Données tronquées sous R dans les modèles linéaires simples et à effets mixtes

D. Thiam[a] et **G. Nuel**[a,b]

[a]Labo de Maths Appliquées (MAP5, CNRS 8145)
Université Paris Descartes
djeneba.thiam@gmail.com

[b] Institut des Maths et Interactions (INSMI)
CNRS Paris
gregory.nuel@parisdescartes.fr

Mots clefs : Données tronquées, modèle Tobit, modèles mixte, algorithme EM.

Nous nous intéressons à une variable réponse tronquée avec la possibilité d'avoir plusieurs seuils. Dans le cadre des modèles linéaires simples, le modèle Tobit [1], très populaire en économie permet la prise en compte des troncatures hautes et basses. Dans le cadre des modèles à intercept aléatoires, autrement appelé données de panel dans le domaine économique, Tobit adapte la présence d'effets aléatoires en utilisant une approximation de l'intégrale sur les effets aléatoires par la méthode de quadrature de Gauss Hermite [2]. D'autres alternatives sont possibles, via des algorithmes itératifs de type EM [3] , Marquart [4], Newton Ralphson [5, 6]. Ces derniers permettent une prise en compte plus générale des effets aléatoires.

Sous R la gestion des données tronquées dans les modèles linéaires se fait à l'aide des librairies `censReg` [7], `AER` [8]. Dans le cadre des données de panel, la librairie `censReg` dispose d'une option permettant la prise en compte de données à intercept aléatoire. Ces librairies sont rapides, et fournissent une estimation rapide des paramètres du modèle. Cependant certaines fonctionnalités pourraient être améliorées. Un premier point est celui du calcul des résidus du modèle en présence des troncatures. Le second point est la gestion des effets aléatoires et troncatures simultanément dans le cadre des modèles linéaires à effets mixtes avec troncatures.

L'objectif de ce travail est de présenter une approche plus générale pour la prise en compte des données tronquées en présences d'effets aléatoires. En effet en combinant l'algorithme EM, les lois conditionnelles et les sorties R des packages `lmer` [9] ou `censReg` [7], nous arrivons à obtenir une estimation approchée des paramètres du modèle linéaire à effets mixtes en présence de troncatures doubles ou multiples.

Considérons le modèle suivant:

$$\mathtt{y_{ij}} = \beta \mathtt{x_{ij}} + \mathtt{z_i} + \varepsilon_{\mathtt{ij}}$$

- $\mathtt{y_{ij}}$: variable de réponse avec possibilité de troncatures.
- $\mathtt{y} = (\mathtt{y_T}, \mathtt{y_0})$ ou $\mathtt{T} = \{i, j \text{ tq } \mathtt{y_{ij}} \text{ tronqué}\}$
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ résidus; $\mathtt{z_i}$ iid $\sim \mathcal{N}(0, \eta^2)$: effet aléatoire
- $\mathtt{x_{ij}}$: covariable ($\in \mathbb{R}^n$)

Nous nous proposons de résoudre ce problème d'estimation par l'algorithme EM. En considérant l'ensemble $(\mathtt{y_T}, \mathtt{z})$ comme donnée non observée, le paramètre θ du modèle à l'itération courante

est obtenu de la manière suivante:

$$M(\theta) = \arg\max_{\theta'} \underbrace{\int_{\mathtt{z}} \int_{\mathtt{y_T}} \mathbb{P}(\mathtt{z}, \mathtt{y_T}|\mathtt{y_0}; \theta) \log \mathbb{P}(\mathtt{y_0}|\mathtt{z}, \mathtt{y_T}; \theta') \, \mathrm{d}\mathtt{z} \, \mathrm{d}y_T}_{Q(\theta'|\theta)}$$

Afin de mettre à jour le paramètre courant en dehors de la procédure d'estimation, nous posons le problème différemment: en remarquant qu'à $\mathtt{z}$ (rep. $\mathtt{y_T}$) donné la vraisemblance $\mathbb{P}(\mathtt{y_0}, \mathtt{z}|\theta)$ (resp. $\mathbb{P}(\mathtt{y_0}, \mathtt{y_T}|\theta)$) est celle obtenue par `lmer` (resp. `tobit`). Ainsi deux techniques permettent d'obtenir le paramètre θ à l'itération courante :
1) EM combiné avec `lmer` du package `lme4`

$$M_{\text{lmer}}(\theta) = \arg\max_{\theta'} \underbrace{\int_{\mathtt{y_T}} \mathbb{P}(\mathtt{y_T}|\mathtt{y_0}; \theta) \log \mathbb{P}(\mathtt{y_0}, \mathtt{y_T}; \theta') \, \mathrm{d}y_T}_{Q(\theta'|\theta)}$$

2) EM combiné avec `tobit` des packages `censReg` ou `AER`

$$M_{\text{tobit}}(\theta) = \arg\max_{\theta'} \underbrace{\int_{\mathtt{z}} \mathbb{P}(\mathtt{z}|\mathtt{y_0}; \theta) \log \mathbb{P}(\mathtt{y_0}, \mathtt{z}; \theta') \, \mathrm{d}\mathtt{z}}_{Q(\theta'|\theta)}$$

Nous comparons ces deux techniques entres elles en terme d'estimation des paramètres, des résidus et des effets aléatoires. Ces méthodes sont aussi comparées aux alternatives R existantes comme `censReg` pour données panel, ou une implémentation d'un algorithme EM stochastique associé à du Gibbs sampling pour l'échantillonnage sous les lois conditionnelles.

Références

[1] J. Tobin (1958). Estimation of relationship for limited dependent variables. *Econometrica*, **26**, 24-36.
[2] J. Pan; R. Thompson (2003). Gauss-Hermite Quadrature Approximation for Estimation in Generalised Linear Mixed Models, **18**, 57-78.
[3] A. P. Dempster; N. M. Laird; D. B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm *Journal of the Royal Statistical Society* , **39**, 1-38.
[4] D. Marquardt (1963). Methods of Conjugate Gradients for Solving Linear Systems. *SIAM Journal on Applied Mathematics*, **11**, 1-462.
[5] F. Cajori(1911). Historical note on the Newton Raphson method of approximation. *Am. Math. Monthly*, **18**, 19-33.
[6] H. W. Richmond (1944). On the Newton-Raphson method of approximation *Edinburgh Math. Notes*, **44**, 5-8.
[7] A. Henningsen (2010). Estimating Censored Regression Models in R using the censReg Package. `http://cran.r-project.org/web/packages/censReg/vignettes/censReg.pdf`.
[8] C. Kleiber; A Zeileis (2010). Applied Econometrics with AER package version 1.1-7. `http://CRAN.R-project.org/package=AER`.
[9] D. Bates; M. Maechler; B. Bolker (2010).Package lmer. `http://cran.r-project.org/web/packages/c`