



## **The Importance of Income on Home Ownership**

*Ajay Raman, Boston University*

Does more income correlate to a higher probability of home ownership? Many relatives and even websites tell you to invest your money in Real Estate. This idea is even deeply rooted in the so-called “American Dream,” where the ideal is a suburban home with a white picket fence. So, one would assume that as households increase their income they will be more likely to purchase a home, but is that really true?

There are two broad explanations for this question. One explanation poses the idea that an increased income is required to buy a home in the first place, thereby making more income increase the chance of ownership of a home. The other explanation is that increases or decreases in income do not significantly change the chance of home ownership, instead there are other unobservable variables that change the chance of home ownership and I believe that home ownership does correlate to a higher household income and the aim of this paper is to use multiple econometric methods to determine whether this is true or not. Also, this paper aims to see how a household’s home ownership probability changes based on their income once race, whether they are in a city or not, and what crisis they are in is factored in. I plan to analyze these variables through a multiple regression model and interaction terms.

After analysis, I found that higher income equates to a higher probability of home ownership in households. This finding proves that home ownership is beneficial for wealth accumulation. I also looked at how race related to this. I found that white households are more likely to own a home compared to non-white households. As income increases, white households become less likely to own a house compared to non-white households. This means that home ownership is a bigger focus for non-white households as they accumulate wealth. Households that lived in metropolitan areas are also less likely to own a home compared to rural households, thus proving that home ownership is more easily available in rural areas compared to urban areas.

Why does this matter though? Owning a home typically means that your income is higher as Table 5 shows us that the average household who owns their home makes 110,033 a year compared to 59425 if they are renting. If my hypothesis is true, Minorities and immigrants who are disadvantaged in America could potentially build generational wealth if they invest in home ownership. Also, I could focus on policy that discourages renting and makes home ownership more accessible for lower-income households.

This paper consists of 9 sections. **Section II** includes a summary and analysis of various papers that talk about relationships like the one I am discussing in this paper. I will specifically talk about how each paper relates to my own paper. **Section III** explains the econometric models that I am running to analyze the relationship between income and home ownership. It also discusses the functional forms I use later in the paper and explains why I picked the variables that I am analyzing. **Section IV** gives a better background to the data I am using and gives a description of the variables used in my analysis. **Section V** is where the result of the analysis is. More specifically, I talk about every result, their importance, and connect some of the data to the literature review. **Section VI** holds the conclusion derived from the results. **Section VII** is the reference section, **Section VIII** is the appendix, and **Section IX** is where the Do-File is.

### **II: Literature Review**

Many papers have researched the relationship between homeownership and wealth. Many of these studies focus on the implications of low-income households to home ownership, particularly about low income minorities' relationship with home ownership. Another thing that many of these papers accounted for was time and how the housing market played a role in the value of home ownership for low-income households. Overall, the main conclusion was that home ownership did benefit low-income households.

Christopher E. Herbert, et al. (2013) considered the decline in housing prices and recessions while conducting their analysis. They specifically also used regression analysis to conduct their research. Their findings showed that home ownership is a significant part of household wealth and is still very important for growing a household's wealth. This paper concludes that low-income households should be given more home ownership opportunities because over time owning a home will increase their wealth. A couple of significant differences from this paper is that Christopher E. Herbert, et al. are looking at this relationship over a much larger time scale of 1999 to 2009 and look at the mechanisms that cause wealth through homeownership. These mechanisms are home appreciation and the large increase in savings that happens when homeownership occurs. The second mechanism is related to this paper because savings occurred more in higher income households. These conclusions are both relevant to my paper because of their relation to my analysis of race, income, and home ownership probability. I will also test Herbert's conclusions in my own paper by looking at how being a white household affects your probability of owning a house compared to a non-white household.

David O'Neil (2021) considered a similar idea but focused more on low-income minorities and their relationship to home ownership during the business cycle. He looked at from 2006 to 2018 which is a much bigger scope of time compared to my paper. He also specifically used regression analysis to conduct his research. He found that being from an African American descent reduces your chances of owning a home by a substantial amount and that single African American citizens are having less and less of a chance of being homeowners as time passes. One other interesting idea that O'Neil discovered was the effect the recession had on the mobility of households with regards to home ownership. I do not delve as much into this idea in my paper, but O'Neil's findings provide a lot of background for some of the analysis I am doing. The fact that African Americans are having lower and lower chances of being homeowners is indicative of a wealth disparity between races. I believe so because of Herbert's conclusions and will also hope to back up my hypothesis further through my own multiple regression. Overall, this paper has a slightly different premise than mine but still is pretty closely related.

Thomas P. Boehm and Alan Schlottmann (2004) also researched the relationship between home ownership and income inequality but focused more on effects over time while also controlling for things such as race. They use regression analysis in their research but also use many other methods like dynamic models and transition matrices. They mainly focus on a data set that was collected between 1984-1992 but also bring up other data in passing. Some interesting conclusions that they came to were that high-income white households' average median level of non-housing wealth accumulation is \$2,650, while low-income minority household's is \$0. This is interesting to me because it could potentially answer the question of why there are negative total household incomes in this paper's dataset. This conclusion is also important to consider in

general because it could also explain the correlation between home ownership and income. Thomas P. Boehm and Alan Schlottmann also found that homeowners often transition back to renting after some time. This is interesting and something this paper currently does not touch on but could potentially implement in the future. Thomas P. Boehm and Alan Schlottmann also found that white high-income households' transition back to ownership a lot more than low-income minority households which is seemingly obvious but an interesting conclusion, nonetheless. One distinction that Thomas P. Boehm and Alan Schlottmann focus on is the differences between white households and minority households. They find a lot of inequality between the two different groups which is something I discuss further in this paper. Overall, this paper delves a lot deeper into the relationship between low-income households and home ownership, specifically how they interact with renting and buying properties. It can provide a lot of reasoning behind my research and the reason why low-income households don't buy properties even if that point may seem obvious.

Elizabeth Kneebone and Mark Trainer (2019) also researched home ownership's relationship to metropolitan areas and their research provides helpful insights to this paper's analysis. Elizabeth Kneebone and Mark Trainer did not use regression analysis in their research so they I am not able to get any useful information about how to compare these variables, but their results do give some helpful background information about the relationship between location and home ownership probability. Elizabeth Kneebone and Mark Trainer found that housing production in the US has slowed down in recent years and that because of housing prices increasing, the housing market has become constrained. This is important to this paper because it provides another viewpoint for the relationship between location and home ownership probability. With this information, datasets taken from the COVID crisis could have a negative impact on home ownership probability compared to the earlier years of the housing crisis.

### III: Econometric Model

#### *Simple Linear Regression Models*

To put it simply, my goal is to see whether a household's income affects their probability of ownership status. My first linear regression model was made to be a base to then create many other functional forms out of. The purpose of doing this is to find the model that is best fit to the data in the dataset. Because my y variable, *ownershp\_dum*, is a dummy variable, I can test a quadratic, cubic, and linear-log functional form while still being able to compare each functional form to one and other. I can compare these functional forms to each other because the y variable will still have the same data points.

$$(1) \text{ownershp\_dum} = \beta_1 + \beta_2 \text{hhincome} + e$$

$$(2) \text{ownershp\_dum} = \beta_1 + \beta_2 \text{hhincome}^2 + e$$

$$(3) \text{ownershp\_dum} = \beta_1 + \beta_2 \text{hhincome}^3 + e$$

$$(4) \text{ownershp\_dum} = \beta_1 + \beta_2 \text{hhincome\_log} + e$$

## Importance of Home Ownership

This paper plans to analyze the relationship found between *ownership\_dum* and *hhincome* and provide  $\beta_1$  and  $\beta_2$  numbers. I am also planning to test many functional forms in the econometric model. I am using income squared, income cubed, and the log of income to run different regressions to try and minimize the sum of squared errors in the model. After running both regressions, I will compare all three of their SSEs to see which is the best functional form for this model. Model 1 - my simple linear regression equation - estimates how the probability of a particular household owning a home changes when their income increases by \$1. Model 2 - my quadratic model - estimates how the probability of owning a home changes when a household's income *squared* increases by \$1. The  $y$  vs  $x^2$  functional form found in Model 2 is used to account for a parabolic relationship. Model 3 - my cubic model - estimates how the probability of owning a home changes when a household's income *cubed* increases by \$1. The  $y$  vs  $x^3$  functional form found in Model 3 is to account for a cubic function. Model 4 - my fourth simple regression equation - estimates how the probability of owning a home changes when a household's log of income increases by \$1. The  $y$  vs  $\log(x)$  functional form found in Model 4 is used to account for a logarithmic function.

### *Multiple regression Models*

I will also be testing two different multiple regression models to control for endogeneity and to look at some other smaller relationships. I used a linear multiple regression and a linear-log multiple regression because these two models were best fit for my simple regression model. I will include the parameters *metro\_a*, which describes whether the household is in a metropolitan area and *white*, which analyzes whether the household is of African American descent. I will also be adding many time variables to analyze how different crises affect the probability of ownership status.

$$(5) \text{ownership\_dum} = \beta_1 + \beta_2 \text{hhincome} + \beta_3 \text{white} + \beta_4 \text{metro\_a} + \beta_5 \text{time2008} + \beta_6 \text{time2009} + \beta_7 \text{time2010} + \beta_8 \text{time2020} + \beta_9 \text{time2021} + e$$

$$(6) \text{ownership\_dum} = \beta_1 + \beta_2 \text{hhincome\_log} + \beta_3 \text{white} + \beta_4 \text{metro\_a} + \beta_5 \text{time2008} + \beta_6 \text{time2009} + \beta_7 \text{time2010} + \beta_8 \text{time2020} + \beta_9 \text{time2021} + e$$

*White*, *metro\_a*, and the different time variables are all dummy variables. I chose the *white* variable because I believe that being white increases the chances of owning a home. My reason for believing this is because of one of the papers that I reviewed which talked about how white people are more likely to buy a house and are also more likely to switch to home owning from renting compared to other races. I chose the *metro\_a* variable because I wanted to see how living in a metro area affects a household's home ownership probability, if at all. Lastly, I chose the time variables because I wanted to see how much a crisis affects a household's home ownership percentage. I decided to take data from the housing crisis of 2008 and the COVID crisis of 2020. I specifically chose those two because one had to do with housing prices while the other did not. I want to see if the fact that one crisis was specifically about housing made a noticeable difference on the home ownership probability. Most importantly, I also chose the years when a recession took place to compare it with normal years. These normal years of 2007, 2019, and 2022 are not included in the dummy variables I added to the multiple regressions.

### *Interaction Between Income and Race*

$$(7) \text{ownership\_dum} = \beta_1 + \beta_2 \text{hhincome} + \beta_3 \text{white} + \beta_4 \text{metro\_a} + \beta_5 \text{time2008} + \beta_6 \text{time2009} + \beta_7 \text{time2010} + \beta_8 \text{time2020} + \beta_9 \text{time2021} + \beta_{10} \text{hhincome\_white} + e$$

Model 7 introduces the interaction term *hhincome\_white* which is the product between *hhincome* and the dummy variable *white*. If *white* equals one that means the respondent is white and if it is 0 it means that the respondent is some other non-white race. This is an interaction between a dummy and continuous variable. I created this new variable to analyze whether the effect of income on home ownership status varies by race, specifically whether the person is white or not. My goal is to see whether higher earning white people are more likely to own a house compared to higher earning people of other races. I also want to see if that difference varies with income level.

#### *Interaction Between Income and Location*

$$(8) \text{ownership\_dum} = \beta_1 + \beta_2 \text{hhincome} + \beta_3 \text{white} + \beta_4 \text{metro\_a} + \beta_5 \text{time2008} + \beta_6 \text{time2009} + \beta_7 \text{time2010} + \beta_8 \text{time2020} + \beta_9 \text{time2021} + \beta_{10} \text{hhincome\_metro\_a} + e$$

Model 8 introduces the interaction term *hhincome\_metro\_a* which is the product between *hhincome* and the dummy variable *metro\_a*. If *metro\_a* equals one that means the respondent lives in a metropolitan area, including towns that neighbor a city, and if it is 0 it means that the respondent lives in some rural area. This is an interaction between a dummy and continuous variable. I created this new variable to analyze whether the effect of income on home ownership status varies by a household's location, specifically whether the person lives near a city or not. My goal is to determine whether higher earning households in metropolitan areas are more likely to own a house compared to higher earning households in rural areas. I also want to see if that difference varies with income level.

## **IV: Data Set Information**

### *Overall Summary*

The data used in this study are from the Current Population Survey (CPS), and are from the years 2007, 2008, 2009, 2010, 2020, 2021, 2022. The total amount of observations I have after cleaning is 1,222,238, which can be seen in Table 1.

### *Simple Regression Model Variables*

In the simple regression models, I use variables *hhincome*, which stands for total household income, and *ownership\_dum*, which stands for home ownership status. I cleaned each of these variables to make the analysis more precise. For *hhincome*, I only took out responses that did not report their income. I did not take out negative income data points because people who make a negative income probably do not own a house so they should be part of this study. For *ownership\_dum*, I decided to take out all the NIU terms because people who were NIU either did not want to respond or also might not be owning or renting and could be doing many things like they could be homeless or in assisted living. But for the importance of this regression, they would not be important, so I am also taking them out. I could later look at the income of people who were NIU on ownership to see a possible trend and could talk about why these people were

## Importance of Home Ownership

NIU for ownership. Another thing I did was make *ownership\_dum* a dummy variable before beginning the regression analysis. To make ownership a dummy variable from *ownership*. I did this by categorizing each response from *ownership* into a binary format. There were three responses in *ownership*: owned or being bought, rent with cash, and rent without cash. I categorized the first response with one and the rest of them with zero because this paper is analyzing how home ownership affects income so renting with cash or without cash does not really matter in this analysis. I could look further into the difference between cash and no cash rent and the implications it has onto a household's income but for right this paper categorizes them as the same.

### *Multiple Regression Model Variables*

In the multiple regression models, in addition to the *hhincome* and *ownership\_dum* variables, I added *white*, *metro\_a*, *time2008*, *time2009*, *time2010*, *time2020*, *time2021*, and *time2021* variables to help account for endogeneity. To clean the *race* and *metro* variables, I dropped all the NIU data points and turned both of those variables in *white* and *metro\_a* dummy variables respectively. When *white* equals one, it means that the household is of white origin, instead, if *white* is equal to zero, it means that the household is some other non-white race. When *metro* equals 1, the household is in a central city or is right outside of a central city. I decided to include the latter because many housing prices are inflated in suburbs next to major cities. When *metro* equals 0, the household does not live in or near a metropolitan area. For every time variable, I created a dummy variable and assigned 1 whenever the variable took place in the year associated with the dummy variable. For example, for *time2008*, it was equal to one whenever the *year* variable was equal to 2008 and was equal to zero in every other datapoint.

### *Descriptive statistics of Variables Together*

The average household respondent is white, lives in a metropolitan area, and makes around \$94,719. This can be found on Table 2 in my appendix. These values are found by looking at the mean of all the variables I use in my regression analysis. An average household is white and lives in a metropolitan area because the dummy variables *white* and *metro\_a* are both more than 0.7 indicating that the average household respondent will probably be white and live in the city.

## **V: Description of results**

### *Interpretation of Models*

Out of all the functional forms, Models 1 and 4 were the best because they had the lowest SSEs. The SSE stands for the sum of squared errors and is one way to determine how good a functional form fits the dataset. If the SSE is small, it means that the line of best fit represents the trend of the dataset well and vice versa applies. On Model 1, the  $y$  vs  $x$  functional form has an SSE of 244938.284. On Model 2, the  $y$  vs  $x^2$  functional form has an SSE of 256376.056 which is more than the  $y$  vs  $x$  functional form. On Model 3, the  $y$  vs  $x^3$  functional form has an SSE of 257756.469, which is the highest SSE of all of the functional forms. On Model 4, the  $y$  vs  $\log(x)$  functional form has a SSE of 229575.856, which is the lowest SSE of all of the functional forms. Model 4 has a slightly smaller SSE compared to Model 1 but Model 1 is easier to interpret because it is linear so I will be using that model when I introduce interaction terms. Model 5 is

the multiple regression model which adds on *white*, *metro\_a*, *time2008*, *time2009*, *time2010*, *time2020*, *time2021*, and *time2022*. Model 6 is the lin-log multiple regression model which adds on the same variables. The linear and lin-log functional forms were only used for the multiple regression model because Models 1 and 4 were the functional forms with the lowest SSE.

Even though I know that the linear and linear-log regression are the most accurate functional forms, I still need to see whether the  $b_1$  and the  $b_2$  variables are BLUE (best linear unbiased estimators). To find this out, I must see if the four OLS assumptions hold true. For the first assumption, I know that the mean of the predicted error is  $-3.40 \times 10^{-11}$ , which is very close to zero. Because the mean of the predicted error is close to zero, I can say that the first assumption is true. For the second assumption, I would have to look at a scatter plot between *hhincome* and *ownershp\_dum* but I know that *ownershp\_dum* is a dummy variable so the graph would basically not be readable. But in general, every linear regression has heteroskedasticity because of how real-world data functions so I will assume that the second assumption is violated. For the third assumption, autocorrelation can be disregarded because I am using data from two different time periods, 2008-2010 and 2019-2022. For the fourth assumption, the skewness is  $-0.806$  and the kurtosis is  $1.834$ . The skewness is a little above  $-0.5$  so it is moderately skewed but because of the central limit theorem this distribution is still normal. The kurtosis is under two, so I do not have too many outliers. Because  $b_1$  and  $b_2$  failed the second assumption they are not the best unbiased estimators. I will try to make up for this by introducing more variables in the multiple regression models though.

#### *Simple Linear Regression Model (Model 1, Table 4)*

My simple linear regression model shows that in the data sample, when income is \$0, the  $b_1$  is equal to 0.603, which means that the probability of a household owning a home when their income is \$0 is equal to 0.603. The  $b_2$  is equal to  $9.96 \times 10^{-7}$ , which means that when income increases by \$100,000 dollars per year, then the probability of home ownership increases by 0.0996. The elasticity is 0.13527, meaning when income increases by 1% the probability of household home ownership increases by 0.13527%, evaluated at the mean. The semi elasticity is equal to 0.00014, meaning that when income increases by \$10,000, the probability of household home ownership increases by 1.4%, evaluated at the mean.

#### *Simple Quadratic Linear Regression Model (Model 2, Table 4)*

My simple quadratic regression model shows that in the data sample, when income is \$0, the  $b_1$  is equal to 0.690, which means that the probability of a household owning a home when their income is \$0 is equal to 0.690. The  $b_2$  is equal to  $3.74 \times 10^{-13}$ , which means that when income squared increases by \$100,000 dollars per year, then the probability of home ownership increases by 0.0000000374. The slope of this regression line is  $7.08498 \times 10^{-8}$ , meaning that when income increases by \$100,000, the probability of household home ownership increases by approximately 0.00708498. The elasticity is 0.00962, meaning when income increases by 1% the probability of household home ownership increases by 0.00962%, evaluated at the mean. The semi elasticity is equal to 0.00001, meaning that when income increases by \$10,000, the probability of household home ownership increases by 0.1%, evaluated at the mean.

## Importance of Home Ownership

### *Simple Cubic Linear Regression Model (Model 3, Table 4)*

My simple quadratic regression model shows that in the data sample, when income is \$0, the  $b_1$  is equal to 0.696, which means that the probability of a household owning a home when their income is \$0 is equal to 0.696. The  $b_2$  is equal to  $7.53 \times 10^{-20}$ , which means that when income cubed increases by \$100,000,000,000 dollars per year, then the probability of home ownership increases by 0.00000000753. The slope of this regression line is  $2.02671 \times 10^{-9}$ , meaning that when income increases by \$100,000, the probability of household home ownership increases by approximately 0.000202671. The elasticity is 0.00027, meaning when income increases by 1% the probability of household home ownership increases by 0.00027%, evaluated at the mean. The semi elasticity is equal to  $2.9061 \times 10^{-7}$ , meaning that when income increases by \$1,000,000, the probability of household home ownership increases by 0.29061%, evaluated at the mean.

### *Simple Linear-Log Regression Model (Model 4, Table 4)*

One interesting quirk with this sample regression line is that its y intercept is negative. The  $b_1$  is -0.6006403, meaning that when `hhincome_log` is at 0, or `hhincome` is at 1, `ownership_dum` equals -0.6. This means that when household income is at 1, the probability of owning a house is -0.6. This makes no sense if household income is 1 because a probability cannot be negative. This means that this regression will only be valid if `ownership_dum` is greater than or equal to 0. Another way to interpret the  $b_2$  is that when household income increases by one percent the chance of owning a home increases by 0.15%. Another notable description is what the slope of the line is at the mean. The slope of the line at the mean is  $9.75089946 \times 10^{-7}$ . An interpretation of this is when household income increases by 1 dollar, the percentage chance of owning a house increases by  $9.75089946 \times 10^{-7}$  percent at the mean. The elasticity at the mean is 0.16552. This means that when household income increases by one percent the percentage chance of owning a house increases by 0.16552% at the mean. These results show that income does greatly affect the percentage of home ownership in a household.

### *Multiple Linear Regression (Model 5, Table 4)*

The multiple linear regression model accounts for variables other than income on home ownership status, therefore yielding different results than the simple regression models. The  $b_1$  is equal to approximately 0.5602. The interpretation behind this value is that it is the probability of home ownership in a household that is not white, does not live in a metro area, has no income, and lives in 2007, 2019, or 2022. This interpretation is not realistic because household's usually have income of some kind. The  $b_2$  is  $1.03 \times 10^{-6}$ , meaning that when household income increases by \$100,000, the probability of household home ownership increases by 0.103, *ceteris paribus*. The  $b_3$  is approximately 0.140, meaning that the probability of household home ownership in a white household is 0.140 more than a non-white household, *ceteris paribus*. This result backs up the hypothesis I made in the literature review on Thomas P. Boehm and Alan Schlottmann paper. I will analyze this relationship further in my interaction term section. The  $b_4$  is approximately -0.097, meaning that the probability of household home ownership in a household that lives in a metropolitan area is 0.097 less than a household that lives in a rural area, *ceteris paribus*. This is substantial because it means that living in a rural area increases your chances of owning a house by almost 10%, which is significant. The  $b_5$  is approximately 0.033, meaning that the probability of home ownership in a household in 2008 is 0.033 more than a



household in 2007, 2019, or 2022. The  $b_6$  is approximately 0.025, meaning that the probability of home ownership in a household in 2009 is 0.025 more than a household in 2007, 2019, or 2022. The  $b_7$  is approximately 0.013, meaning that the probability of home ownership in a household in 2010 is 0.013 more than a household in 2007, 2019, or 2022. The  $b_8$  is approximately -0.009, meaning that the probability of home ownership in a household in 2020 is 0.009 less than a household in 2007, 2019, or 2022. The  $b_9$  is approximately -0.0166, meaning that the probability of home ownership in a household in 2021 is 0.016 less than a household in 2007, 2019, or 2022. By looking at the  $R^2$  one sees that approximately 7.8% of the variation on household home ownership probability is explained by income, race, metropolitan area, and time. I can compare this number with the linear log figure to determine which functional form is more accurate.

#### *Multiple Linear-Log Regression Model (Model 6, Table 4)*

The multiple linear-log regression model accounts for variables other than income on home ownership status, therefore yielding different results than the simple regression models. The  $b_1$  is equal to approximately -0.899. The interpretation behind this value is that it is the probability of home ownership in a household that is not white, does not live in a metro area, has no income, and lives in 2007, 2019, or 2022. This interpretation is not realistic because household's usually have income of some kind. It is also not realistic because it is negative, and a probability cannot be negative. The  $b_2$  is 0.143, meaning that when the log of household income increases by \$1, the probability of household home ownership increases by 0.143, ceteris paribus. The  $b_3$  is approximately 0.116, meaning that the probability of household home ownership in a white household is 0.116 more than a non-white household, ceteris paribus. The  $b_4$  is approximately -0.108, meaning that the probability of household home ownership in a household that lives in a metropolitan area is 0.108 less than a household that lives in a rural area, ceteris paribus. This is substantial because it means that living in a rural area increases your chances of owning a house by more than 10%, which is significant. The  $b_5$  is approximately 0.036, meaning that the probability of home ownership in a household in 2008 is 0.036 more than a household in 2007, 2019, or 2022. The  $b_6$  is approximately 0.029, meaning that the probability of home ownership in a household in 2009 is 0.029 more than a household in 2007, 2019, or 2022. The  $b_7$  is approximately 0.018, meaning that the probability of home ownership in a household in 2010 is 0.018 more than a household in 2007, 2019, or 2022. The  $b_8$  is approximately -0.015, meaning that the probability of home ownership in a household in 2020 is 0.009 less than a household in 2007, 2019, or 2022. The  $b_9$  is approximately -0.021, meaning that the probability of home ownership in a household in 2021 is 0.021 less than a household in 2007, 2019, or 2022. By looking at the  $R^2$  one can see that approximately 12% of the variation on home ownership probability is explained by income, race, metropolitan area, and time. This  $R^2$  figure is slightly more than the multiple linear regression model, but I will still be using the multiple linear regression model for the interaction terms because they are a lot easier to interpret.

#### *Income Interactions with Race (Model 7, Table 5)*

I wanted to see the marginal effect of being a white household on the probability of home ownership. To do this, I took the derivative of the regression line and subtracted the version where  $white = 0$  from when  $white = 1$ . I got the equation  $0.174 - 0.000000388 * hhincome$ . Interpreted this means that when  $hhincome = \$0$ , white households are 0.174 more likely to own

## Importance of Home Ownership

a home compared to a non-white household. Also, with every \$100,000 increase in income, white households become 0.039 less likely to own a house compared to non-white households. This result is interesting because it says that white households become less likely to own a house as they become wealthier compared to non-white households. This means that home ownership is a bigger focus for non-white households as they accumulate wealth.

I also wanted to find the marginal effect of income on home ownership with respect to race. I found that the effect of a \$100,000 increase in a white household is a 0.0942 increase in probability of home ownership. For non-white households, the effect of a \$100,000 increase equates to a 0.133 increase in the probability of home ownership. This further proves my point that having a higher income is much more important for non-white households compared to white households.

### *Income Interactions with Location (Model 8, Table 5)*

I wanted to see the marginal effect of being a household in the metro area on the probability of home ownership. To do this, I took the derivative of the regression line and subtracted the version where  $metro\_a = 0$  from when  $metro\_a = 1$ . I got the equation  $-0.085 - 0.00000015 * hhincome$ . Interpreted this means that when  $hhincome = \$0$ , households in the metropolitan area are 0.085 less likely to own a home compared to a household living in a rural area. Also, with every \$100,000 increase in income, households in metropolitan areas become 0.015 less likely to own a house compared to rural households. This result proves that households in a metropolitan area have a lower chance of home ownership compared to households in rural areas. Logically, this makes sense because houses in rural areas tend to cost less than real estate in and around metropolitan areas.

I also wanted to find the marginal effect of income on home ownership with respect to the area the household resides in. I found that the effect of a \$100,000 increase in a household living in the metropolitan area is a 0.101 increase in probability of home ownership. For non-white households, the effect of a \$100,000 increase equates to a 0.116 increase in the probability of home ownership. This further proves my point that households residing in a metropolitan area have a harder time owning homes compared to households in a rural area.

*Simple Linear Regression Hypothesis Test:*

The hypothesis test is performed on Model 1, a simple linear regression model. I decided to use 0.00000098 as my conjecture because I wanted to pick a value that had a great effect on the probability of home ownership and this conjecture did as it meant that for every \$100,000 increase in income the probability of that household to own a home increased by 0.098. I am using a 95% confidence interval and an  $\alpha$  of 0.05 significance in this hypothesis test. The test is as follows:

$$H_0 : \beta_2 \leq 0.00000098$$

$$H_1 : \beta_2 > 0.00000098$$

Using STATA, I found that the t-stat is 3.99 while the critical value for  $\beta_2$  is 1.645. Since the t-stat is larger than the critical value, I reject the null hypothesis and conclude that the effect of income has on the probability of home ownership is greater than 0.00000098 per dollar of income.

**VI: Conclusion**

My original question is to examine the relationship between income and home ownership status across time. I have shown that income is positively correlated to home ownership status. This meant that the more income a household had, the higher of a probability it would have of home ownership.

Another aim of this paper is to see how a household's home ownership probability changes based on their income on their race, whether they are in a city or not, and what crisis they are in is factored in. I found that white households were more likely compared to own a house compared to a non-white household, but as their income increased, non-white household's probability of home ownership increased at a greater rate than white households. These results were pretty intriguing and showed that home ownership was much more important to non-white households as they gained wealth compared to white households. These findings back up David O'Neil's paper, where he analyzed how low-income black households compared to low-income white households in their probabilities of home ownership.

I also found that households that live in a metropolitan area have a lower probability of owning a house compared to a household living in a rural area. These results made sense logically but also showed us that this effect is exasperated by an increase in income in the household.

One interesting conclusion I came to had to do with how different crises impacted a household's home ownership probability. My original hypothesis was that the 2008 housing crisis will have a negative impact on the probability of home ownership compared to the COVID-19 crisis but my results disproved this hypothesis. In fact, it was the exact opposite. In the years 2020 and 2021, the probability of home ownership was down, while in 2008-2010 the probability of home ownership increased compared to 2007, 2019, and 2022 ceteris paribus. The reason for this discrepancy may be because home ownership may be down so people have to resort to renting

## Importance of Home Ownership

properties. This hypothesis makes sense when you look at the Elizabeth Kneebone and Mark Trainer paper (2019).

Some limitations in this study have to do with the fact that I am using an OLS regression model. If I was able to use a logistic regression model, I may have been able to reduce some of the error that I have currently by using the OLS linear regression model. Also, when using the interaction term model, it would have been slightly more accurate to do a linear-log regression model. The reason I did not do this though is because interpreting this model is very complicated compared to the linear model.

Overall, I have answered the question of whether income is positively correlated to home ownership status in this paper. I have also analyzed how a household's race, location, and time influences their probability of home ownership. These findings are significant, but they do not prove causality. These findings only show correlation between these variables. Even though I tried to control for endogeneity by adding a lot of other variables, there are still a lot of other factors that contribute to probability of home ownership that I have not considered.

## VII: References

Boehm, Thomas P, and Alan Schlottmann. "Wealth Accumulation and Homeownership: Evidence for Low-Income Households." *Wealth Accumulation and Homeownership: Evidence for Low-Income Households* | HUD USER, Abt Associates Inc., Dec. 2004, <https://www.huduser.gov/portal/publications/HOMEOWN/WAccuNHomeOwn.html>.

Herbert, Christopher, et al. "Is Homeownership Still an Effective Means of Building Wealth for Low-income and Minority Households? (Was It Ever?)." *Is Homeownership Still an Effective Means of Building Wealth for Low-income and Minority Households? (Was It Ever?)*, Sept. 2013, [www.jchs.harvard.edu/sites/default/files/hbti-06.pdf](http://www.jchs.harvard.edu/sites/default/files/hbti-06.pdf).

Kneebone, Elizabeth, and Mark Trainer. "How Housing Supply Shapes Access to Entry-Level Homeownership." *Terner Center for Housing Innovation*, UC Berkeley, Nov. 2019, [https://ternercenter.berkeley.edu/wp-content/uploads/pdfs/How\\_Housing\\_Supply\\_Shapes\\_Access\\_to\\_Entry-Level\\_Homeownership\\_2019.pdf](https://ternercenter.berkeley.edu/wp-content/uploads/pdfs/How_Housing_Supply_Shapes_Access_to_Entry-Level_Homeownership_2019.pdf).

O'Neil, David. "'Homeownership Trends of Low-Income Americans Throughout the Current Bu' by David O'Neil." *"Homeownership Trends of Low-Income Americans Throughout the Current Business Cycle,"* by David O'Neil, 1 Sept. 2021, <https://digitalcommons.iwu.edu/parkplace/vol28/iss1/12>

## VIII: Appendix

**Table 1: Description of Variables**

<b>Variable Name</b>	<b>Variable Description</b>
hhincome	Total Household Income
ownersh	Variable which provides information about if a household owns their home, rents with cash, or rents without cash.
ownersh_dum	Dummy variable which is 1 if the household owns their home and 0 if the household rents their home.
race	Variable that provides information concerning the race of the inhabitants of a household.
white	Dummy variable which is 1 if the household is of white origin and 0 if the household is of non-white origin.
metro	Variable which provides information about if a household lives in a city, near a city, or in a rural area.
metro_a	Dummy variable which is 1 if the household lives in and around a metropolitan area and 0 if the household lives in a rural area
time2008	Dummy variable which is 1 if the household is reporting data in 2008 and 0 if the household is reporting data from another year

## Importance of Home Ownership

time2009	Dummy variable which is 1 if the household is reporting data in 2009 and 0 if the household is reporting data from another year
time2010	Dummy variable which is 1 if the household is reporting data in 2010 and 0 if the household is reporting data from another year
time2020	Dummy variable which is 1 if the household is reporting data in 2020 and 0 if the household is reporting data from another year
time2021	Dummy variable which is 1 if the household is reporting data in 2021 and 0 if the household is reporting data from another year
time2022	Dummy variable which is 1 if the household is reporting data in 2022 and 0 if the household is reporting data from another year

**Table 2: Descriptive Statistics of the Variables**

Summary Statistics					
VARIABLES	(1) N	(2) mean	(3) sd	(4) min	(5) max
hhincome	1.222e+06	94,719	103,572	-31,941	2.990e+06
ownersh_p_dum	1.222e+06	0.697	0.459	0	1
white	1.222e+06	0.768	0.422	0	1
metro_a	1.222e+06	0.772	0.419	0	1

**Table 3: Descriptive Statistics according to Ownership Status**

Statistics for households that own their houses and households that rent their houses							
VARIABLE	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	<u>ownersh_p_dum 0</u> N	mean	sd	<u>ownersh_p_dum 1</u> N	mean	sd	Difference in mean
<u>hhincome</u>	369,853	59,425	70,007	852,385	110,033	111,715	50,608
<u>white</u>	369,853	0.671	0.470	852,385	0.810	0.392	0.139
<u>metro_a</u>	369,853	0.821	0.384	852,385	0.751	0.432	-0.07

# Importance of Home Ownership

**Table 4: Regression Models**

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 5	(6) Model 6
<u>hhincome</u>	9.96e-07*** (3.91e-09)				1.03e-06*** (3.92e-09)	
<u>white</u>					0.140*** (0.000953)	0.116*** (0.000938)
<u>metro_a</u>					-0.0967*** (0.000964)	-0.108*** (0.000943)
<u>time2008</u>					0.0329*** (0.00126)	0.0359*** (0.00123)
<u>time2009</u>					0.0252*** (0.00126)	0.0287*** (0.00123)
<u>time2010</u>					0.0128*** (0.00126)	0.0183*** (0.00123)
<u>time2020</u>					-0.00852*** (0.00139)	-0.0151*** (0.00136)
<u>time2021</u>					-0.0166*** (0.00137)	-0.0206*** (0.00134)
<u>hhincome_sq</u>		0*** (0)				
<u>hhincome_cu</u>			0*** (0)			
<u>hhincome_log</u>				0.141*** (0.000392)		0.143*** (0.000394)
Constant	0.603*** (0.000549)	0.690*** (0.000423)	0.697*** (0.000416)	-0.856*** (0.00435)	0.560*** (0.00130)	-0.899*** (0.00439)
Observations	1,222,238	1,222,238	1,222,238	1,210,597	1,222,238	1,210,597
R-squared	0.050	0.006	0.001	0.096	0.079	0.122

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



Table 5: Additional Regression Models

VARIABLES	(1) Model 7	(2) Model 8
hhincome	1.33e-06*** (8.40e-09)	1.16e-06*** (1.15e-08)
white	0.174*** (0.00126)	0.139*** (0.000953)
metro_a	-0.0964*** (0.000963)	-0.0854*** (0.00133)
time2008	0.0332*** (0.00126)	0.0330*** (0.00126)
time2009	0.0255*** (0.00126)	0.0253*** (0.00126)
time2010	0.0130*** (0.00126)	0.0129*** (0.00126)
time2020	-0.00871*** (0.00139)	-0.00860*** (0.00139)
time2021	-0.0169*** (0.00137)	-0.0167*** (0.00137)
hhincome_white	-3.88e-07*** (9.43e-09)	
hhincome_metro_a		-1.50e-07*** (1.22e-08)
Constant	0.534*** (0.00145)	0.551*** (0.00151)
Observations	1,222,238	1,222,238
R-squared	0.080	0.079

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