Journal of Multivariate Analysis 131 (2014) 163–173

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

An analysis of longitudinal data with nonignorable dropout using the truncated multivariate normal distribution

Shahab Jolani

Department of Methodology and Statistics, Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 1 August 2013
Available online 5 July 2014

AMS subject classifications:
62N01
62N02
62H12

Keywords:
Incomplete data
Missing not at random
Nonresponse
Selection bias

ABSTRACT

For a vector of multivariate normal when some elements, but not necessarily all, are truncated, we derive the moment generating function and obtain expressions for the first two moments involving the multivariate hazard gradient. To show one of many applications of these moments, we then extend the two-step estimation of censored regression models to longitudinal studies with nonignorable dropout, in the sense that the probability of dropout depends on unobserved, or missing, observations. With nonignorable dropout, direct maximization of the likelihood function can be computationally intensive or even infeasible. The two-step method in such cases can be an adequate substitute. In a set of simulation studies the developed two-step method and the maximum likelihood (ML) method are compared. It turns out that the proposed method preforms at least as well as the ML and provides a convenient alternative that can easily be implemented in standard software.

1. Introduction

Incomplete data are a practical problem in many areas of social or medical studies. In longitudinal studies, for example, researchers are usually confronted with dropout, that is, some participants leave the study permanently and their outcome measurements are missing. Any meaningful analysis of longitudinal data with dropout should take the process of creating incompleteness into account because the resulting conclusions highly depend on the dropout process. Following the terminology in [16,13,1] provide a useful classification of dropout processes: completely random dropout (CRD), random dropout (RD) and nonignorable (or informative) dropout (NIG). Ignoring the dropout process (i.e., assuming CRD and/or RD) is not always the best solution, or even realistic in many cases. For instance, a longitudinal study on the effect of a treatment on depression can be an example of NIG, where the most severely affected patients tend to be the most dropouts because they are too ill to participate in the study on a regular basis. Consequently, rendering an NIG process cannot generally be disregarded and such incomplete data should be treated with great care.

Under NIG, the analysis of incomplete outcomes should be based on joint modeling of the scientific interest model (e.g., a regression of the outcomes on explanatory variables) and the process governing the dropout (i.e., the dropout process). One possible approach is the so-called selection model [8], which decomposes this joint distribution to the marginal distribution of the outcomes (the scientific interest model) and the conditional distribution of the dropout process given the outcomes.

Heckman [8] and Diggle and Kenward [1], among the others, assumed a (multivariate) normal distribution for the marginal distribution of the outcomes and a probit (or logit) function for the dropout process. Two general procedures exist for parameter estimation of this model: Maximum likelihood (ML) and the Heckman's two-step. Broadly speaking, the former is more efficient than the latter [1]. The direct ML, however, can be computationally more expensive, particularly in

E-mail address: s.jolani@uu.nl.

http://dx.doi.org/10.1016/j.jmva.2014.06.016
0047-259X/© 2014 Elsevier Inc. All rights reserved.

multivariate missing data (e.g., in longitudinal studies with dropout), because it requires evaluation of multiple probability integrals. The estimation procedure is therefore very challenging and even infeasible in such cases, and there are not many available statistical programs to maximize the likelihood function when it involves the evaluation of high-dimensional integrals. The likelihood function may also have several local maxima and one should be prepared to repeat the analysis several times, with different starting values of the parameters, to guard against misleading inference.

The two-step method, on the other hand, offers a straightforward and less computationally challenging alternative that can be implemented with standard statistical programs. The estimators of the two-step method are also asymptotically equivalent to the ML estimates, or at the very least can be considered as start values for the ML estimation. In this paper we thus develop the two-step method for longitudinal studies with dropout.

It should be emphasized that implementation of the two-step method requires numerical computation of multiple integrals in a multivariate normal distribution. However, this is not an issue anymore because several algorithms have been proposed for calculation of these integrals. Genz [3,4], among the others, proposed an algorithm that numerically computes the multivariate normal integral by applying randomized lattice rule on the transformed integral. Details can be found in [5]. The approach of Genz can effectively evaluate multiple integrals up to 100 dimensions, and is widely available in major statistical programs such as R (pmvnorm function), Stata (ghk function), Matlab (mvncdf function), SAS (mvt SAS/IML program), and many more.

The two-step method estimates the model parameters in two steps. The first step models the probability of observing the outcome by a (multivariate) probit model, and the second step models the expectation of the outcome conditional on having been observed, i.e., a regression of the outcome on the explanatory variables only for the observed outcomes. The key element to implement the two-step method is the evaluation of the conditional distribution of the outcomes in the second step. To do so, moments of a truncated multivariate normal distribution have to be calculated. Many authors have reported the calculation of the moments of a truncated multivariate normal distribution (e.g., [9,7]). These calculations, however, have been done by direct integration, which are suitable only for low-dimensional multivariate normal distributions. In contrast, [14] obtained compact expressions for the first two moments of a truncated multivariate normal distribution using the moment generating function (MGF). The author considered a case where all elements of a multivariate normal random vector are truncated on the left. As opposed to [14], we here need to consider situations in which the multivariate normal random vector is truncated in some elements, but not necessarily all, so we derive the MGF along with the first two moments when some elements of the multivariate normal random vector are truncated. To show one of many applications of these moments, we then use them to develop the two-step method in longitudinal studies with dropout.

It is important to mention that the two-step and ML methods require invoking assumptions about the distribution of the outcomes, i.e. the multivariate normality. This assumption, however, cannot be verified from the observed data, and the results are highly sensitive to departure from normality. Therefore, any conclusion can be misleading if this assumption is violated. We will return to this point in the conclusion section.

The remainder of this paper is organized as follows. The MGF for the truncated multivariate normal distribution when some of its elements are truncated from the left is derived in the next section, as well as the expressions for the first two moments. Section 3 uses these expressions to extend the two-step method to longitudinal studies with dropout. A comparison of the ML and two-step methods through a set of simulation studies is presented in Section 4. Some concluding remarks are given in the last section.

2. Truncated multivariate normal distribution and its moments

We start with some notation. Let $\mathbf{Z} = (Z_1, \dots, Z_k)$ be a multivariate standardized normal random variables with zero mean vector and correlation matrix $\mathbf{R}$, where its diagonal and off-diagonal elements are 1 and ρ_{ij}, respectively. Also, let $\nabla_{\mathbf{z}}$ denote the gradient operator $(\partial/\partial z_1, \dots, \partial/\partial z_k)'$ so that $\nabla_{\mathbf{z}} f(\mathbf{z})$ is the gradient vector and $\nabla_{\mathbf{z}}\nabla'_{\mathbf{z}} f(\mathbf{z})$ is the Hessian matrix for any function $f(\mathbf{z})$. We use the abbreviation ∇^2 for $\nabla\nabla'$ too. For a constant vector $\mathbf{a}$, $\nabla(\mathbf{a})$ denotes the gradient of $f(\mathbf{z})$ evaluated at $\mathbf{a}$. For a vector transformation $\mathbf{w}(\mathbf{z})$, $\nabla[\mathbf{w}(\mathbf{z})]'$ is the matrix of the first partial derivatives of the elements of $\mathbf{w}$ with ijth element $\partial w_j/\partial z_i$. We suppress subscripts i and j from the notation for convenience. Using the 'chain rule' of derivation, we can write

$$\nabla_{\mathbf{z}} f(\mathbf{w}(\mathbf{z})) = \nabla_{\mathbf{z}}[\mathbf{w}(\mathbf{z})]' \nabla_{\mathbf{w}} f(\mathbf{w}(\mathbf{z})).$$

Now suppose that some elements of $\mathbf{Z}$, say k_1, are fully observed but the rest$(k_2 = k - k_1)$ is truncated on the left by values in the vector $\mathbf{a} = (a_{k_1+1}, \dots, a_k)'$. Without loss of generality, we assume that the first k_1 variables are not truncated. In order to construct the MGF, $\mathbf{Z}$ needs to be partitioned as

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{pmatrix},$$

where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are the sets of non-truncated and truncated variables, respectively. Accordingly, the correlation matrix $\mathbf{R}$ is partitioned as

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{pmatrix},$$