Mathematical Modelling and Applications
2020; 5(2): 87-93
http://www.sciencepublishinggroup.com/j/mma
doi: 10.11648/j.mma.20200502.14
ISSN: 2575-1786 (Print); ISSN: 2575-1794 (Online)

Comparative Study of Various Methods of Handling Missing Data

Fredrick Ochieng' Odhiambo

Department of Mathematics and Actuarial Sciences, South Eastern Kenya University (Seku), Kitui, Kenya

Email address:
od032003@yahoo.com

To cite this article:
Fredrick Ochieng' Odhiambo. Comparative Study of Various Methods of Handling Missing Data. *Mathematical Modelling and Applications*. Vol. 5, No. 2, 2020, pp. 87-93. doi: 10.11648/j.mma.20200502.14

Received: October 2, 2019; **Accepted**: April 13, 2020; **Published**: April 30, 2020

Abstract: Scientific literature lack straight forward answer as to the most suitable method for missing data imputation in terms of simplicity, accuracy and ease of use among the existing methods. Exploration various methods of data imputation is done, and then a robust method of data imputation is proposed. The paper uses simulated data sets generated for various distributions. A regression function on the simulated data sets is used and obtained the residual standard errors for the function obtained. Data are randomly from the set of independent variables to create artificial data-non response and use suitable methods to impute the missing data. The method of Mean, regression, hot and cold decking, multiple, median imputation, list wise deletion, EM algorithm and the nearest neighbour method are considered. This paper investigates the three most common traditional methods of handling missing data to establish the most optimal method. The suitability is hence determined by the method whose imputed data sample characteristic does not vary considerably from the original data set before imputation. The variation is here determined using the regression intercept and the residual standard error. R statistical package has been used widely in most of the regression cases. Microsoft excel is used to determine the correlation of columns in hot decking method; this is because it is readily available as a component of Microsoft package. The results from data analysis section indicated an intercept and R-squared values that closely mirror those of original data sets, suggesting that median imputation is a better data imputation method among the conventional methods. This finding is important from the research point of view, given the many cases of data missingness in scientific research. Finding and using the median is simple and as such most researchers have a ready tool at hand for handling missing data.

Keywords: Regression, Nearest Neighbor, Hot Decking, Median Substitution, Missing Data

1. Introduction

Research is the driving force behind any development of a Nation. Any endeavourer in this area therefore requires that the people concerned with the research arm themselves with the right kind of tools that shall help them get accurate and relevant information from the survey being undertaken. Missing data is a big challenge in many areas of research, especially in social research. Many researchers when confronted with this scenario feel very helpless. Some may resort to non-scientific ways of addressing this challenge while others compromise reliability of their findings by using procedures that cannot guarantee accuracy. Non-response problem is an issue of great concern to researchers because it pervades almost all survey research, [24].

In any research work, the ultimate goal of researcher is to con- duct the most accurate analysis of data to so as to be able to make valid and efficient inferences about a population to guide users of statistical results and researchers alike, [31]. The most challenging part therefore is to get all the relevant information about the case under investigation. In most cases this fails to materialize, one, because subjects make up their minds to hold back certain information for personal reasons or a few may not have ready answers at the point when they are being interviewed. It is worth noting that when part of a data is missing from a given survey and missing data is ignored by and using only the available sample, the result so yielded may not be representative of the population under study; after all there are some of its characteristics missing. Ignoring missing data occur when there is a wide spread failure to understand the significance of the problem or lack of awareness of the solution to the

problem of missing data, [17].

The higher the non-response rate the greater the bias if the characteristic under study in respondents differ markedly from non- respondents. According to [14], other causes of missing data are, error on the part of the researcher, those collecting or entering data and the participants.

Comparing them on the basis of their ease of use, efficiency and robustness, Median imputation performs better than the other methods. For a detailed review of these approaches. [25]

2. Various Methods of Handling Missing Data

This is a review of various methods that exist which have been used towards addressing the issue of missing data in survey.

2.1. EM Algorithm

This method is described as "archaic". [11] Despite being archaic it is still the quickest, however where accuracy is key, this method must be used cautiously. It is important to note that the method is largely suitable when dealing with data that is MCAR. This method may be suitable if only a small number of cases are missing values. The sentiment is also supported by, [11].

2.2. List Wise Deletion

After the list wise deletion, one cannot be guaranteed that the remaining data is still representative of the original population under study. This systematic loss of data by list wise deletion results into an increased risk of bias. According to [16], list wise deletion method is regarded as the most common and easiest method of dealing with missing data, it is also called complete case analysis according to [11]. This approach there- fore leads to a reduction in sample size which in turn translates into reduced statistical power bringing into question the how representative the remaining sample is of the population being studied. [11]. The list wise deletion according to [6], because of this systematic loss of data with list wise deletion, there is an increased risk of bias, a risk which can only be lessened when the data is MCAR. Some researchers have characterized list wise deletion as the least desirable data imputation method because of these biases and have warned against its use. [11].

2.3. Mean Substitution

The third method in this study is the mean substitution method. This method is "archaic"[12] but still considered. To use this method, the mean of the total sample for a variable is substituted for all the missing values in that variable. Mean substitution is a quick and easy way to recover cases. [20]. Furthermore the estimate of the standard deviation and variance used in calculating other parametric tests is reduced resulting in biased standard errors [36]. There is a debate about using this method because of the inherent bias that result [31]. This method would only be appropriate if only a small number of cases are missing values. The serious disadvantage with this method is that it can distort the distribution hence in underestimating variance and covariance [36]. According to [28] among the drawbacks of mean imputation are (a) Sample size is overestimated (b) Variance is underestimated (c) Correlation is negatively biased, and (d) The distribution of the new values is an incorrect representation of the population values because the shape of the distribution is distorted by adding values equal to the mean.

2.4. Regression Imputation

According to [15], the best predictors (that is, those with the highest correlations) are selected and used as independent variables in a regression equation the variable with missing data is used as independent variable. The predictors from the last round are the ones that are used to replace the missing value [15]. The statistical software is a better solution here, but even this comes a sacrifice that a user has to embrace in terms of time to learn the software apart from the financial constraints on the part of the user [33]. Clearly then, a better method of data imputation needs to be sought.

2.5. Multiple Imputations

Multiple imputations essentially is a way to solve the modeling problem by simulating the distribution of the missing data [30]. Users are free to ignore the imputations, all imputed values are tagged, "Satisfied", if variables that determine the nonresponsive are not included as conditioning variables, [32]. This has been demonstrated in simulation studies, [7]. Furthermore, using simulated and real datasets from different scientific fields and with varying rates of item non-response, existing research emphasizes the robustness of multiple imputation to the specially chosen imputation model, given that appropriate conditioning variables are available in the data set [2]. Multiple imputations create several imputed datasets. If automatic variable selection is then run on each of these datasets separately, the set of variables entering the model can vary across the datasets. This makes it hard to assimilate the results [33].

2.6. Hot Decking

This method works well when the variable used to sort the data is highly predictive of the variable with the missing values and when there is a large sample so that a similar case is easily identified [35]. According to [32], One of the advantages of hot decking, compared with mean substitution, is that the standard deviation of the variable with the inserted values better approximates the standard deviation value for the variable without the substituted values. However, standard, standard deviations are still likely to be lower overall [35]. This method may not work when there exists no correlation between the variables. Thus the method only works very when the variable used to sort the data is highly predictive of the variable with missing values and when there

is a large sample so that a similar case is easily identified [35]. Another drawback with hot decking is that it is difficult to implement; programming requires great time and labor. [39]

2.7. Median Imputation

According to [1], the mean is affected by the presence of outliers ad it seems natural to use the median instead just to assure robustness. The existence of other features in the data set with similar information (high correlation), or similar predicting power can make the missing data imputation useless, or even harmful [1].

2.8. The Nearest Neighbor (NN) Imputation Method

According to [38], this is one commonly used imputation method for item non response. Here the missing value is imputed from the ones at the auxiliary variables. Thus impute the auxiliary value closer to the missing value by considering the previous value and the next value. In which case a single value NN is carried out as follows: Considering a population, $\mathrm{U} = 1,2,3,..\mathrm{N}$. Associated with the k th unit of the population are two variables $(\mathrm{x_k}, \mathrm{y_k}), \mathrm{k} = 1,2,..\mathrm{N}$, where $\mathrm{x_k} > 0, \mathrm{y_k} > 0$. The variable y is unknown and is the variable under study and while is the covariate assumed to be known for all the units of the population. Supposing that in this sample m unit correspond to an item y and n, m do not. Then the value y_{ik} is imputed for the missing value y_k [38]. According to [26], the bias of the population mean is known to be small if the relationship between y and x is linear. Obviously therefore, when the relationship is not linear a serious challenge will arise [26].

3. Methods

Three traditional methods of data imputation are considered in this research. The methods involving substituting a single value [33]. Usually the imputed values are the mean or median of the variable being substituted, [33]. The dependent variables Y_{is} where $Y_{is} \in \varphi$ and the independent variable X_{is} as a linear combination of vectors are considered. A quantitative response is assumed here and a multiple regression as the most common method of statistical adjustment, additive model is proposed.

Let

$$X_{ij} = \mu_{ij} + \left(\beta_j X_{1j} + \beta_j X_{2j} + \cdots + \beta_j X_{nj}\right) \quad (1)$$

Where X_{ij} for $i = 1,2,..,n$ and $j = 1,..,n$ is such that $X_{ij}{}'s$ are normal iid random variables. A linear model for the distribution can be written as

$$Y = X\beta + \varepsilon \quad (2)$$

Equation (1) is used as a linear model with a logistic error, ε.

3.1. The Model

Let us denote each independent variable by Y_τ, let Y_τ, depend on several factors $X_i{}'s$. Each $X_i{}'s$, $X_i{}'s$ is therefore a vector belonging to a vector of random variables.

$Y_i = \beta_o + \sum_{y=1} \beta_{ij} X_{ij} + \varepsilon_{ij}$ Where $i = 1, \ldots, n$ and $k = 1, \ldots 4.$

3.2. Regression for Complete Data Set

Regression for a complete data set is proposed to have the form

$$Y_i = \beta_o + \sum_{y=1} \beta_{io} X_{iojo} + \varepsilon_o \quad (3)$$

Where $i = 1, \ldots, n$ and $k = 1, \ldots 4.$ was obtained for the original sample data set for each distribution.

The regression of data set with median imputation is

$$Y_e = \beta_e + \sum_{y=1} \beta_e X_{ieje} + \varepsilon_e \quad (4)$$

And is obtained from the data set, finally in the same way the regression with list wise deletion is obtained as

$$Y_I = \beta_I + \sum_{Y=1} \beta_{il} X_{IL} + \varepsilon_l \quad (5)$$

The error ε_i is assumed independent and identically distributed with mean zero and unit scale.

$$Y = \beta_o + \sum_{i=1} \sum_{j=1} \beta X + \varepsilon_I \quad (6)$$

3.3. Parameter Estimation

The R –statistical package was used to estimate the $\beta_i{}'s$ and the error term $\varepsilon_i{}'s$. The data non-response imputation that gave the values $\beta_i{}'s$ which closely mirror values from the complete values dataset is deemed to be the most robust method of data imputation. The values of the intercepts $\beta_O{}'s$ for the complete data set are compared with sample $\beta_\alpha{}'s$, $\beta_e{}'s$ and $\beta_i{}'s$ respectively for mean, median and list wise deletion approach. The procedure I repeated for five different distributions; Gamma, Weibull, binomial, Poisson and normal distributions for the 10% missing and 30% missing.

4. Results

The results on the optimal imputation, the main comparison here, are the intercepts, residuals, standard errors, R-squared and intercept standard errors.

Table 1. *Poisson Distribution Summary with 30% data non-response.*

Method	Intercept	RSE	R-Squared	Min. Residual	Max. Residual	Intercept S. E
None	2.93835	1.283	0.03418	2.2502	4.8368	.56097
List-wise	3.061151	1.316	0.03978	-2.1724	4.9180	.772275
Mean	3.051151	1.087	0.03978	-2.172	4.918	.633510
Median	3.004357	1.088	0.03896	-2.1444	4.9555	.628298

Poisson distribution table 1 shows that the method of median imputation does better with a Poisson distribution 30% data missing.

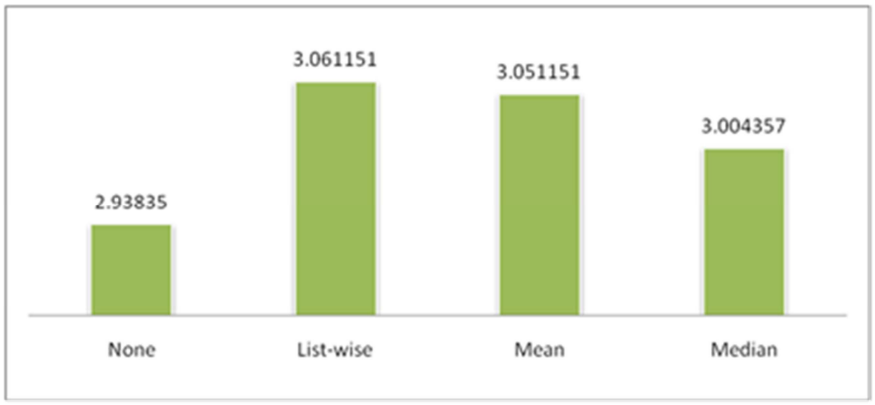

Figure 1. *Poisson Distribution Summary with 30% data non-response.*

Table 2. *Binomial Distribution Summary with 30% data non-response.*

Method	Intercept	RSE	R-Squared	Min. Residual	Max. Residual	Intercept S.
None	578.03200	17.99	0.01987	-47.78	37.242	55.70933
List-wise	534.61384	17.91	0.06202	-43.435	33.255	70.60254
Mean	534.61382	17.79	0.06202	-43.435	33.255	58.31206
Median	535.362	14.80	0.06237	-43.685	32.890	58.2280

Binomial distribution table 2 shows that the method of median imputation outperforms other methods in terms of intercept standard error and R-squared, with a Binomial distribution 30% data missing.

Table 3. *Normal Distribution Summary with 30% data non-response.*

Method	Intercept	RSE	R-Squared	Min. Residual	Max. Residual	Intercept S. E
None	730.05960	2.823	0.05213	-6.965	6.334	40.55204
List-wise	778.48588	2.836	0.07853	-5.2460	6.4261	49.65590
Mean	778.89485	2.343	0.07853	-5.246	6.426	41.01727
Median	740.412964	2.68	0.04145	-7.8321	7.5289	37.143166

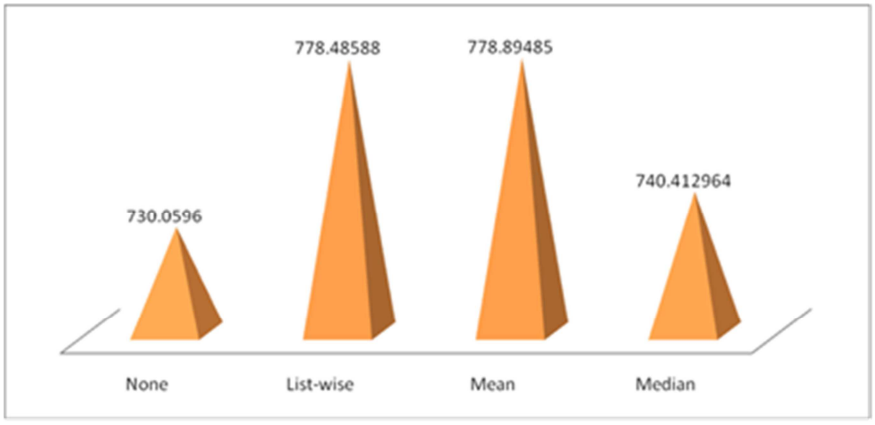

Figure 2. *Normal Distribution intercepts Summary with 30% data non-response.*

The above table of a normal distribution data set with 30% data non-response reveals that the median imputation posts far better intercept values. The R-square value also closely mirror those of original data set compared to the other two methods.

Table 4. *Poisson Distribution Summary with 10% data non-response.*

Method	Intercept	RSE	R-Squared	Min. Residual	Max. Residual	Intercept S. E
None	1.960	1.347	0.05349	-2.116	3.7741	0.592332
List-wise	1.58675	1.359	0.006922	-1.8994	4.2644	0.53011
Mean	2.30150	1.218	0.1075	-2.1419	4.1870	0.47784
Median	2.27339	1.238	0.977	-2.0465	4.23	0.483390

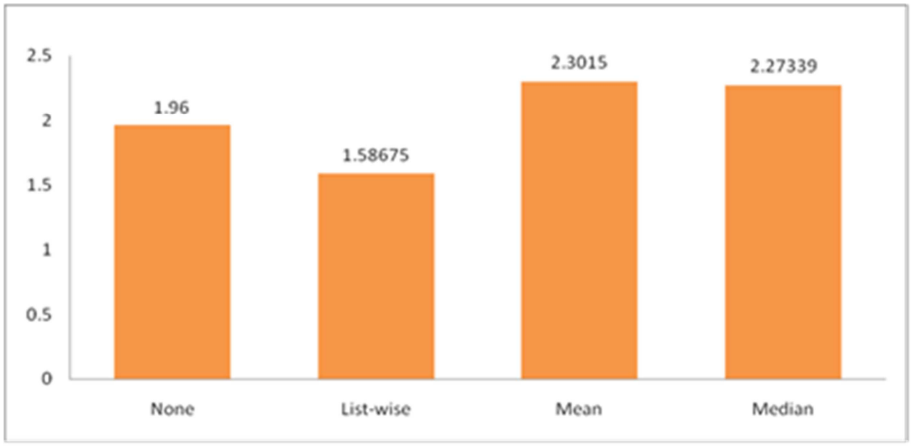

Figure 3. *Poisson distribution intercept Summary with 10% data non-response.*

With 10% missingness for a normal distribution table 4, there is clearly no major difference between the mean substitutions and leastwise deletion, perhaps a clear pointer that when non response is small it can be ignored without much effect on the sample size and results.

Table 5. *Wei bull Distribution Summary with 10% data non-response.*

Method	Intercept	RSE	R-Squared	Min. Residual	Max. Residual	Intercept S. E
None	25.44464	0.01902	0.05794	-0.052649	0.029370	0.99247
List-wise	25.547331	0.01783	0.01107	-0.050277	0.031643	1.019497
Mean	25.54311	0.01686	0.01107	-0.50277	0.31644	0.964014
Median	25.477415	0.0169	0.009958	-0.050805	0.31149	0.960140

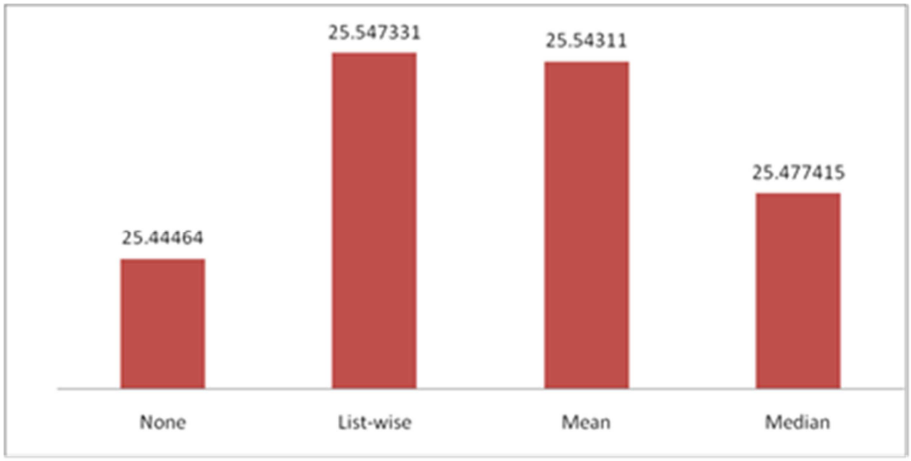

Figure 4. *Weibull Distribution intercept Summary with 10% data non-response.*

From above figures and tables 5 and 6 for both Weibull and Poisson distribution with 10% missingness, median imputation does better. We note that a list-wise deletion is not a good method as can seen from the values of the intercept and standard error.

Table 6. *Normal Distribution Summary with 10% data non-response.*

Method	Intercept	RSE	R-Squared	Min. Residual	Max. Residual	Intercept S. E
None	721.049822	2.682	0.2422	-6.3410	5.2815	33.352123
List-wise	704.30506	2.692	0.01681	-6.1132	5.1475	35.82217
Mean	704.30506	2.546	0.01681	-6.1132	5.1475	33.87254
Median	704.17013	2.548	0.01681	-6.1467	5.1147	33.89641

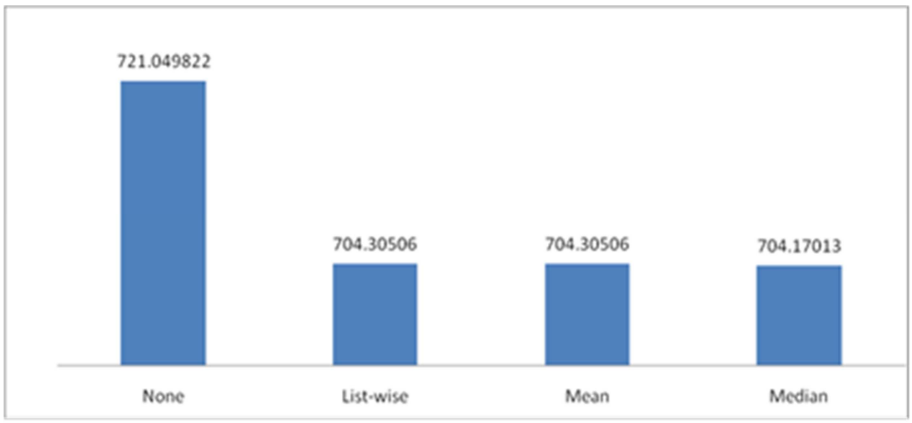

Figure 5. *Normal Distribution intercept Summary with 10% data non-response.*

With 10% missingness for a normal distribution both figure and table 6, there is clearly no major difference between the mean substitutions and leastwise deletion, perhaps a clear pointer that when non response is small with a normal distribution, it can be ignored without much effect on the sample size and results.

5. Conclusion

The aim was to establish the most reliable method of imputing missing data among the conventional methods. Simulation data set analysis show clearly that when data is MAR, Median imputation performed consistently better than list wise deletion. Median imputation equally did better than mean imputation for 30% and 10% non-response for both skewed distribution and the Normal distribution.

Median imputation is proposed as the optimal method of missing data imputation for data non response for up to 30%. It is considered that use of median for data imputation, gives researchers an easy to obtain, ready tool for handling data non-response.

References

[1] Acuna et al. (2008): The Treatment of missing values and its effect in the classifier accuracy. http://www.uprm.edu.

[2] Bernaards, C. A. et al. (2003): Comparison of Two Multiple Imputation Procedures in a Cancer Screening Survey. Journal of Data Science, 1 (3), 293-312.

[3] Biewen, M. (2001): Item non-response and inequality measurement: Evidence from the German earnings distribution. Allgemeines Statistisches Archiv, 85 (4), 409-425.

[4] Bover, O. (2004): The Spanish Survey of Household Finances (EFF): Description and Methods of the 2002 Wave. Documentos Ocasionales N. 0409. Banco de Espana.

[5] Cameron, A. C. and P. K. Trivedi (2005): Microeconometrics. Methods and Applications. New York: Cambridge University Press.

[6] Essig, L. and J. Winter (2003): Item Nonresponse to Financial Questions in Household Surveys: An Experimental Study of Interviewer and Mode Effects. MEA-Discussion Paper 39-03, MEA – Mannheim Research Institute for the Economics of Aging. University of Mannheim.

[7] Ezzati-Rice, T. M., W. Johnson, M. Khare, R. J. A. Little, D. B. Rubin, and J. L. Schafer (1995): Multiple imputation of missing data in NHANES III. Proceedings of theAnnual Research Conference, U.S. Bureau of the Census, 459-487.

[8] Ferber, R. (1966): Item nonresponse in a consumer survey. Public Opinion Quarterly, 30 (3), 399-415.

[9] Frick, J. R. and M. M. Grabka (2005): Item nonresponse on income questions in panelh surveys: Incidence, imputation and the impact on inequality and mobility. Allgemeines Statistisches Archiv, 90 (1), 49-62.

[10] Geman, S. and D. Geman (1984): Stochastic Relaxation, Gibbs Distribution, and theBayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6 (6), 721-741.

[11] Graham, J. W. and J. L. Schafer (1999): On the performance of multiple imputation formultivariate data with small sample size. In: R. Hoyle (Ed.), Statistical Strategies for Small Sample Research, 1-29, Thousand Oaks, CA: Sage.

[12] Groves, R. M., D. A. Dillman, J. L. Eltinge, and R. J. A. Little (2002): Surveynonresponse. New York: Wiley. 41.

[13] Hastings, W. K. (1970): Monte Carlo Sampling Methods Using Markov Chain and Their Applications. Biometrika, 57, 97–109.

[14] Hoynes, H., M. Hurd, and H. Chand (1998): Household Wealth of the Elderly under Alternative Imputation Procedures. In: D. A. Wise (Ed.), Inquiries in the Economics of Aging, 229-257. Chicago: The University of Chicago Press.

[15] Hud, et. al. (2010): Data non-response. http://www.edu.

[16] Johnson, N. and S. Kotz (1970): Distributions in Statistics – Continuous Univariate Distributions. Vol. 2. New York: Wiley.

[17] Kennickell, A. B. (1998): Multiple Imputation in the Survey of Consumer Finances. Proceedings of the 1998 Joint Statistical Meetings, Dallas TX.

[18] Little, R. J. A. and D. B. Rubin (2002): Statistical Analysis with Missing Data. New York: Wiley.

[19] Little, R. J. A. and T. Raghunathan (1997): Should Imputation of Missing Data Condition on All Observed Variables? Proceedings of the Section on Survey Research Methods, Joint Statistical Meetings, Anaheim, California.

[20] Little, R. J. A., I. G. Sande, and F. Scheuren (1988): Missing-data adjustments in large surveys. Journal of Business and Economic Statistics, 6 (3), 117-131.

[21] Manski, C. (2005): Partial Identification with Missing Data: Concepts and Findings. International Journal of Approximate Reasoning, 39 (2-3), 151-165.

[22] Naem et al. (2010). Determinant of Households Demand for Electricity in District of Peshawar. European journal of social sciences-volume 14, number (2010).

[23] Orwa et al. (2006): Non-Response Weighting adjustment approach in survey sampling. East African Journal of Statistics. No 2 pp 143-162.

[24] Othuon, L. A. (2006): Bias in regression coefficient estimates upon different treatments of systematically missing data. East African Journal of Statistics. No 2 pp. 186-197.

[25] Pigot, T D, (2002). A review of Methods for missing data. Education research and evaluation, 7-353-385.

[26] Rancourt, E., Sarndal, C. E., and Lee, H (1994). Estimation of the variance in the presence of nearest neighbor imputation. In 1994 proceedings of the sectionon Survey Research Methods (pp. 888893). Alexandaria, VA: American Statistical Association.

[27] Rässler, S. and R. Riphahn (2006): Survey item nonresponse and its treatment. Allgemeines Statistisches Archiv, 90, 217-232.

[28] Riphahn, R. and O. Serfling (2004): Item Non-response on Income and Wealth Questions. Empirical Economics, 30 (2), 521-538.

[29] Rubin, D. B. (1987): Multiple Imputations for Non response in Surveys. New York: Wiley.

[30] Rubin, D. B. (1996): Multiple Imputation After 18+ Years. Journal of the American Statistical Association, 91 (434), 473-489.

[31] Rubin, D. B. and N. Schenker (1986): Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. Journal of the American Statistical Association, 81 (394), 366-374.

[32] Schafer, J. L. (1997): Analysis of incomplete multivariate data. London: Chapman & Hall.

[33] Siddharth T. Krishna M. Mayank R. Saurabh K. (2007): Implementing multiple imputation in an automatic variable Selection scenario. Inductis inc. 571 central Avenue New Jersey.

[34] Silverman, B. W. (1986): Density Estimation for Statistics and Data Analysis. London: Chapman and Hall.

[35] Smith, J. P. (1995): Racial and Ethnic Differences in Wealth. Journal of Human Resources, 30, 158-183.

[36] Strainer D. L. (2000). The case of missing Data: Method of dealing with drop outs and other research Vagaries, Canadian Journal of Psychiatry, 47, 68-75.

[37] Soley Bori M. (2013), Dealing with missing data: key assumptions and methods for applied analysis, Technical report No. 4.

[38] Wafula C. Otieno R. O., Mwenda M. M: Estimation of variance in the presence of Nearest Neighbour imputation. African Journal of Science and technology (AJST) Science and Engineering series Vol. 4, No 3, pp. 5-11.

[39] www. imputing missing data.