (y_1 dans notre exemple) sur $\hat{X}$, en deuxième etape, en notant que

$$\hat{\beta}_{VI} = (\hat{X}'X)^{-1}\hat{X}'y_1 = (\hat{X}'\hat{X})^{-1}\hat{X}'y_1.$$

Lorsque nous avons plus d'instruments que de variables endogènes, on dira alors que l'on a suridentification. Comme nous l'avons mentionné plus haut, on va alors choisir une combinaison linéaire ZA telle que $A'Z'X$ est de rang égal à K, le nombre de variables explicatives, ainsi $A'Z'X$ est inversible. La combinaison linéaire $A = (Z'Z)^{-1}Z'X$ correspond à l'estimateur des doubles moindes carrés qui a la propriété d'être à variance minimum parmi toutes les combinaisons linéaires possibles de Z.

Preuve :

La variance de l'estimateur des doubles moindres carrés est :

$$var(\hat{\beta}_{DMC}/X, Z) = \sigma^2 \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1}.$$

La variance de l'estimateur à variables instrumentales pour toutes combinaisons linéaires ZA est :

$$var(\hat{\beta}_{VI}/X, Z) = \sigma^2 \left(A'Z'X\right)^{-1} A'Z'ZA \left(X'ZA\right)^{-1}.$$

On doit montrer que $var(\hat{\beta}_{DMC}/X, Z) \leq var(\hat{\beta}_{VI}/X, Z)$ pour toutes matrices A. Ceci consiste donc à montrer que

$$X'Z(Z'Z)^{-1}Z'X \geq X'ZA \left(A'Z'ZA\right)^{-1} A'Z'X.$$

au sens matriciel.

Puisque la matrice $Z'Z$ est symétrique definie positive, on peut la réécrire comme étant : $Z'Z = C\Lambda C'$ où $C'C = I$ et $C' = C^{-1}$. Ceci implique que $(Z'Z)^{-1} = C\Lambda^{-1}C'$. On doit montrer que

$$X'Z(Z'Z)^{-1}Z'X - X'ZA\left(A'Z'ZA\right)^{-1}A'Z'X \geq 0.$$

On peut réécrire l'expression plus haut conmme étant :

$$X'ZC\Lambda^{-1}C'Z'X - X'ZA\left(A'Z'ZA\right)^{-1}A'Z'X \geq 0,$$

ce qui est égale à

$$X'ZC\Lambda^{-1/2}\left[I - \Lambda^{1/2}C'A\left(A'Z'ZA\right)^{-1}A'C\Lambda^{1/2}\right]\Lambda^{-1/2}C'Z'X \quad \geq \quad 0.$$

On a que $A'Z'ZA = A'C\Lambda C'A$, on peut alors définir $D = \Lambda^{1/2}C'A$. L'expression plus haut nous donne :

$$X'ZC\Lambda^{-1/2}\left[I - D\left(D'D\right)^{-1}D'\right]\Lambda^{-1/2}C'Z'X \quad \geq \quad 0.$$

La matrice $\left[I - D\left(D'D\right)^{-1}D'\right]$ est une matrice idempotente, elle est donc positive semi-définie, ce qui complète la preuve.

De façon générale, la matrice des instruments Z est de dimension $N \times L$. Lorsque $L < K$, on a alors sous-identification, lorsque $L = K$, le système est juste-identifié et lorsque $L > K$, on a suridentification.

La condition nécessaire (qu'on appelle **condition d'ordre**) pour l'identification est donc que $L \geq K$. On a alors au moins autant d'instruments que de variables explicatives. La condition suffisante pour l'identification (appelé **condition de rang**) est que $E(Z'X)$ soit de rang K. Malheureusement la condition de rang n'est pas toujours vérifiée. On a souvent ce que l'on appelle des instruments faibles (weak instruments).

6.2.4 Tests de spécification

Test de Hausman

Le test est basé sur le respect ou non de la condition d'orthogonalité suivante des moindres carrés ordinaires

$$E(X'\epsilon) \;=\; 0$$

avec $\hat{\beta}_{MCO} = (X'X)^{-1}X'Y$ et $\hat{\beta}_{VI} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y$. La statistique du test de Hausman est

$$\hat{d} \;=\; \hat{\beta}_{VI} - \hat{\beta}_{MCO}.$$

Les hypothèses nulle et alternative sont alors

$$\begin{aligned} H_0 &: \; d = 0 \\ H_1 &: \; d \neq 0. \end{aligned}$$

Sous l'hypothèse nulle,

$$\text{plim } \hat{d} = \text{plim } \hat{\beta}_{VI} - \text{plim } \hat{\beta}_{MCO} = 0$$

Le test est basé sur la statistique à distance finie

$$H_N = \left(\hat{\beta}_{VI} - \hat{\beta}_{MCO}\right)' \left(\widehat{Var}(\hat{\beta}_{VI} - \hat{\beta}_{MCO})\right)^{-1} \left(\hat{\beta}_{VI} - \hat{\beta}_{MCO}\right) \rightarrow \chi^2(k - k^*)$$

où k^* est le nombre de variables exogènes dans X.

On sait que

$$\begin{aligned} \widehat{Var}(\hat{\beta}_{VI}) &= \tilde{\sigma}^2 \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} \\ \widehat{Var}(\hat{\beta}_{MCO}) &= \hat{\sigma}^2 \left(X'X\right)^{-1} \\ \widehat{Var}(\hat{\beta}_{VI} - \hat{\beta}_{MCO}) &= \widehat{Var}(\hat{\beta}_{VI}) - \widehat{Var}(\hat{\beta}_{MCO}) \\ &= \tilde{\sigma}^2 \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} - \hat{\sigma}^2 \left(X'X\right)^{-1} \\ &= \hat{\sigma}^2 \left[\left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} - \left(X'X\right)^{-1}\right] \end{aligned}$$

Ce résultat provient du fait que

$$\text{plim } Cov(\hat{\beta}_{VI}, \hat{\beta}_{MCO}) = \text{plim } Var(\hat{\beta}_{MCO}).$$

Essayer de démontrer ce résultats. Le test se déroule correctement asymptotiquement, mais il peut y avoir des problèmes en petit échantillon provenant de l'inversion de la matrice $\left[(X'Z(Z'Z)^{-1}Z'X)^{-1} - (X'X)^{-1}\right]$.

Test de suridentification

La convergence en probabilité et la convergence en loi de l'estimateur à variables instrumentales dépendent de l'hypothèse que le modèle est bien spécifié. En particulier, les bonnes propriétés de l'estimateur repose sur l'hypothèse que les instruments sont orthogonaux aux termes d'erreurs. On appelle ceci des conditions d'orthogonalité ou des conditions de moments. On peut effectuer un test de spécification sur ces conditions d'orthogonalité appelé test de Sargan. Ce test consiste à vérifier si les conditions d'orthogonalité $E(Z'\epsilon)$ sont respectées. On utilisera alors l'estimateur de ces conditions, c.a.d. $\frac{1}{N}Z'\hat{\epsilon} = \frac{1}{N}Z'\left(Y - X\hat{\beta}_{IV}\right)$, et on évaluera si celle-ci sont significativement différentes de zéro. Le test est basé sur la statistique suivante :

$$J_N = N\left(\frac{1}{N}Z'\hat{\epsilon}\right)'\left(\hat{\sigma}^2\frac{Z'Z}{N}\right)^{-1}\left(\frac{1}{N}Z'\hat{\epsilon}\right)$$

où $\hat{\sigma}^2\frac{Z'Z}{N}$ est l'estimateur de la matrice de variance-covariance de $\frac{1}{\sqrt{N}}Z'\hat{\epsilon}$ si on suppose que $E(\epsilon\epsilon'/Z) = \sigma^2 I$. Cette statistique suit asymptotiquement une loi du khi-deux à L-K degrés de liberté. On doit donc avoir suridentification pour effectuer le test. Sinon par construction $\frac{1}{N}Z'\hat{\epsilon} = 0$. Pour cette raison, le test est souvent appelé test de suridentification.

Si on soupçonne qu'un sous-ensemble de L_1 instruments sont valides et que $K \leq L_1 < L$, on peut effectuer un test d'orthogonalité pour les $L - L_1$ instruments restants. On effectue premièrement le test J_N avec l'ensemble des L instruments et on effectue ensuite le test pour une estimation basée sur le sous-ensemble de L_1 instruments considérés comme étant valides que nous appellerons J_{1N}. Le test sur

la validité des $L - L_1$ instruments est alors basé sur la statistique $J_N - J_{1N}$ et suit une loi asymptotique du khi-deux à $L - L_1$ degrés de liberté. Si $L_1 = K$, alors ce test est équivalent au test J_N.

6.2.5 Instruments faibles

Si les instruments ne sont pas corrélés ou sont faiblement corrélés avec les variables explicatives, l'estimateur à variables instrumentales n'est pas convergent et sa loi asymptotique peut dévier substantiellement de la loi normale. Le problème provient de la condition suffisante, à savoir lorsque la quantitité $\frac{1}{N}Z'X$ est près de zéro. Il y a deux stratégies possibles lorsqu'on soupçonne la présence d'instruments faibles. La première stratégie est d'effectuer une régression des variables explicatives que l'on considèrent comme étant endogènes sur les instruments. Si les instruments sont significatifs, on peut alors utiliser ces instruments. Staiger et Stock (1997) suggèrent que pour une statistique F plus grande que 10, on a pas à se soucier du problème des instruments faibles. La deuxième stratégie consiste à faire l'estimation par des variables instrumentales, sans se soucier si ces instruments sont faibles ou pas et à utiliser des tests robustes aux instruments faibles. Il existe maintenant plusieurs tests statistiques sur les paramêtres qui sont robustes aux instruments faibles dans un cadre linéaire ou non linéaire.

Réexaminons le système d'équations suivant :

$$\begin{aligned} y_1 &= \beta_1 y_2 + \delta_{11} z_1 + \epsilon_1 \\ y_2 &= \beta_2 y_1 + \delta_{22} z_2 + \epsilon_2 \end{aligned}$$

où z_1 et z_2 sont des variables exogènes. Est-ce que l'on peut estimer β_1, β_2, δ_{11} et δ_{11} par variables instrumentales ?

La forme réduite correspondante est : obtiendra alors :

$$\begin{aligned} y_1 &= \pi_{11} z_1 + \pi_{12} z_2 + \nu_1 \\ y_2 &= \pi_{21} z_1 + \pi_{22} z_2 + \nu_2. \end{aligned}$$

Pour l'estimateur à variables instrumentales dans le cas de la première équation structurelle, $\beta = (\beta_1, \delta_{11})'$, $X = (y_2, z_1)$ et $Z = (z_1, z_2)$. On a vu que la matrice $Z'X$ est de rang plein si et seulement si $\pi_{22} \neq 0$. Pour la première équation, l'estimateur à variables instrumentales est :

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

pour $\beta = (\beta_1, \delta_{11})'$, $X = (y_2, z_1)$ et $Z = (z_1, z_2)$. En prenant les probabilités limites

$$\begin{aligned}
\text{plim } \hat{\beta}_{VI} &= \text{plim } (Z'X)^{-1}Z'Y \\
&= \text{plim } (Z'X)^{-1}Z'(X\beta + \epsilon_1) \\
&= \beta + \text{plim } \left(\frac{Z'X}{N}\right)^{-1} \text{plim } \left(\frac{Z'\epsilon_1}{N}\right) \\
&= \beta
\end{aligned}$$

si $\left(\frac{Z'X}{N}\right)$ est de rang complet. On a donc

$$\begin{aligned}
\text{plim } \left(\frac{Z'X}{N}\right) &= \text{plim } \frac{1}{N}\left(\begin{pmatrix} z_1' \\ z_2' \end{pmatrix}\begin{pmatrix} y_2 & z_1 \end{pmatrix}\right) \\
&= \text{plim } \frac{1}{N}\begin{pmatrix} z_1'y_2 & z_1'z_1 \\ z_2'y_2 & z_2'z_1 \end{pmatrix}
\end{aligned}$$

Regardons chaque terme de cette matrice à partir de la forme réduite

$$\begin{aligned}
\text{plim } \frac{1}{N}\left(z_1'y_2\right) &= \text{plim } \frac{z_1'}{N}(\pi_{21}z_1 + \pi_{22}z_2 + v_2) \\
&= \text{plim } \frac{1}{N}(\pi_{21}z_1'z_1 + \pi_{22}z_1'z_2 + z_1'v_2) \\
&= \text{plim } \frac{1}{N}(\pi_{21}z_1'z_1 + \pi_{22}z_1'z_2) \\
\text{plim } \frac{1}{N}\left(z_2'y_2\right) &= \text{plim } \frac{z_2\prime}{N}(\pi_{21}z_1 + \pi_{22}z_2 + v_2) \\
&= \text{plim } \frac{1}{N}(\pi_{21}z_2\prime z_1 + \pi_{22}z_2\prime z_2)
\end{aligned}$$

On a alors la matrice

$$\begin{aligned}\text{plim}\left(\frac{Z'X}{N}\right) &= \text{plim}\,\frac{1}{N}\begin{pmatrix} z_1'y_2 & z_1'z_1 \\ z_2'y_2 & z_2'z_1 \end{pmatrix} \\ &= \text{plim}\,\frac{1}{N}\begin{pmatrix} \pi_{21}z_1'z_1 + \pi_{22}z_1'z_2 & z_1'z_1 \\ \pi_{21}z_2'z_1 + \pi_{22}z_2'z_2 & z_2'z_1 \end{pmatrix}\end{aligned}$$

Si $\pi_{22} = 0$, on a multicollinéarité entre les deux colonnes, la condition de rang n'est pas satisfaite. On a un problème d'instruments faible si π_{22} approche zéro. On pourrait avoir par exemple que $\pi_{22} = \frac{C}{\sqrt{N}}$. Ainsi la probabilité limite est donnée par

$$\begin{aligned} plim\frac{1}{N}\pi_{22}z_1'z_2 &= plim\frac{1}{N}\frac{C}{\sqrt{N}}z_1'z_2 = 0 \\ plim\frac{1}{N}\pi_{22}z_2'z_2 &= plim\frac{1}{N}\frac{C}{\sqrt{N}}z_2'z_2 = 0. \end{aligned}$$

En petit échantillon, ces expressions ne seront pas égales à zéro mais seront près de zéro. On a alors un instrument faible.

Nous allons maintenant présenter un test robuste aux instruments faibles basé sur la statistique de Anderson et Rubin (1949). Prenons la représentaion générale suivante :

$$\begin{aligned} y_1 &= Y_2\beta + Z_1\delta + \epsilon_1 \\ Y_2 &= Z_1\Pi_1 + Z_2\Pi_2 + V. \end{aligned}$$

la variable d'intérêt y_1 est de dimension $n \times 1$, elle est fonction de M_1 variables endogènes telles que Y_2 est de dimension $N \times M_1$ et les matrices d'instruments Z_1 et Z_2 sont respectivements de dimension $N \times K_1$ et $N \times K_2$ et $Z = [Z_1 Z_2]$ est une matrice de dimension $N \times K$. On s'intéresse à l'hypothèse nulle $H_0 : \beta = \beta_0$ dans la première équation. On peut réécrire la première équation de la façon suivante :

$$y_1 - Y_2\beta_0 = Z_1\theta_1 + Z_2\theta_2 + \eta$$

où $\theta_1 = \delta + \Pi_1(\beta - \beta_0)$, $\theta_2 = \Pi(\beta - \beta_0)$ et $\eta = \epsilon_1 + V(\beta - \beta_0)$. Sous l'hypothèse nulle

$$y_1 - Y_2\beta_0 = Z_1\theta_1 + \eta. \tag{25}$$

Tester que β est égal à β_0 revient donc à tester que $\theta_2 = 0$. On peut donc construire la statistique F de l'hypothèse nulle que $\theta_2 = 0$ pour l'équation (25). Ainsi,

$$AR(\beta_0) = \frac{\left[SS_0(\beta_0) - SS_1(\beta_0)\right]/K_2}{SS_1(\beta_0)/(N-K)}$$

où $SS_0(\beta_0) = (y_1 - Y_2\beta_0)' M_{Z_1}(y_1 - Y_2\beta_0)$ qui est la somme des carrés des résidus sous l'hypothèse nulle et $SS_1(\beta_0) = (y_1 - Y_2\beta_0)' M_Z(y_1 - Y_2\beta_0)$ qui est la somme des carrés des résidus sous l'alternative et $M_{Z_1} = [I - Z_1(Z_1'Z_1)^{-1}Z_1']$, $M_Z = [I - Z(Z'Z)^{-1}Z']$. Sous l'hypothèse que le vecteur de termes d'erreurs suit une normale multivariée centrée à zéro et $E(\epsilon_1\epsilon_1'/Z) = \sigma^2 I$, la statistique $AR(\beta_0)$ suit une Fisher $F(K_2, N-K)$.

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, on peut réécrire le moment théorique sous la forme :

$$E(y_t - \mu) = 0$$

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 convergent en probabilité, il doit respecter la condition suivante :

$$E(\varepsilon_t / x_t) = 0.$$

On peut obtenir des conditions de moments non conditionnelles en prémultipliant par le vecteur x_t et en prenant l'espérance. Ainsi, qui correspond aux conditions de moments suivantes :

$$E x_t E(\varepsilon_t / x_t) = E(x_t \varepsilon_t) = 0.$$

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,

$$\frac{1}{T}\sum_{t=1}^{T} x_t\hat{\varepsilon}_t = \frac{1}{T}\sum_{t=1}^{T} x_t(y_t - \hat{\beta}'x_t) = 0.$$

Puisque le nombre de conditions de moments (ou conditions d'orthogonalité) est égal au nombre de paramètres dans le vecteur β, l'estimateur de la méthode des moments généralisés est donné directement par

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

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 égale 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 convergent en probabilité s'il respecte les conditions d'orthogonalité suivantes :

$$E(z_t\varepsilon_t) = 0.$$

Un vecteur de variables instrumentales doit donc être orthogonale au terme d'erreur ε_t. De plus, ce vecteur doit contenir le plus possible la même information que le vecteur de variables explicatives x_t. Le vecteur de variables instruments doit donc être le plus fortement corrélé possible avec le vecteur de variables explicatives.

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

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

Lorsque le nombre d'instruments est plus grand que le nombre de paramètres d'intérêt, on cherchera $\hat{\beta}$ de sorte à ce que les moments empiriques

$$\frac{1}{T}Z'\hat{\varepsilon} = \frac{1}{T}Z'(Y - X\hat{\beta})$$

soient le plus près possible de zéro. On minimisera alors la forme quadratique suivante :

$$\hat{\beta}_{VI} = \arg\min \left[\frac{1}{T}Z'(Y - X\hat{\beta})\right]' W_T \left[\frac{1}{T}Z'(Y - X\hat{\beta})\right]$$

où W_T est une matrice de poids symétrique positive définie. En résolvant les C.P.O de ce problème de minimisation, on obtient

$$\hat{\beta}_{VI} = (X'ZW_TZ'X)^{-1} X'ZW_TZ'Y. \tag{26}$$

Cet estimateur dépend donc de la matrice de poids W_T. On peut montrer que l'estimateur optimal (donc à variance minimale) sera celui dont la matrice de poids est proportionnelle à l'inverse de la matrice de variance-covariance des moments théoriques. Prenons l'hypothèse que $E(\epsilon\epsilon'/Z) = \sigma^2 I$. La variance des moments théoriques $E(Z'\epsilon)$ est alors donnée par $\sigma^2 E(Z'Z)$. La matrice de poids optimale est alors

$$W_T^o = \left(\hat{\sigma}^2 \frac{Z'Z}{T}\right)^{-1}.$$

En insérant cette matrice de poids dans l'estimateur (26), on obtient

$$\hat{\beta}_{VI} = \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} X'Z(Z'Z)^{-1}Z'Y.$$

Cette estimateur est le même que celui présenté dans le section précédente. On l'appelle souvent en anglais GIVE (generalized instrumental variables estimator). Et comme nous l'avons vu, il correspond à l'estimateur des doubles moindres carrés.

Soit maintenant un modèle général non linéaire :

$$Y = h(X, \beta) + \varepsilon$$

où $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_t)'$ et $E(\varepsilon) = 0, E(\varepsilon\varepsilon') = \Omega$ 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

$$\frac{1}{T}\sum_{t=1}^{T} z_t\hat{\varepsilon}_t = \frac{1}{T}\sum_{t=1}^{T} z_t(y_t - h(x_t, \hat{\beta})) = 0.$$

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

$$S = \lim_{T\to\infty} \frac{1}{T}\sum_{i=1}^{T} \sigma_i^2 z_i z_i'.$$

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 $q \geq p$.

L'estimateur $\hat{\theta}_T$ 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 15 *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 :

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{loi} N\left(0, (D'WD)^{-1}D'WSWD(D'WD)^{-1}\right)$$

où

$$\frac{1}{T}\sum_{t=1}^{T} \frac{\partial f(x_t, \hat{\theta}_T)}{\partial \theta'} \xrightarrow{p} D = E\left[\frac{\partial f(x_t, \theta_0)}{\partial \theta'}\right]$$

et $W_T \xrightarrow{p} W$, 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

$$\sqrt{T}(\hat{\theta_T} - \theta_0) \xrightarrow{loi} N\left(0, \left(D'S^{-1}D\right)^{-1}\right).$$

Pour ce choix de W_T, la matrice de variance-covariance est la plus petite possible. 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.

Lorsque $q > p$, on obtient un test de spécification basé sur les conditions de suridentification. Ce test est donné par la statistique suivante :

$$J_T = T \left(\frac{1}{T} \sum_{t=1}^{T} f(x_t, \hat{\theta}_T) \right)' S_T^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} f(x_t, \hat{\theta}_T) \right)$$

et suit asymptotiquement une loi du $\chi^2(q-p)$. On appelle ce test le test J de Hansen

On peut également effectuer un test sur un sous-ensemble de moments. Si on soupçonne qu'un sous-ensemble q_1 de moments sont valides et que $p \leq q_1 < q$, on peut effectuer un test d'orthogonalité pour les $q - q_1$ conditions de moments suspectent. On ordonne les q_1 conditions de moments valides en premier dans le vecteur des conditions de moments. Ainsi

$$\frac{1}{T} \sum_{t=1}^{T} f(x_t, \theta) = \left(\frac{1}{T} \sum_{t=1}^{T} f_1(x_t, \theta)', \frac{1}{T} \sum_{t=1}^{T} f_2(x_t, \theta)' \right)'.$$

L'ordre ne change rien à l'estimateur $\hat{\theta}_T$ obtenu avec l'ensemble des moments. On effectue premièrement le test J_T avec l'ensemble des q conditions de moments. On effectue ensuite le test de spécification pour une estimation basée sur le sous-ensemble de q_1 moments considérés comme étant valides. Appellons $\tilde{\theta}_T$ l'estimateur obtenu avec seulement le premier sous-ensemble de moments. La statistique de ce test est alors donnée par

$$J_{1T} = T \left(\frac{1}{T} \sum_{t=1}^{T} f_1(x_t, \tilde{\theta}_T) \right)' S_{11,T}^{-1} \left(\frac{1}{T} \sum_{t=1}^{T} f_1(x_t, \tilde{\theta}_T) \right)$$

où $S_{11,T}$ est défini comme étant

$$S_T = \begin{bmatrix} S_{11,T} & S_{12,T} \\ S_{21,T} & S_{22,T} \end{bmatrix}$$

et S_T est la matrice de poids optimale pour l'ensemble des moments. Le matrice $S_{11,T}$ est donc l'inverse de la matrice de variance-covariance du premier sous-ensemble de moments. Le test sur la validité des $q - q_1$ instruments est alors basé sur la statistique $C_T = J_T - J_{1T}$ et suit une loi asymptotique du khi-deux à $q - q_1$ degrés de liberté. Si $q_1 = p$, alors ce test est équivalent au test J_T.

Exemple :

Supposons une économie composée d'un agent représentatif qui maximise une fonction d'utilité intertemporelle sous une contrainte budgétaire.

Le problème de maximisation de l'agent est le suivant :

$$\max_{C_{t+j}, A_{t+j+1}} E_t \left[\sum_{j=0}^{\infty} \beta^j U(C_{t+j}) \right]$$

où β est le taux d'escompte, C_{t+j} la consommation au temps $t+j$ et $U(C_{t+j})$ est la fonction d'utilité dépendant de la consommation en $t+j$. La contrainte budgétaire est donnée par l'expression suivante :

$$C_{t+j} + A_{t+j+1} = (1 + R_{t+j-1,t+j})A_{t+j} + Y_{t+j}$$

où A_{t+j} est la quantité détenue d'un actif financier sans risque venant à échéance en $t+j$, $R_{t+j-1,t+j}$ est le rendement de l'actif financier sans risque détenu de la période $t+j-1$ à la période $t+j$ et Y_{t+j} représente le revenu exogène à la période $t+j$.

On considère la forme fonctionnelle suivante pour la fonction d'utilité :

$$U(C_t) = \frac{C_t^{1-\sigma} - 1}{1 - \sigma}$$

où σ est le paramètre d'aversion au risque.

Par les conditions du premier ordre, on peut obtenir l'équation d'Euler suivante :

$$E_t\left[\beta\left(1+R_{t,t+1}\right)C_{t+1}^{-\sigma}\right]=C_t^{-\sigma}.$$

On peut réécriire cette équation sous la forme suivante :

$$E_t\left[\beta\left(1+R_{t,t+1}\right)\left(\frac{C_t}{C_{t+1}}\right)^{\sigma}-1\right]=0.$$

On a donc une espérance conditionnelle par rapport à l'ensemble d'information en t. On cherche à estimer les paramètres structurelles $\theta=(\beta,\sigma)'$. On peut obtenir une espérance non conditionnelle en projetant l'équation d'Euler sur des éléments de l'ensemble d'information en t. Prenons par exemble les instruments suivants $z_t=(C_t/C_{t-1},C_{t-1}/C_{t-2},R_{t-1,t},R_{t-2,t-1})'$. On aura donc comme moments théoriques :

$$Ez_tE_t\left(\varepsilon_{t+1}/z_t\right)=E\left(z_t\varepsilon_{t+1}\right)=0$$

où

$$\varepsilon_{t+1}=\beta\left(1+R_{t,t+1}\right)\left(\frac{C_t}{C_{t+1}}\right)^{\sigma}-1.$$

On peut donc estimer les paramètres structurelles par GMM avec l'équivalent empirique des moments théoriques. On pourra également effectuer un test de spécification puisque nous avons quatre moments et deux paramètres à estimer.

Problèmes :

– Instruments faibles : le choix des instruments est très important pour les propriétés de l'estimateur GMM en petit et en grand échantillons. Les variables instrumentales doivent être corrélées avec les variables dépendantes sinon le conportement en petit et grand échantillons est très mauvais (voir Staiger et Stock (1997)). Ces deux auteurs proposent d'effectuer une régression des variables dépendantes sur les instruments et de calculer une statistique F pour la significativité des instruments dans cette régression.

– Biais : le biais de l'estimateur GMM augmente avec le nombre de moments. Newey et Smith (2004) donnent une justification asymptotique à ce comportement.
– L'estimateur n'est pas invariant à la normalisation choisie.

Hansen, Heaton et Ogaki (1996) ont proposé un estimateur GMM invariant à la normalisation choisie. Cet estimateur est appelé "Continuous Updated Estimator" (CUE). Il prend la forme suivante :

$$\hat{\theta}_T = \arg\min_{\theta\in\Theta} \left(\frac{1}{T}\sum_{t=1}^{T} f(x_t,\theta)\right)' W_T(\theta) \left(\frac{1}{T}\sum_{t=1}^{T} f(x_t,\theta)\right).$$

La matrice de poids est donc estimée conjointement. La procédure d'estimation s'effectue en une étape. Newey et Smith (2004) montrent également que le biais de cet estimateur n'augmente pas avec le nombre de conditions de moments. Cet estimateur est cependant plus fragile numériquement.