**Ec331: Research in Applied Economics**
**Spring term, 2022**

# Wk. 3 class: Discrete Choice, Ordered Choice

# Resources

Stata search function and pdf documentation

Ec226 Lecture notes

# Resources

Wooldridge "Introductory Econometrics – A Modern Approach", Third edition

- – Sections 7.5, 8.5 (LPM)
- – Ch. 17 Logit, Probit, Tobit, Count Poisson, Censored

Dougherty (2011) "An introduction to Econometrics", Ch. 11

C. Baum (2006). "An introduction to modern econometrics using stata", Ch. 10

Cameron and Trivedi (2009). "Microeconometrics Using Stata", Ch. 14, 15.

Freese and Long (2006). "Regression Models for Categorical Dependent Variables Using Stata".

Mitchell (2012). "Interpreting and Visualising Regression Models Using Stata".

K. Train (2009). "Discrete Choice Methods with Simulation"

Greene and Hensher (2010). "Modeling Ordered Choices: A Primer"

# Discrete Choice, Ordered Choice

Today:

(1) Properties of Discrete Choice Models - Train Ch.2

(2) A Model for Ordered Choices - Greene and Hensher Ch. 3

(3) Estimation, inference and analysis using the Ordered Choice Model - G/H Ch. 5

# (1) Properties of Discrete Choice Models

Common features of all discrete choice models: the choice set, and choice probabilities - which can be derived from utility-maximising behaviour (with implications for specification and normalisation).

The choice set: (i) mutually exclusive
(ii) exhaustive
(iii) finite set of alternatives

Mutual exclusivity and exhaustiveness are not restrictive conditions, and specification is governed by research goals and data availability.

Having a finite number of alternatives is restrictive, and the defining characteristic of discrete choice models.

# Random Utility Models (RUMs)

Decision maker: $n$

Utility from alternative $j$ : $U_{nj}, j = 1, \ldots, J$

Behavioural model:

choose $i$ if and only if $U_{ni} > U_{nj} \;\; \forall j \neq i$

Researcher observes:

(i) attributes of alternatives, $x_{nj}$

(ii) attributes of decision maker, $s_n$

Representative Utility: $V_{nj} = V(x_{nj}, s_n) \;\; \forall j$

Utility: $\mathrm{U}_{nj} = V_{nj} + \varepsilon_{nj}$

Joint density of $\varepsilon_n$ denoted $f(\varepsilon_n)$

# Derivation of Choice Probabilities

Probability decision maker chooses alternative $i$ :

$$
\begin{aligned}
P_{ni} &= \text{Prob}(U_{ni} > U_{nj} \;\; \forall j \neq i) \\
&= \text{Prob}(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \;\; \forall j \neq i) \\
&= \text{Prob}(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \;\; \forall j \neq i)
\end{aligned}
$$

and using density $f(\varepsilon_n)$

$$
= \int_{\varepsilon} I(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj} \;\; \forall j \neq i) f(\varepsilon_n) d\varepsilon_n
$$

which is a multidimensional integral over the density of the unobserved element of utility, with different discrete choice models obtained from different specifications of this density (e.g. logit derived by assuming iid extreme value distribution, probit by assuming multivariate normal)

# Interpretation of choice probabilities

Interpretation of the density of the unobservables determines interpretation of choice probabilities:

(i) If this density is the distribution of the unobserved portion of utility within a population of people who face the same observed portion of utility, the choice probability is the share of people who choose that alternative.

(ii) If the density is considered as the researcher's subjective probability that unobserved utility will take given values for an individual, the choice probability is the probability that individual will choose that alternative, given the researcher's ideas about unobservables.

(iii) If the density represents factors quixotic to the decision maker him/herself, the choice probability is the probability that the quixotic factors will induce choice of that alternative, given observed factors.

# Further issues

Identification of Choice Models

- only differences in utility matter, so only parameters that capture differences across alternatives are identified (and can be estimated): hence normalisation of absolute level of constants (standard: set one to 0), and of the impact of attributes of the decision maker (again, standard: set one parameter to 0).

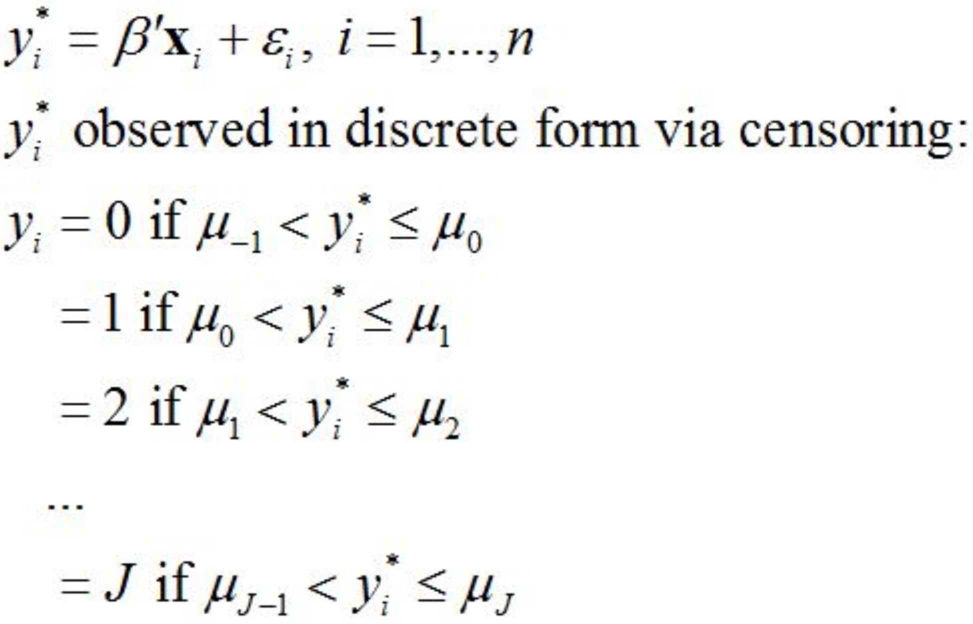
- overall scale of utility is irrelevant, so researcher must normalise the scale (standard: normalise variance of error terms).

Aggregation

- average probabilities not equivalent to probabilities at average characteristics

# (2) A model for ordered choices

A latent regression model (or underlying RUM) for a continuous measure:



$$y_i^* = \beta'\mathbf{x}_i + \varepsilon_i,\ i = 1,\ldots,n$$

$y_i^*$ observed in discrete form via censoring:

$$\begin{aligned} y_i &= 0 \text{ if } \mu_{-1} < y_i^* \le \mu_0 \\ &= 1 \text{ if } \mu_0 < y_i^* \le \mu_1 \\ &= 2 \text{ if } \mu_1 < y_i^* \le \mu_2 \\ &\ldots \\ &= J \text{ if } \mu_{J-1} < y_i^* \le \mu_J \end{aligned}$$

Note: strong assumptions - neither coefficients nor thresholds differ across individuals

# Probability of observed outcome

$$\text{Prob}[y_i = j \mid \mathbf{x}_i] = \text{Prob}[\varepsilon_i \le \mu_j - \beta'\mathbf{x}_i] - \text{Prob}[\varepsilon_i \le \mu_{j-1} - \beta'\mathbf{x}_i],\ j = 0,1,\ldots,J$$
$$= [F(\mu_j - \beta'\mathbf{x}_i) - F(\mu_{j-1} - \beta'\mathbf{x}_i)] > 0,\ j = 0,1,\ldots,J$$

Note: in general, no obvious regression (conditional mean) relationship between observed dependent variable and regressors

Normalisations required to identify model parameters:

$$\mu_j > \mu_{j-1},\ \mu_{-1} = -\infty,\ \mu_J = \infty$$

$$Var[\varepsilon_i] = \sigma_\varepsilon^2 = \bar{\sigma}^2 \quad (\text{Probit, } = 1;\ \text{Logit, } \frac{\pi^2}{3})$$

$$\text{Ass. constant in } \mathbf{x}_i :\ \mu_0 = 0$$

# Ordered probit, ordered logit

Model completed by distributional assumptions over the unobservables:

- continuous random disturbance with conventional CDF, $F(.)$
- disturbances independent from (i.e. exogeneity of) **x**
- standard normal distribution: ordered probit model
- standardized logistic distribution: ordered logit model

Applications well divided between probit and logit; "a compelling case for one distribution over the other remains to be put forth... the motivation for [other distributional choices] is even less persuasive than that for a preference for probits over logits". For recent arguments that preference can be based on a fit measure, see Hahn and Soyer (2009).

Estimation of parameters via maximum likelihood estimation, subject to constraints. Can also collapse categories (even up to binary choice) though sacrificing information may provide a less efficient estimator - see Murad et al (2003) for analysis

# (3) Estimation, inference and analysis using the ordered choice model

Estimation results imply:

$$
\begin{aligned}
&y^{*} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k + \varepsilon \\
&y = 0 \text{ if } \mathrm{y}^{*} \leq 0 \\
&y = 1 \text{ if } 0 < y^{*} \leq \hat{\mu}_1 \\
&\ldots \\
&y = J \text{ if } \hat{\mu}_{J-1} < y^{*} \leq \hat{\mu}_J
\end{aligned}
$$

Note: sample proportions need not provide a histogram of the underlying distribution

# Interpretation of the model - partial effects

Interpretation more complicated than ordinary regression setting; no natural conditional mean function, $E[y|\mathbf{x}]$, to analyse, since the outcome variable is merely a label for ordered, nonquantitative outcomes.

To interpret, typically refer to the probabilities, with partial effects:

$$\delta_j(\mathbf{x}_i) = \frac{\partial \text{Prob}(y = j \mid \mathbf{x}_i)}{\partial \mathbf{x}_i} = [f(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i) - f(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i)]\boldsymbol{\beta}$$

$$\begin{aligned}\Delta_j(D) = {} & [F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i + \gamma) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i + \gamma)] \\ & - [F(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i) - F(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i)]\end{aligned}$$

for continuous and dummy variables respectively (gamma being the dummy variable coefficient). Note: neither sign nor magnitude of the coefficient is directly informative. Effect dependent on all parameters, data, and probability (cell) of interest.

# Cumulative partial effects, scaled coefficients

Might be interested in cumulative values of partial effects:

$$\frac{\partial \text{Prob}(y \leq j \mid \mathbf{x}_i)}{\partial \mathbf{x}_i} = \sum_{m=0}^{j} [f(\mu_{m-1} - \boldsymbol{\beta}'\mathbf{x}_i) - f(\mu_m - \boldsymbol{\beta}'\mathbf{x}_i)]\boldsymbol{\beta}$$

$$= -f(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i)\boldsymbol{\beta}$$

Note: difference in coefficients from probit and logit (mostly) reflects inherent difference in scaling of the underlying variable, and highlights risks in naive direct interpretation and comparison. This problem is eliminated in comparisons of partial effects.

# Nonlinearities in the variables

In the computation of partial effects, it is assumed that the independent variables can vary independently; thus, the computation of interactions and nonlinearities becomes problematic (see Ai and Norton, 2003, for extensive analysis).

e.g. if Ed, Ed squared and Ed*Age terms are included:

$$\delta_j(Ed) = \frac{\partial \text{Prob}(y = j \mid \mathbf{x}_i)}{\partial Ed} = [f(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i) - f(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i)]$$

$$\times(\beta_{Ed} + 2\beta_{EdSq}Ed + \beta_{EdAge}Age)$$

computed using reported results:

$$= \frac{\partial \text{Prob}(y = j \mid \mathbf{x}_i)}{\partial Ed} + (2Ed)\frac{\partial \text{Prob}(y = j \mid \mathbf{x}_i)}{\partial EdSq} + Age\frac{\partial \text{Prob}(y = j \mid \mathbf{x}_i)}{\partial EdAge}$$

# Average partial effects

Generally, indicative partial effects computed by inserting the sample means of the regressors, i.e.

$$\frac{\partial \text{Prob}(y = j \mid \bar{\mathbf{x}})}{\partial Ed} = [f(\mu_{j-1} - \boldsymbol{\beta}'\bar{\mathbf{x}}) - f(\mu_j - \boldsymbol{\beta}'\bar{\mathbf{x}})]\beta_{Ed}$$

but may also consider the average partial effect:

$$APE_j(Ed) = \frac{1}{n}\sum[f(\mu_{j-1} - \boldsymbol{\beta}'\mathbf{x}_i) - f(\mu_j - \boldsymbol{\beta}'\mathbf{x}_i)]\beta_{Ed}$$

In practise, give very similar results unless sample size is very small or data highly skewed and affected by outliers.

# Interpreting the threshold parameters, and the underlying regression

In most treatments, threshold parameters considered necessary but of "no intrinsic interest", but Daykin and Moffatt (2002, p. 162) argue that (in the absence of other information) the cut points associated with attitudinal scales reveal information about the preferences of respondents:

(i) tightly bunched cut points in the middle of the distribution implying most people are in strong agreement or disagreement

(ii) widely dispersed cut points implying less desire to report strong views

(though this is revealed directly from response distribution itself)

The model does imply partial changes for the latent regressand:

$$\frac{\partial E[y^* \mid \mathbf{x}]}{\partial \mathbf{x}} = \boldsymbol{\beta}$$

but the scaling of the dependent variable has been lost due to censoring, so McKelvey and Zavoina (1975) suggest interpretation req. standardised coefficients (multiplying by the s.d. of the regressor over the s.d. of the latent variable), so changes are in standard deviation units (see Greene and Hensher, p149)

# Inference

Given assumptions underlying MLE are met, inference can be based on usual methods.

Inference about a single coefficient based on the standard "z" test.

Inference about the threshold parameters meaningless, and generally not carried out.

Tests regarding more than one coefficient can be carried out using a Wald test ("test" command in Stata) or likelihood ratio test (twice the difference between the log likelihoods for the null and alternative, and asymptotically equivalent to the Wald test). Greene and Hensher prefer the latter in finite samples, since it uses more information (being based on both models).

"Test of the model" in spirit of overall F stat in linear regression is LR test against the null that the model contains only a constant term and threshold parameters (routinely reported in software, inc. Stata).

# Testing for structural change or homogeneity of strata; robust covariance matrix estimation; partial effects

Whether the same model describes two (or more) groups, tested via likelihood ratio:

$$LR = 2[\sum_{g=groups} \log L_g - \log L_{pooled}]$$

with degrees of freedom equal to G-1 times no. of parameters.

Applications often compute "robust" covariance matrix; but Greene and Hensher assert that if model assumptions are correct, this estimator is the same as the conventional estimators, and if incorrect, the estimator of parameters is inconsistent anyway.

Note: for ordered probit, estimation of parameters is inconsistent under: (i) omitted variables, even if orthogonal to regressors (ii) heteroscedasticity (iii) incorrect distributional assumptions (iv) endogeneity (v) omission of latent heterogeneity

Inference regarding significance of partial effects is possible, but may result in contradictions; as such, Greene and Hensher prefer inference regarding structural coefficients - see page 157.

# Prediction

Predicted probabilities may be of interest:

$$\hat{P}_j(\mathbf{x}_i) = F(\hat{\mu}_j - \hat{\boldsymbol{\beta}}'\mathbf{x}_i) - F(\hat{\mu}_{j-1} - \hat{\boldsymbol{\beta}}'\mathbf{x}_i)$$

which could be evaluated for particular observations (e.g. average characteristics), or tabulated against variables of interest.

# Measuring Fit

The search for a scalar measure of fit is "even more difficult" than for binary choice models, due to a lack of dependent variable (beyond a labelling convention) and a lack of "variation" (around the mean) to be explained. As such, caution is advised.

To assess the fit of predictions by the model to the observed data, Greene and Hensher suggest the overall model chi squared (see prev.), which is often reported as transformed in McFadden's (1974) "pseudo R squared", but important to emphasise that not a measure of model fit and not a measure of proportion of variation explained.

Long and Freese (2006) list a variety of possible measures (obtained in Stata using FitStat command). To compare models to each other, other fit measures based on the log likelihood function are often used, most commonly the Akaike Information Criterion.

Can also usefully consider "Count R squared" measures, that focus on the average number of correct predictions (e.g. assuming prediction of the most probable outcome). See Greene and Hensher p160 for further discussion.

# Perfect prediction and further issues

If a variable predicts perfectly one of the implicit dependent variables (i.e. one alternative only ever chosen when a regressor is a certain value) then impossible to fit coefficients, as corresponding threshold parameter inestimable.

Necessary to drop such observations from the sample, which Stata does automatically ("observations completely determined"). This might be a small sample problem, but it could be due to endogeneity, which brings the coefficient estimates into question. This is a similar problem to sample selection: the discarded observations are non-random.

Also:

Accommodating individual heterogeneity (esp. issue for SWB) - G&H Ch7

Parameter variation - G&H Ch8

Ordered choice modeling with panel and time series data - G&H Ch9