Racial Disparities in Mortgage Lending

Ghazal Ayobi

Introduction

Recent incidents in the United States, such as George Floyd's death, have drawn international attention to the country's racial inequalities. This type of segregation has greatly contributed to the racial disparities in a variety of different areas, including financial services; whereas, a mortgage loan is the most popular type of refinancing that has been affected as well. Possessing a home is associated with accumulating more wealth. However, historically, people of color have been denied such mortgages at a higher rate. Thus, The goal of this study is to conduct an empirical analysis of loan approval status on different races in the United States of America. The primary objective of this project is to determine how much people of color are less likely to get a mortgage loan application approved comparing to white citizens, while controlling the same income and loan amounts. To address this topic, I regress the probability of loan approval on different races. Additional confounding variables such as gender, income, loan amount, and collateral are included in the model to support the study. Variables are described as following:

- Dependent variable: Loan approved is binary variable. It is equal to 1 if loan is approved, 0 otherwise
- Explanatory variable: Race is categorical variable that contains information about following races : White, Black, Asian, Hawaiian and Alaskan natives. For each of them I created a binary variable. Confounding Variables:
- Gender: Male is a binary variable (Male = 1, Female = 0)
- Income is applicants annual income
- Loan amount
- Collateral is a binary variable (Secured by first lien = 1, 0 otherwise)

Data Cleaning

Source

The data for this study has been derived from Bureau of Consumer Projection in the United States, 2017. The selected data set is published under Home Mortgage Disclosure Act (HMDA) section which can be found here. To analyze racial inequality in mortgage lending, I selected New Jersey state. The mentioned state is accounted for one of the diverse states containing 50% White, 21% Hispanic or Latino and 12% Black or African American population and it is one of the highly diverse states based on U.S. News *report*. The downloaded financial data for 2017 is the latest available data set regarding this topic. The original data set contains 349563 rows and and 78 columns, which can be found as a zip file in this Github Repository. HMDA is a complicated data for the purpose of this project, these columns are selected for further analysis: applicant id, loan purpose, loan amount, actions taken for the loan application, applicants' race and gender, loan amount, and lien status.
Based on *MPA* the most common loan is Fixed-rate mortgage or conventional home loan. Around 90% of home buyers choose this type of loan. Thus, in this project I only consider Conventional mortgage. While working with this dataset the major problem was its size, due to fact that the HMDA data is really big I could not view it on GitHub repository, so I cleaned the data and uploaded on the mentioned repository.

Filters

Information not provided by the applicant about their gender and race is removed from the data set. Moreover, applicants were filtered if they have withdrawn their loan application. Mortgage loan is the most common loan among lower and middle class families in the United States, and based on Rutgers University in the New Jersey the middle class income is below USD 180,000. Thus, in this research I will be focusing on the American lower and middle class loan applicants. Moreover, as part of data cleaning all N/A values have been dropped.

Dummmy Variables

I created a dummy variable if loan is originated, it is called "loan_approved" and if the loan is approved it is equal to 1 and 0 otherwise. In addition, I created more dummy variables for each racial group such as: White, Black, Asian, and Hawaiian - Alaskan natives. I added more binary variables for gender, and created another column called "Male" which is equal to one if the applicant is male and zero if the applicant is a female. The data set also include other valuable information about loan purpose and lien status for the loan application. I created two more dummy variables, first is "home loan" if the loan purpose is home purchase or home improvement then it is equal to one, zero if the loan purpose is refinancing. Collateral plays a vital role in securing a loan, thus, I created a dummy variable called "collateral" which is equal to one if loan applicantion is secured by first lien or subordinate lien and zero otherwise.

Tranformation

Loan amount and applicant income were multiplied by 1000. As both of the variables are right skewed, their logarithms are considered for further analysis. Figure 1 and Figure 2 exhibits distribution of both variables. As a result the sum of observations is 114622.

Summary Statisitcs

The summary statistics table shows that 70% of loans were approved. 80% of loan applicants are white, and men forms more than half of mortgage applicants. 91% of loan applications are secured by lien.

Table 1: Descriptive statistics

	Mean	Median	SD	Min	Max	P05	P95
Loan Approved	0.70	1.00	0.46	0.00	1.00	0.00	1.00
White	0.80	1.00	0.40	0.00	1.00	0.00	1.00
Black	0.07	0.00	0.26	0.00	1.00	0.00	1.00
Asian	0.12	0.00	0.32	0.00	1.00	0.00	1.00
Hawaiian & Alaskan	0.01	0.00	0.10	0.00	1.00	0.00	0.00
Male	0.65	1.00	0.48	0.00	1.00	0.00	1.00
Home loan	0.60	1.00	0.49	0.00	1.00	0.00	1.00
Collateral	0.91	1.00	0.29	0.00	1.00	0.00	1.00
Income	96 982.42	94 000.00	40 467.21	1000.00	180 000.00	35 000.00	167 000.00
Loan Amount	224 248.64	210 000.00	137 464.01	1000.00	3 750 000.00	23 000.00	464 000.00

Correlation Matrix

A correlation matrix is used to further visualize the association among dependent, explanatory and confounding variables. This matrix helps to predict the evolution of the relationship between variables. Correlation matrix is shown in the Appendix, Figure 3. The correlation matrix shows that if a loan is approved, it is positively associated with log of income, and log of loan amount, collateral and home loan. It can be seen that loan approval status positively correlated with White, on the other hand, it employs negative relationship if race is black or Hawaiian and Alaskan natives.

Model

The main hypothesis of this research is that loan approval status is unequal among races. Thus, First, I estimated a simple linear probability model with loan being approved as a dummy variable regressed on each racial group (binary variables). The linear probability model has the following form.

Model 0

$$LoanApproved^P = \alpha + \beta(race)$$

Table 2 in the Appendix illustrates the probability of loan approval. Column one indicates that we can be 95% confident if someone is white they are 6.8% more likely to get a loan approved and Asians are 3.09% more likely to get it. However, Black or African American, Hawaiian and Alaskan are 18% less likely to get approval for a loan application. For further illustrations, the scatter plots and regression lines that correspond to each regression is shown in the Appendix figures : 4, 5, 6, and 7. However, the scatter plots for each regression line is in corners because both the dependent variable and explanatory variables are dummy variables. The size of the dots are proportional to their frequency in the data. Figure 4 and 6 show that Probability of loan approval for white, same as Asian, have a positive slope. On the contrary, Figures 5 and 7 have negative slopes for Black and Hawaiian-Alaskan natives. To get closer to uncover the effects of racial disparities, loan approval is regressed on all races considering "Asian" as a base category.

Model 1

$$LoanApproved^P = \alpha + \beta_1(hawaiian - alaskan) + \beta_2(black) + \beta_3(white)$$

To address the problem of predicting probabilities that are less than zero or greater than one, there are two models as alternatives to the linear probability model which are logit and probit models. Pseudo R-squared is used to evaluate goodness of fit of a logistic and probit models. It is similar to the R-squared which measures how much goodness-of-fit is compared to what it would be if we were not using any of the right-hand-side variables for prediction. Table 3 shows the result of Pseudo R-squared for Model 1 for both logit and probit regressions which is 1%.

Model 2

To further evaluate the regression model, I added two control variables which are "collateral" and "log of income". Correlation Matrix in Appendix, Figure 3, showed that there is positive correlation between Probability of loan approval, log of income and collateral. As it can be seen from the Table 3 that Pseudo R-squared changes from 1% to 4.7% or both logistic and probit regression of Model 2.

$$LoanApproved^P = \alpha + \beta_1(hawaiian - alaskan) + \beta_2(black) + \beta_3(white) + \beta_4(collateral) + \beta_5(log(income)$$

Model 3

In the Model 3 probability of loan Approved is regressed on Black, White, and Hawaiian or Alaskan. Asian is taken as a base category. Other control variables such as collateral, log of income, gender, loan purpose and log of loan amount are added to the Model. As a result, Table 3 shows that Pseudo R-squared changes to 7.4% for both logit and probit models. Thus, the preferred model is as following.

$$LoanApproved^P = \alpha + \beta_1(hawaiian - alaskan) + \beta_2(black) + \beta_3(white) + \beta_4(collateral) + \beta_5(log(income)$$

$$+\beta_6(male) + \beta_7(home - loan) + \beta_8(log(loan - amount)$$

Table 4 shows results of five regression for Model 3, the preferred model: lpm, logit, marginal logit, probit and marginal probit. Column 1, LPM shows that we can be 95% confident that Hawaiian or Alaskans are 9%, and Black are 10% less likely to get approval for a loan application compared to Asians. However, white are 3.8% more likely to get a loan. Moreover, based on LPM regression we can be 95% confident that loan application with a collateral are 27.4 percentage point more likely to be approved. Under this model 1% change in loan applicants' income, makes them 15% more likely to receive a loan approval. The interesting finding is that male are 1% less likely to get their loan approved. Loan amount is not statistically significant. Based on the heteroskedastic robust standard errors, the results are statistically non different from zero. To show that, a two-sided hypothesis test is provided below:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_A : \beta_1 = \beta_2 = \beta_3 \neq 0$$

The result in the regression confirm the alternative hypothesis holds. It validates the hypothesis of the study where probability of loan approval is statistically significant and different for each included race in the study. To further check the robustness I ran logit and probit regressions for Model. By looking at the logit and probit estimates for the given model, the probability of loan approval to different races, gender, collateral, log of income, and log of loan amount are same as linear probability model. Columns 2 and 3, the Logit Coefficients are almost four and a half times the size of corresponding logit marginal differences. Furthermore, in the column 4 and 5, probit coefficient is almost three times the size of corresponding probit marginal differences. It is interesting to observe that the two marginal differences, logit and probit, are the same with LPM coefficients in column 1. Thus, I will be interpreting the coefficients of marginals differences of both logit and probit models. Figure 8 visualizes the findings of three models with predicted probabilities of logit and probit on y axis and predicted probability of LPM in the x axis. It can be inferred that logit and probit are very similar to each other and very close to LPM as shown by the S-shaped curve lying close to 45 degree line. As a result, it can be said that across all linear probability model, Marginal logit, and marginal probit, we are 95% confident that Hawaiian and Alaskan natives are 9%, 7.7%, 7.9%, consecutively, less likely to get approved loan application. Blacks are 10%, 8.9% and 9% less likely to receive a loan. However, Whites are 3.8%, 4% and 4% more likely to get a loan.

Conclusion

Based on the result of regression analysis, it can be said that the preferred model is Model 3. Supported by increased Pseudo R-squared in the Table 3. Pseudo R-squared significantly increased from 1% to 7.4%. Across all models it is evident that the loan approval is different across races. Thus, the Null hypothesis is invalid. Based on Model 3, by running linear probability model, marginal logit and marginal probit models conveyed a similar result.

Appendix

Figure 1 Income

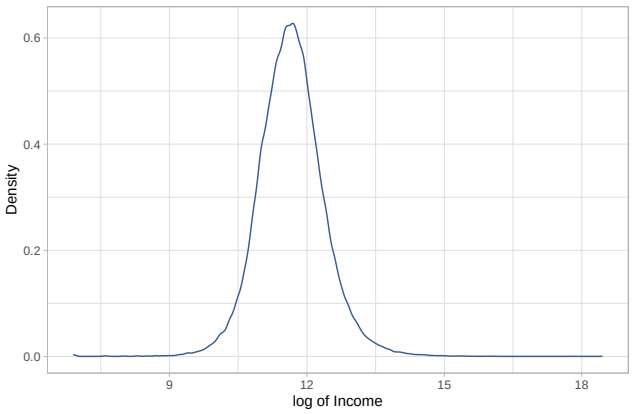
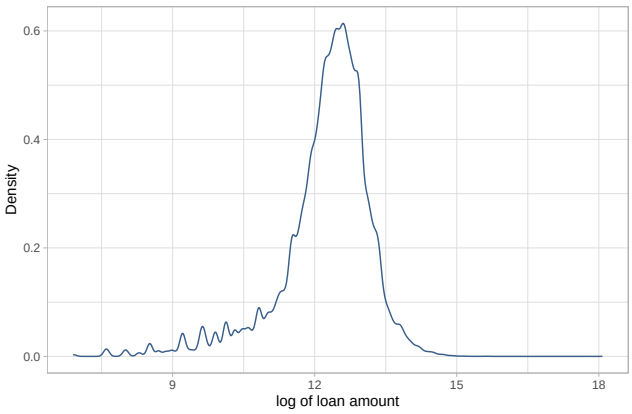
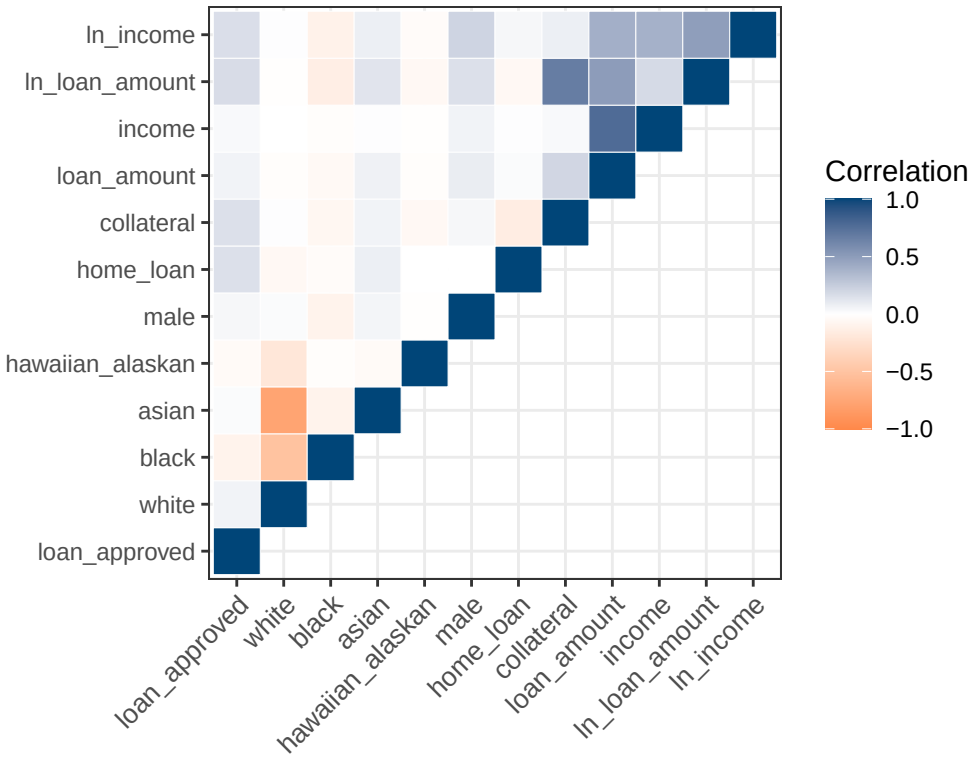


Figure 2 Loan Amount



\

Figure 3



[H]

Table 2: Models to uncover relation between Probability of loan approval and races

	(1)	(2)	(3)	(4)
(Intercept)	0.6439***	0.7116***	0.6947***	0.7004***
	(0.0032)	(0.0014)	(0.0014)	(0.0014)
white	0.0682***			
	(0.0035)			
black		-0.1878***		
		(0.0057)		
asian			0.0309***	
			(0.0041)	
hawaiian_alaskan				-0.1825***
				(0.0145)
Observations	114,622	114,622	114,622	114,622
R2	0.00353	0.01093	0.00048	0.00165

Figure 4 : White

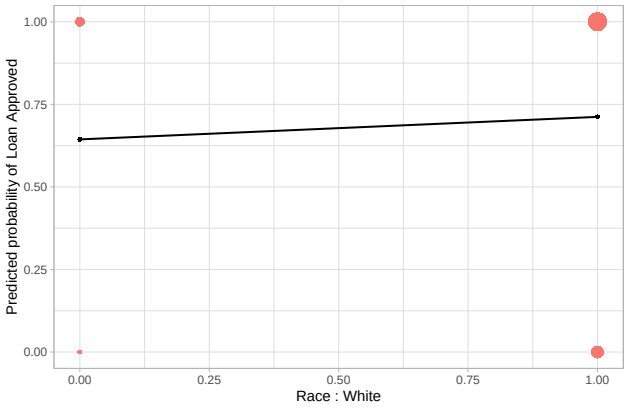


Figure 5 : Black

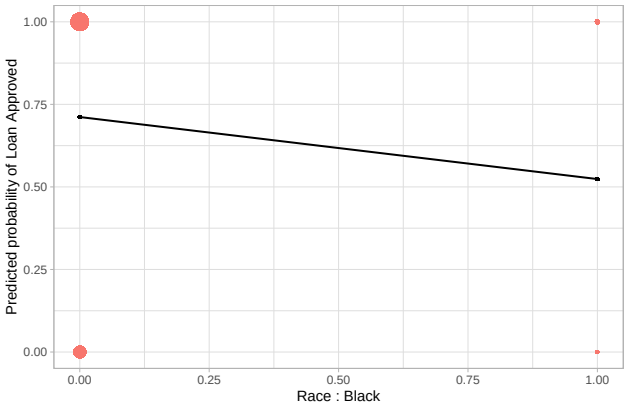


Figure 6 : Asian

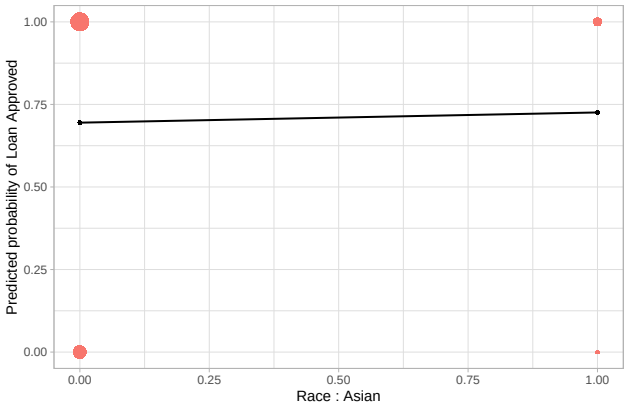


Figure 7 : Hawaiian and Alaskan

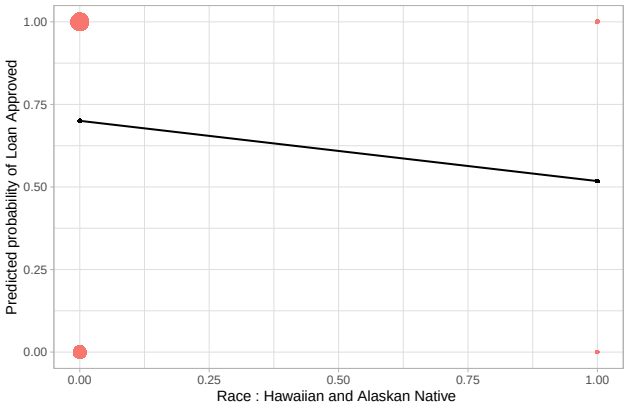


Table 3: Logit, Probit with Pseudo R2

	(M1) logit	(M1) Probit	(M2) logit	(M2) Probit	(M3) logit	(M3) Probit
Constant	0.973**	0.600**	−8.138**	−4.872**	−8.986**	−5.354**
	(0.019)	(0.011)	(0.154)	(0.092)	(0.163)	(0.097)
hawaiian_alaskan	−0.901**	−0.555**	−0.521**	−0.318**	−0.381**	−0.234**
	(0.061)	(0.038)	(0.063)	(0.039)	(0.065)	(0.039)
black	−0.878**	−0.540**	−0.563**	−0.345**	−0.437**	−0.267**
	(0.029)	(0.018)	(0.031)	(0.019)	(0.031)	(0.019)
white	−0.067**	−0.040**	0.086**	0.052**	0.204**	0.123**
	(0.021)	(0.012)	(0.021)	(0.012)	(0.022)	(0.013)
collateral			0.944**	0.578**	1.273**	0.760**
			(0.021)	(0.013)	(0.034)	(0.020)
ln_income			0.716**	0.428**	0.736**	0.435**
			(0.013)	(0.008)	(0.015)	(0.009)
ln_loan_amount					−0.019	−0.008
					(0.012)	(0.007)
male					−0.050**	−0.029**
					(0.014)	(0.009)
home_loan					0.851**	0.509**
					(0.014)	(0.008)
Num.Obs.	114 622	114 622	114 622	114 622	114 622	114 622
PseudoR2	0.010	0.010	0.047	0.047	0.074	0.074

* p < 0.05, ** p < 0.01

Table 4: The Probability of Loan Approval across races- LPM, Logit, and Probit models

	(1)LPM	(2) logit coeffs	(3) logit Marg	(4) Probit	(5) Probit Marg
Constant	−1.301**	−8.986**		−5.354**	
	(0.031)	(0.163)		(0.097)	
hawaiian_alaskan	−0.090**	−0.381**	−0.077**	−0.234**	−0.079**
	(0.013)	(0.065)	(0.014)	(0.039)	(0.014)
black	−0.100**	−0.437**	−0.089**	−0.267**	−0.090**
	(0.006)	(0.031)	(0.007)	(0.019)	(0.007)
white	0.038**	0.204**	0.040**	0.123**	0.040**
	(0.004)	(0.022)	(0.004)	(0.013)	(0.004)
collateral	0.274**	1.273**	0.277**	0.760**	0.272**
	(0.007)	(0.034)	(0.008)	(0.020)	(0.008)
ln_income	0.150**	0.736**	0.141**	0.435**	0.140**
	(0.003)	(0.015)	(0.003)	(0.009)	(0.003)
ln_loan_amount	−0.006*	−0.019	−0.004	−0.008	−0.002
	(0.002)	(0.012)	(0.002)	(0.007)	(0.002)
male	−0.010**	−0.050**	−0.010**	−0.029**	−0.009**
	(0.003)	(0.014)	(0.003)	(0.009)	(0.003)
home_loan	0.168**	0.851**	0.168**	0.509**	0.168**
	(0.003)	(0.014)	(0.003)	(0.008)	(0.003)
Num.Obs.	114 622	114 622	114 622	114 622	114 622

* p < 0.05, ** p < 0.01

Figure 8 : Predicted Probability of LPM, Logit and Probit Models