Statistica Sinica **13**(2003), 111-127

ESTIMATION OF A LOGISTIC REGRESSION MODEL WITH MISMEASURED OBSERVATIONS

K. F. Cheng and H. M. Hsueh

National Central University and National Chengchi University

Abstract: We consider the estimation problem of a logistic regression model. We assume the response observations and covariate values are both subject to measurement errors. We discuss some parametric and semiparametric estimation methods using mismeasured observations with validation data and derive their asypmtotic distributions. Our results are extentions of some well known results in the literature. Comparisons of the asymptotic covariance matrices of the studied estimators are made, and some lower and upper bounds for the asymptotic relative efficiencies are given to show the advantages of the semiparametric method. Some simulation results also show the method performs well.

Key words and phrases: Kernel estimation, estimated likelihood, logistic regression, measurement error, misclassification.

1. Introduction

Logistic regression is the most popular form of binary regression; see Cox (1970) and Pregibon (1981). Researchers often use logistic regression to estimate the effect of various predictors on some binary outcome of interest. Basically, the model assumes that the log of the odds of the outcome is a linear function of the predictors. That is, suppose that the variables (X_i, Y_i) follow the model

$$Pr(Y_i = 1 \mid X_i = x_i) = \frac{\exp(x_i^T \beta^0)}{\{1 + \exp(x_i^T \beta^0)\}} \equiv F(x_i^T \beta^0),$$

where the X_i are random d-vector predictors and the Y_i are Bernoulli response variables. Usually, the maximum likelihood (ML) method is used to estimate the regression coefficients β^0. Under some regularity conditions, the maximum likelihood estimator of β^0 is asymptotically normal.

The ML method requires that the data consist of precise measurements for the binary outcomes and predictors. However, the data are often not measured perfectly. For instance, Golm, Holloran and Longini (1998, 1999) mentioned that collecting information on exposure to infection for estimating vaccine efficacy may be mismeasured. On the other hand, Albert, Hunsberger and Bird (1997) and Bollinger and David (1997) gave examples showing that the binary outcome of interest may also be misclassified. It is generally true that the usual analyses based on the mismeasured observations lead to inconsistent estimation.

The topic of binary regression when predictors X_i are measured with error has been the subject of several recent papers, see Carroll, Spiegelman, Lan, Bailey and Abbott (1984), Carroll and Wand (1991), Reilly and Pepe (1995), and Lawless, Kalbfleisch and Wild(1999), etc. When the binary responses Y_i are subject to misclassification, Pepe(1992) and Cheng and Hsueh (1999) discussed bias correction methods in the estimation of logistic regression parameters. In this paper, we study the estimation problem of β^0 when both observations Y_i and predictors X_i are measured with error. Parametric and semiparametric methods are discussed. We find that the proposed semiparametric estimation method is a generalization of the pseudolikelihood method of Carroll and Wand (1991) and the estimated likelihood method of Pepe (1992). Many other estimating approaches have been proposed in the literature. A method based on the mean score was proposed by Pepe, Reilly and Fleming (1994) and Reilly and Pepe (1995) for the mismeasured outcome data problem and mismeasured covariate data problem, respectively. However, Golm *et al.* (1999) and Lawless *et al.* (1999) argued that this semiparametric method is less efficient. Robins, Rotnitzky and Zhao (1994) proposed a class of semiparametric efficient estimators for the model with mismeasured predictors based on the inverse probability weighted estimating equation approach. When both response variable and predictors are subject to measurement error, the semiparametric efficient estimator has not been formally derived yet. In this paper we only emphasize various imputation approaches. The weighting methods, such as the semiparametric efficient estimation derived by Robins *et al.* (1994), will not be discussed further.

We assume in this paper that the complete data set consists of a primary sample plus a smaller validation subsample which is obtained by double sampling scheme. Extension of the selection probabilities for the validation data set to depend on the observed surrogate covariates is discussed briefly in Section 3. Asymptotically, we also suppose that the validation subsample size is a fraction of the major sample size. The estimators under discussion will be formally defined in Section 2. In Section 3, some asymptotic results are derived for semiparametric and parametric estimators of β^0. Comparisons of their asymptotic covariance matrices are given in Section 4. Further, finite sample properties are explored through a simulation study in Section 5.

2. Estimation Methods

Suppose the true random variables (Y, X) are subject to mismeasurement and the surrogate observations are represented by (Y^0, W), here $X = (1, X_1)^T$ and $W = (1, W_1)^T$. In addition to the primary sample $\{(Y_i^0, W_i), i = 1, \ldots, n\}$, a smaller validation subsample is also observed in order to understand the mismeasurement structure. The sampling scheme is to randomly select k units from the primary sample and at the selected units the true measurement devices are used to obtain the validation data. Here the simple random sampling scheme is considered and for the sake of simplicity, the first k units of the primary sample are assumed to be the selected units. Thus we have the validation subsample $\{(Y_i^0, Y_i, X_i, W_i), i = 1, \ldots, k\}$. Later in Section 3 the theoretical results will be given for the case that the sampling scheme depends on the value of the surrogate predictors W.

From the primary sample, we denote the regression function of Y^0 on $W = w$ as $\pi^0(w) = Pr(Y^0 = 1 \mid w)$. For Y and W being assumed mutually independent given X, and Y^0 and X being mutually independent given Y, W, the above regression function can be rewritten as $\pi^0(w) = \pi(w, \beta^0)$, where $\pi(w, \beta) = \{1 - \theta^0(w)\}E\{F(X^T\beta) \mid w\} + \phi^0(w)E\{\bar{F}(X^T\beta) \mid w\} \equiv 1 - \bar{\pi}(w, \beta)$. Here the expectation is taken with respect to the conditional density $f^0_{X|W}$, $\bar{F}(\cdot) = 1 - F(\cdot)$, and the misclassification probability functions $\theta^0(w)$ and $\phi^0(w)$ are defined by $\theta^0(w) = Pr(Y^0 = 0 \mid Y = 1, w)$ and $\phi^0(w) = Pr(Y^0 = 1 \mid Y = 0, w)$. Note that the surrogate W is often a fallible measurement of X and corresponding to a coarser partition in the sample space. The conditional independence of Y^0 and X given Y, W would be implied if the misclassification scheme depends on X only through the value of W. It is clear that the expectation $E[\{Y^0 - F(W^T\beta^0)\}W^T]$ is not necessarily zero. Hence it is inappropriate to apply the usual likelihood method to the primary sample for inference about β^0. In the following, we discuss some consistent estimates based on different imputation methods.

Suppose first that $f^0_{X|W}$ and $\theta^0(w)$, $\phi^0(w)$ are known a priori. Then for the observation (y_j^0, w_j), $j = k+1, \ldots, n$, in the primary sample, the corresponding log-likelihood and score function are

$$\begin{aligned}
L^*(\beta \mid y_j^0, w_j) &= y_j^0 \ln \pi(w_j, \beta) + (1 - y_j^0) \ln \bar{\pi}(w_j, \beta),\\
S^*(\beta \mid y_j^0, w_j) &\equiv \frac{\partial L^*(\beta \mid y_j^0, w_j)}{\partial \beta}\\
&= \frac{y_j^0 - \pi(w_j, \beta)}{\pi(w_j, \beta)\bar{\pi}(w_j, \beta)}\{1 - \theta^0(w_j) - \phi^0(w_j)\}E\{F(X_j^T\beta)\bar{F}(X_j^T\beta)X_j \mid w_j\}.
\end{aligned}$$

Since,

$$\frac{y_j^0 - \pi(w_j, \beta)}{\pi(w_j, \beta)\bar{\pi}(w_j, \beta)} = \frac{2y_j^0 - 1}{y_j^0\pi(w_j, \beta) + (1 - y_j^0)\bar{\pi}(w_j, \beta)},$$

$$1 - \theta^0(w_j) - \phi^0(w_j) = \{1 - \theta^0(w_j) - \pi(w_j, \beta)\}\frac{E\{F(X_j^T\beta) \mid w_j\}}{E\{F(X_j^T\beta) \mid w_j\}E\{\bar{F}(X_j^T\beta) \mid w_j\}},$$

we have

$$\frac{y_j^0 - \pi(w_j,\beta)}{\pi(w_j,\beta)\bar{\pi}(w_j,\beta)}\{1-\theta^0(w_j)-\phi^0(w_j)\} = \frac{E(Y_j \mid y_j^0, w_j) - E\{F(X_j^T\beta) \mid w_j\}}{E\{F(X_j^T\beta) \mid w_j\}E\{\bar{F}(X_j^T\beta) \mid w_j\}}.$$

Consequently,

$$\begin{aligned} S^*(\beta \mid y_j^0, w_j) &= \frac{E(Y_j \mid y_j^0, w_j) - E\{F(X_j^T\beta) \mid w_j\}}{E\{F(X_j^T\beta) \mid w_j\}E\{\bar{F}(X_j^T\beta) \mid w_j\}} E\left\{F(X_j^T\beta)\bar{F}(X_j^T\beta)X_j \mid w_j\right\} \\ &= \frac{\{A_1^0(y_j^0,w_j) - A_2^0(y_j^0,w_j)\}E\{F(X_j^T\beta)\bar{F}(X_j^T\beta)X_j \mid w_j\}}{\left[A_1^0(y_j^0,w_j)E\{F(X_j^T\beta) \mid w_j\} + A_2^0(y_j^0,w_j)E\{\bar{F}(X_j^T\beta) \mid w_j\}\right]}, \end{aligned}$$

where $A_1^0(y_j^0, w_j) = Pr(Y_j^0 = y_j^0 \mid Y_j = 1, w_j) = \{\theta^0(w_j)\}^{(1-y_j^0)}\{1-\theta^0(w_j)\}^{y_j^0}$, and $A_2^0(y_j^0, w_j) = Pr(Y_j^0 = y_j^0 \mid Y_j = 0, w_j) = \{\phi^0(w_j)\}^{y_j^0}\{1-\phi^0(w_j)\}^{(1-y_j^0)}$. Using this result, we easily see that the MLE $\hat{\beta}_f$ of β^0 can be obtained by solving the likelihood equations

$$\sum_{i=1}^{k} S(\beta \mid y_i, x_i) + \sum_{j=k+1}^{n} S^*(\beta \mid y_j^0, w_j) = 0, \tag{1}$$

where $S(\beta \mid y_i, x_i) = \{y_i - F(x_i^T\beta)\}x_i$.

Unfortunately, in applications $f_{X|W}^0(\cdot)$, $\theta^0(\cdot)$ and $\phi^0(\cdot)$ are rarely known and $\hat{\beta}_f$ can not be obtained. If the subsample size k is large enough, a simple estimate $\hat{\beta}_s$ can be obtained by using only the validation subsample, i.e., $\hat{\beta}_s$ satisfies $\sum_{i=1}^k S(\beta \mid y_i, x_i) = 0$. Such a simple estimate is in general not efficient; see Cheng and Hsueh (1999). The basic reason is that the information contained in the primary sample is not properly exploited. To do this, assume there exist parametric functions and parameters (γ^0, α^0), independent of β^0, such that $\theta^0(W) = \theta(W;\alpha^0)$, $\phi^0(W) = \phi(W;\alpha^0)$ and $f_{X|W}^0(X \mid W) = f(X \mid W;\gamma^0)$ almost surely. See Carroll and Wand (1991) and Cheng and Hsueh (1999) for related discussions. Then one can employ the usual approach to obtain the joint MLE $(\hat{\beta}_m, \hat{\alpha}_m, \hat{\gamma}_m)$. On the other hand, the maximum pseudolikelihood estimate (MPLE) $\hat{\beta}_p$ can be obtained by solving (1) with α^0 and γ^0 in $S^*(\beta \mid y_j^0, w_j)$ being replaced by their MLE $\hat{\alpha}_p$ and $\hat{\gamma}_p$ derived from the validation data.

In general, $\hat{\beta}_m$ can be shown to be more efficient than $\hat{\beta}_p$ asymptotically. However, in solving the equations for finite sample cases, the computational complexity grows (and numerical stability deteriorates) with the number of unknown parameters, so that $\hat{\beta}_m$ may not have proper performance, see the simulation study in Section 5.

The last approach to be discussed is a semiparametric method. Assuming $f^0_{X|W}(\cdot \mid w)$, $\theta^0(w)$ and $\phi^0(w)$ in $S^*(\beta \mid y^0_j, w_j)$ of (1) are unknown functions, a nonparametric method is used to estimate them. Formally, we propose the MPLE $\hat{\beta}_{sp}$ solving the following pseudolikelihood equations

$$\sum_{i=1}^{k} S(\beta \mid y_i, x_i) + \sum_{j=k+1}^{n} \hat{S}^*(\beta \mid y^0_j, w_j) = 0,$$

$$\hat{S}^*(\beta \mid y^0, w) = \frac{\{\hat{A}_1(y^0, w) - \hat{A}_2(y^0, w)\}\hat{E}[F(X^T\beta)\bar{F}(X^T\beta)X \mid w]}{\Big[\hat{A}_1(y^0, w)\hat{E}\{F(X^T\beta) \mid w\} + \hat{A}_2(y^0, w)\hat{E}\{\bar{F}(X^T\beta) \mid w\}\Big]}.$$

Here, for any integrable function $g(x;\beta)$, the estimated expectation $\hat{E}\{g(X;\beta) \mid w\}$ is given by $\hat{E}\{g(X;\beta) \mid w\} = \{\sum_{i=1}^k K_h(w - w_i)g(x_i;\beta)\}/\{\sum_{i=1}^k K_h(w - w_i)\}$. Moreover $\hat{A}_1(y^0, w)$ and $\hat{A}_2(y^0, w)$ are estimates of $A^0_1(y^0, w)$ and $A^0_2(y^0, w)$, with $\theta^0(w)$ and $\phi^0(w)$ being replaced by their nonparametric estimates $\hat{\theta}(w) = \{\sum_{i=1}^k K_h(w - w_i)y_i(1 - y^0_i)\}/\{\sum_{i=1}^k K_h(w - w_i)y_i\}$ and $\hat{\phi}(w) = \{\sum_{i=1}^k K_h(w - w_i)(1 - y_i)y^0_i\}/\{\sum_{i=1}^k K_h(w - w_i)(1 - y_i)\}$. The above kernel function $K(t)$ is taken to be a density function, $K_h(t) = h^{-1}K(t/h)$, the bandwidth h depends on n and tends to zero as $n \to \infty$.

We remark that the proposed semiparametric estimation method generalizes the methods of Carroll and Wand (1991) and Pepe (1992). If $Y = Y^0$ with probability one, i.e., only measurement error occurs in the predictor, $\hat{\beta}_{sp}$ is the MPLE of Carroll and Wand (1991). On the other hand, if $X = W$ with probability one, i.e., only misclassification occurs in the response, $\hat{\beta}_{sp}$ reduces to the maximum estimated likelihood estimate of Pepe (1992).

3. Asymptotic Distributions

The asymptotic distributions of the estimators proposed in the preceding section will be presented here for the case $d = 1$. Extension to general d is simple and the related results will be given in a remark. The asymptotic properties depend on certain regularity conditions.

A.1 $\beta^0 \in \Lambda$, an open set in R^2.

A.2 $E(X^2) < \infty$.

A.3 The misclassification probability functions $\theta^0(w)$ and $\phi^0(w) \in (0, 1)$, and the density function $f^0_W(w)$ of W are strictly positive in the space of W. Furthermore, these functions and their l-th derivatives, $l = 1$, 2, are in L_4 and also satisfy a weighted Lipschitz condition. (A function $\eta(\cdot)$ is said to satisfy a weighted Lipschitz condition if there exists a constant c and a bounded function ψ in L_4 such that $|\eta(x) - \eta(y)| < \psi(x)|x - y|$ for all x, y with $|x - y| < c$.)

A.4 The parametric functions $\theta(w;\alpha)$,$\phi(w;\alpha)$ and $f(x \mid w;\gamma)$, and their partial derivatives $\partial\theta(w;\alpha)/\partial\alpha$, $\partial\phi(w;\alpha)/\partial\alpha$ and $\partial f(x \mid w;\gamma)/\partial\gamma$, are uniformly continuous at (α^0,γ^0) for all x and w in the support of X and W.

A.5 The function $K(\cdot)$ is a second-order kernel (see Gasser and Müller (1979)).

A.6 $h = h_n \to 0$ with $nh^2 \to \infty$ and $nh^4 \to 0$, as $n \to \infty$. Also, as $n \to \infty$, $k = rn\{1 + O(h^2)\}$, where $r \in (0, 1]$.

In the following, we give the basic asymptotic results under the assumption that $\alpha = (\alpha_0, \alpha_1)^T$ and $\gamma = (\gamma_0, \gamma_1)^T$. Except for the asymptotic normality of $\hat\beta_{sp}$, whose proof is given in the Appendix, the proofs for other estimators are standard and thus are omitted.

Theorem 1. *Suppose conditions* (A.1)−(A.6) *hold and let $\tilde\beta$ be any estimator discussed in Section* 2*. Then as $n \to \infty$, $n^{1/2}(\tilde\beta - \beta^0)$ converges in distribution to a normal with mean zero. Their asymptotic covariance matrices are*

$$
\begin{aligned}
&\lim_{n\to\infty} \mathrm{Cov}\,\{n^{1/2}(\hat\beta_f - \beta^0)\} = \Sigma_f = \{rI_v + (1-r)I_v^c\}^{-1},\\
&\lim_{n\to\infty} \mathrm{Cov}\,\{n^{1/2}(\hat\beta_s - \beta^0)\} = \Sigma_s = (rI_v)^{-1},\\
&\lim_{n\to\infty} \mathrm{Cov}\,\{n^{1/2}(\hat\beta_m - \beta^0)\} = \Sigma_m = \Sigma_f + \frac{(1-r)^2}{r}\Sigma_f I_m \Sigma_f,\\
&\lim_{n\to\infty} \mathrm{Cov}\,\{n^{1/2}(\hat\beta_p - \beta^0)\} = \Sigma_p = \Sigma_f + \frac{(1-r)^2}{r}\Sigma_f I_p \Sigma_f,\\
&\lim_{n\to\infty} \mathrm{Cov}\,\{n^{1/2}(\hat\beta_{sp} - \beta^0)\} = \Sigma_{sp} = \Sigma_f + \frac{(1-r)^2}{r}\Sigma_f I_{sp} \Sigma_f.
\end{aligned}
$$

Here $I_v = E\{F(X^T\beta^0)\bar F(X^T\beta^0)X^{\otimes 2}\}$ and

$$
I_v^c = E\left(\frac{\{1-\theta^0(W)-\phi^0(W)\}^2\left[E\{F(X^T\beta^0)\bar F(X^T\beta^0)X \mid W\}\right]^{\otimes 2}}{\pi^0(W)\{1-\pi^0(W)\}}\right),
$$

$$
\begin{aligned}
I_p = {}& E\left[\frac{\{1-\theta^0(W)-\phi^0(W)\}E\{F(X^T\beta^0)\bar F(X^T\beta^0)X| W\}}{\pi^0(W)\{1-\pi^0(W)\}}\begin{pmatrix}\dot\pi^0_\alpha(W;\alpha^0,\gamma^0)\\ \dot\pi^0_\gamma(W;\alpha^0,\gamma^0)\end{pmatrix}^T\right]\\
&\times\begin{pmatrix} E\left[\frac{F(X^T\beta^0)\dot\theta^{\otimes 2}(W;\alpha^0)}{\theta^0(W)\{1-\theta^0(W)\}}+\frac{\bar F(X^T\beta^0)\dot\phi^{\otimes 2}(W;\alpha^0)}{\phi^0(W)\{1-\phi^0(W)\}}\right] & 0\\ 0 & E\{\dot f(X\mid W;\gamma^0)\}^{\otimes 2}\end{pmatrix}^{-1}\\
&\times E\left[\frac{\{1-\theta^0(W)-\phi^0(W)\}E\{F(X^T\beta^0)\bar F(X^T\beta^0)X| W\}}{\pi^0(W)\{1-\pi^0(W)\}}\begin{pmatrix}\dot\pi^0_\alpha(W;\alpha^0,\gamma^0)\\ \dot\pi^0_\gamma(W;\alpha^0,\gamma^0)\end{pmatrix}^T\right]^T,
\end{aligned}
$$

$$I_{sp} = E\left\{\frac{\{1-\theta^0(W)-\phi^0(W)\}\left[E\{F(X^T\beta^0)\bar{F}(X^T\beta^0)X \mid W\}\right]^{\otimes 2}}{[\pi^0(W)\{1-\pi^0(W)\}]^2}\right.$$
$$\times\Big(E\{F(X^T\beta^0)|\,W\}\theta^0(W)\{1-\theta^0(W)\}+E\{\bar{F}(X^T\beta^0)|\,W\}\phi^0(W)\{1-\phi^0(W)\}$$
$$\left.+\ \{1-\theta^0(W)-\phi^0(W)\}^2[E\{F^2(X^T\beta^0)\mid W\}-E^2\{F(X^T\beta^0)\mid W\}]\Big)\right\},$$

$$I_m = \frac{r}{(1-r)^2}\Delta_{(\beta^0,\alpha^0,\gamma^0)}\{\Delta_{(\alpha^0,\gamma^0)}-\Delta^T_{(\beta^0,\alpha^0,\gamma^0)}\Sigma_f\Delta_{(\beta^0,\alpha^0,\gamma^0)}\}^{-1}\Delta^T_{(\beta^0,\alpha^0,\gamma^0)},$$

$$\Delta_{(\beta^0,\alpha^0,\gamma^0)} = (1-r)E\left[\frac{\{1-\theta^0(W)-\phi^0(W)\}E\{F(X^T\beta^0)\bar{F}(X^T\beta^0)X \mid W\}}{\pi^0(W)\{1-\pi^0(W)\}}\right.$$
$$\left.\times\begin{pmatrix}\dot{\pi}^0_\alpha(W;\alpha^0,\gamma^0)\\ \dot{\pi}^0_\gamma(W;\alpha^0,\gamma^0)\end{pmatrix}^T\right],$$

$$\Delta_{(\alpha^0,\gamma^0)} = r\begin{pmatrix}E\left[\frac{F(X^T\beta^0)\dot{\theta}^{\otimes 2}(W;\alpha^0)}{\theta^0(W)\{1-\theta^0(W)\}}+\frac{\bar{F}(X^T\beta^0)\dot{\phi}^{\otimes 2}(W;\alpha^0)}{\phi^0(W)\{1-\phi^0(W)\}}\right] & 0\\ 0 & E\{\dot{f}(X\mid W;\gamma^0)\}^{\otimes 2}\end{pmatrix}$$
$$+(1-r)E\left[\frac{1}{\pi^0(W)\{1-\pi^0(W)\}}\begin{pmatrix}\dot{\pi}^0_\alpha(W;\alpha^0,\gamma^0)\\ \dot{\pi}^0_\gamma(W;\alpha^0,\gamma^0)\end{pmatrix}^{\otimes 2}\right].$$

In this,

$$\dot{\pi}_\alpha(W;\alpha^0,\gamma^0) = \left.\frac{\partial\pi^0(W)}{\partial\alpha}\right|_{(\alpha^0,\gamma^0)}$$
$$= E\{\bar{F}(X^T\beta^0)\mid W\}\dot{\phi}(W;\alpha^0) - E\{F(X^T\beta^0)\mid W\}\dot{\theta}(W;\alpha^0),$$
$$\dot{\pi}_\gamma(W;\alpha^0,\gamma^0) = \left.\frac{\partial\pi^0(W)}{\partial\gamma}\right|_{(\alpha^0,\gamma^0)}$$
$$= \{1-\theta^0(W)-\phi^0(W)\}E\{F(X^T\beta^0)\dot{f}(X\mid W;\gamma^0)\mid W\},$$
$$\dot{\theta}(W;\alpha^0) = \left.\frac{\partial\theta(W;\alpha)}{\partial\alpha}\right|_{\alpha^0}, \qquad \dot{\phi}(W;\alpha^0) = \left.\frac{\partial\phi(W;\alpha)}{\partial\alpha}\right|_{\alpha^0},$$
$$\dot{f}(X\mid W;\gamma^0) = \left.\frac{\partial\ln f(X\mid W;\gamma)}{\partial\gamma}\right|_{\gamma^0},$$

where, for any column vector A, $A^{\otimes 2} = AA^T$.

Remark 1. It can be seen that I_m, I_p and I_{sp} are nonnegative definite matrices, and thus $\frac{(1-r)^2}{r}\Sigma_f I_m\Sigma_f$, $\frac{(1-r)^2}{r}\Sigma_f I_p\Sigma_f$, $\frac{(1-r)^2}{r}\Sigma_f I_{sp}\Sigma_f$ can be regarded as additional variations due to estimating the unknown functions $\theta^0(W)$, $\phi^0(W)$ and $f^0_{X|W}$ by parametric and nonparametric methods.

Remark 2. It is well known that the information matrix I_v is the variance of the score; that is, $I_v = \mathrm{Var}\{S(\beta^0 \mid Y, X)\} = E[\mathrm{Var}\{S(\beta^0 \mid Y, X) \mid X\}]$. We see that the information matrices I_v^c and I_{sp} also have similar expressions: $I_v^c = E[\mathrm{Var}\{S^*(\beta^0 \mid Y^0, W) \mid W\}] = E[\rho(Y, Y^0 \mid W)\mathrm{Var}\{S_*(\beta^0 \mid Y, W) \mid W\}]$, and $I_{sp} = E(\rho(Y, Y^0 \mid W)[1-\rho(Y, Y^0 \mid W)\{1-\rho(Y, F(X^T\beta^0) \mid W)\}]\mathrm{Var}\{S_*(\beta^0 \mid Y, W) \mid W\})$, where $S_*(\beta \mid Y, W) = \partial \ln Pr(Y \mid W)/\partial\beta$, and $\rho(Y, Y^0 \mid W) = Corr^2(Y, Y^0 \mid W) = \{\{1-\theta^0(W)-\phi^0(W)\}^2 E\{F(X^T\beta^0) \mid W\}E\{\bar{F}(X^T\beta^0) \mid W\}\}/[\pi^0(W)\{1-\pi^0(W)\}]$,

$$\begin{aligned}\rho(Y, F(X^T\beta^0) \mid W) &= Corr^2(Y, F(X^T\beta^0) \mid W)\\ &= \frac{E\{F^2(X^T\beta^0) \mid W\} - [E\{F(X^T\beta^0) \mid W\}]^2}{E\{F(X^T\beta^0) \mid W\}E\{\bar{F}(X^T\beta^0) \mid W\}}.\end{aligned}$$

Thus these information matrices depend not only on the score $S_*(\beta^0 \mid Y, W)$ but also on the squared correlation functions.

Remark 3. For obtaining the validation subsample, a simple random sampling design is used. However, the selection probabilities may depend on the value of W. For example, suppose $g(w)$ is the selection probability. Then the asymptotic covariance matrix of $\hat{\beta}_{sp}$ becomes $\Sigma^*_{sp} = \Sigma^* + \Sigma^* I^*_{sp}\Sigma^*$, where

$$\begin{aligned}\Sigma^{*-1} =& E\Big[g(W)E\{F(X^T\beta^0)\bar{F}(X^T\beta^0)X^{\otimes 2} \mid W\}\Big]\\ &+E\Big(\{1-g(W)\}\frac{\{1-\theta^0(W)-\phi^0(W)\}^2\Big[E\{F(X^T\beta^0)\bar{F}(X^T\beta^0)X \mid W\}\Big]^{\otimes 2}}{\pi^0(W)\{1-\pi^0(W)\}}\Big),\\ I^*_{sp} =& E\Big\{\frac{\{1-g(W)\}^2}{g(W)}\frac{\{1-\theta^0(W)-\phi^0(W)\}\Big[E\{F(X^T\beta^0)\bar{F}(X^T\beta^0)X \mid W\}\Big]^{\otimes 2}}{[\pi^0(W)\{1-\pi^0(W)\}]^2}\\ &\times\Big(E\{F(X^T\beta^0) \mid W\}\theta^0(W)\{1-\theta^0(W)\}+E\{\bar{F}(X^T\beta^0) \mid W\}\phi^0(W)\{1-\phi^0(W)\}\\ &+\{1-\theta^0(W)-\phi^0(W)\}^2[E\{F^2(X^T\beta^0) \mid W\} - E^2\{F(X^T\beta^0) \mid W\}]\Big)\Big\}.\end{aligned}$$

Other asymptotic covariance matrices in Theorem 1 can be modified accordingly.

Remark 4. It is clear that Σ_{sp} reduces to the asymptotic covariance matrix of Carroll and Wand's (1991) estimator provided $Y = Y^0$ with probability one. Suppose $X = W$ with probability one, then Σ_{sp} is the asymptotic covariance matrix of Pepe's (1992) estimator; see also Cheng and Hsueh (1999).

Remark 5. The results in Theorem 1 can be extended to $d > 1$ vector predictors, provided the function K is a pth order kernel, $p > d$, and the bandwidth parameter satisfies $nh^{2d} \to \infty$ and $nh^{2p} \to 0$ as $n \to \infty$.

Remark 6. All the asymptotic covariance matrices can be estimated by moment type estimates.

4. Comparisons of Asymptotic Covariance Matrices

In this section, we discuss the behaviors of different estimators by comparing their asymptotic covariance matrices. The general results basically agree with our expectation, but some of their proofs are not trivial. For any matrices A and B, we write $A \geq B$ if $A - B$ is a nonnegative definite matrix.

4.1. Results under general conditions

First, from Remark 1 of Section 3, we have $\Sigma_m \geq \Sigma_f$, $\Sigma_p \geq \Sigma_f$ and $\Sigma_{sp} \geq \Sigma_f$. Further, suppose we let $\{\Delta^v_{(\alpha^0,\gamma^0)}\}^{-1}$ be the asymptotic covariance matrix of the MLE $(\hat{\alpha}_p, \hat{\gamma}_p)$ which is obtained from the validation subsample. Then one can rewrite Σ_p as $\Sigma_p = \Sigma_f[\Sigma_f^{-1} + \Delta_{(\beta^0,\alpha^0,\gamma^0)}\{\Delta^v_{(\alpha^0,\gamma^0)}\}^{-1}\Delta^T_{(\beta^0,\alpha^0,\gamma^0)}]\Sigma_f$, and consequently we have

$$
\begin{aligned}
&\Sigma_p - \Sigma_m \\
&= \Sigma_f \Delta_{(\beta^0,\alpha^0,\gamma^0)} \Big[\{\Delta^v_{(\alpha^0,\gamma^0)}\}^{-1} - \{\Delta_{(\alpha^0,\gamma^0)} - \Delta^T_{(\beta^0,\alpha^0,\gamma^0)}\Sigma_f \Delta_{(\beta^0,\alpha^0,\gamma^0)}\}^{-1}\Big] \Delta^T_{(\beta^0,\alpha^0,\gamma^0)}\Sigma_f.
\end{aligned}
$$

Since $\{\Delta^v_{(\alpha^0,\gamma^0)}\}^{-1}$ and $\{\Delta_{(\alpha^0,\gamma^0)} - \Delta^T_{(\beta^0,\alpha^0,\gamma^0)}\Sigma_f\Delta_{(\beta^0,\alpha^0,\gamma^0)}\}^{-1}$ are respectively, the asymptotic covariance matrices of $(\hat{\alpha}_p, \hat{\gamma}_p)$ and $(\hat{\alpha}_m, \hat{\gamma}_m)$, the above results yield the following theorem.

Theorem 2. *Given conditions* (A.1)−(A.6), *we have* $\Sigma_s \geq \Sigma_f, \Sigma_{sp} \geq \Sigma_f$ *and* $\Sigma_p \geq \Sigma_m \geq \Sigma_f$.

4.2. Results under special constraints

Cheng and Hsueh (1999) compared the asymptotic covariance matrices when $X = W$ with probability one. Here the comparisons focus on the case when $Y = Y^0$ with probability one.

As a consequence of Theorem 1, if $Y = Y^0$ with probability one, the matrices Σ_f and I_{sp} simplify to

$$
\begin{aligned}
\Sigma_f = \Big[&rE\{F(X^T\beta^0)\bar{F}(X^T\beta^0)X^{\otimes 2}\} \\
&+(1-r)E\Big(\frac{[E\{F(X^T\beta^0)\bar{F}(X^T\beta^0)X \mid W\}]^{\otimes 2}}{E\{F(X^T\beta^0) \mid W\}E\{\bar{F}(X^T\beta^0) \mid W\}}\Big)\Big]^{-1},
\end{aligned}
$$

$$I_{sp} = E\left(\rho(Y, F(X^T\beta^0) \mid W)E\left[\{S_*(\beta^0 \mid Y, W)\}^{\otimes 2} \mid W\right]\right).$$

Then Corollary 1 follows easily from

$$\Sigma_s - \Sigma_{sp} = \Sigma_f\Bigg((1-r)E\{S_*(\beta^0 \mid Y, W)\}^{\otimes 2}$$
$$-\frac{(1-r)^2}{r}E\left(\rho(Y, F(X^T\beta^0) \mid W)E[\{S_*(\beta^0 \mid Y, W)\}^{\otimes 2} \mid W]\right)$$
$$+\frac{(1-r)^2}{r}E\{S_*(\beta^0|Y,W)\}^{\otimes 2}\left[E\{S(\beta^0|Y,X)\}^{\otimes 2}\right]^{-1}E\{S_*(\beta^0|Y,W)\}^{\otimes 2}\Bigg)\Sigma_f.$$

Corollary 1. *Suppose conditions* (A.1)$-$(A.6) *are satisfied, that* $Y = Y^0$ *with probability one, and* $\rho(Y, F(X^T\beta^0) \mid W) \leq min(r/(1-r), 1)$. *Then* $\Sigma_s \geq \Sigma_{sp}$.

Accordingly, the semiparametric estimator $\hat{\beta}_{sp}$ is always better than the simple MLE $\hat{\beta}_s$ under our conditions. Similar conclusion can be derived for the case that $X = W$ with probability one. Now suppose β^0 is a scalar. Then under the conditions of Theorem 1 and $\rho(Y, F(X^T\beta^0) \mid W) \leq \rho^*$, with probability one for some constant ρ^*, the ARE of $\hat{\beta}_s$ with respect to $\hat{\beta}_{sp}$ is always smaller than

$$e_1(\rho^*, I, r) = 1 - \frac{(1-r)^2 I[I - \{\rho^* - r/(1-r)\}]}{\{r + (1-r)I\}^2},$$

where the coefficient of reliability $I = E\{S_*(\beta^0 \mid Y, W)\}^2/E\{S(\beta^0 \mid Y, X)\}^2 \leq 1$. This can be used to measure the quality of surrogate predictors W. Some curves of e_1 are given in Figure 1 for $\rho^* = 0.3$. The results clearly show that if the coefficient I is large enough, then even for a smaller sampling fraction r, the semiparametric estimator $\hat{\beta}_{sp}$ is still much more efficient than the simple MLE $\hat{\beta}_s$.

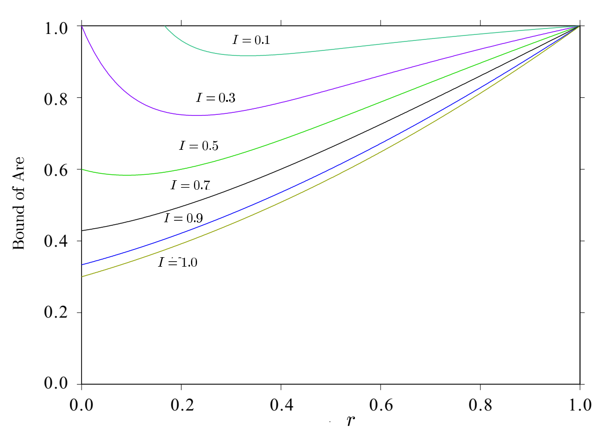

Figure 1. For $\rho^* = 0.3$, curves of $e_1(\rho^*, I, r)$ at various I.

Similar results can be derived for the ARE of $\hat{\beta}_{sp}$ with respect to $\hat{\beta}_f$. We note that under the same conditions, a lower bound for the ARE of $\hat{\beta}_{sp}$ with respect to $\hat{\beta}_f$ is

$$e_2(\rho^*, I, r) = 1 - \frac{(1-r)^2 I \rho^*}{r^2 + r(1-r)I + (1-r)^2 I \rho^*}.$$

Some curves of e_2 are also given in Figure 2 for $\rho^* = 0.3$. The ARE is at least 0.60 for $r \geq 0.3$. Further, for fixed r, e_2 increases as the coefficient of reliability I decreases.

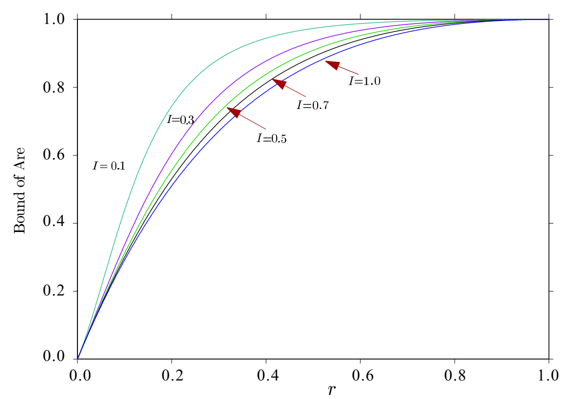

Figure 2. For $\rho^* = 0.3$, curves of $e_2(\rho^*, I, r)$ at various I.

5. Simulation Studies

In order to study the finite sample performance of the estimates, some empirical studies were carried out. The logistic regression model considered was $F(x^T\beta^0) = \exp(x)/\{1+\exp(x)\}$ so $(\beta_0^0, \beta_1^0) = (0, 1)$. A linear error model for the predictor X was assumed: $W_i = X_i + v \cdot U_i$, for some v. X_i and U_i were pseudo independent $N(0, 0.25)$ random variables. Thus given $W_i = w_i$, X_i is a $N(w_i/(1+v^2), (0.25)v^2/(1+v^2))$ variate. We also assumed that given $W_i = w_i$, the misclassification probabilities were independent of X_i and Y_i, i.e., $\theta^0(w_i) = \phi^0(w_i)$, for all w_i, where $\theta^0(w_i) = \theta(w_i, \alpha^0)$ and $\phi^0(w_i) = \phi(w_i, \alpha^0)$. Four different models were considered.

(1a) $W_i = X_i + 0.1U_i$ and $Y_i = Y_i^0$ with probability one. In this case, no misclassification occurs, and hence the only functional form needed for estimating $\hat{\beta}_m$ and $\hat{\beta}_p$ is $f(x|w;\gamma) = N(w/(1+4\gamma), \gamma/(1+4\gamma)), \gamma > 0$.

(1b) $W_i = X_i + 0.1U_i$ and $\theta^0(w_i) = \phi^0(w_i) = 0.1$. In this case, the functional forms needed for computing $\hat{\beta}_m, \hat{\beta}_p$ are $\theta(w, \alpha) = \alpha_0, \phi(w, \alpha) = \alpha_1$, and $f(x|w;\gamma) = N(w/(1+4\gamma), \gamma/(1+4\gamma)), \gamma > 0$.

(2a) $W_i = X_i + U_i$ and $Y_i = Y_i^0$ with probability one. In this case, no misclassification occurs, and hence the only functional form needed for estimating $\hat{\beta}_m$ and $\hat{\beta}_p$ is $f(x|w;\gamma) = N(w/(1+4\gamma), \gamma/(1+4\gamma)), \gamma > 0$.

(2b) $W_i = X_i + U_i$ and $\theta^0(w_i) = \phi^0(w_i) = \exp(w_i - 2.5)/\{1 + \exp(w_i - 2.5)\}$. In this case, the misclassification probabilities follow a logit model with regression coefficients $\alpha_0^0 = -2.5, \alpha_1^0 = 1$. The functional forms needed for estimating $\hat{\beta}_m$ and $\hat{\beta}_p$ are $\theta(w,\alpha) = \exp(\alpha_0 + \alpha_1 w)/\{1 + \exp(\alpha_0 + \alpha_1 w)\}, \phi(w,\alpha) = \exp(\alpha_0 + \alpha_1 w)/\{1 + \exp(\alpha_0 + \alpha_1 w)\}$, and $f(x|w;\gamma) = N(w/(1+4\gamma), \gamma/(1+4\gamma)), \gamma > 0$.

Model (1b) features small measurement error for both responses and covariates. On the other hand, Model (2b) has more severe measurement error. The primary sample size used was $n = 150$, and the sampling fractions for validation subsamples were $r = 0.2$, 0.4 and 0.6. One thousand pseudo data sets were generated to compute the simulated mean squared errors.

The function $K(\cdot)$ used in computing the nonparametric regression estimates was the Epanechnikov kernel, that is, $K(t) = (3/4)(1-t^2)I_{[-1,1]}(t)$; see Eubank (1988). Several bandwidths $h = a\hat{\sigma}_w k^{-1/3}$ were used in our simulations. Here $\hat{\sigma}_w$ is the sample standard deviation of W based on the validation data set, and a is some constant value. Such choice of bandwidth was justified by Sepanski, Knickerbocker and Carroll (1994); see also Carroll and Wand (1991) and Wang and Wang (1997).

First, we investigate the performance of $\hat{\beta}_{sp}$ for various choices of a. The performance of $\hat{\beta}_{sp} = (\hat{\beta}_{sp,0}, \hat{\beta}_{sp,1})^T$ is measured by its simulated total mean squared error (TMSE), defined to be $MSE(\hat{\beta}_{sp,0}) + MSE(\hat{\beta}_{sp,1})$. Table 1 reports the simulation results. It is seen that the performance of $\hat{\beta}_{sp}$ is not sensitive to the choice of a. This agrees with many findings for semiparametric estimation using kernel regression. Here, however, a better choice is $a = 1$ and hence we use this value for h in remaining simulations.

Table 1. Simulated TMSE of $\hat{\beta}_{sp}$ for different a

	$W = X + .1U$						$W = X + U$					
	$\theta^0(W) = 0$			$\theta^0(W) = .1$			$\theta^0(W) = 0$			$\theta^0(W) = \frac{\exp(W-2.5)}{\{1+\exp(W-2.5)\}}$		
a	$r=.2$	$r=.4$	$r=.6$	$r=.2$	$r=.4$	$r=.6$	$r=.2$	$r=.4$	$r=.6$	$r=.2$	$r=.4$	$r=.6$
0.5	.2005	.1860	.1742	.2270	.1972	.1748	.3418	.2444	.2011	.3810	.2584	.2110
1.0	.1899	.1863	.1625	.2112	.1813	.1843	.3091	.2121	.1989	.3158	.2470	.2046
1.5	.1947	.1900	.1715	.1991	.1901	.1870	.3592	.2604	.2182	.3612	.2496	.2238
2.0	.2027	.1900	.1767	.2453	.1867	.1832	.3645	.2665	.2254	.3374	.2505	.2268

Next, we compare the simulated mean squared error of different estimates for $\beta = (\beta_0, \beta_1)^T$. The results are tabulated in Table 2. Obviously, $\hat{\beta}_{sp}$ has the best overall performance among the competing estimates. Moreover, the behavior of the simulated TMSE for $\hat{\beta}_{sp}$ seems insensitive to the selection of r values. This is particularly true when the model has only minor measurement errors. Unreported calculations also show that the simulated standard errors (SE) for $\hat{\beta}_{sp,0}$ and $\hat{\beta}_{sp,1}$ are stable. Under all cases considered, the simulated SE's range from 0.166 to 0.188 for $\hat{\beta}_{sp,0}$ and from 0.366 to 0.506 for $\hat{\beta}_{sp,1}$. In general, $\hat{\beta}_{sp}$ has much better performance than the other estimates, especially when $r \leq 0.4$. From Table 2, we also see that the estimates $\hat{\beta}_f$ and $\hat{\beta}_p$ are quite competitive and better than the simple estimate $\hat{\beta}_s$. The latter statement is especially clear when $r = 0.2$. However, the simulated SE's for $\hat{\beta}_{f,0}$, $\hat{\beta}_{p,0}$ and $\hat{\beta}_{s,0}$ are not very different. These values range from 0.219 to 0.421. On the other hand, the simulated SE's for $\hat{\beta}_{f,1}$, $\hat{\beta}_{p,1}$ are smaller than those of $\hat{\beta}_{s,1}$, particularly when $r = 0.2$. The largest simulated SE for $\hat{\beta}_{f,1}$ and $\hat{\beta}_{p,1}$ is 0.894 compared with the corresponding value 1.292 for $\hat{\beta}_{s,1}$.

Table 2. Simulated TMSE of different estimators

		$\hat{\beta}_s$	$\hat{\beta}_m$	$\hat{\beta}_p$	$\hat{\beta}_f$	$\hat{\beta}_{sp}$
$W = X + .1U$						
$\theta^0(W) = 0.$	$r = .2$	1.2917	14.6527	.8332	.8318	.1899
	$r = .4$	.4917	6.4017	.4144	.4570	.1863
	$r = .6$	.2840	1.5523	.2808	.2799	.1625
$\theta^0(W) = 0.1$	$r = .2$	1.2917	13.6781	.8721	.8039	.2112
	$r = .4$	.4917	4.7618	.4284	.4488	.1813
	$r = .6$	.2840	2.4016	.2967	.2987	.1843
$W = X + U$						
$\theta^0(W) = 0.$	$r = .2$	1.2917	17.0013	.8311	.8945	.3091
	$r = .4$	.4917	5.3543	.4606	.4339	.2121
	$r = .6$	.2840	1.4972	.2697	.2476	.1989
$\theta^0(W) = \frac{\exp(W-2.5)}{1+\exp(W-2.5)}$	$r = .2$	1.2917	36.7846	.8622	.8203	.3158
	$r = .4$	.4917	7.0655	.3740	.3992	.2470
	$r = .6$	.2840	2.2060	.2754	.2530	.2046

Finally, we comment on the comparison between $\hat{\beta}_m$ and $\hat{\beta}_p$. In general, $\hat{\beta}_p$ is better than $\hat{\beta}_m$ unless $r = 0.6$. This happens because there are more parameters to be estimated simultaneously in computing $\hat{\beta}_m$, and the validation subsample size k is not large enough. The largest simulated SE is 3.103 for $\hat{\beta}_{m,1}$ and 1.769 for $\hat{\beta}_{m,0}$, showing that the estimates $\hat{\beta}_m$ are not very stable. The performance of $\hat{\beta}_m$ and $\hat{\beta}_p$ are expected to become better if k is sufficiently large and the parametric functions $\theta(w, \alpha)$, $\phi(w, \alpha)$ and $f_{X|W}(x|w; \gamma)$ are correctly modeled. Cheng and Hsueh (1999) reported that the performances of $\hat{\beta}_m$ and $\hat{\beta}_p$ depend heavily on the correct choice of the models for misclassification probabilities when $X = W$ with probability one. Therefore, $\hat{\beta}_m$ and $\hat{\beta}_p$ are not very robust and extra care needs to be taken to apply the parametric methods.

Acknowledgements

This research was supported in part by the National Science Council, R.O.C. The authors thank the Editor, an associate editor and referees for their useful comments and suggestions which greatly improve the presentation of this paper.

Appendix

Proof of Theorem 1. By Taylor's Theorem and for $n \to \infty$, we have

$$\sqrt{n}(\hat{\beta}_{sp}-\beta^0)=\left[\frac{1}{n}\left\{-\frac{\partial^2\hat{l}(\beta^0)}{\partial\beta^2}\right\}-\frac{1}{2n}\left\{-\frac{\partial^3\hat{l}(\beta^*)}{\partial\beta^3}\right\}(\hat{\beta}_{sp}-\beta^0)\right]^{-1}\frac{1}{\sqrt{n}}\left\{\frac{\partial\hat{l}(\beta^0)}{\partial\beta}\right\}+o_p(1),$$

where β^* is some quantity such that $|\beta^*-\beta^0| \le |\hat{\beta}_{sp}-\beta^0|$. Here $\hat{l}$ is the pseudo-likelihood, and

$$\frac{\partial\hat{l}(\beta^0)}{\partial\beta}=\sum_{i=1}^{k}\left\{\frac{\partial l_{1,i}(\beta^0)}{\partial\beta}\right\}+\sum_{j=k+1}^{n}\left\{\frac{\partial\hat{l}_{2,j}(\beta^0)}{\partial\beta}\right\},$$

$$\frac{\partial l_{1,i}(\beta^0)}{\partial\beta}=\{Y_i-F(X_i^T\beta^0)\}X_i,$$

$$\frac{\partial\hat{l}_{2,j}(\beta^0)}{\partial\beta}=\frac{[\hat{A}_1(Y_j^0,W_j)-\hat{A}_2(Y_j^0,W_j)]\hat{E}[F(X_j^T\beta^0)\bar{F}(X_j^T\beta^0)X_j \mid W_j]}{\hat{E}\{F(X_j^T\beta^0)\mid W_j\}\hat{A}_1(Y_j^0,W_j)+\hat{E}\{\bar{F}(X_j^T\beta^0)\mid W_j\}\hat{A}_2(Y_j^0,W_j)}.$$

Further,

$$-\frac{\partial^2\hat{l}(\beta^0)}{\partial\beta^2}=\sum_{i=1}^{k}\left\{-\frac{\partial^2 l_{1,i}(\beta^0)}{\partial\beta^2}\right\}+\sum_{j=k+1}^{n}\left\{-\frac{\partial^2\hat{l}_{2,j}(\beta^0)}{\partial\beta^2}\right\},$$

$$-\frac{\partial^2 l_{1,i}(\beta^0)}{\partial\beta^2}=F(X_i^T\beta^0)\bar{F}(X_i^T\beta^0)X_iX_i^T,$$

$$\begin{aligned}&-\frac{\partial^2\hat{l}_{2,j}(\beta^0)}{\partial\beta^2}\\&=\left(\frac{[\hat{A}_1(Y_j^0,W_j)-\hat{A}_2(Y_j^0,W_j)]\hat{E}[F(X_j^T\beta^0)\bar{F}(X_j^T\beta^0)X_j\mid W_j]}{\hat{E}\{F(X_j^T\beta^0)\mid W_j\}\hat{A}_1(Y_j^0,W_j)+\hat{E}\{\bar{F}(X_j^T\beta^0)\mid W_j\}\hat{A}_2(Y_j^0,W_j)}\right)^{\otimes 2}\\&\quad-\frac{[\hat{A}_1(Y_j^0,W_j)-\hat{A}_2(Y_j^0,W_j)]\hat{E}[\{1-2F(X_j^T\beta^0)\}F(X_j^T\beta^0)\bar{F}(X_j^T\beta^0)X_j\mid W_j]}{\hat{E}\{F(X_j^T\beta^0)\mid W_j\}\hat{A}_1(Y_j^0,W_j)+\hat{E}\{\bar{F}(X_j^T\beta^0)\mid W_j\}\hat{A}_2(Y_j^0,W_j)}.\end{aligned}$$

By the Law of Large Numbers,

$$\begin{aligned}
&\frac{1}{n}\left\{-\frac{\partial^2 \hat{l}(\beta^0)}{\partial\beta^2}\right\} \\
&= \frac{1}{n}\left\{-\frac{\partial^2 l(\beta^0)}{\partial\beta^2}\right\} + \left[\frac{1}{n}\left\{-\frac{\partial^2 \hat{l}(\beta^0)}{\partial\beta^2}\right\} - \frac{1}{n}\left\{-\frac{\partial^2 l(\beta^0)}{\partial\beta^2}\right\}\right] \\
&= rE[F(X^T\beta^0)\bar{F}(X^T\beta^0)\}XX^T] \\
&\quad +(1-r)E\left\{\frac{\{1-\theta^0(W)-\phi^0(W)\}^2}{\pi^0(W)\{1-\pi^0(W)\}}(E[F(X^T\beta^0)\bar{F}(X^T\beta^0)X \mid W])^{\otimes 2}\right\} \\
&\quad +o_p(1) + O_p(\frac{1}{n\sqrt{h}} + \frac{1}{\sqrt{n}} + h^2) \\
&= \Sigma_f^{-1} + o_p(1), \quad n\to\infty,\ h\to 0 \text{ and } n^2h\to\infty.
\end{aligned}$$

Then as $n\to\infty$, $h\to 0$, $n^2h\to\infty$, $\sqrt{n}(\hat{\beta}_{sp}-\beta^0) = \{\Sigma_f + o_p(1)\}\left\{\frac{\partial\hat{l}(\beta^0)}{\partial\beta}\right\}/\sqrt{n} + o_p(1)$. Further, by Taylor's Theorem again, as $n\to\infty$,

$$\sum_{i=k+1}^{n}\left\{\frac{\partial \hat{l}_{2,j}(\beta^0)}{\partial\beta}\right\} = \sum_{i=k+1}^{n}\left\{\frac{\partial l_{2,j}(\beta^0)}{\partial\beta} + H_{n,j}\right\} + O_p(\frac{1}{nh^{\frac{3}{2}}} + nh^4 + \frac{1}{h} + \sqrt{n}h^{\frac{3}{2}}),$$

where $H_{n,j} = \frac{1}{k}\sum_{i=1}^{k} h_{i,j}$, and

$$\begin{aligned}
h_{i,j} = \frac{K_h(W_i-W_j)}{f(W_j)}&\left(\left\{\frac{\partial^2 l_{2,j}(\beta^0)}{\partial\beta\partial\theta^0(W_j)}\right\}\frac{Y_i\{1-Y_i^0-\theta^0(W_j)\}}{E\{F(X_j^T\beta^0)\mid W_j\}}\right. \\
&+\left\{\frac{\partial^2 l_{2,j}(\beta^0)}{\partial\beta\partial\phi^0(W_j)}\right\}\frac{(1-Y_i)\{Y_i^0-\phi^0(W_j)\}}{E\{\bar{F}(X_j^T\beta^0)\mid W_j\}} \\
&+\left\{\frac{\partial^2 l_{2,j}(\beta^0)}{\partial\beta\partial E\{F(X_j^T\beta^0)\mid W_j\}}\right\}\Big[F(X_i^T\beta^0) - E\{F(X_j^T\beta^0)\mid W_j\}\Big] \\
&+\left\{\frac{\partial^2 l_{2,j}(\beta^0)}{\partial\beta\partial E[F(X_j^T\beta^0)\bar{F}(X_j^T\beta^0)X_j\mid W_j]}\right\} \\
&\left.\times\Big[F(X_i^T\beta^0)\bar{F}(X_i^T\beta^0)X_i - E[F(X_j^T\beta^0)\bar{F}(X_j^T\beta^0)X_j\mid W_j]\Big]\right).
\end{aligned}$$

As a consequence, for $n\to\infty$,

$$\begin{aligned}
\frac{1}{\sqrt{n}}\left\{\frac{\partial\hat{l}(\beta^0)}{\partial\beta}\right\} &= \frac{\sqrt{n}}{k(n-k)}\sum_{i=1}^{k}\sum_{j=k+1}^{n}\left[\frac{k}{n}\left\{\frac{\partial l_{1,i}(\beta^0)}{\partial\beta}\right\} + \frac{n-k}{n}\left\{\frac{\partial l_{2,j}(\beta^0)}{\partial\beta}\right\} + \frac{n-k}{n}h_{i,j}\right] \\
&\quad + O_p(\frac{1}{n^{\frac{3}{2}}h^{\frac{3}{2}}} + n^{\frac{1}{2}}h^4 + \frac{1}{n^{\frac{1}{2}}h} + h^{\frac{3}{2}}).
\end{aligned}$$

Define two i.i.d. random vectors $Z_i = (X_i, W_i, Y_i, Y_i^0)$, $i = 1, \ldots, k$ and $Z_j^* = (W_j, Y_j^0)$, $j = k+1, \ldots, n$ and let

$$Q_n(Z_i; Z_j^*) = \frac{k}{n}\left\{\frac{\partial l_{1,i}(\beta^0)}{\partial\beta}\right\} + \frac{n-k}{n}\left\{\frac{\partial l_{2,j}(\beta^0)}{\partial\beta}\right\} + \frac{n-k}{n} h_{i,j}.$$

Then $k^{-1}(n-k)^{-1}\sum_{i=1}^{k}\sum_{j=k+1}^{n} Q_n(Z_i; Z_j^*)$ is a generalized U-statistic. Applying the Central Limit Theorem for generalized U-statistics (see Serfling(1980)), we have, for $n \to \infty$,

$$\begin{aligned}
&\frac{\sqrt{n}}{k(n-k)}\sum_{i=1}^{k}\sum_{j=k+1}^{n} Q_n(Z_i; Z_j^*) \\
&= \sqrt{n}\left[\frac{1}{k}\sum_{i=1}^{k} E\{Q_n(Z_i; Z_j^*) \mid Z_i\} + \frac{1}{n-k}\sum_{i=k+1}^{n} E\{Q_n(Z_i; Z_j^*) \mid Z_j^*\}\right] + o_p(1) \\
&\xrightarrow{d} N\left(0, \Sigma_f^{-1} + \frac{(1-r)^2}{r} I_{sp}\right),
\end{aligned}$$

$$\begin{aligned}
&E\{Q_n(Z_i; Z_j^*) \mid Z_i\} \\
&= \frac{k}{n}\left\{\frac{\partial l_{1,i}(\beta^0)}{\partial\beta}\right\} + \frac{n-k}{n}\frac{\{1-\theta^0(W_i)-\phi^0(W_i)\}E[F(X_i^T\beta^0)\bar{F}(X_i^T\beta^0)X_i \mid W_i]}{\pi^0(W_i)\{1-\pi^0(W_i)\}} \\
&\times \Big(Y_i\{1-Y_i^0-\theta^0(W_i)\} - (1-Y_i)\{Y_i^0-\phi^0(W_i)\} \\
&\quad -\{1-\theta^0(W_i)-\phi^0(W_i)\}[F(X_i^T\beta^0) - E\{F(X_i^T\beta^0) \mid W_i\}]\Big) + O_p(h^2),
\end{aligned}$$

$$E\{Q_n(Z_i; Z_j^*) \mid Z_j^*\} = \frac{n-k}{n}\left\{\frac{\partial l_{2,j}(\beta^0)}{\partial\beta}\right\}.$$

Consequently, as $n \to \infty$, $\sqrt{n}(\hat{\beta}_{sp} - \beta^0) \xrightarrow{d} N\left(0, \Sigma_f + \frac{(1-r)^2}{r}\Sigma_f I_{sp}\Sigma_f\right)$.

References

Albert, P. S., Hunsberger, S. A. and Bird, F. M. (1997). Modeling repeated measures with monotonic ordinal responses and misclassification, with applications to studying maturation. *J. Amer. Statist. Assoc.* **92**, 1304-1311.

Bollinger, C. R. and David, M. H. (1997). Modeling discrete choice with response error: Food stamp participation. *J. Amer. Statist. Assoc.* **92**, 827-835.

Carroll, R. J., Spiegelman, C. H., Lan, K. K. G., Bailey, K. T. and Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika* **71**, 19-26.

Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *J. Roy. Statist. Soc. Ser. B* **53**, 573-585.

Cheng, K. F. and Hsueh, H. M. (1999). Correcting bias due to misclassification in the estimation of logistic regression models. *Statist. Probab. Lett.* **44**, 229-240.

Cox, D. R. (1970). *The Analysis of Binary Data*. Chapman and Hall, London.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New York.

Gasser, T. and Müller, H. G. (1979). Kernel Estimation of Regression Functions in *Smoothing Techniques for Curve Estimation*. (Edited by T. Gasser and M. Rosenblatt), 23-68. Lecture Notes in Mathematics 757, Springer-Verlag, Berlin.

Golm, G. T., Halloran, M. E. and Longini, I. M. Jr. (1998). Semi-parametric models for mismeasured exposure information in vaccine trials. *Statist. Medicine* **17**, 2335-2352.

Golm, G. T., Halloran, M. E. and Longini, I. M. Jr. (1999). Semiparametric methods for multiple exposure mismeasurement and bivariate outcome in HIV vaccine trials. *Biometrics* **55**, 94-101.

Lawless, J. F., Kalbfleisch, J. D. and Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *J. Roy. Statist. Soc. Ser. B* **61**, 413-438.

Pepe, M. S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika* **79**, 355-365.

Pepe, M. S., Reilly, M. and Fleming, T. R. (1994). Auxiliary outcome data and the mean score method. *J. Statist. Plann. Inference* **42**, 137-160.

Pregibon, D. (1981). Logistic regression diagnostics. *Ann. Statist.* **9**, 705-724.

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

Sepanski, J. H., Knickerbocker, R. and Carroll, R. J. (1994). A semiparametric correction for attenuation. *J. Amer. Statist. Assoc.* **89**, 1366-1373.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley, New York.

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

Graduate Institute of Statistics, National Central University, Chungli, Taiwan, R.O.C.

E-mail: kfcheng@cc.ncu.edu.tw

Department of Statistics, National Chengchi University, Taipei, Taiwan, R.O.C.

E-mail: hsueh@nccu.edu.tw

(Received April 2000; accepted July 2002)