National Research University Higher School of Economics

as a manuscript

Potanin Bogdan Stanislavovich

Multivariate Generalization of regression models with endogenous switch and nonrandom selection

PhD Dissertation Summary
for the purpose of obtaining academic degree Doctor of Philosophy
in Economics

Academic supervisor:

lecturer, PhD in Physics and Mathematics Kossova Elena Vladimirovna

JEL: C34

Moscow – 2020

Sample selection as a problem of econometric analysis

A. Roy was the first to rise the problem of selection bias[1]. He investigated the influence of profession selection mechanism on the distribution of income and skills within professional groups. According to the Roy's model individuals select the profession not by chance but based on their skills and demand for them. As a result, distribution of skills and wages among professional groups differs from the corresponding distribution within the whole population. This distribution may depend on the number of professions, skills joint distribution and their effect on wages. This model served as a foundation for many studies related to the parameters identification and estimation under sample selection[2].

Sample selection may cause censoring or truncation depending on the sample generating mechanism. In both cases there is some sample selection rule determining observations inclusion into the sample. The key distinction between truncated and censored samples is that "in a truncated sample one cannot use the available data to estimate the probability that an observation has complete data. In a censored sample, one can." (Heckman J. , 1976, p. 478). Within the regression analysis framework sample selection usually concerns the dependent variable observability. Then truncation undermines that the support of dependent variable given the sample selection rule will be restricted[3]. If the censoring takes place then probability of observation inclusion does not depend on the values of dependent variable but some of its values are unobservable. Classical example states that wages are observable only for employed individuals[4] (Heckman & Killingsworth, 1987). Consequently, the mechanism determining

[1] A. Roy's model concerns particular case of sample selection, namely self-selection which undermines that individual selects the profession following some (usually wage maximization) criteria.
[2] The review of these studies with a focus on their relation to the Roy's model has been provided by (Heckman & Taber, 2010) and (Heckman & Honore, 1990).
[3] The support of discrete (continuous) random variable is the set of values for which probability (density) function takes values greater than zero.
[4] If the sample has been generated based on the survey among employed individuals only, then it is impossible to estimate the probability of inclusion of wages for randomly chosen individual. Then the sample is truncated. If the survey covers the whole population then it is possible to estimate probability of employment and hence the probability of wages observability which makes the sample censored.

employment will govern the rule of observations inclusion into the sample. Note that this thesis concerns the problem of censoring only.

Econometric methods which do not account for the fact that the sample is censored or truncated may provide inconsistent estimators. Wherein samples being subject to sample selection are frequently arise in practice usually due to individuals (or other observational units) self-selection or when “selection decisions by analysts or data processors operate in much the same fashion as self selection” (Heckman J. , 1979, стр. 1), (Хекман, 2013, стр. 130). This obstacle gave a rise to the extensive literature related to the sample selection models implementation: estimation methods accounting for the fact that sample has been generated following a several of selection rules. Let’s consider the main findings and results of these studies which are relevant to the topic of the thesis.

Brief review of sample selection models for multiple rules of observations selection

“Classical sample selection models take into account single selection rule determined by the value of binary variable (Heckman J. , 1979). However, some issues require the consideration of more complicated selection mechanisms. So selection equation also may be ordinal (Kugler, 1987), (Vella, 1993), continuous (Garen, 1984) or categorical (Jeffrey & McFadden, 1984). The last one undermines that selection rule depends on the values of several binary variables. For example, wages are observable only for employed individuals who are additionally willing to reveal their salary information.” (Коссова & Потанин, 2018).

Furthermore, sometimes researchers may be interested in parameters estimation for several forms of the main equation. Wherein the form depends on the combination of selection rules. For example, the wage of individual with (without) higher education will be observable only if he has (not) graduated from the university. Then wage determining equations for both these states may differ since presence of higher education may influence returns to skills. In order to estimate parameters of this model endogenous switching regression may be used (Lee, 1978). This method undermines that we observe one form of dependent variable when particular selection rule has been satisfied and the other form — otherwise. Herewith the need to account for additional selection rules may arise. For example, employment is additional requirement for wages observability independent of the presence of higher education.

“Since likelihood function maximization for the multivariate sample selection models is technically complicated task there are few studies in this field and most of them consider only two independent selection rules (Vella, 1998).” (Коссова & Потанин, 2018). These studies take into account at most three selection rules with continuous (Poirier, 1980), (Cinzia, 2009), (Ogundimu & Hutton, 2016) and binary (Rosenmana, Mandal, Tennekoon, & Hill, 2010) dependent variables (Коссова & Потанин, 2018). In order to account for multiple selection rules some studies apply non-parametric two-step procedures (De Luca & Peracchi, 2012), (Das, Newey, & Vella, 2003) which do not allow to reconstruct random errors joint distribution therefore sufficiently reducing results interpretability. Furthermore (Ogundimu & Hutton, 2016) and (De Luca & Peracchi, 2012) do not provide sufficient arguments for their two-step estimator consistency as well as do not derive its asymptotic distribution and consistent estimator of the covariance matrix. So, it complicates hypothesis testing.

Therefore, there is need for econometric methods providing consistent estimators given multiple sample selection criteria and several forms of the main equation. Wherein this method should allow to reconstruct random errors joint distribution in order to extend interpretability of the results.

Several new sample selection models have been provided in this thesis. The models assume voluntary number of selection equations and several forms of the main equation. These models generalize classical and some semi-parametric approaches simultaneously accounting for sample selection and endogenous switching. “The model has very general setting. The form of the main equation depends on the combination of selection rules. Particularly, for some of them dependent variable is unobservable. Furthermore, there may be no information concerning the values of selection equation which corresponds to the scheme of consecutive decision making.” (Коссова & Потанин, 2018).

Goals and objectives of the research

The thesis concentrates on models with several sample selection rules. The subject of this research covers econometric methods providing estimators for these models' parameters. The goal of the thesis is to implement econometric methods[5] providing estimators for the models with finite voluntary number of sample selection criteria and several forms of the main equation. In order to achieve this goal, the following tasks should be accomplished:

1. Derive some properties of truncated multivariate normal distribution and distribution introduced by (Gallant & Nychka, 1987) in order to get estimators, standard errors, conditional expectations and marginal effects expressions for the methods being introduced in this thesis.

2. Review most popular and influential parametric, semi-parametric and semi-nonparametric sample selection models.

3. Provide sample selection methods accounting for finite voluntary number of selection criteria determining the form of the main equation. Furthermore, these methods estimators' properties should be investigated.

4. Provide software implementation of these methods in order to apply them to the analysis of real and simulated data.

5. Analyze these methods estimators accuracy using simulated data and various specification assumptions, particularly, concerning random errors joint distribution and presence of exclusion restrictions (unique regressors for some equations).

6. Apply these methods to the analysis of real data that is subject to multivariate sample selection.

[5] Further on let's sometimes refer to them as "these methods" for brevity.

Main findings of the thesis

The set of econometric tools for the analysis of samples generated subject to multiple selection rules has been extended through the introduction of new methods generalizing a couple of classical parametric and semi-parametric sample selection models.

First, endogenous switching and Heckman's models have been generalized to the case of voluntary number of selection equations. Both maximum likelihood and two-step implementations of these methods have been generalized. The last one has been equipped with covariance matrix consistent estimator allowing for hypothesis testing. Furthermore, formulas for conditional expectations and marginal effects have been derived for these models.

Second, semi-nonparametric method of (Gallant & Nychka, 1987) has been generalized. This method requires likelihood function maximization under substitution of true density function with its Hermite form approximation. Furthermore, some properties of distribution generated by this approximating function have been derived[6]. These results might be useful for semi-nonparametric estimation of other models' parameters going beyond the scope of sample selection problem.

Third, nonparametric Newey's method has been generalized for the case of finite voluntary number of selection equations. This method approximates random errors' unknown conditional expectation via polynomial of smooth functions which arguments are linear indexes related to selection equations.

The accuracy of methods being provided in this research has been investigated within simulation study framework. The data has been simulated under the assumption of two sample selection rules. Let's highlight main findings of the experiments with simulated data.

First, generalized parametric methods provide notably more accurate results then their classical counterparts and least squares method.

[6] Statistical package "hpa" has been implemented within R software framework. It allows to estimate truncated, conditional and marginal moments as well as density and cumulative distribution functions for the distribution introduced by (Gallant & Nychka, 1987) .

Second, accuracy of generalized parametric two-step procedure sufficiently declines if exclusion restrictions do not hold. The reason is probably the same as for one dimensional case: inverse mills ratios are rather close to linear function causing losses in efficiency due to multicollinearity problem.

Third, generalized parametric and semi-parametric methods are approximately equally accurate when random errors' distribution deviates from normality in terms of tails thickness and asymmetry. However, if the distribution is bimodal then semi-parametric methods demonstrate notable advantage over parametric approaches.

Fourth, generalized Newey's methods' specification with simple one-dimensional inverse mills ratios and interactions between polynomial terms related to different selection equations seems to be equally accurate as specification with generalized mills ratios. Wherein the first of these specifications has much lower computational complexity since it allows to estimate selection equations parameters separately rather than within the system of equations. Note that the other tested specification with simple mills ratios without interactions seems much less accurate than those mentioned above.

Fifth, generalized method of Gallant and Nychka notably inferior in terms of accuracy to other methods provided in thesis if random errors have Student or Beta distribution. However, this method comparative advantage rises substantially if random errors follow bimodal distribution. Note that the study has investigated this method accuracy only for low polynomial orders because of high computational burden associated with likelihood function maximization given corresponding approximating functions. Therefore, increase in polynomial order probably may result in sufficient accuracy boost.

Parametric methods introduced in this thesis have been applied to Russian Longitude Monitoring Survey data (RLMS) in order to estimate returns to education and marriage for males. The results suggest that marriage and higher education are endogenous to the wage equation: conditional distribution of wage equations' random errors depends on probabilities of being married and having higher education.

Estimate of correlation between random errors of marriage and wage equations is significant and positive indicating that there are some unobservable characteristics positively affecting wages and marriage probability.

Furthermore, it has been found that marriage exhibits negative effect on wage probably because unmarried men may spend more time on job search process since they are not obligated to insure their family financial stability. Finally, note that probably because of selection bias estimate of marriage effect on wages for classical Heckman's and least squares methods is positive.

The effect of education on wages has been found to be positive and greater about 1.25-1.5 times (than for least squares or classical Heckman's method) after accounting for selection bias caused by self-selection to universities. It is consistent with previous findings of studies based on United States labor market data. Furthermore, negative correlation between higher education and wage equations has been revealed. It may be caused by the fact that the data being used has lack of differentiation among various levels of higher education related to universities quality.

Finally, wage equation parameters have been estimated under the assumption that there is additional sample selection rule related to the fact that some individuals do not reveal their salary information. The results suggest that omitted data on wages cause selection bias while its effect on estimates values seems to be insufficient.

Main scientific contributions of the research

The thesis makes the following contribution to the methodology of models' parameters estimation under multivariate sample selection:

1. Parametric generalizations of endogenous switching and Heckman models' have been provided based on maximum likelihood and two-step procedures. These generalizations allow to consider finite voluntary number of selection equations and several forms of the main equation. Estimators for marginal effects and conditional expectations for these methods have been provided. These estimators allow to predict

dependent variable’s expected value given various combinations of selection rules. Furthermore, generalized two-step estimator consistency and asymptotic normality has been insured. Finally, consistent covariance matrix estimator for two-step method has been derived.

2. Multivariate generalization to semi-nonparametric (Gallant & Nychka, 1987) and semi-parametric (Newey, 2009) methods have been provided (the last one following the idea of (De Luca & Peracchi, 2012)). Formulas for random errors’ conditional expectations have been derived for the generalization of Gallant and Nychka method. These formulas being coupled with numeric differentiation technics allow to estimate marginal effects and conditional expected values for different forms of the dependent variable.

3. These methods accuracy has been investigated by the means of simulated data analysis. Particularly it has been demonstrated that generalized parametric methods’ estimates are robust to violation of random errors joint normality assumption.

The research contribution related to the application of these methods to the analysis of RLMS data is as follows:

1. Returns to education estimate has been obtained simultaneously accounting for self-selection of individuals into employment and universities. The results suggest that education is endogenous respect to wage equation of Russian males. Without considering self-selection of individuals to universities the effect of higher education seems to be underestimated that is consistent with findings of several studies based on United States labor market data being analyzed via the method of instrumental variables.

2. It has been shown that marriage status seems to be endogenous variable with respect to the wage equation of Russian males. The results suggest that without accounting for marriage endogeneity selection bias upon the marriage effect estimator may become sufficient enough to change the sign of the corresponding estimate.

3. Statistical evidence has been found in favor of the argument that refusal to reveal wage information may cause selection bias. However, this bias seems to be insufficient.

Note that formulas for truncated, conditional and marginal moments as well as density and cumulative distribution functions of distribution introduced by (Gallant &

Nychka, 1987) may be applied to semi-nonparametric generalization of broad class of classical econometric models which parameters are estimated through the maximization of likelihood function value.

The main results for the defense of the thesis

This paragraph briefly introduces new econometric methods being proposed in the thesis. Since the available space is limited some formulas that have been derived in the thesis are omitted.

Sample selection model with multivariate sample selection and endogenous switching

“Consider m sample selection rules depending on the values of binary variables $z_{si}, s \in \{1...m\}$. If the rule s has been satisfied for the i-th observation then $z_{si} = 1$ and $z_{si} = -1$ — otherwise. There are 2^m possible selection rules combinations and for r of them dependent variable y_i^* is observable, where $(1 \le r \le 2^m)$. So, there are r groups of observations indexed by $\{1, \cdots, r\}$. There is also additional group of observations indexed with 0 value. This group has no observations for y_i^* and may be empty. For each observation i let’s define index function $g_i = g(z_{1i}, ..., z_{mi}) = s \ (0 \le s \le r)$ which value equals to the group index of this observation.” (Коссова & Потанин, 2018).

“The data generating process is as follows:” (Коссова & Потанин, 2018).

$$y_i^* = x_i'\beta_{g_i} + \epsilon_{i,g_i},$$

$$z_{si}^* = w_{si}'\gamma_s + u_{si}, s \in \{1...m\},$$

$$z_{si} = \begin{cases} 1, \text{ если } z_{si}^* \ge 0, \text{ т.е. } u_{si} \ge -w_{si}'\gamma_s, \\ -1, \text{ если } z_{si}^* < 0, \text{ т.е. } u_{si} < -w_{si}'\gamma_s. \end{cases}$$

$$g_i = g(z_{1i}, ..., z_{mi}), \ g_i \in \{0...r\},$$
$$u_i' = (u_{1i}, ..., u_{mi})',$$

$$(\epsilon_{i,g_i}, u_i')' \sim N\left(\begin{bmatrix}0\\ \vdots \\ 0\end{bmatrix}, \Sigma_{g_i}\right), \quad где \quad \Sigma_{g_i} = \begin{bmatrix} \sigma_{g_i}^2 & \rho_{1,g_i}\sigma & \rho_{2,g_i}\sigma & \cdots & \rho_{m,g_i}\sigma \\ \rho_{1,g_i}\sigma & 1 & \rho_{12} & \cdots & \rho_{1m} \\ \rho_{2,g_i}\sigma & \rho_{12} & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{m,g_i}\sigma & \rho_{1m} & \cdots & \cdots & 1 \end{bmatrix};$$

$$y_i = \begin{cases} y_i^*, \text{ если } g_i > 0, \\ не \quad наблюдаем, \text{ если } g_i = 0, \end{cases} \quad i \in \{1...n\}.,$$

"where x_i and w_{is} — are vectors of exogeneous variables for the i-th observation of the main equation and selection equation s correspondingly. These variables effects depend on the coefficients vectors β_{g_i} and γ_s. Finally, there are random errors ϵ_{i,g_i} and u_{si} following multivariate normal distribution." (Коссова & Потанин, 2018).

"Let's define the following vectors:" (Коссова & Потанин, 2018).

$$z_i = \begin{bmatrix} z_{1i} \\ \vdots \\ z_{mi} \end{bmatrix}, \; z_i = \begin{bmatrix} \tilde{z}_{1i} \\ \vdots \\ \tilde{z}_{mi} \end{bmatrix} = \begin{bmatrix} z_{1i}w_{1i}'\gamma_1 \\ \vdots \\ z_{mi}w_{mi}'\gamma_m \end{bmatrix}, \; u_i = \begin{bmatrix} -z_{1i}u_{1i} \\ \vdots \\ -z_{mi}u_{mi} \end{bmatrix} \text{ и } \rho_{g_i} = \begin{bmatrix} \rho_{1,g_i} \\ \vdots \\ \rho_{m,g_i} \end{bmatrix}.$$

According to the properties of truncated multivariate normal distribution conditional expected value and variance of y_i has the following form:

$$E(y_i \mid z_{1i},...,z_{mi}) = x_i'\beta_{g_i} + \sum_{j=1}^{m} \sigma_{g_i}\rho_{j,g_i}\lambda_j^{\tilde{u}_i}(\tilde{z}_i)z_{ji},$$

$$D(y_i \mid z_{1i},...,z_{mi}) = \sigma_{g_i}^2(1 - \sum_{k=1}^{m}\rho_{k,g_i}^2\tilde{z}_{ki}\lambda_k^{\tilde{u}_i}(\tilde{z}_i) + \\ + \sum_{k=1}^{m} z_{ki}\rho_{k,g_i}\sum_{j\neq k} z_{ji}(\rho_{j,g_i} - \rho_{kj}\rho_{k,g_i})\Lambda_{kj}^{\tilde{u}_i}(\tilde{z}_i) - (\sum_{k=1}^{m} z_{ki}\rho_{k,g_i}\lambda_k^{\tilde{u}_i}(\tilde{z}_i))^2).$$

where:

$$\lambda^{\tilde{u}_i}(x) = \frac{\nabla F_{\tilde{u}_i}(x)}{F_{\tilde{u}_i}(x)}, \; \Lambda^{\tilde{u}_i}(x) = \frac{H(F_{\tilde{u}_i}(x))}{F_{\tilde{u}_i}(x)}.$$

(Коссова & Потанин, 2018).

The term $\lambda^{\tilde{u}_i}(x)$ represents generalized inverse mills ratio.

Suppose that there is single selection equation, $g(1) > 0$ and $g(-1) = 0$, then this model will be identical to classical Heckman's model. Furthermore if $g(1) > 0$ and $g(1) \neq g(-1) > 0$, then the model coincides with endogenous switching model. In addition when $m=2$ and $g(1,1)>0, g(1,-1)=g(-1,1)=g(-1,-1)=0$ then the model matches the models of (Poirier, 1980) and (De Luca & Peracchi, 2012). Finally, if $m=3$, $g(1,1,1)>0$ and equals zero otherwise then the model is the same as the model proposed by (Ogundimu & Hutton, 2016).

Let’s briefly discuss the estimators of this model parameters introduced in the thesis.

Parametric maximum likelihood estimator

The model parameters could be estimated via maximization of the following likelihood function:

$$L(\Sigma_1,...,\Sigma_r,\beta_1,...,\beta_r,\gamma_1,...,\gamma_m)=\prod_{i=1}^{n} r_i \ ,$$

$$r_i=\begin{cases} F_{u_i|\epsilon_{g_i}=y_i-x_i'\beta_{g_i}}(z_{1i}w_{1i}'\gamma_1,...,z_{mi}w_{mi}'\gamma_m)\, f_{\epsilon_{g_i}}(y_i-x_i'\beta_{g_i}), & \text{для} \quad i:\ g_i>0, \\ F_{u_i}(z_{1i}w_{1i}'\gamma_1,...,z_{mi}w_{mi}'\gamma_m), & \text{для} \quad i:\ g_i=0. \end{cases},$$

where F_{u_i} — is distribution function of $\tilde{u}_i$ and $F_{u_i|\epsilon_{g_i}=y_i-x_i'\beta_{g_i}}$ — is his conditional distribution function given $\epsilon_{g_i}=y_i-x_i'\beta_{g_i}$.

Maximum likelihood estimator will be consistent, efficient and asymptotically normal if random errors follow multivariate normal distribution and some regularity conditions are satisfied. Simulated data analysis provides an evidence that even if random errors normality assumption has been violated then this method estimates are not less accurate than those obtained by nonparametric and semi-parametric multivariate sample selection methods.

Parametric two-step estimator

“According to the theorem proposed by (Murphy & Topel, 1985) least squares estimator will be consistent and asymptotically normal if inverse generalized mills ratios are substituted with their consistent estimators.“ (Коссова & Потанин, 2018).

“Therefore, two-step estimator is as follows:

1. First, apply multivariate probit model in order to estimate γ_s and ρ_{sk}, where $s,k \in \{1...m\}$. Then use maximum likelihood estimator invariance property in order to get consistent estimator $\lambda_k^{\tilde{u}_i}(\tilde{z}_i)$ of $\lambda_j^{\tilde{u}_i}(\tilde{z}_i)$, where $i \in \{1,...,n\}$.

2. Second, apply least squares estimator for each group of observations $\{i : g_i = c\}$ separately:

$$y_i = x_i'\beta_c + \sum_{j=1}^{m} \beta_{\lambda_j}(\lambda_j^{\tilde{u}_i}(\hat{\tilde{z}}_i) z_{ji}) + v_i \text{, где } \beta_{\lambda_j} = \sigma_c \rho_{j,c},$$

where generalized inverse mills ratios $\lambda_j^{\tilde{u}_i}(\tilde{z}_i)$ should be substituted with their consistent estimators $\lambda_j^{\tilde{u}_i}(\hat{\tilde{z}}_i)$ from the previous step and v_i is heteroscedastic random error.” (Коссова & Потанин, 2018).

Variances and correlations consistent estimators are:

$$\hat{\sigma}_c^{\,2} = \frac{1}{n_c}\left(e^T e + \sum_{i=1}^{n_c}\left(\sum_{k=1}^{m} \beta_{\lambda_k}{}^2 \hat{\tilde{z}}_{ki} \hat{\lambda}_k^{\tilde{u}_i}(\hat{\tilde{z}}_i) - \sum_{k=1}^{m} \beta_{\lambda_k} z_{ki} \sum_{j \neq k} z_{ji}(\beta_{\lambda_j} - \hat{\rho}_{kj}\beta_{\lambda_k})\hat{\Lambda}_{kj}^{\tilde{u}_i}(\hat{\tilde{z}}_i) + \left(\sum_{k=1}^{m} z_{ki}\beta_{\lambda_k}\hat{\lambda}_k^{\tilde{u}_i}(\hat{\tilde{z}}_i)\right)^2\right)\right),$$

$$\hat{\rho}_{k,c} = \frac{\hat{\beta}_{\lambda_k}}{\sigma_c}, \quad 1 \le k \le m.$$

Consistent estimator of the two-step estimator’s covariance matrix has the following form:

$$As.Cov(\hat{\tilde{\beta}}) = \hat{\sigma}_c^2 (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \left[(I - \hat{\Delta}) + \hat{\Gamma} As.Cov\left(\begin{bmatrix}\hat{\gamma} \\ \hat{\rho}\end{bmatrix}\right)\hat{\Gamma}^T\right] \tilde{X}(\tilde{X}^T \tilde{X})^{-1},$$

where properties of truncated multivariate normal distribution guarantee that:

$$\Gamma = J(\sum_{k=1}^{m} \rho_{k,c} z_k \circ \lambda_k^{\tilde{u}}(\tilde{z})),$$

$$\Theta^{\tilde{u}_i}(x) = \frac{D^3(F_{\tilde{u}_i}(x))}{F_{\tilde{u}_i}(x)},$$

$$\left(\Gamma_{\gamma_{ks}}\right)_i = W_{ik}(\sum_{j \neq k} \rho_{j,c} z_{ji} z_{ki}(\Lambda_{kj}^{\tilde{u}_i}(\tilde{z}_i) - \lambda_k^{\tilde{u}_i}(\tilde{z}_i)\lambda_j^{\tilde{u}_i}(\tilde{z}_i)) - \rho_{k,c}(\tilde{z}_{ki}\lambda_k^{\tilde{u}_i}(\tilde{z}_i) + (\lambda_k^{\tilde{u}_i}(\tilde{z}_i))^2 + \sum_{j \neq k} \rho_{kj}\Lambda_{kj}^{\tilde{u}_i}(\tilde{z}_i) z_{ki} z_{ji})).,$$

$$\left(\Gamma_{\rho_{lk}}\right)_i = z_{ki}\rho_{l,c}((\frac{z_{li}z_{ki}\rho_{lk}\tilde{z}_{ki}-\tilde{z}_{li}}{1-\rho_{lk}^2}-\lambda_l^{\tilde{u}_i}(\tilde{z}_i))\Lambda_{lk}^{\tilde{u}_i}(\tilde{z}_i)+\sum_{j\neq l,k}\frac{z_{li}z_{ji}(\rho_{lk}\rho_{kj}-\rho_{lj})}{1-\rho_{lk}^2}\Theta_{ljk}^{\tilde{u}_i}(\tilde{z}_i))+$$
$$+z_{li}\rho_{k,c}((\frac{z_{li}z_{ki}\rho_{lk}\tilde{z}_{li}-\tilde{z}_{ki}}{1-\rho_{lk}^2}-\lambda_k^{\tilde{u}_i}(\tilde{z}_i))\Lambda_{lk}^{\tilde{u}_i}(\tilde{z}_i)+\sum_{j\neq l,k}\frac{z_{ki}z_{ji}(\rho_{lk}\rho_{lj}-\rho_{kj})}{1-\rho_{lk}^2}\Theta_{ljk}^{\tilde{u}_i}(\tilde{z}_i))+ \quad ,$$
$$+\sum_{j\neq l,k} z_{ji}z_{li}z_{ki}\rho_{j,c}(\Theta_{ljk}^{\tilde{u}_i}(\tilde{z}_i)-\lambda_j^{\tilde{u}_i}(\tilde{z}_i)\Lambda_{lk}^{\tilde{u}_i}(\tilde{z}_i)).$$

"and $\hat{\Gamma}$ is matrix Γ estimated with $\hat{z},\hat{\rho}$ and $\hat{\sigma}$." (Коссова & Потанин, 2018). Since two-step estimator is asymptotically normal then covariance matrix consistent estimator is sufficient for hypothesis testing.

This two-step estimator will be consistent and asymptotically normal. However, it is less efficient then maximum likelihood estimator described above. Nonetheless simulated data analysis results suggest that this estimator is more robust to random errors' normality assumption violation. However, its estimates accuracy declines substantially if exclusion restrictions are violated.

Semi-nonparametric estimator

Abovementioned parametric estimators' consistency depends on the random errors' normality assumption. If it does not hold then consistent estimator could be obtained through the maximization of the likelihood function where density functions are substituted with approximating functions suggested by (Gallant & Nychka, 1987). For example, if there is no endogenous switching[7] the function to be maximized will have the following form:

$$L(\alpha^{(-1)},\mu,\sigma,\beta,\gamma)=\prod_{i=1}^{n} r_i.$$

$$r_i=\begin{cases}\overline{F}_{u_i|\epsilon_{g_i}=y_i-x_i'\beta}(-w_{1i}'\gamma_1,\ldots,-w_{mi}'\gamma_m;\infty,\cdots,\infty)f_{\epsilon_{g_i}}(y_i-x_i'\beta), & для \quad i:\ g_i>0\\ F_{u_i}(-w_{1i}'\gamma_1,\ldots,-w_{mi}'\gamma_m), & для \quad i:\ g_i=0\end{cases},$$

where formulas derived in thesis insure that:

[7] It means that $(z_{1i}=\cdots=z_{mi}=1)$, $g(1,\cdots,1)>0$ and equals zero otherwise.

$$\bar{F}_{u_i|\epsilon_{g_i}=y_i-x_i'\beta}(-w_{1i}'\gamma_1,\dots,-w_{mi}'\gamma_m;\infty,\cdots,\infty)=\frac{1}{\psi_1}\prod_{r_0=1}^{m}\Phi\left(\frac{w_{r_0 i}'\gamma_{r_0}+\mu_{r_0+1}}{\sigma_{r_0+1}}\right)*$$
$$*\sum_{i_1=0}^{K_1}\cdots\sum_{i_{m+1}=0}^{K_{m+1}}\sum_{j_1=0}^{K_1}\cdots\sum_{j_{m+1}=0}^{K_{m+1}}\alpha_{(i_1,\dots,i_{m+1})}\alpha_{(j_1,\dots,j_{m+1})}\left(y_i-x_i'\beta\right)^{i_1+j_1}\prod_{r_1=2}^{m+1}\mathcal{M}_{TR}\left(i_{r_1}+j_{r_1};-w_{r_1 i}'\gamma_{r_1},\infty,\mu_{r_1},\sigma_{r_1}\right),$$

$$F_{u_i}(-w_{1i}'\gamma_1,\dots,-w_{mi}'\gamma_m)=\frac{1}{\psi_2}\prod_{r_0=1}^{m_d}\left(1-\Phi\left(\frac{w_{r_0 i}'\gamma_{r_0}+\mu_{r_0+1}}{\sigma_{r_0+1}}\right)\right)*$$
$$*\sum_{i_1=0}^{K_1}\cdots\sum_{i_{m+1}=0}^{K_{m+1}}\sum_{j_1=0}^{K_1}\cdots\sum_{j_{m+1}=0}^{K_{m+1}}\alpha_{(i_1,\dots,i_{m+1})}\alpha_{(j_1,\dots,j_{m+1})}\mathcal{M}(i_1+j_1;\mu_1,\sigma_1)\prod_{r_1=2}^{m+1}\mathcal{M}_{TR}\left(i_{r_1}+j_{r_1};-w_{r_1 i}'\gamma_{r_1},\infty,\mu_{r_1},\sigma_{r_1}\right),$$

$$f_{\epsilon_{g_i}}(y_i-x_i'\beta)=\frac{1}{\psi_3}\phi\left(\frac{y_i-x_i'\beta}{\sigma_1}\right)\sum_{i_1=0}^{K_1}\cdots\sum_{i_{m+1}=0}^{K_{m+1}}\sum_{j_1=0}^{K_1}\cdots\sum_{j_{m+1}=0}^{K_{m+1}}\alpha_{(i_1,\dots,i_m)}\alpha_{(j_1,\dots,j_m)}\left(y_i-x_i\beta\right)^{i_1+j_1}*$$
$$*\prod_{r=2}^{m+1}\mathcal{M}(i_{r_2}+j_{r_2};\mu_{r_2},\sigma_{r_2}),$$

and:

$$\psi_1=\sum_{i_1=0}^{K_1}\cdots\sum_{i_{m+1}=0}^{K_{m+1}}\sum_{j_1=0}^{K_1}\cdots\sum_{j_{m+1}=0}^{K_{m+1}}\alpha_{(i_1,\dots,i_{m+1})}\alpha_{(j_1,\dots,j_{m+1})}\left(y_i-x_i'\beta\right)^{i_1+j_1}\prod_{r=2}^{m+1}\mathcal{M}(i_{r_2}+j_{r_2};\mu_{r_2},\sigma_{r_2}),$$

$$\psi_2=\sum_{i_1=0}^{K_1}\cdots\sum_{i_{m+1}=0}^{K_{m+1}}\sum_{j_1=0}^{K_1}\cdots\sum_{j_{m+1}=0}^{K_{m+1}}\alpha_{(i_1,\dots,i_{m+1})}\alpha_{(j_1,\dots,j_{m+1})}\prod_{r=2}^{m+1}\mathcal{M}(i_r+j_r;\mu_r,\sigma_r),$$

$$\psi_3=\sum_{i_1=0}^{K_1}\cdots\sum_{i_{m+1}=0}^{K_{m+1}}\sum_{j_1=0}^{K_1}\cdots\sum_{j_{m+1}=0}^{K_{m+1}}\alpha_{(i_1,\dots,i_{m+1})}\alpha_{(j_1,\dots,j_{m+1})}\prod_{r=1}^{m+1}\mathcal{M}(i_r+j_r;\mu_r,\sigma_r),$$

$$\alpha^{(-1)}=\left(\alpha_{(1,\cdots,0)},\cdots,\alpha_{(K_1,\cdots,K_{m+1})}\right),\ K_1,\cdots,K_{m+1}\in\left(N\cup\{0\}\right),$$

where $\mathcal{M}_{TR}\left(i_r+j_r;\underline{x},\bar{x},\mu_r,\sigma_r\right)$ are moments of the order (i_r+j_r) of the truncated at lower point $\underline{x}$ and upper point $\bar{x}$ normal variable which expected value and variance equal to μ_r and σ_r^2 correspondingly. Note that $\mathcal{M}_{TR}\left(i_r+j_r;\mu_r,\sigma_r\right)=\mathcal{M}_{TR}\left(i_r+j_r;-\infty,\infty,\mu_r,\sigma_r\right)$. Finally, ϕ and Φ are standard normal density and cumulative distribution functions correspondingly.

In order to insure parameters identification it is necessary to set $\alpha(0,\cdots,0)=1$, exclude constants from the selection equations and fix to 1 one of the regression coefficients for each selection equation. Note that preserving β and γ estimators consistency it is possible to fix μ_k and σ_k parameters where $k\in\{1,\cdots,m+1\}$.

While the estimator is consistent its asymptotic distribution remains unknown that complicates hypothesis testing which should be performed via bootstrap procedure. Simulated data analysis reveals that this method has comparative advantage in terms of estimates accuracy if random errors' follow multimodal distribution.

Semi-parametric two-step estimator

This estimator generalizes one proposed by (Newey, 2009).

Conditional expected value of the dependent variable has the form:

$$\begin{aligned}E(y_i) &= E(y_i^* \mid z_{1i},...,z_{mi}) = E(y_i^* \mid -z_{1i}u_{1i} \le z_{1i}w_{1i}'\gamma_1,...,-z_{mi}u_{mi} \le z_{mi}w_{mi}'\gamma_m) = \\ &= x_i'\beta_{g_i} + E(\epsilon_{i,g_i} \mid -z_{1i}u_{1i} \le z_{1i}w_{1i}'\gamma_1,...,-z_{mi}u_{mi} \le z_{mi}w_{mi}'\gamma_m) = \\ &= x_i'\beta_{g_i} + g^*(w_{1i}'\gamma_1,...,w_{mi}'\gamma_m)\end{aligned}.$$

Following the approach suggested by (De Luca & Peracchi, 2012) let's approximate unknow conditional expectation with the polynomial of generalized mills ratios:

$$g_k(w_{1i}'\gamma_1,...,w_{mi}'\gamma_m) = \sum_{t=1}^{k}\sum_{j=1}^{m}\tau_t^{(v)}\lambda_j^{(-u_i)}(w_{1i}'\gamma_1,...,w_{mi}'\gamma_m)^t + \tau_0$$

Semi-parametric two-step estimator proceeds as follows:

1. Estimate $w_{ji}'\gamma_j$ and correlation between random errors via some semi-parametric or non-parametric method, $j \in \{1,\cdots,m\}$. Then estimate $\lambda_j^{(-u_i)}(w_{1i}'\gamma_1,...,w_{mi}'\gamma_m)$.
2. Substitute g^* with g and insert first step estimates into the equation of dependent variable conditional expected value. Then estimate this equation parameters via least squares method. Polynomial degree k should be selected by cross-validation (for example leave-one-out) while hypothesis testing requires bootstrap procedure application.

Note that the most technically complicated part is estimation of the system of binary equations on the first step. In order to reduce computational burden, it is possible to use the following specifications for approximating function allowing to estimate binary equations parameters separately:

$$g_k(w'_{1i}\gamma_1,...,w'_{mi}\gamma_m)=\sum_{t=1}^{k_1}\tau_t^{(1)}s(w'_{1i}\gamma_1)^t+...+\sum_{t=1}^{k_m}\tau_t^{(m)}s(w'_{mi}\gamma_m)^t+\tau_0,$$

$$g_k(w'_{1i}\gamma_1,...,w'_{mi}\gamma_m)=\sum_{t_1=0}^{k_1}...\sum_{t_m=0}^{k_m}\tau_{(t_1,...t_m)}s(w'_{1i}\gamma_1)^{t_1}...s(w'_{mi}\gamma_m)^{t_m},$$

where $k=(k_1,\cdots,k_m)$ and s could be inverse mills ratio for the standard normal distribution. “Note that the first of these specifications is a special case of the second one. The last one includes interaction terms in order to capture statistical relationship among random errors.” (Коссова, Куприянова, & Потанин, 2020).

It is technically complicated to proof that this estimator will be consistent. The proof for the one dimensional case has been provided by (Newey, 2009) while (De Luca & Peracchi, 2012) generalized his approach to bivariate case without any guarantee that tthe estimator will be consistent. However simulated data analysis suggests that this estimator accuracy is rather high.

Marginal effects

Researchers usually need to measure the effect of the regressors on the dependent variable. If its expected value depends on the regressors nonlinearly then marginal effects should be estimated.

For parametric models proposed above marginal effects have the form:

$$\frac{\partial E(y_i\mid z_{1i},...,z_{mi})}{\partial\psi}=\beta_\psi+\sum_{j=1}^{m}\Upsilon_\psi^j(\tilde{z}_i),$$

where:

$$\begin{aligned}\Upsilon_\psi^j(\tilde{z}_i)&=z_{ji}\sigma_{g_i}\rho_{j,g_i}\sum_{k=1}^{m}z_{ki}\gamma_\psi^k\frac{\partial\lambda_j^{\tilde{u}_i}(\tilde{z}_i)}{\partial\tilde{z}_k}=z_{ji}\sigma_{g_i}\rho_{j,g_i}(\sum_{k\neq j}z_{ki}\gamma_\psi^k(\Lambda_{jk}^{\tilde{u}_i}(\tilde{z}_i)-\lambda_j^{\tilde{u}_i}(\tilde{z}_i)\lambda_k^{\tilde{u}_i}(\tilde{z}_i)))-\\&-\gamma_\psi^j\sigma_{g_i}\rho_{j,g_i}(\tilde{z}_{ji}\lambda_j^{\tilde{u}_i}(\tilde{z}_i)+(\lambda_j^{\tilde{u}_i}(\tilde{z}_i))^2+\sum_{k\neq j}z_{ji}z_{ki}\rho_{jk}\Lambda_{jk}^{\tilde{u}_i}(\tilde{z}_i)).\end{aligned}$$

In order to get marginal effects estimates for semi-nonparametric method numeric differentiation technics could be applied for example Richardson’s method (Richardson, 1911). Note that in order to get estimates for conditional expectations and covariance matrix it is possible to apply formulas derived in the thesis in order to get the following expression (where $k_t \in N$):

$$E\left(\prod_{t=1}^{m}\left(\xi_t^{(\bar{\alpha},\bar{\beta})}\right)^{k_t}\right)=\frac{1}{\psi\bar{F}_{\xi}(\bar{\alpha},\bar{\beta})}\prod_{t=1}^{m}(\Phi(\bar{\beta}_t;\mu_t,\sigma_t)-\Phi(\bar{\alpha}_t;\mu_t,\sigma_t))*$$

$$*\sum_{i_1=0}^{K_1}\dots\sum_{i_m=0}^{K_m}\sum_{j_1=0}^{K_1}\dots\sum_{j_m=0}^{K_m}\alpha_{(i_1,\dots,i_m)}\alpha_{(j_1,\dots,j_m)}\prod_{r=1}^{m}\mathcal{M}_{TR}\left(i_r+j_r+k_r;\bar{\alpha}_r,\bar{\beta}_r;\mu_r,\sigma_r\right)= \quad ,$$

$$=\frac{1}{\tilde{\psi}}\sum_{i_1=0}^{K_1}\dots\sum_{i_m=0}^{K_m}\sum_{j_1=0}^{K_1}\dots\sum_{j_m=0}^{K_m}\alpha_{(i_1,\dots,i_m)}\alpha_{(j_1,\dots,j_m)}\prod_{r=1}^{m}\mathcal{M}_{TR}\left(i_r+j_r+k_r;\bar{\alpha}_r,\bar{\beta}_r;\mu_r,\sigma_r\right)$$

$$\tilde{\psi}=\sum_{i_1=0}^{K_1}\dots\sum_{i_m=0}^{K_m}\sum_{j_1=0}^{K_1}\dots\sum_{j_m=0}^{K_m}\alpha_{(i_1,\dots,i_m)}\alpha_{(j_1,\dots,j_m)}\prod_{r=1}^{m}\mathcal{M}_{TR}\left(i_r+j_r;\bar{\alpha}_r,\bar{\beta}_r;\mu_r,\sigma_r\right).$$

Publication of the results

The results of the thesis have been published in leading domestic and international economic journals included into Scopus citation system:

1. Kossova, E., & Potanin, B. (2018). Heckman method and switching regression model multivariate generalization. Applied Econometrics, 50, 114-143.

Characteristics: Scopus Q4; Size — 1.3 copyright sheet.

2. Potanin, B. (2019). Estimating the Effect of Higher Education on an Employee's Wage. Studies on Russian Economic Development, 30(3), 319-326.

Characteristics: Scopus Q3; Size — 0.7 copyright sheet.

3. Kossova, E., Potanin, B., & Sheluntcova, M. (2020). Estimating effect of marriage on male wages in Russia. Journal of Economic Studies, 47(7). Accepted for publication.

Characteristics: Scopus Q1; Size — 1 copyright sheet.

4. Kossova, E., Potanin, B., & Kupriyanova, L. (2020). Parametric and semiparametric multivariate sample selection models estimators accuracy comparative analysis on simulated data. Applied Econometrics. Accepted for publication.

Characteristics: Scopus Q4; Size — 1.1 copyright sheet.

Also, the candidate has presented the results of the thesis on the following conferences:

1. V Международная конференция «Modern Econometric Tools and Applications – META2018» (Нижний Новогород). Доклад: Estimating the effect of marriage on male wages.

2. XIX Апрельская международная научная конференция (Москва). Доклад: Обобщение метода Хекмана на случай произвольного числа уравнений отбора.

3. XIX Апрельская международная научная конференция (Москва). Доклад: Оценка влияние типа политического режима на приток прямых иностранных инвестиций при помощи многомерной свич-пробит модели с фиксированными эффектами.

4. XIX Апрельская международная научная конференция (Москва). Доклад: Применение многомерной иерархической свич-пробит модели для анализа дискриминации замужних женщин на российском рынке труда по данным РМЭЗ за 2016 год.

5. VI Международная конференция «Modern Econometric Tools and Applications – META2019» (Нижний Новгород). Доклад: Comparative Analysis of Parametric, Semi-parametric and SemiNonparametric Sample Selection Models with Application to Engel's Curve Parameters Estimation.

6. XX Апрельская международная научная конференция НИУ ВШЭ (Москва). Доклад: Оценивание эффекта высшего образования на зарплату в условиях неслучайного отбора.

Literature

Arellano-Valle, R. B., Branco, M. D., & Genton, M. G. (2006). A Unified View on Skewed Distributions Arising from Selections. *The Canadian Journal of Statistics*, *34*(4), 581-601.

Cartinhour, J. (1990). One-dimensional marginal density functions of a truncated multivariate normal density function. *Communications in Statistics - Theory and Methods*, *19*(1), 197-203. doi:https://doi.org/10.1080/03610929008830197

Cinzia, D. N. (2009). Sample selection correction in panel data models when selectivity is due to two sources. *Institute of Public Policy and Public Choice (POLIS) Working Papers*.

Copas, J. B., & Li, H. G. (1997). Inference for non-random samples. *Journal of the Royal Statistical Society. Series B (Methodological), 59*, 55-95.

Das, M., Newey, K., & Vella, F. (2003). Nonparametric Estimation of Sample Selection Models. *The Review of Economic Studies*, *1*(1), 33-58.

De Luca, G., & Peracchi, F. (2012). Estimating Engel curves under unit and item nonresponse. *Journal of Applied Econometrics*, *27*(7), 1076-1099.

Gallant, A., & Nychka, D. (1987). Semi-Nonparametric Maximum Likelihood Estimation. *Econometrica, 55*(2), 363-390.

Garen, J. (1984). The retums to schooling: A selectivity bias approach with a continuous choice variable. *Econometrica*, *52*(5), 1199-1218.

Heckman, J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic Social Measurement, 5*(4), 475-492.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153-161.

Heckman, J. J., & Taber, C. (2010). Roy model. *Durlauf S.N., Blume L.E. (eds) Microeconometrics. The New Palgrave Economics Collection. Palgrave Macmillan*. doi:10.1057/9780230280816_27

Heckman, J., & Honore, B. E. (1990). The Empirical Content of the Roy Model. *Econometrica*, 1121-1149.

Heckman, J., & Killingsworth, M. (1987). Female labor supply: a survey. *O. Ashenfelter and R. Layard (eds.) Handbook of Labor Economics, 1*, 103-204.

Jeffrey, D., & McFadden, D. (1984). An econometric analysis of residential electric appliance holdings and consumption. . *Econometrica*, *52*(2), 345-362.

Kossova, E. V., Kupriyanova, L. A., & Potanin, B. S. (2020). Parametric and semiparametric multivariate sample selection models estimators accuracy comparative analysis on simulated data. *Applied Econometrics (принята к печати)*.

Kossova, E., & Potanin, B. (2018). Heckman method and switching regression model multivariate generalization. *Applied Econometrics, 50*, 114-143.

Kossova, E., Potanin, B., & Sheluntcova, M. (2020). Estimating effect of marriage on male wages in Russia. *Journal of Economic Studies (принята к печати), 47*(7).

Kugler, J. E. (1987). The earnings impact of training duration in a developing country. *Journal of Human Resources, 22*(2), 228-247.

Lee, L. (1978). Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables. *International Economic Review, 19*, 415-433.

Manjunath, B. G., & Wilhelm, S. (2012). Moments Calculation For the Doubly Truncated Multivariate Normal Density. *arXiv*.

Murphy, K. M., & Topel, R. H. (1985). Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics, 3*, 370-379.

Newey, W. K. (2009). Two-step series estimation of sample selection models. *The Econometrics Journal, 12*, 217-229.

Ogundimu, E. O., & Hutton, J. L. (2016). A unified approach to multilevel sample selection models. *Communications in Statistics - Theory and Methods, 45*(9), 2592-2611.

Poirier, D. L. (1980). Partial observability in bivariate probit models. *Journal of Econometrics, 12*(2), 209-217.

Potanin, B. (2019). Estimating the Effect of Higher Education on an Employee's Wage. *Studies on Russian Economic Development, 30*(3), 319-326.

Richardson, L. F. (1911). The approximate arithmetical solution by finite differences of physical problems including differential equations, with an application to the stresses in a masonry dam. *Philosophical Transactions of the Royal Society A, 210*, 307-357.

Rosenmana, R., Mandal, B., Tennekoon, V., & Hill, L. G. (2010). Estimating treatment effectiveness with sample selection. *Working Paper*.

Roy, A. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers, 3*, 135–146.

Vella, F. (1993). A simple estimator for models with censored endogenous regressors. *International Economic Review, 34*(2), 441-457.

Vella, F. (1998). Estimating models with sample selection bias: A survey. *Journal of Human Resources*, *34*(2), 127-169.