

Analyzing non-normal data with categorical response variables

Niloofer Ramezani, University of Northern Colorado; Ali Ramezani, Allameh Tabataba'i University

Abstract

In many applications, the response variable is not normally distributed. If the response variable is categorical with two or more possible responses, it makes no sense to model the outcome as normal. When dealing with categorical outcome variables, the relationship between the outcome and predictors is not linear anymore, hence more advanced models than general linear models need to be used to appropriately model this relationship. Binary, Multinomial, and ordinal logistic regression models are some examples of the robust predictive methods to use for modeling the relationship between non-normal discrete response and the predictors.

This study looks at several methods of modeling binary, categorical and ordinal correlated response variables within regression models. Starting with the simplest case of binary responses, through ordinal response variables, this study discusses different modeling options within SAS®. At the end, some missing data handling techniques are suggested to appropriately account for the high percentages of missing observations which happens a lot in practice when performing these statistical models.

Various statistical techniques such as logistic models, generalized linear mixed models (GLMM) and generalized estimating equations (GEE) are some of the models which are being discussed and applied to real data with categorical outcome variables in this study. This paper discusses different options within SAS 9.4 for the aforementioned models. These procedures include PROC LOGISTIC, PROC GENMOD, PROC GLIMMIX, PROC NLMIXED, PROC CATMOD and PROC GEE.

INTRODUCTION

When modeling non-normal categorical response variables, the robust method to use for modeling the relationship between categorical outcomes and different predictors without assuming a linear relationship between them is logistic regression. In the presence of binomial outcome, binary logistic models can be performed. If the response variable is categorical, then multinomial logistic regression models can be applied. Ordinal logistic regression models have been applied in recent years in analyzing data with ranked multiple response outcomes. When observations are correlated due to clustering or repeated measurements, a complexity is added to the statistical models because of violation of observations' independence assumption. Different conditional and marginal models have been developed to account for the dependence among observations. Models such as generalized linear mixed models (GLMM) and generalized estimating equations (GEE) are some examples of such models. Missing values that are present when dealing with real data most of the time add more complexity to binomial, multinomial and ordinal models. Accounting for all the complexities mentioned above including non-normal response variables, possible correlation among observations 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.

CATEGORICAL RESPONSES AND LOGISTIC MODELS

Logistic regression is useful when modeling binary and categorical response variables based on values of a set of independent variables. The difference between the logistic regression and a usual linear regression model is using a link function in the logistic model. Having this link function in the model will result in a non-constant (non-linear) change in the probability of occurrence of an event with constant changes in the values of the predictors (Agresti, 2007). The relationship between the occurrence of any event (or success) and its dependency on different independent variables can be expressed as

$$p = 1 / (1 + e^{-z}),$$

where p is the probability of the occurrence of an event for example in case of having a binary outcome. The predictors may be either continuous or discrete, or any combination of both types and they do not necessarily have normal distributions. Then, logistic regression fits an equation of the following form to the data

$$z = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik},$$

where β_0 is the model's intercept, β_j 's ($j = 1, 2, \dots, k$) are the slope coefficients of the logistic regression model for k predictors, and x_{ij} 's ($i = 1, 2, \dots, n$; $j = 1, 2, \dots, k$) are the independent variables for n subjects (Hosmer & Lemeshow, 2013).

In logistic regression, the probability of the outcome falling into any of the categories of the response variable is measured by the odds of occurrence of an event. Using the most common form of the logit link function, which some

of them are mentioned in Table 1, within a logistic regression will result in the following logistic regression model

$$\text{logit}(p) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = \beta_0 + \sum_{k=1}^K \beta_k X_k.$$

Depending on the distribution of the outcome which is specified according to the number of categories of the response, the logit functions will vary. The binomial distribution is used for the binary outcomes and when having more than two categories within the outcome, multinomial and ordinal logistic regression models should be used. When the dependent variable has more than two categories, some extensions of the binomial logistic regression should be used. Multinomial and ordinal logistic regression models are some robust models to use under these circumstances. If the categories of the response variable are not ordinal, an unconditional nominal logistic model using a generalized logit link function can be used to model such response variables in which a set of $J - 1$ response functions are modeled and are known as generalized or baseline logits that contrast each level with the last level.

When the response categories are ordered, a multinomial logistic regression model can still be used but it will result in throwing away the information about the ordering of the response (Agresti, 2007). An ordinal logistic regression model is more appropriate for the ordered outcome and provides increased power while keeping the rankings when interpreting the model results. Within such models, three different logit functions can be used to provide useful extensions of the multinomial logistic regression model to data with ordinal outcome. The first one is the cumulative logit function that looks like a binary logistic regression in which categories 1 to j combine to form a single category and categories $j + 1$ to J combine to form a second category. The second one is the adjacent-categories logit function that is used to model two adjacent categories. Within this model, only adjacent categories will be used in odds resulting in using local odds ratios for interpretations, whereas within the cumulative logit models, the entire response scale is used for the model and cumulative odds ratio is used for their interpretation. The third one is the continuation-ratio logit function that contrasts each category with a grouping of categories from higher levels of the response scale. This model is useful when a sequential mechanism determines the response outcome (Agresti, 2013). Table 1 summarizes these logistic models along with the SAS procedures that can be used to fit such models to the data.

Logistic Models	Logit Function	SAS Procedure
Binary Logistic Regression	$\text{logit}(P) = \log\left(\frac{P(Y = j)}{1 - P(Y = j)}\right)$	PROC LOGISTIC & PROC GENMOD
Multinomial Logistic Regression	$\text{logit}(P) = \log\left(\frac{P(Y = j)}{P(Y = j)}\right)$	PROC LOGISTIC & PROC GENMOD (link=glogit)
Ordinal Logistic Regression (Cumulative Logit)	$\text{logit}(P) = \log\left(\frac{P(Y \leq j)}{P(Y > j)}\right)$	PROC LOGISTIC & PROC GENMOD (link=clogit)
Ordinal Logistic Regression (Adjacent Categories Logit)	$\text{logit}(P) = \log\left(\frac{P(Y = j)}{P(Y = j + 1)}\right)$	PROC NL MIXED & PROC CATMOD
Ordinal Logistic Regression (Continuation Ratio Logit)	$\text{logit}(P) = \log\left(\frac{P(Y = j)}{P(Y \geq j + 1)}\right)$	PROC GENMOD & PROC CATMOD

Table 1. Logistic Regression Models: Different Logit Functions and Respective SAS Procedures

Below are some example codes showing how to appropriately use different SAS procedures to fit binomial, multinomial and ordinal logistic models.

Assuming there exist a cross-sectional dataset called Data with a binary dependent variable called DV and two categorical independent variables and two continuous independent variable respectively called IV1, IV2, IV3, and IV4. PROC LOGISTIC and PROC GENMOD can be used to fit a binary logistic regression model. The call to PROC LOGISTIC is displayed:

```
PROC LOGISTIC DATA=Data;
  CLASS IV1 IV2 / PARAM=REF;
  MODEL DV (EVENT='1') = IV1 IV2 IV3 IV4 / LACKFIT CORRB;
RUN;
```

The CLASS statement is used to define all categorical variables so they all need to be listed within the CLASS statement. 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. Using the EVENT= option in the model statement makes it possible to specify the event which is of interest to be modeled. Modeling 1 as event instead of 0 is possible by using the DESCENDING option in the PROC LOGISTIC statement. Using the LACKFIT option after the model statement tests the null hypothesis using Hosmer-Lemeshow test of goodness-of-fit to see whether there is any difference between the observed and predicted values of the response variable or not.

PROC GENMOD can also be used to perform a binary logistic regression as below:

```
PROC GENMOD DATA=Data DESCENDING;
  CLASS IV1 IV2;
  MODEL DV = IV1 IV2 IV3 IV4 / DIST=BIN CORRB;
RUN;
```

Within this procedure specifying DIST=BIN is needed to impose performing a binary logistic regression model. Within GENMOD procedure specifying that we want to model 1 as event instead of 0 for the dependent variable can be done by using the DESCENDING option.

Now, suppose the same dataset mentioned above is used. Assume this time the response variable, DV, has four categories instead of two. The four independent variables are the same as before, IV1, IV2, IV3 and IV4. When having a data set with multiple categories of the response, procedures such as PROC LOGISTIC, PROC GENMOD and PROC NLMIXED can be used to fit multinomial and ordinal logistic regression models as described in Table 1. 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. PROC SURVEYLOGISTIC with the specification of LINK=GLOGIT option can also be used. 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.

When the response variable has multiple ordered categories, using ordinal logistic regression is recommended. To fit such model with a cumulative logit function, PROC LOGISTIC and PROC GENMOD may be used as below:

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

Within the model statement of the PROC LOGISTIC, using LINK=CLOGIT will specify the cumulative logit link function.

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

```
PROC GENMOD DATA= Data RORDER=data DESCENDING;
  CLASS DV (REF="1") IV1 IV2;
  MODEL DV= IV1 IV2 IV3 IV4 / DIST=MULTINOMIAL LINK=CUMLOGIT;
RUN;
```

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

Fitting an ordinal logistic regression with adjacent categories or continuation-ratio logit functions is harder in SAS since there still is not a built in procedure in SAS for this type of analysis. PROC NLMIXED is one of the procedures that can be used to perform the adjacent categories logit model. The likelihood functions need to be typed within the NLMIXED procedure which can be time consuming specially in the presence of a lot of 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).

There also exist 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 but still is not easy to use. Another option when running the continuation ratio model is within PROC LOGISTIC in which various sources (e.g., Allison, 2012) demonstrate how to restructure the original dataset and use

it to create binary response variables. 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).

CORRELATED DATA

One of the conditional models to use when dealing with the correlated binary and categorical responses is the GLMM which is a particular type of mixed-effect models. These models contain fixed effects as well as random effects that usually have normal distributions. More details about it can be found in Agresti (2007) and detailed example SAS codes can be found in Ramezani (2016).

The GLMM model can be written as

$$\eta = X\beta + Z\gamma,$$

where link function is $g(\cdot) = \log_e\left(\frac{p}{1-p}\right)$ and $g(E(y)) = \eta$ (Lee & Nelder, 2004).

Assuming there exist a correlated dataset called Data with a binary dependent variable called DV and two categorical independent variables and two continuous independent variables respectively called IV1, IV2, IV3 and IV4, GLIMMIX and GENMOD procedures in SAS 9.4 can be used to fit a GLMM to this dataset as below. The call to PROC GLIMMIX is displayed:

```
PROC GLIMMIX DATA=Data;
  CLASS IV1 IV2 ID_CODE;
  MODEL DV = IV1 IV2 IV3 IV4 / DIST=BIN LINK=LOGIT SOLUTION;
  RANDOM INTERCEPT / SUBJECT=ID_CODE;
RUN;
```

Notice all the categorical variables need to be listed within the CLASS statement. The binomial distribution should be used in the MODEL statement because the dependent variable is binary. So, within this procedure, options DIST=BIN and LINK=LOGIT are provided to specify a logistic regression model using a generalized linear model link function. Adding the option SUBJECT=ID_CODE to the code will specify the repeated measures that exist for every subject with its unique ID_CODE. This will account for the dependence among the multiple measures per subject. RANDOM statement can be used to specify the random intercept and slope. PROC NLMIXED can also be used to fit the same model.

One of the marginal models to use when dealing with the correlated data is GEE developed by Liang and Zeger (1986). GEE can be used to estimate the regression coefficients without completely specifying the response distribution. A 'working' correlation structure is used instead to explain the correlation between a subject's repeated measurements. When modeling discrete response variables, GEE can be used to model correlated data with binary or categorical responses with more than two categories. If the main interest is estimating the regression parameters rather than variance-covariance structure of the correlated data, this technique is recommended. The desirable characteristic of GEE models is that the estimators of the regression coefficients and their standard errors based on GEE are consistent even if the covariance structure for the data is misspecified. It also allows missing values within a subject without losing all the information from that subject resulting in a higher power of the study (Fitzmaurice, Davidian, Verbeke, & Molenberghs, 2009).

Different within-subject correlation matrices can be used within GEE models. Independent that specifies no correlation among the repeated observations, exchangeable that is used when the same correlation between any two responses of each subject exists, autoregressive that is used if the interval length between any two observations is the same, and unstructured that suggests some sort of unknown correlation between any two responses.

PROC GENMOD can be used to fit GEE models for both binary and categorical correlated outcomes. Considering the dataset and variables introduced above for the correlated binary outcome, the procedure may be performed as below:

```
PROC GENMOD DATA= Data DESCENDING;
  CLASS IV1 IV2 ID_CODE;
  MODEL DV = IV1 IV2 IV3 IV4 / DIST=BIN CORRB;
  REPEATED SUBJECT=ID_CODE / CORR=UN;
RUN;
```

The REPEATED statement indicates the use of the GEE approach in order to account for the correlation that exists among repeated observations. CORR=UN specifies an unstructured within-time correlation matrix which can be replaced by any other structure.

When modeling correlated data with categorical outcomes using GEE, PROC GENMOD can still be used by

specifying DIST=MULT within the MODEL statement to request ordinal multinomial logistic model. Assuming that now DV is a categorical response variable with more than two categories, PROC GENMOD may be performed as below:

```
PROC GENMOD DATA=Data RORDER=data DESCENDING;
  CLASS DV (REF="1") IV1 IV2 ID_CODE;
  MODEL DV= IV1 IV2 IV3 IV4 / DIST=MULTINOMIAL LINK=CUMLOGIT;
  REPEATED SUBJECT= ID_CODE / CORR=UN;
RUN;
```

The REF= option in the CLASS statement determines the reference level for EFFECT which can be used for categorical dependent and independent variables. The CUMULATIVE link is referring to the cumulative logit function which can be used within these models. More details about different logit functions for modeling categorical response variables using data examples can be found in Ramezani (2016).

PROC GEE can also be used for modeling ordinal multinomial responses beginning in SAS 9.4 TS1M3. Using TYPE= option in the REPEATED statement specifies the correlation structure among the repeated measurements within a subject and fits a GEE to the data as below:

```
PROC GEE DATA= Data DESCENDING;
  CLASS DV (REF="1") IV1 IV2 ID_CODE visit;
  MODEL DV= IV1 IV2 IV3 IV4 visit / DIST=MULTINOMIAL;
  REPEATED SUBJECT= ID_CODE / WITHIN=visit;
RUN;
```

Variable “visit” is being added to this procedure to be used in the WITHIN option to specify the order of the measurements being recorded in multiple visits of subjects in the study. If the data are entered in proper order within each subject, there is no need to specify this option but if there are some missing observations at some time points, it needs to be specified to properly order the existing measurements and treat the omitted measures as missing values. If the WITHIN= option is not specified for the standard GEE method, missing values are assumed to be the last values and the remaining observations are ordered in the sequence in which they are entered in the original input data set.

MISSING DATA

Missing data presents a challenge in any type of research. Missing data is associated with numerous statistical concerns (Cheema, 2014), and the severity of the problem depends on the type of missing data (Rubin, 1976) as well as the quantity of missing observations (Gibson & Olejnik, 2003). Various 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). Listwise deletion is easy to use and is often the default in statistical packages, but it can be led to 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 way to impute missing observations. Within this method, each missing value in a variable is replaced with the mean of the observed values for that variable resulting in a very small variation in each variable due to using the same value instead of each missing observation. Mean imputation is not recommended and the user should be aware of the implications (Buuren & Groothuis-Oudshoorn, 2011).

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 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 correct estimates of the standard errors while the precision of the study associations is commonly overestimated 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, then the multiple results are combined using Rubin's rule (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 2 summarizes these missing data handling methods with the appropriate SAS procedure to be used to perform them.

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

Table 2. Missing Data Procedures

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;
```

To perform the multiple imputation as the missing data handling technique for the same analysis, 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;
```

As it is obvious from the SAS code mentioned above, there are three main steps in performing a multiple imputation within SAS. 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. The number of imputations can be specified using NIMPUTE in the PROC MI statement. Intercept as well as the independent variables, which their coefficients need to be estimated, should be specified in the MODELEFFECTS statement in the MIANALYZE procedure.

Below is another multiple imputation example for when modeling correlated ordinal response variables:

```
PROC MI DATA=Data SEED=454 NOPRINT OUT=outimputed;
  VAR DV IV1 IV2 IV3 IV4;
RUN;
```

```
PROC GENMOD DATA=outimputed;
  CLASS DV (REF="1") IV1 IV2 resident_id;
  MODEL DV=IV1 IV2 IV3 IV4 / DIST=MULTINOMIAL LINK=CUMLOGIT covb;
  REPEATED SUBJECT=resident_id / CORR=INDEP;
  BY _Imputation_;
  ODS OUTPUT ParameterEstimates=gmparms
  CovB=gmcovb;
RUN;
```

```
PROC PRINT DATA=gmparms (obs=8);  
VAR _Imputation_ Parameter Estimate;  
TITLE 'GENMOD Model Coefficients (First Two Imputations)';  
RUN;  
  
PROC PRINT DATA=gmcovb (obs=8);  
VAR _Imputation_ RowName Prm1 Prm2 Prm3;  
TITLE 'GENMOD Covariance Matrices (First Two Imputations)';  
RUN;  
  
PROC MIANALYZE PARMS=gmparms ;  
MODELEFFECTS Intercept IV1 IV2 IV3 IV4;  
RUN;
```

This code will result in five imputed data sets and will use Rubin's rule to give the final estimates at the end.

CONCLUSION

Different options for modeling non-normal discrete response variables were discussed above. These response variable types include binary, multinomial and ordinal. 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. When dealing with correlated observations, taking into consideration the correlation that exists among observations is important because the existence of repeated measurements results in the violation of the independence assumption, hence the cross-sectional models cannot appropriately model the correlated data anymore. Procedures such as PROC GENMOD, PROC GLIMMIX, PROC NLMIXED and PROC GEE can appropriately model correlated outcomes. There still are some issues to fit ordinal models to correlated data. Using appropriate models which were discussed above will result in more informative and powerful models.

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 common methods such as listwise deletion. More work need to be done in developing SAS procedures that can easily fit ordinal logistic regression models to cross-sectional and correlated data specifically when trying to use adjacent-categories and continuation-ratio logit functions.

REFERENCES

- Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). New York: Wiley.
- Agresti, A. (2013). *Categorical data analysis* (3rd ed.). New York: Wiley.
- Allison, P. D. (2012). *Logistic regression using SAS: Theory and application*. SAS Institute.
- 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.
- Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (Eds.). (2009). *Longitudinal data analysis*. CRC Press.
- 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.
- High, R. Models for Ordinal Response Data (2013). SAS Global Forum, Paper 445-2013.
- Hosmer Jr, D. W., & Lemeshow, S. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons.
- 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, Y., & Nelder, J. A. (2004). Conditional and marginal models: another view. *Statistical Science*, 19(2), 219-238.
- Liang, K. Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13-22.
- 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. 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.
- 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.
- 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.

CONTACT INFORMATION

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

Niloofar Ramezani
University of Northern Colorado
Niloofar.ramezani@unco.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.