



Thailand Statistician
July 2006; 4: 13-26
<http://statassoc.or.th>
Contributed paper

A Comparison of Estimation Methods for Missing Data in Multiple Linear Regression with Two Independent Variables

Sujitta Suraphee*, Chanchaen Raksmanee, Jaruchat Busaba, Chanchai Chaisorn, and Wilaiwan Nakornthai

Department of Mathematics, Faculty of Science, Mahasarakham University,

Mahasarakham 44150, Thailand.

***Author for correspondence**, e-mail: suraphee@hotmail.com

Received: 10 October 2005

Accepted: 28 February 2006

Abstract

In multiple linear regression, if the incomplete values occur in sample, many researchers will use the statistical software packages which delete the cases that deal with missing value. Then they use the ordinary least square (OLS) to estimate parameters in the regression model. However, the validity of this approach is decreased in making inference because the size of the sample reduces. In this paper, when missing values occur in response variable, we perform a simulation study of imputation – based procedures and indicate that missing values should be filled by the EM algorithm and regression imputation methods, which have been more efficient than deletion of observations. However, in case of having the large sample sizes and small variances of errors and small proportions of missing values, pairwise deletion methods can be used efficiently. Further, we analyze completed data by ordinary least square method to predict regression coefficients in the regression model. This leads to a result with minimum mean squared errors.

Keywords: EM algorithm, Missing Data, Multiple Regression, Pairwise Deletion, Regression Imputation.

1. Introduction

The study of missing values is one of the most important topics in applied statistics, especially in survey problems and experimental data. In this paper, we study the problem of missing values of response variable in multiple linear regression using two methods for filling those missing values. Firstly, a procedure for imputation is regression imputation, where missing variables for units are estimated by values from regression on the known variables for those units. Secondly, we use the EM algorithm, which is an iterative computational method to get a maximum-likelihood estimate. Finally, we use pairwise deletion, so we simply don't include an observation that is missing on one of the items. Moreover, three methods are compared using the mean squared errors (MSE) and show descriptive statistics of MSE. This research is based on Monte Carlo simulation.

2. Materials and Methods

2.1 Regression Imputation Method

The method is an explicit statistical model imputation with the following steps:

1. Calculate regression coefficients of Y on X using complete cases by least squares method. We have that

$$\hat{\beta}^* = (X^{*T} X^*)^{-1} \cdot (X^{*T} Y^*)$$

where X^* is a matrix of independent variables with complete cases,

Y^* is a vector of response variable with complete cases.

2. Compute the fitted value

$$\hat{Y}_i^* = \hat{\beta}_0^* + \hat{\beta}_1^* X_{1i}^* + \hat{\beta}_2^* X_{2i}^*$$

each nonrespondent i . For each missing point, we impute $Y_{mis} = \hat{y}_i^*$.

The values of Y_i after imputation are therefore

$$Y_i = \begin{cases} y_i & , \text{ if } y_i \text{ is observed,} \\ \hat{y}_i^* = \hat{\beta}_0^* + \hat{\beta}_1^* X_{1i} + \hat{\beta}_2^* X_{2i} & , \text{ if } y_i \text{ is missing,} \end{cases}$$

$$i=1,2,\dots,n.$$

After that the completed data are analyzed by the ordinary least squares method to predict regression coefficients in the regression model.

2.2 The EM Algorithm

The EM algorithm is an iterative method for computation of the maximum of the posterior distribution of observed data. The EM algorithm formalizes a relatively old ad hoc idea for handling missing data: (1) replace missing values by estimate values, (2) estimate parameters, (3) reestimate the missing values assuming the new parameter estimates are correct, (4) reestimate parameters, and so forth, iterating until convergence. Such methods are EM algorithms for models where the complete-data loglikelihood $l(\theta | Y_{obs}, Y_{mis})$ indexed by unknown parameter θ is linear in Y_{mis} ; more generally, missing sufficient statistics rather than individual observations to be estimated, and even more generally, the loglikelihood $l(\theta | Y)$ itself needs to be estimated at each iteration of the algorithm. Each iteration of the EM algorithm consists of two steps : expectation (E step) and maximization (M step).

The E step finds the conditional expectation of “missing data” given the observed data and current estimated parameters, and then substitutes these expectations for the “missing data”.

The M step is to describe perform maximum likelihood estimation of θ just as if they have been filled in. Thus the M step of EM uses the identical computational methods as ML estimation from $l(\theta | Y)$.

Specifically, let $\theta^{(t)}$ be the current estimate of the parameter θ . The E step of EM finds the expected loglikelihood if θ were $\theta^{(t)}$. Let

$$Q(\theta|\theta^{(t)}) = \int l(\theta|Y) \cdot f(Y_{\text{mis}}|Y_{\text{obs}}, \theta = \theta^{(t)}) dY_{\text{mis}}.$$

The M step of EM determines $\theta^{(t+1)}$ by maximizing this expected loglikelihood

$$Q(\theta^{(t+1)}|\theta^t) \geq Q(\theta|\theta^{(t)}), \quad \text{for all } \theta \text{ [3].}$$

2.3 Apply EM to Multiple Linear Regression with Missing Values

Consider the multiple linear regression model

$$Y = X\beta + \varepsilon$$

where Y is an $n \times 1$ vector of response variable,

X is an $n \times p$ matrix with $p - 1$ explanatory variables,

β is a $p \times 1$ vector of regression coefficients,

and ε is an $n \times 1$ vector of errors distributed as $N(0, \sigma^2 I)$.

When data are incomplete, some values of the $n \times 1$ vector of response variable (y_i) are "missing at random" (MAR) [2] then the EM will be applied to this problem.

When $\varepsilon_i \sim N(0, \sigma_\varepsilon^2 I)$, it can be proved that $y_i \sim N(\mu_{y|x}, \sigma_y^2)$. So the M step corresponds to the least square analysis and the E step involves finding the expected values of the missing y_i 's given the current estimates of parameters $\theta_{y \cdot x}^{(t)} = \beta_{y \cdot x}^{(t)}$.

$$E(y_i | X, Y_{\text{obs}}, \theta_{y \cdot x}^{(t)}) = \begin{cases} y_i & , \quad \text{if } y_i \text{ is observed } (i = m_0 + 1, \dots, n) \\ \beta_{y \cdot x}^{(t)} \cdot x_i & , \quad \text{if } y_i \text{ is missing } (i = 1, \dots, m_0) \end{cases}$$

where $Y^{(t)}$ is the vector Y with missing components y_i replaced by estimates from the E step at iteration. The M step calculates

$$\beta_{y \cdot x}^{(t+1)} = (X^T X)^{-1} \cdot (X^T Y^{(t)}) \dots \dots \dots [1]$$

(see details in appendix).

2.4 Pairwise Deletion Method

For this method, we will estimate the regression coefficients by deleting cases with missing values. Using the complete responses, we get

$$\hat{\beta} = (X^{*T} X^*)^{-1} \cdot (X^{*T} Y^*)$$

where X^* and Y^* are matrix observed values of X and Y .

2.5 A Simulation Study

In this research, we use sample sizes $n = 20, 50, 70, 100$, and 150 . The parameters of regression coefficients are assigned to $\beta_0 = 10, \beta_1 = 1$ and $\beta_2 = 2$ and $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$; $\sigma_\varepsilon^2 = 25, 50, 75, 100$.

Consider the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

where both X_{1i} and X_{2i} are generated from the standard normal distribution. Once the data are generated, let 10%, 30%, and 50% of the elements of Y vector be randomly missing.

Then apply two methods of imputation, EM algorithm and regression imputation, and use another method, pairwise deletion. Moreover, the ordinary least square method

is used to find parameter estimates. Finally, we calculate MSE from three sets of data in order to consider the minimum value.

3. Simulation Results

The empirical results in Tables 1-3 are the number of minimum mean squared errors of the 500 iterations for comparing the performance of three missing estimates, EM algorithm, regression imputation, and pairwise deletion. Moreover, the Tables 4-5 display descriptive statistics of mean squared errors for considering difference of three methods.

Table 1. The number of minimum mean squared errors of each method in 500 iterations are classified to variances of errors (σ_{ε}^2), size of sample ($n = 20$).

variances of errors (σ_{ε}^2)	Proportions of missing values (%)	Methods		
		Pairwise Deletion	Regression Imputation	EM algorithm
25	10	41	358	398*
	30	62	194	249*
	50	70	178	253*
50	10	19	355	374*
	30	34	199	268*
	50	28	215	255*
75	10	15	350	378*
	30	20	220	261*
	50	15	225	260*
100	10	12	360	386*
	30	16	226	258*
	50	13	227	260*

Note: *maximum frequency of minimum mean squared errors.

From Table 1, we found that when the size of a sample is small ($n=20$), in all values of variances of errors and missing proportions in simulation, the maximum number of minimum mean squared errors in EM algorithm is greater than one in regression imputation and pairwise deletion. It means that the EM algorithm has results better than both regression imputation and pairwise deletion. Similarly, if the sizes of samples are larger than those we already considered before ($n=50,70$) with all values of variances of errors and missing proportions in this simulation, the EM algorithm still gives results better than the both methods.

Table 2. The number of minimum mean squared errors of each method in 500 iterations are classified to variances of errors (σ_{ε}^2), size of sample ($n = 100$).

Variances of errors (σ_{ε}^2)	Proportions of missing values (%)	The number of minimum mean squared errors		
		Pairwise Deletion	Regression Imputation	EM algorithm
25	10	204*	141	179
	30	198	103	203*
	50	148	87	267*
50	10	110	170	220*
	30	100	156	244*
	50	70	165	265*
75	10	77	186	237*
	30	65	172	263*
	50	56	179	265*
100	10	52	202	246*
	30	48	189	263*
	50	40	200	260*

Note: *maximum frequency of minimum mean squared errors.

Fig.1 The number of minimum mean squared errors of 500 simulated data sets (variance of error = 25, proportion of missing value = 10)

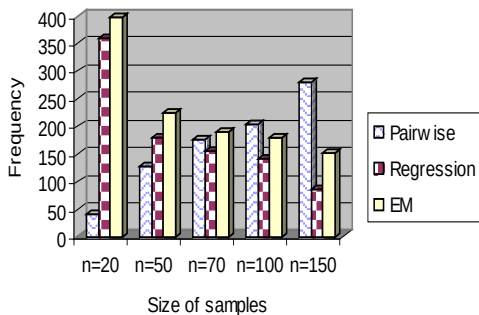


Fig.2 The number of minimum mean squared errors of 500 simulated data sets (variance of error = 25, proportion of missing value = 30)

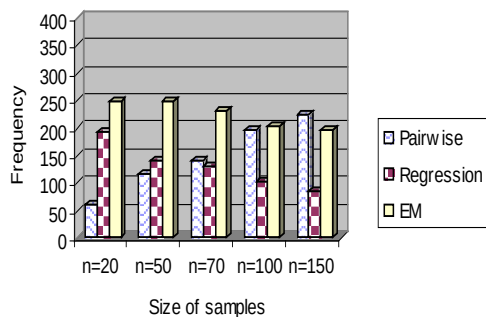


Table 3. The number of minimum mean squared errors of each method in 500 iterations are classified to variance of error (σ_{ε}^2), size of sample ($n = 150$).

Variances of errors (σ_{ε}^2)	Proportions of missing values (%)	The number of minimum mean squared errors		
		Pairwise Deletion	Regression Imputation	EM algorithm
25	10	281*	86	152
	30	223*	84	196
	50	167	97	238*
50	10	128	153	219*
	30	119	145	236*
	50	83	155	262*
75	10	92	180	228*
	30	84	158	258*
	50	63	180	257*
100	10	143	87	270*
	30	190	86	224*
	50	180	98	222*

Note: *maximum frequency of minimum mean squared errors.

From Tables. 2-3 and Fig. 1-2, it is obvious when the sample sizes are large ($n=100,150$) and variances of errors equal to 25 and proportions of missing values equal to 10 or 30, the pairwise deletion has result better than one obtained by EM algorithm and regression imputation. In contrast with this when variances of errors is greater than 25 and proportions of missing values is greater than 30, the EM still gives the best result .

Table 4. Descriptive statistics of mean squared errors of 500 simulated data sets in each method of missing estimations.

Sample size ($n = 20$)							
Variance of errors (σ_{ε}^2)	Proportions of missing values (%)	Methods of missing estimations	Min.	Max.	Mean	Median	Std. Deviation
25	10	P	0.0006	9201.503	685.9540	309.9084	1013.177
		R	0.0006	9021.081	673.0272	304.3563	993.3352
		EM	0.0006	9021.081	672.8611	304.3563	993.3005
	30	P	0.0126	6722.078	875.8084	406.7827	1176.799
		R	0.6519	6214.334	812.6339	379.3737	1087.762
		EM	0.5402	6210.403	811.7162	377.2202	1087.752
	50	P	0.0064	11387.75	860.1505	340.3438	1350.538
		R	1.5338	9381.899	713.7619	286.3381	1112.363
		EM	1.2866	9387.006	711.3840	278.8717	1111.183

Sample size ($n = 100$)

Variance of errors (σ_{ε}^2)	Proportions of missing values (%)	Methods of missing estimations	Min.	Max.	Mean	Median	Std. Deviation
75	10	P	0.0087	48932.36	5894.185	2501.322	8085.501
		R	5.5004	45894.24	5539.814	2357.784	7586.777
		EM	5.5099	45878.90	5539.399	2357.457	7586.141
	30	P	0.0194	52726.23	5388.364	2638.524	7205.619
		R	17.2341	41392.66	4292.118	2171.597	5704.060
		EM	17.1436	41299.69	4287.880	2168.625	5697.666
	50	P	0.2958	75510.20	5702.455	2675.918	8107.536
		R	42.7693	47820.44	3533.593	1601.649	4938.922
		EM	41.9155	47373.58	3518.011	1601.509	4960.734

In Table 4, we can see clearly that pairwise deletion method has minimum values (Min) smaller than one of other methods, but it yield the maximum values (Max) of mean squared errors bigger than one of other methods too. For means and standard deviations of mean squared errors we find that the EM algorithm and the regression imputation give almost the same results, which are lower than one of the pairwise deletion. This implies that filling in missing values by EM algorithm and regression imputation have better results than pairwise deletion.

Table 5. Descriptive statistics of mean squared of 500 simulated data sets in each method of missing estimations.

Sample size($n = 150$)

Variance of errors	Proportions of missing values (%)	Methods of missing estimations	Min.	Max.	Mean	Median	Std. Deviation
25	10	P	0.0004	5737.957	623.9068	294.1397	857.0383
		R	0.3908	5725.814	624.0165	294.4002	855.0483
		EM	0.3845	5725.920	624.4445	294.4787	855.0619
	30	P	0.0068	6755.548	704.2564	321.7841	995.5078
		R	1.2523	6699.218	705.1165	323.1399	986.7803
		EM	1.1008	6695.346	708.2918	322.4110	986.8200
	50	P	0.0032	5454.551	646.6533	336.7967	819.2805
		R	3.9471	5347.603	638.4606	334.3045	802.5258
		EM	2.7640	5337.920	636.4601	332.6712	802.3515

From Table 5, surprisingly ,we find when the size of a sample is large but variance of errors and proportions of missing values are small, pairwise deletion gives us better results.

4. Conclusion

Based on the results of Tables 1-5 and others, using the mean squared error (MSE) in regression, in case of small sample sizes($n = 20, 50, 70$) and all variances of errors in the simulation ($\sigma_{\varepsilon}^2 = 25, 50, 75, 100$) of missing values 10%, 30% and 50%, we found that EM algorithm yields better results than the regression imputation. On the other hand, pairwise deletion method is inferior than the others.

When sample sizes are large ($n=100,150$) and variances of errors are 50, 75, and 100 and the proportions of missing values are 10%, 30%, and 50%, EM algorithm still gives better results than regression imputation and pairwise deletion. However, when both the variance of error and the proportion of missing values are small, pairwise deletion gives better results than other methods. The pairwise deletion gives remarkable minimum of mean squared error when sample size is the large ($n=150$) and the proportions of missing values are small (10 %,30%) [4].

5. Appendix

To facilitate understanding we would like to give step details for EM algorithm

Step 1. From the observation table, the example data are

Y	X ₁	X ₂
y ₁	X ₁₁	X ₂₁
y ₂	X ₁₂	X ₂₂
y ₃ (miss)	X ₁₃	X ₂₃
y ₄	X ₁₄	X ₂₄
y ₅ (miss)	X ₁₅	X ₂₅

Y_(miss) is missing value of Y.

The Regression model is defined by

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \beta \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} + \varepsilon$$

where

$$Y_1 = \begin{bmatrix} y_1 \\ y_2 \\ y_4 \end{bmatrix}, \quad Y_2 = \begin{bmatrix} y_3^* \\ y_5^* \end{bmatrix},$$

$$X_1 = \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ x_{14} & x_{24} \end{bmatrix}, \quad X_2 = \begin{bmatrix} x_{13} & x_{23} \\ x_{15} & x_{25} \end{bmatrix}, \quad X = \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ x_{13} & x_{23} \\ x_{14} & x_{24} \\ x_{15} & x_{25} \end{bmatrix}.$$

Step 2. The regression coefficients have the initial value chosen by least square method from Y_1 and X_1 so that

$$\hat{\beta}^{(0)} = (X_1^T X_1)^{-1} \cdot (X_1^T Y_1).$$

Step 3. E step is calculated by

$$\begin{aligned} E(Y | X, Y_1, \hat{\beta}^{(0)}) &= \begin{cases} y_i & , \quad \text{if } y_i \text{ is observed} \\ X_2 \hat{\beta}^{(0)} & , \quad \text{if } y_i \text{ is missing} \end{cases} \\ &= Y^{(1)}. \end{aligned}$$

Step 4. M step in the first iteration is calculated by

$$\hat{\beta}^{(1)} = (X^T X)^{-1} \cdot (X^T Y^{(1)}).$$

Step 5. If difference between the regression coefficients ($\hat{\beta}^{(t)}$) and the next iterative coefficients ($\hat{\beta}^{(t+1)}$) is greater than 0.001, we return to step 3. Otherwise go to the next step.

Step 6. Compute $\hat{y} = X\hat{\beta}^{(t)}$ and $MSE = \frac{\sum (y_i - \hat{y})^2}{n - 3}$.

Acknowledgements

The authors would like to thank Prof. Dr. Prachoom Suwattee and Asst. Prof. Siriluck Jernjitpornchai for useful comments and valuable suggestions, which help us to improve the paper.

References

- [1] Atkinson, A.C. and Cheng, T.-C., *On robust linear regression with incomplete data*. Comput, 2000 ; 33 : 361-380 .
- [2] David C. Howell, Treatment of Missing Data. www.uvm.edu.
- [3] Little, R.J.A. and Rubin, D.B., *Statistical Analysis with Missing Data*. Wiley, New York, 1987.
- [4] Wilaiwan, N., Chanchai, C., Chancharoen, R., Sujitta, S. and Jaruchat, B., *A comparison of missing estimation in multiple linear regression with missing values in dependent variable*. Senior Project in Statistics, Mahasarakham University, Thailand, 2005.