ECONOMETRIE LINEAIRE

Bruno Crépon

Novembre 2005

Table des matières

Chapitre 1

Introduction

1.1 Le modèle

Le modèle central auquel on s'intéresse dans ce cours est le modèle linéaire que l'on écrit en toute généralité

$$y = \alpha + \beta_1 x_1 + \cdots + \beta_K x_K + u = xb + u$$

Dans ce modèle interviennent différentes grandeurs :
- y la variable expliquée ou dépendante
- $x_1, \ldots, x_K$, K variables explicatives ou indépendantes
- u une perturbation
- $b = (\alpha, \beta_1, \cdots, \beta_K)'$ le paramètre à estimer

Parmi ces éléments les variables y et x sont observées. En revanche le paramètre b est inconnu et la perturbation u inobservée.

1.2 D'où vient le modèle ? - 1 de la théorie économique

- Le modèle vient d'abord d'idées sur les relations entre y et x.... Ces idées peuvent avoir un lien très étroit avec la théorie économique. Il peut s'agir par exemple d'une fonction de production

  $$Y = F(K, L)$$

  On pourrait estimer la fonction de production parmi toutes les fonctions possibles. On ferait alors des régressions dites non paramétriques. Le cadre que l'on considère ici est plus simple et consiste à restreindre l'ensemble des possibilités et de se placer dans un ensemble de fonctions de productions dépendant d'un nombre fini de paramètres. On retient souvent la spécification de Cobb-Douglas, ce qui implique en

particulier une restriction sur les possibilités de substitution par rapport au cadre général :

$$Y = AK^{\alpha}L^{\beta}$$

Cette spécification conduit à une relation log linéaire :

$$y = a + \alpha k + \beta l$$

qui est le modèle auquel on s'intéresse. Dans ce cadre on peut noter que la perturbation a une interprétation naturelle, il s'agit de la constante a représentant le niveau de la technologie, susceptible de varier d'une entreprise à l'autre. En revanche le modèle fait l'hypothèse qu'il y a homogénéité des autres coefficients dans la population d'entreprises.
Un autre exemple de modèle directement déduit de la théorie économique est celui des demandes de facteurs. Si on spécifie une fonction de coût $C(Q, p_X, u)$, où Q est la production, p_X le vecteur des prix et u le niveau de la technologie, la demande pour un facteur donné est donnée par le Lemme de Shephard :

$$X^{0d} = \frac{\partial C(Q, p_X, u)}{\partial p_{X_0}}$$

Comme dans le cas précédent on se restreint en général à une forme paramétrique de la fonction de coût. Une spécification standard est la fonction de coût translog avec deux facteurs, capital de coût $\exp(c)$ et travail de coût $\exp(w)$:

$$LogC = a + \alpha c + \beta w + 0.5\delta_c c^2 + \delta_{w,c} cw + 0.5\delta_w w^2 + \log(Q) - \log(u)$$

La constante représente là aussi le niveau de la technologie. Ce type de spécification conduit à des fonctions de demande spécifiant la part de chaque facteur. Par exemple pour le travail on a

$$\frac{wL}{Q} = \beta + \delta_{w,c} c + \delta_w w$$

On voit que dans cette spécification la perturbation n'a pas d'interprétation aussi naturelle que dans le cas précédent. Il faut considérer que soit le paramètre β est hétérogène, soit la part observée s'écarte de la part théorique pour des raisons non expliquées.
Le modèle peut aussi provenir d'une relation moins structurelle entre les variables. Par exemple un type d'équations très souvent estimé est l'équation de Mincer qui fait dépendre le salaire du nombre d'années d'étude et de l'expérience. Par exemple :

$$\log(w_i) = a_0 + a_s s_i + a_e e_i + u_i$$

où a_s représente le gain lié à une année d'étude supplémentaire et a_e le gain lié à une année d'expérience supplémentaire. Les paramètres économiques auxquels on

s'intéresse alors sont le rendement de l'éducation ou le rendement de l'expérience. La modélisation sous-jacente est celle du capital humain : le capital humain s'accumule d'abord durant la période des études puis durant la vie active par l'expérience, en apprenant sur le tas. Si on fait l'hypothèse d'un marché du travail concurrentiel, les différences de rémunérations entre les agents traduiront des différences dans le capital humain. On peut remarquer concernant cette équation que l'on ne s'intéresse pas seulement à expliquer les différences moyennes de revenus entre les agents mais que l'on souhaite aussi parvenir à une estimation plus ambitieuse qui puisse conduire à une interprétation causale : si on augmente la durée des études de un an d'un individu quel sera son gain en terme de rémunération ?
Un autre exemple dans lequel le modèle entretient des rapports encore plus ténus avec des paramètres structurels mais possède une interprétation causale est celui de l'incidence de la taille d'une classe sur le taux de réussite des élèves de la classe. On peut légitimement se poser la question de savoir si la réduction de la taille des classes conduit à une amélioration du taux de réussite scolaire. On peut ainsi considérer un modèle du type :

$$\tau_i = a_0 + a_t taille_i + x_i a_x + u_i$$

où τ_i représente le taux de réussite d'une classe. Dans cette spécification que l'on pourrait appeler fonction de production scolaire, on introduit un ensemble d'autres variables. En effet on se doute bien que de nombreux facteurs affectent la réussite d'une classe. Par exemple l'environnement scolaire est certainement un facteur important. On pourrait se dire que comme on ne s'intéresse pas à la variable d'environnement on ne la met pas dans la régression. D'un côté on y gagne car on n'a pas à faire l'effort de mesurer cette variable, mais d'un autre côté cette variable contribue aussi à déterminer la taille de la classe. Il est possible que dans certains milieux défavorisés la taille des classes soit plus petites. Si on ignore le rôle de l'environnement scolaire et qu'on ne l'intègre pas dans la régression, on risque de mesurer un effet de la taille de la classe qui soit un mixte de l'effet propre de la taille et de l'effet de l'environnement. Il donc important dans ce type de modèle, entretenant des rapports larges avec la théorie, d'introduire des facteurs annexes qui permettront d'isoler l'effet propre de la taille de la classe. On cherche à contrôler pour un certain nombre de facteurs extérieurs.
Enfin, on peut avoir une approche descriptive des données. Il est important de remarquer que dans ce cas les paramètres n'ont pas d'interprétation structurelle.

1.3 Les données

Les données constituent le cœur de l'économétrie. Leur recueil et leur examen descriptif constituent aussi en général une part importante de tout travail économétrique. Il y a principalement trois grands types de données :

1. Données temporelles ou longitudinales. Elles sont indicées par le temps t. On dispose ainsi de séries dites temporelles : y_t, x_t, par exemple les séries trimestrielles de la consommation et du revenu, de l'inflation... En général le nombre d'observation T est assez réduit, de l'ordre de la cinquantaine. On note en général $\underline{y}$ le vecteur $T \times 1$ $(y_1, \ldots, y_T)'$ et $\underline{x}$ la matrice $T \times (K+1)$: $(x_1', \ldots, x_T')'$ où x_t est le vecteur ligne formé des valeurs des différentes variables explicatives (dont la constante) à la date t.
2. Données en coupe. y_i, x_i. Leur indice correspond à l'identifiant d'un individu ou d'une entreprise. Ces données peuvent représenter par exemple le salaire d'un individu pour y et son diplôme, son expérience... pour les variables explicatives. Les échantillons dont on dispose sont en général de beaucoup plus grande taille : le nombre d'observation N dépasse le plus souvent la centaine et peut aller jusqu'à plusieurs dizaines de milliers. On note là encore en général $\underline{y}$ le vecteur $N \times 1$ $(y_1, \ldots, y_N)'$ et $\underline{x}$ la matrice $N \times (K+1)$: $(x_1', \ldots, x_N')'$ où x_i est le vecteur ligne formé des valeurs des différentes variables explicatives (dont la constante) pour l'individu i.
3. Données à double indice, dites de panel : y_{it}, x_{it}. On dispose d'informations sur des individus $i = 1, \ldots, N$ que l'on suit sur plusieurs périodes, $t = 1, \ldots, T$. Les NT observations z_{it} correspondent à N observations vectorielles "individuelles" $z_{i1}, \ldots z_{iT}$. On note en général $\underline{y}_i$ le vecteur $T \times 1$ $(y_{i1}, \ldots, y_{iT})'$ et $\underline{x}_i$ la matrice $T \times (K+1)$: $(x_{i1}', \ldots, x_{iT}')'$ et $\underline{y}$ le vecteur $NT \times 1$ $\left(\underline{y}_1, \ldots, \underline{y}_N\right)'$ et $\underline{x}$ la matrice $NT \times (K+1)$: $(\underline{x}_1', \ldots, \underline{x}_N')'$ où $\underline{x}_i$ est la matrice formée des valeurs des différentes variables explicatives (dont la constante) pour l'individu i aux différentes dates.

1.4 L'estimation

Estimer le modèle c'est trouver une fonction des observations $\underline{y}$ et $\underline{x}$

$$\widehat{b} = b\left(\underline{y}, \underline{x}\right)$$

dont on souhaite qu'elle vérifie certaines conditions. Par exemple l'estimateur peut être choisi tel

- qu'il soit "sans biais" $E\left(\widehat{b}\right) = \int b\left(\underline{y}, \underline{x}\right) f\left(\underline{y}, \underline{x}\right) d\underline{y} d\underline{x} = b$
- qu'il satisfasse un critère : minimisation de la somme des carrés des résidus $\widehat{b} = \arg\min \sum (y - xb)^2$; maximisation de la log-vraisemblance $\widehat{b} = \arg\max \sum \log l\left(y, x\right)$
- qu'il soit de variance minimale
- qu'il soit convergent, c'est à dire qu'il se rapproche de la vraie valeur du paramètre lorsque le nombre d'observations devient grand.

1.5 Pourquoi estimer le modèle ?

- tester l'existence d'un effet, i.e. vérifier qu'une variable x a un effet spécifique sur une variable y. Par exemple on peut s'interroger sur l'effet des taux d'intérêt sur l'investissement, c'est à dire sur l'existence d'un canal monétaire de la politique monétaire. Dans le cadre d'un modèle accélérateur profit standard, $I = \alpha \Delta Q_t + \beta\pi + \gamma r + v$, on peut s'interroger sur le fait que le coefficient du taux d'intérêt γ soit nul ou non. On s'intéresse donc à l'hypothèse $H_0 : \gamma = 0$, et on souhaite que les données permettent de répondre à cette question. De façon similaire, dans le cas de la fonction de production scolaire on peut s'interroger sur l'existence d'un effet de la taille de la classe sur le taux de réussite. On va alors s'intéresser à l'hypothèse $H_0 : a_t = 0$, et là aussi on souhaite que les données nous permettent de choisir entre oui ou non. L'estimation du modèle et la confrontation du paramètre à zéro est la voie la plus naturelle pour prendre cette décision. La question est ici de savoir si le paramètre est significatif au sens statistique du terme.
- quantifier cet effet, ce qui est utile à des fins de simulations. Par exemple dans les deux cas précédents on est aussi intéressé par donner un ordre de grandeur de l'effet à attendre d'une variation de la variable. Si on voulait par exemple prendre une décision de politique économique consistant à baisser la taille des classes, ce qui est très coûteux, on est intéressé certes à savoir si cela aura un effet non nul mais aussi à savoir l'ordre de grandeur de cet effet. S'il est très faible on ne prendra pas alors aussi facilement la décision de réduire la taille des classes. L'ordre de grandeur du paramètre est aussi important. La question est ici de savoir si le paramètre est significatif au sens économique du terme.
- prévoir. Dans le modèle $y_t = x_t\beta + u_t$, le paramètre β peut être estimé sur les observations $t = 1, \ldots, T$: $\widehat{\beta}$. Connaissant x_{T+1} on calcule la prévision de y à la date $T + 1$: $\widehat{y}_{T+1} = x_{T+1}\widehat{\beta}$

1.6 D'où vient le modèle ? - 2 de relations stochastiques

Le modèle provient aussi de relations stochastiques entre les variables. L'écriture de la relation

$$y = xb + u$$

ne constitue pas en fait un modèle économétrique. Comme on l'a vu il s'agit d'une relation plus ou moins fondée. Si on l'admet fondée, le paramètre b a un sens en lui-même. Il a une définition économique, par exemple l'élasticité de la production au capital. Pour que ce modèle soit un modèle économétrique il faut lui adjoindre une restriction stochastique. Une façon naturelle de procéder est de spécifier la loi jointe des observations $l(y, x; b)$. Ceci revient à spécifier la loi du résidu sachant les variables explicatives : $l(u \,|x)$. La

situation de base est celle dans laquelle cette loi est choisie comme une loi normale ne dépendant pas des variables x. On impose donc dans ce cas une restriction stochastique essentielle pour l'analyse économétrique

$$l(u|x) = l(u) = \varphi(u/\sigma)/\sigma$$

où φ est la densité de la loi normale. Imposer cette restriction permet de définir la densité des observations

$$l(y, x; b) = l(y|x; b)\, l(x) = \varphi((y - xb)/\sigma)\, l(x)/\sigma$$

et donc d'estimer les paramètres en appliquant par exemple la méthode du maximum de vraisemblance. L'estimateur auquel on parvient est alors celui des moindres carrés ordinaires. On peut aussi faire des hypothèses sur la loi de u sachant x qui soient moins fortes que la spécification de la loi complète. Par exemple on peut se contenter de spécifier :

$$E(u|x) = E(u) = 0$$

Cette propriété est satisfaite si on spécifie la loi conditionnelle de u sachant x comme une loi normale indépendante de x. L'inverse est faux et cette spécification est donc moins exigeante que la précédente. Elle permet, elle aussi, d'estimer le modèle. Elle implique en effet des restrictions du type $E(x'(y - xb)) = 0$ appelées intuitivement conditions d'orthogonalité dont on verra qu'elles sont suffisantes pour estimer les paramètres du modèle. On remarque à ce stade que dans cette spécification il y a d'ores et déjà un paramètre de moins : la variance des résidus n'intervient plus.

Ces restrictions stochastiques définissent un paramètre statistique. On pourrait ainsi définir autant de paramètres b qu'il y a de restrictions stochastiques envisageables, c'est à dire une infinité. On pourrait par exemple considérer le paramètre b_Z associé à des restrictions stochastiques $E(z'(y - xb_Z)) = 0$ dont on verra qu'elles aussi peuvent être utilisées souvent pour conduire à une estimation du paramètre. Il n'est pas certain que le paramètre statistique associé à une restriction stochastique coïncide avec le paramètre économique. L'estimation peut ainsi être non convergente, c'est à dire que la valeur du paramètre estimée ne se rapprochera pas de la vraie valeur (économique) du paramètre lorsque le nombre d'observation augmente, ou être biaisée, c'est à dire que l'espérance du paramètre n'est pas la vraie valeur (économique) du paramètre. Une partie importante de l'économétrie, qui passe par une réflexion sur le modèle, les données et les méthodes consiste à rechercher des conditions dans lesquelles le paramètre statistique coïncide avec le paramètre économique. La question est-ce que $p\lim \widehat{b} = b_0$, la vraie valeur économique du paramètre, est en dernier ressort la question la plus centrale et la plus importante de l'économétrie, et assez naturelle : est-ce que j'ai bien mesuré ce que je voulais ? C'est beaucoup moins facile qu'il n'y paraît, car de nombreux facteurs affectent les décisions individuelles et il est difficile d'isoler l'effet d'une unique cause.

1.7 Plan

Le cours débute dans le chapitre 2 par l'estimateur des moindres carrés, c'est à dire le vecteur des coefficients de la projection orthogonale de y sur l'espace vectoriel engendré par les variables explicatives. On présente d'abord les propriétés algébriques de cet estimateur et ses propriétés statistiques sous des hypothèses minimales telles que l'indépendance et l'équidistribution des observations (Théorème de Frish-Waugh, Théorème de Gauss-Markov, estimation des paramètres du second ordre, le R^2 et l'analyse de la variance). On montre ensuite dans le chapitre 3 comment la spécification de la loi des résidus comme une loi normale permet de compléter l'analyse en particulier en permettant d'obtenir la loi des estimateurs, étape incontournable pour procéder à des tests d'hypothèses simples (test de Student) ou définir des intervalles de confiance pour les paramètres. On examine ensuite dans le chapitre 4 et dans le même cadre où la loi des résidus est supposée normale, le cas important des estimations sous contraintes linéaires (dans les paramètres). On présente alors les tests d'hypothèses linéaires sur les paramètres par le biais des tests de Fisher. Ces résultats sont obtenus sous des hypothèses fortes :

- Indépendance des résidus et des variables explicatives : $l(\underline{u}\,|x) = l(\underline{u})$
- Homoscédasticité $V(\underline{u}\,|x) = \sigma^2 I$
- Spécification de la loi des résidus : $l(\underline{u})$ normale.

Les chapitres suivants vont progressivement revenir sur chacune de ces hypothèses. On va d'abord examiner dans un cadre très proche la loi asymptotique des estimateurs, c'est à dire lorsque le nombre d'observations devient grand. On va chercher à développer le même genre de propriétés permettant de faire de l'inférence mais sans spécifier la loi des résidus. Les résultats seront obtenus sous les hypothèses :

- Absence de corrélation entre les résidus et les variables explicatives $E(ux') = 0$
- Homoscédasticité $V(\underline{u}\,|x) = \sigma^2 I$

Le comportement asymptotique des estimateurs est examiné dans le chapitre 5.

Dans le chapitre 6 on revient sur les hypothèses d'indépendance et d'équidistribution des paramètres. On présente l'estimateur des moindres carrés généralisée ainsi que différentes façons de traiter la situation dite d'hétéroscédasticité, i.e. situation dans laquelle la variance des résidus dépend des variables explicatives. On aborde aussi succinctement la question des données de panel et de l'estimation de modèles faisant intervenir des systèmes d'équations. Le cadre dans lequel on se situe est juste basé sur

- Absence de corrélation entre les résidus et les variables explicatives $E(ux') = 0$

Les chapitres 7, 8 et 9 utilisent la méthode des moindres carrés généralisés en s'appuyant sur une connaissance a priori de la structure de corrélation des résidus. Le chapitre 7 s'intéresse plus particulièrement au cas des régressions empilées. Dans le chapitre 8, on considère le cas d'une régression en coupe dans laquelle on a hétéroscédascticité du résidu, ce qui peut être le cas par exemple pour une équation de salaire, la variance du résidu étant généralement croissante avec le revenu. Dans le chapitre 9, on considère le cas d'estimations où le résidu peut être modélisé comme une série temporelle de comportement

connu. On construit l'estimateur les moindres carrés quasi-généralisés en s'appuyant sur la connaissance de la forme de l'autocorrélation du résidu.

Dans le chapitre 10, on considère la situation dans laquelle $E(ux') \neq 0$. On aborde la question de l'identification, fondamentale en économétrie. On montre comment à l'aide de variables extérieures z, dites instrumentales, il est possible d'estimer le paramètre d'intérêt. On revient donc en partie sur certains aspects des généralisations précédentes pour mieux se concentrer sur l'hypothèse d'identification. Les résultats sont obtenus sous les hypothèses

- Absence de corrélation entre les résidus et des variables z : $E(uz') = 0$,
- $Rg(z'x) = \dim x$
- Homoscédasticité $V(\underline{u}\,|x, z) = \sigma^2 I$

On présente aussi deux tests importants : le test d'exogénéité et le test de suridentification qui sont des guides importants dans le choix des variables instrumentales.

Dans le chapitre 11 on présente une généralisation importante de la méthode à variable instrumentale et qui englobe la plupart des méthodes économétriques standards. Il s'agit de la méthode des moments généralisée et on montre en particulier comment elle permet d'étendre la méthode à variables instrumentales au cas dans lequel les perturbations sont hétéroscédastiques et à d'autres cas tels que celui de l'économétrie des données de panel ou l'estimation de systèmes d'équations. Les hypothèses s'écrivent un peu différemment ce qui souligne le caractère général de cette méthode

- $E(g(z, \theta)) = 0$

où z représente l'ensemble des variables du modèle, c'est à dire inclus les y et les x.

Dans le chapitre 12, on présente succinctement certains modèles non linéaires proches des modèle linéaires. On s'intéresse ainsi au modèles dits probit pour lesquels la variable à expliquer n'a plus un support continu sur R mais prend ses valeurs dans $\{0, 1\}$. La modélisation sous-jacente consiste à introduire une variable latente, i.e. non observée complètement

$$I^* = zc + u$$

et dont les réalisations gouvernent l'observation de la variable I :

$$I = 1 \Longleftrightarrow I^* > 0$$

On aborde également d'autres situations importantes permettant d'aborder la questions de la sélectivité des échantillons, c'est à dire la situation dans laquelle on n'observe la variable dépendante que sous une condition liée par ailleurs à la variable dépendante elle-même :

$$\begin{aligned} y^* &= xb + u \\ I^* &= zc + u \end{aligned}$$

les réalisations de I^* gouvernent l'observation de la variable I et de la variable y :

$$\begin{aligned} I^* &> 0 \Rightarrow \left\{ \begin{array}{c} I = 1 \\ y = y^* \end{array} \right. \\ I^* &\leq 0 \Rightarrow I = 0 \end{aligned}$$

Ce type de modèle appelé modèle Tobit est souvent utilisé, en particulier pour aborder l'endogénéité de variables explicatives prenant la valeur 0 ou 1 dans des modèles à coefficients variables

$$y_i = \lambda_i I_i + v_i$$

Ce type de modèle est souvent utilisé pour aborder l'évaluation des effets microéconomiques des politiques de l'emploi comme les stages de formations.

Dans le chapitre 13, on s'intéresse à l'évaluation des politiques publiques. On introduit notamment l'estimateur par différence de différences qui s'applique à une expérience naturelle. On parle d'expérience naturelle lorsqu'une partie de la population a fait l'objet d'une nouvelle politique, tandis qu'une autre partie de la population n'a pas fait l'objet de cette politique et donc peut servir de population témoin. On ne peut observer le comportement des individus touchés par une mesure s'ils n'avaient pas été touchés, on verra comment on peut néanmoins construire des estimateurs évaluant l'impact d'une nouvelle politique.

Chapitre 2

L'estimateur des moindres carrés ordinaires

L'estimateur des moindres carrés ordinaires reste l'un des estimateurs les plus fréquemment utilisés. Il a de nombreux usage. On peut l'utiliser par exemple pour procéder à une description des données : quelles sont les variables rendant compte le mieux de la variabilité d'une variable d'intérêt. On peut aussi l'utiliser dans de nombreuses autres situations pour estimer un paramètre auquel on donne un sens causal : que se passerait-il si on faisait varier une variable donnée d'un montant donné. Il est basé sur l'hypothèse essentielle que les résidus et les variables explicatives sont orthogonaux. Il faut d'autres hypothèses pour dériver les principales propriétés de l'estimateur. On verra d'abord les propriétés algébriques puis les propriétés statistiques. Une partie du cours correspondra à l'extension et la reformulation des propriétés de l'estimateur des mco lorsque l'on remet en cause ces hypothèses. On généralise ou adapte le plus souvent les propriétés de l'estimateur à la condition que l'hypothèse centrale d'absence de corrélation entre perturbations et variables explicatives soit maintenue.

On va voir dans ce chapitre la définition de l'estimateur des mco et son interprétation algébrique comme vecteur des coefficients de la projection orthogonale de la variable dépendante sur les variables explicatives. On va également obtenir deux propriétés importantes de cet estimateur qui sont : la propriété de "sans biais" et une propriété d'optimalité concernant la variance de l'estimateur, connue sous le nom de Théorème de Gauss-Markov.

2.1 Définition et propriétés algébriques

2.1.1 Définition

On considère une variable d'intérêt y appelée variable dépendante et un ensemble de K variables dites explicatives auquel on adjoint une constante. On dispose de N observations. On note $\underline{y} = (y_1, \ldots, y_N)$ l'empilement des N observations de la variable dépendante. On

définit de même les vecteurs $\underline{x_1}, \ldots, \underline{x_K}$ et $\underline{x}$ la matrice des variables explicatives à laquelle on adjoint le vecteur constant $e = (1, \ldots, 1)'$: $\underline{x} = \left(e, \underline{x_1}, \ldots, \underline{x_K}\right)$ est donc une matrice de dimension $N \times (K+1)$.

Definition *L'estimateur des moindres carrés ordinaires est défini comme le vecteur b de dimension $K+1$, $b = (b_0, \ldots, b_K)'$, des coefficients de la combinaison linéaire de $e, \underline{x_1}, \ldots, \underline{x_K}$ réalisant le minimum de la distance de $\underline{y}$ à l'espace vectoriel de R^N engendré par $e, \underline{x_1}, \ldots, \underline{x_K}$, pour la norme euclidienne :* $\widehat{b}_{mco} = \arg\min \left\|\underline{y} - \underline{x}b\right\|^2$

Proposition *Sous l'hypothèse*

$H1$: les vecteurs $e, \underline{x_1}, \ldots, \underline{x_K}$ sont indépendants,

l'estimateur des moindres carrés existe, est unique et a pour expression

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

Démonstration *L'objectif à minimiser est* $Ob(b) = \left\|\underline{y} - \underline{x}b\right\|^2 = \left(\underline{y} - \underline{x}b\right)'\left(\underline{y} - \underline{x}b\right)$. *La condition du premier ordre s'écrit*

$$\frac{dOb}{db} = -2\underline{x}'\left(\underline{y} - \underline{x}b\right) = 0$$

et la condition du second ordre

$$\frac{d^2Ob}{dbdb'} = 2\underline{x}'\underline{x} \text{ définie positive}$$

L'hypothèse d'indépendance de $e, \underline{x_1}, \ldots, \underline{x_K}$ revient à faire l'hypothèse que $\underline{x}'\underline{x}$ est définie positive. La condition du second ordre est ainsi satisfaite et la condition du premier ordre admet une solution

2.1.2 Interprétation géométrique

On associe deux grandeurs importantes à l'estimateur des moindres carrés :

1. Le vecteur prédit $\underline{\widehat{y}} = \underline{x}\widehat{b}$
2. Le vecteur résiduel $\underline{\widehat{u}} = \underline{y} - \underline{\widehat{y}}$

On voit immédiatement compte tenu de la définition de l'estimateur des moindres carrés ordinaires que le vecteur résiduel est orthogonal aux variables explicatives et donc aussi au vecteur prédit :

$$\begin{aligned} \underline{x}'\underline{\widehat{u}} &= 0 \\ \underline{\widehat{y}}'\underline{\widehat{u}} &= 0 \end{aligned}$$

$\underline{\widehat{y}}$ s'interprète donc comme la projection orthogonale de $\underline{y}$ sur l'espace engendré par $e, \underline{x_1}, \ldots, \underline{x_K}$ et l'estimateur des moindres carrés ordinaires comme le vecteur des coefficients de cette projection.

Remarque *Comme la constante appartient à l'ensemble des régresseurs, on a immédiatement* $\underline{e}'\underline{\widehat{u}} = 0$, *soit* $\overline{\widehat{u}} = \frac{1}{N}\sum \widehat{u}_i = 0$ *: la moyenne du vecteur résiduel est nulle.*

Les vecteurs prédits et résiduels peuvent s'écrire directement à partir du vecteur $\underline{y}$. On a en effet

$$\begin{aligned} \underline{\widehat{y}} &= \underline{x}\widehat{b} = \underline{x}\left(\underline{x}'\underline{x}\right)^{-1}\underline{x}'\underline{y} = P_x\underline{y} \\ \underline{\widehat{u}} &= \underline{y} - \underline{\widehat{y}} = \left(I_N - P_{\underline{x}}\right)\underline{y} = M_x\underline{y} \end{aligned}$$

Les matrices P_x et M_x sont les matrices des projecteurs orthogonaux sur respectivement l'espace engendré par $\left(e, \underline{x_1}, \ldots, \underline{x_K}\right)$ et son orthogonal. Comme on le vérifie directement on a en effet

$$\begin{aligned} P_x^2 &= P_x \\ M_x^2 &= M_x \\ P_x + M_x &= I_N \end{aligned}$$

et en outre

$$P_x v = v \Longleftrightarrow \exists\lambda \text{ tq } v = \underline{x}\lambda$$

2.1.3 Théorème de Frish-Waugh

Le théorème de Frish-Waugh est une propriété algébrique de l'estimateur des moindres carrés qui explicite l'interdépendance des coefficients de différentes variables dans une régression. Il permet de répondre à la question : dans quel cas est-il nécessaire d'introduire toutes les variables d'un modèle dans la liste des régresseurs ?

Theoreme *Dans la régression de* $\underline{y}$ *sur un ensemble de variables explicatives* $\underline{x}$, *si* $\underline{x}$ *se décomposent en deux sous-ensembles* $\underline{x_1}$ *et* $\underline{x_2}$: $\underline{x} = \left(\underline{x_1}, \underline{x_2}\right)$, *les coefficients des variables* $\underline{x_1}$ *peuvent être obtenus indirectement en régressant les résidus* $M_{x_2}\underline{y}$ *de la régression de la variable dépendante* $\underline{y}$ *sur les variables explicatives* $\underline{x_2}$, *sur les résidus* $M_{x_2}\underline{x_1}$ *des régressions des variables* $\underline{x_1}$ *sur les variables explicatives* $\underline{x_2}$:

$$\widehat{b}_1 = \left(\left(M_{x_2}\underline{x_1}\right)' M_{x_2}\underline{x_1}\right)^{-1}\left(M_{x_2}\underline{x_1}\right)' M_{x_2}\underline{y}$$

on peut alors retrouver les coefficients des variables x_2 *en régressant la partie inexpliquée* $\underline{y} - \underline{x_1}\widehat{b}_1$ *sur* $\underline{x_2}$:

$$\widehat{b}_2 = \left(\underline{x_2}'\underline{x_2}\right)^{-1}\underline{x_2}'\left(\underline{y} - \underline{x_1}\widehat{b}_1\right)$$

avec $M_{x_2} = I_N - \underline{x_2}\left(\underline{x_2}'\underline{x_2}\right)^{-1}\underline{x_2}'$

Démonstration *Les coefficients de la régression de* $\underline{y}$ *sur* $\underline{x} = \left(\underline{x_1}, \underline{x_2}\right)$ *satisfont*

$$\begin{aligned} \underline{x_1}' \left(\underline{y} - \underline{x_1}\widehat{b}_1 - \underline{x_2}\widehat{b}_2\right) &= 0 \\ \underline{x_2}' \left(\underline{y} - \underline{x_1}\widehat{b}_1 - \underline{x_2}\widehat{b}_2\right) &= 0 \end{aligned}$$

De la deuxième équation on tire directement la deuxième partie du théorème

$$\widehat{b}_2 = \left(\underline{x_2}'\underline{x_2}\right)^{-1} \underline{x_2}' \left(\underline{y} - \underline{x_1}\widehat{b}_1\right)$$

Lorsque l'on réintroduit cette expression dans la première équation il vient

$$\underline{x_1}' \left(\underline{y} - \underline{x_1}\widehat{b}_1 - \underline{x_2} \left(\underline{x_2}'\underline{x_2}\right)^{-1} \underline{x_2}' \left(\underline{y} - \underline{x_1}\widehat{b}_1\right)\right) = 0$$

soit

$$\begin{aligned} \underline{x_1}' M_{x_2} \left(\underline{y} - \underline{x_1}\widehat{b}_1\right) &= 0 \\ \underline{x_1}' M_{x_2} \left(M_{x_2}\underline{y} - M_{x_2}\underline{x_1}\widehat{b}_1\right) &= 0 \end{aligned}$$

compte tenu de $M_{x_2}^2 = M_{x_2}$. *D'où l'expression de* $\widehat{b}_1$

Remarque *La caractéristique importante est d'utiliser les résidus des régressions de* $\underline{x_1}$ *sur* $\underline{x_2}$. *Il n'est pas nécessaire d'utiliser aussi les résidus de la régression de* $\underline{y}$ *sur* $\underline{x_2}$.

Applications du Théorème de Frish-Waugh

1. Dans la régression de $\underline{y}$ sur $\underline{x_1}$ et $\underline{x_2}$ on peut régresser séparément $\underline{y}$ sur $\underline{x_1}$ et $\underline{y}$ sur $\underline{x_2}$ lorsque $\underline{x_1}$ et $\underline{x_2}$ sont orthogonaux.
2. Données de panel. Lorsque la régression introduit des indicatrices spécifiques à chaque individu (donc N variables, spécification dite à effets fixes) en plus d'un ensemble de régresseurs d'intérêt x_1, on peut d'abord régresser les variables d'intérêt et la variable dépendante sur les variables indicatrices puis utiliser les résidus des régressions correspondantes. Dans ces opérations puisque les variables indicatrices sont orthogonales les unes aux autres on peut effectuer les régressions sur les indicatrices séparément. On vérifie aisément que le coefficient de la régression d'une variable sur une variable indicatrice d'individu est la moyenne des observations pour cet individu. Les résidus des régressions sont donc les écarts aux moyennes individuelles des différentes variables d'intérêt. L'estimateur obtenu en régressant les écarts des variables explicatives aux moyennes individuelles sur la quantité analogue pour la variable dépendante est très populaire et connu sous le nom d'estimateur Within (ou Intra).
3. Pour obtenir les coefficients de $\underline{x_1}$ dans la régression de $\underline{y}$ sur $\underline{x_1}$ et $\underline{x_2}$, on peut régresser $\underline{y}$ sur $\underline{x_1}$ et la prévision de $\underline{x_1}$ par $\underline{x_2}$: $P_{x_2}\underline{x_1}$.

2.2 Modèle et propriétés statistiques

L'estimateur des moindres carrés ordinaires a une définition mathématique. Il s'agit du vecteur des coefficients de la projection orthogonale de la variable dépendante sur les variables explicatives. Dans le cadre de l'économétrie on s'intéresse néanmoins à l'estimation des paramètres d'un modèle économétrique. On considère ainsi le modèle linéaire suivant :

$$y = b_0 + b_1 x_1 + \cdots + b_K x_K + u$$

Pour lequel on dispose de N observations. Le modèle s'écrit aussi sous forme matricielle :

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

On s'intéresse aux propriétés statistiques de l'estimateur des mco : quelle est son espérance, sa variance... Comme l'estimateur est une fonction des observations, ses propriétés statistiques dépendent de la loi des observations $l(y, x)$. On les caractérise a partir d'hypothèses sur la loi conditionnelle de y sachant x, c'est à dire dans le cadre du modèle précédent comme des hypothèses concernant la loi de la perturbation u conditionnellement aux variables explicatives.

2.2.1 Quand l'estimateur des mco est-il sans biais ?

On s'intéresse d'abord aux conditions sous lesquelles l'espérance de l'estimateur des mco coïncide avec la vraie valeur du paramètre. On dit alors que l'estimateur est sans biais.

Definition *On dit qu'un estimateur $\widehat{b}\left(\underline{y}, \underline{x}\right)$ est sans biais lorsque*

$$E\left(\widehat{b}\left(\underline{y}, \underline{x}\right)\right) = b$$

Dans cette définition $E\left(\widehat{b}\left(\underline{y}, \underline{x}\right)\right) = \int \widehat{b}\left(\underline{y}, \underline{x}\right) f\left(\underline{y}, \underline{x}\right) d\underline{y} d\underline{x}$ où $f\left(\underline{y}, \underline{x}\right)$ représente la densité jointe des variables explicatives et dependantes.

Proposition *Sous l'hypothèse*
$H2 : E\left(u_n \left| \underline{x}\right.\right) = 0 \ \forall n$
l'estimateur des mco est sans biais.

Démonstration *L'estimateur des mco s'écrit*

$$\begin{aligned} \widehat{b}_{mco} &= \left(\underline{x}'\underline{x}\right)^{-1} \underline{x}'\underline{y} \\ &= \left(\underline{x}'\underline{x}\right)^{-1} \underline{x}' \left(\underline{x} b + \underline{u}\right) \\ &= b + \left(\underline{x}'\underline{x}\right)^{-1} \underline{x}'\underline{u} \end{aligned}$$

on s'intéresse à $E\left(\widehat{b}\left(\underline{y},\underline{x}\right)|\underline{x}\right)$. *On a clairement* $E\left(\widehat{b}\left(\underline{y},\underline{x}\right)|\underline{x}\right) = b + \left(\underline{x}'\underline{x}\right)^{-1}\underline{x}'E\left(\underline{u}\,|\underline{x}\right)$. *Comme* $E\left(\underline{u}\,|\underline{x}\right) = 0$ *par hypothèse on a bien* $E\left(\widehat{b}\left(\underline{y},\underline{x}\right)|\underline{x}\right) = b$. *On en déduit immédiatement* $E\left(\widehat{b}\left(\underline{y},\underline{x}\right)\right) = E\left(E\left(\widehat{b}\left(\underline{y},\underline{x}\right)|\underline{x}\right)\right) = b$

L'hypothèse $H2$ est extêmement forte, puisqu'elle signifie que lorsque les résidus changent, les variables explicatives ne changent pas. Dans de nombreuses situations cette hypothèses ne peut pas être tenu. C'est par exemple le cas si on prend un modèle offre-demande dans lequel on observe les prix et les quantités. Si on considère l'équation de demande par exemple, elle correspond à l'existence d'une relation décroissante entre la variable dépendante, la quantité, et la variable explicative, le prix. Si il y a un choc de demande, le déséquilibre sur le marché va se résoudre par une hausse de la quantité échangée et une hausse du prix. Dans ce modèle on ne peut donc pas tenir l'hypothèse $H2$ par nature même du modèle auquel on s'intéresse. Dans d'autres cas la situation peut être plus favorable. Par exemple dans le cas de la taille de la classe et du taux de réussite scolaire, il est vrai que l'on peut contester le fait que $E\left(u\,|taille\right) = 0$, mais il est possible qu'il existe un ensemble de variables explicatives x tel que l'on ait $u = xc + v$ et $E\left(v\,|taille, x\right) = 0$. Autrement dit, on peut identifier, mesurer et introduire dans la régression les sources de variabilité communes à la taille et au résidu. Le modèle devient $tx = a_0 + a_t taille + xb + v$.

2.2.2 Quelle est la précision de l'estimateur des mco ?

Le fait que la propriété d'absence de biais soit satisfaite est très intéressant mais on a besoin d'informations plus précises. On souhaite savoir si la vraie valeur peut se trouver loin de l'estimateur. Une telle information est donnée par la précision de l'estimateur et on l'étudie en considérant la variance :

Proposition *sous les hypothèses H1, H2,*
H3 : $V\left(u_n\,|\underline{x}\right) = \sigma^2 \quad \forall\ n$
H4 : $E\left(u_n u_m\,|\underline{x}\right) = 0 \quad \forall\ n, m$
la variance de l'estimateur des mco conditionnellement au variables explicatives est donnée par

$$V\left(\widehat{b}_{mco}\,|\underline{x}\right) = \sigma^2\left(\underline{x}'\underline{x}\right)^{-1}$$

La variance non conditionnelle est donnée par

$$V\left(\widehat{b}_{mco}\right) = \sigma^2 E\left[\left(\underline{x}'\underline{x}\right)^{-1}\right]$$

Démonstration *La variance conditionnelle est définie comme*

$$V\left(\widehat{b}_{mco}\,|\underline{x}\right) = E\left(\left[\widehat{b}_{mco} - E\left(\widehat{b}_{mco}\,|\underline{x}\right)\right]\left[\widehat{b}_{mco} - E\left(\widehat{b}_{mco}\,|\underline{x}\right)\right]'|\underline{x}\right)$$

Comme $E\left(\widehat{b}_{mco}\,|\underline{x}\right) = b$ *et* $\widehat{b}_{mco} - b = (\underline{x}'\underline{x})^{-1}\,\underline{x}'\underline{u}$,

$$V\left(\widehat{b}_{mco}\,|\underline{x}\right) = (\underline{x}'\underline{x})^{-1}\,\underline{x}'E\left(\underline{u}\underline{u}'\,|\underline{x}\right)\underline{x}\,(\underline{x}'\underline{x})^{-1}$$

La matrice $E\left(\underline{u}\underline{u}'\,|\underline{x}\right)$ *a pour éléments* n, m $E\left(u_n u_m\,|\underline{x}\right)$. *On déduit directement des hypothèses que* $E\left(\underline{u}\underline{u}'\,|\underline{x}\right) = \sigma^2 I_N$

La matrice de variance a deux composantes : σ^2 et $E\left[(\underline{x}'\underline{x})^{-1}\right]$. Plus σ^2, i.e. la variance résiduelle, est grande, moins l'estimateur est précis. Ceci implique que l'on peut accroître la précision des estimateurs de variables d'intérêt en introduisant des variables additionnelles, satisfaisant les hypothèses du modèle linéaire $H1 - H4$, dès lors qu'elles contribuent à réduire la variance résiduelle. La matrice $\underline{x}'\underline{x}$ joue un rôle central dans la variance de l'estimateur. On peut l'écrire à partir des observations individuelles comme $\underline{x}'\underline{x} = \sum_n x_n' x_n$. On voit qu'une écriture plus adaptée est $\underline{x}'\underline{x} = N\left(\frac{1}{N}\sum_n x_n' x_n\right)$. Dans le cas du modèle linéaire simple avec une unique variable explicative centrée la matrice $\left(\frac{1}{N}\sum_n x_n' x_n\right)^{-1}$ s'écrit simplement comme $1/\overline{x^2} = 1/V(x)$. On voit que dans ce cas la variance de l'estimateur s'écrit $V\left(\widehat{b}\right) = \sigma^2/(NV(x))$. L'estimateur est donc d'autant plus précis que le nombre d'observations est grand. On s'intéresse en général à l'écart-type des paramètres estimés. La formule précédente implique que l'écart type décroît comme $\sqrt{N}$. Lorsque la taille de l'échantillon est multipliée par 4 l'écart-type n'est divisé que par 2. On imagine donc bien que dans un échantillon de petite taille la précision de l'estimateur est un problème important. On voit aussi que dans de grands échantillons de plusieurs centaines de milliers d'observations, la précision des estimations sera très grande. La formule précédente montre aussi que l'estimateur est d'autant plus précis que la variance de la variable explicative est importante. C'est parce que l'on observe des situations différentes au regard des variables explicatives qui ne soient pas corrélées avec les résidus du modèle économique que l'on peut identifier l'effet de ces variables. Enfin un dernier cas permettant d'illustrer les implications de la formule précédente est le cas dans lequel il y a deux variables explicatives par exemple de même variance σ^2 et ayant un coefficient de corrélation ρ. Dans ce cas on calcule simplement

$$\left(\frac{1}{N}\sum_n x_n' x_n\right)^{-1} = \frac{1}{\sigma_x^2\left(1-\rho^2\right)}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$$

On voit que dans ce cas la précision de l'estimateur est d'autant plus faible que les variables sont corrélées. Au total, on voit que si les variables sont presque colinéaires la précision de l'estimateur sera faible.

2.2.3 L'estimateur des mco est-il le plus précis : le théorème de Gauss-Markov

On s'intéresse naturellement à la question de l'optimalité de l'estimation du paramètre b. Ce paramètre, comme on l'a vu, est sans biais et il est en outre défini comme une

fonction linéaire des observations. Ceci forme une classe d'estimateurs. La question à laquelle répond le théorème de Gauss-Markov est celle de l'optimalité (au sens de la précision) de l'estimateur dans la classe des estimateurs linéaires sans biais.

Definition *Un estimateur $\widehat{b}_1$ est optimal dans une classe d'estimateurs $\widehat{b}$ si toute estimation d'une combinaison linéaire du paramètre est estimée plus précisément avec $\widehat{b}_1$ qu'avec n'importe quel estimateur de la classe considérée :*

$$\forall \lambda, \quad V\left(\lambda' \widehat{b}_1\right) \leq V\left(\lambda' \widehat{b}\right)$$

Cette propriété signifie que la matrice de variance $V\left(\widehat{b}_1\right)$ de $\widehat{b}_1$ vérifie $\lambda' V\left(\widehat{b}_1\right) \lambda \leq \lambda' V\left(\widehat{b}\right) \lambda \quad \forall \lambda$, c'est à dire que $V\left(\widehat{b}_1\right) - V\left(\widehat{b}\right)$ est semi-définie négative.

Theoreme *Gauss-Markov : Sous les hypothèses H1-H4 l'estimateur des moindres carrés ordinaires du modèle*

$$y = xb + u$$

est optimal dans la classe des estimateurs sans biais conditionnellement aux variables $\underline{x}$.

Démonstration *Soit $\widetilde{b}$ un estimateur linéaire sans biais du paramètre b. Il existe donc une matrice A tel que cet estimateur s'écrit $\widetilde{b} = A\underline{y}$. L'hypothèse d'absence de biais signifie $E\left(\widetilde{b} \,|\underline{x}\right) = b$ ce qui implique $E\left(A\underline{y} \,|\underline{x}\right) = E\left(A\left(\underline{x}b + \underline{u}\right)|\underline{x}\right) = A\underline{x}b + AE\left(\underline{u} \,|\underline{x}\right) = b$ Comme $E\left(\underline{u} \,|\underline{x}\right) = 0$. L'absence de biais signifie $A\underline{x}b = b$. Ce résultat est vrai pour b quelconque donc pour tout b, c'est-à-dire :*

$$A\underline{x} = I_{K+1}$$

On a en outre $\widetilde{b} - E\left(\widetilde{b} \,|\underline{x}\right) = A\left(\underline{y} - E\left(y \,|\underline{x}\right)\right) = A\underline{u}$. La variance d'un estimateur linéaire sans biais quelconque est donc de la forme $V\left(\widetilde{b} \,|\underline{x}\right) = V\left(A\underline{u} \,|\underline{x}\right) = AV\left(\underline{u} \,|\underline{x}\right) A' = \sigma^2 AA'$ compte tenu de l'hypothèse cruciale $V\left(\underline{u} \,|\underline{x}\right) = \sigma^2 I_N$. Comme $I_N = P_x + M_x = \underline{x}\left(\underline{x}'\underline{x}\right)^{-1}\underline{x}' + M_x$, on a

$$\begin{aligned} V\left(\widetilde{b} \,|\underline{x}\right) &= \sigma^2 AA' = \sigma^2 A\left(\underline{x}\left(\underline{x}'\underline{x}\right)^{-1}\underline{x}' + M_x\right) A' \\ &= \sigma^2\left(A\underline{x}\left(\underline{x}'\underline{x}\right)^{-1}\underline{x}'A' + AM_xA'\right) \end{aligned}$$

comme $A\underline{x} = I_{K+1}$ et $V\left(\widehat{b} \,|\underline{x}\right) = \sigma^2\left(\underline{x}'\underline{x}\right)^{-1}$, on a

$$V\left(\widetilde{b} \,|\underline{x}\right) = V\left(\widehat{b} \,|\underline{x}\right) + \sigma^2 AM_xA'$$

et la matrice AM_xA' est nécessairement semi-définie négative

2.2.4 Estimation des paramètres du second ordre

La variance des résidus, intervenant dans l'hypothèses $H4$, est un paramètre dit du second ordre car il correspond aux moments d'ordre 2 de la variable y conditionnellement aux variables explicatives. C'est un paramètre important à plus d'un titre. D'abord, il permet de mesurer la qualité de l'ajustement. En outre, comme on l'a vu, il intervient dans la matrice de variance-covariance des estimateurs et est à l'origine de nombreux tests d'hypothèses. Il est donc légitime de s'intéresser à son estimation. Cette estimation fait intervenir le vecteur des résidus estimés

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

Proposition *Sous les hypothèses H1 à H4, l'estimateur*

$$\widehat{\sigma}^2 = \frac{\underline{\widehat{u}}'\underline{\widehat{u}}}{N-K-1} = \frac{\sum_n \widehat{u}_n^2}{N-K-1}$$

est un estimateur sans biais du paramètre du second ordre σ^2.

Démonstration *Comme on l'a vu* $\underline{\widehat{u}} = M_x\underline{y} = M_x\underline{u}$. *On a donc*

$$\underline{\widehat{u}}'\underline{\widehat{u}} = \underline{u}'M_x\underline{u} = Tr\left(\underline{u}'M_x\underline{u}\right) = Tr\left(M_x\underline{u}\underline{u}'\right)$$

On a donc

$$\begin{aligned} E\left(\underline{\widehat{u}}'\underline{\widehat{u}}\,|\underline{x}\right) &= E\left(Tr\left(M_x\underline{u}\underline{u}'\right)|\underline{x}\right) = Tr\left(E\left(M_x\underline{u}\underline{u}'\,|\underline{x}\right)\right) \\ &= Tr\left(M_xE\left(\underline{u}\underline{u}'\,|\underline{x}\right)\right) = \sigma^2 Tr\left(M_x\right) \end{aligned}$$

et $M_x = I_N - \underline{x}\left(\underline{x}'\underline{x}\right)^{-1}\underline{x}'$ *d'où*

$$\begin{aligned} Tr\left(M_x\right) &= Tr\left(I_N - \underline{x}\left(\underline{x}'\underline{x}\right)^{-1}\underline{x}'\right) = N - Tr\left(\underline{x}\left(\underline{x}'\underline{x}\right)^{-1}\underline{x}'\right) \\ &= N - Tr\left(\left(\underline{x}'\underline{x}\right)^{-1}\underline{x}'\underline{x}\right) = N-K-1 \end{aligned}$$

Exemple *Application à la prévision. On considère le modèle* $y_n = x_nb + u_n$ *pour lequel on a* $n = 1,\ldots,N$ *observations et satisfaisant les hypothèses H1 à H5. Connaissant* x_{N+1} *et faisant l'hypothèse que le modèle reste valide pour cette observation, on souhaite estimer* y_{N+1}.

Dire que le modèle reste valide signifie que non seulement la relation entre y_n *et* x_n *peut être étendue à l'observation* $N+1$: $y_{N+1} = x_{N+1}b + u_{N+1}$ *mais encore que les hypothèses stochastiques peuvent être étendues à l'inclusion de l'observation* $N+1$ *en particulier ceci impose* $E\left(u_{N+1}\,|\underline{x}, x_{N+1}\right) = 0$, $V\left(u_{N+1}\,|\underline{x}, x_{N+1}\right) = \sigma^2$, $E\left(u_{N+1}u_n\,|\underline{x}, x_{N+1}\right) = 0$.

La prévision de y_{N+1} est

$$\widehat{y}_{N+1} = x_{N+1}\widehat{b}_{mco}$$

Conditionnellement aux variables explicatives la prévision est sans biais :

$$E\left(\widehat{y}_{N+1} - y_{N+1}\,|\underline{x}, x_{N+1}\right) = E\left(x_{N+1}\left(\widehat{b}_{mco} - b\right) - u_{N+1}\,|\underline{x}, x_{N+1}\right) = 0$$

$\widehat{y}_{N+1}$ est le meilleur estimateur sans biais de y_{N+1}, linéaire dans les observations $y_1, \ldots, y_N$. Ceci constitue une application directe du Théorème de Gauss Markov : si on considère un estimateur linéaire sans biais $\widetilde{y}_{N+1}$ de y_{N+1}. La variance de l'erreur de prévision s'écrit $E\left(y_{N+1} - \widetilde{y}_{N+1}\,|\underline{x}, x_{N+1}\right)^2 = E\left(x_{N+1}b + u_{N+1} - \widetilde{y}_{N+1}\,|\underline{x}, x_{N+1}\right)^2 = E\left(x_{N+1}b - \widetilde{y}_{N+1}\,|\underline{x}, x_{N+1}\right)^2 + E\left(u_{N+1}^2\,|\underline{x}, x_{N+1}\right)$ puisque l'estimateur est linéaire en $\underline{y}$ et que $\underline{y}$ n'est pas corrélé à u_{N+1} conditionnellement aux observations de x. Le problème se résume donc à chercher l'estimateur linéaire sans biais de variance minimale de la combinaison linéaire $x_{N+1}b$ du paramètre b. Le théorème de Gauss-Markov indique qu'il s'agit de $x_{N+1}\widehat{b}_{mco}$

La variance de l'erreur de prévision est

$$E\left(\widehat{y}_{N+1} - y_{N+1}\right)^2 = \sigma^2\left[x'_{N+1}\left(\underline{x}'\underline{x}\right)^{-1}x_{N+1} + 1\right]$$

2.2.5 Analyse de la variance

L'analyse de la variance est fondée sur l'orthogonalité entre le vecteur des résidus estimés et de la variable prédite.

$$y = \widehat{y} + \widehat{u}$$

Les régressions que l'on considère ayant un terme constant on a $\overline{y} = \overline{\widehat{y}}$ dont on tire :

$$y - \overline{y}e = \widehat{y} - \overline{\widehat{y}}e + \widehat{u}$$

compte tenu de l'orthogonalité on peut donc écrire l'équation dite équation d'analyse de la variance

$$\sum_n \left(y_n - \overline{y}\right)^2 = \sum_n \left(\widehat{y}_n - \overline{\widehat{y}}\right)^2 + \sum_n \widehat{u}_n{}^2$$

ou encore

$$V\left(y\right) = V\left(\widehat{y}\right) + V\left(\widehat{u}\right)$$

La variance totale est la somme de la variance expliquée et de la variance résiduelle. On introduit une quantité très couramment utilisée qui mesure la part de la variance expliquée par le modèle.

$$R^2 = \frac{\left\|\widehat{y} - \overline{\widehat{y}}e\right\|^2}{\left\|y - \overline{y}e\right\|^2} = 1 - \frac{\left\|\widehat{u}\right\|^2}{\left\|y - \overline{y}e\right\|^2} \quad \in \quad \left[\begin{array}{cc} 0 & 1 \end{array}\right]$$

Le R^2 est fréquemment utilisé pour mesurer la qualité de l'ajustement. Néanmoins deux précautions doivent être prises :

– Le R^2 dépend du calibrage des observations. Par exemple si on considère une fonction de production

$$y = \alpha + \beta l + \gamma k + u$$

l'estimation va fournir un R^2 beaucoup plus important que celui obtenu avec le modèle identique mais expliquant la productivité

$$y - l = \alpha + (\beta - 1) l + \gamma k + u$$

– On montre facilement que plus on étend l'ensemble des variables explicatives plus le R^2 est grand. Ce n'est donc pas nécessairement un bon critère de choix de modèle. Pour cette raison on a introduit une quantité proche mais pas directement reliée qui est le R^2 ajusté. Il est défini d'une façon très voisine du R^2

$$R_a^2 = 1 - \frac{\widehat{\sigma}^2}{V(y)} = 1 - \frac{\|\widehat{u}\|^2 / (N-K-1)}{\|y - \overline{y}e\|^2 / N-1} = 1 - \left(1 - R^2\right) \frac{N-1}{N-K-1}$$

Remarque *Cette équation d'analyse de la variance permet de préciser l'expression de la variance de chacune des composantes de l'estimateur. Dans la formule générale* $V\left(\widehat{b}_{mco} \left| \underline{x} \right.\right) = \sigma^2 \left(\underline{x}'\underline{x}\right)^{-1}$, *la variance de la kième composante de l'estimateur des mco correspond au kième éléments diagonal. Si on utilise les formules d'inversion par bloc*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},\ A^{-1} = \begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{bmatrix},\ A^{11} = \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1}$$

Si on considère une variable x_k *particulière, alors, quitte à réorganiser l'ordre des variables explicatives :* $\underline{x} = (\underline{x_k}, \underline{x_{-k}})$, *où* $\underline{x_{-k}}$ *représente l'ensemble des variables explicatives autres que la kième,*

$$\underline{x}'\underline{x} = \begin{bmatrix} \underline{x_k}'\underline{x_k} & \underline{x_k'}\underline{x_{-k}} \\ \underline{x'_{-k}}\underline{x_k} & \underline{x'_{-k}}\underline{x_{-k}} \end{bmatrix}$$

et on a donc $\left(\underline{x}'\underline{x}\right)^{-1}_{11} = \underline{x_k}'\underline{x_k} - \underline{x_k'}\underline{x_{-k}} \left(\underline{x'_{-k}}\underline{x_{-k}}\right)^{-1} \underline{x'_{-k}}\underline{x_k} = \underline{x_k}' M_{x_{-k}} \underline{x_k} = \left(N.V\left(x_k \left| x_{-k} \right.\right)\right)^{-1}$. $V\left(x_k \left| x_{-k} \right.\right)$ *est la variance résiduelle de la variable* x_k *une fois pris en compte la part de la variance de la variable* x_k *expliquée par les autres variables explicatives du modèle. La variance de chacune des composante de l'estimation du paramètre s'écrit donc*

$$V\left(\widehat{b}_k\right) = \sigma^2 / \left(NV\left(x_k \left| x_{-k} \right.\right)\right) = \sigma^2 / NV_{k|-k}$$

2.3 Variable omise et régresseur additionnel

2.4 Résumé

1. On a vu dans ce chapitre la définition algébrique de l'estimateur des mco comme vecteur des coefficients de la projection orthogonale de la variables dépendante sur

l'espace engendré par les variables explicatives.

2. Cet estimateur existe est unique sous l'hypothèse $H1$ que les vecteurs des variables explicatives soient linéairement indépendant.
3. On a vu sous quelle condition l'estimateur des mco est un estimateur sans biais du paramètre économique b dans le modèle linéaire $y = xb + u.$: Il s'agit de l'hypothèse $H2$ que l'espérance des résidus conditionnellement aux variables observables est nulle.
4. Sous les hypothèses $H3$ et $H4$ que dans ce modèle les perturbations sont conditionnellement aux variables explicatives des variances identiques et sont non corrélées les unes avec les autres, on peut donner l'expression classique de la matrice de variance de l'estimateur $V\left(\widehat{b}\,|\underline{x}\right) = \sigma^2 \left(\underline{x}'\underline{x}\right)^{-1}$.
5. Sous ces même hypothèses l'estimateur des mco est le meilleur estimateur linéaire sans biais, au sens de la minimisation de la variance.
6. L'interprétation de cette formule conduit à la conclusion que plus le nombre d'observations est grand, plus la variance résiduelle σ^2 est faible, plus les variables explicatives présentent de variabilité propre, plus l'estimateur est précis.
7. Le paramètre du second ordre σ^2 peut être estimé sans biais comme la moyenne des carrés des résidus tenant compte des degrés de liberté : $\widehat{\sigma}^2 = \sum \widehat{u}_n^2 / \left(N - K - 1\right)$.
8. Le R^2 est une mesure de la qualité de l'ajustement du modèle aux données : il mesure la part de la variance totale expliquée par le modèle.

Ces résultats sont importants : ils établissent les conditions sous lesquelles les estimateurs sont sans biais et ils permettent de déterminer la précision des estimations. Ils sont néanmoins insuffisants pour donner des intervalles de confiance sur les paramètres estimés et réaliser des tests d'hypothèse. Pour aller plus loin il faut faire des hypothèses supplémentaires. On peut procéder de deux façons :

1. Lorsque le nombre d'observations est faible, on peut spécifier la loi des observations conditionnellement aux variables explicatives. Ceci est fait dans la majeure partie des cas en spécifiant les résidus comme suivant une loi normale. On peut alors caractériser la loi de l'estimateur. On peut aussi dans ce cas estimer le modèle par maximum de vraisemblance. On peut alors tester des hypothèses dites simples (nullité d'un paramètre). Ces tests sont appelés test de Student. Ce cas est examiné dans le chapitre 3. On peut aussi sur la base de cette hypothèse estimer le modèle en imposant des contraintes linéaires sur les paramètres et tester l'hypothèse que ces contraintes sont acceptées. Les tests mis en oeuvres sont alors des test dits de Fisher. Ces aspects sont présentés dans le chapitre 4.
2. La deuxième façon est d'étudier les propriétés asymptotiques de l'estimateur, c'est à dire lorsque le nombre d'observations devient grand. On montre dans le chapitre 5 que sans spécifier la loi des résidus mais en faisant des hypothèses suffisamment

fortes sur l'épaisseur des queues de distribution des résidus, on peut spécifier la loi asymptotique de l'estimateur.

Chapitre 3

Les MCO sous l'hypothèse de normalité des perturbations.

Dans ce chapitre on examine les propriétés de l'estimateur des mco lorsque l'on fait l'hypothèse de normalité des perturbations. Plus précisément on fait l'hypothèse H_n suivante.

H_n : la loi de $\underline{u}$ conditionnellement aux variables explicatives $\underline{x}$ est une loi normale de moyenne nulle et de matrice de variance $\sigma^2 I_N$.

$$\begin{aligned} l\left(\underline{u}\,|\underline{x}\right) &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^N}\varphi\left(-\sum u_n^2/2\sigma^2\right) \\ \underline{u}\,|\underline{x} &\rightsquigarrow N\left(0,\sigma^2 I_N\right) \end{aligned}$$

Remarque *Cette hypothèse est plus forte que les hypothèses H_2-H_4 puisqu'elle implique que le moment d'ordre 1 de u conditionnellement à x est nul. c'est à dire l'espérance*

On va voir que dans ce cas on peut préciser la loi de l'estimateur du paramètre ainsi que celle de l'estimateur de la variance des résidus. On va aussi obtenir un résultat central, le théorème de Cochrane, à la base de tous les tests effectués à partir de l'estimateur des mco.

3.1 Normalité de l'estimateur des mco

Proposition *Sous l'hypothèse H_n, on peut spécifier la loi jointe de l'estimateur des mco et de l'estimateur de la variance des résidus conditionnellement aux variables explicatives :*

1. *L'estimateur du paramètre des mco $\widehat{b}_{mco}$ est distribué comme une loi normale de moyenne b, la vraie valeur du paramètre, et de matrice de variance $\sigma^2\left(\underline{x}'\underline{x}\right)^{-1}$: $\widehat{b}_{mco} \rightsquigarrow N\left(b,\sigma^2\left(\underline{x}'\underline{x}\right)^{-1}\right)$*

2. *L'estimateur* $\widehat{\sigma}^2$, *convenablement normalisé, est distribué suivant une loi du* χ^2

$$[N-(K+1)]\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2\left(N-(K+1)\right)$$

3. $\widehat{b}_{mco}$ *et* $\widehat{\sigma}^2$ *sont indépendants (Théorème de Cochran)*

Démonstration *Le résultat concernant la normalité de l'estimateur est immédiat. Il provient du fait que l'estimateur des mco est linéaire dans les observations de la variable dépendante. Comme conditionnellement à* $\underline{x}$ *la variable dépendante est normale, l'estimateur des mco est une combinaison linéaire de variables normales et est donc lui même un vecteur normal, caractérisé par ces deux premiers moments : son espérance dont on a vu qu'elle était égale à la vraie valeur du paramètre, et sa matrice de variance dont on a donné l'expression au chapitre précédent, sous des hypothèses plus générales que celle de la loi normale.*

De même, les résidus estimés sont eux mêmes normaux. On a en effet $\underline{\widehat{u}} = M_x\underline{y} = M_x\underline{u}$. *Par ailleurs, on a aussi directement* $\widehat{b}-b = (\underline{x}'\underline{x})^{-1}\underline{x}'\underline{u}$. *Finalement on peut spécifier la loi jointe des résidus estimés et de l'estimateur des mco :*

$$\begin{pmatrix} \widehat{b}-b \\ \underline{\widehat{u}} \end{pmatrix} = \begin{pmatrix} (\underline{x}'\underline{x})^{-1}\underline{x}' \\ M_x \end{pmatrix}\underline{u}$$

On en déduit donc que ces deux vecteurs suivent une loi normale jointe, de moyenne visiblement nulle et dont on peut préciser la variance :

$$\begin{aligned} V\left(\begin{array}{c|c} \widehat{b}-b & \\ \underline{\widehat{u}} & \underline{x} \end{array}\right) &= \begin{pmatrix} (\underline{x}'\underline{x})^{-1}\underline{x}' \\ M_x \end{pmatrix} V\left(\begin{array}{c|c} \widehat{b}-b & \\ \underline{\widehat{u}} & \underline{x} \end{array}\right)\begin{pmatrix} (\underline{x}'\underline{x})^{-1}\underline{x}' \\ M_x \end{pmatrix}' \\ &= \sigma^2\begin{pmatrix} (\underline{x}'\underline{x})^{-1}\underline{x}' \\ M_x \end{pmatrix}\begin{pmatrix} \underline{x}(\underline{x}'\underline{x})^{-1} & M_x \end{pmatrix} \\ &= \sigma^2\begin{pmatrix} (\underline{x}'\underline{x})^{-1}\underline{x}'\underline{x}(\underline{x}'\underline{x})^{-1} & \underline{x}(\underline{x}'\underline{x})^{-1}M_x \\ M_x\underline{x}(\underline{x}'\underline{x})^{-1} & M_x \end{pmatrix} \end{aligned}$$

Comme $M_x\underline{x} = 0$, *on en déduit*

$$V\left(\begin{array}{c|c} \widehat{b}-b & \\ \underline{\widehat{u}} & \underline{x} \end{array}\right) = \sigma^2\begin{pmatrix} (\underline{x}'\underline{x})^{-1} & 0 \\ 0 & M_x \end{pmatrix}$$

Dont on déduit

1. *l'expression de la variance de l'estimateur des mco*
2. *l'estimateur des mco et les résidus estimés sont indépendants (car étant tous les deux normaux et non corrélés). L'estimateur des mco et l'estimateur de la variance* $\widehat{\sigma}^2 = \underline{\widehat{u}}'\underline{\widehat{u}}/\left(N-K-1\right)$ *sont donc indépendants.*

3. Les résidus estimés suivent une loi normale de matrice de variance $\sigma^2 M_x$.

Rappel :
- *Si $Z \rightsquigarrow N(0, I_L)$, alors par définition $\|Z^2\| = Z'Z = \sum_{l=1}^{L} Z_l^2 \sim \chi^2(L)$*
- *Si P est un projecteur orthogonal sur un sous espace de dimension L_1 alors $Z'PZ \sim \chi^2(L_1)$ (Voir annexe)*

On applique ce résultat à $Z = \underline{u}/\sigma \rightsquigarrow N(0, I_N)$ et $P = M_x$. On a : $(\widehat{\underline{u}}/\sigma)'(\widehat{\underline{u}}/\sigma) = (\underline{u}/\sigma)' M_x' M_x (\underline{u}/\sigma) = (\underline{u}/\sigma)' M_x (\underline{u}/\sigma)$. On en déduit que $\widehat{\underline{u}}'\widehat{\underline{u}}/\sigma^2 \rightsquigarrow \chi^2(N-K-1)$, puisque M_x est le projecteur orthogonal sur l'orthogonal de l'espace vectoriel engendré par les x donc de dimension $N-K-1$. Finalement, comme $\widehat{\underline{u}}'\widehat{\underline{u}} = (N-K-1)\widehat{\sigma}^2$, $[N-(K+1)]\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(N-(K+1))$

On rappelle qu'une loi du $\chi^2(L)$ à L degrés de libertés a pour premier et second moments $E(\chi^2(L)) = L$, $V(\chi^2(L)) = 2L$. On vérifie donc que $E\left([N-(K+1)]\frac{\widehat{\sigma}^2}{\sigma^2}\right) = N-K-1$. On vérifie donc que l'on a bien $E\left(\widehat{\sigma}^2\right) = \sigma^2$: l'estimateur de la variance est sans biais. On apprend maintenant, grâce à la spécification normale la distribution de l'estimateur de la variance des résidus et donc sa variance : on a $V\left([N-(K+1)]\frac{\widehat{\sigma}^2}{\sigma^2}\right) = 2(N-K-1)$, soit $V\left(\widehat{\sigma}^2\right) = 2\sigma^4/(N-K-1)$. On voit donc que comme pour l'estimateur des mco, lorsque le nombre d'observations devient grand la variance de l'estimateur tend vers zero. Le rythme de convergence est en outre identique à celui de l'estimateur des mco. On remarque en revanche une spécificité de l'estimateur de la variance : plus la dispersion des résidus est importante, plus l'estimateur est imprécis.

L'estimation de la variance des résidus peut être intéressante pour elle-même, mais elle nous intéresse en premier lieu car c'est un paramètre important de la matrice de variance de l'estimateur du paramètre de premier intérêt b. En effet, on a vu que $\widehat{b}_{mco}|\underline{x} \rightsquigarrow N\left(b, \sigma^2(\underline{x}'\underline{x})^{-1}\right)$, mais ce résultat reste insuffisant dans la mesure où on ne connaît pas la variance des résidus.

3.2 Ecart-types estimés, tests et intervalles de confiance

3.2.1 Ecart-type

La formule de la matrice de variance de l'estimateur est utile $V\left(\widehat{b}_{mco}|\underline{x}\right) = \sigma^2(\underline{x}'\underline{x})^{-1}$, mais elle n'est pas directement exploitable car on ne connaît pas la variance des résidus σ^2. Un estimateur naturel de cette matrice consiste à remplacer la quantité inconnue σ^2 par un estimateur.

$$\widehat{V}\left(\widehat{b}_{mco}|\underline{x}\right) = \widehat{\sigma}^2(\underline{x}'\underline{x})^{-1}$$

On a immédiatement le résultat que $\widehat{V}\left(\widehat{b}_{mco}|\underline{x}\right)$ est un estimateur sans biais de la matrice de variance de l'estimateur mco du paramètre.

On s'intéresse en fait plus spécifiquement à la variance de chaque composante de l'estimateur $\sigma_k^2 = V\left(\widehat{b}_k\right) = \sigma^2\left[\left(\underline{x}'\underline{x}\right)^{-1}\right]_{kk} = \sigma^2 x^{kk}$ où dans cette notation x^{kk} est le kième élément diagonal de $\left(\underline{x}'\underline{x}\right)^{-1}$. Dans le chapitre précédent on a vu que ce kième élément était en fait l'inverse de la variance résiduelle de la projection de x_k sur les autres variables du modèle (la variance propre de la kième variable) divisée par le nombre d'observations. Un estimateur naturel de σ_k^2 est

$$\widehat{\sigma}_k^2 = \widehat{\sigma}^2 x^{kk}$$

La quantité $\widehat{\sigma}_k = \sqrt{\widehat{\sigma}_k^2}$ est systématiquement associé à n'importe qu'elle estimation par les mco. Grâce aux résultats portant sur la loi de $\widehat{\sigma}^2$ on peut directement donner la loi de $\widehat{\sigma}_k^2$:

Proposition *Sous l'hypothèse H_n l'estimateur de la variance de la kième composante du vecteur des paramètre suit, convenablement normalisée une loi du $\chi^2(N-K-1)$:*

$$[N-(K+1)]\frac{\widehat{\sigma}_k^2}{\sigma_k^2} \sim \chi^2(N-(K+1))$$

et est indépendant de l'estimateur des mco $\widehat{b}_{mco}$.

3.2.2 Un résultat central

On s'intéresse à l'obtention d'intervalles de confiance et à des tests d'hypothèse simple du type $H_0 : \ b_k = b_k^0$ pour une valeur donnée de b_k^0. Un cas très fréquemment examiné est par exemple celui de la nullité d'un paramètre ($b_k^0 = 0$). Pour obtenir des intervalles de confiance ou pour effectuer des tests, on a besoin d'obtenir une fonction des estimateurs qui ne dépende pas des paramètres.

Proposition *Sous l'hypothèse de normalité des perturbations H_n, pour une composante donnée k du paramètre on a*

$$\frac{\widehat{b}_k - b_k}{\widehat{\sigma}_k} \sim Student(N-K-1)$$

Démonstration *Ce résultat découle directement de la définition des lois de Student : Si X_1 suit une loi normale $N(0,1)$ et X_2 suit une loi du $\chi^2(H)$ à H degrés de liberté, et si X_1 et X_2 sont indépendants alors*

$$S = \frac{X_1}{\sqrt{X_2/H}} \sim Student(H)$$

Ici $\widehat{b}_k \rightsquigarrow N\left(b_k, \sigma_k^2\right)$. *On en déduit donc que* $\left(\widehat{b}_k - b_k\right)/\sigma_k \rightsquigarrow N(0,1)$. *En outre le résultat précédent établit que* $[N-(K+1)]\frac{\widehat{\sigma}_k^2}{\sigma_k^2} \sim \chi^2\left(N-(K+1)\right)$ *et est indépendant de* $\widehat{b}_k$. *On a donc par application directe de la définition*

$$\frac{\left(\widehat{b}_k - b_k\right)/\sigma_k}{\sqrt{\left([N-(K+1)]\frac{\widehat{\sigma}_k^2}{\sigma_k^2}\right)/[N-(K+1)]}} = \frac{\widehat{b}_k - b_k}{\widehat{\sigma}_k} \sim Student\left(N-K-1\right)$$

Les lois de Student sont des lois symétriques de moyenne nulle et de variance $H/(H-2)$ où H est le nombre de degrés de liberté. Plus H est faible, plus les queues de distribution sont épaisses. On voit qu'il y a un nombre minimal de degrés de liberté pour que le moment d'ordre 2 existe : $H > 2$.

3.2.3 Intervalle de confiance

Definition *Un intervalle de confiance pour le paramètre* b_k *au niveau* α *est un intervalle* $[\underline{a}, \overline{a}]$ *tq* $P\left(b_k \in [\underline{a}, \overline{a}]\right) = 1-\alpha$.

Lemme *Soit* z *une variable aléatoire dont la distribution* f *est symétrique autour de zéro, croissante pour* $z<0$, *continue et de fonction de répartition* F, *tout intervalle* $[\underline{z}, \overline{z}]$ *tel que* $P\left(z \in [\underline{z}, \overline{z}]\right) = p_0$ *donné, de longueur minimale est symétrique.*

Démonstration *Ce résultat se montre très facilement. La symétrie de la distribution s'écrit* $f(z) = f(-z)$ *et implique* $F(-z) = 1-F(z)$. *On a* $F(\overline{z}) - F(\underline{z}) = p_0$, *donc la longueur de l'intervalle est* $L = \overline{z} - \underline{z} = F^{-1}\left(F(\underline{z}) + p_0\right) - \underline{z}$. *La dérivée de la longueur de l'intervalle par rapport à* $\underline{z}$ *est* $dL/d\underline{z} = f(\underline{z})/f(\overline{z}) - 1$. *Si* $f(\underline{z}) < f(\overline{z})$, *alors* $dL/d\underline{z} < 0$. *On pourra diminuer la longueur de l'intervalle en augmentant* $\underline{z}$. *Comme* f *est croissante dans le domaine négatif accroître* $\underline{z}$ *conduit à accroître* $f(\underline{z})/f(\overline{z}) - 1$. *L'extremum de la longueur, obtenu pour* $f(\underline{z})/f(\overline{z}) - 1 = 0$ *est donc bien un minimum.*

Pour trouver un intervalle de confiance pour le paramètre b_k on applique directement les résultats du lemme :

Proposition *Sous les hypothèses* H_n, *soit* $\widehat{b}_k$ *la kième composante de l'estimateur des mco et* $\widehat{\sigma}_k = \sqrt{\widehat{\sigma}_k^2}$ *l'estimateur de son écart-type, alors l'intervalle de confiance de longueur minimale du paramètre* b_k *au niveau* α *est*

$$\left[\widehat{b}_k - \widehat{\sigma}_k\, t_{N-K-1}(1-\alpha/2) \quad , \quad \widehat{b}_k + \widehat{\sigma}_k\, t_{N-K-1}(1-\alpha/2)\right]$$

où $t_{N-K-1}(1-\alpha/2)$ *est le quantile d'ordre* $1-\alpha/2$ *d'une loi de Student à* $N-K-1$ *degrés de liberté.*

Le quantile d'ordre $1-\alpha/2$ d'une loi de Student à $N-K-1$ degrés de liberté est la quantité t telle que pour une variable S suivant une loi de Student à $N-K-1$ degrés de liberté, $P(S<t)=1-\alpha/2$, et de façon similaire $P(S>t)=\alpha/2$

Démonstration *Par application des résultats précédents, on a immédiatement que* $S=\frac{\widehat{b}_k-b_k}{\widehat{\sigma}_k} \rightsquigarrow Student\,(N-K-1)$. *Comme la loi de Student est symétrique, on en déduit que l'intervalle de longueur minimale auquel* S *appartienne avec probabilité* $1-\alpha$ *est*

$$P\left(S\in\left[-t_{N-K-1}\left(1-\alpha/2\right),t_{N-K-1}\left(1-\alpha/2\right)\right]\right)=1-\alpha$$

dont on déduit immédiatement l'expression des bornes de l'intervalle de confiance.

Remarque *Ce résultat s'étend directement au cas dans lequel on cherche un intervalle de confiance pour une combinaison linéaire donnée des paramètres :* $\lambda'b$. *En effet, on trouve directement la loi de l'estimateur de la combinaison linéaire* $\lambda'\widehat{b}_{mco}$: $\lambda'\widehat{b}_{mco} \rightsquigarrow N\left(\lambda'b,\sigma^2\lambda'\left(\underline{x}'\underline{x}\right)^{-1}\lambda\right)$. *En notant* $\sigma_{\lambda b}=\sqrt{\sigma^2\lambda'\left(\underline{x}'\underline{x}\right)^{-1}\lambda}$ *et* $\widehat{\sigma}_{\lambda b}=\sqrt{\widehat{\sigma}^2\lambda'\left(\underline{x}'\underline{x}\right)^{-1}\lambda}$, *on vérifie aisément que l'intervalle de confiance pour la combinaison linéaire donnée des paramètres est* $\left[\lambda'\widehat{b}_{mco}-\widehat{\sigma}_{\lambda b}\;t_{N-K-1}\left(1-\alpha/2\right)\quad,\quad\lambda'\widehat{b}_{mco}+\widehat{\sigma}_{\lambda b}\;t_{N-K-1}\left(1-\alpha/2\right)\right]$

3.2.4 Tests de la forme $\lambda'b=\mu$

On rappelle d'abord des éléments basiques concernant les tests. On se réfère pour cela à Gouriéroux-Monfort. Les notions importantes sont celles d'hypothèse nulle, notée H_0, et d'hypothèse alternative, notée H_1. Elles correspondent à une partition de l'ensemble des lois possibles des observations. Ici compte tenu du fait qu'on se situe dans un cadre paramétrique (la loi des observations est spécifiée intégralement), l'ensemble des lois possibles est décrit par l'ensemble des valeurs possibles de tous les paramètres : b,σ^2. Les hypothèses que l'on va considérer ici portent sur la valeur d'une composante du paramètre ou d'une combinaison linéaire du paramètre : $b_k=b_k^0$ pour une valeur donnée de b_k^0, un cas très fréquent étant celui de la nullité, $b_k^0=0$. On examinera dans le chapitre suivant des hypothèses portant sur plusieurs paramètres, mais les rappels que l'on effectue ici valent pour l'une et l'autre situation. D'une façon générale, elles vont s'écrire sous la forme $H_0:\theta\in\Theta_0$ et $H_1:\theta\in\Theta_1$.

Un test pur est une règle de décision pure c'est à dire une fonction des observations conduisant à choisir entre la décision d_0 : H_0 est vraie, et d_1 : H_1 est vraie. A un test pur est associé une région critique, en général notée W définie comme l'ensemble des réalisations des observations conduisant à prendre la décision d_1. Les tests peuvent aussi en théorie être mixtes. Dans ce cas la règle de décision est mixte. Il s'agit alors d'une fonction des observations associant à la décision d_1 une probabilité : compte tenu des observations y on accepte l'hypothèse H_1 avec une probabilité $\phi(y)$. Il y a trois grandeurs essentielles associées à un test : le risque de première espèce, le risque de deuxième espèce et la puissance du test. Le risque de première espèce correspond à la probabilité de de

rejeter H_0 alors que H_0 est vraie (i.e. rejeter H_0 à tort). Pour un test pur caractérisé par une région critique W, il s'agit de la fonction $P_\theta(W)$ définie sur Θ_0 Pour un test aléatoire, elle est définie par $E_\theta(\phi(y))$. On la note $\alpha(\phi,\theta)$. Dans cette notation, ϕ représente le test et θ la valeur du paramètre. Le risque de deuxième espèce est à l'inverse la probabilité d'accepter à tort l'hypothèse nulle (i.e. la probabilité de rejeter H_1 alors que H_1 est vraie. Il est défini comme $1-E_\theta(\phi(y))$ pour $\theta \in \Theta_1$ et dans le cas d'un test pur par $1-P_\theta(W)$. On note en général cette quantité $\beta(\phi,\theta)$ Enfin la puissance du test représente la probabilité de rejeter à raison l'hypothèse nulle. On la note $\gamma(\phi,\theta)$. Cette fonction est définie sur Θ_1 et étroitement liée à la fonction de risque de deuxième espèce $\gamma(\phi,\theta) = 1-\beta(\phi,\theta)$. On préférerait des tests pour lesquels les risques de premières et seconde espèce soient les plus faibles possibles. C'est à dire qu'un test est préféré à un autre si les fonctions de risque de première et seconde espèce sont plus faibles. Il existe clairement des tests minimisant séparément chacun des risques (le test correspondant au rejet systématique de H_1 minimise le risque de première espèce). Néanmoins on montre facilement qu'il n'y a pas de test annulant simultanément les deux risques : il est donc nécessaire de se référer à un principe permettant de sélectionner un test. Le principe retenu est celui de Neyman qui consiste à privilégier la minimisation du risque de seconde espèce. On considère des classes de tests caractérisés par un seuil (ou encore niveau) donné α. Ces tests sont tels que le risque de première espèce soit uniformément inférieur à α. Parmi ces tests, on souhaiterait sélectionné ceux maximisant la puissance. C'est ce que l'on appelle des tests uniformément plus puissants. Ils sont tels qu'ils maximisent parmi les tests de niveaux α la puissance pour toute valeur du paramètre correspondant à l'hypothèse alternative. De tels tests n'existe en général pas et on adjoint d'autres propriétés : tests sans biais, tests invariants... qui permettent de restreindre encore la classe des tests examinés. La propriété de tests sans biais au niveaux α correspond pour les tests de niveau α au fait que la puissance du test pour toute valeur du paramètre sous l'hypothèse alternative soit supérieure à α. On considère le test de l'hypothèse nulle

$$H_0 : b_k = b_k^0$$

contre l'hypothèse

$$H_1 : b_k \neq b_k^0$$

On a alors le résultat suivant

Proposition *Considérant la statistique*

$$\widehat{S} = \frac{\widehat{b}_k - b_k^0}{\widehat{\sigma}_k}$$

le test défini par la région critique

$$W = \left\{ \widehat{S} \middle| \widehat{S} < -t_{N-K-1}(1-\alpha/2) \right\} \cup \left\{ \widehat{S} \middle| \widehat{S} > t_{N-K-1}(1-\alpha/2) \right\}$$

où $t_{N-K-1}(1-\alpha/2)$ *est le quantile d'ordre* $1-\alpha/2$ *d'une loi de Student à* $N-K-1$ *degrés de liberté est un test uniformément plus puissant sans biais au niveau* α *de l'hypothèse* H_0 *contre* H_1*.*

On vérifie aisément que ce test est un test au niveau α. En effet sous l'hypothèse nulle on a vu que $\frac{\widehat{b}_k - b_k^0}{\widehat{\sigma}_k}$ suit une loi de Student à $N-K-1$ degrés de liberté. La probabilité de rejeter l'hypothèse nulle (la probabilité de la région critique) dans ce cas est donc bien α. Montrer la propriété de sans biais et la propriété concernant la puissance est plus compliqué (voir les résultats dans Gourieroux et Monfort sur le modèle exponentiel). On peut aussi définir la région critique par $W = \left\{ \widehat{S} \,\middle|\, \left|\widehat{S}\right| > t_{N-K-1}(1-\alpha/2) \right\}$

Mise en oeuvre du test : on calcule la statistique de Student $\frac{\widehat{b}_k - b_k^0}{\widehat{\sigma}_k}$. Suivant les valeurs prises par cette statistique, on accepte ou rejette l'hypothèse nulle. Si la statistique prend des valeurs extrêmes on rejette l'hypothèse, sinon on l'accepte. Le seuil de rejet dépendent du niveau du test. On considère en général des tests au seuil de 5%. Le quantile d'ordre $97,5\% = 1 - 2,5\%$ d'une loi de Student dépend du nombre de degrés de liberté. lorsque ce nombre devient grand, ce quantile est 1.96. On sera donc amené à rejeter au seuil de 5% une hypothèse dès lors que la statistique de Student en valeur absolue est supérieur à 1.96. Lorsque le nombre de degrés de liberté est plus faible, c'est à dire lorsque le nombre de variables explicatives est plus important ou lorsque le nombre d'observations est plus faible, le seuil augmente. Par exemple pour 5 degrés de liberté, le seuil de la région critique est de 2,56 ; pour 500 degrés de liberté de 1,96 (voire figure 3.1)

Ce test est parfois caractérisé par ce que l'on appelle la p-value. Il s'agit à contrario du niveau du test pour lequel la statistique observée serait le seuil. Elle est donc définie par la quantité $\widehat{p} - value = P\left(|S| > \left|\widehat{S}\right|\right) = 2\left(1 - F\left(\left|\widehat{S}\right|\right)\right)$ lorsque S suit une loi de Student à N-K-1 degrés de liberté. On acceptera l'hypothèse nulle pour un test au niveau α si la $\widehat{p} - value$ est supérieure à α. En effet compte tenu du fait que $F(t_{N-K-1}(1-\alpha/2)) = 1-\alpha/2$, on a $2(1 - F(t_{N-K-1}(1-\alpha/2))) = \alpha$

$$\widehat{p} - value > \alpha \Longleftrightarrow \left|\widehat{S}\right| < t_{N-K-1}(1-\alpha/2)$$

Un test systématiquement mis en oeuvre est le test dit de significativité des paramètres. Il correspond à l'hypothèse nulle $b_k = 0$. La statistique de Student associée à ce test, nommée t de Student est définie par $\widehat{b}_k/\widehat{\sigma}_k$. En général n'importe quelle estimation d'un modèle linéaire fait par défaut l'hypothèse de normalité des résidus. Elle produit la valeur estimée du paramètre la valeur estimée de l'écart-type, la valeur du t de Student (correspondant à l'hypothèse de significativité du paramètre) et la p-value correspondant à ce test.

3.3 Un exemple

Pour illustrer les tests et leur utilisation, on peut calculer la fonction de puissance du test lorsque la vraie valeur du paramètre varie. On va considérer un modèle à une unique

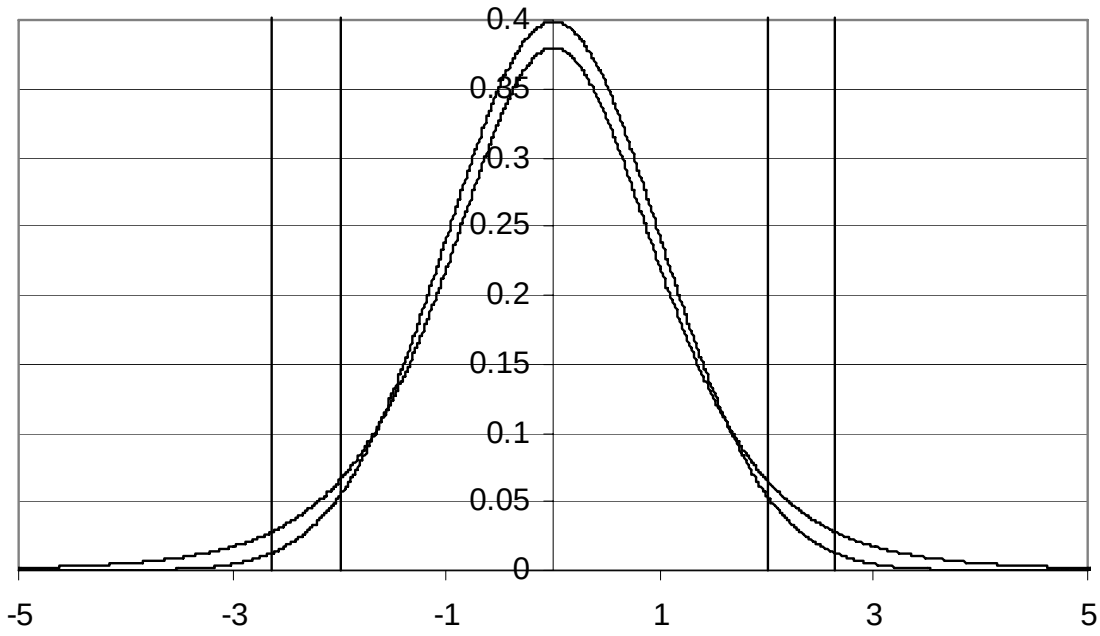


FIG. 3.1 – Distribution de Student pour 5 et 500 degrés de liberté

variable

$$y = 1 + xb_0 + u$$

et on va simuler ce modèle pour différente vraie valeur du paramètre, allant de 0 à 2. On va s'intéresser au test de l'hypothèse $H_0 : b = 1$. Pour calculer la fonction de puissance en un point donné b_0, on utilise des simulations. On titre un échantillon Ech_1 avec b_0 comme vraie valeur du paramètre. Sur cet échantillon on applique le test. On retient la décision $d_1 = 1$ si on rejette et $d_1 = 0$ sinon. On réplique cette opération avec la même vraie valeur sur M échantillons, avec M grand. On a ainsi un ensemble de valeur $(d_i)_{i \leq M}$. On approxime la valeur de la fonction de puissance par $\phi(b_0) = \overline{d_i}$. C'est bien un estimateur du nombre de fois ou on a rejeté à raison l'hypothèse. Bien sur, lorsque $b_0 = 1$, la quantité calculée n'est pas la puissance mais le risque de première espèce. On peut procéder ainsi pour différentes taille d'échantillons. On considère le cas dans lequel il n'y a que 20 observations, puis on augmente progressivement ce nombre. On considère respectivement $N = 50, 100, 500, 2000$. La figure 3.2 montre le résultat de ces estimations. On voit que le graphe de la fonction de puissance a une forme de vasque. Si on se situe au niveau de la valeur testée $b_0 = 1$, on trouve bien que la proportion de rejet est de 5%, correspondant au risque de première espèce, et ce quelque soit le nombre d'observations. Lorsque l'on s'écarte de la vraie valeur on voit que la courbe croît : on rejette de plus en plus souvent le paramètre. La croissance est très vive lorsque le nombre d'observation est grand : si la vraie valeur est de 0.95, on va rejeter l'hypothèse dans 60% des cas. Par contre, dans le cas de 20 observations, il faut que la vraie valeur s'écarte de plus de 0.5 pour que l'on atteigne des taux de rejet similaire. Ce résultat mérite d'être noté : avec un petit nombre d'observations, on est amené à accepter à tort l'hypothèse dans 40% des cas même lorsque la vraie valeur est assez éloignée. Lorsque l'écart à la valeur testée augmente, la probabilité de rejet tend vers 1. Cette valeur est très rapidement atteinte lorsque le nombre d'observations est grand, pour des nombres plus petits il faut des écarts plus importants.

Remarque *Dans le cas où la variance des résidus est connu, on peut très facilement calculer la fonction de puissance. En effet dans ce cas*

$$\sqrt{N}\frac{\widehat{b} - b_0}{\sigma/\sigma_x} \rightsquigarrow \mathcal{N}(0,1)$$

Sous $H_0 : b_0 = 1$, *on a donc*

$$\sqrt{N}\frac{\widehat{b} - 1}{\sigma/\sigma_x} \rightsquigarrow \mathcal{N}(0,1)$$

et a région critique du test est

$$W = \left\{ \sqrt{N}\frac{\widehat{b} - 1}{\sigma/\sigma_x} < q_{n,\alpha/2} \right\} \cup \left\{ \sqrt{N}\frac{\widehat{b} - 1}{\sigma/\sigma_x} > q_{n,1-\alpha/2} \right\}$$

c'est à dire en faisant intervenir la vraie valeur du paramètre

$$W = \left\{ \sqrt{N}\frac{\widehat{b}-b_0}{\sigma/\sigma_x} < q_{n,\alpha/2} + \sqrt{N}\frac{b_0-1}{\sigma/\sigma_x} \right\} \cup \left\{ \sqrt{N}\frac{\widehat{b}-b_0}{\sigma/\sigma_x} > q_{n,1-\alpha/2} + \sqrt{N}\frac{b_0-1}{\sigma/\sigma_x} \right\}$$

On en déduit facilement la fonction de puissance

$$P(b_0) = \Phi\left(q_{n,\alpha/2} + \sqrt{N}\frac{b_0-1}{\sigma/\sigma_x} \right) + 1 - \Phi\left(q_{n,1-\alpha/2} + \sqrt{N}\frac{b_0-1}{\sigma/\sigma_x} \right)$$

On voit qu'au voisinage de $b_0 = 1$, *la fonction de puissance se développe en*

$$P(b_0) = \alpha + q_{n,1-\alpha/2}\phi\left(q_{n,1-\alpha/2}\right) N \left(\frac{b_0-1}{\sigma/\sigma_x} \right)^2$$

Comme la fonction $x\phi(x)$ *est décroissante pour* $x > 1$, *que pour des valeurs de* α *faibles* $q_{n,1-\alpha/2}$ *est plus grand que 1 et que* $q_{n,1-\alpha/2}$ *croit avec* α, *plus* α *est élevé, plus* $q_{n,1-\alpha/2}\phi\left(q_{n,1-\alpha/2}\right)$ *est grand. On voit que dans ces conditions, les tests ayant des risques de première espèce faibles auront peu de puissance pour des vraies valeurs au voisinage de la valeur traitée. On voit aussi que la dépendance dans la taille de l'échantillon est en N. Il est clair que lorsque N tend vers l'infini la puissance du test tend vers 1. Pour étudier la puissance d'un test on s'intéresse en général à ce que l'on appelle des alternatives locales en déterminant la puissance pour*

$$b_0(N) = 1 + \beta/\sqrt{N}$$

où 1 *est la valeur testée et* β *une direction donnée dans l'espace des paramètre (ici comme le paramètre est de dimension 1 cette caractéristique tombe).*

3.4 Comparaison avec l'estimateur du Maximum de Vraisemblance

On peut aussi directement appliquer l'estimateur du maximum de vraisemblance aux données. La vraisemblance s'écrit :

$$L\left(\underline{y}, \underline{x}, b, \sigma^2\right) = -0.5N\log(2\pi) - 0.5N\log\left(\sigma^2\right) + 0.5(y-xb)'(y-xb)/\sigma^2$$

Proposition *L'estimateur du maximum de vraisemblance du paramètre b est identique à l'estimateur des mco. Il a les mêmes propriétés que l'estimateur des mco : sous les hypothèses* $H1 - H4$ *&* H_n, *il suit une loi normale centrée sur la vraie valeur du paramètre et ayant pour matrice de variance* $V_b = \sigma^2 E(x'x)^{-1}$. *L'estimateur du maximum de vraisemblance du paramètre du second ordre* σ^2 *se déduit linéairement de l'estimateur des mco de ce paramètre par application d'un facteur* $(N-K-1)/N$. *Cet estimateur n'est donc pas sans biais, mais il est indépendant de l'estimateur du MV du paramètre b.*

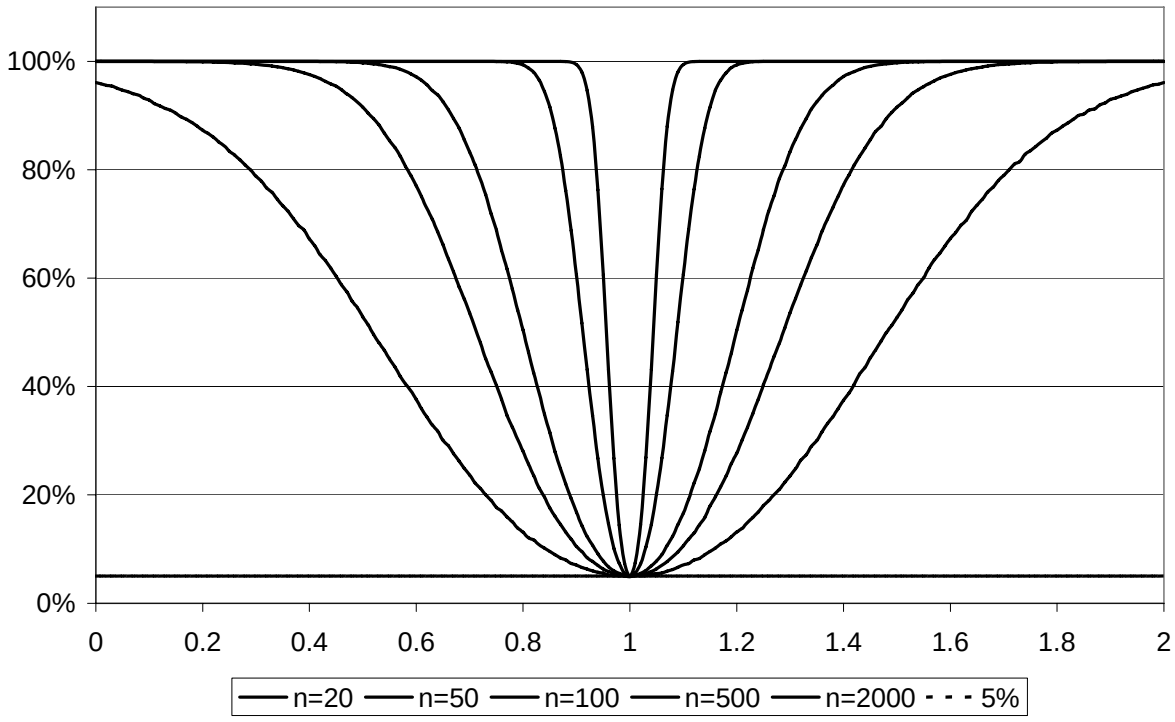


FIG. 3.2 – Fonction de puissance du test de Student en fonction du nombre d'obseravtions

3.5 Résumé

1. Dans ce chapitre on a examiné les propriétés de l'estimateur des mco lorsque la loi de $\underline{u}$ conditionnellement aux variables explicatives $\underline{x}$ est une loi normale de moyenne nulle et de matrice de variance $\sigma^2 I_N$.
2. On a montré que l'estimateur des mco suit une loi normale, que l'estimateur de la variance des résidus suit convenablement normalisé une loi du χ^2 et que ces deux estimateurs sont indépendants.
3. On a vu que l'on pouvait utiliser ces résultats pour obtenir un estimateur sans biais de la matrice de variance de l'estimation du paramètre.
4. On a vu que pour une composante donnée k du paramètre $\frac{\widehat{b}_k - b_k}{\widehat{\sigma}_k} \rightsquigarrow Student\,(N-K-1)$
5. On a appliqué ce résultat pour définir une région de confiance pour le paramètre et mettre en oeuvre des tests.
6. On a vu en particulier que la région critique pour le test de significativité d'un paramètre correspondait à des valeurs extrêmes du t de Student. Le caractère extrême s'appréciant par rapport au niveau du test.

3.6 Annexe : Distribution de la norme de la projection d'un vecteur normal

Considérons $Z \rightsquigarrow N(0, I_L)$, *et* P *est un projecteur orthogonal sur un sous espace de dimension* L_1 *alors* $Z'PZ \rightsquigarrow \chi^2(L_1)$.

L'hypothèse sur P revient à dire que P est une matrice symétrique et que ses valeurs propres sont 0 ou 1. Comme P est symétrique, on peut la diagonaliser dans le groupe orthogonal. On peut donc écrire $P = Q'\widetilde{P}Q$, avec $Q'Q = I_L$ et $\widetilde{P} = Diag(\underbrace{1, \ldots, 1}_{L_1 \text{éléments non nuls}}, 0, \ldots 0)$

On définit $Z^* = QZ$.

Z^* est aussi un vecteur normal $N(0, I_L)$ puisque

1. C'est un vecteur normal puisqu'il est combinaison linéaire d'un vecteur normal
2. Il est d'espérance nulle puisque $E(Z^*) = E(QZ) = QE(Z) = 0$
3. Il est de variance identité puisque $V(Z^*) = E(Z^*Z^{*\prime}) = E(QZZ'Q') = QE(ZZ')Q' = QI_LQ' = QQ' = I_L$

On a alors $Z'PZ = Z'Q'\widetilde{P}QZ = Z^{*\prime}\widetilde{P}Z^* = \sum_{l=1}^{L_1} Z_l^{*2}$. C'est donc la somme du carré de L_1 variables normales indépendantes de moyenne nulle et de variance 1. Par définition elle suit un $\chi^2(L_1)$

Chapitre 4

Estimation sous contraintes linéaires

On peut souhaiter estimer un modèle économétrique linéaire en incorporant une information a priori sur les paramètres prenant la forme de contraintes linéaires. On peut aussi vouloir tester si certaines relations entre les paramètres sont bien acceptées par les données. Les résultats obtenus au chapitre précédent ont montré comment tester des hypothèses très simples, s'écrivant sous la forme $H_0 : b_k = b_k^0$, où b_k^0 est une valeur donnée. On va examiner ici un cas un peu plus général dans lequel les hypothèses que l'on veut tester, ou bien les contraintes que l'on veut imposer font intervenir une ou plusieurs combinaisons linéaires des paramètres. On va montrer obtenir un estimateur différent de celui des moindres carrés ordinaires, appelé estimateur des moindres carrés contraints (mcc) et on va montrer ses deux propriétés principales : l'estimateur des mcc est toujours plus précis que l'estimateur des mco ; l'estimateur des mcc est non biaisé seulement si la vraie valeur du paramètre satisfait les contraintes imposées. Il y a donc un arbitrage entre robustesse et précision des estimateurs. Un tel arbitrage est très fréquent en économétrie. On va aussi introduire un test très utilisé permettant de tester des contraintes linéaire. Ce test est connu sous le nom de test de Fisher, et on va voir comment le mettre en oeuvre simplement à partir de deux régressions, l'une par les mcc et 'autre par les mco.

Exemple *Homogénéité du progrès technique. On considère une fonction de production faisant intervenir le capital et le travail. On fait l'hypothèse que le facteur travail n'est pas homogène. Il fait intervenir différents types de main d'oeuvre, pas tous aussi efficace les uns que les autres.*

$$Y = F\left(A_{CI}CI, A_K K, A_1 L_1, \ldots, A_M L_M\right)$$

La dérivée logarithmique s'écrit donc

$$\begin{aligned} d\log Y \;=\; & \frac{d\log F}{d\log CI}\left(d\log CI + d\log A_{CI}\right) + \frac{d\log F}{d\log K}\left(d\log K + d\log A_K\right) + \\ & \frac{d\log F}{d\log L_1}\left(d\log L_1 + d\log A_1\right) + \cdots + \frac{d\log F}{d\log L_M}\left(d\log L_M + d\log A_M\right) \end{aligned}$$

Sous l'hypothèse de rendements constants et de concurrence parfaite sur le marché des biens et des produits, la part de la rémunération de chaque facteur dans la production est égale à l'élasticité de la production. On peut donc mesurer $\frac{d\log F}{d\log CI} = \pi_{CI} = \frac{c_{CI}CI}{Y}$, $\frac{d\log F}{d\log K} = \pi_K = \frac{c_k K}{Y}$ *et* $\frac{d\log F}{d\log L_m} = \pi_m = \frac{w_m L_m}{Y}$. *On a donc l'équation :*

$$d\log SR = \pi_{CI} d\log A_{CI} + \pi_K d\log A_k + \pi_1 d\log A_1 + \cdots + \pi_M d\log A_M$$

où $d\log SR = d\log Y - \pi_{CI} d\log CI - \pi_K d\log K - \pi_1 d\log L_1 - \cdots - \pi_M d\log L_M$ *mesure le Résidu de Solow, c'est à dire la part de la croissance qui n'est pas expliquée par celle des facteurs de production. On suppose que les entreprises peuvent ou non adopter une innovation. On considère* I *une variable indicatrice prenant la valeur* 1 *si une entreprise a adopté une innovation et 0 sinon. On modélise*

$$d\log A_m = a_{0m} + a_{Im} I + u$$

Les gains d'efficacité des facteurs de production font donc intervenir un terme fixe propre au facteur, un terme dépendant du fait que l'entreprise ait innové et un terme aléatoire commun à tous les facteurs. On obtient alors l'équation

$$\begin{aligned} d\log SR \;=\; & \pi_K.(a_{0K} - a_{0CI}) + \pi_1.(a_{01} - a_{0CI}) + \cdots + \pi_M(a_{0M} - a_{0CI}) + \\ & + I\pi_{CI}.a_{ICI} + I\pi_K.a_{IK} + I\pi_1.a_{I1} + \cdots + I\pi a_{IM} + u \end{aligned}$$

où on utilise le fait que la somme des parts vaut 1. Les régresseurs sont donc les parts des facteurs et les parts des facteurs interragies avec la variable d'innovation. On peut sur cette base formuler un certain nombre d'hypothèses :

- *H0(L) : Homogénéité de l'effet de l'innovation sur le facteur travail.*

$$a_{I1} = \cdots = a_{IM}$$

- *H0(L,K,CI) : Homogénéité de l'effet de l'innovation sur les facteurs.*

$$a_{ICI} = a_{IK} = a_{I1} = \cdots = a_{IM}$$

- *H0(L=K=CI=0) : Absence d'effet de l'innovation sur les facteurs.*

$$a_{ICI} = a_{IK} = a_{I1} = \cdots = a_{IM} = 0$$

- *H0(K=CI=0) : Absence d'effet de l'innovation sur le capital et les consommations intermédiaires.*

$$a_{ICI} = a_{IK} = 0$$

- *H0(K=CI=0,L) : Absence d'effet de l'innovation sur le capital et les consommations intermédiaires et homogénéité sur le travail.*

$$a_{CI} = a_{IK} = 0,\ a_{I1} = \cdots = a_{IM}$$

Le nombre de contraintes est bien sûr différent d'une hypothèse à l'autre

Hypothèse	Nombre de contraintes
$H0(L)$	$M-1$
$H0(L,K,CI)$	$M+1$
$H0(L=K=CI=0)$	$M+2$
$H0(K=CI=0)$	2
$H0(K=CI=0,L)$	$M+1$

Plusieurs questions se posent :

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 : $\widehat{b}_{mcc}$

2. Quelles sont les conséquences de cette prise en compte pour les estimations obtenues ?

   On va voir que les estimations obtenues sont toujours plus précises que celles des mco mais que par contre elles ne sont sans biais que si la contrainte imposée est vérifiée par la vraie valeur du paramètre. Il y a donc un arbitrage que l'on retrouve souvent en économétrie, entre robustesse et efficacité. La robustesse correspond à l'obtention d'estimateurs non biaisés sous des hypothèses plus faibles. Ici l'estimateur des mco est robuste car il est sans biais que les contraintes soient satisfaites ou non par la vraie valeur du paramètre. L'efficacité correspond à l'obtention d'estimateurs les plus précis possibles. Ici l'estimateur des mco n'est pas le plus efficace puisque l'estimateur des mcc a une variance plus faible.

3. Peut-on tester l'information a priori ?

   Dans le cas présent, on pourrait tester l'hypothèse de constance des rendements avec un test de Student. Néanmoins, on va voir que dans le cas général, lorsqu'il y a plus d'une contrainte, un tel test n'est plus suffisant. On va introduire un test très courant qui généralise le test de Student : le test de Fisher. Comme précédemment, alors que l'on peut répondre aux deux questions précédentes dans un cadre général ne faisant des hypothèses que sur les moments d'ordre 1 et 2 des perturbations conditionnellement aux variables explicatives, la possibilité d'effectuer des tests requière de spécifier la loi conditionnelle des perturbations

4.1 Formulation

On considère le modèle linéaire :

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

dans lequel on fait les hypothèses H1-H4 et pour lequel la vraie valeur du paramètre vérifie le système de p contraintes linéaires :

$$Rb = r$$

R est une matrice donnée $p \times (K+1)$, et r un vecteur donné $p \times 1$.
Il y a de toutes évidences des contraintes qui pèsent sur cette formulation.

1. Il ne doit pas y avoir de contraintes redondantes. Ceci impose que $R'\lambda = 0 \Longrightarrow \lambda = 0$
2. Il doit y avoir une solution non unique à l'équation $Rb = r$

Ces deux contraintes imposent que R soit de rang p et que le nombre de contraintes p soit au maximum égal à $(K+1)-1$. En effet si on en avait $K+1$ ou plus, on pourrait en sélectionner $K+1$ par exemple $R_1 b = r_1$ et on pourrait alors calculer le paramètre $b = R_1^{-1}\, r_1$. il n'y aurait plus de problème d'estimation.

Exemple *Considérons à nouveau l'exemple précédent. Le modèle s'écrit*

$$\begin{aligned} d\log SR &= \pi_K.b_{0K} + \pi_1.b_{01} + \cdots + \pi_M b_{0M} + \\ &\quad + I\pi_{CI}.a_{ICI} + I\pi_K.a_{IK} + I\pi_1.a_{I1} + \cdots + I\pi a_{IM} + u \end{aligned}$$

Dans le cas de l'hypothèse H0L : $a_{I1} = \cdots = a_{IM}$, on peut écrire les contraintes sur les paramètres comme $a_{I2} - a_{I1} = 0, \cdots, a_{IM} - a_{I1} = 0$, c'est à dire :

$$\begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & \vdots & 0 & \ddots & 0 \\ 0 & 1 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} (b', a_{ICI}, a_{IK})' \\ a_{I1} \\ a_{I2} \\ \vdots \\ a_{IM} \end{pmatrix} = 0$$

4.2 L'Estimateur des Moindres Carrés Contraints (MCC)

Definition *L'estimateur $\hat{b}_{mcc}$ de b est défini comme le paramètre minimisant la somme des carrés des résidus et satisfaisant les contraintes $Rb = r$:*

$$\begin{aligned} &\min_b\, (\underline{y} - \underline{x}\, b)'(\underline{y} - \underline{x}\, b) \\ &Sc : Rb = r \end{aligned}$$

Proposition *L'estimateur des MCC a pour expression*

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

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

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

On voit directement sur cette expression que l'estimateur des MCC apporte une correction à l'estimateur $\hat{b}_{mco}$ et que cette correction est d'autant plus importante que $R\hat{b}_{mco} - r \neq 0$. Dans le cas où $R\hat{b}_{mco} = r$, les deux estimateurs sont identiques.

Démonstration *Pour trouver l'expression de l'estimateur on écrit le Lagrangien :*

$$L = \frac{1}{2}(\underline{y} - \underline{x}b)'(\underline{y} - \underline{x}b) + (Rb - r)'\lambda$$

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

$$\begin{aligned} \left.\frac{\partial L}{\partial b}\right|_{mcc} &= -\underline{x}'\,\underline{y} + (\underline{x}'\underline{x})\,\hat{b}_{mcc} + R'\,\hat{\lambda} = 0 \\ \left.\frac{\partial L}{\partial \lambda}\right|_{mcc} &= R\hat{b}_{mcc} - r = 0 \end{aligned}$$

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

Introduit dans la deuxième condition il vient l'expression $R\,(\underline{x}'\underline{x})^{-1}\left(\underline{x}'\underline{y} - R'\,\hat{\lambda}\right) = r$ *soit* $R\,(\underline{x}'\underline{x})^{-1}R'\,\hat{\lambda} = R\,(\underline{x}'\underline{x})^{-1}\underline{x}'\underline{y} - r$

dont on tire $\hat{\lambda} = \left[R\,(\underline{x}'\underline{x})^{-1}R'\right]^{-1}\left[R\,(\underline{x}'\underline{x})^{-1}\underline{x}'\underline{y} - r\right]$

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

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

4.3 Espérance et variance de $\hat{b}_{mcc}$

Proposition *Sous l'hypothèse H2 et sous l'hypothèse H_c : $Rb = r$, l'estimateur des mcc est sans biais. En revanche, sous l'hypothèse H2 seule, l'estimateur est biaisé et le biais dépend linéairement de $Rb - r$*

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

Sa variance est donnée sous $H2 - H4$ par

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

indépendamment de l'hypothèse H_c

Ainsi l'estimateur des moindres carrés contraints est potentiellement biaisé, mais on voit qu'il est aussi plus précis que l'estimateur des mco. Sa variance est en effet donnée par :

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

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

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

Il y a donc un arbitrage entre robustesse et efficacité. Introduire plus de contraintes améliore la précision des estimations mais risque de conduire à des estimateurs biaisé. A l'inverse, moins de contraintes produit des estimateurs plus robustes mais moins précis.

Démonstration *Compte tenu de l'expression*

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

et du fait que $\hat{b}_{mco}$ est un estimateur linéaire sans biais de b sous l'hypothèse $H2$:

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

On voit donc que sous l'hypothèse $H_c : Rb = r$, on a $E\left(\hat{b}_{mcc}\,|\underline{x}\right) = b$. En revanche si les contraintes ne sont pas satisfaites il existe un biais

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

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

On vérifie que ce biais est systématiquement différent de 0. En effet, si $Rb - r \neq 0$ alors $\lambda = \left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1}[Rb - r]$ est aussi différent de 0 et donc $B = -(\underline{x}'\underline{x})^{-1}R'\lambda$. Comme les contraintes sont non redondantes, et $\lambda \neq 0$, on ne peut avoir $R'\lambda = 0$.

On a en outre

$$\begin{aligned}\hat{b}_{mcc} - E\left(\hat{b}_{mcc}\,|\underline{x}\right) &= \left(\hat{b}_{mco} - b\right) - (\underline{x}'\underline{x})^{-1}\,R'\left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1}R\,\left(\hat{b}_{mco} - b\right)\\ &= \left[I - (\underline{x}'\underline{x})^{-1}\,R'\left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1}R\right](\underline{x}'\underline{x})^{-1}\underline{x}'\underline{u}\end{aligned}$$

On introduit $H = (\underline{x}'\underline{x})^{-1}R'\left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1}R$. Cette matrice vérifie les propriétés suivantes

$$\begin{aligned}H^2 &= H\\ H(\underline{x}'\underline{x})^{-1} &= (\underline{x}'\underline{x})^{-1}H'\\ H(\underline{x}'\underline{x})^{-1}H' &= H^2(\underline{x}'\underline{x})^{-1} = H(\underline{x}'\underline{x})^{-1}\end{aligned}$$

On a donc

$$\hat{b}_{mcc} - E\left(\hat{b}_{mcc}\,|\underline{x}\right) = [I - H]\,(\underline{x}'\underline{x})^{-1}\underline{x}'\underline{u}$$

Par conséquent comme $E\left[\underline{u}\underline{u}'\,|\underline{x}\right] = \sigma^2 I$ *:*

$$\begin{aligned} V\left(\hat{b}_{mcc}\,|\underline{x}\right) &= E\left[\left(\hat{b}_{mcc} - E\left(\hat{b}_{mcc}\,|\underline{x}\right)\right)\left(\hat{b}_{mcc} - E\left(\hat{b}_{mcc}\,|\underline{x}\right)\right)'\,|\underline{x}\right] \\ &= E\left[\left[\,I - H\right](\underline{x}'\underline{x})^{-1}\underline{x}'\underline{u}\underline{u}'\underline{x}(\underline{x}'\underline{x})^{-1}\left[\,I - H'\right]|\underline{x}\right] \\ &= \sigma^2\left[\,I - H\right](\underline{x}'\underline{x})^{-1}\left[\,I - H'\right] \end{aligned}$$

En développant, compte tenu des propriétés de H

$$\begin{aligned} V\left(\hat{b}_{mcc}\,|\underline{x}\right) &= \sigma^2\left[(\underline{x}'\underline{x})^{-1} - H(\underline{x}'\underline{x})^{-1} - H'(\underline{x}'\underline{x})^{-1} + H(\underline{x}'\underline{x})^{-1}H'\right] \\ &= \sigma^2\left[(\underline{x}'\underline{x})^{-1} - H(\underline{x}'\underline{x})^{-1}\right] \end{aligned}$$

Le résultat provient de l'expression $H(\underline{x}'\underline{x})^{-1} = (\underline{x}'\underline{x})^{-1}R'\left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1}R(\underline{x}'\underline{x})^{-1}$

4.4 Estimateur de la variance des résidus σ^2

Comme pour l'estimateur des mco, on peut définir le vecteur des résidus estimés

$$\widehat{\underline{u}}_c = \underline{y} - \underline{x}\,\hat{b}_{mcc}$$

On peut comme dans le cas des mco définir un estimateur de la variance des résidus à partir de la somme des carrés de ces résidus.

Lemme *On peut écrire le vecteur des résidus estimés dans le modèle contraint comme la somme de deux termes orthogonaux, le vecteur des résidus estimés par les mco d'une part et un terme appartenant à l'espace engendré par les* $\underline{x}$ *d'autre part*

$$\widehat{\underline{u}}_c = \widehat{\underline{u}} + P_c\underline{u} = \widehat{\underline{u}} + \widetilde{\underline{u}}$$

où $P_c = \underline{x}(\underline{x}'\underline{x})^{-1}R'\left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1}R(\underline{x}'\underline{x})^{-1}\underline{x}'$ *est un projecteur orthogonal sur un sous espace de l'espace engendré par les* $\underline{x}$*.*

Démonstration *On a l'expression de* $\widehat{\underline{u}}_c$

$$\begin{aligned} \widehat{\underline{u}}_c &= \underline{x}b + \underline{u} - \underline{x}\,\hat{b}_{mcc} = \left[I - \underline{x}\left[I - H\right](\underline{x}'\underline{x})^{-1}\underline{x}'\right]\underline{u} \\ &= \left[M_x + \underline{x}H(\underline{x}'\underline{x})^{-1}\underline{x}'\right]\underline{u} \end{aligned}$$

avec $M_x = \left(I - \underline{x}\,(\underline{x}'\underline{x})^{-1}\,\underline{x}'\right)$*. On introduit*

$$P_c = \underline{x}H(\underline{x}'\underline{x})^{-1}\underline{x}' = \underline{x}(\underline{x}'\underline{x})^{-1}R'\left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1}R(\underline{x}'\underline{x})^{-1}\underline{x}'$$

On a directement $P_c^2 = P_c$ *et* $P_c' = P_c$*. En outre* $P_c z = \underline{x}\,(H(\underline{x}'\underline{x})^{-1}\underline{x}'z)$ *appartient à l'espace engendré par les* $\underline{x}$*.*

Proposition *Sous les hypothèses $H2-H4$, et H_c, l'estimateur de la variance des résidus*

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

est sans biais.

Une différence importante avec l'estimateur issu des mco correspond au nombre de degrés de liberté. Ici il s'agit de $N-K-1+p$. Avec l'estimateur des mco, le nombre de degrés de liberté est plus faible : $N-K-1$.

Démonstration *L'expression de $\underline{\widehat{u}}_c$: $\underline{\widehat{u}}_c = \underline{\widehat{u}} + P_c\underline{u} = \underline{\widehat{u}} + \underline{\widetilde{u}}$ conduit directement à*

$$\underline{\widehat{u}}_c'\underline{\widehat{u}}_c = \underline{\widehat{u}}'\underline{\widehat{u}} + \underline{\widetilde{u}}'\underline{\widetilde{u}}$$

$\underline{\widehat{u}}$ et $\underline{\widetilde{u}}$ sont en effet orthogonaux puisque $\underline{\widehat{u}}$ est la projection de $\underline{u}$ sur l'orthogonal de $\underline{x}$ et $\underline{\widetilde{u}}$ une projection de $\underline{u}$ sur un sous espace de l'espace engendré par les $\underline{x}$. Donc

$$E\left(\underline{\widehat{u}}_c'\underline{\widehat{u}}_c\,|\underline{x}\right) = E\left(\underline{\widehat{u}}'\underline{\widehat{u}}\,|\underline{x}\right) + E\left(\underline{\widetilde{u}}'\underline{\widetilde{u}}\,|\underline{x}\right) = \sigma^2\left[(N-K-1)+Tr\left(P_c\right)\right]$$

En outre

$$\begin{aligned} Tr\left(P_c\right) &= TR\left(\underline{x}(\underline{x}'\underline{x})^{-1}R'\left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1}R(\underline{x}'\underline{x})^{-1}\underline{x}'\right) \\ &= TR\left(\left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1}R(\underline{x}'\underline{x})^{-1}\underline{x}'\underline{x}(\underline{x}'\underline{x})^{-1}R'\right) \\ &= Tr\left(I_p\right) = p \end{aligned}$$

4.5 Loi de l'estimateur des moindres carrés contraints

Comme dans le cas non contraint, on peut préciser la loi de l'estimateur des moindres carrés contraints lorsque les résidus sont distribués suivant une loi normale. On fait ici l'hypothèse que les contraintes sont satisfaites, c'est à dire que la vraie valeur du paramètre b_0 satisfait effectivement $Rb_0 = r$

Les résultats du Théorème de Cochran se généralisent

Proposition *Sous l'hypothèse H_n :*

1. *L'estimateur du paramètre des mco $\widehat{b}_{mcc}$ est distribué comme une loi normale de moyenne b, la vraie valeur du paramètre, et de matrice de variance $V\left(\hat{b}_{mcc}\,|\underline{x}\right) = \sigma^2\left[(\underline{x}'\underline{x})^{-1} - H(\underline{x}'\underline{x})^{-1}\right]$*
2. *L'estimateur $\widehat{\sigma}_{mcc}^2$, convenablement normalisé, est distribué suivant une loi du χ^2*

$$\left[N-(K+1)+p\right]\frac{\widehat{\sigma}_{mcc}^2}{\sigma^2} \sim \chi^2\left(N-(K+1)+p\right)$$

3. $\widehat{b}_{mcc}$ *et* $\widehat{\sigma}^2_{mcc}$ *sont indépendants*
4. *Considérant la* $k^{ième}$ *composante de l'estimateur, on a* $\widehat{b}_{mcc}(k) - b_0(k) \Big/ \widehat{\sigma}_{mcc}(k)$ *suit une loi de Student à* $N - (K+1) + p$ *degrés de liberté*

Démonstration *Compte tenu de l'expression de* $\hat{b}_{mcc}$

$$\begin{aligned}\hat{b}_{mcc} &= (\underline{x}'\underline{x})^{-1}\,\underline{x}'\underline{y} - (\underline{x}'\underline{x})^{-1}\,R'\left[R(\underline{x}'\underline{x})^{-1}\,R'\right]^{-1}\,\left[R(\underline{x}'\underline{x})^{-1}\,\underline{x}'\underline{y} - r\right] \\ &= b + (\underline{x}'\underline{x})^{-1}\,\underline{x}'\underline{u} - (\underline{x}'\underline{x})^{-1}\,R'\left[R(\underline{x}'\underline{x})^{-1}\,R'\right]^{-1}\,R(\underline{x}'\underline{x})^{-1}\,\underline{x}'\underline{u}\end{aligned}$$

lorsque les contraintes sont satisfaites, on voit directement que l'estimateur est normal lorsque les résidus sont normaux puisque l'estimateur est une combinaison linéaire du résidu. On a en outre

$$\hat{b}_{mcc} - b = (I - H)\,(\underline{x}'\underline{x})^{-1}\,\underline{x}'\underline{u}$$

et

$$\widehat{\underline{u}}_c = [M_x + P_c]\,\underline{u}$$

avec $M_x = \left(I - \underline{x}\,(\underline{x}'\underline{x})^{-1}\,\underline{x}'\right)$ *et* $P_c = \underline{x}H(\underline{x}'\underline{x})^{-1}\underline{x}'$. *On vérifie donc sans peine que* $\hat{b}_{mcc}$ *et* $\widehat{\underline{u}}_c$ *sont non corrélés et donc indépendants :*

$$\begin{aligned}E\left(\left(\hat{b}_{mcc} - b\right)\widehat{\underline{u}}_c'\right) &= E\left((I-H)\,(\underline{x}'\underline{x})^{-1}\,\underline{x}'\underline{u}\underline{u}'\,[M_x + P_c']\right) = \sigma^2\,(I-H)\,(\underline{x}'\underline{x})^{-1}\,\underline{x}'\,[M_x + P_c'] \\ &= \sigma^2\,(I-H)\,(\underline{x}'\underline{x})^{-1}\,\underline{x}'P_c' = \sigma^2\,(I-H)\,(\underline{x}'\underline{x})^{-1}\underline{x}'\underline{x}H(\underline{x}'\underline{x})^{-1}\underline{x}' \\ &= \sigma^2\,(I-H)\,H(\underline{x}'\underline{x})^{-1}\underline{x}' = 0\end{aligned}$$

puisque $H^2 = H$. *Les points qui suivent sont immédiats.*

Exemple *On peut mettre en oeuvre les estimations de la fonction de production avec innovation. On dispose d'un échantillon de 3627 observations. On a introduit une distinction entre travailleurs jeunes et vieux. Le nombre de catégorie de travailleurs considéré est donc* $M = 2$. *On considère la régression sous l'hypothèse alternative* $H1$

$$d\log SR = \pi_K.b_{0K} + \pi_1.b_{01} + \cdots + \pi_M b_{0M} + I\pi_{CI}a_{ICI} + I\pi_K.a_{IK} + I\pi_L a_{IL} + Xd + u$$

ainsi que les différentes spécifications contraintes introduites précédemment :

- *H0(L) : Homogénéité de l'effet de l'innovation sur le facteur travail.* $a_{I1} = \cdots = a_{IM}$
- *H0(L,K,CI) : Homogénéité de l'effet de l'innovation sur les facteurs.* $a_{ICI} = a_{IK} = a_{I1} = \cdots = a_{IM}$
- *H0(L=K=CI=0) : Absence d'effet de l'innovation sur les facteurs.* $a_{ICI} = a_{IK} = a_{I1} = \cdots = a_{IM} = 0$
- *H0(K=CI=0) : Absence d'effet de l'innovation sur le capital et les consommations intermédiaires.* $a_{ICI} = a_{IK} = 0$
- *H0(K=CI=0,L) : Absence d'effet de l'innovation sur le capital et les consommations intermédiaires et homogénéité sur le travail.* $a_{IK} = 0$, $a_{I1} = \cdots = a_{IM}$

Les résultats sont reportés dans le tableau 4.1. Pour chacune des spécifications on reporte la valeur estimée du coefficient ainsi que l'écart-type estimé. Ces deux informations permettent de faire des tests d'hypothèses simples (en particulier de nullité de chaque coefficient pris individuellement). La loi suivie par les t de Student que l'on peut former est une loi de Student à 3627-12 degrés de liberté pour la spécification alternative $H1$. Il varie ensuite d'une colonne à l'autre suivant le nombre de contraintes introduites. Dans la première spécification, le nombre de contrainte est de 1, le nombre de degrés de liberté est donc 3627-12+1. En théorie les valeurs critiques des t de Student pour un test à un niveau α donné diffèrent d'une colonne à l'autre puisque la loi n'est pas la même. Néanmoins ici le nombre de degrés de liberté est grand et dans ce cas la distribution d'une loi de Student se confond avec celle d'une loi normale : la valeur critique est donc la même pour chaque régression. Dans le cas d'un test à 5% la valeur critique est ainsi de 1.96. On acceptera donc l'hypothèse de nullité de chaque paramètre pris individuellement si le ratio entre le coefficient et son écart-type est en valeur absolue inférieur à 1.96.

On voit sur les estimations du modèle non contraint que l'effet de l'innovation sur l'efficacité des facteurs semble assez différentes d'un facteur à l'autre. Le coefficient du capital apparaît négatif et grand en valeur absolue alors que le coefficient pour les jeunes est positif et grand. Néanmoins on voit que les estimations sont imprécises et les tests d'égalité des coefficients pris individuellement sont souvent acceptés. En fait seul le coefficient pour la part des jeunes est significativement différent de zéro. On est typiquement dans une situation dans laquelle les résultats sont robustes mais peu précis. On sent bien qu'il y a là moyen de gagner en précision de façon importante en imposant des contraintes supplémentaires.

On voit néanmoins que chacune des spécifications contraintes conduit à des modifications importantes des coefficients : si on impose l'homogénéité sur l'ensemble des facteurs, on parvient à une efficacité très faible pour chaque facteur. Si on impose en revanche la nullité pour le capital et les consommations intermédiaires et l'homogénéité sur le travail, on voit que l'effet sur le travail est important, de l'ordre de 0.05, significativement différent de zéro. Face à cette forte sensibilité des résultats aux hypothèses effectuées il est important de pouvoir mettre en oeuvre des tests qui permettront de guider le choix vers une spécification plus qu'une autre.

4.6 Estimation par intégration des contraintes

Le problème d'estimation sous contraintes peut se ramener au résultat classique d'estimation par la méthode des moindres carrés en intégrant directement les contraintes dans le modèle. On peut en effet utiliser les p contraintes pour exprimer p paramètres parmi les $k+1$ à estimer en fonction des $(k+1-p)$ autres paramètres.

Par exemple, on ré-écrit les contraintes $Rb = r$ comme :

	H1		H0(L)		H0(L,K,CI)		H0(L=K=CI=0)		H0(K=CI=0)		H0(K=CI=0,L)	
Constante	0.00	(0.01)	0.00	(0.01)	-0.01	(0.01)	0.00	(0.01)	-0.01	(0.01)	-0.01	(0.01)
part capital	0.08	(0.04)	0.08	(0.04)	0.04	(0.03)	0.04	(0.03)	0.04	(0.03)	0.04	(0.03)
part jeunes	0.15	(0.06)	0.18	(0.05)	0.21	(0.05)	0.20	(0.05)	0.17	(0.05)	0.20	(0.05)
part vieux	-0.03	(0.03)	-0.04	(0.03)	-0.01	(0.03)	-0.01	(0.03)	-0.01	(0.03)	-0.02	(0.03)
l*part capital	-0.11	(0.06)	-0.11	(0.06)	**0.01**	(0.01)	***0.00***	(0.00)	***0.00***	(0.00)	***0.00***	(0.00)
l*part CI	-0.01	(0.02)	-0.01	(0.02)	**0.01**	(0.01)	***0.00***	(0.00)	***0.00***	(0.00)	***0.00***	(0.00)
l*part jeunes	0.19	(0.09)	**0.09**	(0.03)	**0.01**	(0.01)	***0.00***	(0.00)	0.12	(0.09)	**0.05**	(0.02)
l*part vieux	0.06	(0.05)	**0.09**	(0.03)	**0.01**	(0.01)	***0.00***	(0.00)	0.02	(0.04)	**0.05**	(0.02)
Sect1	0.01	(0.01)	0.01	(0.01)	0.01	(0.01)	0.01	(0.01)	0.01	(0.01)	0.01	(0.01)
Sect2	0.03	(0.01)	0.03	(0.01)	0.03	(0.01)	0.03	(0.01)	0.03	(0.01)	0.03	(0.01)
Sect3	-0.01	(0.01)	-0.01	(0.01)	-0.01	(0.01)	-0.01	(0.01)	-0.01	(0.01)	-0.01	(0.01)
Sect4	0.02	(0.02)	0.02	(0.02)	0.02	(0.02)	0.02	(0.02)	0.02	(0.02)	0.02	(0.02)

TAB. 4.1 – Résultats des estimations par les MCC

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

avec R_1 une sous matrice de R de dimension $p \times p$ supposée inversible, R_2 une autre sous matrice de dimension $p \times (K+1-p)$, b_1 un vecteur de dimension $p \times 1$ et b_2 un vecteur de dimension $K+1-p \times 1$

On peut alors écrire $r = R_1\, b_1 + R_2\, b_2$ soit encore :

$$b_1 = R_1^{-1}\,[r - R_2\, b_2]$$

Par conséquent, le modèle peut se réécrire :

$$\underline{y} = \underline{x}_1\, b_1 + \underline{x}_2\, b_2 + \underline{u} = \underline{x}_1 \left[R_1^{-1}(r - R_2\, b_2)\right] + \underline{x}_2\, b_2 + \underline{u}$$

Ceci revient à estimer :

$$\underline{y} - \underline{x}_1 R_1^{-1}\, r = \left[\underline{x}_2 - \underline{x}_1\, R_1^{-1}\, R_2\right]\; b_2 + \underline{u}$$

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 : $b_1 = R_1^{-1}\, r - R_2\, b_2$

Exemple *Cette intégration peut en pratique être extrêmement simple. Si on reprend le cas de l'exemple précédent, dans le cas de la spécification* $H0L : a_{I1} = \cdots = a_{IM}$, *on écrit*

$a_{I2} = a_{I1}, \ldots, a_{IM} = a_{I1}$. *On a ainsi*

$$\begin{aligned} d\log SR &= \pi_K.b_{0K} + \pi_1.b_{01} + \cdots + \pi_M b_{0M} + I\pi_{CI}.a_{ICI} + I\pi_K.a_{IK} + \\ &\quad I\pi_1.a_{I1} + I\pi_2 a_{I1} + \cdots + I\pi_M a_{I1} + u \\ &= \pi_K.b_{0K} + \pi_1.b_{01} + \cdots + \pi_M b_{0M} + I\pi_{CI}.a_{ICI} + I\pi_K.a_{IK} + \\ &\quad (I\pi_1 + I\pi_2 + \cdots + I\pi_M)a_{I1} + u \end{aligned}$$

On voit donc que l'estimation par intégration des contraintes dans ce cas spécifique consiste à introduire la somme de toutes les variables concernées par la restriction.

4.7 Tester les contraintes : le test de Fisher

Les résultats précédents sont valables sous les hypothèses $H1 - H4$, qui ne spécifient que les deux premiers moments de la loi des résidus conditionnellement aux variables explicatives. On peut comme dans le cas des mco vouloir apprendre plus sur les paramètres estimés et en particulier sur leur loi pour pouvoir faire des test d'hypothèses. Parmi ces tests potentiels figure naturellement l'hypothèse imposée aux paramètres :

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

Une façon naturelle de tester l'hypothèse consiste à examiner si l'estimateur des mco satisfait approximativement les contraintes. On construit donc la quantité $\widehat{\Delta} = R\widehat{b} - r$, et on examine si elle est proche de zéro. Sous l'hypothèse nulle on sait que $\widehat{\Delta} \sim N\left(0, \sigma^2 R(\underline{x}'\underline{x})^{-1}R'\right)$.

Rappel : $Z \rightsquigarrow N(0, V)$ avec V inversible, alors $Z'V^{-1}Z \sim \chi^2(\dim(Z))$

On sait donc que sous H_0 on a $\widehat{\Delta}'\left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1}\widehat{\Delta}/\sigma^2 \sim \chi^2(p)$. Toutefois, cette relation ne peut être utilisée directement puisque σ^2 est inconnue. Comme pour le test de Student, on remplace cette quantité inconnue par un estimateur : $\widehat{\sigma}^2$. Cette statistique convenablement normalisée suit comme on l'a vu une loi du χ^2.

Definition *La loi de Fisher à q_1 et q_2 degrés de liberté, notée $F(q_1, q_2)$ est définie comme le ratio de deux lois du χ^2, divisées par leurs degrés de liberté : Si $Q_1 \sim \chi^2(q_1)$ et $Q_2 \sim \chi^2(q_2)$ et $Q_1 \perp Q_2$ alors $Z = \frac{Q_1/q_1}{Q_2/q_2} \sim F(q_1, q_2)$*

Proposition *Lorsque les hypothèses H1, H2, H3&H4 ainsi que l'hypothèse H_n de normalité des résidus, on peut effectuer un test de l'hypothèse $H_0 : Rb - r = 0$ en considérant la statistique de Fisher :*

$$\widehat{F} = \frac{1}{p}\frac{\widehat{\Delta}'\left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1}\widehat{\Delta}}{\widehat{\sigma}^2} \sim F(p, N-(k+1))$$

où $\widehat{\Delta} = R\hat{b}_{mco} - r$. Sous l'hypothèse H_0 $\widehat{F}$ suit une loi de Fisher à p et $N-(k+1)$ degrés de liberté. Le test caractérisé par la région critique

$$W = \left\{ \widehat{F} \,\middle|\, \widehat{F} > q_{1-\alpha}\left(F(p, N-(k+1))\right) \right\}$$

est un test UPP dans la classe des tests invariants, où $q_{1-\alpha}\left(F(p, N-(k+1))\right)$ *est le quantile d'ordre* $1-\alpha$ *de la loi de Fisher à* p *et* $N-(K+1)$ *degrés de liberté.*

Démonstration *La preuve du résultat concernant la distribution de la statistique sous* H_0 *découle directement de* $Q_1 = \widehat{\Delta}' \left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1} \widehat{\Delta}/\sigma^2 \sim \chi^2(p)$, *de* $Q_2 = (N-(K+1))\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-(K+1)}$, *et du fait que comme* $\widehat{\Delta}' \left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1} \widehat{\Delta}/\sigma^2$ *est issu de* $\widehat{b}_{mco}$ *qui est indépendant de* $\widehat{\sigma}^2$ Q_1 *et* Q_2 *sont indépendants. On a alors par définition de la loi de Fisher*

$$\frac{\widehat{\Delta}' \left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1} \widehat{\Delta}/\sigma^2}{p} \Bigg/ \frac{(N-(K+1))\frac{\hat{\sigma}^2}{\sigma^2}}{N-K-1} = \frac{\widehat{\Delta}' \left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1} \widehat{\Delta}}{p\hat{\sigma}^2} \sim F(p, N-(k+1))$$

On voit aussi directement que le test est un test au seuil α *puisque le risque de première espèce* $P(W,\theta)$ *pour* $\theta \in \Theta_0$ *est par définition de la région critique* α. *Pour le résultat d'optimalité, il faut noter que le test est optimal dans la classe des tests invariants, c'est à dire dans la classe des tests ne changeant pas lorsque on applique une transformation bijective aux données.*

On peut obtenir une expression de la statistique du test de Fisher la rendant très simple à mettre en pratique. Cette expression ne fait plus intervenir l'écart $R\hat{b}_{mco} - r$ mais uniquement les sommes des carrés des résidus dans les estimations du modèle contraint SCR_C et non contraint SCR.

Proposition *La statistique de Fisher* $\widehat{F} = \frac{1}{p}\frac{\widehat{\Delta}'\left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1}\widehat{\Delta}}{\widehat{\sigma}^2}$ *se réécrit simplement à partir des sommes des carrés des résidus dans le modèle contraint et non contraint*

$$\widehat{F} = \frac{1}{p}\frac{\widehat{\Delta}' \left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1} \widehat{\Delta}}{\widehat{\sigma}^2} = \frac{SCR_C - SCR}{SCR} \times \frac{N-(k+1)}{p}$$

Démonstration *En effet :* $\hat{b} = (\underline{x}'\underline{x})^{-1}\,\underline{x}'\underline{y} = b + (\underline{x}'\underline{x})^{-1}\,\underline{x}'\underline{u}$ *donc sous* H_0, *on a :* $\widehat{\Delta} = R\hat{b} - r = R(\underline{x}'\underline{x})^{-1}\,\underline{x}'\underline{u}$. *La quantité* $\widehat{\Delta}' \left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1} \widehat{\Delta}$ *s'écrit donc simplement :*

$$\widehat{\Delta}' \left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1} \widehat{\Delta} = \underline{u}'\underline{x}(\underline{x}'\underline{x})^{-1}R' \left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1} R(\underline{x}'\underline{x})^{-1}\,\underline{x}'\underline{u}$$

On reconnaît dans cette expression la matrice $P_C = \underline{x}(\underline{x}'\underline{x})^{-1}R' \left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1} R(\underline{x}'\underline{x})^{-1}\,\underline{x}'$ *introduite dans le lemme décomposant le résidu dans le modèle contraint comme*

$$\widehat{\underline{u}}_c = \widehat{\underline{u}} + P_c\underline{u} = \widehat{\underline{u}} + \widetilde{\underline{u}}$$

On a donc $\widehat{\Delta}' \left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1} \widehat{\Delta} = \underline{u}' P_C \underline{u} = \widetilde{\underline{u}}'\widetilde{\underline{u}}$. *On en déduit donc*

$$\widehat{\Delta}' \left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1} \widehat{\Delta} = \underline{u}' P_C \underline{u} = \widehat{\underline{u}}_C'\widehat{\underline{u}}_C - \widehat{\underline{u}}'\widehat{\underline{u}}$$

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

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

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

4.8 Applications du test de Fisher

4.8.1 Un test en deux étapes

La mise en oeuvre du test de Fisher d'un ensemble de contraintes $H_0 : Rb - r = 0$ se fait en plusieurs étapes.

1. On estime le modèle avec et sans contraintes. Dans chacun des cas on récupère les résidus estimés ou directement la somme des carrés des résidus $SCR_C = \underline{\widehat{u}}_C'\underline{\widehat{u}}_C$ et $SCR = \underline{\widehat{u}}'\underline{\widehat{u}}$.
2. On calcule alors la statistique $\widehat{F}$ et on la compare au fractile d'ordre $(1-\alpha)$ de la loi $F(p, N-(K+1))$, noté $F(1-\alpha)$.
3. Si $\widehat{F} > F(1-\alpha)$ alors on rejette H_0 : la somme des carrés des résidus estimés sous contraintes diffère trop de celle des carrés des résidus estimés sans contrainte pour accepter que H_0 est vraie.
4. Si $\widehat{F} \leq F(1-\alpha)$, on accepte l'hypothèse H_0.

Exemple *Homogénéité du progrès technique. On considère la régression non contrainte :*

$$\begin{aligned}d\log SR &= \pi_K.b_{0K} + \pi_1.b_{01} + \cdots + \pi_M b_{0M} + \\ &\quad + I\pi_{CI}.a_{ICI} + I\pi_K.a_{IK} + I\pi_1.a_{I1} + \cdots + I\pi a_{IM} + Xc + u\end{aligned} \tag{4.4}$$

où on introduit en plus des variables de contrôle.

- *H0(L) : Homogénéité de l'effet de l'innovation sur le facteur travail.* $a_{I1} = \cdots = a_{IM}$
- *H0(L,K,CI) : Homogénéité de l'effet de l'innovation sur les facteurs.* $a_{ICI} = a_{IK} = a_{I1} = \cdots = a_{IM}$
- *H0(L=K=CI=0) : Absence d'effet de l'innovation sur les facteurs.* $a_{ICI} = a_{IK} = a_{I1} = \cdots = a_{IM} = 0$
- *H0(K=CI=0) : Absence d'effet de l'innovation sur le capital et les consommations intermédiaires.* $a_{ICI} = a_{IK} = 0$
- *H0(K=CI=0,L) : Absence d'effet de l'innovation sur le capital et les consommations intermédiaires et homogénéité sur le travail.* $a_{IK} = 0$, $a_{I1} = \cdots = a_{IM}$

	SCR	p	F	Seuil à 5%	p-value
H1	97.099	3616			
H0(L)	97.13	1	1.15	3.84	0.28
H0(L,K,CI)	97.384	3	3.53	2.6	0.01
H0(L=K=CI=0)	97.491	4	3.63	2.37	0.005
H0(K=CI=0)	97.246	2	2.73	2.99	0.065
H0(K=CI=0,L)	97.266	3	2.07	3.53	0.10

TAB. 4.2 – Test de Fisher

Pour tester chacune de ces hypothèses contre l'hypothèse nulle H_1 (pas de restrictions sur les coefficients a_{ICI}, a_{IK}, $a_{I1}, \cdots, a_{IM}$) on peut considérer la régression sous l'hypothèse alternative ainsi que les régressions intégrant les différentes contraintes. Pour mettre en oeuvre le test de l'hypothèse d'une spécification contrainte, on considère la somme des carrés des résidus sous l'hypothèse nulle la somme des carrés des résidus sous l'hypothèse alternative ainsi que le nombre de degrés de liberté et le nombre de contraintes. Le tableau 4.2 reporte les informations pertinentes pour mettre en oeuvre le test. Si on prend par exemple le cas de la dernière spécification la somme des carrés des résidus vaut 97.266 sous l'hypothèse nulle et 97.099 sous l'hypothèse alternative. Le nombre de contraintes introduites est 3 et le nombre de degrés de liberté sous l'hypothèse alternative est $N - K + 1 = 3616$. La statistique de Fisher vaut donc

$$\widehat{F} = \frac{SCR_C - SCR}{SCR} \times \frac{N - (k+1)}{p} = \frac{97.266 - 97.099}{97.099} \times \frac{3616}{3} = 2.07$$

Sous l'hypothèse nulle cette quantité est distribuée suivant une loi de Fisher à 3 et 3616 degrés de liberté dont le quantile d'ordre 95% est 3.53. Comme la valeur estimée est inférieure à cette valeur seuil, on accepte l'hypothèse. On peut aussi regarder la p-value qui est la probabilité pour qu'une loi de Fisher à à 3 et 3616 degrés de liberté excède la valeur obtenue (2.07). On trouve une probabilité de 10%que l'on compare à la valeur seuil choisie.

On voit que parmi toutes les contraintes essayées certaines sont rejetées. Statistiquement on ne peut accepter en particulier l'hypothèse que l'effet est homogène entre tous les facteurs (spécification $H0\,(L, K, CI)$). Cette spécification conduisait on l'a vu à des coefficients très faible, loin des valeurs calculées dans la spécification non contrainte. Par contre on voit que les hypothèses d'homogénéité de l'effet sur le travail $H0\,(L)$ et de nullité de l'effet sur le capital et les consommations intermédiaires $H0\,(k = CI = 0)$ sont acceptées. En outre l'hypothèse globale réunissant ces deux contraintes $H0\,(K = CI = 0, L)$: homogénéité de l'effet sur le travail et nullité de l'effet sur le capital et les consommations intermédiaires, est acceptée. Il est intéressant de remarquer que le test de l'hypothèse globale $H0\,(K = CI = 0, L)$ passe un peu mieux que le test de l'hypothèse $H0\,(K = CI = 0)$ comme en témoigne les p-values (10% contre 6.5%). On aurait pu à la limite rejeter l'hypothèse $H\,(K = CI = 0)$ mais accepter l'hypothèse plus contraignante $H0\,(K = CI = 0, L)$.

4.8.2 Test de la nullité globale des paramètres

Dans le modèle

$$y = b_0 e + \sum_{k=1}^{k=J} \underline{x}_k b_k + \sum_{k=J+1}^{k=K} x_k b_k + u$$

on veut tester l'hypothèse de l'égalité à une valeur donnée de plusieurs coefficients. H_0 : $b_1 = b_1^0$, $b_2 = b_2^0, \ldots, b_J = b_J^0$. La différence avec le test de Student standard est qu'on souhaite faire un test global, sur l'identité simultanée des coefficients. Avec le test de Fisher il suffit d'estimer le modèle non contraint

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

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

$$\underline{y} - \sum_{k=1}^{k=J} \underline{x}_k b_k^0 = b_0 e + \sum_{k=J+1}^{k=K} \underline{x}_k b_k + \underline{u}$$

de calculer la somme $SCRC$ des carrés des résidus estimés et de former la statistique

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

Pour un test au niveau α on refusera l'hypothèse nulle si $\widehat{F}$ est supérieur au fractile d'ordre $(1-\alpha)$ de la loi $F(J, N-(K+1))$, noté $F(1-\alpha)$.

On déduit de l'exemple précédent un test systématiquement associé à toute régression et d'utilisation très courante : **le test de la significativité globale des coefficients d'une régression**

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

Il obéit à la même logique que précédemment, mais on montre que dans ce cas la statistique de Fisher est seulement fonction du R^2 dans l'estimation non contrainte du modèle.

Proposition *Dans le modèle*

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

la statistique de Fisher du test de nullité globale des paramètres H_0 s'exprime simplement à partir du R^2

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

Démonstration *Sous H_0, le modèle s'écrit : $\underline{y} = b_0\, e + \underline{u}$, d'où $\hat{b}_0 = \bar{\underline{y}}$ et $\widehat{\underline{u}}_c = \underline{y} - \bar{\underline{y}}\, e$. La SCRC est donc donnée par : $SCRC = \Sigma_n (y_n - \bar{y})^2$. Sous H_1 : $SCR = \widehat{\underline{u}}'\widehat{\underline{u}}$. Or $R^2 = 1 - \frac{\widehat{\underline{u}}'\widehat{\underline{u}}}{\Sigma_n (y_n - \bar{y})^2}$, soit $\widehat{\underline{u}}'\widehat{\underline{u}} = \Sigma_n (y_n - \bar{y})^2 \left(1 - R^2\right)$, on a donc $SCR = SCR_C \left(1 - R^2\right)$, par conséquent, la statistique de Fisher s'écrit*

$$\frac{N - (K+1)}{K} \frac{SCR_C - SCR}{SCR} = \frac{N - (K+1)}{K} \frac{SCR_C - SCR_C \left(1 - R^2\right)}{SCR_C \left(1 - R^2\right)}$$

d'où le résultat

4.8.3 Le Test de Chow de stabilité des paramètres

Une question naturelle est celle de l'homogénéité des paramètres sur deux sous population. On peut s'interroger sur l'existence de rupture temporelle dans les comportements. On peut se demander par exemple si le comportement de consommation estimé sur série temporelles est homogène dans le temps. On peut se demander aussi si les technologies de production, estimées sur un panel d'entreprises sont homogènes entre secteurs. Le Test de Chow formalise ce problème de test et applique les résultat du test de Fisher pour l'obtention de statistique de test.

Supposons que l'on dispose de deux échantillons $(\underline{y}_1, \underline{x}_1)$ et $(\underline{y}_2, \underline{x}_2)$ de tailles respectives N_1 et N_2, relatifs à deux groupes d'observations différents (par exemple deux périodes, deux catégories d'entreprises,...) de la variable dépendante y et des variables explicatives x.

Le modèle relatif au 1er groupe s'écrit

$$\underline{y}_1 = \underline{x}_1\, b_1 + \underline{u}_1$$

où $\underline{y}_1$ vecteur $N_1 \times 1$ des observations de la variable dépendante pour le premier groupe et $\underline{x}_1$ la matrice$N_1 \times (K+1)$ des variables explicatives $(1, \underline{x}_1, \ldots, \underline{x}_K)$ pour le premier groupe.

De même, pour le deuxième groupe :

$$\underline{y}_2 = \underline{x}_2\, b_2 + \underline{u}_2$$

On fait les hypothèses stochastique $l\left(\underline{u}_1, \underline{u}_2 \,|\underline{x}_1, \underline{x}_2\right) \sim N(0, \sigma^2 I_{N_1+N_2})$.

Ce modèle se réécrit dans le cadre du modèle linéaire standard en introduisant les matrices $\widetilde{\underline{x}}$ $(N_1 + N_2) \times (2\,(K+1))$ et $\underline{x}$ $(N_1 + N_2) \times (K+1)$

$$\widetilde{\underline{x}} = \begin{pmatrix} \underline{x}_1 & 0 \\ 0 & \underline{x}_2 \end{pmatrix} \text{ et } \underline{x} = \begin{pmatrix} \underline{x}_1 \\ \underline{x}_2 \end{pmatrix}$$

sous la forme

$$\underline{y} = \widetilde{\underline{x}} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \underline{u}$$

avec $l\left(\underline{u}\,|\widetilde{\underline{x}}\right) \sim N(0, \sigma^2 I_N)$. L'hypothèse d'homogénéité s'écrit alors simplement dans ce cadre :

$$H_0 : \; b_1 = b_2$$

et on peut clairement aborder cette question avec le formalisme du test de Fisher. On effectue la régression dans le modèle contraint

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

pour lequel on récupère la somme des carrés des résidus $SCR_C = SCRT$. On effectue la régression dans le modèle non contraint et on récupère aussi la somme des carrés des résidus SCR. La statistique de Fisher du test d'homogénéité des coefficients est donc

$$\widehat{F} = \frac{SCR_C - SCR}{SCR} \times \frac{(N_1 + N_2) - 2(K+1)}{(K+1)}$$

et on rejettera l'hypothèse nulle lorsque cette statistique est trop élevé : pour un test au niveau α la région critique est ainsi

$$\widehat{F} > f_{(1-\alpha)}\left(K+1,\; N_1 + N_2 - 2(K+1)\right)$$

La statistique se simplifie en fait car on montre facilement que la somme SCR est la somme $SCR1 + SCR2$ des sommes des carrés des résidus sur les modèles estimés librement sur chacun des sous-échantillons. Pour s'en convaincre il suffit de calculer $M_{\widetilde{\underline{x}}} = I - \widetilde{\underline{x}}\left(\widetilde{\underline{x}}'\widetilde{\underline{x}}\right)^{-1}\widetilde{\underline{x}}'$ puisque $SCR = \underline{u}' M_{\widetilde{\underline{x}}} \underline{u}$. On vérifie aisément que $M_{\widetilde{\underline{x}}} = Diag\left(M_{\underline{x}_1, \underline{x}_2}\right)$. La statistique est donc finalement

$$\widehat{F} = \frac{SCRT - (SCR1 + SCR2)}{SCR1 + SCR2} \times \frac{(N_1 + N_2) - 2(K+1)}{(K+1)}$$

et se calcule très simplement à partir des trois régressions : 1) contrainte 2) et 3) sur chacun des sous échantillons pris séparément.

4.9 Résumé

1. Dans ce chapitre on a vu comment étendre l'estimateur des mco au cas dans lequel on impose des contraintes linéaires sur les paramètres du type $Rb = r$.
2. On a vu que lorsque l'on fait les hypothèses $H1 - H2$, l'estimateur est sans biais lorsque les contraintes sont satisfaites par la vraie valeur du paramètre. En revanche, l'estimateur est biaisé lorsque les contraintes sont imposées à tort.
3. On a obtenu sous les hypothèses $H1 - H4$ l'expression de la matrice de variance de l'estimateur. On a vu que cette matrice était toujours plus petite que celle de l'estimateur des mco, que les contraintes soient imposées à tort ou à raison.

4. On en a conclu qu'il y a un arbitrage entre précision des estimations et robustesse.
5. On a également obtenu un estimateur sans biais de la variance des résidus.
6. On a montré comment les résultats sur la loi de l'estimateur pouvaient être étendus dans le cas d'estimations contraintes lorsque la loi des perturbations est spécifiée.
7. On a montré comment dans ce cadre il était possible de tester les contraintes imposées au paramètre.
8. Le test correspondant porte le nom de Test de Fisher, il est basé sur la comparaison des résidus dans le modèle contraint et le modèle non contraint.
9. On a vu deux exemples importants de mise en oeuvre de ce test
   (a) Le test de significativité globale des paramètres
   (b) Le test dit de Chow de stabilité des paramètres sur deux sous-échantillons.

Chapitre 5

Propriétés asymptotiques de l'estimateur des MCO

Dans ce chapitre on montre comment il est possible d'obtenir la loi des estimateurs sans faire d'hypothèses sur la loi des perturbations. On va voir que l'hypothèse de normalité de la distribution conditionnelle peut être remplacée par des hypothèses sur l'existence de moments des variables du modèle lorsque le nombre d'observations devient grand. L'obtention de ces résultats repose sur différentes notions de convergence et certains résultats essentiels comme la Loi des Grands Nombre et le Théorème Central Limite.

5.1 Rappel sur les convergences

Soit (X_n) une suite de variables aléatoires. Soit F_n la fonction de répartition de X_n. Soit X une variable aléatoire 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 les valeurs des $X_n(\omega)$ pour tous les n et de $X(\omega)$.

5.1.1 Définition : Convergence en probabilité, Convergence en loi, Convergence en moyenne quadratique

Definition *On dit que* (X_n) converge en probabilité *vers* X *(*$X_n \xrightarrow{P} X$ *ou* $\operatorname{plim}_{n\to\infty} X_n = X$*) si*

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

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

Cette notion de convergence nous intéressera pour la convergence ponctuelle des estimateurs. Dans ce cas l'élément ω est un état de la nature qui engendre un nombre infini de réalisation du processus étudié. Les suites $X_n(\omega)$ sont les suites d'estimateurs que l'on

peut construire en utilisant l'échantillons des n premières observations du processus. La limite X est une constante. La notion de convergence signifie que pour n'importe quelle boule centrée sur la limite, les états de la nature tels qu'il existe des estimateurs hors de la boule considérée pour des tailles arbitrairement grandes des échantillons sont de mesure nulle.

Definition *On dit que* (X_n) converge en moyenne quadratique *vers* X *(*$X_n \overset{mq}{\to} X$ *) si*

$$E\left\|X_n - X\right\|^2 \underset{n\to\infty}{\longrightarrow} 0.$$

Proposition *La convergence en moyenne quadratique implique la convergence en probabilité et la convergence en moyenne quadratique vers une constante résulte de la convergence du moment d'ordre 1 vers cette constante et du moment d'ordre 2 vers 0 :* $E(X_n) \to a$, *et* $V(X_n) \to 0$

Démonstration *La première partie résulte de l'inégalité de Bienaymé-Tchebitchev*

$$\Pr\left\{\left\|X_n - X\right\| > \varepsilon\right\} < \frac{E\left\|X_n - X\right\|^2}{\varepsilon^2}$$

qui exprime simplement

$$\begin{aligned} E\left\|X_n - X\right\|^2 &= E\left(\left\|X_n - X\right\|^2 \left|\left\|X_n - X\right\| > \varepsilon\right.\right) \Pr\left\{\left\|X_n - X\right\| > \varepsilon\right\} \\ &\quad + E\left(\left\|X_n - X\right\|^2 \left|\left\|X_n - X\right\| \leq \varepsilon\right.\right) \Pr\left\{\left\|X_n - X\right\| \leq \varepsilon\right\} \\ &\geq \varepsilon^2 \Pr\left\{\left\|X_n - X\right\| > \varepsilon\right\} \end{aligned}$$

la deuxième partie résulte de

$$\begin{aligned} E\left\|X_n - a\right\|^2 &= E\left(\left(X_n - EX_n\right)'\left(X_n - EX_n\right)\right) + \left(EX_n - a\right)'\left(EX_n - a\right) \\ &= \left\|EX_n - a\right\|^2 + TraceV\left(X_n\right) \end{aligned}$$

Definition *On dit que* (X_n) converge en loi *vers* X *(*$X_n \overset{L}{\to} X$*) si la suite des fonctions de répartition associées* (F_n) *converge, point par point, vers* F *la fonction de répartition de* X *en tout point où* F *est continue :*

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

5.1.2 Loi des Grands Nombres et Théorème Centrale Limite

On donne maintenant les deux théorèmes centraux sur lesquels reposent toutes les propriétés asymptotiques des estimateurs usuels : la loi des grand nombre qui stipule que sous des hypothèses assez faible la moyenne empirique converge en probabilité vers l'espérance, et le théorème central limite qui précise la loi de l'écart entre la moyenne empirique et l'espérance.

Proposition ***Loi des grands nombres (Chebichev) :*** *Soit (x_i) une suite de va indépendantes telles que $EX_i = m_i$ et $VX_i = \sigma_i^2$ existent. On considère $\overline{X}_N = \frac{1}{N}\sum_{i=1}^N X_i$ la moyenne empirique si la variance de cette moyenne empirique tend vers 0, $\Sigma_N = \frac{1}{N^2}\sum_{i=1}^N \sigma_i^2 \to 0$, alors*

$$\overline{X}_N - \overline{m}_N = \frac{1}{N}\sum_{i=1}^N X_i - \frac{1}{N}\sum_{i=1}^N m_i \xrightarrow{P} 0 \quad qd\ N \to \infty.$$

Démonstration $\frac{1}{N}\sum_{i=1}^N X_i - \frac{1}{N}\sum_{i=1}^N m_i = \frac{1}{N}\sum_{i=1}^N (X_i - m_i)$. *Pour montrer la convergence en probabilité vers zéro, il suffit de montrer la convergence en moyenne quadratique vers 0, qui résulte de la convergence vers 0 de la variance. Ce qui est acquis par hypothèse.*

Corollaire *1. Soit (X_i) une suite de va indépendantes telles que $EX_i = m$ et $VX_i = \Sigma$ existent, alors*

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

Démonstration *La variance de la moyenne empirique est dans ce cas Σ/N. Elle tend bien vers zero.*

On peut étendre la loi faible des grands nombres au cas où les variables X_n sont dans L_1, mais au prix d'une démonstration beaucoup plus compliquée.

Proposition *Soit (X_i) une suite de va indépendantes et équidistribuées telles que $EX_i = m$ et $E|X_i|$ existent, alors*

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

Proposition *Théorème central limite (Lindeberg-Levy) : Soit (X_i) une suite de variables aléatoires indépendantes et équidistribuées telles que $EX_i = m$ et $VX_i = \Sigma$ existent,*

$$\sqrt{N}\left(\overline{X}_N - m\right) \xrightarrow{L} \mathcal{N}(0, \Sigma).$$

Remarque *On sait déjà que $E\left(\sqrt{N}\left(\overline{X}_N - m\right)\right) = 0$, et également que $V\left(\sqrt{N}\left(\overline{X}_N - m\right)\right) = NV\left(\overline{X}_N\right) = V(X_n)$. Le résultat important vient du fait que l'on connaît la loi de la moyenne empirique dilatée $\sqrt{N}\left(\overline{X}_N - m\right)$.*

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

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

Les fonctions caractéristiques ont une 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 répartition 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 $\phi_n(t) = E\left(\exp it'\sqrt{N}\left(\frac{\sum_{i=1}^N X_i}{N} - m\right)\right)$

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

d'où l'approximation

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

Ce théorème est suffisant dans la majeure partie des cas. Néanmoins il fait l'hypothèse que les variables sont équidistribuées et qu'elles ont en particulier des moments d'ordre 1 et 2 identiques. Ce théorème peut être reformulé sous une autre forme. En effet $E\left(\overline{X_n}\right) = m$ et $V\left(\overline{X_n}\right) = V/N$. Le théorème ne stipule donc rien d'autre que $V\left(\overline{X_n}\right)^{-1/2}\left(\overline{X_n} - E\left(\overline{X_n}\right)\right) \xrightarrow{L} \mathcal{N}(0,1)$. Là aussi on peut étendre le théorème centrale limite pour traité des cas plus généraux. En particulier on peut obtenir un théorème de convergence pour des données indépendantes mais non équidistribuées. C'est au prix d'une condition supplémentaire appelée condition de Liapounov et qui concerne les moments d'ordre 3 de la variable.

Proposition *Théorème central limite (Liapounov) : Soit* (X_n) *une suite de variables aléatoires indépendantes de moyenne* μ_n, *de variance* σ_n^2 *et telle que* $w_{3N} = E\left(|X_n - \mu_n|^3\right)$ *existent. Si* $\lim\left(\sum_1^N w_{3n}\right)^{1/3} \Big/ \left(\sum_1^N \sigma_n^2\right)^{1/2} = 0$ *alors*

$$V\left(\overline{X_n}\right)^{-1/2}\left(\overline{X_n} - E\left(\overline{X_n}\right)\right) \xrightarrow{L} \mathcal{N}(0,1)$$

Remarque $V\left(\overline{X_n}\right) = \frac{1}{N}\overline{\sigma_n^2}$, *c'est à dire la variance moyenne divisée par* N.

5.1.3 Différents résultats concernant les convergences

On donne maintenant différents résultats, utiles lorsque l'on souhaite dériver les propriétés asymptotiques des estimateurs.

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

Le théorème de Slutsky et une application

Si $X_n \xrightarrow{L} X$ *et* $Y_n \xrightarrow{P} a$ alors on a

1. $X_n Y_n \xrightarrow{L} Xa$
2. $X_n + Y_n \xrightarrow{L} X + a$
3. $X_n / Y_n \xrightarrow{L} X/a$ si $a \neq 0$

Application : On considère deux variables aléatoires z_{1i} et z_{2i} telles que $E(z_{1i}) = m_1$ et $E(z_{2i}) = 0$. Alors pour un échantillon iid, par application de la loi des grands nombres, $\overline{z_{1i}} \xrightarrow{P} m_1$ et par application du théorème central limite $\sqrt{N}\ \overline{z_{2i}} \xrightarrow{L} N(0, V_2)$. Par application du théorème de Slutsky on a

$$\sqrt{N}\overline{z_{1i}} \times \overline{z_{2i}} \xrightarrow{L} N(0, m_1 V_2 m_1')$$

Les ordres en probabilité.

Soit X_n une suite de variable aléatoire et a_n une suite de réel.

- On dit que X_n est un "petit o de a_n" et on le note $o(a_n)$ si $a_n^{-1} X_n \xrightarrow{P} 0$. Ainsi par exemple, X_n est un $o(1)$ si $X_n \xrightarrow{P} 0$, X_n est un $o(1/n)$ si $nX_n \xrightarrow{P} 0$.
- On dit que X_n est un "grand O de a_n" et on le note $O(a_n)$ si $a_n^{-1} X_n$ est borné en probabilité. Ceci signifie que pour n'importe quel niveau de probabilité α il existe une valeur finie M_α telle que les réalisations de ω satisfaisant $\|a_n^{-1} X_n\| < M_\alpha$ pour tout n sont de mesure supérieure à α : $\forall n$, $P(\|a_n^{-1} X_n\| < M_\alpha) > \alpha$. Ce qui signifie que pour n'importe quel niveau de probabilité α aussi élevé soit il, on peut trouver une quantité bornant $a_n^{-1} X_n$ avec probabilité α uniformément en n. On peut aussi définir cette notion à partir des fonction de répartition F_n de $\|X_n\|$: $F_n(t) = P(\|X_n\| < t)$. Dire que X_n est un grand $O(a_n)$ consiste à dire que pour tout niveau de probabilité α, $\exists M_\alpha$ tel que $\forall n$ $F_n(a_n M_\alpha) > \alpha$, soit $a_n^{-1} F_n^{-1}(\alpha) < M_\alpha$. Donc $X_n = O(a_n)$ si $\underset{n}{Sup} a_n^{-1} F_n^{-1}(\alpha) < \infty$, où encore, si $\forall \alpha$, $\underset{n}{Sup} Q_n(\alpha)/a_n < \infty$ où Q_n est la fonction de quantile.

Proposition *Si* $X_n \xrightarrow{L} X$ *alors* $X_n = O(1)$

Démonstration *On considère $F(t)$ la fonction de répartition de $|X|$ et $F_n(t)$ celle de $|X_n|$. $F_n(t)$ converge en tout point de continuité de F vers F. Pour α donné, on peut définir $M_1(\alpha)$ tel que $F(M_1(\alpha)) = 2\alpha$. Il existe donc un $n(\alpha)$ tel que pour $n > n(\alpha)$ $F_n(M_1(\alpha)) > \alpha$. Pour $n < n(\alpha)$, on peut définir $M_2(\alpha) = \underset{n<n(\alpha)}{supp} F_n^{-1}(\alpha)$. On peut prendre pour $M(\alpha)$ le maximum de $M_1(\alpha)$ et de $M_2(\alpha)$.*

Proposition *Si $Y_n = O(1)$ et $X_n = o(1)$, alors $Y_n X_n = o(1)$*

Démonstration

$$\begin{aligned} P(|X_n Y_n| > \varepsilon) &= P(|X_n Y_n| > \varepsilon \,||Y_n| > M) \times P(|Y_n| > M) + P(|X_n Y_n| > \varepsilon \,||Y_n| \leq M) \times P(|Y_n| \leq \\ &< P(|Y_n| > M) + P(|X_n| > \varepsilon/M) = 1 - P(|Y_n| < M) + P(|X_n| > \varepsilon/M) \end{aligned}$$

Comme Y_n est bornée en probabilité, on peut trouver M tel que $P(|Y_n| < M) > \alpha$ pour tout n et donc $1 - P(|Y_n| < M) < \varepsilon$. Comme X_n est un $o(1)$, $P(|X_n| > \varepsilon/M) \to 0$

Proposition *Si X_n est un $O(a_n)$ alors X_n est un $o(a_n b_n)$ pour n'importe quelle suite b_n tendant vers $+\infty$.*

Démonstration *En effet $\forall \delta \; \exists M_\delta$ tq $P(\|a_n^{-1} X_n\| > M_\delta) < \delta$ i.e. $P(\|a_n^{-1} b_n^{-1} X_n\| > b_n^{-1} M_\delta) < \delta$, et $b_n^{-1} M_\delta \to 0$. Pour ε donné il existe $n(\varepsilon)$ tel que pour $n > n(\varepsilon)$ $b_n^{-1} M_\delta < \varepsilon$ et donc $P(\|a_n^{-1} b_n^{-1} X_n\| > \varepsilon) < P(\|a_n^{-1} b_n^{-1} X_n\| > b_n^{-1} M_\delta) < \delta$*

Le théorème de Slutsky a une implication importante :

Definition *Deux suites de variables aléatoires X_{1n} et X_{2n} sont dites asymptotiquement équivalentes si $X_{1n} - X_{2n} \xrightarrow{P} 0$, i.e. $X_{1n} - X_{2n} = o(1)$.*

Corollaire *du théorème de Slutsky : si X_{1n} et X_{2n} sont asymptotiquement équivalentes et $X_{1n} \xrightarrow{L} X$, alors $X_{2n} \xrightarrow{L} X$*

Démonstration *Ceci résulte directement du fait que suivant le Théorème de Slutsky si $X_{1n} - X_{2n} \xrightarrow{P} 0$ et $X_{1n} \xrightarrow{L} X$ alors $X_{2n} = X_{1n} - (X_{1n} - X_{2n}) \xrightarrow{L} X$*

On présente enfin un dernier résultat très utile, qui permet d'obtenir la loi d'une combinaison dérivable quelconque de paramètres convergeant en loi.

Proposition *Méthode delta : Pour toute fonction g continue, différentiable, si $\sqrt{n}(X_n - m) \xrightarrow{L} N(0, \Sigma)$, 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 $\sqrt{N}(X_N - m) \xrightarrow{L} N(0, \Sigma)$, $\sqrt{N}(X_N - m) = O(1)$ et donc $(X_N - m) = O\left(1/\sqrt{N}\right) = o(1)$. On applique le théorème de la valeur*

moyenne : $\exists\ \theta_n \in [0,1]$ *tq*

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

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

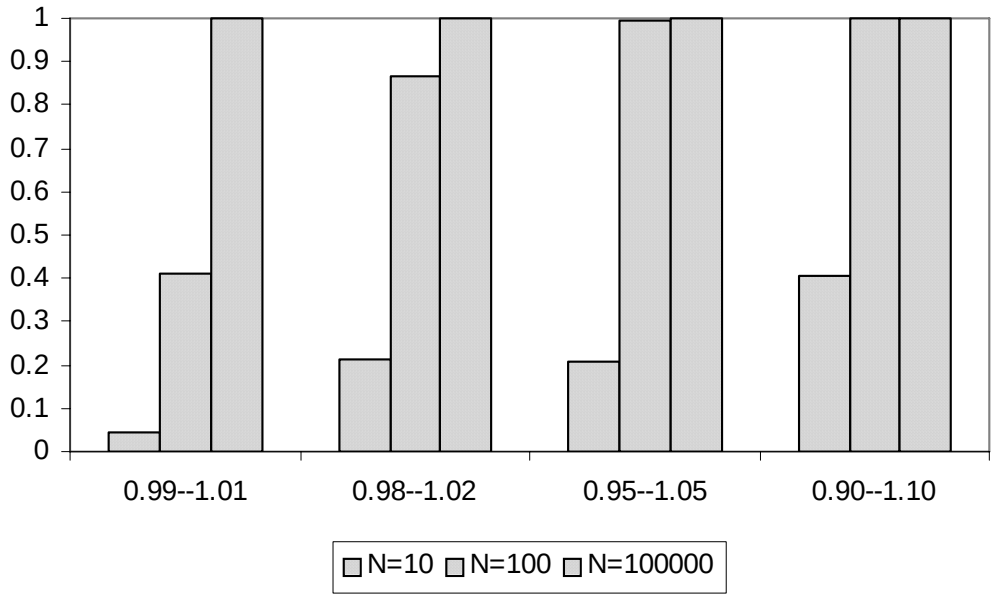
Comme $\sqrt{n}\left(X_n - m\right) \xrightarrow{L} \mathcal{N}(0, \Sigma)$, *et* $Z_n \xrightarrow{P} \frac{\partial g}{\partial m'}\left(m\right)$, *on applique le théorème de Slutsky et on en déduit*

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

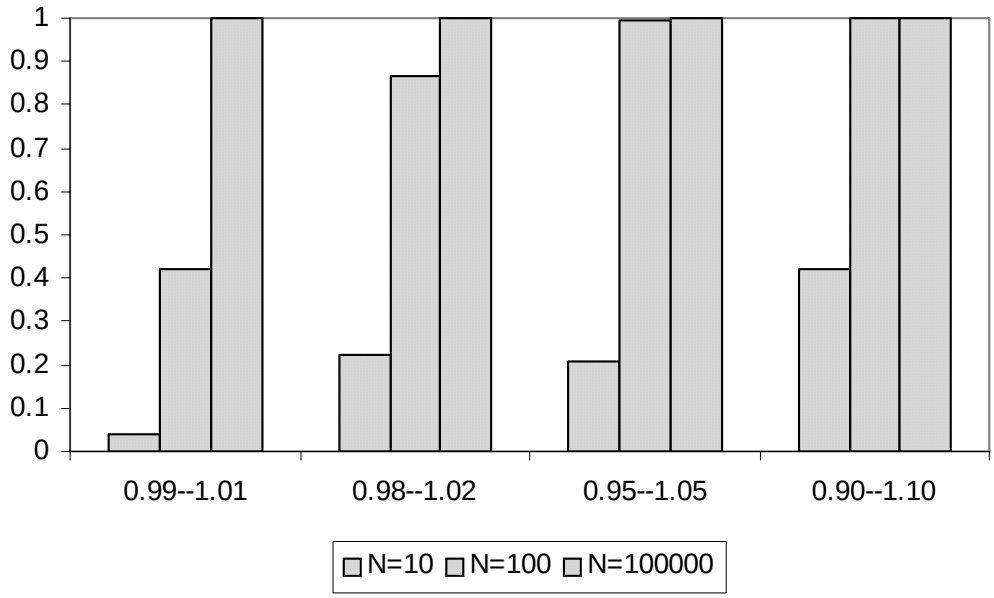
5.1.4 Illustration

On illustre ces propriétés en examinant le comportement asymptotique de moyennes d'un nombre donné d'observations tirées indépendamment dans une même loi. Plus précisément pour une taille d'échantillon donnée 10, 1.000, 100.000 on tire un grand nombre d'échantillons, en pratique 5.000, et pour chaque échantillon on calcule la moyenne empirique. On connaît l'espérance théorique E. La loi des grands nombres dit que pour un intervalle $[E - \delta\ ,\ E + \delta]$ de longueur donnée, la proportion de moyenne empirique tombant dans l'intervalle croit avec la taille de l'échantillon vers 1. Les cas que l'on considère sont $E = 1$, et on examine des intervalles pour $\delta = 0.1$, 0.05, 0.02 et 0.01. On considère deux lois différentes. On prend d'abord une loi symétrique : la loi uniforme sur $[0; 2]$. Son espérance est 1 et sa variance est de $1/3$. On prend ensuite une loi dissymétrique : une loi du $\chi^2(1)$. Cette loi a, elle aussi, une moyenne de 1 mais sa variance est de 2. Pour la rendre plus comparable à la loi précédente, on la normalise de telle sorte que sa variance soit elle aussi de $1/3$, sa moyenne restant de 1. On considère donc $y = 1 + (\chi^2(1) - 1)\frac{1}{\sqrt{6}}$. Le graphique 1 donne les proportions de moyenne empirique tombant dans les intervalles donnés. On voit que ces proportions croissent avec la largeur de l'intervalle et avec la taille de l'échantillon. Pour les plus grandes tailles d'échantillon, toutes les moyennes empiriques tombent dans l'intervalle considéré, aussi étroit soit-il. On voit aussi qu'il n'y a pas grande différence entre la loi du χ^2 et la loi uniforme. On examine ensuite la distribution des écarts à l'espérance théorique, dilatée par $\sqrt{N}$. Plus spécifiquement, on examine la distribution empirique de $\sqrt{N}\left(\overline{y}_i - E\right)/\sigma$. Pour cela on met en oeuvre un estimateur non paramétrique de la densité, dit à noyau. Si la théorie asymptotique est satisfaite, cette distribution doit être approximativement normale pour un grand échantillon. Les résultats sont présentés dans le graphique 2. On voit là des différences importantes entre les deux types de loi. Dans les deux cas pour de grands échantillons, l'approximation normale fonctionne bien. Par contre pour les petits échantillons, l'approximation normale marche très bien pour la loi uniforme, mais beaucoup moins bien, pour la loi du χ^2.

Loi Uniforme



Loi du Chi2



TAB. 5.1 – Convergence en probabilité

5.2 Propriétés asymptotiques de l'estimateur des MCO

On applique maintenant les résultats précédents à l'estimateurs des mco. On va voir que l'écart entre la vraie valeur du paramètre et le paramètre estimé s'écrit sous la forme $\widehat{b}-b=\left(\overline{x_i'x_i}\right)^{-1}\overline{x_i'u_i}$. On va étudier le comportement asymptotique de chacune des deux composantes. D'une façon générale, on va écrire $\overline{x_i'x_i}\overset{P}{\to}Q$ constante. On va donner des conditions sous lesquelles cette matrice est $E\left(x_i'x_i\right)$, comme on s'y attend, mais ce n'est pas le point central. Le point central est que cette matrice converge en probabilité vers une matrice fixe. Pour étudier le deuxième terme on va appliquer le théorème central limite à $x_i'u_i$, c'est à dire que l'on va étudier $\sqrt{N}\overline{x_i'u_i}$ et on va exploiter le fait que $E\left(x_i'u_i.\right)=0$.

Plus précisément, on considère le modèle

$$y_i=x_ib+u_i$$

avec les hypothèses

H1 : Les observations $(y_i,x_i)\in\mathbb{R}\times\mathbb{R}^{K+1}$, $i=1,...,N$, sont IID

H2 : $\forall$ N, $\underline{x}'\underline{x}$ est non singulière

H3 : Les moments de $|x_{ki}x_{li}|$ existent.et $E(x_ix_i')$ est inversible

H3bis $\underline{x}'\underline{x}/N\overset{P}{\to}Q$ inversible

H4 : $E\left(u_i\left|x_i\right.\right)=0$

H5 : $V\left(u_i\left|x_i\right.\right)=V\left(u_i\right)=\sigma^2$

Proposition *Sous les hypothèses H1 à H5, l'estimateur des MCO*

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

1. $\widehat{b}_{mco}\overset{P}{\to}b$,
2. $\sqrt{N}\left(\widehat{b}_{mco}-b\right)\overset{L}{\to}\mathcal{N}\left(0,V_{as}\right)$,
3. $V_{as}=\sigma^2E(x_ix_i')^{-1}$ *(ou* σ^2Q^{-1}*)*
4. $\widehat{\sigma}^2=\frac{1}{N-K-1}\left(\underline{y}-\underline{x}\widehat{b}_{mco}\right)'\left(\underline{y}-\underline{x}\widehat{b}_{mco}\right)\overset{P}{\to}\sigma^2$
5. $N\widehat{V}\left(\widehat{b}_{mco}\right)=\widehat{V}_{as}=\widehat{\sigma}^2\left(\overline{x_ix_i'}\right)^{-1}\overset{P}{\to}V_{as}$
6. $\sqrt{N}\widehat{V}_{as}^{-1/2}\left(\widehat{b}_{mco}-b\right)=\widehat{V}^{-1/2}\left(\widehat{b}_{mco}-b\right)\overset{L}{\to}\mathcal{N}\left(0,I_{K+1}\right)$

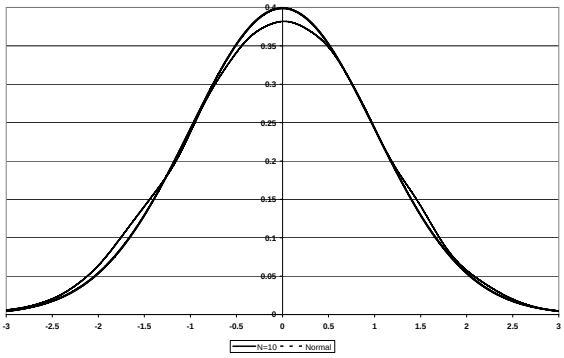
On dit que $\widehat{b}$ *est* ***convergent et asymptotiquement normal****.*

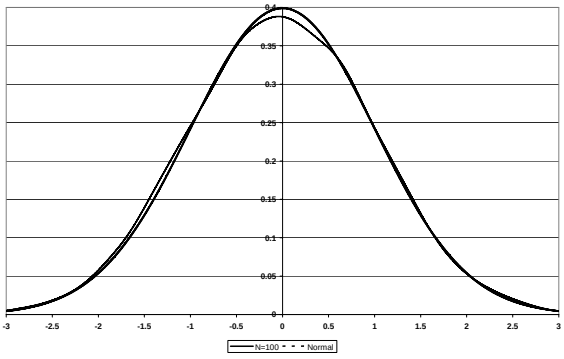
Démonstration ***Convergence en probabilité de l'estimateur.***

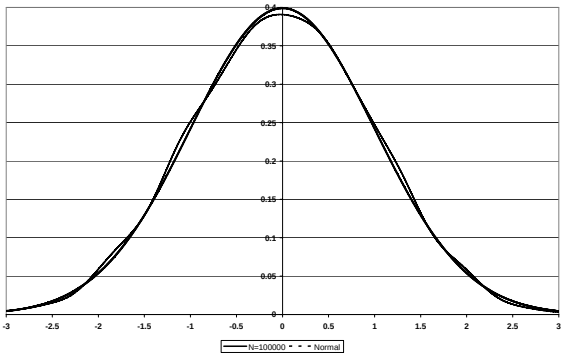
L'estimateur des mco s'écrit

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

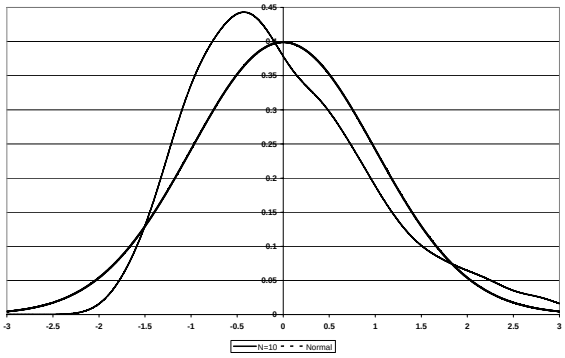
Loi Uniforme

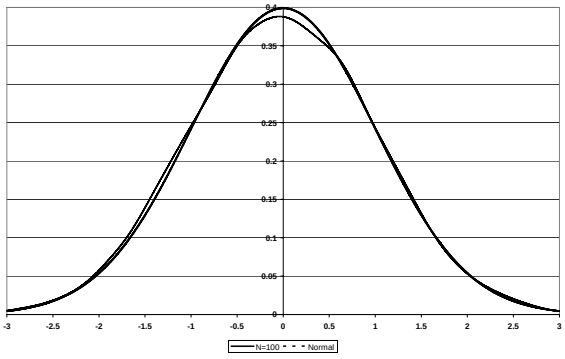


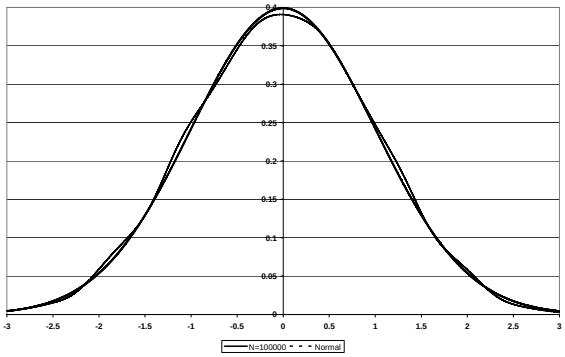




Loi du Chi2







TAB. 5.2 – Convergence en Loi

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

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

Comme les moments $|x_{ki} x_{li}|$ des variables explicatives existent, on peut appliquer la loi des grands nombres à $\overline{x_i' x_i}$. De même on peut appliquer la loi des grands nombre à $\overline{x_i' u_i}$, si $E(x_i' u_i)$ et $V(x_i' u_i)$ existent. Comme $E(x_i' u_i) = E\left(E(x_i' u_i \,|x_i)\right) = 0$ et $V(x_i' u_i) = E\left(V(x_i' u_i \,|x_i)\right) + V\left(E(x_i' u_i \,|x_i)\right) = \sigma^2 E\left(x_i' x_i\right)$, on a

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

On en déduit que

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

car les espérances $E(x_i' x_i)$ et $E(x_i' u_i)$ sont par définition des constantes, que l'application $A \to A^{-1}$ est continue et enfin que le produit et la somme de suite de variables aléatoires convergent en probabilité vers des constantes converge en probabilité.

Comme par ailleurs

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

On a bien

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

Normalité asymptotique

De la formulation $\widehat{b}_{mco}$: $\widehat{b}_{mco} = b + \overline{x_i' x_i}^{-1} \overline{x_i' u_i}$ 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 veut appliquer le Théorème Central Limite à $\sqrt{N} \overline{x_i' u_i}$. Les variables aléatoires $x_i' u_i$ sont indépendantes et équidistribuées. On pourra appliquer le Théorème Central limite si les deux premiers moments de cette variable existent. On sait que

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

Les moments d'ordre 1 et 2 de $x_i' u_i$ existent donc. On sait qu'alors le TCL permet d'affirmer

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

Comme

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

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

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

on a donc bien

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

Estimation de la variance

L'estimateur de la variance des résidus

$$\widehat{\sigma}^2 = \frac{1}{N}\left(\underline{y}-\underline{x}\widehat{b}_{mco}\right)'\left(\underline{y}-\underline{x}\widehat{b}_{mco}\right)$$

s'écrit compte tenu de $\underline{y} = \underline{x}b + \underline{u}$

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{1}{N}\left(\underline{x}\left(b-\widehat{b}_{mco}\right)+\underline{u}\right)'\left(\underline{x}\left(b-\widehat{b}_{mco}\right)+\underline{u}\right) \\ &= \overline{\left(x_i\left(b-\widehat{b}_{mco}\right)+u_i\right)\left(x_i\left(b-\widehat{b}_{mco}\right)+u_i\right)} \\ &= \overline{\left(b-\widehat{b}_{mco}\right)'x_i'x_i\left(b-\widehat{b}_{mco}\right)+2u_ix_i\left(b-\widehat{b}_{mco}\right)+u_i^2} \\ &= \left[\left(b-\widehat{b}_{mco}\right)'\overline{x_i'x_i}\left(b-\widehat{b}_{mco}\right)+2\overline{u_ix_i}\left(b-\widehat{b}_{mco}\right)+\overline{u_i^2}\right] \xrightarrow{P} \sigma^2 \end{aligned}$$

puisque $\widehat{b}_{mco} \xrightarrow{P} b$, $\overline{x_i'x_i} \xrightarrow{P} E\left(x_i'x_i\right)$, $\overline{x_i'u_i} \xrightarrow{P} E\left(x_i'u_i\right)$ *et* $\overline{u_i^2} \xrightarrow{P} E\left(u_i^2\right) = \sigma^2$. *Puisque* u_i^2 *est une variable positive identiquement distribuée sur les individus. On remarque qu'il est ici nécessaire de d'avoir recours à la loi forte des grands nombres dans L1, on devrait sinon faire l'hypothèse que* $E\left(u_i^4\right)$ *existe.*

Estimation de la matrice de variance asymptotique de l'estimateur

On l'obtient directement par le fait que $\widehat{\sigma}^2 \xrightarrow{P} \sigma^2$ *et* $\overline{x_i'x_i}^{-1} \xrightarrow{P} E\left(x_i'x_i\right)^{-1}$

Enfin en appliquant le théorème de Slutsky à $\widehat{V}_{as} = \widehat{\sigma}^2\left(\overline{x_i'x_i}\right)^{-1} \xrightarrow{P} V_{as}$,.*et* $\sqrt{N}\left(\widehat{b}_{mco}-b\right) \xrightarrow{L} \mathcal{N}\left(0,V_{as}\right)$, *on a directement le dernier point.*

Remarque *On peut se passer de l'hypothèse d'équidistribution au prix d'un renforcement des hypothèses sur les moments des variables. pour pouvoir appliquer le Théorème Central Limite de Liapounov à* $\overline{x_i'u_i}$, *il suffit par exemple que l'on ait* $E\left(|u_i^3|\right) = \gamma_3 < +\infty$ *et pour chaque variable explicative* $E\left(|x_{ki}^3|\right) = \gamma_{k3} < +\infty$. *La condition de Liapounov est alors satisfaites et on obtient alors la normalité asymptotique de* $\sqrt{N}\overline{x_i'u_i}$.

Remarque *$\widehat{V}_{as}$ est un estimateur de V_{as}, la variance asymptotique de l'estimateur dilaté par $\sqrt{N}$ qui est une matrice constante. En revanche $\widehat{V}\left(\widehat{b}_{mco}\right)$ est un estimateur de la variance de l'estimateur. C'est une quantité qui tend vers 0 quand N tend vers l'infini : $N\widehat{V}\left(\widehat{b}_{mco}\right) = \widehat{V}_{as} \xrightarrow{P} V_{as}$*

5.3 Tests asymptotiques

On a vu dans les chapitres précédents que connaître la loi de l'estimateur était utile dès lors que l'on veut faire des tests. C'est à nouveau cette question qui nous intéresse. Les tests que l'on considère sont des test dits asymptotiques. La différence essentielle avec les cas précédents est qu'ils sont basés sur une statistique dont on ne connaît la loi qu'asymptotiquement, alors que dans le cadre des chapitres précédents, on connaissait exactement la loi de la statistique à distance finie : Student, Fisher,...

La différence concerne aussi la notion d'optimalité que l'on retient. Comme précédemment, les tests que l'on va considérer sont définis par une région critique W pour une statistique $\widehat{S}$ telle que

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

On introduit aussi les risques de première espèce

$p\lim \Pr\left(\widehat{S} \in W \,|H_0\right)$ est le *risque de première espèce* : il représente asymptotiquement la probabilité de rejeter H_0 à tort.

$p\lim \Pr\left(\widehat{S} \notin W \,|H_a\right)$ est le *risque de deuxième espèce* : la probabilité d'accepter H_0 à tort. On introduit aussi la puissance du test définie comme 1− *risque de deuxième espèce* : *puissance*=$p\lim \Pr\left(\widehat{S} \in W \,|H_a\right)$. Le principe du test est comme précédemment de minimiser le risque de seconde espèce en contrôlant à un niveau donné le risque de première espèce. Ce niveau du maximal du risque de première espèce est appelé la encore le *seuil ou le niveau du test.* Dans le cas normal on avait introduit la notion de tests uniformément plus puissants, c'est à dire de tests qui maintenant un niveau donné du risque de première espèce conduise pour toute valeur de l'hypothèse alternative à une probabilité de rejet maximale. Cette propriété est trop forte et on ne peut pas trouver en toute généralité un tel test. On avait alors introduit des classes de tests plus restreintes, les tests sans biais, les tests invariants pour lesquels on pouvait trouver un test optimal.

La notion que l'on retient ici est celle de test convergent. Elle rejoint la notion de test uniformément plus puissant puisqu'un test convergent est un test dont la puissance tend vers 1.

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

On définit aussi de façon alternative la *p-value*. La statistique $\widehat{S}$ est choisie de telle sorte que sous H_0 $\widehat{S} \to S_0$ dont la loi est connue et à support positif (valeur absolue d'une loi normale, loi du khi deux). La région critique est définie comme

$$W = \left\{ \widehat{S} \middle| \widehat{S} > q\left(1-\alpha, S_0\right) \right\}$$

où $q\left(1-\alpha, S_0\right)$ est le quantile d'ordre $1-\alpha$ de S_0 : $\Pr\left(S_0 > q\left(1-\alpha, S_0\right)\right) = \alpha$
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 si et seulement si $\alpha \geq p\left(\widehat{S}\right)$. En effet, $\alpha \geq p\left(\widehat{S}\right)$ signifie que

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

5.3.1 Test d'hypothèses linéaires

Test de Student asymptotique

Il s'agit du test d'une hypothèse linéaire unidimensionnelle de la forme

$$H_0 : c'b = r$$

où $c \in \mathbb{R}^{K+1}$ et $r \in \mathbb{R}$. Un cas particulièrement important est celui de la significativité du coefficient $b_k = 0$.

Proposition *Si les hypothèses H1-H5 sont satisfaites, sous l'hypothèse nulle $H_0 : c'b = r$ on a*

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

le test défini par la région critique

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

où $q\left(1-\frac{\alpha}{2}\right)$ est le quantile $1-\frac{\alpha}{2}$ de la loi normale $\mathcal{N}(0,1)$ est un test convergent au niveau α.

On retrouve donc un test très proche de celui obtenu dans le cas où on spécifie la loi des résidus. Les seules différences sont que 1/ le résultat n'est valable qu'asymptotiquement, alors qu'il était valable à distance finie dans le cas normal et 2/ la loi considérée est

une loi normale et non plus une loi de Student. Cette dernière différence n'en est une qu'en partie puisque l'on peut montrer que la loi de Student tend vers une loi normale lorsque le nombre de degrés de liberté tend vers l'infini. Les régions critiques sont donc asymptotiquement les mêmes.

Démonstration *Sous les hypothèses H1-H5, on a* $\sqrt{N}\left(\widehat{b}_{mco}-b\right)\xrightarrow{L}\mathcal{N}\left(0,\sigma^2E(x_ix_i')^{-1}\right)$,

Sous l'hypothèse nulle$H_0 : c'b = r$ *on a donc* $\sqrt{N}\left(c'\widehat{b}_{mco}-r\right)\xrightarrow{L}\mathcal{N}\left(0,c'V_{as}\left(\widehat{b}_{mco}\right)c\right)$*ou encore*

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

On rencontre le même problème que dans le cas normal : il faut diviser $c'\widehat{b}_{mco}-r$ *par l'écart-type de* $c'\widehat{b}_{mco}-r$ *qui est inconnu. Comme dans le cas normal on va diviser par un estimateur de cet écart-type. Dans le cas normal la statistique considérée suivait une loi de Student quelque soit le nombre d'observation. Ici on tient compte du fait qu'on divise par un estimateur convergent en probabilité. Le théorème de Slutsky permet alors de définir la loi asymptotique de la statistique.*

Comme

$$\widehat{V}_{as}\left(\widehat{b}_{mco}\right)=\widehat{\sigma}^2\left(\overline{x_i'x_i}\right)^{-1}=\widehat{\sigma}^2\frac{1}{N}\left(\underline{x}'\underline{x}\right)^{-1}\xrightarrow{P}V_{as}\left(\widehat{b}_{mco}\right)=\sigma^2\left[E(x_i'x_i)\right]^{-1}$$

On en déduit que la statistique de Student :

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

puisque $N\widehat{V}\left(\widehat{b}_{mco}\right)=\widehat{V}_{as}\left(\widehat{b}_{mco}\right)$. *On définit la région critique comme*

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

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*

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

Le test défini par la région critique W *est donc un test au niveau* α.

Comme on est dans le cas asymptotique, on étudie beaucoup plus facilement le comportement de la statistique sous l'hypothèse alternative.

Sous H_1 *on a* $c'\widehat{b}_{mco} - r \to c'b - r = m \neq 0$ *donc* $\left|\widehat{S}\right| \Big/ \sqrt{N} = \left|\left(c'\widehat{b}_{mco} - r\right)\right| \Big/ \sqrt{c'\widehat{V}_{as}\left(\widehat{b}_{mco}\right)c} \to |m| \Big/ \sqrt{c'V_{as}\left(\widehat{b}_{mco}\right)c}$ *d'où* $\left|\widehat{S}\right| \to +\infty$. *Il en résulte que*

$$\Pr\left\{\widehat{S} \in W \,|H_1\right\} \to 1$$

le test est donc convergent.

Remarque *On généralise directement ces résultats au cas du test unilatéral* $H_0 : c'b - r = 0$ *contre* $H_1 : c'b - r > 0$. *On définit la région critique comme*

$$W = \left\{\widehat{S} \,\middle|\, \widehat{S} > q\,(1-\alpha)\right\}$$

où $q\,(1-\alpha)$ *est le quantile* $1-\alpha$ *de la loi normale* $\mathcal{N}(0,1)$.*Sous* H_0 *on a*

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

Sous H_1 *on a* $c'\widehat{b} - r \to c'b - r = m > 0$ *donc* $\widehat{S} \Big/ \sqrt{N} = \left(c'\widehat{b} - r\right) \Big/ \sqrt{c'\widehat{V}_{as}\left(\widehat{b}\right)c} \to m \Big/ \sqrt{c'V_{as}\left(\widehat{b}\right)c}$ *d'où* $\left|\widehat{S}\right| \to +\infty$

$$\Pr\left\{\widehat{S} \in W \,|H_1\right\} \to 1$$

Application : test de Student asymptotique de nullité d'un paramètre à 5% Le cas d'application le plus direct est celui du test de la nullité d'un paramètre d'une régression. Dans ce cas le vecteur $c' = (0, \ldots, 0, 1, 0, \ldots, 0)$, $c'b = b_k$, $r = 0$, car on s'intéresse à l'hypothèse nulle de nullité de la kième composante du paramètre et $\sqrt{c'_{as}\widehat{V}_{as}\left(\widehat{b}\right)c/N} = \sqrt{c'\widehat{V}\left(\widehat{b}\right)c} = \sqrt{\widehat{V}\left(\widehat{b}_k\right)} = \widehat{\sigma}_k$. Le résultat de la proposition stipule donc qu'un test asymptotique au seil α de l'hypothèse de nullité du paramètre peut être fait en considérant le t de Student

$$t_k = \frac{\widehat{b}_k}{\widehat{\sigma}_k}$$

Asymptotiquement sous l'hypothèse nulle, cette quantité suit une loi normale. Un Test au seuil α êut être effectué en comparant la valeur du t au quantile d'ordre $1 - \alpha/2$ de la loi normale. Ainsi on rejettera H_0 à $\alpha\%$ si $|t_k| > q\,(1-\alpha/2, N\,(0,1))$.

En pratique on s'intéresse souvent à des tests à 5%. Dans ce cas le quantile auquel on compare est le quantile d'ordre 97,5% dont la valeur est de 1,96. En d'autres termes : on rejette à 5% l'hypothèse de nullité d'un paramètre si le ratio de la valeur estimée du paramètre à son écart-type estimé, le t de Student, est en valeur absolue supérieur à 1,96.

Remarque *Ce test à l'intérêt d'être valable quelque soit la loi des résidus, qu'elle soit normale ou non, tant qu'elle vérifie les hypothèses garantissant les propriétés asymptotiques de l'estimateur des mco. Le test de Student vu dans le chapitre précédent n'est valable que pour le cas de résidus suivant une loi normale. Il est en revanche valable à distance finie. Asymptotiquement les deux test coïncident car une suite de variables aléatoires X_n suivant une loi de Student à n degrés de liberté converge en loi vers une loi normale. On peut le oir facilement. Si X_n suit une loi de Student, elle peut s'écrire sous la forme d'un ratio $Z_{1n}/\sqrt{Z_{2n}/n}$ avec Z_{1n} suivant une loi normale et Z_{2n}, indépendante de Z_{1n} suivant une loi du $\chi^2(n)$. Une loi du $\chi^2(n)$ a pour variance $2n$. On en déduit que $E(Z_{2n}/n) = 1$ et $V(Z_{2n}/n) = 2/n$. On voit donc que $\sqrt{Z_{2n}/n} \xrightarrow{m.q.} 1$. Donc $\sqrt{Z_{2n}/n} \xrightarrow{p} 1$ On en déduit donc que $Z_{1n}/\sqrt{Z_{2n}/n}$ converge en Loi vers une loi normale.*

Test de Wald d'une hypothèse multi-dimensionnelle.

Comme précédemment, on souhaite tester un système de contraintes linéaires :

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

On a vu que dans le cas où les résidus étaient spécifiés comme normaux, on pouvait faire un test de Fisher. Ce test permettait de contrôler le risque de première espèce et avait de bonnes propriétés d'optimalité. Ici on va considérer une statistique analogue et on va étudier son comportement asymptotiquement. Pour la même raison que pour le test de Student, la statistique ne suivra pas une loi de Fisher mais une loi du Chi2.

Proposition *Lorsque les hypothèses H1-H5 sont satisfaites, la statistique $\widehat{S}$ définie par*

$$\begin{aligned}\widehat{S} &= N\left(R\widehat{b}_{mco} - r\right)' \left[R\widehat{V}_{as}\left(\widehat{b}_{mco}\right)R'\right]^{-1}\left(R\widehat{b} - r\right) \\ &= \frac{\left(R\widehat{b}_{mco} - r\right)'\left[R\left(\underline{x}'\underline{x}\right)^{-1}R'\right]^{-1}\left(R\widehat{b}_{mco} - r\right)}{\widehat{\sigma}^2}\end{aligned}$$

converge en loi vers un χ^2_p, sous l'hypothèse nulle H_0. Le test défini par la région critique

$$W = \left\{\widehat{S} \middle| \widehat{S} > q\left((1-\alpha), \chi^2(p)\right)\right\}$$

est un test convergent au niveau α. La statistique peut aussi être calculée comme

$$\widehat{S} = p\widehat{F} = (N - (K+1))\frac{SCRC - SCR}{SCR} \simeq N\frac{\widehat{\sigma}_c^2 - \widehat{\sigma}^2}{\widehat{\sigma}^2}$$

Démonstration *On a :*

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

On en déduit

$$N\left(R\widehat{b}_{mco}-r\right)'\left(RV_{as}\left(\widehat{b}_{mco}\right)R'\right)^{-1}\left(R\widehat{b}_{mco}-r\right)\xrightarrow{L}\chi_p^2.$$

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

$$\begin{aligned}\widehat{S} &= N\left(R\widehat{b}_{mco}-r\right)'\left[R\widehat{V}_{as}\left(\widehat{b}_{mco}\right)R'\right]^{-1}\left(R\widehat{b}_{mco}-r\right)\\ &= \left(R\widehat{b}_{mco}-r\right)'\left[R\widehat{V}\left(\widehat{b}_{mco}\right)R'\right]^{-1}\left(R\widehat{b}_{mco}-r\right)\\ &= \frac{\left(R\widehat{b}_{mco}-r\right)'\left[R\left(\underline{x}'\underline{x}\right)^{-1}R'\right]^{-1}\left(R\widehat{b}_{mco}-r\right)}{\widehat{\sigma}^2}=p\widehat{F}\xrightarrow{L}\chi^2(p),\ \textit{sous } H_0\end{aligned}$$

Ce résultat permet de montrer que le test défini par la région critique donnée est un test au niveau α*.*

Sous H_1*on a en revanche* $R\widehat{b}-r\to Rb-r=m\neq 0$*. Donc* $\widehat{S}/N=\left(R\widehat{b}_{mco}-r\right)'\left[R\widehat{V}_{as}\left(\widehat{b}_{mco}\right)R'\right]^{-1}\left(R\widehat{b}_{mco}-r\right)\to$*constante et donc*

$$\widehat{S}\to\infty$$

donc la puissance du test converge vers 1

Un cas particulier : Test de la nullité des paramètres d'une régression sauf la constante.

On a vu que lorsque l'on spécifiait la loi des résidus comme une loi normale, on avait

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

D'où

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

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

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

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

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

5.3.2 Test d'hypothèses non linéaires

La théorie asymptotique permet de traiter des questions qui ne pouvaient pas être abordées auparavant. En effet, on peut vouloir tester des hypothèses non linéaires dans les paramètres. Le modèle dit à retards échelonnés en constitue un exemple. Dans ce modèle on a une variable dépendante y_t dépendant d'une variable x_t et de ses retards : x_{t-1}, $x_{t-2}, \ldots, x_{t-L}$:

$$y_t = \alpha + \beta_0 x_t + \cdots \beta_L x_{t-L} + u_t$$

Une restriction fréquemment imposée sur ces paramètres est qu'ils soient de la forme : $\beta_k = \beta_0 \lambda^k$. Ceci correspond à imposer $L-1$ contraintes de la forme

$$\frac{\beta_2}{\beta_1} = \frac{\beta_1}{\beta_0}, \ldots, \ \frac{\beta_L - 1}{\beta_{L-2}} = \frac{\beta_1}{\beta_0}, \frac{\beta_L}{\beta_{L-1}} = \frac{\beta_1}{\beta_0}$$

qui sont typiquement non linéaires et ne peuvent donc être testées dans le cadre précédent. On peut s'intéresser d'une façon générale à des hypothèses de la forme :

$$H_0 : g(b_0) = 0,$$

où $g(b)$ est un vecteur de p contraintes non linéaires sur les paramètres telle que $\frac{\partial g(b_0)}{\partial b'}$ est de plein rang. Cette hypothèse équivaut à $\frac{\partial g(b_0)}{\partial b'} \left(\frac{\partial g(b_0)}{\partial b'} \right)'$ inversible, avec b_0 est la vraie valeur du paramètre.

Remarque *Si $g(b) = Rb - r$; alors $\frac{\partial g(b)}{\partial b'} = R$. On retrouve donc la condition sur le rang de R*

Le résultat suivant permet de généraliser les tests précédents au cas non linéaire

Proposition *Si $\widehat{b}_N$ est un estimateur asymptotiquement normal de b :*

$$\sqrt{N} \left(\widehat{b}_N - b \right) \xrightarrow{L} \mathcal{N} \left(0, V_{as} \left(\widehat{b} \right) \right)$$

et si on dispose d'un estimateur convergent de la matrice de variance de l'estimateur,

$$\widehat{V}_{as} \left(\widehat{b} \right) \xrightarrow{P} V_{as} \left(\widehat{b} \right)$$

Alors

$$\sqrt{N} \left[\frac{\partial g(\widehat{b})}{\partial b'} \widehat{V}_{as} \left(\widehat{b} \right) \frac{\partial g(\widehat{b})}{\partial b'} \right]^{-1/2} \left(g(\widehat{b}) - g(b) \right) \xrightarrow{L} \mathcal{N} \left(0, I_p \right).$$

pour toute fonction g continue, dérivable et à dérivée continue, de dimension $p \times 1$

Démonstration *On applique la méthode delta. On sait que*

$$\sqrt{N}\left(g(\widehat{b})-g(b)\right)\xrightarrow{L}\mathcal{N}\left(0,\frac{\partial g(b)}{\partial b'}V_{as}\left(\widehat{b}\right)\frac{\partial g(b)}{\partial b'}\right)$$

C'est à dire

$$\sqrt{N}\left[\frac{\partial g(b)}{\partial b'}V_{as}\left(\widehat{b}\right)\frac{\partial g(b)}{\partial b'}\right]^{-1/2}\left(g(\widehat{b})-g(b)\right)\xrightarrow{L}\mathcal{N}\left(0,I\right)$$

Comme $\frac{\partial g(\widehat{b})}{\partial b'}\widehat{V}_{as}\left(\widehat{b}\right)\frac{\partial g(\widehat{b})}{\partial b'}\xrightarrow{P}\frac{\partial g(b)}{\partial b'}V_{as}\left(\widehat{b}\right)\frac{\partial g(b)}{\partial b'}$, *on obtient le résultat par application du théorème de Slutsky.*

Ce résultat permet d'étendre directement les tests précédents au cas d'hypothèses non linéaires :

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

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

  et on procède comme dans le cas d'une contrainte linéaire.
- Cas de plusieurs contraintes, $p<K+1$. On calcule la statistique de Wald :

$$\widehat{S}=Ng(\widehat{b})'\left[\frac{\partial g(\widehat{b})}{\partial b'}\widehat{V_{as}}\left(\widehat{b}\right)\left(\frac{\partial g(\widehat{b})}{\partial b'}\right)'\right]^{-1}g(\widehat{b})=g(\widehat{b})'\left[\frac{\partial g(\widehat{b})}{\partial b'}\widehat{V}\left(\widehat{b}\right)\left(\frac{\partial g(\widehat{b})}{\partial b'}\right)'\right]^{-1}g(\widehat{b})$$

  que l'on compare au quantile $1-\alpha$ de la loi du chi-deux à p (le nombre de contraintes) degrés de liberté. On est contraint dans ce cas à la mise en oeuvre du test de Wald. Il n'y a pas d'analogue simple du test de Fisher puisque l'estimation du modèle sous l'hypothèse nulle ne peut être faite simplement.

5.4 Exemple

Pour illustrer les propriétés asymptotiques des tests, on reprend le même cadre que celui utilisé pour étudier la puissance du test de Student. On simule donc un modèle un grand nombre de fois avec des vraies valeurs différentes sur l'intervalle $[0,2]$ et on fait le test de l'égalité du paramètre à 1. On va examiner comment les résultats sont modifié lorsque l'on met en oeuvre le test de Student asymptotique, basé sur la distribution d'une loi normale et non plus le test de Student basé sur la loi de Student. on va aussi examiné comment ces résultats sont modifié lorsque les perturbations ne suivent plus une loi normale. On prendra l'exemple d'une loi de Fisher à 1 et 5 degrés de liberté, normalisée pour que son espérance soit nulle et sa variance unitaire. On choisit cette loi car elle est

asymétrique et que les lois de Fisher n'ont un moment d'ordre 2 que si le deuxième degrés de liberté est supérieur à 4. On est donc dans un cas où les hypothèses de convergence sont juste satisfaites.

[**A FAIRE**]

5.5 Résumé

Dans ce chapitre on a :

- rappelé les différents modes de convergence utiles pour l'examen des propriétés asymptotiques des estimateurs : convergence en loi et convergence en probabilité.
- rappelé les propriétés asymptotiques importantes des moyennes empiriques de variables : la loi des grands nombres et le théorème central limite.
- montré que sous des hypothèses très faibles (existence des moments d'ordre 1 et 2), l'estimateur des mco est convergent et asymptotiquement normal.
- Etendu la notion de test pour définir des tests asymptotiques, caractérisés par le fait que leur puissance tend vers 1 et généralisé les notions de test de Student et de test de Fisher au cas asymptotique.

Chapitre 6

Le modèle linéaire sans l'hypothèse d'homoscédasticité

6.1 Présentation : Homoscédasticité et hétéroscédasticité.

Jusqu'à présent on a examiné le cas du modèle linéaire

$$y_i = x_i b + u_i$$

dans lequel les observations étaient supposées Indépendantes et Identiquement Distribuées (IID). On a obtenu des résultats de convergence de distribution d'optimalité sous différentes hypothèses. On a vu qu'il était possible d'assouplir un peu ces hypothèses et de relâcher l'hypothèse ID pour qu'elles ne portent que sur les moments d'ordre 1 et 2 de la loi des perturbations conditionnellement aux variables explicatives. Les hypothèses centrales qui étaient faites portaient $E(u_i \,|\underline{x}) = 0$ qui est une condition d'identification et sur $V(u_i \,|\underline{x}) = \sigma^2$ et $Cov(u_i, u_j \,|\underline{x}) = 0$, soit $V(\underline{u} \,|\underline{x}) = \sigma^2 I$. C'est à dire une variance des perturbations conditionnelle aux variables explicative indépendante des variables explicatives et l'absence de corrélation entre les perturbations. Ces hypothèses sont appelées hypothèses **d'homoscédasticité.** Les situations alternatives sont qualifiées d'**hétéroscédastiques**. On distingue l'hétéroscédasticité relative aux perturbations : $V(\underline{u} \,|\underline{x}) = V(\underline{u}) \neq \sigma^2 I$, de l'hétéroscédasticité relative aux variables explicatives $V(\underline{u} \,|\underline{x}) \neq V(\underline{u})$.

6.1.1 Quelques exemples

Exemple ***Séries temporelles*** *avec 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* $t\neq t'$, $E(\varepsilon_t^2|X)=\sigma_\varepsilon^2$. *Donc*

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

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

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

Exemple ***Séries temporelles*** *avec erreurs distribuées suivant un processus autoregressif :*

$$\begin{aligned} y_t &= x_tb+u_t\\ u_t &= \rho u_{t-1}+\varepsilon_t \end{aligned}$$

$u_t=\sum_{s=0}^{\infty}\rho^s\varepsilon_{t-s}$. *Là encore on suppose* $E(\varepsilon_t|X)=0$, $E(\varepsilon_t\varepsilon_{t'}|X)=0$ *pour* $t\neq t'$, $E(\varepsilon_t^2|X)=\sigma_\varepsilon^2$. *Un calcul similaire au précédent donne*

$$\begin{aligned} E\left(u_tu_{t-k}|X\right) &= E\left(\left(\sum_{s=0}^{\infty}\rho^s\varepsilon_{t-s}\right)\left(\sum_{s=0}^{\infty}\rho^s\varepsilon_{t-k-s}\right)\right)\\ &= E\left(\left(\left(\sum_{s=0}^{k-1}\rho^s\varepsilon_{t-s}\right)+\left(\sum_{s=k}^{\infty}\rho^s\varepsilon_{t-s}\right)\right)\left(\sum_{s=0}^{\infty}\rho^s\varepsilon_{t-k-s}\right)\right)\\ &= E\left(\rho^k\left(\sum_{s=k}^{\infty}\rho^s\varepsilon_{t-s}\right)\left(\sum_{s=0}^{\infty}\rho^s\varepsilon_{t-k-s}\right)\right)=\sigma_\varepsilon^2\rho^k/\left(1-\rho^2\right) \end{aligned}$$

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

$$\begin{aligned} V\left(\underline{u}|\underline{x}\right) &= \sigma_\varepsilon^2/\left(1-\rho^2\right)\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^T\\ \rho & \ddots & \ddots & \ddots & \vdots\\ \rho^2 & \ddots & & \ddots & \rho^2\\ \vdots & \ddots & \ddots & \ddots & \rho\\ \rho^T & \cdots & \rho^2 & \rho & (1+\rho^2) \end{pmatrix}\\ &\neq \sigma^2 I_T \end{aligned}$$

Exemple ***Séries temporelles*** *avec erreurs corrélées sans restrictions :*

$$y_t=x_tb+u_t$$

Là encore on suppose $E(u_t|X) = 0$, *mais par contre on ne fait plus d'hypothèses sur la structure des corrélations. La matrice de variance covariance est quelconque. Dans une spécification plus contrainte, on peut supposer que la variance des résidus est constante et que le coefficient de corrélation entre deux périodes ne dépend que de l'écart entre ses deux périodes :* $Cov(u_t, u_{t-s}) = \sigma_u^2 \rho_s$. *La matrice de variance covariance s'écrit alors pour un échantillon de taille* T

$$\begin{aligned} V(\underline{u}|\underline{x}) &= \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_T \\ \rho_1 & \ddots & \ddots & \ddots & \vdots \\ \rho_2 & \ddots & & \ddots & \rho_2 \\ \vdots & \ddots & \ddots & \ddots & \rho_1 \\ \rho_T & \cdots & \rho_2 & \rho & (1+\rho^2) \end{pmatrix} \\ &\neq \sigma^2 I_T \end{aligned}$$

Le nombre de paramètre de la matrice de variance tend vers l'infini lorsque la taille de l'échantillon augmente.

Exemple ***Modèle à coefficients aléatoires*** $(dim(x_i) = 1)$

$$\begin{aligned} y_i &= a + x_i b_i + v_i \\ b_i &= b + v_{bi} \end{aligned}$$

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

$$\begin{aligned} y_i &= a + x_i b_i + v_i = a + x_i(b + v_{bi}) + v_i \\ &= a + x_i b + x_i v_{bi} + v_i = a + x_i b + u_i \\ u_i &= x_i v_{bi} + v_i \end{aligned}$$

et on a donc les propriétés

$$E(u_i|\underline{x}) = E(x_i v_{bi} + v_i|\underline{x}) = x_i E(v_{bi}|\underline{x}) + E(v_i|\underline{x}) = 0$$

d'où l'expression de la matrice de variance

$$\begin{aligned} E(u_i u_j|\underline{x}) &= 0 \;\forall\; i \neq j \\ &= E((x_i v_{bi} + v_i)(x_j v_{bj} + v_j)|\underline{x}) \\ &= x_i x_j E(v_{bi}v_{bj}|\underline{x}) + x_i E(v_{bi}v_j|\underline{x}) + x_j E(v_i v_{bj}|\underline{x}) + E(v_i v_j|\underline{x}) = 0 \\ E(u_i^2|\underline{x}) &= x_i^2\sigma_b^2 + \sigma_v^2 \\ &= E\left((x_i v_{bi} + v_i)^2|\underline{x}\right) = E\left(\left(x_i^2 v_{bi}^2 + 2x_i v_{bi} v_i + v_i^2\right)|\underline{x}\right) \end{aligned}$$

La matrice de variance covariance s'écrit donc

$$\begin{aligned} V(\underline{u}\,|\underline{x}) &= Diag\left(\sigma_v^2 + x_i^2\sigma_b^2\right) \\ &\neq \sigma^2 I_N \end{aligned}$$

Dans ce cas, la matrice est bien diagonale, mais les éléments diagonaux sont des fonctions de x_i.

Exemple ***Modèle hétéroscédastique en coupe, à forme d'hétéroscédasticité connue***

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

avec , $E(u_i\,|\underline{x}) = 0$, $(u_i u_j\,|\underline{x}) = 0$ pour $i \neq j$, $E(u_i^2\,|\underline{x}) = g(x_i, \theta)$. La forme de la fonction g est connue mais le paramètre θ est inconnu. La matrice de variance covariance s'écrit alors

$$\begin{aligned} V(\underline{u}\,|\underline{x}) &= Diag\left(g(x_i, \theta)\right) \\ &\neq \sigma^2 I_N \end{aligned}$$

Dans ce cas la matrice de variance dépend d'un nombre de paramètre infini.

Exemple ***Modèle hétéroscédastique pur en coupe***

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

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

$$\begin{aligned} V(\underline{u}\,|\underline{x}) &= Diag\left(\sigma_i^2\right) \\ &\neq \sigma^2 I_N \end{aligned}$$

Dans ce cas la matrice de variance dépend d'un nombre de paramètre infini.

Exemple ***Données de panel.*** *D'autres exemples sont fournis par les données à double indice ou encore données de panel*

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

Ces données correspondent à la situation dans laquelle on suit des individus au cours du temp. i est un indice représentant les individus. Le nombre d'individus observés est en général grand. t est l'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 en empilant les observations relatives à un même individu :

$$\underline{y}_i = \underline{x}_i b + \underline{u}_i \quad i = 1, \ldots, N$$

On fait les hypothèses : $E(\underline{u}_i|\underline{x}) = 0$, $E\left(\underline{u}_i\underline{u}'_j|\underline{x}\right) = 0 \; \forall \; i \neq j$, *c'est à dire la condition d'identification est satisfaites, et les observations relatives à deux individus différents sont non corrélées. En revanche* ***on ne fait pas l'hypothèse*** $E(\underline{u}_i\underline{u}'_i|\underline{x}) = \sigma^2 I_T$. *Le résidu* u_{it} *incorpore des éléments inobservés permanent dans le temps. Il est modélisé suivant le* ***Modèle à erreurs composées***

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

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

$$\Omega = V(\underline{u}_i|\underline{x}) = \begin{pmatrix} \sigma_\varepsilon^2 + \sigma_W^2 & \sigma_\varepsilon^2 & \cdots & \sigma_\varepsilon^2 \\ \sigma_\varepsilon^2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma_\varepsilon^2 \\ \sigma_\varepsilon^2 & \cdots & \sigma_\varepsilon^2 & \sigma_\varepsilon^2 + \sigma_W^2 \end{pmatrix}$$

ainsi que la matrice de variance covariance des résidus empilés

$$\begin{aligned} V(\underline{u}|\underline{x}) &= I_N \otimes \Omega \\ &\neq \sigma^2 I_{NT} \end{aligned}$$

On peut remarquer qu'un cas intéressant est celui dans lequel sur le modèle précédent on considère les différences premières $\Delta y_{it} = y_{it} - y_{it-1}$. *Dans ce cas l'effet individuel est éliminé. En notant*

$$\Delta\underline{u}_i = \begin{pmatrix} u_{iT} - u_{iT-1} \\ u_{iT-1} - u_{iT-2} \\ \vdots \\ u_{i2} - u_{i1} \end{pmatrix}$$

le modèle se réécrit

$$\Delta\underline{y}_i = \Delta\underline{x}_i b + \Delta\underline{u}_i \quad i = 1, \ldots, N$$

et la matrice de variance des perturbations est alors :

$$\Omega = V(\Delta\underline{u}_i|\underline{x}) = \sigma_\varepsilon^2 \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & \ddots & 0 \\ 0 & \ddots & \ddots & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$$

La caractéristique intéressante est que cette matrice est connue à une constante multiplicative prés.

Exemple ***Régressions empilées :*** *On a considéré jusqu'à présent le cas dans lequel il n'y avait qu'une équation. On est parfois amené à s'intéresser à un ensemble d'équations.*

On pourrait en toute généralité se dire que l'on va estimer ces équations une par une. Ce serait possible mais parfois ce n'est pas suffisant. En effet, on peut vouloir examiner si certaines propriétés faisant intervenir des coefficients de différentes équations sont satisfaites. On peut en fait généraliser facilement le cadre à une équation au cas d'équations multiples. On considère la situation dans laquelle il y a M *variables à expliquer, et* $K+1$ *variables explicatives :*

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

Le modèle s'écrit pour chaque variable dépendante :

$$y_{mi} = x_i b_m + u_{mi} \quad i = 1, \ldots, N$$

ou encore

$$\begin{pmatrix} y_{1i} \\ \vdots \\ y_{Mi} \end{pmatrix} = \begin{pmatrix} x_i & 0 & \\ 0 & \ddots & 0 \\ & 0 & x_i \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_M \end{pmatrix} + \begin{pmatrix} u_{1i} \\ \vdots \\ u_{Mi} \end{pmatrix}$$
$$\underline{y}_i = Diag\left(\underline{x}_i\right) b + \underline{u}_i \quad i = 1, \ldots, N,$$

On fait les hypothèses $E\left(\underline{u}_i \left| \underline{x}\right.\right) = 0$, $cov\left(\underline{u}_{i\underline{u}j} \left| \underline{x}\right.\right) = 0 \ \forall \ i \neq j, V\left(\underline{u}_i \left| \underline{x}\right.\right) = \Sigma$. *Les résidus* u_{mi} *n'ont pas nécessairement la même variance et peuvent en outre être corrélés entre eux. La matrice de variance covariance des résidus empilés a alors pour expression*

$$\begin{aligned} E\left(\underline{u}\underline{u}' \left| \underline{x}\right.\right) &= I_N \otimes \Sigma \\ &\neq \sigma^2 I_{NT} \end{aligned}$$

Tel qu'il est écrit ce modèle n'impose pas de contraintes entre les paramètres des différentes équations. On pourrait néanmoins se trouver dans une situation dans laquelle les paramètres de la régression sont fonction d'un paramètre alternatif de dimension plus faible : $b = Hc$ *avec* $\dim b > \dim c$ *et* H *une matrice. le modèle s'écrit dans ce cas :*

$$\begin{aligned} \underline{y}_i &= Diag\left(\underline{x}_i\right) Hc + \underline{u}_i \quad i = 1, \ldots, N \\ &= \widetilde{\underline{x}}_i c + u_i \end{aligned}$$

6.1.2 Conclusion des exemples et définition du modèle linéaire hétéroscédastique

On conclut de ces exemples qu'il y a une grande diversité de situations. La matrice de variance des perturbations peut

- dépendre de paramètres additionnels de dimension finie. C'est le cas par exemple des données de panel, des régressions empilées, des modèles de série temporelle avec erreur distribuée suivant un processus autoregressif d'ordre 1 ou une moyenne mobile.

- dépendre ou non des variables explicatives. C'est le cas par exemple du modèle à coefficients aléatoires, du modèle hétéroscédastique avec hétéroscédasticité de forme connue.
- dépendre de paramètres additionnels de dimension infinie. C'est le cas du modèle hétéroscédastique pur en coupe ou des séries temporelles avec structure de corrélation quelconque.

Definition *On appelle modèle linéaire hétéroscédastique le modèle dans lequel un vecteur de variables aléatoires* $\underline{y}$ *dépend linéairement de* $K+1$ *variables explicatives* $\underline{x}$:

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

avec les hypothèses

1. $H1 : E(\underline{u}\,|\underline{x}) = 0$
2. $H2 : V(\underline{u}\,|\underline{x}) = \Omega = \Sigma(\underline{x}, \theta)$ *inversible*
3. $H3 : \underline{x}'\underline{x}$ *inversible*

Le modèle est dit **hétéroscédastique** car on n'a plus l'hypothèse $H2 : V(\underline{u}\,|\underline{x}) = \sigma^2 I$ dans un tel cas le modèle aurait été dit **homoscédastique.**

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 ?
- Quelles sont les propriétés asymptotiques des estimateurs ?

On ne peut pas espérer avoir un cadre général permettant de traiter toutes les situations. Les réponses que l'on va pouvoir apporter à ces questions dépendent du cas considéré.

6.2 Estimation par les MCO et les MCG

6.2.1 Propriétés des moindres carrés ordinaires

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

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

et sa variance sachant $\underline{x}$ *est*

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

Démonstration *On a*

$$\begin{aligned}\widehat{b}_{MCO} &= (\underline{x}'\underline{x})^{-1}\underline{x}'\underline{y} = (\underline{x}'\underline{x})^{-1}\underline{x}'(\underline{x}b+\underline{u}) \\ &= b+(\underline{x}'\underline{x})^{-1}\underline{x}'\underline{u}\end{aligned}$$

On a donc pour l'espérance de l'estimation

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

De plus

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

On voit donc que la propriété de "sans biais" n'est pas affectée par la relaxation de l'hypothèse $H2$. En revanche, on voit que la formule de la variance de l'estimateur est différente. Ce sont donc les écarts-type des paramètres qui sont différents. Cette conclusion est générale. Dans le cadre du modèle linéaire, le principal problème posé par l'existence d'hétéroscédasticité concerne le calcul de la précision des estimateurs et corrélativement la validité des différents tests que l'on peut mettre en oeuvre en transposant directement les procédure issue de l'hypothèse IID.

6.2.2 La méthode des Moindres Carrés Généralisés (MCG)

On introduit un autre estimateur appelé estimateur des moindres carrés généralisé. Il correspond à la minimisation de la distance entre les observations et l'espace engendré par les variables explicatives, non plus dans la métrique canonique de R^N, mais dans celle correspondant à Ω^{-1}.

Definition *L'estimateur des MCG est solution du problème :*

$$\widehat{b}_{MCG} = \arg\min \left\|\underline{y}-\underline{x}b\right\|^2_{\Omega^{-1}}$$

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

$$\widehat{b}_{MCG} = (\underline{x}'\Omega^{-1}\underline{x})^{-1}\underline{x}'\Omega^{-1}\underline{y}$$

Démonstration *Les conditions du premier ordre s'écrivent :*

$$\frac{\partial\left\|\underline{y}-\underline{x}\widehat{b}\right\|^2_{\Omega^{-1}}}{\partial b} = 2\underline{x}'\Omega^{-1}\left(\underline{y}-\underline{x}\widehat{b}\right) = 0 \Leftrightarrow \underline{x}'\Omega^{-1}\underline{x}\widehat{b} = \underline{x}'\Omega^{-1}\underline{y}$$

La matrice hessienne de l'objectif a pour expression

$$\frac{\partial \left\| \underline{y} - \underline{x}\widehat{b} \right\|^2_{\Omega^{-1}}}{\partial b \partial b'} = -2\underline{x}'\Omega^{-1}\underline{x}$$

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

$$\frac{\partial \left\| \underline{y} - \underline{x}\widehat{b} \right\|^2_{\Omega^{-1}}}{\partial b \partial b'} < 0 :$$

Les CN sont nécessaires et suffisantes, $\widehat{b}_{MCG} = (\underline{x}'\Omega^{-1}\underline{x})^{-1}\underline{x}'\Omega^{-1}\underline{y}$ car $\underline{x}'\Omega^{-1}\underline{x}$ inversible

Sphéricisation.

L'analyse des propriétés de l'estimateur des MCG est grandement simplifiée lorsque l'on applique aux observations une opération appelée sphéricisation.

Proposition *Pour toute matrice symétrique et définie positive W il existe une matrice $W^{-1/2}$ telle que*

$$W^{-1/2}WW^{-1/2\prime} = I$$

Cette matrice vérifie aussi

$$W^{-1/2\prime}W^{-1/2} = W^{-1}$$

Démonstration *Comme W est symétrique définie positive, elle est diagonalisable dans le groupe orthogonal. Il existe donc une matrice orthogonale P ($P'P = P^{-1}P = I$) telle que $W = P'DP$, où D est diagonale, les éléments de la diagonale étant strictement positifs puisque W est définie positive. On peut considérer $W^{-1/2} = P'D^{-1/2}P$, où $D^{-1/2}$ est la matrice diagonale dont les élément diagonaux sont les inverses de la racine des éléments diagonaux de D. On a*

$$\begin{aligned} W^{-1/2}WW^{-1/2\prime} &= P'D^{-1/2}PP'DPP'D^{-1/2}P \\ &= P'D^{-1/2}DD^{-1/2}P = P'P = I \end{aligned}$$

En outre si $W^{-1/2}WW^{-1/2\prime} = I$, alors

$$W^{-1/2\prime}W^{-1/2}WW^{-1/2\prime}W^{-1/2} = W^{-1/2\prime}W^{-1/2}$$

et donc

$$WW^{-1/2\prime}W^{-1/2} = I$$

d'où

$$W^{-1/2\prime}W^{-1/2} = W^{-1}$$

Ceci permet donc de définir une matrice $\Omega^{-1/2}$. Cette décomposition n'est pas unique. Par exemple on peut choisir $\Omega^{-1/2}$ semi-définie positive. Mais on peut aussi la choisir de telle sorte qu'elle ait d'autres propriétés, un choix qui peut être utile est celui dans lequel la matrice est triangulaire inférieure.

L'opération de sphéricisation consiste à multiplier le modèle par l'une de ces matrices $\Omega^{-1/2}$. On a :

$$\begin{aligned} \Omega^{-1/2}\underline{y} &= \Omega^{-1/2}\underline{x}b + \Omega^{-1/2}\underline{u} \\ \widetilde{y} &= \widetilde{x}b + \widetilde{u} \end{aligned}$$

Les hypothèses du modèle peuvent se transposer en partie au cas du modèle sphéricisé :

$HS1 : E(\widetilde{u}\,|\widetilde{x}) = E\left(\Omega^{-1/2}\underline{u}\,\middle|\Omega^{-1/2}\underline{x}\right) = \Omega^{-1/2}E(\underline{u}\,|\underline{x}) = 0$

$HS2 : E(\widetilde{u}\widetilde{u}'\,|\widetilde{x}) = E\left(\Omega^{-1/2}\underline{u}\underline{u}'\Omega^{-1/2\prime}\,\middle|\Omega^{-1/2}\underline{x}\right) = \Omega^{-1/2}E(\underline{u}\underline{u}'\,|X)\,\Omega^{-1/2\prime} = \Omega^{-1/2}\Omega\Omega^{-1/2\prime} = I$

$HS3 : \widetilde{x}'\widetilde{x} = \underline{x}'\Omega^{-1/2\prime}\Omega^{-1/2}\underline{x} = \underline{x}'\Omega^{-1}\underline{x}$ inversible

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

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

Exemple *Sphéricisation du modèle hétéroscédastique en coupe. On a vu que pour ce modèle la matrice de variance des perturbations s'écrit :*

$$V(\underline{u}\,|\underline{x}) = Diag\left(g(x_i,\theta)\right)$$

On vérifie directement que pour sphériciser le modèle on peut prendre

$$\Sigma^{-1/2} = Diag\left(g(x_i,\theta)^{-\frac{1}{2}}\right)$$

Exemple *Sphéricisation du modèle à perturbation $AR(1)$. On a vu que pour ce modèle on a*

$$V(\underline{u}\,|\underline{x}) = \sigma_\varepsilon^2/\left(1-\rho^2\right)\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^T \\ \rho & \ddots & \ddots & \ddots & \vdots \\ \rho^2 & \ddots & & \ddots & \rho^2 \\ \vdots & \ddots & \ddots & \ddots & \rho \\ \rho^T & \cdots & \rho^2 & \rho & (1+\rho^2) \end{pmatrix}$$

et on vérifie sans peine que l'on peut prendre

$$\Sigma^{-1/2} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & \cdots & \cdots & \cdots & 0 \\ -\rho & 1 & \ddots & & & \vdots \\ 0 & -\rho & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 1 & 0 \\ 0 & \cdots & \cdots & 0 & -\rho & 1 \end{bmatrix}$$

L'estimateur des MCG peut alors être calculé comme estimateur des mco appliqué au modèle :

$$\begin{pmatrix} y_1\sqrt{1-\rho^2} \\ y_2 - \rho\, y_1 \\ \vdots \\ y_T - \rho\, y_{T-1} \end{pmatrix} = \begin{pmatrix} x_1\sqrt{1-\rho^2} \\ x_2 - \rho x_1 \\ \vdots \\ x_T - \rho\, x_{T-1} \end{pmatrix} b + \begin{pmatrix} u_1\sqrt{1-\rho^2} \\ u_2 - \rho\, u_1 \\ \vdots \\ u_T - \rho\, u_{T-1} \end{pmatrix}$$

Exemple *Sphéricisation des données de panel. On a vu que pour des données de panel lorsque les résidus étaient modélisés comme*

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

avec indépendance des ε_i et des w_{it}, la matrice de variance s'écrivait

$$V(\underline{u}_i) = \begin{pmatrix} \sigma_\varepsilon^2 + \sigma_\omega^2 & \sigma_\varepsilon^2 & \cdots & \sigma_\varepsilon^2 \\ \sigma_\varepsilon^2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma_\varepsilon^2 \\ \sigma_\varepsilon^2 & \cdots & \sigma_\varepsilon^2 & \sigma_\varepsilon^2 + \sigma_\omega^2 \end{pmatrix}$$

Il est commode d'introduire deux matrices permettant de décomposer cette matrice

$$B = \frac{J_T}{T}, \text{ et } W = I_T - B$$

avec $J_T = e_T e_T'$, où $e_T' = (1, \ldots, 1)$. J_T est donc une matrice $T \times T$ dont chaque élément vaut 1. On vérifie sans peine que ces deux matrices sont symétriques et semi-définies positives. Elles vérifient en outre les propriétés

$$\begin{aligned} B^2 &= B \\ W^2 &= W \\ BW &= WB = 0 \end{aligned}$$

Ces deux matrices ont en outre une interprétation simple. En effet on vérifie que $Bz_i = e_T z_{i\cdot}$*, où* $z_{i\cdot}$ *est la moyenne individuelle des observations de l'individu* i *:* $z_{i\cdot} = (z_{i1} + \cdots + z_{iT})/T$*. Il en suit que* $W\underline{z}_i$ *est le vecteur formé des écarts à la moyenne individuelle. On peut exprimer simplement la matrice de variance des perturbations du modèle à erreurs composées à partir de ces deux matrices. On a en effet :*

$$V(\underline{u}_i) = \sigma_\varepsilon^2 J_T + \sigma_\omega^2 I_T = \left(\sigma_\omega^2 + T\sigma_\varepsilon^2\right) B + \sigma_\omega^2 W$$

Les matrices de la forme $\lambda B + \mu W$ *sont stables par multiplication* $(\lambda B + \mu W)(\lambda' B + \mu' W) = \lambda\lambda' B + \mu\mu' W$*. On en déduit sans peine que*

$$V(\underline{u}_i)^{-1/2} = \frac{1}{\sqrt{(\sigma_\omega^2 + T\sigma_\varepsilon^2)}} B + \frac{1}{\sqrt{\sigma_\omega^2}} W \propto W + \sqrt{\frac{\sigma_\omega^2}{(\sigma_\omega^2 + T\sigma_\varepsilon^2)}} B = I + \theta B$$

où $\theta = \sqrt{\sigma_\omega^2/(\sigma_\omega^2 + T\sigma_\varepsilon^2)} - 1$*. On en déduit que pour sphériciser les données il est possible de rajouter aux observations* y_{it} *et* x_{it} $\theta \times$ *la moyenne individuelle des observations (*$y_{i\cdot}$ *ou* $x_{i\cdot}$*). La quantité* θ *est inconnue, mais on peut la calculer aisément à partir de la matrice de variance covariance des résidus estimés par les mco ou à partir de deux estimateurs annexes : l'estimateur Between, estimateur des mco sur les moyennes individuelles dont la variance résiduelle est* $\sigma_B^2 = \sigma_\varepsilon^2 + \sigma_\omega^2/T$ *et l'estimateur Within, estimateur des mco sur les écarts aux moyennes individuelles dont la matrice de variance est* $\sigma_W^2 = \sigma_\omega^2 (T-1)/T$ *On voit donc que*

$$\frac{\sigma_\omega^2}{(\sigma_\omega^2 + T\sigma_\varepsilon^2)} = \frac{\sigma_W^2 T/(T-1)}{\sigma_B^2 T} = \frac{\sigma_W^2}{(T-1)\sigma_B^2}$$

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

Proposition *L'estimateur des MCG vérifie les propriétés suivantes*

- *L'estimateur des MCG est sans biais :* $E\left(\widehat{b}_{MCG}\,|\underline{x}\right) = b$
- *L'estimateur des MCG a pour matrice de variance* $\mathrm{V}(\widehat{b}_{MCG}|\underline{x}) = (\underline{x}'\Omega^{-1}\underline{x})^{-1}$
- *L'estimateur des MCG est l'estimateur linéaire sans biais de variance minimale (Th. de Gauss Markov)*

Démonstration $\widehat{b}_{MCG} = (\underline{x}'\Omega^{-1}\underline{x})^{-1}\underline{x}'\Omega^{-1}\underline{y} = (\underline{x}'\Omega^{-1}\underline{x})^{-1}\underline{x}\Omega^{-1}(\underline{x}b + \underline{u})$

$$\Rightarrow \widehat{b}_{MCG} = b + (\underline{x}'\Omega^{-1}\underline{x})^{-1}\underline{x}\Omega^{-1}\underline{u}$$

On a donc

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

et aussi

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

L'optimalité provient directement du fait que n'importe quel estimateur linéaire sans biais du paramètre est aussi un estimateur linéaire sans biais du paramètre dans le modèle sphéricisé. Or dans ce modèle l'estimateur des MCO est optimal et cet estimateur est celui des MCG : $\widehat{b}_{MCG} = \widehat{\widehat{b}}_{MCO}$ *et* $\widehat{\widehat{b}}_{MCO}$ *optimal*

Les propriétés algébriques de l'estimateur des MCO du cas homoscédastique se transposent directement au cas des MCG. Néanmoins cette transposition est peu utile en pratique car on est rarement dans le cas où la matrice de variance est connue. Rappelons toutefois que dans le cas des données de panel on a vu que pour le modèle à erreurs composées la matrice de variance des erreurs du modèle en différence première était connue à un facteur multiplicatif prés.

6.3 L'estimateur des MCQG

L'estimateur des MCG ne peut en général pas être mis en oeuvre car on ne connaît pas la matrice de variance des perturbations Ω, sauf dans des cas très spécifiques. Il en résulte que l'estimateur des MCG et la matrice de variance des MCO ne sont pas calculables. Une façon de procéder est de chercher à estimer cette matrice et à remplacer dans l'expression de l'estimateur la matrice Ω par son estimateur.

Definition *Soit* $\widehat{\Omega}$ *un estimateur de* Ω*. On appelle estimateur des Moindres Carrés Quasi-Généralisés l'estimateur :*

$$\widehat{b}_{MCQG} = (\underline{x}'\widehat{\Omega}^{-1}\underline{x})^{-1}\underline{x}'\widehat{\Omega}^{-1}\underline{y}.$$

L'estimateur des MCQG n'est en général pas sans biais ni linéaire en $\underline{y}$ puisque $\widehat{\Omega}$ dépend de $\underline{y}$. Les propriétés de $\widehat{b}_{MCQG}$ ne peuvent donc être *qu'asymptotiques*. Ces propriétés vont dépendre du cas considéré. On s'intéresse donc à la convergence et à la distribution asymptotique des paramètres. Il faut en fait examiner les propriétés asymptotiques au cas par cas suivant la nature de l'hétéroscédasticité. On peut alors étudier de façon similaire les propriétés asymptotiques de l'estimateur des mco.

On va dans les trois chapitres suivants considérer les trois formes importantes d'hétéroscédasticité survolées dans la première partie de ce chapitre.

1. Cas où $\Omega = I_N \otimes \Sigma(\theta)$ et θ de dimension finie. C'est le cas des données de panel et des régressions empilées. L'hétéroscédasticité est relative à des corrélations entre observations, mais celle-ci sont suffisamment régulière.
2. Cas où $\Omega = I_N \otimes h(x_i, \theta)$. C'est le cas de l'hétéroscédasticité liée aux variables explicatives.
3. Cas des séries temporelles.

Chapitre 7

Le modèle hétéroscédastique en coupe

La situation que l'on considère est celle d'un modèle de régression en coupe

$$y_i = x_i b + u_i$$

pour lequel on fait certaines des hypothèses précédentes :
$H1$ $E(u_i|x_i) = 0$
$H2$ $\forall$ N $\overline{x_i' x_i}$ est inversible

Ces hypothèses garantissent l'existence de l'estimateur des mco et le fait qu'il soit sans biais. On a vu qu'il y a un grand nombre de situations dans lesquelles on ne peut pas faire l'hypothèse d'homoscédasticité : $V(u_i|x_i) = \sigma^2$. dès que cette hypothèse d'homoscédasticité n'est plus satisfaite, on sait que d'une part le calcul des écart-type est affecté et d'autre part qu'il est en théorie possible de définir des estimateurs plus précis. On peut donc s'intéresser à deux questions distinctes : comment faire de l'inférence robuste à cette situation d'hétéroscédasticité ? Ceci revient à s'interroger sur l'estimation de la matrice de variance de l'estimateur des mco. On peut y répondre sous des hypothèses générales en faisant un effort de spécification minimal du modèle, i.e. en laissant la variance des résidus pour chaque observation être spécifique à l'individu : $V(u_i|x_i) = \sigma_i^2$. Il s'agit du modèle hétéroscédastique pur. La deuxième question correspond à la mise en oeuvre d'estimateurs plus efficaces que les mco. Comme on l'a vu il s'agit de l'estimateur des MCQG. Il est alors nécessaire de spécifier la forme de la variance à partir d'un nombre de paramètre restreint : $V(u_i|x_i) = h(x_i, \theta)$. Comme on va le voir il est possible alors sous certaines hypothèses de mettre en oeuvre des estimateurs asymptotiquement équivalents à l'estimateur des MCG. Néanmoins si les résultats des estimations ne sont pas tellement affectés par ce type de procédure et la spécification de la variance, l'inférence que l'on fait (le résultat des tests) est fortement liée à ces hypothèses faites. Comme en général ces estimations sont faites dans de grands échantillons, le gain d'efficacité est parfois modeste par rapport au risques liés à une mauvaise spécification de la variance conditionnelle des

résidus. Au total la mise en oeuvre de l'estimateur des mCQG dans ce cadre est assez rare et la plupart du temps on se contente d'appliquer les mco et de faire de l'inférence robuste à la présence d'hétéroscédasticité.

7.1 Inférence robuste à l'hétéroscédasticité

On considère le modèle

$$y_i = x_i b + u_i$$

les résultats que l'on va montrer sont vrais sous des hypothèses très générales autorisant par exemple le fait que les observations ne soient pas équidistribuées. C'est par exemple le cas dans le modèle hétéroscédastique pur pour lequel $V(u_i|x_i) = \sigma_i^2$, et dans lequel on pourrait aussi faire l'hypothèse que les régresseurs ne sont pas distribués suivant une même loi. On va néanmoins se situer dans un cadre plus proche du précédent dans lequel on fera des hypothèses d'homogénéité plus fortes :

- $H1$ Les observations $(y_i, x_i) \in \mathbb{R}\times\mathbb{R}^{K+1}$, $i = 1, ..., N$, sont indépendantes et équidistribuées
- $H2$ $E(u_i x_i) = 0$
- $H3$ $V(x_i' u_i) = E(u_i^2 x_i' x_i)$ existe
- $H4$ $H4$ $\forall$ N $\underline{x}'\underline{x}$ et $E(x_i' x_i)$ sont inversibles
- $H5$ Les moment $|x_{ki} x_{li}|$ existent
- $H6$ Pour tout indice des variables explicatives l_1, l_2, l_3, l_4 les moments $u_i^2 |x_{l_1 i} x_{l_2 i}|$ $|u_i| |x_{l_1 i} x_{l_2 i} x_{l_3 i}|$ et $|x_{l_1 i} x_{l_2 i} x_{l_3 i} x_{l_4 i}|$ existent

Comme on le voit la différence essentielle avec le cadre homoscédastique est que l'on ne fait plus l'hypothèse $V(u_i|x_i) = \sigma^2$ on a une situation beaucoup plus générale dans laquelle par exemple $V(u_i|x_i) = g(x_i)$ avec g quelconque pourvu que $E(g(x_i) x_i' x_i)$ existe, ce qui est garanti dès lors que $V(u_i x_i)$ existe. On voit que cette plus grande généralité est néanmoins payée par une exigence plus forte sur la distribution des variable puisqu'il faut que les moments des variables existent jusqu'à l'ordre 4 (hypothèse $H6$). Cette dernière hypothèse est utile pour l'estimation de la matrice de variance. Elle permet d'obtenir la convergence en probabilité des moments d'ordre 4. On voit qu'elle est exigeante et que, même si elle est satisfaite, vraisemblablement il sera nécessaire qu'il y ait un grand nombre d'observations pour que la moyenne empirique d'un polynôme de degrés 4 des observations soit proche de sa valeur limite. N'importe quelle observation dans les queues de distributions aura un effet important sur ces moments qui ne sera résorbé que si le nombre d'observations est grand. C'est pourquoi la notion de propriétés asymptotiques signifie ici plus qu'ailleurs que le nombre d'observations est grand.

7.1.1 Propriétés asymptotiques de l'estimateur

Proposition *Sous les hypothèses H1 à H6, l'estimateur des MCO*

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

vérifie quand $N \to \infty$

1. $\widehat{b}_{mco} \xrightarrow{P} b$, *l'estimateur est convergent*
2. $\sqrt{N}\left(\widehat{b}_{mco} - b\right) \xrightarrow{L} N\left(0, V_{as}\left(\widehat{b}_{mco}\right)\right)$, *l'estimateur est asymptotiquement normal*
3. $V_{as}\left(\widehat{b}_{mco}\right) = \left[E(x_i'x_i)\right]^{-1} E(u_i^2 x_i'x_i)\left[E(x_i'x_i)\right]^{-1}$

   Sous les hypothèses H1-H7 on a en plus
4. $\widehat{V}\left(\widehat{b}_{mco}\right) = \overline{(x_i'x_i)}^{-1}\overline{\widehat{u}_i^2 x_i'x_i}\,\overline{x_i'x_i}^{-1} \xrightarrow{P} V\left(\widehat{b}_{mco}\right)$ *on peut estimer la matrice de variance*
5. $\sqrt{N}\widehat{V}\left(\widehat{b}_{mco}\right)^{-1/2}\left(\widehat{b}_{mco} - b\right) \xrightarrow{L} \mathcal{N}(0, I)$

Démonstration ***Convergence*** $\widehat{b}_{mco} = b + \left(\overline{x_i'x_i}\right)^{-1}\overline{x_i'u_i}$. *L'existence des moments* $|x_{ki}x_{li}|$ *de* x_i *garantie la convergence de* $\overline{x_i'x_i} \xrightarrow{P} E(x_i'x_i)$. *La seule chose qu'il y ait à montrer est la convergence de* $\overline{x_i'u_i}$ *vers* $E(x_i'u_i)$. *Pour cela on applique la loi des grands nombres :* $E(x_i'u_i) = 0$ *et* $V(x_i'u_i) = E(u_i^2 x_i'x_i)$. *On est dans les condition d'application de la loi des grands nombres :* $\overline{x_i'u_i} \xrightarrow{P} E(x_i'u_i) = 0$.

Normalité asymptotique *La encore la seule question est celle de la convergence de* $\sqrt{N}\overline{x_i'u_i}$. *mais comme précédemment, l'existence des moments d'ordre 1 et 2 de* $x_i'u_i$, $E(x_i'u_i) = 0$, $V(x_i'u_i) = E(u_i^2 x_i'x_i)$ *garantissent que* $\sqrt{N}\overline{x_i'u_i} \xrightarrow{L} \mathcal{N}(0, E(u_i^2 x_i'x_i))$. *Il en résulte que* $\sqrt{N}\left(\widehat{b}_{mco} - b\right) = \sqrt{N}\overline{x_i'x_i}^{-1}\overline{x_i'u_i} \xrightarrow{L} N\left(0, E(x_i'x_i)^{-1} E(u_i^2 x_i'x_i) E(x_i'x_i)^{-1}\right)$

Convergence de l'estimation de la matrice de variance.

Le point important est de montrer que $\overline{\widehat{u}_i^2 x_i'x_i} \xrightarrow{P} E(u_i^2 x_i'x_i)$

$$\begin{aligned}
\overline{\widehat{u}_i^2 x_i'x_i} &= \overline{\left(x_i\left(b - \widehat{b}_{mco}\right) + u_i\right)^2 x_i'x_i} \\
&= \overline{u_i^2 x_i'x_i} + \overline{\left(x_i\left(b - \widehat{b}_{mco}\right)\right)^2 x_i'x_i} + \\
&\quad 2\left(b - \widehat{b}_{mco}\right)\overline{x_i'u_i x_i'x_i}
\end{aligned}$$

Pour que le premier terme converge en probabilité vers son espérance, il est nécessaire que les éléments qui la forme $u_i^2 x_{l_1 i} x_{l_2 i}$ *satisfasse la loi de grands nombres. Ce qui est garanti par la propriété H6. Le troisième terme tend alors vers zéro en probabilité puisque* $\overline{x_i'u_i x_i'x_i} \xrightarrow{P} E(x_i'u_i x_i'x_i) = 0$. *Le second terme tend aussi vers zéro puisque les éléments qui le constituent sont de la forme* $\left(b_k - \widehat{b}_{kmco}\right)\left(b_l - \widehat{b}_{lmco}\right)\overline{x_{li}x_{ki}x_{l_1 i}x_{l_2 i}}$ *et* $\overline{x_{li}x_{ki}x_{l_1 i}x_{l_2 i}} \xrightarrow{P} E(x_{li}x_{ki}x_{l_1 i}x_{l_2 i})$ *puisque les moments d'ordre 4 existent et que* $b_{kmco} - b_k \xrightarrow{P} 0$.

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'hétéroscédasticité**. Il est très couramment utilisé et systématiquement proposé dans les logiciels standards (sauf SAS).

Remarque *Là encore les résultats peuvent être généralisés au cas dans lequel on ne fait plus l'hypothèse d'équidistribution. Ceci permet en particulier de traiter le cas du modèle hétéroscédastique pur, dans lequel* $V(u_i|x_i) = \sigma_i^2$. *Tous les résultats découlent de l'application du théorème central limite de Liapounov à* $\overline{x_i' u_i}$. *Il faut donc que la condition de Liapounov soit satisfaite. Si on considère* $\overline{\sigma_N^2} = \sum_{n=1}^N \sigma_n^2 \Big/ N$ *et si on considère* $\gamma_i^3 = E(|u_i^3||x_i)$ *ainsi que* $\overline{\gamma_N^3} = \sum_{n=1}^N \gamma_n^3 \Big/ N$, *il suffit que* $\overline{\gamma_N^3} \Big/ \left(N^{\frac{1}{6}} \overline{\sigma_N^2}\right) \to 0$, *si par exemple les variable explicatives sont iid. On sait qu'alors .* $\left[\overline{\sigma_N^2} E(x_i' x_i)\right]^{-1} \sqrt{N \overline{x_i' u_i}} \xrightarrow{L} \mathcal{N}(0, I)$.

Ces résultats se généralisent directement sans modification au cas des données de panel et au cas des équations empilées. Si on considère le modèle

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

spécifié en terme de vecteur $\underline{y}_i$, $\underline{x}_i$ et $\underline{u}_i$. Sous des hypothèses convenables, dont la condition d'identification $E(\underline{u}_i|\underline{x}_i) = 0$, et l'analogue de la condition précédente pour la variance $E(\underline{x}_i' \underline{u}_i \underline{u}_i' \underline{x}_i)$ existe et des conditions sur l'existence de moments des variables d'un ordre élevé. On a l'extension des résultats précédents :

1. $\widehat{b}_{mco} \xrightarrow{P} b$, l'estimateur est convergent
2. $\sqrt{N}\left(\widehat{b}_{mco} - b\right) \xrightarrow{L} N\left(0, V_{as}\left(\widehat{b}_{mco}\right)\right)$, l'estimateur est asymptotiquement normal
3. $V_{as}\left(\widehat{b}_{mco}\right) = \left[E(\underline{x}_i' \underline{x}_i)\right]^{-1} E\left(\underline{x}_i' \underline{u}_i \underline{u}_i' \underline{x}_i\right) \left[E(\underline{x}_i' \underline{x}_i)\right]^{-1}$
4. $\widehat{V}\left(\widehat{b}_{mco}\right) = \overline{(\underline{x}_i' \underline{x}_i)}^{-1} \overline{\underline{x}_i' \widehat{\underline{u}}_i \widehat{\underline{u}}_i' \underline{x}_i} \overline{\underline{x}_i' \underline{x}_i}^{-1} \xrightarrow{P} V\left(\widehat{b}_{mco}\right)$ on peut estimer la matrice de variance
5. $\sqrt{N}\widehat{V}\left(\widehat{b}_{mco}\right)^{-1/2}\left(\widehat{b}_{mco} - b\right) \xrightarrow{L} N(0, I)$

7.1.2 Test d'hypothèses dans le modèle hétéroscédastique

L'intérêt de ces résultats est bien sur la possibilité d'effectuer des tests. On s'intéresse à des tests d'une hypothèse nulle de la forme $H_0 : Rb = r$.

Proposition *Sous les hypothèses H1-H7,*

$$\sqrt{N}\left(R\overline{(x_i' x_i)}^{-1} \overline{\widehat{u}_i^2 x_i' x_i} \overline{x_i' x_i}^{-1} R'\right)^{-1/2}\left(R\widehat{b}_{mco} - r\right) \xrightarrow{L} N(0, I_p)$$

où p *est le nombre de ligne de la matrice* R. *Sous l'hypothèse* $H_0 : Rb = r$, *la statistique*

$$\widehat{S} = N\left(R\widehat{b}_{mco} - r\right)'\left[R\overline{(x_i' x_i)}^{-1} \overline{\widehat{u}_i^2 x_i' x_i} \overline{x_i' x_i}^{-1} R'\right]^{-1}\left(R\widehat{b}_{mco} - r\right) \xrightarrow{L} \chi^2(p)$$

Un test de H_0 contre $H_1 : Rb \neq r$ peut être effectué à partir de la région critique $W = \left\{ \widehat{S} \left| \widehat{S} > q\left(\chi^2(p), 1-\alpha\right) \right. \right\}$ où $q\left(\chi^2(p), 1-\alpha\right)$ est le quantile d'ordre $1-\alpha$ de la loi du χ^2 à p degrés de libertés.

Remarque *On remarque que dans ce cas le principe du test de Fisher se généralise. Dans le cas homoscédastique, le test de Fisher consistait à regarder comme ici si $R\widehat{b}_{mco} - r$ est proche ou non de zéro. On a vu que dans le cas homoscédastique, il est possible de réécrire la statistique à partir des sommes des carrés des résidus sous les hypothèses nulles et alternatives. Ici cette dernière simplification n'est plus possible. Il faut donc prendre garde au fait que dans de nombreux logiciels on peut simplement mettre en oeuvre les tests de Fisher, mais que ceux-ci sont faits sous l'hypothèse d'homoscédasticité.*

Remarque *Le principe du test se généralise là aussi au test d'hypothèses non linéaire de la forme $H_0 : g(b) = 0$. On utilise là encore la méthode delta. La statistique de test est de la forme $\widehat{S} = N g(\widehat{b})' \left[\frac{\partial g(\widehat{b})}{\partial b'} \widehat{V_{as}} \left(\widehat{b}\right) \left(\frac{\partial g(\widehat{b})}{\partial b'} \right)' \right]^{-1} g(\widehat{b})$. Par rapport au cas homoscédastique, la seule différence est que la matrice de variance à prendre en compte est la matrice de variance robuste.*

7.1.3 Estimation sous contraintes linéaires en présence d'hétéroscédasticité

On ne présente pas ici tous les résultats. L'estimateur des moindres carrés contraints est toujours calculé de la même manière comme

$$\begin{aligned} \hat{b}_{mcc} &= (\underline{x}'\underline{x})^{-1}\,\underline{x}'\underline{y} - (\underline{x}'\underline{x})^{-1}\,R'\left[R(\underline{x}'\underline{x})^{-1}\,R'\right]^{-1}\left[R(\underline{x}'\underline{x})^{-1}\,\underline{x}'\underline{y} - r\right] \\ &= \hat{b}_{mco} - (\underline{x}'\underline{x})^{-1}\,R'\left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1}\left[R\,\hat{b}_{mco} - r\right] \end{aligned}$$

On a

$$\begin{aligned} \hat{b}_{mcc} - b &= \left[I - (\underline{x}'\underline{x})^{-1}\,R'\left[R(\underline{x}'\underline{x})^{-1}R'\right]^{-1}R\right](\underline{x}'\underline{x})^{-1}\underline{x}'\underline{u} \\ &= \left[I - \overline{x_i'x_i}^{-1}\,R'\left[R\overline{x_i'x_i}^{-1}R'\right]^{-1}R\right]\overline{x_i'x_i}^{-1}\overline{x_i'u_i} \end{aligned}$$

Sous les même hypothèses que précédemment, on peut déterminer la loi asymptotique de l'estimateur et un estimateur convergent de la matrice de variance asymptotique.

1. $\widehat{b}_{mco} \xrightarrow{P} b$, l'estimateur est convergent
2. $\sqrt{N}\left(\widehat{b}_{mco} - b\right) \xrightarrow{L} N\left(0, V_{as}\left(\widehat{b}_{mco}\right)\right)$, l'estimateur est asymptotiquement normal
3. $V_{as}\left(\widehat{b}_{mco}\right) = [I-H]\,E(x_i'x_i)^{-1}E(u_i^2x_i'x_i)E(x_i'x_i)^{-1}\,[I-H']$
   avec $H = E(x_i'x_i)^{-1}\,R'\,[RE(x_i'x_i)^{-1}R']^{-1}\,R$

4. $\widehat{V}_{as}\left(\widehat{b}_{mco}\right) = \left[I - \widehat{H}\right] \overline{(x_i'x_i)}^{-1} \overline{\widehat{u}_i^2 x_i' x_i} \overline{x_i'x_i}^{-1} \left[I - \widehat{H}\right] \xrightarrow{P} V_{as}\left(\widehat{b}_{mco}\right)$,
   avec $\widehat{H} = \overline{x_i'x_i}^{-1} R' \left[R\overline{x_i'x_i}^{-1}R'\right]^{-1} R$
5. $\sqrt{N}\widehat{V}_{as}\left(\widehat{b}_{mco}\right)^{-1/2}\left(\widehat{b}_{mco} - b\right) \xrightarrow{L} N(0, I)$

7.2 Test d'hétéroscédasticité

7.2.1 Le test de Breush-Pagan

La différence essentielle entre l'approche avec hétéroscédasticité et l'approche sans hétéroscédasticité est que $E\left(u_i^2 x_i x_i'\right) \neq E\left(u_i^2\right) E\left(\ x_i x_i'\right)$. Un test naturel d'homoscédasticité consiste donc à tester si $E\left(u_i^2 x_i x_i'\right) = E\left(u_i^2\right) E\left(x_i x_i'\right)$. Ce qui revient exactement à tester la nullité globale du vecteur des coefficients de la projection orthogonale de u_i^2 sur les variables explicatives $x_{li}x_{mi}$ $l, m \leq K+1$ sauf la constante. Le test ne fait intervenir que la projection de u_i^2 et pas une modélisation de la forme de l'hétéroscédasticité. On ne spécifie pas en particulier

$$E\left(u_i^2 \left| x_i\right.\right) = \sum_{l,m \leq K+1} x_{li}x_{mi}\gamma_{lm}$$

et le test que l'on fait n'est pas $H_0 : E\left(u_i^2 \left| x_i\right.\right) = \sigma^2$ contre $H_1 : E\left(u_i^2 \left| x_i\right.\right) = \sum_{l,m \leq K+1} x_{li}x_{mi}\gamma_{lm}$ mais simplement celui de

$$H_0 : E\left(u_i^2 x_i x_i'\right) = E\left(u_i^2\right) E\left(x_i x_i'\right)$$

contre

$$H_1 : E\left(u_i^2 x_i x_i'\right) \neq E\left(u_i^2\right) E\left(x_i x_i'\right)$$

Le test se fait néanmoins au moyen de la régression

$$u_i^2 = \sum_{l,m \leq K+1} x_{li}x_{mi}\gamma_{lm} + v_i$$

Ici v_i est défini par la propriété $E\left(v_i x_{li} x_{mi}\right) = 0$. L'idée du test est de procéder au test de la nullité jointe des coefficients de la régressions précédente. Pour cela il faut connaître la loi asymptotique des estimateurs. On pourrait l'obtenir sous des conditions générales par exemple ne faisant pas d'hypothèses sur les moments d'ordre 2 de la forme $E\left(v_i^2 x_{li}x_{mi}x_{l'i}x_{m'i}\right)$. Néanmoins on fait en général le test de la nullité globale sous l'hypothèse d'homoscédasticité des résidus v_i : c'est à dire $E\left(v_i^2 x_{li}x_{mi}x_{l'i}x_{m'i}\right) = E\left(v_i^2\right) E\left(x_{li}x_{mi}x_{l'i}x_{m'i}\right)$. Dans ce cas le test est très simple à mettre en oeuvre il s'agit simplement du test de la nullité globale des coefficients dans une régression. Un problème

vient du fait que le résidu n'est pas observé mais seulement estimé, mais xomme pour les autres résultats asymptotiques que l'on a vu, il suffit de remplacer le résidu par le résidu estimé On a le résultat suivant :

Proposition *Dans le modèle*

$$y_i = x_i b + u_i$$

avec les hypothèses H1-H6, le test de l'hypothèse

$$H_0 : E\left(u_i^2 x_{li} x_{mi}\right)) = E\left(u_i^2\right) E\left(x_{li} x_{mi}\right)$$

peut être fait simplement comme un test de nullité jointe des coefficients sauf celui de la constnate dans le modèle de régression

$$u_i^2 = \sum_{l,m \leq K+1} x_{li} x_{mi} \gamma_{lm} + v_i$$

où v_i est défini par $E\left(v_i x_{li} x_{mi}\right) = 0$ et dans lequel on fait l'hypothèse de régularité $E\left(v_i^2 x_{li} x_{mi} x_{l'i} x_{m'i}\right) = \delta^2 E\left(x_{li} x_{mi} x_{l'i} x_{m'i}\right)$. Le test est mis en oeuvre à partir du modèle de régrsssion

$$\widehat{u}_i^2 = \sum_{l,m \leq K+1} x_{li} x_{mi} \gamma_{lm} + v_i$$

incluant $(K+1)(K+2)/2$ variables, dans lequel on fait un test de nullité jointes de tous les paramètres exceptée la constante. Sous H_0, la statistique NR^2 suit un $\chi^2\left((K+1)(K+2)/2-1\right)$. Un test convergent au niveau α peut être fait à partir de la région critique $\{NR^2 | \, NR^2 > q\left(\chi^2\left((K+1)(K+2)/2-1\right), 1-\alpha\right)\}$

Démonstration *Il est d'abord nécessaire de montrer que si pour une variable z_1 de dimension 1 et une variable z_2 de dimension q, l'hypothèse $E(z_1 z_2) = E(z_1) E(z_2)$ est analogue à l'hypothèse de nullité de la valeur limite des coefficients sauf la constante de la projection orthogonale de z_1 sur $(1, z_2)$. En effet les coefficients de z_2 sont obtenus directement comme ceux de la régression de la variable $z_1 - E(z_1)$ sur $z_2 - E(z_2)$. Ils ont donc pour expression $V\left(z_2^{-1}\right) E\left(\left(z_2 - E(z_2)\right)'\left(z_1 - E(z_1)\right)\right) = V\left(z_2^{-1}\right)\left(E\left(z_2' z_1\right) - E(z_2)' E(z_1)\right) = 0$.*

Le seul point restant à montrer est que sous les hypothèses faites l'estimateur des coefficients γ dans le modèle avec $\widehat{u}$ est asymptotiquement équivalent à celui avec u. Pour cela il suffit de montrer que $\sqrt{N}\left(\overline{z_i \widehat{u}_i^2} - \overline{z_i u_i^2}\right) \xrightarrow{P} 0$, avec z les éléments du type $x_{li} x_{mi}$. Comme $\widehat{u}_i = u_i + x_i\left(b - \widehat{b}\right)$, d'où $z_i \widehat{u}_i^2 = z_i u_i^2 + 2 z_i u_i x_i\left(b - \widehat{b}\right) + z_i x_i^2\left(b - \widehat{b}\right)^2$. Il en résulte que $\sqrt{N}\left(\overline{z_i \widehat{u}_i^2} - \overline{z_i u_i^2}\right) = 2\overline{z_i u_i x_i}\ \sqrt{N}\left(b - \widehat{b}\right) + \overline{z_i x_i^2}\sqrt{N}\left(b - \widehat{b}\right)^2$. Sous les hypothèses $H_0 - H_6$: $\overline{z_i u_i x_i} \xrightarrow{P} E(z_i u_i x_i) = E\left(z_i x_i E\left(u_i | x_i\right)\right) = 0$, donc $\overline{z_i u_i x_i} = o(1)$ et $\overline{z_i x_i^2} \xrightarrow{P} E\left(z_i x_i^2\right)$. En outre $\sqrt{N}\left(b - \widehat{b}\right) \xrightarrow{L} N\left(0, V_{as}\right)$, donc $\sqrt{N}\left(b - \widehat{b}\right) = O(1)$ et

$\overline{z_i x_i^2}\ \sqrt{N}\left(b-\widehat{b}\right) = O(1)$. *Comme* $\left(b-\widehat{b}\right) = o(1)$, $\overline{z_i x_i^2}\sqrt{N}\left(b-\widehat{b}\right)^2 = o(1)$. *Comme* $\overline{z_i u_i x_i} = o(1)$ *et* $\sqrt{N}\left(b-\widehat{b}\right) = O(1)$, $\overline{z_i u_i x_i}\ \sqrt{N}\ \left(b-\widehat{b}\right) = o(1)$.

Remarque *1. L'intérêt de ce test d'hétéroscédasticité est d'informer sur les situations dans lesquelles il est nécessaire d'effectuer la correction de White pour l'hétéroscédasticité. Si on accepte l'hypothèse d'homoscédasticité, alors on pourra estimer la matrice de variance des estimateurs sous sa forme standard, et on pourra effectuer les tests d'hypothèses linéaires comme on a vu à partir des sommes des carrés des résidus sous les hypothèses nulles et alternatives, ce qui présente un intérêt pratique certain. Sinon, on utilise la formule donnant la matrice robuste de White et les tests doivent être effectués comme on l'a montré dans le cadre hétéroscédastique.*

2. *Ce type de test s'étend aussi au cas dans lequel on spécifie un modèle pour l'hétéroscédasticité. On pourrait par exemple spécifier une forme d'hétéroscédasticité particulière, par exemple* $E(u_i^2|x_i)) = \sigma^2 + \sum_{l,m\leq K+1} x_{li}x_{mi}\gamma_{lm}$, *ou plus généralement* $E(u_i^2|x_i)) = \sum_{d<D} P_d(x)\gamma_d$, *avec* P_d *un ensemble de fonction et effectuer un test de la nullité jointe des paramètres pour tester l'absence d'hétéroscédasticité de la forme particulière imposée. Dans ce cas on aura un test de l'hypoyhèse*

$$H_0 : E\left(u_i^2|x_i\right)) = \sigma^2$$

contre

$$H_1 : E\left(u_i^2|x_i\right)) = \sigma^2 + z\gamma$$

dans lequel z *est un sous-ensemble des variables explicatives, peut être fait simplement à partir de la régression*

$$\widehat{u}_i^2 = a_0 + z\gamma + v_i$$

incluant K_Z *variables entrant dans* z*, dans lequel on fait un test de nullité jointes de tous les paramètres exceptée la constante. Sous* H_0*, la statistique* NR^2 *suit un* $\chi^2(K_Z)$*. Un test convergent au niveau* α *peut être fait à de la région critique* $\{NR^2 | NR^2 > q(\chi^2(K_Z), 1-\alpha)\}$

Le sens du test est néanmoins différents. Ces test sont des test portant sur un paramétrage de l'hétéroscédasticité, alors que le premier test ne porte que sur l'absence de covariance entre le résidus au carré et les polynômes d'ordre 2 des variables explicatives. Postuler une forme d'hétéroscédasticité est utile pour la prendre en compte par exemple pour mettre en oeuvre l'estimateur des mcqg. Exaliner l'absence de corrélation au deuxièmeordre est utile pour le choix du calcul de la matrice de variance.

7.2.2 Test de Goldfeld-Quandt

Une forme plus ancienne des tests d'hétéroscédasticité est donnée par le test de Goldfeld Quandt. Il s'agit d'une situation dans laquelle on suspecte qu'une variable donnée z joue sur la variance des régresseurs de façon monotone, c'est à dire $E\left(u_i^2 \left| x_i \right.\right) = \sigma^2 + h\left(z\right)$, avec $\dim z = 1$ et h une fonction croissante. L'idée du test de Goldfeld et Quandt est d'ordonner les observations en fonction de z_i et de partitionner ensuite les observations en deux groupes tels que

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

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 $N_2 \approx 2N/3$. L'idée du test de Goldfeldt et Quandt est de comparer les estimateurs des variances dans chaque sous échantillons

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

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

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

L'hypothèse nulle d'homoscédasticité est rejetée 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. Ce test n'est plus tellement utilisé. Il a été développé

dans le cadre spécifique dans lequel les résidus sont normaux et la statistique de test est exacte et non pas asymptotique. C'est la raison pour laquelle d'ailleurs les estimateurs du paramètre b sont différents dans les deux échantillons. Cela garantit en effet que les deux estimateurs des variances sont indépendants, ce qui est important pour construire la statistique de Fisher. Il en résulte d'ailleurs que le test effectué n'est pas nécessairement le test d'hétéroscédasticité puisque les hypothèses nulles et alternatives du test de Goldfeld et Quandt sont

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ et } b_1 - b_2 \in \Re$$

contre

$$H_1 : \sigma_1^2 \neq \sigma_2^2 \text{ et } b_1 - b_2 \in \Re$$

Alors que le test d'hétérogénéité pur est un test de

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ et } b_1 = b_2$$

contre

$$H_1 : \sigma_1^2 \neq \sigma_2^2 \text{ et } b_1 = b_2$$

En tout état de cause rien n'empêche de considérer des indicatrices d'appartenance aux trois sous échantillons $I_1 = \{i < N_1\}$, $I_2 = \{N_1 \leq i < N_2\}$ et $I_3 = \{N_2 \leq i\}$, et d'examiner les resultats de la régression

$$\widehat{u}_i^2 = \sigma^2 + \delta_1 I_1 + \delta_3 I_3 + v_i$$

et de tester l'égalité $\delta_1 = \delta_3$.

7.3 L'estimateur des MCQG dans le cas où $V(u_i | x_i) = h(\theta, x_i)$

Un cas pouvant se présenter est celui dans lequel on spécifie le moment d'ordre 1 et le moment d'ordre 2 d'une variable conditionnellement à des variables explicatives. On a alors un modèle de la forme

$$\begin{aligned} E(y_i | x_i) &= x_i b \\ V(y_i | x_i) &= h(x_i, \theta) > 0 \end{aligned}$$

où h est une fonction connue, mais θ un paramètre inconnu. On est typiquement dans un cas hétéroscédastique, et on sait que l'estimateur des MCG serait l'estimateur linéaire sans biais le plus efficace du paramètre b. Cet estimateur pourrait être obtenu en sphéricisant d'abord les observations, i.e. en divisant les variables explicatives et la variable dépendante par $\sqrt{h(x_i, \theta)}$ puis en appliquant l'estimateur des MCO. Néanmoins il n'est pas possible de mettre en oeuvre cette méthode directement car le paramètre θ est inconnu. On peut

néanmoins dans certaines situations avoir un estimateur convergent $\widehat{\theta}$ du paramètre θ, et on met alors en oeuvre l'estimateur des MCQG en divisant les variables par $\sqrt{h\left(x_i, \widehat{\theta}\right)}$. On étudie ici les conditions sous lesquelles l'estimateur obtenu est asymptotiquement équivalent à l'estimateur des MCG et sera donc l'estimateur de variance minimale. Il convient néanmoins de remarquer que ce type de démarche est rarement mis en oeuvre. En effet, on a tendance à privilégier la robustesse des estimations et les tailles d'échantillons parfois très grands dont on dispose incitent à le faire. Il s'agit ici non pas de la robustesse de l'estimateur du paramètre b mais de la robustesse et de la convergence de l'estimateur de la variance de ce paramètre. Les résultats de l'inférence faite lorsqu'on spécifie les deux moments sont nécessairement plus fragiles que lorsqu'on ne spécifie qu'un seul de ces deux moments.

- H_0 Les observations $(y_i, x_i) \in \mathbb{R}\times\mathbb{R}^{K+1}$, $i = 1, ..., N$, sont IID
- $H1$ $E(u_i|x_i) = 0$
- $H2$ $H2$ $V(u_i|x_i) = h(\theta, x_i)$ mesurable et dérivable
- $H3$ $H4$ $\forall$ N $\underline{x}'\underline{x}$ et $E(x_i'x_i)$ sont inversibles
- $H5$ Les moment $|x_{ki}x_{li}|$ existent
- $H6$ $\exists$ $\widehat{\theta} = \theta_0 + O\left(1/\sqrt{N}\right)$ où θ_0 est la vraie valeur du paramètre
- $H7$ $\exists$ une fonction $d(x_i)$ telle que $Max\left(|x_{li}'u_i|\,|\nabla h^{-1}(\theta, x_i)|\ , |x_{li}'u_i|\,h^{-1}(\theta, x_i)\,, \left|x_{l_1i}'x_{l_2i}\right| h^{-1}(\theta, x_i)\right) < d(x_i)$ et $E(d(x_i)) < \infty$

Ce modèle spécifie donc à la fois les moments d'ordre 1et 2 des résidus conditionnellement aux variables explicatives. La condition $H7$ garantie la convergence uniforme en θ des moyennes empiriques de fonction de θ vers leur espérance $\overline{h^{-1}(\theta, x_i)\,x_i'x_i} \xrightarrow{P} \mathbb{E}(h^{-1}(\theta, x_i)\,x_i'x_i)$, $\overline{h^{-1}(\theta, x_i)\,x_i'u_i} \xrightarrow{P} \mathrm{E}(h^{-1}(\theta, x_i)\,x_i'u_i)$ et $\overline{\nabla h^{-1}(\theta, x_i)\,x_i'u_i} \xrightarrow{P} E(\nabla h^{-1}(\theta, x_i)\,x_i'u_i)$.

Proposition *Sous les hypothèses H0 à H7, l'estimateur des MCQG*

$$\widehat{b}_{mcqg} = \overline{\left(\frac{x_i'x_i}{h\left(x_i, \widehat{\theta}\right)}\right)}^{-1} \overline{\left(\frac{x_i'y_i}{h\left(x_i, \widehat{\theta}\right)}\right)}$$

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. $\mathrm{V}_{as}\left(\widehat{b}_{mcqg}\right) = \left[\mathrm{E}(\frac{x_i'x_i}{h(x_i,\theta_0)})\right]^{-1} = \underline{\mathrm{V}\left(\widehat{b}_{mcg}\right)}$ *Equivalence des MCQG et des MCG*
4. $\widehat{\mathrm{V}}_{as}\left(\widehat{b}_{mcqg}\right) = \overline{\frac{x_i'x_i}{h(x_i,\theta_0)}}^{-1} \xrightarrow{P} \mathrm{V}_{as}\left(\widehat{b}_{mcqg}\right)$ *Estimation de la matrice de variance asymptotique* V_{as}

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* $\widehat{h}_i = h\left(\widehat{\theta}, x_i\right)$.

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

$\overline{h^{-1}\left(x_i,\widehat{\theta}\right)x_i'z_i} \xrightarrow{P} E\left(h^{-1}\left(x_i,\theta_0\right)x_i'z_i\right)$ *comme* $\theta\xrightarrow{P}\theta_0$ *et par l'hypothèse H7 qui garantit la convergence uniforme*

$$\overline{\widehat{h}_i^{-1}x_i'z_i}\xrightarrow{P}E\left(h_i^{-1}x_i'z_i\right)$$

D'où la convergence de l'estimateur puisque $E\left(h^{-1}\left(x_i,\theta_0\right)x_i'u_i\right)=0$.

Normalité asymptotique

Le seul point à montrer est $\sqrt{N}\overline{\widehat{h}_i^{-1}x_i'u_i}\xrightarrow{L}N\left(0,\mathrm{E}(\frac{x_i'x_i}{h(x_i,\theta_0)})\right)$

$$\sqrt{N}\overline{\widehat{h}_i^{-1}x_i'u_i} = \sqrt{N}\overline{\left(\widehat{h}_i^{-1}-h^{-1}\left(x_i,\theta_0\right)\right)x_i'u_i} + \sqrt{N}\overline{h^{-1}\left(x_i,\theta_0\right)x_i'u_i}$$

Le deuxième terme converge clairement en loi puisque $h^{-1}\left(x_i,\theta_0\right)x_i'u_i$ *a des moments d'ordre 1 et 2. On a d'ailleurs par un calcul immédiat* $V\left(h^{-1}\left(x_i,\theta_0\right)x_i'u_i\right)=\mathrm{E}(\frac{x_i'x_i}{h(x_i,\theta_0)})$. *On applique le théorème de la valeur moyenne* $\widehat{h}_i^{-1}-h^{-1}\left(x_i,\theta_0\right)=\nabla h^{-1}\left(\widetilde{\theta},\underline{x}_i\right)\left(\widehat{\theta}-\theta\right)$, *avec* $\left|\widetilde{\theta}-\theta\right| < \left|\widehat{\theta}-\theta\right|$ *On peut donc écrire* $\sqrt{N}\overline{\left(\widehat{h}_i^{-1}-h^{-1}\left(x_i,\theta_0\right)\right)x_i'u_i} = \overline{x_i'u_i\nabla h^{-1}\left(\widetilde{\theta},\underline{x}_i\right)}\sqrt{N}\left(\widehat{\theta}-\theta\right)$ *et* $\sqrt{N}\left(\widehat{\theta}-\theta\right)$ *est borné en probabilité et par l'hypothèse H7* $\overline{x_i'u_i\nabla h^{-1}\left(\widetilde{\theta},\underline{x}_i\right)}\xrightarrow{P}E\left(x_i'u_i\nabla h^{-1}\left(\widetilde{\theta}_0,\underline{x}_i\right)\right)=0$

Les deux derniers points se démontrent de la même façon que précédemment

7.3.1 Application

On considère le modèle en coupe

$$y_i = x_i b + u_i$$

dans lequel on spécifie la forme de l'hétérogénéité.

Cas : $\mathbf{E}(u_i\,|x_i\,) = \sum_{l,m\leq K+1} x_{li}x_{mi}\gamma_{lm}$

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 $\widehat{u}_i^2$ sur les variables $x_{li}x_{mi}$: $\widehat{u}_i^2 = \sum_{l,m\leq K+1} x_{li}x_{mi}\gamma_{lm} + w_i$

3. Construction d'un estimateur de $\widehat{\sigma}_i$ par $\widehat{\sigma}_i = \sqrt{\sum_{l,m\leq K+1} x_{li}x_{mi}\gamma_{lm}}$
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

On a vu les conditions sous lesquelles l'estimateur $\widehat{\gamma}$ converge bien vers la vraie valeur. Cette spécfication a néanmoins l'inconvénient de ne pas imposer la positivité de u_i^2. Bien qu'elle soit naturelle, on lui rpéfère souvent pour cette raison d'autres traitement de l'hétéroscédasticité en particculier avec des formes exponentielles.

Cas : $u_i = v_i \exp\left(\sum_{l,m\leq K+1} x_{li}x_{mi}\gamma_{lm}\right)$

On suppose de plus que v_i est indépendant de x_i avec $E(v_i) = 0$ et $V(v_i) = 1$. On a donc $E(u_i^2 | x_i) = \exp\left(2\sum_{l,m\leq K+1} x_{li}x_{mi}\gamma_{lm}\right)$. Cette forme est utile et souvent choisie car elle garantit que la variance conditionnelle est positive. Il faut estimer le paramètre γ. Ceci est fait à partir du logarithme des résidus des mco au carré. On a en effet $E(\ln(u_i^2) | x_i) = E(2\ln(|v_i|) | x_i) + 2\sum_{l,m\leq K+1} x_{li}x_{mi}\gamma_{lm}$. Les coefficients $\gamma_{l,m}$, excepté celui de la constante sont donc estimés de façon convergente à partir d'une régression de $\ln(u_i^2)$.

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 $\ln(\widehat{u}_i^2)$ sur les variables z_i : $\ln(\widehat{u}_i^2) = x_{li}x_{mi}\gamma_{lm} + w_i$.
3. Construction d'un estimateur de $\widehat{\sigma}_i$ par $\widehat{\sigma}_i = \exp z_i'\widehat{\theta}$
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

7.4 Exemple : estimation d'une équation de salaire

On illustre les résultats de ce chapitre en estimant une équation de salaire. Cette équation dite de Mincer relie le salaire (en logarithme) au niveau d'éducation et à l'expérience. Le niveau d'éducation est mesuré par le nombre d'année de scolarité, et l'expérience en nombre d'années écoulées depuis la fin des études. La spécification retenue est quadratique :

$$w_i = \alpha_0 + \alpha_s sco_i + \alpha_e \exp_i + \beta_e (\exp_i - 10)^2 + \alpha_h \hom me + u_i$$

le rendement de l'éducation est l'accroissement du salaire lié à l'augmentation d'une année de la scolarité : α_s. Le paramètre α_s représente donc le rendement de l'éducation

	bmco	s(bmco)	sw(bmco)	sw(bmco)/s(bmco)
Cste	4.11090	(0.02932)	(0.03587)	1.224
scolarité	0.06346	(0.00182)	(0.00218)	1.196
expérience	0.02568	(0.00078)	(0.00089)	1.144
expérience²	-0.00052	(0.00004)	(0.00004)	1.049
homme	0.15131	(0.00829)	(0.00832)	1.004

TAB. 7.1 – Estimateur des mco avec écart-types robustes et standards

au bout de 12 années d'étude. De même le rendement de l'expérience est estimé comme $\alpha_e + 2\beta_e(\exp_i - 20)$. Le coefficient α_e s'interprète donc comme le rendement de l'expérience à 20 ans, et le coefficients β_e reflète quant à lui la nature croissante ou non des rendements de l'expérience. L'équation est d'abord estimée par les mco. On calcule pour cette estimation les écarts-type de deux façons : d'abord avec la formule standard des mco $\widehat{V}_{as}(1) = \widehat{\sigma}^2 \overline{x_i' x_i}^{-1}$ et $\widehat{V}_b(1) = \widehat{V}_{as}(1)/N$ puis avec la formule robuste de White $\widehat{V}_{as}(2) = \overline{x_i' x_i}^{-1} \overline{\widehat{u}_i^2 x_i' x_i} \overline{x_i' x_i}^{-1}$ et $\widehat{V}_b(2) = \widehat{V}_{as}/N$. Les résultats sont présentés dans le tableau 7.1

La première colonne donne la valeur estimée du paramètre. La deuxième l'écart-type estimé par la formule ignorant l'hétéroscédasticité, la troisième colonne donne l'écart-type robuste calculé avec la matrice de White. Enfin la dernière colonne donne le ratio entre les deux écarts-type. Les résultats sont obtenus sur un échantillon de 6975 salariés dans le commerce en 2002. Les résultats montrent que le rendement de l'éducation est 6.3%. Une année d'éducation supplémentaire conduit donc à un accroissement du salaire de 6.2%. On observe que le rendement de l'expérience est décroissant avec l'age. Il est de 2.6% pour une année supplémentaire à 10 ans d'ancienneté et de 2.0% à 20 ans. Enfin on voit que les hommes sont payés 15% plus que les femmes. L'intérêt principal de ce tableau réside néanmoins dans les écarts-type estimés. On voit qu'en général les écarts-type tenant compte de l'hétéroscédasticité sont plus élevés et qu'en terme relatif les différences sont élevées. Ainsi pour le coefficient de la scolarité l'erreur est de 20%. On voit néanmoins que dans l'absolu les écarts-type ne sont pas fondamentalement différents. Ainsi pour la scolarité l'intervalle de confiance à 95% calculé avec le premier écart-type est de [5.98 , 6.71] alors qu'avec le second il est de [5.91 , 6.78].

Malgré cette faible différence, on peut faire un test d'hétéroscédasticité. Pour cela on régresse le résidu au carré sur les variables explicatives leurs carrés et leurs produits croisés : c'est à dire sur les treize variables explicatives $\widetilde{x}_i = 1$, sco_i, exp_i, exp_i^2, $Homme$, sco_i^2, $sco_i exp_i$, $sco_i exp_i^2$, $sco_i Homme$, exp_i^3, $exp_i Homme$, $exp_i^2 Homme$. On parvient au résultats reportés dans le tableau 7.2 pour cette régression.

	parametre	écart-type	student
Cste	0.8783	(0.1262)	6.96
scolarité	-0.1024	(0.0158)	-6.50
expérience	-0.0352	(0.0044)	-8.04
expérience²	0.0028	(0.0003)	8.21
homme	-0.0101	(0.0524)	-0.19
scolarité²	0.0028	(0.0005)	5.45
scolarité x expérience	0.0030	(0.0003)	10.03
scolarité x expérience²	-0.0001	(0.0000)	-5.95
scolarité x homme	0.0029	(0.0033)	0.88
expérience3	-0.0001	(0.0000)	-5.50
expérience x homme	-0.0018	(0.0014)	-1.29
expérience4	0.0000	(0.0000)	4.00
expérience² x homme	0.0001	(0.0001)	1.24
	R^2	F	
	0.0287605	187.51859	

TAB. 7.2 – Régression du carré du résidu sur les variables et leurs produits croisés

Le tableau donne le paramètre estimé ainsi que son écart-type. On voit que de nombreux coefficients sont significatifs : la scolarité, l'expérience, l'expérience au carré.... Le test d'hétéroscédasticité consiste à faire un test de nullité globale mis à part la constante. Ce test peut se faire à partir du R^2 de la régression en examinant la statistique $F = NR^2$. La statistique suit est un $\chi^2(12)$. Bien que le R^2 soit très faible, la statistique est très élevée et excède très largement la valeur seuil d'un test à 5% : 21.03. On rejette donc l'hypothèse de nullité globale. L'hypothèse d'homoscédasticité est ainsi très fortement rejetée.

Si on spécifie la forme de l'hétéroscédasticité, on peut mettre en oeuvre l'estimateur des mCQG. On spécifie comme cela est fait en général cette hétérogénéité sous la forme d'une exponentielle. On spécifie alors la perturbation comme

$$u_i = v_i \exp\left(\widetilde{x}_i \phi\right)$$

où $\widetilde{x}_i$ représente l'ensemble des variables explicatives, de leurs carrés et de leurs produits croisés. On fait l'hypothèse

$$v_i \perp x_i$$

Sous cette hypothèse

$$\ln\left(u_i^2\right) = \widetilde{x}_i \phi + \ln\left(v_i^2\right)$$

Le paramètre ϕ est estimé à la constante près à partir de la régression

$$E\left(\ln\left(u_i^2\right)|x_i\right) = \widetilde{x}_i \phi$$

	parametre	écart-type
Cste	-0.1030	(0.9749)
scolarité	-0.5734	(0.1216)
expérience	-0.2728	(0.0338)
expérience²	0.0220	(0.0026)
homme	0.0779	(0.4043)
scolarité²	0.0170	(0.0039)
scolarité x expérience	0.0235	(0.0023)
scolarité x expérience²	-0.0008	(0.0001)
scolarité x homme	0.0018	(0.0256)
expérience3	-0.0004	(0.0001)
expérience x homme	-0.0007	(0.0109)
expérience4	0.0000	(0.0000)
expérience² x homme	0.0000	(0.0005)
	257.72443	12

TAB. 7.3 – Régression du logarithme du carré du résidu sur les variables et leurs produits croisés

puisque $E\left(\ln\left(v_i^2\right)|x_i\right) = E\left(\ln\left(v_i^2\right)|x_i\right)$. Les résultats auxquels on parvient sont reportés dans le tableau 7.3.

On voit que là aussi de nombreux paramètres sont significatifs, et on pourrait comme précédemment faire un test d'hétéroscédasticité correspondant au test de la nullité globale des paramètres, à partir du R^2.de la régression. On parviendrait à la statistique de 255.30, plus élevée que la précédente mais conduisant à la même conclusion que l'on rejette fortement l'hypothèse d'homoscédasticité. Toutefois l'intérêt de cette régression est de récupérer la valeur prédite et d'en déduire une estimation de la variance conditionnelle. A partir de ces estimations on peut en effet calculer $\widehat{\sigma}^2(x_i) = \exp\left(\widetilde{x}_i\widehat{\phi}\right)$, et on sphéricise les données en divisant le modèle par $\exp\left(\widetilde{x}_i\widehat{\phi}/2\right)$. On considère ainsi $y_{isph} = y_i/\widehat{\sigma}(x_i)$ et $x_{isph} = x_i/\widehat{\sigma}(x_i)$, y compris la constante. Pour trouver l'estimateur des mCQG, on procède alors à la régression par les mco. Bien sur il est là aussi possible de calculer un estimateur robuste de la matrice de variance du paramètre exactement comme on le fait en l'absence de correction d'hétéroscédasticité. Normalement les écarts-type doivent être très proches, si la correction à retiré toute l'hétéroscédasticité du modèle. On parvient aux résultats reportés dans le tableau 7.4.

	bmcqg	s(bmcqg)	sw(bmcqg)	s(bmcqg)/sw(bmco)	sw(bmcqg)/sbmcqg)
Cste	4.26942	(0.03118)	(0.03152)	0.869	1.011
scolarité	0.05496	(0.00194)	(0.00197)	0.892	1.015
expérience	0.02275	(0.00080)	(0.00079)	0.899	0.988
expérience²	-0.00046	(0.00003)	(0.00004)	0.904	1.044
homme	0.14501	(0.00769)	(0.00781)	0.924	1.015

TAB. 7.4 – Estimateur des mcqg

On voit que les résultats sont un peu changés. On remarque en particulier une baisse du rendement de l'éducation qui passe de 6.3% à 5,5%. Cette différence faible est inquiétante car là encore les deux paramètres devraient être très proches et là il diffèrent plus que ce qu'implique l'ordre de grandeurs de la précision des estimations. Ceci n'est donc pas une bonne nouvelle en ce qui concerne la convergence des estimateurs. On voit néanmoins que les écarts-type sont modifiés. On vérifie bien la propriété des mCQG que les écarts-type correspondants sont plus petits que ceux des mco : le gain est ici de l'ordre de 10%. Toutefois compte tenu de la taille de l'échantillon, cela ne représente qu'un gain modeste en terme de largeur de l'intervalle de confiance. Les changements ne sont pas bouleversants. On observe par ailleurs une plus grande similitude entre les écarts-type du modèle sphéricisé robuste et directement obtenus que dans le cas précédent.

En conclusion de cet exemple, l'hétéroscédasticité est bien présente ici, mais les différentes façons de la prendre en compte soit dans le calcul des écarts-type, soit par la mise en oeuvre des mCQG, ne conduisent pas à des modifications considérables dans la précision des estimateurs et leur estimation. Là encore on se rend compte que la vraie question est plus l'existence de biais dans les estimations que celle de la possibilité de gains importants dans la précision des estimateurs. On verra par la suite que lorsque l'on aborde cette question, les estimateurs que l'on pourra mettre en oeuvre vont devenir beaucoup moins précis. Dans ce cas, la correction de l'hétéroscédasticité pourra représenter un gain appréciable de précision.

Chapitre 8

Autocorrélation des résidus dans les séries temporelles

Dans les modèles en série temporelles, 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 est donc dans un cadre dans lequel on ne peut plus faire l'hypothèse d'indépendance des observations.

On va voir à ce sujet :

- différentes formes d'autocorrélation,
- les tests permettant de détecter l'autocorrélation,
- les méthodes d'estimation adaptées en présence d'autocorrélation.

8.1 Différentes formes d'autocorrélation des perturbations

8.1.1 Processus stationnaires au premier et au second ordres

Un processus est une série temporelle (z_t). On dit qu'il est stationnaire au premier et au second ordre lorsque les moments d'ordre 1 $E(z_t) = \mu$ est indépendant de t, et $Cov(z_t, z_s) = \sigma_{t-s}$, ne dépend que du nombre de dates séparant les deux observations.

On ne considérera que des processus stationnaires au premier et au second ordre. On peut néanmoins citer quelques exemple de processus non stationnaires. Une variable trendée par exemple ne suit pas de processus stationnaire au premier ordre puisque $E(z_t) = a + bt$. Une marche aléatoire $z_t = z_{t-1} + \varepsilon_t$, avec ε_t, IID de moyenne nulle et de variance σ^2 constante est un processus stationnaire au premier ordre $E(z_t) = E(z_{t-1}) +$

$E(\varepsilon_t) = E(z_{t-1})$, mais pas au second ordre : $E(z_t^2) = E\left(z_{t-1}^2\right) + 2E(z_{t-1}\varepsilon_t) + E(\varepsilon_t^2) = E\left(z_{t-1}^2\right) + \sigma^2$. La variance n'est pas constante et on voit même qu'elle tend vers l'∞.

8.1.2 Perturbations suivant une moyenne mobile (MA)

Perturbations suivant une moyenne mobile d'ordre 1 (MA(1))

La perturbation u_t suit un processus de moyenne mobile d'ordre 1 noté $MA(1)$ si :

$$u_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

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

Les perturbations u_t ne sont plus IID, mais ces hypothèses sont transposées au processus ε_t.

On voit très facilement que le les processus MA(1) sont stationnaire à l'ordre 2. On a en effet $E(u_t | \underline{x}) = 0$, $V(u_t | \underline{x}) = \left(1 + \theta^2\right)\sigma_\varepsilon^2$, $E(u_t u_{t-1} | \underline{x}) = \theta\sigma_\varepsilon^2$ et $E(u_t u_{t-s} | \underline{x}) = 0$ pour $s > 1$. La matrice de variance covariance des perturbations a donc pour expression

$$V(u) = \sigma_\varepsilon^2 \begin{pmatrix} 1+\theta^2 & \theta & 0 & & 0 \\ \theta & 1+\theta^2 & \theta & \ddots & \\ 0 & \theta & \ddots & \ddots & 0 \\ & \ddots & \ddots & & \theta \\ 0 & & 0 & \theta & 1+\theta^2 \end{pmatrix}$$

Perturbations suivant une moyenne mobile d'ordre q (MA(q))

Ce cadre se généralise directement au cas d'un processus moyenne mobile d'ordre q. La perturbation u_t suit un processus de moyenne mobile d'ordre q noté $MA(q)$ si :

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

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

On voit là aussi très facilement que le les processus MA(q) sont stationnaires à l'ordre 2. On a en effet $E(u_t | X) = 0$, et en outre

$$V(u_t | \underline{x}) = \left(1 + \theta_1^2 + \cdots + \theta_q^2\right)\sigma_\varepsilon^2$$

Pour $s > q$, on a clairement $E(u_t u_{t-s} | \underline{x}) = 0$, par ailleurs pour $s \leq q$ on a

$$\begin{aligned} E(u_t u_{t-s} | \underline{x}) &= E\left(\left(\varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q}\right)\left(\varepsilon_{t-s} + \theta_1\varepsilon_{t-s-1} + \cdots + \theta_q\varepsilon_{t-s-q}\right)\right) \\ &= E\left(\left(\theta_s\varepsilon_{t-s} + \theta_{s+1}\varepsilon_{t-s-1} + \cdots + \theta_q\varepsilon_{t-q}\right)\left(\varepsilon_{t-s} + \theta_1\varepsilon_{t-s-1} + \cdots + \theta_{q-s}\varepsilon_{t-q}\right)\right) \\ &= \left(\theta_s + \theta_{s+1}\theta_1 + \cdots + \theta_q\theta_{q-s}\right)\sigma_\varepsilon^2 \end{aligned}$$

ne dépend que de s. Le processus est stationnaire à l'ordre 2.

Une généralisation de ce processus est le processus moyenne mobile ∞. Il est simplement défini

$$u_t = \varepsilon_t + \sum_{q=1}^{\infty} \theta_q \varepsilon_{t-q}$$

Proposition *Un processus moyenne mobile infini défini par*

$$u_t = \varepsilon_t + \sum_{q=1}^{\infty} \theta_q \varepsilon_{t-q}$$

est stationnaire dès que $\left(1 + \sum_{q=1}^{\infty} \theta_q^2\right) < \infty$

On voit directement que

$$V\left(u_t \left| \underline{x}\right.\right) = \left(1 + \sum_{q=1}^{\infty} \theta_q^2\right) \sigma_\varepsilon^2$$

est fini dès que la série θ_q^2 converge. Pour les covariances, on a aussi directement

$$E\left(u_t u_{t-s} \left| \underline{x}\right.\right) = \left(\theta_s + \sum_{q=1}^{\infty} \theta_{s+q}\theta_q\right) \sigma_\varepsilon^2$$

cette quantité ne dépend pas de t et est en outre finie dès lors que la série θ_q^2 converge, de par l'inégalité de Cauchy $\left|\sum_{q=1}^{\infty} a_q b_q\right|^2 \leq \sum_{q=1}^{\infty} a_q^2 \sum_{q=1}^{\infty} b_q^2$.

8.1.3 Perturbations suivant un processus autorégressif (AR)

Perturbations suivant un processus autorégressif d'ordre 1 (AR(1))

Un processus (AR1), est un processus dans lequel les perturbations sont engendrées par le processus :

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

avec :

- $E\left(\varepsilon_t \left| \underline{x}\right.\right) = 0$, $V\left(\varepsilon_t \left| \underline{x}\right.\right) = \sigma_\varepsilon^2$, $cov\ \left(\varepsilon_t, \varepsilon_{t'} \left| X\right.\right) = 0, \quad \forall\ t \neq t'$: les hypothèses d'homoscédasticité et d'indépendance des perturbations du modèle sont là aussi transférées aux ε_t c'est à dire aux innovations du processus :
- $|\rho| < 1$

On peut calculer la matrice de variance covariance d'un processus AR(1). On écrit facilement la façon dont la perturbation u_t dépend des perturbations passées

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

Proposition *Le processus AR(1) u_t est stationnaire si $E(u_0|X) = 0$ et $V(u_0|X) = \sigma_\varepsilon^2/(1-\rho^2)$ 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 : $E(u_t|X) = E(\varepsilon_t|X) + \cdots + \rho^{t-1}E(\varepsilon_1|X) + \rho^t E(u_0|X) = 0$

En outre u_t est non corrélé avec les perturbations futures. En effet, pour $t' > t$, $E(u_t\varepsilon_{t'}|X) = E(\varepsilon_{t'}(\varepsilon_t + \cdots + \rho^{t-1}\varepsilon_1 + \rho^t u_0)|X) = 0$, puisque $E(\varepsilon_{t'}\varepsilon_{t-l}|\underline{x}) = 0$, et $E(\varepsilon_{t'}u_0|\underline{x}) = 0$. Par ailleurs, $u_t = \varepsilon_t + \cdots + \rho^{(t-s-1)}\varepsilon_{s+1} + \rho^{t-s}u_s$, et donc compte tenu du résultat précédent $E(u_t u_s|\underline{x}) = E\left(\left(\varepsilon_t + \cdots + \rho^{(t-s-1)}\varepsilon_{s+1} + \rho^{t-s}u_s\right)u_s|\underline{x}\right) = \rho^{t-s}E(u_s^2|\underline{x})$. Enfin

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

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

$$\begin{aligned} V(u_t|\underline{x}) &= \sigma_\varepsilon^2/\left(1-\rho^2\right) \\ Cov(u_t, u_s) &= \rho^{t-s}\sigma_\varepsilon^2/\left(1-\rho^2\right) \end{aligned}$$

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

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

On a donc

$$V(u_t|\underline{x}) = \sum_{s=0}^{\infty}\rho^{2s}\sigma_\varepsilon^2 = \sigma_\varepsilon^2/\left(1-\rho^2\right)$$

La matrice de variance-covariance des perturbations à donc une expression très simple

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

Ce type de processus est fréquemment postulé, car il traduit l'idée simple et importante qu'un choc exogène à un moment donné a un effet persistant mais décroissant exponentiellement avec le temps. De par la simplicité de l'expression de la matrice de variance, ce

type de spécification permet en outre une mise en oeuvre facile de méthodes d'estimation plus efficaces que les MCO (telles les MCQG).

Perturbations suivant un processus AR(p)

La spécification précédente se généralise au cas où la perturbation u_t dépend des p perturbations précédentes. On note ce type de processus AR(p) si :

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

Ce que l'on représente par

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

avec $A(Z) = 1 - \rho_1 Z - \rho_2 Z^2 - \cdots - \rho_p Z^p$, $E(\varepsilon_t \,|\underline{x}) = 0$. On fait là encore l'hypothèse que $V(\varepsilon_t \,|X) = \sigma_\varepsilon^2$ et $cov(\varepsilon_t, \varepsilon_{t'} \,|X) = 0,\ \forall\, t \neq t'$

Proposition *Pour que le processus $AR(p)$ soit stationnaire à l'ordre 2 il faut que les racines du polynôme $A(X)$ soient de module supérieur à 1.*

Démonstration *On a en effet*

$$\begin{aligned} u_t &= \frac{\varepsilon_t}{A(L)} = \frac{\varepsilon_t}{1 - \rho_1 L - \rho_2 L^2 - \cdots - \rho_p L^p} \\ &= \frac{\varepsilon_t}{\prod_{s=1}^{p}(1 - r_s L)} = \left(\prod\nolimits_{s=1}^{p} \sum\nolimits_{k=0}^{\infty} r_s^k L^k\right) \varepsilon_t = \left(\sum\nolimits_{k=0}^{\infty} \eta_k L^k\right) \varepsilon_t \end{aligned}$$

où r_s est l'inverse de la $s^{ième}$ racine (éventuellement complexe) du polynôme $A(Z)$ et est donc de module strictement inférieur à 1. Le processus apparaît ainsi comme un processus moyenne mobile infini dont les coefficients sont directement déduit des racines r_s. Chacun des processus moyenne mobile $\sum_{k=0}^{\infty} r_s^k L^k$ est stationnaire puisque $|r_s| < 1$. En outre on montre facilement que si on considère deux $MA(\infty)$ $(\sum a_q L^q)$ et $(\sum b_q L^q)$ tels que $(\sum |a_q|) < \infty$ et $(\sum |b_q|) < \infty$ alors le produit de ces deux $MA(\infty)$ est un $MA(\infty)$ ayant la même propriété de sommabilité.

$$\left(\sum a_q L^q\right)\left(\sum b_q L^q\right) = \left(\sum \left(\sum_s b_s a_{q-s}\right) L^q\right)$$

et

$$\sum \left|\sum_s b_s a_{q-s}\right| \leq \sum \sum_s |b_s|\,|a_{q-s}| = \left(\sum |a_q|\right)\left(\sum |b_q|\right) < \infty$$

On en déduit que $(\sum_{k=0}^{\infty} |\eta_k|) < \infty$ et donc $\left(\sum_{k=0}^{\infty} |\eta_k|^2\right) < \infty$. Le processus est donc stationnaire.

L'expression de la matrice de variance covariance peut être néanmoins relativement complexe. Si on considère le cas d'un processus $AR(2)$ par exemple, on peut calculer

$$\begin{array}{l} Vu_t = \sigma_u^2 = \frac{1-\rho_2}{(1+\rho_2)\left[(1-\rho_2)^2-\rho_1^2\right]}\sigma_\varepsilon^2 = \Psi_0, \forall\ t \\ \text{cov}\ (u_t, u_{t-1}) = \frac{\rho_1}{1-\rho_2}\sigma_u^2 = \Psi_1 \\ \text{cov}\ (u_t, u_{t-2}) = \rho_2\sigma_u^2 + \frac{\rho_1^2}{1-\rho_2}\sigma_u^2 = \Psi_2 = \rho_2\Psi_0 + \rho_1\Psi_1 \\ \text{cov}\ (u_t, u_{t-s}) = \Psi_s = \rho_1\Psi_{s-1} + \rho_2\Psi_{s-2},\ s > 2 \end{array}$$

Ces formules illustrent la complexité de la forme de la matrice de variance covariance dans le cas AR(2) . On voit toutefois émerger une certaine régularité dans la détermination des covariances, qui se généralise au cas AR(p) . En effet pour un AR(p) : $u_t = \rho_1 u_{t-1} + \cdots + \rho_p u_{t-p} + \varepsilon_t$, pour des valeurs de s suffisamment élevée($\geq p$), on a

$$\begin{array}{rcl} E\left(u_t u_{t-s}\right) & = & \rho_1 E\left(u_{t-1}u_{t-s}\right) + \cdots + \rho_p E\left(u_{t-p}u_{t-s}\right) + E\left(\varepsilon_t u_{t-s}\right) \\ \gamma_s & = & \rho_1\gamma_{s-1} + \cdots + \rho_p\gamma_{s-p} \end{array}$$

pour $\gamma_s = E\left(u_t u_{t-s}\right)$. Cette équation est connue sous le nom d'équation Yule-Walker. Elle est aussi vraie pour les corrélations (c'est à dire la covariance divisée par la variance puisque le processus est stationnaire)

8.1.4 Perturbation suivant un processus ARMA(p,q)

Une dernière généralisation correspond à la situation combinant les deux processus précédents : on dit que la perturbation u_t suit un processus ARMA(p,q) si l'on peut écrire :

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

avec

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

et

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

On a le même résultat que le processus est stationnaire si les racines du polynôme $A\left(Z\right)$ sont à l'extérieur du cercle unité.

On examine le cas particulier d'un processus ARMA(1,1)

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

Par conséquent

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

Comme $E(u_t\varepsilon_t) = E(\varepsilon_t^2) = \sigma_\varepsilon^2$, 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

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

De même

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

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

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

8.2 Estimateur des MCO lorsque les perturbations suivent un AR(1)

On considère le cas d'un modèle

$$y_t = x_t\, b + u_t$$

dans lequel les perturbations suivent un processus AR(1) et sont indépendantes des variables explicatives. On a donc :

1. $E(\underline{u}\,|\underline{x}) = 0$
2. $V(\underline{u}\,|\underline{x}) = \Sigma$ de dimension $T \times T$ et on a vu que

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

3. $\frac{1}{T}\underline{x}'\underline{x} \xrightarrow{P} Q_{XX}$, et que $\underline{x}'\underline{x}$ et Q_X sont inversibles.

Cette hypothèse n'est pas systématiquement garantit en pratique. En particulier dans le cas de la présence d'un trend ou dans le cas de la présence de variables explicatives distribuées suivant une marche aléatoire les moments d'ordre 2 n'existent pas.

On fait enfin l'hypothèse que la matrice $\frac{1}{T}\underline{x}'\Sigma\underline{x} \xrightarrow{P} Q_{X\Sigma X}$

Sous ces hypothèses l'estimateur des mco

$$\widehat{b}_{mco} = (\underline{x}'\underline{x})^{-1}\underline{x}'\underline{y}$$

vérifie les propriétés suivantes :

1. $E\left(\widehat{b}_{mco}\,|\underline{x}\right) = b$: l'estimateur est sans biais
2. $V\left(\widehat{b}_{mco}\,|\underline{x}\right) = (\underline{x}'\underline{x})^{-1}\underline{x}'\Sigma\underline{x}\,(\underline{x}'\underline{x})^{-1}$
3. $\widehat{b}_{mco} \xrightarrow{P} b$: l'estimateur est convergent
4. $\sqrt{T}\left(\widehat{b}_{mco} - b\right) \xrightarrow{L} N(0, V_{as})$: l'estimateur est asymptotiquement normal.
5. $V_{as} = Q_{XX}^{-1}Q_{X\Sigma X}Q_{XX}^{-1} = p\lim TV\left(\widehat{b}_{mco}\,|\underline{x}\right)$
6. L'estimateur de la variance des résidus

$$\widehat{\sigma}^2 = \frac{1}{T}\Sigma_{t=1}^{T}\,\hat{u}_t^2$$

est convergent : $\widehat{\sigma}^2 \xrightarrow{P} \sigma^2$

7. L'estimateur du coefficient d'autocorrélation des résidus est convergent

$$\hat{\rho} = \frac{\Sigma_{t=2}^{T}\,\hat{u}_t\hat{u}_{t-1}}{\Sigma_{t=2}^{T}\,\hat{u}_{t-1}^2} \xrightarrow{P} \rho$$

8. $\sqrt{T}\,(\hat{\rho} - \rho) \xrightarrow{L} N(0, 1-\rho^2)$ il est asymptotiquement normal

On en déduit que

9. $\Sigma\left(\widehat{\rho}, \widehat{\sigma}^2\right) \xrightarrow{P} \Sigma\left(\rho, \sigma^2\right)$,
10. $\widehat{V}_{as}\left(\widehat{b}_{mco}\,|\underline{x}\right) = \left(\frac{\underline{x}'\underline{x}}{T}\right)^{-1}\frac{\underline{x}'\Sigma\left(\widehat{\rho},\widehat{\sigma}^2\right)\underline{x}}{T}\left(\frac{\underline{x}'\underline{x}}{T}\right)^{-1} \xrightarrow{P} Q_{XX}^{-1}Q_{X\Sigma X}Q_{XX}^{-1}$ On peut donc obtenir un estimateur convergent de la matrice de variance de l'estimateur.
11. $\sqrt{T}\widehat{V}_{as}\left(\widehat{b}_{mco}\,|\underline{x}\right)^{-1/2}\left(\widehat{b}_{mco} - b\right) \xrightarrow{L} N(0, I)$

Remarque *1. Les résultats ne sont pas fondamentalement changés par rapport à ceux du chapitre précédent : l'estimateur est convergent, asymptotiquement normal et on peut estimer de manière convergente sa matrice de variance.*

2. *La définition de l'estimateur du coefficient d'autocorrélation à une interprétation simple. 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

$$\hat{\rho} = \frac{\Sigma_{t=2}^{T} \hat{u}_t \hat{u}_{t-1}}{\Sigma_{t=2}^{T} \hat{u}_{t-1}^2}$$

si les résidus n'étaient pas estimés, on obtiendrait directement la loi asymptotique de l'estimateur en appliquant les résultats standards : $\sqrt{T}(\hat{\rho} - \rho) \xrightarrow{L} N(0, V)$ et $V = V(u_{t-1})^{-1} V(\widetilde{\varepsilon}_t)$. Comme $V(u_t) = \rho^2 V(u_{t-1}) + V(\widetilde{\varepsilon}_t)$ et $V(u_t) = V(u_{t-1})$, $V(u_{t-1})^{-1} V(\widetilde{\varepsilon}_t) = (1 - \rho^2)$

3. *On peut préciser l'allure de l'expression de la matrice $\frac{1}{T}\underline{x}'\Sigma\underline{x}$. Dans le cas d'une seule variable explicative, par exemple, on a*

$$\frac{1}{T}\underline{x}'\Sigma\underline{x} = \frac{\sigma_\varepsilon^2}{1-\rho^2}\left(\frac{\sum_t x_t^2}{T} + 2\sum\nolimits_s \rho^s \frac{\sum_t x_t x_{t-s}}{T}\right)$$

si le processus engendrant les x est stationnaire et de moyenne nulle, et que l'on définit γ_s comme $cov(x_t, x_{t-s}) = \gamma_s V(x_t)$, ce terme s'écrit

$$\frac{1}{T}\underline{x}'\Sigma\underline{x} \xrightarrow{P} \frac{\sigma_\varepsilon^2}{1-\rho^2} V(x_t)\left(1 + 2\sum\nolimits_s \rho^s \gamma_s\right) = V(u_t) V(x_t)\left(1 + 2\sum\nolimits_s \rho^s \gamma_s\right)$$

et la matrice de variance de l'estimateur est alors

$$V_{as}\left(\widehat{b}_{mco} | X\right) = \frac{V(u_t)}{V(x_t)}\left(1 + 2\sum\nolimits_s \rho^s \gamma_s\right)$$

l'erreur sur la matrice de variance est donc d'un facteur multiplicatif $(1 + 2\sum_s \rho^s \gamma_s)$. On voit qu'elle est d'autant plus importante que le coefficient d'autocorrélation est fort. Si $\rho = 0$ on voit que l'on retrouve la formule standard de la variance des mco (dans ce cas spécifique). On voit aussi que l'erreur est d'autant plus importante que les variables explicatives sont elles-mêmes corrélées dans le temps. A la limite si les γ_s sont nuls, il n'y a pas d'erreur.

4. *L'obtention de ces résultats repose sur des théorèmes de convergence étudiant la moyenne de variable dépendante dans le temps. On donne les deux principaux. On considère un processus stationnaire z_t dont la moyenne est $E(z_t) = m$, avec des covariances $E(z_t z_{t-k}) = \gamma_k$ définie pour k allant de $-\infty$ à $+\infty$. On fait l'hypothèse que ces covariances sont absolument sommables :*

$$\sum\nolimits_{-\infty}^{+\infty} |\gamma_k| < \infty$$

(a) $\overline{z_t} \xrightarrow{P} m$ *et* $\lim TE\left(\overline{z_t} - m\right)^2 \to \sum_{-\infty}^{+\infty} \gamma_k$

(b) si $z_t = m + \sum_s \phi_s \varepsilon_{t-s}$, *avec* $\sum_s |\phi_s| < \infty$ *et* ε_t *IID, alors* $\sqrt{T}\left(\overline{z_t} - m\right) \xrightarrow{L} N\left(0, \sum_{-\infty}^{+\infty} \gamma_k\right)$

Le résultat $\lim TE\left(\overline{z_t} - m\right)^2 \to \sum_{-\infty}^{+\infty} \gamma_k$ *présente le changement fondamental avec la situation du chapitre précédent. Dans le chapitre précédent on avait simplement* $NE\left(\overline{z_t}\right)^2 = \sigma^2$, *ici l'analogue de* γ_0. *La différence provient ici du fait qu'il est nécessaire de prendre en comte la corrélation entre les observations aux différentes dates. Le résultat n'a toutefois rien de très surprenant. Dans le cas d'une variable de moyenne nulle, on a :*

$$\begin{aligned} T\overline{z_t}^2 &= \frac{1}{T}\left(z_1 + \cdots + z_T\right)^2 \\ &= \frac{1}{T}\sum_{t=1}^{T} z_t^2 + 2\frac{1}{T}\sum_{t=2}^{T} z_t z_{t-1} + \cdots + 2\frac{1}{T}\sum_{t=T}^{T} z_t z_{t-T+1} \\ &= \frac{1}{T}\sum_{t=1}^{T} z_t^2 + 2\frac{T-1}{T}\frac{1}{T-1}\sum_{t=2}^{T} z_t z_{t-1} + \cdots + 2\frac{1}{T}\frac{T}{T-1}\sum_{t=T}^{T} z_t z_{t-T+1} \end{aligned}$$

et donc

$$\begin{aligned} TE\left(\overline{z_t}^2\right) &= \gamma_0 + 2\frac{T-1}{T}\gamma_1 + \cdots + 2\frac{1}{T}\gamma_{T-1} \\ TE\left(\overline{z_t}^2\right) - \left(\gamma_0 + 2\gamma_1 + \cdots + 2\gamma_{T-1}\right) &= -2(\frac{1}{T}\gamma_1 + \cdots + \frac{T-1}{T}\gamma_{T-1}) \end{aligned}$$

qui tend vers zéro puisque la série $\sum_{-\infty}^{+\infty} |\gamma_k|$ *converge.*
Comme on le voit pour que la loi des grands nombres soit satisfaite il faut que la dépendance temporelle s'estompe suffisamment rapidement. On voit aussi que la variance de l'estimateur de la moyenne prend en compte cette dépendance temporelle. Plus la dépendance temporelle est forte moins les estimations sont précises.

8.3 L'estimateur de Newey-West de la matrice de variance de $\widehat{b}_{mco}$

Les résultats précédents sur la convergence de l'estimateur et l'estimation de sa matrice de variance pourraient être obtenus dans de très nombreuses situations, c'est à dire pour différentes spécifications du processus engendrant les perturbations. La propriété de sans biais, de convergence et de normalité asymptotique ne va pas être fondamentalement remise en cause. L'expression de la matrice de variance de l'estimateur dépend en revanche de la spécification du processus, car dans chacune des spécifications envisageables la matrice de variance covariance des perturbations est différente. Dans toutes ces spécifications toutefois, la matrice de variance des perturbations dépend d'un nombre limité

de paramètres, et ces paramètres pourraient être estimés à partir des résidus de l'estimation ; comme on l'a montré pour le coefficient de corrélation des perturbations. Il est donc possible en théorie d'obtenir une estimation convergente des la matrice Σ, à partir de laquelle on peut estimer la matrice de variance de l'estimateur des mco. Maintenant il est clair que cette matrice va dépendre de l'hypothèse choisie pour des raisons parfois incomplètement explicitées. On peut donc être tenté de rechercher un estimateur de la matrice de variance covariance de l'estimateur des mco qui soit robuste à ce choix plus ou moins arbitraire d'une spécification du processus engendrant les perturbations. En outre dans l'approche précédente, on fait l'hypothèse que la corrélation entre les résidus à différentes dates ne dépend pas des valeurs prises par les variables explicatives. On a pourtant mis l'accent dans le chapitre précédent sur les possibilités de dépendance des moments d'ordre 2 et des variables explicatives. Une telle question se pose pareillement dans le cadre des séries temporelles. Le point important concerne la variance du produit $\frac{1}{\sqrt{T}}\underline{x}'\underline{u} = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t'u_t$. La variance de ce terme s'écrit

$$\begin{aligned}
E\left(\underline{x}'\underline{u}\underline{u}'\underline{x}\right)/T &= E\left(\sum\nolimits_{t=1}^{T} x_t'x_tu_t^2/T + \sum\nolimits_{t,s\neq 0} x_t'x_{t-s}u_tu_{t-s}/T + x_{t-s}'x_tu_{t-s}u_t/T\right) \\
&= E\left(\sum\nolimits_{t=1}^{T} x_t'x_tu_t^2/T\right) + \\
&\quad E\left(\sum\nolimits_{t=2}^{T} x_t'x_{t-1}u_tu_{t-1}/T + x_{t-1}'x_tu_{t-1}u_t/T\right) + \\
&\quad E\left(\sum\nolimits_{t=3}^{T} x_t'x_{t-2}u_tu_{t-2}/T + x_{t-2}'x_tu_{t-2}u_t/T\right) + \cdots + \\
&\quad E\left(\sum\nolimits_{t=q}^{T} x_t'x_{t-q+1}u_tu_{t-q+1}/T + x_{t-q+1}'x_tu_{t-q+1}u_t/T\right) + \cdots + \\
&\quad E\left(\sum\nolimits_{t=T}^{T} x_T'x_1u_Tu_1/T + x_1'x_Tu_1u_T/T\right)
\end{aligned}$$

soit $E\left(x_t'x_tu_t^2\right) + \sum_{s\neq 0}\left(E\left(x_t'x_{t-s}u_tu_{t-s}\right) + E\left(x_{t-s}'x_tu_{t-s}u_t\right)\right)(T-s+1)/T$. Pour un s donné, $\sum_t x_t'x_{t-s}u_tu_{t-s}/T$ est un estimateur convergent de $E\left(x_t'x_{t-s}u_tu_{t-s}\right)(T-s+1)/T$. Le problème est qu'il faut estimer cette quantité pour toutes les valeurs de s de $s=1$ jusqu'à $s=T$, ce qui est impossible dans un échantillon de taille T. L'optique choisie par Newey-West est de n'estimer ces termes que pour les valeurs de s les plus faibles, le nombre de valeurs retenues dépendant de la taille de l'échantillon. Ceci est exact si la série x_tu_t est distribuée suivant une moyenne mobile d'ordre fini. C'est une approximation sinon, mais si le degrés de corrélation temporelle de x_tu_t décroît assez vite et si l'estimateur retenu intègre un nombre de retard croissant avec la taille de l'échantillon on peut montrer que cette matrice est convergente. Ceci est conforme à l'idée que les corrélations entre les perturbations disparaissent à un taux relativement élevé. Par exemple dans le cadre du modèle $AR(1)$ elles disparaissent exponentiellement. L'estimateur de Newey West estime $E\left(\underline{x}'\underline{u}\underline{u}'\underline{x}\right)/T$ par

$$\sum\nolimits_{t=1}^{T} x_t'x_tu_t^2/T + \sum\nolimits_{s\neq 0}\pi_s(T)\sum\nolimits_t\left(x_t'x_{t-s}u_tu_{t-s} + x_{t-s}'x_tu_{t-s}u_t\right)/T$$

avec $\pi_s(T)$ décroissant avec s et croissant avec T. Le poids proposé par Newey-West est linéaire en s, de la forme $\pi_s(T) = (1 - s/(q(T)+1))\,1(s \leq q(T))$. On fait bien sur croître $q(T)$ vers l'infini lorsque T augmente, mais à un rythme beaucoup plus faible que T. On montre que sous des hypothèses de régularité satisfaisante cet estimateur converge vers $E(\underline{x}'\underline{u}\underline{u}'\underline{x})/T$. Au total l'estimateur de la matrice de variance covariance robuste à l'hétéroscédasticité temporelles et liée aux variables explicatives est

$$\widehat{V}_{as}\left(\widehat{b}_{mco}\right) = \left(\frac{\underline{x}'\underline{x}}{T}\right)^{-1}\left(\widehat{\Gamma}_0 + \sum_{s=1}^{q(T)}\left(1 - \frac{s}{q(T)+1}\right)\left(\widehat{\Gamma}'_s + \widehat{\Gamma}_s\right)\right)\left(\frac{\underline{x}'\underline{x}}{T}\right)^{-1}$$

où

$$\begin{aligned}\widehat{\Gamma}_0 &= \sum_{t=1}^{T}\frac{x'_t x_t \widehat{u}_t^2}{T}\\ \widehat{\Gamma}_s &= \sum_{t=s+1}^{T}\frac{x'_t x_{t-s}\widehat{u}_t\widehat{u}_{t-s}}{T}\end{aligned}$$

On rappelle encore que cette matrice est robuste à la fois à la corrélation temporelle des résidus, pourvu qu'elle s'estompe assez vite et à l'existence d'hétéroscédasticité relative aux x. On vérifie bien au passage que si on fait l'hypothèse qu'il n'y a pas de corrélation temporelle dans les perturbations ou les variables explicatives, alors on retrouve la formule de White (dans ce cas on n'a en effet que le terme Γ_0 dans le terme central).

8.4 Les MCQG dans le modèle $AR(1)$: l'estimateur de Prais-Watson.

On sait que sous les hypothèses énoncées :

1. $E(\underline{u}\,|\underline{x}) = 0$,
2. $V(\underline{u}\,|\underline{x}) = \Sigma$ de dimension $T \times T$ inversible
3. $\underline{x}'\underline{x}$ inversible,

l'estimateur des MCO n'est pas l'estimateur optimal. Le meilleur estimateur linéaire sans biais de b est l'estimateur des MCG :

$$\hat{b}_{mcg} = (\underline{x}'\,\Sigma^{-1}\,\underline{x})^{-1}\,\underline{x}'\,\Sigma^{-1}\,\underline{y}$$

dont la variance est donnée par :

$$V\left(\hat{b}_{mcg}\right) = (\underline{x}'\,\Sigma^{-1}\,\underline{x})^{-1}$$

Il peut être obtenu comme estimateur des mco dans le modèle :

$$\Sigma^{-1/2}\,\underline{y} = \Sigma^{-1/2}\,\underline{x}b + \Sigma^{-1/2}\,\underline{u}$$

où $\Sigma^{-1/2}\Sigma\Sigma^{-1/2'} = I_T$. La pre-multilplication du modèle par $\Sigma^{-1/2}$ porte on le rappelle le nom de sphéricisation, ceci parce qu'elle rend les perturbations

indépendantes. 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 appliqué au modèle :

$$\begin{pmatrix} y_1\sqrt{1-\rho^2} \\ y_2 - \rho\, y_1 \\ \vdots \\ y_T - \rho\, y_{T-1} \end{pmatrix} = \begin{pmatrix} x_1\sqrt{1-\rho^2} \\ x_2 - \rho x_1 \\ \vdots \\ x_T - \rho\, x_{T-1} \end{pmatrix} b + \begin{pmatrix} u_1\,\sqrt{1-\rho^2} \\ u_2 - \rho\, u_1 \\ \vdots \\ u_T - \rho\, u_{T-1} \end{pmatrix}$$

Dans d'autre cas, si par exemple les perturbations sont définies suivant un $AR\,(p)$, ou un $MA\,(q)$, on aurait d'autres formules beaucoup plus compliquées, faisant intervenir les p ou q paramètres de la matrices de variance. Néanmoins dans le cas AR(1) comme dans les autres, pour calculer l'estimateur MCG, il faut connaître ρ. Comme celui-ci est inconnu, on utilise l'estimateur des moindres carrés quasi généralisés (mCQG). Le principe de cet estimateur est de remplacer les paramètres inconnus, en nombre fini, par des estimateurs convergents dans l'étape de sphéricisation. Dans le cas AR(1), il faut ainsi remplacer ρ dans la prémultiplication du modèle par $\Sigma^{-1/2}\,(\rho)$ par $\widehat{\rho}$, et donc multiplier le modèle par $\Sigma^{-1/2}\,(\widehat{\rho})$. Comme on l'a vu on dispose à partir de la mise en oeuvre de l'estimateur des mco d'un estimateur convergent de ce coefficient à partir des résidus estimés.

Sous les hypothèses :

- $E\,(\underline{u}\,|\underline{x}) = 0$
- $V\,(\underline{u}\,|\underline{x}) = \Sigma\,(\theta)$ de dimension $T \times T$, θ de dimension finie
- $\frac{1}{T}\underline{x}'\underline{x} \xrightarrow{P} Q_{XX}$, $\underline{x}'\underline{x}$ et Q_X inversibles
- $\frac{1}{T}\underline{x}'\Sigma^{-1}\underline{x} \xrightarrow{P} Q_{X\Sigma^{-1}X}$ inversible
- $\widehat{\theta} \xrightarrow{P} \theta$ on dispose d'un estimateur convergent de θ

L'estimateur des MCQG

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

vérifie

- $\widehat{b}_{mcqg} \xrightarrow{P} b$: convergence
- $\sqrt{T}\left(\widehat{b}_{mcqg} - b\right) \xrightarrow{L} N\left(0, V_{as}(mcqg)\right)$: normalité asymptotique
- $V_{as}\left(mcqg\right) = Q^{-1}_{X\Sigma^{-1}X} = p\lim TV\left(mcg\right)$ équivalence entre mCQG et MCG
- $\widehat{V}_{as}\left(mcqg\right) = \left(\frac{1}{T}\underline{x}'\Sigma\left(\widehat{\theta}\right)^{-1}\underline{x}\right)^{-1} \xrightarrow{P} V_{as}\left(mcqg\right)$ estimation de la matrice de variance

L'estimateur de Prais-Watson, est l'estimateur des mCQG dans le modèle AR(1) . Il est obtenu en plusieurs étapes :

1. estimation par MCO du modèle $y_t = x_t b + u_t, t = 1, ..., T$
2. calcul des résidus estimés : $\hat{u}_t = y_t - x_t\hat{b}_{mco}$
3. estimation de ρ par application des mco au modèle :

$$\hat{u}_t = \rho\hat{u}_{t-1} + \varepsilon_t, \quad t = 2, ..., T$$

soit

$$= \frac{\Sigma_{t=2}^{T}\ \hat{u}_t\hat{u}_{t-1}}{\Sigma_{t=2}^{T}\ \hat{u}_{t-1}^2}$$

On calcule alors les données transformées :

$$\begin{aligned}\tilde{y}_1 &= \sqrt{1-\hat{\rho}^2}\ y_1\ et\ \tilde{y}_t = y_t - \hat{\rho}y_{t-1}, \quad t = 2, ..., T\\ \tilde{x}_1 &= \sqrt{1-\hat{\rho}^2}\ x_1\ et\ \tilde{x}_t = x_t - \hat{\rho}x_{t-1}, \quad t = 2, ..., T\end{aligned}$$

et on estime par les MCO sur ce modèle :

$$\tilde{y}_t = \tilde{x}_t b + \tilde{u}_t, \quad t = 1, ..., T$$

L'estimateur $\hat{b}$ ainsi obtenu est convergent et asymptotiquement aussi efficace que l'estimateur des MCG. Les écarts-type donnés par les logiciels standards peuvent en outre être directement utilisés (Remarque : il ne faut pas oublier de retirer la constante du modèle et ne pas omettre non plus d'appliquer la transformation à toutes les variables du modèle initial, y compris la constante si il en comprend une).

8.5 Détection de l'autocorrélation

8.5.1 Un test asymptotique

On se place dans le cadre du modèle AR(1) : $u_t = \rho\ u_{t-1} + \varepsilon_t$. On souhaite tester l'absence d'autocorrélation c'est à dire tester : $H_0 : \rho = 0$ contre $H_1 : \rho \neq 0$. Si on s'en tient aux résultats précédemment énoncés, on peut estimer le modèle par les mco, récupérer alors les résidus et estimer le coefficient d'auto corrélation comme on l'a vu. On a asymptotiquement $\sqrt{T}(\hat{\rho} - \rho) \xrightarrow{L} N(0, 1-\rho^2)$. Donc sous H_0, on a $\sqrt{T}\hat{\rho} \xrightarrow{L} N(0,1)$. On peut donc former la statistique de test $S = \sqrt{T}\hat{\rho}$, et définir la région critique $W = \left\{S \,\middle|\, |S| > t_{1-\alpha/2}\right\}$. Ce test asymptotique est convergent au seuil α.

8.5.2 Le test de Durbin et Watson

Néanmoins on se trouve parfois dans des échantillons de petite taille dans lesquels l'approximation asymptotique ne vaut pas parfaitement. C'est pourquoi on utilise très fréquemment, souvent par inertie le test dit de Durbin-Watson et qui repose sur la statistique :

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

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

$$p\lim \widehat{d} = 2(1-\rho)$$

En effet :

$$\begin{aligned} p\lim \widehat{d} &= p\lim \frac{\frac{1}{T}\Sigma_{t=2}^{T}\hat{u}_t^2 - 2\frac{1}{T}\Sigma_{t=2}^{T}\hat{u}_t\hat{u}_{t-1} + \frac{1}{T}\Sigma_{t=2}^{T}\hat{u}_{t-1}^2}{\frac{1}{T}\Sigma_{t=1}^{T}\hat{u}_t^2} \\ &= 1 - 2\rho + 1 = 2(1-\rho) \end{aligned}$$

puisque

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

et que

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

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 toutefois difficile à établir car elle dépend des résidus estimés et donc des valeurs prises par les variables explicatives du modèle. On montre néanmoins que :

Sous l'hypothèse $H_0 : \rho = 0$, il existe deux statistiques, d_l 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, le nombre de variables explicatives.

Test de $H_0 : \rho = 0$ contre $H_1 : \rho > 0$

Si $\widehat{d}$ est proche de 2 on accepte l'hypothèse et si $\widehat{d}$ est 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

$$\begin{aligned} P\left(d_l < d_l^\star(\alpha)\right) &= \alpha \\ P\left(d_u < d_u^\star(\alpha)\right) &= \alpha \end{aligned}$$

Comme

$$d_l < d_0 < d_u$$

On a

$$d_l^\star(\alpha) < d^\star(\alpha) < d_u^\star(\alpha)$$

La règle de décision est alors la suivante :

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'autocorrélation. 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. Lorsque le nombre d'observation devient important, on se trouve dans la situation asymptotique et on peut utiliser l'approche précédemment évoquée.

Test de $H_0 : \rho = 0$ **contre** $H_1 : \rho < 0$

La statistique de test à utiliser est $4 - \widehat{d}$, et il faut à nouveau la comparer à 2 : on rejettera l'hypothèse pour des valeurs faibles de la statistique et on l'acceptera si elle prend des valeurs suffisamment importantes. On a en effet dans ce cas :

$$4 - d_u^\star < 4 - d^\star < 4 - d_\ell^\star$$

Par conséquent la règle de décision est donnée par :
- si $4 - \widehat{d} > 4 - d_\ell^\star$, alors $4 - \widehat{d} > 4 - d^\star$: on refuse H_0
- si $4 - \widehat{d} < 4 - d_u^\star$, alors $4 - \widehat{d} < 4 - d^\star$: on accepte H_0
- 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.

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 ;
   exemples : MA(1) ou AR(2). Dans les autres situations, il est préférable de recourir à d'autres tests.

8.6 Résumé

Dans ce chapitre, on a étudié

1. Les différentes formes de corrélations des perturbations
2. Présenté les modèles $AR(p)$ et $MA(q)$ et mis l'accent sur le modèle $AR(1)$ qui modélise simplement une idée simple et importante : les innovations d'un processus peuvent avoir des effets durables mais qui s'estompe progressivement.
3. Examiné les propriétés de convergence de l'estimateur des mco dans le cas $AR(1)$ et étudié en quoi elle diffère du cadre IID.
4. On retrouve le résultat central que la corrélation des résidus n'affecte pas les propriétés de convergence de l'estimateur mais modifie en revanche les écarts-type des estimations.
5. On a proposé une matrice de variance robuste à l'hétéroscédasticité temporelle et relative au x, la matrice de Newey-West, qui généralise au cadre des séries temporelles la matrice de White robuste à l'hétéroscédasticité relative aux x seulement.
6. On a examiné l'estimateur des MCQG dans le cadre du modèle $AR(1)$, estimateur dit de Prais-Watson, simplement mis en oeuvre en deux étapes. une étape mco

permettant de calculer le coefficient de corrélation ρ, une étape mco sur le modèle sphéricisé, cette étape étant particulièrement simple dans le cas $AR(1)$.

7. On a enfin examiné les tests d'auto-corrélation et présenté le test très connu de Durbin -Watson.

Chapitre 9

L'estimateur des MCQG dans le cas où $\Omega = I_N \otimes \Sigma(\theta)$

On examine ici le cas des données de panel et le cas des régressions empilées. On considère d'une façon générale le modèle

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

Le modèle est ici spécifié en terme de vecteur $\underline{y}_i$, $\underline{x}_i$ et $\underline{u}_i$. Comme on va le voir ce cas est en fait une généralisation directe du cas des mco précédemment examiné.

Estimateur des MCO

On montre d'abord comment les résultats obtenus pour l'estimateur des moindres carrés ordinaires se généralisent au cas considéré.

On fait les hypothèses

- $H0$ Les observations $(\underline{y}_i, \underline{x}_i) \in \mathbb{R} \times \mathbb{R}^{K+1}$, $i = 1, ..., N$, sont IID
- $H1$ $E(\underline{u}_i | \underline{x}_i) = 0$
- $H2 V(\underline{u}_i | \underline{x}_i) = V(\underline{u}_i) = \Sigma(\theta)$. Σ est ici une matrice de dim $M \times M$, θ est alors nécessairement un paramètre de dimension finie, de taille au plus égale à $M(M+1)/2$
- $H3$ $H4$ $\forall N$ $\underline{x}'\underline{x}$ et $E(\underline{x}_i'\underline{x}_i)$ sont inversibles
- $H5$ Les moments de $|x_{ki}x_{li}|$ et de $|u_{ti}u_{si}|$ existent.

Proposition *Sous les hypothèses H0 à H6, l'estimateur des MCO*

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

vérifie quand $N \to \infty$

1. $\widehat{b}_{mco} \xrightarrow{P} b$, *l'estimateur est convergent*
2. $\sqrt{N}\left(\widehat{b}_{mco} - b\right) \xrightarrow{L} \mathcal{N}\left(0, \mathrm{V}_{as}\left(\widehat{b}_{mco}\right)\right)$, *l'estimateur est asymptotiquement normal*

3. $V_{as}\left(\widehat{b}_{mco}\right) = \left[E(\underline{x}_i'\underline{x}_i)\right]^{-1} E(\underline{x}_i'\Sigma\underline{x}_i)\left[E(\underline{x}_i'\underline{x}_i)\right]^{-1}$
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 *la matrice de variance des perturbations*
5. $\widehat{V}_{as}\left(\widehat{b}_{mco}\right) = \overline{(\underline{x}_i'\underline{x}_i)}^{-1}\overline{\underline{x}_i'\widehat{\Sigma}\underline{x}_i}\overline{\underline{x}_i'\underline{x}_i}^{-1} \xrightarrow{P} V_{as}\left(\widehat{b}_{mco}\right)$ *Estimation de* V_{as}
6. $\sqrt{N}\widehat{V}_{as}\left(\widehat{b}_{mco}\right)^{-1/2}\left(\widehat{b}_{mco} - b\right) \xrightarrow{L} \mathcal{N}(0, I)$

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

Convergence *Pour montrer la convergence on écrit* $\widehat{b}_{mco} = b + \left(\overline{\underline{x}_i'\underline{x}_i}\right)^{-1}\overline{\underline{x}_i'\underline{u}_i}$. *Comme les observations sont indépendantes et équidistribuées entre deux individus* i *et* j *et que les moments* $|x_{ki}x_{li}|$ *existent* $\overline{\underline{x}_i'\underline{x}_i} \xrightarrow{P} E(\underline{x}_i'\underline{x}_i)$. *Comme dans le cas standard, les moments d'ordre 1 et 2 de* $\underline{x}_i'\underline{u}_i$ *existent. On a en effet* $E(\underline{x}_i'\underline{u}_i) = E(\underline{x}_i'E(\underline{u}_i|\underline{x}_i)) = 0$ *et* $V(\underline{x}_i'\underline{u}_i) = E(\underline{x}_i'V(\underline{u}_i|\underline{x}_i)\underline{x}_i') + V(\underline{x}_i'E(\underline{u}_i|\underline{x}_i)) = E(\underline{x}_i'\Sigma\underline{x}_i')$. *On a donc* $\left(\overline{\underline{x}_i'\underline{x}_i}\right)^{-1}\overline{\underline{x}_i'\underline{u}_i} \xrightarrow{P} E(\underline{x}_i'\underline{x}_i)^{-1}E(\underline{x}_i'\underline{u}_i) = 0$ *par application de la loi faible des grands nombres.*

Normalité asymptotique $\sqrt{N}\left(\widehat{b}_{mco} - b\right) = \left(\overline{\underline{x}_i'\underline{x}_i}\right)^{-1}\sqrt{N}\overline{\underline{x}_i'\underline{u}_i}$

On applique le Théorème central limite à $\underline{x}_i'\underline{u}_i$. *On a déjà vu que les deux premiers moments de ce vecteur existent . On a donc* $\sqrt{N}\overline{\underline{x}_i'\underline{u}_i} \xrightarrow{L} N(0, E(\underline{x}_i'\Sigma\underline{x}_i))$. *On applique alors le théorème de Slutsky* $\left(\overline{\underline{x}_i'\underline{x}_i}\right)^{-1} \xrightarrow{P} E(\underline{x}_i'\underline{x}_i)^{-1}$ *et* $\sqrt{N}\overline{\underline{x}_i'\underline{u}_i} \xrightarrow{L} N(0, E(\underline{x}_i'\Sigma\underline{x}_i))$ *donc*

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

Estimation de Σ

L'estimateur de Σ *est* $\widehat{\Sigma} = \overline{\left(\underline{y}_i - \underline{x}_i\widehat{b}_{mco}\right)\left(\underline{y}_i - \underline{x}_i\widehat{b}_{mco}\right)'} = \overline{\widehat{\underline{u}}_i\widehat{\underline{u}}_i'}$ *et* $\widehat{\underline{u}}_i = \underline{y}_i - \underline{x}_i\widehat{b}_{mco} = \underline{x}_i\left(b - \widehat{b}_{mco}\right) + \underline{u}_i$. *Donc*

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

Le premier terme converge vers Σ par la loi des grands nombres puisque $|u_{si}u_{ti}|$ existent.

Le deuxième terme est une matrice dont les éléments sont somme de termes $\overline{x_{li}^k \left(b-\widehat{b}_{mco}\right)_m \left(b-\widehat{b}_{mco}\right)_{m'} x_{l'i}^{k'}}$ *Comme* $\left(b-\widehat{b}_{mco}\right) \xrightarrow{P} 0$ *et que* $\overline{x_{li}^k x_{l'i}^{k'}} \xrightarrow{P} E\left(x_{li}^k x_{l'i}^{k'}\right)$ *ce terme tend vers zero en probabilité.*

De même pour le troisième et le quatrième terme.

Estimation de la variance de l'estimateur des mco $V\left(\widehat{b}_{mco}\right) = \overline{(\underline{x}_i'\underline{x}_i)}^{-1}\overline{\underline{x}_i'\widehat{\Sigma}\underline{x}_i}\,\overline{\underline{x}_i'\underline{x}_i}^{-1} \xrightarrow{P} V\left(\widehat{b}_{mco}\right)$

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 zéro en probabilité par la loi forte des grands nombres. Le premier terme tend vers zéro en probabilité par le même genre d'argument que précédemment, puisque $\widehat{\Sigma} \xrightarrow{P} \Sigma$

Enfin, comme $\widehat{V}\left(\widehat{b}_{mco}\right) \xrightarrow{P} V\left(\widehat{b}_{mco}\right)$ *et* $\sqrt{N}\left(\widehat{b}_{mco}-b\right) \xrightarrow{L} \mathcal{N}\left(0, V\left(\widehat{b}_{mco}\right)\right)$ *on a directement par le théorème de Slutsky*

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

Remarque *Là encore on peut étendre les résultats au cas où bien que les hypothèses $H1$ à $H5$ soient satisfaites (en particulier identité des moments d'ordre 2, les observations ne sont pas équidistribuées. Ceci correspondrait par exemple au cas dans lequel les moments d'ordre supérieur à deux soient spécifiques à chaque individu. Il faut comme dans le cas des MCO du modèle homoscédastique imposer des restrictions sur les moments d'ordre 3 de la valeur absolue de chaque composante du résidu.*

Estimateur des MCQG

On s'intéresse maintenant à l'estimateur des MCQG. On introduit une hypothèse supplémentaire :

$H6$ $\exists\, \widehat{\theta} \xrightarrow{P} \theta$,

Cette hypothèse n'en est pas vraiment une si on lui adjoint les hypothèses précédentes puisqu'on a vu qu'alors on pouvait construire un estimateur convergent de la matrice de variance. On peut alors a fortiori obtenir un estimateur convergent du paramètre sous jacent θ.

Proposition *Sous les hypothèses H0 à H6, l'estimateur des MCQG*

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

vérifie quand $N \to \infty$

1. $\widehat{b}_{mcqg} \xrightarrow{P} b$, *l'estimateur est convergent*
2. $\sqrt{N}\left(\widehat{b}_{mcqg} - b\right) \xrightarrow{L} \mathcal{N}\left(0, V_{as}\left(\widehat{b}_{mcqg}\right)\right)$, *l'estimateur est asymptotiquement normal*
3. $V_{as}\left(\widehat{b}_{mcqg}\right) = [E(\underline{x}_i' \Sigma^{-1} \underline{x}_i)]^{-1} = V\left(\widehat{b}_{mcg}\right)$ *l'estimateur est asymptotiquement équivalent à l'estimateur des MCG*
4. $\widehat{V}_{as}\left(\widehat{b}_{mcqg}\right) = \overline{\underline{x}_i' \widehat{\Sigma}^{-1} \underline{x}_i}^{-1} \xrightarrow{P} V\left(\widehat{b}_{mcg}\right)$ *Estimation de la matrice de variance*
5. $\sqrt{N}\widehat{V}_{as}\left(\widehat{b}_{mcqg}\right)^{-1/2}\left(\widehat{b}_{mcqg} - b\right) \xrightarrow{L} \mathcal{N}(0, I)$

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

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* $\overline{x_{li}^k \widehat{\Sigma}_{m,m'}^{-1} x_{l'i}^{k'}} = \widehat{\Sigma}_{m,m'}^{-1} \overline{x_{li}^k x_{l'i}^{k'}}$ *qui convergent tous vers* $\widehat{\Sigma}_{m,m'}^{-1} \overline{x_{li}^k x_{l'i}^{k'}} \xrightarrow{P} \Sigma_{m,m'}^{-1} E\left(x_{li}^k x_{l'i}^{k'}\right)$ *qui 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

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. Comme on l'a rappelé au début du chapitre 5, une suite variables aléatoires convergent en loi est borné en probabilité, c'est un* $O(1)$, *et on a vu aussi au début du chapitre 5 que* $o(1) O(1) = o(1)$. *Le comportement asymptotique de* $\sqrt{N}\overline{\underline{x}_i' \widehat{\Sigma}^{-1} \underline{u}_i}$ *est donc le même que celui de* $\sqrt{N}\underline{x}_i' \Sigma^{-1} \underline{u}_i$. *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* $\mathcal{N}\left(0, E\left(\underline{x}_i' \Sigma^{-1} \underline{x}_i\right)\right)$

Les deux derniers points se démontrent de la même façon que précédemment

Application : Données de panel et Régressions empilées

La mise en oeuvre de l'estimateur des MCQG dans le cas des données de panel ou des régressions empilées est très simple. Elle se fait en plusieurs étapes.

- On estime d'abord le modèle

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

par les MCO : $\widehat{b}_{MCO} = (\underline{x}'\underline{x})^{-1}(\underline{x}'\underline{y})$

- On calcule ensuite le résidu pour chaque individu

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

- A partir de cet estimateur 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 estimer la variance asymptotique et la variance de l'estimateur des MCO par

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

- Dans une deuxième étape, on calcule l'estimateur des MCQG

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

  Cette mise en oeuvre peut être facilitée s'il existe un moyen simple de sphériciser le modèle.
- La variance est alors donnée par :

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

Suivant les cas on peut avoir un nombre plus ou moins important de paramètres à estimer. Dans le cas des données de panel la matrice de variance ne dépend que de deux paramètres la variance de l'effet individuel et la variance de l'effet temporel. Différentes méthodes peuvent être utilisées pour estimer ces paramètres et donc mettre en oeuvre l'estimateur des MCQG.

9.1 Le cas des régressions empilées.

Lorsque l'on a affaire à des régressions empilées pour lesquelles il n'existe pas de restriction entre les paramètres, alors le calcul de l'estimateur est facilité par la proposition suivante connue sous le nom de Théorème de Zellner

Proposition *Dans le cas des régressions empilées, lorsqu'il n'existe pas de contraintes entre les paramètres d'une équation à l'autre (et que les régresseurs sont les mêmes) l'estimateur des MCG est identique à l'estimateur des MCO effectué équation par équation. La matrice de variance de l'estimateur a alors pour expressions*

$$V_{as} = \Sigma \otimes \overline{\underline{x}_i' \underline{x}_i}^{-1}$$

Démonstration *Le modèle s'écrit*

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

L'estimateur des MCG s'écrit :

$$b_{mcg} = \overline{(I_M \otimes \underline{x}_i') \Sigma^{-1} (I_M \otimes \underline{x}_i)}^{-1} \overline{(I_M \otimes \underline{x}_i') \Sigma^{-1} \underline{y}_i}$$

On peut réécrire $\Sigma^{-1} = \Sigma^{-1} \otimes 1$. $\underline{x}_i'$ est un vecteur $(K+1) \times 1$. Donc $(I_M \otimes \underline{x}_i') \Sigma^{-1} = (I_M \otimes \underline{x}_i')(\Sigma^{-1} \otimes 1) = \Sigma^{-1} \otimes \underline{x}_i'$. Car $(A \otimes B)(C \otimes D) = AB \otimes CD$ pour des matrices aux dimensions qui conviennent. Donc

$$\overline{(I_M \otimes \underline{x}_i') \Sigma^{-1} (I_M \otimes \underline{x}_i)} = \Sigma^{-1} \otimes \overline{\underline{x}_i' \underline{x}_i}$$

en outre

$$\begin{aligned}
\overline{(I_M \otimes \underline{x}_i') \Sigma^{-1} \underline{y}_i} &= \overline{(I_M \otimes \underline{x}_i') \left(\Sigma^{-1} \underline{y}_i \otimes 1 \right)} \\
&= \overline{\Sigma^{-1} \underline{y}_i \otimes \underline{x}_i'} = \overline{(\Sigma^{-1} \otimes I_K) \left(\underline{y}_i \otimes \underline{x}_i' \right)} \\
&= \left(\Sigma^{-1} \otimes I_K \right) \overline{\left(\underline{y}_i \otimes \underline{x}_i' \right)}
\end{aligned}$$

donc l'estimateur des MCQG s'écrit

$$\begin{aligned}
b_{mcg} &= \Sigma \otimes \overline{\underline{x}_i' \underline{x}_i}^{-1} \left(\Sigma^{-1} \otimes I_K \right) \overline{\left(\underline{y}_i \otimes \underline{x}_i' \right)} = I_M \otimes \overline{\underline{x}_i' \underline{x}_i}^{-1} \overline{\left(\underline{y}_i \otimes \underline{x}_i' \right)} \\
&= I_M \otimes \overline{\underline{x}_i' \underline{x}_i}^{-1} Vec \left(\overline{\underline{x}_i' \underline{y}_i'} \right) \\
&= Vec \left(\overline{\underline{x}_i' \underline{x}_i}^{-1} \overline{\underline{x}_i' \underline{y}_i'} \right)
\end{aligned}$$

On utilise ici la propriété de l'opérateur Vec : $Vec(ABC) = C' \otimes AVecB$

9.2 Illustration : estimation d'une fonction de production sur données individuelles

On considère un échantillon de 381 entreprises observées sur les années 1986-1989, pour lesquelles on dispose de la valeur ajoutée, des effectifs du stock de capital et du stock de capital recherche. On considère une technologie de production de Cobb-Douglas

$$y = \alpha + \alpha_L l + \alpha_C c + \alpha_K k + v$$

les coefficients sont donc les élasticités de la production aux effectifs, au capital et au capital de recherche. Les observations dont on dispose sont des données de panel puisque chacun des 381 individu est suivi sur 4 ans : $\underline{y}_i' = (y_{i86}, y_{i87}, y_{i88}, y_{i89})$. On estime le modèle par les mco. Il est alors possible d'estimer la matrice de variance des perturbations

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

on peut alors calculer les écarts-type de deux façons : soit en ignorant la nature de données de panel des données, i.e. en faisant comme si la matrice Σ était diagonale, soit en prenant cette information en compte. Dans un cas les écarts-type sont simplement donnés par la formule standard $\widehat{V}_{as} = \widehat{\sigma}^2 \left(\overline{\underline{x}_i' \underline{x}_i}\right)^{-1}$ et $\widehat{V}_b(1) = \widehat{V}_{as}/N$. Dans l'autre cas les écarts-type sont calculés suivant la formule $\widehat{V}_{as} = \left(\overline{\underline{x}_i' \underline{x}_i}\right)^{-1} \overline{\underline{x}_i' \widehat{\Sigma} \underline{x}_i} \left(\overline{\underline{x}_i' \underline{x}_i}\right)^{-1}$ et toujours $\widehat{V}_b(2) = \widehat{V}_{as}/N$. Le tableau suivant présente les résultats de cette estimation par les mco et les écarts-type calculés suivant les deux modes de calcul :

	b	$\widehat{\sigma}(1)$	$\widehat{\sigma}(2)$
un	4.78	(0.120)	(0.226)
l	0.509	(0.023)	(0.044)
c	0.235	(0.022)	(0.040)
k	0.229	(0.017)	(0.026)

On voit que les écarts-type sont nettement plus élevé avec la formule qui tient compte des corrélations entre les résidus aux différentes dates. On peut regarder la matrice de variance des perturbations estimée. On parvient à la matrice symétrique suivante :

	86	87	88	89
86	0.209	.	.	.
87	0.191	0.214	.	.
88	0.184	0.186	0.203	.
89	0.176	0.177	0.192	0.210

et on voit qu'elle est très loin d'être une matrice diagonale. Les éléments sur la diagonale sont plus ou moins constants, mais on voit aussi que les éléments hors de la diagonale sont certes plus faibles que ceux sur la diagonale mais d'un ordre de grandeur comparable. L'hétéroscédasticité est ainsi une caractéristique essentielle et l'omettre serait une grave

erreur. Compte tenu de l'ordre de grandeur des coefficients de la matrice de variance covariance on voit qu'on est beaucoup plus prés d'une situation dans laquelle les observations seraient répétées quatre fois que d'une situation dans laquelle les quatre observations de chaque individu constitueraient quatre tirages indépendants. Le nombre total d'observations est donc 381x4=1524 mais on est très loin d'avoir l'information de 1524 observations indépendantes. On est bien plus prés d'avoir 381 observations répliquées 4 fois. De fait les estimateurs étant convergent en $\sqrt{N}$. Comme la dimension temporelle est de 4, on doit se tromper approximativement d'un facteur $\sqrt{4} = 2$ dans les écarts-type. C'est bien ce que l'on observe en gros. La conclusion que l'on doit tirer de cet exemple est que la correction des écarts-type tenant compte de l'hétéroscédasticité est essentielle pour les données de panel.

On peut aussi chercher à mettre en oeuvre l'estimateur des MCQG la formule est :

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

et la matrice de variance peut être estimées par $\widehat{V}_{asmcqg} = \left(\overline{\underline{x}_i' \widehat{\Sigma}^{-1} \underline{x}_i}\right)^{-1}$ et $\widehat{V}_b(3) = \widehat{V}_{asmcqg}/N$. Les résultats sont donnés dans le tableau suivant :

	$\widehat{b}_{mcqg}$	$\widehat{\sigma}_{mcqg}$
C^{ste}	4.67	(0.193)
l	0.505	(0.032)
c	0.352	(0.026)
k	0.086	(0.009)

On voit que par rapport à l'estimateur des mco, cet estimateur est sensiblement plus précis. Le coefficient du capital recherche en particulier est environ 3 fois plus précis. La mise en oeuvre de ce type d'estimation est donc dans ce cas un gain précieux. On remarque aussi que les deux estimateurs sont en fait assez différents en particulier les coefficients concernant le capital physique et le capital de recherche. Le coefficient du capital physique augmente fortement alors que celui du capital recherche baisse au contraire. Ces différences importantes sont en outre grandes devant l'ordre de grandeur des écarts-type. Bien qu'il n'y est pas de test formel ici, il est vraisemblable que ces différences soient significatives. Ceci n'est pas un bon signe, comme on le verra plus tard. En effet on peut remarquer dès maintenant une sorte d'incohérence : normalement sous les hypothèses faites l'estimateur des mco et celui de mCQG sont tous les deux convergents : les valeurs estimées devraient donc être assez proches.

9.3 Résumé

Dans ce chapitre on a :

- exhibé différentes situations fréquentes en pratique dans lesquelles l'hypothèse d'homoscédasticité n'est plus satisfaite.
- présenté un estimateur alternatif à l'estimateur des mco, de variance minimale parmi les estimateurs linéaires sans biais : l'estimateur des MCG
- cet estimateur est fonction de la matrice de variance des perturbations qui est inconnue. L'estimateur n'est donc pas calculable. On a présenté l'estimateur de mCQG dans lequel la matrice de variance des perturbations, inconnue, est remplacée par un estimateur.
- L'estimateur n'est plus sans biais. Ses propriétés ne sont qu'asymptotiques. Dans le meilleur des cas il est asymptotiquement équivalent à l'estimateur des mco.
- On a montré comment dans le cas où la matrice de variance dépend d'un nombre fini de paramètres, il est possible de préciser les propriétés asymptotiques de l'estimateur des mCQG.
- Sous des hypothèses peu exigeantes, cet estimateur et ne peut pas être calculé en pratique réalisant examiné les propriétés asymptotique de l'estimateur des mco rappelé les propriétés asymptotiques importantes des moyennes empiriques de variables : la loi des grands nombres et le théorème central limite.
- montré que sous des hypothèses très faibles (existence des moments d'ordre 1 et 2), l'estimateur des mco est convergent et asymptotiquement normal.
- Etendu la notion de test pour définir des tests asymptotiques, caractérisés par le fait que leur puissance tend vers 1 et généralisé les notions de test de Student et de test de Fisher au cas asymptotique.

Chapitre 10

Variables instrumentales

On a considéré jusqu'à présent le cas de modèles s'écrivant

$$y_i = b_0 + x_i^1 b_1 + \cdots + x_i^K b_K + u_i$$

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 statistique du paramètre b. Le coefficient b s'interprète alors comme le vecteur des coefficients de la régression linéaire de y_i sur le vecteur de variables x_i. Une telle définition présente un intérêt dans une approche descriptive des données. Néanmoins on est fréquemment amené 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 en pourcentage l'incidence d'une augmentation de 1% du stock de capital sur la production. Ce paramètre économique n'a pourtant aucune raison de coïncider avec celui de la régression linéaire, et on peut même avancer de nombreuses raisons pour lesquelles il pourrait ne pas coïncider. On est ainsi fréquemment amené à considérer des modèles structurels pour lesquels on a une équation linéaire entre une variable d'intérêt et des variables explicatives mais pour laquelle on a des raisons de remettre en doute l'hypothèse $E\left(u_i \left| x_i \right.\right) = 0$. Ce chapitre est consacré à la présentation des méthodes d'estimations élémentaires adaptées à l'estimation des paramètres structurels dans ce cas. On va voir que l'on peut identifier le paramètre d'intérêt en ayant recours à des hypothèses alternatives à $E\left(u_i \left| x_i \right.\right) = 0$ qui mobilisent des informations extérieures. Elles vont prendre la forme suivante : il existent des variables extérieures dites instrumentales telles que $E\left(u_i \left| z_i \right.\right) = 0$ et $E\left(z_i' x_i\right)$ de rang $K+1$. On va voir aussi deux tests très importants dits tests de spécifications qui permettent de guider dans le choix des variables extérieures

(test de Sargan) et de tester l'hypothèse des mco : $E(u_i|x_i) = 0$ (test d'exogénéité). Si dans les chapitres précédents on mettait beaucoup l'accent sur l'efficacité des estimateurs (le Théorème de Gauss-Markov), ici on va mettre au contraire l'accent sur l'identification des paramètres et sur la robustesse des estimations, et on va voir qu'il y a un arbitrage entre robustesse et efficacité.

10.1 Trois exemples types d'endogénéité des régresseurs

10.1.1 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 supposée pour simplifier de dimension 1 et centrée comme la variable y_i et on fait l'hypothèse $E(u_i|x_i^*) = 0$.

On suppose en outre que la variable x_i^* est mesurée avec erreur :

$$x_i = x_i^* + e_i$$

avec $E(e_i|x_i^*) = 0$ et u_i et e_i non corrélés.

Dans ces conditions le modèle dont on dispose est

$$y_i = x_i b + u_i - be_i$$

On est dans une situation dans laquelle le résidu de l'équation $v_i = u_i - be_i$ est corrélé avec la variable explicative

$$\begin{aligned} E(v_i x_i) &= E((u_i - be_i)(x_i^* + e_i)) \\ &= E(u_i x_i^*) + E(u_i e_i) - bE(e_i x_i^*) - bE\left(e_i^2\right) \\ &= -b\sigma_e^2 \neq 0 \end{aligned}$$

On voit alors très facilement qu'à la limite le paramètre de la régression linéaire ne coïncide pas avec celui du modèle : l'estimateur des mco n'est pas convergent.

$$b_{mco} \xrightarrow{P} b + \frac{E\left(x_i' v_i\right)}{E\left(x_i' x_i\right)} = b\left(1 - \frac{\sigma_e^2}{\sigma_e^2 + \sigma_{x^*}^2}\right)$$

10.1.2 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 : $E\left(u_i^d p_i\right) \neq 0$. On peut voir aisément que l'estimateur des mco de l'équation de demande ou d'offre sera biaisé. On peut pour cela considérer le graphe représentant l'équilibre offre demande représenté sur la figure 10.1. Les observations correspondent à l'ensemble des intersections des courbes d'offre et de demande. Ces courbes se déplacent, sous l'action des variations des variables explicatives et aussi sous l'action des chocs de demande et d'offre. On voit que s'il n'y a que des chocs de demande, l'ensemble des points d'intersection des courbes d'offre et de demande va décrire la courbe de demande, de même, s'il n'y a que des chocs de demande, l'ensemble des points d'équilibre va décrire la courbe d'offre. Dans le cas général, il y a des chocs d'offre et de demande, et l'ensemble des équilibres ne décrit ni la courbe d'offre ni la courbe de demande, la droite de régression passe au milieu.

10.1.3 Omission de régresseurs, hétérogénéité inobservée

On considère le modèle

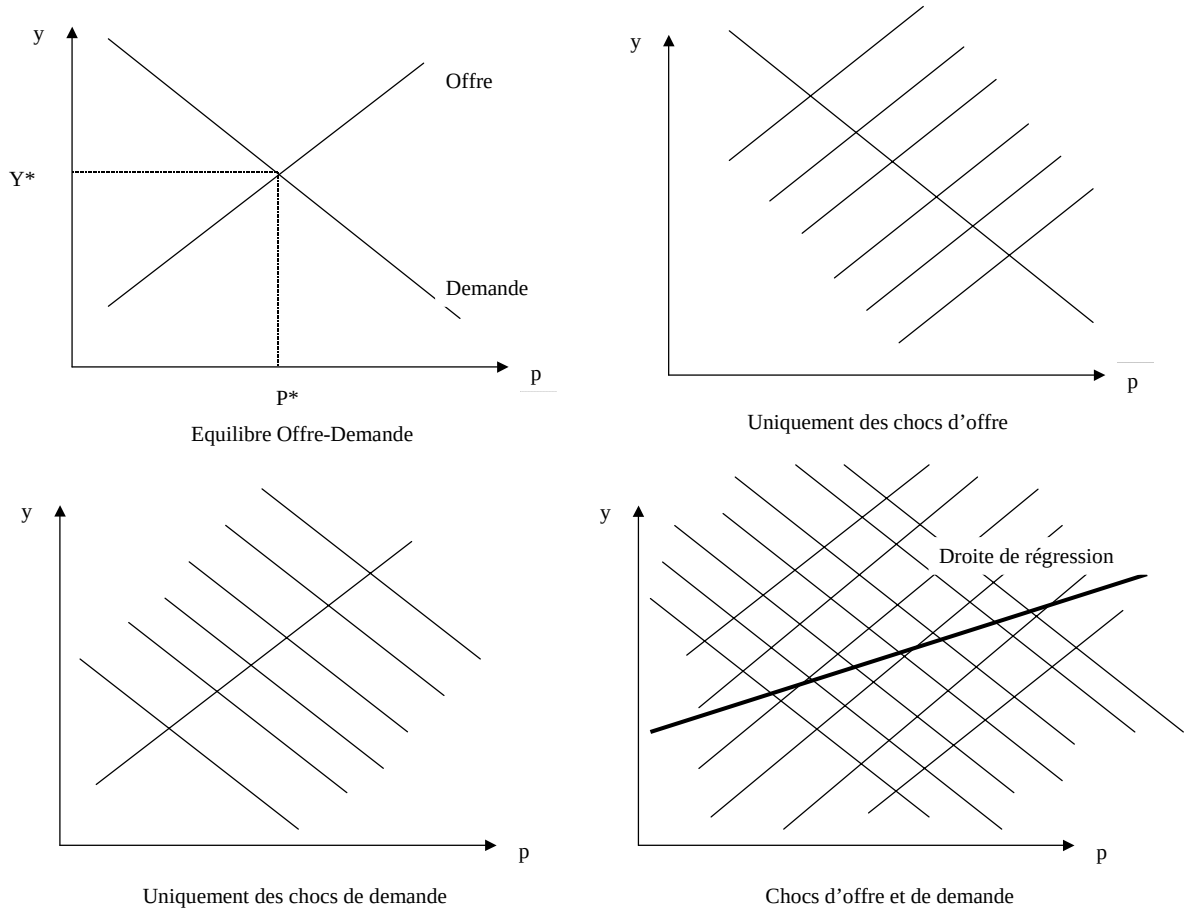
$$y_i = x_i b + z_i c + u_i$$

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 dès lors qu'elle 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 régression linéaire de z_i sur x_i.



TAB. 10.1 – différents équilibre offre-demande

Un exemple important est donné par les équations 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. 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.

On peut considérer aussi le cas d'une fonction de production agricole : y_i est le rendement de la terre, x_i la quantité d'engrais b est le rendement des épandages et z_i la qualité de la terre. L'omission de cette variable biaise l'estimation du paramètre technologique b si les décisions d'épandages d'engrais dépendent de la qualité de la terre. Le paramètre estimé n'identifie pas seulement le paramètre structurel mais une combinaison non désirée de ce paramètre et de celui reflétant le comportement de l'agriculteur.

10.2 La méthode des variables instrumentales

10.2.1 Modèle à variables endogènes et non convergence de l'estimateur des mco

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. En effet, il 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

$$p\lim \widehat{b}_{mco} \neq b$$

Remarque *On a introduit une distinction entre variable endogène et variable exogène, néanmoins l'ensemble des coefficients est biaisé et pas seulement ceux des variables endogènes. Pour le voir on peut considérer l'exemple de la fonction de production que l'on considère en taux de croissance*

$$\Delta y_i = \alpha \Delta l_i + \beta \Delta k_i + u_i$$

On fait en général l'hypothèse que le stock de capital s'ajuste lentement et n'est de ce fait pas corrélé avec la perturbation. Par contre le travail est un facteur variable, positivement corrélé à la perturbation : $E(\Delta l_i u_i) = \theta > 0$. *On calcule sans peine la valeur limite du paramètre :*

$$\begin{aligned} p\lim biais_{mco} &= \frac{1}{V(\Delta l_i)V(\Delta k_i) - cov(\Delta l_i \Delta k_i)} \begin{pmatrix} V(\Delta k_i) & -cov(\Delta l_i \Delta k_i) \\ -cov(\Delta l_i \Delta k_i) & V(\Delta l_i) \end{pmatrix} \begin{pmatrix} \theta \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} V(\Delta k_i)\theta \\ -cov(\Delta l_i \Delta k_i)\theta \end{pmatrix} \Big/ V(\Delta l_i)V(\Delta k_i) - cov(\Delta l_i \Delta k_i) \end{aligned}$$

On constate donc que les deux coefficients sont biaisés : celui du travail sans ambiguïté à la hausse, et celui du capital à la baisse si comme c'est probable le capital et le travail sont corrélés positivement.

10.2.2 Résoudre le problème de l'identification par l'utilisation de variables instrumentales

Sans prétendre produire ici des estimateurs, on s'intéresse aux conditions d'identification. On considère pour cela 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 éléments peuvent être commun aux deux ensembles et n'interviennent dans ce cas qu'une fois dans x_i. On fait les hypothèses

$$E\left(x_i^{'} u_i^d\right) = 0, E\left(x_i^{'} u_i^s\right) = 0 \tag{10.1}$$

c.-à-d. que 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 10.1, on peut exprimer les coefficients des régressions linéaires de y_i et p_i sur x_i à partir des paramètres structurels.

La *modélisation,* c'est à dire la spécification d'une fonction d'offre et de demande et des restrictions stochastiques (exogénéité de x_i), conduit à des *restrictions* sur les paramètres des régressions linéaires des variables endogènes qui sont susceptibles de permettre l'*identification* des paramètres structurels du modèle.

Proposition *S'il existe une variable exogène intervenant spécifiquement dans l'équation d'offre, l'équation de demande est identifiée.*

De même, s'il existe une variable exogène intervenant spécifiquement dans l'équation de demande, l'équation d'offre est identifiée

Démonstration *Si x_{1i}^s est une telle variable, le coefficient de cette variable dans la régression 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 régression 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*

Ce résultat est obtenu en ayant recours à une modélisation de l'ensemble des variables endogènes du modèle : la production et le prix, ou de façon équivalente le système d'équations qui les détermine simultanément. Dans de nombreuses situations on ne s'intéresse qu'à une des deux équations, par exemple 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 et vérifiant $E\left(\left[x_i^d\ x_{1i}^s\right]' u_i^d\right) = 0$. Dans ce cas si on considère les coefficients γ_y et γ_p des régressions linéaires de y_i et p_i sur $\widetilde{x}_i = \left[x_i^d\ x_{1i}^s\right]$ sont

$$\begin{aligned}
\gamma_y &= E\left(\widetilde{x}_i'\widetilde{x}_i\right)^{-1} E\left(\widetilde{x}_i' y_i\right) = E\left(\widetilde{x}_i'\widetilde{x}_i\right)^{-1} E\left(\widetilde{x}_i'\left(-\alpha^d p_i + x_i^d b^d + u_i^d\right)\right) \\
&= -\alpha^d E\left(\widetilde{x}_i'\widetilde{x}_i\right)^{-1} E\left(\widetilde{x}_i' p_i\right) + E\left(\widetilde{x}_i'\widetilde{x}_i\right)^{-1} E\left(\widetilde{x}_i' x_i^d\right) b^d \\
&= -\alpha^d \gamma_p + \begin{pmatrix} b_d & 0 \end{pmatrix}'
\end{aligned}$$

Le vecteur γ_y est identifié par les données : il s'agit du vecteur des coefficients de la régression linéaire de y_i sur $\widetilde{x}_i$. Il en est de même pour le vecteur γ_p. dès lors que le coefficient de la variable x_{1i}^s dans la régression de la variable de prix sur $\widetilde{x}_i$, élément de γ_p, est non nul, et que la variable x_{1i}^s ne figure pas dans la liste des régresseurs exogènes (structurels) de l'équation de demande, on voit que les coefficients de l'équation de demande sont identifiés. Il n'en est pas nécessairement de même pour l'équation d'offre, soit parce que l'on ne mesure pas toutes les variables x_i^s garantissant $E\left(u_i^s x_i^s\right) = 0$, soit parce qu'il n'y a pas de variables affectant la demande qui n'affecte pas directement l'offre. Enfin on remarque qu'il n'est pas nécessaire de spécifier l'équation d'offre.

Cet exemple illustre bien la démarche des variables instrumentales. Celle-ci correspond à la mobilisation de variables extérieures au modèle qui possèdent la particularité de ne pas être corrélées avec le résidu de l'équation structurelle et qui sont néanmoins corrélées

avec la variable endogène. L'identification vient alors du fait que l'effet de la variable instrumentale sur la variable dépendante ne fait que refléter celui de la variable endogène.

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.

On voit donc qu'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.

10.2.3 Identification

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} $(\dim x_{1i} = K_1 = K - K_2)$.

Definition *Un ensemble de variables* $z_i = (z_i^e, x_{2i})$, *de dimension* $H + 1$, *non parfaitement corrélées (*$\operatorname{rang} E\left(z_i^{'} z_i\right) = H + 1$*), est dit ensemble de variables* instrumentales *si les deux conditions suivantes sont satisfaites :*

$$E\left(z_i^{'} u_i\right) = 0. \tag{10.2}$$

et

$$\operatorname{rang} E\left(z_i^{'} x_i\right) = K + 1$$

La première condition, appelée *condition d'orthogonalité*, consiste à supposer que le vecteur des variables instrumentales n'est pas corrélé avec le résidu de l'équation structurelle. Il fait intervenir les $K_2 + 1$ variables exogènes x_2 ainsi que $(H + 1) - (K_2 + 1) = H - K_2$ instruments extérieurs z_i^e.

L'hypothèse (10.2) est parfois introduite sous la forme :

$$E(u_i|z_i) = 0$$

qui est plus forte que la précédente (non corrélation) puisqu'elle implique en particulier $E\left(g\left(z_i\right) u_i\right) = 0$ pour toute fonction g.

La deuxième condition est dite condition de rang. Elle joue un rôle essentiel, parfois oublié, et que l'on détaillera par la suite.

La condition (10.2) peut être réécrite comme suit :

$$E\left(z_i^{'}(y_i - x_i b)\right) = 0$$

Soit encore :

$$E\left(z_i' y_i\right) = E\left(z_i' x_i\right) b \tag{10.3}$$

Cette condition définit un système de $H+1$ équations à $K+1$ inconnues b.

Le modèle est identifié si le système (10.3) 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 $\operatorname{rang} E\left(z_i' x_i\right) = K+1$ le modèle est juste identifié.
- Si $H > K$, $\operatorname{rang} E\left(z_i' x_i\right) = K+1$ le modèle est dit sur-identifié. Dans ce cas il y a plus de variables instrumentales qu'il n'est nécessaire

La condition de rang garantit que l'on se trouve dans l'une des deux dernières situations.

Proposition *Considérant le modèle*

$$y_i = x_i b + u_i$$

Sous les hypothèses

- $\exists\ z_i$ *tel que* $E\left(z_i' u_i\right) = 0$
- $E\left(z_i' x_i\right)$ *est de rang K+1,*

Le paramètre b est identifié.

Démonstration *En multipliant le modèle par* z_i' *et en prenant l'espérance, il vient*

$$E\left(z_i' y_i\right) = E\left(z_i' x_i\right) b + E\left(z_i' u_i\right) = E\left(z_i' x_i\right) b$$

Comme $E\left(z_i' x_i\right)$ *est de rang K+1, il existe nécessairement une matrice A de dimension* $(K+1) \times \dim z_i$ *telle que* $AE\left(z_i' x_i\right)$ *de dimension* $(K+1) \times (K+1)$ *soit inversible (il suffit par exemple de considérer* $A = E\left(z_i' x_i\right)'$*). On en déduit donc que*

$$b = \left(AE\left(z_i' x_i\right)\right)^{-1} AE\left(z_i' y_i\right)$$

b s'exprime donc comme la limite d'une fonction ne dépendant que des observations par exemple $\left(A\overline{z_i' x_i}\right)^{-1} A\left(\overline{z_i' y_i}\right)$

10.2.4 Moindres carrés indirects

Si $H = K$ et si $E\left(z_i' x_i\right)$ est inversible, ce qui est le cas dès lors que la condition de rang est satisfaite, alors on peut résoudre $b = E\left(z_i' x_i\right)^{-1} E\left(z_i' y_i\right)$. On obtient un estimateur

de b appelé Estimateur des Moindres Carrés Indirects en remplaçant les espérances par leurs contreparties empiriques :

$$\begin{aligned}\widehat{b}_{mci} &= \left(\frac{1}{N}\sum_{i=1}^{N} z_i' x_i\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} z_i' y_i \\ &= (\underline{z}'\underline{x})^{-1}\underline{z}'\underline{y}\end{aligned}$$

où $\underline{z}$ est la matrice dont la i-ième ligne est z_i, $\underline{x}$ la matrice dont la i-ième ligne est x_i et $\underline{y}$ 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 $AE(z_i'x_i)$ est inversible,

$$b = (AE(z_i'x_i))^{-1} AE(z_i'y_i).$$

On en déduit une classe d'estimateur :

$$\begin{aligned}\widehat{b}_{mci}(A) &= \left(A\overline{z_i'x_i}\right)^{-1} A\overline{z_i'y_i} \\ &= (A\underline{z}'\underline{x})^{-1}A\underline{z}'\underline{y}\end{aligned}$$

10.2.5 Propriété asymptotiques des estimateurs des MCI

Proposition *Dans le modèle*

$$y_i = x_i b + u_i$$

à $K+1$ variables explicatives. Sous les hypothèses :

H1 $E(z_i'u_i) = 0$ *avec* z_i *de dim* $1 \times H+1$

H2 Les observations (x_i, z_i, y_i) *sont iid*

H3 $E(u_i^2|z_i) = \sigma^2$

H4 Les moments de (x_i, z_i, y_i) *existent jusqu'à un ordre suffisant*

H5 $E\left(z_i'x_i\right)$ *et* $\overline{z_i'x_i}$ *sont de rang* $K+1$

Alors, 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 et toute suite de matrice, éventuellement dépendant des données* $A_n \xrightarrow{p} A$, *on a :*

1. $\widehat{b}_{mci}(A)$ *est convergent :* $p\lim \widehat{b}_{mci}(A) = b$
2. $\widehat{b}_{mci}(A)$ *est asymptotiquement normal :*

$$\sqrt{N}\left(\widehat{b}_{mci}(A) - b\right) \xrightarrow{L} N(0, \Sigma(A)),$$

avec

$$\Sigma(A) = \sigma^2 \left[AE\left(z_i' x_i\right)\right]^{-1} AE\left(z_i' z_i\right) A' \left[E\left(x_i' z_i\right) A'\right]^{-1}$$

3. $\widehat{\Sigma}(A) = \widehat{\sigma}^2 \left[A\overline{z_i' x_i}\right]^{-1} A\overline{z_i' z_i} A' \left[\overline{x_i' z_i} A'\right]^{-1}$ *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* $\operatorname{rang} E\left(z_i' x_i\right) = K + 1$. *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 :

$$\widehat{b}_{mci}(A_N) = \left(A_N \overline{z_i' x_i}\right)^{-1} A_N \overline{z_i' y_i} = b + \left(A_N \overline{z_i' x_i}\right)^{-1} A_N \overline{z_i' u_i}.$$

La convergence découle simplement de la loi des grands nombres :

$$\overline{z_i' u_i} \xrightarrow{p} E\left(z_i' u_i\right) = 0.$$

et du fait que $A_N \xrightarrow{p} A$ *et* $\overline{z_i' x_i} \xrightarrow{p} E\left(z_i' x_i\right)$

Normalité asymptotique

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

Comme $V\left(z_i' u_i\right) = E(z_i' z_i u_i^2) = E\left[z_i' z_i E(u_i^2 \mid z_i)\right] = \sigma^2 E\left(z_i' z_i\right)$, *la normalité asymptotique découle directement du théorème central limite :*

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

et $\left(A_N \overline{z_i' x_i}\right)^{-1} A_N \xrightarrow{p} \left(AE\left(z_i' x_i\right)\right)^{-1} A$

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* $b - \widehat{b}\left(A\right) \to 0$

Remarque *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 H3. Les conclusions sont simplement modifiées en :* $\widehat{b}_{mci}(A)$ *est asymptotiquement normal :*

$$\sqrt{N}\left(\widehat{b}_{mci}(A) - b\right) \xrightarrow{L} N(0, \Sigma_{het}(A)),$$

avec

$$\Sigma_{het}(A) = \left[AE\left(z_i'x_i\right)\right]^{-1} AE\left(u_i^2 z_i'z_i\right) A' \left[E\left(x_i'z_i\right) A'\right]^{-1}$$

et $\widehat{\Sigma}_{het}(A) = \left[A\overline{z_i'x_i}\right]^{-1} A\overline{\widehat{u}\left(A\right)_i^2 z_i'z_i} A' \left[\overline{x_i'z_i} A'\right]^{-1}$ *est un estimateur convergent de la matrice de variance.*

10.3 L'estimateur des doubles moindres carrés

10.3.1 Existence d'un estimateur optimal

On peut se demander s'il n'existe pas une matrice A^* qui conduise à un estimateur de variance minimale, c'est à dire tel que pour toute combinaison linéaire λb, on ait $V\left(\lambda\widehat{b}\left(A^*\right)\right) \leq V\left(\lambda\widehat{b}\left(A\right)\right)$. Une telle matrice existe et mène à l'estimateur des doubles moindres carrés.

Proposition *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

$$\Sigma(A^*) = \sigma^2 \left[E\left(x_i'z_i\right) E\left(z_i'z_i\right)^{-1} E\left(z_i'x_i\right)\right]^{-1}$$

Démonstration *Pour montrer que* $\Sigma(A) \geqslant \Sigma(A^*)$ *au sens des matrices, i.e.* $\forall\ \lambda$ *on a* $\lambda'\left(\Sigma(A) - \Sigma(A^*)\right)\lambda \geqslant 0$ *on peut clairement éliminer le facteur* σ^2. *La matrice de variance* $\Sigma(A^*)$ *s'écrit :*

$$\Sigma(A^*) = \left[E\left(x_i'z_i\right) E\left(z_i'z_i\right)^{-1} E\left(z_i'x_i\right)\right]^{-1} = \left(C'C\right)^{-1}$$

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

$$\Sigma(A) - \Sigma(A^*) = BB' - (C'C)^{-1} = BB' - BC\,(C'C)^{-1}\,C'B'$$

puisque $BC = I$. *On a donc :*

$$\Sigma(A) - \Sigma(A^*) = B\left[I - C\,(C'C)^{-1}\,C'\right]B'$$

Comme $I - C\,(C'C)^{-1}\,C'$ *est une matrice semi-définie positive,* $\Sigma(A) - \Sigma(A^*)$ *est aussi une matrice semi-définie positive*

Remarque *On a vu que dans le cas hétéroscédastique, la variance de l'estimateur des moindres carrés indirects s'écrivait :* $\Sigma_{het}(A) = \left[AE\left(z_i'x_i\right)\right]^{-1} AE\left(u_i^2 z_i'z_i\right) A' \left[E\left(x_i'z_i\right) A'\right]^{-1}$. *O voit par analogie avec le cas précédent homoscédastique que dans ce cas aussi il y a un estimateur optimal et qu'il correspond à la matrice* $A = E\left(x_i'z_i\right) E\left(u_i^2 z_i'z_i\right)^{-1}$.

10.3.2 L'estimateur optimal comme estimateur des doubles moindres carrés

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. $A_N = \overline{x_i'z_i}\ \ \overline{z_i'z_i}^{-1}$ est un choix naturel.

$$\begin{aligned}
\widehat{b}_{mci}(A_N) &= \left(\overline{x_i'z_i}\ \ \overline{z_i'z_i}^{-1}\overline{z_i'x_i}\right)^{-1} \overline{x_i'z_i}\ \ \overline{z_i'z_i}^{-1}\overline{z_i'y_i} \\
&= \left(\underline{x}'\underline{z}\,(\underline{z}'\underline{z})^{-1}\,\underline{z}'\underline{x}\right)^{-1} \underline{x}'\underline{z}\,(\underline{z}'\underline{z})^{-1}\,\underline{z}'\underline{y}
\end{aligned}$$

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 $\underline{z}$, $P_z = \underline{z}\,(\underline{z}'\underline{z})^{-1}\,\underline{z}'$

$$\widehat{b}_{2mc}(A^*) = (\underline{x}'P_z\underline{x})^{-1}\,\underline{x}'P_z\underline{y} = ((P_z\underline{x})'\,P_z\underline{x})^{-1}\,(P_z\underline{x})'\,\underline{y}$$

On voit que la projection des variables explicatives sur les variables instrumentales joue un rôle très important. Il correspond de façon évidente à l'estimateur des mco de la variable endogène $\underline{y}$ sur la projection $\widehat{\underline{x}} = P_z\underline{x}$ des variables explicatives sur l'ensemble des instruments. On peut vérifier directement ce point en considérant à nouveau le modèle et en décomposant les variables explicatives en $\underline{x} = P_z\underline{x} + M_z\underline{x}$. Le modèle s'écrit :

$$\begin{aligned}
\underline{y} &= \underline{x}b + \underline{u} \\
&= P_z\underline{x}b + M_z\underline{x}b + \underline{u} = P_z\underline{x}b + \underline{v}
\end{aligned}$$

Ici la perturbation comprend le vrai résidu mais aussi la partie des variables explicatives orthogonales aux variables instrumentales : $\underline{v} = M_z\underline{x} + \underline{u}$. On voit que pour ce nouveau modèle dans lequel les régresseurs ont été remplacés par leurs projections sur les variables explicatives, il y a orthogonalité entre le résidu et les variables explicatives puisque $(P_z\underline{x})'\,\underline{u}/N = \underline{x}'\underline{z}/N\,(\underline{z}'\underline{z}/N)^{-1}\,\underline{z}'\underline{u}/N \to E\,(x'z)\,E\,(\underline{z}'\underline{z})^{-1}\,E\,(\underline{z}'\underline{u}) = 0$ et $(P_z\underline{x})'\,M_z\underline{x} = \underline{x}'P_zM_z\underline{x} = 0$. On en déduit que l'estimateur des mco de la régression de $\underline{y}$ sur $P_z\underline{x}$ est bien convergent.

C'est pourquoi on appelle cet estimateur *estimateur des doubles moindres carrés* et on le note $\widehat{b}_{2mc}$ puisqu'il pourrait être obtenu à partir d'une première régression des variables explicatives sur les variables instrumentales puis par régression de la variable endogène sur les variables prédites de cette régression.

L'estimateur peut être déterminé en deux étapes :

1. On régresse $\underline{x}$ sur $\underline{z}$ et on récupère $\widehat{\underline{x}}$ la valeur prédite.
2. On régresse y sur $\widehat{\underline{x}}$

La matrice de variance asymptotique de $\widehat{b}_{2mc}$ est

$$V_{as}(\widehat{b}_{2mc}) = \sigma^2\left[E\left(x_i'z_i\right)E\left(z_i'z_i\right)^{-1}E\left(z_i'x_i\right)\right]^{-1}$$

et la matrice de variance de l'estimateur dans un échantillon de taille N est

$$V(\widehat{b}_{2mc}) = V_{as}/N = \sigma^2\left[E\left(x_i'z_i\right)E\left(z_i'z_i\right)^{-1}E\left(z_i'x_i\right)\right]^{-1}/N$$

On peut l'estimer par

$$\widehat{V}(\widehat{b}_{2mc}) = \widehat{\sigma}^2\left(\underline{x}'\underline{z}\left(\underline{z}'\underline{z}\right)^{-1}\underline{z}'\underline{x}\right)^{-1} = \widehat{\sigma}^2(\underline{x}'P_z\underline{x})^{-1} = \widehat{\sigma}^2\left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)^{-1}$$

L'écart-type des résidus à retenir est celui du modèle

$$y_i = x_ib + u_i$$

et peut être estimé par $\overline{\left(y_i - x_i\widehat{b}_{2mc}\right)^2}$. Il faut remarquer qu'ici il s'agit du résidu $\widehat{u}_i = y_i - x_i\widehat{b}_{2mc}$ et non du résidu de la deuxième étape $y_i - \widehat{x}_i\widehat{b}_{2mc}$.

Cette écriture de l'estimateur à variables instrumentales montre qu'on peut l'interpréter comme opérant un filtrage de l'information. On ne retient de la variabilité des variables explicatives que la partie qui correspond à des chocs non corrélés avec la perturbation. Ce filtrage est opéré en projetant les variables explicatives sur un ensemble de variables non corrélées avec la perturbation. La condition de rang garantit que l'on ne perd pas le minimum d'information requis pour identifier le paramètre.

On voit aussi que dans cette opération de filtrage on perd de l'information et que cette perte d'information conduit à une moins grande précision de l'estimateur :

$$V_{as}\left(\widehat{b}_{2mc}\right) = p\lim \sigma^2((P_z\underline{x})'\,P_z\underline{x}/N)^{-1} \gtrsim \sigma^2(\underline{x}'\underline{x}/N)^{-1} = V_{as}\left(\widehat{b}_{mco}\right)$$

La précision de l'estimateur à variables instrumentales ne peut donc dépasser celle qu'aurait l'estimateur des mco si les variables explicatives étaient exogènes. On voit que lorsque la dimension de l'espace sur lequel on projette augmente, la précision de l'estimateur à variables instrumentales s'accroît. A la limite, si la taille de l'espace sur lequel on projette augmente suffisamment, on retrouve la précision de l'estimateur des mco, mais alors on retrouve aussi l'estimateur des mco. Dans la décision d'introduire ou non telle ou telle variable dans la liste des variables instrumentales, il y a donc un arbitrage entre précision de l'estimateur et convergence de l'estimateur : plus il y a de variables instrumentales plus l'estimateur est précis, mais plus les risques de biais sont importants.

10.3.3 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

$$\begin{aligned} V_{as,het}(A^*) &= \left[A^*E\left(z_i'x_i\right)\right]^{-1} A^*E\left(u_i^2z_i'z_i\right)A^{*\prime}\left[E\left(x_i'z_i\right)A^{*\prime}\right]^{-1} \\ &= \left[E\left(x_i'z_i\right)E\left(z_i'z_i\right)^{-1}E\left(z_i'x_i\right)\right]^{-1} E\left(x_i'z_i\right)E\left(z_i'z_i\right)^{-1} \\ &\quad E\left(u_i^2z_i'z_i\right)E\left(z_i'z_i\right)^{-1}E\left(z_i'x_i\right)\left[E\left(x_i'z_i\right)E\left(z_i'z_i\right)^{-1}E\left(z_i'x_i\right)\right]^{-1} \\ &= E\left(\widetilde{x}_i'\widetilde{x}_i\right)^{-1}E\left(u_i^2\widetilde{x}_i'\widetilde{x}_i\right)E\left(\widetilde{x}_i'\widetilde{x}_i\right)^{-1} \end{aligned}$$

où $\widetilde{x}_i = z_iE\left(z_i'z_i\right)^{-1}E\left(z_i'x_i\right)$.

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

$$V_{het}\left(\widehat{b}_{2mc}\right) = V_{as,het}(A^*)/N$$

Elle peut être estimée par

$$\widehat{V}_{het}\left(\widehat{b}_{2mc}\right) = \frac{V_{as,het}(A^*)}{N} = \left(\overline{\widehat{\widetilde{x}}_i\widehat{\widetilde{x}}_i'}\right)^{-1}\left(\sum_{i=1}^{N}\widehat{u}_i^2\widehat{\widetilde{x}}_i\widehat{\widetilde{x}}_i'\right)\left(\sum_{i=1}^{N}\widehat{\widetilde{x}}_i\widehat{\widetilde{x}}_i'\right)^{-1}$$

où $\widehat{\widetilde{x}}_i = z_i\left(\overline{z_i'z_i}\right)^{-1}\left(\overline{z_i'x_i}\right)$ qui est exactement la matrice de White.

10.4 Interprétation de la condition $\operatorname{rang} E\left(z_i' x_i\right) = K+1$

La mise en oeuvre de la méthode des variables instrumentales repose sur la condition $\operatorname{rang} E\left(z_i' x_i\right) = K+1$. Les variables du modèle sont scindées en K_1 variables endogènes x_{1i} et K_2+1 variables exogènes. 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 $\operatorname{rang} E\left(z_i' x_i\right) = K+1$ est analogue à la condition $\operatorname{rang} E\left(z_i' z_i\right)^{-1} E\left(z_i' x_i\right) = K+1$. Cette matrice correspond à la matrice des coefficients des régressions des variables explicatives sur les instruments. Comme les variables du modèle et les instruments 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 $\operatorname{rang} E\left(z_i' z_i\right)^{-1} E\left(z_i' x_i\right) = K+1$ est donc équivalente à la condition

$$\operatorname{rang} \Gamma_{1\widetilde{z}} = K_1$$

Cette condition s'interprète comme le fait que les variables instrumentales extérieures *expliquent suffisamment bien* les variables endogènes. Il n'existe pas de test formel de cette condition qui puisse être facilement mis en oeuvre. Néanmoins il est important de regarder la façon dont les variables instrumentales expliquent les variables endogènes, même si on peut mettre en oeuvre l'estimateur des doubles moindres carrés directement sans faire cette régression intermédiaire. On peut par exemple, bien que cela ne garantisse pas que la condition est satisfaite dès 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 $\operatorname{rang} E\left(z_i' x_i\right) = K+1$ n'est pas satisfaite, on aura néanmoins en général à distance finie $\operatorname{rang} \overline{z_i' x_i} = K+1$ et l'estimateur pourra être numériquement mis en oeuvre. La conséquence du fait que $\operatorname{rang} E\left(z_i' x_i\right) < K+1$ est que

$$\underline{x}' \underline{z} \left(\underline{z}' \underline{z}\right)^{-1} \underline{z}' \underline{x} \to E\left(x_i' z_i\right) E\left(z_i' z_i\right)^{-1} E\left(z_i' x_i\right)$$

non inversible. L'estimateur sera donc très instable et présentera des écarts-type très élevés sur certains coefficients, à l'instar de ce qui se produit avec les mco dans le cas de multicolinéarité.

Lorsque l'on est à la limite de cette situation, c'est à dire lorsque l'on dispose de variables instrumentales expliquant très mal les variables endogènes on parle d'*instruments faibles*.

On peut être tenté de pallier ce manque de pouvoir explicatif des instruments par l'utilisation d'un grand nombre d'entre eux : on est dans la situation ou il y a beaucoup de variables instrumentales mais ou toutes, prises ensemble ont un pouvoir explicatif faible. Cette situation présente des effets indésirables dont on peut avoir facilement l'intuition. Lorsque le nombre d'instruments sur lequel on projette les variables devient grand et mécaniquement, sans que cela résulte d'une propriété statistique, la prédiction de la variable explicative va devenir meilleure : elle va se rapprocher des variables explicatives simplement parce que l'espace sur lequel on projette devient plus grand. On comprend alors que dans ce cas l'estimateur à variables instrumentales se rapproche de l'estimateur des mco. L'utilisation d'un grand nombre de variables instrumentales au pouvoir explicatif médiocre est donc une situation peu souhaitable. On considère pour s'en prémunir qu'il faut que le F de Fisher testant la nullité globale des coefficients des variables instrumentales dans la régression des variables explicatives endogènes soit plus grand que 1.

10.5 Test de suridentification

En pratique, on est souvent amené à effectuer des estimations d'une même équation en étendant ou restreignant la liste des variables instrumentales. On a vu en effet que l'on pouvait avoir intérêt à accroître le nombre de variables instrumentales dans la mesure où cela conduit à des estimateurs plus précis. On a vu aussi qu'accroître indûment l'ensemble des variables instrumentales pouvait conduire à faire apparaître des biais dans l'estimation. On va présenter dans cette section un test très important et très couramment utilisé permettant de contrôler qu'il n'y a pas d'incohérence dans le choix des variables instrumentales. Ce test, appelé test de Suridentification, ou test de Sargan constitue un guide incontournable dans le choix des variables instrumentales. On présente d'abord l'idée et le sens du test de Sargan d'une façon informelle, on aborde ensuite la question plus formellement et de façon plus pratique.

10.5.1 Idée du test

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 très 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

$$b = (AE(z_i'x_i))^{-1} AE(z_i'y_i)$$

Cette dernière relation doit être vraie pour toute matrice A telle que $AE(z_i'x_i)$ 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 $\widehat{b}_{mci}(A)$ doivent converger vers une même valeur.

Considérons par exemple le cas d'un modèle ne présentant qu'une variable explicative et pour lequel il existe h variables instrumentales. On pourrait considérer h estimateurs à variables instrumentales obtenus en utilisant à chaque fois une seule des variables instrumentales.

$$\widehat{b}_{VI}(k) = \frac{\overline{z_i(k)\, y_i}}{\overline{z_i(k)\, x_i}}$$

Si toutes ces variables sont compatibles entre elles, les estimateurs obtenus doivent tous être proches les uns des autres on doit avoir $p\lim \widehat{b}_{VI}(k)$ indépendant de k. L'idée du test de suridentification est de comparer entre eux les différents estimateurs et de juger s'ils sont ou non proches. Ceci constitue l'idée du test de suridentification, cela ne représente nullement la façon dont on le met en oeuvre. On va voir ultérieurement une procédure permettant de tester directement l'hypothèse que pour un jeu de variables instrumentales donné l'ensemble des estimateurs $\widehat{b}_{mci}(A)$ convergent tous vers la même valeur, sans avoir à calculer tous ces estimateurs.

Remarquons que ce test n'est pas à proprement parlé un test de validité des instruments mais un test de compatibilité des instruments. Il signifie en effet uniquement $\exists\ \widetilde{b}$ tq $\widehat{b}_{mci}(A) \to \widetilde{b}$. Ceci est une propriété statistique des données, qui peut être testée. Il ne signifie pas néanmoins $\widehat{b}_{mci}(A) \to \widetilde{b} = b$ le paramètre structurel que l'on souhaite identifier.

10.5.2 Approche formelle

La convergence de chaque estimateur des moindres carrés indirects provient de la propriété $E(z_i'u_i) = 0$. C'est elle que l'on pourrait souhaiter tester directement. Il s'agirait du test de l'hypothèse nulle

$$H_{00} : E(z_i'u_i) = 0$$

Si le résidu était connu un tel test serait très 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.

Néanmoins comme on l'a dit, le test que l'on peut espérer mettre en oeuvre n'est pas le test de H_{00}, soit le test de

$$H_{00} : E(z_i'(y_i - x_i b)) = 0$$

ou b est le paramètre structurel mais simplement le test de

$$\exists \widetilde{b} \text{ tq } E\left(z_i'\left(y_i - x_i\widetilde{b}\right)\right) = 0$$

Il est clair que sous H_{00} : $p\lim \widehat{b}_{2mc} = \widetilde{b}$ et donc que la façon naturelle de tester une telle hypothèse est d'examiner si $\overline{z_i'\widehat{u}_i}$ est proche de zéro.

Remarque *1. Sous l'hypothèse H_{00} on aurait donc en appliquant le théorème centrale limite, et compte tenu de l'hypothèse d'homoscédasticité*

$$\sqrt{N}\overline{z_i'u_i} \to N\left(0, \sigma^2 E\left(z_i'z_i\right)\right)$$

et donc

$$\frac{N}{\sigma^2}\overline{z_i'u_i}' E\left(z_i'z_i\right)^{-1}\overline{z_i'u_i} \to \chi^2\left(\dim\left(z_i\right)\right)$$

ou encore

$$\frac{N}{\widehat{\sigma}^2}\overline{z_i'u_i}'\overline{z_i'z_i}^{-1}\overline{z_i'u_i} \to \chi^2\left(\dim\left(z_i\right)\right)$$

2. 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

$$V\left(W\right)V\left(W\right)^{-}V\left(W\right) = V\left(W\right)$$

Ici on ne peut pas utiliser u_i le résidu "structurel" mais $\widehat{u}_i$.

La statistique de test va rester la même à ceci prêt que :

1. on utilise $\widehat{u}_i$ et non u_i
2. le nombre de degrés de liberté n'est pas le nombre de variables instrumentales $\dim\left(z_i\right) = H + 1$, mais $H - K$, c'est à dire le nombre d'instruments en excès.

Ce dernier point exprime bien le fait qu'une partie des conditions d'orthogonalité est mobilisée pour identifier le paramètre et illustre bien le nom de suridentification donné au test.

Proposition *Sous les hypothèses de régularité garantissant la convergence et la normalité asymptotique de l'estimateur à variables instrumentales, dans le cas de résidus homoscédastiques ($V\left(\left(y_i - x_i\widetilde{b}\right)^2 | z_i\right) = \sigma^2$),*

Sous H_0 : $\exists \widetilde{b} tq E\left(z_i'\left(y_i - x_i\widetilde{b}\right)\right) = 0$, la statistique $\widehat{S}$

$$\widehat{S} = N\overline{z_i'\widehat{u}_i}'\frac{\left(\overline{z_i'z_i}\right)^{-1}}{\widehat{\sigma}^2}\overline{z_i'\widehat{u}_i} \to \chi^2\left(H - K\right)$$

où $\widehat{u}_i = y_i - x_i\widehat{b}_{2mc}$ *et* $\widehat{\sigma}^2 = \overline{\widehat{u}_i^2}$.

Le test de H_0 *contre* $H_1 : \nexists \widetilde{b}$ *tq* $E\left(z_i'\left(y_i - x_i\widetilde{b}\right)\right) = 0$ *basé sur la région critique*

$$W = \left\{\widehat{S}\,\middle|\,\widehat{S} > q\left(1-\alpha, \chi^2\left(H-K\right)\right)\right\}$$

où $q\left(1-\alpha, \chi^2\left(H-K\right)\right)$ *est le quantile d'ordre* $1-\alpha$ *d'un* $\chi^2\left(H-K\right)$ *est un test convergent au seuil* α.

Démonstration *Sous* H_0, *soit* $\widetilde{b}$ *la valeur du paramètre telle que* $E\left(z_i'\left(y_i - x_i\widetilde{b}\right)\right) = 0$ *et soit* $\widetilde{u}_i$ *le résidu correspondant. Ces grandeurs sont a priori distinctes rappelons le des quantités ayant sens sur le plan économique* $\widetilde{b}$ *et* u_i. *Néanmoins, pour ne pas alourdir, on les note* b *et* u_i,

$$\widehat{u}_i = y_i - x_i\widehat{b}_{2mc} = x_i b + u_i - x_i\widehat{b}_{2mc} = u_i - x_i\left(\widehat{b}_{2mc} - b\right)$$

d'où

$$\overline{z_i'\widehat{u}_i} = \frac{1}{N}\underline{z}'\underline{\widehat{u}} = \overline{z_i'u_i} - \overline{z_i'x_i}\left(\widehat{b}_{2mc} - b\right)$$

comme

$$\begin{aligned}\widehat{b}_{2mc} &= \left(\overline{x_i'z_i}\ \ \overline{z_i'z_i}^{-1}\overline{z_i'x_i}\right)^{-1}\overline{x_i'z_i}\ \overline{z_i'z_i}^{-1}\overline{z_i'y_i} \\ &= b + \left(\overline{x_i'z_i}\ \ \overline{z_i'z_i}^{-1}\overline{z_i'x_i}\right)^{-1}\overline{x_i'z_i}\ \overline{z_i'z_i}^{-1}\overline{z_i'u_i}\end{aligned}$$

on a :

$$\overline{z_i'\widehat{u}_i} = \left(I_{H+1} - \overline{z_i'x_i}\left(\overline{x_i'z_i}\ \ \overline{z_i'z_i}^{-1}\overline{z_i'x_i}\right)^{-1}\overline{x_i'z_i}\ \overline{z_i'z_i}^{-1}\right)\overline{z_i'u_i} = \left(I_{H+1} - M_N\right)\overline{z_i'u_i}$$

et $M_N \to M = E\left(z_i'x_i\right)\left(E\left(x_i'z_i\right)\ \ E\left(z_i'z_i\right)^{-1}E\left(z_i'x_i\right)\right)^{-1}\ E\left(x_i'z_i\right)\ \ E\left(z_i'z_i\right)^{-1}$. M *vérifie en outre* $M^2 = M$

On en déduit que

$$\sqrt{N}\overline{z_i'\widehat{u}_i} = \left(I_{H+1} - M\right)\sqrt{N}\overline{z_i'u_i} + o_p\left(1\right) \xrightarrow{L} N\left(0, \Sigma\right)$$

avec $\Sigma = \left(I_{H+1} - M\right)V\left(z_i'u_i\right)\left(I_{H+1} - M'\right) = \sigma^2\left(I_{H+1} - M\right)E\left(z_i'z_i\right)\left(I_{H+1} - M'\right)$. *On vérifie que* $\left(I_{H+1} - M\right)\ E\left(z_i'z_i\right) = \ E\left(z_i'z_i\right)\ \left(I_{H+1} - M'\right)$ *si bien que* $V_{as}\left(\sqrt{N}\overline{z_i'\widehat{u}_i}\right) = \sigma^2\left(I_{H+1} - M\right)\ E\left(z_i'z_i\right)$

Comme $M^2 = M$ *on vérifie immédiatement que* $MV_{as}\left(\sqrt{N}\overline{z_i'\widehat{u}_i}\right) = 0$ *et donc que* $V_{as}\left(\sqrt{N}\overline{z_i'\widehat{u}_i}\right)$ *n'est pas de plein rang. Comme* $V_{as}\left(\sqrt{N}\overline{z_i'\widehat{u}_i}\right) = \sigma^2\left(I_{H+1} - M\right)E\left(z_i'z_i\right)$,

le rang de $V_{as}\left(\sqrt{N}\overline{z_i'\widehat{u}_i}\right)$ *est clairement celui de* $(I_{H+1}-M)$ *et comme* $M^2=M$, *les valeurs propres de* M *sont soit* 1 *soit* 0. *On en déduit que*

$$\begin{aligned} rang V_{as}\left(\sqrt{N}\overline{z_i'\widehat{u}_i}\right) &= Tr\left(I_{H+1}-M\right)=rang\left(I_{H+1}-M\right)=H+1-Tr\left(M\right) \\ &= H+1- \\ &\quad Tr\left(E\left(z_i'x_i\right)\left(E\left(x_i'z_i\right)\ E\left(z_i'z_i\right)^{-1}E\left(z_i'x_i\right)\right)^{-1}E\left(x_i'z_i\right)\ E\left(z_i'z_i\right)^{-1}\right) \\ &= H+1- \\ &\quad Tr\left(\left(E\left(x_i'z_i\right)\ E\left(z_i'z_i\right)^{-1}E\left(z_i'x_i\right)\right)^{-1}E\left(x_i'z_i\right)\ E\left(z_i'z_i\right)^{-1}E\left(z_i'x_i\right)\right) \\ &= H-K \end{aligned}$$

On a aussi

$$\begin{aligned} V_{as}\left(\sqrt{N}\overline{z_i'\widehat{u}_i}\right)\frac{1}{\sigma^2}E\left(z_i'z_i\right)^{-1}V_{as}\left(\sqrt{N}\overline{z_i'\widehat{u}_i}\right) &= \sigma^2\left(I_{H+1}-M\right)E\left(z_i'z_i\right) \\ &\quad \frac{1}{\sigma^2}E\left(z_i'z_i\right)^{-1}\sigma^2\left(I_{H+1}-M\right)E\left(z_i'z_i\right) \\ &= \sigma^2\left(I_{H+1}-M\right)\left(I_{H+1}-M\right)E\left(z_i'z_i\right) \\ &= \sigma^2\left(I_{H+1}-M\right)E\left(z_i'z_i\right) \\ &= V_{as}\left(\sqrt{N}\overline{z_i'\widehat{u}_i}\right) \end{aligned}$$

on en déduit que $\frac{1}{\sigma^2}E\left(z_i'z_i\right)^{-1}$ *est un inverse généralisé de la matrice de variance asymptotique* $\sqrt{N}\overline{z_i'\widehat{u}_i}$. *On a donc*

$$N\overline{\widehat{u}_i z_i}\frac{1}{\sigma^2}E\left(z_i'z_i\right)^{-1}\overline{z_i'\widehat{u}_i}\xrightarrow{L}\chi^2\left(H-K\right)$$

et on peut clairement remplacer en appliquant le théorème de Slutsky $E\left(z_i'z_i\right)$ *par* $\overline{z_i'z_i}$ *et* σ^2 *par* $\widehat{\sigma}^2$.

Donc, sous H_0 : $\widehat{S}=\overline{z_i'\widehat{u}_i}'\frac{\left(\overline{z_i'z_i}\right)^{-1}}{\widehat{\sigma}^2}\overline{z_i'\widehat{u}_i}\xrightarrow{L}\chi^2\left(H-K\right)$.

En outre sous H_1, $\overline{z_i'\widehat{u}_i}=\overline{z_i'\left(y_i-x_i\widehat{b}_{2mc}\right)}=\overline{z_i'\left(y_i-x_i p\lim\widehat{b}_{2mc}\right)}+o_p\left(1\right)\xrightarrow{P}\delta\neq 0$, *comme* $\frac{\left(\overline{z_i'z_i}\right)^{-1}}{\widehat{\sigma}^2}\xrightarrow{P}\Theta$ *inversible,* $\overline{z_i'\widehat{u}_i}'\frac{\left(\overline{z_i'z_i}\right)^{-1}}{\widehat{\sigma}^2}\overline{z_i'\widehat{u}_i}\xrightarrow{P}\delta'\Theta\delta$, *sous* H_1, *donc* $\widehat{S}\to\infty$ *et il en résulte que* $P\left(W\left|H_1\right.\right)\to 1$.

10.5.3 Mise en oeuvre du test

Le test de suridentification est très 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 effet, si on considère le modèle

$$\widehat{u}_i = z_i\psi + w_i$$

l'estimateur des mco de ψ est $\widehat{\psi} = \left(\overline{z_i'z_i}\right)^{-1}\overline{z_i'\widehat{u}_i}$, $V\left(\widehat{\psi}\right) = V(w_i)\left(\overline{z_i'z_i}\right)^{-1}/N$. Sous l'hypothèse $H_\psi : \psi = 0$, $V(w_i) = V(\widehat{u}_i) = \widehat{\sigma}^2$ et donc le test de $\psi = 0$ doit être mené à partir de $\widehat{\psi}'V\left(\widehat{\psi}\right)^{-1}\widehat{\psi} = \overline{\widehat{u}_i z_i}\left(\overline{z_i'z_i}\right)^{-1}\left(N\left(\overline{z_i'z_i}\right)/\widehat{\sigma}^2\right)\left(\overline{z_i'z_i}\right)^{-1}\overline{z_i'\widehat{u}_i} = N\overline{\widehat{u}_i z_i}\left(\overline{z_i'z_i}\right)^{-1}\overline{z_i'\widehat{u}_i}/\widehat{\sigma}^2$ qui est la statistique. Le test est donc formellement équivalent au test de la nullité globale des coefficients de la régression de $\widehat{u}_i$ sur les variables instrumentales z_i. On sait que ce même test peut être effectué (asymptotyiquement) à partir du R^2 de la régression. La statistique de test est NR^2 et est équivalente sous l'hypotyèse nulle au F de la régression. Le test peut donc être effectué à partir du F de cette régression. Néanmoins il convient d'être prudent en ce qui concerne le calcul de cette statistique et celui de la $p-value$. Ceci tient au nombre de degrés de liberté retenu dans le calcul. Considérons $\widehat{S}$ la statistique de test de la proposition précédente. La statistique donnée par le logiciel F_{Log} est reliée à cette statistique $\widehat{S}$ par la formule $F_{Log} = \widehat{S}/H$. On divise par H car le logiciel prend en compte le nombre de régresseurs. La $p-value$ qui accompagne le F de la régression donné dans tous les logiciels, fait l'hypothèse que cette statistique suit une loi $F(H, N-H-1)$ degrés de liberté, où H est le nombre de variables explicatives non constantes de la régression, ici on a $N \to \infty$. Pour $N \to \infty$ $F(k, N-k-1) \to \chi^2(k)/k$. La $p-value$ indiquée correspond donc à une statistique $\chi^2(H)/H$. Elle n'est donc pas correcte, la statistique non plus. On sait que $HF_{Log} \to \chi^2(H-K)$ et donc $F_{Rec} = (H/(H-K))F_{Log}$ suit une loi $F(H-K, N-(H-K)-1)$. On doit donc considérer soit la statistique HF_{Log} et calculer la p-value à partir d'une loi du $\chi^2(H-K)$, soit considérer F_{Rec} et calculer la p-value à partir d'une loi $F(H-K, \infty)$.

Remarque *1. On a a priori toujours intérêt à avoir un ensemble d'instrument 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 nécessairement moins ou aussi bon que l'estimateur des doubles moindres carrés avec l'ensemble d'instrument 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.*

2. 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

$$V(b_{mco}) = \sigma^2(\underline{x}'\underline{x})^{-1} \quad et\ V(b_{2mc}) = \sigma^2(\underline{x}'P_z\underline{x})^{-1}$$

En outre, on voit aussi en comparant les expressions des estimateurs

$$b_{mco} = (\underline{x}'\underline{x})^{-1}\,\underline{x}'\underline{y} \text{ et } b_{2mc} = (\underline{x}'P_z\underline{x})^{-1}\,\underline{x}'P_z\underline{y}$$

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.

10.6 Test d'exogénéité des variables explicatives

10.6.1 Intérêt et idée du test

Ayant estimé le modèle par les doubles moindre carrés, c'est à dire sous l'hypothèse

$$H_1 : \exists\ c\ tq\ E\left(z_i^{'}\left(y_i - x_i c\right)\right) = 0$$

on peut vouloir tester l'hypothèse que les régresseurs x_i sont exogènes.

On considère donc l'hypothèse

$$H_0 : \exists\ c\ tq\ E\left(z_i^{'}\left(y_i - x_i c\right)\right) = 0 \ \ et\ E\left(x_i^{'}\left(y_i - x_i c\right)\right) = 0$$

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. Une idée naturelle consiste à examiner si les coefficients estimés sous l'hypothèse nulle et sous l'hypothèse alternative sont identiques, c'est à dire si $p\lim \widehat{b}_{2mc} = p\lim \widehat{b}_{mco}$. Notons que là encore il ne s'agit que d'un test de compatibilité des conditions d'orthogonalité entre elles et non pas un test de leur validité dans le cadre de l'estimation d'un paramètre structurel.

10.6.2 Approche formelle

Test de Hausman

L'idée précédemment avancée de tester l'hypothèse $p\lim \widehat{b}_{2mc} = p\lim \widehat{b}_{mco}$ peut être mise en oeuvre en se fondant sur la comparaison de $\widehat{b}_{2mc} - \widehat{b}_{mco}$ avec 0. Pour faire ce test on va donc examiner $N\left(\widehat{b}_{2mc} - \widehat{b}_{mco}\right)' V_{as}\left(\widehat{b}_{2mc} - \widehat{b}_{mco}\right)^{-}\left(\widehat{b}_{2mc} - \widehat{b}_{mco}\right)$. Plusieurs questions

se posent naturellement. On a vu qu'au sein des variables explicatives x s'opérait une distinction naturelle entre les K_1 variables endogènes x_1 et les $1+K_2$ variables exogènes x_2. On peut s'attendre à ce que le test ne porte que sur les coefficients des variables potentiellement endogènes. En outre se pose les questions du rang de la matrice de variance $V_{as}\left(\widehat{b}_{2mc}-\widehat{b}_{mco}\right)$ qui conditionne le nombre de degrés de liberté de la loi limite de la statistique et de la détermination d'un inverse généralisé. On examine tour à tour chacune de ces questions.

Le test peut être basé sur les coefficients des endogènes

Lemme *On a*

$$\left(\widehat{b}_{2mc}-\widehat{b}_{mco}\right)=\left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)^{-1}\begin{pmatrix}\left(\left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)^{11}\right)^{-1}\\ 0_{K_2+1,K_1}\end{pmatrix}\left(\widehat{b}^{(1)}_{2mc}-\widehat{b}^{(1)}_{mco}\right)$$

Le test de $p\lim\widehat{b}_{2mc}=p\lim\widehat{b}_{mco}$ *est identique à celui de* $p\lim\widehat{b}^{(1)}_{2mc}=p\lim\widehat{b}^{(1)}_{mco}$. *En outre*

$$\widehat{b}^{(1)}_{2mc}-\widehat{b}^{(1)}_{mco}=\left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)^{11}\widehat{\underline{x}}'_1M_x\underline{y}$$

Démonstration *En effet* $\widehat{b}_{2mc}=\left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)^{-1}\widehat{\underline{x}}'\underline{y}$ *et* $\widehat{b}_{mco}=\left(\underline{x}'\underline{x}\right)^{-1}\underline{x}'\underline{y}$ *donc*

$$\begin{aligned}\widehat{\underline{x}}'\widehat{\underline{x}}\left(\widehat{b}_{2mc}-\widehat{b}_{mco}\right) &= \widehat{\underline{x}}'\widehat{\underline{x}}\left[\left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)^{-1}\widehat{\underline{x}}'\underline{y}-\left(\underline{x}'\underline{x}\right)^{-1}\underline{x}'\underline{y}\right]\\ &= \left[\widehat{\underline{x}}'\underline{y}-\widehat{\underline{x}}'\widehat{\underline{x}}\left(\underline{x}'\underline{x}\right)^{-1}\underline{x}'\underline{y}\right]\\ &= \left[\widehat{\underline{x}}'\underline{y}-\widehat{\underline{x}}'\underline{x}\left(\underline{x}'\underline{x}\right)^{-1}\underline{x}'\underline{y}\right]=\widehat{\underline{x}}'M_x\underline{y}\end{aligned}$$

Puisque $\widehat{\underline{x}}'\widehat{\underline{x}}=\left(P_z\underline{x}\right)'\left(P_z\underline{x}\right)=\underline{x}'P_zP_z\underline{x}=\left(P_z\underline{x}\right)'\underline{x}=\widehat{\underline{x}}'\underline{x}$ *et avec* $M_x=I_N-\underline{x}\left(\underline{x}'\underline{x}\right)^{-1}\underline{x}'$.

Comme $x_2\in z$, $\widehat{\underline{x}}_2=\left(P_z\underline{x}_2\right)=\underline{x}_2$ *et donc* $\widehat{\underline{x}}'_2M_x=\underline{x}'_2M_x=0$.

$$\widehat{\underline{x}}'\widehat{\underline{x}}\left(\widehat{b}_{2mc}-\widehat{b}_{mco}\right)=\begin{pmatrix}\widehat{\underline{x}}'_1M_x\underline{y}\\ 0\end{pmatrix}$$

Dont on en déduit que

$$\left(\widehat{b}_{2mc}-\widehat{b}_{mco}\right)=\left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)^{-1}\begin{pmatrix}\widehat{\underline{x}}'_1M_x\underline{y}\\ 0\end{pmatrix}$$

soit, avec $b^{(1)}$ *le vecteurs des coefficients de* x_{1i} *et symétriquement pour* $b^{(2)}$, *et les notations standards*

$$\begin{bmatrix}\left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)_{11} & \left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)_{12}\\ \left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)_{21} & \left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)_{22}\end{bmatrix}^{-1}=\begin{bmatrix}\left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)^{11} & \left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)^{12}\\ \left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)^{21} & \left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)^{22}\end{bmatrix}$$

$$\left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)_{21}\left(\widehat{b}_{2mc}^{(1)}-\widehat{b}_{mco}^{(1)}\right)+\left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)_{22}\left(\widehat{b}_{2mc}^{(2)}-\widehat{b}_{mco}^{(2)}\right)=0$$

et

$$\widehat{b}_{2mc}^{(1)}-\widehat{b}_{mco}^{(1)}=\left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)^{11}\widehat{\underline{x}}_1'M_x\underline{y}$$

Le test de $p\lim\widehat{b}_{2mc}=p\lim\widehat{b}_{mco}$ *et donc équivalent à celui de* $p\lim\widehat{b}_{2mc}^{(1)}=p\lim\widehat{b}_{mco}^{(1)}$. *Ce test peut en outre être pratiqué à partir de l'expression* $\widehat{b}_{2mc}^{(1)}-\widehat{b}_{mco}^{(1)}=\left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)^{11}\widehat{\underline{x}}_1'M_x\underline{y}$

Rang de la matrice de variance de $\widehat{b}_{2mc}^{(1)}-\widehat{b}_{mco}^{(1)}$

Lemme *Sous l'hypothèse* $\operatorname{rang}\left(\underline{z}'\underline{x}\right)=K+1$, *le rang de la matrice de variance de* $\widehat{b}_{2mc}^{(1)}-\widehat{b}_{mco}^{(1)}$ *est* K_1, *le nombre de variables explicatives endogènes.*

Démonstration *L'expression précédente montre que la matrice de variance de* $\widehat{b}_{2mc}^{(1)}-\widehat{b}_{mco}^{(1)}$ *est* $\sigma^2\left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)^{11}\widehat{\underline{x}}_1'M_x\widehat{\underline{x}}_1\left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)^{11}$. *Son rang est donc égal à celui de* $\widehat{\underline{x}}_1'M_x\widehat{\underline{x}}_1$, *donc à celui de* $M_x\widehat{\underline{x}}_1$. *Supposons que l'on ait pour un vecteur* $\lambda: M_x\widehat{\underline{x}}_1\lambda=0$ *alors* $P_x\widehat{\underline{x}}_1\lambda=\widehat{\underline{x}}_1\lambda$ *il existe donc un vecteur* μ *tel que* $\widehat{\underline{x}}_1\lambda=\underline{x}\mu$. *Comme* $\widehat{\underline{x}}_1$ *appartient à l'espace engendré par* $\underline{z}=\left[\widetilde{\underline{z}},\ \underline{x}_2\right]$, *la combinaison linéaire* $\underline{x}\mu$ *est nécessairement une combinaison linéaire des seules variables explicatives exogènes :* $\underline{x}\mu=\underline{x}_2\mu_2$. *Notant comme précédemment* $\Gamma_1=\left[\Gamma_{1\widetilde{z}},\Gamma_{1x_2}\right]$, *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. L'équation* $\widehat{\underline{x}}_1\lambda=\underline{x}_2\mu_2$, *s'écrit* $\widetilde{\underline{z}}\Gamma_{1\widetilde{z}}\lambda+\underline{x}_2\left(\Gamma_{1x_2}\lambda-\mu_2\right)=0$. *Comme* Z *est de rang* $K+1$ *ceci nécessite* $\Gamma_{1\widetilde{z}}\lambda=0$. *Et on a vu que la condition* $\operatorname{rang}\left(\underline{z}'\underline{x}\right)=K+1$ *est 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* $\widehat{b}_{2mc}^{(1)}-\widehat{b}_{mco}^{(1)}$ *est inversible : le nombre de degrés de liberté du test d'exogénéité est égal à* K_1.

Le test de Hausman

Proposition *Lorsque l'hypothèse d'homoscédasticité,* $E(u_i^2|x_i,z_i)=\sigma^2$ *est satisfaite, sous l'hypothèse nulle d'exogénéité de* x_i, *la statistique*

$$\widehat{S}=\frac{N}{\widehat{\sigma}^2}\left(\widehat{b}_{2mc}^{(1)}-\widehat{b}_{mco}^{(1)}\right)'\left[\left(\frac{\widehat{\underline{x}}'\widehat{\underline{x}}}{N}\right)^{11}-\left(\frac{\underline{x}'\underline{x}}{N}\right)^{11}\right]^{-1}\left(\widehat{b}_{2mc}^{(1)}-\widehat{b}_{mco}^{(1)}\right)\xrightarrow{L}\chi^2\left(K_1\right)$$

Un test convergent au niveau α *de* H_0 *peut être effectué en comparant la valeur de la statistique* $\widehat{S}$ *au quantile d'ordre* $1-\alpha$ *d'une loi du* χ^2 *à* K_1 *degrés de liberté*

Démonstration *Sous l'hypothèse d'homoscédasticité et sous l'hypothèse nulle,* $\widehat{b}_{mco}$ *est l'estimateur de variance minimale dans la classe des estimateurs sans biais dont fait parti l'estimateur des doubles moindres carrés. On a donc*

$$V_{as}\left(\widehat{b}_{2mc}-\widehat{b}_{mco}\right)=V_{as}\left(\widehat{b}_{2mc}\right)-V_{as}\left(\widehat{b}_{mco}\right)$$

Un estimateur convergent de la matrice de variance de la différence $\widehat{b}_{2mc} - \widehat{b}_{mco}$ *est donc donné par*

$$\widehat{V}_{as}\left(\widehat{b}_{2mc} - \widehat{b}_{mco}\right) = \widehat{\sigma}^2 \left[\frac{\left(\widehat{\underline{x}}'\widehat{\underline{x}}\right)^{11}}{N} - \frac{\left(\underline{x}'\underline{x}\right)^{11}}{N} \right]$$

On en déduit que $\widehat{S}$ *suit une loi du* χ^2 *à* K_1 *degrés de liberté sous* H_0*. Sous l'hypothèse alternative* $p\lim \widehat{b}_{2mc}^{(1)} - p\lim \widehat{b}_{mco}^{(1)} \neq 0$ *et donc* $\widehat{S} \to +\infty$

Test d'exogénéité par le biais de la régression augmentée

Le test d'exogénéité peut être mis en oeuvre très simplement par le biais d'une simple régression de la variable dépendante $\underline{y}$ sur les variables potentiellement endogènes du modèle et les variables exogènes $\underline{x}_1$ et $\underline{x}_2$ et sur la projection des variables endogènes sur les variables instrumentales $\widehat{\underline{x}}_1$:

$$\underline{y} = \underline{x}_1 c_1 + \underline{x}_2 c_2 + \widehat{\underline{x}}_1 \gamma + \underline{w}$$

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 $\widehat{\underline{x}}_1$ sur les autres variables, c'est à dire $\underline{x}$. On a donc

$$\widehat{\gamma} = \left(\widehat{\underline{x}}_1' M_x \widehat{\underline{x}}_1\right)^{-1} \widehat{\underline{x}}_1' M_x \underline{y}$$

or on a vu précédemment

$$\widehat{b}_{2mc}^{(1)} - \widehat{b}_{mco}^{(1)} = \left(\widehat{\underline{x}}\widehat{\underline{x}}\right)^{11} \widehat{\underline{x}}_1' M_x \underline{y}$$

On en déduit que l'on a :

$$\widehat{b}_{2mc}^{(1)} - \widehat{b}_{mco}^{(1)} = \left(\widehat{\underline{x}}\widehat{\underline{x}}\right)^{11} \left(\widehat{\underline{x}}_1' M_x \widehat{\underline{x}}_1\right) \widehat{\gamma}$$

le test de $p\lim \widehat{b}_{2mc}^{(1)} - p\lim \widehat{b}_{mco}^{(1)} = 0$ est donc équivalent au test de $\gamma = 0$ et peut être effectué à partir de l'estimateur $\widehat{\gamma}$. Il peut donc être effectué très simplement par l'intermédiaire d'un test de Wald ou d'un test de Fisher.

Remarquons enfin que le test peut être mené de façon analogue sur les résidus des régressions des variables explicatives endogènes sur les instruments $\varepsilon\left(\underline{x}_1\right) = \underline{x}_1 - \widehat{\underline{x}}_1$. L'équation

$$\underline{y} = \underline{x}_1 c_1 + \underline{x}_2 c_2 + \widehat{\underline{x}}_1 \gamma + \underline{w}$$

se réécrit de façon analogue comme

$$\underline{y} = \underline{x}_1 \left(c_1 + \gamma\right) + \underline{x}_2 c_2 - \varepsilon\left(\underline{x}_1\right) \gamma + \underline{w}$$

10.7 Illustrations

10.7.1 Réduction du temps de travail et gains de productivité

Une des questions importantes dans l'effet du passage à 35 heures sur l'économie est son effet sur les gains de productivité. Par exemple si on considère que la production reste inchangée, l'effet sur l'emploi sera important si il y a peu de gains de productivité. Les résultats présentés ici ne sont qu'illustratifs et ne prétendent pas donner un avis sur la réduction du temps de travail. Ils montrent néanmoins si besoin était que l'économétrie permet de répondre à des questions importantes et soulignent le rôle essentiel des hypothèses identificatrices dans la réponse que l'on apporte. On peut considérer le modèle suivant :

$$\Delta \ln y_i = \alpha \Delta \ln l_i + (1-\alpha) \Delta \ln k_i + \gamma RTT_i + u_i$$

où Δl_i représente l'évolution des effectifs entre 1997 et 2000, Δk_i celle capital et RTT_i une indicatrice correspondant au fait que l'entreprise ait signé un accord de réduction du temps de travail sur cette période. u_i représente un choc de productivité ou de demande. Ce modèle est structurel, c'est à dire que l'on s'intéresse à l'effet de la réduction du temps de travail et des variations des facteurs sur l'activité. Dans un tel contexte il est clair que d'importants problèmes d'endogénéité des facteurs se posent : $\Delta \ln l_i$ en premier lieu, mais aussi $\Delta \ln k_i$ sont susceptibles d'incorporer les nouvelles conditions d'activité u_i : ces variables sont très certainement endogènes. La variable RTT_i est, elle aussi, probablement endogène : toutes les entreprises sont sensées passer à terme à 35 heures. Les entreprises ayant de bonnes perspectives de productivité peuvent plus facilement et donc plus rapidement trouver un moyen avantageux de le faire. Ceci a pu être particulièrement vrai dans le contexte de la fin des années 1990 où après une longue stagnation, la croissance qui avait déjà soutenu longuement l'activité aux USA, arrivait en France. Compte tenu des déclarations des dirigeants politiques, il n'y avait aucun doute qu'un jour ou l'autre il faudrait passer aux 35 heures. La question n'était donc pas faut-il ou non passer à la réduction du temps de travail, mai quand faut-il passer aux 35 heures. Pour se concentrer sur l'effet de la RTT on élimine le problème de l'estimation du paramètre α en le mesurant comme la part des salaires dans la valeur ajoutée dans le secteur. L'équation s'écrit alors :

$$\Delta \ln y_i - \alpha \Delta \ln l_i - (1-\alpha) \Delta \ln k_i = \Delta PGF_i = \gamma RTT_i + u_i$$

Pour atténuer l'endogénéité potentielle de la variable RTT_i on peut introduire certains régresseurs X_i : le secteur, la taille, la part des salaires dans la valeur ajoutée, la structure des qualifications... Le modèle s'écrit alors

$$\Delta PGF_i = X_i b + \gamma RTT_i + v_i$$

où v_i représente le choc de productivité résiduel, c'est à dire une fois pris en compte les facteurs X_i.

Pour estimer ce modèle on fait l'hypothèse que les aides potentiellement reçue par les entreprises si elles signent un accord de réduction du temps de travail $Aide_i$ affectent sa décision de passer à la RTT, mais pas les gains de productivité. On considère aussi que l'information dont disposaient les entreprises sur la réduction du temps de travail affecte la décision de passage mais pas la productivité. On mesure cette variable par la part des entreprises Robien dans le secteur Inf_i. On considère de même que le taux d'endettement des entreprises affecte la décision de passage mais pas la productivité $Endt_i$. Enfin, on considère que la part des femmes dans l'entreprise Pf_i n'affecte pas les gains de productivité mais influence la décision de passage. On a ainsi quatre variables instrumentales potentielles : $Aide_i$, Inf_i, $Endt_i$ et Pf_i.

On vérifie d'abord la condition de rang en régressant la variable RTT_i sur X_i et les variables instrumentales. On voit clairement sur le tableau 10.2 que les coefficients des variables instrumentales sont significatifs ce qui garantit que la condition de rang soit satisfaite.

Le tableau 10.3 donnent le résultat des estimations par les mco et par les variables instrumentales. On ne fait figurer que la variable RTT, mais les régressions comprennent toutes les variables de contrôle qui figurent dans le tableaux précédent. On observe deux résultats importants sur ces tableaux : d'une part les coefficients estimés pour la variable RTT sont très différents suivant la méthode d'estimation. Dans le cas mco on a -0.036, ce qui signifie que les entreprises ayant signé un accord de réduction du temps de travail on vu leur production baisser de 3.6% à facteurs inchangés. Pour une baisse de 10.3% (4/39) de la durée du travail, c'est assez peu et cela correspondrait à l'existence d'important gains de productivité dans les entreprises passées aux 35 heures. Le coefficient estimé par la méthode des variables instrumentales est très différent. Il est de -0.107 ce qui correspond à une baisse de la production de 10.7%. Ceci signifierait qu'il n'y a pas eu de gains de productivité associés au passage à 35 heures. On voit donc que la conclusion à laquelle on parvient dépend très fortement des hypothèses identificatrices effectuées. Un autre enseignement des deux tableaux est la différence importante entre les écarts-type estimés : l'écart-type est de 0.003 pour la régression par les mco et de 0.032 pour les variables instrumentales. Il y a donc un facteur 10 dans la précision. Il faudrait pour obtenir un estimateur aussi précis que celui des mco multiplié la taille de l'échantillon par 100 ! Les régressions présentées sont effectuées sur 30.000 observations. On voit donc clairement le prix des variables instrumentales en terme de précision.

Le tableaux 10.4 présentent le test de suridentification. Il est réalisé en régressant le résidu de la régression à variable instrumentale sur les variables exogènes du modèle c'est à dire les instruments et les régresseurs exogènes. On récupère le F de cette régression donné par le logiciel F_{Log}, et on applique la correction présentée $F_{Rec} = (H/(H-K))\, F_{Log}$. Ici H est le nombre de variables exogènes (régresseurs plus instruments) i.e. $H = 25$, et K

Régression de la variable RTT sur les instruments

Instruments	parametres	écart-type	Student	Pvalue
un	-0.801	0.082	-9.732	0.000
N16b	0.163	0.017	9.868	0.000
N16c	0.205	0.015	13.568	0.000
N16d	0.065	0.032	2.034	0.042
N16e	0.027	0.014	1.940	0.052
N16f	0.055	0.012	4.448	0.000
N16g	0.510	0.053	9.680	0.000
N16h	0.096	0.014	7.072	0.000
N16j	0.119	0.011	10.544	0.000
N16k	-0.014	0.015	-0.945	0.344
N16n	0.167	0.013	12.483	0.000
taille1	-0.240	0.027	-8.856	0.000
taille2	-0.187	0.027	-6.909	0.000
taille3	-0.164	0.027	-6.011	0.000
taille4	-0.077	0.032	-2.433	0.015
eja1	0.413	0.037	11.203	0.000
eja2	0.211	0.026	8.132	0.000
eja3	0.294	0.031	9.508	0.000
ejq1	0.022	0.018	1.209	0.227
ejq2	0.000	0.019	0.021	0.983
pi97	-0.031	0.014	-2.223	0.026
Taux d'endettement	**0.013**	**0.006**	**2.211**	**0.027**
robien	**1.466**	**0.161**	**9.095**	**0.000**
aide	**0.113**	**0.009**	**12.711**	**0.000**
part des hommes	**-0.086**	**0.015**	**-5.772**	**0.000**

TAB. 10.2 – Condition de rang

Estimation par les mco

variables	parametres	écart-type	Student	Pvalue
RTT	**-0.036**	**0.003**	**144.387**	**0.000**

Estimation par les variables instrumentales

variables	parametres	écart-type	Student	Pvalue
RTT	**-0.107**	**0.032**	**11.564**	**0.001**

TAB. 10.3 – Estimation pas les MCO et le VI

Test de Sargan

Instruments	parametres	écart-types
Taux d'endettement	**-0.00201**	**0.00329**
robien	**0.17451**	**0.06910**
aide	**-0.00826**	**0.00373**
part des hommes	**-0.00254**	**0.00753**

Statistique	degrés	p-value
7.57	3	5.6%

TAB. 10.4 – Test de Sargan

est le nombre de variables explicatives exogène et endogènes du modèle. Ici $K = 22$, la régression inclue en effet les variables de contrôle qui ne sont pas montrées ici. La correction est donc très importante puisqu'on multiplie la statistique du logiciel par $25/3 = 8.33$. Le nombre de degrés de liberté est le nombre d'instrument en excès c'est à dire 3. On voit que ce test n'est que légèrement accepté, puisque la statistique est de 7.57 ce qui conduit à une p-value de 5.6% pour 3 degrés de liberté. Notons que si on accepte l'hypothèse ($5.6\% > 5\%$ on pourrait donc accepter à la limite pour un test à 5%) ce que l'on accepte n'est pas le fait que les instruments sont valides, c'est à dire qu'ils vérifient la condition $E(z_i u_i) = 0$, autrement dit que le paramètre estimé converge vers le vrai paramètre. Ce que l'on accepte c'est que les estimateurs auxquels conduirait chacune des variables instrumentales prise séparément ne seraient pas statistiquement différents : en résumé on accepte que si il y a biais, le biais sera le même avec n'importe lequel de ces instruments. On insiste ici à dessein sur le fait qu'il s'agit d'un test de compatibilité des instruments et pas un test de validité des instruments. L'identification repose nécessairement sur une hypothèse non testable. On peut en vérifier la cohérence interne le cas échéant, c'est à dire lorsqu'il y a suridentification, mais pas la validité. Les tests de spécification sont un guide très utile mais pas une réponse définitive.

Le tableau 10.5 présente le résultat du test d'exogénéité. L'hypothèse testée est : conditionnellement au fait que l'on accepte la validité des instruments (ce qui n'a de sens que si le test de suridentification a été accepté, et qui n'est le cas qu'à 5,6% ici) peut on accepter que la variable supposée endogène est en fait exogène. C'est à dire peut on se baser sur l'estimateur des mco. La différence de précision des estimations motive de façon convaincante l'utilité de se poser cette question. Le test est effectué par le biais de la régression augmentée. On introduit la variable supposée endogène et la variable prédite par la régression de la variable endogène sur les instruments (celle du tableau1) l'hypothèse est rejetée si cette dernière variable est significative. C'est nettement le cas ici. Ce test signifie que si on croit à la validité des instruments, on ne peut pas croire à

Test d'Exogénéité

variables	parametres	écart-type	Chi2	Pvalue
RTT prédit	**-0.072**	**0.031**	**5.208**	**0.022**
RTT	**-0.036**	**0.003**	**136.164**	**0.000**

TAB. 10.5 – Test d'exogénéité

Variables Instrumentales

Variables	BIV	SBIV0	CHIBIV0	PROBBIV0
RTT	**-0.161**	**0.039**	**17.317**	**0.000**

Test de Sargan

Instruments	parametres	écart-types
Endt	**-0.0012**	**0.0033**
aide	**-0.0026**	**0.0030**
Hommes	**-0.0075**	**0.0074**

Statistique	degrés	p-value
1.152	**2**	**56.2%**

Test d'Exogénéité

variables	parametres	écart-type	Chi2	Pvalue
RTT prédit	**-0.126**	**0.038**	**10.993**	**0.001**
RTT	**-0.035**	**0.003**	**135.507**	**0.000**

TAB. 10.6 – Résultat sans la part des Robien

l'exogénéité de la variable de RTT.

Le tableau 10.6 montre le résultat des estimations lorsque l'on retire la variable Inf_i de la liste des instruments. Le hypothèse de compatibilité des variables instrumentales est beaucoup plus largement acceptée. L'hypothèse d'exogénéité est quant à elle rejetée et le coefficient estimé pour la variable de RTT est un peu modifié. Il atteint un niveau de -16%, ce qui est très élevé et signifie qu'il n'y a pas eu de gains de productivité horaire mais plutôt des pertes. Il est aussi moins précis.

10.8 Résumé

Dans ce chapitre, on a étudié

1. Différentes raisons de remettre en cause l'hypothèse identificatrice fondamentale $E(x_i' u_i) = 0$
2. Certaines variables apparaissent ainsi endogènes et d'autres restent exogènes.
3. On a montré que l'on peut recourir à des hypothèses identifiantes alternatives à celle des moindres carrés ordinaires basées sur des variables instrumentales. Il s'agit de variables corrélées avec les variables explicatives mais non corrélées avec les perturbations.
4. On a vu que parmi l'ensemble des estimateurs possibles il en existait, dans le cadre homoscédastique étudié, un plus efficace que les autres appelé estimateur à variables instrumentales.
5. Cet estimateur s'interprète comme l'estimateur obtenus en régressant la variable dépendante sur la ; projection des variables explicatives sur les variables instrumentales.
6. Cet estimateur est toujours moins précis que l'estimateur des moindres carrés ordinaires
7. On a vu un test très courant : le test de suridentification, ou test de Sargan, qui teste la compatibilité des variables instrumentales. Il ne s'agit pas d'un test de validité des instruments mais d'un test permettant de vérifier qu'il n'y a pas d'incompatibilité entre les différents instruments utilisés.
8. On a vu aussi qu'il était possible de tester l'exogénéité des variables instrumentales ce qui permet d'avoir recours, le cas échéant, à l'estimateur des moindres carrés ordinaires.

Chapitre 11

La Méthode des moments généralisée

11.1 Modèle structurel et contrainte identifiante : restriction sur les moments

Les méthodes d'estimation que l'on a vu jusqu'à présent exploitaient sans le dire explicitement l'existence de fonctions des paramètres et des variables du modèle dont l'espérance est nulle. Par exemple dans le cas du modèle linéaire vu jusqu'à présent

$$y_i = x_i b + u_i$$

On a vu que l'estimateur des mco exploitait largement l'hypothèse de non covariance entre les variables explicatives et le résidu :

$$E\left(x_i' u_i\right) = 0$$

Cette restriction se réécrit de façon analogue comme

$$E\left(x_i'\left(y_i - x_i b\right)\right) = 0$$

Elle est directement liée à l'expression de l'estimateur des mco. Celui-ci peut en effet être vu comme la valeur du paramètre qui annule la contrepartie empirique des conditions d'orthogonalité :

$$\overline{x_i'\left(y_i - x_i \widehat{b}_{mco}\right)} = 0$$

Il en va de même pour les variables instrumentales. La contrainte identifiante centrale prenait en effet la forme :

$$E\left(z_i^{VI\prime} u_i\right) = 0$$

et on a alors des relations du type

$$E\left(z_i^{VI\prime}\left(y_i - x_i b\right)\right) = 0$$

Les estimateurs de mci sont caractérisés par le fait qu'ils annulent une combinaison linéaire donnée de la contrepartie empirique des conditions d'orthogonalité :

$$A.\overline{z_i^{VI\prime}\left(y_i - x_i\widehat{b}_{mci}\left(A\right)\right)} = 0$$

Ces restrictions ont en commun le fait qu'un vecteur de fonctions des observations et des paramètres dont l'espérance est égale à zéro pour la vraie valeur du paramètre. Dans le premier cas il s'agit de $x_i^{'}\left(y_i - x_i b\right)$ et dans le second cas de $z_i^{'}\left(y_i - x_i b\right)$. La méthode des moments généralisée est la méthode adaptée pour estimer des modèles économétriques définis par l'existence de fonctions des observations et des paramètres d'espérance nulle. La méthode des moments généralisée va avoir pour nous plusieurs avantages :

- On va pouvoir étendre les procédure d'estimation et de test à des domaines plus généraux. Dans le cas des variables instrumentales par exemple, on va pouvoir définir des estimateurs optimaux ne reposant que sur les contraintes identifiantes $E\left(z_i^{VI\prime}\left(y_i - x_i b\right)\right) = 0$. En particulier, ils ne feront pas d'hypothèses de régularité sur la constance des moments d'ordres supérieurs. On va aussi pouvoir étendre les procédures de tests de suridentification et d'exogénéité au cas dans lequel les résidus sont hétéroscédastiques.
- On va aussi pouvoir aborder des situations plus générales que celle examinées jusqu'à présent en considérant pas exemple des systèmes d'équations à variables instrumentales. Ce type de généralisation est essentiel dans l'économétrie des données de panel. Là aussi on va pouvoir discuter les conditions d'identification des paramètres, définir des estimateurs optimaux, développer des procédure de tests de suridentification.
- La méthode des moments généralisée va aussi être l'occasion d'estimer et d'étudier des modèles se présentant sous des formes moins standards que celle d'une équation ou d'un système d'équation avec résidu. Dans certains cas, c'est spontanément sous la forme de fonctions des paramètres et des variables d'espérance nulle qu'un modèle émerge de la théorie. C'est le cas en particulier des équations d'Euler. Considérons par exemple le cas d'une entreprise décidant de son investissement. Notons $F\left(K_t, L_t, \theta\right)$ la fonction de production, et $M\left(K_t, I_t, \zeta\right)$ la fonction de coût d'ajustement. L'équation d'accumulation du capital s'écrit $K_t = \left(1-\delta\right)K_{t-1} + I_t$. La fonction de profit de l'entreprise s'écrit

$$E_t\left(\sum_{\tau=0}^{+\infty}\frac{1}{\left(1+r\right)^{\tau}}\left(p_\tau F\left(K_\tau, L_\tau, \theta\right) - w_\tau L_\tau - p_{I\tau} I_\tau - M\left(K_t, I_t, \zeta\right)\right)\right)$$

  L'entreprise cherche à maximiser ce profit sous contrainte d'accumulation. Le Lagrangien de l'objectif de l'entreprise s'écrit

$$E_t\left(\sum_{\tau=0}^{+\infty}\frac{1}{\left(1+r\right)^{\tau}}\left(p_\tau F\left(K_\tau, L_\tau, \theta\right) - w_\tau L_\tau - p_{I\tau} I_\tau - M\left(K_\tau, I_\tau, \zeta\right)\right) + \lambda_\tau\left(K_\tau - \left(1-\delta\right)K_{\tau-1} - I_t\right)\right)$$

On en déduit les conditions du premier ordre :

$$\begin{aligned} E_t\left(p_\tau \frac{\partial F\left(K_\tau, L_\tau, \theta\right)}{\partial K_\tau} - \frac{\partial M\left(K_\tau, I_\tau, \zeta\right)}{\partial K_\tau} + \lambda_\tau - \lambda_{\tau+1}\frac{1-\delta}{1+r}\right) &= 0 \\ E_t\left(p_{I\tau} + \frac{\partial M\left(K_\tau, I_t, \zeta\right)}{\partial I_\tau} + \lambda_\tau\right) &= 0 \\ E_t\left(p_\tau \frac{\partial F\left(K_\tau, L_\tau, \theta\right)}{\partial L_\tau} - w_\tau\right) &= 0 \end{aligned}$$

On en déduit en particulier pour la date $\tau = t$ la relation

$$\begin{aligned} 0 = E_t\Bigg[p_t \frac{\partial F\left(K_t, L_t, \theta\right)}{\partial K_t} - \frac{\partial M\left(K_t, I_t, \zeta\right)}{\partial K_t} + p_{It} + \frac{\partial M\left(K_t, I_t, \zeta\right)}{\partial I_t} - \\ \left(\left(\frac{1-\delta}{1+r}\right)\left(p_{It+1} + \frac{\partial M\left(K_{t+1}, I_{t+1}, \zeta\right)}{\partial I_{t+1}}\right)\right)\Bigg] \end{aligned}$$

Ce qui signifie que pour n'importe quelle variable z_t appartenant à l'ensemble d'information de la date t, on a

$$\begin{aligned} 0 = E\Bigg[\Bigg\{p_t \frac{\partial F\left(K_t, L_t, \theta\right)}{\partial K_t} - \frac{\partial M\left(K_t, I_t, \zeta\right)}{\partial K_t} + p_{It} + \frac{\partial M\left(K_t, I_t, \zeta\right)}{\partial I_t} - \\ \left(\left(\frac{1-\delta}{1+r}\right)\left(p_{It+1} + \frac{\partial M\left(K_{t+1}, I_{t+1}, \zeta\right)}{\partial I_{t+1}}\right)\right)\Bigg\} z_t\Bigg] \end{aligned}$$

On voit donc que dans ce cas le modèle conduit à un grand nombre (a priori infini) de relations entre les variables et les paramètres dont l'espérance est égale à zéro. L'un des intérêts de la méthode des moments généralisée est justement associé à cette particularité du modèle. Si le modèle est juste alors on doit avoir la propriété qu'il existe un paramètre de dimension finie annulant les conditions d'orthogonalité bien qu'elles soient en très grand nombre. Dans une certaine mesure peu importe la valeur du paramètre, ce qui compte vraiment est de savoir si l'ensemble des restrictions que la théorie économique impose aux données sont bien vérifiées empiriquement ; c'est à dire que l'on puisse trouver une valeur du paramètre telle que l'on accepte l'hypothèse de nullité de la contrepartie empirique des conditions d'orthogonalité lorsqu'elles sont évaluées en ce point.

11.2 Définir un modèle par le biais de conditions d'orthogonalit é

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 :

$$E\left(g\left(z_i,\theta\right)\right) = 0 \Leftrightarrow \theta = \theta_0$$

et pour θ_0 seulement. De telles relations portent le nom de conditions d'orthogonalité.

C'est un cadre très général englobant de nombreuses situations spécifiques :

11.2.1 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 vraisemblance est maximale pour $\theta = \theta_0$:

$$E\frac{\partial \log L\left(z_i,\theta\right)}{\partial\theta} = 0 \Leftrightarrow \theta = \theta_0$$

11.2.2 Modèle d'espérance conditionnelle, moindres carrés non linéaires

Il s'agit de la situation dans laquelle le modèle défini l'espérance d'une variable aléatoire y_i conditionnellement à des variables explicatives x_i :

$$E\left(y_i\left|x_i\right.\right) = f\left(x_i,\theta_0\right)$$

Les moindres carrés non linéaires définissent le paramètre comme celui minimisant la somme des carrés des résidus : $\left[\overline{\left(y_i - f\left(x_i,\theta\right)\right)^2}\right]$. On peut montrer que la vraie valeur du paramètre θ_0 minimise $E\left[\left(y_i - f\left(x_i,\theta\right)\right)^2\right]$ En effet, comme

$$\begin{aligned}
E\left[\left(y_i - f\left(x_i,\theta\right)\right)^2\right] &= E\left[y_i - f\left(x_i,\theta_0\right) + f\left(x_i,\theta_0\right) - f\left(x_i,\theta\right)\right]^2 \\
&= E\left[\left(y_i - f\left(x_i,\theta_0\right)\right)^2\right] \\
&\quad +2E\left[\left(y_i - f\left(x_i,\theta_0\right)\right)\left(f\left(x_i,\theta_0\right) - f\left(x_i,\theta\right)\right)\right] \\
&\quad +E\left[\left(f\left(x_i,\theta_0\right) - f\left(x_i,\theta\right)\right)^2\right] \\
&> E\left[\left(y_i - f\left(x_i,\theta_0\right)\right)^2\right]
\end{aligned}$$

on en déduit que $E\left[\left(y_i - f\left(x_i, \theta\right)\right)^2\right]$ est minimal pour $\theta = \theta_0$. On en déduit que la vraie valeur du paramètre et la vraie valeur seulement vérifie

$$E\left[\left(y_i - f\left(x_i, \theta\right)\right)\frac{\partial f\left(x_i, \theta\right)}{\partial \theta}\right] = 0 \Leftrightarrow \theta = \theta_0$$

11.2.3 Méthode à variables instrumentales pour une équation seule

Il s'agit de la généralisation du cas vu au chapitre précédent dans lequel on fait l'hypothèse qu'il existe un système de variables extérieures dites instrumentales, non corrélés avec les résidus :

$$E\left(z_i^{VI\prime}\left(y_i - x_i\theta_0\right)\right) = 0$$

où y_i est la variable dépendante, x_i le vecteur ligne des variables explicatives de dimension $1 \times \dim\left(\theta\right)$ et z_i le vecteur ligne des instruments de dimension $1 \times H$.

On a

$$E\left(z_i^{'}\left(y_i - x_i\theta\right)\right) = E\left(z_i^{'}x_i\right)\left(\theta_0 - \theta\right)$$

dès lors que $E\left(z_i^{'}x_i\right)$ est de rang $\dim\left(\theta\right)$

$$E\left(z_i^{'}\left(y_i - x_i\theta\right)\right) = 0 \Leftrightarrow \theta = \theta_0$$

Il s'agit d'une généralisation du cas du chapitre précédent dans la mesure où on ne fait plus que les hypothèses minimales : existence des conditions d'orthogonalité et condition de rang. En particulier on ne fait plus l'hypothèse d'homoscédasticité. De ce fait comme on va le voir l'estimateur optimal ne sera plus l'estimateur des doubles moindres carrés, le test de suridentification se généralise mais ne prend plus la même forme, le test d'exogénéité peut être mis en oeuvre mais fait partie d'une classe plus générale de tests de spécification. Le but principal de ce chapitre est tout en présentant les éléments généraux de la méthode des variables instrumentales de présenter l'extension des résultats précédents à cette situation plus générale.

11.2.4 Méthode à variables instrumentales pour un système d' équations.

La situation précédente peut être généralisée à un système d'équations. On considère ainsi le cas où les conditions d'orthogonalité sont données par :

$$E\left(\underline{Z}_i^{'}\left(\underline{y}_i - \underline{x}_i\theta_0\right)\right) = 0$$

où $\underline{y}_i$ est un vecteur de variables dépendantes de dimension $M \times 1$, $\underline{x}_i$ une matrice de variables explicatives de dimension $M \times \dim(\theta)$ et $\underline{Z}_i$ une matrice d'instruments de dimension $M \times H$ où la ligne m contient les instruments z_m de l'équation m : $\underline{Z}_i = diag(z_{mi})$ de telle sorte que

$$\underline{Z}_i'\underline{\varepsilon}_i = \begin{bmatrix} z_{1i}' & & \\ & \ddots & \\ & & z_{Mi}' \end{bmatrix} \begin{bmatrix} \varepsilon_{1i} \\ \vdots \\ \varepsilon_{Mi} \end{bmatrix} = \begin{bmatrix} z_{1i}'\varepsilon_{1i} \\ \vdots \\ z_{Mi}'\varepsilon_{Mi} \end{bmatrix}$$

On a

$$E\left(\underline{Z}_i'\left(\underline{y}_i - \underline{x}_i\theta\right)\right) = E\left(\underline{Z}_i'\underline{x}_i\right)(\theta_0 - \theta)$$

dès lors que $E\left(\underline{Z}_i'\underline{x}_i\right)$ est de rang $\dim(\theta)$

$$E\left(\underline{Z}_i'\left(\underline{y}_i - \underline{x}_i\theta\right)\right) = 0 \Leftrightarrow \theta = \theta_0$$

Ce cas simple, linéaire, englobe lui-même de très 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.

11.2.5 L'économétrie des données de panel

Le cadre précédent constitue un cadre général dans lequel il est possible de traiter l'économétrie des données de panel. Le modèle considéré est le suivant :

$$y_{it} = x_{it}b + \varepsilon_i + \omega_{it}$$

Les perturbations suivent donc le modèle à erreurs composées. On s'intéresse aux différentes possibilités de corrélation entre les variables explicatives et les perturbations, c'est à dire à la matrice

$$\Sigma = E\left(\underline{u}_i Vec\left(\underline{x}_i\right)'\right)$$

L'opérateur Vec est l'opérateur qui transforme une matrice en vecteur en empilant les colonnes de la matrice les unes en dessous des autres. D'une façon générale, les différentes possibilité de corrélation vont conduire à des paramétrages différents de la matrice Σ. On aura des matrices $\Sigma(\beta)$ différentes suivant la nature des corrélations entre les variables explicatives et les perturbations. L'ensemble de conditions d'orthogonalité que l'on considère est

$$E\left(\left(\underline{y}_i - \underline{x}_i b\right) Vec\left(\underline{x}_i\right)'\right) = \Sigma(\beta)$$

Tel quel cet ensemble est exprimé comme une matrice. On peut le transformer pour l'exprimer sous forme vectorielle en appliquant l'opérateur Vec. On voit que mis sous cette

forme, il y a toujours le même nombre de conditions d'orthogonalité : $\dim g = (K+1)T^2$ et un nombre de paramètre variant d'une spécification à l'autre. On voit bien que plus on va paramétrer la matrice de variance $\Sigma(\beta)$, moins on va conserver d'information pour estimer le paramètre d'intérêt b. Des situations dans lesquelles la matrice $\Sigma(\beta)$ est nulle par exemple vont exploiter toutes les covariances entre les perturbations et les variables explicatives à toutes les dates pour estimer le paramètre. Cette situation est très exigeante vis à vis des données. En revanche, dans la situation extrême dans laquelle la matrice $\Sigma(\beta)$ serait laissée totalement libre, on voit que le paramètre b ne serait plus identifié. En pratique les paramètres β sont des paramètres de nuisance et on n'a pas toujours envie de les estimer car ils peuvent être nombreux et leur examen serait long fastidieux et pas nécessairement très utile. Dans les cas considérés il est en général possible d'éliminer tout ou partie de ces paramètres de nuisance en appliquant des transformations aux données. On a ainsi en général des relations prenant la forme

$$E\left(H\left(\underline{y}_i - \underline{x}_i b\right) Vec\left(\underline{x}_i\right)'\right) = 0$$

On voit que formellement, la situation considérée est analogue à celle d'un système d'équations avec variables instrumentales. Les instruments ici sont dits internes dans la mesure où ce sont les valeurs passées présentes ou futures des variables explicatives qui sont utilisées comme instruments. On voit aussi que ce cadre est très général, et qu'il est susceptible de délivrer des estimateurs des paramètres dans des situations nouvelles pour lesquelles il n'était pas possible de le faire auparavant, dans le cadre standard. On détaille maintenant les différentes situations possibles et on donne l'ensemble de conditions d'orthogonalité correspondant.

Exogénéité forte

La première situation que l'on considère est celle dite d'exogénéité forte et correspond au cas dans lequel il n'y a pas de corrélations entre les perturbations et les variables explicatives passées présentes et futures, soit

$$\begin{aligned} E\left(\varepsilon_i x_{ks,i}\right) &= 0 \quad \forall s = 1, \ldots, T \\ E\left(\omega_{t,i} x_{ks,i}\right) &= 0 \quad \forall s, t = 1, \ldots, T \end{aligned}$$

Ces hypothèses sont autant de restrictions intervenant explicitement dans les estimations. Sous ces hypothèses ne nombreux estimateurs standards : MCO, Between, Within, MCQG, sont tous convergents. On va voir comment ces estimateurs s'interprètent maintenant dans le cadre plus général considéré ici. On peut remarquer qu'il y a ici $(K+1)T^2$ conditions d'orthogonalité :

$$E\left(u_{t,i} x_{ks,i}\right) = 0, \forall t, \forall s$$

et que ces conditions d'orthogonalité peuvent de réécrire de façon équivalente comme

$$\begin{aligned} E\left(u_{1i} x_{ks,i}\right) &= 0, \forall s \\ E\left(\Delta u_{ti} x_{ks,i}\right) &= 0 \forall t > 1, \forall s \end{aligned}$$

Effets corrélés

Une deuxième situation correspond au cas où l'on ne souhaite pas faire reposer les estimations sur l'hypothèse $E(\varepsilon_i x_{ks,i}) = 0$. On introduit donc des paramètres de nuisance $\delta_{ks} = E(\varepsilon_i x_{ks,i})$. On autorise donc le fait que les éléments d'hétérogénéité individuelles soient corrélés avec les variables explicatives, d'où le nom d'effets corrélés. Il y a donc $(K+1)T$ paramètres de nuisance. On maintient par contre l'hypothèse $E(\omega_{t,i} x_{ks,i}) = 0$. On a donc comme condition d'orthogonalité :

$$E(u_{t,i} x_{ks,i}) = \delta_{ks}, \forall t, s$$

De façon équivalente, on peut éliminer les paramètres de nuisance, éliminant au passage certaines conditions d'orthogonalité. Les $(K+1)T^2$ conditions d'orthogonalité peuvent ainsi être réécrites après élimination des $(K+1)T$ paramètres de nuisance comme

$$E(\Delta u_{t,i} x_{ks,i}) = 0, \forall t > 1, \forall s$$

Il y a alors $(K+1)T(T-1)$ conditions d'orthogonalité. On remarque en outre qu'il s'agit aussi du deuxième ensemble de conditions d'orthogonalité identifié dans le cas de l'exogénéité forte.

Exogénéité faible

L'hypothèse $E(\omega_{t,i} x_{ks,i}) = 0 \quad \forall s, t = 1, \ldots, T$ peut paraître excessive elle aussi. Ainsi dans le cas des conditions d'Euler on est plutôt amené à utiliser comme variables instrumentales des variables passées. On peut ainsi préférer ne retenir comme restriction identifiante que $E(\omega_{t,i} x_{ks,i}) = 0 \quad \forall t = 1, \ldots, T$ et $s < t$. On autorise ainsi que les chocs passés affectent les décisions concernant le niveau de la variable $x_{ks,i}$. C'est cette spécification qui porte le nom d'exogénéité faible. Elle consiste donc à introduire $(K+1)T + (K+1)T(T+1)/2$ paramètres de nuisance :

$$\begin{aligned} E(\varepsilon_i x_{ks,i}) &= \delta_{ks} \\ E(\omega_{t,i} x_{ks,i}) &= \gamma_{t,ks} \text{ pour } s \geq t \end{aligned}$$

On maintient en revanche

$$E(\omega_{t,i} x_{ks,i}) = 0 \quad \forall t = 1, \ldots, T \quad \text{et} \quad s < t$$

Finalement les conditions d'orthogonalité s'écrivent dans ce cas sous la forme

$$E(u_{t,i} x_{ks,i}) = \delta_{ks} + \gamma_{t,ks} 1(t \geq s), \quad \forall t, s$$

Là aussi on peut de façon équivalente réécrire ces conditions d'orthogonalité pour éliminer les paramètres de nuisance. Les $(K+1)T^2$ conditions d'orthogonalité peuvent ainsi être réécrites après élimination des paramètres de nuisance comme

$$E(\Delta u_{t,i} x_{ks,i}) = 0, \forall t > s+1, \forall s$$

	Exogénéité forte	Effets Corrélés	Exogénéité faible
Restrictions relâchées	-	$E(\varepsilon_i x_{ks,i}) = 0$	E($\varepsilon_i x_{ks,i}$) =0, E($\omega_{t,i} x_{ks,i}$) = 0 $\forall s \geq t$
Restrictions maintenues	E($\varepsilon_i x_{ks,i}$) =0, $E(\omega_{t,i} x_{ks,i}) = 0 \quad \forall s,t$	$E(\omega_{t,i} x_{ks,i}) = 0$ $\forall s < t$	$E(\omega_{t,i} x_{ks,i}) = 0$ $\forall s,t$
Conditions d'orthogonalité	$g_F = \{g_{F/C}, g_{C/f}, g_f\}$	$g_C = \{g_{C/f}, g_f\}$	g_f

TAB. 11.1 – Conditions d'orthogonalité et choix d'une spécification

Il y a alors $(K+1)T(T-1)/2$ conditions d'orthogonalité. On remarque en outre qu'il s'agit aussi d'une sous partie de l'ensemble de conditions d'orthogonalité de celui obtenu dans le cas des effets corrélés.

Synthèse

On voit que l'on peut synthétiser les résultats précédents en introduisant trois ensembles de conditions d'orthogonalité :

$$\begin{aligned} g_f &= (\Delta u_{t,i} x_{ks,i})_{t>s+1} \\ g_{C/f} &= (\Delta u_{t,i} x_{ks,i})_{t\leq s+1} \\ g_{F/C} &= (u_{1i} x_{ks,i}) \end{aligned}$$

Le tableau 11.1 récapitule les trois situations examinées. Les différentes spécifications sont emboîtées les unes dans les autres. La plus générale est la spécification exogénéité faible. Dans ce cas les estimations ne reposent que sur un ensemble minimal d'information. La spécification effets corrélés introduit plus d'information. L'ensemble des conditions d'orthogonalité inclus outre celles déjà présentes dans la spécification exogénéité faible certaines conditions supplémentaires spécifiques aux effets corrélés. Enfin dans le cas de l'exogénéité forte, on adjoint à l'ensemble de conditions d'orthogonalité précédent des conditions additionnelles, spécifiques à l'exogénéité forte. On va pouvoir définir des estimateurs ne reposant que sur ces différents sous-ensembles de conditions d'orthogonalité. On va aussi pouvoir, comme dans le cas des variables instrumentales, tester la cohérence de chacun de ces sous-ensembles de conditions d'orthogonalité. Le test effectué sera analogue au test de Sargan. Enfin, on va pouvoir tester la compatibilité des différents sous-ensembles d'information entre eux. Ainsi on va pourvoir tester si par exemple lorsque l'on a estimé le modèle sous l'hypothèse effets corrélés, les conditions d'orthogonalité additionnelles spécifiques à l'exogénéité fortes sont compatibles avec les conditions déjà mobilisées. Le test s'apparente alors au test d'exogénéité examiné dans le cas homoscédastique univarié.

11.3 Principe de la méthode :

Le principe des 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, de la méthode des variables instrumentales lorsqu'il y a autant d'instruments que de variables endogènes)

- 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* $E\left(g\left(z_i,\theta\right)\right)$ *même dans le cas de la suridentification, alors que c'est en général impossible à distance finie pour la contrepartie empirique des conditions d'orthogonalité.*

Dans le cas de la 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 :

$$\widehat{\theta}_{S_N} = Arg\min_{\theta} \overline{g\left(z_i,\theta\right)}' S_N \overline{g\left(z_i,\theta\right)}$$

Remarque *Dans le cas des variables instrumentales, on réglait le problème de la suridentification en considérant des combinaisons linéaires des conditions d'orthogonalité. Ceci conduisait aux estimateurs des moindres carrés indirects* $\widehat{b}_{mci}\left(A\right)$ *, définis par*

$$\overline{Az_i^{VI\prime}\left(y_i - x_i\widehat{b}_{mci}\left(A\right)\right)} = 0$$

Ici on aurait pu procéder de même et définir des estimateurs basés sur une combinaison linéaire des conditions d'orthogonalité. On aurait alors défini des estimateurs de la forme

$$\overline{Ag\left(z_i,\widehat{\theta}_{A_N}\right)} = 0$$

Les deux approches sont en fait analogues.

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*

$$g(z_i,\theta) = \underline{Z}_i^{'}\left(\underline{y}_i - \underline{x}_i\theta\right) = \underline{Z}_i^{'}\underline{y}_i - \underline{Z}_i^{'}\underline{x}_i\theta = g_1(z_i) - g_2(z_i)\,\theta$$

On note $\overline{g_1} = \overline{g_1(z_i)}$ *et* $\overline{g_2} = \overline{g_2(z_i)}$. *L'estimateur est alors défini par :*

$$\widehat{\theta}_S = Arg\min_{\theta}\left(\overline{g_1} - \overline{g_2}\theta\right)^{'} S_N\left(\overline{g_1} - \overline{g_2}\theta\right)$$

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

$$\widehat{\theta}_S = \left(\overline{\underline{x}_i^{'}\underline{Z}_i} S_N \overline{\underline{Z}_i^{'}\underline{x}_i}\right)^{-1} \overline{\underline{Z}_i^{'}\underline{x}_i} S_N \overline{\underline{Z}_i^{'}\underline{y}_i}$$

Dans le cas d'une seule équation, les estimateurs obtenus par la méthode des moments généralisée sont ainsi :

$$\widehat{\theta}_S = \left(\overline{x_i^{'}z_i} S_N \overline{z_i^{'}x_i}\right)^{-1} \overline{x_i^{'}z_i} S_N \overline{z_i^{'}y_i}$$

Si on prend par exemple pour métrique $S_N = \overline{z_i^{'}z_i}^{-1}$ *On obtient l'estimateur des doubles moindres carrés. On en conclut que dans le cas où les conditions d'orthogonalité sont* $E\left(z_i^{'}\left(y_i - x_i\theta_0\right)\right) = 0$, *c'est à dire celles vues dans le chapitre précédent sur les variables instrumentales, on retrouve comme estimateur GMM particulier l'estimateur des doubles moindres carrés. Néanmoins le cadre dans lequel on se situe est plus général puisqu'on ne fait plus l'hypothèse d'homoscédasticité. On va voir que pour cette raison, l'estimateur des doubles moindres carrés n'est plus l'estimateur de variance minimal.*

11.4 Convergence et propriétés asymptotiques

Comme dans les cas examinés précédemment on va voir que les estimateurs GMM présentés sont convergents et asymptotiquement normaux. Comme précédemment l'obtention de ces résultats nécessite des hypothèses. Elles vont porter ici sur les moments des variables z_i mis aussi sur la régularité de la fonction $g(z_i,\theta)$.

Proposition *Sous les hypothèses*

1. *H*1 *L'espace des paramètres* Θ *est compact. La vraie valeur est* θ_0 *intérieure à* Θ,
2. *H*2 $E\left(g(z_i,\theta)\right) = 0 \Leftrightarrow \theta = \theta_0$,

3. *H3* $g(z_i,\theta)$ *est continûment dérivable en* θ,
4. *H4* $E\left[\sup_\theta |g(z_i,\theta)| + \sup_\theta |g(z_i,\theta)|^2 + \sup_\theta |\nabla_\theta g(z_i,\theta)|\right] < \infty$,
5. *H5* $g_k(z_i,\theta_0)$ *a des moments finis d'ordre* 1 *et* 2,
6. *H6 Le Jacobien* $G = E(\nabla_\theta g(z_i,\theta_0))$ *de dimension* $\dim g \times \dim\theta$ *est de* $\operatorname{rang}\dim\theta$,
7. *H7* $S_N \xrightarrow{P} S_0$ *définie positive.*

L'estimateur GMM $\widehat{\theta}_{SN}$ *minimisant* $Q_N(\theta)$ *défini par* $Q_N(\theta) = \overline{g(z_i,\theta)}' S_N \overline{g(z_i,\theta)}$, *est convergent et asymptotiquement normal. Sa matrice de variance asymptotique est fonction de* S_0 *et de la matrice de variance des conditions d'orthogonalité. Elle peut être estimée de façon convergente.*

1. $\widehat{\theta}_S \xrightarrow{P} \theta_0$ *convergence*
2. $\sqrt{N}\left(\widehat{\theta}_S - \theta_0\right) \xrightarrow{L} N\left(0, V_{as}\left(\widehat{\theta}(S)\right)\right)$ *normalité asymptotique*
3. $V_{as}\left(\widehat{\theta}_S\right) = [G'S_0G]^{-1} G'S_0V(g(z_i,\theta_0)) S_0G [G'S_0G]^{-1}$ *où* $S_0 = p\lim S_N$ *et* $V(g(z_i,\theta_0)) = E\left[g(z_i,\theta_0)\, g(z_i,\theta_0)'\right]$
4. $\widehat{V}(g(z_i,\theta_0)) = \overline{g\left(z_i,\widehat{\theta}_S\right) g\left(z_i,\widehat{\theta}_S\right)'} \to V(g(z_i,\theta_0))$ *et* $\widehat{G} = \overline{\frac{\partial g}{\partial\theta}\left(z_i,\widehat{\theta}_S\right)} \to G$
5. $\widehat{V}_{as}\left(\widehat{\theta}_S\right) = \left[\widehat{G}'S_0\widehat{G}\right]^{-1} \widehat{G}'S_N\widehat{V}(g(z_i,\theta_0)) S_N\widehat{G}\left[\widehat{G}'S_0\widehat{G}\right]^{-1} \to V_{as}\left(\widehat{\theta}(S)\right)$

Parmi ces conditions la deuxième est de loin la plus importante puisque c'est elle qui définit l'identification du paramètre. C'est sur le choix des fonctions $g(z_i,\theta)$ que porte le travail du modélisateur. La condition 3 est essentielle pour obtenir la loi asymptotique des paramètres. En effet il est central de pouvoir linéariser autour de la vraie valeur du paramètre. La condition 4 est technique. Elle garantit qu'il y a convergence uniforme en probabilité de $\overline{g(z_i,\theta)}$ vers $E(g(z_i,\theta))$ (et pareil pour les autres fonctions concernées $\nabla_\theta g(z_i,\theta)$ et $g(z_i,\theta)\, g(z_i,\theta)'$). La condition 5 est l'analogue de la condition z_iu_i a des moments d'ordre 1 et 2, dans le cas des variables instrumentales. Elle est essentielle dans l'application du théorème central limite dans la dérivation de l'expression de la matrice de variance. La condition 6 sert aussi pour dériver l'expression de la matrice de variance. Dans le cas linéaire, elle est analogue à la condition d'identification 2.

Démonstration *Convergence : Soit* $Q_N(\theta) = \overline{g(z_i,\theta)}' S_N \overline{g(z_i,\theta)}$ *et* $Q(\theta) = E(g(z_i,\theta))' S_0\, E(g(z_i,\theta))$. *On peut écrire*

$$\begin{aligned} Q\left(\widehat{\theta}_S\right) - Q(\theta_0) &= \left[Q_N\left(\widehat{\theta}_S\right) + \left(Q\left(\widehat{\theta}_S\right) - Q_N\left(\widehat{\theta}_S\right)\right)\right] - \\ &\quad [Q_N(\theta_0) + (Q(\theta_0) - Q_N(\theta_0))] \end{aligned}$$

comme $Q_N\left(\widehat{\theta}_S\right) \leq Q_N(\theta_0)$ *et* $Q(\theta_0) \leq Q\left(\widehat{\theta}_S\right)$, *on a*

$$\begin{aligned} 0 &\leq Q\left(\widehat{\theta}_S\right) - Q(\theta_0) \leq \left(Q\left(\widehat{\theta}_S\right) - Q_N\left(\widehat{\theta}_S\right)\right) - (Q(\theta_0) - Q_N(\theta_0)) \\ &\leq 2\sup_\theta |Q(\theta) - Q_N(\theta)| \end{aligned}$$

La condition $E\left[\sup_\theta |g(z_i,\theta)|\right] < +\infty$ *permet de montrer qu'il y a convergence uniforme de* $\overline{g(z_i,\theta)}$ *vers* $E(g(z_i,\theta))$, *et donc de* $Q_N(\theta)$ *vers* $Q(\theta) = E(g(z_i,\theta))' S_0 E(g(z_i,\theta))$. *On en déduit donc que* $Q\left(\widehat{\theta}_S\right) \xrightarrow{P} Q(\theta_0)$. *Comme la fonction* Q *est continue, que* Θ *est compact, que* $Q(\theta_0) = 0$ *et* $Q(\theta) = 0 \Leftrightarrow E(g(z_i,\theta)) = 0 \Leftrightarrow \theta = \theta_0$ *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* $0 = \sqrt{N}\overline{g\left(z_i,\widehat{\theta}_S\right)}\sqrt{N\,g(z_i,\theta_0)} + \overline{\nabla_\theta g\left(z_i,\widetilde{\theta}_S\right)}\sqrt{N}\left(\widehat{\theta}_S - \theta_0\right)$, *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* $\overline{\nabla_\theta g\left(z_i,\widehat{\theta}_S\right)}' S_N \overline{\nabla_\theta g\left(z_i,\widetilde{\theta}_S\right)}\sqrt{N}\left(\widehat{\theta}_S - \theta_0\right) = -\overline{\nabla_\theta g\left(z_i,\widehat{\theta}_S\right)}' S_N \sqrt{N\,g(z_i,\theta_0)}$

La condition $E\left[\sup_\theta |\nabla_\theta g(z_i,\theta)|\right] < +\infty$ *garantit la convergence uniforme en probabilité de* $\overline{\nabla_\theta g(z_i,\theta)}$ *vers* $E(\nabla_\theta g(z_i,\theta))$. *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* $\operatorname{rang} G = \dim\theta$. *La condition que* $g_k(z_i,\theta_0)$ *a des moments d'ordre 1 et 2 permet d'appliquer le théorème central limite à* $\sqrt{N\,g(z_i,\theta_0)}$: $\sqrt{N\,g(z_i,\theta_0)} \xrightarrow{Loi} N(0, V(g(z_i,\theta_0)))$. *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 :*

$$\sqrt{N}\left(\widehat{\theta}_S - \theta_0\right) = -\left(G' S_N G\right)^{-1} G' S_N \sqrt{N\,g(z_i,\theta_0)} + o(1)$$

Estimation de la matrice de variance asymptotique

Le seul point à montrer est que $\overline{g\left(z_i,\widehat{\theta}_S\right) g\left(z_i,\widehat{\theta}_S\right)'} \to V(g(z_i,\theta_0))$. *La condition* $E\left[\sup_\theta |g(z_i,\theta)|^2\right] < \infty$, *permet de montrer qu'il y a convergence uniforme de* $\overline{g(z_i,\theta)\,g(z_i,\theta)'}$

vers $E\left(g\left(z_i,\theta\right)g\left(z_i,\theta\right)'\right)$

11.5 Estimateur optimal

Comme dans les cas précédemment abordés, on montre qu'il existe un estimateur GMM optimal.

11.5.1 Existence d'un estimateur optimal

Proposition *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 dans la classe des estimateurs GMM, ils 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.

11.5.2 Mise en oeuvre de l'estimateur optimal : 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 ne faisant pas intervenir le paramètre. En fait on a intérêt à réfléchir et à chercher une matrice qui ne soit pas trop loin de la matrice optimale. $S_N = I_{\dim g}$ est un choix possible mais certainement pas le meilleur. La mise en oeuvre des GMM avec cette métrique 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 conditions d'orthogonalité :

$$\widehat{V}\left(g\right)_N = \overline{g\left(z_i, \widehat{\theta}_1\right) g\left(z_i, \widehat{\theta}_1\right)'} \xrightarrow{P} V\left(g\left(z_i, \theta_0\right)\right)$$

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 dès lors déterminer un estimateur de la matrice de variance asymptotique de ce premier estimateur

$$\widehat{V}_{as}\left(\widehat{\theta}_1\right)_N = \left(\widehat{G}' S_N \widehat{G}\right)^{-1} \widehat{G}' S_N \widehat{V}\left(g\right)_N S_N \widehat{G} \left(\widehat{G}' S_N \widehat{G}\right)^{-1}$$

Deuxième étape : On met à nouveau en oeuvre l'estimateur des GMM avec la métrique $S_N^* = \widehat{V}\left(g\right)_N^{-1}$. On obtient ainsi un estimateur convergent et asymptotiquement efficace dont on peut estimer la matrice de variance asymptotique

$$\widehat{V}_{as}\left(\widehat{\theta}^*\right)_N = \left(\widehat{G}' S_N^* \widehat{G}\right)^{-1}$$

11.6 Application aux Variables Instrumentales

11.6.1 Variables instrumentales dans un système d'équations - cas général

On considère le cas d'un système d'équations avec variables instrumentales

$$g\left(z_i, \theta\right) = \underline{Z}_i' \left(\underline{y}_i - \underline{x}_i \theta\right) = \underline{Z}_i' \underline{y}_i - \underline{Z}_i' \underline{x}_i \theta$$

Vérification des hypothèses de convergence des estimateurs GMM

$H2$ $E\left(\underline{Z}_i'\underline{y}_i\right) - E\left(\underline{Z}_i'\underline{x}_i\right)\theta = 0$ admet une unique solution si $\operatorname{rang} E\left(\underline{Z}_i'\underline{x}_i\right) = \dim\theta$. Il s'agit là d'une simple généralisation de la condition déjà vue dans le cadre univarié.

$H3$ est satisfaite du fait de la linéarité.

$H4$ et $H5$ sont satisfaites si $E\left[\left(\sup\left|\underline{Z}_i'\underline{y}_i\right| + \sup\left|\underline{Z}_i'\underline{x}_i\right|\right)^2\right] < +\infty$, c'est à dire si les moments de $\underline{Z}_i$, $\underline{x}_i$ et $\underline{y}_i$ existent jusqu'à un ordfre suffisant.

$H6$ $\nabla_\theta g(z_i,\theta_0) = -\underline{Z}_i'\underline{x}_i$. Si $E\left(\underline{Z}_i'\underline{x}_i\right)$ est de rang $\dim\theta$ $G = E\left(\nabla_\theta g(z_i,\theta_0)\right) = -E\left(\underline{Z}_i'\underline{x}_i\right)$ est de rang $\dim\theta$

Expression de la matrice de variance des conditions d'orthogonalité :

La variance des conditions d'orthogonalité s'écrit :

$$\begin{aligned} V\left(g(z_i,\theta_0)\right) &= E\left(\underline{Z}_i'\left(\underline{y}_i - \underline{x}_i\theta_0\right)\left(\underline{y}_i - \underline{x}_i\theta_0\right)'\underline{Z}_i\right) \\ &= E\left(\underline{Z}_i'\underline{u}_i\underline{u}_i'\underline{Z}_i\right) \end{aligned}$$

Cette expression est très proche de celle vue dans le cadre des variables instrumentales. Néanmoins, comme on le voit elle fait en général intervenir le paramètre θ. Il est donc souvent nécessaire de mettre en oeuvre une méthode en deux étapes.

Mise en oeuvre de l'estimation

Première étape : Il faut choisir une métrique pour l'estimateur de première étape. La métrique optimale est l'inverse de la matrice de variance des conditions d'orthogonalité. Elle a l'expression donnée précédemment. On a intérêt à choisir pour métrique de première étape une métrique qui soit proche de la métrique optimale. Pour cela on peut choisir pour métrique ce qu'aurait été la métrique optimale en présence d'hypothèses de régularité supplémentaires. Une hypothèse de régularité importante pourrait être l'homoscédasticité

$$E\left(\underline{u}_i\underline{u}_i' \,|\underline{Z}_i\right) = E\left(\underline{u}_i\underline{u}_i'\right)$$

Qui pourra être utilisée si

$$E\left(\underline{u}_i\underline{u}_i'\right) = \sigma^2 D$$

où D est une matrice donnée. Par exemple $D = I_M$, ce qui correspondrait à l'hypothèse que les résidus des équations sont indépendants et équidistribués. On utiliserait alors pour métrique de première étape

$$S_N = \overline{\underline{Z}_i' D \underline{Z}_i}$$

On peut se trouver dans des situations ou spontanément la matrice de variance des résidus aurait une allure différente. C'est en particulier le cas parfois dans le cas de l'économétrie des données de panel. Quel que soit le choix effectué, l'estimateur de première étape a pour expression :

$$\widehat{\theta}_S = \left(\overline{\underline{x}_i'\underline{Z}_i} S_N \overline{\underline{Z}_i'\underline{x}_i}\right)^{-1} \overline{\underline{x}_i'\underline{Z}_i} S_N \overline{\underline{Z}_i'\underline{y}_i}$$

La matrice de variance des conditions d'orthogonalité peut être alors être estimée par

$$\widehat{V}(g) = \overline{\underline{Z}_i' \left(\underline{y}_i - \underline{x}_i\widehat{\theta}_S\right) \left(\underline{y}_i - \underline{x}_i\widehat{\theta}_S\right)' \underline{Z}_i} = \overline{\underline{Z}_i'\widehat{\underline{u}}_i\widehat{\underline{u}}_i'\underline{Z}_i}$$

A partir de cette estimation, on peut aussi estimer la variance de l'estimateur de première étape :

$$\widehat{V}\left(\widehat{\theta}(S)\right) = \left(\overline{\underline{x}_i'\underline{Z}_i} S_N \overline{\underline{Z}_i'\underline{x}_i}\right)^{-1} \overline{\underline{Z}_i'\underline{x}_i} S_N \widehat{V}(g) S_N \overline{\underline{x}_i'\underline{Z}_i} \left(\overline{\underline{Z}_i'\underline{x}_i} S_N \overline{\underline{Z}_i'\underline{x}_i}\right)^{-1}$$

ainsi que l'estimateur optimal :

$$\widehat{\theta}_S^* = \left(\overline{\underline{x}_i'\underline{Z}_i}\widehat{V}(g)^{-1}\overline{\underline{Z}_i'\underline{x}_i}\right)^{-1} \overline{\underline{x}_i'\underline{Z}_i}\widehat{V}(g)^{-1}\overline{\underline{Z}_i'\underline{y}_i}$$

et sa variance asymptotique :

$$\widehat{V}_{as}\left(\widehat{\theta}_S^*\right) = \left(\overline{\underline{x}_i'\underline{Z}_i}\widehat{V}(g)^{-1}\overline{\underline{Z}_i'\underline{x}_i}\right)^{-1}$$

11.6.2 Régressions à variables instrumentales dans un système homoscédastique

On a vu que dans le cas de M régressions empilées homoscédastiques, lorsque les régresseurs étaient les mêmes et qu'il n'existait pas de restrictions sur les paramètres, la mise en oeuvre de la méthode des MCQG conduit aux mêmes estimateurs que ceux obtenus par les moindres carrés ordinaires équation par équation. On peut voir que ce résultat se généralise au cas des variables instrumentales dans un système d'équation.

Si les régresseurs sont les mêmes, si il n'existe pas de contraintes entre les paramètres des équations $(\underline{x}_i = I_M \otimes x_i)$, et si les instruments sont les mêmes d'une équation à l'autre $(\underline{Z}_i = I_M \otimes z_i)$, dans le cas d'homoscédasticité des perturbations : $E(\underline{u}_i\underline{u}_i' | \underline{Z}_i) = \Sigma$, l'estimateur GMM optimal est identique à l'estimateur à variables instrumentales équation par équation. Sous l'hypothèse d'homoscédasticité, la matrice de variance des conditions d'orthogonalité a pour expression $E\left(\underline{Z}_i'\Sigma\underline{Z}_i\right) = \Sigma \otimes E\left(z_i'z_i\right)$. (Rappel : pour des matrices aux tailles appropriées $(A \otimes B)(C \otimes D) = AC \otimes BD$). On a donc $\Sigma\underline{Z}_i = (\Sigma \otimes 1)(I_M \otimes z_i) = \Sigma \otimes z_i$. D'où $\underline{Z}_i'\Sigma\underline{Z}_i = \left(I_M \otimes z_i'\right)(\Sigma \otimes z_i) = \Sigma \otimes z_i'z_i$. On a

donc

$$\begin{aligned}\overline{\underline{x}_i'\underline{Z}_i}S^*\overline{\underline{Z}_i'\underline{x}_i} &= \left(I_M\otimes\overline{x_i'z_i}\right)\left(\Sigma\otimes E\left(z_i'z_i\right)\right)^{-1}\left(I_M\otimes\overline{z_i'x_i}\right)\\ &= \Sigma^{-1}\otimes\left(\overline{x_i'z_i}E\left(z_iz_i'\right)^{-1}\overline{z_i'x_i}\right)\end{aligned}$$

et

$$\begin{aligned}\overline{\underline{x}_i'\underline{Z}_i}S^*\overline{\underline{Z}_i'\underline{y}_i} &= \left(I_M\otimes\overline{x_i'z_i}\right)\left(\Sigma\otimes E\left(z_i'z_i\right)\right)^{-1}\overline{\left(I_M\otimes z_i'\right)\underline{y}_i}\\ &= \left[\Sigma^{-1}\otimes\left(\overline{x_i'z_i}E\left(z_iz_i'\right)^{-1}\right)\right]\begin{bmatrix}\overline{z_i'y_{1i}}\\ \vdots\\ \overline{z_i'y_{Mi}}\end{bmatrix}\end{aligned}$$

puisque $\left(I_M\otimes z_i'\right)\underline{y}_i=\begin{bmatrix}z_i'y_{1i}\\ \vdots\\ z_i'y_{Mi}\end{bmatrix}$. L'estimateur optimal a donc pour expression

$$\begin{aligned}\widehat{\theta}_S^* &= \Sigma\otimes\left(\overline{x_i'z_i}E\left(z_iz_i'\right)^{-1}\overline{z_i'x_i}\right)^{-1}\times\Sigma^{-1}\otimes\left(\overline{x_i'z_i}E\left(z_iz_i'\right)^{-1}\right)\begin{bmatrix}\overline{z_i'y_{1i}}\\ \vdots\\ \overline{z_i'y_{Mi}}\end{bmatrix}\\ &= I_M\otimes\overline{x_i'z_i}\left(\Sigma\otimes E\left(z_iz_i'\right)\right)^{-1}\begin{bmatrix}\overline{z_i'y_{1i}}\\ \vdots\\ \overline{z_i'y_{Mi}}\end{bmatrix}=\begin{bmatrix}\widehat{b}_{2mc1}\\ \vdots\\ \widehat{b}_{2mcM}\end{bmatrix}\end{aligned}$$

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

$$V\left(\widehat{\theta}^*\right)=\Sigma\otimes\left(E\left(x_i'z_i\right)E\left(z_iz_i'\right)^{-1}E\left(z_i'x_i\right)\right)^{-1}$$

on voit donc que les estimateurs ne sont pas indépendants les uns des autres dès que la matrice de variance Σ n'est pas diagonale.

11.6.3 Application aux données de panel

Le cas des variables instrumentales dans un système d'équation correspond aussi données de panel. On a vu dans la première section Les différents types de spécification que l'on pouvait retenir. On a examiné le cas de l'exogénéité forte, des effets corrélés et de

l'exogénéité faible. Dans ce dernier cas, on a vu que le modèle était mis en différence première et que l'on utilisait les variables explicatives retardées à partir de l'ordre 2 comme instrument. On a ainsi la spécification matricielle suivante :

$$\underline{Z}_i'\Delta\underline{u}_i = \begin{pmatrix} x_{1i} & 0 & 0 & \\ 0 & x_{1i} & 0 & \\ & x_{2i} & x_{1i} & \\ & 0 & x_{2i} & \\ & & x_{3i} & \\ \vdots & & 0 & \\ & \vdots & & x_{1i} \\ & & & \vdots \\ 0 & 0 & 0 & x_{T-2i} \end{pmatrix} \begin{pmatrix} \Delta u_{3i} \\ \Delta u_{4i} \\ \Delta u_{5i} \\ \vdots \\ \Delta u_{Ti} \end{pmatrix}$$

De même pour les effets corrélés, on a

$$\underline{Z}_i'\Delta\underline{u}_i = \begin{pmatrix} \underline{x}_i & 0 & 0 & \\ 0 & \underline{x}_i & 0 & \\ & 0 & \underline{x}_i & \\ & & 0 & \\ & & & \\ \vdots & & & \\ & \vdots & & 0 \\ & & & \underline{x}_i \\ 0 & 0 & 0 & \end{pmatrix} \begin{pmatrix} \Delta u_{2i} \\ \Delta u_{4i} \\ \Delta u_{5i} \\ \vdots \\ \Delta u_{Ti} \end{pmatrix}$$

et enfin pour l'exogénéité forte on a

$$\left(\underline{Z}_i' \begin{pmatrix} u_{1i} \\ \Delta\underline{u}_i \end{pmatrix}\right) = \begin{pmatrix} \underline{x}_i & 0 & 0 & \\ 0 & \underline{x}_i & 0 & \\ & 0 & \underline{x}_i & \\ & & 0 & \\ & & & \\ \vdots & & & \\ & \vdots & & 0 \\ & & & \underline{x}_i \\ 0 & 0 & 0 & \end{pmatrix} \begin{pmatrix} u_{1i} \\ \Delta u_{2i} \\ \Delta u_{4i} \\ \Delta u_{5i} \\ \vdots \\ \Delta u_{Ti} \end{pmatrix}$$

Pour mettre en oeuvre l'estimateur optimal on applique la méthode exposée précédemment. On peut remarquer que dans le cas de l'exogénéité faible et des effets corrélés, la structure des conditions d'orthogonalité est telle qu'elle ne fait intervenir que la différence première des résidus. Ceci est à l'origine d'une possibilité d'un choix judicieux de la

matrice de variance de première étape. En effet, sous l'hypothèse d'homoscédasticité des résidus On aurait

$$E\left(\underline{Z}_i'\Delta\underline{u}_i\Delta\underline{u}_i'\underline{Z}_i\right) = E\left(\underline{Z}_i'E\left(\Delta\underline{u}_i\Delta\underline{u}_i'\right)\underline{Z}_i\right)$$

Or $E\left(\Delta\underline{u}_i\Delta\underline{u}_i'\right) = \sigma_\omega^2 D$, où

$$D = \begin{pmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix}$$

ne dépend pas des paramètres. On va que dans ce cas on peut choisir comme matrice de première étape une matrice approximant à l'hypothèse d'homoscédasticité prés la matrice de variance des conditions d'orthogonalité. La matrice S_1 a ainsi pour expression

$$S_1 = \overline{\underline{Z}_i' D \underline{Z}_i}$$

11.6.4 Estimateur VI 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. Les conditions d'orthogonalité sont donc

$$E\left(z_i'\left(y_i - x_i\theta\right)\right) = 0$$

Les résultats du chapitre 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'orthogonalité 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

$$\widehat{\theta}_{2mc} = \left(\overline{x_i'z_i}\ \overline{z_i'z_i}^{-1}\overline{z_i'x_i}\right)^{-1}\overline{x_i'z_i}\ \overline{z_i'z_i}^{-1}\overline{z_i'y_i}$$

Là aussi il faut mettre en oeuvre la méthode en deux étapes. Un bon choix dans ce cas est l'estimateur des 2mc, qui est certainement proche de l'estimateur optimal. On peut alors calculer un estimateur de la matrice de variance des conditions d'orthogonalité :

$$\widehat{V}(g) = \overline{\widehat{u}_{2mci}^2 z_i' z_i}$$

puis déterminer l'estimateur optimal,

$$\widehat{\theta}_S^* = \left(\overline{x_i' z_i}\ \overline{\widehat{u}_{2mci}^2 z_i' z_i}^{-1} \overline{z_i' x_i}\right)^{-1} \overline{x_i' z_i}\ \overline{\widehat{u}_{2mci}^2 z_i' z_i}^{-1} \overline{z_i' y_i}$$

ainsi que les matrice de variance de chacun des estimateurs :

$$V_{as}\left(\widehat{\theta}_{2mc}\right) = \left(\overline{x_i' z_i}\ \overline{z_i' z_i}^{-1} \overline{z_i' x_i}\right)^{-1} \overline{x_i' z_i}\ \overline{z_i' z_i}^{-1}\ \overline{\widehat{u}_i^2 z_i' z_i z_i' z_i}^{-1} \overline{x_i z_i'}\ \left(\overline{x_i' z_i}\ \overline{z_i' z_i}^{-1} \overline{z_i' x_i}\right)^{-1}$$

pour l'estimateur des doubles moindres carrés, et

$$V_{as}\left(\widehat{\theta}^*\right) = \left(\overline{x_i' z_i}\ \overline{\widehat{u}_i^2 z_i' z_i}^{-1} \overline{z_i' x_i}\right)^{-1}$$

pour l'estimateur optimal.

11.7 Test de spécification

11.7.1 Test de suridentification

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 :

$$\exists \quad \theta \quad | \qquad E\left(g\left(z_i, \theta\right)\right) = 0$$

Tous les estimateurs obtenus avec différentes métriques doivent converger vers une même valeur. Le principe est ici analogue à celui des variables instrumentales. La suridentification exprime la même idée qu'à la limite l'estimateur ne dépend pas de l'importance que l'on accorde à telle condition d'orthogonalité, tout comme le test de spécification avec les variables instrumentales exprimait qu'à la limite l'estimateur ne dépend pas de telle variable instrumentale. Il ne s'agit en fait que d'une généralisation valable pour des cas dans lesquels les conditions d'orthogonalité prennent une forme différente de celle du produit d'un résidu et d'un instrument.

Le principe du test reste le même que celui que l'on appliquerait pour tester la nullité de l'espérance d'une variable aléatoire : regarder si la moyenne empirique est proche de zéro $\overline{g\left(z_i, \theta_0\right)}$ 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

$$N\overline{\widehat{g}_i}' V_{as}\left(\overline{\widehat{g}_i}\right)^{-} \overline{\widehat{g}_i} \to \chi^2\left(\operatorname{rang} V\left(\overline{\widehat{g}_i}\right)\right)$$

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. Pour ce qui est du rang, on retrouve la même idée que pour les variables instrumentales : on teste la suridentification, c'est à dire la compatibilité du surcroît d'information introduit dans le modèle par rapport au minimum requis pour estimer le paramètre. Le rang va donc être la différence entre le nombre de conditions d'orthogonalité et la dimension du paramètre à estimer.

Proposition *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*

$$\sqrt{N}\overline{\widehat{g}_i} \simeq \sqrt{N}\overline{g_{i_0}} + G\sqrt{N}\left(\widehat{\theta}^* - \theta_0\right)$$

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

$$\sqrt{N}\overline{\widehat{g}_i} \simeq \left(I_{\dim g} - G\left(G' S^* G\right)^{-1} G' S^*\right)\sqrt{N}\overline{g_{i_0}} = \left(I_{\dim g} - P_G\right)\sqrt{N}\overline{g_{i_0}}$$

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.* $\dim\theta$ *par hypothèse. 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*

$$V_{as}\left(\overline{\widehat{g}_i}\right) = \left(I_{\dim g} - P_G\right) S^{*-1}\left(I - P_G\right)' = \left(I_{\dim g} - P_G\right) S^{*-1}$$

On en déduit immédiatement le rang de $V_{as}\left(\overline{\widehat{g}_i}\right)$ *:*

$$\operatorname{rang} V\left(\overline{\widehat{g}_i}\right) = \dim g - \dim\theta$$

et un inverse généralisé :

$$\begin{aligned} V_{as}\left(\overline{\widehat{g}_i}\right) S^* V_{as}\left(\overline{\widehat{g}_i}\right) &= \left(I_{\dim g} - P_G\right) S^{*-1} S^* \left(I_{\dim g} - P_G\right) S^{*-1} \\ &= \left(I_{\dim g} - P_G\right)^2 S^{*-1} = \left(I_{\dim g} - P_G\right) S^{*-1} \\ &= V_{as}\left(\overline{\widehat{g}_i}\right) \end{aligned}$$

d'où

$$S^* = V_{as}\left(\overline{\widehat{g}_i}\right)^-$$

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 uniformément 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^*

11.7.2 Tester la compatibilité de conditions d'orthogonalité additionnelles

On peut être amener à vouloir adjoindre à un ensemble de conditions d'orthogonalité des conditions additionnelles. Cette adjonction peut en effet conduire à des estimations plus précises. L'exemple le plus manifeste est celui dans lequel on adjoint à une liste de variables instrumentales supposées vérifier les conditions d'orthogonalité, des conditions d'orthogonalité formées en utilisant les variables explicatives comme instrument. Dans le cas homoscédastique on avait déjà envisagé ce type de test que l'on avait appelé test d'exogénéité. Cette notion peut en fait se généraliser.

Proposition *On s'intéresse au test de l'hypothèse nulle*

$$H_0 : \exists\ \theta_0\ tq\ E\left(g_1\left(z_i,\theta_0\right)\right) = 0\ et\ E\left(g_2\left(z_i,\theta_0\right)\right) = 0$$

soit

$$\exists\ \theta_0\ tq\ E\left(g\left(z_i,\theta_0\right)\right) = 0$$

où $g' = \left(g_1', g_2'\right)$ *contre l'hypothèse alternative*

$$H_1 : \exists\ \theta_0\ tq\ E\left(g_1\left(z_i,\theta_0\right)\right) = 0$$

Sous H_0 *la statistique*

$$\begin{aligned} \widehat{S} &= N\overline{g\left(z_i,\widehat{\theta}^*\right)}' \widehat{V}\left(g\left(z_i,\theta_0\right)\right)^{-1} \overline{g\left(z_i,\widehat{\theta}^*\right)} - N\overline{g_1\left(z_i,\widehat{\theta}_1^*\right)}' \widehat{V}\left(g_1\left(z_i,\theta_0\right)\right)^{-1} \overline{g_1\left(z_i,\widehat{\theta}_1^*\right)} \\ &= Q_N^0\left(\widehat{\theta}^*\right) - Q_N^1\left(\widehat{\theta}_1^*\right) \to \chi^2\left(\dim g - \dim g_1\right) \end{aligned}$$

où $\widehat{\theta}^*$ *est l'estimateur GMM optimal sous* H_0 *et* $Q_N^0\left(\widehat{\theta}^*\right) = N\overline{g\left(z_i,\widehat{\theta}^*\right)}'\widehat{V}\left(g\left(z_i,\theta_0\right)\right)^{-1}\overline{g\left(z_i,\widehat{\theta}^*\right)}$ *la valeur atteinte par l'objectif à l'optimum sous* H_0, *et* $\widehat{\theta}_1^*$ *l'estimateur GMM optimal sous* H_1 *et* $Q_N^1\left(\widehat{\theta}_1^*\right) = N\overline{g\left(z_i,\widehat{\theta}_1^*\right)}'\widehat{V}\left(g_1\left(z_i,\theta_0\right)\right)^{-1}\overline{g\left(z_i,\widehat{\theta}_1^*\right)}$ *la valeur atteinte par l'objectif à l'optimum sous* H_1.

Le test défini par la région critique $\left\{\widehat{S}\,\middle|\,\widehat{S} > q_{1-\alpha}\left(\chi^2\left(\dim g - \dim g_1\right)\right)\right\}$ *est un test convergent au niveau* α.

Ce type de test est proche des tests du rapport des maxima de vraisemblance. On pourrait en donner des équivalents correspondants au test de Hausman ou au test du multiplicateur de Lagrange.

11.7.3 Application test de suridentification et d'exogénéité pour un estimateur à variables instrumentales dans le cas univarié et hétéroscédastique

Test de suridentification

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

Corollaire *Sous l'hypothèse nulle,* $H_0 : \exists \quad \theta \quad | \qquad E\left(z_i'\left(y_i - x_i\theta\right)\right) = 0$, *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 $\widehat{S}_\chi > Q\left(1-\alpha, \chi^2\left(\dim z - \dim x\right)\right)$. On voit que l'expression de la statistique est très proche de celle vue précédemment dans le cas homoscédastique mais néanmoins différente car : elle n'est pas basée sur le même estimateur, elle n'a pas exactement la même expression, faisant intervenir $\overline{\widehat{u}_i^{*2}z_i'z_i}^{-1}$ et non $\overline{z_i'z_i}^{-1}\Big/\overline{\widehat{u}_i^{*2}}$, ce qui est une conséquence directe de l'abandon de l'hypothèse d'homoscédasticité et enfin qu'elle ne peut plus être mise en oeuvre de façon aussi directe et simple que précédemment par le biais de la régression des résidus estimés sur les variables instrumentales.

Test d'exogénéité des variables explicatives.

L'hypothèse nulle s'écrit

$$H_0 : \exists\ b_0\ tq\ E\left(z_i'\left(y_i - x_i b_0\right)\right) = 0\ et\ E\left(x_{1i}'\left(y_i - x_i b_0\right)\right) = 0$$

et l'hypothèse alternative

$$H_1 : \exists\ b_0\ tq\ E\left(z_i'\left(y_i - x_i b_0\right)\right) = 0$$

où x_{1i} représente les variables endogènes. On lui associe $\widehat{b}_0^*$ l'estimateur GMM basé sur l'ensemble les conditions d'orthogonalité de H_0 ainsi que la valeur $\widehat{S}_0$ atteinte par l'objectif à l'optimum. Dans la mesure où on ne fait plus l'hypothèse d'homoscédasticité, cet estimateur n'est pas nécessairement l'estimateur des mco : les conditions d'orthogonalité portant sur les variables instrumentales extérieures peuvent apporter une information ne se trouvant pas dans les conditions d'orthogonalité fondées sur les seules variables explicatives. On considère aussi $\widehat{b}_1^*$ l'estimateur GMM basé sur les conditions d'orthogonalité sous $H1$ ainsi que la valeur $\widehat{S}_1$ atteinte par l'objectif à l'optimum. Le résultat stipule que la statistique

$$\widehat{S}_0 - \widehat{S}_1 \rightarrow \chi^2\left(K_1\right)$$

où K_1 est le nombre de variables explicatives endogènes.

11.7.4 Application aux données de panel

On peut appliquer ces résultats à l'économétrie des données de panel. On a vu en effet que les spécifications que l'on était susceptible de retenir étaient emboîtées. Il est ainsi possible d'estimer le modèle avec l'ensemble d'information minimal, c'est à dire avec la spécification exogénéité faible. On obtient alors des estimateurs robustes à de nombreuses sources de corrélations entre variables explicatives et perturbations. En revanche, les estimateurs n'incluant que peu de restrictions ont de grandes chances d'être imprécis. On peut donc chercher à améliorer leur précision en faisant des hypothèses restrictives supplémentaires comme l'hypothèse d'effets corrélés. On peu tester les hypothèses restrictives supplémentaires par la méthode que l'on vient de détailler. Ici elle prendra la forme suivante :

1. Estimation du modèle sous la spécification exogénéité faible : On retient la valeur de l'objectif à l'optimum : $V_f = \left\| \underline{Z}_{fi}' \Delta \underline{u}_i^{f*} \right\|_{S_f^*}^2$, où S_f^* est la métrique optimale pour cette spécification.
2. Sous l'hypothèse nulle que la spécification est adaptée, la statistique V_f suit un χ^2 dont le nombre de degrés de liberté d est la différence entre le nombre de conditions d'orthogonalité et le nombre de paramètres à estimer. On peut donc calculer la

p-value associée à la statistique de test $(1 - F^{-1}(V_f, d))$ et on accepte l'hypothèse nulle si la p-value excède la valeur seuil retenue. Si on rejette l'hypothèse nulle, il faut réfléchir à une spécification alternative. Si en revanche l'hypothèse nulle est acceptée, on peut tester si des contraintes additionnelles sont compatibles avec celles d'ores et déjà retenues.

3. Estimation du modèle sous la spécification d'effets corrélés : On retient la valeur de l'objectif à l'optimum : $V_C = \left\| \underline{Z}'_{Ci} \Delta \underline{u}_i^{C*} \right\|^2_{S_C^*}$,
4. On forme la différence $V_C - V_f$ qui suit sous l'hypothèse nulle de compatibilité des conditions d'orthogonalité additionnelles un χ^2 dont le nombre de degrés de liberté est la différence entre les nombre de conditions d'orthogonalité dans les deux spécifications. On calcule la p-value de cette statistique et on accepte l'hypothèse nulle si la p-value excède le seuil retenu.
5. Si on rejette l'hypothèse on conserve l'estimateur avec exogénéité faible, sinon on peut estimer le modèle avec l'hypothèse d'exogénéité forte. On retient la valeur de l'objectif à l'optimum : $V_F = \left\| \underline{Z}'_{Fi} \Delta \underline{u}_i^{F*} \right\|^2_{S_F^*}$,
6. On procède comme au 3 et 4 en comparant les valeurs atteintes à l'optimum. On peut remarquer qu'il est possible de tester l'hypothèse de compatibilité avec soit les conditions de l'exogénéité faible soit celles des effets corrélés. Si ceci n'affecte pas la puissance du test, il n'en est pas de même avec le risque de première espèce.

11.8 Illustrations

11.8.1 Réduction du temps de travail et gains de productivité

On reprend l'illustration du chapitre précédent et on montre comment les résultats sont modifiés. Par la mise en oeuvre de la méthode des moments généralisée. On rappelle que l'équation que l'on estime s'écrit :

$$\Delta PGF_i = X_i b + \gamma RTT_i + v_i$$

où v_i représente le choc de productivité résiduel, c'est à dire une fois pris en compte les facteurs X_i.

Les variables instrumentales retenues sont : $Aide_i$, Inf_i, $Endt_i$ et Pf_i. L'intérêt de la mise en oeuvre de la méthode des moments généralisé est de pouvoir traiter le cas d'une possible (et vraisemblable) hétéroscédasticité du résidu.

On ne présente pas la condition de rang qui est la même que dans le cas précédent (tableau 10.2 du chapitre précédent). On ne présente pas de tableau de résultat mais seulement certains d'entre eux. L'estimateur à variable instrumentale usuel sert d'estimateur de première étape. Il est identique à celui du chapitre précédent : le coefficient de la variable de RTT est -0.107 et son écart-type est de 0.032, calculé avec la méthode

standard. On peut aussi calculer cet ecart-type sans faire l'hypothèse d'homoscédasticité comme on l'a expliqué plus haut. On voit qu'il n'y a pas de différence dans le calcul de cet écart-type : On trouve à nouveau 0.032. Le biais lié à la présence d'hétéroscédasticité dans l'estimation des écarts-type de l'estimateur à variables instrumentales est très faible dans le cas présent. On peut aussi calculer l'estimateur GMM optimal et son écart-type. Là aussi on ne trouve pas de différence les coefficients estimés sont les même et l'écart-type également. La seule différence notable entre les deux estimations réside en fait dans la statistique de Sargan : elle est plus faible lorsque l'on prend en compte l'hétéroscédasticité. La statistique avec l'estimateur standard (basé sur la régression du résidus sur toutes les variables exogènes) donne une statistique de 7.57 soit une p-value de 5.6% pour un $\chi^2(3)$. Avec l'estimateur optimal elle est de 6.58 soit une p-value de 8.7% : on accepte beaucoup plus facilement l'hypothèse de compatibilité des instruments. On peut aussi mettre en oeuvre le test d'exogénéité. Avec la méthode du chapitre précédent, sous hypothèse d'homoscédasticité, on procédait à une régression augmentée. Ici on fait une régression par VI par la méthode des GMM en incluant la variable de RTT dans la liste des instruments. On s'intéresse d'abord au test de compatibilité des instruments Cette hypothèse est très fortement rejetée la statistique est de 11.53 pour 4 degrés de liberté soit une p-value très faible de 2%. La statistique du test d'exognéité est la différence entre les deux statistiques de suridentification de la régression GMM avec et sans la variable de RTT. On trouve une statistique de 11.53-6.58=4.95 la aussi fortement rejeté pour un degrés de liberté de 1 (4-1).

11.8.2 Salaires et heures

On peut aussi aborder la question de la relation entre productivité et heures en examinant un équation de salaire sur des données de salarié. En effet, sous l'hypothèse que la rémunération est égale à la productivité marginale le salaire peut être utilisé comme une mesure de la productivité marginale. On peut donc considérer l'équation

$$w_i = h_i + x_i b + u_i \tag{11.1}$$

où w_i représente le logarithme du salaire et h_i le logarithme des heures. Les variable x_i sont celles qui affectent le niveau de productivité et donc les variables de capital humain : niveau d'éducation et expérience. Néanmoins dans cette régression la variable d'heure est, elle aussi, endogène. Le salaire et le nombre d'heure reflètent également un choix du salarié qui arbitre entre rémunération et loisir. Parmi toutes les offres d'emploi qu'a reçu l'individu, celle que l'on observe est celle qui est préférée (on n'aborde pas ici la question pourtant centrale du choix entre emploi et non emploi qui sera traitée dans le chapitre suivant). Pour la rémunération proposée les agents sont prêts à travailler un certain nombre d'heures qui leur est propre. Dans les préférences des salariés interviennent les caractéristiques familiales : nombre d'enfants, revenus alternatifs (conjoint, autres membres du ménage),

	parametres	std robuste	std sandards
Constante	3.8236	(0.1138)	(0.0803)
scolarité	0.0541	(0.0030)	(0.0026)
expérience	0.0197	(0.0012)	(0.0011)
(expérience-10)²	-0.0004	(0.0001)	(0.0000)
heures (log)	1.1422	(0.0315)	(0.0210)

TAB. 11.2 – Régression par les MCO

célibataire... Ces variables sont susceptibles de jouer le rôle de variables instrumentales dans la régression 11.1.

On considère un échantillon de femmes employées dans le commerce. On se restreint à la population féminine car c'est sur elle que les variables instrumentales retenues ont le plus de chance de jouer fortement. L'échantillon retenu provient de l'Enquête Emploi faite par l'INSEE et comprend 3192 individus. Le tableau 11.2 présente les résultats de la régression par les moindres carrés ordinaires. La première colonne présente le paramètre, la seconde l'écart-type robuste et la dernière l'écart-type obtenu avec la formule standard. L'intérêt principal de ce tableau est de fournir la valeur du coefficient des heures, qui s'élève ici à 1.14. Ceci signifie qu'une augmentation des heures de 1% conduit à une hausse du salaire (et donc de la productivité de 1,14%). Le coefficient est significativement différent de 1, ce qui implique qu'il y a de légers gains de productivité horaire lorsque les heures augmentent.

.

Le tableau 11.3 présente la régression de la variable explicative endogène, le logarithme des heures, sur les variables explicatives exogènes : le nombre d'année d'étude, l'expérience et l'expérience au carré et les variables instrumentales : le nombre d'enfant, l'existence de revenus alternatifs dans le foyer (salaire du conjoint, allocations chômage), le logarithme de ce revenu le cas échéant (zéro sinon), le nombre de revenus salariés dans le ménage et une indicatrice indiquant si l'individu vit seule ou non. Le tableau donne le coefficient estimé, son écart-type et son écart-type robuste. On examine l'apport des différentes variables instrumentales à l'explication de la variable endogène. On observe comme on s'y attend que plus le nombre d'enfants est élevé, plus l'incitation à travailler est faible. On observe aussi que le fait d'être célibataire conduit à des heures plus élevées. L'effet du salaire annexe sur les heures est en revanche non significatif, bien que positif.

Le tableau 11.4 présente les résultats de l'estimation du modèle par les variables instrumentale, en ignorant l'hétéroscédasticité dans la détermination de l'estimateur. L'expression de l'estimateur est donc $\widehat{b}_{IV} = \left(\overline{x_i' z_i z_i' z_i}^{-1} \overline{z_i' x_i}\right)^{-1} \overline{x_i' z_i z_i' z_i}^{-1} \overline{z_i' y_i}$. La deuxième colonne présente l'écart-type robuste et la dernière l'écart-type obtenu avec la formule valable

	parametres	std robuste	std sandards
Constante	3.3186	(0.0380)	(0.0360)
scolarité	0.0102	(0.0022)	(0.0021)
expérience	0.0045	(0.0010)	(0.0010)
(expérience-10)²	-0.0002	(0.0000)	(0.0000)
nombre d'enfants	-0.0568	(0.0070)	(0.0061)
vit seule	0.0609	(0.0167)	(0.0164)
revenu alternatif	0.0026	(0.0015)	(0.0015)

TAB. 11.3 – Régression de la variable d'heure sur les exogènes et les instruments

	parametres	std robuste	std sandards
Constante	2.5613	(0.4393)	(0.3891)
scolarité	0.0494	(0.0034)	(0.0031)
expérience	0.0193	(0.0013)	(0.0011)
(expérience-10)²	-0.0004	(0.0001)	(0.0001)
heures (log)	1.5252	(0.1312)	(0.1173)

TAB. 11.4 – Régression par les variables instrumentales

pour l'homoscédasticité du résidu. Les matrices de variance correspondantes s'écrivent $\widehat{V}_{\hom o}\left(\widehat{b}_{IV}\right) = \widehat{\sigma}^2\left(\overline{x_i'z_i}\overline{z_i'z_i}^{-1}\overline{z_i'x_i}\right)^{-1}$ et $\widehat{V}_{hetero}\left(\widehat{b}_{IV}\right) = \left(\overline{x_i'z_i}\overline{z_i'z_i}^{-1}\overline{z_i'x_i}\right)^{-1}\ \overline{x_i'z_i}\ \overline{z_i'z_i}^{-1}$ $\overline{\widehat{u}_i^2 z_i'z_i}\ \overline{z_i'z_i}^{-1}\ \overline{z_i'x_i}\ \left(\overline{x_i'z_i}\overline{z_i'z_i}^{-1}\overline{z_i'x_i}\right)^{-1}$. On observe que la variable d'heure est sensiblement plus élevé que dans la régression par les mco. Alors que la régression par les mco donne un coefficient de 1.14, le chiffre obtenu ici est nettement plus élevé puisqu'il s'élève à 1.52. Cela signifie que lorsque l'allongement du temps de travail s'accompagne de gains de productivité horaire important : une augmentation de 1% des heures conduit à une augmentation des rémunérations de 1.5%. On peut noter que ce coefficient n'est pas éloigné de celui trouvé dans l'approche par les fonctions de production lorsque l'on n'utilisait pas la variable Robien, comme instrument. On remarque aussi que le coefficient est là aussi statistiquement différent de 1 mais que l'écart-type estimé est quatre fois plus important que celui des moindres carrés ordinaires. On remarque qu'il existe des différences liées à la prise en compte de l'hétéroscédasticité mais qu'elles ne sont pas phénoménales.

Le tableau 11.5 présente les résultats obtenus par la méthode des moments généralisée. L'estimateur est donc $\widehat{b}_{IV} = \left(\overline{x_i'z_i}\widehat{\Omega}^*\overline{z_i'x_i}\right)^{-1}\overline{x_i'z_i}\widehat{\Omega}^*\overline{z_i'y_i}$, avec $\Omega^* = E\left(u_i^2 z_i'z_i\right)^{-1}$ et $\widehat{\Omega}^* = \overline{\widehat{u}_i^2 z_i'z_i}^{-1}$, où $\widehat{u}_i$ est le résidu estimé obtenu à partir d'une première étape utilisant une matrice de pondération quelconque. Le choix naturel qui est celui qui a été effectué ici consiste à se baser sur l'estimateur par variable instrumentale. On voit que les changements sont modestes par rapport au tableau précédent. C'est une bonne nouvelle à priori. Si entre

	parametres	std robuste
Constante	2.6139	(0.4373)
scolarité	0.0498	(0.0034)
expérience	0.0195	(0.0013)
(expérience-10)²	-0.0004	(0.0001)
heures (log)	1.5081	(0.1305)

TAB. 11.5 – Régression par la méthode des moments généralisée

la première et la deuxième étape, il y avait des changements importants, cela signifierait que vraisemblablement les conditions d'orthogonalité ne sont pas compatibles entre elles. Ici le fait que les résultats soient très proches signifie aussi peut être que l'hétéroscédasticité n'est pas un phénomène de premier ordre. Le coefficient auquel on parvient est de 1.51 et on observe qu'il n'est pas beaucoup plus précis que l'estimateur précédent. Dans le cas présent, les gains liés à l'utilisation de l'estimateur GMM sont assez faibles.

Enfin, on peut examiner la question de la spécification, en procédant aux tests de suridentification et d'exogénéité. Les tests ont la même interprétation que dans le cas variables instrumentales, mais la mise en oeuvre est différente. Les tests dans le cas homoscédastiques, sont effectués à partir de régressions auxiliaires : régression du résidu estimé sur les instruments et test de la nullité globale des coefficients pour le test de suridentification et régression étendue dans laquelle on introduit en plus des variables explicatives la prévision des variables endogènes par les instruments et les variables exogènes. Dans le cas GMM, on n'a pas ce genre de simplification et les tests sont basés sur l'objectif atteint par l'estimateur optimal : $S = \overline{z_i'\widehat{u}_i^*}'\widehat{\Omega}^*\overline{z_i'\widehat{u}_i^*}$. Les tests de suridentification compare la valeur obtenue de S à la valeur seuil pour un test de niveau donné. Le test d'exogénéité compare quant à lui la valeur S à la valeur S_e, obtenue avec pour ensemble d'instruments z, x_{end}. La statistique de test $S_e - S$ suit un χ^2 dont le nombre de degrés de liberté est le nombre de variables endogènes. On voit dans le tableau 11.6 que l'hypothèse de suridentification est acceptée mais pas celle d'exogénéité. Il y a en outre là aussi peu de différence entre la méthode à variables instrumentales et la méthode des moments généralisée. Les statistiques de suridentification sont très proches et les statistiques pour le test d'exogénéité, bien que non directement comparables, conduisent aux mêmes conclusions.

Enfin le tableau 11.7 présente les résultats pour différents secteurs. Les deux premières colonnes donnent la valeur du paramètre et son écart-type en utilisant pour instruments le fait d'être célibataire, le nombre d'enfants et le revenu alternatif. Les deux colonnes suivantes présentent le test de Sargan et sa p-value. On présente le test d'exogénéité. Ceci n'est pas effectué pour les Industries Agricoles, le Transport et la Finance puisque dans ces secteurs, le test de validité de suridentification conduit au rejet de l'hypothèse de

Test	Statistique	degrés	pvalue
		GMM	
Suridentification	2.522	2.000	0.283
Exogénéité	8.650	1.000	0.003
		VI	
Suridentification	2.805	2.000	0.246
Exogénéité	1.128		
	(0.021)		

TAB. 11.6 – Tests de spécification

	Par	std	S	p	S(e)	p(e)	Par	std	S	p
Industries Agricoles	0.51	(0.67)	8.33	0.02						
Biens de consommation	1.68	(0.71)	2.85	0.24	0.91	0.34	1.13	(0.09)	3.76	0.29
Automobiles et Equipements	0.79	(0.38)	4.13	0.13	2.01	0.16	1.22	(0.07)	6.15	0.10
Biens Intermédiaires	1.04	(0.26)	0.77	0.68	0.08	0.77	0.98	(0.05)	0.85	0.84
Commerce	1.51	(0.13)	2.52	0.28	8.65	0.00				
Transport	1.92	(0.52)	2.42	0.30	2.76	0.10	1.19	(0.08)	5.18	0.16
Finance	1.20	(0.24)	6.02	0.05						
Services Entreprises	1.23	(0.16)	10.09	0.01						
Services Particuliers	2.69	(0.48)	0.14	0.93	82.10	0.00				
Education Santé	1.18	(0.11)	4.76	0.09	18.02	0.00				
Administration	1.30	(0.15)	3.13	0.21	4.87	0.03				

TAB. 11.7 – Résultats Sectoriels

compatibilité des instruments. On ne peut donc pas tester la compatibilité de restrictions identifiantes supplémentaires. Les colonnes 7 et 8 présentent la valeur du paramètre estimé en utilisant comme instruments les trois variables retenues et la variable d'heure. Enfin les deux dernières colonnes présentent le test de suridentification lorsque l'on utilise tous ces instruments. On vérifie que la valeur de la statistique est la somme des statistiques obtenus dans les colonnes (3) et (5). Ce que montre ce tableau est que les instruments ne sont pas toujours considérés comme compatibles. Lorsqu'ils le sont les valeurs sont assez différentes d'un secteur à l'autre, quoique toujours supérieure à 1. On voit aussi que les estimations sont peu précises et que lorsque l'hypothèse d'exogénéité est acceptée, on obtient des gains d'efficacité non négligeables.

11.9 Résumé

Dans ce chapitre on a présenté une méthode d'estimation très générale, englobant la totalité des méthodes vues jusqu'à présent. Elle permet aussi de considérer facilement des généralisations utiles des situations envisagées jusqu'à présent. En particulier elle permet

de généraliser la méthode des variables instrumentales aux cas hétéroscédastiques et au cas de systèmes d'équations.

1. Cette méthode est basée sur l'exploitation de conditions d'orthogonalité, qui sont des fonctions des variables et des paramètres du modèle dont l'espérance est nulle.
2. Le principe de la méthode des moments généralisée consiste à choisir le paramètre de telle sorte que la contrepartie empirique des conditions d'orthogonalité soit le plus proche possible de zéro.
3. Lorsqu'il y a juste identification, c'est à dire lorsque le nombre de paramètre à estimer est le même que le nombre de conditions d'orthogonalité, on peut exactement annuler (en général) les contreparties empiriques des conditions d'orthogonalité.
4. Lorsqu'il y a plus de conditions d'orthogonalité que de paramètres à estimer, on est dans la situation dite de suridentification. On ne peut en général pas annuler directement la contrepartie empirique des conditions d'orthogonalité. On minimise alors la norme de ces contreparties.
5. Les estimateurs auxquels on parvient sont sous certaines hypothèses de régularité convergents et asymptotiquement normaux. La convergence ne dépend pas de la métrique choisie pour estimer mais la matrice de variance de l'estimateur si.
6. Parmi tous les estimateurs envisageable, il en existe un plus précis que tous les autres : c'est l'estimateur GMM optimal. Il est obtenu en utilisant pour métrique l'inverse de la matrice de variance des conditions d'orthogonalité.
7. La méthode des moments généralisée permet comme la méthode des variables instrumentale de procéder à des tests de spécification. Il est ainsi possible de tester la compatibilité des conditions d'orthogonalité entre elles (à l'instar des tests de compatibilité des variables instrumentales). Ce test est un test de compatibilité et pas un test de validité.
8. La méthode permet aussi de tester la compatibilité d'un ensemble de conditions d'orthogonalité additionnel avec un ensemble de conditions d'orthogonalité initial dont la validité constitue l'hypothèse alternative.

Chapitre 12

Variables dépendantes limitées

On a examiné jusqu'à présent le cas de modèles linéaires pour lesquels la variable dépendante y_i avait pour support $\Re$. On examine dans ce chapitre trois types de modèles aux applications très nombreuses et qui sont des extensions directes du modèle linéaire : Les modèles dichotomiques, les modèles Tobit et le modèle Logit Multinomial

- 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". On va présenter dix modèles très couramment utilisés pour modéliser ce type de situation : les modèles Logit et les modèles Probit et on va insister sur la relation entre la modélisation statistique des variables prenant leurs valeurs dans $\{0, 1\}$ et la modélisation économique. Ceci va nous conduire à introduire la notion importante de variable latente : une variable dont le support peut être $\mathcal{R}$ mais qui n'est qu'en partie observée. On est ainsi conduit à modéliser cette variable, ce qui correspond à une modélisation économique (dans le cas de la faillite d'une entreprise il peut s'agir de la valeur des profits futurs de l'entreprise), et à modéliser aussi la façon dont une censure s'opère dans les observations, ce qui peut résulter là aussi d'un comportement économique (dans le cas de la faillite il peut s'agir du fait que la valeur de l'entreprise passe sous un certain seuil) mais aussi d'une caractéristique statistique des données.
- Le modèle logit Multinomial 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 $y_i \in \{0, 1, 2, \ldots, M\}$. Le modèle que l'on va introduire est très utilisé dans de nombreux domaines appliqués. Il insiste lui aussi sur la modélisation économique. L'idée générale est qu'à chaque modalité est associée une valeur dépendant des préférences intrinsèques d'un individu mais aussi de caractéristiques économiques telles que les prix ou le revenu. Le choix sélectionné

par un individu est celui correspondant à la valorisation maximale. Ce type de modélisation, du à l'origine à Mac Fadden, est très utilisé dans la modélisation des systèmes de demande pour des biens différenciés et intervient souvent en économie industrielle empirique.

– Le Modèle Tobit est un modèle central dans l'analyse économique. Il correspond à la prise en compte de sélectivité dans les observations : le fait que l'on observe un phénomène n'est pas indépendant de ce phénomène. Pour l'analyser il faut donc modéliser le phénomène et les conditions qui conduisent à son observation. 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. Cette modélisation fait comme le modèle Probit appelle à des variables latentes. Il existe différents types de modèles Tobit qui correspondent à autant de situations économiques. Le classement de ces situations en différents types de modèles Tobit est du à Amemiya. Il y a ainsi des modèles Tobit de type I, de type II, de type III, IV et V. On va voir dans ce chapitre les modèles de type I à III.

12.1 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.

$$P\left(y_i = 1 \left| x_i \right.\right) = \widetilde{G}\left(x_i\right)$$

qui définit complètement la loi conditionnelle de y_i sachant x_i. Cette probabilité est aussi l'espérance conditionnelle de la variable y_i :

$$\begin{aligned} E\left(y_i \left| x_i \right.\right) &= \sum_{y_i \in \{0,1\}} y_i \left[1_{(y_i=1)} P\left(y_i = 1 \left| x_i \right.\right) + 1_{(y_i=0)} \left(1 - P\left(y_i = 1 \left| x_i \right.\right)\right)\right] \\ &= P\left(y_i = 1 \left| x_i \right.\right) = \widetilde{G}\left(x_i\right) \end{aligned}$$

On spécifie en général cette fonction comme dépendant d'un indice linéaire en x_i :

$$\widetilde{G}\left(x_i\right) = G\left(x_i b\right)$$

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.

12.1.1 Modèle à probabilités linéaires

C'est la situation dans laquelle on spécifie simplement

$$E\left(y_i\left|x_i\right.\right)=P\left(y_i=1\left|x_i\right.\right)=x_ib$$

Le modèle peut alors être estimé par les MCO.

En dépit de sa simplicité attractive, ce choix de modélisation présente néanmoins l'inconvénient majeur que le modèle ne peut contraindre $P\left(y_i=1\left|x_i\right.\right)=x_ib$ à appartenir à l'intervalle $[0,1]$. Il y a donc une incohérence dans cette modélisation.

Un autre 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 général, c'est à dire basée sur la spécification $E\left(y_i\left|x_i\right.\right)=G\left(x_ib\right)$, doit prendre en compte le fait que le modèle de régression correspondant

$$y_i=G\left(x_ib\right)+u_i$$

est hétéroscédastique. En effet on a :

$$\begin{aligned}V\left(y_i\left|x_i\right.\right) &= E\left(y_i^2\left|x_i\right.\right)-E\left(y_i\left|x_i\right.\right)^2=E\left(y_i\left|x_i\right.\right)-E\left(y_i\left|x_i\right.\right)^2\\ &= E\left(y_i\left|x_i\right.\right)\left[1-E\left(y_i\left|x_i\right.\right)\right]=G\left(x_ib\right)\left[1-G\left(x_ib\right)\right]\end{aligned}$$

L'estimateur des mco dans le cas linéaire a donc pour variance

$$V_{as}\left(\widehat{b}_{mco}\right)=E\left(x_i^{'}x_i\right)^{-1}E\left(u_i^2x_i^{'}x_i\right)E\left(x_i^{'}x_i\right)^{-1}$$

que l'on estime par la méthode de White

$$\widehat{V}_{as}\left(\widehat{b}_{mco}\right)=\overline{x_i^{'}x_i}^{-1}\overline{\widehat{u}_i^2x_i^{'}x_i}\ \overline{x_i^{'}x_i}^{-1}$$

On pourrait être tenté d'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 connaît l'expression de la matrice de variance des résidus conditionnellement à x_i : $E\left(u_i^2\left|x_i\right.\right)=G\left(x_ib\right)\left(1-G\left(x_ib\right)\right)=\sigma^2\left(x_ib\right)$. 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 avec le modèle de probabilité linéaire puisqu'il n'est pas exclu que $x_ib\left(1-x_ib\right)$ soit négatif.

12.1.2 Les modèles probit et logit.

Il est 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**

Effet marginal d'une variation d'un régresseur continu x

L'un des avantages majeurs du modèle de probabilité linéaire est qu'une variation marginale d'un régresseur a un effet constant dans la population. Cette propriété simple et attractive n'existe plus dans le cas des modèles probit ou logit. On peut néanmoins préciser l'effet d'une variable sur la probabilité conditionnelle d'observer l'événement modélisé. 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

$$\frac{\partial Log E(y_i | x_i)}{\partial x_i^k} = \frac{G'(x_i b)}{G(x_i b)} b_k$$

Pour le modèle Probit on a ainsi :

$$\frac{\partial E(y_i | x_i)}{\partial x_i^k} = \varphi(x_i b) b_k, \quad \frac{\partial Log E(y_i | x_i)}{\partial x_i^k} = \frac{\varphi(x_i b)}{\Phi(x_i b)} b_k$$

et pour le modèle Logit

$$\begin{aligned}\frac{\partial E\left(y_i\left|x_i\right.\right)}{\partial x_i^k} &= F\left(x_i b\right)\left(1-F\left(x_i b\right)\right)b_k \\ \frac{\partial LogE\left(y_i\left|x_i\right.\right)}{\partial x_i^k} &= \left(1-F\left(x_i b\right)\right)b_k\end{aligned}$$

puisqu'on vérifie facilement $F^{'}=F\left(1-F\right)$.

L'effet marginal de l'accroissement d'un facteur dépend donc du point où l'on se situe. En pratique on est amené à considérer une situation de référence qui peut être un groupe d'individus lorsque les variables explicatives sont elles mêmes des variables de catégories, ou bien le point moyen de l'échantillon. Dans ce cas par exemple, on calculerait

$$\frac{\partial E\left(y_i\left|\overline{x}_i\right.\right)}{\partial x_i^k}=G^{'}\left(\overline{x}_i b\right)b_k$$

12.2 Variables latentes

La modélisation précédente est une modélisation statistique. Les modèles à variables dépendantes discrètes peuvent souvent être introduits en rendant plus explicites les hypothèses économiques sous-jacentes à la modélisation. Ceci est effectué par le biais de ce que l'on appelle une variable latente, c'est à dire une variable inobservée mais qui détermine complètement la réalisation de la variable indicatrice étudiée. Dans le cas présent, on modélise la réalisation de la variable indicatrice étudiée par le biais d'une variable :

$$y_i^*=x_i b+u_i$$

Dans cette modélisation on suppose que le résidu intervenant dans l'expression de la variable latente est indépendant des variables explicatives. La variable latente y_i^* n'est jamais observée complètement mais elle est liée à la réalisation de la variable d'intérêt par :

$$y_i=1\Leftrightarrow y_i^*>0\Leftrightarrow x_i b+u_i>0$$

Lorsque l'on spécifie la loi du résidu u_i, on est capable de définir complètement la probabilité $P\left(y_i=1\left|x_i\right.\right)$. Si on suppose que le résidu intervenant dans modélisation de la variable latente est normal, on obtient le modèle Probit. 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 $v_i = u_i/\sigma \rightsquigarrow N(0,1)$. Les paramètres b sont identifiables à un facteur multiplicatif prêt. Si on pose $c = b/\sigma$, on a

$$\begin{aligned} P(y_i = 1 \,|x_i) &= P\left(x_i \frac{b}{\sigma} + \frac{u_i}{\sigma} > 0\right) = P(v_i > -x_i c) = P(v_i < x_i c) \\ &= \Phi(x_i c) \end{aligned}$$

où on utilise le fait que la loi normale est symétrique, et que donc $P(v > a) = P(v < -a)$.

Exemple *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 là encore que l'on soit capable 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 alors la variable latente associée au modèle.

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 $P(y_i = 1 \,|x_i)$.

12.3 Estimation des modèles dichotomiques

Mis à part le modèle de probabilité linéaire qui s'estime directement par les MCO, les modèles dichotomiques s'estiment par le maximum de vraisemblance. En effet la spécification de la probabilité conditionnelle conduit à spécifier entièrement la loi des observations. Compte tenu d'une modélisation conduisant à

$$P\left(y_i = 1 \left| x_i \right.\right) = G\left(x_i b\right)$$

avec G une fonction de répartition connue, de densité g. La probabilité d'observer y_i pour un individu peut s'écrire comme

$$\begin{aligned} P\left(y_i \left| x_i \right.\right) &= P\left(y_i = 1 \left| x_i \right.\right)^{y_i} \left[1 - P\left(y_i = 1 \left| x_i \right.\right)\right]^{1-y_i} \\ &= G\left(x_i b\right)^{y_i} \left[1 - G\left(x_i b\right)\right]^{1-y_i} \end{aligned}$$

La vraisemblance de l'échantillon s'écrit donc

$$L\left(y \left| x \right.\right) = \prod_{i=1}^{N} P\left(y_i \left| x_i \right.\right) = \prod_{i=1}^{N} G\left(x_i b\right)^{y_i} \left[1 - G\left(x_i b\right)\right]^{1-y_i}$$

compte tenu de l'hypothèse d'indépendance. La $\log -$vraisemblance s$'$écrit alors

$$\log L_N = \sum_{i=1}^{N} \left[y_i \log G\left(x_i b\right) + \left(1 - y_i\right) \log\left(1 - G\left(x_i b\right)\right)\right]$$

Lorsque l'on fait l'hypothèse que les observations sont indépendantes, la maximisation de la vraisemblance conduit à des estimations convergentes. On a vu en effet dans le chapitre précédent que la méthode du maximum de vraisemblance, basée sur la nullité de l'espérance du score

$$E \frac{\partial \log L\left(z_i, \theta\right)}{\partial \theta} = 0 \Leftrightarrow \theta = \theta_0$$

est une méthode de type GMM et que l'on peut étudier les propriétés asymptotiques des estimateurs dans le cadre général de la convergence des estimateurs GMM. On rappelle ici les principaux résultats de la méthode des moments généralisée et leur transcription au cas et leur transcription au cas du maximum de vraisemblance.

On considère un modèle dont la vraisemblance s'écrit $L\left(z_i, \theta\right)$

Proposition *Sous les hypothèses*

1. *H*1 *L'espace des paramètres* Θ *est compact. La vraie valeur est* θ_0 *intérieure à* Θ,
2. *H*2 $\exists$, $\theta_0 \in \Theta$ *tq* $L\left(z_i, \theta_0\right)$ *est la vraie densité des observations*
3. *H*3 $L\left(z_i, \theta\right)$ *est deux fois continûment dérivable en* θ,

4. $H4$ $E\left[\sup_{\theta}\left|\partial \log L\left(z_i,\theta\right)/\partial\theta\right|+\sup_{\theta}\left|\partial \log L\left(z_i,\theta\right)/\partial\theta\right|^2+\sup_{\theta}\left|\partial^2 \log L\left(z_i,\theta\right)/\partial\theta\partial\theta'\right|\right]<\infty$,
5. $H5$ $\partial \log L\left(z_i,\theta\right)/\partial\theta_k$ *a des moments finis d'ordre* 1 *et* 2,
6. $H6$ *Le Jacobien* $J=E\left(\partial^2 \log L\left(z_i,\theta_0\right)/\partial\theta\partial\theta'\right)$ *de dimension* $\dim\theta\times\dim\theta$ *est de* $\operatorname{rang}\dim\theta$,

Alors l'estimateur du maximum de vraisemblance $\widehat{\theta}_{SN}$ *maximisant* $Q_N\left(\theta\right)=\overline{LogL\left(z_i,\theta\right)}$, *vérifie les propriétés :*

1. $\widehat{\theta}_S\xrightarrow{P}\theta_0$ *convergence*
2. $\sqrt{N}\left(\widehat{\theta}_S-\theta_0\right)\xrightarrow{L}N\left(0,V_{as}\left(\widehat{\theta}\left(S\right)\right)\right)$ *normalité asymptotique*
3. $V_{as}\left(\widehat{\theta}_S\right)=J^{-1}=I^{-1}$ *où* $I=E\left[\partial \log L\left(z_i,\theta\right)/\partial\theta\;\partial \log L\left(z_i,\theta\right)/\partial\theta'\right]$
4. $\widehat{I}=\overline{\partial \log L\left(z_i,\widehat{\theta}\right)\Big/\partial\theta\;\partial \log L\left(z_i,\widehat{\theta}\right)\Big/\partial\theta'}\rightarrow I$ *et* $\widehat{J}=\overline{\partial^2 \log L\left(z_i,\widehat{\theta}\right)\Big/\partial\theta\partial\theta'}\rightarrow J$

Démonstration *Il s'agit d'une transcription directe des résultats concernant la convergence de l'estimateur de la méthode des moments généralisée au_ cas du score* $E\frac{\partial \log L(z_i,\theta)}{\partial\theta}=0$, *à quelques exception prés. On a vu que si le modèle est bien spécifié, c'est à dire si effectivement la densité des observations peut être paramètrée par le modèle utilisé, alors la vraisemblance est maximale pour la vraie valeur des paramètres. C'est le sens de la condition* $HMV2$ *analogue de la condition* $H2$ *de la méthode des moments généralisée. Par rapport à la méthode des moments généralisée, une caractéristique importante provient du fait que le modèle est juste identifié. L'expression de la matrice de variance en est simplifiée.*

Dans le cas général son expression est $V_{as}\left(\widehat{\theta}_S\right)=\left[G'S_0G\right]^{-1}G'S_0V\left(g\left(z_i,\theta_0\right)\right)S_0G\left[G'S_0G\right]^{-1}$. *Ici les notations sont différentes,* $G=J$ *et* $V=I$ *et en outre* G *est de dimension* $\dim\theta\times\dim\theta$ *puisque* $\dim g=\dim\theta$ *et de rang* $\dim\theta$ *par hypothèse.* G *est donc inversible, d'où une expression plus simple* $V_{as}\left(\widehat{\theta}_S\right)=J^{-1}IJ^{-1}$.

Une simplification supplémentaire provient du fait qu'il s'agit d'une vraisemblance. On a alors :

$$E\left(\frac{\partial^2 \log L\left(z_i,\theta\right)}{\partial\theta\partial\theta'}\right)=-E\left[\frac{\partial \log L\left(z_i,\theta\right)}{\partial\theta}\frac{\partial \log L\left(z_i,\theta\right)}{\partial\theta}'\right].$$

Cette dernière relation provient simplement du fait que pour une famille de densité de probabilité $f\left(x,\theta\right)$,

$$\int f\left(x,\theta\right)dx=1$$

donc

$$\int\frac{\partial f}{\partial\theta}\left(x,\theta\right)dx=0\;\;soit\;\int f\left(x,\theta\right)\frac{\partial Logf}{\partial\theta}\left(x,\theta\right)=0,\;\;i.e.E_\theta\left(\frac{\partial Logf}{\partial\theta}\right)=0$$

En dérivant à nouveau en θ, il vient

$$\begin{aligned}
\int f(x,\theta)\frac{\partial^2 Logf}{\partial\theta\partial\theta'}(x,\theta)\,dx+\int\frac{\partial Logf}{\partial\theta}(x,\theta)\frac{\partial f}{\partial\theta'}(x,\theta)\,dx &= 0\\
\int f(x,\theta)\frac{\partial^2 Logf}{\partial\theta\partial\theta'}(x,\theta)\,dx+\int\frac{\partial Logf}{\partial\theta}(x,\theta)\frac{\partial Logf}{\partial\theta'}(x,\theta)\,f(x,\theta)\,dx &= 0\\
E_\theta\left(\frac{\partial^2\log f(x,\theta)}{\partial\theta\partial\theta'}\right)+E_\theta\left[\frac{\partial\log f(x,\theta)}{\partial\theta}\frac{\partial\log f(x,\theta)'}{\partial\theta}\right] &= 0
\end{aligned}$$

Finalement on retrouve a à partir des formules GMM que dans le cas du maximum de vraisemblance

$$V_{as}\left(\widehat{\theta}\right)=-E\left(\frac{\partial^2\log L(z_i,\theta)}{\partial\theta\partial\theta'}\right)^{-1}=E\left(\frac{\partial\log L(z_i,\theta)}{\partial\theta}\frac{\partial\log L(z_i,\theta)'}{\partial\theta}\right)^{-1}$$

12.3.1 Conditions de 1er ordre pour la maximisation

L'estimateur du maximum de vraisemblance est défini par :

$$\frac{\partial\log L_N}{\partial\,\beta}=\sum_{i=1}^{N}\left[y_i\frac{g\left(x_i\widehat{b}\right)}{G\left(x_i\widehat{b}\right)}+(1-y_i)\frac{-g\left(x_i\widehat{b}\right)}{1-G\left(x_i\widehat{b}\right)}\right]x_i^{'}=0$$

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 assez similaires aux conditions vues pour les moindres carrés, mis à part la pondération et la non linéarité. On remarque également que la pondération s'interprète naturellement par le fait que $V(y_i|x_i)=G(x_i,b)(1-G(x_i,b))$, et que $g(x_i,b)\,x_i'$ est la dérivée par rapport à b de $G(x_ib)$. La pondération est donc analogue à la sphéricisation pratiquée dans la méthode des mCQG du modèle linéarisé autour de la vraie valeur du paramètre.

Pour le modèle Logit on a $G(z) = F(z) = 1/(1+\exp(-z))$, et $g(z) = \exp(-z)/(1+\exp(-z))^2 = F(z)(1-F(z))$. 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 $G(z) = \Phi(z)$, et $g(z) = \varphi(z)$. On a donc simplement

$$\left.\frac{\partial \log L_N}{\partial\, b}\right|_{\Pr obit} = \sum_{i=1}^{N}\left[y_i - \Phi\left(x_i\widehat{b}\right)\right]\frac{\varphi\left(x_i\widehat{b}\right)}{\Phi\left(x_i\widehat{b}\right)\left[1-\Phi\left(x_i\widehat{b}\right)\right]}x_i^{'} = 0$$

12.3.2 Dérivées secondes de la log-vraisemblance - condition de concavité

On sait qu'asymptotiquement, la vraisemblance a un maximum global unique. Ceci ne signifie pas qu'il n'y ait pas de maximum local. Ceci ne signifie pas non plus qu'il n'y ait pas à distance fini des maxima locaux. Il est donc important d'examiner les conditions du second ordre de l'objectif maximisé qui permettent d'étudier l'existence d'optima multiples. On montre que dans le cas du modèle probit et du modèle logit on est dans un cas favorable dans lequel la matrice hessienne est toujours négative : la log-vraisemblance est donc globalement concave. Ceci garantit donc que l'optimum trouvé est bien celui qu'il faut considérer.

Pour le modèle Logit, on le vérifie directement aisément. La matrice des dérivées secondes de l'objectif a en effet pour expression :

$$H = \left.\frac{\partial^2 \log L_N}{\partial b \partial b'}\right|_{Logit} = -\sum_{i=1}^{N}\left[1-F\left(x_i\widehat{b}\right)\right]F\left(x_i\widehat{b}\right)x_i x_i^{'}$$

Pour le modèle probit on montre plus généralement une proposition basée sur la log concavité de la densité. On présente d'abord un lemme :

Lemme *Si* $\log(g)$ *est concave, alors le ratio* $g(z)/G(z)$ *est une fonction décroissante de* z.

Démonstration $\frac{g(z)}{G(Z)}$ *est décroissant si* $g'G < g^2$ *c'est à dire si* $\frac{g'}{g}G < g$. *Si* $\log(g)$ *est concave alors* $\frac{g'}{g}$ *décroissante. 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* $g(z) > \frac{g'(z)}{g(z)}G(z)$.

Proposition *Si* $\log(g)$ *est concave et si* g *est symétrique, alors le hessien de la vraisemblance du modèle dichotomique à probabilité* $G(x_{ib})$ *est défini négatif.*

Démonstration *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 :

$$\frac{\partial^2 \log L_N}{\partial b \partial b'} = \sum_{I_1} \left(\frac{g_i}{G_i}\right)' x_i' x_i + \sum_{I_0} \left(-\frac{g_i}{1 - G_i}\right)' x_i' x_i$$

Comme g est symétrique $G(-z) = 1 - G(z)$, on a $-\frac{g(z)}{1-G(z)} = -\frac{g(-z)}{G(-z)}$, il en résulte que si $\frac{g}{G}$ est une fonction décroissante, alors $-\frac{g(z)}{1-G(z)}$ est aussi une fonction décroissante. Le Hessien est négatif puisque les dérivées des ratios $\frac{g_i}{G_i}$ et $-\frac{g_i}{1-G_i}$ sont négatives.

Dans le cas Probit, $g(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z^2\right)$, c'est bien une fonction symétrique et $\log g(z) = -\log \sqrt{2\pi} - \frac{1}{2} z^2$, est bien une fonction concave. L'objectif est donc globalement concave.

12.3.3 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 $\hat{\beta}$:

$$\hat{V}_{as}(\widehat{b}) = \left(\overline{\frac{\partial \log L\left(y_i, x_i, \widehat{b}\right)}{\partial b} \left(\frac{\partial \log L\left(y_i, x_i, \widehat{b}\right)}{\partial b}\right)'}\right)^{-1}$$

Compte tenu de l'expression donnée plus haut

$$\frac{\partial \log L\left(y_i, x_i, \widehat{b}\right)}{\partial b} = \omega\left(x_i, \widehat{b}\right)\left[y_i - E\left(y_i \left| x_i, \widehat{b}\right.\right)\right] x_i'$$

avec $\omega\left(x_i, \widehat{b}\right) = g\left(x_i\widehat{b}\right) \Big/ G\left(x_i\widehat{b}\right)\left[1 - G\left(x_i\widehat{b}\right)\right]$, on note que dans ce cas la matrice de variance s'écrit sous une forme 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 $\widehat{\varepsilon}_i = y_i - G\left(x_i, \widehat{b}\right)$

La matrice de variance covariance de l'estimateur est dans tous les cas estimée par

$$\hat{V}(\widehat{b}) = \hat{V}_{as}(\widehat{b})/N$$

12.4 Illustration : participation des femmes sur le marché du travail

On peut mettre en oeuvre les méthodes d'estimation précédentes en examinant le comportement de participation des femmes sur le marché du travail. La modélisation de la décision de participation fait intervenir le salaire de marché w_i et le salaire de réservation $\overline{w}_i$. Le salaire de marché est modélisé comme une fonction du capital humain, c'est à dire comme une fonction de la scolarité et l'expérience sur le marché du travail. Le salaire de réservation est fonction lui de la situation familiale : revenu alternatif, célibat, nombre d'enfants... Au lieu de modéliser le capital humain par l'expérience, fonction des décisions passées de participation sur le marché du travail, on peut faire intervenir directement l'age. Au total on a une décision de participation prenant la forme :

$$\begin{aligned} I &= 1 \Longleftrightarrow w_i > \overline{w}_i \\ w_i &= \alpha_0 + \alpha_1 sco_i + \alpha_2 age_i + \alpha_3 age_i^2 + u_i \\ \overline{w}_i &= \beta_0 + \beta_1 wa_i + \beta_2 \sin gle_i + \beta_3 nenf_i + \beta_4 age_i + \beta_5 age_i^2 v_i \end{aligned}$$

On a donc la modélisation de participation :

$$I = 1 \Longleftrightarrow \gamma_0 + \gamma_1 sco_i + \gamma_2 age_i + \gamma_3 age_i^2 + \gamma_4 wa_i + \gamma_5 \sin gle_i + \gamma_6 nenf_i + \omega_i > 0$$

On peut estimer ce modèle en faisant l'hypothèse que les résidus sont distribués de telle sorte que l'on ait un modèle Probit, Logit ou à probabilité linéaire. On met en oeuvre cette estimation sur un échantillon de femmes en 2002, tiré de l'enquête emploi. L'échantillon comprend 36249 femmes. Les résultats sont présentés dans le tableau 12.1. On voit que

	Probit		Logit		Linéaire		
	b	sb	b	sb	b	sbh	sb
Constante	-0.207	(0.057)	-0.379	(0.095)	0.441	(0.020)	(0.019)
Nenf	-0.317	(0.008)	-0.530	(0.013)	-0.108	(0.002)	(0.002)
wa	0.043	(0.002)	0.071	(0.003)	0.015	(0.001)	(0.001)
single	0.297	(0.024)	0.490	(0.039)	0.103	(0.008)	(0.008)
scolarité	0.089	(0.003)	0.151	(0.005)	0.029	(0.001)	(0.001)
age	-0.006	(0.001)	-0.010	(0.001)	-0.002	(0.000)	(0.000)
age²/1000	-0.237	(0.008)	-0.401	(0.013)	-0.081	(0.003)	(0.003)

TAB. 12.1 – Estimation du modèle de participation des femmes

les paramètres sont distincts d'une régression à l'autre mais que les sens de variations sont toujours les mêmes. On note aussi que les estimations sont très précises, ce qui tient à la taille importante de l'échantillon. Les résultats sont bien ceux auxquels on s'attend : plus le capital humain est important : âge et scolarité élevés, plus la participation est importante. De même plus le nombre d'enfants est élevé, moins la participation est élevée. Le célibat conduit aussi comme on s'y attend à une participation plus importante. On remarque enfin que le revenu alternatif (celui du conjoint) n'a pas le signe attendu. On aurait pu penser en effet que le salaire du conjoint conduisait à une participation plus faible. Ceci pourrait être lié au fait que dans la décision de mise en couple les capacités sur le marché du travail des deux individus sont corrélées positivement.

Pour aller plus loin dans la comparaison des estimateurs entre eux, il faudrait comparer les effets marginaux, c'est à dire calculer en chaque point l'effet prédit par le modèle d'un accroissement marginal de la variable.

12.5 Sélectivité : le modèle Tobit

12.5.1 Présentation de la sélectivité

La sélectivité est une des causes principales de biais dans les estimations des modèles linéaires. Elle correspond à la situation dans laquelle le phénomène que l'on étudie est observé uniquement sous certaines conditions qui ne sont pas indépendantes du phénomène étudié. Pour certains individus, on n'observe pas le phénomène étudié, il y a donc un problème de "données manquantes", et la raison pour laquelle on n'observe pas le phénomène est elle même liée à ce phénomène. Le fait de ne pas observer le phénomène apporte donc paradoxalement une information sur le phénomène lui-même. On dit dans ce cas que le processus de sélection n'est pas ignorable.

Exemple *Le modèle d'offre de travail d'Heckman. Pour illustrer le problème de la sélectivité on présente le modèle d'offre de travail d'Heckman. On modélise le salaire de marché*

d'un individu comme :

$$w_i^* = x_i b + u_i$$

avec x_i comprenant les variables affectant le capital humain : la scolarité et l'âge (à la place de l'expérience) et le salaire de réserve comme

$$\overline{w}_i = x_{ri} b_r + u_{ri}$$

avec x_{ri} comprenant le nombre d'enfant, une indicatrice valant 1 en cas de célibat, le cas échéant, le revenu du conjoint. On introduit en plus de ces variables un polynôme de l'âge pour prendre en compte les spécificités du marché du travail français qui subventionne le retrait d'activité des travailleurs âgés. On introduit en outre une modélisation des heures. Les heures de travail offertes dépendent de l'écart entre le salaire de marché et le salaire de réserve :

$$h_i^* = \gamma \left(w_i^* - \overline{w}_i \right)$$

et on a donc un nombre d'heures non nul, donc observé si $w_i^* > \overline{w}_i$. Le paramètre γ est particulièrement intéressant puisqu'il correspond à l'élasticité de l'offre de travail au salaire. A cette modélisation correspond différentes possibilités d'observation.

1. *On n'observe que la décision de participation :*

$$\begin{cases} p_i = 1 & si\ h_i^* > 0 \\ p_i = 0 & si\ h_i^* \leq 0 \end{cases}$$

   Il s'agit du modèle Probit déjà examiné.

2. *On observe la décision de participation et le nombre d'heures :*

$$\begin{cases} \begin{cases} h_i = h_i^* = \gamma x_i b - x_{ri} b_r + \gamma u - u_{ri} = z_{ic} + v_i \\ p_i = 1 \end{cases} & \begin{array}{l} si\ h_i^* > 0 \\ si\ h_i^* \leq 0 \end{array} \\ p_i = 0 \end{cases}$$

   Il s'agit du modèle Tobit dit simple ou de type I car la variable définissant la censure est aussi celle qui est observée lorsqu'il n'y a pas censure. Dans le cas considéré ici, il est clair que l'estimation de ce modèle ne permet pas l'estimation simple du paramètre d'élasticité d'offre de travail au salaire. On peut identifier $l\left(h_i^* \left| z_i, h_i^* > 0\right.\right)$ qui est bien sur différente de $l\left(h_i^* \left| z_i\right.\right)$. Le processus de sélection n'est donc pas ignorable dans ce cas de façon évidente.

3. *On observe le salaire et la décision de participation*

$$\begin{cases} \begin{cases} w_i = x_i b + u_i \\ p_i = 1 \end{cases} & \begin{array}{l} si\ h_i^* > 0 \\ si\ h_i^* \leq 0 \end{array} \\ p_i = 0 \end{cases}$$

   Il s'agit du modèle Tobit dit de type II car la variable définissant la censure n'est pas celle qui est observée lorsqu'il n'y a pas censure. On peut identifier ici $l\left(w_i^* \left| z_i, h_i^* > 0\right.\right)$

qui peut être différente ou non de $l(w_i^* | z_i)$. Le processus de sélection peut donc être ignorable ou non dans ce cas. On voit que si $l(w_i^* | z_i, h_i^*) = l(w_i^* | z_i)$, c'est à dire si la variable réalisant la censure est indépendante de la variable étudiée conditionnellement aux variables explicatives, le processus de sélection sera ignorable.

4. *On observe le salaire, le nombre d'heures et la décision de participation*

$$\begin{cases} \begin{cases} w_i = x_i b + u_i \\ h_i = h_i^* = \gamma x_i b - x_{ri} b_r + \gamma u_i - u_{ri} & si\ h_i^* > 0 \\ p_i = 1 & si\ h_i^* \leq 0 \end{cases} \\ p_i = 0 \end{cases}$$

Ce modèle est dit modèle Tobit de Type III. Il permet sous certaines conditions d'estimer le paramètre d'élasticité de l'offre de travail aux heures.

L'estimation de ce type de modèles est en général complexe lorsque l'on ne spécifie pas la loi des résidus. On va examiner ici la situation dans laquelle la loi jointe des deux résidus u_{wi} de l'équation de salaire et u_{hi} de l'équation d'heure, conditionnellement aux variables explicatives, est une loi normale bivariée :

$$\begin{pmatrix} u_{wi} \\ u_{hi} \end{pmatrix} \rightsquigarrow N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_w^2 & \rho\sigma_w\sigma_h \\ \rho\sigma_w\sigma_h & \sigma_h^2 \end{pmatrix} \right]$$

Une caractérisitique importante de cette modélisation est de laisser possible une corrélation entre les deux équations de salaire et de participation. C'est justement dans le cas où il y a corrélation que le processus de sélection n'est pas ignorable dans le cas du modèle de type II.

Definition 1. *On appelle Modèle Tobit de type I, ou modèle Tobit simple le modèle dans lequel une variable d'intérêt modélisée comme*

$$y_i^* = x_i b + u_i$$

avec $u_i \rightsquigarrow \mathcal{N}(0, \sigma_u^2)$, est observée sous la condition, elle même observée,

$$y_i^* > 0$$

C'est à dire, on observe :

$$\begin{cases} \begin{cases} y_i = y_i^* = x_i b + u_i & si\ y_i^* > 0 \\ I_i = 1 & \text{sin}\, on \end{cases} \\ I_i = 0 \end{cases}$$

2. *On appelle Modèle Tobit de type II, le modèle dans lequel une variable d'intérêt, modélisée comme*

$$y_i^* = x_i b + u_i$$

est observée sous la condition elle même observée

$$I_i^* = z_i c + v_i > 0$$

avec (u_i, v_i) *distribués suivant une loi normale de moyennes nulle et de variance* σ_u^2 *et* σ_v^2 *et de corrélation* ρ. *On observe donc*

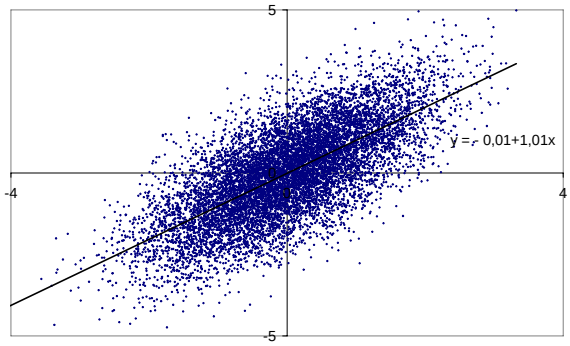
$$\begin{cases} y_i = y_i^* = x_i b + u_i & si\ I_i^* > 0 \\ I_i = 1 & \\ I_i = 0 & \text{sin}\, on \end{cases}$$

Pour mesurer l'importance potentielle des biais auquel peut conduire une information incomplète, on considère la situation dans laquelle il y a deux variables aléatoires

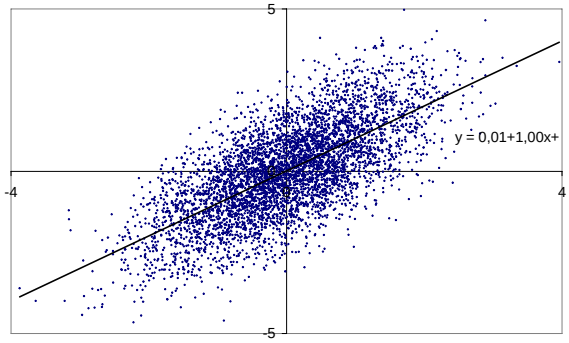
$$\begin{cases} y_1^* = x + u_1 \\ y_2^* = x + u_2 \end{cases}$$

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 reportés dans le tableau 12.2 montrent les nuages de points observés.

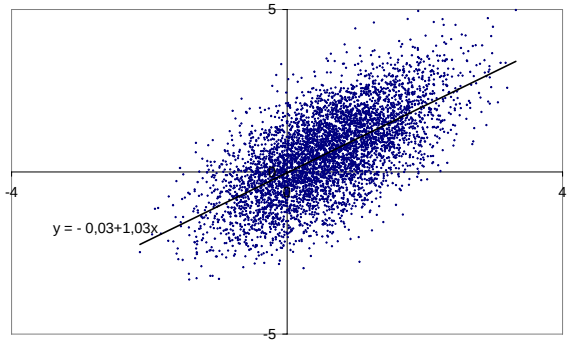
On voit que les nuages de points dans les échantillons non tronqués se ressemblent beaucoup, que la corrélation soit nulle ou de 0.9. Les droites de régressions linéaires donnent toutes deux des coefficients proches des vraies valeurs : 1 pour la variable x et 0 pour la constante. On voit aussi que la 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 sélection des variables explicatives n'affecte pas la propriété 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 là encore très proches des vraies valeurs. En revanche les changements pour le cas $\rho = 0.9$ en présence de troncature sont très 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 retenu que les observations avec une valeur importante de u_2 et donc de u_1 puisque ces variables sont fortement corrélées. On en déduit que à x donné, on a retenu des observations pour lesquelles u_1 est suffisamment 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.



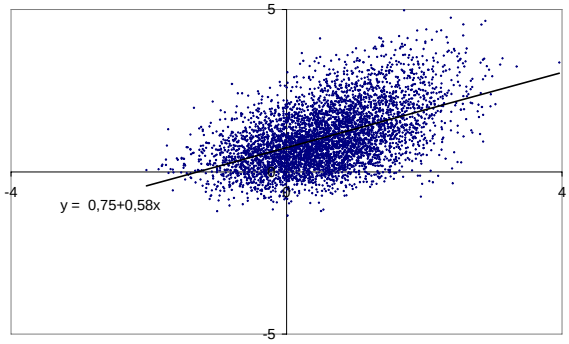
Complet $\rho = 0$



Complet $\rho = 0.9$



Tronqué $\rho = 0$



Tronqué $\rho = 0.9$

TAB. 12.2 – Nuages de points et troncatures : différentes configurations

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 très 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.

12.5.2 Rappels sur les lois normales conditionnelles.

Quelques rappels sur les lois normales sont nécessaires pour étudier le modèle de sélectivité.

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 multivariée de moyenne μ et de variance Σ : $y \sim N(\mu, \Sigma)$ 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

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \rightsquigarrow N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right]$$

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

Definition *On appelle inverse du ratio de Mills la quantité*

$$M(c) = \frac{\varphi}{\Phi}(c)$$

Ce ratio est central dans l'analyse des biais de sélectivité. On a vu précédemment en étudiant le modèle probit que ce ratio est une fonction décroissante de c.

Proposition *Soit $u \sim N(0,1)$, et c un scalaire. On s'intéresse aux moments de la loi normale tronquée $E(u|u > c)$ et $E(u|u < c)$, ainsi que $V(u|u > c)$ et $V(u|u < c)$. On a*

$$\begin{aligned} E(u|u &> c) = M(-c) \\ E(u|u &< c) = -M(c) \end{aligned}$$

et

$$\begin{aligned} V(u|u &> c) = 1 + cM(-c) - M(-c)^2 < 1 \\ V(u|u &< c) = 1 - cM(c) - M(c)^2 < 1 \end{aligned}$$

Démonstration *u a pour densité $\varphi(u)$. Compte tenu de $\varphi'(u) = -u\varphi(u)$, on a :*

$$E(u|u > c) = \frac{\int_c^\infty u\varphi(u)du}{1-\Phi(c)} = \frac{[-\varphi(u)]_c^\infty}{1-\Phi(c)} = \frac{\varphi(c)}{1-\Phi(c)} = \frac{\varphi(-c)}{\Phi(-c)} = M(-c)$$

de même

$$E(u|u < c) = -E(-u|-u > -c) = -M(c)$$

Pour les moments d'ordre 2 on a :

$$E\left(u^2|u > c\right) = \frac{\int_c^\infty u^2\varphi(u)du}{1-\Phi(c)} = 1 + cM(-c)$$

où on intègre par partie $\int_c^\infty u^2\varphi(u)du = [-u\varphi(u)]_c^\infty + \int_c^\infty \varphi(u)du = c\varphi(c) + 1 - \Phi(c)$. 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

$$\begin{aligned} E(u^2|u &< c) = E((-u)^2 \mid -u > -c) = 1 - cM(c) \\ V(u|u &< c) = 1 - cM(c) - M(c)^2 \end{aligned}$$

Le lemme que l'on avait pour une loi normale $z + \frac{\varphi}{\Phi}(z) > 0$ *et aussi* $-z + \frac{\varphi}{1-\Phi}(z) > 0$ *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* $V(u|u \lesseqgtr c) < 1$.

Lemme *Quelque soit* z, *on a*

$$z + \frac{\varphi}{\Phi}(z) > 0$$

et

$$-z + \frac{\varphi}{1-\Phi}(z) > 0$$

Démonstration *Compte tenu de* $\varphi'(z) = -z\varphi(z)$ *on déduit de* φ/Φ *décroissant* $\varphi'(z)/\Phi - \varphi^2/\Phi^2 < 0$, *soit* $-z\varphi(z)/\Phi - \varphi^2/\Phi^2 < 0$. *En multipliant cette inégalité par* $-\frac{\varphi}{\Phi}(z)$, *on en déduit un résultat qui sera utile par la suite :* $z + \frac{\varphi}{\Phi}(z) > 0$. *En appliquant cette inégalité à* $-z$, *on en déduit aussi* $-z + \frac{\varphi}{1-\Phi}(z) > 0$.

Remarque *Dans le cas d'une variable non centrée réduite* $v \sim N(\mu, \sigma^2)$, *on peut déduire des résultats précédents les moments des lois tronquées en notant que* $(v-\mu)/\sigma$ *suit une loi* $N(0,1)$ *et que* $v \lesseqgtr c \Leftrightarrow u = (v-\mu)/\sigma \lesseqgtr \widetilde{c} = (c-\mu)/\sigma$. *on a donc*

$$\begin{aligned} E(v|v &> c) = E(\sigma u + \mu | u > \widetilde{c}) = \mu + \sigma M\left(-\frac{c-\mu}{\sigma}\right) \\ E(v|v &< c) = E(\sigma u + \mu | u < \widetilde{c}) = \mu - \sigma M\left(\frac{c-\mu}{\sigma}\right) \end{aligned}$$

et

$$V(v|v > c) = \sigma^2\left(1 + \frac{c-\mu}{\sigma}M\left(-\frac{c-\mu}{\sigma}\right) - M\left(-\frac{c-\mu}{\sigma}\right)^2\right)$$

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 $y_1 > 0$.

Proposition *On a*

$$\begin{aligned} E(y_2 | y_1 > 0) &= \mu_2 + \rho\sigma_2 M\left(\frac{\mu_1}{\sigma_1}\right) \\ E(y_2 | y_1 > 0) &= \mu_2 - \rho\sigma_2 M\left(-\frac{\mu_1}{\sigma_1}\right) \end{aligned}$$

et

$$\begin{aligned} V(y_2 | y_1 > 0) &= \sigma_2^2 - \rho^2\sigma_2^2\left(\frac{\mu_1}{\sigma_1} M\left(\frac{\mu_1}{\sigma_1}\right) + M\left(\frac{\mu_1}{\sigma_1}\right)^2\right) \\ V(y_2 | y_1 < 0) &= \sigma_2^2 - \rho^2\sigma_2^2\left(-\frac{\mu_1}{\sigma_1} M\left(-\frac{\mu_1}{\sigma_1}\right) + M\left(-\frac{\mu_1}{\sigma_1}\right)^2\right) \end{aligned}$$

Démonstration *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 $\sigma_2^2(1-\rho^2)$. 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 $V(y_2|y_1) = (1-\rho^2)\sigma_2^2$.

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* $-\rho$

12.6 Estimation du modèle Tobit

On considère à nouveau le modèle Tobit

$$\begin{aligned} y_i^* &= x_i b + u_i \\ I_i^* &= zc + v_i \end{aligned}$$

dans lequel la loi jointe des résidus conditionnellement aux variables explicatives est une loi normale bivariée

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \rightsquigarrow N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix}\right]$$

Les observations sont régies par :

$$\begin{cases} \begin{cases} y_i = y_i^* & \text{si } I_i^* > 0 \\ I_i = 1 & \text{si } I_i^* \leq 0 \end{cases} \\ I_i = 0 \end{cases}$$

12.6.1 Pourquoi ne pas estimer un modèle Tobit par les MCO ?

Si on se restreint aux observations pour lesquelles le salaire est renseigné, on a

$$E(y_i|x_i, z_i, I_i = 1) = E(y_i^*|x_i, z_i, I_i^* > 0)$$

En appliquant les résultats précédents à $y_2 = y^*$, et $y_1 = I^*$ on a directement :

$$E\left(y_i^* \left| x_i, z_i, I_i^* > 0\right.\right) = x_i b + \rho\sigma_u M\left(\frac{z_i c}{\sigma_v}\right)$$

On voit donc que dès lors que la corrélation entre les éléments inobservés de l'équation de salaire et de l'équation de participation sont corrélés, c'est à dire dès que $\rho \neq 0$, ne pas prendre en compte la sélectivité revient à oublier une variable dans la régression : $M\left(\frac{z_i c}{\sigma_v}\right)$. Cet oubli est donc susceptible de conduire à une estimation biaisée des paramètres dès lors que les variables $M\left(\frac{z_i c}{\sigma_v}\right)$ et x_i sont corrélées.

Si on considère à titre illustratif que l'équation de sélection s'écrit $y_i^* > \overline{y}$, on a $\rho = 1$ et $\frac{z_i c}{\sigma_v} = \frac{x_i b - \overline{y}}{\sigma_u}$. L'équation précédente s'écrit alors

$$E\left(y_i^* \left| x_i, z_i, I_i^* > 0\right.\right) = x_i b + \sigma_u M\left(\frac{x_i b - \overline{y}}{\sigma_u}\right)$$

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 dépend de ρ et de la corrélation entre le ratio de Mills et $M\left(\frac{z_i c}{\sigma_v}\right)$ les variables explicative entrant dans la modélisation de y_i^*.

Si on introduit également les observations pour lesquelles $y_i = 0$, on a

$$\begin{aligned} E\left(y_i \left| x_i, z_i\right.\right) &= E\left(y_i \left| x_i, z_i, I_i = 1\right.\right) P\left(I_i = 1 \left| x_i, z_i\right.\right) + \\ &\quad E\left(y_i \left| x_i, z_i, I_i = 0\right.\right) P\left(I_i = 0 \left| x_i, z_i\right.\right) \\ &= E\left(w_i \left| x_i, z_i, I_i = 1\right.\right) P\left(I_i = 1 \left| x_i, z_i\right.\right) \\ &= \left(x_i b\right) \Phi\left(\frac{z_i c}{\sigma_v}\right) + \rho\sigma_u \varphi\left(\frac{z_i c}{\sigma_v}\right) \end{aligned}$$

et on voit que la forme linéaire n'est pas non plus adaptée.

12.6.2 Estimation par le maximum de vraisemblance

Comme on a spécifié la loi des perturbations, on a spécifié la loi des observations. L'estimateur du maximum de vraisemblance est donc le plus efficace. Les estimations vont être basées sur la densité des observations. celle-ci se calcule de la façon suivante : on écrit la probabilité d'observer chaque réalisation du couple (y_i, I_i).

– Pour $I_i = 0$ on n'observe pas y_i la seule probabilité est $P\left(I_i^* < 0\right)$, c'est à dire $P\left(z_i c + v_i < 0\right) = \Phi\left(-\frac{zc}{\sigma_v}\right) = 1 - \Phi\left(\frac{zc}{\sigma_v}\right)$

Pour $I_i = 1$ on observe $y_i = y_i^*$ et $I_i^* > 0$. La densité correspondante est

$$f\left(y_i^* = w_i, i_i = 1\right) = \int_{I_i^* > 0} f\left(y_i, I_i^*\right) dI_i^* = f\left(y_i\right) \int_{I_i^* > 0} f\left(I_i^* \left| y_i\right.\right) dI_i^*$$

et la loi de I_i^* conditionnelle à $y_i^* = y_i$ est pas définition une loi normale de moyenne $\widetilde{\mu}_I(y_i) = \mu_I + \rho\sigma_v \frac{y_i - \mu_y}{\sigma_u}$ et de variance $\widetilde{\sigma}_v^2 = \sigma_v^2(1-\rho^2)$ la probabilité pour qu'une telle variable aléatoire soit positive est $\Phi\left(\frac{\widetilde{\mu}_I(y_i)}{\widetilde{\sigma}_v}\right) = \Phi\left(\frac{\mu_I + \rho\sigma_v \frac{y_i-\mu_y}{\sigma_u}}{\sigma_v\sqrt{(1-\rho^2)}}\right)$. Finalement, la densité des observations est

$$\begin{aligned} L &= \prod_{I_i=0}\left[1-\Phi\left(\frac{zc}{\sigma_v}\right)\right] \times \prod_{I_i=1}\frac{1}{\sigma_u}\varphi\left(\frac{y_i - x_i b}{\sigma_u}\right)\Phi\left(\frac{z_i c + \rho\sigma_v \frac{y_i - x_i b}{\sigma_u}}{\sigma_v\sqrt{(1-\rho^2)}}\right) \\ &= \prod_i \left[1-\Phi\left(\frac{zc}{\sigma_v}\right)\right]^{1-I_i} \times \left[\frac{1}{\sigma_u}\varphi\left(\frac{y_i - x_i b}{\sigma_u}\right)\Phi\left(\frac{z_i c + \rho\sigma_v \frac{y_i - x_i b}{\sigma_u}}{\sigma_v\sqrt{(1-\rho^2)}}\right)^{I_i}\right] \end{aligned}$$

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 $\widetilde{c} = \frac{c}{\sigma_u}$ est identifiable. Compte tenu de cette redéfinition des paramètres du modèle, la vraisemblance s'écrit :

$$L = \prod_i \left[1-\Phi\left(z_i\widetilde{c}\right)\right]^{1-I_i} \times \left[\frac{1}{\sigma_u}\varphi\left(\frac{y_i - x_i b}{\sigma_u}\right)\Phi\left(\frac{z_i\widetilde{c} + \rho\frac{y_i - x_i b}{\sigma_u}}{\sqrt{(1-\rho^2)}}\right)^{I_i}\right]$$

Remarque *1. Dans le cas où $\rho = 0$ on voit que la vraisemblance est séparable entre une contribution correspondant à l'observation de $I_i = 0/1$ et une contribution associée aux observations de w_i :*

$$L = \left(\prod_i \left[1-\Phi\left(z_i\widetilde{c}\right)\right]^{1-I_i}\Phi\left(z_i\widetilde{c}\right)^{I_i}\right) \times \left(\prod_i\left[\frac{1}{\sigma_u}\varphi\left(\frac{y_i - x_i b}{\sigma_u}\right)\right]^{I_i}\right)$$

On retrouve donc le fait que dans le cas $\rho = 0$ on peut ignorer la sélection des observations. On voit aussi que dans le cas général où $\rho \neq 0$ la sélectivité importe.

2. *La fonction de vraisemblance n'est pas globalement concave en $(\rho, \sigma_u, b, \widetilde{c})$.Elle est concave globalement en $\theta = (\sigma_u, b, \widetilde{c})$ pour ρ fixé.*
3. *Une solution consiste à fixer la valeur de ρ et estimer les paramètre correspondant $\widehat{\theta}(\rho)$ et à balayer sur les valeur possible de ρ.*

12.6.3 Estimation en deux étapes par la méthode d'Heckman

Il existe une méthode d'estimation très simple et très largement utilisée dans le cas où les perturbations sont normales. Elle ouvre aussi la voie à des spécifications plus générales dans lesquelles on laisse non spécifiées la loi des perturbations. Cette méthode est basée sur l'équation précédente

$$E\left(y_i \left| x_i, z_i, I_i = 1\right.\right) = x_i b + \rho\sigma_u M\left(z_i\widetilde{c}\right) = x_i b + \rho\sigma_u M_i\left(\widetilde{c}\right)$$

Le principe de la méthode d'Heckman consiste à estimer d'abord le modèle Probit associé à I_i. De l'estimation de $\widetilde{c} = c/\sigma_v$ on tire un estimateur $M_i\left(\widehat{\widetilde{c}}\right) = M\left(z_i\widehat{\widetilde{c}}\right)$. On procède ensuite à la régression augmentée sur les seules observations pour lesquelles les données sont disponibles :

$$y_i = x_i b + \rho\sigma_u M_i\left(\widehat{\widetilde{c}}\right) + \varpi_i$$

Ces estimateurs sont asymptotiquement sans biais, mais ils ne sont pas asymptotiquement efficaces. Par exemple, cette méthode permet d'estimer seulement le produit $\rho\sigma_u$, alors que la méthode du maximum de vraisemblance permet d'estimer ρ et σ_u séparément.

Remarque *Le calcul des écarts-type est un peu compliqué. Il fait intervenir deux aspects. D'une part le modèle est hétéroscédastique. En effet, compte tenu des résultats obtenus précédemment pour* $V(y_2 \,|y_1 > 0)$, *on a :*

$$\begin{aligned} V(y_i \,|x_i, z_i, I_i = 1) &= V(y_i^* \,|x_i, z_i, I_i^* > 0) \\ &= \sigma_u^2 - \rho^2\sigma_u^2\left(z_i\widetilde{c}M_i(\widetilde{c}) + M_i(\widetilde{c})^2\right) \end{aligned}$$

Cette formule montre bien la présence d'hétéroscédasticité. Elle donne aussi une voie pour estimer le modèle de façon plus efficace en utilisant l'estimateur des mCQG. Néanmoins ce n'est pas le seul problème, en effet la variable additionnelle introduite dans la régression fait intervenir le paramètre $\widetilde{c}$ *qui n'est pas connu et est remplacé par une estimation. L'introduction de ce paramètre estimé est aussi une source de complication dans le calcul des écarts-type. Plus précisément, le paramètre est lui même issu d'une estimation (par le MV) que l'on peut résumer par l'annulation de la contrepartie empirique de conditions d'orthogonalité*

$$E\left(h_{\widetilde{c}}(I_i, z_i, \widetilde{c})\right) = 0$$

L'estimation du modèle par les mco conduit quant à elle à l'annulation de la contrepartie empirique de

$$\begin{aligned} & E\left(\begin{pmatrix} x_i' \\ M_i(\widetilde{c}) \end{pmatrix}\left[y_i - x_i b - \rho\sigma_u M_i(\widetilde{c})\right] 1_{I_i=1}\right) \\ = \;& E\left(h_{b,\rho\sigma_u}(I_i, y_i, x_i, b, \rho\sigma_u)\right) = 0 \end{aligned}$$

Le calcul des écarts-type 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

$$E\begin{pmatrix} h_{\widetilde{c}}(I_i, z_i, \widetilde{c}) \\ h_{b,\rho\sigma_u}(I_i, y_i, x_i, b, \rho\sigma_u) \end{pmatrix} = 0$$

On utilise parfois l'estimateur de Heckman comme une première valeur pour le calcul de l'estimateur du maximum de vraisemblance. On utilise l'estimateur du modèle Probit, l'estimateur du modèle de Heckman et l'expression de la variance des résidus qui permet d'obtenir une estimation convergente de ρ *et* σ_w.

12.6.4 Des extensions paramétriques simples

Le cas normal conduit à des spécifications particulièrement simple. La loi normale peut néanmoins paraître trop restrictive et on peut vouloir spécifier encore la loi des résidus mais dans des ensembles de lois plus générales.

Loi quelconque donnée pour le résidu de l'équation de sélection.

Tant que la loi du terme de l'équation de sélection a une fonction de répartition F strictement croissante, on peut reformuler le modèle de telle sorte qu'il entre dans le cadre précédent. Cette reformulation repose sur la propriété suivante :

Proposition *Si une variable aléatoire à une fonction de répartition F strictement croissante, alors la variable aléatoire $\widetilde{v} = F(v)$ suit une loi uniforme sur $[0,1]$.*

Démonstration *En effet, comme F est à valeurs dans $[0,1]$ le support de $\widetilde{v}$ est bien $[0,1]$. De plus on a*

$$P(\widetilde{v} \leq t) = P(F(v) \leq t) = P\left(v \leq F^{-1}(t)\right) = F \circ F^{-1}(t) = t$$

On en déduit alors la proposition suivante concernant le modèle de sélection : En appliquant ce résultat à la transformation : $\widetilde{v} = \Phi^{-1} \circ F(v)$, on en déduit que $\widetilde{v}$ suit une loi normale. Le modèle de sélection $I = 1 \Longleftrightarrow I^* = zc + v \geq 0$ est donc équivalent à $I = 1 \Longleftrightarrow \widetilde{v} = \Phi^{-1} \circ F(v) \geq \Phi^{-1} \circ F(-zc)$ soit encore à $-\Phi^{-1} \circ F(-zc) + \widetilde{v} \geq 0$, avec dans ce cas $\widetilde{v}$ normal. On peut donc généraliser les résultats précédents en substituant $-\Phi^{-1} \circ F(-zc)$ à zc. On parvient alors au résultat que

$$E(y\,|I = 1, x, z) = xb + \rho\sigma_u \frac{\phi}{\Phi}\left(-\Phi^{-1} \circ F(-zc)\right)$$

Compte tenu du fait que

$$P(z) = P(zc + v \geq 0) = P(v \geq -zc) = 1 - F(-zc)$$

on a

$$E(y\,|I = 1, x, z) = xb + \rho\sigma_u \frac{\phi}{\Phi}\left(-\Phi^{-1}(1 - P(z))\right)$$

En utilisant le fait que $\Phi(-x) = 1 - \Phi(x)$, soit $\Phi^{-1}(P) = -\Phi^{-1}(1 - P)$, on a :

$$E(y\,|I = 1, x, z) = xb + \rho\sigma_u \frac{\phi \circ \Phi^{-1} P(z)}{P(z)}$$

Des lois plus générales que la loi normale

On peut considérer le modèle de sélection précédent en faisant l'hypothèse que les éléments inobservés ont pour loi jointe une loi de Student de degrés η et non pas une loi normale.

La densité de la loi jointe des éléments inobservés s'écrit alors :

$$h(u,v) = \frac{1}{2\pi\left(1-\rho^2\right)^{1/2}}\frac{\eta}{\eta-2}\left[1+\frac{1}{\left(\eta-2\right)\left(1-\rho^2\right)}\left(u^2-2\rho uv+v^2\right)\right]^{-(1/2)(\eta+2)}$$

On peut montrer la propriété suivante sur la loi jointe de u et v :

$$E(u\,|v) = \rho v$$

La loi de u, $g_\eta(u)$ a pour expression :

$$g_\eta(u) = \sqrt{\frac{\Gamma\left((\eta+1)/2\right)}{\pi\eta\Gamma\left(\eta/2\right)}}\left(1+t^2\right)^{-(\eta+1)/2}$$

On note $G_\eta(u)$ sa fonction de répartition.

On peut montrer que l'expression de l'espérance de la loi de Student de degrés η tronquée est :

$$E(v\,|v<t) = -\frac{\eta+t^2}{\eta-1}\frac{g_\eta}{G_\eta}(t)$$

D'où

$$\begin{aligned} E(v\,|v>-t) &= -E(v\,|v<-t)\frac{G_\eta(-t)}{(1-G_\eta(-t))} = \frac{G_\eta(-t)}{(1-G_\eta(-t))}\frac{\eta+t^2}{\eta-1}\frac{g_\eta}{G_\eta}(-t) \\ &= \frac{\eta+t^2}{\eta-1}\frac{g_\eta}{1-G_\eta}(-t) = \frac{\eta+t^2}{\eta-1}\frac{g_\eta}{G_\eta}(t) \end{aligned}$$

Ceci permet de généraliser les résultats obtenus précédemment pour le modèle de sélection

$$\begin{aligned} E(y\,|I=1,x,z) &= xb+E(u\,|d=1,x,z) \\ &= xb+E(u\,|zc+v>0,x,z) \\ &= xb+E\left(E(u\,|v,x,z)\,|zc+v>0,x,z\right) \\ &= xb+\rho\sigma E(v\,|v>-zc) \\ &= xb+\rho\sigma\frac{\eta+zc^2}{\eta-1}\frac{g_\eta}{G_\eta}(zc) \end{aligned}$$

On peut obtenir une généralisation supplémentaire en combinant les deux approches et en considérant que l'équation de sélection à un résidu d'une loi quelconque connue. Par le même genre d'argument que dans la première situation envisagée, on a

$$E\left(y\left|I=1,x,z\right.\right)=xb+\rho\sigma\frac{\eta+G_{\eta}^{-1}\left(P\left(z\right)\right)^{2}}{\eta-1}\frac{g_{\eta}\circ G_{\eta}^{-1}\left(P\left(z\right)\right)}{P\left(z\right)}$$

12.6.5 Le modèle de sélection semi paramétrique.

On reprend le modèle de sélectivité sur inobservables :

$$y=xb+u$$

avec la modélisation de l'affectation au traitement :

$$\begin{aligned}T^{*} &= zc+v\\ T &= 1\Longleftrightarrow T^{*}\geq 0\end{aligned}$$

on suppose comme précédemment l'indépendance entre les variables de conditionnement et les éléments inobservés.

$$\left(u,v\right)\perp\left(x,z\right)$$

mais on ne fait plus d'hypothèse sur la loi jointe des perturbations. On montre que l'on obtient une relation pour l'espérance conditionnelle qui s'apparente à celles obtenues dans les cas précédents :

Proposition *Dans le cas du modèle de sélectivité sur inobservables, si les fonctions de répartition de v est strictement croissante, il existe une fonction $K\left(P\left(zc\right)\right)$ telle que*

$$E\left(y\left|I=1,x,z\right.\right)=xb+K\left(P\left(zc\right)\right)$$

où

$$P\left(zc\right)=P\left(T=1\left|r,z\right.\right)$$

Démonstration *On montre d'abord que $P\left(I=1\left|r,z\right.\right)=P\left(zc\right)$. On a*

$$P\left(I=1\left|r,z\right.\right)=E\left(1\left(zc+v>0\right)\left|r,z\right.\right)=\int_{v>-zc}f\left(v\left|r,z\right.\right)=\int_{v>-zc}f\left(v\right)=1-F\left(zc\right)=P\left(zc\right)$$

On en déduit en outre que $zc=H_{1}\left(P\left(zc\right)\right)$, puisque F est strictement croissante. On écrit ensuite l'espérance de la variable d'intérêt

$$E\left(y\left|I=1,x,z\right.\right)=xb+E\left(u\left|I=1,x,z\right.\right)$$

et on montre que $E(u|I=1,x,z)$ *est une fonction de* $P(zc)$

$$\begin{aligned} E(u|I=1,x,z) &= E(uI|x,z)\,P(I=1|r,z)^{-1} = \int 1(zc+v \lesseqgtr 0)\,uf(u,v)\,dudvP(I=1|zc)^{-1} \\ &= H_2(zc) = K(P(zc)) \end{aligned}$$

Remarque *On peut voir à partir des expressions précédentes un point très important. Dans le cas de la normalité, on a une relation non linéaire déterminée entre l'espérance de la variable à laquelle on s'intéresse et la probabilité de sélection. Cette non linéarité permet l'obtention d'estimation même dans le cas où les variables entrant dans l'équation de sélection et l'équation d'intérêt principal sont identiques. Dans le cas plus général, on voit néanmoins que ce n'est plus le cas. En effet quelque soit la fonction de probabilité retenue P, si la fonction K est quelconque, et que* x_p *est identique à* x_w*, on ne pourra dissocier l'effet des variables intervenant au travers de la sélectivité de leur effet intervenant directement : le modèle n'est pas identifié. Ce n'est que lorsque l'on introduit dans l'équation de sélectivité une variable intervenant dans la sélectivité mais pas dans l'équation principale que l'on peut identifier le modèle. Le raisonnement est ici très proche de celui fait dans le cas des variables instrumentales : il faut postuler une relation d'exclusion. Cette nécessité est un peu masquée dans le cas de la normalité par la non linéarité du modèle, mais elle n'en est pas moins essentielle.*

Ce type de modèle peut être estimé sans faire d'hypothèse sur la forme de la fonction K. On considère l'équation :

$$E(y|I=1,x,z) = xb + K(P(zc))$$

Une première façon d'estimer le modèle consiste à utiliser des séries. L'idée est très simple elle consiste à introduire différentes puissance du score : $P(zc), P(zc)^2, \ldots$. Les propriétés asymptotiques de ce type d'estimateur ont été étudiée par Andrews (1991).

$$E(y|I=1,x,z) = xb + \alpha_1 P(zc) + \cdots + \alpha_{d_N} P(zc)^{d_N}$$

Cette méthode est très simple à mettre en oeuvre, et de ce fait très utile. Ses propriétés asymptotiques ont été clairement établies, par Newey (1999) qui montre en particulier que les paramètres d'intérêt de la partie linéaire du modèle sont convergent en $\sqrt{N}$. Le problème de ce type de méthode réside dans le choix du degré du polynôme retenu.

Une méthode d'estimation alternative est fournie par la méthode d'estimation de (Robinson 1988) c'est une sorte de super méthode de Frish-Waugh. L'Idée de la méthode de Robinson est de projeter cette équation sur l'ensemble des fonctions de $P(zc)$

$$\begin{aligned} E(y|I=1,P(zc)) &= E(E(y|I=1,x,z)|I=1,P(zc)) \\ &= E(x|I=1,P(zc))\,b + K(P(zc)) \end{aligned}$$

En prenant la différence avec l'équation précédente on peut éliminer la fonction $K\left(P\left(zc\right)\right)$. On a alors :

$$E\left(y - E\left(y\left|I=1,P\left(zc\right)\right.\right)\left|I=1,x,z\right.\right) = \left(x - E\left(x\left|I=1,P\left(zc\right)\right.\right)\right)b$$

En notant $\varepsilon_y^P = y - E\left(y\left|I=1,P\left(zc\right)\right.\right)$ et $\varepsilon_x^P = x - E\left(x\left|T=1,P\left(zc\right)\right.\right)$ les résidus des régressions non paramétriques de y et des variables explicatives r sur le score $P\left(zc\right)$, on a clairement

$$E\left(\varepsilon_y^P\left|\varepsilon_r^P\right.\right) = \varepsilon_r^P b$$

On peut estimer le paramètre b en régressant ε_y^P sur ε_r^P. Dans ce cas, on peut montrer que l'estimateur de b obtenu est convergent en $\sqrt{N}$ bien qu'il incorpore un intermédiaire de calcul non paramétrique. Toutefois sa variance est difficile à calculer et on est amené à utiliser des méthodes de bootstrap très intensives en calculs, notamment pour ce type d'estimateur par noyaux.

Remarque *Cette méthode permet d'estimer le paramètre b. Néanmoins ceci n'est pas vrai pour tous les paramètres : la constante du modèle n'est pas identifiée. Ceci se voit très bien puisque la fonction K est estimée en toute généralité, donc à une constante prés. Ceci n'est en général pas grave car on n'accorde que peu d'intérêt à la constante, sauf dans certains cas précis qui peuvent être très importants. C'est en particulier le cas de l'évaluation des politique publiques que l'on aborde dans le chapitre suivant. On reviendra alors sur cette question délicate.*

12.6.6 Illustration : le modèle d'offre de travail d'Heckman

Pour illustrer les résultats du cadre précédent on estime le modèle d'offre de travail présenté dans l'exemple de la page 217. Il s'agit d'un modèle Tobit dit de TypeIII, dans la terminologie de Amemiya. La forme réduite de ce modèle s'écrit :

$$\begin{aligned} w_i^* &= x_i b + u_i \\ h_i^* &= \gamma x_i b - x_{ri} b_r + \gamma u_i - u_{ri} = z_i c + v_i \end{aligned}$$

En appliquant le formalisme de la méthode d'Heckman, on voit que l'on a :

$$\begin{aligned} E\left(w_i\left|z_i,h_i^*>0\right.\right) &= x_i b + \left(u_i\left|z_i,h_i^*>0\right.\right) \\ &= x_i b + \rho\sigma\frac{\phi}{\Phi}\left(z_i c\right) \\ E\left(h_i\left|z_i,h_i^*>0\right.\right) &= \gamma x_i b - x_{ri} b_r + \rho_h\sigma_h\frac{\phi}{\Phi}\left(z_i c\right) \end{aligned}$$

On voit clairement que les paramètres b, γ et b_r sont identifiés. En effet, le modèle Probit identifie le paramètre c, la régression de salaire identifie b et $\rho\sigma$, la régression d'heure identifie γb, b_r et $\rho_h\sigma_h$. On voit que l'on peut en déduire une estimation de γ dès lors

qu'il y a une variable entrant dans la liste des variables affectant le salaire de marché mais pas le salaire de réserve. La variable retenue ici assurant cette identification est la variable de scolarité. En effet on fait intervenir la variable d'âge dans le salaire de réserve et dans le salaire de marché. Néanmoins l'identification du paramètre γ est liée ici à la forme fonctionnelle, c'est à dire à la forme du ratio de Mills. On voit que si on avait retenu une autre loi et que pour cette loi le terme analogue au ratio de Mills avait été linéaire le modèle ne serait pas identifié puisqu'il impose que $z_i c$ soit proportionnel à $\gamma x_i b - x_{ri} b_r$. Même si le modèle impose des restrictions qui peuvent être testées comme le fait que les paramètres de la partie $\gamma x_i b - x_{ri} b_r$ sont bien proportionnels à ceux de la partie $z_i c$, on ne peut en déduire d'estimateur de ces paramètres, sauf à faire une hypothèse comme celle faite ici que les variables inobservées sont distribuées suivant une loi normale. On peut noter que le modèle de salaire de marché peut lui aussi faire intervenir les heures. Dans ce cas l'identification porte comme pour le modèle d'heures offertes sur la forme fonctionnelle. Enfin, on voit aussi que l'estimation s'apparente ici à une estimation par la méthode des moments généralisée. En effet, on peut réécrire l'équation d'offre de travail par exemple sous la forme

$$E\left(h_i^* - \gamma w_i^* + x_{ri} b_r \left| z_i, h_i^* \geq 0\right.\right) = E\left(-u_{ri} \left| z_i, h_i^* \geq 0\right.\right) = \widetilde{\rho}_h \widetilde{\sigma}_h \frac{\phi}{\Phi}\left(z_i c\right)$$

Soit

$$E\left(h_i^* - \gamma w_i^* + x_{ri} b_r - \widetilde{\rho}_h \widetilde{\sigma}_h \frac{\phi}{\Phi}\left(z_i c\right) \left| z_i, h_i^* \geq 0\right.\right) = 0$$

avec $\widetilde{\rho}_h \widetilde{\sigma}_h = cov(-u_{ri}, \gamma u_i - u_{ri}) / \sigma\left(\gamma u_i - u_{ri}\right)$. Il en résulte que les paramètres peuvent être estimés en utilisant comme conditions d'orthogonalité

$$E\left(\left(h_i^* - \gamma w_i^* + x_{ri} b_r - \widetilde{\rho}_h \widetilde{\sigma}_h \frac{\phi}{\Phi}\left(z_i c\right)\right) \begin{pmatrix} z_i \\ \frac{\phi}{\Phi}\left(z_i c\right) \end{pmatrix} \middle| h_i^* \geq 0\right) = 0$$

De même, pour l'équation de salaire, on a

$$E\left(\left(w h_i^* - \lambda h_i^* - x_i b - \rho\sigma \frac{\phi}{\Phi}\left(z_i c\right)\right) \begin{pmatrix} z_i \\ \frac{\phi}{\Phi}\left(z_i c\right) \end{pmatrix} \middle| h_i^* \geq 0\right) = 0$$

qui peut être utilisée avec λ contraint à 1 (l'identification des autres paramètres est alors garanti quelle que soit la forme fonctionnelle retenue) ou librement estimé (l'identification des paramètres repose alors sur l'hypothèse de normalité).

Remarque *Pour la détermination des écarts-type, il faut tenir compte de deux aspects importants. Le premier est que le modèle est hétéroscédastique. L'utilisation de la méthode des moments généralisée permet de traiter ce problème. Le deuxième est que le ratio de Mills fait intervenir l'estimation de l'équation de participation. Il faut en théorie corriger les écarts-type pour cette estimation intermédiaire. Ceci peut être fait en considérant l'estimation comme un problème d'estimation par la méthode des moments généralisée. On*

adjoint à l'ensemble de condition d'orthogonalité précédent les conditions d'orthogonalité correspondant à l'estimation préliminaire, et qui sont les conditions du premier ordre du maximum de vraisemblance. Ici, compte tenu du fait que le modèle Probit est estimé sur 36249 femmes et que les estimations sont effectués dans le secteur du commerce sur seulement 3164 femmes, on néglige le problème.

On présente dans le tableau 12.3 les résultats obtenus pour l'estimation de l'équation de salaire. On voit que le ratio de Mills joue significativement et que son coefficient est négatif. Le signe est celui de la corrélation entre $\gamma u_i - u_{ri}$ et u_i. Si on écrit $u_{ri} = \eta u_i + \varepsilon_i$, avec u_i et ε_i non corrélé, on a $cov(\gamma u_i - u_{ri}, u_i) = (\gamma - \eta)\,\sigma_u^2$. Le signe négatif s'interprète donc comme le fait que les éléments inobservés dans l'équation de salaire et l'équation de salaire de réserve sont fortement corrélés. On voit qu'ignorer la sélectivité, oublier la variable de ratio de Mills, conduit à biaiser les coefficients. Ici il s'agit surtout de celui de la scolarité. Le coefficient est en effet de 0.03 avec prise en compte de la sélectivité au lieu de 0.04 lorsqu'on l'ignore. On voit que lorsque l'on introduit la variable d'heures comme régresseur l'erreur liée au fait d'oublier la variable de sélectivité est encore plus forte. En effet l'élasticité du salaire de marché (donc de la productivité) aux heures est élevée et significativement différente de 0 lorsque l'on ignore la sélectivité. Par contre lorsqu'on prend en compte la sélectivité, on voit que cette variable est deux fois plus faible et qu'elle n'est plus significativement différente de 0. Ceci est susceptible de remettre fortement en cause les résultats présentés dans le chapitre sur la méthode des moments généralisée. Toutefois, il ne faut pas oublier que lorsque l'on introduit la variable d'heure, l'identification des paramètres repose sur le choix de la normalité pour distribution jointe des résidus.

Le tableau 12.4 présente les résultats de l'équation d'offre de travail. On voit là aussi que la variable de sélectivité est significativement différente de zéro. Son signe est celui de $\widetilde{\rho}_h \widetilde{\sigma}_h = cov(-u_{ri}, \gamma u_i - u_{ri})$. Soit pour $u_{ri} = \eta u_i + \varepsilon_i$, celui de $\sigma_\varepsilon^2 + (\eta - \gamma)\,\eta \sigma_u^2$. Le signe obtenu est donc compatible avec le précédent. On voit que là aussi les changements sont importants lorsque l'on estime le modèle avec et sans prise en compte de la sélectivité. En effet sans prise en compte de la sélectivité, on a un coefficient faible de l'ordre de 0.10. Une baisse de la rémunération de 10% conduit à une baisse des heures offertes de 1%. Lorsque l'on prend en compte la sélectivité, on parvient à une valeur beaucoup plus élevée de 0.4 : une baisse de la rémunération de 10% conduit à une baisse des heures de 4%.

	Sans les heures			
	Avec Sélectivité		Sans Sélectivité	
	b	sb	b	sb
Constante	4.6368	(0.0768)	4.4496	(0.0555)
Age	0.0096	(0.0008)	0.0098	(0.0008)
Age²	-0.0004	(0.0001)	-0.0005	(0.0001)
Scolarité	0.0333	(0.0034)	0.0414	(0.0026)
Ratio de mills	-0.1662	(0.0456)	--	--
	Avec les heures			
Constante	3.7674	(0.8199)	2.6204	(0.5044)
Age	0.0094	(0.0008)	0.0094	(0.0008)
Age²	-0.0004	(0.0001)	-0.0005	(0.0001)
Scolarité	0.0346	(0.0035)	0.0369	(0.0029)
Ratio de mills	-0.0967	(0.0708)	--	--
h	0.2380	(0.2251)	0.5454	(0.1496)

TAB. 12.3 – Estimation de l'équation de salaire avec et sans prise en compte de la sélectivité, avec et sans prise en compte des heures

	Avec Sélectivité		Sans Sélectivité	
	b	sb	b	sb
Constante	-0.0805	(1.1674)	2.3980	(0.2713)
Age	-0.0051	(0.0015)	-0.0019	(0.0004)
Age²	-0.0002	(0.0001)	-0.0001	(0.0001)
Nenf	-0.0665	(0.0150)	-0.0349	(0.0054)
wa	0.0071	(0.0025)	0.0022	(0.0012)
single	0.0672	(0.0133)	0.0554	(0.0133)
Ratio de mills	0.3055	(0.1421)	--	--
w	0.4124	(0.1314)	0.1332	(0.0309)

TAB. 12.4 – Estimation de l'équation d'offre de travail avec et sans prise en compte de la sélectivité

12.7 Modèles de choix discrets : le Modèle Logit Multinomial

On s'intéresse dans cette dernière section à un modèle de choix entre différentes alternatives. Le choix d'un type de véhicule, d'un lieu de vacances, etc... Ce modèle, appelé modèle Logit Multinomial est très simple et très facile à estimer. Il est très largement employé. Il est en outre susceptible de généralisations importantes qui permettent notamment de prendre en compte l'existence de caractéristiques inobservées des individus opérant les choix. Le développement et l'estimation de ce type de modèle est aujourd'hui un thème de recherche très actif aux nombreuses applications.

Supposons qu'un individu i ait à choisir, parmi un ensemble de K modalités, une et une seule de ces modalités, notée k.

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 maximale.

$$y_i = Arg \max_k (U_{ik})$$

Proposition *Si les* $\{\varepsilon_{ik}\}_{k=1,...K}$ *sont des v.a. indépendantes et identiquement distribuées selon une loi des valeurs extrêmes de fonction de répartition.*

$$G(x) = \exp[-\exp(-x)],$$

de support $]-\infty, +\infty[$ *alors la probabilité de choisir la modalité* k *s'écrit :*

$$P[y_i = k] = \frac{\exp(\mu_{ik})}{\Sigma_{l=1}^{K} \exp(\mu_{il})} = \frac{\exp(x_i b_k)}{\Sigma_{l=1}^{K} \exp(x_i b_l)}$$

Ce modèle est appelé modèle logit multinomial.

Démonstration *Notons* g *la fonction de densité des* ε *:*

$$g(z) = G'(z) = \frac{d}{dz} \exp[-\exp(-z)] = \exp(-z) \exp(-\exp(-z)) = \exp(-z) G(z)$$

On peut remarquer en préliminaire la propriété suivante :

$$E \exp(-t \exp(-z)) = \frac{1}{1+t}$$

En effet :

$$E \exp(-t \exp(-z)) = \int_{-\infty}^{+\infty} \exp(-t \exp(-z)) \exp(-z) \exp(-\exp(-z))\, dz$$

en faisant le changement de variable $v = \exp(-z)$, *on a*

$$E\exp(-t\exp(-z)) = \int_0^{+\infty} \exp(-tv)\, v \exp(-v)\, v = \frac{1}{1+t}$$

On peut écrire par exemple la probabilité de choisir la première solution

$$\begin{aligned} P(y=1) &= E\left(\prod_{k=2}^{K} 1(U_k < U_1)\right) = E\left(E\left(\prod_{k=2}^{K} 1(U_k < U_1 \,|U_1)\right)\right) \\ &= E\left(\prod_{k=2}^{K} E(1(U_k < U_1 \,|U_1))\right) \end{aligned}$$

Puisque les valeurs des différentes options sont indépendantes les unes des autres. Comme $P(\mu_k + \varepsilon_k < \mu_1 + \varepsilon_1 \,|\varepsilon_1) = G(\mu_1 - \mu_k + \varepsilon_1) = \exp[-\exp(-\mu_1 + \mu_k - \varepsilon_1)]$, *on a*

$$\begin{aligned} P(y=1) &= E\left(\prod_{k=2}^{K} \exp[-\exp(-\mu_1 + \mu_k - \varepsilon_1)]\right) \\ &= E\left(\exp\left[-\sum_{k=2}^{K} \exp(-\mu_1 + \mu_k - \varepsilon_1)\right]\right) = E(\exp[-t\exp(-\varepsilon_1)]) \end{aligned}$$

avec $t = \sum_{k=2}^{K} \exp(-\mu_1 + \mu_k)$. *On en déduit que*

$$P(y=1) = \frac{1}{1+t} = \frac{1}{\sum_{k=1}^{K} \exp(-\mu_1 + \mu_k)}$$

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 modifiées 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'interprètent alors comme des écarts à la référence* 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.*

12.7.1 Estimation du modèle logit multinomial :

Proposition *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 *La vraisemblance s'écrit* $\log L = \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} \log P_{ik} = \log L = \sum_{i=1}^{n} \left(\sum_{k=2}^{K} y_{ik} x_{ki} b_k - \log \left(1 + \sum_{l=2}^{K} \exp(x_{li} b_l) \right) \right)$. *On calcule facilement la dérivée par rapport à* b_l :

$$\frac{\partial \log L}{\partial b_l} = \sum_{i=1}^{n} y_{il} x'_{li} - \frac{\exp(x_{li} b_l)}{\left(1 + \sum_{l=2}^{K} \exp(x_{li} b_l) \right)} x'_{li} = \sum_{i=1}^{n} \left(y_{il} - P_{li} \right) x'_{li}$$

On détermine ensuite la dérivée seconde

$$\frac{\partial^2 \log L}{\partial b_l \partial b'_m} = \sum_{i=1}^{n} \frac{\partial}{\partial b'_m} \left(y_{il} x'_{li} - \frac{\exp(x_{li} b_l)}{\left(1 + \sum_{l=2}^{K} \exp(x_{li} b_l) \right)} x'_{li} \right) = - \sum_{i=1}^{n} \frac{\partial}{\partial b'_m} \frac{\exp(x_{li} b_l)}{\left(1 + \sum_{l=2}^{K} \exp(x_{li} b_l) \right)} x'_{li}$$

Pour $m \neq l$, on a

$$\frac{\partial}{\partial b'_m} \frac{\exp(x_{li} b_l)}{\left(1 + \sum_{l=2}^{K} \exp(x_{li} b_l) \right)} x'_{li} = - \frac{\exp(x_{li} b_l) \exp(x_{mi} b_m)}{\left(1 + \sum_{l=2}^{K} \exp(x_{li} b_l) \right)^2} x'_{li} x_{mi} = - P_{mi} P_{li} x'_{li} x_{mi}$$

Pour $m = l$, on a

$$\begin{aligned}\frac{\partial}{\partial b_l'}\frac{\exp(x_{li}b_l)}{\left(1+\sum_{l=2}^{K}\exp(x_{li}b_l)\right)}x_{li}' &= \frac{\exp(x_{li}b_l)}{\left(1+\sum_{l=2}^{K}\exp(x_{li}b_l)\right)}x_{li}'x_{li} - \frac{\exp(x_{li}b_l)^2}{\left(1+\sum_{l=2}^{K}\exp(x_{li}b_l)\right)^2}x_{li}'x_{li} \\ &= \left(P_{li}-P_{li}^2\right)x_{li}'x_{li}\end{aligned}$$

Pour montrer la concavité de l'objectif, on calcule $\lambda' H\lambda$, pour un vecteur λ quelconque. La matrice H a pour dimension $\dim b_2 + \cdots + \dim b_K$. On peut donc écrire $\lambda' = (\lambda_2', \ldots, \lambda_K')$. Comme H est une matrice bloc dont les blocs sont de la forme : $H_{l,m} = \theta_{mli}x_{li}'x_{mi}$, avec $\theta_{mli} = P_{mi}P_{li}$ et $\theta_{mmi} = -P_{mi} + P_{mi}^2$, $\lambda' H\lambda = \sum_{l,m}\lambda_l' H_{l,m}\lambda_m = \sum_{l,m}\theta_{mli}\lambda_l' x_{li}' x_{mi}\lambda_m$. En définissant v_i le vecteur de dimension $K-1$ dont la mième composante est $x_{mi}\lambda_m$, on a $\lambda' H\lambda = \sum_{l,m}\theta_{mli}v_{mi}v_{li}$ et compte tenu de l'expression de θ_{mli}, on a $\sum_{l,m}\theta_{mli}v_{mi}v_{li} = \sum_m\left(-P_{mi}+P_{mi}^2\right)v_{mi}^2 + 2\sum_{m\neq l}P_{mi}P_{li}v_{mi}v_{li} = -\left(\sum_m P_{mi}v_{mi}^2 - \left(\sum_m P_{mi}v_{mi}\right)^2\right) \leq 0$ et égal à zero seulement si $v_i = 0$. On en déduit que $\lambda' H\lambda \leq 0$ et $\lambda' H\lambda = 0$ si et seulement si $v_i = 0 \forall i$, ce qui signifie que $\exists\lambda$ tel que $\forall i\ x_{mi}\lambda_m = 0$ ce qui correspond au fait que les variables explicatives ne sont pas indépendantes.

12.8 Résumé

Dans ce chapitre on a présenté trois exemples de modèles non linéaires généralisant directement les modèles linéaires vus précédemment. On a ainsi examiné

1. Les modèles dichotomiques, caractérisés par le fait que la variable explicative prend ses valeurs dans $\{0,1\}$. On a vu que des modélisations adaptées faisaient intervenir des variables latentes i.e. des variables dont seulement une partie de la réalisation est observée.
2. Deux exemples types sont les modèles Logit et les modèles Probit. Ces deux modèles s'estiment par le maximum de vraisemblance et nécessitent une étape d'optimisation.
3. On a également présenté les modèles Tobit. Ce sont des modèles dans lesquels on observe une variable conditionnellement à la valeur prise par une autre variable.
4. La situation standard est celle dans laquelle il y a une variable d'intérêt et une variable décrivant la sélection.
5. Un exemple typique est celui du salaire : on n'observe le salaire que conditionnellement au fait que le nombre d'heures de travail soit strictement positif.
6. Ces modèles nécessitent en général des hypothèses sur la loi des résidus des équations de sélection et de la variable d'intérêt.
7. On fait en souvent l'hypothèse de résidus normaux. Dans ce cas le modèle peut être estimé simplement soit par la méthode du maximum de vraisemblance, soit par une méthode alternative, dite de Heckman. Cette méthode donne simplement des

estimateurs mais est moins efficace que la méthode de maximum de vraisemblance. Elle consiste à estimer d'abord un modèle Probit pour l'équation de sélection, puis à partir des estimations à calculer un terme correctif dit ratio de Mills introduit ensuite dans la régression de la variable d'intérêt.

8. Dans ces modèles à sélection endogène il faut traiter la sélection comme on traiterait un régresseur endogène dans une équation linéaire. Il est ainsi nécessaire de disposer d'une variable intervenant dans l'équation de sélection et n'intervenant pas dans l'équation d'intérêt, faute de quoi les paramètres ne sont estimés que sur la non linéarité de la forme fonctionnelle.
9. Différentes généralisations ont été proposées pour obtenir des estimations avec des lois plus générales que la loi normale. Le modèle de sélection semiparamétrique généralise ainsi l'approche de Heckman. Une fonction polyméries de la probabilité de sélection est ainsi introduite au lieu du ratio de Mills. Ces modèles ne permettent pas en général l'estimation de la constante et nécessitent une fois abandonnée l'hypothèse de normalité l'exclusion d'un régresseur de la liste des variables explicatives affectant la variable d'intérêt.
10. Enfin on a présenté succinctement les modèles de choix discrets qui offrent une modélisation de la situation dans laquelle un individu doit arbitrer entre plusieurs choix possibles. L'intérêt de ces modèles est de présenter un lien étroit entre la théorie des choix et l'économétrie.

Chapitre 13

Evaluation

L'évaluation des politiques publiques nécessite souvent la connaissance de paramètres de comportements des agents qui sont inconnus. La mesure de l'effet d'une politique instaurant une taxe sur certains produits fait ainsi intervenir les élasticités d'offre et de demande de ces biens. De même, l'effet d'une politique favorisant le retour à l'emploi, tel que l'Earning Income Taxe Credit aux Etats Unis ou la Prime pour l'Emploi en France font intervenir l'élasticité de l'offre de travail. La mesure de ces paramètres est une préoccupation importante de l'économétrie. Les chapitres précédents ont montré la difficulté de l'estimation de ces paramètres et la nécessité de contextes observationnels très exigeants. La connaissance de ces paramètres permet d'apporter de nombreux éclairages sur les effets des politiques publiques. Par exemple l'estimation d'équations d'offre de travail permet de mesurer la valeur que les agents accordent au temps libre. L'évolution d'une telle valeur et sa dispersion dans la population est bien sur intéressante dans le contexte de la réduction du temps de travail. Connaître les paramètres structurels du comportements des agents permet de mesurer ex ante les effets probables d'une mesure de politique économique. Elle permet aussi de mesurer l'effet de politiques ayant déjà été mises en oeuvre.

Exemple *Laroque Salanié (2000) Modélisation de l'offre de travail en fonction de la rémunération et des transferts(modélisation d'un salaire de réserve), modélisation de la demande de travail (productivité d'un travailleur). Il y a emploi si le salaire offert (la productivité) est supérieur au salaire de réserve et au smic. On peut alors examiner l'effet d'un relèvement du smic ou l'effet d'une modification des transferts.*

Ces évaluations reposent sur la spécification de modèles de comportement et leur estimation. De nombreux paramètres structurels sont susceptibles d'intervenir et il est probable que les conditions de l'identification de ces paramètres ne soient pas réunies pour chacun d'entre eux. On peut être tenté d'apporter une réponse plus précise à une question plus générale. Plutôt que l'évaluation d'une politique basée sur la décomposition et la mesure des différentes composantes d'une politique (effet via l'offre et via la demande par exemple) et qui nécessitent l'estimation de tous les paramètres structurels (élasticités

d'offre et de demande par exemple) on peut chercher à répondre à la question globale quel a été l'effet de la politique au total ? Ceci ne nécessite que l'estimation de combinaisons des paramètres structurels et pas leur identification individuelle. Une branche de l'économétrie s'est développée fortement au cours des dernières années qui cherche à répondre à cette question. C'est essentiellement aux travaux de James Heckman que l'on doit ces avancées. Elle ne s'intéresse qu'à des évaluations ex-post et aux situations dans laquelle la politique in fine a concerné une partie de la population seulement. Par exemple effet du relèvement du salaire minimum dans certains états aux Etats Unis. Mise en place d'un système de formation pour les chômeurs, ou d'un système d'aide à la recherche d'emploi (PAP) etc... L'idée centrale est qu'une partie de la population bénéficie de la mesure et l'autre non. On peut sous certaines hypothèses, là aussi parfois exigeantes, retrouver l'effet de la politique sur les individus qui en ont bénéficiés, à partir de comparaisons entre les deux populations. On voit bien que mesurer l'effet global de la politique mise en oeuvre de cette façon est moins exigeant que la mesure de l'ensemble des paramètres structurels sous-jacents. Seule la façon dont ils se combinent pour conduire au résultat final compte. En pratique, on considère des politiques se traduisant par le fait que la population va être répartie dans différents états. On introduit ainsi une variable appelée variable de *traitement* T prenant ses valeurs dans $\{0, 1, ..., M\}$. L'état $T = 0$ correspondant au fait de n'être pas directement touché par la politique. On va s'intéresser principalement à la situation dans laquelle il n'y a que deux états : $T \in \{0, 1\}$. Les évaluations auxquelles on procède sont des évaluations ex post : elles concernent les politiques qui ont été déjà mises en oeuvre et ont déjà produit leurs effets. Le but est de définir et de mesurer l'ampleur de ces effets sur la base des information dont on dispose pour les individus traités et les individus non traités. Cette approche est ainsi dite "*observationnelle*" car ancrée dans l'observation des effets d'une politique.

Exemple *Stage de formation. La population va se décomposer en deux types d'individus : ceux bénéficiant du stage* $T = 1$,*dits traités, et ceux n'en bénéficiant pas* $T = 0$, *dits non traités. Il s'agit en fait du cas type qui a été largement étudié par Heckman (voir Heckman Lalonde et Smith (1999))*

Exemple *Modification de certains paramètres de la législation. Certains individus ne sont pas concernés par le changement de législation, d'autres le sont. Un exemple pourrait être le relèvement du Smic : les individus dont la rémunération avant le relèvement se trouve entre l'ancien et le nouveau smic sont dits traités et ceux dont la rémunération se trouve au delà du nouveau smic avant son relèvement sont dits non traités. Abowd, Kramarz et Margolis (1999) utilisent les augmentations successives du Smic depuis 1981 pour comparer chaque année les pertes d'emploi des salariés rattrapés par le Smic avec celle des autres salariés.*

13.1 Le Modèle causal

On définit pour chaque individu deux *outputs potentiels* y_1 et y_0. y_1 est la variable aléatoire caractérisant la situation de l'individu s'il bénéficie de la mesure, par exemple s'il suit le stage de formation. y_0 est la situation de l'individu lorsqu'il ne bénéficie pas de la mesure par exemple s'il ne suit pas le stage.

Ces deux grandeurs existent pour chaque individu, qu'il bénéficie ou non de la mesure. On définit l'effet causal comme étant :

$$\Delta = y_1 - y_0$$

Il s'agit donc de la différence entre la situation d'un individu lorsqu'il suit le stage avec sa situation lorsqu'il ne le suit pas.

13.1.1 Choix de la variable d'intérêt et choix de l'état de référence

Le choix de la variable y est important. Lorsqu'il s'agit d'évaluer une politique il est nécessaire de définir un critère. Concernant les stages de formation ce critère n'est pas nécessairement évident. Il peut s'agir de la situation vis à vis de l'emploi, du salaire, de la valeur d'un individu sur le marché du travail, du bien être de l'individu... Chacune de ces caractéristiques correspond à une valorisation différente du passage par un stage de formation et qui représente aussi le point de vue de différents agents.

La définition de l'état de référence est aussi une question importante. On peut au moins distinguer deux types de définitions pour l'état de référence :

- le traitement existe et on n'y participe pas y_0.
- le traitement n'existe pas $\widetilde{y_0}$.

On pourrait définir un effet causal $\widetilde{\Delta} = y_1 - \widetilde{y_0} = (y_1 - y_0) + (y_0 - \widetilde{y_0}) = \Delta + (y_0 - \widetilde{y_0})$. Le fait que y_0 puisse être différent de $\widetilde{y_0}$ correspond à l'existence d'effets indirects. Le fait qu'une mesure de politique économique soit prise peut affecter un individu même s'il n'est pas directement concerné par la mesure. Si on considère la situation dans laquelle deux individus sont en concurrence pour un emploi et qu'il y a un stage disponible seulement, on conçoit que les deux grandeurs y_0 et $\widetilde{y_0}$ soient différentes, et qu'omettre les effets indirects puisse conduire à une évaluation erronnée de la politique mise en oeuvre. Dans le cas du relèvement du smic examiné par Abowd Kramarz et Margolis, il est possible que la situation des individus non concernés directement par le relèvement du smic, c'est à dire les individus dont la rémunération avant le relèvement du smic est au dessus de la nouvelle valeur soient affectés malgré tout par le relèvement du smic. En effet ils ne sont plus en concurrence avec ceux dont la rémunération était en dessous du nouveau smic.

13.1.2 Paramètres d'intérêt

On s'intéresse en général à deux types de paramètres :
- $\Delta^{TT}(x) = E(y_1 - y_0 | T = 1, x)$
- $\Delta^{ATE}(x) = E(y_1 - y_0 | x)$

Le premier paramètre est l'effet moyen du traitement sur les individus de caractéristiques x ayant bénéficié de la mesure (Average Treatment Effect). Le second paramètre est l'effet moyen du traitement sur les individus de caractéristiques x qu'ils aient ou non bénéficié de la mesure (Treatment on the Treated). L'interprétation des ces deux paramètres est différente. Le premier ne concerne que la mesure des gains pour les individus ayant bénéficié du traitement alors que le second mesure l'effet du traitement s'il était étendu à l'ensemble de la population. Ils ont toutes les chances d'être différents puisque vraisemblablement le gain que l'on retire du traitement conditionne la décision de participation.

Ces paramètres ne sont pas directement identifiés. Dans l'idéal on souhaiterait pouvoir identifier la distribution jointe :

$$l(y_1, y_0, T)$$

Ceci permettrait d'identifier la loi jointe de l'effet causal et du traitement $l(\Delta, T)$, à la source du calcul de nombreux paramètres présentant un intérêt. On observe en effet un individu soit s'il bénéficie du traitement soit s'il n'en bénéficie pas, mais jamais dans les deux situations à la fois. Les observations sont ainsi :

$$\begin{cases} T \in \{1, 0\} \\ y = Ty_1 + (1 - T) y_0 \end{cases}$$

Les données ne permettent d'identifier que $l(T)$, $l(y_1 | T = 1) = l(y | T = 1)$ et $l(y_0 | T = 0) = l(y | T = 0)$. On voit que c'est toujours insuffisant pour estimer n'importe lequel des deux paramètres. En effet le premier paramètre s'écrit $\Delta^{ATE} = E(y_1 - y_0 | T = 1, x) = E(y | T = 1, x) - E(y_0 | T = 1, x)$, de telle sorte qu'il est nécessaire d'identifier $E(y_0 | T = 1, x)$ qui est inobservé. Le second paramètre nécessite l'identification non seulement de $E(y_0 | T = 1, x)$ mais aussi de $E(y_1 | T = 0, x)$.

Remarque *Ces paramètres s'interprètent comme les gains de surplus liés à la mise en oeuvre de la politique ou à son extension. Si on considère les trois outputs potentiels pertinents : y_1, y_0 et $\widetilde{y}_0$, et les surplus $\widetilde{W}_0$, W, W_T, associés respectivement aux situations sans la politique, avec la politique telle qu'elle a été mise en oeuvre et lorsque la politique est étendue. On calcule simplement les gains associés aux deux situations :*

$$W - \widetilde{W}_0 = N\left(P(T = 1) E\left(\Delta^{TT}(x) | T = 1\right) + E(y_0) - E(\widetilde{y}_0) - E(c | T = 1)\right)$$

Lorsque la politique est mise en oeuvre et que les individus y participent librement, et

$$W_T - \widetilde{W}_0 = N\left(E\left(\Delta^{ATE}(x)\right) + E(y_0) - E(\widetilde{y}_0) - E(c)\right)$$

Lorsque la politique est étendue à toute la population. On voit que le premier paramètre est pertinent pour juger de l'efficacité de la politique telle qu'elle a été mise en oeuvre alors que le second est pertinent pour juger de l'opportunité de son extension. On voit également qu'une évaluation complète doit aussi prendre en compte les effets indirects : la situation de référence change par l'instauration même du dispositif. De même une évaluation complète doit aussi faire intervenir les coûts du traitement. Les deux paramètres considérés, bien que centraux n'apporte donc qu'une partie de l'information nécessaire à des évaluations complètes. Enfin on remarque que pour la première situation, la probabilité de suivre le traitement intervient aussi de façon importante.

– **Remarque** *Modélisation des outputs potentiels Une modélisation permet de mieux comprendre la nature des paramètres* $\Delta^{TT}(x)$ *et* $\Delta^{ATE}(x)$ *et leurs différences. On modélise :*

$$\begin{cases} y_1 = \alpha_1 + x\beta_1 + u_1 \\ y_0 = \alpha_0 + x\beta_0 + u_0 \end{cases}$$

où on fait l'hypothèse que $(u_1, u_0) \perp x$. *Les coefficients* β_0 *et* β_1 *sont des paramètres susceptibles de recevoir une interprétation économique : ce sont des paramètres structurels caractérisant le comportement des agents. Les deux paramètres sont alors :*

$$\Delta^{ATE}(x) = E(y_1 - y_0 | x) = \alpha_1 - \alpha_0 + x(\beta_1 - \beta_0)$$

et

$$\Delta^{TT}(x) = E(y_1 - y_0 | x, T = 1) = \alpha_1 - \alpha_0 + x(\beta_1 - \beta_0) + E(u_1 - u_0 | x, T = 1)$$

On voit que le premier paramètre ne fait intervenir que les variables observées et les paramètres structurels et est donc de ce fait un paramètre standard de l'économétrie. Il n'en est pas de même en revanche du second paramètre qui fait intervenir les caractéristiques inobservées u_1 *et* u_0. *Les deux paramètres ne sont identiques que lorsqu'il y n'a pas d'hétérogénéité inobservable dans l'effet du traitement, soit* $u_1 = u_0$ *ou lorsqu'une telle hétérogénéité existe mais n'est pas prise en compte par les individus lors de la décision de participation au programme* $T \perp (u_1 - u_0) | x = 1$.

13.1.3 Biais de sélectivité

Definition *L'estimateur naïf est celui qui correspond à la comparaison de la situation moyenne des individus ayant fait l'objet d'un traitement et celle de ceux n'en ayant pas fait l'objet*

$$\widehat{\Delta}^0 = \overline{y}^{T=1} - \overline{y}^{T=0}$$

C'est estimateur est très populaire, largement répandu mais potentiellement biaisé. En effet la situation moyenne des individus ne bénéficiant pas du traitement n'est pas nécessairement la même que celle qui est pertinente : la situation moyenne des individus ayant bénéficié du traitement s'il n'en avait pas bénéficié.

Proposition *Pour que l'estimateur naïf soit un estimateur convergent de TT il faut que l'affectation au traitement soit indépendante de l'output potentiel y_0. Pour qu'il soit un estimateur convergent de ATE, il faut que l'affectation au traitement soit indépendante des deux outputs potentiels y_0 et y_1.*

Démonstration

$$\begin{aligned}\widehat{\Delta}^0 &\rightarrow \Delta^0 = E(y_1|T=1) - E(y_0|T=0)\\ &= E(y_1|T=1) - E(y_0|T=1) + E(y_0|T=1) - E(y_0|T=0)\\ &= \Delta^{TT} + B^{TT}\end{aligned}$$

On voit qu'il apparaît un biais de sélectivité : $B^{TT} = E(y_0|T=1) - E(y_0|T=0)$. Il est nul si $y_0 \perp T$. Pour Δ^{ATE} on a

$$E(y_1) = P(T=1)E(y_1|T=1) + (1 - P(T=1))E(y_1|T=0)$$

d'où

$$E(y_1|T=1) = E(y_1) + (1 - P(T=1))[E(y_1|T=1) - E(y_1|T=0)]$$

de même

$$E(y_0|T=0) = E(y0) - P(T=1)[E(y_0|T=1) - E(y_0|T=0)]$$

d'où

$$\begin{aligned}\Delta^0(x) &= \Delta^{ATE} + (1 - P(T=1))[E(y_1|T=1) - E(y_1|T=0)] +\\ &\quad P(T=1)[E(y_0|T=1) - E(y_0|T=0)]\\ &= \Delta^{ATE} + B^{ATE}\end{aligned}$$

Il est nécessaire d'avoir l'indépendance de T et du couple d'output potentiels (y_1, y_0) conditionnellement aux x.

Remarque *On voit en outre que*

$$B^{ATE} = B^{TT} + (1 - P(T=1))[E(y_1 - y_0|T=1) - E(y_1 - y_0|T=0)]$$

La deuxième source de biais provient de l'hétérogénéité du traitement, alors que la première source de biais provient du problème classique d'endogénéité de la variable de traitement.

13.2 L'estimateur des Différences de Différences

Le cadre des régressions précédentes permet de présenter certains estimateurs standards très fréquemment utilisés. Il s'agit de l'estimateur "*Avant-Après*" ou "*Before-After*", de l'estimateur en *coupe* "*Cross section*" et de l'estimateur par *double différence* "*Difference in difference*"

Les estimateurs Avant Après et par Différence de différence font intervenir le temps. On suppose donc que le traitement est mis en oeuvre à une date t_0 et que l'on dispose d'informations sur les individus en $\underline{t} < t$ et en $\overline{t} > t$ pas nécessairement pour des individus similaires.

13.2.1 Estimateur en coupe

L'estimateur en coupe est très proche de celui que l'on pourrait déduire du modèle de régression précédent. Le modèle sous sa forme générale s'écrirait comme :

$$y_{\overline{t}} = \alpha_0 + x_{\overline{t}}\beta_0 + T\left(\alpha_1 - \alpha_0 + x\left(\beta_1 - \beta_0\right)\right) + \underbrace{u_{0,\overline{t}} + T\left(u_{1,\overline{t}} - u_{0,\overline{t}}\right)}_{v_{\overline{t}}}$$

Les estimateurs standards en coupe ignorent l'hétérogénéité de l'effet du traitement. L'équation précédente se réécrit donc :

$$y_{\overline{t}} = \alpha + x_{\overline{t}}\beta + Tc + u_{\overline{t}}$$

Dans ce cadre le biais est simplement lié au fait que l'on n'a pas forcément $E\left(u_{\overline{t}}\,|x, T\right) = 0$. Le biais a pour expression :

$$B^{Cross} = E\left(u_{\overline{t}}\,|x, T = 1\right) - E\left(u_{\overline{t}}\,|x, T = 0\right)$$

La mise en oeuvre de cet estimateur ne nécessite que des informations en coupe sur une période suivant le traitement. Une version encore plus simple de cet estimateur consiste à négliger les variables de conditionnement. Dans ce cas l'estimateur est simplement

$$\widehat{\Delta}^{Cross} = \overline{y_{\overline{t}}}^{T=1} - \overline{y_{\overline{t}}}^{T=0}$$

13.2.2 Estimateur Avant-Après

L'estimateur avant après est basé sur des informations sur des données temporelles d'individus ayant fait l'objet du traitement. L'idée générale est que les informations dont on dispose sur les individus avant le traitement permettent de reconstituer ce qu'aurait été leur situation en l'absence de traitement. L'estimateur est défini comme la différence des moyennes des individus traités après et avant le traitement. Dans sa forme la plus simple son expression est donnée par :

$$\widehat{\Delta}^{BA} = \overline{y_{\overline{t}}}^{T=1} - \overline{y_{\underline{t}}}^{T=1}$$

Dans le cadre des régressions précédentes il s'écrirait à partir des régressions :

$$\begin{aligned} y_{\overline{t}} &= \alpha + x_{\overline{t}}\beta + c + u_{\overline{t}} \text{ pour } T = 1 \\ y_{\underline{t}} &= \alpha + x\underline{t}\beta + u_{\underline{t}} \text{pour } T = 1 \end{aligned}$$

Soit le modèle de régression :

$$y_t = \alpha + x_t\beta + c1\left(t = \overline{t}\right) + u_{\overline{t}}1\left(t = \overline{t}\right) + u_{\underline{t}}1\left(t = \underline{t}\right) \text{ pour } T = 1$$

Il y a deux problèmes principaux avec cet estimateur. Le premier problème provient du biais classique déjà analysé. Le biais pour cet estimateur est donné par :

$$B_{BA} = E\left(u_{\overline{t}} \middle| x, T = 1\right) - E\left(u_{\underline{t}} \middle| x, T = 1\right)$$

Supposons que le terme de perturbation soit la somme de deux éléments : $u_t = u + \varepsilon_t$ avec ε_t non corrélé dans le temps, alors le terme de biais précédent se réécrit :

$$\begin{aligned} E\left(u_{\overline{t}} \middle| x, T = 1\right) - E\left(u_{\underline{t}} \middle| x, T = 1\right) &= E\left(u \middle| x, T = 1\right) + E\left(\varepsilon_{\overline{t}} \middle| x, T = 1\right) - \\ &\quad E\left(u \middle| x, T = 1\right) - E\left(\varepsilon_{\underline{t}} \middle| x, T = 1\right) \\ &= E\left(\varepsilon_{\overline{t}} \middle| x, T = 1\right) - E\left(\varepsilon_{\underline{t}} \middle| x, T = 1\right) \end{aligned}$$

Si la décision de participation dépend de la chronique des éléments inobservés alors ce terme est non nul. En particulier on a observé que la participation à des programme de formation aux Etats-Unis était en général associée à une baisse des revenus passés, c'est à dire à des éléments $\varepsilon_{\underline{t}}$ faibles.

Le second terme de biais est encore plus radical. Supposons qu'en l'absence de politique le modèle s'écrive

$$y_t = \alpha_t + x_t\beta + u_t$$

Le α_t représente par exemple des chocs macroéconomiques. Alors le modèle précédent se réécrit :

$$y_t = \alpha_{\underline{t}} + x_t\beta + \left(c + \alpha_{\overline{t}} - \alpha_{\underline{t}}\right)1\left(t = \overline{t}\right) + \left\{u_{\overline{t}}1\left(t = \overline{t}\right) + u_{\underline{t}}1\left(t = \underline{t}\right)\right\} \text{ pour } T = 1$$

Il est impossible de séparer l'effet du traitement de l'effet de chocs macroéconomiques.

Remarquons que si le traitement s'adresse à des individus qui sont repérables ex ante : $T = \{z \in Z\}$, alors l'estimateur précédent ne nécessite pas de données temporelles. Seules des coupes successives pour les individus tels que $\{z \in Z\}$ sont nécessaires.

13.2.3 Estimateur par différence de différence.

Cet estimateur combine les deux estimateurs précédents. Il correspond à la situation dans laquelle le traitement correspond à la réalisation à partir d'une date donné d'un certain nombre de conditions d'éligibilité qui sont observables. On peut donc définir une variable T correspondant aux conditions d'éligibilité, sur des observations temporelles.

Elle ne correspond au traitement que pour $t = \overline{t}$ postérieur à la date de traitement. Dans le cadre du modèle de régression précédent, il correspond à la situation dans laquelle on introduit une indicatrice correspondant à la date, une indicatrice correspondant aux conditions d'éligibilité et le produit croisé indicatrice temporelle post et conditions d'éligibilité :

$$y_t = x_t\beta + \gamma_c + \gamma_{\overline{t}} 1\left(t = \overline{t}\right) + \gamma_T T + \gamma_{\overline{t},T} T 1\left(t = \overline{t}\right) + v_t$$

Proposition *Lorsque le biais d'une estimation en coupe est constant dans le temps ce qui est équivalent au fait que le biais avant-après soit le même pour les éligibles et les non éligibles, la régression introduisant comme variables une indicatrice temporelle post, une indicatrice pour les conditions d'éligibilité et le produit de ces deux variables permet d'estimer l'effet du traitement.*

Démonstration *On peut examiner à quoi correspondent ces différents termes dans le cadre du modèle précédent :*

$$y_t = x_t\beta + \alpha_t + cT + u_t$$

On a

$$E\left(y_t \left| x_t, t, T\right.\right) = x_t\beta + \alpha_t + cT + E\left(u_t \left| x_t, t, T\right.\right) = x_t\beta + \alpha_t + cT + E\left(u_t \left| t, T\right.\right)$$

On introduit $m_{t,T} = E\left(v_t \left| t, T\right.\right)$, *on a*

$$\begin{aligned}
E\left(u_t \left| t, T\right.\right) &= m_{\overline{t},1} T 1\left(t = \overline{t}\right) + m_{\overline{t},0}\left(1 - T\right) 1\left(t = \overline{t}\right) + m_{\underline{t},1} T 1\left(t = \underline{t}\right) + m_{\underline{t},0}\left(1 - T\right) 1\left(t = \underline{t}\right) \\
&= m_{\overline{t},0} 1\left(t = \overline{t}\right) + m_{\underline{t},0} 1\left(t = \underline{t}\right) + \left(m_{\overline{t},1} - m_{\overline{t},0}\right) T 1\left(t = \overline{t}\right) + \left(m_{\underline{t},1} - m_{\underline{t},0}\right) T 1\left(t = \underline{t}\right) \\
&= m_{\underline{t},0} + \left(m_{\overline{t},0} - m_{\underline{t},0}\right) 1\left(t = \overline{t}\right) + \left(m_{\underline{t},1} - m_{\underline{t},0}\right) T \\
&\quad + \left[\left(m_{\overline{t},1} - m_{\overline{t},0}\right) - \left(m_{\underline{t},1} - m_{\underline{t},0}\right)\right] T 1\left(t = \overline{t}\right)
\end{aligned}$$

On voit donc que les coefficients de la régression s'écrivent :

$$\begin{aligned}
\gamma_c &= m_{\underline{t},0} \\
\gamma_{\overline{t}} &= \left(m_{\overline{t},0} - m_{\underline{t},0}\right) = B_{BA}\left(T = 0\right) \\
\gamma_T &= \left(m_{\underline{t},1} - m_{\underline{t},0}\right) = B^{Cross}\left(\overline{t}\right) \\
\gamma_{\overline{t},T} &= \left[\left(m_{\overline{t},1} - m_{\overline{t},0}\right) - \left(m_{\underline{t},1} - m_{\underline{t},0}\right)\right] = B^{Cross}\left(\overline{t}\right) - B^{Cross}\left(\underline{t}\right) \\
\gamma_{\overline{t},T} &= \left[\left(m_{\overline{t},1} - m_{\underline{t},1}\right) - \left(m_{\overline{t},0} - m_{\underline{t},0}\right)\right] = B_{BA}\left(T = 1\right) - B_{BA}\left(T = 0\right)
\end{aligned}$$

L'estimateur par différence de différence résout donc directement le problème précédent d'instabilité du modèle sous-jacent.

On en conclut que la régression en incluant une indicatrice correspondant au traitement, capture le biais de sélectivité de la coupe, en incluant une indicatrice temporelle capture le biais de l'estimation Before After, et qu'en introduisant le produit croisé condition d'éligibilité $\times$ *indicatrice post elle va estimer le coefficient* $\Delta + B^{Cross,\overline{t}} - B^{Cross,\underline{t}} = \Delta + B^{BA,T=1} - B^{BA,T=0}$. *Le biais est donc nul dans le cas de l'estimateur par différence de différence lorsque* $B^{Cross,\overline{t}} - B^{Cross,\underline{t}} = 0$ *ou encore si* $B^{BA,T=1} - B^{BA,T=0}$.

Si on reprend la modélisation simple des perturbations présentées pour l'estimateur Avant Après : $u_t = u + \varepsilon_t$ La différence des termes de biais s'écrit :

$$\begin{aligned} B^{BA,T=1} - B^{BA,T=0} &= \{E(\varepsilon_{\overline{t}}|x, T=1) - E(\varepsilon_{\underline{t}}|x, T=1)\} - \\ &\quad \{E(\varepsilon_{\overline{t}}|x, T=0) - E(\varepsilon_{\underline{t}}|x, T=0)\} \end{aligned}$$

On voit que si la participation au traitement est conditionnée par des chocs négatifs sur la variable d'output, alors ce terme n'est pas nul.

On appelle cet estimateur différence de différence car dans le cas où il n'y a pas de variables explicatives il s'écrit simplement. Il nécessite aussi en général des informations longitudinales sur les individus traités et non traités. Dans sa forme la plus simple cet estimateur s'écrit simplement

$$\begin{aligned} \widehat{\Delta}^{DD} &= \left(\overline{y_{\overline{t}}}^{T=1} - \overline{y_{\underline{t}}}^{T=1}\right) - \left(\overline{y_{\overline{t}}}^{T=0} - \overline{y_{\underline{t}}}^{T=0}\right) \\ &= \widehat{\Delta}^{BA,T=1} - \widehat{\Delta}^{BA,T=0} \\ &= \left(\overline{y_{\overline{t}}}^{T=1} - \overline{y_{\overline{t}}}^{T=0}\right) - \left(\overline{y_{\underline{t}}}^{T=1} - \overline{y_{\underline{t}}}^{T=0}\right) \\ &= \widehat{\Delta}^{Cross,\overline{t}} - \widehat{\Delta}^{Cross,\underline{t}} \end{aligned}$$

13.2.4 Exemple : La Contribution Delalande

La contribution Delalande est une taxe sur le licenciement des travailleurs âgés. Elle a été créée en 1987 à l'instigation du député Delalande. Dans le schéma initial, le licenciement d'un salarié de plus de 50 ans conduisait à une taxe correspondant à 3 mois de salaire. Ce schéma initial a été profondément modifié à deux reprises, une fois en 1992 et une fois en 1998. Le schéma final est particulièrement désincitatif puis qu'il conduit à une taxe correspondant à un an de salaire pour les salariés de plus de 56 ans. dès 1992 l'âge seuil d'entrée dans le dispositif a été abaissé à 50 ans. Ce type de politique est susceptible d'avoir deux effets, l'un direct et l'autre indirect. L'effet direct correspond au fait que le licenciement des travailleurs âgés deviennent moins attractif et donc se réduise. L'effet indirect correspond au fait que ce type de politique est susceptible de rendre l'embauche de salariés moins attractive et donc réduise les embauches. A ce titre la modification du dispositif Delalande en 1992 introduisait une spécificité qui permet de mesurer l'ampleur de ce phénomène. A partir de 1992 les employeurs embauchant un salarié de plus de 50 ans ne sont plus redevable de la contribution Delalande en cas de licenciement de ce salarié. Une façon naturelle d'étudier l'effet désincitatif de la contribution Delalande consiste donc à comparer les taux d'embauche de salariés de plus de 50 ans et de moins de 50 ans autour de 1992. L'idée est que le renforcement important du dispositif en 1992 a conduit réduire les embauches de salariés de moins de 50 ans. Dans la mesure où les demandeurs d'emploi de plus de 50 ans ont été exclus de ce dispositif, on ne doit pas observer de dégradation similaire de l'embauche de chômeurs de plus de 50 ans. On peut donc examiner l'effet de la contribution Delalande de différentes façons :

	Sans contrôles			Avec contrôles		
	48-51 ans	46-53 ans	44-55 ans	48-51 ans	46-53 ans	44-55 ans
Avant 1992, <50 ans	20,0	20,3	19,7	19,4	20,0	18,8
	2,9	1,7	1,3	2,8	1,7	1,2
Avant 1992, >50 ans	20,5	14,9	13,7	19,1	14,5	13,9
	2,9	1,4	1,0	2,7	1,4	1,0
Après 1992, <50 ans	14,3	14,6	14,9	14,6	14,7	14,8
	1,7	1,0	0,8	1,7	1,0	0,8
Après 1992, >50 ans	14,6	15,2	13,0	15,3	15,5	13,4
	1,8	1,1	0,8	1,8	1,1	0,8
Avant 1992, différence - 50/+50	-0,5	5,4	6,0	0,3	5,5	4,9
	4,1	2,2	1,6	3,9	2,1	1,6
Après 1992, différence - 50/+50	-0,3	-0,7	2,0	-0,6	-0,8	1,4
	2,5	1,5	1,1	2,5	1,5	1,1
Différence de différence	0,2	-6,1	-4,1	-0,9	-6,3	-3,5
	4,7	2,7	2,0	4,6	2,6	1,9
Nombre d'observations	1 211	3 661	6 179	1 211	3 661	6 179

TAB. 13.1 – Contribution Delalande - Estimation de l'effet indirect par la méthode des différences de différences

- Avant après : Comparaison de la variation du taux d'embauche des moins de 50 ans entre avant et après 1992
- En coupe : Comparaison des taux d'embauche des moins de 50 ans et des plus de 50 ans après 1992
- En Différence de Différence : Comparaison de la variation du taux d'embauche des moins de 50 ans et des plus de 50 ans avant et après 1992

On peut examiner cette question à partir des transitions Chômage-Emploi. L'Enquête Emploi fournit les informations nécessaires. Dans l'idéal on souhaiterait comparer les taux d'embauche de chômeurs de juste moins de 50 ans et de juste plus de 50 ans. En pratique ceci n'est pas possible car il n'y a pas suffisamment d'observations de ce type dans l'enquête emploi. On est amené à considérer des fenêtres plus larges. On parvient aux résultats reportés dans le tableaux 13.1

Le tableau se présente en deux parties droite et gauche. La partie droite reporte les résultats portant sur des comparaisons brutes, celle de gauche ceux obtenus lorsque l'on corrige des caractéristiques inobservables des agents. Chaque partie comprend trois colonnes correspondant aux différentes fenêtres considérées : étroite, moyenne, large. Les quatre premières lignes présentent les taux de retour à l'emploi en CDI pour les moins de 50 ans et pour les plus de cinquante ans avant 1992, puis après 1992.

On constate que le taux annuel de retour à l'emploi des hommes de 48 ans, avant 1992, était de 20% en moyenne, quantité estimée de façon peu précise comme en témoigne l'écart-type (2,9%). Le taux de retour à l'emploi des plus de cinquante ans s'élève alors à 20,5% et est lui aussi peu précisément estimé. Cette imprécision tient largement à la taille

de l'échantillon mobilisé (1 211 individus-années). Introduire des variables de contrôle ne change les ordres de grandeur ni des paramètres, ni des écarts-type. C'est cette imprécision qui motive le choix de fenêtres plus larges. Ceci conduit à introduire des individus moins directement représentatifs de la comparaison effectuée mais permet d'obtenir des écarts-type plus réduits. L'élargissement conduit au résultat attendu : les taux bruts ou nets estimés sont beaucoup plus précis

Les cinquièmes et sixièmes lignes présentent les différences entre les taux de retour à l'emploi des plus et des moins de 50 ans, avant et après 1992. Avant 1992, le taux de retour à l'emploi des moins de 50 ans est généralement plus élevé que celui des plus de 50 ans (différence de 5,4 points pour la fenêtre 46-53 ans). On constate que les écarts-type sont beaucoup plus importants que pour les estimations des taux eux-mêmes, ce qui provient du fait que (pour les taux bruts) les estimateurs sont indépendants et que de ce fait la variance de leur différence est la somme des variances. L'imprécision est très sensible pour la fenêtre étroite si bien que la différence entre les taux n'est pas statistiquement significative. Dans les échantillons plus larges (pour les deux autres fenêtres), on voit apparaître un écart positif et significatif entre les taux de retour à l'emploi des plus et moins de 50 ans, avant 1992. Ce résultat n'est pas totalement satisfaisant, dans la mesure où le choix des fenêtres d'observation était motivé par le fait que les deux catégories d'individus devaient être très proches. Les différences de taux de retour à l'emploi s'inversent ou s'atténuent après 1992, et restent plus sensibles au choix de la fenêtre.

La dernière ligne du tableau présente les résultats en différence de différence, c'est-à-dire compare la façon dont les écarts de taux de retour à l'emploi des plus et des moins de 50 ans ont évolué entre les périodes antérieures et postérieures à 1992. La fenêtre de 46-53 ans est un bon compromis entre taille et comparabilité des échantillons. Selon cet estimateur, le taux relatif de retour à l'emploi se serait dégradé pour les moins de 50 ans de 6,1 points (6,3 points après contrôle des effets de structure). Cet effet est statistiquement différent de 0, et il est d'une ampleur conséquente. Il convient néanmoins de noter que l'effet n'apparaît pas sur une petite fenêtre d'âge, peut-être en raison d'échantillons trop petits (les écarts-type sont plus élevés), et apparaît atténué et à la limite de la significativité si on considère la fenêtre d'âges élargie.

13.3 Indépendance conditionnelles à des observables

13.3.1 Identification sous l'hypothèse d'indépendance conditionnelles à des observables

L'effet moyen du traitement pour les individus de caractéristiques x n'est pas identifié sans hypothèses sur la loi jointe des outputs potentiels et du traitement conditionnellement à x. En effet, pour estimer l'effet moyen du traitement sur les traités $E(y_1 - y_0 | x, T = 1)$,il est nécessaire d'identifier $E(y_0 | x, T = 1)$ alors que les données

ne permettent d'identifier que $E(y_0|x, T=0) = E(y|x, T=0)$. De même pour identifier l'effet du traitement dans la population, il est nécessaire d'identifier $E(y_0|x, T=1)$ et également $E(y_1|x, T=0)$, alors que concernant y_1 seul $E(y_1|x, T=1) = E(y|x, T=1)$ est identifiable.

Un premier ensemble d'hypothèses identifiantes consiste à faire l'hypothèse que ces quantités sont égales :

Definition *On dit qu'il y a indépendance forte conditionnellement à des observables s'il existe un ensemble de variables observables $\widetilde{x}$ tel que :*

$$l(y_1, y_0|T, \widetilde{x}) = l(y_1, y_0|\widetilde{x})$$

On dit qu'il y a indépendance faible conditionnellement à des observables s'il existe un ensemble de variables observables $\widetilde{x}$ tel que :

$$l(y_0|T, \widetilde{x}) = l(y_0|\widetilde{x})$$

Proposition *L'hypothèses d'indépendance faible est suffisante pour identifier le paramètre Δ^{TT}, en revanche, pour identifier le paramètre Δ^{ATE} il est nécessaire d'avoir recours à l'hypothèse d'indépendance forte.*

Proposition *En effet dans ces conditions, $l(y_0|\widetilde{x}) = l(y_0|T, \widetilde{x}) = l(y_0|T=0, \widetilde{x}) = l(y|T=0, \widetilde{x})$ la densité de l'output potentiel est identifiée et on peut donc estimer $E(y_0|\widetilde{x}, T=1) = E(y|\widetilde{x}, T=0)$*

Pour comprendre la signification de cette hypothèse, on peut revenir à la modélisation des outputs précédentes :

$$\begin{cases} y_1 = \alpha_1 + x\beta_1 + u_1 \\ y_0 = \alpha_0 + x\beta_0 + u_0 \end{cases}$$

On a pour y_0 par exemple :

$$E(y_0|T, x) = \alpha_0 + x\beta_0 + E(u_0|T, x) = g_0(x, T)$$

si il existe une source de variabilité commune à u_0 et T conditionnellement à x alors on aura $E(y_0|T=1, x) \neq E(y_0|T=0, x)$. Si néanmoins on est capable d'étendre l'ensemble des variables observables en $\widetilde{x}$ de telles sorte que l'on puisse épuiser les sources de variabilité commune entre u_0 et T alors on aura

$$E(y_0|T, \widetilde{x}) = \alpha_0 + x\beta_0 + E(u_0|T, \widetilde{x}) = g_0(\widetilde{x})$$

L'hypothèse d'indépendance conditionnellement à des observables consiste à supposer que l'on est capable de contrôler pour ces sources de variabilité. Remarquons qu'alors la fonction $g_0(\widetilde{x})$ ne reçoit plus d'interprétation économique alors que cela pouvait être le cas pour $\alpha_0 + x\beta_0$. Dans cette approche on accepte de perdre des informations sur le

comportement des individus : on ne peut plus distinguer l'effet spécifique de x sur y_0 de son effet transitant par $E(u_0|\widetilde{x})$. Le point important est qu'à ce prix, il est possible de construire pour chaque individu traité de caractéristique $\widetilde{x}$ un contrefactuel, c'est à dire une estimation de ce qu'aurait pu être sa situation en l'absence de traitement, par le biais de $g_0(\widetilde{x})$.

13.3.2 Le score de propension (propensity score)

La dimension de l'ensemble des variables de contrôle à introduire pour assurer l'indépendance entre le traitement et les outputs potentiels est souvent élevé, ce qui peut conduire à des complications importantes, notamment pour la mise en oeuvre de version semi paramétrique des estimateurs. Rubin et Rosenbaum (1983) ont montré un résultat important permettant de nombreuses simplifications pratiques :

Proposition *S'il y a indépendance conditionnellement à des observable, alors il y a indépendance conditionnellement au score :* $P(T_i = 1|x_i)$:

$$y_0 \perp T|\widetilde{x} \Longrightarrow y_0 \perp T|P(T=1|\widetilde{x})$$

Démonstration *On note* $s = P(T=1|\widetilde{x})$

$$\begin{aligned} P(T=1|s,y_0) &= \int P(T=1|\widetilde{x},y_0)\, l(\widetilde{x}|s,y_0)\, dx = \int P(T=1|\widetilde{x})\, l(\widetilde{x}|s,y_0)\, d\widetilde{x} \\ &= \int sl(\widetilde{x}|s,y_0)\, d\widetilde{x} = s \end{aligned}$$

De même, $P(T=1|s) = s$
On a donc : $P(T=1|s,y_0) = P(T=1|s)$

Ainsi le problème de la dimension peut être résolu de façon drastique : il est seulement nécessaire de conditionner par une unique variable quelque soit la dimension de l'ensemble initialement introduit.

Ainsi une étape initiale de toute évaluation consiste en une régression expliquant l'affectation au traitement. Elle est faite par exemple en utilisant un modèle Logit.

Remarque *Si* $\widetilde{s}$ *est un ensemble d'information plus large que* s*, par exemple* $\widetilde{s} = \{s, g(\widetilde{x})\}$*, le résultat demeure :* $P(T=1|\widetilde{s},y_0) = P(T=1|\widetilde{s})$*. un tel ensemble d'information est appelé "**balancing score**". La propriété de Rosenbaum et Rubin est en toute généralité que lorsqu'il y a indépendance conditionnelle à des observables, il y a aussi indépendance conditionnellement à n'importe quel balancing score.*

13.3.3 Méthodes d'estimation

Il y a principalement trois méthodes d'estimation. Une basée sur des régressions, une basée sur des appariements entre individus traité et individus non traités et une basée sur

des pondérations. Toutes ces méthodes mettent l'accent sur l'hétérogénéité de l'effet du traitement au sein de la population.

Les deux premières estimations ont des caractéristiques communes. Pour chaque individu traité de caractéristique x_i on cherche un estimateur de ce qu'aurait pu être sa situation en l'absence de traitement, i.e $E(y_0 | T = 1, x = x_i)$. La propriété d'indépendance permet d'écrire $E(y_0 | T = 1, x = x_i) = E(y_0 | T = 0, x = x_i) = E(y | T = 0, x = x_i)$. Les procédures d'estimation consiste à estimer de façon aussi peu restrictive que possible la fonction $E(y | T = 0, x = x_i)$. L'estimateur calculé in fine est alors défini par

$$\widehat{E}(\Delta | T = 1, x_i \in X) = \frac{1}{N_{1,X}} \sum_{\{T_i=1, x_i \in X\}} y_i - \widehat{E}(y | T = 0, x = x_i)$$

La fonction $E(y | T = 0, x = x_i)$ peut être estimée de différente façon correspondant aux approche par régression ou par appariement.

Régression :

Une première façon d'estimer l'effet du traitement consiste à procéder à la régression de la variable d'output observée sur le traitement et les variables de contrôle.

Proposition *Dans la régression*

$$E(y | T, x) = h(x) + Tg(x)$$

La propriété d'indépendance faible $E(y_0 | T, x) = E(y_0 | x)$ *permet d'identifier* $g(x) = E(y_1 - y_0 | T = 1, x)$. *On peut estimer* $\Delta^{TT} = E(g(x) | T = 1)$ *à partir d'une estimation convergente de* g *comme*

$$\widehat{\Delta}^{TT} = \frac{1}{N_1} \sum_{T_i=1} \widehat{g}(x_i)$$

La propriété d'indépendance forte $E(y_0 | T, x) = E(y_0 | x)$ *et* $E(y_1 | T, x) = E(y_1 | x)$ *permet d'identifier* $g(x) = E(y_1 - y_0 | T = 1, x) = E(y_1 - y_0 | T = 1, x)$. *On peut estimer* $\Delta^{TT} = E(g(x) | T = 1)$ *à partir d'une estimation convergente de* g *comme précédemment et*$\Delta^{ATE} = E(g(x))$

$$\widehat{\Delta}^{ATE} = \frac{1}{N} \sum \widehat{g}(x_i)$$

Démonstration *Comme* $y = y_0(1 - T) + y_1 T = y_0 + T(y_1 - y_0)$, *on a :*

$$E(y | T, x) = E(y_0 | T, x) + TE(y_1 - y_0 | T, x) = E(y_0 | T, x) + TE(y_1 - y_0 | T = 1, x)$$

Comme $E(y_0 | T, x) = E(y_0 | x)$, *on a donc*

$$E(y | T, x) = E(y_0 | x) + TE(y_1 - y_0 | T = 1, x)$$

et on a bien $g(x) = E(y_1 - y_0 | T = 1, x)$

Une estimation non paramétrique de y sur la variable de traitement et les variables de conditionnement permet donc en présence de la seule hypothèse $y_0 \perp T \,|x$ d'identifier le paramètre $\Delta^{TT}(x)$. En pratique : si la propriété d'indépendance est vraie, elle est aussi vraie pour le score (propriété de Rosenbaum et Rubin) Les régressions peuvent donc être basées sur le score et non sur l'ensemble des variables explicatives. On peut en pratique procéder aux régressions suivantes sur les populations séparées :

$$\begin{aligned} y &= \sum_{j=1}^{J} \alpha_j^1 f_j(s) + w^1 \qquad \text{pour } T=1 \\ y &= \sum_{j=1}^{J} \alpha_j^0 f_j(s) + w^0 \qquad \text{pour } T=0 \end{aligned}$$

où s est le score. Pour l'effet du traitement sur les traités, on estime alors :

$$\widehat{E}(\Delta \,|T=1) = \frac{1}{N_1} \sum_{T_i=1} y_{1i} - \sum_{j=1}^{J} \widehat{\alpha}_j^0 f_j(s_i)$$

ou aussi :

$$\widehat{E}(\Delta \,|T=1) = \frac{1}{N_1} \sum_{T_i=1} \sum_{j=1}^{J} \left(\widehat{\alpha}_j^1 - \widehat{\alpha}_j^0\right) f_j(s_i)$$

Le deuxième estimateur est un peu moins précis puisqu'il incorpore la variance du résidu mais il évite d'avoir à spécifier et estimer l'équation d'output pour les individus traités.

Remarque *L'intérêt de cette méthode est qu'elle apparaît comme un prolongement naturel de la régression à variables de contrôle* $y = xb + \Delta T + u$.

Appariement

Pour chaque individu traité $\widetilde{i}$, ayant des caractéristiques $x_{\widetilde{i}}$, on cherche un individu non traité $j\left(\widetilde{i}\right)$, ayant les mêmes caractéristiques observables, i.e $j\left(\widetilde{i}\right) \in \{j \,|T_j = 0, x_j = x_{\widetilde{i}}\}$ .On estime alors l'effet du traitement pour l'individu i par $\widehat{\Delta}_i = y_i - y_{j\left(\widetilde{i}\right)}$. On compare ainsi l'output de l'individu considéré et l'output d'un individu non traité ayant les mêmes caractéristiques observables. Le terme d'appariement provient de l'idée que chaque individu traité est apparié avec son jumeau non traité.

La quantité $y_{j\left(\widetilde{i}\right)}$ est un estimateur (non paramétrique) de

$$E(y \,|T=0, x=x_{\widetilde{i}}) = E(y_0 \,|T=0, x=x_{\widetilde{i}}) = E(y_0 \,|x=x_{\widetilde{i}}) = E(y_0 \,|T=1, x=x_{\widetilde{i}})$$

L'estimateur calculé finalement est obtenu en prenant la moyenne de la quantité $c_{\widetilde{i}} = y_{\widetilde{i}} - y_{j\left(\widetilde{i}\right)}$ sur la population traitée à laquelle on s'intéresse :

$$\widehat{E}\left(\Delta \left| T=1\right.\right) = \frac{1}{N_1}\sum_{T_i=1} y_i - y_{j(i)}$$

En pratique il n'est pas toujours possible de trouver pour chaque individu traité, un individu non traité ayant les mêmes caractéristiques que l'individu traité considéré. On peut alors choisir l'individu apparié de telle sorte que $\left\| x_{\widetilde{i}} - x_{j\left(\widetilde{i}\right)} \right\|_{\Sigma}$ soit minimal, pour Σ une métrique donnée. Une métrique naturelle dans ce cas est la métrique de Mahalanobis $\Sigma = V\left(x\right)^{-1}$.

Néanmoins la qualité de cet appariement peut être mauvaise en pratique : pour certains individus traités, il n'existe pas d'individu proche non traité notamment dans le cas où il y a un grand nombre de variables de conditionnement. La propriété de Rosenbaum et Rubin simplifie beaucoup l'appariement dans ce cas. En effet cette propriété permet de procéder à des appariements sur la base du seul résumé des variables de conditionnement que constitue le score. On peut ainsi apparier des individus dont les caractéristiques peuvent être très éloignées, mais qui ont des scores proches.

Ceci constitue le principe de l'appariement tel qu'il a été développé par les statisticiens. De nombreuses questions restent néanmoins non résolues : doit on faire l'appariement avec ou sans rejet ? Un individu non traité une fois apparié doit-il être évincé de l'ensemble des individus susceptibles d'être appariés avec les individus non traités restants. Si on choisit qu'un individu ne peut être apparié qu'une seule fois alors la qualité de l'appariement se dégradera progressivement. La question est alors de savoir par où commencer. De même, si on dispose d'un échantillon d'individu non traité très vaste, ne peut on pas tirer partie des individus qui in fine n'auront pas été appariés. Enfin, ce principe d'appariement tel qu'il est exprimé ne permet pas de préciser le comportement asymptotique de l'estimateur proposé.

Extension Kernel matching estimator Les méthodes d'appariement se généralisent directement dès lors que l'on interprète $y_{j\left(\widetilde{i}\right)}$ comme un estimateur non paramétrique de $E\left(y_0 \left| T=0, x=x_{\widetilde{i}}\right.\right)$. Différents autres types d'estimateurs non paramétriques peuvent être envisagés. Ils consistent tous à remplacer $y_{j\left(\widetilde{i}\right)}$ par une moyenne pondérée des observations de l'échantillon de contrôle :

$$\widehat{E}\left(y_0 \left| T=1, x=x_{\widetilde{i}}\right.\right) = \sum_{T_j=0} w_N\left(\widetilde{i}, j\right) y_j$$

On peut ainsi considérer une moyenne pondérée d'un nombre donné n, à choisir, de voisins les plus proches. *n nearest neighbours*. L'estimateur proposé par Rubin est en fait celui du voisin le plus proche. Considérer un nombre plus important de voisins affecte l'erreur

quadratique moyenne de l'estimateur, elle même somme du carré du biais et de la variance de l'estimateur. Lorsque le nombre d'individus considéré augmente le biais augmente : on prend en compte des individus dont les caractéristiques sont plus éloignées que celle de l'individu traité. En revanche la variance baisse car on prend la moyenne sur un ensemble plus important d'individus. On peut montrer que le nombre optimal d'individus à prendre en compte croît avec la taille de l'échantillon.

L'estimateur proposé par Heckmann Ichimura and Todd (1998) est un estimateur à noyau de la quantité $E\left(y_0 \left| T=1, x=x_{\widetilde{i}}\right.\right)$.

$$\widehat{E}\left(y_0 \left| T=1, x=x_{\widetilde{i}}\right.\right) = \frac{\sum\limits_{T_j=0} K_h\left(x_j - x_{\widetilde{i}}\right) y_j}{\sum\limits_{T_j=0} K_h\left(x_j - x_{\widetilde{i}}\right)} = \sum_{T_j=0} \frac{K_h\left(x_j - x_{\widetilde{i}}\right)}{\sum\limits_{T_j=0} K_h\left(x_j - x_{\widetilde{i}}\right)} y_j = \sum_{T_j=0} w_N\left(j, \widetilde{i}\right) y_j$$

dans cette expression $K_h(z) = \frac{1}{h} K\left(\frac{z}{h}\right)$ ou K est un noyau et h un paramètre appelé la fenêtre. Le noyau est une fonction maximale en zéro, positive en zéro, symétrique autour de zéro et d'intégrale unitaire (cette condition ne joue pas de rôle dans le cas de l'estimation d'une fonction de régression). Il existe de multiples exemples de noyau, par exemple le noyau uniforme valant 0.5 sur $[-1, 1]$, Dans ce cas l'estimateur non paramétrique correspondant consiste simplement à prendre la moyenne des observations pour des individus dont les caractéristiques se situent dans l'intervalle $[x - h_N, x + h_N]$. Un autre exemple correspond à $\phi(z)$ la densité de la loi normale. Ce noyau présente l'avantage d'avoir $\Re$ pour support Un noyau fréquemment choisi en pratique dans le cas unidimensionnel est le noyau quartique : $K(z) = \frac{15}{16}\left(1 - z^2\right)^2 1\{|z| \leq 1\}$

Dans les expressions précédentes, h est la *fenêtre*. Plus elle est faible, moins on prend en compte les observations s'éloignant de $x_{\widetilde{i}}$. Dans ce cas l'estimateur sera très peu précis mais le biais sera en revanche faible. A l'inverse, lorsque la fenêtre s'élargit l'estimateur considéré devient plus précis autour de sa valeur limite, mais cette valeur limite tend elle même à s'écarter de la quantité que l'on cherche à estimer. Le choix de la fenêtre est tel qu'il minimise l'erreur quadratique moyenne, somme du carré du biais et de la variance de l'estimateur. On peut montrer que lorsque elle est choisie comme une fonction croissante de la dispersion des variables x et décroissante du nombre d'individu. Un choix possible pour la fenêtre est dans le cas unidimensionnel : $h(N) = \sigma_x / N^{1/5}$. En général les estimateurs non paramétriques ont une vitesse de convergence plus faible que les estimateurs paramétriques. Ici le rythme de convergence est en $\sqrt{Nh}$ soit une vitesse de convergence en $N^{\frac{2}{5}}$.

Finalement l'estimateur de l'effet moyen du traitement sur les traités est estimé par :

$$\widehat{E}\left(\Delta \left| T=1\right.\right) = \frac{1}{N\{T_i = 1\}} \sum_{\{T_i=1\}} \left(y_i - \sum_{T_j=0} w_N(j, i)\, y_j \right)$$

Bien que basé sur des estimateurs non paramétriques qui donc convergent lentement, Heckman Ichimura et Todd ont montré que la vitesse de convergence de cet estimateur est en $\sqrt{N}$. Ceci tient au fait que l'estimateur final est une moyenne d'estimateurs non paramétriques. Il est dit semi -paramétrique. L'expression de la variance de cet estimateur est complexe et son estimation à partir de sa formule littérale nécessite là aussi le calcul d'intermédiaires non paramétrique. En pratique, on détermine la variance de cet estimateur par bootstrap. Ceci consiste à tirer avec remise un grand nombre d'échantillons aléatoires dans la population, et à appliquer sur chacun de ces échantillons toute la procédure d'estimation. La distribution des estimateurs que l'on obtient in fine est la distribution exacte de l'estimateur. On peut l'utiliser pour déterminer les écarts-type ou les intervalles de confiance.

Là aussi la propriété de Rubin est très importante. En effet elle autorise à procéder à la régression non paramétrique sur la seule variable que constitue le score $s(x)$. On est ainsi amené à calculer pour chaque individu : $\widehat{E}(y_0 | T=1, s(x)=s(x_{\tilde{i}}))$ et non plus $\widehat{E}(y_0 | T=1, x=x_{\tilde{i}})$. Cette simplification ne remet pas en cause la validité de l'estimateur alternatif basé sur l'appariement sur chacune des caractéristiques. La vitesse de convergence n'est pas plus élevée avec l'un qu'avec l'autre estimateur. Néanmoins le nombre d'observations nécessaires pour que ce comportement asymptotique soit obtenu est vraisemblablement plus faible avec l'appariement sur le score. Cet estimateur apparaît plus fiable à ce titre.

Remarque *: Les résultats précédents peuvent être appliqués en sens inverse pour appariés chaque individu non traité avec un (des) individus traités. On estime alors* $E(\Delta | T=0, x_i \in X)$. *On peut donc par appariement estimer l'effet moyen du traitement.*

Pondérations

Une dernière méthode d'estimation est basée sur des pondérations.

Proposition *Sous l'hypothèse d'indépendance faible conditionnelle aux observables, l'effet moyen du traitement vérifie la relation*

$$E(c) = E\left(y\left(\frac{T}{P(x)} - \frac{(1-T)}{(1-P(x))}\right)\right)$$

Sous l'hypothèse d'indépendance faible conditionnelle aux observables, l'effet du traitement sur les traités vérifie la relation

$$E(c | T=1) = E\left(y\frac{P(x)}{P(T=1)}\left(\frac{T}{P(x)} - \frac{(1-T)}{(1-P(x))}\right)\right)$$

Démonstration *En effet, les propriétés d'indépendance conditionnelles permettent d'identifier très simplement les espérances des outputs potentiels.*

$$y_k \perp T | x \Longrightarrow E(y_k 1(T=k) | x) = E(y_k | x) E((1(T=k)) | x) = E(y_k | x) P(T=k | x)$$

On a donc :

$$E\left(y_k\left|x\right.\right)=E\left(y_k\frac{1\left(T=k\right)}{P\left(T=k\left|x\right.\right)}\left|x\right.\right)$$

D'où la première relation. Par ailleurs on a

$$\begin{aligned}E\left(y_0T\left|x\right.\right) &= P\left(x\right)E\left(y_0\left|T=1,x\right.\right)=P\left(x\right)E\left(y_0\left|T=1,x\right.\right)E\left(\frac{1-T}{1-P\left(x\right)}\left|x\right.\right)\\ &= E\left(P\left(x\right)E\left(y_0\left|T=1,x\right.\right)\frac{1-T}{1-P\left(x\right)}\left|x\right.\right)\end{aligned}$$

D'où

$$E\left(y_0T\right)=E\left(y_0\left|T=1\right.\right)P\left(T=1\right)=E\left(P\left(x\right)E\left(y_0\left|T=1,x\right.\right)\frac{1-T}{1-P\left(x\right)}\right)$$

Comme $E\left(y_0\left|T=1,x\right.\right)=E\left(y_0\left|T=0,x\right.\right)$

$$\begin{aligned}E\left(y_0\left|T=1\right.\right) &= E\left(P\left(x\right)E\left(y_0\left|T=1,x\right.\right)\frac{1-T}{1-P\left(x\right)}\right)\Big/P\left(T=1\right)\\ &= E\left(P\left(x\right)E\left(y_0\left|T=0,x\right.\right)\frac{1-T}{1-P\left(x\right)}\right)\Big/P\left(T=1\right)\\ &= E\left(E\left(P\left(x\right)y_0\frac{1-T}{1-P\left(x\right)}\left|T=0,x\right.\right)\right)\Big/P\left(T=1\right)\\ &= E\left(P\left(x\right)y_0\frac{1-T}{1-P\left(x\right)}\right)\Big/P\left(T=1\right)\end{aligned}$$

13.3.4 Vraisemblance de l'hypothèse d'indépendance conditionnelle à des observables.

Plusieurs questions se posent concernant la méthode par appariement. La première concerne de savoir s'il est raisonnable de faire l'hypothèse d'indépendance conditionnelle à des observables. La deuxième est comment choisir en pratique les variables de conditionnement ? Faut-il retenir toute l'information à disposition ? On présente d'abord un résultat permettant de répondre en partie à ces questions :

Proposition $z_1\perp z_2\left|w_1,w_2\right.$ *et* $w_2\perp z_2\left|w_1\right.\Longrightarrow z_1\perp z_2\left|w_1\right.$

Démonstration *En effet :*

$$l\left(z_1,z_2\left|w_1\right.\right)=\int l\left(z_1,z_2\left|w_1,w_2\right.\right)l\left(w_2\left|w_1\right.\right)dw_2$$

en outre : $l\left(z_1,z_2\left|w_1,w_2\right.\right)=l\left(z_1\left|w_1,w_2\right.\right)l\left(z_2\left|w_1,w_2\right.\right)=l\left(z_1\left|w_1,w_2\right.\right)l\left(z_2\left|w_1\right.\right)$, *d'où :*

$$\begin{aligned}l\left(z_1,z_2\left|w_1\right.\right) &= \int l\left(z_1\left|w_1,w_2\right.\right)l\left(z_2\left|w_1\right.\right)l\left(w_2\left|w_1\right.\right)dw_2=l\left(z_2\left|w_1\right.\right)\int l\left(z_1\left|w_1,w_2\right.\right)l\left(w_2\left|w_1\right.\right)dw_2\\ &= l\left(z_2\left|w_1\right.\right)l\left(z_1\left|w_1\right.\right)\end{aligned}$$

Prise en compte d'effets individuels : l'apport de données temporelles

L'hypothèse d'indépendance conditionnelle à des observables a en fait peu de chance d'être satisfaite dès lors que les variables sont en niveau. Il y a en effet une hétérogénéité très forte dans les situations individuelles. Il est peu vraisemblable que l'on puisse par adjonction de variable de contrôle épuiser toute la partie de cette hétérogénéité qui est prise en compte dans la décision de participation. La majeur partie de cette hétérogénéité correspond à la présence de caractéristiques inobservées permanentes dans le temps semblable à un effet individuel. Les résultats dont on dispose en économétrie des données de panel montrent bien que premièrement, les effets individuels ont une très forte variance, même dans les modèles dans lesquels on a cherché à introduire de nombreux contrôles et que deuxièmement l'hypothèse d'indépendance entre les variables explicatives et les effets individuels est très fréquemment rejetée. Une hypothèse plus vraisemblable consisterait à introduire dans les variables de conditionnement un terme d'hétérogénéité constant dans le temps :

$$\begin{aligned} H_{Forte} &: y_0, y_1 \perp T \,|x, u \\ H_{Faible} &: y_0 \perp T \,|x, u \end{aligned}$$

Prendre en compte cette hétérogénéité dans le cadre précédent n'est pas directement possible justement parce qu'elle est inobservable.

Néanmoins, à l'instar de ce qui est effectué dans le cadre de l'économétrie des données de panel, elle peut être éliminée par différentiation. Plus précisément, prenant par exemple le cas de l'indépendance faible, on a la proposition suivante qui découle directement de la proposition précédente :

Proposition *Dans le cas où il existe un élément inobservé u tel que la condition*

$$y_0 \perp T \,|x, u$$

est vérifiée. Si :

1. *Il existe des observations disponibles y^p de l'output antérieures au traitement*
2. *$y_0 - y^p \perp T \,|x, u$, ce qui est vrai dès lors que $y^p \in \{x\}$ dans la condition $y_0 \perp T \,|x, u$*
3. *$y_0 - y^p \perp u \,|x$,*

alors la condition d'indépendance,

$$y_0 - y^p \perp T \,|x$$

est vérifiée

On voit que dans ce cas l'effet individuel peut être éliminé par différentiation et on retrouve une propriété d'indépendance conditionnelle à des observables. En pratique, ceci revient à introduire les variables passées de l'output dans la liste des variables de conditionnement et à considérer comme variable d'output non les outputs eux mêmes, mais leur évolutions. Notant $\Delta y_1 = y_1 - y^p$ et $\Delta y_0 = y_0 - y^p$, on estime

$$E\left(\Delta y_1 - \Delta y_0 \,|T = 1, x\right) = E\left((y_1 - y^p) - (y_0 - y^p) \,|T = 1, x\right) = E\left(y_1 - y_0 \,|T = 1, x\right)$$

qui est donc bien le paramètre cherché.

Sélection des observables

On peut être tenté de considérer un grand nombre de variables de conditionnement. Ceci n'est pas nécessairement une bonne propriété comme on le verra et il vaut mieux chercher l'ensemble de variables de conditionnement le plus petit possible tel que la condition d'indépendance soit satisfaite.

Proposition *Supposons*

$$y_0, y_1 \perp T \,|x_1, x_2$$

Si seule une partie de ces variables affecte la variable de traitement :

$$T \perp x_2 \,|x_1$$

Alors on a

$$y_0, y_1 \perp T \,|x_1$$

La liste des variables de conditionnement peut être amputée de toutes les variables qui n'affectent pas la variable de traitement, ce qui peut être aisément testé sur les données.

Problème de support

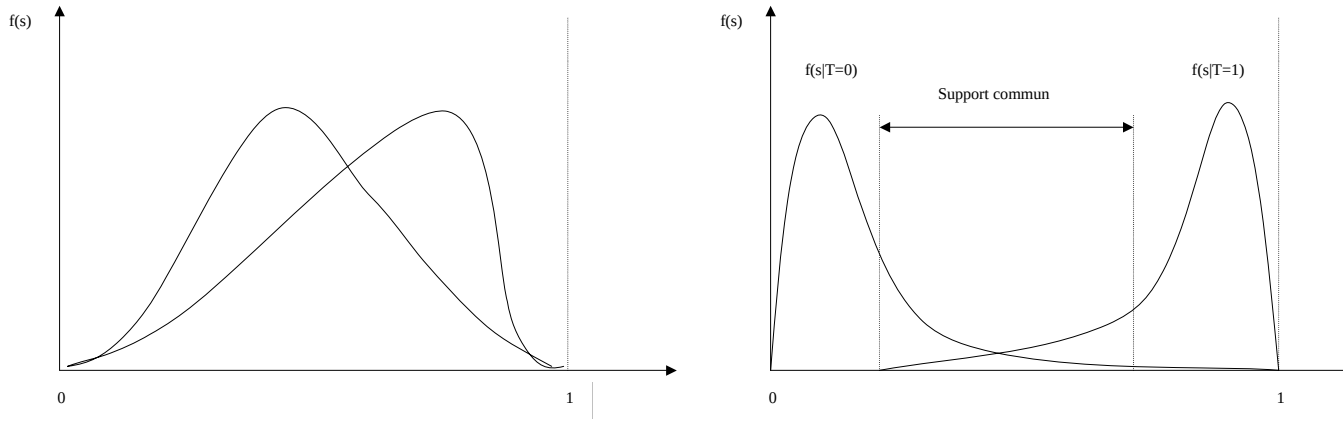
La question du support des distributions du score conditionnellement au traitement est essentielle dans ce type d'analyse. Son importance a été soulignée par Heckman et al. (1998) qui ont montré qu'elle constitue une source forte de biais dans l'estimation de l'effet causal du traitement.

Dans les méthodes d'estimation par appariement ou par régression, il est nécessaire de pouvoir construire pour chaque individu traité un contrefactuel à partir des individus non traités, c'est-à-dire de pouvoir estimer $E\left(y \,|s, T = 0\right)$ pour déterminer l'effet causal du traitement sur la population des individus traités. En outre, il est nécessaire d'estimer $E\left(y \,|s, T = 1\right)$ dès qu'on s'intéresse à l'effet causal du traitement dans la population totale.

Une estimation non paramétrique de cette quantité, donc sans restriction sur la forme qu'elle prend, impose que l'on dispose pour un individu traité de score s d'individus non traités ayant des valeurs du score proche de s. Dit d'une autre manière, la densité du

score pour les individus non traités ne doit pas être nulles pour les valeurs du score des individus traités considérés. On ne peut donc construire de contrefactuel que pour les individus dont le score appartient à l'intersection des supports de la distribution du score des individus traités et des individus non traités.

Ceci conduit à la conclusion que même sous l'hypothèse d'indépendance conditionnelle à des observables, on ne peut pas systématiquement estimer $E(\Delta)$ ou $E(\Delta|T=1)$ dans la mesure où $E(\Delta|s)$ ne peut être estimé que pour les individus dont le score appartient au support commun de la distribution du score pour les individus traités et non traités. L'estimateur obtenu in fine est alors un estimateur local : $E(c|s \in S_\cap)$ ou $E(c|s \in S_\cap, T=1)$, avec $S_\cap$ le support commun défini par $S_\cap = S_{T=1} \cap S_{T=0}$ avec $S_{T=1}$ le support de la distribution du score des individus traités et $S_{T=0}$ celui des individus non traités.



Cette condition du support a une autre implication : le modèle servant à la construction du score, c'est à dire expliquant le traitement à partir des variables de conditionnement, ne doit pas être trop bon. Dans le cas extrême où on expliquerait parfaitement le traitement, les densités du score conditionnellement au traitement seraient toutes deux des masses de Dirac, l'une en zéro pour les individus non traités, l'autre en 1 pour les individus traités. Les supports seraient alors disjoints et aucun appariement ne serait possible.

Pour bien comprendre cette condition importante du score, il faut garder présente à l'esprit l'idée initiale de Rubin : conditionnellement à un ensemble de variables explicatives x (ou le score), on se trouve dans le cas d'une expérience contrôlée, c'est à dire dans laquelle on dispose d'individus traités et non traités qui sont affectés aléatoirement à chacun de ces groupes. Il faut dans chaque cellule dans laquelle on se trouve dans des conditions d'expérience qu'il y ait un fort aléa sur l'affectation au traitement. La persistance de cette composante aléatoire de l'affectation au traitement conditionnellement à des observables est ainsi essentielle dans la procédure d'appariement.

Remarque *Il peut être utile d'utiliser des restrictions a priori. Les modèles précédents sont en effet purement statistiques. Fréquemment on a une idée de modélisation de la*

variable d'output à partir d'un ensemble de variables explicatives.

$$y_0 = r\beta + u_0 \text{ avec } r \perp u_0$$

Le problème d'endogénéité provient alors du fait que la variable de traitement est corrélée à la perturbation conditionnellement à r*. On peut supposer que la propriété d'indépendance est vraie lorsque l'on adjoint un ensemble de variables* z *à* r*.*

$$u_0 \perp T \,|r, z$$

On fait l'hypothèse $r \perp T \,|z$ *, ce qui revient à supposer* $P(T = 1 \,|r, z) = P(T = 1 \,|z) = P(z)$*. En outre on étend la condition d'indépendance :* $r \perp u_0$ *à* $r \perp u_0 \,|z$ *. Dans ces conditions on a le résultat suivant*

$$E(y_0 \,|T, r, P(z)) = r\beta + h(P(z))$$

En effet

$$E(y_0 \,|T, r, P(z)) = r\beta + E(u_0 \,|T, r, P(z))$$

Comme $u_0 \perp T \,|r, z$ *on a en raison de la propriété de Rubin et Rosenbaum* $u_0 \perp T \,|P(T = 1 \,|r, z)$*. et* $P(T = 1 \,|r, z) = P(T = 1 \,|z)$*. On a donc :*

$$u_0 \perp T \,|P(z), r$$

d'où

$$E(u_0 \,|T, r, P(z)) = E(u_0 \,|r, P(z)) = E(u_0 \,|P(z))$$

où la dernière égalité provient du fait que $r \perp u_0 \,|z$ *. On peut donc transposer tous les estimateurs précédents au cas présent. On peut en particulier procéder comme suit. On estime d'abord le paramètre* β*. Pour cela on remarque que comme :*

$$E(y \,|T = 0, r, P(z)) = r\beta + h(P(z))$$

on a

$$E(y \,|T = 0, P(z)) = E(r \,|T = 0, P(z))\,\beta + h(P(z))$$

En prenant la différence des deux équations, on en déduit :

$$E(y - E(y \,|T = 0, P(z)) \,|T = 0, r, P(z)) = (r - E(r \,|T = 0, P(z)))\,\beta$$

Dont on déduit que

$$E(y - E(y \,|T = 0, P(z)) \,|T = 0, r) = (r - E(r \,|T = 0, P(z)))\,\beta$$

Ce qui signifie qu'on peut estimer β *en régressant simplement les résidus des régressions non paramétriques* $y - \widehat{E}(y \,|T = 0, P(z))$ *et* $r - \widehat{E}(r \,|T = 0, P(z))$ *l'un sur l'autre. La fonction* h *peut alors être estimée à partir* $y - r\beta$*. En effet :*

$$E(y - r\beta \,|T = 0, r, P(z)) = h(P(z)) = E(y - r\beta \,|T = 0, P(z))$$

Le contrefactuel pour un individu traité i de caractéristiques r_i et P_i est alors

$$\widehat{E}\left(y_{0i}\left|T=1,r_i,P_i\right.\right)=r_i\widehat{\beta}+\frac{1}{N_0}\sum_{T_j=0}\left(y_j-r_j\widehat{\beta}\right)\frac{K_h\left(P_j-P_i\right)}{\sum\limits_{T_j=0}K_h\left(P_j-P_i\right)}$$

et l'estimateur de l'effet du traitement est alors

$$\widehat{\Delta}^{TT}=\frac{1}{N_1}\sum_{T_i=1}\left[y_i-r_i\widehat{\beta}-\frac{1}{N_0}\sum_{T_j=0}\left(y_j-r_j\widehat{\beta}\right)\frac{K_h\left(P_j-P_i\right)}{\sum\limits_{T_j=0}K_h\left(P_j-P_i\right)}\right]$$

13.4 Le modèle de sélectivité sur inobservables

L'approche précédente présente des attraits non négligeables. Le premier est qu'elle est assez naturelle : on compare des individus traités et non traités aussi similaires que possible. Le second avantage est qu'elle ne nécessite pas la modélisation du comportement des agents. En revanche, elle présente des limites certaines. Ainsi elle n'est pas toujours réalisable. L'obtention de la condition d'indépendance peut requérir l'introduction d'un grand nombre de variables de conditionnement qui ne sont pas toujours accessibles d′une part et réduisent aussi la pertinence de l'analyse dans la mesure où les possibilités de comparaison d'un individu à l'autre se réduisent lorsque l'on explique de mieux en mieux l'affectation au traitement, i.e. lorsque croît le nombre de variables de conditionnement. Enfin et surtout, les méthodes d'appariement sur observables présentent un caractère mécanique qui fait reposer l'évaluation sur une propriété purement statistique, en pratique difficile à justifier à partir du comportement des agents. Dans une certaine mesure l'intérêt que présente le fait de ne pas modéliser les comportements comporte aussi un revers qui est celui de conduire à des évaluations dont les fondements peuvent paraître peu étayés. Il peut être préférable de modéliser les output potentiel et la décision de participation de façon jointe. On parvient alors au modèle de sélectivité sur inobservable. On l'écrit sous la forme suivante. Les deux outputs potentiels $y_1$et y_0 sont modélisés sous la forme :

$$\begin{aligned} y_1 &= \alpha_1+r\beta_1+u_1 \\ y_0 &= \alpha_0+r\beta_0+u_0 \end{aligned}$$

On modélise également l'affectation au traitement par le biais d'une variable latente, T^* :

$$\begin{aligned} T^* &= zc+v \\ T &= 1\Longleftrightarrow T^*\geq 0 \end{aligned}$$

T^* peut représenter par exemple le gain net du coût du traitement $c\left(z,\eta\right)+v$: $T^*=y_1-y_0-c\left(z,\eta\right)-v$

La principale hypothèse identifiante effectuée consiste à supposer l'indépendance entre les variables de conditionnement et les éléments inobservés.

$$(u_1, u_0, v) \perp (x, z)$$

Definition *Le modèle de sélectivité sur inobservable est défini par la modélisation jointe des outputs potentiels et de l'affectation au traitement*

$$\begin{aligned} y_1 &= \alpha_1 + r\beta_1 + u_1 \\ y_0 &= \alpha_0 + r\beta_0 + u_0 \\ T &= 1 \Longleftrightarrow zc + v \geq 0 \end{aligned}$$

avec en outre l'hypothèse d'indépendance

$$(u_1, u_0, v) \perp (r, z)$$

Remarque *Ces hypothèse sont très différentes de celle du modèle de sélectivité sur observables. Dans le modèle de sélectivité sur observables, on faisait l'hypothèse que la corrélation entre la variable de traitement T et les éléments inobservés u_0 pouvait être éliminée par en introduisant des variables de conditionnement supplémentaires. Ces variables étaient par définition des variables affectant à la fois le traitement et la perturbation. L'hypothèse est ici diamétralement opposée dans la mesure ou elle consiste à dire qu'à l'inverse il existe une variable z affectant le traitement mais pas les éléments inobservés. Elle est donc très proche d'une variable instrumentale, alors que dans l'approche précédente il s'agissait de variable de contrôle.*

Dans cette approche, le score $P(T = 1 | r, z)$ est encore amené à joué un rôle central. Sous les hypothèses effectuées le score ne dépend que des variables z. En effet

$$P(T = 1 | r, z) = P(zc + v > 0 | r, z) = P(zc + v > 0 | z) = P(z)$$

Toutefois, ces hypothèses ne sont pas suffisantes pour assurer l'identification des paramètres d'intérêt et il existe en fait une différence importante avec les variables instrumentales, sur laquelle on reviendra plus tard. Les paramètres d'intérêt sont définis par :

$$\begin{aligned} \Delta^{ATE} &= E(y_1 - y_0) = E(\alpha_1 - \alpha_0 + r(\beta_1 - \beta_0)) \\ \Delta^{TT} &= E(y_1 - y_0 | T = 1) = E(y_1 - (\alpha_0 + r\beta_0 + u_0) | T = 1) \end{aligned}$$

13.4.1 Expression des paramètres d'intérêt dans le cas général

Proposition *Dans le cas du modèle de sélectivité sur inobservables, si les fonctions de répartition de v est strictement croissante, il existe deux fonctions $K_0(P(zc))$ et*

$K_1(P(zc))$ telles que

$$\begin{aligned} E(y_0 \,|T=0,r,z) &= \alpha_0 + r\beta_0 + K_0(P(zc)) \\ E(y_1 \,|T=1,r,z) &= \alpha_1 + r\beta_1 + K_1(P(zc)) \end{aligned}$$

Les paramètres d'intérêt sont alors définis par

$$\begin{aligned} \Delta^{TT} &= E\left(y - \left(\alpha_0 + r\beta_0 - \frac{1-P(z)}{P(z)}K_0(P(zc))\right)\middle| T=1\right) \\ \Delta^{ATE} &= E(\alpha_1 - \alpha_0 + r(\beta_1 - \beta_0)) \end{aligned}$$

où

$$P(zc) = P(T=1\,|r,z)$$

Démonstration *La forme des fonctions retenues est une application directe du modèle de sélection sur inobservables vu précédemment. Pour ce qui concerne le paramètre Δ^{TT}, l'identification porte donc essentiellement sur l'output potentiel y_0. Les données sur cet output concernent les individus pour lesquels $T=0$. On a :*

$$E(y_0 \,|T=0,r,z) = \alpha_0 + r\beta_0 + E(u_0 \,|T=0,r,z) = \alpha_0 + r\beta_0 + K_0(P(zc))$$

et on souhaite identifier

$$E(y_0 \,|T=1,r,z) = \alpha_0 + r\beta_0 + E(u_0 \,|T=1,r,z)$$

Les quantités $E(u_0 \,|T=0,r,z)$ et $E(u_0 \,|T=1,r,z)$ sont liées par :

$$0 = E(u_0 \,|r,z) = E(u_0 \,|T=0,r,z)(1-P(zc)) + E(u_0 \,|T=1,r,z)\,P(zc)$$

d'où

$$E(u_0 \,|T=1,r,z) = -\frac{(1-P(zc))}{P(zc)}K_0(P(zc))$$

En toute généralité on ne peut donner la forme des fonctions K_0 et K_1. Elle font en effet intervenir la loi jointe des élément (u_0, v) et (u_1, v). Ceci est à l'origine d'un problème important pour l'estimation puisque comme les expressions précédentes le montrent clairement, il est nécessaire de pouvoir séparer les fonctions K des constantes α.

On va voir d'abord comment il est possible de résoudre ce problème en spécifiant la loi jointes des observations. Puis on examinera le cas dans lequel on ne fait pas d'hypothèse et on verra qu'il faut des conditions particulières et au total assez restrictives pour identifier chacun des deux paramètres d'intérêt.

13.4.2 Le cas Normal

La spécification de la loi jointe des observations comme des lois normales permet d'identifier aisément le modèle. On peut soit recourir à la méthode du maximum de vraisemblance soit recourir à une méthode en deux étapes due à l'origine à Heckman, basée sur les résultats précédents. C'est cette dernière méthode que l'on présente car elle est d'un emploi plus facile et est directement liée à la présentation précédente. Elle présente en outre un degrés de généralité légèrement supérieure. On reprend le modèle d'outputs potentiels précédents :

$$\begin{aligned} y_1 &= \alpha_1 + r\beta_1 + u_1 \\ y_0 &= \alpha_0 + r\beta_0 + u_0 \end{aligned}$$

avec la règle d'affectation au traitement basée sur la variable latente, T^* :

$$\begin{aligned} T^* &= zc + v \\ T &= 1 \Longleftrightarrow T^* \geq 0 \end{aligned}$$

Outre l'hypothèse d'indépendance déjà évoquée, on fait l'hypothèse que les deux couples (u_0, v) et (u_1, v) suivent une loi normale.

Les résultats précédents permettent d'écrire que :

$$\begin{aligned} E\left(y_0 \left| r, z, T = 0\right.\right) &= \alpha_0 + r\beta_0 - \rho_0\sigma_0 \frac{\phi}{1-\Phi}\left(zc\right) \\ E\left(y_1 \left| r, z, T = 1\right.\right) &= \alpha_1 + r\beta_1 + \rho_1\sigma_1 \frac{\phi}{\Phi}\left(zc\right) \end{aligned}$$

Par rapport aux expressions obtenues dans le cas général

$$E\left(y_0 \left| T = 0, r, z\right.\right) = \alpha_0 + r\beta_0 + K_0\left(P\left(zc\right)\right)$$

et compte tenu du fait que $P\left(zc\right) = \Phi\left(z\widetilde{c}\right)$, on voit que le fait de spécifier la loi des observations comme une loi normale revient à imposer que les fonctions $K_0\left(P\left(zc\right)\right)$ et $K_1\left(P\left(zc\right)\right)$ ont pour expressions :

$$\begin{aligned} K_0\left(P\left(zc\right)\right) &= -\rho_0\sigma_0 \frac{\phi \circ \Phi^{-1}\left(P\left(zc\right)\right)}{1 - P\left(zc\right)} \\ K_0\left(P\left(z\right)\right) &= \rho_1\sigma_1 \frac{\phi \circ \Phi^{-1}\left(P\left(zc\right)\right)}{P\left(zc\right)} \end{aligned}$$

Elle ne dépend donc que d'un paramètre supplémentaire $\rho_0\sigma_0$. Les paramètres d'intérêt Δ^{TT} et Δ^{ATE} ont alors pour expressions :

$$\begin{aligned}
\Delta^{TT} &= E\left(y - \left(\alpha_0 + r\beta_0 - \frac{1-P(z)}{P(z)}K_0\left(P(zc)\right)\right)\middle| T=1\right) \\
&= E\left(y - \left(\alpha_0 + r\beta_0 + \rho_0\sigma_0\frac{\phi\circ\Phi^{-1}\left(P(zc)\right)}{P(zc)}\right)\middle| T=1\right) \\
&= E\left(y - \left(\alpha_0 + r\beta_0 + \rho_0\sigma_0\frac{\phi}{\Phi}\left(z\widetilde{c}\right)\right)\middle| T=1\right) \\
\Delta^{ATE} &= E\left(\alpha_1 - \alpha_0 + r\left(\beta_1 - \beta_0\right)\right)
\end{aligned}$$

Mise en oeuvre :

1. Estimation du modèle probit associé au traitement et détermination des variables de biais $\frac{\phi}{\Phi}(zc)$ et $\frac{\phi}{1-\Phi}(zc)$
2. Estimation des régressions sur chacune des populations traitées et non traitées : identification des paramètres $\alpha_1, \alpha_0, \beta_1, \beta_0$ et des paramètres $\rho_1\sigma_{u_1}$ et $\rho_0\sigma_{u0}$.
3. Estimation des paramètres d'intérêt

$$\begin{aligned}
\widehat{\Delta}^{TT} &= \frac{1}{N_1}\sum_{d_i=1}\left(y_i - \left(\widehat{\alpha}_0 + r_i\widehat{\beta}_0 + \widehat{\rho_0\sigma_0}\frac{\phi}{\Phi}\left(z_i\widehat{c}\right)\right)\right) \\
\widehat{\Delta}^{ATE} &= \frac{1}{N}\sum\left(\widehat{\alpha}_1 - \widehat{\alpha}_0 + r_i\left(\widehat{\beta}_1 - \widehat{\beta}_0\right)\right)
\end{aligned}$$

4. Calcul des écarts-type, on doit prendre en compte le fait que le paramètres du modèle Probit a été estimé dans une première étape.

13.4.3 Des extensions paramétriques simples

Comme dans le cas du modèle de sélection du chapitre précédent, on peut étendre d'abord les résultats obtenus avec la loi normale à des familles de lois plus générales.

Loi quelconque donnée pour le résidu de l'équation de sélection.

On a vu dans le chapitre précédent que le modèle de sélection pouvait être facilement étendu en considérant une loi quelconque pour l'équation de sélection. Elle donne alors lieu à une probabilité de sélection notée $P(z)$

$$E\left(y\,|I=1,x,z\right) = xb + \rho\sigma_u\frac{\phi\circ\Phi^{-1}P(z)}{P(z)}$$

Ce résultats se transposent directement au cas du modèle causal. Les équations des outputs potentiels sont :

$$\begin{aligned} P(T=1|z) &= P(z) \\ E(y_0|T=0,r,z) &= \alpha_0 + r\beta_0 - \rho_0\sigma_0 \frac{\phi \circ \Phi^{-1} P(z)}{1-P(z)} \\ E(y_1|T=1,r,z) &= \alpha_1 + r\beta_1 + \rho_1\sigma_1 \frac{\phi \circ \Phi^{-1} P(z)}{P(z)} \end{aligned}$$

Les paramètres d'intérêt ont alors pour expression :

$$\begin{aligned} \Delta^{TT} &= E\left(y - \left(\alpha_0 + r\beta_0 + \rho_0\sigma_0 \frac{\phi \circ \Phi^{-1}(P(z))}{P(z)} \right) \middle| T=1 \right) \\ \Delta^{ATE} &= E(\alpha_1 - \alpha_0 + r(\beta_1 - \beta_0)) \end{aligned}$$

Des lois plus générales que la loi normale

On peut considérer le modèle de sélection précédent en faisant l'hypothèse que les éléments inobservés ont pour loi jointe une loi de Student de degrés η et non pas une loi normale. On a vu dans le chapitre précédent que ceci conduisait à la spécification suivante pour l'équation d'output :

$$E(y|d=1,x,z) = xb + \rho\sigma \frac{\eta + G_\eta^{-1}(P(z))^2}{\eta - 1} \frac{g_\eta \circ G_\eta^{-1}(P(z))}{P(z)}$$

Là aussi les résultats se transposent directement au cas du modèle causal. Les équations des outputs potentiels sont :

$$\begin{aligned} P(T=1|z) &= P(z) \\ E(y_0|T=0,r,z) &= \alpha_0 + r\beta_0 - \rho_0\sigma_0 \frac{\eta + G_\eta^{-1}(P(z))^2}{\eta - 1} \frac{g_\eta \circ G_\eta^{-1}(P(z))}{1-P(z)} \\ E(y_1|T=1,r,z) &= \alpha_1 + r\beta_1 + \rho_1\sigma_1 \frac{\eta + G_\eta^{-1}(P(z))^2}{\eta - 1} \frac{g_\eta \circ G_\eta^{-1}(P(z))}{P(z)} \end{aligned}$$

Les paramètres d'intérêt ont alors pour expression :

$$\begin{aligned} \Delta^{TT} &= E\left(y - \left(\alpha_0 + r\beta_0 + \rho_0\sigma_0 \frac{\eta + G_\eta^{-1}(P(z))^2}{\eta - 1} \frac{g_\eta \circ G_\eta^{-1}(P(z))}{P(z)} \right) \middle| T=1 \right) \\ \Delta^{ATE} &= E(\alpha_1 - \alpha_0 + r(\beta_1 - \beta_0)) \end{aligned}$$

On dispose ainsi d'un ensemble très vaste de possibilités d'estimation des paramètres correspondant à différentes hypothèses sur la loi des perturbations. Ces choix reviennent

tous à introduire des termes différents dans les équations des outputs potentiels. Ils ont des conséquences importantes sur l'estimation des paramètres d'intérêt. Il est en outre difficile de réaliser des tests permettant d'examiner quelle spécification est préférable dans la mesure où les hypothèses ne sont pas emboîtées. On peut donc souhaiter estimer ces modèles sans avoir recours à la spécification de la loi jointe des perturbations.

13.4.4 Le modèle de sélection semi paramétrique.

On reprend le modèle de sélectivité sur inobservables :

$$\begin{aligned} y_1 &= \alpha_1 + r\beta_1 + u_1 \\ y_0 &= \alpha_0 + r\beta_0 + u_0 \end{aligned}$$

avec la modélisation de l'affectation au traitement :

$$\begin{aligned} T^* &= zc + v \\ T &= 1 \Longleftrightarrow T^* \geq 0 \end{aligned}$$

on suppose comme précédemment l'indépendance entre les variables de conditionnement et les éléments inobservés.

$$(u_1, u_0, v) \perp (x, z)$$

On a vu qu'en l'absence d'hypothèses sur la loi jointe des perturbations, les équations des outputs potentiels prenaient la forme :

$$\begin{aligned} E(y_0 | T = 0, r, z) &= \alpha_0 + r\beta_0 + K_0(P(z)) \\ E(y_1 | T = 1, r, z) &= \alpha_1 + r\beta_1 + K_1(P(z)) \end{aligned}$$

avec K_0 et K_1 des fonctions non spécifiées. Les paramètres d'intérêt s'écrivent simplement comme :

$$\begin{aligned} \Delta^{TT} &= E\left(y - \left(\alpha_0 + r\beta_0 - \frac{1 - P(z)}{P(z)} K_0(P(z))\right) \middle| T = 1\right) \\ \Delta^{ATE} &= E(\alpha_1 - \alpha_0 + r(\beta_1 - \beta_0)) \end{aligned}$$

La difficulté de l'estimation est double. D'une part il est nécessaire d'estimer les paramètres α et β en laissant la fonction K non spécifiée. En deuxième lieu il faut estimer la fonction K elle même. On procède en plusieurs étapes. Dans un premier temps, on estime le paramètre β. Dans un deuxième temps, on estime la fonction $G = \alpha + K$. Enfin dans un dernier temps on sépare α de K.

Identification des paramètres β

Pour les paramètres β_0 et β_1, on applique la méthode d'estimation de Robinson vue dans le chapitre précédent. Ceci consiste à prendre rappelons comme dans le théorème de Frish-Waugh, l'écart des variables y et r à leur espérance conditionnellement au score (la différence avec le théorème de Frish-Waugh est qu'il ne s'agit plus d'une simple projection linéaire). Il suffit ensuite de régresser le résidus obtenu pour y sur ceux obtenus pour les variables r.

Identification des constantes et des termes de biais de sélectivité K_0 et K_1.

Dans un premier temps on identifie les quantités $\widetilde{K}_0(P(z)) = \alpha_0 + K_0(P(z))$ et $\widetilde{K}_1(P(z)) = \alpha_1 + K_1(P(z))$. Pour cela on forme le résidu $\widehat{v}_0 = y - r\widehat{\beta}_0$ et on utilise le fait que

$$E(v_0 | T = 0, P(z)) = E(y - r\beta_0 | T = 0, P(z)) = \alpha_0 + K_0(P(z)) = \widetilde{K}_0(P(z))$$

la régression non paramétrique du résidu sur le score fournit un estimateur de $\widetilde{K}_0$. Par exemple pour une valeur donnée de p_0 de $P(z)$ on estime :

$$\widehat{\widetilde{K}}_0(p_0) = \frac{\sum_{j \in I_0} K_h(P(z_j) - p_0)\,\widehat{v}_{0i}}{\sum_{j \in I_0} K_h(P(z_j) - p_0)}$$

Pour identifier les constantes p.e. α_0 il est nécessaire de disposer de valeurs de $P(z)$ telle que $K_0(P(z)) = 0$.

Il existe une possibilité d'identification naturelle. On a les relations :

$$K_0(0) = 0 \text{ et } K_1(1) = 0$$

En effet, on utilise le fait que $E(u_0 | z) = 0$ et $E(u_1 | z) = 0$. Pour la fonction K_0 par exemple, on a

$$E(u_0 | z) = 0 = E(u_0 | z, T = 1) P(z) + E(u_0 | z, T = 0)(1 - P(z))$$

et la fonction K_0 est définie par :

$$K_0(P(z)) = E(u_0 | z, T = 0)$$

On a donc :

$$E(u_0 | z, T = 1) P(z) + K_0(P(z))(1 - P(z)) = 0$$

On a donc bien $K(0) = 0$:

Une façon de tirer parti de ces restrictions est de considérer la moyenne des"résidus" $y - r\widehat{\beta}_0$ pour les individus non traités ayant une faible probabilité d'être traité. Plus précisément, un estimateur de la constante α_0 pourrait être :

$$\widehat{\alpha}_0 = \frac{\sum_i \left(y_i - r_i\widehat{\beta}_0\right)(1 - T_i)\,1\left(z_i\widehat{c} < \gamma_n^-\right)}{\sum_i (1 - T_i)\,1\left(z_i\widehat{c} < \gamma_n^-\right)}$$

où γ_n^- est une suite tendant vers $-\infty$.

Remarque *Ces hypothèses permettent d'identifier "à l'infini" la constante α_0, et donc la fonction $K_0(\cdot)$. Il est possible d'identifier ainsi $E(y_0)$ et $E(y_0|T=1)$. Ces hypothèses suffisent donc pour identifier Δ. On peut remarquer que dans ce cas la détermination du paramètre d'intérêt fait intervenir la détermination de la fonction K_0 en chaque point du support du score pour les individus traités. La forme finale de l'estimateur est ainsi*

$$\begin{aligned}
\widehat{\Delta}^{TT} &= \frac{1}{N_1}\sum_{T_i=1}\left[y_i - \widehat{\alpha}_0 - r_i\widehat{\beta}_0 + \frac{1 - P(z_i)}{P(z_i)}\left(\frac{\sum_{j\in I_0} K_h\left(P(z_j) - P(z_i)\right)\left(y_j - r_j\widehat{\beta}_0\right)}{\sum_{j\in I_0} K_h\left(P(z_j) - P(z_i)\right)} - \widehat{\alpha}_0\right)\right] \\
&= \frac{1}{N_1}\sum_{T_i=1}\left[y_i - \frac{\widehat{\alpha}_0}{P(z_i)} - r_i\widehat{\beta}_0 + \frac{1 - P(z_i)}{P(z_i)}\left(\frac{\sum_{j\in I_0} K_h\left(P(z_j) - P(z_i)\right)\left(y_j - r_j\widehat{\beta}_0\right)}{\sum_{j\in I_0} K_h\left(P(z_j) - P(z_i)\right)}\right)\right]
\end{aligned}$$

Dans ce cas il est possible d'identifier la constante α_1 et donc la fonction K_1. On peut sous l'ensemble de ces hypothèses identifier le paramètre $E(y_1)$ et donc l'effet moyen du traitement qui sera simplement défini comme

$$\widehat{\Delta}^{ATE} = \frac{1}{N}\sum_i\left[\widehat{\alpha}_1 - \widehat{\alpha}_0 + r_i\left(\widehat{\beta}_1 - \widehat{\beta}_0\right)\right]$$

En pratique la probabilité de recevoir le traitement est souvent concentrée vers des valeurs faibles. Si les hypothèses sur les queues de distribution, concernant l'identification de α_0 sont vraisemblables, il n'en est pas de même de celles concernant l'identification de α_1. Il est donc vraisemblable qu'en général l'identification de l'effet moyen du traitement échappe à ce type d'approche.