Estimating the Effects of Housing Burdens on Educational Attainment

Names

- Nyan Aye
- Serene Xie
- Angela Shen
- Bofu Zou
- Kyounghyun So

Research Question

Do household housing cost burdens in California predict educational attainment in Californian adults?

Background

Housing costs have always been a rising concern in California, especially in densely populated major cities such as Los Angeles and San Francisco. In fact, as the population grows and the housing supply is unable to keep up, California has become the third most expensive state to live in as of 2024. As housing costs grow, households become more unlikely to spend on other matters, such as education, which will be the main focus of our study. The goal of our study is to find out if we can draw a correlation between the percentage monthly household income paid towards housing costs versus the percentage of adults in an area pursuing a 4-year college degree or higher. In broader strokes, we hope that finding said correlation would display the importance of housing issues by showing such issues can alleviate other issues, starting with education, in California.

Previous Research

There is previous research that discusses topics similar to this one. One such example is a study conducted by A Habitat for Humanity U.S. Research and Measurement Team that shows that there is a negative correlation between household income and children's educational outcome, measured by attendance[1].

[1] 21-81776_RD_EvidenceBrief-6-Education_FASH-lores_1.pdf (habitat.org)

Hypothesis

There exists a correlation between the percentage of households with ‘high’ household burden (50% of household income goes towards housing) and the percentage of individuals in the area pursuing a 4-year college degree or higher within the same area.

Data

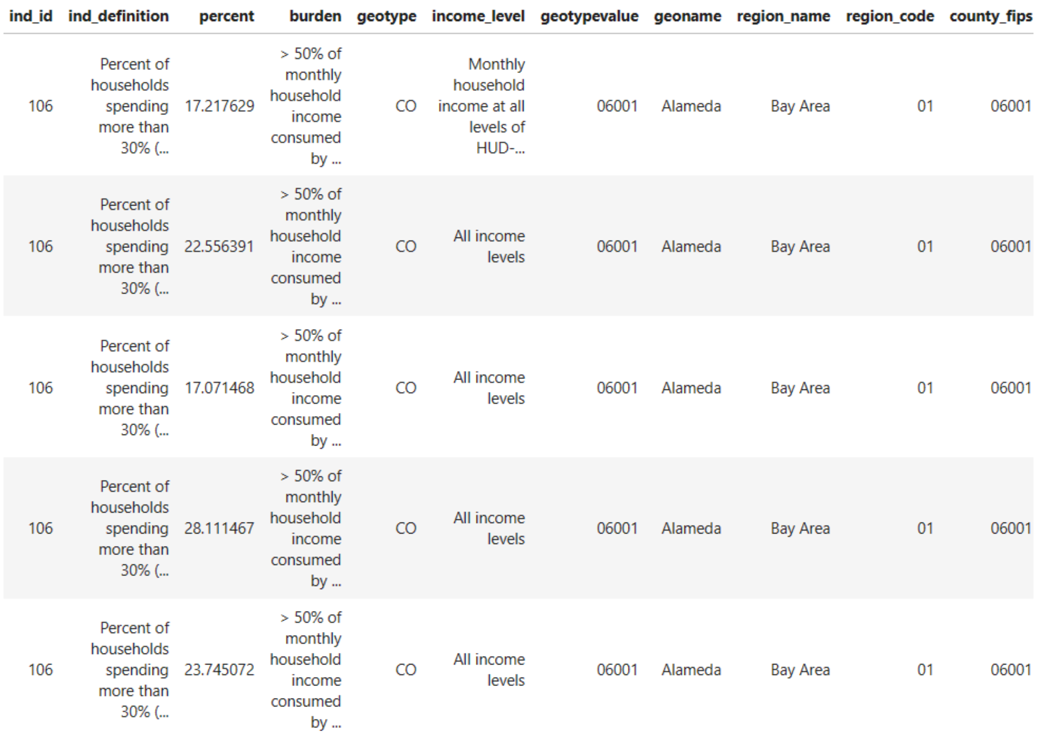
We found two datasets for this project from the data provided by the California government. The first one is about California housing cost burdens[2], with a focus on the following variables: percentage of income burden in household, county and region. The second dataset was the educational attainment dataset[3], with variables county, region, and estimated percentage of educational attainment in the region. Each dataset has at least 1000 observations to be able to draw accurate estimations, with the housing cost burdens dataset utilizing a sample size of 521,264 and education attainment dataset utilizing a sample size of 166,662. We want to draw a present correlation between housing cost and education from previous historical trends, so both datasets will have a temporal coverage of 2006-2010.

Data Cleaning

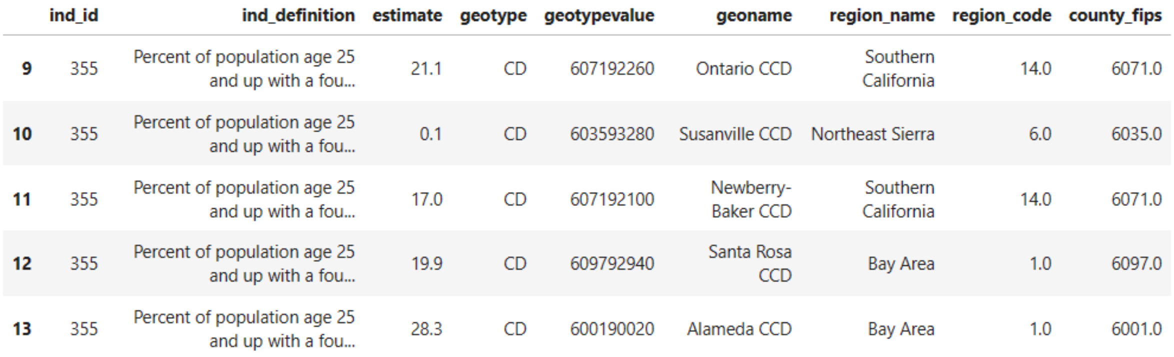
To make our data usable, we cleaned the data based on our needs. For the educational attainment dataset, we removed the first few rows that aggregate education level by race, since this is not relevant to our goals. For the housing burden dataset, we only kept observations classified under “>50% of monthly income…” and “all income levels/monthly income at all levels…income”. We also removed some columns from the table and only kept relevant data such as estimates since the dataset is too large. We also used a function to change the strings from unicode into plain text. We grouped the dataset by the column “county_fips” and got the mean percentage of housing burden and the mean percent of each county’s population aged 25 and up with a four-year college degree or higher, which has a number representing each county in California. The first table is the cleaned dataset for household burdens and the second table is the cleaned dataset for educational attainment.

[2] Housing Cost Burden - Dataset - California Open Data

[3] Educational Attainment - Dataset - California Open Data



ind_id	ind_definition	percent	burden	geotype	income_level	geotypevalue	geoname	region_name	region_code	county_fips
106	Percent of households spending more than 30% (...	17.217629	> 50% of monthly household income consumed by ...	CO	Monthly household income at all levels of HUD-...	06001	Alameda	Bay Area	01	06001
106	Percent of households spending more than 30% (...	22.556391	> 50% of monthly household income consumed by ...	CO	All income levels	06001	Alameda	Bay Area	01	06001
106	Percent of households spending more than 30% (...	17.071468	> 50% of monthly household income consumed by ...	CO	All income levels	06001	Alameda	Bay Area	01	06001
106	Percent of households spending more than 30% (...	28.111467	> 50% of monthly household income consumed by ...	CO	All income levels	06001	Alameda	Bay Area	01	06001
106	Percent of households spending more than 30% (...	23.745072	> 50% of monthly household income consumed by ...	CO	All income levels	06001	Alameda	Bay Area	01	06001



	ind_id	ind_definition	estimate	geotype	geotypevalue	geoname	region_name	region_code	county_fips
9	355	Percent of population age 25 and up with a fou...	21.1	CD	607192260	Ontario CCD	Southern California	14.0	6071.0
10	355	Percent of population age 25 and up with a fou...	0.1	CD	603593280	Susanville CCD	Northeast Sierra	6.0	6035.0
11	355	Percent of population age 25 and up with a fou...	17.0	CD	607192100	Newberry-Baker CCD	Southern California	14.0	6071.0
12	355	Percent of population age 25 and up with a fou...	19.9	CD	609792940	Santa Rosa CCD	Bay Area	1.0	6097.0
13	355	Percent of population age 25 and up with a fou...	28.3	CD	600190020	Alameda CCD	Bay Area	1.0	6001.0

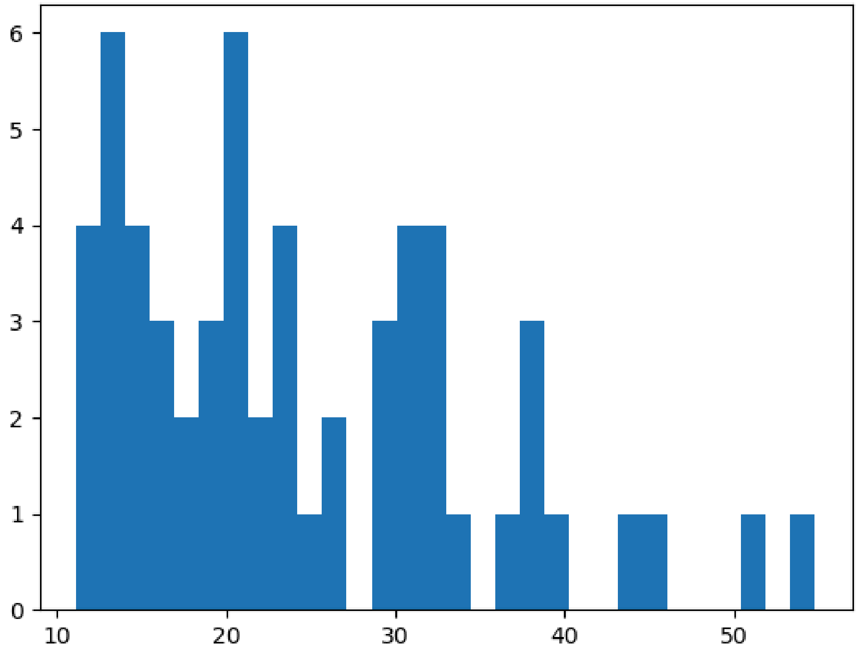
Exploratory Data Analysis

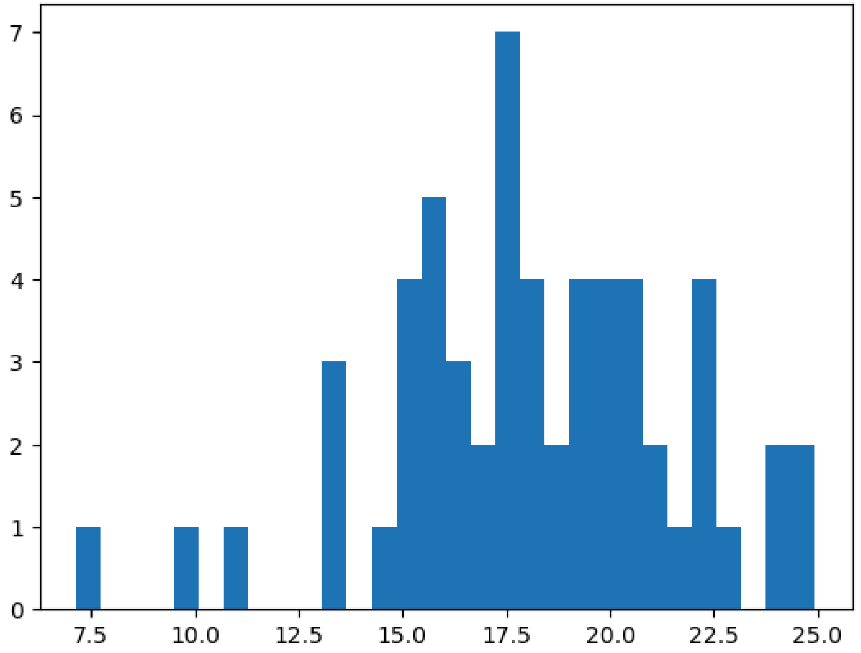
Before moving onto hypothesis testing, we first use exploratory data analysis techniques to identify relationships between the two variables.

- Histogram to do a cursory check of the distribution of the two variables.
- Side-by-side choropleth map of housing burden and education attainment will show how the distribution of the two datasets look in California when viewed side by side.
- A heat map will be used to communicate relationships between the variables.
- A scatterplot will also be included as visualization to show the strength and direction of the correlation.

Histogram

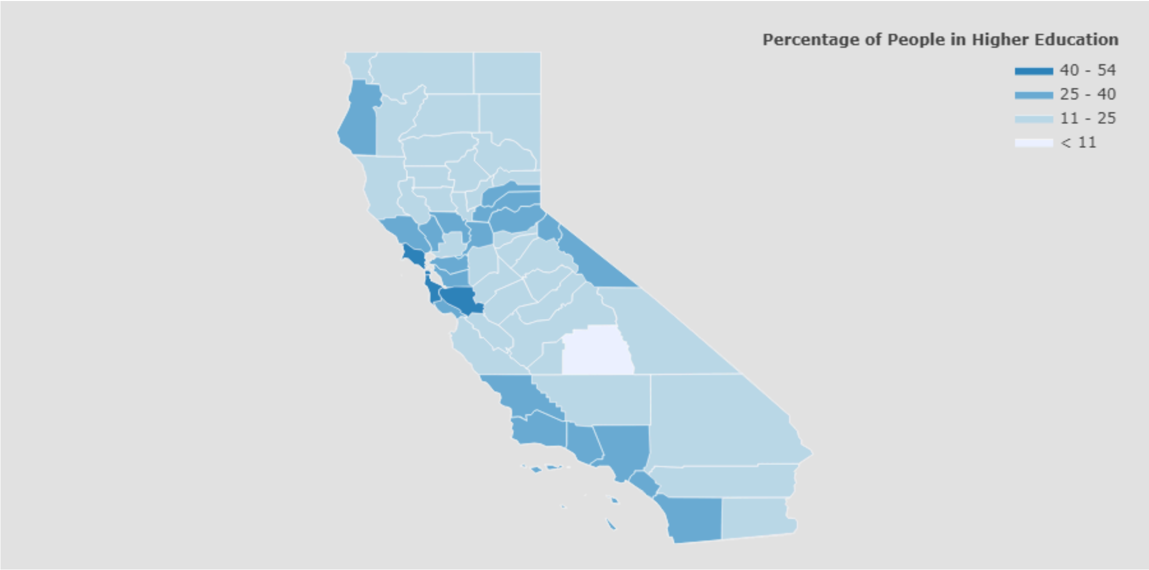
The x-axis on the histogram are the proportion of households with high housing burden/population with higher education and the y-axis are the number of counties corresponding to a range of proportions that each bar represents. The first histogram is about educational attainment and the second histogram is about housing burden. A quick look shows that the distributions appear to be very different.

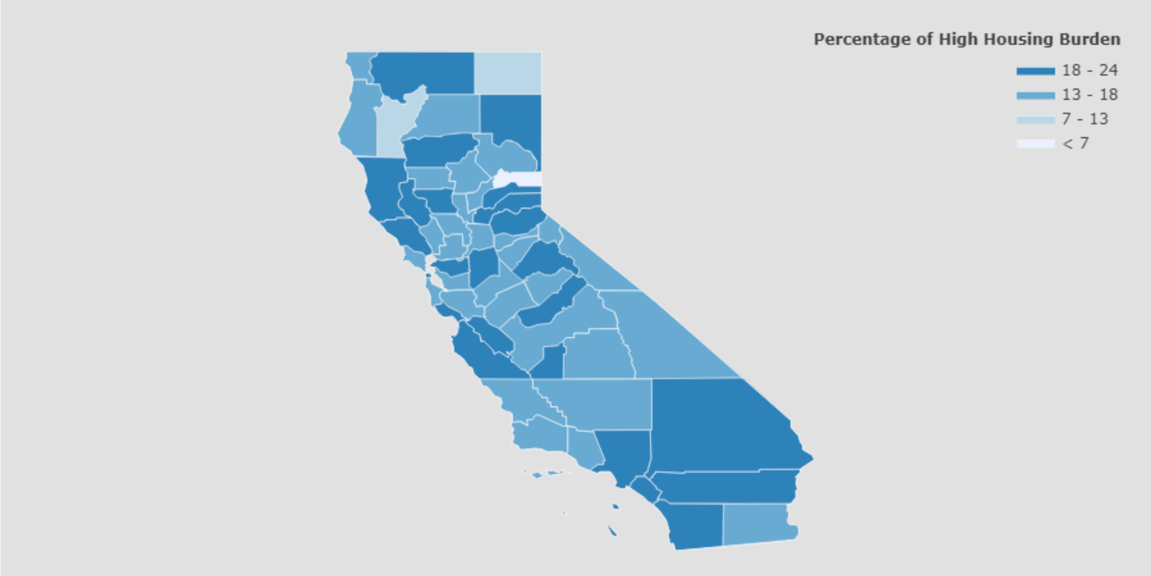




Choropleth Map

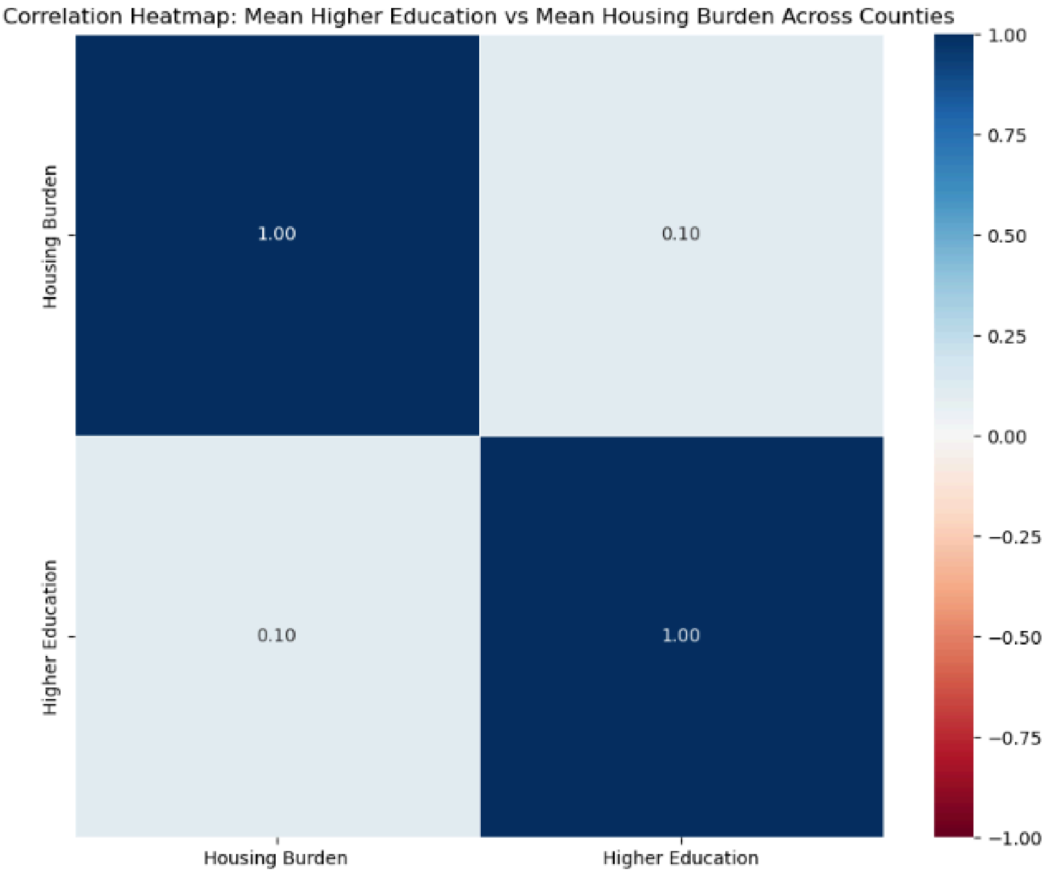
The choropleth maps are divided into 58 counties of California. Looking at the two maps side by side, it can be observed that some places with a greater percentage of households with high housing burden have lower education attainment and vice versa.





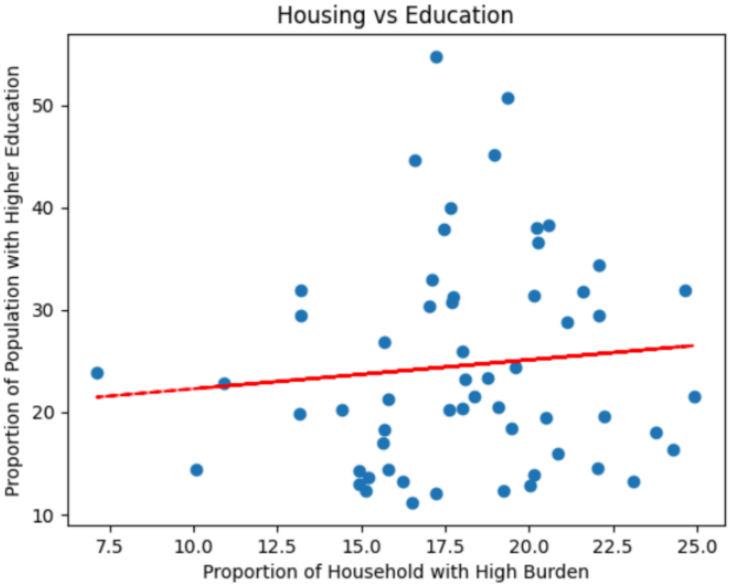
Heat Map

There is a slight positive correlation of 0.1 between housing burden and education. We will also see this weak positive correlation from the scatterplot later.



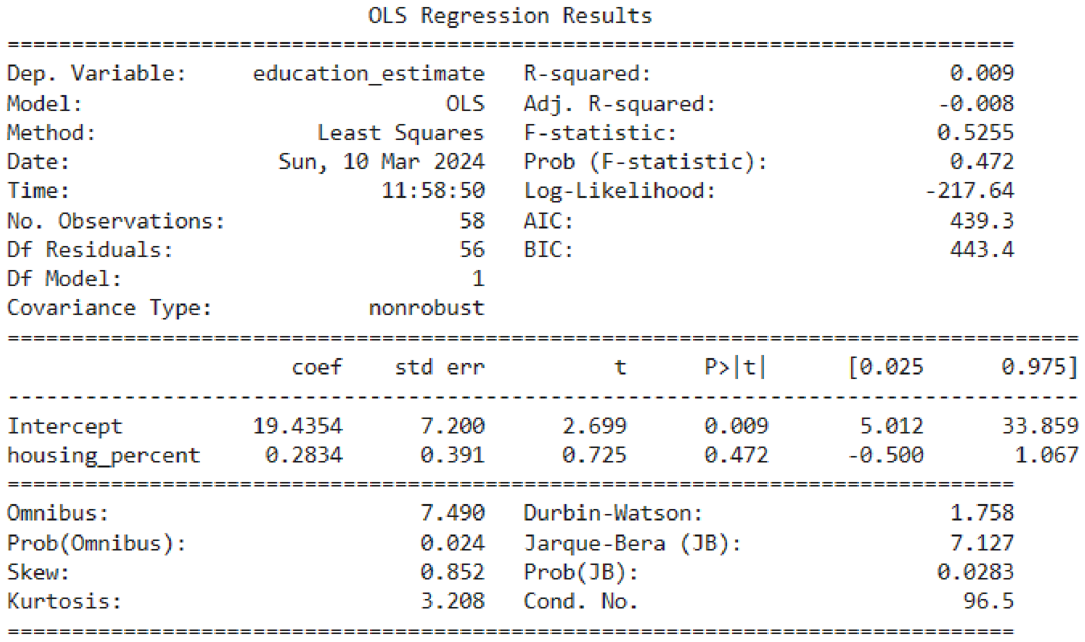
Scatterplot

The scatterplot shows that the observations don’t seem to follow any particular trend. Even with a regression line, only a faint weakly positive correlation can be drawn from the graph.



Ordinary Least Squares Regression

Going further with the scatterplot, we decided to run the OLS regression to see how strong the correlation is. We tested assumptions first, namely homoscedasticity and normality, to ensure that we can draw inferences from OLS regression. Both of these assumptions are met so we can go onto the regression. After running the OLS test, there isn't sufficient evidence to prove that there is a correlation between the two variables due to high p-values. Even if there were, the R-squared value of 0.009 means that the correlation would be very weak regardless. We also drew a scatter plot to show the weak correlation.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     education_estimate   R-squared:                       0.009
Model:                            OLS   Adj. R-squared:                 -0.008
Method:                 Least Squares   F-statistic:                    0.5255
Date:                Sun, 10 Mar 2024   Prob (F-statistic):              0.472
Time:                        11:58:50   Log-Likelihood:                -217.64
No. Observations:                  58   AIC:                             439.3
Df Residuals:                      56   BIC:                             443.4
Df Model:                           1
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
Intercept          19.4354      7.200      2.699      0.009       5.012      33.859
housing_percent     0.2834      0.391      0.725      0.472      -0.500       1.067
==============================================================================
Omnibus:                        7.490   Durbin-Watson:                   1.758
Prob(Omnibus):                  0.024   Jarque-Bera (JB):                7.127
Skew:                           0.852   Prob(JB):                       0.0283
Kurtosis:                       3.208   Cond. No.                         96.5
==============================================================================
```

Hypothesis Test

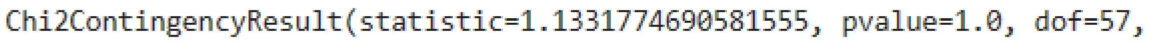
Chi-square for Independence

Hypothesis test we did was the chi-square for independence test. While we could have used the OLS regression from earlier, it might suffer from omitted variable bias since we removed a lot of other variables during data cleaning. We wanted to know if there is any relation between the percentage of households with high housing burden versus the percentage of the population that attained higher education within each region.

- Null hypothesis: No relationship between the two variables
- Alternative hypothesis: There is a relationship between the two variables.

We get a p-value of 1 running the test, so we do not reject the null hypothesis. There appears to be no relationship between education attainment and housing burden. Very similar results to the OLS test from before.

The following is the output from the test:



```
Chi2ContingencyResult(statistic=1.1331774690581555, pvalue=1.0, dof=57,
```

Interpretation

Exploratory Data Analysis

Our EDA through choropleth, heat map, and scatterplot seem to show a weak positive correlation between percentage of households with high housing burden and percentage of the population with high education attainment in each county. This may seem to support our hypothesis that there may exist a correlation between the two variables, but this appears rather weak when looking at the scatterplot. Due to omitting all of the confounding variables in the datasets, we cannot fully interpret the reasoning behind the positive correlation and any attempts to do so would be speculation.

Hypothesis Testing Both the hypothesis tests - chi-square test for independence and OLS - firmly conclude that the null hypothesis cannot be rejected. First, the OLS regression gives the R-squared value of 0.009, so even though the correlation can be positive, it is a very weak correlation that is very close to 0, to the point where the correlation might not even exist at all. Second, the chi-square tests give us a p-value of 1, so regardless of confidence level, there is no sufficient evidence to support our proposed hypothesis that there exists a correlation between education attainment and housing burden.

Conclusion

Our investigations did little to suggest that we reject the null hypothesis as there seems to be a very weak, if any at all, correlation between housing burden and higher education attainment. Both the exploratory data analysis and hypothesis testing showed overwhelming evidence supporting our null hypothesis. There may be multiple reasons for this:

- It’s possible that the results may change if we redefine the high housing burden as households spending 30% of income on housing instead of 50%.
- There may be a lot of confounding variables at play, such as the role that family size or culture may have on education attainment as opposed to housing burden. It’s possible that omitting those variables from the cleaned dataset can also lead to omitted variable bias, meaning that even if the alternative hypothesis is supported, it may be infeasible to draw any meaningful conclusions from it.
- Furthermore, the housing burden doesn’t tell us the overall income of a family (e.g a household with six figure income with high burden may do better than a poor household with subsidized housing.

In the future, we would like to redo this analysis with the aforementioned flaws better controlled to find out if there truly exists a correlation between housing burden and education attainment.