Regression Analysis with Block Missing Values and Variables Selection

Chien-Pai Han
Department of Mathematics
University of Texas at Arlington
United States
cphan@uta.edu

Yan Li
Department of Mathematics
University of Texas at Arlington
United States
liyanna@uta.edu

Abstract

We consider a regression model when a block of observations is missing, i.e. there are a group of observations with all the explanatory variables or covariates observed and another set of observations with only a block of the variables observed. We propose an estimator of the regression coefficients that is a combination of two estimators, one based on the observations with no missing variables, and the other the set all observations after deleting of the block of variables with missing values. The proposed combined estimator will be compared with the uncombined estimators. If the experimenter suspects that the variables with missing values may be deleted, a preliminary test will be performed to resolve the uncertainty. If the preliminary test of the null hypothesis that regression coefficients of the variables with missing value equal to zero is accepted, then only the data with no missing values are used for estimating the regression coefficients. Otherwise the combined estimator is used. This gives a preliminary test estimator. The properties of the preliminary test estimator and comparisons of the estimators are studied by a Monte Carlo study.

Keywords: Missing data; Combined estimator; Preliminary test estimator; Comparisons of regression coefficient estimators; Missing values.

1. Introduction

We consider a regression model with a block of observations missing. The model can be written as follows,

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad i=1, 2, \ldots, n_1 \qquad (1)$$

and

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \quad i=1, 2, \ldots, n_2 \qquad (2)$$

where *y* is the response variable, *x*'s are explanatory variables, ε is the random error, n_1+n_2=*n*, and $1 \leq k < p$.

It is seen that the first n_1 observations in equation (1) have no missing value. The last n_2 observations in equation (2) have the observations in the last *p*−*k* variables missing. This situation can happen in practice. For example most graduate schools in United States universities receive admission applications from both US residents and foreign students. The scores of TOEFL (Test of English as a Foreign Language) are required for foreign students, but not for US students. A regression equation of grade point average (GPA) in graduate study on undergraduate GPA, Graduate Record Examination (GRE) scores and TOEFL score can be considered. In this case the TOEFL scores on all US students are missing. We will consider this example later.

When there are missing values in a regression, the usual estimators of the regression coefficients are not obtainable. We propose an estimator that is a combination of two regression coefficient estimators, one based on the $n_1$observations with no missing variables, and the other on all *n* observations after deleting the block of variables with missing values. This estimating procedure is different from the usual procedure of imputation. There are numerous research on missing values, for example, papers by Ibrahim, J. G. (1990), Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (1999), Meng, X.-L. (2000), Raghunathan, T. E. (2004), Rubin, D. B. (1976, 1996) and others; and books by Allison, P. D. (2002), Little, R. J. A., and Rubin, D. B. (2002), Rubin, D. B. (1987). The most popular technique dealing with missing data is imputation. But imputation may not be appropriate in some practical situations. The above mentioned admission data is one of such situation. The US students never take TOEFL, so TOEFL scores do not exist for US students. Hence imputation does not make sense. Our proposed procedure avoids using imputation and has good properties.

If the experimenter suspects that the variables with missing values may be deleted, a preliminary test will be performed to resolve the uncertainty. If the preliminary test of the null hypothesis that regression coefficients of the variables with missing value equal to zero is accepted, then only the data with no missing values are used for estimating the regression coefficients. Otherwise the combined estimator is used. This gives a preliminary test estimator. Preliminary test procedures have been studied extensively since the first paper by Bancroft (1944), see e.g. Han and Bancroft (1968, 1978), Johnson, Bancroft, and Han (1977), Giles and Giles (1993) Kennedy and Bancroft (1971), among others. Two bibliographies, Bancroft and Han (1977) and Han, Rao and Ravichandran (1988) are published. The two books by Judge and Bock (1978) and Saleh (2006), give excellent treatment of preliminary test and shrinkage estimation.

This paper is organized as following. Section 2 discusses the proposed combined estimator. The variable selection using preliminary test and the preliminary test estimator are given in Section 3. Section 4 compares the various estimators in terms of bias and mean square error by a Monte Carlo study. The admission data is given as an example in Section 5. Section 6 gives the conclusion.

2. Combined Estimator of Regression Coefficients

We consider two regression equations separately, one based on the first n_1 observations in equation (1) with no missing value and the other on all *n* observation with only the first *k* explanatory variables. The following vectors and matrices are partitioned.

Let

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix} \tag{3}$$

where

$$\boldsymbol{Y}_1 = \begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \end{bmatrix} \qquad \boldsymbol{Y}_2 = \begin{bmatrix} y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \end{bmatrix}$$

$$\boldsymbol{\beta}_1 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \boldsymbol{\beta}_2 = \begin{bmatrix} \beta_{k+1} \\ \vdots \\ \beta_p \end{bmatrix}$$

$$\boldsymbol{\varepsilon}_1 = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{n_1} \end{bmatrix} \qquad \boldsymbol{\varepsilon}_2 = \begin{bmatrix} \varepsilon_{n_1+1} \\ \vdots \\ \varepsilon_{n_1+n_2} \end{bmatrix}$$

The model in equation (1) can be written as

$$\boldsymbol{Y}_1 = \boldsymbol{Z}_b\boldsymbol{\beta} + \boldsymbol{\varepsilon}_1 \tag{4}$$

where **β** is the *p* × 1 vector of regression coefficients,

$$\boldsymbol{Z}_b = [\boldsymbol{X}_{11} \quad \boldsymbol{X}_{12}] \tag{5}$$

$$\boldsymbol{X}_{11} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & & \ddots & \vdots \\ 1 & x_{n_1 1} & \cdots & x_{n_1 k} \end{bmatrix}, \quad \boldsymbol{X}_{12} = \begin{bmatrix} x_{1\,k+1} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n_1\,k+1} & \cdots & x_{n_1 p} \end{bmatrix}$$

We assume that $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\boldsymbol{V}(\varepsilon) = \sigma^2\mathbf{I}$. The least squares estimator of **β** is

$$\boldsymbol{b} = (\boldsymbol{Z}_b'\boldsymbol{Z}_b)^{-1}\boldsymbol{Z}_b'\boldsymbol{Y}_1 \tag{6}$$

assuming the inverse exists. This estimator uses the first n_1 observations and includes all the *p* explanatory variables. It is unbiased and the covariance matrix of ***b*** is

$$\boldsymbol{V}(\boldsymbol{b}) = (\boldsymbol{Z}_b'\boldsymbol{Z}_b)^{-1}\sigma^2 \tag{7}$$

Now let us consider the model based on all *n* observations but using only the first *k* explanatory variables,

$$\boldsymbol{Y} = \boldsymbol{Z}_k\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \tag{8}$$

Where

$$\boldsymbol{Z}_k = \begin{bmatrix} \boldsymbol{X}_{11} \\ \boldsymbol{X}_{21} \end{bmatrix} \tag{9}$$

and

$$\boldsymbol{X}_{21} = \begin{bmatrix} 1 & x_{n_1+1\,1} & \cdots & x_{n_1+1\,k} \\ \vdots & & \ddots & \vdots \\ 1 & x_{n\,1} & \cdots & x_{n\,k} \end{bmatrix}$$

The least squares estimator of $\boldsymbol{\beta}_1$ is

$$\widehat{\boldsymbol{\beta}_1} = (\boldsymbol{Z}_k'\boldsymbol{Z}_k)^{-1}\boldsymbol{Z}_k'\boldsymbol{Y} \tag{10}$$

The estimator $\widetilde{\boldsymbol{\beta}_1}$ is biased. The bias will be obtained by simulation in Section 4. The covariance matrix of $\widetilde{\boldsymbol{\beta}_1}$ is

$$\mathbf{V}(\widetilde{\boldsymbol{\beta}_1}) = (\boldsymbol{Z}_k'\boldsymbol{Z}_k)^{-1}\sigma^2 \tag{11}$$

We have two estimators of $\boldsymbol{\beta}_1$, one from the model in equation (4) and the other from the model in equation (8). If we partition

$$\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix}$$

as $\boldsymbol{\beta}$, then the two estimators of $\boldsymbol{\beta}_1$ are $\widetilde{\boldsymbol{\beta}_1}\ and\ \boldsymbol{b}_1$. We now construct the combined estimator of $\boldsymbol{\beta}_1$ as the weighted average of these two estimators. We use the weight that is proportional to the inverse of the determinant of the covariance matrix. Let

$$\boldsymbol{C} = (\boldsymbol{Z}_b'\boldsymbol{Z}_b)^{-1} = \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{bmatrix}$$

where $\boldsymbol{C}_{11}$ is a $k \times k$ matrix, i.e. the $\boldsymbol{C}$ matrix is partitioned according to the block of missing values. So the covariance matrix of $\boldsymbol{b}_1$ is

$$V(\boldsymbol{b}_1) = \boldsymbol{C}_{11}\sigma^2$$

The combined estimator of $\boldsymbol{\beta}_1$ is defined as

$$\widetilde{\boldsymbol{\beta}_{1c}} = w\widetilde{\boldsymbol{\beta}_1} + (1 - w)\boldsymbol{b}_1$$

where

$$w = \frac{1/v_1}{1/v_1 + 1/v_2}$$

and

$$v_1 = \det(\boldsymbol{Z}_k'\boldsymbol{Z}_k)^{-1}$$

$$v_2 = \det \boldsymbol{C}_{11}$$

Note that there is only one estimator of $\boldsymbol{\beta}_2$ i.e. $\boldsymbol{b}_2$. Hence a combined estimator of $\boldsymbol{\beta}$ is

$$\widetilde{\boldsymbol{\beta}_c} = \begin{bmatrix} \widetilde{\boldsymbol{\beta}_{1c}} \\ \boldsymbol{b}_2 \end{bmatrix} \tag{12}$$

The bias and mean square error of $\widetilde{\boldsymbol{\beta}_c}$ will be studied in Section 4.

3. Variable Selection after Testing the Regression Coefficients

The variables with the block of missing values may be viewed as candidates for deleting. In particular when the experimenter suspect that $\boldsymbol{\beta}_2$ may be zero, he/she can test H_o: $\boldsymbol{\beta}_2 = \mathbf{0}$ against H_1: $\boldsymbol{\beta}_2 \neq \mathbf{0}$ to resolve the uncertainty. An *F* test can be used in the model given in equation (4) when the error has a normal distribution. The *F* statistic is

$$F = \frac{(n-p-1)\left[R(x_o, x_1, \dots, x_p) - R(x_o, x_1, \dots, x_k)\right]}{(p-k)(\boldsymbol{Y} - \boldsymbol{Z}_b\boldsymbol{b})'(\boldsymbol{Y} - \boldsymbol{Z}_b\boldsymbol{b})}$$

where R(·) denotes the reduction in sum of squares due to fitting the model involving the explanatory variable in the parenthesis, and $x_o \equiv 1$. The *F* statistic has an F($p-k$, $n-p-1$) distribution under H_o. If H_o is accepted, we delete the last $p-k$ explanatory variable; otherwise all variables are kept in the model. Hence an estimator of **β** is

$$\widehat{\boldsymbol{\beta}_{PTE}} = \begin{cases} \begin{bmatrix} \widehat{\boldsymbol{\beta}_c} \\ \mathbf{0} \end{bmatrix} & \text{if } F \leq F_\alpha \\ \begin{bmatrix} \widehat{\boldsymbol{\beta}_c} \\ \boldsymbol{b}_2 \end{bmatrix} & \text{if } F > F_\alpha \end{cases} \qquad (13)$$

This is a preliminary test estimator. The exact bias and MSE of $\widehat{\boldsymbol{\beta}_{PTE}}$ are difficult to obtain. Hence we will use simulation to study them that is given in Section 4.

Here we select or delete the block of explanatory variables when those variables with missing values are doubtful. One may also consider the sequential deletion procedure and forward selection procedure such as given in Kennedy and Bancroft (1971). We will consider these procedures in a future paper.

4. Comparison of Estimators

In this section we compare the estimators of **β** given in Section 3 by a Monte Carlo study.

We considered the following linear regression model with three covariates, $x_1, x_2,$ and x_3, and sample size *n* = 50:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i, \text{for } i = 1, \dots, n, \qquad (14)$$

where ε iid $N(0, \sigma_\varepsilon^2)$. We first generate the three covariates from a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \sim \text{tri-normal}\left(\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{bmatrix} \right)$$

so that the correlations between any pair of covariates $\rho(x_l, x_{l'}) = 0.5$ for *l* ≠ *l'* = 1, 2, 3. We fix β_0 = 0.5, $\beta_1 = 1$, $\beta_2 = 1$, and $\sigma_\varepsilon^2 = 1$. For given ($x_1, x_2,$ and x_3), we generate the response variable *y* from model (14) with selected values of β_3 given in Tables 1 and 2. The simulated data $\{y_i, x_{1i}, x_{2i}, x_{3i}), i = 1,2, \dots, n\}$ are then used to estimate regression coefficients $\boldsymbol{\beta}$. The variable x_3 will have a block of missing values. For comparison purpose, five estimators are considered:

1) The estimator ***b*** defined in (6), estimated by the simulated data $\{(y_i, x_{1i}, x_{2i}, x_{3i}), i = 1,2, \dots, n_1\}$ and $n_1 = (1 - \text{missing rate of } x_3) \times n,$ where missing rate of x_3 is specified in Tables 1-2;
2) The estimator $\widehat{\boldsymbol{\beta}_1}$ defined in (10), estimated by the simulated data $\{(y_i, x_{1i}, x_{2i}), i = 1,2, \dots, n\}$;
3) The combined estimator $\widehat{\boldsymbol{\beta}_c}$ defined in (12);
4) The PTE defined in (13) with significance level 0.05 for testing H$_0$: $\beta_3 = 0$; and

5) The PTE defined in (13) with significance level 0.25 for testing H$_0$: $\beta_3 = 0$.

For each simulation condition in Tables 1-2, we conduct R (=10,000) simulation runs and evaluate the performance of the five estimators in terms of bias and the mean squared error (MSE), defined as, respectively,

$$Bias = \frac{1}{R}\sum_{r=1}^{R}\widehat{\boldsymbol{\beta}_r} - \boldsymbol{\beta}, \text{ and } MSE = \frac{1}{R-1}\sum_{r=1}^{R}\left(\widehat{\boldsymbol{\beta}_r} - \boldsymbol{\beta}\right)^2,$$

where $\widehat{\boldsymbol{\beta}_r}$ is the estimate of $\boldsymbol{\beta}$ in the r^{th} simulation run and r = 1, 2,…, R.

It is expected that **b** is unbiased, but produces larger variance than $\widehat{\boldsymbol{\beta}_1}$. The estimator $\widehat{\boldsymbol{\beta}_1}$ considers all the observations in the analysis, therefore more efficient. However, $\widehat{\boldsymbol{\beta}_1}$ is biased because variable x_3 is not included in the analysis model. The advantage of efficiency of $\widehat{\boldsymbol{\beta}_1}$ is becoming more dramatic as missing block gets larger. As a result, the combined estimator is constructed by weighting heavily toward the biased estimator $\widehat{\boldsymbol{\beta}_1}$. Tables 1-2 show the bias and MSE, respectively, of the five estimators with varying missing rates on x_3 and varying effects of x_3 on y (β_3). Consistent with our expectation, $\widehat{\boldsymbol{\beta}_c}$ and $\widehat{\boldsymbol{\beta}_1}$ produce similar bias, especially when missing rate is high (80%). When missing rate is small (20%), the combined estimator $\widehat{\boldsymbol{\beta}_c}$ are less biased due to the strength borrowed from the unbiased estimator ***b***. Estimators of $\widehat{\boldsymbol{\beta}_1}$ and $\widehat{\boldsymbol{\beta}_c}$ produce similar MSE, both smaller than the unbiased estimator ***b*** and PTE estimators when missing block is large or the effect of missing variable x_3 on y (β_3) is small; otherwise, unbiased estimator **b** and PTE (0.25) are superior, producing smaller bias and smaller MSE. We have also considered other values of the total number of covariates (p), the number of covariates without missing values (k), and the sample size (n). The results have similar patterns to those given above, hence not given here to save space.

5. Numerical Example

We illustrate our proposed methods using data given by admissions office at the University of Texas at Arlington. The data sets contain two populations. One population is the Success Group in which the students receive their master's degree with the master GPA (y) greater than or equal to 3.0. The other population is the Failure group (y<3.0) where the students don't complete their master degree. For each population, there are 10 foreign students and 10 United States students. Each foreign student has 5 variables which are x_1 = undergraduate GPA, x_2 = GRE verbal, x_3 = GRE quantitative, x_4 = GRE analytic, and x_5 = TOEFL score. For each United States student, one variable, x_5 = TOEFL score is missing. We standardized the variables of x_1 ~ x_5 by subtracting their mean and divided by their standard error. Table 3 presents the regression coefficient estimates using the five methods. The results are consistent with what we found in simulation studies. The pair of ($\widehat{\boldsymbol{\beta}_c}$ and $\widehat{\boldsymbol{\beta}_1}$) or (***b*** and PTE (0.25)) produces similar results. Considering the moderate missing rate (50%) and the small estimated effect of x_5 on y ($\widehat{\beta_5}$=0.22), we recommend $\widehat{\boldsymbol{\beta}_c}$ or $\widehat{\boldsymbol{\beta}_1}$ should be reported for this example.

Table 1: Bias of regression coefficient estimators

	missing rate = 80%				missing rate = 50%				missing rate = 20%			
	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
	$\beta_3 = 0$											
$\boldsymbol{b}$	0.002	0.005	0.000	-0.007	-0.002	-0.003	-0.002	0.004	-0.002	-0.002	0.003	0.002
$\widetilde{\boldsymbol{\beta}_1}$	0.002	-0.002	0.001	0.000	0.000	0.000	0.001	0.000	-0.003	0.000	0.004	0.000
$\widetilde{\boldsymbol{\beta}_c}$	0.002	-0.002	0.001	-0.007	-0.001	0.000	0.000	0.004	-0.003	0.000	0.003	0.002
PTE(.05)	0.001	-0.001	0.001	0.000	-0.002	-0.002	0.000	0.005	-0.004	0.000	0.004	0.001
PTE(.25)	0.002	-0.002	0.004	-0.001	-0.002	-0.004	0.001	0.005	-0.003	-0.001	0.003	0.001
	$\beta_3 = 0.1$											
$\boldsymbol{b}$	-0.014	0.013	0.003	-0.004	0.007	-0.001	-0.002	-0.002	-0.001	0.001	-0.002	0.005
$\widetilde{\boldsymbol{\beta}_1}$	0.058	0.022	0.020	-0.100	0.061	0.022	0.019	-0.100	0.061	0.022	0.018	-0.100
$\widetilde{\boldsymbol{\beta}_c}$	0.058	0.022	0.020	-0.004	0.060	0.021	0.019	-0.002	0.051	0.018	0.015	0.005
PTE(.05)	0.044	0.012	0.024	-0.079	0.047	0.011	0.016	-0.072	0.046	0.015	0.011	-0.069
PTE(.25)	0.020	-0.006	0.030	-0.037	0.027	-0.001	0.007	-0.030	0.021	0.004	0.002	-0.023
	$\beta_3 = 0.3$											
$\boldsymbol{b}$	0.002	-0.001	0.002	-0.002	0.004	-0.001	0.002	-0.004	0.001	-0.003	0.001	0.001
$\widetilde{\boldsymbol{\beta}_1}$	0.181	0.063	0.057	-0.300	0.182	0.059	0.060	-0.300	0.180	0.058	0.061	-0.300
$\widetilde{\boldsymbol{\beta}_c}$	0.180	0.063	0.057	-0.002	0.176	0.057	0.058	-0.004	0.150	0.048	0.051	0.001
PTE(.05)	0.141	0.028	0.068	-0.225	0.129	0.026	0.044	-0.199	0.110	0.027	0.032	-0.166
PTE(.25)	0.058	-0.022	0.077	-0.087	0.055	-0.010	0.023	-0.068	0.042	0.002	0.008	-0.047
	$\beta_3 = 0.5$											
$\boldsymbol{b}$	-0.005	-0.001	0.003	0.005	0.002	-0.002	-0.003	0.003	-0.003	0.003	0.000	0.000
$\widetilde{\boldsymbol{\beta}_1}$	0.297	0.102	0.100	-0.500	0.302	0.099	0.099	-0.500	0.298	0.103	0.099	-0.500
$\widetilde{\boldsymbol{\beta}_c}$	0.296	0.102	0.100	0.005	0.293	0.096	0.095	0.003	0.248	0.086	0.083	0.000
PTE(.05)	0.210	0.046	0.111	-0.348	0.164	0.028	0.056	-0.248	0.105	0.029	0.028	-0.156
PTE(.25)	0.078	-0.029	0.106	-0.117	0.055	-0.012	0.022	-0.066	0.025	0.006	0.005	-0.031

Table 2: MSE of regression coefficient estimators

	missing rate = 80%				missing rate = 50%				missing rate = 20%			
	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
						$\beta_3 = 0$						
$\boldsymbol{b}$	0.473	0.354	0.304	0.268	0.194	0.099	0.084	0.117	0.114	0.050	0.051	0.061
$\widehat{\boldsymbol{\beta}_1}$	0.084	0.043	0.043	0.000	0.086	0.043	0.044	0.000	0.086	0.044	0.044	0.000
$\widehat{\boldsymbol{\beta}_c}$	0.084	0.043	0.043	0.268	0.087	0.043	0.044	0.117	0.087	0.044	0.044	0.061
PTE(.05)	0.118	0.072	0.062	0.059	0.096	0.048	0.047	0.028	0.090	0.045	0.045	0.017
PTE(.25)	0.213	0.147	0.117	0.178	0.125	0.063	0.055	0.082	0.098	0.047	0.047	0.043
						$\beta_3 = 0.1$						
$\boldsymbol{b}$	0.475	0.361	0.301	0.263	0.195	0.100	0.082	0.121	0.113	0.050	0.052	0.062
$\widehat{\boldsymbol{\beta}_1}$	0.088	0.045	0.044	0.010	0.089	0.044	0.044	0.010	0.089	0.044	0.045	0.010
$\widehat{\boldsymbol{\beta}_c}$	0.088	0.045	0.044	0.263	0.090	0.044	0.044	0.121	0.089	0.044	0.045	0.062
PTE(.05)	0.125	0.070	0.061	0.063	0.103	0.051	0.047	0.043	0.094	0.045	0.046	0.029
PTE(.25)	0.223	0.157	0.119	0.179	0.131	0.066	0.056	0.093	0.102	0.047	0.049	0.051
						$\beta_3 = 0.3$						
$\boldsymbol{b}$	0.477	0.359	0.297	0.269	0.192	0.100	0.082	0.120	0.114	0.050	0.052	0.061
$\widehat{\boldsymbol{\beta}_1}$	0.118	0.048	0.047	0.090	0.117	0.047	0.046	0.090	0.117	0.047	0.048	0.090
$\widehat{\boldsymbol{\beta}_c}$	0.118	0.048	0.047	0.269	0.117	0.047	0.046	0.120	0.118	0.047	0.048	0.061
PTE(.05)	0.160	0.085	0.071	0.154	0.137	0.058	0.055	0.133	0.127	0.048	0.053	0.121
PTE(.25)	0.256	0.168	0.135	0.231	0.155	0.073	0.067	0.130	0.118	0.049	0.056	0.082
						$\beta_3 = 0.5$						
$\boldsymbol{b}$	0.467	0.355	0.293	0.275	0.193	0.098	0.083	0.119	0.114	0.050	0.050	0.062
$\widehat{\boldsymbol{\beta}_1}$	0.174	0.054	0.054	0.250	0.171	0.053	0.053	0.250	0.172	0.054	0.053	0.250
$\widehat{\boldsymbol{\beta}_c}$	0.174	0.054	0.054	0.275	0.171	0.053	0.053	0.119	0.173	0.054	0.053	0.062
PTE(.05)	0.226	0.103	0.088	0.302	0.203	0.070	0.070	0.295	0.177	0.053	0.062	0.253
PTE(.25)	0.315	0.197	0.162	0.307	0.189	0.080	0.082	0.172	0.128	0.051	0.061	0.098

Table 3: Regression coefficient estimates from real data analysis

	β_0	β_1	β_2	β_3	β_4	β_5
$\boldsymbol{b}$	2.772	-0.133	-0.534	0.115	0.203	0.220
$\widehat{\boldsymbol{\beta}_1}$	3.128	0.123	-0.100	0.092	0.159	
$\widehat{\boldsymbol{\beta}_c}$	3.126	0.122	-0.102	0.092	0.159	0.220
PTE(.05)	3.128	0.123	-0.100	0.092	0.159	
PTE(.25)	2.772	-0.133	-0.534	0.115	0.203	0.220

6. Conclusion

Missing value occurs in real data analysis frequently. We consider the case that observations of a block of variables are missing for part of the data. The unbiased estimator ***b*** uses only n_1 out of *n* observations, so some information is lost and it has larger variance. The estimator $\widetilde{\boldsymbol{\beta}_1}$ uses all *n* observations and ignores the last *p* – *k* variables; so $\widetilde{\boldsymbol{\beta}_1}$ is biased. It does not give an estimate of $\boldsymbol{\beta}_2$. The combined estimator $\widetilde{\boldsymbol{\beta}_c}$ is a weight average of ***b*** and $\widetilde{\boldsymbol{\beta}_1}$ so it is a compromise of the two estimators. It has good properties, i.e. controls the bias and MSE well. If the experimenter has prior information the $\boldsymbol{\beta}_2$ may be zero but is not certain, he can use a preliminary test to resolve the uncertainty. The preliminary test estimator can be used in such a case. In general the preliminary test estimator is good when the value of the parameter is close to the null hypothesis.

References

1. Allison, P. D. (2002). *Missing Data*, SAGE University Papers.
2. Bancroft, T. A. (1944). On biases in estimation due to use of preliminary tests of significance. Ann Math Stat 15:190–204.
3. Bancroft T. A. and Han, C-P. (1977). Inference based on conditional specification: a note and a bibliography. *International Statistical Review*, 45, 117-127, 1977.
4. Giles J. A., Giles D.E.A. (1993). Pre-test estimation in econometrics: recent developments. J Econ Surv 7:145–197.
5. Han,C-P. and Bancroft TA (1968) On pooling means when variance is unknown. JAmStat Assoc 63:1333–1342.
6. Han, C.-P. and Bancroft, T. A. (1978). Estimating regression coefficients under conditional specification. *Communications in Statistics, Part A - Theory and Methods*, A7, 47-56, 1978.
7. Han C-P., Rao CV, Ravichandran J. (1988). Inference based on conditional specification: a second bibliography. Commun Stat Theory Methods 17: 1945–1964.
8. Ibrahim, J. G. (1990). Incomplete Data in Generalized Linear Models, *Journal of the American Statistical Association*, 85, 765–769.
9. Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (1999). Monte Carlo EM for Missing Covariates in Parametric Regression Models, *Biometrics*, 55, 591–596.
10. Johnson, J. P., Bancroft, T. A. and Han, Chien-Pai (1977). A pooling methodology for regressions in prediction, *Biometrics*, 33, 57-67.
11. Judge, G. G. and Bock, M. E. (1978). *The Statistical Implication of Pre-test and Stein-Rule Estimators in Econometrics*. North-Holland.
12. Kennedy, W. J. and Bancroft, T. A. (1971). Model building for prediction in regression based upon repeated significance tests, *The Annals of Mathematical Statistics*, 42, 1273-1284.
13. Little, R. J. A., and Rubin, D. B. (2002). *Statistical Analysis With Missing Data* (2nd ed.), New York: Wiley.

14. Meng, X. L. (2000). Missing Data: Dial M for ???, *Journal of the American Statistical Association*, 95, 1325–1330.

15. Raghunathan, T. E. (2004). What Do We Do With Missing Data? Some Options for Analysis of Incomplete Data, *Annual Review of Public Health*, 25, 99–117.

16. Rubin, D. B. (1976). “Inference and Missing Data,” *Biometrika*, 63, 581–590.

17. Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

18. Rubin, D. B. (1996). Multiple Imputation After 18+Years, *Journal of the American Statistical Association*, 91, 473–489.

19. Saleh A. K. Md (2006). Theory of preliminary test and stein-type estimations with applications. Wiley, New York.