No 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2006 : Enjeux méthodologiques reliés à la mesure de la santé des populations

2006

Statistics Canada Statistique Canada

Canada

Recueil du Symposium 2006 de Statistique Canada
Enjeux méthodologiques reliés à la mesure de la santé des populations

Estimation des paramètres de régression au moyen de données d'enquête

Wayne A. Fuller et Yu Y. Wu[1]

Résumé

Les coefficients des équations de régression sont souvent des paramètres d'intérêt dans le cas des enquêtes sur la santé et ces dernières sont habituellement réalisées selon un plan de sondage complexe avec l'utilisation des taux d'échantillonnage différentiels. Nous présentons des estimateurs des coefficients de régression applicables aux enquêtes complexes qui sont supérieurs aux estimateurs à facteur d'extension ordinaires selon le modèle en question, mais retiennent aussi les propriétés souhaitables du plan. Nous présentons les propriétés théoriques et celles qui sont simulées par la méthode Monte Carlo.

MOTS-CLÉS : Variables instrumentales; pondération probabiliste; enquêtes complexes

1. Introduction

Nous considérons l'estimation des coefficients de régression à l'aide de données recueillies selon un plan de sondage complexe. Nous supposons que l'équation de régression fait partie d'un modèle spécialisé qui spécifie la population finie qu'il convient de générer par un processus stochastique, lequel est appelé modèle de superpopulation. Nous utilisons la graphie $\mathcal{F}$ pour représenter la population finie, U pour représenter l'ensemble des indices de la population finie et A pour représenter l'ensemble des indices de l'échantillon. Nous supposons qu'il existe une fonction $p(.)$ telle que $p(\mathrm{A})$ donne la probabilité de sélectionner l'échantillon A à partir de U.

Un modèle de superpopulation pour la régression est donné par

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + e_i,$$

où $(y_i, \mathbf{x}_i)$ sont des vecteurs indépendants et identiquement distribués (*iid*), et e_i est indépendante de $\mathbf{x}_i$. Soit un ensemble de N vecteurs définissant une population finie. Le modèle de la population finie peut alors s'écrire sous la forme

$$\begin{aligned} \mathbf{y}_N &= \mathbf{X}_N\boldsymbol{\beta} + \mathbf{e}_N, \\ e_N &\sim (\mathbf{0}, \mathbf{I}_N\sigma^2), \end{aligned} \tag{1}$$

où $\mathbf{y}_N = (y_1, y_2, ..., y_N)'$ est le vecteur de dimension N des valeurs de la variable dépendante, $\mathbf{X}_N = (x_1', x_2', ..., x_N')'$ est la matrice de dimensions $N \times k$ des valeurs des variables explicatives, et le vecteur d'erreurs $\mathbf{e}_N = (e_1, e_2, ..., e_N)'$ est un vecteur de dimension N qui est indépendant de $\mathbf{X}_N$. Supposons que l'on sélectionne un échantillon probabiliste à partir de la population finie avec les probabilités de sélection π_i.

[1] Wayne A. Fuller est professeur distingué émérite, Département de statistique, Iowa State University, Ames, Iowa, États-Unis. 50011 (courriel : waf@iastate.edu); Yu Y. Wu, Département de statistique, Iowa State University, Ames, Iowa, États-Unis 50011 (courriel : yuwu@iastate.edu).

2. Estimateurs

L'estimateur par les moindres carrés ordinaires (MCO pour *ols* pour *ordinary least squares* en anglais) de $\boldsymbol{\beta}$ est

$$\hat{\boldsymbol{\beta}}_{ols} = \left(\sum_{i \in A} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \sum_{i \in A} \mathbf{x}_i' y_i = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}, \tag{2}$$

où $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ est le vecteur colonne de dimension *n* des observations, et $\mathbf{X} = (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n')'$ est la matrice de dimensions $n \times k$ des observations des variables explicatives. L'erreur de l'estimateur par les moindres carrés ordinaires est

$$\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{e}, \tag{3}$$

où $\mathbf{e}$ est le vecteur de dimension *n* des erreurs et, selon des hypothèses faibles,

$$\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta} = \left(\sum_{i \in U} \mathbf{x}_i' \pi_i \mathbf{x}_i \right)^{-1} \sum_{i \in U} \mathbf{x}_i' \pi_i e_i + O_p(n^{-1}). \tag{4}$$

Donc, si $\mathbf{x}_i \pi_i$ et e_i sont corrélés, l'estimateur par les MCO est biaisé.

L'estimateur pondéré par les probabilités (PP) (*PW* pour *probability weighted* en anglais),construit en utilisant les inverses des probabilités de sélection, est donné par

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{PW} &= \left(\sum_{i \in A} \mathbf{x}_i' \pi_i^{-1} \mathbf{x}_i \right)^{-1} \sum_{i \in A} \mathbf{x}_i' \pi_i^{-1} y_i \\ &= (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \mathbf{X}'\mathbf{W}\mathbf{y}, \end{aligned} \tag{5}$$

où $\mathbf{W} = diag(\pi_1^{-1}, \pi_2^{-1}, \ldots, \pi_n^{-1}) =: diag(w_1, w_2, \ldots, w_n)$. Selon des hypothèses faibles concernant la population et pour de nombreux plans de sondage,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{PW} - \boldsymbol{\beta} &= \left(\sum_{i \in U} \mathbf{x}_i' \mathbf{x}_i \right)^{-1} \sum_{i \in U} \mathbf{x}_i' e_i + O_p(n^{-1}) \\ &= O_p(n^{-1/2}). \end{aligned} \tag{6}$$

Donc, $\hat{\boldsymbol{\beta}}_{PW}$ est un estimateur convergent du paramètre d'intérêt. Voir aussi Fuller (2002).

Posons que le modèle spécialisé spécifie les vecteurs $\mathbf{q}_i$ de telle façon que

$$E\left\{ \sum_{i \in A} \mathbf{q}_i' e_i \right\} = E\left\{ \sum_{i \in U} \pi_i \mathbf{q}_i' e_i \right\} = \mathbf{0}. \tag{7}$$

Nous définissons un estimateur à variables instrumentales (*IV* pour *instrumental variable* en anglais) par

$$\hat{\boldsymbol{\beta}}_{IV} = [(\mathbf{Q}'\mathbf{X})' \hat{\mathbf{V}}_{bb}^{-1} \mathbf{Q}'\mathbf{X}]^{-1} (\mathbf{Q}'\mathbf{X})' \hat{\mathbf{V}}_{bb}^{-1} \mathbf{Q}'\mathbf{y} \tag{8}$$

où $\hat{\mathbf{V}}_{bb}$ est une matrice définie positive symétrique. Le choix privilégié pour $\hat{\mathbf{V}}_{bb}$ est un estimateur de la variance de $\mathbf{Q}'\mathbf{e}$. À titre d'exemple, posons que $\mathbf{x}_i = (1, \mathbf{x}_{1,i})$, et que $\mathbf{q}_i = (w_i, w_i\mathbf{x}_{1i})$, où $w_i = \pi_i^{-1}$, et notons que $E\{\mathbf{Q}'\mathbf{e}\} = \mathbf{0}$. S'il existe une corrélation modérée entre e_i et π_i, et que $e_i \sim ind\ (0, \sigma^2)$, alors $\mathbf{V}_{bb} = V\{\mathbf{Q}'\mathbf{e}\} \doteq \mathbf{Q}'\mathbf{Q}\sigma^2$. L'estimateur (8) lorsque $\mathbf{V}_{bb} = \mathbf{Q}'\mathbf{Q}\sigma^2$ est l'estimateur par les doubles moindres carrés,

$$\hat{\boldsymbol{\beta}}_{IV} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}, \tag{9}$$

où $\hat{\mathbf{X}} = \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{X}$. Par exemple, voir Wooldridge (2000). En utilisant (9) et $\mathbf{q}_i = (w_i, w_i\mathbf{x}_{1,i})$, on obtient l'estimateur à variables instrumentales

$$\hat{\boldsymbol{\beta}}_{\mathrm{IV}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}. \tag{10}$$

Donc, l'estimateur PP est un estimateur à variables instrumentales avec $\mathbf{q}_i = (w_i, w_i\mathbf{x}_{1,i})$. Le cadre des variables instrumentales permet d'ajouter des instruments et d'effectuer des tests sur des instruments éventuels. Par exemple, l'estimateur de Pfeffermann-Sverchkov(1999) pour le modèle pour lequel $\mathbf{x}_i = (1, \mathbf{x}_{1,i})$ est un estimateur à variables instrumentales avec $\mathbf{q}_i = (w_i / \hat{w}_i, w_i / \hat{w}_i\mathbf{x}_{1,i})$, où $\hat{w}_i$ est le prédicteur par les moindres carrés de w_i fondé sur $\mathbf{x}_i$.

Dans un certain nombre de situations, il pourrait être raisonnable de penser que la probabilité de sélection est corrélée à l'erreur e_i, mais que

$$E\{(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})'e_i \mid i \in A\} = \mathbf{0}. \tag{11}$$

Considérons l'hypothèse selon laquelle les probabilités de sélection sont représentées par

$$\pi_i = g_1(\mathbf{x}_i) + g_2(e_i) + u_i, \tag{12}$$

où $g_1(.)$ et $g_2(.)$ sont des fonctions continues dérivables et u_i est indépendant de $(\mathbf{x}_i, e_i)$. Un exemple où le modèle (12) est raisonnable est celui des probabilités de sélection reliées à une valeur antérieure de *y*. Sachant (12),

$$E\left\{\sum_{i\in A}(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})'e_i\right\} = E\left\{\sum_{i\in U}\pi_i(\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})'e_i\right\} = \mathbf{0}, \tag{13}$$

parce que, selon le modèle (1), e_i est indépendante de $\mathbf{x}_i$. Il découle de (13) que (11) tient et que l'estimateur défini par

$$\left[\sum_{i\in A}\left[w_i, (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})\right]'(1, x_{1i})\right]\hat{\boldsymbol{\beta}}_{IV} = \sum_{i\in A}\left[w_i, (\mathbf{x}_{1,i} - \bar{\mathbf{x}}_{1,N})\right]'y_i \tag{14}$$

est convergent pour $\boldsymbol{\beta}$. Si $\bar{\mathbf{x}}_{1,N}$ est inconnu, il peut être remplacé par un estimateur convergent.

Pour étudier les tests applicables aux instruments, nous partitionnons le vecteur $\mathbf{q}_i$ en $(\mathbf{q}_{1i}, \mathbf{q}_{2i})$, en supposant que $E\{\Sigma_{i\in A}\mathbf{q}_{1i}'e_i\} = \mathbf{0}$ et nous souhaitons vérifier que

$$E\left\{\sum_{i\in A}\mathbf{q}_{2i}'e_i\right\} = \mathbf{0}. \tag{15}$$

Pour vérifier que $E\{\mathbf{Q}_2'\mathbf{e}\} = \mathbf{0}$, en utilisant l'estimateur par les doubles moindres carrés (9) comme estimateur de base, nous calculons

$$\hat{\boldsymbol{\gamma}} = [(\hat{\mathbf{X}}, \mathbf{R}_2)'(\hat{\mathbf{X}}, \mathbf{R}_2)]^{-1}(\hat{\mathbf{X}}, \mathbf{R}_2)'\mathbf{y}, \tag{16}$$

où $\mathbf{R}_2 = \mathbf{Q}_2 - \mathbf{Q}_1(\mathbf{Q}_1'\mathbf{Q}_1)^{-1}\mathbf{Q}_2$ et $\mathbf{Q} = (\mathbf{Q}_1, \mathbf{Q}_2)$. Si nous pouvons ignorer la correction pour population finie, une matrice de covariance estimée pour $\hat{\boldsymbol{\gamma}}$ est donnée par

$$\hat{V}\{\hat{\boldsymbol{\gamma}}\} = [(\hat{\mathbf{X}}, \mathbf{R}_2)'(\hat{\mathbf{X}}, \mathbf{R}_2)]^{-1}\hat{V}\{(\hat{\mathbf{X}}, \mathbf{R}_2)'\mathbf{e}\}[(\hat{\mathbf{X}}, \mathbf{R}_2)'(\hat{\mathbf{X}}, \mathbf{R}_2)]^{-1}. \tag{17}$$

Un estimateur de $V\{(\hat{\mathbf{X}}, \mathbf{R}_2)'\mathbf{e}\}$ est l'estimateur d'Horvitz-Thompson calculé avec $(\hat{\mathbf{X}}, \mathbf{R_2})'\hat{\mathbf{e}}$, où $\hat{e}_i = y_i - (\mathbf{x}_i, \mathbf{r}_{2i})\hat{\boldsymbol{\gamma}}$ et $\mathbf{r}_{2i}$ est la i^e ligne de $\mathbf{R}_2$. Selon l'hypothèse nulle $E\{\mathbf{Q}_2'\mathbf{e}\} = \mathbf{0}$, $\hat{\boldsymbol{\gamma}}_1$, le coefficient de $\hat{\mathbf{X}}$, estime $\boldsymbol{\beta}$ et $\hat{\boldsymbol{\gamma}}_2$, le coefficient de $\mathbf{r}_{2i}$, estime **0.** Par conséquent, un test statistique est donné par

$$F(k_2, n-k) = k_2^{-1}\hat{\boldsymbol{\gamma}}_2'\hat{\mathbf{V}}_{\gamma\gamma 22}^{-1}\hat{\boldsymbol{\gamma}}_2, \tag{18}$$

où $\hat{\mathbf{V}}_{\gamma\gamma 22}$ est le bloc inférieur droit $k_2 \times k_2$ de $\hat{V}\{\hat{\boldsymbol{\gamma}}\}$, k est la dimension de $\mathbf{q}_i$ et k_2 est la dimension de $\mathbf{r}_{2i}$. Selon l'hypothèse nulle, la statistique de test suit approximativement la loi F tabulée avec k_2 et $\mathrm{n-k}$ degrés de liberté.

3. Étude de Monte Carlo

Nous avons procédé à une étude par simulation pour évaluer les propriétés de deux estimateurs à VI et d'un estimateur de prétest en deux étapes. Le modèle est

$$\begin{aligned} y_i &= \beta_0 + x_{1i}\beta_1 + e_i \\ &= \mathbf{x}_i\boldsymbol{\beta} + e_i, \end{aligned}$$

où $\mathbf{x}_i = (1, x_{1i})$.

Nous créons chaque échantillon en générant le vecteur (x_{1i}, e_i, a_i, u_i), où x_{1i} est une variable aléatoire de loi normale (0, 0,5), e_i est une variable aléatoire de loi normale (0, 0,5), a_i est une variable aléatoire de loi normale (0, 0,5), u_i est une variable aléatoire de loi uniforme (0, 1), et les variables x_{1i}, e_i, a_i et u_i sont mutuellement indépendantes. La probabilité de sélection p_i est une fonction de x_{1i}, e_i et a_i,

$$p_i = p(x_{1i,}\, e_i, a_i) = 0,25r(x_{1i}) + 1,75r(\psi^{0,5}e_i + [1-\psi]^{0,5}a_i), \tag{19}$$

où

$$\mathrm{r(x)} = \begin{cases} 0,025 & \text{si x} < 0,2 \\ 0,475(\mathrm{x} - 0,20) + 0,025 & \text{si } 0,2 \le \mathrm{x} \le 1,2 \\ 0,5 & \text{si x} > 1,2 \end{cases}$$

et ψ est un paramètre dont on fait varier la valeur dans l'expérience. Le paramètre ψ détermine la corrélation entre p_i et e_i. Si $u_i \le p_i$ le vecteur est retenu pour l'échantillon; sinon, il est écarté.

Le premier estimateur à VI utilise un vecteur de quatre variables instrumentales, $\mathbf{z}_{1,i} = (w_i, w_i x_{1i}, w_i \hat{p}_i, w_i \hat{p}_i x_{1i})$, où $\hat{p}_i$ est la valeur prédite par la régression par les MCO de p_i sur $(1, r(x_{1i}))$. Le deuxième estimateur à VI est basé sur le vecteur $\mathbf{z}_{2,i} = (w_i, w_i x_{1i}, w_i \hat{p}_i, w_i \hat{p}_i x_{1i}, x_{1i})$. Les estimateurs à VI de $\boldsymbol{\beta}$ sont

$$\hat{\boldsymbol{\beta}}_{IVj} = [\mathbf{X}'\mathbf{Z}_j(\mathbf{Z}_j'\mathbf{Z}_j)^{-1}\mathbf{Z}_j'\mathbf{X}]^{-1}\mathbf{X}'\mathbf{Z}_j(\mathbf{Z}_j'\mathbf{Z}_j)^{-1}\mathbf{Z}_j'\mathbf{y}, \tag{20}$$

où $\mathbf{Z}_j = (\mathbf{z}_{j,1}', \mathbf{z}_{j,2}', \ldots, \mathbf{z}_{j,n}')'$, pour $j = 1, 2$. La matrice de covariance estimée de $\hat{\boldsymbol{\beta}}_{IVj}$ est

$$\hat{V}(\hat{\boldsymbol{\beta}}_{IVj}) = (\hat{\mathbf{X}}_j'\hat{\mathbf{X}}_j)^{-1}\hat{\mathbf{X}}_j'\hat{\mathbf{D}}_{ee,IVj}\hat{\mathbf{X}}_j(\hat{\mathbf{X}}_j'\hat{\mathbf{X}}_j)^{-1}, \tag{21}$$

où $\hat{\mathbf{X}}_j = \mathbf{Z}_j(\mathbf{Z}_j'\mathbf{Z}_j)^{-1}\mathbf{Z}_j'\mathbf{X}$, $\hat{\mathbf{D}}_{ee,IVj} = diag(\hat{e}_{1,IVj}^2, \hat{e}_{2,IVj}^2, \ldots, \hat{e}_{n,IVj}^2)$, et $\hat{e}_{i,IVj} = y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}}_{IVj}$.

L'estimateur de prétest est un estimateur en deux étapes basé sur l'estimateur par les MCO et les deux estimateurs à VI. Le premier test, qui est un test de l'importance des pondérations, est obtenu à partir de deux régressions : la régression de y_i sur $(1, x_{1i}, w_i, w_i x_{1i})$ (modèle complet) et la régression de y_i sur $(1, x_{1i})$ (modèle réduit). Si la statistique F

$$\mathrm{F}_{\mathrm{n-4}}^2 = \frac{(\mathrm{SCE}_{\text{réd.}} - \mathrm{SCE}_{\text{comp.}})/2}{\mathrm{EQM}_{\text{comp.}}} \tag{22}$$

n'est pas statistiquement significative, $\hat{\boldsymbol{\beta}}_{ols}$ est l'estimateur et sinon, un deuxième test doit être effectué. On calcule la régression par les MCO de y_i sur $(\tilde{\mathbf{x}}_i, x_{1i} - \hat{x}_{1i})$, où $\tilde{\mathbf{x}}_i$ est la valeur prévue d'après la régression par les MCO de

$\mathbf{x}_i$ sur $\mathbf{z}_{2,i}$, et $\hat{x}_{1i}$ est la valeur prévue d'après la régression par les MCO de x_{1i} sur $\mathbf{z}_{1,i}$. La statistique de test pour $\gamma_2 = 0$ est définie en (18) où γ_2 est le coefficient par les MCO pour $x_{1i} - \hat{x}_{1i}$. Comme γ_2 est un scalaire, représentons par t^2 la statistique de (18). Alors, l'estimateur de prétest en deux étapes est

$$\hat{\boldsymbol{\beta}}_{\text{pre}} = \begin{cases} \hat{\boldsymbol{\beta}}_{\text{ols}} & \text{si } \mathrm{F} < \mathrm{F}_{2,\mathrm{n}-4}(\alpha) \\ \begin{cases} \hat{\boldsymbol{\beta}}_{\text{IV2}} & \text{si } |\mathrm{t}| < \mathrm{Z}(\alpha/2) \\ \hat{\boldsymbol{\beta}}_{\text{IV1}} & \text{si } |\mathrm{t}| \geq \mathrm{Z}(\alpha/2) \end{cases} & \text{et } \mathrm{F} \geq \mathrm{F}_{2,\mathrm{n}-4}(\alpha), \end{cases} \tag{23}$$

où α est la taille du test.

Le calcul de l'erreur-type de $\hat{\boldsymbol{\beta}}_{pre}$ suit la méthode d'estimation de la variance appropriée pour l'estimateur choisi. Un estimateur de la variance est donné par

$$\hat{\mathrm{V}}\{\hat{\boldsymbol{\beta}}_{\text{pre}}\} = \begin{cases} \hat{\mathrm{V}}\{\hat{\boldsymbol{\beta}}_{\text{ols}}\} & \text{si } \mathrm{F} < \mathrm{F}_{2,\mathrm{n}-4}(\alpha) \\ \begin{cases} \hat{\mathrm{V}}\{\hat{\boldsymbol{\beta}}_{\text{IV2}}\} & \text{si } |\mathrm{t}| < \mathrm{Z}(\alpha/2) \\ \hat{\mathrm{V}}\{\hat{\boldsymbol{\beta}}_{\text{IV1}}\} & \text{si } |\mathrm{t}| \geq \mathrm{Z}(\alpha/2) \end{cases} & \text{et } \mathrm{F} \geq \mathrm{F}_{2,\mathrm{n}-4}(\alpha), \end{cases} \tag{24}$$

où $\hat{V}\{\hat{\boldsymbol{\beta}}_{IVj}\}$ est défini en (21). L'estimateur de la variance $\hat{V}\{\hat{\boldsymbol{\beta}}_{pre}\}$ n'est pas un estimateur de variance sans biais. Nous appelons la statistique

$$t_{\beta_{pre,m}} = [\hat{V}\{\hat{\beta}_{pre,m}\}]^{-1/2}(\hat{\beta}_{pre,m} - \beta_m) \tag{25}$$

pour β_m, $m = 0, 1,$ la statistique *t*, quoiqu'elle ne suive pas la loi *t* de Student.

Les tableaux 1 et 2 donnent l'erreur quadratique moyenne des estimateurs. Un échantillon a été créé en générant 1 000 vecteurs donnant une taille d'échantillon prévue de 221. Une taille de $\alpha = 0{,}10$ a été utilisée pour les estimateurs de prétest. La deuxième colonne du tableau 1 donne la corrélation entre e_i et p_i. Pour la corrélation modérée de 0,077 associée à une valeur de ψ de 0,01, l'estimateur par les MCO (*ols*) est inférieur à l'estimateur PP (*PW*). L'estimateur à variables instrumentales IV1 est plus efficace que l'estimateur PP, parce que l'estimateur IV1 contient un plus grand nombre de variables instrumentales que l'estimateur PP. L'estimateur IV2 est approprié pour notre processus de génération de données et est celui qui utilise le plus d'information. Par conséquent, l'estimateur IV2 est systématiquement supérieur à l'estimateur IV1. Les erreurs quadratiques moyennes de l'estimateur de prétest sont comprises entre l'erreur quadratique moyenne de l'estimateur par les MCO et celle de l'estimateur IV1. À mesure que la valeur de ψ augmente, l'erreur quadratique moyenne de l'estimateur de prétest se rapproche de l'erreur quadratique moyenne de l'estimateur IV1, parce que la procédure de prétest aboutit plus fréquemment au rejet de l'hypothèse nulle à mesure qu'augmente la corrélation entre p_i et e_i.

Tableau 1 : Erreur quadratique moyenne de Monte Carlo (×1000) des estimateurs de β_0 (10 000 échantillons)

ψ	Corr. (p_i, e_i)	$\hat{\beta}_{ols,0}$	$\hat{\beta}_{Pw,0}$	$\hat{\beta}_{IV1,0}$	$\hat{\beta}_{IV2,0}$	$\hat{\beta}_{pre,0}$ $\alpha = 0.10$
0	0,000	2,33	5,92	5,71	5,33	3,39
0,01	0,077	6,77	5,71	5,55	5,14	6,97
0,02	0,108	10,82	5,75	5,53	5,10	8,94
0,05	0,171	23,94	5,60	5,41	4,99	9,35
0,07	0,203	32,45	5,65	5,47	5,02	8,01
0,10	0,243	45,11	5,58	5,42	5,06	6,55
0,20	0,343	88,22	5,67	5,55	5,18	5,41
0,30	0,420	131,22	5,44	5,34	4,89	5,11
0,50	0,542	217,28	5,26	5,23	4,88	5,07

Tableau 2 : Erreur quadratique moyenne de Monte Carlo (×1000) des estimateurs de β_1 (10 000 échantillons)

ψ	$\hat{\beta}_{ols,1}$	$\hat{\beta}_{PW,1}$	$\hat{\beta}_{IV1,1}$	$\hat{\beta}_{IV2,1}$	$\hat{\beta}_{pre,1}$ $\alpha = 0.10$
0	4,16	9,62	8,53	4,29	5,12
0,01	4,30	9,87	8,61	4,32	5,61
0,02	4,41	9,71	8,63	4,32	5,93
0,05	4,66	9,54	8,49	4,34	6,18
0,07	4,94	9,80	8,64	4,46	6,49
0,10	5,32	9,69	8,57	4,58	6,52
0,20	6,47	9,48	8,39	4,84	6,56
0,30	7,91	9,30	8,25	5,20	6,66
0,50	10,29	9,10	8,25	5,76	6,97

Comme l'illustre les résultats des simulations présentés aux tableaux 3 et 4, les valeurs de la statistique *t* excèdent presque toutes la valeur tabulaire $t_{.025}$ pour le *t* de Student. On se souviendra qu'il existe une large gamme de probabilités de sélection, de sorte que la variance de l'estimateur de variance est supérieure à celle obtenue dans le cas d'un échantillon aléatoire simple. Les propriétés de la statistique de test sont généralement meilleures pour l'estimateur IV1 que pour les autres estimateurs. Comme prévu, pour l'estimateur de prétest, la statistique *t* est très biaisée lorsque l'ordonnée à l'origine réelle est proche de 1,5 écart-type de l'estimateur. À mesure que ψ augmente, $P(|\, t_{\beta_{pre},0} \,| > t_{.025})$ s'approche de $P(|\, t_{\beta_{IV1,0}} \,| > t_{.025})$.

Tableau 3 : Probabilité de Monte Carlo que $|\, t_{\beta_0} \,| > t_{.025}$ (10 000 échantillons)

ψ	$\hat{\beta}_{ols}$	$\hat{\beta}_{PW}$	$\hat{\beta}_{IV1}$	$\hat{\beta}_{IV2}$	$\hat{\beta}_{pre}$ $\alpha = 0.10$
0	0,049	0,058	0,057	0,053	0,065
0,01	0,282	0,065	0,066	0,061	0,237
0,02	0,486	0,061	0,061	0,057	0,320
0,05	0,870	0,055	0,056	0,051	0,247
0,07	0,950	0,065	0,065	0,062	0,167
0,10	0,990	0,059	0,059	0,055	0,086
0,20	1,000	0,058	0,060	0,055	0,059
0,30	1,000	0,059	0,064	0,059	0,063
0,50	1,000	0,060	0,064	0,060	0,065

Tableau 4 : Probabilité de Monte Carlo que $|\, t_{\beta_1} \,| > t_{.025}$ (10 000 échantillons)

ψ	$\hat{\beta}_{ols}$	$\hat{\beta}_{PW}$	$\hat{\beta}_{IV1}$	$\hat{\beta}_{IV2}$	$\hat{\beta}_{pre}$ $\alpha = 0.10$
0	0,049	0,069	0,066	0,051	0,063
0,01	0,054	0,073	0,070	0,055	0,072
0,02	0,057	0,073	0,068	0,053	0,074
0,05	0,072	0,070	0,065	0,056	0,080
0,07	0,077	0,070	0,068	0,057	0,085
0,10	0,083	0,073	0,069	0,054	0,081
0,20	0,119	0,071	0,067	0,052	0,074
0,30	0,154	0,076	0,072	0,053	0,076
0,50	0,233	0,074	0,072	0,054	0,070

Remerciements

Ces travaux ont été financés en partie par le contrat de coopération n° 68-3A75-4-122 conclu entre le Natural Resources Conservation Service de l'USDA et le Center for Survey Statistics and Methodology de la Iowa State University.

Références

Fuller, W. A. (2002), "Estimation par régression appliquée à l'échantillonage", *Techniques d'enquête,* 28:5-25.

Pfeffermann, D., et M. Sverchkov (1999), "Parametric and semi-parametric estimation of regression models fitted to survey data", *Sankhyā Series B: The Indian Journal of Statistics,* 61:166-186.

Wooldridge, J. M. (2000), *Introductory Econometrics: A Modern Approach,* South-Western Educational Publishing.