Polychotomous Logistic Regression and the Creation of Odds Ratios and Confidence Intervals Using the SAS PROC CATMOD

Besa Nushi, Naval Health Research Center, San Diego, CA

ABSTRACT

Objective To determine if differences in odds exist among levels in categorical independent variables with respect to a categorical multi-level dependent variable.

Methods Using a multivariate polychotomous logistic approach and the SAS System's PROC CATMOD, risk factors were assessed for a multi-level outcome variable over the time period of October 1, 1997 through September 30, 1998. This modeling was followed by using a short algorithm created to calculate odds ratios and 95% confidence intervals for each of the independent variables using the adjusted parameter estimates. This provided insight into the relationship between the variables and allowed determination as to whether certain strata were at increased or decreased odds of having a response at a particular level over the reference strata.

Conclusions Multiple level response variables can be handled efficiently by using a polychotomous logistic regression technique and the SAS System's PROC CATMOD. This procedure will output parameter estimates adjusted for the other independent variables, which can then be used to create odds ratios and associated confidence intervals.

INTRODUCTION

Logistic regression is a statistical procedure used to evaluate many independent variables (X_1, X_2, ..., X_p) in order to predict a dichotomous outcome. Generally this outcome is denoted as Y = 1 or Y = 0 for the two possibilities. Often, however, we desire to look at a multilevel response. Polychotomous logistic regression is an application of normal logistic regression extended from a binary outcome to greater than two outcomes. Individual logistic regressions can be calculated for each of the pairwise comparisons, however the polychotomous approach results in more accurate maximum likelihood estimations. Polychotomous logistic regression is valuable in that covariate parameters can be simultaneously estimated, in addition to allowing the direct testing of multiple categories. Primarily the application deals with nominal data, that is, data which can be classified into categories without any specific order value.

The SAS System's PROC CATMOD offers numerous approaches to analyzing categorical data. One such approach is polychotomous logistic regression. We use PROC CATMOD to output desired data sets, and through a set of systematic steps and the use of PROC TRANSPOSE, we can output odds ratios and confidence intervals.

A condensed form of the polychotomous method is offered here. First, recall that in simple logistic regression the probability of occurrence of one of the two outcomes is defined as:

$$P(Y=1) = \frac{1}{1+\exp\left[-\beta_0 + \left(\sum_{k=1}^{p} \beta_k X_k\right)\right]}$$

where k = 1, 2,...,p. The logit function is defined as:

$$\ln \frac{\left[P(Y=1|V)\right]}{\left[P(Y=0|V)\right]} = \beta_0 + \sum_{k=1}^{p} \beta_k X_k$$

Where Y = 0 denotes the reference category and V is a vector of the independent variables X_1 through X_p. We can now apply this to the polychotomous logit model where there are q mutually exclusive categories and again we denote Y = 0 as the reference category. We let Y denote the outcome coded as 0, 1,....., q-1. We then form the following q-1 logits where t = 1, 2,..., q-1:

$$g_t(V) = \ln \frac{\left[P(Y=t|V)\right]}{\left[P(Y=0|V)\right]} = \beta_{t0} + \sum_{k=1}^{p} \beta_{tk} X_k$$

If we wish to compare, for instance, Y = 1 to Y = 2, the logit can be attained by taking the difference of logit Y = 1 versus Y = 0 and the logit Y = 2 versus Y = 0.

The general expression for the conditional probability for each outcome given the vector of covariates can be written as:

$$P(Y= t \mid V) = \frac{\exp [g_t (V)]}{\sum_{j=0}^{q-1} \exp [g_j (V)]}$$

and therefore $\beta_0 = 0$ and $g_0 (V) = 0$.

Next we formulate the likelihood function. In doing so, we construct q binary variables. The variables are coded as 0 or 1 to designate whether the observation is a member of that group. When Y = q, $Y_q = 1$ and all other Y subcategories = 0. For example when Y = 0, $Y_0 = 1$, and all Y_1 through Y_{q-1} = 0. When Y = 1, $Y_0 = 0$, $Y_1 = 1$, and all Y_2 through $Y_{q-1} = 0$, and so on. Mathematically we find that the summation of all Y_q equals one for each observation. We can now continue with the likelihood function for a sample of n independent observations. P_t represents the conditional probability for each outcome given the vector of covariates; $P_t = P(Y= t \mid V)$. For the i^{th} observation, where β^* denotes the vector of the parameter coefficients:

$$l(\beta^*) = \prod_{i=1}^{n} \left[P_0^{y0i} P_1^{y1i} \ldots P_{q-1}^{y(q-1)i}\right]$$

Recall that $\sum Y_t = 1$, therefore $\sum Y_{ti} = 1$ for each i, and $Y_{0i} = 1 - Y_{1i} - Y_{2i} - \ldots - Y_{q-1i}$. We take the natural log of the likelihood function in order to find the log-likelihood function:

$$L(\beta^*) = \sum_{i=1}^{n} - \ln[1+ \exp(g_1) + \exp(g_2) + \ldots + \exp(g_{q-1})] + y_{1i}g_1 + y_{2i}g_2 + \ldots + y_{q-1i}g_{q-1}$$

We obtain the likelihood equation by taking the first partial derivative of the log-likelihood function:

$$\frac{\partial L(\beta^*)}{\partial \beta_{tk}} = - \sum_{i=1}^{n} x_{ki} (y_{ti} - P_{ti})$$

The maximum likelihood estimator, $\hat{\beta}$, is found by setting the above equation equal to zero and solving for β^*.

We can obtain the information matrix [2(p +1) by 2(p + 1)] if we take the second partial derivative of the log-likelihood function:

$$\frac{\partial^2 L(\beta^*)}{\partial \beta_{tk} \partial \beta_{tk'}} = - \sum_{i=1}^{n} x_{k'i} x_{ki} P_{ti}(1 - P_{ti})$$

and

$$\frac{\partial^2 L(\beta^*)}{\partial \beta_{tk} \partial \beta_{t'k'}} = - \sum_{i=1}^{n} x_{k'i} x_{ki} P_{ti} P_{t'i}$$

where t' and k' range over the same set of values as t and k. In a normally distributed large sample, the maximum likelihood estimators will have a covariance matrix equal to the inverse of the information matrix.

The most useful, and consequently the most widely used, application of logistic regression is its interpretation in the form of odds ratios. Recall our general logit form:

$$g_t (V) = \ln \frac{\left[P(Y= t \mid V)\right]}{\left[P(Y=0 \mid V)\right]} = \beta_{t0} + \sum_{k=1}^{p} \beta_{tk} X_k$$

where there is one logit for each of q-1 categories. We may use our estimation of β^*, the vector of the parameter coefficients to calculate the odds ratio:

$$\hat{OR}\ (tk) = \exp(\hat{\beta}_{tk})$$

For example, when comparing Y=1 to the reference category Y=0, we obtain the following logit:

$$g_1 (V) = \ln \frac{\left[P(Y=1 \mid V)\right]}{\left[P(Y=0 \mid V)\right]} = \beta_{10} + \sum_{k=1}^{p} \beta_{1k} X_k$$

from which we can estimate the odds ratio for the k^{th} independent variable coded dichotomously as 1 or 0:

$$\hat{OR}\ (1k) = \frac{P(Y= 1 \mid X_k = 1) / P(Y=0 \mid X_k = 1)}{P(Y= 1 \mid X_k = 0) / P(Y=0 \mid X_k = 1)}$$

or more simply:

$$\hat{OR}\ (1k) = \exp(\hat{\beta}_{1k})$$

For each odds ratio we can calculate a 100(1- α) level confidence interval:

$$\left[\exp(\hat{\beta}_{ik} - Z_{1-\alpha/2}\, S_{\hat{\beta}_{ik}}),\ \exp(\hat{\beta}_{ik} + Z_{1-\alpha/2}\, S_{\hat{\beta}_{ik}})\right]$$

where $S_{\hat{\beta}_{ik}}$ is the estimated standard error for $\hat{\beta}_{ik}$, and Z represents the standard normal for the specified level of alpha.

For hypothesis testing we utilize a log-likelihood ratio test statistic which compares the full model to a reduced model (all variables except those being tested). Let *d* represent the degrees of freedom for the independent variable(s) being tested. For example, a three category outcome covariate has (3-1) = 2 degrees of freedom. The statistic is as follows:

$$\chi^2 = -2\,[\text{log-likelihood (reduced model)} - \text{log-likelihood (full model)}]$$

When the null hypothesis is true, this test statistic follows a chi-squared distribution with $d(q\text{-}1)$ degrees of freedom.

ANALYSIS

We were interested in looking at an outcome that had clinical meaning at different levels.
Polychotomous logistic regression was used to analyze the cohort of 44,128 subjects. The outcome of interest was the result of a standardized medical test. A subject could be categorized as having a negative reaction, a low positive reaction, a high positive reaction, or having a history of a positive reaction. Negatives were used as the reference category. Covariates in the model included the following demographics: age, race/ethnicity, and gender. Place of birth was also included to assess geographical risk.

Age was categorized into approximate thirds in the following manner: 17 to 18 years, 19 to 20 years, or 21 years and older. Race and ethnicity were collapsed into one covariate with five categories: White, Black, Hispanic, Asian or Pacific Islander, and other. Gender was coded as male or female, and place of birth was coded as U.S. born or foreign-born. Dummy variables were created for all covariates.

The SAS code for the polychotomous logistic regression looked like this:

```
PROC CATMOD DATA=POLY;
  DIRECT GENDER AGEGRP1 AGEGRP2
    RACETH1 RACETH2 RACETH3 RACETH4 US;
  RESPONSE LOGIT / OUTEST=PARAMET;
  MODEL INDYPOL= GENDER AGEGRP1
    AGEGRP2 RACETH1 RACETH2 RACETH3
    RACETH4 US / NOPROFILE;
RUN;
```

Since we have formed our own dummy variables, DIRECT indicates to SAS to use the variables that we created, rather than making new ones.

RESPONSE indicates the function we want to model.

OUTEST option generates a SAS data set containing estimated β coefficients and the covariance matrix.

It is important to note, the number of parameters created in the output from PROC CATMOD. There should be one for each level of outcome variable (not including the reference category) at each of the dummy covariates and the intercept. In this case we had a 3 level outcome (excluding the reference category) and 8 covariates = 24 parameters + 3 for the intercept = 27 total parameters.

As PROC CATMOD outputs beta parameters, we created the following program to generate odds ratios and 95% confidence intervals:

```
DATA BETAS (DROP=_METHOD_ _TYPE_
    _NAME_);
 SET PARAMET (WHERE=(_TYPE_='PARMS'));
RUN;

PROC TRANSPOSE DATA=BETAS OUT=DBETAS;
RUN;

DATA BETAS2 (KEEP=ID BETA);
  SET DBETAS (RENAME=(COL1=BETA));
  ID=_N_;
RUN;

DATA COV;
  SET PARAMET (WHERE=(_TYPE_='COV'));
  ID=_N_;
RUN;

DATA COVAR (KEEP= ID COVAR _NAME_);
  SET COV;
```

```
   ARRAY B(27) B1 B2 B3 B4 B5 B6 B7 B8
    B9 B10 B11 B12 B13 B14 B15 B16 B17
    B18 B19 B20 B21 B22 B23 B24 B25 B26
    B27;
   DO I=1 TO 27;
    IF ID=I THEN DO;
       COVAR=B(I);
    END;
   END;
RUN;

DATA VARS (KEEP= ID SE _NAME_);
  SET COVAR;
  SE= (SQRT(COVAR));
RUN;

DATA VARPARM;
  MERGE
   VARS
   BETAS2;
  OR= EXP(BETA);
  LB= EXP(BETA - (1.96*SE));
  UB= EXP(BETA + (1.96*SE));
RUN;
```

From the dataset PARAMET, we created two sub-datasets; BETAS, which contains the beta coefficients, and COV, which contains the covariance matrix.

Since the beta coefficients were assembled in a row, we used PROC TRANSPOSE to reconstruct them into column format so they could be used in the calculations that followed.

Recall earlier we took note of the number of total parameters in our model. We used this to construct an array that enabled us to extract the diagonals from the covariance matrix in the COV dataset. Then we took the square root to obtain an estimate of the standard error. We then had the estimated beta coefficients and standard errors and could create the desired odds ratios and 95% confidence intervals.

INTERPRETATION OF OUTPUT

The following table is a summary of findings from the polychotomous logistic regression:

Characteristic	Standardized medical test										Overall p-value
	Negative[a]	Low positive			High positive			History of positive			
	N	N	OR[b]	CI[b]	N	OR[b]	CI[b]	N	OR[b]	CI[b]	
Gender											≤0.0001
Female[a]	8143	56			111			86			
Male	34,418	363	1.5	1.1-2.0	665	1.4	1.1-1.7	241	0.6	0.5-0.8	
Age											≤0.0001
17-18[a]	17,391	103			154			90			
19-20	15,068	146	1.5	1.1-1.9	259	1.7	1.4-2.1	89	1.0	0.8-1.4	
21 and over	10,102	170	2.0	1.6-2.6	363	2.7	2.2-3.3	148	2.0	1.5-2.6	
Race/Ethnicity											≤0.0001
White[a]	24,123	115			114			50			
Black	7940	74	1.7	1.3-2.3	154	3.3	2.6-4.3	51	2.4	1.6-3.6	
Hispanic	4745	82	1.8	1.3-2.5	181	3.6	2.8-4.7	83	4.0	2.8-5.9	
Asian/Pacific Islander	1826	86	2.2	1.6-3.1	213	4.7	3.6-6.2	98	5.6	3.7-8.4	
Other	3927	62	2.0	1.5-2.8	114	3.4	2.6-4.5	45	3.2	2.1-4.8	
Place of birth											≤0.0001
U.S.[a]	39,270	216			316			134			
Foreign	3291	203	7.4	5.8-9.5	460	8.6	7.2-10.3	193	7.6	5.8-10.0	

[a] reference category.
[b] OR, adjusted odds ratios from polychotomous logistic regression model; CI, 95%confidence interval.

The strongest relationship found was that foreign born subjects were 8.6 times more likely to have had a high positive reaction than U.S. born subjects. Almost all odds ratios were significant at the $\alpha = 0.05$ level.

As in other regression models, this model adjusted for all other covariates. Interestingly, in this example, this means that those with Asian or Pacific Islander race were 5.6 times more likely to have a history of positive reaction than Whites, even after adjusting for their place of birth, age and gender.

Another interesting finding was that, in the multivariable logistic regression model (not shown), where the outcome was dichotomously categorized into positive or negative, there was only a borderline significant difference between females and males. That is, males were 1.2 times more likely than females to be positive (p=0.0471). In the polychotomous logistic regression, we saw a stronger difference between genders. Males were, in fact, more likely to be either low or high positives (OR = 1.5, 1.4 respectively). However, males were less likely to have had a history of positive reaction (OR = 0.6). This finding was obscured in the multivariable logistic regression when outcome categories were collapsed.

CONCLUSION

Polychotomous logistic regression is an extremely useful tool in analyzing categorical data with a multi-level outcome. PROC CATMOD and further SAS coding enabled the production of odds ratios and 95% confidence intervals to assess demographic and geographic risk factors for reaction to a standardized medical test. The more in depth analysis of polychotomous logistic regression allowed for the determination of particular risk factors that differed between genders, and were not apparent in classic dichotomous logistic regression.

ACKNOWLEDGMENTS

Thanks to LCDR Margaret Ryan, CAPT Greg Gray, and Tyler Smith at the Naval Health Research Center, San Diego, and Dr. Donald Slymen at San Diego State University for their support and ideas.

This research was supported by the Department of Defense, Health Affairs, under work unit no. NMRDC.WR.00098(6423).

SAS software is a registered trademark of SAS Institute Inc. in the USA and other countries.

REFERENCES

Agresti, A. (1990). Models for Binary Response Variables, Categorical Data Analysis (pp. 79-129). New York, NY: John Wiley and Sons, Inc.

Dubin N, Pasternack BS. (1986). Risk assessment for case-control subgroups by polychotomous logistic regression. *Am J Epidemiol*, 123, 1101-1117.

Hosmer, D. W., & Lemeshow, S. (1989). Special Topics: Polytomous Logistic Regression, Applied Logistic Regression (pp. 216-238). New York, NY: John Wiley and Sons, Inc.

SAS Institute Inc., SAS/STAT® *User's Guide, Version 6, Fourth Edition, Volume 1,* Cary, NC: SAS Institute Inc., 1989. 943 pp.

ABOUT THE AUTHOR

Besa Nushi has used SAS for 3 years including work as an undergraduate student in Biology and Chemistry at California State University, Chico, as a graduate student at the San Diego State University Graduate School of Public Health in Biostatistics, and currently as a biostatistician with the Naval Health Research Center, San Diego. Her responsibilities include management and documentation of large military and demographic data sets, in addition to mathematical modeling and statistical analysis.