On New Techniques for Testing Incomplete Data Using Regression Models.

O.B. Aladeniyi, O.E. Olowofeso, and O.A. Fasoranbaku.

Department of Mathematical Sciences, School of Science,
Federal University of Technology, Akure, Nigeria.

E-mail: omoolabodunde@yahoo.com

ABSTRACT

In this study, the effect of missing values on a data set is investigated to see how it impacts the robustness of the estimated parameters obtained from the data. Three cases of missing data are considered which are: extreme, randomly distributed, and different cases. The cases are treated in phases of different sample sizes. For each phase, three existing techniques and two new techniques namely, $\sum \log x_i / n$ and $1/\sqrt{x_i}$.

The ordinary Least Squares (OLS) technique is used to estimate the parameters of the model from the complete data after the missing data have been replaced. Results presented in this paper are readily applicable in decision making especially when dealing with small sample size.

(Keywords: missing data, heteroscedasticity, regression model, explanatory variables, parameters)

INTRODUCTION

The phenomenon of missing data in a data set, whether small or large, is peculiar to all areas of research that deal with observed data from experiments and it has constituted a serious problem in many fields of research. An added complication is that the more data that are missing in a database, the more likely it is that one will need to address the problem of incomplete cases, yet those are precisely the situations where inputting or filling in values for the missing data points is most questionable due to the small proportion of valid data points relative to the size of the data matrix. The goal of statistical procedures must be clear to aid efficient and statistical inference that will enhance robust statistical modeling. When missing values are present in a data set, the probability of making valid and efficient inferences about a population of interest is low; this is vital to successful planning and implementation.

There exist large bodies of literature devoted to various techniques of solving missing data problems. Prominent among these are:

Cohen and Cohen (1983): Dummy Variables Approach; this approach indicate the use of dummy variables for identifying missing observations. However, it does not produce unbiased parameter estimates.

Little and Rubin (1987): Case Deletion otherwise known as Listwise Deletion method. In this approach, it was indicated that if a record has missing data for any variable used in particular analysis, omit that entire record from the analysis. If the discarded cases form a representative and relatively small portion of the entire dataset, this may indeed be a reasonable approach. It leads to valid inferences that implicitly assume that the discarded cases are like a random sub sample. This approach can result in different magnitudes or signs of casual or descriptive inferences. When the discarded cases differ systematically from the rest estimates, it may be seriously biased. Moreover, in multivariate problems, case deletion often results in a large portion of the data being discarded and an unacceptable loss of power.

Raghunathan (2004): Single Imputation method. In a similar development, this is an alternate approach based on filling in or imputing the missing values in the data set. Again, this approach can be traced back to survey practices adopted by the U.S. Bureau of Census, Ford (1983). This approach of filling in the missing

values is attractive for several reasons. First, imputation adjusts for differences between non-respondents and respondents on variables observed for both and included in the imputation process, as well as differences on variables not included in the model that are predicted by the model. Second, the complete data software can be used to process the data to obtain descriptive statistics and other statistical measures. Third, when a data set is being produced for analysis by the public, imputation by the data producer allows the incorporation of specialized knowledge about the reasons for missing data in the imputation procedure.

Raghunathan and Siscovick (1996) demonstrate that using an auxiliary variable in the imputation process can improve the efficiency considerably. Therefore, when the proportion of missing values is small, single imputation may be quite reasonable.

Although single imputation enjoys the positive attributes just mentioned, analysis of single imputed data tend to produce estimated standard errors that are too small, confidence intervals that are too narrow, and significant tests with p-values that are too small. Without special corrective measures, single imputation inference tends to overstate precision because it omits the between-imputation components of variability. Rao and Shao (1992) and Rao (1996) proposed the Jacknife method for computing correct standard errors in this approach.

Schafer and Graham (2002): Mean Substitution method; this is a process of replacing each missing value for a variable with the average of the observed values. This method may accurately predict missing data but distort estimated variances and correlations. A missing value cannot be properly evaluated apart from the modeling, estimation, or testing procedure in which it is embedded.

Rubin (1977): Multiple Imputation method; this approach generates actual raw data values suitable for filling in gaps in an existing database. Multiple imputation methods have been around for about two decades and are now the choice of most statisticians. It is a means of generating multiple simulated values for each incomplete datum, and then iteratively analyzing datasets with each simulated value substituted in turn. The purpose is, arguably, to generate estimates that better reflect true variability and uncertainty in the data than do regression methods. This was further improved on in Rubin (1987) and Schafer (1997) that Multiple Imputation replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. Therefore, it is a technique that seeks to retain the advantages of Single Imputation while also allowing the uncertainty due to imputation to be incorporated into the analysis.

This technique involves three distinct phases:

- The missing data are filled in *m* times to generate *m* complete data sets
- The *m* complete data sets are analyzed by using standard procedures
- The results from the *m* complete data sets are combined for inference.

Little and Rubin (1989); Little and Schenker (1995) say that inferences based on Multiple Imputation are more efficient than Listwise and they are not biased. It is worth nothing that the method of choice depends on the type of missing data pattern.

Rubin (1987): Regression Method; In this approach, a regression equation based on complete case data for a given variable will be developed, treating it as the outcome and using all other relevant variables as predictors. For cases where Y is missing, the available data is plugged into the regression equation as predictors, then substitute the equation predicting Y value into the database for use in other analyses. A regression model is fitted for each variable with missing values, with the previous variables as covariates. Based on the resulting model, a new regression model is then fitted and is used to impute the missing values for each variable. The process is repeated sequentially for variables with missing values.

Another existing method by David (2007): Expectation-Maximization algorithm, abbreviated as the EM algorithm which is one of the ways of obtaining maximum likelihood estimators. We will first estimate the parameters on the basis of the data we do have. Then we will estimate the missing data on the basis of those parameters. Then we will re-estimate the parameters based on the filled-in data, and so on. We would first

take estimates of the variances, co-variances and means. We would then use those estimates to solve for the regression coefficients, and then estimate missing data based on those regression coefficients (for example, we would use whatever data we have to estimate the regression $\hat{Y} = bX + a$, and then use X to estimate Y wherever it is missing). This is the "estimation step" of the algorithm. Having filled in missing data with these estimates, we would then use the complete data (including estimated values) to recalculate the regression coefficients. The EM algorithm gets around underestimating error by adding a bit of error to the variances it estimates, and then uses those new estimates to impute data, and so on until the solution stabilizes. At that point we have maximum likelihood estimates of the parameters, which is the "maximization step", and we can use those to make the final maximum likelihood estimates of the regression coefficients.

Though impressive, the solution techniques employed by these aforementioned studies are only capable of handling missing data problems when the sample size is large. However, situations arise when the sample size of an observation data is relatively small. To the best of authors' knowledge, studies concerning the effects of missing values on a data set when the sample is small are not found in literature. Thus, this paper assesses the problem of analyzing the effect of missing data on a data set when the sample size is small. In order to ensure a level of confidence in the results obtained in this study, Goldfeld-Quandt test is used to check for the presence of heteroscedasticity. The specific objective is to develop novel techniques or mechanisms which are capable of handling missing values problem for both large and small sample sizes.

PROBLEM STATEMENT

In this work, the model considered for the simulation is as defined below:

Let $Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + e_i$

Where, Y = dependent variable, x_i, i = 1, 2, 3 are the explanatory variables, b_i, i = 0, 1, 2, 3 are the parameters to be estimated, and e_i is the error term which is identically and independently normally distributed with mean zero and variance σ^2_i.

By solving the normal equations, we have estimators to be:

$$\hat{\beta} = \left(X'X\right)^{-1} X'Y$$

MATERIALS AND METHODS

The matrices of dimensions 31x4, 62x4, and 95x4 were used. These matrices were selected in order to determine the effect of various sample sizes on missing data and the ability to draw good inference based on the dimensional specification. The data used for this work were secondary data obtained from the *Statistical Digest on Nigerian Universities, 4th Edition*, a publication of National Universities Commission (NUC), Abuja, Nigeria.

Two different techniques were proposed in the study. The techniques are mean of log of the explanatory variables ($\sum \log x_i / n$), and the inverse of the root of the mean of the explanatory variables $1/\sqrt{\bar{x}_i}$. These techniques will be examined for Extreme, Randomly Distributed, and Different cases.

For the technique $\sum \log x_i / n$, the following procedures were employed:

(i) take the log of the incomplete data set;

(ii) find the mean;

(iii) replace the missing values with the mean and analyze the complete data set using SPSS computer package.

For the technique $1/\sqrt{\bar{x}_i}$, the following procedures were employed:

(i) find the mean of the incomplete data set;

(ii) take the inverse of the root of the mean

(iii) replace the missing values with the new values and analyze the complete data set using SPSS computer package.

The Ordinary Least Squares Method (OLS) was used for this research to estimate the complete data in which the best proposed technique has been used to replace the missing data. Furthermore, this was extended to generalized least square test for heteroscedasticity by considering the Goldfeld-Quandt.

Test For Heteroscedasticity
Goldfeld-Quandt Test

The following procedures were required; Firstly: order the observations according to the magnitude of the explanatory variable X.

Secondly: select arbitrarily a certain number (c) of central observations which will be omitted from the analysis. The remaining (n-c) observations were divided into two sub-samples of equal size [(n-c)/2].

Thirdly: fit separate regressions to each sub-sample and obtain the sum of squared residuals from each of them i.e $\sum e_1^2$ and $\sum e_2^2$

Each of these sums is divided by the appropriate degrees of freedom to obtain estimates of the variances of the $\mu's$ in two sub-samples. The ratio of the two variances:

$$F^* = \frac{\sum e_2^2 / [\{(n-c)/2\} - k]}{\sum e_1^2 / [\{(n-c)/2\} - k]} = \frac{\sum e_2^2}{\sum e_1^2}$$

has an F distribution with,

$v_1 = v_2 = [\{n-c\}/2] - k = [\{n-c-2k\}/2]$ degrees of freedom,

where n =total number of observations, c =central observations omitted i.e. ¼ of n, and k = number of parameters estimated from each regression.

The calculated F is compared with the theoretical value of F with $v_1 = v_2 = (n-c-2k)/2$ degrees of freedom at a chosen level of significance.

Fourth: if F* > F we have heteroscedasticity
if F* < F we have homoscedasticity

DISCUSSION AND ANALYSIS OF RESULTS

The results of the estimated parameters of models formulated for mean of Xi's, Expectation Maximization, Regression, $\sum \log x_i / n$ and $1/\sqrt{\bar{x}_i}$ with sample sizes of 31, 62 and 95 were presented below. The results of the heteroscedasticity were also presented.

For Table 1 when n is 31 under Phase I, the estimated parameters b_1, b_2, and b_3 are statistically significant at 0.05 for the existing techniques of Mean of X_i's, and Regression but statistically significant in only b_1 and b_2 for Expectation Maximization.

The value for b_0 is not statistically significant for all the existing techniques and b_3 not significant at 0.05 significance-level for Expectation Maximization. Also for the proposed technique, $\sum \log x_i / n$, b_1 and b_3 are statistically significant at 0.05. Also, b_0 for the proposed technique $1/\sqrt{\bar{x}_i}$ is statistically significant at 0.05 significance-level.

In Phases II and III, the estimated parameters b_1 and b_3 for both the existing methods and the proposed techniques are statistically significant at 0.05 significance-level. For all the techniques, the result shows that the remaining coefficients of b_0 and b_2 are not significant at 0.05 significance-level except for the proposed technique $1/\sqrt{\bar{x}_i}$ where b_0 is significant.

The R-squared obtained for the proposed techniques is 0.939 when missing data is randomly distributed. This shows that the goodness of fit at this point is better than the existing techniques.

The proposed technique of $1/\sqrt{\bar{x}_i}$ approach produced the minimum standard error for all the

estimated parameters which shows that it is more consistent and efficient than the other techniques.

Table 1 also shows the comparative analysis of the different phases. For Phase I which is the extreme cases, the R-squared value for Mean of X_i's and Regression is 0.897, while the remaining techniques of Expectation Maximization, $\sum \log x_i / n$ and $1/\sqrt{\bar{x}_i}$ have the following magnitudes 0.897, 0.48 and 0.149, respectively.

In addition, for Phase II the Randomly distributed cases, the existing techniques Mean of X_i's, Expectation Maximization and Regression have the same magnitude of 0.922 while the proposed techniques have the same magnitude of 0.939. Likewise, for Phase III the different cases, existing techniques have different magnitudes of R-squared values of 0.902, 0.906 and 0.907 respectively. While the proposed techniques $\sum \log x_i / n$ and $1/\sqrt{\bar{x}_i}$ have different magnitudes for R-squared 0.93 and 0.889 respectively.

From this analysis, it was also discovered that all the three existing techniques produced approximately the same estimated parameters and corresponding standard errors and R-squared values of $\sum \log x_i / n$ and $1/\sqrt{\bar{x}_i}$.

For Table 2 when n is 62 under Phase I, the estimated parameters b_1 and b_3 for the existing techniques Mean of X_i's, Expectation Maximization and Regression are statistically significant at 0.05. While the remaining coefficients b_0 and b_2 are not significant at 0.05 significance-level. Also for the proposed technique $\sum \log x_i / n$, all the estimated parameters b_0, b_1, b_2 and b_3 are statistically significant at 0.05. While b_0 and b_3 for the proposed technique $1/\sqrt{\bar{x}_i}$ is statistically significant at 0.05 significance-level.

In Phase II, the estimated parameters b_1 and b_3 for both the existing and proposed techniques are statistically significant at 0.05 significance-level. For all the techniques, the result shows that the remaining coefficients of b_0 and b_2 are not significant at 0.05 significance-level except for the proposed technique $1/\sqrt{\bar{x}_i}$ which is significant at b_0.

In Phase III, the estimated parameters b_1 and b_3 for Mean of X_i's, Expectation Maximization and Regression techniques are statistically significant at 0.05 significance-level. While for $\sum \log x_i / n$ and $1/\sqrt{\bar{x}_i}$, the estimated parameters b_0, b_1 and b_3 are statistically significant at 0.05 significance-level.

The R-squared obtained for $\sum \log x_i / n$ is 0.914 when missingness is randomly distributed. This shows that the goodness of fit at this point is better than the existing techniques.

It also produced the minimum standard error for all the estimated parameters which shows that, it is more consistent and efficient than the other techniques.

This Table 2 also shows the comparative analysis of the different phases. For Phase I which is the extreme cases, the R-squared values for the existing techniques Mean of X_i's, Expectation Maximization and Regression have the following magnitudes of 0.678, 0.717 and 0.71 respectively, while the proposed techniques $\sum \log x_i / n$ and $1/\sqrt{\bar{x}_i}$ have the following magnitudes 0.529 and 0.416 respectively.

In addition, for Phase II the randomly distributed cases, the existing techniques have the same magnitude of 0.868 while the proposed techniques have the following magnitude 0.914 and 0.865 respectively. Likewise, for Phase III the different cases, existing technique Mean of X_i's has the magnitude 0.869 while the remaining existing techniques have the same magnitude 0.87. The proposed techniques have different magnitudes for R-squared to be 0.88 and 0.847, respectively.

From this analysis, it was discovered that all the three existing techniques produced approximately the same estimated parameters and corresponding standard errors. However, there is significant difference in the estimated parameters, standard errors and R-squared values of the proposed techniques when missing data is at the extreme.

Table 1: The estimated parameters of models for Mean of X_i's, Expectation Maximization, Regression, $\sum \log x_i / n$ and $1/\sqrt{\bar{x}_i}$ when n is 31

Techniques of Analysis		PHASE I (Extreme cases)			PHASE II (Randomly distributed cases)			PHASE III (Different cases)		
		Extreme Cases	Standard Error	P-Value	Randomly distributed cases	Standard Error	P-Value	Different cases	Standard Error	P-Value
Existing Techniques	Mean of Xis	b_0=133.366 b_1=2.427 b_2=6.595 b_3=-1.382 R^2=89.9%	473.086 0.238 1.741 0.591	0.780 0.000* 0.001* 0.027*	b_0=5.058 b_1=1.734 b_2=0.768 b_3=1.122 R^2=92.2%	403.429 0.264 1.591 0.318	0.990 0.000* 0.633 0.002*	b_0=-114.417 b_1=1.670 b_2=0.594 b_3=1.154 R^2=90.2%	475.143 0.282 2.038 0.459	0.815 0.000* 0.773 0.018
	EM	b_0=89.247 b_1=2.388 b_2=6.496 b_3=-1.249 R^2=89.7%	475.156 0.251 1.892 0.655	0.852 0.000* 0.002* 0.067	b_0=6.730 b_1=1.724 b_2=0.737 b_3=1.134 R^2=92.2%	403.952 0.265 1.609 0.326	0.987 0.000* 0.651 0.002*	b_0=-103.492 b_1=1.708 b_2=0.452 b_3=1.143 R^2=90.6%	463.409 0.279 2.035 0.456	0.825 0.000* 0.826 0.018*
	Regression	b_0=109.791 b_1=2.395 b_2=6.621 b_3=-1.319 R^2=89.9%	472.160 0.244 1.820 0.622	0.818 0.000* 0.001* 0.043*	b_0=5.021 b_1=1.731 b_2=0.745 b_3=1.128 R^2=92.2%	403.634 0.264 1.597 0.321	0.990 0.000* 0.645 0.002*	b_0=-102.928 b_1=1.711 b_2=0.436 b_3=1.143 R^2=90.7%	462.930 0.279 2.020 0.454	0.826 0.000* 0.831 0.018*
Proposed Techniques	$\sum \log x_i / n$	b_0=1735.958 b_1=1.801 b_2=-6.292 b_3=2.053 R^2=48%	1039.035 0.617 3.679 0.616	0.106 0.007* 0.099 0.003*	b_0=330.631 b_1=1.661 b_2=1.876 b_3=0.894 R^2=93.9%	336.046 0.222 1.279 0.218	0.334 0.000* 0.154 0.000*	b_0=625.083 b_1=1.536 b_2=-0.934 b_3=1.495 R^2=93%	403.269 0.267 2.048 0.476	0.133 0.000* 0.652 0.004*
	$1/\sqrt{\bar{x}_i}$	b_0=4511.661 b_1=1.122 b_2=-6.242 b_3=1.549 R^2=14.9%	1310.783 0.862 5.255 0.872	0.002* 0.203 0.244 0.086	b_0=531.455 b_1=1.632 b_2=1.753 b_3=0.883 R^2=93.9%	296.015 0.217 1.254 0.215	0.083 0.000* 0.173 0.000*	b_0=994.567 b_1=1.294 b_2=-9.712 b_3=1.619 R^2=88.8%	393.647 0.304 2.544 0.561	0.017* 0.000* 0.781 0.001*

* P-value estimates that are significant at 0.05 significant level

Table 2: The estimated parameters of models for Mean of X_i's, Expectation Maximization, Regression, $\sum \log x_i / n$ and $1/\sqrt{\bar{x}_i}$ when n is 62.

Techniques of Analysis		PHASE I (Extreme cases)			PHASE II (Randomly distributed cases)			PHASE III (Different cases)		
		Extreme Cases	Standard Error	P-Value	Randomly distributed cases	Standard Error	P-Value	Different cases	Standard Error	P-Value
Existing Techniques	Mean of Xis	b_0=948.761 b_1=1.655 b_2=-1.820 b_3=-1.820 R^2=67.8%	550.786 0.431 2.926 0.506	0.090 0.000* 0.531 0.001*	b_0=309.646 b_1=1.318 b_2=0.872 b_3=1.434 R^2=86.8%	360.175 0.252 1.768 0.413	0.393 0.000* 0.624 0.001*	b_0=316.674 b_1=1.470 b_2=-0.376 b_3=1.570 R^2=86.9%	351.103 0.246 1.750 0.385	0.371 0.000* 0.831 0.000*
	EM	b_0=722.748 b_1=1.636 b_2=-1.058 b_3=1.748 R^2=71.7%	522.246 0.417 2.884 0.498	0.172 0.000* 0.715 0.001*	b_0=311.320 b_1=1.317 b_2=0.889 b_3=1.431 R^2=86.8%	359.763 0.252 1.788 0.416	0.390 0.000* 0.621 0.001*	b_0=324.446 b_1=1.469 b_2=-0.460 b_3=1.591 R^2=87%	350.291 0.246 1.753 0.386	0.358 0.000* 0.794 0.000*
	Regression	b_0=777.821 b_1=1.634 b_2=-1.491 b_3=1.831 R^2=71%	527.025 0.419 2.876 0.498	0.145 0.000* 0.606 0.001*	b_0=310.200 b_1=1.316 b_2=0.880 b_3=1.434 R^2=86.8%	360.499 0.253 1.785 0.416	0.393 0.000* 0.624 0.001*	b_0=322.370 b_1=1.469 b_2=-0.426 b_3=1.591 R^2=87%	350.555 0.246 1.751 0.386	0.362 0.000* 0.808 0.000*
Proposed Techniques	$\sum \log x_i / n$	b_0=1845.558 b_1=1.502 b_2=-6.388 b_3=2.412 R^2=52.9%	640.413 0.417 2.656 0.503	0.006* 0.001* 0.019* 0.000*	b_0=422.866 b_1=1.674 b_2=-1.132 b_3=1.467 R^2=91.4%	277.009 0.211 1.726 0.383	0.132 0.000* 0.514 0.000*	b_0=731.067 b_1=1.169 b_2=1.526 b_3=1.539 R^2=88%	314.279 0.186 0.988 0.216	0.024* 0.000* 0.128 0.000*
	$1/\sqrt{\bar{x}_i}$	b_0=2349.054 b_1=0.084 b_2=3.793 b_3=1.619 R^2=41.6%	696.147 0.505 2.426 0.599	0.001* 0.868 1.123 0.009*	b_0=856.374 b_1=0.867 b_2=-0.172 b_3=2.161 R^2=86.5%	335.287 0.188 1.621 0.360	0.013* 0.000* 0.916 0.000*	b_0=748.434 b_1=1.036 b_2=1.749 b_3=1.650 R^2=84.7%	356.886 0.188 1.132 0.249	0.040* 0.000* 0.128 0.000*

* P-value estimates that are significant at 0.05 significant level

Table 3: The estimated parameters of models for Mean of X_i's, Expectation Maximization, Regression, $\sum \log x_i / n$ and $1/\sqrt{\bar{x}_i}$ when n is 95

Techniques of Analysis		PHASE I (Extreme cases)			PHASE II (Randomly distributed cases)			PHASE III (Different cases)		
		Extreme Cases	Standard Error	P-Value	Randomly distributed cases	Standard Error	P-Value	Different cases	Standard Error	P-Value
Existing Techniques	Mean of Xis	b_0=701.040 b_1=1.389 b_2=1.721 b_3=1.317 R^2=75.3%	388.906 0.293 1.787 0.486	0.075 0.000* 0.338 0.008*	b_0=306.499 b_1=1.402 b_2=-1.536 b_3=2.107 R^2=91%	233.991 0.157 1.145 0.264	0.194 0.000* 0.183 0.000*	b_0=196.562 b_1=1.535 b_2=0.077 b_3=1.649 R^2=89.6%	254.154 0.172 1.013 0.244	0.441 0.000* 0.940 0.000*
	EM	b_0=603.389 b_1=1.453 b_2=1.232 b_3=1.404 R^2=77.2%	374.394 0.285 1.750 0.473	0.111 0.000* 0.483 0.004*	b_0=304.413 b_1=1.413 b_2=-1.577 b_3=2.103 R^2=91.1%	233.298 0.157 1.145 0.264	0.195 0.000* 0.172 0.000*	b_0=180.923 b_1=1.577 b_2=-0.034 b_3=1.625 R^2=90%	250.302 0.171 1.003 0.241	0.472 0.000* 0.973 0.000*
	Regression	b_0=603.240 b_1=1.445 b_2=1.485 b_3=1.346 R^2=77.2%	374.817 0.285 1.751 0.473	0.111 0.000* 0.399 0.006*	b_0=307.135 b_1=1.408 b_2=-1.583 b_3=2.110 R^2=91.1%	233.388 0.157 1.147 0.265	0.191 0.000* 0.171 0.000*	b_0=182.746 b_1=1.569 b_2=-0.010 b_3=1.629 R^2=89.9%	251.136 0.171 1.012 0.243	0.469 0.000* 0.992 0.000*
Proposed Techniques	$\sum \log x_i / n$	b_0=1875.974 b_1=0.289 b_2=2.123 b_3=2.054 R^2=55.7%	490.174 0.377 1.746 0.452	0.000* 0.445 0.227 0.000*	b_0=697.702 b_1=1.233 b_2=2.578 b_3=1.201 R^2=87.2%	266.602 0.178 0.962 0.217	0.010* 0.000* 0.009* 0.000*	b_0=633.180 b_1=1.431 b_2=1.985 b_3=1.111 R^2=88.1%	260.112 0.159 0.848 0.224	0.017* 0.000* 0.021* 0.000*
	$1/\sqrt{\bar{x}_i}$	b_0=1941.942 b_1=0.552 b_2=3.288 b_3=1.334 R^2=53%	504.801 0.390 2.030 0.475	0.000* 0.160 0.109 0.006*	b_0=621.331 b_1=1.369 b_2=3.036 b_3=0.914 R^2=84%	302.906 0.200 1.192 0.246	0.043* 0.000* 0.013* 0.000*	b_0=472.153 b_1=1.591 b_2=2.446 b_3=0.878 R^2=88.7%	254.674 0.156 0.886 0.231	0.067 0.000* 0.007* 0.000*

* P-value estimates that are significant at 0.05 significant level

For Table 3 when n is 95 Phase I, the estimated parameters b_1 and b_3 for the existing techniques Mean of X_i's, Expectation Maximization and Regression are statistically significant at 0.05. While the remaining coefficients b_0 and b_2 are not significant at 0.05 significance-level. Also for the proposed techniques $\sum \log x_i / n$ and $1/\sqrt{\bar{x}_i}$, b_0 and b_3 are statistically significant at 0.05. While b_1 and b_2 are not statistically significant at 0.05 significance-level.

In Phases II and III the estimated parameters b_1 and b_3 for the existing methods are statistically significant at 0.05 significance-level. While all the estimated parameters are significant at 0.05 significance-level for the proposed techniques, except for $1/\sqrt{\bar{x}_i}$ of Phase III that is not significant at 0.05 significance-level for b_0.

The R-squared obtained for both Expectation Maximization and Regression is 0.911 when missing data is randomly distributed. This shows that the goodness of fit at this point is better than the proposed techniques.

The Expectation Maximization approach produced the minimum standard error for all the estimated parameters which shows that, it is more consistent and efficient than the other techniques.

Table 3 also shows the comparative analysis of the different phases. For Phase I which is the extreme cases, the R-squared values for the existing techniques Mean of X_i's, Expectation Maximization and Regression have the following magnitudes 0.753 and 0.772 respectively, while the proposed techniques $\sum \log x_i / n$ and $1/\sqrt{\bar{x}_i}$ have the following 0.557 and 0.53 respectively.

In addition, for Phase II the randomly distributed cases, existing techniques have the same magnitude of 0.91 and 0.911, respectively, while the proposed techniques have the following magnitude 0.872 and 0.84 respectively. Likewise, for Phase III the different cases, existing techniques have different magnitudes 0.896, 0.90, and 0.899, respectively. The proposed techniques also have different magnitudes for R-squared 0.881 and 0.887, respectively.

From this analysis, it was discovered that all the three existing techniques produced approximately the same estimated parameters and corresponding standard errors. However, there is a significant difference in the estimated parameters, standard errors and R-squared values of the proposed techniques when missing data are at the extreme point of the data set. It was also noted that almost all the estimated variables under the proposed techniques were significant at 0.05 significance-level.

Test for Heteroscedasticity
The Goldfeld and Quandt Test

The Goldfeld and Quardt test yields the following results. By ordering the observations in ascending order of the X's, and omitting the twenty five central observations when n is 95, we are left with
two subsets of the data, one with the lower values of X with dimension 4x35 and one with the higher values of X with dimension 4x35.

Applying OLS to each subset we obtain:

(a) For subset 1

$\hat{y}_1 = 4649.853 + 48.38x_1 - 110.45x_2 - 33.48x_3$

$R^2 = 0.225$ and $\sum e_1^2 = 610703395.71$

(b) For subset 2

$\hat{y}_2 = 14469.84 - 2.38x_1 - 6.13x_2 + 1.78x_3$

$R^2 = 0.287$ and $\sum e_2^2 = 540255875.61$

We form the ratio of the two unexplained variations:

$$F^* = \frac{\sum e_2^2}{\sum e_1^2} = \frac{540255875.61}{610703395.71}$$

$= 0.88 \approx 1$

The theoretical value of F at the 5 percent level of significance with

$$v_1 = v_2 = \frac{n - c - 2k}{2} = \frac{95 - 25 - 2(4)}{2} = 32$$

F = 1.74 but F* = 0.88

:. F*<F as 0.88<1.74

Table 4: Subsets of data for Goldfeld and Quardt Test.

	Subset 1				Subset 2				
n1	Y1	X1	X2	X3	n2	Y2	X1	X2	X3
1	11986	252	42	100	1	1071	2684	435	1652
2	17511	278	45	107	2	1820	2741	436	1717
3	13900	355	58	145	3	12902	2789	437	1752
4	11120	355	61	147	4	12934	2831	446	1781
5	13175	357	61	149	5	15985	2831	448	1782
6	10668	384	67	197	6	14280	2831	449	1790
7	9263	396	73	197	7	15017	2905	450	1803
8	5154	412	83	197	8	12570	2921	456	1919
9	5178	431	83	335	9	10456	2959	457	1927
10	7014	436	89	335	10	7138	3109	468	1967
11	3799	441	89	335	11	6681	3139	475	2090
12	6638	472	92	362	12	8089	3139	478	2111
13	6972	472	92	366	13	4134	3250	502	2132
14	1999	519	92	373	14	8815	3308	519	2155
15	896	549	101	390	15	8317	3319	617	2177
16	1925	602	108	393	16	2432	3463	631	2397
17	779	609	111	398	17	1394	3590	642	2406
18	1306	609	125	412	18	2435	3820	646	2462
19	1227	609	126	435	19	1670	3870	716	2478
20	759	631	131	457	20	1849	3920	716	2527
21	6616	652	144	464	21	2106	3966	755	2994
22	2688	679	146	470	22	1556	3967	775	2994
23	3640	700	146	524	23	278	4028	784	3311
24	6328	702	151	527	24	9112	4295	879	3567
25	2582	704	155	535	25	6616	4327	886	3814
26	3700	715	157	542	26	3650	4327	899	3816
27	6641	734	159 .	549	27	3640	4369	912	3816
28	6655	739	160	553	28	9328	4528	920	3816
29	1153	753	165	561	29	3941	4542	1049	3914
30	1041	755	170	572	30	4620	4623	1129	4451
31	1820	811	171	579	31	4018	5031	1135	4452
32	12162	829	172	634	32	436	5094	1137	4528
33	12447	860	173	648	33	1484	656	1186	4779
34	14951	878	177	653	34	1071	6756	1186	4988
35	11120	995	183	660	35	1820	6756	1732	5674

The two variances are the same since the value of F* tends to 1. We can then conclude that the $\mu' s$ are homoscedastic.

CONCLUSION

In this study, the effect of missing data on regression models is investigated. New techniques namely $\sum \log x_i / n$ and $1/\sqrt{\bar{x}_i}$ are proposed to handle the problem of missing data. Results obtain from these new novel techniques perfectly agree with existing results. These two techniques have the advantage over the existing techniques in the sense that they can handle more efficiently the problem of missing data when the data set is small. Another interesting feature of these new techniques is that the problem of missing data in any multivariate data set can be handled efficiently. Results also show that when the missing values are randomly distributed the

best goodness of fit is obtained, while in the case of extreme missing values, the weakest goodness of fit is obtained.

Furthermore, Goldfield-Quandt test is used to test for the presence of heteroscedasticity in the missing data used. This test shows that there is no difference in the variances of sample data used. Thus, the two proposed techniques can be used to address the biasedness and the heteroscedasticity problems associated with estimated parameters of the model derived from incomplete data since the two techniques are found to be consistent with the existing techniques.

REFERENCES

1. Cohen, J. and Cohen, P. 1983. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (2nd Edition)*. Eribaum: Hillsdale, NJ.

2. Howell, D.C. 2006. *Statistical Methods for Psychology*. Wadsworth Publishing: New York, NY.

3. Little, R.J. A. and Schenker, N. 1995. "Missing Data". In: Arminger, Clogg and Sobel (eds). *Handbook of Statistical Modeling for the Social & Behavioral Sciences*. Plenum: New York, NY.

4. Little, R.J.A. and Rubin, D.B. 1987. *Statistical Analysis with Missing Data*. J. Wiley & Sons: New York, NY.

5. Raghunatha, T.E. 2004. "What Do We Do With Missing Data? Some Options for Analysis of Incomplete Data". *Annual Review of Public Health Journal*. 25:99-117.

6. Raghanathan, T.E. and Siscovick, D.S. 1996. "A Multiple Imputation Analysis of a Case-control Study of the Risk of Primary Cardinarrest among Pharmacologically Treated Hypertensive". *Applied Statistics*. 45:335-352

7. Rao, J.N.K. 1996. "On Variance Estimation with Imputes Survey Data". *Journal of the American Statistical Association*. 91

8. Rao, J.N.K. and Shao, J. 1992. "Jacknife Variance Estimation with Survey Data under Hot Deck Imputation". *Biometrika*. 79:811-822

9. Rubin, D.B. 1987. *Multiple Imputation for Non-response in Surveys*. Wiley: New York, NY.

10. Rubin, D.B. 1977. "Formalizing Subjective Notions about the Effect of Non-respondents in Sample Surveys". *Journal of the American Statistical Association*. 72: 538-543

11. Schafer, J.L. and Graham, J.W. 2002. "Missing Data: Our View of the State of the Art". *The American Psychological Association Incorporated*. 7(2):147-177.

12. Schafer, J.L. 1997. *Analysis of Incomplete Multivariate Data*. Chapman and Hall: London, UK.

13. Ford, B.N. 1983. *An Overview of Hot Deck Procedures*. U.S. Bureau of Census: Washington, D.C.

SUGGESTED CITATION

Aladeniyi, O.B., O.E. Olowofeso, and O.A. Fasoranbaku. 2009. "On New Techniques for Testing Incomplete Data Using Regression Models". *Pacific Journal of Science and Technology*. 10(1):149-160.


Pacific Journal of Science and Technology