Indian Journal of Science and Technology Vol. 5 No. 2 (Feb 2012) ISSN: 0974- 6846

Methods of analysing missing values in a regression model

P. Ogoke Uchenna and E.C. Nduka*
Department of Mathematics and Statistics University of Port Harcourt-Nigeria
etelnduka@yahoo.com*

Abstract

Different methods of imputation are adopted in this study to compensate for missing values encountered in the data collected. The imputation methods considered are the overall mean value, Random Overall, Logistic Regression, Linear Regression, Predictive Match, Multiple Imputations and the Hot Deck Imputation. The various values obtained by the methods were analysed and compared using Bartlett's test statistic for equality of variances among groups (Mean Square Errors of the seven methods). The software packages used for this research work are Winmice, Solas and SAS. Different values were estimated applying the various methods. However, results obtained from the test showed that the variances among the groups have no significant differences, that is, any of the imputation methods could be used. Further test using relative variance revealed that the multiple imputation method may be preferred.

Keywords: Missing at Random, Imputation, Bartlett's Test, Coefficient of Relative Variance.

Introduction

In carrying out surveys, the aim has always been to produce estimates based on all units of the sample. However, this is rarely possible since a number of factors may lead to missing values. Proper handling of missing values is very important in all experiments. Improper handling of missing values will distort analysis and results. The problem with missing values is not so much reduced sample size as is the possibility that the remaining data set is biased. The imputation of values where data are missing is an area of statistics which has developed much since 1980s (http://www.mendeley.com).

An obvious desire of both the data collector and the data analyst is to get rid of the missing values and thereby restore the ability to use standard complete data method to draw inferences. Item non-response may occur in a sample survey because a sample unit may refuse or be unable to answer a particular question or due to fatigue sensitivity or lack of knowledge or other factors, respondents not infrequently leave a particular item blank on mail or questionnaires or decline to give any response during interviews. Other reasons for missing observations in an experiment could be transcription errors, drop outs in follow up studies, and clinical trials. Various methods have therefore been proposed to address missing data in experiments. We will apply these methods to estimate the missing observations, in order to draw proper inferences on the data set.

The three main approaches to handling missing values include Discarding the missing values and analyzing the rest (Rubin, 1986), weighting adjustment and imputation/estimation of missing values. In this study therefore, we aim to impute missing values in an observational data using different imputation methods, to compare the mean square errors (MSE) of the analysed data for various imputation procedures employed and to test, if there is any significant difference in the imputation methods adopted.

There are several methods for handling missing data in sample surveys. Afifi and Elashoff (1966) highlighted different methods of handling missing data which include complete-case analysis or list wise deletion and available - case analysis or pair wise deletion. The complete case analysis can be very inefficient since it reduces the sample size (Little & Rubin, 2002). For available-case analysis, different subsets of cases are used to estimate individual parameters. When cases with missing data are ignored, the parameter estimates may differ from the target population. Biases may be introduced. Moreover, pair wise deletion can yield an estimated covariance matrix that is not necessarily positive definite (Kim & Curry 1977). Thus statistical analysis such as regression analysis may be problematic.

Lepkowski *et al*. (1987) analyzed imputed data from a sample survey, the National Medical Care Utilization and Expenditure Survey (NMCUES) which was designated to collect data about the United States Civilian Non-Institutionalized Population in 1980's. They presented four different strategies for handling imputed survey data which the result of their paper allows. For a large complex survey like theirs, they recommended that the first strategy "use all the data, real as well as imputed, in all analysis," be adopted. Our desired option for handling missing value is imputation. Although imputation techniques have not been used widely, a number of applications pertaining to demographic and health research has appeared in the statistical literature.

Methodology

The data for this study were obtained from Chukwuma Farms Enterprises, Abia State, Nigeria, on the production capacity of different breeds of poultry birds for a period of 18 months. Since the work has to do with the different methods of analysing missing values in a regression model, one of the objectives of this study is to impute missing values in an observational data using different imputation methods.

Indian Journal of Science and Technology Vol. 5 No. 2 (Feb 2012) ISSN: 0974- 6846

Table 1. Imputation methods and results

	Overall Mean	Random Overall	Logistic regression	Linear regression	Predictive match	Multiple imputation	Hot deck imputation
YAFFA	53.55714	51	49	51.3631	54	51.54	62
	53.55714	50	49	41.60622	56	48.01	55
NIGER.P	53.76056	53	46	63.287331	61	43.93	55
HARCO	54.1594	55	48	58.54	50	50	50
	54.1594	46	48	48.27	50	49	62
	54.1594	60	48	52.23	50	51	57
BLACK P	53.2428	61	47	52.348	50	64.57	56
	53.2428	63	47	49.62	57	62.75	51

The data set consist of the following information:
Y1…Y4, are the dependent variables (different breeds of poultry birds) while X is the independent variable (quantity of feed given to the poultry birds). (See the main work for the data).
Y1 = Yaffa
Y2 = Niger pullets (Niger. P)
Y3 = Harco
Y4 = Black pullets (Black. P)
X = Different quantities of feed given to the poultry birds
* = The row that has missing values.
In this study we considered seven methods of imputation namely: Overall Mean Value, Random Overall, Logistic Regression, Linear Regression, Predictive Match, Multiple Imputation and Hot Deck Imputation method.

This work considers the effect of imputed values from different methods of simple linear regression analysis. Suppose that r of n dependent y-values are actually measured, the remaining m = (n - r) y-values are missing and all independent x-values are observed. For simplicity, we assume that $y_{r+1} \ldots y_n$ are missing.

We need to impute the missing y-values from the known values of x's and the observed r values of y's.
Where
r = non missing values.
n =number of values.
m =number of missing values in the dependent y-values.
Out of the four breeds of poultry birds cited, the outcome of one (1) was not recorded in Y_1 resulting in 1.38% missing responses, Y_2 has 2.78% missing responses,Y3 has 4.17% while Y4 has 2.78% missing responses. We assume that the missing observations are missing at random (MAR). Then, the appropriate values (or estimates) of non respondents were obtained by the use of the soft-wares (SOLAS & WINMICE) using the procedures discussed in Table 1.

Summary of the estimation of the missing values using the various imputation methods

This was done with the use of the software known as 'WINMICE' & 'SOLAS' (Solas Version 3.2; Winmice Prototype Version 0.1.)
This information is shown in the Table one below:
A standard data set was obtained with the use of Table one, then used for our analysis.

Data analysis

We wish to estimate the various parameters using the simple linear model. $Y_i = \beta_0 + \beta_1 X_1 + e_i$, where β_0 and β_1 are unknown regression parameters and e_i is the error term that describes the effect of $Y_1, \ldots, Y_4$ other than the value of the independent factor x. To get the Mean Square Error(MSE) of our regression estimates, we used a software known as SAS as shown in Table 2.

Tests for equatily of variances

Hypothesis: $H_o: \sigma_1^2 = \sigma_2^2 = \ldots, = \sigma_k^2$

Table 2. Mse for the seven imputation methods

	Overall Mean	Random Overall	Logistics regression	Linear regression	Predictive match	Multiple imputation	Hot deck imputation
YAFFA	40.81609	41.04725	41.33848	42.79729	40. 91793	41.25784	41.89404
NIGER.P	36.98106	36.98984	37.83312	38.24656	37.71031	38.34599	37.00154
HARCO	39.23840	40.68772	40.71582	40.07132	40.07132	39.92802	40.46262
BLACK P	38.25707	40.48491	39.22757	38.40127	38.63158	41.36474	38.42526

Table 3. Bartlett's test for equality of k - variances

Dependent variables (Y)	Overall Mean	Random Overall	Logistics regression	Linear regression	Predictive match	Multiple imputation	Hot deck
Y1 (YAFFA)	40.81609	41.04725	41.33848	42.79729	40.91793	41.25784	41.89404
Y2 (NIGER. P)	36.98106	36.98984	37.83312	38.24656	37.71031	38.34599	37.00154
Y3 (HARCO)	39.2384	40.68772	40.71582	40.07132	40.07132	39.92802	40.46262
Y4 (BLACK. P)	38.25707	40.48491	39.22757	38.40127	38.63158	41.36474	38.42526
VARIANCES	$\hat{S}_1^2 = 2.6193$	$\hat{S}_2^2 = 3.5699$	$\hat{S}_3^2 = 2.4667$	$\hat{S}_4^2 = 4.4673$	$\hat{S}_5^2 = 2.0607$	$\hat{S}_6^2 = 2.0101$	$\hat{S}_7^2 = 4.6812$

Indian Journal of Science and Technology Vol. 5 No. 2 (Feb 2012) ISSN: 0974- 6846

Vs H_1: $\sigma_1^2 \neq \sigma_2^2 = \ldots, = \sigma_k^2$

where k is the number of imputation methods adopted k = 7

and $\sigma_i^2 =$ Variances of imputation methods, k=1,2,.., 7.

The hypothesis is rejected if $B \geq \chi^2_{k-1}(0.05)$

$\Rightarrow \quad B \geq \chi^2_6(0.05)$, otherwise we accept Ho.

The test statistics is

$$B = \frac{2.30259}{C}\left[\sum(n_{i-1}) \log S_p^2 - \sum(n_{i-1}) \log S_i^2\right]$$

In the above, S_i^2 is the variance of the i[th] group, n is the total sample size, n_i is the sample size of the i[th] group, k is the number of imputation methods considered and S_p^2 is the pooled variance, (Table 3).

The pooled variance is a weighted average of the group variances and is defined as

(pooled variance)=

$$\frac{\sum_{i=1}^{k}(n_i - 1)S_i^2}{\sum_{i=1}^{k}(n_i - 1)} =$$

$$\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \ldots + (n_7 - 1)S_7^2}{(n_1 - 1) + (n_2 - 1) + \ldots + (n_7 - 1)}$$

We have

$$S_p^2 = \frac{65.6256}{21} = 3.1250$$

$$\log S_p^2 = 0.4949$$

N/B: $n_1 = 4$, $n_2 = 4 = n_3 = n_4 = n_5 = n_6 = n_7 = 4$

The test statistics is

$$B = \frac{2.30259}{C}\left[\sum(n_{i-1}) \log S_p^2 - \sum(n_{i-1}) \log S_i^2\right]$$

where: $\sum(n_{i-1}) \log S_p^2 - \sum(n_{i-1}) \log S_i^2 = 0.4901$

$$C = 1 + \frac{1}{3(k+1)}\left[\frac{1}{\sum(n_{i-1})} - \frac{1}{\sum(n_{i-1})}\right] = 1.0952$$

$$B = \frac{(2.3026)(0.4901)}{1.0952}$$

$$B = 1.0304$$

we reject Ho if $B \geq \chi^2_6(0.05)$ otherwise accept Ho.

Table 4. Comparison of variance estimates for the seven imputation methods

S/N	Imputation methods	Variances	Ratio	Deviation from 1	Rank
1.	Overall Mean Value	$\hat{S}_1^2 = 2.6193$	0.8382	-0.1618	4
2.	Random Overall	$\hat{S}_2^2 = 3.5699$	1.1424	0.1424	5
3.	Logistic Regression	$\hat{S}_3^2 = 2.4667$	0.7893	-0.2107	3
4.	Linear Regression	$\hat{S}_4^2 = 4.4673$	1.4295	0.4295	6
5.	Predictive Match	$\hat{S}_5^2 = 2.0607$	0.6594	-0.3406	2
6.	Multiple Imputation	$\hat{S}_6^2 = 2.0101$	0.6432	-0.3568	1
7.	Hot Deck	$\hat{S}_7^2 = 4.6812$	1.4980	0.4980	7
		Pooled Variance 3.1250	1.0000	0.0000	

$X_6^2(0.05) = 12.592$

B = $1.0304 \leq X_6^2(0.05) = 12.592$

:. we accept Ho since there is no significant difference among the methods.

Coefficient of relative variance

The coefficient of relative variance (CRV) is a square of coefficient of variation. This method is used to show how much each method of imputation varies from the pooled variance.

Thus: Ratio = $\frac{S_i^2}{S_p^2}$

where S_i^2 is the variance of the ith group and S_p^2 is the pooled variance

The coefficient of relative variance was computed to compare how much each imputation method varies from the pooled variance. The coefficient of relative variance was computed by dividing each estimated variance method by the pooled variance $\frac{S_i^2}{S_p^2}$ and then we evaluated the deviation from 1 by subtracting each variance method from the pooled variance ratio. Finally, the ranking was done in ascending order to show the method of imputation that is more consistent and efficient. From Table 4, the result of the coefficient of relative variance points to the multiple imputation method as

Indian Journal of Science and Technology Vol. 5 No. 2 (Feb 2012) ISSN: 0974- 6846

being more consistent and efficient since it has the least relative variance. This in-turn minimizes bias and eliminates its effects.

Conclusion

The Bartlett's test statistic for the equality of variance shows that there is no significant difference among the imputation methods. In addition, the coefficient of relative variance (CRV) was computed and the results were ranked which points to the multiple imputation as being consistent and efficient. Hence it is most preferred.

References

1. Afifi AA and Elashoff RM (1966) Missing observations in multivariate statistics. *J. Am. Stat. Asso.* 61(2), 595-604.
2. Kim J and Curry J (1977) The Treatment of Missing Data in Multivariate Analysis. *Sociol. Methods & Res*. 6, 215-240.
3. Lepkowski JM, J Richard Landis, Sharon A Stehouwer (1987) Strategies for Analyzing of Imputed Data from Sample Survey. The *Med. Care utilization & Expenditure*. 25(8), 705-715.
4. Little RJA and Rubin DB (2002) Statistical analysis with missing data 2nd Edition, Wiley, NY.
5. Rubin DB (1986) Statistical matching using file concentration with adjusted weights and multiple imputation. *J. Bus. & Econ. Stat.* 4, 87-94.
6. SAS Learning Edition Version 4.1. Bringing the power of data analytics to individuals. http://www.amazon. com/SAS-Learning-4-1-Little-nterprise/dp/ 1590479173.