MULTINOMIAL AND ORDERED RESPONSE MODELS

Econometric Analysis of Cross Section and Panel Data, 2e
MIT Press
Jeffrey M. Wooldridge

1. Introduction
2. Multinomial Logit
3. Probabilistic Choice Models
4. Ordered Response Models

1. INTRODUCTION

- Two ways to extend the binary response: unordered and ordered outcomes. In both cases, it is convenient to label the possible outcomes on y as $\{0, 1, \ldots, J\}$, so y takes on $J + 1$ different values.
- In the unordered (or nominal) case, the labeling of outcomes is totally arbitrary. For example, if y is mode of transportation to work, we might use the follow labels: 0 is by car without pooling, 1 is car pooling, 2 is bus, and 3 is rapid transit (train). Nothing changes if we switch the labels.
- Another example of an unordered outcome is different kinds of health insurance.

- In other cases the order matters. For example, each person applying for a mortage is given a credit rating in the set $\{0, 1, 2, 3, 4, 5, 6\}$. The fact that a credit rating of 5 is better than 4, and that 1 is better than 0, is important.
- Such outcomes are ordinal because we could replace the values by any other set that preserves the ranking. In other words, cardinality does not matter, but the order does.

2. MULTINOMIAL LOGIT

- In the basic multinomial logit (MNL) model, y is an unordered response and we have a set of conditioning variables, $\mathbf{x}$, that change by unit but not alternative. For example, in modeling type of health insurance, we include observable characteristics of the individual but not of the different kinds of health plans.
- In this setting, we are interested in the response probabilities,

$$p_j(\mathbf{x}) = P(y = j|\mathbf{x}), j = 0, \ldots, J.$$

Since exactly one choice is possible,

$$p_0(\mathbf{x}) + p_1(\mathbf{x}) + \ldots + p_J(\mathbf{x}) = 1 \text{ for all } \mathbf{x}$$

- We are interested in how changing elements of $\mathbf{x}$ affects the response probabilities.
- The MNL response probabilities are

$$P(y = j|\mathbf{x}) = \frac{\exp(\mathbf{x}\boldsymbol{\beta}_j)}{\left[1 + \sum_{h=1}^{J} \exp(\mathbf{x}\boldsymbol{\beta}_h)\right]}, \; j = 1, \ldots, J$$

$$P(y = 0|\mathbf{x}) = \frac{1}{\left[1 + \sum_{h=1}^{J} \exp(\mathbf{x}\boldsymbol{\beta}_h)\right]}$$

where in almost all applications $x_1 \equiv 1$.

• We can write the response probabilities in common form, using the first equation, by defining $\boldsymbol{\beta}_0 \equiv \mathbf{0}$.

• Unless $J = 1$ (binary response logit), the partial effects are complicated. For a continuous x_k,

$$\frac{\partial p_j(\mathbf{x})}{\partial x_k} = p_j(\mathbf{x})\left\{\beta_{jk} - \frac{\left[\sum_{h=1}^{J} \beta_{hk} \exp(\mathbf{x}\boldsymbol{\beta}_h)\right]}{\left[1 + \sum_{h=1}^{J} \exp(\mathbf{x}\boldsymbol{\beta}_h)\right]}\right\},$$

which need not be the same sign as β_{jk}.

• Easier to interpret:

$$\frac{p_j(\mathbf{x})}{p_0(\mathbf{x})} = \exp(\mathbf{x}\boldsymbol{\beta}_j)$$

- The log odds of response j relative to response 0 is

$$\log\left[\frac{p_j(\mathbf{x})}{p_0(\mathbf{x})}\right] = \mathbf{x}\boldsymbol{\beta}_j,$$

and so β_{jk} measures the partial effect of x_k on the log odds of j relative to outcome 0.

- A key feature of the MNL model is that if we condition on any two outcomes, the resulting model for choosing between the outcomes is a binary response logit. That is, suppose we condition on the event that $y \in \{j, h\}$.

- Then

$$P(y = j | y = j \text{ or } y = h) = p_j(\mathbf{x}, \boldsymbol{\beta}) / [p_j(\mathbf{x}, \boldsymbol{\beta}) + p_h(\mathbf{x}, \boldsymbol{\beta})]$$
$$= \frac{\exp(\mathbf{x}\boldsymbol{\beta}_j)}{[\exp(\mathbf{x}\boldsymbol{\beta}_j) + \exp(\mathbf{x}\boldsymbol{\beta}_h)]} = \frac{\exp[\mathbf{x}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_h)]}{\{[\exp[\mathbf{x}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_h)] + 1\}}$$
$$= \Lambda[\mathbf{x}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_h)].$$

- In other words, $P(y = j | y = j \text{ or } y = h)$ has the logit form with parameter vector $\boldsymbol{\beta}_j - \boldsymbol{\beta}_h$.
- This is an artifact of the MNL functional form.

• Maximum likelihood estimation of the $\boldsymbol{\beta}_j$ is straightforward. The log likelihood for random draw $(\mathbf{x}_i, y_i)$ is

$$\ell_i(\boldsymbol{\beta}) = \sum_{j=0}^{J} 1[y_i = j] \log[p_j(\mathbf{x}_i, \boldsymbol{\beta})].$$

• Inference is standard. The expected Hessian given $\mathbf{x}_i$ is easy to compute.

• In terms of goodness of fit and prediction, the MNL model often works well. Can choose $\mathbf{x}$ to be flexible functions of underlying explanatory variables.

3. PROBABILISTIC CHOICE MODELS

- Again, let there be $J + 1$ choices, but now explicitly view the response (choice) as maximizing underlying utility. For a random draw i, the latent utilities are

$$y_{ij}^{*} = \mathbf{x}_{ij}\boldsymbol{\beta} + a_{ij},\ j = 0, \ldots, J,$$

where $\mathbf{x}_{ij}$ can vary by unit (i) and choice (j). Notice that $\boldsymbol{\beta}$, in this formulation, does not depend on j.

- For example, $\mathbf{x}_{ij}$ can include the costs of various modes of transportation for each unit i. Its coefficient measures the effect of cost on utility across any mode of transportation.

• Sometimes a variable will change only by choice and not individual (such as the price of a car). In more sophisticated settings, another dimension – such as market (often measured by geographic location) – is added to the problem. Then, price can change by market and brand, but not by individual.

• Let $\mathbf{x}_i$ include all nonredundant elements of $(\mathbf{x}_{i0}, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iJ})$. Let $\mathbf{a}_i = (a_{i0}, a_{i1}, \ldots, a_{iJ})$ and assume $\mathbf{a}_i$ is independent of $\mathbf{x}_i$ (exogeneity).

• The observed choice $y_i \in \{0, 1, \ldots, J\}$ is the one that maximizes utility:

$$y_i = argmax(y_{i0}^*, y_{i1}^*, \ldots, y_{iJ}^*),$$

that is, $y_i = j$ if choice j yields the highest utility.

• McFadden (1974) showed that if the $\{a_{ij} : j = 0, 1, \ldots, J\}$ are independent, identically distributed with the *type I extreme value distribution*, that is, with cdf $F(a) = \exp[-\exp(-a)]$, then it can be shown that

$$P(y_i = j|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_{ij}\boldsymbol{\beta})}{\left[1 + \sum_{h=1}^{J} \exp(\mathbf{x}_{ih}\boldsymbol{\beta})\right]}, \; j = 0, 1, \ldots, J.$$

• In the context of probabilistic choice models, usually called the *conditional logit model* (the name given by McFadden).

- Easy to estimate β, which is common to all choices, by MLE.
- The type I extreme value distribution is not especially natural because it is not symmetric – it has a thicker right tail. But it does roughly have a "bell shape."

• Can encompass the MNL in the CL model. Suppose we have a MNL model with covariates $\mathbf{w}_i$ and parameters $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots, \boldsymbol{\delta}_J$. Let $d1, d2, \ldots, dJ$ be dummy variables for all but the zero alternative. Define $\mathbf{x}_{ij} = (d1_j \mathbf{w}_i, d2_j \mathbf{w}_i, \ldots, dJ_j \mathbf{w}_i)$ and $\boldsymbol{\beta} = (\boldsymbol{\delta}_1', \boldsymbol{\delta}_2', \ldots, \boldsymbol{\delta}_J')'$.

• So the focus is often on CL model.

• In many applications, allow for choice-specific and individual-specific covariates:

$$y_{ij}^* = \mathbf{z}_{ij}\boldsymbol{\gamma} + \mathbf{w}_i\boldsymbol{\delta}_j + a_{ij},\ j = 0, 1, \ldots, J$$

with $\boldsymbol{\delta}_0 = \mathbf{0}$.

• A key restriction of CL model is independence from irrelevant alternatives (IIA), which for the pure CL model follows from

$$P(y = j | y = j \text{ or } y = h) = \frac{\exp(\mathbf{x}_{ij}\boldsymbol{\beta})}{[\exp(\mathbf{x}_{ij}\boldsymbol{\beta}) + \exp(\mathbf{x}_{ih}\boldsymbol{\beta})]}$$

• This means that the probability of selecting between two alternatives given only those two choices does not depend on characteristics of other choices – that is, $\mathbf{x}_{im}$ for $m \notin \{j, k\}$ – do not appear.

- Can have unattractive implications for the probabilities when alternatives are similar, and for predicting substitution patterns when new alternatives are introduced or old choices are taken away.
- Another way to characterize the problem: in

$$y_{ij}^{*} = \mathbf{x}_{ij}\boldsymbol{\beta} + a_{ij},\ j = 0, \ldots, J,$$

the $a_{ij}, j = 0, 1, \ldots, J$, are assumed to be independent. This is an unrealistic assumption when some choices are similar.

- See Imbens' "Discrete Choice" lecture from NBER Summer Course.

Three restaurants in Berkeley, Chez Panisse (C), Lalime's (L), and the Bongo Burger (B).

- Suppose the two characteristics of the restaurants are price, with

$$P_C = 95,\ P_L = 80, \text{ and } P_B = 5,$$

and quality, with

$$Q_C = 10,\ Q_L = 9, \text{ and } Q_B = 2$$

- Utility is given by

$$y_{ij}^* = -.2P_j + 2Q_j + a_{ij}$$

• If we compute the choice probabilties – which can be thought of the market shares – they are roughly

$$S_C = .09,\ S_L = .24, \text{ and } S_B = .67$$

For example,

$$S_C = \frac{\exp(-.2 \cdot 95 + 2 \cdot 10)}{[\exp(-.2 \cdot 95 + 2 \cdot 10) + \exp(-.2 \cdot 80 + 2 \cdot 9) + \exp(-.2 \cdot 5 + 2 \cdot 2)]}$$

(Note: In this case, there is no normalization of setting P_j and Q_j for one of the choices equal to zero.)

• Now suppose Lalime's goes out of business. The new shares for Chez Panisse and Bongo Burger predicted by the CL model are

$$P(y = C|y = C \text{ or } B) \approx \frac{.09}{.09 + .67} \approx .12$$
$$P(y = B|y = C \text{ or } B) = \frac{.67}{.09 + .67} \approx .88$$

In other words, C gets about $(.09/.76)(.24) \approx .03$ of B's share and C gets $(.67/.76)(.24) \approx .21$.

• Seems much more likely that most of B's customers will patronize restaurant A, so the shares should be closer to $.33$ and $.67$ (but might be, say, $.30$ and $.70$).

• Three popular ways to relax IIA.

1. Allow correlation among the a_{ij}. Usually done by specifying multivariate normal. That is, assume $\mathbf{a}_i$ has a multivariate normal distribution (with unit variances) and an unrestricted correlation matrix. This leads to *multinomial probit* (which is better called *conditional probit*, in the spirit of the probabilistic choice framework).

• Mutinomial probit is computationally very difficult, although simulation methods and fast computers help. More importantly, it is not clear it does what we want. If we only ever observe a single choice for each unit, difficult to estimate many correlation parameters when the choice set is large.

2. Nested logit. Suppose we can group alternatives into S groups of "similar" alternatives. Let there be G_s alternatives in subgroup s, $s = 1, \ldots, S$. Now specify a nested structure

$$P(y \in G_s|\mathbf{x}) = \frac{\left\{\alpha_s \left[\sum_{j \in G_s} \exp(\rho_s^{-1}\mathbf{x}_j\boldsymbol{\beta})\right]^{\rho_s}\right\}}{\sum_{r=1}^{S} \alpha_r \left[\sum_{j \in G_r} \exp(\rho_r^{-1}\mathbf{x}_j\boldsymbol{\beta})\right]^{\rho_r}}$$

$$P(y = j|y \in G_s, \mathbf{x}) = \frac{\exp(\rho_s^{-1}\mathbf{x}_j\boldsymbol{\beta})}{\left[\sum_{h \in G_s} \exp(\rho_s^{-1}\mathbf{x}_h\boldsymbol{\beta})\right]}$$

- Need a normalization, usually $\alpha_1 = 1$. Get standard CL model by $\rho_s = 1$, all s.

- Important Issue: How can the nesting structure be chosen? Also, gets more complicated with more than one level of nesting.
- Structure does lead to a simple two-step estimation method. Let $\boldsymbol{\lambda}_s = \rho_s^{-1}\boldsymbol{\beta}$, $s = 1, \ldots, s$. These can be easily estimated by applying conditional logit within each group.
- Then estimate the α_s and ρ_s by maximizing

$$\sum_{i=1}^{N}\sum_{s=1}^{S} 1[y_i \in G_s] \log[q_s(\mathbf{x}_i; \hat{\boldsymbol{\lambda}}, \boldsymbol{\alpha}, \boldsymbol{\rho})]$$

where $q_s(\mathbf{x}; \boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\rho})$ is $P(y \in G_s|\mathbf{x})$.

- Can easily bootstrap the standard errors and inference for the two-step estimation method.
- Problem with method is that, by specifying the groups, we are assuming the extreme value errors within each group are independent. Results can be sensitive to those choices.

3. An approach that fits well in the utility maximization framework is random coefficient models. So consider models of the form

$$\begin{aligned} y_{ij}^{*} &= \mathbf{x}_{ij}\mathbf{b}_i + a_{ij},\ j = 0, \ldots, J \\ &= \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{x}_{ij}\mathbf{d}_i + a_{ij} \\ &\equiv \mathbf{x}_{ij}\boldsymbol{\beta} + u_{ij} \end{aligned}$$

where $u_{ij} = \mathbf{x}_{ij}\mathbf{d}_i + a_{ij}$.

- Does not require us to group ahead of time, as in nested logit.
- Even thought the a_{ij} are assumed to be independent across j – usually with indentical extreme value distributions – the u_{ij} are correlated through $\mathbf{d}_i$, and the correlation depends on $\mathbf{x}_{ij}$.

- If the intercept in $\mathbf{b}_i$ is the only heterogeneous parameter, can write $\mathbf{x}_{ij}\mathbf{d}_i = c_i$ with $E(c_i) = 0$, which gives a kind of random effects structure across choices:

$$y_{ij}^* = \mathbf{x}_{ij}\boldsymbol{\beta} + c_i + a_{ij}$$

- The presence of c_i breaks the IIA property conditional on $\mathbf{x}_i$.

- In the general model $y_{ij}^{*} = \mathbf{x}_{ij}\mathbf{b}_i + a_{ij}$, often assume that, conditional on $\mathbf{b}_i$, the model is conditional logit. Then specify a distribution for $\mathbf{b}_i$, such as assume a finite number of types. Or, use a continuous distribution, such as multivariate normal. Can even allow $\mathbf{b}_i$ to depend on observed individual-specific characteristics, $\mathbf{w}_i$.
- Estimation is computationally very intensive, and simulation methods of estimation are often used.

- Extensions of conditional logit, and its extensions, to allow for endogenous characteristics is possible but can be very difficult. Petrin and Train (2010, *Journal of Marketing Research*) show how simple control function methods can be used for continuous endogenous explanatory variables.
- Panel data harder to handle, too, but the CRE approach of Chamberlain can be used. As in the Petrin and Train approach, easiest to assume that the model conditional on observables follows a MNL functional form, or some other convenient model.

4. ORDERED RESPONSE MODELS

- Here we discuss cases where the ordered response, y, is the variable we wish to explain. A setting with a similar statistical structure, but a different interpretation, is *interval regression*, which is a data censoring problem that arises from observing an underlying continuous response only in cells. Here, y is the response of interest.
- When the response probabilities are of interest, we can take the outcomes to be $\{0, 1, \ldots, J\}$ without loss of generality.

• Underlying *ordered probit* is a latent variable model that looks just like binary response:

$$y^* = \mathbf{x}\boldsymbol{\beta} + e,\ e|\mathbf{x} \sim Normal(0,1)$$

where, for reasons to be seen, **x** does not include a constant. Let $\alpha_1 < \alpha_2 < \ldots < \alpha_J$ be J unknown *cut points*. These are parameters that we estimate these along with $\boldsymbol{\beta}$.

- Assume

$$
\begin{aligned}
y &= 0 \text{ if } y^* \leq \alpha_1 \\
y &= 1 \text{ if } \alpha_1 < y^* \leq \alpha_2 \\
&\vdots \\
y &= J-1 \text{ if } \alpha_{J-1} < y^* \leq \alpha_J \\
y &= J \text{ if } y^* > \alpha_J .
\end{aligned}
$$

- The response probabilities are easy to obtain:

$$
\begin{aligned}
P(y = 0|\mathbf{x}) &= P(\mathbf{x}\boldsymbol{\beta} + e \leq \alpha_1|\mathbf{x}) = \Phi(\alpha_1 - \mathbf{x}\boldsymbol{\beta}) \\
P(y = 1|\mathbf{x}) &= P(\alpha_1 < \mathbf{x}\boldsymbol{\beta} + e \leq \alpha_2|\mathbf{x}) = \Phi(\alpha_2 - \mathbf{x}\boldsymbol{\beta}) - \Phi(\alpha_1 - \mathbf{x}\boldsymbol{\beta}) \\
&\vdots \\
P(y = J-1|\mathbf{x}) &= \Phi(\alpha_J - \mathbf{x}\boldsymbol{\beta}) - \Phi(\alpha_{J-1} - \mathbf{x}\boldsymbol{\beta}) \\
P(y = J|\mathbf{x}) &= 1 - \Phi(\alpha_J - \mathbf{x}\boldsymbol{\beta})
\end{aligned}
$$

- Of course, when we add them all up, we get one.

- For random draw i the log likelihood is

$$\begin{aligned}\ell_i(\boldsymbol{\alpha},\boldsymbol{\beta}) &= 1[y_i = 0]\log[\Phi(\alpha_1 - \mathbf{x}_i\boldsymbol{\beta})] \\ &+ 1[y_i = 1]\log[\Phi(\alpha_2 - \mathbf{x}_i\boldsymbol{\beta}) - \Phi(\alpha_1 - \mathbf{x}_i\boldsymbol{\beta})] \\ &+ \ldots + 1[y_i = J]\log[1 - \Phi(\alpha_J - \mathbf{x}_i\boldsymbol{\beta})]\end{aligned}$$

- MLE is well behaved: computation is usually straightforward, inference is standard.
- When $J = 1$, $P(y = 0|\mathbf{x}) = \Phi(\alpha_1 - \mathbf{x}\boldsymbol{\beta}) = 1 - \Phi(\mathbf{x}\boldsymbol{\beta} - \alpha_1)$, $P(y = 1|\mathbf{x}) = \Phi(\mathbf{x}\boldsymbol{\beta} - \alpha_1)$, and so $-\alpha_1$ plays the role of the intercept in standard probit.

- For *ordered logit*, replace $\Phi(\cdot)$ with $\Lambda(\cdot)$.
- Interpreting coefficients requires some care.

$$\frac{\partial p_0(\mathbf{x})}{\partial x_k} = -\beta_k \phi(\alpha_1 - \mathbf{x}\boldsymbol{\beta}), \quad \frac{\partial p_J(\mathbf{x})}{\partial x_k} = \beta_k \phi(\alpha_J - \mathbf{x}\boldsymbol{\beta})$$

$$\frac{\partial p_j(\mathbf{x})}{\partial x_k} = \beta_k [\phi(\alpha_{j-1} - \mathbf{x}\boldsymbol{\beta}) - \phi(\alpha_j - \mathbf{x}\boldsymbol{\beta})]$$

- The sign of $\partial p_j(\mathbf{x})/\partial x_k$ is ambiguous. It depends on $|\alpha_{j-1} - \mathbf{x}\boldsymbol{\beta}|$ versus $|\alpha_j - \mathbf{x}\boldsymbol{\beta}|$ (remember, $\phi(\cdot)$ is symmetric about zero).
- As in other nonlinear models, can compute PEAs or APEs. Bootstrap standard errors.

• For ordered logit or probit,

$$P(y \leq j|\mathbf{x}) = P(y^* \leq \alpha_j|\mathbf{x}) = G(\alpha_j - \mathbf{x}\boldsymbol{\beta}), \quad j = 0, 1, \ldots, J-1,$$

where $G(\cdot) = \Phi(\cdot)$ or $G(\cdot) = \Lambda(\cdot)$. Probabilities differ across j only because of the cut parameters, α_j. In effect, an intercept shift inside the nonlinear cdf determines the differences in probabilities. Sometimes called the *parallel assumption*.

• Some have proposed replacing $\boldsymbol{\beta}$ with $\boldsymbol{\beta}_j$, which means estimating a sequence of binary responses: $P(y \leq j|\mathbf{x}) = G(\alpha_j - \mathbf{x}\boldsymbol{\beta}_j)$, $P(y > j|\mathbf{x}) = 1 - G(\alpha_j - \mathbf{x}\boldsymbol{\beta}_j)$. But the resulting estimates of $P(y \leq j|\mathbf{x})$ need not increase in j.

• Can construct a likelihood ratio test (say) comparing OP or OL against the more general model. If we reject the OP or OL models against the general alternative, what would we do? Is a statistical rejection important for computing partial effects?

• The OP and OL models allow us to sign partial effects on $P(y > j|\mathbf{x})$: for a continuous variable x_h,

$$\frac{\partial P(y > j|\mathbf{x})}{\partial x_h} = \beta_h g(\alpha_j - \mathbf{x}\boldsymbol{\beta}),$$

where $g(\cdot)$ is the density associated with $G(\cdot)$. If $\beta_h > 0$, an increase in x_h increases the probability that y is greather than any value j.

- It is sometimes useful to compute the conditional mean, and partial effects on the mean, especially if the the ordered variable can be (roughly) assigned magnitudes. The estimates of the probabilities in each category will be the same provided the order is preserved.
- As an example, suppose on a survey about retirement investments, people are asked whether their assets are in “all bonds,” “mostly bonds,” “mix of stocks and bonds,” “mostly stocks,” and “all stocks.” We could just estimate an ordered probit or logit with $J = 4$.

• But we also might assign approximate numerical values for the fraction held in stocks, for example

$$m_0 = 0, m_1 = .2, m_2 = .5, m_3 = .8, m_4 = 1$$

• Using these values in ordered probit or logit has no effect on the estimates of $\boldsymbol{\beta}$ or the α_j; that is the nature of an ordered response.

• But, after estimation, we might compute an estimate of $E(y|\mathbf{x})$ because its magnitude has some meaning.

• Generally, let $\{m_0, m_1, \ldots, m_J\}$ be the J values assigned to y, where $m_{j-1} < m_j$. Then, for ordered probit,

$$\begin{aligned} E(y|\mathbf{x}) &= m_0 P(y = m_0|\mathbf{x}) + m_1 P(y = m_1|\mathbf{x}) + \ldots + m_J P(y = m_J|\mathbf{x}) \\ &= m_0 \Phi(\alpha_1 - \mathbf{x}_i\boldsymbol{\beta}) + m_1[\Phi(\alpha_2 - \mathbf{x}\boldsymbol{\beta}) - \Phi(\alpha_1 - \mathbf{x}\boldsymbol{\beta})] \\ &\quad + \ldots + m_J[1 - \Phi(\alpha_J - \mathbf{x}\boldsymbol{\beta})] \\ &= (m_0 - m_1)\Phi(\alpha_1 - \mathbf{x}\boldsymbol{\beta}) + (m_1 - m_2)\Phi(\alpha_2 - \mathbf{x}\boldsymbol{\beta}) \\ &\quad + \ldots + (m_{J-1} - m_J)\Phi(\alpha_J - \mathbf{x}\boldsymbol{\beta}) + m_J \end{aligned}$$

• It is easy to see that the signs of the partial effects on $E(y|\mathbf{x})$ are unambiguously the same sign as a coefficient:

$$\frac{\partial E(y|\mathbf{x})}{\partial x_k} = \beta_k[(m_1 - m_0)\phi(\alpha_1 - \mathbf{x}\boldsymbol{\beta}) + (m_2 - m_1)\phi(\alpha_2 - \mathbf{x}\boldsymbol{\beta}) + \ldots + (m_J - m_{J-1})\phi(\alpha_J - \mathbf{x}\boldsymbol{\beta})]$$

and each term in $[\cdot]$ is positive because $m_j > m_{j-1}$ and $\phi(\cdot) > 0$.

• The estimated partial effects, when averaged across $\mathbf{x}_i$, can be compared with OLS estimates of a linear model. The linear model estimates make some sense when y is assigned one of the m_j values.

• One way to extend the basic model that preserves ordering of probabilities is to allow heteroskedasticity in the latent variable model, as in binary case:

$$e|\mathbf{x} \sim Normal(0, \exp(2\mathbf{x}_1\boldsymbol{\delta}))$$

where $\mathbf{x}_1$ can be a subset of $\mathbf{x}$.

• Can use the Rivers-Vuong control function approach to allow endogeneity when y_2 is continuous.

$$y_1^* = \mathbf{z}_1\boldsymbol{\delta}_1 + \gamma_1 y_2 + u_1$$
$$y_2 = \mathbf{z}\boldsymbol{\delta}_2 + v_2,$$

where (u_1, v_2) is independent of $\mathbf{z}$ and jointly normally distributed. (As in the binary case, we can relax these assumptions a bit.)

- Again, $\mathbf{z}_1$ does not contain an intercept. Instead, there a cut points, α_j, $j = 1, \ldots, J$. We define the observed ordered response, y_1, in terms of the latent response, y_1^*.
- Write $u_1 = \theta_1 v_2 + e_1$ and plug in:

$$y_1^* = \mathbf{z}_1 \boldsymbol{\delta}_1 + \gamma_1 y_2 + \theta_1 v_2 + e_1,$$

where $\theta_1 = \eta_1 / \tau_2^2$, $\eta_1 = Cov(v_2, u_1)$, $\tau_2^2 = Var(v_2)$, $e_1 | \mathbf{z}, v_2$ $\sim Normal(0, 1 - \rho_1^2)$, and $\rho_1 = \theta_1^2 \tau_2^2 = \eta_1^2 / \tau_2^2$.

• So (1) Obtain the OLS residuals, $\hat{v}_{i2}$, from the first-stage regression y_{i2} on $\mathbf{z}_i$, $i = 1, \ldots, N$. (2) Run ordered probit of y_{i1} on $\mathbf{z}_{i1}, y_{i2}$, and $\hat{v}_{i2}$ in a second stage. Consistently estimate the scaled coefficients $\boldsymbol{\delta}_{\rho 1} \equiv \boldsymbol{\delta}_1/(1-\rho_1^2)^{1/2}$, $\gamma_{\rho 1} \equiv \gamma_1/(1-\rho_1^2)^{1/2}$, $\theta_{\rho 1} \equiv \theta_1/(1-\rho_1^2)^{1/2}$, and $\alpha_{\rho j} = \alpha_j/(1-\rho_1^2)^{1/2}$.

• A simple test of the null hypothesis that y_2 is exogenous is just the standard t statistic on $\hat{v}_{i2}$.

• Can estimate the original parameters by dividing each of the scaled coefficients by $(1 + \hat{\theta}_{\rho 1}^2 \hat{\tau}_2^2)^{1/2}$.

• As usual, can obtain the average structural function by averaging out the $\hat{v}_{i2}$ from the equation with scaled coefficients. For example, with $0 < j < J$,

$$\widehat{ASF}_j(\mathbf{x}_1) = N^{-1}\sum_{i=1}^{N}[\Phi(\hat{\alpha}_{\rho 2} - \mathbf{x}_1\hat{\boldsymbol{\beta}}_{\rho_1} - \hat{\theta}_{\rho 1}\hat{v}_{i2}) - \Phi(\hat{\alpha}_{\rho 1} - \mathbf{x}_1\hat{\boldsymbol{\beta}}_{\rho_1} - \hat{\theta}_{\rho 1}\hat{v}_{i2})]$$

where $\mathbf{x}_1$ can be any function of $(\mathbf{z}_1, y_2)$.

• As always, partial effects are obtained by taking derivatives or differences.

• Bootstrapping is a natural way to obtain standard errors; the delta method can also be used.

• Panel data versions of ordered probit are easily specified and estimated. We add unobserved heterogeneity to the model and subsume its mean into the cut points.

$$
\begin{aligned}
y_{it}^{*} &= \mathbf{x}_{it}\boldsymbol{\beta} + c_i + e_{it} \\
e_{it} | \mathbf{x}_i, c_i &\sim Normal(0,1)
\end{aligned}
$$

$$
\begin{aligned}
y_{it} &= 0 \text{ if } y_{it}^{*} \leq \alpha_1 \\
y_{it} &= 1 \text{ if } \alpha_1 < y_{it}^{*} \leq \alpha_2 \\
&\vdots \\
y_{it} &= J \text{ if } y_{it}^{*} > \alpha_J .
\end{aligned}
$$

- Notice that the assumption on e_{it} in incorporates strict exogeneity conditonal on c_i.
- Again, a convenient assumption is

$$c_i|\mathbf{x}_i \sim Normal(\psi + \bar{\mathbf{x}}_i\boldsymbol{\xi}, \sigma_a^2)$$

- Under these assumptions, we can estimate the coefficients scaled by $(1 + \sigma_a^2)^{-1/2}$ because, for example, for $0 < j < J$,

$$P(y_{it} = j|\mathbf{x}_i) = \Phi(\alpha_{a,j+1} - \mathbf{x}_{it}\boldsymbol{\beta}_a - \bar{\mathbf{x}}_i\boldsymbol{\xi}_a) - \Phi(\alpha_a - \mathbf{x}_{it}\boldsymbol{\beta}_a - \bar{\mathbf{x}}_i\boldsymbol{\xi}_a).$$

- The APEs can be obtained from

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^{N}[\Phi(\hat{\alpha}_{a,j+1} - \mathbf{x}_t\hat{\boldsymbol{\beta}}_a - \bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}_a) - \Phi(\hat{\alpha}_{aj} - \mathbf{x}_t\hat{\boldsymbol{\beta}}_a - \bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}_a)]$$

- Use the panel bootstrap for standard errors.
- If we add conditional independence, we can estimate the orginal parameters and σ_a^2 separately. Called (correlated) random effects ordered probit.

• We can extend the basic dynamic probit model to the ordered case, too. Because y_{it} is an ordered response, a dynamic model should allow the current probabilities to depend on the past in a flexible way. Let $w_{itj} = 1[y_{it} = j], j = 1, \ldots, J$, and $\mathbf{w}_{it} = (w_{it1}, ..., w_{itJ})$ and write the latent variable model as

$$y_{it}^{*} = \mathbf{z}_{it}\boldsymbol{\delta} + \mathbf{w}_{i,t-1}\boldsymbol{\rho} + c_i + u_{it}, \quad t = 1, \ldots, T$$

where y_{it} is defined as before.

• We assume the dynamics are correctly specified, which means that

$$D(u_{it}|\mathbf{z}_i, y_{i,t-1}, \ldots, y_{i0}, c_i) = D(u_{it}) = Normal(0, 1).$$

where $\mathbf{z}_i = (\mathbf{z}_{i1}, \mathbf{z}_{i2}, \ldots, \mathbf{z}_{iT})$.

- To account for the initial conditions problem, the unobserved effect, c_i, is modeled as $c_i = \psi + \mathbf{w}_{i0}\boldsymbol{\eta} + \mathbf{z}_i\boldsymbol{\xi} + a_i$, where $\mathbf{w}_{i0}$ is the J-vector of initial conditions, w_{i0j}.
- Assume

$$a_i|\mathbf{z}_i, \mathbf{w}_{i0} \sim Normal(0, \sigma_a^2).$$

- We can apply random effects ordered probit to the equation

$$y_{it}^* = \mathbf{z}_{it}\boldsymbol{\delta} + \mathbf{w}_{i,t-1}\boldsymbol{\rho} + \mathbf{w}_{i0}\boldsymbol{\eta} + \mathbf{z}_i\boldsymbol{\xi} + a_i + u_{it}, \quad t = 1, \ldots, T,$$

where we absorb the intercept into the cut parameters, α_j.

• Any software that estimates random effects ordered probit models can be applied directly to estimate all parameters, including σ_a^2; we simply specify the explanatory variables at time t as $(\mathbf{z}_{it}, \mathbf{w}_{i,t-1}, \mathbf{w}_{i0}, \mathbf{z}_i)$. (Pooled ordered probit does *not* consistently estimate any interesting parameters.)

• Average partial effects are easily computed. Not surprisingly, the APEs depend on the coefficients multiplied by $(1 + \hat{\sigma}_a^2)^{-1/2}$; see Wooldridge (2005b, Journal of Applied Econometrics).

• Using the same approach for dynamic probit, the mean and variance of c_i can also be estimated.