Logistic Regression Model with Surrogate Covariate

by **Q. T. Thach** and **N. G. N. Prasad**

Q. T. Thach
Department of Community Medicine
The University of Hong Kong
Patrick Manson Building South Wing
7 Sassoon Road
Hong Kong

N. G. N. Prasad
Department of Mathematical Sciences
University of Alberta
Edmonton, Alberta, T6G–2G1
Canada

In the present paper, we propose an alternative estimator under a logistic regression model with a surrogate variable using the empirical likelihood technique. This estimator is shown to be asymptotically normal and more efficient relative to the partial, imputation and bootstrap estimators.

---

1 Introduction

Logistic models are often used to model the conditional mean of a binary response Y with a covariate of interest X. That is, the conditional mean of $Y|X$ is modeled as $F(\beta_0 + \beta_1 X)$ where β_0 and β_1 are unknown parameters and $F(x) = (1+e^{-x})^{-1}$. In such applications, values of the covariate X for a subset study subjects may be missing; either the measurement is difficult or expensive to obtain. A closely related variable Z may be used as a surrogate for the covariate X. For example, consider the study done by Gladen and Rogan (1979). They examine the disease risk due to body burden of accumulated chemical pollutants in body tissues. Two classes of environmental pollutants which exhibit this "accumulation" phenomenon are the metals, such as DDT's, PCB's and PBB's. Body burden is measured by the levels of the chemicals. For the metals, depot tissues are teeth and bones. For the halogenated hydrocarbon, fat is the depot tissue. The depot tissue is usually impossible or difficult to obtain from living subjects. As a result, a surrogate measurement is necessary, such as blood levels. However, blood levels are usually lower than depot tissue and are affected by nutritional or metabolic state of the individual. Use of such measurements is, therefore, subject to criticism. One approach to this situation is to obtain the data in the form of two independent samples. In the first sample (validation data set) only the information on the response variable Y and the surrogate variable Z are measured. While in the second one (primary data set) information on the covariate X is measured in addition to the information on the response Y and surrogate variable Z. The use of surrogate variables is common, particularly in medical research, and there has been considerable discussion to identify "valid" surrogates. For a review of the use of surrogate variables in clinical trials, see Prentice (1989) and Wittes, Lakatos, and Probstfield (1989).

Although relatively few papers have addressed the missing value problem specifically in the context of logistic regression, there are four general methods for the analysis of incomplete data with surrogate variable that can be widely used, namely, partial case, imputation, maximum likelihood and semi-parametric methods. Perhaps the simplest approach to this problem is the partial case method, which discards cases with missing values. This is the default method used by most statistical software packages such as SAS and SPSS. Since we can measure the outcome and surrogate variables for the discarded units, these units still carry some information on the effect of the covariate. Hence, partial case analysis is not efficient for not using all the available information. Especially, large missing rate in the covariate can add up to a substantial loss of data. A second general approach is to replace the missing values with reasonable estimates (imputed values) and then analyze the data. Several strategies to construct such estimates have been suggested. However, estimates of the variance of the estimated regression parameters from the artificially completed data set are invalid in general. This is because variance estimates have to be corrected for variations due to imputation. One solution to this problem is to assess this variance by computing repeated estimates; following multiple imputation method, see Rubin (1987) and Kalton and Kasprzyk (1986).

A third general approach is to parameterize the conditional probability relationship between X and Z through model $P_{\boldsymbol{\eta}}(X|Z)$ and to maximize the likelihood

$$L(\boldsymbol{\beta}, \boldsymbol{\eta}) = \prod_{i \in S_m} P_{\boldsymbol{\beta}}(Y_i|X_i) P_{\boldsymbol{\eta}}(X_i|Z_i) \prod_{i \in S_{n-m}} P_{\boldsymbol{\beta},\boldsymbol{\eta}}(Y_i|Z_i),$$

where $P_{\boldsymbol{\beta},\boldsymbol{\eta}}(Y_i|Z_i) = \int P_{\boldsymbol{\beta}}(Y_i|X_i) P_{\boldsymbol{\eta}}(X_i|Z_i) dX$, $\boldsymbol{\beta} = (\beta_0, \beta_1)$, S_m and S_{n-m} denote the primary and the validation sets, respectively. However, this parametric method is not generally used in applied work, in part, because misspecification of the nuisance function $P_{\boldsymbol{\eta}}(X|Z)$ can lead to an inconsistent estimator of $\boldsymbol{\beta}$. Moreover, except for some special cases, implementation of the likelihood based approach is cumbersome; requiring either numerical integration to calculate $P_{\boldsymbol{\beta},\boldsymbol{\eta}}(Y|Z)$ and its derivative, or other complicated algorithms such as expectation maximization (EM) and data augmentation algorithm. For more details, see Schafer (1987) and Tanner and Wong (1987).

The fourth general method for analyzing incomplete data with surrogate variable is to use a non-parametric kernel regression method on the validation data set S_{n-m} to estimate the probability $P_{\boldsymbol{\beta}}(Y|Z)$ (see Carroll and Wand, 1991). They proposed an estimate of $\boldsymbol{\beta}$ as a solution to

$$\sum_{i \in S_m} S_{\boldsymbol{\beta}}(Y_i|X_i) + \sum_{i \in S_{n-m}} \hat{H}_{\boldsymbol{\beta}}(Y_i|Z_i) = \mathbf{0},$$

where $S_{\boldsymbol{\beta}}(Y|X)$ was the score function of $(Y|X)$ and $\hat{H}_{\boldsymbol{\beta}}(Y|Z)$ was a kernel regression estimate of $(Y|Z)$. A semi-parametric estimate of $\boldsymbol{\beta}$ based on this method was asymptotically normally distributed. Although this method is generally more robust than the others, it has the disadvantage of requiring a bandwidth selection. Pepe and Fleming (1991) considered a similar problem with discrete covariate Z. Stefanski and Carroll (1985) discussed the case in which all the X_i's were unobserved while $\{Z_i, i \in S_n\}$ were available.

Mak, Li, and Kuk (1986) assumed a model for $P_{\boldsymbol{\eta}}(X|Z)$ and then proposed a bias-corrected estimator of the form $\hat{\boldsymbol{\beta}}_I - \boldsymbol{c}_0$, where $\hat{\boldsymbol{\beta}}_I$ was an estimator based on the imputation method and $\boldsymbol{c}_0$ was an estimate of bias obtained from a bootstrap method. However, in their bootstrap procedure, the information on Y and Z in the validation data set is not being used in the resampling process. Furthermore, it is non-robust with respect to a misspecification of the conditional density $P_{\boldsymbol{\eta}}(X|Z)$. For these reasons, the resulting bootstrap procedure is questionable and hence, can lead to an inefficient estimator.

In the present paper, we propose an alternative estimator under a logistic regression model with a surrogate variable using the empirical likelihood technique. This estimator is shown to be asymptotically normal and more efficient relative to the partial, imputation and bootstrap estimators.

Section 2 of this paper presents the logistic regression model with surrogate covariate and the three methods of estimation. In Section 3, a brief introduction is given to the empirical likelihood and it is shown, explicitly, how empirical likelihood can be applied to the present problem. Section 4 derives the asymptotic variance of partial, imputation and empirical likelihood estimators. Some simulation results that compare these methods are given in Section 5.

2 The model

Let Y denote a binary outcome variable, X be a covariate of interest and $P_\beta(Y|X) = F(\beta_0 + \beta_1 X)$ be the logistic regression model for the conditional distribution of Y given X. In the remainder of this paper, we consider the case $\beta_0 = 0$ and without lost of generality we let $\beta = \beta_1$. The objective is to estimate the parameter β when some units of X are missing. The data sets available for analysis consist of m observations $\{(Y_i, X_i, Z_i), i \in S_m\}$ and $n - m$ observations $\{(Y_i, Z_i), i \in S_{n-m}\}$, where Z_i is the measurement on the surrogate variable Z for the i-th unit, $i \in S_n$ with $S_n = S_m \cup S_{n-m}$. We assume that the validation set, S_{n-m}, is a simple random sample from S_n.

2.1 Partial case method

The partial case method estimates the logistic parameter β by maximizing the likelihood function of m complete cases of X, ignoring Z. This likelihood function is written as

$$L_m(\beta|X_1,\ldots,X_m) = \prod_{i\in S_m} F(\beta x_i)^{y_i}[1-F(\beta x_i)]^{1-y_i}. \tag{2.1}$$

In practice, the partial case estimate of β, denoted by $\hat{\beta}_m$, and its estimated standard error can be computed by using some standard statistical packages.

2.2 Imputation method

This method involves imputing missing values of X for units in S_{n-m} with predicted values obtained from a simple regression model $X = a + bZ + \varepsilon$, where ε denotes a random vector with mean 0 and variance σ^2. That is, $\{X_i, i \in S_{n-m}\}$ are imputed by

$$\hat{X}_i = \hat{a} + \hat{b}Z_i, \quad i = m+1, m+2, \ldots, n, \tag{2.2}$$

where $\hat{a}$ and $\hat{b}$ are the least square estimators based on $\{X_i, Z_i; i \in S_m\}$. It is common practice to treat these imputed values as if they are true values and then compute the variance estimate of β using standard likelihood theory. This procedure can lead to serious underestimation of the true variance of the estimate when the proportion of missing values is appreciable. As a result, the confidence interval based on the resulting estimate will have coverage probability smaller than its corresponding nominal level since the method ignores errors in the estimation of X from Z. To describe this method, consider the likelihood function,

$$L_I(\beta|X_1,\ldots,X_m,\hat{X}_{m+1},\ldots,\hat{X}_n) = \prod_{i\in S_n} F(\beta\tilde{x}_i)^{y_i}[1-F(\beta\tilde{x}_i)]^{1-y_i}, \tag{2.3}$$

where

$$\tilde{x}_i = \begin{cases} x_i & i \in S_m \\ \hat{x}_i & i \in S_{n-m}. \end{cases}$$

The estimate of β, denoted by $\hat{\beta}_I$, is then obtained by maximizing the likelihood equation (2.3). It can also be noted that the above estimator will be biased due to imputation of X_i's.

2.3 Bootstrap method

Mak *et al.* (1986) proposed a bootstrap procedure to estimate the incurred bias $\hat{\beta}_I$ due to imputation. They suggested a bias-corrected estimator $\hat{\beta}_B = \hat{\beta}_I - c_0$, where c_0 was a correction for the bias induced by the bootstrap sampling. To describe their bootstrap sampling, let $\hat{G}$ and $\hat{H}$ be the empirical probability distributions with mass m^{-1} each at $\{X_i, i \in S_m\}$ and $\{\hat{\varepsilon}_i = X_i - \hat{a} - \hat{b}Z_i, i \in S_m\}$, respectively, where $\hat{a}$ and $\hat{b}$ are the least square estimators based on $\{X_i, Z_i; i \in S_m\}$.

1. Draw a sample $\{X_i^*, i \in S_m\}$ from $\hat{G}$ and generate

$$Y_i^* = \begin{cases} 0 & \text{with probabilty} \quad 1 - F(\hat{\beta}_m X_i^*) \\ 1 & \text{with probabilty} \quad F(\hat{\beta}_m X_i^*), \end{cases}$$

   where $\hat{\beta}_m$ is the partial case estimator obtained from the logistic regression analysis based on $\{Y_i, X_i; i \in S_m\}$.

2. Draw a sample $\{\hat{\varepsilon}_i^*, i \in S_n\}$ from $\hat{H}$ and let $\{X_i^* = \hat{a} - \hat{b}Z_i^* + \hat{\varepsilon}_i^*, i \in S_n\}$.

3. Then the "bootstrap sample " will consist of $\{Y_i^*, X_i^*, Z_i^*; i \in S_m\}$ and $\{Y_i^*, Z_i^*; i \in S_{n-m}\}$.

4. Compute $\hat{\beta}_I$ using the imputation method outlined in Section 2.2.

5. Repeat steps (1)–(4) above k times to obtain $c_0 = k^{-1} \sum_{h=1}^{k} \hat{\beta}_I^{*h} - \hat{\beta}_m$, where $\hat{\beta}_I^{*h}$ is the value of the estimator $\hat{\beta}_I$ computed on the h-th bootstrap sample, $h = 1, \ldots, k$.

Note that in their bootstrap algorithm described above, $\{Y_i, X_i; i \in S_{n-m}\}$ are not being used in the resampling procedure. It is often the case that the number of units in S_{n-m} is much larger than the number of units in S_m; therefore, the procedure can lead to an unstable variance estimator.

3 The empirical likelihood method

For the present problem, we employ the empirical likelihood method to use all the information from both S_m and S_{n-m} through the following constraint

$$\sum_{i \in S_n} S_\theta(Y_i | Z_i) = 0, \tag{3.1}$$

where $S_\theta(Y|Z) = Z[Y - F(\theta Z)]$ is the score function obtained from the likelihood function

$$L(\theta|Y_1, \ldots, Y_n; Z_1, \ldots, Z_n) = \prod_{i \in S_n} F(\theta z_i)^{y_i}[1 - F(\theta z_i)]^{1-y_i}. \tag{3.2}$$

Here, the goodness-of-fit of the logistic regression model of Y on Z is not relevant. The idea of fitting the above model is only to extract association between X and Y through the associated information between Z and Y when X and Z are correlated.

3.1 The proposed method

We apply empirical likelihood method for the model $P_\beta(Y|X) = F(\beta X)$ by maximizing the conditional likelihood

$$L(p_1, \ldots, p_m|S_n) = \prod_{i \in S_m} p_i, \tag{3.3}$$

with respect to p_i, $i = 1, \ldots, m$ subject to restrictions

$$\sum_{i \in S_m} p_i = 1, \quad p_i \geq 0, \quad i \in S_m \quad \text{and} \quad \sum_{i \in S_m} S_\theta(Y_i|Z_i)p_i = 0. \tag{3.4}$$

where $S_\theta(Y|Z) = Z[Y - F(\theta Z)]$. The last restriction follows from the fact that

$$\sum_{i \in S_n} S_\theta(Y_i|Z_i) = 0. \tag{3.5}$$

Then the maximum of $\log L(p_1, p_2, \ldots, p_m|S_n)$ may be found via Lagrange multipliers by letting

$$H = \sum_{i \in S_m} \log p_i + \lambda_1 \left(1 - \sum_{i \in S_m} p_i\right) - m\lambda_2 \sum_{i \in S_m} p_i S_{\hat{\theta}}(Y_i|Z_i), \tag{3.6}$$

where the λ's are Lagrange multipliers and $\hat{\theta}$ is the maximum likelihood estimator of θ obtained as the solution to the equation (3.5). Taking the derivatives with respect to p_i, we have

$$\frac{\partial H}{\partial p_i} = \frac{1}{p_i} - \lambda_1 - m\lambda_2 S_{\hat{\theta}}(Y_i|Z_i) = 0. \tag{3.7}$$

Hence,

$$\sum_{i \in S_m} p_i \frac{\partial H}{\partial p_i} \quad = m - \lambda_1 = 0 \Rightarrow \lambda_1 = m. \tag{3.8}$$

Replacing $\lambda_1 = m$ in (3.7), we have

$$p_i = \left(\frac{1}{m}\right) \frac{1}{1 + \lambda_2 S_{\hat{\theta}}(Y_i|Z_i)}, \quad i \in S_m. \tag{3.9}$$

Now, the restriction from the third part of (3.4) is

$$0 = \sum_{i \in S_m} p_i S_{\hat{\theta}}(Y_i|Z_i) = \left(\frac{1}{m}\right) \sum_{i \in S_m} \frac{S_{\hat{\theta}}(Y_i|Z_i)}{1 + \lambda_2 S_{\hat{\theta}}(Y_i|Z_i)}. \tag{3.10}$$

from which λ_2 and hence the p_i's can be obtained . After obtaining the optimal p_i's, $i = 1 \ldots, m$ we obtain the empirical likelihood estimator of β from the estimating equation

$$S_m(\beta) = m \sum_{i \in S_m} S_\beta(Y_i|X_i) p_i = 0, \tag{3.11}$$

where $S_\beta(Y|X) = X[Y - F(\beta X)]$.

The solution $\hat{\beta}_E$ to equation (3.11) can be evaluated by implementing a root finding algorithm such as Brent's method (see Press, Flannery, Teukolsky, and Vetterling, 1993). In the next section, we consider the asymptotic variance of $\hat{\beta}_E$ along with the other estimators.

4 Asymptotic variances

This section is devoted to the derivation of the asymptotic variance of $\hat{\beta}_m$, $\hat{\beta}_I$ and $\hat{\beta}_E$. In addition, the asymptotic variance of the maximum likelihood estimator will also be given for the case when all the X_i's, $i \in S_n$ are observed. In the rest of this paper, the asymptotic results are obtained by letting $m \to \infty$ and $n \to \infty$ such that $m/n \to k'$ where $k' \in (0, 1)$. Further, we assume that there exists a positive constant C such that $|X_i| \leq C$ and $|Z_i| \leq C$ for all i. The asymptotic distribution theory in a general case with density $f_Y(y; \beta)$ relies upon the following assumptions (see Cox and Hinkley, 1974 and Pepe, Reilly, and Fleming, 1994).

(a) The parameter space Ω has finite dimension, is closed and compact, and the true parameter value β is interior to Ω.

(b) The first three derivatives of the log likelihood $l(Y;\beta)$ with respect to β exist in the neighborhood, N_0, of the true parameter value almost surely. Further, in such a neighborhood, n^{-1} times the absolute value of the third derivative is bounded above by a function of Y whose expectation exists. The absolute value of the third derivative of the log likelihood $l(Y;\beta)$ with respect to β is bounded away from 0 in a neighborhood, N_0, almost surely.

(c) $-E\{(\partial^2 l(Y;\beta)/\partial^2\beta)\}$ is finite and positive in the neighborhood, N_0, of the true parameter β.

It can be noted that for the model considered in this paper, the above conditions hold. The first and second derivatives with respect to β of the log-likelihood $l_m(\beta)=\log L_m(\beta|X_1,\dots,X_m)$ defined in (2.1) are given by

$$\frac{\partial l_m(\beta)}{\partial\beta}=\sum_{i\in S_m}x_i[y_i-F(\beta x_i)],\qquad \frac{\partial^2 l_m(\beta)}{\partial\beta^2}=-\sum_{i\in S_m}x_i^2F(\beta x_i)[1-F(\beta x_i)].$$

From standard likelihood theory, the asymptotic (unconditional) variance of $\hat\beta_m$, $Var(\hat\beta_m)$, is then given by

$$\begin{aligned} Var(\hat\beta_m) &= -\left\{E\left[\frac{\partial^2 l_m(\beta)}{\partial\beta^2}\right]\right\}^{-1} \\ &= \frac{1}{m}\left\{\frac{1}{n}\sum_{i\in S_n}x_i^2F(\beta x_i)[1-F(\beta x_i)]\right\}^{-1}. \end{aligned} \tag{4.1}$$

If all the X_i's, $i\in S_n$ were observed, the maximum likelihood estimator of β could be obtained by maximizing the function $L(\beta|X_1,\dots,X_n)$ with respect to β. We denote the resulting estimator of β by $\hat\beta_C$. Then the asymptotic variance of $\hat\beta_C$ is given by

$$Var(\hat\beta_C)=\frac{1}{n}\left\{\frac{1}{n}\sum_{i\in S_n}x_i^2F(\beta x_i)[1-F(\beta x_i)]\right\}^{-1}. \tag{4.2}$$

Turning to the asymptotic variance of $\hat\beta_I$, consider

$$Var(\hat\beta_I)=E\{Var(\hat\beta_I|S_n)\}+Var\{E(\hat\beta_I|S_n)\}. \tag{4.3}$$

Since for large m, $E(\hat{\beta}_I|S_n) \doteq \beta$ implies that $Var\{E(\hat{\beta}_I|S_n)\} \doteq 0$. Hence,

$$Var(\hat{\beta}_I) \doteq \frac{1}{n}\left\{\frac{1}{n}\left[\sum_{i\in S_m} x_i^2 F(\beta x_i)[1-F(\beta x_i)] + \sum_{i\in S_{n-m}} E\left\{\hat{x}_i^2 F(\beta \hat{x}_i)[1-F(\beta \hat{x}_i)]\right\}\right]\right\}^{-1}. \tag{4.4}$$

To establish the consistency of $\hat{\beta}_E$, we first obtain the consistency of $\tilde{\beta}_E$, which is the same as $\hat{\beta}_E$ with the condition that the p_i's are fixed known constants. We use the following result on estimating functions due to Foutz (1977).

Theorem 4.1. *There exists a unique consistent solution to an estimating equation $S_m(\beta)$ given in (3.11) in a neighborhood N_0 if*

(i) $|\partial S_m(\beta)/\partial\beta|$ exists and is continuous in a neighborhood N_0;

(ii) $m^{-1}\partial S_m(\beta)/\partial\beta$ converges uniformly in probability to $E\{m^{-1}\partial S_m(\beta)/\partial\beta\}$ in N_0;

(iii) with probability converging to 1, the quantity $\partial S_m(\beta)/\partial\beta$ evaluated at the true parameter is negative as $m \to \infty$;

(iv) $E\{S_m(\beta)\} \doteq 0$.

The next two theorems are along the lines of Theorems 3.1 and 3.2 in Pepe, Reilly, and Fleming (1994).

Theorem 4.2. *An estimator $\tilde{\beta}_E$ that satisfies equation (3.11) exists and is unique in a neighborhood N_0, with probability converging to 1 as $m \to \infty$ and $n \to \infty$ such that $m/n \to k'$ where $k' \in (0,1)$. Furthermore, $\tilde{\beta}_E$ is consistent for the true parameter β.*

Proof. First, we assume that the p_i's are fixed and let $I_\beta(Y|X) = -\partial^2 \log L(\beta|X_1,\ldots,X_m)/\partial\beta^2$. In this case, the score function is $S_m(\beta) = m\sum S_\beta(Y_i|X_i)p_i$, where $S_\beta(Y_i|X_i) = X_i[Y_i - F(\beta X_i)]$. Condition (i) of Foutz above follows from assumption (b).

Consider $m^{-1}\partial S_m(\beta)/\partial\beta) \doteq -m^{-1}\sum_{i\in S_m} x_i^2 F(\beta x_i)[1-F(\beta x_i)]$ which is the average of independent and non-identically distributed random variables. Then, the Kolmogorov strong law of large numbers for independent non-identically distributed random variables applies and yields

$m^{-1}(\partial S_m(\beta)/\partial\beta) - \{-n^{-1}\sum_{i\in S_n} x_i^2 F(\beta x_i)[1 - F(\beta x_i)]\} \xrightarrow{p} 0$ as $m \to \infty$ (see Serfling, 1980, p.27). The pointwise convergence of $\partial S_m(\beta)/\partial\beta$ can be extended to uniform convergence on a neighborhood, N_0. This follows by noting that the assumption (b) is satisfied for our model, i.e., $\partial S_m(\beta)/\partial\beta$ has bounded derivative in a neighborhood, N_0, almost surely and by the application of the dominated convergence theorem to establish that $E\{I_\beta(Y_i|X_i)\}$ also has bounded derivatives. The pointwise convergence of $m^{-1}\partial S_m(\beta)/\partial\beta$ at the true parameter value together with assumption (c) implies condition (iii) of Foutz.

Finally, turning to condition (iv) we note that

$$\begin{aligned} E\{S_m(\beta)\} &= E\{m \sum_{i\in S_m} S_\beta(Y_i|X_i)p_i\} \\ &\doteq \sum_{i\in S_m} E\{S_\beta(Y_i|X_i)\} \\ &= 0. \end{aligned} \tag{4.5}$$

$\square$

Hence the result of Theorem 4.2 follows for p_i's fixed.

We now give the asymptotic variance of $\tilde{\beta}_E$ in the following theorem.

Theorem 4.3. *As $m \to \infty$ and $n \to \infty$ such that $m/n \to k'$ where $k' \in (0,1)$, $m^{1/2}(\tilde{\beta}_E - \beta)$ converges in distribution to a normally distributed random variable such that for large m and n, $E(\tilde{\beta}_E) = 0$ and variance given by*

$$\begin{aligned} Var(\tilde{\beta}_E) = {} & \frac{1}{n}\left(\frac{1}{n}\sum_{i\in S_n} x_i^2 F(\beta x_i)[1 - F(\beta x_i)]\right)^{-1} \\ & + \left(\frac{1}{m}\right)\left(\frac{1}{n}\sum_{i\in S_n} x_i^2 F(\beta x_i)[1 - F(\beta x_i)]\right)^{-2} \\ & \left[\frac{1}{n}\sum_{i\in S_n} x_i^2 F(\beta x_i)[1 - F(\beta x_i)] - \left(\frac{1}{n}\sum_{i\in S_n} x_i z_i F(\beta x_i)[1 - F(\beta x_i)]\right)^2 \right. \\ & \left. \times \left(\frac{1}{n}\sum_{i\in S_n} z_i^2 F(\theta z_i)[1 - F(\theta z_i)]\right)^{-1}\right]. \end{aligned} \tag{4.6}$$

Proof. Consider a second order Taylor series expansion of $S_m(\tilde{\beta}_E)$ around β. Then we have

$$0 = S_m(\tilde{\beta}_E) = S_m(\beta) + \frac{\partial S_m(\beta)}{\partial \beta}(\tilde{\beta}_E - \beta) + o_p(m^{-1/2}),$$

so that

$$m^{1/2}(\tilde{\beta}_E - \beta) = [-m^{-1}\frac{\partial S_m(\beta)}{\partial \beta}]^{-1}\{m^{-1/2}S_m(\beta)\} + o_p(1).$$

Observe that $m^{-1}\partial S_m(\beta)/\partial\beta \doteq m^{-1}\sum_{i\in S_m} -I_\beta(Y_i|X_i)$ which is the mean of independent and non-identically distributed random variables. It was previously proven that $m^{-1}\partial S_m(\beta)/\partial\beta - \{-n^{-1}\sum_{i\in S_n} x_i^2\ F(\beta x_i)[1-F(\beta x_i)]\} \xrightarrow{p} 0$, which is negative. Therefore, we look at the asymptotic distribution of $S_m(\beta)$. Consider

$$S_m(\beta) = m \sum_{i\in S_m} S_\beta(Y_i|X_i)p_i, \tag{4.7}$$

which is the sum of independent and non-identically distributed random variables with mean 0. Asymptotic normality of $S_m(\beta)$ follows then from the Lindeberg-Feller central limit theorem by noting that $X_i^* = mX_i[Y_i - F(\beta X_i)]p_i$ with $EX_i^* = 0$ and for some $\nu > 2$, $B_m^{-\nu}\sum_{i=1}^m E|X_i^*|^\nu = o(1)$ where $B_m^2 = \sum_{i=1}^m E(X_i^*)^2$. Upon simplification, it is shown in the Lemma below that the variance of $S_m(\beta)$ is given by expression (4.9). It follows that the asymptotic distribution of $m^{1/2}(\tilde{\beta}_E - \beta)$ is normal with mean 0 and variance given by

$$Var(\sqrt{m}\tilde{\beta}_E) \doteq \left[-m^{-1}\frac{\partial S_m(\beta)}{\partial \beta}\right]^{-2} m^{-1}Var(S_m(\beta)), \tag{4.8}$$

which is the desired result. □

Lemma 4.1. *For large m and n, we have*

$$\begin{aligned} Var\{S_m(\beta)\} \doteq \left(\frac{m}{n}\right)^2 \sum_{i\in S_n} x_i^2 F(\beta x_i)[1-F(\beta x_i)] + m\Bigg\{\frac{1}{n}\sum_{i\in S_n} x_i^2 F(\beta x_i) \\ \times[1-F(\beta x_i)] - \left(\frac{1}{n}\sum_{i\in S_n} x_i z_i F(\beta x_i)[1-F(\beta x_i)]\right)^2 \\ \times\left(\frac{1}{n}\sum_{i\in S_n} z_i^2 F(\theta z_i)[1-F(\theta z_i)]\right)^{-1}\Bigg\}. \end{aligned} \tag{4.9}$$

Proof. To prove the lemma, we use the standard formula,

$$Var\{S_m(\beta)\} = E[Var\{S_m(\beta)\}|S_n] + Var[E\{S_m(\beta)\}|S_n]. \tag{4.10}$$

First, consider

$$\begin{aligned} S_m(\beta) &= m \sum_{i \in S_m} x_i[y_i - F(\beta x_i)]p_i \\ &\doteq \frac{m}{m} \sum_{i \in S_m} x_i[y_i - F(\beta x_i)][1 - \lambda_2 S_{\hat{\theta}}(Y_i|Z_i)], \end{aligned} \tag{4.11}$$

and observe that by the strong law of large numbers $n^{-1}\sum_{i \in S_n} S_\theta(Y_i|Z_i) \xrightarrow{p} 0$, since $E[S_\theta(Y_i|Z_i)] = 0$ and then by noting that $S_{\hat{\theta}}(Y|Z) \xrightarrow{p} S_\theta(Y|Z)$, we have

$$E\{S_m(\beta)|S_n\} \doteq \frac{m}{n} \sum_{i \in S_n} x_i[y_i - F(\beta x_i)]. \tag{4.12}$$

It follows that

$$Var[E\{S_m(\beta)\}|S_n] \doteq \left(\frac{m}{n}\right)^2 \sum_{i \in S_n} x_i^2 F(\beta x_i)[1 - F(\beta x_i)]. \tag{4.13}$$

Turning to $E[Var\{S_m(\beta)\}|S_n]$, consider

$$\begin{aligned} S_m(\beta) &= m \sum_{i \in S_m} x_i[y_i - F(\beta x_i)]p_i \\ &\doteq m \left(\frac{1}{m} \sum_{i \in S_m} y_i^* - \frac{\hat{\lambda}_2}{m} \sum_{i \in S_m} S_{\hat{\theta}}(Y_i|Z_i) y_i^* \right) \\ &\doteq m \left(\bar{y}_m^* - \frac{\bar{x}^*}{m \hat{S}_{x^*}^2} \sum_{i \in S_m} x_i^* y_i^* \right) \\ &= m(\bar{y}_m^* - \hat{\beta}^* \bar{x}_m^*), \end{aligned} \tag{4.14}$$

where

$$y_i^* = x_i[y_i - F(\beta x_i)], \quad \bar{y}_m^* = m^{-1} \sum y_i^*, \tag{4.15}$$

$$x_i^* = z_i(y_i - F(\theta z_i)], \quad \bar{x}_m^* = m^{-1} \sum x_i^*, \tag{4.16}$$

$$\hat{\beta}^* = (m\hat{S}_{x^*}^2)^{-1} \sum x_i^* y_i^*, \quad \hat{S}_{x^*}^2 = m^{-1} \sum x_i^{*2}, \tag{4.17}$$

i.e. the empirical likelihood estimator is asymptotically equivalent to the regression estimator $\hat{\beta}_{lr}$, in the sense that $m^{1/2}(\tilde{\beta}_E - \hat{\beta}_{lr}) = o_p(1)$ (see Hartley and Rao, 1967), and $\hat{\lambda}_2$ satisfy (3.10). Along the lines of argument given in Chen and Qin (1993), it can be shown that

$$\begin{aligned}\hat{\lambda}_2 &= \left[\frac{1}{m}\sum_{i\in S_m} S^2_{\hat{\theta}}(Y_i|Z_i)\right]^{-1}\left[\frac{1}{m}\sum_{i\in S_m} S_{\hat{\theta}}(Y_i|Z_i)\right] + o_p(m^{-1/2})\\ &= O_p(m^{-1/2}). \end{aligned} \tag{4.18}$$

Hence, the conditional variance of $S_m(\beta)|S_n$ is given by

$$\begin{aligned} Var\{S_m(\beta)|S_n\} &\doteq m^2\left(\frac{1}{m}\right)(1-\hat{\rho}^2_{y^*x^*})\hat{S}^2_{y^*}\\ &= m^2\left(\frac{1}{m}\right)\left(\hat{S}^2_{y^*} - \frac{\hat{S}^2_{y^*x^*}}{\hat{S}^2_{x^*}}\right), \end{aligned} \tag{4.19}$$

where $\hat{\rho}_{y^*x^*}$ and $\hat{S}_{y^*x^*}$ are the correlation coefficient and covariance between y^* and x^*, respectively and $\hat{S}^2_{y^*}$ is the variance of y^*. Taking expectation of (4.19), we have

$$\begin{aligned} E[Var\{S_m(\beta)|S_n\}] \doteq m &\left\{\frac{1}{n}\sum_{i\in S_n} x_i^2 F(\beta x_i)[1-F(\beta x_i)]\right.\\ &- \left(\frac{1}{n}\sum_{i\in S_n} x_i z_i F(\beta x_i)[1-F(\beta x_i)]\right)^2\\ &\left.\times\left(\frac{1}{n}\sum_{i\in S_n} z_i^2 F(\theta z_i)[1-F(\theta z_i)]\right)^{-1}\right\}. \end{aligned} \tag{4.20}$$

Combining (4.13) and (4.20), we get

$$\begin{aligned} Var\{S_m(\beta)\} \doteq \left(\frac{m}{n}\right)^2 &\sum_{i\in S_n} x_i^2 F(\beta x_i)[1-F(\beta x_i)] + m\left\{\frac{1}{n}\sum_{i\in S_n} x_i^2 F(\beta x_i)\right.\\ &\times[1-F(\beta x_i)] - \left(\frac{1}{n}\sum_{i\in S_n} x_i z_i F(\beta x_i)[1-F(\beta x_i)]\right)^2\\ &\left.\times\left(\frac{1}{n}\sum_{i\in S_n} z_i^2 F(\theta z_i)[1-F(\theta z_i)]\right)^{-1}\right\}. \end{aligned} \tag{4.21}$$

□

We note that the first component of the variance $Var(\tilde{\beta}_E)$ is the expected information based on $L(\beta|X_1, \dots, X_n)$, the likelihood for observed data if $\{X_i, i \in S_n\}$ were known. The second term is, therefore, the penalty induced for not knowing $\{X_i, i \in S_{n-m}\}$.

All of the above results regarding $\tilde{\beta}_E$ assume that p_i's are fixed. Define $\hat{\beta}_E$ the same as $\tilde{\beta}_E$ except the p_i's are replaced by their respective estimates

$$\hat{p}_i = \left(\frac{1}{m}\right) \frac{1}{1 + \hat{\lambda}_2 S_{\hat{\theta}}(Y_i|Z_i)}, \quad i \in S_m. \tag{4.22}$$

Now, consider

$$m^{1/2}(\hat{\beta}_E - \beta) = m^{1/2}(\tilde{\beta}_E - \beta) + m^{1/2}(\hat{\beta}_E - \tilde{\beta}_E). \tag{4.23}$$

By noting that the second term of the right hand side of (4.22) converges to 0 in probability, we conclude that $m^{1/2}(\hat{\beta}_E - \beta)$ has the same asymptotic distribution as $m^{1/2}(\tilde{\beta}_E - \beta)$.

5 A simulation study

We conduct a simulation study to investigate the properties of the estimators studied in this paper. In particular, the absolute bias and relative efficiencies are considered for (a) the partial case estimator ($\hat{\beta}_m$), (b) the imputed estimator ($\hat{\beta}_I$), (c) bootstrap estimator ($\hat{\beta}_B$), and (d) the empirical likelihood estimator ($\hat{\beta}_E$). The values of (X, Z) are generated according to the following three different models:

(I) Linear: $X_i = 1+2Z_i+\varepsilon_i$ with $Z_i \overset{ind}{\sim}$ Unif$(0, 2)$ and $\varepsilon_i \sim \sqrt{Z_i}N(0, 0.25)$ for $i \in S_n$.

(II) Quadratic: $X_i = 2.0 - 2.0Z_i + 2.0Z_i^2 + \varepsilon_i$ with Z_i and ε_i are defined as in Model I for $i \in S_n$.

(III) Linear-Quadratic: $X_i = 1 + 2Z_i + \varepsilon_i$ for $i \in S_m$ and $X_i = 1 + 2Z_i^2 + \varepsilon_i$ for $i \in S_{n-m}$ with Z_i and ε_i are defined as in Model I.

Other parameters covered in this study are $\beta = 1$ and $n = 5,000$. The outcome Y is generated from a Bernoulli random variable conditional on X, whereby $Y = 1$ with probability $(1 + \exp\{-\beta x)\})^{-1}$ and $Y = 0$ otherwise. Furthermore, for each combination of the above parameter values, $N = 200$ independent samples are generated for $n = 5,000$ according to the above three models. A simple random sampling of size $m = 100, 200$ and 500 are then taken without replacement from these populations of n elements. This sampling is repeated for $S = 50$ times. To run the bootstrap procedure with 200 simulations and 100 bootstrap samples with $S = 50$ on a Sun SPARC station 20 model 712, with dual 75 MHZ super SPARC CPU's and 192 MB of RAM, 600 hours of CPU time are required. Hence, to minimize the computer time, the bootstrap method is implemented only for the Model III. The Mak's bootstrap estimator is computed based on $B = 100$ bootstrap samples. The absolute bias for each estimator is computed using the following formula:

$$\text{Absolute bias}(\hat{\beta}_{s,n}) = \frac{1}{NS}\sum_{n=1}^{N}\sum_{s=1}^{S}|\hat{\beta}_t^{s,n} - \beta|, \tag{5.1}$$

$$\tag{5.2}$$

where $\hat{\beta}_t^{s,n}$ is the estimator based on the method $t = m, I, B$ and E. The relative efficiencies of $\hat{\beta}_m$, $\hat{\beta}_B$ and $\hat{\beta}_E$ are defined as the ratios of the mean square error of $\hat{\beta}_I$ to their respective mean square errors., i.e.,

$$\text{Relative Efficiency}(\hat{\beta}_{s,n}) = \frac{\text{MSE}(\hat{\beta}_I^{s,n})}{\text{MSE}(\hat{\beta}_t^{s,n})}, \text{ where} \tag{5.3}$$

$$\text{MSE}(\hat{\beta}_t^{s,n}) = \frac{1}{NS}\sum_{n=1}^{N}\sum_{s=1}^{S}(\hat{\beta}_t^{s,n} - \beta)^2, \tag{5.4}$$

The results are reported in Tables 1–3. The simulation results show the advantages of the empirical likelihood approach over its competitors on both grounds: biasedness and efficiency. The absolute bias values under the Model I are in the range of 0.068–0.167 for $\hat{\beta}_m$, 0.348–0.381 for $\hat{\beta}_I$ and 0.062–0.149 for $\hat{\beta}_E$. While the efficiencies relative to $\hat{\beta}_I$ are in the range of 3.003–17.720 for $\hat{\beta}_m$ and 3.973–21.857 for $\hat{\beta}_E$. Similarly, the absolute bias and efficiencies values under Model II have the similar pattern as in Model I. The absolute bias values under the Model III are in the range of 0.079–0.155 for $\hat{\beta}_m$, 0.140–0.160 for $\hat{\beta}_I$, 0.055–0.094 for $\hat{\beta}_B$ and 0.036–0.076 for $\hat{\beta}_E$. While the efficiencies relative to $\hat{\beta}_I$ are in the range of 0.599–2.039 for $\hat{\beta}_m$, 1.514–4.146 for $\hat{\beta}_B$ and 2.753–10.250 for $\hat{\beta}_E$. There is a negligible absolute bias in $\hat{\beta}_E$ with greater relative efficiency over all other estimators.

The empirical likelihood estimator performs well even when the relationship between X and Z is not linear for $i \in S_{n-m}$, whereas, the performance of the bootstrap estimator in this case is rather poor. These results are encouraging and imply that the proposed method is robust to misspecification of the relationship between X and Z. It is interesting to note that the bootstrap estimator still performs better than the imputed estimator.

In summary, the method we have proposed is a useful alternative to the standard procedure and it is not as computationally intensive as the bootstrap method. The computations can be carried out with an existing maximization subroutine such as Brent's method.

Table 1: Absolute bias and efficiencies of three estimators relative to $\hat{\beta}_I$ for $n = 5,000$, $S = 50$ and $N = 200$ under linear model $X_i = 1.0 + 2.0Z_i + \varepsilon_i$ for $i \in S_n$.

	Absolute bias			Rel. Efficiency	
m	$\hat{\beta}_m$	$\hat{\beta}_I$	$\hat{\beta}_E$	$\hat{\beta}_m$	$\hat{\beta}_E$
100	0.167	0.381	0.149	3.003	3.973
200	0.118	0.348	0.102	4.838	6.586
500	0.068	0.355	0.062	17.720	21.857

Table 2: Absolute bias and efficiencies of three estimators relative to $\hat{\beta}_I$ for $n = 5,000$, $S = 50$ and $N = 200$ under quadratic model $X_i = 2.0 - 2.0Z_i + 2.0Z_i^2 + \varepsilon_i$ for $i \in S_n$.

	Absolute bias			Rel. Efficiency	
m	$\hat{\beta}_m$	$\hat{\beta}_I$	$\hat{\beta}_E$	$\hat{\beta}_m$	$\hat{\beta}_E$
100	0.195	0.357	0.171	1.521	1.915
200	0.112	0.371	0.095	6.745	9.351
500	0.074	0.317	0.063	11.121	15.143

Table 3: Absolute bias and efficiencies of three estimators relative to $\hat{\beta}_I$ for $n = 5,000$, $S = 50$, $B = 100$ and $N = 200$ under linear model $X_i = 1+2Z_i+\varepsilon_i$ for $i \in S_m$ and quadratic model $X_i = 1 + 2Z_i^2 + \varepsilon_i$ for $i \in S_{n-m}$.

	Absolute bias				Rel. Efficiency		
m	$\hat{\beta}_m$	$\hat{\beta}_I$	$\hat{\beta}_B$	$\hat{\beta}_E$	$\hat{\beta}_m$	$\hat{\beta}_B$	$\hat{\beta}_E$
100	0.155	0.160	0.094	0.076	0.599	1.514	2.753
200	0.122	0.153	0.071	0.049	0.996	2.693	6.513
500	0.079	0.140	0.055	0.036	2.039	4.146	10.25

References

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

Chen, J., and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, **80**, 107–116.

Cox, D. R., and Hinkley, D. V. (1974). *Theoretical Statistics*, Chapman and Hall, London.

Foutz, R. V. (1977). On the unique consistent solution to the likelihood equation. *J. Am. Statist. Assoc.*, **72**, 147–148.

Gladen, B., and Rogan, W. (1979). Misclassification and the design of environmental studies. *Am. J. Epidem.*, **109**, 607–616.

Hartley, H. O., and Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, **54**, 93–108.

Kalton, G., and Kasprzyk, D. (1986). Imputing for missing surveys responses. *Survey Methodology*, **12**, 1–16.

Mak, T. K., Li, W. K., and Kuk, Y. C. (1986). The use of surrogate variables in binary regression models. *J. Statist. Comput. Simul.*, **24**, 245–254.

Pepe, M. S., and Fleming, T. (1991). A general nonparametric method for dealing with errors missing or surrogate data. *J. Am. Statist. Assoc.*, **86**, 108–121.

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

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statist. Med.*, **8**, 431–440.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1993). *Numerical Recipes in C: The Art of Scientific Computing*, 2nd edition, Cambridge University Press, New York.

Rubin, D. B. (1987). *Multiple Imputation for Non-Response in Surveys*, John Wiley and Sons, New York.

Schafer, D. W. (1987). Covariate measurement error in generalized linear models. *Biometrika*, **74**, 385–391.

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

Stefanski, L. A., and Carroll, R. J. (1985). Covariate measurement error in generalized linear models. *Ann. Statist.*, **13**, 1335–1351.

Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Am. Statist. Assoc.*, **82**, 528–540.

Wittes, J., Lakatos, E., and Probstfield, J. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statist. Med.*, **8**, 415–425.