Ordered Choice Modelling with Discretized Continuous Dependent Variable

Ágoston Reguly

Central European University

Joint work with: Felix Chan, László Mátyás

Brown Bag Seminars

11.02.2020

The problem

Let us assume that we have a simple linear regression model such as

$$y_i = x_i'\beta + \epsilon_i,$$

where

- β is the parameter vector of interest
- ϵ_i is an idiosyncratic disturbance term
- with $E(\epsilon_i|x_i) = 0\,,\ \forall i,$

Our goal is to estimate

$$\frac{\partial \mathbb{E}\left[y|x\right]}{\partial x} = \beta,$$

The main problem is we do not observe y_i.

The problem

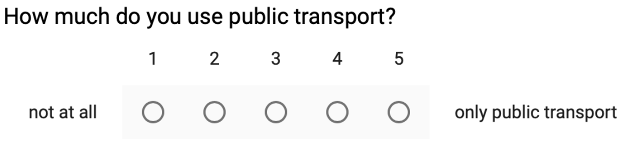
Instead we observe y_i^*, where

$$y_i^* = \begin{cases} z_1 = 1, & \text{if} \quad c_0 \leq y_i < c_1 \\ z_2 = 2, & \text{if} \quad c_1 \leq y_i < c_2 \\ z_3 = 3, & \text{if} \quad c_2 \leq y_i < c_3 \\ z_4 = 4, & \text{if} \quad c_3 \leq y_i < c_4 \\ z_5 = 5, & \text{if} \quad c_4 \leq y_i \leq c_5 \end{cases}$$

Remark: In our previous paper x was discretized, while here the dependent variable y is.

An illustrative example

For example let y_i be the commuting in percentage points in public transport. Instead y_i, we observe the individual's choice from a survey (y_i^*),



and let x be some price variable. We are then looking for the price elasticity i.e., *What is the partial effects of the price on transportation usage?*

Current practice

Current practice is to model the probability of an observation taking the arbitrarily assigned choice value m, such as

$$\Pr\left[y_i^* = m | x_i\right] .$$

Restriction: need an assumption about $F(y_i^*)$. Commonly used distributions are the probit and logit. Occasionally Gumbel I, II. and generalised extreme value distributions.

$$\Pr\left[y_i^* = m \mid x_i\right] = \left[F(c_m - \beta' x_i) - F(c_{m-1} - \beta' x_i)\right] > 0, \;\; m = 1, \ldots, M.$$

The exogeneity assumption in this case translates to $F(\epsilon_i | x_i) = F(\epsilon_i)$.

Main drawbacks of the current practices

Current models approximate *conditional probabilities*. This has some important implications:

- It predicts the probability of an observation lying in an interval, unlike the true model which refers to the conditional mean function.
- For maximum likelihood estimation, one need identification restrictions:

	β	$F(\cdot, \sigma^2)$	c_m
1.	unconstrained	fixed e.g., $\sigma^2 = 1$	one of c_m is fixed, e.g., $c_0 = 0$
2.	no intercept	fixed e.g., $\sigma^2 = 1$	unconstrained
3.	unconstrained	unconstrained	two c_m are fixed

- Partial effects in the ordered choice models are,

$$\frac{\partial \Pr\left[y_i^* = m \mid x_i\right]}{\partial x} = \left[f(c_{m-1} - \beta' x_i) - f(c_m - \beta' x_i)\right] \beta,$$

- Finally, the ML estimates are $\hat{\beta}^{ML} \neq \beta$.

Main drawbacks of the current practices

Current models approximate *conditional probabilities*. This has some important implications:

- It predicts the probability of an observation lying in an interval, unlike the true model which refers to the conditional mean function.
- For maximum likelihood estimation, one need identification restrictions:

	β	$F(\cdot, \sigma^2)$	c_m
1.	unconstrained	fixed e.g., $\sigma^2 = 1$	one of c_m is fixed, e.g., $c_0 = 0$
2.	no intercept	fixed e.g., $\sigma^2 = 1$	unconstrained
3.	unconstrained	unconstrained	two c_m are fixed

- Partial effects in the ordered choice models are,

$$\frac{\partial \Pr\left[y_i^* = m \mid x_i\right]}{\partial x} = \left[f\left(c_{m-1} - \beta' x_i\right) - f\left(c_m - \beta' x_i\right)\right] \beta,$$

- Finally, the ML estimates are $\hat{\beta}^{ML} \neq \beta$.

Main drawbacks of the current practices

Current models approximate *conditional probabilities*. This has some important implications:

- It predicts the probability of an observation lying in an interval, unlike the true model which refers to the conditional mean function.
- For maximum likelihood estimation, one need identification restrictions:

	β	$F(\cdot, \sigma^2)$	c_m
1.	unconstrained	fixed e.g., $\sigma^2 = 1$	one of c_m is fixed, e.g., $c_0 = 0$
2.	no intercept	fixed e.g., $\sigma^2 = 1$	unconstrained
3.	unconstrained	unconstrained	two c_m are fixed

- Partial effects in the ordered choice models are,

$$\frac{\partial \Pr\left[y_i^* = m \mid x_i\right]}{\partial x} = \left[f\left(c_{m-1} - \beta' x_i\right) - f\left(c_m - \beta' x_i\right)\right] \beta,$$

- Finally, the ML estimates are $\hat{\beta}^{ML} \neq \beta$.

Why $\hat{\beta}^{ML} \neq \beta$ in an ordered choice model?

- Ordered choice models gives different $\hat{\beta}^{ML}$ values, depending on the underlying distribution and on the identifying restriction. E.g. difference between ordered probit and ordered logit estimates, is $\approx 40\%$ due to $\sigma_N^2 = 1$ and $\sigma_L^2 = \pi^2/3$.
- We call the difference between β and $\hat{\beta}^{ML}$: *"distortion"*, while these estimates theoretically are not the same due to 'scaling effect'.
- Although, many people use these naive estimates and interpret them as partial effects of the conditional expectation or probability.

Example for distortions

Three choice options with $c_1 = -8, c_2 = -2, c_3 = 2, c_4 = 8$, one regressor $x_i \sim N(0,1)$ with $\beta = 0.5$, $N = 10,000$ and $1,000$ Monte Carlo replications. ϵ_i is generated by different distributions:

	Probit		Logit	
	Difference	SD	Difference	SD
$N(0,1)$	0.0004	0.0206	0.5737	0.0414
$Logistic(0, \pi/\sqrt{3})$	-0.2317	0.0128	0.0004	0.0236
$N(1,2)$	-0.3878	0.0113	-0.3145	0.0187
$Logistic(1, 0.5)$	0.0342	0.0187	0.5008	0.0341
$U(-5,5)$	-0.3549	0.0112	-0.2607	0.0185
$Exp(0.5)$	-0.2439	0.0133	-0.0807	0.0221
$Weibull(1, 0.5)$	-0.3388	0.0149	-0.2164	0.0261

Table: Distortion from using naive parameter estimates as the true β

Literature

There is a large literature on ordered choice models.

- McKelvey and Zavoina (1971, 1975) introduced the basic model.
- Bayesian methods for many parameters (Congdon, 2005), bi-variate models (Biswas and Das, 2002), auto-correlation (Girard and Parent, 2001) or endogeneous dummy variables (Czado et al., 2011).
- IV estimation and ordered probit (Chevalier and Fielding, 2011)
- Individual variation in the set of thresholds (Greene, 2007; Greene and Hensher, 2010; Eluru and Yasmin, 2015)
- Varying coefficients (Boes and Winkelmann, 2010) and latent class models (McLachlan and Peel, 2004; Williams, 2016).
- Semi/non-parametric methods (Chen and Khan, 2003; Lewbel, 2000; Stewart, 2005).

One can find great overview in Koop and Tobias (2006) for Bayesian methods and in Greene and Hensher (2010) for frequentists. Again, we emphasise we are after the partial effects on conditional expectations!

Our debate with choice modeller

Recent heated exchange with Bill Greene.
Greene's argument:

- No interest in conditional expectation if one can estimate the conditional probabilities $\rightarrow$ elasticities are not relevant.
- Parameters and conditional probabilities are not so easy to interpret, one need to carefully scale the estimators, before compare them.

Our standpoint:

- The interest of conditional expectation is depending on the research question $\rightarrow$ this is purely subjective. There are many cases where one is interested in elasticities.
- Occam's razor: simple model, which nests the complicated non-linear models.

The main reason behind the heated exchange: if our method works, some part of the literature becomes irrelevant.....

Proposed solution

Instead of making restrictive assumptions (e.g., parametric), we exploit:

- These surveys are usually large.
- Estimators are consistent as the number of choices increases (M).

How to convert the M consistency into the "usual" consistency in the sample size N to gain the necessary information?

Remark: Similar method as with RHS variable, but different application! We return this question at slide 15.

Proposed solution

Instead of making restrictive assumptions (e.g., parametric), we exploit:

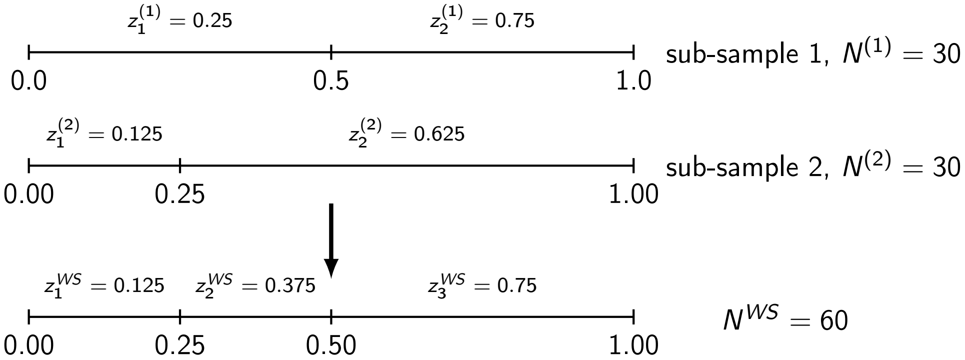
- These surveys are usually large.
- Estimators are consistent as the number of choices increases (M).

How to convert the M consistency into the "usual" consistency in the sample size N to gain the necessary information?

Remark: Similar method as with RHS variable, but different application! We return this question at slide 15.

Sub-sampling - idea

Use sub-samples, which measures different responses based on different questionnaires.



Sub-sampling - issues

- Characterisation of the sub-sampling methods is given by the choice of boundary points (c_m).
- In this paper, we focus on
  - Two methods of creating boundary points in sub-samples:
    - Magnifying method
    - Shifting method
  - We assume the *domain* for y_i is known.
  - We use the mid-values, $z_m = (c_m + c_{m-1})/2$
- Working sample's boundary points are the the union of boundary points in the sub-samples

$$\bigcup_{i=0}^{B} c_i^{WS} = \bigcup_{s=1}^{S} \bigcup_{j=0}^{M} c_j^{(s)}$$

- We create observations y_i^{WS}.

Sub-sampling - issues

- Characterisation of the sub-sampling methods is given by the choice of boundary points (c_m).
- In this paper, we focus on
  - Two methods of creating boundary points in sub-samples:
    - Magnifying method
    - Shifting method
  - We assume the *domain* for y_i is known.
  - We use the mid-values, $z_m = (c_m + c_{m-1})/2$
- Working sample's boundary points are the the union of boundary points in the sub-samples

$$\bigcup_{i=0}^{B} c_i^{WS} = \bigcup_{s=1}^{S} \bigcup_{j=0}^{M} c_j^{(s)}$$

- We create observations y_i^{WS}.

Sub-sampling - issues

- Characterisation of the sub-sampling methods is given by the choice of boundary points (c_m).
- In this paper, we focus on
  - Two methods of creating boundary points in sub-samples:
    - Magnifying method
    - Shifting method
  - We assume the *domain* for y_i is known.
  - We use the mid-values, $z_m = (c_m + c_{m-1})/2$
- Working sample's boundary points are the the union of boundary points in the sub-samples

$$\bigcup_{i=0}^{B} c_i^{WS} = \bigcup_{s=1}^{S} \bigcup_{j=0}^{M} c_j^{(s)}$$

- We create observations y_i^{WS}.

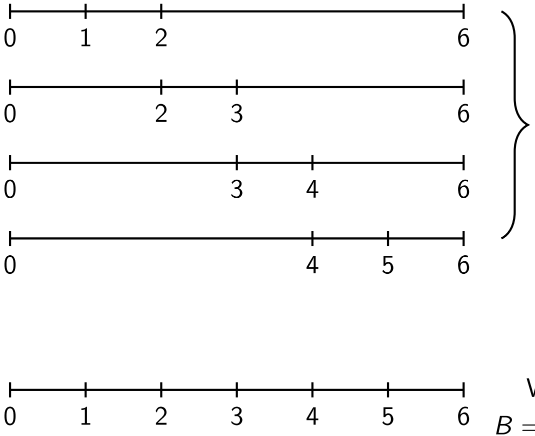
Sub-sampling - issues

- Characterisation of the sub-sampling methods is given by the choice of boundary points (c_m).
- In this paper, we focus on
  - Two methods of creating boundary points in sub-samples:
    - Magnifying method
    - Shifting method
  - We assume the *domain* for y_i is known.
  - We use the mid-values, $z_m = (c_m + c_{m-1})/2$
- Working sample's boundary points are the the union of boundary points in the sub-samples

$$\bigcup_{i=0}^{B} c_i^{WS} = \bigcup_{s=1}^{S} \bigcup_{j=0}^{M} c_j^{(s)}$$

- We create observations y_i^{WS}.

Magnifying method

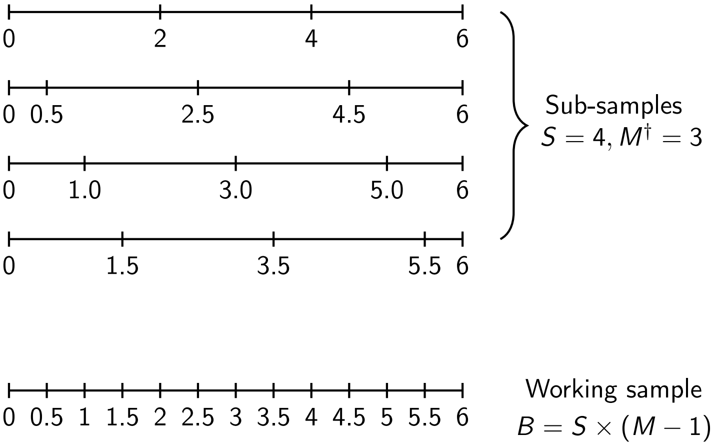


Sub-samples
$S = 4, M = 3$

Working sample
$B = S \times (M - 2) + 2$

Two types: '*DTO only*' and '*all observations with DTO*'.

Shifting method



Intuition for theory I.

With sub-sampling, we can reconstruct the underlying distribution,

$$y^{\dagger} \xrightarrow{d} y$$

However, we will not achieve pointwise-convergence!
The main challenge is to keep track of the correlation structure.

Remark: Difference when DOC variable is on RHS, we can not calculate simple conditional expectations based on C_m.

Intuition for theory I.

With sub-sampling, we can reconstruct the underlying distribution,

$$y^{\dagger} \xrightarrow{d} y$$

However, we will not achieve pointwise-convergence!
The main challenge is to keep track of the correlation structure.

Remark: Difference when DOC variable is on RHS, we can not calculate simple conditional expectations based on C_m.

Intuition for theory II.

Let D_l denote a set containing $l = 1, \ldots, L$ mutually exclusive partitions of the domain of x. Then

$$\mathbb{E}\left(y|x \in D_l\right) = \mathbb{E}\left(x|x \in D_l\right)\beta \qquad l = 1, \ldots, L. \tag{1}$$

Let $\tilde{y}_l$ and $\tilde{x}_l$ denote consistent estimates for the conditional expectations based on $y_i^{\dagger}, x_i$. We can estimate,

$$\tilde{y}_i = \tilde{x}_i\hat{\beta} + u_i, \tag{2}$$

where $\hat{\beta}$ is given by the OLS estimator. We have shown,

$$\hat{\beta} - \beta = o_p(1)\,,$$

Monte Carlo simulation baseline

We have used a simple DGP,

$$y_i = 0.5x_i + \epsilon_i,$$

$$x_i \sim N(0, 0.25, -1, 1)$$

$$\epsilon_i \sim Exp(0.5, 0, 4) - 0.5$$

y_i^* is generated with, $c_0 = -2, c_M = 4$ and used $M = 5$.

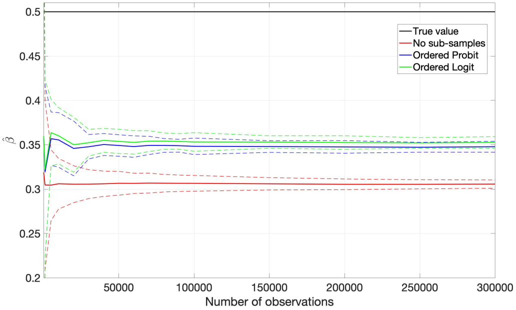


Figure: Different $\hat{\beta}$ estimators, with changing N, using $M = 5$

Monte Carlo simulation results - MAE

	N=1,000	N=10,000	N=100,000	N=500,000
Ordered Probit	0.1588	0.1523	0.1523	0.1533
Ordered Logit	0.1544	0.1483	0.1477	0.1488
No sub-samples	0.1969	0.1937	0.1938	0.1944
Magnifying - only DTO				
S=3	0.0584	0.0150	0.0046	0.0022
S=10	0.0884	0.0279	0.0085	0.0038
S=50	0.2279	0.0648	0.0204	0.0089
Magnifying - all observations with DTO				
S=3	0.0672	0.0153	0.0046	0.0021
S=10	0.1079	0.0296	0.0085	0.0038
S=50	0.2530	0.0752	0.0206	0.0089
Shifting				
S=3	0.0283	0.0099	0.0031	0.0014
S=10	0.0287	0.0097	0.0031	0.0014
S=50	0.0276	0.0094	0.0030	0.0013

Table: Mean absolute error for the estimated $\hat{\beta}$, with $M = 5$

Monte Carlo simulation results - revisited

$M = 3$ with $c_1 = -8, c_2 = -2, c_3 = 2, c_4 = 8$, one regressor $x_i \sim N(0, 1)$ with $\beta = 0.5$, $N = 10,000$ and 1000 Monte Carlo replications.

	Probit		Logit		Magnifying*		Shifting	
	Diff	SD	Diff	SD	Diff	SD	Diff	SD
$N(0,1)$	0.0004	0.0206	0.5737	0.0414	-0.0136	0.0293	0.0040	0.0236
$Logistic(0, \pi/\sqrt{3})$	-0.2317	0.0128	0.0004	0.0236	-0.0229	0.0533	-0.0022	0.0238
$N(1,2)$	-0.3878	0.0113	-0.3145	0.0187	-0.0175	0.0610	-0.0025	0.0230
$Logistic(1, 0.5)$	0.0342	0.0187	0.5008	0.0341	-0.0124	0.0293	-0.0049	0.0221
$U(-5,5)$	-0.3549	0.0112	-0.2607	0.0185	-0.0176	0.0901	-0.0268	0.0355
$Exp(0.5)$	-0.2439	0.0133	-0.0807	0.0221	0.0064	0.0168	0.0017	0.0089
$Weibull(1, 0.5)$	-0.3388	0.0149	-0.2164	0.0261	-0.0692	0.0568	-0.0344	0.0210

Table: Monte Carlo average distortion and standard deviation for $\hat{\beta}$ using different distributions

Perception Effect

Evidence in the behavioural literature: answers to a question depend on the way the question is asked.

Remark: Thanks for Botond, pointing out this effect and how to think about this problem.

$$y^{**} = \begin{cases} z_1 & \text{if} \quad c_0 < y_i + B_s < c_1 \\ \vdots & \\ z_m & \text{if} \quad c_{m-1} < y_i + B_s < c_M, \end{cases}$$

Leads to a model,

$$\tilde{y}^{**} = \tilde{y}_i^* + B_s = \beta \tilde{x}_i + u_i.$$

In the spirit of a FE estimator,

$$\hat{\beta} = \left(\tilde{x}' M_D \tilde{x}\right)^{-1} \tilde{x}' M_D \tilde{y}^{**}$$

which is a consistent, unbiased estimator for β.

Some other extensions

- Interacting class and sub-sample effects, such as B_{sm}
- Panel data: use of sub-sample effects and individual fixed effects, such as B_{is}
- Test the impacts of the perception effects on the estimator.
  - Use type of Wald test.
  - Exact regularity conditions and the construction of the test statistic would depend on the nature of the perception effect: B_s is fixed or stochastic.

What we have learned

- Modelling with ordered choice variables can be easy!
  - No need for complicated non-linear models with fancy estimation technique!
  - Only need a good survey-design!
- One can use OLS!
- Fixed effect estimator can take care of different forms of perception effects.
  - Test can be constructed for these effects.

Thank you for your attention!