Likelihood methods for missing covariate data in highly stratified studies
Paul J. Rathouz
University of Chicago, Chicago, IL, USA

Summary. This paper considers canonical link generalized linear models with stratum-specific nuisance intercepts and missing covariate data. This family includes the conditional logistic regression model. Existing methods for this problem, each of which uses a conditioning argument to eliminate the nuisance intercept, model either the missing covariate data or the missingness process. This paper compares these methods under a common likelihood framework. The semiparametric efficient estimator is identified, and a new estimator, which reduces dependence on the model for the missing covariate, is proposed. A simulation study compares the methods with respect to efficiency and robustness to model misspecification.

Keywords: Conditional likelihood; conditional logistic regression; fixed effects; matched case-control; missing data; nuisance parameter; semiparametric efficiency.

Address for correspondence: Paul J. Rathouz, Department of Health Studies, 5841 South Maryland Avenue, MC 2007, Chicago, IL 60637.
E-mail: prathouz@health.bsd.uchicago.edu

1 Introduction

We consider independent data \((Y_i, X_i, Z_i)\) for records \(i = 1, \ldots, n\), where interest lies in the conditional distribution of \((Y_i|X_i, Z_i)\), and where part or all components of covariate \(X_i\) may be missing on some records. Here, \(Y_i\) is a univariate response,
while $X_i$ and $Z_i$ are possibly multivariate covariate vectors. When all records are complete, such data are often analyzed via a generalized linear model with canonical link function (McCullagh and Nelder, 1989),

$$f(Y_i|X_i,Z_i; \beta, \phi) = \exp \left\{ \frac{Y_i \eta_i - b(\eta_i)}{a(\phi)} + c^*(Y_i, \phi) \right\},$$  \hspace{1cm} (1)

where $\eta_i$ is a linear function of covariates $(X_i, Z_i)$, $\phi$ is a scale parameter, and $a(\cdot)$, $b(\cdot)$ and $c^*(\cdot)$ are known functions. Examples include logistic, Poisson and linear regression. When the data are stratified, clustered, or longitudinal, and the $n$ observations belong to one among many strata $s = 1, \ldots, J$, the linear predictor $\eta_i$ is often assumed to be of the form

$$\eta_i = q_s + \beta^T_x Z_i + \beta^T_x X_i,$$  \hspace{1cm} (2)

in what is sometimes referred to as a fixed effects model (Greene, 2000). Stratum effects are accounted for by the stratum specific intercept $q_s$, which is considered a nuisance parameter. Because for fixed $\beta = (\beta^T_x, \beta^T_x)^T$ and $\phi$, $\sum_i Y_i$ within stratum $s$ is a sufficient statistic for the nuisance $q_s$, conditioning on $\sum_i Y_i$ eliminates $q_s$ in the likelihood resulting from (1) (Godambe, 1976; Diggle, et al., Ch.9). When $Y_i$ is a binary disease variable, (1) and (2) comprise the model underlying the conditional logistic regression (CLR) method for matched case control studies (Breslow and Day, 1980, p.248). Each matched set is its own stratum, and $\sum_i Y_i$ is the number of cases in a matched set. Conditioning on $\sum_i Y_i$ not only eliminates $q_s$, but also reflects the case-control sampling strategy.

The problem of making inferences on $(\beta, \phi)$ in this model when $X_i$ is missing for some records has been addressed in recent papers by Satten and Kupper (1993), Lipsitz et al. (1998), Satten and Carroll (2000), Paik and Sacco (2000), and Rathouz et al. (2002). While all of those authors develop methods for the CLR model for binary $Y_i$, their methods would apply equally well to other canonical-link generalized
linear models of form (1) and (2). These methods fall under two general approaches. The first of these involves modelling the distribution of the missing covariates $X_i$. Paik and Sacco, Satten and Kupper, and Satten and Carroll each propose conditional likelihoods which rely on a model for the distribution of $X_i$ among the control subjects. Satten and Carroll model $(X_i|Y_i = 0, Z_i)$ non-parametrically for $(X_i, Z_i)$ with finite support, and in a similar approach, Satten and Kupper exploit a surrogate for $X_i$. By contrast, Paik and Sacco assume that $(X_i|Y_i, Z_i)$ is univariate with a distribution belonging to an exponential family model.

When a model for the distribution of $X_i$ given $Z_i$ is difficult to specify, perhaps because either $X_i$ or $Z_i$ is of high dimension, then an alternative is to model the process giving rise to missing data. Lipsitz et al. (1998) propose such an approach, modelling the case-control status only among the subjects with observed $X_i$, conditioning on whether or not each subject has complete data. Rathouz et al. (2002) extend this likelihood approach to a class of estimators which substantially increase efficiency in estimating $\beta_Z$, but without much improvement for $\beta_X$. These approaches require at most weak knowledge of the distribution of $X_i$.

Using the stratum-level likelihood for the observed data as a unifying framework, we will distinguish among these methods and provide some guidance as to which one the user should select for data analysis. In the following section, we present the conditional likelihood estimator in its classic form and in an alternative form that applies when it is possible to model the distribution of $X_i$. In Section 3, we compare methods that have been previously proposed when $X_i$ may be missing. We show in Section 3.1 that, when it is possible to model the distribution of $X_i$, the likelihood of Satten and Kupper yields the semiparametric efficient estimator in this problem. Section 3.2 introduces a new suboptimal likelihood and estimator related to those proposed by Paik and Sacco. These methods use the data on all records, whether $X_i$
is observed or not. When data analysis is limited to records with observed \(X_i\), either by choice or by data constraints, we show in Section 3.3 that the likelihood due to Lipsitz \textit{et al.} yields the semiparametric efficient estimator for this problem. Lipsitz \textit{et al.}'s approach requires knowledge of the probability of \(X_i\) being missing for each record. Section 4 contains a simulation study comparing the estimators in Section 3 with one another with respect to efficiency and robustness to misspecification of the model for \(X_i\) and the missingness model. In Section 5, we present conclusions in the form of recommendations to the user of these methods.

In developing our results, we recast the conditional logistic regression model in the more general canonical-link generalized linear model family (1) and (2). Our results therefore apply to highly stratified problems where the response data are other than binary. Throughout, following the authors mentioned above, we assume that \(X_i\) is missing at random (MAR; Little and Rubin, 1987), although we examine robustness to this assumption in our simulation work.

2 Semiparametric efficiency of the conditional likelihood estimator

Consider the data from a given stratum \(s\) for the setting in which there are no missing data; that is, suppose we only have covariates \(Z_i\) in (2). To denote the data on stratum \(s\), write \(Z = (Z_1, \ldots, Z_i, \ldots, Z_n)^T\) for the matrix of \(n\) row vectors of covariates \(Z_i\); similarly define the vector of responses \(Y\). Throughout, we consider inferences conditional on \(Z\). From (1) and (2), and letting \(f(\cdot)\) denote the density or probability mass function (pmf) of either \(Y_i\) or \(Y\), the likelihood from stratum \(s\) is \(f(Y|Z) = L^*(\beta, \phi, q_s) = L^*,\) where

\[
\log\{L^*(\beta, \phi, q_s)\} = \left\{ q_s \sum_i Y_i + \beta^T Z \sum_i Z_i Y_i - \sum_i b(\eta_i) \right\} / a(\phi) + \sum_i c^*(Y_i, \phi),
\]

(3)
and where the sums are over $i$ within $s$. Now note that for fixed $\beta_z$ and $\phi$, $\sum_i Y_i$ is a sufficient statistic for the nuisance parameter $q_s$. Therefore, the conditional likelihood for $\xi^* = (\beta_z, \phi)$, $L^*_c(\xi^*) = f(Y|\sum_i Y_i, Z)$, is free of $q_s$.

Now, considering data across strata $s = 1, \ldots, J$, let $\hat{\xi}^*$ denote the estimator for $\xi^*$ obtained by maximizing the conditional likelihood $\prod_s L^*_c(\xi^*)$. Lindsay (1983) showed that $\hat{\xi}^*$ is the semiparametric efficient estimator for $\xi^*$ in the presence of the nuisance $q_s$'s, in the following sense. Suppose that instead of treating the $q_s$'s as nuisance parameters, we model them as random variables from arbitrary unknown mixing distributions $Q_Z$ which may depend on $Z$. The mixture model for the distribution of $(Y|Z)$, marginally over $q_s$, is now semiparametric in that $f(Y|Z; \xi^*, q_s)$ is a regular parametric model, while $Q_Z$ is non-parametric. Theorem 1, proved in Appendix A, extends Lindsay's (1983) result by establishing the optimality of $\hat{\xi}^*$ in this more general context.

**Theorem 1.** Let $f(Y|Z; \xi^*, q_s)$, $\hat{\xi}^*$, $q_s$ and $Q_Z$ be defined as above, so that the model $f(\cdot; \xi^*, q_s)$ admits $\sum_i Y_i$ as a complete sufficient statistic for $q_s$. Then, under regularity conditions given in Appendix A, as $J \to \infty$, $\hat{\xi}^*$ achieves the Cramèr-Rao lower bound for estimation of $\xi^*$ in the presence of unknown $Q_Z$.

To further extend this result to include a model for $X_i$, define the matrix $X$ analogously to $Z$, and extend (3) to include the term $\beta^T_x \sum_i X_i Y_i/a(\phi)$. Now, define $p_0 \equiv p_0(X_i|Z_i; \alpha)$ to be the density or pmf of $(X_i|Y_i = 0, Z_i)$ in stratum $s$, governed by a finite-dimensional parameter $\alpha$. Note that using $Y_i = 0$ is arbitrary; one could define $p_0$ for $Y_i = y_0$ for any $y_0$ in the support of $Y_i$. We assume that $p_0$ does not depend on the stratum intercept $q_s$, although it may depend in other parametric ways on $s$. Finally, let $\xi = (\beta, \phi, \alpha)^T$. We now develop the conditional likelihood for $\xi$ arising from data $(X_i, Y_i|Z_i)$.
Without loss of generality, model (1) can be re-expressed in terms of the odds
\[
\theta(Y_i|X_i, Z_i) = \frac{f(Y_i|X_i, Z_i)}{f(Y_i = 0|X_i, Z_i)} = \exp \left\{ q_s Y_i + \beta_x^T Z_i Y_i + \beta_x^T X_i Y_i \right\}/a(\phi) + c(Y_i, \phi)
\]
where \( c(Y_i, \phi) = c^*(Y_i, \phi) - c^*(0, \phi) \). Now, define the odds \( \tilde{\theta}(Y_i|Z_i) = f(Y_i|Z_i)/f(Y_i = 0|Z_i) \), marginally over \( X_i \). Then, following the approach of Satten and Kupper (1993) for the logistic regression model, we have the following two results. First, it can be shown that
\[
\tilde{\theta}(Y_i|Z_i) = \int \theta(Y_i|x, Z_i) p_0(x|Z_i) \, dx.
\]
For the exponential family model given by (1) and (2),
\[
\tilde{\theta}(Y_i|Z_i) = \exp \left\{ q_s Y_i + \beta_x^T Z_i Y_i \right\}/a(\phi) + c(Y_i, \phi)
\]
\[
\times \int_x \exp \left\{ \beta_x^T x Y_i/a(\phi) \right\} p_0(x|Z_i; \alpha) \, dx.
\]
(4)
Second, letting \( p(X_i|Y_i, Z_i) \) be the density or pmf of \((X_i|Y_i, Z_i)\), it can be shown that
\[
p(X_i|Y_i, Z_i) = p_0(X_i|Z_i) \theta(Y_i|X_i, Z_i)/\tilde{\theta}(Y_i|Z_i),
\]
(5)
which is free of \( q_s \) and simplifies to
\[
p(X_i|Y_i, Z_i) = \frac{p_0(X_i|Z_i; \alpha) \exp \left\{ \beta_x^T X_i Y_i/a(\phi) \right\}}{\int_x \exp \left\{ \beta_x^T x Y_i/a(\phi) \right\} p_0(x|Z_i; \alpha) \, dx}.
\]
(6)
We are now in a position to write the likelihood \( L \) for \((\xi, q_s)\) arising from the joint distribution of \((X, Y|Z)\). This is conveniently expressed via the decomposition
\[
p(X|Y, Z)f(Y|Z) = L(\xi, q_s) = L.
\]
(7)
By expansion of (7) and analogy to (3), it is easy to see that, for fixed \( \xi \), \( \sum_i Y_i \) is a complete sufficient statistic for the nuisance \( q_s \) in \( L \). Again, the conditional likelihood
\[
L^C(\xi) = p(X|Y, Z)f(Y|\sum_i Y_i, Z) = \left\{ \prod_i p(X_i|Y_i, Z_i) \right\} \left\{ \frac{\prod_i \tilde{\theta}(Y_i|Z_i)}{\sum_{y \in Y} \prod_i \tilde{\theta}(y_i|Z_i)} \right\},
\]
(8)
is free of \( q_s \). Here, \( Y = Y(Y) \) is the set of vectors \( y = (y_1, \ldots, y_n)^T \) such that \( \sum_i y_i = \sum_i Y_i \). When \( Y_i \) is continuous, the sum \( \sum_{y \in Y} \) is replaced by an integral.
Now, the results of Lindsay (1983) and Theorem 1 again apply, and under similar regularity conditions as $J \to \infty$, the conditional likelihood estimator $\hat{\xi}$ obtained by maximizing $\Pi_s L_s^\xi(\xi)$ is semiparametric efficient for $\xi$ in the presence of the nuisance parameters $q_s$. A corollary of this result is that the estimator $(\hat{\beta}, \hat{\phi})$ in $\hat{\xi} = (\hat{\beta}, \hat{\phi}, \hat{\alpha})$ is semiparametric efficient for $(\beta, \phi)$ in the presence of the $q_s$'s and $\alpha$.

3 Conditional likelihood estimators when $X_i$ may be missing

3.1 Efficient estimator

Consider the setting where $X_i$ may be missing and define $R_i \in \{0, 1\}$ to be an indicator variable for whether or not $X_i$ is completely observed for the $i$th record. We assume that within stratum $s$, $X_i$ is missing at random (MAR), i.e., $R_i \perp X_i | (Y_i, Z_i, s)$, and that $R_i \perp R_{i'}, i \neq i'$. Similarly to $Y$, define the vector $R = (R_1, \ldots, R_n)^T$. Further define $X_{\text{obs}}$ to be the observed rows of $X$. Let $\Pr(R_i = 1 | Y_i = y, X_i, Z_i) = \pi(y, Z_i; \gamma)$, where $\gamma$ is a finite-dimensional nuisance parameter, and where MAR ensures that $\pi(\cdot)$ does not depend on $X_i$. The full likelihood $L$ arising from stratum-level data $(X_{\text{obs}}, R, Y | Z)$ can then be written

$$L = L(\xi, \gamma, q_s) = p(X_{\text{obs}} | R, Y, Z) \Pr(R | Y, Z) f(Y | Z).$$

(9)

Note that $p(X_{\text{obs}} | R, Y, Z)$ would be ambiguous without conditioning on $R$, which indicates which components of $X$ are included in $X_{\text{obs}}$. This factor is explicitly written

$$p(X_{\text{obs}} | R, Y, Z) = \prod_i^p p(X_i | R_i = 1, Y_i, Z_i)^R = \prod_i p(X_i | Y_i, Z_i)^R,$$

the second equality resulting from the MAR assumption.

Now, by analogy to the case of no missing $X_i$ and equation (3), $\sum_i Y_i$ is a complete sufficient statistic for $q_s$ in $L$. The $q_s$'s are therefore eliminated from $L$ by conditioning
on $\sum_i Y_i$, resulting in the conditional likelihood

$$L^c(\xi, \gamma) = p(\mathbf{X}_{\text{obs}}| \mathbf{R}, \mathbf{Y}, \mathbf{Z}; \xi) \Pr(\mathbf{R}| \mathbf{Y}, \mathbf{Z}; \gamma) f(\mathbf{Y}| \sum_i Y_i, \mathbf{Z}; \xi).$$

Note that parameters $\gamma$ and $\xi$ are completely separable in $L^c$. Therefore, for inferences on $\xi$ ignoring $\gamma$,

$$L^c(\xi) \propto \left\{ \prod_i p(X_i|Y_i, Z_i)^{R_i} \right\} \left\{ \frac{\prod_i \tilde{\theta}(Y_i|Z_i)}{\sum_{y \in y} \prod_i \bar{\theta}(y_i|Z_i)} \right\}, \quad (10)$$

just as in (8).

The conditional likelihood $L^c(\xi)$ was proposed by Satten and Kupper (1993) and Satten and Carroll (2000) for the case of logistic regression. Theorem 1 and the subsequent development of Section 2 apply directly to $L^c(\xi)$ even when $X_i$ may be missing, so that the conditional likelihood estimator for $\xi$ obtained by maximizing $\prod_i L^c(\xi)$ is semiparametric efficient for $(\beta, \phi)$ in the presence of the nuisance parameters $q_s$'s and $\alpha$. This result was suggested by Satten and Carroll in their discussion. Note that if $X_i$ is never missing, $L^c$ still holds and is given by (8). It does not reduce to the standard conditional likelihood, and is in fact more efficient because it exploits the model for $p_0$ to extract information on $(\beta, \phi)$ contained in $(X_i|Y_i, Z_i)$.

### 3.2 Sub-optimal estimators

In Section 3.1, $L^c$ was derived by considering the joint distribution of $(\mathbf{X}_{\text{obs}}, \mathbf{R}, \mathbf{Y}| \mathbf{Z})$, resulting in a likelihood which depends on the nuisance distribution $p_0$ of $(X_i|Y_i = 0, Z_i)$, even when $X_i$ is never missing. To reduce the dependence of $\beta$-inferences on $p_0$, one might begin with the conditional distribution

$$f(\mathbf{Y}|\mathbf{X}_{\text{obs}}, \mathbf{R}, \mathbf{Z}) = \prod_i f(Y_i|R_i = 1, X_i, Z_i)^{R_i} f(Y_i|R_i = 0, Z_i)^{1-R_i}, \quad (11)$$

where again $\mathbf{R}$ plays the dual roles of conditioning statistic and selection operator.

As in (9), $\sum_i Y_i$ is sufficient for $q_s$ in (11), so that conditioning on it will eliminate $q_s$. The resulting conditional likelihood

$$L^c_{\text{subopt}}(\xi, \gamma) = f(\mathbf{Y} | \sum_i Y_i, \mathbf{X}_{\text{obs}}, \mathbf{R}, \mathbf{Z}; \xi, \gamma)$$
could therefore be used instead of \( L^c \) for inferences on \( \beta \). Noting that the odds
\[
\frac{f(Y_i|R_i=1, X_i, Z_i)}{f(Y_i=0|R_i=1, X_i, Z_i)} = \frac{\theta(Y_i|X_i, Z_i) - \pi(Y_i, Z_i)}{\pi(Y_i=0, Z_i)},
\]
(12)
we can write
\[
L^c_{subopt} = \frac{\prod_i \pi(Y_i, Z_i)^{R_i} \theta(Y_i|X_i, Z_i)^{R_i} \{1 - \pi(Y_i, Z_i)\}^{1-R_i} \tilde{\theta}(Y_i|Z_i)^{1-R_i}}{\sum_y \prod_i \pi(y_i, Z_i)^{R_i} \theta(y_i|X_i, Z_i)^{R_i} \{1 - \pi(y_i, Z_i)\}^{1-R_i} \theta(y_i|Z_i)^{1-R_i}}.
\]
(13)
Now, because \((\sum_i Y_i, X_{obs}, R)\) is not minimal sufficient, and therefore not complete, for \( q_s \) in likelihood \( L \), the maximum likelihood estimator for \((\beta, \phi)\) from \( L^c_{subopt} \) will not be semiparametric efficient. However, an advantage to \( L^c_{subopt} \) is that when \( X_i \) is never missing, it reduces to the standard conditional likelihood, reflecting the fact that, relative to \( L^c \), \( L^c_{subopt} \) is less dependent on the assumed model for \( p_0 \).

Likelihood \( L^c_{subopt} \) is similar to an expression proposed by Paik and Sacco (2000) and Paik (2002). The difference is that those authors’ proposal, which we denote \( L^c_{ps} \), omits the factors \( \pi(Y_i, Z_i)^{R_i}, \{1 - \pi(Y_i, Z_i)\}^{1-R_i}, \pi(y_i, Z_i)^{R_i}, \) and \( \{1 - \pi(y_i, Z_i)\}^{1-R_i} \) from (13). We study the performance of \( L^c_{ps} \) via simulations in the next section.

Using conditional likelihood \( L^c_{subopt} \), we now propose a new estimator for \((\beta, \phi)\) constructed as follows. Note that \( L^c_{subopt} \) contains nuisance parameters \( \alpha \) in \( \tilde{\theta} \) and \( \gamma \) in \( \pi(\cdot) \), so that estimation of \((\beta, \phi)\) requires estimation of \( \alpha \) and \( \gamma \) either simultaneously with, or prior to, estimation of \( \beta \). First, we use the likelihood \( \Pr(R|Y, Z; \gamma) \) to compute the maximum likelihood estimator \( \hat{\gamma} \) and plug \( \hat{\gamma} \) into \( L^c_{subopt} \). Then, while one might consider \( p(X_{obs}|R, Y, Z; \xi) \) for estimation of \( \alpha \), note that by (6), \( p(X_{obs}|R, Y, Z; \xi) \) depends not only on \( \alpha \), but also on \( (\beta_x, \phi) \). We propose handling this problem by first doing maximum likelihood estimation of \((\alpha, \beta_x, \phi)\) using \( p(X_{obs}|R, Y, Z; \alpha, \beta_x, \phi) \), yielding estimates \((\tilde{\alpha}, \tilde{\beta}_x, \tilde{\phi})\). The estimate \( \tilde{\alpha} \) is then plugged into (4) to compute the integral over \( x \) in \( \tilde{\theta}(Y_i|Z_i) \). The \( \tilde{\theta}(Y_i|X_i)'s \), as functions of \((\beta, \phi)\), are subsequently plugged into \( L^c_{subopt} \), which is then used for estimation of \((\beta, \phi)\). Note that \( \beta_x \) is estimated twice in this procedure, and this would evidently
give rise to a loss of efficiency. However, by not combining the two estimators of \(\beta_x\), the method reduces to standard maximum conditional likelihood when \(X_i\) is never missing, thereby reducing bias in estimation of \(\beta_x\) due to misspecification of \(p_0\).

A similar procedure can be used with \(L_{ps}^c\). It does not require estimation of \(\gamma\), and reduces to standard maximum conditional likelihood estimation when \(X_i\) is always observed. The estimator proposed by Paik and Sacco (2000) for binary \(Y_i\) also relies on likelihood \(L_{ps}^c\) and involves pre-estimation of \((\alpha, \beta_x)\), although the way in which \((\tilde{\alpha}, \tilde{\beta}_x)\) is plugged into \(L_{ps}^c\) for estimation of \(\beta\) differs somewhat from our proposal.

### 3.3 Efficient estimation using complete-record data

Suppose that the analyst only has access to data on records with observed \(X_i\). This situation may occur, for example, in studies using two-phase sampling strategies (Breslow and Cain, 1988; Yates, 1981) in which the publicly-released data contain only those records for which \(X_i\) was measured. Analysis is then conditional on the vector \(R\) of non-missing data indicators, so that an appropriate likelihood is

\[
\prod_i f(Y_i|R_i = 1, X_i, Z_i)^{R_i} = f(Y_{obs}|R, X_{obs}, Z_{obs}),
\]

where \(Y_{obs}\) and \(Z_{obs}\) are the components of \(Y\) and rows of \(Z\) corresponding to those in \(X_{obs}\). An alternative motivation for using only the complete record data \(Y_{obs}\), even if data on all records is available, is that it is too difficult or impractical to model the distribution \(p_0\) of \((X_i|Y_i = 0, Z_i)\), because either \(X_i\) and/or \(Z_i\) is of high dimension. Conditioning on \(X_{obs}\) and only modelling \(Y_{obs}\) avoids the need for \(p_0\).

To derive (14), define the odds of \(Y_i\) conditional on \(X_i\) being observed as \(\theta^*(Y_i|X_i, Z_i) = f(Y_i|R_i = 1, X_i, Z_i)/f(Y_i = 0|R_i = 1, X_i, Z_i)\). Then

\[
\theta^*(Y_i|X_i, Z_i) = \exp \left\{ \frac{q_i Y_i + \beta^T Z_i Y_i + \beta^{T*} X_i Y_i}{a(\phi) + c(Y_i, \phi) + B(Y_i, Z_i; \gamma)} \right\},
\]

where \(B(Y_i, Z_i; \gamma) = \log \{\pi(Y_i, Z_i; \gamma)/\pi(0, Z_i; \gamma)\}\). Therefore, the odds \(\theta(Y_i|X_i, Z_i)\) when there is no missing data can be corrected for possibly missing \(X_i\) by adding
the term $B(Y_i, Z_i, \gamma)$ to the linear predictor when conditioning on $X_i$ being observed (Breslow and Cain, 1988; Lipsitz et al., 1998).

From the form of $\theta^*(Y_i|X_i, Z_i)$ and by analogy to (3), it is seen that the nuisance parameter $q_s$ admits $\sum_i Y_i R_i$ as a complete sufficient statistic in (14). The complete-data conditional likelihood is therefore

$$L_{\text{complete}}^e = f(Y_{\text{obs}}|\sum_i Y_i R_i, R, X_{\text{obs}}, Z_{\text{obs}}),$$

and by Theorem 1, maximization yields the semiparametric efficient estimator for $\beta$ among estimators conditioning on $X_{\text{obs}}$ and using only complete-data records.

Computing $\theta^*(Y_i|X_i, Z_i)$ requires knowledge of the probabilities $\pi(Y_i, Z_i)$. In a public-use data set which only contains data on records with observed $X_i$, the $\pi(\cdot)$’s or some estimates thereof would presumably be released with the data as sampling probabilities. Alternatively, if one is using $L_{\text{complete}}^e$ to avoid specification of $p_0$, but all data are available, the likelihood for $\gamma$, $\Pr(R|Y, Z; \gamma)$, can be used to obtain the maximum likelihood estimator $\hat{\gamma}$ which can then be plugged into $L_{\text{complete}}^e$ for making inferences on $\beta$. Rathouz et al. (2002) have shown that using estimated $\hat{\gamma}$ in $L_{\text{complete}}^e$ yields more efficient $\beta$-inferences than using the true $\gamma$.

4 Simulation Study

4.1 Design

To compare the finite sample performance of the estimators presented in Section 3 under various assumptions about the distribution of $(X_i|Y_i, Z_i)$ and the missingness mechanism $\pi(\cdot)$, we conducted a simulation wherein (1) is a logistic model for binary $Y_i$. We sample from a population uniformly distributed among 200 strata ($s = 1, \ldots, 200$), letting $q_s = \{(s - 1)/199\}^2 - 1.5$, so that some strata will be at higher risk for $Y_i = 1$ than most. We consider two versions of (1). In both, covariate
$Z_i$ is a standard normal random variable. In the first model, $X_i$ is Bernoulli with logit\{$\Pr(X_i = 1|Z_i)\} = \log(0.3/0.7) + 0.6Z_i$, so that corr($X_i, Z_i) = 0.26$. In the second, $X_i = \min(X_i^*, 5.0)$, where $X_i^*|Z_i$ follows an exponential distribution such that log\{$E(X_i^*|Z_i)\} = -1/(2 \times 1.7^2) + Z_i/1.7$. This yields corr($X_i, Z_i) = 0.46$. Since $Z_i$ is standard normal, $E(X_i^*) = 1$. For both models, $\beta_z = \log(1.5)$. For binary $X_i$, $\beta_x = \log(2.0)$, while, for censored exponential $X_i$, $\beta_x = \log(1.3)$. In both cases, $E(Y_i) = 0.3$ marginally over $(X_i, Z_i, s)$. For each replicate dataset, four subjects were sampled from each of the 200 population strata, yielding a sample size of 800.

Missingness of $X_i$ was generated according to logit\{$\Pr(R_i = 1|Y_i, Z_i, X_i; \gamma)\} = \gamma_0 + \gamma_1 Y_i + \gamma_2 Z_i + \gamma_3 X_i$, allowing for a variety of missingness mechanisms. For MAR data generation, where missingness depends only on $(Y_i, Z_i)$, we set $\gamma = (1.6, -1, -1, 0)$ and $\gamma = (1.25, 0, -1, 0)$, refering to these cases as MAR-YZ and MAR-Z respectively. Note that under MAR-Z, $L^c_{ps}$ is equivalent to $L^c_{subopt}$ with known $\gamma$, as the factors containing $\pi(\cdot)$ cancel out of (12) and (13). Also, for $L^c_{complete}$ under MAR-Z, the term $B(Y_i, Z_i; \gamma) = 0$ in $\theta^*$, so that $L^c_{complete}$ is equivalent to the naive conditional likelihood obtained by simply dropping the records with missing $X_i$ from the analysis. In order to investigate the robustness properties of the estimators considered, we also considered two data generating mechanisms where missingness is not at random (NMAR). For missingness depending only on $(X_i, Z_i)$ (NMAR-XZ), we set $\gamma = (1.65, 0, -1, -1)$ for binary $X_i$, and $\gamma = (1.6, 0, -1, -1.3)$ for censored exponential $X_i$. As with MAR-Z, when $\pi(\cdot)$ depends on $(X_i, Z_i)$ but not on $Y_i$, $L^c_{subopt}$ with known $\pi(\cdot)$ reduces to $L^c_{ps}$, and the naive likelihood based only on complete records is equivalent to $L^c_{complete}$ with known $\pi(\cdot)$. Finally, we let missingness depend on $(Y_i, X_i, Z_i)$ (NMAR-YXZ), setting $\gamma = (2.0, -1, -1, -1)$ for binary $X_i$ and $\gamma = (1.95, -1, -1, -1.3)$ for censored exponential $X_i$. All four missing data mechanisms yielded 26% missingness.

Nine estimators of $\beta$ were computed for each replicate, some of which involved
misspecification of the distribution \( p_0(X_i|Z_i; \alpha) \) and/or the model \( \pi(Y_i, Z_i; \gamma) \). First, the efficient conditional likelihood estimator was obtained by maximizing \( \prod_i L^c(\xi) \) jointly over \( \xi = (\beta, \alpha) \). Assumed models for \( p_0 \) were logit\{Pr\( (X_i = 1|Y_i = 0, Z_i) \)\} = \( \alpha_0 + \alpha_1 Z_i + \alpha_2 Z_i^2 \) for binary \( X_i \) and log\{E\( (X_i^*|Y_i = 0, Z_i) \)\} = \( \alpha_0 + \alpha_1 Z_i + \alpha_2 Z_i^2 \) for censored exponential \( X_i \). Second, we maximized \( \prod_i L^c_{\text{subopt}, s}(\beta, \tilde{\alpha}, \gamma) \) for \( \beta \), as described in Section 3.2, using the same model for \( p_0 \) and the MAR model logit\{\( \pi_i(Y_i, Z_i; \gamma) \)\} = \( \gamma_0 + \gamma_1 Y_i + \gamma_2 Z_i \) for \( (R_i|Y_i, Z_i) \). Third, we maximized \( \prod_i L^c_{ps,s}(\beta, \tilde{\alpha}) \), again plugging in \( \tilde{\alpha} \) as described in Section 3.2. The next three estimators also used \( L^c \), \( L^c_{\text{subopt}} \) and \( L^c_{ps} \), but assumed incorrectly that \( X_i \Pi Z_i \), setting \( \alpha_1 = \alpha_2 = 0 \) in the model for \( p_0 \). Finally, we computed three estimators using only the complete-record data. The naive estimator was obtained by dropping all records with missing \( X_i \) and maximizing \( \prod_i L^c_{\text{complete}, s}(\beta) \) with \( B(\cdot) = 0 \). The complete-record estimator similarly used \( L^c_{\text{complete}}(\beta, \gamma) \), where \( \gamma \) was either known or was replaced by \( \tilde{\gamma} \), estimated just as for \( L^c_{\text{subopt}} \). Note that for the MAR-Z and NMAR-XZ mechanisms, \( L^c_{\text{complete}} \) with known \( \pi(\cdot) \) is equivalent to \( L^c_{\text{complete}} \) with \( B(\cdot) = 0 \). For each data generating mechanism and estimator, we report percent bias in \((\hat{\beta}_z, \hat{\beta}_x)\) and mean-square error efficiency relative to the efficient conditional likelihood estimator. Likelihoods were programmed in Fortran, and maximization was performed using nlm\( \text{inb()} \) in S-Plus v. 6.0 (MathSoft, 2000); software is available from the author upon request.

4.2 Results

Under the MAR data generating mechanisms (Table 1), when the distribution \( p_0 \) is correctly modelled, both \( L^c \) and \( L^c_{\text{subopt}} \) perform well in terms of bias. There is some loss of efficiency in using \( L^c_{\text{subopt}} \), presumably due to the fact that \( L^c \) optimally uses information on \( \beta_x \) in \( p(X_{obs}|R, Y, Z) \). By contrast, \( L^c_{ps} \) exhibits bias under MAR-YZ, which appears to be restricted to \( \hat{\beta}_z \) for binary \( X_i \), corroborating the findings in Paik
and Sacco (2000, Table 2). Under MAR-Z, \( L_{\text{subopt}}^c \) and \( L_{\text{ps}}^c \) perform similarly, as the two likelihoods are equivalent for known \( \pi(\cdot) \) and asymptotically equivalent when \( \pi(\cdot) \) is estimated. When the distribution \( p_0 \) is misspecified by assuming that \( X_i \| Z_i \), the \( L^c \) estimators exhibit bias which is in some cases very severe for each of the MAR data mechanisms. The bias is more controlled when using \( L_{\text{subopt}}^c \) for estimation, presumably due to reduced dependence on the assumed model for \( p_0 \). Note that the maximization algorithm did not always converge for the estimators based on \( L_{\text{ps}}^c \). Results are presented only for the replicates which did achieve convergence.

Under the NMAR data mechanisms (Table 2), all estimators based on \( L^c \), \( L_{\text{subopt}}^c \) and \( L_{\text{ps}}^c \) are biased. As expected, the estimators using the correct model for \( p_0 \) perform better than the ones assuming that \( X_i \| Z_i \), and generally \( L_{\text{subopt}}^c \) and \( L_{\text{ps}}^c \) outperform \( L^c \) in terms of bias. Again, this is most likely due to the fact that the \( L^c \) estimators rely heavily on the estimator \( \hat{\alpha} \) for \( p_0 \), and the NMAR mechanisms result in biased estimators of \( \alpha \). In neither of these settings is the estimator based on \( L_{\text{subopt}}^c \) clearly better or worse than that based on \( L_{\text{ps}}^c \), both being subject to some bias in estimation of \( \alpha \). It is interesting to note that under NMAR-XZ, \( L_{\text{subopt}}^c \) and \( L_{\text{ps}}^c \) are equivalent to one another, and both likelihoods are valid. So the only difference between the resulting estimators is the inconsistent estimator of \( \pi(\cdot) \) plugged into \( L_{\text{subopt}}^c \). However, both estimators are biased due to inconsistent estimation of \( p_0 \) when MAR does not hold. Again, the algorithm did not always converge for \( L_{\text{ps}}^c \).

For the methods based on \( L_{\text{complete}}^c \) (Tables 1 and 2), under MAR-YZ, the naive estimator with \( B(\cdot) = 0 \) is severely biased for \( \beta_z \), but performs well for \( \beta_x \). Under either MAR mechanism, use of \( L_{\text{complete}}^c \) with known or estimated \( \pi(\cdot) \) corrects this bias, but is much less efficient than \( L^c \) or \( L_{\text{subopt}}^c \). Under MAR-YZ, estimation of \( \pi(\cdot) \) slightly improves efficiency in \( \tilde{\beta}_z \) relative to when \( \pi(\cdot) \) is known. As expected, the naive method performs well in terms of bias for both MAR-Z and NMAR-XZ.
5 Conclusion

In this paper, we have compared the conditional likelihoods for several methods that have appeared in the literature for inference in conditional logistic regression models with missing covariates. Our approach uses the more general canonical exponential family formulation, so that the methods presented extend beyond conditional logistic regression to other fixed effects models with nuisance intercepts. The following presents our conclusions in the form of recommendations to users of these models.

First, if it is possible to model the distribution $p_0$ of the missing covariate $X_i$, likelihood $L^c$ of Satten and Kupper (1993) or Satten and Carroll (2000) will yield the semiparametric efficient estimator for $(\beta, \phi)$ in the presence of the nuisance parameters $\alpha$ and $q_a$. If available, and the analyst is confident of the assumed model for $p_0$, this is the method of choice. An alternative is to employ a new estimator which maximizes a sub-optimal conditional likelihood $L^c_{\text{subopt}}$. This method depends on the missingness model $\pi(\cdot)$ as well as on $p_0$, with some loss of efficiency, especially in $\tilde{\beta}_a$. However, its merit is that it is considerably more robust to misspecification of $p_0$ and reduces to standard maximum conditional likelihood when $X_i$ is never missing. The likelihood $L^c_{ps}$ due to Paik and Sacco (2000) is similar in form to $L^c_{\text{subopt}}$, but can exhibit bias due to omission of terms involving $\pi(\cdot)$. Unless it is impossible to model $\pi(\cdot)$, we do not recommend use of $L^c_{ps}$, as it is out-performed by $L^c_{\text{subopt}}$.

When only records with observed $X_i$ are available, the complete-record method of Lipsitz et al. (1998) using likelihood $L^c_{\text{complete}}$ is semiparametric efficient among estimators which are conditional on the observed $X_{\text{obs}}$. This method requires that the probabilities $\pi(\cdot)$ of observed $X_i$ be known. When $X_i$ is sometimes missing, but $(Y_i, Z_i)$ is available on all records, use of $L^c_{\text{complete}}$ is considerably less efficient than the other methods, although it does not require any distributional assumptions on $X_i$. 

15
Our simulations show that efficiency can be mildly improved by modelling the probabilities \( \pi(Y_i, Z_i; \gamma) \) and using an estimated value for \( \gamma \) specific to the data being analyzed, even if \( \gamma \) is known. A theoretical reason for this is given in Rathouz et al. (2002). However, when the analyst wishes to avoid distributional assumptions on \( X_i \), the recommended approach is that of Rathouz et al. (2002), who use a projection argument to obtain considerable efficiency improvement in \( \beta_z \) estimation as compared to \( L_{\text{complete}}^c \), with no further critical modelling assumptions. Finally, the naive complete-record estimator obtained by setting \( B(\cdot) = 0 \) in \( L_{\text{complete}}^c \) is the only one that is consistent when missingness depends on \( (X_i, Z_i) \), but not on \( Y_i \). Although this condition is not testable with data, if the investigator can justify it via external data or scientific considerations, this estimator may be of interest.

We pointed out in § 2 that \( p_0 \) may depend on the stratum variable \( s \). In some applications, stratum level information is in fact available, and if it is of concern that \( X_i \) is not independent of \( q_s \), then it would be important to include \( s \) in the model for \( p_0 \). The likelihood functions \( L^c, L_{\text{subopt}}^c \) and \( L_{ps}^c \) are easily extended to allow \( p_0 \) to depend on \( s \) in some parametric way with no additional development required. In a similar spirit, the model for \( \pi(\cdot) \) can incorporate parametric effects of \( s \) on the missingness probabilities.

We stated in § 1 that \( X_i \) may be vector-valued, but we have assumed throughout that \( X_i \) is either completely observed or completely missing. Some of the methods presented here extend to the more general case wherein \( X_i \) is partially observed. If \( L^c \) is being used for inferences, then the right-hand factor of (10) would remain the same. As presented, the left-hand factor in (10) contains the full joint distribution of \( (X_i|Y_i, Z_i) \) for records in which \( R_i = 1 \). By extension, in records where \( X_i \) is only partially observed, \( p(X_i|Y_i, Z_i)^{R_i} \) would be replaced by the distribution of the observed part of \( X_i \), conditional on \( (Y_i, Z_i) \), but marginal over the unobserved part.
of $X_i$. Similarly, for $L^c_{\text{subopt}}$, (13) could be extended to include a different version of $\tilde{\theta}$ for every $i$ depending on the missingness pattern observed for record $i$. As long as $p_0(x|Z_i; \alpha)$ is estimated for the full vector $X_i$, then the distribution of any missing subvector, given the observed part of $X_i$ and $Z_i$ is also available. For $L^c_{\text{subopt}}$, the factors $\pi$ and $1 - \pi$ in (12) would be replaced by the corresponding probabilities of the observed missingness patterns. A similar approach would apply to $L^c_{\text{PS}}$. In contrast to these methods, the complete-record estimator does not easily extend to be able to exploit records with partially observed $X_i$.

**Acknowledgments**

The author thanks an associate editor and two referees for helpful suggestions that improved the paper considerably. This material is based upon work supported by the National Science Foundation under Grant No. 0096412.

**Appendix A: Technical details**

**A.1 Preliminaries**

Let $Z(s)$ denote the matrix $Z$ for stratum $s$. For each $J$, assume that all inferences are conditional on the sequence $\{Z(s)\}_s = Z(1), \ldots, Z(J)$. In using $L^c(\xi^*)$ for inferences on $\xi^*$, the maximum conditional likelihood estimator $\hat{\xi}^*$ is the solution to

$$
\sum_{s=1}^{J} U^c_s(\xi^*) = 0,
$$

where $U^c = (\partial \log L^c / \partial \xi^*)$. Following standard asymptotic theory, as $J \to \infty$, $\sqrt{J}(\hat{\xi}^* - \xi^*)$ converges in law to a Gaussian random variable with variance equal to the inverse of the limiting information,

$$
\lim_{J \to \infty} \frac{1}{J} \sum_{s=1}^{J} I^c_s,
$$

(A.1)

where $I^c_s = E(U^{s c} U^{s T c})$. In order to ensure regular $\xi^*$-inferences from $U^{s c}$, suppose that $\{Z(s)\}_s$ is such that the limit (A.1) is positive definite.
For ease of presentation, we assume that the sample size \( n \) is constant across strata \( s \), although it is possible to relax this assumption. Suppose that for every \( z \in \text{support}(Z) \), the distribution \( Q_z \) of \( (q_s | Z = z) \) is absolutely continuous on the real line, and that the \( q_s \)'s are independent across \( s \) given \( \{Z(s)\} \). Lindsay (1983) studied the problem of efficient estimation of \( \xi^* \) in the presence of the nuisance mixing distribution \( Q_z \); however, his results for exponential family model (1) and (2) were restricted to the case where \( Z(s) \) is constant in \( s \). In what follows, we sketch his development and generalize his result to the setting where \( Z(s) \) varies in \( s \). Although we assume for ease of presentation that \( \xi^* \) is uni-dimensional, Lindsay also showed that the extension to \( \text{dim}(\xi^*) > 1 \) is straightforward.

Lindsay’s Sections 2 and 3 treat the generic problem of estimation of a one-dimensional parameter in the presence of a nonparametric nuisance function. He defined a “modified minimal Fisher information,” \( J \) (Lindsay denoted this quantity \( i^{**} \)) and showed that the inverse of \( J \) is the Cramér-Rao lower variance bound for consistent estimators of \( \xi^* \), so that when \( \mathcal{I}^c = J \), \( \hat{\xi}^* \) is semiparametric efficient. These results apply directly to our setting, where the nuisance function will be the mapping \( Q \) defined in A.2 below.

In his Section 4, Lindsay treated the case where the nuisance function is the mixing distribution \( Q_z \), but where \( Z(s) = z \) is fixed in \( s \), providing conditions under which \( \mathcal{I}^c = J \). Suppose that for fixed \( \xi^* \), the complete sufficient statistic \( \Sigma_i Y_i \) has an exponential family density with canonical parameter \( q_s \), as in (3). Given that the true mixing distribution \( Q_{z,0} \) is absolutely continuous on an interval of the real line, then \( \mathcal{I}^c = J \) (Lindsay, 1983, Corollary 4.4). To see the key steps in the development of this result, define the mixture model

\[
f^M(Y|Z; \xi^*, Q_z) = \int_q f(Y|Z; \xi^*, q)\, dQ_z(q),
\]
parameterized by $\xi^*$ and $Q_z$. Now, define the class of centered likelihood ratio scores

$$V(c_z, P_z) = c_z \left[ \frac{f^M(Y|Z; \xi^*_0, P_z^*)}{f^M(Y|Z; \xi^*_0, Q_{z,0})} - 1 \right],$$

(A.2)

indexed by $c_z \geq 0$ and $P_z$, where $\xi^*_0$ and $Q_{z,0}$ are the true parameter values. In (A.2), $P_z$ is any mixing distribution for $q_s$ on the real line such that $V(c_z, P_z)$ has finite variance $(\xi^*_0, Q_{z,0})$. Scores (A.2) arise as the right-handed $\tau$ derivative from the one-parameter (in $\tau$) mixture model $f^M\{Y|Z; \xi^*_0, (1 - c_z \tau) Q_{z,0} + c_z \tau P_z\}$ evaluated at $\tau = 0^+$ for given $c_z$ and $P_z$. Define $C_z$ to be the $L^2$-closure of the set of all possible centered likelihood ratio scores $V(c_z, P_z)$ over $c_z$ and $P_z$. Finally, let $U^* = (\partial \log L^*/\partial \xi^*)$, where $L^*$ is given in (3). Then, $\mathcal{T}^c = i$ follows from the facts that $U^{*c} = U^* - E(U^*| \sum_i Y_i, Z)$ and that $E(U^*| \sum_i Y_i, Z) \in C_z$.

**A.2 Proof of Theorem 1**

To extend this result to where $Z(s)$ varies across $s$, define the sequence $\{Y(s)\}_s = Y(1), \ldots, Y(J)$. Define $Q$ to be the mapping from the support of $Z$ to the space of absolutely continuous distributions on the real line such that $Q(z) = Q_z$. Similarly, let $P$ be any mapping from support($Z$) to the space of mixing distributions on the real line such that for each $z$ and $P(z) = P_z$, (A.2) has finite variance. Let $c(z) = c_z$ be any mapping from support($Z$) to the non-negative real line. Now for any given $P$ and $c(\cdot)$, define the one-parameter product mixture model

$$\prod_{s=1}^J f^M\{Y(s)|Z(s); \xi^*_0, (1 - c_{Z(s)} \tau) Q_{Z(s),0} + c_{Z(s)} \tau P_{Z(s)}\}.$$

(A.3)

Then, the $\tau$-score at $\tau = 0^+$ from (A.3) is

$$V\{c(\cdot), P\} = \sum_{s=1}^J c_{Z(s)} \left[ \frac{f^M(Y|Z; \xi^*_0, P_{Z(s)})}{f^M(Y|Z; \xi^*_0, Q_{Z(s),0})} - 1 \right].$$

Let $C$ be the $L^2$-closure of the set of all possible scores $V\{c(\cdot), P\}$ over $c(\cdot)$ and $P$.

Now, summing over strata $s$, the conditional score $\sum_{s=1}^J U^{*c}_s$ is

$$\sum_{s=1}^J U^{*c}_s = \sum_{s=1}^J \left[ U^*_s - E\{U^*_s| \sum_i Y_i(s), Z(s)\} \right] = \sum_{s=1}^J U^*_s - \sum_{s=1}^J E\{U^*_s| \sum_i Y_i(s), Z(s)\}.$$
So, to show that $\hat{\xi}$ is semiparametric efficient when $Z(s)$ varies across $s$, it suffices to show to that $\sum_{s=1}^{J} E\{U^*_s|\sum_i Y_i(s), Z(s)\}$ is in $C$. But this is obviously true since (i) $C$ is simply the set of positive-coefficient linear combinations of elements of the $C_{Z(s)}$'s; (ii) $E\{U^*_s|\sum_i Y_i(s), Z(s)\} \in C_{Z(s)}$ for all $s$; and (iii) $\sum_{s=1}^{J} E\{U^*_s|\sum_i Y_i(s), Z(s)\}$ is a positive-coefficient linear combination of the quantities $E\{U^*_s|\sum_i Y_i(s), Z(s)\}$.

References


Table 1. Simulation results for missing at random (MAR) data generation based on 1000 replicates. True values are $\beta_z = 0.405$, $\beta_x = 0.693$ for binary $X_i$, and $\beta_x = 0.262$ for censored exponential $X_i$.

<table>
<thead>
<tr>
<th>Missing Mechanism</th>
<th>Method</th>
<th>$\beta_z$</th>
<th>$\beta_x$</th>
<th>% Bias</th>
<th>% Rel. Eff.</th>
<th>$\beta_z$</th>
<th>$\beta_x$</th>
<th>% Bias</th>
<th>% Rel. Eff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAR-YZ</td>
<td>$L_c$, $X \parallel Z$</td>
<td>1.2</td>
<td>-1.7</td>
<td>100</td>
<td>100</td>
<td>0.3</td>
<td>-3.2</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>$L_{subopt}^c$, $X \parallel Z$</td>
<td>1.1</td>
<td>-0.2</td>
<td>95</td>
<td>69</td>
<td>-2.1</td>
<td>-4.8</td>
<td>82</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>$L_{ps}^c$, $X \parallel Z$</td>
<td>-2.6</td>
<td>19.1</td>
<td>88</td>
<td>37</td>
<td>-21.1$^1$</td>
<td>29.2$^1$</td>
<td>46$^1$</td>
<td>34$^1$</td>
</tr>
<tr>
<td></td>
<td>$L^c$, $X \parallel Z$</td>
<td>20.5</td>
<td>12.4</td>
<td>66</td>
<td>92</td>
<td>33.0</td>
<td>67.3</td>
<td>50</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>$L_{subopt}^c$, $X \parallel Z$</td>
<td>9.7</td>
<td>-1.5</td>
<td>91</td>
<td>69</td>
<td>17.4</td>
<td>-9.4</td>
<td>83</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>$L_{ps}^c$, $X \parallel Z$</td>
<td>10.2</td>
<td>-0.2</td>
<td>88</td>
<td>40</td>
<td>21.1$^2$</td>
<td>-33.9$^2$</td>
<td>57$^2$</td>
<td>16$^2$</td>
</tr>
<tr>
<td></td>
<td>$L_{complete}^c$, $B(\cdot) = 0$</td>
<td>-37.9</td>
<td>0.7</td>
<td>23</td>
<td>60</td>
<td>-39.9</td>
<td>0.8</td>
<td>28</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>$L_{complete}^c$, known $\pi(\cdot)$</td>
<td>2.5</td>
<td>0.9</td>
<td>48</td>
<td>60</td>
<td>0.1</td>
<td>2.5</td>
<td>56</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>$L_{complete}^c$, est. $\pi(\cdot)$</td>
<td>2.5</td>
<td>0.9</td>
<td>51</td>
<td>60</td>
<td>0.1</td>
<td>2.5</td>
<td>59</td>
<td>51</td>
</tr>
</tbody>
</table>

| MAR-Z             | $L_c$, $X \parallel Z$ | 1.2 | -1.9 | 100 | 100 | 0.5 | -3.3 | 100 | 100 |
|                   | $L_{subopt}^c$, $X \parallel Z$ | 0.9 | -0.4 | 96 | 71 | -2.4 | -2.4 | 81 | 73 |
|                   | $L_{ps}^c$, $X \parallel Z$ | 0.9 | -0.2 | 96 | 70 | -2.6 | -2.0 | 80 | 71 |
|                   | $L^c$, $X \parallel Z$ | 20.5 | 22.6 | 65 | 68 | 33.0 | 99.1 | 47 | 13 |
|                   | $L_{subopt}^c$, $X \parallel Z$ | 9.5 | -1.5 | 92 | 70 | 16.9 | -8.5 | 80 | 70 |
|                   | $L_{ps}^c$, $X \parallel Z$ | 9.4 | -1.5 | 92 | 69 | 17.0 | -8.7 | 80 | 68 |
|                   | $L_{complete}^c$, $B(\cdot) = 0$ | 1.9 | 0.3 | 54 | 60 | 0.0 | 3.2 | 56 | 55 |
|                   | $L_{complete}^c$, est. $\pi(\cdot)$ | 1.9 | 0.3 | 57 | 60 | 0.0 | 3.1 | 59 | 55 |

% Rel. Eff.: Mean squared error relative efficiency ($\times 100$) compared to $L_c$ with the correct model for $X_i$, $X_i \parallel Z_i$.

Note: For MAR-Z, the naive estimator, $L_{complete}^c$, $B(\cdot) = 0$ is equal to the $L_{complete}^c$ estimator with known $\pi(\cdot)$.

1Result for 997 replicates; did not converge for 3 replicates.

2Result for 986 replicates; did not converge for 14 replicates.
Table 2. Simulation results for not missing at random (NMAR) data generation
based on 1000 replicates. True values are $\beta_z = 0.405$, $\beta_x = 0.693$ for binary $X_i$, and

$$
\beta_x = 0.262
$$

for censored exponential $X_i$.

<table>
<thead>
<tr>
<th>Missing Mechanism</th>
<th>Method</th>
<th>Binary $X_i$</th>
<th>Cens. Exp. $X_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\beta_z$</td>
<td>% Bias</td>
</tr>
<tr>
<td>NMAR-XZ</td>
<td>$L^c$, $X \parallel Z$</td>
<td>8.9</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>$L^c_{\text{subopt}}$, $X \parallel Z$</td>
<td>8.7</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>$L^c_{\text{ps}}, X \parallel Z$</td>
<td>8.5</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>$L^c$, $X \parallel Z$</td>
<td>20.5</td>
<td>15.2</td>
</tr>
<tr>
<td></td>
<td>$L^c_{\text{subopt}}, X \parallel Z$</td>
<td>13.6</td>
<td>-0.9</td>
</tr>
<tr>
<td></td>
<td>$L^c_{\text{ps}}, X \parallel Z$</td>
<td>13.6</td>
<td>-0.2</td>
</tr>
<tr>
<td></td>
<td>$L^c_{\text{complete}}, B(\cdot) = 0$</td>
<td>1.9</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>$L^c_{\text{complete}}, \text{est. } \pi(\cdot)$</td>
<td>8.3</td>
<td>1.0</td>
</tr>
<tr>
<td>NMAR-XYZ</td>
<td>$L^c$, $X \parallel Z$</td>
<td>11.9</td>
<td>-28.2</td>
</tr>
<tr>
<td></td>
<td>$L^c_{\text{subopt}}, X \parallel Z$</td>
<td>12.0</td>
<td>-26.5</td>
</tr>
<tr>
<td></td>
<td>$L^c_{\text{ps}}, X \parallel Z$</td>
<td>9.8</td>
<td>-8.0</td>
</tr>
<tr>
<td></td>
<td>$L^c$, $X \parallel Z$</td>
<td>20.5</td>
<td>-17.4</td>
</tr>
<tr>
<td></td>
<td>$L^c_{\text{subopt}}, X \parallel Z$</td>
<td>15.8</td>
<td>-26.7</td>
</tr>
<tr>
<td></td>
<td>$L^c_{\text{ps}}, X \parallel Z$</td>
<td>15.9</td>
<td>-24.1</td>
</tr>
<tr>
<td></td>
<td>$L^c_{\text{complete}}, B(\cdot) = 0$</td>
<td>-34.1</td>
<td>-25.0</td>
</tr>
<tr>
<td></td>
<td>$L^c_{\text{complete}}, \text{est. } \pi(\cdot)$</td>
<td>13.4</td>
<td>-25.0</td>
</tr>
</tbody>
</table>

% Rel. Eff.: Mean squared error relative efficiency ($\times100$) compared to $L^c$
with the correct model for $X_i, X_i \parallel Z_i$.

Note: For NMAR-XZ, the naive estimator, $L^c_{\text{complete}}, B(\cdot) = 0$ is equal to
the $L^c_{\text{complete}}$ estimator with known $\pi(\cdot)$.

\(^\dag\)Result for 992 replicates; did not converge for 8 replicates.

\(^\dag\)Result for 989 replicates; did not converge for 11 replicates.