Statistica Sinica **12**(2002), 555-574

JOINT CONDITIONAL LIKELIHOOD ESTIMATOR IN LOGISTIC REGRESSION WITH MISSING COVARIATE DATA

C. Y. Wang, J. C. Chen, S. M. Lee and S. T. Ou

Fred Hutchinson Cancer Research Center, Fu Jen Catholic University, Feng Chia University and Tamkang University

Abstract: This article considers semiparametric estimation in logistic regression with missing covariates. In a validation subsample, we assume covariates are measured without error. Some covariates are missing in the non-validation set, while surrogate variables may be available for all study subjects. We consider the case when a covariate variable is missing at random such that the selection probability of the validation set depends only on observed data. Breslow and Cain (1988) proposed a conditional likelihood approach based on the validation set. We combine the conditional likelihoods of the validation set and the non-validation set. The proposed estimator is easy to implement and is semiparametric since no additional model assumption is imposed. Large sample theory is developed. For the estimation of the parameter for the missing covariate, simulations show that, under various situations, the proposed estimator is significantly more efficient than the validation likelihood estimator of Breslow and Cain and the inverse selection probability weighted estimator. Under moderate sample sizes and moderate values of relative risk parameters, our estimator remains competitive when compared with the nonparametric maximum likelihood estimator of Scott and Wild (1997). The proposed method is illustrated by a real data example.

Key words and phrases: Conditional likelihood, errors in variable, logistic regression, two-phase design.

1. Introduction

We consider logistic regression to investigate factors related to disease incidence. The missing covariate problem arises when certain components of the covariate vector are too difficult to measure on all study subjects. As an example, in a clinical study, the missing covariate problem may arise because some medical procedures required to ascertain certain covariates are too invasive to be performed on all individuals. This problem may also be from the study design. For example, in case-cohort studies, while general covariate information such as demographics are available for all study subjects, certain covariate data such as blood samples are only assembled on a subset of the study cohort. In general, let Y be the binary response, Z be a covariate which is always observed and X be

another covariate that may be missing. Assume that W is a surrogate variable for X, such that W is independent of Y given (X, Z). Let n be the sample size. For $i = 1, \ldots, n$, consider the logistic regression model

$$\text{pr}(Y_i = 1|X_i, Z_i, W_i) = H(\beta_0 + \beta_1^t X_i + \beta_2^t Z_i), \tag{1}$$

where $H(u) = \{1+e^{-u}\}^{-1}$ is the logistic distribution function and $\beta = (\beta_0, \beta_1^t, \beta_2^t)^t$ is a vector of parameters. Let δ_i indicate whether X_i is observed ($\delta_i = 1$) or not ($\delta_i = 0$). The validation data set ($\delta_i = 1$) consists of (Y_i, X_i, Z_i, W_i), and the nonvalidation data set ($\delta_i = 0$) consists of (Y_i, Z_i, W_i). In this paper we consider the case where X_i is assumed to be missing at random (MAR, Rubin (1976)) such that the probability of X_i being observed (selection probability), $\text{pr}(\delta_i = 1|Y_i, X_i, Z_i, W_i) = \pi(Y_i, Z_i, W_i)$, depends on (Y_i, Z_i, W_i) but not on X_i. In some cases, the data are obtained in two stages. At the first stage, $(Y_i, Z_i, W_i), i = 1, \ldots, n$ are observed from all subjects, and at the second stage X_i's are measured in the validation data set. This is sometimes called a two-phase design.

When Z and W are discrete, Breslow and Cain (1988) proposed a conditional likelihood method for a two-stage case-control study such that at the second stage some X's are observed on each stratum classified by (Y, Z, W). This approach is based on a validation likelihood. When X is also discrete, Schill, Jöckel, Drescher and Timm (1993) considered a refined estimator, and a constrained maximum likelihood estimator was later developed by Breslow and Holubkov (1997). As in Breslow and Chatterjee (1999), this is the same as a nonparametric maximum likelihood estimator of Scott and Wild (1997). A mean-score estimator was proposed by Reilly and Pepe (1995) for discrete (Z, W). In this case with discrete (Z, W), the mean-score estimator is the same as the Horvitz and Thompson (1952) inverse selection probability estimator when the weights are estimated nonparametrically. Robins, Rotnitzky and Zhao (1994) proposed efficient estimation by computing an optimal score function in semiparametric models. See Carroll, Ruppert and Stefanski (1995) for a general review. When the missingness mechanism does not depend on Y, Chen and Chen (2000) proposed a general approach to the problem.

In this paper we consider discrete (Z, W). The proposed estimator is semiparametric because additional model assumptions for nuisance components, such as the selection probabilities of the validation data set or the probability density of $X|(Y, Z, W)$, are not necessary. In Section 2, we review the conditional likelihood estimator based on validation data. We note that the conditional likelihood estimator utilizes the likelihood of Y given $(X, Z, W, \delta = 1)$ only, but it does not include the contribution of the likelihood of Y given $(Z, W, \delta = 0)$. The proposed estimator is described in Section 3. The asymptotic distribution theory is given in Section 4. Covariance estimation of the proposed estimator

is provided. In Section 5, we consider a joint unconditional likelihood estimator and discuss the advantage of our proposed estimator over the unconditional one. The inverse selection weighted estimator and the nonparametric maximum likelihood (NPML) estimator of Scott and Wild (1997) are also reviewed here. In Section 6, we examine the finite sample performance of the proposed estimator. In comparison with the validation likelihood estimator and a weighted estimator, the proposed estimator is shown to be generally more efficient in estimating the effect of the missing covariate, while it remains competitive for the other parameters. Compared to the NPML estimator, the proposed estimator is rather competitive under the situation when the sample sizes and relative risk parameters are moderate. Note that the computation of the proposed estimator is similar to that of Breslow and Cain (1988) except for adding a similar term based on the non-validation likelihood, and hence it remains easy to compute. An example is presented in Section 7. Some concluding remarks are given in Section 8 and technical details for the asymptotic normal theory are provided in the Appendix.

2. Review of the Validation Likelihood Estimator

For notational simplicity, let $\mathcal{X} = (1, X^t, Z^t)^t, V = (Z^t, W^t)^t$, $\mathrm{pr}(\delta = 1|Y, V) = \pi(Y, V)$. Recall that the likelihood function (1) is correct when all covariates are observed. When X is MAR, it can be easily shown that $\mathrm{pr}(Y = 1|V, X, \delta = 1) = H[\beta^t\mathcal{X} + \ln\{\pi(1, V)/\pi(0, V)\}]$. Breslow and Cain (1988) proposed a conditional likelihood estimator of β, which solves the estimating equation

$$U_{1n} = n^{-1/2}\sum_{i=1}^{n}\delta_i\mathcal{X}_i\left[Y_i - H\{\beta^t\mathcal{X}_i + \ln\frac{\pi(1, V_i)}{\pi(0, V_i)}\}\right] = 0. \tag{2}$$

The conditional likelihood estimator is a validation likelihood estimator. When $\pi(1, V)$ and $\pi(0, V)$ are known, the estimating equation (2) is unbiased since it is likelihood-based from the validation sample, as can be shown from direct calculations. Let $H_+(X_i, V_i) = H[\beta^t\mathcal{X}_i + \ln\{\pi(1, V_i)/\pi(0, V_i)]$. Then

$$\begin{aligned}\mathrm{E}\left[\delta_i\mathcal{X}_i\{Y_i - H_+(X_i, V_i)\}\right] &= \mathrm{pr}(\delta_i = 1)\mathrm{E}\left[\delta_i\mathcal{X}_i\{Y_i - H_+(X_i, V_i)\}|\delta_i = 1\right]\\ &\quad + \mathrm{pr}(\delta_i = 0)\mathrm{E}\left[\delta_i\mathcal{X}_i\{Y_i - H_+(X_i, V_i)\}|\delta_i = 0\right]\\ &= \mathrm{pr}(\delta_i = 1)\mathrm{E}\left(\mathrm{E}\left[\delta_i\mathcal{X}_i\{Y_i - H_+(X_i, V_i)\}|X_i, V_i, \delta_i = 1\right]|\delta_i = 1\right) + 0 = 0. \end{aligned}\tag{3}$$

If the selection probabilities are not known, then $\pi(y, v)$ may be estimated by $\widehat{\pi}(y, v) = (\sum_{i=1}^{n}\delta_i I[Y_i = y, V_i = v])/(\sum_{i=1}^{n} I[Y_i = y, V_i = v])$. Define $\widehat{H}_+(X_i, V_i) = H[\beta^t\mathcal{X}_i + \ln\{\widehat{\pi}(1, V_i)/\widehat{\pi}(0, V_i)\}]$. The resulting estimating score of the conditional likelihood estimator is

$$\widehat{U}_{1n}(\beta) = n^{-1/2}\sum_{i=1}^{n}\delta_i\mathcal{X}_i\{Y_i - \widehat{H}_+(X_i, V_i)\}. \tag{4}$$

It is clear that (4) is the estimating score from the likelihood of Y given $(X, V, \delta = 1)$, but not Y given $(V, \delta = 0)$. The proposed estimator, as will be described next, is to combine both conditional likelihood functions so that higher efficiency can be achieved.

3. The Joint Conditional Likelihood Estimator

We now describe how to make use of the likelihood of $Y|V$ from the non-validation set. By Satten and Kupper (1993),

$$\mathrm{pr}(Y = 1|V) = H\{\beta_0 + \beta_2^t Z + R(V)\}, \tag{5}$$

where $R(V) = \ln\left[\mathrm{E}\{e^{\beta_1^t X}|V, Y = 0\}\right]$. By direct calculation,

$$\begin{aligned}\mathrm{pr}(Y = 1|V, \delta = 0) &= \frac{\mathrm{pr}(Y = 1|V)\mathrm{pr}(\delta = 0|Y = 1, V)}{\sum_{y=0}^{y=1} \mathrm{pr}(Y = y|V)\mathrm{pr}(\delta = 0|Y = y, V)} \\ &= H\{\beta_0 + \beta_2^t Z + R(V) + \ln\frac{\overline{\pi}(1, V)}{\overline{\pi}(0, V)}\},\end{aligned}$$

where $\overline{\pi}(y, v) = 1 - \pi(y, v)$ for $y = 0, 1$. Define $H_-(V) = H\{\beta_0 + \beta_2^t Z + R(V) + \ln\frac{\overline{\pi}(1,V)}{\overline{\pi}(0,V)}\}$, $\mathcal{T}_i = (1, \frac{\partial}{\partial\beta_1^t} R(V_i), Z_i^t)^t$. Similar to the calculations in (3), we have

$$\mathrm{E}\left[(1 - \delta_i)\mathcal{T}_i\{Y_i - H_-(V_i)\}\right] = 0. \tag{6}$$

When the selection probabilities and the conditional probabilities of X given $(Y{=}0,V)$ are known, an estimating equation combining the validation and non-validation conditional likelihoods may be obtained by solving $U_n(\beta) = 0$, where $U_n(\beta) = n^{-1/2}\sum_{i=1}^n \left[\delta_i \mathcal{X}_i\{Y_i - H_+(X_i, V_i)\} + (1 - \delta_i)\mathcal{T}_i\{Y_i - H_-(V_i)\}\right]$. The unbiasedness of the estimating score $U_n(\beta)$ can be seen directly from (3) and (6). When the selection probabilities and the conditional probabilities of X given $(Y{=}0,V)$ are not known, we propose a semiparametric estimator $\widehat{\beta}$ for β which solves the following estimating equations,

$$\widehat{U}_n(\beta) = n^{-1/2}\sum_{i=1}^n \left[\delta_i \mathcal{X}_i\{Y_i - \widehat{H}_+(X_i, V_i)\} + (1 - \delta_i)\widehat{\mathcal{T}}_i\{Y_i - \widehat{H}_-(V_i)\}\right],$$

where $\widehat{H}_-(V_i) = H[\beta_0 + \beta_2^t Z_i + \hat{R}(V_i) + \ln\{\widehat{\overline{\pi}}(1, V_i)/\widehat{\overline{\pi}}(0, V_i)\}]$, $\widehat{R}(V_i) = \ln\{(\sum_{j=1}^n \delta_j e^{\beta_1^t X_j} I[V_j = V_i, Y_j = 0])/(\sum_{j=1}^n \delta_j I[V_j = V_i, Y_j = 0])\}$, $\widehat{\mathcal{T}}_i = (1, \widehat{R}^{(1)}(V_i), Z_i^t)^t$ and $\widehat{R}^{(1)}(V_i) = (\sum_{j=1}^n \delta_j X_j e^{\beta_1^t X_j} I[V_j = V_i, Y_j = 0])/(\sum_{j=1}^n \delta_j e^{\beta_1^t X_j} I[V_j = V_i, Y_j = 0])$.

Observe that the proposed estimator $\widehat{\beta}$ is based on the joint conditional likelihoods (JCL) $\mathcal{L}^\delta(Y|X, Z, \delta = 1)\mathcal{L}^{1-\delta}(Y|V, \delta = 0)$, where $\mathcal{L}(\cdot)$ denotes the (conditional) likelihood function of a specified random variable.

4. Asymptotic Distribution Theory

To obtain $\widehat{\beta}$ we write the estimating equation $\widehat{U}_n(\beta) = \widehat{U}_{1n}(\beta) + \widehat{U}_{2n}(\beta)$, where $\widehat{U}_{1n}(\beta)$ was defined in (4) and

$$\widehat{U}_{2n}(\beta) = n^{-1/2} \sum_{i=1}^{n} (1 - \delta_i) \widehat{\mathcal{T}}_i \{Y_i - \widehat{H}_-(V_i)\}. \tag{7}$$

We assume the following:

(A1) For any $y = 0, 1$ and $v \in \text{supp}(V)$, the support of V, the selection probability $\pi(y, v) > 0$.

(A2) For any $v \in \text{supp}(V)$, $\text{E}\{\exp(\beta_1^t X) | V = v, Y = 0\}$ exists in a neighborhood of the true β.

(A3) For any $y = 0, 1$ and $v \in \text{supp}(V)$, the selection probability $\pi(y, v) < 1$.

(A4) $\text{E}\{\delta \mathcal{X} \mathcal{X}^t H_+^{(1)}(X, V) + (1 - \delta) \mathcal{T} \mathcal{T}^t H_-^{(1)}(V)\}$ is positive definite in a neighborhood of the true β, where $H_+^{(1)}(.) = H_+(.)\{1 - H_+(.)\}$ and $H_-^{(1)}(.) = H_-(.)\{1 - H_-(.)\}$.

(A5) The second derivatives of $\widehat{U}_n(\beta)$ with respect to β exist in a neighborhood of the true β almost surely. Further, in such a neighborhood, the second derivatives are bounded above by a function of (Y, X, V), whose expectation exists.

4.1. Limit distribution

We first express $\widehat{U}_{1n}(\beta)$ as the sum of independent variables. For convenience, let $S_c(Y_i, X_i, V_i) = \delta_i \mathcal{X}_i \{Y_i - H_+(X_i, V_i)\}$, $\mathcal{E}_c(Y_i, V_i) = (-1)^{Y_i} \{\delta_i - \text{pr}(Y_i, V_i)\} \{\pi(Y_i, V_i) \text{pr}(Y_i | V_i)\}^{-1} \text{E}\{\pi(Y, V) \mathcal{X} H_+^{(1)}(X, V) | V = V_i\}$. Note that S_c stands for the score from a complete case since it is the derivative of the logarithm of the conditional likelihood of Y_i given $(X_i, V_i, \delta_i = 1)$, *i.e.*, $(\partial/\partial\beta)[Y_i \ln\{H_+(X_i, V_i)\} + (1 - Y_i) \ln\{1 - H_+(X_i, V_i)\}]$. Also, $\mathcal{E}_c$ stands for the approximation error from the complete data score S_c due to the estimation of nuisance parameters (selection probabilities).

Lemma 1. *Under Conditions* (A1)-(A2), $\widehat{U}_{1n}(\beta) = n^{-1/2} \sum_{i=1}^{n} \{S_c(Y_i, X_i, V_i) + \mathcal{E}_c(Y_i, V_i)\} + o_p(1)$.

The proof of Lemma 1 is given in the Appendix. We note that an alternative but equivalent linearization of $\widehat{U}_{1n}$ was given in Wang and Wang (1997, p.1107).

We now linearize $\widehat{U}_{2n}(\beta)$. Let $r(V_i) = \mathrm{E}\{\exp(\beta_1^t X_i)|Y_i = 0, V_i\}$ and $\widehat{r}(V_i) = (\sum_{j=1}^n \delta_j e^{\beta_1^t X_j} I[Y_j = 0, V_j = V_i])/(\sum_{j=1}^n \delta_j I[Y_j = 0, V_j = V_i])$. Define $S_m(Y_i, V_i) = (1 - \delta_i)\mathcal{T}_i\{Y_i - H_-(V_i)\}$,

$$\begin{aligned}
&\mathcal{E}_m(Y_i, X_i, V_i)\\
&= (-1)^{(1-Y_i)}\{\delta_i - \pi(Y_i, V_i)\}\{\overline{\pi}(Y_i, V_i)\mathrm{pr}(Y_i|V_i)\}^{-1}\mathcal{T}_i H_-^{(1)}(V_i)\mathrm{E}[\{\overline{\pi}(Y, V)\}|V = V_i]\\
&\quad -I[Y_i = 0, \delta_i = 1]\{e^{\beta_1^t X_i} - r(V_i)\}\{r(V_i)\}^{-1}\mathcal{T}_i H_-^{(1)}(V_i)\frac{\mathrm{pr}(\delta_i = 0, V_i)}{\mathrm{pr}(\delta_i = 1, Y_i = 0, V_i)}.
\end{aligned}$$

Here S_m stands for the score from a case with missing covariate, while $\mathcal{E}_m$ stands for the approximation error from the incomplete data score S_m due to the estimation of nuisance parameters (selection probabilities and $R(V)$).

Lemma 2. *Under Conditions* (A2)-(A3), $\widehat{U}_{2n}(\beta) = n^{-1/2}\sum_{i=1}^n\{S_m(Y_i, V_i) + \mathcal{E}_m(Y_i, X_i, V_i)\} + o_p(1)$.

The proof of Lemma 2 is given in the Appendix.

We now present the limiting distribution for the proposed JCL estimator. For any vector a, define $a^{\otimes 2} = aa^t$.

Theorem 1. *Let $\widehat{\beta}$ be the joint conditional likelihood estimator solving $\widehat{U}_n(\beta) = 0$. Under Conditions* (A1)-(A5), *$\widehat{\beta}$ is consistent and unique in a neighborhood of β with probability converging to 1 as $n \to \infty$. Furthermore, $n^{1/2}(\widehat{\beta} - \beta)$ has an asymptotically normal distribution with mean 0 and covariance matrix $G^{-1}(\beta)M(\beta)\{G^{-1}(\beta)\}^t$, where*

$$\begin{aligned}
G(\beta) &= \mathrm{E}\{\delta\mathcal{X}\mathcal{X}^t H_+^{(1)}(X, V) + (1 - \delta)\mathcal{T}\mathcal{T}^t H_-^{(1)}(V)\};\\
M(\beta) &= \mathrm{E}\Big[\{S_c(Y, X, V) + \mathcal{E}_c(Y, V) + S_m(Y, V) + \mathcal{E}_m(Y, X, V)\}^{\otimes 2}\Big].
\end{aligned}$$

The proof of Theorem 1 is given in the Appendix. Note that the asymptotic covariance matrix is a sandwich type because of the linearization involved in $\widehat{U}_n(\beta)$, in which $\pi(y, v)$, $R(v)$ and $R^{(1)}(v)$ are estimated nonparametrically. If these nuisance components were known, then, instead of $\widehat{U}_n(\beta)$, one should solve $U_n(\beta) = 0$ for the joint conditional likelihood estimator. In this case, $\mathcal{E}_c(Y, V)$ and $\mathcal{E}_m(Y, X, V)$ would not be in $M(\beta)$. Under this generally impractical case, the asymptotic covariance can further be simplified to $G^{-1}(\beta)$.

4.2. Covariance estimation

We now present consistent estimation of the standard error of $\widehat{\beta}$. To simplify matters, write $H_+(X, V, \beta)$ for $H_+(X, V)$ (and similarly for $H_-(\cdot)$, $\mathcal{T}(\cdot)$, $S_c(\cdot)$, $\mathcal{E}_c(\cdot)$, $S_m(\cdot)$ and $\mathcal{E}_m(\cdot)$) to emphasize dependence on β. Let $\widehat{G}(\beta) =$

$n^{-1}\sum_{i=1}^n[\delta_i\mathcal{X}_i\mathcal{X}_i^t\widehat{H}_+^{(1)}(X_i,V_i,\beta)+(1-\delta_i)\widehat{\mathcal{T}}_i(\beta)\widehat{\mathcal{T}}_i^t(\beta)\widehat{H}_-^{(1)}(V_i,\beta)]$. Then it can be shown that $\widehat{G}(\widehat{\beta})\to G(\beta)$ in probability. Let

$$\widehat{M}(\beta)=n^{-1}\sum_{i=1}^n\{S_c(Y_i,X_i,V_i,\beta)+\widehat{\mathcal{E}}_c(Y_i,V_i,\beta)+S_m(Y_i,V_i,\beta)+\widehat{\mathcal{E}}_m(Y_i,X_i,V_i,\beta)\}^{\otimes 2},$$

where $\widehat{\mathcal{E}}_c(Y_i,V_i,\beta)=(-1)^{Y_i}\{\delta_i-\widehat{\pi}(Y_i,V_i)\}\{\sum_{j=1}^n\delta_j\mathcal{X}_j\widehat{H}_+^{(1)}(X_j,V_j,\beta)I[V_j=V_i]\}/(\sum_{j=1}^n\delta_jI[Y_j=Y_i,V_j=V_i])$ and

$$\begin{aligned}&\widehat{\mathcal{E}}_m(Y_i,X_i,V_i,\beta)\\&=(-1)^{(1-Y_i)}\frac{\delta_i-\widehat{\pi}(Y_i,V_i)}{\widehat{\overline{\pi}}(Y_i,V_i)}\widehat{\mathcal{T}}_i(\beta)\widehat{H}_-^{(1)}(V_i,\beta)\{\widehat{\overline{\pi}}(Y_i,V_i)\}^{-1}\frac{\sum_{j=1}^n(1-\delta_j)I[V_j=V_i]}{\sum_{j=1}^nI[Y_j=Y_i,V_j=V_i]}\\&-\delta_iI[Y_i=0]\{e^{\beta_1^tX_i}-r(V_i,\beta)\}\{r(V_i)\}^{-1}\widehat{\mathcal{T}}_i(\beta)\widehat{H}_-^{(1)}(V_i,\beta)\frac{\sum_{j=1}^n(1-\delta_j)I[V_j=V_i]}{\sum_{j=1}^n\delta_jI[Y_j=0,V_j=V_i]}.\end{aligned}$$

Then it can be shown that $\widehat{M}(\widehat{\beta})\to M(\beta)$ in probability. As a result, the covariance estimates for $n^{1/2}(\widehat{\beta}-\beta)$ can be consistently estimated by $\widehat{G}^{-1}(\widehat{\beta})\widehat{M}(\widehat{\beta})\{\widehat{G}^{-1}(\widehat{\beta})\}^t$.

5. Some Other Semiparametric Estimators

Observe that $\mathcal{L}$(observed data) $=\mathcal{L}_1\mathcal{L}_2$, where $\mathcal{L}_1=\{\mathcal{L}(Y|X,Z,W,\delta=1)\}^\delta\{\mathcal{L}(Y|Z,W,\delta=0)\}^{1-\delta}$; $\mathcal{L}_2=\{\mathcal{L}(X,Z,W,\delta=1)\}^\delta\{\mathcal{L}(Z,W,\delta=0)\}^{1-\delta}$. Hence, the proposed estimator is a semiparametric estimator using the partial likelihood $\mathcal{L}_1$. We note that $\mathcal{L}_2$ may be related to β and hence this leads to the requirement of providing the consistency result of the JCL estimator in the last two sections. Next, we describe a slightly different estimator which is based on *joint unconditional likelihoods*.

5.1. Joint unconditional likelihood estimator

Because the distribution of Y given V can be computed from (5), one intuitive approach is to consider estimation based on the partial likelihood $\{\mathcal{L}(Y|X,V)\}^\delta\{\mathcal{L}(Y|V)\}^{1-\delta}$. The resulting estimating equation for this joint unconditional likelihood estimator is

$$\sum_{i=1}^n\Big(\delta_i\mathcal{X}_i\{Y_i-H(\beta^t\mathcal{X}_i)\}+(1-\delta_i)\widehat{\mathcal{T}}_i[Y_i-H\{\beta_0+\beta_2^tZ_i+\widehat{R}(V_i)\}]\Big)=0. \quad (8)$$

This is different from our proposed estimator in that it is unconditional on the selection indicator δ. To understand the estimator, we rewrite the full likelihood

as

$$\begin{aligned}&\{\mathcal{L}(\delta=1,Y,V,X)\}^{\delta}\{\mathcal{L}(\delta=0,Y,V)\}^{1-\delta}\\ &=\{\mathcal{L}(\delta=1|Y,V,X)\}^{\delta}\{\mathcal{L}(Y|V,X)\}^{\delta}\{\mathcal{L}(V,X)\}^{\delta}\{\mathcal{L}(\delta=0|Y,V)\}^{1-\delta}\{\mathcal{L}(Y|V)\}^{1-\delta}\\ &\quad\{\mathcal{L}(V)\}^{1-\delta}.\end{aligned}$$

Therefore, under MAR the joint unconditional likelihood estimator is consistent if (i) $\mathcal{L}(\delta|Y,V)$ does not depend on β; and (ii) $\mathcal{L}(V,X)$ and $\mathcal{L}(V)$ do not depend on β. Generally, (i) seems to be a reasonable assumption under MAR. However, (ii) may be violated. For example, as in Wang, Wang and Carroll (1997), if $X|(V,Y=0)$ is normal with variance σ^2, then $X|(V,Y=1)$ is still normal with the same variance but $\mathrm{E}(X|V,Y=1)=\mathrm{E}(X|V,Y=0)+\beta_1^2\sigma^2$. Furthermore, direct calculation of the expectation of the joint unconditional likelihood estimating function (8) may not be zero in general. We also see this bias problem in the simulation study.

5.2. Weighted estimator and mean-score estimator

Inverse selection probability weighted estimation has gained much attention recently. In the spirit of Horvitz and Thompson (1952), the simplest weighted estimator uses subjects in the validation set and applies $\{\pi(Y_i,V_i)\}^{-1}$ as the weight for subject i. As in Robins, et al. (1994), Wang, et al. (1997), using estimated $\pi(Y_i,V_i)$ in general leads to a more efficient estimator of β compared to that using the true selection probabilities. The estimating equation for the weighted estimator is

$$n^{-1/2}\sum_{i=1}^{n}\frac{\delta_i}{\widehat{\pi}(Y_i,V_i)}\mathcal{X}_i\{Y_i-H(\beta^t\mathcal{X}_i)\}=0.$$

Let $\phi_i=\mathcal{X}_i\{Y_i-H(\beta_0+\beta_1^tX_i+\beta_2^tZ_i)$ and $\phi_i^*=E(\phi_i|Y_i,V_i)$. For discrete (Y_i,V_i), this weighted estimator can be seen to be equivalent to the mean-score estimator of Reilly and Pepe (1995), which solves $n^{-1/2}\sum_{i=1}^{n}\{\delta_i\phi_i+(1-\delta_i)\widehat{\phi}_i^*\}=0$, where $\widehat{\phi}_i^*=\{\sum_{j=1}^{n}\delta_j\phi_jI[Y_j=Y_i,V_j=V_i]\}/\{\sum_{j=1}^{n}\delta_jI[Y_j=Y_i,V_j=V_i]$. Both estimators share the property that for discrete (Y_i,V_i), if X_i is not observed, then the average score $\widehat{\phi}_i^*$ obtained from the validation set is used for the estimation of β.

The covariance for the weighted estimator may be obtained by using a simple sandwich estimator. Let $\mathcal{G}=n^{-1}\sum_{i=1}^{n}(\delta_i/\widehat{\pi}_i)\mathcal{X}_i\mathcal{X}_i^tH^{(1)}(\beta_0+\beta_1^tX_i+\beta_2^tZ_i)$. Then the covariance of the weighted estimator can be consistently estimated by

$$\mathcal{G}^{-1}\left[\sum_{i=1}^{n}\{\frac{\delta_i}{\widehat{\pi}(Y_i,V_i)}\phi_i+(1-\frac{\delta_i}{\widehat{\pi}(Y_i,V_i)})\widehat{\phi}_i^*\}\{\frac{\delta_i}{\widehat{\pi}(Y_i,V_i)}\phi_i+(1-\frac{\delta_i}{\widehat{\pi}(Y_i,V_i)})\widehat{\phi}_i^*\}^t\right]\mathcal{G}^{-1}.$$

5.3. Nonparametric maximum likelihood estimator

As described in Section 2, the validation likelihood estimator of Breslow and Cain (1988) applies $\sum_{i=1}^{n} \delta_i I[Y_i = y, V_i = v]$ to estimate $\pi(y, v)$ in the calculation of the offset $\ln\{\pi(1, V)/\pi(0, V)\}$. The NPML estimator of Scott and Wild (1997) considers the NPML estimator in the sense that the marginal distribution of (X, V) is not specified. As in Breslow and Chatterjee (1999), the NPML estimator is the same as the constrained ML estimator of Breslow and Holubkov (1997). Suppose V contains B strata and, for notational simplicity, let $V = j$ if V is at the jth stratum. Let γ_j be the logarithm of the odds ratio when $V = j$. That is, γ_j satisfies $\mathrm{pr}(Y_i = 1|V_i = j) = H(\gamma_j)$. Define $\xi_j = \xi_j(\gamma_j) = \ln[\{m_{1,j} - n_{1,j} + n_{\cdot,j}H(\gamma_j)\}/\{m_{0,j} - n_{0,j} + n_{\cdot,j}\overline{H}(\gamma_j)\}] - \gamma_j$, where $m_{y,j} = \sum_{i=1}^{n} \delta_i I[Y_i = y, V_i = j]$, $n_{y,j} = \sum_{i=1}^{n} I[Y_i = y, V_i = j]$ and $n_{\cdot,j} = n_{0,j} + n_{1,j}$. For given β, the NPML estimator of γ_j solves

$$\sum_{i=1}^{n}\Big[I[Y_i{=}1, V_i{=}j]{-}I[V_i{=}j]H(\gamma_j){-}\delta_i I[V_i{=}j]\{Y_i{-}H(\beta_0{+}\beta_x^t X_i{+}\beta_2^t Z_i{+}\xi_j)\}\Big]{=}0.$$

For given $\gamma_j, j = 1, \ldots, B$, the NPML estimator of β solves $\sum_{i=1}^{n} \delta_i \mathcal{X}_i\{Y_i - H(\beta^t \mathcal{X}_i + \xi_{V_i})\} = 0$. In the above equation, we note that $\xi_{V_i} = \xi_j$ if $V_i = j$. Scott and Wild used an iterative algorithm. Note that β and $\gamma_j, j = 1, \ldots, B$ are finite dimensional. We solve the joint estimating equations simultaneously by the standard Newton-Raphson algorithm.

6. Simulation Study

We evaluated small sample properties of the proposed joint conditional likelihood estimator ($\widehat{\beta}_{JCL}$). We compared it with a naive complete-case analysis ($\widehat{\beta}_{CC}$), inverse-selection probability weighted estimator ($\widehat{\beta}_{IPW}$), the conditional validation likelihood estimator of Breslow and Cain ($\widehat{\beta}_{VL}$), the joint unconditional likelihood estimator ($\widehat{\beta}_{JUL}$) and the NPML estimator of Scott and Wild ($\widehat{\beta}_{NPML}$).

Here are the simulation results. In the tables, "bias" was calculated by taking the average of $\widehat{\beta} - \beta$ from replicates, "SD" denotes the sample standard deviation of the estimators, "ASE" denotes the average of the estimated standard errors of the estimators. We have also calculated the 95% confidence interval coverage probabilities (CP). The standard errors for the proposed JCL estimator were obtained from the sandwich estimator described in Section 4. The standard errors for the weighted estimator, the validation likelihood estimator and the joint unconditional likelihood estimator were from sandwich estimates which have also taken into consideration the estimation of the nuisance parameters. Bootstrap resampling 50 times was used to estimate the standard errors of the NPML estimates.

Table 1. Simulation study with univariate covariate.

		Estimates						Relative efficiency		
		$\widehat{\beta}_{CC}$	$\widehat{\beta}_{IPW}$	$\widehat{\beta}_{VL}$	$\widehat{\beta}_{JCL}$	$\widehat{\beta}_{JUL}$	$\widehat{\beta}_{NPML}$	RE1	RE2	RE3
$n = 200$										
β_0	Bias	−0.643	−0.013	−0.016	−0.007	−0.007	−0.009			
	SD	0.315	0.170	0.196	0.167	0.168	0.167	1.04	1.38	1.00
	ASE	0.309	0.166	0.200	0.165	0.164	0.219			
	CP	0.424	0.953	0.962	0.953	0.951	0.964			
β_1	Bias	0.233	0.045	0.038	0.024	0.061	0.029			
	SD	0.548	0.393	0.404	0.326	0.341	0.329	1.45	1.54	1.02
	ASE	0.534	0.351	0.406	0.320	0.320	0.413			
	CP	0.944	0.921	0.960	0.947	0.939	0.968			
$n = 500$										
β_0	Bias	−0.633	−0.007	−0.010	−0.005	−0.004	−0.005			
	SD	0.194	0.106	0.118	0.103	0.104	0.103	1.06	1.31	1.00
	ASE	0.192	0.105	0.119	0.103	0.103	0.105			
	CP	0.061	0.954	0.961	0.957	0.957	0.947			
β_1	Bias	0.205	0.013	0.019	0.009	0.043	0.008			
	SD	0.345	0.245	0.251	0.199	0.208	0.198	1.52	1.59	0.99
	ASE	0.334	0.230	0.244	0.200	0.199	0.205			
	CP	0.920	0.930	0.947	0.955	0.946	0.952			

NOTE: The logit of pr$(Y|X)$ had parameters $\beta = (\beta_0, \beta_1)^t = (-\ln(2), \ln(3))^t$ and the selection probability for the validation set was $\{1 + \exp(0.5 + Y - W)\}^{-1}$. Covariate X had a uniform [-1,1] distribution. There were 2000 replicates with an average of 58% missing X. Relative efficiencies were defined by RE1 = var($\widehat{\beta}_{IPW}$'s)/var($\widehat{\beta}_{JCL}$'s), RE2 = var($\widehat{\beta}_{VL}$'s)/var($\widehat{\beta}_{JCL}$'s), RE3 = var($\widehat{\beta}_{NPML}$'s)/var($\widehat{\beta}_{JCL}$'s). ASE is the average of the estimated standard errors, CP is the coverage probability of the 95% confidence intervals.

In Table 1 we consider a univariate covariate, with the logit of pr$(Y|X)$ linear in parameters $\beta = (-\ln(2), \ln(3))$. We generated $X_i, i = 1, \ldots, n$ from a uniform [-1,1] distribution. The surrogate variable W was then generated by $I[X > 0]$, where $I[\cdot]$ is an indicator function. The selection probability of the validation sample follows pr$(\delta = 1|Y, W) = \{1 + \exp(0.5 + Y - W)\}^{-1}$. On average, there were 58% missing X. Total sample sizes were $n = 200$ and $n = 500$, respectively. Relative efficiency comparisons were included. The relative efficiencies were defined by RE1 = var($\widehat{\beta}_{IPW}$'s) / var($\widehat{\beta}_{JCL}$'s), RE2 = var($\widehat{\beta}_{VL}$'s) / var($\widehat{\beta}_{JCL}$'s), and RE3 = var($\widehat{\beta}_{NPML}$'s) / var($\widehat{\beta}_{JCL}$'s). We do not calculate the efficiency comparison to the joint unconditional likelihood estimator because it may be inconsistent when the distribution of $X|(Y, V)$ depends on β; this will be demonstrated later. It is seen that the proposed estimator is more efficient than the weighted estimator and the validation likelihood estimator for the estimation

of β_1. For the estimation of β_0, the proposed estimator was more efficient than the validation likelihood estimator, and was about as efficient as the weighted estimator. The complete-case analysis had serious bias because the selection of the validation sample depends on the binary outcome variables. The JCL estimator is in general slightly more efficient than the JUL estimator. In this setup, the JCL estimator performs slightly better than the NPML estimator when $n = 200$, and is almost as efficient when $n = 500$.

Table 2. Simulation study when distribution of $X|V$ depends on β.

		Estimates						Relative efficiency		
		$\widehat{\beta}_{CC}$	$\widehat{\beta}_{IPW}$	$\widehat{\beta}_{VL}$	$\widehat{\beta}_{JCL}$	$\widehat{\beta}_{JUL}$	$\widehat{\beta}_{NPML}$	RE1	RE2	RE3
$n = 200$										
β_0	Bias	−0.428	−0.035	−0.032	−0.001	−0.126	−0.032			
	SD	0.294	0.255	0.258	0.245	0.321	0.248	1.08	1.11	1.03
	ASE	0.273	0.238	0.243	0.234	0.251	0.265			
	CP	0.703	0.938	0.940	0.940	0.904	0.954			
β_1	Bias	0.084	0.041	0.036	0.012	0.238	0.038			
	SD	0.238	0.243	0.232	0.222	0.312	0.229	1.19	1.09	1.06
	ASE	0.232	0.229	0.225	0.218	0.263	0.229			
	CP	0.954	0.942	0.949	0.947	0.892	0.954			
$n = 500$										
β_0	Bias	−0.404	−0.011	−0.010	0.003	−0.145	−0.010			
	SD	0.176	0.153	0.155	0.152	0.213	0.150	1.02	1.05	0.98
	ASE	0.173	0.148	0.150	0.149	0.164	0.157			
	CP	0.357	0.940	0.940	0.942	0.839	0.960			
β_1	Bias	0.060	0.015	0.014	0.003	0.254	0.014			
	SD	0.145	0.147	0.141	0.139	0.209	0.139	1.12	1.04	1.01
	ASE	0.144	0.143	0.140	0.138	0.171	0.140			
	CP	0.948	0.944	0.946	0.949	0.712	0.950			

NOTE: The logit of $\mathrm{pr}(Y|X)$ had parameters $\beta = (\beta_0, \beta_1)^t = (-\ln(2), \ln(3))^t$ and the selection probability for the validation set was $\{1 + \exp(-0.5 + Y - W)\}^{-1}$. Covariate $X_i|(W_i, Y_i = 0)$ was normal with mean $-0.5 + W_i$ and variance 0.25; $X_i|(W_i, Y_i = 1)$ was normal with mean $-0.5 + W_i + 0.5\beta_1$ and variance 0.25; There were 2000 replicates with an average of 38% missing X.

In Table 2, surrogates were first generated randomly such that $\mathrm{pr}(W_i = 1) = 0.5$. We assumed that $X|(W, Y = 0)$ was normally distributed with mean -0.5+W, variance 0.25. Then outcomes Y's were generated using (5). Finally covariates X's were generated by noting that $X|(W, Y)$ has mean $-0.5 + W + 0.5\beta_1 Y$ and variance 0.25. This setup may be more practical for case-control studies in which data are from retrospective sampling. See also Wang, Wang and

Carroll (1997) for details about the relationship between $Y|(X,V)$ and $X|(Y,V)$. As discussed in Section 5.1, under this situation the joint unconditional likelihood estimator may be inconsistent. It is then seen that $\widehat{\beta}_{JUL}$ has a bias problem. The comparisons between other estimators are similar but in this case the efficiency gain over the weighted estimator and the validation likelihood estimator is not as significant as that in Table 1. We note that from Table 1, $\widehat{\beta}_{IPW}$ was generally better than $\widehat{\beta}_{VL}$, but this is not the case for Table 2.

Table 3. Simulation study with bivariate covariate.

		Estimates						Relative efficiency		
		$\widehat{\beta}_{CC}$	$\widehat{\beta}_{IPW}$	$\widehat{\beta}_{VL}$	$\widehat{\beta}_{JCL}$	$\widehat{\beta}_{JUL}$	$\widehat{\beta}_{NPML}$	RE1	RE2	RE3
$n=300$										
β_0	Bias	−0.597	−0.007	−0.015	0.001	0.002	−0.004			
	SD	0.307	0.193	0.199	0.191	0.192	0.192	1.02	1.09	1.01
	ASE	0.310	0.188	0.201	0.187	0.187	0.200			
	CP	0.511	0.942	0.955	0.944	0.944	0.945			
β_1	Bias	0.178	0.065	0.070	0.033	0.067	0.038			
	SD	0.437	0.321	0.309	0.254	0.264	0.256	1.60	1.48	1.02
	ASE	0.408	0.286	0.299	0.259	0.259	0.279			
	CP	0.944	0.941	0.955	0.975	0.961	0.983			
β_2	Bias	−0.067	0.009	0.013	0.009	0.015	0.008			
	SD	0.474	0.276	0.280	0.275	0.277	0.272	1.01	1.04	0.98
	ASE	0.443	0.264	0.278	0.264	0.266	0.275			
	CP	0.932	0.931	0.953	0.944	0.938	0.943			
$n=600$										
β_0	Bias	−0.580	−0.009	−0.009	−0.007	−0.007	−0.008			
	SD	0.219	0.134	0.138	0.133	0.134	0.133	1.02	1.08	1.00
	ASE	0.213	0.132	0.137	0.131	0.131	0.130			
	CP	0.206	0.953	0.953	0.949	0.948	0.948			
β_1	Bias	0.106	0.008	0.004	−0.002	0.037	0.000			
	SD	0.287	0.212	0.205	0.180	0.188	0.179	1.39	1.30	0.99
	ASE	0.281	0.203	0.204	0.180	0.181	0.179			
	CP	0.947	0.948	0.962	0.963	0.951	0.950			
β_2	Bias	−0.105	0.005	0.003	0.006	0.012	0.004			
	SD	0.322	0.185	0.190	0.186	0.187	0.184	0.99	1.04	0.98
	ASE	0.306	0.184	0.190	0.185	0.186	0.184			
	CP	0.926	0.952	0.944	0.945	0.945	0.942			

NOTE: The logit of $\mathrm{pr}(Y|X,Z)$ had parameters $\beta=(\beta_0,\beta_1,\beta_2)^t=(-\ln(2),\ln(3),\ln(3))^t$ and the selection probability for the validation set was $\{1+\exp(Y+0.5Z-0.5W)\}^{-1}$. There were 1000 replicates with an average of 61% missing X.

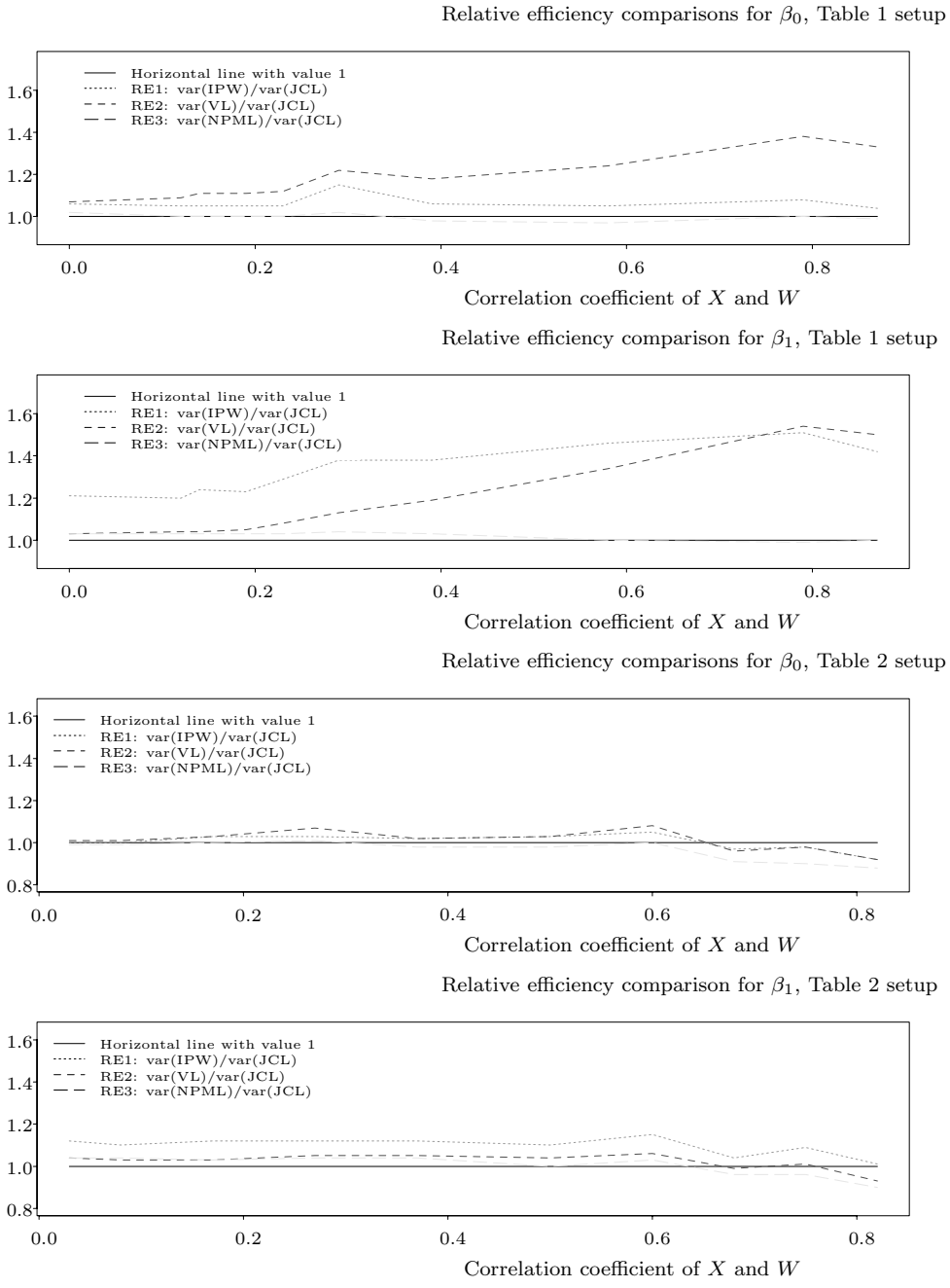

Figure 1. Relative Efficiency Comparisons with Various Correlation Coefficients of $(X, W), n = 500$.

In Table 1, the correlation coefficient between X and W, denoted by cor(X,

W), was about 0.87. We also investigated the efficiency comparison under various $\mathrm{cor}(X, W)$ values. The top portion of Figure 1 presents the result when $n = 500$. It was seen that when W provides good information, our proposed estimator is preferred to the weighted estimator and the validation likelihood estimator; while almost as good as the NPML estimator. When $\mathrm{cor}(X, W)$ was small, including 0 as an extreme case, the proposed estimator was still slightly better. However, analytical efficiency comparison with other estimators does not appear to have a simple form. The bottom portion of Figure 1 considers the setup of Table 2 in which the conditional distribution of $X|(W, Y = 1)$ depends on β_1. The distribution of $X|(W, Y)$ has mean $-0.5+\gamma W+0.5\beta_1$, variance 0.25. We selected various γ values so that $\mathrm{cov}(X, W)$ ranged from 0.03 to 0.82. Interestingly, under this setup, the JCL estimator may not be as efficient as the validation likelihood estimator if $\mathrm{cov}(X, W)$ is as large as 0.82. Nevertheless, the proposed estimator, although inferior under some situations, performs well in general. The proposed estimator may be slightly better than the NPML estimator under some situations when $n = 500$.

Table 3 considers the case with bivariate covariates, and the logit of $\mathrm{pr}(Y|X, Z)$ is linear with parameters $\beta = (\ln(2), \ln(3), \ln(3))$. Covariates X_i was generated from a uniform [-1,1] distribution and binary Z_i was independent of X_i such that $\mathrm{pr}(Z_i = 1) = \mathrm{pr}(Z_i = 0) = 0.5$. Surrogate variables W_i were generated by $I[X_i > 0]$. The selection probability of the validation sample follows $\mathrm{pr}(\delta = 1|Y, W, Z) = \{1 + \exp(Y + 0.5Z - 0.5W)\}^{-1}$. On average, there were 61% missing X. Total sample sizes were $n = 300$ and $n = 600$, respectively. Similarly, for the estimation of β_1, the proposed estimator is more efficient than the weighted estimator and the validation likelihood estimator; but slightly less efficient than the NPML estimator. Overall, it performs quite well.

In summary, in the standard (prospective) logistic regression setup when the distribution of covariates does not depend on β, the proposed JCL estimator is more efficient than the weighted estimator and the validation likelihood estimator for the estimation of the parameter for the missing covariate X. The efficiency gain is less in the retrospective case-control data setup, but in general the proposed estimator still performs well. Numerical comparisons with the efficient NPML estimator show that under various situations the proposed JCL estimator is quite efficient for moderate sample sizes.

7. Example

We consider a case-control study of bladder cancer conducted at the Fred Hutchinson Cancer Research Center. This population based case-control study

was designed to evaluate the association between bladder cancer and some nutrient intakes (Bruemmer, White, Vaughan and Cheney (1996)). In this study, eligible subjects were residents of 3 counties of western Washington state who were diagnosed between January 1987 and June 1990 with invasive or noninvasive bladder cancer. In this section, we consider 498 current and past smokers to demonstrate the methodology. The response variable is bladder cancer history (Y). There were 215 cases. We are interested in the covariates smoking packet year (X) and obesity indicator (Z). Approximately (US NIH Consensus Development Conference Statement, 1995), we define the obesity indicator as 1 if a subject's body mass index (weight/height^2) is above the 85 percentile (29.6 Kg/m^2) of the study samples. The smoking packet year of a participant is defined as the average number of cigarette packets smoked per day multiplied by the years one has been smoking. In this study, smoking packet year is not available for some study subjects since they did not respond to the question of the number of cigarettes smoked per day. A surrogate for X is $W = 1$ if the smoking year is larger than the median year (32), 0 otherwise. The majority of subjects provided the years they had smoked. There were 178 subjects in the validation set.

We first examine the missingness mechanism. We ran a logistic regression with outcome $\delta_i, i = 1, \ldots, n = 498$, covariates (Y, Z, W, ZW). The parameter estimates for these 4 factors were (1.754, 0.175, 2.027, -0.609), with s.e.s (0.229, 0.449, 0.255, 0.623). It is seen that the missingness mechanism depends significantly on Y and W.

Table 4. Analysis of bladder cancer study data.

	$\widehat{\beta}_{CC}$	$\widehat{\beta}_{IPW}$	$\widehat{\beta}_{VL}$	$\widehat{\beta}_{JCL}$	$\widehat{\beta}_{JUL}$	$\widehat{\beta}_{NPML}$
β_0	0.486	-0.686	-0.598	-0.504	-0.537	-0.651
(s.e.)	(0.221)	(0.120)	(0.187)	(0.112)	(0.114)	(0.137)
β_1	0.272	0.848	0.570	0.457	0.516	0.612
(s.e.)	(0.205)	(0.201)	(0.227)	(0.144)	(0.136)	(0.161)
β_2	0.694	0.665	0.592	0.650	0.658	0.710
(s.e.)	(0.546)	(0.298)	(0.351)	(0.280)	(0.282)	(0.391)

Note: Parameters β_1 and β_2 are the coefficients for smoking packet year and obesity indicator, respectively.

Table 4 demonstrates the results from various methods. Recall that the missingness mechanism depends significantly on (Y, W) and hence the CC analysis

may have a bias problem, along with lack of efficiency. Interestingly, except for the CC analysis, the other 5 estimators show significant effects of smoking packet year and obesity on bladder cancer incidence (the obesity effect from the validation likelihood is on the boundary). The difference between the CC analysis and the other methods is primarily because the missingness depends on the outcome variable Y. Consistent with most of the simulation results, the proposed JCL estimator is significantly more efficient than the validation likelihood and the weighted estimators for the estimation of the effect of smoking packet year. As a final remark, we note that the smoking effect based on all the methods would have been stronger if we had included non-smokers in the analysis.

8. Discussion

The paper provides a method for logistic regression when complete covariate information is known only for part of the study cohort. The idea is to combine the conditional likelihood of Y given $(X, Z, W, \delta = 1)$ and that of Y given $(Z, W, \delta = 0)$. The main result was presented for discrete (Z, W). Extension to continuous (Z, W) can be done by an approach similar to that of Wang and Wang (1997). However, the linearization techniques require more complicated calculations since the nuisance components involve both the selection probabilities and the estimation of the relative risk $E(\exp(\beta_1^t X)|Y_i = 0, V_i)$. An important feature of the proposed method is that distributional assumptions with respect to covariates are unnecessary.

The idea of the approach is based on *conditional* likelihood, by which we meant conditional on δ. This is slightly different from the parametric maximum likelihood estimator based on an additional model assumption. By the result of Satten and Kupper (1993), one may consider the likelihood of $\mathcal{L}(Y|Z, W)\mathcal{L}^{\delta}(X|Y, Z, W)$, which is equivalent to $\mathcal{L}^{\delta}(Y, X|Z, W)\mathcal{L}^{1-\delta}(Y|Z, W)$. Note that for continuous X, one may model $\mathcal{L}(X|Y, Z, W)$ from the validation set since δ and X are conditionally independent given (Y, Z, W). Certainly the advantage of the parametric maximum likelihood estimator is efficiency, at least for large sample sizes, but estimation of β may be sensitive to this additional modeling of $X|(Y, Z, W)$.

Acknowledgement

The research of C. Y. Wang was supported by US National Institutes of Health grants CA 53996, AG 15026 and a travel fund from the National Science Council and Feng Chia University, Taiwan, ROC. The research of S. T. Ou was supported by National Science Council grant of Taiwan, ROC, 87-2118-M-126-01. We are thankful to Professor Norm Breslow and two anonymous referees for helpful comments.

Appendix. Technical Proofs

Proof of Lemma 1. For any $y = 0, 1$ and v in the support of V,

$$\begin{aligned}\widehat{\pi}(y,v) - \pi(y,v) &= \frac{\sum_{j=1}^n \{\delta_j - \pi(y,v)\} I[Y_j = y, V_j = v]}{\sum_{j=1}^n I[Y_j = y, V_j = v]} \\ &= \frac{n^{-1}\sum_{j=1}^n \{\delta_j - \pi(y,v)\} I[Y_j = y, V_j = v]}{\mathrm{pr}(Y = y, V = v)} + o_p(n^{-1/2}),\end{aligned}$$

where the $o_p(n^{-1/2})$ term is uniform in (y, v) since the support of (Y, V) is finite. Note that if $\widehat{a} \to a$ and $\widehat{b} \to b$, then $\ln(\widehat{a}/\widehat{b}) = \ln(a/b) + (\widehat{a} - a)/a - (\widehat{b} - b)/b + o(\widehat{a} - a) + o(\widehat{b} - b)$. By the definition of $U_{1n}(\beta)$ and $\widehat{U}_{1n}(\beta)$, $\widehat{U}_{1n}(\beta) - U_{1n}(\beta) = n^{-1/2}\sum_{i=1}^n \delta_i \mathcal{X}_i H_+^{(1)}(X_i, V_i)$ ($[\{\widehat{\pi}(0, V_i) - \pi(0, V_i)\}/\pi(0, V_i)] - [\{\widehat{\pi}(1, V_i) - \pi(1, V_i)\}/\pi(1, V_i)]$) $+ o_p(1) \equiv A_n - B_n + o_p(1)$. Now,

$$\begin{aligned}A_n &= n^{-3/2}\sum_{i=1}^n \delta_i \mathcal{X}_i H_+^{(1)}(X_i, V_i) \frac{\sum_{j=1}^n \{\delta_j - \pi(0, V_i)\} I[Y_j = 0, V_j = V_i]}{\pi(0, V_i)\mathrm{pr}(Y_i = 0, V_i)} + o_p(1) \\ &= n^{-1/2}\sum_{j=1}^n I[Y_j = 0] n^{-1}\sum_{i=1}^n \frac{\delta_i \mathcal{X}_i H_+^{(1)}(X_i, V_i)\{\delta_j - \pi(0, V_i)\} I[V_i = V_j]}{\pi(0, V_i)\mathrm{pr}(Y_i = 0, V_i)} + o_p(1) \\ &= n^{-1/2}\sum_{j=1}^n I[Y_j = 0]\{\delta_j - \pi(0, V_j)\}\{\pi(0, V_j)\mathrm{pr}(Y_j = 0 | V_j)\}^{-1} \\ &\quad \times \mathrm{E}\{\pi(Y, V)\mathcal{X} H_+^{(1)}(X, V) | V = V_j\} + o_p(1); \\ B_n &= n^{-3/2}\sum_{i=1}^n \delta_i \mathcal{X}_i H_+^{(1)}(X_i, V_i) \frac{\sum_{j=1}^n \{\delta_j - \pi(1, V_i)\} I[Y_j = 1, V_j = V_i]}{\pi(1, V_i)\mathrm{pr}(Y_i = 1, V_i)} + o_p(1) \\ &= n^{-1/2}\sum_{j=1}^n I[Y_j = 1]\{\delta_j - \pi(1, V_j)\}\{\pi(1, V_j)\mathrm{pr}(Y_j = 1 | V_j)\}^{-1} \\ &\quad \times \mathrm{E}\{\pi(Y, V)\mathcal{X} H_+^{(1)}(X, V) | V = V_j\} + o_p(1).\end{aligned}$$

Therefore $\widehat{U}_{1n}(\beta) - U_{1n}(\beta) = n^{-\frac{1}{2}}\sum_{i=1}^n (-1)^{Y_i}\{\delta_i - \pi(Y_i, V_i)\}\{\pi(Y_i, V_i)\mathrm{pr}(Y_i | V_i)\}^{-1} \times \mathrm{E}\{\pi(Y, V)\mathcal{X} H_+^{(1)}(X, V) | V = V_i\} + o_p(1)$. This completes the proof of Lemma 1.

Proof of Lemma 2. For any $\widehat{a}, \widehat{b}$, we note that $\widehat{a}\widehat{b} = ab + a(\widehat{b} - b) + b(\widehat{a} - a) + (\widehat{a} - a)(\widehat{b} - b)$. Thus

$$\widehat{U}_{2n}(\beta) = n^{-1/2}\sum_{i=1}^n (1 - \delta_i)\mathcal{T}_i\{Y_i - H_-(V_i)\} + Q_{1n} + Q_{2n} + Q_{3n}, \tag{9}$$

where $Q_{1n} = n^{-1/2}\sum_{i=1}^n (1 - \delta_i)\mathcal{T}_i\{H_-(V_i) - \widehat{H}_-(V_i)\}$, $Q_{2n} = n^{-1/2}\sum_{i=1}^n (1 - \delta_i)(\widehat{\mathcal{T}}_i - \mathcal{T}_i)\{Y_i - H_-(V_i)\}$, $Q_{3n} = n^{-1/2}\sum_{i=1}^n (1 - \delta_i)(\widehat{\mathcal{T}}_i - \mathcal{T}_i)\{H_-(V_i) - \widehat{H}_-(V_i)\}$.

Again, if $\widehat{a} \to a$ and $\widehat{b} \to b$, then $\ln(\widehat{a}/\widehat{b}) = \ln(a/b) + (\widehat{a}-a)/a - (\widehat{b}-b)/b + o(\widehat{a}-a) + o(\widehat{b}-b)$. First,

$$\begin{aligned} Q_{1n} &= n^{-1/2}\sum_{i=1}^{n}(1-\delta_i)\mathcal{T}_i H_-^{(1)}(V_i)\Big[\{\overline{\pi}(1,V_i)\}^{-1}\{\widehat{\pi}(1,V_i)-\pi(1,V_i)\} \\ &\quad -\{\overline{\pi}(0,V_i)\}^{-1}\{\widehat{\pi}(0,V_i)-\pi(0,V_i)\} - \{r(V_i)\}^{-1}\{\widehat{r}(V_i)-r(V_i)\}\Big] + o_p(1) \\ &\equiv D_{1n} - D_{2n} - D_{3n} + o_p(1). \end{aligned}$$

By direct calculation,

$$\begin{aligned} D_{1n} &= n^{-\frac{3}{2}}\sum_{i=1}^{n}(1-\delta_i)\mathcal{T}_i H_-^{(1)}(V_i)\frac{\sum_{j=1}^{n}\{\delta_j-\pi(1,V_i)\}I[Y_j=1,V_j=V_i]}{\overline{\pi}(1,V_i)\mathrm{pr}(Y_i=1,V_i)} + o_p(1) \\ &= n^{-\frac{1}{2}}\sum_{j=1}^{n} I[Y_j=1]n^{-1}\sum_{i=1}^{n}(1-\delta_i)\mathcal{T}_i H_-^{(1)}(V_i)\frac{\{\delta_j-\pi(1,V_i)\}I[V_i=V_j]}{\overline{\pi}(1,V_i)\mathrm{pr}(Y_i=1,V_i)} + o_p(1) \\ &= n^{-\frac{1}{2}}\sum_{j=1}^{n} I[Y_j=1]\{\delta_j-\pi(1,V_j)\}\{\overline{\pi}(1,V_j)\mathrm{pr}(Y_j=1|V_j)\}^{-1}\mathcal{T}_j H_-^{(1)}(V_j) \\ &\quad \mathrm{E}[\overline{\pi}(Y,V)|V=V_j] + o_p(1); \\ D_{2n} &= n^{-\frac{3}{2}}\sum_{i=1}^{n}(1-\delta_i)\mathcal{T}_i H_-^{(1)}(V_i)\frac{\sum_{j=1}^{n}\{\delta_j-\pi(0,V_i)\}I[Y_j=0,V_j=V_i]}{\overline{\pi}(0,V_i)\mathrm{pr}(Y_i=0,V_i)} + o_p(1) \\ &= n^{-\frac{1}{2}}\sum_{j=1}^{n} I[Y_j=0]\{\delta_j-\pi(0,V_j)\}\{\overline{\pi}(0,V_j)\mathrm{pr}(Y_j=0|V_j)\}^{-1}\mathcal{T}_j H_-^{(1)}(V_j) \\ &\quad \mathrm{E}[\overline{\pi}(Y,V)|V=V_j] + o_p(1); \\ D_{3n} &= n^{-3/2}\sum_{i=1}^{n}(1-\delta_i)\mathcal{T}_i H_-^{(1)}(V_i)\{r(V_i)\}^{-1} \\ &\qquad \times\left\{\frac{\sum_{j=1}^{n}\delta_j e^{\beta_1^t X_j}I[Y_j=0,V_j=V_i]}{\mathrm{pr}(\delta_i=1,Y_i=0,V_i)} - r(V_i)\right\} + o_p(1) \\ &= n^{-1/2}\sum_{j=1}^{n} n^{-1}\sum_{i=1}^{n}(1-\delta_i)\mathcal{T}_i H_-^{(1)}(V_i)\{r(V_i)\}^{-1} \\ &\qquad \times\frac{\{e^{\beta_1^t X_j}-r(V_i)\}I[V_i=V_j]I[Y_j=0,\delta_j=1]}{\mathrm{pr}(\delta_i=1,Y_i=0,V_i)} + o_p(1) \\ &= n^{-1/2}\sum_{j=1}^{n} I[Y_j=0,\delta_j=1]\{e^{\beta_1^t X_j}-r(V_j)\}\{r(V_j)\}^{-1}\mathcal{T}_j H_-^{(1)}(V_j) \\ &\qquad \times\frac{\mathrm{pr}(\delta=0,V_j)}{\mathrm{pr}(\delta=1,Y=0,V_j)} + o_p(1). \end{aligned}$$

We have thus shown $Q_{1n} = n^{-1/2}\sum_{i=1}^{n}\mathcal{E}_m(Y_i,X_i,V_i)+o_p(1)$, where $\mathcal{E}_m(Y_i,X_i,V_i)$

was defined in Lemma 2.

Now consider Q_{2n}. Note that

$$Q_{2n} = n^{-1/2}\sum_{i=1}^{n}(1-\delta_i)\{Y_i - H_-(V_i)\}\begin{pmatrix} 0 \\ \widehat{R}^{(1)}(V_i) - R^{(1)}(V_i) \\ 0 \end{pmatrix}.$$

Because V_i has a finite support, and hence $\widehat{R}^{(1)}(V_i) - R^{(1)}(V_i) = o_p(1)$ uniformly on V_i. Observe that $n^{-1/2}\sum_{i=1}^{n}(1-\delta_i)\{Y_i - H_-(V_i)\} = O_p(1)$. Therefore, $Q_{2n} = o_p(1)$.

Similarly, it can be shown that $Q_{3n} = o_p(1)$. By (9), the proof of Lemma 2 is thus complete.

Proof of Theorem 1. Let $\dim(\beta_1) = k_1$, $\dim(\beta_2) = k_2$ and

$$\begin{aligned} G_n(\beta) &= n^{-1/2}\frac{\partial}{\partial\beta}\widehat{U}_n(\beta) \\ &= n^{-1}\sum_{i=1}^{n}\Bigg[\delta_i\mathcal{X}_i\mathcal{X}_i^t H_+^{(1)}(X_i,V_i) + (1-\delta_i)\mathcal{T}_i\mathcal{T}_i^t H_-^{(1)}(V_i) \\ &\quad -(1-\delta_i)\begin{pmatrix} 0_{1\times1} & 0_{1\times k_1} & 0_{1\times k_2} \\ 0_{k_1\times1} & \widehat{R}^{(2)}_{k_1\times k_1}(V_i) & 0_{k_1\times k_2} \\ 0_{k_2\times1} & 0_{k_2\times k_1} & 0_{k_2\times k_2} \end{pmatrix}\{Y_i - H_-(V_i)\}\Bigg], \end{aligned}$$

$$\widehat{R}^{(2)}(V_i) = \frac{\sum_{j=1}^{n}\delta_j X_j X_j^t e^{\beta_1^t X_j} I[Y_j = 0, V_j = V_i]}{\sum_{j=1}^{n}\delta_j e^{\beta_1^t X_j} I[Y_j = 0, V_j = V_i]} - \left\{\frac{\sum_{j=1}^{n}\delta_j X_j e^{\beta_1^t X_j} I[Y_j = 0, V_j = V_i]}{\sum_{j=1}^{n}\delta_j e^{\beta_1^t X_j} I[Y_j = 0, V_j = V_i]}\right\}^{\otimes 2}.$$

Then it can be shown that $G_n(\beta) \to G(\beta)$ in probability, where $G(\beta)$ was defined in Theorem 1. We note that $n^{-1/2}\{\widehat{U}_n(\widehat{\beta}) - \widehat{U}_n(\beta)\} = G_n(\beta^\star)(\widehat{\beta} - \beta)$, for some $\beta^\star$ between $\widehat{\beta}$ and β. It is easy to see that $n^{-1/2}\{\widehat{U}_n(\beta)\} \to 0$ in probability since $\mathrm{E}\{S_c(Y,X,V) + \mathcal{E}_c(Y,V) + S_m(Y,V) + \mathcal{E}_m(Y,X,V)\} = 0$ when evaluated at the true parameter. By Condition (A5), the convergence of $G_n(\beta)$ to $G(\beta)$ is uniform in a neighborhood of the true β. By the Inverse Function Theorem as in Foutz (1977), along with Condition (A4), there exists a unique consistent solution to the estimating equation $\widehat{U}_n(\beta) = 0$ in a neighborhood of the true β.

We now derive the asymptotic distribution of $n^{1/2}(\widehat{\beta} - \beta)$. By a Taylor expansion of $\widehat{U}_n(\widehat{\beta})$, it is easily seen that $0 = \widehat{U}_n(\widehat{\beta}) = \widehat{U}_n(\beta) - G_n(\beta)n^{1/2}(\widehat{\beta} - \beta) + o_p(1)$. By the consistency of $G_n(\beta)$ to $G(\beta)$, we have $n^{1/2}(\widehat{\beta} - \beta) = G^{-1}(\beta)\widehat{U}_n(\beta) + o_p(1)$. By Lemmas 1 and 2, $\mathrm{Cov}\{\widehat{U}_n(\beta)\} = M(\beta)$ defined in Theorem 1, since $S_c(Y_i,X_i,V_i) + \mathcal{E}_c(Y_i,V_i) + S_m(Y_i,V_i) + \mathcal{E}_m(Y_i,X_i,V_i)$ are independent variables for $i = 1,\ldots,n$. By the Taylor expansion above, the proof of the asymptotic covariance matrix is thus complete.

References

Bruemmer, B., White, E., Vaughan, T. and Cheney, C. (1996). Nutrient intake in relationship to bladder cancer among middle aged men and women. *Amer. J. Epidemiology* **144**, 485-495.

Breslow, N. E. and Cain, K. C. (1988). Logistic regression for two–stage case–control data. *Biometrika* **75**, 11-20.

Breslow, N. E. and Chatterjee N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Appl. Statist.* **48**, 457-468.

Breslow, N. E. and Holubkov, R. (1997). Maximum likelihood estimation for logistic regression parameters under two-phase, outcome-dependent sampling. *J. R. Statist. Soc. Ser. B* **59**, 447-461.

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. Chapman and Hall, London.

Chen, Y. H. and Chen H. (2000). A unified approach to regression under double sampling design. *J. R. Statist. Soc. Ser. B* **62**, 449-460.

Foutz, R. V. (1977). On the unique consistent solution to the likelihood equations. *J. Amer. Statist. Assoc.* **72**, 147-148.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47**, 663-685.

Reilly, M. and Pepe, M. S. (1995). A mean-score method for missing and auxiliary covariate data in regression models. *Biometrika* **82**, 299-314.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846-866.

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

Satten, G. A. and L. L. Kupper (1993). Inferences about exposure-disease association using probability of exposure information. *J. Amer. Statist. Assoc.* **88**, 200-208.

Schill, W., Jöckel, K.-H., Drescher, K. and Timm, J. (1993). Logistic analysis in case-control studies under validation sampling. *Biometrika* **80**, 339-352.

Scott, A. J. and Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84**, 57-71.

Wang, C. Y. and Wang, S. (1997). Semiparametric methods in logistic regression with measurement error. *Statist. Sinica* **7**, 1103-1120.

Wang, C. Y., Wang, S. and Carroll, R. J. (1997). Estimation in choice-based sampling with measurement error and bootstrap analysis. *J. Econometrics* **77**, 65-86.

Wang, C. Y., Wang, S., Zhao, L. P. and Ou, S. T. (1997). Weighted semiparametric estimation in regression analysis with missing covariate data. *J. Amer. Statist. Assoc.* **92**, 512-525.

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, P.O. Box 19024, Seattle, WA 98109-1024, U.S.A.

E-mail: cywang@fhcrc.org

Institute of Applied Statistics, Fu Jen Catholic University, Taipei, Taiwan.

E-mail: stat1014@mails.fju.edu.tw

Department of Statistics, Feng Chia University, Taichung, Taiwan.

E-mail: smlee@stat.fcu.edu.tw

Department of Statistics, Tamkang University, Tamshui, Taiwan.

E-mail: 073336@mail.tku.edu.tw

(Received November 1999; accepted June 2001)