POLI 8501
Models for Ordinal Responses I

It is often the case that we want to model variables that take the form of a small number of discrete, ordered categories. WE can think of two types ...

- *Grouped continuous* data
  - Data originally measured at the interval/ratio level, then "grouped" into ordered categories.
  - such as measurements of age ("$18-24$", "$25-34$", *etc*), income, etc. which have been "clumped" into categories.
- *Assessed ordered* data
  - Likert items
  - Agree/disagree
  - Other ordinal categories

Before we start, consider a few things about ordinal dependent variables ...

1. Ordinality is often something the researcher has to decide for herself.
2. Some things *can* be ordered, but *shouldn't* be (e.g., color preferences).
3. Some things should be ordered in some circumstances, but not in others (e.g., party identification - typically ordered vis-á-vis ideology, but not vis-á-vis something else (say, region)).
4. Some things will have one ordering for one application, and another for another ... Consider the choice of whether to vote for Nader, Gore, Bush, or Buchanan ...
   - They can be ordered ideologically vis-á-vis say, the environment, BUT
   - They also have cross-cutting issues (e.g., international trade).

Issues with Analyzing Ordinal Responses

Clearly, ordinal response data are *discrete* – they take on a finite number of specific values, corresponding to the ordered categories. At the same time, a high number of such categories might make it possible to consider the data as effectively interval-level.

As an example, consider a *feeling thermometer*, of the sort currently used on the American National Election Studies (ANES):

> "I'd like to get your feelings toward some of our political leaders and other people who are in the news these days. I'll read the name of a person and I'd like you to rate that person using something we call the feeling thermometer. Ratings between 50 and 100 degrees mean that you feel favorably and warm toward the person; ratings between 0 and 50 degrees mean that you don't feel favorably toward the person and that you don't care too much for that person. You would rate the person at the 50 degree mark if you don't feel particularly warm or cold toward the person."

The result is a scale that runs from zero to 100. Note a few things about it:

- On one hand, the scale is clearly ordinal. We know, for example, that a rating of 60 is "higher" than one of 50; at the same time, whether the difference between 50 and 60 is the same as that between 0 and 10 can't be known for sure. (It probably is, but may not be).
- On the other hand, the fact that the scale has (potentially) 101 different possible values means that, for all practical purposes, we can treat it as continuous.

When and Why You Don't Want to Apply OLS To Ordinal Responses

One rule of thumb, then, is that as the number of ordinal categories in our response variable increases, the more justifiable the use of continuous-variable models like OLS (and, all else equal, the more valid the results you'll get using them). When the number of ordinal categories is relatively small, however, OLS and other continuous-response models don't work as well.

Beyond the number of categories, there is also the matter of how the categories are (for lack of a better word) "distributed." More specifically, OLS will *only* give relatively "good" results if the "cut points" for the ordinal categories are about the same distance apart (which they're often not...). Consider an example, where I generated 1000 "fake" observations on Y^*, according to:

$$Y_i^* = 6.0 + 1.0X_i + u_i,$$

where both X_i and $u_i \sim N(0,1)$; Y^* thus has a mean of six and a variance of about two. I then created two ordered, categorical variables Y_1 and Y_2, where:

$$\begin{aligned} Y_{1i} &= 1 \text{ if } Y_i^* < 4 \\ &= 2 \text{ if } 4 \leq Y_i^* < 6 \\ &= 3 \text{ if } 6 \leq Y_i^* < 8 \\ &= 4 \text{ if } Y_i^* > 8 \end{aligned}$$

and

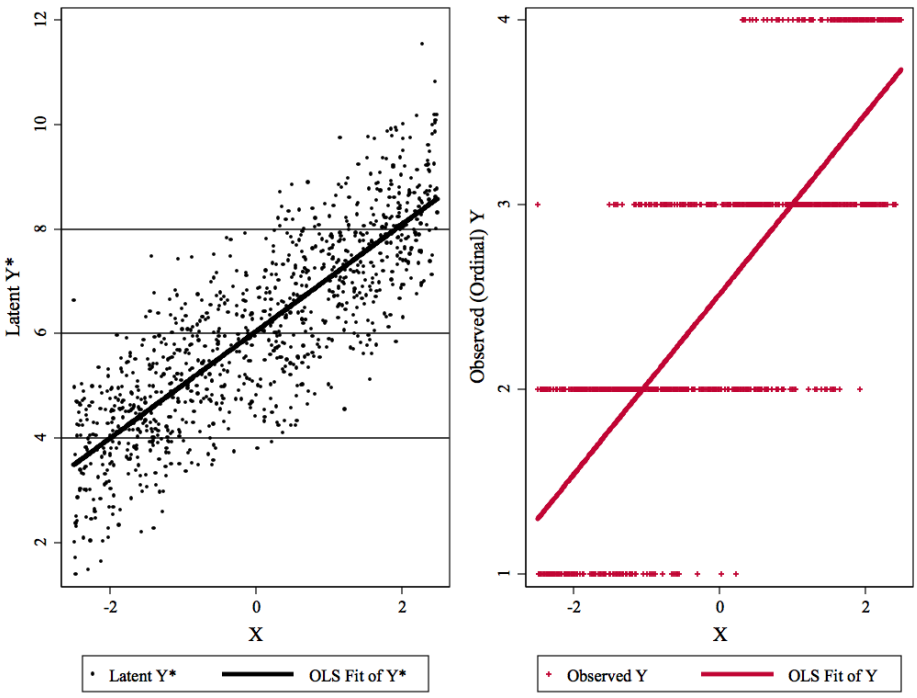
$$\begin{aligned} Y_{2i} &= 1 \text{ if } Y_i^* < 3 \\ &= 2 \text{ if } 3 \leq Y_i^* < 8 \\ &= 3 \text{ if } 8 \leq Y_i^* < 9 \\ &= 4 \text{ if } Y_i^* > 9 \end{aligned}$$

The latent values Y^* are the points in the left-hand panels of Figures 1 and 2, as regressed on X; the horizontal lines are the "cut points" by which Y^* is categorized into Y_1 and Y_2. The discrete values are the panels on the right-hand side of the two figures. Here, we "know" that the regression of Y^* on X should give us an estimate of $\hat{\beta} = 1.0$; and, in fact, it does:

```
. regress Ystar X

      Source |       SS       df       MS              Number of obs =    1000
-------------+------------------------------           F(  1,   998) = 2288.59
       Model |  2219.80531     1  2219.80531           Prob > F      =  0.0000
    Residual |  968.003275   998  .969943161           R-squared     =  0.6963
-------------+------------------------------           Adj R-squared =  0.6960
       Total |  3187.80859   999  3.19099959           Root MSE      =  .98486
------------------------------------------------------------------------------
       Ystar |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |   1.021344   .0213495    47.84   0.000      .979449    1.063239
       _cons |     6.0383   .0311446   193.88   0.000     5.977184    6.099416
------------------------------------------------------------------------------
```

Note that, in going from Y^* to Y_1, we've cut the variability of Y roughly in half; while the range of values on Y^* is from about 1.5 to 11 or so, that on Y_1 is one to four. Accordingly, regressing Y_1 on X should give us a $\hat{\beta}$ of roughly 0.5, which, as it happens, it does:

Figure 1: Regression of Continuous (Latent) Y^* and Discrete Y on X, Symmetrical "Cut-Points"

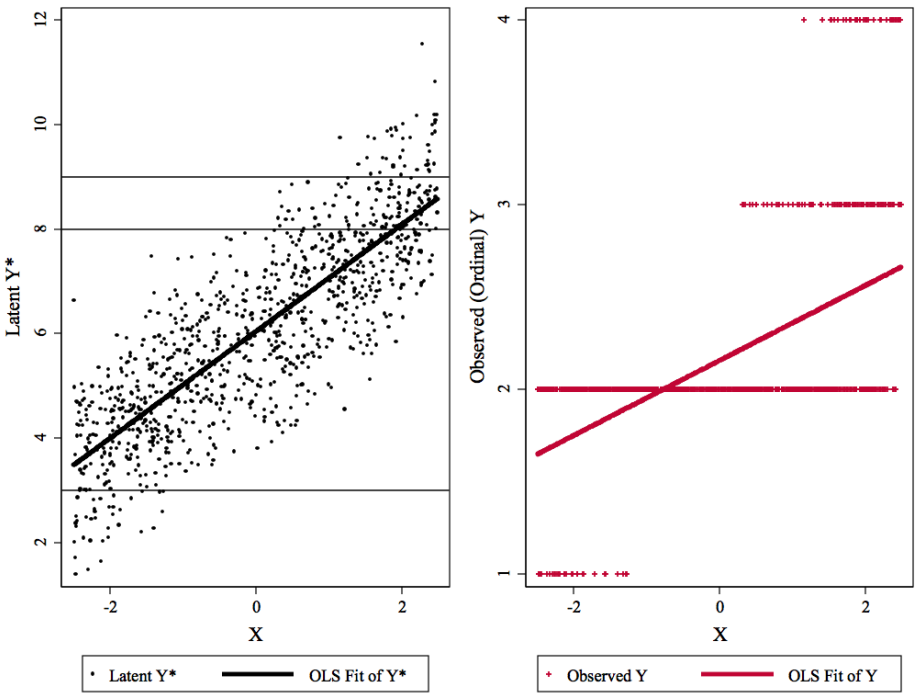



```
. regress Y1 X

      Source |       SS       df       MS              Number of obs =    1000
-------------+------------------------------           F(  1,   998) = 1650.66
       Model |   505.79375     1   505.79375           Prob > F      =  0.0000
    Residual |   305.80625   998  .306419088           R-squared     =  0.6232
-------------+------------------------------           Adj R-squared =  0.6228
       Total |       811.6   999  .812412412           Root MSE      =  .55355
------------------------------------------------------------------------------
          Y1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |     .48753   .0119998    40.63   0.000     .4639824    .5110777
       _cons |   2.515301   .0175052   143.69   0.000      2.48095    2.549653
------------------------------------------------------------------------------
```

This regression is illustrated in the right-hand panel of Figure 1. The two models (of Y^* and Y_1) are largely similar, in terms of significance levels, R^2, model fit, etc. OLS works well here because relatively little distortion of the data occurs in the process of "ordinalizing" them.

By comparison, consider what happens when we regress Y_2 on X:

Figure 2: Regression of Continuous (Latent) Y^* and Discrete Y on X, Asymmetrical "Cut-Points"



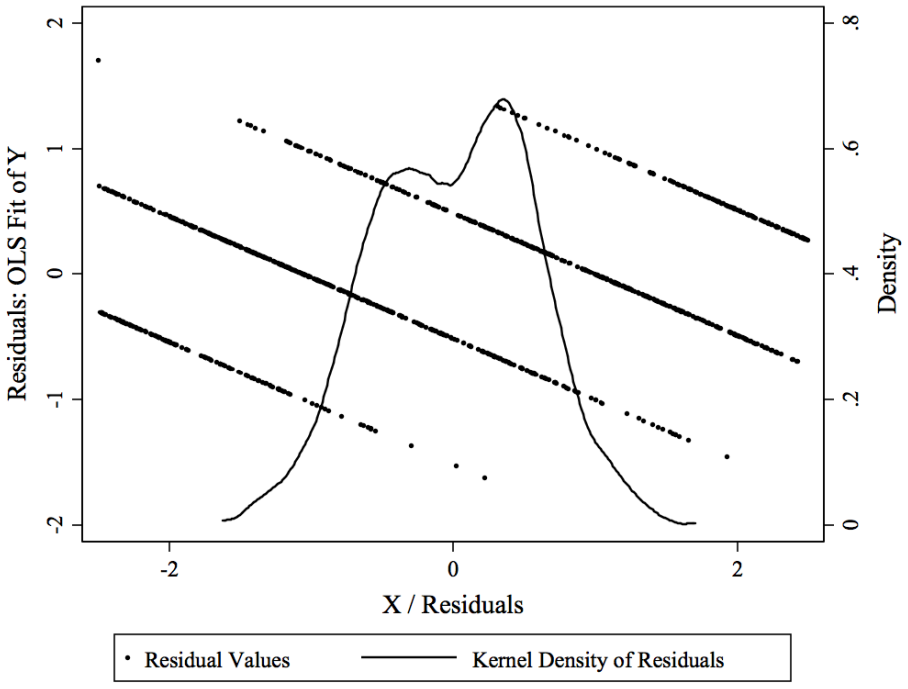

```
. regress Y2 X

      Source |       SS       df       MS              Number of obs =    1000
-------------+------------------------------           F(  1,   998) =  431.90
       Model |   87.508985     1   87.508985           Prob > F      =  0.0000
    Residual |  202.210015   998  .202615246           R-squared     =  0.3020
-------------+------------------------------           Adj R-squared =  0.3013
       Total |     289.719   999  .290009009           Root MSE      =  .45013
------------------------------------------------------------------------------
          Y2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |   .2027874   .0097578    20.78   0.000     .1836393    .2219355
       _cons |   2.157046   .0142346   151.54   0.000     2.129112    2.184979
------------------------------------------------------------------------------
```

Here again, we'd expect the marginal effect of X on Y_2 to be about 0.5. By collapsing the data unevenly, however, we've both reduced and distorted the variability in Y^*, such that the estimate $\hat{\beta}$ is only 0.2 or so. Also, the model does not fit as well; X does not account for variation in Y_2 nearly as well as it does in Y_1.

Non-Normal Errors

Beyond the possibility of bias due to using linear models on grouped/ordered outcomes, applying OLS to ordinal data will always yield heteroscedastic, nonnormal errors. Figure 3 shows the residuals that result from the OLS regression of Y_1 on X, plotted against values of X:

Figure 3: Residual Plot: OLS Regression of Y on X (Symmetrical "Cut-Points")



Notice a few things:

1. The residuals are grouped into a series of bands; this is because they all look like either $1 - \mathbf{X}\hat{\boldsymbol{\beta}}$, $2 - \mathbf{X}\hat{\boldsymbol{\beta}}$, $3 - \mathbf{X}\hat{\boldsymbol{\beta}}$, etc. As a result,

2. The distribution of the errors (the kernel density plot) indicates that the errors are slightly bimodal. In fact, errors from such models will often be multimodal, because of the categorical nature of Y.

3. Finally (while it really isn't the case in Figure 3), the OLS errors from an ordinal Y will often be heteroscedastic, particularly if (as was the case in Y_2 above) the categories are relatively imbalanced relative to the covariates.

The implications of all this are pretty straightforward:

- If your response variable is ordinal, but has many, many categories, continuous-linear models can generally work well.
- If your ordinal response variable has relatively few categories, but the categories have relatively balanced / symmetrical "cut points" (as might be the case if you had a measure of income grouped into $25,000 bands: 0-$25,000, $25,000-$50,000, etc.), then OLS/linear models can also work well.
- Finally, if you either (a) know or suspect that your ordinal response has asymmetrical "cut points," or if the "cut points" themselves are of some interest (as would likely be the case for, say, a Likert-type item), OLS is probably a bad option.

Motivation

What to do, then? Once again, start by considering a latent variable Y^*:

$$Y_i^* = \mu + u_i \tag{1}$$

with a corresponding observed indicator Y defined as:

$$Y_i = j \text{ if } \tau_{j-1} \leq Y_i^* < \tau_j,\ j \in \{1, ...J\} \tag{2}$$

- So, Y has J ordered outcome categories and $J-1$ "cut points" (usually denoted as τs).
- The "endpoint" categories 1 and J correspond to $\tau_0 = -\infty$ and $\tau_J = \infty$.
- Thus, if we have, say, four categories ($J = 4$), we get:

$$\begin{aligned} Y_i &= 1 \text{ if } -\infty \leq Y_i^* < \tau_1 \\ &= 2 \text{ if } \tau_1 \leq Y_i^* < \tau_2 \\ &= 3 \text{ if } \tau_2 \leq Y_i^* < \tau_3 \\ &= 4 \text{ if } \tau_3 \leq Y_i^* < \infty \end{aligned}$$

More generally, we can always write the probability of any particular discrete outcome on Y as equal to:

$$\begin{aligned} \Pr(Y_i = j) &= \Pr(\tau_{j-1} \leq Y^* < \tau_j) \\ &= \Pr(\tau_{j-1} \leq \mu + u_i < \tau_j) \end{aligned} \tag{3}$$

Now, if we allow the mean of Y^* to vary linearly with a vector of k covariates $\mathbf{X}$ (and their associated parameters $\boldsymbol{\beta}$):

$$\mu_i = \mathbf{X}_i\boldsymbol{\beta}$$

then we can rewrite Equation (3) as:

$$\begin{aligned}
\Pr(Y_i = j|\mathbf{X}, \boldsymbol{\beta}) &= \Pr(\tau_{j-1} \leq Y_i^* < \tau_j|\mathbf{X}) \\
&= \Pr(\tau_{j-1} \leq \mathbf{X}_i\boldsymbol{\beta} + u_i < \tau_j) \\
&= \Pr(\tau_{j-1} - \mathbf{X}_i\boldsymbol{\beta} \leq u_i < \tau_j - \mathbf{X}_i\boldsymbol{\beta}) \\
&= \int_{-\infty}^{\tau_j - \mathbf{X}_i\boldsymbol{\beta}} f(u_i)du - \int_{-\infty}^{\tau_{j-1} - \mathbf{X}_i\boldsymbol{\beta}} f(u_i)du \\
&= F(\tau_j - \mathbf{X}_i\boldsymbol{\beta}) - F(\tau_{j-1} - \mathbf{X}_i\boldsymbol{\beta}) \qquad (4)
\end{aligned}$$

where $f(\cdot)$ is the density for u, $F(\cdot)$ is the corresponding CDF, and the last result holds because of the Fundamental Theorem of Calculus.

The intuition is that we "cut" the density at (say) two points, τ_{j-1} and τ_j. The probability of a given observation i receiving the value of Y associated with this interval is just the area under the density curve between the two cut points (which we can get by integrating the entire area up to τ_j and then getting rid of the bit that falls below τ_{j-1} – see Figure 4).

Similar to what we did with binary logit/probit, we can then assume a distribution for the errors, and this gives us the probability statement we need to form a likelihood.

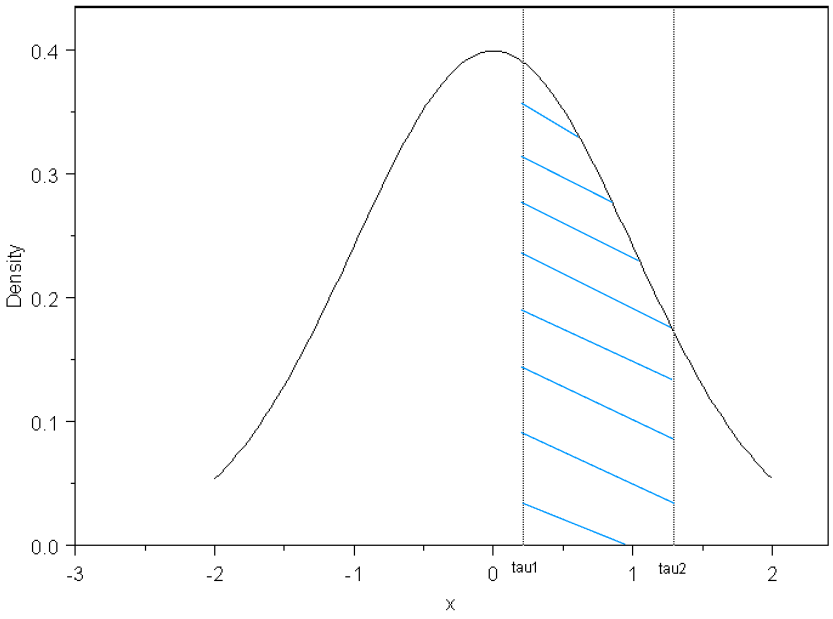
- Not surprisingly, we usually use either $N(0, 1)$ (that is, Φ) or a standard logistic (that is, Λ) for F.
- As in the binary case, which one we choose really doesn't matter...

Once we do this, we can evaluate the above probability statement (4) for each of the possible categories...

- Since $\tau_0 = -\infty$ and $\tau_J = \infty$, these correspond to probabilities of 0 and 1, respectively.
- This means that for, say, a four-category probit example ($J = 4$), we get:

$$\begin{aligned}
\Pr(Y_i = 1) &= \Phi(\tau_1 - \mathbf{X}_i\boldsymbol{\beta}) - 0 \\
\Pr(Y_i = 2) &= \Phi(\tau_2 - \mathbf{X}_i\boldsymbol{\beta}) - \Phi(\tau_1 - \mathbf{X}_i\boldsymbol{\beta}) \\
\Pr(Y_i = 3) &= \Phi(\tau_3 - \mathbf{X}_i\boldsymbol{\beta}) - \Phi(\tau_2 - \mathbf{X}_i\boldsymbol{\beta}) \\
\Pr(Y_i = 4) &= 1 - \Phi(\tau_3 - \mathbf{X}_i\boldsymbol{\beta})
\end{aligned}$$

Figure 4: "Cutting up" the density of u.



These then become our basic probability statements about Y. The general likelihood can then be written as:

$$L(Y|\mathbf{X}, \boldsymbol{\beta}, \tau) = \prod_{i=1}^{N} \prod_{j=1}^{J} [F(\tau_j - \mathbf{X}_i\boldsymbol{\beta}) - F(\tau_{j-1} - \mathbf{X}_i\boldsymbol{\beta})]^{\delta_{ij}} \tag{5}$$

where $\delta_{ij} = 1$ if $Y_i = j$ and 0 otherwise. The log–likelihood is then

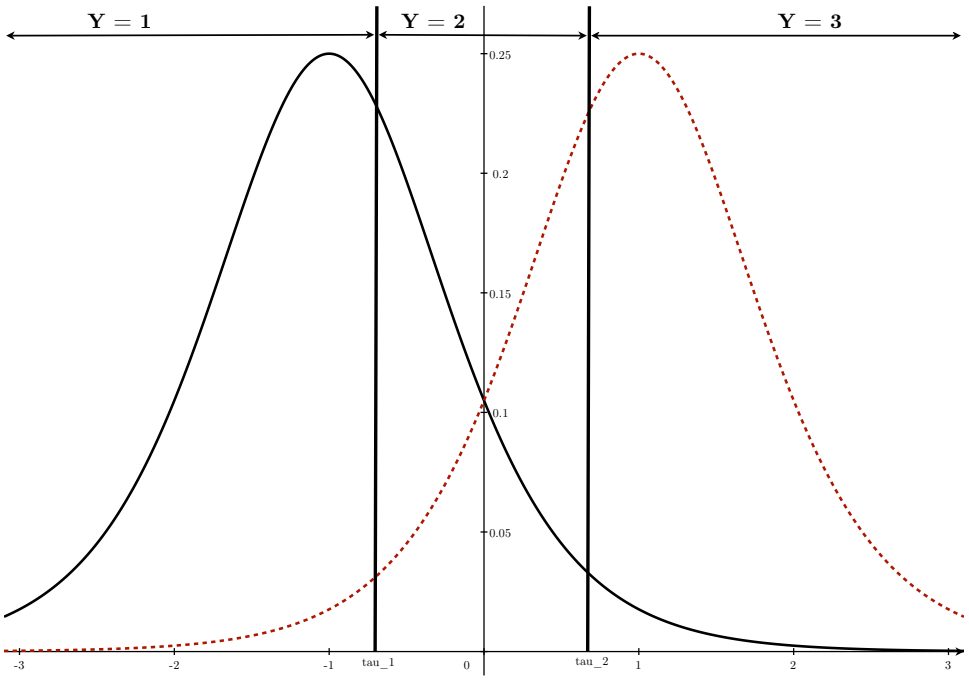
$$\ln L(Y|\mathbf{X}, \boldsymbol{\beta}, \tau) = \sum_{i=1}^{N} \sum_{j=1}^{J} \delta_{ij} \ln[\Phi(\tau_j - \mathbf{X}_i\boldsymbol{\beta}) - \Phi(\tau_{j-1} - \mathbf{X}_i\boldsymbol{\beta})] \tag{6}$$

for the *ordered probit* model, and

$$\ln L(Y|\mathbf{X}, \boldsymbol{\beta}, \tau) = \sum_{i=1}^{N} \sum_{j=1}^{J} \delta_{ij} \ln[\Lambda(\tau_j - \mathbf{X}_i\boldsymbol{\beta}) - \Lambda(\tau_{j-1} - \mathbf{X}_i\boldsymbol{\beta})] \tag{7}$$

for the ordered logit model. We can then estimate this model in the usual MLE way, look at the inverse of the Hessian to get standard errors, and so forth. We'll talk all about interpretation of these estimates next time...

Figure 5: Ordinal-Response Models: Shift in μ



Intuitively, we can think of these models as "shifting" the density of u along the X-axis, while holding the cut-points fixed: Changes in $\mathbf{X}$ (which, of course, change μ) can be thought of as moving the density of the errors relative to the cut-points, and therefore changing the relative probabilities of each of the various outcomes.

Consider Figure 5, which illustrates the three-category ordinal case. The two cut-points divide the space into regions of $Y = 1$, $Y = 2$, and $Y = 3$. The solid density plot has most of its mass to the left of the zero point, suggesting that the lower values of Y have greater probability; in fact, it's clear from looking at the areas under the curve relative to the various τs that $\Pr(Y = 1) > \Pr(Y = 2) > \Pr(Y = 3)$.

Assuming that a positive change in $\mathbf{X}$ increases the value of Y^*, an increase in $\mathbf{X}$ is represented by the dotted density. Note that the probabilities have reversed in magnitude: $\Pr(Y = 1) < \Pr(Y = 2) < \Pr(Y = 3)$. Had we shifted the density a bit less far to the right, we could also have had the case where (e.g.) $\Pr(Y = 2) > \Pr(Y = 3) > \Pr(Y = 1)$, or $\Pr(Y = 2) > \Pr(Y = 1) > \Pr(Y = 3)$.

One implication of this is that, for "middle" values of y (those not equal to 1 or J), the marginal change in $\Pr(Y = j)$ associated with a change in $\mathbf{X}$ can be positive or negative, *irrespective of the sign of* $\hat{\beta}$. We'll talk more about this next week.

Identification

There are two issues to deal with concerning identification of ordinal-response models like these. The first is that – as in the binary case – we cannot know or estimate σ_u^2 (the stochastic variance of the latent variable Y^*) from data on Y. As in our binary-response models, this is typically circumvented by assuming a value for σ_u^2 (e.g., $\sigma_u^2 = 1$ for the ordered probit model). And, as in the binary case, we can generalize this to allow σ_u^2 to vary with a set of covariates; we'll talk about that next time.

The second identification issue goes to the "cut-points." One can think of the τs in these ordered models as a series of "intercepts"...

- In a standard (linear or binary-response) model, the intercept is the "baseline" probability, that which holds when $\mathbf{X} = \mathbf{0}$. Think of this as the probability of being in each of the various categories for an observation with $\mathbf{X} = \mathbf{0}$.
- Intuitively, if we include all $J-1$ "cut points" in the model, we can change the intercept (that is, the location of the density curve, a la in Figure 5) arbitrarily and always "make up for" that change by shifting the thresholds in parallel.
- The result is that attempting to estimate the model with an intercept term *and* all $J-1$ of the cut-points renders the model unidentified.
- The nut of it is, you can either have an intercept term in $\mathbf{X}$, *or* you can estimate all $J-1$ "cut points" τ.

Mathematically, the two are equivalent, so that's not an issue. As a practical matter, both Stata (commands `-ologit-` and `-oprobit-`) and S-Plus/R (command `-polr-`) drop the intercept, and estimate all $j-1$ of the cut-points, which is probably the easier way to think about the model; LIMDEP does the opposite (and retains a "constant term" β_0).

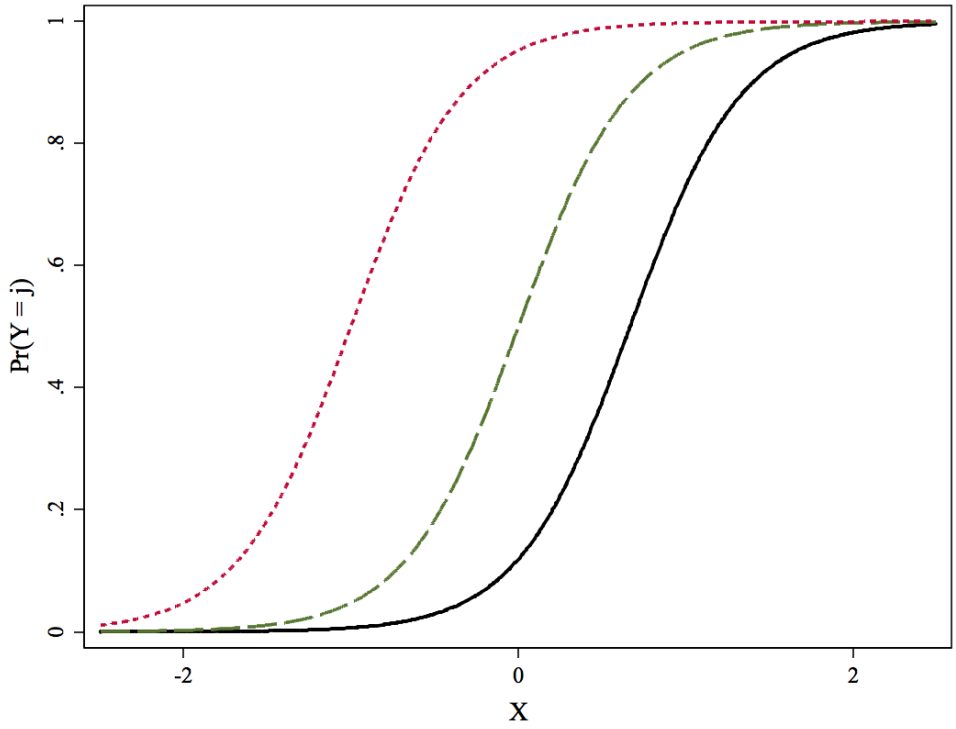
"Parallel Regressions"

Notice that, for any particular covariate X, we are estimating a single $\hat{\beta}$. This means that the effect of a variable is assumed to have a constant effect on the probability of $Y = j \forall j$. Long calls this the "parallel regressions" assumption. Formally,

$$\frac{\partial \Pr(Y_i = j)}{\partial X} = \frac{\partial \Pr(Y_i = j')}{\partial X} \; \forall \; j \neq j' \tag{8}$$

This means that the impact of X is the same across all J possible values of Y; the CDF "curve" simply "shifts" to the left/right. Intuitively, this means that the "slope" of the S-curve associated with a particular covariate X is the same across different values of Y; you can get a sense of this in Figure 6.

Figure 6: Parallel Regressions – Identical βs for each j.



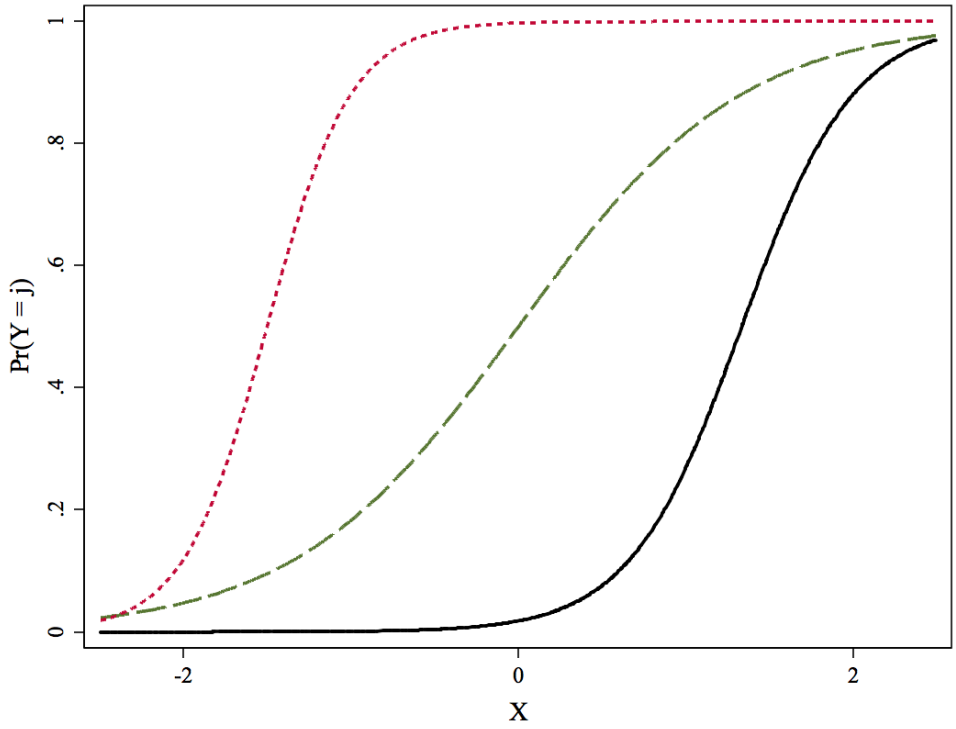
This assumption (which is also occasionally called the *proportional odds assumption*) is somewhat restrictive, in the sense that we may not expect this to be the case. Formally,

$$\frac{\partial \Pr(Y_i = j)}{\partial X} \neq \frac{\partial \Pr(Y_i = j')}{\partial X} \; \forall \; j \neq j' \tag{9}$$

That is, we might think that the effect of a covariate varies (perhaps, increases) as you move "higher up" the ordered categories. For example, suppose our Y variable is an ordinal indicator of the degree of conflict between two nations (say, from no conflict, to a low-level diplomatic dispute, to border clashes, to full-scale war). In such an example, we might expect the influence of certain covariates on the level/intensity of the dispute to be different at different levels.

So, whether or not the two countries have a formal military alliance might have a large (negative) effect on whether or not they go to war with one another, but it may not have much influence at all on whether the two get into low-level disputes. Conversely, whether two countries share a land border might have (comparatively) little influence on whether they wind up in a diplomatic dispute, but it could have a large (positive) impact on whether such a dispute escalates to an armed conflict, since contiguity makes such armed conflict more possible.

Figure 7: Non-Parallel Regressions – Different βs for each j.



The idea of nonparallel regressions is illustrated in Figure 7. There, the marginal effect of a change in X is different for each of the three categories j. Assuming the data follow (8) (and estimating a model that fixes them to be equal) when in fact (9) is the case amounts to a form of specification bias, and thus should be avoided.

As we'll see next week, you can test this by comparing the ordered probit/logit results to a series of binary regressions, using a Wald or LR test – we talk more about this next class.