Effect of Cosigners on Mortgage Default in IL Mortgages between 2011 and 2016

Final Project for PPHA 41430

Andre Oviedo Mendoza

University of Chicago

2024-12-03

Introduction

- Mortgages are a key component of the US housing market and a major source of household wealth.
- The ability to predict mortgage defaults is crucial for risk management in the financial industry.
- This project aims to use debiased machine learning methods to estimate causal effects of borrower characteristics on mortgage performance.
- We will focus on the effect of the number of borrowers on the probability of default.

Motivation

- Importance of a healthy housing market for economic growth.
  - Housing wealth accounts for 34% of the net worth of the median household.
  - The effect of a housing downturn on household wealth is equivalent to losing 10% of GDP.
- Literature on the effect of borrower characteristics on mortgage performance is mixed.
- We do not aim to predict default, but to estimate the effect of borrower characteristics on the probability of default
- Mortgages are an unique type of loan as they are secured by real estate and the possibility of foreclosure is a concern for both lenders and borrowers.

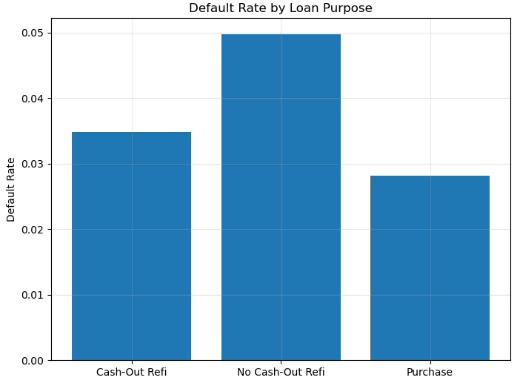
Data Processing

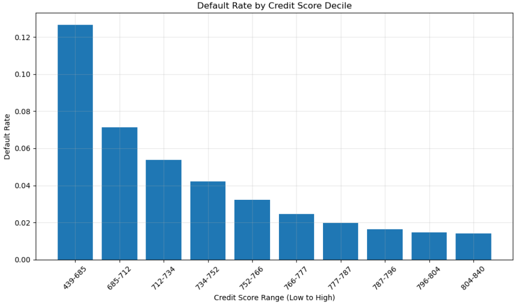
- Source: Freddie Mac Single-Family dataset
- Information from 2011 to 2016 on 96M loans
- Chose to work with a sample of 306K loans (from IL) to speed up computation
- Information on origination and performance. Around 100 variables available for each loan. Choice of covariates focused on borrower and loan characteristics.

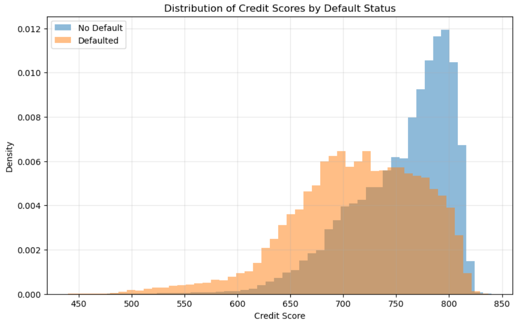
Characteristics of the processed data

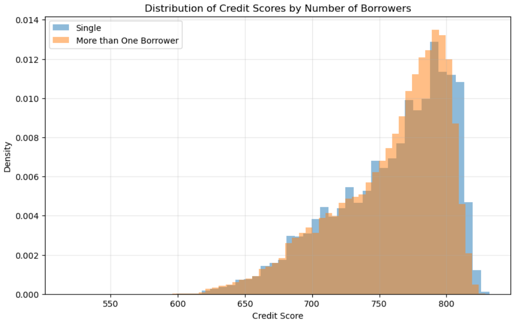
- Origination characteristics:
  - **Number of borrowers**
  - Loan purpose
  - Property type
  - Occupancy type
  - Score
  - Loan amount
  - Loan term
  - Origination date
- Performance:
  - Monthly payment history
  - Current delinquency status
  - **Default status**
- Really high quality data which is not only updates new information but also checks for errors in past publications.

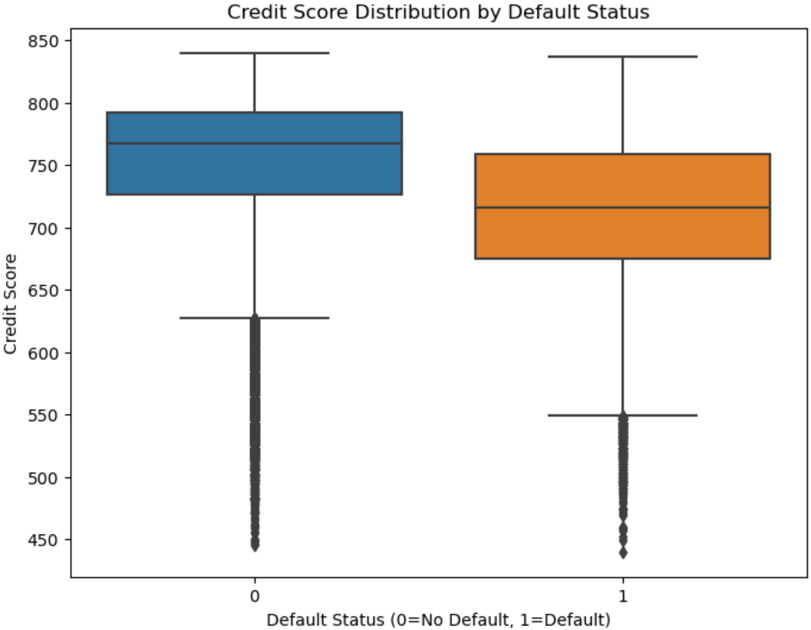
Some graphs


Default Rate by Loan Purpose
Default Rate
0.00
0.01
0.02
0.03
0.04
0.05
Cash-Out Refi
No Cash-Out Refi
Purchase


Default Rate by Credit Score Decile
Default Rate
0.00
0.02
0.04
0.06
0.08
0.10
0.12
439-685
685-712
712-734
734-752
752-766
766-777
777-787
787-796
796-804
804-840
Credit Score Range (Low to High)


Distribution of Credit Scores by Default Status
No Default
Defaulted
Density
Credit Score
0.000
0.002
0.004
0.006
0.008
0.010
0.012
450
500
550
600
650
700
750
800
850


Distribution of Credit Scores by Number of Borrowers
Single
More than One Borrower
Density
Credit Score
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
550
600
650
700
750
800


Credit Score Distribution by Default Status
Credit Score
850
800
750
700
650
600
550
500
450
0
1
Default Status (0=No Default, 1=Default)

Methodology

- Usage of DML to estimate the effect of borrower characteristics on the probability of default.
- Default follows a logistic process.
- Follows proposed methdology by Chernozhukov, Hansen, et al for DML:
    - Modifications: included cluster-robust standard errors at the MSA level and fixed-effects to control for different vintages of loans.

We perform inference on β in the following the partially linear model:

$$default_i = \beta D_i + g(Z_i) + \epsilon_{j,t}.$$

In the first stage, using cross-fitting, we employ modern regression methods to build estimators $\hat{\ell}(Z_i)$ and $\hat{m}(Z_i)$, where

- $\ell(Z_i) := E(Y_i|Z_i)$
- $m(Z_i) := E(D_i|Z_i)$

Using these, we obtain the estimates of the residualized quantities

- $\tilde{Y}_i = Y_i - E(Y_i|Z_i)$
- $\tilde{D}_i = D_i - E(D_i|Z_i)$

Using these residualized quantities, we note our model can be written as

$$\tilde{Y}_i = \beta \tilde{D}_i + \epsilon_i, \quad E(\epsilon_i|\tilde{D}_i) = 0.$$

In the final stage, using ordinary least squares of $\tilde{Y}_i$ on $\tilde{D}_i$, we obtain the estimate of β. We also add cluster-robust standard errors at the MSA level and fixed-effects to control for different vintages of loans.

Main Results

Model	Estimate	Std. Error
Basic Regression	-0.0211	0.0056
Controls	-0.0194	0.0047
DML No Controls	-0.0212	0.0056
DML Basic Controls	-0.0197	0.0046
DML RF	-0.0187	0.0048
DML Boosted Trees	-0.0192	0.0049
DML NN (Early Stopping)	-0.0193	0.0050

Note: All coefficients are significant at the 5% level. Standard errors are clustered at the MSA level. RMSE Y and RMSE D show the prediction accuracy of the first-stage models.

The results show that:

- Having more than one borrower consistently reduces default probability by about 1.9-2.1 percentage points

Conclusions

- Showed the importance of methods such as DML to estimate causal effects of borrower characteristics on mortgage performance in a context of high-dimensional data and nonlinearities.
- The results suggest that the number of borrowers has a positive effect on the probability of default, but this effect is reduced once we control for borrower characteristics and non-linearities.
- The results are robust to the inclusion of different controls and model specifications on the positive effect of the number of borrowers on the probability of default.

Further work

- Include more variables in the analysis:
  - Actually publicly available variables: economic variables, demographics, etc.
  - Variables not publicly available: loan-level variables, debtor characteristics, etc.
- Only worked with a sample of the data: from 96M observations, only worked with 306K.
- Include more heterogeneity in the analysis:
  - Different sub-samples of the data.
  - Data from recent years has a modified number of borrowers variable which goes up to 10 borrowers
- Include different definition of default: used the most conservative one to analyze the effect but in terms of predicting defaults, this might not be the best definition.