A Unified Approach to Regression Analysis under Double Sampling Design

By Yi-Hau Chen† and Hung Chen

Department of Mathematics, National Taiwan Normal University and National Taiwan University, Taiwan, R.O.C.

Summary. We propose a unified approach to the estimation of the regression parameters under the double sampling design, in which a primary sample consisting of data on the rough or proxy measures for the response and/or explanatory variables, as well as a validation subsample consisting of data on the exact measurements, are available. We assume that the validation sample is a simple random subsample from the primary sample. Our proposal utilizes a specified parametric model to extract the partial information contained in the primary sample. The resulting estimator is consistent even if such model is misspecified, and achieves higher asymptotic efficiency than the estimator based only on the validation data. Specific cases are discussed to illustrate the application of the proposed estimator.

Keywords: Double sampling; Estimating equation; Generalized linear model; Incomplete data; Validation sample.

†*Address for correspondence*: Department of mathematics, National Taiwan Normal University, No 88, Sec 4, Ting Chou Road, Taipei 117, Taiwan, R.O.C.
E-mail: yhchen@math.ntnu.edu.tw

1 Introduction

It is common in many fields of applications to study the relationship between a response variable y and a set of explanatory variables $\mathbf{x}$ through a regression model for the conditional mean of y given $\mathbf{x}$. Due to the limitation of cost, sometimes it is prohibitive to collect data on $(y, \mathbf{x})$ for all study participants, therefore a double or two-stage sampling design may be adopted by investigators as a cost-effective strategy. By this design, a primary sample is drawn from the target population at the first stage of the study, wherein data on rough or proxy, but less expensive, measurement $(\tilde{y}, \tilde{\mathbf{x}})$ for $(y, \mathbf{x})$ is collected. The exact information on $(y, \mathbf{x})$ is then ascertained for a validation subsample drawn from the primary sample. Thus, two sets of data, the primary data $(\tilde{y}, \tilde{\mathbf{x}})$ on all study subjects, along with the validation data $(y, \mathbf{x}, \tilde{y}, \tilde{\mathbf{x}})$ on a validation subset, are available. The setup thus can be put into the context of 'incomplete data', in the sense that data $(y, \mathbf{x})$ is incompletely observed for all subjects, except for validation members.

Several approaches have been developed to estimate the regression parameters $\boldsymbol{\beta}$ from both the primary and the validation sample. The maximum likelihood (ML) method based on both samples of data requires a parametric specification for the full model regarding relationship among $(y, \mathbf{x}, \tilde{y}, \tilde{\mathbf{x}})$, and some special computational algorithms, e.g. 'EM' (Dempster, et al., 1977; Ibrahim, 1990), to implement the work. In general the model specification is very crucial for the ML method to yield the consistent estimation (Pepe, 1992), and the computation is cumbersome.

Alternative semi-parametric approaches have been proposed under several specific settings. When covariates are subject to missing values or non-differential measurement error, and the validation sample is a simple random subsample, Carroll and Wand (1991) and Pepe and Fleming (1991) proposed an estimated likelihood approach whereby the likelihood based on $(y, \tilde{\mathbf{x}})$ is estimated nonparametrically. Their method can only be applied to cases where the continuous components in $\tilde{\mathbf{x}}$ is of low dimension. In essentially the same context, Breslow and Cain (1988) focused mainly on the logistic regression but allowed the validation sample to be drawn via stratification on $(y, \tilde{\mathbf{x}})$, used a pseudolikelihood approach, see also Schill et al. (1995) for a related method. Reily and Pepe (1995) proposed a 'mean score' method dealing with the case where $(y, \tilde{\mathbf{x}})$ is categorical and the validation sample is drawn by stratification on $(y, \tilde{\mathbf{x}})$. Robins et al. (1994) provided a general class of estimating functions for 'incomplete' covariates data. Their estimators include all possible regular asymptotic linear estimators (Robins et al., proposition 8.1), and can be extended to cases where both y and $\mathbf{x}$ are incompletely observed. This approach requires that π, the probability that one subject is selected for validation, to be known up to a finite set of parameters. The efficiency of this approach depends on the specification of a function $\boldsymbol{\kappa} = \boldsymbol{\kappa}(\tilde{y}, \tilde{\mathbf{x}})$ of $(\tilde{y}, \tilde{\mathbf{x}})$. Although an 'optimal' $\boldsymbol{\kappa}$ does exist, in general it is difficult to implement. See Robins et al. (1994) and section 4 for details.

This work introduces a simple method to estimate regression parameters under the double sampling scheme. It is a general and unified approach allowing both the outcome and explanatory variables to be incompletely observed, so that it can be applied to a wide variety of study designs. The proposed estimating procedure utilizes available data from both the primary and validation samples. It is robust in the sense that the consistency of the estimation and inference does not require a correct specification of the association relating complete and incomplete data. The asymptotic variance of the estimator proposed can be readily derived and estimated.

In the following discussions, some assumptions and notations are kept in use.

- The observations $(\tilde{y}_i, \tilde{\mathbf{x}}_i)$, $i = 1, \ldots, N$, of the primary data are independently and identically distributed (iid) with finite first and second moments, where N denotes the sample size of the primary sample.

- The observations $(y_i, \mathbf{x}_i, \tilde{y}_i, \tilde{\mathbf{x}}_i)$, $i \in V$, of the validation data are iid with finite first and second moments, where V denotes the validation sample. Denote by n the sample size of the validation sample, and $\rho \equiv$plim n/N the validation fraction.

- The validation sample is a simple random subsample from the primary sample.

- The underlying regression model for the conditional mean of y given $\mathbf{x}$ is given by $E(y|\mathbf{x}) = g(\mathbf{x}; \boldsymbol{\beta})$, where $g(\cdot)$ is a known scalar function and $\boldsymbol{\beta}$ is a q-vector of unknown regression parameters.

Note that the restriction that the validation sample is a simple random subset of the main study excludes the applicability of our method to some more complex settings, such as the case where the sampling of validation sample can depend on $(\tilde{y}, \tilde{\mathbf{x}})$.

2 Estimation

Let $\hat{\boldsymbol{\beta}}$ solve the estimating equation of $\boldsymbol{\beta}$ based on data $(y, \mathbf{x})$ in the validation sample

$$0 = \sum_{i \in V} \mathbf{s}_i(\boldsymbol{\beta}) \equiv \sum_{i \in V} \mathbf{w}_i \{y_i - g(\mathbf{x}_i; \boldsymbol{\beta})\}, \tag{1}$$

where $\mathbf{w} = \mathbf{w}(\mathbf{x})$ is a $q \times 1$ function of $\mathbf{x}$. It is known that $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}^*$, the true value of $\boldsymbol{\beta}$, provided that some regularity conditions hold (Robins et al, 1994; Foutz, 1977). A special case for (1) is the score equation corresponding to the generalized linear model (GLM) (Nelder and Wedderburn, 1972).

Recall that $(\tilde{y}, \tilde{\mathbf{x}})$ is a proxy measure of $(y, \mathbf{x})$, hence the primary sample still contains partial information on $\boldsymbol{\beta}$. If a model $h(\tilde{\mathbf{x}}; \boldsymbol{\gamma})$ is specified to relate $\tilde{\mathbf{x}}$ to $\tilde{y}$, where $h(\cdot)$ is a known scalar function and $\boldsymbol{\gamma}$ is an r-vector of unknown regression parameters, and $\hat{\boldsymbol{\gamma}}$ is the estimate of $\boldsymbol{\gamma}$ obtained from the validation sample by solving the estimating equation

$$0 = \sum_{i \in V} \tilde{\mathbf{s}}_i(\boldsymbol{\gamma}) \equiv \sum_{i \in V} \tilde{\mathbf{w}}_i \{\tilde{y}_i - h(\tilde{\mathbf{x}}_i; \boldsymbol{\gamma})\}, \tag{2}$$

where $\tilde{\mathbf{w}} = \tilde{\mathbf{w}}(\tilde{\mathbf{x}})$ is an $r \times 1$ function of $\tilde{\mathbf{x}}$. Then under some regularity conditions, $\hat{\boldsymbol{\gamma}}$ is consistent for $\boldsymbol{\gamma}^*$ satisfying $0 = E[\tilde{\mathbf{w}}\{\tilde{y} - h(\tilde{\mathbf{x}}, \boldsymbol{\gamma}^*)\}]$, the expectation is taken with respect to $\boldsymbol{\gamma}^*$. The following well known lemma gives the asymptotic joint distribution of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$, which is immediate by the Taylor expansion and the central limit theorem. Henceforth a prime will denote matrix transposition, and the expectation is taken with respect to $(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$.

Lemma. *Suppose that the regularity conditions in Appendix hold. Then,* $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ *is consistent for* $(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ *, and* $n^{1/2}\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)', (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)'\}'$ *is asymptotically normal with mean zero and variance given by* $D^{-1}CD^{-1}$*, where* $D = diag(D_1, D_2)$ *with*

$$D_1 \equiv E\{\partial \mathbf{s}(\boldsymbol{\beta}^*)/\partial \boldsymbol{\beta}\}, \quad D_2 \equiv E\{\partial \tilde{\mathbf{s}}(\boldsymbol{\gamma}^*)/\partial \boldsymbol{\gamma}\},$$

$$\textit{and} \quad C \equiv E[\{\mathbf{s}(\boldsymbol{\beta}^*)', \tilde{\mathbf{s}}(\boldsymbol{\gamma}^*)'\}'\{\mathbf{s}(\boldsymbol{\beta}^*)', \tilde{\mathbf{s}}(\boldsymbol{\gamma}^*)'\}]. \diamondsuit$$

Denote submartrices of C by

$$C_{11} \equiv E\{\mathbf{s}(\boldsymbol{\beta}^*)\mathbf{s}'(\boldsymbol{\beta}^*)\}, \quad C_{12} \equiv E\{\mathbf{s}(\boldsymbol{\beta}^*)\tilde{\mathbf{s}}'(\boldsymbol{\gamma}^*)\}, \quad C_{22} \equiv E\{\tilde{\mathbf{s}}(\boldsymbol{\gamma}^*)\tilde{\mathbf{s}}'(\boldsymbol{\gamma}^*)\}.$$

By the distribution theory for multivariate normal variates, the conditional distribution of $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ given $n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)$ is asymptotic normal with mean

$$n^{1/2} D_1^{-1} C_{12} C_{22}^{-1} D_2 (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*).$$

Note that this conditional mean can be 'estimated' by substituting empirical estimates based on the validation sample for D_1, D_2, C_{12}, and C_{22}, and replacing $\boldsymbol{\gamma}^*$ with some estimate of $\boldsymbol{\gamma}^*$ based on the primary sample. To be specific, let

$$\hat{D}_1 = n^{-1} \sum_{i \in V} \partial \mathbf{s}_i(\hat{\boldsymbol{\beta}})/\partial \boldsymbol{\beta}, \quad \hat{D}_2 = n^{-1} \sum_{i \in V} \partial \tilde{\mathbf{s}}_i(\hat{\boldsymbol{\gamma}})/\partial \boldsymbol{\gamma},$$

$$\hat{C}_{12} = n^{-1} \sum_{i \in V} \mathbf{s}_i(\hat{\boldsymbol{\beta}})\tilde{\mathbf{s}}_i(\hat{\boldsymbol{\gamma}})', \quad \hat{C}_{22} = n^{-1} \sum_{i \in V} \tilde{\mathbf{s}}_i(\hat{\boldsymbol{\gamma}})\tilde{\mathbf{s}}_i(\hat{\boldsymbol{\gamma}})',$$

which is a set of consistent estimates for D_1, D_2, C_{12}, and C_{22} respectively. Also let $\bar{\boldsymbol{\gamma}}$ be the estimate of $\boldsymbol{\gamma}^*$ based on the primary sample solving

$$0 = \sum_{i=1}^{N} \tilde{\mathbf{s}}_i(\boldsymbol{\gamma}) = \sum_{i=1}^{N} \tilde{\mathbf{w}}_i \{\tilde{y}_i - h(\tilde{\mathbf{x}}_i; \boldsymbol{\gamma})\}. \tag{3}$$

By equating $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ with its estimated conditional mean

$$n^{1/2} \hat{D}_1^{-1} \hat{C}_{12} \hat{C}_{22}^{-1} \hat{D}_2 (\hat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}),$$

and solving for $\boldsymbol{\beta}^*$, we therefore obtain a reasonable estimate $\bar{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}^*$ given by

$$\bar{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \hat{D}_1^{-1} \hat{C}_{12} \hat{C}_{22}^{-1} \hat{D}_2 (\hat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}). \tag{4}$$

The estimator given by (4) is quite general without reference to any specific structure of the relationship among $(y, \mathbf{x}, \tilde{y}, \tilde{\mathbf{x}})$. In some special cases, model-based estimates for C_{12} and/or C_{22} may be available and appropriate, hence one can also use them instead to yield an alternative version of $\bar{\beta}$. See Section 5 for several examples.

To sum up, what we proposed amounts to the following procedure:

1. estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ from validation sample by solving (1) and (2) respectively, obtain quantities $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$, $\hat{D}_1$, $\hat{C}_{12}$, $\hat{C}_{22}$, and $\hat{D}_2$;

2. estimate $\boldsymbol{\gamma}$ from the primary sample by solving (3), obtain $\bar{\boldsymbol{\gamma}}$;

3. calculate $\bar{\boldsymbol{\beta}}$ using the expression (4).

The large sample properties of $\bar{\boldsymbol{\beta}}$ will be investigated in the next section.

3 Asymptotic Theory

Proposition. *Suppose the conditions in Appendix hold. Then $\bar{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}^*$, and $n^{1/2}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ is asymptotic normal with mean zero and variance Ω given by*

$$\Omega = D_1^{-1} C_{11} D_1^{-1} - (1 - \rho) D_1^{-1} C_{12} C_{22}^{-1} C_{12}' D_1^{-1}. \tag{5}$$

A sketch of the proof is given in Appendix. Note that the consistency of $\bar{\boldsymbol{\beta}}$ does not require the correct specification of the regression model $h(\tilde{\mathbf{x}}; \boldsymbol{\gamma})$. In other words, $h(\tilde{\mathbf{x}}; \boldsymbol{\gamma})$ serves as a 'working' model by which we can extract information on $\boldsymbol{\beta}$ contained in data $(\tilde{y}, \tilde{\mathbf{x}})$. The misspecification of $h(\tilde{\mathbf{x}}; \boldsymbol{\gamma})$, however, will increase the magnitude of C_{22} and hence that of the asymptotic variance for $\bar{\boldsymbol{\beta}}$. Fortunately, in the double sampling scheme, the primary sample is usually of large size, so that $h(\cdot; \boldsymbol{\gamma})$ may be adequately specified in the light of some conventional model-selection strategies, and therefore satisfactory efficiency can be anticipated for a suitably specified $h(\cdot; \boldsymbol{\gamma})$.

Note that in the context of surrogate covariates data, the conventional 'surrogate condition' or 'non-differential error' assumption imposed by most of the existing methodologies, but might be inappropriate in some cases, is not necessary for our proposal.

It can be seen that any estimator of the form

$$\hat{\boldsymbol{\beta}} - \Theta(\hat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}})$$

with Θ (or its limit) being any finite, constant $q \times r$ matrix is consistent for $\boldsymbol{\beta}^*$. Analogous to the well known result for the best linear predictor (e.g. Bickel and Doksum,1977, p. 40), it can be shown that the estimator $\bar{\boldsymbol{\beta}}$ we proposed attains highest asymptotic efficiency within this class of estimators.

Observe that the first term of (5) is definitely the asymptotic variance for $\hat{\boldsymbol{\beta}}$. Thus the expression (5) shows that, asymptotically, the proposed estimate $\bar{\boldsymbol{\beta}}$ is more efficient than $\hat{\boldsymbol{\beta}}$. The asymptotic relative efficiency of $\hat{\beta}_k$ to $\bar{\beta}_k$, where the subindex k denotes the kth component, is given by $\{1-(1-\rho)\lambda_k^2\}$ with

$$\lambda_k^2 = \frac{\operatorname{cov}\{s_k(\boldsymbol{\beta}^*), \tilde{\mathbf{s}}(\boldsymbol{\gamma}^*)'\}\operatorname{var}^{-1}\{\tilde{\mathbf{s}}(\boldsymbol{\gamma}^*)\}\operatorname{cov}\{s_k(\boldsymbol{\beta}^*), \tilde{\mathbf{s}}(\boldsymbol{\gamma}^*)\}}{\operatorname{var}\{s_k(\boldsymbol{\beta}^*)\}},$$

which is the squared multiple correlation between $s_k(\boldsymbol{\beta}^*)$ and $\tilde{\mathbf{s}}(\boldsymbol{\gamma}^*)$.

The asymptotic variance can be consistently estimated by replacing the population quantities in (5) with their empirical or, when available and appropriate, model based estimates based on the validation sample data. The latter version may not be positive-definite when the validation proportion n/N is rather small.

4 Relation to existing approaches

As a referee and the associate editor commented, since Robins et al. (1994) have proposed a general class of estimators that include all possible regular asymptotic linear estimators, the relationship between the estimator $\bar{\boldsymbol{\beta}}$, which is also asymptotic linear, and those of Robins et al. (1994) is thus of great interest. Recall that the estimators of Robins et al. (1994), making essentially the same assumptions as our proposal, are asymptotic linear with influence function of the form $D_1^{-1}R(\mathbf{w}, \boldsymbol{\kappa})$, where

$$R(\mathbf{w}, \boldsymbol{\kappa}) = \delta\mathbf{s}(\boldsymbol{\beta}^*)/\pi - (\delta-\pi)\boldsymbol{\kappa}/\pi, \tag{6}$$

with $\delta = 1$ if an observation belongs to the validation sample and 0 otherwise, $\pi =$pr$(\delta = 1|\tilde{y}, \tilde{x})$ and $\boldsymbol{\kappa} = \boldsymbol{\kappa}(\tilde{y}, \tilde{\mathbf{x}})$ being a $q \times 1$ function of $(\tilde{y}, \tilde{\mathbf{x}})$. Write $\mathbf{s} = \mathbf{s}(\boldsymbol{\beta}^*)$, $\tilde{\mathbf{s}} = \tilde{\mathbf{s}}(\boldsymbol{\gamma}^*)$. Observe that

$$N^{1/2}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = N^{-1/2} D_1^{-1} \sum_{i=1}^{N} \{\delta_i(\mathbf{s}_i - B\tilde{\mathbf{s}}_i)/\rho + B\tilde{\mathbf{s}}_i\} + o_p(1),$$

where $B = C_{12}C_{22}^{-1}$ =cov$(\mathbf{s}, \tilde{\mathbf{s}})\{\text{var}(\tilde{\mathbf{s}})\}^{-1}$. It thus can be seen that the estimator $\bar{\beta}$ corresponds to a member of the class of Robins et al. (1994) with π being a constant ρ and $\boldsymbol{\kappa}$ chosen to be $\boldsymbol{\kappa}^{reg} \equiv B\tilde{\mathbf{s}}$, which is the population least squares regression of $\mathbf{s}$ on $\tilde{\mathbf{s}}$. For given $\mathbf{w}$, Robins et al (1994) showed that the optimal choice of $\boldsymbol{\kappa}$ (in terms of asymptotic efficiency) is given by $\boldsymbol{\kappa}^{\mathbf{w}} \equiv E(\mathbf{s}|\tilde{y}, \tilde{\mathbf{x}})$, and even further efficiency can be reached with great effort by replacing $\mathbf{w}$ with $\mathbf{w}^{eff}$ given by equation (23) of Robins et al. (1994). However, for a given $\mathbf{w}$, the 'optimal' $\boldsymbol{\kappa}^{\mathbf{w}}$ is usually not feasible since it depends on the unknown probability law generating the data. A convenient solution suggested by Robins et al. (1994) is to specify a parametric model for $E(\mathbf{s}|\tilde{y}, \tilde{\mathbf{x}})$ and estimate it from validation data. In contrast, instead of specifying a possibly rather complicated function $\boldsymbol{\kappa}^{\mathbf{w}}$, our method alternatively specifies a parametric model relating $\tilde{\mathbf{x}}$ to $\tilde{y}$. Therefore the main difference between the proposed approach and that of Robins et al. (1994) is at the conceptual level.

Observe that the computation of D_1, C_{12}, C_{22}, and D_2 is straightforward by using the existing softwares. Also, as manifested in section 5 and 6, our method, albeit inefficient overall, does achieve satisfactory efficiency without great effort in a variety of different settings.

The method proposed in this paper utilizing a working parametric model to extract the partial information contained in the primary sample. A similar strategy has also been adopted by Lee and Sepanski (1995) and Chen and Chen (1998). Those methods differs from the present one mainly in that, taking the case $\tilde{y} \equiv y$ for illustration, the working parametric model is specified for $E\{g(\mathbf{x};\boldsymbol{\beta}^*)|\tilde{\mathbf{x}}\}$, not for $E(y|\tilde{\mathbf{x}})$. This allows the information on β^* to be directly extracted from the primary sample, rather than indirectly through the estimation of $\boldsymbol{\gamma}^*$. For example, one of the methods of Lee and Sepanski (1995) obtains consistent estimates for $\boldsymbol{\beta}^*$ respectively from the validation and the primary samples, and then pools the two estimates together by a linear combination of them with the 'minimum distance'. In certain cases this may improve the efficiency but requires making more assumptions about the relationship between $(y,\mathbf{x},\tilde{y},\tilde{\mathbf{x}})$ such as the 'surrogate condition' $E(y|\mathbf{x},\tilde{\mathbf{x}})=E(y|\mathbf{x})=g(\mathbf{x};\boldsymbol{\beta}^*)$.

5 Examples

In this section three examples are considered to elucidate the application of the proposed method.

Example 1. (Surrogate covariates in GLM) Consider the case where $\tilde{y}\equiv y$, and given $\mathbf{x}$, y follows a generalized linear model (GLM) with canonical link $g(\mathbf{x};\boldsymbol{\beta})$ and scale parameter ϕ. Assume that $E(y|\mathbf{x},\tilde{\mathbf{x}})=E(y|\mathbf{x})$, namely, $\tilde{\mathbf{x}}$ is a surrogate for $\mathbf{x}$ satisfying the usual 'surrogate' or 'non-differential error' condition, and $E(\tilde{\mathbf{x}}|\mathbf{x})=\Theta'\mathbf{x}$ with Θ a finite constant $q\times r$ matrix. Let $\hat{\beta}$ be the maximum likelihood estimate based on data $(y,\mathbf{x})$ in the validation sample, then we have $D_1=-E\{\mathbf{x}v(\boldsymbol{\beta}^*)\mathbf{x}'\}$, where $v(\boldsymbol{\beta}^*)=\phi^{-1}\mathrm{var}(y|\mathbf{x})$. If the model relating y and $\tilde{\mathbf{x}}$ is specified by a GLM with canonical link $h(\tilde{\mathbf{x}};\boldsymbol{\gamma})$ and scale parameter ψ, and $\hat{\gamma}$ is the solution of the corresponding score equation based on data $(y,\tilde{\mathbf{x}})$ in the validation sample, then by the surrogate condition,

$$\begin{aligned}C_{12} &= E[\mathbf{x}\{y-g(\mathbf{x};\boldsymbol{\beta}^*)\}\{y-h(\tilde{\mathbf{x}};\boldsymbol{\gamma}^*)\}\tilde{\mathbf{x}}'] \\ &= E\{\mathbf{x}\mathrm{var}(y|\mathbf{x})\tilde{\mathbf{x}}'\}=\phi E\{\mathbf{x}v(\boldsymbol{\beta}^*)\tilde{\mathbf{x}}'\}.\end{aligned}$$

So that $D_1^{-1}C_{12} = -\phi\Theta$. If the model $h(\tilde{\mathbf{x}};\boldsymbol{\gamma})$ is indeed appropriate, then $D_2 = -E\{\tilde{\mathbf{x}}\tilde{v}(\boldsymbol{\gamma}^*)\tilde{\mathbf{x}}'\}$ with $\tilde{v}(\boldsymbol{\gamma}^*) = \psi^{-1}\mathrm{var}(y|\tilde{\mathbf{x}})$, and $C_{22} = -\psi D_2$. So, in this case $\bar{\boldsymbol{\beta}}$ is simplified to

$$\bar{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \frac{\phi}{\psi}\Theta(\hat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}). \tag{7}$$

Here we assume that values of ϕ, ψ, and Θ are known. Otherwise they can be replaced with their respective estimates. The asymptotic variance of $\bar{\beta}$ is then given by

$$\phi D_1^{-1} - (1-\rho)\frac{\phi^2}{\psi}\Theta D_2^{-1}\Theta'.$$

For the even simpler scenario that $y = \beta x + \epsilon$, $\epsilon \sim N(0,1)$, $x \sim N(0,1)$, and $\tilde{x} = x + e$ with $e \sim N(0,\sigma^2)$, the asymptotic variance for the 'ideal' maximum likelihood estimate (MLE) $\hat{\beta}_{MLE}$ for β, assuming that $\mathrm{pr}(x|\tilde{x})$ is fully known, is

$$\{1 + \frac{(1-\rho)}{\rho}\frac{1}{(1+\sigma^2)\psi} + \frac{(1-\rho)}{\rho}\frac{2\sigma^4\beta^2}{(1+\sigma^2)^2\psi^2}\}^{-1},$$

with $\psi = 1+\beta^2\sigma^2/(1+\sigma^2)$. Figure 1 displays the asymptotic relative efficiency (ARE) of $\bar{\beta}$ to $\hat{\beta}_{MLE}$ at $\beta = 0$ under different values of ρ and σ^2. It can be seen that, except for the cases where ρ is rather small and σ^2 is rather large, the loss of efficiency of $\bar{\boldsymbol{\beta}}$ is within a reasonable range. The ARE of $\bar{\beta}$ will be even higher when parameters in $\mathrm{pr}(x|\tilde{x})$ are also estimated in calculating the MLE.

(• *Figure 1 is about here.*)

Example 2. (Missing covariates in GLM) We still consider the preceding GLM setting, but assume that $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')'$ and $\tilde{\mathbf{x}} \equiv \mathbf{x}_1$. That is, $\tilde{\mathbf{x}}$ consists of the first r elements of $\mathbf{x}$. This is the setting of missing covariates data that the last $(q-r)$ elements of $\mathbf{x}$ are missing for the whole study units except for those in the validation sample. Note that this corresponds to a special case for Example 1 with $\Theta = \{I_r, O\}'$, where I_r is the r-dimensional identity matrix, and O is the $r \times (q-r)$ matrix whose elements are all zeros. When y is binary, $\tilde{\mathbf{x}}$ is discrete and $g(\cdot;\boldsymbol{\beta})$ is a logistic link, in such missing covariate setting $\bar{\boldsymbol{\beta}}$ is identical to the pseudolikelihood estimator $\hat{\boldsymbol{\beta}}_B$ proposed by Breslow and Cain (1988). Since in this case $\hat{\boldsymbol{\beta}}_B$ is also the MLE (Schill et al., 1995), $\bar{\boldsymbol{\beta}}$ is fully efficient in this case.

Example 3. (Seemingly unrelated regression) Let $\tilde{\mathbf{x}} \equiv \mathbf{x}$. Suppose that $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$, $E(\epsilon_i|\mathbf{x}_i) = 0$, for $i \in V$, and $\tilde{y}_i = \mathbf{x}_i'\boldsymbol{\gamma} + e_i$, $E(e_i|\mathbf{x}_i) = 0$, for $i = 1, \ldots, N$. Let $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$ be respectively the ordinary least squares (OLS) estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ based on the validation sample, and $\bar{\boldsymbol{\gamma}}$ be the OLS estimate of $\boldsymbol{\gamma}$ based on the primary sample. The well-known procedure for the 'seemingly unrelated regression' (Schmidt, 1977; Conniffe, 1996) provides the estimator $\tilde{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ given by

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \tilde{\theta}(\hat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}), \quad \text{with} \;\; \tilde{\theta} = \sum_{i \in V}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})(\tilde{y}_i - \mathbf{x}_i'\hat{\boldsymbol{\gamma}}) / \sum_{i \in V}(\tilde{y}_i - \mathbf{x}_i'\hat{\boldsymbol{\gamma}})^2.$$

In contrast, in this case the proposed estimator $\bar{\boldsymbol{\beta}}$ can be written as

$$\bar{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \hat{\Theta}(\hat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}),$$

which differs from $\tilde{\boldsymbol{\beta}}$ in that $\tilde{\theta}$ is replaced by

$$\hat{\Theta} = \{\sum_{i \in V} \mathbf{x}_i(\tilde{y}_i - \mathbf{x}_i'\hat{\boldsymbol{\gamma}})^2\mathbf{x}_i'\}^{-1}\{\sum_{i \in V} \mathbf{x}_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})(\tilde{y}_i - \mathbf{x}_i'\hat{\boldsymbol{\gamma}})\mathbf{x}_i'\}$$

Note that when both $\text{cov}(\epsilon, e|\mathbf{x})$ and $\text{var}(e|\mathbf{x})$ have constant values independent of $\mathbf{x}$, $\bar{\boldsymbol{\beta}}$ will be asymptotically equivalent to $\tilde{\boldsymbol{\beta}}$. It has been shown by Conniffe (1985) that, when $\{(\epsilon_i, e_i)\}$ are iid bivariate normal with a constant covariance matrix, $\tilde{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$. Hence in this case, $\bar{\boldsymbol{\beta}}$ will be fully efficient. In general cases, neither $\tilde{\boldsymbol{\beta}}$ nor $\bar{\boldsymbol{\beta}}$ will be the MLE, whereas the latter will be more efficient than the former as discussed in Section 3.

6 Simulation Studies

Two simulation studies are demonstrated in this section. The first study was to investigate the finite sample properties of the proposed estimator. In each experiment 500 primary and validation data sets were generated. Each primary sample consisted of $N = 500$ observations, from which a validation subsample was randomly sampled with the sampling probability ρ chosen to be either 0.1 or 0.2. Based on the 500 replications, the Monte Carlo estimates for the sampling mean and variance of $\bar{\boldsymbol{\beta}}$, and for the mean of the estimated variance were examined to assess the adequacy of $\bar{\boldsymbol{\beta}}$ as well as its variance estimator $\bar{\Omega}$. The coverage probability was given by the percentage of 500 samples for which the 90% confidence interval for $\boldsymbol{\beta}^*$ based on $\bar{\boldsymbol{\beta}}$ included $\boldsymbol{\beta}^*$. For comparison, the relative MSE (mean squared error) of $\bar{\boldsymbol{\beta}}$ relative to the estimator $\hat{\boldsymbol{\beta}}$ based only on validation data, was also computed empirically based on the 500 simulated samples.

Two simple cases were considered in this study. The first was related with the errors-in-variables problem. The explanatory variable x was generated from a lognormal distribution with $\ln(x)$ having mean -0.5 and variance 1. The response y was a binary 0-1 variable generated by the logistic regression model

$$g(\mathbf{x}) = \mathrm{pr}(y = 1|\mathbf{x}) = 1/\{1 + \exp(-\beta_0^* - \beta_1^* x)\}, \tag{8}$$

where $\beta_0^* = 0$, $\beta_1^* = 1$. The proxy variable $\tilde{x}$ for x was a lognormal variable such that given x, $\ln(\tilde{x})$ had mean $\ln(x)$ and variance 0.5. The primary sample consisted of observations on $(y, \tilde{x})$, and the validation sample consisted of observations on $(y, x, \tilde{x})$. To obtain $\bar{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}$ was computed from the score equation for logistic regression model (8) based on data (y_i, x_i), $i \in V$. Also, the score equation for the 'working' logistic regression $h(\tilde{\mathbf{x}}; \boldsymbol{\gamma}) = 1/\{1 + \exp(-\gamma_0 - \gamma_1 \tilde{x})\}$ was used to obtain $\hat{\boldsymbol{\gamma}}$ and $\bar{\boldsymbol{\gamma}}$, each was based on data $(y, \tilde{x})$ respectively from the validation and the primary sample. The second case was in the context of the surrogate outcome problem. The response variable y of interest was assumed to be

$$y = g(\mathbf{x};\boldsymbol{\beta}^*) + \epsilon, \quad g(\mathbf{x};\boldsymbol{\beta}^*) = \beta_0^* + \beta_1^* x, \tag{9}$$

with $\beta_0^* = 0$, $\beta_1^* = 1$, x and ϵ being independent standard normal variates. A binary 0-1 variable $\tilde{y}$ was a surrogate for y such that $\mathrm{pr}(\tilde{y} = 1|y) = 1/\{1+\exp(-3y)\}$. Primary data $(\tilde{y}, x)$ and validation data (y, x) were then created according to the above models. $\hat{\boldsymbol{\beta}}$ was obtained by the OLS estimate for the model (9) , and $\hat{\boldsymbol{\gamma}}$ and $\bar{\boldsymbol{\gamma}}$ were obtained from the score equations for the 'working' logistic model $h(\mathbf{x};\boldsymbol{\gamma})$ $=\mathrm{pr}(\tilde{y} = 1|\mathbf{x}) = 1/\{1 + \exp(-\gamma_0 - \gamma_1 x)\}$ based respectively on the validation and the primary data.

The results are shown in Table 1. It can be seen that the proposed estimator $\bar{\boldsymbol{\beta}}$ is nearly unbiased in all cases we considered. The variance estimator exhibits mild degrees of negative bias in this small sample case. The coverage probability is near its nominal value in each case. Compared with $\hat{\boldsymbol{\beta}}$, the proposed estimator clearly performs better in terms of mean squared error.

(• *Table 1 is about here.*)

Simulation study 2 was to compare $\bar{\boldsymbol{\beta}}$ with some of existing estimators. In this study, a linear model $y = \beta_0 + \beta x + \epsilon$ was considered, with $\epsilon \sim N(0, 1)$ and $\beta_0^* = -1$. Also $\tilde{y} \equiv y$, $x = \tilde{x} + e$ and $\tilde{x} \sim N(0, 1)$. Two cases were examined: (i) $e \sim N(0, 1)$ and (ii) $\ln(e) \sim N(-0.5, 0.5)$. The primary sample size was 1000. Simulation size was 500. In each replication, in addition to $\bar{\boldsymbol{\beta}}$, we also calculated the estimator $\hat{\boldsymbol{\beta}}_L$ proposed in Lee and Sepanski (1995), and two variants of estimators, $\hat{\boldsymbol{\beta}}_{R0}$ and $\hat{\boldsymbol{\beta}}_R$, proposed in Robins et al. (1994). The estimator $\hat{\boldsymbol{\beta}}_{R0}$ corresponds to the one with $\boldsymbol{\kappa} = 0$ in (6), and $\hat{\boldsymbol{\beta}}_R$ corresponds to the one with $\boldsymbol{\kappa} = \boldsymbol{\kappa}^{\mathbf{w}}$ in (6). Here $\boldsymbol{\kappa}^{\mathbf{w}}$ was obtained by the quadratic model

$$E(\mathbf{s}|y, \tilde{x}) = E\{x(y - \boldsymbol{\beta}^* x)|y, \tilde{x}\} = \mathbf{z}'\boldsymbol{\lambda},$$

with $\mathbf{z} = (1, \tilde{x}, y, \tilde{x}^2, y^2, \tilde{x}y)$ and $\boldsymbol{\lambda}$ a set of unknown parameters. The estimation of $\boldsymbol{\lambda}$ was based on the validation data with $\boldsymbol{\beta}^*$ replaced by $\hat{\boldsymbol{\beta}}_{R0}$. Also the estimated value $\hat{\pi}$ for π was used when calculating $\hat{\boldsymbol{\beta}}_{R0}$ and $\hat{\boldsymbol{\beta}}_R$ to improve the efficiency as suggested in Robins et al. (1994). Note that $\boldsymbol{\kappa}^{\mathbf{w}}$ was correctly specified in case (i), but was misspecified in case (ii). The calculation of $\hat{\boldsymbol{\beta}}_L$ was based on the chosen basis $\mathbf{z}_V \equiv (1, \tilde{x})'$, and the 'minimum distance' estimate pooling the two estimates based respectively on the validation and the primary data was reported, see Lee and Sepanski (1995) for details. The results are shown in Table 2 for various choices of β^* and ρ.

As exhibited in Table 2, in case (i) where $\boldsymbol{\kappa}^{\mathbf{w}}$ is correctly specified, the relative efficiency of $\bar{\boldsymbol{\beta}}$ to $\hat{\boldsymbol{\beta}}_R$ is rather satisfactory (from 70% to 100%). When $\boldsymbol{\kappa}^{\mathbf{w}}$ is misspecified, case (ii), $\bar{\boldsymbol{\beta}}$ can be more efficient than $\hat{\boldsymbol{\beta}}_R$ (when $\beta^* = 2$ and $\rho = 0.1$). In all the cases we considered, $\bar{\boldsymbol{\beta}}$ achieves higher efficiency than $\hat{\boldsymbol{\beta}}_{R0}$ and $\hat{\boldsymbol{\beta}}_L$. Note that the variances of $\hat{\boldsymbol{\beta}}_{R0}$ and $\hat{\boldsymbol{\beta}}_R$ can be decreased by a regression technique (Rotnitzky, Robins, and Scharfstein, 1998), which is logically equivalent to our proposal to obtain more efficiency than the validation-data-only method.

(• *Table 2 is about here.*)

7 Concluding Remarks

We have proposed a simple and general methodology to estimation of the regression parameter $\boldsymbol{\beta}$ relating $\mathbf{x}$ to y when data on $(y, \mathbf{x})$ is only available for a simple random subsample, while proxy or surrogate information $(\tilde{y}, \tilde{\mathbf{x}})$ is available for the whole primary sample.

The advantages of the proposed method are: (i) it is a general and unified approach that can deal simultaneously with surrogate outcome as well as incomplete covariates data in a wide class of regression models; (ii) it uses a parametric model $h(\cdot;\boldsymbol{\gamma})$ relating $\tilde{\mathbf{x}}$ to $\tilde{y}$ to extract information contained in the primary sample, whereas the consistency of the estimation and inference does not rely on the specification of the model; (iii) reasonable estimation efficiency can be achieved by specifying an appropriate model for $h(\cdot;\boldsymbol{\gamma})$ using existing model-selection tools; and (iv) it is conceptually and computationally convenient, hence can be easily implement in daily data analysis.

Appendix

Regularity conditions

Let $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$, $\mathbf{u}(\boldsymbol{\alpha}) = \{\mathbf{s}(\boldsymbol{\beta}), \tilde{\mathbf{s}}(\boldsymbol{\gamma})\}$. The regularity conditions required in proving the proposition are listed below.

(R1) $\boldsymbol{\alpha}^* = (\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ lies in the interior of a compact parameter space.

(R2) $E\{\mathbf{u}(\boldsymbol{\alpha})\} \neq 0$ if $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$.

(R3) There exists a neighborhood $\mathbf{A}$ of $\boldsymbol{\alpha}^*$ such that $E\{\sup_{\boldsymbol{\alpha}\in\mathbf{A}} \|\mathbf{u}(\boldsymbol{\alpha})\|\}$, $E\{\sup_{\boldsymbol{\alpha}\in\mathbf{A}} \|\partial\mathbf{u}(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}\|\}$, $E\{\sup_{\boldsymbol{\alpha}\in\mathbf{A}} \|\mathbf{u}(\boldsymbol{\alpha})\mathbf{u}'(\boldsymbol{\alpha})\|\}$ are all finite, where $\|M\| = (\Sigma_{ij} m_{ij}^2)^{1/2}$ for any matrix M with elements m_{ij}.

(R4) $E\{\mathbf{u}(\boldsymbol{\alpha}^*)\mathbf{u}'(\boldsymbol{\alpha}^*)\}$ is finite and positive definite.

Proof of the Proposition

Following the arguments in Foutz (1977), from (R1), (R2), (R3), together with $E\{\mathbf{u}(\boldsymbol{\alpha}^*)\} = 0$, the consistency of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ and $\bar{\boldsymbol{\gamma}}$ are obtained by the uniform law of large numbers and the inverse function theory. Let $\Delta = D_1^{-1} C_{12} C_{22}^{-1} D_2$, and $\hat{\Delta} = \hat{D}_1^{-1} \hat{C}_{12} \hat{C}_{22}^{-1} \hat{D}_2$ as defined in Section 2. By (R3) and (R4), with probability going to one, $\hat{\Delta}$ converges uniformly to the finite matrix Δ. The consistency of $\bar{\boldsymbol{\beta}}$

thus follows.

The asymptotic normality of $\bar{\boldsymbol{\beta}}$ is obtained by the asymptotic normality of $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$, and $\bar{\boldsymbol{\gamma}}$, and by Slutsky's Theorem. Let $\mathcal{V}(\cdot)$ and $\mathcal{C}(\cdot,\cdot)$ denote asymptotic variance and covariance respectively. The asymptotic variance of $\bar{\boldsymbol{\beta}}$ is readily obtained by noting that

$$\mathcal{V}(n^{1/2}\hat{\boldsymbol{\beta}}) = D_1^{-1}C_{11}D_1^{-1},\ \mathcal{V}(n^{1/2}\bar{\boldsymbol{\gamma}}) = \rho\mathcal{V}(n^{1/2}\hat{\boldsymbol{\gamma}}) = \rho D_2^{-1}C_{22}D_2^{-1},$$

and

$$\mathcal{C}(n^{1/2}\hat{\boldsymbol{\beta}}, n^{1/2}\bar{\boldsymbol{\gamma}}) = \rho\mathcal{C}(n^{1/2}\hat{\boldsymbol{\beta}}, n^{1/2}\hat{\boldsymbol{\gamma}}) = \rho D_1^{-1}C_{12}D_2^{-1}.$$

Acknowledgements

We are grateful to the referees and the Associate Editor for their insightful and helpful comments.

References

Bickel, P. J., and Doksum, K. A. (1977). *Mathematical Statistics.* San Francisco: Holden-Day.

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

Carroll, R. J. and Wand, M. P. (1991). Semiparametric estimation in logistic measurement error models. *J. R. Statist. Soc.* **B, 53**, 573-587.

Chen, Y.-H. and Chen, H. (1999). Incomplete covariates data in generalized linear models. *J. Statist. Plann. Inference*, **79**, 247-258.

Conniffe, D. (1996). A comment on the use of auxiliary or surrogate outcome data.

J. Statist. Plann. Inference, **55**, 353-359.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via EM algorithm (with discussion). *J. R. Statist. Soc.* **B, 39**, 1-38.

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

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *J. Am. Statist. Assoc.*, **85**, 765-769.

Lee, L.-F. and Sepanski, J. H. (1995). Estimation of linear and nonlinear errors-in-variables models using validation data. *J. Am. Statist. Assoc.*, **90**, 130-140.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc.* **A, 135**, 370-384.

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

Pepe, M. S. and Fleming, T. R. (1991). A non-parametric method for dealing with mismeasured covariate data. *J. Am. Statist. Assoc.*, **86**, 108-113.

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

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

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. Am. Statist. Assoc.*, **93**, 1321-1339.

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

Schmidt, P. (1977). Estimation of seemingly unrelated regressions with unequal numbers of observations. *J. Econom.*, **5**, 365-377.

Illustration

Figure 1. Asymptotic relative efficiency (ARE) of $\bar{\beta}$ to MLE at $\beta = 0$ under the model $y = \beta x + \epsilon$, $\epsilon \sim N(0,1)$, $x \sim N(0,1)$, and $\tilde{x} = x + e$ with $e \sim N(0, \sigma^2)$. The validation fraction is denoted by ρ.

Table 1. *Results of simulation study 1. Case (i): logistic regression with mismeasured covariate. Case (ii): linear regression with binary surrogate outcome. In each case* $N = 500$.

	Case (i). $\beta^* = (0,1)$				Case (ii). $\beta^* = (0,1)$			
	$\rho = 0.1$		$\rho = 0.2$		$\rho = 0.1$		$\rho = 0.2$	
	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_0$	$\bar{\beta}_1$	$\bar{\beta}_0$	$\bar{\beta}_1$
Mean^a	-0.047	1.108	-0.024	1.080	-0.006	1.016	-0.006	1.005
var $^b \times 10^2$	13.18	36.68	7.34	17.31	1.36	1.88	0.67	0.94
$\widehat{\text{var}}^c \times 10^2$	12.30	34.61	6.20	15.90	1.36	1.71	0.71	0.84
CP^d	90%	88%	90%	88%	89%	87%	90%	88%
RMSE^e	48%	69%	48%	69%	70%	85%	74%	88%

a: simulation mean of of $\bar{\beta}$
b: simulation variance of $\bar{\beta}$
c: simulation mean of the estimated asymptotic variance of $\bar{\beta}$
d: simulation coverage probability of 90% confidence interval for β^* based on $\bar{\beta}$
e: simulation mean squared error of $\bar{\beta}$ relative to that of $\hat{\beta}$

Table 2. *Results of simulation study 2, where* $\tilde{y} = y = \beta_0 + \beta x + \epsilon$, $\epsilon \sim N(0,1)$, $\beta_0^* = -1$, $x = \tilde{x} + e$, *and* $\tilde{x} \sim N(0,1)$. $N = 1000$. *The upper entry:* $\rho = 0.1$, *the lower entry:* $\rho = 0.4$.

	$\beta^* = 0$		$\beta^* = 1$		$\beta^* = 2$	
	bias$\times 10^3$	var$\times 10^3$	bias$\times 10^3$	var$\times 10^3$	bias$\times 10^3$	var$\times 10^3$
			Case (i). $e \sim N(0,1)$			
$\bar{\beta}$	3.81	3.14	3.90	4.23	4.14	4.91
	1.32	0.91	1.72	1.08	1.66	1.17
$\hat{\beta}_{R0}$†	3.30	5.51	2.72	5.49	2.82	5.48
	1.31	1.23	1.54	1.22	1.50	1.22
$\hat{\beta}_R$‡	4.40	3.41	12.78	2.90	13.25	4.02
	1.50	0.92	2.87	0.77	2.85	0.95
$\hat{\beta}_L$¶	3.28	5.48	3.08	5.32	3.34	5.30
	1.29	1.23	1.32	1.22	1.30	1.22
			Case (ii). $\ln(e) \sim N(-0.5, 0.5)$			
$\bar{\beta}$	2.59	2.43	4.07	3.68	4.51	5.45
	0.35	0.97	1.08	1.14	1.13	1.38
$\hat{\beta}_{R0}$	2.16	7.40	1.75	7.44	1.51	7.48
	0.37	1.69	0.51	1.68	0.49	1.68
$\hat{\beta}_R$	2.70	2.56	10.38	3.22	19.12	6.19
	0.70	0.98	2.73	1.03	4.41	1.32
$\hat{\beta}_L$	2.15	7.37	1.79	7.25	1.79	7.16
	0.37	1.69	0.39	1.67	0.37	1.67

†: Robins et al. (1994) with $\boldsymbol{\kappa} = 0$; ‡: Robins et al. (1994) with $\boldsymbol{\kappa} = \boldsymbol{\kappa}^{\mathbf{w}}$, which is based on the estimated quadratic model for $E(s|y,\tilde{x})$, with $s = x(y - \beta_0^* - \beta^* x)$; ¶: Lee and Sepanski (1995).