Multichotomous Dependent Variables I

Last time we looked at binary dependent variables. We now move on to multichotomous dependent variables:

1. Ordered Dependent Variables:
    - Presidential Approval - Approve, Indifferent, Disapprove
    - Political Interest - very, somewhat, not much, not at all
    - International Conflict - war, diplomatic conflict, peace.
2. Unordered Dependent Variables
    - Which of three parties did you vote for?
    - Did you travel by car, plane or train?
    - Which potential government coalition formed?
3. Count Dependent Variables
    - Number of Wars
    - Number of Government Formation Attempts
    - Number of Court Cases Heard

We begin by looking at ordered dependent variables.

Ordered Dependent Variables

When we impose ordering, we are making a strong assumption - that 2 has a bigger effect than 1. This may or may not make sense - you should always think about this before using an ordered model. Some things can be ordered (spectrum of color) but probably shouldn't be in most applications. Moreover, you need to think about what the appropriate order is if you are going to use an ordered variable. For example, while the categories strongly agree, agree, neutral, disagree, strongly disagree are ordered, someone interested in the intensity of opinion should use the following ordering: strongly agree or strongly disagree, agree or disagree, neutral. The bottom line is that you should be careful when thinking about whether your dependent variable is ordered and how it is ordered. Ordered models obviously use 'ordinal' data - we can say whether some observation is more or less 'something' than some other observation, but we cannot say how much more or how much less 'something' it is. Note that if your dependent variable is truly ordered, then you should never try to dichotomize it because this is throwing away useful information.

OLS

You might think of treating the ordered variable as an interval variable and using OLS. However, this runs into problems: (i) the errors will be heteroskedastic and (ii) the OLS model will only correspond (roughly) to the correct ordered model if the thresholds are all about the same distance apart - when this is not the case, the OLS model will give very misleading results (Long 1997, 118).

As an introduction to the ordered models, let's start by considering the following example.[1]

1 Example

Consider the responses to the following question.

> Would you say that over the past year the nation's economy has gotten better, stayed about the same, or gotten worse? (Would you say much better or somewhat better?) (Would you say much worse or somewhat worse?)

Table 1: An Example of an Ordered Variable

Responses	Category	Category Name
18%	1	Much Worse
12%	2	Somewhat Worse
15%	3	Stayed the Same
20%	4	Somewhat Better
35%	5	Much Better

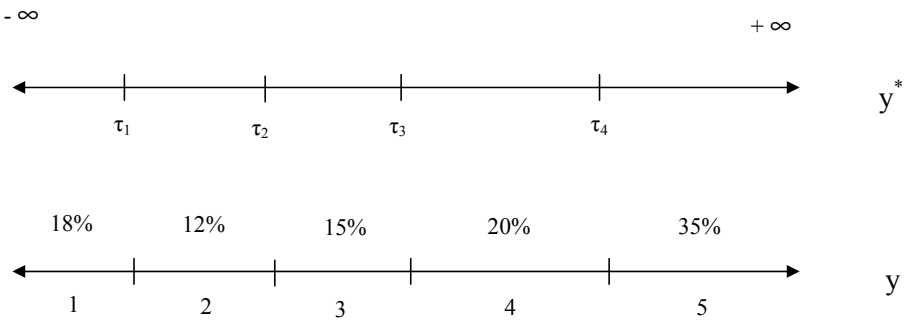
Just as we did with logit and probit, we can think in terms of an underlying latent variable indicating a respondent's view of the economy. Figure 1 indicates how ordinal responses can be placed along the real number line where each category of y is defined as the region between one threshold and the next.

1.1 Standard Normal

We now need to figure out how to represent these outcomes in terms of some familiar distribution. We might want to pick the normal distribution. At this point, all we want to do is model the marginal proportions of y as the probabilities associated with draws from a normal distribution. Put differently, we need to divide the normal distribution into regions which define the probabilities 0.18, 0.12, 0.14, 0.20, and 0.35. This is relatively straightforward. Start with a standard normal density and then find the thresholds that will divide this density into regions with areas of 0.18, 0.12, 0.14,

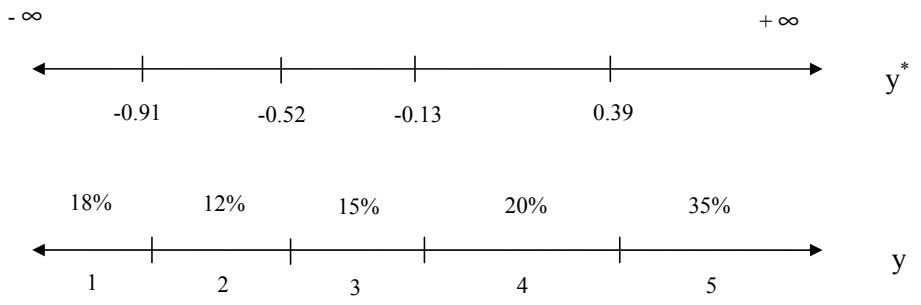
[1] This example is drawn from the notes of Charles Franklin.

Figure 1: Mapping Latent Variable into Ordered Observed Variable



0.20, and 0.35. In other words, we need to solve for $\Phi^{-1}(0.18)$ - this is τ_1 and equals -0.9154.[2] The next threshold is $\Phi^{-1}(0.18+0.12) = -0.5244 = \tau_2$. Note that $\Phi(\tau_2) - \Phi(\tau_1) = 0.12$ - the proportion of cases in the second response category. Continuing, we have $\Phi^{-1}(0.18+0.12+0.15) = -0.1257 = \tau_3$ and $\Phi^{-1}(0.18 + 0.12 + 0.15 + 0.20) = 0.3853 = \tau_4$. Thus, we have Figure 2.

Figure 2: Mapping Latent Variable into Ordered Observed Variable



1.2 Transforming the Space - σ

Can we simply assume a standard normal? In the standard normal density, we have assumed that $\mu = 0$ and $\sigma = 1$. But what happens if we have $\sigma = 2$? Could you still find the corresponding thresholds? Yes. Recall that the following transformation $z = \frac{x-\mu}{\sigma}$ converts any normal into a standard normal. Thus, to get back to the original scale, we would just undo the transformation: $x = z\sigma + \mu$. Therefore, if we knew that $\sigma = 2$, $\mu = 0$, we can get the thresholds in the correct units

[2] In STATA, type 'display invnormal(0.18)'. In other words, we have $\Phi(y^*) = 0.18$ and $y^* = \Phi^{-1}(0.18)$ where $y^* = -0.9154$.

just by applying the transformation, $\tau_j^* = \tau_j\sigma + \mu = 2\tau_j$. As you can see, for any σ that you can think of, you can solve for the probabilities (the regions representing the marginals of y) using the standard normal and then transform them back into your original scale so long as you know what μ and σ are. More importantly, this means that you don't actually even need to know μ and σ – you can still represent the probabilities perfectly well with the standard normal and never need to know the true scale. The point to take away is that the model represents the probabilities equally well no matter what the true scale of the thresholds underlying the measurement. The probabilities are invariant with respect to any linear transformation of the underlying scale. As a result, we don't actually need to know or estimate σ or μ in order to get the probabilities.

1.3 What happens when we let μ vary?

In the example that we have used so far, we have allowed the mean of the distribution to be a constant, μ. Now we can let this vary, so that μ_i becomes a mean for each observation, which differs across different observations. How does this affect the probabilities of each outcome?

Suppose we have two observations - $\mu_1 = 0$ and $\mu_2 = 1$ and let $\sigma = 1$ for both observations. We already know the probabilities implied by μ_1 and the thresholds since we just calculated them. What probabilities result if μ_2 is used but with the original thresholds? For this second distribution, what is the probability that $y = 1$? As before, we have

$$\begin{aligned} p(y=1) &= \Phi(\tau_1 - \mu_2) \\ &= \Phi(-0.9154 - 1) \\ &= \Phi(-1.9154) \\ &= 0.0277 \end{aligned}$$

$$\begin{aligned} p(y=2) &= \Phi(\tau_2 - \mu_2) - \Phi(\tau_1 - \mu_2) \\ &= \Phi(-0.5244 - 1) - \Phi(-0.9154 - 1) \\ &= 0.0637 - 0.0277 \\ &= 0.0360 \end{aligned}$$

$$\begin{aligned} p(y=3) &= \Phi(\tau_3 - \mu_2) - \Phi(\tau_2 - \mu_2) \\ &= \Phi(-0.1257 - 1) - \Phi(-0.5244 - 1) \\ &= 0.1301 - 0.0637 \\ &= 0.0664 \end{aligned}$$

$$\begin{aligned} p(y=4) &= \Phi(\tau_4 - \mu_2) - \Phi(\tau_3 - \mu_2) \\ &= \Phi(0.3853 - 1) - \Phi(-0.1257 - 1) \\ &= 0.2694 - 0.1301 \\ &= 0.1393 \end{aligned}$$

$$\begin{aligned} p(y=5) &= 1-\Phi(\tau_4-\mu_2) \\ &= 1-\Phi(0.3853-1) \\ &= 1-0.2694 \\ &= 0.7306 \end{aligned}$$

If the thresholds are held fixed, then shifting μ results in different response probabilities. These are shown in Table 2 below.

Table 2: Letting μ_i Vary

y	$\mu=0$	$\mu=0.125$	$\mu=0.25$	$\mu=0.5$	$\mu=1$
1	0.18	0.1491	0.1219	0.0785	0.0277
2	0.12	0.1089	0.0974	0.0743	0.0360
3	0.15	0.1430	0.1343	0.1133	0.0064
4	0.20	0.2017	0.2002	0.1885	0.1393
5	0.35	0.3973	0.4462	0.5457	0.7306

As you can see, the parameter μ_i shifts the distribution around while the thresholds remain constant. As we have done with previous models, we reparameterize μ_i in terms of observed exogenous variables that capture individual differences across observations so that we have $\mu_i = x_i\beta$. In this model, we continue to assume $\sigma = 1$.

2 General Model

Now let's move beyond our example to a general model. Consider the formula for $y = 1$. We observe $y = 1$ when y^* falls between $\tau_0 = -\infty$ and τ_1. This implies that

$$P(y_i = 1|x_i) = P(\tau_0 \leq y_i^* < \tau_1|x_i) \tag{1}$$

If we substitute $y^* = x\beta + \epsilon$, we have

$$P(y_i = 1|x_i) = P(\tau_0 \leq x_i\beta + \epsilon_i < \tau_1|x_i) \tag{2}$$

If we then subtract $x\beta$ within the inequality, we have

$$P(y_i = 1|x_i) = P(\tau_0 - x_i\beta \leq \epsilon_i < \tau_1 - x_i\beta|x_i) \tag{3}$$

The probability that a random variable is between two values is the different between the cdf evaluated at these values. Thus, we have

$$\begin{aligned} P(y_i = 1|x_i) &= P(\epsilon_i < \tau_1 - x_i\beta|x_i) - P(\epsilon_i \leq \tau_0 - x_i\beta|x_i) \\ &= F(\tau_1 - x_i\beta) - F(\tau_0 - x_i\beta) \end{aligned} \tag{4}$$

These steps can be generalized to compute the probability of any observed outcome $y = j$ given x.

$$P(y_i = j|x_i) = F(\tau_j - x_i\beta) - F(\tau_{j-1} - x_i\beta) \tag{5}$$

Note that when computing $P(y = 1|x_i)$, the second term on the right-hand side drops out since $F(\tau_0 - x_i\beta) = F(-\infty - x_i\beta) = 0$. Note also that when computing $P(y = J|x_i)$, the first term equals 1 since $F(\tau_J - x_i\beta) = F(\infty - x_i\beta) = 1$. As a result, our generalized model for J categories with an ordered probit model can be written as:

$$\begin{aligned}
P(y_i &= 1|x_i) = \Phi(\tau_1 - x_i\beta) \\
P(y_i &= 2|x_i) = \Phi(\tau_2 - x_i\beta) - \Phi(\tau_1 - x_i\beta) \\
&\vdots \\
P(y_i &= j|x_i) = \Phi(\tau_j - x_i\beta) - \Phi(\tau_{j-1} - x_i\beta) \\
&\vdots \\
P(y_i &= J|x_i) = 1 - \Phi(\tau_{J-1} - x_i\beta)
\end{aligned} \tag{6}$$

where the βs and τs are to be estimated. The probability for this last category is sometimes equivalently expressed as $1 - \Phi(\tau_{J-1} - x_i\beta) = \Phi(-(\tau_{J-1} - x_i\beta))$ thanks to the symmetry of the normal distribution and then as $\Phi(x_i\beta - \tau_{J-1})$.[3] Note that we have written this using the standard normal distribution so that we have an ordered probit model. However, there is no reason why we couldn't use a standard logistic function and have an ordered logit model – we would just switch the Φs for Λs.

Again, it is easy to see that if the thresholds are fixed, increasing $x\beta$ results in different response probabilities (look back at Table 2). As you can see, increasing our Xs shifts the distribution around while the thresholds remain constant.

2.1 The Likelihood and Log-Likelihood Function

Under the assumption of independent observations, the sample likelihood is just the product of these probabilities. Note that the probability of an observation depends on which category of y it falls into. So, if a case in in $y = 2$, it has a probability of $\Phi(\tau_2 - x_i\beta) - \Phi(\tau_1 - x_i\beta)$ and if it falls in

[3] A further extension might be to define two thresholds, $\tau_0 = \infty$ and $\tau_J = +\infty$, so that we can write the first and last probabilities as

$$\begin{aligned}
P(y_i = 1|x_i) &= \Phi(\tau_1 - x_i\beta) - \Phi(\tau_0 - x_i\beta) \\
&= \Phi(\tau_1 - x_i\beta) - 0 \\
&= \Phi(\tau_1 - x_i\beta)
\end{aligned} \tag{7}$$

and

$$\begin{aligned}
P(y_i = J|x_i) &= \Phi(\tau_J - x_i\beta) - \Phi(\tau_{J-1} - x_i\beta) \\
&= 1 - \Phi(\tau_{J-1} - x_i\beta)
\end{aligned} \tag{8}$$

Some textbooks write it like this.

$y = 1$, it has a probability of $\Phi(\tau_1 - x_i\beta)$ etc. To write the joint likelihood function of the sample, we have to write it so that it associates the correct probability for each observation. An easy way to do this is to define J dummy variables for each of the J categories so that $d_{ij} = 1$ if observation i falls in category j and $d_{ij} = 0$ otherwise. The likelihood is then the product over all j and all i of the probabilities raised to d_{ij} (which just picks out the correct probability associated with each observation).

$$\begin{aligned} \mathrm{L} &= \prod_{j=1}^{J}\prod_{i=1}^{N} P(y=j)^{d_{ij}} \\ &= \prod_{i=1}^{N} P(y=1)^{d_{i1}} \times \prod_{i=1}^{N} P(y=1)^{d_{i2}} \times \ldots \times \prod_{i=1}^{N} P(y=J)^{d_{iJ}} \end{aligned} \tag{9}$$

Thus, the log-likelihood is

$$\begin{aligned} \ln\mathrm{L} &= \sum_{j=1}^{J}\sum_{i=1}^{N} \ln[P(y=j)^{d_{ij}}] \\ &= \sum_{j=1}^{J}\sum_{i=1}^{N} d_{ij}\ln[\Phi(\tau_J - x_i\beta) - \Phi(\tau_{j-1} - x_i\beta)] \\ &= \sum_{i=1}^{N} d_{i1}\ln[\Phi(\tau_1 - x_i\beta)] \\ &+ d_{i2}\ln[\Phi(\tau_2 - x_i\beta) - \Phi(\tau_1 - x_i\beta)] + \ldots \\ &+ d_{ij}\ln[\Phi(\tau_j - x_i\beta) - \Phi(\tau_{j-1} - x_i\beta)] + \ldots \\ &+ d_{iJ-1}\ln[1 - \Phi(\tau_{J-1} - x_i\beta)] \end{aligned} \tag{10}$$

This is often written as

$$\begin{aligned} \ln\mathrm{L} &= \sum_{y=1} \ln[\Phi(\tau_1 - x_i\beta)] \\ &+ \sum_{y=2} \ln[\Phi(\tau_2 - x_i\beta) - \Phi(\tau_1 - x_i\beta)] \\ &\vdots \\ &+ \sum_{y=j} \ln[\Phi(\tau_j - x_i\beta) - \Phi(\tau_{j-1} - x_i\beta)] \\ &\vdots \\ &+ \sum_{y=J} \ln[1 - \Phi(\tau_{J-1} - x_i\beta)] \end{aligned} \tag{11}$$

2.2 Identification Issues

For estimation, we have k parameters in β and $J-1$ thresholds in τ. By setting $\text{var}(\epsilon|x) = 1$ (ordered probit), we have already avoided having to estimate this parameter - as we saw earlier, this assumption does not affect the probabilities.[4] However, there is one more identification issue to deal with. How do we pin down the location of τ if the underlying scale is unknown? The problem is that the probabilities are unaffected if we shift the entire distribution and thresholds left or right by any constant amount (think back to our earlier example and diagrams). To see this, rewrite the expression for the first category (the same holds for the other categories)

$$P(y=1) = \Phi(\tau_1 - x_i\beta) = \Phi(\tau_1 - \beta_0 - x_{i*}\beta_*) \tag{12}$$

where β_0 is the intercept term and $x_{i*}\beta_*$ is the $x_i\beta$ but with the unit vector intercept removed. You'll see that if we add a constant to both the threshold and the intercept, we get exactly the same equation

$$\begin{aligned} P(y=1) &= \Phi(\tau_1 + c - (\beta_0 + c) - x_{i*}\beta_*) \\ &= \Phi(\tau_1 - \beta_0 + c - c - x_{i*}\beta_*) \\ &= \Phi(\tau_1 - \beta_0 - x_{i*}\beta_*) \end{aligned} \tag{13}$$

Substantively all the addition of a constant does is shift the scale left or right by c units. The result, though, shows that we can never tell what c is since the results are exactly the same i.e. we cannot know if we have estimated τ_1 or $\tau_1 + c$ and β_0 or $\beta_0 + c$. This is a classic case of under-identification. The relative distances between τ_j and $x_i\beta$ will be unchanged and it is these relative distances which determine the probabilities. Hence to identify either τ or β_0, we must constrain one of them to some fixed value. This is equivalent to pinning down the distribution at some reference point and then estimating the model relative to that reference point. The most common identifying restriction is to either set $\tau_1 = 0$ or $\beta_0 = 0$. We can identify the model with either restriction - we estimate β_0 with the first restriction or we estimate τ_1 with the second restriction. Since the estimated probabilities will be identical, the choice between these identifying restrictions is purely a matter of convenience and convention - the estimates of the slopes in β are unaffected by the restriction chosen. Different software packages choose different restrictions. For example, STATA assumes that $\beta_0 = 0$ whereas I think LIMDEP assumes that $\tau_1 = 0$. So don't be concerned that although you use a constant in your model of y^*, STATA does not report it - it essentially becomes the first cutpoint (τ_1).

[4] If we were using ordered logit, we would assume that $\text{var}(\epsilon|x) = \frac{\pi^2}{3}$.

2.3 Assumption of Parallel Regression

There is a hidden assumption in the general model that we have presented. Imagine that we were computing cumulative probabilities from our ordered probit model. We would have:

$$\begin{aligned} P(y_i \leq 1|x_i) &= \Phi(\tau_1 - x_i\beta) \\ P(y_i \leq 2|x_i) &= \Phi(\tau_2 - x_i\beta) \\ &\vdots \\ P(y_i \leq j|x_i) &= \Phi(\tau_j - x_i\beta) \\ &\vdots \\ P(y_i \leq J|x_i) &= 1 \end{aligned} \tag{14}$$

What this tells us is that as an independent variable x increases, the cumulative distribution function shifts to the right or left *but that there is no shift in the slope of the distribution* i.e. the βs are the same for each category. In other words, there is a parallel shift of the probability curve and this parallel shift is due to an assumption that the βs are equal for each equation in Eq. 14 i.e. we are assuming that we have the same β for all j categories. It is possible to test this assumption by estimating $J-1$ binary regressions (the first one would code an observation 1 if $y \leq 1$, 0 otherwise etc.) and test the hypothesis that $\beta_1 = \beta_2 = \ldots = \beta_{j-1} = \beta$.[5] If the parallel regression assumption does not hold, you could look at the Generalized Ordered Logit Model that allows the βs to differ across the categories of y (Long 2001). For more information on the assumption of parallel regression see Long (1997, 140-145).

2.4 Continuation Ratio Model

Some ordinal data have the specific characteristic that the categories represent the progression of events or stages through which an individual can advance. For example, the outcome could be faculty rank, where the stages are assistant professor, associate professor and full professor. Or the outcome might be conflict, where the stages are dispute, low-level conflict, and high-level conflict. The key characteristic is that an individual must pass through each stage. The appropriate ordinal model for this type of situation is called the Continuation Ratio Model. This model can be fitted in STATA with the 'ocratio' command. You will probably have to install this by typing: 'net search ocratio' and follow the prompts.

[5] STATA does not have an automatic command for this test. However, if in STATA you type 'net search omodel' and download omodel, you can test the assumption of parallel regression by tying in my case: omodel probit y x's.

3 Interpretation

3.1 Coefficients and Threshold Parameters

3.1.1 Coefficients

As with logit and probit, the coefficients do not indicate the marginal effect of the independent variables on the probabilities of $y = 0, 1, 2, 3$ etc. However, recall that with probit and logit you could infer the direction and statistical significance associated with increasing x on the probability of y from the coefficients. Is this true of the coefficients from ordered probit or logit? What can you infer from the ordered probit coefficients?

Recall that increasing one of the independent variables while holding the βs and τs constant is equivalent to shifting the distribution slightly to the right (see Greene (2003, 738)). The effect of the shift is unambiguously to shift some mass out of the leftmost cell (y=0). Thus, if the β for the independent variable is positive, P(y=0|x) will decline. We can also infer from an increase in some independent variable that there will be more mass in the rightmost cell i.e $y = J$. Of course, if the coefficient on our independent variable is negative we can say exactly the opposite of this. **NOTE** though that we cannot infer from the coefficients what the effect of a change in the independent variable is on the middle categories i.e $1 < y < J$. This is because more mass is moving into these categories from the left but some is leaving to the right - we don't know whether more mass is being added or subtracted just by looking at the coefficients. Thus, analysts who infer that a positive coefficient on some x_k means that an increase in x_k increases the probability of some intermediary category is possibly making a mistake – the signs of the coefficients can only tell us about how an independent variable affects the probability of the end categories.

3.1.2 Threshold Parameters

As we saw earlier, the τs provide the cutpoints on y^* that determine the observed value of y. Remember that one can make the assumption that β_0 or τ_1=0. If one makes the latter assumption, then all of the τs must be positive. However, this will not necessarily be the case with the first parameterization. The threshold parameters should not be treated as nuisances but should be interpreted as an interesting part of the model.

1. **Are the threshold parameters significantly different?**

   Recall that we have assumed that our ordered categories are truly different. You can test whether this is the case simply by determining whether the threshold parameters can be distinguished from one another. If they cannot, then the implication is that observed categories are not really different. At this point, one might want to collapse those categories that are indistinguishable. A quick way to examine whether the τs are different is to look to see if the confidence intervals overlap. If they do not overlap, then the τs are different. If they do

overlap, one will want to test to see if the τs are actually indistinguishable. To do this one would calculate the difference between the τs, the standard error of this difference, and do a z-test.[6]

$$z = \frac{\hat{\tau}_1 - \hat{\tau}_2}{\sqrt{var(\hat{\tau}_1) + var(\hat{\tau}_2) - 2cov(\hat{\tau}_1, \hat{\tau}_2)}} \tag{15}$$

In this particular test, we will not be able to tell the difference between $y = 1$ and $y = 2$ if τ_1 and τ_2 are indistinguishable. The general test of the null hypothesis that $\tau_j - \tau_{j-1} = 0$ is

$$z = \frac{\hat{\tau}_j - \hat{\tau}_{j-1}}{\sqrt{var(\hat{\tau}_j) + var(\hat{\tau}_{j-1}) - 2cov(\hat{\tau}_j, \hat{\tau}_{j-1})}} \tag{16}$$

You should always test all of your adjacent thresholds and report whether they are statistically different when you use ordered logit or probit. The automatic command in STATA is

- test _b[_cut1]=_b[_cut2] etc.

2. **Are the thresholds equally spread out?**

   Depending on your question, it might be interesting to evaluate whether your threshold parameters are equally spread out and if so (or if not) what this means substantively. For example, substantially unequal widths might suggest that the political meanings of some categories are more expansive than others - which might be important. Alternatively, unequal widths could indicate that the semantic distinctiveness of adjectives used to label the response categories are unequal, implying different ranges of inter-subjective meaning. Note also that narrow intervals must produce greater probability of change out of a category even when $x_i\beta$ remains fixed i.e. it will appear that certain responses are more apt to change than others - however, this is only due to the narrow intervals between the thresholds and shouldn't necessarily be interpreted as evidence of an unstable response.

3.2 Going Beyond the Table of Results

In terms of inferences, one should probably go beyond the table of results i.e. beyond simply interpreting the coefficients and threshold parameters.

3.2.1 Odds Ratios

If you use ordered logit, you can use an odds-ratio interpretation of the coefficients. For that model, the change in the odds of Y being equal to or less than j (versus greater than j) associated with a δ-unit change in X_k is equal to $e^{\delta\hat{\beta}_k}$. This means that a change in odds associated with a one-unit

[6]The results from an ordered probit model do report the τs with standard errors and z-scores. However, these results do not tell us whether the thresholds are statistically different from each other - they only tell us whether the thresholds are statistically different from zero, something that is almost never of substantive interest.

change in X_k is just $e^{\hat{\beta}_k}$. So, if we had a coefficient that was, say, -2.39 on our Black variable, then the effect of a one unit increase in this variable would be $e^{-2.39} = 0.09$. In other words, the odds of being in some category rather than some other category are 91% lower if Black= 1 than if Black= 0.

3.2.2 Predicted Probabilities

You could present the predicted probabilities of $y = 0, 1, 2, 3$ etc. for interesting values of the independent variables. Since there are J categories of y, there will be J predicted probabilities - one for each outcome category.

$$\begin{aligned}
P(y_i &= 1|x_i) = \Phi(\hat{\tau}_1 - x_i\hat{\beta}) \\
P(y_i &= 2|x_i) = \Phi(\hat{\tau}_2 - x_i\hat{\beta}) - \Phi(\hat{\tau}_1 - x_i\hat{\beta}) \\
&\vdots \\
P(y_i &= j|x_i) = \Phi(\hat{\tau}_j - x_i\hat{\beta}) - \Phi(\hat{\tau}_{j-1} - x_i\hat{\beta}) \\
&\vdots \\
P(y_i &= J|x_i) = 1 - \Phi(\hat{\tau}_{J-1} - x_i\hat{\beta})
\end{aligned} \tag{17}$$

You could present the results in a table like Table 3.

3.2.3 First Differences

You could also calculate first differences i.e. calculate the predicted probability of y= 0, 1, 2, 3 etc. for two different scenarios of the independent variables. You would again use the equations shown above and calculate the change in probability. You would then get the confidence intervals via the same simulation methods we have used in the past. You could present the results in a table like Table 3 along with the predicted probabilities.

3.2.4 Marginal Effects

You can also calculate marginal effects. Since there are J outcomes, there are J marginal effects i.e. the marginal effect of an instantaneous change in x_k on the probability that $y = 1, 2, 3$ etc.

These are:

$$\begin{aligned}\frac{\partial P(y_i = 1|x_i)}{\partial x} &= \frac{\partial \Phi(\tau_1 - x_i\beta)}{\partial x} \\ &= \phi(\tau_1 - x_i\beta)(-\beta) \qquad (18)\end{aligned}$$

$$\begin{aligned}\frac{\partial P(y_i = j|x_i)}{\partial x} &= \frac{\partial[\Phi(\tau_j - x_i\beta) - \Phi(\tau_{j-1} - x_i\beta)]}{\partial x} \\ &= \phi(\tau_j - x_i\beta)(-\beta) - \phi(\tau_{j-1} - x_i\beta)(-\beta) \\ &= [\phi(\tau_j - x_i\beta) - \phi(\tau_{j-1} - x_i\beta)](-\beta) \qquad (19)\end{aligned}$$

$$\begin{aligned}\frac{\partial P(y_i = J|x_i)}{\partial x} &= \frac{\partial[1 - \Phi(\tau_{J-1} - x_i\beta)]}{\partial x} \\ &= \phi(\tau_{J-1} - x_i\beta)(+\beta) \qquad (20)\end{aligned}$$

The first of the marginal effects is always opposite the sign of β_k i.e. a positive β_k indicates that an increase in x_k reduces the probability of the lowest category whereas a negative β_k indicates that an increase in x_k increases the probability of the lowest category. Similarly, the last marginal effect is always the same sign as β_k i.e a positive β_k indicates that an increase in x_k increases the probability of the highest category whereas a negative β_k indicates that an increase in x_k reduces the probability of the lowest category. Thus, if β_k is positive, then an increase in x_k *has* to reduce $P(y = 1)$ and *has* to increase $P(y = J)$ - we noted this earlier. Recall, though, that the effects in the interior categories is *not* monotonic - the effect of a change in x_k will depend on the relative sizes of $\phi(\tau_j - x_i\beta)$ and $\phi(\tau_{j-1} - x_i\beta)$ for these interior categories.

Table 3: Predicted Probabilities, Marginal Effects, First Differences

	Strongly Disapprove	Disapprove	Approve	Strongly Approve
Predicted Probability	0.22	0.23	0.35	0.20
(Scenario 1)	[0.07, 0.41]	[0.15, 0.27]	[0.25, 0.38]	[0.07, 0.41]
Predicted Probability	0.42	0.25	0.25	0.08
(Scenario 2)	[0.18, 0.66]	[0.19, 0.27]	[0.12, 0.37]	[0.02, 0.22]
Difference	0.20	0.02	-0.10	-0.12
	[0.08, 0.31]	[-0.06, 0.09]	[-0.17, -0.002]	[-0.21, -0.04]
Marginal Effect (Black)	0.16	0.05	-0.07	-0.15
(Scenario 1)	[0.09, 0.22]	[-0.02, 0.09]	[-0.13, 0.04]	[0.08, 0.22]

Scenario 1: Respondent is white; Scenario 2: Respondent is black
Education=13 years, Income=$40,000, Union=1, EconomyWorse=3
(95% Confidence Interval in parentheses)

While marginal effects may be useful in some circumstances and may be the quantity of interest, this

is not always the case and first differences and predicted probabilities are probably more useful. In Table 3 I show how you might want to show predicted probabilities, first differences, and marginal effects. You could also use graphs if this is more helpful.

In Table 3, I present the marginal effect of *Black* in scenario 1. I present this particular marginal effect to highlight interpretational issues with this quantity of interest. First, note that the marginal effect of *Black* is not the same as the first difference between scenario 1 (white) and scenario 2 (black) as one might have thought. The reason for this is quite straightforward. Because the ordered probit and logit models are non-linear, it is not possible to interpret the partial derivative (marginal effect) as the change in probability associated with a one unit change in some independent variable. If you think about it, the marginal effect is capturing the slope of the distribution at scenario 1 i.e. at a given point - the slope at this particular point is linear since we are looking for a very, very small change in the independent variable *Black*. However, our distribution function is not linear and when we calculate the first difference (changing Black from 0 to 1) we are making a relatively large change in the independent variable *Black* and the slope between these two points will be different. Second, this particular example highlights that we probably should not report the marginal effect for dummy variables. This is because a very, very small change in a dummy variable is substantively meaningless - the dummy variable is either 0 or 1 and so we should use first differences for this type of independent variable. Marginal effects only really make sense in the context of continuous variables, and even then they are not always entirely useful.[7]

4 Extensions

There are a variety of ways to extend these ordered models.

4.1 Heteroskedastic Ordered Probit

Alvarez and Brehm (1998) extend the heteroskedastic probit model to the case of multiple ordered outcomes. See their article. The log-likelihood is just:

$$\text{lnL} = \sum_{j=1}^{J}\sum_{i=1}^{N} d_{ij}\ln\left[\Phi\left(\frac{\tau_J - X_i\beta}{e^{Z_i\gamma}}\right) - \Phi\left(\frac{\tau_{j-1} - X_i\beta}{e^{Z_i\gamma}}\right)\right] \tag{21}$$

where the d_{ij}s are indicator variables as before and $Z_i\gamma$ are the variables believed to influence the variability of the latent variable, Y^*.

[7] While you should have calculated marginal effects by hand if you reported them here, the STATA code is:

- mfx compute, at(·) predict(outcome(1))

This will give the marginal effects for each independent variable on the probability of $y = 1$. Note that STATA recognizes the problem of reporting marginal effects for dummy variables and actually reports a first difference i.e. switching from Black=0 to Black=1.

4.2 Variable 'Cut-Points'

Sanders (2000) extend the ordered models even further by allowing for both heteroskedasticity and for cut-points to vary across individuals. We have a set of variables $X_i\beta$ that influence the mean of Y^*, another set $Z_i\gamma$ that affect its variance, and another set $W_i\eta$ that influence where the cut points are. The log-likelihood is just:

$$\ln L = \sum_{j=1}^{J}\sum_{i=1}^{N} d_{ij} \ln \left[\Phi\left(\frac{W_i\eta - X_i\beta}{e^{Z_i\gamma}}\right) - \Phi\left(\frac{W_i\eta - X_i\beta}{e^{Z_i\gamma}}\right)\right] \tag{22}$$

References

Alvarez, R. Michael & John Brehm. 1998. "Speaking in Two Voices: American Equivocation about the Internal Revenue Service." *American Journal of Political Science* 42:418–452.

Greene, William. 2003. *Econometric Analysis*. New Jersey: Prentice Hall.

Long, J. Scott. 1997. *Regression Models for Categorical and Limited Dependent Variables*. London: Sage Publications.

Long, J. Scott. 2001. *Regression Models for Categorical and Limited Dependent Variables Using STATA*. Texas: STATA Corporation.

Sanders, Mitchell S. 2000. "Uncertainty and Turnout." *Political Analysis* 9:45–57.