Paper SD-87

Don't be binary! Tales of Non-binary Categorical Regression

Charlotte Baker, Florida Agricultural and Mechanical University

ABSTRACT

It is not always optimal to reorganize your data into two levels for regression. To prevent the loss of information that occurs when categories are collapsed, polytomous regression can be used. This paper will discuss situations in which polytomous regression can be used and how you can write the code.

INTRODUCTION

Regression analyses are regularly used to study the relationship between at least two variables. In particular, the analysis is completed to determine how much (if any) influence one variable (the independent variable) has on another variable (the dependent variable). Though linear, logistic, poisson, and cox proportional hazards regressions are regularly discussed and utilized in education and practice, it is not always beneficial to force data to fit one of these. For example, one may have conducted a survey where responses were captured on a five point likert scale (Very Good, Good, Neutral, Bad, Very Bad). Putting these categories together in a two level categorical variable might be easy to do, but the importance and value of the individual responses are lost in doing so. The solution may instead be the use of polytomous regression. This paper will discuss what polytomous regression is and how to use it.

WHAT EXACTLY IS POLYTOMOUS REGRESSION?

Also known as multinomial regression, polytomous regression is similar to logistic regression but with more than two levels of the dependent variable and a slightly different interpretation of the results.

There are many situations in which one might choose to use polytomous regression. The most common among these are times where valuable data can be lost if categories are dichotomized or when data is continuous and one intends to group parts together but two groups will not suffice. Polytomous regression can be used for ordinal or nominal data including non-continuous time categories, multi-response questions, and disease severity. Particularly when considering nominal data, polytomous regression can be useful to compare multiple levels of a response without creating additional models. Just as with any other regression, it is important to be familiar with the data and make sure your data meets the assumptions.

ASSUMPTIONS

The assumptions of polytomous regression include the independence of outcome levels (levels that are not too similar) and that the data are a random sample from a population with a multinomial distribution. The assumptions do not include normality, linearity, or homogeneity. If the data is ordinal, the assumptions also include meeting the proportional odds assumption. Methods of testing for these assumptions in SAS® are well documented elsewhere.

EXAMPLES

All examples in this paper are shown using the 2012 BRFSS data from the Centers for Disease Control and Prevention. All data and documentation used in these examples is freely available and the location of this information can be found in the References section of this paper.

POLYTOMOUS REGRESSION WITH ORDINAL DATA

If the data is ordinal, one is modeling the likelihood of each additional response versus the reference. For example, there is a multilevel outcome variable in the 2012 BRFSS data set named *_bmi5cat*. It indicates what category a respondent's body mass index (BMI) falls into. The variable has the levels Underweight, Normal Weight, Overweight, and Obese (coded in the data set as 1, 2, 3, and 4 respectively). The intended reference is Obese. In order to answer the question "How do exercise and sex influence BMI?" a polytomous regression model using cumulative logits might be in order. We will use a modified version of the variable *exerany2* for exercise and the variable *sex*.

There are some important things to remember when writing the code to answer this question using PROC LOGISTIC. First, SAS® automatically picks lowest alphabetical or numerical category as the referent category for each variable. This can easily be changed if that category is not the reference you want to use. It is very important to be sure that SAS® has picked the appropriate reference or the answers obtained will be incorrect. When using ordinal data and PROC LOGISTIC, changing the reference category is usually done by using the DESCENDING option if one wants to model the highest level of the dependent variable. If you have used formats for the dependent variable, the

categories may be out of the intended order. To solve this, one can use the ORDER = INTERNAL option or not use the formats for the dependent variable. If ORDER = INTERNAL and DESCENDING are used together, SAS® performs ORDER first then performs DESCENDING. If modeling the lowest level as the reference, the DESCENDING option is not needed because that is the default. For independent variables, one can also use similar and more specific options such as REF = or EVENT = to specify the reference category. If you have used formats, the value following the equal sign in REF = needs to be exactly the same as the formatted value or SAS® will not recognize it as a valid value for the reference.

Second, the CLASS statement is necessary to indicate what variables in the model are categorical variables. It must come before the MODEL statement.

Third, if one wants to use dummy coding instead of effect coding, the PARAM = REF option is necessary on the CLASS statement. Dummy coding is preferred if there may be interaction between categorical variables in the model. If there is no interaction, either coding type can be utilized with similar results. For more information on effect and dummy coding, see the UCLA Statistical Computing Group website in the References section.

Fourth, SAS® will automatically use the cumulative logits if it detects that there are more than 2 levels of the dependent variable. However, one can specifically request the use of the cumulative logits by using the LINK = CLOGIT option on the MODEL statement.

Below is an example of what the code may look like to answer the example question if we do not use any options:

```
proc format;
value exer 1="Yes" 2="No";
value _bmi5cat 1 = "Underweight" 2 = "Normal Weight" 3 = "Overweight" 4 = "Obese";
value sex 1 = "Male" 2 = "Female";
run;

data brfss2;
set brfss (keep = _bmi5cat sex exerany2);
exercise = exerany2;
if exercise in (7, 9) then exercise = .;
format exercise exer. _bmi5cat _bmi5cat. sex sex.;
run;

proc logistic data = brfss2;
class exercise _bmi5cat sex;
model _bmi5cat = exercise sex;
run;
```

The response profile table of the output indicates that, by default, SAS® is modeling the probability of being a lower BMI (underweight, normal weight, and overweight) as indicated by the ordered values 1, 2, and 3.

Response Profile

Ordered Value	_BMI5CAT	Total Frequency
1	Normal Weight	151545
2	Obese	127323
3	Overweight	162299
4	Underweight	7765

Probabilities modeled are cumulated over the lower Ordered Values.

Output 1. Response Profile Table in Polytomous Regression Output for Ordinal Data

If we are interested in predicting the likelihood of having a higher BMI (not a lower BMI), we need to make sure that SAS® uses the appropriate references. We alter our code to make SAS® reverse the levels of BMI so that the ordered values of 1, 2, and 3 refer to obese, overweight, and normal weight respectively.

```
proc logistic data = brfss2 descending order = internal;
class exercise (ref = "No") _bmi5cat sex (ref = "Female") / param = ref;
model _bmi5cat = exercise sex;
format _bmi5cat _bmi5cat.;
run;
```

As discussed above, we use ORDER = INTERNAL so that SAS® utilizes the order given by the values in the data set (1, 2, 3, and 4) instead of the formatted names. Because we are using a format, we also use DESCENDING so that SAS® will reverse the order of those numeric values and consider the numerically highest levels of the data to be the lowest ordered values.

Other options and a FORMAT statement can also be used in the PROC LOGISTIC if desired. A full listing of options available can be found in the SAS® 9.3 User's Guide.

We obtain the following results from PROC LOGISTIC:

Response Profile

Ordered Value	_BMI5CAT	Total Frequency
1	Obese	127323
2	Overweight	162299
3	Normal Weight	151545
4	Underweight	7765

Probabilities modeled are cumulated over the lower Ordered Values.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	Obese	1	-0.7057	0.00613	13249.3660	<.0001
Intercept	Overweight	1	0.8486	0.00618	18872.1853	<.0001
Intercept	Normal Weight	1	4.3205	0.0127	115851.980	<.0001
exercise	Yes	1	-0.5295	0.00642	6793.4315	<.0001
SEX	Male	1	0.3761	0.00560	4513.8409	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
exercise Yes vs No	0.589	0.582	0.596
SEX Male vs Female	1.457	1.441	1.473

Output 2. Response Profile, Analysis of Maximum Likelihood Estimates, and Odds Ratio Estimates Tables in Polytomous Regression Output for Ordinal Data

The output from the second example includes information on the number of response levels, the reference group, parameter estimates for the likelihood of each event, and odds ratios. We can confirm with the response profile table that SAS has reversed the order of the BMI variable and that we are indeed modeling the probability of having a high BMI. The remainder of the output looks very similar to the output for a logistic regression. Two differences from logistic regression can be found in the tenth and eleventh tables of the output.

In the tenth table in the output, we see a table of Analysis of Maximum Likelihood Estimates. Instead of having a single intercept as one gets with a logistic regression, there are N-1 intercepts (N being the number of dependent variable categories). The parameter estimates are interpreted similarly to those from a logistic regression. For every unit increase in the independent variable, there is a logit increase or decrease equal to the parameter estimate for the lower categories of the dependent variable versus the reference category given everything else in the model stays constant. In this case, for those that exercise compared to those that do not exercise the logit is 0.5295 lower for being underweight to overweight versus being obese. Simply put – those that exercise are less likely to be obese than those that do not exercise.

If we look at the eleventh table in the output, we see a table of Odds Ratio Estimates. Because we are not comparing two level variables, SAS® produces and labels these ratio estimates odds ratios but because they have a slightly different interpretation, they are best referred to as odds-like ratios instead of odds ratios. Since we modeled the likelihood of the lower ordered levels of BMI (in this case obese, overweight, and normal weight), we would say that the odds of being underweight to overweight versus being obese for those that exercise are 0.589 times more likely than for those that do not exercise, given all other variables in the model are held constant. Again – those that exercise are less likely to be obese than those that do not exercise.

POLYTOMOUS REGRESSION WITH NOMINAL DATA

If the data is nominal, one would be modeling the likelihood of each response level versus the other response levels. Just as with ordinal data, one of the levels of the dependent variable has to be the referent group. By default the lowest level is used as the referent but this can be changed to be any level of the variable. Polytomous regression with nominal data uses generalized logits instead of cumulative logits. This regression can be assessed with PROC LOGISTIC and PROC CATMOD. This paper will only discuss the use of PROC LOGISTIC to complete the analysis with the LINK = GLOGIT option on the MODEL statement.

We are interested in finding out if women are more likely to use their seatbelt when riding in a car. We will use the variable *seatbelt* and the variable *sex* from the 2012 BRFSS.

```
proc format;
value sex 1 = "Male" 2 = "Female";
value seatbelt 1 = "always" 2 = "nearly always" 3 = "sometimes" 4 = "seldom"
5 = "never";
run;

data brfss3;
set brfss2 (keep = sex seatbelt);
format seatbelt seatbelt. sex sex.;
label SEATBELT = 'HOW OFTEN USE SEATBELTS IN CAR?';
if seatbelt in (7,8,9) then seatbelt = .;
run;

proc logistic data = brfss3;
class sex (ref = 'Male') / param = ref;
model seatbelt (ref = 'Never') = sex /link = glogit;
run;
```

There is a slight difference in the PROC LOGISTIC output for nominal data. However, we still interpret the results by looking for the parameter estimates and the odds ratios (tables nine and ten of the output).

Analysis of Maximum Likelihood Estimates							
Parameter		SEATBELT	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		always	1	3.3981	0.0145	54584.3055	<.0001
Intercept		nearly always	1	1.4672	0.0159	8545.0520	<.0001
Intercept		seldom	1	-0.2491	0.0216	132.7432	<.0001
Intercept		sometimes	1	0.5367	0.0180	887.8869	<.0001
SEX	Female	always	1	1.0349	0.0237	1906.2253	<.0001
SEX	Female	nearly always	1	0.4466	0.0255	307.3172	<.0001
SEX	Female	seldom	1	-0.0293	0.0356	0.6763	0.4109
SEX	Female	sometimes	1	0.2645	0.0287	84.6863	<.0001

Odds Ratio Estimates				
Effect	SEATBELT	Point Estimate	95% Wald Confidence Limits	
SEX Female vs Male	always	2.815	2.687	2.949
SEX Female vs Male	nearly always	1.563	1.487	1.643
SEX Female vs Male	seldom	0.971	0.906	1.041
SEX Female vs Male	sometimes	1.303	1.231	1.378

Output 3. Analysis of Maximum Likelihood and Odds Ratio Estimates Tables in Polytomous Regression Output for Nominal Data

Notice that the output includes a parameter estimate for every comparison of the dependent variable categories versus the reference level. The parameter estimates are interpreted similarly to those from a logistic regression. For every unit increase in the independent variable, there is a logit increase equal to the parameter estimate for the category of the dependent variable versus the reference category given everything else in the model stays constant. In this case, for females compared to males the logit is 1.0349 higher for always wearing a seatbelt versus never wearing a seatbelt. Simply put - females are more likely than males to always wear a seatbelt than never wear a seatbelt. This remains evident when looking at the odds ratio estimates. More accurately these should be referred to as odds-like ratio estimates because the dependent variable is not binary. In this case, females are 2.815 times more likely to always wear a seatbelt than never wear a seatbelt compared to males. The other parameter estimates and odds ratio estimates can be interpreted in the same way.

OTHER CONSIDERATIONS

It is important to take into account model fit and model selection when utilizing any regression model. These can be *a priori*, determined from automated methods, manual methods or a combination of the above. Be sure that your model actually is representing and predicting the outcome that you are interested in predicting. If not, you may need to go back and change the references of your outcome or independent variables.

CONCLUSION

It is not always beneficial to dichotomize variables for analysis due to loss of information. Using polytomous regression allows one to take advantage of more than two levels of ordinal or nominal data and can be very easy to perform and interpret.

REFERENCES

- Centers for Disease Control and Prevention. BRFSS 2012 Survey Data and Documentation. 2013. Available at http://www.cdc.gov/brfss/annual_data/annual_2012.html
- UCLA: Statistical Consulting Group. What is effect coding? 2014. Available at http://www.ats.ucla.edu/stat/mult_pkg/faq/general/effect.htm
- UCLA: Statistical Consulting Group. SAS Data Analysis Examples - Ordinal Logistic Regression. 2014. Available at http://www.ats.ucla.edu/stat/sas/dae/ologit.htm

RECOMMENDED READING

- *Base SAS® Procedures Guide*
- Panik, Michael. 2009. *Regression Modeling*. pp 232-250. Boca Raton, FL: Chapman & Hall/CRC

CONTACT INFORMATION

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

Charlotte Baker
Florida A&M University Institute of Public Health
1515 S. Martin Luther King, Jr Blvd
Tallahassee, FL 32307
charlotte.m.h.baker@gmail.com