Imputation Procedures in Exponential Regression Analysis in the presence of missing values

Young Sool Park[1]

Abstract

A data set having missing observations is often completed by using imputed values. In this paper, performances and accuracy of five imputation procedures are evaluated when missing values exist only on the response variable in the exponential regression model. Our simulation results show that adjusted exponential regression imputation procedure can be well used to compensate for missing data, in particular, compared to other imputation procedures. An illustrative example using real data is provided.

Keywords : MCAR, Imputation procedure, Hot deck, Adjusted exponential regression imputation, Exponential regression model

1. INTRODUCTION

The problem of missing values in incomplete data arises frequently in many data sets and appears particularly 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 result of nonresponse for a sample unit and typically arise from subjects who refuse or are inaccessible. This type of nonresponse is also called unit nonresponse. On the other hand, missing items may result when some individuals provide information, but fail to answer some of the questions. This type of nonresponse is called item nonresponse.

When analyzing data with missing values, it is common practice to either 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 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 compared to discarding all units which have missing values on any variables used in the particular analysis. Once the missing values are imputed, methods of analysis that require complete data on all

[1] Associate professor, Dept. of Information Statistics, Kwandong University, Kangnung, Kangwondo 210-701, Korea,
E-mail : yspark@kwandong.ac.kr

variables are then used to analyze the data.

Efron (1994) used nonparametric bootstrap approaches to assess the accuracy of an estimator in a missing data situation and found that the simplest form of nonparametric bootstrap confidence interval turns out to give convenient and accurate answers. Bello (1995) examined several numerical 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. He found that imputed values based on both the response and explanatory variables may give spurious impression of high precision especially as the proportion of missing data increases, and overestimation of residual mean square error may arise when imputed values are based on only the explanatory variables. 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. The results of the study provide that the grand mean procedure is not appropriate for handling missing data and when the prediction of the dependent variable is of interest, the regression procedure should be used.

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 $x_1, \cdots x_k$ 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:

$$P\{m_i = 1 \mid y_i, x_{1i}, \cdots, x_{ki}\} = P\{m_i = 1\}$$

(2) missing at random(MAR) if

$$P\{m_i = 1 \mid y_i, x_{1i}, \cdots, x_{ki}\} = P\{m_i = 1 \mid x_{1i}, \cdots, x_{ki}\}$$

(3) non-ignorable(NI) if

$$P\{m_i = 1 \mid y_i, x_{1i}, \cdots, x_{ki}\} = P\{m_i = 1 \mid y_i, x_{1i}, \cdots, x_{ki}\} \text{ or}$$

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

The primary purpose of this paper is to examine the behavior of and to investigate the accuracy of five different imputation procedures on the estimates of the Exponential regression coefficients when we are considering only the situation where time to event, Y, is missing and the mechanism of missingness for the variable, Y, is MCAR. Five imputation procedures we examine are the grand mean imputation procedure(GM), conditional mean imputation procedure(CM), hot-deck (HD), exponential regression imputation procedure(EI) and adjusted exponential regression imputation procedure(AEI).

In section 2 the exponential regression model is formulated and in section 3 we briefly describe the imputation procedures under consideration. Monte Carlo design and results are presented in section 4. Finally, an application to an example data set and conclusions are given in sections 5 and 6, respectively.

2. THE EXPONENTIAL REGRESSION MODEL

Let t_i denote the failure time of the i-th observation. The hazard rate of the proportional hazard model for the Exponential distribution function is given by

$$h(t) = \lambda \exp\{\beta_1' x_1 + \cdots + \beta_p' x_p\},\ t \geq 0 \text{ and } \lambda > 0$$

where λ is the scale parameter, $x_1, \cdots, x_p$ is a set of covariates, and $\beta_1', \cdots, \beta_p'$ are regression coefficients. If we let $Y = \ln t$, then Y has a standard extreme-value distribution. The accelerated failure-time(AFT) model for the same distribution function is also given by

$$t = \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\}\,\varepsilon$$

where $x_1, \cdots, x_p$ is a set of covariates, and ε has an exponential distribution with parameter 1.

There are relationship between the proportional hazard model and the accelerated failure-time model as next follow.

$$\beta_i = -\beta_i' \text{ for } i = 1, 2, \cdots, p,$$
$$\lambda = \exp\{-\beta_0\}.$$

3. SOME IMPUTATION-BASED PROCEDURES

In this section we present a brief description of five imputation procedures that use all covariates to obtain imputed values for missing values in response variable Y. Suppose that some individuals do not have complete variables and that the missing variables are missing completely at random as described in Rubin (1976).

3.1 Traditionary Method

(a) Grand Mean Imputation

The grand mean imputation procedure (GM) is perhaps the simplest imputation procedure. This procedure involves replacing missing values on a particular variable by the mean value of the observed data on that variable.

(b) Conditional Mean Imputation

The conditional mean imputation (CM) uses collateral information to provide an estimate for the missing value. This procedure partitions the sample into homogeneous groups based on responses to collateral information. For example, groups might be formed on the basis of stated responses (or dependent variable) to a question about class rank. The responses are categorized into quartiles (top 25%, second quarter, third quarter and the last quarter).

(c) Hot Deck Method

An individuals missing value, Y, is replaced by the value of another individual sampled in the survey whose Y value was not missing. There are many ways to select the donor individuals for a hot-deck (HD) 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 set 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 set, the cell of each record is identified and, if a certain value of a variable is missing, then the value for that cells register is substituted for the missing value.

On the other hand, if the individuals 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.

3.2 Exponential Regression Imputation and Adjusted Exponential Regression Imputation

Regression equations are fit from a data set consisting of complete records with the variable to be imputed serving as the dependent variable. The fitted regression line may be of the form

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k.$$

where $\hat{y}$ is the response variable to be imputed for a given record and $x_1, x_2, \cdots, x_k$ are covariates known for the individual. This method is called regression imputation.

Exponential regression imputation (EI) is studied similarly and can be discussed in a manner similar to regression imputation. The AFT models are fit from a data set consisting of complete individuals with the variable to be imputed serving as the dependent variable. The resulting Exponential regression model may be of the form

$$\hat{t} = \exp(b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k), \tag{3.1}$$

where $\hat{t}$ is the fitted survival time and $x_1, x_2, \cdots, x_k$ are covariates known for the individual. EI will be a procedure to replace missing values by the fitted values.

The log-likelihood function $L(\beta)$ of the AFT model is obtained as

$$L(\beta) = \sum(c_iz_i - e^{z_i}).$$

where $z_i = y_i - x_i'\beta$, $\beta = (\beta_0, \beta_1, \cdots, \beta_k)'$ and $x_i = (1, x_{1i}, \cdots, x_{ki})'$. Note that the score equation for the intercept term is

$$\sum c_i = \sum\left(\frac{t_i}{\hat{t}_i}\right) \tag{3.2}$$

where $\hat{t}$ is given in equation (3.1).

In the usual linear regression model, the sum of the observed and the fitted values are equal. However, the first term in (3.2) is equal to the number of non-censored observations. For this reason, the adjusted predicted value $\tilde{t}_i$ of the Exponential regression model can be proposed as

$$\tilde{t}_i = \left(\sum \frac{c_i}{n}\right)\hat{t}_{i..}$$

Adjusted Exponential regression imputation procedure (AEI) can be proposed by replacing missing values by the adjusted predicted values $\tilde{t}_i$. As expected, in section 4 we will show that AEI becomes a slightly improved procedure compared to EI. Note that the distribution of the survival time t is Exponential.

4. MONTE CARLO STUDY

We have performed a simulation study to empirically examine the comparative performance of the imputation procedures described in section 3. The objective of this work was to investigate the accuracy of five different imputation procedures on the estimates of the Exponential regression coefficients in the prediction system below.

Data were generated from a Exponential distribution $E(\lambda)$ with fixed parameter, where λ is a scale parameter using IMSL subroutines RNEXP and SSCAL. The resulting survival time t is of the form

$$t = E(\lambda = 1)\ \exp(\beta_0 + c_1\beta_1 + c_2\beta_2).$$

Values of $\beta_0, \beta_1, \beta_2$ were set to 1.5, 5.5 and -1.5 respectively. The missing scheme is based on MCAR. Two covariates were generated: the first, c_1, is a categorical variable which takes on the values 0 for half the subject and 1 for the other half; the second, c_2, is a random variable generated from a uniform distribution $U(0,1)$ using IMSL subroutine RNUN. 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. Reasonable proportions(k) of missing data are considered as 0.05, 0.10, 0.15 and 0.20 in this study as these would seem to cover values likely to occur in real practical situations.

Tables 1 and Tables 2 present the estimates of five different parameters $(\beta_0, \beta_1, \beta_2)$ and the corresponding mean squared errors(MSEs), respectively, arising from the use of the five imputation procedures on incomplete Exponential regression data.

We note several commonalities among the five imputation procedures: (1) When the missing proportion, k, increases, the estimate based on each of the imputation procedures gets more remote from the true values and the corresponding MSEs tend to increase. (2) When the size of the sample, n, increases, the corresponding estimates of all parameters seem to get closer to their true values and, as a result, the corresponding MSEs decrease.

Differences among five different imputation procedures can be summarized as follows: (1) When the sizes of the sample, 40 or 60, MSEs based on CM and AEI has smallest value. When the size of the sample, 100, and the missing proportions, 0.05 and 0.10, MSEs based on AEI and CM has smallest value. When the size of the sample, 100, and the proportions of missing data, 0.15 and 0.20, MSEs based on AEI and HD has smallest value. (2) In the estimates of parameters $(\beta_0, \beta_1, \beta_2)$, AEI and HD has a smaller bias than CM. (3) MSEs based on GM tend to be bigger than those based on other four imputation procedures, regardless of sample size and missing proportion. So GM is not good. MSEs based on AEI becomes a slightly smaller than those based on EI.

Table 1. Comparison of five imputation procedures based on the estimates, where GM means the grand mean imputation, CM the conditional mean imputation, HD the hot deck method, EI the Exponential regression imputation and AEI the adjusted Exponential regression imputation. The 3×1 vector in each cell denotes the estimates of the shape parameter, intercept term β_0, β_1 and β_2, respectively when the true values of the 3×1 vector are -1.5, -5.5, 1.5, respectively.

n	*k*	GM	CM	HD	EI	AEI
40	5%	-3.635123 -4.873974 3.571461	-1.500239 -5.618687 1.499434	-1.531679 -5.611927 1.540333	-1.533301 -5.622315 1.540218	-1.526753 -5.620828 1.532117
	10%	-4.323903 -3.811954 3.089545	-1.475153 -5.641269 1.451774	-1.536520 -5.606614 1.540844	-1.535152 -5.618607 1.536323	-1.520455 -5.616261 1.526687
	15%	-3.864768 -2.526356 0.289728	-1.395989 -5.566090 1.149003	-1.542747 -5.604462 1.554565	-1.529388 -5.620047 1.528153	-1.521009 -5.624607 1.562556
	20%	-4.415373 -2.220041 0.775722	-1.325754 -5.584087 1.098648	-1.545991 -5.598067 1.550889	-1.551888 -5.624381 1.562415	-1.508503 -5.622331 1.561946
60	5%	-4.200583 -4.031991 3.305291	-1.540261 -5.539330 1.447723	-1.609492 -5.498826 1.556051	-1.619278 -5.494202 1.558645	-1.615212 -5.495996 1.557929
	10%	-4.519772 -3.364414 2.817064	-1.500836 -5.564879 1.368722	-1.605988 -5.496493 1.550907	-1.621527 -5.493703 1.557850	-1.613282 -5.494158 1.562955
	15%	-4.473425 -2.373850 1.141999	-1.431176 -5.545960 1.198603	-1.611064 -5.493121 1.556271	-1.633810 -5.485446 1.563964	-1.611423 -5.492766 1.575350
	20%	-4.521713 -2.119308 0.792166	-1.426651 -5.525581 1.139993	-1.610302 -5.487852 1.546610	-1.633040 -5.492458 1.573598	-1.610801 -5.495729 1.615629
100	5%	-3.334405 -3.219060 0.520675	-1.527537 -5.442117 1.314367	-1.558673 -5.454051 1.402601	-1.567348 -5.447823 1.403379	-1.564289 -5.449037 1.403888
	10%	-3.693627 -2.747298 0.353834	-1.508374 -5.433132 1.255393	-1.559703 -5.455675 1.406137	-1.567142 -5.445173 1.393682	-1.557994 -5.445399 1.396877
	15%	-3.966180 -2.332092 0.038747	-1.459855 -5.427676 1.105867	-1.560780 -5.458047 1.410002	-1.560498 -5.438385 1.360402	-1.549130 -5.442465 1.389138
	20%	-4.458676 -2.028631 0.509578	-1.392729 -5.474834 1.029996	-1.565741 -5.457628 1.420577	-1.568136 -5.434446 1.370499	-1.537009 -5.436823 1.396470

Table 2. Comparison of five imputation procedures based on MSEs, where GM means the grand mean imputation, CM the conditional mean imputation, HD the hot deck method, EI the Exponential regression imputation and AEI the adjusted Exponential regression imputation.

n	k	GM	CM	HD	EI	AEI
40	5%	3.290644	0.204909	0.242663	0.211084	0.211013
	10%	4.657721	0.203242	0.254296	0.216319	0.215266
	15%	5.382910	0.193223	0.266210	0.244044	0.245374
	20%	6.657419	0.222159	0.271079	0.268105	0.266821
60	5%	4.352593	0.119442	0.145434	0.139835	0.139517
	10%	5.244156	0.118754	0.156914	0.146505	0.145638
	15%	6.295015	0.134957	0.167482	0.163793	0.162409
	20%	7.072670	0.145135	0.177154	0.173941	0.174901
100	5%	3.209503	0.084957	0.089958	0.085220	0.085003
	10%	4.616873	0.092629	0.093761	0.090034	0.089019
	15%	6.132301	0.117365	0.098049	0.097506	0.094475
	20%	7.301789	0.140754	0.109728	0.104334	0.101085

5. AN EXAMPLE

To illustrate some patterns for the values imputed by the five imputation procedures, the data set for the HMO-HIV+study shown in Hosmer & Lemeshow (1999, p.4) was used. Four variables, time(days between entry date and end date), age, drug(history of IV drug use) and a censoring status variable measured on 100 different subjects were considered. For the purpose of demonstration, we have chosen the same missing proportions k=5%, 10%, 15% and 20% as was used in the simulation study in section 4. Missing values were chosen at random from the response variable, time. Results of applying some imputation procedures on the incomplete data are shown in Tables 1-2.

Table 3. Comparison of five imputation procedures based on the estimates, where GM means the grand mean imputations, CM the conditional mean imputation, HD the hot deck method, EI the Exponential regression imputation and AEI the adjusted Exponential regression imputation. The 3×1 vector in each cell denotes the estimates of the shape parameter, intercept term β_0, β_1 and β_2, respectively when the true values of the 3×1 vector are -6.151630, 0.092092, 1.009856, respectively (HMO-HIV+ study).

k	nonmissing	GM	CM	HD	EI	AEI
5%	-6.151630 0.092092 1.009856	-5.713845 0.077556 1.022298	-5.609066 0.073270 1.130116	-6.079307 0.089992 0.994101	-5.995275 0.086656 1.035377	-6.006202 0.087053 1.034384
10%	-6.151630 0.092092 1.009856	-5.597359 0.073935 0.911987	-5.579041 0.071069 1.111660	-5.946630 0.085387 1.011745	-5.995272 0.086041 1.015121	-6.018138 0.087083 1.016440
15%	-6.151630 0.092092 1.009856	-5.565361 0.072042 0.866754	-5.579041 0.069943 1.143497	-5.940540 0.084477 0.974614	-6.039474 0.085652 1.048344	-6.067515 0.087179 1.048025
20%	-6.151630 0.092092 1.009856	-5.516619 0.069418 0.818621	-5.514032 0.066406 1.177710	-5.753061 0.076774 1.029930	-6.014347 0.083538 1.060857	-6.050556 0.085804 1.066151

Table 4. Comparison of five imputation procedures based on MSEs, where GM means the grand mean imputations, CM the conditional mean imputation, HD the hot deck method, EI the Exponential regression imputation and AEI the adjusted Exponential regression imputation (HMO-HIV+study).

k	GM	CM	HD	EI	AEI
5%	0.0640072	0.1030640	0.0018277	0.0083759	0.0072588
10%	0.1057082	0.1205968	0.0140225	0.0081707	0.0059632
15%	0.1215305	0.1154028	0.0152863	0.0046980	0.0028521
20%	0.1467745	0.1451218	0.0531648	0.0071736	0.0044749

The pattern of the imputed values shown in these tables lends support to our simulation results. For *k*=10%, 15% or 20% and regardless of sample sizes, in particular, AEI is the most efficient of the imputation procedures considered in the sense of closeness of Exponential regression coefficients to their true values and smallness of MSEs as comparative criteria. In addition, for $k=5\%$, HD and AEI are more efficient than any other imputation procedures. That is, MSEs based on HD and AEI are relatively smaller than GM and CM.

6. CONCLUSIONS

In this paper, the simulation study clearly shows that the five different imputation procedures may lead to different results, depending on sample sizes, shape parameter and proportion of missing values in the Exponential regression model. The results of the study show that GM is not an appropriate procedure for handling missing data. In particular, AEI may be recommended as a useful and practical guide even in the Exponential regression models when estimating the Exponential regression coefficients of the covariates. But when determined to use AEI, we should be careful which variables to choose the given data set to apply AEI properly. For the use of AEI, we propose to select all significant covariates in Exponential regression model. I would like to recommend AEI because using CM might cause various limitations in selecting variables and grouping data.

REFERENCES

1. Bello, A. L. (1995). Imputation techniques in regression analysis: Looking closely at their implementation, *Computational Statistics & Data Analysis*, 20, 45-57.
2. Efron, B. (1994). Missing data, imputation, and bootstrap, *Journal of the American Statistical Association*, 89, 463-474.
3. 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.
4. Hosmer, D. W., and Lemeshow, S. (1999). *Applied survival analysis*: John Wiley and Sons, New York.
5. Little, R. J. A., and Rubin, D. B. (1987). *Statistical analysis with missing data*: John Wiley and Sons, New York.
6. Rubin, D. B. (1976), Inference and missing data, *Biometrika*, **63**, 581-592.