# *ENDOGENEITE ET VARIABLES INSTRUMENTALES DANS LES SCIENCES SOCIALES*

*J-M. ROBIN*

INRA-LEA et CREST-INSEE

**Résumé**

Dans ce manuel nous essayons d'abord de recenser les situations pouvant être à l'origine d'une corrélation entre une ou plusieurs variables explicatives et le terme d'erreur d'un régression, puis nous montrons comment la technique des variables instrumentales permet une estimation convergente des paramètres du modèle. Après avoir examiné le cas des modèles linéaires, nous expliquons, pour quelques cas particuliers importants, comment cette technique peut s'adapter au cas des modèles non linéaires. Enfin, nous illustrons les différentes techniques développées en procédant à l'estimation de modèles d'activité et de salaires sur un échantillon tiré de l'enquête Emploi à l'aide de procédures SAS de base.

# 1 Introduction

La régression est un outil commode pour résumer la force du "lien" statistique entre deux variables dans une population.[1] Par exemple, le signe du coefficient de la régression du revenu des fils sur celui des pères, ou l'inverse, est une information sur le sens de l'association statistique entre ces deux variables et le $R^2$ de la régression est un indicateur de sa "force".[2] Considérons maintenant trois variables $x$, $y$ et $z$. Le coefficient de $x$ dans cette régression multiple s'obtient de façon équivalente en régressant d'abord $y$ sur $z$, puis en régressant les résidus $\widehat{u}$ de cette régression sur la variable $x$. La régression multiple fournit donc une information sur ce qu'une variable explique *en propre* dans un ensemble de variables supposées explicatives, où "expliquer" a le sens que lui donne l'interprétation du $R^2$ comme part de variance que le modèle permet de reproduire. C'est aussi le sens de la formule *"ceteris paribus"*. Par exemple, si on régresse une consommation (une pratique) sur un croisement du revenu et du diplôme et que l'ajout d'une variable de PCS dans cette régression contribue à augmenter sensiblement le $R^2$ de la régression, on conclura que l'information contenue dans la PCS est sensiblement différente de celle contenue dans le croisement du revenu et du diplôme.

La régression multiple et l'analyse de la variance ne sont toutefois applicables que si $x$ varie encore lorsqu'on a fixé $z$. Sinon, à l'évidence, toute l'information contenue dans la variable $x$ est déjà contenue dans la variable $z$. C'est cependant mal comprendre l'outil que de croire que placer simultanément revenu, diplôme et PCS dans une régression c'est forcément donner une réalité à cet être social improbable que serait une épicière polytechnicienne pauvre, car régresser une variable de pratique sur les variables de revenu, de diplôme et de PCS ne préjuge en rien des relations que ces trois variables entretiennent entre elles. Certains croisements

[1] Je remercie Daniel Verger pour ses commentaires attentifs et ses conseils quant à la définition du contenu de ce manuel. Je remercie Stéphan Lollivier, Philippe Zamora et Sébastien Roux pour leurs commentaires, et Anne Flipo pour ses commentaires sur différentes versions de ce texte, ainsi que pour son aide précieuse à la fois dans la définition et la réalisation de l'illustration empirique.

[2] Un mot ici sur la significativité du coefficient d'une régression : la significativité d'une statistique n'est pas une information sur la relation entre $x$ et $y$ dans la population, c'est une information sur la capacité de l'échantillon à fournir une information crédible sur cette relation. N'importe quelle estimation devient significative pour peu que la taille de l'échantillon soit assez grande ! A taille d'échantillon donnée, un coefficient de régression sera d'autant plus significatif que le $R^2$ sera grand.

peuvent ainsi fort bien ne pas exister.

L'omission d'une variable dans une régression produit les ravages que l'on sait : les variables qui lui sont corrélées prenant une partie du pouvoir explicatif qui lui revient.[3] La crainte de la fameuse variable cachée responsable d'associations statistiques fallacieuses (la cigogne de Lazarsfeld !) hante le chercheur en sciences sociales. Le pire est sans doute atteint lorsqu'on a de bonnes raisons de penser que cette variable cachée existe mais on ne peut l'exhiber. Une telle situation arrive souvent lorsque le travail empirique procède d'une entreprise de validation ou de falsification des conclusions logiques d'un raisonnement théorique à partir de prédicats (hypothèses) dont la falsification n'est pas directement à la porté du travail empirique. On qualifie cette démarche de *structurelle*.

La question dont nous traitons dans la suite de ce texte a à voir avec la problématique suivante. Nous avons certaines raisons, théoriques ou expérimentales, de penser qu'il existe une relation de cause à effet (déterministe) entre une variable $x$ et une autre variable $y$. Si $x$ n'est pas la seule cause de $y$ et que cette autre cause possible, disons $u$, n'est pas indépendante de $x$, alors la corrélation statistique entre $x$ et $y$ capture plus ou autre chose que la relation de causalité entre $x$ et $y$. Evidemment, $u$ n'est pas directement observable, sinon la régression multiple serait la solution. Nous explorons dans ce manuel la voie suivante : $u$ n'est pas observable, mais supposons qu'il soit possible d'exhiber une autre variable $z$ déterminant $x$ indépendamment de $u$. On qualifie une telle variable d'*instrumentale*. La variable $z$ ne causant $y$ qu'à travers $x$, en faisant varier $z$ on fait donc varier $x$ indépendamment de $u$, ce qui rend à nouveau possible une méthode "toutes choses égales d'ailleurs". C'est la technique des variables instrumentales.

La suite de l'exposé est composé de la façon suivante. Nous définissons d'abord une classe de modèles, dits apparemment linéaires, suffisamment générale pour fournir une base rigoureuse à la discussion de ces situations où la façon dont une variable $y$ dépend d'une variable $x$ n'est pas indépendante de la façon dont $x$ est elle-même déterminée. Puis nous donnons quelques exemples supplémentaires de situations de ce type. Les trois sections suivantes traitent des solutions à apporter au problème d'endogénéité de variables explicatives. Tout d'abord nous traitons

[3] Attention, c'est important : une variable omise ne biaise pas les paramètres d'une régression dont les variables explicatives sont non corrélées avec la variable omise.

brièvement le cas des données de panels. Nous décrivons ensuite l'estimateur à variables instrumentales ainsi que les tests d'exogénéité et de validité des variables instrumentales. La section suivante montre comment faire dans le cas de modèles non linéaires de structure plus complexe que les modèles apparemment linéaires. Enfin nous developpons dans la dernière partie une illustration empirique concrète.

## 2 Modèles linéaires avec variables explicatives endogènes

Pour poser le problème nous commencerons par l'étude du modèle classique d'équilibre offre-demande. Soit un échantillon de $n$ couples $(x_i, y_i)$, $i = 1, ..., n$, de variables de prix $(x_i)$ et de ventes $(y_i)$. La théorie économique de l'équilibre entre l'offre et la demande nous apprend qu'il n'y a pas à proprement parler de relation de cause à effet entre prix et ventes : ce sont deux résultats simultanés de l'équilibre concurrentiel de marché qui fait que le prix trouve sa valeur d'équilibre lorsque la demande de bien est égale à l'offre. On pose ainsi quatre équations : une équation de demande :

$$D_i(p_i, u_i) = a + bp_i + u_i, \tag{1}$$

qui décrit la réponse de la demande $D_i(p_i, u_i)$ à un choc de prix $p_i$, causal, et un choc de demande $u_i$ (avec vraisemblablement $b$ négatif) ; une équation d'offre :

$$S_i(p_i, v_i) = \alpha + \beta p_i + v_i \tag{2}$$

dont l'interprétation est similaire (avec $\beta > 0$) ; enfin, deux relations d'équilibre :

$$y_i = D_i(p_i, u_i) = S_i(p_i, v_i), \tag{3}$$

$$x_i = p_i, \tag{4}$$

où $x_i$ est le prix d'équilibre, i.e. celui qui permet d'ajuster l'offre à la demande, alors toutes les deux égales à $y_i$, la quantité échangée.

Le modèle précédent fait intervenir deux variables supplémentaires $u_i$ et $v_i$ qui ont bien une interprétation "résiduelle" puisqu'elles ont vocation à représenter l'ensemble des variables susceptibles de conditionner les variations d'offre et de demande indépendamment des prix. Par exemple, $u_i$ peut qualifier les changements dans les préférences des ménages et $v_i$ les changements technologiques des

processus de production. Il ne s'agit toutefois pas de résidus statistiques mais de variables économiques ayant une interprétation structurelle. L'économie est ici une mécanique où tout mouvement résulte d'un choc propre.

Il importe de bien comprendre la différence entre la notion de résidu statistique et celle de perturbation ou de choc dans une équation économique comportementale. On a l'habitude de poser le problème de l'inférence dans le modèle de régression linéaire normal comme celui de l'estimation des paramètres $a$ et $b$ dans l'équation :

$$y_i = a + bx_i + u_i, \tag{5}$$

où $u_i$ est une variable aléatoire indépendante de $x_i$ et distribuée selon une loi normale de moyenne 0 et de variance $\sigma^2$. Cependant, pour ce qui concerne l'inférence statistique, on aurait tout aussi bien pu se passer d'introduire la variable $u_i$. En fait, rien n'empêche l'équation (5) de n'être qu'une tautologie, définissant la variable $u_i$ comme la différence $y_i - a - bx_i$. Le modèle linéaire normal est ainsi entièrement contenu dans les deux hypothèses suivantes :

1. les observations $(y_i, x_i)$ sont deux à deux indépendantes ;
2. la loi de $y_i - a - bx_i$ est normale de moyenne 0 et de variance $\sigma^2$.

Sous ces deux hypothèses, l'estimateur des moindres carrés ordinaires (MCO), $(\widehat{a}, \widehat{b})$, est normal, sans biais, et convergent et efficace puisqu'il est l'estimateur du maximum de vraisemblance. Le *résidu statistique* de la régression est ensuite défini par l'expression : $\widehat{u}_i = y_i - \widehat{a} - \widehat{b}x_i$, et sert dans la construction de différents tests statistiques de la spécification du modèle.

Le modèle linéaire ne dépend donc d'aucune interprétation particulière du terme $u_i$ dans l'équation (5), et admet tout aussi bien chacune des deux interprétations suivantes de la relation entre $y_i$ et $x_i$ :

1. ou bien il s'agit d'une description de la façon dont $y_i$ est distribué dans la population de référence conditionnellement à $x_i$ ;
2. ou bien il s'agit d'une équation déterministe, spécifiant comment la variable $y_i$ se déduit de la variable $x_i$, observée par le statisticien, et d'une autre variable $u_i$, inobservée mais bien réelle, et dont on précise la distribution dans la population de référence.

Dans le premier cas, on parlera de *modèle descriptif* (et peu importe si $y_i$ résulte d'un coup de dés ou non), dans le second cas, de *modèle structurel.*

Cette parenthèse étant fermée, revenons au modèle d'équilibre offre-demande. Les restrictions (3) et (4) imposent aux prix d'équilibre de satisfaire la relation suivante :

$$x_i = \frac{1}{b-\beta}\left[\alpha - a + v_i - u_i\right], \tag{6}$$

qui caractérise la façon dont les chocs d'offre et de demande déterminent les prix d'équilibre. Les ventes s'obtiennent quant à elles en substituant $x_i$ à $p_i$ dans les équations d'offre et de demande :

$$\begin{aligned} y_i &= a + bx_i + u_i \\ &= \alpha + \beta x_i + v_i \\ &= \frac{\alpha b - a\beta}{b-\beta} - \frac{\beta}{b-\beta}u_i + \frac{b}{b-\beta}v_i. \end{aligned} \tag{7}$$

Dans la terminologie des modèles à équations simultanées, le modèle formé des équations comportementales et d'équilibre (1) à (4) définit le modèle *structurel* et celui formé des équations (6) et (7) définit le modèle *réduit.* D'une façon générale, on qualifie habituellement de structurelle toute démarche empirique ayant pour objectif premier l'estimation des paramètres d'un modèle économique théorique. Les modèles dont les paramètres sont des fonctions (souvent non spécifiées) des paramètres structurels fondamentaux tombent dans la catégorie des modèles réduits. C'est un premier critère. J'en ajouterai un autre : pour qu'un modèle économétrique puisse être justement qualifié de structurel, il faut que l'aléa lui-même ait sa justification théorique, qu'il résulte d'une loterie (comme dans le cas d'une stratégie mixte), où de processus déterministes trop complexes pour être décrits exhaustivement. Dans ce dernier cas, l'erreur aléatoire a le statut d'une variable économique synthétique. Mais on ne construit pas un modèle économétrique en ajoutant en toute fin de construction des erreurs aléatoires aux équations d'un modèle économique (sauf si l'on croit que le bruit enregistré résulte essentiellement d'erreurs de mesure additives).

Considérons maintenant les relations structurelles : $y_i = a + bx_i + u_i$ et $y_i = \alpha + \beta x_i + v_i$. Le passage du modèle économique au modèle économétrique se fait au moment où l'on fait des hypothèses sur la loi statistique des variables inob-

servables. Supposons les couples $(u_i, v_i)$ mutuellement indépendants, de même loi, et d'espérance nulle. Les équations (6) et (7) impliquent que les couples $(y_i, x_i)$ sont alors eux-mêmes mutuellement indépendants et de même loi. La régression de $y_i$ sur $x_i$ dans un échantillon d'observations $\{(y_i, x_i), i = 1, ..., n\}$ produira deux estimations $\widehat{\theta}_0$ et $\widehat{\theta}_1$, pour la constante de la régression et le coefficient de $x_i$, qui convergeront, lorsque $n$ croît, vers deux valeurs $\theta_0$ et $\theta_1$, fonctions des paramètres $a, b, \alpha$ et $\beta$, et de la variance de $(u_i, v_i)$. Par exemple, supposons que

$$\text{var}(u_i) = \sigma_u^2, \ \text{var}(v_i) = \sigma_v^2 \text{ et } \text{cov}(u_i, v_i) = \rho\sigma_u\sigma_v.$$

Le coefficient de la pente de la régression théorique de $y_i$ sur $x_i$ (ce vers quoi converge $\widehat{\theta}_1$ lorsque le nombre d'observations tend vers l'infini) est alors :

$$\begin{aligned} \theta_1 &= \frac{\text{cov}(y_i, x_i)}{\text{V}\, x_i} \\ &= \frac{\beta\sigma_u^2 - (\beta + b)\rho\sigma_u\sigma_v + b\sigma_v^2}{\sigma_u^2 - 2\rho\sigma_u\sigma_v + \sigma_v^2}. \end{aligned}$$

Le moins que l'on puisse dire est que le lien avec les paramètres structurels est plutôt complexe ! Il est toujours possible de régresser $y_i$ sur $x_i$ mais on voit que l'interprétation des résultats est délicate.

Cependant, pourquoi le coefficient de la régression ne converge-t-il ni vers $b$ ni vers $\beta$ alors que les perturbations $u_i$ et $v_i$ sont d'espérance nulle et que les équations $y_i = a + bx_i + u_i$ et $y_i = \alpha + \beta x_i + v_i$ ont toutes les apparences d'une régression linéaire, y compris lorsque les perturbations $u_i$ et $v_i$ sont indépendantes $(\rho = 0)$ ? C'est que

$$\begin{aligned} \theta_1 &= \frac{\text{cov}(y_i, x_i)}{\text{V}\, x_i} \\ &= b + \frac{\text{cov}(u_i, x_i)}{\text{V}\, x_i} \\ &= \beta + \frac{\text{cov}(v_i, x_i)}{\text{V}\, x_i} \end{aligned}$$

n'est ni $b$ ni $\beta$ dès lors que $\text{cov}(u_i, x_i) \neq 0$ et $\text{cov}(v_i, x_i) \neq 0$ (y compris si les deux chocs $u_i$ et $v_i$ sont indépendants).

En résumé, soit la régression

$$y_i = a + bx_i + u_i. \tag{8}$$

On dira que la variable $x_i$ est endogène pour $u_i$ si $x_i$ est liée à $u_i$ par une relation structurelle qui se traduit empiriquement par le fait que $x_i$ est corrélée à $u_i$. Le modèle (8) n'est ainsi qu'*apparemment* linéaire. Bien sûr il est toujours possible d'estimer la régression (8) par la méthode des moindres carrés ordinaires. On obtiendra une description des corrélations entre la variable de gauche et les variables de droite. Mais l'interprétation des corrélations en règles causales suppose une modélisation structurelle préalable, formalisée ou non.

**Définition (Modèle apparemment linéaire ou modèle linéaire à variables explicatives endogènes)** *Soit n triplets de variables aléatoires $(y_i, x_i, u_i)$ indépendants, de même loi, et à valeurs dans $R \times R^K \times R$. Un modèle apparemment linéaire est une relation $y_i = x_i'b + u_i$ où $\mathrm{E}\, u_i = 0$ et $\mathrm{cov}(u_i, x_i) \neq 0$. La variable $x_i$ est alors dite endogène par rapport à $u_i$.*

L'objet de la note qui suit est de montrer comment produire une estimation convergente des paramètres de tels modèles. On commencera par développer quelques exemples supplémentaires de modèles structurels apparemment linéaires pour mieux faire comprendre les différentes sources possibles d'endogénéité. Puis, après avoir brièvement décrit en quoi les données de panel constituent une première réponse au problème, nous verrons comment la méthode des variables instrumentales permet d'estimer de tels modèles en toute généralité. Enfin, nous examinerons l'extension de la méthode aux modèles non-linéaires.

# 3 Différentes causes possibles de l'endogénéité

Il convient de distinguer les cas d'échantillonnages d'individus dans une population de ceux d'échantillonnages temporels (séries temporelles), les données de panels cumulant les deux propriétés. Dans le cas d'un échantillonnage individuel, la source fondamentale d'endogénéité est l'existence d'*hétérogénéité inobservée*. Dans celui d'un échantillonnage temporel, la source d'endogénéité est, comme on l'a déjà vu avec le modèle d'équilibre, la *simultanéité* (les chocs d'offre et de demande déterminaient simultanément les ventes et les prix). Lorsque l'on effectue une régression, il faut toujours se demander si la régression est descriptive ou structurelle. Cherche-t-on à décrire des corrélations, ou à tester une relation causale déterministe ? Dans le dernier cas, il faut ensuite se demander si la présence d'hétérogénéité inobservée ou de simultanéité n'est pas susceptible de biaiser l'estimation.

## 3.1 Hétérogénéité inobservée

On comprendra l'effet dévastateur de la présence d'hétérogénéité inobservée en observant la figure ??. L'échantillon total est la réunion de deux sous-échantillons homogènes. Dans chacun de ces échantillons, une variable $Y$ est liée positivement à une variable $X$. Mais la régression dans l'échantillon entier fait apparaître une corrélation négative.

On observe ainsi que la corrélation entre la quantité achetée d'un bien stockable ($Y$) comme les céréales de petit déjeuner, par exemple, et la durée séparant la date d'achat du prochain renouvellement ($X$) est négative. Or, ce résultat est contraire à l'intuition qui suggère que si un ménage achète plus que d'habitude pour une raison ou pour une autre, il va retarder le moment du prochain achat, car le fait d'acheter plus ne signifie vraisemblablement pas qu'il s'est subitement mis à consommer plus.[4] On peut expliquer ce phénomène en suggérant que la population des acheteurs/consommateurs de céréales est hétérogène : les petits consommateurs achètent à la fois en moins grande quantité et moins souvent (durées inter-achats plus courtes ; la patate de droite de la figure ??) que les gros consommateurs (la patate de gauche). Dans chaque groupe de consommateurs, la relation entre du-

[4] L'intuition est ici l'expression d'un raisonnement économique décrivant les approvisionnements des ménages comme l'expression d'un comportement de renouvellement de stocks.

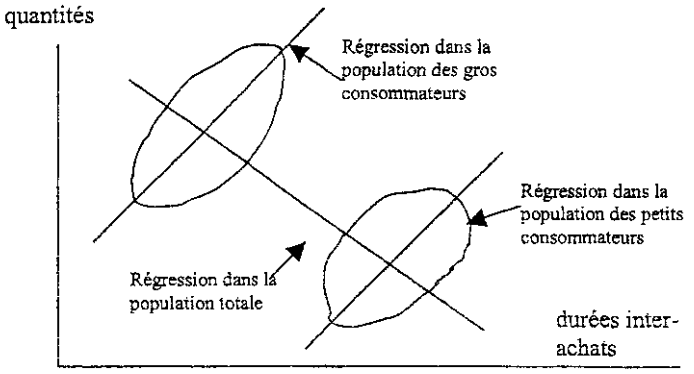


FIG. 1: Régression dans un mélange de populations

rées et quantités est positive (une régression dans chaque sous-échantillon produit une droite de pente positive), mais dans l'échantillon entier, l'écart entre les deux groupes est tel qu'une régression produit une droite de pente négative. L'hétérogénéité individuelle explique ainsi une plus grande part de la variance totale du couple $(Y, X)$ dans la population totale que les fluctuations d'approvisionnement suscitées, par exemple, par les opérations promotionnelles des grandes surfaces.

Autre exemple : il s'agit d'estimer le rendement du diplôme sur le marché du travail. Soit $\{(y_i, x_i), i = 1, ..., n\}$ un échantillon d'observations individuelles de salaires et de nombres d'années d'études. Supposons que l'équilibre sur le marché du travail induise une relation déterministe linéaire entre le salaire $y_i$, l'éducation $x_i$ et d'autres caractéristiques des individus et des entreprises, que nous résumerons par la variable $u_i$ : $y_i = a + bx_i + u_i$. Si certaines des caractéristiques individuelles inobservées par l'économètre le sont cependant par l'employeur et contribuent ainsi à la formation des salaires, ces caractéristiques ont elles mêmes des chances d'être corrélées avec le diplôme, parce que si un individu est efficace au travail, il y a des chances pour qu'il l'ait été aussi à l'école, corrélant par là même $x_i$ avec $u_i$. Dans ce cas, le coefficient de la régression linéaire de $y_i$ sur $x_i$ surestimera $b$, le rendement propre du diplôme. Cet exemple montre très bien combien il est dangereux d'interpréter les résultats d'une régression sans faire l'effort d'expliciter au

préalable les mécanismes économiques et sociaux que l'on pense être au fondement des relations entre les variables de l'analyse statistique.

Ceux qui contestent la qualité de l'estimateur des MCO de $b$ pour estimer le rendement des études, ont en tête le modèle Beckerien suivant. Le salaire est le prix des compétences individuelles sur le marché du travail. Supposons que ces compétences s'agrégent dans un critère unique de qualité : le capital humain. Dans un marché concurrentiel, le salaire est égal à la productivité marginale, elle-même directement fonction du capital humain. Les agents naissent avec une dotation initiale en capital humain (laquelle dépend des hasards génétiques et familiaux) qui rend inégales les chances de réussite scolaire. Posons alors les trois équations suivantes :

$$\begin{aligned} y_i &= a + bH_i + \eta_i, \\ H_i &= x_i + \varepsilon_i, \\ x_i &= \beta\varepsilon_i + e_i, \end{aligned}$$

qui décrivent le salaire $y_i$ comme une fonction du capital humain $H_i$ et d'une variable, $\eta_i$, représentant les aléas de marché local, les erreurs de mesure sur le salaire, les imperfections de marché diverses, qu'on suppose exogènes. La seconde équation spécifie le capital humain comme la somme du niveau scolaire, $x_i$, et d'un terme, $\varepsilon_i$, qui symbolise la dotation initiale en compétences de l'individu. La troisième équation dit que le niveau scolaire varie selon les capacités individuelles, $\varepsilon_i$, et l'offre locale d'éducation $e_i$, qu'on supposera aussi exogène.

Eliminons $H_i$ du système :

$$\begin{aligned} y_i &= a + bx_i + b\varepsilon_i + \eta_i, \\ x_i &= \beta\varepsilon_i + e_i. \end{aligned}$$

On obtient un système de deux équations qui montre que la source de l'endogénéité de $x_i$ par rapport au terme inobservable de l'équation de salaire : $u_i = b\varepsilon_i + \eta_i$, est le terme d'hétérogénéité individuelle de compétences non acquises par l'éducation scolaire, $\varepsilon_i$.

Substituons enfin $x_i$ dans la première équation par sa valeur dans la seconde :

$$y_i = a + be_i + (b + \beta)\varepsilon_i + \eta_i.$$

On voit que le rendement spécifique du diplôme pourrait être identifié à condition de disposer de données et d'une variabilité suffisante dans les conditions d'offre scolaire. Encore faut-il que l'hypothèse d'exogénéité de ces conditions par rapport à la dotation initiale de capital humain $\varepsilon_i$ soit effectivement satisfaite. Or, c'est assez douteux, lorsqu'on sait les stratégies efficaces que les parents développent pour détourner le principe de la carte scolaire! Il arrivera souvent que le travail d'investigation économétrique se heurte à des difficultés de ce type. Constatant un problème d'endogénéité, on ne saura pas le réduire, car on n'arrivera pas à trouver une variable exogène déterminant en propre la variable endogène à laquelle on s'intéresse (on appelera plus loin de telles variables exogènes des instruments). Il s'agit là d'une des difficultés fondamentales du travail empirique dans les sciences sociales.[5]

Avant de passer à l'exemple suivant, remarquons que si l'effet que l'on cherche à estimer est le rendement descriptif du diplôme comme "prédicteur" du salaire, alors le problème d'endogénéité du diplôme dans l'équation de salaire disparaît comme par enchantement. Reprenons l'équation : $y_i = a + bx_i + u_i$, avec $x_i$ et $u_i$ corrélés. Supposons ces deux variables normales. Il s'ensuit que la loi conditionnelle de $u_i$ sachant $x_i$ est encore normale et l'on peut écrire :

$$u_i = \gamma x_i + v_i,$$

avec $x_i$ et $v_i$ deux variables normales indépendantes et $\gamma = \frac{\text{cov}(u_i, x_i)}{Vx_i}$. De sorte que l'équation de salaire devient :

$$y_i = a + (b + \gamma)x_i + v_i,$$

qui montre que la régression de $y_i$ sur $x_i$ surestime le rendement du diplôme ($\gamma > 0$ si $\text{cov}(u_i, x_i) > 0$) et qui montre aussi combien on ne peut parler de variable explicative endogène sans modèle structurel.

On pourrait sans doute résumer les choses de la façon suivante : à tout modèle apparemment linéaire

$$y_i = a + bx_i + u_i,$$

[5] On peut trouver le développement qui précède un peu complexe. Mais il ne s'agit pas seulement de donner un exemple de variable endogène, il s'agit aussi de faire comprendre la nature du raisonnement qui permet d'apporter une solution au problème d'endogénéité, d'où ici la discussion jointe de l'hétérogénéité des capacités individuelles et de l'offre scolaire.

avec $x_i$ et $u_i$ corrélés, correspond un modèle linéaire

$$y_i = a + (b + \gamma)x_i + v_i,$$

avec $x_i$ et $v_i$ non corrélés.[6] Un modèle apparemment linéaire ne prend sens que dès lors qu'on peut le compléter par une relation complémentaire exprimant la relation entre $x_i$, $u_i$ et une tierce variable exogène, qu'on qualifiera d'*instrumentale*.

Un dernier exemple (tiré d'une liste inépuisable) nous est fourni par le modèle de participation au marché du travail. La variable $y_i$ représente le nombre d'heures travaillées dans la semaine et $x_i$ est le salaire horaire payé (attention au changement de notation par rapport à l'exemple précédent). Plaçons nous dans le cadre d'un modèle d'arbitrage entre consommation et loisir. Supposons pour simplifier que la consommation soit celle d'un seul bien composite, de prix unitaire. La maximisation de l'utilité du couple (consommation, loisir) sous la contrainte que la consommation est égale aux revenus (salaire plus revenu non salarial) conduit nécessairement à un nombre optimal d'heures ouvrées (nombre d'heures disponibles moins le nombre d'heures de loisir), fonction du salaire horaire offert et de certaines caractéristiques de la fonction d'utilité. Disons que $u_i$ dans la régression $y_i = a + bx_i + u_i$ résume l'hétérogénéité des goûts des individus pour la consommation et le loisir.[7]

C'est, mine de rien, un modèle assez riche, puisqu'il permet de prédire une relation des heures travaillées au salaire horaire, soit uniformément croissante, soit d'abord croissante puis décroissante pour des salaires élevés. (De plus le nombre d'heures ouvrées décroît nécessairement lorsque s'accroît le revenu non salarial.) On suppose alors généralement que le salaire est endogène dans l'équation d'heure parce que toute caractéristique individuelle favorisant le goût pour le travail a des chances d'être une caractéristique recherchée par les employeurs et peut donc par là même engendrer des salaires plus élevés. Ici, cependant, on ne se préoccupera

[6] Nous reviendrons plus rigoureusement sur ce point dans la section suivante.

[7] Nous avons laissé volontairement de côté la contrainte de participation proprement dite, qui conduit certains agents à ne pas participer si l'activité n'est pas assez profitable. La question de la sélection *endogène* d'échantillon n'est qu'indirectement liée à la notion d'endogénéité qui nous intéresse ici. Nous reviendrons plus longuement sur le traitement de l'endogénéité dans les modèles de sélection endogène dans la section 7.

généralement pas de l'endogénéité potentielle du diplôme dans l'équation de salaire car il ne s'agit pas d'estimer le rendement propre du diplôme sur les salaires.

## 3.2 La simultanéité

Nous avons déjà vu l'exemple du modèle d'équilibre. Un autre exemple nous est fourni par l'étude des systèmes de demande. Un système de demande est une relation entre un vecteur de demandes pour une liste de biens, leurs prix et la dépense totale. C'est aussi le résultat d'une opération de maximisation de l'utilité pour la consommation d'un panier de biens sous la contrainte de budget. Oublions les prix qui sont ici généralement considérés comme exogènes, ou encore, considérons une coupe d'observations des dépenses d'un échantillon de ménages soumis au même système de prix. Pour simplifier, supposons qu'il n'y a que deux biens. Soit $(y_{1i}, y_{2i})$ le vecteur des dépenses du ménage $i$ en chaque bien. Par définition, la dépense totale est $y_i = y_{1i} + y_{2i}$. Un système de demande linéaire postule que

$$y_{1i} = a_1 + b_1 y_i + u_{1i}, \tag{9}$$

$$y_{2i} = a_2 + b_2 y_i + u_{2i}, \tag{10}$$

où $u_{1i}$ et $u_{2i}$ sont deux composantes d'hétérogénéité individuelle des goûts pour chaque bien. Puisque $y_i = y_{1i} + y_{2i}$ pour tout $i$ (contrainte dite d'additivité), il faut donc bien que

$$a_1 + a_2 = 0, \tag{11}$$

$$b_1 + b_2 = 1, \tag{12}$$

$$u_{1i} + u_{2i} = 0, \quad \forall i. \tag{13}$$

De sorte que la seconde équation du système de demande n'apporte aucune information supplémentaire par rapport à la première. Les systèmes de demande s'estiment en ôtant une équation du système, les paramètres de l'équation écartée étant récupérés grâce à la contrainte d'additivité.

Le fait que les erreurs correspondant aux différentes équations de demande sont nécessairement liées entre elles par la condition (13) n'est pas la cause de l'endogénéité de $y_i$ (sauf en présence d'erreurs de mesure sur les dépenses), parce qu'en sommant les dépenses on élimine *de facto* les perturbations spécifique à

chaque équation ($u_{1i} + u_{2i} = 0$). La raison qui fait qu'on considère généralement la dépense totale comme une variable endogène dans un système de demande, est que la dépense totale est vraisemblablement l'objet d'une décision préalable, le partage du revenu entre consommation et épargne par exemple. Cette décision dépend des préférences du ménage au même titre que le partage du budget entre les différentes fonctions de consommation. Il faut donc ajouter au système (9)-(10) une troisième équation :

$$y_i = a + bx_i + u_i, \tag{14}$$

par exemple, décrivant l'effet du revenu $x_i$ sur le budget $y_i$, où $u_i$ est une autre variable de choc sur les préférences.

On peut légitimement supposer les chocs de revenu $x_i$ orthogonaux aux chocs de préférences $u_i$, $u_{1i}$ et $u_{2i}$. En revanche, on voit bien que si $\text{cov}(u_i, u_{1i}) \neq 0$, par exemple, alors $y_i$ est endogène par rapport à $u_{1i}$ dans l'équation (9).[8]

## 3.3 Les erreurs de mesure

Outre la présence d'hétérogénéité inobservée et la réponse simultanée à des chocs inobservés, l'existence d'erreurs de mesure sur une variable explicative rend l'observation endogène.

Considérons un modèle linéaire :

$$y_i = a + bx_i^* + u_i,$$

avec $\text{E}(u_i \mid x_i^*) = 0$. Supposons que l'on observe $x_i^*$ avec erreur :

$$x_i = x_i^* + \varepsilon_i.$$

avec $\text{E}(\varepsilon_i \mid x_i^*, u_i) = 0$. Alors,

$$y_i = a + bx_i + u_i - b\varepsilon_i,$$

[8] Le revenu salarial est endogène puisque l'offre de travail est, on l'a vu, endogène. Toutefois, il y a des degrés dans l'endogénéité. La simultanéité de la détermination du budget total et des dépenses qui en dépendent paraît ici plus évidente que la dépendance du revenu aux composantes inobservées des dépenses associées à chaque bien.

Entre parenthèses, l'"accusation" d'endogénéité, comme celle de sélection endogène, ne doivent pas être portées à la légère. Si tout est endogène, alors il n'y plus d'identification possible des relations structurelles. On est alors condamné à estimer des formes réduites, et il n'y a plus guère de place pour une théorie des faits économiques et sociaux empiriquement falsifiable.

où $x_i$, la mesure, est évidemment corrélée avec la perturbation $u_i - b\varepsilon_i$ :

$$\begin{aligned} \mathrm{cov}(x_i, u_i - b\varepsilon_i) &= \mathrm{cov}(x_i^* + \varepsilon_i, u_i - b\varepsilon_i) \\ &= -b\,\mathrm{V}\,\varepsilon_i. \end{aligned}$$

Ainsi dans le cas d'une seule variable explicative, l'erreur de mesure biaise, asymptotiquement, l'estimateur des MCO vers 0. Notons qu'une erreur de mesure sur la variable dépendante diminue la précision des estimateurs mais ne les biaise pas asymptotiquement.

# 4 Le cas des données de panel

Nous ne traiterons pas ici en détail l'économétrie des données de panel. Il importe cependant d'expliquer en quoi le fait de disposer de séries d'observations pour chaque individu ou ménage du panel, permet de résoudre en grande partie les problèmes liés à la présence d'hétérogénéité inobservée. Reprenons l'exemple précédent des consommateurs/acheteurs de céréales de petit déjeuner. Soit $i$ l'indice individuel et $t$ le numéro d'ordre de l'achat de l'individu $i$ dans le panel. Soit $y_{it}$ la durée séparant le $t$-ième du $(t+1)$-ième achat. Soit $x_{it}$ la quantité achetée au cours du $t$-ième achat. Supposons un lien structurel entre $y_{it}$ et $x_{it}$ de la forme

$$y_{it} = a + bx_{it} + u_i + v_{it}, \tag{15}$$

où $u_i$ est le paramètre d'hétérogénéité individuelle et $v_{it}$ un choc représentant l'ensemble des sources non prévisibles de délai dans le renouvellement "normal" des achats. Un échantillon d'observations est un ensemble $\{(x_{i1}, ..., x_{iT_i}, y_{i1}, ..., y_{iT_i}), i = 1, ..., n\}$, où $T_i$ est le nombre d'achats enregistrés au cours de la période d'enquête pour le ménage $i$. Supposons les vecteurs de variables $(x_{i1}, u_i, v_{i1}, ..., v_{iT_i})$, quand $i$ varie, deux à deux indépendants : les comportements d'approvisionnement des ménages sont indépendants[9].

Même si la variable $x_{it}$ et le choc $v_{it}$ sont orthogonaux, $x_{it}$ peut-être endogène dans l'équation (15) si $x_{it}$ et $u_i$ sont corrélées. Vraisemblablement, dans l'exemple du processus d'approvisionnement en céréales :

$$\text{cov}(x_{it}, u_i) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_{it} u_i < 0,$$

quel que soit $t$, où l'égalité (presque sûre) provient de la loi forte des grands nombres. Car les durées faibles sont associées aux quantités élevées, en moyenne.[10] Une façon simple de se débarrasser du paramètre d'hétérogénéité est de calculer la relation qui se déduit de (15) entre les écarts de $y_{it}$ et $x_{it}$ et leurs moyennes individuelles.

Soient en effet :

$$y_{i\cdot} = \frac{1}{T_i} \sum_{t=1}^{T_i} y_{it}, \; x_{i\cdot} = \frac{1}{T_i} \sum_{t=1}^{T_i} x_{it}, \; v_{i\cdot} = \frac{1}{T_i} \sum_{t=1}^{T_i} v_{it}.$$

[9] bien qu'en cas de soldes il puissent ne pas l'être.

[10] Les individus avec des $u_i$ faibles auront des $y_{it}$ faibles (équation (15)) et des $x_{it}$ élevées.

Alors on a :

$$y_{it} - y_{i\cdot} = b\,(x_{it} - x_{i\cdot}) + v_{it} - v_{i\cdot} \tag{16}$$

pour tout $i = 1, ..., n$ et tout $t = 1, ..., T_i$. L'estimateur des MCO appliqués à la régression (16),[1] dite *within* ou intra, sera convergent à condition que :

$$\mathrm{cov}(x_{it} - x_{i\cdot}, v_{it} - v_{i\cdot}) = 0,$$

ce qui, en général, suppose que, pour chaque ménage, chaque $x_{it}$ dans $x_{i\cdot}$ soit orthogonal à chaque $v_{it'}$ dans $v_{i\cdot}$ :

$$\forall t = 1, ..., T_i,\ \forall \tau = 1, ..., T_i,\ \mathrm{cov}(v_{it}, x_{i\tau}) = 0. \tag{17}$$

La perturbation $v_{it}$ est orthogonale au passé, au présent et au futur de $x_{it}$. Si la série des $(x_{i1}, ..., x_{iT_i})$ et celle des $(v_{i1}, ..., v_{iT_i})$ satisfont cette propriété, on dit que $(x_{it})$ est *fortement*, ou *strictement*, *exogène* par rapport à la série des chocs $(v_{it})$.

Dans quelle mesure l'hypothèse d'exogénéité forte est-elle acceptable ? Dans l'exemple considéré, elle est sans doute discutable. Si un empêchement quelconque oblige un ménage à retarder un achat programmé ($v_{it} > 0$), le stock peut baisser en dessous du seuil de renouvellement habituel, et le ménage peut être amené à acheter plus que de coutume la fois suivante ($x_{i,t+1}$ tend à croître avec $v_{it}$). En d'autres termes, vraisemblablement :

$$\mathrm{cov}(v_{it}, x_{i,t+1}) > 0.$$

En revanche, $v_{it}$ s'interprétant comme un allongement ou un raccourcissement *non prévisible* de la période normale de renouvellement d'achat (un raccourcissement si le ménage, profitant d'une promotion par exemple, renouvelle son stock de façon anticipée), il est naturel de supposer que

$$\forall t = 1, ..., T_i,\ \forall \tau = 1, ..., t,\ \mathrm{cov}(v_{it}, x_{i\tau}) = 0. \tag{18}$$

La perturbation $v_{it}$ est orthogonale au passé et au présent de $x_{it}$. Cette propriété est la propriété d'*exogénéité faible* de $(x_{it})$ par rapport à $(v_{it})$.

Considérons finalement cette autre transformation des variables, la différentiation première, qui élimine aussi l'effet individuel $u_i$ :

$$y_{it} - y_{i,t-1} = b\,(x_{it} - x_{i,t-1}) + v_{it} - v_{i,t-1}. \tag{19}$$

[1] Cette estimation correspond à la fonction absorb de la procédure GLM de SAS.

Les MCO appliqués à cette régression sont convergents si

$$\operatorname{cov}(x_{it} - x_{i,t-1}, v_{it} - v_{i,t-1}) = 0.$$

L'hypothèse d'exogénéité faible ne suffit pas à valider cette condition puisque, sous la condition (18),

$$\operatorname{cov}(x_{it} - x_{i,t-1}, v_{it} - v_{i,t-1}) = -\operatorname{cov}(x_{it}, v_{i,t-1}),$$

qui n'est généralement pas nulle. La méthode des variables instrumentales, que nous décrivons dans la section suivante, peut alors être employée pour obtenir un estimateur convergent de $b$ dans l'équation (19) en instrumentant, par exemple, $x_{it} - x_{i,t-1}$ par $x_{i,t-1}$ et $x_{i,t-2}$.

# 5 La méthode des variables instrumentales

Dans cette section, nous étudions la méthode dite des variables instrumentales qui permet d'estimer de façon convergente les modèles apparemment linéaires. Plus précisément, nous allons étudier le modèle :

$$y_{it} = x_i' b + u_i, \quad i = 1, ..., n, \tag{20}$$

où $\mathrm{E}\, u_i = 0$, $b$ est un paramètre de dimension $K$, et les couples $(x_i, u_i)$ sont à valeur dans $\mathbb{R}^K \times \mathbb{R}$ et sont mutuellement indépendants. On décompose le vecteur des $K$ variables explicatives en une partie $x_{1i} \in \mathbb{R}^{K_1}$ endogène et une partie $x_{2i} \in \mathbb{R}^{K_2}$ exogène $(K_1 + K_2 = K)$, i.e. $\mathrm{cov}(u_i, x_{2i}) = 0$. Le vecteur $x_{2i}$ admet en particulier 1 comme composante si la régression (20) contient une constante.

## 5.1 Non-convergence de l'estimateur des moindres carrés ordinaires

Considérons tout d'abord l'estimateur des MCO de $b$. Celui-ci est

$$\begin{aligned}
\widehat{b}_{MCO} &= \left( \sum_{i=1}^{n} x_i x_i' \right)^{-1} \sum_{i=1}^{n} x_i y_i \\
&= \left( \sum_{i=1}^{n} x_i x_i' \right)^{-1} \sum_{i=1}^{n} x_i (x_i' b + u_i) \\
&= b + \left( \sum_{i=1}^{n} x_i x_i' \right)^{-1} \sum_{i=1}^{n} x_i u_i.
\end{aligned}$$

La loi des grands nombres s'applique et lorsque la taille de l'échantillon tend vers l'infini,

$$\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \longrightarrow \mathrm{E}\, x_i x_i'$$

et

$$\frac{1}{n} \sum_{i=1}^{n} x_i u_i \longrightarrow \mathrm{E}\, x_i u_i = \mathrm{cov}(x_i, u_i).$$

Et donc

$$\widehat{b}_{MCO} \longrightarrow b + (\mathrm{E}\, x_i x_i')^{-1}\, \mathrm{E}\, x_i u_i.$$

L'estimateur des MCO est donc biaisé asymptotiquement, le biais étant égal à

$$\gamma = (\mathrm{E}\, x_i x_i')^{-1}\, \mathrm{E}\, x_i u_i.$$

## 5.2 Instruments

On appelle *instrument* ou *variable instrumentale* toute variable non corrélée avec la perturbation $u_i$. Un instrument est donc une variable exogène. Un vecteur $z_i \in \mathbb{R}^H$ d'instruments satisfait donc la condition :

$$\operatorname{cov}(z_i, u_i) = \mathrm{E}\, z_i u_i = 0. \tag{21}$$

Le vecteur $x_{2i}$ est donc un vecteur d'instruments.

L'hypothèse (21) est la condition qui permet de construire un estimateur à variables instrumentales à partir d'un vecteur d'instruments. On la trouvera souvent écrite sous la forme suivante :

$$\mathrm{E}(u_i | z_i) = 0. \tag{22}$$

Cette hypothèse revient exactement à supposer que $u_i$ est non corrélée avec n'importe quelle fonction de $z_i$. Dans ce cas, toute fonction de variables instrumentales est donc encore un instrument. En pratique, comme on se limite toujours à un choix fini de fonctions, les faire figurer toutes dans la liste $z_i$ dans (21) ou supposer (22) n'a que peu de conséquence.

## 5.3 Identification

La condition (21) peut être réécrite comme suit :

$$\mathrm{E}\, z_i(y_{it} - x_i' b) = 0.$$

Soit encore :

$$\mathrm{E}\, z_i y_{it} = (\mathrm{E}\, z_i x_i')\, b. \tag{23}$$

Cette condition définit un système de $H$ équations à $K$ inconnues $b$. On dira que le système (23) *identifie* le paramètre structurel $b$ s'il existe une solution unique, égale à la vraie valeur des paramètres, à ce système. Si $H < K$, à l'évidence, $b$ n'est pas identifié, puisqu'il y a moins d'équations que de variables. Le modèle est dit *sous-identifié*. Si $H = K$ et le rang de $\mathrm{E}\, z_i x_i'$ est égal à $K$, on dit que le modèle est *juste* identifié. Si $H > K$, $\operatorname{rang}(\mathrm{E}\, z_i x_i') = K$ et aucun instrument n'est une combinaison linéaire des autres, alors le modèle est dit *sur-identifié*, ce qui est souhaitable car la sur-identification autorise le test de la validité des instruments (cf *infra*).

Le vecteur des instruments $z_i$ contient au moins le vecteur des variables explicatives exogènes $x_{2i}$. Pourquoi en effet laisser de côté une source potentielle d'identification ? Il est donc nécessaire de trouver au moins $K_1$ instruments supplémentaires pour identifier le modèle, un, au moins, pour chaque variable explicative endogène. Par ailleurs, si le vecteur des variables endogènes contient plusieurs transformations non linéaires d'une même variable, il est légitime d'introduire dans la liste des instruments de ces variables les mêmes transformations non linéaires. Par exemple, on pourra instrumenter la dépense totale et son carré, dans un système de demande quadratique, par le revenu et son carré.

## 5.4 Moindres carrés indirects

Si $H = K$ et si $\mathrm{E}\, z_i x_i'$ est inversible, alors $b = (\mathrm{E}\, z_i x_i')^{-1} \mathrm{E}\, z_i y_i$, et l'on obtient l'estimateur des Moindres Carrés Indirects en remplaçant les espérances par leurs contreparties empiriques :

$$\begin{aligned} \widehat{b}_{MCI} &= \left( \frac{1}{n} \sum_{i=1}^{n} z_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} z_i y_i \\ &= (Z'X)^{-1} Z'y, \end{aligned}$$

en notant $Z$ la matrice dont la $i$-ième ligne est $z_i'$, $X$ la matrice dont la $i$-ième ligne est $x_i'$ et $y$ le vecteur dont la $i$-ième composante est $y_i$.

Si $H > K$, on se ramène au cas précédent en sélectionnant $K$ combinaisons linéaires des instruments : $Az_i$, où $A$ est une matrice $K \times H$, de plein rang. On construit de la sorte une classe d'estimateur :

$$\begin{aligned} \widehat{b}_{MCI}(A) &= \left( \frac{1}{n} \sum_{i=1}^{n} Az_i x_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} Az_i y_i \\ &= (AZ'X)^{-1} AZ'y. \end{aligned}$$

**Proposition 1** *Sous l'hypothèse que 1) les triplets $(x_i, z_i, u_i)$ sont mutuellement indépendants, 2) de même loi, et 3) $\mathrm{E}(u_i^2|z_i) = \sigma^2$, l'estimateur des MCI, $\widehat{b}_{MCI}(A)$, pour tout matrice de sélection $A$, est convergent et asymptotiquement normal :*

$$\sqrt{n} \left( \widehat{b}_{MCI}(A) - b \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma(A)),$$

*avec*

$$\Sigma(A) = \sigma^2 \left[ A\, \mathrm{E}\,(z_i x_i') \right]^{-1} A\, \mathrm{E}\,(z_i z_i')\, A' \left[ \mathrm{E}\,(x_i z_i')\, A' \right]^{-1}.$$

**Preuve :** On a :

$$\begin{aligned}\widehat{b}_{MCI}(A) &= \left(\frac{1}{n}\sum_{i=1}^{n} Az_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} Az_i y_i \\ &= b + \left(\frac{1}{n}\sum_{i=1}^{n} Az_i x_i'\right)^{-1} A\frac{1}{n}\sum_{i=1}^{n} z_i u_i.\end{aligned}$$

La convergence découle simplement de la loi des grands nombres :

$$\frac{1}{n}\sum_{i=1}^{n} z_i u_i \underset{n\longrightarrow\infty}{\longrightarrow} \mathrm{E}\,(z_i u_i) = 0.$$

De plus,

$$\sqrt{n}\left(\widehat{b}_{MCI}(A) - b\right) = \left(\frac{1}{n}\sum_{i=1}^{n} Az_i x_i'\right)^{-1} A\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} z_i u_i.$$

Or, puisque $u_i$ est indépendant des instruments :

$$\begin{aligned}\mathrm{V}\,(z_i u_i) &= \mathrm{E}\,(z_i z_i' u_i^2) \\ &= \mathrm{E}\left[z_i z_i' \mathrm{E}\,(u_i^2 \mid z_i)\right] \\ &= \sigma^2 \mathrm{E}\,(z_i z_i')\end{aligned}$$

et la normalité asymptotique découle directement du théorème limite centrale :

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} z_i u_i \overset{loi}{\longrightarrow} N(0, \sigma^2 \mathrm{E}\, z_i z_i'). \blacksquare$$

**Remarques :**

- Un estimateur de la matrice de variance-covariance asymptotique de $\widehat{b}_{MCI}(A)$ est

$$\begin{aligned}\frac{\widehat{\Sigma}(A)}{n} &= \widehat{\sigma}(A)^2 \left(\sum_{i=1}^{n} Az_i x_i'\right)^{-1} \left(\sum_{i=1}^{n} Az_i z_i' A'\right) \left(\sum_{i=1}^{n} x_i z_i' A'\right)^{-1} \\ &= \widehat{\sigma}(A)^2 (AZ'X)^{-1} AZ'ZA'(X'ZA')^{-1},\end{aligned}$$

où

$$\begin{aligned}\widehat{\sigma}(A)^2 &= \frac{1}{n}\sum_{i=1}^{n} \left(y_i - x_i'\widehat{b}_{MCI}(A)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i(A)^2.\end{aligned}$$

- De même que pour les 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 $z_i$, à savoir

$$\begin{aligned}\frac{\widehat{\Sigma}_{het}(A)}{n} &= \left(\sum_{i=1}^{n} Az_ix_i'\right)^{-1}\left(\sum_{i=1}^{n}\widehat{u}_i(A)^2 Az_iz_i'A'\right)\left(\sum_{i=1}^{n} x_iz_i'A'\right)^{-1} \\ &= (AZ'X)^{-1}AZ'\,\mathrm{diag}[\widehat{u}_i(A)^2]ZA'(X'ZA')^{-1},\end{aligned}$$

où $\mathrm{diag}[\widehat{u}_i(A)^2]$ est la matrice diagonale avec $\widehat{u}_i(A)^2$ en $i$-ième position dans la diagonale et des zéros en dehors de la diagonale.

## 5.5 Estimateur à variables instrumentales optimal ou estimateur des doubles moindres carrés

On vient de définir une classe d'estimateurs convergents et asymptotiquement normaux, celle des estimateurs des MCI pour toute matrice de sélection $A \in \mathbb{R}^{K\times H}$ de plein rang. L'estimateur à variables instrumentales optimal ou estimateur des doubles moindres carrés (2MC) est l'estimateur des MCI de variance minimale.

**Proposition 2** *Il existe une matrice de sélection $A^*$ optimale au sens où $\widehat{b}_{MCI}(A^*)$ est de variance minimale dans la classe des estimateurs $\widehat{b}_{MCI}(A)$ :*

$$A^* = \sum_{i=1}^{n} x_iz_i'\left(\sum_{i=1}^{n} z_iz_i'\right)^{-1} = X'Z(Z'Z)^{-1}.$$

*L'estimateur $\widehat{b}_{MCI}(A^*)$ est appelé* estimateur des doubles moindres carrés *(2MC). Il s'écrit :*

$$\widehat{b}_{2MC} = (X'P_ZX)^{-1}X'P_Zy$$

*où $P_Z = Z(Z'Z)^{-1}Z'$ est le projecteur orthogonal dans l'espace engendré par les colonnes de $Z$.*

**Preuve** : Notons $\gg$ la relation d'ordre sur les matrices symétriques semi définies-positives $A \gg B$ si et seulement si $A - B$ est symétrique semi définie-positive, soit encore : $\forall x$, $x'(A-B)x \geq 0$. On veut montrer que

$$(AZ'X)^{-1}AZ'ZA'(X'ZA')^{-1} \gg (X'P_ZX)^{-1},$$

soit encore, en prenant l'inverse de chaque membre et en inversant le sens de l'inégalité :

$$X'P_Z X \gg X'ZA'(AZ'ZA')^{-1}AZ'X,$$

et en regroupant à gauche les deux membres :

$$X'[P_Z - P_{ZA'}]X \gg 0.$$

où $P_{ZA'} = ZA'(AZ'ZA')^{-1}AZ'$. Cette inégalité est vraie pour tout $A$ puisque $P_Z - P_{ZA'} = P_Z(I_H - P_{ZA'})P_Z$ et puisque $I_H - P_{ZA'}$ est la matrice d'un projecteur orthogonal, donc positive. ■

**Corollaire** *Sous l'hypothèse que 1) les triplets $(x_i, z_i, u_i)$ sont mutuellement indépendants, 2) de même loi, et 3) $\mathbb{E}(u_i^2|z_i) = \sigma^2$, l'estimateur des 2MC est convergent et asymptotiquement normal, i.e. :*

$$\sqrt{n}(\widehat{b}_{2MC} - b) \xrightarrow{loi} N(0, \Sigma^*),$$

*où*

$$\Sigma^* = \sigma^2(X'P_Z X)^{-1}.$$

**Remarques :**

- L'estimateur des Doubles Moindres Carrés $\widehat{b}_{2MC}$ s'appelle ainsi parce qu'il s'obtient en deux étapes de MCO : on régresse d'abord chaque variable explicative endogène sur les instruments, pour obtenir la prévision $\widehat{X} = P_Z X$, puis on régresse $y$ sur $\widehat{X}$. Si $Z$ contient au moins les colonnes de $X_2$, alors $P_Z X_2 = X_2$. On ne projettera donc que le sous ensemble $X_1$ de variables explicatives endogènes sur les instruments.
- Un estimateur de la matrice de variance-covariance asymptotique de $\widehat{b}_{2MC}$ est

$$\begin{aligned} \frac{\widehat{\Sigma}(A)}{n} &= \widehat{\sigma}^2_{2MC}\left(\sum_{i=1}^{n}\widehat{x}_i\widehat{x}_i'\right)^{-1} \\ &= \widehat{\sigma}^2_{2MC}(\widehat{X}'\widehat{X})^{-1}, \end{aligned}$$

où

$$\begin{aligned}\widehat{\sigma}^2_{2MC} &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i'\widehat{b}_{2MC}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i^2.\end{aligned}$$

- Noter bien que en régressant $\widehat{y} = P_Z y$ sur $\widehat{X} = P_Z X$ par MCO, on obtient l'estimateur des 2MC, $\widehat{b}_{2MC} = (\widehat{X}'\widehat{X})^{-1}\widehat{X}'\widehat{y}$, mais une évaluation des écarts-types asymptotiques érronée, à savoir : $\widetilde{\sigma}^2(\widehat{X}'\widehat{X})^{-1}$, où

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - \widehat{x}_i'\widehat{b}_{2MC})^2.$$

  C'est pourquoi il vaut mieux utiliser les programmes de 2MC des logiciels statistiques quand ils existent si l'on veut s'épargner la corrections des écarts-type par le facteur multiplicatif $\widehat{\sigma}_{2MC}/\widetilde{\sigma}$.

- La matrice de variance-covariance asymptotique pour résidus hétéroscédastiques est

$$\begin{aligned}\frac{\widehat{\Sigma}_{het}(A^*)}{n} &= \left(\sum_{i=1}^{n}\widehat{x}_i\widehat{x}_i'\right)^{-1}\left(\sum_{i=1}^{n}\widehat{u}_i^2\widehat{x}_i\widehat{x}_i'\right)\left(\sum_{i=1}^{n}\widehat{x}_i\widehat{x}_i'\right)^{-1} \\ &= (\widehat{X}'\widehat{X})^{-1}\left(\widehat{X}'\,\mathrm{diag}[\widehat{u}_i^2]\widehat{X}\right)(\widehat{X}'\widehat{X})^{-1},\end{aligned}$$

  qui est exactement la matrice de White.

## 5.6 Test d'Hausman d'exogénéité

Un test naturel d'exogénéité est le test d'Hausman (*Econometrica*, 1978). Il se fonde sur la comparaison de $\widehat{b}_{2MC} - \widehat{b}_{MCO}$ avec 0. Calculons cette différence :

$$\begin{aligned}\widehat{b}_{2MC} - \widehat{b}_{MCO} &= \left[(X'P_ZX)^{-1}X'P_Z - (X'X)^{-1}X'\right]y \\ &= \left[(X'P_ZX)^{-1}X'P_Z - (X'X)^{-1}X'\right](Xb + u) \\ &= \left[(X'P_ZX)^{-1}X'P_Z - (X'X)^{-1}X'\right]u.\end{aligned}$$

Sous l'hypothèse nulle d'exogénéité de $X$, si $\mathrm{E}(u_i^2|x_i, z_i) = \sigma^2$,

$$\mathrm{V}\left(\widehat{b}_{2MC} - \widehat{b}_{MCO}|X, Z\right) = \sigma^2\left[(X'P_ZX)^{-1}X'P_Z - (X'X)^{-1}X'\right]$$

$$
\begin{aligned}
& \times \left[(X'P_Z X)^{-1} X'P_Z - (X'X)^{-1}X'\right]' \\
= \; & \sigma^2 \left[(X'P_Z X)^{-1} - (X'X)^{-1}\right] \\
= \; & \mathrm{V}\left(\widehat{b}_{2MC}|X,Z\right) - \mathrm{V}\left(\widehat{b}_{MCO}|X,Z\right).
\end{aligned}
$$

**Proposition 3** *Sous l'hypothèse nulle d'exogénéité de $X$, la statistique*

$$
\frac{1}{\widehat{\sigma}^2}\left(\widehat{b}_{2MC} - \widehat{b}_{MCO}\right)' \left[(X'P_Z X)^{-1} - (X'X)^{-1}\right]^{-} \left(\widehat{b}_{2MC} - \widehat{b}_{MCO}\right)
$$

*suit une loi du $\chi^2$ à $K_1$ degrés de liberté ($K_1$ est le nombre de variables explicatives endogènes).*

**Remarque :**

- $[\cdot]^-$ est l'opérateur "inverse généralisée" : $AA^-A = A$ et $A^-AA^- = A^-$. Si $A$ est de plein rang, on prendra $A^- = A^{-1}$.
- Il peut arriver que la matrice $(X'P_Z X)^{-1} - (X'X)^{-1}$ soit de rang inférieur à $K_1$. Dans ce cas, la statistique d'Hausman converge vers une loi du $\chi^2$ dont le nombre de degrés de liberté est égal au rang de la matrice $(X'P_Z X)^{-1} - (X'X)^{-1}$.

## 5.7 La régression augmentée

Le test d'Hausman d'exogénéité s'implémente très simplement en pratique de la façon suivante. Notons $\widehat{v}_{1i}$ le vecteur des résidus de la régression des variables endogènes $x_{1i}$ sur les instruments $z_i$. Notons aussi $X_1 = (x'_{1i})$ et $\widehat{V}_1 = (\widehat{v}'_{1i})$ :

$$
\widehat{V}_1 = (I - P_Z)X_1 = M_Z X_1,
$$

avec $M_Z = I - P_Z$.

La proposition suivante est alors vraie.

**Proposition 4** *L'estimateur des MCO des paramètres de $x_i$ dans la régression augmentée de $y_i$ sur $x_i$ et $\widehat{v}_{1i}$ est l'estimateur des 2MC. On peut tester l'exogénéité de $x_{1i}$ de façon équivalente au test d'Hausman d'égalité des estimateurs MCO et 2MC en testant la nullité de l'estimateur des MCO de $\widehat{v}_{1i}$ dans la régression augmentée au moyen des tests habituels (Student, Fisher, Wald).*

**Preuve :** L'estimateur des MCO des paramètres de $x_i$ dans cette régression est l'estimateur des Doubles Moindres Carrés : en effet, celui-ci s'obtient en régressant $y_i$ sur la projection orthogonale des $x_i$ sur l'orthogonal de l'espace engendré par les résidus $\widehat{v}_{1i}$ (théorème de Frisch-Waugh) :

$$\begin{aligned}\left[I - M_Z X_1 \left(X_1' M_Z X_1\right)^{-1} X_1' M_Z\right] X &= \left[I - M_Z X_1 \left(X_1' M_Z X_1\right)^{-1} X_1' M_Z\right] (X_1 : X_2) \\ &= ((I - M_Z) X_1 : X_2) \\ &= (P_Z X_1 : X_2) \\ &= P_Z X.\end{aligned}$$

où la seconde égalité provient de ce que $M_Z X_2 = 0$ si le vecteur des instruments contient au moins les variables exogènes $x_{2i}$.

De même l'estimateur des MCO $\widehat{c}$ des paramètres de $\widehat{v}_{1i}$ s'obtient en régressant $y_i$ sur la projection orthogonale des $\widehat{v}_{1i}$ sur l'orthogonal de l'espace engendré par les variables $x_i$ :

$$\begin{aligned}(I - P_X)\widehat{V}_1 &= M_X M_Z X_1 \\ &= M_X X_1 - M_X P_Z X_1 \\ &= -M_X P_Z X_1 \quad \text{(parce que } M_X X_1 = 0).\end{aligned}$$

C'est-à-dire :

$$\widehat{c} = -\left(X_1' P_Z M_X P_Z X_1\right)^{-1} X_1' P_Z M_X y.$$

Or

$$\begin{aligned}(X' P_Z X)\left(\widehat{b}_{2MC} - \widehat{b}_{MCO}\right) &= (X' P_Z X)\left[(X' P_Z X)^{-1} X' P_Z y - (X'X)^{-1} X' y\right] \\ &= X' P_Z y - X' P_Z X (X'X)^{-1} X' y \\ &= X' P_Z (I - P_X) y \\ &= X' P_Z M_X y \\ &= \begin{pmatrix} X_1' P_Z M_X y \\ 0_{K_2 \times 1} \end{pmatrix},\end{aligned}$$

où $0_{K_2 \times 1}$ est le vecteur de zéros de dimension $K_2$, puisque

$$\begin{aligned}X_2' P_Z M_X &= X_2' M_X \\ &= 0.\end{aligned}$$

On voit donc que $\widehat{c}$ est lié bijectivement à $\widehat{b}_{2MC} - \widehat{b}_{MCO}$.[12] Le test d'exogénéité fondé sur la comparaison des estimateurs $\widehat{b}_{2MC}$ et $\widehat{b}_{MCO}$ est donc équivalent au test de $\widehat{c} = 0$. ■

Sous l'hypothèse nulle d'exogénéité de $X_1$, $\widehat{c}$ converge presque sûrement vers $c = 0$. La statistique de Wald associée au test de $c = 0$,

$$\xi = \frac{1}{\widehat{\sigma}^2} \widehat{c}' \, (X_1' P_Z M_X P_Z X_1)^{-1} \, \widehat{c}$$

suit, sous l'hypothèse nulle, une loi du $\chi^2$ à $K_1$ degrés de liberté, où $K_1$ est le nombre de variables explicatives endogènes (i.e. le nombre de colonnes de $X_1$).

**Remarques :**

- Dans la formule précédente $\widehat{\sigma}$ est n'importe quel estimateur convergent sous l'hypothèse nulle de la variance de $u_i$ sachant $X$ et $Z$. On peut prendre en particulier l'estimateur fondé sur les résidus de la régression augmentée.
- Dans le cas $K_1 = 1$, une seule variable potentiellement endogène, la racine carrée de $\xi$ est une statistique de Student suivant asymptotiquement sous la nulle une loi normale $\mathcal{N}(0, 1)$.
- Lorsque $K_1 > 1$ la régression augmentée a l'avantage sur le test d'Hausman de permettre le test d'exogénéité de chaque variable séparément au moyen d'un simple test de Student. Si le coefficient du résidu de la régression instrumentale d'une seule des variables de $X_1$ sort significatif dans la régression augmentée, alors on rejettera l'exogénéité de cette variable seulement.

On a vu que la régression augmentée produisait un estimateur de $b$ exactement identique à l'estimateur des 2MC. Il importe de savoir que, comme pour l'estimateur des 2MC, les écarts-types calculés par les programmes de MCO sont cependant incorrects lorsque les variables $X_1$ sont effectivement endogènes. En effet, en régressant $y_i$ sur $x_i$ et $\widehat{v}_{1i}$, les programmes usuels de MCO produisent le calcul suivant pour l'écart-type de $\widehat{b}_{2MC}$ :

$$\widehat{\omega}^2 (X' P_Z X)^{-1}$$

[12] Dans un espace de dimension $K_1$ ; ce qui signifie que $K_2$ paramètres de $\widehat{b}_{2MC} - \widehat{b}_{MCO}$ sont redondants.

avec

$$\widehat{\omega}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\widehat{b}_{2MC} - \widehat{v}_{1i}'\widehat{c})^2$$

alors qu'un calcul correct, on l'a vu, utilise $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\widehat{b}_{2MC})^2$ à la place de $\widehat{\omega}^2$. Les écarts-types seront donc sous-estimés, et ceci d'autant plus que la corrélation de $x_{1i}$ avec la perturbation $u_i$ sera élevée, ou que les résidus $\widehat{v}_{1i}$ des régressions instrumentales auront un fort pouvoir explicatif dans la régression augmentée. On peut corriger les écarts-types asymptotiques des MCO en les multipliant par le ratio $\widehat{\sigma}/\widehat{\omega}$. Noter que sous l'hypothèse nulle d'exogénéité, cependant, l'erreur est asymptotiquement négligeable puisque $\widehat{c}$ tend vers 0 et donc $\widehat{\omega}^2$ et $\widehat{\sigma}^2$ sont asymptotiquement équivalents.

D'où vient l'erreur ? Une première façon de répondre est de dire que la régression n'est qu'un outil de calcul et n'a pas d'interprétation structurelle. On peut toutefois lui en donner une à condition de donner d'abord à la régression instrumentale des variables explicatives endogènes sur les instruments une interprétation structurelle. Considérons ainsi le modèle à équations simultanées suivant :

$$\begin{aligned} y_i &= x_i'b + u_i, \\ x_i &= Az_i + v_i, \end{aligned}$$

où le fait de poser la régression instrumentale de cette façon donne au résidu $v_i$ le statut d'une variable inobservée. Supposons que

$$\begin{aligned} \mathrm{E}(v_i|z_i) &= 0, \\ \mathrm{E}(u_i|z_i, v_i) &= v_i'c. \end{aligned}$$

On voit que

$$\mathrm{E}(y_i|z_i, v_i) = x_i'b + v_i'c$$

ou

$$y_i = x_i'b + v_i'c + \varepsilon_i$$

avec $\mathrm{E}(\varepsilon_i|z_i, v_i) = 0$. Cette dernière équation indique que le choc $u_i$ peut se décomposer linéairement en une fonction du choc $v_i$ et une partie spécifique à $y_i$, $\varepsilon_i$. Elle ressemble de très près à la régression augmentée. Il y a malgré tout une différence essentielle. Ici $v_i$ est une variable inobservée, une perturbation au même

titre que $u_i$. La remplacer par l'approximation convergente $\widehat{v}_i = x_i - \widehat{A}z_i$, où $\widehat{A}$ est l'estimateur des MCO de $A$ dans la régression linéaire de $x_i$ sur $z_i$, bruite le signal d'une valeur exactement égale à l'écart entre $\widehat{\omega}^2$ et $\widehat{\sigma}^2$ dans le paragraphe précédent. (Noter que $\widehat{\omega}^2$ est maintenant un estimateur convergent de la variance de $\varepsilon_i$ sachant $z_i, v_i$, alors que $\widehat{\sigma}^2$ est un estimateur convergent de la variance de $u_i$ sachant $z_i$.)

## 5.8 Test de validité des variables instrumentales en cas de suridentification

Une autre question naturelle consiste à se demander si les instruments sont bien non corrélés avec les résidus. Lorsque le nombre d'instruments supplémentaires (c'est-à-dire en plus des variables explicatives exogènes) est **strictement supérieur** au nombre de variables endogènes (cas de suridentification[13]), on procèdera au test joint de validité, c'est-à-dire d'exogénéité, des instruments, de la façon suivante. Pour tester si le vecteur $z_i$ des $H$ instruments est exogène, on régressera les résidus $\widehat{u}_i$ sur le vecteur $z_i$, puis on calculera le coefficient de détermination non centré de cette régression ($R^2$). Sous l'hypothèse nulle de non corrélation des perturbations $u_i$ avec les variables $z_i$, $nR^2$ suit une loi du $\chi^2$ à $H - K$ ($= H_1 - K_1$ où $H_1$ est le nombre d'instruments autres que les $K_2$ variables explicatives exogènes).

En effet, ici

$$\begin{aligned} nR^2 &= \frac{\|P_Z\widehat{u}\|^2}{\|\widehat{u}\|^2/n} \\ &= \frac{1}{\widehat{\sigma}^2}\widehat{u}'P_Z\widehat{u} \\ &= \frac{1}{\widehat{\sigma}^2}u'M'P_ZMu \end{aligned}$$

avec $M = I - X(X'P_ZX)^{-1}X'P_Z$. Or, $M$ n'est pas un projecteur mais

$$M'P_ZM = P_Z - P_{P_ZX}$$

est un projecteur orthogonal dans un espace de dimension $H - K$. D'où le résultat.

[13] Il faut au moins autant d'instruments que de variables endogènes pour identifier et estimer le modèle, et calculer ensuite les résidus $\widehat{u}_i$. On ne peut intuitivement espérer tester la validité des instruments strictement nécessaires à ce calcul.

(Si l'on régresse $\widehat{u}_i$ sur $x_i$ au lieu de $z_i$ on obtient la matrice $M'P_{X_1}M$ qui n'est pas un projecteur orthogonal.)

Si le modèle est juste identifié ($H = K$, ou $H_1 = K_1$), alors $nR^2 = 0$ : le résidu de la régression de $\widehat{u}_i$ sur $z_i$ est par construction orthogonal aux $z_i$.

Pratiquement, on régresse $\widehat{u}_i$ sur $z_{1i}$. Si l'estimateur des MCO des coefficients de $z_{1i}$ est significatif, on rejette l'hypothèse d'exogénéité de $z_{1i}$ (i.e. de validité ou d'admissibilité des variables instrumentales). On peut par ailleurs procéder par élimination successive : si certains seulement des coefficients de $z_{1i}$ sont significatifs, on les rejette et on recommence la procédure d'estimation des 2MC à zéro.

On peut aussi, de façon équivalente, régresser les résidus de la régression augmentée $\widehat{\varepsilon}_i$ sur les instruments. En effet

$$\begin{aligned} \widehat{\varepsilon}_i &= y_i - x_i'\widehat{b}_{2MC} - \widehat{v}_{1i}'\widehat{c} \\ &= \widehat{u}_i - \widehat{v}_{1i}'\widehat{c} \end{aligned}$$

et, sous forme matricielle,

$$\begin{aligned} \widehat{\varepsilon} &= \widehat{u} - \widehat{V}_1\widehat{c} \\ &= \widehat{u} - M_Z X_1 \widehat{c}. \end{aligned}$$

L'estimateur des MCO de la régression de $\widehat{\varepsilon}_i$ sur $z_i$ vaut donc

$$\begin{aligned} (Z'Z)^{-1}Z'\widehat{\varepsilon} &= (Z'Z)^{-1}Z'\widehat{u} - (Z'Z)^{-1}Z'M_Z X_1\widehat{c} \\ &= (Z'Z)^{-1}Z'\widehat{u}. \end{aligned}$$

Ce qui prouve que régresser $\widehat{\varepsilon}_i$ ou $\widehat{u}_i$ sur $z_i$ est équivalent.

## 5.9 Résumé de la marche à suivre

Soit la régression :

$$y_i = x_{1i}'b_1 + x_{2i}'b_2 + u_i = x_i'b + u_i \tag{24}$$

où $x_{1i}$ est un vecteur de variables endogènes et $x_{2i}$ est un vecteur de variables explicatives exogènes. Soit $z_{1i}$ un vecteur d'instruments autres que ceux qui composent $x_{2i}$, et de dimension supérieure (ou égale) à celle de $x_{1i}$.

Pratiquement, on procèdera de la façon suivante :

1. **Régression instrumentale :** Régresser (par MCO) $x_{1i}$ sur $z_i = (x_{2i}, z_{1i})$. (On insiste ici encore une fois sur le fait que cette régression instrumentale n'a pas nécessairement d'interprétation structurelle. Il peut ainsi arriver qu'à l'évidence certains instruments dans $x_{2i}$ ne conditionnent pas $x_{1i}$. On peut ne pas les inclure dans la régression instrumentale mais on transforme alors par là-même la regression instrumentale en équation structurelle.)
2. **Régression augmentée et test d'exogénéité :** Calculer $\widehat{v}_{1i}$ le résidu de la régression instrumentale précédente, et régresser (par MCO) $y_i$ sur $x_{1i}$, $x_{2i}$ et $\widehat{v}_{1i}$. On obtient ainsi l'estimateur des doubles moindres carrés $\widehat{b}_{2MC}$. Si le paramètre associé à $\widehat{v}_{1i}$ est significatif, on rejette l'hypothèse d'exogénéité de $x_{1i}$.
3. **Test de validité des instruments :** Dans le cas seulement où le modèle est suridentifié (plus d'instruments que de variables explicatives), régresser (par MCO) $\widehat{u}_i = y_i - x_i'\widehat{b}_{2MC}$ ou les résidus de la régression augmentée sur $z_{1i}$. Si l'estimateur des MCO des coefficients de $z_{1i}$ est significatif, on rejette l'hypothèse d'exogénéité de $z_{1i}$ (de validité ou d'admissibilité des variables instrumentales). On peut procéder par élimination successive : si certains seulement des coefficients de $z_{1i}$ sont significatifs, on les rejette et on recommence la procédure à zéro.

Nombreux sont les logiciels de statistiques qui produisent un calcul de l'estimateur des doubles moindres carrés, la `PROC SYSLIN` de SAS par exemple. Ces logiciels ne fournissent généralement pas de test d'exogénéité ni de test de validité des instruments.

Un programme SAS type enchaîne les ordres suivants :

```
/* DOUBLES MOINDRES CARRES */
/* (pour un calcul correct des écarts-types) */
PROC SYSLIN [DATA = A] ;
ENDOGENOUS Y X1 ;
INSTRUMENTS X2 Z1 ;
MODEL Y = X1 X2 ;
```

```
/* TEST D'EXOGENEITE */
/* 1) Régression instrumentale */
PROC REG ;
MODEL X1 =X2 Z1 ;
OUTPUT OUT = OUT1 R = V ;

/* 2) Régression augmentée */
PROC REG DATA = OUT1 ;
MODEL Y = X1 X2 V ;
OUTPUT OUT = OUT2 R = E ;

/* TEST DE VALIDITE DES INSTRUMENTS */
/* (si suridentification) */
PROC REG DATA = OUT2 ;
MODEL E = Z ;
```

# 6 Modèles apparemment linéaires avec transformations non linéaires de variables explicatives endogènes

On considère ici le cas de modèles du type :

$$y_i = f(x_i)'b + u_i$$

avec $\mathrm{E}\, u_i = 0$ et $\mathrm{E}(u_i|x_i) \neq 0$, et où la variable $x_i$ est continue.

## 6.1 Premier exemple : modèles polynomiaux

Supposons par exemple que $f(x)'b = a+bx+cx^2$. Soit $z$ un instrument de $x$. On peut alors appliquer simplement la méthode des variables instrumentales : on traite chaque puissance de $x$ comme une variable endogène différente et on génère autant d'instruments que nécessaire en prenant les puissances de $z$ correspondantes. Ainsi la méthode de la régression augmentée s'adapte au modèle quadratique de la façon suivante : 1) on régresse d'abord $x_i$ ET $x_i^2$, séparemment, sur $z_i$ et $z_i^2$ pour calculer deux résidus $\widehat{v}_{1i}$ et $\widehat{v}_{2i}$ ; 2) puis on régresse $y_i$ sur $x_i$, $x_i^2$, $\widehat{v}_{1i}$ et $\widehat{v}_{2i}$.

En pratique, on constate que le résidu de la régression instrumentale de $x_i^2$ sur les instruments est souvent non significatif. Ceci s'explique de la façon suivante. Supposons que

$$\begin{aligned} x_i &= z_i'\alpha + v_i, \\ u_i &= \gamma v_i + \varepsilon_i \end{aligned}$$

pour un vecteur $z_i$ d'instruments quelconques, avec $\mathrm{E}(\varepsilon_i|z_i, v_i) = 0$. On obtient par exemple une telle structure de résidus dans le cas où le couple $(u_i, v_i)$ suit une loi normale multidimensionnelle.

On peut donc réécrire la régression structurelle de la façon suivante :

$$y_i = a + bx_i + cx_i^2 + \gamma v_i + \varepsilon_i$$

où $v_i = x_i - z_i'\alpha$ peut être remplacé de façon équivalente par le résidu de la régression de $x_i$ sur $z_i$.

Ainsi, dans le cas d'une régression structurelle d'une variable $y_i$ sur un polynôme d'une variable $x_i$ endogène, si les variables $y_i$ et $x_i$ sont liées par le système

d'équations simultanées précédentes (ou approximativement distribuées selon des lois gaussiennes), alors il suffit pour estimer les paramètres du modèle structurel de façon convergente de procéder à la régression augmentée de $y_i$ sur le polynôme de $x_i$ et le résidu de la régression instrumentale de $x_i$. Noter que comme c'était déjà le cas pour la régression augmentée dans la section précédente (voir §5.7), les écarts-types calculés par le programme de MCO ne tiennent pas compte de ce que l'on a remplacé $v_i$ par une approximation. On peut ici les corriger comme on l'a expliqué au §5.7 en les multipliant par le ratio

$$\frac{\left(\sum_{i=1}^{n}(y_i - \widehat{a} - \widehat{b}x_i - \widehat{c}x_i^2)^2\right)^{1/2}}{\left(\sum_{i=1}^{n}(y_i - \widehat{a} - \widehat{b}x_i - \widehat{c}x_i^2 - \widehat{\gamma}\widehat{v}_i)^2\right)^{1/2}}.$$

## 6.2 Cas d'une variable explicative dichotomique endogène

Considérons un modèle apparemment linéaire dans lequel une variable explicative dichotomique est endogène :

$$y_i = a + bx_i + u_i, \tag{25}$$

avec $x_i \in \{0, 1\}$ et $\mathrm{E}(u_i x_i) \neq 0$. Comme dans le cas précédent, le caractère qualitatif de la variable endogène n'interdit pas d'appliquer la méthode des 2MC. Le fait que l'on ait à régresser $x_i$ sur $z_i$ n'est en rien modifié par la nature qualitative de $x_i$. Il importe de comprendre que la régression instrumentale de première étape des 2MC n'est pas structurelle. Elle n'implique nullement que le vrai modèle de $x_i$ sachant $z_i$ soit un modèle de probabilité linéaire. Les deux régressions de la méthode des 2MC constituent la manière la plus simple de produire l'estimateur des Moments fondés sur la contrainte identifiante : $\mathrm{E}(z_i u_i) = 0$.

Toutefois, on peut alternativement utiliser un modèle PROBIT ou LOGIT à la place de la régression instrumentale si l'on est prêt à faire cette hypothèse structurelle sur la loi de $x_i$ sachant $z_i$. Soit $F(z_i'\widehat{c})$ la valeur prédite de $\mathrm{E}(x_i|z_i) = \Pr\{x_i = 1|z_i\} = F(z_i'c)$, où $F$ est par exemple la fonction de répartition de la loi normale centrée, réduite dans le cas d'un modèle PROBIT. On obtiendra un estimateur convergent des paramètres $a$ et $b$ du modèle apparemment linéaire (25) en effectuant la régression de $y_i$ sur $F(z_i'\widehat{c})$, ou la régression augmentée de $y_i$ sur

$x_i$ et $\widehat{v}_i = x_i - F(z_i'\widehat{c})$. Cette fois, cependant, ces deux estimateurs de $b$ ne seront pas identiques parce que $F(z_i'\widehat{c})$ n'est pas une projection orthogonale.

L'estimateur $\widehat{b}$ de $b$ issu de la régression de $y_i$ sur $F(\widehat{c}z_i)$ par la méthode des MCO est convergent sous l'hypothèse que $\mathrm{E}(u_i|z_i) = 0$. Notons $\mathrm{cov}_e$ et $\mathrm{V}_e$ les covariances et variances empiriques. Alors,

$$\begin{aligned}
\widehat{b} &= \frac{\mathrm{cov}_e(y_i, F(z_i'\widehat{c}))}{\mathrm{V}_e F(z_i'\widehat{c})} \\
&\to \frac{\mathrm{cov}(y_i, F(z_i'c))}{\mathrm{V}\, F(z_i'c)} \\
&= b\frac{\mathrm{cov}(x_i, F(z_i'c))}{\mathrm{V}\, F(z_i'c)} + \frac{\mathrm{cov}(u_i, F(z_i'c))}{\mathrm{V}\, F(z_i'c)}.
\end{aligned}$$

Or,

$$\begin{aligned}
\mathrm{cov}(x_i, F(z_i'c)) &= \mathrm{E}\,[x_i F(z_i'c)] - \mathrm{E}\,x_i\ \mathrm{E}\,F(z_i'c) \\
&= \mathrm{E}\,[\mathrm{E}(x_i|z_i)F(z_i'c)] - \mathrm{E}\,[\mathrm{E}(x_i|z_i)]\,\mathrm{E}\,F(z_i'c) \\
&= \mathrm{E}\,[F(z_i'c)^2] - [\mathrm{E}\,F(z_i'c)]^2 \\
&= \mathrm{V}\,F(cz_i)
\end{aligned}$$

puisque $\mathrm{E}(x_i|z_i) = \Pr\{x_i = 1|z_i\} = F(z_i'c)$. Par ailleurs,

$$\frac{\mathrm{cov}(u_i, F(z_i'c))}{\mathrm{V}\, F(z_i'c)} = 0$$

sous l'hypothèse que $\mathrm{E}(u_i|z_i) = 0$ (hypothèse plus forte que $\mathrm{cov}(u_i, z_i) = 0$).

On voit donc que les hypothèses assurant la convergence de cet estimateur sont plus fortes que celles garantissant la convergence de l'estimateur des 2MC puisqu'il s'agit de spécifier $\mathrm{E}(x_i|z_i)$ et imposer $\mathrm{E}(u_i|z_i) = 0$. Par ailleurs, il reste encore à calculer la variance asymptotique de cet estimateur en deux étapes. L'estimateur $\widehat{b}$ dépend de l'estimateur $\widehat{c}$ de première étape et la variance asymptotique de $\widehat{b}$ dépend de la forme exacte de $\widehat{c}$. Ce calcul dépasse les limites de ce manuel (tout comme l'emploi de techniques de bootstrap qui permettent de se passer du calcul de la variance asymptotique de l'estimateur en deux étapes). Mais il faut cependant savoir que les formules habituelles du calcul de la variance d'un estimateur des MCO, comme pour la régression augmentée, doivent être corrigées pour tenir compte de ce que $c$ est estimé. Il n'y a que sous l'hypothèse nulle d'exogénéité que les écarts types seront correctement calculés. De sorte que le test d'exogénéité

obtenu en testant la significativité de $x_i - F(z_i'\widehat{c})$ dans la régression augmentée se fera comme auparavant. Mais dès l'instant que ce test rejette l'hypothèse nulle, il faudrait en toute rigueur recalculer les écarts types.

Sauf à faire une hypothèse complète sur la loi jointe de $(y_i, x_i)$ sachant $z_i$ et utiliser la méthode du maximum de vraisemblance pour estimer les paramètres, il n'est pas clair du tout que cette méthode, utilisant un modèle PROBIT ou LOGIT au lieu du modèle de probabilité linéaire pour construire une prédiction de $x_i$, ait le moindre avantage. Je ne connais pas en tous cas de simulation de Monte Carlo permettant de se faire une idée des performances comparées de ces deux estimateurs à distance finie et il n'existe pas d'argument théorique permettant de les classer.

## 6.3 Cas d'une variable explicative endogène qualitative ordinale

Supposons par exemple que l'on souhaite régresser les heures travaillées sur le salaire et que l'on observe le salaire en tranches. Soit $y_i$ la variable d'heures. Soit $x_i$ le salaire et soient $x_{1i}, ..., x_{Ki}$ les variables indicatrices indiquant la classe de salaire : $x_{ji} = \mathbf{1}(x_i \in ]s_j, s_{j+1}])$ = 1 si $x_i$ appartient à la $j$ième tranche : $]s_j, s_{j+1}]$, = 0 sinon.

Une première question est celle de savoir quel est le modèle d'intérêt. Est-ce

$$y_i = a + bx_i + u_i \tag{26}$$

ou

$$y_i = a + b_1 x_{1i} + ... + b_K x_{Ki} + u_i ? \tag{27}$$

Si c'est (26) alors on a affaire à un double problème de variable endogène et d'erreur de mesure. D'un autre côté, on peut voir dans le modèle (27) une façon d'approximer une relation non linéaire inconnue entre $y_i$ et $x_i$.

Si le modèle que l'on cherche à estimer est le modèle (27), alors on est dans une situation où les 2MC s'appliquent. Le seul problème est de trouver $K$ instruments (ou $K-1$ puisque $x_{1i} + ... + x_{Ki} = 1$) pour instrumenter les $K$ indicatrices $x_{ji}$. Pourtant, on se dit qu'un seul devrait suffire puisqu'il n'y a au fond qu'une seule variable endogène : $x_i$. Pour fabriquer $K$ instruments à partir d'une seule variable

instrumentale continue, on fera comme pour $x_i$ : on découpera l'instrument en classes, comme on instrumenterait la dépense totale et son carré, dans un système de demande quadratique, par le revenu et son carré.

Supposons maintenant que le modèle à estimer soit (26). Soit $z_i$ un instrument. On peut dans une première étape estimer un modèle PROBIT polytomique ordonné à seuils connus (en supposant $x_i = z_i'\alpha + \sigma v_i$ et $v_i$ normal, centré et réduit). Soit $\widehat{c}$ l'estimateur du modèle PROBIT polytomique ordonné. On estimera $a$ et $b$ dans une seconde étape en régressant par la méthode des MCO $y_i$ sur $\widehat{x}_i = z_i'\widehat{\alpha}$, ou sur $x_i$ et $x_i - \widehat{x}_i$ (la régression augmentée). Comme précedemment, un calcul correct des écarts types de l'estimateur de seconde étape nécessite un développement théorique spécifique qui dépasse le cadre de cette note. Noter ici qu'à la différence de l'estimateur précédent, où il s'agissait d'instrumenter les indicatrices de classe, la première étape de l'estimation (le PROBIT ordonné) implique une hypothèse structurelle sur la loi de $x_i$ sachant $z_i$. On n'est donc plus dans le cadre d'un modèle apparemment linéaire mais dans celui d'un modèle à équations simultanées non linéaire.

# 7 Modèles à variable dépendante qualitative

On considère ici le cas encore plus général de liaisons non linéaires entre $y_i$ et $x_i$ faisant intervenir des paramètres de façon non linéaire.

## 7.1 Cas d'une transformation qualitative d'un modèle apparemment linéaire

J'entends ici le cas suivant. Soit un modèle apparemment linéaire

$$y_i^* = a + bx_i + u_i \tag{28}$$

avec $\mathrm{E}(u_i|x_i) \neq 0$. La variable dépendante $y_i^*$ est une variable latente et l'on observe $y_i = g(y_i^*)$ une transformation non linéaire de cette variable.

Par exemple $y_i^*$ est l'offre Marshallienne de travail, fonction du salaire offert $x_i$ *a priori* corrélé avec l'hétérogénéité des préférences $u_i$, et $y_i$ est soit la décision d'activité :

$$y_i = 1 \text{ si } y_i^* > 0$$

$$= 0 \text{ sinon,}$$

soit le nombre d'heures effectivement travaillées :

$$\begin{aligned} y_i &= y_i^* \text{ si } y_i^* > 0 \\ &= 0 \text{ sinon.} \end{aligned}$$

Nous traiterons cet exemple en détail dans la section 7.1.2.

La technique de la régression augmentée s'applique ici à condition de supposer cette fois (contrairement à ce qui était nécessaire dans le cas de l'estimateur des 2MC) le couple $(x_i, u_i)$ normal conditionnellement aux instruments $z_i$. Supposons en effet que

$$x_i = z_i'\alpha + v_i$$

avec $(u_i, v_i)$ distribués selon une loi normale bidimensionnelle de moyenne $(0, 0)$ et de matrice de variance-covariance

$$\begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix}.$$

On sait alors que

$$\begin{aligned} u_i &= \frac{\rho\sigma_u}{\sigma_v} v_i + \varepsilon_i \\ &= \gamma v_i + \varepsilon_i \quad \text{(disons)} \end{aligned}$$

où $\varepsilon_i$ est une autre variable normale indépendante de $v_i$.

On peut alors réécrire le modèle (28) sous la forme de la régression augmentée suivante :

$$y_i^* = a + bx_i + \gamma v_i + \varepsilon_i$$

où l'on remplacera $v_i$ par le résidu de la régression instrumentale de $x_i$ sur $z_i$. La régression latente peut être augmentée d'autant de résidus de régressions instrumentales qu'il y a de variables explicatives potentiellement endogènes. Noter qu'à nouveau, l'estimation en deux étapes oblige à une correction des écarts-types des estimateurs de seconde étape (ou à bootstrapper). Enfin, cette procédure ne s'applique qu'au cas de variables (explicatives) continues pour lesquelles l'hypothèse de normalité est raisonnable.

### 7.1.1 Doubles moindres carrés non linéaires

La méthode des doubles moindres carrés non linéaires permettent de traiter le cas de modèles du type :

$$y_i = g(x_i, b) + u_i, \quad i = 1, ..., n,$$

avec $x_i \in \mathbb{R}^K$ endogène. Soit $z_i \in \mathbb{R}^H$ un vecteur d'instruments, c'est à dire tel que :

$$\mathrm{E}(z_i u_i) = 0.$$

La méthode des Doubles Moindres Carrés Non Linéaires (2MCNL) est une extension naturelle de l'estimateur des 2MC. On appelle estimateur des doubles moindres carrés non linéaires, la solution au problème :

$$\widehat{b} = \arg\min_b [y - g(X, b)]' P_Z [y - g(X, b)]$$

en notant $y$ le vecteur des $y_i$ et $g(X, b)$ le vecteur dont le $i$ième élément est $g(x_i, b)$. Sous l'hypothèse que $\mathrm{E}(u_i^2 | z_i) = \sigma^2$ indépendant de $z_i$, cet estimateur est un estimateur de Moments Généralisés optimal. Il est convergent et asymptotiquement normal :

$$\sqrt{n}(\widehat{b} - b) \xrightarrow{loi} \mathcal{N}(0, \Sigma)$$

avec

$$\Sigma = \sigma^2 \lim \left[ \frac{1}{n} \frac{\partial g(X, b)'}{\partial b} P_Z \frac{\partial g(X, b)}{\partial b'} \right]^{-1}.$$

Ce problème n'admet pas en général de solution analytique mais on peut assez facilement le résoudre au moyen de la méthode numérique suivante. L'estimateur est solution des conditions du premier ordre :

$$\frac{\partial g(X, \widehat{b})'}{\partial b} P_Z \left[ y - g(X, \widehat{b}) \right] = 0.$$

L'algorithme de Gauss-Newton permet d'obtenir la solution comme la limite $\widehat{b} = \lim b_n$ de l'itération :

$$b_{n+1} - b_n = \left[ \frac{\partial g(X, b_n)'}{\partial b} P_Z \frac{\partial g(X, b_n)}{\partial b'} \right]^{-1} \frac{\partial g(X, b_n)'}{\partial b} P_Z [y - g(X, b_n)].$$

Cette formule a une interprétation intéressante puisqu'il apparaît que $b_{n+1} - b_n$ est exactement l'estimateur des 2MC de $\theta$ dans la régression :

$$y - g(X, b_n) = \frac{\partial g(X, b_n)}{\partial b'} \theta + \varepsilon.$$

où $\frac{\partial g(X,b_n)}{\partial b'}$ est la matrice $n \times K$ telle que l'élément $(i,k)$ (iième ligne, $k$ième colonne) est la dérivée de $g(x_i, b_n)$ par rapport à la $k$ième composante de $b_n$.

### 7.1.2 Illustration d'un cas plus complexe : le modèle d'activité

Nous terminerons cette note par un exemple montrant comment on peut faire dans des cas plus complexes. Considérons pour cela le modèle de participation au marché du travail d'Heckman. Soit $y_i$ le nombre d'heures travaillées par un individu et soit $w_i$ son salaire s'il travaille. On modélise la décision de participation en supposant qu'il existe deux variables latentes $y_i^*$ et $w_i^*$ s'interprétant comme le nombre d'heures $y_i^*$ que l'individu souhaiterait travailler au salaire $w_i^*$ s'il pouvait librement choisir le nombre d'heures ouvrées, dont en particulier des valeurs négatives, lesquelles correspondent à un temps de loisir supérieur à la valeur maximale autorisée (24×7 heures par semaine par exemple).

On supposera que l'individu réalise son idéal si celui-ci est réalisable et ne travaille pas dans tous les autres cas :

$$\begin{aligned} y_i &= y_i^* \quad \text{si } y_i^* > 0 \\ &= 0 \quad \text{sinon.} \end{aligned} \tag{29}$$

Par ailleurs, le salaire observé $w_i$ est lié au salaire offert par la relation :

$$\begin{aligned} w_i &= w_i^* \quad \text{si } y_i^* > 0 \\ &= 0 \quad \text{sinon,} \end{aligned} \tag{30}$$

où l'on a affecté une observation nulle à tout salaire offert non observé.

Enfin, on écrit que

$$y_i^* = a + bw_i^* + cx_i + u_i \tag{31}$$

en supposant que la relation du nombre d'heures optimal au salaire offert $w_i^*$ (ou son logarithme) est linéaire, et où $x_i$ est un vecteur d'autres variables exogènes conditionnant les préférences des agents pour le loisir. Habituellement, $x_i$ contient des informations sur le diplôme de l'individu, ses revenus non salariaux (le salaire du mari s'il s'agit d'un modèle d'activité féminine), des caractéristiques socio-démographiques du ménage.

Quant au salaire offert, on le suppose déterminé linéairement par un vecteur de caractéristiques $z_i$ (education, expérience) :

$$w_i^* = \alpha + \beta z_i + v_i. \tag{32}$$

Et l'on suppose le couple de perturbations $(u_i, v_i)$ (hétérogénéité des goûts, hétérogénéité des qualifications) distribué selon une loi normale bivariée

$$\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \rho\sigma_u\sigma_v \\ \rho\sigma_u\sigma_v & \sigma_v^2 \end{pmatrix}\right).$$

On peut ensuite écrire la vraisemblance jointe de l'échantillon complet $\{(y_i, w_i, x_i, z_i), i = 1, ..., n\}$. C'est bien sûr la méthode la plus efficace. Cependant, certain logiciels, comme SAS, ne permettent pas de programmer facilement du maximum de vraisemblance (contrairement à GAUSS, S+ ou MATLAB ou même STATA) et ne livrent pas un ensemble de procédures permettant d'estimer des modèles à variables qualitatives trés complet. Plaçons nous dans la situation où le seul modèle à variables qualitatives que le logiciel sache estimer soit le modèle dichotomique (modèles PROBIT ou LOGIT). On estimera les paramètres de façon convergente en procédant par étapes de la manière suivante.

En remplaçant $w_i^*$ dans (31) par son expression dans (32), la décision de participation $y_i^* > 0$ se réduit à la condition de participation :

$$a + b\alpha + b\beta z_i + cx_i + bv_i + u_i > 0,$$

où $bv_i + u_i$ est une variable aléatoire normale de moyenne 0 et de variance

$$\begin{aligned} \omega^2 &= b^2 \mathrm{V}\, v_i + \mathrm{V}\, u_i + 2b \operatorname{cov}(v_i, u_i) \\ &= b^2\sigma_v^2 + \sigma_u^2 + 2b\rho\sigma_u\sigma_v \\ &= (b\sigma_v + \rho\sigma_u)^2 + (1 - \rho^2)\sigma_u^2. \end{aligned}$$

Et le modèle (30) prend alors la forme du modèle TOBIT généralisé suivant :

$$\begin{aligned} w_i &= \alpha + \beta z_i + v_i \quad \text{si } a + b\alpha + b\beta z_i + cx_i + bv_i + u_i > 0 \\ &= 0 \quad \text{sinon.} \end{aligned} \tag{33}$$

On peut estimer le modèle (33) par la méthode du maximum de vraisemblance ou par la méthode d'Heckman en deux étapes. Puisque les salaires offerts ne sont

pas observés pour les inactifs, l'idée de l'estimateur en deux étapes consiste à régresser les salaires observés $w_i$ sur les variables explicatives $z_i$ pour le sous-échantillon des actifs seulement.

Or, une régression sert à estimer une moyenne. Calculons donc l'espérance des salaires offerts $w_i$ sachant l'ensemble des variables explicatives exogènes $x_i$ et $z_i$ et la sélection des seuls actifs $(y_i > 0)$ :

$$\begin{aligned} \mathrm{E}(w_i |\ x_i, z_i, y_i > 0\ ) &= \mathrm{E}(w_i^* | x_i, z_i, y_i > 0) \\ &= \alpha + \beta z_i + \tau \cdot \lambda(x_i, z_i) \end{aligned}$$

où $\lambda(x_i, z_i)$ est le "ratio de Mills" et corrige du "biais de sélectivité" :[14]

$$\lambda(x_i, z_i) = \frac{\varphi\left(\frac{a+b\alpha+b\beta z_i+cx_i}{\omega}\right)}{\Phi\left(\frac{a+b\alpha+b\beta z_i+cx_i}{\omega}\right)}$$

et

$$\begin{aligned} \tau &= \frac{\mathrm{cov}(v_i, bv_i + u_i)}{\sqrt{V(bv_i + u_i)}} \\ &= \frac{b\sigma_v^2 + \rho\sigma_u\sigma_v}{\omega}. \end{aligned}$$

La procédure d'estimation de $\alpha$ et $\beta$ en deux étapes consiste à estimer d'abord le modèle PROBIT

$$\begin{aligned} y_i &> 0 \text{ si } a + b\alpha + b\beta z_i + cx_i + bv_i + u_i > 0 \\ &= 0 \text{ sinon,} \end{aligned}$$

[14] Soit $(X_1, X_2)$ un couple de variables aléatoires normales, alors on montre que (voir, par exemple, Gouriéroux, *Econométrie des Variables Qualitatives*, Economica, 1984) :

$$\mathrm{E}(X_1 | X_2 > 0) = \mathrm{E}(X_1) + \frac{\mathrm{cov}(X_1, X_2)}{\sqrt{\mathrm{V}(X_2)}}\lambda$$

où $\lambda$ est le ratio de Mills :

$$\lambda = \varphi\left(\frac{\mathrm{E}(X_1)}{\sqrt{\mathrm{V}(X_2)}}\right) / \Phi\left(\frac{\mathrm{E}(X_1)}{\sqrt{\mathrm{V}(X_2)}}\right)$$

où $\varphi$ et $\Phi$ sont respectivement la densité et la fonction de répartition de la loi normale centrée réduite :

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

La fonction $\Phi$ est en général proposée par les logiciels de statistiques les plus courant (fonction PROBNORM de SAS).

qui fournit une estimation convergente de

$$\frac{a+b\alpha}{\omega}, \frac{c}{\omega} \text{ et } \frac{b\beta}{\omega},$$

et donc du ratio de Mills $\lambda(x_i, z_i)$. Puis on estime $\alpha$, $\beta$ et $\tau$ en appliquant la méthode des MCO à la régression :

$$w_i = \alpha + \beta z_i + \tau\widehat{\lambda}_i + résidu,$$

sur l'échantillon des observations pour lesquelles on dispose d'une observation de salaire, et où

$$\begin{aligned} \tau &= \frac{\text{cov}(v_i, bv_i + u_i)}{\sqrt{V(bv_i + u_i)}} \\ &= \frac{b\sigma_v^2 + \rho\sigma_u\sigma_v}{\omega}. \end{aligned}$$

On obtient ainsi une estimation convergente de $\alpha, \beta$ et $\tau$.

On procède de même pour estimer $a, b$ et $c$. Considérons maintenant les seules observations pour lesquelles $y_i > 0$ et calculons la moyenne des heures $y_i$ sachant les exogènes pour les seuls actifs :

$$\begin{aligned} \mathrm{E}(y_i|x_i, z_i, y_i > 0) &= \mathrm{E}(y_i^*|x_i, z_i, y_i > 0) \\ &= a + b\,\mathrm{E}(w_i|x_i, z_i, y_i > 0) + cx_i + \mathrm{E}(u_i|x_i, z_i, y_i > 0) \\ &= a + b(\alpha + \beta z_i) + cx_i + \mathrm{E}(bv_i + u_i|x_i, z_i, y_i > 0) \\ &= a + b(\alpha + \beta z_i) + cx_i + \omega \cdot \lambda(x_i, z_i). \end{aligned}$$

On procèdera donc comme pour les salaires, en régressant $y_i$ sur la prédiction $\widehat{w}_i = \widehat{\alpha} + \widehat{\beta} z_i$ du salaire, les variables $x_i$ et le ratio de Mills estimé :

$$y_i = a + b\widehat{w}_i + cx_i + \omega\widehat{\lambda}_i + résidu,$$

sur l'échantillon des agents actifs. On obtient ainsi une estimation convergente de $a, b, c$ et $\omega$.

Pour finir, il convient de préciser que cette méthode d'estimation en plusieurs étapes produit donc une estimation convergente des paramètres du modèle structurel mais que les écarts-types que les logiciels usuels calculent au cours des différentes étapes de PROBIT ou de MCO ne sont en aucun cas une estimation convergente de la précision de ces estimateurs.

En l'absence d'observations de salaires offerts aux inactifs, on ne peut pas proposer de méthode de régression augmentée comme on l'avait fait dans le cadre de modèles apparemment linéaires, permettant de tester simplement l'exogénéité du salaire dans l'équation d'heures. Par ailleurs, la procédure en plusieurs étapes précédente ne s'applique plus dès lors que la dépendance des heures $y_i^*$ au salaire offert $w_i^*$ (ou son logarithme) n'est plus linéaire, quadratique par exemple. Il faut alors recourir à la méthode du maximum de vraisemblance.

## 8. Illustration

### *8.1. Estimation d'une équation d'heures ouvrées par MCO*

On utilise ici le sous-échantillon de l'enquête « Emploi » de 1998, composé des femmes inactives ou actives salariées, vivant en couple avec un salarié occupé, pour estimer par MCO l'équation :

$$NBHSEM = a + b_1 LSAL + b_2 LSAL^2 + c_1 LSALH + c_2 LSALH^2 + c_3 ENF6 + c_4 FINDET + c_5 FINDET^2 + u \quad (1)$$

où *NBHSEM* est la variable de nombre d'heures travaillées habituellement par semaine, *LSAL* est le log du salaire horaire de la femme, *LSALH* celui de l'homme, *ENF6* est le nombre d'enfants de moins de 6 ans, et *FINDET* est l'âge de fin d'études.

**Le programme SAS**

```
proc reg data=exlc;
model nbhsem=lsal lsal2 enf6 lsalh lsalh2 findetud findet2;
output out=est1 p=nbhest1;
title "modèle descriptif d'offre de travail estimé par les MCO";
run;
```

**La sortie SAS**

```
Dependent Variable: NBHSEM
Analysis of Variance

                           Sum of          Mean
Source              DF     Squares        Square     F Value     Prob>F

Model                7  39033.38940   5576.19849      54.802     0.0001
Error            15616 1588957.0558 101.75186064
C Total          15623 1627990.4452

    Root MSE       10.08721     R-square       0.0240
    Dep Mean       33.77125     Adj R-sq       0.0235
    C.V.           29.86923

Parameter Estimates

                   Parameter       Standard     T for H0:
Variable  DF        Estimate          Error    Parameter=0    Prob > |T|

INTERCEP   1      10.914649     2.47218110          4.415        0.0001
LSAL       1       2.698435     0.62832935          4.295        0.0001
LSAL2      1      -0.668246     0.08402993         -7.952        0.0001
ENF6       1      -1.001221     0.16477256         -6.076        0.0001
LSALH      1       1.085611     0.54036494          2.009        0.0446
LSALH2     1      -0.172267     0.06790138         -2.537        0.0112
FINDETUD   1       1.834048     0.18886389          9.711        0.0001
FINDET2    1      -0.036423     0.00472495         -7.709        0.0001
```

Le signe de *LSAL2* étant négatif, l'effet du salaire est négatif pour toute valeur de *LSAL* supérieure au seuil qui annule la dérivée de la fonction $2{,}698 * LSAL - 0{,}668 * LSAL^2$, soit un salaire horaire égal à 2.698/(2*0,668) = 7.53, valeur inférieure au plus petit salaire de

l'échantillon. On conclurait donc à un effet fortement dépressif du salaire offert sur la participation féminine.

### *8.2. Instrumentation du salaire dans l'équation d'heures*

On suppose maintenant que le salaire offert est endogène dans l'équation d'heures ouvrées et satisfait la régression suivante :

$$LSAL = \alpha + \beta_1 EXPE + \beta_2 EXPE^2 + \beta_3 EXPE * FINDET + \beta_4 FINDET^2 + \beta_5 FINDET + \beta_6 ENF6 + \nu, \quad (2)$$

où *EXPE* (l'expérience, i.e. l'âge moins l'âge de fin d'étude, *FINDET*), $EXPE^2$, l'expérience au carré, et *EXPE*FINDET* sont les instruments spécifiques de *LSAL*. On suppose en effet que l'âge (et l'expérience) n'ont d'effet sur les préférences en matière de loisir que par le lien implicite avec le nombre d'enfants.

On estime l'équation d'heures par la technique de la régression augmentée décrite dans la première partie de ce document. Pour faire simple, nous n'augmenterons la régression des heures que du résidu de la régression de *LSAL* et en omettant celui de la régression instrumentale de *LSAL2*. On a vu à quelles conditions ceci était légitime.

**Le programme SAS**

```
proc reg data=exlc ;
model lsal= expe expe2 enf6 findetud findet2 expfind ;
output out=est  r=vchap;
title "Estimation du salaire par les MCO";
run;

proc reg ;
model nbhsem=lsal lsal2 enf6 lsalh lsalh2 findetud findet2 vchap ;
title "équation d'offre de travail : régression augmentée après
instrumentation du salaire";
output out=est2 p=nbhest2 r=uchap;
run;

proc reg ;
model uchap=expe expe2 expfind;
title "pour un test d'exogénéité des instruments";
run;
```

**La sortie SAS**

### 1. Equation de salaire

Dependent Variable: LSAL
Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Prob>F |
|---|---|---|---|---|---|
| Model | 6 | 934.56670 | 155.76112 | 848.912 | 0.0001 |
| Error | 15824 | 2903.43862 | 0.18348 | | |
| C Total | 15830 | 3838.00532 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.42835 | R-square | 0.2435 |
| Dep Mean | 3.81862 | Adj R-sq | 0.2432 |
| C.V. | 11.21740 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| |
|---|---|---|---|---|---|
| INTERCEP | 1 | 1.349400 | 0.15907691 | 8.483 | 0.0001 |
| EXPE | 1 | 0.012992 | 0.00447278 | 2.905 | 0.0037 |
| EXPE2 | 1 | -0.000170 | 0.00004179 | -4.059 | 0.0001 |
| ENF6 | 1 | 0.066790 | 0.00732606 | 9.117 | 0.0001 |
| FINDETUD | 1 | 0.138228 | 0.01270824 | 10.877 | 0.0001 |
| FINDET2 | 1 | -0.001516 | 0.00026028 | -5.823 | 0.0001 |
| EXPFIND | 1 | 0.000672 | 0.00016570 | 4.053 | 0.0001 |

### 2. Régression augmentée de l'équation d'offre de travail

Dependent Variable: NBHSEM

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Prob>F |
|---|---|---|---|---|---|
| Model | 8 | 41579.11708 | 5197.38963 | 51.158 | 0.0001 |
| Error | 15615 | 1586411.3281 | 101.59534602 | | |
| C Total | 15623 | 1627990.4452 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 10.07945 | R-square | 0.0255 |
| Dep Mean | 33.77125 | Adj R-sq | 0.0250 |
| C.V. | 29.84625 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| |
|---|---|---|---|---|---|
| INTERCEP | 1 | 4.189153 | 2.81201289 | 1.490 | 0.1363 |
| LSAL | 1 | 5.690484 | 0.86686946 | 6.564 | 0.0001 |
| LSAL2 | 1 | -0.721236 | 0.08462994 | -8.522 | 0.0001 |
| ENF6 | 1 | -0.943795 | 0.16504496 | -5.718 | 0.0001 |
| LSALH | 1 | 1.060902 | 0.53997175 | 1.965 | 0.0495 |
| LSALH2 | 1 | -0.199023 | 0.06805935 | -2.924 | 0.0035 |
| FINDETUD | 1 | 1.615719 | 0.19369312 | 8.342 | 0.0001 |
| FINDET2 | 1 | -0.034534 | 0.00473638 | -7.291 | 0.0001 |
| VCHAP | 1 | -2.873715 | 0.57408285 | -5.006 | 0.0001 |

On voit maintenant que l'effet du salaire sur le nombre d'heures ouvrées est positif pour tout salaire inférieur à exp(5.690484/(2*0.721236)) = 51,67 Francs de l'heure, et négatif sinon. Cette valeur de 51,67F/h est proche du salaire médiant.

Nous avons aussi effectué l'estimation de ce modèle sans le terme quadratique *LSAL2*. L'effet estimé du salaire sur le nombre d'heure est non significatif, l'effet positif compensant exactement l'effet négatif. Il paraît donc particulièrement important de modéliser un effet non linéaire du salaire offert sur le nombre d'heures ouvrées par semaine. Sinon on estime une élasticité des heures au salaire non significative.

**3. Test de validité des instruments**

Dependent Variable: UCHAP Residual
Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Prob>F |
|---|---|---|---|---|---|
| Model | 3 | 205.71468 | 68.57156 | 0.675 | 0.5671 |
| Error | 15620 | 1586205.6135 | 101.54965515 | | |
| C Total | 15623 | 1586411.3281 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 10.07718 | R-square | 0.0001 |
| Dep Mean | -0.00000 | Adj R-sq | -0.0001 |
| C.V. | -1.103351E16 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| |
|---|---|---|---|---|---|
| INTERCEP | 1 | -0.090325 | 0.27722068 | -0.326 | 0.7446 |
| EXPE | 1 | 0.052827 | 0.04484814 | 1.178 | 0.2388 |
| EXPE2 | 1 | -0.000503 | 0.00075566 | -0.666 | 0.5057 |
| EXPFIND | 1 | -0.002002 | 0.00144101 | -1.390 | 0.1647 |

On constate donc que les variables d'expérience n'ont pas d'effet direct sur la participation. Ce sont donc des instruments valides.

### *8.3. Estimation de l'effet du travail à temps partiel sur les salaires par MCO*

On s'intéresse maintenant au cas d'une variable explicative qualitative.

Soit le cas d'une équation de salaire dans laquelle on a introduit une variable de temps partiel :

$$LSAL = \alpha + \beta_1 EXPE + \beta_2 EXPE^2 + \beta_3 EXPE * FINDET + \beta_4 FINDET + \beta_5 FINDET^2 + \beta_6 ENF6 + \beta_7 TP + u \quad (3)$$

où *TP* est une variable indiquant s'il s'agit d'un emploi à temps partiel ou non.

### Le programme SAS

```
proc reg data=trav;
model lsal= tpsp expe expe2 findetud findet2 expfind;
output out=est1 p=lsalest1;
title "équation de salaire brute de décoffrage";
run;
```

### La sortie SAS

```
Dependent Variable: LSAL

Analysis of Variance

                           Sum of          Mean
Source                DF   Squares        Square      F Value      Prob>F

Model                  6   957.02271    159.50378      876.086      0.0001
Error              15824   2880.98261      0.18206
C Total            15830   3838.00532

    Root MSE        0.42669     R-square       0.2494
    Dep Mean        3.81862     Adj R-sq       0.2491
    C.V.           11.17394

Parameter Estimates

                   Parameter       Standard    T for H0:
Variable  DF        Estimate          Error   Parameter=0    Prob > |T|

INTERCEP   1       1.462918     0.15874690         9.215        0.0001
TPSP       1      -0.104268     0.00724529       -14.391        0.0001
EXPE       1       0.014951     0.00444576         3.363        0.0008
EXPE2      1      -0.000222     0.00004135        -5.371        0.0001
FINDETUD   1       0.133754     0.01266514        10.561        0.0001
FINDET2    1      -0.001423     0.00025940        -5.487        0.0001
EXPFIND    1       0.000604     0.00016463         3.671        0.0002
```

On voit ici que les femmes qui travaillent à temps partiel ont tendance à percevoir des salaires horaires environ 10% plus faibles que celles qui travaillent à temps complet.

### *8.4. Instrumentation de la variable de temps partiel*

On se dit que la décision de travailler à temps partiel peut être endogène. On instrumente donc cette variable par le nombre d'enfants du couple (*ENFC90*) et la variation d'une année sur l'autre du taux de chômage des femmes entre 20 et 60 ans du département (*VARCHO*):

(4) $$TP = a + b_1 ENFC90 + b_2 VARCHO + v$$

### Le programme SAS

```
proc reg data=trav;
model tpsp= enfc90 varcho;
output out=est p=tpspest r=vchap;
title "Estimation du temps partiel par les MCO";
run;

proc reg ;
model lsal=tpsp expe expe2 findetud findet2 expfind vchap ;
```

```
title "équation de salaire: régression augmentée après instrumentation du
salaire";
output out=est2 p=lsalest2 r=uchap;
run;

proc reg ;
model uchap= enfc90 varcho;
title "pour un test d'exogénéité des instruments";
run;
```

## La sortie SAS

### 1. Régresser le temps partiel par les MCO sur les instruments

Dependent Variable: TPSP

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Prob>F |
|---|---|---|---|---|---|
| Model | 2 | 110.15803 | 55.07902 | 255.361 | 0.0001 |
| Error | 15983 | 3447.38282 | 0.21569 | | |
| C Total | 15985 | 3557.54085 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.46443 | R-square | 0.0310 |
| Dep Mean | 0.33429 | Adj R-sq | 0.0308 |
| C.V. | 138.92774 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| |
|---|---|---|---|---|---|
| INTERCEP | 1 | 0.228835 | 0.00606589 | 37.725 | 0.0001 |
| ENFC90 | 1 | 0.079043 | 0.00351872 | 22.463 | 0.0001 |
| VARCHO | 1 | 0.000579 | 0.00021139 | 2.740 | 0.0062 |

### 2. Régression augmentée de l'équation de salaire

Dependent Variable: LSAL

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Prob>F |
|---|---|---|---|---|---|
| Model | 7 | 967.36192 | 138.19456 | 761.729 | 0.0001 |
| Error | 15823 | 2870.64340 | 0.18142 | | |
| C Total | 15830 | 3838.00532 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.42594 | R-square | 0.2520 |
| Dep Mean | 3.81862 | Adj R-sq | 0.2517 |
| C.V. | 11.15422 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| |
|---|---|---|---|---|---|
| INTERCEP | 1 | 1.491750 | 0.15851282 | 9.411 | 0.0001 |
| TPSP | 1 | -0.435508 | 0.04446983 | -9.793 | 0.0001 |
| EXPE | 1 | 0.022420 | 0.00454688 | 4.931 | 0.0001 |
| EXPE2 | 1 | -0.000340 | 0.00004413 | -7.701 | 0.0001 |
| FINDETUD | 1 | 0.136020 | 0.01264635 | 10.756 | 0.0001 |
| FINDET2 | 1 | -0.001420 | 0.00025895 | -5.483 | 0.0001 |
| EXPFIND | 1 | 0.000460 | 0.00016544 | 2.783 | 0.0054 |
| VCHAP | 1 | 0.340922 | 0.04516028 | 7.549 | 0.0001 |

L'effet du temps partiel est donc beaucoup plus fort qu'il n'apparaissait précédemment. Les salaires horaires des employés à temps partiel sont en réalité inférieurs d'un bon tiers aux salaires des temps plein.

**3. Test de validité des instruments**

Dependent Variable: UCHAP Residual

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Prob>F |
|---|---|---|---|---|---|
| Model | 2 | 0.16108 | 0.08054 | 0.444 | 0.6414 |
| Error | 15828 | 2870.48231 | 0.18135 | | |
| C Total | 15830 | 2870.64340 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.42586 | R-square | 0.0001 |
| Dep Mean | 0.00000 | Adj R-sq | -0.0001 |
| C.V. | 7.9968477E15 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| |
|---|---|---|---|---|---|
| INTERCEP | 1 | 0.001073 | 0.00559206 | 0.192 | 0.8479 |
| ENFC90 | 1 | -0.000336 | 0.00324887 | -0.103 | 0.9177 |
| VARCHO | 1 | 0.000182 | 0.00019469 | 0.935 | 0.3496 |

### *8.5. Le modèle de participation d'Heckman avec une fonction d'heures ouvrées linéaire par rapport au log du salaire offert*

On estime ici trois modèles : un modèle PROBIT réduit de participation, une équation de salaire avec correction du biais de sélection par le ratio de Mills, et une équation d'heure avec instrumentation du salaire et correction du biais de sélection ; l'instrumentation se faisant ici en remplaçant le salaire par sa prédiction issue de l'étape précédente.

## Le programme SAS

```
/****** Forme réduite de l'équation de participation ********/

 proc logistic data=ex3 ;
 model exprof = expe expe2 expfind
                findetud findet2 lsalh lsalh2 enf6
                   / link=normit maxiter=100;
  output out=mills xbeta=xb;
title "Equation de participation";
  run;

data mills;set mills; /*** calcul du ratio de Mills ****/
ratio=(1/sqrt(8*atan(1)))*exp(-xb*xb/2)/probnorm(xb);
run;

proc reg data=mills;
model lsal = expe expe2 expfind enf6 findetud findet2 ratio;
output out=etape2 p=lsalest r=vchap;
title "équation de salaire corrigée du biais de sélection";
run;

data final;/*** forme structurelle de l'équation d'offre ****/
set etape2;
if nbhsem>0;
 proc reg ;
 model nbhsem = lsalest /*** prédiction du log du salaire ***/
                findetud findet2 lsalh lsalh2 enf6 ratio;
 title "Equation d'offre de travail corrigée du biais de sélection";
  run;
```

## La sortie SAS

### 1. L'équation de participation.

```
The LOGISTIC Procedure
Data Set: WORK.EX3
Response Variable: EXPROF
Response Levels: 2
Number of Observations: 23977
Link Function: Normit

     Response Profile

Ordered
  Value  EXPROF     Count

      1       1     15762
      2       2      8215

WARNING: 367 observation(s) were deleted due to missing values for the response or
         explanatory variables.

   Model Fitting Information and Testing Global Null Hypothesis BETA=0

                          Intercept
            Intercept         and
Criterion     Only       Covariates    Chi-Square for Covariates

AIC          30824.819     28870.967          .
SC           30832.904     28943.731          .
-2 LOG L     30822.819     28852.967     1969.852 with 8 DF (p=0.0001)
Score              .             .       1953.533 with 8 DF (p=0.0001)
```

Analysis of Maximum Likelihood Estimates

| Variable | DF | Parameter Estimate | Standard Error | Wald Chi-Square | Pr > Chi-Square | Standardized Estimate |
|---|---|---|---|---|---|---|
| INTERCPT | 1 | -6.3563 | 0.4099 | 240.4340 | 0.0001 | . |
| EXPE | 1 | 0.1823 | 0.0106 | 295.7147 | 0.0001 | 1.993230 |
| EXPE2 | 1 | -0.00199 | 0.000096 | 433.1858 | 0.0001 | -0.950661 |
| EXPFIND | 1 | -0.00530 | 0.000405 | 171.0038 | 0.0001 | -0.916350 |
| FINDETUD | 1 | 0.4759 | 0.0323 | 217.4757 | 0.0001 | 1.513257 |
| FINDET2 | 1 | -0.00739 | 0.00066 | 117.9126 | 0.0001 | -0.918749 |
| LSALH | 1 | 0.1035 | 0.0618 | 2.8047 | 0.0940 | 0.066744 |
| LSALH2 | 1 | -0.0261 | 0.00768 | 11.5928 | 0.0007 | -0.095617 |
| ENF6 | 1 | -0.4745 | 0.0158 | 898.1601 | 0.0001 | -0.280569 |

Association of Predicted Probabilities and Observed Responses

| | | | |
|---|---|---|---|
| Concordant | = 66.1% | Somers' D | = 0.326 |
| Discordant | = 33.5% | Gamma | = 0.327 |
| Tied | = 0.4% | Tau-a | = 0.147 |
| (129484830 pairs) | | c | = 0.663 |

## 2. L'équation de salaire corrigée du biais de sélection

Dependent Variable: LSAL
Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Prob>F |
|---|---|---|---|---|---|
| Model | 7 | 1017.74397 | 145.39200 | 823.272 | 0.0001 |
| Error | 15616 | 2757.82588 | 0.17660 | | |
| C Total | 15623 | 3775.56984 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.42024 | R-square | 0.2696 |
| Dep Mean | 3.81847 | Adj R-sq | 0.2692 |
| C.V. | 11.00548 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| |
|---|---|---|---|---|---|
| INTERCEP | 1 | -9.420361 | 0.47551299 | -19.811 | 0.0001 |
| EXPE | 1 | 0.261973 | 0.01126283 | 23.260 | 0.0001 |
| EXPE2 | 1 | -0.002861 | 0.00011933 | -23.971 | 0.0001 |
| EXPFIND | 1 | -0.006732 | 0.00034880 | -19.301 | 0.0001 |
| ENF6 | 1 | -0.593032 | 0.02832224 | -20.939 | 0.0001 |
| FINDETUD | 1 | 0.808933 | 0.03068675 | 26.361 | 0.0001 |
| FINDET2 | 1 | -0.012390 | 0.00052282 | -23.698 | 0.0001 |
| RATIO | 1 | 2.556381 | 0.10508258 | 24.098 | 0.0001 |

Noter la forte augmentation du rendement des études et de l'expérience lorsque l'on corrige l'estimation de l'équation de salaire du biais de sélection.

### 3. L'équation d'offre de travail corrigée du biais de sélection

Dependent Variable: NBHSEM

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Prob>F |
|---|---|---|---|---|---|
| Model | 7 | 21323.18073 | 3046.16868 | 29.621 | 0.0001 |
| Error | 15656 | 1610036.9847 | 102.83833577 | | |
| C Total | 15663 | 1631360.1655 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 10.14092 | R-square | 0.0131 |
| Dep Mean | 33.77375 | Adj R-sq | 0.0126 |
| C.V. | 30.02605 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| |
|---|---|---|---|---|---|
| INTERCEP | 1 | 22.770182 | 3.46032175 | 6.580 | 0.0001 |
| LSALEST | 1 | -0.041532 | 0.60022573 | -0.069 | 0.9448 |
| FINDETUD | 1 | 1.237913 | 0.22705312 | 5.452 | 0.0001 |
| FINDET2 | 1 | -0.027497 | 0.00523584 | -5.252 | 0.0001 |
| LSALH | 1 | 0.896436 | 0.56132599 | 1.597 | 0.1103 |
| LSALH2 | 1 | -0.239232 | 0.07651549 | -3.127 | 0.0018 |
| ENF6 | 1 | -0.166692 | 0.27738361 | -0.601 | 0.5479 |
| RATIO | 1 | -3.330507 | 0.92176785 | -3.613 | 0.0003 |

Le salaire ne sort donc pas significatif dans l'équation d'heures.

### *8.6. Cas d'un équation d'heure avec effet non linéaire du log du salaire offert*

Dans le cas où le salaire apparaît non linéairement dans l'équation d'heures, sous une forme quadratique par exemple, on ne peut comme on l'a fait précédemment simplement substituer la prédiction du log du salaire (et son carré) en lieu et place du log du salaire observé (et de son carré). Il n'y a malheureusement pas d'autre moyen de procéder dans ce cas que par l'application de la méthode du maximum de vraisemblance, méthode que SAS ne permet pas d'appliquer simplement. Il faut alors recourrir à d'autres logiciels plus complets en algorithmes statistiques comme GAUSS ou MATLAB.

### *8.7. Test de suridentification : une erreur à ne pas faire*

**Le test de suridentification pour valider les instruments n'est pas valable lorsqu'il y a juste identification (un instrument pour une variable).**

**Exemple 1**

A titre d'exemple, reprenons la spécification de la section 6.4. On instrumente TPSP par ENF6 uniquement (exemple 1). On pourrait croire que ENF6 est un bon instrument (ie exogène) en lisant la 3ème étape. Or si le student vaut 0, c'est que ENF6 est orthogonal au résidu, par construction.

**Le programme :**

```
proc reg data=trav;
model tpsp= enf6;
output out=est p=tpspest r=vchap;
title "Estimation du temps partiel par les MCO";
run;

proc reg ;
model lsal=tpsp expe expe2 findetud findet2 expfind
                vchap ;
title "équation de salaire: régression augmentée après instrumentation
du salaire";
output out=est2 p=lsalest2 r=uchap;
run;

proc reg ;
model uchap= enf6;
title "pour un test d'exogénéité des instruments";
run;
```

**La sortie SAS :**

```
Dependent Variable: TPSP
Analysis of Variance

                           Sum of          Mean
Source           DF       Squares        Square       F Value        Prob>F

Model             1      10.67384      10.67384        48.102        0.0001
Error         15984    3546.86701       0.22190
C Total       15985    3557.54085

    Root MSE        0.47106     R-square        0.0030
    Dep Mean        0.33429     Adj R-sq        0.0029
    C.V.          140.91365

Parameter Estimates

                  Parameter       Standard    T for H0:
Variable  DF       Estimate          Error   Parameter=0     Prob > |T|

INTERCEP   1       0.321746     0.00414169         77.685         0.0001
ENF6       1       0.051336     0.00740194          6.936         0.0001

équation de salaire: régression augmentée après instrumentation du
salaire

Model: MODEL1
Dependent Variable: LSAL

Analysis of Variance

                           Sum of          Mean
Source           DF       Squares        Square       F Value        Prob>F

Model             7     976.05368     139.43624        770.907        0.0001
Error         15823    2861.95164       0.18087
C Total       15830    3838.00532

    Root MSE        0.42529     R-square        0.2543
    Dep Mean        3.81862     Adj R-sq        0.2540
    C.V.          11.13732

Parameter Estimates

                  Parameter       Standard    T for H0:
Variable  DF       Estimate          Error   Parameter=0     Prob > |T|
```

```
INTERCEP     1       1.038940      0.16353632          6.353          0.0001
TPSP         1       1.347556      0.14172118          9.509          0.0001
EXPE         1       0.011849      0.00444149          2.668          0.0076
EXPE2        1      -0.000173      0.00004149         -4.172          0.0001
FINDETUD     1       0.128813      0.01263282         10.197          0.0001
FINDET2      1      -0.001364      0.00025862         -5.274          0.0001
EXPFIND      1       0.000734      0.00016457          4.461          0.0001
VCHAP        1      -1.457216      0.14206268        -10.258          0.0001

pour un test d'exogénéité des instruments

Model: MODEL1
Dependent Variable: UCHAP       Residual

Analysis of Variance

                          Sum of          Mean
Source                DF  Squares        Square      F Value      Prob>F

Model                  1        0             0        0.000      1.0000
Error              15829  2861.95164    0.18080
C Total            15830  2861.95164

    Root MSE       0.42521     R-square       0.0000
    Dep Mean       0.00000     Adj R-sq      -0.0001
    C.V.      2.8141782E15

Parameter Estimates

                   Parameter       Standard    T for H0:
Variable  DF        Estimate          Error   Parameter=0    Prob > |T|

INTERCEP   1  -3.98846E-13     0.00375742         -0.000        1.0000
ENF6       1  1.691619E-12     0.00671144          0.000        1.0000
```

**Exemple 2**

En introduisait un deuxième instrument (ce qui permet de suridentifier le modèle), on découvre que ENF6 n'est pas un bon instrument (exemple 2).

**Le programme**

```
proc reg data=trav;
model tpsp=enf3 enf3_6;
output out=est p=tpspest r=vchap;
title "Estimation du temps partiel par les MCO";
run;

proc reg ;
model lsal=tpsp expe expe2 findetud findet2 expfind vchap ;
title "équation de salaire: régression augmentée après instrumentation
du salaire";
output out=est2 p=lsalest2 r=uchap;
run;
```

```
proc reg ;
model uchap= enf3 enf3_6;
title "pour un test d'exogénéité des instruments";
run;
```

## La sortie SAS

*Estimation du temps partiel par les MCO*
Dependent Variable: TPSP

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Prob>F |
|---|---|---|---|---|---|
| Model | 2 | 12.18395 | 6.09198 | 27.464 | 0.0001 |
| Error | 15983 | 3545.35689 | 0.22182 | | |
| C Total | 15985 | 3557.54085 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.47098 | R-square | 0.0034 |
| Dep Mean | 0.33429 | Adj R-sq | 0.0033 |
| C.V. | 140.88806 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| |
|---|---|---|---|---|---|
| INTERCEP | 1 | 0.321668 | 0.00414105 | 77.678 | 0.0001 |
| ENF3 | 1 | 0.026812 | 0.01196315 | 2.241 | 0.0250 |
| ENF3_6 | 1 | 0.069776 | 0.01023287 | 6.819 | 0.0001 |

*Equation de salaire: régression augmentée après instrumentation du salaire*

Model: MODEL1
Dependent Variable: LSAL

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Prob>F |
|---|---|---|---|---|---|
| Model | 7 | 969.82519 | 138.54646 | 764.325 | 0.0001 |
| Error | 15823 | 2868.18013 | 0.18127 | | |
| C Total | 15830 | 3838.00532 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.42575 | R-square | 0.2527 |
| Dep Mean | 3.81862 | Adj R-sq | 0.2524 |
| C.V. | 11.14943 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| |
|---|---|---|---|---|---|
| INTERCEP | 1 | 1.144935 | 0.16285517 | 7.030 | 0.0001 |
| TPSP | 1 | 0.992065 | 0.13065323 | 7.593 | 0.0001 |
| EXPE | 1 | 0.012101 | 0.00444895 | 2.720 | 0.0065 |
| EXPE2 | 1 | -0.000175 | 0.00004164 | -4.200 | 0.0001 |
| FINDETUD | 1 | 0.130277 | 0.01264413 | 10.303 | 0.0001 |
| FINDET2 | 1 | -0.001385 | 0.00025887 | -5.352 | 0.0001 |
| EXPFIND | 1 | 0.000704 | 0.00016469 | 4.274 | 0.0001 |
| VCHAP | 1 | -1.100728 | 0.13097602 | -8.404 | 0.0001 |

*Pour un test d'exogénéité des instruments*

Model: MODEL1
Dependent Variable: UCHAP Residual

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Prob>F |
|---|---|---|---|---|---|
| Model | 2 | 7.44447 | 3.72223 | 20.595 | 0.0001 |
| Error | 15828 | 2860.73566 | 0.18074 | | |
| C Total | 15830 | 2868.18013 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.42513 | R-square | 0.0026 |
| Dep Mean | -0.00000 | Adj R-sq | 0.0025 |
| C.V. | -4.11006E16 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > \|T\| |
|---|---|---|---|---|---|
| INTERCEP | 1 | -0.003552 | 0.00375684 | -0.945 | 0.3444 |
| ENF3 | 1 | 0.066414 | 0.01085515 | 6.118 | 0.0001 |
| ENF3_6 | 1 | -0.023264 | 0.00926968 | -2.510 | 0.0121 |