NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Agatha M. Poroshina

CREDIT RISK MODELING OF RESIDENTIAL MORTGAGE LENDING IN RUSSIA

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: Financial Economics
WP BRP 30/FE/2014

SERIES: *FINANCIAL ECONOMICS*

Agatha M. Poroshina[1]

CREDIT RISK MODELING OF RESIDENTIAL MORTGAGE LENDING IN RUSSIA[2]

This paper analyzes the problems of credit risk modeling of residential mortgage lending in Russia. Using unique mortgage loan and macro data from a regional branch of the Agency of Home Mortgage Lending (2008-2012), we find that borrower and mortgage loan characteristics affect the loan performance and play an important role in predicting default as well as a macroeconomic situation. On the residential mortgage market, borrowers with undeclared income have the lowest probability of default, mainly explained by the difference in declared and real income. Obtained results are robust under parametric and semiparametric specifications with correction for selectivity bias.

JEL Classification: C14, D12, R20

Keywords: credit risk, default, mortgage lending, sample selection, Russia

[1] National Research University Higher School of Economics. Department of Applied Mathematics and Modeling in Social Systems, Group for Applied Markets and Enterprises Studies. Lecturer, Junior Research Assistant. E-mail: aporoshina@hse.ru. Perm, Russia.

[2] This study was carried out within "The National Research University Higher School of Economics' Academic Fund Program in 2013-2014, research grant No. 12-01-0130". The author is responsible for any errors that remain. The author would like to thank Anil K. Bera (Univeristy of Illinois), Andreas A. Woudenberg (Inholland Univeristy), Alexander M. Karminsky (Higher School of Economics), and Evgeniy M. Ozhegov (Higher School of Economics) for their helpful comments.

Introduction

During the years 2005-2008, mortgage for residential lending was one of the fastest-growing segments of the Russian credit market. Different factors contributed to that substantial growth. Especially, several federal acts were passed and a safe regulatory environment was created for the development of mortgage lending. As a result of these strong governmental processes, the volume of residential mortgage lending increased steadily and reached 1.36% of total GDP in 2008 (see Fig. 1 and 2). Additionally, in 2008 the average weighted mortgage interest rate and maturity were 12.9% and 215.3 months, respectively. By comparison, the corresponding figures in the year 2005 were 14.9% and 174.6 months, respectively.

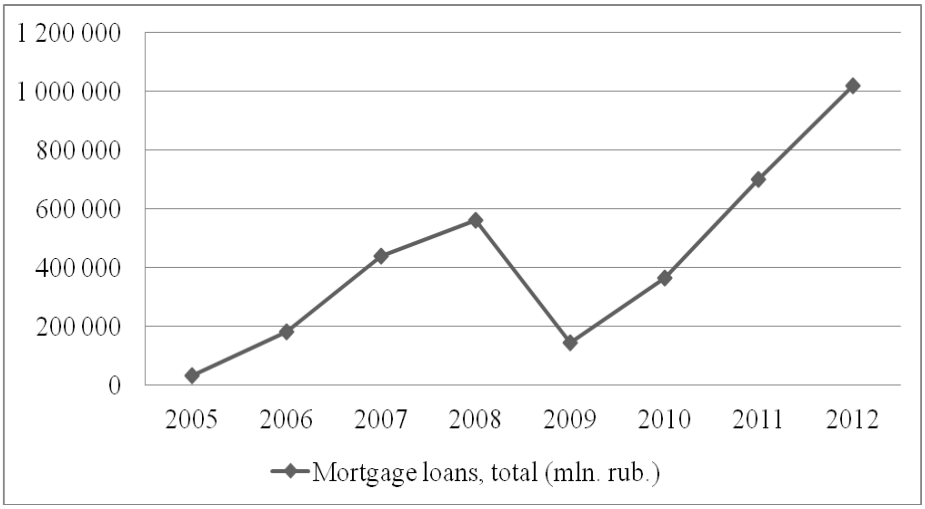

Fig. 1. Mortgage loans (mln. rub.)

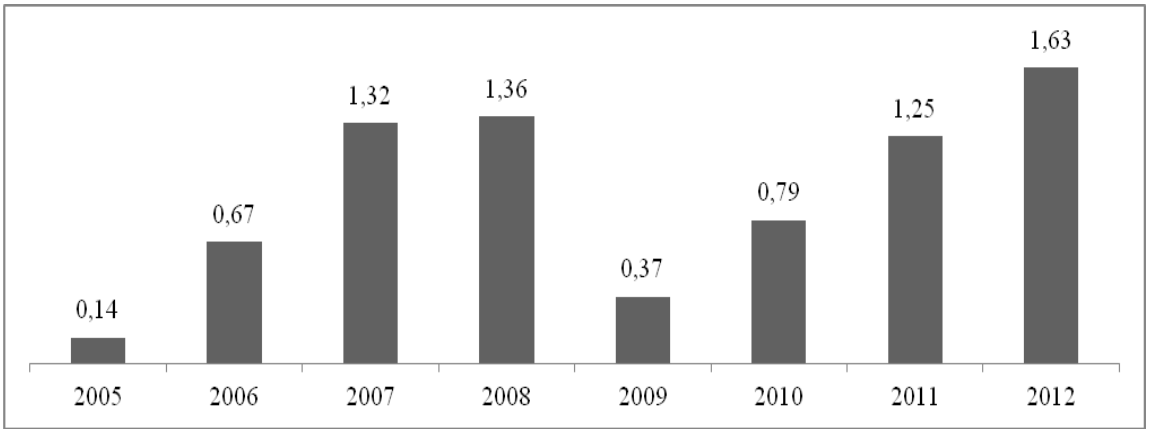

Fig. 2. The volume of mortgage residential lending (% from total GDP)

However, during 2009 a loss of capital by credit organizations, a decrease in bank credit capacity, and a tightening of credit conditions along with reduction of effective demand of households and a transition to "waiting-and-saving" strategy were observed. These events initially affected not only a substantial reduction in the volume of mortgage lending in 2009, but growth of the average weighted mortgage interest rate to 14.3%, together with an increase in

overdue mortgage loans and defaults (Stolbov, 2012; AHML, 2009) (see Fig.1-3). Triggers of the Russian mortgage crisis in the years 2008-2009 have been discussed in detail, for example by Stolbov (2012), Sternik (2009).

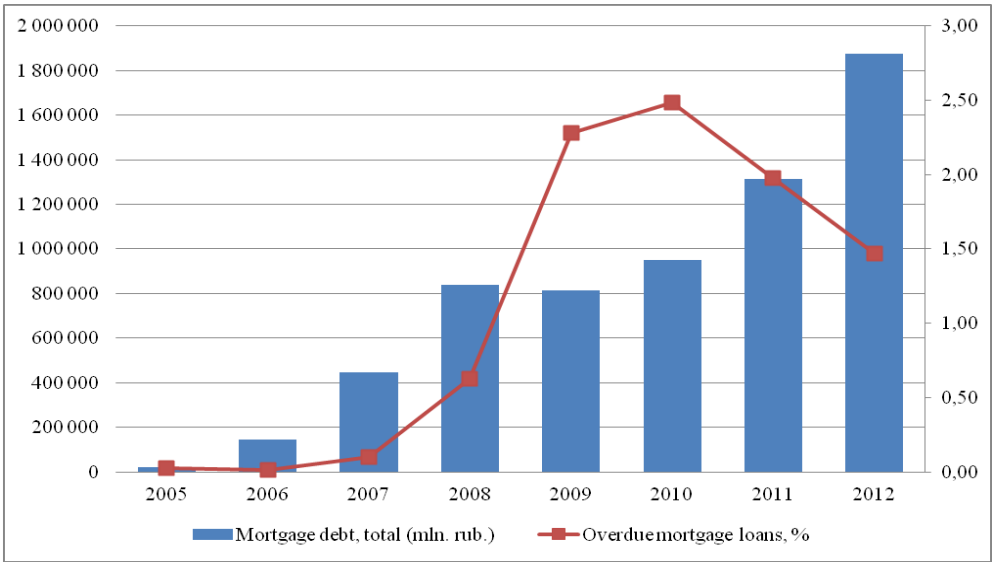

Fig. 3. Overdue mortgage loans (% from mortgage debt)

On the one hand, the residential mortgage market opens up opportunities for rapid growth in real estate development, in the production of construction materials and for the development of related sectors. Moreover, financing housing construction with a mortgage loan creates collateral value that can be used by households as collateral for future further loans. As a result this causes an increase in total consumption (Stolbov, 2012).

On the other hand, the negative effects of the Russian mortgage crisis during the years 2008-2009 spread quickly across the financial sector and to other sectors of the economy. This has emphasized the importance of understanding drivers to default on mortgages and the shortcomings of the current credit risk practices. For these reasons regulators, credit organizations, policymakers, and researchers pay special attention to the problems of mortgage lending, credit risk evaluation and developing effective risk management systems, especially under the Internal Ratings-Based (IRB) approach. The IRB approach is an alternative way to a standardized approach to assess credit risk, based on the internal ratings of credit organizations (BIS, 2006; Bank of Russia, 2012). However, the experience in developing effective internal rating systems for credit risk evaluation of mortgage borrower in the Russian bank practice is rather limited.

This research investigates factors which are influenced by the mortgagor's default decision on the residential mortgage market that can be used to a develop credit risk system under IRB approach. The paper has the following structure. The next section contains a literature review and some generalizations about recent studies of the default process. Section 3 contains a

description of the collected data and an estimation strategy, which allows correcting for sample selection bias. In Section 4 the results are discussed, followed by Section 5, in which the semiparametric approach is employed to ensure robustness of results. The final section describes the conclusions and directions for further work.

Literature Review

Different concepts are used to measure credit risk, such as probability of default (PD), loss given default (LGD), exposure at default (EAD), maturity (M) and correlated defaults (CD). But default is regarded as the worst event of credit risk and it is arguably most relevant to the recent Russian mortgage crisis and related spillover effects.

The notion of mortgage default has not yet been incorporated in the Russian legislation. According to BIS (2006) and the Bank of Russia (2012) a borrower is in default if any of the following credit events happen:

- a borrower cannot repay a loan without selling collateral (for mortgage loans, collateral is real property);
- monthly payments are not met for 90 days or more.

From the beginning of the 1960s and extending to the present, an important stream of literature has addressed the default problem. Many researchers have proposed simple single-equation models of the mortgage lending process, but many papers have emphasized that the mortgage lending process consists of related or sequentially dependent mortgage lending decisions. It means that far more complex econometric specifications (multiple-equation models) and econometric techniques allowing dealing with sequential selectivity with limited dependent variable are needed.

In the pioneer paper, Follain (1990) presented a theoretical model that explained mortgage choice by a homeowner that is more general than mortgage demand. It includes three components: (1) the choice of a Loan-To-Value ratio (LTV), (2) the refinancing and default decisions, and (3) the choice of mortgage instrument (adjustable [ARM] or fixed [FRM] interest rates).

Then Rachlis and Yezer (1993) suggested a theoretical model of the mortgage lending process followed by Maddala and Trost (1982) which consists of a system of four simultaneous equations indicating equations for: (1) the probability of applying for credit, (2) the choice of mortgage contract terms (LTV and maturity), (3) the probability of the endorsement, and (4) the probability of foreclosure.

Both papers discussed estimation techniques that can be used for such kind of

simultaneous equation systems. They also explored the necessity for better understanding of mortgage choices to answer important policy questions, but without any empirical framework.

Since the mid-1990s, mortgage data were made publicly available, e.g. American mortgage datasets from the Federal Housing Authority (FHA) foreclosure, the Boston Fed Study, the Home Mortgage Disclosure Act (HMDA), and several empirical studies about the interdependency of a bank endorsement and the borrower's decisions modeled by bivariate probit model (BVP) appeared.

As an extension of study (Rachlis, Yezer, 1993), Yezer et al. (1994) used a simulation method to estimate the above-listed theoretical model. They empirically have shown that isolated modeling processes of the credit underwriting and default leads to biased parameter estimates. Their findings were supported by Phillips, Yezer (1996) and Ross (2000). The first ones compared the estimation results of the single equation approach with those of BVP. By using demographic characteristics of the household, income, loan amount and characteristics of mortgage to underwriting and default decisions. They showed that isolated modeling processes of the credit underwriting and default lead to biased estimates. For this reason, corrections for the sample selection bias should not be ignored. The second one compared a single-equation default model and a bivariate probit model jointly estimated loan denials with loan performance. Ross (2000) 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 single-equation default model suffers from substantial selection bias mainly due to the omission of credit history that is only in the application sample. The author used two sets of data – for conventional loans (HMDA data) and FHA loans. However, the assumption that their underwriting models are similar seems questionable.

Bajari et al. (2008) applied BVP with partial observability. Empirical findings indicate that borrower characteristics, terms of credit contract, and fundamental characteristics play important roles in explaining the default. Specifically, due to the lack of sociodemographic information at the individual level, authors included country sociodemographic information and country unemployment rate as proxy variables. Moreover, the main driver of default is the nationwide decrease in home prices. The estimated effect of housing prices on default behavior implies that default will be highly geographically correlated when home prices decline nationwide. Several empirical studies found that highly statistically significant variables in explaining mortgage default are not only micro-level sociodemographic factors such as age, race, marital status, and income of borrower, but also regional unemployment rate, divorce rate and loan terms (Archer et al., 1996; Deng at al., 2000; Pavlov, 2001; Goldberg, Harding, 2003; Clapp et al., 2006).

A new stream of mortgage studies has come from the financial turmoil of 2007-2009 in the USA. Several recent empirical findings reported by Mian and Sufi (2009), Demyanyk and Van Hemert (2011), Dell'Ariccia et al. (2012) confirm the high statistical significance of macroeconomic conditions in explaining mortgage default. Obtained results are consistent with the notion that a relaxation of lending standards, triggered by an increased demand for loans, contributed to the boom and the ensuing crisis, together with other supply-side explanations such as house price appreciation and mortgage securitization (Keys et al., 2010; Dell'Ariccia et al., 2012).

Other important default determinants are trigger events or shocks like loss of a job, divorce or death, changes in marital status, education, neighborhood effects, and job-hopping. (Vandell, 1995; Archer et al., 1997; Deng et al., 2000). Also, several studies show that mortgage default is determined by the desire or the necessity to move that are determined by transaction costs (time costs, moving costs, sell and purchase costs), real estate investment incentives, household life-cycle (age), education, occupation, income, and race (Archer et al., 1996; Pavlov, 2001; Deng et al., 2005).

The default decision might be driven not only by economic factors, but emotional considerations. Guiso et al. (2013) found that people who are angrier about the current economic situation are more willing to express their desire to default, as are people who trust banks less. Moreover, the authors emphasized the effect of a social contagion. For example, people acquainted with somebody who defaulted strategically are more likely to declare their intention to do so. From the point of behavioral finance, borrowers vary in their personal circumstances, and in their ability to manage their financial affairs in their own long-run interest. Three particularly important types of heterogeneity are in moving propensity, financial sophistication, and present-biased preferences (Campbell, 2012).

Russian mortgage studies are mainly focused on the triggers of the Russian mortgage crisis from 2008-2009 (Stolbov, 2012; Sternik, 2009) and the discussion of different strategies to develop the mortgage market in Russia (Polterovich, Starkov, 2007). Yet, little is currently known about the mortgage default drivers on the Russian residential mortgage market. A potential explanation for this fact is the lack of micro-level data on mortgage loans.

This study is to purpose a mortgage default model. The goal of the paper is not to try to derive an optimal credit underwriting system or an optimal mortgage contract (Piskorski, Tchistyi, 2010), to solve the equilibrium model of mortgage choice (Corba, Quintin, 2010), or to develop dynamic mortgage default model (Campebell, Cocco, 2011). But instead, this paper investigates determinants of mortgage default within an empirical application to the Russian residential mortgage market. Comparing results across parametric and semiparametric

econometric models are performed as robustness checks of our results.

Methodology and Data

The main goal of this paper is to estimate PD, which is a discrete dependent variable y_1 (*flag of default*).

$$y_1^* = x_1\beta_1 + \varepsilon_1, \quad (1)$$

$$y_1 = \begin{cases} 1, & if \quad y_1^* > 0, \\ 0, & otherwise. \end{cases} \quad (2)$$

$$x_1 = (D_1, C_1, M_1) \quad (3)$$

where y_1^* is a vector of unobserved latent variable. In credit underwriting process y_1^* is defined as the probability of approval of an applicant by the credit organization. x_1 is a vector of exogenous independent variables[3], which includes vectors of sociodemographic characteristics (D_1), terms of credit contract (C_1) and macrovariables (M_1) which are discussed further. β_1 is a vector of regression coefficients. Finally, ε_1 is an unobserved error term.

The data used in this research contains two sets. The first data set is aggregated regional monthly data on the Agency of Home Mortgage Lending (AHML) branch performance, mortgage market characteristics and regional macroeconomic variables for the period from 08/2008 to 08/2012. This data set is publicly available.

AHML is a national institute for the development of housing activity that was established in 1997. It helps to implement strong government housing policies and anti-recessionary measures to support mortgage lending in Russia. AHML is a state-owned provider of government-insured loans, which uses a two-level system of lending (see Fig. 4). In the first step banks and non-credit organizations provide mortgage loans to households according the common standards of AHML. Network of AHML partners consists of about 136 banks and 200 non-banking organizations. 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 that operate such programs. The list of programs contains "Young Researchers", "Young Teachers", "Mortgage for Soldiers", "Mothers' Capital" and other social credit programs. All of them have relatively high risk that is insured by government. Considering this, the demand for these kinds of mortgage programs and behavior of borrowers is

[3] Estimation of a specification with selection and endogeneity issues, especially for semiparametric or nonparametric models, is challenging. Ozhegov, Poroshina (2014) assumed that terms of credit contract are endogenous and followed Attanasio et al. (2008) approach of endogeneity and sample selection modeling to the Russian mortgage residential loans for strictly parameterized models.

generated by some special subsample of potential borrowers, different from the general population.

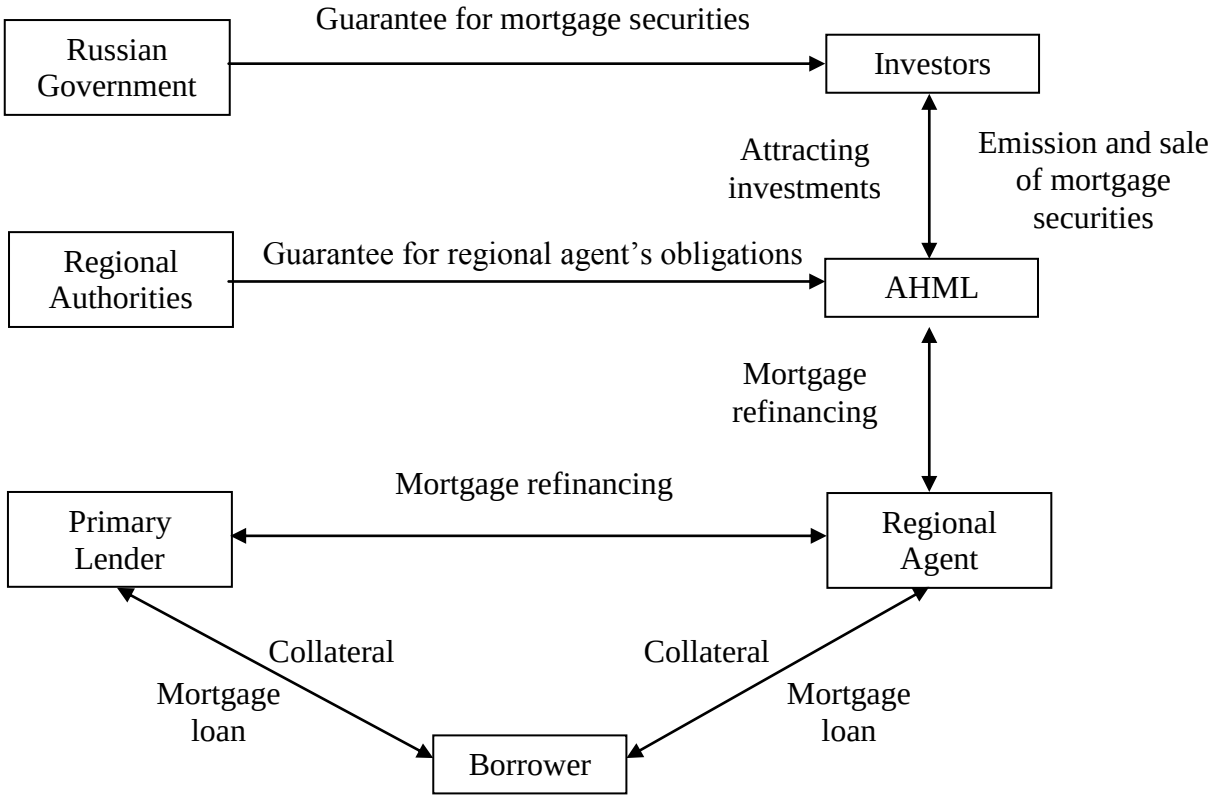

Fig. 4. AHML's lending system

The second set is provided by one regional AHML branch and includes loan-level data on 4298 applicants for mortgage loans (after data cleaning from initial 4897 observations). Our estimation uses daily data on applicants that applied for mortgage loans in the period 08/2008 – 08/2012. For each applicant, we observe an applicant's sociodemographic characteristics at the time of application, the flag of credit organization's approval decision based on the credit underwriting process, and the applicant's flag of contract agreement. For each originated mortgage loan, we observe the loan terms and property characteristics at the time of origination, and the flag of default. The flag of default is only observed (delinquent monthly payments for over 90 days), but not the date of default. The variables used in the estimation are defined in Table A1 and A2.

The data set of 4298 applicants includes both approved and rejected ones in total proportion 86:14 that corresponds to a reject rate of 14%. Although Fig. 5-6 show that it was non-stationary. However, only 2799 borrowers (75.7% from total number of approved applicants) have mortgage loans. From this amount, 5.9% were defaulted. The problem of data disproportion is typical in credit risk modeling. According to Maddala (1992, p.325), in the

estimation of a binary choice model or even a linear probability model it influences only the estimated intercept, but not other estimated parameters.

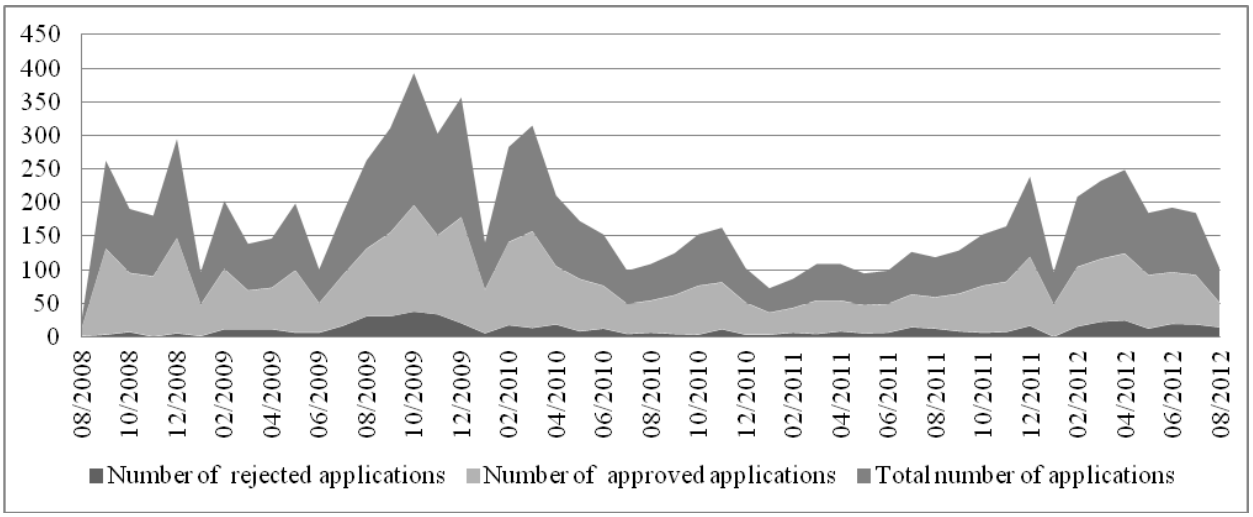

Fig. 5. Number of approved and rejected applications for the time period 08/2008 – 08/2012

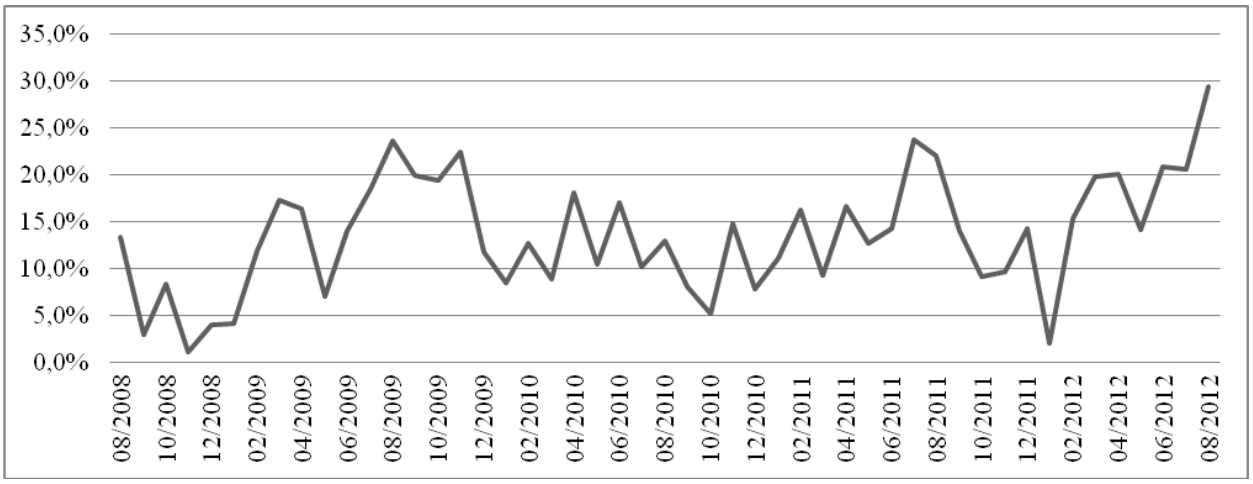

Fig. 6. Dynamics of reject rate for the time period 08/2008 – 08/2012

When trying to estimate a classical binary choice model (probit and logit) for PD (1)-(3), we are faced with a number of sample selection problems leading to biased parameter estimates. First, the sample selection bias is due to a simultaneity bias. This problem arises when mortgage default modeling does not take into consideration the underwriting process. The decision of endorsement or decline of a credit application is based on the latter process. Second, the truncation or the partial observability causes this bias induced by the nature of our data. PD and terms of credit contracts are observed only for approved borrowers. We are faced with this issue when information about denied applicants is absent. Therefore, the magnitude of bias depends on the degree of correlation between two processes – the default process and the credit underwriting process.

For these reasons, we model these selection issues as an extension of the classic Heckman model (1976, 1979), known as a bivariate probit model (BVP) with sample selection correction, by adding to probit model (1)-(3) for PD following conditions:

$$y_2^* = x_2\beta_2 + \varepsilon_2 \tag{4}$$

$$y_2 = \begin{cases} 1, & if \quad y_2^* > 0, \\ 0, & otherwise. \end{cases} \tag{5}$$

$$x_2 = (D_2, M_2) \tag{6}$$

$$y_1 = \begin{cases} x_1\beta_1 + \varepsilon_1, & if \quad y_2^* = 1, \\ unobserved, & otherwise. \end{cases} \tag{7}$$

$$y_2^* \ is \ observed \ for \ all \ classes. \tag{8}$$

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho \\ \rho & 1 \end{pmatrix}\right] \tag{9}$$

where y_2^* is a vector of unobserved latent variable. y_2 is the probability of endorsement (*flag of endorsement*[4]). x_2 is a vector of exogenous independent variables[5], which includes vectors of sociodemographic characteristics (D_2) and macrovariables (M_2). β_2 is a vector of regression coefficients. ε_2 is an unobserved error term. Typically the errors are assumed independent and identically distributed as a standard bivariate normal with correlation ρ.

If the errors between the two probit models (outcome equation (1) for PD and selection equation (4) for the probability of endorsement[6]) are independent of one another i.e. $\rho=0$, then we can just estimate the two probit models separately. In this case, it is possible to obtain consistent estimates. However, when $\rho\neq0$, it means that two decisions (lender's decision in the credit underwriting process and borrower's default decision) are interrelated and the joint estimation of these equations gives more efficient parameter estimates.

The credit underwriting process is one of the main steps in the mortgage lending process that includes an assessment of the applicant's ability to pay off a mortgage. It is a complex procedure that includes analysis of applicant's creditworthiness, liquidity of property to be mortgaged, terms of mortgage loan, and credit risk evaluation. Based on the credit underwriting process, a credit organization makes a decision to approve or reject an application for a mortgage loan.

The lender's decision to approve an applicant is mainly based on the analysis of LTV and

[4] Italicized variables are used in estimation.
[5] For more details see Ozhegov, Poroshina (2014).
[6] We do not attempt to model application process, mainly because we do not have the micro-level data to do so. The paper (Ozhegov, Poroshina, 2013a) presented the demand-supply model for the Russian residential mortgage market based on the aggregated data. Robust estimates of demand function evidenced that decision on application for mortgage loan is lagged and depends not only on the current macroeconomic situation, but also on the shocks of previous periods.

DTI ratios, and credit history of potential borrowers. Both ratios include mortgage loan terms that are observed in our sample only in case of signed mortgage contracts. For this reason, we control for sociodemographic characteristics (D_2) that are included in an application form and can potentially influence an applicant's creditworthiness, such as *age of borrower* and *sex* (to capture possible discrimination), *categorical income of main borrower, educational level, family status, activity category, categorical income of co-borrowers* and macrovariables (M_2) such as *probability of application* and *unemployment rate*. Tests of equality of group means and medians in Tab. A3-A4 show that most of them have discriminating power across the approved and rejected applicants.

The important identification assumption (excluded restriction) is that at least one variable (*monthly income of co-borrowers*) enters the credit underwriting process and does not influence on the mortgage default process. It is possible to estimate the BVP model (1)-(9) by maximum likelihood (ML), but sometimes it is difficult to get convergence in the estimation process. For this reason, the Heckman's two-step procedure is used. It is based on using the consistent estimates of the inverse Mill's ratio[7] (10) to estimate PD model (1)-(3).

$$\hat{\lambda}(x_2\hat{\beta}_2) = \frac{\phi(x_2\hat{\beta}_2)}{\Phi(x_2\hat{\beta}_2)}, \tag{10}$$

where $\phi(x_2\hat{\beta}_2)$ and $\Phi(x_2\hat{\beta}_2)$ are, respectively, the density and cumulative density function of N(0,1).

According to the mortgage literature, categories of credit risk factors generally include sociodemographic information (D_1), mortgage loan terms (C_1) and macroeconomic conditions (M_1).

Firstly, we control for sociodemographic factors (D_1), specifically *age of borrower*, *sex*, *education level*, *family status*, *activity category* and *monthly income of main borrower*. It is assumed that males and females as individuals of different age-groups have different credit discipline as a consequence of different levels of PD. 92.2% of total sample are middle-aged hired employees, married individuals are 56.9% and females are 56.3% of total sample. Age, family status and activity category are potentially connected with such trigger events as death or illness, divorce and loss of job (income shocks) that can influence a borrower's default decision. As mentioned by Moody's, PD for entrepreneurs (*entrepreneur*) (0.9% of total sample) is higher because their income is more sensitive to economic downturns (Moody's, 2008). In addition, they do not have wide job skills and are more likely to face problems finding a job. We suggest also that the level of education could be regarded as a proxy for a borrower's financial literacy

[7] For more details see Greene (p.781-782, 2003).

which could influence credit discipline. In our sample, most individuals have higher education (49.8%) and secondary one (40.7%).

Naturally, the income of the main borrower has a direct effect on credit repayment. The inclusion of a borrower's income in the estimation is potentially problematic because 86.6% of borrowers do not declare their income. The reason is that AHML branch offers credit programs that do not require declaring some sociodemographic characteristics such as income of main borrower (*not declared income of main borrower* - known as "low doc" and "self-certified" loans), family status (*not declared income of main borrower*), activity category (*not declared activity category*) and education level (*not declared educational level*). We suggest that these are nonrandom missing values that are potential trigger events of a borrower's default. For this reason, we generate special dummies for these categories of individuals and investigate how they affect default probabilities. We also experimented with alternative specifications, in which a set of dummies for *categorical DTI* (including *not declared DTI*) is used instead of dummies for the categorical monthly income of main borrowers. The results were almost unchanged. We therefore report only specifications with income (Tab. 2).

Secondly, loan terms (C_1) determine the extent of financial burden for borrowers. In practice, they are used as proxy variables to estimate the credit risk of a particular borrower. The correlation analysis in Tab. A6-A7 shows that loan terms are highly correlated. To avoid multicollinearity and identification problems, we choose factors that are more likely to influence default probabilities and ones with influence that are less well understood or discussed in mortgage literature.

A borrower defaults when he/she is faced with a substantial loan debt burden that is defined as well by mortgage loan terms. Specifically, a borrower defaults when he/she cannot repay monthly payments that consist of debt amount and interests charged on loan. Monthly payments are defined by a credit program and depend also on *downpayment*, loan amount, contract rate (*rate*), *maturity* and flat value. For example, a longer maturity reduces the size of monthly payment and increases the uncertainty about a borrower's future. More than 70% of observations are mortgages with maturity exceeding 15 years. It allows a consumer to borrow a larger mortgage loan, but increase the financial burden for the borrower. At the same time, in our sample history of mortgage credits are observed during different amounts of days. We suggest that the total amount of days observed (*duration*) in credit has a positive effect on PD following the methodology of Moody's (2009).

We investigate the extent to which *categorical LTV* and *DTI* ratios at mortgage origination affect mortgage PD. Higher LTV leads to smaller downpayment, but increases monthly payments and contract rate. As a result, a higher LTV ratio reduces borrower's

incentives to meet mortgage payments and increases PD that was empirically documented by several researchers, for example, Mayer et al. (2009). To control for the incidence of mortgage default, regulators in many countries ban high LTV ratios. Additionally, credit organizations require borrowers to take out mortgage insurance. The data in Tab. 1 shows that the sample contains borrowers with different LTV. On average, borrowers are not high risky, because the sample mean LTV equals 56%, which is much less than 90%. The sample assessed property value approximately 2 mln. Russian rub., which is common for secondary real market in local area.

DTI ratio is a measure of mortgage affordability that is often used by credit organizations to determine loan limit and the contract rate. An average 45% of monthly income is spent to repay mortgage payments. In the Tab. 1 it shows DTI, which has a larger effect on borrowers with low credit quality. With a significant increase in the share of mortgage payments, PD increases even at the slight reduction of a borrower's income due to unexpected life circumstance.

Finally, the exiting literature on mortgage default has emphasized the role of macrovariables (M_1) in explaining PD. Macrovariables characterize the market demand and supply. Dell'Ariccia et al. (2012) confirmed the highly statistical significance of macroeconomic conditions in explaining mortgage default. Specifically, such supply factors include house price appreciation (*mean m2 value*) and mortgage securitization. Moreover, the contribution of local economic conditions and change of credit underwriting standards to default are significant, too (Cutts, Merrill 2008). Regional unemployment rate (*unemployment rate*) is supposed to capture expectations about future macroeconomic developments that may affect loan markets and the risk of job loss (Attanasio et al., 2008). Other factors that can influence the default decision include changes in marital status, education, neighborhood effects, job-hopping, emotional state, financial sophistication and present-based preferences, but they are unobservable.

In addition, tests of equality of group means (t-test and ANOVA-test[8]) and medians (Wilcoxon-Mann-Whitney test), and the Chi-square test (and Fisher's exact test when one or more of the cells in cross-tabs have an expected frequency of five or less) in Tab. A3-A4 show that all of the above-mentioned sociodemographic factors such as loan terms and macrovariables have discriminating power across the defaulted and non-defaulted borrowers.

In the estimation of PD, we used BVP with correction for sample selection and basic specifications are presented in Tab. 1-2.

[8] Tests assume the normality distributed independent variables. Skweness and kurtosis normality test rejected the hypothesis that sociodemographic characteristics and terms of credit contract in Tab. A1-A2 are normally distributed. However, due to the Central Limit Theorem, the normality assumption is not problematic when sample size is sufficiently large. In practice, the sum of observations in both groups has to be more then 30.

Tab. 1. Model specifications for the probability of endorsement

Parametric approach[9]	
(1)	Probit $probability\ of\ endorsement = f_1(D_2, M_2)$, $M_2 = (probability\ of\ application)$
(2)	Probit $probability\ of\ endorsement = f_2(D_2, M_2)$, $M_2 = (probability\ of\ application, unemployment\ rate)$
(3)	Probit $probability\ of\ endorsement = f_3(D_2, M_2)$, $M_2 = (unemployment\ rate)$
(4)	Probit $probability\ of\ endorsement = f_4(D_2, M_2)$, $M_2 = (probability\ of\ application, unemployment\ rate, unemployment\ rate^2)$
(5)	Probit $probability\ of\ endorsement = f_5(D_2, M_2)$, $M_2 = (unemployment\ rate, unemployment\ rate^2)$

Tab. 2. Model specifications for PD

Parametric approach[10]	
(1)	BVP1 $PD = h_1(D_1, C_1, M_1, \hat{f}_1)$
(2)	BVP2 $PD = h_2(D_1, C_1, M_1, \hat{f}_2)$
(3)	BVP3 $PD = h_3(D_1, C_1, M_1, \hat{f}_3)$
(4)	BVP4 $PD = h_4(D_1, C_1, M_1, \hat{f}_4)$
(5)	BVP5 $PD = h_5(D_1, C_1, M_1, \hat{f}_5)$

Results

Tab. 3 presents the results of the parametric estimation of the PD[11]. In actual estimation, we always found that borrower age squared, LTV squared, and unemployment rate had no explanatory power. For this reason we dropped it from the model and do not report these specifications. In the following discussion we focus on the coefficients of the default equation according to our primary aim of research; we do not report the results of the selection equation[12].

Columns 2-6 in Tab. 3 reported estimates of average marginal effects for BVP models for PD that correct for the sample selection. We estimate various specifications for the probability of endorsement, and confirm the robustness of our results to the functional form of selection equation[13]. The percent of correct predictions remains unchanged – 86.9% and areas under ROC curves (AUC) do not differ significantly. Fitted probabilities of endorsement are used to estimate BVP models for PD.

[9] We estimated logit models for the probability of application. Because the results of logit models are mostly the same to probit models, we report only probit ones. D_2 includes linear sociodemographic characteristics that are discussed earlier.

[10] D_1, C_1, M_1 consisted of sociodemographic information, mortgage loan terms and macroeconomic conditions that are linear and discussed earlier.

[11] We estimated both parameters of PD models and average marginal effects. Because the signs and statistical significance of most parameters are mostly the same, we report only estimates of average marginal effects.

[12] The discussion about credit underwriting process and other stages of mortgage lending process is presented in (Ozhegov, Poroshina, 2013a) and (Ozhegov, Poroshina, 2013b).

[13] Independent variables that enter the selection equation are common for all specifications and discussed in methodology and data section. They vary in terms of macrovariables: probability of application and unemployment rate that is presented in Tab 1.

Tab. 3. Average marginal effects for probability of default (Parametric approach)

	BVP_1	BVP_2	BVP_3	BVP_4	BVP_5
Age of borrower	3.58×10^{-4}	3.58×10^{-4}	3.52×10^{-4}	3.47×10^{-4}	3.24×10^{-4}
	(0.64)	(0.64)	(0.63)	(0.63)	(0.59)
Sex	0.030***	0.030***	0.030***	0.030***	0.029***
	(3.43)	(3.43)	(3.41)	(3.45)	(3.41)
Not declared educational level	-0.013	-0.013	-0.012	-0.014	-0.011
	(-0.43)	(-0.43)	(-0.40)	(-0.44)	(-0.36)
Secondary education	0.019	0.019	0.020	0.019	0.020
	(0.77)	(0.77)	(0.79)	(0.76)	(0.78)
Complete higher education	0.026	0.026	0.026	0.026	0.025
	(1.00)	(1.00)	(0.99)	(1.00)	(0.97)
Not declared family status	0.071***	0.071***	0.070***	0.070***	0.069***
	(3.13)	(3.13)	(3.12)	(3.17)	(3.12)
Single	0.023**	0.023**	0.023**	0.023**	0.023**
	(2.51)	(2.51)	(2.51)	(2.46)	(2.47)
Widowed	6.48×10^{-3}	6.48×10^{-3}	7.42×10^{-3}	5.96×10^{-3}	7.89×10^{-3}
	(0.19)	(0.19)	(0.22)	(0.18)	(0.24)
Divorced	7.51×10^{-3}	7.51×10^{-3}	7.85×10^{-3}	7.18×10^{-3}	7.93×10^{-3}
	(0.60)	(0.60)	(0.63)	(0.58)	(0.64)
Hired employee	0.030	0.030	0.029	0.030	0.028
	(1.02)	(1.02)	(0.99)	(1.05)	(0.96)
Entrepreneur	0.054	0.054	0.053	0.055	0.051
	(1.26)	(1.26)	(1.24)	(1.32)	(1.23)
State employee	0.096*	0.096*	0.095*	0.099**	0.094**
	(1.96)	(1.96)	(1.93)	(2.05)	(1.97)
Not declared income of main borrower	-0.117***	-0.117***	-0.115***	-0.118***	-0.110***
	(-2.61)	(-2.61)	(-2.59)	(-2.80)	(-2.72)
Income of main borrower 10000-19999	-0.034*	-0.034*	-0.034*	-0.035**	-0.033*
	(-1.95)	(-1.95)	(-1.92)	(-2.01)	(-1.93)
Income of main borrower 20000-39999	-0.057**	-0.057**	-0.056**	-0.058***	-0.055**
	(-2.49)	(-2.49)	(-2.46)	(-2.60)	(-2.52)
Income of main borrower 40000 and more	-0.046	-0.046	-0.045	-0.047	-0.041
	(-1.55)	(-1.55)	(-1.52)	(-1.64)	(-1.54)
Rate	0.027***	0.027***	0.027***	0.027***	0.027***
	(10.39)	(10.39)	(10.36)	(10.30)	(10.27)
Maturity<120 months	0.044*	0.044*	0.044*	0.043*	0.043*
	(1.78)	(1.78)	(1.79)	(1.76)	(1.76)
Maturity 120-179 months	0.036*	0.036*	0.036*	0.035	0.035
	(1.66)	(1.66)	(1.66)	(1.63)	(1.64)
Maturity 180-239 months	0.030	0.030	0.031	0.030	0.030
	(1.46)	(1.46)	(1.47)	(1.44)	(1.45)
Maturity 240-299 months	0.027	0.027	0.028	0.027	0.028
	(1.25)	(1.25)	(1.27)	(1.24)	(1.26)
LTV<0.5	6.53×10^{-3}	6.53×10^{-3}	6.53×10^{-3}	6.40×10^{-3}	6.43×10^{-3}
	(0.63)	(0.63)	(0.63)	(0.62)	(0.62)
LTV>0.7	1.71×10^{-3}	1.71×10^{-3}	1.58×10^{-3}	1.41×10^{-3}	1.17×10^{-3}
	(0.17)	(0.17)	(0.15)	(0.14)	(0.11)
Downpayment	-9.87×10^{-10}	-9.87×10^{-10}	-9.44×10^{-10}	-9.96×10^{-10}	-9.43×10^{-10}
	(-0.19)	(-0.19)	(-0.18)	(-0.19)	(-0.18)
Duration	9.14×10^{-5}***	9.14×10^{-5}***	9.18×10^{-5}***	8.82×10^{-5}***	8.93×10^{-5}***
	(3.26)	(3.26)	(3.29)	(3.21)	(3.25)
Mean m2 value	1.46×10^{-6}**	1.46×10^{-6}**	1.49×10^{-6}**	2.64×10^{-6}***	2.00×10^{-6}***
	(2.33)	(2.33)	(2.37)	(2.81)	(2.75)
Fitted probability of endorsement 2	-0.050*	-0.050*			
	(-1.93)	(-1.93)			
Fitted probability of endorsement 3			-0.049*		
			(-1.89)		
Fitted probability of endorsement 4				-0.051**	
				(-2.09)	
Fitted probability of endorsement 5					-0.046**
					(-1.99)
Observations	2728	2728	2728	2728	2728
Pseudo R^2	0.444	0.445	0.444	0.445	0.445
AIC	748.2	747.8	747.9	747.1	747.5
BIC	913.7	913.3	913.5	912.6	913.1
Log likelihood	-346.1	-345.9	-346.0	-345.6	-345.8
AUC	0.9430	0.9432	0.9432	0.9432	0.9431

% Right predictions	94.5	94.5	94.5	94.5	94.5
p-value (link test)	0.412	0.470	0.487	0.616	0.612

Note: t statistics in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Categories that were taken as base outcomes: education level "Incomplete higher education"; family status "Married"; activity category "Unemployed"; income level 0-9999; maturity >=300 months; LTV 0.5-0.7.

Our results of parametric estimation for various PD specifications are almost unchanged in terms of statistical significance of estimated average marginal effects and their signs. We empirically confirm the necessity to correct for sample selection bias by finding strong evidence of high statistical significance of fitted probability of endorsement. This effect remains the same across alternative specifications for BVP models. The significant negative correlation of credit underwriting and default processes not only support the hypothesis about sample selection bias, but also show that credit organizations tend to approve less risky borrowers in term of payment discipline. Put differently, higher PD decreases the probability of endorsement of a particular borrower. However, this result is unstable in semiparametric estimation that will explain further.

We found high statistical significance in the explanation of PD sex, not declared family status and income of main borrower, and rate and age of the credit. Dummies for borrowers with monthly income 40,000 rub. and more do not have statistical power in all BVP models as well as dummy for maturity exceeding 15 years. These findings will explain further. To give further economic interpretation of main determinants of PD in mortgage lending we rather use estimated parameters from parametric estimation of BVP_4 with the highest AUC and log likelihood function from various parametric specifications.

Educational level is not significant which means that financial literacy does not play a significant role, as we expected, in the mortgage default decision.

The results confirm that PD is higher for males, borrowers with not declared family status or single, and state employees. Average marginal effects of these explanatory variables on PD are 0.03, 0.07, 0.02 and 0.09, correspondingly. These facts should be explained.

The number of females in the total sample is higher than males that correspond to the current demographic situation. Most of female borrowers (41.4%) are single who tend to be more responsible for their mortgage obligations and have strong credit discipline. While male individuals are mostly married (72.6%) and may divorce. Moreover, male life expectancy is relatively short and they are more subject to illness that leads to the risk of job loss. These results correspond to Deng et al. (2000).

Not declared family status is highly statistically significant and should be regarded as a nonrandom missing value. This category of borrower may consist of individuals with a high risk of divorce that is a potential trigger event for mortgage default, as we mentioned earlier. In addition, single borrowers are statistically significant, but on the less significant level than

borrowers with not declared family status and have a smaller value of average marginal effect. Taking into account that most of them are males, the explanation is the same regarding the higher PD for males.

Also we found a significant and unstable activity category "State employee". In different specifications with BVP, this variable is statistically significant on 5% and 10%. The positive influence of this activity category on the PD is mostly explained by consequences of global financial crisis. It is accompanied by mass laying off and staff reduction of state employees. For this reason, state employees have a high risk of dismissal that is related to unstable borrower's income and delinquent mortgage payments.

The relationship between PD and borrower's income is statistically significant, as expected. However, in Russia a substantial proportion of income comes from off-the-books economy, which amounts to about 15 % of total GDP. In developing countries, it can reach up to 60-70% of GDP. Moreover declared income often does not correspond to real income. Borrowers who do not declare their income most probably have real income that cannot be officially confirmed. This fact explains that they meet mortgage obligations and have the lowest PD with corresponding average marginal effect -0.11.

The results show that lower income borrowers have the highest PD. This could be due to the fact that this subgroup of individuals may have insufficient and unstable income and most probably are faced with problems in mortgage loan performance.

As we expected, PD is higher with a higher contract rate that becomes an additional financial burden for borrowers, but average marginal effect of the rate on PD is relatively small (0.03). Despite the fact that maturity is another significant factor, it is difficult to make a statistical inference. Firstly, most mortgages are long term (maturity exceeding 15 years). Taking into account that database dates are from 2008, we observe these mortgages during a short period of time and default has not yet taken place. This explains low statistical significance for 10- and 15-year mortgages. Secondly, this assumption is confirmed by the statistical significance of loan age. A positive sign of it means that the probability to observe default increases with the number of days observed in mortgage loans, but the average marginal effect is close to zero.

Mortgages with low LTV are attractive for non-liquid borrowers. These are also special credit programs for young teachers and households with "Mothers' capital". The probability that they could encounter a serious repayment problem with a loan is much higher. Moreover, borrowers with LTV higher than 70 % think as holders, because they do not invest a lot of their own capital and are 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. Surprisingly, LTV is not statistically significant for the residential mortgage

market[14]. These results correspond to Campebell, Cocco (2011) who modeled mortgage default using a dynamic model based on the concept of a rational utility-maximizing household.

Shocks of house prices in a region contribute significantly in the explanation of PD, but its average marginal positive effect is quite small. It shows the strong relationship between the residential housing market and mortgage lending market.

Robustness Checks

Next, we check the robustness of our results. Despite the fact that BVP models allow controlling for the sample selection bias, they are strictly parameterized. The main limitation is that there is no prior knowledge about true distribution. As a result, the misspecification problem leads to misestimating and wrong inferences (Creel, 2008). In principle, to be more flexible in specification, we relax into parametric approach the assumptions of know a joint normal distribution for the error terms ε_1 and ε_2 and specific functional form of some variables in vectors x_1 and x_2 We apply the approach of Attanasio et al. (2008) by assuming that both the credit underwriting (4) and mortgage default (1) models are semiparametric models. It means that they have parametric and nonparametric components.

$$y_1^* = x_1^1\beta_1 + g_1(x_1^2) + \varepsilon_1 \quad (11)$$

$$y_2^* = x_2^1\beta_2 + g_2(x_2^2) + \varepsilon_2 \quad (12)$$

where $x_1 = (x_1^1, x_1^2) = (D_1, C_1, M_1)$ and $x_2 = (x_2^1, x_2^2) = (D_2, M_2)$ are vectors of exogenous independent variables. They include vectors of independent variables x_1^1 and x_2^1, that consist of parametric components β_1, β_2 and vectors of independent variables x_2^1 and x_2^2 that consist of nonparametric component $g_1(\cdot), g_2(\cdot)$, correspondingly. These variables are discussed further.

Thus, the conditional mean is:

$$E(y_1 \mid x_1, y_2^* > 0) = x_1^1\beta_1 + f(x_1^2) + \lambda(\varepsilon_1, \varepsilon_2, \hat{y}_2) \quad (13)$$

where $\hat{y}_2$ is fitted probability of endorsement (the selection probability or propensity score). The idea is to nonparametrically estimate the selection correction terms in (13), say via series approximation.

We do not restrict functional form both $g_1(\cdot), g_2(\cdot)$ and $\lambda(\cdot)$. In fact, they need to be parametrically specified. In principle $g_1(\cdot), g_2(\cdot)$ and $\lambda(\cdot)$ may be approximated with a polynomial in its scalar arguments.

A variety of specifications by employed semiprametric approach are used to ensure the

[14] The possible explanation of this fact is an endogenous nature of LTV that is discussed in detail in study (Ozhegov, Poroshina, 2013b).

robustness of our results. Default evaluation on the mortgage market using nonparametric techniques as baseline modeling tools are rarely applied (LaCour-Little, Maxam, 2001; LaCour-Little et al., 2002). Creel (2008) mentioned several reasons for that:

- Estimation results very often are too complicated and may not have obvious economic sense.
- Less restrictive distributional assumptions lead to the loss of efficiency.
- Applying semiparametric and nonparametric techniques lead to large computational costs.
- These techniques require tweaking that can influence the estimation results.

In the semiparametric approach, we experimented with more flexible forms that are presented in Tab. 4-5.

Tab. 4. Model specifications for the probability of endorsement

Semiparametric approach[15]	
(6)	$probability\ of\ endorsement = f_6(D_2, M_2)$, $D_2 = (linear\ and\ cross-products\ sociodemographics)$ $M_2 = (probability\ of\ application, probability\ of\ application^2, probability\ of\ application \times unemployment\ rate,$ $unemployment\ rate, unemployment\ rate^2)$
(7)	$probability\ of\ endorsement = f_7(D_2, M_2)$, $D_2 = (linear\ and\ cross-products\ sociodemographics)$ $M_2 = (unemployment\ rate, unemployment\ rate^2)$
(8)	$probability\ of\ endorsement = f_8(D_2, M_2)$, $D_2 = (linear\ and\ cross-products\ sociodemographics)$ $M_2 = (probability\ of\ application, probability\ of\ application^2)$
(9)	$probability\ of\ endorsement = f_9(D_2, M_2)$, $D_2 = (linear\ sociodemographics)$ $M_2 = (probability\ of\ application, probability\ of\ application^2, probability\ of\ application \times unemployment\ rate,$ $unemployment\ rate, unemployment\ rate^2)$
(10)	$probability\ of\ endorsement = f_{10}(D_2, M_2)$, $D_2 = (linear\ sociodemographics)$ $M_2 = (unemployment\ rate, unemployment\ rate^2)$
(11)	$probability\ of\ endorsement = f_{11}(D_2, M_2)$, $D_2 = (linear\ sociodemographics)$ $M_2 = (probability\ of\ application, probability\ of\ application^2)$
(12)	$probability\ of\ endorsement = f_{12}(D_2, M_2)$, $D_2 = (linear\ and\ cross-products\ sociodemographics)$ $M_2 = (probability\ of\ application, unemployment\ rate)$

Tab. 5. Model specifications for PD

Semiparametric approach[16]	
(6)	$PD = h_6(D_1, C_1, M_1, \hat{f}_4, \hat{f}_4^2, \hat{f}_4^3)$, $C_1 = (linear\ contract\ terms\ , LTV \times maturity, rate^2, rate^3)$
(7)	$PD = h_7(D_1, C_1, M_1, \hat{f}_4)$, $C_1 = (linear\ contract\ terms\ , LTV \times maturity, rate^2, rate^3)$
(8)	$PD = h_8(D_1, C_1, M_1, \hat{f}_4)$, $C_1 = (linear\ contract\ terms)$

[15] We estimated models for the probability of endorsement with semiparametric probability of application (probability of application squared) and unemployment rate (unemployment rate squared). Because probability of application squared was not statistically significant we do not report these results.

[16] D_1, M_1 consist of sociodemographic information and macroeconomic conditions that are linear and were discussed earlier.

(9)	$PD = h_9(D_1, C_1, M_1, \hat{f}_4, \hat{f}_4^2, \hat{f}_4^3),\ C_1 = (linear\ \ contract\ \ terms)$
(10)	$PD = h_{10}(D_1, C_1, M_1, \hat{f}_5, \hat{f}_5^2, \hat{f}_5^3),\ C_1 = (linear\ \ contract\ \ terms\ \ , LTV \times maturity, rate^2, rate^3)$
(11)	$PD = h_{11}(D_1, C_1, M_1, \hat{f}_5),\ C_1 = (linear\ \ contract\ \ terms\ \ , LTV \times maturity, rate^2, rate^3)$
(12)	$PD = h_{12}(D_1, C_1, M_1, \hat{f}_5),\ C_1 = (linear\ \ contract\ \ terms)$
(13)	$PD = h_{13}(D_1, C_1, M_1, \hat{f}_5, \hat{f}_5^2, \hat{f}_5^3),\ C_1 = (linear\ \ contract\ \ terms)$

We specify linear and quadratic polynomials for macrovariables (M_2) (first and second degree polynomial in the probability of application and unemployment rate, cross-products of them) and cross-products for sociodemographic characteristics (D_2)[17] in the probability of endorsement – Equation (12) (see Tab. 4, Tab. A1). The selection correction term $\lambda(\cdot)$ is approximated via first, second and third order polynomials in the fitted probability of endorsement $\hat{y}_2$. In the semiparametric estimation of PD (11) we use polynomial approximation with degrees 1, 2, 3 for some mortgage loan terms (C_1) (rate) and fitted probability of endorsement ($\hat{f}$), otherwise it would complicate our model considerably (Vella, 1998; Attanasio at al., 2008) (see Tab.5). Cross-products LTV and maturity, and cross-products of demographics are used to capture the joint effect of independent variables. Once the unknown functions of $g_1(\cdot), g_2(\cdot)$ and $\lambda(\cdot)$ is decided, the least squares estimation is applied (Attanasio et al., 2008). The large difference in the results of parametric and semiparametric estimations detects the misspecification problems of strictly parameterized models (Creel, 2008). We estimate also models with linear and semiparametric components of the above-mentioned factors. The results for the PD equation are reported in Tab. 6.

In alternative semiparametric specifications, signs of average marginal effects are almost the same and such factors sex, single borrowers, contract rate, loan age and average house price remain statistically significant. On the one hand, results based on specifications (6)-(7) and (10)-(11) with third and second degrees polynomials in the fitted probability of endorsement from (4) and (5) correspondingly, produced a similar pattern. However, model (10) (with semiparametric contract terms and semiparametric the fitted probability of endorsement) and model (11) (with semiparametric contract terms and linear the fitted probability of endorsement) have slightly higher percent of correct predictions – 94.4%. On the other hand, the percent of correct prediction in (8) is less than 0.4%, but it supports the results of parametric estimation about the significant negative correlation of credit underwriting and default processes. As evident from the comparison of alternative specifications, the selection bias correction based on the fitted

[17] Based on the results of goodness of fit test (Hosmer-Lemeshow test) and specification test (link test), in semiparametric estimation of PD with correction for sample selection we experimented with fitted probability of endorsement from (4) and (5) (see Tab.1). All above-mentioned socio-demographic independent variables remain the same.

probability of endorsement from (4) and (5) has a negligible effect that is not statistically significant.

The results of parametric and semiparametric estimations of PD (Tab. 3 and Ta6) are not remarkably different. Estimated parameters only slightly differ in absolute value as average marginal effects. In some semiparametric specifications we found less statistically significant borrowers with not declared family status or single, not declared income and unstable statistical significance of activity category "State employee" and other categories of borrower's income. The same evidence is found for mortgage loans with a maturity of less than 120 months or between 120-179 months.

Empirical results show that sample selection term does not have semiparametric statistical significance. However, comparing the percent of correct predictions suggests that BVP models have a high predictive power, but it is slightly high (approximately 0.1%). Overall, the results are robust to alternative model specifications, but they are conditional on data that are used. When the size of the sample is relatively small, nonparametric and semiparametric methods may lead to substantial parameter variation even if they are less biased compared to a biased, but incorrectly specified parametric model (Creel, 2008). The result of the specification test (link test) for parametric estimation of PD indicates that strictly parameterized models are specified correctly.

Tab. 6. Average marginal effects for probability of default (Semiparametric approach)

	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Age of borrower	2.58×10^{-4}	2.35×10^{-4}	1.07×10^{-4}	1.12×10^{-4}	-1.59×10^{-4}	-1.73×10^{-4}	5.42×10^{-4}	5.90×10^{-4}
	(0.37)	(0.34)	(1.43)	(1.49)	(-0.21)	(-0.23)	(0.67)	(0.73)
Sex	0.032***	0.032***	0.043***	0.043***	0.027***	0.027***	0.036***	0.036***
	(3.15)	(3.16)	(3.86)	(3.86)	(2.63)	(2.65)	(3.15)	(3.15)
Not declared educational level	-0.037	-0.033	-0.113*	-0.123*	0.002	0.005	-0.063	-0.073
	(-0.66)	(-0.61)	(-1.87)	(-1.96)	(0.03)	(0.09)	(-0.98)	(-1.10)
Secondary education	6.16×10^{-3}	6.28×10^{-3}	-8.06×10^{-3}	-8.44×10^{-3}	0.011	0.011	-8.61×10^{-3}	-0.001
	(0.27)	(0.28)	(-0.33)	(-0.34)	(0.48)	(0.48)	(-0.04)	(-0.05)
Complete higher education	0.018	0.018	0.035	0.035	0.004	0.004	0.018	0.018
	(0.71)	(0.71)	(1.31)	(1.31)	(0.15)	(0.17)	(0.63)	(0.63)
Not declared family status	0.118*	0.121*	0.136**	0.128*	0.104	0.106*	0.116*	0.109
	(1.85)	(1.91)	(2.01)	(1.89)	(1.63)	(1.67)	(1.70)	(1.60)
Single	0.021**	0.021**	0.024**	0.025**	0.020*	0.020*	0.022*	0.022*
	(2.02)	(2.02)	(2.12)	(2.13)	(1.83)	(1.83)	(1.88)	(1.90)
Widowed	0.015	0.015	-0.022	-0.024	0.031	0.031	-0.002	-0.004
	(0.30)	(0.32)	(-0.49)	(-0.54)	(0.61)	(0.62)	(-0.03)	(-0.09)
Divorced	0.015	0.015	0.006	0.006	0.019	0.019	0.010	0.010
	(1.23)	(1.23)	(0.44)	(0.45)	(1.47)	(1.47)	(0.76)	(0.77)
Hired employee	0.065	0.048	0.152**	0.194**	-6.91×10^{-4}	-0.015	0.068	0.110
	(0.91)	(0.73)	(1.99)	(2.34)	(-0.01)	(-0.20)	(0.80)	(1.27)
Entrepreneur	0.0984	0.0812	0.285***	0.328***	0.0162	0.001	0.180	0.223*
	(0.94)	(0.81)	(2.58)	(2.84)	(0.14)	(0.01)	(1.51)	(1.85)
State employee	0.0930	0.0763	0.222**	0.264***	0.005	-0.009	0.108	0.150
	(1.09)	(0.93)	(2.35)	(2.67)	(0.06)	(-0.09)	(1.02)	(1.41)
Not declared income of main borrower	-0.112*	-0.112*	-0.252***	-0.251***	-0.061	-0.062	-0.185**	-0.184**
	(-1.80)	(-1.79)	(-3.56)	(-3.57)	(-0.88)	(-0.88)	(-2.41)	(-2.42)
Income of main borrower 10000-19999	-0.069	-0.069	-0.109**	-0.109**	-0.065	-0.065	-0.104**	-0.104**
	(-1.44)	(-1.45)	(-2.18)	(-2.17)	(-1.35)	(-1.35)	(-2.06)	(-2.06)
Income of main borrower 20000-39999	-0.089*	-0.086*	-0.131***	-0.131***	-0.076	-0.076	-0.117**	-0.117**
	(-1.82)	(-1.81)	(-2.59)	(-2.60)	(-1.57)	(-1.56)	(-2.29)	(-2.30)
Income of main borrower 40000 and more	-0.042	-0.042	-0.093*	-0.094*	-0.025	-0.025	-0.071	-0.072
	(-0.83)	(-0.82)	(-1.72)	(-1.73)	(-0.49)	(-0.47)	(-1.29)	(-1.30)
Rate	0.074**	0.074**	0.008***	0.008***	0.073**	0.072**	0.007***	0.007***
	(2.26)	(2.26)	(6.25)	(6.25)	(2.20)	(2.20)	(5.97)	(5.95)
Maturity<120 months	0.0221	0.0224	0.0434*	0.0429*	0.0232	0.0236	0.0464**	0.0458**
	(0.91)	(0.93)	(1.87)	(1.85)	(0.96)	(0.98)	(2.01)	(1.99)
Maturity 120-179 months	0.0172	0.0177	0.0256*	0.0247*	0.0180	0.0185	0.0270**	0.0261**
	(1.21)	(1.24)	(1.93)	(1.86)	(1.26)	(1.30)	(2.04)	(1.98)
Maturity 180-239 months	0.016	0.016	0.002	0.002	0.017	0.017	0.003	0.003
	(1.37)	(1.38)	(0.19)	(0.15)	(1.45)	(1.47)	(0.28)	(0.26)
Maturity 240-299 months	0.014	0.014	0.009	0.009	0.015	0.015	0.010	0.001

	(1.26)	(1.29)	(0.82)	(0.79)	(1.34)	(1.37)	(0.93)	(0.89)
LTV<0.5	-0.002	-0.002	0.004	0.004	-0.002	-0.002	0.003	0.003
	(-0.07)	(-0.06)	(0.37)	(0.37)	(-0.08)	(-0.08)	(0.29)	(0.28)
Downpayment	1.21×10^{-9}	1.25×10^{-9}	2.66×10^{-9}	2.64×10^{-9}	1.21×10^{-9}	1.27×10^{-9}	2.83×10^{-9}	2.80×10^{-9}
	(0.13)	(0.13)	(0.25)	(0.25)	(0.13)	(0.14)	(0.27)	(0.26)
LTV×Maturity	-2.91×10^{-4}	-2.89×10^{-4}			-3.05×10^{-4}	-2.97×10^{-4}		
	(-0.14)	(-0.13)			(-0.14)	(-0.14)		
Rate squared	-0.015^{***}	-0.015^{***}			-0.015^{***}	-0.015^{***}		
	(-2.87)	(-2.88)			(-2.85)	(-2.85)		
Rate cubed	$7.70\times10^{-4***}$	$7.70\times10^{-4***}$			$7.68\times10^{-4***}$	$7.68\times10^{-4***}$		
	(3.51)	(3.52)			(3.50)	(3.51)		
Duration	$8.13\times10^{-5***}$	$8.12\times10^{-5***}$	$9.30\times10^{-5***}$	$9.30\times10^{-5***}$	$8.98\times10^{-5***}$	$8.95\times10^{-5***}$	$1.08\times10^{-4***}$	$1.08\times10^{-4***}$
	(4.74)	(4.71)	(5.33)	(5.36)	(5.01)	(4.97)	(6.32)	(6.34)
Mean m2 value	$5.32\times10^{-6***}$	$5.32\times10^{-6***}$	$1.07\times10^{-5***}$	$1.07\times10^{-5***}$	$4.88\times10^{-6***}$	$4.87\times10^{-6***}$	$1.03\times10^{-5***}$	$1.03\times10^{-5***}$
	(4.28)	(4.28)	(10.09)	(10.07)	(3.92)	(3.91)	(9.97)	(9.97)
Fitted probability of endorsement 4	-9.362	-0.241	-0.816^{***}	-19.22				
	(-0.49)	(-0.93)	(-2.79)	(-0.99)				
Fitted probability of endorsement 4 squared	6.358			12.66				
	(0.47)			(0.93)				
Fitted probability of endorsement 4 cubed	-1.473			-2.896				
	(-0.46)			(-0.90)				
Fitted probability of endorsement 5					-10.65	0.066	-0.407	-19.94
					(-0.54)	(0.21)	(-1.17)	(-0.99)
Fitted probability of endorsement 5 squared					7.574			13.49
					(0.54)			(0.94)
Fitted probability of endorsement 5 cubed					-1.778			-3.098
					(-0.53)			(-0.91)
Observations	2728	2728	2728	2728	2728	2728	2728	2728
R^2	0.285	0.284	0.199	0.199	0.284	0.284	0.196	0.197
Adjusted R^2	0.276	0.277	0.191	0.191	0.276	0.277	0.188	0.188
AIC	-933.2	-934.7	-629.7	-627.5	-930.3	-933.8	-621.1	-619.1
BIC	-749.9	-757.4	-464.2	-450.2	-741.1	-756.4	-455.6	-441.7
Log likelihood	497.6	497.4	342.8	343.8	497.1	496.9	338.6	339.5
% Right predictions	94.3	94.3	94.0	94.0	94.4	94.4	94.0	94.0

Note: t statistics in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Categories that were taken as base outcomes: education level “Incomplete higher education”; family status “Married”; activity category “Unemployed”; income level 0-9999; maturity >=300 months; LTV 0.5-0.7.

Conclusion

To summarize our results, we have estimated a model of mortgage default by using unique micro-level data on contracts from a regional branch of the Agency of Home Mortgage Lending (2008-2012) and macro-level data about mortgage market and labor market performance. The level of our data allows us to consider loan terms and borrower risk factors. To control for the possible correlation of some components in credit underwriting and default processes, we employed a parametric two-step estimation approach that corrected for sample selection bias. The results confirm significant negative correlation between these mortgage lending processes and show that risky applicants have a lesser chance of being approved for a mortgage loan.

Then to confirm the robustness of obtained results, we relax the parametric assumptions and employ a semiparametric estimation procedure by approximating the true joint density with polynomial both on some exogenous variables in the probability of endorsement and probability of default equations and selection correction term. The comparative analysis of results derived from different estimation approaches suggests that semiparametric models of probability of default are not substantially better than the fully parameterized ones. Both of them have relatively high predictive power and estimated parameters and average marginal effects are not remarkably different. However, in terms of empirical results, we find that sample selection term does not remain statistical significance.

Empirical results suggest that borrower and mortgage loan characteristics affect loan performance and play an important role in predicting default as well as macrovariables.

Borrower's income has a great impact on probability of default because it is the main source of making mortgage payments. However, a large proportion of borrowers in our sample did not declare income for different unobserved reasons. The fact that they have the lowest probability of default is connected with the difference between declared and real income. Credit organizations require an official confirmation of applicant's income. However, real income is not always can be officially confirmed. In contrast, lower income borrowers with insufficient or unstable income had the highest probabilities of default. That is why they face problems meeting their mortgage obligations. But the relationship between probability of default and different income levels on the residential mortgage market requires further investigation.

Males, single borrowers (or with not declared), and state employees had a higher probability of default. This fact is mostly explained by higher potential risks of these borrowers that are caused by different reasons. Most of them are connected with risks of divorce, suffering from illness or losing their job. Additionally, we did not find any significant evidence that the

education level of a borrower, used as a proxy for a borrower's financial literacy, influences on probability of default.

Moreover, the contract rate and shocks of regional house prices are indeed important drivers of mortgage default, positively connected with number of days observed in mortgage loans. Mortgages with low and high Loan-to-Value ratio are attractive for non-liquid or risky borrowers that have a high probability of default, but this effect is not statistically significant.

Finally, the findings in this paper have practical implementations. In particular, they can be used to develop effective risk management systems in credit organizations under Internal Rating Based (IRB) - systems for credit risk evaluation recommended by regulators.

The framework of this paper is based on the assumption of exogenous nature of explanatory variables because semiparametric estimation with correction for selection and endogeneity issues is rather challenging. In addition, the collected data set suffers from a lack of credit history data, quality of service of AHML and other credit organizations, and low variation in data. Further research should attempt to avoid these challenges and employ more flexible estimation techniques such as nonparametric ones. Despite all advantages of semiparametric models, however, parametric ones are computationally unchallenging, easily interpreted and more efficiency in case of appropriate model specification. For these reasons, semiparametric procedures should be regarded not as a substitution of parametric ones, but as complementation to them (Creel, 2008).

References:

AHML: Agency of Home Mortgage Lending (2009). Godovoy Otchet [Annual Report] // AHML.

Archer, Wayne R., David C. Ling, and Gary A. McGill (1996). The Effect of Income and Collateral Constraints on Residential Mortgage Terminations. // Regional Science and Urban Economics – Vol. 26, No. 3/4, P. 235–261.

Archer, Wayne R., David C. Ling, and Gary A. McGill (1997). Demographic Versus Option-Driven Mortgage Terminations. // Journal of Housing Economics – Vol. 6, No. 2, P. 137–163.

Attanasio, P. Orazio, Goldberg, Pinelopi Koujianou, and Kyriazidou, Ekaterini (2008). Credit Constraints in the Market for Consumer Durables: Evidence from Micro Data on Car Loans. // International Economic Review – Vol. 49, No. 2, P. 401–436.

Bajari, Patrick, Chu, Chenghuan S., and Park, Minjung (2008). An Empirical Model of Subprime Mortgage Default from 2000 to 2007. // NBER Working paper 14625.

Bank of Russia (2012). Pis'mo Banka Rossii ot 29.12.2012 N 192-T «O Metodicheskikh Rekomendatsiyakh po Realizatsii Podkhoda k Raschetu Kreditnogo Riska na Osnove Vnutrennikh Reytingov Bankov» [Letter of Bank of Russia from 29.12.2012 N 192-T «The guidance on implementation of the approach to the calculation of credit risk based on internal ratings of banks»] // Bank of Russia.

BIS: Basel Committee on Banking Supervision. Basel II: International Convergence of Capital Measurement and Capital Standards: a Revised Framework (2006) // Bank for International Settelments document.

Campbell, John Y., and Cocco, Joao F. (2011). A Model of Mortgage Default. // NBER Working Paper 17516.

Campbell, John Y. (2012). Mortgage Market Design. // Review of Finance – Vol. 17, P. 1–33.

Clapp, John M., Yongheng Deng, and Xudong An. (2006). Unobserved heterogeneity in models of competing mortgage termination risks. // Real Estate Economics – Vol. 34(2). P. 243–273.

Corbae, Dean, and Quintin, Erwan (2010). Mortgage Innovation and the Foreclosure Boom. // Working Paper, University of Texas and University of Wisconsin.

Creel, Michel (2008). Some Possible Pitfalls of Parametric Inference. // Quantile – No. 4, P. 1–6.

Cutts, Ammy C., and Merrill, William (2008). Interventions in Mortgage Default: Policies and Practices to Prevent Home Loss and Lower Costs. // Freddie Mac Working Paper 08-01.

Dell'Ariccia Giovanni, Igan, Deniz, and Laeven, Luc. (2012). Credit Booms and Lending Standards: Evidence from the Subprime Mortgage Market. // Journal of Money, Credit and Banking –Vol. 44, No. 2–3, P. 367–384.

Dell'Ariccia, Giovanni, Deniz Igan, and Luc Laeven (2012). Credit Booms and Lending Standards: Evidence from the Subprime Mortgage Market. // Journal of Money, Credit and Banking – Vol. 44. No. 2–3. P. 367–384.

Demyanyk, Yuliya, and Otto Van Hemert (2011). Understanding the subprime mortgage crisis. // Review of Financial Studies – Vol. 24 (6). P. 1848–1880.

Deng, Yongheng, John M. Quigley, and Robert Order. (2000). Mortgage terminations, heterogeneity and the exercise of mortgage options. // Econometrica – Vol. 68. P. 275–307.

Deng, Yongheng, Andrey D. Pavlov, and Lihong Yang (2005). Spatial Heterogeneity in Mortgage Terminations by Refinance, Sale and Default. // Real Estate Economics – Vol. 33(4). P. 739–764.

Follain, James. R. (1990). Mortgage Choice // AREUEA Journal – Vol. 18, No. 2, P. 125–144.

Goldberg, Gerson M., and John P. Harding (2003). Investment Characteristics of Low- and Moderate-income Mortgage Loans. // Journal of Housing Economics – Vol. 12(3). P. 151–180.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales (2013). The Determinants of Attitudes towards Strategic Default Mortgages. // The Journal of Finance – Vol. 68, No. 4, P. 1473–1515.

Heckman, James (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 – Vol. 5, No. 4, P. 475–492.

Heckman, James (1979). Sample Selection Bias as a Specification Error. // Econometrica – Vol.47, No. 1, P. 153–161.

Greene, William, H. (2003). Econometric Analysis, Fifth Edition. Upper Saddle River, NJ: Prentice Hall.

Keys, Benjamin J., Mukherjee, T., Seru, A., and Vig, V. (2010). Did securitization lead to lax screening? Evidence from subprime loans. // Quarterly Journal of Economics – Vol. 125. P. 307–362.

LaCour-Little, Michael, and Maxam, Clark L. (2001). Applied Nonparametric Regression Techniques: Estimating Prepayments on Fixed-Rate Mortgage-Backed Securities. // Journal of Real Estate Finance and Economics – Vol. 23, No. 2, P. 139–160.

LaCour-Little, Michael, Marshoun, Michael, and Maxam, Clark L. (2002). Improving Parametric Mortgage Prepayment Models with Non-parametric Kernel Regression. // Journal of Real Estate Research – Vol. 24, No. 3, P. 299–327.

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

Maddala, Gangadharrao S., and Trost, Robert P. (1982). On Measuring Discrimination in Loan Markets. // Housing Finance Review – Vol. 1, No. 3, P. 245–266.

Mayer, Christopher, Karen Pence and Shane Sherlund (2009). The Rise in Mortgage Defaults. // Journal of Economic Perspectives – Vol. 23, No. 1, P. 27–50.

Mian, Atif, and Amir Sufi (2009). The Consequences of Mortgage Credit Expansion: Evidence from the U.S. Mortgage Default Crisis. // Quarterly Journal of Economics – Vol. 124. P. 1449–96.

Moody's MILAN Methodology for Rating Russian RMBS (2008) // Moody's.

Ozhegov, Evgeniy, and Poroshina, Agatha (2013a). The Lagged Structure of Dynamic Demand Function for Mortgage Loans in Russia. // Journal of Corporate Finance – Vol. 3, No. 27, P. 37–49.

Ozhegov, Evgeniy, and Poroshina, Agatha (2014). Otsenka Kreditnogo Riska pri Iipotechnom Zhilishchnom Kreditovanii [Credit Risk Evaluation in the Mortgage Residential Lending]. // Applied Econometrics (forthcoming).

Ozhegov, Evgeniy, and Poroshina, Agatha (2013b). Bank Risk Preferences on the Government-Insured Mortgage Market. // Working paper, Higher School of Economics, Group for Applied Markets and Enterprises Studies.

Pavlov, Andrey D. (2001). Competing Risks of Mortgage Termination: Who Refinances, Who Moves, and Who Defaults? // Journal of Real Estate Finance and Economics – Vol. 23(2). P. 185–211.

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 – Vol. 11, No. 1, P. 87–102.

Piskorski, Tomasz, and Tchistyi, Alexei (2010). Optimal Mortgage Design // Review of Financial Studies – Vol. 23, No. 8, P. 3098–3140.

Polterovich, Victor M., and Starkov, Oleg, U. (2007). Strategiya Formirovaniya Ipotechnogo Rynka v Rossii [Strategy of Formation of the Mortgage Market in Russia]. // Economics and Mathematical Methods – Vol. 43, No. 4, P. 3–22.

Rachlis, Mitchell B., and Yezer, Anthony M. J. (1993). Serious Flaws in Statistical Tests for Discrimination in Mortgage Markets. // Journal of Housing Research – Vol. 4, P. 315–336.

Ross, Stephen. L. (2000). Mortgage Lending, Sample Selection and Default. // Real Estate Economics – Vol. 28, No. 4, P. 581–621.

Sternik, Gennadij M. (2009). Spad na Rynke Stroitel'stva i Prodazhi Zhil'ya v Rossii [Downturn in the Construction Market and Home Sales to Russia]. // Journal of New Economic Association – Vol. 3-4, P. 185–207.

Stolbov, Michael (2012). Teoriya Finansovogo Akseleratora i Rossiyskiy Ipotechnyy Rynok [Financial Accelerator Theory and the Russian Mortgage Market]. // Journal of New Economic Association – Vol. 1, No. 13, P. 79–98.

Vandell, Kerry D. (1995). How Ruthless Is Mortgage Default? A Review and Synthesis of the Evidence. // Journal of Housing Research – Vol. 6, No. 2, P. 245–264.

Vella, Francis (1998). Estimating Models with Sample Selection Bias: a Survey. // Journal of Human Resources – Vol. 33, No.1, P. 127– 169.

Yezer, Anthony, Philips, Robert F., and Trost, Robert P. (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 – Vol. 9, No. 3, P. 197–215.

Appendix

Tab. A1. Summary statistics[18]

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)	-	-	-	-
Sociodemogrpahic characteristics (4298 applicants)					
Age of borrower	Age of borrower, years	34	7.6	21	61
Age squared	Age of borrower squared, years	-	-	-	-
Age cubed	Age of borrower cubed, years	-	-	-	-
Sex	Sex, =1 male	-	-	-	-
Declared income of main borrower	Monthly income of main borrower, Russian rub.	30 663.6	26 203.2	1 658.7	38 5531.4
Declared income of co-borrowers	Sum of monthly co-borrowers main income, Russian rub.	17 654.3	11 555.9	38.3	72 800.5
Sex×Family status	Product of sex and family status	-	-	-	-
Sex×Activity category	Product of sex and activity category	-	-	-	-
Sex×Education level	Product of sex and education level	-	-	-	-
Sex×Income	Product of sex and monthly income of main borrower	-	-	-	-
Family status×Activity category	Product of family status and activity category	-	-	-	-
Family status×Education level	Product of family status and education level	-	-	-	-
Family status×Income	Product of family status and monthly income of main borrower	-	-	-	-
Education level×Income	Product of education level and monthly income of main borrower	-	-	-	-
Education level× Activity category	Product of education level and activity category	-	-	-	-
Income×Activity category	Product of monthly income of main borrower and activity category activity category	-	-	-	-
Co-borrower income×Sex	Product of monthly income of co-borrowers and sex	-	-	-	-
Co-borrower income×Family status	Product of monthly income of co-borrowers and family status	-	-	-	-
Co-borrower income× Activity category	Product of monthly income of co-borrowers and activity category	-	-	-	-
Co-borrower income× Education level	Product of monthly income of co-borrowers and education level	-	-	-	-
Co-borrower income× Income	Product of monthly income of co-borrowers and monthly income of main borrower	-	-	-	-
Terms of credit contract (2799 contracts)					
Loan limit	Maximum loan limit, Russian rub.	1 087 933	616 643.1	120 000	12 700 000
Loan amount	Loan amount, Russian rub.	1 040 037	573 665.9	120 000	10 000 000
Rate	Contract rate (when fixed), %	11.59	1.64	9.55	19
Rate squared	Contract rate squared (when fixed), %	-	-	-	-
Rate cubed	Contract rate cubed (when fixed), %	-	-	-	-
Type of rate	Type of rate, =1 adjusted	-	-	-	-
Maturity	Maturity of credit, months	189.05	62.17	26	360
Downpayment	Downpayment, Russian rub.	854 962.3	706 635.4	40 000	13 820 000
Flat value	Assessed value, Russian rub.	1 894 999	1 049 502	330 000	15 290 000
Monthly payment	Monthly payment, Russian rub.	12 610.96	7 324.47	1 872.44	14 0381
LTV	Loan-to-value ratio	0.56	0.17	0.11	0.94
DTI	Debt-to-income ratio (for declared income)	0.45	0.18	0.06	1
LTV×Matutirty	Cross product of categorical LTV on categorical maturity	-	-	-	-
Duration	Total amount of days observed in credit, days	867.2	419.7	18	1 487

[18] Cross products of sociodemographic characteristics, Unemployment rate × Probability of application, LTV × maturity such as rate, probability of application, unemployment rate squared and cubed are used in the semiparametric estimation of corresponding equations for the probability of endorsement and the probability of default.

Macrovariables (50 months)					
Unemployment rate	Quarterly regional unemployment rate, %	8.4	1.5	6.3	10.9
Unemployment rate squared	Quarterly regional unemployment rate squared, %	-	-	-	-
Probability of application *1000	The probability of application on aggregated data (number of applications in month t divided by the amount of households), %	38.4	16.5	16.3	83.7
Probability of application squared	Probability of application squared	-	-	-	-
Unemployment rate × Probability of application	Product of unemployment rate and probability of application	-	-	-	-
Mean m2 value	Average price for 1 square meters in region, Russian rub.	38 622.8	6 165.8	28 782	51 304

Tab. A2. Summary of categorical variables[19]

Variables	Total	%
Sociodemogrpahic characteristics (4298 applicants)		
Sex		
male	1879	43.7
female	2419	56.3
Family status		
not declared	46	1.1
single	1220	28.4
married	2358	54.9
widowed	56	1.3
divorced	618	14.4
Activity category		
not declared	138	3.2
unemployed	1	0.0
soldier	13	0.3
hired employee	3963	92.2
entrepreneur	39	0.9
state employee	144	3.4
Education level		
not declared	205	4.8
elementary education	65	1.5
secondary education	1748	40.7
incomplete higher education	138	3.2
higher education	2142	49.8
Monthly income of main borrower (Russian rub.)		
not declared	2918	67.9
0-9999	118	2.8
10000-19999	376	8.8
20000-39999	597	13.9
>=40000	289	6.7
Monthly income of co-borrowers (Russian rub.)		
not declared	3724	86.6
0-9999	159	3.7
10000-19999	225	5.2
>=20000	190	4.4
Terms of credit contract[20] (2799 contracts)		
Type of rate		
fixed rate	2421	86.5
adjusted rate	378	13.5
Maturity		
< 120 months	181	6.5
120-179 months	595	21.3
180-239 months	1106	39.5
240-299 months	690	24.6
>=300 months	227	8.1
LTV		
<0.5	968	34.6
0.5-0.7	1531	54.7
>=0.7	300	10.7
DTI		
not declared	1651	59.0
<0.2	41	1.5
0.2-0.4	505	18.0
0.4-0.6	379	13.5
0.6-0.8	160	5.7
>=0.8	63	2.3

[19] In the estimation of models categorical variables were transformed into a set of dummy variables.
[20] Type of rate, maturity, LTV and DTI are available only for issued mortgages.

Tab. A3. Tests on discriminating power for continuous variables

Variables	p-value for Bartlett's test	p-value for t-test[21]	p-value for ANOVA-test	p-value for the Wilcoxon-Mann-Whitney test
Approved/rejected applicants				
Age of borrower	0.000***	0.0009***	0.0021***	0.0163**
Declared income of main borrower	0.983	0.0132**	0.0132**	0.0014***
Declared income of co-borrowers	0.020**	0.0086***	0.0086***	0.0323**
Unemployment rate	0.006***	0.9028	0.8968	0.3654
*Probability of application*1000*	0.384	0.0689*	0.0689*	0.0608*
Mean m2 value	0.007***	0.0118**	0.0177**	0.0726*
Defaulted/non-defaulted borrowers				
Age of borrower	0.926	0.0002***	0.0002***	0.0001***
Declared income of main borrower	0.000***	0.3777	0.1680	0.1258
Declared income of co-borrowers	0.884	0.2544	0.2544	0.1089
Loan limit	0.000***	0.0247**	0.0004***	0.0000***
Loan amount	0.000***	0.0647*	0.0013***	0.0000***
Rate	0.000***	0.0000***	0.0000***	0.0000***
Maturity	0.016**	0.0000***	0.0000***	0.0000***
Downpayment	0.000***	0.5972	0.4188	0.0000***
Flat value	0.000***	0.1720	0.0214**	0.0000***
Monthly payment	0.000***	0.8053	0.6145	0.0001***
LTV	0.000***	0.6558	0.6558	0.8306
DTI	0.591	0.9232	0.9232	0.6728
Duration	0.000***	0.0000***	0.0000***	0.0000***
Unemployment rate	0.000***	0.000**	0.000***	0.0000***
*Probability of application*1000*	0.000***	0.0127**	0.0935*	0.0185**
Mean m2 value	0.003***	0.000***	0.0000***	0.0000***

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

Tab. A4. Tests on discriminating power for categorical variables

Variables	p-value for the Chi-square test	p-value the Fisher's exact test
Approved/rejected applicants		
Sex	0.677	-
Activity category	-	0.000***
Education level	0.000***	-
Family status	-	0.508
Monthly income of main borrower (categorical)	-	0.000***
Monthly income of co-borrowers (categorical)	-	0.000***
Defaulted/non-defaulted borrowers		
Sex	0.031**	-
Activity category	-	0.000***
Education level	-	0.001***
Family status	-	0.004***
Monthly income of main borrower (categorical)	0.000***	-
Monthly income of co-borrowers (categorical)	0.000***	-
Maturity (categorical)	-	0.000***
LTV (categorical)	0.000***	-
DTI (categorical)	0.000***	-
Type of rate	-	0.000***

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

[21] When the assumption of equal variances between groups is violated, t-test for samples with unequal variances has to be used. To test this assumption we performed Bartlett's test and in case when p-value less then critical level (the second column), we reported t-test for samples with unequal variances.

Tab. A5. Spearman's Correlation matrix

	Flag of endorsement	Flag of default
Flag of endorsement	1.00	-
Flag of default	-	1.00
Borrower age	0.04**	0.07**
Sex	0.01	0.06**
Activity category	0.17***	0.02
Education level	0.16***	-0.03
Familystatus	-0.02	-0.03
Monthly income of main borrower (categorical)	0.22***	-0.03
Loan limit	-	-0.11***
Loan amount	-	-0.10***
Rate	-	0.41***
Maturity	-	-0.09***
Maturity (categorical)	-	-0.08***
Monthly payment	-	-0.05
Downpayment	-	-0.09***
Flat value	-	-0.11***
LTV	-	0.01
LTV (categorical)	-	0.03
DTI	-	0.01
DTI (categorical)	-	0.02
Duration	-	0.32***
Unemployment rate	0.01	0.13***
Probability of application*1000	-0.03*	0.05**
Mean m2 value	0.03*	0.22***

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

Tab. A6. Correlation matrix for independent variables[22]

	Borrower age	Sex	Activity category	Education level	Family status	Monthly income of main borrower (categorical)	Loan limit	Loan amount	Rate	Maturity (categorical)	Monthly payment	Down-payment	Flat value
Borrower age	1.00												
Sex	-0.01***	1.00											
Activity category	0.01	-0.04	1.00										
Education level	-0.07**	-0.12***	0.02	1.00									
Familystatus	0.23***	0.06**	0.00	-0.10***	1.00								
Monthly income of main borrower (categorical)	0.10***	0.07**	0.06**	0.24***	0.00	1.00							
Loan limit	0.02	0.05***	-0.02	0.16***	0.00	0.01	1.00						
Loan amount	0.02	0.03	-0.04*	0.18***	0.00	0.03*	**0.90**	1.00					
Rate	0.06***	-0.03*	-0.06***	-0.05***	-0.00	0.25***	-0.18***	-0.16***	1.00				
Maturity (categorical)	-0.27***	0.04	0.02	-0.03	-0.04	-0.04	0.21***	0.22***	-0.04	1.00			
Monthly payment	0.12***	0.03	-0.03*	0.18***	0.01	0.11***	**0.84*** **	**0.93*** **	-0.08***	-0.06***	1.00		
Downpayment	0.11***	0.02	0.01	0.10***	0.03*	0.03	0.31 ***	0.34***	-0.14***	-0.08***	0.35***	1.00	
Flat value	0.09***	0.03	-0.01	0.17***	0.03	0.04*	**0.70*** **	**0.77*** **	-0.18***	0.05***	**0.75*** **	**0.86*** **	1.00
LTV	-0.12***	0.00	-0.01	0.02	-0.04**	0.07***	0.29***	0.33***	0.07***	0.22***	0.28***	**-0.60*** **	-0.22***
LTV (categorical)	-0.09***	0.01	-0.00	0.00	-0.04	0.22***	0.31***	0.31***	0.26***	0.18***	0.32***	**-0.72*** **	-0.27***
DTI	-0.01	-0.07**	0.06**	-0.05*	0.02	-0.43***	0.12***	0.13***	0.02	0.02	0.12***	-0.03	0.05
DTI (categorical)	-0.02	-0.07**	-0.02	-0.07**	0.06**	-0.39***	0.15***	0.17***	0.11***	-0.00	0.19***	-0.06*	0.06*
Maturity	-0.29***	-0.00	0.01	-0.02	-0.03	-0.11***	0.15***	0.20***	-0.04*	**0.98*** **	-0.06***	-0.08***	0.05***
Duration	0.07***	0.01	-0.08***	-0.02	-0.05**	0.39***	-0.22***	-0.22***	**0.69*** **	-0.09***	-0.13***	-0.09***	-0.18***
Type of rate	-0.10***	0.01	0.02	0.01*	0.05	0.03***	0.14***	0.22***	-0.41***	-0.01***	0.21***	0.18***	0.25***
Unemployment rate	-0.00	0.05	-0.03	-0.04	0.02	-0.11 ***	-0.19***	-0.18***	0.48***		-0.12***	-0.08***	-0.15***
Probability of application*1000	-0.04***	0.02	-0.00	0.00	-0.06**	0.02	-0.08***	-0.08***	-0.07***	-0.08***	-0.08***	0.00	-0.04**
Mean m2 value	0.09***	-0.00	-0.01	-0.10***	0.02	-0.08***	0.03*	0.03	-0.11***	-0.06***	0.07***	0.02	0.03

Note: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

[22] For categorical variables, Spearman rank correlations are calculated.

Tab. A7. Correlation matrix for independent variables (continued)

	LTV	LTV (categorical)	DTI	DTI (categorical)	Maturity	Duration	Type of rate	Unemployment rate	Probability of application* 1000	Mean m2 value
Borrower age										
Sex										
Activity category										
Education level										
Family status										
Monthly income of main borrower (categorical)										
Loan limit										
Loan amount										
Rate										
Maturity (categorical)										
Monthly payment										
Downpayment										
Flat value										
LTV	1.00									
LTV (categorical)	**0.92*** **	1.00								
DTI	0.13***	0.10***	1.00							
DTI (categorical)	0.16***	0.14***	**0.94*** **	1.00						
Maturity	0.22***	0.17***	0.04	-0.09***	1.00					
Duration	-0.02	0.03	-0.04	0.41***	-0.11***	1.00				
Type of rate	-0.02	-0.04	0.01	0.00	-0.02	-0.41***	1.00			
Unemployment rate	-0.03	-0.00	-0.04	0.00	-0.09***	**0.80*** **	-0.41***	1.00		
Probability of application*1000	-0.05***	-0.10***	-0.10***	-0.10***	-0.08***	0.31***	-0.02	0.42***	1.00	
Mean m2 value	0.03	0.08***	0.09***	0.12***	-0.05***	-0.11***	-0.19***	-0.57***	-0.34***	1.00

Note: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

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. Lecturer, Junior Research Assistant. Perm, Russia.
E-mail: aporoshina@hse.ru, amporoshina@gmail.com Tel. +7 (342) 200-95-52