Journal of Nonparametric Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gnst20

A functional method for the conditional logistic regression with errors-in-covariates

Samiran Sinha [a]
[a] Department of Statistics, Texas A&M University, College Station, TX, 77843, USA

Available online: 25 May 2012

To cite this article: Samiran Sinha (2012): A functional method for the conditional logistic regression with errors-in-covariates, Journal of Nonparametric Statistics, DOI:10.1080/10485252.2012.687735

To link to this article: http://dx.doi.org/10.1080/10485252.2012.687735

Journal of Nonparametric Statistics
2012, 1–19, iFirst article

Taylor & Francis
Taylor & Francis Group

A functional method for the conditional logistic regression with errors-in-covariates

Samiran Sinha*

Department of Statistics, Texas A&M University, College Station, TX 77843, USA

(*Received 7 July 2011; final version received 9 April 2012*)

In this article, we develop a functional approach for handling errors-in-covariates in matched case–control studies which are commonly analysed through the conditional logistic regression. We propose to estimate the parameters from a set of unbiased estimating equations. We require that the moment-generating function of the measurement errors exists. We investigate the asymptotic properties of the estimators. The finite sample performance of the method is judged via simulation studies. The proposed methodology is illustrated by analysing the data from the NIH-AARP Diet and Health study.

Keywords: Bootstrap variance; case–control; conditional likelihood; estimating-equation; moment-generating function; NIH-AARP Diet and health study

1. Introduction

In nutrition epidemiology, the association between nutrient intakes and a disease outcome, such as cancer, is investigated, and many such studies show a lack of significant association between the disease and nutrient intakes. One of the possible causes of lack or weak association is the presence of errors in the reported intakes. Usually, the estimator of the association parameter is biased if one ignores the measurement errors-in-covariates. In this paper, we propose a functional approach to handling additive measurement errors in a matched case–control data. This design is routinely used for studying association between a disease and risk factors after controlling some confounding variables which are markedly associated with the disease and the potential risk factors. A matched case–control study consists of several matched strata which are formed by a set of confounding variables, and each stratum contains a case (diseased subject) and a number of controls (non-diseased subjects), and a logistic regression is used to model the disease risk in terms of the covariates and the confounding (matching) variables.

Measurement errors in the logistic regression model have drawn considerable attention and the work can be broadly classified into two categories: the structural and functional approaches.

*Email: sinha@stat.tamu.edu

ISSN 1048-5252 print/ISSN 1029-0311 online

http://dx.doi.org/10.1080/10485252.2012.687735
http://www.tandfonline.com

In the structural approaches (Guolo 2008; Carroll, Ruppert, Stepanski, and Crainiceanu 2006, pp. 181–203), the unknown true exposure is treated as a random variable, and it requires the knowledge of the distribution of the true exposure and the measurement errors. In the functional approach (Stefanski and Carroll 1987; Buzas and Stefanski 1996; Huang and Wang 2001), the unobserved true covariate is unknown, but considered to be fixed, and consequently no assumption is made regarding the distribution of unobserved true covariate. For dealing with measurement errors in the logistic regression, Wang and Wang (1997) extended the pseudo-conditional likelihood (Breslow and Cain 1988) and the mean-score approach (Reilly and Pepe 1995) using a kernel-based method where the observed surrogate variable was a continuous variable. Cheng and Hsueh (2003) proposed a semiparametric efficient estimator based on inverse probability-weighted estimating equation (EE) when a covariate is mismeasured and the binary response variable is misclassified. Rabe-Hesketh, Pickles, and Skrondal (2003) proposed a nonparametric maximum likelihood method for handling normal additive measurement error, and their study design included replicated measurements of the surrogate variable. In this paper, we develop a functional approach to handle the measurement errors in the conditional logistic regression.

Alternative to the structural and functional approaches is the regression calibration (RC) method where the unobserved true exposure is replaced by its conditional expectation given the surrogate and other error-free covariates (Carroll et al. 2006, p. 65). The RC method is easy to apply but lacks theoretical justification. Sugar, Wang, and Prentice (2007) proposed some modifications to the RC method.

It is to be noted that a potential difficulty in analysing a matched data is that the distribution of the exposure in the underlying population differs from that in the case–control sample, a feature that makes the structural inferential analysis difficult and potentially nonrobust. Furthermore, the distribution of the exposure among the cases and controls potentially depends on the matching variables, and a parametric modelling of this distribution may risk model misspecification. Also, the methods for handling measurement errors in a prospective study are not directly applicable to a matched design due to (1) stratum-level dependence among the case and controls and (2) the functional form of the effect of the matching variables on the logit of the disease risk is left unspecified.

For handling additive errors-in-covariates in a matched case–control study, Armstrong, Whittemore, and Howe (1989) proposed a simple method when the measurement errors and the unobserved true covariates both follow normal distribution. Their method had the flexibility to handle the differential measurement error, that means, the measurement errors that may have different distributions for the case and control groups. Forbes and Santner (1995) developed a method of estimating log-odds ratio parameter for a dichotomous exposure variable while the continuous confounding variables are measured with errors. McShane, Midthune, Dorgan, Freedman, and Carroll (2001) proposed a conditional score approach for handling errors in the potential risk factors. They assumed that the measurement errors followed a normal distribution, otherwise the method did not require any assumption on the distribution of the unobserved true covariate. The last three articles used external calibration data sets to calibrate the error variance. Guolo and Brazzale (2008) presented a simulation-based comparison of the RC, SIMEX, and the likelihood-based approach with a known distribution for the unobserved covariate for the additive measurement errors.

All the above-mentioned methods are particularly designed to handle normal measurement errors. Of course in the full likelihood-based method of Guolo and Brazzale (2008), one can incorporate any parametric distribution for the measurement errors. We propose a functional approach where parameter estimates are obtained by solving a set of unbiased EEs and the method works as long as the moment-generating function (MGF) of the measurement errors exists. The key concept is to form an unbiased EE in terms of the observed data such that its conditional

expectation is a weighted average of the score functions or the EEs which are used when none of the covariates are measured with errors. Originally, this idea was used by Buzas (1998) for estimating parameters of the Cox proportional hazard model when covariates were measured only with random errors whose means are zero. Along that line Huang and Wang (2000) proposed nonparametric method for analysing the Cox regression model in the presence of replicated measurements of a surrogate variable. Their method is nonparametric because their approach does not require that the distribution of the errors be known. Due to some similarities between the Cox partial likelihood and the conditional logistic likelihood, we adopt Buzas's (1998)'s idea in matched case–control studies, and solve the problem when covariates are measured with systematic and random errors. Like Huang and Wang (2000), we also do not require that the distribution of the errors be known, but our design is somewhat different from that of the Huang and Wang (2000) as we do not have replicated measurements of the surrogate variable in the main study. In addition, in our set-up the true covariate is never observed, even in the calibration data. The important advantage of the proposed method is that we get consistent estimators without making any assumption regarding the distribution of the unobserved true covariate and the errors associated with the main surrogate variable.

In order to illustrate the proposed methodology, we construct a matched case–control data from the NIH-AARP Diet and Health cohort (Schatzkin et al. 2001) using the age at entry into the study and age at menopause as the matching variables, and study the association between breast cancer and non-alcohol energy from total fat. The nutrient intakes are measured via food frequency questionnaire (FFQ) and they involve substantial amount of errors.

A brief outline of the remainder of this article is as follows. In Section 2 we present model and assumptions while the proposed methodology is described in Section 3. In Section 4 we study the asymptotic properties of the estimators. Section 5 contains the data analysis. Section 6 contains a simulation study followed by a discussion given in Section 7.

2. Model and assumption

Suppose that we have a 1:M matched case–control data with n strata, and V is the set of matching variables which are potentially associated with the disease of interest and the exposure variables. Typically in a matched case–control study, we observe Y, a binary disease indicator variable, Z, a $q \times 1$ vector of error-free covariates, and W, an erroneous version of X along with the matching variables. Here, we assume that the dimensions of X and W are the same. We will be using i and j as the index for strata and the subjects within a stratum, respectively, and thus $j = 1, \ldots, (M+1)$ and $i = 1, \ldots, n$. The disease model is

$$\mathrm{pr}(Y_{ij} = 1 | V_i, X_{ij}, Z_{ij}) = H\{\beta_0(V_i) + \beta_1^{\mathrm{T}} X_{ij} + \beta_2^{\mathrm{T}} Z_{ij}\},$$

where $H(u) = \exp(u)/\{1 + \exp(u)\}$, and β_1 and β_2 are the log-odds ratio parameters corresponding to X and Z, respectively. The effect of the matching variables on the disease risk is captured through $\beta_0(V_i)$, which is completely unspecified. For no measurement error scenario, that means when $W = X$, $\beta = (\beta_1^{\mathrm{T}}, \beta_2^{\mathrm{T}})^{\mathrm{T}}$ is estimated by solving $S(\beta) = \sum_{i=1}^{n} S_i(\beta) = 0$, where

$$S_i(\beta) = \sum_{j=1}^{M+1} (Y_{ij} - p_{ij}) \begin{pmatrix} X_{ij} \\ Z_{ij} \end{pmatrix}, \tag{1}$$

and

$$p_{ij} = \frac{\exp(\beta_1^{\mathrm{T}} X_{ij} + \beta_2^{\mathrm{T}} Z_{ij})}{\sum_{k=1}^{M+1} \exp(\beta_1^{\mathrm{T}} X_{ik} + \beta_2^{\mathrm{T}} Z_{ik})}$$

represents the conditional probability that the jth subject in the ith stratum is a case given that there is only one case in the ith stratum. Note that $S(\beta)$ is the score function derived from the conditional logistic regression likelihood (Breslow and Day 1980, p. 251) and $E\{S_i(\beta) \mid V, X, Z\} = 0$. It is clear from (1) that due to the conditioning on $\sum_{j=1}^{M+1} Y_{ij} = 1$, the summands within $S_i(\beta)$ are not independent which makes it different from the score function of the logistic likelihood for a prospectively collected data set.

Suppose that S_i is partitioned into two parts corresponding to β_1 and β_2:

$$S_{1,i}(\beta) = \frac{\sum_{k=1}^{M+1} \exp(\beta_1^{\mathrm{T}} X_{ik} + \beta_2^{\mathrm{T}} Z_{ik}) \sum_{j=1}^{M+1} Y_{ij} X_{ij} - \sum_{k=1}^{M+1} X_{ik} \exp(\beta_1^{\mathrm{T}} X_{ik} + \beta_2^{\mathrm{T}} Z_{ik})}{\sum_{k=1}^{M+1} \exp(\beta_1^{\mathrm{T}} X_{ik} + \beta_2^{\mathrm{T}} Z_{ik})},$$

$$S_{2,i}(\beta) = \frac{\sum_{k=1}^{M+1} \exp(\beta_1^{\mathrm{T}} X_{ik} + \beta_2^{\mathrm{T}} Z_{ik}) \sum_{j=1}^{M+1} Y_{ij} Z_{ij} - \sum_{k=1}^{M+1} Z_{ik} \exp(\beta_1^{\mathrm{T}} X_{ik} + \beta_2^{\mathrm{T}} Z_{ik})}{\sum_{k=1}^{M+1} \exp(\beta_1^{\mathrm{T}} X_{ik} + \beta_2^{\mathrm{T}} Z_{ik})}.$$

For the errors-in-covariates situation, we observe W instead of X. Since we assume that W is a surrogate for X, given X, W and Y are independent. The following linear model is assumed for W:

$$W = \Delta_0 + \Delta_1 X + \Delta_2 Z + \Delta_3 V + U_W, \tag{2}$$

where U_W has a mean of zero and are assumed to be independent across strata and subjects within a stratum. Assume that the MGF for U_W exists but unknown, and we denote it by $\mathcal{M}_W(\cdot)$. Also we assume that the errors U_W is non-differential, i.e. its distribution is independent of the disease status Y. Importantly, it is assumed that the dimensions of W and X are the same, and Δ_1^{-1} exists. Observe that W becomes an unbiased surrogate for X when $\Delta_0 = 0$, $\Delta_2 = 0$, $\Delta_3 = 0$, and $\Delta_1 = I_p$. Define $\theta = (\mathrm{vec}^{\mathrm{T}}(\Delta_0), \mathrm{vec}^{\mathrm{T}}(\Delta_1), \mathrm{vec}^{\mathrm{T}}(\Delta_2), \mathrm{vec}^{\mathrm{T}}(\Delta_3))^{\mathrm{T}}$, where vec represents the vectorisation operator.

Following our data example, we consider the scenario where the external data contain repeated measures of an unbiased surrogate variable T along with W, Z, V. More specifically, for our data example, the calibration data are D_l^{calib}, $l = 1, \ldots, m$, where $D_l^{\mathrm{calib}} = (V_l, W_l, Z_l, T_{lk}, k = 1, \ldots, K)$. Along with model (2), we assume that

$$T_{lk} = X_l + U_{T,lk},$$

and the distribution of U_T is known. We assume $U_{T,lk} \overset{\mathrm{iid}}{\sim} \mathrm{Normal}(0, \Sigma_T)$. Hence, the MGF of U_T is $\mathcal{M}_T(t) = \exp(t^{\mathrm{T}} \Sigma_T t/2)$. Observe that $T_{lk} - T_{lk'} = U_{T,lk} - U_{T,lk'}$ for $k \neq k' = 1, \ldots, K$. Therefore, based on the calibration data we can readily estimate $\mathcal{M}_T(t)\mathcal{M}_T(-t)$. But the estimation of $\mathcal{M}_T(t)$ is not so obvious. Therefore, the scenario of unknown $\mathcal{M}_T(t)$ is definitely a non-trivial problem, and we will pursue it in a future article. In addition, we assume that all the models hold for both the external calibration and the matched case–control data with the same parameter values (transportability assumption).

3. Estimation methodology

3.1. *EE method*

Define $\psi(\beta,\theta)=\partial\log\{\mathcal{M}_W(\beta_1^{\mathrm{T}}\Delta_1^{-1})\}/\partial\beta_1$, $S_{\mathrm{new}}=\sum_{i=1}^{n}S_{i,\mathrm{new}}$ and $S_{i,\mathrm{new}}\equiv S_{i,\mathrm{new}}(\beta,\theta,\psi(\beta,\theta))=\{S_{1,i,\mathrm{new}}^{\mathrm{T}}(\beta,\theta,\psi(\beta,\theta)),S_{2,i,\mathrm{new}}^{\mathrm{T}}(\beta,\theta)\}^{\mathrm{T}}$, where

$$
\begin{aligned}
S_{1,i,\mathrm{new}}(\beta,\theta,\psi(\beta,\theta)) &= \left\{\sum_{j=1}^{M+1}Y_{ij}X_{ij}^{+}\sum_{k=1}^{M+1}\exp(\beta_1^{\mathrm{T}}X_{ik}^{+}+\beta_2^{\mathrm{T}}Z_{ik})+\psi(\beta,\theta)\sum_{j=1}^{M+1}(1-Y_{ij})\right.\\
&\quad\left.\times\exp(\beta_1^{\mathrm{T}}X_{ij}^{+}+\beta_2^{\mathrm{T}}Z_{ij})-\sum_{j=1}^{M+1}X_{ij}^{+}\exp(\beta_1^{\mathrm{T}}X_{ij}^{+}+\beta_2^{\mathrm{T}}Z_{ij})\right\}\\
&\quad\div\sum_{k=1}^{M+1}\exp\{\beta_1^{\mathrm{T}}E(X_{ik}|Z_{ik},V_i)+\beta_2^{\mathrm{T}}Z_{ik}\}\\
S_{2,i,\mathrm{new}}(\beta,\theta) &= \left\{\sum_{k=1}^{M+1}\exp(\beta_1^{\mathrm{T}}X_{ik}^{+}+\beta_2^{\mathrm{T}}Z_{ik})\sum_{j=1}^{M+1}Y_{ij}Z_{ij}-\sum_{k=1}^{M+1}Z_{ik}\exp(\beta_1^{\mathrm{T}}X_{ik}^{+}+\beta_2^{\mathrm{T}}Z_{ik})\right\}\\
&\quad\div\sum_{k=1}^{M+1}\exp\{\beta_1^{\mathrm{T}}E(X_{ik}|Z_{ik},V_i)+\beta_2^{\mathrm{T}}Z_{ik}\},
\end{aligned}
$$

and $X_{ik}^{+}=\Delta_1^{-1}(W_{ik}-\Delta_0-\Delta_2Z_{ik}-\Delta_3V_i)$. Let $Q_i\equiv\sum_{j=1}^{M+1}\exp(\beta_1^{\mathrm{T}}X_{ij}+\beta_2^{\mathrm{T}}Z_{ij})/\sum_{j=1}^{M+1}\exp\{\beta_1^{\mathrm{T}}E(X_{ij}\,|Z_{ij},V_i)+\beta_2^{\mathrm{T}}Z_{ij}\}$. Then it is easy to check that for $r=1,2$, $E\{S_{r,i,\mathrm{new}}(\beta,\theta,\psi(\beta,\theta))\mid V,X,Y,Z\}=S_{r,i}(\beta)\mathcal{M}_W(\beta_1^{\mathrm{T}}\Delta_1^{-1})Q_i$. That means the conditional expectation of the estimating functions S_{new} with respect to the conditional distribution of W given V,X, and Z yields a weighted average of the score functions which are obtained from the conditional logistic likelihood function when all covariates are measured without any error. Since $E\{S_{r,i}(\beta)\mid V,X,Z\}=0$, the proposed EEs S_{new} are unbiased.

The denominators of $S_{i,\mathrm{new}}(\beta,\theta,\psi(\beta,\theta))$ act as a weight function, and the choice of this weight will not affect the consistency of the estimators. The following argument clarifies the issue related to the choice of the denominator of $S_{i,\mathrm{new}}$. Define $S_i\equiv S_i(\beta)$ and $S_i^{*}\equiv S_{i,\mathrm{new}}(\beta,\theta,\psi(\beta,\theta))\sum_{k=1}^{M+1}\exp\{\beta_1^{\mathrm{T}}E(X_{ik}|Z_{ik},V_i)+\beta_2^{\mathrm{T}}Z_{ik}\}/G(Z,V,\beta)$. The optimal choice of $G(Z,V,\beta)$, the denominator of S_i^{*}, can be obtained by minimising $E(S_i-S_i^{*})(S_i-S_i^{*})^{\mathrm{T}}$. Following the arguments of Buzas (1998), one can show that the approximate optimal choice of the denominator is $\sum_{k=1}^{M+1}\exp\{\beta_1^{\mathrm{T}}E(X_{ik}|Z_{ik},V_i)+\beta_2^{\mathrm{T}}Z_{ik}\}$. Now, $E(X|V,Z)$ along with θ and Σ_T need to be estimated from the calibration data.

Suppose that we model $E(X|V,Z)$ in terms of a finite-dimensional parameter, and assume that $E(X|V,Z)=\Gamma_0+\Gamma_1V+\Gamma_2Z$. Define $\gamma\equiv(\mathrm{vec}^{\mathrm{T}}(\Gamma_0),\mathrm{vec}^{\mathrm{T}}(\Gamma_1),\mathrm{vec}^{\mathrm{T}}(\Gamma_2))^{\mathrm{T}}$, and from now and on we will refer to $S_{\mathrm{new}}(\beta,\theta,\psi(\beta,\theta))$ by $S_{\mathrm{new}}(\beta,\theta,\gamma,\psi(\beta,\theta))$. Suppose that $\hat{\theta}$ $\hat{\gamma}$ and $\hat{\Sigma}_T$ are the $\sqrt{m}$-consistent regular estimator of θ, γ and Σ_T, respectively, obtained from the calibration data, and by $\hat{\psi}(\beta,\theta,\Sigma_T)$ we denote a $\sqrt{m}$ consistent estimator of $\psi(\beta,\theta)$ when β,θ, and Σ_T are known.

We propose to estimate β by solving $S_{\mathrm{new}}(\beta,\hat{\theta},\hat{\gamma},\hat{\psi}(\beta,\hat{\theta},\hat{\Sigma}_T))=\sum_{i=1}^{n}S_{i,\mathrm{new}}(\beta,\hat{\theta},\hat{\gamma},\hat{\psi}(\beta,\hat{\theta},\hat{\Sigma}_T))=0$, where $S_{i,\mathrm{new}}(\beta,\hat{\theta},\hat{\gamma},\hat{\psi}(\beta,\hat{\theta},\hat{\Sigma}_T))=\{S_{1,i,\mathrm{new}}^{\mathrm{T}}(\beta,\hat{\theta},\hat{\gamma},\hat{\psi}(\beta,\hat{\theta},\hat{\Sigma}_T)),S_{2,i,\mathrm{new}}^{\mathrm{T}}(\beta,\hat{\theta},\hat{\gamma})\}^{\mathrm{T}}$. For solving the above equation, we adopted the Newton–Raphson procedure.

The proposed EEs not necessarily admit an unique solution because for a given β_2, $S_{1,\mathrm{new}}$ is not a monotone function of β_1, when X is a scalar variable. However, the asymptotic unbiasedness of the EE and the regularity conditions ensure that there is a sequence of roots of the EE,

$S_{\text{new}}(\beta, \hat{\theta}, \hat{\gamma}, \hat{\psi}(\beta, \hat{\theta}, \hat{\Sigma}_T)) = 0$ which converges to the true parameter with probability 1. For our data examples and simulation studies, we found the root by iterating the Newton– Raphson procedure and there we set the RC estimates as the initial value of the parameters. However, we suggest to follow Stefanski and Carroll (1987)'s recommendation in case of multiple solutions.

3.2. *Estimation of the secondary model parameters θ, γ, and Σ_T*

The true nutrient intakes (X) are never observed. We will obtain $\sqrt{m}$-consistent estimator of θ and γ by solving the following set of *corrected* EEs (Carroll et al. 2006, Chapter 7):

$$\begin{aligned}
&\sum_{l=1}^{m}(W_l - \Delta_0 - \Delta_1\bar{T}_l - \Delta_2 Z_l - \Delta_3 V_l) = 0,\\
&\sum_{l=1}^{m}\left\{(W_l - \Delta_0 - \Delta_1\bar{T}_l - \Delta_2 Z_l - \Delta_3 V_l)\bar{T}_l^{\mathrm{T}} + \frac{1}{K}\hat{\Sigma}_T\Delta_1\right\} = 0,\\
&\sum_{l=1}^{m}(W_l - \Delta_0 - \Delta_1\bar{T}_l - \Delta_2 Z_l - \Delta_3 V_l)Z_l^{\mathrm{T}} = 0,\\
&\sum_{l=1}^{m}(W_l - \Delta_0 - \Delta_1\bar{T}_l - \Delta_2 Z_l - \Delta_3 V_l)V_l^{\mathrm{T}} = 0,\\
&\sum_{l=1}^{m}(\bar{T}_l - \Gamma_0 - \Gamma_1 Z_l - \Gamma_2 V_l) = 0,\\
&\sum_{l=1}^{m}(\bar{T}_l - \Gamma_0 - \Gamma_1 Z_l - \Gamma_2 V_l)Z_l^{\mathrm{T}} = 0,\\
&\sum_{l=1}^{m}(\bar{T}_l - \Gamma_0 - \Gamma_1 Z_l - \Gamma_2 V_l)V_l = 0,
\end{aligned} \tag{3}$$

where $\hat{\Sigma}_T \equiv \sum_{l=1}^{m}\sum_{j=1}^{K}(T_{lj} - \bar{T}_l)(T_{lj} - \bar{T}_l)^{\mathrm{T}}/\{m(K-1)\}$ and $\bar{T}_l = \sum_{j=1}^{K} T_{lj}/K$. Due to measurement errors in T, instead of ordinary least-squares method we use the corrected EEs in (3).

3.3. *Handling unknown MGF of U_W*

Observe that for given β_1, θ, and Σ_T,

$$\widehat{\mathcal{M}_W(\beta_1^{\mathrm{T}}\Delta_1^{-1})} \equiv \left\{\mathcal{M}_T\left(-\frac{\beta_1}{K}\right)\right\}^{-K}\frac{1}{m}\sum_{l=1}^{m}\exp(\beta_1^{\mathrm{T}}\Delta_1^{-1}W_l^*) \xrightarrow{\text{a.s}} \mathcal{M}_W(\beta_1^{\mathrm{T}}\Delta_1^{-1}),$$

$$\begin{aligned}
\frac{\partial\widehat{\mathcal{M}_W(\beta_1^{\mathrm{T}}\Delta_1^{-1})}}{\partial\beta_1} &\equiv \left\{\mathcal{M}_T\left(-\frac{\beta_1}{K}\right)\right\}^{-K}\frac{1}{m}\sum_{l=1}^{m}\Delta_1^{-1}W_l^*\exp(\beta_1^{\mathrm{T}}\Delta_1^{-1}W_l^*) - \frac{\Sigma_T\beta_1}{K}\widehat{\mathcal{M}_W(\beta_1^{\mathrm{T}}\Delta_1^{-1})}\\
&= \left\{\mathcal{M}_T\left(-\frac{\beta_1}{K}\right)\right\}^{-K}\frac{1}{m}\sum_{l=1}^{m}\left(\Delta_1^{-1}W_l^* - \frac{\Sigma_T\beta_1}{K}\right)\\
&\quad\times\exp(\beta_1^{\mathrm{T}}\Delta_1^{-1}W_l^*) \xrightarrow{\text{a.s}} \frac{\partial\mathcal{M}_W(\beta_1^{\mathrm{T}}\Delta_1^{-1})}{\partial\beta_1},
\end{aligned}$$

where $W_l^* \equiv (W_l - \Delta_0 - \Delta_1 \bar{T}_l - \Delta_2 Z_l - \Delta_3 V_l)$. Then

$$\hat{\psi}(\beta, \theta, \hat{\Sigma}_T) \equiv \frac{\sum_{l=1}^m (\Delta_1^{-1} W_l^* - \hat{\Sigma}_T \beta_1 / K) \exp(\beta_1^{\mathrm{T}} \Delta_1^{-1} W_l^*)}{\sum_{l=1}^m \exp(\beta_1^{\mathrm{T}} \Delta_1^{-1} W_l^*)}$$

is a consistent estimator of $\psi(\beta, \theta)$ provided $\mathcal{M}_W(\beta_1^{\mathrm{T}} \Delta_1^{-1}) > 0$. Apparently, it seems that $\hat{\psi}(\beta, \theta, \hat{\Sigma}_T)$ is free from $\mathcal{M}_T(\cdot)$. However, the second term in the numerator of $\hat{\psi}(\beta, \theta, \hat{\Sigma}_T)$ is $\partial \mathcal{M}_T^K(-\beta_1/K)/\partial \beta_1 = \Sigma_T \beta_1 / K$.

4. Asymptotic results

For asymptotic derivations, we assume that there is one case and M controls in every stratum, i.e. $\sum_{j=1}^{M+1} Y_{ij} = 1$. Also, we assume that M, the number of controls in each stratum, is fixed while $n, m \to \infty$. However, the ratio (n/m) converges to a finite positive constant as $n, m \to \infty$.

First, we define a few notations needed for the asymptotic derivations. Define $\bar{X}_i = \sum_{j=1}^{M+1} X_{ij} p_{ij}$, $\bar{Z}_i = \sum_{j=1}^{M+1} Z_{ij} p_{ij}$, and

$$A \equiv E\left\{ \frac{\partial S_{i,\mathrm{new}}(\beta, \theta, \gamma, \psi(\beta, \theta))}{\partial \beta} \right\} = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^{\mathrm{T}} & A_{22} \end{pmatrix},$$

where $A_{11} = -\mathcal{M}_W(\beta_1^{\mathrm{T}} \Delta_1^{-1}) E[\{\sum_{j=1}^{M+1} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^{\mathrm{T}} p_{ij}\} Q_i]$, $A_{12} = -\mathcal{M}_W(\beta_1^{\mathrm{T}} \Delta_1^{-1}) E[\{\sum_{j=1}^{M+1} (X_{ij} - \bar{X}_i)(Z_{ij} - \bar{Z}_i)^{\mathrm{T}} p_{ij}\} Q_i]$ and $A_{22} = -\mathcal{M}_W(\beta_1^{\mathrm{T}} \Delta_1^{-1}) E[\{\sum_{j=1}^{M+1} (Z_{ij} - \bar{Z}_i)(Z_{ij} - \bar{Z}_i)^{\mathrm{T}} p_{ij}\} Q_i]$. Observed data in stratum i are denoted by $D_i^{\mathrm{obs}} = \{Y_{ij}, Z_{ij}, W_{ij}, j = 1, \ldots, (M+1), V_i\}$. Define $A_2 \equiv E\{\partial S_{i,\mathrm{new}}(\beta, \theta, \gamma, \hat{\psi}(\beta, \theta, \Sigma_T))/\partial \theta\}$ and $A_3 \equiv E\{\partial S_{i,\mathrm{new}}(\beta, \theta, \gamma, \psi(\beta, \theta))/\partial \psi(\beta, \theta)\}$.

PROPOSITION 1 *Under the regularity conditions listed in the appendix,*

(i) $S_{\mathrm{new}} = 0$ *admits a sequence of consistent solutions* $\hat{\beta}_n$.
(ii) $\sqrt{n}(\hat{\beta}_n - \beta) = A^{-1} n^{-1/2} \sum_{i=1}^n S_{i,\mathrm{new}}\ (\beta, \theta, \gamma, \psi(\beta, \theta)) + \rho A^{-1} m^{-1/2} \sum_{l=1}^m \{A_2 F^{-1} S_{l,\mathrm{calib}} + A_3 a(D_l^{\mathrm{calib}})\} + o_p(1)$, *where* $\rho = \lim_{n,m \to \infty} \sqrt{n/m}$, *and* $F = E(F_l)$. *The expressions for* $S_{l,\mathrm{calib}}$, F_l *and* $a(D_l^{\mathrm{calib}})$ *are given in the appendix.*
(iii) *If* $\mathrm{var}\{S_{1,\mathrm{new}}(\beta, \theta, \gamma, \psi(\beta, \theta))\} < \infty$ *and* $\mathrm{var}\{A_2 F^{-1} S_{1,\mathrm{calib}}(\theta) + A_3 a(D_1^{\mathrm{calib}})\} < \infty$, *then the asymptotic variance of* $\sqrt{n}(\hat{\beta}_n - \beta)$ *is* $A^{-1}[\mathrm{var}\{S_{1,\mathrm{new}}(\beta, \theta, \gamma, \psi(\beta, \theta))\} + \rho^2 \mathrm{var}\{A_2 F^{-1} S_{1,\mathrm{calib}}(\theta) + A_3 a(D_1^{\mathrm{calib}})\}] A^{-\mathrm{T}}$.

The variability due to the estimation of θ is accounted in $\mathrm{var}(\hat{\beta}_n)$ through $S_{l,\mathrm{calib}}$, and $a(D_l^{\mathrm{calib}})$ can be attributed to the estimation of $\psi(\beta, \theta)$. A brief outline of the proof of Proposition 1 is given in the appendix.

A consistent estimate of the asymptotic variance can be obtained by replacing the true parameters by their estimates, and by replacing ρ by $\hat{\rho} = \sqrt{n/m}$, F by $\hat{F} = m^{-1} \sum_{l=1}^m F_l$, $\psi(\beta, \theta)$ by $\hat{\psi}(\hat{\beta}, \hat{\theta}, \hat{\Sigma}_T)$, A_2 by $\hat{A}_2 = n^{-1} \sum_{i=1}^n \{\partial S_{i,\mathrm{new}}(\beta, \theta, \gamma, \hat{\psi}(\beta, \theta, \Sigma_T))/\partial \theta\}$, A_3 by $\hat{A}_3 =$

$n^{-1}\sum_{i=1}^{n}\{\partial S_{i,\mathrm{new}}(\beta,\theta,\gamma,\psi(\beta,\theta))/\partial\psi(\beta,\theta)\}$, A by

$$\hat{A}=\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix}\left[\frac{\partial}{\partial\beta}\left\{\sum_{j=1}^{M+1}Y_{ij}X_{ij}^{+}\sum_{k=1}^{M+1}\exp(\beta_1^{\mathrm{T}}X_{ik}^{+}+\beta_2^{\mathrm{T}}Z_{ik})+\hat{\psi}(\beta,\theta,\hat{\Sigma}_T)\sum_{j=1}^{M+1}(1-Y_{ij})\right.\right.\\ \left.\left.\times\ \exp(\beta_1^{\mathrm{T}}X_{ij}^{+}+\beta_2^{\mathrm{T}}Z_{ij})-\sum_{j=1}^{M+1}X_{ij}^{+}\exp(\beta_1^{\mathrm{T}}X_{ij}^{+}+\beta_2^{\mathrm{T}}Z_{ij})\right\}\right]\\ \div\sum_{k=1}^{M+1}\exp\{\beta_1^{\mathrm{T}}E(X_{ik}|Z_{ik},V_i)+\beta_2^{\mathrm{T}}Z_{ik}\}\\ \left[\frac{\partial}{\partial\beta}\left\{\sum_{k=1}^{M+1}\exp(\beta_1^{\mathrm{T}}X_{ik}^{+}+\beta_2^{\mathrm{T}}Z_{ik})\sum_{j=1}^{M+1}Y_{ij}Z_{ij}-\sum_{k=1}^{M+1}Z_{ik}\exp(\beta_1^{\mathrm{T}}X_{ik}^{+}+\beta_2^{\mathrm{T}}Z_{ik})\right\}\right]\\ \div\sum_{k=1}^{M+1}\exp\{\beta_1^{\mathrm{T}}E(X_{ik}|Z_{ik},V_i)+\beta_2^{\mathrm{T}}Z_{ik}\}\end{pmatrix},$$

and $\mathcal{M}_T^K(-\beta_1/K)\mathcal{M}_W(\beta_1^{\mathrm{T}}\Delta_1^{-1})$ in $a(D_l^{\mathrm{calib}})$ by $m^{-1}\sum_{l'=1}^{m}\exp(\beta_1^{\mathrm{T}}\Delta_1^{-1}W_{l'}^{*})$.

5. Analysis of the NIH-AARP diet and health study

5.1. *Background information of the data*

The NIH-AARP Diet and Health Study was developed at the National Cancer Institute of the National Institutes of Health to improve our understanding of the relationship between diet and health (Schatzkin et al. 2001). From 1995 through 1996, 3.5 million questionnaires were mailed to current members of the AARP who were aged between 50 and 71 years, and the data were collected prospectively on the exposure of risk factors and occurrence of any type of cancer. Here, we are particularly interested in the risk of breast cancer and its association with the percentage of non-alcohol energy from total fat. A part of this data set has been analysed by Sinha, Mallick, Kipnis, and Carroll (2010) in the context of measurement error in a nonparametric regression set-up of a logistic model.

There were 226,736 women in the study, and after excluding the subjects with missing values and or very extreme values of a variable we are left out with 167,331 women of which 4049 developed breast cancer disease as of 31 December 2000. However, due to lack of matched controls we are able to include only 4007 case subjects in the analysis, and for each case we choose three controls by matching the age at natural menopause (V_1) and the age at entry (V_2). Thus, we analysed an 1:3 matched case–control data set with $n = 4007$ strata.

5.2. *Description of the variables*

The disease variable Y takes on 1 or 0 for the presence and absence of breast cancer, respectively. The age at natural menopause was a categorical variable which took values 0, 1, 2, and 3 for actual age of menopause between 50 and 54 years, less than 45 years, between 45 and 49 years, and more than 55 years, respectively. We considered $Z_1 \equiv (\mathrm{BMI}-25)/6$, total years of replacement hormone used (Z_2), and the number of live born children (Z_3) as the other potential risk factors which were measured without error. Here, Z_2 takes on 0, 1, 2, and 3 for never used hormones, less than 5 years, between 5 and 9 years, more than or equal to 10 years of replacement hormone

used, respectively, and Z_3 takes on 0, 1, 2, 3, and 4 for no children, one child, two children, three children, and four or more children, respectively.

Since the nutrient intakes measured via FFQ involve errors, the cohort data are accompanied by a prospectively collected calibration data set which contained two 24-h recalls along with other risk factor questionnaires. The calibration data contained 1953 subjects including men and women, and we considered only the postmenopausal women who had not used medicine or surgery to have menopause. The logarithm of the percentage of non-alcohol energy from total fat measured via 24-h recall will be considered as an unbiased surrogate variable (T) for the true logarithm of percentage of non-alcohol energy from total fat (X), and it is defined as $X = \log[\text{totalfatintakeingrams} \times 900/\{\text{totalenergy} - (\text{alcoholintakeingrams} \times 7)\}]$. Figure 1 shows the histogram of W, the logarithm of percentage of non-alcohol energy from total fat measured via FFQ, among the cases and controls of the matched data set.

5.3. *Method of analyses*

We analyse the data by using the three methods: (1) naive (NV), (2) RC, and (3) the proposed EE method. The standard errors of the estimators for the RC method are calculated by drawing 100 Bootstrap samples from the calibration data and from the matched case–control data.

5.4. *Results of the analyses*

The results of all analyses are presented in Table 1. The parameter estimates are very much consistent for all three methods except that for the percentage of non-alcohol energy from total fat which has odds-ratio of $\exp(0.11) = 1.116$, $\exp(0.198) = 1.219$, and $\exp(0.253) = 1.287$, for the NV, RC, the EE approach, respectively. However, none of the approaches shows any significant effect of non-alcohol energy from total fat at 5% level. The standard error of the estimator for the EE approach is higher than the other approaches. If all the other factors remain fixed, the risk of breast cancer is increased by 10% when the BMI increases 6 units. Overall, the disease risk

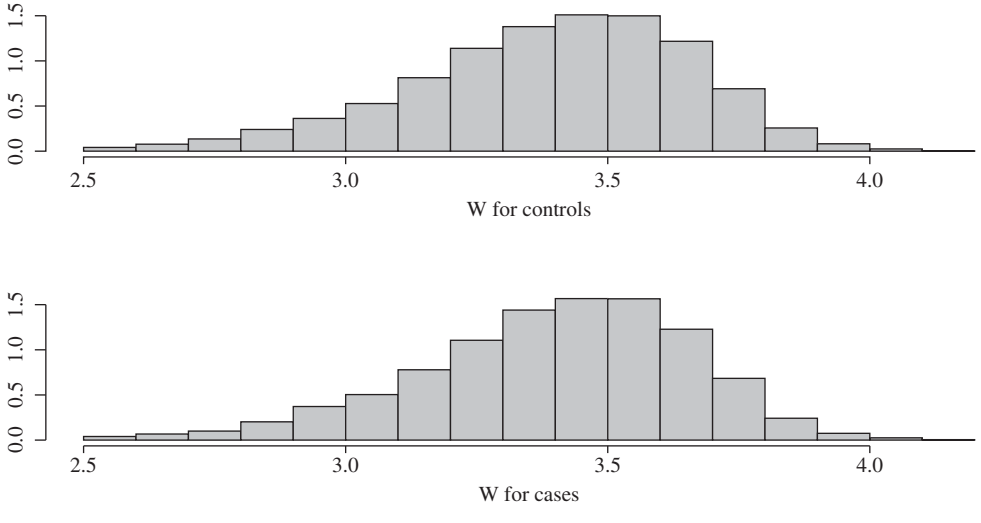

Figure 1. Histogram of logarithm of percentage of non-alcohol energy from total fat (W) in the cohort data.

Table 1. Analysis of the matched case–control data constructed from the NIH-AARP Diet and Health study.

	NV			RC			EE		
Variable	EST	SE	p-Value	EST	SE	p-Value	EST	SE	p-Value
Z_1	0.093 (0.052, 0.134)	0.021	<0.001	0.092 (0.053, 0.131)	0.020	<0.001	0.093 (0.052, 0.134)	0.021	<0.001
Hormone years 1 *vs.* 0	0.219 (0.117, 0.321)	0.052	<0.001	0.226 (0.129, 0.322)	0.049	<0.001	0.225 (0.121, 0.328)	0.053	<0.001
Hormone years 2 *vs.* 0	0.341 (0.227, 0.455)	0.058	<0.001	0.347 (0.231, 0.462)	0.059	<0.001	0.359 (0.243, 0.475)	0.059	<0.001
Hormone years 3 *vs.* 0	0.284 (0.180, 0.387)	0.053	<0.001	0.287 (0.191, 0.383)	0.049	<0.001	0.292 (0.188, 0.396)	0.053	<0.001
Live child 1 *vs.* 0	−0.205 (−0.346, −0.063)	0.072	0.005	−0.219 (−0.368, −0.070)	0.076	0.004	−0.221 (−0.366, −0.076)	0.074	0.003
Live child 2 *vs.* 0	−0.248 (−0.362, −0.134)	0.058	<0.001	−0.253 (−0.367, −0.139)	0.058	<0.001	−0.256 (−0.372, −0.140)	0.059	<0.001
Live child 3 *vs.* 0	−0.306 (−0.411, −0.200)	0.054	<0.001	−0.313 (−0.419, −0.207)	0.054	<0.001	−0.313 (−0.420, −0.205)	0.055	<0.001
Live child 4 or more *vs.* 0	−0.524 (−0.667, −0.381)	0.073	<0.001	−0.527 (−0.688, −0.366)	0.082	<0.001	−0.525 (−0.670, −0.379)	0.074	<0.001
X	0.110 (−0.029, 0.249)	0.071	0.120	0.198 (−0.027, 0.423)	0.115	0.085	0.253 (−0.082, 0.588)	0.171	0.139

Note: Here, X represents the log of the percentage of non-alcohol energy from total fat and $Z_1 = (\text{BMI} - 25)/6$, and BMI is measured in kg/m^2. Hormone years is a categorical variable for total years of replacement hormone used. Hormone years 0, 1, 2, and 3 represent no replacement hormone used, replacement hormone used for less than 5 years, replacement hormone used for 5–9 years, and replacement hormone used for more than 10 years, respectively. Live child represents the number of live children. Here, EST and SE represent the estimate and the estimated standard error, respectively. Also, NV, RC, and EE represent the naive, regression calibration, and the EE approach, respectively. The Wald-type 95% confidence intervals are given right beneath the point estimates.

increases with the years of replacement hormone used, and the risk decreases with the number of live born children. Table 1 contains theoretical p-values based on the asymptotic distribution of the estimators. For the proposed method, we also calculated the empirical p-values based on 10,000 Bootstrap samples, and the distributions of the estimators under the null hypothesis of no association between the disease and the covariates are presented in Figure 2. According to the order of the covariates presented in Table 1, the empirical p-values are $0, 0, 0, 0, 0.0024, 0, 0, 0$, and 0.1324, respectively. For creating Bootstrap samples under the null hypothesis (Field and Welsh 2007), within each stratum we randomly assign $Y = 1$ to one of the four subjects and the other three subjects receive $Y = 0$.

6. Simulation study

6.1. *Simulation designs*

First, we simulated a cohort, and from there we constructed a matched case–control data. We simulated a cohort of size $N = 20,000$ by simulating V, Z, X, and Y. From each cohort, we constructed an 1:2 matched case–control data with $n = 300$ strata. Here, $V \sim \text{Normal}(0, 1)$, and mimicking the distribution of the body mass index in the real data we simulated Z from $\text{Weibull}(3, 2.95) - 2.54$ distribution. For scenarios 1 and 2, we set $X = 3.322 + 0.1\cos(\pi V) + \vartheta$, where $\vartheta \sim \text{Gamma}(2, 3.5)$ with mean $E(\vartheta) = \frac{2}{3.5}$. We took $\text{logit}\{\text{pr}(Y = 1|V, X, Z)\} = \beta_0 + 0.2V + \beta_1 X + 0.1Z$, with two choices for β_1, 0.5 and 1, and varied β_0 so that marginally $\text{pr}(Y = 1) \approx 10\%$. For the surrogate variable, we took the model $W = 0.538 + 0.083V + 0.1Z + 0.84X + U_W$, and we considered six different distributions (1) $U_W \sim \text{Normal}(0, \sigma_W^2)$, (2) $U_W \sim \sigma_W \text{Uniform}(-1.8, 1.8)$, (3) $U_W \sim \sigma_W\{\text{Gamma}(3, \sqrt{3}) - \sqrt{3}\}$, (4) $U_W \sim R\text{Normal}(-0.38, 0.12^2) + (1 - R)\text{Normal}(0.38, 0.12^2)$, (5) $U_W = U_W^* - E(U_W^*)$, where $U_W^* \sim R\text{Gamma}(4.4, 6) + (1 - R)\text{Gamma}(1, 5)$, and (6) $U_W \sim R\text{Normal}(-0.16, 0.48^2) + (1 - R)\text{Gamma}(0.8, 5)$. For scenarios 4, 5, and 6, $R \sim \text{Bernoulli}(0.5)$. We took $\sigma_W^2 = 0.15$. The parameter values somewhat mimic the estimated parameters of the real data analysis.

For all simulation scenarios, we construct a calibration data by drawing a random sample of size $m = 100$ from the simulated cohort. Along with W, Z, V, the calibration data contained

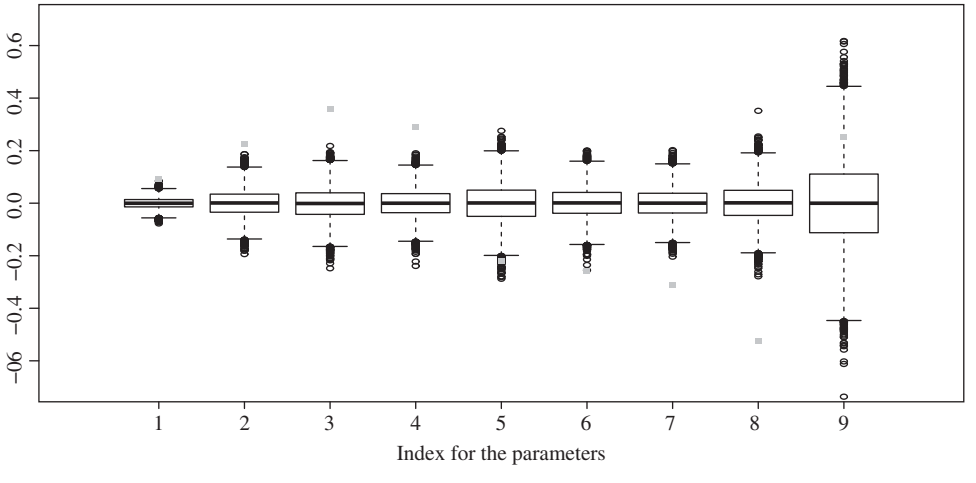

Figure 2. Boxplot of the distribution of the estimators under the null hypothesis of no association. The grey squares denote the estimates of the parameters.

Table 2. Simulation results for the naive (NV), regression calibration (RC), and the EE approach.

	NV		RC		EE	
	$\beta_1 = 0.5$	$\beta_2 = 0.1$	$\beta_1 = 0.5$	$\beta_2 = 0.1$	$\beta_1 = 0.5$	$\beta_2 = 0.1$
$U_W \sim \text{Normal}(0, \sigma_W^2)$						
Mean	0.281	0.071	0.554	0.098	0.510	0.100
Emp. SE	0.138	0.073	0.296	0.075	0.282	0.081
Est. SE	0.136	0.075	0.314	0.081	0.288	0.086
CP	0.629	0.942	0.98	0.963	0.965	0.967
MSE × 10	0.667	0.063	0.905	0.057	0.793	0.066
Median	0.276	0.072	0.532	0.101	0.478	0.102
MAD	0.098	0.049	0.196	0.049	0.172	0.054
	$\beta_1 = 1$	$\beta_2 = 0.1$	$\beta_1 = 1$	$\beta_2 = 0.1$	$\beta_1 = 1$	$\beta_2 = 0.1$
Mean	0.574	0.041	1.132	0.098	1.038	0.105
Emp. SE	0.137	0.078	0.341	0.087	0.376	0.105
Est. SE	0.137	0.077	0.372	0.091	0.368	0.110
CP	0.141	0.874	0.984	0.958	0.925	0.958
MSE × 10	1.998	0.095	1.336	0.076	1.427	0.110
Median	0.570	0.038	1.102	0.099	0.978	0.102
MAD	0.092	0.051	0.223	0.054	0.223	0.064
$U_W \sim \sigma_W \text{Uniform}(-1.8, 1.8)$						
Mean	0.261	0.071	0.537	0.097	0.489	0.098
Emp. SE	0.134	0.076	0.292	0.077	0.269	0.083
Est. SE	0.134	0.075	0.319	0.080	0.271	0.084
CP	0.559	0.937	0.974	0.959	0.951	0.956
MSE × 10	0.753	0.067	0.865	0.059	0.725	0.069
Median	0.259	0.069	0.525	0.098	0.471	0.099
MAD	0.049	0.089	0.189	0.053	0.171	0.056
	$\beta_1 = 1$	$\beta_2 = 0.1$	$\beta_1 = 1$	$\beta_2 = 0.1$	$\beta_1 = 1$	$\beta_2 = 0.1$
Mean	0.552	0.044	1.141	0.101	1.027	0.103
Emp. SE	0.139	0.075	0.348	0.086	0.343	0.107
Est. SE	0.134	0.076	0.374	0.091	0.326	0.107
CP	0.105	0.897	0.973	0.956	0.938	0.958
MSE × 10	2.198	0.088	1.411	0.075	1.185	0.115
Median	0.551	0.043	1.103	0.098	0.988	0.102
MAD	0.097	0.053	0.223	0.058	0.215	0.067

Note: Here, Emp. SE, MAD, Est. SE, MSE, and CP denote empirical standard error, median absolute deviation, estimated standard error, mean-squared error, and 95% coverage probability based on the Wald-type confidence intervals, respectively. Here, $\sigma_W^2 = 0.15$ and $\sigma_T^2 = 0.07$.

$T_1 = X + U_{T1}$ and $T_2 = X + U_{T2}$, where $U_{T1}, U_{T2} \sim \text{Normal}(0, \sigma_T^2)$ with $\sigma_T^2 = 0.07$. Note that X were used only for data generation, and were no longer used in the analyses.

6.2. *Method of analyses*

We simulated $R = 1000$ matched case–control data sets and each data set was accompanied by a calibration data set. Each data set was analysed by the (1) NV, (2) RC, and (3) the proposed EE approach. In the RC approach, we fitted a linear regression model of the average of T_1 and T_2 on V, W, and Z based on the calibration data. The fitted regression was then used to estimate the unobserved X in the matched study. In the EE approach, $E(X|V, Z)$ was modelled as a linear function of V and Z. However, for both scenarios it was a model misspecification as the true regression of X was not a linear function of V. We present the mean, median, empirical standard error (Emp. SE), estimated standard error (Est. SE) based on the asymptotic standard error formula, 95% empirical coverage probabilities based on the Wald-type intervals, and the

Table 3. Simulation results for the naive (NV), regression calibration (RC), and the EE approach.

	NV		RC		EE	
	$\beta_1 = 0.5$	$\beta_2 = 0.1$	$\beta_1 = 0.5$	$\beta_2 = 0.1$	$\beta_1 = 0.5$	$\beta_2 = 0.1$
$U_W \sim \sigma_W\{\text{Gamma}(3, \sqrt{3}) - \sqrt{3}\}$						
Mean	0.270	0.075	0.532	0.103	0.496	0.101
Emp. SE	0.131	0.075	0.278	0.077	0.319	0.089
Est. SE	0.135	0.075	0.311	0.080	0.332	0.090
CP	0.600	0.939	0.972	0.957	0.931	0.963
MSE × 10	0.697	0.062	0.785	0.060	1.018	0.081
Median	0.271	0.073	0.519	0.102	0.456	0.102
MAD	0.088	0.049	0.181	0.051	0.172	0.052
	$\beta_1 = 1$	$\beta_2 = 0.1$	$\beta_1 = 1$	$\beta_2 = 0.1$	$\beta_1 = 1$	$\beta_2 = 0.1$
Mean	0.561	0.041	1.105	0.098	0.935	0.097
Emp. SE	0.132	0.075	0.345	0.086	0.407	0.223
Est. SE	0.135	0.076	0.376	0.091	0.462	0.163
CP	0.089	0.878	0.973	0.96	0.891	0.971
MSE × 10	2.103	0.092	1.301	0.075	1.724	0.109
Median	0.563	0.043	1.068	0.102	0.842	0.103
MAD	0.088	0.049	0.208	0.057	0.217	0.069
	$\beta_1 = 0.5$	$\beta_2 = 0.1$	$\beta_1 = 0.5$	$\beta_2 = 0.1$	$\beta_1 = 0.5$	$\beta_2 = 0.1$
$U_W \sim R\text{Normal}(-0.38, 0.12^2) + (1 - R)\text{Normal}(0.38, 0.12^2)$						
Mean	0.278	0.072	0.571	0.099	0.510	0.098
Emp. SE	0.129	0.076	0.290	0.078	0.256	0.082
Est. SE	0.134	0.075	0.316	0.081	0.261	0.085
CP	0.279	0.992	0.973	1	0.96	0.961
MSE × 10	0.661	0.066	0.893	0.061	0.660	0.068
Median	0.279	0.069	0.558	0.099	0.491	0.098
MAD	0.092	0.052	0.199	0.055	0.177	0.057
	$\beta_1 = 1$	$\beta_2 = 0.1$	$\beta_1 = 1$	$\beta_2 = 0.1$	$\beta_1 = 1$	$\beta_2 = 0.1$
Mean	0.561	0.041	1.146	0.095	1.035	0.094
Emp. SE	0.132	0.076	0.318	0.086	0.341	0.105
Est. SE	0.135	0.076	0.362	0.091	0.317	0.107
CP	0.094	0.875	0.978	0.962	0.942	0.954
MSE × 10	2.094	0.093	1.226	0.074	1.170	0.112
Median	0.559	0.039	1.117	0.097	0.992	0.096
MAD	0.085	0.052	0.213	0.058	0.203	0.066

Note: Here, Emp. SE, MAD, Est. SE, MSE, and CP denote empirical standard error, median absolute deviation, estimated standard error, mean-squared error, and 95% coverage probability based on the Wald-type confidence intervals, respectively. Here, $\sigma_W^2 = 0.15$ and $\sigma_T^2 = 0.07$, and $R \sim \text{Bernoulli}(0.5)$.

median absolute deviation ($\text{MAD} = \text{median}_{1 \le j \le R}|\hat{\beta}_j - \text{median}_{1 \le j \le R}(\hat{\beta}_j)|$). For the RC method, we estimate standard error using 100 Bootstrap samples with replacement.

6.3. *Summary of results*

Tables 2–4 contain results for scenarios 1–6, and they can be summarised as follows. The naive method performed poorly compared with the other methods in terms of bias of $\hat{\beta}_1$ and also of $\hat{\beta}_2$. Overall the proposed method performs well in all the scenarios that we considered here. The results indicate that the EE approach has significantly less bias than the naive and the RC approach, especially when $\beta_1 = 1$. Interestingly, for scenarios 1, 2, 4, and 6 with $\beta_1 = 0.5$, the EE method has somewhat less variance than the RC method. Overall the empirical standard errors for the RC method are smaller than that for the EE method. All tables suggest that the asymptotic standard error formula for the EE method works well. Based on 100 Bootstrap samples, the estimated standard error for the RC method is somewhat higher than the empirical standard error. When U_W

Table 4. Simulation results for the naive (NV), regression calibration (RC), and the EE approach.

	NV		RC		EE	
	$\beta_1 = 0.5$	$\beta_2 = 0.1$	$\beta_1 = 0.5$	$\beta_2 = 0.1$	$\beta_1 = 0.5$	$\beta_2 = 0.1$
$U_W = U_W^* - E(U_W^*)$, $U_W^* \sim R\text{Gamma}(4.4, 6) + (1-R)\text{Gamma}(1, 5)$						
Mean	0.280	0.072	0.559	0.101	0.517	0.100
Emp. SE	0.137	0.075	0.303	0.077	0.333	0.089
Est. SE	0.134	0.075	0.316	0.081	0.340	0.098
CP	0.611	0.931	0.967	0.955	0.946	0.963
MSE × 10	0.673	0.064	0.957	0.060	1.116	0.080
Median	0.277	0.075	0.541	0.102	0.478	0.104
MAD	0.093	0.049	0.197	0.051	0.176	0.055
	$\beta_1 = 1$	$\beta_2 = 0.1$	$\beta_1 = 1$	$\beta_2 = 0.1$	$\beta_1 = 1$	$\beta_2 = 0.1$
Mean	0.563	0.047	1.122	0.102	0.942	0.111
Emp. SE	0.131	0.078	0.343	0.091	0.400	0.127
Est. SE	0.135	0.076	0.373	0.091	0.437	0.144
CP	0.106	0.891	0.974	0.954	0.910	0.966
MSE × 10	2.086	0.089	1.327	0.082	1.634	0.164
Median	0.556	0.047	1.086	0.106	0.948	0.107
MAD	0.085	0.052	0.202 3	0.059	0.211	0.072
	$\beta_1 = 0.5$	$\beta_2 = 0.1$	$\beta_1 = 0.5$	$\beta_2 = 0.1$	$\beta_1 = 0.5$	$\beta_2 = 0.1$
$U_W = R\text{Normal}(-0.16, 0.48^2) + (1-R)\text{Gamma}(0.8, 5)$						
Mean	0.274	0.069	0.560	0.096	0.510	0.094
Emp. SE	0.136	0.079	0.310	0.082	0.308	0.087
Est. SE	0.135	0.075	0.336	0.081	0.321	0.090
CP	0.619	0.914	0.974	0.950	0.949	0.952
MSE × 10	1.967	0.102	1.673	0.085	1.733	0.148
Median	0.275	0.068	0.541	0.092	0.481	0.092
MAD	0.089	0.054	0.186	0.058	0.164	0.058
	$\beta_1 = 1$	$\beta_2 = 0.1$	$\beta_1 = 1$	$\beta_2 = 0.1$	$\beta_1 = 1$	$\beta_2 = 0.1$
Mean	0.565	0.045	1.158	0.102	1.029	0.101
Emp. SE	0.136	0.078	0.384	0.093	0.433	0.121
Est. SE	0.137	0.076	0.411	0.092	0.452	0.127
CP	0.123	0.886	0.981	0.962	0.917	0.952
MSE × 10	2.078	0.090	1.725	0.088	1.882	0.145
Median	0.561	0.045	1.105	0.103	0.965	0.107
MAD	0.092	0.056	0.221	0.066	0.220	0.077

Note: Here, Emp. SE, MAD, Est. SE, MSE, and CP denote empirical standard error, median absolute deviation, estimated standard error, mean-squared error, and 95% coverage probability based on the Wald-type confidence intervals, respectively. Here, $\sigma_W^2 = 0.15$ and $\sigma_T^2 = 0.07$, and $R \sim \text{Bernoulli}(0.5)$.

follows a gamma distribution (scenario 2) and the true $\beta_1 = 1$, the bias in $\hat{\beta}_1$ due to the EE method is somewhat larger than that of the other scenarios. However, further numerical investigation reveals that this bias is due to small sample size of the calibration data because the mean, median, empirical standard error, and estimated standard error of $\hat{\beta}_1$ due to the RC and the EE methods are 1.079, 1.078, 0.270, 0.279, and 0.997, 0.943, 0.337, 0.329, respectively when $m = 500$, and 1.073, 1.075, 0.261, 0.269, and 1.016, 0.979, 0.332, 0.321, respectively for $m = 1000$. Finally, we want to point out that simulation results validate the two aspects of the asymptotic results, the asymptotic normal distribution, and the asymptotic standard error expression (Bosley 1996). The first aspect is validated through the fact that Wald-type 95% coverage probabilities are reasonably close to 0.95. The second aspect is validated through the fact that the estimated standard error measured via the asymptotic formula is close to the empirical standard error. We found that as both n and m increase keeping n/m fixed, the asymptotic and the empirical standard errors get close (results are not presented here) (Figure 3).

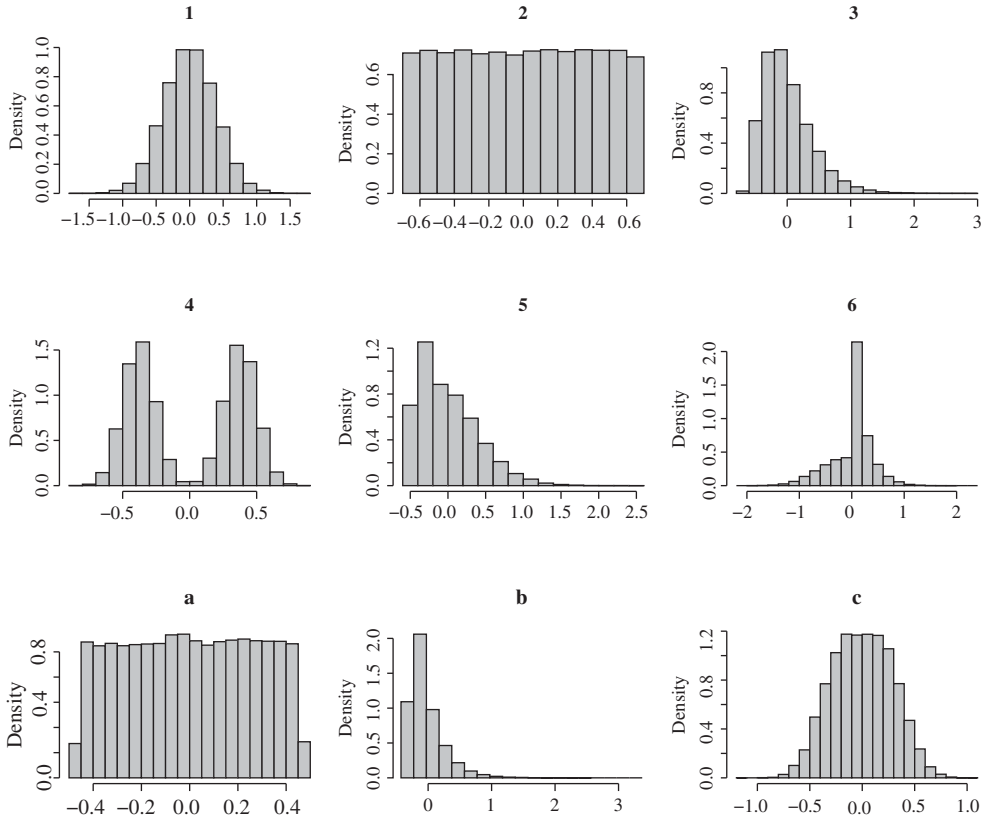

Figure 3. The first two rows contain the distribution of U_W under six different scenarios, and the last row contains the three non-normal distributions of U_T. In all cases $E(U_W) = E(U_T) = 0$, $\text{var}(U_W) = 0.15$, and $\text{var}(U_T) = 0.07$.

6.4. *Robustness study towards the normality assumption of U_T:*

For assessing the performance of the EE approach when the normality assumption is violated for U_T we considered scenario 1 with $\beta_1 = 1$. But for the calibration data U_T is simulated from three non-normal distributions: (a) $\sigma_T\text{Uniform}(-1.75, 1.75)$, (b) $\sigma_T\{\text{Gamma}(1, 1) - 1\}$, and (c) $(\sigma_T/0.26)\{R\text{Normal}(0.2, 0.2^2) + (1 - R)\text{Normal}(-0.2, 0.2^2)\}$ with $R \sim \text{Bernoulli}(0.5)$. However, we analysed the simulated data sets under the normal distribution assumption of U_T. The results presented in Table 5 indicate that the EE method works reasonably well without any visible impact of model misspecification, and in all these cases the RC method shows high bias in the estimates of β_1.

7. Discussion

We have proposed an EE approach for handling errors-in-covariates in matched case–control studies. In particular, here we handle the additive measurement errors. The proposed method produces consistent and asymptotically normally distributed estimator without the knowledge of the distribution of the unobserved true covariates. One of the main advantages of the proposed approach is that the method can handle different types of error distributions, symmetric or asymmetric. Of course the method requires that the MGF of the errors exists. The simulation study indicates the

Table 5. Simulation results for the naive (NV), regression calibration (RC), and the EE approach.

	NV		RC		EE	
	$\beta_1 = 1$	$\beta_2 = 0.1$	$\beta_1 = 1$	$\beta_2 = 0.1$	$\beta_1 = 1$	$\beta_2 = 0.1$
$U_T \sim \sigma_T \text{Uniform}(-1.75, 1.75)$						
Mean	0.574	0.040	1.140	0.099	1.032	0.100
Emp. SE	0.144	0.077	0.359	0.088	0.403	0.117
Est. SE	0.137	0.076	0.369	0.091	0.407	0.116
CP	0.134	0.876	0.97	0.962	0.917	0.959
MSE × 10	2.025	0.096	1.489	0.077	1.631	0.125
Median	0.571	0.039	1.086	0.099	0.967	0.101
MAD	0.093	0.051	0.211	0.056	0.218	0.064
$U_T \sim \sigma_T\{\text{Gamma}(1, 1) - 1\}$						
Mean	0.584	0.038	1.155	0.096	1.026	0.102
Emp. SE	0.142	0.078	0.356	0.092	0.366	0.113
Est. SE	0.137	0.076	0.371	0.091	0.386	0.118
CP	0.167	0.868	0.972	0.954	0.928	0.965
MSE × 10	1.935	0.100	1.507	0.085	1.348	0.129
Median	0.580	0.035	1.116	0.095	0.962	0.103
MAD	0.094	0.055	0.221	0.061	0.204	0.072
$U_T \sim (\sigma_T/0.26)\{R\text{Normal}(0.2, 0.2^2) + (1 - R)\text{Normal}(-0.2, 0.2^2)\}$						
Mean	0.573	0.039	1.135	0.097	1.038	0.102
Emp. SE	0.141	0.076	0.351	0.088	0.389	0.110
Est. SE	0.137	0.076	0.370	0.091	0.378	0.112
CP	0.139	0.866	0.968	0.95	0.929	0.959
MSE × 10	2.002	0.094	1.416	0.079	1.529	0.121
Median	0.576	0.042	1.108	0.100	0.968	0.100
MAD	0.094	0.051	0.216	0.056	0.218	0.068

Note: Here, Emp. SE, MAD, Est. SE, MSE, and CP denote empirical standard error, median absolute deviation, estimated standard error, mean-squared error, and 95% coverage probability based on the Wald-type confidence intervals, respectively. Here, $U_W \sim \text{Normal}(0, \sigma_W^2)$, $\sigma_W^2 = 0.15$, and $\sigma_T^2 = 0.07$, and $R \sim \text{Bernoulli}(0.5)$.

advantage of the proposed method over the naive and the RC approach. For estimating parameters using the naive and the RC method, we used `clogit` function of the `R` package `Survival`. For the EE method, we used the Newton–Raphson algorithm using `R` and `Fortran`, and the computer code is available from the author upon request.

Some limitations of the method are as follows. In principle, the EEs can produce multiple roots. Although, this method can handle any type of error distribution, the method requires the existence of the MGF of the errors. We also assume that the unbiased surrogate variables observed only in the calibration data follow a normal distribution conditional on X. The proposed approach does not work well when X and W are weakly associated. Although not considered in this paper, the multiplicative structure of errors may arise in observational studies, and development of a flexible method for handling multiplicative errors-in-covariates is a part of our future research.

Acknowledgements

The author wishes to thank the NIH-AARP study group for providing the NIH-AARP cohort data. He also thanks the editor, associate editor, and a referee for very constructive suggestions which have led to a much improved manuscript. This research was partially supported by NSF grant SES-0961618.

References

Armstrong, B.G., Whittemore, A.S., and Howe, G.R. (1989), 'Analysis of Case–Control Data with Covariate Measurement Error: Application to Diet and Colon Cancer', *Statistics in Medicine*, 8, 1151–1163.

Bosley, D.L. (1996), 'A Technique for the Numerical Verification of Asymptotic Expansions', *SIAM Review*, 38, 128–135.
Breslow, N.E., and Cain, K.C. (1988), 'Logistic Regression for Two-Stage Case-Control Data', *Biometrika*, 75, 11–20.
Breslow, N.E., and Day, N.E. (1980), *Statistical Methods in Cancer Research* (Vol. 1), Lyon, France: International Agency for Research on Cancer.
Buzas, J.S. (1998), 'Unbiased Scores in Proportional Hazards Regression with Covariate Measurement Error', *Journal of Statistical Planning and Inference*, 67, 247–257.
Buzas, J.S., and Stefanski, L.A. (1996), 'A Note on Corrected Score Estimation', *Statistics and Probability Letters*, 28, 1–8.
Carroll, R.J., Ruppert, D., Stepanski, L.A., and Crainiceanu, C. (2006), *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd ed.), New York: Chapman & Hall.
Cheng, K.F., and Hsueh, H.M. (2003), 'Estimation of a Logistic Regression Model with Mismeasured Observations', *Statistica Sinica*, 13, 111–127.
Field, C.A., and Welsh, A.H. (2007), 'Bootsrapping Clustered Data', *Journal of the Royal Statistical Society, Series B*, 69, 369–390.
Forbes, A.B., and Santner, T.J. (1995), 'Estimators of Odds Ratio Regression Parameters in Matched Case–Control Studies with Covariate Measurement Error', *Journal of the American Statistical Association*, 90, 1075–1084.
Foutz, R.V. (1977), 'On the Unique Consistent Solution to the Likelihood Equations', *Journal of the American Statistical Association*, 72, 147–148.
Guolo, A. (2008), 'A Flexible Approach to Measurement Error Correction in Case Control Studies', *Biometrics*, 64, 1207–1214.
Guolo, A., and Brazzale, A.R. (2008), 'A Simulation-Based Comparison of Techniques to Correct for Measurement Error in Matched Case–Control Studies', *Statistics in Medicine*, 27, 3755–3775.
Huang, Y., and Wang, C.Y. (2000), 'Cox Regression with Accurate Covariates Unascertainable: A Nonparametric-Correction Approach', *Journal of the American Statistical Association*, 95, 1209–1219.
Huang, Y., and Wang, C.Y. (2001), 'Consistent Functional Methods for Logistic Regression with Errors in Covariates', *Journal of the American Statistical Association*, 96, 1469–1482.
McShane, L.M., Midthune, D.N., Dorgan, J.F., Freedman, L.S., and Carroll, R.J. (2001), 'Covariate Measurement Error Adjustment for Matched Case–Control Studies', *Biometrics*, 57, 62–73.
Rabe-Hesketh, S., Pickles, A., and Skrondal, A. (2003), 'Correcting for Covariate Measurement Error in Logistic Regression Using Nonparametric Maximum Likelihood Estimation', *Statistical Modelling*, 3, 215–232.
Reilly, M., and Pepe, M.S. (1995), 'A Mean Score Method for Missing and Auxiliary Covariate Data in Regression Models', *Biometrika*, 82, 299–314.
Schatzkin, A., Subar, A.F., Thompson, F.E., Harlan, L.C., Tangrea, J., Hollenbeck, A.R., Hurwitz, P.E., Coyle, L.S.N.M.D.S., Freedman, L.S., Brown, C.C., Midthune, D., and Kipnis, V. (2001), 'Design and Serendipity in Establishing a Large Cohort with Wide Dietary Intake Distributions', *American Journal of Epidemiology*, 154, 1119–1125.
Sinha, S., Mallick, B.K., Kipnis, V., and Carroll, R.J. (2010), 'Semiparametric Bayesian Analysis of Nutritional Epidemiology Data in the Presence of Measurement Error', *Biometrics*, 66, 444–454.
Stefanski, L.A., and Carroll, R.J. (1987), 'Conditional Scores and Optimal Scores for Generalized Linear Measurement-Error Models', *Biometrika*, 74, 703–716.
Sugar, E.A., Wang, C.-Y., and Prentice, R.L. (2007), 'Logistic Regression with Exposure Biomarkers and Flexible Measurement Error', *Biometrics*, 63, 143–151.
van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge, UK: Cambridge University Press.
Wang, C.Y., and Wang, S. (1997), 'Semiparametric Methods in Logistic Regression with Measurement Error', *Statistica Sinica*, 7, 1103–1120.

Appendix

Before we describe the asymptotic properties of the proposed estimator, we write the sufficient regularity conditions. Assume that for each β in an open ball of the Euclidean space

C.1 each component of $E\{S_{i,\text{new}}S_{i,\text{new}}^{\mathrm{T}}\}$ is finite,
C.2 $A = E(\partial S_{i,\text{new}}/\partial\beta)$ exists and is non-singular,
C.3 $\{\partial^2 S_{i,\text{new}}/\partial\beta\partial\beta^{\mathrm{T}}\}$ exists for every data and each entry of this matrix is bounded by a measurable and integrable function, i.e. $|\partial^2 S_{i,\text{new}}/\partial\beta_k\partial\beta_k'| \leq \Psi_{kk'}$, where $\Psi_{kk'}$ is a measurable and integrable function,
C.4 $F = E\{\partial S_{i,\text{calib}}/\partial\theta\}$ exists and is non-singular,
C.5 $\{\partial^2 S_{i,\text{calib}}/\partial\theta\partial\theta^{\mathrm{T}}\}$ exists for every data and each entry of this matrix is bounded by a measurable and integrable function of the calibration data,
C.6 the MGF $\mathcal{M}_W(t)$ exists and is positive at $t = \Delta_1^{-1}\beta_1$.

LEMMA A.1 *For known θ and β,*

$$\sqrt{m}\{\hat{\psi}(\beta,\theta,\hat{\Sigma}_T) - \psi(\beta,\theta)\} = m^{-1/2}\sum_{i=1}^{m} a(D_i^{\text{calib}}) + o_p(1),$$

where

$$a(D_i^{\text{calib}}) \equiv \left\{\mathcal{M}_T^K\left(-\frac{\beta_1}{K}\right)\mathcal{M}_W(\beta_1\Delta_1^{-1})\right\}^{-1}\exp(\Delta_1^{-1}\beta_1 W_i^*)\left(\Delta_1^{-1}W_i^* - \Sigma_T\frac{\beta_1}{K} - \psi(\beta,\theta)\right)$$
$$-\left\{\frac{1}{K-1}\sum_{j=1}^{K}(T_{lj}-\bar{T}_l)(T_{lj}-\bar{T}_l)^{\mathrm{T}} - \Sigma_T\right\}\frac{\beta_1}{K}.$$

Proof Using the Taylor series expansion $\sqrt{m}\{\hat{\psi}(\beta,\theta,\hat{\Sigma}_T)-\psi\} = \sqrt{m}\{\hat{\psi}(\beta,\theta,\Sigma_T)-\psi\} + \sqrt{m}(\hat{\Sigma}_T-\Sigma_T)\{\partial\hat{\psi}(\beta,\theta,\Sigma_T)/\partial\Sigma_T\} + o_p(1)$, where $\{\partial\hat{\psi}(\beta,\theta,\Sigma_T)/\partial\Sigma_T\} = -\beta_1/K$, and

$$\sqrt{m}(\hat{\Sigma}_T-\Sigma_T) = \frac{1}{\sqrt{m}}\sum_{l=1}^{m}\left\{\frac{1}{K-1}\sum_{j=1}^{K}(T_{lj}-\bar{T}_l)(T_{lj}-\bar{T}_l)^{\mathrm{T}} - \Sigma_T\right\}.$$

Now consider the first term of the Taylor expansion. Let $dP_m(\cdot) = m^{-1}\sum_{i=1}^{m} I(\cdot)$, and P be the corresponding population distribution. Then $\hat{\psi}(\beta,\theta,\Sigma_T)$ can be written as $\int(\Delta^{-1}W^* - \Sigma_T\beta_1/2)\exp(\beta_1\Delta_1^{-1}W^*)\,\mathrm{d}P_m/\int\exp(\beta_1\Delta_1^{-1}W^*)\,\mathrm{d}P_m$. Now the Lemma follows from the Hadamard differentiation (van der Vaart 1998) of the function $\int(\Delta^{-1}W^* - \Sigma_T\beta_1/2)\exp(\beta_1\Delta_1^{-1}W^*)\,\mathrm{d}P/\int\exp(\beta_1\Delta_1^{-1}W^*)\,\mathrm{d}P$. ■

LEMMA A.2 *The $\sqrt{m}$-consistent estimator of θ can be expressed as $\sqrt{m}(\hat{\theta}-\theta) = m^{-1/2}F^{-1}(\theta)\sum_{l=1}^{m}S_{l,\text{calib}}(\theta) + o_p(1)$, where $F(\theta) = E\{F_l(\theta)\}$ with*

$$S_{l,\text{calib}}(\theta) = \begin{bmatrix} (W_l - b_l^{\mathrm{T}}\theta)\otimes 1 \\ (W_l - b_l^{\mathrm{T}}\theta)\otimes\bar{T}_l + \dfrac{1}{K(K-1)m}\displaystyle\sum_{j=1}^{K}(T_{lj}-\bar{T}_l)(T_{lj}-\bar{T}_l)^{\mathrm{T}}\Delta_1 \\ (W_l - b_l^{\mathrm{T}}\theta)\otimes Z_l \\ (W_l - b_l^{\mathrm{T}}\theta)\otimes V_l \end{bmatrix},$$

$$F_l(\theta) = \begin{bmatrix} b_l^{\mathrm{T}}\otimes 1 \\ b_l^{\mathrm{T}}\otimes\bar{T}_l + \left(0, \dfrac{1}{K(K-1)}\displaystyle\sum_{j=1}^{K}(T_{lj}-\bar{T}_l)(T_{lj}-\bar{T}_l)^{\mathrm{T}}, 0, 0\right) \\ b_l^{\mathrm{T}}\otimes Z_l \\ b_l^{\mathrm{T}}\otimes V_l \end{bmatrix},$$

and $b_l^{\mathrm{T}} = [I_p : I_p\otimes\bar{T}_l : I_p\otimes Z_l^{\mathrm{T}} : I_p\otimes V_l^{\mathrm{T}}]$, and $\otimes$ denotes the Kronecker product. The proof of this result easily follows from the EEs given in Section 3.2.

Proof of Proposition 1 It can be easily shown that due to the law of large number $(1/n)S_{\text{new}}(\beta,\hat{\theta},\hat{\gamma}\ \hat{\psi}(\beta,\hat{\theta},\hat{\Sigma}_T)) \to 0$ in probability. The key fact for this purpose is the unbiasedness of $S_{\text{new}}(\beta,\theta,\gamma,\psi(\beta,\theta))$. Under the conditions listed above, the consistency of the estimators now follows from straightforward application of Foutz (1977, Theorem 2).

Before we derive the asymptotic distribution of the estimator, we introduce few other notations. Now, using the Taylor series expansion we write

$$\sqrt{n}(\hat{\beta}_n-\beta) = \frac{A^{-1}}{\sqrt{n}}\sum_{i=1}^{n}S_{i,\text{new}}(\beta,\hat{\theta},\hat{\gamma},\hat{\psi}(\beta,\hat{\theta},\hat{\Sigma}_T)) + o_p(1)$$
$$\stackrel{d}{=} A^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}S_{i,\text{new}}(\beta,\theta,\gamma,\psi(\beta,\theta)) + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}S_{i,\text{new}}(\beta,\theta,\gamma,\hat{\psi}(\beta,\theta,\Sigma_T))\frac{\sqrt{n}}{\sqrt{m}}\sqrt{m}(\hat{\theta}-\theta)\right.$$
$$+\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\gamma}S_{i,\text{new}}(\beta,\theta,\gamma,\hat{\psi}(\beta,\theta,\Sigma_T))\frac{\sqrt{n}}{\sqrt{m}}\sqrt{m}(\hat{\gamma}-\gamma)$$
$$\left.+\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\psi}S_{i,\text{new}}(\beta,\theta,\gamma,\psi(\beta,\theta))\frac{\sqrt{n}}{\sqrt{m}}\sqrt{m}\{\hat{\psi}(\beta,\theta,\hat{\Sigma}_T)-\psi(\beta,\theta)\}\right] + o_p(1).$$

Since $E\{\partial S_{i,\text{new}}(\beta,\theta,\gamma\psi(\beta,\theta))/\partial\gamma\} = 0$, and using Lemmas A.1 and A.2, we can write

$$\sqrt{n}(\hat{\beta}_n - \beta) \stackrel{d}{=} \frac{A^{-1}}{\sqrt{n}} \sum_{i=1}^{n} S_{i,\text{new}}(\beta,\theta,\gamma,\psi(\beta,\theta)) + \rho A^{-1} \left\{ \frac{A_2}{\sqrt{m}} F^{-1}(\theta) \sum_{l=1}^{m} S_{l,\text{calib}}(\theta) + \frac{A_3}{\sqrt{m}} \sum_{l=1}^{m} a(D_l^{\text{calib}}) \right\} + o_p(1). \tag{A1}$$

Observe that the first term is independent of the second and third terms on the right-hand side of Equation (A1) as the calibration data are drawn independently of the main data from the population. Now, using the Central Limit Theorem, these two terms asymptotically follow normal distribution. Hence, asymptotically $\sqrt{n}(\hat{\beta}_n - \beta)$ follows a normal distribution with mean zero and variance

$$A^{-1}[\text{var}\{S_{1,\text{new}}(\beta,\theta,\gamma,\psi(\beta,\theta))\} + \rho^2 \text{var}\{A_2 F^{-1} S_{1,\text{calib}}(\theta) + A_3 a(D_1^{\text{calib}})\}]A^{-\text{T}}.$$

Note that when m is very large compared with n, the contribution of the second term to the variance of $\hat{\beta}_n$ is negligible. ■