0230

A Protective Estimator for Linear Regression with Nonignorably Missing Gaussian Outcomes

S.R. Lipsitz, G. Molenberghs, G.M. Fitzmaurice, J.G. Ibrahim

A Protective Estimator for Linear Regression with Nonignorably Missing Gaussian Outcomes

Stuart R. Lipsitz[1], Geert Molenberghs[2], Garrett M. Fitzmaurice[3], and Joseph G. Ibrahim[3,4*]

(1) Department of Biometry and Epidemiology,
Medical University of South Carolina

(2) Center for Statistics,
Limburgs Universitair Centrum, tUL, Belgium

(3) Department of Biostatistics, Harvard School of Public Health,
Boston MA 02115, U.S.A.

(4) Division of Biostatistical Science, Dana-Farber Cancer Institute
Boston MA 02115, U.S.A.

Summary

We propose a method for estimating the regression parameters in a linear regression model for Gaussian data when the outcome variable is missing for some subjects and missingness is thought to be nonignorable. Throughout, we assume that missingness is restricted to the outcome variable and that the covariates are fully observed. Although maximum likelihood estimation of the regression parameters is possible once joint models for the outcome variable and the nonignorable missing data mechanism have been specified, these models are fundamentally non-identifiable unless unverifiable modeling assumptions are imposed. In this paper, rather than explicitly modeling the nonignorable missingness mechanism, we consider the use of a "protective" estimator of the regression parameters (Brown, 1990). To implement the proposed method, it is necessary to assume that the outcome variable and one of the covariates have an approximate bivariate normal distribution, conditional on the remaining covariates. In addition, it is assumed that the missing data mechanism is conditionally independent of this covariate, given the outcome variable and the remaining covariates; the latter is referred to as the "protective" assumption. A method of moments approach is used to obtain the protective estimator of the regression parameters; the jackknife (Quenouille, 1956) is used to estimate the variance. The method is illustrated using data on the persistence of maternal smoking from the Six Cities study of the health effects of air pollution (Ware, et. al., 1984).

Key words: EM-algorithm, method of moments, ordinary least squares.

[*] Address for correspondence: Geert Molenberghs, Center for Statistics, Limburgs Universitair Centrum, tUL, Universitaire Campus, B–3590 Diepenbeek, Belgium, E-mail: geert.molenberghs@luc.ac.be

1 Introduction

Linear regression, assuming Gaussian errors, is probably one of the most widely used statistical models for relating the mean of an outcome variable to covariates. A common problem in the application of linear regression is that the outcome variable is often missing for a subset of the subjects in the study. The problem of missing outcome data arises in a wide variety of fields of applications, from sample surveys to controlled clinical trials. For example, consider a subsample of data on the persistence of maternal smoking from the Six Cities study of the health effects of air pollution (Ware, et. al., 1984). In this data set the outcome variable of interest is a measure of the mother's smoking (in cigarettes per day) when her child is 10 years old. It is of interest to determine how changes in the mother's smoking behavior (from the previous year) are related to her child's wheeze status at age 9 (yes or no) and the city of residence (there are two participating cities here). Preliminary analyses have shown that the square root transformation of the maternal smoking variable is approximately normal. As a result, it is of interest to estimate the parameters in the linear regression of the outcome ('square root of maternal smoking when child is age 10') on the prior measure of maternal smoking ('square root of maternal smoking when child is age 9'), the child's wheeze status at age 9, and the city of residence. Of note, 208 (or 35%) of the 574 subjects have missing outcome data. Furthermore, with this amount of missing data, a 'complete case analysis', based only on the 366 subjects with no missing outcome data, could potentially yield quite biased estimates of the regression parameters if missingness is related to the outcome. The data for 30 randomly selected subjects are shown in Table 1.

When nonresponse in the outcome variable is unrelated to the value of the possibly unobserved outcome, the nonresponse is said to be ignorable (e.g., see Little 1982, Little and Rubin 1987). Commonly, however, there may be concern that nonresponse is related to the values of the possibly unobserved outcome variable. For example, in the Six Cities study, mothers who smoke more may be less likely to report their cigarette consumption. When nonresponse is related to the value of the possibly unobserved outcome, the nonresponse is said to be ignorable. When there is nonignorable nonresponse, serious biases in the estimates of the parameters may result if the missing data mechanism is not modeled. Little and Rubin (1987) and Rubin (1987) discuss this issue in detail and provide excellent examples illustrating this point.

A standard linear regression analysis, based on the complete cases, discards data on subjects for whom the outcome is missing. When the missing data mechanism is nonignorable, the complete case analysis could yield very biased estimates of the regression parameters (e.g., Vach and Blettner, 1991; Ibrahim and Lipsitz, 1996). To reduce or remove the potential bias a nonignorable missing data mechanism, relating the probability that the outcome is missing to both the outcome and covariates, could be incorporated into the model. However, caution must be exercised with models for nonignorable nonresponse since these models are fundamentally non-identifiable unless unverifiable modeling assumptions are imposed (e.g., Ibrahim and Lipsitz, 1996;Ibrahim, Lipsitz and Chen, 1999). This is due to the fact that there is a lack of information in the observed data to estimate specific parameters in the missing data model (e.g., Little and Rubin, 1987, p.239; and Baker and Laird, 1988).

Because models for nonignorable nonresponse are known to be heavily dependent on unverifiable modeling assumptions, it is desirable to consider alternative approaches. One alternative to explicitly modeling the missingness mechanism in a nonignorable model is to consider the use of a “protective” estimator of the regression parameters (Brown, 1990). In this paper we propose a protective estimator of the linear regression parameters. To implement the proposed method, it is necessary to assume that the outcome variable and one of the covariates have an approximate bivariate normal distribution, conditional on the remaining covariates. In addition, it is assumed that the missing data mechanism is conditionally independent of this covariate, given the outcome variable and the remaining covariates; the latter is referred to as the "protective" assumption. A method of moments approach is used to obtain the protective estimator of the regression parameters; the jackknife (Quenouille, 1956) is used to estimate the variance. The proposed method only requires the application of two separate ordinary least squares regressions. The remainder of this article is organized as follows. In Section 2, we develop notation and present the complete data model and the observed data likelihood. In Section 3, a protective estimator for the linear regression parameters is developed. In section 4, the results of a simulation study, comparing the protective estimator to the maximum likelihood (under a correctly specified nonignorable model) and the complete case estimator, are presented. Finally, in Section 5, the proposed method is illustrated using the data on the persistence of maternal smoking from the Six Cities study.

2 Notation and Maximum Likelihood

Consider a linear regression model with n independent subjects, $i = 1, \ldots, n$. Let Y_i denote the outcome variable for the ith subject and let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ denote a $p \times 1$ vector of covariates. The primary interest is in estimation of the vector of regression coefficients $\boldsymbol{\beta}' = [\beta_0, \boldsymbol{\beta}']$ for the linear regression model

$$\mu_i = E[Y_i|\mathbf{x}_i, \boldsymbol{\beta}] = \beta_0 + \mathbf{x}_i'\boldsymbol{\beta}. \tag{1}$$

Although the covariates are fully observed, Y_i is missing for a subset of the subjects. Furthermore, the missing data mechanism is thought to be nonignorable.

Note that maximum likelihood estimation of $\boldsymbol{\beta}$ (and σ^2) requires specification of the conditional distribution of y_i given $\mathbf{x}_i$; furthermore, it is assumed that,

$$f(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2}} e^{-(y_i-\mu_i)^2/2\sigma^2}, \tag{2}$$

where $\sigma^2 = Var[Y_i|\mathbf{x}_i]$ and $\mu_i = \mu_i(\boldsymbol{\beta})$ is given by (1). However, since Y_i can be missing, we also define the indicator random variable R_i, which equals 1 if Y_i is observed and 0 if Y_i is missing. With nonignorable missing data, Ibrahim and Lipsitz (1996) and Ibrahim, Lipsitz and Chen (1999) propose using the joint distribution $(y_i, r_i|\mathbf{x}_i)$ to estimate $\boldsymbol{\beta}$, i.e.,

$$f(r_i, y_i|\mathbf{x}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2) = f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) f(r_i|\mathbf{x}_i, y_i, \boldsymbol{\alpha}) \ , \tag{3}$$

where $\boldsymbol{\alpha}$ is the parameter vector of the 'missing data mechanism' $f(r_i|\mathbf{x}_i, y_i, \boldsymbol{\alpha})$. For example, a logistic regression model could be specified for the Bernoulli random variable R_i given $(\mathbf{x}_i, y_i)$,

$$f(r_i|\mathbf{x}_i, y_i, \boldsymbol{\alpha}) = \pi_i^{r_i}(1-\pi_i)^{(1-r_i)}, \tag{4}$$

where

$$\pi_i = \frac{\exp(\alpha_0 + \boldsymbol{\alpha}_1'\mathbf{x}_i + \alpha_2 y_i)}{1 + \exp(\alpha_0 + \boldsymbol{\alpha}_1'\mathbf{x}_i + \alpha_2 y_i)}. \tag{5}$$

Note that in (5), if $\alpha_2 = 0$, then $f(r_i|\mathbf{x}_i, y_i, \boldsymbol{\alpha})$ does not depend on y_i; this implies that the missing data are missing at random (Rubin, 1976) and the missing data mechanism is ignorable (provided $\boldsymbol{\beta}$, σ, and $\boldsymbol{\alpha}$ are variation independent). If $\alpha_2 \neq 0$, then the missing data mechanism depends on y_i and is nonignorable.

When Y_i is missing ($R_i = 0$), the observed data are simply $(r_i, \mathbf{x}_i)$; if Y_i is observed, then the observed data are $(r_i, y_i, \mathbf{x}_i)$. Recall that the main interest lies in making inferences about the parameter $\boldsymbol{\beta}$ from the density $f(y_i|\mathbf{x}_i, \boldsymbol{\beta})$. However, since Y_i can be nonignorably missing, we must also consider the random variable R_i when using the observed data to make inferences about $\boldsymbol{\beta}$ (e.g., Ibrahim, Lipsitz and Chen, 1999). In particular, for likelihood-based inference about $\boldsymbol{\beta}$, the density of the observed data is (3) if Y_i is observed and is

$$f(r_i|\mathbf{x}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2) = \int_{y_i} f(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) f(r_i|\mathbf{x}_i, y_i, \boldsymbol{\alpha}) dy_i$$

if Y_i is missing. As a result, the observed data log-likelihood is

$$\sum_{i=1}^{n} r_i \log[f(r_i, y_i|\mathbf{x}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2)] + (1 - r_i) \log[f(r_i|\mathbf{x}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2)],$$

and the MLE for $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2)$ is obtained by directly maximizing the log-likelihood of observed data by solving

$$\frac{\partial}{\partial(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma^2)} \sum_{i=1}^{n} \left\{ r_i \log[f(r_i, y_i|\mathbf{x}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2)] + (1 - r_i) \log[f(r_i|\mathbf{x}_i, \boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma^2)] \right\} = 0,$$

for $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \hat{\sigma}^2)$ using a Newton-Raphson algorithm. Alternatively, the EM-algorithm (Dempster et al., 1977; Ibrahim, Chen, and Lipsitz, 1999) can be used to obtain the MLE. Unfortunately, as discussed in the introduction, caution must be exercised with models for nonignorable missing data because certain parameters may be inestimable unless unverifiable modeling assumptions are imposed. As an alternative to explicitly modeling the missingness mechanism in a nonignorable model, in Section 3 we propose a protective estimator (Brown, 1990).

3 Protective Estimator

To develop the protective estimator we must assume that one of the covariates, say x_{i1}, has a normal distribution. In particular, we partition $\mathbf{x}_i$ into $\mathbf{x}_i' = [x_{i1}, \mathbf{x}_{i2}']$, and assume that $f(y_i, x_{i1}|\mathbf{x}_{i2})$ has a bivariate normal distribution. Next, consider the distribution of (y_i, x_{i1}) given $\mathbf{x}_{i2}$ when no data are missing. The density $f(y_i, x_{i1}|\mathbf{x}_{i2})$ is given by,

$$\left(\begin{array}{c} Y_i \\ X_{i1} \end{array} \middle| \mathbf{x}_{i2} \right) \sim N \left[\left(\begin{array}{c} \theta_0 + \boldsymbol{\theta}_1 \mathbf{x}_{i2} \\ \gamma_0 + \boldsymbol{\gamma}_1 \mathbf{x}_{i2} \end{array} \right), \left(\begin{array}{cc} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{array} \right) \right] . \tag{6}$$

Then, in terms of the parameters in (6), the regression model of interest in (1) is given by,

$$\begin{aligned} E[Y_i|\mathbf{x}_i] &= \theta_0 + \boldsymbol{\theta}_1 \mathbf{x}_{i2} + \tfrac{\sigma_{12}}{\sigma_{22}^2}[x_{i1} - \gamma_0 - \boldsymbol{\gamma}_1 \mathbf{x}_{i2}] \\ &= \left(\theta_0 - \tfrac{\sigma_{12}}{\sigma_{22}^2}\gamma_0\right) + \tfrac{\sigma_{12}}{\sigma_{22}^2} x_{i1} + \left(\boldsymbol{\theta}_1 - \tfrac{\sigma_{12}}{\sigma_{22}^2}\boldsymbol{\gamma}_1\right) \mathbf{x}_{i2} \qquad (7) \\ &= \beta_0 + \beta_1 x_{i1} + \boldsymbol{\beta}_2 \mathbf{x}_{i2}, \end{aligned}$$

where

$$\beta_0 = \theta_0 - \frac{\sigma_{12}}{\sigma_{22}^2}\gamma_0,$$

$$\beta_1 = \frac{\sigma_{12}}{\sigma_{22}^2},$$

and

$$\boldsymbol{\beta}_2 = \boldsymbol{\theta}_1 - \frac{\sigma_{12}}{\sigma_{22}^2}\boldsymbol{\gamma}_1 \ .$$

Further, the conditional variance is

$$Var[Y_i|\mathbf{x}_i] = \sigma_{11}^2 - \frac{\sigma_{12}}{\sigma_{22}^2} \ . \tag{8}$$

In the presence of nonignorable missing outcome data, if the parameters

$$(\theta_0, \boldsymbol{\theta}_1, \gamma_0, \boldsymbol{\gamma}_1, \sigma_{11}^2, \sigma_{12}, \sigma_{22}^2)$$

in (6) can be consistently estimated, they can be substituted in (7) to consistently estimate the regression parameters of interest. The protective estimator of $\boldsymbol{\beta}$ uses the conditional distributions of $f(x_{i1}|\mathbf{x}_{i2})$ and $f(x_{i1}|y_i, \mathbf{x}_{i2})$ to estimate these parameters. Since x_{i1} and $\mathbf{x}_{i2}$ are both fully observed, it is straightforward to estimate $f(x_{i1}|\mathbf{x}_{i2})$ using all observations. However, since y_i is observed only when $r_i = 1$, it is not straightforward to estimate $f(x_{i1}|y_i, \mathbf{x}_{i2})$ unless an additional assumption about the nonignorable missing data mechanism is made.

¿From an examination of (6), note that the conditional mean of X_{i1} given $\mathbf{x}_{i2}$ is

$$E(X_{i1}|\mathbf{x}_{i2}, \boldsymbol{\gamma}) = \gamma_0 + \boldsymbol{\gamma}_1 \mathbf{x}_{i2}, \tag{9}$$

with conditional variance

$$Var(X_{i1}|\mathbf{x}_{i2}) = \sigma_{22}^2. \tag{10}$$

Since there are no missing data on X_{i1} or $\mathbf{x}_{i2}$, $(\gamma_0, \boldsymbol{\gamma}_1, \sigma_{22}^2)$ can be consistently estimated using ordinary least squares, where the outcome variable is X_{i1} and the regression model is given by

(9). Suppose we denote the ordinary least squares estimate of these parameters by $(\widehat{\gamma}_0, \widehat{\boldsymbol{\gamma}}_1, \widehat{\sigma}^2_{22})$. Estimation of the remaining parameters, $(\theta_0, \boldsymbol{\theta}_1, \sigma^2_{11}, , \sigma_{12})$, can be based on the conditional distribution $f(x_{i1}|y_i, \mathbf{x}_{i2})$.

However, without additional assumptions, it is possible to estimate relationships between Y_i and other variables only when Y_i is observed $(R_i = 1)$. Consider the density

$$f(x_{i1}|y_i, \mathbf{x}_{i2}, R_i = 1) = \frac{pr(R_i = 1|x_{i1}, y_i, \mathbf{x}_{i2}) f(x_{i1}|y_i, \mathbf{x}_{i2})}{pr(R_i = 1|y_i, \mathbf{x}_{i2})} \,. \tag{11}$$

If, given $(y_i, \mathbf{x}_{i2})$, the missing data mechanism does not depend on x_{i1}, i.e.,

$$pr(R_i = 1|x_{i1}, y_i, \mathbf{x}_{i2}) = pr(R_i = 1|y_i, \mathbf{x}_{i2}), \tag{12}$$

then (11) reduces to

$$f(x_{i1}|y_i, \mathbf{x}_{i2}, R_i = 1) = f(x_{i1}|y_i, \mathbf{x}_{i2}). \tag{13}$$

Equation (12) is the protective assumption. For appropriate choices of x_{i1}, this assumption is often quite reasonable, since for many nonignorable missing data mechanisms, missingness depends primarily on the unobserved value of the outcome Y_i. In particular, the protective assumption asserts that, conditional on Y_i (and $\mathbf{x}_{i2}$), missingness is independent of x_{i1}. At the very least, the assumption in (12) can form the basis of a sensitivity analysis, that assesses the sensitivity of inferences to departures from the assumption that missingness is ignorable. Under the protective assumption, the result in (13) implies that the complete cases $(R_i = 1)$ can be used to consistently estimate the parameters of the conditional distribution of X_{i1} given $(y_i, \mathbf{x}_{i2})$. In particular,

$$E(X_{i1}|\mathbf{x}_{i2}, y_i, R_i = 1, \boldsymbol{\gamma}, \boldsymbol{\theta}) = E(X_{i1}|\mathbf{x}_{i2}, y_i, \boldsymbol{\gamma}, \boldsymbol{\theta}).$$

Using (6), the conditional mean of X_{i1} given $(\mathbf{x}_{i2}, y_i)$, is

$$\begin{aligned} E(X_{i1}|\mathbf{x}_{i2}, y_i, \boldsymbol{\gamma}, \boldsymbol{\theta}) &= (\gamma_0 + \boldsymbol{\gamma}_1 \mathbf{x}_{i2}) + \tfrac{\sigma_{12}}{\sigma^2_{11}}[y_i - \theta_0 - \boldsymbol{\theta}_1 \mathbf{x}_{i2}] \\ &= \phi_0 + \boldsymbol{\phi}_1 \mathbf{x}_{i2} + \phi_2 y_i, \end{aligned} \tag{14}$$

where

$$\phi_2 = \frac{\sigma_{12}}{\sigma_{11}^2},$$

$$\phi_0 = \gamma_0 - \frac{\sigma_{12}}{\sigma_{11}^2}\theta_0 = \gamma_0 - \phi_2\theta_0,$$

and

$$\boldsymbol{\phi}_1 = \boldsymbol{\gamma}_1 - \frac{\sigma_{12}}{\sigma_{11}^2}\boldsymbol{\theta}_1 = \boldsymbol{\gamma}_1 - \phi_2\boldsymbol{\theta}_1 \ .$$

Also, the conditional variance is given by

$$Var(X_{i1}|\mathbf{x}_{i2}, y_i) = \sigma_{22}^2 - \frac{\sigma_{12}^2}{\sigma_{11}^2}. \qquad (15)$$

Then, the parameters $[\phi_0, \boldsymbol{\phi}_1, \phi_2, Var(X_{i1}|\mathbf{x}_{i2}, y_i)]$ can be estimated, based on the complete cases, via ordinary least squares regression with outcome variable X_{i1} and covariates $(\mathbf{x}_{i2}, y_i)$. Given the ordinary least squares estimates $[\widehat{\phi}_0, \widehat{\boldsymbol{\phi}}_1, \widehat{\phi}_2, \widehat{\mathrm{Var}}(X_{i1}|\mathbf{x}_{i2}, y_i)]$ from the latter regression model, and the estimate $(\widehat{\gamma}_0, \widehat{\boldsymbol{\gamma}}_1, \widehat{\sigma}_{22}^2)$ from the regression model in (9), $(\theta_0, \boldsymbol{\theta}_1, \sigma_{11}^2, \sigma_{12})$ can be estimated as follows,

$$\widehat{\theta}_0 = (\widehat{\gamma}_0 - \widehat{\phi}_0)/\widehat{\phi}_2,$$

and

$$\widehat{\boldsymbol{\theta}}_1 = (\widehat{\boldsymbol{\gamma}}_1 - \widehat{\boldsymbol{\phi}}_1)/\widehat{\phi}_2.$$

¿From an examination of the residual variance in (15), note that

$$\frac{\sigma_{12}^2}{\sigma_{11}^2} = \sigma_{22}^2 - Var(X_{i1}|\mathbf{x}_{i2}, y_i),$$

so that

$$\sigma_{12} = \frac{\sigma_{12}^2/\sigma_{11}^2}{\sigma_{12}/\sigma_{11}^2} = \frac{\sigma_{12}^2/\sigma_{11}^2}{\phi_2} = \frac{\sigma_{22}^2 - Var(X_{i1}|\mathbf{x}_{i2}, y_i)}{\phi_2}.$$

Then, σ_{12} can be estimated using

$$\widehat{\sigma}_{12} = \frac{\widehat{\sigma}_{22}^2 - \widehat{\mathrm{Var}}(X_{i1}|\mathbf{x}_{i2}, y_i)}{\widehat{\phi}_2}$$

and $\sigma_{11}^2 = \sigma_{12}/\phi_2$, σ_{11}^2 can be estimated using

$$\widehat{\sigma}_{11}^2 = \frac{\widehat{\sigma}_{12}}{\widehat{\phi}_2} = \frac{\widehat{\sigma}_{22}^2 - \widehat{\mathrm{Var}}(X_{i1}|\mathbf{x}_{i2}, y_i)}{\widehat{\phi}_1^2}.$$

Then, the protective estimator of $\boldsymbol{\beta}' = [\beta_0, \beta_1, \boldsymbol{\beta}_2']$ in (7) is given by

$$\widehat{\beta}_0 = \widehat{\theta}_0 - \frac{\widehat{\sigma}_{12}}{\widehat{\sigma}_{22}^2}\widehat{\gamma}_0,$$

$$\widehat{\beta}_1 = \frac{\widehat{\sigma}_{12}}{\widehat{\sigma}_{22}^2},$$

and

$$\widehat{\boldsymbol{\beta}}_2 = \widehat{\boldsymbol{\theta}}_1 - \frac{\widehat{\sigma}_{12}}{\widehat{\sigma}_{22}^2}\widehat{\boldsymbol{\gamma}}_1 \ .$$

When the assumptions about the missing data mechanism and the distribution of $f(y_i, x_{i1}|\mathbf{x}_{i2})$ are correct, results from method of moments can be used to show that $\widehat{\boldsymbol{\beta}}$ is consistent and has an asymptotic multivariate normal distribution, with mean vector $\boldsymbol{\beta}$ and a covariance matrix which can be consistently estimated using the delta method or the jackknife (Quenouille, 1956). Because $\boldsymbol{\beta}$ is a complicated function of $(\theta_0, \boldsymbol{\theta}_1, \gamma_0, \boldsymbol{\gamma}_1, \sigma_{11}^2, \sigma_{12}, \sigma_{22}^2)$, and ordinary least squares regression is not computationally demanding, it is preferable to use the jackknife variance estimator. The jackknife variance estimate can be obtained as follows,

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) = \left(\frac{n-1}{n}\right)\sum_{i=1}^{n}(\widehat{\boldsymbol{\beta}}_{-i} - \widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_{-i} - \widehat{\boldsymbol{\beta}})', \tag{16}$$

where $\widehat{\boldsymbol{\beta}}_{-i}$ is the estimate of $\boldsymbol{\beta}$ obtained by deleting the data for the i^{th} subject. Because OLS regression is computationally simple, the additional CPU time required to obtain the jackknife variance estimate is minimal. For example, in the illustrative example presented in Section 5, where $n = 574$, it took only 11 seconds (real time) on a SPARC Ultra-80 workstation to obtain the jackknife variance estimate.

4 Simulation Study

We performed a modest simulation study to compare the estimates obtained using the complete cases (CC), maximum likelihood with a correctly-specified nonignorable missing data model (ML-NI), and the proposed protective estimate (PR). In the simulation study, there were two covariates, (x_{i1}, x_{i2}) and the true model for the simulations was formulated by specifying each term on the right side of

$$f(r_i, y_i, x_{i1}, x_{i2}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = f(r_i|y_i, x_{i1}, x_{i2}, \boldsymbol{\alpha}) f(y_i|x_{i1}, x_{i2}, \boldsymbol{\beta}) f(x_{i1}|x_{i2}, \boldsymbol{\gamma}). \tag{17}$$

For the covariate distributions, x_{i2} was fixed to be a binary variable with half of the observations set equal to 0 and the other half set equal to 1. The distribution of X_{i1} given x_{i2} was assumed to be normal, with mean x_{i2} and variance 1. The distribution of Y_i given (x_{i1}, x_{i2}) was assumed to be normal, with mean

$$\mu_i = E(Y_i|x_{i1}, x_{i2}) = 1 + x_{i1} + x_{i2}, \tag{18}$$

and variance 1, so that $\beta' = (\beta_0, \beta_1, \beta_2) = (1, 1, 1)$. The true model for $\pi_i = pr(R_i = 1|y_i, x_{i1}, x_{i2})$ was

$$\text{logit}(\pi_i) = -y_i. \tag{19}$$

For each of $n = 125$, $n = 250$ and $n = 500$, we performed 1000 simulation replications. The results of the simulations are summarized in Table 2. We note that for ML-NI (maximum likelihood), (19) is correctly specified. For the protective estimator, $f(y_i, x_{i1}|x_{i2}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is assumed to be bivariate normal.

The results in Table 2 indicate that the protective estimator is approximately unbiased, and displays the least bias for all sample sizes considered. Note that the CC estimator is biased for all parameters and for all sample sizes; the bias does not appear to depend on sample size.

The bias in the ML-NI estimator is modest and decreases as n gets larger; the magnitude of the bias is within 10% for all sample sizes considered. Of note, the ML-NI estimator had the smallest variance (and, although not shown, the smallest mean square error) for all sample sizes. For example, when $n = 500$, the efficiency of the protective estimator versus ML-NI is 52% for β_0, 29% for β_1, and 20% for β_2. However, the main motivation for the protective estimator is in its use in sensitivity analysis for assessing the impact of departures from the assumption that missingness is ignorable; the issue of efficiency is of less concern.

5 Illustrative Example

In this section, we present an illustration of the use of the protective estimator using data on the persistence of maternal smoking from the Six Cities study. Recall that the outcome variable is a measure of maternal smoking when her child is 10 years old. Specifically, it is assumed that

$$\text{Smoke}_{i2} = \sqrt{\text{maternal cigarettes smoked per day when child is age 10}}$$

has an approximate normal distribution and it is of interest to estimate the parameters in the following linear regression model,

$$E(\text{Smoke}_{i2}|\text{city}_i, \text{wheeze}_i, \text{Smoke}_{i1}) = \beta_0 + \beta_1\text{Smoke}_i + \beta_2\text{city}_i + \beta_2\text{wheeze}_i \ , \tag{20}$$

where

$$\text{Smoke}_{i1} = \sqrt{\text{maternal cigarettes smoked per day when child is age 9}} \ ;$$

$$\text{wheeze}_i = \begin{cases} 1 & \text{if child wheezed at age 9} \\ 0 & \text{if child did not wheeze at age 9} \end{cases} \ ,$$

and city_i equals 0 or 1 (for the two participating cities).

Using maximum likelihood, we considered a nonignorable missing data model, with

$$\log(\pi_i/(1-\pi_i)) = \alpha_0 + \alpha_1\text{Smoke}_{i2} + \alpha_2\text{Smoke}_{i1} + \alpha_3\text{wheeze}_i + \alpha_3\text{city}_i, \tag{21}$$

where $\pi_i = pr(R_i = 1|\text{Smoke}_{i2}, \text{Smoke}_{i1}, \text{wheeze}_i, \text{city}_i, \alpha)$. Table 3 gives the ML estimate of $\boldsymbol{\alpha}$ and there is evidence that the outcome variable is nonignorably missing. Specifically, at the 5% level of significance, missingness in the outcome appears to be significantly related to the outcome Smoke_{i2} ($p < .0001$), the city of residence ($p < .0001$), and marginally related to wheeze ($p \approx 0.096$). Conditional on the outcome Smoke_{i2}, missingness does not appear to be related to maternal smoking at the previous visit (Smoke_{i1}) ($p \approx 0.758$).

For the protective estimator it is assumed that $f(\text{Smoke}_{i2}, \text{Smoke}_{i1}|\text{wheeze}_i, \text{city}_i,)$ is bivariate normal. Since the probability of missingness in (21) does not appear to be related to Smoke_{i1}, and the marginal distributions of Smoke_{i2} and Smoke_{i1} each appear to be approximately normal in preliminary analyses, the two assumptions required for the protective estimator to be consistent appear to hold. We note, however, that because of the inherent difficulties in fitting nonignorable models, ordinarily it will not be possible to determine whether the nonresponse probabilities are conditionally independent of x_{i1}; in general, the assumption must often be made on subject-matter grounds.

Table 4 provides the estimates of β obtained from the three different approaches; ML-NI (nonignorable), PR (protective), and CC (complete case). In Table 4, all three approaches yield very similar estimates of the intercept, and the effects of Smoke_{i1} and wheeze (although, for the latter, the CC estimate is negative, whereas the other two estimates are positive). However, we note that the estimate of the city effect is similar for ML-NI and the protective estimate, but discernibly different from the CC estimate. Also, complementing the results from the simulations, the ML estimator appears to be the most efficient, and most notably so for estimation of the city effect, in which the protective estimate is estimated to be only 16% as efficient as the ML-NI estimate. Thus, even though the ML-NI and the PR estimates of the city effect are quite similar, the larger sampling variability of the PR estimate yields a non-significant $p-$value.

Finally, we note that the ML-NI estimate was obtained by programming a Newton-Raphson algorithm in SAS, using numerical integration with a trapezoid rule. The convergence criterion used for the Newton-Raphson algorithm was that the distance between the t^{th} iteration and the $(t+1)^{th}$ iteration in each parameter was less than 10^{-7}. The number of iterations required for convergence was 6, and it took 82 seconds (real time) to obtain the estimates using a SPARC-80 Workstation. This does not compare favorably to the 11 seconds required for the protective estimate and jackknife variance estimate.

6 Conclusion

In this paper we have proposed a protective estimator of the parameters in a linear regression model with nonignorably missing Gaussian outcomes. We note that maximum likelihood with a nonignorable missing data model has been proposed previously to estimate the regression parameters, but, as discussed in Ibrahim and Lipsitz (1996), caution must be exercised since these models are fundamentally non-identifiable unless unverifiable modeling assumptions are imposed. The use of a protective estimator avoids the need to jointly specify and estimate the parameters of a nonignorable missing data model. However, the protective estimator is not free of assumptions. In particular, the protective estimator assumes that the outcome variable and one of the covariates have an approximate bivariate normal distribution, conditional on the remaining covariates. In addition, it is assumed that the missing data mechanism is conditionally independent of this covariate, given the outcome variable and the remaining covariates; the latter is referred to as the "protective" assumption. The property of the bivariate normal distribution that is crucial to the protective estimator is that the conditional mean of the covariate can be expressed as a linear function of the outcome variable (and the remaining covariates). In cases where the bivariate normal assumption may not hold, but the conditional mean is nonetheless approximately linear in the outcome variable, the protective estimator should yield valid estimates of the regression parameters (provided the additional conditional independence assumption holds). In general, we view the protective estimator as providing a simple means for conducting sensitivity analysis, for assessing the sensitivity of inferences to departures from the assumption that missingness is ignorable.

Because of the broad range of possible missing data configurations and underlying probability distributions generating the data, it is very difficult to draw definitive conclusions from the modest simulation study that was conducted; we can only make some general suggestions. The results of the simulation study suggest that the bias in estimating $\boldsymbol{\beta}$ can be reduced when using the protective estimator as compared to the CC estimator. However, there is a tradeoff between using the protective estimator and the ML-NI estimator. For the protective estimator to be consistent, it is necessary to correctly specify $f(y_i, x_{i1}|\mathbf{x}_{i2})$, but the missing data mechanism does not need to be specified; the only requirement is that missingness is conditionally independent of x_{i1}. For the ML-NI estimator to be consistent, it is necessary to correctly specify $f(y_i, r_i|\mathbf{x}_i)$. The results of the simulation study indicate that the ML-NI estimator can be more efficient than the protective estimator. However, as noted earlier, a basic problem with nonignorable missing data models is that they are fundamentally non-identifiable unless unverifiable modeling assumptions are made. Moreover, it is well-known that the resulting estimates are quite sensitive to modeling assumptions. For example, from the data at hand, it is simply not possible to distinguish between nonignorably missing outcome data arising from a normal distribution and ignorably missing outcome data arising from a distribution with positive or negative skewness.

Acknowledgments

We are grateful for the support provided by grants HL69800, AHRQ 10871, HL52329, HL61769, GM 29745, MH 54693, and CA 70101 from the US National Institutes of Health. The second author is grateful for support from FWO-Vlaanderen Research Project "Sensitivity Analysis for Incomplete and Coarse Data," and Belgian IUAP/PAI network "Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data".

References

Baker, S.G. and Laird, N. M. (1988). Regression Analysis for Categorical Variables With Outcome Subject to Nonignorable Nonresponse. *Journal of the American Statistical Association*, **83**, 62-69.

Brown, C.H. (1990). Protecting against nonrandomly missing data in longitudinal studies. *Biometrics* **46**, 143-55.

Dempster, A.P., Laird, N.M., & Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B*, **39**, 1–38.

Ibrahim, J.G. (1990). Incomplete Data in Generalized Linear Models. *Journal of the American Statistical Association*, **85**, 765-769.

Ibrahim, J.G., Chen, M.H., and Lipsitz S.R., (1999). Monte Carlo EM for Missing Covariates in Parametric Regression Models, *Biometrics*, **99**, 104-111.

Ibrahim, J. G., and Lipsitz, S. R. (1996), Parameter Estimation From Incomplete Data in Binomial Regression When the Missing Data Mechanism is Nonignorable, *Biometrics*, 1071-1078.

Ibrahim J, Lipsitz SR, and Chen M. (1999) Missing Covariates in Generalized Linear Models When The Missing Data Mechanism is Nonignorable. *Journal of the Royal Statistical Society, Series B*, **61**, 173–190.

Little, R.J.A. (1982). Models for Nonresponse in Sample Surveys. *Journal of the American Statistical Association*, **77**, 237-250.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical analysis with missing data.* New York: John Wiley.

Quenouille, M. H., (1956), Notes on bias in estimation, *Biometrika* **43**, 353-60.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581-592.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley.

Vach, W. and Blettner, M. (1991). Biased Estimation of the Odds Ratio in Case-control Studies Due to the Use of Ad Hoc Methods of Correcting for Missing Values for Confounding Variables, *American Journal of Epidemiology*, **134**, 895–907.

Ware, J.H., Dockery, D.W., Spiro, A. III, Speizer, F.E. and Ferris, B.G. Jr. (1984). Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases*, **129**, 366-374.

Table 1. Data on 30 randomly selected subjects from the Six Cities Study.

Subject	City	Child wheeze at age 9	Maternal Smoking: cigarettes at age 9	Maternal Smoking: cigarettes at age 10
1	1	NO	0.0000	.
2	1	NO	1.0000	.
3	1	YES	10.0000	.
4	0	NO	13.0000	.
5	0	YES	15.0000	.
6	0	NO	20.0000	.
7	1	YES	20.0000	.
8	0	NO	30.0000	.
9	0	YES	35.0000	.
10	0	YES	39.9999	.
11	1	NO	39.9999	.
12	0	NO	0.0000	0.0000
13	0	YES	0.0000	0.0000
14	1	NO	0.0000	0.0000
15	1	YES	0.0000	0.0000
16	0	NO	12.0000	0.0000
17	0	NO	4.0000	3.0000
18	0	NO	10.0000	6.0000
19	0	NO	1.0000	7.0000
20	1	NO	10.0000	10.0000
21	0	YES	20.0000	10.0000
22	1	NO	20.0000	10.0000
23	0	NO	20.0000	20.0000
24	0	YES	20.0000	20.0000
25	0	NO	30.0000	20.0000
26	1	NO	20.0000	21.0000
27	1	NO	35.0000	25.0000
28	0	YES	39.9999	30.0000
29	1	YES	39.9999	30.0000
30	1	YES	2.0000	39.9999

Table 2. Summary of results from the simulation study.

	n	Method	$\beta_0 = 1$	$\beta_1 = 1$	$\beta_2 = 1$
Estimate	125	CC	0.426	0.868	0.848
		ML	0.918	1.035	0.971
		PR	1.036	1.030	1.027
	250	CC	0.419	0.862	0.845
		ML	0.923	1.005	0.969
		PR	1.021	1.004	1.019
	500	CC	0.421	0.857	0.846
		ML	0.953	0.985	0.974
		PR	1.011	1.002	1.010
Simulation Variance	125	CC	0.0585	0.2719	0.0487
		ML	0.0798	0.1887	0.0320
		PR	0.3723	0.8775	0.3023
	250	CC	0.0298	0.1227	0.0214
		ML	0.0061	0.0992	0.0038
		PR	0.1125	0.3258	0.0863
	500	CC	0.0132	0.0546	0.0101
		ML	0.0189	0.0397	0.0065
		PR	0.0364	0.1391	0.0318

Table 3. Maximum likelihood estimates for logistic regression for the missing data model, $pr(R_i = 1|y_i, \mathbf{x}_i)$.

PARAMETER	ESTIMATE	SE	Z	$p-$value
Intercept	2.359	0.260	9.08	$< .0001$
Smoke$_{i2}$	-0.405	0.075	-5.38	$< .0001$
Smoke$_{i1}$	-0.016	0.051	-0.31	0.758
Wheeze	-0.382	0.229	-1.67	0.096
City	-1.521	0.210	-7.26	$< .0001$

Table 4. Estimates for linear regression model for $E(Y_i|\mathbf{x}_i)$.

Effect	Method	$\hat{\beta}$	SE	Z	$p-$value
Intercept	CC	0.892	0.162	5.51	$< .0001$
	PR	1.045	0.286	3.65	$< .0001$
	ML	1.169	0.156	7.50	$< .0001$
Smoke$_{i1}$	CC	0.294	0.047	6.30	$< .0001$
	PR	0.416	0.086	4.83	$< .0001$
	ML	0.301	0.042	7.12	$< .0001$
Wheeze	CC	-0.045	0.245	-0.19	0.853
	PR	0.083	0.556	0.15	0.881
	ML	0.075	0.222	0.34	0.736
City	CC	-0.025	0.221	-0.11	0.910
	PR	0.424	0.495	0.86	0.392
	ML	0.460	0.199	2.31	0.021