Regression calibration using response variables in linear model

Y. H. Steve Huang

October 31, 2001

Department of Mathematics, Tamkang University, Tamsui, Taipei, Taiwan.

Abstract

In errors-in-variables models, the model may be unidentifiable and the naive estimator is usually not satisfactory. When a validation subsample is available, Regression calibration is an easy way to improve the estimation. Regression calibration replaces a missing covariate by a estimate which is more accurate than the surrogate. One might expect a better parameter estimation if one has a better covariate estimator. This is possible by using response variables together with surrogates to estimate or to predict the missing values. However the introducing of response variables will raise bias in the resulting estimating function. In this article, we use response variables to calibrate the true covariate and provide an estimation method for the regression parameter in linear model. While the errors in variables are small, we show that regression calibration using response variable is better than the conventional regression calibration and the complete case analysis for large sample . A small simulation study for examining their performances in finite sample is also provided.
Key words: Regression calibration; Missing Data; Validation sample; Errors in variables; Response variable.

1 Introduction

In a applied problem, the relationship between response variables and covariates is an important thing one wants to know. In some situations, the true covariates are hard to measure or expensive to collect and only part of them are available. If a surrogate W of the true covariates are collected for all subjects. Then we can partition the whole sample as a union of two subsample, one with true covariates (the validation data set) and the other one with surrogate only (the primary data set). If one ignored data containing any missing value, this is called the complete data analysis. Under some conditions the process causing missing can be ignored (Rubin, 1976) and a complete case analysis is a valid one. But it wastes information that is contained in the primary data set. On the other hand, one can fit the primary data as a measurement error model, and analyses with the help of "extra information" like variance of measurement error or ratio of errors' variance, which are available from the validation data. An easy way to combine the two data set is the estimation using regression calibration (R.C.). It is a kind of imputation, it estimate the "missing covariate" first, and replace missing values by their estimates, then proceed the analysis as if there are no missing values. R. C. is simple and potentially applicable to any regression model, provided the approximation is sufficiently accurate (Carroll et al. 1995, Chp. 3.).

Errors-in-variables model with some validation data can be view as a missing value problem with the features of surrogacy, and only part of true covariates are missing. Hence R.C. can be treat as a method that deal with missing covariates in a special kind of missing values problem. There are many literatures for missing value problems. Beale and Little (1975) maximum the likelihood when data are from multivariate normal distribution. Pepe and Fleming (1991) estimate the likelihood function when some independent varibles or surrogates are discrete. Rubin (1987) proposed multiple imputation instead of imputing a single data for nonresponses. Little (1992) provide a widely review on regression with missing covariates. An analytic technique for inference about multiple imputation can be found on Schafer and Schenker (2000).

R.C. methods are simpler both in concept and in implementation. R.C. means to "calibrate" the missing variables with available data, and then replace the missing values by them. Estimation can be done by using this "complete data" with cautions about the variance estimates. Conventional R.C on regression model usually need assumption about conditional expectation in addition to the regression function. Carroll and Stefanski (1990) provide a quasi-likelihood estimation base on small measurement error approximation, they using Taylor series to expand the expectation of the covariates condition on surrogate but not on response. Lee and Sepanski (1995) project the regression function to the space of integrable and linear function of surrogate as a wide-sense conditional expectation. A important reason for these method being better than the naive estimations is that they using a better covariates estimates than the surrogate. Following this idea, we can have a more accurate estimates of missing values by using the response variables in the calibration stage especially when they are highly correlated. We also find that we can doing this in linear model without any additional distribution or conditional expectation except the original regression function itself.

The idea of using response variables to estimate missing values and then proceed to regression analysis can be found as early as Afifi (1967), but they estimate each missing value by a degenerate random variable, i.e. a constant, and their model is a simple linear model without measurement errors and covariates are assumed to be normal distributed. Here we consider the simple errors-in-variables model, with some moments conditions on error terms and true covariates. The estimation method can extend to multiple regression easily. Though the estimator we propose is an asymptotically linear estimator as described in Robin et al. (1994), and an optimal estimator among this class is provided in the same article. But computation and some expectation assumption is needed in order to compute the optimal one. The method we propose has the advantages of being very easy to calculate and no distribution assumptions is needed. The variances' estimator are also simple base on small error assumption. Furthermore, we show that improving the conventional R.C. by introducing response variables is feasible. This can motivate further studies on nonlinear models.

In section 2 we motivate and propose the estimating function. Some asymptotic results is derived there. Section 3 is theoretical comparisons of propose one with conventional R.

C. and complete case analysis. Section 4 provides some simulation studies. Section 5 will discuss the possibility to extend the R.C. with response variables to multiple regression model, and the appendix presents the technical details of the theorems.

2 Regression calibration using response variables

We assume three random variables W, X and Y have the relation

$$Y = \alpha + \beta X + \sigma_\epsilon \epsilon, \quad W = X + \sigma_\delta \delta \qquad (2.1).$$

where X, ϵ and δ are independently distributed with finite fourth moments. Both ϵ and δ are random variables of mean 0 and variance one. W is a surrogate of X, which means that condition on X, W and Y are independent. No normality is assumed here, but $E(\epsilon^3) = 0$ is assumed 0 for technical reason.

We divide the observations into two sets depends on whether there are missingness or not, one comprises observations like (Y, X, W) is denoted by V (validation data set), and one comprises (Y, W) only is denoted by P (primary data set). That is, (Y_i, W_i, X_i) is observed if $i \in V$, and (Y_i, W_i) is observed if $i \in P$. Denote the size of set P by N_p and let N_v denote the size of set V. The total sample size is $N(= N_p + N_v)$. The probability of being missing is constant for each individual, i.e. the covariates are missing completely at random (MCAR).

Since the validation data is a random subsample. The analysis using only complete cases is valid, but may lose some efficiency because no information is extracted from primary data.

The conventional R.C. consists of two steps. At first, using validation data to estimate the regression function of X on W, which may be accomplished by solving

$$\sum_{i \in V} (X_i - \gamma_0 - \gamma_1 W_i) \begin{pmatrix} 1 \\ W_i \end{pmatrix} = 0,$$

for γ_0 and γ_1. Let $\hat{\gamma}_0, \hat{\gamma}_1$ be the solutions of the equation, and $m_i = m(W_i, \hat{\gamma}_0, \hat{\gamma}_1) = \hat{\gamma}_0 + \hat{\gamma}_1 W_i$. Secondly, replace unobserved X_i by m_i and proceed as if there is no measurement error. That is to solve

$$\sum_{i \in V} (Y_i - \beta_0 - \beta_1 X_i) \begin{pmatrix} 1 \\ X_i \end{pmatrix} + \sum_{i \in V} (Y_i - \beta_0 - \beta_1 m_i) \begin{pmatrix} 1 \\ m_i \end{pmatrix} = 0$$

for estimates of β_0 and β_1, if least square estimation is adopted.

We notice that in the first step of estimating "missing X_i" no information of response Y_i is used. This may causes some deficiencies in estimating X especially when they are highly correlated. It is sensible to expect a "better" estimation of β_0 and β_1 if we use a "better" estimator of "missing X" in the second step. Such conjecture is partly verified in this paper.

2.1 The Estimating Procedure

The most efficient way to predict the unobserved X_i is the conditional expectation $E(X_i \mid Y_i, W_i)$, but it can't be computed without assumptions about joint distribution of $(X_i, \epsilon_i, \delta_i)$. Even the distribution is known it may be difficult to derive any explicit form. In stead of $E(X_i \mid Y_i, W_i)$, we use the best linear predictor H_i of X_i, where $H_i = a + bW_i + cY_i$ and (a, b, c) makes $E(X_i - H_i)^2$ attains its minimum. Thus (a, b, c) can be estimated by minimizing

$$\sum_{i \in V} (X_i - H_i)^2,$$

with respectively to a, b and c. It is easy to know that (a, b, c) satisfy the equation

$$E \begin{bmatrix} 1 & W & Y \\ W & W^2 & WY \\ Y & WY & Y^2 \end{bmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = E \begin{pmatrix} X \\ WX \\ XY \end{pmatrix}. \tag{2.2}$$

If we replace any unobserved X_i by H_i like conventional regression calibration. We have the estimating equation

$$\sum_{i \in V} (Y_i - \beta_0 - \beta_1 X_i) \begin{pmatrix} 1 \\ X_i \end{pmatrix} + \sum_{i \in P} (Y_i - \beta_0 - \beta_1 H_i) \begin{pmatrix} 1 \\ H_i \end{pmatrix} = 0. \tag{2.3}$$

Unfortunately, this is a biased estimating equation because the regressor H_i is correlated to the error $Y_i - \beta_0 - \beta_1 H_i$, both of them contain the error ϵ_i and inconsistent estimator will result. To correct the bias, we see that

$$E[\sum_{i \in V} (Y_i - \beta_0 - \beta_1 X_i) \begin{pmatrix} 1 \\ X_i \end{pmatrix} + \sum_{i \in P} (Y_i - \beta_0 - \beta_1 H_i) \begin{pmatrix} 1 \\ H_i \end{pmatrix}]$$

$$= N_P E(Y_i - \beta_0 - \beta_1 H_i) \begin{pmatrix} 1 \\ H_i \end{pmatrix} + N_p E\epsilon_i \begin{pmatrix} 1 \\ H_i \end{pmatrix} = N_P \beta_1 E(X_i - H_i) \begin{pmatrix} 1 \\ H_i \end{pmatrix} + N_p E\epsilon_i \begin{pmatrix} 1 \\ H_i \end{pmatrix}.$$

The last 2nd term is $\mathbf{0}$ by the property of H and the last term is $N_p\binom{0}{c\sigma_\epsilon^2}$. We subtract this from the equation and derive a $\mathbf{0}$-unbiased estimating function

$$\sum_{i\in V}(Y_i-\beta_0-\beta_1X_i)\binom{1}{X_i}+\sum_{i\in P}[(Y_i-\beta_0-\beta_1H_i)\binom{1}{H_i}-\binom{0}{c\sigma_\epsilon^2}], \tag{2.4}$$

and this is the estimation function we will use. We propose an estimation method for β_0 and β_1 as follows:

1. Using the validation data set to estimate the coefficients of X regress on (W,Y). Obtain the best linear predictor H_i for each missing X_i.
2. Using the validation data to compute the ordinary least estimates of β_0, β_1 and σ_ϵ^2 as initial estimates.
3. Using current estimates of σ_ϵ^2 to solve (2.4) with respect to β_0 and β_1, derive new estimates of β_0 and β_1.
4. With new estimates of β_0 and β_1 we can recompute the residuals and update the estimates of σ_ϵ^2.
5. With new estimates of σ_ϵ^2, go back to step 3 until we find that the estimates of β_0 and β_1 are converge.

This procedure is equivalent to solve

$$\begin{gathered}\sum_{i\in V}[(Y_i-\beta_0-\beta_1X_i)^2-\sigma_\epsilon^2]=0\\ \sum_{i\in V}(X_i-a-bW_i-cY_i)\begin{pmatrix}1\\ W_i\\ Y_i\end{pmatrix}=\mathbf{0}, \\ \sum_{i\in V}(Y_i-\beta_0-\beta_1X_i)\binom{1}{X_i}+\sum_{i\in P}[(Y_i-\beta_0-\beta_1H_i)\binom{1}{H_i}-\binom{0}{c\sigma_\epsilon^2}]=\mathbf{0}\end{gathered} \tag{2.5}$$

where $H_i=a+bW_i+cY_i$.

Transform the original parameters $\boldsymbol{\theta}=(\mu,\sigma_X^2,\sigma_\epsilon^2,\sigma_\delta^2,\beta_0,\beta_1)$ into $\boldsymbol{\eta}=(\sigma_\epsilon^2,a,b,c,\beta_0,\beta_1)$. and denote

$$A_i=(Y_i-\beta_0-\beta_1X_i)^2-\sigma_\epsilon^2,\quad B_i=(X_i-a-bW_i-cY_i)\begin{pmatrix}1\\ W_i\\ Y_i\end{pmatrix},$$

$$C_i = (Y_i - \beta_0 - \beta_1 X_i)\begin{pmatrix} 1 \\ X_i \end{pmatrix}, \quad D_i = (Y_i - \beta_0 - \beta_1 H_i)\begin{pmatrix} 1 \\ H_i \end{pmatrix} - \begin{pmatrix} 0 \\ c\sigma_\epsilon^2 \end{pmatrix}$$

and $\rho = N_v/N_p$. The asymptotic variance-covariance condition on ρ can written as

$$NE(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})' \approx \begin{pmatrix} \rho E\frac{\partial A}{\partial \boldsymbol{\eta}} \\ \rho E\frac{\partial B}{\partial \boldsymbol{\eta}} \\ \rho E\frac{\partial C}{\partial \boldsymbol{\eta}} + (1-\rho)E\frac{\partial D}{\partial \boldsymbol{\eta}} \end{pmatrix}^{-1}$$

$$\begin{pmatrix} \rho EAA' & \rho EAB' & \rho EAC' \\ \rho EB'A & \rho EBB' & \rho EBC' \\ \rho EC'A & \rho EC'B & \rho ECC' + (1-\rho)EDD' \end{pmatrix} \begin{pmatrix} \rho E\frac{\partial A}{\partial \boldsymbol{\eta}} \\ \rho E\frac{\partial B}{\partial \boldsymbol{\eta}} \\ \rho E\frac{\partial C}{\partial \boldsymbol{\eta}} + (1-\rho)E\frac{\partial D}{\partial \boldsymbol{\eta}} \end{pmatrix}^{'-1}. \tag{2.6}$$

This is the sandwich estimator of covariance matrix (Carroll et.al. (1995), Appendix A). The matrix in (2.6) is too complicate to give any interesting result and is not suitable for comparison purpose. Hence we consider the small error approximation.

2.2 Small error approximation

Here we assume that the ratio of two errors' variances $\frac{\sigma_\delta^2}{\sigma_\epsilon^2}$, denote by k, remain fixed when σ_ϵ^2 approaches 0. The matrix in (2.6) after computations becomes to $H^{-1}MH^{-1'} + \mathbf{O}(\sigma_\epsilon^3)$, where

$$H = -\begin{pmatrix} \rho & 0 & 0 & 0 & 0 & 0 \\ 0 & \rho & 0 & \rho\beta_0 & 0 & 0 \\ 0 & 0 & \rho\sigma_X^2 + \sigma_\delta^2 & \rho\beta_1\sigma_X^2 & 0 & 0 \\ 0 & \rho\beta_0 & \rho\beta_1\sigma_X^2 & \rho(\beta_0^2 + \beta_1^2\sigma_X^2 + \sigma_\epsilon^2) & 0 & 0 \\ 0 & (1-\rho)\beta_1 & 0 & (1-\rho)\beta_0\beta_1 & 1 & 0 \\ (1-\rho)c & 0 & (1-\rho)\beta_1\sigma_X^2 & (1-\rho)\beta_1^2\sigma_X^2 & 0 & \sigma_X^2 \end{pmatrix},$$

$$M = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \rho\frac{k}{1+\beta_1^2 k} & 0 & \rho\beta_0\frac{k}{1+\beta_1^2 k} & -\rho c & 0 \\ 0 & 0 & \rho\sigma_X^2\frac{k}{1+\beta_1^2 k} & \rho\beta_1\sigma_X^2\frac{k}{1+\beta_1^2 k} & 0 & -\rho c\sigma_X^2 \\ 0 & \frac{\rho\beta_0 k}{1+\beta_1^2 k} & \frac{\rho\beta_1\sigma_X^2 k}{1+\beta_1^2 k} & \rho(\beta_0^2 + \beta_1^2\sigma_X^2)\frac{k}{1+\beta_1^2 k} & -\rho\beta_0 c & -\rho\beta_1 c\sigma_X^2 \\ 0 & -\rho c & 0 & -\rho\beta_0 c & -\rho - \frac{1-\rho}{1+\beta_1^2 k} & 0 \\ 0 & 0 & -\rho c\sigma_X^2 & -\rho c\beta_1\sigma_X^2 & 0 & -\rho\sigma_X^2 - \frac{1-\rho}{1+\beta_1^2 k}\sigma_X^2 \end{pmatrix}\sigma_\epsilon^2,$$

and $\mathbf{O}(\sigma_\epsilon^3)$ denote a matrix with every components being $O(\sigma_\epsilon^3)$. Hence $H^{-1}MH^{-1'}$ is an approximation of the asymptotic covariance matrix of the solutions in (2.6) when errors are "small". The lower-right part of $H^{-1}MH^{-1'}$ corresponds to the covariance matrix of the regression parameters' estimators is given in the following theorem.

Theorem 1. Under the assumption of model (2.1), the solutions defined by equations (2.5) are consistent. Moreover, if $\sigma_\delta^2/\sigma_\epsilon^2$ (denote by k) remain fixed when $\sigma_\epsilon^2 \to 0$, then these estimators have asymptotic matrix covariance $H^{-1}MH^{-1'} + \mathbf{O}(\sigma_\epsilon^3)$, where H and M are defined above. In particular, the asymptotic covariance matrix (standardized by multiplying N) of $\hat{\beta}_0$ and $\hat{\beta}_1$ after simplification is

$$\begin{pmatrix} 1 & -\mu \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{\rho+\beta_1^2 k}{\rho+\rho\beta_1^2 k} & 0 \\ 0 & \frac{1}{\sigma_X^2}\frac{\rho+\beta_1^2 k}{\rho+\rho\beta_1^2 k} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\mu & 1 \end{pmatrix} \sigma_\epsilon^2 + \mathbf{O}(\sigma_\epsilon^3),$$

where μ denote the mean of X.

To compare with conventional R. C., which not uses response variables in the calibration stage. A similar handling was apply to find the approximate asymptotic covariance matrix for estimator defined as solutions of

$$\sum_{i \in V}(X_i - \gamma_0 - \gamma_1 W_i) \begin{pmatrix} 1 \\ W_i \end{pmatrix} = \mathbf{0}, \tag{2.7}$$

$$\sum_{i \in V}(Y_i - \beta_0 - \beta_1 X_i)\begin{pmatrix} 1 \\ X_i \end{pmatrix} + \sum_{i \in P}(Y_i - \beta_0 - \beta_1(\gamma_0 + \gamma_1 W_i))\begin{pmatrix} 1 \\ (\gamma_0 + \gamma_1 W_i) \end{pmatrix} = \mathbf{0}.$$

Theorem.2 With the same conditions stated in theorem 1. The estimators derived by conventional regression calibration, which are solutions of (2.7) have asymptotic covariance (standardized by multiplying N)

$$\begin{pmatrix} 1 & -\mu \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{(1-\rho)\beta_1^2 k+\rho}{\rho} & 0 \\ 0 & \frac{(1-\rho)\beta_1^2 k+\rho}{\rho\sigma_X^2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\mu & 1 \end{pmatrix} \sigma_\epsilon^2 + \mathbf{O}(\sigma_\epsilon^3).$$

Now we are ready to compare the R.C. with response variables, without response and the complete case analysis.

3 Efficiency comparison

Here we compare three methods in estimating β_0 and β_1. The complete case analysis, the conventional R.C and the R.C using response variables. Let $\hat{\boldsymbol{\beta}}_c$, $\hat{\boldsymbol{\beta}}_{rc}$ and $\hat{\boldsymbol{\beta}}$ denote the vector of the estimators from complete case analysis, conventional R.C and R.C. using response variables, respectively. It is well known that

$$N_v E(\hat{\boldsymbol{\beta}}_c - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_c - \boldsymbol{\beta})' = \begin{pmatrix} 1 & -\mu \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\sigma_X^2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\mu & 1 \end{pmatrix} \sigma_\epsilon^2 + \mathbf{o}(\sigma_\epsilon^2).$$

Also from theorem 2 and theorem 1, we have

$$NE(\hat{\boldsymbol{\beta}}_{rc}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{rc}-\boldsymbol{\beta})' = \begin{pmatrix} 1 & -\mu \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{(1-\rho)\beta_1^2k+\rho}{\rho} & 0 \\ 0 & \frac{(1-\rho)\beta_1^2k+\rho}{\rho\sigma_X^2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\mu & 1 \end{pmatrix} \sigma_\epsilon^2 + \mathbf{O}(\sigma_\epsilon^3).$$

$$NE(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})' = \begin{pmatrix} 1 & -\mu \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{\rho+\beta_1^2k}{\rho+\rho\beta_1^2k} & 0 \\ 0 & \frac{1}{\sigma_X^2}\frac{\rho+\beta_1^2k}{\rho+\rho\beta_1^2k} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -\mu & 1 \end{pmatrix} \sigma_\epsilon^2 + \mathbf{O}(\sigma_\epsilon^3),$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ is the true parameter. They are standardized by multiplying different number N_v and N. In order to compare them, we set $\mu = 0$ and ignore terms that smaller than $O(\sigma_\epsilon^2)$ and standardize every covariance matrix by N_v. We found that the three matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\sigma_X^2} \end{pmatrix} \sigma_\epsilon^2, \quad [(1-\rho)\beta_1^2k+\rho] \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\sigma_X^2} \end{pmatrix} \sigma_\epsilon^2, \quad \frac{\rho+\beta_1^2k}{1+\rho\beta_1^2k} \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\sigma_X^2} \end{pmatrix} \sigma_\epsilon^2 \qquad (2.8)$$

stands for the major parts of asymptotic covariance matrices of $\hat{\boldsymbol{\beta}}_c$, $\hat{\boldsymbol{\beta}}_{rc}$ and $\hat{\boldsymbol{\beta}}$, respectively. Under the assumption of small σ_ϵ and when N is sufficient large, we observed the followings

1. $\hat{\boldsymbol{\beta}}_{rc}$ is better than $\hat{\boldsymbol{\beta}}_c$ only if $[(1-\rho)\beta_1^2k+\rho] < 1$, which can happen when $\mid \beta_1 \mid$ or k is small.

2. $\hat{\boldsymbol{\beta}}$ is better than $\hat{\boldsymbol{\beta}}_c$ since $\frac{\rho+\beta_1^2k}{1+\beta_1^2k} < 1$.

3. $\hat{\boldsymbol{\beta}}$ is better than $\hat{\boldsymbol{\beta}}_{rc}$ since $\frac{\rho+\beta_1^2k}{1+\beta_1^2k} - [(1-\rho)\beta_1^2k+\rho] = -\frac{(1-\rho)\beta_1^4k^2}{1+\beta_1^2k} < 0$.

We also note that when ρ close to 1, the three estimations does not differ much, and the advantage of R.C. using response variable over complete case will be more obvious if ρ, β_1^2 or k are getting small. R.C. using response variable can be much better than without response variable if ρ is small and β_1 and k are not close to 0. In conclusion, the R.C. using response variables is preferable in any cases when σ_ϵ is small.

4 Simulation studies

We conduct some simulations to see the performances of each estimation method in finite sample. In the simulation, X_i are either draw form a standard normal or from a standardized uniform distribution. Both σ_δ^2 and σ_ϵ^2 are set to 0.25, thus the ratio of error's variance is 1. Sample size is 300 and the mean square errors in the following tables are the average values from 1,000 replications. The values of formulas in (2.8) evaluate at estimated parameters is also recorded. These values is helpful in judging the small error approximations of variance matrices are good or not.

Table 1. $X \sim N(0,1)$

$\rho = 0.2,\quad \beta_1 = 0.5$			$\rho = 0.5,\quad \beta_1 = 0.5$		
	MSE*10^3	Estimates of variances $*10^3$		MSE*10^3	Estimates of variances $*10^3$
$\hat{\beta}_c$	(3.95, 4.52)*	4.13°(0.76), 4.16$^\diamond$(0.88)	$\hat{\beta}_c$	(1.65,1.50)	1.66(0.19), 1.68(0.28)
$\hat{\beta}_{rc}$	(1.54,1.86)	1.68(0.67), 1.66(0.63)	$\hat{\beta}_{rc}$	(1.10,1.05)	1.04(0.13), 1.05(0.15)
$\hat{\beta}$	(1.42,1.86)	1.43(0.48), 1.42(0.47)	$\hat{\beta}$	(1.05,1.01)	0.992(0.11), 1.000(0.14)
$\rho = 0.2,\quad \beta_1 = 1.5$			$\rho = 0.5,\quad \beta_1 = 1.5$		
	MSE*10^3	Estimates of variances $*10^3$		MSE*10^3	Estimates of variances $*10^3$
$\hat{\beta}_c$	(3.95,4.52)	4.13(0.76), 4.16(0.88)	$\hat{\beta}_c$	(1.65,1.50)	1.66(0.19), 1.68(0.28)
$\hat{\beta}_{rc}$	(6.95,8.07)	8.41(5.70), 8.26(5.44)	$\hat{\beta}_{rc}$	(2.56,2.39)	2.70(0.83), 2.68(0.70)
$\hat{\beta}$	(2.92,3.25)	3.02(5.81), 3.03(6.31)	$\hat{\beta}$	(1.41,1.28)	1.37(0.20), 1.38(0.22)

Table 2. $X \sim Uni(-0.5, 0.5) * 3.6461$

$\rho = 0.2,\quad \beta_1 = 0.5$			$\rho = 0.5,\quad \beta_1 = 0.5$		
	MSE*10^3	Estimates of variances $*10^3$		MSE*10^3	Estimates of variances $*10^3$
$\hat{\beta}_c$	(3.86, 4.10)	4.10(0.77), 4.11(0.81)	$\hat{\beta}_c$	(1.64,1.57)	1.66(0.19), 1.67(0.23)
$\hat{\beta}_{rc}$	(1.61,1.73)	1.68(0.50), 1.67(0.48)	$\hat{\beta}_{rc}$	(1.03,1.03)	1.04(0.13), 1.04(0.13)
$\hat{\beta}$	(1.50,1.63)	1.47(0.34), 1.46(0.33)	$\hat{\beta}$	(1.00,0.955)	0.997(0.11), 0.999(0.12)
$\rho = 0.2,\quad \beta_1 = 1.5$			$\rho = 0.5,\quad \beta_1 = 1.5$		
	MSE*10^3	Estimates of variances $*10^3$		MSE*10^3	Estimates of variances $*10^3$
$\hat{\beta}_c$	(3.86,4.10)	4.10(0.77), 4.11(0.81)	$\hat{\beta}_c$	(1.64,1.57)	1.66(0.19), 1.67(0.23)
$\hat{\beta}_{rc}$	(7.20,7.26)	8.54(4.13), 8.44(3.95)	$\hat{\beta}_{rc}$	(2.42,2.23)	2.72(0.66), 2.71(0.59)
$\hat{\beta}$	(3.09,2.91)	3.01(2.83), 3.00(3.03)	$\hat{\beta}$	(1.39,1.22)	1.39(0.16), 1.40(0.17)

- Remark: The "$\star$" represents the mean square errors in estimation of β_0 and β_1. "$\circ$" and "$\diamond$" represent the average values of variances' estimates from (2.8) of estimators of β_0 and β_1, respectively. The standard deviation of these variances estimates are recorded in the parentheses.

From the results, we first note that there is no much difference in Table 1 and Table 2. This reflects that normality is not necessary for the proposed method. We also note that when $\beta_1 = 0.5$, $\hat{\boldsymbol{\beta}}_{rc}$ is better than $\hat{\boldsymbol{\beta}}_c$. But when $\beta_1 = 1.5$, the advantage is on $\hat{\boldsymbol{\beta}}$. However in the simulated cases, $\hat{\boldsymbol{\beta}}$ seems to be the best as expected. The values in column $\circ$ and column $\diamond$ should close be to the MSE (column $\star$) if the approximation of variances under small error are "good". From the values of standard deviation, we found that these estimates of variances of $\hat{\boldsymbol{\beta}}_{rc}$ and $\hat{\boldsymbol{\beta}}$ are more variable when ρ is small and β_1 is not close to zero, but works fine in others. In summary, the R.C. with response variables should be prefer according to the simulation result and theorems in estimation parameters. But the variance estimates under the small error assumptions should be use with cautions when the size of validation data is relative small and the regression parameter seems far away from 0.

5 Generalization to multiple regression

The idea of using response variables to calibrate missing covariate can extend to multiple regression under some restrictions. Consider the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_m X_{im} + \sigma_\epsilon \epsilon_i,$$

and only a portion of data has missing X_{i1}'s completely at random and surrogates W_{i1} of X_{i1} are available for all individuals. Then one can consider the relationship

$$Y_i - \beta_2 X_{i2} - \cdots - \beta_m X_{im} = \beta_0 + \beta_1 X_{i1} + \sigma_\epsilon \epsilon_i$$

as model (2.1) after $\beta_2, \cdots, \beta_m$ are estimated from the validation data set. However the covariance matrices are no longer the same as theorems due to the variability of estimators of $\beta_2, \cdots, \beta_m$. It will be more complicate and will be pursued in the future.

6 Appendix

Before we give the proofs of theorems, two lemmas and some notations can help us in reading the proofs. The lemma 3 take care of the parameter $E(X)$, and lemma 4 lists the cumbersome computation results. We denote H_i as $X_i - \Delta X_i - \eta$, where $\Delta = (1 - b - c\beta_1)$ and $\eta = -(b\sigma_\delta\delta_i + c\sigma_\epsilon\epsilon_i)$. This can be done because $H_i = a + bW_i + cY_i$ is a linear combination of constant, X_i and errors ϵ_i and δ_i. Denote $e_i = X_i - H_i = \Delta + \eta$, the difference of true covaritae and its predictor. Also the index "i" will be suppressed for simplicity in notations hereafter.

Lemma 3. Let the asymptotic covariance matrix of estimators of β_0 and β_1 by solving (2.5) be V_0 when evaluate at the situation $\mu = 0$, and be V when μ is not zero, where $\mu = E(X)$, then

$$V = \begin{pmatrix} 1 & -\mu \\ 0 & 1 \end{pmatrix} V_0 \begin{pmatrix} 1 & 0 \\ -\mu & 1 \end{pmatrix}.$$

Proof. Note that the equations in (2.5) can be rewritten as

$$\sum_{i\in V}[(Y_i - (\beta_0 + \beta_1\mu) - \beta_1 X_i^*)^2 - \sigma_\epsilon^2] = 0$$

$$\sum_{i\in V}(X_i^* - a^* - bW_i^* - cY_i)\begin{pmatrix} 1 \\ W^* \\ Y \end{pmatrix} = \mathbf{0},$$

$$\sum_{i\in V}(Y_i - (\beta_0 + \beta_1\mu) - \beta_1 X_i^*)(\begin{matrix} 1 \\ X_i^* \end{matrix}) + \sum_{i\in P}[(Y_i - (\beta_0 + \beta_1\mu) - \beta_1 H_i^*)(\begin{matrix} 1 \\ H_i^* \end{matrix}) - (\begin{matrix} 0 \\ c\sigma_\epsilon^2 \end{matrix})] = \mathbf{0}$$

where $X_i^* = X_i - \mu, W_i^* = W_i - \mu, H_i^* = a^* + bW_i^* + cY_i$, and $a^* = a - (1 - b)\mu$. The quantities $(b_0 + b_1\mu, b_1)$ has asymptotic variance covariance V_0 according to the assumption. Since $\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} 1 & -\mu \\ 0 & 1 \end{pmatrix}\begin{pmatrix} b_0 + b_1\mu \\ b_1 \end{pmatrix}$, the conclusion follows easily.

Lemma 4. Assume that X, ϵ and δ all have finite fourth moments, $\mu = 0$ and $\sigma_\delta^2/\sigma_\epsilon^2 = k$ remain fixed when σ_ϵ^2 is varying, then we have

$$a = \frac{-\beta_0\beta_1 k}{1 + \beta_1^2 k} + O(\sigma_\epsilon^2), b = \frac{1}{1 + \beta_1^2 k} + O(\sigma_\epsilon^2), c = \frac{\beta_1 k}{1 + \beta_1^2 k} + O(\sigma_\epsilon^2)$$

(terms involve e)

$$\begin{aligned}
E(e) &= E(eh) = E(eW) = E(eY) = 0,\ E(eX) = \Delta\sigma_X^2 = O(\sigma_\epsilon^2), \\
E(e^2) &= \frac{k}{1 + \beta_1^2 k}\sigma_\epsilon^2 + O(\sigma_\epsilon^3), E(e^2 W) = O(\sigma_\epsilon^3), \\
E(e^2 Y) &= \frac{\beta_0 k}{1 + \beta_1^2 k}\sigma_\epsilon^2 + O(\sigma_\epsilon^3),\ E(e^2 X) = O(\sigma_\epsilon^3),\ \ E(e^2 H) = O(\sigma_\epsilon^3), \\
E(e^2 W^2) &= \frac{k\sigma_X^2}{1 + \beta_1^2 k}\sigma_\epsilon^2 + O(\sigma_\epsilon^3),\ E(e^2 Y^2) = (\beta_0^2 + \beta_1^2\sigma_X^2)\frac{k}{1 + \beta_1^2 k}\sigma_\epsilon^2 + O(\sigma_\epsilon^3),
\end{aligned}$$

(terms involve ϵ)

$$E(\epsilon H) = c\sigma_\epsilon, E(\epsilon^2 H) = 0, E(\epsilon^2 H^2) = \sigma_X^2\sigma_\epsilon^2 + O(\sigma_\epsilon^3),$$

(terms involve e and ϵ)

$$\begin{aligned}
E(e\epsilon) &= -c\sigma_\epsilon,\ E(e\epsilon H) = 0,\ E(e\epsilon H^2) = -c\sigma_\epsilon\sigma_X^2 + O(\sigma_\epsilon^3), \\
E(e\epsilon W) &= 0,\ E(e\epsilon X) = 0,\ E(e\epsilon Y) = -\beta_0 c\sigma_\epsilon, \\
E(e\epsilon XW) &= -c\sigma_\epsilon\sigma_X^2,\ E(e\epsilon XY) = -c\sigma_\epsilon\beta_1\sigma_X^2 + O(\sigma_\epsilon^3) \\
E(e^2WY) &= \frac{\beta_1^2 k\sigma_X^2}{1+\beta_1^2 k}\sigma_\epsilon^2 + O(\sigma_\epsilon^3),\ E(e^2H^2) = \frac{k\sigma_X^2}{1+\beta_1^2 k}\sigma_\epsilon^2.
\end{aligned}$$

Proof of Lemma 4. We give the proofs of equations about a, b, c and $E(e^2WY)$ and omit the others, because their proofs are either similar to or simpler than the one of $E(e^2WY)$.

From (2.2),

$$\begin{aligned}
\begin{pmatrix} a \\ b \\ c \end{pmatrix} &= (E\begin{bmatrix} 1 & W & Y \\ W & W^2 & WY \\ Y & WY & Y^2 \end{bmatrix})^{-1} E\begin{pmatrix} X \\ WX \\ XY \end{pmatrix} \\
&= \begin{bmatrix} 1 & 0 & \beta_0 \\ 0 & \sigma_X^2 + \sigma_\delta^2 & \beta_1\sigma_X^2 \\ \beta_0 & \beta_1\sigma_X^2 & \beta_0^2 + \beta_1^2\sigma_X^2 + \sigma_\epsilon^2 \end{bmatrix}^{-1} \begin{pmatrix} 0 \\ \sigma_X^2 \\ \beta_1\sigma_X^2 \end{pmatrix} \\
&= \frac{\sigma_X^2}{\sigma_X^2(\sigma_\epsilon^2 + \beta_1^2\sigma_\delta^2) + \sigma_\delta^2\sigma_\epsilon^2} \begin{pmatrix} -\beta_0\beta_1\sigma_\delta^2 \\ \sigma_\epsilon^2 \\ \beta_1\sigma_\delta^2 \end{pmatrix} = \frac{\sigma_X^2}{\sigma_X^2(1+\beta_1^2 k) + k\sigma_\epsilon^2} \begin{pmatrix} -\beta_0\beta_1 k \\ 1 \\ \beta_1 k \end{pmatrix} \\
&= \frac{1}{1+\beta_1^2 k} \begin{pmatrix} -\beta_0\beta_1 k \\ 1 \\ \beta_1 k \end{pmatrix} + \mathbf{O}(\sigma_\epsilon^2).
\end{aligned}$$

Recall that $e = X - H = \Delta x + \eta$ where $\eta = -b\sigma_\delta\delta - c\sigma_\epsilon\epsilon = O_p(\sigma_\epsilon)$, and $\Delta = 1 - b - c_1\beta_1$, which is $k\sigma_\epsilon^2/[\sigma_X^2(1+\beta_1^2 k) + k\sigma_\epsilon^2] = O(\sigma_\epsilon^2)$ after computation, it follows that

$$\begin{aligned}
E(e^2WY) &= E[(\Delta^2X^2 + \eta^2 + 2\Delta X\eta)(X + \sigma_\delta\delta)(\beta_0 + \beta_1 X + \sigma_\epsilon\epsilon)] \\
&= E[(\eta^2(X)(\beta_0 + \beta_1 X)] + O(\sigma_\epsilon^3) = E(\eta^2\beta_0 X + \beta_1\eta^2X^2) + O(\sigma_\epsilon^3) \\
&= E(\beta_1\eta^2X^2) + O(\sigma_\epsilon^3) = \beta_1\sigma_\eta^2\sigma_X^2 + O(\sigma_\epsilon^3) = \beta_1(b^2\sigma_\delta^2 + c^2\sigma_\epsilon^2)\sigma_X^2 + O(\sigma_\epsilon^3) \\
&= \beta_1[(\frac{1}{1+\beta_1^2 k} + O(\sigma_\epsilon^2))^2 k\sigma_\epsilon^2 + (\frac{\beta_1 k}{1+\beta_1^2 k} + O(\sigma_\epsilon^2))^2\sigma_\epsilon^2)\sigma_X^2 + O(\sigma_\epsilon^3) \\
&= \frac{\beta_1^2 k\sigma_X^2}{1+\beta_1^2 k}\sigma_\epsilon^2 + O(\sigma_\epsilon^3).
\end{aligned}$$

Proof of Theorem 1. We assume that $E(X) = 0$ to derive the formula of asymptotic covariance matrix, then we apply lemma 3 for the case $E(X) \neq 0$.

Since the expectations of the equations in the right-hand side of (2.5) is 0 if and only if evaluated at true parameter $\boldsymbol{\eta}$. The consistency follows.

The formula (2.6) is a kind of "sandwich" estimator described in Appendix A of Carroll et. al. (1995). To compute the matrices in (2.6), we find that it is easy to show that

$$\begin{aligned} E\frac{\partial A}{\partial \boldsymbol{\eta}} &= (\frac{\partial A}{\partial \sigma_\epsilon^2}, \frac{\partial A}{\partial a}, \frac{\partial A}{\partial b}, \frac{\partial A}{\partial c}, \frac{\partial A}{\partial \beta_0}, \frac{\partial A}{\partial \beta_1}) = (-1, 0, 0, 0, 0, 0), \\ E\frac{\partial B}{\partial \boldsymbol{\eta}} &= E\begin{pmatrix} 0 & -1 & -W & -Y & 0 & 0 \\ 0 & -W & -W^2 & -WY & 0 & 0 \\ 0 & -Y & -WY & -Y^2 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & -1 & 0 & -\beta_0 & 0 & 0 \\ 0 & 0 & -(\sigma_X^2 + \sigma_\delta^2) & -\beta_1\sigma_X^2 & 0 & 0 \\ 0 & -\beta_0 & -\beta_1\sigma_X^2 & -(\beta_0^2 + \beta_1^2\sigma_X^2 + \sigma_\epsilon^2) & 0 & 0 \end{pmatrix}, \end{aligned}$$

and

$$E\frac{\partial C}{\partial \boldsymbol{\eta}} = E\begin{pmatrix} 0 & 0 & 0 & 0 & -1 & -X \\ 0 & 0 & 0 & 0 & -X & -X^2 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\sigma_X^2 \end{pmatrix}.$$

The term

$$E\frac{\partial D}{\partial \boldsymbol{\eta}} = E\partial\begin{pmatrix} Y - \beta_0 - \beta_1 H \\ (Y - \beta_0 - \beta_1 H)H - c\sigma_\epsilon^2 \end{pmatrix}/\partial\boldsymbol{\eta} = E\partial\begin{pmatrix} Y - \beta_0 - \beta_1 H \\ (\beta_1 e + \sigma_\epsilon \epsilon)H - c\sigma_\epsilon^2 \end{pmatrix}/\partial\boldsymbol{\eta},$$

comprises

$$\begin{aligned} \frac{\partial D}{\partial \sigma_\epsilon^2} &= \begin{pmatrix} 0 \\ -c \end{pmatrix}, \quad \frac{\partial D}{\partial a} = \begin{pmatrix} -\beta_1\frac{\partial H}{\partial a} \\ (\beta_1 e + \sigma_\epsilon\epsilon)\frac{\partial H}{\partial a} - \beta_1 H\frac{\partial H}{\partial a} \end{pmatrix} \\ \frac{\partial D}{\partial b} &= \begin{pmatrix} -\beta_1\frac{\partial H}{\partial b} \\ (\beta_1 e + \sigma_\epsilon\epsilon)\frac{\partial H}{\partial b} - \beta_1 H\frac{\partial H}{\partial b} \end{pmatrix}, \frac{\partial D}{\partial c} = \begin{pmatrix} -\beta_1\frac{\partial H}{\partial c} \\ (\beta_1 e + \sigma_\epsilon\epsilon)\frac{\partial H}{\partial c} - \beta_1 H\frac{\partial H}{\partial c} - \sigma_\epsilon^2 \end{pmatrix}, \end{aligned}$$

$$\frac{\partial D}{\partial \beta_0} = \begin{pmatrix} -1 \\ -H \end{pmatrix} \text{ and } \frac{\partial D}{\partial \beta_1} = \begin{pmatrix} -H \\ -H^2 \end{pmatrix}.$$

Replace $\frac{\partial H}{\partial a}, \frac{\partial H}{\partial b}, \frac{\partial H}{\partial c}$ by $1, W, Y$ and H by $(1-\Delta)X - \eta$, respectively. And using the results of lemma 4 we have

$$E\frac{\partial D}{\partial \boldsymbol{\eta}} = \begin{pmatrix} 0 & -\beta_1 & 0 & -\beta_0\beta_1 & -1 & 0 \\ -c & 0 & -\beta_1\sigma_X^2 & -\beta_1^2\sigma_X^2 & 0 & -\sigma_X^2 \end{pmatrix} + \mathbf{O}(\sigma_\epsilon^2).$$

In conclusion, we have

$$\begin{pmatrix} \rho E\frac{\partial A}{\partial \boldsymbol{\eta}} \\ \rho E\frac{\partial B}{\partial \boldsymbol{\eta}} \\ \rho E\frac{\partial C}{\partial \boldsymbol{\eta}} + (1-\rho)E\frac{\partial D}{\partial \boldsymbol{\eta}} \end{pmatrix}$$

$$= - \begin{pmatrix} \rho & 0 & 0 & 0 & 0 & 0 \\ 0 & \rho & 0 & \rho\beta_0 & 0 & 0 \\ 0 & 0 & \rho(\sigma_X^2+\sigma_\delta^2) & \rho\beta_1\sigma_X^2 & 0 & 0 \\ 0 & \rho\beta_0 & \rho\beta_1\sigma_X^2 & \rho(\beta_0^2+\beta_1^2\sigma_X^2+\sigma_\epsilon^2) & 0 & 0 \\ 0 & (1-\rho)\beta_1 & 0 & (1-\rho)\beta_0\beta_1 & 1 & 0 \\ (1-\rho)c & 0 & (1-\rho)\beta_1\sigma_X^2 & (1-\rho)\beta_1^2\sigma_X^2 & 0 & \sigma_X^2 \end{pmatrix} + \mathbf{O}(\sigma_\epsilon^2). \quad (A.1)$$

On the other hand, concerning the quadratic term in (2.6), we see that

$$\begin{aligned}
E(AA^{'}) &= E(\epsilon^2\sigma_\epsilon^2-\sigma_\epsilon^2)^2 = \sigma_\epsilon^4 Var(\epsilon^2) = O(\sigma_\epsilon^3), \\
E(BB^{'}) &= E(e^2\begin{pmatrix} 1 & W & Y \\ W & W^2 & WY \\ Y & WY & Y^2 \end{pmatrix}) = \begin{pmatrix} 1 & 0 & \beta_0 \\ 0 & \sigma_X^2 & \beta_1\sigma_X^2 \\ \beta_0 & \beta_1\sigma_X^2 & (\beta_0^2+\beta_1^2\sigma_X^2) \end{pmatrix} \frac{k\sigma_\epsilon^2}{1+\beta_1^2 k} + O(\sigma_\epsilon^3), \\
E(CC^{'}) &= E(\sigma_\epsilon^2\epsilon^2\begin{pmatrix} 1 & X \\ X & X^2 \end{pmatrix}) = \begin{pmatrix} 1 & 0 \\ 0 & \sigma_X^2 \end{pmatrix}\sigma_\epsilon^2, \\
E(DD^{'}) &= \begin{pmatrix} \beta_1 e+\sigma_\epsilon\epsilon \\ \beta_1 eh+\sigma_\epsilon\epsilon h - c\sigma_\epsilon^2 \end{pmatrix} (\beta_1 e+\sigma_\epsilon\epsilon, \beta_1 eh+\sigma_\epsilon\epsilon h - c\sigma_\epsilon^2) \\
&= E\begin{pmatrix} \beta_1^2e^2+\sigma_\epsilon^2\epsilon^2+2\beta_1 e\sigma_\epsilon\epsilon & \beta_1^2e^2h+\beta_1 e\sigma_\epsilon\epsilon h-\beta_1 ec\sigma_\epsilon^2+\sigma_\epsilon\epsilon(\beta_1 eh+\sigma_\epsilon\epsilon h-c\sigma_\epsilon^2) \\ \beta_1^2e^2h+\beta_1 e\sigma_\epsilon\epsilon h-\beta_1 ec\sigma_\epsilon^2+\sigma_\epsilon\epsilon(\beta_1 eh+\sigma_\epsilon\epsilon h-c\sigma_\epsilon^2) & \beta_1^2e^2h^2+\sigma_\epsilon^2\epsilon^2h^2+c^2\sigma_\epsilon^4+2\beta_1 e\sigma_\epsilon\epsilon h^2-2\beta_1 ehc\sigma_\epsilon^2-2c\epsilon h\sigma_\epsilon^3 \end{pmatrix} \\
&= \begin{pmatrix} \beta_1^2\frac{k}{1+\beta_1^2k}-2\beta_1 c+1 & 0 \\ 0 & \beta_1^2\sigma_X^2\frac{k}{1+\beta^2k}-2\beta_1 c\sigma_X^2+\sigma_X^2 \end{pmatrix}\sigma_\epsilon^2 + \mathbf{O}(\sigma_\epsilon^3) \\
&= \begin{pmatrix} \frac{1}{1+\beta_1^2k} & 0 \\ 0 & \frac{\sigma_X^2}{1+\beta_1^2k} \end{pmatrix}\sigma_\epsilon^2 + O(\sigma_\epsilon^3), \\
E(AB^{'}) &= E(\epsilon^2\sigma_\epsilon^2-\sigma_\epsilon^2)e(1,W,Y) = (0,0,0), \\
E(AC^{'}) &= E(\epsilon^2\sigma_\epsilon^2-\sigma_\epsilon^2)e(1,X) = (0,0),
\end{aligned}$$

and

$$E(BC^{'}) = E[\sigma_\epsilon e\epsilon\begin{pmatrix} 1 & X \\ W & XW \\ Y & XY \end{pmatrix}] = -c\begin{pmatrix} 1 & 0 \\ 0 & \sigma_X^2 \\ \beta_0 & \beta_1\sigma_X^2 \end{pmatrix}\sigma_\epsilon^2 + O(\sigma_\epsilon^3),$$

by finding the corresponding terms in lemma 4 for every expectation. We also note that

$$E(AD^{'}) = \mathbf{0}, \quad E(BD^{'}) = \mathbf{0}, \quad E(CD^{'}) = \mathbf{0},$$

because the individuals in validation data set and in primary data set are independent. In conclusion, the quadratic term in (2.6) is

$$\begin{pmatrix} \rho EAA' & \rho EAB' & \rho EAC' \\ \rho EB'A & \rho EBB' & \rho EBC' \\ \rho EC'A & \rho EC'B & \rho ECC' + (1-\rho)EDD' \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{\rho k}{1+\beta_1^2 k} & 0 & \frac{\rho\beta_0 k}{1+\beta_1^2 k} & -\rho c & 0 \\ 0 & 0 & \frac{\rho k\sigma_X^2}{1+\beta_1^2 k} & \frac{\rho\beta_1 k\sigma_X^2}{1+\beta_1^2 k} & 0 & -\rho c\sigma_X^2 \\ 0 & \frac{\rho\beta_0 k}{1+\beta_1^2 k} & \frac{\rho\beta_1 k\sigma_X^2}{1+\beta_1^2 k} & \frac{\rho(\beta_0^2+\beta_1^2\sigma_X^2)k}{1+\beta_1^2 k} & -\rho\beta_0 c & -\rho\beta_1 c\sigma_X^2 \\ 0 & -\rho c & 0 & -\rho c\beta_0 & \frac{1+\rho\beta_1^2 k}{1+\beta_1^2 k} & 0 \\ 0 & 0 & -\rho c\sigma_X^2 & -\rho c\beta_1\sigma_X^2 & 0 & \frac{1+\rho\beta_1^2 k}{1+\beta_1^2 k}\sigma_X^2 \end{pmatrix} \sigma_\epsilon^2 + \mathbf{O}(\sigma_\epsilon^3). \qquad (A.2)$$

If terms of $O(\sigma_\epsilon^2)$ in (A.1) and terms of $O(\sigma_\epsilon^3)$ in (A.2) are ignored, then these terms becomes the matrices H and M in (2.6). Multiplying the matrix $H^{-1}MH^{-1'}$ and simplify the results, we have completed the proof of theorem 1.

Proof of Theorem 2. The proofs of theorem 2 is either similar or simpler than the one of theorem 1 in each steps, Hence only important computation results are presented. Lemma 3 is also applicable here, thus we assume $\mu = 0$.

The conventional R.C. seeks the solution of

$$\sum_{i\in V}(X_i - \gamma_0 - \gamma_1 W_i)\begin{pmatrix} 1 \\ W \end{pmatrix} = \mathbf{0},$$

$$\sum_{i\in V}(Y_i - \beta_0 - \beta_1 X_i)(\begin{matrix} 1 \\ X_i \end{matrix}) + \sum_{i\in P}(Y_i - \beta_0 - \beta_1(\gamma_0 + \gamma_1 W_i))\begin{pmatrix} 1 \\ (\gamma_0 + \gamma_1 W_i) \end{pmatrix} = \mathbf{0}, \qquad (A.3)$$

which equivalently, to solve

$$\sum_{i\in V}(X_i - \gamma_0 - \gamma_1 W_i)\begin{pmatrix} 1 \\ W \end{pmatrix} = \mathbf{0},$$

$$\sum_{i\in V}(Y_i - \beta_0 - \beta_1 X_i)(\begin{matrix} 1 \\ X_i \end{matrix}) + \sum_{i\in P}(Y_i - \beta_0 - \beta_1(\gamma_0 + \gamma_1 W_i))\begin{pmatrix} 1 \\ W_i \end{pmatrix} = \mathbf{0},$$

Denote

$$B_i^* = (X_i - \gamma_0 - \gamma_1 W_i)\begin{pmatrix} 1 \\ W \end{pmatrix}, C_i^* = (Y_i - \beta_0 - \beta_1 X_i)(\begin{matrix} 1 \\ X_i \end{matrix}),$$

and

$$D_i^* = (Y_i - \beta_0 - \beta_1(\gamma_0 + \gamma_1 W_i))\begin{pmatrix} 1 \\ W_i \end{pmatrix},$$

respectively. Let $\boldsymbol{\eta}^* = (\gamma_0, \gamma_1, \beta_0, \beta_1)$ be a part of parameters, then the asymptotic covariance of the estimator define by solving $(A.3)$ is

$$\begin{pmatrix} \rho E\frac{\partial B}{\partial \boldsymbol{\eta}^*} \\ \rho E\frac{\partial C}{\partial \boldsymbol{\eta}^*} + (1-\rho)E\frac{\partial D}{\partial \boldsymbol{\eta}^*} \end{pmatrix}^{-1}$$

$$\begin{pmatrix} \rho EBB' & \rho EBC' \\ \rho EC'B & \rho ECC' + (1-\rho)EDD' \end{pmatrix} \begin{pmatrix} \rho E\frac{\partial B}{\partial \boldsymbol{\eta}^*} \\ \rho E\frac{\partial C}{\partial \boldsymbol{\eta}^*} + (1-\rho)E\frac{\partial D}{\partial \boldsymbol{\eta}^*} \end{pmatrix}^{-1'}.$$

After straightforward computations, we found that

$$\begin{pmatrix} \rho E\frac{\partial B}{\partial \boldsymbol{\eta}^*} \\ \rho E\frac{\partial C}{\partial \boldsymbol{\eta}^*} + (1-\rho)E\frac{\partial D}{\partial \boldsymbol{\eta}^*} \end{pmatrix} = -\begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & \rho\sigma_X^2 & 0 & 0 \\ (1-\rho)\beta_1 & 0 & 1 & 0 \\ 0 & (1-\rho)\beta_1\sigma_X^2 & 0 & \sigma_X^2 \end{pmatrix} + O(\sigma_\epsilon^2),$$

and

$$\begin{pmatrix} \rho EBB' & \rho EBC' \\ \rho EC'B & \rho ECC' + (1-\rho)EDD' \end{pmatrix}$$

$$= \begin{pmatrix} \rho k & 0 & 0 & 0 \\ 0 & \rho k\sigma_X^2 & 0 & 0 \\ 0 & 0 & (1-\rho)\beta_1^2 k + 1 & 0 \\ 0 & 0 & 0 & (1-\rho)\beta_1^2 k\sigma_X^2 + \sigma_X^2 \end{pmatrix} \sigma_\epsilon^2 + O(\sigma_\epsilon^3).$$

Theorem 2 follows easily after multiplication of these matrices.

References

[1] Afifi, A. A., and Elashoff, R. M. (1967), Missing observations in multivariate statistics: II: Point estimation in simple linear regression. *Journal of the American Statistical Association,* **62**, 10-29.

[2] Beale, E. M. L., and Little, R. J. A. (1975), Missing values in multivariate analysis. *Journal of Royal Statistics Society*, Ser. B, **37**, 129-145

[3] Carroll, R. J., Rupper, D. and Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models*. Chapman & Hall, London.

[4] Carroll, R. J. and Stefanski, L. A. (1990). Approximate quasilikelihood estimation in models with surrogate predictor. *Journal of the American Statistical Association.* **85**, 652-663.

[5] Lee, L. and Sepanski, J. H. (1995). Estimation of linear and nonlinear errors-in-variables models using validation data. *Journal of the American Statistical Association.* **90**, 130-140.

[6] Little R. J. A. (1992). Regressio with missing X's: A Review. *Journal of the American Statistical Association.* **87**, 1227-1237.

[7] Pepe, M. S. and Fleming, T. R. (1991) A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association.* **86**, 108-113.

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

[9] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63** , 581-592

[10] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

[11] Schafer, J. L. and Schenker, N. (2000). Inference with imputed conditonal means. *Journal of the American Statistical Association.* **95**, 144-154.