SAS® GLOBAL FORUM 2020 #SASGF

Paper 5146 -2020

Analyzing Non-normal Data: Application to Missing Data Problems

Niloofar Ramezani, George Mason University

ABSTRACT

In many applications, the response variable is neither continuous nor normally distributed. If the outcome variable is binary, count, multinomial, or ordinal, the relationship between predictors and response variable is non-linear and using ordinary linear models developed for continuous data is inappropriate. Therefore, more advanced models adopting classification algorithms should be applied. Binomial, multinomial, and ordinal logistic models, as well as Poisson regression, for low-dimensional data and classification random forest for high-dimensional data are among robust predictive methods discussed for such scenarios. Starting with the simplest case of binary outcomes, through count, multinomial, and ordinal response variables, this study discusses various modeling options for low- and high-dimensional data while handling non-normal responses and missing data issues in SAS®. Three missing data techniques including multiple imputation are considered to appropriately account for the high percentages of missing observations, which are present in the majority of applied studies. Various techniques are discussed that can be applied to data with non-normal outcomes and missing observations. This paper discusses different options within SAS 9.4 for the aforementioned models using procedures such as PROC LOGISTIC, PROC GENMOD, PROC HPFOREST, and PROC STANDARD, PROC MI, and PROC MIANALYZE.

INTRODUCTION

In many data analysis scenarios, the response variable is not continuous or normally distributed. When that is the case, modeling the outcome using regular linear regression predictive models is not appropriate. When dealing with such response variables, generalized linear models (GLM) are appropriate predictive models to use for modeling the relationship between response and the predictors. Through GLM, different distributions and link functions can be used to accommodate varying types of response variables.

When modeling non-normal categorical responses, the robust GLM to use for modeling the relationship between categorical outcomes and different types of predictors, without assuming a linear relationship between them, is called logistic regression. In the presence of binomial outcomes, binary logistic models can be fitted using the logit link function. If the response variable is categorical, then another GLM that uses multinomial distribution and different link functions can be applied; this model is multinomial logistic regression. Ordinal logistic regression models have been applied in recent years in analyzing data with ranked multiple response outcomes. When modeling count data, assuming Poisson distribution, a Poisson regression may be fitted. Negative binomial distribution can also be adopted if there exists over-dispersion. If the number of variables or features is higher than the number of samples, the data are considered high-dimensional and different analytical methods are needed to handle non-normal data. One of the classification approaches which is appropriate for such scenarios is classification random forest, which is discussed in this paper.

Missing values that are present when dealing with real data most of the time add more complexity to binomial, multinomial, ordinal, and count models. Accounting for all the complexities mentioned above including non-normal response variables and missing data is important to build valid models. Some of these models and the appropriate methods of taking care of these complexities are discussed in this paper.

BINARY, CATEGORICAL, ORDINAL, AND COUNT DATA

When the outcome variable has two possible categories or levels, binary logistic regression is appropriate to model the association of the risk factors with the existence or absence of a desired event. In the presence of a categorical response variable with more than two possible outcomes, an extension of a binary logistic model needs to be used to account for multiple response categories. Multinomial logistic regression is an appropriate model which can be adopted for modeling categorical response variables with no order of the multiple outcomes. However, when analyzing data with ranked multiple response outcomes, ordinal logistic regression models have been applied in recent years (Ramezani, 2016).

LOGISTIC MODELS

Logistic regression allows building a predictive model between a categorical response variable and multiple input variables. Logistic regression, which is a GLM, helps predicting the presence or absence of a characteristic or outcome based on values of one or a set of predictors. The power of logistic regression over a liner model, through the addition of an appropriate link function to an ordinary linear model, is that the response variable does not need to follow a normal distribution while the predictors may be any combination of discrete and continuous variables. When the dependent variable is binary following a binomial distribution, the logit link function is widely used within the GLM, making the predictive model a binary logistic regression (Atkinson & Massari, 1998). Logistic regression coefficients can be used to estimate ratios for each of the independent variables in the model (Lee, 2005).

Binary Logistic Regression

In Logistic Regression, instead of measuring the average change of the response, as it is done in linear regression models, probability of the outcome is measured by the odds of occurrence of an event. Change in probability of the occurrence of one category of the response (Y) is not constant (linear) with constant changes in each predictor (X). This means the probability of success ($Y = 1$), given X, is a non-linear function, usually a logistic function (Ramezani, 2019). The most common form of logistic regression uses the logit link function so it is easily understandable to show the logistic regression equation as

$$logit\ (p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k,$$

where p is the probability of success and $logit(p)$ is the natural logarithm of probability of success over probability of failure. PROC LOGISTIC and PROC GENMOD are two of the SAS procedures that can be adopted to fit a binary logistic regression model. The call to PROC LOGISTIC can be written as below:

```
PROC LOGISTIC DATA=(mention the dataset name here);
        CLASS (list the categorical variables here)/PARAM=REF;
        MODEL (specify the response variable here) (EVENT='1')=(list all the predictor
        variables used here)/LACKFIT CORRB;
RUN;
```

The CLASS statement is to specify all categorical variables used in the model. PARAM=REF option will result in creating dummy variables for a categorical variable as oppose to the default within the LOGISTIC procedure which is effect coding. MODEL statement is where the response variable is specified followed by the predictors after an equal sign. Using the EVENT=’1’ option in the model statement, defines the event of interest to be modeled as 1, which generally is used to show the occurrence of the event. 1 can be replaced with 0 if modeling the nonoccurrence is of interest. Modeling category 1 as the event of interest instead of the nonoccurrence category, 0, is also possible by using the DESCENDING option in the PROC LOGISTIC statement. LACKFIT is another option that can be added within the model statement to test the goodness of fit of the model using Hosmer-Lemeshow test. This test will check to see if there is any difference between the observed and predicted values of the response variable (Ramezani, 2019).

PROC GENMOD is another SAS procedure that can be used to perform a similar binary logistic regression as below:

```
PROC GENMOD DATA=(mention the dataset name here) DESCENDING;
        CLASS (list the categorical variables here);
        MODEL (specify the response variable here)=(list all the predictor variables
        here)/DIST=BIN CORRB;
RUN;
```

Within this procedure, specifying DIST=BIN is needed to impose performing a binary logistic regression model. This is because the GENMOD procedure can be used for performing a few other models too. As a result, he distribution of the response needs to be specified to define the type of the performed model. Within the GENMOD procedure specifying that we want to model 1 as the event of interest instead of 0 for the dependent variable can be done by simply adding the DESCENDING option to the procedure (Ramezani, 2019).

Multinomial and Ordinal Logistic Regression

The multinomial logistic regression model, which can be considered as an extension of the binomial logistic regression model, is a proper model to be used when the dependent variable has more than two nominal, yet unordered, categories. When the response categories are not ordered, procedures such as PROC LOGISTIC, with the specification of LINK=GLOGIT option in the MODEL statement, can be used to fit a multinomial logistic regression. A sample code is shown below:

```
PROC LOGISTIC DATA=(mention the dataset name here);
        CLASS (specify the response variable here)(REF="1") (list the categorical predictor
        variables here)/PARAM=REF;
        MODEL (specify the response variable here)=(list the predictor variables
        here)/LINK=GLOGIT;
RUN;
```

PROC SURVEYLOGISTIC with the specification of LINK=GLOGIT option can also be used to perform the same analysis. The GLIMMIX and HPGENSELECT procedures can also be used to fit this model by specifying the DIST=MULT and LINK=GLOGIT options in the MODEL statement. All of the aforementioned procedures fit the model using maximum likelihood estimation. PROC CATMOD can also be used to fit the multinomial logistic model using maximum likelihood by default or using weighted least squares after specifying the WLS option (Ramezani, 2019).

As discussed in Ramezani (2019), when the response categories are ordered, a multinomial regression model still can be used but it will result in the loss of some information about the way the response categories are ordered (Agresti, 2007). A better choice, when modeling data with ranked multiple response outcomes, is ordinal logistic regression model, which preserves the ordering related information, but the model is slightly more complex (Ramezani, 2015). Due to this complexity, assumptions validation, and limitations of modeling options offered by statistical packages, the use of such models for ordered information is still rare. Consequently, it is important to highlight the importance of such models that will result in robust estimates when dealing with ordinal responses.

In ordinal logistic regression models, there are different logit functions that should be used within the GLM framework. Cumulative logit, adjacent–categories logit, and continuation ratio

logit are briefly explained below, based on the description of Ramezani (2015), and notations used in Agresti (2007).
The cumulative logit function used in the ordinal logistic models basically models categories $\leq j$ versus categories $> j$, where j is the cut-off point category decided by the researcher based on the research question (Hosmer & Lemeshow, 2013). To fit this model using a cumulative logit function, PROC LOGISTIC and PROC GENMOD may be used as below:

```
PROC LOGISTIC DATA=(mention the dataset name here);
        CLASS (specify the response variable here)(REF="1") (list the categorical predictor
        variables here)/PARAM=REF;
        MODEL (specify the response variable here)=(list the predictor variables
        here)/LINK=CLOGIT SCALE=NONE AGGREGATE RSQ LACKFIT;
RUN;
```

Within the model statement of the PROC LOGISTIC, LINK=CLOGIT is added to enforce the using of the cumulative logit link function.

PROC GENMOD may also be used to fit the same model as below:

```
PROC GENMOD DATA= (mention the dataset name here) RORDER=data DESCENDING;
        CLASS (specify the response variables here)(REF="1") (list the categorical predictor
        variables here);
        MODEL (specify the response variables here)=(list the predictor variables
        here)/DIST=MULTINOMIAL LINK=CUMLOGIT;
RUN;
```

Within the MODEL statement of the PROC GENMOD, DIST=MULTINOMIAL is added to enforce the use of the multinomial distribution for the categorical outcome variable and the LINK=CUMLOGIT specifies the adoption of the cumulative logit link function in the ordinal logistic regression model.

The adjacent-categories logit function used in ordinal logistic models is to model two adjacent categories**.** They will use local odds ratios for interpretations, whereas within the cumulative logit models, the entire response scale is used for the model and therefore, cumulative odds ratio is used for interpreting the results (Ramezani, 2019).
Fitting an ordinal logistic regression with adjacent categories logit function in SAS is not as straight forward as when cumulative logit link is used. There still is not a SAS built-in procedure for this type of analysis. Requiring more effort, PROC NLMIXED can be used to perform the adjacent categories logit model. The likelihood functions need to be defined within the NLMIXED procedure, which can be time consuming specially in the presence of several independent variables in the model. Using PROC CATMOD to perform this type of analysis is also recommended in some books including Allison (2012) but it causes some issues in the output reported by this procedure (Ramezani, 2015).

The continuation-ratio logit function can also be used in ordinal multinomial logistic models**.** This model is useful when a sequential mechanism determines the response outcome (Agresti, 2007). Mechanisms like survival through various age periods of subjects would be suitable for such models. For more details and examples about the application of these models to real data, see Ramezani (2016).
Fitting an ordinal logistic regression with continuation-ratio logit function is not easy either due to not having a built in procedure in SAS to perform this type of analysis. There exists some issues when running models using a continuation-ratio logit function using PROC CATMOD. Agresti (2013) suggests using PROC GENMOD for the continuation ratio logit models that performs better than PROC CATMOD, yet is not easy to use. One cans also use PROC LOGISTIC and then restructure the original dataset to create binary response (Allison, 2012). Having this new binary outcome, PROC LOGISTIC produces can provide the same results as NLMIXED. Within this procedure the PARAM=GLM coding in the CLASS statement should be used (High, 2013). For more details, see Ramezani and Ramezani (2016).

COUNT MODEL

When working with count data, one wishes to predict the number of certain events of occurrences of any outcome of interest for instance within a fixed period of time. Such data do not follow a normal distribution; therefore, appropriate GLM and distribution should be applied. Poisson regression is often used for modeling count data. Negative binomial regression is another model that can be used for over-dispersed count data, that is when the conditional variance exceeds the conditional mean. It is similar to Poisson regression but has an extra parameter to model the over-dispersion. If the conditional distribution of the outcome variable is over-dispersed, the confidence intervals for Negative binomial regression are likely to be narrower as compared to those from a Poisson regression.

PROC GENMOD is usually used for Poisson regression analysis in SAS as below:

```
PROC GENMOD DATA=(mention the dataset name here);
        CLASS (list categorical variables here)/PARAM=glm;
        MODEL (specify the count response variables here)=(list the predictor variables
        here)/TYPE3 DIST= POISSON;
        STORE par;
RUN;
```

Like previous models, the CLASS statement is to list the categorical variables. The global option PARAM = glm is added to specify the GLM model and to save the model using the STORE statement for future post estimations. The TYPE3 option in the model statement is used to get the test of the categorical variables listed on the class statement. The DIST=POISSON option is used to indicate that a Poisson distribution should be used in this GLM. Statement STORE allows us to store the parameter estimates to a data set, which we call par, so we can perform post estimation without rerunning the model.

HIGH-DIMENSIONAL NON-NORMAL DATA

When dealing with high dimensional data, methods such as principal component analysis, random forest(s) are highly recommended (Wickham & Grolemund, 2016). Radom forest models can be applied to both categorical and continuous response variables in two forms of classification random forest and regression random forest, respectively. Breiman (2001) proposed random forest (or random forests), which uses a group of decision trees. Within random forests, in addition to constructing each tree using a different bootstrap sample of the data, how the classification or regression trees are constructed are altered in a way to build the most optimal group of trees for each data set (Ramezani, 2019). This strategy turns out to perform fairly well compared to many other classifiers, including discriminant analysis, support vector machines and neural networks, and is robust against overfitting of the data (Liaw & Wiener, 2002).

DECISION TREES AND RANDOM FOREST

Random forest is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the output of this class by individual trees. A tree is called a classification tree when the target variable is categorical and it is called a regression tree when the target variable is continuous. When classification decision trees are used within the forest building, the respective random forest is referred to as classification random forest (Hartshorn, 2016). This is the only type of random forests discussed here since this paper is dedicated to modeling non-continuous response variables. For details on both types of random forest models, please see Ramezani (2019).

In order to run a random forest in SAS, PROC HPFOREST can be used. Using this procedure, the response variable and their nature needs to be specified. TARGET is to list the target or response variable and INPUT is to specify the input or predictor variables. LEVEL after each of these commands should be added to specify the type of the variable. The LEVEL option for the INPUT statement defines the predictors as binary, nominal, or ordinal. LEVEL=binary is to specify a binary variable where there are two categories of the variable, LEVEL=nominal is for noting that the variable is un-ordered categorical mainly for more than two categories of the variable, and LEVEL=ordinal to specify the categorical variable where there exists an order among the categories.

```
PROC HPFOREST DATA=(mention the dataset name here);
        TARGET (response variable name should be mentioned here)/LEVEL=(specify type of
        the variable here);
        INPUT (list the categorical variables here)/LEVEL=(specify type of the variables here;
        nominal should be specified for categorical variables);
RUN;
```

In the presence of different types of predictor variables, instead of writing multiple lines of INPUT statement, the INPUT lines can be condensed into one line to include the list of all the variables together as below:

```
input <predictor1, predictor2, predictor3>/
level= <binary, nominal, ordinal>;
```

After running the code, a series of tables will appear in the SAS output that include the model information such as, number of randomly selected variables to test each node or possible split in each tree and the maximum number of trees. By default, the “Inbag Fraction” is set at 60% and the out of bag (OBB) is at a rate of 40%. The “Prune Fraction” is defaulted at “0″ and the closer it is set to “1″, the lower the level of growth the tree will have.

HPFOREST automatically uses only the valid variables that have no missing records under any of the observations. Table of “Baseline Fit Statistics” will include information such as the percentage of mis-classification the model identified. 100 percent minus this percentage, or the remaining percentage, will be the rate of accurate classification the random forest identifies. The higher percentage of the accurate classification shows that the majority of the sample has been classified correctly in each of the randomly selected samples. The “Fit Statistics” table of the decision trees shows the result of the analysis of the fitness of the trees within the forest. The column “Miscalculation Rate (Train)” shows the miscalculation rate for each tree and the rate tends to decrease by the increase in the tree number. The closer to the bottom of the table is where the minimum miscalculation rate is generally observed. Another table the model generates is the “Loss Reduction Variable Importance” table, which outlines the rank of importance of variables in terms of how each variable contributes to the predictability of the model (Ramezani, 2019).

By performing random forest as a data-mining algorithm, we can select important explanatory variables in the process of predicting the outcome, response, or target variable. In addition, this exercise allows us to use a combination of categorical and quantitative variables. In sum, this forest lets us know which variables are important, but not the relationship to one another.

MISSING DATA

Missing data presents a challenge in data analysis and research and is associated with many statistical concerns (Cheema, 2014). The severity of the impact of missing data on the results depends on the type of missingness (Rubin, 1987) in addition to the quantity of missing data (Gibson & Olejnik, 2003). Many missing data handling procedures are available to researchers, but the procedures vary in regards to overall effectiveness and technical skill required for implementation (Gibson & Olejnik, 2003). Methods such as listwise deletion, mean imputation and multiple imputation are some of these methods.

Listwise deletion deletes any individual in a data set that involves missing data on any of the variables used in the study. Listwise deletion is the most common missing data handling procedure in different fields of research (Cheema, 2014) because it is very easy to use and is often the default in statistical packages. However, it can result in a loss in power, especially if missing values are distributed across several variables (Schafer & Olsen, 1998). This missing data handling procedure can also bias parameter estimates if data is missing at random (MAR) or missing not at random (MNAR) (Roth, 1994).

Mean imputation is another technique of handling missing data that is known as the easiest imputation technique. Within this method, each missing value in a variable is replaced with the mean of the observed values for that variable. As easy as it is to use, this method will result in a very small variation in each variable due to using the same value instead of each missing observation. Due to this under-estimation of the variance, mean imputation is not recommended and the user should be aware of the implications (Buuren & Groothuis-Oudshoorn, 2011).

Here is an example of performing a mean imputation. The ordinal logistic model example with cumulative logit from above is used here and the only thing that has been added to it is the PROC STANDARD to perform the mean imputation within the regression models. The STANDARD procedure outputs the new mean imputed data (outdata here) which should be used within the PROC LOGISTIC as the inputted data set. The SAS code can be written as below:

```
PROC STANDARD DATA=Data OUT=outdata REPLACE;
RUN;

PROC LOGISTIC DATA= outdata DESCENDING;
        CLASS DV (REF="1") IV1 IV2 / PARAM = REF;
        MODEL DV= IV1 IV2 IV3 IV4 / LINK=CLOGIT SCALE=NONE AGGREGATE RSQ
        LACKFIT;
RUN;
```

Multiple imputation is the recommended technique when dealing with data sets with missing values. It is a popular and useful way of handling missing data under MAR assumption (Little and Rubin, 2002). Instead of filling in a single value for each missing value, like it is done the mean imputation, within Rubin’s (1987) multiple imputation method, each missing value is replaced with a set of plausible values representing the uncertainty about the right value to impute (Yuan, 2010). Multiple imputation results in accurate estimates of the standard errors while the precision of the study associations is commonly over-estimated with a single imputation due to obtaining very low estimates of the standard error (Koopman, Heijden, Grobbee, & Rovers, 2008). In multiple imputation, the missing data are stochastically imputed multiple times. In the commonest approach, the multiple completed data sets are then analyzed using methods appropriate for complete data sets, then the multiple results are combined using Rubin's rule and a single set of output is created (Rubin, 1987). More modern approaches such as multiple imputation and full information maximum likelihood are preferable to traditional approaches such as listwise deletion (Buhi & Goodson, 2008). Table 1 summarizes these missing data handling methods with the appropriate SAS procedure to perform them.

Missing Data Handling Technique	**SAS Procedure**
Listwise Deletion	Default
Mean Imputation	PROC STANDARD
Multiple Imputation	PROC MI/PROC MIANALYZE

Table 1. Missing Data and Appropriate SAS Procedures

To perform the multiple imputation as the missing data handling technique for the same analysis as the above mean imputation method was performed, the SAS code is provided as below in three steps:

```
PROC MI DATA=Data NIMPUTE=10 SEED=454 OUT=outimputedex1;
RUN;

PROC LOGISTIC DATA=outimputedex1 DESCENDING OUTEST=outreg;
CLASS DV (ref="1") IV1 IV2 / PARAM = ref;
MODEL DV= IV1 IV2 IV3 IV4 / LINK=clogit SCALE=none AGGREGATE RSQ LACKFIT;
BY _imputation_;
ODS OUTPUT ParameterEstimates=lgsparms;
RUN;

PROC MIANALYZE PARMS=lgsparms;
    MODELEFFECTS Intercept IV1 IV2 IV3 IV4;
RUN;
```

In this SAS code, there are three main steps in performing a multiple imputation. First, using PROC MI to impute data, then running the actual analysis (i.e., PROC LOGISTIC), and finally PROC MIANALYZE to pool the results from all imputations together and get the final results. Unfortunately, PROC MI/PROC MIANALYZE is not compatible with ordinal models using PROC LOGISTIC and it will cause some issues when outputting the results but it is compatible with the rest of the models discussed in this paper. The number of imputations can be specified using NIMPUTE in the PROC MI statement. Intercept and predictors, which their coefficients need to be estimated, should be specified in the MODELEFFECTS statement in the MIANALYZE procedure.
This code will result in ten imputed data sets and will use Rubin's rule to give the final estimates at the end.

CONCLUSION

Different options for modeling non-normal non-continuous responses or target variables were discussed above. These response variable types include binary, multinomial, ordinal, and count. Procedures such as PROC LOGISTIC and PROC GENMOD can be used to perform binary logistic models. PROC LOGISTIC, PROC GENMOD and PROC NLMIXED can be used to fit multinomial and ordinal logistic models. PROC GENMOD can be used to appropriately model count outcomes. Using appropriate models such as random forests, high dimensions of such response variables can be modeled using PROC HPFOREST.

At the end, some missing data handling techniques were introduced as there exist missing data within any of the models mentioned above that needs to be considered in order to get unbiased results. The use of multiple imputation techniques are recommended rather than the more commonly used or preferred methods such as listwise deletion. More work need to be done in developing SAS procedures that can easily fit ordinal logistic regression models specifically when trying to use adjacent-categories and continuation-ratio logit functions.

REFERENCES

Agresti, A. (2007), An Introduction To Categorical Data Analysis. Willey Series in Probability and Statistics, second edition.

Agresti, A. (2013). Categorical data analysis (3rd ed.). New York: Willey.

Allison, P. D. (2012). *Logistic regression using SAS: Theory and application*. SAS Institute.

Atkinson, P. M., & Massari, R. (1998). Generalised linear modelling of susceptibility to landsliding in the central Apennines, Italy. *Computers & Geosciences*, *24*(4), 373-385.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Buhi, E. R., Goodson, P., & Neilands, T. B. (2008). Out of sight, not out of mind: strategies for handling missing data. American journal of health behavior, 32(1), 83-92.

Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of statistical software, 45(3).

Cheema, J. R. (2014). A Review of Missing Data Handling Methods in Education Research. Review of Educational Research, 84(4), 487-508.

Gibson, N. M., & Olejnik, S. (2003). Treatment of missing data at the second level of hierarchical linear models. Educational and Psychological Measurement, 63(2), 204-238.

Hartshorn, S. (2016). Machine Learning With Random Forests And Decision Trees: A Visual Guide For Beginners. *Kindle Edition*.

High, R., Models for Ordinal Response Data (2013). SAS Global Forum, Paper 445-2013

Hosmer, D., Lemeshow, S., Applied Logistic Regression (2013). Willey Series in Probability and Statistics, third edition.

Koopman, L., van der Heijden, G. J., Grobbee, D. E., & Rovers, M. M. (2008). Comparison of methods of handling missing data in individual patient data meta-analyses: an empirical example on antibiotics in children with acute otitis media. American journal of epidemiology, 167(5), 540-545.

Lee, S. (2005). Application of logistic regression model and its validation for landslide susceptibility mapping using GIS and remote sensing data. *International Journal of Remote Sensing*, *26*(7), 1477-1491.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

Little, R. J. A., and Rubin, D. B. (2002), Statistical Analysis With Missing Data (2nd ed.), New York: Wiley.

Ramezani, N. (2015). Approaches for missing data in ordinal multinomial models*. In JSM Proceedings,* Biometrics section, New Methods for Studies with Missing Data Session. Alexandria, VA: American Statistical Association Journal.

Ramezani, N. (2016). Analyzing non-normal binomial and categorical response variables under varying data conditions. *In proceedings of the SAS Global Forum Conference.* Cary, NC: SAS Institute Inc.

Ramezani, N., Ramezani, A. (2016). Analyzing non-normal data with categorical response variables. In proceedings of the Southeast SAS Users Group Conference. Cary, NC: SAS Institute Inc.

Ramezani, N. (2019). Advanced Statistical Modeling within Machine Learning and Big Data Analytics. In proceedings of the Southeast SAS Users Group Conference. Cary, NC: SAS Institute Inc.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. Personnel psychology, 47(3), 537-560.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. John Wiley & Sons.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate behavioral research, 33(4), 545-571.

Wickham, H., & Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.".

Yuan, Y. C. (2010). Multiple imputation for missing data: Concepts and new development (Version 9.0). SAS Institute Inc, Rockville, MD, 49.

RECOMMENDED READING

- *Allison, P. D. (2012). Logistic regression using SAS: Theory and application. SAS Institute*
- *Ramezani, N. (2015). Approaches for Missing Data in Ordinal Multinomial Models. In JSM Proceedings, Biometrics Section. Alexandria, VA: American Statistical Association, pp. 2809-2823.*
- *Ramezani, N. (2016). Analyzing non-normal binomial and categorical response variables under varying data conditions. In proceedings of the SAS Global Forum Conference. Cary, NC: SAS Institute Inc.*
- *Ramezani, N. (2016). How to analyze correlated and longitudinal data?. In proceedings of the Western Users of SAS Software Conference. Cary, NC: SAS Institute Inc.*
- Ramezani, N. (2019). Advanced Statistical Modeling within Machine Learning and Big Data Analytics. In proceedings of the Southeast SAS Users Group Conference. Cary, NC: SAS Institute Inc.
- Ramezani, N. (2020). Modern Statistical Modeling in Machine Learning and Big Data Analytics: Statistical Models for Continuous and Categorical Variables. In Handbook of Research on Big Data Clustering and Machine Learning (pp. 135-151). IGI Global.
- Ramezani, N., Ramezani, A. (2016). Analyzing non-normal data with categorical response variables. In proceedings of the Southeast SAS Users Group Conference. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Niloofar Ramezani
nramezan@gmu.edu