Ordered Variables

Michael D. Ward[1]

[1]Department of Political Science
Duke University, Durham, NC

February 13, 2012

Outline of Topics

1 Extension

2 Summary

Agenda

- Ordinal and categorical data
- Motivating the ordered probit and logit models
- Latent Variables and Thresholds
- The ordered probit model: The log-likelihood function
- An example
- Interpreting coefficients & model fit considerations

Ordered Categorical Data

- Ordered probit and logit models extend what we've learned about dichotomous categorical data to ordered categorical data
- For discrete outcomes (i.e., dependent variables) that are inherently ordered, we employ ordered logit or ordered probit models.
- If the categories were not ordered (e.g., "north", "south", "east", "west"), multinomial logit would be more appropriate.

An Example of Ordered Categorical Data

- Consider responses to the 2000 National Election Survey (NES)question:
- “How much attention do you pay to newspaper articles about the campaign for President – a great deal, quite a bit, some, very little, or none?”
- Responses to this question could be coded as a dummy variable ranging from 1 to 5 with "none" $=1$ and "a great deal" $=5$. The data thus represent responses that fall into discrete, mutually exclusive categories.
- These categories are *ordered*, with "none" $<$ "very little" $<$ $\ldots <$ "a great deal" .

A Few Recent Examples from American Politics

- Volden (*AJPS* 2002)
- Looks at US states' propensities to adjust AFDC benefits for inflation.
- Uses ordered logit on the number of years since last increasing benefits.
- Hetherington and Globetti (*AJPS* 2002)
- Uses NES data to evaluate the support for affirmative action policies as a function of race and trust in the federal government.

A Few Recent Examples from IR

- Chiozza and Choi (*Journal of Conflict Resolution*, 2003)
- Examines leadership tenure and propensity to make compromises in territorial disputes.
- DV is a 3-pt. scale of dispute resolution, ranging from military action to accommodation
- Drury (*Journal of Peace Research*, 1998)
- Investigates the determinants of economic sanctions effectiveness. Effectiveness ranges between 1 and 4
- Yoon (*JCR*, 1997)
- Examines US intervention in Third World conflicts.
- "Intervention" takes values in {no intervention, non-military intervention, indirect military intervention, military intervention}

Latent Variables

- In the NES example, the questionnaire only allows us to observe Y_i, the category into which respondent i in $i = 1, \ldots, N$ falls.
- The observed data, however, are a function of the unobserved variable, Y_i^*.
- In the NES case we can think of Y_i^* as representing the actual amount of time spent reading relevant newspaper articles.
- Y_i^* is a *latent variable*.
- We assume Y_i^* is continuous (whereas the observed data are discrete).

Threshold Models

- In order to use the latent variable construction we need some way of mapping the values of Y_i^* into the $j = 1, \ldots, m$ categories of Y_i.
- In the NES example $m = 5$
- Consider a threshold value τ_1 such that $Y_i = 1$ if $Y_i^* < \tau_1$. This is simply the standard logit/probit model with dichotomous dependent variable.
- By adding more τs we increase the number of thresholds and categories

Thresholds & Hypothetical NES Responses

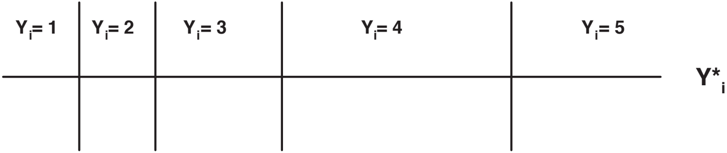
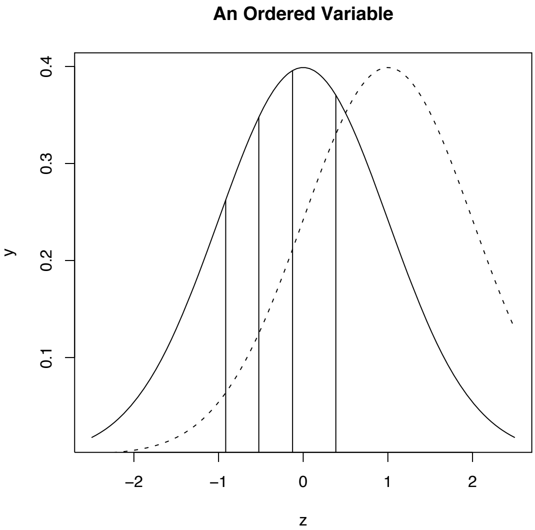


Figure: NES Response categories along the (continuous) real number line with hypothetical response proportions

Thresholds & the Standard Normal Distribution

- Assume observations are drawn from a standard normal distribution ($\mu = 0$ and $\sigma^2 = 1$).
- We can then choose thresholds that divide the density into region matching the (hypothetical) response data.
- e.g.,`qnorm(.18 + .12)` for the second threshold
- Q: What happens if we shift the mean of the distribution to the right 1 unit ($\mu = 1$) while keeping the thresholds constant?
- A: The mass of the distribution becomes more concentrated in the upper categories.
- We'll return to this idea when we interpret coefficients

Normal Densities



Latent Model

Like our general approach, this combines a stochastic component, which is more detailed than in the case of many other linear models:

$$\begin{aligned} Y_i^* &\sim \text{logit}(y_i^* \mid \mu_i), \quad \text{where} \\ Y_i^* &= j \iff \tau_{j-1} \leq Y_i^* \leq \tau_j \quad \forall \quad j \in \{1, \ldots, j\} \end{aligned}$$

where τ_j are the threshold parameters which divide the unobserved, latent variable into observed, modeled categories j.
This lead to a stochastic component with the following form:

$$P(Y \leq j) = P(Y^* \leq \tau_j) \quad = \quad \frac{\exp(\tau_j - \mathbf{x}_i\beta)}{1 + \exp(\tau_j - \mathbf{x}_i\beta)}$$

More Orders

To find the probability that y_i^* falls in any specific category j, you just need to difference adjacent categories:

$$\pi_j \quad = \quad \frac{\exp(\tau_j - \mathbf{x}_i\beta)}{1+\exp(\tau_j - \mathbf{x}_i\beta)} - \frac{\exp(\tau_{j-1} - \mathbf{x}_i\beta)}{1+\exp(\tau_{j-1} - \mathbf{x}_i\beta)}$$

.

This is because

$$\int_{-\infty}^{\tau_j - \mathbf{X}_i\beta} f(u_i)du - \int_{-\infty}^{\tau_{j-1} - \mathbf{X}_i\beta} f(u_i)du = F(\tau_j\mathbf{X}_i\beta) - F(\tau_{j-1}\mathbf{X}_i\beta),$$

where F is the density for u, in this case the logistic function Λ.

Generalizing the Model

- μ_i moves the distribution around, representing the individual differences across observations.
- If we write μ_i as linear function of covariates x_i and parameter vector β we have a linear predictor term:
- $\boxed{\mu_i = x_i\beta}$
- In ordered probit Y_i^* is modeled as stylized normal:
- $\boxed{Y_i^* \sim f_{stn}(y_i^* \mid \mu_i, \sigma^2 = 1)}$
- σ^2 is fixed at 1 for identification purposes.
- Ordered logit proceeds similarly, except we replace the stylized normal distribution with the logistic distribution.

Generalizing the Model

- Define thresholds $\tau_j = 1, \ldots, m$ such that $\tau_1 = -\infty$ and $\tau_m = +\infty$ and $\tau_1 < \tau_2 < \ldots < \tau_m$
- Thus y_i^* is mapped into category j of Y_i if
- $\boxed{\tau_{j-1} < y_i^* \leq \tau_j}$
- We can then break y_i into several dichotomous variables y_{ji} where
- $$\boxed{y_{ji} = \begin{cases} 1 \text{ if } \tau_{j-1} < y_i^* \leq \tau_j \\ 0 \text{ otherwise} \end{cases}}$$

The Likelihood Function

Now we derive the probability that y_{ji} is in category j:

$$\begin{aligned} Pr(Y_{ji} = 1) &= Pr(\tau_{j-1} < Y_i^* \leq \tau_j) \\ &= \int_{\tau_{j-1}}^{\tau_j} f_{stn}(y_i^* \mid \mu_i, 1) dy_i^* \\ &= F_n(\tau_j \mid \mu_i, 1) - F_n(\tau_{j-1} \mid \mu_i, 1) \\ &= F_n(\tau_j \mid x_i\beta, 1) - F_n(\tau_{j-1} \mid x_i\beta, 1) \end{aligned}$$

where the integral of $f_{stn}(\cdot, \cdot)$ is the Normal CDF, $F_n(\cdot, \cdot)$.

The Likelihood Function (cont'd)

- We must take this probability over all m categories and n observations.
- The full likelihood function (assuming y_i is i.i.d.) is thus

$$\boxed{\mathcal{L}(\tilde{\tau}, \tilde{\beta} \mid y) = \prod_{i=1}^{n} \prod_{i=1}^{m} \left[F_n(\tilde{\tau}_j \mid x_i \tilde{\beta}, 1) - F_n(\tilde{\tau}_{j-1} \mid x_i \tilde{\beta}, 1) \right]^{y_{ji}}}$$

The Log Likelihood

- Taking the natural log we get the log-likelihood:

$$\boxed{\ln \mathcal{L}(\tilde{\tau}, \tilde{\beta} \mid y) = \sum_{i=1}^{n} \sum_{i=1}^{m} y_{ji} \ln[F_n(\tilde{\tau_j} \mid x_i\tilde{\beta}, 1) - F_n(\tilde{\tau_{j-1}} \mid x_i\tilde{\beta}, 1)]}$$

- Note that we are estimating *both* the $(m-2) \times 1$ vector of thresholds, τ, *and* the $(k+1) \times 1$ vector of β coefficients at the same time, for a total of $m+k-1$ parameters (this includes an intercept).
- ordered probit is sometimes referred to as the "proportional odds" model because the $\frac{Pr(Y_i \leq m)}{Pr(Y_i > m)}$ is assumed to be constant independent of m
- The MLE is then calculated by maximizing the log likelihood.

Identification Issues and other issues

- We estimate $J-1$ thresholds if there are J categories. This is because we set the expected value for one category, so that we are avoiding an identification problem.
- But how do we pin down the location of the thresholds when the underlying scale is unknown? Shifting the scale doesn't affect the probabilities. This is typically accomplished by setting one of the τ to zero, or setting $\beta_0 = 0$.
- Parallel Regression, this means that as an independent variable increases (decreases), for example, the cumulative distribution shifts to the right (left). However, there is no shift in the slope of the distribution, thus the $\hat{\beta}_k$ are equal for every category. You can test this by estimating $J-1$ binary regressions, and testing whether all the β across regressions are equal.
- We need to test whether the τs are different? Do their credible intervals overlap? You can do a z-test for this.

Computation: Maximizing The Log Likelihood

- There are several different off-the-shelf ordered logit/probit functions in various `R` libraries:
- `polr()` in the `MASS` library for ordered *logit*
- `nordr()` in the `gnlm` library also employs a logit link.

An Example: the success of economic sanctions

- Drury (1998) re-examines the factors that determine whether or not economic sanctions are ”successful”
- 117 sanction episodes are examined
- His dependent variable is a scale of policy change on by the sanction target, ranging from 1 (failure) to 4 (success).
- For the first model he estimates, the linear predictor is specified as follows:
- Success = distress + GNP ratio−cooperation−black knight+trade+target GNP cost−sender cost+policies+year-US
- black knight, policies, cooperation, sender cost, and US are dummy variables
- In `R`: `result` $\sim$ `dstres` + `log(gnprat)`+ `coop` + `blknight` + `trade` + `gnpcst` + `cost` + `policies` + `year` + `us`

Results: This is what you get

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	24.487	1.065	0.000
dstres	−0.4889	0.195	0.012
coop	−0.147	0.115	0.202
blknight	−0.436	0.279	0.119
trade	0.004	0.003	0.151
cost	−0.207	0.178	0.245
policies	0.197	0.279	0.481
year	−0.011	0.000	0.000
us	0.125	0.275	0.649
log(gnprat)	−0.108	0.067	0.108
gnpcst	0.089	0.054	0.097
τ_1	0.000	NA	
τ_2	0.889	0.137	
τ_3	1.407	0.160	
Log-likelihood	−145.47		

Interpreting Coefficients & Goodness of Fit

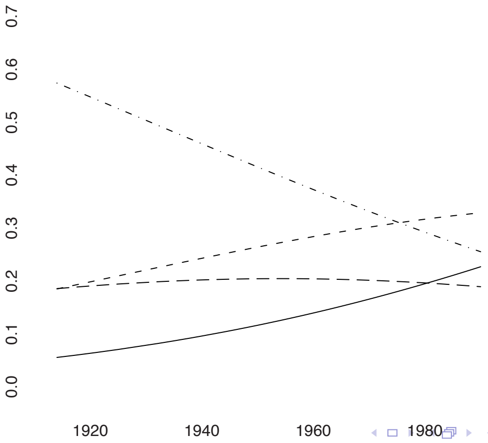
- Interpreting coefficients in ordered logit and probit analysis is similar to standard binary logit/probit. Interpretations can be presented using:
- marginal effects (holding other variables constant)
- graphical depictions of the relationship between particular IVs and the DV (holding other variables constant)
- "scenario" construction and comparison
- Goodness of fit can be evaluated using:
- Likelihood ratio tests for nested models
- Internal predictive power: how well does the model classify observations?
- Rather than a 2-by-2 chart, you will get a *m*-by-*m* chart of actual vs. predicted categories.

Predicted Probabilities

- To find $Pr(\texttt{result} = m)$, holding the other variables at some constant level, we calculate: $Pr(\texttt{result} = m|\mathbf{x}_*) = F(\widehat{\tau_m} - \mathbf{x}_*\widehat{\beta}) - F(\widehat{\tau_{m-1}} - \mathbf{x}_*\widehat{\beta})$
- $\mathbf{x}_*$ is the $n \times k+1$ matrix of $k-1$ covariates held at some "reasonable" value along with the vector of the variable of interest and the intercept
- $F(\cdot)$ is the Normal CDF with $\sigma^2 = 1$
- $\widehat{\tau_m}$ and $\widehat{\beta}$ are the estimated threshold and regression parameters
- for $m = 1$: $F(\tau_1 - \mathbf{x}_*\widehat{\beta})$ (remember $\tau_1 = 0$)
- for $m = 2$: $F(\tau_2 - \mathbf{x}_*\widehat{\beta}) - F(\tau_1 - \mathbf{x}_*\widehat{\beta})$
- for $m = 3$: $F(\tau_3 - \mathbf{x}_*\widehat{\beta}) - F(\tau_2 - \mathbf{x}_*\widehat{\beta})$
- for $m = 4$: $1 - F(\tau_3 - \mathbf{x}_*\widehat{\beta})$

Predicted Probabilities by Year in the Drury Example

There seems to be an effect over time, with sanctions decreasing in effectiveness.



How'd he do that?

```
x.star.year<-cbind(rep(1,n),rep(median(dstres),n)...
    ...year,...,rep(mean(gnpcst),n))
#
pprob.year.1<-pnorm(lambda[1]-(x.star.year %*% beta.hat))
pprob.year.2<-pnorm(lambda[2]-(x.star.year %*% beta.hat))
                       -pnorm(lambda[1]-(x.star.year %*% beta.hat))
pprob.year.3<-pnorm(lambda[3]-(x.star.year %*% beta.hat))
                       -pnorm(lambda[2]-(x.star.year %*% beta.hat))
pprob.year.4<-1-pnorm(lambda[3]-(x.star.year %*% beta.hat))
#
plot(year,pprob.year.1 )
lines(year,pprob.year.2,lty=2 )
lines(year,pprob.year.3,lty=5,lwd=2 )
lines(year,pprob.year.4,lty=4,lwd=2 )
```

First Differences

- We want to describe how much the $Pr(\texttt{result} = m|\mathbf{x})$ changes as some x goes from one value of interest to another, holding all others constant.
- more formally, we want
- $\frac{\Delta Pr(\texttt{result}=m|\mathbf{x}_*)}{\Delta x_k} = Pr(y = m|\mathbf{x}_*, x_k = x_k^{max}) - Pr(y = m|\mathbf{x}_*, x_k = x_k^{min})$
- x_k is the variable of interest
- in this case, we are examining it as it changed from it's minimum to maximum values
- this is often of interest for examining the marginal effects of dummy variables

First Differences in the Drury example

How much does a change in distress from 1 to 3 alter the predicted effectiveness of sanctions?

result	$m = 1$	$m = 2$	$m = 3$	$m = 4$
$\Delta Pr(y = m)$	0.649	0.439	0.091	-0.179

When `dstres` changes from 1 to 3, the predicted probability that `result`=4 changes by -0.179, holding the year constant at 1973 and all other variables at their mean values (medians for dummies). This makes sense: we estimated a negative coefficient, which shifts the mass into the lower categories.

How'd he do that? Part Deux

```
x.star.dstres.1<-cbind(rep(1,n),rep(1,n),rep(median(coop),n)...)
x.star.dstres.3<-cbind(rep(1,n),rep(3,n),rep(median(coop),n)...)
pprob.dstres.1.1<-pnorm(lambda[1]-
     (x.star.dstres.1\%*\%beta.hat)
    $\vdots$
pprob.dstres.1.4<-1-pnorm(lambda[3]-(x.star.dstres.1 \%*\% beta.hat))
pprob.dstres.3.1<-pnorm(lambda[1]-
    (x.star.dstres.3 \%*\%beta.hat))
 $\vdots$
pprob.dstres.3.4<-1-pnorm(lambda[3]-
    (x.star.dstres.3 \%*\%beta.hat))
Diff.1<-2*(pprob.dstres.3.1-pprob.dstres.1.1)
Diff.2<-2*(pprob.dstres.3.2-pprob.dstres.1.2)
Diff.3<-2*(pprob.dstres.3.3-pprob.dstres.1.3)
Diff.4<-2*(pprob.dstres.3.4-pprob.dstres.1.4)
```

CHOPIT

- There is a wide variety of sculpted ordered models to deal with heteroscedasticity and to model explicitly model the τ_is as functions of other variables, for example.
- Recently, Gary King and colleagues at the WHO introduced the idea of a conditional probit model that was calibrated across different situations via special sets of questions called anchoring vignettes
- he basic idea is that different groups of people may have different levels at which they calibrate their ordinal responses. King et alia found, for example, that self-assessments of political efficacy in the June 2002 World Health Organization surveys illustrated that about 50% of Mexican respondents perceived that they had no say in governance via elections, while only about 28% of Chinese had similar self-assessments. This leads to the conclusion that on average, political efficacy is higher in China than in Mexico, an assessment that is widely believed to be erroneous.

Vignettes

Anchored

- Anchoring vignettes are survey vignettes that are asked of sets of respondents in surveys to calibrate their ordinal responses to self-assessment questions.

Vignettes

Anchored

- Anchoring vignettes are survey vignettes that are asked of sets of respondents in surveys to calibrate their ordinal responses to self-assessment questions.
- The basic idea is that a vignette can be constructed that has the same meaning to all–since it is hypothetical–and that the responses to such vignettes can be used to calibrate across different groups through the use of a conditional hierarchical model.

- Anchoring vignettes are survey vignettes that are asked of sets of respondents in surveys to calibrate their ordinal responses to self-assessment questions.
- The basic idea is that a vignette can be constructed that has the same meaning to all–since it is hypothetical–and that the responses to such vignettes can be used to calibrate across different groups through the use of a conditional hierarchical model.
- The Likelihoods needed for this analysis are available in the appendix to the King et alia paper (page 205) and are also available in the $\mathcal{R}$ package called `anchors`.

- Anchoring vignettes are survey vignettes that are asked of sets of respondents in surveys to calibrate their ordinal responses to self-assessment questions.
- The basic idea is that a vignette can be constructed that has the same meaning to all–since it is hypothetical–and that the responses to such vignettes can be used to calibrate across different groups through the use of a conditional hierarchical model.
- The Likelihoods needed for this analysis are available in the appendix to the King et alia paper (page 205) and are also available in the $\mathcal{R}$ package called `anchors`.
- I used this approach to model political trust in surveys undertaken in the North Caucasus Region of Russia and Bosnia in the Winter of 2005 (Ward et al 2005.

Big Table of Numbers

Table: *Estimates of Stated Trust of National Groups via ordered Probit as well as Conditional Hierarchical Ordered Probit (CHOPit); N=3017. Estimates significant at $p < .05$ are shown in bold. These results need to be updated.*

	Ordered Probit			Chopit		
Variable	Estimate	σ	t-value	Estimate	σ	t-value
Bosnia	-0.62	0.05	-13.37	-0.75	0.08	
Gender	0.00	0.04	0.03	-0.05	0.06	
Age	0.00	0.00	0.34	-0.00	0.00	
Education	0.08	0.02	4.70	0.10	0.03	
Material status	0.02	0.03	0.91	-0.01	0.04	
Ethnic relations	-0.17	0.03	-5.12	-0.06	0.06	
Current situation	-0.07	0.04	-2.01	-0.11	0.06	
Ethnic friends	-0.16	0.02	-8.16	-0.05	0.03	
Closest friends	0.24	0.03	9.40	0.29	0.04	
Pride	0.04	0.02	1.78	0.10	0.03	
	0.14	0.10	1.45	-0.07	0.16	
Violence	0.15	0.04	3.39	0.30	0.07	
cut-points	$\hat{\tau}$	σ	t-value			
1—2	-1.73	0.17	-10.40	-1.68	0.26	
2—3	-1.00	0.16	-6.08	0.49	0.22	
3—4	-0.30	0.16	-1.81	0.64	0.15	
4—5	0.65	0.16	3.98	1.62	0.18	

Chopit Graph

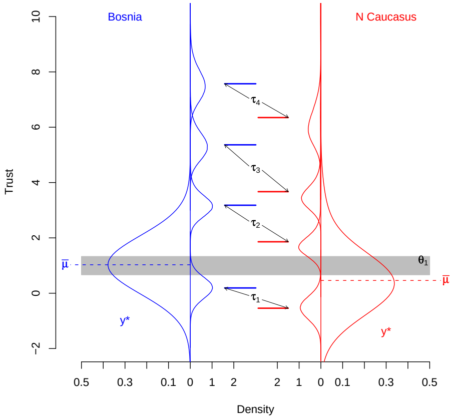


Figure: *Plot of relation between ordered probit model of trust in Bosnia and the North Caucasus Region of Russia, corrected by anchoring vignette.*

Summing Up

- Ordered logit and probit are the appropriate model structures when modeling categorical data that can be described in ordinal terms
- These models are based on the concept of an unobserved ("censored") latent variable. We don't observe the "real" variable, only whether it falls between certain thresholds.
- Probit assumes the error term is distributed Normally. Logit models assume the error is distributed logistically. There is rarely a theoretical reason for choosing one over the other.
- Models are estimated via maximum likelihood estimation. Be careful with "off the shelf" code.
- Model fit can be examined via likelihood ratio and internal prediction methods
- Interpretation and presentation can take the form of predicted probabilities, marginal effects plots, and first differences.