Exercices de travaux dirigés
Cours d'économétrie
Maîtrise d'économétrie

September 21, 2004

1 Le modèle linéaire - Rendements d'une fonction de production Cobb-Douglas

Présentation du problème: On considère la fonction de production suivante à deux facteurs, le travail L et le capital K, correspondant à une technologie de type Cobb-Douglas:

$$Y(L,K) = AL^{\alpha}K^{\beta} \tag{1}$$

où α et β sont des réels compris entre 0 et 1.

Soit un échantillon $\{(y_i, \ell_i, k_i)\,,\ i = 1, ..., n\}$ d'observations indépendantes de logarithmes d'outputs et d'inputs de n entreprises. On suppose que

$$y_i = a + \alpha l_i + \beta k_i + u_i, \tag{2}$$

où les u_i sont *iid* et **normaux** d'espérance nulle et de variance σ^2. On supposera de plus u_i orthogonal à l_i et k_i.

Données numériques : $n = 1000$
$\sum_i \ell_i = 500$
$\sum_i k_i = 490$
$\sum_i y_i = 1490$
$\sum_i \ell_i^2 = 330$
$\sum_i k_i^2 = 320$
$\sum_i y_i^2 = 3200$
$\sum_i \ell_i y_i = 800$
$\sum_i k_i y_i = 770$

Questions:

1. Justifier l'équation (2). Interpréter en particulier le sens de la variable u_i. Le fait de la traiter comme une variable aléatoire signifie-t-il que la valeur de u_i est le produit du hasard pour l'entreprise i?

2. Ecrire le modèle (2) sous la forme matricielle suivante :

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{u}, \tag{3}$$

où $\mathbf{y}$, $\mathbf{X}$, $\mathbf{b}$ et $\mathbf{u}$ sont des matrices que l'on déterminera et dont on indiquera les dimensions.

3. Donner l'expression de l'estimateur des MCO $\widehat{\mathbf{b}}$ de $\mathbf{b}$ en fonction de $\mathbf{y}$ et de $\mathbf{X}$.

4. **Estimation des coefficients:** Dans cette question, on cherche à estimer les coefficients $\widehat{\alpha}$, $\widehat{\beta}$ et $\widehat{a}$. On **centre** les variables du problème. Pour cela, on définit :

$$\widetilde{y}_i = y_i - \overline{y}, \quad \widetilde{\ell}_i = \ell_i - \overline{\ell}, \quad \text{et} \quad \widetilde{k}_i = k_i - \overline{k},$$

où $\overline{y}, \overline{\ell}$ et $\overline{k}$ sont les moyennes arithmétiques de y_i, ℓ_i et k_i dans l'échantillon.

   (a) Déduire du modèle (2) le modèle des variables $\left(\widetilde{y}_i, \widetilde{\ell}_i, \widetilde{k}_i\right)$.

   (b) L'écrire sous forme matricielle :

$$\widetilde{\mathbf{y}} = \widetilde{\mathbf{X}}\widetilde{\mathbf{b}} + \widetilde{\mathbf{u}}. \tag{4}$$

   Précisez les dimensions de $\widetilde{\mathbf{b}}$ et $\widetilde{\mathbf{X}}$.

   (c) On fait l'hypothèse :

$$\sum_i \widetilde{\ell}_i \widetilde{k}_i = \sum_i (\ell_i - \overline{\ell})(k_i - \overline{k}) = 0. \tag{5}$$

   Interpréter cette hypothèse.

   (d) Calculer la matrice $\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}$, ainsi que son inverse, sachant que l'hypothèse (5) est vérifiée. En déduire l'estimateur des MCO des coefficients de la régression de $\widetilde{y}_i$ sur $\widetilde{\ell}_i$ et $\widetilde{k}_i$ sans constante.

   (e) En déduire, par l'application du théorème de régression partitionnée, l'estimateur des MCO des coefficients de la régression de y_i sur ℓ_i et k_i avec constante.

   (f) **Application numérique**.

5. **Significativité des coefficients:** On teste ici la significativité de $\widehat{\alpha}$ et $\widehat{\beta}$.

   (a) Ecrire l'équation d'orthogonalité entre $\mathbf{X}\widehat{\mathbf{b}}$ et $\mathbf{y} - \mathbf{X}\widehat{\mathbf{b}}$.

   (b) Donner l'expression de la somme des carrés des résidus, SCR, en fonction de y_i, $\overline{y}$, ℓ_i, $\overline{\ell}$, k_i, $\overline{k}$, $\widehat{\alpha}$ et $\widehat{\beta}$. En déduire l'expression de l'estimateur $\widehat{\sigma}^2$. **Application numérique**.

   (c) Donner l'expression de $\widehat{\sigma}^2_{\widehat{\alpha}}$ et $\widehat{\sigma}^2_{\widehat{\beta}}$ en fonction de $\widehat{\sigma}^2$, ℓ_i, $\overline{\ell}$, k_i et $\overline{k}$. **Application numérique**.

   (d) Tester la significativité de $\widehat{\alpha}$ et de $\widehat{\beta}$ à 10% et 5% près. Conclusion?

6. **Test de l'hypothèse de rendements constants:** On teste l'hypothèse nulle suivante

$$(H_0) : \alpha + \beta = 1$$

Contre l'alternative :

$$(H_a) : \alpha + \beta \neq 1$$

(a) Ecrire le modèle **contraint** associé à H_0, modèle dans lequel β n'intervient plus. Quelles sont les variables dépendantes et indépendantes de ce nouveau problème?

(b) Calculer la somme des carrés des résidus du modèle contraint SCR_c. **Application numérique :** On vérifiera que $\alpha_c \simeq 0.594$.

(c) Calculer alors la statistique de Fischer associée au test de rendements constants. Tester H_0 à 5% près. Conclusion ?

2 Théorie asymptotique des MCO

2.1 Lancer d'une pièce de monnaie pipée

On lance un grand nombre de fois une pièce de monnaie déséquilibrée, dont la probabilité d'obtenir "face" est égale à $\alpha \in [0, 1]$.

1. On modélise le problème par le modèle suivant:

$$y = \beta + \varepsilon,$$

où y est la réalisation du lancer ($y = 0$ si "pile", $y = 1$ si "face"), et ε est d'espérance nulle.
Que vaut le scalaire β? Dessiner la fonction de répartition de ε. Sa loi est-elle normale?

2. Calculer l'estimateur des MCO de β: $\hat{\beta}$. Préciser les hypothèses ("naturelles") que vous faites sur les résidus.
Vers quelle valeur converge $\hat{\beta}$ lorsque le nombre de lancers tend vers $+\infty$?

3. Donner l'expression de la variance de ε en fonction de α.

4. Déduire de la question précédente la densité asymptotique de $\hat{\beta}$ (NB: on pourra noter n le nombre de lancers).

2.2 Inclusion et oubli de variables non pertinentes

On dispose d'un échantillon $\{(y_i, x_{1i}, x_{2i}),\ i = 1, ..., n\}$ d'observations indépendantes.

1. On suppose tout d'abord que le modèle de y_i sachant x_{1i}, x_{2i} est

$$y_i = \alpha + \beta_1 x_{1i} + u_i, \tag{6}$$

avec $\mathbb{E}(u_i|x_{1i}, x_{2i}) = 0$. On cherche à mesurer les conséquences de l'introduction d'une seconde variable explicative, x_{2i}.

   (a) Calculer les estimateurs des MCO $\widehat{\beta}_1$ et $\widehat{\beta}_2$ des coefficients de x_{1i} et x_{2i} dans la régression de y_i sur x_{1i} et x_{2i}.

   (b) Montrer que $\widehat{\beta}_1$ est asymptotiquement équivalent à l'estimateur des MCO de la régression de y_i sur x_{1i} **sans** x_{2i}. Ces deux estimateurs sont-ils pour autant identiques?

2. On suppose maintenant que le vrai modèle est

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + v_i \tag{7}$$

avec $\mathbb{E}(v_i|x_{1i}, x_{2i}) = 0$.

   (a) Poser $u_i = \beta_2 x_{2i} + v_i$ et calculer $\mathbb{E}(u_i|x_{1i}, x_{2i})$. A quelle condition $\mathbb{E}(u_i|x_{1i}, x_{2i}) = 0$?

(b) On suppose d'abord que x_1 et x_2 sont orthogonaux ($cov(x_1, x_2) = 0$). Montrer que l'estimateur des MCO du coefficient de x_{1i} dans la régression de y_i sur x_{1i} **sans** x_{2i} est convergent mais moins efficace (asymptotiquement moins précis) que l'estimateur des MCO du coefficient de x_{1i} dans la régression de y_i sur x_{1i} **et** x_{2i}.

(c) Si x_1 et x_2 sont corrélés, montrer que l'estimateur des MCO du coefficient de x_{1i} dans la régression de y_i sur x_{1i} **sans** x_{2i} est non convergent (asymptotiquement biaisé).

2.3 Test d'égalité de deux moyennes

On cherche à tester l'influence du sexe sur le salaire. Soit $\{w_{1i},\ i = 1, ..., T_1\}$ un échantillon de salaires d'hommes et $\{w_{2i},\ i = 1, ..., n_2\}$ un échantillon de salaires de femmes. On suppose les observations *iid.* Pour cela, on divise le panel en deux sous-échantillons, femmes (1, taille n_1) et hommes (2, taille n_2). On considère alors les deux modèles suivants

- Le modèle *non contraint* s'écrit :

$$w_{1i} = \alpha_1 + u_{1i}, \qquad \mathbb{E}u_{1i} = 0,$$

$$w_{2i} = \alpha_2 + u_{2i}, \qquad \mathbb{E}u_{2i} = 0,$$

  où α_1 et α_2 ne sont pas supposés égaux *a priori*.

- Le modèle *contraint* impose l'égalité des coefficients α_1 et α_2:

$$H_0 : \alpha_1 = \alpha_2.$$

Données numériques: $n_1 = 32728$
$n_2 = 35144$
$\sum_i w_{1i} = 2.11 * 10^8$
$\sum_i w_{2i} = 2.865 * 10^8$
$\sum_i w_{1i}^2 = 1.66 * 10^{12}$
$\sum_i w_{2i}^2 = 2.84 * 10^{12}$
$\widehat{\sigma_1} = 916$
$\widehat{\sigma_2} = 1266$

Questions:

1. On empile les deux échantillons pour former l'échantillon $\{w_i,\ i = 1, ..., n\}$ où $n = n_1 + n_2$ et

$$w_i = \left| \begin{array}{l} w_{1i},\ i = 1, ..., n_1, \\ w_{2i},\ i = n_1 + 1, ..., n, \end{array} \right.$$

   (a) Ecrire le modèle de w_i sous la forme:

$$w_i = \alpha_1 S_i + \alpha_2 (1 - S_i) + u_i, \quad i = 1, ..., n \tag{8}$$

   où l'on précisera S_i et u_i.

(b) Comment s'interprètent les variables S_i et $1 - S_i$ et combien vaut $\mathbb{E}(u_i|S_i)$?

(c) Montrer que l'estimateur des MCO de α_1 (resp. de α_2) dans la régression de w_i sur S_i et $1 - S_i$ est identique à l'estimateur des MCO de α_1 dans la régression de w_{1i} (resp. de w_{2i}) sur la constante 1. Expliciter $\widehat{\alpha}_1$ et $\widehat{\alpha}_2$.

2. **Le test à distance finie (rappels de licence).** On suppose les résidus du modèle (8) sont iid et *normaux*, de variance $\mathbb{V}u_i = \sigma^2$.

(a) L'hypothèse $\mathbb{V}u_i = \sigma^2$ pour tout $i = 1, ..., n$ vous paraît-elle discutable?

(b) Ecrire le modèle non contraint sous forme matricielle.

(c) Montrer que la somme des carrés des résidus non contraints SCR_{nc} s'écrit très simplement en fonction des sommes des carrés des résidus u_1 et u_2. La calculer. Montrer de plus que $\frac{SCR_{nc}}{\sigma^2}$ suit un χ^2 dont on précisera le nombre de degrés de liberté.

(d) Montrer que, sous l'hypothèse d'absence de changement structurel que l'on précisera, l'estimateur $(\widehat{\alpha}_1, \widehat{\alpha}_2)$ suit une loi normale dont on calculera la moyenne et la variance.

(e) Exprimer la différence $SCR_c - SCR_{nc}$ en fonction de n_1, n_2, et de la différence $\widehat{\alpha}_1 - \widehat{\alpha}_2$.

(f) En déduire que cette quantité, convenablement normalisée, suit un χ^2 à un degré de liberté. Interpréter.

(g) Calculer la statistique de Fischer associé au problème. Tester à 5% l'égalité entre α_1 et α_2. Conclusion?

3. **Le test asymptotique.** On abandonne ici l'hypothèse de normalité des résidus, ceux-ci restant *iid*.

(a) Calculer l'estimateur des MCO de $\alpha_1 - \alpha_2$.

(b) Sous quelle hypothèse quant au comportement asymptotique de la statistique $\overline{S} = \frac{1}{n}\sum_{i=1}^{n} S_i = \frac{n_1}{n}$ cet estimateur est-il convergent? Interpréter.

(c) Sous ces hypothèses, montrer que

$$\sqrt{n}\widehat{\alpha_1 - \alpha_2} \xrightarrow[n\to\infty]{L} \mathcal{N}\left(0, \frac{\sigma^2}{p(1-p)}\right)$$

où $p = \mathbb{E}S_i$.
Pour cela, montrer successivement que:

i. $\overline{S} \xrightarrow[n\to\infty]{P} p$.

ii. $\frac{1}{n}\sum_{i=1}^{n}\left(S_i - \overline{S}\right)^2 = \overline{S} - \overline{S}^2 \xrightarrow[n\to\infty]{P} p(1-p)$.

iii. $\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\left(S_i - \overline{S}\right)(u_i - \overline{u}) \xrightarrow[n\to\infty]{L} \mathcal{N}(0, \sigma^2 p(1-p))$.

(d) Comment estimer la variance asymptotique de $\widehat{\alpha_1 - \alpha_2}$, $\mathbb{V}_{as}\widehat{\alpha_1 - \alpha_2}$?

(e) Tester alors l'absence de changement structurel, à 5%, à l'aide d'un test de Student (asymptotique).

3 Le Modèle Hétéroscédastique

3.1 Test d'égalité de deux moyennes (suite de (2.3))

On suppose mantenant que $\mathbb{V}u_{1i} = \sigma_1^2$ et $\mathbb{V}u_{2i} = \sigma_2^2$ avec σ_1 et σ_2 quelconques. On estime α_1 et α_2 séparément sur les échantillons de filles et de garçons.

1. Montrer que $\widehat{\alpha}_1$ et $\widehat{\alpha}_2$ sont deux variables aléatoires indépendantes et que

$$\sqrt{n_1}\left(\widehat{\alpha}_1 - \alpha_1\right) \xrightarrow[n_1\to\infty]{L} \mathcal{N}\left(0, \sigma_1^2\right),$$
$$\sqrt{n_2}\left(\widehat{\alpha}_2 - \alpha_2\right) \xrightarrow[n_2\to\infty]{L} \mathcal{N}\left(0, \sigma_2^2\right).$$

2. Montrer de plus que

$$\widehat{\sigma}_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(w_{1i} - \widehat{\alpha}_1\right)^2 \xrightarrow[n_1\to\infty]{P} \sigma_1^2,$$
$$\widehat{\sigma}_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_1}\left(w_{12} - \widehat{\alpha}_2\right)^2 \xrightarrow[n_2\to\infty]{P} \sigma_2^2.$$

3. On suppose que $n_1/n_2 \to k \neq 0$ lorsque $n_1 \to \infty$.

   (a) Montrer que
   $$\sqrt{n_1}\left(\widehat{\alpha}_1 - \widehat{\alpha}_2\right) \xrightarrow[n_1\to\infty]{L} \mathcal{N}\left(0, \sigma_1^2 + k\sigma_2^2\right).$$

   (b) En déduire que
   $$T = \frac{\widehat{\alpha}_1 - \widehat{\alpha}_2}{\sqrt{\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}}} \xrightarrow[n_1\to\infty]{L} \mathcal{N}(0,1)$$
   sous l'hypothèse nulle $H_0 : \alpha_1 = \alpha_2$.

   (c) Vérifier que
   $$\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2} \xrightarrow[n_1\to\infty]{P} \frac{\sigma^2}{np\left(1-p\right)}$$
   si $\sigma_1 = \sigma_2 = \sigma$.

4. On suppose que $n_1 \to \infty$ et $n_1/n_2 \to k = 0$.

   (a) Interpréter.

   (b) Montrer que
   $$\frac{\widehat{\alpha}_1 - \widehat{\alpha}_2}{\frac{\widehat{\sigma}_1^2}{n_1}} \xrightarrow[n_1\to\infty]{L} \mathcal{N}(0,1).$$

3.2 Un test de Goldfeld-Quandt

Dans cet exercice, on cherche à quantifier l'influence du diplôme sur le salaire. On considérera donc le modèle linéaire suivant :

$$w_i = \alpha x_i + \beta + u_i$$

Où x_i repère le niveau d'éducation de l'individu i. On supposera les résidus u_i inid et orthogonaux aux variables explicatives.
On teste l'hétéroscédasticité des résidus en divisant l'échantillon en deux sous-échantillons I_1 et I_2 correspondant aux "non-diplômés" (niveau inférieur au bac) et aux "diplômés" (au moins le bac). On retire ensuite de ces deux sous-échantillons une proportion des individus telle que tous deux aient la même taille. On calcule ensuite les sommes des résidus des deux sous-modèles, soit SR_1 et SR_2.

1. Montrer que $\frac{SR_2}{SR_1}$ suit, sous l'hypothèse d'homoscédasticité, une loi de Fisher dont on précisera le nombre de degrés de liberté.

2. Tester l'hypothèse d'homoscédasticité aux niveaux 1%, 5% et 10%. Qu'en déduire ?

3. On considère alors le modèle :

$$ln(w_i) = \alpha x_i + \beta + u_i. \tag{9}$$

   Justifier la forme choisie: pourquoi utilise-t-on le logarithme?

4. Tester l'hypothèse d'homoscédasticité sur ce modèle. Conclusion ?

3.3 Utilisation à tort des MCO

On évalue l'erreur qui est faite lorsque l'on estime un modèle linéaire *hétéroscédastique* par les MCO.

On considère le modèle hétéroscédastique :

$$y_i = a + bx_i + u_i, \tag{10}$$

où la matrice de variance-covariance des u_i est diagonale, égale à $\sigma^2 diag(\omega_1, ..., \omega_n)$. On suppose que les poids ω_j sont positifs, et somment à 1.

1. Calculer $\hat{b}_{MCO}$. Est-il sans biais? Convergent?

2. Mêmes questions pour $\hat{b}_{MCP}$, l'estimateurs des Moindres Carrés Pondérés du paramètre b.

3. Calculer les variances de $\hat{b}_{MCO}$ et $\hat{b}_{MCP}$.

4. Montrer que $\hat{b}_{MCP}$ est plus précis.

5. On calcule un estimateur de la variance de $\hat{b}_{MCP}$ par la méthode des MCP. Montrer que cet estimateur n'est pas biaisé.

6. On calcule maintenant un estimateur de la variance de $\hat{b}_{MCO}$ par les MCO. Montrer que cet estimateur est biaisé. Dans quelle direction?

3.4 Observations Groupées

Soit un échantillon d'observations iid $\{(y_i, d_i), i = 1, ..., N\}$, avec $y_i \in \mathbb{R}$ et $d_i \in \{1, ..., J\}$. La variable d_i est une variable discrète qui indique un groupe social d'appartenance de l'individu i (les diplômés par opposition aux non diplômés, différentes PCS, etc.). Chaque groupe social $j \in \{1, ..., J\}$ est caractérisé par un vecteur de constantes $z_j \in \mathbb{R}^K$ (revenu moyen, âge moyen, etc.). Pour tout $j \in \{1, ..., J\}$, on note N_j le nombre d'individus i dans le groupe j et $\overline{y}_j$ la moyenne de y_i dans le groupe j. Enfin, $\delta_i^j = \mathbf{1}\{d_i = j\}$ dénote la variable indiquant si le groupe d'appartenance est le groupe j ($\delta_i^j = 1$ si $d_i = j$, $= 0$ sinon).

1. Soit $\widehat{\beta} = (\widehat{\beta}_1, ..., \widehat{\beta}_J)'$ l'estimateur des MCO de la régression de y_i sur le vecteur $x_i = (\delta_i^1, ..., \delta_i^J)'$.

   (a) Remplacer les ? dans les deux équations suivantes par l'expression appropriée:

   $$\begin{aligned} N_j &= \sum_{i=1}^{N} ? \\ \overline{y}_j &= \frac{\sum_{i=1}^{N} ? y_i}{\sum_{i=1}^{N} ?} \end{aligned}$$

   (b) Interprétez géométriquent le système des équations normales définissant l'estimateur des MCO:

   $$\sum_{i=1}^{N} x_i \left(y_i - x_i' \widehat{\beta}\right) = 0.$$

   (c) Déduire des équations normales que $\widehat{\beta}_j = \overline{y}_j$.

2. On considère maintenant le modèle de régression linéaire suivant:

   $$y_i = a + z_{d_i}' b + u_i \tag{11}$$

   où $z_{d_i} = \sum_{j=1}^{J} z_j \delta_i^j$ et avec $\mathbb{E}(u_i | d_i) = 0$ et $\mathbb{V}(u_i | d_i) = \sigma^2$.

   (a) Montrer que les équations normales définissant l'estimateur des MCO de a et b s'écrivent:

   $$\begin{cases} \sum_{j=1}^{J} N_j \left(\overline{y}_j - \widehat{a} - z_j' \widehat{b}\right) = 0 \\ \sum_{j=1}^{J} N_j z_j \left(\overline{y}_j - \widehat{a} - z_j' \widehat{b}\right) = 0 \end{cases} \tag{12}$$

   où N_j est le nombre d'individus i appartenant au groupe j.

   (b) Combien y-a-t'il d'équations et de variables dans le système (12)?

   (c) Montrer que

   $$\widehat{a} = \overline{y} - \overline{z}' \widehat{b}., \tag{13}$$

   $$\widehat{b} = \left(\sum_{j=1}^{J} N_j (z_j - \overline{z})(z_j - \overline{z})'\right)^{-1} \left(\sum_{j=1}^{J} N_j (z_j - \overline{z})\left(\overline{y}_j - \overline{y}\right)\right), \tag{14}$$

   où $\overline{y}$ et $\overline{z}$ sont les moyennes de y_i et z_{d_i} dans l'échantillon,.

3. On considère ensuite le modèle de régression linéaire suivant:

$$\overline{y}_j = a + z_j' b + v_j, \quad j = 1, ..., J. \tag{15}$$

(a) Montrer que l'équation (15) se déduit de l'équation (11) pour un choix de v_j que vous expliciterez.

(b) Montrer que $\mathbb{E}(v_j|X) = 0$ où $X = (\delta_i^1, ..., \delta_i^J)_{i=1,...,N}$.

(c) Montrer que $\mathbb{V}(v_j|X) = \frac{\sigma^2}{N_j}$.

(d) Montrer que $\text{Cov}(v_j, v_{j'}|X) = 0$, $\forall j \neq j' \in \{1, ...J\}$.

(e) Calculer l'estimateur des MCO de la régression de $\overline{y}_j$ sur 1 et z_j.

(f) Calculer l'estimateur des MCG et montrer que c'est le même estimateur que celui obtenu dans la question 2c.

4 Endogénéité des variables explicatives

4.1 Rendements de l'éducation

On considère l'équation de salaire suivante :

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

où x_i représente le niveau d'éducation de l'individu i, et y_i le logarithme de son salaire. On s'intéresse aux éventuels problèmes d'endogénéité posés par cette formulation.

1. Pour mettre ces problèmes en évidence, on postule dans cette question l'existence d'une caractéristique inobservée, z_i, qui influence à la fois u_i et x_i. Soit :

$$\begin{cases} u_i = bz_i + \eta_i, \\ x_i = \alpha + \beta z_i + e_i. \end{cases}$$

   (a) Interpréter ce modèle structurel.

   (b) En supposant η_i et e_i non corrélés et de moyenne nulle, déterminer le biais asymptotique de l'estimateur des MCO $\hat{b}_{MCO}$. Montrer qu'il est vraisemblablement positif.

   (c) On calcule empiriquement le biais de $\hat{b}$. Quelle méthode peut-on utiliser? On trouve alors un biais significativement négatif.

2. On interprète le paradoxe des questions précédentes en postulant la présence d'erreurs de mesure. On suppose que le vrai modèle s'écrit :

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

   où y_i^* est le salaire mesuré par y_i avec erreur:

$$y_i = y_i^* + \nu_i,$$

   et x_i^* le vrai niveau d'éducation mesuré avec erreur par x_i:

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

   On suppose les erreurs de mesure ε_i et ν_i non corrélées entre elles, *iid* et non corrélées avec x_i^*.

   (a) Soit $\widehat{b}$ l'estimateur des MCO du coefficient de x_i dans la régression de y_i sur x_i avec constante. Exprimer le biais asymptotique sur b et montrer que l'erreur de mesure biaise l'estimateur vers 0.

   (b) Soit $\widehat{c}$ l'estimateur des MCO du coefficient de y_i dans la régression de x_i sur y_i avec constante. Exprimer le biais asymptotique de $1/\widehat{c}$ sur b. Montrer que le biais est positif.

   (c) En déduire que l'on peut obtenir un encadrement du vrai rendement de l'éducation, et discuter la précision de cet encadrement. Montrer en particulier que $1/\widehat{c}$ reste biaisé même lorsqu'il n'y a pas d'erreur de mesure.

4.2 Un modèle d'offre de travail

Dans cet exercice, on considère le modèle suivant:

$$y_i = a + bx_i + u_i. \tag{16}$$

La variable y_i représente le nombre d'heures travaillées par l'individu i dans la semaine précédant l'enquête, et x_i est le salaire horaire de ce même individu.

1. Quelles sont les deux interprétations possibles du résidu u_i vues en cours. Pour quelle raisons, dans l'éventualité d'une interprétation causale de cette relation, la variable x_i est-elle susceptible d'être corrélée au résidu u_i?

2. On suppose dans cette question que x_i est endogène dans l'équation (16). Montrer qu'alors l'estimateur des MCO $\hat{b}$ de b est biaisé, et calculer son biais en fonction de x_i et u_i. S'attend-on à un biais positif ou négatif? Justifier.

3. Parmi les variables suivantes, lesquelles peut-on rejeter immédiatement comme n'étant pas des instruments convenables pour le modèle (16) : indicatrice de temps partiel, profession de l'individu, région de résidence, diplôme, salaire hebdomadaire. Justifier chacune de vos affirmations.

4. On retient dans cette question et la suivante la profession des parents comme instrument. Expliquer comment on obtient l'estimateur des doubles moindres carrés de b, associé au modèle (16) et à l'instrument considéré (que l'on pourra noter z_i pour les besoins de l'explication).
   On effectue le calcul de $\hat{b}_{2MC}$ sur un échantillon de 10000 individus. La régression augmentée de y_i sur x_i, $\widehat{v}_i$ et la constante donne :

$$\hat{y}_i = \underset{(12)}{174}x_i - \underset{(12)}{204}\widehat{v}_i + \underset{(3)}{14},$$

   où les écarts-types sont entre parenthèse. Que représente la variable v_i? Donner la moyenne et l'écart-type de $\hat{b}_{2MC}$. Tester ensuite l'exogénéité de x_i pour le modèle (1) à 5%. Dans quel sens l'estimateur des MCO est-il biaisé? Commenter.

5. Peut-on tester la validité de l'instrument "profession des parents" à partir des informations contenues dans l'énoncé ? Comment pourrait-on s'y prendre pour la tester? Expliquer.

6. On s'intéresse maintenant à l'éventuelle hétéroscédasticité du modèle (16). Expliquer pourquoi la variance conditionnelle $\mathbb{V}(y_i|x_i)$ est vraisemblablement monotone en x_i. Quelle méthode peut-on appliquer pour tester l'hétéroscédasticité du modèle?

7. On suppose le modèle (16) hétéroscédastique. On instrumente alors par la profession des parents, comme dans la question 4. Le coefficient $\hat{b}_{2MC}$ est-t-il convergent? Que dire de son écart-type?

8. Donner une méthode permettant d'éliminer asymptotiquement le biais mis en évidence à la question précédente. Expliquer son fonctionnement.

9. D'après les conclusions de l'exercice, quel effet, revenu ou substitution, est dominant dans l'échantillon? Proposer une autre forme pour le modèle (16) qui permette de prendre en compte ces deux effets simultanément.

4.3 Régression vers la Moyenne?

Soit un échantillon d'observations iid $\{(y_i, x_i), i = 1, ..., N\}$, avec $y_i \in \mathbb{R}$, $x_i \in \mathbb{R}$. On suppose qu'il existe une variable $d_i \in \{1, ..., J\}$, inobservée, qui partitionne les individus en J groupes. La variable x_i est la taille du père de l'individu i et y_i est sa propre taille. En régressant $y_i - \overline{y}$ sur $x_i - \overline{x}$ le statisticien Galton a trouvé un coefficient inférieur à un, phénomène qu'il a qualifié de régression vers la moyenne. En réalité, il s'agit d'un artefact statistique qu'on va chercher à comprendre.

Soit $z_1, ..., z_J \in \mathbb{R}$. On suppose vérifié le modèle suivant:

$$\begin{aligned} y_i - \overline{y} &= z_{d_i} + u_i, \\ x_i - \overline{x} &= z_{d_i} + v_i, \end{aligned}$$

où u_i et v_i sont deux perturbations de moyennes nulle et de variances constantes non nulles conditionnellement à d_i:

$$\begin{aligned} \mathbb{E}(u_i|d_i) &= 0 \quad \text{et} \quad \mathbb{V}(u_i|d_i) = \sigma_u^2, \\ \mathbb{E}(v_i|d_i) &= 0 \quad \text{et} \quad \mathbb{V}(v_i|d_i) = \sigma_v^2. \end{aligned}$$

1. Calculer $\mathbb{E}(y_i - \overline{y}|d_i = j)$ et $\mathbb{E}(x_i - \overline{x}|d_i = j)$.

2. Interprétez z_j. Quelle justification donner au fait que l'on suppose que c'est le même z_j qui apparaît dans les deux équations?

3. Calculer l'estimateur des MCO $\widehat{b}$ du coefficient de la régression sans constante de $y_i - \overline{y}$ sur $x_i - \overline{x}$.

4. Montrer que $0 < \text{plim}_{N\to\infty} \widehat{b} < 1$.

5. Les économistes de la croissance ont souvent régressé le taux de croissance moyen du PIB (sur une période donnée) sur le PIB de début de période:

$$\ln PIB_{i1} - \ln PIB_{i0} = a + b \ln PIB_{i0} + u_i$$

pour un échantillon de pays $i = 1, ..., N$. Une estimation négative du coefficient b est souvent interprétée comme le signe d'une convergence vers un niveau de PIB commun. Montrer à l'aide du modèle précédent qu'une telle interprétation peut être fallacieuse.

5 Equations simultanées

5.1 Modèle de Haavelmo

On considère le modèle d'équilibre général formé des deux équations suivantes:

$$\begin{aligned} c &= \alpha y + \beta + u, \\ y &= c + i, \end{aligned}$$

où y est la production, c la consommation et l'investissement i est considéré comme *exogène*.

1. Interpréter ces deux équations.
2. Exprimer les formes réduites de ce système.
3. Calculer la limite en probabilité de $\widehat{\alpha}_{MCO}$, l'estimateur de α par les MCO.
4. Quelles remarques pouvez-vous faire sur l'estimation des modèles d'équilibre général. Proposez une méthode d'estimation convergente des paramètres.

5.2 Offre et demande

On estime dans cet exercice un modèle Offre/Demande. Soit :

$$S_i(p_i, v_i) = \alpha + \beta p_i + v_i,$$

$$D_i(p_i, u_i) = a + b p_i + u_i.$$

On suppose de plus que les résidus suivent une loi normale bivariée dont les paramètres sont :

$$E(u_i) = E(v_i) = 0,$$

$$V(u_i) = \sigma_u^2 \quad ; \quad V(v_i) = \sigma_v^2 \quad ; \quad Cov(u_i, v_i) = \rho \sigma_u \sigma_v.$$

1. La loi conditionnelle $v_i | u_i$ est normale. Calculer sa moyenne et sa variance. En déduire par symétrie la loi de $u_i | v_i$.
2. Calculer la loi marginale du prix d'équilibre p_i.
3. Calculer $E(u_i|p_i)$ et $E(v_i|p_i)$.
4. Vérifier que :
$$a + b p_i + E(u_i|p_i) = \alpha + \beta p_i + E(v_i|p_i)$$

5.3 Identification

1. Soit le système d'équations simulatanées :

$$\begin{cases} y_{1t} = a_1 + b_1 \cdot y_{2t} + c_1 \cdot x_{1t} + u_{1t} \\ y_{2t} = a_2 + b_2 \cdot y_{1t} + c_2 \cdot x_{2t} + u_{2t} \end{cases}$$

   avec u_{1t} et u_{2t} corrélés.

   (a) Ecrire les formes structurelle et réduite correspondantes.

   (b) Que veut dire : "les paramètres du modèle sont identifiés." Donner la définition.

   (c) A l'aide de la condition d'ordre, dire si les équations sont identifiables.

   (d) Montrer à l'aide de la forme réduite que les paramètres sont en effet identifiés.

2. Soit le modèle :

$$\begin{cases} y_{1t} = a_1 + b_1 \cdot y_{2t} + c_1 \cdot x_{1t} + u_{1t} \\ y_{2t} = a_2 + c_2 \cdot x_{1t} + u_{2t} \end{cases}$$

   (a) La condition d'ordre reste-t-elle satisfaite pour chaque équation ?

   (b) Quels paramètres ou fonctions des paramètres sont identifiables?