BIOMETRICS 56, 98–105
March 2000

# Conditional-Cumulant-of-Exposure Method in Logistic Missing Covariate Regression

**C. Y. Wang**

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center,
P.O. Box 19024, Seattle, Washington 98109-1024, U.S.A.
*email:* cywang@fhcrc.org

**and**

**Wen-Tao Huang**

Institute of Statistical Science, Academia Sinica, Taipei, Taiwan 115, Republic of China

SUMMARY. We consider estimation in logistic regression where some covariate variables may be missing at random. Satten and Kupper (1993, *Journal of the American Statistical Association* **88,** 200–208) proposed estimating odds ratio parameters using methods based on the probability of exposure. By approximating a partial likelihood, we extend their idea and propose a method that estimates the cumulant-generating function of the missing covariate given observed covariates and surrogates in the controls. Our proposed method first estimates some lower order cumulants of the conditional distribution of the unobserved data and then solves a resulting estimating equation for the logistic regression parameter. A simple version of the proposed method is to replace a missing covariate by the summation of its conditional mean and conditional variance given observed data in the controls. We note that one important property of the proposed method is that, when the validation is only on controls, a class of inverse selection probability weighted semiparametric estimators cannot be applied because selection probabilities on cases are zeros. The proposed estimator performs well unless the relative risk parameters are large, even though it is technically inconsistent. Small-sample simulations are conducted. We illustrate the method by an example of real data analysis.

KEY WORDS: Cumulant-generating function; Likelihood; Measurement error; Missing data.

## 1. Introduction

In epidemiologic studies, estimation of an exposure effect on disease incidence is a frequent objective. Due to practical considerations, an exposure variable may not be available for all subjects from the study cohort. For 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 address the association between bladder cancer and some nutrients (Bruemmer et al., 1996). We are interested in covariate variables smoking packet year and body mass index (weight/height$^2$). 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. Nevertheless, the majority of subjects did have the data on the year that they had smoked. Smoking year in this example is a surrogate for smoking packet year. A surrogate variable does not have an effect on the disease incidence given true covariates. To this problem, a naive approach to analysis is the complete-case (CC) analysis, which is to perform the usual analysis using observations in a subsample with complete data, called the validation set. This approach, however, can give rise to bias and reduced statistical efficiency.

In addition to the CC estimator, regression calibration is also a practical approach. This method replaces a missing variable by its conditional expectation on surrogates and covariates. It is easy to implement and it often performs well, although it may be technically inconsistent under some situations (Carroll, Ruppert, and Stefanski, 1995, Chapter 3). Alternatively, maximum likelihood (ML) may be useful, but sometimes it requires an additional model to relate a missing covariate to observed data. For example, likelihood methods were studied in Satten and Kupper (1993) by modeling the probability of exposure. In general, solving ML estimates may need some iterative procedures, such as the EM algorithm. Recently, semiparametric approaches have gained considerable attention to avoid the misspecification of a submodel that specifies the relationship between covariate variables. One may consider estimating the likelihood nonparametrically (Carroll and Wand, 1991; Pepe and Fleming, 1991)

or a class of inverse selection probability weighted estimators (Robins, Rotnitzky, and Zhao, 1994).

The goal of this paper is to extend the idea of Satten and Kupper (1993) to approximate a partial likelihood. We consider the case that a covariate variable may be missing at random (MAR). Under this assumption, the missingness mechanism does not depend on the missing value given observed data. We note that approximating a partial likelihood in this problem was previously discussed in a similar setting by Vach and Schumacher (1993). Our proposal is to use some lower order conditional cumulants (e.g., conditional mean and conditional variance) to estimate the cumulant-generating function of true covariates given observed covariates in the controls. It is important to note that the proposed estimator is an approximate approach and hence is not consistent in general. However, it performs quite well under a variety of circumstances. The proposed estimator also works in a situation where the validation set is only available for controls.

In Section 2, we introduce the model and a likelihood-based estimator from Satten and Kupper (1993). We note that a conditional cumulant-generating function is involved in the ML estimation procedure. The proposed method is described in Section 3. The idea of our method is to estimate the conditional cumulants of the unobserved covariate given observed data in the controls. The asymptotic distribution theory is given. Section 4 summarizes results from a simulation study. In Section 5, we illustrate the proposed method on the bladder cancer case–control study described above.

## 2. Estimation Using Conditional Distribution of Exposure

Let $Y$ denote the disease incidence variable, $X$ be a covariate variable that is missing in some of the total of $n$ subjects, and $Z$ be covariates always observable. Let $W$ be surrogate for $X$, which is relevant to $X$ but is independent of $Y$ given $(Z, X)$. Consider the logistic regression model

$$\text{pr}(Y = 1 \mid Z, X) = H\left(\beta_0 + \beta_1^t Z + \beta_2^t X\right), \qquad (1)$$

where $H(u) = \{1 + \exp(-u)\}^{-1}$. Define $\delta_i = 1$ if $X_i$ is available and equal zero otherwise. Because we assume that $X$ is MAR, $X$ and $\delta$ are conditionally independent given other available data. Our goal is to estimate regression parameters $\beta \equiv (\beta_0, \beta_1^t, \beta_2^t)^t$. The CC estimator is based on data $\{(Y_i, Z_i, X_i) : \delta_i = 1, i = 1, \ldots, n\}$, ignoring cases with one or more missing variables. The CC analysis of $\beta$ is not valid if the selection probabilities depend on $Y$.

Satten and Kupper (1993) proposed estimation using information on the probability of exposure. Let $X \mid (Z, W, Y = 0)$ be modeled up to an unknown parameter $\alpha$. Satten and Kupper showed that

$$\text{pr}(Y = 1 \mid Z, W) = H\left\{\beta_0 + \beta_1^t Z + R(Z, W, \beta_2)\right\}, \qquad (2)$$

where $R(Z, W, \beta) = \ln[\text{E}\{\exp(\beta_2^t X) \mid Z, W, Y = 0\}]$. They also showed that $f_{X|(Z,W,Y=1)}(x) = f_{X|(Z,W,Y=0)}(x) \times \exp\{\beta_2^t x - R(z, w, \beta_2)\}$. Wang, Wang, and Carroll (1997) noted that modeling the conditional distribution of exposure in the controls ($Y = 0$) as a member of the exponential family of distributions implies that $X$ given $(Z, W)$ is a mixture of two distributions from the same family. In particular, if $X$ given $(Z, W)$ is normal with variance $\sigma^2$ among the controls with intercept $\alpha_0$, then it is normal among the cases but with the intercept $\alpha_0 + \sigma^2\beta_2$.

Let $P(Y \mid Z, W)$ denote the likelihood of $Y$ given $(Z, W)$ and $P(Y, \delta X)$ denote $P(Y \mid Z, W)$ if $\delta = 0$ and $P(Y, X \mid Z, W)$ if $\delta = 1$. Note that $P[X \mid Z, W, Y] = P[X \mid Z, W, Y, \delta = 1]$ because $X$ is MAR. Therefore, $P[Y, \delta X \mid Z, W] = P(Y \mid Z, W)\{P(X \mid Z, W, Y, \delta = 1)\}^{\delta}$. The ML estimator may be obtained by the method of scoring since $R(Z, W, \beta)$ has an explicit form as long as $X$ given $(Z, W, Y)$ follows the exponential family. The ML estimator under this setup is tractable and does not require complicated calculations.

## 3. Conditional-Cumulant-of-Exposure Method

Our proposal in this paper is to relax modeling the conditional distribution of $X$ given $(Z, W, Y = 0)$. As mentioned in the previous section, if $X \mid (Z, W, Y = 0)$ is normal, then $X \mid (Z, W)$ is a mixture of normals. On the other hand, if $X \mid (Z, W)$ is normal, then the model assumption of $X \mid (Z, W, Y = 0)$ does not have a simple form. In this case, assuming normality of $(X \mid Z, W, Y = 0)$ may lead to bias estimation of $\beta$. Therefore, in the development of our method, instead of fully modeling the distribution of $X \mid (Z, W, Y = 0)$, we need only a regression model of $X \mid (Z, W, Y = 0)$. We assume that $\text{E}(X \mid Z, W, Y = 0)$ is parametrized by $\alpha$.

### 3.1 *The Proposed Estimator*

Our estimation is to approximate the partial likelihood $\{P(Y \mid X, Z)\}^{\delta}\{P(Y \mid Z, W)\}^{1-\delta}$, discussed in Vach and Schumacher (1993). They observed that, if the missingness does not depend on $Y$, then it is the same as $\{P(Y \mid X, Z, \delta = 1)\}^{\delta}\{P(Y \mid Z, W, \delta = 0)\}^{1-\delta}$. Furthermore, when the missingness depends on $(Y, Z, W)$, it is proportional to the likelihood $\{P(Y, X \mid Z, W)\}^{\delta}\{P(Y \mid Z, W)\}^{1-\delta}$ if $P(X \mid Z, W)$ does not depend on $\beta$. This is closely related to Carroll and Wand (1991) and Pepe and Fleming (1991).

By (2), approximating the induced model $P(Y \mid Z, W)$ involves estimating $R(Z, W, \beta_2)$. Because the development of the new method involves the cumulant-generating function of $X \mid (Z, W, Y = 0)$, we consider scalar $X$ for simplicity. The extension to vector $X$ needs only more complicated notation. Let $\text{E}(X \mid Z, W, Y = 0) \equiv \mu(Z, W)$ and, for $j = 2, \ldots, \infty$, let $\text{E}[\{X - \mu(Z, W)\}^j \mid Z, W, Y = 0] \equiv \mu_j(Z, W)$. Let $\kappa_1 \equiv \mu(Z, W), \kappa_2(Z, W) = \sigma^2(Z, W), \ldots$ be cumulants of $X \mid (Z, W, Y = 0)$. We note that $R(Z, W, \beta_2) = \beta_2\mu(Z, W) + \Sigma_{j=2}^{\infty}\{\beta_2^j/j!\}\kappa_j(Z, W)$ (Kendall and Stuart, 1977, Chapter 3). Assume that there is a positive $k$ such that $R(Z, W, \beta_2) = \beta_2\mu(Z, W) + \Sigma_{j=2}^{k}(\beta_2^j)\kappa_j(Z, W)/(j!)$. Let $X_i^* = X_i^{\delta_i}\{\mu(Z_i, W_i) + \Sigma_{j=2}^{k}\beta_2^{j-1}\kappa_j(Z, W)/(j!)\}^{1-\delta_i}$. Thus, the induced model $\text{pr}(Y = 1 \mid Z, W) = H[\beta_0 + \beta_1^t Z + \beta_2\{\mu(Z, W) + \Sigma_{j=2}^{k}\beta_2^{j-1}\kappa_j(Z, W)/(j!)\}]$. As discussed above, the argument of approximating the partial likelihood suggests the replacement of an unobserved $X_i$ by $\mu(Z_i, W_i) + \Sigma_{j=2}^{k}\beta_2^{j-1}\kappa_j(Z, W)/(j!)$ in the nonvalidation set. Let $\hat{X}_i^*$ be $X_i^*$ but with estimated $\mu(Z, W)$ and $\kappa_j(Z, W)$. The resulting estimator $\hat{\beta}$, called the conditional-cumulant-of-exposure (CCOE) estimator, solves

$$\sum_{i=1}^{n}(1, Z_i^t, \hat{X}_i^*)^t\left\{Y_i - H\left(\beta_0 + \beta_1^t Z_i + \beta_2\hat{X}_i^*\right)\right\} = 0. \qquad (3)$$

To implement the proposed estimator, we may solve estimating equation (3) by the Newton–Raphson algorithm and the starting value for $\beta$ may be the CC estimator. Alternatively, one may apply a usual logistic regression using covariate variable $X_i$ (if $\delta_i = 1$) or the replacement $\hat{X}_i^*$ (if $\delta_i = 0$). However, in practice, a large $k$ value will significantly increase the dimension of the nuisance parameters modeling $\kappa_j(Z, W)$, $j = 1, \ldots, k$. Therefore, we suggest a practical approximation that assumes $k = 2$. Let $\alpha$ be the parameter for $\mu(Z, W)$ and $\gamma$ be the parameter for $\sigma^2(Z, W)$. To estimate $\alpha$, we consider the idea of an estimating equation and assume $\hat{\alpha}$ solves

$$\sum_{i=1}^{n} \delta_i I[Y_i = 0] \mu_\alpha(Z_i, W_i, \alpha)\{X_i - \mu(Z_i, W_i, \alpha)\} = 0, \quad (4)$$

where $\mu_\alpha(Z_i, W_i, \alpha) = (\partial/\partial\alpha)\mu(Z_i, W_i, \alpha)$ and $I[\cdot]$ is an indicator function. Similarly, to estimate $\gamma$ in $\sigma^2(Z, W, \gamma) = \sigma^2(Z, W)$, we may solve

$$\sum_{i=1}^{n} \delta_i I[Y_i = 0] \sigma_\gamma^2(Z_i, W_i, \gamma) \times \left[\{X_i - \mu(Z_i, W_i, \alpha)\}^2 - \sigma^2(Z_i, W_i, \gamma)\right] = 0, \quad (5)$$

where $\sigma_\gamma^2(Z_i, W_i, \gamma) = (\partial/\partial\gamma)\sigma^2(Z_i, W_i, \gamma)$. In this case, a missing $X_i$ is replaced by $\hat{X}_i^* = \mu(Z_i, W_i, \hat{\alpha}) + (1/2) \times \beta_2 \sigma^2(Z_i, W_i, \hat{\gamma})$. Modeling $\mu(Z_i, W_i, \alpha)$ and $\sigma^2(Z_i, W_i, \gamma)$ may be examined from controls in the validation set. We will further illustrate this in the simulation study.

### 3.2 *Distribution Theory*

In this section, we assume that $\kappa_j(Z, W) = 0$ for $j > 2$. Using the notation above, write (3) as $U_n(\beta, \hat{\alpha}, \hat{\gamma}) = 0$ and let $\beta_*$ be the root of $\mathrm{E}\{U_n(\beta, \alpha, \gamma)\} = 0$. Then the solution to (3) converges to $\beta_*$. The convergence can be proved by using a standard technique of M-estimators. As mentioned in the Introduction, the estimand $\beta_*$ is only an approximate value for $\beta$, but in general, it may be different from $\beta$. The asymptotic distribution theory for the proposed estimator $\hat{\beta}$ is summarized in the following result.

PROPOSITION 1: *Let* $\mathrm{pr}(Y = 1 \mid Z, X) = H(\beta_0 + \beta_1^t Z + \beta_2 X)$ *and assume that* $X_i$ *is observed under the selection probability* $\mathrm{pr}(\delta_i = 1 \mid Z_i, W_i, Y_i)$. *Assume that* $\mathrm{E}(X \mid Z, W, Y = 0) = \mu(Z, W, \alpha)$ *for some unknown* $\alpha$ *and* $\mathrm{var}(X \mid Z, W, Y = 0) = \sigma^2(Z, W, \gamma)$ *for some unknown* $\gamma$, *such that* $\log \mathrm{E}\{\exp(\beta_2 X) \mid Z, W, Y = 0\} = \beta_2 \mu(Z, W, \alpha) + (\beta_2^2/2)\sigma^2(Z, W, \gamma)$. *If* $\hat{\beta}$ *solves* (3), *then* $n^{1/2}(\hat{\beta} - \beta_*)$ *is asymptotically normally distributed with mean zero and asymptotic variance given in expression* (6) *of the Appendix.*

A sketch of the proof of Proposition 1 is given in the Appendix. We observe that the moment conditions in Proposition 1 take into consideration the heteroscedastic error of $X$ given $(Z, W, Y = 0)$ but ignores higher order moments. The approximation error due to ignoring higher order moments may result but may be limited compared to $\beta$.

## 4. Simulation Study

In this section, we demonstrate the numerical performance of the estimator that we have developed in this paper. Covariate variables $X$, $Z$, and $W$ are scalars. We compared the following estimators:

(i) CC estimator.

(ii) Inverse selection probability weighted estimator. We estimate $\pi_i$ by parametrically modeling the logistic regression of $\delta_i$ given $(Y_i, Z_i, W_i)$.

(iii) ML estimator assuming $X \mid (Z, W, Y = 0)$ is normal with mean $\alpha_0 + \alpha_1 Z + \alpha_2 W$ and a constant variance $\sigma^2$ (described in Section 2).

(iv) Approximate CCOE estimator assuming $k = 2$. It solves (3) by using the first two conditional cumulants and by assuming that $X \mid (Z, W, Y = 0)$ has mean $\mu(Z, W) = \alpha_0 + \alpha_1 Z + \alpha_2 W$ and a nonconstant variance function $\sigma^2(Z, W) = \gamma_0 + \gamma_1 Z + \gamma_2 W + \gamma_3 Z^2 + \gamma_4 W^2 + \gamma_5 ZW$.

In Table 1, we consider the case that $X \mid (Z, W, Y = 0)$ is normal. We first generated both $Z$ and $W$ from $\mathrm{N}(0, 1)$ such that $\mathrm{corr}(Z, W) = 0.25$, with size $n = 200$, 500, respectively. We next generated binary outcome $Y$'s from (2), with $R(Z, W, \beta) = \beta_2 \mu(Z, W) + (1/2)\beta_2^2 \sigma^2$, where $\mu(Z, W) = \alpha_0 + \alpha_1 Z + \alpha_2 W$. We then generated $X$ such that $X \mid Z, W, Y = 0$ is normally distributed with mean $\mu(Z, W)$ and variance $\sigma^2$. Parameters are $\beta = (0, \ln(2), \ln(2))^t$, $\alpha = (0, 1, -1)^t$, and $\sigma = 0.5$. There were 1000 simulations in each experiment. The validation data indicator $\delta_i$ in Table 1 was generated by the selection probability such that $\mathrm{pr}(\delta_i = 1 \mid Y_i, Z_i, W_i) = \{1 + \exp(-1 + Y_i + Z_i + W_i)\}^{-1}$. On average, about 58% of the observations are validation data in which $X$ was observed. The bias means the average of $\hat{\beta} - \beta$, where $\hat{\beta}$ is an estimator of $\beta$. SD denotes the square root of the sample variance of the 1000 estimates. Mean(SE) denotes the average of the 1000 standard error estimates. We also include 95% coverage probabilities of the true parameters for the estimates.

We note from Table 1 that the CC analysis that uses only the validation data has large bias since the selection probabilities depend on $(Y, Z, W)$. This bias problem can be seen more clearly for the case with $n = 500$. Note that both the ML estimator and CCOE estimator perform well in terms of bias and efficiency. The weighted estimator appears to have a small sample performance problem, although it is theoretically consistent. The coverage probabilities of the 95% confidence interval seem too low, especially for $n = 200$. This problem diminished as we increased the sample size, and it can hardly be seen when $n = 2000$.

In Table 2, data were generated similar to those of Table 1, but $X$ given $(Z, W, Y = 0)$ is normal with a nonconstant variance function that $\sigma^2(Z, W) = \{\sigma_* \mu(Z, W)\}^2$, where $\sigma_* = 0.75$. The validation set consists of a random sample of 80% of controls only, leading to 60% of missing $X$. In this case, the CC and weighted estimators are not applicable. Similar to a result in Section 2, we note that $X$ given $(Z, W, Y = 1)$ is normal with mean $\mu(Z, W, \alpha) + \beta_2 \sigma^2(Z, W)$ and the same variance $\sigma^2(Z, W)$. Parameters are $\beta = (0, \ln(3), \ln(3))^t$ and $\alpha = (0, 1, -1)^t$. The ML estimator based on a constant variance assumption has a moderate model violation, and this led to the estimation bias in $\beta_0$ and $\beta_2$. In this situation, the CCOE estimator is still satisfactory.

**Table 1**
*$X$ given $(Z, W, Y = 0)$ is normal and $\sigma = 0.5$. Parameters $\beta = (0, \ln(2), \ln(2))$*

| | | $n = 200$ | | | | $n = 500$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | CC | Weighted | ML | CCOE | CC | Weighted | ML | CCOE |
| $\beta_0$ | Bias | −0.411 | −0.022 | −0.004 | 0.000 | 0.401 | −0.014 | −0.003 | 0.002 |
| | SD | 0.276 | 0.255 | 0.181 | 0.184 | 0.164 | 0.151 | 0.113 | 0.116 |
| | Mean(SE) | 0.265 | 0.215 | 0.178 | 0.183 | 0.164 | 0.139 | 0.111 | 0.114 |
| | 95% coverage probability | 0.667 | 0.903 | 0.954 | 0.956 | 0.307 | 0.929 | 0.943 | 0.937 |
| $\beta_1$ | Bias | −0.259 | 0.014 | 0.018 | 0.019 | −0.275 | −0.003 | 0.015 | 0.017 |
| | SD | 0.369 | 0.480 | 0.258 | 0.258 | 0.222 | 0.306 | 0.164 | 0.166 |
| | Mean(SE) | 0.373 | 0.385 | 0.257 | 0.263 | 0.230 | 0.264 | 0.160 | 0.164 |
| | 95% coverage probability | 0.887 | 0.891 | 0.953 | 0.954 | 0.779 | 0.911 | 0.937 | 0.941 |
| $\beta_2$ | Bias | 0.158 | 0.066 | 0.021 | 0.021 | 0.143 | 0.035 | 0.012 | 0.011 |
| | SD | 0.242 | 0.298 | 0.178 | 0.181 | 0.152 | 0.193 | 0.113 | 0.114 |
| | Mean(SE) | 0.247 | 0.257 | 0.182 | 0.184 | 0.152 | 0.172 | 0.113 | 0.114 |
| | 95% coverage probability | 0.950 | 0.906 | 0.962 | 0.966 | 0.883 | 0.922 | 0.952 | 0.950 |

Table 3 demonstrates the asymptotic bias for a variety of parameters. We used $n = 5000$ for all entries and rechecked some entries with $n = 10{,}000$, and a difference of <0.005 was noticed. Covariates $Z, W$ were generated similar to Table 1, but $X$ given $(Z, W)$ was normal with mean $Z - W$ and variance $\sigma^2$. The validation set selection is the same as in Table 2. Under this setup, the bias of the CCOE estimator is mainly from approximating $R(Z, W, \alpha)$. The maximal absolute bias happens when $\beta_1 = \beta_2 = \ln(6)$ and $\sigma = 1$. Among these factors, larger $\beta_2$ is the most sensitive one to the bias. Nevertheless, the magnitude of the bias seems to be limited compared to the magnitude of $\beta_2$.

## 5. Data Analysis

We now apply the methods to the case–control study of bladder cancer described in the Introduction. Issues on the prospective analysis of case–control data were addressed in Carroll, Wang, and Wang (1995). In this study, eligible subjects were residents of three counties of western Washington state who were diagnosed between January 1987 and June 1990 with invasive or noninvasive bladder cancer. In our demonstration, the response variable is the bladder cancer history $(Y)$, and we are interested in covariate variables smoking packet year $(X)$ and body mass index $(Z)$. There were a total of 215 cases and 283 controls in this data analysis, where $Z$ is

**Table 2**
*Validation sample available only for controls. $X$ given $(Z, W, Y = 0)$ is normal with* $\mathrm{var}(X \mid Z, W, Y = 0) = \{\sigma_* \mathrm{E}(X \mid Z, W, Y = 0)\}^2$, *where $\sigma_* = 0.75$. $\beta = (0, \ln(3), \ln(3))$.*

| | | $n = 200$ | | | | $n = 500$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | CC | Weighted | ML | CCOE | CC | Weighted | ML | CCOE |
| $\beta_0$ | Bias | — | — | −0.225 | −0.005 | — | — | −0.200 | 0.024 |
| | SD | — | — | 0.322 | 0.252 | — | — | 0.182 | 0.143 |
| | Mean(SE) | — | — | 0.327 | 0.266 | — | — | 0.193 | 0.149 |
| | 95% coverage probability | — | — | 0.956 | 0.962 | — | — | 0.906 | 0.958 |
| $\beta_1$ | Bias | — | — | 0.016 | 0.031 | — | — | −0.001 | 0.004 |
| | SD | — | — | 0.366 | 0.380 | — | — | 0.223 | 0.233 |
| | Mean(SE) | — | — | 0.341 | 0.408 | — | — | 0.213 | 0.244 |
| | 95% coverage probability | — | — | 0.954 | 0.976 | — | — | 0.944 | 0.960 |
| $\beta_2$ | Bias | — | — | −0.036 | 0.071 | — | — | −0.072 | 0.013 |
| | SD | — | — | 0.297 | 0.396 | — | — | 0.181 | 0.227 |
| | Mean(SE) | — | — | 0.295 | 0.472 | — | — | 0.177 | 0.249 |
| | 95% coverage probability | — | — | 0.916 | 0.960 | — | — | 0.888 | 0.948 |

**Table 3**
*Asymptotic bias analysis: covariates $Z$ and $W$ are standard bivariate normal with correlation 0.25. $X$ given $Z, W$ is normal with mean $Z - W$ and variance $\sigma^2$.*

| | | $\beta_1 = \ln(2)$ $\beta_2$ | | | $\beta_1 = \ln(4)$ $\beta_2$ | | | $\beta_1 = \ln(6)$ $\beta_2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $n_v/n$ | $\sigma$ | ln(2) | ln(4) | ln(6) | ln(2) | ln(4) | ln(6) | ln(2) | ln(4) | ln(6) |
| | | **Bias of the CCOE Estimator for $\beta_1$** | | | | | | | | |
| 0.4 | 0.5 | −0.003 | −0.013 | −0.026 | −0.005 | −0.029 | −0.048 | −0.007 | −0.030 | −0.065 |
| | 1 | −0.010 | −0.041 | −0.065 | −0.022 | −0.075 | −0.106 | −0.030 | −0.101 | −0.137 |
| 0.2 | 0.5 | −0.003 | −0.013 | −0.026 | −0.006 | −0.033 | −0.057 | −0.009 | −0.048 | −0.079 |
| | 1 | −0.007 | −0.029 | −0.043 | −0.023 | −0.082 | −0.112 | −0.034 | −0.121 | −0.167 |
| | | **Bias of the CCOE Estimator for $\beta_2$** | | | | | | | | |
| 0.4 | 0.5 | 0.000 | −0.005 | −0.020 | 0.002 | −0.004 | −0.020 | 0.004 | −0.002 | −0.019 |
| | 1 | 0.007 | 0.021 | 0.061 | 0.010 | 0.012 | 0.027 | 0.014 | 0.011 | 0.019 |
| 0.2 | 0.5 | −0.001 | −0.016 | −0.045 | 0.000 | −0.022 | −0.053 | 0.000 | −0.022 | −0.061 |
| | 1 | −0.001 | −0.033 | −0.055 | −0.002 | −0.051 | −0.099 | −0.002 | −0.060 | −0.120 |

available for all subjects. However, the smoking packet year information of 1 case and 159 controls were missing. In addition, we treated past smokers as in the nonvalidation set since we were primarily interested in the smoking effect of current smokers. As a result, there were 121 cases and 58 controls in the validation set. We consider the years that a subject had smoked ($W$) as the surrogate variable for $X$. Standardized measurements for $X, Z$, and $W$ were used. Note that the original study has a total of 667 subjects, but the analyses here are based on 498 current and past smokers.

Figure 1 shows the scatterplot of $X$ versus $W$ from the validation controls. We note that the variation of $X$ increases as $W$ increases. From the regression of $X$ given $(Z, W)$ in the controls, the least square estimates of $\alpha$ are (−0.628, 0.116, 0.670) with standard errors (0.173, 0.116, 0.186) and the regression standard deviation assuming constant residual variance is 0.805. However, $X \mid (Z, W, Y = 0)$ seems to have a heteroscedastic variance, which can be seen from the residual plot (Figure 2), with the regression residual against $X$.

We now discuss the missing data mechanism. By running a logistic regression of $\delta$ on $(Y, Z, W)$, we obtained the parameter estimates (−1.847, 1.808, −0.179, 1.643) with standard errors (0.197, 0.247, 0.123, 0.184). The result of the parametric estimation of the selection probabilities suggests that the missingness has strong association with $Y$ and $W$.

The estimates of $\beta_1$ and $\beta_2$ are given in Table 4, while those of $\beta_0$ are not valid because of the case–control sampling scheme. The $\mu(Z, W)$ and $\sigma^2(Z, W)$ are identical to the ones used in the simulation study. We note that the ML estimate assuming normality with a constant variance of $X \mid (Z, W, Y = 0)$ might have a model misspecification problem in this data analysis. From this result, the CC analysis does not show a significant adverse effect of smoking on bladder

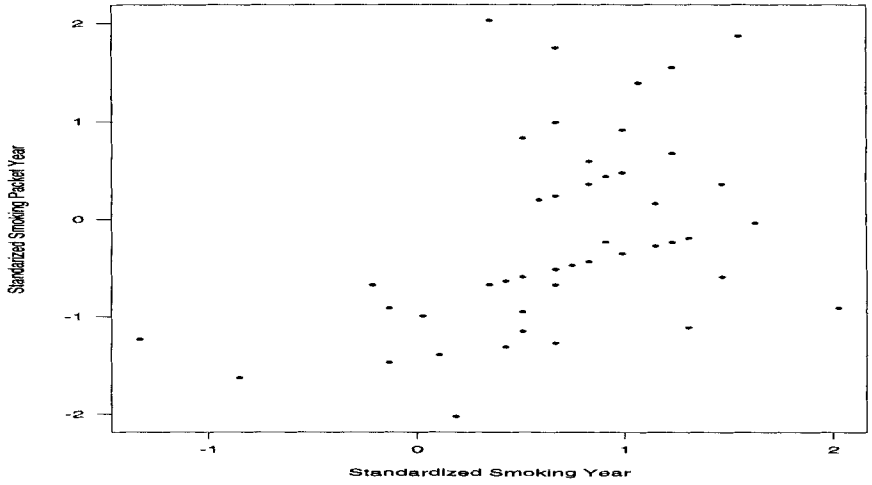


**Figure 1.** Scatterplot of the standardized smoking packet year versus the standardized smoking year in the controls.

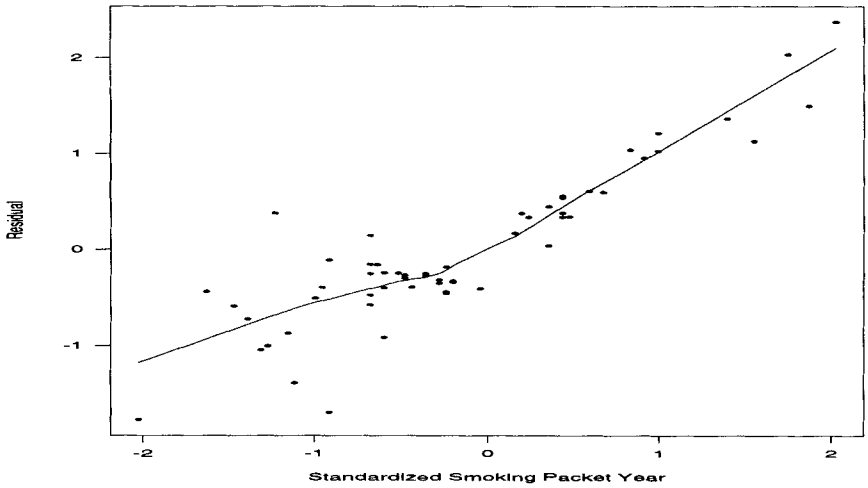


**Figure 2.** Residuals versus the standardized smoking packet year in the controls. The solid line is a smoothing curve to fit the residuals.

cancer risk among current or past smokers. This might be because the selection of the validation set is strongly positively associated with the surrogate variable. Because of the missingness mechanism, we tend to believe that the estimates from the weighted estimate, the ML estimate, and the CCOE estimates are more reliable than the CC analysis. Some differences, although not large, between the ML estimates and the CCOE estimates were observed. This supports that smoking, measured by packet year, increases the risk of bladder cancer. From the ML or CCOE estimate, an increased risk of bladder cancer incidence was also noted for an increased body mass index at the 5% significance level.

## 6. Discussion

We have proposed an estimation method for logistic regression when some covariates may be missing and the missingness is ignorable. In this paper, we have illustrated approximation based on modeling the conditional mean and the conditional variance of the exposure variable. One important feature of the proposed estimator is that it is applicable if the validation set includes only controls. Our analytic calculations and simulations suggest that, under a variety of circumstances, the proposed estimator performs quite well for the estimation of the logistic regression parameter $\beta$. This proposed CCOE estimator can be viewed as a method that replaces a missing $X$ value by $\hat{X}^*$, defined in Section 3. The estimator can be implemented by either solving estimating equations or using an algorithm that needs only standard subroutines (e.g., from SAS or Splus). To estimate the standard errors, the former can be obtained by a sandwich estimator and the latter may be obtained by bootstrap. Naturally, there are situations where higher order cumulants may be needed to reduce the bias of the CCOE estimator. For example, if the distribution of $X$ given $(Z, W, Y = 0)$ is highly skewed and the measurement error is large, then we suggest using $k = 4$.

The proposed approach can be applied to Cox regression with missing covariates. Briefly, for a subject with missing covariate $X$, one needs to approximate $\mathrm{E}[\exp\{\beta_2 X \mid Z, W, \mathcal{G}(t)\}]$, where $\mathcal{G}(t)$ is an at-risk indicator. (The details are in an unpublished manuscript available from the first author.)

**Table 4**
*Bladder cancer analysis*

| | CC | Weighted | ML | CCOE |
|---|---|---|---|---|
| $\beta_1$ | 0.088 | 0.191 | 0.278 | 0.226 |
| (SE) | (0.142) | (0.189) | (0.113) | (0.115) |
| $\beta_2$ | 0.251 | 0.596 | 0.443 | 0.508 |
| (SE) | (0.177) | (0.248) | (0.105) | (0.146) |

Note: Parameters $\beta_1$ and $\beta_2$ are the coefficients of the standardized body mass index and the standardized smoking packet year, respectively.

## Acknowledgements

CYW's research was supported by U.S. National Cancer Institute grants CA 53996 and AG 15026 and a travel award from Academia Sinica, Taiwan. The authors are grateful to Barbara Bruemmer and Emily White for the case–control data and to two referees for helpful comments.

## Résumé

Nous considérons le problème de l'estimation dans une régression logistique où certaines covariables peuvent être aléatoirement manquantes. Satten et Kupper (1993) ont proposé d'estimer les paramètres que sont les rapports de chances en utilisant des méthodes fondées sur la probabilité d'exposition. En approchant une vraisemblance partielle, nous étendons leur idée et proposons une méthode qui permet d'estimer la fonction génératrice des cumulants de la covariable manquante sachant les covariables observées et les variables suppléantes dans les contrôles. Notre méthode consiste à estimer d'abord quelques cumulants d'ordre faible de la distribution conditionnelle des données non observées et ensuite à résoudre une équation d'estimation du paramètre de la régression logistique qui en résulte. Une version simple de la méthode pro-

posée est de remplacer la covariable manquante par la sommation de ses moyenne conditionnelle et variance conditionnelle sachant les valeurs observées dans les contrôles. Nous remarquons qu'une propriété importante de la méthode proposée est que, lorsque la validation est faite seulement sur les contrôles, une classe d'estimateurs semi paramétriques inversement pondérés par la probabilité de sélection ne peut être utilisée car alors les probabilités de sélection de ces cas sont nulles. L'estimateur proposé a de bonnes performances à moins que les paramètres de risques relatifs soient grands; même s'il est techniquement non convergent. Des simulations sur de petits échantillons sont faites. Nous illustrons la méthode par un exemple d'analyse de données réelles.

## References


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

Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *Journal of the Royal Statistical Society, Series B* **53,** 573–585.

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

Carroll, R. J., Wang, S., and Wang, C. Y. (1995). Prospective analysis of stratified logistic case–control studies. *Journal of the American Statistical Association* **90,** 157–169.

Kendall, M. G. and Stuart, A. (1977). *The Advanced Theory of Statistics,* Volume 1, 4th edition. New York: Hafner.

Pepe, M. S. and Fleming, T. R. (1991). A general nonparametric method for dealing with errors in missing or surrogate data. *Journal of the American Statistical Association* **86,** 108–113.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89,** 846–866.

Satten, G. A. and Kupper, L. L. (1993). Inferences about exposure–disease association using probability of exposure information. *Journal of the American Statistical Association* **88,** 200–208.

Vach, W. and Schumacher, M. (1993). Logistic regression with incomplete observed categorical covariates: A comparison of three approaches. *Biometrika* **80,** 353–362.

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


*Received May* 1998. *Revised November* 1998. *Accepted April* 1999.

## Appendix

*Sketch of the Proof of Proposition* 1

Define $V_i = X_i^{\delta_i}\{\mu(Z_i, W_i, \alpha) + \beta_2\sigma^2(Z_i, W_i, \gamma)\}^{1-\delta_i}$, $\mathcal{X}_i = (1, Z_i^t, X_i^*)^t$, and $\mathcal{T}_i = (1, Z_i^t, V_i)^t$. We note that $U_n(\beta, \alpha) = n^{-1/2}\Sigma_{i=1}^n \mathcal{X}_i\{Y_i - H(\beta_0 + \beta_1^t Z_i + \beta_2 X_i^*)\}$ is unbiased when the expectation is evaluated at $\beta = \beta_*$. By a Taylor expansion,

$$\begin{aligned} 0 &= U_n(\hat\beta, \hat\alpha, \hat\gamma) \\ &= U_n(\beta_*, \alpha, \gamma) - G_n(\beta_*, \alpha, \gamma)n^{1/2}(\hat\beta - \beta_*) \\ &\quad - A_n(\beta_*, \alpha, \gamma)n^{1/2}(\hat\alpha - \alpha) \\ &\quad - B_n(\beta_*, \alpha, \gamma)n^{1/2}(\hat\gamma - \gamma) + o_p(1), \end{aligned}$$

where

$$\begin{aligned} &G_n(\beta, \alpha, \gamma) \\ &\quad = n^{-1}\sum_{i=1}^n \mathcal{X}_i\mathcal{T}_i^t H^{(1)}\left(\beta_0 + \beta_1^t Z_i + \beta_2 X_i^*\right) \\ &\qquad + n^{-1}\sum_{i=1}^n (1-\delta_i)(0,0,1)^t\{0, 0, (\beta_2/2)\sigma^2(Z_i, W_i, \gamma)\} \\ &\qquad\qquad \times \left\{Y_i - H\left(\beta_0 + \beta_1^t Z_i + \beta_2 X_i^*\right)\right\}, \\ &A_n(\beta, \alpha, \gamma) \\ &\quad = n^{-1}\sum_{i=1}^n (1-\delta_i)\mathcal{X}_i\beta_2\mu_\alpha^t(Z_i, W_i, \alpha) \\ &\qquad\qquad \times H^{(1)}\left(\beta_0 + \beta_1^t Z_i + \beta_2 X_i^*\right) \\ &\qquad - n^{-1}\sum_{i=1}^n (1-\delta_i)(0,0,1)^t\mu_\alpha^t(Z_i, W_i, \alpha) \\ &\qquad\qquad \times \left\{Y_i - H\left(\beta_0 + \beta_1^t Z_i + \beta_2 X_i^*\right)\right\}, \\ &B_n(\beta, \alpha, \gamma) \\ &\quad = n^{-1}\sum_{i=1}^n (1-\delta_i)\mathcal{X}_i\left\{\beta_2^2/2\right\}\sigma_\gamma^2(Z_i, W_i, \gamma) \\ &\qquad\qquad \times H^{(1)}\left(\beta_0 + \beta_1^t Z_i + \beta_2 X_i^*\right) \\ &\qquad - n^{-1}\sum_{i=1}^n (1-\delta_i)\{0, 0, (\beta_2/2)\}\sigma_\gamma^2(Z_i, W_i, \gamma) \\ &\qquad\qquad \times \left\{Y_i - H\left(\beta_0 + \beta_1^t Z_i + \beta_2 X_i^*\right)\right\}, \end{aligned}$$

where $\mu_\alpha(Z_i, W_i, \alpha) = (\partial/\partial\alpha)\mu(Z_i, W_i, \alpha)$, $\sigma_\gamma^2(Z_i, W_i, \gamma) = (\partial/\partial\gamma)\sigma^2(Z_i, W_i, \gamma)$, and $H^{(1)}(\cdot) = H(\cdot)\{1 - H(\cdot)\}$. Denote the probability limits of $G_n(\beta, \alpha, \gamma)$, $A_n(\beta, \alpha, \gamma)$, and $B_n(\beta, \alpha, \gamma)$ by $G(\beta, \alpha, \gamma)$, $A(\beta, \alpha, \gamma)$, and $B(\beta, \alpha, \gamma)$, respectively. Let $C(\alpha)$ be the probability limit of

$$C_n(\alpha) = n_{v0}^{-1}\Sigma_{i=1}^n \delta_i I[Y_i = 0]\mu_\alpha(Z_i, W_i, \alpha)\mu_\alpha^t(Z_i, W_i, \alpha),$$

where $n_{v0} = \Sigma_{i=1}^n \delta_i I[Y_i = 0]$ is the sample size of controls in the validation set. Let $D(\gamma)$ be the probability limit of $D_n(\gamma) = n_{v0}^{-1}\Sigma_{i=1}^n \delta_i I[Y_i = 0]\sigma_\gamma^2(Z_i, W_i, \gamma)\sigma_\gamma^2(Z_i, W_i, \gamma)$. By some

calculations, $n^{1/2}(\hat{\beta} - \beta_*)$ is asymptotically normally distributed with mean zero and asymptotic variance

$$G^{-1}(\beta_*, \alpha, \gamma) \left\{ n^{-1} \sum_{i=1}^{n} \phi_i(\beta_*, \alpha, \gamma) \phi_i^t(\beta_*, \alpha, \gamma) \right\} \times G^{-1}(\beta_*, \alpha, \gamma), \tag{6}$$

where

$$\begin{aligned}
&\phi_i(\beta, \alpha, \gamma) \\
&\quad = \mathcal{X}_i \{ Y_i - H(\beta_0 + \beta_1^t Z_i + \beta_2 X_i^*) \} \\
&\qquad - n n_{v0}^{-1} A(\beta, \alpha, \gamma) C^{-1}(\alpha) \delta_i I[Y_i = 0] \mu_\alpha(Z_i, W_i, \alpha) \\
&\qquad \times \{ X_i - \mu(Z_i, W_i, \alpha) \} \\
&\qquad - n n_{v0}^{-1} B(\beta, \alpha, \gamma) D^{-1}(\gamma) \delta_i I[Y_i = 0] \sigma_\gamma^2(Z_i, W_i, \gamma) \\
&\qquad \times \left[ \{ X_i - \mu(Z_i W_i, \alpha) \}^2 - \sigma^2(Z_i, W_i, \gamma) \right].
\end{aligned}$$