Paper 11702-2016

Analyzing non-normal binomial and categorical response variables under varying data conditions

Niloofar Ramezani

University of Northern Colorado

ABSTRACT

When dealing with non-normal categorical response variables, logistic regression is the robust method to use for modeling the relationship between categorical outcomes and different predictors without assuming a linear relationship between them. Within such models, the categorical outcome may be binary, multinomial, or ordinal and predictors may be continuous or categorical. Another complexity that might be added to such studies is when data are longitudinal, such as when outcomes are collected at multiple follow-up times. Learning about modeling such data within any statistical method is beneficial because it allows researchers to look at changes over time. This study looks at several methods of modeling binary and categorical response variables within regression models using real-world data. Starting with the simplest case of binary outcomes, through ordinal outcomes, this study looks at different modeling options within SAS® and includes longitudinal cases for each model. To assess binary outcomes, the current study models binary data in the absence and presence of correlated observations under regular logistic regression and mixed logistic regression. To assess multinomial outcomes, the current study uses multinomial logistic regression. When responses are ordered, using ordinal logistic regression is required as it allows for interpretations based on inherent rankings. Different logit functions for this model include the Cumulative Logit, Adjacent–Categories Logit, and Continuation Ratio Logit. Each of these models is also considered for longitudinal (panel) data using methods such as mixed models and Generalized Estimating Equations (GEE). The final consideration, which cannot be addressed by GEE, is the conditional logit to examine bias due to omitted explanatory variables at the cluster level. Different procedures for the aforementioned within SAS 9.4 are explored and their strengths and limitations are specified for applied researchers finding similar data characteristics. These procedures include PROC LOGISTIC, PROC GLIMMIX, PROC GENMOD, PROC NLMIXED, and PROC PHREG.

Keywords: LOGISTIC REGRESSION, GENERALIZED MIXED MODEL, ORDINAL MODEL, GEE

INTRODUCTION

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 response variables, logistic regression is the robust predictive method to use for modeling the relationship between response and the predictors.

When the outcome variable has two possible categories, binary logistic regression is appropriate to model the association of the risk factors with the occurrence or nonoccurrence of an event. In the presence of a categorical response variable with more than two possible outcomes, some extensions of binary logistic models need 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 but their use in some fields is still rare. This may be attributed to these models' complexity, assumptions validation, and limitations of modeling options offered by statistical packages (Lall et al., 2002). Regardless of their complexity, ordinal hypothesis tests provide increased power and ordinal logistic models allow for interpretations based on inherent rankings; therefore, increased accessibility of these models, particularly choosing among link functions such as cumulative logits, adjacent-category logits, and continuation-ratio logits, is important.

When dealing with correlated and longitudinal data, another complexity is added to the model which requires some adjustments. This is because by violating the assumption of the observations independence, some problems such as overestimation of the statistical significance and underestimation of variance may arise if the appropriate models are not used (Williams, 1995). Binary, multinomial, and ordinal logistic regression assume independence of the observations; therefore, they won't work with the correlated data anymore (Bena & McIntyre, 2008). There are different models that can be applied when dealing with correlated data and among those, Generalized Linear Mixed Model (GLMM) which is a particular type of mixed models is a useful approach to be applied instead of the binary logistic regression. Generalized Estimating Equations (GEE), Alternating Logistic Regression (ALR) and Fixed Effects with Conditional Logit Analysis are some other methods which may be used to account for the correlation among observations. Using two real data sets, this study compares various statistical models and evaluates different procedures that can be used within SAS 9.4 (The SAS Institute, Cary, NC).

BINARY LOGISTIC MODELS

Logistic regression is useful for predicting the presence or absence of a characteristic or an outcome based on values of a set of predictor variables through the addition of an appropriate link function to the usual linear regression model. The variables may be either continuous or discrete, or any combination of both types and they do not necessarily have normal distributions. As explained by Hosmer and Lemeshow (2013), the relationship between the occurrence of any event 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. Then, logistic regression fits an equation of the following form to the data

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

where β_0 is the model's intercept, $\beta_j{}'s$ $(\mathrm{j} = 1, 2, \ldots, \mathrm{k})$ are the slope coefficients of the logistic regression model, and x_{ij}'s $(i = 1, 2, \ldots, n\,;\; j = 1, 2, \ldots, k)$ are the independent variables.

In logistic regression, the probability of the outcome is measured by the odds of occurrence of an event. Change in probability is not constant (linear) with constant changes in x. This means that the probability of a success given the predictor variables is a non-linear function, specifically a logistic function. The most common form of logistic regression uses the logit link function which gives us the logistic regression equation as

$$logit\ (p_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}.$$

Example (Binary Outcome): Transitional Housing Facility

Data are presented from a transitional housing facility (THF) in the greater Rocky Mountain Region that works as a temporary facility that helps family units find stable housing and employment while living within the shelters. The THF aims to help these first-time and episodic (experiencing episodes of homelessness) homeless families regain stable housing and employment with the goal of mitigating long-term occurrence of the problem. This research outlines the changing condition of the homeless in terms of the time it takes until finding a job and leaving the transitional house based on the job training they had at the center, the number of hours they spent with case worker each month, either they had temporary assistance for needy families or not, and some demographic variables that seem effective on the length of their stay at the transitional house like child abuse, number of children in the family, being either a single parent or not, and finally being either unemployed or employed for families living in the THF from 2006 to 2010.

To fit logistic regression to this dataset, 60 days can be considered as the end point because families were supposed to leave the THF within 60 days while not everyone did so. The number of nights they stayed at the THF is used to define the binary response variable; if this number is larger than 60, then the length of stay is defined as censored (0) for the related family, but if this number is less than 60, the length of stay is defined as non-censored (1) borrowing the concept of censoring from survival models. A logistic regression is applied to the aggregated data looking into the binary censoring variable to model whether the family left the THF within 60 days or stayed longer. The reason for aggregating the data for this step is to have only one record per family since some families stayed multiple times at the THF. This repeated nature of the dataset will be taken into consideration later in this paper. Performing a logistic regression on the THF data, the likelihood-ratio test (p-value < .0001) and Wald test (p-value < .0001) showed a significant model. The model fit was also good based on Hosmer and Lemeshow goodness of fit test (p-value = .9923). To check the association of predicted probabilities and observed response, Somer's D value was calculated which showed that the proportion of variance explained by the variables in the model is 95% which is very high. All these results are part of the output of the below SAS procedures but are not included in this paper due to lack of space. Based on analysis of maximum likelihood estimates illustrated in Table 1, only the numbers of hours (summation of the hours for aggregated data) families spent with the case worker within the THF is significant (Job_sum; p-value < .0001) at the significance level of .05. PROC LOGISTIC is used for this part of analysis. The call to this procedure is displayed:

```
PROC LOGISTIC DATA=homeless;
        CLASS TANF Abuse Unemployed Single_Parent censoring  / PARAM=REF;
        MODEL censoring (EVENT='1') = TANF Abuse Num_Children Unemployed Single_Parent
        Job_sum Case_sum /LACKFIT CORRB;
RUN;
```

Notice that the CLASS statement is used for all categorical variables. SAS will create dummy variables for a categorical variable with the default of effect coding for all the categorical variables in PROC LOGISTIC. By using the PARAM = REF option, it is changed to dummy coding. Notice that specifying the event which is of interest to be modeled is possible by using the EVENT = option in the model statement. Using a pair of quotes surrounding 1 is necessary even when the variable is numeric and it will come in handy when dealing with character response variables. It is also doable to specify that we want to model 1 as event instead of 0 by using the DESCENDING option in the PROC LOGISTIC statement. Using the LACKFIT option after the model statement gives the results of the Hosmer-Lemeshow test of goodness-of-fit. It tests the null hypothesis that there is no difference between the observed and predicted values of the response variable. Requesting the generalized R-square measure for this model is also possible by using RSQUARE option after the model statement.

Table 1: Results of logistic regression analysis

Analysis of Maximum Likelihood Estimates Logistic Regression

Parameter		*DF*	*Estimate*	*Standard Error*	*Wald Chi-Square*	*Pr > ChiSq*
Intercept		1	14.3722	2.2156	42.0805	<.0001
TANF	*0*	1	0.1554	0.6818	0.0520	0.8197
Abuse	*0*	1	0.0539	0.5116	0.0111	0.9161
Num_Children		1	-0.0865	0.1925	0.2020	0.6531
Unemployed	*0*	1	-0.7524	0.5134	2.1482	0.1427
Single_Parent	*0*	1	1.0044	0.5529	3.3001	0.0693
Job_sum		1	-0.2633	0.2151	1.4976	0.2210
Case_sum		1	-0.6833	0.1029	44.0752	<.0001

PROC GENMOD is also applied as another option to perform the same analysis which provides the same results. Within this procedure specifying DIST=BIN is needed to impose performing a binary logistic regression model with the outcome variable which follows a binomial distribution. Within GENMOD procedure specifying that we want to model 1 as event instead of 0 for the dependent variable is done by using the DESCENDING option. PROC GENMOD to perform a binary logistic regression can be conducted as below:

```
PROC GENMOD DATA=homeless DESCENDING;
        CLASS TANF Abuse Unemployed Single_Parent censoring;
        MODEL censoring = TANF Abuse Num_Children Unemployed Single_Parent Job_sum
        Case_sum / DIST=BIN CORRB;
RUN;
```

CORRELATED DATA WITHIN BINARY LOGISTIC MODELS

As mentioned in Bena and McIntyre (2008), an added complexity occurs when not all of the observations are independent. When there are multiple observations per subject in a model, standard errors of the estimates underestimate the true amount of variability that exists.

The modelling of correlated binary outcomes in a way that the marginal response probabilities are still logistic has been discussed in different articles along with the association measures for the dependence between correlated observations. For paired correlated data, the full likelihood can be evaluated and for an arbitrary number of correlated observations, a pseudo likelihood approach for obtaining parameter estimates is proposed by Cessie and Houwelingen (1994). A discussion of the various approaches to model correlated binary observations can be found in Prentice (1988), Zeger, Liang, and Albert (1988), and also in Neuhaus, Kalbfleisch, and Hauck (1991).

GEE is a population average model. Liang and Zeger (1986) and Zeger et al. (1988) first used the GEE with the binary data. The set of equations used in the GEE approach look like weighted versions of the likelihood equations. Requiring an assumption about the structure of this correlation, the weights involve an approximation of the underlying covariance matrix of the correlated within-cluster observations. Under the independent model, $Cor(Y_{ij}, Y_{i1}) = 0$ for $j = 1$ and the GEE equations simplify to the likelihood equations obtained from the binomial likelihood (Hosmer & Lemeshow, 2013).

The correlation among responses depends on the lag between the observations and is assumed to be constant for equally lagged observations. Settings where there is an explicit time component are more specialized and need additional approaches to handling such data covered in texts such as Diggle et al. (2002, as cited by Hosmer & Lemeshow, 2013) or Hedeker and Gibbons (2006, as cited by Hosmer & Lemeshow, 2013). In the unstructured correlation case, one assumes that the correlation of the possible pairs of responses is different, $Cor(Y_{ij}, Y_{i1}) = \rho_{j1}$ for $j = 1$. The disadvantage of using this method is that it requires estimating a large number of parameters that are, for the most part, of secondary importance. In most applications researchers are only interested in estimating the regression coefficients and need to account for correlation in the responses to obtain the correct estimates of the standard errors of the estimated coefficients. The idea is to choose a correlation structure for estimation that seems plausible for the setting and then this structure is used in adjusting the variance's estimator (Hosmer & Lemeshow, 2013). For data without a clear choice of structure, a reasonable and parsimonious choice is the "exchangeable correlation" structure. One of the advantages of the GEE approach is the "robustness" of the estimates to choice of correlation. In other words, even if the correlation structure chosen is not the true structure, the parameter estimates from the GEE are often still valid.

Pseudo likelihood estimation and related estimation techniques are also very helpful if the full underlying distribution of the data is unknown or if the true likelihood is difficult to evaluate. The pseudo-likelihood method is a very easy method to understand and the multivariate correlation model has the advantage that estimates of the joint probabilities can be generated relatively easily. There is some loss in efficiency by using pseudo-likelihood but because it equals the full likelihood for $p = 0$, only small losses are expected when p is small (Cessie & Houwelingen, 1994).

One of the other models to use when dealing with the correlated data is the GLMM which is a particular type of mixed models. It is an extension to the generalized linear model in which the linear predictor contains random effects in addition to the usual fixed effects. These random effects are usually assumed to have a normal distribution. In the GLMM, the default optimization technique that is used is the Quasi-Newton method. Because a residual likelihood technique is used to compute the objective function, only the covariance parameters are participating in the optimization. This model is not complicated and more details about it can be found in Agresti (2007).

The GLMM can be written as

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

where link function is $g(.) = log_e(\frac{p}{1-p})$ and $g(E(y)) = \eta$.

Example (Binary Outcome, Correlated Data): Transitional Housing Facility

When first analyzing the THF dataset above, the repeated measures were being aggregated. However, if the repeated records of the families who returned to the THF multiple times are kept in the model, more information can be provided which needs some adjustments to account for the added complexity which occurs when not all the observations are independent. This should be addressed by applying appropriate models to the correlated THF data which does not satisfy the independence assumption anymore because of not aggregating different observations of the same family over time for the second analysis to keep the repeated measure nature of the data that provides more observations including 926 families.

As mentioned before to generalize the logistic regression for the autocorrelated data, one of the best models is the GLMM which is a particular type of mixed models. A lower boundary constraint is placed on the variance component for the random center effect. The solution for this variance cannot be less than zero. After the initial optimization using the GLIMMIX procedure, it performed 16 updates before the convergence criterion was met. At convergence, the largest absolute value of the gradient was near zero which indicates that the process stopped at an extremum of the objective function. This model is statistically significant as twice the negative of the residual log likelihood in the final pseudo-model equaled 4764.52. The ratio of the generalized chi-square statistic and its degrees of freedom is close to 0.3. This is a measure of the residual variability in the marginal distribution of the data. From the covariance parameter estimation procedure, the variance of the random center intercepts on the logit scale is estimated as $\hat{\sigma}_c^2$=5.2757. Finally looking into Table 2 to check the significance of different variables, it is obvious that the number of hours families spent with case worker each month is significant at the .01 significance level (Case_Hours; p-value = .0032) and employment is also significant at the .05 significance level (Unemployed; p-value = .0211). The call to PROC GLIMMIX is displayed:

```
PROC GLIMMIX DATA=REPhomeless;
        CLASS TANF Abuse Unemployed Single_Parent censoring ID_CODE;
        MODEL censoring = TANF Abuse Num_Children Unemployed Single_Parent Job_hour
        Case_hour / DIST=BIN LINK=LOGIT SOLUTION;
        RANDOM INTERCEPT / SUBJECT=ID_CODE;
RUN;
```

Within this procedure, options DIST=BIN and LINK=LOGIT are provided to specify a logistic regression model using a generalized linear model link function. Notice that for this model only intercept is being specified as random; random slope can be used for this model as well by simply adding the name of the independent variables in front of the random statement within PROC GLIMMIX. The problem of non-convergence happens a lot when fitting mixed-effect models to different data sets which in that case PROC NLMIXED has more flexibility and is a good option. This problem did not arise in this data analysis so there was no need to use PROC NLMIXED.

PROC GENMOD can also be used to take care of the correlation among observations by performing the GEE analyses to account for the dependence among data. The procedure may be performed as below:

```
PROC GENMOD DATA= REPhomeless DESCENDING;
        CLASS TANF Abuse Unemployed Single_Parent censoring ID_CODE;
        MODEL censoring = TANF Abuse Num_Children Unemployed Single_Parent Job_hour
        Case_hour/ DIST=BIN CORRB;
        REPEATED SUBJECT=ID_CODE / CORR=UN;
RUN;
```

The REPEATED statement indicates the GEE approach, and CORR=UN specifies an unstructured within-time correlation matrix which can be replaced by other structures. If the repeated observations were fixed for every subject in this study at the same number of time points, the option CORRW could be added after CORR=UN to display the working correlation matrix between the time point measurements in the output. Adding this option was not necessary for this data analysis due to having an unbalanced data set for this study meaning that not every subject has the same number of repeated observations, if any.

Table 2: Results of the GLMM analysis

Solutions for Fixed Effects					
Effect	*Estimate*	*Standard Error*	*DF*	*t Value*	*Pr > |t|*
Intercept	3.2436	0.5461	413	5.94	<.0001
Child_Abuse	-0.2864	0.4396	505	-0.65	0.5150
Case_Hours	-0.1541	0.05205	505	-2.96	0.0032
Job_Training	-0.6988	0.4233	505	-1.65	0.0994
Single_Parent	-0.03898	0.3794	505	-0.10	0.9182
Unemployed	0.7847	0.3391	505	2.31	0.0211
Num_Children	-0.1358	0.1308	505	-1.04	0.2998
TANF	-0.3568	0.4594	505	-0.78	0.4378

If one is willing to take into consideration how long it takes for each family to leave the THF which is an important factor, this could be presented by an additional variable recording the time to the event of interest (leaving THF here). Some researchers argue the necessity of using this variable due to the significant importance of accounting for time to the occurrence of an event. In order to use this piece of information in the analysis, survival analysis for independent data and shared frailty models for correlated observations may be adopted to provide more information and result in models with higher power. As this was not the main focus of this study, it has not been reported here. To find out more about such models and their comparison to the aforesaid techniques of modeling binary outcomes, see Ramezani (2014).

MULTINOMIAL AND ORDINAL LOGISTIC MODELS

The multinomial logistic regression model is an extension of the binomial logistic regression discussed above. This type of model is used when the dependent variable has more than two nominal categories. If the categories of the response variable are not ordinal, an unconditional nominal logistic model 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. This generalized logit function may be written as below:

$$GLogit = log\left(\frac{P(Y=j)}{P(Y=J)}\right), for\ j = 1,2,\ldots,J.$$

When the response categories are ordered, a multinomial logistic regression model can still be used. According to Agresti (2007), the disadvantage is that some information about the ordering is thrown away. An ordinal logistic regression model which is more appropriate for the ordered outcomes preserves that information and it is slightly more informative. Regardless of their complexity, ordinal hypothesis tests provide increased power while keeping the rankings when interpreting the model results. Therefore, increased accessibility of these models is important, particularly choosing among link functions such as cumulative logits, adjacent-category logits, and continuation-ratio logits. An ordinal multinomial logistic regression is as below

$$logit\ (P) = \beta_{0j} + \sum_{k=1}^{K} \beta_k X_k.$$

Three different logit functions which were mentioned above are used within regression models to provide useful extensions of the multinomial logistic model to ordinal response data. To evaluate these logit functions and compare their performance within SAS procedures, the author proposes fitting these models to a new data set with multiple categories of the response and describes the existing limitations of modeling options within procedures such as PROC LOGISTIC, PROC GENMOD and PROC NLMIXED. Each of these models is briefly explained below according to the notations used in Agresti (2013):

Cumulative Logit

The cumulative logit function used in ordinal multinomial logistic models looks like a binary logistic regression in which categories $1\ to\ j$ combine to form a single category and categories $j+1$ to $\boldsymbol{J}$ form a second category. This logit function is as below modeling categories $\leq j$ versus categories $> j$

$$logit\ (P) = log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right), \qquad for\ j = 1, \cdots, J-1.$$

Adjacent–Categories Logit

The adjacent-categories logit function used in ordinal multinomial logistic models is used to model two adjacent categories. This approach forms logits for all pairs of adjacent categories as below

$$logit\ (P) = log\left(\frac{P(Y = j)}{P(Y = j+1)}\right), \qquad for\ j = 1, \cdots, J-1.$$

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.

Continuation–ratio Logit

The continuation-ratio logit function used in ordinal multinomial logistic models contrasts each category with a grouping of categories from higher levels of the response scale and is as below

$$logit\ (P) = log\left(\frac{P(Y = j)}{P(Y \geq j+1)}\right), \qquad for\ j = 1, \cdots, J-1$$

As described in Agresti (2013), this model is useful when a sequential mechanism determines the response outcome. Mechanisms like survival through various age periods would be suitable for such models.

Example (ordinal multinomial outcome): Mental Health Data

Data are presented from a mental health facility in North Carolina that works as a healing community to help individuals with a mental health challenge or emotional distress to learn new ways to gain coping skills, learn to become independent, and attain fulfillment in life through a comprehensive program. This research tries to predict the length of stay (LOS) of the residents of this facility. This study describes length of stay and other baseline variables, selects relevant variables, and selects logit functions to be used within ordinal multinomial logistic regression.

One of the challenges of working with this dataset is that the LOS is right-skewed and truncated at zero. There also are some limits of possible LOS values due to their nature. Another challenge is having numerous baseline measures available to be used in LOS prediction. Replicated observations is another challenge as it violates the assumption of the observations independence; therefore, some problems such as overestimation of the statistical significance and underestimation of variance may arise if the correlated observations are ignored (Williams, 1995). The correlated measurements add a complexity to the statistical model which requires some adjustments. Due to the fact that those more complex models were not the purpose of this section, for this part of analysis aggregation is used to take care of the correlated observations issues. In the next section of the study which is allocated to the modeling options of correlated ordinal data, all the repeated measurements are kept in the model which will be discussed later.

The original dataset included 322 observations of 40 baseline variables. In order to start this analysis, the baseline measures that effectively predict LOS needed to be identified to be used in the future models. For the initial analysis, a log-linear Poisson regression model was applied to account for the right-skewness inherent in the LOS, and LASSO estimation was used to account for multicollinearity and to select the appropriate predictors using SAS. Missing values were eliminated using listwise deletion at this stage for simplicity so this part of analysis was done on 242 complete observations. It was found that seven variables can be used to effectively predict length of stay with the correlation of about 0.806 between predictors and observed LOS. These variables are health survey score (HS), spirituality survey score (SS) representing the measure of spirituality, race / ethnicity indicators for Caucasian (C) and Hispanic (H), marital status (MS), depression (D), and anxiety (A). The resulting formula for predicting length of stay is

$$\widehat{LOS} = \exp(5.10135 + 0.06884 \times HS - 0.01032 \times SS - 0.11066 \times C + 0.54479 \times H + 0.08172 \times MS \\ + 0.00842 \times D + 0.00101 \times A)$$

Eight additional variables were selected with lighter restrictions giving the correlation of about 0.843 between predictions and observed LOS. They are the indicators for Schizophrenia, Personality Disorder, Future Snyder Hope Scale at baseline (Futs_00), Mental Health Recovery Measure at baseline (mhrm_00) which is a 30-item self-report survey developed to measure five constructs of recovery with a traditional five-point Likert response scale indicating more positive recovery outcomes for higher scores, Obsessive-Compulsive Disorder Indicator at baseline (O_C0), Health Outcomes Survey score at baseline (HOS0) which includes items related to patients taking care of their own health and hygiene, Global Scale Inventory at baseline (GSI0), and Positive Symptom Distress Index measured at baseline (PSDI0).

Within SAS, listwise deletion is default for most of the procedures including the ones used for this analysis within ordinal multinomial logistic models so no specific procedure needed to be added for listwise deletion. Other missing data handling techniques can be applied within the logistic models such as mean imputation which is easily done using PROC STANDARD; however, mean imputation is not recommended due to underestimation of standard errors. On the other hand, multiple imputation is more informative but more complicated and needs to be done in three steps; first using PROC MI to impute data, then running the actual analysis (i.e., PROC LOGISTIC for this data analysis), and finally PROC MIANALYZE to pool the results from all imputations together and get the final results. This part of analysis is done but not reported in this paper as it requires a separate discussion of missing data type, missing data handling techniques, and possible procedures within SAS which is not the main purpose of the current study. More details regarding missing data and the appropriate missing data handling techniques within these logistic models in SAS can be found in Ramezani (2015).

Model Grouped Length of Stay

Instead of predicting the actual LOS (in days) which was done as the initial analysis and involves a lot of random noise, it may be of interest to model LOS in greater groups. For example, models can be constructed to predict the chance of a client staying for less than three months, between three and six months, etc. This "coarser" view may give a more reliable and useful indication of how long clients tend to stay based on initial measures, and it also is more interesting to the mental healing facility and families of the clients in different aspects including the financial planning.

The LOS was categorized into four groups which are up to three months (group 1), three to six months (group 2), six to twelve months (group 3), and finally more than twelve months (group 4). When modeling LOS using ordered groups for individuals with persistent psychological health conditions at the aforesaid live-in healing community in North Carolina, there is a need for appropriate models of ordinal data. Using an Ordinal Multinomial Logistic Regression with 14 predictors, the chance of falling into each of these groups is predicted. It is important to predict LOS to allocate appropriate funding to this community. The variables affecting the length of stay in the healing community, such as race, gender, and health conditions are presented in the logistic models so their significance and effect on the LOS response can be evaluated and used as a way of comparing different models. The LOS response variable which was originally reported in days may be categorized in four ordinal groups using this SAS code:

```
DATA mental;
        SET mental;
        IF LOS<=180 THEN length=1;
        IF (LOS>=181 AND LOS<=270) THEN length=2;
        IF (LOS>=271 AND LOS<=365) THEN length=3;
        IF LOS>365 THEN length=4;
RUN;
```

The new created dataset with the additional categorical outcome may be used within SAS for the future analysis procedures or may be exported and used in any other statistical software packages in different formats which is recommended to do. Exporting this dataset in a csv format can be done as below:

```
PROC EXPORT DATA=mental
        OUTFILE="C:\Users\Documents\Cat_mental.csv" DBMS=CSV;
RUN;
```

When fitting an ordinal logistic regression, different logit functions within these models can be used. In SAS, PROC GENMOD or PROC LOGISTIC can be easily used to perform an ordinal multinomial logistic regression model using a cumulative logit. Table 3 shows one of the output results for the Cumulative Logit within an ordinal logistic model.

As mentioned above, PROC LOGISTIC may be used for fitting such model as below:

```
PROC LOGISTIC DATA=Cat_mental;
        CLASS length (REF="1") SA Personality race (REF = "Caucasian/") marital_status (REF="S") /
        PARAM = REF;
        MODEL length= SA Personality futs_00 hsur_00 mhrm_00 Sibr_00 race marital_status O_C0
        DEP0 ANX0 HOS0 GSI0 PSDI0 / 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=Cat_mental RORDER=data DESCENDING;
        CLASS length (REF="1") SA Personality race (REF = "Caucasian/") marital_status (REF="S");
        MODEL length= SA Personality futs_00 hsur_00 mhrm_00 Sibr_00 race marital_status O_C0
        DEP0 ANX0 HOS0 GSI0 PSDI0 / 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.

After performing the above analysis, test of global null hypothesis for the ordinal multinomial logistic regression model for this data analysis is significant. The significant predictors when using listwise deletion as the technique of handling missing data under cumulative logit model were baseline Health Survey Measure on Admission (hsur_00), Depression on Admission (DEP0), Positive Symptom Distress Index (PSDI0), and the baseline measure of Obsessive Compulsive disorder (O_C0). Higher values of all of these variables result in higher chance of longer stay.

If the response categories were not ordered, PROC LOGISTIC with the specification of LINK=GLOGIT option in the MODEL statement could be used to fit a multinomial logistic regression. PROC SURVEYLOGISTIC can also be used to fit this model with the LINK=GLOGIT option. The GLIMMIX and HPGENSELECT procedures fit this model as well by specifying the DIST=MULT and LINK=GLOGIT options in the MODEL statement. All of the above procedures fit the model using maximum likelihood estimation. When the response has more than two levels, PROC CATMOD can fit the model using maximum likelihood by default or using weighted least squares after specifying the WLS option; however the use of PROC CATMOD is not recommended in general.

Table 3: Ordinal Multinomial Logistic Regression - Cumulative Logit (Listwise Deletion)

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	4	1	-4.1602	1.1393	13.3339	0.0003
Intercept	3	1	-2.7307	1.1142	6.0066	0.0143
Intercept	2	1	-1.0112	1.1002	0.8448	0.3580
SA	0	1	0.3520	0.2740	1.6503	0.1989
Personality	0	1	-0.5296	0.2961	3.1999	0.0736
futs_00		1	0.1486	0.3530	0.1771	0.6739
hsur_00		1	0.8512	0.2584	10.8472	0.0010
mhrm_00		1	-0.1174	0.3099	0.1435	0.7048
Sibr_00		1	0.0699	0.1610	0.1885	0.6641
race	Asian/Paci	1	1.4664	1.3156	1.2424	0.2650
race	Hispanic	1	1.8181	1.1461	2.5164	0.1127
race	Middle Eas	1	-12.4552	1027.2	0.0001	0.9903
race	Multi-raci	1	-10.4533	1027.2	0.0001	0.9919
race	Native Ame	1	-0.8489	1.0789	0.6191	0.4314
marital_status	D	1	0.8101	1.3384	0.3663	0.5450
marital_status	M	1	-0.5878	0.5349	1.2075	0.2718
O_C0		1	0.1049	0.0389	7.2859	0.0069
DEP0		1	0.1196	0.0400	8.9453	0.0028
ANX0		1	0.0318	0.0394	0.6490	0.4205
HOS0		1	0.0517	0.0483	1.1475	0.2841
GSI0		1	-0.1381	0.1123	1.5100	0.2191
PSDI0		1	-1.0958	0.2854	14.7358	0.0001

PROC NLMIXED can be used to perform the adjacent categories logit model, but due to the fact that there still is not a straightforward built in procedure in SAS for when using this type of logit function, the likelihood functions need to be typed within the NLMIXED procedure which can be time consuming specially when there are a lot of independent variables in the model. Using PROC CATMOD to perform this type of analysis is mentioned in some books including Allison (2012) but is not recommended for such models due to being outdated which causes some issues in the output reported by this procedure. These issues were observed for the current data analysis using CATMOD procedure so is not reported here. A sample code using the NLMIXED procedure in the presence of a few of the predictors is presented here as including every predictor and writing the corresponding likelihood function is very time consuming. The call to a sample PROC MLMIXED is displayed:

```
PROC NLMIXED DATA= Cat_mental;
        PARMS a1=0.1 a2=0.1 a3=0.1 b1=0.1 b2=0.1 b3=0.1;
        /* Linear predictors */
        eta1 = a1 + b1*(4-1)*hsur_00 + b2*(4-1)*O_C0 + b3*(4-1)*DEP0;
        eta2 = a2 + b1*(4-2)*hsur_00 + b2*(4-2)*O_C0 + b3*(4-2)*DEP0;
        eta3 = a3 + b1*(4-3)*hsur_00 + b2*(4-3)*O_C0 + b3*(4-3)*DEP0;
        /* Define likelihood */
        IF length=1 THEN prob= exp(eta1)/(1 + exp(eta1) + exp(eta2)+exp(eta3));
        IF length=2 THEN prob= exp(eta2)/(1 + exp(eta1) + exp(eta2)+exp(eta3));
        IF length=3 THEN prob= exp(eta3)/(1 + exp(eta1) + exp(eta2)+exp(eta3));
        IF length=4 THEN prob= 1 /(1 + exp(eta1) + exp(eta2)+exp(eta3));
        /* To make sure that probabilities are valid ones */
        p = (prob>0 and prob<=1)*prob + (prob<=0)*1e-8 + (prob>1);
        loglike = log(p);
        /* Specify distribution for response variable */
        MODEL length ~ GENERAL(LOGLIKE);
RUN;
```

There exist even more problems when running models using a continuation ratio logit function using PROC CATMOD which is suggested to be used by some references due to the same reason of being outdated and so not providing very reasonable output. Agresti (2013) suggests using PROC GENMOD for the continuation-ratio logit models which performs better than PROC CATMOD but still is not as easy to use as when using cumulative logit as the logit function within the ordinal logistic models. 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. With the restructured dataset and the created binary response variable PROC LOGISTIC produces can provide the same results as NLMIXED. Within this procedure the PARAM=GLM coding in the CLASS statement should be used rather than as an option on the MODEL statement (High, 2013).

Unfortunately use of ordered information is not very common among researchers in different fields probably due to these software packages' limitations and models' complexity which were discussed above, but this should not stop using such models when dealing with ordered responses due to the higher power of the ordinal hypothesis tests. Applying the different options provided in this study on the real data within SAS shows that cumulative logit function within ordinal logistic models is the easiest one to perform in SAS. However, sometimes due to the type of results we are hoping to get based on the research questions and also based on specific interpretations we want to present, other logit functions might be more appropriate. Hopefully more straightforward procedures within SAS will be available to researchers for the adjacent-categories and continuation-ratio logit functions in future.

CORRELATED DATA WITHIN ORDINAL AND MUTINOMIAL LOGISTIC MODELS

When dealing with longitudinal (or repeated measures) data, the regular ordinal multinomial logistic regression models which were discussed above do not work for the categorical responses. The situation is almost similar to when modeling the correlated binary outcome in which the regular binary logistic models were not appropriate anymore since the assumption of observations independence was violated. Specific models are required for a multinomial response that is observed more than once on each subject, either at multiple times or under multiple conditions. Three primary types of models which can be used to handle this kind of complexity are marginal, mixed-effect, and transitional models which some of them are discussed below.

The GEE method which was mentioned above when discussing the models for correlated data with binary response is also appropriate for modeling correlated responses with more than two possible outcomes. GEE is a common choice for marginal modeling of ordinal response for correlated data if one is interested in the regression parameters rather than variance-covariance structure of the longitudinal data. The covariance structure is considered as nuisance within a GEE model. In this regard, 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. GEE has the advantage of allowing missing values within a subject without losing all the information from the subject which results in a higher power. Note that the GEE estimation method is not a maximum likelihood method.

ALR is another modeling option that models the association among responses with odds ratios rather than correlations resulting in parameter estimates of the model on the log odds ratios among the measurements. Similar to GEE model, the ALR method provides estimates of the marginal model parameters; however, it does not restrict the correlation among the repeated measurements as the GEE method does.

Cluster model with variance adjustment is another marginal model that uses maximum likelihood to fit the model to categorical correlated responses. In order to take into consideration the correlated nature of the data, the variances need to be adjusted within this model using the cluster structure of the data. Using the Morel adjustment within this model will result in a better fit than the GEE model for a small number of clusters. This adjustment can be easily made by using VARADJUST=MOREL option within the SAS procedure used to perform this model which will be discussed in an example later.

Random effects logistic model also known as a generalized linear (or nonlinear) mixed model allows random effects in a logistic model resulting in a subject-specific model. This is a conditional model that can also be used to model longitudinal or repeated measures data.

Fixed effects with conditional logit analysis is a conditional subject-specific modeling technique for correlated observations which treats each measurement of each subject as a separate observation. The set of subject coefficients that would appear in an unconditional model are eliminated by conditional methods. This model is more appropriate for binary correlated response rather than multinomial correlated outcome which is the topic of this section. This model however has some advantages which makes it important to discuss and even to use for modeling longitudinal ordinal responses by adopting the cumulative logit within the model which in fact dichotomizes the categories of the response. According to Allison (2012), GEE does not correct for the bias resulting from omitted explanatory variables of the cluster level. When working with correlated data as a result of repeated measurements per subject, statistically controlling for all stable characteristics of subjects of the study is possible by adding an intercept, α_i, which is the same for a given subject at multiple time points. This term implements a positive correlation among the observed outcome. The general model will be as below

$$\log\left(\frac{p_{it}}{1-p_{it}}\right) = \alpha_i + \beta x_{it},$$

where p_{it} within the logit function is the probability of having a outcome of interest for subject i at time point t and α_i represents all differences among individuals that are stable over time. If this term is treated as a random effect with any distribution such as normal, this model will be a random-effects or mixed model which was explained before within the homeless data example and can be modeled using GLIMMIX macro. Within this model, a conditional likelihood may be used which is identical to conditional logit model which is very similar to certain discrete-time survival models that can be estimated within a PHREG procedure.

Different procedures within SAS using aforementioned models are discussed below to provide some sample codes for the analysis of the mental health data introduced above.

Example (ordinal multinomial outcome, correlated data): Mental Health Data

Using the mental health data introduced before, by not aggregating the repeated records of some individuals who had to come back to the mental healing facility for more treatments, multiple observations from the same clients appear in this dataset. This means a higher sample size and increased power of the study because of additional information which are being added to the model by subjects with multiple visits; therefore, models that account for the correlation among the measurements of the same individuals are required.

As described above, GEE is an appropriate model in the presence of longitudinal data. To fit the GEE model, one needs to specify the REPEATED statement within the GENMOD procedure. In PROC GENMOD, the DIST=MULT option must be used within the MODEL statement to request ordinal multinomial logistic models for the ordinal length of stay in the mental healing facility which is used as the response variable for this study. PROC GENMOD may be performed as below:

```
PROC GENMOD DATA=Cat_RepMental RORDER=data DESCENDING;
        CLASS length (REF="1") SA Personality race (REF = "Caucasian/") marital_status (REF="S")
        subject_ID;
        MODEL length= SA Personality futs_00 hsur_00 mhrm_00 Sibr_00 race marital_status O_C0
        DEP0 ANX0 HOS0 GSI0 PSDI0 / DIST=MULTINOMIAL LINK=CUMLOGIT;
        REPEATED SUBJECT=subject_ID / CORR=UN;
RUN;
```

The ordinal multinomial model is available in PROC GEE beginning in SAS 9.4 TS1M3. One can use the TYPE= option in the REPEATED statement to specify the correlation structure among the repeated measurements within a subject.

```
PROC GEE DATA= Cat_RepMental DESCENDING;
        CLASS length (REF="1") SA Personality race (REF = "Caucasian/") marital_status (REF="S")
        subject_ID visit;
        MODEL length= SA Personality futs_00 hsur_00 mhrm_00 Sibr_00 race marital_status O_C0
        DEP0 ANX0 HOS0 GSI0 PSDI0 visit/ DIST=MULTINOMIAL;
        REPEATED SUBJECT=subject_ID / WITHIN=visit;
RUN;
```

Variable “visit” is being added to this procedure only to be used in the WITHIN option to specify the order of the measurements being recorded in multiple visits. WITHIN defines an effect that specifies the order of measurements within subjects. Each distinct level of the within-subject-effect defines a different response from the same subject. If the data are in proper order within each subject, specifying this option is not necessary. If some measurements do not appear in the data for some subjects which is what has happened within the current dataset, this option properly orders the existing measurements and treats 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 are not used; the remaining observations are then ordered in the sequence in which they are provided in the original input data set. Not specifying the WITHIN= option in a weighted GEE method will impose that the observations are ordered in the sequence in which they are provided in the input data set.

The cumulative logit link function is the default option that is used to fit the model. If the LINK=GLOGIT is specified in the MODEL statement, the generalized logit link function would be used and the nominal multinomial model would be performed here. Due to the fact that the outcome is ordered here, the generalized logit link function is not specified, so the responses are treated as ordinal multinomial data.

By default, TYPE=IND is used within the GEE procedure mentioned above. When trying to fit the ALR method, using the option LOGOR= is required and TYPE= should not be specified anymore. The ALR method can be fitted to the dataset as below:

```
PROC GEE DATA= Cat_RepMental DESCENDING;
        CLASS length (REF="1") SA Personality race (REF = "Caucasian/") marital_status (REF="S")
        subject_ID visit;
        MODEL length= SA Personality futs_00 hsur_00 mhrm_00 Sibr_00 race marital_status O_C0
        DEP0 ANX0 HOS0 GSI0 PSDI0 visit/ DIST=MULTINOMIAL;
        REPEATED SUBJECT=subject_ID / WITHIN=visit LOGOR=EXCH;
RUN;
```

Specifying LOGOR=EXCH in the REPEATED statement to select the ALR method emphasizes the fact that ALR has a fully exchangeable model for the log odds ratio. In addition to using PROC GEE and specifying the LOGOR= rather than TYPE= in the REPEATED statement of PROC GEE to fit the ALR model, the same specifications may be made within PROC GENMOD to fit the same model.

PROC GEE also implements the weighted GEE method when missing responses depend on previous responses. Weighted GEE is not available in PROC GENMOD. A GEE model, estimated by residual pseudo-likelihood, can also be fit using PROC GLIMMIX by specifying the EMPIRICAL option in the PROC GLIMMIX statement. Furthermore, specifying the RANDOM _RESIDUAL_ statement with the subject variable in the SUBJECT= option is required.

In order to fit the cluster model with variance adjustment which was discussed above, the same steps need to be taken only within PROC SURVEYLOGISTIC with the CLUSTER statement being used and VARADJUST= being optionally stated in the MODEL statement.

Random effects logistic model can also be fitted to this dataset through PROC GLIMMIX. A logistic regression model for categorical response variables can be imposed by specifying DIST=MULT LINK=CLOGIT for an ordinal model with a cumulative logit function, or by stating DIST=MULT LINK=GLOGIT for a nominal logit model in the MODEL statement. RANDOM statement can be used to define random effects. This model can also be fitted in PROC NLMIXED by using a different methodology that typically limits the number of random effects to one or two. Only binary responses are directly supported by specifying BINARY(p) or BINOMIAL(n,p) in the MODEL statement, though multinomial models can be accommodated by defining the multinomial log likelihood via the GENERAL distribution type in the MODEL statement which is a little more complicated.

Finally the fixed effects with conditional logit analysis which was discussed above may be fitted in SAS through a PHREG procedure due to the similarity between the likelihood function used in this model and the likelihood function for stratified Cox regression analysis. The sample code may be written as below:

```
PROC PHREG DATA= Cat_RepMental;
        MODEL length= SA Personality futs_00 hsur_00 mhrm_00 Sibr_00 race marital_status O_C0
        DEP0 ANX0 HOS0 GSI0 PSDI0 / TIES=DISCRETE;
        STRATA subject_ID;
RUN;
```

If the clients were falling into the same category of the length of stay each time they were visiting the mental healing facility, the TIES=DISCRETE option would be unnecessary.

CONCLUSION AND FUTURE WORK

Different options for modeling binary responses for cross-sectional and longitudinal data were discussed above as there are different procedures developed in SAS that can appropriately model dichotomous outcomes. The important point which needs to be considered by applied researchers is the necessity of accounting for the correlation that exists among observations when modeling longitudinal data as the existence of repeated measurements results in the violation of the independence assumption; therefore, the regular logistic models cannot appropriately model longitudinal data. Correctly using the methods that take into consideration the autocorrelation among observations will result in more informative and powerful models which some were described in this study.

On the other hand, ordinal multinomial response variables can create many challenges in data analysis in the presence of correlated data. While current methods which some of them were mentioned above are able to address repeated measurement issues in ordinal logistic models, many are limited in different types of logit functions when it comes to adjacent-categories and continuation ratio logit functions. Unfortunately, there still exists no straightforward procedure in SAS that could perform the ordinal logistic models using the aforesaid two logit functions as easy as using the cumulative logit function in the presence of repeated measurements especially when the data is unbalanced. Therefore, more general models such as GEE needs to be performed to model ordinal longitudinal responses which still have some limitations if one desires to use some logit functions which are not defined within these procedure.

The author is still working on models that could easily handle this type of data such as multilevel proportional odds model to compare them with the existing models. By comparing such models and evaluating their performance, more straightforward recommendations in terms of modeling options within SAS can be provided in early future.

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: Willey.

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

Bena, J., McIntyre, Sh. 2008. Survival Methods for Correlated Time-to-Event Data.

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.

Lall, R., Campbell, M. J., Walters, S. J., Morgan, K., & Co-operative, M. C. (2002). A review of ordinal regression models applied on health-related quality of life assessments. *Statistical methods in medical research*, *11*(1), 49-67.

Le Cessie, S., and Van Houwelingen, J. C. 1994. Logistic regression for correlated binary data. *Applied Statistics*, 95-108.

Liang, K. Y., and Zeger, S. L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13-22.

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review/Revue Internationale de Statistique*, 25-35.

Prentice, R. L. 1988. Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 1033-1048.

Ramezani, N. 2014. Comparison of Survival Analysis and Logistic Regression for Correlated Data. In JSM Proceedings Biometrics Section. Alexandria, VA: American Statistical Association, pp. 2495-2504.

Ramezani, N. 2015. Approaches for Missing Data in Ordinal Multinomial Models. In JSM Proceedings, Biometrics Section. Alexandria, VA: American Statistical Association, pp. 2809-2823.

Williams, R. L. (1995). Product-limit survival functions with correlated survival times. *Lifetime data analysis*, *1*(2), 171-186.

Zeger, S. L., Liang, K. Y., & Albert, P. S. 1988. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 1049-1060.

CONTACT INFORMATION

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

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