Family Size and Mortgage Loan Amounts

Georgia Institute of Technology

ECON 3161 Econometrics

Fall 2022

Dr. Shatakshee Dhongde

Rylee Calhoun and Maggie Xia

Abstract

This paper seeks to uncover potential mortgage loan discrimination related to family size for Boston mortgage loan applicants in 1989. We utilize mortgage loan amounts as the primary dependent variable and an applicant's number of dependents as the primary independent variable to explore this relationship. The dataset analyzed in this paper, *loanapp*, comes from Dr. Jeffrey Wooldridge's introductory econometric data repository that records data from the 1989 Boston Federal Reserve Bank's study of the Boston metropolitan area's mortgage practices. After implementing multiple regression models, we find no evidence of statistical discrimination concerning family size, suggestive by the insignificance of our primary independent variable, *dep*, in regression analysis. Specifically, we find that a single increase in an applicant's number of dependents raises mortgage loan amounts by about 0.381% overall and raises loan amounts by 0.137% for applicants who have at least one dependent. Ultimately, we reject our null hypothesis that increases in family size negatively influence mortgage loan amounts.

I. Introduction

In 1968, President Lyndon B. Johnson enacted the Fair Housing Act of 1968 that criminalized housing discrimination related to demographic characteristics such as, but not limited to, familial status (U.S. Department of Housing and Urban Development). Much of this act's passing concerned the rampant redlining and blockbusting practices throughout the Jim Crow era which deterred historically disadvantaged communities from attaining homeownership at similar rates to their white counterparts. Despite the act's passage, many studies have unveiled discriminatory practices that prevail in mortgage lending markets. Given that homeownership is arguably one of the most integral components of the *American Dream*, it is imperative to identify and ameliorate any persisting mortgage lending discrimination that deters individuals from purchasing homes. We will investigate familial status to examine this possibility.

We hypothesize that as family size (i.e. number of dependents) increases, mortgage loan amounts decrease as lenders do not want to assume additional risk on the potential mortgage. Our primary independent variable is thus the number of dependents a family has while our dependent variable is the loan amount approved by the lender and received by the applicant. As the number of dependents increases, applicant income is more thinly divided to support the family. Lenders may view the increased dependency as increased risk, as applicants may miss payments to provide for family members in need of monetary assistance. Thus, there is potential discrimination in mortgage loan amounts with regard to family size.

We control for the variables race, gender, marriage status, loan interest rates (fixed or adjusted), loan term, applicant income, and whether an applicant has any missed mortgage payments (credit history). Lenders may participate in statistical discrimination resulting in gender and/or racial inequality as well. Additional factors that may also affect loan amounts pertaining to family size are the initial income that supports the family and whether the family has one or two income sources. Families with low income or with one income source may also face discrimination from lenders in the form of lesser loan amounts. General factors that also affect loan amounts are the interest rates, which are generally lower with a high down payment as lenders view a large initial stake in property as lower risk; credit history, which is typically higher with low-interest rates and larger loan amounts due to lower risk ascertained in our model with missed payments; and loan term, which indicates the amount of time needed to pay off the loan. By accounting for these variables, we control for other factors that may affect loan amounts other than family size.

II. Literature Review

Ambrose and Diop (2014) explore how increases in subprime mortgage lending opportunities affected the rental and leasing market before the 2008 Financial Crisis, thereby increasing the feasibility and attractiveness of American homeownership. In their analysis, they argue that mortgage credit expansion enabled those previously unable to afford home mortgages to exit the rental market, borrow mortgage loans, and impose negative externalities on individuals forced to remain in the leasing market (2014). Ambrose & Diop utilize residential lease data from Experian RentBureau from January 2002 to November 2009, mortgage data from the Home Mortgage Disclosure Act, and subprime mortgage data from the Department of Housing and Urban Development to achieve two tasks: examine the impact of mortgage credit expansion (through subprime lending) and changes in homeownership rates on residential lease risk (2014). To do so, they employ an ordinary least squares regression model that predicts aggregate monthly lease default indices with the percentage of annual subprime mortgages to the number of purchased mortgages at the metropolitan level as the primary explanatory variable as well as a set of robust control variables with fixed state and year effects (2014). Among these controls, Ambrose & Diop include differences in credit risk, lease characteristics, housing market conditions, and local demographic and economic conditions to improve the validity of their model. Per their estimation results, the primary negative externality that subprime mortgage lending created was an overwhelming influx and saturation of high-risk renters in the leasing market, ultimately increasing rental payments as lenders sought to mitigate the increased risk of rent default by doing so (2014). As low-risk renters that could not previously afford home mortgages migrated to homeownership, which subprime mortgage lending and underwriting made possible, only high-risk renters remained in the leasing market, forcing them to confront expensive rental terms with a high likelihood of defaulting on said terms. Thus, Ambrose and Diop conclude that the implementation of a progressive social policy intended to improve homeownership affordability potentially decreased leasing affordability for those in need of housing most.

Robinson (2002) investigates potential racial, gender, and familial discrimination from lenders in mortgage lending markets for Boston homes in 1992 and uncovers both motherhood penalties and negative child effects. Given the relatively new inception of the Home Mortgage Disclosure Act of 1988, Robinson's paper wields the findings and dataset of the 1992 Boston Federal Reserve Study in efforts to uncover patterns of the aforementioned discriminatory practices that many critics and researchers refused to believe persisted. Robinson analyzes potential mortgage lending discrimination by employing a logistic regression model to predict the probability of loan denial among several subcategories divided by race– Type 1 (married couples) and Type 2 (unmarried male-female couples) that fell under White-only or African-American/Hispanic distinction (2002). Moreover, an important aspect of Robinson's analysis concerned the working status of the woman, whether she was a mother or not, and the subsequent impact

on loan denial. Among these four groups and female specification, Robinson's model found that white couples with a non-working female partner realized a decrease in the probability of loan denial when children were present while white couples with a working female partner realized an increase in the probability of loan denial when children were present (2002). Contrastingly, this trend was reversed for African-American/Hispanic couples where couples with a working female partner experienced decreases in loan denial with children present and increases in loan denial for couples where the woman worked and had children (2002). Moreover, single white women were more disadvantaged than white men in loan denial while single black/Hispanic benefitted from having children (2002). These findings, according to Robinson, potentially accentuate a stark revelation in the beliefs and approaches to mortgage applications for mortgage lenders–they may be operating under the subconscious and inadvertent bias of deeply imperfect social customs coupled with taste discrimination (2002). Robinson's analysis provides early evidence for discrimination to exist in mortgage lending markets, despite American law rendering such practices unlawful.

Berkovec et al (1998) attempt to uncover potential taste-based or *noneconomic* discrimination within the FHA mortgage lending market by addressing the probability of loan default and the value lost due to defaulting on a mortgage loan. Due to the common tendency of discrimination studies to encounter substantial omitted variable bias issues, Berkovec et al employ a logit model with nearly fifty control variables encompassing demographic information (race), loan-specific controls, borrower-specific controls (first-time homeowner), and location-specific controls to estimate the probability of default and mitigate the likely impact of unobserved factors on potential poor loan performance of minority borrowers (1998). Furthermore, Berkovec et al introduce interaction terms to ascertain the true effect of taste-based discrimination on loan defaults, such as whether a borrower is Black and its interaction with the Herfindal-Hirschmann index (HERF), which stands as a measure of market concentration (1998). Before their statistical analysis, Berkovec et al hypothesize that there is no noneconomic discrimination in FHA mortgage lending. Following the results of the probit model, Berkovec et al assert that negative coefficients on the interaction term between race (Black, American Indian, Asian, and Hispanic) and 'HERF' lack statistical significance and therefore indicate no presence of noneconomic discrimination (1998). Unlike related literature, Berkovec et al's paper contradicts the detection of taste-based discrimination in the mortgage lending market as they highlight the oversight of researchers to overestimate the impact of omitted variable bias on their analyses' results. Therefore, Berkovec et al implore future researchers engaging in studies on mortgage loan performance and discrimination to keep the aforementioned issues salient in their methodology.

In this paper, we will analyze multiple ordinary least squares regressions to determine the impact of family size on mortgage loan amounts, whereas other studies have focused their research on the

probability of loan default, the probability of loan acceptance, and/or the probability of loan denial. We will use regression analysis to reveal whether there is a 'child effect' (Robinson 2002) on mortgage loan amounts for loan applicants and if so, quantify that effect through the causal nature of OLS estimations. Moreover, prior literature has greatly emphasized the role of taste-based or noneconomic discrimination concerning race and/or gender in mortgage lending decisions, however, our study will control for variables commonly identified in these studies to discern the true effect of the number of dependents on the total loan amount that mortgage applicants received. Opposite to Berkovec et al's (1998) and Robinson's (2002) research, we hypothesize that any observed familial-status discrimination will be largely statistical in essence as mortgage lenders will attempt to maximize their profits by awarding larger loan amounts to applicants with the least likelihood of default.

III. Data

Table 1: Variable Descriptions

Variable	Description	Year	Units	Source
loanamt	Loan amount the applicant received	1989	$1000s (USD)	Wooldridge/Boston Federal Reserve Bank
dep	Number of dependents the applicant indicated	1989	Dependents	Wooldridge/Boston Federal Reserve Bank
white	Binary; =1 if applicant white	1989	N/A	Wooldridge/Boston Federal Reserve Bank
male	Binary; =1 if applicant male	1989	N/A	Wooldridge/Boston Federal Reserve Bank
married	Binary; =1 if applicant married	1989	N/A	Wooldridge/Boston Federal Reserve Bank
atotinc	Total monthly income	1989	$1000s (USD)	Wooldridge/Boston Federal Reserve Bank
term	Loan term length	1989	Months	Wooldridge/Boston Federal Reserve Bank
fixadj	Binary; =1 if adjusted rate on loan, fixed otherwise	1989	N/A	Wooldridge/Boston Federal Reserve Bank
mortlat	Binary; =1 if missed any mortgage payments	1989	N/A	Wooldridge/Boston Federal Reserve Bank
unit	Number of units in the property	1989	Units	Wooldridge/Boston Federal Reserve Bank
sch	Binary; =1 if applicant has > 12 years of schooling	1989	N/A	Wooldridge/Boston Federal Reserve Bank

Despite Berkovec et al's finding of statistical insignificance concerning racial discrimination in mortgage lending, we control for race to capture any cultural/social biases that may influence lender

decision-making and to avoid prejudicing our ability to isolate the true 'child effect' on mortgage loan amounts (Berkovec et al 1998). Per Robinson, we include male and married indicators to control for biased 'child-effect' estimates by the gender penalty on mortgage loans for women, the motherhood penalty on mortgage loans for white women, the motherhood advantage on mortgage loans for minority women, and the mortgage loan disadvantage associated with single-parent families (Robinson 2002). Further, we include applicants' monthly income as a control to proxy for general economic standing as applicants who earn greater salaries are more likely to purchase more expensive homes and receive larger loan amounts than applicants who earn lesser salaries. Similarly, we control for the term of the loan in months as lenders will typically award greater loan amounts to accommodate longer mortgages.

To ensure that we do not bias the 'child effect' we observe in our estimates with loan quality and type, we follow suit with Ambrose and Diop and incorporate a binary variable indicating whether there was a fixed or adjustable rate on the mortgage loan as a control (Ambrose & Diop 2014). Because the dataset does not record a continuous loan rate, we use the fixed or adjusted loan to compensate for this lapse in data collection. We additionally incorporate late mortgage payment occurrences as a control variable to intercede any profit-maximizing behavior–per statistical discrimination–on part of the mortgage lender. Lastly, we add the number of units on the mortgaged property to control for variations in housing size and lending amounts and a binary variable to assess the education level of the loan applicant, which may influence the lender's perception of the applicant's ability to repay the mortgage loan. The binary element of the education variable outlines whether an applicant has completed secondary education, as it is set to one if the applicant has at least twelve years of schooling. In the United States, almost all students go through twelve years of schooling before graduating high school, completing their secondary education, and choosing to pursue higher-level education.

The data used in this analysis is an introductory econometric dataset supplied by Jeffrey Wooldridge's instructional textbook, *Introductory Econometrics: A Modern Approach*. The raw data was initially gathered and compiled in the Boston Federal Reserve Bank's 1989 Study of Mortgage Lending which followed mortgage lending trends throughout the metropolitan Boston area. From the data provided by Wooldridge, the dataset contains 1,989 observations with 59 variables. Not all variables provided in this dataset will be used in our statistical analysis. Because this dataset is a byproduct of the Boston Federal Reserve Bank's 1989 study, it is cross-sectional in nature and outlines mortgage lending figures for the year 1989. From the dataset, we use the variables *loanamt* (loan amount), *dep* (number of dependents), *fixadj* (interest rates), *term* (loan term in months), *atotinc* (monthly applicant income), and *unit* (number of units in property). In our dataset, we observed an anomaly in the variable *term*, which originally had four observations with a 999,999.4-month loan. Because the next highest loan term is 480

months, we removed those observations to prevent skewing our model and thus use 1,985 observations from the dataset. We also use the binary variables *white* (1 if white), *male* (1 if male), *sch* (1 if >12 years of schooling), and *married* (1 if married). We create and append the variable *mortlat* to the dataset from the original variables *mortlat1* and *mortlat2* which respectively indicated whether an applicant had 1-2 missed payments or more than two missed payments. Our combined variable *mortlat* is a binary variable where 0 indicates no missed payments and 1 indicates any missed payments. We take the logarithmic value of *loanamt* to generate the variable *logloan*, creating a log-level model to analyze the unit increase in dep resulting in 100*β_k% increase in loan amounts.

Table 2: Summary Statistics

Summary statistics

	N	Mean	SD	Median	Min	Max
loanamt	1985	143.332	80.576	126	2	980
logloan	1985	4.846	.488	4.836	.693	6.888
dep	1982	.77	1.104	0	0	8
white	1985	.846	.361	1	0	1
male	1970	.813	.39	1	0	1
married	1982	.658	.474	1	0	1
atotinc	1985	5197.291	5273.538	3813	0	81000
logtinc	1981	8.338	.598	8.247	5.011	11.302
term	1985	341.086	64.494	360	6	480
logterm	1985	5.795	.349	5.886	1.792	6.174
fixadj	1985	.307	.461	0	0	1
mortlat	1985	.03	.17	0	0	1
unit	1981	1.123	.438	1	1	4
sch	1985	.772	.419	1	0	1

Table 2 outlines the summary statistics for the variables used in our analysis. Within the sample, the average mortgage loan amount the applicant received is $143,332. Because the standard deviation of the received loan amounts is fairly large, $80,520, we log-transformed our primary dependent variable, *loanamt*, to reduce the large spread of awarded loan amounts observed. Our independent variable, *dep*, records a range of zero to eight dependents across the sample with a mean of 0.771, suggesting that most applicants have zero or one dependent. We also account for loan terms and applicants' monthly income as continuous variables. On average, applicants in our sample requested loans that were about 341 months (approximately 28.42 years) long and generated about $5,197 in monthly income (1989 nominal USD). We include six binary variables: *white*, *male*, *married*, *fixadj*, *mortlat*, and *sch*. As shown above, most applicants are white males, with nearly half of this sample being married and not having missed any mortgage payments, and over three-quarters of the sample having education beyond high school completion. Only 30.7% of the sample received an adjusted-rate mortgage loan, suggesting that most applicants received fixed-rate mortgage loans. The average number of units in the property is about 1,

with minimum and maximum values of 1 and 4, respectively. Because our sample is primarily married men with no dependents, it seems sensible that most applicants sought one-unit homes to accommodate themselves and their spouses.

Figure 1: Scatterplot

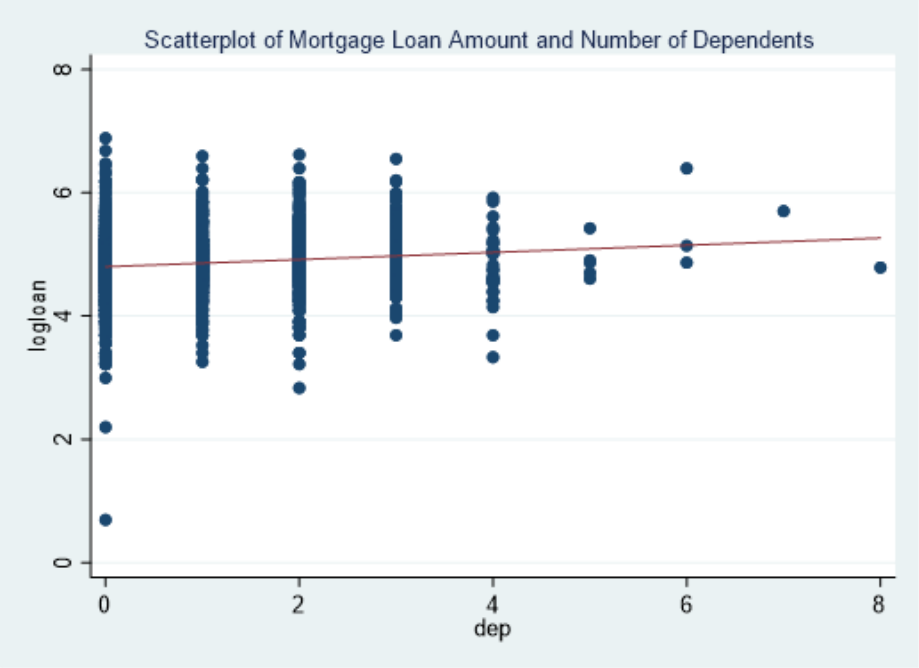


Figure 1 illustrates the scatterplot of *logloan* and *dep*. The trend line shows a slightly positive relationship between the two variables, suggesting that an increase in the number of dependents could somewhat increase the percentage of the loan amount awarded. This preliminary discovery is contrary to our hypothesis that loan amounts decrease with increased numbers of dependents, but we note possible sample bias in the variation of applicants' number of dependents that may skew regression estimates. Figure 1 reveals that very few applicants have more than four dependents, whereas many have no dependents. Although the range of *logloan* when *dep* = 0 is large, demonstrating variation in loan amounts, the variation in *dep* is small in nature.

Our model is linear in parameters with an error term 'u', indicated by the relatively linear trend observed in the scatterplot of mortgage loan amount to the number of dependents. Our dataset meets the random sampling condition as the data is drawn as a random sample from the overall population in the city of Boston, Massachusetts, by the Boston Federal Reserve Bank. Although the majority of the sample is clustered around 0 and 1 dependents, the sample variation in family size does not have a standard deviation of zero.

The model also meets the no perfect collinearity condition as the correlation table (Appendix B.a) shows relatively weak associations between the controls we include. The largest correlation has a coefficient of 0.518 between *logtinc* and *logloan*, which is to be expected, as income is a large factor in lenders approving loans. The next highest correlation is between *dep* and *married* with a coefficient of 0.357. The remaining variables have less than 0.2 correlation coefficients, demonstrating the lack of perfect correlation.

To satisfy the zero conditional mean assumption, we include controls that we believe to be strong influencers on our dependent variable, mortgage loan amount, and thus minimize the error term. Nonetheless, we acknowledge that there is potential for small omitted variable bias as we are incapable of providing controls to accommodate each observed factor influencing mortgage loan amounts. Initially, our model did not meet the homoscedastic assumption of constant variance in the error terms and yielded a classic cone shape in a residual vs. fitted plot. To correct this, we applied log transformations to non-binary variables with large standard deviations, namely *loanamt*, *term*, and *atotinc*. The correction resulted in a relatively constant variance in the error terms, satisfying homoscedasticity (demonstrated with Model 3 in Appendix C). Additionally, our error term exhibits normal distribution (demonstrated with Model 3 in Appendix C), thereby satisfying the assumptions of both the Classical Linear Model and Gauss-Markov.

IV. Results

Our results can be classified into two sections. Models 1 – 3 include all observations, where dependents range from 0 to 8. Models 4 and 5 match their 1 and 2 parts respectively but include less observations, where we removed *dep* = 0 observations to attempt correcting possible sample bias. Model 6 is similar to Model 3, with the difference of the variable *white* being removed instead of *male*, as *white* was insignificant for Model 6 but not Model 3.

Table 3: Regression Models

Models	Equation	Estimated Equation
Model 1	$logloan = \beta_0 + \beta_1 dep + u$	$logloan = 4.801 + 0.0585 dep$
Model 2	$logloan = \beta_0 + \beta_1 dep + \beta_2 white + \beta_3 male + \beta_4 married + \beta_5 fixadj + \beta_6 logterm + \beta_7 mortlat + \beta_8 logtinc + \beta_9 unit + \beta_{10} sch + u$	$logloan = 0.546 + 0.00269 dep - 0.038 white + 0.0945 male + 0.148 married + 0.0474 fixadj + 0.124 logterm - 0.0003 mortlat + 0.388 logtinc + 0.0887 unit + 0.109 sch$
Model 3	$logloan = \beta_0 + \beta_1 dep + \beta_2 unit + \beta_3 married + \beta_4 fixadj + \beta_5 logterm + \beta_6 logtinc + \beta_7 sch + \beta_8 male + u$	$logloan = 0.516 + 0.00381 dep + 0.0929 unit + 0.148 married + 0.0483 fixadj + 0.127 logterm + 0.386 logtinc + 0.107 sch + 0.0925 male$
Model 4	$logloan = \beta_0 + \beta_1 dep + u$	$logloan = 4.86 + 0.0333 dep$
Model 5	$logloan = \beta_0 + \beta_1 dep + \beta_2 white + \beta_3 male + \beta_4 married + \beta_5 fixadj + \beta_6 logterm + \beta_7 mortlat + \beta_8 logtinc + \beta_9 unit + \beta_{10} sch + u$	$logloan = 0.26 + 0.00186 dep - 0.0619 white + 0.00183 male + 0.223 married + 0.0551 fixadj + 0.135 logterm - 0.0339 mortlat + 0.42 logtinc + 0.0789 unit + 0.115 sch$
Model 6	$logloan = \beta_0 + \beta_1 dep + \beta_2 unit + \beta_3 married + \beta_4 fixadj + \beta_5 logterm + \beta_6 logtinc + \beta_7 sch + \beta_8 white + u$	$logloan = 0.244 + 0.00137 dep + 0.0795 unit + 0.218 married + 0.0566 fixadj + 0.135 logterm + 0.422 logtinc + 0.111 sch - 0.0578 white$

Table 4: Regression Output

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 5	(6) Model 6
dep	0.0585***	0.00269	0.00381	0.0333*	0.00186	0.00137
	(0.00985)	(0.00905)	(0.00900)	(0.0187)	(0.0154)	(0.0153)
logterm		0.124***	0.127***		0.135***	0.135***
		(0.0262)	(0.0262)		(0.0412)	(0.0411)
logtinc		0.388***	0.386***		0.420***	0.422***
		(0.0162)	(0.0161)		(0.0241)	(0.0238)
white		-0.0380			-0.0619*	-0.0578
		(0.0260)			(0.0373)	(0.0369)
male		0.0945***	0.0925***		0.00183	
		(0.0252)	(0.0251)		(0.0509)	
married		0.148***	0.148***		0.223***	0.218***
		(0.0218)	(0.0217)		(0.0490)	(0.0434)
fixadj		0.0474**	0.0483**		0.0551*	0.0566*
		(0.0201)	(0.0201)		(0.0313)	(0.0312)
mortlat		-0.000281			-0.0339	
		(0.0546)			(0.0754)	
unit		0.0887***	0.0929***		0.0789**	0.0795**
		(0.0214)	(0.0212)		(0.0330)	(0.0329)
sch		0.109***	0.107***		0.115***	0.111***
		(0.0226)	(0.0226)		(0.0345)	(0.0343)
Constant	4.801***	0.546***	0.516**	4.860***	0.260	0.244
	(0.0133)	(0.205)	(0.203)	(0.0394)	(0.316)	(0.315)
Observations	1,982	1,961	1,961	809	801	807
R-squared	0.018	0.323	0.322	0.004	0.382	0.383

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3 outlines the regression output for the models we constructed for this analysis. In model 1, the simple regression model, *dep* is statistically significant at the 1% level with a positive coefficient estimate. For this model, each additional increase in the number of dependents will increase the loan amount by about 5.85%. The coefficient of determination is quite small, indicating that our primary independent variable, *dep*, does not have much prediction power in explaining variations in loan amounts.

Model 2 serves as the initial multiple regression model while incorporating each control variable of interest. The primary independent variable, *dep*, is no longer significant, but still maintains a positive coefficient that is lesser in value than the estimate observed in the simple regression model. All remaining control variables aside from *white* and *mortlat* show significance at the 1% level. *White* records a negative coefficient, suggesting that if the applicant lists their race as white, the loan amount awarded decreases by approximately 3.8%, which could be a result of lesser need of loan amounts due to increased income compared to non-white applicants. The coefficient of greatest magnitude is *logtinc*, which we anticipated seeing that income and the applicant's ability to afford the mortgage loan should be one of the primary elements of a lender's decision to award the mortgage loan. Thus, a 1% increase in an applicant's monthly income will increase the loan amount by approximately 38.8%.

For our qualitative variables, there appears to be a gender premium on mortgage loan amounts with male applicants receiving a 9.45% increase in loan amounts compared to women. Applicants who have greater than 12 years of schooling realize a 10.9% increase in loan amount compared to those who have less education. Regarding home size, each additional increase of units in the home raises the loan amount by 8.87%. This estimate appears intuitive as larger homes are listed with greater prices and require larger mortgage loans. Additionally, fixed-rate mortgage loans receive about a 4.74% increase in loan amounts as opposed to adjusted-rate mortgage loans. With this model, the coefficient of determination increased greatly from 0.018 to 0.323, revealing a model with stronger prediction power than model 1.

Because *mortlat* and *white* lacked significance at the 1%, 5%, and 10% levels in model 2, we excluded these variables in model 3 and re-estimated the coefficients. Upon this change, the coefficient on *dep* increased to approximately 0.004 and *unit* increased to about 0.0929, while the remaining control variables listed similar estimates as model 2. In general, we find that an applicant's number of dependents is not significant in the decision of awarding various mortgage loan amounts. Although models 2 and 3 record much higher values for the coefficient of determination, the large coefficient estimate for *logtinc* suggests that monthly income remains an integral predictor of mortgage loan amounts for applicants.

Considering that over 50% of our sample are applicants that list no dependents, we sought to repeat models 1-3 while restricting *dep* to observations where *dep* > 0. This restriction produces a sample of applicants with at least one dependent and allows us to measure a potential marginal child effect within already-established families. In the simple regression model, *dep* is significant and positive at the 10% level but shows insignificance in multiple regression models 5 and 6. In model 5, *male* is no longer significant and *mortlat* remains insignificant as previously seen in model 2. Model 6 removes these two variables and recalculates the coefficient estimates. For the sample restricted to applicants with at least one dependent, we observe that a single increase in dependents raises mortgage loan amounts by 0.137%. *Logtinc* remains significant at the 1% level with a 1% increase in monthly income increasing loan amounts by 42.2%. A similar marriage premium is realized, but greater in magnitude than model 3. Consistent estimates when compared to model 3 result for *sch*, *unit*, and *fixadj*, but *fixadj* and *unit* decrease from significant at the 5% level to 10% level and 1% level to 5% level, respectively. The increased coefficient of determination suggests that this model is more suited to applicants who already have dependents as opposed to those who do not.

Tables 4 and 5 below depict the statistical inferences of Model 3 and Model 6, which we consider our final models. Our null hypothesis is H_0: $\beta_{dep} = 0$, while our alternative hypothesis is H_a: $\beta_{dep} < 0$. The null hypothesis states that increases in family size have no effect on mortgage loan amounts, and the

alternative states as family size increases, mortgage loan amounts decrease. Both Model 3 and Model 6 t-statistics for *dep* are smaller in comparison to the other variables. The independent variable *dep* is also statistically insignificant for both models, with a p-value greater than 0.1. The remaining variables (controls) in the models are all significant at an alpha of 5%, with most being significant at an alpha of 1%. The exception is *white* in Model 6, which is also statistically insignificant. The confidence intervals are given at a 95% confidence level. For both of our final models, we fail to reject our null hypothesis due to the given t-statistics, which is further corroborated by our confidence intervals for dep. These intervals range from negative to positive and include 0, indicating that we do not have enough evidence to conclude that there is an effect of an applicant's increased number of dependents on loan amounts.

Table 5: Model 3 Stat Inference

	(1)	(2)
VARIABLES	T-Stat	Confidence Interval
logloan		-
dep	0.424	-0.0138 - 0.0215
male	3.685***	0.0433 - 0.142
married	6.784***	0.105 - 0.190
fixadj	2.409**	0.00898 - 0.0876
logterm	4.852***	0.0756 - 0.178
logtinc	23.97***	0.354 - 0.418
unit	4.375***	0.0512 - 0.135
sch	4.718***	0.0622 - 0.151
Constant	2.538**	0.117 - 0.915
Observations	1,961	
R-squared	0.322	

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 6: Model 6 Stat Inference

	(1)	(2)
VARIABLES	T-Stat	Confidence Interval
logloan		-
dep	0.0900	-0.0286 - 0.0313
white	-1.565	-0.130 - 0.0147
married	5.034***	0.133 - 0.304
fixadj	1.815*	-0.00461 - 0.118
logterm	3.295***	0.0547 - 0.216
logtinc	17.69***	0.375 - 0.469
unit	2.419**	0.0150 - 0.144
sch	3.246***	0.0440 - 0.179
Constant	0.775	-0.374 - 0.862
Observations	807	
R-squared	0.383	

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

V. Extensions

In model 3, we exclude *white* and *mortlat* from regression analysis as they lack significance in our kitchen sink regression model, model 2. Despite the insignificance associated with these variables when tested individually, we acknowledge that joint significance between *white* and *mortlat* could drive these results. Given the United States' historical redlining practices that exacerbated fair housing opportunities and continuing wage inequality between races, *white* and *mortlat* may be highly correlated as mortgage lenders assume that white applicants earn greater salaries than their minority counterparts, leading them to miss fewer mortgage payments. We explore this hypothesis further using a multicollinearity check with variance inflation factors and an F-test to detect potential joint significance.

a) Multicollinearity

Table 7: Variance Inflation Factor

Variance inflation factor

	VIF	1/VIF
married	1.279	.782
dep	1.2	.834
male	1.154	.867
logtinc	1.136	.88
sch	1.088	.919
white	1.057	.946
fixadj	1.037	.965
unit	1.032	.969
logterm	1.021	.98
mortlat	1.013	.987
Mean VIF	1.101	.

Table 6 shows the variance inflation factor estimates for each of our control variables used in model 2. Of particular interest is *white* and *mortlat* which record values of 1.057 and 1.013, respectively. Given that VIF values closer to 1 represent less multicollinearity among each control variable, *white* and *mortlat* do not appear to have alarming associations with one another or our other control variables of interest. We continue to investigate this finding in a correlation matrix (reference Appendix B.a), where the correlation between *white* and *mortlat* is approximately -0.028. This weakly negative relationship between the two variables suggests that whether the applicant is white decreases whether the applicant misses any mortgage payments. Thus, despite the findings of the VIF estimates, the correlation coefficient leads us to explore this potential multicollinear relationship further by implementing an F-test.

b) F-test

$$H_0: \beta_{white} = 0, \beta_{mortlat} = 0$$

$$H_1: H_0 \; is \; not \; true$$

$$df = n - k - 1 = 1961 - 10 - 1 = 1950$$

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} = \frac{(318.036 - 317.689)/2}{317.689/(1961-10-1)} \approx 1.065$$

$$F \sim F_{2,1950} \rightarrow \; c_{0.05} = 3.00$$

To conduct the F-test, we state our null hypothesis as *white* and *mortlat* being jointly insignificant. Our alternate hypothesis states that our null hypothesis is not true, suggesting that the two variables are jointly significant. The degrees of freedom we use for our calculations is 1,950. To calculate the F-statistic, we utilize the squared sum of residuals from model 2, our unrestricted model, and model 3, our restricted model that excludes *white* and *mortlat*. We test our F-statistic at the 5% significance level using the number of restrictions in the restricted model, 2, and the overall degrees of freedom, 1,950.

Because our F-statistic does not fall in the rejection region of the 5% significance level, we fail to reject the null hypothesis that *white* and *mortlat* are jointly insignificant ($F = 1.065 < c_{0.05} = 3.00$). This suggests that there is little association between whether the applicant lists their race as white and if the applicant has previously missed any mortgage payments. This result renders model 3 valid as we can conclude that whether the applicant was white and missed any mortgage payments has no effect on our primary dependent variable, mortgage loan amounts, and can be excluded from the regression.

We abstain from including additional models with more dummy variables, as we feel that an overall model with eight control variables suffices. Moreover, we note that Model 2 is a kitchen sink model including ten variables, thus reinforcing the idea to exclude additional dummy variables. We also hesitate to include different functional forms, as our three continuous variables are already log-transformed, and the remaining variables are discrete or binary.

VI. Conclusion

a. Limitations

We acknowledge several limitations in our paper, notably that the dataset is 33 years old and constrained to the Boston metropolitan area, a fairly urban environment. Thus, we are hesitant to extend our findings to other United States cities, especially rural locations, and in modern contexts. Additionally, we attempted to decrease any possible sample bias by restricting the sample to applicants listing at least one dependent (Models 4-6), but there is a likelihood that sample bias still exists as the dataset also contains a large number of applicants with one or two dependents. In contrast, there are only five applicants with greater than five dependents. Our primary independent variable, *dep*, is a discrete variable ranging from zero to eight, so this possible sample bias could potentially skew our dataset. Additionally, the age of the applicant was not recorded, so it could be possible that our dataset contains individuals in the early stages of their professional careers who have just begun family planning, especially considering that 59.16% of applicants in the sample listed no dependents. We also lack data on the loan amount requested and specific interest rates. The control *fixadj* mitigated possible omitted variable bias from lacking specific interest rates, but it is a binary variable and the analysis likely would benefit from a continuous variable with interest rate amounts, which can influence the amount loaned. Due to the age of the study and the lack of data itself, we could not provide these variables.

b. Conclusion of Results

In the final analysis, we find that a mortgage loan applicant's number of dependents has no effect on the mortgage loan amount the applicant receives. Even in statistical models restricted to already established families, we found little evidence of the adverse impact of an increased number of dependents

lowering mortgage loan amounts for the applicant. Instead, our models revealed that monthly income is the greatest predictor of mortgage loan amounts, followed by the overall loan term, whether the applicant has greater than 12 years of schooling, and whether the applicant is married. Monthly income, loan term, and education seem rather intuitive in a lender's decision to award a specific loan amount as a lender will not award an applicant a loan they cannot afford based on their income, will not agree to long loan terms where they will not realize profits in adequate time, and do not believe in the applicant's income potential, signaled by education level, to repay the loan. The marriage premium, contrastingly, is quite interesting in our analysis. In both models 3 and 6, marriage was a significant and positive predictor of the percent change in the loan amount awarded. This may result as lenders envision dual-income households as less risky compared to single-headed households that depend on one stream of income to pay off the mortgage loan. Nonetheless, the lack of significance corresponding to an applicant's number of dependents in predicting loan amounts suggests that there was no familial discrimination occurring at the time of this study. This finding contradicts our initial hypothesis which states that as family size increases, mortgage loan amounts decrease as lenders seek to minimize risk on mortgage loan repayment.

c. Future Research

Despite detecting no familial discrimination in this sample, we believe that analyses like such are integral to upholding the framework of equal opportunity established by the Fair Housing Act. Follow-up studies should seek to collect more data that extends to a diverse array of U.S locations in more recent years to detect whether there is any ongoing potential familial discrimination in mortgage lending. Larger samples across the country may help generalize results, but many cities can also benefit from fairly local studies to pinpoint and investigate any discrimination specific to that particular region, as seen in the Boston Federal Reserve Bank study.

VII. References

Ambrose, B. W., & Diop, M. (2014). Spillover effects of subprime mortgage originations: The effects of single-family mortgage credit expansion on the multifamily rental market. *Journal of Urban Economics*, 81, 114-135.

Berkovec, J. A., Canner, G. B., Gabriel, S. A., & Hannan, T. H. (1998). Discrimination, competition, and loan performance in FHA mortgage lending. *Review of Economics and Statistics*, 80(2), 241-250.

History of fair housing - HUD. HUD.gov / U.S. Department of Housing and Urban Development (HUD). (n.d.). Retrieved October 13, 2022, from https://www.hud.gov/program_offices/fair_housing_equal_opp/aboutfheo/history

Robinson, J. K. (2002). Race, gender, and familial status: discrimination in one US mortgage lending market. *Feminist Economics*, 8(2), 63-85.

10.7 - detecting multicollinearity using variance inflation factors. 10.7 - Detecting Multicollinearity Using Variance Inflation Factors | STAT 462. (n.d.). Retrieved November 17, 2022, from https://online.stat.psu.edu/stat462/node/180/

VIII. Appendix

A) STATA Code

```
// generate variable descriptions
describe loanamt dep white male married atotinc term fixadj mortlat1 mortlat2 unit sch

// create mortlat variable
gen mortlat = mortlat1 + mortlat2

// generate variable descriptions
describe loanamt dep white male married atotinc term fixadj mortlat unit sch

// summarize chosen variables
summarize loanamt dep white male married atotinc term fixadj mortlat unit sch

// take log of all non-binary variables
gen logloan = log(loanamt)
gen logterm = log(term)
gen logtinc = log(atotinc)

//summarize chosen variables with log transformed variables
summarize loanamt logloan dep white male married atotinc logtinc term logterm fixadj mortlat
unit sch

// explore outliers
tabulate term
tabulate atotinc

// investigate dep further
tabulate dep
asdoc tabulate dep

// drop outliers from term value
drop if term > 90000

//summarize chosen variables after dropping outliers
summarize loanamt logloan dep white male married atotinc logtinc term logterm fixadj mortlat
unit sch

//export summary statistics table
asdoc sum loanamt logloan dep white male married atotinc logtinc term logterm fixadj mortlat
unit sch, stat(N mean sd median min max)

// variable correlations
corr loanamt logloan dep white male married atotinc logtinc term logterm fixadj mortlat unit sch

//export correlation table
```

```
asdoc corr logloan dep white male married logtinc logterm fixadj mortlat unit sch

// scatter logloan and dep
graph twoway (scatter logloan dep) (lfit logloan dep)

// simple regression model
regress logloan dep
outreg2 using myreg.doc, replace ctitle(Model 1)

// multiple regression model 1 (kitchen sink model)
regress logloan dep logterm logtinc white male married fixadj mortlat unit sch
outreg2 using myreg.doc, append ctitle(Model 2)

// multiple regression model 2 (exclude white and mortlat)
regress logloan dep logterm logtinc male married fixadj unit sch
outreg2 using myreg.doc, append ctitle(Model 3)

// simple regression model when dep > 0
regress logloan dep if dep > 0
outreg2 using myreg.doc, append ctitle(Model 4)

// multiple regression model when dep > 0 (kitchen sink model)
regress logloan dep logterm logtinc white male married fixadj mortlat unit sch if dep > 0
outreg2 using myreg.doc, append ctitle(Model 5)

// multiple regression model 2 when dep > 0 (exclude logterm, male, and mortlat)
regress logloan dep logterm logtinc white married fixadj unit sch if dep > 0
outreg2 using myreg.doc, append ctitle(Model 6)

// construct residual plot
plot residuals vs. fitted values: rvfplot, yline(0)

// construct histogram with predicted values
plot residual histogram: predict resid, residuals

// construct histogram with residuals and normalized values
hist resid, normal

// variance inflation factor for model 2
regress logloan dep logterm logtinc white male married fixadj mortlat unit sch
vif
asdoc vif
```

B) Additional Tables

   a. Correlation Matrix

Matrix of correlations

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
(1) logloan	1.000										
(2) dep	0.136	1.000									
(3) white	0.022	-0.064	1.000								
(4) male	0.167	0.167	0.055	1.000							
(5) married	0.199	0.357	0.006	0.348	1.000						
(6) logtinc	0.518	0.161	0.114	0.105	0.083	1.000					
(7) logterm	0.082	-0.046	-0.076	-0.027	-0.023	-0.023	1.000				
(8) fixadj	0.128	-0.017	-0.019	0.003	0.013	0.136	0.106	1.000			
(9) mortlat	0.040	0.074	-0.028	0.036	0.009	0.073	-0.020	0.035	1.000		
(10) unit	0.063	-0.003	-0.140	-0.021	-0.063	-0.005	0.022	0.004	-0.006	1.000	
(11) sch	0.184	-0.079	0.102	-0.042	-0.053	0.226	0.009	0.073	0.015	-0.086	1.000

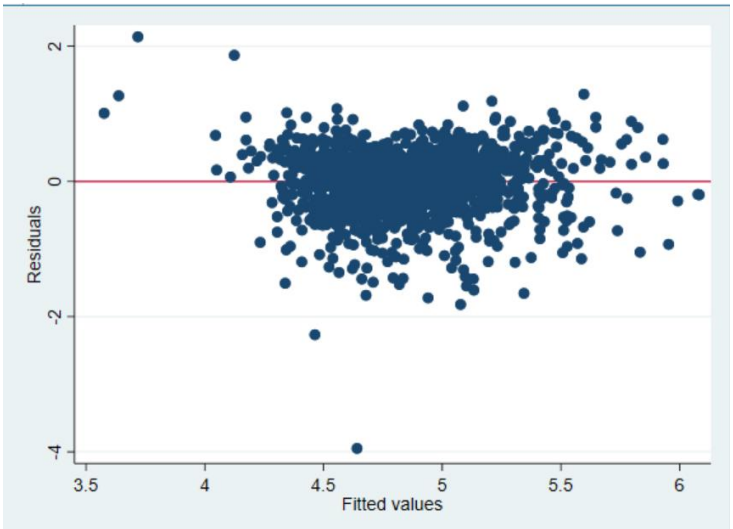
b. Frequency table for dep

Tabulation of dep

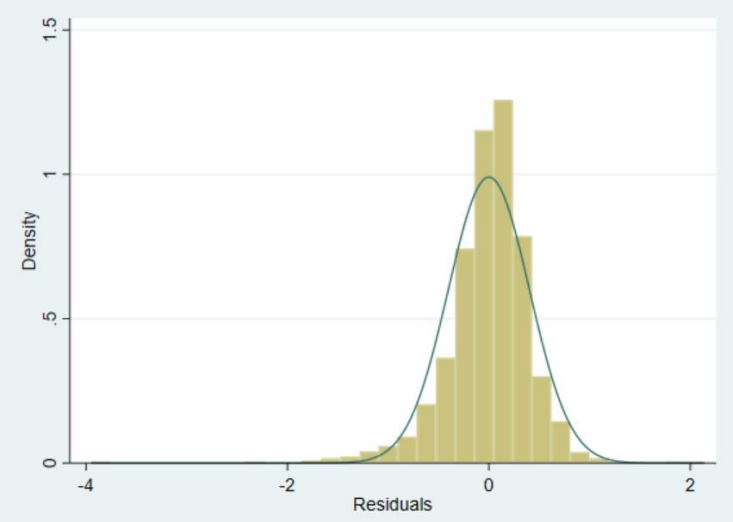
number of dependents	Freq.	Percent	Cum.
0	1175	59.16	59.16
1	317	15.96	75.13
2	327	16.47	91.59
3	126	6.34	97.94
4	31	1.56	99.50
5	5	0.25	99.75
6	3	0.15	99.90
7	1	0.05	99.95
8	1	0.05	100.00
Total	1986	100.00	

C) Additional Figures

a. Model 3 Residual vs. Fitted Plot



b. Model 3 Normal Residuals



D) Stata output

a. Model 2

```
. // multiple regression model 1 (kitchen sink model)
. regress logloan dep logterm logtinc white male married fixadj mortlat unit sch
```

Source	SS	df	MS	Number of obs	=	1,961
				F(10, 1950)	=	92.94
Model	151.413674	10	15.1413674	Prob > F	=	0.0000
Residual	317.688568	1,950	.162917214	R-squared	=	0.3228
				Adj R-squared	=	0.3193
Total	469.102241	1,960	.239337878	Root MSE	=	.40363

logloan	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
dep	.0026855	.0090466	0.30	0.767	-.0150565	.0204275
logterm	.1242117	.0262433	4.73	0.000	.0727438	.1756796
logtinc	.3884849	.0162208	23.95	0.000	.3566729	.4202969
white	-.038006	.0260278	-1.46	0.144	-.0890512	.0130392
male	.0944916	.0251543	3.76	0.000	.0451595	.1438237
married	.1476261	.0217551	6.79	0.000	.1049604	.1902918
fixadj	.0473636	.0200731	2.36	0.018	.0079966	.0867305
mortlat	-.0002809	.0546084	-0.01	0.996	-.1073778	.106816
unit	.0887192	.0214193	4.14	0.000	.0467121	.1307263
sch	.108619	.0226261	4.80	0.000	.0642451	.1529929
_cons	.5461052	.2045023	2.67	0.008	.1450392	.9471712

b. Model 3

```
. // multiple regression model 2 (exclude white and mortlat)
. regress logloan dep logterm logtinc male married fixadj unit sch
```

Source	SS	df	MS	Number of obs	=	1,961
				F(8, 1952)	=	115.90
Model	151.065975	8	18.8832469	Prob > F	=	0.0000
Residual	318.036266	1,952	.162928415	R-squared	=	0.3220
				Adj R-squared	=	0.3193
Total	469.102241	1,960	.239337878	Root MSE	=	.40364

logloan	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
dep	.0038143	.0089953	0.42	0.672	-.013827	.0214556
logterm	.1269716	.0261706	4.85	0.000	.0756464	.1782969
logtinc	.3859587	.0161041	23.97	0.000	.3543756	.4175419
male	.092532	.025109	3.69	0.000	.0432888	.1417752
married	.1475187	.0217466	6.78	0.000	.1048696	.1901677
fixadj	.0483134	.0200545	2.41	0.016	.0089831	.0876438
unit	.0928728	.0212293	4.37	0.000	.0512383	.1345072
sch	.1065267	.022581	4.72	0.000	.0622413	.1508122
_cons	.5164374	.2034799	2.54	0.011	.1173767	.9154981