Parametric and semiparametric estimation of ordered response models with sample selection and individual-specific thresholds

Giuseppe De Luca, ISFOL
Valeria Perotti, ISFOL
Claudio Rossetti, Tor Vergata University

Italian Stata Users Group Meeting,
October 20, 2008

1 Introduction

This paper focuses on the estimation of a sample selected ordered choice model where

- observations are subject to **binary section mechanism**.
- the outcome variable of interest is measured on an **ordinal scale**.

Examples:

- **Outcome variable:** educational attainment, job satisfaction, self-reported health assessments, cognitive ability measures, ...
- **Selection mechanism:** participation into a training program, self-selection in the labor market, missing data problems, ...

Novelties of our paper: we provide new Stata commands for

- **parametric ML** estimation of a sample selected ordered probit model.
- **parametric ML** estimation of a sample selected ordered probit model with **individual heterogeneity**
- **semi-nonparametric (SNP)** estimation of a sample selected ordered choice model.

2 Sample selected ordered probit model

Our baseline model is a straightforward variation of a classical sample selection model (Heckman 1979) where the outcome equation is non-linear,

$$Y_j^* = X_j^\top \beta_j + U_j, \qquad j = 1, 2, \tag{1}$$

$$Y_1 = I(Y_1^* \geq 0), \tag{2}$$

$$Y_2 = \sum_{h=1}^{H} h\, I(\alpha_{h-1} < Y_2^* \leq \alpha_h) \qquad \text{if } Y_1 = 1, \tag{3}$$

- the Y_1 is the binary selection mechanism,
- the Y_2 is the observed outcome variable of interest,
- the X_j are k_j-vectors of exogenous variables,
- the β_j are k_j-vectors of unknown parameters,
- $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_H)$, with $\alpha_0 = -\infty$, $\alpha_H = +\infty$ and $\alpha_{h-1} > \alpha_h$, is a vector of ordered threshold coefficients.
- the U_j are latent regression errors independent of (X_1, X_2).

Assumption: the joint distribution function of (U_1, U_2) is Gaussian, with zero means, unit variances and correlation coefficient ρ.

Data allow to identify $(H+1)$ possible events. Under the Gaussian distributional assumption, the probabilities of these events are

$$\begin{aligned}\pi_0(\theta) &= \Pr\{Y_1 = 0\} = 1 - \Phi(\mu_1), \\ \pi_h(\theta) &= \Pr\{Y_1 = 1, Y_2 = h\} = \\ &= \Phi_2(\mu_1, \alpha_h - \mu_2; -\rho) - \Phi_2(\mu_1, \alpha_{h-1} - \mu_2; -\rho).\end{aligned} \tag{4}$$

with $h = 1, \ldots, H$, $\theta = (\beta_1, \beta_2, \alpha, \rho)$ and $\mu_j = X_j^\top \beta_j$.

A **parametric ML estimator** of θ maximizes the likelihood function

$$L(\theta) = \prod_{i=1}^{n} \pi_{0i}(\theta)^{1-Y_{1i}} \prod_{h=1}^{H} \pi_{hi}(\theta)^{Y_{1i} I(Y_{2i}=h)}. \tag{5}$$

3 Extension 1: Modelling Individual Heterogeneity

There are at least three approaches:

- **Approach 1:** using a random coefficient specification for the slope coefficients β_j (Greene, 2002; Boes and Winkelmann, 2006).
- **Approach 2:** allowing the threshold coefficients α_h to depend on a set of observable covariates (Terza 1985).
- **Approach 3:** using anchoring vignette questions to account for individual heterogeneity in the response scale of Y_2^* (King *et al.* 2004).

Here, we focus on Approaches 2 and 3.

3.1 Individual specific thresholds

The thresholds coefficients are allowed to depend on a set of observable covariates Z according to

$$\begin{aligned} \alpha_1 &= Z^\top \delta_1 \\ \alpha_h &= \alpha_{h-1} + \exp(Z^\top \delta_h), \qquad h = 2, \ldots, H-1 \end{aligned} \tag{6}$$

where $\delta_1, \ldots, \delta_H$ are threshold-specific vectors of parameters to be estimated jointly with (β_1, β_2, ρ).

Model (6) guarantees that:

- thresholds are defined over the whole real line,
- monotonicity of the thresholds: $\alpha_h > \alpha_{h-1}$ for every h.

Identification: Model (6) is **identified** only if Z and X_2 do **not** have common variables.

3.2 The model with vignettes

People of different groups may judge similar conditions in quite different ways.

Vignette questions can be considered as an instrument to control for individual heterogeneity in the response scale of Y_2^*.

- a **self-assessment question:** where respondents evaluate their own subjective outcome using an ordered response scale,
- some **vignette questions:** where respondents evaluate the subjective outcome of a hypothetical individual using the same response scale.

Data availability: Vignette data have been recently collected in sample surveys like SHARE, ELSA, HRS, among others.

Example: Depression Problems - SHARE 2004

- **Self-assessment question:** Overall in the last 30 days, how much of a problem did you have with feeling sad, low, or depressed?

  Answers: 1. None, 2. Mild, 3. Moderate, 4. Severe, 5. Extreme

- **Vignette question 1:** Anna feels depressed most of the time. She weeps frequently and feels hopeless about the future. She feels that she has become a burden on others and that she would be better dead.

  Overall in the last 30 days, how much of a problem did Anna have with feeling sad, low, or depressed?

  Answers: 1. None, 2. Mild, 3. Moderate, 4. Severe, 5. Extreme

- **Vignette question 2:** Maria feels nervous and anxious. She worries and thinks negatively about the future, but feels better in the company of people or when doing something that really interests her.....

  Overall in the last 30 days, how much of a problem did Maria have with feeling sad, low, or depressed?

  Answers: 1. None, 2. Mild, 3. Moderate, 4. Severe, 5. Extreme

Assumptions: Following King *et al.* (2004), we assume that

- **Vignette equivalence:** levels of vignette variables are perceived by all respondents in the same way, apart from random measurement error.
- **Response consistency:** respondents use response categories in the same way when answering self-assessment and vignette questions.

Under these assumptions, vignette data provide repeated observations on the scale of the latent variable Y_2^*.

Our baseline model can be extended to include $J-2$ vignette variables

$$
\begin{aligned}
Y_j^* &= X_j^\top \beta_j + U_j, && j = 1, \ldots, J, && (7)\\
Y_1 &= I(Y_1^* \geq 0), && && (8)\\
Y_j &= \sum_{h=1}^{H} h\, I(\alpha_{h-1} < Y_j^* \leq \alpha_h) \qquad \text{if } \; Y_1 = 1, && j = 2, \ldots, J. && (9)
\end{aligned}
$$

where

- Y_1 is the binary selection mechanism,
- Y_2 is the observed outcome of the self-assessment question,
- the Y_j, $j = 3, \ldots, J$, are the observed outcomes of the vignette questions,
- the thresholds coefficients are specified as follows

$$
\begin{aligned}
\alpha_1 &= Z^\top \delta_1 + \eta \\
\alpha_h &= \alpha_{h-1} + \exp(Z^\top \delta_h), \qquad h = 2, \ldots, H-1
\end{aligned}
\tag{10}
$$

  where η is a random effect independent of (X_1, X_2, Z, U) and distributed according to $N(0, \varphi^2)$.

- $U = (U_1, \ldots, U_J)$ is a vector of error terms which follow a J-variate Gaussian distribution with zero means and covariance matrix

$$
\Omega = \begin{bmatrix}
1 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1J} \\
 & 1 & \sigma_{23} & \cdots & \sigma_{2J} \\
 & & \sigma_3^2 & \cdots & \sigma_{3J} \\
 & & & \ddots & \vdots \\
 & & & & \sigma_J^2
\end{bmatrix}.
$$

For parsimony reasons and to reduce the computational burden of the estimation process, we assume that:

- Error terms of the vignette equations are mutually uncorrelated and with the same variance,

$$\begin{aligned} \sigma_{ks} &= 0 \qquad k, s = 3, \ldots, J \\ \sigma_k^2 &= \sigma^2 \qquad k = 3, \ldots, J \end{aligned}$$

- Error terms of the vignette equations are uncorrelated with the error term of the selection equation,

$$\sigma_{2s} = 0 \qquad s = 3, \ldots, J$$

- Error term of the vignette equations are equally correlated with the error term of the selection equation

$$\sigma_{1s} = \sigma_{1v} \qquad s = 3, \ldots, J$$

Thus,

$$\Omega = \begin{bmatrix} 1 & \sigma_{12} & \sigma_{1v} & \cdots & \sigma_{1v} \\ & 1 & 0 & \cdots & 0 \\ & & \sigma^2 & \cdots & 0 \\ & & & \ddots & \vdots \\ & & & & \sigma^2 \end{bmatrix}.$$

A parametric ML estimator of $\theta = (\beta_1, \beta_2, \delta_1, \ldots, \delta_H, \sigma, \sigma_{12}, \sigma_{1v}, \varphi)$ maximizes the likelihood function

$$L(\theta) = \prod_{i=1}^{n} \int L_i^s(\theta_1 \,|\, \eta) L_i^v(\theta_2 \,|\, \eta) \frac{1}{\varphi} \phi\left(\frac{\eta}{\varphi}\right) d\eta. \tag{11}$$

where

- the random effect η is integrated-out by approximating the integral in (11) through Gauss-Hermite quadrature method.
- $L^s(\theta_1 \,|\, \eta)$ is the conditional likelihood of the self-assessed component

$$L^s(\theta_1 \,|\, \eta) = \pi_0^{1-Y_1} \prod_{h=1}^{H} \pi_{2h}(\eta)^{Y_1\, I(Y_2=h)}.$$

  with $\theta_1 = (\beta_1, \beta_2, \delta_1, \ldots, \delta_H, \sigma_{12})$ and
  $\pi_{2h}(\eta) = \Phi_2(\mu_1, \alpha_h - \mu_2; -\sigma_{12}) - \Phi_2(\mu_1, \alpha_{h-1} - \mu_2; -\sigma_{12})$.

- $L^v(\theta_2 \,|\, \eta)$ is the conditional likelihood of the vignette component

$$L^v(\theta_2 \,|\, \eta) = \pi_0^{1-Y_1} \prod_{j=3}^{J} \prod_{h=1}^{H} \pi_{jh}(\eta)^{Y_1\, I(Y_j=h)},$$

  with $\theta_2 = (\beta_1, \delta_1, \ldots, \delta_H, \sigma, \sigma_{1v})$, $\rho_{1v} = \sigma^{-1}\sigma_{1v}$, and
  $\pi_{jh}(\eta) = \Phi_2(\mu_1, \sigma^{-1}(\alpha_h - \mu_j); -\rho_{1v}) - \Phi_2(\mu_1, \sigma^{-1}(\alpha_{h-1} - \mu_j); -\rho_{1v})$. .

4 Extension 2: SNP Estimation

The literature on semiparametric estimation has been mainly concerned with the estimation of a standard ordered choice model without a selection mechanism.

We generalize the SNP estimator by Stewart (2004) to our baseline model with fixed thresholds.

- This is a straightforward generalization of the SNP estimator for bivariate binary choice models proposed by De Luca and Peracchi (2007).
- Our estimator accounts for problems of sample selectivity without requiring strong parametric assumptions on the error terms distribution.

Nonparametric specification of the outcome probabilities

If we denote by F the joint distribution function of (U_1, U_2) and by F_j the marginal distribution function of U_j, then

$$\begin{aligned}\pi_0(\theta) &= F_1(-\mu_1),\\ \pi_h(\theta) &= F_2(\alpha_h - \mu_2) - F(-\mu_1, \alpha_h - \mu_2)\\ &\quad - [F_2(\alpha_{h-1} - \mu_2) - F(-\mu_1, \alpha_{h-1} - \mu_2)],\end{aligned} \tag{12}$$

with $h = 1, \ldots, H$, $\theta = (\beta_1, \beta_2, \alpha)$ and $\mu_j = X_j^\top \beta_j$.

SNP Model - Gallant & Nychka(1987)

The basic idea of the SNP estimators is that of approximating the unknown density of U_1 and U_2 by an Hermite polynomial expansion of the form

$$f^*(u_1, u_2; \tau) = \frac{1}{\psi_R(\tau)} \, \tau_R(u_1, u_2; \tau)^2 \, \phi(u_1) \, \phi(u_2), \tag{13}$$

where

- $\tau = (\tau_{11}, \ldots, \tau_{R_1 R_2})$ is a $(R_1 \times R_2)$–vector of unknown parameters,
- $\tau_R(u_1, u_2; \tau) = 1 + \sum_{h=1}^{R_1} \sum_{k=1}^{R_2} \tau_{hk} u_1^h u_2^k$ is a polynomial in u_1 and u_2 of order $R = (R_1, R_2)$,
- $\psi_R(\tau)$ is a normalization factor to ensure that f^* is a proper density.

As shown by Gallant and Nychka (1987), the class of densities that can be approximated by this polynomial expansion is very large and includes densities with any form of skewness and kurthosis.

Analytical approximations

De Luca and Peracchi (2007) derive by integration the following analytical approximations

- f_1^* and f_2^* to the **marginal densities** f_1 and f_2,
- F^* to the **bivariate cdf** F,
- F_1^* and F_2^* to the **marginal cdf's** F_1 and F_2.

The approximations F^*, F_1^* and F_2^* are all that is needed to approximate the nonparametric outcome probabilities in (12).

Compared with the SNP routines by De Luca (2008), our estimator is more computational demanding because F^* and F_2^* must be evaluated at H different points instead of a single point.

We used MATA to significantly speed up the bivariate SNP routine provided by De Luca (2008).

Estimation & asymptotic properties

The SNP estimator of $\theta = (\beta_1, \beta_2, \alpha, \tau)$ is obtained by maximizing the pseudo-likelihood function in (5) where F, F_1 and F_2 are replaced by their approximations F^*, F_1^* and F_2^*.

In principle, the resulting estimator is $\sqrt{n}$-consistent provided that both R_1 and R_2 increase with sample size.

In practice, for a given sample size, inference is conducted conditional on fixed values of R_1 and R_2 that are selected on the basis of standard model selection criteria (LRT, AIC, BIC, CV).

Thus, the SNP model is treated as a flexible parametric model and it is estimated in a standard ML environment.

5 STATA commands

We provide three new Stata command:

- `opsel` fits a parametric sample selected ordered probit model with constant thresholds coefficients.
- `opselth` fits a parametric sample selected ordered probit model with individual specific thresholds coefficients.
- `snpopsel` fits a semi-nonparametric sample selected ordered choice model with constant thresholds coefficients.

The general syntax of these commands is as follows

`opsel` *equation1* [*weight*] [*if*] [*in*] , `select(`*equation2*`)` [`robust` `from(`*matname*`)` `level(`#`)` *maximize_options*]

`opselth` *equation1* [*weight*] [*if*] [*in*] , `select(`*equation2*`)` [`thrcovariate(`*varlist*`)` `vignette(`*equation3*`)` `re` `robust` `from(`*matname*`)` `level(`#`)` *maximize_options*]

`snpopsel` *equation1* [*weight*] [*if*] [*in*] , `select(`*equation2*`)` [`order1(`#`)` `order2(`#`)` `robust` `dplot(`*filename*`)` `from(`*matname*`)` `level(`#`)` *maximize_options*]

6 Empirical application

We used the first wave of the SHARE data to study determinant of **Depression problems** across 9 European countries.

SHARE data

- Target population: 50+ individuals.
- Structure of the interview: CAPI interview + self-administered questionnaire (Drop-off or Vignette).
- Selection mechanism: nonresponse of the vignette questionnaire is about 23 percent.

```
. describe 'main_variables'
               storage  display     value
variable name    type   format      label      variable label
-------------------------------------------------------------------------------
Resp            float  %9.0g                  Response indicator
Depression      byte   %8.0g       v6         Depression self-assessment
Depression_V1   byte   %8.0g       v25        Depression Vignette 1
Depression_V2   byte   %8.0g       v23        Depression Vignette 2
Depression_V3   byte   %8.0g       v21        Depression Vignette 3
Female          byte   %8.0g       female     Female dummy
Age             byte   %9.0g                  Age
Education       byte   %8.0g                  Year of education
Couple          byte   %9.0g                  Living with spouse or partner
Income          float  %9.0g                  Log per-capita income
Numeracy        byte   %8.0g                  Numeracy indicator
Fluency         byte   %8.0g       cf010_     Verbal fluency score
Recall          byte   %8.0g       cf016tot   Ten words list learning
Heart_att       byte   %8.0g       ph006d01   Heart attack dummy
Cancer          byte   %8.0g       ph006d10   Cancer dummy
Ulcer           byte   %8.0g       ph006d11   Ulcer dummy
Arthritis       byte   %8.0g       ph006d08   Arthritis dummy
Bmi             float  %9.0g                  Body Mass Index
Be              byte   %9.0g                  Country dummy: Belgium
De              byte   %9.0g                  Country dummy: Germany
Es              byte   %9.0g                  Country dummy: Spain
Gr              byte   %9.0g                  Country dummy: Greece
It              byte   %9.0g                  Country dummy: Italy
Fr              byte   %9.0g                  Country dummy: France
Nl              byte   %9.0g                  Country dummy: Netherland
Sw              byte   %9.0g                  Country dummy: Sweden
Iv_female       byte   %9.0g                  Interviewer female
Iv_age          byte   %9.0g                  Interviewer age
Iv_educ         byte   %8.0g                  Interviewer year of education
Int_home        byte   %9.0g       yesno      Interview done at the
                                                respondent's home
Int_afc         byte   %9.0g                  Asked for clarification during
                                                the interview
Int_duq         byte   %9.0g                  Difficulties to understand
                                                questions during the interview
```

ORDERED PROBIT ESTIMATES

```
. oprobit Depression ‘predictors_depression’, nolog
Iteration 0:   log likelihood = -2590.2867
Iteration 1:   log likelihood =  -2429.286
Iteration 2:   log likelihood = -2427.1703
Iteration 3:   log likelihood = -2427.1662
Ordered probit regression                         Number of obs   =       3988
                                                  LR chi2(21)     =     326.24
                                                  Prob > chi2     =     0.0000
Log likelihood = -2427.1662                       Pseudo R2       =     0.0630
```

Depression	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Female	.2555797	.0488502	5.23	0.000	.1598349	.3513244
Age	-.1002231	.0313856	-3.19	0.001	-.1617378	-.0387084
Age2	.0006671	.0002377	2.81	0.005	.0002012	.001133
Education	-.0113293	.0065522	-1.73	0.084	-.0241714	.0015128
Couple	-.2020253	.0525743	-3.84	0.000	-.305069	-.0989815
Income	-.044331	.0222328	-1.99	0.046	-.0879065	-.0007554
Numeracy	-.0484411	.0253373	-1.91	0.056	-.0981013	.0012191
Fluency	-.0018849	.0041286	-0.46	0.648	-.0099767	.0062069
Recall	-.0572161	.0137178	-4.17	0.000	-.0841025	-.0303298
Heart_att	.2257344	.0677936	3.33	0.001	.0928614	.3586073
Cancer	.412443	.0894593	4.61	0.000	.2371061	.5877799
Ulcer	.2644155	.0924341	2.86	0.004	.0832481	.445583
Arthritis	.2960677	.0548926	5.39	0.000	.1884802	.4036552
Bmi	.0120502	.0051369	2.35	0.019	.001982	.0221184
Be	-.0650164	.0829868	-0.78	0.433	-.2276675	.0976348
De	.1789262	.0885434	2.02	0.043	.0053843	.352468
Es	-.0201746	.0893131	-0.23	0.821	-.195225	.1548758
Gr	.2818607	.0794337	3.55	0.000	.1261736	.4375478
It	.0677741	.0896302	0.76	0.450	-.1078979	.2434461
Nl	-.3534414	.0994058	-3.56	0.000	-.5482732	-.1586096
Sw	.5034566	.0902287	5.58	0.000	.3266116	.6803015
/cut1	2.514114	.7360755			1.071433	3.956795
/cut2	3.286234	.7368596			1.842015	4.730452

SNEOP ESTIMATES (Stewart 2004)

```
. sneop Depression ‘predictors_depression’ , order(3) nolog
SNP Estimation of Extended Ordered Probit Model   Number of obs   =       3988
                                                  Wald chi2(21)   =     276.60
Log likelihood = -2422.7219                       Prob > chi2     =     0.0000

------------------------------------------------------------------------------
  Depression |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Female |   .3420927   .0883448     3.87   0.000     .1689401    .5152453
         Age |  -.1516375   .0156396    -9.70   0.000    -.1822907   -.1209844
        Age2 |   .0009982   .0001124     8.88   0.000     .0007779    .0012185
   Education |  -.0218237     .01085    -2.01   0.044    -.0430894    -.000558
      Couple |  -.3558346   .1031736    -3.45   0.001     -.558051   -.1536181
      Income |  -.0628891   .0346305    -1.82   0.069    -.1307637    .0049855
    Numeracy |  -.0701948   .0396038    -1.77   0.076    -.1478169    .0074273
     Fluency |  -.0032831   .0063323    -0.52   0.604    -.0156943     .009128
      Recall |  -.0827437   .0251564    -3.29   0.001    -.1320494   -.0334381
   Heart_att |   .3993993   .1204811     3.32   0.001     .1632608    .6355379
      Cancer |    .617471   .1614144     3.83   0.000     .3011047    .9338374
       Ulcer |   .3966443   .1422177     2.79   0.005     .1179027    .6753858
   Arthritis |   .4406035   .0993848     4.43   0.000     .2458129     .635394
         Bmi |   .0189622   .0085877     2.21   0.027     .0021307    .0357936
          Be |  -.0580071   .1276143    -0.45   0.649    -.3081267    .1921124
          De |   .3519141   .1515091     2.32   0.020     .0549617    .6488666
          Es |   .0236211   .1448331     0.16   0.870    -.2602465    .3074887
          Gr |   .5047906    .143543     3.52   0.000     .2234514    .7861297
          It |   .2181361    .148231     1.47   0.141    -.0723913    .5086635
          Nl |   -.491965    .161028    -3.06   0.002     -.807574   -.1763559
          Sw |   .8212817   .1896703     4.33   0.000     .4495347    1.193029
-------------+----------------------------------------------------------------
Thresholds 1 |   2.514114      Fixed
           2 |   3.708243   .1951454    19.00   0.000     3.325765    4.090721
-------------+----------------------------------------------------------------
SNP coefs: 1 |   -.216212    .056312    -3.84   0.000    -.3265814   -.1058425
           2 |   .3631399   .1069436     3.40   0.001     .1535343    .5727455
           3 |  -.1105704   .0582851    -1.90   0.058     -.224807    .0036663
------------------------------------------------------------------------------
Likelihood ratio test of OP model against SNP extended model:
Chi2(1) statistic =      8.888549      (p-value =  .0028696)
------------------------------------------------------------------------------
Estimated moments of error distribution:
Variance =             1.691049      Standard Deviation =      1.300403
3rd moment =            .832315      Skewness =                .3784893
4th moment =           7.699837      Kurtosis =                2.692585
------------------------------------------------------------------------------
```

A LRT rejects the Gaussian assumption for the marginal distribution of U_2.

PROBIT ESTIMATES

```
. probit Resp ‘predictors_response’, nolog
Probit regression                               Number of obs   =       5052
                                                LR chi2(28)     =     474.26
                                                Prob > chi2     =     0.0000
Log likelihood = -2363.4542                     Pseudo R2       =     0.0912
```

Resp	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Female	.0443005	.0449789	0.98	0.325	-.0438565	.1324576
Age	.0775479	.0292711	2.65	0.008	.0201776	.1349182
Age2	-.0006184	.0002204	-2.81	0.005	-.0010505	-.0001864
Education	.0034673	.0059117	0.59	0.558	-.0081195	.015054
Couple	.0416932	.0506499	0.82	0.410	-.0575787	.1409651
Income	.0221285	.0202153	1.09	0.274	-.0174926	.0617497
Numeracy	.047422	.0234705	2.02	0.043	.0014206	.0934235
Fluency	.0091942	.003564	2.58	0.010	.002209	.0161795
Recall	.0355678	.0128881	2.76	0.006	.0103076	.0608281
Heart_att	.0467097	.0650326	0.72	0.473	-.0807518	.1741712
Cancer	-.0345096	.0892269	-0.39	0.699	-.2093912	.1403719
Ulcer	-.0166674	.0925371	-0.18	0.857	-.1980369	.164702
Arthritis	-.0191385	.0520441	-0.37	0.713	-.1211431	.082866
Bmi	-.0032935	.0047727	-0.69	0.490	-.0126478	.0060609
Be	.1914947	.0687669	2.78	0.005	.0567142	.3262753
De	-.0118648	.0775302	-0.15	0.878	-.1638212	.1400917
Es	.8383638	.0906079	9.25	0.000	.6607757	1.015952
Gr	1.570667	.1452673	10.81	0.000	1.285948	1.855386
It	.5457544	.0809563	6.74	0.000	.387083	.7044259
Nl	.5050427	.0825242	6.12	0.000	.3432982	.6667873
Sw	.5918062	.0945888	6.26	0.000	.4064155	.777197
Iv_female	-.0093097	.0472878	-0.20	0.844	-.1019922	.0833727
Iv_age	.0044314	.0025201	1.76	0.079	-.000508	.0093708
Iv_age2	-.0002332	.0001341	-1.74	0.082	-.000496	.0000297
Iv_educ	.0173886	.0092213	1.89	0.059	-.0006849	.0354621
Int_home	.2837169	.1316416	2.16	0.031	.0257042	.5417296
Int_afc	-.1801218	.0827431	-2.18	0.029	-.3422952	-.0179484
Int_duq	-.2381832	.0890168	-2.68	0.007	-.412653	-.0637134
_cons	1.609192	.6946259	2.32	0.021	.2477506	2.970634

SNP ESTIMATES (De Luca 2008)

```
. snp Resp ‘predictors_response’, order(3) nolog
SNP Estimation of Binary-Choice Model              Number of obs   =      5052
                                                   Wald chi2(28)   =    230.21
Log likelihood = -2358.5284                        Prob > chi2     =    0.0000
```

Resp	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Resp						
Female	.0850976	.0474901	1.79	0.073	-.0079812	.1781765
Age	.1186613	.0111825	10.61	0.000	.0967439	.1405786
Age2	-.0009171	.0000805	-11.39	0.000	-.001075	-.0007592
Education	.0023852	.0059022	0.40	0.686	-.0091829	.0139534
Couple	.0606559	.0512204	1.18	0.236	-.0397342	.161046
Income	.0160659	.0197113	0.82	0.415	-.0225675	.0546993
Numeracy	.0530117	.0253282	2.09	0.036	.0033693	.1026541
Fluency	.0105391	.0042772	2.46	0.014	.002156	.0189222
Recall	.0376825	.0144699	2.60	0.009	.0093221	.066043
Heart_att	.0314388	.065426	0.48	0.631	-.0967939	.1596715
Cancer	-.0487122	.0902612	-0.54	0.589	-.2256208	.1281965
Ulcer	-.0068388	.0927953	-0.07	0.941	-.1887144	.1750367
Arthritis	-.013973	.0528335	-0.26	0.791	-.1175248	.0895788
Bmi	-.0047818	.004901	-0.98	0.329	-.0143875	.004824
Be	.1700677	.0703657	2.42	0.016	.0321535	.3079818
De	-.0329154	.0716218	-0.46	0.646	-.1732917	.1074608
Es	1.003207	.2056473	4.88	0.000	.6001454	1.406268
Gr	3.779515	.5223575	7.24	0.000	2.755714	4.803317
It	.5126377	.1130683	4.53	0.000	.2910279	.7342475
Nl	.4629334	.1130496	4.09	0.000	.2413602	.6845067
Sw	.6335128	.1789543	3.54	0.000	.2827688	.9842568
Iv_female	.0183094	.0468021	0.39	0.696	-.0734209	.1100397
Iv_age	.0049834	.002647	1.88	0.060	-.0002047	.0101714
Iv_age2	-.0003012	.0001403	-2.15	0.032	-.0005761	-.0000263
Iv_educ	.0199313	.0100652	1.98	0.048	.0002039	.0396587
Int_home	.32516	.1341219	2.42	0.015	.062286	.5880341
Int_afc	-.1963799	.0825964	-2.38	0.017	-.3582657	-.034494
Int_duq	-.1948027	.0906001	-2.15	0.032	-.3723757	-.0172298
_cons	1.609192	Fixed				
SNP coefs: 1	.8743181	.5688327	1.54	0.124	-.2405735	1.98921
2	-.2904175	.0707551	-4.10	0.000	-.4290949	-.1517401
3	-.3214101	.1261423	-2.55	0.011	-.5686445	-.0741758

```
Likelihood ratio test of Probit model against SNP model:
Chi2(1) statistic =      9.851559      (p-value =  .0016969)

Estimated moments of error distribution:
Variance =              3.312728      Standard Deviation =      1.82009
3rd moment =           -3.149483      Skewness =              -.5223488
4th moment =            36.90722      Kurtosis =                 3.3631
```

A LRT rejects the Gaussian assumption for the marginal distribution of U_1. Accordingly, we reject the Gaussian assumption for joint distribution of (U_1, U_2)

SAMPLE SELECTED ORDERED PROBIT

```
. opsel Depression ‘predictors_depression’, select(Resp=‘predictors_response’) nolog
```

```
oprobit with sample selection                     Number of obs   =       5052
                                                  Wald chi2(28)   =     404.79
Log likelihood = -4785.9468                       Prob > chi2     =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Resp						
Female	.0422228	.0448718	0.94	0.347	-.0457243	.1301698
Age	.0705579	.0290094	2.43	0.015	.0137006	.1274153
Age2	-.0005676	.0002182	-2.60	0.009	-.0009953	-.0001399
Education	.0027181	.0058585	0.46	0.643	-.0087644	.0142006
Couple	.0313134	.0507334	0.62	0.537	-.0681222	.130749
Income	.0236014	.0201799	1.17	0.242	-.0159505	.0631532
Numeracy	.0501842	.0234047	2.14	0.032	.0043119	.0960565
Fluency	.009465	.0035339	2.68	0.007	.0025386	.0163914
Recall	.0302413	.012876	2.35	0.019	.0050048	.0554777
Heart_att	.0488032	.0648283	0.75	0.452	-.0782578	.1758643
Cancer	-.0384383	.0888111	-0.43	0.665	-.2125048	.1356282
Ulcer	-.0214096	.0918523	-0.23	0.816	-.2014368	.1586175
Arthritis	-.0188678	.0518604	-0.36	0.716	-.1205124	.0827768
Bmi	-.0026742	.0047413	-0.56	0.573	-.0119669	.0066185
Be	.2005433	.0685259	2.93	0.003	.0662349	.3348517
De	-.0033055	.0772961	-0.04	0.966	-.1548031	.148192
Es	.8474469	.0904227	9.37	0.000	.6702217	1.024672
Gr	1.624603	.1421723	11.43	0.000	1.345951	1.903256
It	.5473331	.0802542	6.82	0.000	.3900378	.7046285
Nl	.5113241	.0825559	6.19	0.000	.3495175	.6731307
Sw	.5992546	.0943902	6.35	0.000	.4142532	.784256
Iv_female	.0058937	.0465798	0.13	0.899	-.085401	.0971885
Iv_age	.0059676	.0024775	2.41	0.016	.0011119	.0108234
Iv_age2	-.0002892	.0001317	-2.20	0.028	-.0005474	-.0000311
Iv_educ	.0207292	.0090487	2.29	0.022	.0029941	.0384643
Int_home	.355853	.1297213	2.74	0.006	.1016039	.6101021
Int_afc	-.1892702	.0804071	-2.35	0.019	-.3468652	-.0316752
Int_duq	-.2421075	.0870216	-2.78	0.005	-.4126666	-.0715484
_cons	1.385376	.6877518	2.01	0.044	.0374076	2.733345

(*Continued on next page*)

Depression						
Female	.2526661	.0467943	5.40	0.000	.1609509	.3443812
Age	-.0739322	.0303707	-2.43	0.015	-.1334577	-.0144068
Age2	.0004666	.0002297	2.03	0.042	.0000164	.0009168
Education	-.0096527	.0062551	-1.54	0.123	-.0219125	.0026071
Couple	-.1755344	.0508041	-3.46	0.001	-.2751086	-.0759601
Income	-.0356816	.0212246	-1.68	0.093	-.077281	.0059179
Numeracy	-.034707	.0243297	-1.43	0.154	-.0823922	.0129783
Fluency	.0007155	.0039448	0.18	0.856	-.0070162	.0084472
Recall	-.0456305	.013312	-3.43	0.001	-.0717216	-.0195395
Heart_att	.2205001	.0648538	3.40	0.001	.0933889	.3476113
Cancer	.3830455	.0856163	4.47	0.000	.2152406	.5508504
Ulcer	.2447054	.0884874	2.77	0.006	.0712734	.4181374
Arthritis	.272961	.0527791	5.17	0.000	.1695158	.3764062
Bmi	.0105323	.0049036	2.15	0.032	.0009214	.0201432
Be	-.0048574	.0788268	-0.06	0.951	-.1593551	.1496404
De	.1854824	.0833758	2.22	0.026	.0220689	.348896
Es	.1594108	.0887949	1.80	0.073	-.014624	.3334456
Gr	.5077336	.0800716	6.34	0.000	.3507961	.664671
It	.1976268	.0865598	2.28	0.022	.0279727	.367281
Nl	-.2172447	.0970253	-2.24	0.025	-.4074108	-.0270785
Sw	.6028789	.0868161	6.94	0.000	.4327225	.7730354
Thresholds:						
/cut1	2.304863	.7029817	3.28	0.001	.9270445	3.682682
/cut2	3.034511	.7050969	4.30	0.000	1.652546	4.416476
/athrho	.8300984	.2668969	3.11	0.002	.30699	1.353207
rho	.6805288	.1432918			.2976963	.8748081

LR test of indep. eqns. (rho = 0): chi2(1) = 9.35 Prob > chi2 = 0.0022

According to the parametric model there is a positive and strongly significant selectivity effect.

SNP - SAMPLE SELECTED ORDERED CHOICE MODEL

```
. snpopsel Depression ‘predictors_depression’, select(Resp=‘predictors_response’) ///
order1(3) order2(3) nolog dplot(Depression)
SNP oprobit with sample selection                       Number of obs   =       5052
                                                        Wald chi2(28)   =     136.46
Log likelihood = -4778.5462                             Prob > chi2     =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Resp						
Female	.0811668	.0640688	1.27	0.205	-.0444058	.2067393
Age	.089961	.0236852	3.80	0.000	.043539	.1363831
Age2	-.0007228	.0001842	-3.92	0.000	-.0010838	-.0003618
Education	.0014684	.0081735	0.18	0.857	-.0145513	.0174881
Couple	.0625062	.0696909	0.90	0.370	-.0740854	.1990977
Income	.0269049	.0288216	0.93	0.351	-.0295845	.0833943
Numeracy	.0747514	.0353191	2.12	0.034	.0055273	.1439755
Fluency	.0162641	.005521	2.95	0.003	.0054432	.027085
Recall	.0453201	.0201513	2.25	0.025	.0058243	.0848159
Heart_att	.0439783	.0904012	0.49	0.627	-.1332048	.2211614
Cancer	-.0809667	.1229418	-0.66	0.510	-.3219282	.1599947
Ulcer	.0017273	.1279123	0.01	0.989	-.2489762	.2524308
Arthritis	-.0242151	.07548	-0.32	0.748	-.1721532	.1237229
Bmi	-.0045705	.006968	-0.66	0.512	-.0182275	.0090865
Be	.2751613	.1056756	2.60	0.009	.068041	.4822817
De	.0125617	.1031817	0.12	0.903	-.1896708	.2147941
Es	1.268681	.2664907	4.76	0.000	.7463688	1.790993
Gr	3.021355	.36275	8.33	0.000	2.310378	3.732332
It	.751133	.1842058	4.08	0.000	.3900963	1.11217
Nl	.6761311	.175772	3.85	0.000	.3316243	1.020638
Sw	.8390023	.2344596	3.58	0.000	.3794699	1.298535
Iv_female	.0325488	.0646155	0.50	0.614	-.0940953	.1591929
Iv_age	.008111	.0036077	2.25	0.025	.0010401	.0151819
Iv_age2	-.0003874	.0002013	-1.92	0.054	-.0007819	7.07e-06
Iv_educ	.0287211	.0134891	2.13	0.033	.0022829	.0551594
Int_home	.5334485	.1891887	2.82	0.005	.1626454	.9042516
Int_afc	-.2737216	.1243663	-2.20	0.028	-.517475	-.0299681
Int_duq	-.3719013	.1541008	-2.41	0.016	-.6739333	-.0698693

(*Continued on next page*)

Depression						
Female	.2896597	.0927033	3.12	0.002	.1079646	.4713548
Age	-.0803036	.0203834	-3.94	0.000	-.1202543	-.040353
Age2	.0005018	.0001425	3.52	0.000	.0002226	.0007811
Education	-.0113432	.0080069	-1.42	0.157	-.0270363	.00435
Couple	-.2343764	.0857356	-2.73	0.006	-.4024152	-.0663377
Income	-.0468131	.0274396	-1.71	0.088	-.1005937	.0069675
Numeracy	-.0523231	.0313714	-1.67	0.095	-.1138099	.0091637
Fluency	-.000231	.0047504	-0.05	0.961	-.0095417	.0090796
Recall	-.0479577	.0209285	-2.29	0.022	-.0889769	-.0069385
Heart_att	.278413	.1016694	2.74	0.006	.0791445	.4776814
Cancer	.4418695	.1507948	2.93	0.003	.1463171	.7374219
Ulcer	.3270908	.1297249	2.52	0.012	.0728346	.5813469
Arthritis	.3216127	.1029851	3.12	0.002	.1197656	.5234599
Bmi	.0102584	.0067922	1.51	0.131	-.0030541	.0235708
Be	-.0066066	.0899833	-0.07	0.941	-.1829706	.1697575
De	.2147648	.107916	1.99	0.047	.0032532	.4262763
Es	.1567523	.1130407	1.39	0.166	-.0648034	.3783079
Gr	.5249029	.2206698	2.38	0.017	.092398	.9574079
It	.2512606	.1174173	2.14	0.032	.0211269	.4813943
Nl	-.256891	.1240597	-2.07	0.038	-.5000435	-.0137384
Sw	.7249231	.2038873	3.56	0.000	.3253115	1.124535
Intercept:						
_cons1	1.385376	Fixed				
Thresholds:						
/cut1	2.304863	Fixed				
/cut2	3.166303	.2154854	14.69	0.000	2.743959	3.588646
SNP coefs:						
g_1_1	.7853773	.3739819	2.10	0.036	.0523863	1.518368
g_1_2	-.1363932	.2721251	-0.50	0.616	-.6697486	.3969622
g_1_3	-.0325345	.1097879	-0.30	0.767	-.2477149	.1826458
g_2_1	.0413806	.1793519	0.23	0.818	-.3101428	.3929039
g_2_2	-.0030359	.0427787	-0.07	0.943	-.0868805	.0808088
g_2_3	-.0229809	.0378333	-0.61	0.544	-.0971327	.0511709
g_3_1	-.1522812	.1342403	-1.13	0.257	-.4153873	.1108249
g_3_2	.1101614	.0530694	2.08	0.038	.0061472	.2141756
g_3_3	-.0023378	.0325391	-0.07	0.943	-.0661133	.0614376

```
Estimated moments of errors distribution
    Main equation                           Selection equation
    Standard Deviation =  1.42271           Standard Deviation = 1.688109
    Variance =           2.024105           Variance =           2.849711
    Skewness =          -.0107897           Skewness =          -.1015669
    Kurtosis =           2.524022           Kurtosis =           2.892402

Estimated correlation coefficient
    rho =                .0085209

(file Depression.gph saved)
```

To compare estimates of these different models, we set the coefficient of the `Age` variable equal to $-.1$ in the selection equation and to $.1$ in the outcome equation by using the `nlcom` command.

Here results for the selection equation...

Variable	probit_c	snp_c	opsel_sel_c	snpopsel_s~c
Female	0.057	0.072	0.060	0.090
Age2	-0.001***	-0.001***	-0.001***	-0.001***
Education	0.004	0.002	0.004	0.002
Couple	0.054	0.051	0.044	0.069
Income	0.029	0.014	0.033	0.030
Numeracy	0.061	0.045*	0.071	0.083*
Fluency	0.012	0.009*	0.013	0.018**
Recall	0.046	0.032**	0.043	0.050*
Heart_att	0.060	0.026	0.069	0.049
Cancer	-0.045	-0.041	-0.054	-0.090
Ulcer	-0.021	-0.006	-0.030	0.002
Arthritis	-0.025	-0.012	-0.027	-0.027
Bmi	-0.004	-0.004	-0.004	-0.005
Be	0.247	0.143*	0.284	0.306**
De	-0.015	-0.028	-0.005	0.014
Es	1.081*	0.845***	1.201*	1.410***
Gr	2.025**	3.185***	2.303*	3.359***
It	0.704*	0.432***	0.776*	0.835***
Nl	0.651*	0.390***	0.725*	0.752***
Sw	0.763*	0.534***	0.849*	0.933***
Iv_female	-0.012	0.015	0.008	0.036
Iv_age	0.006	0.004	0.008	0.009*
Iv_age2	-0.000	-0.000*	-0.000	-0.000
Iv_educ	0.022	0.017	0.029	0.032*
Int_home	0.366	0.274*	0.504	0.593***
Int_afc	-0.232	-0.165*	-0.268	-0.304*
Int_duq	-0.307	-0.164*	-0.343	-0.413*

legend: * p<0.05; ** p<0.01; *** p<0.001

Here results for the main equation...

Variable	op_c	sneop_c	opsel_Dep_c	snpopsel_D~c
Female	0.255**	0.226**	0.342*	0.361**
Age2	0.001***	0.001***	0.001***	0.001***
Education	-0.011	-0.014*	-0.013	-0.014
Couple	-0.202*	-0.235***	-0.237	-0.292**
Income	-0.044	-0.041	-0.048	-0.058
Numeracy	-0.048	-0.046	-0.047	-0.065
Fluency	-0.002	-0.002	0.001	-0.000
Recall	-0.057*	-0.055**	-0.062*	-0.060*
Heart_att	0.225*	0.263**	0.298*	0.347**
Cancer	0.412**	0.407***	0.518*	0.550**
Ulcer	0.264*	0.262**	0.331	0.407*
Arthritis	0.295**	0.291***	0.369*	0.400**
Bmi	0.012	0.013*	0.014	0.013
Be	-0.065	-0.038	-0.007	-0.008
De	0.179	0.232*	0.251	0.267*
Es	-0.020	0.016	0.216	0.195
Gr	0.281*	0.333**	0.687*	0.654*
It	0.068	0.144	0.267	0.313
Nl	-0.353*	-0.324**	-0.294	-0.320*
Sw	0.502**	0.542***	0.815*	0.903**
cut1	2.509***		3.118***	
cut2	3.279***	2.445***	4.104***	3.943***

legend: * p<0.05; ** p<0.01; *** p<0.001

Our estimator accounts for both departure from the Gaussian distributional assumption and selectivity effect due to nonresponse.

7 Conclusions

In this paper, we provide 3 new Stata commands for estimation of sample selected ordered probit model

- `opsel` fits a parametric sample selected ordered probit model with constant thresholds coefficients.
- `opselth` fits a parametric sample selected ordered probit model with individual specific thresholds coefficients.
- `snpopsel` fits a semi-nonparametric sample selected ordered choice model with constant thresholds coefficients.

Improvements and extensions:

- Combining SNP with individual heterogeneity.
- Individual heterogeneity: random coefficient model for the slope coefficient.
- SNP: Cross Validation routine for optimal choice of R_1 and R_2.
- Routines for predicted probabilities and marginal effects.
- Empirical application...to be completed