KEY DETERMINANTS OF CHOOSING GOVERNMENT-INSURED MORTGAGE IN RUSSIA

Evgeniy M. Ozhegov
Agatha M. Poroshina

National Research University Higher School of Economics
Department of Applied Mathematics and Modeling in Social Systems
Group for Applied Markets and Enterprises Studies

Abstract

The paper presents the structural model of decision-making process on the residential mortgage market. We empirically estimates key drivers of mortgage borrowing, underwriting, and default process by jointly using market-level monthly data and loan-level data from regional branch of Agency of Home Mortgage Lending (AHML). The multistep estimation procedure allows correcting for sample selection bias and endogeneity and provides consistent parameter estimates. Obtained results shows that risk preferences are changing during the time and AHML borrowers are relatively high risky.

Key words: demand, default, mortgage lending, sample selection, endogeneity

JEL classification: C36; D12; R20.

Introduction

The mortgage crisis that started in USA in 2007 and lasted until 2009 was characterized by an unusually large number of defaults on the subprime mortgage market. As a result, it overgrew in to the global economic recession and placed the stability of the world banking system in jeopardy. It caused strong government processes to support mortgage lending and residential housing as a part of all anti-recessionary measures. Such activities include support of citizens with mortgages and the refinancing system of mortgage lending, helping to buy property by citizens, providing living quarters for particular categories of Russian citizens. Key issues of government policy include providing of affordable housing, identifying the main drivers of mortgage borrowing and performance of mortgage loans. Therefore the problem of developing optimal credit contracts and effective risk management systems, especially on the residential mortgage market, is becoming crucial.

National institute for development of housing activity - Agency of Home Mortgage Lending (AHML) helps to implement strong government housing policy and anti-recessionary measures to support mortgage lending in Russia. AHML is state-owned provider of government-insured loans, which uses two-level system of lending. In the first step banks and non-credit organizations provide mortgage loans to households according the common standards of AHML. The second step is refinancing (redemption) of mortgage receivables by AHML. AHML develops special mortgage programs and refinances risks from its regional branches and commercial banks, which operates such programs. The list of programs contains “Young researchers”, “Young teachers”, “Mortgage for Soldiers”, “Mothers’ capital” and other social and subprime programs. All of them have relatively high risk that is insured by government. Considering this the demand for such kind of mortgage programs and behavior of borrowers are generated by some special subsample of potential borrowers that is different from the general population. This research investigates the key drivers of self-selection of borrowers to participate in AHML programs, choosing particular terms of credit contract and loan performance.

This paper has the following structure. It starts with literature review and some generalization of recent studies of mortgage borrowing process. The second part contains the description of collected data and estimation strategy, which allows correcting for sample selection bias and endogeneity. Finally, we discuss the empirical results and conclude with further work.

Literature review

Demand for mortgage loan is the function of probability of credit contract agreement and functions of credit contract terms, on characteristics of borrower, aim of lending, expected loan performance and some macroeconomic variables. However, econometric estimation of parameters of these functions facing with inconsistency driven by endogeneity and sample selection bias.

Endogeneity is generated by simultaneity in borrower and credit organization decisions on explanatory variables in demand and credit risk equations. Sample selection arises when decision-making process of borrowing is made sequentially and some explanatory variables are observed partially in different stages of lending process. These challenges in estimation process were avoided in recent papers that studied lending process.

The following papers focused on the structure of borrowing process and have not got any empirical evidence.

Mortgage borrowing as a sequence of consumer and bank decisions firstly introduced by Follain (1990). He defines the borrowing process as a choice of how much to borrow (the Loan-To-Value ratio, LTV decision), if and when to refinance or default (the termination decision), and the choice of mortgage instrument itself (the contract decision). The main contribution of this study is focusing on possible self-selection of borrowers in borrowing process.

Rachlis and Yezer (1993) then suggested a theoretical model of mortgage lending process, which consists of a system of four simultaneous equations: (1) borrower's application, (2) borrower's selection of mortgage terms, (3) lender's endorsement, and (4) borrower's default. This paper investigates the nature of inconsistency of estimates of recent researches on borrower's discrimination and showed that all of four equations (and decisions) should be considered as interdependent.

From the middle of 90s XX, data publicly were available such as, American mortgage datasets from the Federal Housing Authority (FHA) foreclosure, The Boston Fed Study, The Home Mortgage Disclosure Act (HMDA), and several empirical studies, which analyzed mortgage lending process and studied the interdependency of bank endorsement decision and borrower's decisions modeled by bivariate probit model.

As an extension of study (Rachlis, Yezer, 1993), Yezer, Phillips, Trost, (1994) applied Monte-Carlo experiment to estimate above-listed theoretical model. They empirically shown that isolated modeling processes of the credit underwriting and default leads to the biased parameter estimates. Later on Phillips and Yezer (1996) and Ross (2000) supported these findings.

Phillips and Yezer (1996) compared the estimation results of the single equation approach with those of the bivariate probit model. They showed that discrimination estimation is biased if the lender's rejection decision is decoupled from the borrower's self-selection of loan programs, or if the lender's underwriting decision is decoupled from the borrower's refusal decision.

Ross (2000) studied the link between loan approval and loan default by bivariate probit and found that most of the approval equation parameters have the opposite sign compared with the same from the default equation after correction for the sample selection.

The following earlier papers studied the borrower's choice of mortgage contract terms by probability models.

Shear and Yezer (1983) estimated the linear probability model of choosing FHA insured loans by OLS.

Gabriel and Rosenthal (1991) estimated the probability of choosing between FHA insured loans and conventional loans by Maximum-Likehood estimation (MLE) of parameters of simple probit model.

Canner, Gabriel and Wooley (1991) studied the link between the probability of choosing conventional loan as self-measure of risk and probability of delinquency and showed that minority borrowers are more likely to take a mortgage insured by FHA has less probability of delinquency. This paper deals with two logit equations that estimated independently by MLE.

Coulibaly and Li (2009) using survey data found the evidence that more risk-averse, with risky income and low probability of future move borrowers prefers the fixed rate mortgage contracts than the adjusted rate ones also by estimating logit model.

Leece (2001) investigated the choice of ARM-FRM in the UK market dependent on the expected level of rates. Thus with sustainable low interest rates households intends to lock into fixed rate mortgage. In order to construct consistent and unbiased estimates he used linear additive model with time-dependent explanation variables and generalized linear probit model.

Firestone et al. (2007) analyzed the default and prepayment behavior of low- and moderate-income borrowers. Main finding of this research is that non-white borrowers prepay more slowly than white ones. Results are stable during the time. The data contains the performance of 1.3 million loans originated from 1993 to 1997. To construct consistent estimates they used proportional hazard models for probability of default and prepayment and estimated it by MLE.

Forthowski, LaCour-Little, Rosenblatt and Yao (2011) studied the demand for mortgage loans from the point of choosing of adjusted rate mortgage versus fixed rate mortgage as a function on expected mobility. They find that, with all else equal, who self-select into ARM estimates their probability of moving in the future as relatively high. Choice of ARM-FRM modeled by logit but expected mobility in that equation is endogenous that's why it is predicted by proportional hazard model. The main contribution of this paper is finding that expected mobility and, respectively, choice of mortgage terms are functions on macrovariables.

LaCour-Little (2007) was also focused on the question of choosing the credit program among low- and moderate-income borrowers. Using the loan-level from only one financial organization he founds that LMI borrowers are more likely to choose Federal Housing Administration insured mortgage programs and Special programs that assumed less down payments and higher score of expected risks due to high levels of current debt or weaker credit history. He also finds that nonprime loans preferred for those borrowers who are time limited to provide full documentation. This paper contains multinomial logit (MNL) model for the probability of choosing one of the credit programs. In order to deal with endogeneity at several first steps authors constructed OLS estimated linear models and MLE binary choice models to predict fitted values for endogenous variables in MNL equation.

Previous models that tackled sample selection bias in lending analysis are not appropriate to estimate the loan amount or LTV ratio. The bivariate probit model of Ross (2000) and bivariate probit model used by Yezer, Phillips, and Trost (1994) and Phillips and Yezer (1996) are suitable for estimating a binary outcome. The following papers studied the dependence of the decision on loan amount as well as different endogenous variables on the exogenous ones.

Zhang (2010) investigated the sample selection bias and interaction between pricing and underwriting decisions using classical Heckman model.

Courchane (2007) studied differences in pricing for different ethnicities after controlling of other pricing and underwriting parameters by estimating the Heckman.

Karlan and Zinman (2009) found different method for solve the endogeneity problem when modeling the loan amount equation in microfinance crediting. They generated the truly random sample of credit proposals by sending letters with it to former borrowers. Using the classical Heckman model they estimated the elasticities of demand for consumer credits to maturity and interest rate for different risk types of borrowers.

Attanazio, Goldberg and Kyriazidou (2008) were followed Das et al. (2003) and introduced more progressive approach of managing the sample selection problem when modeling the empirical demand for loan equation. They studied the existence of credit constraints in different income segments. Using loan-level data of car loans they found that low-income households has positive elasticity of demand for car loans on the maturity and zero reaction of demand to interest rate change that means that those households are credit constraint. For doing that they used three-stage estimation methodology. At the first stage they estimated the selection equation. At the second stage the endogenous variables equations are being estimated by semi-parametric regression with correction for self-selection. Then endogenous variables in the demand equation was replaced by fitted values and the parameters were estimated also by semi-parametric regression. The only one motivation of using semiparametric regression is that the error terms of the loan amount, endogenous variables error terms and error term from the participation equation are correlated in non-linear way.

Bocian, Ernest, and Li (2008) used 3SLS for the simultaneous decisions on pricing and credit rating and found the empirical evidence that non-white borrowers are more likely to receive higher-priced subprime credits than similar white borrowers. Ambrose et al. (2004) constructed a simultaneous equation system of LTV and house value, which is used as a proxy for loan amount to account for endogeneity.

As a generalization of recent papers, mortgage-lending process can be represented by following sequence of decisions:

1. *Application of borrower.*

Potential borrower realizes the necessity of borrowing, chooses the credit organization and credit program that reflects her/his preferences, fills an application form with demographic characteristics.

2. *Approval of borrower.*

Considering application form and recent credit history, credit organization endorses the application or not, inquires the form data.

3. *Set the limit loan amount*

When credit organization endorsed a particular borrower, it sets the limit loan amount.

4. *Contract agreement*

The approved borrower makes a choice on contract agreement and when agreed.

5. *Choice of credit terms.*

The approved borrower makes a choice on property to buy and credit terms from feasible set: loan amount not more than limit, down payment, monthly payment and maturity determined by credit program.

6. *Loan performance.*

Borrower chooses the strategy of loan performance: to pay in respect to contract terms or to default, prepay or refinance the loan.

Estimation strategy and data

Econometric model repeats steps of the structural one. The functional form of regression function is taken unspecified following (Das et al., 2003). Particular assumptions on specification form of regression functions and distribution of error terms will be introduced further.

1. Since we do not have micro-level data on who not applied to AHML, the first step of estimation process is modeling the probability of application on aggregated data:

$$y_{1it} = g_{1t}(z_{1t}) + e_{1it} \tag{1}$$
$$e_{1it} \sim IID(0, \sigma^2_{e_{it}})$$
$$z_{1t} = (D_t, M_t),$$

where

$t = 1,..., T, T$ – a set of time moments,

y_{1it} – the probability of application as number of applications in month *t* divided by the amount of households,

D_t – a vector of strictly exogenous aggregated demographics,

M_t – exogenous macrovariables.

2. Modeling the probability of endorsement for all applied applicants:

$$y^*_{2it} = g_2(z^*_{2it}) + e_{2i} \tag{2}$$
$$y_{2it} = y_{1it} y^*_{2it} - \text{is observed}$$
$$\text{cov}(e_{1it}, e_{2i}) = \sigma_{12}, i \in N_t$$

$$\mathrm{cov}(e_{1it}, e_{2i}) = 0, i \notin N_t$$
$$z^*_{2it} = (D^*_i, M_t),$$

where

y^*_{2i} – the probability of endorsement for all applied applicants,

N – a set of individuals, $N = (N_1,..., N_T)$,

N_t – a set of individuals, who applied for a mortgage loan in the time moment *t*,

y_{2it} – an endorsement decision of *i* individual,

z^*_{2it} – a vector of exogenous individual demographics and macrovariables on the date of application.

3. Since loan amount limit is chosen by credit organization it is endogenous and needed to be instrumented (as well as all further endogenous variables) for all endorsed borrowers:

$$\overline{L}^*_{it} = g_{\overline{L}}(z^*_{\overline{L}it}) + e_{\overline{L}i} \tag{3}$$
$$\overline{L}_{it} = y_{2it}\overline{L}^*_{it} \text{ is observed}$$
$$\mathrm{cov}(e_{2i}, e_{\overline{L}i}) = \sigma_{2\overline{L}}$$

where

$\overline{L}$ – a decision on loan limit,

$z^*_{\overline{L}it}$ – a vector of instrumental demographics and macrovariables on the date of application.

4. Modeling the probability of contract agreement:

$$y^*_{3it} = g_3(z^*_{3it}, \hat{L}^*_{it}) + e_{3i} \tag{4}$$
$$y_{3it} = y_{2it}y^*_{3it} \text{ is observed}$$
$$\mathrm{cov}(e_{2i}, e_{3i}) = \sigma_{23}$$
$$z^*_{3it} = (D^*_i, M_t)$$
$$\hat{L}^*_{it} = \hat{g}_{\overline{L}}(z^*_{\overline{L}it})$$

where

$y_{3i} = 1$ an agreement decision,

z^*_{3it} – a vector of individual demographics and macrovariables on the date of application.

$\hat{L}^*_{it}$ – a fitted value of loan amount limit.

5. Simultaneous choice of the credit terms and property for all agreed contracts:

$$\begin{cases} C^*_{jit} = g_{C_j}(z^*_{C_jit}, \hat{L}^*_{it}, C^*_{-jit}, V^*_i) + e_{ji} \\ V^*_{it} = g_V(z^*_{Vit}, \hat{L}^*_{it}, C^*_{it}) + e_{Vi} \end{cases} \tag{5}$$
$$(C_{it}, V_{it}) = y_{3it}(C^*_{it}, V^*_{it}) \text{ is observed}$$
$$\mathrm{cov}(e_{3i}, e_{ji}) = \sigma_{3j}$$
$$\mathrm{cov}(e_{3i}, e_{Vi}) = \sigma_{3V}$$
$$\mathrm{cov}(e_{ji}, e_{ki}) = \sigma_{jk}$$
$$\mathrm{cov}(e_{ji}, e_{Vi}) = \sigma_{jV}$$
$$z^*_{it} = (D^*_i, M_t, F_i)$$

where

V – property value,

$C = (C_j, C_{-j})$ – vector of contract terms (loan amount, maturity, down payment, interest rate, type of rate),

z_{it}^* – a vector of individual demographics D_i^*, macrovariables M_t on the date of application and property characteristics F_i.

6. Modeling the probability of contract events and loss given credit event:

$$y_{4it}^* = g_4(z_{4it}^*, \hat{C}_{it}^*, \hat{V}_{it}^*) + e_{4i} \quad (6)$$

$$y_{4it} = y_{3it} y_{4it}^* \text{ is observed}$$

$$\text{cov}(e_{3i}, e_{4i}) = \sigma_{34}$$

where

y_{4it}^* – the probability of default,

$\hat{C}_{it}^*$ – fitted value of the credit terms,

$\hat{V}_{it}^*$ – fitted property value.

In (Das et al., 2003) it was showed that with some light assumptions the simple sample selection model with endogenous repressors like

$$y_i^* = g_0(z_{1i}^*, x_i) + e_i \quad (7)$$

$$x_i^* = \pi(z_{1i}^*, z_{2i}^*) + v_i$$

$$y_i = dy_i^* \text{ is observed}$$

$$x_i = dx_i^* \text{ is observed}$$

$$\text{cov}(e_i, d) \neq 0$$

$$\text{cov}(v_i, d) \neq 0$$

can be estimated up to additive constant by following three-step procedure:

1. Consistent estimation of $\hat{p} = E[d \mid z_d]$;
2. Estimation of $\hat{x} = E[x \mid z, d = 1] = \pi(z_1, z_2) + \mu_0(\hat{p})$;
3. Estimation of $E[y \mid x, z, d = 1] = g_0(z, \hat{x}) + \lambda_0(\hat{p})$.

The paper (Ozhegov, 2013) contains an extension of this model and Newey, Powell, Vella (1999) for the case of simultaneous equations with sample selection (5). It reflects this method to non-triangular system of simultaneous equations with sample selection and adds one estimation step. For the model like

$$\begin{cases} y_{1i}^* = g_1(z_{1i}^*, y_{-1,i}^*, x_i) + e_{1i} \\ \ldots \\ y_{Ki}^* = g_K(z_{Ki}^*, y_{-K,i}^*, x_i) + e_{1i} \end{cases} \quad (8)$$

$$y^* = (y_1^*, \ldots, y_K^*) = (y_j^*, y_{-j}^*)$$

$$z^* = (z_1^*, \ldots, z_K^*)$$

$$x_i^* = \pi(z_i^*, z_{xi}^*) + v_i$$

$$y_{ji} = dy_{ji}^* \text{ is observed}, \ j = 1, \ldots, K$$

$$x_i = dx_i^* \text{ is observed}$$

$$\text{cov}(e_{ji}, e_{ki}) = \sigma_{jk}, \ j, k \in \{1, \ldots, K\}$$

$$\text{cov}(e_{ji}, d) = \sigma_{jd}, \ j \in \{1, \ldots, K\}$$

$$\text{cov}(v_i, d) = \sigma_{vd}$$

the estimation procedure will now be:

1. Consistent estimation of $\hat{p} = E[d \mid z_d]$;
2. Estimation of $\hat{x} = E[x \mid z, z_x, d = 1] = \pi(z, z_x) + \mu_0(\hat{p})$;

3. Estimation of $\hat{y}_j = E[y_j \mid z, x, d = 1] = g'_j(z, \hat{x}) + \lambda_j(\hat{p})$;
4. Estimation of $E[y_j \mid z, y_{-j}, x, d = 1] = g_j(z, \hat{y}_{-j}, \hat{x}) + \lambda_j(\hat{p})$.

This procedure provides an identification of g up to K additive constants when μ_0, λ, π, g are continuously differentiable with continuous distribution functions almost everywhere and with probability one $\frac{\partial \hat{p}(z_d)}{\partial z_d} \neq 0$ and $rank[\frac{\partial \hat{x}}{\partial z_x}] = \dim(x)$.

Estimation of model (1–6) considers following assumptions. Error terms in (1–6) have jointly normal distribution with zero-vector of first moments. Matrix of second moments has each diagonal element equal to one (because of identifiability of the model up to the set of constants, it need to be specified) and non-zero covariates between error terms that is being estimated. This assumption implies using Heckman's lambda (or Inverse Mills ratio) as μ_0 and λ functions in (7–8). Combined with first degree polynomial approximation functions for π and g it satisfies the first condition of identifiability of model. It was shown in Attanasio et al. (2008) that the estimates in demand-for-credit equation with higher degree polynomial approximation functions are no less consistent but less efficient. Evidence of this statement for the data set of this research for all borrowing process stages will be provided in further researches.

The data collected for this research contains two sets. The first data set is aggregated regional monthly data on the AHML branch performance, mortgage market characteristics and regional macroeconomic variables for the period from 01/08/2008 to 31/08/2012. This data set is publicly available.

The second set includes the loan-level data from one regional AHML branch on 4300 applications for mortgage loans. This data set contains information on socio-demographic characteristics of each particular applicant, the date of application, the flag of credit organization's approval decision, the flag of contract agreement, the credit terms agreed, property characteristics, which was bought, the flag of default, the date of default. Socio-demographic characteristics are fixed on the date of application.

Initially the second data set included 4897 observations. However we cleaned data and excluded outliers. Borrowers whom age was not specified or fewer 21 years old, mortgages with negative down payment or/and null monthly payment or/and contract rate, and observations with LTV exceeds 1 or close to 0 and DTI equals 0 or exceeds 1 were dropped. The variables in models are defined in Table 1-2. Tables 1-2 contain variable description, descriptive statistics: sample means and standard deviations after dropping outliers.

Specifically, data set of 4300 individuals includes both approved and denied ones in the proportion 86:14. However only 2801 borrowers (76,6 % from total number of approved applicants) have mortgages. 5% of approved loans were defaulted (90 days or more delinquent). The problem of data disproportion is typical in the credit risk modeling. According to Maddala (1992), in the estimation of binary choice model or even linear probability model it influences only estimated intercept, but not other estimated parameters.

The terms of credit contract practically are used as proxy variables to estimate the risk of a particular borrower. For example, mortgages with low loan-to-value ratio (LTV) are attractive for non-liquid borrowers. The probability that they could face with serious problem of repayment of a loan is much higher. Moreover, borrowers with LTV higher 90%, think as holders, because they do not invest a lot of own capital and have less motivated to overcome obstacles with repayment of a loan. For this reason mortgages with high LTV are riskier and lenders offer higher interest rates for these mortgage products. The data in Table 1 shows that sample contains borrowers with different LTV, but an average these are not high risky borrowers because sample mean LTV equals 56 %, which is much less 90 % and sample assessed property value approximately 2 million Russian rubles, which is common for secondary real market.

Typically mortgages have two types of interest rates – adjustable (ARM) and fixed (FRM). Adjusted-rate mortgages are riskier, and practically the level of such interest rate depends on one of the stock index. Only 13,5% mortgages in sample are that ones. Approximately 39,54% of observations are 15-year and 20-year mortgages.

From socio-demographic characteristics is an income of a particular borrower plays significant role to predict the probabilities of application, approval, contract agreement, and default, because it directly influences on the ability to repay a mortgage. Noticeably among 4300 individuals, approximately 68 % do not have information about income. 8,74 % and 13,93 % people have monthly income 10 000-19999 Russian rubles, and 20000-39999 Russian rubles correspondingly. An average 45% of monthly income spends to repay mortgage payments. In the Table 1 it shows the debt-to-income ratio (DTI), which has larger effect on borrowers with low credit quality.

The level of education could be regarded as a proxy for the level of financial literacy of particular borrower, which could influence the probability of default too. Most of borrowers in the sample have higher education (52,36%) and secondary one (42,69%). 95,27% of total sample are middle-aged hired employees.

Macrovariables characterizes the market demand and supply characteristics. Sample mean housing to price ratio reaches 3,48 years that means during this period household can able to save money from current income for buying property. In Europe, this index ranges from 3 to 6 years. However, there is a significant difference between Russia and the developed countries in terms of the conditions to save money for buying property and access to credit resources (Kosareva, 2006).

TABLE 1
Summary statistics

Variables	Description	Mean	Std. Dev.	Min	Max
Flag of endorsement	=1 if loan approved	-	-	-	-
Flag of contract agreement	=1 if client agreed to have mortgage	-	-	-	-
Flag of default	=1 if borrower defaults on an approved loan (delinquent payments more than 90 days)	-	-	-	-
Age of borrower	Age of borrower, years	34	7.6	21	61
Age squared	Age of borrower squared, years	-	-	-	-
Male	Sex, =1 male	-	-	-	-
Family status	Family status, 1 - single; 2 - married; 3 - widowed; 4 - divorced	-	-	-	-
Activity category	Type of work, 1 - unemployed; 2 - retiree[1]; 3 - soldier; 4 - hired employee; 5 - entrepreneur; 6 - state employee	-	-	-	-
Education level	Education level, 1 - elementary education; 2 - secondary education; 3 - incomplete higher education; 4 - higher education	-	-	-	-
Income category	Monthly income of borrower (in Russian rubles), 1 - no data on income; 2 - income 0-9999; 3 - income 10000-19999; 4 - income 20000-39999; 5 - income >=40000	-	-	-	-
Sum of co-borrowers main income	Sum of co-borrowers main income (in Russian rubles), 0 - no data co-borrower's income; 1 - co-borrower's income 10000-19999; 3 - co-borrower's income >=20000	-	-	-	-
# of co-borrowers	Number of co-borrowers	0.62	0.57	0	3
Loan limit	Maximum loan limit, Russian rubles	936059,4	684952	0	12700000
Rate	Contract rate, %	11.58	1.62	9.55	19
Type of rate	Type of contract rate, 0 - fixed rate, 1 - adjusted rate	-	-	-	-
Loan amount	Loan amount, Russian rubles	1039966	573503.1	120000	10000000
Maturity	Maturity of credit, 1 - maturity < 120 months; 2 – maturity 120-179 months; 3 - maturity 180-239 months; 4 - maturity 240-299; months; 5 - maturity >=300 months	-	-	-	-
Down payment	Down payment, Russian rubles	854494.6	706638.9	0	13800000
LTV	Loan-to-value ratio	0.56	0.17	0.02	0.94
DTI	Debt-to-income ratio	0.45	0.18	0.06	1
Flat value	Assessed value, Russian rubles	1894460	1049331	330000	15300000
Days of observation	Total amount of days observed in credit, days	786.65	430.77	15	1487
Unemployment rate	Quarterly regional unemployment, %	8.43	1.51	6.3	10.9
Mean loan	Average size of mortgage in region, Russian rubles	1160.27	252.23	899.31	1908.2
Median maturity	Median maturity for mortgage in region, Russian rubles	201,64	12.7	173	222.2
Median rate	Median contract rate for mortgage in region, %	13.1	0.82	12	14,3
Mean DTI	Average DTI in region	34.81	0.7	33.44	36.68
Mean m2 value	Average price for 1 square meters in region, Russian rubles	37617.03	6395.77	28782	51304
Lodging coefficient in years	Housing price to income ratio, years	3.48	0.68	2.57	4.65
Mortgage volume	Total amount of mortgages in region, millions Russian rubles	885948.4	563161.6	116100	2191000
Mortgage amount	Total amount of mortgages in the region	896.57	528.89	134	2112

[1] There is 1 retiree individual in the sample. After cleaning data that observation was dropped.

TABLE 2
Summary of categorical variables[2]

Variables	Total	%
Male		
male	1881	43.74
female	2419	56.26
	4300	100
Family status		
single	1221	28.70
married	2359	55.45
widowed	56	1.32
divorced	618	14.53
	4254	100
Activity category		
unemployed	1	0.02
retiree	0	0.00
soldier	13	0.31
hired employee	3965	95.27
entrepreneur	39	0,94
state employee	144	3.46
	4162	100
Education level		
elementary education	65	1.59
secondary education	1748	42.69
incomplete higher education	138	3.37
higher education	2144	52.36
	4095	100
Income category		
no data on income	2918	67.86
0-9999	118	2.74
10000-19999	376	8.74
20000-39999	599	13.93
>=40000	289	6.72
	4300	100
Sum of co-borrowers main income		
no data co-borrower's income	3725	86.63
co-borrower's income 0-9999	160	3.72
co-borrower's income 10000-19999	225	5.23
co-borrower's income >=20000	190	4.42
	4300	100
Type of rate		
fixed rate	2423	86.50
adjusted rate	378	13.50
	2801	100
Maturity		
< 120 months	181	6.46
120-179 months	595	21.25
180-239 months	1107	39.54
240-299 months	690	24.64
>=300 months	227	8.11
	2800	100

[2] Such variables as the family status, the activity category, and the education level have missing data. Percentages are calculated as percent from total available data. The type of rate and the maturity are available only for issued mortgages.

Results

The parameters of model (1–6) was estimated with linear *g*-functions for continuous outcomes, probit *g*-functions for discrete outcomes and sample selection bias correction term in form of Heckman's lambda function. Consistent estimation of model (1–6) is presented in Table 3–8. These estimates were compared with those that were not corrected for sample selection bias or endogeneity and both. Standard errors of parameters were estimated in robust (controlling for hidden heterogeneity) and bootstrap (controlling for correlation between households who take a mortgage in the same month) way with 100 repetitions. It compared with simple standard errors estimates and parameters remain significance in all specifications. Then the bootstrap standard errors are reported.

Application of borrower

TABLE 3

Estimated parameters for probability of application equation (eq.1)

	(1) OLS
Mean loan	–0.019***
	(0.000)
Median maturity	–1.561***
	(0.004)
Median rate	9.033***
	(0.047)
Mean DTI	–22.608***
	(0.046)
Mean m2 value/1000	0.164***
	(0.005)
Lodging coefficient in years	20.152***
	(0.067)
Constant	976.993***
	(2.143)
Observations	4284
R^2	0.481
Adjusted R^2	0.481

Note:
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The second column in Table 3 contains OLS estimates of linear probability model for the probability of household to apply to AHML for mortgage (eq.1). The market demand and supply determinants and aggregated demographic characteristics were used as explanatory variables. The set of demographics is coming out to be insignificant due to low variation between the time moments (more information can be mined from regional-monthly panel structured data from several regional branches). The exception is the lodging coefficient which is computed as the number of years that needed to work with mean nominal earnings to buy median flat. This variable receives the information on income and price variation. Increasing of this coefficient (the affordability of lodging) causes increasing the probability of application for mortgage. The increasing of this coefficient is a main proxy for negative demand shock. Mortgage market negative supply shocks (increasing of median interest rate and decreasing of median maturity as proxies of mortgage market collapsing) also increases the probability of application to AHML. It can be easily explained by AHML's social functions of providing accessible housing even when mortgage and real estate markets drop.

Approval of borrower

TABLE 4

Estimated parameters for probability of endorsement (eq.2)

	(1) Probit corrected for sample selection	(2) Probit
Fitted probability of application	–0.005**	
	(0.002)	
Age of borrower	–0.041	–0.037
	(0.031)	(0.031)
Age squared	0.001	0.001
	(0.000)	(0.000)
Male	0.124**	0.119**
	(0.061)	(0.061)
Single	0.041	0.036
	(0.068)	(0.068)
Widowed	–0.486**	–0.486**
	(0.216)	(0.216)
Divorced	–0.078	–0.076
	(0.084)	(0.084)
Entrepreneur	0.534	0.517
	(0.461)	(0.459)
State employee	0.565***	0.570***
	(0.194)	(0.194)
Elementary education	0.074	0.104
	(0.246)	(0.245)
Secondary education	–0.055	–0.032
	(0.143)	(0.142)
Complete higher education	0.339**	0.350**
	(0.143)	(0.142)
No data on income	–1.529***	–1.520***
	(0.380)	(0.378)
Income 10000-19999	-0.117	–0.137
	(0.425)	(0.423)
Income 20000-39999	–0.507	–0.518
	(0.396)	(0.394)
Income >=40000	–0.648	–0.656
	(0.412)	(0.410)
Constant	3.006***	2.690***
	(0.680)	(0.664)
Observations	3987	3987
AIC	2716.3	2719.9
BIC	2829.6	2826.8
Log-likelihood	–1340.2	–1342.9
% of right predictions	87.5	87.5

Note:
Family status “Married” was taken as base outcome
Activity category “Hired employee” was taken as base outcome
Education level “Incomplete higher” was taken as base outcome
Income level 0-9999 was taken as base outcome

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Underwriting decision is, first of all, determined by clear recent credit history and providing of correct documentation (that is unobserved in collected data, gaining low explaining ability of other borrower characteristics). The second driver of underwriting decision is the number of applications that negatively affects the probability of borrower’s approval. It is supported by the negative significant correlation (–0.005) between fitted probability of application and the probability of endorsement. This is due to the fact that regional AHML branches have the limits of funds provided by AHML in each period and branches must take into account and regulate the number of applications by controlling the rigidity of underwriting. Controlling for probability of application is necessary to correct the sample for sample selection bias caused by different rigidity of underwriting in different months and provides some additional information of variation of probability of application endorsement. Relatively high percentage of right predictions (87.5%) appeared in this model. All coefficients in both specifications have same signs and approximately same values.

The coefficient on gender is statistically significant. This finding provides evidence of gender discrimination in approval process. However, problems of discrimination on the Russian mortgage market are not developed yet. In addition, the presence of complete higher education and work in state-owned organization increase the likelihood of approval. All other coefficients are intuitive. There is a point about the preference of relatively high endorsement of low-income applicants but it evidenced in next part of research that low-income borrowers have relatively small credit risk (See comments on Table 6).

Set the limit loan amount

Credit limit or maximum affordable loan amount for borrower is determined firstly by the characteristics of borrowers such as age (mid-aged preferred), sex of main borrower, family status of main borrower (married strictly preferred to single and then to widowed and divorced), education level (complete high and higher education preferred), type of work (entrepreneurs preferred to hired employees and state employees) and positively correlated with level of income and co-corrower's income level. Moreover, the less value of credit limit affected by the relatively high probability of approval due to the reasons of funds controlling described above.

Contract agreement

TABLE 6
Estimated parameters for probability of contract agreement (eq.4)

	(1) Probit with correction for endogeneity and sample selection	(2) Probit with correction for endogeneity	(3) Probit with correction for sample selection	(4) Probit
Probability of application	0.236***		0.221***	
	(0.049)		(0.048)	
Probability of endorsement	0.377*		0.218	
	(0.202)		(0.182)	
Loan limit $x10^{-4}$	0.003*	0.000	0.000	0.000
	(0.001)	(0.001)	(0.000)	(0.000)
Mean loan	0.004***	0.001***	0.004***	0.001***
	(0.001)	(0.000)	(0.001)	(0.000)
Median maturity	0.402***	0.030***	0.377***	0.030***
	(0.078)	(0.005)	(0.076)	(0.005)
Median rate	–2.101***	–0.200**	–2.019***	–0.202***
	(0.412)	(0.079)	(0.409)	(0.077)
Mean LTV	0.017	0.012	0.015	0.011
	(0.017)	(0.017)	(0.017)	(0.017)
Mean DTI	5.453***	0.175**	5.117***	0.174**
	(1.099)	(0.073)	(1.081)	(0.073)
Lodging coefficient in years	–4.439***	0.343***	–4.105***	0.346***
	(0.988)	(0.090)	(0.969)	(0.088)
Single	–0.231***	–0.264***	–0.262***	-0.266***
	(0.059)	(0.058)	(0.056)	(0.056)
Widowed	–0.031	–0.116	–0.116	–0.122
	(0.231)	(0.227)	(0.226)	(0.224)
Divorced	–0.194***	–0.241***	–0.230***	–0.244***
	(0.075)	(0.073)	(0.072)	(0.071)
Entrepreneur	–0.432	–0.197	–0.285	–0.188
	(0.310)	(0.295)	(0.297)	(0.290)
State employee	0.375***	0.389***	0.357***	0.385***
	(0.132)	(0.131)	(0.132)	(0.130)
Elementary education	0.751***	0.714***	0.654***	0.705***
	(0.249)	(0.249)	(0.243)	(0.242)
Secondary education	0.360**	0.317**	0.281**	0.311**
	(0.145)	(0.145)	(0.138)	(0.138)
Complete higher education	0.281**	0.310**	0.272**	0.309**
	(0.136)	(0.136)	(0.136)	(0.135)
Constant	–245.250***	–11.424***	–229.221***	–11.296***
	(48.545)	(3.156)	(47.651)	(3.065)
Observations	3487	3487	3487	3487
Pseudo R^2	0.096	0.089	0.095	0.089
AIC	3501.2	3523.4	3504.4	3523.3
BIC	3612.1	3621.9	3615.3	3621.8
Log-likelihood	–1732.6	–1745.7	–1734.2	–1745.7
% of right predictions	76.1	76.2	76.0	76.2

Note: Family status “Married” was taken as base outcome
Activity category “Hired employee” was taken as base outcome
Education level “Incomplete higher education” was taken as base outcome

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

In different specifications the estimates remain statistical significance and are generally consistent. Sample selection bias terms from selection equations is need to be included due to its significance (significant correlation between error terms in application, endorsement and contract agreement decisions). Correction for endogeneity of loan limit gains significance of this variable for contract agreement decision and that is so only with sample selection correction.

Relatively high probability of contract agreement is observed for borrowers who are tend to be settled(married, work in state-owned organization) and has positive expectations on the terms of credit contract (high mean DTI, loan amount and maturity, low rate and gained higher credit limit). Borrowers with elementary, secondary, or complete higher education are more likely to contract compared with borrowers with incomplete higher education. This equation has relatively small goodness-of-fit reasoned by strong dependency on such unobservables as alternative offers, quality of service in AHML and presence of suited property.

Choice of credit terms

Demand for mortgage loan or desired loan amount less determined by characteristics of borrower (and from underwriting and contract agreement decisions that firstly determined by demographics) but on desired flat (more expensive flat, larger loan), macrovariables (determines the probability of applications) and (fitted) contract terms such as type of rate (larger loans with fixed rate, smaller loans with adjusted rates), down payment (more down payment needed, less loan) and loan limit (with positive dependency) and does not affected by mortgage rate and maturity. It is needed to be pointed out that correction for endogeneity strongly corrects the coefficients suffered from inconsistency due to simultaneity bias.

Loan performance

TABLE 8

Estimated parameters for probability of default (eq.6)

	(1) Probit with correction for endogeneity and sample selection	(2) Probit with correction for endogeneity	(3) Probit with correction for sample selection	(4) Probit
Probability of application	0.061***		–0.016***	
	(0.021)		(0.006)	
Probability of endorsement	1.580*		–0.224	
	(0.952)		(0.710)	
Probability of contract agreement	0.512		–0.026	
	(0.643)		(0.418)	
Rate is adjusted	–11.584	4.473	0.322***	0.331***
	(20.484)	(3.204)	(0.043)	(0.041)
Rate	1.659***	0.305	0.563	0.620*
	(0.371)	(0.523)	(0.379)	(0.376)
Maturity <120 months	8.313	–19.045**	0.563	0.620*
	(8.659)	(8.039)	(0.379)	(0.376)
Maturity 120-179 months	15.305**	–20.623**	0.481	0.527*
	(7.715)	(8.236)	(0.317)	(0.317)
Maturity 180-239 months	14.134*	–19.633***	0.359	0.394
	(7.263)	(7.424)	(0.304)	(0.305)
Maturity 240-299 months	29.289***	–12.513***	0.142	0.194
	(10.947)	(3.695)	(0.323)	(0.323)
Flat value $x10^{-4}$	–0.091***	–0.018	–0.001	–0.001
	(0.032)	(0.012)	(0.001)	(0.001)
Loan amount $x10^{-4}$	0.175***	–0.002	0.001	0.001
	(0.065)	(0.005)	(0.002)	(0.002)
LTV	–36.064**	–0.003	–0.077	–0.098
	(14.351)	(0.002)	(0.451)	(0.443)
Age of borrower	–0.742**	0.441**	0.055	0.060
	(0.367)	(0.212)	(0.070)	(0.067)
Age squared	0.011*	–0.005*	–0.001	–0.001
	(0.006)	(0.003)	(0.001)	(0.001)
Male	0.255**	0.438***	0.287**	0.269**
	(0.124)	(0.160)	(0.124)	(0.120)
Entrepreneur	0.775	0.353	0.696	0.537
	(0.519)	(0.452)	(0.488)	(0.450)
State employee	–0.349***	–0.334***	–0.350***	–0.345***
	(0.112)	(0.109)	(0.118)	(0.110)
# of coborrowers	0.391	-2.001***	–0.826**	–0.760***
	(0.670)	(0.483)	(0.398)	(0.268)
No data on income	–0.427*	–1.170***	–0.377	–0.406*
	(0.230)	(0.330)	(0.232)	(0.228)
Income 10000-19999	–0.080	–1.056***	–0.661**	–0.658***
	(0.397)	(0.295)	(0.275)	(0.243)
Income 20000-39999	–0.130	–0.414	–0.446	–0.389
	(0.659)	(0.478)	(0.357)	(0.306)
Income >=40000	–0.054**	0.027	0.000	0.015
	(0.022)	(0.020)	(0.012)	(0.009)
Mean m2 value	0.001	0.003***	0.002***	0.001***
	(0.001)	(0.001)	(0.001)	(0.000)
Constant	–9.455	3.662	–7.622***	–8.768***
	(8.161)	(4.478)	(1.626)	(1.422)

Observations	2229	2229	2229	2229
Pseudo R^2	0.377	0.341	0.419	0.411
AIC	695.8	727.9	651.1	653.2
BIC	832.8	847.8	782.4	767.4
Log-likelihood	–323.9	–342.9	–302.5	–306.6
% of right predictions	93.8	93.8	94.0	94.3

Note: Activity category “Hired employee” was taken as base outcome
Income level 0-9999 was taken as base outcome
Maturity >=300 months was taken as base outcome

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

One of the most unfavorable credit events is default. The probability of default is used as a measure of credit risk. The second column in Table 8 reports the results of probit model with correction endogeneity and sample selection for the probability of borrower’s default (eq. 6). High predictive power of the model is supported by high percentage of right predictions, which is close to 94%.

Credit risk increases with increasing in (fitted) mortgage rate, loan amount, deceasing of flat value, linked with initial low and high (but not moderate) income level and negatively with number of co-borrowers. The last two facts should be explained. Mortgage programs suppose debt-to-income ratio not more than the upper bound determined by particular program. In order to obtain it low-income borrowers pick co-borrowers which may take some risk in case of delinquency. Additional explanation of high payment discipline of low-income borrowers is that a mortgage is the only chance to obtain housing and bankruptcy will cause deprivation of property. Parabolic with branches up dependency of credit risk on age may be explained by higher moral cost for young-aged (deterioration of credit history) and old-aged (soviet stark discipline mentality) borrowers when default.

Relatively high level of credit risk of AHML borrowers is empirically evidenced by significant positive correlation of error terms in probability of default equation and equation of probability of application to AHML. The reason is that AHML orients to achieve social goals and providing affordable housing even in negative mortgage market demand and supply shocks. Negative shocks determine high probability of application to AHML which is positively correlated with risk of delinquency.

Low significance of probability of endorsement and its’ positive correlation with credit risk error terms are giving evidence that endorsement process is not aimed to bring light to potential unfair borrowers but takes into account rather different than risk factors. Since all risk refinanced by government, AHML borrowers are high risk household which can not take affordable mortgage loan in commercial banks.

Conclusion

To summarize our results, we estimated the model of borrowing process on the stages of application, underwriting, contract agreement and loan performance. The estimation strategy relies on several assumptions on joint distribution of error terms and approximation functions for regression equations. However, corrected for sample selection and simultaneity biases estimates appear to be consistent.

Results are, of course, conditional on data and estimation strategy. Based on this estimates we find that (1) probability of application for mortgage to AHML increases with negative mortgage market shocks; (2) underwriting process rely on characteristics of borrower and amount of applications in particular period; (3) probability of contract agreement determined by loan limit, expectations of credit terms and stability of demographics; (4) demand for mortgage is a function of loan limit and characteristics of desired flat and less determined by contract characteristics and demographics; (5) credit risk is higher with higher rate, for larger loans, moderate-income and middle-aged borrowers; (6) AHML borrowers is relatively more risky than the general sample.

The collected data set suffers from lack of credit history data, reasons of disagreement on contract, particular rival offers on mortgage market, quality of service of AHML and other credit organizations, low variation in aggregate demographic characteristics. Further research should attempt to avoid these challenges. More flexible econometric techniques like semiparametric and nonparametric estimation should apply. Cross-validation allows conducting robustness check of models.

Acknowledgements

The authors would like to thank Anil K. Bera for his helpful comments. This study was carried out with support from "The National Research University Higher School of Economics Academic Fund Program in 2013-2014, Research Grant No. 12-01-0130 ". The authors are responsible for any errors that remain.

References

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), pp. 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), pp. 401-436.

Bocian, D., Ernst, K., Li, W. (2008). Race, Ethnicity and Subprime Home Loan Pricing. Journal of Economics and Business, 60, pp. 110–124.

Campbell, J. Y., Cocco, J. (2011). A Model of Mortgage Default. NBER Working Paper Series.

Canner, G.B., Gabriel, S.A. Wooley, G.M. (1991). Race, Default Risk, and Mortgage Lending: A Study of the FHA and Conventional Loan Markets. Southern Economics Journal, 58(1), pp. 249-262.

Clapp, M. (2006). Unobserved Heterogeneity in Models of Competing Mortgage Termination Risks. Real Estate Economics, 34(2), pp. 243-273.

Coulibaly, B., Li, G. (2009). Choice of Mortgage Contracts: Evidence from the Survey of Consumer Finance. Real Estate Economics, 37(4), pp. 659-673.

Courchane, M. (2007). The Pricing of Home Mortgage Loans to Minority Borrowers: How Much of the APR Differential Can We Explain? Journal of Real Estate Research, 29, pp. 399-439.

Gabriel, S. A., Rosenthal, S. S. (1991). Credit rationing, race, and the mortgage market. Journal of Urban Economics, 29(3), pp. 371-379.

Das, M., Newey, W. K., Vella, F. (2003). Nonparametric estimation of sample selection models. The Review of Economic Studies, 70(1), pp. 33-58.

DeMarzo, P. M., Sannikov, Y. (2006). A Continuous-Time Agency Model of Optimal Contracting and Capital Structure. Journal of Finance, 61, pp.2681–2724.

Firestone, S., Van Order, R., Zorn, P. (2007). The Performance of Low-Income and Minority Mortgages. Real Estate Economics, 35(4), pp. 479-504.

Follain, J. R. (1990). Mortgage choice. AREUEA Journal, 18(2), pp. 125–144.

Forthowski, E., LaCour-Little, M., Rosenblatt, E., Yao, V. (2011). Housing Tenure and Mortgage Choice. Journal of Real Estate Finance and Economics, 42, pp. 162-180.

Ghent, A., (2011). Subprime Mortgages, Mortgage Choice, and Hyperbolic Discounting, Working Paper, Zicklin School of Business.

Imbens, G. W., Newey, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. Econometrica, 77(5), pp. 1481-1512.

Karlan, D., Zinman J. (2009). Observing Unobservables: Identifying Information Assymetries with a Consumer Credit Field Experiment. Econometrica, 77(6), pp.1993-2008.

Kosareva NB (2006). Osnovi ipotechnogo credtovanya [Basics of Mortgage Lending]. Moscow, Found «Institute for Urban Economics».

LaCour-Little, M. (2007). The Home Purchase Mortgage Preferences of Low- and Moderate-Income Households, Real Estate Economics, 35(3), pp. 265–290.

Leece, D. (2001). Regressive Interest Rate Expectations and Mortgage Instrument Choice in the United Kingdom Housing Market, Real Estate Economics, 29(4), pp. 589-613.

Leece, D. (2006). Testing a Theoretical Model of Mortgage demand on UK Data. Applied Economics, 38, pp. 2037-2051.

Maddala, G.S. (1992). Introduction to Econometrics. 2nd ed., Macmillan.

Maddala, G., Trost R. (1982). On Measuring Discrimination in Loan Markets. Housing Finance Review, 1, pp. 245–268.

Newey, W. K., Powell, J. L., Vella, F. (1999). Nonparametric estimation of triangular simultaneous equations models. Econometrica, 67(3), pp. 565-603.

Nichols, J., Pennington-Cross, A., Yezer, A. (2005). Borrower Self-selection, Underwriting Costs, and a Subprime Mortgage Credit Supply, Journal of Real Estate Finance and Economics, 30(2), pp. 197-202.

Ozhegov, E.M. 2013. Nonparametric estimation of simultaneous equations models with sample selection (Working paper, Higher School of Economics, Group for Applied Markets and Enterprises Studies).

Piskorski, T., Tchistyi, A. (2010). Optimal Mortgage Design. Review of Financial Studies, 23, pp. 3098—3140.

Piskorski, T., Tchistyi, A. (2011). Stochastic House Appreciation and Optimal Subprime Lending. Review of Financial Studies, 24, pp. 1407—1446.

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, pp. 87–102.

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

Ross, S. L. (2000). Mortgage Lending, Sample Selection and Default. Real Estate Economics, 28(4), pp. 581-621.

Shear, W. B., Yezer, A. M. (1982). An indirect test for differential treatment of borrowers in mortgage markets. *Real Estate Economics*, *10*(4), pp. 405-420.

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, pp. 197-215.

Zhang, Y. (2010). Fair Lending Analysis of Mortgage Pricing: Does Underwriting Matter?, Office of the Comptroller of the Currency, OCC Economics Working Paper 2010)