Using your FREQ effectively:

Displays for Deciphering Proportional Odds in Ordinal Regression

Robert G. Downer, Grand Valley State University, Allendale, MI

ABSTRACT

The proportional odds assumption of the cumulative logit model is an intriguing challenge in the modeling of an ordinal response. Valuable insight for modeling decisions can be gained by further investigation of why the proportional odds assumption has been satisfied or not. This investigation can be exploratory and completely separate from the logistic modeling. Empirical cumulative logit plots are one possibility but the interpretation is not intuitive with respect to odds or proportions. This paper presents exploratory methods which enhance the toolbox for understanding the proportional odds test. In conjunction with other SAS procedures, effective tabular and graphical options of PROC FREQ are used to support the findings of the proportional odds test. An informal application of the Breslow Day Test is introduced. Model development is not a focus of the paper but some PROC LOGISTIC details are discussed.

INTRODUCTION

As described in Downer (2013) and applied statistics references such as Agresti (2013), a standard binary logistic regression model with two response categories expresses the log odds of one response category (versus the other) as a linear function of the predictor variables. With p representing the probability of the characteristic of interest, the binary logistic regression model for predictors $X_1....X_k$ is expressed as:

$$\log\left(\frac{p}{1-p}\right) = \beta_o + \beta_1 x_1 + ... \beta_k x_k$$

The estimated coefficients $\hat{\beta}_1, \hat{\beta}_2, ... \hat{\beta}_k$ can be interpreted in terms of the log-odds (logit, the left hand side) or odds of the response of interest.

As described in Agresti (2013) and previously discussed in Downer (2016), a nominal response of J categories typically motivates the use of a generalized logit model where there are (J-1) intercepts and (J-1) slopes for each predictor coefficient. The interpretation of the modeling involves comparison to the Jth response category. Such a model is specified by the option LINK = GLOGIT in PROC LOGISTIC.

A cumulative logit model has traditionally been utilized for modeling an ordinal response. For a single continuous predictor x, the cumulative logit

$$logit(P(Y \le i)) = log\left(\frac{P(Y \le i)}{1 - P(Y \le i)}\right) = \alpha_i + \beta x$$

modelsthe log-odds of the cumulative probability of falling in the ith category i or below (as opposed to its complement). The model can also be expressed in the reverse manner where one considers the log-odds of falling in a given category or above. For J levels of the ordinal response, there are still J-1 intercepts as in a generalized logit model. However, this traditional cumulative model assumes proportional odds for each predictor coefficient. There is only one slope for each predictor term and hence the cumulative log-odds for a change in x will be the same for each of the J-1 response categories.

The default multi-category model in PROC LOGISTIC is this proportional odds cumulative logit model. Regardless of the number and nature of the predictors, a single proportional odds test is provided as part of the default output from a model fit. A generalized logit model no longer has to be the next modeling step if the proportional odds test is significant. As described in Derr (2013), Hilliard (2017) and Lund (2017), partial proportional cumulative logit models are now possible through the UNEQUALSLOPES option on the model statement in PROC LOGISTIC.

This paper focuses on exploratory techniques which facilitate an understanding of the proportional odds test in a traditional cumulative logit model. The illustrations presented can be conducted separate from the modeling itself. Options and output associated with PROC FREQ are discussed in conjunction with other SAS procedures. The displays support the outcome of proportional odds testing with an objective of providing interpretation that is appropriate for a general audience. The illustrations of this paper are discussed within the context of the proportional odds assumption associated with cumulative logit modeling. However, the tabulations and graphical displays have general application for data involving proportions, odds and relative risk. For more detailed applications of some recent tabular and graphical features of PROC FREQ, see McHawthorn (2017).

APPLICATION DATA

A randomly chosen subset of 2000 observations from the New Immigrant Survey (NIS) was utilized to illustrate the displays of interest.. Versions of this public data set are available via registration through the Office of Population Research (OPR). The goals and description of the study can be found at http://nis.princeton.edu/project.html.   Research papers and goals focusing on immigration can be found in Guillemena et al (2006), (2014).

This data was also used in Downer (2016) for investigating SAS modeling performance involved with collapsing a multi-category response into a binary response. As a measure of current living conditions of the immigrants (a categorization of whether housing is paid by the immigrant) was utilized as the response. This categorization is considered ordinal in this paper. A response value of 3 for *pydwell* indicates that the immigrant owns a home, a 2 indicates that the immigrant is renting and pydwell = 1 indicates that the housing status is some other living situation. The predictors considered in this paper are *age* (continuous), *numinh* (number of individuals in household (an integer considered as continuous) and *adjustee* (1=visa status changed after entering the USA, 0 otherwise)

Two different sets of predictors are used within cumulative logit models to illustrate the exploratory techniques. In the first example, the investigation considers just a single continuous predictor in modeling the odds of an immigrant being in a better (or worse) housing situation. In the second example, a different continuous predictor (numinh) is considered along with one categorical predictor (adjustee) and their interaction. In each modeling investigation, all the predictor terms are significant.

EXAMPLE 1: ONE CONTINUOUS PREDICTOR

An initial cumulative model in PROC LOGISTIC used age as the only predictor of housing status and it resulted in a significant proportional odds test (p < .0001). An application of PROC MEANS gave the first, second and third quartiles of the immigrant age distribution of this data set as 29, 36, and 47 respectively. Subsets of the data based on quartiles were explored. A model using only the interquartile range (IQR) of the age distribution (the middle 50 percent of the ages) also gave a significant proportional odds test (p=.0006). Insignificance was found within models based only on the first and last quartiles of the age distribution.

As suggested in SAS Note (2012), Derr (2013), Hilliard (2017) and Kelly (2017), empirical logits have

become a way to illustrate the possible violation of the proportional odds assumption. The categorical variable *agecat* was defined with 4 levels based on the age quartiles and the observations were binned accordingly. The cumulative empirical logits based on the 4 age categories are displayed in Figure 1. The syntax and code for the empirical logit plot displayed below are provided as an Appendix to this paper.

The variable cumulpd1 in Figure 1 is a computation of the empirical logit for Pydwell =1 (versus 2 or 3) while cumulpd2 is an empirical logit of pydwell =1 or 2 versus pydwell = 3. The non-parallel empirical logits before and after the median (categories 2 and 3) support the significance of the proportional odds test for the model which used the interquartile range of the age distribution. Unfortunately, the y axis scale of an empirical logit plot is not intuitive and difficult to interpret for most audiences.
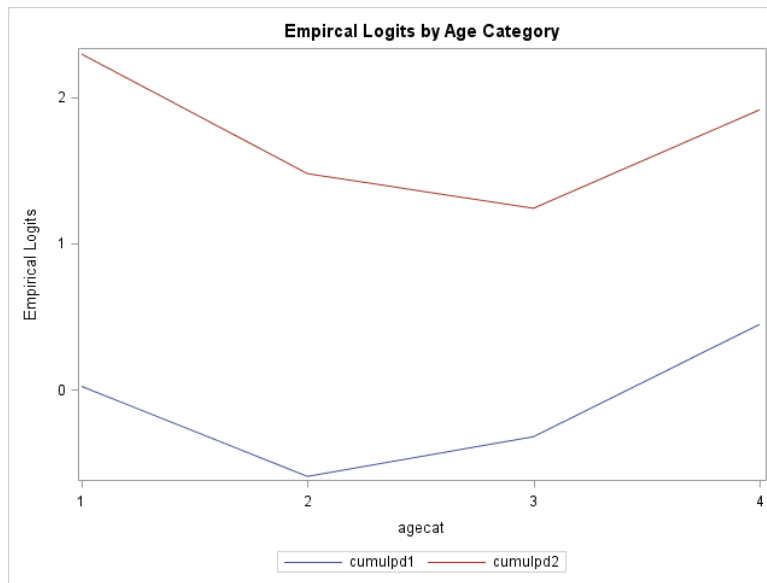

Figure 1: Empirical Cumulative Logits by quartile age category

Output providing cumulative percentage is an improvement with respect to interpretation and can be produced in PROC FREQ using a CUMCOL option on the TABLES statement. In this example model, quartiles are a reasonable step for binning observations of a continuous variable and the quartiles of the age distribution were used for binning as agecat = 1,2,3,4 respectively. For ease of understanding of the cumulative percentages, other table cell entries are suppressed by using NOROW NOPERCENT and NOCOL as options on the TABLES statement below:

```
ods graphics on;
proc freq ;
tables pydwell*agecat / norow nopercent nocol cumcol plots=mosaicplot;
run;
ods graphics off;
```

In the resulting PROC FREQ output given as Table 1 below, the pattern of the cumulative percentage within pydwell = 1 and pydwell = 2 differ as one goes through the quartile age categories (observations with missing age values were removed). For pydwell = 1, the cumulative percentage increases as one moves from quartile 2 to quartile 3 in age (35% to 42%) . For cumulative age percentage defined by the renting response category (Y<=2), the reverse is true (81 % to 77%). The odds of being less than or equal to a particular value of Y appears to not be the same as we go from the 2nd to 3rd age categories. This exploratory evidence again supports the significance of the proportional odds test of the logistic

model within the IQR of the continuous predictor age.

Table of pydwell by agecat					
pydwell	**agecat**				
Frequency Cumulative Col%	**1**	**2**	**3**	**4**	**Total**
1	265 50.57	178 35.60	206 42.04	289 60.97	938 47.18
2	211 90.84	229 81.40	174 77.55	124 87.13	738 84.31
3	48 100.00	93 100.00	110 100.00	61 100.00	312 100.00
Total	524	500	490	474	1988

Table 1: PROC FREQ output of cumulative response percentages by quartile age category

With ods graphics invoked in the previous PROC FREQ run, the MOSAICPLOT graphics option within the TABLES produced Figure 2 below. This display also supports the proportional odds test but the interpretation is more difficult without the cumulative percentages given in Table 1. The green area for pydwell = 1 increases as one moves from age category 2 to age category 3 but the cumulative area of red and green (pydwell = 1 or 2) decreases as we move between these same age categories.
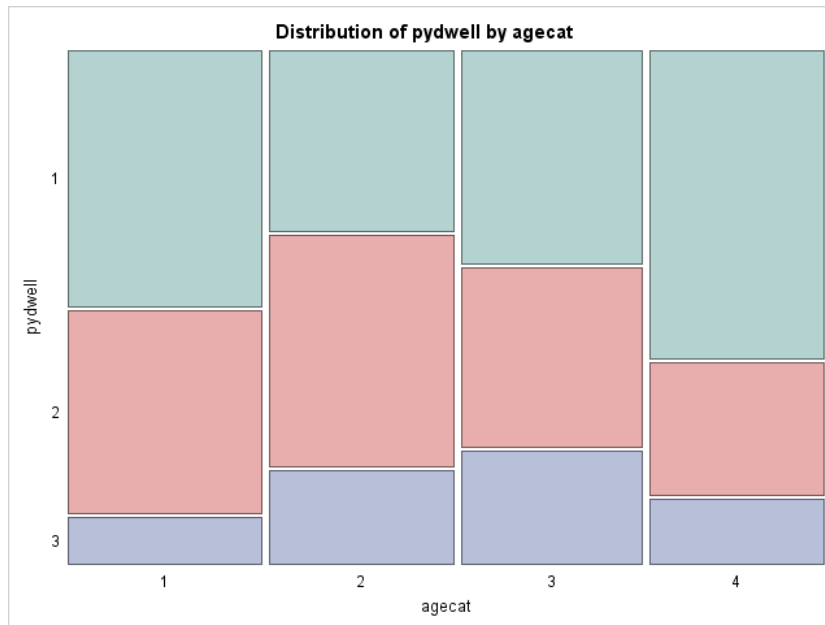

Figure 2: Mosaic display of housing status response proportions by age quartile

EXAMPLE 2: ONE CONTINUOUS & ONE CATEGORICAL PREDICTOR

The modeling of the housing status response with 2 predictors (number in household as continuous , adjustee status as categorical) and their interaction is discussed in this example The following cumulative logit model fit from PROC LOGISTIC resulted in a significant interaction (p=.0007) and a significant proportional odds test (p < .0001):

```
proc logistic descending;
class adjustee/ param = glm descending;
model pydwell = adjustee numinh adjustee*numinh;
run;
```

Modeling could utilize an UNEQUALSLOPES option on the above MODEL statement and continue accordingly. For details on partial proportional models, see Derr (2013), Hilliard (2017) and Lund (2017).

As a result of the significant interaction, fitting separate models for adjustees and non-adjustees (with numinh as the only predictor) is still a logical next step. For non-adjustees, the following output was obtained:

Score Test for the Proportional Odds Assumption		
Chi-Square	DF	Pr > ChiSq
1.7312	1	0.1883

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	3	1	-1.8198	0.2171	70.2858	<.0001
Intercept	2	1	0.6298	0.1818	11.9964	0.0005
numinh		1	-0.2882	0.0434	44.0388	<.0001

Odds Ratio Estimates		
Effect	Point Estimate	95% Wald Confidence Limits
numinh	0.750	0.688	0.816

while the following was obtained for adjustees:

Score Test for the Proportional Odds Assumption		
Chi-Square	DF	Pr > ChiSq
5.4984	1	0.0190

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	3	1	-0.7532	0.1502	25.1614	<.0001
Intercept	2	1	0.8859	0.1509	34.4752	<.0001
numinh		1	-0.0756	0.0355	4.5245	0.0334

For illustration and ease of interpretation, the DESCENDING option was used on the PROC LOGISTIC line in the modeling. This allows interpretation of the cumulative logit modeling to now be the direction of a better housing situation (pydwell = 3) in both models. For the non-adjustees, the odds of being in a more positive housing situation significantly decreased as the number of household members increased (coefficient is negative and odds ratio confidence interval is completely less than 1). Since the proportional odds test is insignificant for non-adjustees, the same odds interpretation can be made regardless of whether we're discussing odds of owning a home (pydwell =3) as well as renting or owning (pydwell = 2 or pydwell = 3). The model fit for adjustees above shows some similarity in the nature of the relationship but the proportional odds assumption is violated for the adjustee group. For the adjustee group, an UNEQUAL SLOPES option for the numinh predictor could be also be attempted here (See Derr(2013), Hilliard (2017)

Why did the assumption fail?. Exploratory work by adjustee status may reveal the nature of the violation of the proportional odds assumption. It should be informative with respect to how adjustees and non-adjustees differed with respect the odds of a better living situation based on number of household members.

PROC MEANS revealed that the first quartile, median and third quartile of number of household members in the data set was 3, 4. And 5 respectively. For binning of this continuous predictor, using the quartiles is again a reasonable first step with respect to understanding the cumulative proportions.

PROC FREQ output of tables of the response (Pydwell) cross-classified with the quartile categories of the number of household members was produced. Output from this run of PROC FREQ was utilized as input for the displays produced by PROC SGPLOT in Figure 3 and Figure 4. As we move through categories of number of household members in the plot each adjustee group, the cumulative proportion does differ.
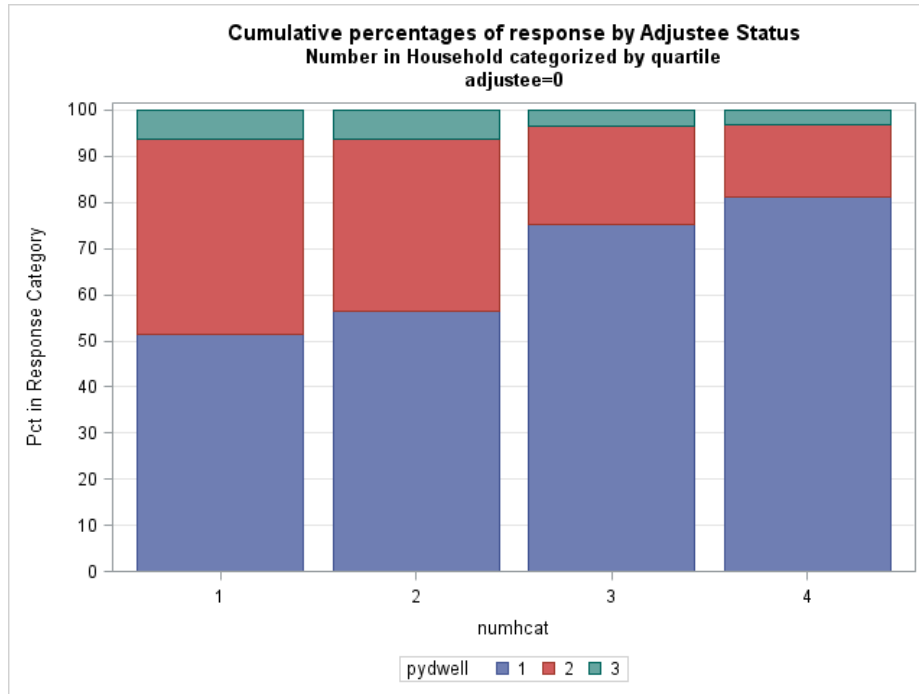
Figure 3: SGPLOT display of response fractions by number in household for non-adjustees
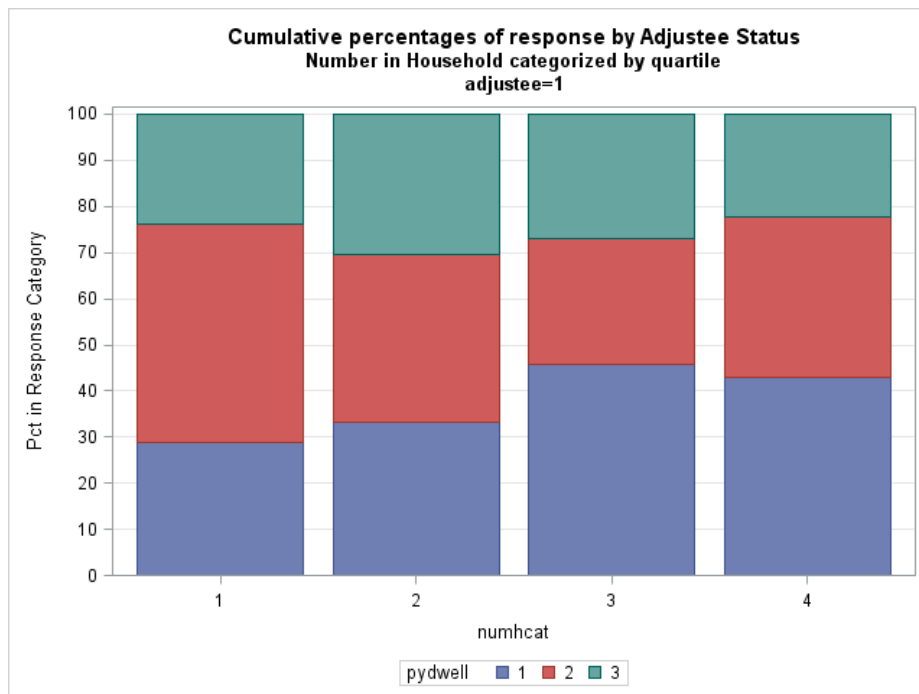

Figure 4: SGPLOT display of response fractions by number in household for adjustees

The SGPLOT displays given above in Figure 3 and Figure 4 are visually appealing and they may be more informative than mosaic plots that could be produced using similar PROC FREQ code as produced

Figure 2. The accompanying cross-tabulations from PROC FREQ with the CUMCOL option can supplement these figures and give some sample size context to the cumulative percentages. The proportional odds test is known to be more liberal as sample size increases.

The number of adjustees and non-adjustees associated with Figure 3 and Figure 4 (and corresponding cross-tabulation) is similar (960 versus 1028 after missing responses are delete).A much greater percentage of adjustees (66%) are either renting or owning (pydwell = 2 or 3) as compared to non-adjustees (38.3%) This cumulative percentage difference is even more pronounced if the number of household members is in the upper half of the household member distribution This corresponds to at least 5 household members. If we consider the sample sizes involved, there are 147 adjustees in these cells of that 3 x 4 table while there are only 70 non-adjustees with pydwell = 2 or 3 and at least 5 household members. As we change the number of household members, the proportional odds test (investigating the odds of renting or owning) was able to detect a difference with adjustees as a result of the larger sample size.

This interpretation was not evident from the LOGISTIC output on its own and the cross-tabulation enhanced the interpretation of the information in Figure 3 and Figure 4. The information provided by PROC FREQ is removing the mystery associated with the proportional test violation within PROC LOGISTIC.

For some contexts, having a supplementary proportional odds test with a straight forward interpretation may be desired. The standard Breslow Day test has the homogeneity of the k odds ratios as its null hypothesis (for k 2 x 2 tables). A split at the median number of household members made sense in this example based on the earlier exploration. Binning with a split at the median number of household members it would result in two 2 x2 tables for each of the 2 categories of adjustee.   One of the other quartiles could also be used but the median makes sense as an initial binning.

With 3 categories for the ordinal response, we can create two binary categories for Y regardless of the row categorization of Y for each table. The rows could be done as Y<=1 (with the complement as the other bin) and Y <=2 (with the complement as the other row bin) or the binning could be reverse based on Y=3 and Y<=2. The same two odds ratios will result. It's an informal test because the all of the observations will be in each of the two 2 x2 tables for each level of adjustee. A formal Breslow Day test would have independent strata and an observation would only be in one table.

For nonadjustees, a run of PROC FREQ with a RELRISK option for a Y split of Y=1 versus Y>=1 (and household members split as 4 or less (hcat =1, at most the median) versus at least 5 gave the following confidence limits for the odds ratio:

Statistic	Value	95% Confidence Limits	
Odds Ratio	2.2780	1.7468	2.9708

The odds of being in the worst housing category (as opposed to renting or owning) did change as we went from at most 4 to at least 5 household members. This supports the plot in Figure 3A for non-adjustees. However, a Y split Y<=2 versus Y =3 through PROC FREQ (and a measures option) gave odds ratio confidence limits below. This confidence interval contains 1, suggesting that the odds of being in a housing category of renting or worse does not differ as we change categories of number of household members.

Statistic	Value	95% Confidence Limits	
Odds Ratio	1.4300	0.8113	2.5205

The confidence intervals above overlap and this evidence supports homogeneity of the true odds ratios

for the 2 tables.. The Breslow Day test produced by the following code for non-adjustees is insignificant

```
proc freq
weight count;
tables cut*hcat*pd / cmh;
where ajustee= 0;
run;
```

Breslow-Day Test for Homogeneity of the Odds Ratio Adjustee = 0	
Chi-Square	2.1443
DF	1
Pr > ChiSq	0.1431

For adjustees, similar applications of PROC FREQ give confidence limits of (1.2481,2.1010) for the first Ysplit (Y<=1 versus Y > 1) and (0.6344, 1.1137) for the other split. This is consistent with the pattern shown in Figure 3B and the direction of the relationships shown for non-adjustees. The odds of being in the worse housing category (Y=1) differs significantly as we go to a higher number of household members (1 is not contained within the confidence interval). For the other split of Y (less than or equal to 2 versus 3), this significance is again not eveident (as with non-adjustees). The Breslow Day test is given below:

Breslow-Day Test for Homogeneity of the Odds Ratios Adjustee = 1	
Chi-Square	11.2781
DF	1
Pr > ChiSq	0.0008

For adjustees, the significant Breslow Day test is supported by non-overlapping confidence intervals and supports the proportional odds test within PROC LOGISTIC. For adjustees, we cannot assume that the odds of being in a worse (or better) housing situation is going to same as we change the number of household members. The odds ratios significantly differed for the two 2 x 2 tables which were formed by the two possible binary splits of Y and using a median binning of the number of household members as the other categorical variable.

CONCLUSION

Careful examination of appropriate displays can lead to solid understanding of the proportional odds test associated with the fit of a traditional cumulative logit model. Tabular and graphical exploration using PROC FREQ provides ample information to enhance the interpretation of this important aspect of ordinal regression modeling.

REFERENCES

Agresti, A. 2013. *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: John Wiley & Sons

Derr, B. 2013. Ordinal Regression Modeling, *Proceedings of the SAS Global Forum 2013 Conference*, Paper 446--2013. Cary, NC: SAS Institute, Inc.

Downer, R. G. 2013. "Improved Interaction Interpretation: Application of the EFFECTPLOT statement and other useful features in PROC LOGISTIC", MWUG 2013, *Proceedings of the Midwest SAS Users Group 2013 Meeting,* Paper AA08-2013. Cary, NC: SAS Institute, Inc.

Downer, R.G. (2016), "To be two or not be two, that is a LOGISTIC question", *Proceedings of the Midwest SAS Users Group 2016 Meeting*, Paper AA18-2016.

Guillermina, J., Douglas S. Massey, Mark R. Rosenzweig and James P. Smith (2006). "The New Immigrant Survey 2003 Round 1 (NIS-2003-1) Public Release Data". Funded by NIH HD33843, NSF, USCIS, ASPE & Pew. http://nis.princeton.edu.

Guillermina, J, Douglas S. Massey, Mark R. Rosenzweig and James P. Smith.(2014) "The New Immigrant Survey 2003 Round 2 (NIS-2003-2) Public Release Data". Funded by NIH HD33843, NSF, USCIS, ASPE & Pew. http://nis.princeton.edu.

Hilliard, P. (2017) "Using New SAS 9.4 Features for Cumulative Logit Models with Partial Proportional Odds"*Proceedings of the SAS Global Forum 2017 Conference*, Paper 406-2017. Cary, NC: SAS Institute, Inc.

Kelly, S. (2017) "Fitting a Cumulative Logistic Regression Model". *Proceedings of the SAS Global Forum 2017 Conference*, Paper 1108-2017. Cary, NC: SAS Institute, Inc.

Lund, B. (2017) "Weight of Evidence Coding for the Cumulative Logit Model". *Proceedings of the SAS Global Forum 2017 Conference*, Paper 785-2017. Cary, NC: SAS Institute, Inc.

Mclawhorn (2017) "Tables and Graphics That Will "FREQ" You Out". *Proceedings of the SAS Global Forum 2017 Conference*, Paper 404-2017. Cary, NC: SAS Institute, Inc.

SAS Institute Inc. (2012), "SAS Note 37944: Plots to Assess the Proportional Odds Assption in an Ordinal Logistic Model," http://support.sas.com/kb/37944. Cary, NC.

CONTACT INFORMATION:

Robert G. Downer

Biostatistics Director & Professor

Department of Statistics, Grand Valley State University

downerr@gvsu.edu

APPENDIX: EMPIRICAL LOGIT PLOT

```
proc transpose out= cumul;
by agecat; var count;
run;

data empirical;
```

```
   set cumul;
    cumulpd1=log(    sum(of col1-col1)/sum(of col2-col3)     );
    cumulpd2=log(    sum(of col1-col2)/sum(of col3-col3)     );
    run;

    title 'Empircal Logits by Age Category';
    proc sgplot data = a;
    series y=cumulpd1 x= agecat ;
    series y=cumulpd2 x= agecat;
    xaxis label= "Age Category";
    xaxis values= (1,2,3,4);
    run;


proc sgplot ;
     series y=cumulpd1 x= agecat;
     series y=cumulpd2 x= agecat;
     xaxis label= "Age Category";
     yaxis label="Empirical Logits";
    run;
```