НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Identification and estimation of nonparametric simultaneous equations models with sample selection

Evgeniy M. Ozhegov
National Research University Higher School of Economics
Research Group for Applied Market and Enterprises Studies

Neighbors in Research Conference, NRU HSE, Perm, 12.05.2014

Outline

1. Model and estimation challenges
2. Literature review
3. Methodology
4. Conclusion

Model

$$d_i = \begin{cases} 1, g_0(w_{0i}, x_{0i}) + e_{0i} \geq 0 \\ 0, g_0(w_{0i}, x_{0i}) + e_{0i} < 0 \end{cases}$$

$$\begin{cases} y^*_{1i} = g_1\left(x_i^{1^*}, x_i^{2^*}, w_{1i}, y^*_{-1i}\right) + e_{1i} \\ \ldots \\ y^*_{ki} = g_k\left(x_i^{1^*}, x_i^{2^*}, w_{ki}, y^*_{-ki}\right) + e_{ki} \end{cases}$$

$$x_i^{2^*} = \pi\left(x_i^{1^*}, z_i\right) + \nu_i$$

$$(y_i, x_i) = d_i(y_i^*, x_i^*) \text{ is observed}$$

Model. Estimation challenges

If $f(e_0, e_1, \dots, e_k, w_0, w, x)$ is density function of joint error and regressors distribution then separate estimation of each j-th equation of the system using OLS may be inconsistent due to:

1. Sample selection if $cov(e_j, e_0|x, w) \neq 0$;
2. Simultaneity of equations if $\exists s \in \{1, \dots, k\}, s \neq j, cov(e_j, e_s|x, w) \neq 0$;
3. Endogeneity of regressors if $cov(e_j, x|x_0, e_0) \neq 0$.

Literature review

Sample selection bias problem: Gronau (1973), Heckman (1974);

2-step estimation strategy: Heckman (1976, 1979);

Estimation strategy for unknown disturbances distribution: Heckman, Robb (1985), Newey (1988);

Endogenous variables and unknown regression functions: Newey, Powell (1989);

2-step procedure for estimation of triangular systems of simultaneous equations: Newey, Powell, Vella (1999);

Sample selection and endogenous variables in outcome equation: Das, Newey, Vella (2003).

Literature review

Matzkin (2010) proposed technique for identification and estimation for nonseparable simultaneous equations model through estimation of kernel density of exogenous and unobservable variables distribution;

In (Matzkin, 2012) it was extended for simultaneous equations with Tobit-type truncated endogenous variables;

Matzkin and Blundell (2010) also proposed identification conditions for nonseparable simultaneous equation models when the number of excluded instruments is less then number of endogenous variables:

Imbens and Newey (2009) for triangular simultaneous equations without additivity using quantile IV approach.

This paper

Extends Matzkin (2012) for the case of simultaneous equations with sample selection;

Uses more simple and efficient estimator for small samples (ext. of DNV, 2003);

Uses assumption on separability of unobservables (will be relaxed in further work).

Methodology. Procedure

1. Firstly, we need to estimate the propensity score for the selection equation:

$$p = E[d|x_0, w_0] = g_0(w_0, x_0)$$

Methodology. Procedure

2. On the second step we will estimate the prediction of endogenous regressors corrected for sample selection using propensity score:

$$E[x^2|x^1, z, w_0, d = 1] = \pi(x^1, z) + \lambda(p)$$

Methodology. Procedure

3. Then we will estimate each equation of the system in the reduced form corrected for sample selection and endogeneity of regressors using propensity score and errors from the endogenous regressors equations:

$$E\left[y_j \middle| x^1, x^2, z, w, w_0, d = 1\right] = \gamma_j(x^1, x^2, w) + \mu(p, v)$$

Methodology. Procedure

4. On the last step we will estimate the structural form equations corrected for sample selection, endogeneity and simultaneity using propensity score, errors from endogenous regressors equations and reduced form errors:

$$E\left[y_j \middle| x^1, x^2, w_j, y_{-j}, z, w_{-j}, w_0, d = 1\right] = g_j\left(x^1, x^2, w_j, y_{-j}\right) + \varphi(p, v, e_{-j})$$

Methodology. Identification

Theorem 1. *If functions* $g_0(w_0, x_0)$, $\pi(x^1, z)$, $\lambda(p)$, $\gamma_j(x^1, x^2, w)$, $\mu(p, \nu)$, $g_j(x^1, x^2, w_j, y_{-j})$, $\varphi(p, \nu, e_{-j})$ *are continuously differentiable with continuous distribution functions almost everywhere and with probability one* $\frac{\partial g_0(w_0, x_0)}{\partial w_0} \neq 0$, $rank\left[\frac{\partial \pi(x^1, z)}{\partial z}\right] \geq dim(x^2)$ *and for each* $j \in \{1, \dots, k\}$ *there is at least one* w_j *with* $\frac{\partial \gamma_j(x^1, x^2, w)}{\partial w_j} \neq 0$ *exists then each regressions function* g_0, π, γ_j, g_j *identified up to an additive constant.*

Methodology. Consistency

Theorem 2. *If with probability one* $\frac{\partial g_0(w_0,x_0)}{\partial w_0} \neq 0$, $rank\left[\frac{\partial \pi(x^1,z)}{\partial z}\right] \geq dim(x^2)$ *and for each* $j \in \{1, \dots, k\}$ *there is at least one* w_j *with* $\frac{\partial \gamma_j(x^1,x^2,w)}{\partial w_j} \neq 0$, *regression and correction functions at each stage is approximated by polynom, propensity score and error terms are replaced with its estimates from previous stages and* (x^1, z, w_0, w) *is independent on the distribution of* $(e_0, e_1, \dots, e_k, v\ )$ *then OLS estimate of polynomial approximation of regression function will be consistent.*

Methodology. Requirements

1. Large support of regression functions;
2. Approximation of regression and control functions by polynoms;
3. One excluded instrument (relevant and valid) for selection equation and each endogenous variable;
4. Exogeneity of included instruments.

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

Thank you!

tos600@gmail.com

Hey! Last slide!