Cornell University
Cornell Statistical Consulting Unit

StatNews #91

Ordinal Logistic Regression Models and Statistical Software: What You Need to Know

June 2016

Ordinal logistic regression is a statistical analysis method that can be used to model the relationship between an ordinal response variable and one or more explanatory variables. An ordinal variable is a categorical variable for which there is a clear ordering of the category levels. The explanatory variables may be either continuous or categorical. Estimating ordinal logistic regression models with statistical software is not difficult, but the interpretation of the model output can be cumbersome.

Ordinal logistic regression is an extension of logistic regression (see StatNews #81) where the logit (i.e. the log odds) of a binary response is linearly related to the independent variables. If instead the response variable has k levels, then there are k-1 logits. A major assumption of ordinal logistic regression is the assumption of proportional odds: the effect of an independent variable is constant for each increase in the level of the response. Hence the output of an ordinal logistic regression will contain an intercept for each level of the response except one, and a single slope for each explanatory variable.

There are several ways in which an ordinal regression model can be parameterized and different statistical software packages use different parameterizations. Thus, great care should be taken when interpreting the output from ordinal regression models. We will consider an example to illustrate the different model parameterizations and corresponding interpretation for several commonly used statistical software packages.

Suppose that customers at a bedding store are asked to rate how comfortable they find a newly engineered mattress on a scale from 1 to 3; 1 for uncomfortable, 2 for comfortable, 3 for very comfortable. The categorical explanatory variable of interest is the gender of the respondent; 0 for female, 1 for male. The simulated dataset consists of 400 total observations. Below is a tabulation of the response variable and the explanatory variable, which shows the count of the number of participants that answered with each rating as well as the proportion of these counts conditional on the gender.

Table 1

	Females (0)	Males (1)
Uncomfortable (1)	28 (0.136)	30 (0.155)
Comfortable (2)	63 (0.306)	64 (0.330)
Very Comfortable (3)	115 (0.558)	100 (0.515)

For ordinal logistic regression, software packages do not have to determine a reference level for the dependent variable. However, the software does select an order for the levels of the dependent variable. It is wise to use numerical encoding for ordinal logistic regression since some programs, such as Stata, use alphanumeric ordering, which may not coincide with the intended ordering of the variable.

A cumulative logit parameterization is used in ordinal logistic regression models. However, there are several ways in which this can be done. Table 2 shows the common parameterizations for the cumulative logit model, where *J* represents the number of levels in the categorical response variable, and *p* represents the number of explanatory variables. The most common parameterizations are models 1 and 2 where the outcome of interest is observing a particular value of the response variable or *less.* For model 3, the cumulative logit parameterization specifies that the outcome of interest is observing a particular value of the response variable or *greater.* Regardless of the parameterization, the model will have *J-1* cutoffs (also referred to as intercepts or threshold values), denoted by α_j in the parameterizations below, and one parameter for each explanatory variable. This allows for the intercept to vary for each cumulative logit. However, the model assumes that each explanatory variable exerts the same effect on each cumulative logit. This is why the ordinal logistic regression model is also known as a proportional-odds model.

Table 2

Model	Parameterization
1	$log\left(\frac{P(Y \leq j)}{1-P(Y \leq j)}\right) = S_j = \alpha_j - \left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right) \quad j = 1,2, \dots, J-1$
2	$log\left(\frac{P(Y \leq j)}{1-P(Y \leq j)}\right) = S_j = \alpha_j + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \quad j = 1,2, \dots, J-1$
3	$log\left(\frac{P(Y > j)}{1-P(Y > j)}\right) = S_j = \alpha_j + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \quad j = 2, \dots, J-1, J$

Model 1 incorporates a negative sign so that there is a direct correspondence between the slope and the ranking. Thus a positive coefficient indicates that as the value of the explanatory variable increases, the likelihood of a higher ranking increases. This is also the case for the parameterization of model 3, but notice that the intercepts will differ between model 1 and model 3.

Ordinal logistic regression models can be estimated in most statistical software packages. Some possible implementations include:

- SAS: proc logistic or proc genmod
- R: clm in the "ordinal" package, vglm in the "VGAM" package, polr in the "MASS" package, and lrm in the "rms" package
- Stata: ologit command
- JMP: fit model menu with the response variable classified as ordinal
- SPSS: generalized linear model menu or the ordinal regression menu

Besides knowing the parameterization of the cumulative logit implemented by a software package, a researcher must also be aware of the coding scheme and choice of reference level for categorical explanatory variables. R, Stata, SPSS, and SAS (using proc genmod) use dummy coding, while JMP and SAS (using proc logistic) use effect coding (see StatNews #72 for more information on these two coding schemes). Both R and Stata use the first level alphanumerically as the reference level, whereas SAS, JMP, and SPSS use the last level as the reference level. However, it is possible to customize the reference level in each of these programs.

Table 3 contains the results from an ordinal logistic regression model fit in the various statistical software programs using their default settings with rating as the ordinal response variable and gender as the explanatory variable. The model number refers to the different model parameterizations presented in Table 2.

Table 3

	Stata, R (polr, clm)	R (vglm)	R (lrm)	SPSS	JMP or SAS (proc logistic)	SAS (proc genmod)
Model	1	2	3	1	2	2
Coding	Dummy	Dummy	Dummy	Dummy	Effect	Dummy
Threshold 1	-1.858	-1.858	1.858	-1.690	-1.774	-1.691
Threshold 2	-0.232	-0.232	0.232	-0.064	-0.148	-0.064
Gender=1	-0.168	0.168	-0.168	--	--	--
Gender=0	--	--	--	0.168	-0.084	-0.168

As an example, using the Stata output we can write the functional form of the ordinal regression as follows:

$$\log\left(\frac{\Pr(Y \leq 1)}{1 - \Pr(Y \leq 1)}\right) = -1.858 + 0.168 * Gender$$

$$log\left(\frac{\Pr(Y \leq 2)}{1-\Pr(Y \leq 2)}\right) = -0.232 + 0.168 * Gender$$

One way to interpret the coefficients is via a proportional odds ratio. The model parameterization dictates the interpretation of the odds ratio. Using Stata's estimates, the odds ratio for gender is *exp(*$-\beta_1$*) = exp(0.168) = 1.18*. Thus men are 1.18 times more likely than women to rate the mattress with a lower score.

For R (vglm), the same interpretation holds but the odds ratio is computed by exponentiating the parameter estimate without adding the negative sign, exp (β_1*) = exp(0.168) = 1.18*

However, for SAS proc genmod we would say that women are 0.84 times as likely as men to rate the mattress at a higher score, *exp(*β_1*) = exp(-0.168) = 0.84*. Note this is the same interpretation as above because the frame of reference changes from women to men and *0.84 = 1/1.18*.

Similarly to a logistic regression model, once the model has been estimated, most software packages have an option to compute predicted probabilities, which are estimated using the following formula:

$$P(Y \leq j \mid x) = \frac{e^{S_j}}{1+e^{S_j}}.$$

When the assumption of proportional odds is satisfied, the predicted probabilities from the model will be similar to the observed proportions. Table 4 shows the predicted probabilities from the ordinal logistic regression model as well as the observed proportions (in parentheses-from Table 1) of the ratings based on gender. Note that although the model outputs in Table 3 are different due to the parameterizations used by each software package, they all agree in interpretation and estimate the same predicted probabilities.

Table 4

	Females (0)	Males (1)
Uncomfortable (1)	0.135 (0.136)	0.156 (0.155)
Comfortable (2)	0.307 (0.306)	0.328 (0.330)
Very Comfortable (3)	0.558 (0.558)	0.516 (0.515)

Tests are available to assess the assumption of proportional odds. In Stata, the brant command applied after an ordinal logistic model provides one method for testing the assumption of proportional odds. In R, the nominal_test() function in the ordinal package can be used to test this assumption. SAS includes the test for the proportional odds assumption automatically in the output, as does SPSS's ordinal regression menu. JMP does not offer a test of proportional odds. In the absence of a test, one can fit both an ordinal logistic regression and a multinomial logistic regression to compare the AIC values. If the proportional odds assumption is not met, one can use a multinomial logistic regression model, an adjacent-categories logistic model, or a partial proportional odds model.

If you need assistance with the implementation or interpretation of an ordinal logistic model or have any other statistical consulting questions, please feel free to contact the statistical consultants at CSCU.

References

Agresti, Alan. “Categorical Data Analysis.” New York: Wiley, 2002. Print.

Le, Chap T. “Applied Categorical Data Analysis.” New York: Wiley, 1998. Print.

Author: Stephen Parry