Italian Statistical Society

Statistical Methods for the analysis of large data-sets

UNIVERSITY G. D'ANNUNZIO
CHIETI-PESCARA

September 23-25, 2009

Book of short papers

University student performance analysis with non-ignorable drop-out

Filippo Belloc
Dip. di Economia Politica
Università di Siena
Email: filippo.belloc@unisi.it

Antonello Maruotti
Dip. di Istituzioni Pubbliche, Economia e Società
Università di Roma Tre
Email: antonello.maruotti@uniroma3.it

Lea Petrella
Dip. di Studi Geoeconomici, Linguistici, Statistici, Storici per l'Analisi Regionale
Sapienza Università di Roma
Email: lea.petrella@uniroma1.it

Abstract: Frequently, the interpretation of the data in university students' performance analysis is complicated by students' drop-out, so that some subjects are lost to follow-up before completion of their sequence of measurements, leading to an attrition problem. If the drop-out is non-ignorable, i.e. it depends on either unobserved values or an underlying response process, it may be a pervasive problem. In this paper, we consider the dependence between the primary response (student performance) and the drop-out mechanism with a suitable random effect model. We use data from the individual records of the faculty of Economics of Sapienza University of Rome to perform the empirical analysis.

Keywords: University student performance, mixed-effect hybrid models, non-ignorable drop-out

1. Introduction

This paper discusses a regression model for the analysis of longitudinal data in a generalized linear mixed models (GLMMs) framework; attention is focused on empirical situations where some measurements are *missing* for some units, due to attrition.
When complete follow-up data are not available for all subjects, inferences based on only observed data may be not valid. In particular, a drop-out mechanism depending on either unobserved values or an underlying response process can result in *non-ignorable missing data* (Little and Rubin, 2002).
In this paper, we perform a statistical analysis on longitudinal data in which the response is to be measured at progressive time points and some subjects are lost to follow-up because of drop-out. We use data from the individual students' records of the faculty of Economics of Sapienza University of Rome and consider individual performance as the response variable. The factors that affect the performance of graduates may differ from those affecting the performance of dropped-out students. Moreover, both individual student's drop-out and performance may depend on the same unobservable characteristics,

so that these characteristics simultaneously shape students' performance and sample selection. As a consequence, the estimated parameters may be far from the real parameters, if the underlying drop-out process is not taken into consideration. To tackle with such a problem, in our empirical analysis, we try to consider the dependence between the primary response (the observed student performance) and the drop-out mechanism with a suitable random effect model.
The model we propose is not a tool to state the non-ignorability of the drop-out; in fact, we need to conduct a sensitivity analysis under a range of different assumptions.

2. Statistical modeling

In longitudinal studies the problem of drop-out of some of the observed individuals is an important one. Likelihood based and estimation equation methods have been proposed to handle this problem. In particular, for the likelihood-based approach, Little (1995) identifies two broad classes of models: selection models (Diggle and Kenward, 1994; Follman and Wu, 1995) and pattern-mixture models (Little, 1994; Fitzmaurice et al., 2001). Little (2008) defines a new class of likelihood-based models, namely mixed-effect hybrid models (MEHMs), based on a new factorization of the likelihood of the outcome process and the drop-out process. Unlike selection models and pattern-mixture models, MEHMs factorize the likelihood of the outcome process and the drop-out process into the marginal distribution of random effects, the conditional distribution of the drop-out pattern given random effects, and the conditional distribution of the outcome given both random effects and the drop-out pattern. The resulting MEHMs have features of selection models in that they directly model the drop-out process, and also have features of pattern-mixture models in that the sample is stratified by the missing data patterns and the outcome process is modeled over these patterns. As a result, the MEHM, on the one hand, directly models the drop-out mechanism and, on the other hand, shares with pattern-mixture models the feature of computational simplicity (Yuan and Little, 2009). Suppose to collect K repeated measurements of a count response variable Y and coviariates X for each of the n individuals such that $\mathbf{Y}_i$=$(Y_{i1}, Y_{i2}, \ldots, Y_{iK})$ and $\mathbf{X}_i = (\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{iK})$ with $\mathbf{X}_{ik}$= $(X_{ik1}, X_{ik2}, \ldots, X_{ikp})$ denote the associated $K \times p$ covariates matrix. Since we consider only monotone missing data patterns, i.e., irretrievable dropout, let D_i index drop-out patterns such that $D_i = K$ for complete cases and $D_i = k$ if the subject i drops out between the kth and $(k+1)$th measurement time, for $k = 1, \ldots, K$; in formulas

$$D_i = K - \sum_{k=1}^{K} R_{ik} = \sum_{k=1}^{K} (1 - R_{ik})$$

where $R_{ik} = 1$ if the i-th unit drops out at any point within $(k-1, k)$, $k = 1, ..., K$, $R_{ik} = 0$ otherwise.
Let $\mathbf{b}_i$ be the random effects which model the correlation of repeated measurements on the same subject. The factorization of the joint distribution of $\mathbf{Y}_i$, $\mathbf{b}_i$ and D_i is

$$f(D_i, \mathbf{Y}_i, \mathbf{b}_i|\mathbf{X}_i) = f_B(\mathbf{b}_i|\mathbf{X}_i) f_{D|B}(D_i|\mathbf{b}_i, \mathbf{X}_i) f_{Y|D,B}(\mathbf{Y}_i|\mathbf{b}_i, \mathbf{X}_i);$$

in particular, the first two factors model the drop-out process, a feature of mixed-effects selection models, and the third factor models the longitudinal outcome process conditional on the pattern of *missing* data, a feature of pattern-mixture models. In other words, the

attrition is addressed in a straightforward way by the use of *potential* outcomes with a joint distribution (see e.g. Rubin, 2000).
In particular, let us assume that for some link function ζ the following model holds:

$$\zeta\left[E\left(D_i\,|\mathbf{b}_i\right)\right] = \mathbf{v}_i^{\mathsf{T}}\boldsymbol{\phi} + \mathbf{w}_i^{T}\mathbf{b}_i$$

where $\mathbf{v}_i$ is a (dropout-specific) covariate vector, and $\boldsymbol{\phi}$ represents the corresponding vector of model parameters, while $\mathbf{w}_i$ is a (dropout-specific) covariate whose effect is variable across subjects.
Without loss of generality, we will focus on random effect models, including some form of autoregression; this may help us distinguish between sources of true and spurious contagion, i.e. between dependence on past outcomes and the effects of individual, unobserved, characteristics.
Thus, assuming that variables whose effects are fixed and variable across subjects are collected in $\mathbf{x}_{ik}$ and $\mathbf{z}_{ik}$ (respectively), responses Y_{ik}, $i = 1, ..., n$, $k = 1, ..., K$ are assumed to be conditionally independent Poisson variates with canonical parameters defined by the following linear function

$$\theta_{ik} = \gamma d_i + \mathbf{x}_{ik}^{\mathsf{T}}\boldsymbol{\beta} + \alpha y_{i,k-1} + \mathbf{z}_{ik}^{T}\mathbf{b}_i, \quad k = 2, \ldots, K_i$$

where K_i is the number of measurements for each unit. A different model structure is defined for the first occasion, i.e.

$$\theta_{i1} = \mathbf{x}_{i1}^{\mathsf{T}}\boldsymbol{\beta}^{*} + \mathbf{z}_{i1}^{T}\mathbf{b}_i^{*}$$

$\mathbf{b}_i^{*} = \lambda\mathbf{b}_i$, to account for potential overdispersion in the random effect distribution when the lagged term is not available. To approximate the high-dimensional integration over the distribution of the random effects, non-parametric maximum likelihood estimation of the mixing distribution can be achieved in a finite mixture framework (see e.g. Aitkin, 1999). The use of finite mixtures has several significant advantages over parametric models; for instance, the discrete nature of the estimates helps us to classify subjects in clusters characterized by homogeneous values of random parameters. This is particularly appealing in social sciences, where components can be interpreted as groups with similar behaviors.

3. Data and Variables

In order to perform the empirical analysis, we use administrative data from the individual records of the faculty of Economics of Sapienza University of Rome, which cover 1639 students enrolled in the academic year 2003/2004 in a three-years bachelor program. We observe the students' characteristics by four-months data, within the third academic year of their program.
We consider the student performance as the primary response. Student performance may be defined in a variety of ways. To measure it, in this paper we use the number of ECTS credits obtained by the individual student every four-months. Moreover, we explicitly include faculty drop-out in our analysis and the related missing data generation process. Since our dataset is faculty specific, we do not make any distinction between the student withdraw from the university and student transfer to another faculty of the same

athenaeum. In particular, our definition of drop-out includes both the cases of who officially withdraws from the faculty and of who does not renew his or her registration in the next academic year.

We consider three dimensions shaping student performance: the degree course chosen by the individual student, his or her average mark on four-months basis and time invariant personal characteristics. First, we include the individual student's degree course (by means of a set of dummies) as one of the explanatory variables to tackle with the heterogeneity of the programs. Second, the individual student's average mark, measured at every observation time point, is considered to control for the part of the student success that remains overlooked by our index of performance: by measuring student performance by means of the number of credits, we do not evaluate potential qualitative differences between students with the same quantitative outcomes. Third, in the set of personal characteristics, we consider sex, citizenship, place of residence, type of diploma, secondary education final mark, age, latency period (as number of years between secondary education diploma and enrollment in the university) and a measurement of the student's household economic situation (ISEE).

Finally, we add the one-period-lagged response variable as one of the covariates. Every student has to reach the same amount of credits (180) to complete the program; thus, students with a high early performance may show lower outcomes in the last periods of their programs, and students with a low early performance may show higher outcomes later. By including a lagged variable we try to control for such dynamics.

References

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55:117–128.

Diggle, P. and Kenward, M G. (1994). Informative Drop-Out in Longitudinal Data Analysis, *Applied Statistics*, 43:49-73.

Fitzmaurice, G.M., Laird, N.M. and Schneyer, L. (2001). An Alternative Parameterization of the General Linear Mixture Model for Longitudinal Data with Non-Ignorable Drop-Outs, *Statistics in Medicine*, 20:1009-1021.

Follman, D. and Wu, M.C. (1995). An Approximate Generalized Linear Model with Random Effects for Informative Missing Data, *Biometrics*, 51:151-168.

Little, R.J.A. (1994). A Class of Pattern-Mixture Models for Normal Missing Data, *Biometrics*, 81:471-483.

Little, R.J.A. (1995). Modeling the Drop-Out Mechanism in Repeated-Measures Studies, Journal of the American Statistical Association, 90:1112-1121

Little, R.J.A. (2008). Selection and Pattern-Mixture Models, in *Advances in Longitudinal Data Analysis*, G. Fitzmaurice, M. Davidian, G. Verbeke and G. Nolenberghs (eds.), London: CRC Press.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd edition, New York: Wiley.

Rubin, D.B. (2000). The Utility of Counterfactuals for Causal Inference-Discussion of Causal Inference Without Counterfactuals by A. P. Dawid, Journal of the American Statistical Association, 95, 435–438.

Yuan, Y. and Little, R.J.A. (2008). Mixed-Effect Hybrid Models for Longitudinal Data with Nonignorable Dropout, *Biometrics*, DOI: 10.1111/j.1541-0420.2008.01102x.