Chapitre 2

Enquêtes longitudinales

1. Le traitement combiné des effets de non-réponse non-ignorable et de sondage informatif dans l'analyse des données issues des enquêtes longitudinales

Gad NATHAN [1] et Abdulhakeem EIDEH [2]

1.1 Introduction

Les données issues des enquêtes par sondage, et surtout par les enquêtes longitudinales, sont employées fréquemment pour inférer sur des modèles supposés. Souvent on ne tient pas compte des traits du plan du sondage (stratification, sondage par grappes ou à probabilités inégales) et les données venant de l'enquête par sondage sont analysées en employant des méthodes classiques, basées sur le plan de sondage aléatoire simple. Cette approche peut mener à des inférences erronées à cause du biais de sélection, impliqué par un plan de sondage informatif. Pour traiter les effets de tirage par probabilités inégales sur l'analyse de données issues des enquêtes longitudinales, Feder, Nathan et Pfeffermann (2000) ont appliqué des modèles hiérarchiques en combinaison avec des modèles de séries chronologiques. Pfeffermann, Krieger et Rinott (1998) ont proposé l'emploi de la distribution

1 Département de Statistique, Université Hébraïque de Jérusalem, gad@huji.ac.il.
2 Département de Mathématique, Université Alquds, Palestine, msabdul@ppu.edu.

dans l'échantillon induite par un modèle supposé pour la population, sous un plan de sondage informatif, pour une enquête en temps unique, et ont développé des expressions pour son calcul. Une approche similaire est employée par Nathan et Eideh (2004) et par Eideh et Nathan (2006), en proposant des modèles de séries chronologiques pour l'analyse des données issues des enquêtes longitudinales sous un plan de sondage informatif général.

En plus de l'effet du plan de sondage complexe, un des problèmes principaux pour l'analyse des données issues des enquêtes longitudinales est celui des données manquantes. Pour l'analyse longitudinale on cherche à mesurer une série d'observations pour chaque unité dans l'échantillon. Des données manquantes peuvent apparaître quand des observations sont indisponibles pour un ou plusieurs des temps de la série, ou par intermittence, ou pour une période continue jusqu'à la fin de la série.

Dans le contexte d'enquêtes par sondage, le traitement des données manquantes dans les enquêtes longitudinales est considéré, sur la base du plan de sondage, par Kalton (1986) et Lepkowski (1989). Pfeffermann et Nathan (2001) développent des méthodes de redressement des données manquantes dans les enquêtes longitudinales, par un modèle multiniveau intégré dans un modèle autorégressif. Skinner et Holmes (2003) proposent un modèle hiérarchique avec un effet aléatoire permanent au niveau de l'unité et des effets aléatoires temporaires, qui sont autocorrélés, pour les différentes périodes de l'enquête.

Dans cet article nous étudions le traitement combiné de non-réponse non-ignorable et de sondage informatif pour l'analyse des données issues des enquêtes longitudinales, par la spécification de la distribution jointe des observations quand le plan de sondage est informatif. Cette distribution décrit simultanément l'effet du plan de sondage informatif et celui de la réponse informative.

1.2 La distribution dans la population

Soit y_{it} la valeur observée pour l'unité $i(=1,...,N)$ en période $t(=1,...,T)$. Avec chaque valeur, y_{it}, sont associées les valeurs (connues), x_{itk}, $k(=1,...,p)$, de p variables explicatives. On suppose que les valeurs y_{it} suivent le modèle de régression : $y_{it} = \beta_1 x_{it1} + ... + \beta_p x_{itp} + \varepsilon_{it}$, où les valeurs de ε_{it} pour $t = 1,...,T$, sont une série aléatoire de longueur T, associée à chacun des N unités. La structure longitudinale des données suggère que les valeurs de ε_{it} sont corrélées à l'intérieur des unités.

Soit $\mathbf{y}_i = (y_{i1},...,y_{iT})'$, $\mathbf{x}_{it} = (x_{it1},...,x_{itp})'$, et soit $\boldsymbol{\beta} = (\beta_1,...,\beta_p)'$ le vecteur des coefficients de régression inconnus. Le modèle linéaire général multivarié pour les données longitudinales considère les vecteurs aléatoires $\mathbf{y}_i$, $i = 1,...,N$, comme des variables normales multivariées, qui sont distribuées $\mathbf{y}_i \mid \mathbf{x}_i \sim MVN(\mathbf{x}_i\boldsymbol{\beta}, \mathbf{V})$, où $\mathbf{x}_i$ est la matrice de taille $T \times p$ de variables

explicatives pour l'unité i, et $\mathbf{V}$ a pour élément $(jk): v_{jk} = \text{cov}_p(y_{ij}, y_{ik}),\ j, k = 1, ..., T,$ (Diggle, Liang et Zeger, 1994).

1.3 La distribution dans l'échantillon

Pour beaucoup d'exemples d'études longitudinales on emploie un sondage de panel, où les unités sélectionnées pour la première période restent dans l'échantillon jusqu'à la fin de l'étude (voir, par exemple, Nathan, 1999). Nous supposons, donc, un plan de sondage informatif à un degré pour un échantillon de panel sélectionné à temps $t = 1$ et que toutes les unités restent dans l'échantillon jusqu'au temps $t = T$. Il est raisonnable, alors, de supposer que les probabilités d'inclusion du premier ordre, π_i, dépendent des valeurs de la variable de réponse à la première occasion seulement, y_{i1}, et des valeurs des variables explicatives pour la première période, $\mathbf{x}_{i1} = (x_{i11}, ..., x_{i1p})'$. Si $\mathbf{y}_i \sim f_p(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta})$ est la distribution conditionnelle dans la population, la distribution marginale dans l'échantillon de $\mathbf{y}_i$, étant donné $\mathbf{x}_i$, est donnée par :

$$f_s(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{E_p(\pi_i|y_{i1}, \mathbf{x}_{i1}, \boldsymbol{\theta})}{E_p(\pi_i|\mathbf{x}_{i1}, \boldsymbol{\theta})} f_p(y_{i1}|\mathbf{x}_{i1}, \boldsymbol{\theta})\, f_p(y_{i2}, y_{i3}, ..., y_{iT}|y_{i1}, \mathbf{x}_i; \boldsymbol{\theta}) \qquad (1)$$

où $E_p(\pi_i \mid \mathbf{x}_{i1}, \boldsymbol{\theta}) = \int E_p(\pi_i|y_{i1}, \mathbf{x}_{i1}, \boldsymbol{\gamma})\, f_p(y_{i1}|\mathbf{x}_{i1}, \boldsymbol{\theta})\, dy_{i1}$. La démonstration de ce résultat est donnée par Eideh et Nathan (2006).

En supposant l'indépendance des observations dans la population, Pfeffermann, Krieger, et Rinott (1998) démontrent l'indépendance asymptotique des valeurs des unités sélectionnées sous la distribution dans l'échantillon, pour les plans de sondage avec des probabilités inégales, souvent employés. En conséquence, l'emploi de la distribution dans l'échantillon permet l'utilisation des procédures efficaces d'inférence standards, comme l'inférence basée sur le maximum de vraisemblance.

Notons qu'étant donnée la distribution dans la population, $f_p(y_{i1}|\mathbf{x}_{i1}, \boldsymbol{\theta})$, la distribution dans l'échantillon, $f_s(y_{i1}|\mathbf{x}_{i1}, \boldsymbol{\theta})$, est entièrement déterminée par les valeurs des espérances des probabilités d'inclusion, $E_p(\pi_i|y_{i1}, \mathbf{x}_{i1})$. Nous considérons les modèles approximatifs suivants pour ces espérances des probabilités d'inclusion, proposés par Pfeffermann, Krieger, et Rinott (1998) et par Skinner (1994) :

(a) Modèle exponentiel :

$$E_p(\pi_i|y_{i1}, \mathbf{x}_{i1}) = \exp(a_0^* + a_0 y_{i1} + a_1 x_{i11} + a_2 x_{i12} + ... + a_p x_{i1p}) \qquad (2)$$

(b) Modèle linéaire :

$$E_p(\pi_i|y_{i1}, \mathbf{x}_{i1}) = b_0^* + b_0 y_{i1} + b_1 x_{i11} + b_2 x_{i12} + ... + b_p x_{i1p} \qquad (3)$$

Eideh et Nathan (2004) considèrent, en plus, les modèles logit et probit pour les espérances des probabilités d'inclusion.