Handling Incomplete Data with Regression Imputation

S Annas1*, P Kartikasari2, R Arisandi1

1 Department of Statistics, UniversitasNegeri Makassar, Makassar, South Sulawesi, 90222, Indonesia.
2 Department of Statistics, FSM Universitas Diponegoro, Semarang, Center Java, 50275, Indonesia.

*e-mail: suwardi_annas@unm.ac.id

Abstract. Regression analysis is widely used in various fields due to easy-to-understand. One of purposes of regression analysis is to predict the response variable using the predictor variables. Unfortunately, in real cases, some values may be missing. This circumstance will produce large error, indeed, poor prediction. Missing value lead us to trade-off remove the paired data points or replace. The purpose of this study was to estimate the missing value with regression imputation. This study conducted two scenarios of amount of missing values, 10% and 15%. The study results showed that the higher of amount of missing values, the higher the value of MSE was.

Keyword: Handling incomplete data, regression imputation

1. Introduction
Regression is an analysis widely used in various fields because of easy-to-use. The purpose of regression analysis is to find out the functional relationship between the dependent variable and the independent variables [1]. If the functional relationship is known or there is prior information then estimation of regression coefficient will be conducted by parametric regression method. The relationship between the dependent variable and the independent variable, in the model, can occur with linear functions in parameters or nonlinear [2]. In estimation of regression coefficients, information about the functional relationship between the variables must satisfy otherwise require assumptions in the analysis to get the regression model. Linear regression involves some assumptions i.e. linearity, normally distributed and homogenous error [3]. A condition that must be satisfied in the regression analysis is that the data used must be paired between values of dependent variable and independent variable [4]. [5], [6], [7].

Unfortunately, in some cases like experiment and survey cases, missing value could be encountered. The validity of data set can be identified by the missingness on the data set. The missingness will produce a large error and indeed poor prediction. The missing values lead to trade-off i.e. remove the paired point on data set or replace them [8], [9]. Remove paired data points will lose some information, so that the way to keep the information is replacing the missing values. The method that can be used to replace them is imputation. Imputation is a process used to determine and assign replacement values to resolve problems of missing, invalid or inconsistent data [10].

Some of basic statistical methods to replace missing values are item mean substitution, person mean substitution, regression imputation, and hot deck imputation. The advanced method are Expectation maximization, full information maximum likelihood and multiple imputation. Regression imputation is
a better approach to predict missing value. [11]. Therefore, this study applied regression imputation for countering the missing value. The estimation of regression equation obtained from simulation using generated data set and used it to calculate the missing value. The last step was comparison between imputation result and complete data in regression analysis.

2. Data and Method

2.1. Data
The data used were generated data. The number of observation was 100. This data set would be used in simulation scenarios of missing value. The amount of missing value was 10% as the first scenario. Other scenario of amount of missing values, the proposed scenarios as comparison, was 15%.

2.2. Method
The partial linear regression model is given as

\[ Y = X^T \beta + \nu(Z) + \varepsilon, \]

where \( Y \) is dependent variable, \( X \) is transpose of independent variable, \( \beta \) is unknown parameter, and \( f(Z) \) is a known function of the shape, then the method of research related to the research objectives were as follows:

- Simulate a function from \( f(Z) \) and parameters \( \beta \) in partial linear models for data whose parametric components are missing.
- Apply the method obtained by the simulation data. The steps were as follows:
  - Establish a simulation model. Generate distributed error \( N(0, \sigma^2) \), with \( n = 100, \sigma^2 = 0.01 \) and \( \beta = 0.375 \).
  - Eliminate data randomly on parametric components.
  - Compare between the results of complete data and incomplete data regression based on R-square and mean square error (MSE).

3. Data and Method

3.1. Estimation of the regression function
The missing value on the parametric component in the regression analysis, it is necessary to estimate the regression function. For example in the regression model \( Y = X^T \beta + \nu(Z) + \varepsilon \). There are missing value on the components of parametric define \( \delta = 1 \) for \( X \) as fixed and \( \delta = 0 \) for \( X \) as random [12]. It is assumed that random data is missing at variable. The first step in imputation was to estimate i.e. determine the weight. Furthermore, to obtain the estimator by using weighted least square method as follows

\[
\sum_{i=1}^{n} W_i \left\{ Y_i^* - \alpha_0 - \alpha_1(D_i) \right\}^2
\]

or

\[
f = (Y - Z \alpha)^T W (Y - Z \alpha)
\]

\[
Y = (Y_1^*, Y_2^*, ..., Y_n^*)^T, \quad \alpha = (\alpha_0, \alpha_1)^T,
\]

\[
Z = \begin{bmatrix}
1 & D_1 \\
1 & D_2 \\
\vdots & \vdots \\
1 & D_n
\end{bmatrix}
\]
Furthermore, Equation (2) is derived from $\alpha$ to obtain the estimator

$$\hat{\alpha} = (Z^T WZ)^{-1} Z^T WY$$

where $\hat{\alpha}$ is the minimum point of a positive definite function.

Estimator for the curve $v(Z)$ in the case of missing value parametric components, it was obtained as follows:

$$v(Z) = \frac{1}{A} \sum_{i=1}^{n} \left( \hat{\delta}_i (Z_i; h) - \hat{\delta}_i (Z_i; h)(Z_i - Z) \right) K_h (Z_i - Z) \frac{\delta_i}{\pi(Y_i, Z_i)} (Y_i - x_i^T \beta)$$

(3)

or

$$\hat{v}(Z) = P + Q = H_{\delta}Y + K_{\delta} \beta$$

(4)

Furthermore, the model was obtained:

$$Y = X^T \beta + \hat{v}(Z) + \varepsilon,$$

(5)

So for the regression estimator in the case of missing value on the parametric component was estimated by using

$$\hat{\beta} = \left[ X (I + H_{\delta})^T (I + H_{\delta}) X^T \right]^{-1} X (I + H_{\delta})^T (I - H_{\delta}) Y$$

(6)

3.2. Simulation

Suppose that a simple linear regression function model $Y = X^T \beta + v(Z) + \varepsilon$ with error that normally distributed. Generated 100 samples where $\beta = 0.375$ and suppose that the conditions of missing value on the independent variable were 10%, and 15%. Then an indicator $\delta_i$ was assigned to mark the missing value and be weighted. The plot of generated data set as follows.

**Figure 1.** Plot of relationship of generated data set
The generated data set had positive relationship between dependent and independent variables in the other words was direct proportion. The strength of the relationship was 0.375. The data set was used in two built scenarios of percentage of missing values. The estimation of those scenarios discussed on next section.

3.3. Estimation of missing value
This section presented the estimation of missing values where the first scenario including 10% missing value where $\sigma^2$ was 0.01. This properties yielded $\hat{\beta} = 3.076$. The regression line of this scenario tent to be over the regression line of complete data. There was a quite difference of their slopes as depicted on Figure 1. The second scenario included 15% missing value where $\sigma^2$ was 0.01. This properties yielded $\hat{\beta} = 3.351$. In line with the first scenario, the regression line of the second scenario tent to be over the regression line of complete data.

The difference of the first and the second scenarios was the difference of distance of their regression lines to the regression line of complete data. The first scenario had little closer to the complete data as depicted on Figure 2. Contrary, the second scenario had been far from the complete data as depicted Figure 3.

![Figure 2. Plot of the estimated 10% missing value](image1)

![Figure 3. Plot of the estimated 15% missing value](image2)
Table 1. Performance of imputation regression on complete and incomplete data

| Data                  | $R^2$  | MSE   |
|-----------------------|--------|-------|
| Complete Data         | 84.42  | 21.737|
| 10% Missing value     | 78.02  | 24.090|
| 15% Missing value     | 73.13  | 27.388|

Table 1 showed that the prediction performance using complete data and both of scenarios. The percentage of missing value was greater, then the value of the MSE would be even greater. On the other hand, the percentage of missing value was greater, the coefficient of determination ($R^2$) would be smaller. The amount of missing values determined MSE and $R^2$. Regression imputation was better on first scenario than the second was. This was in line with [3] stated that imputation given 10% of the amount of missing values can use regression to predict the missing values. Meanwhile, the use of regression imputation given 15% of the amount of missing values had low performance.

4. Conclusion

Based on the simulation, it obtained the estimator used in the regression model was

$$\hat{\beta} = \left(X(I + H_\delta)^\top(I + H_\delta)X^{-1}\right)^{-1}X(I + H_\delta)^\top(I - H_\delta)Y.$$

Using imputation regression on the first scenario was closer to the complete data than the second scenario. The estimator tent to obtain small coefficient of determination as the increasing of amount of missing values. In line with that, the MSE value would increase as the amount of missing values.

References

[1] Hastie, T., Tibshirani, R., Friedman, J. (2008). *The elements of statistical learning: data mining, inference, and prediction 2nd edition*. California: Springer.

[2] Draper, N., Smith, H. (1998). *Applied Regression Analysis*. New York: John Wiley & Sons.

[3] Seber, G. A. F., Lee, A. J. (2003). *Linear Regression Analysis*. New Jersey: John & Sons.

[4] Liang, H., Wang, S., Robins M, and Carroll R, J. (2004) Estimation in Partially Linear Models with Missing Covariates. *Journal of the American Statistical Association. Vol 99*: 357-367.

[5] Robins, J. M., Mark, S. D., and Newey, W. K. (1992). Estimating Exposure Effect by Modelling the Expectation of Exposure Condition on Confounders. *Biometrics. Vol 48*: 479-495.

[6] Little, R. A., Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Canada: A Wiley-Interscience Publication.

[7] Hardle, G. (1990). *Applied Nonparametric Regression*. New York: Cambridge University Press.

[8] Enders, C. K. (2011). Analyzing Longitudinal Data with Missing Values. *Rehabilitation Psychology. Vol 56(4)*: 267–288.

[9] Horton, N. J., Kleinman, K. P. (2007). Much do about Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. *Am Stat. Vol 61(1)*: 79–90

[10] Statistics Canada. (2010). *Survey Methods and Practices*. Ottawa: Minister of Industry.

[11] Yu, G., Zhu, L., Sun, Jand, Robinson, L., 2018 Regression analysis of incomplete data from event history studies with the proportional rate model *Stat Inference Vol 11(1)*: 91-97.

[12] Robins, J. M., Rotnitzky, A., and Zhao, L, P., (1994). Estimation of Regression Coefficients When Some Regressor are Not Always Observed. *Journal of the American Statistical Association. Vol 89*: 846-866.