NATIONAL RESEARCH UNIVERSITY

Sample Selection Bias in the Mortgage Market Credit Risk Modeling

Agatha Poroshina

Department of Applied Mathematics and Modeling in Social Systems,
Lab of Investment Analysis

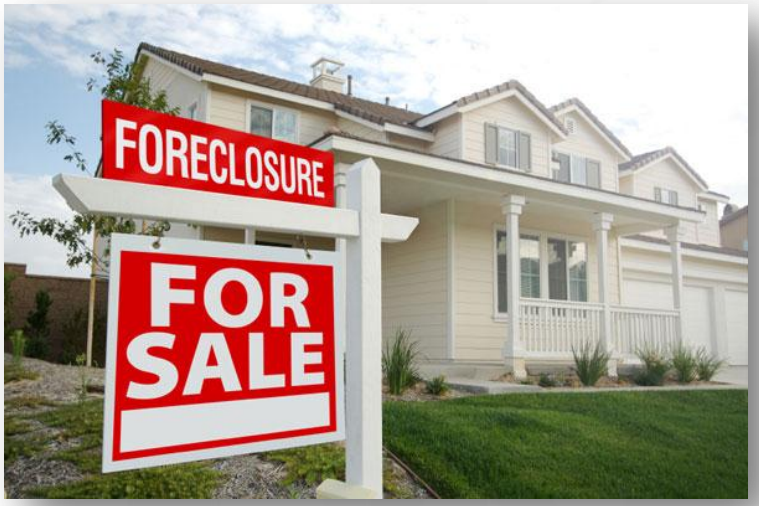
Perm Winter School, February 5th 2013

Outline

1. Motivation
2. Stylized Facts
3. Sample Selection Bias
4. Literature Review
5. Research Questions
6. Analytical Approach
7. Data Description
8. Conclusions

This study has been 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"

Motivation of Research



The impact of financial crisis on a sub-prime mortgage market

Almost 10 % mortgages were delinquent (USA)

(Querica, 2011, Center of Community Capital)

The shortcomings of credit risk techniques

Stylized Facts

1. Probability of default (PD).
2. Default – 90 days delinquent.
3. The absence of the concept of ‘mortgage default’ in Russian legislation.
4. Default drivers:
   - sociodemographic information
   - terms of mortgage
   - mortgage performance
   - macroeconomic conditions

5. Classical binary choice models (Bhutta, Dokko, Shan, 2010, Federal Reserve Board): mortgage default estimates are subject to sample selection bias.
6. Mortgage default estimates could be biased and inconsistent.
7. Sample selection bias due to:
   - simultaneity bias (not considering the underwriting process)
   - trunction (partial observability)
8. A size of bias depends on the degree of correlation between the default process and the credit underwriting process.

Accepted/Rejected applicants	Lender's decision - the underwriting process	Borrower's default process

Sample Selection Bias

The Heckman Model (1976, 1979)

$$y_i = x_i^{/}\beta + \varepsilon_i$$

Assume that y_i is observed only when unobserved latent z_i^* variable exceeds a particular threshold:

Outcome equation

$$\begin{cases} y_i = x_i^{/}\beta + \varepsilon_i, z_i^* > 0, \\ y_i = unobserved, otherwise \end{cases}$$

$\varepsilon_i \sim N(0, \sigma^2)$

Selection equation

$$z_i^* = w_i^{/}\alpha + u_i$$

$$z_i = \begin{cases} 1, if \;\; z_i^* > 0 \\ 0, otherwise \end{cases}$$

$u_i \sim N(0,1)$

$corr(\varepsilon_i, u_i) = \rho_{\varepsilon u}$

$$E(y_i \mid y_i \text{ is observed}) = E(y_i \mid z_i^* > 0) =$$
$$= x_i'\beta + E[\varepsilon_i \mid u_i > -w_i'\alpha] = x_i'\beta + \rho_{\varepsilon u}\sigma_\varepsilon \lambda_i(w_i'\alpha)$$

Heckman's λ (Inverse Mills ratio)

$$\lambda_i(w_i'\alpha) = \frac{\varphi(w_i'\alpha)}{\Phi(w_i'\alpha)}$$

The coefficient on the λ indicate if there is sample selection bias

1) Heckman's two-step procedure
2) MLE version

Bivariate Probit Model with Selection

The bivariate probit model with selection

The classic bivariate probit model

$$y_1^* = x_1\beta_1 + \varepsilon_1$$

$$y_2^* = x_2\beta_2 + \varepsilon_2$$

$$y_1 = \begin{cases} 1, if \;\; y_1^* > 0, \\ 0, if \;\; y_1^* \le 0. \end{cases}$$

Outcome equation (default-pay on time)

$$y_2 = \begin{cases} 1, if \;\; y_2^* > 0, \\ 0, if \;\; y_2^* \le 0. \end{cases}$$

Selection equation (accept-reject)

$$E(\varepsilon_1) = E(\varepsilon_2) = 0, \;\; corr(\varepsilon_1, \varepsilon_2) = \rho$$

$$\varepsilon_1, \varepsilon_2 \; are \; \phi(0,0,1,1,\rho) standard \; bivariate \; normal \; distribution$$

$$y_1^* \; is \; observed \; only \; if \;\; y_2^* = 1$$

$$y_2^* \; is \; observed \;\; for \;\; all \;\; classes$$

	Defaulted	*Non-defaulted*	*Total*
Accepted	Observed	Observed	Observed
Rejected	Not observed	Not observed	Observed

MLE and Heckman's λ

Literature Review

Rachils, Yezer (1993, Journal of Housing Research)

- Unbiased tests for discrimination require multiple-equations models (correction for sample selection bias)
- 4 decisions: the selection of the originator, the application for a particular mortgage product (loan-to-value ratio, maturity), lender's decision to approve/reject application, borrower's decision to repay/default

Phillips, Yezer, Trost (1994, 1996, Journal of Real Estate Finance and Economics)

- Lenders from Boston (The Home Mortgage Disclosure Act data (HMDA)), The Boston Fed data
- Modeling processes of the credit underwriting and default separately leads to the biased parameter estimates
- Provide empirical evaluations of the endogenity mortgage terms (Ross, Yinger, 1999)

Ross (2000, Journal of Real Estate Economics):

- Boston Federal Reserve data, Federal Housing Authority (FHA) foreclosure data
- Bivariate probit with selection (the probability of denial, PD) fits data in the best way
- The use of more borrower characteristics including credit history and others risk factors will directly minimize concerns about sample selection bias

Bajari, Chu, Park (2008, National Bureau of Economic Research):

- LoanPerformance data (USA), 2000-Census data, Bureau of Labor and Statistics
- Bivariate probit model with selection (ability and willingness to pay, PD) gives better parameter estimates than the univariate probit
- Key default drivers are borrower and loan characteristics
- The nationwide decrease in home prices as the deterioration are important driver behind the recent surge in defaults

Research Questions

1. What are key drivers of a borrower's default?
2. What is the impact of sample selection bias on the default estimates on **the Perm mortgage market**?
3. What would be the effect of correction for sample selection bias?

Analytical Approach

C - Client

B - Bank

Lend money

Repay

Do not repay

C

B

Do not lend money

Go

C

Do not go

PD (y_1)

The probability of approval (y_2)

$$y_1^* = x_1\beta_1 + \varepsilon_1$$

$$y_2^* = x_2\beta_2 + \varepsilon_2$$

$$y_1 = \begin{cases} 1, if \;\; y_1^* > 0, \\ 0, if \;\; y_1^* \le 0. \end{cases}$$

$$y_2 = \begin{cases} 1, if \;\; y_2^* > 0, \\ 0, if \;\; y_2^* \le 0. \end{cases}$$

y_1^* *is observed only if* $y_2^* = 1$

y_2^* *is observed for all classes*

Data Description

Perm Mortgage Company

Total sample 4913 applicants, 2007-2012

- Reject rate = 18%
- Acceptance rate = 82%
  - Default rate = 7%

1. Micro-level data about accepted/rejected and defaulted/non-defaulted clients.
2. Borrower characteristics, terms of the mortgage contract, mortgage characteristics, and the mortgage performance are available.

Sociodemographic information	•Age •Place of birth •Gender •Marital status •Occupation •Education •Income •Numbers of persons in the household
Terms of credit	•Numbers of co-applicants •Income of co-applicants •Loan amount •Maturity •Date of credit •Monthly payment •Cash down •Plan of payments

Mortgage characteristics	•Appraised value •Purchase value •Total area •Numbers of rooms •Location
Credit performance	•Arrear information •Default information
The decision of credit underwriting process	•Maximum loan amount limit •Decision of approval/denial

Conclusions

1. There are few published studies about default modeling on Russian mortgage market (including sample selection bias).
2. Available data set includes recent observations allowing to focus on the drivers behind the recent wave of mortgage defaults.
3. The level of detail in data allows to control for various loan terms and borrower risk factors to control for a more comprehensive list of potential drivers of default.

NATIONAL RESEARCH UNIVERSITY

Thank you for your attention!

AMPoroshina@gmail.com

27, Lebedeva str., Perm, Russia, 614000

www.hse.ru