Selection endogenous dummy ordered probit, and selection endogenous dummy dynamic ordered probit models

Massimiliano Bratti & Alfonso Miranda

Motivation

In many fields of applied work researchers need to model an **ordinal** dependent variable that is only observed for a proportion of the sample and that is a function of an endogenous variable

- Smoking and Education, but not everybody smokes
- Drinking and Education, but not everybody drinks
- Job type (unskilled/skilled/proffesional) and Education, but not everybody works

Selection can have different sources:

- Entry into an activity (smoking / drinking)
- Survey and/or item non-response

Fundamentally, the DGP of selection into missingness and the DGP of the ordinal variable are essentially two different although related processes (e.g., extensive vs. intensive margin decisions)

Previous work

- Terza, Kenkel, Tsui-Fang and Shinichi (2008) suggest a two-step method for estimating a selection endogenous dummy model for an interval coded dependent variable (grouped). This is an extension of Mullahy (1998) Modified Two Part Model and Mullahy (1986) Hurdle Model. Despite being a two-step approach, this is not a LIML estimator but relies on joint multivariate normality.
- Miranda and Rabe-Hesketh (2006) consider a model for an ordinal dependent variable with either an endogenous dummy or sample selection. Cannot deal with the two problems at the same time.
- Harris and Zhao (2007) suggests a zero-inflated ordered probit model, which is quite similar to Mullahy's Hurdle model and is related to Lambert (1992) zero-inflated count data models.
- Bratti and Miranda (2009) use the BCS70 and the methods described here to analyse how higher education affects smoking intensity in the UK.

Selection endogenous dummy ordered probit

Let y_i be the ordinal variable of interest. Variable y_i is generated according to a continuous latent variable model

$$y_i^* = \mathbf{x_i'}\boldsymbol{\beta} + \delta G_i + v_i, \tag{1}$$

where the observed response y_i is determined by a threshold model

$$y_i = \begin{cases} \text{missing} & \text{if} \quad S_i = 0 \\ 1 & \text{if} \quad y_{it}^* \leq k_1 \ \& \ S_i = 1 \\ 2 & \text{if} \quad k_1 < y_{it}^* \leq k_2 \ \& \ S_i = 1 \\ . & . \quad . \\ . & . \quad . \\ H & \text{if} \quad k_{H-1} < y_{it}^* \ \& \ S_i = 1, \end{cases}$$

G_i represents a potentially endogenous dummy and the main response y_i is only observed if a selection rule $S_i = 1$ is met. Both G_i and S_i are always observed and $\{k_1, \cdots, k_{H-1}\} \in \mathbb{R}^{H-1}$ are constants to be estimated along other parameters.

Selection endogenous dummy ordered probit II

The endogenous dummy and the selection dummy are also generated according to a continuous latent variable model

$$\begin{aligned} S_i^* &= \mathbf{r_i}'\boldsymbol{\theta} + \varphi G_i + q_i \\ G_i^* &= \mathbf{z_i}'\boldsymbol{\gamma} + w_i, \end{aligned} \tag{2}$$

with $S_i = 1(S_i^* > 0)$ and $G_i = 1(G_i^* > 0)$. Although the model is identified by functional form it is valuable to specify a set of exclusion restrictions to avoid problems of *tenuous identification*. Hence, when possible, some elements of the $\mathbf{z_i}$ should not enter $\mathbf{x_i}$ or $\mathbf{r_i}$, and some elements of $\mathbf{r_i}$ should not enter $\mathbf{x_i}$ or $\mathbf{z_i}$.

Correlation among the three outcomes is allowed by imposing some structure to the error terms,

$$\begin{aligned} w_i &= u_i + \zeta_i, \\ q_i &= \lambda_1 u_i + \eta_i \\ v_i &= \lambda_2 u_i + \nu_i, \end{aligned} \tag{3}$$

Selection endogenous dummy ordered probit III

To close the model we require the covariates to be all strictly exogenous, and the idiosyncratic errors to be orthogonal given the individual heterogeneity term u,

$$
\begin{aligned}
D(u|\mathbf{x},\mathbf{z},\mathbf{r}) &= D(u) && (4)\\
D(\zeta|\mathbf{x},\mathbf{z},\mathbf{r},u) &= D(\zeta|u) && (5)\\
D(\eta|\mathbf{x},\mathbf{z},\mathbf{r},u) &= D(\eta|u) && (6)\\
D(\nu|\mathbf{x},\mathbf{z},\mathbf{r},u) &= D(\nu|u) && (7)\\
\zeta|u \perp \eta|u &\perp \nu|u. && (8)
\end{aligned}
$$

To ease estimation we suppose that u, ζ, η, and ν are all independent standard normal, in which case $D(\zeta|u) = D(\zeta)$, $D(\eta|u) = D(\eta)$, and $D(\nu|u) = D(\nu)$.

The *factors loadings* $\{\lambda_1\ \lambda_2\} \in \mathbb{R}^2$ are free parameters and allow any type of correlation (positive, negative or null) between y_i^*, G_i^* and S_i^*.

Selection endogenous dummy ordered probit IV

Correlation between unobservables entering G_i^* and S_i^* is given by:

$$\rho_{gs} = \frac{\lambda_1}{\sqrt{2\left(1+\lambda_1^2\right)}}.$$

Similarly, correlation between unobservables entering G^* and y_i^* is given by:

$$\rho_{gy} = \frac{\lambda_2}{\sqrt{2\left(1+\lambda_2^2\right)}}.$$

Finally, correlation between unobservables entering main response y^* and selection S_i^* is given by:

$$\rho_{sy} = \frac{\lambda_1\lambda_2}{\sqrt{\left(1+\lambda_1^2\right)\left(1+\lambda_2^2\right)}}.$$

In this model G is exogenous wrt S if $\rho_{gs} = 0$, G is exogenous wrt y if $\rho_{gy} = 0$, and y is observed at random if $\rho_{sy} = 0$.

Selection endogenous dummy ordered probit V

- Estimate the system by Maximum Simulated Likelihood
- Analytical first derivatives and numerical second derivatives
- Can also do OPG approx. of the Hessian (much faster!)
- Halton sequences cover the (0,1) interval better and require fewer draws to achieve high precision than random samples from uniform distribution
- Program written in Stata/Mata
- Really fast!
  - Stata 10/SE + 400 Halton draws + 2,792 indv / 8,043 pers-obs + numerical 2nd derivatiives = 1.6hrs
  - Stata 10/SE + 400 Halton draws + 2,792 indv / 8,043 pers-obs + OPG Hessian = less than 5min

Selection Endogenous dummy dynamic ordered probit model

Suppose that y_i is observed for two periods $t = \{1, 2\}$. We now extend the model to accommodate the fact that the outcome of the ordinal response in period 2 can be a function of the value that the variable took in period 1. In other words, we consider the possibility of having autoregressive dynamics in the ordered variable y_{i2} such that

$$y_{i1}^* = \mathbf{m}_i'\boldsymbol{\delta} + \delta_1 G_i + v_i \tag{9}$$

$$y_{i2}^* = \mathbf{x}_i'\boldsymbol{\beta} + \delta_2 G_i + \sum_{j=1}^{H} \pi_j 1(y_{i1} = j) + \xi_i, \tag{10}$$

Selection endogenous dummy dynamic ordered probit model II

As usual we suppose the model is complemented by a threshold rule,

$$y_{it} = \begin{cases} \text{missing} & \text{if} \quad S_{it} = 0 \\ 1 & \text{if} \quad y_{it}^* \leq k_1 \ \& \ S_{it} = 1 \\ 2 & \text{if} \quad k_1 < y_{it}^* \leq k_2 \ \& \ S_{it} = 1 \\ . & . \quad . \\ . & . \quad . \\ H & \text{if} \quad k_{H-1} < y_{it}^* \ \& \ S_{it} = 1, \end{cases}$$

To ease presentation we suppose that G and S do not have dynamics themselves. Further, we suppose that y is always observed in the first period and that G is a time invariant variable. Hence, The selection and endogenous dummies are generated by the following latent variable models,

$$G_{it}^* = G_i^* = \mathbf{z_i'}\gamma + w_i \tag{11}$$

$$S_{it}^* = S_i^* = \mathbf{r_i'}\boldsymbol{\theta} + \varphi_1 G_i + q_i \tag{12}$$

Selection endogenous dummy dynamic ordered probit model III

Like in the SED-OP model, we impose some structure to the error terms,

$$\begin{aligned} w_i &= u_i + \zeta_i, \\ q_i &= \lambda_1 u_i + \eta_i \\ v_i &= \lambda_2 u_i + \nu_i \\ \xi_i &= \lambda_3 u_i + \varepsilon_i, \end{aligned} \tag{13}$$

where u_i, ζ_i, η_i, ν_i, and ε_i are all supposed to be independent standard normal, and $\lambda_1, \ldots, \lambda_3$ are free factor loadings.

Notice that the model fully recognises the fact that the initial state dummies$1(y_{i1} = 1), \ldots, 1(y_{i1} = H)$ are potentially endogenous in equation (10) by modelling together the *initial conditions* and the dynamic equation, and allowing any type of correlation among unobservables entering both equations. There is true state dependance if the coefficients on the initial state dummies in (10) are different from zero.

Selection endogenous dummy dynamic ordered probit model IV

Unobservables entering the system are now correlated in six different ways:

$$\rho_{g,s} = \frac{\lambda_1}{\sqrt{2\left(1+\lambda_1^2\right)}} \qquad \rho_{g,y_1} = \frac{\lambda_2}{\sqrt{2\left(1+\lambda_2^2\right)}}$$

$$\rho_{g,y_2} = \frac{\lambda_3}{\sqrt{2\left(1+\lambda_3^2\right)}} \qquad \rho_{s,y_1} = \frac{\lambda_1\lambda_2}{\sqrt{\left(1+\lambda_1^2\right)\left(1+\lambda_2^2\right)}}$$

$$\rho_{s,y_2} = \frac{\lambda_1\lambda_3}{\sqrt{\left(1+\lambda_1^2\right)\left(1+\lambda_3^2\right)}} \qquad \rho_{y_1,y_2} = \frac{\lambda_2\lambda_3}{\sqrt{\left(1+\lambda_2^2\right)\left(1+\lambda_3^2\right)}}.$$

Selection endogenous dummy dynamic ordered probit model V

- The initial state dummies can be included into the S^* without further complications
- The model can be extended to allow dynamics in S^* and G^*
- The model can deal with more than two periods with relative minor modifications
- As before we use MSL for estimation
- Program written in Stata/Mata
- Really fast!

Example I - Variables definition

We apply the SED-OP model to study **the effect of higher education (HE) on drinking frequency** using the British Cohort Study 1970 (BCS70), 29-year follow-up survey.

Variables definition:

- y_i (drinking frequency) $= 1$ (2 to 3 times a month); 2 (once a week); 3 (2 to 3 days a week); 4 (on most days)
- G_i (higher education) $= 1$ (HE); 0 (lower than HE)
- S_i (usual drinker) $= 1$ (drinks more than 2 to 3 times a month); 0 (drinks less than 2 to 3 times a month, i.e. less often or in special occasions, not nowadays, never drunk)

Example II - Descriptive statistics

Descriptive statistics, BCS70, 29-year follow-up survey

	% of usual drinkers	
	Women	Men
No HE	69.9	85.78
HE	83.56	92.02

	Drinking frequency per month - usual drinkers				
	2-3 times per month	once a week	2-3 times a week	most days	Total
Women					
no HE	25.04	34.4	31.63	8.92	100
HE	16.01	24.53	43.53	15.93	100
Total	21.7	30.75	36.04	11.51	100
Men					
no HE	14.14	24.95	43.27	17.63	100
HE	8.95	19.13	48.35	23.57	100
Total	12.35	22.94	45.03	19.69	100

Example III - SED-OP specification

- **HE**: parents' absence, parents' education, highest social class, school type, state, *BAS score*, ethnicity, religion, height, *teacher's assessment of child's knowledge and parental interest in child's education, teacher's homework style*, all at age 10.
- **Usual drinker**: HE, parents' absence, parents' education, highest social class, school type, state, ethnicity, religion, height, all at age 10; *height at age 30*, *mother's drinking during pregnancy*, homework on parents' demand, *month of 29-year follow-up interview*
- **Drinking frequency**: same controls as the selection equation (usual drinker)

Since the selection and the main outcome variables refer to the same process (drinking) we preferred not to impose exclusion restrictions between the two equations. In other cases, for instance item non-reponse and panel attrition, it can be easier to find valid exclusion restrictions.

Example III - SED-OP Results

	Be a usual drinker -s- (ME)	Drinking frequency -y- (ME) 2-3 times a month	once a week	2-3 times a week	most days
Men					
HE -g-	0.156*** [0.015]	-0.129*** [0.032]	-0.109*** [0.021]	0.054*** [0.009]	0.183*** [0.046]
ρ_{gs}		-0.484*** [0.056]			
ρ_{gy}		-0.320*** [0.091]			
ρ_{sy}		0.310*** [0.063]			
No. obs.		3300			
Women					
HE -g-	0.259*** [0.024]	-0.302*** [0.046]	-0.087*** [0.010]	0.181*** [0.015]	0.208*** [0.039]
ρ_{gs}		-0.418*** [0.061]			
ρ_{gy}		-0.479*** [0.087]			
ρ_{sy}		0.401*** [0.058]			
No. obs.		3525			

***(**) Significant at 1% (5%). Eicker-Huber-White robust standard errors reported in square brackets. Marginal effects (ME) are computed at the sample mean. For dummy variables, they show the change in the relevant probability when the variable changes from 0 to 1.

Concluding remarks

Hence, our empirical application shows that:

- HE is endogenous wrt both drinking participation (i.e. be an usual drinker) and drinking frequency
- Unobservables affecting drinking participation and drinking frequency are positively correlated (positive selection)
- HE has a positive causal effect on both drinking participation and drinking frequency (i.e., educated people drink more)
- The effect is much larger for females than for males

References

- Bratti, M and Miranda, A (2009) Non-pecuniary returns to higher education: The effect on smoking intensity in the UK. *Health Economics* Published Online: Jul 13 2009 12:08PM (Early View) `http://dx.doi.org/10.1002/hec.1529`.
- Harris, MN and Zhao, X (2007) A zero-inflated ordered probit model, with an application to modelling tobacco consumption. *Journal of Econometrics* **141**:1073–1099.
- Heckman, JJ (1979) Sample selection bias as a specication error. *Econometrica* **47**: 153–61.
- Lambert,D (1992) Zero-Inflated Poisson Regression with an Application to Defects in Manufacturing. *Technometrics* **34**:1-14.
- Miranda, A. and Rabe-Hesketh, S. (2006) Maximum likelihood estimation of endogenous switching and sample selection models for binary, ordinal, and count variables. *Stata Journal* **6** (3): 285–308.
- Terza, JV and Kenkel, DS and Lin, TF and Sakata, S (2008) Care-giver advice as a preventive measure for drinking during pregnancy: zeros, categorical outcome responses, and endogeneity. *Health Economics* **17**: 41-54.
- Train, KE (2003). Discrete choice methods with simulation. Cambridge university press.