Journal of Statistical Research *ISSN 0256 - 422 X*
2010, Vol. 44, No. 2, pp. 233-242
Bangladesh

USING DIAGNOSTIC MEASURES TO DETECT NON-MCAR PROCESSES IN LINEAR REGRESSION MODELS WITH MISSING COVARIATES

Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur, Kanpur-208016, India
Email: shalab@iitk.ac.in, shalabh1@yahoo.com

Helge Toutenburg[1]
Department of Statistics, University of Munich, Munich, Germany

Andreas Fieger
Service Barometer AG Gottfried-Keller-Str. 12, 81245 Mnchen, Germany

summary

This paper considers the problem of missing data in a linear regression model. It presents a method to analyze and detect the missing completely at random (MCAR) process when some values of covariates are missing but corresponding values of response variable are available. The idea of using outlier detection method in linear regression model is proposed to be employed to detect a non-MCAR processes. Such an idea is utilized and a graphical method is proposed to visualize the problem.

Keywords and phrases: Linear regression model, missing data, missing completely at random, diagnostic.

AMS Classification: 62J05

1 Introduction

A fundamental assumption in any regression analysis is that all the observations on response variable and covariates are available. In many applications, the observations on covariates may be missing due to one reason or the other. Under such a case, one simple approach is to use only the available observations and conduct the regression analysis. Another approach is to use the imputation techniques and impute the missing values. The imputation can be carried out using hot-deck imputation, cold deck imputation, mean imputation, regression imputation etc., see Rao et al. (2008, Chapter 8) for more details on imputation methods for incomplete covariate matrix. We consider here the case when observation are missing in covariates only. The imputation techniques

[1]This work started before Professor Toutenburg passed away in 2009.

depends mainly on the pattern of missingness in the observations, i.e., whether the observations on covariates follow Missing Completely At Random (MCAR) pattern or not. Heitjan and Basu (1996) have explained and reviewed the issues and concepts related to missing at random, observed at random and parameter distinctness. Nittner (2003) compares different methods of nonparametric estimation in an additive model when some observations on explanatory variables are missing at random but corresponding observations on response variables are available. Guobing and Copas (2004) provides further insight into the concept of missing at random and envisage two models with separable parameters: a model for the response of interest and a model for the missing data mechanism. It is shown that if the response model is given by a complete density family, then frequentist inference from the likelihood function ignoring the missing data mechanism is valid if and only if the missing data mechanism is missing at random. Carpenter et al. (2007) propose an imputation technique for multiple imputation case when missing values are not missing at random. Wang (2009) develops two approaches, viz., model calibration approach and weighting approach, to define the estimators of the parametric and nonparametric parts in the partial linear model with the covariates missing at random. Yang et al. (2009) considered the partial linear model with covariates missing at random and investigated the empirical likelihood ratios for the regression coefficients and baseline function. Sun et al. (2009) considered the model checking problem for a general linear model with response missing at random, see also Allison (2001), Scheuren (2005), Zhou et al. (2008) for more details on missing data and multiple imputation.

A fundamental assumption in all these work is that the data is missing at random. An important question arises at this stage is that how to decide whether the data on covariates is missing completely at random or not on the basis of sample data. Little (1988) proposed a single global test statistic for testing the whether the data is missing completely at random that uses all of the available data. The test reduces to a standard t test when the data are bivariate with missing data confined to a single variable. Little and Chen (1999) proposed a test for missing completely at random to decide whether or not the generalized estimating equations should be adjusted to correct the possible bias introduced by a missing-data mechanism that is not missing completely at random. Potthoff et al. (2006) clearly claims that there are no direct tests available to test whether the missing data mechanism is missing at random (MAR) or not and have proposed an alternate assumption, MAR+, that can be tested. MAR+ always implies MAR, so inability to reject MAR+ bodes well for MAR. In contrast, MAR implies MAR+ not universally, but under certain conditions. All such developments are related to analytical methods but graphical methods are not known to the best of our knowledge. We have utilized the set up of mixed regression estimation procedure along with the diagnostic tools for the detection of outliers in linear regression model to diagnose the missing pattern is MCAR or not through a graphical technique. The data may be missing in some or all the explanatory variables.

The plan of the paper is as follows. The model set up and estimators of regression coefficients are described in Section 2. The development of mixed regression estimator to deal with missing values is discussed in Section 3. The use of outlier detection tools for diagnosing MCAR pattern is discussed in Section 4. The graphical procedure to use the method and tools graphically is illustrated using a linear regression model in Section 5. Finally, some comments are placed in

Section 6

2 Model Setup and Estimators

Consider the classical linear regression model

$$y = X\beta + \epsilon$$

where y is a $n \times 1$ vector of response variable, X is a $n \times p$ matrix of n observations on each of the p covariates, β is a $p \times 1$ vector of associated regression coefficients and ϵ is a $n \times 1$ vector of random errors. Suppose some observations on covariates are missing corresponding to which observations on response variable y are available. Reorganizing the the n rows of the data matrix X with respect to missing observations and accordingly the corresponding observations in response y and error term ϵ leads to the following structure

$$\begin{pmatrix} y_{\mathrm{c}} \\ y_{\mathrm{mis}} \end{pmatrix} = \begin{pmatrix} X_{\mathrm{c}} \\ X_{\mathrm{mis}} \end{pmatrix} \beta + \begin{pmatrix} \epsilon_{\mathrm{c}} \\ \epsilon_{\mathrm{mis}} \end{pmatrix} \tag{2.1}$$

The index c indicates the completely observed submodel whereas the index mis corresponds to the submodel with missing values in the covariate matrix X_{mis} (note that y_{mis} is completely observed).

Writing the data as $Z_{ij} = (y_i, X_{ij})$, the missing data indicator matrix R introduced by Rubin (1976) is given by

$$R_{ij} = \begin{cases} 1 & \text{if } Z_{ij} \text{ is observed} \quad (i = 1, 2, \ldots, n; j = 1, 2, \ldots, p) \\ 0 & \text{if } Z_{ij} \text{ is missing} \end{cases}$$

The missing data mechanism can then be characterized by the conditional distribution $f(R|Z, \phi)$ of R given the data Z and an unknown parameter ϕ. The $n \times (p+1)$ matrix Z consists of observed data Z_{obs} and unobserved values Z_{mis}.

The data is said to be missing completely at random (MCAR) if the distribution of R given Z and ϕ depends only on the unknown parameter ϕ for any Z, i.e.,

$$f(R|Z, \phi) = f(R|\phi) \quad \forall Z\,.$$

If the conditional distribution of R depends only on Z via the observed values Z_{obs} for all Z_{mis}, i.e.,

$$f(R|Z, \phi) = f(R|Z_{\mathrm{obs}}, \phi) \quad \forall Z_{\mathrm{mis}}\,,$$

then the data is called missing at random (MAR).

The optimal estimator in this case is the Gauss-Markov estimator b of β that is obtained by applying the principle of least squares to the data in (2.1):

$$\begin{aligned} b &= \left(\begin{pmatrix} X_{\mathrm{c}} \\ X_{\mathrm{mis}} \end{pmatrix}' \begin{pmatrix} X_{\mathrm{c}} \\ X_{\mathrm{mis}} \end{pmatrix} \right)^{-1} \begin{pmatrix} X_{\mathrm{c}} \\ X_{\mathrm{mis}} \end{pmatrix}' \begin{pmatrix} y_{\mathrm{c}} \\ y_{\mathrm{mis}} \end{pmatrix} \\ &= (X_{\mathrm{c}}' X_{\mathrm{c}} + X_{\mathrm{mis}}' X_{\mathrm{mis}})^{-1} (X_{\mathrm{c}}' y_{\mathrm{c}} + X_{\mathrm{mis}}' y_{\mathrm{mis}})\,. \end{aligned} \tag{2.2}$$

This estimator can not be used directly due to the involvement of unknown values in X_{mis}. There are various methods which deal with this problem.

3 Dealing with Missing Values

A simple and commonly used method to deal with missing data is to discard all the information available in $(y_{\text{mis}}, X_{\text{mis}})$ and to use the completely observed data in $(y_{\text{c}}, X_{\text{c}})$ only. This gives the least squares estimator of β as

$$b_{\text{c}} = (X_{\text{c}}'X_{\text{c}})^{-1}X_{\text{c}}'y_{\text{c}}\,.$$

Obviously, some of the available information in terms of observations on response variables corresponding to which the observations on covariates are missing is being discarded in this case which is not a good strategy.

The maximum likelihood procedures address the missing data problem by factorizing the joint distribution as

$$f(Z, R|\theta, \xi) = f(Z|\theta)f(R|Z, \xi)\,.$$

Integration over the missing data Z_{mis} yields

$$f(Z_{\text{obs}}, R|\theta, \xi) = \int f(Z, R|\theta, \xi)\, dZ_{\text{mis}} = \int f(R|Z, \xi)f(Z|\theta)\, dZ_{\text{mis}}\,.$$

If $f(R|Z)$ depends only on the observed data Z_{obs}, i.e., the MAR assumption holds, then we have

$$f(Z_{\text{obs}}, R|\theta, \xi) = f(R|Z_{\text{obs}}, \xi)\int f(Z|\theta)\, dZ_{\text{mis}} = f(R|Z_{\text{obs}}, \xi)f(Z_{\text{obs}}|\theta)\,,$$

and that is why the missing data mechanism in this case is also called ignorable.

The imputation procedures present a different approach to the problem. The missing values in X_{mis} are replaced by the values that are generated by some imputation procedure, say X_{R}. Now the estimator (2.2) becomes operational. Various imputation methods are proposed in literature to find the missing values X_{R} to replace the unknown values in X_{mis}. For example, mean imputation or zero order regression (ZOR) replaces an unknown value x_{ij} by the mean $\bar{x}_j$, either formed of the complete cases in X_{c} or the available cases in X_{c} and X_{mis}.

Conditional mean imputation or first order regression (FOR) uses auxiliary regressions to find replacements for the missing values. Regressing the covariate with missing values on the remaining covariates (with parameters estimates based on the complete cases) yields predictions of the missing values that are used as substitutes. If the response y is also used in these regressions then a stochastic element is introduced (see Buck (1960), or Toutenburg and Shalabh (2002).

Multiple imputation repeats the imputation step and averages the results, see Rubin (1987), Schafer (1997). While a single imputation is too smooth, the differences between the individual imputation steps can be properly used to estimate the variance as the sum of the average variance within the imputed data sets and the between imputation variance. This strategy reflects the uncertainty about the imputation process which is ignored in a single imputation strategy.

By replacing a missing value by x_R in (2.2), the model becomes a mixed model as

$$\begin{pmatrix} y_{\mathrm{c}} \\ y_{\mathrm{R}} \end{pmatrix} = \begin{pmatrix} X_{\mathrm{c}} \\ X_{\mathrm{R}} \end{pmatrix} \beta + \begin{pmatrix} 0 \\ \delta \end{pmatrix} + \begin{pmatrix} \epsilon_{\mathrm{c}} \\ \epsilon_{\mathrm{R}} \end{pmatrix},$$

where δ is the difference between the true but unknown values in X_{mis} and their replacements by X_R. Using the mixed estimator due to Theil and Goldberger (1961), see also Rao et al. (2008), we have

$$\begin{aligned} b &= \left(\begin{pmatrix} X_{\mathrm{c}} \\ X_{\mathrm{R}} \end{pmatrix}' \begin{pmatrix} X_{\mathrm{c}} \\ X_{\mathrm{R}} \end{pmatrix} \right)^{-1} \begin{pmatrix} X_{\mathrm{c}} \\ X_{\mathrm{R}} \end{pmatrix}' \begin{pmatrix} y_{\mathrm{c}} \\ y_{*} \end{pmatrix} \\ &= (X_{\mathrm{c}}'X_{\mathrm{c}} + X_{\mathrm{R}}'X_{\mathrm{R}})^{-1}(X_{\mathrm{c}}'y_{\mathrm{c}} + X_{\mathrm{R}}'y_{*}), \end{aligned}$$

The weighted mixed estimator introduced by Schaffrin and Toutenburg (1990) uses a weight $\lambda < 1$ for the values in (y_{mis}, X_R) and is given by

$$b(\lambda) = (X_{\mathrm{c}}'X_{\mathrm{c}} + \lambda X_{\mathrm{R}}'X_{\mathrm{R}})^{-1}(X_{\mathrm{c}}'y_{\mathrm{c}} + \lambda X_{\mathrm{R}}'y_{\mathrm{R}}). \tag{3.1}$$

This estimator may be interpreted in terms of popular mixed estimator in the model

$$\begin{pmatrix} y_{\mathrm{c}} \\ \sqrt{\lambda} y_{*} \end{pmatrix} = \begin{pmatrix} X_{\mathrm{c}} \\ \sqrt{\lambda} X_{\mathrm{R}} \end{pmatrix} \beta + \begin{pmatrix} \epsilon_{\mathrm{c}} \\ \sqrt{\lambda}\phi \end{pmatrix}.$$

We will use the weighted mixed regression estimator (3.1) in the graphical procedures for the diagnosis of the missing mechanism which is described in the next section 4.

4 MCAR Diagnosis with Outlier Measures

Popular diagnostics measures to detect non-MCAR processes consist of the comparison of correlation or covariance matrices, the comparison of means ($\bar{y}_{\mathrm{c}}$ vs. $\bar{y}_{\mathrm{mis}}$) or a more general test as described by Little (1988). For the situations with only one column affected by missing values, Simon and Simonoff (1986) present diagnostic plots where 'envelopes' are compared.

The idea first presented by Simonoff (1988) combines the missing data problem with statistics that is derived from the outlier detection field. A comparison of the values of a statistic computed with and without imputation is the comparison of the sub-samples Z_{c} and Z_{mis}.

If the imputation of values can be considered appropriate under MCAR and if the observations are really MCAR (which is the null hypothesis H_0), then the statistics should be 'more or less' the same. If we have something other than MCAR, the statistics should reflect this departure by having different values.

Simonoff (1988) uses the Cook's distance, which is based on the confidence ellipsoid

$$C = \frac{(\hat{\beta}_* - \hat{\beta}_{\mathrm{c}})'(X_*'X_*)(\hat{\beta}_* - \hat{\beta}_{\mathrm{c}})}{p s_*^2},$$

the residual sum of squares *DRSS* due to Andrews and Pregibon (1978) is

$$DRSS = \frac{(RSS_* - RSS_c)/n_{mis}}{RSS_c/(n_c - n_{mis} - p + 1)},$$

and the determinant of the $X'X$ matrix (denoted by DXX) is

$$DXX = \frac{|X_c'X_c|}{|X_*'X_*|},$$

see Andrews and Pregibon (1978).

The distribution of the measures under H_0 is needed for the construction of tests. As this distribution also depends on the X values, Monte-Carlo methods are used to determine it by computing the complete case statistics first and then imputing the missing values under MCAR-assumption. The generation of new response values

$$y_{mis}^{MC} = \hat{X}_{mis}\hat{\beta}_c + \epsilon^{MC}$$

with $\epsilon^{MC} \sim N(0, s^2 I)$ generates a new data set where the 'missing values' are drawn from a model using a MCAR mechanism.

The diagnostic measures are computed after applying the imputation procedure to these data. Repeating the 'data deletion' and imputation steps, a null distribution of the diagnostic measure is generated. Finally the measure can be applied to the imputed original data and the resulting values can be compared with the null distribution.

5 Graphical Diagnosis of the Missing Mechanism

Various methods exist for the estimation of parameters that adjust for the missing data if the mechanism is ignorable, i.e., if MAR holds. If in addition, the missing data mechanism itself is of interest then the procedure described in the following section may give insight in the structure of missingness.

Animated residual plots are presented in Cook and Weisberg (1989). In a stepwise procedure, the weights between 0 and 1 are used to include one case into the regression. The plots thus represent the influence of that single case. Park, Kim and Toutenburg (1992) present a similar approach to visualize the inclusion of another variable in the regression model.

The adaption for a situation with missing data shows a close relationship to the procedures of the preceding section is described in Fieger (2000). Like in the above procedures, imputation is performed under an MCAR assumption. Having filled the gaps in X_{mis}, the weighted mixed estimator (3.1) is computed for certain values $\lambda \in [0, 1]$. Again the idea is that if we really have MCAR, then there should not be any tendency in the residual plots, when including Z_R in the model stepwise by increasing the weight from 0 to 1.

Figure 1 shows a small program that visualizes the following procedure:

```
for (λ = 0; λ ≤ 1; λ+=step) {
    compute regression parameters;
    compute estimated residuals;
    display residual-plot;
}
```

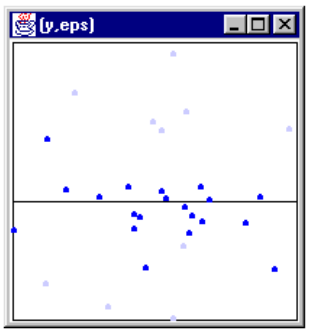

Figure 1: Java program for visualization on computer screen. Reads data for each frame of the animation and draws the single plots of the animation.

The residual plot in figure 2 shows an example of an animated plot of $\hat{y}$ (on X-axis) versus $\hat{\epsilon}$ (on Y-axis) for a model $y = \beta_0 + \beta_1 x_i + \beta_2 x_2 + \epsilon$ with missing data generated by a non-MAR process where $P(R_{i2} = 0)$ (a value x_{i2} is missing) depends on x_2.

An increasing value of λ gives higher weight to the imputed data in the estimation of the regression parameters. The center of the residual plot in figure 2 then shifts towards the origin. For $\lambda = 1$ the imputed data have the same weight as the complete data and biased estimated result. On the other hand, $\lambda = 0$ (the complete case estimator) gives consistent estimates as the missing process is independent of the response y.

6 Conclusions

We have proposed here a possible diagnostic graphical measure to check if the missing pattern in the data is MCAR or not. The diagnostic tool is derived by studying the relationship with the diagnostic measure for outlier detection.

The ideas of animated residual plots could be extended in various ways. Imagine a simultaneous plot of $\hat{\epsilon}$ vs. $\hat{y}$ vs. X_i in different windows where the windows are linked. By brushing, selected points of the plot could be highlighted in all the windows and their location or movement can be studied by changing the value of λ.

Univariate plots of y vs. X_j for all j together with the estimated regression line $\hat{\beta}_0 + \hat{\beta}_i X_i$, where the points are static (as the imputation does not depend on the weight λ) and the estimated regression line is dynamic. Again, these plots could be linked as described in Section 5.

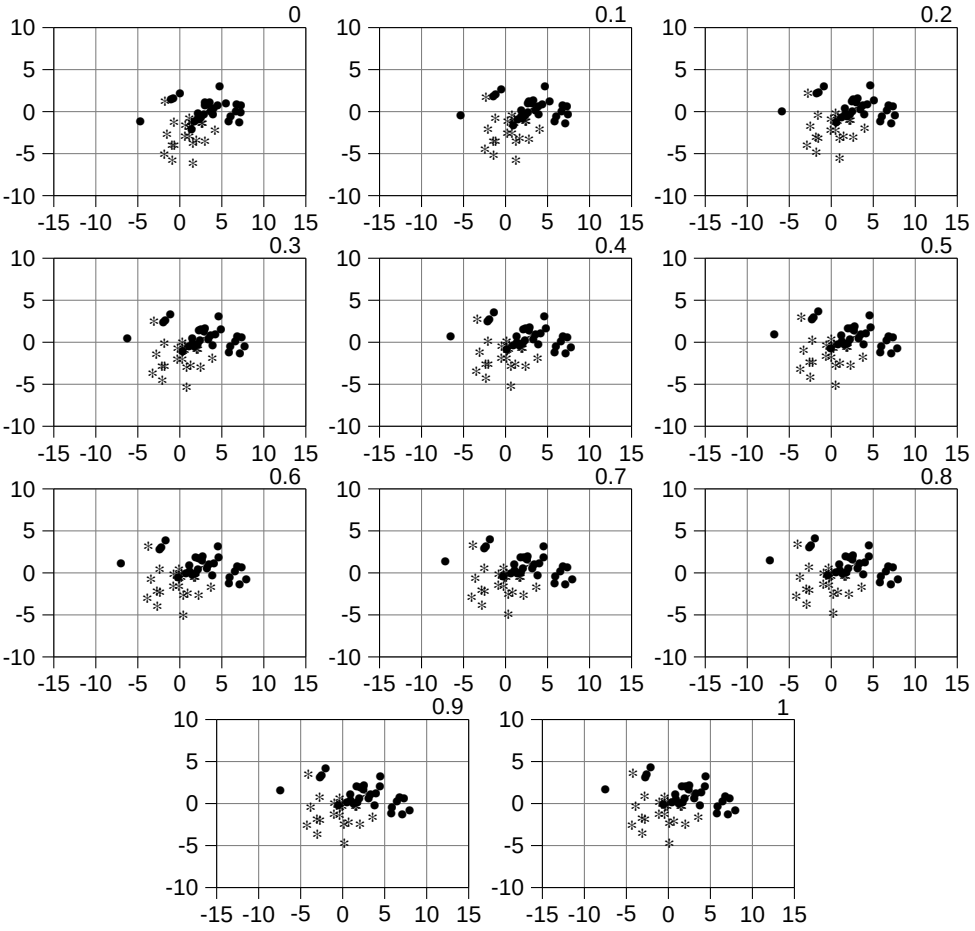

Figure 2: Residual plot of $\hat{y}$ (X-axis) versus $\hat{\epsilon}$ (Y-axis) for a model $y = \beta_0 + \beta_1 x_i + \beta_2 x_2 + \epsilon$ with a non-MAR process where $P(R_{i2} = 0)$ (a value x_{i2} is missing) depends on x_2. Value of λ from 0 (top left) to 1 (bottom right).

Create a null plot where missing values are created artificially by a known MCAR mechanism. This plot can be used as a means of comparison in order to have an idea of what the plot should look like under MCAR.

Acknowledgement

The authors are grateful to the referee for the comments which improved the exposition of the paper.

References

[1] Allison, Paul D. (2001): Missing Data. Sage Publication.

[2] Andrews, D.F. and Pregibon, D. (1978). Finding outliers that matter, *Journal of the Royal Statistical Society, Series B.* **40**, *85-93.*

[3] Buck. S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer, *Journal of the Royal Statistical Society, Series B.* **22**, *302-307.*

[4] Carpenter, James R., Kenward, Michael G., and White, Ian R. (2007): Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research*, 16, **3**, *259-275.*

[5] Chen, Hua Yun and Little, Roderick (1999): A test of missing completely at random for generalised estimating equations with missing data. *Biometrika*, **86**, 1, *1-13.*

[6] Cook, R.D. and Weisberg, S. (1989). Regression diagnostics with dynamic graphics, *Technometrics.* **31**, *277-291.*

[7] Fieger, A. (2000). Fehlende Kovariablenwerte bei Linearen Regressionmodellen, Ph.D. thesis, University of Munich, Munich, Germany.

[8] Heitjan, Daniel F. and Basu, Srabashi (1996): Distinguishing "missing at random" and "missing completely at random". *American Statistician*, **50**, 3, *207-213.*

[9] Little, R.J.A. (1988). A test of missing completely at random for multivariate data with missing values, *Journal of the American Statistical Association*, **83**, 404, *1198-1202.*

[10] Lu, Guobing and Copas, John B. (2004): Missing at random, likelihood ignorability and model completeness. *Annals of Statistics*, **32**, 2, *754-765.*

[11] Nittner, Thomas (2003): Missing at random (MAR) in nonparametric regression—a simulation experiment. *Statistical Methods and Applications*, **12**, 2, *195-210.*

[12] Park, S.H. Kim, Y.H. and Toutenburg, H. (1992). Regression diagnostics for removing an observation with animating graphics, *Statistical Papers*, **33**, *227-240.*

[13] Potthoff, Richard F., Tudor, Gail E., Pieper, Karen S. and Hasselblad, Vic (2006): Can one assess whether missing data are missing at random in medical studies? *Statistical Methods in Medical Research*, **15**, 3, *213-234.*

[14] Rao, C.R., Toutenburg, H., Shalabh, and Heumann, C. (2008). Linear Models and Generalizations - Least Squares and Alternatives. Springer.

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

[16] Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Sample Surveys. Wiley.

[17] Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. Chapman and Hall.

[18] Schaffrin, B. and Toutenburg, H. (1990). Weighted mixed regression, *Zeitschrift für Angewandte Mathematik and Mechanik*. **70**, *735-738*.

[19] Scheuren, Fritz (2005): Multiple imputation: how it began and continues. *American Statistian*, **59**, 4, *315-319*.

[20] Simon, G.A. and Simpnoff, J.S. (1986). Diagnostic plots for missing data in least squares regression, *Journal of the American Statistical Association*, **81**, *501-509*.

[21] Simonoff, J.S. (1988). Regression diagnostics to detect nonrandom missingness in linear regression, *Technometrics*. **30**, *205-214*.

[22] Sun, Zhihua and Wang, Qihua (2009): Checking the adequacy of a general linear model with responses missing at random. *Journal of Statistical Planning and Inference*, **139**, 10, *3588-3604*.

[23] Theil, H. and Goldberger. A.S. (1961). On pure and mixed estimation in econometrics, *International Economic Review*. **2**, *65-78*.

[24] Toutenburg, H. and Shalabh (2002). Prediction of response values in linear regression models from replicated experiments, *Statistical Papers*, **43**, *423-433*.

[25] Wang, Qi-Hua (2009): Statistical estimation in partial linear models with covariate data missing at random. *Annals Institute of Statistical Mathematics*, **61**, 1, *47-84*.

[26] Yang, Yiping, Xue, Liugen, and Cheng, Weihu (2009). Empirical likelihood for a partially linear model with covariate data missing at random. *Journal of Statistical Planning and Inference*, 139, **12**, *4143-4153*.

[27] Zhou, Y., Wan, A.T.K. and Wang, X.J. (2008). Estimating equations inference with missing data. *Journal of the American Statistical Association*, **103**, *1187-1199*.