Objective:

Our objective is to obtain an estimate of the predicted impact of not having enough capital to qualify for a conventional loan. FHA loans have lower capital requirements than conventional loans. And looking at the difference between the homes purchased by FHA homebuyers and conventional homebuyers we should be able to get some idea of how impactful it is to be lacking in initial capital. There are several challenges to ensuring the estimate is reliable. Some challenges can be overcome, others can perhaps be overlooked at least for this analysis.

Hypothesis:

In the hypothetical scenario where we present two identical homebuyers who can both qualify for a conventional loan with the opportunity to buy the same house then randomly take away money from one until he can no longer qualify for a conventional loan, we expect that the homebuyer who cannot qualify for the conventional loans will be forced to pay a lower price. Our model should predict a negative coefficient on our FHA financing “treatment” variable.

Data Description:

The data used for this analysis are from the US Census. The [Survey of Construction for 2017](https://www.census.gov/construction/chars/microdata.html) is used as this is the most recent available dataset. This census data included a [detailed codebook](https://www.census.gov/construction/chars/pdf/socmicro_info.pdf) explaining how each variable was recorded. It is worth noting that each observation in this dataset is a [sample of a strata in the housing market](https://www.census.gov/construction/chars/how_the_data_are_collected/) and is meant to sit in for a certain number of new homes sold in an area. The stratified sampling also comes with a sample weight indicating what share of the population of new houses is represented by each sample. Any analysis must be weighted to reflect the relative importance of each observation. Data collected for the Survey of Construction are also truncated at the tops and bottoms of the distribution, but due to how few observations are present at these extremes the truncation is unlikely to have a lot of impact on the analysis. Properties that were recorded as having a zero-dollar value sale price were omitted from the study.

figure A – Histogram of Housing Prices by Financing

Experimental datasets are ideal for this type of analysis, but this dataset is observational and therefore exposed to selection bias. An obvious source of bias can be seen in the distributions in figure A. The distributions of conventional loans are more evenly distributed, and virtually no one towards the higher end of the distribution used FHA financing. This is an indicator that towards the high end of the market observations opt not to use FHA financing in a non-random manner. The data were not collected and designed for this study so there is also a potential for omitted variable bias. Ideally the data would include information about the homebuyer’s financial situations and other characteristics that could better estimate the final price of the home.

Methods:

FHA homebuyers and Conventional homebuyers are on separate distributions. Because the FHA homebuyer distribution is grouped closer to zero than the Conventional homebuyer distribution, the two are not directly comparable. The two distributions are subject to selection bias perhaps because wealthier people are more likely not to use FHA financing and self-select out of FHA financing by having too much capital. However, if we condition the mean prices on other observable characteristics (number of bathrooms and total square feet) we can eliminate some of the bias that is causing our predictor to be overstated. By taking the log of price and looking at price paid in percent terms instead of absolute dollar amounts we can also come to a more useful result. Our conventional financing distribution skews the to the high side. While our estimate of the “treatment” (having less initial capital) would be high and would reflect an average that is only accurate for the wealthy portion of the distribution. Conversely a percent change in price is relevant no matter where you lie on the distribution. Most of the controls are categorical or dummy variables and unique specifications were not use.

Results:

figure B – Summary of results

------------------------------------------------------------

(1) (2) (3)

logFSLPR logFSLPR logFSLPR

------------------------------------------------------------

FIN -0.384\*\*\* -0.0802\*\*\* -0.0934\*\*\*

(-194.47) (-17.37) (-19.85)

...

------------------------------------------------------------

N 376938 56609 56609

R-sq 0.091 0.629 0.608

------------------------------------------------------------

t statistics in parentheses

\* p<0.05, \*\* p<0.01, \*\*\* p<0.001

The summary of the models in figure B show the results of regression models run using the sample weights provided in the data. First results showed that having reduced capital due to the “treatment” of holding less initial capital is a statistically significant result and not just a result produced by chance. The initial result predicts that home prices in the FHA financing group are lower by -38.4% when compared to the conventional financing group, but such results are subject to omitted variable bias. After including control variables, the impact of the treatment coefficient decreased to -8.02% because the control variables were better fitted for explaining the variation in sale price than the “treatment” alone. R-squared also improved substantially, and significance of the “treatment” was not lost. Analysis of the VIF showed that some of the control variables were inflating standard errors and indicating that there could be some collinearity. The data set contains a few fields that could be redundant like counting final square foot of the house as well as number of bedrooms, and so they were removed in the final model. The final model also held significance and the treatment coefficient increased slightly to -9.34%. The results of this model indicate a predicted negative 9.34% price penalty for homebuyers buying a home using FHA financing.

Discussion:

First it should be said that the results presented are far from definitive and it would not be wrong to question these results from several angles. Because of the observational nature of the data we cannot infer causality. We have tried to condition the data on several controls to ensure that the “treatment” and “control” groups are as comparable as possible, but because we only have access to this data set for the analysis there is a large potential for omitted variable bias. Several other variables relating to the identity of the homebuyer like salary, employment history, credit and family background could be better predictors of the price the homebuyer is willing to pay for a home. Including these variables could decrease the impact of the type of financing the homebuyer is using on sale price because FHA is just a stand in for some of these variables. Were these data available it might open the door to an instrumental variable analysis as well. Another area of concern has to do with the way the control variables are specified in the model. More careful specifications of the control variables, including interaction terms or squared or cubed specifications, could change the results of the analysis potentially invalidating the findings.

Perhaps the biggest criticism could be that our observational dataset is comprised of only new homes sold. This distribution of sales used in this analysis is not necessarily representative of the distribution of all home sales. The results, if they were proved to be causal, could only be applied to new homebuyers. This criticism is also the greatest strength of this data set because using objectively collected data on newly built homes frees the data set from attenuation bias that arises from subjective errors related to data collection of existing homes. Existing homes records tend not to be uniform due to unrecorded depreciation, customization, and general updates that have an impact on sale price beyond what can be easily recorded. Using new sales data restricted the scope of the analysis, but benefitted from increased precision.