

# Semiparametric Inference in Matched Case-Control Studies with Missing Covariate Data

BY PAUL J. RATHOUZ

Department of Health Studies  
University of Chicago  
5841 South Maryland Avenue, MC2007  
Chicago, IL 60637 USA  
prathouz@health.bsd.uchicago.edu

GLEN A. SATTEN

Centers for Disease Control and Prevention  
Atlanta, GA 30333 USA  
gsatten@cdc.gov

RAYMOND J. CARROLL

Department of Statistics  
Texas A&M University  
College Station TX 77843-3143 USA  
carroll@stat.tamu.edu

## Summary

We consider the problem of matched studies with a binary outcome that are analyzed using conditional logistic regression, and for which data on some covariates is missing for some study participants. Methods for this problem involve either modeling the distribution of missing covariates or modeling the probability of data being missing. For this second approach, the previously proposed method did not make use of data for those persons with missing covariate data except in the model for the missingness. We propose a new class of estimators that use outcome and available covariate data for all study participants, and show that a particular member of this class always has better efficiency than the previously-proposed estimator. We illustrate the efficiency gains that are possible with our approach using simulated data.

*Some key words:* Case-control study; Conditional inference; Estimating equations; Missing data; Projections; Robustness; Semiparametric; Two-stage studies.

**Short title:** Matched Studies with Missing Data

# 1 Introduction

In highly stratified cohort studies or matched case-control studies, conditional logistic regression (Breslow & Day, 1980, p.248) is the usual method of analysis of effects of covariates  $(X_i, Z_i)$  on a binary case-control outcome  $D_i$ . The conditional logistic regression model for  $(D_i|X_i, Z_i)$  accounts for stratum effects via a stratum-specific intercept which is considered a nuisance parameter. Conditioning both on the covariate set,  $\{(X_i, Z_i) : i \in s\}$ , and on the number of case subjects,  $\sum_{i \in s} D_i$ , within stratum  $s$  not only reflects the case-control sampling strategy, but also eliminates the nuisance parameter. Indeed, because  $\sum_i D_i$  is a complete sufficient statistic for the stratum intercept, the maximum conditional likelihood estimator of the coefficients  $\beta$  of  $(X_i, Z_i)$  is semiparametric efficient in this problem (Godambe, 1976; Lindsay, 1985). In many studies, however, direct application of this method is hindered because  $X_i$  may be missing on some subjects. The missing covariate problem is especially acute in stratified studies because missing  $X_i$  on one subject can in a naive analysis render the entire stratum uninformative. Gibbons & Hosmer (1991) pointed out this problem and studied several approximate methods of addressing it.

Two general approaches have been developed for obtaining consistent inferences in problems of missing covariates. The most direct of these is to model the distribution of the missing covariates  $X_i$ . To this end, Satten & Kupper (1993), Paik & Sacco (2000) and Satten & Carroll (2000) each propose conditional likelihood methods which rely on a model for the distribution of  $X_i$  among the control subjects. Satten & Carroll (2000) model  $(X_i|D_i = 0, Z_i)$  nonparametrically for  $(X_i, Z_i)$  with finite support, and, in a similar approach, Satten & Kupper (1993) exploit a surrogate for  $X_i$ . By contrast, Paik & Sacco (2000) assume that  $(X_i|D_i = 0, Z_i)$  is univariate with a distribution belonging to a canonical exponential family model. All of these authors assume that

$X_i$  is missing at random (Little & Rubin, 1987).

In each of these approaches, it can be shown via the full likelihood function that  $\sum_i D_i$  remains a complete sufficient statistic for the stratum-level intercept, even in the presence of missing  $X_i$ . By conditioning on  $\sum_i D_i$ , Satten & Kupper (1993) and Satten & Carroll (2000) develop a conditional likelihood function that provides for joint inferences on  $\beta$  and on the distribution of  $(X_i|D_i = 0, Z_i)$  that are semiparametric efficient in the presence of the stratum-level nuisance parameters. The approach by Paik & Sacco (2000) is suboptimal because their conditional likelihood is not of the same form, although it appears to work well in practice. Thus, when it is possible to model the distribution of  $X_i$ , the problem appears to be well solved.

In many problems, however, estimation of the distribution of  $X_i$  is seriously hindered because  $(X_i, Z_i)$  is of high dimension, and/or the distribution of  $X_i$  cannot be reasonably assumed to take a parametric form. In these cases, modelling the missingness process offers an alternative, albeit less efficient, approach to obtaining inferences on  $\beta$ . Lipsitz et al. (1998) propose such an approach, modelling the case-control status only among the subjects with complete data. They accomplish this by first conditioning on the observed pattern of missingness and then conditioning on the number of case subjects among those with complete data in each stratum. Conditioning on the missingness pattern requires a model for the missingness process, but requires no knowledge of the distribution of  $X_i$ . As with conventional conditional logistic regression, conditioning on both the covariates and the number of case subjects among those with complete data eliminates the stratum-level nuisance parameter, and yields a likelihood which depends neither on the marginal distribution of  $(X_i|Z_i)$  nor on that of  $(X_i|D_i = 0, Z_i)$ . We present this method in more detail in § 2.

A problem with the method of Lipsitz et al. (1998) is that it only weakly exploits data  $(D_i, Z_i)$  on subjects with missing  $X_i$ , thereby incurring a loss in efficiency. The

goal of this paper is to address this problem. Section 3 presents a class of estimating functions which potentially improve on the efficiency of the Lipsitz et al. (1998) approach. Within this class, a projected score does increase efficiency, but requires knowledge of the distribution of  $X_i$ . A practical approximation to the projection is proposed that exploits a working parametric model for the distribution of  $X_i$  and is easily computed. Section 4 presents simulations that suggest that the efficiency improvement using this approximation can be substantial and indeed requires only weak modelling assumptions. We remark on semiparametric efficiency in § 5.

## 2 Complete Subject Analysis

Consider a source population divided into strata indexed by  $s$ . Under the matched case-control sampling plan, a fixed number of persons with disease, case subjects, are sampled, and for each case subject a fixed number of persons without disease, control subjects, are sampled from the same stratum as the case subjects. The study sample comprises  $J$  such matched sets, and the numbers of case and control subjects in each set are considered fixed by design. Methodological development in this paper will focus on the stratum level data, and we will generally suppress the index  $s$  denoting stratum. Let  $D_i = 1$  denote case status and  $D_i = 0$  denote control status for the  $i$ th subject in a typical stratum  $s$ . Covariates  $Z_i$  or  $(X_i, Z_i)$  are collected retrospectively from each sampled subject, where  $Z_i$  is an always-observed vector of covariates and  $X_i$  is a vector of covariates which is possibly missing. Let  $R_i = 1$  or  $0$  indicate that  $X_i$  is observed or missing for the  $i$ th subject. Within stratum  $s$ , we are therefore sampling from the distributions of  $(R_i, Z_i|D_i)$  and  $(X_i|R_i = 1, Z_i, D_i)$ . We assume that  $X_i$  is missing at random; that is  $R_i \perp\!\!\!\perp X_i \mid (D_i, Z_i)$  (Little & Rubin, 1987).

Define the population prospective odds of disease given  $(X_i, Z_i)$  within stratum

$s$  to be  $\theta_i = \theta(X_i, Z_i) = \text{pr}(D_i = 1|X_i, Z_i)/\text{pr}(D_i = 0|X_i, Z_i)$ . Interest is in the finite-dimensional parameter  $\beta = (\beta_x^T, \beta_z^T)^T$  in the prospective logistic disease model

$$\log(\theta_i) = q(s) + \beta_z^T Z_i + \beta_x^T X_i, \quad (1)$$

where  $q(s)$  is a stratum-specific intercept which permits disease risk to vary by strata.

To account for missing data, let  $\text{pr}(R_i = 1|D_i = d, X_i, Z_i) = \pi(D_i = d, Z_i; \gamma) = \pi_i(d; \gamma)$  independently of stratum  $s$ , where  $\gamma$  is a finite-dimensional nuisance parameter, where missing at random ensures that  $\pi_i(\cdot)$  does not depend on  $X_i$ , and where the subscript  $i$  reflects the dependence of  $\pi_i(d; \gamma)$  on  $Z_i$ . For example,  $\pi_i(\cdot)$  might depend on  $(D_i, Z_i)$  via a logistic regression model. This induces a model for  $\theta_i^* = \theta^*(X_i, Z_i) = \text{pr}(D_i = 1|R_i = 1, X_i, Z_i)/\text{pr}(D_i = 0|R_i = 1, X_i, Z_i)$ , the prospective disease odds among the subjects with observed  $X_i$ . Given (1), Breslow & Cain (1988) and Lipsitz et al. (1998), show that

$$\log(\theta_i^*) = q(s) + \beta_z^T Z_i + \beta_x^T X_i + B_i = \log(\theta_i) + B_i, \quad (2)$$

where  $B_i = B(Z_i; \gamma) = \log\{\pi_i(1; \gamma)/\pi_i(0; \gamma)\}$ . Therefore, if only the subjects with observed  $X_i$  are analysed, consistent inferences are obtained by adding the offset term  $B_i$  to the linear predictor in model (1).

To denote the data on a given stratum, write  $Z = (Z_1, \dots, Z_i, \dots, Z_n)^T$  for the matrix of  $n$  row vectors of covariates  $Z_i$ . Similarly, define the matrix  $X$ , including the missing rows, and the vectors  $D$  and  $R$ . The vector  $R$  indicates which subjects have complete data. Further define  $X_{\text{obs}}$  to be the observed rows of  $X$ , and  $D_{\text{obs}}$  and  $Z_{\text{obs}}$  to be the components of  $D$  and rows of  $Z$  corresponding to those in  $X_{\text{obs}}$ .

Under the matched case-control sampling plan, the likelihood for  $\beta$  from the stratum  $s$  complete subject data is  $\text{pr}(X_{\text{obs}}, Z_{\text{obs}}|D_{\text{obs}}, R)$ , where,  $R$  not only selects the subjects with observed  $X_i$ , but is also a conditioning statistic. To eliminate the nuisance  $q(s)$ , we further condition on the unordered set of observed covariates,

$Q_{\text{obs}} = \{(X_i, Z_i) : R_i = 1\}$ , decoupled from their case-control status, and base inferences on the conditional likelihood  $L^{(1)} = \text{pr}(X_{\text{obs}}, Z_{\text{obs}} | Q_{\text{obs}}, D_{\text{obs}}, R)$ . To see how  $q(s)$  is eliminated in  $L^{(1)}$ , note that we can rewrite it as

$$L^{(1)} = \text{pr}(D_{\text{obs}} | \sum_i D_i R_i, X_{\text{obs}}, R, Z_{\text{obs}}) \quad (3)$$

and that

$$\begin{aligned} \text{pr}(D_{\text{obs}} | X_{\text{obs}}, R, Z_{\text{obs}}) = \exp \left\{ q(s) \sum_i D_i R_i + \beta_z^T \sum_i Z_i D_i R_i \right. \\ \left. + \beta_x^T \sum_i X_i D_i R_i + \sum_i B_i R_i - \sum_i R_i \log(1 + \theta_i^*) \right\}. \end{aligned} \quad (4)$$

As  $\sum_i D_i R_i$  is a complete sufficient statistic for  $q(s)$  in (4), the conditional likelihood (3) does not depend on  $q(s)$  (Lipsitz et al., 1998). Indeed,  $L^{(1)}$  provides efficient inferences on  $\beta$  in the presence of  $q(s)$  among all estimators which are functions only of the complete subject data  $(D_{\text{obs}}, X_{\text{obs}}, R, Z_{\text{obs}})$ , and condition on  $(X_{\text{obs}}, R, Z_{\text{obs}})$  (Godambe, 1976, Theorem 3.2). The score functions  $U^{(1)} = \partial \log L^{(1)} / \partial \beta^T$  take the same form as those in conventional conditional logistic regression (Breslow & Day, 1980), applied only to subjects with observed  $X_i$  and using the odds  $\theta_i^*$  instead of  $\theta_i$ .

Lipsitz et al. (1998) proposed to estimate  $\beta$  via the solution to  $\sum_s U^{(1)}(\beta, \gamma) = 0$  for fixed  $\gamma$ , where  $\sum_s$  denotes summation over strata  $s = 1, \dots, J$ . When  $\gamma$  is unknown, they estimate  $\beta$  by solving  $\sum_s U^{(1)}(\beta, \hat{\gamma}) = 0$  to obtain  $\hat{\beta}$ , where  $\hat{\gamma}$  is the solution to  $\sum_s T^\gamma(\gamma) = 0$ ,  $T^\gamma = \partial \log L^\gamma / \partial \gamma^T$ , and  $L^\gamma$  is the likelihood

$$L^\gamma = \text{pr}(R | D, Z) = \prod_i \pi_i(D_i; \gamma)^{R_i} \{1 - \pi_i(D_i; \gamma)\}^{1-R_i}.$$

In some applications,  $Z_i$  may include a surrogate vector  $Z_{Si}$  for  $X_i$ . Then  $Z_i = (Z_{Mi}^T, Z_{Si}^T)^T$ , where  $Z_{Mi}$  are the covariates of interest in model (1), and we make the surrogacy assumption that  $D_i \perp\!\!\!\perp Z_{Si} \mid (Z_{mi}, X_i)$  (Carroll et al., 1995). The foregoing method applies by replacing  $Z_i$  with  $Z_{Mi}$  in (1), (2) and (4). The surrogacy assumption can be tested by adding covariates  $Z_{Si}$  to (1), (2) and (4).

## 3 Efficiency Improvements

### 3.1 Introduction

Conditional likelihood  $L^{(1)}$  is inefficient because it discards information in  $(D_i, Z_i)$  from subjects for whom  $X_i$  is missing. Furthermore,  $L^{(1)}$  contains information on  $(R_i|D_i, Z_i)$ , which is ancillary for  $\beta$  and therefore is a threat to efficiency. This suggests that more efficient estimation of  $\beta$  can be achieved by using an estimating function where this information has been removed. The heuristic notion is to identify an estimating function  $U^a$  that (i) contains no  $\beta$ -information, i.e. is ancillary for  $\beta$ , (ii) is unbiased without any further modelling assumptions, and (iii) is positively correlated with  $U^{(1)}$ . Here, we take  $U^a$  as ‘ancillary for  $\beta$ ’ to mean that  $E(-\partial U^a/\partial\beta) = 0$ . If such a function can be found, a new estimating function  $U = U^{(1)} - U^a$  will provide inferences on  $\beta$  with a potential increase in efficiency.

In § 3.2, we propose a class of estimating functions  $U^a$  that satisfy conditions (i) and (ii). Then, using a projection argument, we identify a member of that class, denoted  $U_{\text{proj}}^a$ , that satisfies (iii) as well, and thereby define  $U^{\text{proj}} = U^{(1)} - U_{\text{proj}}^a$ . Although computing  $U^{\text{proj}}$  requires a model for the distribution of  $(X_i|Z_i)$ , we show that when this distribution is correctly specified,  $U^{\text{proj}}$  is at least as efficient as  $U^{(1)}$ . Finally, in § 3.4, using a working model for  $(X_i|Z_i)$ , we propose an approximation  $U^{\text{appr}}$  to  $U^{\text{proj}}$  for which computation is simpler. In § 4, two different simulations show that inferences about  $\beta$  using  $U^{\text{appr}}$  are more efficient than those made using  $U^{(1)}$ . Proofs of the Theorems are in an Appendix.

### 3.2 A class of estimating functions

For the  $n$  observations in a given stratum, there are  $2^n$  possible missingness patterns. Let these patterns be indexed by  $k = 0, \dots, 2^n - 1$  and let  $r_k$  be the  $n$ -vector with

$i$ th component equal to 0, if  $X_i$  is missing in the  $k$ th pattern, and 1 otherwise. Let  $\Delta_k = I(R = r_k)$ , where we recall that  $R = (R_1, \dots, R_n)^\top$  is the random vector that generates the observed missingness pattern.

Define a class of  $\beta$ -ancillary estimating functions of the form

$$U^a = \sum_{k=0}^{2^n-1} (\Delta_k - \epsilon_k) \psi_k, \quad (5)$$

where the  $\psi_k$ 's are any functions of  $(D, Z)$ , and where

$$\epsilon_k = E(\Delta_k \mid D, Z; \gamma) = \prod_i \pi_i(D_i; \gamma)^{r_{ki}} \{1 - \pi_i(D_i; \gamma)\}^{1-r_{ki}}.$$

Note that the  $\psi_k$ 's may be functions of  $(\beta, \gamma)$  as well as additional nuisance parameters  $\alpha$ . Importantly, the  $\psi_k$ 's are not functions of  $X$ , so that  $U^a$  can be computed using observed data. Since the  $\psi_k$ 's are not functions of  $R$ ,  $E(U^a) = E_D\{E_R(U^a \mid D, Z) \mid Z\} = 0$ , where  $E_A$  denotes expectation over  $A$ . Hence, any function  $U^a$  of the form (5) satisfies criteria (i) and (ii) in § 3.1. We therefore propose a class of estimating functions for  $\beta$  of the form

$$U(\beta, \gamma, \alpha) = U^{(1)}(\beta, \gamma) - U^a(\beta, \gamma, \alpha). \quad (6)$$

To use  $U$  for inferences for a given choice of the  $\psi_k$ 's, assume that there exists an  $\alpha$ -estimating function  $T^\alpha = T^\alpha(D, X_{\text{obs}}, R, Z; \alpha)$  and let  $\hat{\alpha}$  be the solution to  $\sum_s T^\alpha(\alpha) = 0$ . Suppose we estimate  $\gamma$  as in § 2, and  $\beta$  by solving  $\sum_s U(\beta, \hat{\gamma}, \hat{\alpha}) = 0$ . Theorem 1 characterises the asymptotic distribution of  $\hat{\beta}$ .

**THEOREM 1.** *Suppose that the modelling assumptions in § 2 hold, that  $\hat{\gamma}$  solving  $\sum_s T^\gamma(\gamma) = 0$  is  $\sqrt{J}$ -consistent, that  $\hat{\alpha}$  as defined above is  $\sqrt{J}$ -consistent for some  $\alpha^*$ , and that  $\hat{\beta}$  solves  $\sum_s U(\beta, \hat{\gamma}, \hat{\alpha}) = 0$ . Then, under mild regularity conditions as  $J \rightarrow \infty$ ,  $\hat{\beta} \rightarrow \beta$  in probability, and  $\sqrt{J}(\hat{\beta} - \beta) \rightarrow N(0, \mathcal{V})$  in distribution,  $\mathcal{V} = \lim_{J \rightarrow \infty} \mathcal{V}_J$ , where*

$$\mathcal{V}_J = J(\sum_s \mathcal{I}_\beta^{(1)})^{-1}(\sum_s \tilde{U}\tilde{U}^\top)(\sum_s \mathcal{I}_\beta^{(1)})^{-1}, \quad (7)$$

$\mathcal{I}_\beta^{(1)} = E(U^{(1)}U^{(1)\top})$ , and  $\tilde{U}$  is the residual from the least-squares regression of  $U$  on to  $\gamma$  scores  $T^\gamma$ ; that is,  $\tilde{U} = U - CT^\gamma$ , where  $C = (\sum_s UT^\gamma^\top)\{\sum_s T^\gamma T^\gamma^\top\}^{-1}$ .

*Remark.* We note that the form of the asymptotic variance  $\mathcal{V}$  follows from the facts that both  $U^{(1)}$  and  $T^\gamma$  are likelihood scores, so that minus the expected value of their derivatives is equal to their variances. Also, if we replace  $U$  by  $U^{(1)}$  in the Theorem, then the asymptotic variance of  $\hat{\beta}$  under the Lipsitz et al. (1998) approach with estimated  $\gamma$  is obtained here for the first time. From (7), we see that the asymptotic efficiency of  $U^{(1)}$  is improved by estimation of  $\gamma$ , even if  $\gamma$  is already known. This phenomenon has been previously noted in missing data problems (Robins et al., 1994).

### 3.3 Improved efficiency via projection

Although only one pattern of missingness is observed, we can imagine writing a likelihood corresponding to each of the  $2^n$  possible missingness patterns. Let  $D_{(k)}$  denote the components of  $D$ , and  $X_{(k)}$  the rows of  $X$ , for the subjects with observed values of  $X_i$  under the  $k$ th missingness pattern. Define the conditional likelihood (3) we would have computed for this missingness pattern to be  $L_k^{(1)} = \text{pr}(D_{(k)} \mid \sum_i D_i r_{ki}, X_{(k)}, R = r_k, Z)$ , and define  $U_k^{(1)} = \partial \log L_k^{(1)} / \partial \beta^T$ . Since only one  $\Delta_k$  is nonzero, we can write  $L^{(1)} = \prod_{k=0}^{2^n-1} (L_k^{(1)})^{\Delta_k}$  and  $U^{(1)} = \sum_k \Delta_k U_k^{(1)}$ .

One technique for obtaining  $U^a$  of form (5) is to project  $U^{(1)}$  on to the  $\mathcal{L}^2$ -space of functions of  $(R, D, Z)$  which are unbiased conditional on  $(D, Z)$ , i.e., the tangent space for the nuisance parameter  $\gamma$ ; see Robins et al. (2000) and references therein. We show in an Appendix that this projection is

$$E_X(U^{(1)} \mid R, D, Z) - E_{R,X}(U^{(1)} \mid D, Z) = \sum_{k=0}^{2^n-1} (\Delta_k - \epsilon_k) \psi_k^{\text{proj}}, \quad (8)$$

where

$$\psi_k^{\text{proj}} = E_X(U_k^{(1)} \mid D, Z). \quad (9)$$

Thereby we define a new estimating function

$$U^{\text{proj}} = U^{(1)} - \sum_{k=0}^{2^n-1} (\Delta_k - \epsilon_k) \psi_k^{\text{proj}} \quad (10)$$

for inferences about  $\beta$ . Since  $\psi_k^{\text{proj}}$  is a function only of  $(D, Z)$ ,  $U^{\text{proj}}$  is in the class defined by (5) and (6).

Note that  $U^{\text{proj}}$  depends on the distribution of  $X_i$  given  $Z_i$ . Consider a model for this distribution governed by a finite-dimensional parameter  $\alpha$ . Then, as in (6),  $U^{\text{proj}} = U^{\text{proj}}(\beta, \gamma, \alpha)$ . Theorem 2 elucidates the efficiency benefit in using  $U^{\text{proj}}$  over  $U^{(1)}$  for inferences about  $\beta$  and implies that the projection (8) satisfies criterion (iii) in § 3.1. In addition, a reviewer has pointed out that  $U^{\text{proj}}$  is a doubly robust estimating function for  $\beta$  (Robins et al., 2000, Lemma 1) in the sense that, for any  $(\gamma^\dagger, \alpha^\dagger)$ , if either  $\gamma^\dagger = \gamma$  or  $\alpha^\dagger = \alpha$ , then  $U^{\text{proj}}(\beta, \gamma^\dagger, \alpha^\dagger)$  is unbiased; a sketch proof is in an Appendix.

**THEOREM 2.**  *$U^{\text{proj}}(\beta, \gamma, \alpha)$  is more efficient than  $U^{(1)}(\beta, \gamma)$  in the sense that  $\mathcal{I}_\beta^{\text{proj}} - \mathcal{I}_\beta^{(1)}$  is positive semidefinite, where the  $U^{\text{proj}}$  information matrix  $\mathcal{I}_\beta^{\text{proj}}$  is*

$$\mathcal{I}_\beta^{\text{proj}} = \mathcal{I}_\beta^{(1)} E(U^{\text{proj}} U^{\text{proj} \text{ T}})^{-1} \mathcal{I}_\beta^{(1)}$$

and  $\mathcal{I}_\beta^{(1)}$  is the corresponding information matrix for  $U^{(1)}$ .

An implication of Theorem 2 is that, if one can approximate  $\psi_k^{\text{proj}}$  via a ‘working’ model for the distribution of  $(X_i|Z_i)$ , one might expect an increase in efficiency over  $U^{(1)}$ . Such a working model may or may not contain the true distribution. For example, one might incorrectly assume that  $X_i$  is independent of  $Z_i$ , or one might replace the true continuous support of  $X_i$  with a finite support. The distribution of  $X_i$  is specified in an effort to improve efficiency, but correct specification is not required to ensure consistency of  $\beta$ -inferences.

### 3.4 A practical estimator

Since it is difficult to compute the expected values (9) required for  $U^{\text{proj}}$ , requiring up to  $n$ -fold integration for each stratum, we seek an approximation to  $U^{\text{proj}}$  that is easily computed but may still result in efficiency improvements over  $U^{(1)}$ . To accomplish

this, we consider the likelihood based on the distribution of  $(D_i|R_i, Z_i)$ , marginally over  $X_i$ . Define the disease odds  $\tilde{\theta}_i = \tilde{\theta}(Z_i) = \text{pr}(D_i = 1|Z_i)/\text{pr}(D_i = 0|Z_i)$ . Using a result from Satten & Carroll (2000), we have that

$$\tilde{\theta}_i = \int \theta(x, Z_i) dF(x|D_i = 0, Z_i), \quad (11)$$

where  $F$  is the distribution of  $X_i$  given  $Z_i$  among subjects with  $D_i = 0$ . As with  $\theta_i$  in equation (2), the odds  $\tilde{\theta}_i$  can be adjusted to be conditional on  $R_i = 1$ . Let  $\tilde{\theta}_i^* = \tilde{\theta}^*(Z_i) = \text{pr}(D_i = 1|R_i = 1, Z_i)/\text{pr}(D_i = 0|R_i = 1, Z_i)$ . By analogy with (2), it can be shown that

$$\log(\tilde{\theta}_i^*) = \log(\tilde{\theta}_i) + B_i, \quad (12)$$

where  $B_i = B(Z_i; \gamma) = \log\{\pi_i(1; \gamma)/\pi_i(0; \gamma)\}$  as in equation (2).

Using (12), and by analogy with (3) and  $L_k^{(1)}$ , we may specify a likelihood for  $D_{(k)}$  conditional on  $Z, R = r_k$  and  $\sum_i D_i r_{ki}$ . Denote this likelihood by  $L_k^{(2)} = \text{pr}(D_{(k)} | \sum_i D_i r_{ki}, R = r_k, Z)$ , and its score function by  $U_k^{(2)} = \partial \log L_k^{(2)} / \partial \beta^T$ . Note that, once  $\tilde{\theta}_i$  is specified using (11) and an assumed model for  $F$ , the  $U_k^{(2)}$ 's are easily computed functions of always-observed variables  $D$  and  $Z$ . Hence, our proposal is to use

$$U^{\text{appr}} = U^{(1)} - \sum_{k=0}^{2^n-1} (\Delta_k - \epsilon_k) U_k^{(2)}$$

for estimating  $\beta$ . Note that, while  $U^{\text{appr}}$  is in the class defined by (5) and (6), it is not doubly robust.

To justify approximating the quantities  $\psi_k^{\text{proj}}$  in (9) and (10) with  $U_k^{(2)}$ , note that

$$\begin{aligned} & \text{pr}(D_{(k)}|X_{(k)}, \sum_i D_i r_{ki}, R = r_k, Z) dF_k(X_{(k)} | \sum_i D_i r_{ki}, R = r_k, Z) \\ &= dF_k(X_{(k)}|D_{(k)}, \sum_i D_i r_{ki}, R = r_k, Z) \text{pr}(D_{(k)} | \sum_i D_i r_{ki}, R = r_k, Z), \end{aligned} \quad (13)$$

where  $F_k$  is the distribution of  $X_{(k)}$ . Writing the integral of (13) over  $x_{(k)}$  as

$$\int_{x_{(k)}} L_k^{(1)} dF_k(x_{(k)} | \sum_i D_i r_{ki}, R = r_k, Z) = \int_{x_{(k)}} dF_k(x_{(k)} | D_{(k)}, R = r_k, Z) L_k^{(2)}, \quad (14)$$

we show in an Appendix that

$$U_k^{(2)} - \psi_k^{\text{proj}} = E_X \left( \frac{\partial \log dF_k(X_{(k)} | \sum_i D_i r_{ki}, R = r_k, Z)}{\partial \beta} \mid D, Z \right). \quad (15)$$

If we approximate the expectation  $E_X(\cdot | D, Z)$  on the right-hand side of (15) by  $E_X(\cdot | \sum_i D_i r_{ki}, R = r_k, Z)$ , then this quantity is zero. To the extent that this approximation holds,  $U^{\text{appr}}$  will approximate  $U^{\text{proj}}$ .

Computing the  $U_k^{(2)}$ 's, and hence  $U^{\text{appr}}$ , requires knowledge of the distribution  $F$  of  $(X_i | D_i = 0, Z_i)$ . As such, let  $F^*(x | D_i = 0, Z_i; \alpha)$  denote a parametric working model for  $F$  governed by  $\alpha$ . We propose a simple version of  $F^*$  as follows. First, create a mapping  $h(\cdot)$  of  $X_i$  into a finite set  $\mathcal{C} = \{1, \dots, C\}$ . Now posit a multinomial logistic model, parameterised by  $\alpha$ , for  $dF^*(c | D_i = 0, Z_i; \alpha) = \text{pr}\{h(X_i) = c | D_i = 0, Z_i; \alpha\}$  for  $c \in \mathcal{C}$ . Using the 'data'  $h(X_i)$  on subjects for whom  $(R_i = 1, D_i = 0)$ , define the working  $\alpha$ -score functions

$$T^\alpha = \sum_i (1 - D_i) R_i [\partial \log dF^*\{h(X_i) | D_i = 0, Z_i; \alpha\} / \partial \alpha^T].$$

Finally, solve  $\sum_s T^\alpha(\alpha) = 0$  to obtain  $\hat{\alpha}$ , and take  $F^*(\cdot | D_i = 0, Z_i; \hat{\alpha})$ , with support  $\mathcal{C}$ , to be the working distribution of  $(X_i | D_i = 0, Z_i)$  for computing  $U^{\text{appr}}$ .

## 4 Simulation Study

To illustrate the use of  $U^{\text{appr}}$ , and to compare the performance of  $U^{(1)}$  and  $U^{\text{appr}}$ , we simulate a population uniformly distributed among 200 strata ( $s = 1, \dots, 200$ ). We let  $q(s) = \{(s - 1)/199\}^2 - 4.5$ , so that by prospective model (1) a few strata will be at higher risk for  $D_i = 1$  than most. We consider two versions of (1) for the odds of disease. In both, covariate  $Z_i$  is a standard normal random variable. In the first model,  $X_i$  is Bernoulli with  $\text{logit}\{\text{pr}(X_i = 1 | Z_i)\} = \log(0.3/0.7) + 0.6Z_i$ , so that  $\text{corr}(X_i, Z_i) = 0.26$ . 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  $\text{corr}(X_i, Z_i) = 0.46$ . Since  $Z_i$  is standard normal,  $E(X_i^*) = 1$ . For both models,  $\beta_z = \log(1.5)$ . For the binary  $X_i$  model,  $\beta_x = \log(2.0)$ , while, for the censored exponential model,  $\beta_x = \log(1.3)$ . In both models, population disease prevalence is approximately 2.3%. The missingness model is  $\text{logit}\{\pi_i(D_i)\} = \gamma_0 + \gamma_1 D_i + \gamma_2 Z_i$ , where  $\gamma_0 = 1.5$ ,  $\gamma_1 = -1.0$  and  $\gamma_2 = -1.0$ , so that  $X_i$  is missing for 50% of case, and for 22% of control, subjects. For each replicate dataset, 200 case subjects, for whom  $D_i = 1$ , were randomly sampled from the population, ignoring strata. For each sampled case subject, the stratum  $s$  was then identified, and 3 matched control subjects,  $D_i = 0$ , were sampled from the same stratum. This was accomplished by sampling from stratum  $s$  until 3 subjects with  $D_i = 0$  were obtained.

Six estimators were computed for each replicate. The first two were maximum conditional likelihood estimators obtained using the method of Satten & Carroll (2000). Assumed models for the distribution of  $X_i$  among the control subjects were  $\text{logit}\{\text{pr}(X_i = 1|D_i = 0, Z_i)\} = \alpha_0 + \alpha_1 Z_i + \alpha_2 Z_i^2$  for binary  $X_i$  and  $\log\{E(X_i^*|D_i = 0, Z_i)\} = \alpha_0 + \alpha_1 Z_i + \alpha_2 Z_i^2$  for censored exponential  $X_i$ . The first maximum conditional likelihood estimator assumed that  $X_i \perp\!\!\!\perp Z_i$  by setting  $\alpha_1 = \alpha_2 = 0$ , while in the second  $(\alpha_0, \alpha_1, \alpha_2)$  were estimated simultaneously with  $\beta$ . Thirdly, we computed the naive complete-subject estimator. The fourth estimator was the solution to  $\sum_s U^{(1)}(\beta, \hat{\gamma}) = 0$  due to Lipsitz et al. (1998), where  $\gamma$  is estimated as in § 2. The last two were the solutions to two versions of  $\sum_s U^{\text{appr}}(\beta, \hat{\gamma}, \hat{\alpha}) = 0$ , one where  $X_i$  is incorrectly assumed to be independent of  $Z_i$  and one where  $X_i$  depends on  $Z_i$ . For binary  $X_i$ ,  $\text{pr}(X_i = 1|D_i = 0, Z_i)$  was modelled as a constant ( $X_i \perp\!\!\!\perp Z_i$ ), or as logistic in  $Z_i$  and  $Z_i^2$ . For exponential  $X_i$ , a working support  $\mathcal{C}$  of  $X_i$  was taken to contain three values, namely the means of the three observed tertiles of  $X_i$  among the controls. This trinomial distribution was modelled in a multinomial logistic model either independently of  $Z_i$  or as a function of  $Z_i$  and  $Z_i^2$ . In both cases, parameter  $\alpha$  in the

model for  $X_i$  was estimated using the data on the control subjects,  $D_i = 0$ , with observed  $X_i$ , as described in § 3.4. For estimators solving  $\sum_s U^{(1)} = 0$  and  $\sum_s U^{\text{appr}} = 0$ , 95% Wald-type confidence intervals were constructed using variance estimator (7).

Table 1 shows that, for binary  $X_i$ , the maximum conditional likelihood estimator is the most efficient when the distribution of  $(X_i|D_i = 0, Z_i)$  is correctly modelled ( $X_i \perp\!\!\!\perp Z_i$ ), but is biased when this distribution is misspecified ( $X_i \amalg Z_i$ ). The complete case estimator for  $\beta_z$  is clearly biased, but not that for  $\beta_x$ . Bias-correction is achieved using  $U^{(1)}$  and  $U^{\text{appr}}$ , and the efficiency for  $\beta_x$  is similar for all four of these estimators. A slight loss in efficiency in  $\hat{\beta}_x$  is incurred when we assume incorrectly that  $X_i \amalg Z_i$ . As these estimators for  $\beta_x$  are 77% efficient relative to the maximum conditional likelihood estimator, there is a price to pay for not modelling the distribution of  $X_i$ . For  $\beta_z$  estimation,  $U^{\text{appr}}$  is 61% more efficient than  $U^{(1)}$ , with only a small loss due to incorrectly assuming that  $X_i \amalg Z_i$ . That the efficiency gain in using  $U^{\text{appr}}$  occurs for  $\beta_z$  and not for  $\beta_x$  reflects the fact that  $U^{\text{appr}}$  incorporates information in  $(D_i, Z_i)$  in subjects with missing  $X_i$ . While  $U^{\text{appr}}$  is only 77% efficient compared to the maximum conditional likelihood estimator, no bias results when the distribution of  $X_i$  is misspecified. Results were similar for the population where  $X_i$  is continuous; see Table 2. For  $\beta_z$ ,  $U^{\text{appr}}$  is 39% more efficient than  $U^{(1)}$  and achieves 78% efficiency relative to the maximum conditional likelihood estimator. The maximum conditional likelihood estimator is substantially biased for misspecified distribution of  $X_i$ .

## 5 Discussion

A reviewer made several useful comments about the efficiency of  $U^{\text{proj}}$  and  $U^{\text{appr}}$ . First, greater efficiency than that of  $U^{\text{proj}}$  could be obtained by replacing the projection (8) with the projection of  $U^{(1)}$  on to the space of functions of the observed

data  $(R, X_{\text{obs}}, D, Z)$  which are unbiased conditional on the complete data  $(X, D, Z)$ . However, this projection is not easily computed and is not of closed form. Our approximate projection employing the  $U_k^{(2)}$ 's is always available in closed form. Furthermore, even if the optimal projection were available, the resulting estimator would not be semiparametric efficient because  $U^{(1)}$  is not the efficient preliminary estimating function in this problem. It is difficult to find this preliminary estimating function because of the presence of the nuisance parameters  $q(s)$ . Finally, to be certain of an efficiency improvement in  $U$  over  $U^{(1)}$  in (5) and (6) for a given set of  $\psi_k$ 's, one could use the estimating function  $U^* = U^{(1)} - C^*U^a$ , where  $C^*$  is the coefficient from the least squares regression of  $U^{(1)}$  on  $U^a$ . If  $\psi_k = \psi_k^{\text{proj}}$ , then  $C$  will converge to one; see Robins et al. (1994, 1995).

## Acknowledgment

The authors thank James Robins for insightful comments on double robustness and semiparametric efficiency, and Ronald Thisted, the editor, associate editor and referees for helpful comments that improved the presentation considerably. Rathouz's research was supported in part by a grant from the National Science Foundation. Carroll's research was supported by a grant from the National Cancer Institute and through the Texas A&M Center for Environmental and Rural Health by a grant from the National Institute of Environmental Health Sciences.

# Appendix

## Technical Details

Proofs of Theorems 1 and 2 are entirely analogous to those of Proposition 6.1(a,b,c) of Robins et al. (1994). Salient details of arguments specific to this setting are given in the sketch proofs which follow.

*Proof of Theorem 1.* The consistency of  $\hat{\beta}$  follows from standard pseudo-likelihood theory (Gong and Samaniego, 1981). For the asymptotic normality, define the following information quantities. Let  $\mathcal{I}_\beta = E(-\partial U/\partial\beta)$ . Now,  $E(-\partial U^a/\partial\beta) = E\{-\sum_{k=0}^{2^n-1}(\Delta_k - \epsilon_k)(\partial\psi_k/\partial\beta)\} = 0$ , by taking expectation over  $R$  first, conditional on  $(D, Z)$ . Therefore  $\mathcal{I}_\beta = E(-\partial U^1/\partial\beta) = \mathcal{I}_\beta^1$ . Because  $U^1$  is a likelihood score,  $\mathcal{I}_\beta^1 = E(U^1 U^{1T})$ . Note that  $T^\gamma$  is ancillary for  $\beta$ , i.e.,  $E(-\partial T^\gamma/\partial\beta) = 0$ . Let  $\mathcal{I}_\gamma = E(-\partial U/\partial\gamma)$  and  $\mathcal{I}_\gamma^\gamma = E(-\partial T^\gamma/\partial\gamma)$ . Because  $T^\gamma$  is a likelihood score for  $\gamma$ ,  $\mathcal{I}_\gamma = E(UT^\gamma T^{\gamma T})$  and  $\mathcal{I}_\gamma^\gamma = E(T^\gamma T^{\gamma T})$ . Finally,  $E(-\partial U^1/\partial\alpha) = E(0) = 0$  and  $E(-\partial U^a/\partial\alpha) = E\{-\sum_{k=0}^{2^n-1}(\Delta_k - \epsilon_k)(\partial\psi_k/\partial\alpha)\} = 0$ , by taking expectation over  $R$  first, conditional on  $(D, Z)$ . Therefore  $\mathcal{I}_\alpha = E(-\partial U/\partial\alpha) = 0$ .

Now, adding subscripts  $s$  to index strata and following the usual Taylor-series arguments,  $(\hat{\gamma} - \gamma) = (\sum_s \mathcal{I}_{\gamma,s}^\gamma)^{-1} \sum_s U_s^\gamma + o_p(1/\sqrt{J}) = O_p(1/\sqrt{J})$ , and similarly  $(\hat{\alpha} - \alpha) = O_p(1/\sqrt{J})$ . By Taylor series expansion,

$$\begin{aligned} \sum_s U_s(\beta, \hat{\gamma}, \hat{\alpha}) &= \sum_s U_s + \left\{ \sum_s (\partial U_s / \partial \gamma) \right\} (\hat{\gamma} - \gamma) + \left\{ \sum_s (\partial U_s / \partial \alpha) \right\} (\hat{\alpha} - \alpha) + o_p(\sqrt{J}) \\ &= \sum_s U_s - \left( \sum_s \mathcal{I}_{\gamma,s} \right) \left( \sum_s \mathcal{I}_{\gamma,s}^\gamma \right)^{-1} \sum_s U_s^\gamma + O_p(1) + o_p(\sqrt{J}) \\ &= \sum_s \left\{ U_s - \left( \sum_{s'} \mathcal{I}_{\gamma,s'} \right) \left( \sum_{s'} \mathcal{I}_{\gamma,s'}^\gamma \right)^{-1} U_s^\gamma \right\} + o_p(\sqrt{J}), \end{aligned}$$

which then leads to the asymptotic normality of  $\hat{\beta}$  with mean 0 and variance

$$J \left\{ - \sum_s \partial U_s(\beta, \hat{\gamma}, \hat{\alpha}) / \partial \beta \right\}^{-1} \left\{ \sum_s U_s(\beta, \hat{\gamma}, \hat{\alpha})^{\otimes 2} \right\} \left\{ - \sum_s \partial U_s(\beta, \hat{\gamma}, \hat{\alpha}) / \partial \beta \right\}^{-1T}.$$

Finally,  $-\sum_s \partial U_s(\beta, \hat{\gamma}, \hat{\alpha}) / \partial \beta = -\sum_s \partial U_s / \partial \beta + o_p(J) = \sum_s \mathcal{I}_{\beta,s} + o_p(J) = \sum_s \mathcal{I}_{\beta,s}^1 +$

$o_p(J)$  and  $\sum_s U_s(\beta, \hat{\gamma}, \hat{\alpha})^{\otimes 2} = \sum_s E(U_s U_s^T) - (\sum_s \mathcal{I}_{\gamma,s})(\sum_s \mathcal{I}_{\gamma,s}^{-1})(\sum_s \mathcal{I}_{\gamma,s})^T + o_p(J)$ .

These last two expressions yield (7) as a “sandwich estimator” of  $\mathcal{V}$  (Huber, 1967).

*Development of (8).* To derive (8), first write

$$E_X(U^{(1)} | R, D, Z) = \sum_{k=0}^{2^n-1} E_X(\Delta_k U_k^{(1)} | R, D, Z) = \sum_{k=0}^{2^n-1} \Delta_k E_X(U_k^{(1)} | R, D, Z),$$

where the latter equality is due to the fact that  $\Delta_k$  is a function of  $R$ . Now,  $E_X(U_k^{(1)} | R, D, Z)$  does not contain  $R$  because missingness is at random, so write  $E_{R,X}(U^{(1)} | D, Z) = \sum_{k=0}^{2^n-1} \epsilon_k E_X(U_k^{(1)} | D, Z)$ . Defining  $\psi_k^{\text{proj}} = E_X(U_k^{(1)} | D, Z)$ , we have (8).

*Proof that  $U^{\text{proj}}$  is doubly-robust.* First, consider the case where  $\gamma^\dagger = \gamma$ , and  $\alpha^\dagger \neq \alpha$ . Then  $E_D(U^{(1)} | R, X, Z) = 0$ , as  $U^{(1)}$  is a likelihood score for  $D_{\text{obs}}$  given  $(R, X_{\text{obs}}, Z_{\text{obs}})$ , and unbiasedness of  $U^{(1)}$  only depends on parameters  $(\beta, \gamma)$ . Also,

$$E_R\{(\Delta_k - \epsilon_k)\psi_k^{\text{proj}} | D, X, Z\} = \psi_k^{\text{proj}} E_R(\Delta_k - \epsilon_k | D, Z) = 0,$$

as  $\psi^{\text{proj}}$  does not contain  $R$ ,  $\epsilon_k$  only depends on parameter  $\gamma$ , and missingness is at random. Therefore, (10) is unbiased, completing the proof for  $\gamma^\dagger = \gamma$ .

For the case where  $\alpha^\dagger = \alpha$ , but  $\gamma^\dagger \neq \gamma$ , write (10) as

$$U^{\text{proj}} = \sum_k \Delta_k \{U_k^{(1)} - E_X(U_k^{(1)} | D, Z)\} + \sum_k \epsilon_k E_X(U_k^{(1)} | D, Z). \quad (\text{A.1})$$

Taking the first set of terms in (A.1), regardless of the value of  $\gamma^\dagger$  plugged into  $U_k^{(1)}$ ,

$$\begin{aligned} E_R \left( E_X \left[ \Delta_k \{U_k^{(1)} - E_X(U_k^{(1)} | D, Z)\} | R, D, Z \right] | D, Z \right) \\ = E_R \left[ \Delta_k \{E_X(U_k^{(1)} | R, D, Z) - E_X(U_k^{(1)} | D, Z)\} | D, Z \right] = 0. \end{aligned}$$

The fact that  $E_X(U_k^{(1)} | R, D, Z) = E_X(U_k^{(1)} | D, Z)$  follows immediately from missingness at random. For the second part of (A.1), note that  $U_k^{(1)}$  depends on  $(\beta, \gamma^\dagger)$  and will therefore not be unbiased. Proof of unbiasedness of these terms relies on a careful accounting of parameter values under which expectations are taken. We denote the relevant parameter values after the conditioning statistics in each expected value.

$$E_D \left[ \epsilon_k(\gamma^\dagger) E_X \left\{ U_k^{(1)}(\beta, \gamma^\dagger) | D, Z; \beta, \alpha \right\} | Z; \beta, \alpha \right]$$

$$\begin{aligned}
&= E_D \left[ E_{X,R} \left\{ \Delta_k U_k^{(1)}(\beta, \gamma^\dagger) | D, Z; \beta, \alpha, \gamma^\dagger \right\} | Z; \beta, \alpha \right] \\
&= E_{X,R} \left[ E_D \left\{ \Delta_k U_k^{(1)}(\beta, \gamma^\dagger) | R, X, Z; \beta, \alpha, \gamma^\dagger \right\} | Z \right] = 0.
\end{aligned}$$

The last equality follows because  $U_k^{(1)}(\beta, \gamma^\dagger)$  is a score function from the distribution of  $(D_{(k)} | R = r_{(k)}, X, Z)$ , and expectation is taken over  $D$  assuming parameters  $(\beta, \gamma^\dagger)$ .

*Proof of Theorem 2.* Assuming for ease of exposition that  $\beta$  is scalar, write (8) as  $U_{\text{proj}}^a = U^b - U^c$ , where  $U^b = E_X(U^1 | R, D, Z)$  and  $U^c = E_{R,X}(U^1 | D, Z)$ . Write  $U^{\text{proj}} = U^1 - U_{\text{proj}}^a$ . Then  $\text{var}(U_{\text{proj}}^a) = \text{var}(U^b - U^c) = \text{var}(U^b) - 2\text{cov}(U^b, U^c) + \text{var}(U^c)$ . Now  $\text{cov}(U^b, U^c) = E(U^c U^c) + E\{(U^b - U^c)U^c\} = \text{var}(U^c)$ , the last term having expected value equal to zero, by taking expectation of  $U^b$  over  $R$  first. So  $\text{var}(U_{\text{proj}}^a) = \text{var}(U^b) - \text{var}(U^c)$ . Similarly,  $\text{var}(U^{\text{proj}}) = \text{var}(U^1 - U_{\text{proj}}^a) = \text{var}(U^1) - 2\text{cov}(U^1, U_{\text{proj}}^a) + \text{var}(U_{\text{proj}}^a)$ . Then,  $\text{cov}(U^1, U_{\text{proj}}^a) = E(U^1 U^b) - E(U^1 U^c) = E(U^b U^b) - E(U^c U^c) + E\{(U^1 - U^b)U^b\} - E\{(U^1 - U^c)U^c\}$ . Again, the last two terms are zero by successive conditioning. So,  $\text{cov}(U^1, U_{\text{proj}}^a) = \text{var}(U^b) - \text{var}(U^c) = \text{var}(U_{\text{proj}}^a)$  and  $\text{var}(U^{\text{proj}}) = \text{var}(U^1) - \text{var}(U_{\text{proj}}^a)$ . Therefore, the  $\beta$  information in  $U$  is

$$E(-\partial U^1 / \partial \beta)^T E(U^1 U^{1T} - U_{\text{proj}}^a U_{\text{proj}}^{aT})^{-1} E(-\partial U^1 / \partial \beta),$$

while that in  $U^1$  is  $E(-\partial U^1 / \partial \beta)^T E(U^1 U^{1T})^{-1} E(-\partial U^1 / \partial \beta)$ . Since  $E(U_{\text{proj}}^a U_{\text{proj}}^{aT})$  is positive semi-definite, the proof is complete.

*Derivation of (15).* Denoting by LHS and RHS the left and right-hand sides of (15),

$$\begin{aligned}
\frac{\partial \log(\text{LHS})}{\partial \beta} &= E_{X_{(k)}} \left( U_k^{(1)} | D_{(k)}, R = r_k, Z \right) \\
&\quad + E_{X_{(k)}} \left( \frac{\partial \log dF_k(X_{(k)} | \sum_i D_i r_{ki}, R = r_k, Z)}{\partial \beta} | D_{(k)}, R = r_k, Z \right) \\
&= E_X \left( U_k^{(1)} | D, Z \right) + E_X \left( \frac{\partial \log dF_k(X_{(k)} | \sum_i D_i r_{ki}, R = r_k, Z)}{\partial \beta} | D, Z \right).
\end{aligned}$$

For the second equality, only  $X_{(k)}$  is inside the expected values. So taking expectation over  $X_{(k)}$  conditional on  $D_{(k)}$  is equivalent to taking it over  $X$  conditional on  $D$ . Second, due to missingness at random, conditioning on  $R$  is irrelevant. Also,

$E_X(U_k^{(1)}|D, Z) = \psi_k^{\text{proj}}$ . For the RHS of (14), we have

$$\text{RHS} = L_k^{(2)} \int_{x^{(k)}} dF_k(x^{(k)}|D^{(k)}, \sum_i D_i r_{ki}, R = r_k, Z) = L_k^{(2)},$$

so  $\partial \log(\text{RHS})/\partial \beta = U_k^{(2)}$ , which yields (15).

## References

- BRESLOW, N.E. & CAIN, K.C. (1988). Logistic regression for two-stage case-control data. *Biometrika* **75**, 11–20.
- BRESLOW, N.E. & DAY, N.E. (1980). *Statistical Methods in Cancer Research*, v.1, *The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer.
- CARROLL, R.J., RUPPERT, D. & STEFANSKI, L.A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.
- GIBBONS, L.E. & HOSMER, D.W. (1991). Conditional logistic regression with missing data. *Commun. Statist. B* **20**, 109–20.
- GODAMBE, V.P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika* **63**, 277–84.
- GONG, G. & SAMANIEGO, F.J. (1981). Pseudo maximum likelihood estimation: theory and applications. *Ann. Statist.* **9**, 861–9.
- HUBER, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the 5th Berkeley Symposium* **1**, 221–33.
- LINDSAY, B.G. (1985). Using empirical partially Bayes inference for increased efficiency. *Ann. Statist.* **13**, 914–31.
- LIPSITZ, S.R., PARZEN, M. & EWELL, M. (1998). Inference using conditional logistic regression with missing covariates. *Biometrics* **54**, 295–303.

- LITTLE, R.J.A. & RUBIN, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley and Sons.
- PAIK, M.C. & SACCO, R.L. (2000). Matched case-control data analyses with missing covariates. *Appl. Statist.* **49**, 145–56.
- ROBINS, J.M., ROTNITZKY, A. & VAN DER LAAN, M. (2000). Comment on ‘On profile likelihood,’ by Murphy, S.A. & van der Vaart, A.W. *J. Am. Statist. Assoc.* **95**, 477–82.
- ROBINS, J.M., ROTNITZKY, A. & ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–66.
- ROBINS, J.M., ROTNITZKY, A. & ZHAO, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Statist. Assoc.* **90**, 106–21.
- SATTEN, G.A. & CARROLL, R.J. (2000). Conditional and unconditional categorical regression models with missing covariates. *Biometrics* **56**, 384–8.
- SATTEN, G.A. & KUPPER, L.L. (1993). Inferences about exposure-disease association using probability of exposure information. *J. Am. Statist. Assoc.* **88**, 200–8.

Table 1. Simulation results for population model with binary  $X_i$ , based on 1000 replicates. Upper entries are for  $\beta_z$  and lower entries are for  $\beta_x$ . True values are

$$\beta_z = 0.405, \beta_x = 0.693.$$

Method	Mean( $\hat{\beta}$ )	% Bias	CV <sup>2</sup>	Cov. %
CMLE, $X \amalg Z$	0.501	23.6	4.7	–
	0.794	14.5	10.3	–
CMLE, $X \overline{\amalg} Z$	0.410	1.1	5.4	–
	0.687	-0.9	11.0	–
Complete Case	0.231	-43.0	11.7	–
	0.688	-0.7	14.3	–
$\sum_s U^{(1)} = 0$	0.408	0.6	11.3	95.7
	0.689	-0.6	14.3	95.3
$\sum_s U^{\text{appr}} = 0, X \amalg Z$	0.411	1.4	7.1	94.4
	0.688	-0.7	14.6	94.9
$\sum_s U^{\text{appr}} = 0, X \overline{\amalg} Z$	0.411	1.4	7.0	94.7
	0.690	-0.5	14.3	95.2

CV, coefficient of variation of  $\hat{\beta}$  relative to  $\beta$ .

Cov. %, coverage percent for 95% Wald-type confidence intervals.

CMLE, maximum conditional likelihood estimator.

Table 2. Simulation results for population model with censored exponential  $X_i$ , based on 1000 replicates. Upper entries are for  $\beta_z$  and lower entries are for  $\beta_x$ . True values are  $\beta_z = 0.405$ ,  $\beta_x = 0.262$ .

Method	Mean( $\hat{\beta}$ )	% Bias	CV <sup>2</sup>	Cov. %
CMLE, $X \text{ II } Z$	0.586	44.5	5.0	–
	0.436	66.2	20.3	–
CMLE, $X \text{ III } Z$	0.413	1.9	8.3	–
	0.257	-2.0	15.5	–
Complete Case	0.233	-42.5	15.0	–
	0.259	-1.3	24.1	–
$\sum_s U^{(1)} = 0$	0.410	1.1	14.9	95.4
	0.263	0.2	24.0	94.9
$\sum_s U^{\text{appr}} = 0, X \text{ II } Z$	0.412	1.6	11.2	95.6
	0.263	0.2	26.6	95.6
$\sum_s U^{\text{appr}} = 0, X \text{ III } Z$	0.411	1.4	10.7	94.9
	0.263	0.2	24.8	95.3

CV, coefficient of variation of  $\hat{\beta}$  relative to  $\beta$ .

Cov. %, coverage percent for 95% Wald-type confidence intervals.

CMLE, maximum conditional likelihood estimator.