# Estimation of ordinal response models, accounting for sample selection bias.

Alfonso Miranda
University of Keele

Sophia Rabe-Hesketh
University of California, Berkeley

Correspondence to: A.Miranda@econ.keele.ac.uk

11th UK Stata Users Group meeting, May 2005.

## Ordered responses

- A limited number of H response categories $y_h, h = 1, 2, ..., H$.
- Categories are ordered,

$$y_1 < y_2 < ... < y_H$$

- Some examples:
  - Health condition status (excellent, good, regular, bad).
  - Opinions of a candidate in an election (strongly support, neutral, strongly opposed).
  - Job satisfaction (highly satisfied, satisfied, not satisfied).

# Latent regression models

- The observed response for individual $i$, $y_i$ is determined by a latent continuous variable process,

$$y_i^* = \mathbf{x_i}'\boldsymbol{\beta} + u_i \tag{1}$$

- A threshold model determines the observed response:

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

- No constant is included in the covariate vector $\mathbf{x_i}$.

## The sample selection problem

- the response variable is only observed if a particular condition $(sel = 1)$ is met.
- A latent regression model for the selection variable is specified,

$$sel_i^* = \mathbf{z_i}'\boldsymbol{\gamma} + v_i, \tag{2}$$

  where $v_i$ is assumed to be normally distributed ($z_i$ should include some variables not in $x_i$ to secure identification).
- If $Cov(u_i, v_i) \neq 0$, using the observed sample of $y$ and ordered Probit (ordered Logit) to estimate $\boldsymbol{\beta}$ will deliver biased estimators.
- This is known as the 'sample selection bias' problem (Heckman 1979).

## The sample selection problem (Cont.)

- Notice that the correlation coefficient, $\rho$, is the only aspect of the covariance matrix that is identified. We impose therefore,

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

<u>Example</u>

- Health condition only available for respondents who exercise at least twice a week. If healthier people exercise more, estimating a model for health condition on the basis of the observed sample will clearly deliver biased estimates.

## Generalised linear latent and mixed models

- May use GLLAMMs to estimate the sample selection ordinal variable regression. We can write the ordinal variable model as,

$$\begin{aligned} y_i &\sim multinomial\left(1,\{\pi_{hi},h=1,\ldots,H\}\right); \\ g_1\left(\gamma_{hi}\right) &= \eta_{1hi} = \mathbf{x_i}'\boldsymbol{\beta}-\kappa_h+\lambda\varepsilon_i; \\ & \quad h=1,...,H-1. \end{aligned}$$

where

$$\gamma_{hi}=\sum_{s=h+1}^{H}\pi_{si}=Pr\left(y>h\right),$$

## GLLAMMs (Cont. 1)

And the selection model as,

$$\begin{aligned} sel_i &\sim binomial\,(1, \pi_i) \\ g_2\,(\pi_i) &= \eta_{2i} = \mathbf{z_i}'\boldsymbol{\gamma} + \varepsilon_i \end{aligned}$$

where $\varepsilon_i \sim N(0,1)$ is a latent variable representing unobserved heterogeneity and $\lambda$ is a factor loading. This reparametrization reduces the dimensions of integration from 2 to 1.

<u>A mixed response model</u>

- Stack $y_i$ and $sel_i$ into a single variable $q_{ji}$, $j = 1, 2$.
- Viewing the ordinal variable j = 1 and the selection status j = 2 as clustered within individuals i, define the dummies $d_{1ji} = 1$ for the ordinal variable and $d_{2ji} = 1$ for the selection.

## GLLAMMs (Cont. 2)

- Now we can define a mixed response model for $q_{ji}$

$$
\begin{aligned}
q_{ji} &\sim \begin{cases} \textit{multinomial} & \text{if} \quad d_{1ji} = 1 \\ \textit{binomial} & \text{if} \quad d_{2ji} = 1 \end{cases} \\
\eta_{jhi} &= d_{1ji} \left[ \mathbf{x_i}' \boldsymbol{\beta} - \kappa_h + \lambda \varepsilon_i \right] + d_{2ji} \left[ \mathbf{z_i}' \boldsymbol{\gamma} + \varepsilon_i \right]; \\
& j = 1, 2, \quad h = 1, \ldots, H - 1; \qquad (3)
\end{aligned}
$$

- $g_1$ can be either the ordered Probit or the ordered Logit link. We use always a Probit link for $g_2$.

## GLLAMMs (Cont. 3)

- Due to the increase in the residual variance in (3) we expect $\boldsymbol{\beta}$ to increase by a factor of $\sqrt{1+\lambda^2}$ if $g_1$ is oprobit or $\sqrt{\frac{\pi^2}{3}+\lambda^2}$ if $g_1$ is ologit.
- Similarly, we expect $\boldsymbol{\gamma}$ to increase by a factor of $\sqrt{2}$.
- Hence, after estimation $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ must be rescaled!!
- Notice finally that,

$$\rho = \begin{cases} \frac{\lambda}{\sqrt{2(1+\lambda^2)}} & \text{if} \quad g_1 \quad \text{is} \quad \text{oprobit} \\ \frac{\lambda}{\sqrt{2\left(\frac{\pi^2}{3}+\lambda^2\right)}} & \text{if} \quad g_1 \quad \text{is} \quad \text{ologit} \end{cases}$$

## The **osm** command

- **osm** is a **gllamm** (Rabe-Hesketh, Skrondal & Pickles 2004) 'wrapper' program that fits endogenous switching and sample selection models for ordinal and count variables (endogenous switching is the default option).
- Accepts data in the usual wide format and then does the required changes to call **gllamm**.
- After estimation coefficients are rescaled and an output table that is easily interpretable is presented.
- **osm** exploits the adaptive quadrature capability of **gllamm**...one of the major **gllamm** strengths.

## Syntax

**osm** *depvar* [*varlist*] [`if` *exp*] [`in` *range*], `i`(*varname*)

`switch`(*varname*= *varlist*) `switch`(*varlist*) `Family`(*familyname*)

`selection` `quadrature`(#) `Link`(*linkname*) `From`(*initial values*)

`Trace` `nolog` `Trace` `Eval` `Commands`

Table 1: Avaibale Families and links

| Family | Link |
|---|---|
| Poisson | log |
| Binomial | ordinal Probit |
| | ordinal Logit |

# Sample selection Ordered Probit: An example

```
. osm ordvar x1 x2, id(id) s(sel = x1 x2 x3 x4) q(15) adapt family(bin) link(oprobit) sel

Running adaptive quadrature
Iteration 0:    log likelihood =  -5444.942
  (output omitted )
Iteration 4:    log likelihood = -5175.5835

Adaptive quadrature has converged, running Newton-Raphson
Iteration 0:   log likelihood = -5175.5835
  (output omitted )
Iteration 3:   log likelihood = -5175.5765

Sample Selection Ordered Probit Regression
(Adaptive quadrature -- 15 points)

                                                     Number of obs  =      3500
                                                     Wald chi2(6)   =   1114.42
Log likelihood = -5175.5765                          Prob > chi2    =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ordvar       |
          x1 |   .3154251   .0408026     7.73   0.000     .2354535    .3953968
          x2 |   .1470225    .026887     5.47   0.000     .0943249    .1997202
-------------+----------------------------------------------------------------
```

```
selection    |
          x1 |   .9573865   .0356374    26.86   0.000     .8875385    1.027234
          x2 |   .4217439   .0286755    14.71   0.000      .365541    .4779468
          x3 |  -.5968153   .0303954   -19.64   0.000    -.6563893   -.5372414
          x4 |   .6372245   .0308598    20.65   0.000     .5767403    .6977087
       _cons |   .5448698   .0288654    18.88   0.000     .4882947     .601445
-------------+----------------------------------------------------------------
aux_ordvar   |
       _cut1 |  -.4012284   .0325979   -12.31   0.000    -.4651192   -.3373376
       _cut2 |   .1583416    .048699     3.25   0.001     .0628932    .2537899
       _cut3 |   .4265045   .0598836     7.12   0.000     .3091348    .5438743
       _cut4 |   .7873888   .0763759    10.31   0.000     .6376948    .9370827
       _cut5 |   1.229029   .0981156    12.53   0.000     1.036726    1.421332
-------------+----------------------------------------------------------------
         rho |   .2901614   .0654419     4.43   0.000     .1458488    .4012554
------------------------------------------------------------------------------
Likelihood ratio test for rho=0: chi2(1)= 21.32 Prob>=chi2 = 0.000
```

# Sample selection Ordered Logit: An example

```
. osm ordvar x1 x2, id(id) s(sel = x1 x2 x3 x4) q(15) adapt family(bin) link(ologit) sel

Running adaptive quadrature
Iteration 0:    log likelihood = -5468.3146
  (output omitted )
Iteration 6:    log likelihood = -5180.7342

Adaptive quadrature has converged, running Newton-Raphson
Iteration 0:   log likelihood = -5180.7342
  (output omitted )
Iteration 3:   log likelihood = -5180.7303

Sample Selection Ordered Logit Regression
(Adaptive quadrature -- 15 points)

                                                     Number of obs  =      3500
                                                     Wald chi2(6)   =   1123.24
Log likelihood = -5180.7303                          Prob > chi2    =    0.0000

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ordvar       |
          x1 |   .3267404   .0369461     8.84   0.000     .2543274    .3991535
          x2 |   .1510283   .0248903     6.07   0.000     .1022441    .1998125
-------------+----------------------------------------------------------------
```

```
selection    |
          x1 |   .9563906   .0356895    26.80   0.000     .8864404    1.026341
          x2 |   .4199398   .0287014    14.63   0.000     .3636862    .4761935
          x3 |  -.5964406   .0305075   -19.55   0.000    -.6562341    -.536647
          x4 |   .6383798   .0309312    20.64   0.000     .5777557    .6990038
       _cons |   .5446456   .0288797    18.86   0.000     .4880425    .6012486
-------------+----------------------------------------------------------------
aux_ordvar   |
       _cut1 |   -.447559   .0312788   -14.31   0.000    -.5088643   -.3862536
       _cut2 |   .0840034   .0416869     2.02   0.044     .0022986    .1657082
       _cut3 |   .4016877   .0529198     7.59   0.000     .2979668    .5054085
       _cut4 |   .7749008   .0673403    11.51   0.000     .6429162    .9068854
       _cut5 |   1.123638   .0803356    13.99   0.000     .9661834    1.281093
-------------+----------------------------------------------------------------
         rho |   .1214521   .0392716     3.09   0.002     .0426992    .1958209
------------------------------------------------------------------------------
Likelihood ratio test for rho=0: chi2(1)= 11.43 Prob>=chi2 = 0.001
```

## Final remarks

- Besides estimating sample selection models **osm** fits endogenous switching models (i.e., when an endogenous dummy is present in the main equation).
- Using the Poisson Family and the Log link **osm** can fit models for count data.
- In the near future **osm** will be extended to allow for:
  - Probit/Logit links.
  - Weights.