NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Evgeniy M. Ozhegov

THE UNDERWRITING, CHOICE AND PERFORMANCE OF GOVERNMENT-INSURED MORTGAGES IN RUSSIA

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: Financial Economics
WP BRP 31/FE/2014

SERIES: *FINANCIAL ECONOMICS*

Evgeniy M. Ozhegov[1]

THE UNDERWRITING, CHOICE AND PERFORMANCE OF GOVERNMENT-INSURED MORTGAGES IN RUSSIA[2]

This paper analyzes the mortgage borrowing process from a Russian state-owned provider of residential housing mortgages concentrating on the choice of having government insurance. This analysis takes into account the underwriting process and the choice of loan limit by the bank, the choice of contract terms and the performance of all loans issued from 2008 to 2012. Our dataset contains demographic, financial, loan terms and the performance information for all applications. We use a multistep nonparametric approach to estimate the determinants of bank and borrower choice. The main finding that the probability of having government insurance is linked to riskier loans, but insured loans also are more likely to be approved by the bank. The bank, when approving a borrower, takes into account not the probability of default but the difference between the probability of default and having government insurance.

JEL classification: C14, C30, C51, G21.

Keyword: mortgage, terms choice, default, nonparametrics.

[1] National Research University Higher School of Economics. Research group for applied markets and enterprises studies. Young research fellow. E-mail: tos600@gmail.com

[2] This study (research grant No 14-01-0104) was supported by The National Research University–Higher School of Economics' Academic Fund Program in 2014- 2015

Introduction

During the previous decade, the Russian housing market was affected by two events. First, the worldwide financial crisis caused housing prices to fall by 30% in September 2009 compared with prices in July 2008[3]. The volume of loans issued in 2009 fell to 25% of the level of 2008. Secondly, Agency of Housing Mortgage Lending (AHML), the state-owned mortgage provider, increased the volume of mortgages issued by 120% from 2006 to 2012 without any spillover during the crisis[4] and now holds 7-12% of the market share. This means that demand for government-issued loans is rising despite the rises and falls in the economy and financial markets.

When applying for an AHML loan the potential borrower chooses whether to have government loan insurance in case of delinquency, along with other mortgage terms. If loan-to-value ratio (LTV) is more than 70% then the loan must be insured. While credit risk in the Russian residential mortgage market has been stable over the past 8 years and the mean probability of default varied from 4 to 5%[5], government-insured AHML loans performed substantially worse and showed a 16% probability of default. This means that government insurance covers potential losses from such loans and may affect its approval process.

We are interested in the conditions leading to having a government-insured loan, its performance and the underwriting process of such loans. In this paper, we will estimate the demand function for AHML loans and the probability of default equation taking into account the personal characteristics of the borrower and their choice of mortgage contract terms. We also control for the selection bias which arises during the underwriting process.

This paper uses unique loan-level data on applications, contract terms and the performance of mortgages from one regional AHML subsidiary. We use a nonparametric sequential estimation approach in order to control for endogeneity in contract terms and the selection process by bank and borrower.

Next section describes borrowing process in AHML and modeling issues. Section 3 details the data. Section 4 contains the econometric model. Section 5 describes the results of estimation. Section 6 concludes.

[3] By the Indicators of housing market's Price Index, www.irn.ru
[4] Agency of Housing Mortgage Lending data, www.ahml.ru
[5] Agency of Housing Mortgage Lending data, www.ahml.ru

Theoretical background

The demand for mortgages from a particular bank is usually dependent on the probability of taking a loan, its size, the characteristics of a potential borrower, the credit terms and the strategies of other banks (Ozhegov, 2014).

This research analyzes the demand for mortgages from a Russian regional bank which offers mortgage programs developed by AHML. AHML is a fully state-owned company which develops mortgage programs for special groups of borrowers ("young families", "young teachers", "soldiers" and so on) and higher risk borrowers who are unable to get a mortgage from commercial banks. These programs are developed for commercial banks. If a bank issues a mortgage on the AHML program with documentation satisfying the "AHML Standards", then this loan will be automatically refinanced by AHML. The bank is paid a fixed reward.

The borrowing process has 4 steps:

1. Application

A potential borrower chooses a credit organization and credit program that reflects their preferences, fills out an application form with their demographic and financial characteristics.

2. Approval

Considering the application and recent credit history, the credit organization approves or disapproves the application, inquires the form data. When approving a particular borrower, the credit organization sets the limit of loan amount.

3. Contract agreement and choice of credit terms

The approved borrower chooses a contract agreement, a particular property to buy and credit terms: loan amount, downpayment, monthly payment, maturity, and government insurance in case of insolvency. As mentioned, if LTV is 70% or more, then the loan must be insured. The interest rate is determined by the credit program and depends on the other terms.

4. Performance of loan

The borrower chooses to repay the loan according to the contract, or deviates from it in some way: becomes delinquent and defaults, or prepays and refinances the loan.

Traditional models for demand estimation on the residential mortgage market used a parametric approach to estimate the loan amount or LTV equation. The two main challenges for those models have been widely discussed. The first is sample selection and the second is the endogeneity of the other contract terms.

Sample selection issues arise when decisions on a loan are made sequentially and some explanatory variables are partially observed at various stages of the lending process. If the approval process is correlated with the choice of contract terms then the magnitude of selection bias depends

on the strength of the correlation between the LTV choice and the underwriting process and also on the available data in the application sample (Ross 2000).

Mortgage borrowing as a sequence of consumer and bank decisions was introduced by Follain (1990). He defines the borrowing process as the choice of how much to borrow, if and when to refinance or default, and the choice of mortgage instrument itself. Rachlis and Yezer (1993) suggest a theoretical model of the mortgage lending process, which consists of a system of four simultaneous equations: (1) borrower application, (2) borrower selection of mortgage terms, (3) lender endorsement, and (4) borrower default. This paper investigates the nature of the inconsistency of estimates of recent research on borrower discrimination and shows that all four equations (and decisions) should be considered interdependent.

Public data, such as American mortgage datasets from the Federal Housing Authority (FHA) foreclosure, The Boston Fed Study, The Home Mortgage Disclosure Act (HMDA) was published in the middle of 1990s. Using this data a few empirical studies analyzed the mortgage lending process and studied the interdependency of bank endorsement decisions and borrower decisions modeled by the bivariate probit model. As an extension of the study Rachlis and Yezer, (1993), Yezer, Philips and Trost (1994) applied a Monte-Carlo experiment to estimate the theoretical model. They empirically show that isolated modeling of the processes of credit underwriting and default lead to biased parameter estimates. Phillips and Yezer (1996) and Munnell, Tootell, Browne and McEneaney (1996) supported these findings.

Later papers studied the dependence of credit term choices on the other endogenous variables. Ambrose, LaCour-Little and Sanders (2004) outlined the endogeneity of the loan amount and LTV.

As key determinants for the demand for the residential mortgage market, authors usually select socio-demographic characteristics of borrower and contract terms. Bajari, Chu and Park (2008) also use district-level aggregated demographic and economic variables as proxies for individual characteristics when they are unavailable.

Attanazio, Goldberg, Kyriazidou, (2008) applied the Das et al. (2003) three-step nonparametric approach to estimate the demand for car loans corrected for sample selection and the endogeneity of rate and maturity. First, they estimated the probability of taking a loan, then residuals from the endogenous variable equations and then the demand equation corrected for sample selection and endogeneity. They found empirical evidence of nonlinearity in the demand function and the non-normality of the joint distribution of error terms.

Not only LTV depends on the other contract terms. The choice of LTV may also affect all the other contract terms. Higher risk loans relate to the credit programs with a higher rate. A higher loan amount with a fixed rate requires larger monthly payments or a maturity extension for credit

constrained borrowers (Attanasio et al., 2008). Higher LTV also implies a higher probability of government insurance. The choice of all credit terms is modeled as structurally interdependent.

The loan limit set by the bank should also be considered as endogenous when modeling the choice of contract terms. If the bank predicts a borrower's decision on credit terms then it may adjust the loan limit in order to achieve the optimal contract.

Recent papers showed that the approval process affects borrower decisions. Table 1 also gives evidence of a biased sample of the characteristics of the borrowers who did not sign a credit contract. In general, when modeling the contract term choice we may consider the subsample as biased because: 1) some applicants were considered uncreditworthy and rejected; 2) some approved borrowers did not sign a contract because of better alternatives in other banks or the loan available was too small. With the data available we cannot separate these two reasons since we do not know the approval decision and loan limit for all the applicants who did not sign a contract.

To sum up, borrowing process is represented by the following econometric model:

$$
\begin{gathered}
d_i = \begin{cases} 1, g_0(w_{0i}, x_i^1) + e_{0i} \geq 0 \\ 0, g_0(w_{0i}, x_i^1) + e_{0i} \leq 0 \end{cases} \\
\begin{cases} y_{1i}^* = g_1\left(x_i^1, x_i^{2^*}, w_{1i}, y_{-1i}^*\right) + e_{1i} \\ \dots \\ y_{ki}^* = g_k\left(x_i^1, x_i^{2^*}, w_{ki}, y_{-ki}^*\right) + e_{ki} \end{cases} \\
x_i^{2^*} = \pi(x_i^1, z_i) + \nu_i \\
def_i^* = \begin{cases} 1, g_{def}\left(y_i^*, x_i^1, x_i^{2^*}\right) + e_{def,i} \geq 0 \\ 0, g_{def}\left(y_i^*, x_i^1, x_i^{2^*}\right) + e_{def,i} \leq 0 \end{cases} \\
(y_i, x_i^1, x_i^2, def_i) = d_i(y_i^*, x_i^1, x_i^{2^*}, def_i^*) \text{ is observed}
\end{gathered}
\tag{1}
$$

where d_i is a binary indicator of contract signing, x_i^1 is a set of demographic and financial characteristics of the borrower and co-borrowers, y_i is the set of credit terms, x_i^2 is the logarithm of the loan limit, $(w_{0i}, w_{1i}, \dots, w_{ki}, z_i)$ is the set of excluded instruments for the contract signing decision, credit terms and loan limit respectively. The set of credit terms y_i contains LTV, logarithm of rate, logarithm of maturity and the probability of government insurance. def_i is a binary indicator of default.

Data description

One of the regional AHML operators provided the data set of all applications for mortgage collected from 2008 to 2012. We know the demographic and financial characteristics of each of

the 3870 applicants as main borrowers and their co-borrowers on the date of application, we also know the date of application. For all signed contracts we know the loan limit set by the bank, the contract terms, and the value of property. The characteristics of the borrower are fully observable and the contract characteristics are partially observable for only the subsample of applicants who signed the contract.

Tab. 1. Descriptive statistics for applicants' characteristics.

Variable	Full sample (3366 obs.)	Signed a contract (2041 obs.)	Did not sign a contract (1325 obs.)
Age[6], years	33.8 (7.57)	34.0 (7.67)	33.5 (7.41)
Sex			
Male	1858 (55.2%)	1161 (56.9%)	697 (52.6%)
Female	1508 (44.8%)	880 (43.1%)	628 (47.4%)
Marital status			
Single	1017 (30.2%)	590 (29.0%)	426 (32.2%)
Married	1807 (53.7%)	1146 (56.1%)	661 (49.9%)
Widowed	42 (1.2%)	20 (1.0%)	22 (1.7%)
Divorced	500 (14.8%)	284 (13.9%)	216 (16.3%)
Category of employment			
Hired employee	3229 (95.9%)	1942 (95.1%)	1287 (97.1%)
Entrepreneur	25 (0.7%)	19 (0.9%)	6 (0.5%)
State-owned employee	112 (3.3%)	80 (3.9%)	32 (2.4%)
Level of education			
Elementary	53 (1.6%)	33 (1.6%)	20 (1.5%)
Secondary	1425 (42.3%)	816 (40.0%)	609 (50.0%)
Incomplete higher	120 (3.6%)	64 (3.1%)	56 (4.2%)
Complete higher	1768 (52.5%)	1128 (55.3%)	640 (48.3%)
Number of co-borrowers			
0	1423 (42.3%)	1012 (49.6%)	593 (44.8%)
1	1809 (53.7%)	1120 (54.9%)	689 (52.0%)
2	134 (4.0%)	91 (4.5%)	43 (3.2%)
Declared income of co-borrowers			
Not declared	2949 (87.6%)	1687 (82.7%)	1262 (95.2%)
From 0 to 9999 rub.	111 (3.3%)	103 (5.0%)	8 (0.6%)
From 10000 to 19999 rub.	161 (4.8%)	133 (6.5%)	28 (2.1%)
More than 20000 rub.	145 (4.3%)	118 (5.8%)	27 (2.0%)
Declared income of main borrower			
Not declared	2337 (69.4%)	1227 (60.1%)	1110 (83.8%)
From 0 to 9999 rub.	91 (2.7%)	53 (2.6%)	38 (2.9%)
From 10000 to 19999 rub.	283 (8.4%)	241 (11.8%)	42 (3.2%)
From 20000 to 39999 rub.	445 (13.2%)	361 (17.7%)	84 (6.3%)
More than 40000 rub.	210 (6.2%)	159 (7.8%)	51 (3.8%)

The outliers from the sample were excluded. We treat an observation as an outlier if the age, level of education, marital status or other characteristics were missing. We exclude observations with borrowers under age 21, with LTV or DTI (debt-to-income ratio) less than 0 or

[6] Mean and standard deviation in the parenthesis.

more than 1. We consider those outliers as random and due to the errors in the database. After excluding the outliers the sample was 3366 observations. 2041 applicants signed the mortgage contract, while 1325 of them did not. The descriptive statistics of the available variables are shown in the Tables 1 and 2.

Some mortgage programs allow the applicants not to provide any information on their income. These programs are usually linked with a higher contract rate. The reason for this choice may be explained by a temporary or changeable income (LaCour-Little, 2007), for instance, for entrepreneurs. Generally, income should be considered endogenous while modeling the approval of borrower or contract terms. However, we can control for employment category, which rejects the inconsistency due to possible endogeneity of income. Moreover, co-borrower income may also be endogenous and we cannot provide any proxy for co-borrower income since we do not have any characteristics of co-borrowers. This is a limitation of the research. But we may consider it as insignificant for the choice of contract terms compared to the income of the main borrower.

Tab. 2. Descriptive statistics of the issued loans (2041 contracts).

Variable		Mean	St. dev.	Min	Max
Loan amount, rub.		1 040 037	573 665	120 000	10 000 000
Downpayment, rub.		854 962	706 635	40 000	13 820 000
Flat value, rub.		1 894 999	1 049 502	330 000	15 290 000
Monthly payment, rub.		12 610	7 324	1 872	140 381
Loan limit, rub.		1 046 023	587 762	150 000	10 000 000
Loan-to-value ratio (LTV)		0.56	0.17	0.11	0.94
Maturity, months		189.05	62.17	26	360
Rate, %		11.59	1.64	9.55	19
Insurance	Is insured		1851 (90.7%)		
	Not insured		190 (9.3%)		
			Insured	Not insured	Total
Indicator of default	Not defaulted		1783 (96.3%)	159 (83.7%)	1942 (95.1%)
	Defaulted		68 (3.7%)	31 (16.3%)	99 (4.9%)

To estimate the model we need to find a set of relevant excluded instruments for the probability of signing a contract, the loan limit and each credit term.

Bajari et al. (2008) discussed the possibility of using aggregated district-level variables as proxies for unavailable data. We will use the same strategy to find the set of instruments. Since we have data without spatial variation we can use time variation in applications. We have data from July 2008 to August 2012 and we know the application date for each applicant. Each application was matched with the set of aggregated mortgage and housing market characteristics for the same month. On average, the process takes two months from the date of application to the date of contract agreement. Also Ozhegov and Poroshina (2013) showed that aggregated demand on

mortgage reacts to changes in supply within two months. Then we need to fix the aggregated market characteristics for each application not only in the month of application, but also the 1-2 months prior the application, and use these as instruments.

Table 3 represents the descriptive statistics of aggregated mortgage and housing market characteristics for the period from July 2008 to August 2012 (50 months).

About 15% of issued loans were refinanced by AHML, but not all of them were issued by the bank supplying the data. Generally, the number of applications to the bank is fewer than the number refinanced by AHML by all the regional banks.

Tab. 3. Aggregated mortgage and housing markets characteristics.

Variable	Mean	St. dev.	Min	Max
Volume of issued mortgage in region, mln. rub.	921.8	562.3	116.1	2191.0
Volume of issued mortgage in region, number	894.40	529.27	134	2112
Mean loan amount, rub.	1 152 568	251 993	899 310	1 908 200
Median maturity, months	200.79	12.81	173	222.2
Median rate, %	12.97	0.80	12	14.3
Mean LTV	0.58	0.03	0.48	0.65
Mean DTI[7]	0.35	0.01	0.33	0.37
Mean $м^2$ value, rub.	38 622	6 165	28 782	51 304
Affordability of housing coefficient[8]	0.287	0.055	0.215	0.389
Number of refinanced in AHML loans	129.1	83.7	30	310
Number of application to the bank	121.4	51.9	43	222

The difference between the number of loans refinanced by AHML and the number of applications to the bank within a particular month may be the excluded variable which explains the probability of contract agreement, but it does not affect the contract term choice. Since every commercial bank operates with the same AHML programs, the difference in the approval process does not affect the term choice. But an increase in the number of refinanced loans shows the changes in the underwriting process in other banks and may correlate with the probability of a contract agreement with the bank. This variable should be considered as exogenous since each individual decision explains an insignificant variation of the aggregated market characteristic (less than 1%).

As an excluded instrument for the loan limit we use the mean Debt-to-Income ratio (DTI). The positive dependence of these two variables is because the mean DTI for all issued loans reflects the evaluation of the mean credit risk (the higher the DTI of issued loans, the less risk). It positively correlates with the loan limit, which reflects the willingness to issue a larger loan for a

[7] DTI – ratio between monthly payment and monthly income.
[8] Affordability coefficient reflects the ratio between an income of mean household and a value of mean flat.

particular borrower. This variable is valid since individual shocks of loan limit do not affect the aggregated characteristic of issued loans.

As excluded instruments for credit terms, LTV, rate, maturity and insurance, we used mean LTV, median rate, median maturity for issued loans and the housing affordability coefficient. The relevance of the first three instruments is implied by the interdependence of mortgage market characteristics and the AHML credit programs conditions. Validity is implied by the exogeneity of the program terms for each particular borrower.

The affordability coefficient is relevant for the probability of insurance because the increase of affordability should lead to the choice of a lower LTV and consequently to a lower probability of loan insurance. Validity is also implied by the independence of individual preference on insurance shocks and the aggregated affordability of housing. All the variables are relevant and valid and may be used as instruments. The relevance will be also proved for each model with the *F*-test for the excluded instrument in section 5.

Econometric model

Model (1) contains a system of simultaneous equations when we model the choice of contract terms. Moreover, contract terms are observable only for the subsample of borrowers who have signed contract. This means that we have selection bias problem.

The sample selection bias problem was discussed in Gronau (1973) and Heckman (1974). Heckman proposed methods to estimate these models using maximum likelihood or the two-step procedure in Heckman (1976, 1979) which corrects the error term in the outcome equation on covariance with the selection equation error term. However, both approaches have been limited by an assumption on the joint error distribution. Further papers deal with a relaxation of the distribution assumption for the two-step procedure using a nonparametric approach for model estimation, for instance, using a Fourier decomposition of unknowns in terms of a functional form error correction function (Heckman and Robb, 1985), or an approximation by a series of power functions (Newey, 1988).

While modeling the borrower choice of contract terms we need to allow regressors to be endogenous and represent the system of equations for each endogenous variable in structural form. Regression functions are unknown and not limited by any assumptions. Newey and Powell (1989) introduced a nonparametric procedure for the estimation of a triangular system of simultaneous equations with unknown regression functions. Vella (1993) elaborated on this procedure for the case of a limited dependent variable. Newey, Powell and Vella (1999) proposed a two-step procedure for the correction of an error term on the endogeneity of regressors approximating the

control function by power series on reduced form residuals. Then Newey (2013) provided an overview of nonparametric instrumental variable methods for simultaneous equations and discussed the problem of weak instruments.

Das, Newey and Vella (2003) proposed a model with both sample selection and endogenous regressors, and its estimation procedure. They also approximated a control function using a power series which depended on the propensity score from the selection equation and the endogenous variable reduced form equation residuals.

We extend the proposed methods for the consistent estimation of a non-triangular system of simultaneous equations with sample selection, endogenous regressors and arbitrary joint error distribution and the functional form of regression and the control functions in reduced and structural forms. We may apply this method to estimate model (1) with the following steps.

1. We need to estimate the propensity score for the contract agreement equation:

$$p = E[d|x_0, w_0] = g_0(w_0, x_0) \tag{2}$$

2. We estimate the prediction of endogenous regressors which is logarithm of the loan limit corrected for sample selection using the estimate of propensity score:

$$E[x^2|x^1, z, w_0, d = 1] = \pi(x^1, z) + \lambda(\hat{p}) \tag{3}$$

3. We estimate each contract term equation in the reduced form corrected for sample selection and the endogeneity of the loan limit using estimates of propensity score and residuals from the loan limit equation:

$$E\left[y_j\middle|x^1, x^2, z, w, w_0, d = 1\right] = \gamma_j(x^1, x^2, w) + \mu(\hat{p}, \hat{v}) \tag{4}$$

4. We estimate the structural form contract term equations corrected for sample selection, endogeneity and simultaneity using the estimates of propensity score, residuals from the loan limit equation and reduced form contract term residuals:

$$E\left[y_j\middle|x^1, x^2, w_j, y_{-j}, z, w_{-j}, w_0, d = 1\right] = g_j\left(x^1, x^2, w_j, y_{-j}\right) + \varphi(\hat{p}, \hat{v}, \hat{e}_{-j}) \tag{5}$$

5. We estimate the probability of default equation corrected for sample selection and the endogeneity of contract terms using the propensity score, residuals from the loan limit equation and structural form residuals:

$$E[def|x^1, x^2, y, z, w, d = 1] = g_{def}(x^1, x^2, y) + \varphi_{def}(\hat{p}, \hat{v}, \hat{e}) \tag{6}$$

Identification conditions for equations (2-6) are formulated with the following theorem.

Theorem 1. *If functions* $g_0(w_0, x_0)$, $\pi(x^1, z)$, $\lambda(p)$, $\gamma_j(x^1, x^2, w)$, $\mu(p, v)$, $g_j\left(x^1, x^2, w_j, y_{-j}\right)$, $\varphi(p, v, e_{-j})$, $g_{def}(x^1, x^2, y)$, $\varphi_{def}(p, v, e)$ *are continuously differentiable with continuous distribution functions almost everywhere and with probability one* $\frac{\partial g_0(w_0, x_0)}{\partial w_0} \neq$

$0, rank\left[\frac{\partial\pi(x^1,z)}{\partial z}\right] = dim(x^2)$ *and for each* $j \in \{1, \dots, k\}$ *at least one* w_j *with* $\frac{\partial\gamma_j(x^1,x^2,w)}{\partial w_j} \neq 0$ *exists then each regression function in (2-6) is identified up to an additive constant.*

Proof. See appendix.

To sum up all the necessary identification conditions, the assumptions of the model restricts the regression function and control function at each step to be functions from different variables and to be separable. The control function also must be a function from the variables which were obtained from the previous steps of estimation procedure.

The first group of Theorem 1 conditions ($\frac{\partial g_0(w_0,x_0)}{\partial w_0} \neq 0$, $rank\left[\frac{\partial\pi(x^1,z)}{\partial z}\right] = dim(x^2)$ and for all $j \in \{1, \dots, k\}$ must be w_j with $\frac{\partial\gamma_j(x^1,x^2,w)}{\partial w_j} \neq 0$) restricts the data. Thus, there must be at least one significant variable in the selection equation excluded from the system and at least one relevant excluded instrument for each endogenous variable (x^2, y).

The last group restricts all regression and control functions to be continuously differentiable.

An estimation procedure is based on an approximation by a series of power functions which depend on the initial set of regressors. This family of regression functions satisfies the differentiability conditions of Theorem 1.

Let $\omega = (\omega_1, \dots, \omega_\chi)$ be a set of variables with $\chi = dim(\omega)$.

$\kappa(\rho, \chi) = \frac{(p+\chi)!}{p!\chi!}$ will be the number of polynomial terms with a power no more than ρ which may be obtained from χ variables.

Let $Q^\rho(\omega) = (q_1(\omega), \dots, q_\kappa(\omega))$ be a vector of κ power functions, which are a full set of polynomial terms with a power no more than ρ obtained from ω, i.e. $q_j(\omega) = \prod_{\tau=1}^{\chi} \omega_\tau^{s_\tau}, \sum_{\tau=1}^{\chi} s_\tau \leq \rho, s_\tau \in \{0,1, \dots, p\}\ \forall\tau = \overline{1,\chi}$.

Let $Q^\rho(\omega)$ be a polynomial approximating series with power ρ.

Then the propensity score of the selection equation may be estimated by OLS as

$$\hat{p_i} = E[d_i|x_{0i}, w_{0i}] = Q^{\rho_0}(w_{0i}, x_{0i})[(Q^{\rho_0}(w_0, x_0))'Q^{\rho_0}(w_0, x_0)]^{-1}(Q^{\rho_0}(w_0, x_0))'d \qquad (7)$$

Let $a = (a_1, a_2)$, $Q^{Z_1,Z_2}(x^1, z, \hat{p}) = (Q^{Z_1}(x^1, z), Q^{Z_2}(\hat{p}))$, then a may be obtained by OLS as

$$\hat{a} = [(Q^{Z_1,Z_2}(x^1, z, \hat{p}))'Q^{Z_1,Z_2}(x^1, z, \hat{p})]^{-1}(Q^{Z_1,Z_2}(x^1, z, \hat{p}))'x^2 \qquad (8)$$

Then the residuals of the loan limit equation may be obtained as

$$\hat{v}_i = x_i^2 - Q^{Z_1,Z_2}(x_i^1, z_i, \hat{p}_i)\hat{a} \qquad (9)$$

Let $b_j = (b_{1j}, b_{2j})$ and $Q^{M_1,M_2}(\mathcal{W}) = Q^{M_1,M_2}(x^1, x^2, w, \hat{p}, \hat{v}) = (Q^{M_1}(x^1, x^2, w), Q^{M_2}(\hat{p}, \hat{v}))$ then b_j may be obtained by OLS as

$$\hat{b}_j = [(Q^{M_1,M_2}(\mathcal{W}))' Q^{M_1,M_2}(\mathcal{W})]^{-1} (Q^{M_1,M_2}(\mathcal{W}))' y_j \quad (10)$$

Then the reduced form contract terms residuals will be

$$\hat{e}_{ji} = y_{ji} - Q^{M_1,M_2}(x_i^1, x_i^2, w_i, \hat{p}_i, \hat{v}_i)\hat{b}_j \quad (11)$$

Let $\beta_j = (\beta_{1j}, \beta_{2j})$ and $Q^{\xi_1,\xi_2}(\mathcal{X}) = Q^{\xi_1,\xi_2}\left(x^1, x^2, w_j, y_{-j}, \hat{p}, \hat{v}, \hat{e}_{-j}\right) = \left(Q^{\xi_1}\left(x^1, x^2, w_j, y_{-j}\right), Q^{\xi_2}\left(\hat{p}, \hat{v}, \hat{e}_{-j}\right)\right)$ then the estimate for β_j may be obtained by OLS as

$$\hat{\beta}_j = [\left(Q^{\xi_1,\xi_2}(\mathcal{X})\right)' Q^{\xi_1,\xi_2}(\mathcal{X})]^{-1} (Q^{\xi_1,\xi_2}(\mathcal{X}))' y_j \quad (12)$$

Then the structural form contract terms residuals will be

$$\hat{e}_{ji} = y_{ji} - Q^{\xi_1,\xi_2}(\mathcal{X})\hat{\beta}_j \quad (13)$$

Let $\alpha = (\alpha_1, \alpha_2)$ and $Q^{\theta_1,\theta_2}(x^1, x^2, y, \hat{p}, \hat{v}, \hat{e}) = \left(Q^{\theta_1}(x^1, x^2, y), Q^{\theta_2}(\hat{p}, \hat{v}, \hat{e})\right)$ then the estimate for α may be obtained by OLS as

$$\hat{\alpha} = [\left(Q^{\theta_1,\theta_2}(x^1, x^2, y, \hat{p}, \hat{v}, \hat{e})\right)' Q^{\theta_1,\theta_2}(x^1, x^2, y, \hat{p}, \hat{v}, \hat{e})]^{-1} * \\ * (Q^{\theta_1,\theta_2}(x^1, x^2, y, \hat{p}, \hat{v}, \hat{e}))' def \quad (14)$$

The next theorem introduces conditions for the consistency of the proposed estimation procedure.

Theorem 2. *If equations (2-6) are identified through theorem 1 and the set of variables* $(w_0, w_1, \dots, w_k, z, x^2)$ *is independent from the distribution of* $(e_0, e_1, \dots, e_k, v, e_{def})$ *then* $\hat{a}, \hat{\alpha}, \hat{b}_j$ *and* $\hat{\beta}_j$ *are consistent.*

Proof. See in Appendix.

Results

Model (1) was estimated with the proposed procedure (7-14).

First, we estimated the model of the probability of a contract agreement based on the characteristics of the borrower and co-borrowers and the difference between the number of AHML refinanced loans and the number of applications.

The last variable which was taken as an excluded instrument is significant at the 1% level. The sign and significance of borrower characteristics are consistent with recent research. The demographic characteristics, such as age, sex, marital status and the level of education of the borrower are insignificant, which supports the absence of discrimination. The probability of a contract agreement is positively correlated with the income of the main borrower and co-borrowers and, on the contrary, negatively correlates with the failure to provide income details. Entrepreneurs have a higher probability of a contract agreement *ceteris paribus*.

These estimates were obtained from the linear probability model and were compared with the probit model. The comparison showed an insignificant difference in the significance of the parameter estimates and predictive power (with slightly higher predictive power for the linear probability model). The propensity score $\widehat{p_i} = E[d_i|x_i, w_{0i}]$ was obtained from the linear probability model.

The model of the logarithm of the loan limit was estimated for all the signed contracts. The excluded instrument (DTI) is significant at the 1% level. We used polynomials up to the third power as approximations for the control function on $\hat{p}$. The hypothesis of its significance was rejected (at 29% level) which suggests there is no selection bias of underwriting on the set of loan limits. The estimated parameters for borrower characteristics are also not counterintuitive. The bank proposed a higher limit for the mid-aged borrowers (38 years) with a higher number of co-borrowers, their income, the higher income of the main borrower, and the higher level of education. Sex, marital status and employment did not affect the loan limit.

For each credit term we estimated the reduced form equation. The control function was approximated by the polynomial with power M_2 on the estimate of the propensity score and the loan limit equation residuals. The regression function was estimated as partially polynomial. It was linear for the characteristics of the borrower and polynomial for the excluded instruments for contract terms with power M_1. The proof of relevance of excluded instruments is provided in Table 4.

Tab. 4. Proof of instrument relevance.

	Eq 1. LTV on mean LTV			Eq. 2. Log. of rate on the median rate		
	(1)	(2)	(3)	(1)	(2)	(3)
Marginal effect	-0.003**	-0.001	0.017	-0.110***	-0.047***	-0.026
	(0.002)	(0.003)	(0.146)	(0.006)	(0.011)	(1.137)
t-stat	1.873	0.411	0.128	17.72	4.196	0.023
k	33	43	63	33	43	63
N	2041	2041	2041	2041	2041	2041
	Eq. 3. Log. of maturity on median maturity			Eq. 4. Probability of insurance on affordability coefficient		
	(1)	(2)	(3)	(1)	(2)	(3)
Marginal effect	0.0016**	-0.002	0.046	-0.580**	-0.701	-0.827
	(0.0009)	(0.002)	(0.117)	(0.288)	(0.623)	(0.715)
t-stat	1.810	0.742	0.392	2.013	1.126	1.157
k	33	43	63	33	43	63
N	2041	2041	2041	2041	2041	2041

Note: In the table cells there are marginal effects of changing of dependent variable on a change of its excluded instrument. Bootstrap standard errors for 100 replications clustered on the month of application are in the parenthesis. Significance level obtained from bootstrap distribution,
* - 10%, ** - 5%, *** - 1%.
k – number of estimated parameters, *N* – number of observations.
For each equation, model (1) was estimated for $M_1 = 1$, $M_2 = 3$, model (2) for $M_1 = 2$, $M_2 = 3$, model (3) for $M_1 = 3$, $M_2 = 3$.
Cragg-Donald Wald *F*-statistics for the joint significance of contract instrument for models (1) is 4.73.

All the excluded instruments are relevant for the use of the first power polynomial for the regression function. The joint significance of the marginal effect of excluded instrument is not rejected at the 5% level. We use the reduced form residuals obtained from the models with $M_1 = 1$, $M_2 = 3$.

We estimated the contract term equations in structural form using a polynomial approximation with power ξ_2 for the control function on $\hat{p}$, residuals from the loan limit equation and the reduced form of the contract term equations. The regression function was partially polynomial, linear for the characteristics of the borrower and polynomial with power ξ_1 for the credit terms and loan limit. Estimation results are provided in Table 5.

The sign and significance for the majority of marginal effects remain the same with the increase of the polynomial power. This supports the robustness of the results. While the marginal effects in models without correction ($\xi_2 = 0$) for sample selection, the endogeneity of the loan limit and the simultaneity of contract term choice are significantly different from the corrected ones (comparing models (3) and (4) for each equation). This result is evidence of the inconsistency of the estimates without correction and the necessity of using the proposed estimation procedure.

The estimates of the marginal effects for the LTV equation are consistent with intuition and recent research. LTV negatively depends on the rate and positively correlates with the maturity, loan limit and having insurance. The demand for government insurance rises with LTV and the rate, for shorter maturity and a lower loan limit. These results are also not counterintuitive. The link between higher risk borrowers and the probability of insurance is also supported by the significantly negative covariance between the error terms in the approval and insurance equations. This means that borrowers that are more likely to be approved are less likely to have insurance. The probability of having government insurance rises with the declared income of the main borrower. This result is not very obvious because insurance is linked with higher risk borrowers which are usually not borrowers with a higher income.

Tab. 5. Estimates for the contract terms equations in structural form.

	Eq 1. LTV				Eq 2. Log. of rate				Eq. 3. Log. of maturity				Eq 4. Probability of insurance			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
LTV	-	-	-	-	-0.431***	-0.359***	-0.232***	-0.018	0.720***	0.046*	0.087***	0.093	0.686***	0.534***	0.104***	1.147***
					(0.055)	(0.047)	(0.026)	(0.037)	(0.143)	(0.036)	(0.016)	(0.120)	(0.089)	(0.040)	(0.022)	(0.052)
Log. of rate	-0.204***	-0.068***	-0.172***	-0.026	-	-	-	-	0.245***	0.276***	0.448***	0.126	0.485***	0.492***	0.436***	0.223***
	(0.017)	(0.013)	(0.018)	(0.043)					(0.046)	(0.018)	(0.008)	(0.209)	(0.018)	(0.019)	(0.018)	(0.043)
Log. of maturity	0.208***	0.161***	0.254***	0.031***	0.157***	0.636***	1.657***	0.025***	-	-	-	-	-0.199***	-0.260***	-0.780***	0.009***
	(0.031)	(0.020)	(0.031)	(0.009)	(0.054)	(0.052)	(0.025)	(0.010)					(0.053)	(0.040)	(0.036)	(0.012)
Probability of insurance	0.396***	0.223***	0.281***	0.301***	0.720***	0.669***	1.041***	0.130	-0.387***	-0.153***	-0.437***	0.045	-	-	-	-
	(0.021)	(0.018)	(0.040)	(0.048)	(0.054)	(0.037)	(0.057)	(0.211)	(0.059)	(0.036)	(0.045)	0.449				
Log. of loan limit	0.115***	0.148***	0.193***	0.120***	-0.355***	-0.333***	0.197***	-0.027***	-0.159***	-0.026*	-0.154***	0.191***	-0.124***	0.075***	-0.133***	0.003
	(0.014)	(0.010)	(0.012)	(0.006)	(0.036)	(0.029)	(0.010)	(0.007)	(0.039)	(0.023)	(0.009)	(0.020)	(0.025)	(0.019)	(0.014)	(0.008)
k	24	60	132	49	24	60	132	49	24	60	132	49	24	60	132	49
N	2041	2041	2041	2041	2041	2041	2041	2041	2041	2041	2041	2041	2041	2041	2041	2041

Note: In the table cells there are marginal effects of changing of dependent variable on a change of another endogenous variable. Bootstrap standard errors for 100 replications clustered on the month of application are in the parenthesis.
Significance level obtained from bootstrap distribution,
* - 10%, ** - 5%, *** - 1%.
k – number of estimated parameters, *N* – number of observations.
For each equation, model (1) was estimated for $\xi_1 = 1$, $\xi_2 = 1$, model (2) for $\xi_1 = 2$, $\xi_2 = 2$, model (3) for $\xi_1 = 3$, $\xi_2 = 3$, model (4) for $\xi_1 = 3$, $\xi_2 = 0$.

Finally we estimated the probability of default equation using a polynomial approximation with power θ_2 for the control function on $\hat{p}$, the residuals from the loan limit equation and structural form contract term equations. The regression function was partially polynomial, linear for the characteristics of the borrower and polynomial with power θ_1 for the credit terms and loan limit. The estimation results are provided in Table 6.

Tab. 6. Estimates for probability of default equation.

	(1)	(2)	(3)
Prop. score from selection equation	-0.082***	-0.038	-0.027
	(0.033)	(0.084)	(0.092)
LTV	-0.006	-0.083**	-0.123*
	(0.034)	(0.049)	(0.082)
Log. of rate	0.561***	0.436***	0.498***
	(0.061)	(0.051)	(0.077)
Log of. maturity	-0.046***	-0.023**	-0.022*
	-0.013	(0.013)	(0.015)
Probability of insurance	0.040*	0.222**	-0.0356
	(0.027)	(0.112)	(0.044)
Log. of loan limit	-0.022	0.096	-0.022
	(0.023)	(0.153)	(0.17)
k	33	68	154
N	2041	2041	2041
% of correct predictions	95.1	95.9	96.4

Note: In the table cells there are marginal effects of changing of dependent variable on a change of another endogenous variable.
Bootstrap standard errors for 100 replications clustered on the month of application are in the parenthesis.
Significance level obtained from bootstrap distribution,
* - 10%, ** - 5%, *** - 1%.
k – number of estimated parameters, N – number of observations.
For each equation, model (1) was estimated for $\theta_1 = 1$, $\theta_2 = 1$, model (2) for $\theta_1 = 2$, $\theta_2 = 2$, model (3) for $\theta_1 = 3$, $\theta_2 = 3$.

The results are also consistent with increasing the power of the approximation. However, increasing the number of parameters improves the predictive power of model but leads to less efficient estimates. The estimate results are intuitive for rate, maturity, propensity score and insurance: a higher rate, less maturity, less probability of approval and higher probability of choosing insurance are linked with a higher probability of default. However, we also have some controversial results. An increase of the co-borrower's declared income decreases the probability of default and at the same time the probability of being approved. It looks like an non-optimal underwriting decision for this category of borrowers, although at the same time it is linked to a lower probability of the loan to be insured.

Table 7 represents the estimation results for the probability of approval, having insurance and the default equation with approximation by polynomials with a power 1. We show only the relevant estimates of the parameters. The table implies that the underwriting process takes into account not only the default probability but the insurance decision too. The probability of contract agreement is negatively and significantly correlated with the probability of default and having insurance. Thus, the probability of having insurance is bad sign for underwriter because it is linked

with a higher probability of default. Then, in order to estimate the potential loss from a delinquent borrower we estimate the difference between the probabilities of default and having insurance. This probability difference shows the probability of a potential default without government insurance. Uninsured defaults cause a loss for the bank, and insured default for government, then the bank will tend to select potential borrowers with the lowest probability difference.

Tab. 7. Estimates for probability of approval, having insurance and default equations.

	Selection	Insurance	Default	Prob. Diff. (Def.-Ins.)
Number of co-borrowers (No co-borrowers is base level):				
1 co-borrower	-0.011	-0.051**	-0.035***	0.015
	(0.016)	(0.032)	(0.006)	(0.039)
2 co-borrowers	0.029*	-0.045	-0.090***	-0.044
	(0.023)	(0.117)	(0.030)	(0.087)
Income of co-borrowers (From 0 to 9999 rub. is base level):				
Not declared	-0.185***	-0.097*	-0.052***	0.045
	(0.052)	(0.091)	(0.022)	(0.069)
From 10000 to 19999 rub.	-0.098***	-0.044	-0.040*	0.004*
	(0.019)	(0.067)	(0.036)	(0.004)
More than 20000 rub.	-0.081***	-0.046*	-0.015*	0.031***
	(0.028)	(0.038)	(0.012)	(0.014)
Income of main borrower (From 0 to 9999 rub. is base level):				
Not declared	-0.020	0.101	-0.116	-0.218*
	(0.046)	(0.125)	(0.142)	(0.190)
From 10000 to 19999 rub.	0.234***	0.094**	0.026	-0.068
	(0.046)	(0.062)	(0.082)	(0.103)
From 20000 to 39999 rub.	0.202***	0.138*	-0.052	-0.190*
	(0.049)	(0.101)	(0.118)	(0.156)
More than 40000 rub.	0.167***	0.210*	-0.020	-0.230
	(0.052)	(0.163)	(0.181)	(0.247)
Prop. score	-	-0.078***	-0.346***	-
		(0.017)	(0.084)	
Prob. of insurance	-	-	0.031*	-
			(0.019)	
k	24	31	33	-
N	3366	2041	2041	

Note: In the table cells there are estimates of parameters.
Bootstrap standard errors for 100 replications clustered on the month of application are in the parenthesis.
Significance level obtained from bootstrap distribution,
* - 10%, ** - 5%, *** - 1%.
k – number of estimated parameters, *N* – number of observations.

Table 7 shows that relatively riskier borrowers (in terms of the probability difference) are less likely to be selected by bank and *vice versa*. Borrowers with insignificant probability difference in all cases are also more likely to be approved. Moreover, not declaring the income of the main borrower is not a risk factor because it has significant, small but negative probability difference. The same is for the co-borrower's income. It does not affect the probability difference but these borrowers are less likely to be approved by bank. This means that bank need to improve the underwriting process and approve more borrowers without declared income.

As the result, government insurance on AHML loans tends to be an important determinant of bank and borrower decisions. The bank takes into account the predicted probability of insurance along with the probability of borrower default. In order to compensate for a high probability of default and to have more chance of being approved, potential borrowers choose to insure the loan, especially when loans have riskier terms such as high LTV, rate and short maturity.

Conclusion

This paper analyzes the borrowing process in one Russian bank which is a regional subsidiary of AHML, a national provider of residential housing mortgages. This analysis takes into account the underwriting process and the choice of loan limit by the bank, the choice of contract terms including having government insurance and the performance of all loans issued by the bank from 2008 to 2012. The dataset contains information about the demographic and financial characteristics of the borrower for all applications, the loan limit set by bank, the contract terms and the property value, and the indicator of default for all signed contracts. We also used regional-level aggregated housing and mortgage market characteristics as instrumental variables for the selection equation and endogenous variables.

We model the demand for loans as a simultaneous choice of loan terms and represent this as a system of simultaneous equations. We observe the choice only for those borrowers who were approved by the bank and chose to get a mortgage from this particular bank. While approving the borrower the bank sets the loan limit which affects the choice of credit terms and is considered an endogenous variable. This structure of borrowing process determines the use of the multistep nonparametric approach.

The main finding is that the probability of having government insurance is linked to riskier loans, such as loans with higher LTR, higher interest rate and lower maturity. Insured loans also are more likely to be approved by the bank. The bank, when approving a borrower, takes into account not the probability of default, but the difference between the probabilities of default and having government insurance. The probability of having insurance positively correlates with the probability of default. It may be explained in two ways: 1) Riskier borrowers in order to compensate for their credit risk choose to have insurance to increase the probability of being approved; 2) Riskier borrowers make a strategic choice to be insured and try to reduce their loss via a potential default.

The obtained estimates depend on the data. We used data from only one regional operator of AHML programs and do not have enough space variation. Our dataset is not big enough to apply nonparametric procedures with high-order polynomial approximations for regression and

correction functions. Therefore the estimates with increasing polynomial order remains consistent but is inefficient. However, we may rely on the obtained results since the estimation procedure is based on the minimum assumptions for the consistency of estimates.

References

Andrews, D. W., Schafgans, M. M. (1998). Semiparametric Estimation of the Intercept of a Sample Selection Model. *The Review of Economic Studies*, *65*(3), 497-517.

Ambrose, B., LaCour-Little M., Sanders, A. (2004). The Effect of Conforming Loan Status on Mortgage Yield Spreads: A Loan Level Analysis. *Real Estate Economics*, 32(4), 541–569.

Attanasio, O.P., Goldberg, P.K., Kyriazidou, E. (2008). Credit Constraints in the Market for Consumer Durables: Evidence from Micro Data on Car Loans. *International Economic Review*, 49(2), 401–436.

Bajari, P., Chu, C. S., Park, M. (2008). An Empirical Model of Subprime Mortgage Default from 2000 to 2007. *NBER working paper* 14625.

Das, M., Newey, W.K., Vella, F. (2003). Nonparametric Estimation of Sample Selection Models. *The Review of Economic Studies*, 70(1), 33–58.

Follain, J. R. (1990). Mortgage Choice. *Real Estate Economics*, 18(2), 125–144.

Gronau, R. (1973). Wage Comparisons: a Selectivity Bias. *NBER Working Paper №13.*

Heckman, J. (1974). Shadow Prices, Market Wages, and Labor Supply. *Econometrica: journal of the econometric society*, 679-694.

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

Heckman, J. (1979). Sample Selection Bias as a Specification Error. *Econometrica: Journal of Econometric Society*, 47(1), 153–161.

Heckman, J. (1990). Varieties of Selection Bias. *The American Economic Review*, 313-318.

Heckman, J. J., Robb Jr, R. (1985). Alternative Methods for Evaluating the Impact of Interventions: An Overview. *Journal of Econometrics*, *30*(1), 239-267.

LaCour-Little, M. (2007). The Home Purchase Mortgage Preferences of Low- and Moderate-Income Households. *Real Estate Economics*, 35, 265-290.

Munnell, A., G. Tootell, L. Browne, McEneaney, J. (1996). Mortgage Lending in Boston: Interpreting HMDA Data. *American Economic Review*, 86, 25–53.

Newey, W. K. (1997). Convergence Rates and Asymptotic Normality for Series Estimators. *Journal of Econometrics*, *79*(1), 147-168.

Newey, W. K. (1999). Two-step Series Estimation of Sample Selection Models. *Working paper, MIT, Department of Economics*.

Newey, W. K. (2013). Nonparametric Instrumental Variables Estimation. *The American Economic Review*, *103*(3), 550-556.

Newey, W. K., Powell, J. L. (1989). Nonparametric Instrumental Variables Estimation. *Working paper, MIT, Department of Economics*.

Newey, W. K., Powell, J. L., Vella, F. (1999). Nonparametric Estimation of Triangular Simultaneous Equations Models. *Econometrica*, *67*(3), 565-603.

Ozhegov E.M. (2014). Modelling Demand for Mortgage Loans Using Loan-Level Data. In: S.V. Ivliev, A.K. Bera, F.Lillo (ed.). *Financial Econometrics and Empirical Market Microstructure*, Springer.

Ozhegov E. M., Poroshina A. M. (2013). The Lagged Structure of Dynamic Demand Function for Mortgage Loans in Russia. *EJournal of Corporate Finance*, 27, 37-49.

Phillips, R., Yezer, A. (1996). Self-Selection and Tests for Bias and Risk in Mortgage Lending: Can You Price the Mortgage If You Don't Know the Process? *Journal of Real Estate Research*, 11, 87–102.

Rachlis, M., Yezer A. (1993). Serious Flaws in Statistical Tests for Discrimination in Mortgage Markets. *Journal of Housing Research*, 4, 315–336.

Ross, S.L. (2000). Mortgage Lending, Sample Selection and Default. *Real Estate Economics*, 8, 581–621.

Vella, F. (1993). A Simple Estimator for Simultaneous Models with Censored Endogenous Regressors. *International Economic Review*, 441-457.

Vella, F. (1998). Estimating Models with Sample Selection Bias: A Survey.*Journal of Human Resources*, *33*(1).

Yezer, A., Philips, R., Trost R. (1994). Bias in Estimates of Discrimination and Default in Mortgage Lending: the Effects of Simultaneity and Self-Selection. *Journal of Real Estate Finance and Economics*, 9, 197–215.

Appendix

Tab. A1. Estimated parameters for selection equation.

Variable	(1) OLS	(2) Probit
Age of borrower	-0.004	-0.012
	(0.009)	(0.024)
Age squared	0.000	0.000
	(0.000)	(0.000)
Male	0.026	0.075
	(0.018)	(0.050)
Family status (Single is base level):		
Married	0.030	0.094
	(0.025)	(0.070)
Divorced	-0.104	-0.278
	(0.075)	(0.204)
Widowed	-0.012	-0.034
	(0.027)	(0.074)
Category of activity (Hired employee is base level):		
Entrepreneur	0.086	0.246
	(0.095)	(0.287)
State employee	0.133***	0.377***
	(0.045)	(0.131)
Level of education (Elementary is base level):		
Secondary education	-0.060	-0.174
	(0.066)	(0.185)
Incomplete higher education	-0.076	-0.208
	(0.077)	(0.216)
Complete higher education	0.013	0.026
	(0.066)	(0.184)
Number of co-borrowers (No co-borrowers is base level)		
1 co-borrower	-0.004	-0.027
	(0.024)	(0.068)
2 co-borrowers	0.020	0.053
	(0.048)	(0.139)
Declared income of co-borrowers (From 0 to 9999 rub. is base level):		
Not declared	-0.193***	-0.852***
	(0.050)	(0.194)
From 10000 to 19999 rub.	-0.087	-0.507
	(0.058)	(0.617)
More than 20000 rub.	-0.115	-0.596
	(0.261)	(0.722)
Declare income of main borrower (From 0 to 9999 is base level):		
Not declared	-0.019	0.004
	(0.053)	(0.147)
From 10000 to 19999 rub.	0.180***	0.531***
	(0.061)	(0.172)
From 20000 to 39999 rub.	0.229***	0.700***
	(0.055)	(0.156)
More than 40000 rub.	0.257***	0.830***
	(0.081)	(0.151)
Difference between AHML loans number and number of applications	-0.000***	-0.001***
	(0.000)	(0.000)

Constant	0.781*** (0.178)	1.028** (0.516)
N	3366	3366
k	22	22
% of correct predictions	64.3	63.7
F-statistics for difference between the number of refinanced loans and applications	13.59	12.13

Note: Robust standard errors are in parenthesis,
significance level obtained from *t*-statistics,
* - 10%, ** - 5%, *** - 1%.
k – number of estimated parameters, *N* – number of observations

Tab. A2. Estimated parameters for loan limit equation.

Variable	(1) OLS
Age of borrower	0.017* (0.010)
Age squared	-0.000* (0.000)
Male	-0.006 (0.021)
Family status (Single is base level):	
Married	0.045 (0.028)
Divorced	-0.041 (0.098)
Widowed	-0.005 (0.031)
Category of activity (Hired employee is base level):	
Entrepreneur	0.076 (0.098)
State employee	-0.066 (0.054)
Level of education (Elementary s base level):	
Secondary education	0.044 (0.074)
Incomplete higher education	0.255*** (0.090)
Complete higher education	0.225*** (0.073)
Number of co-borrowers (No co-borrowers is base level)	
1 co-borrower	0.082*** (0.027)
2 co-borrowers	0.133** (0.053)
Declared income of co-borrowers (From 0 to 9999 rub. is base level):	
Not declared	0.047 (0.064)
From 10000 to 19999 rub.	0.093 (0.059)
More than 20000 rub.	0.259*** (0.064)

Declare income of main borrower (From 0 to 9999 is base level):	
Not declared	0.942***
	(0.065)
From 10000 to 19999 rub.	0.486***
	(0.081)
From 20000 to 39999 rub.	0.893***
	(0.078)
More than 40000 rub.	1.346***
	(0.078)
Mean DTI	-0.000***
	(0.000)
Prop. score	2.292
	(1.736)
Prop. score squared	-4.977
	(3.766)
Prop. score cubed	3.136
	(2.600)
Constant	9.379***
	(0.750)
N	2041
K	25
F-statistics for mean DTI	20.72

Note: Robust standard errors are in parenthesis,
significance level obtained from *t*-statistics,
* - 10%, ** - 5%, *** - 1%.
k – number of estimated parameters, *N* – number of observations

Lemma 1. *If functions* $g_0(w_0, x_0)$, $\pi(x^1, z)$, $\lambda(p)$ *are continuously differentiable with continuous distribution functions almost everywhere and with probability one* $\frac{\partial g_0(w_0,x_0)}{\partial w_0} \neq 0$, *then* $\pi(x^1, z)$ *is identified up to an additive constant.*

Proof (is similar to T.2.1 in Das et al. (2003)): Any observationally equivalent model for (3) must have $E[x^2|x^1, z, w_0, d = 1] = \hat{\pi}(x^1, z) + \hat{\lambda}(p)$. Consider $f_1(x^1, z) + f_2(p) = 0$, where $f_1(x^1, z) = \pi(x^1, z) - \hat{\pi}(x^1, z)$, and $f_2(p) = \lambda(p) - \hat{\lambda}(p)$. If g_0, π and λ are differentiable, then f_1 and f_2 are also differentiable. Then we may differentiate $f_1 + f_2 = 0$ by the set of (w_0, x^1, z):

$$0 = \frac{\partial f_2(p)}{\partial p}\frac{\partial g_0(w_0,x_0)}{\partial w_0}$$

$$0 = \frac{\partial f_1(x^1,z)}{\partial x^1} + \frac{\partial f_2(p)}{\partial p}\frac{\partial p(w_0,x_0)}{\partial x^1} \quad \text{(A.1)}$$

$$0 = \frac{\partial f_1(x^1,z)}{\partial z}$$

First condition and $\frac{\partial g_0(w_0,x_0)}{\partial w_0} \neq 0$ implies $\frac{\partial f_2(p)}{\partial p} = 0$, then f_2 is constant.

Then the second condition gives $\frac{\partial f_1(x^1,z)}{\partial(x,z)} = 0$. It means that $f_1(x^1,z)$ is constant and $\hat{\pi}(x^1,z) = \pi(x^1,z) + \mathrm{C}$. ||

Lemma 2. *If functions* $g_0(w_0,x_0)$, $\pi(x^1,z)$, $\lambda(p)$, $\gamma_j(x^1,x^2,w)$, $\mu(p,v)$ *are continuously differentiable with continuous distribution functions almost everywhere and with probability one* $\frac{\partial g_0(w_0,x_0)}{\partial w_0} \neq 0$ *and* $rank\left[\frac{\partial \pi(x^1,z)}{\partial z}\right] = dim(x^2)$, *then every* $\gamma_j(x^1,x^2,w)$ *is identified up to an additive constant.*

Proof: Any observationally equivalent model for (4) must have $E\left[y_j \middle| x^1,x^2,z,w_0,d=1\right] = \hat{\gamma}_j(x^1,x^2,w) + \hat{\mu}(p,v)$. Consider $f_3(x^1,x^2,w) + f_4(p,v) = 0$ where $f_3(x^1,x^2,w) = \gamma_j(x^1,x^2,w) - \hat{\gamma}_j(x^1,x^2,w)$, and $f_4(p,v) = \mu(p,v) - \hat{\mu}(p,v)$. If all the conditions of this lemma are met, then so are the ones of lemma 1 then $\pi(x^1,z)$ and v are identified up to an additive constant.

If g_0, π, λ, γ_j and μ are differentiable then $f_3(x^1,x^2,w)$ and $f_4(p(w_0,x_0),v)$ are also differentiable. Then we may differentiate $f_3 + f_4 = 0$ on the set of variables (x^1,z,w,w_0):

$$
\begin{aligned}
0 &= \frac{\partial f_3(x^1,x^2,w)}{\partial x^1} + \frac{\partial f_3(x^1,x^2,w)}{\partial x^2}\frac{\partial \pi(x^1,z)}{\partial x^1} + \frac{\partial f_4(p(w_0,x_0),v)}{\partial p}\frac{\partial p(w_0,x_0)}{\partial x^1} \\
0 &= \frac{\partial f_3(x^1,x^2,w)}{\partial x^2}\frac{\partial \pi(x^1,z)}{\partial z} + \frac{\partial f_4(p(w_0,x_0),v)}{\partial p}\frac{\partial p(w_0,x_0)}{\partial z} \\
0 &= \frac{\partial f_3(x^1,x^2,w)}{\partial w} \\
0 &= \frac{\partial f_4(p(w_0,x_0),v)}{\partial p}\frac{\partial p(w_0,x_0)}{\partial w_0}
\end{aligned}
\qquad \text{(A.2)}
$$

$\frac{\partial g_0(w_0,x_0)}{\partial w_0} \neq 0$ and the last condition imply $\frac{\partial f_4(p(w_0,x_0),v)}{\partial p} = 0$.

Then the second condition and $rank\left[\frac{\partial \pi(x^1,z)}{\partial z}\right] = dim(x^2)$ give $\frac{\partial f_3(x^1,x^2)}{\partial x^2} = 0$.

The first condition implies $\frac{\partial f_3(x^1,x^2)}{\partial x^1} = 0$. It means that $f_3(x^1,x^2,w) = \gamma_j(x^1,x^2,w) - \hat{\gamma}_j(x^1,x^2,w)$ is constant, $\hat{\gamma}_j(x^1,x^2,w) = \gamma_j(x^1,x^2,w) + \mathrm{C}_j$. ||

Lemma 3. *If functions* $g_0(w_0, x_0)$, $\pi(x^1, z)$, $\lambda(p)$, $\gamma_j(x^1, x^2, w)$, $\mu(p, v)$, $g_j(x^1, x^2, w_j, y_{-j})$, $\varphi(p, v, e_{-j})$ *are continuously differentiable with continuous distribution functions almost everywhere and with probability one* $\frac{\partial g_0(w_0, x_0)}{\partial w_0} \neq 0$, $rank\left[\frac{\partial \pi(x^1, z)}{\partial z}\right] = dim(x^2)$ *and for each* $j \in \{1, \dots, k\}$ *at least one* w_j *with* $\frac{\partial \gamma_j(x^1, x^2, w)}{\partial w_j} \neq 0$ *exists then each* $g_j(x^1, x^2, w_j, y_{-j})$ *is identified up to an additive constant.*

Proof. As soon as the conditions of the lemma are more rigid than the ones of the lemmas 1 and 2 are also satisfied then $\pi(x^1, z)$, $\gamma_{-j}(x^1, x^2)$, v and e_{-j} are identified up to a set of constants.

Any observationally equivalent model for (5) must have $E[y_j | x^1, x^2, y_{-j}, z, w, w_0, d = 1] = \hat{g}_j(x^1, x^2, w_j, y_{-j}) + \hat{\varphi}(p, v, e_{-j})$. Consider $f_5(x^1, x^2, w_j, y_{-j}) + f_6(p, v, e_{-j}) = 0$, where $f_5(x^1, x^2, w_j, y_{-j}) = g_j(x^1, x^2, w_j, y_{-j}) - \hat{g}_j(x^1, x^2, w_j, y_{-j})$ and $f_6(p, v, e_{-j}) = \varphi(p, v, e_{-j}) - \hat{\varphi}(p, v, e_{-j})$.

If g_0, π, λ, γ_j, μ, g_j, φ are continuously differentiable then $f_5(x^1, x^2, w_j, y_{-j})$ and $f_6(p, v, e_{-j})$ are also continuously differentiable then we may differentiate $f_5 + f_6 = 0$ by the set of exogenous variables $(w_j, w_{-j}, x^1, z, w_0)$:

$$
\begin{aligned}
0 &= \frac{\partial f_5(x^1, x^2, w_j, y_{-j})}{\partial w_j} + \frac{\partial f_5(x^1, x^2, w_j, y_{-j})}{\partial y_{-j}} \frac{\partial \gamma_{-j}(x^1, x^2, w)}{\partial w_j} \\
0 &= \frac{\partial f_5(x^1, x^2, w_j, y_{-j})}{\partial y_{-j}} \frac{\partial \gamma_{-j}(x^1, x^2, w)}{\partial w_{-j}} \\
0 &= \frac{\partial f_5(x^1, x^2, w_j, y_{-j})}{\partial x^1} + \frac{\partial f_5(x^1, x^2, w_j, y_{-j})}{\partial x^2} \frac{\partial \pi(x^1, z)}{\partial x^1} + \frac{\partial f_5(x^1, x^2, w_j, y_{-j})}{\partial y_{-j}} \left[\frac{\partial \gamma_{-j}(x^1, x^2, w)}{\partial x^1} + \right. \\
&\left. + \frac{\partial \gamma_{-j}(x^1, x^2, w)}{\partial x^2} \frac{\partial \pi(x^1, z)}{\partial x^1}\right] + \frac{\partial f_6(p, v, e_{-j})}{\partial p} \frac{\partial g_0(x_0, w_0)}{\partial x^1} \\
0 &= \frac{\partial \pi(x^1, z)}{\partial z} \left[\frac{\partial f_5(x^1, x^2, w_j, y_{-j})}{\partial x^2} + \frac{\partial f_5(x^1, x^2, w_j, y_{-j})}{\partial y_{-j}} \frac{\partial \gamma_{-j}(x^1, x^2, w)}{\partial x^2}\right] \\
0 &= \frac{\partial f_6(p, v, e_{-j})}{\partial p} \frac{\partial g_0(x_0, w_0)}{\partial w_0}
\end{aligned}
\tag{A.3}
$$

The last condition and $\frac{\partial g_0(x_0, w_0)}{\partial w_0} \neq 0$ imply that $\frac{\partial f_6(p, v, e_{-j})}{\partial p} = 0$.

As soon as for every j there is w_j with $\frac{\partial \gamma_j(x^1,x^2,y_{-j})}{\partial w_j} \neq 0$ give $\frac{\partial \gamma_{-j}(x^1,x^2,w)}{\partial w} \neq 0$ and make the second condition equivalent to $\frac{\partial f_5(x^1,x^2,w_j,y_{-j})}{\partial y_{-j}} = 0$.

Replacing $\frac{\partial f_5(x^1,x^2,w_j,y_{-j})}{\partial y_{-j}} = 0$ in the fourth condition and using $\frac{\partial \pi(x^1,z)}{\partial z} \neq 0$ we have $\frac{\partial f_5(x^1,x^2,w_j,y_{-j})}{\partial x^2} = 0$.

And $\frac{\partial f_5(x^1,x^2,w_j,y_{-j})}{\partial y_{-j}} = 0$ in the first condition gives $\frac{\partial f_5(x^1,x^2,w_j,y_{-j})}{\partial w_j} = 0$.

All the obtained results in the third condition give $\frac{\partial f_5(x^1,x^2,w_j,y_{-j})}{\partial x^1} = 0$ which implies that $f_5(x^1,x^2,w_j,y_{-j}) = g_j(x^1,x^2,w_j,y_{-j}) - \hat{g}_j(x^1,x^2,w_j,y_{-j})$ is constant, consequently $\hat{g}_j(x^1,x^2,w_j,y_{-j}) = g_j(x^1,x^2,w_j,y_{-j}) + \mathrm{C}_j'$. ‖

Proof of Theorem 1. By lemmas 1-3 equations (2-5) is identified. Let us prove the identification of equation (6).

Any observationally equivalent model for (5) must have $E[def|x^1,x^2,y,z,w,d=1] = \hat{g}_{def}(x^1,x^2,y) + \hat{\varphi}_{def}(p,v,e)$. Consider $f_7(x^1,x^2,y) + f_8(p,v,e) = 0$, where $f_7(x^1,x^2,y) = g_{def}(x^1,x^2,y) - \hat{g}_{def}(x^1,x^2,y)$ and $f_8(p,v,e) = 0 = \varphi_{def}(p,v,e) - \hat{\varphi}_{def}(p,v,e)$.

If g_{def} and φ_{def} are continuously differentiable then $f_7(x^1,x^2,y)$ and $f_8(p,v,e)$ are also continuously differentiable then we may differentiate $f_7 + f_8 = 0$ by the set of exogenous variables (w,x^1,z,w_0):

$$0 = \frac{\partial f_7(x^1,x^2,y)}{\partial y_j}\frac{\partial g_j(y_{-j},x^1,x^2,w_j)}{\partial w_j}$$

$$0 = \frac{\partial f_7(x^1,x^2,y)}{\partial x^1} + \frac{\partial f_7(x^1,x^2,y)}{\partial x^2}\frac{\partial \pi(x^1,z)}{\partial x^1} + \frac{\partial f_7(x^1,x^2,y)}{\partial y_j}\left[\frac{\partial g_j(y_{-j},x^1,x^2,w_j)}{\partial x^1} + \right. \quad \text{(A.5)}$$

$$\left. + \frac{\partial g_j(y_{-j},x^1,x^2,w_j)}{\partial x^2}\frac{\partial \pi(x^1,z)}{\partial x^1}\right] + \frac{\partial f_8(p,v,e)}{\partial p}\frac{\partial g_0(x_0,w_0)}{\partial x^1}$$

$$0 = \frac{\partial \pi(x^1,z)}{\partial z}\left[\frac{\partial f_7(x^1,x^2,y)}{\partial x^2} + \frac{\partial f_7(x^1,x^2,y)}{\partial y}\frac{\partial g_j(y_{-j},x^1,x^2,w_j)}{\partial x^2}\right]$$

$$0 = \frac{\partial f_8(p,v,e)}{\partial p}\frac{\partial g_0(x_0,w_0)}{\partial w_0}$$

The last condition and $\frac{\partial g_0(x_0,w_0)}{\partial w_0} \neq 0$ imply that $\frac{\partial f_8(p,v,e)}{\partial p} = 0$.

As soon as for every j there is w_j with $\frac{\partial \gamma_j(x^1,x^2,y_{-j})}{\partial w_j} \neq 0$ give $\frac{\partial g_j(y_{-j},x^1,x^2,w_j)}{\partial w_j} \neq 0$ and make the first condition equivalent to $\frac{\partial f_7(x^1,x^2,y)}{\partial y_j} = 0$.

Replacing $\frac{\partial f_7(x^1,x^2,y)}{\partial y_j} = 0$ in the fourth condition and using $\frac{\partial \pi(x^1,z)}{\partial z} \neq 0$ we have $\frac{\partial f_7(x^1,x^2,y)}{\partial x^2} = 0$.

And then the second condition gives $\frac{\partial f_7(x^1,x^2,y)}{\partial x^1} = 0$. This means that $\frac{\partial f_7(x^1,x^2,y)}{\partial (x^1,x^2,y)} = 0$ and $f_7(x^1,x^2,y)$ is constant. $\hat{g}_{def}(x^1,x^2,y) = g_{def}(x^1,x^2,y) + C_{def}$ means that $g_{def}(x^1,x^2,y)$ is identified up to additive constant.‖

Proof of Theorem 2. Consider a procedure of model (1) identification. It will take 4 steps:

1. On the first step we will estimate the propensity score $p = E[d|x_0,w_0]$ from the selection equation:

$$d_i = \begin{cases} 1, g_0(w_{0i},x_{0i}) + e_{0i} \geq 0 \\ 0, g_0(w_{0i},x_{0i}) + e_{0i} < 0 \end{cases} \quad \text{(A.6)}$$

For every marginal distribution f_{e_0}, $E[d|x_0,w_0] = E[d=1|x_0,w_0] = \int_{-g_0(w_0,x_0)}^{\infty} f_{e_0}(s)ds = \gamma_0(w_0,x_0)$. γ_0 with arbitrary distribution of e_0 and functional form of g_0 will be a function with arbitrary functional form but will depend only on the known set of variables, w_0, x_0.

We may decompose γ_0 into the Taylor series in a neighborhood of each (w_{0i},x_{0i}). $p_i = E[d_i|x_{0i},w_{0i}]$ may be approximated by a polynom $Q^{\rho_0}(w_{0i},x_{0i})\alpha_0$, where $Q^{\rho_0}(w_0,x_0)$ is

polynomial approximating series for $\gamma_0(w_0, x_0)$ with ρ_0 and α_0 is a vector of parameters with dimensiality $\kappa = \frac{(\rho_0+\chi_0)!}{\rho_0!\chi_0!}$, $\chi_0 = dim(w_0, x_0)$.

Estimate of α_0 may be obtained by OLS as

$$\hat{\alpha}_0 = [(Q^{\rho_0}(w_0, x_0))' Q^{\rho_0}(w_0, x_0)]^{-1} (Q^{\rho_0}(w_0, x_0))' d \qquad \text{(A.7)}$$

For all fixed ρ_0 we may prove the consistency of $\hat{\alpha}_0$.

$$\begin{aligned} &\operatorname*{plim}_{n\to\infty} \hat{\alpha}_0 = \operatorname*{plim}_{n\to\infty} [(Q^{\rho_0}(w_0, x_0))' Q^{\rho_0}(w_0, x_0)]^{-1} (Q^{\rho_0}(w_0, x_0))' d = \\ &= \operatorname*{plim}_{n\to\infty} [(Q^{\rho_0}(w_0, x_0))' Q^{\rho_0}(w_0, x_0)]^{-1} (Q^{\rho_0}(w_0, x_0))' (Q^{\rho_0}(w_0, x_0)\alpha_0 + \eta_0) \qquad \text{(A.8)} \\ &= \alpha_0 + \operatorname*{plim}_{n\to\infty} [(Q^{\rho_0}(w_0, x_0))' Q^{\rho_0}(w_0, x_0)]^{-1} (Q^{\rho_0}(w_0, x_0))' \eta_0 = \alpha_0 \end{aligned}$$

with the exogeneity of (w_0, x_0).

This is obvious that a convergence speed to true $\gamma_0(w_0, x_0)$ depends on the power ρ_0 of approximation function. The higher ρ_0 gives the slower speed of convergence due to increase in the number of parameters being estimated. Das et al. (2003) showed that with the upper limit to an approximation polynom power the estimate is asymptotically normal. In this paper we will not prove the asymptotic normality and return to the issue of standard errors calculation in results section. In this section we point out that it may be obtained by bootstrap. The basics of asymptotic theory for two-step correction procedures provided by Newey (1997). It is also mentioned in Das et al. (2003) that regression function may be represented as partially linear in regressors then all identification conditions should be held only for nonlinear part of regression function. Then the assumption of differentiability of regression functions may be relaxed when we include all discrete regressors only to linear part of regression function.

Then the propensity score will be

$$\widehat{p_i} = E[d_i | x_{0i}, w_{0i}] = Q^{\rho_0}(w_{0i}, x_{0i})[(Q^{\rho_0}(w_0, x_0))' Q^{\rho_0}(w_0, x_0)]^{-1} (Q^{\rho_0}(w_0, x_0))' d \qquad \text{(A.9)}$$

2. On the second step we will estimate the residuals from endogenous variables equations corrected for sample selection:

$$\nu = x^2 - E[x^2|x^1, z, w_0, d = 1] \tag{A.10}$$

For every marginal joint distribution of e_0 and ν, $f_{e_0,\nu}$:

$$E[\nu|x^1, z, w_0, d = 1] = E[\nu|x^1, z, w_0, d = 1] = E[\nu|g_0(w_0, x_0)+e_0 \geq 0] =$$
$$= \int_{-\infty}^{\infty}\int_{-g_0(w_0,x_0)}^{\infty} \nu f_{e_0,\nu}(s, r)ds\, dr = \lambda(p) \tag{A.11}$$

where λ is a function on propensity score with arbitrary function form.

If $\hat{p}$ is a predicted propensity score obtained on the previous step then $\hat{p}$ on this step will be fixed. (x^1, z) and $\hat{p}$ are the sets of different variables since there is at least one w_0 with $\frac{\partial Q^{\rho_0}(w_{0i},x_{0i})\hat{a}_0}{\partial w_0} \neq 0$.

Every arbitrary functions $\pi(x^1, z)$ and $\lambda(\hat{p})$ may be approximated by $Q^{Z_1}(x^1, z)a_1$ and $Q^{Z_2}(\hat{p})a_2$ where $Q^{Z_1}(x^1, z)$ is polynomial approximation series with a power Z_1, $Q^{Z_2}(\hat{p})$ is polynomial approximation series with a power Z_2, then x^2 may be approximated by

$$x^2 = Q^{Z_1}(x^1, z)a_1 + Q^{Z_2}(\hat{p})a_2 + \eta_z \tag{A.12}$$

Equation (A.12) is identified up to an additive constant due to the Theorem 1 since polynomial approximations for π and λ are continuously differentiable and $\frac{\partial Q^{\rho_0}(w_{0i},x_{0i})\hat{a}_0}{\partial w_0} \neq 0$.

Let $a = (a_1, a_2)$, $Q^{Z_1,Z_2}(x^1, z, \hat{p}) = (Q^{Z_1}(x^1, z), Q^{Z_2}(\hat{p}))$, then a may be obtained by OLS as

$$\hat{a} = [(Q^{Z_1,Z_2}(x^1, z, \hat{p}))' Q^{Z_1,Z_2}(x^1, z, \hat{p})]^{-1}(Q^{Z_1,Z_2}(x^1, z, \hat{p}))' x^2 \tag{A.13}$$

For some large enough Z_1, Z_2, $Q^{Z_1}(x^1, z)\hat{a}_1$ will be an approximation for $\pi(x^1, z)$. And $\hat{a} = (\hat{a}_1, \hat{a}_2)$ will be consistent with the exogeneity of $(x^1, z, \hat{p})$ due to

$$\operatorname*{plim}_{n\to\infty} \hat{a} = \operatorname*{plim}_{n\to\infty}[(Q^{Z_1,Z_2}(x^1, z, \hat{p}))' Q^{Z_1,Z_2}(x^1, z, \hat{p})]^{-1}(Q^{Z_1,Z_2}(x^1, z, \hat{p}))' x^2 =$$
$$= \operatorname*{plim}_{n\to\infty}[(Q^{Z_1,Z_2}(x^1, z, \hat{p}))' Q^{Z_1,Z_2}(x^1, z, \hat{p})]^{-1}(Q^{Z_1,Z_2}(x^1, z, \hat{p}))'(Q^{Z_1,Z_2}(x^1, z, \hat{p})a \tag{A.14}$$
$$+ \eta_z) = a + \operatorname*{plim}_{n\to\infty}[(Q^{Z_1,Z_2}(x^1, z, \hat{p}))' Q^{Z_1,Z_2}(x^1, z, \hat{p})]^{-1}(Q^{Z_1,Z_2}(x^1, z, \hat{p}))' \eta_z = a$$

Identification of an additive constant in this equation is an additional research question when its true value is a point of interest. Heckman (1990) provided examples when

identification of constant is essential. Andrews and Schafgans (1998) discussed also the identification strategy. When the identification of constant is not a point of interest then we only need to fix a value of some parameter. For example, let the parameter behind $(\hat{p})^0$ in $Q^{Z_2}(\hat{p})$ be equal to 0. On the next steps we will also put 0 as a value of parameter behind the polynomial term with 0 power in control function.

Then the residuals of endogenous variables equations may be obtained as

$$\hat{v}_i = x_i^2 - Q^{Z_1,Z_2}(x_i^1, z_i, \hat{p}_i)\hat{a} \tag{A.15}$$

3. On the third step we will estimate the reduced form residuals corrected for sample selection and endogeneity of x^2:

$$e_j = y_j - E\left[y_j \middle| x^1, x^2, z, w, w_0, d = 1\right] \tag{A.16}$$

If e_j has joint marginal distribution with v and e_0 with density function $f_{e_0,e_j,v}$ then

$$\begin{aligned} E[e_j|x^1, x^2, z, w, w_0, d = 1] &= E[e_j|v, g_0(w_0, x_0) + e_0 \geq 0] \\ &= \int_{-\infty}^{\infty}\int_{-g_0(w_0,x_0)}^{\infty} e_j f_{e_0,e_j,v}(s, r|v)dsdr = \mu(p, v) \end{aligned} \tag{A.17}$$

y_j is decomposed into regression and control functions:

$$y_j = \gamma_j(x^1, x^2, w) + \mu_j(p, v) + \eta_j \tag{A.18}$$

The error term in this equation η_j is independent on (x^1, x^2, w).

If $\hat{p}$ is a propensity score and $\hat{v}$ is a residuals of endogenous variables equations then on this stage $\hat{p}$ and $\hat{v}$ will be fixed. (x^1, x^2, w) and $(\hat{p}, \hat{v})$ are two sets of different variables if $\frac{\partial Q^{\rho_0}(w_{0i}, x_{0i})\hat{a}_0}{\partial w_0} \neq 0$ and $rank(\frac{\partial Q^{Z_1}(x^1,z)\hat{a}_1}{\partial z}) = dim(x^2)$.

Every arbitrary functions $\gamma_j(x^1, x^2, w)$ and $\mu_j(\hat{p}, \hat{v})$ may be approximated by $Q^{M_1}(x^1, x^2, w)b_{1j}$ and $Q^{M_2}(\hat{p}, \hat{v})b_{2j}$ respectively, where $Q^{M_1}(x^1, x^2, w)$ is polynomial approximating series with a power M_1, $Q^{M_2}(\hat{p}, \hat{v})$ is polynomial approximating series with a power M_2. Then y_j may be approximated by the following equation:

$$y_j = Q^{M_1}(x^1, x^2, w)b_{1j} + Q^{M_2}(\hat{p}, \hat{v})b_{2j} + \eta_j \tag{A.19}$$

Equation (A.19) is identified up to an additive constant when conditions of Theorem 1 are satisfied. Polynomial approximations for γ_j and μ_j satisfy differentiability condition. And we also need $\frac{\partial Q^{\rho_0}(w_{0i},x_{0i})\hat{\alpha}_0}{\partial w_0} \neq 0$ and $rank(\frac{\partial Q^{Z_1}(x^1,z)\hat{a}_1}{\partial z}) = dim(x^2)$.

Let $b_j = (b_{1j}, b_{2j})$ and $Q^{M_1,M_2}(\mathcal{W}) = Q^{M_1,M_2}(x^1, x^2, w, \hat{p}, \hat{v}) = (Q^{M_1}(x^1, x^2, w), Q^{M_2}(\hat{p}, \hat{v}))$ then b_j may be obtained by OLS as

$$\hat{b}_j = [(Q^{M_1,M_2}(\mathcal{W}))' Q^{M_1,M_2}(\mathcal{W})]^{-1} (Q^{M_1,M_2}(\mathcal{W}))' y_j \tag{A.20}$$

With some large enough M_1, M_2, $Q^{M_1}(x^1, x^2, w)\hat{b}_{1j}$ is an approximation for $\gamma_j(x^1, x^2, w)$. And $\hat{b}_j = (\hat{b}_{1j}, \hat{b}_{2j})$ are consistent with independency of η_j and (x^1, x^2, w) due to

$$\begin{aligned}
&\operatorname*{plim}_{n\to\infty} \hat{b} = \operatorname*{plim}_{n\to\infty}[(Q^{M_1,M_2}(\mathcal{W})' Q^{M_1,M_2}(\mathcal{W})]^{-1} (Q^{M_1,M_2}(\mathcal{W}))' y_j = \\
&= \operatorname*{plim}_{n\to\infty}[\left(Q^{M_1,M_2}(\mathcal{W})\right)' Q^{M_1,M_2}(\mathcal{W})]^{-1} \left(Q^{M_1,M_2}(\mathcal{W})\right)' \left(Q^{M_1,M_2}(\mathcal{W}) b_j + \eta_j\right) = \\
&= b_j + \operatorname*{plim}_{n\to\infty}[(Q^{M_1,M_2}(\mathcal{W}))' Q^{M_1,M_2}(\mathcal{W})]^{-1} (Q^{M_1,M_2}(\mathcal{W}))' \eta_j = b_j
\end{aligned} \tag{A.21}$$

Then the reduced form residuals will be

$$\hat{e}_{ji} = y_{ji} - Q^{M_1,M_2}(x_i^1, x_i^2, w_i, \hat{p}_i, \hat{v}_i)\hat{b}_j \tag{A.22}$$

4. On the fourth step we will estimate the structural equations corrected for sample selection, endogeneity of x^2 and simultaneity in y.

If e_j has joint distribution with ν, e_0 and e_{-j} with density function $f_{e_0,e,\nu}$ then

$$\begin{aligned}
E[e_j | x^1, x^2, w_j, y_{-j}, z, w_{-j}, w_0, d = 1] &= E[e_j | \nu, e_{-j}, g_0(w_0, x_0) + e_0 \geq 0] \\
&= \int_{-\infty}^{\infty} \int_{-g_0(w_0,x_0)}^{\infty} e_j f_{e_0,e,\nu}(s, r | \nu, e_{-j}) ds dr = \varphi(p, \nu, e_{-j})
\end{aligned} \tag{A.23}$$

y_j is decomposed into

$$y_j = g_j(x^1, x^2, w_j, y_{-j}) + \varphi_j(p, \nu, e_{-j}) + \varepsilon_j \tag{A.24}$$

The error term ε_j in this equation will be independent on (x^1, x^2, w_j, y_{-j}).

If $\hat{p}$ is the propensity score, $\hat{v}$ are residuals of endogenous variables equations and $\hat{e}_{-j}$ are reduced form residuals then $\hat{p}$,$\hat{v}$ and $\hat{e}_{-j}$ on this step are fixed. And (x^1, x^2, w_j, y_{-j}) and $(\hat{p}, \hat{v}, \hat{e}_{-j})$ are sets of different variables if $\frac{\partial Q^{\rho_0}(w_{0i}, x_{0i})\hat{\alpha}_0}{\partial w_0} \neq 0$, $rank(\frac{\partial Q^{Z_1}(x^1, z)\hat{a}_1}{\partial z}) = dim(x^2)$ and $\forall j \in \{1, \dots, k\}\ \exists \tilde{w} \in w_j$, $\frac{\partial Q^{M_1}(x^1, x^2, w)\hat{b}_{1j}}{\partial \tilde{w}} \neq 0$.

Every functions $g_j(x^1, x^2, w_j, y_{-j})$ and $\varphi_j(p, v, e_{-j})$ may be approximated by $Q^{\xi_1}(x^1, x^2, w_j, y_{-j})\beta_{1j}$ and $Q^{\xi_2}(\hat{p}, \hat{v}, \hat{e}_{-j})\beta_{2j}$ respectively, where $Q^{\xi_1}(x^1, x^2, w_j, y_{-j})$ is polynomial approximating series with a power ξ_1, $Q^{\xi_2}(\hat{p}, \hat{v}, \hat{e}_{-j})$ is polynomial approximating series with a power ξ_2. Then y_j may be approximated by

$$y_j = Q^{\xi_1}(x^1, x^2, w_j, y_{-j})\beta_{1j} + Q^{\xi_2}(\hat{p}, \hat{v}, \hat{e}_{-j})\beta_{2j} + \varepsilon_j \qquad \text{(A.25)}$$

Equation (A.25) will be identified up to an additive constant if Theorem 1 conditions are satisfied. Polynomial approximations for g_j and φ_j satisfy differentiability condition. And we also need $\frac{\partial Q^{\rho_0}(w_{0i}, x_{0i})\hat{\alpha}_0}{\partial w_0} \neq 0$, $rank(\frac{\partial Q^{Z_1}(x^1, z)\hat{a}_1}{\partial z}) = dim(x^2)$ and $\forall j \in \{1, \dots, k\}\ \exists \tilde{w} \in w_j$, $\frac{\partial Q^{M_1}(x^1, x^2, w)\hat{b}_{1j}}{\partial \tilde{w}} \neq 0$.

Let $\beta_j = (\beta_{1j}, \beta_{2j})$ and $Q^{\xi_1, \xi_2}(\mathcal{X}) = Q^{\xi_1, \xi_2}(x^1, x^2, w_j, y_{-j}, \hat{p}, \hat{v}, \hat{e}_{-j}) = \left(Q^{\xi_1}(x^1, x^2, w_j, y_{-j}), Q^{\xi_2}(\hat{p}, \hat{v}, \hat{e}_{-j})\right)$ then the estimate for β_j may be obtained by OLS as

$$\hat{\beta}_j = [(Q^{\xi_1, \xi_2}(\mathcal{X}))' Q^{\xi_1, \xi_2}(\mathcal{X})]^{-1} (Q^{\xi_1, \xi_2}(\mathcal{X}))' y_j \qquad \text{(A.26)}$$

For some large enough ξ_1, ξ_2, $Q^{\xi_1}(x^1, x^2, w_j, y_{-j})\hat{\beta}_{1j}$ will be an approximation for $g_j(x^1, x^2, w_j, y_{-j})$. Estimate $\hat{\beta}_j = (\hat{\beta}_{1j}, \hat{\beta}_{2j})$ is consistent with independence of ε_j and (x^1, x^2, w_j, y_{-j}) due to

$$\begin{aligned} \operatorname*{plim}_{n\to\infty} \hat{\beta}_j &= \operatorname*{plim}_{n\to\infty} [\left(Q^{\xi_1, \xi_2}(\mathcal{X})\right)' Q^{\xi_1, \xi_2}(\mathcal{X})]^{-1} \left(Q^{\xi_1, \xi_2}(\mathcal{X})\right)' y_j = \\ &= \operatorname*{plim}_{n\to\infty} [\left(Q^{\xi_1, \xi_2}(\mathcal{X})\right)' Q^{\xi_1, \xi_2}(\mathcal{X})]^{-1} \left(Q^{\xi_1, \xi_2}(\mathcal{X})\right)' (Q^{\xi_1, \xi_2}(\mathcal{X})\beta_j + \varepsilon_j) = \end{aligned} \qquad \text{(A.27)}$$

$$= \beta_j + \operatorname*{plim}_{n\to\infty}[\left(Q^{\xi_1,\xi_2}(\mathcal{X})\right)' Q^{\xi_1,\xi_2}(\mathcal{X})]^{-1} \left(Q^{\xi_1,\xi_2}(\mathcal{X})\right)' \varepsilon_j = \beta_j$$

Then the structural form residuals will be

$$\hat{e}_{ji} = y_{ji} - Q^{\xi_1,\xi_2}(\mathcal{X})\hat{\beta}_j \tag{A.28}$$

5. On the last step we will estimate the probability of default equation corrected for sample selection and endogeneity of contract terms using propensity score, residuals from loan limit equation and structural form residuals:

$$E[def|x^1, x^2, y, z, w, d = 1] = g_{def}(x^1, x^2, y) + \varphi_{def}(\hat{p}, \hat{v}, \hat{e}) \tag{A.29}$$

If e_j has joint distribution with v, e_0 and e with density function $f_{e_0,e,v,e_{def}}$ then

$$E[e_{def}|x^1, x^2, y, z, w, d = 1] = E[e_{def}|v, e, g_0(w_0, x_0)+e_0 \geq 0]$$
$$= \int_{-\infty}^{\infty}\int_{-g_0(w_0,x_0)}^{\infty} e_{def} f_{e_0,e,v,e_{def}}(s, r|v, e)dsdr = \varphi_{def}(p, v, e) \tag{A.30}$$

def is decomposed into

$$def = g_{def}(x^1, x^2, y) + \varphi_{def}(p, v, e) + \varepsilon_{def} \tag{A.31}$$

The error term ε_{def} in this equation will be independent on (x^1, x^2, y).

If $\hat{p}$ is the propensity score, $\hat{v}$ are residuals of endogenous variables equations and $\hat{e}$ are structural form residuals then $\hat{p}$,$\hat{v}$ and $\hat{e}$ on this step are fixed. And (x^1, x^2, y) and $(\hat{p}, \hat{v}, \hat{e}\,)$ are sets of different variables if $\frac{\partial Q^{\rho_0}(w_{0i},x_{0i})\hat{\alpha}_0}{\partial w_0} \neq 0$, $rank(\frac{\partial Q^{Z_1}(x^1,z)\hat{a}_1}{\partial z}) = dim(x^2)$ and $\forall j \in \{1, \dots, k\}\ \exists \widetilde{w} \in w_j$, $\frac{\partial Q^{M_1}(x^1,x^2,w)\hat{b}_{1j}}{\partial \widetilde{w}} \neq 0$.

Every functions $g_{def}(x^1, x^2, y)$ and $\varphi_{def}(\hat{p}, \hat{v}, \hat{e})$ may be approximated by $Q^{\theta_1}(x^1, x^2, y)\alpha_1$ and $Q^{\theta_2}(\hat{p}, \hat{v}, \hat{e})\alpha_2$ respectively, where $Q^{\theta_1}(x^1, x^2, y)$ is polynomial approximating series with a power θ_1, $Q^{\theta_2}(\hat{p}, \hat{v}, \hat{e})$ is polynomial approximating series with a power θ_2. Then def may be approximated by

$$def = Q^{\theta_1}(x^1, x^2, y)\alpha_1 + Q^{\theta_2}(\hat{p}, \hat{v}, \hat{e})\alpha_2 + \varepsilon_{def} \tag{A.32}$$

Equation (A.32) will be identified up to an additive constant if Theorem 1 conditions are satisfied. Polynomial approximations for g_{def} and φ_{def} satisfy differentiability condition. And we also need $\frac{\partial Q^{\rho_0}(w_{0i},x_{0i})\hat{\alpha}_0}{\partial w_0} \neq 0$, $rank(\frac{\partial Q^{Z_1}(x^1,z)\hat{a}_1}{\partial z}) = dim(x^2)$ and $\forall j \in \{1,\dots,k\}$ $\exists \tilde{w} \in w_j$, $\frac{\partial Q^{M_1}(x^1,x^2,w)\hat{b}_{1j}}{\partial \tilde{w}} \neq 0$.

Let $\alpha = (\alpha_1, \alpha_2)$ and $Q^{\theta_1,\theta_2}(\mathcal{K}) = Q^{\theta_1,\theta_2}(x^1, x^2, y, \hat{p}, \hat{v}, \hat{e}) = \left(Q^{\theta_1}(x^1, x^2, y), Q^{\theta_2}(\hat{p}, \hat{v}, \hat{e})\right)$ then the estimate for α may be obtained by OLS as

$$\hat{\alpha} = [\left(Q^{\theta_1,\theta_2}(\mathcal{K})\right)' Q^{\theta_1,\theta_2}(\mathcal{K})]^{-1} (Q^{\theta_1,\theta_2}(\mathcal{K}))' def \tag{A.33}$$

For some large enough θ_1, θ_2, $Q^{\theta_1}(x^1, x^2, y)\hat{\alpha}_1$ will be an approximation for $g_{def}(x^1, x^2, y)$. Estimate $\hat{\alpha} = (\hat{\alpha}_1, \hat{\alpha}_2)$ is consistent with independence of ε_{def} and (x^1, x^2, y) due to

$$\begin{aligned}
\operatorname*{plim}_{n\to\infty} \hat{\alpha} &= \operatorname*{plim}_{n\to\infty}[\left(Q^{\theta_1,\theta_2}(\mathcal{K})\right)' Q^{\theta_1,\theta_2}(\mathcal{K})]^{-1} \left(Q^{\theta_1,\theta_2}(\mathcal{K})\right)' def = \\
&= \operatorname*{plim}_{n\to\infty}[\left(Q^{\theta_1,\theta_2}(\mathcal{K})\right)' Q^{\theta_1,\theta_2}(\mathcal{K})]^{-1} \left(Q^{\theta_1,\theta_2}(\mathcal{K})\right)' (Q^{\theta_1,\theta_2}(\mathcal{K})\alpha \\
&\quad + \varepsilon_{def}) = \\
&\alpha + \operatorname*{plim}_{n\to\infty}[\left(Q^{\theta_1,\theta_2}(\mathcal{K})\right)' Q^{\theta_1,\theta_2}(\mathcal{K})]^{-1} \left(Q^{\theta_1,\theta_2}(\mathcal{K})\right)' \varepsilon_{def} = \alpha
\end{aligned} \tag{A.34}$$

Author:

Evgeniy M. Ozhegov

National Research University Higher School of Economics (Perm, Russia). Research group for applied markets and enterprises studies. Young research fellow;

E-mail: tos600@gmail.com, Tel. +7 (952) 652-45-25