Techniques d'enquête, Juin 1997
Vol. 23, n° 1, pp. 61-68
Statistique Canada

Estimation des variables instrumentales des flux bruts en présence d'erreur de mesure

K. HUMPHREYS et C. J. SKINNER[1]

RÉSUMÉ

Les auteurs étudient le problème qui consiste à estimer les taux de transition au moyen des données d'une enquête longitudinale, lorsqu'il existe des erreurs de classification. Ils examinent les approches faisant appel à des données auxiliaires sur les taux de classification erronée et ainsi que d'autres approches pour modéliser l'erreur de mesure. À partir de variables instrumentales nominales, ils suggèrent comment identifier et estimer les modèles qui comprennent des variables de ce genre en considérant un modèle à structure latente restreinte. Enfin, ils étudient les propriétés numériques des estimateurs implicites des variables instrumentales pour les taux des flux grâce aux données de l'étude par panel sur la dynamique de revenu.

MOTS CLÉS: Structure latente; longitudinal; erreur de classification; taux de transition.

1. INTRODUCTION

Un des principaux avantages d'une enquête longitudinale est qu'elle permet d'estimer les flux bruts, par exemple le nombre de chômeurs qui trouvent un emploi (lire Hogue et Flaim 1986). Quand on estime ces flux, le biais introduit par l'erreur de mesure pose un problème majeur. En effet, si les erreurs de classification peuvent avoir tendance à s'annuler lorsqu'on estime des proportions transversales (Chua et Fuller 1987), on ne peut en dire autant quand il s'agit de flux longitudinaux.

La première tentative en vue de résoudre le problème de l'erreur de mesure devrait clairement être de réduire cette erreur dans les méthodes de mesure de l'enquête. Biemer, Groves, Lyberg, Mathiowetz et Sudman (1991) examinent les approches pertinentes, aussi ne les reprendrons-nous pas ici. Néanmoins, aussi bonnes que puissent être les méthodes d'enquête, l'erreur de mesure reste inévitable et on devra en compenser les effets à l'analyse.

En général, les méthodes utilisées pour compenser l'erreur de mesure reposent sur un modèle hypothétique quelconque du processus d'erreur. Plusieurs auteurs décrivent divers modèles qu'on examinera à la partie 2. Pour les identifier et les estimer, on a habituellement besoin d'informations complémentaires, par exemple les renseignements recueillis lors des réinterviews (lire Meyer 1988). Les réinterviews coûtant cher et leur but étant rarement d'estimer les paramètres de la distribution des erreurs de mesure dans la pratique (Forsman et Schreiner 1991), d'autres méthodes s'avèrent nécessaires en l'absence de données issues des réinterviews. Quand l'erreur de mesure porte sur des variables continues et qu'on manque d'informations auxiliaires sur la distribution de cette erreur, on recourt couramment à l'estimation des variables instrumentales (lire Fuller 1987, partie 1.4). Par variable instrumentale, on entend une variable faisant partie des données d'enquête associées à la variable réelle, comprenant une erreur de mesure, mais non corrélée avec cette dernière. Les variables instrumentales et les hypothèses qui s'y rattachent nous fournissent l'information susceptible de remplacer celle des réinterviews et de faciliter l'identification et l'estimation des paramètres du modèle, parmi lesquels se retrouve la variable réelle. Dans cet article, nous verrons comment adapter la méthode d'estimation des variables instrumentales pour estimer les flux entre états discrets. On se rend compte que les modèles à structure latente (lire Bartholomew 1987, Ch. 2) procurent un cadre général dans lequel les hypothèses sur la variable instrumentale correspondent à certaines restrictions imposées aux paramètres du modèle. Notre approche rappelle donc celles qui imposent des restrictions aux modèles à structure latente (van de Pol et de Leeuw 1986; van de Pol et Langeheine 1990).

2. MODÈLES

Nous n'étudierons que le cas à deux possibilités, $t = 1$ et $t = 2$. Soit r, le nombre d'états dans lequel une personne peut être classée à chaque occasion. Appelons les états classés à $t = 1$ et $t = 2$ X et Y, respectivement, et les états véritables, x et y. On suppose l'existence d'un modèle où la valeur vectorielle de (X, Y, x, y) est le résultat indépendant d'un vecteur aléatoire commun de distribution $\mathrm{pr}(X = i, Y = j, x = u, y = v)$.

La première hypothèse sur cette distribution, formulée par divers auteurs (lire Abowd et Zellner 1985; Poterba et Summers 1986 et Chua et Fuller 1987), et que nous reprendrons ici, est que les erreurs de classification effectuées aux deux occasions sont conditionnellement indépendantes, l'état véritable étant connu, c'est-à-dire

$$\mathrm{pr}(X = i, Y = j \mid x = u, y = v) = \mathrm{pr}(X = i \mid x = u, y = v)\,\mathrm{pr}(Y = i \mid x = u, y = v). \quad \text{(A1)}$$

[1] K. Humphreys, Department of Psychology, Stockholm University, S-106 91 Stockholm, Sweden; C. J. Skinner, Department of Social Statistics, University of Southampton, Southampton, SO17 1BJ, United Kingdom.

Pareille hypothèse se retrouve couramment dans les modèles généraux à variables latentes (lire Anderson 1959). Elle semble raisonnable quand les méthodes de mesure de l'enquête sont indépendantes aux deux occasions. Si on obtient X rétrospectivement de la même interview qui a servi à mesurer Y, il est raisonnable de croire que la tendance des répondants à fournir des réponses exagérément cohérentes lors d'une interview entraînera une association positive entre les erreurs de classification. Marquis et Moore (1990) le prouvent avec l'enquête sur le revenu et la participation aux programmes. La possibilité d'une hétérogénéité individuelle dans les probabilités d'erreur de classification (certains répondants pouvant se montrer plus fiables que d'autres) est une deuxième raison pour questionner l'hypothèse d'indépendance conditionnelle. On lira à ce sujet Skinner et Torelli (1993), et Singh et Rao (1995). À la partie 4, nous accepterons l'hétérogénéité en supposant seulement que le modèle se confirme dans les cellules d'un tableau des variables observées.

La seconde hypothèse fondamentale est que l'erreur de classification ne dépend que de l'état réel courant, si bien que

$$\mathrm{pr}(X = i \mid x = u, y = v) = \mathrm{pr}(X = i \mid x = u) = K_{xiu}, \quad \text{(A2)}$$

par exemple, et

$$\mathrm{pr}(Y = j \mid x = u, y = v) = \mathrm{pr}(Y = j \mid y = v) = K_{yjv}, \text{ par exemple.}$$

K_{xiu} et K_{yjv} définissent les matrices $r \times r$ des erreurs de classification $\boldsymbol{K}_x = [K_{xiu}]$ et $\boldsymbol{K}_y = [K_{yiv}]$. Soit $\boldsymbol{P}$, la matrice $r \times r$ ayant pour ij-ième élément $\mathrm{pr}(X = i, Y = j)$ et Π, la matrice $r \times r$ dont le uv-ième élément est $\mathrm{pr}(x = u, y = v)$. On obtient l'équation matricielle

$$\boldsymbol{P} = \boldsymbol{K}_x \Pi \boldsymbol{K}_y'. \quad (1)$$

La matrice Π renferme les paramètres auxquels on s'intéresse alors que c'est la matrice $\boldsymbol{P}$ qu'on peut estimer à partir des valeurs X et Y venant de l'échantillon. Si on dispose des estimations complémentaires $\boldsymbol{K}_x$ et $\boldsymbol{K}_y$ et qu'il ne s'agit pas de matrices singulières, il est possible de résoudre l'équation (1) pour estimer Π. Par ailleurs, si on peut vérifier l'état réel grâce à une réinterview, on peut estimer directement $\boldsymbol{K}_x$ et $\boldsymbol{K}_y$ (Abowd et Zellner 1985). D'un autre côté, si on n'obtient qu'une nouvelle classification indépendante à la réinterview, on ne pourra estimer que les matrices interview-réinterview

$$\boldsymbol{K}_x \Delta_x \boldsymbol{K}_x' \text{ et } \boldsymbol{K}_y \Delta_y \boldsymbol{K}_y'$$

où $\Delta_x = \mathrm{diag}[\mathrm{pr}(x = u)]$ et $\Delta_y = \mathrm{diag}[\mathrm{pr}(y = v)]$ (Chua et Fuller 1987). Chaque matrice interview-réinterview est symétrique et la somme de leurs éléments donne un. Elles ne renferment donc que $r(r+1)/2 - 1$ éléments d'information «indépendants». Puisque chaque colonne de la matrice $\boldsymbol{K}$ et la diagonale de la matrice Δ ont pour somme un, le nombre de paramètres inconnus à chaque occasion est $r(r-1) + r - 1 = r^2 - 1$. Par conséquent, la différence entre le nombre de paramètres et le nombre d'éléments d'information est $r^2 - 1 - r(r+1)/2 + 1 = r(r-1)/2$ à chaque occasion, et le modèle est mal identifié lorsque $r \geq 2$. Chua et Fuller (1987) proposent qu'on ajoute une hypothèse naturelle afin de faciliter l'identification, à savoir supposer que les erreurs de mesure ne sont pas biaisées ou que

$$\mathrm{pr}(x = i) = \mathrm{pr}(X = i), \mathrm{pr}(y = i) = \mathrm{pr}(Y = i)\, i = 1, \ldots, r. \quad (2)$$

Dans ce cas, les valeurs faussement positives et faussement négatives ont tendance à s'annuler lors de l'estimation transversale des proportions. La nouvelle hypothèse réduit le nombre de paramètres de $r - 1$ à chaque occasion. Malgré cela, le modèle reste mal identifié lorsque $r \geq 3$, et Chua et Fuller (1987) sont forcés d'introduire de nouvelles hypothèses.

Voyons maintenant comment identifier le modèle sans les données d'une réinterview. Dans une régression linéaire simple où la covariable comporte une erreur de mesure la covariable, l'approche de la variable instrumentale (Fuller 1987, partie 1.4) suppose l'existence d'une variable instrumentale «observée» W, corrélée à la covariable mais indépendante de l'erreur de mesure et de l'erreur de l'équation de régression. Nous avons élargi cette hypothèse au cadre actuel en faisant en sorte que W soit une *variable instrumentale* si elle n'est pas indépendante de x et si

W et (X, Y) sont conditionnellement indépendants étant donné (x, y), (A3)

si W et y sont conditionnellement indépendants étant donné x. (A4)

En général, nous laisserons W être une variable nominale comprenant un nombre arbitraire s de catégories même si, en pratique, $s = r$ puisqu'on désire que W soit étroitement lié à x. Une possibilité spécifique consiste à faire correspondre W à l'état classé au temps $t - 1$. L'usage de la valeur déphasée d'une «covariable» comme variable instrumentale remonte aux débuts des travaux sur l'estimation des variables instrumentales (lire Reiersol 1941; Durbin 1954). Dans le cas présent, l'hypothèse A4 se confirme si les états réels suivent un processus de Markov et si les erreurs de classification sont conditionnellement indépendantes comme dans A1.

On peut représenter le modèle issu des hypothèses (A1) à (A4) par le graphique d'indépendance conditionnelle de la figure 1. Chaque point représente une variable. Les points ne sont pas reliés si les variables correspondantes présentent une indépendance conditionnelle avec les autres variables.

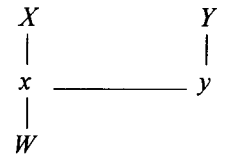

Figure 1. Représentation graphique de l'indépendance conditionnelle dans le modèle de base.

Le modèle illustre un modèle restreint à structure latente (Goodman 1974) où les variables X, Y et W observées sont conditionnellement indépendantes étant donné les variables latentes x et y, bref sont indépendantes dans les r^2 classes latentes définies par les couples (x, y). Ce modèle comprend

$2(r-1)r^2 + (s-1)r^2 + (r^2-1)$ paramètres donnés par les $(r-1)r^2$ paramètres $\mathrm{pr}(X = i \mid x = u, y = v)$, les $(r-1)r^2$ paramètres $\mathrm{pr}(Y = j \mid x = u, y = v)$, les $(s-1)r^2$ paramètres $\mathrm{pr}(W = k \mid x = u, y = v)$ et les $r^2 - 1$ paramètres libres $\mathrm{pr}(x = u, y = v)$. Ces paramètres sont soumis aux $2r(r-1)^2$ restrictions de (A2) et aux $(s-1)r(r-1)$ restrictions que suppose (A4). Nous nous attaquerons d'abord au cas $r = 2$. Dans ce cas, $4s + 7$ paramètres sont sujets à $2s + 2$ restrictions, ce qui laisse $2s + 5$ paramètres libres

$$\{K_{x2u}, K_{y2u}, \varphi_{2u}, \ldots, \varphi_{su}, \theta_u, \pi;\ u = 1,2\ v = 1,2\},$$

où $\varphi_{ku} = \mathrm{pr}(W = k \mid x = u)$, $\theta_u = \mathrm{pr}(y = 2 \mid x = u)$, et $\pi = \mathrm{pr}(x = 2)$. Le nombre probable de cellules «libres» dans la grille de X par Y par W est $r^2 s - 1$, ou $4s - 1$ quand $r = 2$. Une condition essentielle à l'identification quand $r = 2$ est donc que $4s - 1 \geq 2s + 5$ ou $s \geq 3$. Malheureusement, cette condition ne suffit pas. Soit

$$R_u = \mathrm{pr}(Y = 2 \mid x = u) = \sum_{v=1}^{2} K_{y2v}\, \theta_u^{v-1}(1 - \theta_u)^{2-v}, \qquad (3)$$

alors

$$\mathrm{pr}(X = i, Y = j, W = k) = \sum_{u=1}^{2} K_{xiu}\varphi_{ku} R_u^{j-1}(1 - R_u)^{2-j}\pi^{u-1}(1 - \pi)^{2-u}. \qquad (4)$$

Les $4s - 1$ probabilités des cellules libres sont donc déterminées par seulement les $2s + 3$ paramètres

$$\{K_{x2u}, \varphi_{2u}, \ldots, \varphi_{su}, R_u, \pi;\ u = 1, 2\}$$

qui ne seront identifiés que si $4s - 1 \geq 2s + 3$ ou $s \geq 2$. En réalité, cette condition suffit pour identifier les paramètres, sauf dans quelques combinaisons exceptionnelles. (Lire Madansky (1960) pour le cas où $s = 2$, et Goodman (1974) pour le cas général où $s \geq 2$.)

Toutefois, bien que les $2s + 3$ paramètres qui précèdent soient généralement identifiés lorsque $s \geq 2$, on ne peut déterminer les quatre paramètres K_{y21}, K_{y22}, θ_1 et θ_2, car ils ne sont associés qu'à deux paramètres connus, R_1 et R_2, au moyen de l'équation (3). Les paramètres principaux qui nous intéressent, notamment θ_1 et θ_2, restent mal identifiés, peu importe la valeur de s.

Pour les raisons qui précèdent, il faut imposer au moins deux autres restrictions au modèle si l'on veut identifier θ_1 et θ_2. À l'instar de Chua et Fuller (1987), on pourrait supposer que les erreurs de mesure ne sont pas biaisées comme dans (2), ce qui impose deux contraintes:

$$\pi = K_{x21}(1 - \pi) + K_{x22}\pi \qquad (5)$$

$$\theta_1(1 - \pi) + \theta_2\pi = R_1(1 - \pi) + R_2\pi. \qquad (6)$$

Malheureusement, la première contrainte ne s'applique qu'aux paramètres déjà identifiés lorsque $s \geq 2$, si bien qu'on ne peut identifier θ_1 et θ_2. Nous formulerons donc une hypothèse supplémentaire, c'est-à-dire que le processus d'erreur est constant dans le temps, de sorte que

$$K_{xiu} = K_{yiu} = K_{iu}, \text{ par exemple, pour } i, u = 1, 2, \ldots, r. \qquad (A5)$$

Pareille hypothèse semble élémentaire et naturelle si on applique la même procédure de mesure dans le cadre de l'enquête d'année en année. L'hypothèse élimine le problème de sous-identification du cas $r = 2$ examiné précédemment, puisqu'en identifiant $K_{xiu} = K_{iu}$ et R_u, on peut déterminer θ_u à partir de (3)

$$\theta_u = (R_u - K_{21})/(K_{22} - K_{21}) \qquad (7)$$

(en excluant le cas ordinaire où les variables quantifiées sont indépendantes des variables réelles, si bien que $K_{22} = K_{21}$).

En résumé, quand les hypothèses (A1) à (A5) se confirment et que $r = 2$, le modèle comporte $2s + 3$ paramètres libres $\{K_{2u}, \varphi_{2u}, \ldots, \varphi_{su}, \theta_u, \pi;\ u = 1,2\}$ que l'on peut identifier si $s \geq 2$, sauf dans les cas exceptionnels comme ceux analysés par Madansky (1960).

Enfin, revenons au cas général r. Puisque (A5) impose $(r-1)r$ restrictions, il s'ensuit qu'on compte $2(r-1)r^2 + (s-1)r^2 + (r^2-1) - [2r(r-1)^2 + (s-1)r(r-1)] - (r-1)r = 2r^2 + sr - 2r - 1$ paramètres libres. Il y a donc $r^2 s - 1$ cellules libres probables dans le tableau de X par Y par W et, dans l'ensemble, on pourra identifier le modèle si $r(r-1)(s-2) \geq 0$. La condition relative à l'identification des paramètres reste donc $s \geq 2$, pour une valeur quelconque de $r \geq 2$. On peut dire d'autre part que

$$R_{ju} = \Pr(Y = j \mid x = u) = \sum_{v=1}^{r} K_{jv}\theta_{uv}$$

où $\theta_{uv} = \mathrm{pr}(y = v \mid x = u)$. Par conséquent, il est possible de déterminer θ_{uv} à partir de R_{ju} et K_{jv}, également identifiés, pourvu que la matrice $[K_{iu}]$ ne soit pas singulière. Pour le cas général r, le modèle est identifié en vertu des hypothèses (A1) à (A5), sauf dans les cas exceptionnels examinés par Goodman (1974).

3. ESTIMATION

Supposons que pour un échantillon de taille n on retrouve le chiffre n_{ijk} dans les cellules du tableau de contingence $r \times r \times s$ de $X \times Y \times W$, et que les cellules aient une distribution multinomiale selon les paramètres n et $p_{ijk} = \mathrm{pr}(X = i, Y = j, W = k)$. La vraisemblance logarithmique implicite est

$$l = \sum_i \sum_j \sum_k n_{ijk} \log p_{ijk}.$$

En vertu d'un plan d'échantillonnage complexe, on peut considérer que n_{ijk} est un chiffre pondéré, ce qui débouche sur une pseudo-vraisemblance logarithmique (Skinner 1989). Les estimateurs des paramètres obtenus lorsqu'on maximise l porteront le nom d'estimateurs de *variables instrumentales* (VI).

Dans le reste de l'article, nous n'examinerons que le cas $r = s = 2$ en vertu du quel le modèle vient d'être identifié (sauf pour les valeurs aberrantes des paramètres). On pourrait établir que $p_{ijk} = n_{ijk}/n$ et résoudre les équations (6) et (7) pour les paramètres inconnus. Si elles se retrouvent dans les limites autorisées, c'est-à-dire dans la fourchette [0,1], les solutions résultantes correspondront aux estimations des VI. Dans la pratique cependant, on se rend compte que les échantillons de taille moyenne débouchent sur des solutions souvent irréalisables. De plus, les calculs sont complexes. Nous avons jugé plus facile de maximiser l directement au moyen des méthodes numériques de l'ensemble GAUSS (Edlefsen et Jones 1984) ou d'ensembles qui conviennent aux modèles à structure latente reposant sur l'algorithme EM, tel PANMARK (van de Pol, Langeheine et de Jong 1991). Un ensemble à structure latente permettrait d'ajuster un modèle à deux classes sans restriction, puis d'estimer θ_1 et θ_2 avec (7). Rien ne garantira néanmoins que les estimations résultantes tombent dans la fourchette [0,1] acceptable. S'ajouterait en outre la complication qu'on doive établir l'erreur-type de l'estimation de θ_1 et de θ_2 à partir de la matrice des covariances des estimations de $(R_1, R_2, K_{21}, K_{22})$. Il nous a semblé plus commode d'ajuster directement le modèle comme un modèle restreint à structure latente. Un autre avantage de cette approche est qu'elle s'étend naturellement à l'ajustement de modèles analogues d'un sous-groupe à l'autre, sous réserve de la contrainte éventuelle que certains paramètres restent constants entre les sous-groupes. Nous y reviendrons à la partie 4.

Avec les hypothèses multinomiales, il se pourrait que l'erreur-type repose sur la dérivée seconde du logarithme du rapport de vraisemblance obtenu pour l'estimation des VI. Pareille approche devient toutefois problématique quand la valeur maximale de l se trouve à la limite de l'espace paramétrique. Une solution consiste simplement à prendre les valeurs des paramètres telles que connues, à la limite. Le risque est qu'on sous-estime l'incertitude. Baker et Laird (1988) examinent deux autres approches au calcul de l'intervalle estimatif des paramètres dans des circonstances de ce genre: une méthode bootstrap et la méthode du profil de vraisemblance. La première méthode suppose le prélèvement réitéré d'échantillons multinomiaux où p_{ijk} est égal à n_{ijk}/n, et l'inscription de la distribution des paramètres estimatifs pour une suite d'échantillons bootstrap. La méthode du profil de vraisemblance donne les intervalles estimatifs d'un paramètre sous la forme des jeux de valeurs du paramètre non rejetées par un test du ratio de vraisemblance. Ces méthodes sont illustrées à la fin de la partie 4.

4. EXEMPLES NUMÉRIQUES

Aux fins d'illustration numérique, nous nous servirons des données du sous-échantillon à probabilité identique de l'étude américaine par panel sur la dynamique du revenu (EPDR) (lire Hill 1992). Nous examinerons deux états, occupé et non occupé, codés 1 et 2 respectivement, donc limiterons encore une fois notre attention au cas binaire. Pour plus de commodité, nous ne tiendrons pas compte des non-réponses et prendrons l'échantillon de 5,357 personnes de 18 à 64 ans de 1986 avec les valeurs entières pour les variables situation d'emploi en 1985, en 1986 et en 1987, possession d'une voiture, âge, sexe et degré de scolarité.

Les propriétés de l'estimateur VI sont évaluées de deux façons. Tout d'abord, à la partie 4.1, on compare le biais et l'erreur-type de l'estimateur VI à ceux de l'estimateur "non rajusté" des variables instrumentales hypothétiques, pour diverses associations avec x. Deuxièmement, à la partie 4.2, on étudie les conséquences de l'utilisation de différentes variables réelles de l'EPDR comme variables instrumentales.

4.1 Propriétés du biais et de l'erreur-type des estimateurs des variables instrumentales hypothétiques

Les paramètres qui nous intéressent le plus sont les probabilités conjointes $\mathrm{pr}(x = i, y = j)$ ou les probabilités conditionnelles $\mathrm{pr}(y = j \mid x = i)$ qui en dérivent. Les estimateurs simples "non rajustés" des paramètres reposent sur les proportions de l'échantillon correspondantes des variables X et Y classées et ont pour espérance $\mathrm{pr}(X = i, Y = j)$, pour un échantillonnage multinomial. Puisque $\mathrm{Pr}(X = i, Y = j)$ diffère généralement de $\mathrm{pr}(x = i, y = j)$, les estimateurs non rajustés sont habituellement biaisés. En supposant que les hypothèses (A1) à (A5) du modèle se vérifient, les estimateurs VI de $\mathrm{pr}(x = i, y = j)$ ne seront pas biaisés de façon asymptotique, même avec une variance supérieure à celle des estimateurs non rajustés. Nous essayerons ici de déterminer l'importance du compromis entre le biais des estimateurs non rajustés et le relèvement de la variance des estimateurs VI. On supposera que les hypothèses du modèle (A1) à (A5) tiennent et que l'échantillon est assez important pour que l'estimateur VI ne présente pas de biais.

Dans le cadre de notre analyse numérique, il est préférable d'utiliser certaines valeurs "réalistes" des paramètres. Celles-ci viennent de l'arrondissement des estimations du flux annuel survenu entre 1986 et 1987 selon les analyses de la partie 4.2 (voir le tableau 3). Les valeurs des cinq paramètres libres du modèle indépendant de W qui ont été retenues sont $K_{21} = 0.03$, $K_{22} = 0.94$, $\mathrm{pr}(x = 2) = \pi = 0.22$, $\mathrm{pr}(y = 2, x = 1) = \theta_1(1 - \pi) = 0.03$ et $\mathrm{pr}(y = 2, x = 2) = \theta_2 \pi = 0.19$. Les valeurs des deux autres paramètres libres $\varphi_{11} = \mathrm{pr}(W = 1 \mid x = 1)$ et $\varphi_{12} = \mathrm{pr}(W = 1 \mid x = 2)$ apparaissent dans des colonnes différentes du tableau 1. La statistique V de Cramér, qui définit le lien entre deux variables binaires en répartissant essentiellement la valeur chi carré sur un intervalle [0,1], résume la robustesse du lien entre les variables W et x. Pour chaque valeur possible des paramètres, le tableau 1 donne l'erreur-type estimative des estimateurs VI pour un échantillon $n = 5{,}357$ de l'EPDR. Le tableau 1 indique aussi le biais et l'erreur-type de l'estimateur non rajusté pour les mêmes valeurs paramétriques $K_{21}, K_{22}, \pi, \theta_1$ et θ_2, avec un échantillon de taille identique.

Pour montrer comment on calcule le biais des estimateurs non rajustés, considérons $\mathrm{pr}(x = 1, y = 1)$. L'espérance de l'estimateur non rajusté du paramètre, soit $\mathrm{pr}(X = 1, Y = 1)$, qu'on obtient à partir des valeurs $K_{21}, K_{22}, \pi, \theta_1$ et θ_2 et des

hypothèses (A1) à (A5), est 0.71. On doit la comparer à la valeur hypothétique de 0.75 pour $\mathrm{pr}(x = 1, y = 1)$. Le biais est donc de 0.71 – 0.75 = –0.04. Tel qu'indiqué précédemment, on suppose que le biais des estimateurs VI est égal à zéro. L'erreur-type des estimateurs non rajustés vient de la formule binomiale habituelle. Par exemple, l'erreur-type de l'estimateur non rajusté de $\mathrm{pr}(x = 1, y = 1)$ est $\sqrt{0.71 \times 0.29/5{,}357} = 0.0062$, où 0.71 est la valeur de $\Pr(X = 1, Y = 1)$. On obtient l'erreur-type des estimateurs VI en prenant la valeur inverse de la matrice d'information prévue, donnée par $n \sum p_{ijk} H_{ijk}$, où H_{ijk} est la matrice 7×7 des dérivées secondes de $\log p_{ijk}$ pour les sept paramètres libres. Après différenciation, on donne aux paramètres leur valeur hypothétique, tel qu'indiqué précédemment. Notons que l'erreur-type obtenue grâce à la matrice d'information multinomiale est sans doute sous-estimée étant donné le plan d'échantillonnage complexe de l'EPDR.

L'erreur-type de l'estimateur VI suit une tendance évidente puisqu'elle augmente à mesure que le lien entre W et x s'affaiblit. La hausse est assez semblable pour tous les paramètres, par exemple le ratio de V = 0.20 contre V = 1.00 se situe entre 3 et 4 pour l'ensemble des paramètres. Dans tous les cas, l'erreur-type de l'estimateur VI est supérieure à celle de l'estimateur non rajusté. La perte d'efficacité entre l'estimateur VI "idéal" (association parfaite entre W et x) et l'estimateur rajusté se situe entre ces deux pôles. Grosso modo, la perte est plus importante pour les paramètres conditionnels que pour les paramètres inconditionnels. Cette perte d'efficacité peut être vue comme la conséquence du rajustement résultant de l'erreur de mesure de y, laquelle reste nécessaire même quand W mesure x à la perfection. En vertu d'une telle interprétation, il est plausible que l'efficacité des paramètres conditionnels diminue davantage, car ils "dépendent moins" des paramètres de la distribution des x marginaux que les données sur W permettent d'estimer.

Pour étudier le compromis entre le biais de l'estimateur non rajusté et la plus forte variance de l'estimateur VI, nous avons calculé la taille minimale n de l'échantillon nécessaire pour que l'EQM de l'estimateur VI soit inférieure à celle de l'estimateur non rajusté. Pour les plans d'échantillonnage complexes, on suppose que la taille de l'échantillon correspond à la taille de l'échantillon après avoir prix en compte l'effet de plan. Le tableau 2 indique la valeur minimale de l'échantillon pour des forces d'association variables entre W et x. Sans erreur de classification, toutes les entrées se trouvent à l'infini puisque l'efficacité des estimateurs non rajustés est toujours supérieure à celle des estimateurs VI. Pour un nombre hypothétique d'erreurs, délimité par $K_{21} = 0.03$ et $K_{12} = 0.06$, la taille de l'échantillon requise augmente rapidement à mesure que V diminue. La variation dans les rangées du tableau 1 et la variation du biais des estimateurs non rajustés expliquent en partie l'écart entre les rangées du tableau 2. Le biais de l'estimateur non rajusté de $\mathrm{pr}(x = 2, y = 2)$ est donc relativement faible, ce qui se traduit par les fortes valeurs de la rangée correspondante, au tableau 2. Soulignons que la valeur de 1 pour $\mathrm{pr}(x = 2, y = 1)$ et pour le facteur V de Cramér résulte du fait que les deux estimateurs ont la même erreur-type (voir le tableau 1), si bien que le biais des estimateurs non rajustés suppose que l'EQM de l'estimateur VI est plus faible pour $n \geq 1$.

La principale conclusion qu'on retire du tableau 2 reste que dans un certain nombre de situations pratiques, l'estimation des VI s'avère utile pourvu que les hypothèses du modèle soient valables, même s'il faut légèrement augmenter la taille de l'échantillon afin d'accepter des plans d'échantillonnage complexes.

Tableau 1
Biais et erreur-type sous diverses VI hypothétiques

		Valeur des paramètres hypothétiques de l'estimateur VI						
$\mathrm{pr}(W = 1 \mid x = 1)$		1.0	0.1	0.1	0.1	0.3	0.1	0.5
$\mathrm{pr}(W = 1 \mid x = 2)$		0.0	0.9	0.7	0.5	0.7	0.3	0.3
V de Cramér		1.0	0.74	0.59	0.42	0.34	0.24	0.17
		Valeur des paramètres hypothétiques de l'estimateur VI						
Paramètre estimatif	Bias (× 100) de l'estimateur non rajusté	Estimateur non rajusté	Estimateur VI					
$\mathrm{pr}(x = 1, y = 1)$	-4.0	0.62	0.68	0.75	0.88	1.13	1.16	1.82	2.05
$\mathrm{pr}(x = 1, y = 2)$	3.0	0.32	0.39	0.43	0.51	0.64	0.69	1.03	1.24
$\mathrm{pr}(x = 2, y = 1)$	3.0	0.32	0.32	0.37	0.44	0.57	0.66	0.95	1.27
$\mathrm{pr}(x = 2, y = 2)$	-2.0	0.51	0.59	0.65	0.73	0.89	1.06	1.42	1.99
$\mathrm{pr}(y = 1 \mid x = 1)$	-3.9	0.37	0.55	0.55	0.64	0.81	0.88	1.30	1.58
$\mathrm{pr}(y = 1 \mid x = 2)$	12.4	0.60	1.40	1.63	1.95	2.56	2.90	4.30	5.55

Nota: 1 = occupé, 2 = non occupé, n = 5,357; échantillonnage multinomial; biais des estimateurs VI égal à zéro.

Tableau 2
Taille de l'échantillon nécessaire pour que l'EQM de l'estimateur VI soit inférieure à celle de l'estimateur non rajusté (échantillonnage multinomial)

	Valeur présumée du V de Cramér hypothétique pour les estimateurs VI						
	1.0	0.74	0.59	0.42	0.34	0.24	0.17
Paramètre estimé	Taille n de l'échantillon requise						
$\mathrm{pr}(x = 1, y = 1)$	28	59	132	300	320	971	1273
$\mathrm{pr}(x = 1, y = 2)$	31	50	91	184	219	573	843
$\mathrm{pr}(x = 2, y = 1)$	1	20	51	129	198	476	811
$\mathrm{pr}(x = 2, y = 2)$	112	227	366	720	1184	2397	5070
$\mathrm{pr}(y = 1 \mid x = 1)$	42	60	97	183	219	541	818
$\mathrm{pr}(y = 1 \mid x = 2)$	57	81	121	216	281	633	1061

4.2 Résultats avec les variables instrumentales réelles

Les résultats de la partie qui précède reposaient sur des variables instrumentales hypothétiques. Pour illustrer de façon plus réaliste notre propos, nous envisagerons maintenant des variables instrumentales réelles éventuelles. La principal difficulté consiste à trouver une variable W qui respecte les hypothèses (A3) et (A4). Il est apparemment plus facile de trouver une variable qui satisfera l'hypothèse (A3) que (A4), essentiellement parce qu'il est plausible que beaucoup de variables quantifiables sans erreur se plient à l'hypothèse (A3). Des variables sans lien avec un changement de la situation d'emploi, donc qui respectent l'hypothèse (A4), paraissent plus difficiles à trouver.

Aux fins d'illustration, nous avons envisagé deux possibilités. Tout d'abord, nous avons retenu la possession d'une voiture pour W (W = 2 si l'intéressé possède une voiture, W = 1 dans le cas contraire). Il se peut que la mesure de cette variable présente une erreur, mais a priori, il est raisonnable de supposer que l'erreur en question n'est pas liée à l'erreur de mesure concernant la situation d'emploi. Ainsi, lorsqu'ils analysent les erreurs résultant de l'inscription de la possession d'une voiture lors du recensement britannique de 1981, Britton et Birch (1985, p. 67) concluent que la principale difficulté associée au faible nombre d'aberrations dérive des véhicules hors d'usage ou disponibles de façon temporaire - ceux loués, par exemple - mais on peut au moins supposer qu'en réalité, ces erreurs présentent peu de liens avec le genre d'erreurs qui survient quand on enregistre la situation d'emploi. Parallèlement, la possession d'une voiture pourrait servir de valeur de remplacement à la situation sociale ou économique associée à un changement de la situation d'emploi, de sorte que l'hypothèse (A4) semble plus douteuse. Dans notre exemple, nous supposons toutefois qu'il y a vérification de (A3) et (A4).

Comme deuxième exemple, nous avons pris pour W la situation d'emploi déphasée en 1985. Dans ce cas, la difficulté est que l'hypothèse (A4) suppose effectivement un processus de Markov pour l'historique de l'emploi et que les taux de transition sont identiques alors qu'en réalité, ils varient d'une personne à l'autre, ce qui infirme l'hypothèse (A4) (lire van de Pol et Langeheine 1990). Pour pouvoir s'écarter de l'hypothèse (A4), il faut désagréger l'échantillon en 16 groupes, définis par classement croisé entre l'âge (quatre groupes), le sexe et le degré de scolarité (études collégiales ou non). On suppose ensuite que le modèle s'applique à l'intérieur d'un sous-groupe et on se sert des tests du rapport de vraisemblance pour déterminer quels paramètres restent constants d'un sous-groupe à l'autre. Ces tests ne donnent qu'une idée très générale, car ils ne tiennent pas compte du plan d'échantillonnage de l'EPDR. On n'a pas relevé de preuve sensible d'écart dans les probabilités d'erreur de classification K_{ij} d'un sous-groupe à l'autre. Il n'existe pas non plus d'indication sensible d'écart dans $\Pr(W \mid x$, sous-groupe) entre les deux sous-groupes du degré de scolarité, pour chacun des huit sous-groupes définis par l'âge et le sexe. En supposant l'égalité des paramètres, on obtient une valeur chi carré non significative du rapport de vraisemblance de la qualité de l'ajustement égale à 52.9 pour 46 dl (46 correspondant au nombre de cellules $= 16 \times 8 = 128$, moins les deux paramètres K_{ij}, moins $16 \times 4 = 64$ $\mathrm{pr}(x, y, \text{sous-groupe})$ paramètres moins $8 \times 2 = 16$ pr ($W \mid x$, sous-groupe) paramètres). Lorsqu'on combine les paramètres estimatifs du modèle désagrégé, on obtient l'estimation du flux global $\mathrm{pr}(x,y)$.

Le tableau 3 présente l'estimation des paramètres principaux pour les deux sortes de variables instrumentales et la version désagrégée du second choix. On remarque d'abord que l'erreur-type de l'estimateur VI reposant sur la possession d'une voiture est relativement élevée. Le tableau 1 le laissait prévoir, car x et W sont faiblement liés (le V de Cramér est de 0.12). Malgré cela, les rajustements résultants, qui relèvent la valeur estimative des entrées diagonales, sont plausibles et les intervalles de confiance issus de cet estimateur VI semblent plus réalistes que ceux de l'estimateur non rajusté.

L'erreur-type de la deuxième variable instrumentale choisie est plus faible, ainsi qu'on pouvait s'y attendre puisque le lien avec X est plus fort (V de Cramér de 0.73). En

réalité, l'erreur-type n'est pas beaucoup plus importante que celle de l'estimateur non rajusté. L'intervalle de confiance (de deux erreurs-types) ne chevauche plus l'intervalle correspondant de l'estimateur non rajusté des quatre paramètres.

Tableau 3
Estimation non rajustée et estimation VI selon les données de l'EPDR

Paramètre	Estimation non rajustée	Estimations VI		
		IV = possession d'une voiture	IV = emploi déphasé	IV = emploi déphasé (Désagrégé)
$pr(x = 1, y = 1)$	0.719	0.773	0.766	0.757
	(0.006)	(0.033)	(0.008)	(0.007)
$pr(x = 1, y = 2)$	0.055	0.011	0.017	0.025
	(0.003)	(0.020)	(0.005)	(0.003)
$pr(x = 2, y = 1)$	0.061	0.018	0.024	0.032
	(0.003)	(0.019)	(0.004)	(0.003)
$pr(x = 2, y = 2)$	0.166	0.198	0.193	0.186
	(0.005)	(0.027)	(0.007)	(0.006)

Nota: Erreurs-types en vertu d'une hypothèse multinomiale entre parenthèses. Désagrégation selon l'âge (4 groupes), le sexe et le degré de scolarité (2 groupes).

Ainsi qu'on l'a vu précédemment, on peut remettre en question l'hypothèse (A4) pour la variable correspondant à l'emploi déphasé. La version désagrégée de l'estimateur pose des hypothèses "plus faibles", n'exigeant seulement que (A4) se confirme à l'intérieur d'un sous-groupe. Les estimations qui en résultent se rapprochent passablement de l'estimateur VI original et leur erreur-type est légèrement plus faible, peut-être parce qu'on s'est servi des données complémentaires sur le sexe, l'âge et le degré de scolarité (voir l'analyse, plus loin). Fait intéressant, la désagrégation atténue les effets du rajustement d'une valeur relativement faible dans chaque cas. Il se pourrait qu'en s'écartant de (A4), on ait tendance à trop rajuster l'estimateur VI et que l'approche par désagrégation retenue ici contribue à atténuer le biais et permette d'évaluer la sensibilité des résultats selon la spécification du modèle, quand la désagrégation repose sur d'autres variables.

Ainsi qu'on a pu le lire à la partie 3, les estimations VI se retrouvent souvent à la limite de l'intervalle [0,1]. En réalité, parmi les analyses présentées au tableau 3, seule l'analyse par désagrégation donne des estimations limites. Pour les 64 paramètres $pr(x = i, y = j,$ sous-groupe) de $i, j = 1, 2,$ sous-groupe $= 1, ..., 16$, cinq estimations se retrouvent à la limite (contre aucune pour les 18 autres paramètres $pr(W = 1 \mid X = 1)$, et le reste). L'erreur-type indiquée au tableau 3 suppose que les paramètres sont connus, si bien qu'on pourrait sous-estimer l'incertitude de l'estimation des paramètres agrégés $pr(x = i, y = j)$.

Le tableau 4 donne d'autres estimations de l'erreur-type pour un sous-groupe, soit les hommes de 26 à 35 ans sans études collégiales. L'estimation de $pr(x = 1, y = 2)$ et celles qui en dérivent, notamment $pr(y = 1 \mid x = 1)$, se situent à la limite. Comme au tableau 3, l'estimation "type" de l'erreur-type repose sur la matrice des observations recueillies et on considère que les paramètres estimés à la limite sont connus. Les estimations bootstrap de l'erreur-type (pour 10,000 répétitions) se rapprochent beaucoup des estimations types pour les paramètres dont l'estimation ne se retrouve pas à la limite. Ainsi on ne possède pas d'estimation type de l'erreur-type pour l'estimation VI de $pr(x = 1, y = 2)$ située à la limite. En fait, estimer l'écart-type de la distribution de l'échantillon aurait peu de sens dans un tel cas. Il serait plus sensé d'établir un intervalle de confiance unilatéral ce qu'on peut faire avec la méthode du profil de vraisemblance, qui donne [0.016], ou avec la méthode bootstrap du percentile qui donne [0.009]. Les intervalles correspondant pour $pr(y = 1 \mid x = 1)$ sont [0.983,1] et [0.990,1].

Tableau 4
Autres estimations de l'erreur-type pour les hommes de 26 à 35 ans sans études collégiales

Paramètre	Estimations VI	Erreur-type estimative	
		Type	Bootstrap
$pr(W = 1 \mid x = 1)$	0.947	0.011	0.011
$pr(W = 1 \mid x = 2)$	0.107	0.089	0.091
$pr(X = 1 \mid x = 1)$	0.969	0.006	0.007
$pr(X = 1 \mid x = 2)$	0.084	0.088	0.075
$pr(x = 1, y = 1)$	0.953	0.011	0.012
$pr(x = 1, y = 2)$	0	*	*
$pr(x = 2, y = 1)$	0.006	0.007	0.006
$pr(x = 2, y = 2)$	0.041	0.012	0.011
$pr(x = 1)$	0.953	0.011	0.011
$pr(y = 1 \mid x = 1)$	1	*	*
$pr(y = 1 \mid x = 2)$	0.128	0.139	0.117

Nota: $n = 455$; les estimateurs «types» reposent sur la matrice des informations recueillies en vertu de laquelle on présume que les paramètres à la limite sont connus; 10,000 répétitions pour la méthode bootstrap; hypothèses multinomiales.

5. CONCLUSION

L'erreur de mesure peut introduire un biais important dans l'estimation type des taux de transition issus de données longitudinales. Si on peut estimer les taux d'erreur de classification de façon indépendante, il est possible de recourir à diverses méthodes de rajustement. Le présent article montre comment ajuster l'erreur de mesure par l'estimation des variables instrumentales en l'absence de telles données.

Comme c'est le cas pour la méthode d'estimation classique des variables instrumentales, la principale difficulté consiste à trouver une variable qui satisfera les conditions d'une variable instrumentale. En outre, même quand ces conditions sont satisfaites, il est bon que la variable présente un lien robuste avec la situation réelle si on veut obtenir des valeurs raisonnablement précises. Lorsqu'on trouve une variable de

ce genre, la méthode d'estimation des variables instrumentales peut donner de bons résultats.

REMERCIEMENTS

Les auteurs remercient Wayne Fuller de qui vient l'idée à la base de cet article. Les recherches n'auraient pu être menées à bien sans l'aide fournie par l'Economic and Social Research Council (subvention H519 25 5005) dans le cadre de son programme Analysis of Large and Complex Datasets.

BIBLIOGRAPHIE

ABOWD, J.M., et ZELLNER, A. (1985). Estimating gross labor force flows. *Journal of Business and Economic Statistics,* 3, 254-283.

ANDERSON, T.W. (1959). Some scaling models and estimation procedures in the latent class model. Probability and Statistics, (Éd. U. Grenander). Stockholm: Wiksell and Almquist.

BAKER, S.G., et LAIRD, N.M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association,* 83, 62-69.

BARTHOLOMEW, D.J. (1987). *Latent Variable Models and Factor Analysis.* London: Griffin.

BIEMER, P.P., GROVES, R.M., LYBERG, L.E., MATHIOWETZ, N.A., et SUDMAN, S. (1991). *Measurement Errors in Surveys* New York: Wiley.

BRITTON, M., et BIRCH, F. (1985). *1981 Census Post-Enumeration Survey.* London: Her Majesty's Stationery Office.

CHUA, T., et FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association,* 82, 46-51.

DURBIN, J. (1954). Errors in variables. *Revue de l'Institut International de Statistique,* 22, 23-31.

EDLEFSEN, L.E., et JONES, S.D. (1984). Reference Guide to GAUSS. Applied Technical Systems.

FORSMAN, G., et SCHREINER, I. (1991). The design and analysis of reinterview: an overview. Dans *Measurement Errors in Surveys.* (Éds., Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A. et Sudman, S.). New York: Wiley.

FULLER, W.A. (1987). *Measurement Error Models.* New York: Wiley.

GOODMAN, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika,* 61, 215-231.

HILL, M.S. (1992). *The Panel Study of Income Dynamics: A User's Guide.* Newbury Park, CA: Sage.

HOGUE, C.R., et FLAIM, P.O. (1986). Measuring gross flows in the labor force: an overview of a special conference. *Journal of Business and Economic Statistics,* 41, 111-21.

MADANSKY, A. (1960). Determinental methods in latent class analysis. *Psychometrika,* 25, 183-198.

MARQUIS, K.H., et MOORE, J.C. (1990). Measurement errors in the Survey of Income and Program Participation (SIPP): Program Reports. *Proceedings of the 1990 Annual Research Conference.* US Bureau of the Census, 721-745.

MEYER, B.D. (1988). Classification-error models and labor-market dynamics. *Journal of Business and Economic Statistics,* 6, 385-390.

POTERBA, J.M., et SUMMERS, L.H. (1986). Reporting errors and labor market dynamics. *Econometrica,* 54, 1319-1338.

REIERSOL, D. (1941). Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica,* 9, 1-24.

SINGH, A.C., et RAO, J.N.K. (1995). On the adjustment of gross flow estimates for classification error with application to data from the Canadian Labour Force Survey. *Journal of the American Statistical Association,* 90, 478-488.

SKINNER, C.J. (1989). Domain means, regression and multivariate analysis. Dans *Analysis of Complex Surveys,* (Ch. 3) (Éds. Skinner, C.J., Holt, D., et Smith, T.M.F). Chichester: Wiley.

SKINNER, C.J., et TORELLI, N. (1993). Measurement error and the estimation of gross flows from longitudinal economic data. *Statistica,* 53, 391-405.

VAN DE POL, F., et DE LEEUW, J. (1986). A latent Markov model to correct for measurement error. *Sociological Methods and Research,* 15, 118-141.

VAN DE POL, F., et LANGEHEINE, R. (1990). Mixed Markov latent class models. Dans *Sociological Methodology 1990,* (Éd. C.C. Clogg). Oxford: Basil Blackwell, 213-247.

VAN DE POL, F., LANGEHEINE, R., et DE JONG, W. (1991). PANMARK User Manual. Panel analysis using Markov chains. Version 2.2. Netherlands Central Bureau of Statistics.