Paper 1108-2017

Fitting a Cumulative Logistic Regression Model

Shana Kelly, Spectrum Health: Healthier Communities, Grand Rapids, MI

ABSTRACT

Cumulative logistic regression models are used to predict an ordinal response, and have the assumption of proportional odds. Proportional odds means that the coefficients for each predictor category must be consistent, or have parallel slopes, across all levels of the response. The paper uses a sample dataset to demonstrate how to test the proportional odds assumption, and use the UNEQUALSLOPES option when the assumption is violated. A cumulative logistic model is built, and then the performance of the cumulative logistic model on a test set is compared to the performance of a generalized multinomial model. This shows the utility and necessity of the UNEQUALSLOPES option when building a cumulative logistic model. The procedures shown are produced using SAS® Enterprise Guide 7.1.

INTRODUCTION

This paper focuses on building a cumulative logistic regression model that predicts the probability of a certain level of response on an ordinal scale. The type of model that can be built depends on satisfaction or violation of the proportional odds assumption. A sample dataset was used to predict the likelihood of a college student applying to graduate school. The response variable (APPLY) is whether a college junior will apply to graduate school, and is measured on a three point scale: unlikely, somewhat likely, and very likely. There are three explanatory variables in this dataset: whether at least one of the student's parents has a graduate degree (PARED), if the student went to a public or private undergraduate institution (PUBLIC), and the student's GPA on a continuous, four-point scale (GPA).

PROPORTIONAL ODDS ASSUMPTION

One assumption that should not be violated when performing a cumulative logistic regression is the assumption of proportional odds. This states that the relationship between every possible pair of response levels is the same. In other words, the coefficients for each predictor category must be consistent, or have parallel slopes, across all levels of the response. When the proportional odds assumption is violated in a cumulative logistic regression model, the model is typically run as a generalized multinomial logistic regression. Running a generalized multinomial model removes the ordinal aspect of the response variable, which may not be ideal in all situations, and reduces the quality of information that can be gathered from the response.

The proportional odds assumption can be checked using the LOGISTIC procedure. When the proportional odds assumption is not met for the overall model, each explanatory variable may then be checked individually in a simple logistic regression with the response:

```
proc logistic data=data;
   class apply pared public/param=reference;
   model apply=public;
run;
```

Repeat the above procedure for each explanatory variable. PROC LOGISTIC tests the proportional odds assumption and gives the corresponding chi-square p-value. If the p-value is significant, the proportional odds assumption is violated and a traditional cumulative logistic regression should not be run. When checking assumptions, it is better to be more conservative and use a higher alpha level of .10 or even .20.

Figure 1 shows the PROC LOGISTIC output that tests the proportional odds assumption for the explanatory variable PUBLIC.

Score Test for the Proportional Odds Assumption		
Chi-Square	DF	Pr > ChiSq
4.5887	1	0.0322

Figure 1: Test of the proportional odds assumption for the variable PUBLIC.

The proportional odds assumption can also be checked using graphical methods. Using PROC FREQ, a mosaic plot can be created to visually check violation of the proportional odds assumption. A mosaic plot displays the proportion of observations in the explanatory variable of interest versus the response. Violation of the proportional odds assumption occurs when the proportion of observations in each level of the response is not consistent across each level of the explanatory variable.

This process is demonstrated in the following code, and produces the output in Figure 2:

```
proc freq data=data order=formatted;
    table apply*public/plots=mosaicplot;
run;
```

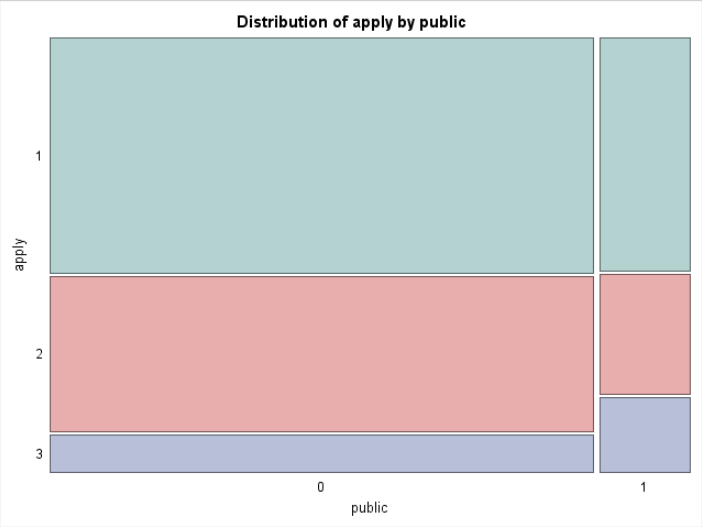

Figure 2: Mosaic plot to test the proportional odds assumption for the variable PUBLIC

Figure 2 shows that for those students with a response of “very likely” or “3”, the proportion of observations with a value of “1” for PUBLIC is much greater than the proportion with a value of “0”. This indicates that the proportional odds assumption is violated, and is consistent with the significant p-value of .0322.

Another graphical method to assess the proportional odds assumption is an empirical logit plot. For the assumption to not be violated, the curves of a predictor plotted against the empirical logits need to be parallel. There will be one less cumulative logit line than there are response categories. In this case, there are three response categories, so there will be two lines on each plot.

The process to calculate the empirical logits is demonstrated in the following code, and produces the output in Figure 3:

```
proc freq data=data;
    table public*apply / out=os;
run;

proc transpose data=os out=trana;
    by public;
    var count;
run;

data a;
    set trana;
    c1=log(sum(of col1-col1)/sum(of col2-col3);
    c2=log(sum(of col1-col2)/sum(of col3-col3);
run;

proc sgplot data=a;
    series y=c1 x=public;
    series y=c2 x=public;
    yaxis values=(-1 to 4);
run;
```

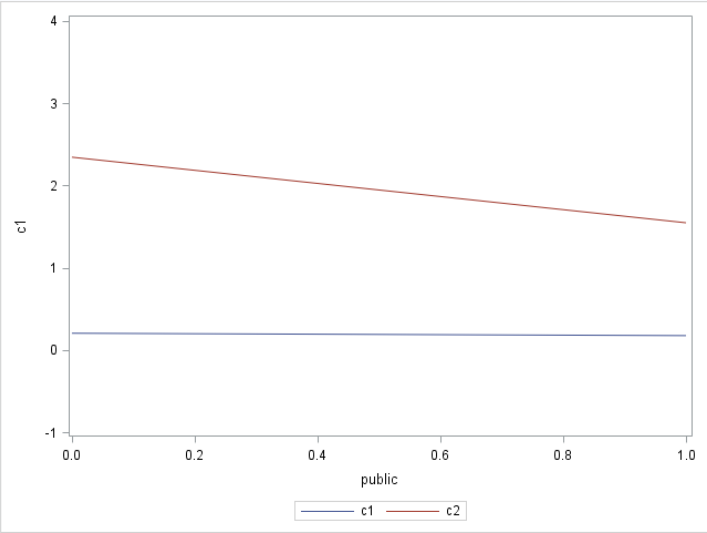

Figure 3: Empirical logit plot to test the proportional odds assumption for the variable PUBLIC

Figure 3 shows that the cumulative logits for the explanatory variable PUBLIC are not parallel, and the proportional odds assumption is violated. This is consistent with the conclusion from the mosaic plot, and the significant p-value of .0322.

BUILDING THE TWO MODELS

Testing each of the explanatory variables in simple logistic regressions with the response results in the variable PUBLIC violating the assumption with a p-value of .0322. PARED and GPA do not violate the assumption with p-values of .8964 and .3198; respectively. Before the UNEQUALSLOPES option was introduced in the LOGISTIC procedure, a generalized logistic regression had to be run if any of the explanatory variables violated the proportional odds assumption.

This process is demonstrated in the following code, and produces the output in Figures 4 and 5:

```
proc logistic data=data;
   class apply pared public/param=reference;
   model apply=pared public gpa/link=glogit;
run;
```

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
pared	2	13.8522	0.0010
public	2	3.0371	0.2190
gpa	2	4.9539	0.0840

Figure 4: Type III analysis of effects for Generalized Multinomial model

Due to the small number of explanatory variables, and the practical relationship between each and the response, all variables will be kept in the model regardless of their p-value.

Analysis of Maximum Likelihood Estimates							
Parameter		apply	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		very likely	1	-3.1132	1.6783	3.4408	0.0636
Intercept		somewhat likely	1	-1.3462	1.0469	1.6534	0.1985
pared	0	very likely	1	-1.3742	0.4222	10.5965	0.0011
pared	0	somewhat likely	1	-0.9516	0.3171	9.0082	0.0027
public	0	very likely	1	-0.3601	0.4435	0.6594	0.4168
public	0	somewhat likely	1	0.4188	0.3433	1.4884	0.2225
gpa		very likely	1	0.9240	0.4742	3.7972	0.0513
gpa		somewhat likely	1	0.4488	0.2902	2.3911	0.1220

Figure 5: Parameter estimates for generalized multinomial model

Using “unlikely” as the reference category, the final generalized models are:

$$log(\frac{\pi(very\ likely)}{\pi(unlikely)})=-3.11-1.37(PARED)-.36(PUBLIC)+.92(GPA)$$

$$log(\frac{\pi(somewhat\ likely)}{\pi(unlikely)})=-1.35-.95\ (PARED)+.42(PUBLIC)+.45(GPA)$$

Building a generalized logistic model with ordinal data diminishes the amount of information that can be gathered from the response, and disregards the ordinal nature of the data. Utilizing the UNEQUALSLOPES option in PROC LOGISTIC allows a cumulative logistic regression to be run, even when all the explanatory variables do not meet the proportional odds assumption.

This process is demonstrated in the follow code, and produces the output in Figures 4 and 5:

```
proc logistic data=data;
   class apply pared public/param=reference;
   model apply=pared public gpa/unequalslopes=public;
run;
```

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
pared	1	15.7449	<.0001
public	2	4.5935	0.1006
gpa	1	5.4821	0.0192

Figure 6: Type III analysis of effects for Partial Proportional Odds Cumulative Logistic Model

All explanatory variables will be kept in the model regardless of their p-value for consistency and better comparison of the two models.

Analysis of Maximum Likelihood Estimates							
Parameter		apply	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		unlikely	1	1.3432	0.9319	2.0776	0.1495
Intercept		somewhat likely	1	2.7797	0.9662	8.2769	0.0040
pared	0		1	1.0576	0.2665	15.7449	<.0001
public	0	unlikely	1	-0.2350	0.3053	0.5927	0.4414
public	0	somewhat likely	1	0.5733	0.4106	1.9490	0.1627
gpa			1	-0.6106	0.2608	5.4821	0.0192

Figure 7: Parameter estimates for cumulative logistic model

Modeling in the direction of "unlikely", the final cumulative logistic regression models are:

$$log(P(Y \leq \text{“Unlikely”}) = 1.34 + 1.06(PARED) - .24(PUBLIC) - .61(GPA)$$

$$log(P(Y \leq \text{“Somewhat Likely”}) = 2.78 + 1.06(PARED) + .57(PUBLIC) - .61(GPA)$$

CONCLUSION

Using the UNEQUALSLOPES option in PROC LOGISTIC when the proportional odds assumption is violated allows a cumulative logistic regression to still be run, and maintains the ordinal aspect of the response. This retains all information that can be gathered from the data, while maintaining the integrity of the model.

Plans for future work on this topic include further testing of the models' performance. This includes comparing concordance between observed and predicted responses in each type of model. Also, comparisons of model fit statistics such as the Akaike Information Criterion (AIC) will be done.

CONTACT INFORMATION

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

Shana Kelly
Spectrum Health: Healthier Communities
665 Seward Ave NW, Suite 110
Grand Rapids, MI 49504
Shana.kelly@spectrumhealth.org