Semi-parametric Bayesian analysis of binary responses with a continuous covariate subject to non-random missingness

Frederico Z. Poleto[1], Carlos Daniel Paulino[2,5], Julio M. Singer[3], and Geert Molenberghs[4]

[1]IME, University of São Paulo, BRAZIL
[2]CEAUL and IST, Technical University of Lisbon, PORTUGAL
[3]IME, University of São Paulo, BRAZIL
[4]Hasselt University, BELGIUM
[5] Corresponding author: Carlos Daniel Paulino, e-mail: dpaulino@math.ist.utl.pt

Abstract

Missingness in explanatory variables requires a model for the covariates even if the interest lies only in a conditional model for the outcomes given the covariates. An incorrect specification of the models for the covariates or for the missingness mechanism may lead to biased inferences for the parameters of interest. Previously published articles either use semi-/non-parametric flexible distributions for the covariates and identify the model via a MAR assumption, or employ parametric distributions for the covariates and allow a more general non-random missingness mechanism. We consider the analysis of binary responses, combining a MNAR mechanism with a non-parametric model based on a Dirichlet process mixture for the continuous covariates. We illustrate the proposal with simulations and by analyzing a real dataset.

Keywords: Dirichlet process mixture, incomplete data, MNAR, non-ignorable missingness mechanism.

1. Introduction

In many studies, data are missing for some explanatory variables ($\mathbf{X}$), and in order not to exclude either these sampling units or these variables from the analysis, we need to specify a model for their marginal distribution, or at least, for the conditional distribution of the explanatory variables that may be missing given the explanatory variables that are always observed, even if the interest lies only on the conditional distribution of the response variables ($\mathbf{Y}$) given $\mathbf{X}$.

In cases where at least one explanatory variable is continuous, we may not have a priori any information for a plausible parametric model. Incorrect assumptions, either for the missingness mechanism or for the distribution of the covariates, may generate biased inferences for the conditional distribution of the responses given the covariates. Therefore, we pragmatically adopt a Bayesian methodology for the sensitivity analysis of the missingness mechanism allowing it to be non-random and consider also a flexible distribution for $\mathbf{X}$ via a non-parametric model based on a Dirichlet process mixture (Ishwaran and James, 2002). For simplicity, we restrict ourselves to the case of a single missing continuous covariate. In Section 2, we describe a semi-parametric model, which is to be compared to alternative parametric models by means of a simulation study in Section 3. A real dataset is analysed in Section 4 through a semi-parametric model that attempts to particularly incorporate a few prior judgements about the missingness mechanism.

2. A semi-parametric model for binary responses with a continuous covariate subject to non-random missingness

Let Y_i denote a binary response always observed, X_i, a continuous covariate with potentially missing values, and R_i, an indicator variable assuming the value of 1 if X_i is observed or 0, if X_i is missing, $i = 1, \ldots, n$. Although interest lies only in the conditional distribution of Y_i given X_i, it is necessary to consider a model for X_i, as we do not want to discard the portion of the sample wherein X_i is missing. As we admit that the missing data generating mechanism may depend on the unobserved values, we also need to model R_i.

Employing the so-called selection model factorization, we consider the model

$$R_i|(Y_i, X_i, \delta_0, \delta_1, \delta_2, \delta_3) \overset{\text{ind.}}{\sim} \text{Bern}(\theta_i),\ \text{logit}(\theta_i) = \delta_0 + \delta_1 X_i + \delta_2 Y_i + \delta_3 X_i Y_i, \quad (1)$$

$$Y_i|(X_i, \beta_0, \beta_1) \overset{\text{ind.}}{\sim} \text{Bern}(\pi_i),\ \text{logit}(\pi_i) = \beta_0 + \beta_1 X_i, \quad (2)$$

$$X_i|(\mu_i, V) \overset{\text{ind.}}{\sim} N(\mu_i, V), \quad (3)$$

where $\text{Bern}(\theta_i)$ denotes the Bernoulli distribution with success probability θ_i, $i = 1, \ldots, n$, along with the prior distributions $\delta_j|(\mu_{\delta_j}, \sigma_{\delta_j}) \overset{\text{ind.}}{\sim} N(\mu_{\delta_j}, \sigma_{\delta_j})$, $j = 0, 1, 2, 3$, $\beta_j|(\mu_{\beta_j}, \sigma_{\beta_j}) \overset{\text{ind.}}{\sim} N(\mu_{\beta_j}, \sigma_{\beta_j})$, $j = 0, 1$, $(\mu_1, \ldots, \mu_n)|G \overset{\text{i.i.d.}}{\sim} G$, $G|\alpha, G_0, M \sim \text{TDP}(\alpha, G_0, M)$, $V|T \sim \text{Unif}[0, T]$, $\alpha|(\lambda_1, \lambda_2) \sim Ga(\lambda_1, \lambda_2)$, G_0 $|(\mu_0, \tau) = N(\mu_0, \tau)$, $\mu_0|(a, A) \sim N(a, A)$, all mutually independent. The symbol TDP means a truncated Dirichlet process where the truncation point M was based on the argument put forward by Antoniak (1974), and Ishwaran and James (2002).

The model is considered semi-parametric because it employs a non-parametric structure for the marginal model of X_i and conventional parametric structures for the conditional distributions of Y_i given X_i and R_i given Y_i and X_i.

The missingness mechanism (1) is non-random because it considers that the probability of having missing covariates may depend on their unobserved values. On the other hand, if we include the missing at random assumption $\delta_1 = \delta_3 = 0$, the missingness mechanism becomes ignorable under the viewpoint of Bayesian inferences for β_0 and β_1 due to the assumed prior independence between (δ_0, δ_2) and the other parameters (Little and Rubin, 2002). A subclass of the MAR model is the missing completely at random (MCAR) mechanism that can be formulated by setting $\delta_1 = \delta_2 = \delta_3 = 0$. In this setup with missingness in explanatory variables, it is important to note that the so-called complete case analysis (CCA), where units with missing data are discarded, commonly generates unbiased inferences for β_0 and β_1 not only under the MCAR mechanism but also under any other missingness mechanisms that do not depend on the response Y_i such as in the reduced version of the missing not at random mechanism, $\text{MNAR}_{\text{red}} : \delta_2 = \delta_3 = 0$. A CCA of data generated under the non-random missingness mechanism MNAR_{red} results in biased inferences for the marginal distribution of X_i, but not for the conditional distribution of Y_i given X_i. Also, the CCA does not require to specify a marginal model for X_i if the interest lies only in the conditional distribution of Y_i given X_i.

3. Some results from a simulation study

We consider the following distributions for the explanatory variable

$$\begin{aligned}
&X^N \sim N(12, 3^2), && (4)\\
&X^L \sim \text{Log-normal}(2.45, 0.246^2), && (5)\\
&X^C = 0.8 \times X^{C1} + 0.2 \times X^{C2}, && (6)\\
&\qquad X^{C1} \sim \text{Unif}[8, 12], X^{C2} \sim \text{Log-normal}(2.79, 0.642^2),
\end{aligned}$$

where Log-normal(μ, σ^2) denotes a log-normal distribution, and μ and σ are, respectively, the mean and the standard deviation of the underlying variable on the logarithmic scale. The mean and the standard deviation of X^L and X^C coincide with the corresponding parameters of X^N, although the densities are very different.

In order to assess the impact of results obtained under different distributional assumptions for the covariate, we generated a sample of X of size $n = 10,000$ from each of the three distributions (4), (5) and (6); then, for each value generated under each of the distributions of the covariates, we generated Y from (2) with $\beta_0 = 6$ and $\beta_1 = -0.5$; finally, we generated R from (1) with $\delta_0 = -3$, $\delta_1 = 0.5$ and $\delta_2 = \delta_3 = 0$. For each of the three generated datasets (with X^N, X^L e X^C), we fitted the semi-parametric model of the previous section as well as normal and log-normal parametric models. For normal and log-normal parametric models, the non-parametric model (3) is replaced, respectively, by

$$X_i|\mu_0, \tau \overset{\text{i.i.d.}}{\sim} N(\mu_0, \tau), \quad i = 1, \ldots, n, \qquad (7)$$

$$X_i|\mu_0, \tau \overset{\text{i.i.d.}}{\sim} \text{Log-normal}(\mu_0, \tau), \quad i = 1, \ldots, n, \qquad (8)$$

provided with vague priors. For all models, we adopted vague prior distributions for δ_j and β_j employing the hyper-parameters $\mu_{\delta_j} = \mu_{\beta_j} = 0$ and $\sigma_{\delta_j} = \sigma_{\beta_j} = 10^3$, $j = 0, 1$. Furthermore, we always assumed the correct structure for the missingness mechanism, i.e., $\delta_2 = \delta_3 = 0$, so that the only varying components in the study are the distribution employed to generate the covariate and the distribution adopted for the covariate in the analysis.

The samples obtained from the posterior distributions of the parameters β_0 and β_1 indicate that the non-parametric model for the covariate generates results very close to those obtained with the corresponding true parametric model under either the normal or the log-normal distributions. In these cases, the credible intervals contain the true values; this does not occur in the analyses under the incorrect parametric models for X^N and X^L. On the other hand, in the case of X^C, only the credible intervals of the analysis under the non-parametric model for the covariate contained the true values of β_0 and β_1.

4. Analysis of pulmonary embolism data

Wicki *et al.* (2001) analyzed data from 1,090 patients that were consecutively admitted to the emergency ward of the University Hospital of Geneva for suspected pulmonary embolism, i.e., blockage of the main artery of the lung or one of its branches. The objective of their study was to develop a scoring system that would indicate the probability of occurrence of this cardiovascular disease based on diagnostic tests and other easily obtained information. For simplicity, we consider here only some of the explanatory variables included in the final model presented by these authors.

The indicator of the presence of pulmonary embolism (response variable) as well as four explanatory variables (age, previous pulmonary embolism or deep vein thrombosis, recent surgery, and pulse rate) were observed for all patients, while two variables that indicate presence of certain characteristics (platelike atelectasis and elevation of hemidiaphragm) had missing values for a single patient who, for this reason, was removed from the data set. On the other hand, the partial pressure of carbon dioxide (PaCO$_2$), obtained from arterial blood gas analysis, was missing for 103 (9%) patients.

Preliminary analyses allow us to show that the observed data for PaCO$_2$ seem to be better accommodated by the posterior predictive distribution of the non-parametric model than by the corresponding densities of normal, log-normal and gamma models. On the other hand, they showed no evidence of association between PaCO$_2$ and the other explanatory variables. Having this in mind, we considered a marginal rather than a conditional non-parametric model for PaCO$_2$.

Information obtained from authors of Wicki *et al.* (2001) allows us to come to a missingness model for PaCO$_2$ found appropriate in the light of the data. We have considered a conditional model for the response indicating pulmonary embolism given the covariates that followed the structure mentioned in Section 2.

Analyses of all available data based on the global model referred to above prove to be more suitable than complete case analyses, because, by embedding assumptions about missing data, they should provide less biased results on the association between pulmonary embolism and PaCO$_2$, and generate more precise results for the other associations.

Acknowledgements

We gratefully acknowledge the financial supports to this research: Frederico Z. Poleto and Julio M. Singer, from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brazil, and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil; Carlos Daniel Paulino, from Fundação para a Ciência e Tecnologia (FCT) through the research centre CEAUL-FCUL, Portugal and project Pest-OE/MAT/UI0006/2011; Geert Molenberghs, from the IAP research Network P6/03 of the Belgian Government (Belgian Science Policy).

References

Antoniak, C.E. (1974) “Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems,” *The Annals of Statistics*, 2, 1152–1174.

Ishwaran, H. and James, L.F. (2002) “Approximate Dirichlet process computing finite normal mixtures: smoothing and prior information,” *Journal of Computational and Graphical Statistics*, 11, 508–532.

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

Wicki, J., Perneger, T.V., Junod, A.F., Bounameaux, H., and Perrier, A. (2001) “Assessing clinical probability of pulmonary embolism in the emergency ward,” *Archives of Internal Medicine*, 161, 92–97.