Effects of Missing Value Estimation Methods in Correlation Matrix- A Case Study of Concrete Compressive Strength Data

Azizur Rahman*[1] and Ajit Kumar Majumder[2]

[1]*Lecturer, Department of Statistics, Jagannath University, Dhaka-1100, Bangladesh*
[2]*Professor, Department of Statistics,*
Jahangirnagar University, Savar, Dhaka, Bangladesh
rahman.aziz83@gmail.com, ajit@juniv.edu

Abstract

Concrete compressive strength is one of the most important factors leading to building construction, in the civil engineering context. While evaluating such data, quantitative analysis required. As it is known that, concrete as a non-homogeneous material, consists of separate phases. The more complicated the concrete, the higher is the compressive strength. But if missing value exists in the microstructure of concrete, then it may provide some unusual effect on the compressive strength of concrete. Thus it is required to deal with the analysis of missing values. In this study traditional and modern estimation techniques of missing values are performed and the effect of these methods on correlation matrix is observed along with their comparison. The result shows that, modern techniques provide efficient estimates compared to traditional method. .The analysis described here were undertaken in the SPSS 13.0 packages.

Keywords: *Missing values, Estimation of missing values, Correlation matrix*

1. Introduction

In most of statistical analysis, it is assumed that the data has been 'tidy'; that is normally distributed with no anomalous and/or missing results. However, in the real world, we often need to deal with 'messy' data for example data sets that contain missing values, unexpected extreme results or are skewed. No matter how well our experiments are planned, for example even in the context of concrete compressive strength data, there will always be times when something goes wrong, resulting in gaps in the data set. Some standard statistical procedures will not work as well or at all, with some data missing [1, 2] and [3]. It is come to know that concrete has been used as a construction material for more than a century. During this period of times, concrete has undergone a continuous development, *e.g.*, the growing use of secondary cementitious material in the binding phase. The use of binder admixtures in the production of concrete with enhanced performance (also known as High Performance Concrete or simply HPC) has received a great amount of attention recently. Concrete, as a non-homogeneous material, consists of separate phases; hydrated cement paste, transition zone and aggregate. Although most of the characteristics of concrete are associated with average characteristics of a component microstructure, the compressive strength are related with the weakest part of the microstructure. The more complicated the concrete, the higher is the compressive strength. Sometimes data of microstructure may contain missing values and provide some unusual effect on the compressive strength of concrete. Thus the best resources always to repeat the experiment to generate the complete data set. Sometimes, however, this is not feasible, particularly where reading/ measurements are taken at a time or cost of retesting is prohibitive, so alternative ways of addressing these problems are of great concern

[5]. The main ideas of missing values with their classification are necessary for this paper is presented in the next section.

2. Missing Values

In statistical analysis the phenomena of interest (i.e. missing value) is commonly represented by a rectangular $(n \times k)$ matrix $Y = y_{ij}$ where rows represent a sample of n observations, cases, or subjects. The column represents variables measured for each case. Each variable may be continuous or categorical such as amount of cement, concrete compressive strength etc. Some cells in such a matrix may be missing. It may happen if a measure is not collected or is not applicable [11]. There is several classification of missing values. These classifications influence the optimal strategy for working with missing values. At this stage, data that are missing completely at random (MCAR), missing at random (MAR) and non-ignorable (NI) missing values covered to get the in-depth idea of missing values classification.

2.1 Missing Completely At Random (MCAR)

The idea of values missing completely at random (MCAR) appears in every technical paper on missing values [6 and 7]. The term has a precise meaning [4 and 9]; thinking of data set as a large matrix, the missing values are randomly distributed throughout the matrix. Let us denote all values of the observations that are missing as Y_{mis} and the rest of Y_{obs}. Thus according to Little & Rubin [4] data are missing completely at random is a stronger condition if holds: $f\left(R/_{Y_{obs}, Y_{mis},\ \theta}\right) = f(R/\theta)$,

where, $R_{ij} = \begin{cases} 1;\ y_{ij}\ observed \\ 0;\ y_{ij}\ misssing \end{cases}$ denote the response indicator. The MCAR assumptions rarely hold in practice, it needs to be tested rarely.

2.2 Missing At Random (MAR)

The data are missing at random (MAR) according to Little & Rubin [4] if $f\left(R/_{Y_{obs}, Y_{mis},\ \theta}\right) = f(R/Y_{obs}, \theta)$, where Y_{mis}, Y_{obs}. and R are discussed in earlier classification. The MAR assumption allows the probability that a datum is missing to depend on the datum itself indirectly through quantities that are observed. For example, in our described data, the investigator might have less possibility of collecting a measure about a predictor from the planned experiment, resulting in higher likelihood that some of the values are missing. The MAR assumption would apply, because the predictor say 'fly Ash' explains the likelihood that the value will be missing. However, if we do not have a measure of fly Ash or simply do not include it in our estimation model, then we can say that the assumption is not satisfied. Another typical example, where the MAR assumption is not satisfied is personal income obtained via survey. It is well known that extreme values of personal income are less likely to be reported. Consequently, the MAR assumption is violated, unless the survey can reliably measure variables that are strongly related to income (an instrumental variable approach). Thus, the MAR assumption is valid if it can be assumed that the pattern of missing values is conditionally random, given the observed values in the mechanism variables. These variables that serve as mechanisms explaining missing may or may not be part of the theoretical model the researcher is using to explain the outcome variable.

2.3 Non Ignorable (NI) Missing Values

Data may be missing in ways that are neither MAR nor MCAR, but nevertheless are systematic. In a panel study of college students where an outcome variable is academic performance, there is likely to be attrition because the students who drop out of college and are lost to the study are more likely to have low scores on academic performance. Ways to model NI data are beyond the scope of this paper but are addressed in Muthen and Muthen [10].

The main objective of this study is to adopt the traditional and modern approaches of missing values, so that effect of missing values in correlation matrix using concrete data can be obtained effectively. Moreover comparison of existing modern approaches (say single imputation using EM only) over other traditional approaches carried out here.

In Section 3, for completeness, we briefly describe most of traditional techniques along with existed modern techniques or approaches for dealing with missing values. In Section 4, results of application of methodology are illustrated to the real world data. Here, we consider 99 concrete compressive strength data from experiments. By doing so, it was aimed to reveal the effect of different methodologies in correlation structure. Section 5, the concluding section summaries the results and make some recommendation.

3. Traditional Approaches of Missing Values

In this section we describe traditional approaches to working with missing values which include deletion techniques and imputation techniques such as list wise deletion, pair wise deletion and mean substitution, and inclusion of an indicator variable respectively.

3.1 Deletion Techniques

Deletion techniques remove some of the cases in order to compute the mean vector and the covariance matrix. Case wise deletion, complete case, or list wise deletion method is the most common solution to missing values. It is so common that it is the default in standard statistical packages. It is the simplest technique where all cases missing at least one observation are removed. This approach is applicable only when a small fraction of observations is discarded. If deleted cases do not represent a random sample from the entire population, the inference will be biased. Also, fewer cases result in less efficient inference [11]. Moreover, this is the usual way of dealing with missing data, but it does not guarantee correct answers. This is particularly so, in complex (multivariate) data sets where it is possible to end up deleting the majority of our data if the missing data are randomly distributed across cases and variables. Many researchers comment that this approach is conservative and that they do not want to "make up" data, but list wise deletion typically results in the loss of 20%-50% of the data. Of greater concern, it often addresses missing values in a systematic way [16].

Pair wise deletion or available case method retains all non missing cases for each pair of variables. We need at least three variables for this approach to be different from list wise deletion. For example, consider the simplest example where the first of three variables is missing in the first case and the remaining cases are complete. Then, the sample covariance matrix would use all cases for the sub matrix representing sample co variances of the second and third variables. The entry representing the sample variance of the first variable and sample co variances between the first and the remaining variables would use only complete cases. More generally, the sample covariance matrix is:

$S_{jk} = \frac{\sum_{jk} R_{ik}R_{ij}\left(y_j - \bar{y}_j^k\right)\left(y_k - \bar{y}_k^j\right)}{\sum_i R_{ij}R_{ik} - 1}$, where $\bar{y}_j^k = \sum_i R_{ij}R_{ik}y_{ij} / \sum_i R_{ij}R_{ik}$ and $R_{ij}\ and\ R_{ik}$ are indicators of missing values as defined in earlier section. Although such method uses more observations, it may lead to a covariance matrix that is not positive-definite and unsuitable for further analysis. Thus the pairwise deletion can be used as an alternative to case wise deletion in situations where parameters(in this paper we consider correlation coefficients, for example) are calculated on successive pairs of variables(e.g., in a civil engineering experiment we may be interested in the correlations between concrete compressive strength and cement, blast furnace slag, fly ash etc. With pair wise deletion, if one fly ash measurement was missing only this single pair would be deleted from the correlation and the correlations for compressive strength versus amount of cement and blast furnace slag would be unaffected [11].

3.2 Imputation Techniques

The substitution or imputation fill (impute) the values that are missing. Any standard analysis may then he done on the complete dataset. Many such techniques would typically provide underestimated standard errors. The simplest substitution technique fills in the average value over available cases (mean substitution). It replaces all missing data in a variable by the mean value for that variable. Though this looks as if the data set is now complete, mean substitution has its own disadvantages. The variability in the data set is artificially decreased in direct proportion to the number of missing data points, leading to underestimates of dispersion (the spread of the data). Mean substitution may also considerably change the values of some other statistics, such as linear regression statistics, particularly where correlations are strong [2]. This also underestimates variances and covariance in MCAR. Other substitution methods include group mean substitution that calculates means over groups of cases known to have homogeneous values within the group. A variation of group mean substitution when the group size is one is called hot-deck imputation. In hot-deck imputation for each case that has a missing value, a similar case is chosen at random. The missing value is then substituted using the value obtained from that case. Similarity may be measured using a Euclidean distance function for numeric variables that are most correlated with the variable that has a missing value. The following two reasons prevent us from recommending simple deletion and imputation methods when a substantial proportion of cases (more than 10 percent) are missing:

i. It is not clear when they do not work.

ii. They give incorrect precision estimates making them unsuitable for interval estimation and hypothesis testing.

As the percentage of missing data increases to higher levels, the assumptions and techniques have a more significant impact on results. Consequently, it becomes very important to use a model based technique with a carefully chosen model [16].

While there is no consensus among all experts about what techniques should be recommended, a fairly detailed set of recommendations is presented in [12 and 13], where factors such as proportion of missing data and the type of missing data (MCAR, MAR, NI) are considered. Roth [12] recommends using the simplest techniques, such as pair wise deletion, in the MCAR case and model based techniques when the MAR assumption does not hold or when the percent of missing data exceeds 15 percent. Because we doubt the validity of the MCAR assumption in most practical cases we do not recommend using techniques that rely o it unless the percent of missing data is small.

Although often used, none of the traditional approaches described is an optimal solution for missing values except under specialized circumstances. These approaches can result in serious biases in a positive or a negative direction, increase Type II errors, and underestimate correlations and β weights.

3.3 Modern Alternatives For Working With Missing Values

Several newer approaches for dealing with missing values exist, and most software programs now offer options that are more reasonable than the traditional approaches. Note that hot-deck imputation (discussed earlier) has long been available and has advantages over other traditional approaches, but is has rarely been used in family studies [20]. Expectation maximization (EM) as implemented in SPSS can impute a single new data set that has no missing values. Multiple imputations improves on this approach by using the consistency of estimations derived from multiple imputations as additional information, and it can estimate standard errors that are unbiased. A growing variety of software packages offer slightly different implementations of this approach. Structural equation modeling software and some multilevel software offer a full information maximum likelihood solution to missing values. In this approach, missing values are not imputed, but all observed information is used to produce the maximum likelihood estimation of parameters. Advocates of each approach are typically critics of alternatives, but often the criticisms have little consequence for practical data analysis. These approaches represent improvements over traditional approaches.

3.3.1 Single Imputation Using EM: EM is a maximum likelihood approach that can be used to create a new data set in which all missing values are imputed with maximum likelihood values. This approach is based on the observed relationships among all the variables and injects a degree of random error to reflect uncertainty of imputation. A explication is available in [14], and a short summary is available athttp://www.cs.brown.edu/research/ai/dynamics/tutorial/Documents/ExpectationMaximization.html. Here, values are imputed iteratively until successive iterations are sufficiently similar. Each successive iteration has are information because it utilizes the information from the preceding iteration. This iterative process is continued until the covariance matrix for the next iterations is virtually the same as that for the preceding iteration. This iterative maximum likelihood process usually converges quickly, but if there are many missing values and many variables, it can involve a great deal of computer time.

One way to do single imputation is to use a missing values module that is optional with the SPSS package. This SPSS MVA module will impute missing values using a variation of the EM approach. In addition to providing the imputed values, SPSS's implementation of EM provides useful information on patterns of missing data and differences between cases with and without imputed values [16].

3.3.2 Multiple Imputations: Single imputation using EM is an important advance over traditional approaches, but it has one inherent flaw. Because single imputation omits possible differences between multiple imputations, single imputation will tend to under-estimate the standard errors and thus overestimate the level of precision. Thus, single imputation gives the researcher more apparent power than the data justify. Multiple imputations (*m* separate data sets are imputed) allow pooling of the parameter estimates to obtain an improved parameter estimate. Multiple imputations produce a somewhat different solution for each imputation. If these *m* solutions were very similar, this would be evidence supporting the imputation. If these solutions differed markedly, however, then it is important to incorporate this uncertainty

into the standard errors. Multiple imputations allow a researcher to incorporate this missing data uncertainty.

Multiple imputations involve a three-step process. Depending on the software being used, the process can be tedious. At the first step, the package creates five to ten data sets using data augmentation. After generating such data set, at the second step, estimate the model (e.g., regression, logistic regression, SEM) separately for each of the five to ten data sets using data augmentation. At the final step, compute pooled estimates of the parameters and standard errors using the five to ten solutions. It is reasonable to expect that all major software packages will incorporate multiple imputation methods over the next few years. If a researcher does not have access to software that can handle multiple imputations in an integrated way, one solution is to do a single imputation for the preliminary analysis with that data set. Then, once the researcher is confident in the model, it is possible to use the multiple imputations only on the final model. The technical advantages of multiple imputations compared to single imputation are unarguable because multiple imputations allow for unbiased standard errors and single imputation does not. Now the question may be raised, How are the imputations combined? Each parameter estimate is simply the mean of the *m* estimates, where *m* is the number of replications. The standard error, however, incorporates the uncertainty by adding to the mean of the error variances the variance between the solutions. Simulations reported by Schafer [15] show that with the above number of imputations, when the MAR assumption is correct, multiple imputation is 94% as efficient as if there were no missing values when actually 30% of the values are missing. A similar efficiency is achieved with ten imputations, when 50% of the values are missing [16]. Here to combine the results of *m* analyses the following rules are used [17]. Denote the quantities of interest produced by the analyses as $P_1, \cdots, P_m$ and their estimated variances as $S_1, \cdots, S_m$.

- The overall estimate for P is an average value of $P_i's$: $\hat{P} = {\sum_i P_i}/{m}$;
- The overall estimate for S is $\hat{S} = {\sum_i S_i}/{m} + \frac{m+1}{m(m-1)}\sum_i \left(\hat{P} - P_i\right)^2$;

A refinement of the rules for small datasets is represented in [18]. Sometimes the inference is performed on multiple quantities simultaneously, for example, if we want to compare two nested multiple regression models, where the more general model has one or more extra parameters that are equal to zero in the simpler model. The rules for combining MI results in such a case are quite complicated [15].

3.3.3 Patterns of Missing Values: Various software packages provide information about the patterns of missing values. This information may indicate how many cases missed each possible combination of variables. Results are shown for each combination of two variables, three variables, four variables, and so on. SPSS's MVA module may give the most information, even providing t tests on differences in the means derived from imputed values and the means derived from observed values. Examining the patterns of missing values can be helpful. It provides a way to see whether there might be one or two problematic variables. Some programs show the proportion of data present for each pair of variables, but the more complex patterns tell the best way for working with missing values because they pinpoint where missing values are a problem [16].

4. Results of Empirical Application

Data on the concrete compressive strength plays important role in the building construction for civil engineering context. The data were obtained from [19]. A total of 99 data we collected from an experiment. A total of six factors relating to concrete compressive strength data were collected. The factors are cement (C1), blast furnace slag (C2), water (C3), super plasticizer (C4), coarse aggregate (C5), fine aggregate (C6) respectively. To measure the traditional and modern methodological effect on simple correlation matrix the missing values of some measurement were introduced. The following tables represent the application of such methodologies in concrete data set in case of missing values. The results are obtained through the use of SPSS13.0 version. Table 1 represents correlation matrix of data set having no missing values and it shows actual correlation structure of the data set. On the other hand Table 2 through Table 4 indicates the correlation matrix and provides the effects of different traditional and modern techniques of missing value estimates. In addition, Table 3.1 gives the pair wise frequency matrix used in calculation of correlation matrix applied in pair wise method. From these we observe that modern techniques give more efficient estimates results comparatively to the traditional methods.

Table 1. Correlation matrix of No Missing Data (99 cases)

	C2	C3	C4	C5	C6	Com Strength
C1	-.634	-.237	.386	-.336	-.085	.311
C2		-.178	.138	.140	.076	-.393
C3			-.844	-.025	-.789	.186
C4				.363	.614	-.135
C5					-.123	-.057
C6						.241

Table 2. Correlation matrix in case of List wise Deletion (only 71 cases remaining)

	C2	C3	C4	C5	C6	Comp Strength
C1	-.654	-.154	-.308	-.303	-.101	.209
C2		-.283	.238	.156	.161	-.518
C3			-.884	-.092	-.822	.292
C4				-.225	.659	-.255
C5					-.102	-.041
C6						-.155

Table 3. Correlation matrix in case of Pair wise Deletion

	C2	C3	C4	C5	C6	Comp Strength
C1	-.634	-.243	.376	-.290	-.100	.209
C2		-.222	.137	.189	.047	-.329
C3			-.835	-.069	-.783	.173
C4				-.338	.619	-.171
C5					-.128	.093
C6						-.249

Table 3.1. Matrix representation of Pair wise Frequency

	C2	C3	C4	C5	C6	Comp Strength
C1	92	92	95	91	94	93
C2		89	92	88	91	91
C3			92	88	91	90
C4				91	94	93
C5					90	89
C6						92

Table 4. Correlation matrix in case of Mean Substitution (99 cases)

	C2	C3	C4	C5	C6	Comp Strength
C1	-.623	-.236	.393	-.342	-.091	.278
C2		-.197	.122	.203	.073	-.330
C3			-.838	-.009	-.795	.193
C4				-.390	.605	-.165
C5					-.144	.021
C6						-.243

5. Conclusion

The results of the two most common traditional approaches and the modern imputation (EM) approaches, for the calculation of a correlation matrix, where the correlation coefficient(r) is determined for each approaches shows the increase, diminish or even reverse sign depending on which method is chosen to handle the missing data. Thus from the above analysis we may say that modern approaches of missing values provides efficient result than the traditional approaches and gives reliable value of the actual correlation coefficient matrix in case of full data set. Here some of the recommendations rely on the statistical processes and potential problems rather than on the particular empirical illustration used in this paper. There are three sets of recommendations: data management, less than ideal strategies and strategies to implement. The best solution is to minimize missing values when the data are

being collected. A researcher should explain how cases are dropped from analysis and the percentage of observations dropped by different approaches to working with missing values.

References

[1] S. Burk, "Scientific Data Management", vol. 1, Issue 1, **(2003)**, pp. 32-38**.**
[2] S. Burk, "Scientific Data Management", vol. 2, Issue 1, **(1998)**, pp. 36-41**.**
[3] S. Burk, "Scientific Data Management", vol. 2, Issue 2, **(1998)**, pp. 32-40**.**
[4] J. A. Little and D. B. Rubin, "Statistical Analysis With Missing Data", 2nd edition, John Wiley & Sons, **(2002).**
[5] E. Rasa, H. Ketabchi and M. H. Afshar, "Predicting Density and Compressive Strength of Concrete Cement Paste Containing Silica Fume Using Artificial Neural Networks", Transaction A: Civil Engineering, vol. 16, Issue 1, **(2007)** July, pp. 33-42**.**
[6] J. A. Little, "A test of missing completely at random for multivariate data with missing values", Journal of the American Statistical Association, vol. 83, Issue 404, **(1988)** December, pp. 1198-1202**.**
[7] J. Kim and J. Curry, "The treatment of missing data in multivariate analysis", Sociological Methods and Research, vol. 6, Issue 2, **(1977)** November, pp. 215-240**.**
[8] R. J. A. Little and D. B. Rubin, "Statistical Analysis with Missing Data", Willey Series in Probability and Mathematical Statistics, 2nd edition, John Willey & Sons **(1987).**
[9] D. B. Rubin, "Formalizing subjective notions about the effect of non respondents in sample surveys", Journal of the American Statistical Association, vol. 72, Issue 359, **(1977)** November, pp. 538-543**.**
[10] L. Muthen and B. Muthen, "Mplus user guide", Los Angeles; Statmodel.com, Version 3, **(2004).**
[11] M. H. Halstead, "Elements of Software Science", Elsevier North-Holland Inc., NY, **(1977).**
[12] P. L. Roth, "Missing data: A conceptual review for applied psychologist", Personel Psychology, vol. 47, **(1994)** April, pp. 537-560.
[13] R. Little and A. Hyonggin, "Robust likelihood-based analysis of multivariate data with missing values", Technical Report Working Paper 5, The University of Michigan Department of Biostatistics Working Paper Series, **(2003)** July.
[14] A. P. Dempster, M. N. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, vol. 39, series B, **(1977)** September, pp. 1-39.
[15] J. L. Schafer, "Analysis of incomplete multivariate data", 2nd edition, London: Chapman & Hall, **(1997)**.
[16] A. C. Acock, "Measurement error in secondary data analysis", Research in Sociology of education and socialization, vol. 8, **(1989)** December, pp. 201-230.
[17] D. B. Rubin, "Multiple imputation for survey non response", 2nd edition, New York: Wiley, **(1987)**.
[18] J. Barnard and D. B. Rubin, "Small sample degrees of freedom with multiple imputation", Biometrika, vol. 86, Issue 4, **(1999)** April, pp. 948-955.
[19] http://www.tc.gc.ca/eng/concrete/resources-researchstats-menue-848.html.
[20] I. G. Sande, "Hot-deck imputation procedures: Incomplete data in sample surveys", vol. 3, New York, Academic Press, **(1983)**.

Authors

Azizur Rahman

He is a Lecturer in Statistics department at Jagannath University, Dhaka-1000, Bangladesh. He finished his M. Sc. Study from Jahangirnagar University, in September, 2010, with the major in Statistics. His research interests focus on statistical theory and inference, Biostatistics, Neural Network methodologies, Econometrics, Time series analysis.

Ajit Kumar Majumder

He is a professor in Statistics department at Jahangirnagar University, Savar, Bangladesh. He finished his Phd degree from Monash University, Asustralia. He has worked for many years in statistical theory and application in different fields. His research interests focus on statistical theory and inference, Support vector machine, Neural Network methodologies, Econometric time series analysis.