Regression Models for Ordinal Dependent Variables

Thus far the logistic and probit regression have involved a binary outcome variable, but an important advantage of these models is that they 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 except for heavily skewed distributions (e.g., Kromrey & Rendina-Gobioff, 2002; Taylor, West, & Aiken, 2006).[1]

Ordinal Logistic and Probit Regression

*Ordinal logistic (*or sometimes called *ordered logit models*) are logistic regressions that model the change among the several ordered values of the dependent variable as a function of each unit increase in the predictor. (With a binary variable, the ordinal logistic model is the same as logistic regression.) In SPSS and R, ordinal logistic 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 by exponentiation of the coefficients (e^{B}).[2] In the case of an ordinal outcome with three or more categories, the odds ratio for the logistic 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. It is assumed that the same effect occurs for each level comparison of the ordered responses, so that the increase or decrease in odds for each unit increase in X is the same for the increment from $\ln[P(Y \leq = 1)]$ to $\ln[P(Y \leq = 2)]$ as from $\ln[P(Y \leq = 2)]$ to $\ln[P(Y \leq = 3)]$. In other words, “slopes” for predicting the logit are parallel over all of the ordered categories of the response (sometimes called a "proportional odds model").

The ordinal logistic regression model can be written in two parts as

$$\ln\left(\frac{\hat{p}_{i1} + \hat{p}_{i2}}{\hat{p}_{i0}}\right) = \tau_1 + BX_i$$

$$\ln\left(\frac{\hat{p}_{i2}}{\hat{p}_{i0} + \hat{p}_{i1}}\right) = \tau_2 + BX_i$$

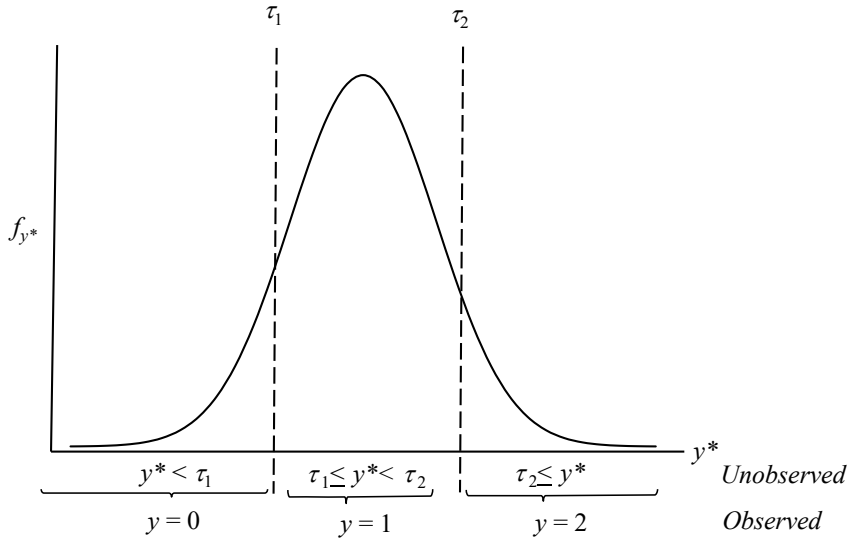
The probabilities $\hat{p}_{i0}$, $\hat{p}_{i1}$ and $\hat{p}_{i2}$ are for the observed values $Y = 0$, $Y = 1$, and $Y = 2$.[3] The slope B represents the average change in the logit for each level increase in the dependent variable, where there is only one slope coefficient. A novel aspect of the ordered logit model, however, is that there are multiple $(J - 1)$ intercepts so that there will always be one fewer intercepts than response categories. Each intercept represent an estimate of the threshold in the generalized linear modeling Y* framework. Above, the thresholds are referred to as τ_1 and τ_2 above and in general we can call τ_j, an estimate of the threshold from the Y* distribution (logistic cdf in logistic case). The Y* value falls between any two intercepts, $\tau_{j-1} \leq Y^* < \tau_j$. If the unknown Y* value is equal to or greater than the threshold value, then Y is observed to be the next higher value (e.g., $Y = 2$ instead of $Y = 1$).

---

[1] See also the "Levels of Measurement and Choosing the Correct Statistical Test" handout for my univariate statistics course for more detail and references.

[2] Note that with the ordinal regression procedure in SPSS and R 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 and R.

[3] My subscripts, 0, 1, and 2 correspond with the subscripts D for "disagree", U for "undecided" and A for "agree" in the Cohen, Cohen, West, and Aiken (2003) example (p. 523). The text also uses t for the sample estimate instead of τ.



Predicted probabilities for each category response on Y can be obtained by using the exponential formula $(1/(1+e^{BX+\tau_j})$, where the appropriate threshold value (e.g., τ_1 for $\hat{p}_1$), a chosen X value, and the slope coefficient are inserted.

A second approach to regression with ordinal outcomes is probit regression, which assumes normally distributed errors (see the "Link Functions and the Generalized Linear Models" handout). Most of the other aspects of the probit model parallel the logistic ordinal model, including multiple thresholds and the assumption of equal slopes across each increment of Y. Because probit models involve a normal distribution for Y^*, the thresholds are standardized score values, with most values occurring between approximately -3 and +3. As with a binary outcome, the logit and probit analysis will nearly always lead to the same conclusions (Long, 1997). 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.

For outcomes that can be considered ordinal, it is generally better to use all of the ordinal values rather than collapsing into fewer categories or dichotomizing variables, even with a sparse number of responses in some categories. Collapsing categories has been shown to reduce statistical power (Ananth & Kleinbaum 1997; Manor, Mathews, & Power, 2000) and increase Type I error rates (Murad, Fleischman, Sadetzki, Geyer, & Freedman, 2003).

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 is more generally related to Poisson models (Agresti, 2013). 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 (2013) is a somewhat more technical source on the topic. See also my handout "Ordinal Analyses" under my Univariate Quantitative Methods class.

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 smart phone purchased, such as Apple, Google, or Samsung, the outcome is not easily ordered in any way. A multinomial (or polytomous) logistic regression model can estimate the odds of choosing one category of phone over

another (e.g., Apple coded as 0). Multinomial logistic models provide multiple sets of coefficients for comparisons of 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. Please see the subsequent handout "Multinomial Logistic Regression Models" for more information.

References and Further Reading

Agresti, A. (2013). *Categorical data analysis (Third Edition).* New York: John Wiley & Sons.
Agresti, A. (2015). *Foundations of linear and generalized linear models*. New York: John Wiley & Sons.
Aldrich, J.H., & Nelson, F.D. (1984). *Linear probability, logit, and probit models.* Newbury Park, CA: Sage.
Ananth, C. V., and Kleinbaum, D. G. (1997), "Regression Models for Ordinal Responses: A Review of Methods and Applications," International Journal of Epidemiology, 26, 1323-1333.
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.
Kromrey, J. D., & Rendina-Gobioff, G. (2002). An empirical comparison of regression analysis strategies with discrete ordinal variables. *Multiple linear regression viewpoints*, *28*(2), 30-43.
Manor, O., Matthews, S., & Power, C. (2000). Dichotomous or Categorical Response? Analysing Self-Rated Health and Lifetime Social Class," *International Journal of Epidemiology, 29*, 149-157
McCullagh, P. (1980). Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, 109-142.
Murad, H., Fleischman, A., Sadetzki, S., Geyer, O., & Freedman, L. S. (2003). Small samples and ordered logistic regression: does it help to collapse categories of outcome?. The American Statistician, 57(3), 155-160.
Taylor, A. B., West, S. G., & Aiken, L. S. (2006). Loss of power in logistic, ordinal logistic, and probit regression when an outcome variable is coarsely categorized. *Educational and psychological measurement*, *66*(2), 228-239.
Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences.* Hillsdale, NJ: Erlbaum