Économétrie

Inspiré du cours de Bruno Crepon

21 février 2003

1 Introduction : le modèle linéaire

On considère le modèle :

$$y = b_0 + x_1 b_1 + \cdots + x_k b_k + u$$

Où

$$\begin{cases} y &= \text{variable dépendante} \\ x &= \text{variables explicatives} \\ u &= \text{terme d'erreur} \\ b &= \text{inconnue du problème} \end{cases}$$

Le but de l'économétrie est d'estimer ce modèle, c'est-à-dire de trouver une fonction $\hat{b} = b(Y, X)$ qui satisfasse les conditions suivantes :

- sans biais : $\mathbb{E}(\hat{b}) = b$;
- obéissance à un principe, comme le maximum de vraisemblance : si la loi des résidus est connue, on connaît la loi conditionnelle $Y|X$ et on choisit $\hat{b}$;
- optimisation d'un critère, comme min $\left((y - xb)^2\right)$;
- minimisation de $\mathbb{V}\text{ar}(\widehat{b})$.

On travaille sur des données appartenant à trois grands types :

- Données temporelles, par observation du même phénomène dans le temps, c'est-à-dire des variables $y_t, x_t, u_t, t \in [1, \ldots, T]$. T doit alors être moyennement grand, de l'ordre de 50 périodes.
  Exemple (Consommation et revenu). $C_t = \alpha + \beta R_t + \alpha_t + u$
- Données en coupe :$y_i, x_i, u_i, i = 1, ..., N$. N peut être grand, voire très grand (plusieurs milliers d'observations).
  L'ajustement est en général beaucoup moins bon que dans le cas des données temporelles.
  Exemple (Enquête-emploi). On a plus de 150000 personnes enquêtées, avec un grand nombre de questions.
  $\Rightarrow \omega_i = \alpha_0 + \alpha_1 sco_i + \beta_1 exp_i + \beta_2 exp_i^2 + u_i$
  C'est le type de données le plus adapté au calibrage macro-économique.
- Données de panel, doublement indicées :
  $y_{i,t}, x_{i,t}, i = 1, ..., N\text{grand}(> 100), t = 1, ..., T\text{petit}(< 10)$.
  Exemple (Fonctions de production d'entreprises).

  $$Y_{it} = A_{it} K_{it}^{\alpha} L_{it}^{\beta} \rightarrow y_{it} = a_{it} + \alpha k_{it} + \beta l_{it}$$

  Le résidu, dit résidu de Solow, est alors a_{it}, et l'observation unitaire, ou unité statistique le T-uplet $(y_{i1}, ..., y_{iT})$.

1.1 À quoi sert l'estimation ?

Il s'agit de vérifier qu'une variable X a bien un effet sur la variable Y, et de quantifier cet effet.

L'estimation peut aussi avoir un but de simulation. Si la consommation des bien modélisée par $C_t = \alpha_0 + \alpha_1 R_t + \alpha_2 T_t + u_t$, quel est l'effet de prélèvements fiscaux T sur la consommation, autrement dit, quel est le signe de α_2 ? La théorie de l'équivalence ricardienne dit que $\alpha_2 = 0$.

On peut enfin vouloir faire de la prévision : si $Y_t = \widehat{b} X_t$, alors il y a une probabilité, à déterminer, pour que $Y_{t+1} = \widehat{b} X_{t+1}$.

1.2 D'où vient le modèle ?

Le modèle vient de la théorie économique ;

Exemple. – Fonction de production $Y = F(X)$, la théorie donnant une idée de la modélisation, comme $Y = \prod_{k=1}^{K} X_k^{\alpha_k}$.
– fonction translog : $\log C = \log Q + \alpha' \log P_X + \log P_X' \Omega \log P_X$.

Pour pouvoir évaluer le modèle, il faut souvent imposer une restriction stochastique.

Exemple. On spécifie la loi de $u|X$ (en général une loi Normale), ce qui permet d'estimer le modèle, puisque cela donne la loi de $Y|X$.

Comme $\mathbb{E}(u|X) = 0$ est une hypothèse forte, on préfère en général faire des hypothèses moins fortes.

Exemple. $Y^d = \alpha p + X^d \beta_d + u_d, Y^0 = \gamma_p + X^0 \beta_0 + u_0$. On observe (Y, P, X), et on s'intéresse principalement à la première équation avec un choc $u_d = 0$.

Si $p = f(u)$, la connaissance du prix donne une information sur le résidu, donc $\mathbb{E}(u|X) \neq 0$.

On doit donc faire la réduction sctochastique $\mathbb{E}(u_d|X^d, X^0) = 0$.

On peut également essayer de spécifier la loi des observations. Cependant, spécifier la loi de $u_i|X_i$ ne suffit pas. Il faut une hypothèse supplémentaire pour passer à $\mathcal{L}(y_1, ..., y_N|x_1, ..., x_N)$. On peut par exemple supposer que les (y_i, x_i) sont iid.

Table des matières

2 Le Modèle linéaire standard

Modèle : $y_i = b_0 + b_1 x_{1i} + \cdots + b_K x_{ki} + u_i$

2.1 Hypothèses

Hypothèse (H_1). $\mathbb{E}(u_i) = 0$

Hypothèse (H_2). $\mathbb{V}\text{ar}(u_i) = \sigma^2$

Hypothèse (H_3). $i \neq i' \Rightarrow \mathbb{C}\text{ov}(u_i, u_{i'}) = 0$

Ces hypothèses reviennent à dire que les observations sont indépendantes les unes des autres.

Hypothèse (H_4). La matrice des observations X est connue.

Cette hypothèse est étrange : tout se passe comme si on pouvait modifier X à sa guise. Elle n'est cependant pas indispensable sous sa forme forte. On peut en effet l'assouplir en formulant les autres hypothèses paramétrées par la connaissance de X, comme $\mathbb{E}(u_i|X) = 0$.

Hypothèse (H_5). Les vecteurs d'observation X_i sont non colinéaires.

Matriciellement, on écrit ce modèle :

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} X = \begin{pmatrix} 1 & x_{11} & \dots & x_{K1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1N} & \dots & x_{KN} \end{pmatrix} u = \begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix}$$

$$Y = Xb + u$$

Avec les hypothèses :

Hypothèse (H_1). $\mathbb{E}(u) = 0$

Hypothèse (H_2 et H_3). $\mathbb{V}\text{ar}(u) = \sigma^2 I_N$

Hypothèse (H_4). La matrice des observations X est connue.

Hypothèse (H_5). $\text{Rang}(X) = K + 1$, ce qui revient à dire que $X'X$ est inversible.

Démonstration. Supposons $X'X$ non inversible. Alors,
$\exists \lambda \neq 0 / X'X\lambda = 0 \Rightarrow \lambda' X'X\lambda = 0 \Leftrightarrow \|X\lambda\|^2 = 0 \Rightarrow X\lambda = 0$
Il existe donc une combinaison linéaire nulle des X_i, ce qui est contraire à notre hypothèse. ■

2.2 L'Estimateur des moindres carrés ordinaires

2.2.1 Définition

Définition 2.1 (Estimateur MCO). On définit l'estimateur des moindres carrées ordinaires comme :

$$\widehat{b}_{mco} = \text{Arg}\min_b \left(\sum_{i=1}^{N} (y_i - x_i b)^2 \right) = \text{Arg}\min \left(\sum_{i=1}^{N} (u_i(b))^2 \right)$$

Comme $u_i(b) = y_i - x_i b$, on a

$$\widehat{b}_{mco} = \operatorname{Arg\,min}_b \left((Y - Xb)'(Y - Xb)\right)$$

On dit que l'estimateur MCO minimise le critère $c = (Y - Xb)'(Y - Xb)$.

Démonstration.

$$\frac{dc}{db} = -2(Y - Xb)'X = 0$$

$$dc'e = 2c'dc$$

$$dXb = \lambda db \Rightarrow d(Y - Xb) = -Xdb$$

On optimise le critère :

$\widehat{b}$ tel que $-2(Y - X\widehat{b}_{MCO})'X = 0 : K + 1$ équations et $K + 1$ inconnues

$$\begin{array}{rcll} (Y - Xb)'x = 0 & \Leftrightarrow & (Y' - b'X')X = 0 & \\ & \Leftrightarrow & Y'X = b'(X'X) & \overset{H_5}{\Rightarrow} \widehat{b} = (X'X)^{-1}(X'Y) \\ & \Leftrightarrow & (X'X)\widehat{b} = X'Y & \end{array}$$

NB : $(Y - X\widehat{b})'X = 0 \Leftrightarrow$ les résidus sont orthogonaux à X. ■

2.2.2 Interprétation géométrique

Définition 2.2 (Valeurs prédites). Soit
- $\widehat{y} = \widehat{b}X$ la valeur prédite par le modèle ;
- $\widehat{u} = y - \widehat{y}$ les résidus estimés par le modèle.

Proposition 2.1. *Étant données ces définitions,*
- *$\widehat{y}$ est la projection orthogonale de y sur Vect(X) ;*
- *matriciellement, $\widehat{Y} = X\widehat{b} = X(X'X)^{-1}X'Y$.*

Proposition 2.2. *On a alors :*

1. *On pose : $P_X = X(X'X)^{-1}X'$ la matrice de projection orthogonale sur Vect(X). Elle vérifie :*
   - *$P_X = P'_X$;*
   - *$P_X^2 = P_X$.*
2. *On a alors, en notant M_X la matrice de projection orthogonale sur l'orthogonal de Vect(X), avec $\widehat{U} = Y - P_X Y = (I - P_X)Y = M_X Y$. Elle vérifie :*
   - *$M_X = M'_X$;*
   - *$M_X^2 = M_X$.*

Proposition 2.3. *Il s'ensuit :*

1. *$P_X M_X = 0$;*
2. *$\widehat{U}'X = 0$;*
3. *$\widehat{Y}'\widehat{u} = 0$: valeur prédite et résidus estimés sont orthogonaux.*

Démonstration.

1. Vect(X) et Vect$(X)^{\perp}$ sont orthogonaux et supplémentaires ;
2. *idem* ;

3. $\widehat{Y}'\widehat{U} = Y'P_X'M_XY$.

■

Proposition 2.4. *Dans le cas d'un modèle avec terme constant, soit :*

$$\begin{aligned}\overline{Y} &= \frac{1}{N}\sum_{i=1}^{N} Y_i \\ \overline{\widehat{Y}} &= \frac{1}{N}\sum_{i=1}^{N} \widehat{Y}_i\end{aligned}$$

On a alors : $\overline{Y} = \overline{\widehat{Y}}$
De plus,

$$\overline{\widehat{U}} = \frac{1}{N}\sum_{i=1}^{N} \widehat{u}_i = 0$$

.

Démonstration.
Soit $e' = (1, \ldots, 1), e \in \mathcal{M}_{N,1}$

$$\begin{aligned}\overline{Y} &= \frac{1}{N}e'Y \\ \overline{\widehat{Y}} &= \frac{1}{N}e'\widehat{Y}\end{aligned}$$

$\widehat{Y} = P_XY$, donc $\overline{\widehat{Y}} = \frac{1}{N}e'P_XY$, puisque $P_Xe = e$, donc

$$\overline{\widehat{Y}} = \frac{1}{N}P_Xe'Y = \frac{1}{N}e'Y = \overline{Y}$$

De plus, $\overline{Y} = \overline{\widehat{Y}} + \overline{\widehat{U}}$ ■

2.3 Propriétés algébriques de l'estimateur des MCO

Théorème 2.5 (Théorème de Frish et Waught). *Soit* $Y = Xb + u$ *le modèle.*

On pose $X = [X_1X_2] \qquad X \in \mathcal{M}_{N,K+1} \qquad X_1 \in \mathcal{M}_{N,K_1} \qquad X_2 \in \mathcal{M}_{N,K_2}$
On écrit donc le modèle : $Y = X_1b_1 + X_2b_2 + u$

On a alors :

$$\left\{\begin{array}{lcl}\widehat{b}_1 &=& (X_1'M_{X_2}X_1)^{-1}X_1'M_{X_2}Y \\ \widehat{b}_2 &=& (X_2'X_2)^{-1}X_2'(Y - X_1\widehat{b}_1)\end{array}\right.$$

D'où :

$$\begin{aligned}\widehat{b}_1 &= (X_1'M_{X_2}'M_{X_2}X_1)'X_1'M_{X_2}'Y \\ &= \left[(M_{X_2}X_1)'M_{X_2}X_1\right]^{-1}(M_{X_2}X_1)'M_{X_2}Y\end{aligned}$$

Donc $\widehat{b}_1$ *est l'estimateur* MCO *de la régression de* $M_{X_2}Y$*, résidu de la régression de* Y *sur* X_2*, sur* $M_{X_2}X_1$*, matrice des résidus de la régression de* X_1 *sur* X_2*.*

En d'autres termes, l'estimateur $\widehat{b}_1$ peut être obtenu comme la régression du résidu de la régression de Y sur X_2 sur les résidus des régressions des variables présentes dans X_1 sur X_2.

Exemple. Soit le modèle : $Y_{it} = X_{it}b + u_i + u_{it}$ (données de panel), où u_i est un paramètre propre à chaque entreprise.

Pour le modèle complet,

$$\widehat{b}_c = \begin{pmatrix} b \\ u_1 \\ \vdots \\ u_N \end{pmatrix} \in \mathcal{M}_{N+K,1} \qquad X_c = [X, I_N \otimes e_T]$$

Le théorème de Frish-Waught dit qui si
- on régresse Y sur $I_N \otimes e_T$;
- on régresse chacun des x_k sur $I_N \otimes e_T$ et on récupère les différents résidus, qui sont orthogonaux à $I_N \otimes e_T$,

On a alors, en notant $\bar{x_i} = \frac{1}{N}\sum_{t=1}^{T} x_{it}$, on peut sans perte d'information considérer $y_{it} - \bar{y_i} = (x_{it} - \bar{x_i})b + u_{it} + u_i$, les écarts à la moyenne temporelle pour chaque individu.

Autrement dit, le théorème indique que quand on a une foultitude d'indicateurs, on peut se simplifier la vie en régressant d'abord les variables explicatives sur les indicatrices.

Démonstration. On part des équations normales pour ce modèle :

$$X'(Y - Xb) = 0 \Leftrightarrow (X_1 X_2)'(Y - X_1\widehat{b}_1 - X_2\widehat{b}_2) = 0$$

D'où

$$X_1'(Y - X_1\widehat{b}_1 - X_2\widehat{b}_2) = 0 \tag{1}$$

$$X_2'(Y - X_1\widehat{b}_1 - X_2\widehat{b}_2) = 0 \tag{2}$$

On considère d'abord (2) :

$$\begin{array}{lll} X_2'(Y - X_1\widehat{b}_1) - (X_2'X_2)\widehat{b}_2 = 0 & \Rightarrow & \widehat{b}_2 = (X_2'X_2)^{-1}X_2'(Y - X_1\widehat{b}_1) \\ & \Rightarrow & X\widehat{b}_2 = X_2(X_2'X_2)^{-1}X_2'(Y - X_1\widehat{b}_1) \\ & \Rightarrow & X\widehat{b}_2 = P_{X_2}(Y - X_1\widehat{b}_1) \end{array}$$

On réintègre cela dans (1) :

$$X_1'(Y - X_1\widehat{b}_1 - P_{X_2}(Y - X_1\widehat{b}_1)) = 0 \Leftrightarrow X_1'(I - P_{X_2})(Y - X_1\widehat{b}_1) = 0$$

$$\begin{array}{ll} \Leftrightarrow & X_1'M_{X_2}(Y - X_1\widehat{b}_1) = 0 \\ \Leftrightarrow & \widehat{b}_1 = (X_1'M_{X_2}X_1)^{-1}X_1'M_{X_2}Y \\ \Leftrightarrow & \widehat{b}_1 = \left[(M_{X_2}x_1)'(M_{X_2}X_1)\right]^{-1}(M_{X_2}X_1)'M_{X_2}Y \end{array}$$

On purge ainsi X_1 des variables de X_2 corrélées avec X_1. ■

Remarque. Soient les modèles : $Y = X_1\widehat{b}_1 + X_2\widehat{b}_2 + u$ et $Y = X_1\widehat{b}_1 + v$ L'estimateur $\widehat{b}_1$ issu du seul second modèle est non biaisé $\Leftrightarrow M_{X_2}X_1 = X_1$, c'est-à-dire $X_1 \perp X_2$. C'est pourquoi on commence par régresser X_1 sur X_2 et qu'on prend le résidu.

2.4 Propriétés statistiques de l'estimateur MCO

Proposition 2.6. *$\widehat{b}_{MCO}$ est sans biais.*

Démonstration.
– Si X est connu :

$$\begin{aligned}\widehat{b}_{MCO} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}(X'Xb + X'u) \\ &= b + (X'X)^{-1}X'u\end{aligned}$$

Donc :

$$\mathbb{E}(\widehat{b}_{MCO}) = \mathbb{E}\left((X'X)^{-1}X'u)\right) \overset{H_1}{=} \mathbb{E}(b)$$

– Si X est inconnu, on a par le même calcul, $\mathbb{E}(\widehat{b}_{MCO}|X) = b$.

■

Proposition 2.7. $\mathbb{V}\text{ar}(\widehat{b}_{MCO}) = \sigma^2(X'X)^{-1}$

Démonstration.
$\mathbb{V}\text{ar}(\widehat{b}) = \mathbb{E}\left[(\widehat{b} - b)(\widehat{b} - b)'\right]$

Comme $\widehat{b} = (X'X)^{-1}X'Y$, $\widehat{b} - b = (X'X)^{-1}X'u$.

Donc $\mathbb{V}\text{ar}(\widehat{b}|X) = \mathbb{E}\left[(X'X)^{-1}X'uu'X(X'X)^{-1}|X\right]$. Or, d'après H_2 et H_3, $\mathbb{E}(uu') = \sigma^2 I$.

Si X est aléatoire, on a : $\mathbb{V}\text{ar}(\widehat{b}) = \sigma^2\mathbb{E}_X\left((X'X)^{-1}\right)$. ■

Exemple (Le modèle linéaire simple $y = xb + u$). Supposons les variables centrées : $\mathbb{E}(y) = \mathbb{E}(x) = 0$.

On a alors : $x'x = \sum x_i^2 = \frac{1}{N}\sum \frac{x_i^2}{N} = N\sigma_x^2$ Donc, $\mathbb{V}\text{ar}(\widehat{b}) = \frac{\sigma^2}{N\sigma_x^2}$, donc quand N augmente, $\mathbb{V}\text{ar}(\widehat{b})$ décroît au rythme de $1/N$, ce qui signifie que σ décroît en $1/\sqrt{N}$, qui est la vitesse standard de convergence des estimateurs.

En outre, σ_x^2 joue un rôle essentiel. Si $\sigma_x^2 = 0$, $\widehat{b}_{MCO}$ n'a pas de sens : il faut que la variable explicative soit suffisemment dispersée.

Exemple (Modèle à deux variables explicatives). On a x_1 et x_2, avec $\sigma_{x_1}^2 = \sigma_{x_2}^2$ et $\mathbb{C}\text{ov}(x_1, x_2) = \rho\sigma^2$.

$$(X'X)^{-1} = \frac{1}{N\sigma^2(1-\rho^2)}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$$

Si ρ est proche de 1, les estimateurs sont très imprécis.

2.5 Optimalité de $\widehat{b}_{MCO}$

Définition 2.3 (Critère d'optimalité). On prend comme critère d'optimalité la minimisation de la variance.

Soit $\widetilde{b}$ un estimateur de b. On dit que $\widetilde{b}$ est optimal ssi $\forall\lambda, \mathbb{V}\text{ar}(\lambda'\widetilde{b})$ est minimale, c'est-à-dire que la variance de toute compostion linéaire des composantes est minimale.

Théorème 2.8 (Théorème de Gauss-Markov). *Sous les hypothèses H_1 à H_5, dans la classe des estimateurs de b linéaires dans les variables à expliquer et sans biais, $\widehat{b}_{MCO}$ est optimal au sens du critère de minimisation de la variance.*

Démonstration.
$\widetilde{b}$ linéaire en $Y \Leftrightarrow \widetilde{b} = AY$
$\widetilde{b}$ sans biais $\Leftrightarrow \mathbb{E}(AXb + Au) = b$
Comme $\mathbb{E}(u) = 0$, $AXb = b, \forall b, \Rightarrow AX = I$.
En outre, $\widetilde{b} - b = AY - b = AXb + Au - b = Au$,
donc $\mathbb{V}\text{ar}(\widetilde{b}) = \mathbb{E}\left((\widetilde{b} - b)(\widetilde{b} - b)'\right) = \mathbb{E}(Auu'A') = A\mathbb{E}(uu')A'$.
Or, on a supposé que $\mathbb{E}(uu') = \sigma^2 I$, donc $\mathbb{V}\text{ar}(\widetilde{b}) = \sigma^2 AIA'$.
Écrivons

$$I = P_X + M_X \qquad \left\{ \begin{array}{rcl} P_X & = & X'(X'X)^{-1}X' \\ M_X & = & I - P_X \end{array} \right.$$

$$\mathbb{V}\text{ar}(\widetilde{(b)}) = \sigma^2(AP_XA' + AM_XA')$$

Or, $\sigma^2 AP_XA' = \sigma^2 AX(X'X)^{-1}X'A'$.

Comme $\widetilde{(b)}$ est sans biais, $AX = I = X'A'$, donc $\sigma^2 AP_XA' = \sigma^2(X'X)^{-1}$, et donc :

$$\mathbb{V}\text{ar}(\widetilde{b}) = \underbrace{\sigma^2(X'X)^{-1}}_{\mathbb{V}\text{ar}(\widehat{b}_{MCO})} + AM_XA'$$

Comme, AM_XA' est symétrique définie positive, on a :

$$\forall \lambda, \mathbb{V}\text{ar}(\lambda'\widetilde{b}) = \mathbb{V}\text{ar}(\lambda'\widehat{b}_{MCO}) + \sigma^2(A'\lambda')M_X(A'\lambda)$$

Donc $\mathbb{V}\text{ar}(\lambda'\widetilde{b}) \geqslant \mathbb{V}\text{ar}(\widehat{b}_{MCO})$.

Il faut noter que cette démonstration repose très fortement sur l'homoscédasticité de u. ■

2.6 Estimation de σ^2

Il est important de bien estimer ce paramètre, car $\mathbb{V}\text{ar}(\widehat{b}_{MCO}) = \sigma^2(X'X)^{-1}$ en dépend. On va avoir :

$$\widehat{\mathbb{V}\text{ar}(\widehat{b})} = \widehat{\sigma^2}(X'X)^{-1}$$

D'autre part, $\widehat{\sigma^2} = \mathbb{V}\text{ar}(u)$, et donc constitue une mesure de la qualité de l'ajustement.

Définition 2.4.

$$\widehat{\sigma^2}_{MCO} = \frac{\sum \widehat{u_i^2}}{N - K - 1} = \frac{\widehat{u}'\widehat{u}}{N - K - 1}$$

Proposition 2.9 (Propriétés de $\widehat{\sigma^2}_{MCO}$). *$\widehat{\sigma^2}_{MCO}$ vérifie :*

1. $\mathbb{E}(\widehat{\sigma^2}_{MCO}) = \widehat{\sigma^2}$ *: $\widehat{\sigma^2}_{MCO}$ est sans biais ;*
2. *$\widehat{u}$ et $\widehat{b}_{MCO}$ sont non corrélés.*

Démonstration.

1. Sans biais :
$$\widehat{\sigma^2}_{MCO} = \frac{\widehat{u}'\widehat{u}}{N-K-1} = \frac{u'M_Xu}{N-K-1}$$
Or, $u'M_Xu$ est un scalaire, donc $u'M_Xu = \mathrm{Tr}(u'M_Xu) = \mathrm{Tr}(M_Xuu')$. Donc,
$$\mathbb{E}_X(\widehat{\sigma^2}_{MCO}) = \frac{\mathbb{E}(\mathrm{Tr}(M_Xu'u))}{N-K-1} = \frac{\mathrm{Tr}(M_X\mathbb{E}(uu'|X))}{N-K-1}$$
Or, $\mathbb{E}(uu'|X) = \sigma^2 I$, donc $\mathbb{E}_X(\widehat{\sigma^2}_{MCO}) = \frac{\sigma^2\mathrm{Tr}(M_X)}{N-K-1}$.
Comme M_X est la matrice de projection sur un espace de dimension $N-K-1$, $\mathrm{Tr}(M_X) = N-K-1$, donc $\mathbb{E}_X(\widehat{\sigma^2}_{MCO}) = \sigma^2$.
2. Non-corrélation :
$$\begin{aligned}\mathbb{E}_X(\widehat{u}\underbrace{(\widehat{b}-b)'}_{\text{on centre}}) &= \mathbb{E}_X(M_Xuu'X(X'X)^{-1}) \\ &= M_X\mathbb{E}_X(u'u)X(X'X)^{-1} \\ &= \sigma^2 M_X X(X'X)^{-1}\end{aligned}$$
Comme $M_XX = 0$, $\mathbb{E}_X = (u'(\widehat{b}-b)') = 0$.
Les paramètres du premier et du second ordre sont donc indépendants.

■

2.7 Application à la prévision

Modèle : $\left\{\begin{array}{ccl} Y_i & = & bX_i + u_i \\ H_1 & \text{à} & H_5 \\ N & & \text{observations}\end{array}\right.$

On suppose que pour une observation $N+1$, le modèle reste vrai :

$$Y_{N+1} = bX_{N+1} + u_{N+1}$$

$$H_1\text{à}H_5 : \left\{\begin{array}{rcll}\mathbb{E}(U_{n+1}) & = & 0 & \\ \mathbb{C}\mathrm{ov}(u_{N+1}, u_i) & = & 0 & \forall i = 1\dots N\end{array}\right.$$

On connaît donc X_{N+1}, et on veut prévoir Y_{N+1}.

Définition 2.5. La prévision MCO de Y est :

$$Y^p_{N+1} = X_{N+1}\widehat{b}_{MCO}$$

Proposition 2.10. *$Y^p_{N+1} = X_{N+1}\widehat{b}_{MCO}$ est le meilleur prédicteur linéaire en Y sans biais de Y_{N+1}.*

Démonstration.
– Sans biais :
$$\begin{aligned}\mathbb{E}(Y^p_{N+1} - Y_{N+1}) &= \mathbb{E}(X_{N+1}\widehat{b}_{MCO} - X_{N+1}b - u_{N+1} \\ &= X_{N+1}\mathbb{E}(\widehat{b}_{MCO} - b) - \mathbb{E}(u_{N+1}) \\ &= 0\end{aligned}$$

– Soit $\widetilde{Y}_{N+1}$ prédicteur linéaire sans biais de Y_{N+1}.

$$\mathbb{E}\left((\widetilde{Y}_{N+1} - Y_{N+1})^2\right) = \mathbb{E}\left((\widetilde{Y}_{N+1} - X_{N+1}b + u_i)\right)$$

Comme $\widetilde{Y}_{N+1}$ est une combinaison linéaire des $y_1, \ldots, y_N$, c'en est une des $u_1, \ldots, u_N$, donc $(\widetilde{Y}_{N+1} - X_{N+1}b)$ et u_{N+1} ne sont pas corrélés, d'où

$$\mathbb{E}\left((\widetilde{Y}_{N+1} - Y_{N+1})^2\right) = \mathbb{E}\left((\widetilde{Y}_{N+1} - X_{N+1}b)^2\right) + \mathbb{E}\left((u_{N+1})^2\right)$$

En raison du théorème de Gauss-Markov (2.8), le meilleur estimateur $\widetilde{Y}_{N+1}$ de $X_{N+1}b$ est $X_{N+1}\widehat{b}_{MCO}$.

■

On peut calculer la variance de la prévision :

$$\begin{aligned}\mathbb{V}\text{ar}(Y_{N+1} - X_{N+1}\widehat{b}_{MCO}) &= \mathbb{V}\text{ar}(X_{N+1}(b - \widehat{b}_{MCO}) + u_{N+I}) \\ &= \mathbb{V}\text{ar}\left(X_{N+1}(\widehat{b}_{MCO} - b)\right) + \mathbb{V}\text{ar}(u_{N+1}) \\ &= \sigma^2 X_{N+1}(X'X)^{-1}X'_{N+1} + \sigma^2\end{aligned}$$

Le second terme est l'erreur standard du modèle, le premier représente l'erreur due à l'estimation de b sur les seuls $x_1, \ldots, x_N$.

2.8 Analyse de la variance

Hypothèse. On suppose que la constante est incluse dans les variables explicatives

Théorème 2.11 (Décomposition de la variance). *Si la constante est incluse dans les variables explicatives, la variance se décompose comme :*

$$\underbrace{\frac{1}{N}\sum\left((y_i - \bar{y})^2\right)}_{\textit{Variance totale}} = \underbrace{\frac{1}{N}\sum\left((\hat{y}_i - \bar{\hat{y}})^2\right)}_{\textit{Variance expliquée}} + \underbrace{\frac{1}{N}\sum u_i^2}_{\textit{Variance résiduelle}}$$

Démonstration.

On a : $y = \hat{y} + \hat{u}$. Comme la constante est incluse dans la régression, $\bar{y} = \bar{\hat{y}}$, et $\bar{\hat{u}} = 0$. D'où :

$$y - \bar{y}e = \hat{y} - \bar{\hat{y}}e + \hat{u}$$

$$(y - \bar{y}e)'(y - \bar{y}e) = \sum_{i=1}^{N}\left((y_i - \bar{y}^2\right)$$

$$\left((\hat{y} - \bar{\hat{y}}e) + \hat{u}\right)'\left((\hat{y} - \bar{\hat{y}}e) + \hat{u}\right) = (\hat{y} - \bar{\hat{y}}e)'(\hat{y} - \bar{\hat{y}}e) + \hat{u}'(\hat{y} - \bar{\hat{y}}e) + \hat{u}'\hat{u}$$

Or, $\hat{u} = M_X u$, $\hat{y} = P_X y$ et $e \in X \Rightarrow \hat{u}'(\hat{y} - \bar{\hat{y}}) = u'M_X(P_X y - e\bar{\hat{y}})$

Or, $M_X P_X = 0$, d'où le résulat. ■

Cette équation permet de définir une mesure synthétique de l'ajustement du modèle :

Définition 2.6 (R^2).

$$R^2 = \frac{\text{Variance expliquée}}{\text{Variance totale}} = \frac{\sum\left((\hat{y}_i - \bar{\hat{y}})^2\right)}{\sum\left((y_i - \bar{y})^2\right)}$$

Du fait du théorème de décomposition, $R^2 \in [0,1]$, et

$$R^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum ((y_i - \bar{y})^2)}$$

Comme R^2 fait intervenir la variance de Y, il est sensible à la forme de la modélisation. Ainsi, si on compare les deux modèles :

$$y = \alpha \log(L) + \beta \log(K) + u \tag{3}$$

$$y - l = (\alpha - 1 = \log L + \beta \log K + u \tag{4}$$

Le modèle (3) aura une variance beaucoup plus importante que le modèle (4), alors que les deux modélisations (en production ou en productivité par tête) sont équivalentes en termes de théorie économique.

En outre, on a le problème que le R^2 augmente mécaniquement quand la liste des variables explicatives augmentent. On peut cependant essayer de l'améliorer :

$$R^2 = 1 - \frac{\|\hat{u}\|^2}{\|y - \bar{y}e\|^2} \qquad \|\hat{u}\|^2 = \sum {\hat{u}_i}^2 = \widehat{\sigma^2}_{MCO}(N - K - 1)$$

Donc :

$$R^2 = 1 - \frac{\widehat{\sigma^2}_{MCO}(N - K - 1)}{\widehat{\sigma^2}_y (N - 1)}$$

Où : $\widehat{\sigma^2}_y = \frac{\|y - \bar{y}e\|^2}{N-1}$ est un estimateur non biaisé de $\mathbb{V}\mathrm{ar}(Y)$. En conséquence :

Définition 2.7 (R^2 ajusté).

$$R^2{}_{\text{ajsuté}} = 1 - \frac{\widehat{\sigma^2}}{\widehat{\sigma^2}_y}$$

On se débarasse ainsi de l'influence des degrés de liberté.

2.9 Le Modèle linéaire statistique

On part du modèle et du jey d'hypothèses de la section précédente. On suppose en outre :

Hypothèse (H_6).

$$u \sim \mathcal{N}(0, \sigma^2)$$

Proposition 2.12 (Propriétés). *Sous H_6, les estimateurs* MCO *vérifient les propriétés suivantes :*

1. $\widehat{b}_{MCO} \sim \mathcal{N}(b, \sigma^2 (X'X)^{-1})$
2. *Loi de $\widehat{\sigma^2}$:*
   $u = M_X u \Rightarrow \hat{u} \sim \mathcal{N}(.,.)$, *d'où :*

$$\begin{pmatrix} \widehat{b} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} (X'X)^{-1}X'Y \\ M_X u \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix} + \begin{pmatrix} (X'X)^{-1}X' \\ M_X u \end{pmatrix} u$$

La loi jointe de $(\widehat{b}, \hat{u})$ est une loi normale, or $\widehat{b}$ et $\hat{u}$ ne sont pas corrélés, donc $\widehat{b}$ est indépendant de $\hat{u}$.
Or, $\widehat{\sigma^2} = \frac{\|\hat{u}\|^2}{N-K-1}$ et $\hat{u} \perp \widehat{b}$, donc $\widehat{\sigma^2}$ est indépendant de $\widehat{b}$.
Alors,

$$(N-K-1)\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2(N-K-1)$$

Démonstration de la loi de $\widehat{\sigma^2}$.

Lemme. Si $Z \sim \mathcal{N}(0, I_L) \Rightarrow {Z_1}^2 + \cdots + {Z_L}^2 \sim \chi^2(L)$.
Soit P un projecteur sur un espace de dimensionL_1, alors :

$$Z'PZ \sim \chi^2(L_1)$$

Démonstration. P est diagonalisable dans le groupe orthogonal : $\exists D$ diagonale et Q orthogonale lettes que $P = Q'DQ$, avec :

$$D = \begin{pmatrix} I_{L_1} & 0 \\ 0 & 0 \end{pmatrix}$$

D'où : $Z'PZ = Z'Q'DQZ$. On pose : $Z^* = QZ$, et donc $Z'PZ = {Z^*}'DZ^*$.

$$\mathbb{V}\text{ar}(Z^* = \mathbb{E}(Z^*{Z^*}') = Q\mathbb{E}(ZZ')Q' = QQ' = I \Rightarrow Z^* \sim \mathcal{N}(0, I)$$

Donc ${Z^*}'DZ^* = Z_1^* + \cdots + Z_{L_1}^* \sim \chi^2(L_1)$. ■

$$(N-K-1)\frac{\widehat{\sigma^2}_{MCO}}{\sigma^2} = (N-K-1)\frac{\hat{u}'\hat{u}}{\sigma^2} = (N-K-1)\frac{u'M_Xu}{\sigma^2}$$

$$u \sim \mathcal{N}(0, \sigma^2 I) \Rightarrow v = \frac{u}{\sigma} \sim \mathcal{N}(0,1)$$

$$(N-K-1)\frac{\widehat{\sigma^2}_{MCO}}{\sigma^2} = (N-K-1)\frac{v'M_Xv}{N-K-1} = v'M_Xv$$

Le lemme donne le résultat voulu. ■

2.9.1 Intervalles de confiance

Définition 2.8 (Intervalle de confiance). Un intervalle de confiance au seuil $(1-\alpha)$ pour un paramètre b_k est la données d'un intervalle $[a_1, a_2]$ tel que : $P(b_k \in [a_1, a_2]) = (1-\alpha)$.

Proposition 2.13. *Soit v_x^k le $k^{ième}$ élément de la diagonale de $(X'X)^{-1}$.*

$$\frac{\widehat{b}_k - b_k}{\hat{\sigma}\sqrt{v_x^k}} \sim \mathcal{S}t(N-K-1)$$

Démonstration.
On sait que $\widehat{b} \sim \mathcal{N}(b, \sigma^2(X'X)^{-1})$, donc $\widehat{b}_k \sim \mathcal{N}(b_k, \sigma^2 v_x^k)$ et $\frac{\widehat{b}_k - b_k}{\sigma\sqrt{v_x^k}} \sim \mathcal{N}(0,1)$.

Seulement, σ est un paramètre inconnu, mais on sait que :

$$\frac{\widehat{\sigma^2}}{\sigma^2}(N-K-1) \sim \chi^2(N-K-1)$$

.

Or,

$$\left.\begin{array}{rcl} X & \sim & \mathcal{N}(0,1) \\ Y & \sim & \chi^2(L) \\ X,Y & & \text{indépendantes} \end{array}\right\} \Rightarrow \frac{X}{\sqrt{V/L}} \sim \mathcal{S}t(L)$$

Donc,

$$\frac{\frac{\widehat{b}_k - b_k}{\sigma\sqrt{v_x^k}}}{\sqrt{\frac{(N-K-1)\widehat{\sigma^2}}{(N-K-1)\sigma^2}}} \sim \mathcal{S}t(N-K-1).$$

■

Donc si on cherche un intervalle de confiance au seuil $(1-\alpha)$, on va charcher des bornes $[-t_{1-\alpha/2}, t_{1-\alpha/2}]$ telles que l'intégrale hors de ces bornes soit égale à α.

Si $S \sim \mathcal{S}t(L)$, $P\left(s \in [-t_{1-\alpha/2}, t_{1-\alpha/2}]\right) = 1-\alpha$.

Donc, on connaît $[-t_{1-\alpha/2}, t_{1-\alpha/2}]$ par la lecture d'une table des quantiles de $\mathcal{S}t$.

$$P\left(-t_{1-\alpha/2} < \frac{\widehat{b}_k - b_k}{\sigma\sqrt{v_x^k}} < t_{1-\alpha/2}\right) = 1-\alpha$$

$$P\left(\widehat{b}_k - \sigma\sqrt{v_x^k}t_{1-\alpha/2} < b_k < \widehat{b}_k + \sigma\sqrt{v_x^k}t_{1-\alpha/2}\right) = 1-\alpha$$

On peut se demander quelle est l'influence du nombre de degrés de liberté. Graphiquement, on voit que moins il y a de degrés de liberté, plus la courbe est étalée. Au contraire, quand le nombre de degrés de liberté est très grand, elle tend vers une $\mathcal{N}(0,1)$.

De même, si on considère une combinaison linéaire des paramètres, $\lambda'\widehat{b}$, $\lambda'\widehat{b} \sim \mathcal{N}(\lambda' b, \sigma^2\lambda'(X'X)^{-1}\lambda)$, donc :

$$\frac{\lambda'\widehat{b} - \lambda' b}{\sigma^2\sqrt{\lambda'(X'X)^{-1}\lambda}} \sim \mathcal{S}t(N-K-1)$$

2.10 Test d'hypothèses

On a une hypothèse $H_0 : b_k = b_k^0$ avec b_k^0 une valeur donnée, et $H_1 = \bar{H}_0$.

On définit une region critique W à un seuil $1-\alpha$ donné telle que :

$$\tilde{b}_k \in W \Rightarrow \text{on rejette } H_0$$

et

$$P(\tilde{b}_k \in W | b_k = b_k^0) = \alpha$$

α représente donc le risque de rejeter à tort H_0.

On utilise le résultat précedent : sous H_0,

$$\frac{\widehat{b}_k - b_k^0}{\hat{\sigma}\sqrt{v_x^k}} \sim \mathcal{S}t(N-K-1)$$

D'où la région critique :

$$W \text{ telle que : } \left|\frac{\widehat{b}_k - b_k^0}{\hat{\sigma}\sqrt{v_x^k}}\right| > \delta \text{avec } \delta \text{ tel que : } P\left(\left|\frac{\widehat{b}_k - b_k^0}{\hat{\sigma}\sqrt{v_x^k}}\right| > \delta\right) = \alpha$$

δ est le quantile d'ordre $1-\alpha/2 = t_{1-\alpha/2}$ d'une $\mathcal{S}t(N-K-1)$. D'où :

$$W = \left\{ \begin{array}{rcl} \widehat{b}_k & > & b_k^0 + \hat{\sigma}\sqrt{v_x^k} t_{1-\alpha/2}(N-K-1) \\ \widehat{b}_k & < & b_k^0 - \hat{\sigma}\sqrt{v_x^k} t_{1-\alpha/2}(N-K-1) \end{array} \right.$$

2.11 Estimateurs MCO et estimateurs du maximum de vraisemblance

On sait que :

$$\mathcal{L}(y_i|x_i;b) = \frac{e^{-\frac{(y_i - x_i b)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$$

Par indépendance des observations,

$$\mathcal{L}(y|x,b) = \prod_i l(y_i|x_i,b)$$

$$\Rightarrow \mathcal{L}(y|x,b) = \frac{e^{-\frac{\sum (y_i - x_i b)^2}{2\sigma^2}}}{(\sqrt{2\pi}\sigma)^N}$$

$$\Rightarrow \ln\left(\mathcal{L}(y|x,b)\right) = -\frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum (y_i - x_i b)^2 - N\ln\left(\sqrt{2\pi}\right)$$

L'estimateur du maximum de vraisemblance de b réalise donc le programme :

$$\max_b \left\{ \frac{1}{2\sigma^2}\sum (y_i - x_i b)^2 \right\}$$

Il s'agit donc de l'estimateur MCO.

L'estimateur de $\widehat{\sigma^2}$ est obtenu par :

$$\max_{\sigma^2} \left\{ -\frac{N}{2}\ln\sigma^2 - \frac{\sum \left((y_i - x_i\widehat{b})^2\right)}{2\sigma^2} \right\}$$

D'où :

$$\widehat{\sigma^2}_{MV} = \frac{\sum \left((y_i - x_i\widehat{b})^2\right)}{N} = \frac{N-K-1}{N}\widehat{\sigma^2}_{MCO}$$

3 Estimation sous contraintes linéaires

3.1 Introduction

On souhaite estimer un modèle économétrique linéaire en incorporant une *information a priori sur les paramètres* prenant la forme de *contraintes linéaires.*

Exemple. Fonction de production Cobb-Douglas à k facteurs, et à rendements d'échelle constants :

$$\log\ y = \log\ \alpha + \beta_1\ \log\ x_1 + ... + \beta_k\ \log\ x_k + u$$

c.a.d un modèle linéaire standard
mais avec $\sum_{j=1}^{k}\ \beta_j = 1$

3.1.1 Questions :

1. Comment tenir compte de cette information a priori dans la procédure d'estimation des paramètres du modèle ?
   → On va introduire un nouvel estimateur : l'estimateur des moindres carrés contraints : $\widehat{b}_c$
2. Quelles sont les conséquences de cette prise en compte pour les estimations obtenues ? Les estimations sont-elles biaisées, sont elles plus précises ? → On va voir qu'il y a un *arbitrage entre robustesse et efficacité*
3. Peut-on tester l'information a priori ?
   → On va introduire un test très courant : *Le test de Fisher*

3.1.2 Formulation : Exemple

Supposons qu'on souhaite estimer le modèle :

$$y_n = b_0 + b_1\ x_{1n} + b_2\ x_{2n} + b_3\ x_{3n} + b_4\ x_{4n} + b_5\ x_{5n} + b_6\ x_{6n} + u_n,$$

avec les hypothèses habituelles
$H_1 : E\,(u_n\,|X\,) = 0, H_2 : V\,(u_n\,|X\,) = \sigma^2,\ \forall n,$
$H_3 : E(u_n u_{n'}\,|X\,) = 0,\ \forall\ n' \neq n,$
H_4 : X de plein rang
avec des contraintes linéaires sur les paramètres :

$$\begin{array}{ll} C_1 : b_1 + b_2 + b_3 = 1 & \\ C_2 : b_4 = b_5 & \text{soit } b_4 - b_5 = 0 \end{array}$$

3.1.3 Réécriture sous forme matricielle :

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

soit

$$R\, b = r$$

avec R une matrice $2 \times (6+1)$ et r un vecteur 2×1

3.1.4 Formulation générale

On considère le modèle linéaire :

$$y = X\, b + u$$

sous les contraintes

$$\underset{(p,k+1)}{R} \quad \underset{(k+1,p)}{b =} \quad \underset{(p,1)}{r}$$

Le nombre de contraintes p doit être *au maximum égal à* $(k+1)-1$. Si on en a $k+1$ ou plus, on en sélectionne $k+1$ et on peut alors *calculer* le paramètre $b = R^{-1}\, r$

3.2 L'Estimateur des Moindres Carrés Contraints (MCC)

L'estimateur $\hat{b}_{mcc}$ de b est défini comme celui minimisant la somme des carrés des résidus sous *les contraintes* :

$$\begin{array}{l} \min_b \left((y - Xb)'(Y - Xb)\right) \\ \text{Sous les contraintes } Rb = r \end{array}$$

Lagrangien :

$$\min_{b,\lambda} L = (Y - Xb)'(Y - Xb) + 2(Rb - r)'\lambda$$

λ multiplicateur de Lagrange : vecteur de dimension $p \times 1$

3.2.1 Expression de l'estimateur des MCC

L'estimateur des MCC a pour expresssion

$$\hat{b}_{mcc} \quad = \quad (X'X)^{-1}\, X'Y - (X'X)^{-1}\, R' \left[R(X'X)^{-1}\, R'\right]^{-1} \left[R(X'X)^{-1}\, X'Y - r\right]$$

Il s'exprime simplement à partir de $\hat{b}_{mco}$

$$\hat{b}_{mcc} = \hat{b}_{mco} - (X'X)^{-1}\, R' \left[R(X'X)^{-1}R'\right]^{-1} \left[R\, \hat{b}_{mco} - r\right]$$

L'estimateur des MCC apporte une correction à l'estimateur $\hat{b}_{mco}$ d'autant plus importante que $R\hat{b}_{mco} - r \neq 0$.

Si $R\hat{b}_{mco} = r$, les deux estimateurs sont identiques.

Démonstration.

$$\begin{aligned} \left.\frac{\partial L}{\partial b}\right|_{mcc} &= -2\ X'\ Y + 2\ X'\ X\ \hat{b}_{mcc} + 2\ R'\ \hat{\lambda} = 0 \\ \left.\frac{\partial L}{\partial \lambda}\right|_{mcc} &= R\hat{b}_{mcc} - r = 0 \end{aligned}$$

De la première condition on tire : $\hat{b}_{mcc} = (X'X)^{-1}\left(X'Y - R'\ \hat{\lambda}\right)$

Introduit dans la deuxième condition il vient l'expression

$R\ (X'X)^{-1}\left(X'Y - R'\ \hat{\lambda}\right) = r$ soit $R\ (X'X)^{-1} R'\ \hat{\lambda} = R\ (X'X)^{-1} X'Y - r$

dont on tire $\hat{\lambda} = \left[R\ (X'X)^{-1} R'\right]^{-1}\left[R\ (X'X)^{-1} X'Y - r\right]$

réintroduit dans on trouve l'expression de $\widehat{b}_{mcc}$

$$\hat{b}_{mcc} = (X'X)^{-1}\ X'Y - (X'X)^{-1}\ R'\left[R(X'X)^{-1}\ R'\right]^{-1}\ \left[R(X'X)^{-1}\ X'Y - r\right]$$

■

3.2.2 Propriétés Statistiques de $\hat{b}_{mcc}$.

Proposition 3.1 (Expression de l'espérance de $\hat{b}_{mcc}$).

$$E\ \left(\hat{b}_{mcc}\,|X\right) = b - (X'X)^{-1}\ R'\left[R(X'X)^{-1}\ R'\right]^{-1}\ [Rb - r]$$

- *Si les contraintes $Rb = r$ sont valides*, l'estimateur $\hat{b}_{mcc}$ est *sans biais*

$$E\ \left(\hat{b}_{mcc}\,|X\right) = b$$

- *Si ces contraintes sont imposés à tort* (i.e. si $Rb \neq r$), l'estimateur des MCC est *biaisé* :

$$\begin{aligned} E\ \left(\hat{b}_{mcc}\,|X\right) &= b - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right][Rb - r] \\ &= b + B \end{aligned}$$

avec $B = -(X'X)^{-1}R'\left[R(X'X)^{-1}R'\right][Rb - r]$

Proposition 3.2 (Expression de la variance de $\hat{b}_{mcc}$). *Que l'estimateur soit biaisé ou non sa variance est donnée par :*

$$V\ \left(\hat{b}_{mcc}\,|X\right) = \sigma^2\ \left[(X'X)^{-1} - (X'X)^{-1}\ R'\ \left[R\ (X'X)^{-1}\ R'\right]^{-1}\ R\ (X'X)^{-1}\right]$$

soit :

$$V\ \left(\hat{b}_{mcc}\,|X\right) = V\ \left(\hat{b}_{mco}\,|X\right) - \sigma^2(X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}$$

Comme $(X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}$ est une matrice symétrique et positive on en conclut que

$$V\left(\hat{b}_{mco}\,|X\right) \preceq V\left(\hat{b}_{mcc}\,|X\right)$$

3.2.3 Interprétation

L'estimateur des mcc $\hat{b}_{mcc}$ est potentiellement biaisé

$$E\left(\hat{b}_{mcc}\,|X\right) = b + B$$

mais est toujours plus efficace que l'estimateur des mco

$$V\left(\hat{b}_{mcc}\,|X\right) \preceq V\left(\hat{b}_{mcc}\,|X\right)$$

Il y a donc un arbitrage entre robustesse et efficacité. Introduire plus de contraintes améliorent la précision des estimations mais risque de conduire à des estimateurs biaisé.

A l'inverse, moins de contrainte produit des estimateurs plus robustes mais moins précis.

Démonstration.

En remplaçant Y par $(Xb+U)$, dans l'expression de $\hat{b}_{mcc}$ on peut ré-écrire l'estimateur des MCC comme :

$$\hat{b}_{mcc} = b+(X'X)^{-1}\,X'U-\,(X'X)^{-1}\,R'\left[R(X'X)^{-1}R'\right]^{-1}\,\left[R(X'X)^{-1}X'u+Rb-p\right]$$

soit

$$\begin{aligned}\hat{b}_{mcc} &= b-\,(X'X)^{-1}\,R'\left[R(X'X)^{-1}R'\right]^{-1}\,[Rb-p]\\ &\quad+\left[(X'X)^{-1}\,X'-\,(X'X)^{-1}\,R'\left[R(X'X)^{-1}R'\right]^{-1}\,R(X'X)^{-1}X'\right]U\\ &= b+B+(X'X)^{-1}\left[\,I-\,R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}\right]X'U\\ &= b+B+(X'X)^{-1}\,[\,I-\,C]\,X'U\end{aligned}$$

où $B = -\,(X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}[Rb-p]$ et
$C = R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}$

– *Expression de l'espérance de* $\hat{b}_{mcc}$ Compte tenu de $H_1E\,(U\,|X\,)=0$

$$E\,\left(\hat{b}_{mcc}\,|X\right) = b-(X'X)^{-1}\,R'\left[R(X'X)^{-1}\,R'\right]^{-1}\,[Rb-r] = b+B$$

– *Expression de la variance de* $\hat{b}_{mcc}$

$$\hat{b}_{mcc}-E\left(\hat{b}_{mcc}\,|X\right) = (X'X)^{-1}\,[\,I-\,C]\,X'U$$

Par conséquent comme $E\,[UU'\,|X\,] = \sigma^2 I$:

$$V\,\left(\hat{b}_{mcc}\,|X\right) = E\left[\left(\hat{b}_{mcc}-E\,\left(\hat{b}_{mcc}\,|X\right)\right)\left(\hat{b}_{mcc}-E\,\left(\hat{b}_{mcc}\,|X\right)\right)'\,|X\right]$$

$$
\begin{aligned}
&= E\left[(X'X)^{-1}\left[\,I - \,C\right]X'UU'X\left[\,I - \,C'\right](X'X)^{-1}\,|X\right] \\
&= \sigma^2(X'X)^{-1}\left[\,I - \,C\right]X'X\left[\,I - \,C'\right](X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}\left[\,X'X - \,CX'X - X'XC' + CX'XC'\right](X'X)^{-1}
\end{aligned}
$$

Compte tenu de l'expression de $C = R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}$ on a

$$
\begin{aligned}
CX'X &= R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}X'X \\
&= R'\left[R(X'X)^{-1}R'\right]^{-1}R = CX'X \\
CX'XC' &= CR'\left[R(X'X)^{-1}R'\right]^{-1}R \\
&= R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R \\
&= X'XC' = CX'X
\end{aligned}
$$

Il en résulte que

$$
\begin{aligned}
V\left(\hat{b}_{mcc}\,|X\right) &= \sigma^2(X'X)^{-1}\left[\,X'X - CX'XC'\right](X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}\left[\,X'X - R'\left[R(X'X)^{-1}R'\right]^{-1}R\right](X'X)^{-1} \\
&= \sigma^2\left[\,(X'X)^{-1} - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}\right]
\end{aligned}
$$

■

3.3 Estimateur de la Variance des résidus σ^2

L'estimateur de la variance des résidus est donné par :

$$\hat{\sigma}_c^2 = \frac{\widehat{U}_c'\widehat{U}_c}{N-(k+1)+p} = \frac{\sum_n \widehat{u}_{nc}'\widehat{u}_{nc}}{N-(k+1)+p}$$

C'est un estimateur sans biais de σ^2 si les contraintes $Rb = r$ sont satisfaites par le vrai modèle.

Démonstration.

A partir de l'expression de $\hat{b}_{mcc} = b + B + (X'X)^{-1}\left[\,I - \,C\right]X'U$ où $C = R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}$, on exprime le residu estimé

$$
\begin{aligned}
\widehat{U}_c &= Y - X\,\hat{b}_{mcc} \\
&= Xb + U - X\,\left(b + B + (X'X)^{-1}\left[\,I - \,C\right]X'U\right) \\
&= -XB + \left[I - X(X'X)^{-1}\left[\,I - \,C\right]X'\right]U \\
&= -XB + (M + X(X'X)^{-1}CX')U = -XB + (M + P_c)U
\end{aligned}
$$

avec $M = \left(I - X\,(X'X)^{-1}\,X'\right)$ et

$$P_c = X(X'X)^{-1}CX' = X(X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}X'$$

Les matrices M et P_c satisfont les propriétés suivantes :

$$\begin{array}{ll} M = M' & P_C = P_C' \\ M^2 = M & P_C^2 = P_C \\ Tr\,(M) = N - (K+1) & Tr\,(P_C) = p \end{array}$$
$$MP_C = P_C M = 0$$

On vérifie facilement $P_C = P_C'$ et $P_C^2 = P_C$. En outre

$$\begin{aligned} Tr\,(P_C) &= Tr\left(X(X'X)^{-1}\,R'\left[R(X'X)^{-1}R'\right]^{-1}\,R(X'X)^{-1}X'\right) \\ &= Tr\left(\left[R(X'X)^{-1}R'\right]^{-1}\,R(X'X)^{-1}X'X(X'X)^{-1}\,R'\right) \\ &= Tr\left(I_{\dim(R(X'X)^{-1}R')}\right) \end{aligned}$$

d'où $Tr\,(P_C) = p$ enfin comme $P_C = XZ$ on a aussi donc $P_C M = 0$
On en déduit que

$$E\left(\widehat{U}_c'\widehat{U}_c\,|X\right) = E\left(-B'X' + U'(M+P_c)\right)\left(-XB + (M+P_c)\,U\,|X\right)$$

$$\begin{aligned} &= E\left(B'X'XB - U'(M+P_c)XB - B'X'\,(M+P_c)\,U + U'(M+P_c)^2U\,|X\right) \\ &= E\left(B'X'XB + U'(M+P_c)U\,|X\right) \end{aligned}$$

Finalement

$$\begin{aligned} E\left(U'(M+P_c)U\,|X\right) &= TrE\left(U'(M+P_c)U\,|X\right) \\ &= TrE\left((M+P_c)UU'\,|X\right) \\ &= \sigma^2 Tr(M+P_c) = \sigma^2\left(N - (K+1) + p\right) \end{aligned}$$

■

3.4 Estimation par intégration des contraintes

Le problème d'estimation sous contrainte peut se ramener au résultat classique d'estimation par la méthode des moindres carrés en *intégrant directement les contraintes dans le modèle.*

On utilise les p contraintes pour exprimer p paramètres parmi les $k+1$ à estimer en fonction des $(k+1-p)$ autres paramètres.

On ré-écrit les contraintes $Rb = r$ de la façon suivante :

$$r = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = [R_1, R_2] \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

$R_1 : p \times p$, $R_2 : p \times (K+1-p)$,
r_1 et $b_1 : p \times 1$, r_2 et $b_2 : K+1-p \times 1$
R_1 est supposée régulière. On peut alors écrire :

$$r_1 = R_1\,b_1 + R_2\,b_2 \text{ soit encore } b_1 = R_1^{-1}\left[r_1 - R_2\,b_2\right]$$

Par conséquent, en partageant le modèle de façon analogue, on obtient :

$$Y = X_1\,b_1 + X_2\,b_2 + U = X_1\left[R_1^{-1}(r_1 - R_2\,b_2)\right] + X_2\,b_2 + U$$

Ceci revient à estimer :

$$Y - X_1 R_1^{-1}\, r_1 = \left[X_2 - X_1\, R_1^{-1}\, R_2\right]\, b_2 + U$$

Le modèle ainsi écrit ne dépend plus alors que de $(k+1-p)$ *paramètres à estimer sans contraintes*. Les p autres paramètres se déduisent de ceux-ci par la relation : $b_1 = R_1^{-1}\, r - R_2\, b_2$

3.5 Test d'un Ensemble de Contraintes

On souhaite tester la validité des contraintes imposées, soit

$$H_0 : \Delta = Rb - r = 0$$

On fait *l'hypothèse de normalité* des résidus : $U \sim \mathcal{N}(0,\ \sigma^2\, I)$
Sous l'hypothèse H_0 on a

$$\begin{aligned} \widehat{F} &= \frac{1}{p} \frac{\widehat{\Delta}' \left[R(X'X)^{-1}R'\right]^{-1} \widehat{\Delta}}{\widehat{\sigma^2}} = \frac{\widehat{U}_C'\widehat{U}_C - \widehat{U}'\widehat{U}}{\widehat{U}'\widehat{U}} \times \frac{N-(K+1)}{p} \\ &= \frac{SCR_c - SCR}{SCR} \times \frac{N-(k+1)}{p} \sim F(p,N-(k+1)) \end{aligned}$$

où $\widehat{\Delta} = R\hat{b}_{mco} - r$ et $SCR_C = \widehat{U}_C'\widehat{U}_C$ et $SCR = \widehat{U}'\widehat{U}$ sont la somme des carrés des résidus du modèle contraint et non contraint.

Démonstration.

Le principe du test est d'examiner si l'estimateur des mco $\hat{b}_{mco}$ est proche de satisfaire les contraintes, c.a.d il concerne la quantité

$$\widehat{\Delta} = R\hat{b}_{mco} - r,$$

en utilisant le fait que l'on connait la loi de $\widehat{\Delta}$: $\widehat{\Delta} \sim N\ (\Delta, \sigma^2\ R(X'X)^{-1}R')$ puisque $\hat{b}_{mco} \sim N\ (b, \sigma^2\ (X'X)^{-1})$ à cause de l'hypothèse de normalité des résidus.

Rappel :

1. Si Z vecteur de dimension h suit une loi normale $N(0,V)$ avec V inversible alors $Z'V^{-1}Z \sim \chi\,(h)$
2. Si $Q_1 \sim \chi\,(q_1)$ et $Q_2 \sim \chi\,(q_2)$ et $Q_1 \perp Q_2$ alors $Z = \frac{Q_1/q_1}{Q_2/q_2} \sim F(q_1,q_2)$ loi de Fisher à q_1 et q_2 degrés de liberté.

Sous H_0, $\Delta = 0$,on a donc :

$$\widehat{Q}_\Delta = \widehat{\Delta}' \left[\sigma^2\ R(X'X)^{-1}R'\right]^{-1} \widehat{\Delta} = \frac{\widehat{\Delta}' \left[R(X'X)^{-1}R'\right]^{-1} \widehat{\Delta}}{\sigma^2} \sim \chi\,(p)$$

σ^2 est inconnue, on la remplace par $\hat{\sigma}^2 = \frac{\widehat{U}'\widehat{U}}{N-(K+1)}$

On sait qu'en outre $\frac{\widehat{U}'\widehat{U}}{\sigma^2} = (N-(K+1))\ \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-(K+1)}$ et que $\hat{\sigma}^2 \perp \hat{b}_{mco}$ d'où $\frac{\widehat{U}'\widehat{U}}{\sigma^2} \perp \widehat{Q}_\Delta$

sous $H_0 : Rb = r$, la statistique :

$$\begin{aligned}\widehat{F} &= \frac{Q_\Delta/p}{(N-(K+1))\,\frac{\widehat{\sigma}^2}{\sigma^2}/\,(N-(K+1))} = \frac{1}{p}\frac{\widehat{\Delta}'\left[R(X'X)^{-1}R'\right]^{-1}\widehat{\Delta}}{\widehat{\sigma}^2} \\ &\sim F(p, N-(k+1))\end{aligned}$$

■

3.5.1 Expression simplifiée de la statistique

La statistique précédente, fonction de $\hat{b}_{mco}$ et $\hat{\sigma}^2$ peut être réécrite sous une forme plus simple à partir de $\hat{b}_{mco}$ et $\hat{\sigma}^2$ et $\hat{b}_{mcc}$ et $\hat{\sigma}^2_{mcc}$.

En effet : $\hat{b} = (X'X)^{-1}\ X'Y = b + (X'X)^{-1}\ X'U$ donc sous H_0, on a : $\widehat{\Delta} = R\hat{b} - r = R(X'X)^{-1}\ X'U$, d'où

$$\widehat{\Delta}'\left[R(X'X)^{-1}R'\right]^{-1}\widehat{\Delta} = U'X(X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}\ X'U$$

On reconnait $P_C = X(X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}\ X'$

On a donc $\widehat{\Delta}'\left[R(X'X)^{-1}R'\right]^{-1}\widehat{\Delta} = U'P_CU$.

Comme sous H_0 $\widehat{U}_C = (M+P_C)\,U$, et $\widehat{U} = MU$ et $(M+P_C)^2 = (M+P_C)$,on a $\widehat{U}'_C\widehat{U}_C = U'\,(M+P_C)\,U = U'MU + U'P_CU = \widehat{U}'\widehat{U} + U'P_CU$

Soit

$$\widehat{\Delta}'\left[R(X'X)^{-1}R'\right]^{-1}\widehat{\Delta} = U'P_CU = \widehat{U}'_C\widehat{U}_C - \widehat{U}'\widehat{U}$$

D'où l'expression de la statistique communément utilisée :

$$\begin{aligned}\widehat{F} &= \frac{SCR_c - SCR}{SCR} \times \frac{N-(k+1)}{p} \\ &\sim F(p,N\text{-}(k+1))\end{aligned}$$

SCR est la somme des carrés des résidus estimés sans contraintes et SCR_c est la somme des carrés des résidus estimés sous contrainte.

3.5.2 Mise en oeuvre du test

1. On estime le modèle avec et sans contraintes, et on déduit $\widehat{U}'_C\widehat{U}_C$ et $\widehat{U}'\widehat{U}$ (i.e. SCR_c et SCR).
2. On calcule $\widehat{F}$ et on la compare au *fractile d'ordre* $(1-\alpha)$ de la loi $F(p, N-(k+1))$, noté $F(1-\alpha)$.
3. *Si* $Q_c > F(1-\alpha)$; *on rejette* H_0 : la somme des carrés des résidus estimés sous contraintes diffère trop de celle des carrés des résidus estimés sous contrainte pour admettre que H_0 est vraie.
4. Si $Q_c \leq F(1-\alpha)$, on accepte l'hypothèse H_0.

3.5.3 Application : Test de l'égalité à une valeur donnée de plusieurs coefficicents :

On veut tester $H_0 : \left| \begin{array}{l} b_1 = b_1^0 \\ b_2 = b_2^0 \\ \vdots \\ b_J = b_J^0 \end{array} \right.$

contre $H_1 : H_0^c$

C'est à dire un test d'égalité de J coefficients à des valeurs données. La différence avec le test de Student standard est qu'on souhaite faire un test global, sur l'identité simultannée des coefficients

Avec le test de Fisher il suffit d'estimer le modèle non contraint

$$Y = Xb + U$$

de calculer la somme SCR des carrés des résidus estimés, d'estimer le modèle contraint

$$Y - \sum_{k=1}^{k=J} X_k b_k^0 = b_0 e + \sum_{k=J+1}^{k=K} X_k b_k + U$$

de calculer la somme SCR_C des carrés des résidus estimés et de former la statistique

$$\widehat{F} = \frac{N-(K+1)}{J} \frac{SCR_C - SCR}{SCR} \sim F\left(J, N-(K+1)\right)$$

3.6 Test de la significativité globale des coefficients d'une régression

$H_0 : b_1 = b_2 = b_3 = ... = b_K = 0$

Sous H_0, le modèle s'écrit : $Y = b_0\ e + U$, d'où $\hat{b}_0 = \bar{y}$ et $\widehat{U}_c = Y - \bar{y}\ e$. La SCR_c est donc donnée par : $SCR_c = \Sigma_n (y_n - \bar{y})^2$. Sous H_1, $SCR = \widehat{U}'\widehat{U}$. Par conséquent, sous H_0, $\frac{\Sigma_n (y_n - \bar{y})^2 - \widehat{U}'\widehat{U}}{\widehat{U}'\widehat{U}} \times \frac{N-(K+1)}{K} \sim F(K, N-(K+1))$. Or $R^2 = 1 - \frac{\widehat{U}'\widehat{U}}{\Sigma_n (y_n - \bar{y})^2}$, on obtient donc :

$$\widehat{F} = \frac{R^2}{1-R^2} \times \frac{N-(K+1)}{K} \sim F\left(K, N-(K+1)\right)$$

Si $\widehat{F}$ est supérieure au Fractile d'ordre $(1-\alpha)$ de la loi de Fisher à $(K, N-(K+1))$ ddl, on refuse l'hypothèse H_0.

3.7 Le Test de Chow

Question : le modèle est-il homogène entre deux groupes d'observation ?

Exemple, dans le domaine de la consommation, on peut se demander si les comportements de ménages appartenant à divers groupes socio-professionnels sont similaires ou bien si, au contraire, des différences marquées peuvent être constatées.

C'est par la mise en oeuvre du *test de Chow* que l'on peut tenter d'apporter une réponse à ces questions.

3.7.1 Formalisme

Supposons que l'on dispose de *deux échantillons* (Y_1, X_1) et (Y_2, X_2) de tailles respectives N_1 et N_2, relatifs à deux groupes d'observations différents (i.e. deux périodes, deux catégories de ménages,...).

1. Modèle relatif au 1er groupe :$Y_1 = X_1\, b_1 + U_1$
   Y_1 vecteur $N_1 \times 1$ des observations pour le premier groupe
   X_1 matrice$N_1 \times (K+1)$ des variables explicatives $(1, x_1, \ldots, x_K)$pour le premier groupe
2. Modèle relatif au 2ème groupe :$Y_2 = X_2\, b_2 + U_2$
   avec $U_1 \sim N(0, \sigma^2 I_{N_1})$, $U_2 \sim N(0, \sigma^2 I_{N_2})$ et $U_1 \perp U_2 = 0$
   La question posée est de savoir si le comportement modélisé est identique pour les deux groupes d'observations.
   i.e. $H_0 : b_1 = b_2$ contre $H_1 : b_1 \neq b_2$

On *empile* les deux régressions définies ci-dessus. Ceci nous amène à écrire :

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}$$

Le test de Chow est donc un cas particulier du test de Fisher : on test ici *l'égalité de deux groupes de coefficients.*

Par conséquent, on refuse H_0 si

$$\frac{SCR_c - SCR}{SCR} \times \frac{(N_1 + N_2) - 2(K+1)}{(K+1)} > f_{(1-\alpha)}\,(K+1,\; N_1 + N_2 - (K+1))$$

où SCR_C est la somme des carrés des résidus associées à la régresion sous l'hyptothèse $H_0 : b_1 = b_2$, SCR est la somme des carrés des résidus associées à la régression sous l'hypothèse $H_1 = b_1 \neq b_2$.

Si cette inégalité est vérifiée, on rejette l'hypothèse d'homogénéité des comportements.

Simplification du calcul des SCR et SCR_c Sous l'hypothèse $H_0 : b_1 = b_2 = b_0$, on peut écrire :

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} b_0 + \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}$$

On estime donc un seul modèle à partir des deux échantillons pris ensemble et on calcule la somme des carrés des résidus SCR_c

Sous l'hypothèse H_1 on retrouve le modèle défini plus haut :

$$\begin{aligned} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} &= \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \\ &= \widetilde{X} b + U \end{aligned}$$

On vérifie aisément que $M_{\widetilde{X}} = I - \widetilde{X}\left(\widetilde{X}'\widetilde{X}\right)^{-1}\widetilde{X}' = \begin{pmatrix} M_{X_1} & 0 \\ 0 & M_{X_2} \end{pmatrix}$

Donc

$$SCR = Y'M_{\tilde{X}}Y = Y_1'M_{X_1}Y_1 + Y_2'M_{X_2}Y_2 = SCR_1 + SCR_2$$

où SCR_1 est la somme des carrés des résidus associée à la régression sur le premier groupe et idem pour SCR_2. La SCR sous H_1 peut s'obtenir comme sommation des SCR associées aux régressions sur chacun des sous-échantillons.

3.7.2 Principe d'application du test de Chow (sous hypothèse d'homosc édasticité et non-corrélation des résidus).

1. Calculer SCR_c en estimant un seul modèle pour les $N_1 + N_2$ observations.
2. Calculer SCR en estimant le modèle sur chaque échantillon et additionnant les SCR associées à chacune de ces régressions.
3. Comparer la quantité $\frac{SCR_c - SCR}{SCR} \times \frac{N_1+N_2-2(K+1)}{(K+1)}$ au seuil théorique $f(K+1, N_1 + N_2 - 2(K+1))$

4 Propriétés asymptotiques de l'estimateur des MCO

4.1 Rappel sur les convergences

Soit (X_n) une suite de va. Soit F_n la fonction de répartition de X_n. Soit X une va de fonction de répartition F.

Toutes ces va sont définies sur le même espace probabilisé, c'est-à-dire qu'un même événement ω détermine une valeur de $X_n(\omega), X(\omega)$.

4.1.1 Convergence en loi

Définition 4.1. On dit que (X_n) *converge en loi* vers X ($X_n \xrightarrow{L} X$) si la suite de fonctions (F_n) converge, point par point, vers F :

$$\forall x, F_n(x) \to F(x).$$

4.1.2 Convergence en probabilité

Définition 4.2. On dit que (X_n) *converge en probabilité* vers X ($X_n \xrightarrow{P} X$ où $\text{plim}_{n\to\infty} X_n = X$) si

$$\forall \varepsilon > 0,\ \Pr\{|X_n - X| > \varepsilon\} \underset{n\to\infty}{\longrightarrow} 0.$$

(NB : $\Pr\{|X_n - X| > \varepsilon\} = \Pr\{\omega,\ |X_n(\omega) - X(\omega)| > \varepsilon\}$.)

4.1.3 Différents résultats

- $X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{L} X$.
- $\forall a$ constant, $X_n \xrightarrow{P} a \Leftrightarrow X_n \xrightarrow{L} a$.
- $X_n \xrightarrow{L} X$ et $Y_n \xrightarrow{L} Y \Rightarrow X_n + Y_n \xrightarrow{L} X + Y$ et $X_n Y_n \xrightarrow{L} XY$.
- Pour toute fonction g continue, $X_n \xrightarrow{L} X \Rightarrow g(X_n) \xrightarrow{L} g(X)$ et $X_n \xrightarrow{P} a \Rightarrow g(X_n) \xrightarrow{P} g(a)$.

Théorème 4.1 (Théorème de Slutsky).

$$X_n \xrightarrow{L} X \text{ et } Y_n \xrightarrow{P} a \Rightarrow \begin{array}{l} X_n Y_n \xrightarrow{L} Xa \\ X_n + Y_n \xrightarrow{L} X + a \\ X_n / Y_n \xrightarrow{L} X/a \text{ si } a \neq 0 \end{array}.$$

Théorème 4.2 (Loi des grands nombres (Chebichev)). *Soit (X_i) une suite de va indépendantes telles que $\mathrm{E}X_i = m$ et $\mathrm{V}X_i = \Sigma$ existent,*

$$\frac{1}{N}\sum_{i=1}^{N} X_i \xrightarrow{P} m \quad qd\ N \to \infty.$$

Démonstration.
Pour toute va positive X on a le résultat

$$\Pr(X > a) < \frac{E(X)}{a}$$

en effet

$$\begin{aligned} E(X) &= \int_0^a Xf(X)\,dX + \int_a^{+\infty} Xf(X)\,dX > \int_a^{+\infty} Xf(X)\,dX \\ &> a\int_a^{+\infty} f(X)\,dX = a\Pr(X > a) \end{aligned}$$

On a donc

$$\begin{aligned} \Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N} X_i - m\right| > \varepsilon\right) &= \Pr\left(\left(\frac{1}{N}\sum_{i=1}^{N}(X_i - m)\right)^2 > \varepsilon^2\right) \\ &< \frac{E\left[\left(\frac{1}{N}\sum_{i=1}^{N}(X_i - m)\right)^2\right]}{\varepsilon^2} \end{aligned}$$

Comme

$$E\left[\left(\frac{1}{N}\sum_{i=1}^{N}(X_i - m)\right)^2\right] = \frac{1}{N^2}E\left[\left(\sum_{i=1}^{N}(X_i - m)\right)^2\right] = \frac{\Sigma}{N}$$

On voit que

$$\Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N} X_i - m\right| > \varepsilon\right) < \frac{\Sigma}{N\varepsilon^2} \to 0$$

■

4.1.4 Théorème central limite (Lindeberg-Levy)

Théorème 4.3 (Théorème central-limite). *Soit (X_i) une suite de va iid telles que $\mathrm{E}X_i = m$ et $\mathrm{V}X_i = \Sigma$ existent,*

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} X_i - m\right) \xrightarrow{L} \mathcal{N}(0, \Sigma).$$

Démonstration. La demonstration se fait à partir des fonctions caractéristiques. On appelle *fonction caracté ristique* d'une variable aléatoire Z la fonction

$$\phi_Z(t) = E(\exp(it'Z))$$

■

Proposition 4.4 (Propriété d'injectivité). *Si $\phi_{Z_1}(t) = \phi_{Z_2}(t)$ alors $F_{Z_1} = F_{Z_2}$, soit $Z_1 \overset{d}{=} Z_2$*

On peut calculer la fonction de caractéristique d'une loi normale

$$z \sim \mathcal{N}(0, \Sigma) \Leftrightarrow \phi_z(t) = \exp\left(-\frac{t'\Sigma t}{2}\right)$$

On a alors directement avec $\phi_n(t) = E\left(\exp it'\sqrt{N}\left(\frac{\sum_{i=1}^{N} X_i}{N} - m\right)\right)$

$$\begin{aligned}\phi_n(t) &= E\left(\exp\sum_{i=1}^{N}\frac{it'(X_i-m)}{\sqrt{N}}\right)=E\left(\prod_{i=1}^{i=n}\exp\frac{it'(X_i-m)}{\sqrt{N}}\right)\\ &= \prod_{i=1}^{i=n}E\left(\exp\frac{it'(X_i-m)}{\sqrt{N}}\right)=\left[E\left(\exp\frac{it'(X_i-m)}{\sqrt{N}}\right)\right]^N\end{aligned}$$

$$\begin{aligned}\phi_n(t) &\approx \left[E\left(1+\frac{it'(X_i-m)}{\sqrt{N}}-\frac{1}{2N}\left(t'(X_i-m)\right)\right)\right]^N\\ &= \left[1-\frac{1}{2N}t'\Sigma t\right]^N\to\exp-\frac{t'\Sigma t}{2}\end{aligned}$$

Théorème 4.5 (Méthode delta). *Pour toute fonction g continue, differentiable, si*

$$\sqrt{n}(X_n-m)\xrightarrow{L}\mathcal{N}(0,\Sigma),$$

alors

$$\sqrt{n}(g(X_n)-g(m))\xrightarrow{L}\mathcal{N}\left(0,\left(\frac{\partial g(m)}{\partial m'}\right)\Sigma\left(\frac{\partial g(m)}{\partial m'}\right)'\right).$$

Démonstration.

– On a d'abord $X_n\xrightarrow{P}m$ puisque

$$\Pr(|X_n-m|>\varepsilon)<\frac{E(X_n-m)^2}{\varepsilon^2}=\frac{V(\sqrt{n}(X_n-m))}{n\varepsilon^2}\approx\frac{\Sigma}{n\varepsilon^2}$$

– On applique le théorème de la valeur moyenne : $\exists\,\theta_n\in[0,1]$ tq

$$\begin{aligned}g(X_n) &= g(m)+\frac{\partial g}{\partial m'}(m+\theta_n(X_n-m))(X_n-m).\\ \sqrt{n}(g(X_n)-g(m)) &= \frac{\partial g}{\partial m'}(m+\theta_n(X_n-m))\sqrt{n}(X_n-m)\end{aligned}$$

– $m+\theta_n(X_n-m)\xrightarrow{P}m$ donc $Z_n=\frac{\partial g}{\partial m'}(m+\theta_n(X_n-m))\xrightarrow{P}\frac{\partial g}{\partial m'}(m)$.

– Comme $\sqrt{n}(X_n-m)\xrightarrow{L}\mathcal{N}(0,\Sigma)$, et $Z_n\xrightarrow{P}\frac{\partial g}{\partial m'}(m)$,

$$\sqrt{n}(g(X_n)-g(m))=Z_n\sqrt{n}(X_n-m)\xrightarrow{L}\mathcal{N}\left(0,\left(\frac{\partial g(m)}{\partial m'}\right)\Sigma\left(\frac{\partial g(m)}{\partial m'}\right)'\right).$$

■

4.2 Propriétés asymptotiques de l'estimateur des MCO

On considère le modèle

$$y_i=x_ib+u_i$$

avec les hypothèses

Hypothèse (H_1). $E(u_i|x_i)=0$

Hypothèse (H_2). $V(u_i|x_i) = V(u_i) = \sigma^2$Les observations$(y_i, x_i) \in \mathbb{R}\times\mathbb{R}^{K+1}, i = 1, ..., N,$sont iid

Hypothèse (H_3). $\forall N, X'X$ est non singulière

Hypothèse (H_4). $\mathbb{E}(x_i x_i')$est inversible

Hypothèse (H_5). Les moments de (y_i, x_i) existent au moins jusqu'à l'*ordre 4*.

Théorème 4.6. *Sous les hypothèses H_1 à H_5,*
L'estimateur des MCO

$$\widehat{b}_{mco} = (X'X)^{-1} X'Y = \left(\overline{x_i'x_i}\right)^{-1} \overline{x_i'y_i}$$

1. $\widehat{b} = (X'X)^{-1}X'Y \xrightarrow{P} b$,
2. $\sqrt{N}\left(\widehat{b} - b\right) \xrightarrow{L} \mathcal{N}\left(0, \sigma^2 \left[\mathrm{E}(x_i x_i')\right]^{-1}\right)$,
3. $\widehat{\sigma}^2 = \frac{1}{N-K-1}\left(Y - X\widehat{b}\right)'\left(Y - X\widehat{b}\right) \xrightarrow{P} \sigma^2$,

qd $N \to \infty$.
On dit que $\widehat{b}$ est ***convergent et asymptotiquement normal.***

Démonstration.

1. *Convergence de l'estimateur*
   L'estimateur des mco s'écrit

$$\widehat{b}_{mco} = (X'X)^{-1} X'Y = \overline{x_i'x_i}^{-1}\overline{x_i'y_i}$$

   On remplace y_i par sa valeur : $y_i = x_i b + u_i$. On a donc

$$\widehat{b}_{mco} = \overline{x_i'x_i}^{-1}\overline{x_i'(x_i b + u_i)} = \overline{x_i'x_i}^{-1}\left(\overline{x_i'x_i}b + \overline{x_i'u_i}\right) = b + \overline{x_i'x_i}^{-1}\overline{x_i'u_i}$$

   La loi des grands nombre appliquée à $\overline{x_i'x_i}$ et $\overline{x_i'u_i}$ montre que

$$\overline{x_i'x_i} = \frac{1}{N}\sum_{i=1}^{N} x_i'x_i \xrightarrow{P} \mathrm{E}(x_i x_i'), \text{ et } \overline{x_i'u_i} = \frac{1}{N}\sum_{i=1}^{N} x_i'u_i \xrightarrow{P} \mathrm{E}(x_i'u_i).$$

   Remarque : Importance de l'hypothèse d'existence des moments d'ordre 4.
   On en déduit que

$$\begin{aligned} & \overline{x_i'x_i}^{-1} \xrightarrow{P} \mathrm{E}(x_i x_i')^{-1} \\ & \overline{x_i'x_i}^{-1}\overline{x_i'u_i} \xrightarrow{P} \mathrm{E}(x_i x_i')^{-1}\mathrm{E}(x_i'u_i) \\ \widehat{b}_{mco} \quad = \quad & b + \overline{x_i'x_i}^{-1}\overline{x_i'u_i} \xrightarrow{P} b + \mathrm{E}(x_i x_i')^{-1}\mathrm{E}(x_i'u_i) \end{aligned}$$

   puisque $\mathrm{E}(x_i x_i')$ et $\mathrm{E}(x_i'u_i)$ sont constants, que l'application $A \to A^{-1}$ est continue et que le produit et la somme de suites de va convergent en probabilité vers des constantes converge en probabilité.
   Comme

$$\mathrm{E}(x_i u_i) = \mathrm{E}\left[x_i \mathrm{E}(u_i|x_i)\right] = 0$$

   On a bien

$$\widehat{b} \xrightarrow{P} b$$

2. *Normalité asymptotique*
   De $\widehat{b}_{mco} = b + \overline{x_i' x_i}^{-1} \overline{x_i' u_i}$ on déduit

$$\sqrt{N}\left(\widehat{b}_{mco} - b\right) = \sqrt{N}\overline{x_i' x_i}^{-1}\overline{x_i' u_i} = \overline{x_i' x_i}^{-1}\sqrt{N}\overline{x_i' u_i}$$

   On applique le Théorème Central Limite à $\sqrt{N}\overline{x_i' u_i}$. On sait que

$$\begin{aligned} \mathrm{E}\left(x_i' u_i\right) &= 0 \\ \mathrm{V}\left(x_i' u_i\right) &= \mathrm{V}\left(\mathrm{E}\left(x_i' u_i \left| x_i \right.\right)\right) + \mathrm{E}\left(\mathrm{V}\left(x_i' u_i \left| x_i \right.\right)\right) = \mathrm{E}\left(x_i' \mathrm{V}\left(u_i \left| x_i \right.\right) x_i\right) = \sigma^2 \mathrm{E}\left(x_i' x_i\right) \end{aligned}$$

   Les moments d'ordre 1 et 2 de $x_i' u_i$ existent donc.
   Le TCL permet alors d'affirmer

$$\sqrt{N}\overline{x_i' u_i} \xrightarrow{L} \mathcal{N}\left(0, \sigma^2 \mathrm{E}(x_i x_i')\right)$$

   Comme

$$\overline{x_i' x_i}^{-1} \xrightarrow{P} \mathrm{E}(x_i x_i')^{-1}.$$

   qui est une matrice constante, on peut donc appliquer le théorème de Slutsky à $\overline{x_i' x_i}^{-1}$ et $\sqrt{N}\overline{x_i' u_i}$:

$$\begin{aligned} & \overline{x_i' x_i}^{-1}\sqrt{N}\overline{x_i' u_i} \xrightarrow{L} \mathrm{E}(x_i x_i')^{-1}\mathcal{N}\left(0, \sigma^2 \mathrm{E}(x_i x_i')\right) \\ = \; & \mathcal{N}\left(0, \mathrm{E}(x_i x_i')^{-1}\sigma^2 \mathrm{E}(x_i x_i')\mathrm{E}(x_i x_i')^{-1}\right) \\ = \; & \mathcal{N}\left(0, \sigma^2 \mathrm{E}(x_i x_i')^{-1}\right) \end{aligned}$$

   on a donc bien

$$\sqrt{N}\left(\widehat{b} - b\right) \xrightarrow{L} \mathcal{N}\left(0, \sigma^2 \left[\mathrm{E}(x_i x_i')\right]^{-1}\right)$$

3. *Estimation de la variance*
   L'estimateur de la variance des résidus

$$\widehat{\sigma}^2 = \frac{1}{N-K-1}\left(Y - X\widehat{b}\right)'\left(Y - X\widehat{b}\right)$$

   s'écrit compte tenu de $Y = Xb + U$

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{1}{N-K-1}\overline{\left(X\left(b-\widehat{b}\right)+U\right)'\left(X\left(b-\widehat{b}\right)+U\right)} \\ &= \frac{N}{N-K-1}\overline{\left(x_i\left(b-\widehat{b}\right)+u_i\right)'\left(x_i\left(b-\widehat{b}\right)+u_i\right)} \\ &= \frac{N}{N-K-1}\overline{\left(b-\widehat{b}\right)' x_i' x_i\left(b-\widehat{b}\right) + \left(b-\widehat{b}\right)' x_i u_i + u_i x_i\left(b-\widehat{b}\right) + u_i u_{i'}'} \\ &= \frac{N}{N-K-1}\left[\left(b-\widehat{b}\right)'\overline{x_i' x_i}\left(b-\widehat{b}\right) + 2\left(b-\widehat{b}\right)'\overline{x_i' u_i} + \overline{u_i^2}\right] \xrightarrow{P} \sigma^2 \end{aligned}$$

   puisque $\widehat{b} \xrightarrow{P} b\overline{x_i' x_i} \xrightarrow{P} \mathrm{E}\left(x_i' x_i\right), \overline{x_i' u_i} \xrightarrow{P} \mathrm{E}\left(x_i' u_i\right), \overline{u_i^2} \xrightarrow{P} \mathrm{E}\left(u_i^2\right) = \sigma^2$

■

4.3 Estimation de la variance de l'estimateur

La matrice de variance-covariance asymptotique de l'estimateur "dilaté " $\sqrt{N}\widehat{b}$ est

$$\mathrm{V}_{as}\left(\sqrt{N}\widehat{b}\right) = \sigma^2 \left[\mathrm{E}(x_i' x_i)\right]^{-1}.$$

Cette matrice peut être estimée de façon convergente par

$$\widehat{\mathrm{V}}_{as}\left(\sqrt{N}\widehat{b}\right) = \widehat{\sigma}^2 \left(\overline{x_i' x_i}\right)^{-1} = \widehat{\sigma}^2 \left(\frac{1}{N} X'X\right)^{-1}.$$

La matrice de variance-covariance de $\widehat{b}$ est approxiamtivement

$$\mathrm{V}\left(\widehat{b}\right) \simeq \frac{1}{N}\sigma_0^2 \left[\mathrm{E}(x_i' x_i)\right]^{-1}.$$

Cette matrice peut être estimée de façon convergente par

$$\widehat{\mathrm{V}}\left(\widehat{b}\right) \simeq \frac{1}{N}\widehat{\sigma}^2 \left(\overline{x_i' x_i}\right)^{-1} = \frac{1}{N}\widehat{\sigma}^2 \left(\frac{1}{N} X'X\right)^{-1} = \widehat{\sigma}^2 \left(X'X\right)^{-1}.$$

5 Tests asymptotiques

On définit une *région critique RC* pour une statistique $\widehat{S}$ telle que

$$\widehat{S} \in \mathcal{RC} \Rightarrow \text{ on rejette } H_0 \text{ contre } H_1$$

Définition 5.1. On dit que le test de région critique $\mathcal{RC}$ est *asymptotique* si ses propriétés sont valables pour N grand ; qu'il est de *niveau asymptotique* α si $\lim_{N\to\infty} \Pr\left(\widehat{S} \in \mathcal{RC} \,|H_0\right) = \alpha$; et qu'il est *convergent* si sa *puissance* tend vers un ($\lim_{N\to\infty} \Pr\left(\widehat{S} \in \mathcal{RC} \,|H_a\right) = 1$).

$\Pr\left(\widehat{S} \in \mathcal{RC} \,|H_0\right)$ est le *risque de première espèce* : la probabilité de rejeter H_0 à tort. α est choisi petit : (5% , 1%).

$\Pr\left(\widehat{S} \in \mathcal{RC} \,|H_a\right)$ est le *risque de deuxième espèce* : la probabilité d'accepter H_0 à tort c'est à dire la *puissance* du test.

5.0.1 p-value

La statistique $\widehat{S}$ est choisie de telle sorte que sous H_0 $\widehat{S} \to S_0$

et la loi de S_0 est connue et positive (valeur absolue d'une loi normale, loi du khi deux). La région critique est définit comme

$$RC = \{S \,|S > q\,(1-\alpha, S_0)\,\}$$

où $q\,(1-\alpha, S_0)$ est le quantile d'ordre $1-\alpha$ de S_0.

$$\Pr\left(S_0 > q\,(1-\alpha, S_0)\right) = \alpha$$

Définition 5.2 (p-value). On définit la p-value $p\left(\widehat{S}\right)$ comme $\widehat{S} = q\left(1 - p\left(\widehat{S}\right), S_0\right)$ i.e.

$$p\left(\widehat{S}\right) = \Pr\left\{S_0 > \widehat{S}\right\}.$$

Pour tout seuil α, on rejette H_0 au seuil α ssi $\alpha \geq p\left(\widehat{S}\right)$. En effet, si $\alpha \geq p\left(\widehat{S}\right)$ c'est que

$$\alpha = \Pr\left\{S_0 > q\left(1-\alpha, S_0\right)\right\} \geq \Pr\left\{S_0 > \widehat{S}\right\} \Rightarrow \left\{\widehat{S} > q\left(1-\alpha, S_0\right)\right\}$$

5.1 Test d'hypothèses linéaires

On teste un système de contraintes linéaires. Pour $R \in \mathbb{R}^{p\times(K+1)}$, une matrice dont les lignes sont linéairement indépendantes, et $r \in \mathbb{R}^p$, on teste

$$H_0 : Rb = r \text{ contre } H_a : Rb \neq r.$$

L'estimateur des MCO étant asymptotiquement normal,

$$\sqrt{N}\left(\widehat{b} - b\right) \xrightarrow{L} \mathcal{N}\left(0, V_{\text{as}}\left(\sqrt{N}\widehat{b}\right) = \sigma^2\left[\mathrm{E}(x_i'x_i)\right]^{-1}\right)$$

on a sous H_0

$$\sqrt{N}\left(R\widehat{b} - r\right) \xrightarrow{L} \mathcal{N}\left(0, \mathrm{V}_{\text{as}}\left(\sqrt{N}R\widehat{b}\right) = \sigma^2 R\left[\mathrm{E}(x_i'x_i)\right]^{-1}R'\right)$$

5.1.1 Cas d'une seule contrainte, $p = 1$: test de Student.

On écrit $R = c' \in \mathbb{R}^{K+1}$ et $r \in \mathbb{R}$. Sous l'hypothèse nulle

$$H_0 : c'b = r$$

On a donc

$$\sqrt{N}\left(c'\widehat{b} - r\right) \xrightarrow{L} \mathcal{N}\left(0, c'\mathrm{V}_{\text{as}}\left(\sqrt{N}\widehat{b}\right)c\right)$$

ou encore

$$\sqrt{N}\frac{c'\widehat{b} - r}{\sqrt{c'\mathrm{V}_{\text{as}}\left(\sqrt{N}\widehat{b}\right)c}} \xrightarrow{L} \mathcal{N}(0,1).$$

$\mathrm{V}_{\text{as}}\left(\sqrt{N}\widehat{b}\right) = \sigma^2\left[\mathrm{E}(x_i'x_i)\right]^{-1}$ est inconnue mais on en a un estimateur convergent $\widehat{\mathrm{V}}_{as}\left(\sqrt{N}\widehat{b}\right) = \widehat{\sigma}^2\left(\overline{x_i'x_i}\right)^{-1} = \widehat{\sigma}^2\left(\frac{1}{N}X'X\right)^{-1}$. On applique le théorème de Slutsky. On en déduit que la statistique de Student :

$$\widehat{T} = \sqrt{N}\frac{c'\widehat{b} - r}{\sqrt{c'\widehat{\mathrm{V}}_{as}\left(\sqrt{N}\widehat{b}\right)c}} = \frac{c'\widehat{b} - r}{\sqrt{c'\widehat{\mathrm{V}}\left(\widehat{b}\right)c}} \xrightarrow{L} \mathcal{N}(0,1).$$

Test bilateral. $H_0 : c'b - r = 0$ **contre** $H_1 : c'b - r \neq 0$ On définit la région critique comme

$$W = \left\{T \,\middle|\, |T| > q\left(1 - \frac{\alpha}{2}\right)\right\}$$

où $q\left(1-\frac{\alpha}{2}\right)$ est le quantile $1-\frac{\alpha}{2}$ de la loi normale $\mathcal{N}(0,1)$

Sous H_0 on a

$$\Pr\left\{\widehat{T} \in W \,|H_0\right\} \to \Pr\left\{|\mathcal{N}(0,1)| > q\left(1-\frac{\alpha}{2}\right)\right\} = \alpha$$

Sous H_1 on a $c'\widehat{b} - r \to c'b - r = m \neq 0$ donc

$$\left|\widehat{T}\right| \Big/ \sqrt{N} = \left|\left(c'\widehat{b} - r\right)\right| \Big/ \sqrt{c'\widehat{\mathrm{V}}_{as}\left(\sqrt{N}\widehat{b}\right)c} \to |m| \Big/ \sqrt{c'\mathrm{V}_{as}\left(\sqrt{N}\widehat{b}\right)c}$$

d'où $\left|\widehat{T}\right| \to +\infty \Rightarrow \Pr\left\{\widehat{T} \in W \,|H_1\right\} \to 1$

Test unilatéral $H_0 : c'b - r = 0$ **contre** $H_1 : c'b - r > 0$ On définit la région critique comme

$$W = \{T \,|T > q\,(1-\alpha)\,\}$$

où $q\,(1-\alpha)$ est le quantile $1-\alpha$ de la loi normale $\mathcal{N}(0,1)$

Sous H_0 on a

$$\Pr\left\{\widehat{T} \in W \,|H_0\right\} \to \Pr\left\{\mathcal{N}(0,1) > q\,(1-\alpha)\right\} = \alpha$$

Sous H_1 on a $c'\widehat{b} - r \to c'b - r = m > 0$ donc

$$\widehat{T} \Big/ \sqrt{N} = \left(c'\widehat{b} - r\right) \Big/ \sqrt{c'\widehat{\mathrm{V}}_{as}\left(\sqrt{N}\widehat{b}\right)c} \to m \Big/ \sqrt{c'\mathrm{V}_{as}\left(\sqrt{N}\widehat{b}\right)c}$$

d'où $\left|\widehat{T}\right| \to +\infty \Rightarrow \Pr\left\{\widehat{T} \in W \,|H_1\right\} \to 1$

5.1.2 Cas de plusieurs contraintes, $p \leq K$: test de Wald.

Rappel $Z \sim \mathcal{N}(0,\Sigma)$, Σ inversible $\Longrightarrow Z'\Sigma^{-1}Z \sim \chi^2_K$. D'où

$$N\left(R\widehat{b} - r\right)'\left(R\mathrm{V}_{as}\left(\sqrt{N}\widehat{b}\right)R'\right)^{-1}\left(R\widehat{b} - r\right) \xrightarrow{L} \chi^2_p.$$

On peut remplacer $\mathrm{V}_{as}\left(\sqrt{N}\widehat{b}\right)$ par un estimateur convergent et appliquer Slutsky. D'où, sous l'hypothèse nulle, $H_0 : Rb_0 = r$, et après simplification des N,

$$\begin{aligned}
\widehat{W} &= N\left(R\widehat{b} - r\right)'\left[R\widehat{\mathrm{V}}_{as}\left(\sqrt{N}\widehat{b}\right)R'\right]^{-1}\left(R\widehat{b} - r\right) \\
&= \left(R\widehat{b} - r\right)'\left[R\widehat{\mathrm{V}}\left(\widehat{b}\right)R'\right]^{-1}\left(R\widehat{b} - r\right) \\
&= \frac{\left(R\widehat{b} - r\right)'\left[R\left(X'X\right)^{-1}R'\right]^{-1}\left(R\widehat{b} - r\right)}{\widehat{\sigma}^2} = p\widehat{F} \xrightarrow{L} \chi^2_p, \text{ sous } H_0
\end{aligned}$$

Région critique et p-value On rejettera H_0 au seuil α si la statistique de Wald, $\widehat{W}$, est supérieure au quantile $1-\alpha$ de la loi du χ^2 à p (le nombre de contraintes) degrés de liberté :

$$\widehat{W} > q\left((1-\alpha), \chi^2_p\right)$$

Sous H_0 on a

$$\Pr\left(\widehat{W} > q\left((1-\alpha), \chi^2_p\right)\right) \to \Pr\left(\chi^2_p > q\left((1-\alpha), \chi^2_p\right)\right) = \alpha$$

Sous H_1on a $R\widehat{b} - r \to Rb - r = m \neq 0$
Donc $\widehat{W}/N = \left(R\widehat{b} - r\right)' \left[R\widehat{\mathrm{V}}_{as}\left(\sqrt{N}\widehat{b}\right) R'\right]^{-1} \left(R\widehat{b} - r\right) \to$constante
et donc

$$\widehat{W} \to \infty$$

La p value est définie comme $p = \Pr\left(S_0 > \widehat{W}\right)$

Application : Test de la nullité des paramètres d'une régression sauf la constante. Pour tester la nullité de tous les paramètres d'une régression sauf la constante, on peut former la statistique de Fisher comme

$$\widehat{F} = \frac{\left(SCR_C - SCR\right)/K)}{SCR/(N-K-1)} = \frac{R^2}{1-R^2}\frac{N-K-1}{K}.$$

D'où

$$\widehat{W} = K\widehat{F} = \frac{R^2}{1-R^2}\left(N-K-1\right).$$

Sous H_0 il est facile de voir que $R^2 \xrightarrow{P} 0$ qd $N \to \infty$. On a donc

$$\widehat{W} \simeq NR^2$$

On peut utiliser la statistique NR^2 et rejetter l'hypothèse nulle si

$$NR^2 > q\left((1-\alpha), \chi^2_p\right).$$

5.2 Test d'hypothèses non linéaires

Le principe du test de Wald s'applique au test d'hypothèses non liné aires générales de la forme :

$$H_0 : g(b) = 0,$$

où $g(b)$ est un vecteur de p contraintes non linéaires sur les paramètres telle que

$$\frac{\partial g(b)}{\partial b'} \text{ est de plein rang} \Leftrightarrow \frac{\partial g(b_0)}{\partial b'}\left(\frac{\partial g(b_0)}{\partial b'}\right)' \text{ inversible.}$$

Remarque $g(b) = Rb - r$; alors $\frac{\partial g(b)}{\partial b'} = R$.
En appliquant la méthode delta :

$$\sqrt{N}\left(g(\widehat{b}) - g(b)\right) \xrightarrow{L} \mathcal{N}\left(0, \sigma^2 \frac{\partial g(b)}{\partial b'}\left[\mathrm{E}(x_i x_i')\right]^{-1}\left(\frac{\partial g(b)}{\partial b'}\right)'\right).$$

Cas d'une seule contrainte, $p = 1$. On forme la statistique de student :

$$\widehat{T} = \frac{g(\widehat{b})}{\sqrt{\frac{\partial g(\widehat{b})}{\partial b'}\widehat{\mathrm{V}}\left(\widehat{b}\right)\left(\frac{\partial g(\widehat{b})}{\partial b'}\right)'}}$$

et on procède comme dans le cas d'une contrainte linéaire.

Cas de plusieurs contraintes, $p < K+1$. On calcule la statistique de Wald :

$$\widehat{W} = g(\widehat{b})' \left[\frac{\partial g(\widehat{b})}{\partial b'} \widehat{\mathbb{V}} \left(\widehat{b} \right) \left(\frac{\partial g(\widehat{b})}{\partial b'} \right)' \right]^{-1} g(\widehat{b})$$

que l'on compare au quantile $1-\alpha$ de la loi du χ^2 à p (le nombre de contraintes) degrés de liberté.

6 Le modèle linéaire sans l'hypothèse IID

6.1 Présentation

On considère le cas dans lequel une variable aléatoire y_i dépend de $K+1$ variables explicatives x_i :

$$y_i = x_i b + u_i$$

On maintient l'hypothèse

Hypothèse (H_1). $E(u_i | x_i) = 0$

En revanche, *on ne fait plus l'hypothèse iid* :

Hypothèse (Hypothèse iid).

$$\mathbb{V}\text{ar}(u_i | x_i) = \sigma^2$$
$$\mathbb{C}\text{ov}(u_i, u_j | x_i) = 0$$

6.2 Exemples :

Exemple (Séries temporelles). Erreurs distribuées suivant une moyenne mobile :

$$\begin{aligned} y_t &= x_t b + u_t \\ u_t &= \varepsilon_t + \rho \varepsilon_{t-1} \end{aligned}$$

et $E(\varepsilon_t | X) = 0$, $E(\varepsilon_t \varepsilon_{t'} | X) = 0$ pour $t \neq t'$, $E\left(\varepsilon_t^2 | X\right) = \sigma_\varepsilon^2$
donc

$$\begin{aligned} E\left(u_t^2 | X\right) &= E(\varepsilon_t + \rho\varepsilon_{t-1})^2 = E\left(\varepsilon_t^2 + 2\rho\varepsilon_t\varepsilon_{t-1} + \rho^2\varepsilon_{t-1}^2\right) = \sigma_\varepsilon^2\left(1+\rho^2\right) \\ E\left(u_t u_{t-1} | X\right) &= E(\varepsilon_t + \rho\varepsilon_{t-1})(\varepsilon_{t-1} + \rho\varepsilon_{t-2}) = \sigma_\varepsilon^2 \rho \\ E(u_t u_{t'} | X) &= 0 \quad |t - t'| > 1 \end{aligned}$$

La matrice de variance covariance s'écrit alors pour un échantillon de taille T

$$\begin{aligned} V(U | X) &= \sigma_\varepsilon^2 \begin{pmatrix} \left(1+\rho^2\right) & \rho & 0 & \cdots & 0 \\ \rho & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \rho \\ 0 & \cdots & 0 & \rho & \left(1+\rho^2\right) \end{pmatrix} \\ &\neq \sigma^2 I_T \end{aligned}$$

Exemple (Données de panel). Données à double indice :

$$y_{it}, x_{it} \quad i = 1, \ldots, N,\ t = 1, \ldots, T$$

i indice représentant les individus en général grand,
t indice temporel, en général faible
Le modèle s'écrit comme d'habitude :

$$y_{it} = x_{it} b + u_{it} \quad i = 1, \ldots, N,\ t = 1, \ldots, T$$

ou encore

$$\begin{array}{rcl} \underline{y}_i & = & \underline{x}_i b + \underline{u}_i \quad i = 1, \ldots, N, \\ \underline{z}_i' & = & \left(\begin{array}{ccc} z_{i1} & \cdots & z_{iT} \end{array} \right) \end{array}$$

On fait les hypothèses
$E\left(\underline{u}_i \,|X\right) = 0$
$E\left(\underline{u}_i \underline{u}_j' \,|X\right) = 0 \ \forall\ i \neq j$
En revanche *on ne fait pas l'hypothèse*
$E\left(\underline{u}_i \underline{u}_i' \,|X\right) = \sigma^2 I_T$
Le résidu u_{it} incorpore des éléments inobservés permanent dans le temps.

Exemple (Modèle à erreurs composées).

$$u_{it} = \varepsilon_i + w_{it}$$

avec
$E\left(\underline{w}_i \underline{w}_i' \,|X\right) = \sigma_W^2 I_T$, $E\left(\varepsilon_i \underline{w}_i' \,|X\right) = 0$, $E\left(\varepsilon_i^2 \,|X\right) = \sigma_\varepsilon^2$
On détermine facilement la matrice de variance

$$\Omega = E\left(\underline{u}_i \underline{u}_i' \,|X\right) = \left(\begin{array}{cccc} \sigma_\varepsilon^2 + \sigma_W^2 & \sigma_\varepsilon^2 & \cdots & \sigma_\varepsilon^2 \\ \sigma_\varepsilon^2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma_\varepsilon^2 \\ \sigma_\varepsilon^2 & \cdots & \sigma_\varepsilon^2 & \sigma_\varepsilon^2 + \sigma_W^2 \end{array} \right)$$

ainsi que la matrice de variance covariance des résidus empilés

$$\begin{array}{rcl} E\left(UU' \,|X\right) & = & I_N \otimes \Omega \\ & \neq & \sigma^2 I_{NT} \end{array}$$

Exemple (Régressions empilées). M variables à expliquer, $K_m + 1$ variables explicatives x_{mi} dans l'équation de y_{mi} :

$$y_{mi}, x_{mi} \quad i = 1, \ldots, N,\ m = 1, \ldots, M$$

Le modèle s'écrit pour chaque variable dépendante :

$$y_{mi} = x_{mi} \widetilde{b}_m + u_{mi} \quad i = 1, \ldots, N$$

ou encore

$$\begin{pmatrix} y_{1i} \\ \vdots \\ y_{Mi} \end{pmatrix} = \begin{pmatrix} x_{1i} & 0 & \\ 0 & \ddots & 0 \\ & 0 & x_{Mi} \end{pmatrix} \begin{pmatrix} \widetilde{b}_1 \\ \vdots \\ \widetilde{b}_M \end{pmatrix} + \begin{pmatrix} u_{1i} \\ \vdots \\ u_{Mi} \end{pmatrix}$$
$$\underline{y}_i = \widetilde{X}_i \widetilde{b} + \underline{u}_i \quad i = 1, \ldots, N,$$

où $\widetilde{X}_i$ est la matrice bloc diagonale dont les élé ments de la diagonale sont x_{mi}. Un tel système porte le non de SUR system, SUR signifiant Seemingly Unrelated Regressions. Elle correspnd à la situation dans laquelle il n'y a pas de restrictions entre les coefficients intervenant dans chaque équation. Un cas particulier est donné par le fait que dans chaque équation l'ensemble des variables explicatives soit le même $x_{mi} = x_i$. Dans ce cas la matrice $\widetilde{X}_i$ sécrit simplement $\widetilde{X}_i = I_M \otimes x_i$

Il peut y avoir à l'inverse des spécifications plus contraintes. On peut par exemple introduire des restrictions entre les paramètres des équations : égalité de coefficients entre deux équations, nullité de la somme de coefficients d'une variable intervenant dans chaque équation...Ces restrictions peuvent s'écrire sous la forme $\exists\ b$ et H tel que $\widetilde{b} = Hb$. L'équation générale se réécrit donc :

$$\underline{y}_i = \widetilde{X}_i H b + \underline{u}_i \quad i = 1, \ldots, N,$$
$$\underline{y}_i = X_i b + \underline{u}_i \quad i = 1, \ldots, N,$$

avec $X_i = \widetilde{X}_i H$

On fait les hypothèses

$E(\underline{u}_i | X) = 0$

$E(\underline{u}_i \underline{u}'_j | X) = 0 \ \forall\ i \neq j$

$E(\underline{u}_i \underline{u}'_i | X) = \Sigma$

Les résidus u_{mi} n'ont pas nécessairement la même variance et peuvent en outre être corrélés entre eux. On peut distinguer le cas particulier où $\Sigma = diag(\sigma_1^2, \ldots, \sigma_M^2)$

La matrice de variance covariance des résidus empilés a alors pour expression

$$E(UU' | X) = I_N \otimes \Sigma$$
$$\neq \sigma^2 I_{NT}$$

Exemple (Modèle à coefficent aléatoire). $(dim(x_i) = 1)$

$$y_i = a + x_i b_i + v_i$$
$$b_i = b + v_{bi}$$

avec , $E(v_i | X) = 0$, $E(v_i v_j | X) = 0$ pour $i \neq j$, $E(v_i^2 | X) = \sigma_v^2$,
et $E(v_{bi} | X) = 0$, $E(v_{bi} v_{bj} | X) = 0$ pour $i \neq j$, $E(v_{bi}^2 | X) = \sigma_b^2$,
et $E(v_{bi} v_j | X) = 0 \ \ \forall\ i, j$
Le modèle se réécrit donc

$$y_i = a + x_i b_i + v_i = a + x_i (b + v_{bi}) + v_i$$
$$= a + x_i b + x_i v_{bi} + v_i = a + x_i b + u_i$$
$$u_i = x_i v_{bi} + v_i$$

On a donc les propriétés

$$E\left(u_i\left|X\right.\right) = E\left(x_i v_{bi} + v_i\left|X\right.\right) = x_i E\left(v_{bi}\left|X\right.\right) + E\left(v_i\left|X\right.\right) = 0$$

et

$$E\left(u_i u_j\left|X\right.\right) = 0 \;\forall\; i \neq j$$

$$\begin{aligned} &= E\left(\left(x_i v_{bi} + v_i\right)\left(x_j v_{bj} + v_j\right)\left|X\right.\right) \\ &= x_i x_j E\left(v_{bi} v_{bj}\left|X\right.\right) + x_i E\left(v_{bi} v_j\left|X\right.\right) + x_j E\left(v_i v_{bj}\left|X\right.\right) + E\left(v_i v_j\left|X\right.\right) = 0 \end{aligned}$$

$$\begin{aligned} E\left(u_i^2\left|X\right.\right) &= x_i^2\sigma_b^2 + \sigma_v^2 \\ &= E\left(\left(x_i v_{bi} + v_i\right)^2\left|X\right.\right) = E\left(\left(x_i^2 v_{bi}^2 + 2x_i v_{bi} v_i + v_i^2\right)\left|X\right.\right) \end{aligned}$$

La matrice de variance covariance s'écrit donc

$$\begin{aligned} E\left(UU'\right) &= Diag\left(\sigma_v^2 + x_i^2\sigma_b^2\right) \\ &\neq \sigma^2 I_N \end{aligned}$$

Exemple (Modèle hétéroscédastique en coupe).

$$y_i = a + x_i b + u_i$$

avec ,
$E\left(u_i\left|X\right.\right) = 0$,
$E\left(v_i v_j\left|X\right.\right) = 0$ pour $i \neq j$,
$E\left(v_i^2\left|X\right.\right) = \sigma_i^2$,
La matrice de variance covariance s'écrit donc

$$\begin{aligned} E\left(UU'\left|X\right.\right) &= Diag\left(\sigma_i^2\right) \\ &\neq \sigma^2 I_N \end{aligned}$$

6.3 Conclusion des exemples

Une grande diversité de situations
La matrice de variance des perturbations peut
- dépendre ou non des variables explicatives :
  cas par exemple du modèle à coefficients aléatoires
- dépendre de paramètres additionnel de dimension finie :
  cas par exemple des données de panel, des régressions empilées
- dépendre de paramètres additionnels de dimension infinie :
  cas du modèle hétéroscédastique en coupe

6.4 Le modèle linéaire hétéroscédastique

6.4.1 Définition et hypothèses

On considère le cas dans lequel une variable aléatoire y_i dépend de $K+1$ variables explicatives x_i :

$$y_i = x_i b + u_i$$

soit

$$Y = Xb + U$$

avec les hypothèses

Hypothèse (H_1). $E(U|X) = 0$

Hypothèse (H_2). $E(UU'|X) = \Omega = \Sigma(X, \theta)$ inversible

Hypothèse (H_3). $X'X$ inversible

Le modèle est dit **hétéroscédastique** car on n'a plus l'hypothèse

Hypothèse (Non-H_2). $E(UU'|X) = \sigma^2 I$

Dans un tel cas le modèle aurait été dit *homoscédastique*.
On peut distinguer deux types d'hétéroscédasticité

- hétéroscédasticité due au fait que les données ne sont pas iid : corrélation des perturbations, hétérogéné ité de la variance

$$E(UU'|X) = \Sigma(\theta)$$

  c'est le cas du modèle à moyenne mobile du modèle de données de panel, du modèle de régressions empilées et du modèle hétéroscédastique en coupe.
- hétéroscédasticité due aux variables explicatives

$$E(UU'|X) = \Sigma(X, \theta), \text{ dépend de } X$$

  c'est le cas du modèle à coefficients variables

On se pose les questions suivantes

- Les propriétés statistiques de l'estimateur des MCO sont elles modifiées ?
  - L'estimateur est-il toujours sans biais et convergent ?
  - Quelle est sa matrice de variance et comment l'estimer ?
- L'estimateur des MCO est-il toujours optimal ?
- Comment détecter la présence d'hétéroscédasticité ?

6.5 Estimation par les MCO

Proposition 6.1. *Sous les hypothèses $H1$, $H2$, $H3$, l'estimateur des MCO, $\widehat{b}_{MCO} = (X'X)^{-1}X'Y$, est sans biais :*

$$E\left(\widehat{b}_{MCO}|X\right) = 0,$$

et sa variance sachant X est

$$V\left(\widehat{b}_{MCO}|X\right) = (X'X)^{-1}X'\Omega X(X'X)^{-1}.$$

Démonstration.
On a

$$\begin{aligned}\widehat{b}_{MCO} &= (X'X)^{-1}X'Y = (X'X)^{-1}X'(Xb+U)\\ &= b+(X'X)^{-1}X'U\end{aligned}$$

On a donc pour l'espérance de l'estimation

$$\begin{aligned}E\left(\widehat{b}_{MCO}|X\right) &= b+E\left((X'X)^{-1}X'U|X\right)\\ &= b+(X'X)^{-1}X'E\left(U|X\right)=b\end{aligned}$$

De plus

$$\begin{aligned}V\left(\widehat{b}_{MCO}|X\right) &= V\left((X'X)^{-1}X'U|X\right)\\ &= (X'X)^{-1}X'V\left(U|X\right)X(X'X)^{-1}\\ &= (X'X)^{-1}X'\Omega X(X'X)^{-1}.\end{aligned}$$

■

6.6 La méthode des Moindres Carrés Généralisés (MCG)

Définition 6.1. L'estimateur des MCG est solution du problème :

$$\min_b \left\{SCRG(b) \equiv (Y-Xb)'\,\Omega^{-1}\,(Y-Xb)\right\}$$

Proposition 6.2. *Sous les hypothèses $H1$, $H2$, $H3$, l'estimateur des MCG existe, il est unique et est donné par :*

$$\widehat{b} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y.$$

Démonstration.
Les conditions du premier ordre CN s'écrivent :

$$\frac{\partial SCRG(\widehat{b})}{\partial b} = 2X'\Omega^{-1}\left(Y-X\widehat{b}\right)=0 \Leftrightarrow X'\Omega^{-1}X\widehat{b}=X'\Omega^{-1}Y.$$

La matrice hessienne de l'objectif a pour expression

$$\frac{\partial SCRG(\widehat{b})}{\partial b\partial b'} = -2X'\Omega^{-1}X$$

Sous $H1$, $H2$, $H3$, $X'\Omega^{-1}X$ est inversible symé trique et positive : $\forall\ a\neq 0\in\mathbb{R}^{K+1}$, a, $Xa\neq 0$ sinon $X'X$ non inversible. Comme Ω est inversible on a $(Xa)'\,\Omega^{-1}Xa>0$. D'où

- $\frac{\partial SCRG(\widehat{b})}{\partial b\partial b'}<0$: Les CN sont nécessaires et suffisantes,
- $\widehat{b}_{MCG}=(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$
  car $X'\Omega^{-1}X$ inversible

■

Sphéricisation. Pour toute matrice symétrique et définie positive W il existe une matrice $W^{1/2}$ telle que $W = \left(W^{1/2}\right)^2$. Cette matrice n'est pas unique. On peut clairement la choisir comme symétrique et semi-définie positive (Puisque W est symétrique et semie dé finie positive elle est diagonalisable dans le groupe orthogonal : $W = P^{'}DP$, avec $P'P = I$ et $D = Diag\left(\lambda_k\right)$ la matrice diagonale formée des valeurs propres de W. $D^{1/2} = Diag\left(\sqrt{\lambda_k}\right)$ existe et vérifie $\left(D^{1/2}\right)^2 = D$. On peut définir $W^{1/2} = P^{'}D^{1/2}P$, et on a $W^{1/2}$ symétrique semi définie positive.). D'autres choix sont néanmoins possible et peuvent se révéler int éressant, comme le fait de choisir $W^{1/2}$ triangulaire inférieure ou supérieur. D'une façon ou d'une autre, on définit ainsi $\Omega^{-1/2}$ et $\Omega^{1/2}$ vérifiant

$$\Omega^{-1/2} = \left(\Omega^{1/2}\right)^{-1}$$

d'où

$$\Omega^{-1} = \left[\Omega^{1/2}\Omega^{1/2}\right]^{-1} = \Omega^{-1/2}\Omega^{-1/2}.$$

Si on multiplie le modèle par $\Omega^{-1/2}$ on a :

$$\begin{aligned} \Omega^{-1/2}Y &= \Omega^{-1/2}Xb + \Omega^{-1/2}U \\ \widetilde{Y} &= \widetilde{X}b + \widetilde{U} \end{aligned}$$

Cette transformation des variables Y et X en $\Omega^{-1/2}Y$ et $\Omega^{-1/2}X$ est dite opération de **sphéricisation**. On dit : sphériciser un modèle .On a

$H1 : E\left(\widetilde{U}\left|\widetilde{X}\right.\right) = E\left(\Omega^{-1/2}U\left|\Omega^{-1/2}X\right.\right) = \Omega^{-1/2}E\left(U\left|X\right.\right) = 0$

$H2 : E\left(\widetilde{U}\widetilde{U}'\left|\widetilde{X}\right.\right) = E\left(\Omega^{-1/2}UU'V\left|\Omega^{-1/2}X\right.\right) = \Omega^{-1/2}E\left(UU'\left|X\right.\right)\Omega^{-1/2}$

$= \Omega^{-1/2}\Omega\Omega^{-1/2} = I$

$H3 : \widetilde{X}'\widetilde{X} = X'\Omega^{-1/2\prime}\Omega^{-1/2}X = X'\Omega^{-1}X$ inversible

L'estimateur des MCG est l'estimateur des MCO des coefficients de la régression de $\widetilde{Y}$ sur les colonnes de $\widetilde{X}$:

$$\begin{aligned} \widehat{\widetilde{b}}_{MCO} &= \left(\widetilde{X}'\widetilde{X}\right)^{-1}\widetilde{X}'\widetilde{Y} = \left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1/2\prime}\Omega^{-1/2}Y \\ &= \left(X'\Omega^{-1}X\right)^{-1}X'\Omega^{-1}Y = \widehat{b}_{MCG} \end{aligned}$$

6.7 Propriétés statistiques de l'espérance et de la variance conditionnelle des MCG

Proposition 6.3. *L'estimateur des MCG vérifie les propriétées suivantes*

1. *L'estimateur des MCG est sans biais* : $E\left(\widehat{b}_{MCG}\left|X\right.\right) = b$
2. *L'estimateur des MCG a pour matrice de variance*

$$\mathrm{V}(\widehat{b}_{MCG}|X) = (X'\Omega^{-1}X)^{-1}$$

3. L'estimateur des MCG est le meilleur estimateur linéaire sans biais *(Th. de Gauss Markov)*

Démonstration.

$\widehat{b}_{MCG} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}(Xb+U)$

$$\Rightarrow \widehat{b}_{MCG} = b + (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}U$$

1. *Sans biais :*

$$\begin{aligned} E\left(\widehat{b}_{MCG}\,|X\right) &= b + E\left((X'\Omega^{-1}X)^{-1}X'\Omega^{-1}U\,|X\right) \\ &= b + (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}E\left(U\,|X\right) = b \end{aligned}$$

2. *Variance*

$$\begin{aligned} V\left(\widehat{b}_{MCG}\,|X\right) &= V\left((X'\Omega^{-1}X)^{-1}X'\Omega^{-1}U\,|X\right) \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}V\left(U\,|X\right)\Omega^{-1}X(X'\Omega^{-1}X)^{-1} \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\Omega\Omega^{-1}X(X'\Omega^{-1}X)^{-1} \\ &= (X'\Omega^{-1}X)^{-1} \end{aligned}$$

3. *Optimalité :* Provient directement de $\widehat{b}_{MCG} = \widehat{\widehat{b}}_{MCO}$ et $\widehat{\widehat{b}}_{MCO}$ optimal

■

7 L'estimateur des MCQG

La matrice Ω est inconnue. L'estimateur des MCG et la matrice de variance des MCO ne sont pas calculables. Il faut donc estimer cette matrice. Soit $\widehat{\Omega}$ un estimateur de Ω. On appelle estimateur des *Moindres Carrés Quasi-Généralisés* l'estimateur :

$$\widehat{b}_{MCQG} = (X'\widehat{\Omega}^{-1}X)^{-1}X'\widehat{\Omega}^{-1}Y.$$

L'estimateur des MCQG n'est en général pas sans biais ni linéaire en Y puisque $\widehat{\Omega}$ dépend de Y.

Les propriétés de $\widehat{b}_{MCQG}$ ne peuvent donc être *qu'asymptotiques.*

7.0.1 Cas où $\Omega = \Sigma(\theta)$ et θ de dimension finie

On considère le modèle

$$\underline{y}_i = \underline{x}_i b + \underline{u}_i,\ \underline{y}_i \text{ de dim } M \times 1,\ \underline{x}_i \text{ de dim } M \times K+1$$

avec les hypothèses

Hypothèse (H_1). $E(\underline{u}_i \,|\underline{x}_i) = 0$

Hypothèse (H_2). $V(\underline{u}_i \,|\underline{x}_i) = V(\underline{u}_i) = \Sigma(\theta)$ Σ de dim $M \times M$, θ est alors nécessairement de dimension finie

Hypothèse (H_3). Les observations $(\underline{y}_i, \underline{x}_i) \in \mathbb{R} \times \mathbb{R}^{K+1}$, $i = 1, ..., N$, sont iid

Hypothèse (H_4 et H_5). $\forall\ N$ $X'X$ et $\mathrm{E}(\underline{x}_i\underline{x}_i')$ sont inversibles

Hypothèse (H_6). Les moments de $(\underline{y}_i, \underline{x}_i)$ existent au moins jusqu'à l'*ordre 4.*

Théorème 7.1. *Sous les hypothèses H1 à H6, l'estimateur des MCO*

$$\widehat{b}_{mco} = (X'X)^{-1}X'Y = \left(\overline{\underline{x}_i'\underline{x}_i}\right)^{-1}\overline{\underline{x}_i'\underline{y}_i}$$

verifie quand $N \to \infty$

1. $\widehat{b}_{mco} \xrightarrow{P} b$, *convergence ;*
2. $\sqrt{N}\left(\widehat{b}_{mco} - b\right) \xrightarrow{L} \mathcal{N}\left(0, \mathrm{V}_{as}\left(\widehat{b}_{mco}\right)\right)$, *Normalité asymptotique ;*
3. $\mathrm{V}_{as}\left(\widehat{b}_{mco}\right) = [\mathrm{E}(\underline{x}_i'\underline{x}_i)]^{-1}\,\mathrm{E}(\underline{x}_i'\Sigma\underline{x}_i)\,[\mathrm{E}(\underline{x}_i'\underline{x}_i)]^{-1}$
4. $\widehat{\Sigma} = \overline{\left(\underline{y}_i - \underline{x}_i\widehat{b}_{mco}\right)\left(\underline{y}_i - \underline{x}_i\widehat{b}_{mco}\right)'} = \overline{\widehat{\underline{u}}_i\widehat{\underline{u}}_i'} \xrightarrow{P} \Sigma$, *Estimation de* Σ
5. $\widehat{\mathrm{V}}_{as}\left(\widehat{b}_{mco}\right) = \overline{(\underline{x}_i'\underline{x}_i)}^{-1}\overline{\underline{x}_i'\widehat{\Sigma}\underline{x}_i}\,\overline{\underline{x}_i'\underline{x}_i}^{-1} \xrightarrow{P} \mathrm{V}_{as}\left(\widehat{b}_{mco}\right)$ *Estimation de* V
6. $\sqrt{N}\widehat{\mathrm{V}}_{as}\left(\widehat{b}_{mco}\right)^{-1/2}\left(\widehat{b}_{mco} - b\right) \xrightarrow{L} \mathcal{N}(0, I)$

Démonstration.
Si M est la longueur du vecteur $\underline{y}_i$: $\underline{y}_i' = \begin{pmatrix} y_{1i} & \cdots & y_{Mi} \end{pmatrix}$
$X'X = \sum_{i=1,m=1}^{i=N,m=M} x_{im}'x_{im} = \sum_{i=1}^{i=N}\sum_{m=1}^{m=M} x_{im}'x_{im} = \sum_{i=1}^{i=N} \underline{x}_i'\underline{x}_i$
d'où l'expression de $\widehat{b}_{mco}$

1. *Convergence* On a $\widehat{b}_{mco} = b + \left(\overline{\underline{x}_i'\underline{x}_i}\right)^{-1} \overline{\underline{x}_i'\underline{u}_i}$
   Comme les observations sont indépendantes entre deux individus i et j et que les moments d'ordre 4 existent, d'où l'existence de moments d'ordre 2 pour $\underline{x}_i'\underline{x}_i$et $\underline{x}_i'\underline{u}_i$ en appliquant la loi des grands nombre $\left(\overline{\underline{x}_i'\underline{x}_i}\right)^{-1} \overline{\underline{x}_i'\underline{u}_i} \xrightarrow{P} E\left(\underline{x}_i'\underline{x}_i\right)^{-1} E\left(\underline{x}_i'\underline{u}_i\right)$ et $E\left(\underline{x}_i'\underline{u}_i\right) = E\left(\underline{x}_i' E\left(\underline{u}_i \left|\underline{x}_i\right.\right)\right) = 0$
2. *Normalité asymptotique* $\sqrt{N}\left(\widehat{b}_{mco} - b\right) = \left(\overline{\underline{x}_i'\underline{x}_i}\right)^{-1} \sqrt{N\underline{x}_i'\underline{u}_i}$
   Théorème central limite appliqué à $\underline{x}_i'\underline{u}_i$ $E\left(\underline{x}_i'\underline{u}_i\right) = 0$ et $V\left(\underline{x}_i'\underline{u}_i\right) = E\left(V\left(\underline{x}_i'\underline{u}_i \left|x_i\right.\right)\right) = E\left(\underline{x}_i' V\left(\underline{u}_i \left|x_i\right.\right) \underline{x}_i\right) = E\left(\underline{x}_i'\Sigma\underline{x}_i\right)$ existent. On a donc $\sqrt{N\underline{x}_i'\underline{u}_i} \xrightarrow{L} N\left(0, E\left(\underline{x}_i'\Sigma\underline{x}_i\right)\right)$
   On applique le théorème de Slutsky $\left(\overline{\underline{x}_i'\underline{x}_i}\right)^{-1} \xrightarrow{P} E\left(\underline{x}_i'\underline{x}_i\right)^{-1}$ et $\sqrt{N\underline{x}_i'\underline{u}_i} \xrightarrow{L} N\left(0, E\left(\underline{x}_i'\Sigma\underline{x}_i\right)\right)$
   donc

$$\begin{aligned}\sqrt{N}\left(\widehat{b}_{mco} - b\right) &= \left(\overline{\underline{x}_i'\underline{x}_i}\right)^{-1} \sqrt{N\underline{x}_i'\underline{u}_i} \\ &\xrightarrow{L} N\left(0, E\left(\underline{x}_i'\underline{x}_i\right)^{-1} E\left(\underline{x}_i'\Sigma\underline{x}_i\right) E\left(\underline{x}_i'\underline{x}_i\right)^{-1}\right)\end{aligned}$$

3. *Estimation de* Σ
   $\widehat{\Sigma} = \overline{\left(\underline{y}_i - \underline{x}_i\widehat{b}_{mco}\right)\left(\underline{y}_i - \underline{x}_i\widehat{b}_{mco}\right)'} = \overline{\widehat{\underline{u}}_i\widehat{\underline{u}}_i'}$
   $\widehat{\underline{u}}_i = \underline{y}_i - \underline{x}_i\widehat{b}_{mco} = \underline{x}_i\left(b - \widehat{b}_{mco}\right) + \underline{u}_i$

$$\begin{aligned}\widehat{\Sigma} &= \overline{\left(\underline{x}_i\left(b - \widehat{b}_{mco}\right) + \underline{u}_i\right)\left(\underline{x}_i\left(b - \widehat{b}_{mco}\right) + \underline{u}_i\right)'} \\ &= \overline{\underline{u}_i\underline{u}_i'} + \overline{\underline{x}_i\left(b - \widehat{b}_{mco}\right)\left(b - \widehat{b}_{mco}\right)'\underline{x}_i'} + \\ &\quad \overline{\underline{x}_i\left(b - \widehat{b}_{mco}\right)\underline{u}_i'} + \overline{\underline{u}_i\left(b - \widehat{b}_{mco}\right)'\underline{x}_i'}\end{aligned}$$

   Le premier terme converge vers Σ par la loi des grands nombres.
   Le deuxième terme est une matrice dont les éléments sont somme de termes $x_{li}^k\left(b - \widehat{b}_{mco}\right)_m \left(b - \widehat{b}_{mco}\right)_{m'} x_{l'i}^{k'} = \left(b - \widehat{b}_{mco}\right)_m \left(b - \widehat{b}_{mco}\right)_{m'} \overline{x_{li}^k x_{l'i}^{k'}}$ comme $\left(b - \widehat{b}_{mco}\right) \xrightarrow{P} 0$ et que $\overline{x_{li}^k x_{l'i}^{k'}} \xrightarrow{P} E\left(x_{li}^k x_{l'i}^{k'}\right)$ le deuxième terme tend vers zero en probabilité. De même pour le troisième et le quatrième terme.
4. *Estimation de la variance de l'estimateur des mco*

$$\widehat{\mathrm{V}}\left(\widehat{b}_{mco}\right) = \overline{\left(\underline{x}_i'\underline{x}_i\right)}^{-1} \overline{\underline{x}_i'\widehat{\Sigma}\underline{x}_i} \overline{\underline{x}_i'\underline{x}_i}^{-1} \xrightarrow{P} \mathrm{V}\left(\widehat{b}_{mco}\right)$$

   Le seul terme important est $\overline{\underline{x}_i'\widehat{\Sigma}\underline{x}_i}$ et on a

$$\begin{aligned}\overline{\underline{x}_i'\widehat{\Sigma}\underline{x}_i} - E\left(\underline{x}_i'\Sigma\underline{x}_i\right) &= \left(\overline{\underline{x}_i'\widehat{\Sigma}\underline{x}_i} - \overline{\underline{x}_i'\Sigma\underline{x}_i}\right) + \left(\overline{\underline{x}_i'\Sigma\underline{x}_i} - E\left(\underline{x}_i'\Sigma\underline{x}_i\right)\right) \\ &= \left(\overline{\underline{x}_i'\left(\widehat{\Sigma} - \Sigma\right)\underline{x}_i}\right) + \left(\overline{\underline{x}_i'\Sigma\underline{x}_i} - E\left(\underline{x}_i'\Sigma\underline{x}_i\right)\right)\end{aligned}$$

Le deuxième terme tend vers zero en probabilité par la loi forte des grands nombres.

Le premier terme tend vers zero en probabilité par le même genre d'argument que précédemment, puisque $\widehat{\Sigma} \xrightarrow{P} \Sigma$

Enfin, comme

$$\widehat{\mathrm{V}}\left(\widehat{b}_{mco}\right) \xrightarrow{P} \mathrm{V}\left(\widehat{b}_{mco}\right) \text{ et } \sqrt{N}\left(\widehat{b}_{mco} - b\right) \xrightarrow{L} \mathcal{N}\left(0, \mathrm{V}\left(\widehat{b}_{mco}\right)\right)$$

on a directement par le théorème de Slutsky

$\sqrt{N}\widehat{\mathrm{V}}\left(\widehat{b}_{mco}\right)^{-1/2}\left(\widehat{b}_{mco} - b\right) \xrightarrow{L} \mathcal{N}(0, I)$

■

Hypothèse (H_7). $\exists\ \widehat{\theta} \xrightarrow{P} \theta$, l'estimateur des MCQG

Théorème 7.2. *Sous les hypothèses H1 à H7, et si*

$$\widehat{b}_{mcqg} = \left(\overline{\underline{x}_i'\Sigma\left(\widehat{\theta}\right)^{-1}\underline{x}_i}\right)^{-1}\overline{\underline{x}_i'\Sigma\left(\widehat{\theta}\right)^{-1}\underline{y}_i}$$

vérifie quand $N \to \infty$

1. $\widehat{b}_{mcqg} \xrightarrow{P} b$, *Convergence ;*
2. $\sqrt{N}\left(\widehat{b}_{mcqg} - b\right) \xrightarrow{L} \mathcal{N}\left(0, \mathrm{V}_{as}\left(\widehat{b}_{mcqg}\right)\right)$, *Normalité asymptotique ;*
3. $\mathrm{V}_{as}\left(\widehat{b}_{mcqg}\right) = \left[\mathrm{E}(\underline{x}_i'\Sigma^{-1}\underline{x}_i)\right]^{-1} = \underline{\mathrm{V}\left(\widehat{b}_{mcg}\right)}$ *Equivalence asymptotique entre MCQG et MCG*
4. $\widehat{\mathrm{V}}_{as}\left(\widehat{b}_{mcqg}\right) = \overline{\underline{x}_i'\widehat{\Sigma}^{-1}\underline{x}_i}^{-1} \xrightarrow{P} \mathrm{V}\left(\widehat{b}_{mcg}\right)$ *Estimation de la variance ;*
5. $\sqrt{N}\widehat{\mathrm{V}}_{as}\left(\widehat{b}_{mcqg}\right)^{-1/2}\left(\widehat{b}_{mcqg} - b\right) \xrightarrow{L} \mathcal{N}(0, I)$.

Démonstration. Soit $\widehat{\Sigma} = \Sigma\left(\widehat{\theta}\right)$. Comme $\widehat{\theta} \xrightarrow{P} \theta$, $\widehat{\Sigma} \xrightarrow{P} \Sigma$

1. *Convergence* $\widehat{b}_{mcqg} = b + \left(\overline{\underline{x}_i'\widehat{\Sigma}^{-1}\underline{x}_i}\right)^{-1}\overline{\underline{x}_i'\widehat{\Sigma}^{-1}\underline{u}_i}$

   Chaque terme de $\overline{\underline{x}_i'\widehat{\Sigma}^{-1}\underline{x}_i}$ est somme de termes de la forme $\overline{x_{li}^k\widehat{\Sigma}^{-1}_{m,m'}x_{l'i}^{k'}} = \widehat{\Sigma}^{-1}_{m,m'}\overline{x_{li}^k x_{l'i}^{k'}}$ converge vers $\widehat{\Sigma}^{-1}_{m,m'}\overline{x_{li}^k x_{l'i}^{k'}} \xrightarrow{P} \Sigma^{-1}_{m,m'}E\left(x_{li}^k x_{l'i}^{k'}\right)$ et est le terme correspondant de $E\left(\underline{x}_i'\Sigma^{-1}\underline{x}_i\right)$. On a donc

   $$\overline{\underline{x}_i'\widehat{\Sigma}^{-1}\underline{x}_i} \xrightarrow{P} E\left(\underline{x}_i'\Sigma^{-1}\underline{x}_i\right)$$

   De même

   $$\overline{\underline{x}_i'\widehat{\Sigma}^{-1}\underline{u}_i} \xrightarrow{P} E\left(\underline{x}_i'\Sigma^{-1}\underline{u}_i\right) = E\left(\underline{x}_i'\Sigma^{-1}E\left(\underline{u}_i\,|\underline{x}_i\right)\right) = 0$$

   D'où la convergence de l'estimateur

2. *Normalité asymptotique*
   Le seul point à montrer est $\sqrt{N}\overline{\underline{x}_i'\widehat{\Sigma}^{-1}\underline{u}_i} \xrightarrow{L} N\left(0, E\left(\underline{x}_i'\Sigma^{-1}\underline{x}_i\right)\right)$

$$\sqrt{N}\overline{\underline{x}_i'\widehat{\Sigma}^{-1}\underline{u}_i} = \sqrt{N}\overline{\underline{x}_i'\left(\widehat{\Sigma}^{-1}-\Sigma^{-1}\right)\underline{u}_i} + \sqrt{N}\overline{\underline{x}_i'\Sigma^{-1}\underline{u}_i}$$

   Chaque terme de $\sqrt{N}\overline{\underline{x}_i'\left(\widehat{\Sigma}^{-1}-\Sigma^{-1}\right)\underline{u}_i}$ est de la forme

$$\sqrt{N}\overline{x_{li}^k\left(\widehat{\Sigma}_{m,m'}^{-1}-\Sigma_{m,m'}^{-1}\right)u_{l'i}} = \left(\widehat{\Sigma}_{m,m'}^{-1}-\Sigma_{m,m'}^{-1}\right)\sqrt{N}\overline{x_{li}^k u_{l'i}}$$

   Le premier terme converge en probabilité vers 0. Le deuxième terme converge en loi vers une loi normale.
   Elle est donc bornée en probabilité :
   X_N bornée en probabilité si $\forall\delta > 0\ \exists M_\delta$ *et* N_δ *tq* $N > N_\delta \Rightarrow P\left(|X_N| > M_\delta\right) < \delta$
   On peut montrer que le produit d'une suite convergeant en probabilité vers 0 et une suite bornée en probabilité converge en probabilité vers 0. Le comportement asymptotique de $\sqrt{N}\overline{\underline{x}_i'\widehat{\Sigma}^{-1}\underline{u}_i}$ est donc le même que celui de $\sqrt{N}\overline{\underline{x}_i'\Sigma^{-1}\underline{u}_i}$. Comme $V\left(\underline{x}_i'\Sigma^{-1}\underline{u}_i\right) = E\left(\underline{x}_i'\Sigma^{-1}\underline{x}_i\right)$, il converge donc en loi vers une loi normale $N\left(0, E\left(\underline{x}_i'\Sigma^{-1}\underline{x}_i\right)\right)$
3. Les deux derniers points se démontrent de la même façon que pr écédemment.

■

7.0.2 Application

Données de panel et Régressions empilées
- On estime le modèle

$$\underline{y}_i = \underline{x}_i b + \underline{u}_i$$

par les MCO : $\widehat{b}_{MCO} = (X'X)^{-1}(X'Y)$
- On calcule le résidu pour chaque individu

$$\widehat{\underline{u}}_i = \underline{y}_i - \underline{x}_i\widehat{b}_{MCO}$$

- On calcule un estimateur de la matrice de variance des résidus

$$\widehat{\Sigma} = \overline{\widehat{\underline{u}}_i\widehat{\underline{u}}_i'}$$

- On peut alors déterminer la variance asymptotique et la variance de l'estimateur des MCO par

$$\begin{aligned}\widehat{\mathrm{V}}_{as}\left(\widehat{b}_{mco}\right) &= \overline{\left(\underline{x}_i'\underline{x}_i\right)}^{-1}\overline{\underline{x}_i'\widehat{\Sigma}\underline{x}_i}\,\overline{\underline{x}_i'\underline{x}_i}^{-1}\\ \widehat{\mathrm{V}}\left(\widehat{b}_{mco}\right) &= \frac{1}{N}\widehat{\mathrm{V}}_{as}\left(\widehat{b}_{mco}\right)\end{aligned}$$

- On calcule l'estimateur des MCQG

$$\widehat{b}_{mcqg} = \left(\overline{\underline{x}_i'\widehat{\Sigma}^{-1}\underline{x}_i}\right)^{-1}\overline{\underline{x}_i'\widehat{\Sigma}^{-1}\underline{y}_i}$$

– On calcule la variance asymptotique et la variance de l'estimateur des MCQG

$$\begin{aligned}\widehat{\mathrm{V}}_{as}\left(\widehat{b}_{mcqg}\right) &= \overline{\underline{x}_i'\widehat{\Sigma}^{-1}\underline{x}_i}^{-1} \\ \widehat{\mathrm{V}}\left(\widehat{b}_{mcqg}\right) &= \frac{1}{N}\widehat{\mathrm{V}}_{as}\left(\widehat{b}_{mcqg}\right)\end{aligned}$$

7.0.3 Retour sur les régressions SUR

On considère la situation dans laquelle l'ensemble des régresseurs intervenant dans chaque équation est le même, lorsqu'il n'y a pas de contrainte entre les paramètres d'une équation à l'autre. Dans une telle situation on a,

Théorème 7.3 (Théorème de Zellner). *L'estimateur des* MCG *est équivalent à l'estimateur des* MCO *effectué équation par équation.*

Démonstration.
Un tel modèle s'écrit sous la forme :

$$\underline{y}_i = (I_M \otimes x_i)\, b + \underline{u}_i$$

et $b' = \left(b_1', \ldots, b_M'\right)$ est de dimension $M(K+1)$. Dans ce cas l'estimateur des MCG est donné par

$$\widehat{b}_{MCG} = \overline{(I_M \otimes x_i)'\,\Sigma^{-1}\,(I_M \otimes x_i)}^{-1}\overline{(I_M \otimes x_i)'\,\Sigma^{-1}\underline{y}_i}$$

Rappel sur les produits de Kronecker de matrices : si AC et BD existent, c'est à dire si leurs dimensions sont conformes aux produits matriciels, on a

$$(A \otimes B)(C \otimes D) = (AC \otimes BD)$$

On rappelle aussi que $(A \otimes B)' = (A' \otimes B')$.

Dans ces conditions, puisque $\Sigma^{-1} = \Sigma^{-1} \otimes 1$ et que x_i est de dimension $1 \times (K+1)$ on a $(I_M \otimes x_i)'\,\Sigma^{-1} = \left(I_M \otimes x_i'\right)\left(\Sigma^{-1} \otimes 1\right) = \left(\Sigma^{-1} \otimes x_i'\right)$. On a de même $(I_M \otimes x_i)'\,\Sigma^{-1}\,(I_M \otimes x_i) = \left(\Sigma^{-1} \otimes x_i'\right)(I_M \otimes x_i) = \left(\Sigma^{-1} \otimes x_i'x_i\right)$ et $(I_M \otimes x_i)'\,\Sigma^{-1}\underline{y}_i = \left(I_M \otimes x_i'\right)\left(\Sigma^{-1}\underline{y}_i \otimes 1\right) = \left(\Sigma^{-1}\underline{y}_i \otimes x_i'\right)$. On a donc

$$\begin{aligned}\widehat{b}_{MCG} &= \overline{\Sigma^{-1} \otimes x_i'x_i}^{-1}\overline{\left(\Sigma^{-1}\underline{y}_i \otimes x_i'\right)} = \Sigma \otimes \overline{x_i'x_i}^{-1}\overline{\left(\Sigma^{-1}\underline{y}_i \otimes x_i'\right)} \\ &= \overline{\Sigma \otimes \overline{x_i'x_i}^{-1}\left(\Sigma^{-1}\underline{y}_i \otimes x_i'\right)} = \overline{\left(\underline{y}_i \otimes \left(\overline{x_i'x_i}^{-1}x_i'\right)\right)}\end{aligned}$$

Comme

$$\underline{y}_i \otimes \left(\overline{x_i'x_i}^{-1}x_i'\right) = \begin{bmatrix} y_{1i}\left(\overline{x_i'x_i}^{-1}x_i'\right) \\ \vdots \\ y_{Mi}\left(\overline{x_i'x_i}^{-1}x_i'\right) \end{bmatrix} = \begin{bmatrix} \overline{x_i'x_i}^{-1}x_i'y_{1i} \\ \vdots \\ \overline{x_i'x_i}^{-1}x_i'y_{Mi} \end{bmatrix}$$

on voit que

$$\widehat{b}_{MCG} = \overline{\underline{y}_i \otimes \left(\overline{x_i' x_i}^{-1} x_i'\right)} = \begin{bmatrix} \overline{x_i' x_i}^{-1} \overline{x_i' y_{1i}} \\ \vdots \\ \overline{x_i' x_i}^{-1} \overline{x_i' y_{Mi}} \end{bmatrix} = \begin{bmatrix} \widehat{b}_{MCO}(1) \\ \vdots \\ \widehat{b}_{MCO}(M) \end{bmatrix}$$

où $\widehat{b}_{MCO}(m) = \overline{x_i' x_i}^{-1} \overline{x_i' y_{mi}}$ est l'estimateur des mco sur l' équation m prise individuellement.

Remarquons toutefois que la variance de l'estimateur s'écrit

$$\begin{aligned} V_{as}\left(\widehat{b}_{MCG}\right) &= E\left[\overline{\underline{u}_i \otimes \left(\overline{x_i' x_i}^{-1} x_i'\right)} \underline{u}_i' \otimes \left(\overline{x_i' x_i}^{-1} x_i'\right)'\right] \\ &= \Sigma \otimes \left(E\left(x_i' x_i\right)^{-1}\right) \end{aligned}$$

Bien que pouvant être calculés simplement équation par é quation, les estimateurs pour chaque équations sont corrélés entre eux. ■

7.0.4 Cas où $\Omega = \Sigma(\theta, X)$ et θ de dimension finie

On considère le modèle

$$\underline{y}_i = \underline{x}_i b + \underline{u}_i$$

avec les hypothèses

Hypothèse (H_1). $E(\underline{u}_i | \underline{x}_i) = 0$

Hypothèse (H_2). $V(\underline{u}_i | \underline{x}_i) = V(\underline{u}_i) = \Sigma(\theta, \underline{x}_i)$ régulière : C_∞

Hypothèse (H_3). Les observations $(\underline{y}_i, \underline{x}_i) \in \mathbb{R} \times \mathbb{R}^{K+1}$, $i = 1, ..., N$, sont iid

Hypothèse (H_4 et H_5). $\forall\ N$ $X'X$ et $\mathrm{E}(\underline{x}_i \underline{x}_i')$ sont inversibles

Hypothèse (H_6). Les moments de $(\underline{y}_i, \underline{x}_i)$ existent au moins jusqu'à n'importe quel ordre

Hypothèse (H_7). $\exists\ \widehat{\theta} \xrightarrow{P} \theta$

Théorème 7.4. *Sous les hypothèses H1 à H7, l'estimateur des MCQG*

$$\begin{aligned} \widehat{b}_{mcqg} &= \left(X' I_N \otimes \Sigma\left(\widehat{\theta}, X\right)^{-1} X\right)^{-1} X' I_N \otimes \Sigma\left(\widehat{\theta}, X\right)^{-1} Y \\ &= \left(\overline{\underline{x}_i' \Sigma\left(\widehat{\theta}, X\right)^{-1} \underline{x}_i}\right)^{-1} \overline{\underline{x}_i' \Sigma\left(\widehat{\theta}, X\right)^{-1} \underline{y}_i} \end{aligned}$$

vérifie quand $N \to \infty$

1. $\widehat{b}_{mcqg} \xrightarrow{P} b$, *Convergence*
2. $\sqrt{N}\left(\widehat{b}_{mcqg} - b\right) \xrightarrow{L} \mathcal{N}\left(0, \mathrm{V}_{as}\left(\widehat{b}_{mcqg}\right)\right)$, *Normalité asymptotique*
3. $\mathrm{V}_{as}\left(\widehat{b}_{mcqg}\right) = \left[\mathrm{E}(\underline{x}_i' \Sigma(\underline{x}_i, \theta)^{-1} \underline{x}_i)\right]^{-1} = \underline{\mathrm{V}\left(\widehat{b}_{mcg}\right)}$ *Equivalence MCQG et MCG*

4. $\widehat{\mathrm{V}}_{as}\left(\widehat{b}_{mcqg}\right) = \overline{\underline{x}_i'\Sigma\left(\underline{x}_i,\widehat{\theta}\right)^{-1}\underline{x}_i}^{-1} \xrightarrow{P} \mathrm{V}_{as}\left(\widehat{b}_{mcqg}\right)$ *Estimation de V*
5. $\sqrt{N}\widehat{\mathrm{V}}_{as}\left(\widehat{b}_{mcqg}\right)^{-1/2}\left(\widehat{b}_{mcqg}-b\right) \xrightarrow{L} \mathcal{N}(0,I)$

Démonstration. Soit $\widehat{\Sigma}_i = \Sigma\left(\widehat{\theta},\underline{x}_i\right)$.

1. *Convergence* $\widehat{b}_{mcqg} = b + \left(\overline{\underline{x}_i'\widehat{\Sigma}_i^{-1}\underline{x}_i}\right)^{-1}\overline{\underline{x}_i'\widehat{\Sigma}_i^{-1}\underline{u}_i}$

   $\overline{\underline{x}_i'\widehat{\Sigma}_i^{-1}\underline{z}_i} = \overline{\underline{x}_i'\Sigma\left(x_i,\theta\right)_i\underline{z}_i} + \overline{\underline{x}_i'\left(\Sigma\left(\underline{x}_i,\widehat{\theta}\right)-\Sigma\left(\underline{x}_i,\theta\right)\right)\underline{z}_i}$ comme $\theta \xrightarrow{P} \theta$

   $$\overline{\underline{x}_i'\widehat{\Sigma_i}^{-1}\underline{z}_i} \xrightarrow{P} E\left(\underline{x}_i'\Sigma\left(\underline{x}_i,\theta\right)^{-1}\underline{z}_i\right)$$

   D'où la convergence de l'estimateur puisque $E\left(\underline{x}_i'\Sigma\left(\underline{x}_i,\theta\right)^{-1}\underline{u}_i\right) = 0$
2. *Normalité asymptotique*

   Le seul point à montrer est $\sqrt{N}\overline{\underline{x}_i'\widehat{\Sigma}_i^{-1}\underline{u}_i} \xrightarrow{L} N\left(0, E\left(\underline{x}_i\Sigma\left(\underline{x}_i,\theta\right)^{-1}\underline{x}_i\right)\right)$

   $$\sqrt{N}\overline{\underline{x}_i'\widehat{\Sigma}^{-1}\underline{u}_i} = \sqrt{N}\overline{\underline{x}_i'\left(\widehat{\Sigma}_i^{-1}-\Sigma\left(\underline{x}_i,\theta\right)^{-1}\right)\underline{u}_i} + \sqrt{N}\overline{\underline{x}_i'\Sigma\left(\underline{x}_i,\theta\right)^{-1}\underline{u}_i}$$

   $\widehat{\Sigma}_{m,m'}^{-1} - \Sigma_{m,m'}^{-1} = \partial\Sigma_{m,m'}/\partial\theta\left(\widetilde{\theta},\underline{x}_i\right)\left(\widehat{\theta}-\theta\right)$, avec $\left|\widetilde{\theta}-\theta\right| < \left|\widehat{\theta}-\theta\right|$

   Chaque terme de $\sqrt{N}\overline{\underline{x}_i'\left(\widehat{\Sigma}^{-1}-\Sigma^{-1}\right)\underline{u}_i}$ est somme de termes de la forme $\sqrt{N}\overline{x_{li}^k\left(\widehat{\Sigma}_{m,m'}^{-1}-\Sigma_{m,m'}^{-1}\right)u_{l'i}} = \sqrt{N}\overline{x_{li}^k u_{l'i}\partial\Sigma_{m,m'}/\partial\theta\left(\widetilde{\theta},\underline{x}_i\right)}\left(\widehat{\theta}-\theta\right)$ Le deuxième terme converge en probabilité vers 0. Le premier terme converge en loi vers une loi normale si $x_{li}^k u_{l'i}\partial\Sigma_{m,m'}/\partial\theta\left(\widetilde{\theta},\underline{x}_i\right)$ a des moments d'ordre 1 et 2. Elle est donc bornée en probabilité et on procède comme précédemment.
3. Les deux derniers points se démontrent de la même façon que pr écédemment.

■

7.0.5 Application :

Modèle en coupe

$$y_i = x_i b + u_i$$

dans lequel on spécifie la forme de l'hétérogénéité (p.e. modèle à coefficient aléatoire). On suppose qu'il existe des variables z_i formées à partir de x_i telles que

$$\begin{aligned}\sigma_i^2 &= \exp z_i\theta \\ \log\left(\sigma_i^2\right) &= z_i\theta\end{aligned}$$

On procède de la façon suivante :

1. Calcul de $\widehat{b}_{MCO}$ et des résidus : $\widehat{u}_i = y_i - x_i\widehat{b}_{MCO}$.

2. Régression de $\log\left(\widehat{u}_i^2\right)$ sur les variables z_i : $\log\left(\widehat{u}_i^2\right) = z_i\theta + w_i$.
3. Construction d'un estimateur de $\widehat{\sigma}_i$ par $\widehat{\sigma}_i = \exp z_i'\widehat{\theta}/2$
4. Calcul des données sphéricisées : $\widetilde{y}_i = y_i/\widehat{\sigma}_i$, $\widetilde{x}_i = x_i/\widehat{\sigma}_i$
5. Calcul de l'estimateur des MCO sur ces données

7.0.6 Cas où $\Omega = \Sigma(\theta)$ et θ de dimension quelconque

On considère le modèle

$$\underline{y}_i = \underline{x}_i b + \underline{u}_i$$

avec les hypothèses

Hypothèse (H_1). $E(\underline{u}_i \,|\underline{x}_i) = 0$

Hypothèse (H_2). $V(\underline{u}_i \,|\underline{x}_i) = \Sigma(\theta)$ et θ de dimension quelconque

Hypothèse (H_3). Les observations $(\underline{y}_i, \underline{x}_i) \in \mathbb{R}\times\mathbb{R}^{K+1}$, $i = 1, ..., N$, sont iid

Hypothèse (H_4). $\forall\ N$ $X'X$ est non singulière

Hypothèse (H_5). $\mathrm{E}(\underline{x}_i\underline{x}_i')$ est inversible

Hypothèse (H_6). Les moments de $(\underline{y}_i, \underline{x}_i)$ existent au moins jusqu'à l'*ordre 8.*

Théorème 7.5. *Sous les hypothèses H1 à H6, l'estimateur des MCO*

$$\widehat{b}_{mco} = (X'X)^{-1}X'Y = \left(\overline{\underline{x}_i'\underline{x}_i}\right)^{-1}\overline{\underline{x}_i'\underline{y}_i}$$

verifie quand $N \to \infty$

1. $\widehat{b}_{mco} \xrightarrow{P} b$,
2. $\sqrt{N}\left(\widehat{b}_{mco} - b\right) \xrightarrow{L} \mathcal{N}\left(0, \mathrm{V}\left(\widehat{b}_{mco}\right)\right)$,
3. $\mathrm{V}\left(\widehat{b}_{mco}\right) = [\mathrm{E}(\underline{x}_i'\underline{x}_i)]^{-1}\,\mathrm{E}(\underline{x}_i'u_iu_i'\underline{x}_i)\,[\mathrm{E}(\underline{x}_i'\underline{x}_i)]^{-1}$
4. $\widehat{\mathrm{V}}\left(\widehat{b}_{mco}\right) = \overline{(\underline{x}_i'\underline{x}_i)}^{-1}\overline{\underline{x}_i'\widehat{u}_i\widehat{u}_i'\underline{x}_i\underline{x}_i'\underline{x}_i}^{-1} \xrightarrow{P} \mathrm{V}\left(\widehat{b}_{mco}\right)$
5. $\sqrt{N}\widehat{\mathrm{V}}\left(\widehat{b}_{mco}\right)^{-1/2}\left(\widehat{b}_{mco} - b\right) \xrightarrow{L} \mathcal{N}(0, I)$,

Démonstration.

1. Le premier point se démontre comme précédemment
2. Pour le deuxième point $\sqrt{N}\left(\widehat{b}_{mco} - b\right) = \left(\overline{\underline{x}_i'\underline{x}_i}\right)^{-1}\sqrt{N\underline{x}_i'\underline{u}_i}$
3. Théorème central limite appliqué à $\underline{x}_i'\underline{u}_i$: $E(\underline{x}_i'\underline{u}_i) = 0$ et $V(\underline{x}_i'\underline{u}_i) = E(\underline{x}_i'\underline{u}_i\underline{u}_i'\underline{x}_i)$ existent. On a donc $\sqrt{N\underline{x}_i'\underline{u}_i} \xrightarrow{L} N(0, E(\underline{x}_i'\underline{u}_i\underline{u}_i'\underline{x}_i))$
   On a donc

$$\begin{aligned}\sqrt{N}\left(\widehat{b}_{mco} - b\right) &= \left(\overline{\underline{x}_i'\underline{x}_i}\right)^{-1}\sqrt{N\underline{x}_i'\underline{u}_i} \\ &\xrightarrow{L} N\left(0, E(\underline{x}_i'\underline{x}_i)^{-1}E(\underline{x}_i'\underline{u}_i\underline{u}_i'\underline{x}_i)E(\underline{x}_i'\underline{x}_i)^{-1}\right)\end{aligned}$$

4. *Estimation de la matrice de variance*
   Le point important est de montrer que $\overline{\underline{x}_i'\widehat{u}_i\widehat{u}_i'\underline{x}_i} \xrightarrow{P} E\left(\underline{x}_i'\underline{u}_i\underline{u}_i'\underline{x}_i\right)$

$$\begin{aligned}
\overline{\underline{x}_i'\widehat{u}_i\widehat{u}_i'\underline{x}_i} &= \overline{\underline{x}_i'\left(\underline{x}_i\left(b-\widehat{b}_{mco}\right)+\underline{u}_i\right)\left(\underline{x}_i\left(b-\widehat{b}_{mco}\right)+\underline{u}_i\right)'\underline{x}_i} \\
&= \overline{\underline{x}_i'\underline{u}_i\underline{u}_i'\underline{x}_i} + \overline{\underline{x}_i'\underline{x}_i\left(b-\widehat{b}_{mco}\right)\left(b-\widehat{b}_{mco}\right)'\underline{x}_i'\underline{x}_i} + \\
&\quad \overline{\underline{x}_i'\underline{x}_i\left(b-\widehat{b}_{mco}\right)\underline{u}_i'\underline{x}_i} + \overline{\underline{x}_i'\underline{u}_i\left(b-\widehat{b}_{mco}\right)'\underline{x}_i'\underline{x}_i}
\end{aligned}$$

   Le premier terme converge vers $E\left(\underline{x}_i'\underline{u}_i\underline{u}_i'\underline{x}_i\right)$ car les moments d'ordre 8 existent.
   Le deuxième terme est une matrice dont les éléments sont somme de termes

$$\overline{(\underline{x}_i'\underline{x}_i)_{l_1l_2}\left(b-\widehat{b}_{mco}\right)_m\left(b-\widehat{b}_{mco}\right)_{m'}(\underline{x}_i'\underline{x}_i)_{l_1'l_2'}} =$$

   $\left(b-\widehat{b}_{mco}\right)_m\left(b-\widehat{b}_{mco}\right)_{m'}\overline{(\underline{x}_i'\underline{x}_i)_{l_1l_2}(\underline{x}_i'\underline{x}_i)_{l_1'l_2'}}$ comme $\left(b-\widehat{b}_{mco}\right) \xrightarrow{P} 0$ et que $\overline{(\underline{x}_i'\underline{x}_i)_{l_1l_2}(\underline{x}_i'\underline{x}_i)_{l_1'l_2'}} \xrightarrow{P} E\left((\underline{x}_i'\underline{x}_i)_{l_1l_2}(\underline{x}_i'\underline{x}_i)_{l_1'l_2'}\right)$ le deuxi ème terme tend vers zero en probabilité. De même pour le troisi ème et le quatrième terme.
   Cet estimateur de la matrice de variance de l'estimateur des mco est connu sous le nom de **matrice de variance de White robuste à l'hét éroscédasticité**. Il est trés couramment utilisé et syst ématiquement proposé dans les logiciels standards.
   Il faut néanmoins conservé à l'esprit que cet estimateur n'est convergeant que pour pour des échantillons de grandes taille pour lesquels on peut espérer que les moments d'ordre quatre calculés soient proches de leurs valeur moyenne

■

7.0.7 Application

Modèle hétéroscédastique en coupe
$V(u_i) = \sigma_i$

7.1 Tests d'hétéroscédasticité

On considère le cas des régressions en coupe
$y_i = x_i b + u_i$
$V(u_i) = \sigma_i^2$
(y_i, x_i) indépendants

7.1.1 Test de Goldfeld-Quandt

Si la variance σ_i^2 varie de façon monotone en fonction d'**emph** des variables explicatives (appelons-la $z_i \in \mathbb{R}$), on peut ordonner les observations en fonction

de z_i et supposer que $z_i \leq z_{i+1}$. On partitionne ensuite les observations en deux groupes tels que :

$$\begin{aligned} \underline{y}_1 &= \begin{pmatrix} y_1 \\ \vdots \\ y_{N_1} \end{pmatrix}, \quad X_1 = \begin{pmatrix} x_1' \\ \vdots \\ x_{N_1}' \end{pmatrix}, \\ \underline{y}_2 &= \begin{pmatrix} y_{N_2+1} \\ \vdots \\ y_N \end{pmatrix}, \quad X_2 = \begin{pmatrix} x_{N_2+1}' \\ \vdots \\ x_N' \end{pmatrix}. \end{aligned}$$

Les seuils N_1 et N_2 sont choisis de façon à écarter les deux échantillons. En pratique on prend $N_1 \approx N/3$ et $N_2 \approx 2N/3$.

On estime le modèle linéaire par la méthode des MCO sur chaque sous-échantillon. Soient

$$\begin{aligned} \widehat{\sigma}_1^2 &= \frac{1}{N_1 - K - 1} \sum_{i=1}^{N_1} (y_i - x_i' \widehat{b}_1)^2, \\ \widehat{\sigma}_2^2 &= \frac{1}{N - N_2 - K - 1} \sum_{i=N_2+1}^{N} (y_i - x_i' \widehat{b}_1)^2 \end{aligned}$$

les deux estimateurs de la variance.

Sous l'hypothèse d'homoscédasticité,

$$\begin{aligned} \widehat{\sigma}_1^2 &\sim \frac{\sigma_0^2}{N_1 - K - 1} \chi^2_{N_1 - K - 1}, \\ \widehat{\sigma}_2^2 &\sim \frac{\sigma_0^2}{N - N_2 - K - 1} \chi^2_{N - N_2 - K - 1}. \end{aligned}$$

Si bien que

$$\frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2} \sim F_{N_1 - K - 1, N - N_2 - K - 1}.$$

On rejettera l'hypothèse nulle d'homoscédasticité (sous l'hypothèse maintenue de normalité) au seuil α si :

$$\frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2} > F_{N_1 - K - 1, N - N_2 - K - 1}(1 - \alpha)$$

où $F_{N_1 - K - 1, N - N_2 - K - 1}(1-\alpha)$ est le quantile $1-\alpha$ de la loi de Fisher à $N_1 - K - 1$ et $N - N_2 - K - 1$ degrés de liberté

7.1.2 Test de Breusch-Pagan

On considère une hypothèse alternative à l'hypothèse d'homoscédasticité de la forme :

$$H_a : \sigma_i^2 = \sigma_0^2 + z_i' \gamma_0$$

où $\sigma_0^2 \in \mathbb{R}$ et $\gamma_0 \in \mathbb{R}^M$ sont deux paramètres et où z_i est maintenant un vecteur quelconque de M variables explicatives formées à partir de x_i (par exemple, les

variables de x_i et leurs produits croisés). Attention, on ne garde dans z_i que des variables, pas de terme constant. L'hypothèse nulle d'homoscédaticité s'écrit :

$$H_0 : \gamma_0 = 0.$$

Le test de Breusch-Pagan se fait de la façon suivante :

1. Estimer le modèle linéaire par MCO et calculer le carré des résidus : $\widehat{u}_i^2$;
2. Régresser par MCO $\widehat{u}_i^2$ sur les variables z_i avec une constante. Soit R^2 le coefficient de détermination de cette régression ;
3. Sous l'hypothèse nulle, $NR^2 \xrightarrow{L} \chi^2_M$. On rejette H_0 au seuil α si $NR^2 > \chi^2_1(M)$.

Remarque. Le test se fait à partir des résidus estimés ($\widehat{u}_i^2/\widehat{\sigma}^2$). Montrer que tout se passe comme si l'on travaillait avec u_i^2/σ_0^2 nécessite des hypothèses supplémentaires.

8 Autocorrelation des résidus

Dans les modèles en série temporelles et en données de panel, l'hypothèse de *non-autocorrélation des perturbations* est assez forte et fréquemment non-vérifiée.

On considère les modèles sur série temporelle :

$$y_t = x_t b + u_t, \quad t = 1, \ldots, T$$

On va voir à ce sujet :
- les principales formes d'autocorrélation ;
- les tests permettant de détecter l'autocorrelation ;
- les méthodes d'estimation adaptées en présence d'autocorrélation.

8.1 Les diverses formes d'autocorrélation des perturbations

8.1.1 Perturbations suivant un processus autorégressif d'ordre 1 (AR1)

Selon cette hypothèse (AR1), les perturbations du modèle sont engendrées par le processus :

$$u_t = \rho\, u_{t-1} + \varepsilon_t, t = 1, ...T$$

avec :
- $E(\varepsilon_t | X) = 0$, $V(\varepsilon_t | X) = \sigma_\varepsilon^2$, cov $(\varepsilon_t, \varepsilon_{t'} | X) = 0, \quad \forall\, t \neq t'$: les hypothèses d'homoscédasticité et d'indépendance sont transférées *aux innovations du processus* : ε_t
- $|\rho| < 1$

8.1.2 Stationnarité au premier et au second ordre d'un processus AR1

$$\begin{aligned} u_t &= \rho u_{t-1} + \varepsilon_t = \rho(\rho\, u_{t-2} + \varepsilon_{t-1}) + \varepsilon_t = \varepsilon_t + \rho\, \varepsilon_{t-1} + \rho^2(\rho\, u_{t-3} + \varepsilon_{t-2}) \\ &= \varepsilon_t + \rho\, \varepsilon_{t-1} + \cdots + \rho^{t-1}\varepsilon_1 + \rho^t u_0 \end{aligned}$$

Le processus u_t est dit *stationnaire au premier ordre* et *au second ordre* si et seulement si :

$$E(u_t | X) = E(u_{t-1} | X) = \cdots = E(u_0 | X) = \theta, \ \forall\, t$$
$$V(u_t | X) = V(u_{t-1} | X) = \cdots = V(u_0 | X) = \sigma_u^2, \ \forall\, t.$$

Le processus AR(1) u_t est stationnaire si $E(u_0 | X) = 0$ et $V(u_0 | X) = \sigma_\varepsilon^2 / \left(1 - \rho^2\right)$ et $cov(\varepsilon_t, u_0) = 0$. Ces conditions sont satisfaites si le processus engendrant u_t débute en $-\infty$.

Compte tenu de l'expression : $u_t = \varepsilon_t + \cdots + \rho^{t-1}\varepsilon_1 + \rho^t u_0$.

On a : $E(u_t | X) = E(\varepsilon_t | X) + \cdots + \rho^{t-1} E(\varepsilon_1 | X) + \rho^t E(u_0 | X) = 0$

De même, compte tenu de l'indépendance des chocs ε_s entre eux et leur indépendance avec u_0

$$\begin{aligned} V(u_t|X) &= V(\varepsilon_t|X) + \rho^2 V(\varepsilon_{t-1}|X) + \cdots + \rho^{2(t-1)} V(\varepsilon_1|X) + \rho^{2t} V(u_0|X) \\ &= \sigma_\varepsilon^2 \left(1 + \rho^2 + \cdots + \rho^{2(t-1)}\right) + \rho^{2t}\sigma_{u_0}^2 \\ &= \sigma_\varepsilon^2 \frac{1-\rho^{2t}}{1-\rho^2} + \rho^{2t}\sigma_{u_0}^2 = \frac{\sigma_\varepsilon^2}{1-\rho^2} - \rho^{2t}\frac{\sigma_\varepsilon^2}{1-\rho^2} + \rho^{2t}\sigma_{u_0}^2 \end{aligned}$$

Si $\sigma_{u_0}^2 = \sigma_\varepsilon^2/\left(1-\rho^2\right)$ on a

$$V(u_t|X) = \sigma_\varepsilon^2/\left(1-\rho^2\right)$$

Si le processus remonte en $-\infty$ on a :

$$u_t = \lim \sum_{s=0}^{\infty} \rho^s \varepsilon_{t-s}$$

On a donc

$$V(u_t|X) = \lim \sum_{s=0}^{\infty} \rho^{2s}\sigma_\varepsilon^2 = \frac{\sigma_\varepsilon^2}{(1-\rho^2)}$$

Réciproquement si le processus est stationnaire on a :

$$V(u_t|X) = V(\rho u_{t-1} + \varepsilon_t|X) = \rho^2 V(u_{t-1}|X) + V(\varepsilon_t)$$

$$\begin{aligned} V(u_t|X) &= \rho^2 V(u_{t-1}|X) + \sigma_\varepsilon^2 \\ \sigma_u^2\left(1-\rho^2\right) &= \sigma_\varepsilon^2 \end{aligned}$$

8.1.3 Covariance entre deux perturbations d'un processus AR(1)

$$Cov(u_t, u_{t-s}|X) = \rho^s \frac{\sigma_\varepsilon^2}{1-\rho^2}$$

En effet, on a :

$$u_t = \rho u_{t-1} + \varepsilon_t = \rho\left[\rho\, u_{t-2} + \varepsilon_{t-1}\right] + \varepsilon_t = \rho^s\, u_{t-s} + \rho^{s-1}\, \varepsilon_{t-(s-1)} + ... + \varepsilon_t$$

Par conséquent

$$\begin{aligned} cov(u_t, u_{t-s}|X) &= E\left(\left(\rho^s u_{t-s} + \rho^{s-1}\varepsilon_{t-s+1} + ... + \varepsilon_t\right) u_{t-s}|X\right) \\ &= \rho^s E\left(u_{t-s}^2|X\right) + \rho^{s-1} E\left(\varepsilon_{t-s+1}u_{t-s}|X\right) + ... + E\left(\varepsilon_t u_{t-s}|X\right) \end{aligned}$$

Comme $E\left(\varepsilon_{t-(s-i)}, u_{t-s}|X\right) = 0,\ \forall\, i \neq 0$ on a bien l'expression cherchée.

8.1.4 Matrice de variances-covariances des perturbations

$$V(U|X) = \frac{\sigma_\varepsilon^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \vdots & \vdots & & & \vdots \\ \rho^{T-2} & & \cdots & 1 & \rho \\ \rho^{T-1} & \rho^{T-2} & \cdots & \rho & 1 \end{bmatrix}$$

Expression simple :
- traduisant une idée simple : un *choc exogène* à un moment donné, a un effet *persistant* mais décroissant exponentiellement avec le temps.
- permettant la mise en oeuvre facile de méthodes d'estimation plus efficaces que les MCO (telles les MCQG).

8.1.5 Perturbations suivant un processus AR(p)

u_t suit un processus autorégressif d'ordre p noté AR(p) si :

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_p u_{t-p} + \varepsilon_t$$

soit

$$A(L)\, u_t = \varepsilon_t$$

avec $A(Z) = 1 - \rho_1 Z - \rho_2 Z^2 - \cdots - \rho_p Z^p$, $E(\varepsilon_t|X) = 0$, $V(\varepsilon_t|X) = \sigma_\varepsilon^2$ et $cov(\varepsilon_t, \varepsilon_{t'}|X) = 0$, $\forall\, t \neq t'$

On montre que pour que le processus $AR(p)$ soit stationnaire

$$Vu_t = \sigma_u^2,\ \text{cov}(u_t, u_{t-s}) = \theta_s$$

il faut que les racines du polynome $\Phi(Z)$ soient de module supérieur à 1.

Exemple (Cas d'un processus $AR(2)$). Les contraintes sur ρ_1 et ρ_2 sont :

$$\rho_1 + \rho_2 < 1,\ \rho_2 - \rho_1 < 1 \text{ et } |\rho_2| < 1$$

Les variances et covariances des perturbations u_t sont alors :
$Vu_t = \sigma_u^2 = \frac{1-\rho_2}{(1+\rho_2)\left[(1-\rho_2)^2-\rho_1^2\right]}\sigma_\varepsilon^2 = \Psi_0, \forall\, t$
$\text{cov}\,(u_t, u_{t-1}) = \frac{\rho_1}{1-\rho_2}\sigma_u^2 = \Psi_1$
$\text{cov}\,(u_t, u_{t-2}) = \rho_2\sigma_u^2 + \frac{\rho_1^2}{1-\rho_2}\sigma_u^2 = \Psi_2 = \rho_2\Psi_0 + \rho_1\Psi_1$
$\text{cov}\,(u_t, u_{t-s}) = \Psi_s = \rho_1\Psi_{s-1} + \rho_2\Psi_{s-2},\ s > 2$
Exemple

$$u_t = -0.5u_{t-1} + 0.3u_{t-2} + e_t$$

Soit : $(1 + 0.5L - 0.3L^2)u_t = e_t$
On détermine les racines du polynome $1 + 0.5z - 0.3z^2$
Le discriminant vaut

$$\Delta = (0.5)^2 - 4(-0.3) = 0.25 + 1.2 = 1.45 = (1.204)^2 > 0$$

et les racines sont donc

$$z_1 = \frac{-0.5 - 1.204}{2(-0.3)} = 2.84 \text{ et } z_2 = \frac{-0.5 + 1.204}{2x(-0.3)} = -1.17$$

Le processus est donc stationnaire puisque les racines sont supérieures à 1 en valeur absolue.

8.1.6 Perturbations suivant un processus de moyenne mobile d'ordre q MA(q)

La perturbation u_t suit un processsus de moyenne d'ordre q noté $MA(q)$ si :

$$u_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

avec $E\varepsilon_t = 0$, $V\varepsilon_t = \sigma_\varepsilon^2$ et cov $(\varepsilon_t, \varepsilon_{t'}) = 0 \quad \forall\, t \neq t'$
Là encore les hypothèses iid sont transposées au processus ε_t.
Le modèle se réécrit donc :

$$u_t = B(L)\,\varepsilon_t$$

avec $B(Z) = 1 + \theta_1 Z + \theta_2 Z^2 + \cdots + \theta_q Z^q$

Application : Les *valeurs anticipées* de variables interviennent souvent dans les modèles économétriques. Elles sont toujours *non-observables* et il faut donc les modéliser. On retiend souvent un schéma adaptatif. L'anticipation $x_t^\star$ de la variable x_t est modélisée suivant un processus adaptatif

$$\begin{aligned} x_t^\star - x_{t-1}^\star &= \theta^\star (x_{t-1} - x_{t-1}^\star),\ |\theta^\star| < 1 \\ x_t^\star &= (1-\theta^\star)\, x_{t-1}^\star + \theta^\star x_{t-1} \end{aligned}$$

Les anticipations sont révisées d'une période à l'autre en fonction de l'erreur d'anticipation commise à la période précédente.
Le processus s'écrit encore

$$[1 - (1-\theta^\star)\, L]\, x_t^\star = \theta^\star x_{t-1} = \theta^\star L x_t$$

et on peut le résoudre comme

$$\begin{aligned} x_t^\star &= \frac{\theta^\star L}{[1-(1-\theta^\star)\,L]} x_t = \theta^\star \left[L \sum_{s=0}^{\infty} (1-\theta^\star)^s\, L^s \right] x_t \\ &= \sum_{s=0}^{\infty} \theta^\star\, (1-\theta^\star)^s\, x_{t-s-1} \end{aligned}$$

Les anticipations $x_t^\star$ apparaissent ainsi comme une somme pondérée infinie (avec des poids décroissants exponentiellement) des valeurs passées de x_t.
Si le modèle que l'on souhaite estimer s'écrit :

$$y_t = a x_t^\star + \varepsilon_t$$

en le prémultipliant par $[1-(1-\theta^\star)\,L]$,on obtient :

$$[1-(1-\theta^\star)\,L]\, y_t = a\,[1-(1-\theta^\star)\,L]\, x_t^\star + [1-(1-\theta^\star)\,L]\,\varepsilon_t$$

Le modèle se réécrit donc

$$\begin{aligned} y_t &= (1-\theta^\star)\, y_{t-1} + a\theta^\star x_{t-1} + [\varepsilon_t - (1-\theta^\star)\,\varepsilon_{t-1}] \\ &= \theta y_{t-1} + a' x_{t-1} + u_t \end{aligned}$$

avec $\theta = 1 - \theta^\star$, $a' = a\theta^\star$ et $u_t = \varepsilon_t - \theta\varepsilon_{t-1}$.

La perturbation u_t suit donc un processus $MA\,(1)$ et on a dans ce cas particulier :

$Vu_t = V\,(\varepsilon_t - \theta\varepsilon_{t-1}) = \sigma_\varepsilon^2(1+\theta^2)$

$\text{cov}(u_t, u_{t-1}) = -\theta\sigma_\varepsilon^2$

$\text{cov}\,(u_t,\ u_{t-s}) = 0, \quad \forall\ s > 1$

soit la matrice de variance covariance :

$$V\ u = \sigma_\varepsilon^2 \begin{bmatrix} 1+\theta^2 & -\theta & 0 & \cdots & 0 \\ -\theta & 1+\theta^2 & -\theta & & \vdots \\ 0 & -\theta & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & -\theta \\ 0 & \cdots & 0 & -\theta & 1+\theta^2 \end{bmatrix}$$

8.1.7 Perturbation suivant un processus ARMA(p,q)

La perturbation u_t suit un processus ARMA(p,q) si l'on peut écrire :

$$A(L)u_t = B(L)\varepsilon_t$$

avec

$$\begin{aligned} A(L) &= 1 - \rho_1 L - \rho_2 L^2 - \cdots - \rho_p L^p \\ B(L) &= 1 + \theta_1\ L + \theta_2\ L^2 + \cdots + \theta_q L^q \end{aligned}$$

et

$$E\,(\varepsilon_t) = 0,\ V\,(\varepsilon_t) = \sigma_\varepsilon^2, Cov\,(\varepsilon_t, \varepsilon_{t'}) = 0\ \ \forall\ t \neq t'$$

Exemple (processus ARMA(1,1)).

$$u_t = \rho u_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}$$

Par conséquent

$$\sigma_u^2 = Vu_t = \rho^2 E\left(u_{t-1}^2\right) + E\left(\varepsilon_t^2\right) + \theta^2 E\left(\varepsilon_{t-1}^2\right) + 2\theta\rho E\,(u_{t-1}\varepsilon_{t-1})$$

Comme $E\,(u_t\varepsilon_t) = E\left(\varepsilon_t^2\right) = \sigma_\varepsilon^2$, on a $\sigma_u^2 = \rho^2\sigma_u^2 + \sigma_\varepsilon^2 + \theta^2\sigma_\varepsilon^2 + 2\theta\rho\sigma_\varepsilon^2$, d'où

$$Vu_t = \sigma_\varepsilon^2\left(\tfrac{1+\theta^2+2\theta\rho}{1-\rho^2}\right) = \sigma_\varepsilon^2 w_0,\ \ \forall\ t.$$

De même

$$\begin{aligned} cov(u_t, u_{t-1}) &= \rho E\left(u_{t-1}^2\right) + \theta E\,(u_{t-1}\varepsilon_{t-1}) \\ &= \sigma_u^2 + \theta\sigma_\varepsilon^2 = \sigma_\varepsilon^2\frac{(1+\theta\rho)(\theta+\rho)}{1-\rho^2} = \sigma_\varepsilon^2 w_1 \end{aligned}$$

et $\forall\ s > 1$

$$cov(u_t, u_{t-s}) = \rho cov(u_{t-1}, u_{t-s}) = \rho cov(u_t, u_{t-(s-1)}) = \rho^{s-1}\sigma_\varepsilon^2 w_1$$

soit

$$V\ u = \sigma_\varepsilon^2 \begin{bmatrix} w_0 & w_1 & \rho w_1 & \rho^2 w_1 & \cdots & \rho^{T-2} w_1 \\ w_1 & w_0 & w_1 & \rho w_1 & \ddots & \vdots \\ \rho w_1 & w_1 & \ddots & \ddots & \ddots & \rho^2 w_1 \\ \rho^2 w_1 & \rho w_1 & \ddots & \ddots & w_1 & \rho w_1 \\ \vdots & \ddots & \ddots & w_1 & w_0 & w_1 \\ \rho^{T-2} w_1 & \cdots & \rho^2 w_1 & \rho w_1 & w_1 & w_0 \end{bmatrix}$$

8.1.8 Détection de l'autocorrélation : le test de Durbin et Watson (1950, 1951)

Considérons le *modèle AR(1)* : $u_t = \rho\ u_{t-1} + \varepsilon_t$

Pour ce modèle, tester l'absence d'autocorrélation revient à tester : $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$

Le test le plus fréquemment utilisé est celui de Durbin-Watson, reposant sur la statistique :

$$\widehat{d} = \frac{\sum_{t=2}^{T} \left(\widehat{u}_t - \widehat{u}_{t-1}\right)^2}{\sum_{t=1}^{T} \widehat{u}_t^2}$$

Cette statistique est liée asymptotiquement au paramètre ρ par la relation suivante :

$$p \lim_{T\to\infty} \widehat{d} = 2(1-\rho)$$

En effet :

$$p \lim_{T\to\infty} \widehat{d} = p \lim \frac{\frac{1}{T}\sum_{t=2}^{T} \widehat{u}_t^2 - 2\frac{1}{T}\sum_{t=2}^{T} \widehat{u}_t \widehat{u}_{t-1} + \frac{1}{T}\sum_{t=2}^{T} \widehat{u}_{t-1}^2}{\frac{1}{T}\sum_{t=1}^{T} \widehat{u}_t^2}$$

$$= 1 - 2\rho + 1 = 2(1-\rho)$$

puisque

$$p \lim \frac{1}{T}\sum_{t=2}^{T} \widehat{u}_t^2 = p \lim \frac{1}{T}\sum_{t=2}^{T} \widehat{u}_{t-1}^2 = p \lim \frac{1}{T}\sum_{t=1}^{T} \widehat{u}_t^2$$

et que

$$\frac{p \lim \frac{1}{T}\sum \widehat{u}_t \widehat{u}_{t-1}}{p \lim \frac{1}{T}\sum_{t=1}^{T} \widehat{u}_t^2} = \frac{Cov\left(u_t, u_{t-1}\right)}{V\left(u_t\right)} = \rho$$

Par conséquent :

– si ρ est nul (absence d'autocorrélation), $\widehat{d}$ est proche de 2,
– si ρ est proche de 1 (forte autocorrélation positive), $\widehat{d}$ est proche de 0
– si ρ est proche de -1 (forte autocorrélation négative), $\widehat{d}$ est proche de 4

La loi de probabilité de la statistique $\widehat{d}$ est difficile à é tablir car elle dépend des résidus estimés et donc des valeurs prises par les variables explicatives du modèle.

Sous l'hypothèse H_0 *:*$\rho = 0$, il existe deux statistiques, d_e et d_u, qui encadrent toujours $\widehat{d}$:

$$d_\ell < \widehat{d} < d_u,$$

et dont la loi ne dépend que de T et K.

Test de $H_0 : \rho = 0$ **contre** $H_1 : \rho > 0$ Si $\widehat{d}$ est proche de 2 on accepte l'hypothèse. Si $\widehat{d}$ est en revanche trop faible on rejette l'hypothèse. Si on connaissait la loi d_0 de $\widehat{d}$, on pourrait déterminer le fractile $d^\star(\alpha)$ de cette loi permettant de conclure au rejet ou à l'acceptation de l'hypothèse H_0 de non-autocorrélation pour un test au seuil α.

$$P\left(d_0 < d^\star(\alpha)\right) = \alpha$$

Ne connaissant pas la loi asymptotique de $\widehat{d}$ on détermine les fractiles correspondants $d_\ell^\star(\alpha)$ de d_l et $d_u^\star(\alpha)$ de d_u

$$\begin{aligned} P\left(d_l < d_\ell^\star(\alpha)\right) &= \alpha \\ P\left(d_u < d_u^\star(\alpha)\right) &= \alpha \end{aligned}$$

Comme

$$d_l < d_0 < d_u$$

On a

$$d_\ell^\star(\alpha) < d^\star(\alpha) < d_u^\star(\alpha)$$

– *Si* $\widehat{d}$ *est inférieure à* $d_\ell^\star(\alpha)$, alors $\widehat{d} < d^\star(\alpha)$: on refuse H_0
– *Si* $\widehat{d}$ *est supérieure à* $d_u^\star(\alpha)$, alors $\widehat{d} > d^\star(\alpha)$: on accepte H_0
– *Si* $d_\ell^\star < \widehat{d} < d_u^\star$, on se trouve dans la zone dite *inconclusive* : le test ne permet pas de conclure au rejet ou à l'acceptation de H_0.

La pratique courante consiste à inclure la zone inconclusive dans la zone de rejet de l'hypothèse H_0 pour se garantir contre le risque d'accepter à tort l'absence d'autocorrélations. L'amplitude de la zone inconclusive, $d_u^\star - d_\ell^\star$, est d'autant plus importante que le nombre T d'observations est faible et que le nombre de variables explicatives est important.

Test de $H_0 : \rho = 0$ **contre** $H_1 : \rho < 0$ On utilise la statistque $4 - \widehat{d}$. Sous H_0 $\widehat{d} = 2$ sous H_1 $\rho < 0$, alors $plim\widehat{d}$=$2(1-\rho) > 2$ donc $plim\left(4 - \widehat{d}\right) < 2$ On rejettera l'hypothèses pour des valeurs faibles de $4 - \widehat{d}$ par rapport à 2. On a :

$$4 - d_u^\star < 4 - d^\star < 4 - d_\ell^\star$$

Par conséqent :
– *si* $4 - \widehat{d} > 4 - d_\ell^\star$, alors $4 - \widehat{d} > 4 - d^\star$: *on accepte* H_0.
– *si* $4 - \widehat{d} < 4 - d_u^\star$, alors $4 - \widehat{d} < 4 - d^\star$: *on refuse* H_0.

– enfin, *si* $4-d_u^\star < 4-\widehat{d} < 4-d_\ell^\star$: on est dans la zone *inconclusive*.

On inclut comme précédemment la zone inconclusive dans la zone de rejet de H_0.

Remarque.
1. Les lois (tabulées) de d_ℓ et d_u ont été établies par Durbin et Watson pour un modèle avec constante et perturbations AR(1) ;
2. Bien qu'il soit spécifiquement destiné à tester l'absence d'autocorrélation contre l'hypothèse alternative d'une autocorrélation associée à un processus AR(1), le test de D.W. se révèle capable de détecter d'autres formes d'autocorrélations ;

   Exemple. MA(1) ou AR(2). Dans les autres situations, il est préférable de recouvrir à d'autres tests.

8.2 Estimateurs des MCO, des MCG et des MCQG dans un modèle dont les perturbations sont autocorrélées

On considère le cas d'un modèle

$$y_t = x_t\, b + u_t$$

avec
$E(U|X) = 0$
$V(U|X) = \Sigma$ de dimension $T \times T$
$\frac{1}{T}X'X \xrightarrow{P} Q_{XX}$, $X'X$ et Q_X inversibles
$\frac{1}{T}X'\Sigma X \xrightarrow{P} Q_{X\Sigma X}$
Alors l'estimateur des mco

$$\widehat{b}_{mco} = (X'X)^{-1} X'Y$$

vérifie
$E\left(\widehat{b}_{mco}|X\right) = b$: l'estimateur est sans biais
$V\left(\widehat{b}_{mco}|X\right) = (X'X)^{-1} X'\Sigma X (X'X)^{-1}$
$\widehat{b}_{mco} \xrightarrow{P} b$: convergence
$\sqrt{T}\left(\widehat{b}_{mco} - b\right) \xrightarrow{L} N\left(0, Q_{XX}^{-1} Q_{X\Sigma X} Q_{XX}^{-1}\right)$: normalité asymptotique

8.2.1 Estimation de la matrice de variance

Si la matrice Σ dépend d'un nombre fini de paramètres : $\Sigma = \Sigma(\theta)$, cas par exemple du modèle $AR(1)$, du modèle $MA(1)$, ou du modèle $ARMA(1,1)$, et si on dispose d'un estimateur $\widehat{\theta}$ convergent de θ, on peut estimer de manière convergente la matrice de variance asymptotique $Q_{XX}^{-1} Q_{X\Sigma X} Q_{XX}^{-1}$ par

$$\widehat{V}_{as} = \left(\frac{X'X}{T}\right)^{-1} \frac{X'\Sigma\left(\widehat{\theta}\right)X}{T} \left(\frac{X'X}{T}\right)^{-1}$$

Un tel estimateur $\widehat{\theta}$ peut être obtenu en général à partir de l'estimateur des mco.

Exemple. Dans le cas du modèle $AR(1)$ on a

$$u_t = \rho u_{t-1} + \varepsilon_t$$

La variance des résidus s'écrit

$$V\ u = \sigma_u^2\ \Omega = \frac{\sigma_\varepsilon^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & & \rho^{T-2} & \rho^{T-1} \\ \rho & 1 & \ddots & & \rho^{T-2} \\ & \ddots & \ddots & \ddots & \\ \rho^{T-2} & & \ddots & \ddots & \rho \\ \rho^{T-1} & \rho^{T-2} & & \rho & 1 \end{bmatrix}$$

On peut construire le résidu estimé

$$\widehat{u}_t = y_t - x_t \widehat{b}_{mco}$$

et on estime ρ par application des mco sur le modèle

$$\widehat{u}_t = \rho \widehat{u}_{t-1} + \widetilde{\varepsilon}_t$$

soit

$$\hat{\rho} = \frac{\Sigma_{t=2}^T\ \widehat{u}_t \widehat{u}_{t-1}}{\Sigma_{t=2}^T\ \widehat{u}_{t-1}^2}$$

L'estimateur des MCG Sous les hypothèses
$E(U|X) = 0$, $V(U|X) = \Sigma$ de dimension $T \times T$ inversible, $X'X$ inversible
Le meilleur estimateur linéaire sans biais de b est :

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

Sa variance est donnée par :

$$V\ \hat{b}_{mcg} = (X'\ \Sigma^{-1}\ X)^{-1}$$

Il peut être obtenu comme estimateur des mco dans le modèle :

$$\Sigma^{-1/2}\ Y = \Sigma^{-1/2}\ Xb + \Sigma^{-1/2}\ U$$

où $\Sigma^{-1/2}\Sigma^{-1/2'} = \Sigma^{-1}$

Dans le cas particulier où les perturbations suivent un processus AR(1), une telle transformation peut être donnée par :

$$\Sigma^{-1/2} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & \cdots & \cdots & \cdots & 0 \\ -\rho & 1 & \ddots & & & \vdots \\ 0 & -\rho & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 1 & 0 \\ 0 & \cdots & \cdots & 0 & -\rho & 1 \end{bmatrix}$$

L'estimateur des MCG peut alors être calculé comme estimateur des mco appliqu é au modèle :

$$\begin{bmatrix} y_1\sqrt{1-\rho^2} \\ y_2-\rho\, y_1 \\ \vdots \\ y_T-\rho\, y_{T-1} \end{bmatrix} = \begin{pmatrix} x_1\sqrt{1-\rho^2} \\ x_2-\rho x_1 \\ \vdots \\ x_T-\rho\, x_{T-1} \end{pmatrix} b + \begin{pmatrix} u_1\,\sqrt{1-\rho^2} \\ u_2-\rho\, u_1 \\ \vdots \\ u_T-\rho\, u_{T-1} \end{pmatrix}$$

Remarque.
1. Si le modèle initial comporte une variable constante, le modèle transformé n'en comporte plus.
2. Pour calculer cet estimateur MCG, il faut *connaître* ρ

L'estimateur des MCQG Sous les hypothèses
$E(U|X)=0$
$V(U|X)=\Sigma(\theta)$ de dimension $T\times T$, θ de dimension finie
$\frac{1}{T}X'X\xrightarrow{P}Q_{XX}$, $X'X$ et Q_X inversibles
$\frac{1}{T}X'\Sigma^{-1}X\xrightarrow{P}Q_{X\Sigma^{-1}X}$ inversible
$\widehat{\theta}\xrightarrow{P}\theta$ on dispose d'un estimateur convergent de θ
L'estimateur des MCQG

$$\widehat{b}_{mcqg}=\left(X'\Sigma\left(\widehat{\theta}\right)^{-1}X\right)^{-1}X'\Sigma\left(\widehat{\theta}\right)^{-1}Y$$

vérifie
$\widehat{b}_{mcqg}\xrightarrow{P}b$: convergence
$\sqrt{T}\left(\widehat{b}_{mcqg}-b\right)\xrightarrow{L}N(0,V_{as}(mcqg))$: normalité asymptotique
$V_{as}(mcqg)=Q^{-1}_{X\Sigma^{-1}X}=p\lim TV(mcg)$ équivalence entre mcqg et mcg
$\widehat{V}_{as}(mcqg)=\left(\frac{1}{T}X'\Sigma\left(\widehat{\theta}\right)^{-1}X\right)^{-1}\xrightarrow{P}V_{as}(mcqg)$ estimation de la matrice de variance

Cas des perturbations $AR(1)$: **L'estimateur de Prais-Watson** (1954).
C'est un estimateur en plusieurs étapes :
- estimation par MCO du modèle $y_t=x_tb+u_t, t=1,...,T$
- calcul des résidus estimés : $\widehat{u}_t=y_t-x_t\hat{b}_{mco}$
- estimation de ρ par application des mco au modèle :

$$\widehat{u}_t=\rho\widehat{u}_{t-1}+\varepsilon_t,\quad t=2,...,T$$

soit

$$\hat{\rho}=\frac{\Sigma_{t=2}^{T}\,\widehat{u}_t\widehat{u}_{t-1}}{\Sigma_{t=2}^{T}\,\widehat{u}_{t-1}^2}$$

- calcul des données transformées :
  $\tilde{y}_1=\sqrt{1-\hat{\rho}^2}\,y_1$ et $\tilde{y}_t=y_t-\hat{\rho}y_{t-1},\quad t=2,...,T$
  $\tilde{x}_1=\sqrt{1-\hat{\rho}^2}\,x_1$ et $\tilde{x}_t=x_t-\hat{\rho}x_{t-1},\quad t=2,...,T$
- estimation des MCO du *modèle transformé sans constante* :

$$\tilde{y}_t=\tilde{x}_tb+\tilde{u}_t,\quad t=1,...,T$$

L'estimateur $\hat{b}$ ainsi obtenu est convergent et asymptotiquement aussi efficace que l'estimateur des MCG.

9 Introduction aux variables instrumentales

On a considéré jusqu'à présent le cas de modèle s'é crivant

$$y_i = b_0 + x_i^1 b_1 + \cdots + x_i^K b_K + u_i$$

avec l'hypothèse

$$E\left(x_i^{'} u_i\right) = 0 \text{ ou } E\left(u_i \left| x_i \right.\right) = 0$$

Cette hypothèse peut aussi constituer une définition du paramètre b. Dans ce cas le coefficient b s'interprete comme le vecteur des coefficients de la regression linéaire de y_i sur le vecteur de variables x_i. Une telle définition présente un intérêt dans une approche descriptive des données.

Neanmoins on est fréquemment amener à estimer des modèles structurels dans lesquels les paramètres ont un sens économique. Le plus simple d'entre eux est certainement la fonction de production

$$y_i = a + \alpha k_i + \beta l_i + u_i$$

le paramètre α mesure l'incidence d'une augmentation de 1 du stock de capital sur la production. Ce paramètre n'a aucune raison de coïncider avec celui de la regression linéaire. On est ainsi fréquemment amener à considerer des modèles structurels pour lesquels on a une équation linéaire entre une variable d'intérêt et des variables explicatives mais pour laquelle on n'a pas nécessairement la relation $E\left(u_i \left| x_i \right.\right) = 0$.

On donne trois exemples type dans lesquels on a ce type d'endogeneité des regresseurs

9.0.2 Erreur de mesure sur les variables

On considère la situation dans laquelle on a un modèle structurel

$$y_i = x_i^* b + u_i$$

La variable x_i^* est supposé pour simplifier de dimension 1 et centrée comme la variable y_i et on fait l'hypothè se $E\left(u_i \left| x_i^* \right.\right) = 0$

On suppose en outre que la variable x_i^* est mesurée avec erreur :

$$x_i = x_i^* + e_i$$

avec $E\left(e_i \left| x_i^* \right.\right) = 0$ et u_i et e_i non corrélées.

Dans ces conditions le modèle dont on dispose est

$$y_i = x_i b + u_i - b e_i$$

On est dans une situation dans laquelle le résidu de l'équation $v_i = u_i - b e_i$ est corrélé avec la variable explicative

$$\begin{aligned} E\left(v_i x_i\right) &= E\left(\left(u_i - b e_i\right)\left(x_i^* + e_i\right)\right) \\ &= E\left(u_i x_i^*\right) + E\left(u_i e_i\right) - bE\left(e_i x_i^*\right) - bE\left(e_i^2\right) \\ &= -b\sigma_e^2 \neq 0 \end{aligned}$$

On voit alors trés facilement qu'à la limite le parmètre de la regression linéaire ne coïncide pas avec celui du modèle : l'estimateur des mco n'est pas convergent.

$$b_{mco} \xrightarrow{P} b + \frac{E\left(x_i^{'} v_i\right)}{E\left(x_i^{'} x_i\right)} = b\left(1 - \frac{\sigma_e^2}{\sigma_e^2 + \sigma_{x^*}^2}\right)$$

9.0.3 Omission de régresseur, hétérogénéité inobservée

On considère le modèle

$$y_i = x_i b + z_i c + u_i$$

Il y a donc un facteur z_i dont on sait qu'il explique la variable y_i. On considère la situation dans laquelle cette variable n'est pas observée.

L'omission de cette variable conduit à une estimation non convergente du modèle par les mco dés lors que cette variable est corrélée avec les régresseurs. On a en effet

$$\begin{aligned} \widehat{b}_{mco} \xrightarrow{P} b + E\left(x_i^{'} x_i\right)^{-1} E\left(x_i^{'}\left(z_i c + u_i\right)\right) &= b + E\left(x_i^{'} x_i\right)^{-1} E\left(x_i^{'} z_i\right) c \\ &= b + \lambda_{z_i/x_i} c \end{aligned}$$

Avec $E\left(x_i^{'} u_i\right) = 0$ et λ_{z_i/x_i} le coefficient de la regression linéaire de z_i sur x_i.

On peut considérer par exemple le cas d'une fonction de production agricole : y_i est le rendement de la terre, x_i la quantité d'engrais, b le rendement des épendages et z_i la qualité de la terre. L'omission de cette variable biaise l'estimation du paramètre technologique b si les décisions d'épendage d'engrais dépendent de la qualité de la terre.

Un autre exemple est donné par les équation dites de Mincer reliant le salaire à l'éducation

$$w_i = \alpha_0 + \alpha_s s_i + u_i$$

Le paramètre α_s mesure l'effet d'une année d'étude supplémentaire sur le niveau de salaire. Dans l'ensemble des causes inobservées affectant le salaire se trouve entre autres le niveau d'aptitude de l'individu. Mais le choix d'un niveau d'étude s_i est une décision rationnelle de la part de l'agent, fonction de l'aptitude de l'individu.

9.0.4 La simultanéité

La simultanéité est la situation dans laquelle certains des régresseurs et la variable à expliquer sont déterminés simultanément. Un exemple typique est celui d'un équilibre offre-demande. Une équation de demande va ainsi s'écrire

$$y_i = -\alpha^d p_i + x_i^d b^d + u_i^d$$

La variable de prix p_i ne peut pas être considérée comme exogène. En effet, il y a aussi une équation d'offre

$$y_i = \alpha^s p_i + x_i^s b^s + u_i^s$$

On peut résoudre ce système pour exprimer

$$p_i = \frac{1}{\alpha_s + \alpha_d}\left(x_i^d b^d - x_i^s b^s + u_i^d - u_i^s\right)$$

un choc de demande u_i^d est transmis dans les prix : $E\left(u_i^d p_i\right) \neq 0$

9.0.5 La méthode des variables instrumentales

Modèle à variables endogènes : Le modèle

$$y_i = x_i b + u_i$$

est dit à variables endogènes si on n'a pas la propriété

$$E\left(x_i^{'} u_i\right) = 0$$

Les variables x_i^k pour lesquelles $E\left(u_i x_i^k\right) \neq 0$ sont dites endogènes, les autres sont dites exogènes

Dans ce modèle

- L'estimateur des mco n'est pas convergent ;
- L'identification du modèle nécessite des hypothèses supplémentaires ;
- La méthodes des variables instrumentales est un moyen privilégié pour formuler et exploiter de telles hypothèses.

L'estimateur des mco n'est pas convergent L'estimateur des MCO de b est donné par :

$$\begin{aligned}
\widehat{b}_{mco} &= \left(\sum_{i=1}^{N} x_i' x_i\right)^{-1} \sum_{i=1}^{N} x_i' y_i = \left(\sum_{i=1}^{N} x_i' x_i\right)^{-1} \sum_{i=1}^{N} x_i'(x_i b + u_i) \\
&= b + \left(\sum_{i=1}^{N} x_i' x_i\right)^{-1} \sum_{i=1}^{N} x_i' u_i \longrightarrow b + E\left(x_i' x_i\right)^{-1} E\left(x_i' u_i\right).
\end{aligned}$$

comme $E\left(x_i' u_i\right) \neq 0$ on a $E\left(x_i' x_i\right)^{-1} E\left(x_i' u_i\right) \neq 0$ et donc

$$p\lim \widehat{b}_{mco} \neq b$$

9.1 Instruments

On considère à nouveau le modèle d'offre et de demande

$$\begin{aligned}
y_i &= -\alpha^d p_i + x_i^d b^d + u_i^d \\
y_i &= \alpha^s p_i + x_i^s b^s + u_i^s
\end{aligned}$$

On note $x_i = \left(x_i^d, x_i^s\right)$, certains élé ments peuvent être commun aux deux ensembles et n'interviennent dans ce cas qu'une fois dans x_i. On fait les hypothèses

$$E\left(x_i^{'} u_i^d\right) = 0, E\left(x_i^{'} u_i^s\right) = 0 \tag{5}$$

c.a.d les variables observables qui déplacent l'offre et la demande sont exogènes pour u_i^d et u_i^s.

On peut résoudre comme précédemment en p_i mais aussi en y_i :

$$\begin{aligned} p_i &= \frac{1}{\alpha_s+\alpha_d}\left(x_i^d b^d - x_i^s b^s + u_i^d - u_i^s\right) \\ y_i &= \frac{\alpha_s}{\alpha_s+\alpha_d}x_i^d b^d + \frac{\alpha_d}{\alpha_s+\alpha_d}x_i^s b^s + \frac{\alpha_s}{\alpha_s+\alpha_d}u_i^d + \frac{\alpha_d}{\alpha_s+\alpha_d}u_i^s \end{aligned}$$

Compte tenu des relations (5), on peut exprimer les coefficients des régressions linéaires de y_i et p_i sur x_i à partir des paramètres structurels.

La *modélisation* conduit à des *restrictions* sur les paramètres des régressions linéaires qui sont suceptibles de permettre *l'identification des paramètres structurels* du modèle.

Plus précisément :

- Si il existe une variable exogène intervenant spécifiquement dans l'équation d'offre, l'équation de demande est identifiée. Si x_{1i}^s est une telle variable, le coefficient de cette variable dans la regression linéaire de p_i sur x_i^s et x_i^d est $-\frac{1}{\alpha_s+\alpha_d}b_1^s$, et le coefficient de cette variable dans la regression linéaire de y_i sur x_i^s et x_i^d est $\frac{\alpha_d}{\alpha_s+\alpha_d}b_1^s$. La comparaison de ces deux coefficients permet l'identification de α_d
- De même, si il existe une variable exogène intervenant spécifiquement dans l'équation de demande, l'équation d'offre est identifiée.

Si on ne s'intéresse qu'à une des deux équations, p.e. l'équation de demande, les hypothèses identificatrices peuvent être assouplies. Il suffit qu'il existe au moins une variable x_{1i}^s entrant dans l'équation d'offre qui vérifie $E\left(\left[x_i^d\ x_{1i}^s\right]' u_i^d\right) = 0$. Dans ce cas les coefficients γ_y de la regressions linéaires de y_i sur $\widetilde{x}_i = \left[x_i^d\ x_{1i}^s\right]$ sont

$$\begin{aligned} \gamma_y &= E\left(\widetilde{x}_i'\widetilde{x}_i\right)^{-1} E\left(\widetilde{x}_i' y_i\right) = E\left(\widetilde{x}_i'\widetilde{x}_i\right)^{-1} E\left(\widetilde{x}_i'\left(-\alpha^d p_i + x_i^d b^d + u_i^d\right)\right) \\ &= -\alpha^d E\left(\widetilde{x}_i'\widetilde{x}_i\right)^{-1} E\left(\widetilde{x}_i' p_i\right) + E\left(\widetilde{x}_i'\widetilde{x}_i\right)^{-1} E\left(\widetilde{x}_i' x_i^d\right) b^d \\ &= -\alpha^d \gamma_p + \begin{pmatrix} b_d & 0 \end{pmatrix}' \end{aligned}$$

Dés lors que le coefficient de la variable x_{1i}^s dans la regression de la variable de prix sur $\widetilde{x}_i$, élément de γ_p, est non nul, on voit que le modèle est identifié.

Cet exemple illustre bien, la démarche des variables instrumentales. Celle-ci correspond à la mobilisation de variables exterieures au modèle et qui possèdent la particularité de n'être pas corrélées avec le résidu de l'équation.

Dire qu'une variable est une variable instrumentale revient à *postuler une relation d'exclusion* : il existe une variable affectant la variable à expliquer et la variable explicative endogène et dont tout l'effet sur la variable à expliquer transite par son effet sur la variable explicative endogène.

Une variable instrumentale ne tombe pas du ciel. Dans l'exemple on justifie le choix de la variable comme étant une variable appartenant à un modèle plus général, le système offre-demande, conduisant à l'équation structurelle de demande et à une équation réduite expliquant la formation de la variable endogène.

On considère le modèle structurel

$$y_i = x_{1i}b_1 + x_{2i}b_2 + u_i$$

les variables x_{2i}, $(dim = K_2 + 1)$ contiennent la constante et sont exogènes, mais on ne fait pas l'hypothèse d'exogénéité de la variable x_{1i} $(dim = K_1 = K - K_2)$.

On fait l'hypothèse qu'il existe un ensemble de variables dites *instrumentales* de dimension $H+1$, non parfaitement corrélées ($rangE\left(z_i^{'}z_i\right) = H+1$), car vérifiant :

$$E\left(z_i^{'}u_i\right) = 0. \tag{6}$$

Le vecteur x_{2i} fait trivialement parti de l'ensemble des variables instrumentales

L'hypothèse (6) est parfois écrite sous la forme suivante :

$$E(u_i|z_i) = 0$$

9.1.1 Identification

La condition (6) peut être réécrite comme suit :

$$E\left(z_i^{'}(y_i - x_i b)\right) = 0$$

Soit encore :

$$E\left(z_i^{'}y_i\right) = E\left(z_i^{'}x_i\right) b \tag{7}$$

Cette condition définit un système de $H+1$ équations à $K+1$ inconnues b.

Le modèle est identifié si le système (7) admet pour unique solution le paramètre structurel b

On distingue trois situations

- Si $H < K$, le modèle est sous identifié, puisqu'il y a moins d'équations que de variables. Il n'y a pas suffisamment de variables instrumentales.
- Si $H = K$ et $\lim rangE\left(z_i^{'}x_i\right) = K+1$ le modèle est juste identifié.
- Si $H > K$, $\lim rangE\left(z_i^{'}x_i\right) = K+1$ le modèle est dit sur-identifié. Dans ce cas il y a plus de variables instrumentales qu'il n'est nécessaire.

9.2 Moindres carrés indirects

Si $H = K$ et si $Ez_i^{'}x_i$ est inversible, alors $b = E\left(z_i^{'}x_i\right)^{-1} E\left(z_i^{'}y_i\right)$. On obtient un estimateur de b appelé Estimateur des Moindres Carrés Indirects en remplaçant les espérances par leurs contreparties empiriques :

$$\begin{aligned} \widehat{b}_{mci} &= \left(\frac{1}{N}\sum_{i=1}^{N} z_i^{'}x_i\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} z_i^{'}y_i \\ &= (Z'X)^{-1}Z'Y \end{aligned}$$

où Z est la matrice dont la i-ième ligne est z_i, X la matrice dont la i-ième ligne est x_i et Y le vecteur dont la i -ième composante est y_i.

Si $H > K$, on se ramène au cas précédent en sélectionnant $K+1$ combinaisons linéaires des instruments : Az_i, où A est une matrice $K+1 \times H+1$, de rang $K+1$. L'hypothèse que l'ensemble des $H+1$ variables dans z_i est un ensemble de variables instrumentales conduit à la propriété que pour A tel que $AE\left(z_i^{'}x_i\right)$ est inversible,

$$b = \left(AE\left(z_i^{'}x_i\right)\right)^{-1} AE\left(z_i^{'}y_i\right).$$

On en déduit une classe d'estimateur :

$$\begin{aligned}\widehat{b}_{mci}(A) &= \left(A\overline{z_i'x_i}\right)^{-1} A\overline{z_i'y_i} \\ &= (AZ'X)^{-1}AZ'Y.\end{aligned}$$

9.2.1 Propriété asymptotiques des estimateurs des MCI

Dans le modèle

$$y_i = x_i b + u_i$$

à $K+1$ variables explicatives.

Sous les hypothèses

Hypothèse (H_1). $E\left(z_i'u_i\right) = 0$ avec z_i de dim $1 \times H+1$

Hypothèse (H_2). Les observations (x_i, z_i, y_i) sont iid

Hypothèse (H_3). $E(u_i^2|z_i) = \sigma^2$

Hypothèse (H_4). Les moments de (x_i, z_i, y_i) existent jusqu'à un ordre suffisant

Hypothèse (H_5). $E\left(z_i'x_i\right)$ et $\overline{z_i'x_i}$ sont de rang $K+1$

Théorème 9.1. *Sous ces hypothèses, il existe au moins une matrice A de dimension $K+1 \times H+1$ pour laquelle l'estimateur $\widehat{b}_{mci}(A) = \left(A\overline{z_i'x_i}\right)^{-1} A\overline{z_i'y_i}$ existe, et pour toute matrice A telle que l'estimateur des MCI existe, on a :*

- *$\widehat{b}_{mci}(A)$ est convergent : $p\lim \widehat{b}_{mci}(A) = b$*
- *$\widehat{b}_{mci}(A)$ est asymptotiquement normal :*

$$\sqrt{N}\left(\widehat{b}_{mci}(A) - b\right) \xrightarrow{L} N(0, \Sigma(A)),$$

avec

$$\Sigma(A) = \sigma^2 \left[AE\left(z_i'x_i\right)\right]^{-1} AE\left(z_i'z_i\right)A' \left[E\left(x_i'z_i\right)A'\right]^{-1}$$

- *$\widehat{\Sigma}(A) = \widehat{\sigma}^2 \left[A\overline{z_i'x_i}\right]^{-1} A\overline{z_i'z_i}A' \left[\overline{x_i'z_i}A'\right]^{-1}$ où $\widehat{\sigma}^2 = \overline{\widehat{u}\left(A\right)_i^2}$, est un estimateur convergent de $\Sigma(A)$*

Démonstration.

- *Existence* d'au moins un estimateur des MCI : Il suffit de prendre $A = E\left(z_i'x_i\right)'$ on a alors $E\left(z_i'x_i\right)'\overline{z_i'x_i} \to E\left(z_i'x_i\right)'E\left(z_i'x_i\right)$ qui est inversible puisque $rang E\left(z_i'x_i\right) = K+1$ Comme le déterminant est une fonction continue $\det A\overline{z_i'x_i} \to \det AA' \neq 0$ et donc la matrice $A\overline{z_i'x_i}$ est inversible pour N assez grand.
- *Convergence* :

$$\widehat{b}_{mci}(A) = \left(A\overline{z_i'x_i}\right)^{-1} A\overline{z_i'y_i} = b + \widehat{b}_{mci}(A) = b + \left(A\overline{z_i'x_i}\right)^{-1} A\overline{z_i'u_i}.$$

La convergence découle simplement de la loi des grands nombres :

$$\overline{z_i'u_{ii}} \longrightarrow E\left(z_i'u_i\right) = 0.$$

– *Normalité asymptotique*

$$\sqrt{N}\left(\widehat{b}_{mci}(A)-b\right)=\left(A\overline{z_i'x_i}\right)^{-1}A\sqrt{N}\overline{z_i'u_i}$$

Comme $V\left(z_i'u_i\right)=E(z_i'z_iu_i^2)=E\left[z_i'z_iE(u_i^2\mid z_i)\right]=\sigma^2E\left(z_i'z_i\right)$, la normalité asymptotique découle directement du théorème cental-limite :

$$\sqrt{N}\overline{z_i'u_i}\overset{loi}{\to}N(0,\sigma^2Ez_iz_i')$$

et $\left(A\overline{z_i'x_i}\right)^{-1}A\to\left(AE\left(z_i'x_i\right)\right)^{-1}A$

– *Estimation de la matrice de variance-covariance asymptotique*
Comme pour l'estimateur des mco, on vérifie facilement que $\overline{\widehat{u}\left(A\right)_i^2}=\overline{\left(u_i+x_i\left(b-\widehat{b}\left(A\right)\right)\right)^2}\to\sigma^2$ puisque $b-\widehat{b}\left(A\right)\to0$

■

9.2.2 Estimation robuste de la matrice de variance

Comme pour l'estimateur des mco, il existe une version de la matrice de variance-covariance $\Sigma(A)$ pour le cas de résidus hétéroscédastiques, i.e. lorsque $E(u_i^2|z_i)$ dépend de z_i. On peut donc supprimer l'hypothèse H_3. Les conclusions sont simplement modifiées en :

– $\widehat{b}_{mci}(A)$ est asymptotiquement normal :

$$\sqrt{N}\left(\widehat{b}_{mci}(A)-b\right)\overset{L}{\to}N(0,\Sigma_{het}(A)),$$

avec

$$\Sigma_{het}(A)=\left[AE\left(z_i'x_i\right)\right]^{-1}AE\left(u_i^2z_i'z_i\right)A'\left[E\left(x_i'z_i\right)A'\right]^{-1}$$

– $\widehat{\Sigma}_{het}(A)=\left[A\overline{z_i'x_i}\right]^{-1}A\overline{\widehat{u}\left(A\right)_i^2z_i'z_i}A'\left[\overline{x_i'z_i}A'\right]^{-1}$

9.2.3 Estimateur à variables instrumentales optimal ou estimateur des doubles moindres carrés

Théorème 9.2. *Il existe une matrice A^* optimale au sens où pour toute suite de matrice $A_N\to A^*$, la variance asymptotique de $\widehat{b}_{mci}(A_N)$ est de variance minimale dans la classe des estimateurs $\widehat{b}_{mci}(A)$.Cette matrice a pour expression :*

$$A^*=E\left(x_i'z_i\right)E\left(z_i'z_i\right)^{-1}$$

La matrice de variance correspondante a pour expression

$$\Sigma(A^*)=\sigma^2\left[E\left(x_i'z_i\right)E\left(z_i'z_i\right)^{-1}E\left(z_i'x_i\right)\right]^{-1}$$

qui s'obtient directement en remplaçant A par $E\left(x_i'z_i\right)E\left(z_i'z_i\right)^{-1}$

$$\Sigma(A)=\sigma^2\left[AE\left(z_i'x_i\right)\right]^{-1}AE\left(z_i'z_i\right)A'\left[E\left(x_i'z_i\right)A'\right]^{-1}$$

et en opérant des simplifications.

Demonstration de l'optimalité.
Pour montrer que $\Sigma(A) \geqslant \Sigma(A^*)$ au sens des matrices, i.e.

$$\forall \lambda, \lambda^{'} \left(\Sigma(A) - \Sigma(A^*)\right) \lambda \geqslant 0$$

On peut clairement laisser tomber le facteur σ^2. La matrice de variance $\Sigma(A^*)$ s'écrit :

$$\Sigma(A^*) = \left[E\left(x_i^{'} z_i\right) E\left(z_i' z_i\right)^{-1} E\left(z_i^{'} x_i\right)\right]^{-1} = (C'C)^{-1}$$

avec $C = E\left(z_i' z_i\right)^{-1/2} E\left(z_i^{'} x_i\right)$ de dim $H+1 \times K+1$.La matrice $\Sigma(A)$ s'écrit :

$$\Sigma(A) = \left[AE\left(z_i^{'} x_i\right)\right]^{-1} AE\left(z_i' z_i\right) A' \left[E\left(x_i^{'} z_i\right) A'\right]^{-1} = BB'$$

avec $B = \left[AE\left(z_i^{'} x_i\right)\right]^{-1} AE\left(z_i' z_i\right)^{1/2}$ de dim $K+1 \times H+1$.On a la relation

$$\begin{aligned} BC &= \left[AE\left(z_i^{'} x_i\right)\right]^{-1} AE\left(z_i' z_i\right)^{1/2} E\left(z_i' z_i\right)^{-1/2} E\left(z_i^{'} x_i\right) \\ &= \left[AE\left(z_i^{'} x_i\right)\right]^{-1} AE\left(z_i^{'} x_i\right) = I_{K+1} \end{aligned}$$

On a donc

$$\Sigma(A) - \Sigma(A^*) = BB' - (C'C)^{-1} = BB' - BC\left(C'C\right)^{-1} C'B'$$

puisque $BC = I$. On a donc :

$$\Sigma(A) - \Sigma(A^*) = B\left[I - C\left(C'C\right)^{-1} C'\right] B'$$

Comme $I - C\left(C'C\right)^{-1} C'$ est une matrice semi definie positive, $\Sigma(A) - \Sigma(A^*)$ est aussi une matrice semi définie positive. ■

9.2.4 Expression de l'estimateur optimal

La matrice $A^* = E\left(x_i^{'} z_i\right) E\left(z_i' z_i\right)^{-1}$ est inconnue. Pour mettre l' estimateur en oeuvre, on la remplace par un estimateur convergent $A_N = \overline{x_i^{'} z_i}\ \ \overline{z_i' z_i}^{-1}$

$$\begin{aligned} \widehat{b}_{mci}(A_N) &= \left(\overline{x_i^{'} z_i}\ \ \overline{z_i' z_i}^{-1} \overline{z_i^{'} x_i}\right)^{-1} \overline{x_i^{'} z_i}\ \ \overline{z_i' z_i}^{-1} \overline{z_i^{'} y_i} \\ &= \left(X'Z\left(Z'Z\right)^{-1} Z'X\right)^{-1} X'Z\left(Z'Z\right)^{-1} Z'Y \end{aligned}$$

Cet estimateur a les mêmes propriétés asymptotiques que l'estimateur $\widehat{b}_{mci}(A^*)$ puisque $A_N \to A^*$.

On peut réécrire l'estimateur en faisant intervenir la matrice de projection orthogonale sur Z, $P_Z = Z\left(Z'Z\right)^{-1} Z'$

$$\widehat{b}_{mci}(A^*) = (X'P_Z X)^{-1} X'P_Z Y = \left(\left(P_Z X\right)' P_Z X\right)^{-1} \left(P_Z X\right)' Y$$

Il correspond à l'estimateur des mco de la variable endogène Y sur la projection $\widehat{X} = P_Z X$ des variables explicatives sur l'ensemble des instruments. C'est

pourquoi on appelle cet estimateur *estimateur des doubles moindres carrés* et on le note $\widehat{b}_{2mc}$.

Il résulte d'une première régression par les mco des variables explicatives X sur l'ensemble des instruments, permettant de déterminer les prédictions $\widehat{X} = P_Z X = Z\left((Z'Z)^{-1} Z'X\right)$ des X par les instruments puis d'une seconde régression par les mco de la variable à expliquer sur les prédictions $\widehat{X}$.

La matrice de variance asymptotique de $\widehat{b}_{2mc}$ est

$$\Sigma(\widehat{b}_{2mc}) = \sigma^2 \left[E\left(x_i' z_i\right) E\left(z_i' z_i\right)^{-1} E\left(z_i' x_i\right)\right]^{-1}$$

et la matrice de variance de l'estimateur dans un échantillon de taille N est

$$V(\widehat{b}_{2mc}) = \Sigma(\widehat{b}_{2mc})/N = \sigma^2 \left[E\left(x_i' z_i\right) E\left(z_i' z_i\right)^{-1} E\left(z_i' x_i\right)\right]^{-1}/N$$

On peut l'estimer par

$$\widehat{V}(\widehat{b}_{2mc}) = \widehat{\sigma}^2 \left(X'Z\left(Z'Z\right)^{-1} Z'X\right)^{-1} = \widehat{\sigma}^2 (X' P_Z X)^{-1} = \widehat{\sigma}^2 \left(\widehat{X}'\widehat{X}\right)^{-1}$$

L'écart-type des résidus à retenir est celui du modèle

$$y_i = x_i b + u_i$$

il peut être estimé par $\overline{\left(y_i - x_i \widehat{b}_{2mc}\right)^2}$.

9.2.5 Cas des résidus hétéroscédastiques

Dans ce cas l'estimateur des doubles moindres carrés n'est plus optimal, et la formule de sa variance n'est plus correcte.

La formule exacte est donnée comme dans le cas général par

$$\begin{aligned}
\Sigma_{het}(A^*) &= \left[A^* E\left(z_i' x_i\right)\right]^{-1} A^* E\left(u_i^2 z_i' z_i\right) A^{*\prime} \left[E\left(x_i' z_i\right) A^{*\prime}\right]^{-1} \\
&= \left[E\left(x_i' z_i\right) E\left(z_i' z_i\right)^{-1} E\left(z_i' x_i\right)\right]^{-1} E\left(x_i' z_i\right) E\left(z_i' z_i\right)^{-1} \\
&\quad E\left(u_i^2 z_i' z_i\right) E\left(z_i' z_i\right)^{-1} E\left(z_i' x_i\right) \left[E\left(x_i' z_i\right) E\left(z_i' z_i\right)^{-1} E\left(z_i' x_i\right)\right]^{-1} \\
&= E\left(\widetilde{x}_i' \widetilde{x}_i\right)^{-1} E\left(u_i^2 \widetilde{x}_i' \widetilde{x}_i\right) E\left(\widetilde{x}_i' \widetilde{x}_i\right)^{-1}
\end{aligned}$$

où $\widetilde{x}_i = z_i E\left(z_i' z_i\right)^{-1} E\left(z_i' x_i\right)$.

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

$$V_{het}\left(\widehat{b}_{2mc}\right) = \Sigma_{het}(A^*)/N$$

Elle peut être estimée par

$$\begin{aligned}
\widehat{V}_{het}\left(\widehat{b}_{2mc}\right) &= \frac{\widehat{\Sigma}_{het}(A^*)}{N} = \left(\overline{\widehat{\widetilde{x}}_i \widehat{\widetilde{x}}_i'}\right)^{-1} \left(\sum_{i=1}^{N} \widehat{u}_i^2 \widehat{\widetilde{x}}_i \widehat{\widetilde{x}}_i'\right) \left(\sum_{i=1}^{N} \widehat{\widetilde{x}}_i \widehat{\widetilde{x}}_i'\right)^{-1} \\
&= (\widehat{X}'\widehat{X})^{-1} \left(\widehat{X}' \lim diag[\widehat{u}_i^2] \widehat{X}\right) (\widehat{X}'\widehat{X})^{-1},
\end{aligned}$$

qui est exactement la matrice de White.

9.2.6 Interprétation de la condition $\text{rang}E\left(z_i'x_i\right) = K+1$

La mise en oeuvre de la méthode des variables instrumentales repose sur la condition $\text{rang}E\left(z_i'x_i\right) = K+1$. Les variables du modèle sont scindées en K_1 variables endogènes x_{1i} et K_2+1 variables esxogènes. Ces variables interviennent également dans la liste des instruments qui contient en outre $H-K_2$ variables extérieures $\widetilde{z}_i$: $z_i = \left[\begin{array}{cc} \widetilde{z}_i & x_{2i} \end{array}\right]$. Compte tenu de l'hypothèse $E\left(z_i'z_i\right)$ inversible, la condition $\text{rang}E\left(z_i'x_i\right) = K+1$ est analogue à la condition $\text{rang}E\left(z_i'z_i\right)^{-1}E\left(z_i'x_i\right) = K+1$. Cette matrice correspond à la matrice des coefficients des régressions des variables explicatives sur les instruments. Comme les variables du modèle et les instrument ont les variables x_2 en commun, on a :

$$\begin{aligned} E\left(z_i'z_i\right)^{-1}E\left(z_i'x_i\right) &= \left[\begin{array}{cc} E\left(z_i'z_i\right)^{-1}E\left(z_i'x_{1i}\right) & 0 \\ & I_{K_2+1} \end{array}\right] \\ &= \left[\begin{array}{cc} \Gamma_{1\widetilde{z}} & 0 \\ \Gamma_{1x_2} & I_{K_2+1} \end{array}\right] \end{aligned}$$

où $\Gamma_{1\widetilde{z}}$ et Γ_{1x_2} sont les coefficients de $\widetilde{z}$ et x_2 des régressions des variables endogènes sur les instruments. La condition $\text{rang}E\left(z_i'z_i\right)^{-1}E\left(z_i'x_i\right) = K+1$ est donc équivalente à la condition

$$\text{rang}\Gamma_{1\widetilde{z}} = K_1$$

Cette condition s'interprète comme le fait que les variables instrumentales extérieures expliquent sufisamment bien les variables endogènes. Il n'existe pas de test formel de cette condition. Néanmoins il est important de regarder la façon dont les variables instrumentales expliquent les variables endogènes. On peut par exemple, bien que cela ne garantisse pas que la condition est satisfaite dés qu'il y a plus d'une variable endogène, effectuer chaque régression des variables endogènes sur l'ensemble des variables instrumentales et faire un test de la nullité globale des coefficients des variables instrumentales extérieures.

Dans le cas où la condition $\text{rang}E\left(z_i'x_i\right) = K+1$ n'est pas satisfaite, on aura néanmoins en général à distance finie $\text{rang}z_i'x_i = K+1$ et l'estimateur pourra être numériquement mis en oeuvre. La conséquence du fait que $\text{rang}E\left(z_i'x_i\right) < K+1$ est que

$$X'Z\left(Z'Z\right)^{-1}Z'X \to E\left(x_i'z_i\right)E\left(z_i'z_i\right)E\left(z_i'x_i\right)$$

non inversible. L'estimateur sera donc trés instable et présentera des écart-types trés élevés sur certains coefficients, à l'instar de ce qui se produit avec les mco dans le cas de multicolinéarité.

9.2.7 Test de suridentification

Lorsqu'il y a plus d'instruments que de variables explicatives le modèle est suridentifié. On a vu que dans le modèle

$$y_i = x_i b + u_i$$

avec pour restriction identifiante

$$E\left(z_i^{'}u_i\right)=0,$$

on pouvait estimer le modèle par les MCI de trés nombreuses façons, l'estimateur le plus performant étant celui des doubles moindres carrés. On avait

$$\widehat{b}_{mci}\left(A\right)=\left(A\overline{z_i'x_i}\right)^{-1}A\overline{z_i'y_i}$$

contrepartie empirique de la relation

$$b=\left(AE\left(z_i'x_i\right)\right)^{-1}AE\left(z_i'y_i\right)$$

Cette dernière relation doit être vraie pour toute matrice A telle que $AE\left(z_i'x_i\right)$ est inversible. Elle montre bien que le modèle impose plus de structure entre les données qu'il n'est nécessaire pour identifier le modèle : Tous les paramètres $\widehat{b}_{mci}\left(A\right)$ doivent converger vers une même valeur.

Par exemple dans le cas où il y a une variable endogène et où en plus des variables exogènes du modèle on a mobilisé h variables instrumentales extérieures au modèle, les h estimateurs que l'on peut obtenir en choisissant comme vecteur de variables instrumentales les exogènes du modèle et l'une des variables instrumentales extérieures doivent être proches. En pratique, on est souvent amené à effectuer des estimation d'une même équation en étendant ou restreignant la liste des variables instrumentales.

Pour rendre cette démarche plus transparente, il est utile d'avoir une procédure qui permette de tester l'hypothèse que pour un jeu de variables instrumentales donné l'ensemble des estimateurs $\widehat{b}_{mci}\left(A\right)$ convergent tous vers la même valeur.

On peut considéré le test de l'hypothèse nulle

$H_0 : E\left(z_i'u_i\right)=0$

On considère le cas standard dans lequel les résidus sont homoscédastiques.

Si le résidu était connu un tel test serait trés facile à mettre en oeuvre. Il consisterait simplement à regarder si la moyenne empirique $\overline{z_i'u_i}$ de $z_i'u_i$ est proche de zéro, c'est à dire si la norme de ce vecteur est proche de zéro.

On rappelle le résultat suivant

$$W\rightsquigarrow N\left(0,V\left(W\right)\right)\Rightarrow W'V\left(W\right)^{-}W'\rightsquigarrow\chi^2\left(\text{rang}\left(V\left(W\right)\right)\right)$$

où $V\left(W\right)^{-}$ est un inverse généralisé de la matrice $V\left(W\right)$, i.e tel que

$$V\left(W\right)V\left(W\right)^{-}V\left(W\right)=V\left(W\right)$$

Sous l'hypothèse H_0 on aurait donc en appliquant le téorème central-limite, et compte tenu de l'hypothèse d'homoscédasticité

$$\sqrt{N}\overline{z_i'u_i}\rightarrow N\left(0,\sigma^2E\left(z_i^{'}z_i\right)\right)$$

et donc

$$\frac{N}{\sigma^2}\overline{z_i'u_i}^{'}E\left(z_i^{'}z_i\right)^{-1}\overline{z_i'u_i}\rightarrow\chi^2\left(\dim\left(z_i\right)\right)$$

ou encore

$$\frac{N}{\widehat{\sigma}^2}\overline{z_i'u_i}^{'}\overline{z_i^{'}z_i}^{-1}\overline{z_i'u_i}\rightarrow\chi^2\left(\dim\left(z_i\right)\right)$$

Le problème vient ici du fait que l'on n'observe pas u_i. On est en revanche capable de déterminer $\widehat{u}_i = y_i - x_i\widehat{b}_{2mc}$. Le test que l'on met en oeuvre est donc basé sur $\overline{z_i'\widehat{u}_i}$.

Détermination de la matrice de variance de $\overline{z_i'\widehat{u}_i}$

On ne peut pas transposer directement le test, il faut calculer la matrice de variance de $\overline{z_i'\widehat{u}_i}$

On a

$$\widehat{u}_i = y_i - x_i\widehat{b}_{2mc} = x_ib + u_i - x_i\widehat{b}_{2mc} = u_i - x_i\left(\widehat{b}_{2mc} - b\right)$$

d'où

$$\overline{z_i'\widehat{u}_i} = \frac{1}{N}Z'\widehat{U} = \frac{1}{N}\left(Z'U - Z'X\left(\widehat{b}_{2mc} - b\right)\right)$$

comme $\widehat{b}_{2mc} = (\widehat{X}'\widehat{X})^{-1}\widehat{X}'Y = b + (\widehat{X}'\widehat{X})^{-1}\widehat{X}'U$, avec $\widehat{X} = P_ZX$, la projection orthogonale de X sur Z, on a :

$$\overline{z_i'\widehat{u}_i} = \frac{1}{N}\left(Z'U - Z'X(\widehat{X}'\widehat{X})^{-1}\widehat{X}'U\right)$$

en outre $X = P_ZX + (I - P_Z)X = \widehat{X} + (I - P_Z)X$ et donc $Z'X = Z'\widehat{X}$. Finalement

$$\begin{aligned}\overline{z_i'\widehat{u}_i} &= \frac{1}{N}\left(Z'U - Z'\widehat{X}(\widehat{X}'\widehat{X})^{-1}\widehat{X}'U\right) = \frac{1}{N}\left(Z'U - Z'P_{\widehat{X}}U\right)\\ &= \frac{1}{N}Z'\left(I_N - P_{\widehat{X}}\right)U\end{aligned}$$

On en déduit que

$$V\left(\overline{z_i'\widehat{u}_i}\right) = \frac{\sigma^2}{N^2}Z'\left(I_N - P_{\widehat{X}}\right)Z = \frac{\sigma^2}{N^2}\left(\left(I_N - P_{\widehat{X}}\right)Z\right)'\left(I_N - P_{\widehat{X}}\right)Z$$

Détermination du rang de la matrice $V\left(\overline{z_i'\widehat{u}_i}\right)$

Le vecteur $\left(I_N - P_{\widehat{X}}\right)Z$ est le résidu de la projection de Z sur $\widehat{X}$. Comme $\widehat{X}$ est la projection de X sur Z l'espace vectoriel engendré par les colonnes de $\widehat{X}$ de dimension $K+1$ est inclus dans celui engendré par les colonnes de Z de dimension $H+1$. La matrice $\left(I_N - P_{\widehat{X}}\right)Z$ est donc de rang $H-K$. Il en résulte que :

$$\text{rang}V\left(\overline{z_i'\widehat{u}_i}\right) = H - K$$

Inverse généralisé de la matrice $V\left(\overline{z_i'\widehat{u}_i}\right)$

La matrice n'est pas inversible, pour mettre le test en oeuvre en déterminer un inverse généralisé. L'un d'entre eux est

$$V\left(\overline{z_i'\widehat{u}_i}\right)^- = \frac{N^2}{\sigma^2}(Z'Z)^{-1}$$

En effet , la matrice de variance s'écrit de façon alternative comme $\frac{\sigma^2}{N^2}Z'\left(P_Z - P_{\widehat{X}}\right)Z$, et on a

$$\begin{aligned}&\frac{\sigma^2}{N^2}Z'\left(P_Z - P_{\widehat{X}}\right)Z\frac{N^2}{\sigma^2}(Z'Z)^{-1}\frac{\sigma^2}{N^2}Z'\left(P_Z - P_{\widehat{X}}\right)Z\\ =\ &\frac{\sigma^2}{N^2}Z'\left(P_Z - P_{\widehat{X}}\right)P_Z\left(P_Z - P_{\widehat{X}}\right)Z\end{aligned}$$

le résultat découle du fait que $P_{\widehat{X}}P_Z = P_Z P_{\widehat{X}} = P_{\widehat{X}}$ et que donc

$$\begin{aligned}\left(P_Z - P_{\widehat{X}}\right) P_Z \left(P_Z - P_{\widehat{X}}\right) &= \left(P_Z - P_{\widehat{X}}\right)\left(P_Z - P_Z P_{\widehat{X}}\right) \\ &= \left(P_Z - P_Z P_{\widehat{X}}\right) - P_{\widehat{X}}P_Z - P_{\widehat{X}}P_Z P_{\widehat{X}} \\ &= \left(P_Z - P_{\widehat{X}}\right)\end{aligned}$$

Le test et son interprétation

Finalement, sous l'hypothèse $H_0 : E\left(z_i' u_i\right) = 0$, on a

$$\begin{aligned}\widehat{S} &= \overline{z_i'\widehat{u}_i}' V \left(\overline{z_i'\widehat{u}_i}\right)^- \overline{z_i'\widehat{u}_i} = \frac{1}{N}\widehat{U}' Z \frac{N^2}{\sigma^2}\left(Z'Z\right)^{-1}\frac{1}{N}Z'\widehat{U} \\ &= \frac{1}{\sigma^2}\widehat{U}' P_Z \widehat{U} \approx N\frac{\widehat{U}' P_Z \widehat{U}}{\widehat{U}'\widehat{U}} \rightsquigarrow \chi^2\left(H - K\right)\end{aligned}$$

Sous l'hypothèse alternative, on a

$$\widehat{u}_i = y_i - x_i\widehat{b}_{2mc} = x_i b + u_i - x_i\widehat{b}_{2mc} = u_i - x_i\left(\widehat{b}_{2mc} - b\right)$$

d'où

$$\overline{z_i'\widehat{u}_i} = \overline{z_i'u_i} - \overline{z_i'x_i}\left(A^*\overline{z_i'x_i}\right)^{-1} A^*\overline{z_i'u_i} = \overline{z_i'u_i} - \overline{z_i'x_i}\left(A^*\overline{z_i'x_i}\right)^{-1} A^*\overline{z_i'u_i}$$

où $A^* = E\left(x_i' z_i\right) E\left(z_i' z_i\right)^{-1}$

Comme $\overline{z_i'u_i}$ ne converge plus vers zéro, cette quantité va converger vers une limite non nulle en général, mais pas toujours. On peut se trouver dans la situation dans laquelle

$$\overline{z_i'u_i} = \overline{z_i'x_i}\left(A^*\overline{z_i'x_i}\right)^{-1} A^*\overline{z_i'u_i}$$

soit

$$\overline{z_i'\left(u_i - x_i\left(A^*\overline{z_i'x_i}\right)^{-1} A^*\overline{z_i'u_i}\right)} = 0$$

soit encore

$$\overline{z_i'\left(y_i - x_i\left(A^*\overline{z_i'x_i}\right)^{-1} A^*\overline{z_i'y_i}\right)} = 0$$

ce qui signifie que le résidu de la régression de y_i sur x_i par les doubles moindres carré peut être orthogonal à z_i, alors qu'on n'a pas $E\left(z_i'u_i\right) = 0$.

Ceci provient du fait que le test que l'on met en oeuvre n'est pas un test de la validité des instruments dans le modèle structurel

$$y_i = x_i b + u_i$$

c'est à dire le test de l'hypothèse

$$E\left(z_i'\left(y_i - x_i b\right)\right) = 0$$

mais le test d'une hypothèse moins forte :

$$\exists \;\; c \;\; tq \;\; Ez_i'\left(y_i - x_i c\right) = 0$$

Pour cette hypothèse nulle, sous H_0 la statistique converge vers la loi qu'on a déterminé, et sous l'hypothèse alternative, elle tend vers $+\infty$.

Résultat :

Sous l'hypothèse nulle

$H_0 : \exists \ \ c \ \ tq \ \ Ez_i' (y_i - x_i c) = 0$, la statistique

$$\widehat{S} = N \frac{\widehat{U}' P_Z \widehat{U}}{\widehat{U}' \widehat{U}} \xrightarrow{L} \chi^2 (H - K)$$

Sous l'hypothèse alternative $\widehat{S} \to +\infty$

Le test est donc un test convergent. Pour un test au niveau α, la r égion critique est $W_\alpha = \left[Q_{1-\alpha} \left(\chi^2 (H-K) \right), +\infty \right[$, où $Q_{1-\alpha} \left(\chi^2 (H-K) \right)$ est le quantile d'ordre $1-\alpha$ d'une loi du χ^2 à $H-K$ degrés de liberté.

Mise en oeuvre du test. Le test de suridentification est trés simple à mettre en oeuvre. Il correspond au test de la nullité globale des coefficients de la régression de $\widehat{u}_i$ sur les variables instrumentales, y compris la constante. En pratique on applique les doubles moindres carrés, on construit les résidus estimés et on les régressent sur les variables instrumentales. La statistique de test est NR^2 de cette régression.

Remarque.

– On a a priori toujours intérêt à avoir un ensemble d'instruments le plus large possible. En effet retirer une variable instrumentale et mettre en oeuvre l'estimateur des doubles moindres carrés correspond à sélectionner une matrice particulière pour l'estimateur des moindres carrés indirects avec le jeu complet d'instruments. Comme on l'a montré cet estimateur est alors nécésairement moins ou aussi bon que l'estimateur des doubles moindres carrés avec l'ensemble d'instruments complet. Quand on étend l'ensemble des variables instrumentales, il est important de bien vérifier la compatibilité globale des instruments utilisés et de mettre en oeuvre le test de suridentification.

– La matrice de variance de l'estimateur des doubles moindres carrés est toujours plus grande que celle de l'estimateur des mco. Ceci se voit immédiatement en examinant l'expression des variances

$$V (b_{mco}) = \sigma^2 (X'X)^{-1} \text{ et } V (b_{2mc}) = \sigma^2 (X' P_Z X)^{-1}$$

En outre, on voit aussi en comparant les expressions des estimateurs

$$b_{mco} = (X'X)^{-1} X' Y \text{ et } b_{2mc} = (X' P_Z X)^{-1} X' P_Z Y$$

que lorsque l'on étend la liste des variables instrumentales la dimension de l'espace sur lequel on projette les variables du modèle augmente et qu'on en a donc une représentation de plus en plus fidèle. La variance de l'estimateur des doubles moindres carrés va s'améliorer, mais l'estimateur des doubles moindres carrés va se rapprocher de l'estimateur des moindres carrés ordinaires. Il y a donc un risque à étendre trop la liste des instruments. A distance finie, on pourrait avoir une mise en oeuvre fallacieuse conduisant à un estimateur proche de celui des mco. Il est utile pour se prémunir de ce risque de regarder la régression des variables endogènes sur les instruments et de contrôler la significativité globales des instruments.

9.2.8 Test d'exogénéité des variables explicatives

Ayant estimé le modèle par les double moindre carrés, c'est à dire sous l'hypothèse

$$H_1 : \exists c / E\left(z_i^{'}\left(y_i - x_i c\right)\right) = 0$$

On peut vouloir tester l'hypothèse que les régresseurs x_i sont exogènes.
On considère donc l'hypothèse

$$H_0 : \exists c / E\left(z_i^{'}\left(y_i - x_i c\right)\right) = 0 \text{et} E\left(x_i^{'}\left(y_i - x_i c\right)\right) = 0.$$

L'intérêt de tester une telle hypothèse est immédiat compte tenu du fait que sous cette hypothèse l'estimateur optimal sera l'estimateur des mco qui domine n'importe quel estimateur à variables instrumentales.

Un test naturel d'exogénéité est le test d'Hausman fondé sur la comparaison de $\widehat{b}_{2mc} - \widehat{b}_{mco}$ avec 0.

Le test peut être fondé sur les coefficients des endogènes

En effet $\widehat{b}_{2mc} = \left(\widehat{X}^{'}\widehat{X}\right)^{-1}\widehat{X}^{'}Y$ et $\widehat{b}_{mco} = \left(X^{'}X\right)^{-1}X^{'}Y$ donc

$$\begin{aligned}\widehat{X}^{'}\widehat{X}\left(\widehat{b}_{2mc} - \widehat{b}_{mco}\right) &= \widehat{X}^{'}\widehat{X}\left[\left(\widehat{X}^{'}\widehat{X}\right)^{-1}\widehat{X}^{'}Y - \left(X^{'}X\right)^{-1}X^{'}Y\right] \\ &= \left[\widehat{X}^{'}Y - \widehat{X}^{'}\widehat{X}\left(X^{'}X\right)^{-1}X^{'}Y\right]\end{aligned}$$

Comme $\widehat{X}^{'}\widehat{X} = \widehat{X}^{'}X$ puisque $X = P_Z X + (I - P_Z)X = \widehat{X} + (I - P_Z)X$

$$\widehat{X}^{'}\widehat{X}\left(\widehat{b}_{2mc} - \widehat{b}_{mco}\right) = \widehat{X}^{'}M_X Y = \begin{pmatrix}\widehat{X}_1^{'}M_X Y \\ 0\end{pmatrix}$$

On en déduit que

$$\left(\widehat{b}_{2mc}^{(2)} - \widehat{b}_{mco}^{(2)}\right) = \left(\widehat{X}^{'}\widehat{X}\right)^{21}\left(\left(\widehat{X}^{'}\widehat{X}\right)^{11}\right)^{-1}\left(\widehat{b}_{2mc}^{(1)} - \widehat{b}_{mco}^{(1)}\right)$$

avec $b^{(1)}$ le vecteurs des coefficients de x_{1i} et sym étriquement pour $b^{(2)}$, et les notations standards

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22}\end{bmatrix}^{-1} = \begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22}\end{bmatrix}$$

On peut donc se contenter de se fonder sur

$$\widehat{b}_{2mc}^{(1)} - \widehat{b}_{mco}^{(1)} = \widehat{X}^{'}\widehat{X}^{11}\widehat{X}_1^{'}M_X Y$$

pour effectuer le test.

Rang de la matrice de variance de $\widehat{b}_{2mc}^{(1)} - \widehat{b}_{mco}^{(1)}$

L'expression précédente montre que la matrice de variance de $\widehat{b}_{2mc}^{(1)} - \widehat{b}_{mco}^{(1)}$ est $\sigma^2 = \widehat{X}^{'}\widehat{X}^{11}\widehat{X}_1^{'}M_X\widehat{X}_1\widehat{X}^{'}\widehat{X}^{11}$. Son rang est donc égal à celui de $\widehat{X}_1^{'}M_X\widehat{X}_1$, donc à celui de $M_X\widehat{X}_1$. Supposons que l'on ait pour un vecteur λ $M_X\widehat{X}_1\lambda = 0$ alors $P_X\widehat{X}_1\lambda = \widehat{X}_1\lambda$ il existe donc un vecteur μ tel que $\widehat{X}_1\lambda = X\mu$. Comme $\widehat{X}_1$

appartient à l'espace engendré par $Z = \left[\widetilde{Z},\ X_2\right]$, nécéssairement $X\mu = X_2\mu_2$. Notant comme précédemment où $\Gamma_{1\widetilde{z}}$ et Γ_{1x_2} les coefficients de $\widetilde{z}$ et x_2 des r égressions des variables endogènes sur les instruments. L'é quation $\widehat{X}_1\lambda = X_2\mu_2$, s'écrit $\widetilde{Z}\Gamma_{1\widetilde{z}}\lambda + X_2\left(\Gamma_{1x_2}\lambda - \mu_2\right) = 0$. Comme Z est de rang $K+1$ ceci nécessite $\Gamma_{1\widetilde{z}}\lambda = 0$. Et on a vu que la condition rang $\left(Z'X\right) = K+1$ était équivalente à $\Gamma_{1\widetilde{z}}$ de rang K_1 on a donc nécessairement sous cette condition $\lambda = 0$ et donc la matrice de variance de $\widehat{b}^{(1)}_{2mc} - \widehat{b}^{(1)}_{mco}$ est inversible : le nombre de degrés de liberté du test d'exogénéité est égal à K_1.

Le test de Hausman Sous l'hypothèse d'homoscédasticité, $E(u_i^2|x_i, z_i) = \sigma^2$, $\widehat{b}_{mco}$ est l'estimateur de variance minimale dans la classe des estimateur sans biais dont fait parti l'estimateur des doubles moindres carrés. On a donc

$$\begin{aligned} V\left(\widehat{b}_{2mc} - \widehat{b}_{mco}\right) &= V\left(\widehat{b}_{2mc}\right) - V\left(\widehat{b}_{mco}\right) \\ \widehat{V}\left(\widehat{b}_{2mc} - \widehat{b}_{mco}\right) &= \widehat{\sigma}^2\left[\left(\widehat{X}'\widehat{X}\right)^{-1} - \left(X'X\right)^{-1}\right]. \end{aligned}$$

On en déduit que sous l'hypothèse nulle d'exogénéité de x_i, la statistique

$$\begin{aligned} \widehat{S} &= \frac{1}{\widehat{\sigma}^2}\left(\widehat{b}^{(1)}_{2mc} - \widehat{b}^{(1)}_{mco}\right)'\left[\left(\widehat{X}'\widehat{X}\right)^{11} - \left(X'X\right)^{11}\right]^{-1}\left(\widehat{b}^{(1)}_{2mc} - \widehat{b}^{(1)}_{mco}\right) \\ &\overset{Loi}{\rightarrow} \chi^2\left(K_1\right) \end{aligned}$$

suit une loi du χ^2 à K_1 degrés de liberté

Un test au niveau α sera donc effectué en comparant la valeur de de la statistique $\widehat{S}$ au quantile d'ordre $1-\alpha$ d'une loi du χ^2 à K_1 degrés de liberté.

Test d'exogénéité par le biais de la régression augmentée Le test d'Hausman d'exogénéité peut être mis en oeuvre tr és simplement par le biais d'une simple régression des la variable dépendante Y sur les variables endogènes et exogènes du modèle X_1 et X_2 et sur la projection des variables endogènes sur les variables instrumentales $\widehat{X}_1$:

$$Y = X_1c_1 + X_2c_2 + \widehat{X}_1\gamma + W$$

L'estimateur MCO du coefficient de γ s'obtient aisément à partir de théorème de Frish-Waugh : il s'agit du coefficient de la régression des mco sur le résidu de la régression de $\widehat{X}_1$ sur les autres variables, c'est à dire X. On a donc

$$\widehat{\gamma} = \left(\widehat{X}_1M_X\widehat{X}_1\right)^{-1}\widehat{X}_1M_XY$$

or on a vu précédemment

$$\widehat{b}^{(1)}_{2mc} - \widehat{b}^{(1)}_{mco} = \widehat{X}'\widehat{X}^{11}\widehat{X}_1'M_XY$$

On en déduit que l'on a :

$$\widehat{b}^{(1)}_{2mc} - \widehat{b}^{(1)}_{mco} = \widehat{X}'\widehat{X}^{11}\left(\widehat{X}_1M_X\widehat{X}_1\right)\widehat{\gamma}$$

Le test de $p\lim \widehat{b}_{2mc}^{(1)} - p\lim \widehat{b}_{mco}^{(1)} = 0$ est donc équivalent au test de $\gamma = 0$.

Le test peut donc être effectué trés simplement par l'intermédiaire d'un test de Wald ou d'un test de Fisher.

Remarquons en fin que le test peut être mené de façon analogue sur sur les résidus des régressions des variables explicatives endogènes sur les instruments $\varepsilon(X_1) = X_1 - \widehat{X}_1$. L'équation

$$Y = X_1 c_1 + X_2 c_2 + \widehat{X}_1 \gamma + W$$

se réécrit de façon analogue comme

$$\begin{aligned} Y &= X_1 (c_1 + \gamma) + X_2 c_2 - \varepsilon(X_1)\gamma + W \\ &= X_1 \widetilde{c}_1 + X_2 c_2 + \varepsilon(X_1)\widetilde{\gamma} + W \end{aligned}$$

le test de $\gamma = 0$ est donc équivalent à celui de $\widetilde{\gamma} = 0$.

10 La Méthode des moments généralisée

10.1 Modèle structurel et contrainte identifiante : restriction sur les moments

Une équation :

$$y_i = x_i b + u_i$$

peut provenir du comportement d'optimisation d'un individu et de ce fait associer au paramètre b un sens économique : élasticité de substitution, élasticité de la demande aux prix, mais telle qu'elle est écrite, elle ne constitue pas pour autant un modèle économétrique.

Il faut pour cela ajouter à cette écriture *une contrainte identifiante*. Si par exemple on fait l'hypothèse est l'indépendance des perturbations et des variables explicatives, on a :

$$E\left(x_i^{'} u_i\right) = 0$$

C'est sous cette dernière forme que le modèle peut être considéré comme un modèle économétrique.

Cette contrainte identifiante conduit à des restrictions de moments, qui sont à la base de l'estimation.

$$E\left(x_i^{'} \left(y_i - x_i b\right)\right) = 0$$

Dans certains cas, c'est spontanément sous cette forme qu'un modèle émerge de la théorie. C'est le cas en particulier des équations d'Euler.

10.2 La méthode des moments généralisée

La méthode des moments généralisée concerne la situation dans laquelle on dispose d'un vecteur de fonctions g de dimension $\dim g$ d'un paramètre d'intérêt θ de dimension $\dim \theta$ et de variables aléatoires observables z_i dont l'espérance est nulle pour $\theta = \theta_0$ la vraie valeur du paramètre :

$$E\left(g\left(z_i, \theta\right)\right) = 0 \Leftrightarrow \theta = \theta_0$$

de telles relations portent le nom de *conditions d'orthogonalité*.

C'est un cadre très général englobant de nombreuses situations spécifiques :

– *maximum de vraisemblance* : On a des observations z_i et un modèle dont la vraisemblance s'écrit $LogL\left(z_i, \theta\right)$. Comme

$$E\left(\frac{L\left(z_i, \theta\right)}{L\left(z_i, \theta_0\right)}\right) = \int \frac{L\left(z_i, \theta\right)}{L\left(z_i, \theta_0\right)} L\left(z_i, \theta_0\right) dz_i = \int L\left(z_i, \theta\right) dz_i = 1 \quad \forall \theta$$

et que du fait de l'inégalité de Jensen

$$\log\left(E\left(\frac{L\left(z_i, \theta\right)}{L\left(z_i, \theta_0\right)}\right)\right) > E\left(\log\left(\frac{L\left(z_i, \theta\right)}{L\left(z_i, \theta_0\right)}\right)\right)$$

pour $\theta \neq \theta_0$, on a

$$0 > E\left(\log L\left(z_i, \theta\right)\right) - E\left(\log L\left(z_i, \theta_0\right)\right)$$

L'espérance de la vraisembleance est maximale pour $\theta = \theta_0$:

$$E\frac{\partial \log L(z_i, \theta)}{\partial \theta} = 0 \Leftrightarrow \theta = \theta_0$$

– *modèle d'espérance conditionnelle, moindres carrés non linéaires*
On a une variable y_i dont l'espérance conditionnelle à des variables explicatives x_i s'écrit

$$E(y_i | x_i) = f(x_i, \theta_0)$$

comme

$$\begin{aligned} E\left[(y_i - f(x_i, \theta))^2\right] &= E[y_i - f(x_i, \theta_0) + f(x_i, \theta_0) - f(x_i, \theta)]^2 \\ &= E\left[(y_i - f(x_i, \theta_0))^2\right] \\ &\quad +2E[(y_i - f(x_i, \theta_0))(f(x_i, \theta_0) - f(x_i, \theta))] \\ &\quad +E\left[(f(x_i, \theta_0) - f(x_i, \theta))^2\right] \\ &> E\left[(y_i - f(x_i, \theta_0))^2\right] \end{aligned}$$

on en déduit

$$E\left[(y_i - f(x_i, \theta))\frac{\partial f(x_i, \theta)}{\partial \theta}\right] = 0 \Leftrightarrow \theta = \theta_0$$

– *méthode à variables instrumentales pour un système d'équations.*

$$E\left(Z_i^{'}(y_i - x_i\theta_0)\right) = 0$$

où y_i est un vecteur de variables dépendantes de dimension $M \times 1$, x_i une matrice de variables explicatives de dimension $M \times \dim(\theta)$ et Z_i une matrice d'instruments de dimension $M \times H$ où la ligne m contient les instruments z_m de l'éqution $m : Z_i = diag(z_{mi})$ de telle sorte que

$$Z_i'\varepsilon_i = \begin{bmatrix} z_{1i}' & & \\ & \ddots & \\ & & z_{Mi}^{'} \end{bmatrix} \begin{bmatrix} \varepsilon_{1i} \\ \vdots \\ \varepsilon_{Mi} \end{bmatrix} = \begin{bmatrix} z_{1i}^{'}\varepsilon_{1i} \\ \vdots \\ z_{Mi}^{'}\varepsilon_{Mi} \end{bmatrix}$$

On a

$$E\left(Z_i^{'}(y_i - x_i\theta)\right) = E\left(Z_i^{'}\ x_i\right)(\theta_0 - \theta)$$

Dés lors que $E\left(Z_i^{'}x_i\right)$ est de rang $\dim(\theta)$

$$E\left(Z_i^{'}(y_i - x_i\theta)\right) = 0 \Leftrightarrow \theta = \theta_0$$

Ce cas simple, linéaire, englobe lui même de trés nombreuses situations, comme celles vues jusqu'à présent mco, variables instrumentales dans le cas univarié mais bien d'autres encore comme l' économétrie des données de panel, l'estimation de système de demande, ou encore l'estimation de systèmes offre-demande.

10.3 Principe de la méthode :

Le principe de la méthode GMM est de trouver $\widehat{\theta}$, rendant

$$\overline{g\left(z_i,\widehat{\theta}\right)},$$

la contrepartie empirique de $E\left(g\left(z_i,\theta\right)\right)$ aussi proche que possible de zéro.

- Si $\dim\left(g\right)=\dim\left(\theta\right)$ on peut exactement annuler $\overline{g\left(z_i,\widehat{\theta}\right)}$: *le modèle est juste identifié* (cas des mco, du maximum de vraisemblance, des moindres carrés non linéaires)
- Si $\dim\left(g\right)>\dim\left(\theta\right)$ On ne peut pas annuler exactement la contrepartie empirique des conditions d'orthogonalité. *Le modèle est dit suridentifié.* C'est le cas le plus fréquent lorsque l'on met en oeuvre des méthodes de type variables instrumentales.

Remarque. L'écriture du modèle signifie qu'on peut annuler exactement l'espérance $E\left(g\left(z_i,\theta\right)\right)$ m ême dans le cas de la suridentification, quand bien même c'est impossible à distance finie pour la contrepartie empirique des conditions d'orthogonalité.

Dans le cas de suridentification, la méthode consiste à rendre aussi proche de zéro que possible la norme de la contrepartie empirique des conditions d'orthogonalité dans une certaine métrique :

$$\left\|\overline{g\left(z_i,\theta\right)}\right\|_{S_N}=\overline{g\left(z_i,\theta\right)}^{'}S_N\overline{g\left(z_i,\theta\right)}$$

L'estimateur est alors défini par :

$$\widehat{\theta}=Arg\min_{\theta}\overline{g\left(z_i,\theta\right)}^{'}S_N\overline{g\left(z_i,\theta\right)}$$

Exemple. Cas où les conditions d'orthogonalité sont linéaires dans le paramètre d'intérêt. C'est par exemple le cas des variables instrumentales dans un système d'équations puisqu'alors

$$g\left(z_i,\theta\right)=Z_i^{'}\left(y_i-x_i\theta\right)=Z_i^{'}y_i-Z_i^{'}x_{\ i}\theta=g_1\left(z_i\right)-g_2\left(z_i\right)\theta$$

On note $\overline{g_1}=\overline{g_1\left(z_i\right)}$ et $\overline{g_2}=\overline{g_2\left(z_i\right)}$. L'estimateur est alors défini par :

$$\widehat{\theta}_S=\text{Arg}\min_{\theta}\left(\overline{g_1}-\overline{g_2}\theta\right)^{'}S_N\left(\overline{g_1}-\overline{g_2}\theta\right)$$

Il existe dans ce cas une solution explicite :

$$\widehat{\theta}_S=\left(\overline{g_2}^{'}S_N\overline{g_2}\right)^{-1}\overline{g_2}^{'}S_N\overline{g_1}$$

Dans le cas des variables instrumentales, on a par exemple

$$\widehat{\theta}_S=\left(\overline{x_i^{'}Z_i}S_N\overline{Z_ix_i}\right)^{-1}\overline{Z_ix_i}S_N\overline{Z_iy_i}$$

10.4 Convergence et propriétés asymptotiques

Théorème 10.1. *Sous les hypothèses*

Hypothèse (H_1). L'espace des paramètres Θ est compact. La vraie valeur est θ_0 intérieure à Θ,

Hypothèse (H_2). $E\left(g\left(z_i,\theta\right)\right)=0\Leftrightarrow\theta=\theta_0$,

Hypothèse (H_3). $g\left(z_i,\theta\right)$ est deux fois continuement dérivable en θ,

Hypothèse (H_4). $E\left[\underset{\theta}{sup}\left|g\left(z_i,\theta\right)\right|+\underset{\theta}{sup}\left|g\left(z_i,\theta\right)\right|^2+\underset{\theta}{sup}\left|\nabla_\theta g\left(z_i,\theta\right)\right|\right]<\infty$,

Hypothèse (H_5). $g_k\left(z_i,\theta_0\right)$ a des moments finis d'ordre 1 et 2,

Hypothèse (H_6). Le Jacobien $G=E\left(\nabla_\theta g\left(z_i,\theta_0\right)\right)$ de dimension $\dim g\times\dim\theta$ est de rang $\dim\theta$,

Hypothèse (H_7). $S_N\xrightarrow{P}S_0$ définie positive.

L'estimateur GMM $\widehat{\theta}_{SN}$ minimisant $Q_N\left(\theta\right)$ défini par $Q_N\left(\theta\right)=\overline{g\left(z_i,\theta\right)}'S_N\overline{g\left(z_i,\theta\right)}$, est convergent et asymptotiquement normal. Sa matrice de variance asymptotique est fonction de S_0 et de la matrice de variance des condition d'orthogonalité et peut être estimée de façon convergente :

- $\widehat{\theta}_S\xrightarrow{P}\theta_0$ *convergence*
- $\sqrt{N}\left(\widehat{\theta}_S-\theta_0\right)\xrightarrow{L}N\left(0,V_{as}\left(\widehat{\theta}\left(S\right)\right)\right)$ *normalité asymptotique*
- $V_{as}\left(\widehat{\theta}_S\right)=\left[G'S_0G\right]^{-1}G'S_0V\left(g\left(z_i,\theta_0\right)\right)S_0G\left[G'S_0G\right]^{-1}$ *où* $S_0=p\lim S_N$ *et* $V\left(g\left(z_i,\theta_0\right)\right)=E\left[g\left(z_i,\theta_0\right)g\left(z_i,\theta_0\right)'\right]$
- $\widehat{V}\left(g\left(z_i,\theta_0\right)\right)=\overline{g\left(z_i,\widehat{\theta}_S\right)g\left(z_i,\widehat{\theta}_S\right)'}\rightarrow V\left(g\left(z_i,\theta_0\right)\right)$ *et* $\widehat{G}=\overline{\frac{\partial g}{\partial\theta}\left(z_i,\widehat{\theta}_S\right)}\rightarrow G$
- $\widehat{V}_{as}\left(\widehat{\theta}_S\right)=\left[\widehat{G}'S_0\widehat{G}\right]^{-1}\widehat{G}'S_N\widehat{V}\left(g\left(z_i,\theta_0\right)\right)S_N\widehat{G}\left[\widehat{G}'S_0\widehat{G}\right]^{-1}$

Démonstration.

- *Convergence* :

$$\begin{aligned}Q\left(\widehat{\theta}_S\right)-Q\left(\theta_0\right) &= \left[Q_N\left(\widehat{\theta}_S\right)+\left(Q\left(\widehat{\theta}_S\right)-Q_N\left(\widehat{\theta}_S\right)\right)\right]-\\ &\quad\left[Q_N\left(\theta_0\right)+\left(Q\left(\theta_0\right)-Q_N\left(\theta_0\right)\right)\right]\end{aligned}$$

comme $Q_N\left(\widehat{\theta}_S\right)\leq Q_N\left(\theta_0\right)$ et $Q\left(\theta_0\right)\leq Q\left(\widehat{\theta}_S\right)$,on a

$$\begin{aligned}0 &\leq Q\left(\widehat{\theta}_S\right)-Q\left(\theta_0\right)\leq\left(Q\left(\widehat{\theta}_S\right)-Q_N\left(\widehat{\theta}_S\right)\right)-\left(Q\left(\theta_0\right)-Q_N\left(\theta_0\right)\right)\\ &\leq 2\underset{\theta}{\sup}\left|Q\left(\theta\right)-Q_N\left(\theta\right)\right|\end{aligned}$$

La condition $E\left[\underset{\theta}{sup}\left|g\left(z_i,\theta\right)\right|\right]<+\infty$ permet de montrer qu'il y a *convergence uniforme* de $\overline{g\left(z_i,\theta\right)}$ vers $E\left(g\left(z_i,\theta\right)\right)$, et donc de $Q_N\left(\theta\right)$ vers $Q\left(\theta\right)=E\left(g\left(z_i,\theta\right)\right)'SE\left(g\left(z_i,\theta\right)\right)$ .On en déduit donc que $Q\left(\widehat{\theta}_S\right)\xrightarrow{P}Q\left(\theta_0\right)$. Comme la fonction Q est continue, que Θ est compact, que $Q\left(\theta_0\right)=0$ et $Q\left(\theta\right)=0\Leftrightarrow E\left(g\left(z_i,\theta\right)\right)=0\Leftrightarrow\theta=\theta_0$ on en déduit $\widehat{\theta}_S\xrightarrow{P}\theta_0$.

– *Normalité asymptotique*
La condition du premier ordre définissant le paramètre $\widehat{\theta}_S$ est définie par $\overline{\nabla_\theta g\left(z_i,\widehat{\theta}_S\right)}' S_N \overline{g\left(z_i,\widehat{\theta}_S\right)} = 0$. En appliquant le théorème de la valeur moyenne à $\overline{g\left(z_i,\widehat{\theta}_S\right)}$, on a

$$0 = \sqrt{N}\overline{g\left(z_i,\widehat{\theta}_S\right)}\sqrt{N}\overline{g\left(z_i,\theta_0\right)} + \overline{\nabla_\theta g\left(z_i,\widetilde{\theta}_S\right)}\sqrt{N}\left(\widehat{\theta}_S - \theta_0\right)$$

, où $\widetilde{\theta}_S$ se trouve entre $\widehat{\theta}_S$ et θ_0 converge donc aussi en probabilité vers θ_0. En multipliant par $\overline{\nabla_\theta g\left(z_i,\widehat{\theta}_S\right)}' S_N$, on a

$$\overline{\nabla_\theta g\left(z_i,\widehat{\theta}_S\right)}' S_N \overline{\nabla_\theta g\left(z_i,\widetilde{\theta}_S\right)}\sqrt{N}\left(\widehat{\theta}_S - \theta_0\right) = -\overline{\nabla_\theta g\left(z_i,\widehat{\theta}_S\right)}' S_N \sqrt{N}\overline{g\left(z_i,\theta_0\right)}$$

La condition $E\left[\sup_\theta \left|\nabla_\theta g\left(z_i,\theta\right)\right|\right] < +\infty$ garantit la convergence uniforme en probabilité de $\overline{\nabla_\theta g\left(z_i,\theta\right)}$ vers $E\left(\nabla_\theta g\left(z_i,\theta\right)\right)$. On en déduit que

$$\overline{\nabla_\theta g\left(z_i,\widehat{\theta}_S\right)}' S_N \xrightarrow{P} G'S$$

et que

$$\left(\overline{\nabla_\theta g\left(z_i,\widehat{\theta}_S\right)}' S_N \overline{\nabla_\theta g\left(z_i,\widetilde{\theta}_S\right)}\right) \xrightarrow{P} G'S_0G$$

, matrice $\dim\theta \times \dim\theta$ inversible compte tenu de $\text{rang}G = \dim\theta$. La condition que $g_k\left(z_i,\theta_0\right)$ a des moments d'ordre 1 et 2 permet d'appliquer le théorème central limite à $\sqrt{N}\overline{g\left(z_i,\theta_0\right)}$: $\sqrt{N}\overline{g\left(z_i,\theta_0\right)} \xrightarrow{Loi} N\left(0, V\left(g\left(z_i,\theta_0\right)\right)\right)$. On en déduit la normalité asymptotique de l'estimateur et l'expression de sa matrice de variance. Remarquons que le développement précédent conduit aussi à une approximation de l'écart entre l'estimateur et la vraie valeur :

$$\sqrt{N}\left(\widehat{\theta}_S - \theta_0\right) \simeq -\left(G' S_N G\right)^{-1} G' S_N \sqrt{N}\overline{g\left(z_i,\theta_0\right)}$$

– *Estimation de la matrice de variance asymptotique*
Le seul point à montrer est que $\overline{g\left(z_i,\widehat{\theta}_S\right) g\left(z_i,\widehat{\theta}_S\right)'} \to V\left(g\left(z_i,\theta_0\right)\right)$. La condition $E\left[\sup_\theta \left|g\left(z_i,\theta\right)\right|^2\right] < \infty$, permet de montrer qu'il y a convergence uniforme de $\overline{g\left(z_i,\theta\right) g\left(z_i,\theta\right)'}$ vers $E\left(g\left(z_i,\theta\right) g\left(z_i,\theta\right)'\right)$

■

10.5 Estimateur optimal

Théorème 10.2. *Les estimateurs $\widehat{\theta}^*$ obtenus à partir de matrice de poids $S_N^* \to S^*$ avec*

$$S^* = V\left(g\left(z_i,\theta_0\right)\right)^{-1}$$

sont optimaux, au sens où il conduisent à des estimateurs de variance minimale.

La matrice de variance asymptotique de cet estimateur est

$$V_{as}\left(\widehat{\theta}^*\right) = \left[G'S^*G\right]^{-1} = \left[G'V\left(g\left(z_i,\theta_0\right)\right)^{-1}G\right]^{-1}$$

et peut être estimée par

$$\widehat{V}_{as}\left(\widehat{\theta}^*\right) = \left[\widehat{G}'S_N^*\widehat{G}\right]^{-1}$$

ou $\widehat{G}$ est comme précédemment un estimateur convergent de G.

Démonstration.

La démonstration se fait comme dans le cas des variables instrumentales. La variance asymptotique de l'estimateur optimal s'écrit

$$V_{as}\left(\widehat{\theta}^*\right) = \left[G'V^{-1}G\right]^{-1} = \left(C'C\right)^{-1}$$

avec $C = V^{-1/2}G$ de dimension $\dim g \times \dim \theta$

La variance asymptotique de l'estimateur général s'écrit

$$V_{as}\left(\widehat{\theta}_S\right) = \left[G'S_0G\right]^{-1}G'S_0VS_0G\left[G'S_0G\right]^{-1} = BB'$$

avec $B = \left[G'S_0G\right]^{-1}G'S_0V^{1/2}$ de dimension $\dim\theta \times \dim g$. On a

$$BC = \left[G'S_0G\right]^{-1}G'S_0V^{1/2}V^{-1/2}G = I_{\dim\theta}$$

d'où

$$V_{as}\left(\widehat{\theta}_S\right) - V_{as}\left(\widehat{\theta}^*\right) = BB' - \left(C'C\right)^{-1} = BB' - BC\left(C'C\right)^{-1}C'B'$$

puisque $BC = I_{\dim\theta}$. On voit donc que

$$V_{as}\left(\widehat{\theta}_S\right) - V_{as}\left(\widehat{\theta}^*\right) = B\left(I_{\dim g} - C\left(C'C\right)^{-1}C'\right)B'$$

est une matrice semi définie positive, d'où l'optimalité. ■

10.6 Mise en oeuvre : deux étapes

Dans le cas général, la mise en oeuvre de la méthode des moments généralisée pour obtenir un estimateur optimal présente un problème : la métrique optimale faire intervenir le paramètre à estimer et est donc inconnue.

$$S_0^* = V\left(g\left(z_i,\theta_0\right)\right)^{-1}$$

Pour mettre cet estimateur en oeuvre on a recours à une méthode en deux étapes :

- *Première étape* : On utilise une métrique quelconque (en fait pas si quelconque, intérêt à réfléchir) ne faisant pas intervenir le paramètre. $S_N = I$ est un choix possible mais certainement pas le meilleur. La mise en oeuvre des GMM avec cette m étrique permet d'obtenir un estimateur convergent mais pas efficace $\widehat{\theta}_1$.

A partir de cet estimateur on peut déterminer un estimateur de la matrice de variance des condition d'orthogonalité :

$$\widehat{V}(g)_N = \overline{g\left(z_i,\widehat{\theta}_1\right) g\left(z_i,\widehat{\theta}_1\right)'} \xrightarrow{P} V\left(g\left(z_i,\theta_0\right)\right)$$

ainsi que

$$\widehat{G} = \overline{\nabla_\theta g\left(z_i,\widehat{\theta}_1\right)} \xrightarrow{P} E\left(\nabla_\theta g\left(z_i,\theta_0\right)\right)$$

On peut dés lors déterminer un estimateur de la matrice de variance asymptotique de ce premier estimateur

$$\widehat{V}_{as}\left(\widehat{\theta}_1\right)_N = \left(\widehat{G}' S_N \widehat{G}\right)^{-1} \widehat{G}' S_N \widehat{V}(g)_N S_N \widehat{G}\left(\widehat{G}' S_N \widehat{G}\right)^{-1}$$

– *Deuxième étape* : On met à nouveau en oeuvre l'estimateur des GMM avec la métrique $S_N^* = \widehat{V}(g)_N^{-1}$. On obtient ainsi un estimateur convergent et asymptotiquement efficace dont on peut estimer la matrice de variance asymptotique

$$\widehat{V}_{as}\left(\widehat{\theta}^*\right)_N = \left(\widehat{G}' S_N^* \widehat{G}\right)^{-1}$$

10.7 Application aux variables instrumentales dans un système d' équations

On considère le cas d'un système d'équations avec variables instrumentales

$$g\left(z_i,\theta\right) = Z_i'\left(y_i - x_i\theta\right) = Z_i' y_i - Z_i' x_{\ i}\theta$$

– *Vérification des hypothèses*

1. $H_2 : E\left(Z_i' y_i\right) - E\left(Z_i' x_i\right)\theta = 0$ admet une unique solution si rang$E\left(Z_i' x_i\right) = \dim\theta$, simple généralisation de la condition déjà vue dans le cadre univarié.
2. H_3 : est satisfaite du fait de la linéarité.
3. H_4 et H_5 sont satisfaites si $E\left[\left(\sup\left|Z_i' y_i\right| + \sup\left|Z_i' x_i\right|\right)^2\right] < +\infty$, c'est à dire si les moments d'ordre quatres de Z_i, x_i et y_i existent.
4. $H_6 : \nabla_\theta g\left(z_i,\theta_0\right) = -Z_i' x_i$. Si $E\left(Z_i' x_i\right)$ est de rang $\dim\theta$, $G = E\left(\nabla_\theta g\left(z_i,\theta_0\right)\right) = -E\left(Z_i' x_i\right)$ est de rang $\dim\theta$.

– *Expression de la matrice de variance des conditions d'orthogonalité*
La variance des conditions d'orthogonalité s'écrit

$$\begin{aligned} V\left(g\left(z_i,\theta_0\right)\right) &= E\left(Z_i'\left(y_i - x_i\theta_0\right)\left(y_i - x_i\theta_0\right)' Z_i\right) \\ &= E\left(Z_i' u_i u_i' Z_i\right) \end{aligned}$$

Expression trés proche de celle vue dans le cadre des variables instrumentales. Cette expression fait bien intervenir en général le paramètre θ et il est alors nécessaire de mettre en oeuvre une méthode en deux étapes.

– *Mise en oeuvre de l'estimation*
Première étape : l'estimateur a pour expression :

$$\widehat{\theta}_S = \left(\overline{x_i' Z_i} S_N \overline{Z_i' x_i}\right)^{-1} \overline{x_i' Z_i} S_N \overline{Z_i' y_i}$$

La matrice de variance des conditions d'ortogonalité peut être estimée par

$$\widehat{V}(g) = \overline{Z_i' \left(y_i - x_i \widehat{\theta}_S\right) \left(y_i - x_i \widehat{\theta}_S\right)' Z_i} = \overline{Z_i' \widehat{u}_i} \overline{\widehat{u}_i' Z_i}$$

A partir de cette estimation, on peut aussi estimer la variance de l'estimateur de première étape :

$$\widehat{V}\left(\widehat{\theta}(S)\right) = \left(\overline{x_i' Z_i} S_N \overline{Z_i' x_i}\right)^{-1} \overline{Z_i' x_i} S_N \widehat{V}(g) S_N \overline{x_i' Z_i} \left(\overline{Z_i' x_i} S_N \overline{Z_i' x_i}\right)^{-1}$$

ainsi que l'estimateur optimal :

$$\widehat{\theta}_S^* = \left(\overline{x_i' Z_i} \widehat{V}(g)^{-1} \overline{Z_i' x_i}\right)^{-1} \overline{x_i' Z_i} \widehat{V}(g)^{-1} \overline{Z_i' y_i}$$

et sa variance asymptotique :

$$\widehat{V}_{as}\left(\widehat{\theta}_S^*\right) = \left(\overline{x_i' Z_i} \widehat{V}(g)^{-1} \overline{Z_i' x_i}\right)^{-1}$$

10.7.1 Régressions à variables instrumentales dans un système homoscédastique

Dans le cas où on fait l'hypothèse d'homoscédasticité : $E(u_i u_i' | Z_i) = \Sigma = E\left((y_i - x_i\theta_0)(y_i - x_i\theta_0)'\right)$, on a $V(g(z_i, \theta_0)) = E\left(Z_i' \Sigma Z_i\right)$. Si les régresseurs sont les mêmes, si il n'existe pas de contraintes entre les paramètres des équations $x_i = I_M \otimes x_i$, et si les instruments sont les mêmes d'une équation à l'autre $Z_i = I_M \otimes z_i$, on a $x_i' Z_i = I_M \otimes x_i' z_i$.

Sous l'hypothèse d'homoscédasticité, la matrice de variance des conditions d'orthogonalité a pour expression $E\left(Z_i' \Sigma Z_i\right) = \Sigma \otimes E\left(z_i' z_i\right)$.

Rappel : pour des matrices aux tailles appropriées $(A \otimes B)(C \otimes D) = AC \otimes BD$. On a donc $\Sigma Z_i = (\Sigma \otimes 1)(I_M \otimes z_i) = \Sigma \otimes z_i$. D'où $Z_i' \Sigma Z_i = \left(I_M \otimes z_i'\right)(\Sigma \otimes z_i) = \Sigma \otimes z_i' z_i$. On a donc

$$\begin{aligned}
\overline{x_i' Z_i} S^* \overline{Z_i' x_i} &= \left(I_M \otimes \overline{x_i' z_i}\right) \left(\Sigma \otimes E\left(z_i' z_i\right)\right)^{-1} \left(I_M \otimes \overline{z_i' x_i}\right) \\
&= \Sigma^{-1} \otimes \left(\overline{x_i' z_i} E\left(z_i z_i'\right)^{-1} \overline{z_i' x_i}\right)
\end{aligned}$$

et

$$\begin{aligned}
\overline{x_i' Z_i} S^* \overline{Z_i' y_i} &= \left(I_M \otimes \overline{x_i' z_i}\right) \left(\Sigma \otimes E\left(z_i' z_i\right)\right)^{-1} \overline{(I_M \otimes z_i') y_i} \\
&= \left[\Sigma^{-1} \otimes \left(\overline{x_i' z_i} E\left(z_i z_i'\right)^{-1}\right)\right] \begin{bmatrix} \overline{z_i' y_{1i}} \\ \vdots \\ \overline{z_i' y_{Mi}} \end{bmatrix}
\end{aligned}$$

puisque $\left(I_M \otimes z_i^{'}\right) y_i = \begin{bmatrix} z_i' y_{1i} \\ \vdots \\ z_i' y_{Mi} \end{bmatrix}$

L'estimateur optimal a donc pour expression

$$\begin{aligned} \widehat{\theta}_S^* &= \Sigma \otimes \left(\overline{x_i' z_i} E\left(z_i z_i^{'}\right)^{-1} \overline{z_i' x_i} \right)^{-1} \times \Sigma^{-1} \otimes \left(\overline{x_i' z_i} E\left(z_i z_i^{'}\right)^{-1} \right) \begin{bmatrix} \overline{z_i' y_{1i}} \\ \vdots \\ \overline{z_i' y_{Mi}} \end{bmatrix} \\ &= I_M \otimes \overline{x_i' z_i} \left(\Sigma \otimes E\left(z_i z_i^{'}\right) \right)^{-1} \begin{bmatrix} \overline{z_i' y_{1i}} \\ \vdots \\ \overline{z_i' y_{Mi}} \end{bmatrix} = \begin{bmatrix} \widehat{b}_{2mc1} \\ \vdots \\ \widehat{b}_{2mcM} \end{bmatrix} \end{aligned}$$

On voit que dans ce cas, l'estimateur optimal est *identique* à *l'estimateur des doubles moindres carrés* effectué *équation* par *équation*. Il n'y a donc pas non plus dans ce cas de méthode en deux étapes à mettre en oeuvre. La matrice de variance des paramètres a pour expression

$$V\left(\widehat{\theta}^*\right) = \Sigma \otimes \left(E\left(x_i' z_i\right) E\left(z_i z_i^{'}\right)^{-1} E\left(z_i' x_i\right) \right)^{-1}$$

on voit donc que les estimateurs ne sont pas indépendants les uns des autres dés que la matrice de variance Σ n'est pas diagonale.

10.7.2 Estimateur à variables instrumentales optimal dans le cas univarié et hétéroscédastique

On considère la situation d'un modèle linéaire univarié

$$y_i = x_i \theta + u_i$$

avec un ensemble d'instruments z_i : Le sconditions d'orthogonalité sont donc

$$E\left(z_i^{'}\left(y_i - x_i\theta\right)\right) = 0$$

Le résultat précédent montre que dans le cas univarié homoscédastique, i.e. $E\left(u_i^2 \left| z_i \right.\right) = E\left(u_i^2\right)$, l'estimateur GMM optimal coïncide avec l'estimateur des 2mc. On examine la situation dans laquelle il n'y a plus homoscédasticité.

La matrice de variance des conditions d'ortogonalité est donnée par

$$V(g) = E\left(\left(y_i - x_i\theta_0\right)^2 z_i^{'} z_i\right) = E\left(u_i^2 z_i^{'} z_i\right)$$

et l'estimateur optimal a pour expression

$$\widehat{\theta}_S^* = \left(\overline{x_i' z_i} V(g)^{-1} \overline{z_i' x_i} \right)^{-1} \overline{x_i' z_i} V(g)^{-1} \overline{z_i' y_i}$$

on voit qu'il est différent de l'estimateur des $2mc$ dont l'expression est

$$\widehat{\theta}_{2mc} = \left(\overline{x_i' z_i}\ \overline{z_i' z_i}^{-1} \overline{z_i' x_i} \right)^{-1} \overline{x_i' z_i}\ \overline{z_i' z_i}^{-1} \overline{z_i' y_i}$$

Il faut donc mettre en oeuvre la méthode en deux étapes. On peut par exemple partir de l'estimateur des 2mc, qui est certainement proche de l'estimateur optimal, et calculer un estimateur de la matrice de variance des conditions d'orthogonalité,

$$\widehat{V}(g) = \overline{\widehat{u}_i^2 z_i' z_i}$$

puis déterminer l'estimateur optimal,

$$\widehat{\theta}_S^* = \left(\overline{x_i' z_i}\ \overline{\widehat{u}_i^2 z_i' z_i}^{-1} \overline{z_i' x_i}\right)^{-1} \overline{x_i' z_i}\ \overline{\widehat{u}_i^2 z_i' z_i}^{-1} \overline{z_i' y_i}$$

ainsi que les matrice de variance de chacun des estimateurs :

$$\begin{aligned} V_{as}\left(\widehat{\theta}_{2mc}\right) &= \left(\overline{x_i' z_i}\ \overline{z_i' z_i}^{-1} \overline{z_i' x_i}\right)^{-1} \overline{x_i' z_i}\ \overline{z_i' z_i}^{-1}\ \overline{\widehat{u}_i^2 z_i' z_i} \\ &\quad \overline{z_i' z_i}^{-1} \overline{x_i z_i'}\ \left(\overline{x_i' z_i}\ \overline{z_i' z_i}^{-1} \overline{z_i' x_i}\right)^{-1} \\ V_{as}\left(\widehat{\theta}^*\right) &= \left(\overline{x_i' z_i}\ \overline{\widehat{u}_i^2 z_i' z_i}^{-1} \overline{z_i' x_i}\right)^{-1} \end{aligned}$$

10.8 Test de spécification.

Comme pour les variables instrumentales, dans le cas où il y a plus de conditions d'orthogonalité que de paramètres à estimer, le modèle impose des restrictions aux données. Elles doivent vérifier la propriété :

$$\exists \quad \theta \quad | \qquad E\left(g\left(z_i, \theta\right)\right) = 0$$

Intuitivement : on peut éliminer le paramètre en se servant d'une partie des équations. L'hypothèse $\exists \quad \theta_0 \quad tq \quad E\left(g\left(z_i, \theta_0\right)\right) = 0$ peut être reformulée de façon équivalente sous la forme $E\left(\phi\left(z_i\right)\right) = 0$ avec $\dim\left(\phi\right) = \dim\left(g\right) - \dim\left(\theta\right)$. Ce sont ces restrictions additionnelles que l'on teste.

Le principe reste le même : regarder si $\overline{g\left(z_i, \theta_0\right)}$ est proche de 0, mais on ne connaît pas θ_0 .

Plus précisément : on regarde si $\overline{\widehat{g}_i} = \overline{g\left(z_i, \widehat{\theta}^*\right)}$ est proche de 0, c'est à dire si la contrepartie empirique des conditions d'orthogonalité *évaluée avec l'estimateur optimal* est proche de zéro.

Le résultat général s'applique

$$N\overline{\widehat{g}_i}' V_{as}\left(\overline{\widehat{g}_i}\right)^{-} \overline{\widehat{g}_i} \to \chi^2\left(\text{rang} V\left(\overline{\widehat{g}_i}\right)\right)$$

Pour effectuer le test il faut donc déterminer le rang de $V_{as}\left(\overline{\widehat{g}_i}\right)$ ainsi qu'un inverse généralisé et un estimateur convergent de cet inverse.

Théorème 10.3. *Sous* $H_0 : \exists \quad \theta \quad | \qquad E\left(g\left(z_i, \theta\right)\right) = 0$*, on a*

$$NQ_N^*\left(\theta^*\right) = N\overline{\widehat{g}_i}' S_N^* \overline{\widehat{g}_i} \xrightarrow{L} \chi^2\left(\dim\left(g\right) - \dim\left(\theta\right)\right)$$

où $\overline{\widehat{g}_i} = \overline{g\left(z_i, \widehat{\theta}^*\right)}$ *et* $S_N^* = \widehat{V}\left(g\left(z_i, \theta_0\right)\right)^{-1} = \overline{g\left(z_i, \widehat{\theta}^*\right) g\left(z_i, \widehat{\theta}^*\right)'}^{-1}$

On remarque que la statistique utilisée pour le test est N fois la valeur de l'objectif à l'optimum.

Démonstration. Comme

$$\sqrt{N}\widehat{\overline{g_i}} \simeq \sqrt{N}\overline{g_{i_0}} + G\left(\widehat{\theta}^* - \theta_0\right)$$

et

$$\sqrt{N}\left(\widehat{\theta}^* - \theta_0\right) \simeq -\left(G' S_N G\right)^{-1} G' S^* \sqrt{N}\overline{g_{i_0}}$$

on a

$$\sqrt{N}\widehat{\overline{g_i}} \simeq \left(I_{\dim g} - G\left(G' S^* G\right)^{-1} G' S^*\right)\sqrt{N}\overline{g_{i_0}} = \left(I_{\dim g} - P_G\right)\sqrt{N}\overline{g_{i_0}}$$

avec $P_G = G\left(G' S^* G\right)^{-1} G' S^*$. $P_G^2 = P_G$. P_G est donc un projecteur dont le rang est celui de G, i.e $\dim\theta$. Comme en outre $P_G S^{*-1} P_G' = P_G S^{*-1}$, et $V_{as}\left(g_{i_0}\right) = S^{*-1}$, on a

$$V_{as}\left(\widehat{\overline{g_i}}\right) = \left(I_{\dim g} - P_G\right) S^{*-1}\left(I - P_G\right)' = \left(I_{\dim g} - P_G\right) S^{*-1}$$

On en déduit immédiatement le rang de $V_{as}\left(\widehat{\overline{g_i}}\right)$:

$$\text{rang}V\left(\widehat{\overline{g_i}}\right) = \dim g - \dim\theta$$

et un inverse généralisé :

$$\begin{aligned} V_{as}\left(\widehat{\overline{g_i}}\right) S^* V_{as}\left(\widehat{\overline{g_i}}\right) &= \left(I_{\dim g} - P_G\right) S^{*-1} S^* \left(I_{\dim g} - P_G\right) S^{*-1} \\ &= \left(I_{\dim g} - P_G\right)^2 S^{*-1} = \left(I_{\dim g} - P_G\right) S^{*-1} \\ &= V_{as}\left(\widehat{\overline{g_i}}\right) \end{aligned}$$

d'où

$$S^* = V_{as}\left(\widehat{\overline{g_i}}\right)^-$$

Estimation convergente de l'inverse généralisée : Comme la matrice $\overline{g\left(z_i,\theta\right) g\left(z_i,\theta\right)'}$ est une fonction continue de θ convergent uniformémént vers $E\left(g\left(z_i,\theta\right) g\left(z_i,\theta\right)'\right)$, $S_N^* = \overline{g\left(z_i,\widehat{\theta}^*\right) g\left(z_i,\widehat{\theta}^*\right)'}$ converge vers S^* ■

10.8.1 Application test de suridentification pour un estimateur à variables instrumentales dans le cas univarié et hétéroscédastique

Le test est effectué sur la contrepartie empirique des conditions d'orthogonalité évaluées en $\theta = \widehat{\theta}^*$, l'estimateur optimal. On calcule donc :

$$\overline{z_i'\left(y_i - x_i\widehat{\theta}^*\right)} = \overline{z_i'\widehat{u_i}^*}$$

et sa norme

$$\overline{z_i'\widehat{u_i}^*}'\,\overline{\widehat{u}_i^2 z_i' z_i}^{-1}\,\overline{z_i'\widehat{u_i}^*}$$

où $\widehat{u}_i = y_i - x_i\widehat{\theta}_1$ est le résidu de l'équation estimé à partir d'une première étape

Le résultat stipule que sous l'hypothèse nulle, $H_0 : \exists\theta \,|\mathbb{E}\left(z_i^{'}\left(y_i - x_i\theta\right)\right) = 0$, la statistique

$$\widehat{S}_\chi = N\overline{z_i'\widehat{u}_i^*}'\overline{\widehat{u}_i^2 z_i' z_i}^{-1}\overline{z_i'\widehat{u}_i^*} \to \chi^2\left(\dim z - \dim x\right)$$

On rejettera l'hypothèse nulle si $\widehat{S}_\chi$ est trop grand, i.e. pour un test au niveau α $\widehat{S}_\chi > Q\left(1-\alpha, \chi^2\left(\dim z - \dim x\right)\right)$

11 Variables dépendantes limitées

On a examiné jusqu'à présent le cas de modèles linéaires pour lesquels la variable dépendante y_i avait pour support $\mathbb{R}$. On examine dans ce chapitre la spécification et l'estimation de modèles dans des situations plus générales.

On examine trois cas

- *Modèle dichotomique :* $y_i \in \{0, 1\}$. Par exemple : participation au marché du travail, à un programme de formation, faillite d'une entreprise, défaut de paiement, signature d'un accord de passage aux 35 heures etc. Les informations dont on dispose dans les enquêtes sont souvent de cette nature : avez vous au cours de la période du tant au tant effectué telle ou telle action .
- *Modèle de choix discret* comme par exemple le choix du lieu de vacances (pas de vacances, montagne, mer, campagne) ou le choix du moyen de transport domicile-travail (bus, auto, metro, à pied). Ces situations conduisent à des variables prenant un nombre fini de modalités $y_i \in \{0, 1, 2, \ldots, M\}$.
- *Données tronquées :* on observe une variable y_i uniquement conditionnellement à la réalisation d'une autre variable. Par exemple le salaire n'est observé que conditionnellement au fait que l'individu ait un emploi. On a alors deux variables à modéliser : la variable de censure $I_i \in \{0, 1\}$ indiquant si le salaire est observé ou non et la variable de salaire w_i lorsqu'il est observé.

11.1 Modèle dichotomique

On souhaite expliquer une variable endogène y_i prenant les valeurs 1 ou 0 en fonction de variables explicatives exogènes x_i,

D'une façon générale on spécifie la probabilité d'observer $y_i = 1$ conditionnellement aux variables explicatives x_i.

$$P(y_i = 1 \,|x_i\,) = \widetilde{G}(x_i)$$

qui définit complètement la loi conditionnelle de y_i sachant x_i. Cette probabilité est aussi l'espérance conditionnelle de la variable y_i :

$$\begin{aligned} E(y_i \,|x_i\,) &= \sum_{y_i \in \{0,1\}} y_i \left[1_{(y_i=1)} P(y_i = 1 \,|x_i\,) + 1_{(y_i=0)} (1 - P(y_i = 1 \,|x_i\,))\right] \\ &= P(y_i = 1 \,|x_i\,) = \widetilde{G}(x_i) \end{aligned}$$

On spécifie en général cette fonction comme dépendant d'un indice linéaire en x_i :

$$\widetilde{G}(x_i) = G(x_i b)$$

Les différentes solutions que l'on peut apporter à la modélisation de la variable dichotomique y_i correspondent à différents choix pour la fonction G.

11.1.1 Modèle à probabilités linéaires

C'est la situation dans laquelle on sépcifie simplement

$$E(y_i \,|x_i\,) = P(y_i = 1 \,|x_i\,) = x_i b$$

Le modèle peut alors être estimé par les MCO.

En dépit de sa simplicité attractive, ce choix de modélisation présente néanmoins des inconvénients :

Deux inconvénients de ce modèle

– Un premier problème vient de l'estimation. Compte tenu du fait que $y_i^2 = y_i$, toute estimation de modèle de choix discret par les moindres carrés, linéaire dans le cas présent ou non linéaire dans le cas gén éral, c'est à dire basée sur la spécification $E(y_i|x_i) = G(x_i b)$, doit prendre en compte le fait que le modèle de régression correspondant

$$y_i = G(x_i b) + u_i$$

est hétéroscédatique. En effet on a :

$$\begin{aligned} V(y_i|x_i) &= E\left(y_i^2|x_i\right) - E(y_i|x_i)^2 = E(y_i|x_i) - E(y_i|x_i)^2 \\ &= E(y_i|x_i)\left[1 - E(y_i|x_i)\right] = G(x_i b)\left[1 - G(x_i b)\right] \end{aligned}$$

L'estimateur des mco dans le cas linéaire a donc pour variance

$$V_{as}\left(\widehat{b}_{mco}\right) = E\left(x_i' x_i\right)^{-1} E\left(u_i^2 x_i' x_i\right) E\left(x_i' x_i\right)^{-1}$$

que l'on estime par la méthode de White

$$\widehat{V}_{as}\left(\widehat{b}_{mco}\right) = \overline{x_i' x_i}^{-1} \overline{\widehat{u}_i^2 x_i' x_i} . \overline{x_i' x_i}^{-1}$$

On pourrait aussi songer à estimer plus directement cette matrice compte tenu de la forme de l'hétéroscédasticité, ou même à mettre en oeuvre l'estimateur des MCQG puisque l'on connait l'expression de la matrice de variance des résidus conditionnellement à x_i :

$$E\left(u_i^2|x_i\right) = G(x_i b)(1 - G(x_i b)) = \sigma^2(x_i b)$$

Par exemple pour l'estimateur des MCQG

$$\widehat{b}_{mcqg} = \overline{\widetilde{x}_i' \widetilde{x}_i}^{-1} \overline{\widetilde{x}_i' \widetilde{y}_i}$$

avec $\widetilde{z}_i = z_i \Big/ \sqrt{\sigma^2\left(x_i \widehat{b}_{mco}\right)}$. Ceci est en pratique impossible et soulève un second problème associé à la spécification d'un modèle de probabilité linéaire

– Le modèle ne peut contraindre $P(y_i = 1|x_i) = x_i b$ à appartenir à l'intervalle $[0, 1]$.

11.1.2 Les modèles probit et logit.

Il est donc préférable de faire un autre choix que l'identité pour la fonction G. On souhaite que cette fonction soit croissante, qu'elle tende vers 1 en $+\infty$ et vers 0 en $-\infty$. En principe, la fonction de répartition de n'importe quelle loi de probabilité pourrait convenir. En pratique les modèles de choix discret sont spécifiés en utilisant deux fonctions de répartition :

– Φ, la fonction de répartition de la loi normale :

$$G(z) = \int_{-\infty}^{z} \varphi(t)dt = \Phi(z)$$

où $\varphi(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}t^2\right)$. On a donc dans ce cas

$$P(y_i | x_i) = \Phi(x_i b)$$

Un tel modèle est appelé **Modèle Probit.**

– F, la fonction logistique

$$F(z) = \frac{1}{1+\exp(-z)}$$

Dans ce cas

$$P(y_i | x_i) = F(x_i b) = \frac{1}{1+\exp(-x_i b)}$$

Un tel modèle est appelé **Modèle Logit**

Relation entre les 3 modèles :

Dans la plupart des applications les différences sont néanmoins assez faibles entre les résultats. On peut pour le voir effectuer un développement limité à l'ordre 3 de chacune des fonction F et Φ

On a

$$\begin{aligned} F(x) &\sim \frac{1}{2} + \frac{1}{4}x - \frac{1}{8}\frac{x^3}{6} = \frac{1}{2} + \frac{x}{4} - \frac{4}{3}\left(\frac{x}{4}\right)^3 \\ \Phi(x) &\sim \frac{1}{2} + \phi(0)\,x - \phi(0)\,\frac{x^3}{6} = \frac{1}{2} + \frac{1}{\sqrt{2\pi}}x - \frac{1}{\sqrt{2\pi}}\frac{x^3}{6} \\ &= \frac{1}{2} + \frac{x}{\sqrt{2\pi}} - \frac{2\pi}{6}\left(\frac{x}{\sqrt{2\pi}}\right)^3 \end{aligned}$$

Donc

$$\begin{aligned} F\left(\frac{4}{\sqrt{2\pi}}x\right) &= \frac{1}{2} + \frac{1}{\sqrt{2\pi}}x - \frac{1}{8}\frac{x^3}{6}\left(\frac{4}{\sqrt{2\pi}}\right)^3 \\ &= \frac{1}{2} + \frac{1}{\sqrt{2\pi}}x - \frac{1}{\sqrt{2\pi}}\frac{x^3}{6}\frac{4}{\pi} \\ &\sim \Phi(x) + \frac{1}{\sqrt{2\pi}}\frac{x^3}{6}\left(\frac{4}{\pi} - 1\right) \sim \Phi(x) + 0.02x^3 \end{aligned}$$

On en conclut que :

1. $\widehat{b}_{Probit} \sim \left(\sqrt{2\pi}/4\right) \times \widehat{b}_{Logit}$, $\sqrt{2\pi}/4 \sim 0.625$
2. $\widehat{b}_{Linéaire} \sim 0.25 \times \widehat{b}_{Logit}$ (+0.5 pour la constante)
3. $\widehat{b}_{Linéaire} \sim 0.4 \times \widehat{b}_{Probit}$ (+0.5 pour la constante)
4. La différence entre la fonction logistique et la fonction probit à l'ordre 3 est trés faible, ce qui suggère que dés lors qu'il n'y a pas de différences trop importantes entre les effectifs des deux populations correspondant aux réalisations de y et que les variables explicatives ne sont pas trop dispersées, l'approximation entre les deux estimations Logit et Probit sera bonne.

5. Les approximations faisant intervenir l'estimations linéaires seront en général moins bonnes, surtout si les effectifs des deux populations sont d éséquilibrés et si les variables explicatives sont dispersées.

Effet marginal d'une variation d'un régresseur continu x Comme $E(y_i|x_i) = G(x_i b)$, on a

$$\frac{\partial E(y_i|x_i)}{\partial x_i^k} = G'(x_i b)\, b_k$$

et l'élasticité

$$\frac{\partial LogE(y_i|x_i)}{\partial x_i^k} = \frac{G'(x_i b)}{G(x_i b)} b_k$$

Pour le modèle Probit on a ainsi :

$$\frac{\partial E(y_i|x_i)}{\partial x_i^k} = \varphi(x_i b)\, b_k, \quad \frac{\partial LogE(y_i|x_i)}{\partial x_i^k} = \frac{\varphi(x_i b)}{\Phi(x_i b)} b_k$$

et pour le modèle Logit

$$\begin{aligned} \frac{\partial E(y_i|x_i)}{\partial x_i^k} &= F(x_i b)(1 - F(x_i b))\, b_k \\ \frac{\partial LogE(y_i|x_i)}{\partial x_i^k} &= (1 - F(x_i b))\, b_k \end{aligned}$$

puisqu'on vérifie facilement $F' = F(1 - F)$

11.1.3 Variables latentes

La modélisation précédente est une modélisation statistique. Les modèles à variables dépendantes discrètes peuvent être souvent introduit par le biais d'une variable latente, c'est à dire une variable inobservée mais qui détermine complètement la réalisation de la variable indicatrice étudiée. Une telle approche permet de rendre plus explicite les hypothèses économiques sous-jacentes à la modélisation.

Exemple. Considérons la décision de participer à un stage de formation. Ce stage représente un gain futur G_i pour l'individu dont le capital humain aura augmenté. Supposons que l'on soit capable de modéliser ce gain à partir de variables explicatives

$$G_i = x_i^g b_g + u_i^g$$

La participation au stage comporte aussi un coût à court-terme C_i, incluant le fait qu'il faut d'abord apprendre, et donc fournir un effort, mais aussi souvent payer pour la formation et subir des coûts indirects comme des coûts de transport. Supposons la encore que l'on soit capables de modéliser ce coût

$$C_i = x_i^c b_c + u_i^c$$

Le gain net pour l'individu est donc $y_i^* = G_i - C_i$.

$$y_i^* = x_i^g b_g - x_i^c b_c + u_i^g - u_i^c = x_i b + u_i$$

On peut modéliser la participation comme le fait que le gain net soit positif :

$$y_i = 1 \Leftrightarrow y_i^* > 0 \Leftrightarrow x_i b + u_i > 0$$

y_i^* est la variable latente associée au modèle. Si on suppose que le résidu intervenant dans modélisation de la variable latente est normal et qu'il est indépendant des variables explicatives, on obtient le modèle Probit. Les paramètres b sont identifiables à un facteur multiplicatif près. Supposons $u_i \rightsquigarrow N\left(0, \sigma^2\right)$

$$y_i = 1 \Leftrightarrow x_i \frac{b}{\sigma} + \frac{u_i}{\sigma} > 0$$

et $v_i = u_i/\sigma \rightsquigarrow N(0,1)$. On pose $c = b/\sigma$, on a donc

$$\begin{aligned} P(y_i = 1 \,|x_i\,) &= P\left(x_i \frac{b}{\sigma} + \frac{u_i}{\sigma} > 0\right) = P(v_i > -x_i c) = P(v_i < x_i c) \\ &= \Phi(x_i c) \end{aligned}$$

où on utilise le fait que la loi normale est symétrique, et que donc $P(v > a) = P(v < -a)$

Le modèle logit est lui aussi compatible avec cette modélisation. On suppose alors que u_i suit une loi logistique de variance σ. La variable u_i/σ suit alors une loi logistique de densité $f(x) = \exp(-x)/(1+\exp(-x))^2$ et de fonction de répartition $F(x) = 1/(1+\exp(-x))$. Cette densité est là encore symétrique en zéro, et on aura

$$\begin{aligned} P(y_i = 1 \,|x_i\,) &= P\left(x_i \frac{b}{\sigma} + \frac{u_i}{\sigma} > 0\right) = P(v_i > -x_i c) = P(v_i < x_i c) \\ &= F(x_i c) \end{aligned}$$

On pourrait considérer d'autres cas comme par exemple le fait que la loi de u_i suive une loi de Student, on obtiendrait alors d'autres expressions pour $P(y_i = 1 \,|x_i\,)$

11.1.4 Estimation des modèles dichotomiques

Les modèles dichotomiques s'estiment par le maximum de vraisemblance. On fait l'hypothèse que les observations sont indépendantes. Compte tenu d'une modélisation conduisant à

$$P(y_i = 1 \,|x_i\,) = G(x_i b)$$

avec G une fonction de répartition connue, de densité g. La probabilité d'observer y_i pour un individu peut s'écrire comme

$$\begin{aligned} P(y_i \,|x_i\,) &= P(y_i = 1 \,|x_i\,)^{y_i} \left[1 - P(y_i = 1 \,|x_i\,)\right]^{1-y_i} \\ &= G(x_i b)^{y_i} \left[1 - G(x_i b)\right]^{1-y_i} \end{aligned}$$

La vraisemblance de l'échantillon s'écrit donc

$$L(Y \,|X\,) = \prod_{i=1}^{N} P(y_i \,|x_i\,) = \prod_{i=1}^{N} G(x_i b)^{y_i} \left[1 - G(x_i b)\right]^{1-y_i}$$

compte tenu de l'hypothèse d'indépendance. La log −vraisemblance s′écrit alors

$$\log L_N = \sum_{i=1}^{N} \left[y_i \log G\left(x_i b\right) + \left(1 - y_i\right) \log\left(1 - G\left(x_i b\right)\right) \right]$$

Conditions de 1er ordre pour la maximisation : L'estimateur du maximum de vraisemblance est défini par :

$$\frac{\partial \log L_N}{\partial\, \beta} = \sum_{i=1}^{N} \left[y_i \frac{g\left(x_i \widehat{b}\right)}{G\left(x_i \widehat{b}\right)} + \left(1 - y_i\right) \frac{-g\left(x_i \widehat{b}\right)}{1 - G\left(x_i \widehat{b}\right)} \right] x_i^{'} = 0$$

soit

$$\frac{\partial \log L_N}{\partial\, b} = \sum_{i=1}^{N} \left[y_i - G\left(x_i \widehat{b}\right) \right] \frac{g\left(x_i \widehat{b}\right)}{G\left(x_i \widehat{b}\right)\left[1 - G\left(x_i \widehat{b}\right)\right]} x_i^{'} = 0$$

Ces équations sont en général non linéaires et nécessitent la mise en oeuvre d'un algorithme d'optimisation.

On voit que ces équations dans le cas général s'expriment sous la forme

$$\sum_{i=1}^{N} \omega\left(x_i, \widehat{b}\right) \left[y_i - E\left(y_i \left| x_i, \widehat{b} \right. \right) \right] x_i^{'} = 0$$

Elles sont donc dans le fond assez similaires aux conditions vues pour les moindres carrés, mis à part la pondération et la non linéarité. On remarque é galement que la pondération s'interprète naturellement par le fait que $V\left(y_i \left| x_i \right.\right) = G\left(x_i, b\right)\left(1 - G\left(x_i, b\right)\right)$, et que $g\left(x_i, b\right) x_i'$ est la dérivée par rapport à b de $G\left(x_i b\right)$. La pondération est donc analogue à une sphéricisation analogue à celle pratiquée dans la méthode des MCQG du modèle linéarisé autour de la vraie valeur du paramètre.

Pour le modèle Logit on a $G\left(z\right) = F\left(z\right) = 1/\left(1 + \exp\left(-z\right)\right)$, et $g\left(z\right) = \exp\left(-z\right)/\left(1 + \exp\left(-z\right)\right)^2 = G\left(z\right)\left(1 - G\left(z\right)\right)$. On a donc simplement

$$\left.\frac{\partial \log L_N}{\partial\, b}\right|_{Logit} = \sum_{i=1}^{N} \left[y_i - F\left(x_i \widehat{b}\right) \right] x_i^{'} = 0$$

Pour le modèle Probit on a $G\left(z\right) = \Phi\left(z\right)$, et $g\left(z\right) = \varphi\left(z\right)$. On a donc simplement

$$\left.\frac{\partial \log L_N}{\partial\, b}\right|_{\Pr obit} = \sum_{i=1}^{N} \left[y_i - \Phi\left(x_i \widehat{b}\right) \right] \frac{\varphi\left(x_i \widehat{b}\right)}{\Phi\left(x_i \widehat{b}\right)\left[1 - \Phi\left(x_i \widehat{b}\right)\right]} x_i^{'} = 0$$

Dérivées secondes de la log-vraisemblanc

– *Pour le modèle logit* : On trouve directement

$$H = \left.\frac{\partial^2 \log L_N}{\partial b \partial b'}\right|_{Logit} = -\sum_{i=1}^{N} \left[1 - F\left(x_i \widehat{b}\right) \right] F\left(x_i \widehat{b}\right) x_i x_i^{'}$$

La matrice hessienne est toujours négative : la fonction de log-vraisemblance est donc globalement concave. La méthode de Newton permettra de converger vers l'optimum en quelques itérations.

– *D'une façon générale*, on peut montrer que si $\log(g)$ est concave, alors le hessien est négatif. En effet, on peut réécrire la log vraisemblance en séparant les observations pour lesquelles $y_i = 1$ de celles pour lesquelles $y_i = 0$, on note I_1 et I_0 les ensembles d'individus correspondants. En notant $g_i = g(x_i b)$ et $G_i = G(x_i b)$, on a alors

$$\begin{aligned}\frac{\partial \log L_N}{\partial\, b} &= \sum_{i=1}^{N} [y_i - G_i] \frac{g_i}{G_i [1 - G_i]} x_i^{'} \\ &= \sum_{I_1} [1 - G_i] \frac{g_i}{G_i [1 - G_i]} x_i^{'} + \sum_{I_0} [0 - G_i] \frac{g_i}{G_i [1 - G_i]} x_i^{'} \\ &= \sum_{I_1} \frac{g_i}{G_i} x_i^{'} + \sum_{I_0} -\frac{g_i}{1 - G_i} x_i^{'}\end{aligned}$$

On a alors :

$$\frac{\partial^2 \log L_N}{\partial b \partial b'} = \sum_{I_1} \left(\frac{g_i}{G_i}\right)^{'} x_i^{'} x_i + \sum_{I_0} \left(-\frac{g_i}{1 - G_i}\right)^{'} x_i^{'} x_i$$

et $\left(\frac{g_i}{G_i}\right)^{'} = \frac{g_i^{'} G_i - g_i^2}{G_i^2}$ et $\left(-\frac{g_i}{[1-G_i]}\right)^{'} = \frac{-g_i^{'}(1-G_i) - g_i^2}{(1-G_i)^2}$. Comme g est symétrique $G(-z) = 1 - G(z)$, donc $\frac{g(-z)}{G(-z)} = \frac{g(z)}{1-G(z)}$, il s'ensuit que $\frac{d}{dz}\left(\frac{g(z)}{1-G(z)}\right) = \frac{d}{dz}\left(\frac{g(-z)}{G(-z)}\right) = -\frac{d}{dz}\left(\frac{g}{G}\right)\Big|_{-z}$, si $\frac{g}{G}$ est une fonction décroissante, alors $-\frac{g(z)}{1-G(z)}$ est aussi une fonction décroissante. Pour montrer que le Hessien est négatif il suffit de montrer que $\frac{g}{G}$ est décroissante, c'est à dire si $g'G < g^2$ soit encore $\frac{g^{'}}{g} G < g$. $\log(g)$ est concave est équivalent à $\frac{g^{'}}{g}$ dé croissante. Dans ce cas $g'(t) = \frac{g'(t)}{g(t)} g(t) > \frac{g'(z)}{g(z)} g(t)$ pour $t \leq z$ donc $\int_{-\infty}^{z} g'(t)\, dt > \frac{g'(z)}{g(z)} \int_{-\infty}^{z} g(t)$ soit $g(z) > \frac{g'(z)}{g(z)} G(z)$.
Dans le cas Probit, $g(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right)$, on a donc $\log g(z) = -\log\sqrt{2\pi} - \frac{1}{2} z^2$, qui est bien une fonction concave. L'objectif est donc globalement concave.

Remarque. Compte tenu de $\phi^{'}(z) = -z\varphi(z)$ on en déduit $z + \frac{\varphi}{\Phi}(z) > 0$ et aussi $-z + \frac{\varphi}{1-\Phi}(z) > 0$.

Matrice de variance-covariance de $\widehat{b}$ La matrice de variance-covariance asymptotique est égale à

$$V_{as}\left(\widehat{b}\right) = \left[-E\left(\frac{\partial^2 \log L}{\partial b \partial b'}\right)\right]^{-1} = \left[E\left(\frac{\partial \log L}{\partial b} \frac{\partial \log L}{\partial b'}\right)\right]^{-1}$$

Elle peut être estimée à partir des dérivée secondes évaluées en $\widehat{b}$:

$$\hat{V}_{as}(\widehat{b}) = \left(-\overline{\frac{\partial^2 \log L\left(y_i, x_i, \widehat{b}\right)}{\partial b \partial b'}}\right)^{-1}$$

ou des dérivées premières évaluée en $\hat{\beta}$:

$$\hat{V}_{as}(\widehat{b}) = \left(\overline{\frac{\partial \log L\left(y_i, x_i, \widehat{b}\right)}{\partial b} \left(\frac{\partial \log L\left(y_i, x_i, \widehat{b}\right)}{\partial b} \right)'} \right)^{-1}$$

On note que dans ce cas la matrice de variance s'écrit sous une forme connue, s'apparentant à celle des mcqg $\hat{V}_{as}(\widehat{b}) = \left(\overline{\widehat{\omega}_i^2 \widehat{\varepsilon}_i^2 x_i' x_i}\right)^{-1}$, où $\widehat{\varepsilon}_i = y_i - G\left(x_i, \widehat{b}\right)$ et $\widehat{\omega}_i = \frac{g\left(x_i \widehat{b}\right)}{G\left(x_i \widehat{b}\right)\left[1 - G\left(x_i \widehat{b}\right)\right]}$. La matrice de variance covariance de l'estimateur est dans tous les cas estimée par

$$\hat{V}(\widehat{b}) = \hat{V}_{as}(\widehat{b})/N$$

11.2 Modèles de choix discrets : le Modèle Logit Multinomial

Supposons qu'un individu i ait à choisir, parmi un ensemble de K modalités, une et une seule de ces modalités, notée k.

Exemple.
- choix du lieu de vacances (montagne, mer, campagne) ;
- choix du moyen de transport domicile-travail (bus, auto, metro) ;
- choix d'un article particulier pour les décisions d'achat de biens différenciés (type de voiture, marque de céréale, type de télé viseur...).

Pour modéliser cette situation on associe à chaque modalité un niveau d'utilité

$$U_{ik} = \mu_{ik} + \varepsilon_{ik} = x_i b_k + \varepsilon_{ik} \quad k = 1, ...K$$

où ε_{ik} est une variable aléatoire non observable. L'individu choisit la modalité que lui procure l'utilité maximal.

$$y_i = Arg \max_k \left(U_{ik}\right)$$

Théorème 11.1 (Mac Fadden, 1974). *Si les $\{\varepsilon_{ik}\}_{k=1,...K}$ sont des v.a. indépendantes et identiquement distribuées selon une loie des valeurs extrêmes de fonction de répartition.*

$$G(x) = \exp[-\exp(-x)],$$

alors la probabilité de choisir la modalité k s'écrit :

$$P[Y_i = k] = \frac{\exp(\mu_{ik})}{\Sigma_{l=1}^{K} \exp\left(\mu_{il}\right)} = \frac{\exp(x_i b_k)}{\Sigma_{l=1}^{K} \exp\left(x_i b_l\right)}$$

Ce modèle est appelé modèle logit multinomial.

Démonstration. Notons g la fonction de densité des ε :

$$g\left(z\right) = G^{'}\left(z\right) = \frac{d}{dz} \exp\left[-\exp\left(-z\right)\right] = \exp\left(-z\right) G\left(z\right)$$

On peut écrire par exemple la probabilité de choisir la première solution

$$\begin{aligned} P(y=1) &= P(U_2 < U_1, \ldots, U_K < U_1) \\ &= P(\mu_2 + \varepsilon_2 < \mu_1 + \varepsilon_1, \ldots, \mu_K + \varepsilon_K < \mu_1 + \varepsilon_1) \\ &= \int_{-\infty}^{+\infty} P(\mu_2 + \varepsilon_2 < \mu_1 + \varepsilon_1, \ldots, \mu_K + \varepsilon_K < \mu_1 + \varepsilon_1 \,|\varepsilon_1)\, g(\varepsilon_1)\, d\varepsilon_1 \end{aligned}$$

Comme les aléas sont indépendants, on a

$$\begin{aligned} & P(\mu_2 + \varepsilon_2 < \mu_1 + \varepsilon_1, \ldots, \mu_K + \varepsilon_K < \mu_1 + \varepsilon_1 \,|\varepsilon_1) \\ = & \prod_{k=2}^{K} P(\mu_k + \varepsilon_k < \mu_1 + \varepsilon_1 \,|\varepsilon_1) = \prod_{k=2}^{K} G(\mu_1 - \mu_k + \varepsilon_1) \\ = & \prod_{k=2}^{K} \exp[-\exp(-\mu_1 + \mu_k - \varepsilon_1)] = \exp\left[-\sum_{k=2}^{K} \exp(-\mu_1 + \mu_k - \varepsilon_1)\right] \\ = & \exp\left[-\exp(-\varepsilon_1) \sum_{k=2}^{K} \exp(\mu_k - \mu_1)\right] \end{aligned}$$

Donc

$$\begin{aligned} P(y=1) &= \int_{-\infty}^{+\infty} \exp\left[-\exp(-\varepsilon_1) \sum_{k=2}^{K} \exp(\mu_k - \mu_1)\right] g(\varepsilon_1)\, d\varepsilon_1 \\ &= \int_{-\infty}^{+\infty} \exp\left[-\exp(-\varepsilon_1) \sum_{k=2}^{K} \exp(\mu_k - \mu_1)\right] \exp(-\varepsilon_1)\, G(\varepsilon_1)\, d\varepsilon_1 \\ &= \int_{-\infty}^{+\infty} \exp\left[-\exp(-\varepsilon_1) \left(\sum_{k=2}^{K} \exp(\mu_k - \mu_1) + 1\right)\right] \exp(-\varepsilon_1)\, d\varepsilon_1 \\ &= \int_{-\infty}^{+\infty} \exp\left[-\exp(-\varepsilon_1) \sum_{k=1}^{K} \exp(\mu_k - \mu_1)\right] \exp(-\varepsilon_1)\, d\varepsilon_1 \end{aligned}$$

puisque $G(\varepsilon_1) = \exp[-\exp(-\varepsilon_1)]$ et $\exp(\mu_1 - \mu_1) = 1$. Si on définit $P_1 = \left[\sum_{k=1}^{K} \exp(\mu_k - \mu_1)\right]^{-1}$, on a

$$P(y=1) = \int_{-\infty}^{+\infty} \exp[-\exp(-\varepsilon_1)/P_1] \exp(-\varepsilon_1)\, d\varepsilon_1$$

On fait le changement de variable $v = \exp(-\varepsilon_1)/P_1$. On a $dv = -\exp(-\varepsilon_1)\, d\varepsilon_1/P_1$, d'où

$$P(y=1) = -\int_{\exp(-(-\infty))/P_1}^{\exp(-(+\infty))/P_1} \exp(-v)\, P_1 dv = -\int_{\infty}^{0} \exp(-v)\, P_1 dv = P_1$$

■

Remarque. 1. Les probabilités ne dépendent que des différences

$$\mu_l - \mu_k = x(b_l - b_k), \;\; l \neq k$$

Elles ne sont pas modifiés si tous les b_l sont translatés en $\widetilde{b}_l = b_l + c$.

2. En conséquence, les b_k sont non identifiables sauf à poser par exemple $b_1 = 0$
3. Les paramètres estimés s'interprétent alors comme des écarts à la réfé rence b_1. Un signe positif signifie que la variable explicative accroît la probabilité de la modalité associée relativement à la probabilité de la modalité de référence.

11.2.1 Estimation du modèle logit multinomial :

Posons

$$\begin{aligned} y_{ki} &= 1\left(y_i = k\right) \\ P_{ki} &= P\left(y_i = k \left| x_i \right.\right) = \frac{\exp(x_{ki} b_k)}{\sum_{l=1}^{K} \exp(x_{li} b_l)} \\ b_1 &= 0 \end{aligned}$$

La log-vraisemblance de l'échantillon s'écrit :

$$\log L = \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} \log P_{ik}$$

Cette fonction est *globalement concave*. Les conditions du premier ordre pour la détermination du paramètre $b^{'} = (b_2, ..., b_K)^{'}$, s'écrivent simplement sous la forme

$$\frac{\partial \log L}{\partial b} = \sum_{i=1}^{n} \begin{pmatrix} \left(y_{i2} - P_{i2}\right) x_{2i} \\ \vdots \\ \left(y_{iK} - P_{iK}\right) x_{Ki} \end{pmatrix} = 0$$

Démonstration. Déterminons d'abord le gradient. On redéfinit les probabilité à partir d'un vecteur de variables observables spécifique à chaque modalité auquel s'applique le vecteur de paramètre b complet. C'est à dire tel que $\widetilde{x}_{ik} b = x_{ik} b_k$, $\widetilde{x}_{ik} = (0, ..., 0, x_{ik}, 0, ..., 0)$ x_{ik} est un vecteur ligne dont le nombre de colonne est n_{b_k}, la dimension de b_k, tandis que $\widetilde{x}_{ik}$ est un vecteur dont la dimension est celle de l'ensemble des paramètres, c'est à dire $n_{b_2} + \cdots + n_{b_K}$. Les probabilité s'écrivent donc sous la forme $P_{ki} = P\left(y_i = k \left| x_i \right.\right) = \frac{\exp(\widetilde{x}_{ki} b)}{\sum_{l=1}^{K} \exp(\widetilde{x}_{li} b_l)}$ et on a $\widetilde{x}_{1i} = 0$. La condition du premier ordre est donnée par

$$\frac{\partial \log L}{\partial b} = \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} \frac{\partial}{\partial b} \log P_{ik} = 0$$

et on a d'une part

$$\begin{aligned} \frac{\partial \log P_{ik}}{\partial b} &= \frac{\partial}{\partial b} \left[(\widetilde{x}_{ik} b) - \log \sum\nolimits_{l=1}^{K} \exp(\widetilde{x}_{il} b) \right] \\ &= \widetilde{x}_{ik} - \frac{\sum_{l=1}^{K} \frac{\partial}{\partial b} \exp(\widetilde{x}_{il} b)}{\left(\sum_{l=1}^{K} \exp(x_i b_l) \right)} \\ &= \widetilde{x}_{ik} - \sum\nolimits_{l=1}^{K} P_{il} \widetilde{x}_{il} = \widetilde{x}_{ik} - \overline{x}_i \end{aligned}$$

avec $\overline{x}_i = \sum_{l=1}^{K} P_{il}\widetilde{x}_{il}$, comme $\sum_{l=1}^{K} P_{il} = 1$, $\overline{x}_i$ représente une moyenne des observations pour l'individu i. Le gradient s'écrit donc

$$\begin{aligned}
\frac{\partial \log L}{\partial b} &= \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\left(\widetilde{x}_{ik} - \sum\nolimits_{l=1}^{K} P_{il}\widetilde{x}_{il}\right) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\widetilde{x}_{ik} - \sum\nolimits_{l=1}^{K}\sum_{k=1}^{K} y_{ik}P_{il}\widetilde{x}_{il} \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\widetilde{x}_{ik} - \sum\nolimits_{l=1}^{K} P_{il}\widetilde{x}_{il} \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \left(y_{ik} - P_{ik}\right)\widetilde{x}_{ik}
\end{aligned}$$

On voit en outre que

$$\begin{aligned}
\frac{\partial^2 \log L}{\partial b \partial b'} &= \sum_{i=1}^{n}\sum_{k=1}^{K} \frac{\partial}{\partial b'}\left(y_{ik} - P_{ik}\right)\widetilde{x}_{ik} = -\sum_{i=1}^{n}\sum_{k=1}^{K} \frac{\partial P_{ik}}{\partial b'}\widetilde{x}_{ik} \\
&= -\sum_{i=1}^{n}\sum_{k=1}^{K} P_{ik}\left(\widetilde{x}_{ik} - \overline{x}_i\right)'\widetilde{x}_{ik}
\end{aligned}$$

Comme $\overline{x}_i = \sum_{k=1}^{K} P_{ik}\widetilde{x}_{ik}$, $\sum_{k=1}^{K} P_{ik}\left(\widetilde{x}_{ik} - \overline{x}_i\right) = 0$ et donc aussi

$$\sum\nolimits_{k=1}^{K} P_{ik}\left(\widetilde{x}_{ik} - \overline{x}_i\right)\overline{x}_i = 0$$

On a donc

$$\frac{\partial^2 \log L}{\partial b \partial b'} = -\sum_{i=1}^{n}\sum_{k=1}^{K} P_{ik}\left(\widetilde{x}_{ik} - \overline{x}_i\right)'\left(\widetilde{x}_{ik} - \overline{x}_i\right)$$

Comme $P_{ik}\left(\widetilde{x}_{ik} - \overline{x}_i\right)'\left(\widetilde{x}_{ik} - \overline{x}_i\right)$ est une matrice semi définie positive le Hessien est une somme de matrice semie définie positive. Pour que $\frac{\partial^2 \log L}{\partial b \partial b'}\alpha = 0$, il faut que pour tout i et pour tout k on ait $P_{ik}\left(\widetilde{x}_{ik} - \overline{x}_i\right)\alpha = 0$ décomposant le vecteur $\alpha' = (\alpha_2, ..., \alpha_K)'$ et compte tenu de $\overline{x}_i = \sum_{k=1}^{K} P_{ik}\widetilde{x}_{ik}$, $\overline{x}_i = \left(P_{i2}x_{i2}, ..., P_{iK}x_{iK}\right)$, $P_{ik}\left(\widetilde{x}_{ik} - \overline{x}_i\right)\alpha = 0$ est équivalent à $P_{ik}\left(1 - P_{ik}\right)x_{ik}\alpha_k = 0$ pour tout i et pour tout k. ■ Ce modèle trés simple et trés facile à estimer est susceptible de généralisations importantes permettant notamment de prendre en compte l'existence de caractéristiques inobservées des individus. Le développement et l'estimation de ce type de modèle est aujourd'hui un thème de recherche trés actif aux nombreuses applications.

11.3 Sélectivité, le modèle Tobit

On prend l'exemple des équations de salaire.

Chaque individu peut travailler et percevoir alors un salaire w_i^*, et en retire une utilité $U\left(w_i^*\right)$, il peut aussi décider de s'abstenir de travailler son utilité est alors c. Sa décision de participer au marché du travail sera donc fonction de

l'écart $p_i^* = U(w_i^*) - U(b_i)$. Les deux variables latentes du modèle : w_i^* et p_i^* sont toutes deux observées partiellement. Plus précisément, on observe

$$\left\{ \begin{array}{ll} \left\{ \begin{array}{ll} w_i = w_i^* & \\ p_i = 1 & \text{si } p_i^* > 0 \end{array} \right. & \\ p_i = 0 & \text{si } p_i^* \leq 0 \end{array} \right.$$

On peut associer une modélisation à chacune de ces variables latentes :

$$\begin{aligned} w_i^* &= x_{wi} b_w + u_{wi} \\ p_i^* &= x_{pi} b_p + u_{pi} \end{aligned}$$

L'estimation de ce type de modèle est en général complexe lorsque l'on ne spécifie pas la loi des résidus. On examine ici la situation dans laquelle la loi jointes des deux résidus u_{wi} et u_{pi}, conditionnellement aux variables explicatives, est une loi normale bivariée :

$$\begin{pmatrix} u_{wi} \\ u_{zi} \end{pmatrix} \rightsquigarrow N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_w^2 & \rho\sigma_w\sigma_p \\ \rho\sigma_w\sigma_p & \sigma_p^2 \end{pmatrix} \right]$$

Une caractérisitique importante de cette modélisation est de laisser possible une corrélation entre les deux équations de salaire et de participation. Un tel modèle porte le nom de *Modèle Tobit*

Les données dans un tel modèle sont dites tronquées. Cette troncature est susceptible de conduire à des biais importants. A titre d$'$ exemple, on considère la situation

$$\left\{ \begin{array}{l} y_1^* = x + u_1 \\ y_2^* = x + u_2 \end{array} \right.$$

Les variables x, u_1 et u_2 sont toutes trois normales, centrée et réduites. x est choisie indépendante de u_1 et u_2. En revanche on envisage deux situations polaires pour la corrélation de u_1 et u_2 : corrélation nulle et corrélation de 0.9. On s'intéresse à la relation entre y_1 et x, et on considère deux cas. Dans le premier cas on observe y_1^* et x sans restriction, dans le second cas on observe y_1^* et x uniquement pour y_2^* positif. Les graphiques suivant montrent les nuages de points observés :

On voit que les nuages de points dans les échantillons non tronqués se ressemblent beaucoup que la corrélation soit nulle ou de 0.9. Les droites de régression linéaire donnent toutes deux des coefficients proches des vraies valeurs : 1 pour la variable x et 0 pour la constante. On voit aussi que la troncature par la variable y_2^* ne change pas beaucoup l'allure de l'échantillon dans le cas de la corrélation nulle. On observe néanmoins que comme on a sélectionné les observations pour lesquelles $x + u_2 > 0$, on a eu tendance à retenir plus de valeurs élevées de x. Néanmoins, cette sélections des variables explicatives n'affecte pas la proprié té d'indépendance des variables explicatives et du résidu dans l'équation de y_1. On vérifie que les coefficients de la droite de régression sont la encore trés proches des vraies valeurs. En revanche les changements pour le cas $\rho = 0.9$ en présence de troncature sont trés importants. On a été amené à ne retenir que les observations pour lesquelles $x + u_2 > 0$ là encore on a eu tendance à retenir plus souvent les observations de x avec des valeurs élevées. Pour une observation retenue pour une valeur de x donnée, on n'a retenue que les observations avec une valeur importante de u_2 et donc de u_1 puisque ces variables sont fortement

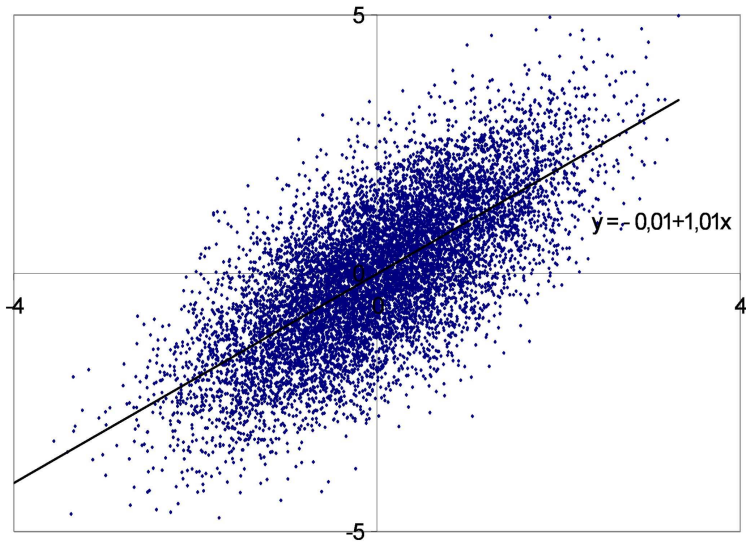

FIG. 1 – Complet $\rho = 0$

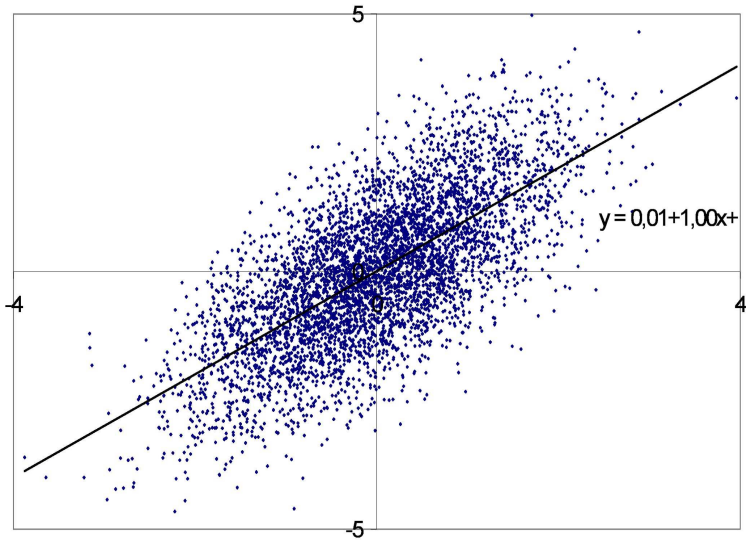

FIG. 2 – Complet $\rho = 0,9$

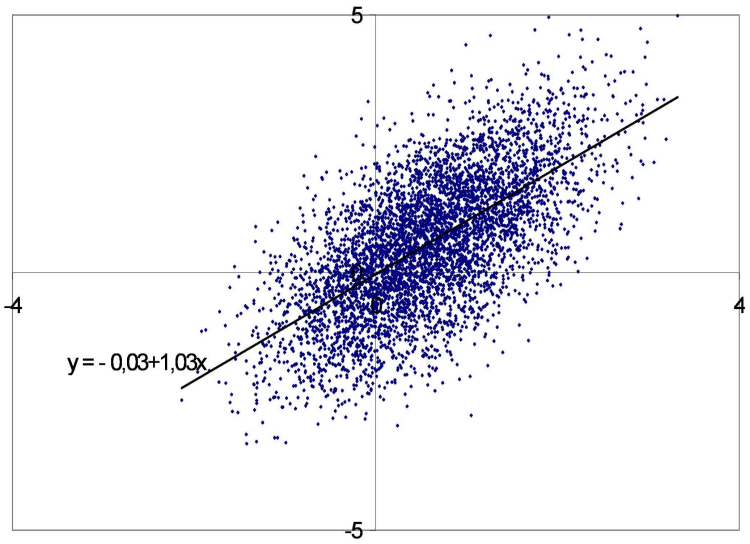

FIG. 3 – Tronqué $\rho = 0$

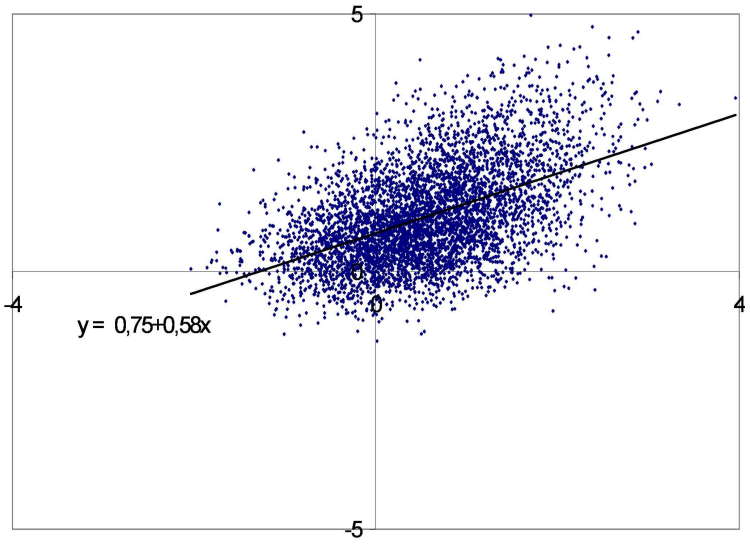

FIG. 4 – Tronqué $\rho = 0,9$

corrélées. On en déduit que à x donné, on a retenu des observations pour lesquelles u_1 est suffisament important. Pour une valeur donnée de x la moyenne des résidus des observations sélectionnées sera donc positive contrairement à ce qu'implique l'hypothèse d'indépendance. En outre, si on considère une valeur de x plus importante, on sera amené à sélectionner des observations de u_2 de façon moins stricte, et la moyenne des résidus de u_1 sélectionnés sera donc toujours positive, mais plus faible. On en déduit que l'espérance des résidus conditionnelle à une valeur donnée de x est une fonction décroissante de x : le résidu de l'équation de y_1 sur les observations sélectionnés ne sont plus indépendants de la variable explicative. Ce résultat se matérialise par une droite de régression de pente beaucoup plus faible que dans le cas précédent : le biais dit de sélectivité est ici trés important. Une autre conséquence que l'on peut voir sur le graphique et qui est intimement liée dans ce cas à la sélection, est que la relation entre y_1 et x est hétéroscédastique.

11.3.1 Rappels sur les lois normales conditionnelles.

Densité La densité d'une loi normale centrée réduite est notée φ et a pour expression

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$

La fonction de répartition est notée $\Phi(u) = \int_{-\infty}^{u} \varphi(t)\, dt$. Compte tenu de la symétrie de la fonction φ on a $\Phi(-u) = 1 - \Phi(u)$

Une variable aléatoire de dimension k suivant une loi normale mutivariée de moyenne μ et de variance $\Sigma : y \sim N(\mu, \Sigma)$, a pour densité :

$$f(y) == \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp\left(-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right)$$

On considère une loi normale bivariée

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \rightsquigarrow N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right]$$

la densité de la loi jointe de u_1 et u_2 est donc donnée par

$$f(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{(\varepsilon_1^2 + \varepsilon_2^2 - 2\,\rho\,\varepsilon_1\varepsilon_2)}{2(1-\rho^2)}\right]$$

avec $\varepsilon_1 = \frac{y_1-\mu_1}{\sigma_1}$ et $\varepsilon_2 = \frac{y_2-\mu_2}{\sigma_2}$.

La loi marginale de y_1 est donnée par

$$f(u_1) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{1}{2}\varepsilon_1^2\right)$$

un calcul simple permet de montrer que la loi y_2 conditionnelle à y_1 donnée par $f(y_2|y_1) = \frac{f(y_1,y_2)}{f(y_1)}$ est aussi une loi normale, mais de moyenne et de variance différente. La moyenne dépend de la valeur prise par y_1, mais pas la variance :

$$f(y_2|y_1) \rightsquigarrow N\left(\mu_2 + \frac{\sigma_2\rho}{\sigma_1}(y_1 - \mu_1), \sigma_2^2(1-\rho^2)\right)$$

Moments d'une loi normale tronquée Soit $u \sim N(0,1)$, elle a pour densité $\varphi(u)$. Compte tenu de $\varphi'(u) = -u\varphi(u)$, on a :

$$\begin{aligned} E(u|u>c) &= \frac{\int_c^\infty u\varphi(u)du}{1-\Phi(c)} = \frac{[-\varphi(u)]_c^\infty}{1-\Phi(c)} = \frac{\varphi(c)}{1-\Phi(c)} = \frac{\varphi(-c)}{\Phi(-c)} \\ &= M(-c) \end{aligned}$$

de même

$$E(u|u<c) = \mu - E(-(u-\mu)\,|\,-u>-c) = -M(c)$$

Et les moments d'ordre 2

$$E\left(u^2|u>c\right) = \frac{\int_c^\infty u^2\varphi(u)du}{1-\Phi(c)} = 1 + cM(-c)$$

où on intègre par partie $\int_c^\infty u^2\varphi(u)du = [-u\varphi(u)]_c^\infty + \int_c^\infty \varphi(u)du$. On en déduit la variance conditionnelle

$$V(u|u>c) = E(u^2|u>c) - [E(u|u>c)]^2 = 1 + cM(-c) - M(-c)^2$$

de façon similaire on a pour la loi normale tronquée supérieurement

$$\begin{aligned} E(u^2|u &< c) = E((-u)^2\,|\,-u>-c) = 1 - cM(c) \\ V(u|u &< c) = 1 - cM(c) - M(c)^2 \end{aligned}$$

Remarque on a vu précédemment que l'on avait pour une loi normale $z + \frac{\phi}{\Phi}(z) > 0$ et aussi $-z + \frac{\phi}{1-\Phi}(z) > 0$ soit encore $zM(z) + M(z)^2 > 0$ et $zM(-z) - M(-z)^2 < 0$ on en déduit que l'on a toujours comme on s'y attend $V(u|u \lesseqgtr c) < 1$.

Dans le cas d'une variable non centrée réduite $v \sim N\left(\mu, \sigma^2\right)$, on peut déduire des résultats précédents les moments des lois tronquées en notant que $(v-\mu)/\sigma$ et que $v \lesseqgtr c \Leftrightarrow u = (v-\mu)/\sigma \lesseqgtr \widetilde{c} = (c-\mu)/\sigma$. on a donc

$$\begin{aligned} E(v|v &> c) = E(\sigma u + \mu | u > \widetilde{c}) = \mu + \sigma M\left(-\frac{c-\mu}{\sigma}\right) \\ E(v|v &< c) = E(\sigma u + \mu | u < \widetilde{c}) = \mu - \sigma M\left(\frac{c-\mu}{\sigma}\right) \end{aligned}$$

En calculant $E(v^2|v>c) = E(\sigma^2u^2 + 2u\sigma\mu + \mu^2|u>\widetilde{c})$, on trouve sans peine l'expression de la variance

$$V(v|v>c) = \sigma^2\left(1 + \frac{c-\mu}{\sigma}M\left(-\frac{c-\mu}{\sigma}\right) - M\left(-\frac{c-\mu}{\sigma}\right)^2\right)$$

Pour les moments de la loi tronquée supérieurement on a également

$$V(v|v<c) = \sigma^2\left(1 - \frac{c-\mu}{\sigma}M\left(\frac{c-\mu}{\sigma}\right) - M\left(\frac{c-\mu}{\sigma}\right)^2\right)$$

On a aussi comme on s'y attend pour toute transformation linéaire

$$\begin{aligned} V(a+bv|v &> c) = b^2 V(v|v>c) \\ V(a+bv|v &< c) = b^2 V(v|v<c) \end{aligned}$$

Moments d'une variable normale tronquée par une autre variable normale On s'intéresse au cas d'une variable aléatoire suivant une loi normale bivariée

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \rightsquigarrow N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right]$$

et on cherche les moments d'ordre 1 et 2 de la variable y_2 tronquée par $y_1 > 0$. On a vu que la loi de y_2 conditionnelle à y_1 est une loi normale de moyenne $\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(y_1 - \mu_1)$ et de variance $\sigma_2^2\left(1-\rho^2\right)$. On en déduit que

$$\begin{aligned} E(y_2|y_1>0) &= E\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(y_1-\mu_1)|y_1>0\right) \\ &= \mu_2 + \rho\sigma_2 E\left(\frac{y_1-\mu_1}{\sigma_1}|y_1>0\right) \\ &= \mu_2 + \rho\sigma_2 E\left(\frac{y_1-\mu_1}{\sigma_1}\middle|\frac{y_1-\mu_1}{\sigma_1} > -\frac{\mu_1}{\sigma_1}\right) \\ &= \mu_2 + \rho\sigma_2 M\left(\frac{\mu_1}{\sigma_1}\right) \end{aligned}$$

De même,

$$\begin{aligned} V(y_2|y_1>0) &= V(E(y_2|y_1)|y_1>0) + E(V(y_2|y_1)|y_1>0) \\ &= V\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(y_1-\mu_1)|y_1>0\right) + \left(1-\rho^2\right)\sigma_2^2 \\ &= \rho^2\sigma_2^2 V\left(\frac{y_1-\mu_1}{\sigma_1}\middle|\frac{y_1-\mu_1}{\sigma_1} > -\frac{\mu_1}{\sigma_1}\right) \\ &= \rho^2\sigma_2^2\left(1 - \frac{\mu_1}{\sigma_1}M\left(\frac{\mu_1}{\sigma_1}\right) - M\left(\frac{\mu_1}{\sigma_1}\right)^2\right) + \left(1-\rho^2\right)\sigma_2^2 \\ &= \sigma_2^2 - \rho^2\sigma_2^2\left(\frac{\mu_1}{\sigma_1}M\left(\frac{\mu_1}{\sigma_1}\right) + M\left(\frac{\mu_1}{\sigma_1}\right)^2\right) \end{aligned}$$

Compte tenu du résultat précédent sur la loi normale unidimensionnelle et puisque $V(y_2|y_1) = \left(1-\rho^2\right)\sigma_2^2$.

On obtient directement les moments de la loi normale y_2 tronquée par $y_1 < 0$ en remplaçant μ_1 par $-\mu_1$ et ρ par $-\rho$

$$E(y_2|y_1<0) = \mu_2 - \rho\sigma_2 M\left(-\frac{\mu_1}{\sigma_1}\right)$$

De même,

$$V(y_2|y_1<0) = \sigma_2^2 - \rho^2\sigma_2^2\left(-\frac{\mu_1}{\sigma_1}M\left(-\frac{\mu_1}{\sigma_1}\right) + M\left(-\frac{\mu_1}{\sigma_1}\right)^2\right)$$

11.3.2 Pourquoi ne pas estimer un modèle Tobit par les MCO ?

– Si on se restreint aux observations positives, on a

$$E\left(w_i \left| x_{wi}, x_{pi}, p_i = 1\right.\right) = E\left(w_i^* \left| x_{wi}, x_{pi}, p_i^* > 0\right.\right)$$

En appliquant les résultats précédents à $y_2 = w^*$, et $y_1 = p^*$

$$\begin{aligned} E\left(w_i^* \left| x_{wi}, x_{pi}, p_i^* > 0\right.\right) &= \mu_w + \rho\sigma_w M\left(\frac{\mu_p}{\sigma_p}\right) \\ &= x_{wi}b_w + \rho\sigma_w M\left(\frac{x_{pi}b_p}{\sigma_p}\right) \end{aligned}$$

On voit donc que dés lors que la corrélation entre les éléments inobservés de l'équation de salaire et de l'équation de participation sont corrélés, ne pas prendre en compte la séléctivité recient à oublier une variable dans la r égression : $M\left(\frac{x_{pi}b_p}{\sigma_p}\right)$ aussi appelé ratio de Mills. Cet oubli est donc susceptible de conduire à une estimation biaisée des paramètres dés lors que les variables $M\left(\frac{x_{pi}b_p}{\sigma_p}\right)$ et x_{wi} sont corrélées. Si on considère à titre illustratif que l'équation de séléction s'écrit $w_i^* > \overline{w}$, on a $\rho = 1$ et $\frac{x_{pi}b_p}{\sigma_p} = \frac{x_{wi}b_w - \overline{w}}{\sigma_w}$. L'équation précédente s'écrit alors

$$E\left(w_i^* \left| x_{wi}, x_{pi}, p_i^* > 0\right.\right) = x_{wi}b_w + \sigma_w M\left(\frac{x_{wi}b_w - \overline{w}}{\sigma_w}\right)$$

Dans ce cas comme $M(z) = \frac{\varphi(z)}{\Phi(z)}$ est une fonction décroissante de z le biais est négatif. Dans le cas général tout dépent de ρ et de la corrélation entre le ratio de Mills et $M\left(\frac{x_{pi}b_p}{\sigma_p}\right)$ les variables explicative entrant dans la modélisation de w_i^*.

– Si on introduit également les observations pour lesquelles $w_i = 0$, on a

$$\begin{aligned} E\left(w_i \left| x_{wi}, x_{pi}\right.\right) &= E\left(w_i \left| x_{wi}, x_{pi}, p_i = 1\right.\right) P\left(p_i = 1 \left| x_{wi}, x_{pi}\right.\right) + \\ &\quad E\left(w_i \left| x_{wi}, x_{pi}, p_i = 0\right.\right) P\left(p_i = 0 \left| x_{wi}, x_{pi}\right.\right) \\ &= E\left(w_i \left| x_{wi}, x_{pi}, p_i = 1\right.\right) P\left(p_i = 1 \left| x_{wi}, x_{pi}\right.\right) \\ &= \left(x_{wi}b_w\right)\Phi\left(\frac{x_{pi}b_p}{\sigma_p}\right) + \rho\sigma_w\varphi\left(\frac{x_{pi}b_p}{\sigma_p}\right) \end{aligned}$$

et on voit que la forme linéaire n'est pas non plus adaptée.

11.3.3 Estimation par le maximum de vraisemblance

On écrit la probabilité d'observer chaque réalisation du couple (w_i, p_i).

– Pour $p_i = 0$ on n'observe pas w_i la seule probabilté est $P\left(p_i^* < 0\right)$, c'est à dire $P\left(x_{pi}b_p + u_{pi} < 0\right) = \Phi\left(-\frac{x_{pi}b_p}{\sigma_p}\right) = 1 - \Phi\left(\frac{x_{pi}b_p}{\sigma_p}\right)$

– Pour $p_i = 1$ on observe $w_i = w_i^*$ et $p_i^* > 0$. La densité correspondante est

$$f\left(w_i^* = w_i, p_i = 1\right) = \int_{p_i^*>0} f\left(w_i, p_i^*\right) dp_i^* = f\left(w_i\right) \int_{p_i^*>0} f\left(p_i^* \left| w_i\right.\right) dp_i^*$$

et la loi de p_i^* conditionnelle à $w_i^* = w_i$ est pas d éfinition une loi normale de moyenne $\widetilde{\mu}_p(w_i) = \mu_p + \rho\sigma_p \frac{w_i - \mu_w}{\sigma_w}$ et de variance $\widetilde{\sigma}_p^2 = \sigma_p^2\left(1-\rho^2\right)$ la probabilityé pour qu'une telle variable aléatoire soit positive est $\Phi\left(\frac{\widetilde{\mu}_p(w_i)}{\widetilde{\sigma}_p}\right) = \Phi\left(\frac{\mu_p + \rho\sigma_p \frac{w_i-\mu_w}{\sigma_w}}{\sigma_p\sqrt{(1-\rho^2)}}\right)$. Finalement, la densité des observations est

$$\begin{aligned} L &= \prod_{p_i=0}\left[1-\Phi\left(\frac{x_{pi}b_p}{\sigma_p}\right)\right]\times \\ &\quad \prod_{p_i=1}\frac{1}{\sigma_w}\varphi\left(\frac{w_i - x_{wi}b_w}{\sigma_w}\right)\Phi\left(\frac{x_{pi}b_p + \rho\sigma_p\frac{w_i - x_{wi}b_w}{\sigma_w}}{\sigma_p\sqrt{(1-\rho^2)}}\right) \\ &= \prod_i\left[1-\Phi\left(\frac{x_{pi}b_p}{\sigma_p}\right)\right]^{1-p_i}\times \\ &\quad \left[\frac{1}{\sigma_w}\varphi\left(\frac{w_i - x_{wi}b_w}{\sigma_w}\right)\Phi\left(\frac{x_{pi}b_p + \rho\sigma_p\frac{w_i - x_{wi}b_w}{\sigma_w}}{\sigma_p\sqrt{(1-\rho^2)}}\right)^{p_i}\right] \end{aligned}$$

On voit que comme dans le cas du modèle Probit, on ne peut pas identifier la totalité des paramètres de l'équation de sélection : seul le paramètre $\widetilde{b}_p = \frac{b_p}{\sigma_p}$ est identifiable. Compte tenu de cette redéfinition des paramètre du modèle, la vraisemblance s'écrit :

$$\begin{aligned} L &= \prod_i\left[1-\Phi\left(x_{pi}\widetilde{b}_p\right)\right]^{1-p_i}\times \\ &\quad \left[\frac{1}{\sigma_w}\varphi\left(\frac{w_i - x_{wi}b_w}{\sigma_w}\right)\Phi\left(\frac{x_{pi}\widetilde{b}_p + \rho\frac{w_i - x_{wi}b_w}{\sigma_w}}{\sqrt{(1-\rho^2)}}\right)^{p_i}\right] \end{aligned}$$

Dans le cas où $\rho = 0$ on voit que la vraisemblance est séparable entre une contribution correspondant à l'observation de $p_i = 0/1$ et une contribution associée aux observations de w_i :

$$\begin{aligned} L &= \prod_i\left[1-\Phi\left(x_{pi}\widetilde{b}_p\right)\right]^{1-p_i}\times\Phi\left(x_{pi}\widetilde{b}_p\right)^{p_i} \\ &\quad \left[\frac{1}{\sigma_w}\varphi\left(\frac{w_i - x_{wi}b_w}{\sigma_w}\right)\right]^{p_i} \end{aligned}$$

On retrouve donc le fait que dans le cas $\rho = 0$ on peut ignorer la sélection des observation. On voit aussi que dans les cas général où $\rho \neq 0$ la sélectivité importe.

Remarque. 1. La fonction de vraisemblance n'est pas globalement concave en $\left(\rho, \sigma_w, b_w, \widetilde{b}_p\right)$.

2. Elle est concave globalement en $\theta = \left(\sigma_w, b_w, \widetilde{b}_p\right)$ *pour* ρ *fixé.*

3. Une solution consiste à fixer la valeur de ρ et estimer les paramètres correpondant $\widehat{\theta}(\rho)$ et à balayer sur les valeur possibles de ρ.

Estimation en deux étapes par la méthode d'Heckman

- Méthode en deux étapes dans laquelle on estime d'abord le Probit associé à $p_i = 1/0$ et ensuite une régression augmentée prenant en compte la sélectivité ;
- Il s'agit d'une méthode d'estimation convergente, mais non efficace ;
- Le calcul des écart-types associés à cette méthode est un peu compliqué ;
- Elle peut être utilisée telle qu'elle ou pour fournir des valeurs initiales pour la maximisation de la vraisemblance ;
- Elle permet une généralisation facile au cas d'autres lois que la loi normale.

1ere étape : estimation de $\widetilde{b}_p = b_p/\sigma_p$ par MV du modèle Probit (sur la partie discrète) soit

$$P\left(p_i = 1\right) = P\left(p_i^* > 0\right) = \Phi\left(x_{pi}\widetilde{b}_p\right)$$

Ceci fournit un estimateur convergent de $\widetilde{b}_p$

2ème étape : on exploite la relation :

$$E(y_{wi}|y_{pi}^\star > 0) = X_{wi}b_w + \rho\sigma_w \frac{\varphi\left(X_{pi}\widetilde{b}_p\right)}{\Phi\left(X_{pi}\widetilde{b}_p\right)}$$

La variable $\frac{\varphi\left(X_{pi}\widetilde{b}_p\right)}{\Phi\left(X_{pi}\widetilde{b}_p\right)}$ est inconnue, on la remplace par

$$\widehat{\lambda}_i = \frac{\varphi\left(X_{pi}\widehat{\widetilde{b}}_p\right)}{\Phi\left(X_{pi}\widehat{\widetilde{b}}_p\right)}$$

et on estime les paramètres b_w, et $\rho\sigma_w$ à partir de la relation :

$$y_{wi} = x_{wi}b_w + (\rho\sigma_w)\hat{\lambda}_i + v_1$$

sur les observations positives

Ces estimateurs sont asymptotiquement sans biais, mais ils ne sont pas asymptotiquement efficaces.

Pour le calcul des écart-types, deux problèmes se présentent

- Le modèle est héréroscédastique. En effet :

$$\begin{aligned} V\left(u_w\,|p_i = 1\right) &= V\left(u_w\,|p_i^* > 0\right) \\ &= \sigma_w^2 - \rho^2\sigma_w^2\left(x_i\widetilde{b}_p M\left(x_i\widetilde{b}_p\right) + M\left(x_i\widetilde{b}_p\right)^2\right) \end{aligned}$$

  dépend des variables observables
- Le paramètres $\widetilde{b}_w$ n'est pas connu et est remplacé par une estimation. Il est lui même issu d'une estimation (par le MV) que l'on peut résumé par l'annulation de la contrpartie empirique de condition d'orthogonalité

$$E\left(h_{\widetilde{b}_p}\left(p_i, x_{pi}, \widetilde{b}_p\right)\right) = 0$$

L'estimation du modèle par les mco conduit quant à elle à l'annulation de la contrepratie empirique de

$$\begin{aligned} & E\left(\begin{pmatrix} x_{wi}^{'} \\ \lambda_i\left(\widetilde{b}_p\right) \end{pmatrix}\left[w_i - x_{wi}b_w - \rho\sigma_w\lambda_i\left(\widetilde{b}_p\right)\right]1_{p_i=1}\right) \\ = \quad & E\left(h_{b_w,\rho\sigma_w}\left(p_i, w_i, x_{wi}, b_w, \rho\sigma_w\right)\right) = 0 \end{aligned}$$

Le calcul des écart-types doit se faire en considérant les formules de l'estimation par la méthode des moments généralisée associée à la totalité des conditions d'orthogonalité, c'est à dire

$$E\begin{pmatrix} h_{\widetilde{b}_p}\left(p_i, x_{pi}, \widetilde{b}_p\right) \\ h_{b_w,\rho\sigma_w}\left(p_i, w_i, x_{wi}, b_w, \rho\sigma_w\right) \end{pmatrix} = 0$$

– Cette dernière façon d'estimer le modèle est inefficace, mais elle est aussi la voie à l'estimation de modèle plus généraux dans lesquels on ne fait plus d'hypothèses sur la loi des observations. On peut montrer qu'on a en général une relation de la forme

$$E\left(w_i \left| p_i = 1, x_{wi}, P\left(x_{pi}\right)\right.\right) = x_{wi}b_w + K\left(P\left(x_{pi}\right)\right)$$

où $P\left(x_{pi}\right) = P\left(p_i = 1 \left| x_{pi}\right.\right)$ et K une fonction quelconque. Dans le cas normal, cette fonction s'écrit simplement $K\left(P\right) = \rho\sigma_w \frac{\varphi\circ\Phi^{-1}(P)}{P}$ et on a en plus $P = \Phi\left(x_{pi}\widetilde{b}_p\right)$. L'estimation de ce type de modèle est néanmoins délicate.