Proceedings of the Autumn Conference in 2001
and the 30th Anniversary of the Korean Statistical Society

Imputation Procedures in Weibull Regression Analysis in the presence of missing values *

Soon-kwi Kim[1] and Dong-Bin Jeong[2]

Abstract

A dataset having missing observations is often completed by using imputed values. In this paper the performances and accuracy of complete case methods and four imputation procedures are evaluated when missing values exist only on the response variables in the Weibull regression model. Our simulation results show that compared to other imputation procedures, in particular, hotdeck and Weibull regression imputation procedure can be well used to compensate for missing data. In addition an illustrative real data is given.

Keywords: MCAR; Imputation procedure; Hot deck; Weibull regression imputation; Weibull regression model

1. Introduction

The problem of missing values (incomplete data) arises frequently in many data sets and appears in particular common in practical situations such as the medical and social sciences. Incomplete data consist of two types – missing units and missing items. Missing units are the results of nonresponse for a sample unit and thus, consist of refusals and inaccessibility. This type of nonresponse is also called unit nonresponse. Missing items refer to those units that have missing values but also have some collected values. This type is also called item nonresponse. When analyzing data with missing values, it is common practice either to eliminate all units with missing data or to use other information to replace the missing data. Procedures using other information to replace missing values are referred to as imputation procedures. Imputation methods

* The first author was partially supported by the Kangnung National University research fund in 2001.

[1] Department of Statistics, Kangnung National University, Kangnung, 210-702, Korea.
[2] Department of Statistics, Kangnung National University, Kangnung, 210-702, Korea.

have been well accepted and widely exploited over the years both in major surveys as well as in small surveys in that they may provide less biased estimates of parameters and result in a less concomitant loss in precision comparing to discarding all units which have missing values on any variables used in the particular analysis. Namely once the missing values are imputed, methods of analysis that require complete data on all variables are then used to analyze the data.

Efron (1994) worked nonparametric bootstrap approaches to assessing the accuracy of an estimator in a missing data situation and found the simplest form of confidence interval that provides convenient and accurate clue. Bello (1995) examined several imputation procedures (the mean substitution method, EM algorithm, principal component method, general iterative principal component method and singular value decomposition method) and investigated their comparative performances. Hegamin-Younger and Forsyth (1998) compared the effectiveness of four imputation procedures (mean, conditional mean, hot deck and regression) in a two-variable regression by including 18,869 participants in the sample.

Let y_i be the true (but possibly missing) value of a variable Y, for an individual i, and let $m_i=1$ if y_i is missing and 0 if it is observed. Let $x_1, \ldots, x_k$ be a set of variables that are observed. Then the mechanism of missingness for the variable Y is called

(1) missing completely at random(MCAR) if the following is true:

$$P\{m_i = 1 | y_i, x_{1i}, \ldots, x_{ki}\} = P\{m_i = 1\}$$

(2) missing at random(MAR)

$$\text{if } P\{m_i = 1 | y_i, x_{1i}, \ldots, x_{ki}\} = P\{m_i = 1 | x_{1i}, \ldots, x_{ki}\} \text{ and}$$

(3) non-ignorable

$$\text{if } P\{m_i = 1 | y_i, x_{1i}, \ldots, x_{ki}\} = P\{m_i = 1 | y_i, x_{1i}, \ldots, x_{ki}\} \text{ or}$$

$$P\{m_i = 1 | y_i, x_{1i}, \ldots, x_{ki}\} = P\{m_i = 1 | y_i\}$$ (Little and Rubin, 1987).

The primary purpose of this paper is to examine the behavior of and to investigate the accuracy of complete case methods and four different imputation procedures on the estimates of the Weibull regression coefficients. Four imputation procedures we examine are the grand mean procedure (GM), conditional mean procedure (CM), hot-deck (HD), and Weibull regression imputation procedure (WI). In section 2 the Weibull regression model is formulated and section 3 briefly describes some imputation procedures. Monte Carlo design and results are presented in section 4 and 5, respectively. Finally, an application to an example dataset and conclusions are given in section 6 and 7, respectively.

2. The Weibull Regression Model

Let T_i denote the failure time of the ith observation, then the hazard rate of the multiplicative or proportional hazard model for the Weibull distribution function is given by

$$h(t) = \lambda r(\lambda t)^{r-1} \exp\{\beta_1' x_1 + \Lambda + \beta_p' x_p\}, \quad t \geq 0 \; and \; r, \lambda > 0$$

where r is a shape parameter, λ is a scale parameter, respectively and $x_1, \ldots, x_p$ is a set of covariates. If we let $Y = \ln T$, then Y has an extreme-value distribution. The accelerated failure-time (AFT) model for the same distribution function is also given by

$$T = \exp\{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\} \; \varepsilon^{\sigma},$$

where σ is a scale parameter, and ε has an exponential distribution with parameter 1. Then we can see

$$r = 1/\sigma$$
$$\beta_i = -\beta_i' \times \sigma \quad \text{for } i=1,2,\ldots,p$$
$$\lambda = \exp\{-\beta_0\}.$$

Let X denote an $(n{\times}p)$ data matrix of n individuals on p variables and suppose $n>p$. Suppose that some individuals do not have complete variables and that the missing variables are missing at random as described in Rubin (1976). We assume that the missing values in the data, X, are estimated exploiting one of the imputation procedures. Let the first column of X be the response variable denoted by Y and the remaining p-$1=q$ columns be the covariates also denoted by X_2.

3. Some Imputation Procedures

In this section we present the brief description of complete case methods and four imputation procedures that impute certain values for the missing values only the response variable.

(1) Complete Case Methods

These methods discard all units that have missing values on any variable used in the particular analysis. This is one of the most popular methods of dealing with incomplete analysis and permits immediate use of statistical methods that require complete data.

(2) Some Imputation-Based Methods

The missing values are filled in and the resultant completed data are analyzed by standard methods. Imputation based methods have been well accepted and widely used over the years both in major surveys and in smaller surveys.

(a) Grand Mean Imputation

The grand mean imputation procedure (GM) is the most common and the simplest imputation procedure. This procedure involves replacing missing values on a particular variable by the-mean value of the observed data on it. Namely means from sets of all recorded values are substituted.

(b) Conditional Mean Imputation

The conditional imputation uses collateral information to provide an estimate for the missing value. This procedure partitions the sample into homogenous groups based on responses to collateral information.

(c) Hot Deck Method

An individual with a missing value has imputed for his missing Y value the value of another individual sampled in the survey whose Y value was nonmissing. There are many ways to select the "donor" individuals for a hot-deck method.

In the hot deck method, cells are defined on the basis of variables that are considered important for imputation. These are generally variables that relate to the particular sample design used or to demographic or other variables. The data file is then sorted first according to these defined cells and secondly by other variables that are considered relevant for imputation. For each cell, a register is defined as the record of an individual on whom all variables are recorded. In a single pass through the data file, the cell of each record is identified and, if a variable is missing, the value for the missing value for that cell's register is substituted for the missing value.

On the other hand, if the individual's record is complete, then the values of the variables for this record replace the values in the registry for that cell. This process is repeated until all missing values are imputed.

(d) Weibull Regression Imputation

In this method regression equations are fitted from a data file consisting of complete records with the variable to be imputed being as the dependent variable. The fitted line may be of the form

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_k x_k,$$

where y is the response variable to be imputed for a given record and $x_1, x_2, \ldots, x_k$ are covariates known for the individual. Mean imputation can be regarded as a special case of regression imputation where the predictor variables are dummy indicator variables for the cells within which the means are imputed.

Weibull regression imputation is proposed similarly and can be discussed as done in regression imputation. The AFT models are fitted from a data set consisting of complete individuals with the variable to be imputed serving as the dependent variable. The fitted Weibull regression line may be of the form

$$T = \exp(b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k),$$

where T is the survival time to be imputed for a given record and x_1, x_2, ..., x_k are covariates known for the individual. Note that the distribution of the survival time T is Weibull.

4. Monte Carlo Scheme

We have performed a simulation study to empirically examine the comparative performances on several imputation procedures as mentioned in section 3. The criterion of this work was to investigate the accuracy of four different imputation procedures on the estimates of the logistic regression coefficients in the prediction system below.

Data were generated from a Weibull distribution W(r,B) with two fixed parameters where B is a scale parameter with B=$1/\lambda$ and r is a shape parameter using IMSL subroutines RNWIB and SSCAL. And thus the survival time t is of the form

$$t = \mathrm{W}(r,B)\exp(\beta_1c_1 + \beta_2c_2).$$

Both values of β_1 and β_2 were set to 1. The missing scheme is based on MAR. Two covariates were generated, one, c_1, is a categorical variable which takes 1 and 0 as its value and the other, c_2, is a random sample from a uniform distribution U(0,1) when c_1 takes the value 0 and when c_2 takes the value 1, respectively. Ten percent of the values were randomly generated from the status variable, which take the value 0 as the censored status.

The results presented here are all based on 5000 replications for sample sizes 40, 60 and 100, and for shape parameters 1.0, 1.5, 2.0 and 2.5. Reasonable proportions (k) of missing data are considered as 0.05, 0.10, 0.15 and 0.20 in this study and these seem to cover all the values likely to occur in real practical situations.

References

Bello, A. L. (1995), Imputation techniques in regression analysis: Looking closely at their implementation. *Computational Statistics & Data Analysis*, **20**, 45-57.

Efron, B. (1994), Missing data, imputation, and bootstrap. Journal of the American Statistical Association, **89** 463-474.

Hegamin-Younger, C. and Forsyth, R. (1998), A comparison of four imputation procedures in a two-variable prediction system, *Educational and Psychological Measurement*, **58**, 197-210.

Little, R. J. A. and Rubin, D. B. (1987), Statistical analysis with missing data: John wiley and sons, New York

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