Regression Models for Ordinal Dependent Variables

The Concept of Propensity and Threshold

Binary responses can be conceptualized as a type of propensity for Y to equal 1. For example, we may ask respondents whether or not they use public transportation with a "yes" or "no" response. But individuals may differ on the likelihood that they will use public transportation with some occasionally using it and others using it very regularly. Even those who do not use public transportation at all might have a high favorability toward public transportation and under the right conditions might use it. This tendency or propensity that might underlie a "yes" or "no" response or many other dichotomous variables is sometimes referred to as a "latent" variable. The latent variable, often referred to as Y* in this context, is thought of as an unobserved continuous variable. When the value of this variable reaches a certain point, a "threshold," the observed response is "yes." In the following equation, τ is the threshold value, often considered to be 0 for convenience.

$$observed\ Y = \begin{cases} 0 & when\ Y^* \leq \tau \\ 1 & when\ Y^* > \tau \end{cases}$$

If we imagine a z-distribution of scores, 0 is in the middle. For a binary variable, when the latent variable, Y*, exceeds 0, then we observed a response of 1. When it is less than or equal to the threshold value, then we observe a response of 0. The same conceptualization is used for a special type of correlation called *polychoric correlations* used to estimate the correlation between two latent continuous distributions when then observed variables are discrete (Olsson, 1979).[1] It is not mathematically necessary to assume an underlying latent variable, it is merely a conceptual convenience to do so.

Thus far the discussion has used the example of a binary outcome variable, but an important advantage of this approach is that it can be generalized to a situation in which there are more than two ordered categories, such as response options of "never," "sometimes," and "a lot." Typically variables analyzed as ordinal have 3 or 4 rank-ordered categories that do not necessarily have equal distance between the values. Once there are 5 or more categories and particularly with larger sample sizes and fairly normally distributed variables, there will be little difference between results obtained with ordinal regression and OLS regression approaches.

Ordered Logit

With a binary variable, the logit model is the same as logistic regression. Three or more ordinally ranked categories can be used for the outcome, however. In both SPSS and SAS, ordinal logit analysis can be obtained through several different procedures. SPSS does not provide odds ratios using the ordinal regression procedure, but odds ratios can be obtained

[1] There are several related terms. *Tetrachoric* correlations are special correlations between two binary variables, *polyserial* correlations are between a binary and continuous variable, and *polychoric* correlations are between ordinal variables. All of these can be considered corrections for correlations among discrete variables to values expected if the variable was actually continuous and normally distributed.

by exponentiation of the coefficients.[2] In the case of an ordinal outcome with three or more categories, the odds ratio for the logit model represents the odds of the higher category as compared to all lower categories combined. In other words, it is a cumulative odds ratio representing the increased likelihood to the next highest category relative to the lower categories for each unit increase in the predictor.

Probit Regression

A second approach to regression with ordinal outcomes is *probit* regression. Probit regression assumes that the errors are distributed normally, whereas logit regression assumes the errors are distributed according to a logistic distribution. Both modeling approaches are acceptable, and researchers tend to choose the approach with which they are the most familiar. Some researchers prefer logistic to probit regression because odds ratios can be computed, but some researchers prefer probit to logit because standardized coefficients can be obtained.

With probit analysis, there is no odds ratio, but authors often use the standardized coefficient. The standardized coefficient is interpreted much the same way as it is in OLS except it represents the change in the latent variable, *Y** rather than the observed variable, Y. The unstandardized coefficients are not as easily interpreted as they represent the probit-transformed change in the predicted score for every unit change in the predictor. Probit and logistic regression will usually produce very similar results, especially with large sample sizes. The figure below compares the logistic and normal (used with probit) cumulative distribution functions (CDFs). There are two lines for the logistic curve, one in which the variance has been standardized so that it is comparable to the normal distribution. As can be seen in the figure, the standardized cumulative logistic distribution is nearly identical to the normal distribution.

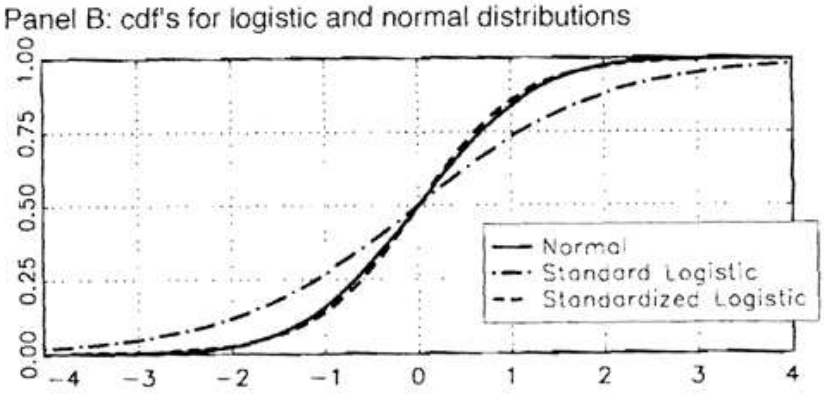

Figure 3.3. Normal and Logistic Distributions

From J. S. Long, 1997, p. 43

[2] Note that with the ordinal regression procedure in SPSS using the logit link function, the threshold is -1 times the constant obtained in the logistic regression, so you will see opposite signed constant values in SPSS compared with SAS.

As with logistic regression, logit and probit models estimate a chi-square (i.e., likelihood ratio) which compares deviances for the full model to the deviance for the baseline or null model (i.e., constant only with no predictors). The significance test of the likelihood ratio indicates whether the predictors in the model together account for significant variance in the dependent variable. Pseudo R^2 values are also printed to estimate the approximate amount of variance accounted for by the predictors.

Loglinear Models

Loglinear models, which can also be used for ordinal variables, are not predictive models. Rather they are like chi-square models in that there is no need to specify an independent and dependent variable. In simple cases, the loglinear model is equivalent to the logit model and more generally to be related to Poisson models (Agresti, 2002). Loglinear models can be used for cases in which there are two or more ordinal categories for the independent or dependent variable. Wickens (1989) provides a gentle introduction to loglinear models and Agresti (2002) is an authoritative source on the topic.

Multinomial Logistic for Multicategory Nominal Outcomes

Not all multicategory outcomes can be ordinally ranked, but a variant on logistic regression can be used to predict such outcomes. For example, if one wanted to predict the type of car purchased, such as compact, sedan, or SUV, the outcome is not easily ordered in anyway. A multinomial (or polytomous) logistic regression model can estimate the odds of choosing one category of car over another (coded as 0). A multinomial logistic estimates separate logistic models comparing each of the other groups to this baseline or comparison group. If there are *g* groups, then there will be *g – 1* logistic models estimated.

References and Further Reading

Agresti, A. (200). *Categorical Data Analysis, 2nd edition*. New York: Wiley.

Aldrich, J.H., & Nelson, F.D. (1984). *Linear probability, logit, and probit models.* Newbury Park, CA: Sage.

Fox, J. (2008). *Applied regression analysis and generalized linear models, second edition.* Los Angeles, CA: Sage.

Long, J.S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44*, 443-460.

Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences.* Hillsdale, NJ: Erlbaum