MUSA500 Homework 1: Using OLS Regression to Predict Median House Values in Philadelphia

Ling Chen, Hang Zhao, Jiahang Li

2023-10-19

Contents

Introduction

Philadelphia, renowned for its rich history and pulsating present, is home to a continually evolving real estate environment. However, a study by the Economic League reveals a nuanced picture: the overall proportion of Philadelphia households grappling with housing cost burden experienced a decrease from 29.8% to 26.7%

between 2016 and 2021(Economic League, 2023). Nevertheless, this alteration in cost burden manifested divergently across various income brackets. Considering housing is a fundamental human necessity, ensuring affordability is crucial for maintaining well-being and quality of life. Consequently, comprehending the factors that influence housing values is essential for exerting better control over the housing market and making more strategic, informed decisions.

This report aims to explore the relationship between median house values and various neighborhood characteristics within the city of Philadelphia. It is widely understood that property values are influenced by both the conditions of the housing and the economic status of the property owners. By analyzing data at the Census block group level, we aim to comprehend the relationship between median house value and several neighborhood characteristics, including the proportion of residents in the Block Group with at least a bachelor's degree, housing vacancy, percentage of housing units that are detached single-family houses, and number of households living poverty.

Methods

Data Cleaning

The data focuses on a variety of demographic variables in the census data, starting with 1,816 census data by census block level. To further refine the dataset, a systematic data cleansing process is employed that ultimately cleans the dataset into 1,720 observations. Firstly, block groups with a population of less than 40, those without any housing units, and those with median house values lower than $10,000 are identified and flagged for further action. Additionally, an outlier block group in North Philadelphia, characterized by an unusually high median house value (over $800,000) and very low median household income (less than $8,000), is isolated. These identified anomalies are either removed from the dataset or corrected as needed, ensuring that the final dataset consists of 1720 clean and validated observations. Comprehensive documentation and quality checks are performed throughout the process to maintain data integrity and transparency.

Exploratory Analysis

The first step involves importing a dataset from "RegressionData.csv" into R, examining the distribution of the dependent variable (MEDHVAL) and predictors (PCBACHMORE, NBELPOV100, PCTVACANT, PCTSINGLES) using histograms, and calculating their mean and standard deviation. Additionally, logarithmic transformations are applied to these variables, with a special transformation ($\log(1 + [\text{VAR}])$) used for variables with zero values. The histograms of both the original and transformed variables are created to assess normality. Finally, a summary statistics table is constructed to present the mean and standard deviation of each variable.

Understanding the characteristics of the data set and the distribution of the variables facilitates the assessment of linearity and normality. The method involves plotting scatter plots as well as using histograms to examine the distribution of the data. This process helps to determine the applicability of different regression models, as various models have different assumptions, such as the normality of residuals. This comprehensive approach enables a thorough exploration of the dataset's characteristics, facilitates data normalization where necessary, and prepares the data for subsequent regression analysis. While it is possible for a non-normally distributed variable to have normally distributed values, it is more likely that if the variable itself is not normally distributed, its residuals will not be normally distributed either. This effort is consistent with the goal of creating interpretable regression models, as normally distributed variables and residuals are easier to interpret and comply with regression assumptions, ultimately improving the reliability and utility of the model for understanding relationships in the data.

The next step is to assess the linearity of the relationships between the dependent variable (MEDHVAL) and each of the predictors (PCBACHMORE, NBELPOV100, PCTVACANT, PCTSINGLES). The methodology involves creating four scatter plots, one for each predictor, to visually examine the patterns and associations

between these variables. This process enables a qualitative assessment of whether the relationships appear to be linear or exhibit other types of trends, which is crucial for determining the suitability of a linear regression model for subsequent analysis. The scatter plots provide a visual representation of the data, aiding in the decision-making process regarding the choice of regression techniques and the understanding of how predictors may influence the dependent variable.

$$y = \beta_0 + \beta_1 * x + \varepsilon$$

The third step is assessing the relationships between predictor variables by calculating Pearson correlations, with a focus on identifying multicollinearity among them. The methodology involves using the cor function in R to compute these correlations, producing a correlation matrix. The process entails examining the values in the correlation matrix to determine if any predictors exhibit strong pairwise correlations, which could indicate multicollinearity. Pearson's correlation coefficient is from -1 to 1, where -1 indicates a strong negative linear relationship, 1 indicates a strong positive linear relationship, and 0 implies no linear relationship. Multicollinearity, where predictor variables are highly correlated with each other, can lead to unstable and unreliable regression results. The aim is to decide whether it's appropriate to include all four variables as predictors in the regression model based on the observed correlations, ensuring a robust and interpretable model for subsequent analysis.

$$r = \frac{\sum((X_i - \bar{X}) \cdot (Y_i - \bar{Y}))}{\sqrt{\sum(X_i - \bar{X})^2} \cdot \sqrt{\sum(Y_i - \bar{Y})^2}}$$

Finally, visualizing spatial patterns and relationships within geographic data by creating choropleth maps for five variables. The methodology involves utilizing the R programming language and the sf package for importing and handling shapefile data, and the ggplot2 package for creating the choropleth maps. The process begins with importing the shapefile and then plotting each variable individually with color scales chosen for clarity and consistency. The final step combines all five maps into a single figure for presentation, facilitating a visual exploration of spatial distributions and correlations among these variables, and enhancing the understanding of geographic patterns in the dataset.

Multiple Regression Analysis

Ordinary Least Squares (OLS) regression is a statistical technique to determine the relationships between a variable of interest, known as the dependent variable, and one or more independent explanatory variables, often referred to as predictors. It is often used to assess the strength and direction of the correlations between variables, indicating whether it's positive, negative, or no correlation. It also evaluates how well the model fits the data, providing goodness of fit information. Each beta coefficient of the predictors demonstrates to what extent the dependent variable will change when one unit changes in one of the predictors, holding all other predictors constant. However, while significant predictor variables indicate a certain relationship, they do not establish causation between variables.

We use regression analysis to determine the correlation between the dependent variable, which is the natural log of median house value, represented as LNMEDHVAL, and the predictors which are a proportion of housing units that are vacant PCTVACANT, percent of housing units that are detached single-family houses PCTSINGLES, proportion of residents in Block Group with at least a bachelor's degree PCTBACHMOR, and the natural log of number of households that income below 100% poverty level LNNBELPOV100. Our equation is shown as follows:

$$LNMEDHVAL = \beta_0 + \beta_1 PCTVACANT + \beta_2 PCTSINGLES + \beta_3 PCTBACHMOR + \beta_4 LNNBELPOV100 + \epsilon$$

Where β_0 is the y-intercept, interpreting the value of the dependent variable when the predictors are 0; β_1, β_2, β_3, β_4 are the slope coefficients of the predictors.

For a linear regression model, for any fixed value of independent variable x, there are parameters β_0, β_i, and , where i = 1 in a simple regression and i>1 in a multiple regression, such that y = β_0 + $\beta_i$xi +ϵ, where $\epsilon \sim$ N(0, σ^2). The term ϵ is known as an error term or residual, and for each observation i, is defined as a vertical deviation (distance) between the observed value of y and the predicted value of y, denoted by ŷ. In addition, ϵ~N(0, σ^2) means that the error terms have a normal distribution with a mean of 0 and variance σ^2. This holds for any given value of x, the average error term will be 0, and a typical deviation from the regression line will be σ^2 units.

There are several assumptions we have to make prior to the linear regression. 1. Check the linearity of each predictor and the dependent variable by creating scatter plots. If no linearity can be observed from the plots, variable transformation or polynomial regression might be better. 2. Examine the normality of residuals by plotting out the histogram. If the histogram is not normally distributed, log transformation may be used to normalize both the dependent variable and predictor. However, sometimes log transformation is not appropriate, especially when there are high zero inflations. 3. Confirm homoscedasticity, which means that the variance of residuals should be constant throughout the different values of x. 4. Predictors should not be strongly correlated with each other, which is also called to prevent multicollinearity. 5. No fewer than 10 observations per predictor.

Given n observations on y, and k predictors x1 … xk, the estimates β_0, β_1, β_2, …. β_k are chosen simultaneously to minimize the expression for the Error Sum of Squares (SSE), given by:

$$SSE = \sum_{i=1}^{n} \epsilon^2 = \sum_{i=1}^{n} (y - \hat{y})^2 = \sum_{i=1}^{n} (y_i - \hat{\beta_0} - \hat{\beta_1} x_1 i - \hat{\beta_2} x_2 i - ... - \hat{\beta_k} x_k i)^2$$

where ŷ is the predicted y of the model, which equals β_0+$\beta_1 x_1$+$\beta_2 x_2$+…..+$\beta_k x_k$, with the minus sign before it would be demonstrated as the equation in the bracket. SSE represents the sum of squared error, or the sum of squared residuals ϵ, which is the amount of variability in y that is not explained when accounting for x in the model. There is another term SST, which means the total sum of squares, is demonstrated as the following equation:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \frac{\sum_{i=1}^{n} y_i}{n})^2$$

where $\bar{y}$ here represents the overall mean of y values, therefore SST is interpreted as the squared deviation of that observation from the overall mean of y, and then summing those squared deviations across all observations i, without any regard to the value of x.

$$\hat{\rho} = r = Corr(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Sample correlation coefficient R is a point estimator of the population correlation ρ.

If we use the formula 1 – SSE/SST, we can get the coefficient of determination R^2, which is the proportion of observed variation in the dependent variable y that was explained by the model. Also, it always ranges between 0 and 1.

$$R^2 = 1 - \frac{SSE}{SST}$$

To assess our model, we examine the F-statistic and its corresponding p-value. The F-test, often referred to as the omnibus test, evaluates whether any of the independent variables in the model significantly predict the dependent variable. It tests the null hypothesis that none of the independent variables are significant predictors against the alternative hypothesis that at least one of them is. A model that fails to reject the

null hypothesis is typically considered less effective. We then focus on the p-value associated with each independent variable. If the p-value for a specific independent variable is below 0.05, we can reject the null hypothesis, indicating that this particular predictor significantly influences the dependent variable. In this case, our null hypothesis, or H0 is that the coefficient β equals zero, and the alternative hypothesis, or Ha states that the coefficient β does not equal zero, demonstrating as H0: β=0 and Ha: $\beta \neq 0$;

Additional Analyses

To further test the relationship between median house values and studied neighborhood characteristics, we also run the stepwise regression. Stepwise regression is a statistical method that allows us to understand the statistical relationship between independent and dependent variables. The process of stepwise regression screens candidate variables and automatically identifies influential variables. In this scenario, stepwise regression is used to examine the statistical relationship between the dependent variable (MEDHVAL), and predictors (PCBACHMORE, NBELPOV100, PCTVACANT, and PCTSINGLES) based on the Akaike Information Criterion (a mathematical method for evaluating how well a model fits the data it was generated from). Specifically, the algorithm adds or removes predictors to see if there is a significant change in the model fit determined by the AIC value and retains all predictors resulting in significant changes. The model with a smaller ACI is usually regarded as a better one. However, there are limitations as well. Firstly, stepwise regression often leads to overfitting. To be more specific, sometimes the dataset does not contain enough data samples to accurately represent all possible input data values, leading to poor generalization to new datasets. Furthermore, rather than relying on professional knowledge, the model relies on an automatic process of selecting predictive variables. Therefore, it may overlook a more comprehensive model.

To test the problem of overfitting, we implement the K-fold cross-validation, a method used for evaluating the model performance. To further explain this, in this scenario(k=5), the sample dataset is randomly divided into five folds for training and validation. During each run, one-fold is selected for validation, and the rest are used for training and further iterations. This process is repeated five times, each with a different fold serving as the validation set and the other four as the training set. After this process, we will get five different performance values for each fold, the average of which serves as a holistic performance metric to determine how generalizable our model is. In this scenario, we will use the root mean squared error (RMSE) as the referencing performance value to evaluate the model's performance. The RMSE measures the average magnitude of errors between the predicted and observed values in a dataset. In other words, it tells us the standard deviation of the residuals (prediction errors).

Turning to the discussion of the formula of the RMSE calculation, firstly, we need to get the SSE. In the formula below, Xi stands for the observed values, Xn stands for the corresponding predicted values. The SSE is calculated as:

$$SSE = \sum_{i=1}^{n} (x_i - x_n)^2$$

We can then get the mean squared error (MSE) by dividing SSE by the number of observations n:

$$MSE = \frac{\sum_{i=1}^{n} (x_i - x_n)^2}{n}$$

After taking the square root of the MSE, we get the value for RMSE:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_i - x_n)^2}{n}}$$

In this study, we initially conduct cross-validation on the regression model incorporating all four predictors. Subsequently, for comparative purposes, we also perform cross-validation on a model using only PCTBANT(housing vacancy) and MEDHHINC(median household income) as predictors.

Software

The software we used is R, a programming language with powerful statistical analysis capabilities.

Results

Exploratory Results

Summary Statistics

First of all, we import the csv and shp data files. To see the fundamental statistical information of the original data, we perform summary() function to see the mean values, and the sd() function to see the standard deviation of each variable. The results are shown below in the table.

Variable	Mean	SD
Dependent Variable		
Median House Value	66287.73	60006.08
Predictors		
N of Households Living in Poverty.	189.7709	164.3185
% of Individuals with Bechelor's Degrees or Higher	16.08137	17.76956
% of Vacant Houses	11.28853	9.628472
% of Single House Units	9.226473	13.24925

```
# look at mean values in summary table
summary(data)
```

```
##     POLY_ID          AREAKEY                MEDHVAL          PCTBACHMOR
##  Min.   :   1.0   Min.   :421010000000   Min.   :  10000   Min.   : 0.000
##  1st Qu.: 430.8   1st Qu.:421010000000   1st Qu.:  35075   1st Qu.: 4.847
##  Median : 860.5   Median :421010000000   Median :  53250   Median :10.000
##  Mean   : 860.5   Mean   :421010000000   Mean   :  66288   Mean   :16.081
##  3rd Qu.:1290.2   3rd Qu.:421010000000   3rd Qu.:  78625   3rd Qu.:20.074
##  Max.   :1720.0   Max.   :421010000000   Max.   :1000001   Max.   :92.987
##     MEDHHINC        PCTVACANT        PCTSINGLES        NBELPOV100
##  Min.   :  2499   Min.   : 0.000   Min.   :  0.000   Min.   :   0.0
##  1st Qu.: 21061   1st Qu.: 4.372   1st Qu.:  2.110   1st Qu.:  72.0
##  Median : 29719   Median : 9.091   Median :  5.714   Median : 147.0
##  Mean   : 31542   Mean   :11.289   Mean   :  9.226   Mean   : 189.8
##  3rd Qu.: 38750   3rd Qu.:16.282   3rd Qu.: 11.056   3rd Qu.: 257.0
##  Max.   :200001   Max.   :77.119   Max.   :100.000   Max.   :1267.0
```

```
# print out all the standard deviations
sd(data$MEDHVAL)
```

```
## [1] 60006.08
```

```
sd(data$NBELPOV100)
```

```
## [1] 164.3185
```

```r
sd(data$PCTBACHMOR)
```

```
## [1] 17.76956
```

```r
sd(data$PCTVACANT)
```

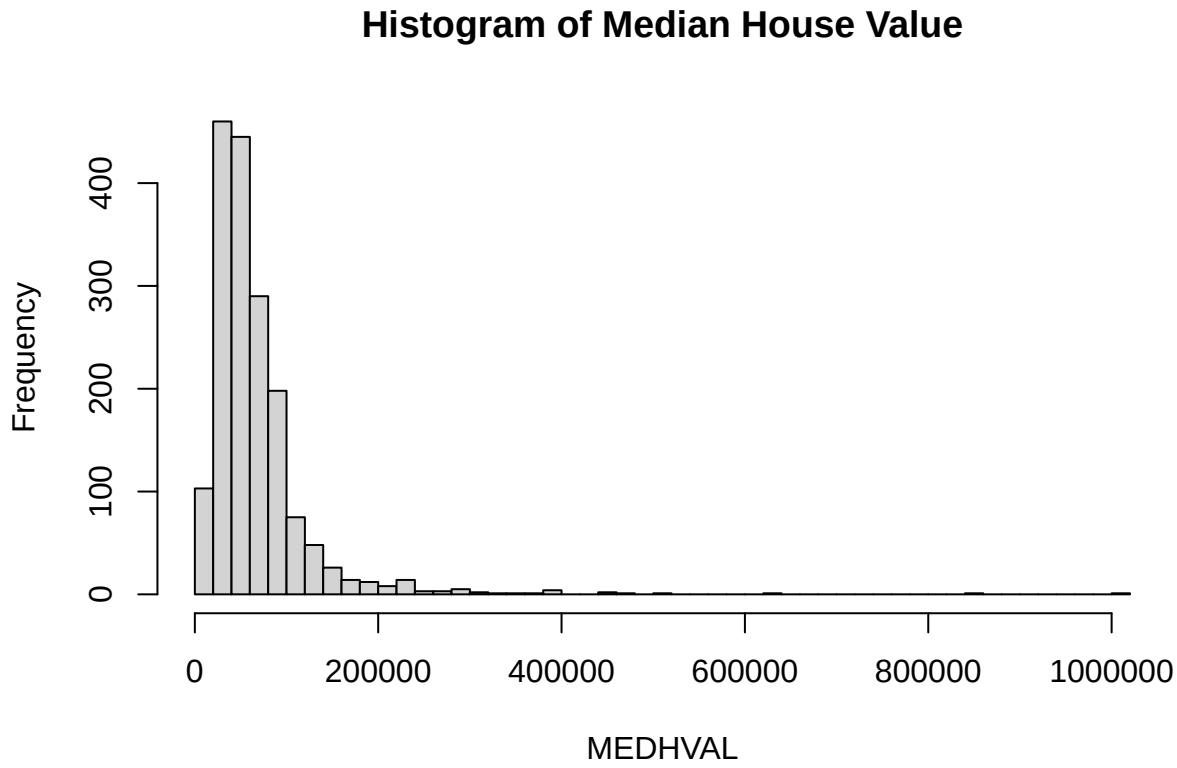
```
## [1] 9.628472
```

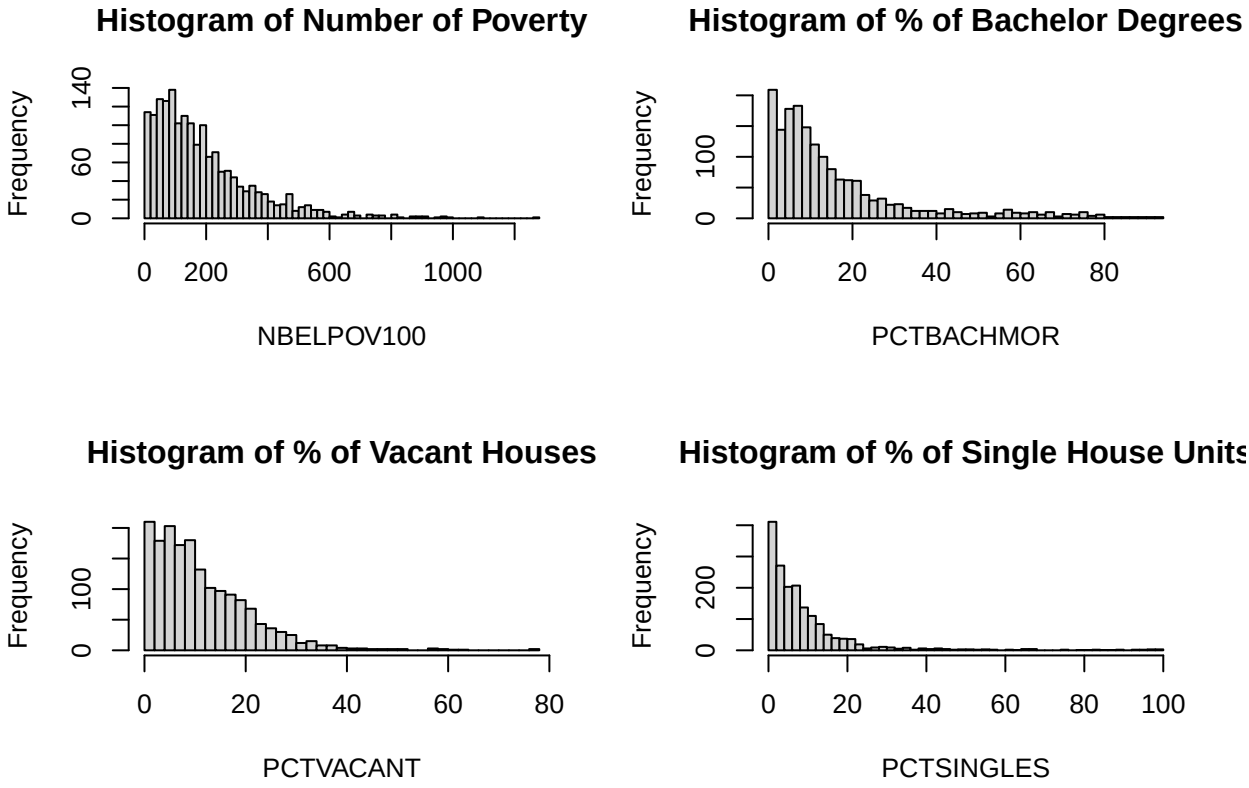
```r
sd(data$PCTSINGLES)
```

```
## [1] 13.24925
```

Histograms and Log Transformation

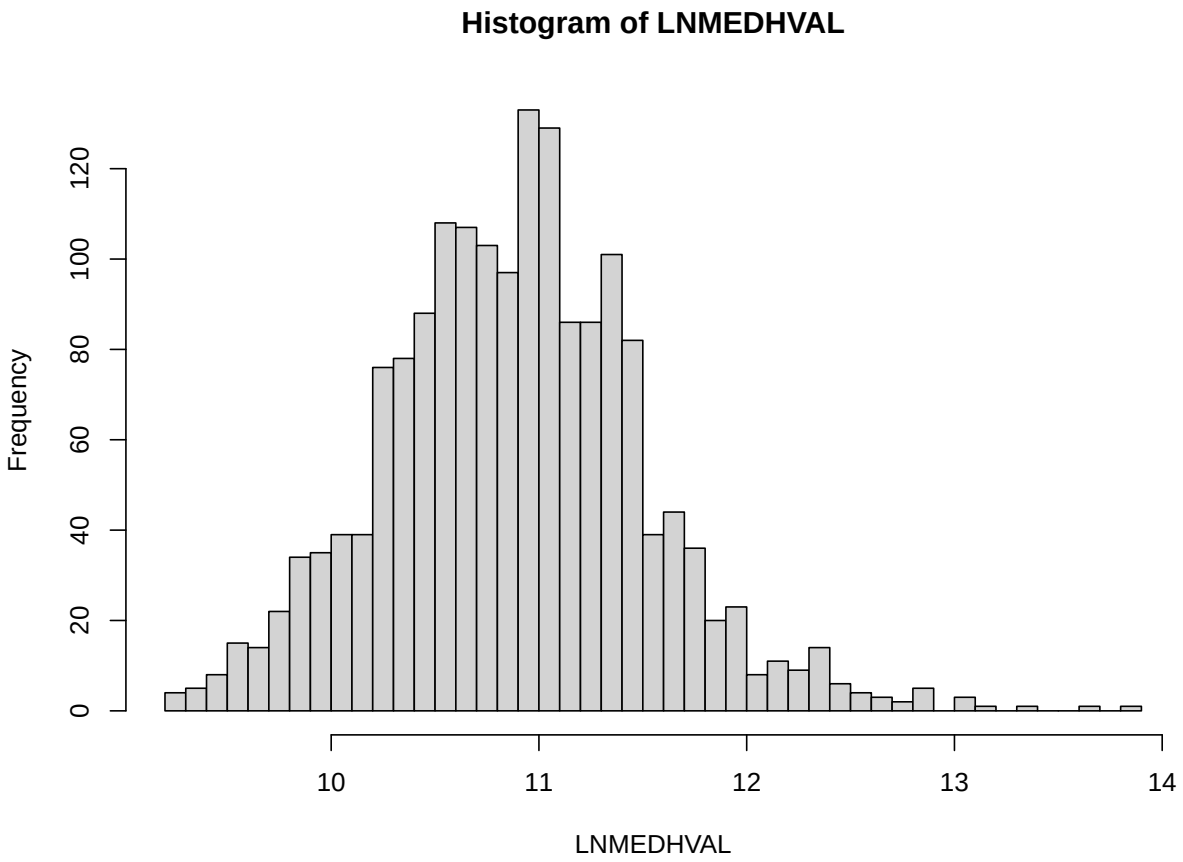
The histograms below illustrate the distribution of the dependent variable and the four predictors, where all of these histograms are positively skewed. Consequently, we have applied Log transformation to the original variables to normalize their distributions. We have added 1 to the log-transformed data to avoid Log(0) which is undefined. Following that, we present new histograms depicting the distribution of the log-transformed variables.

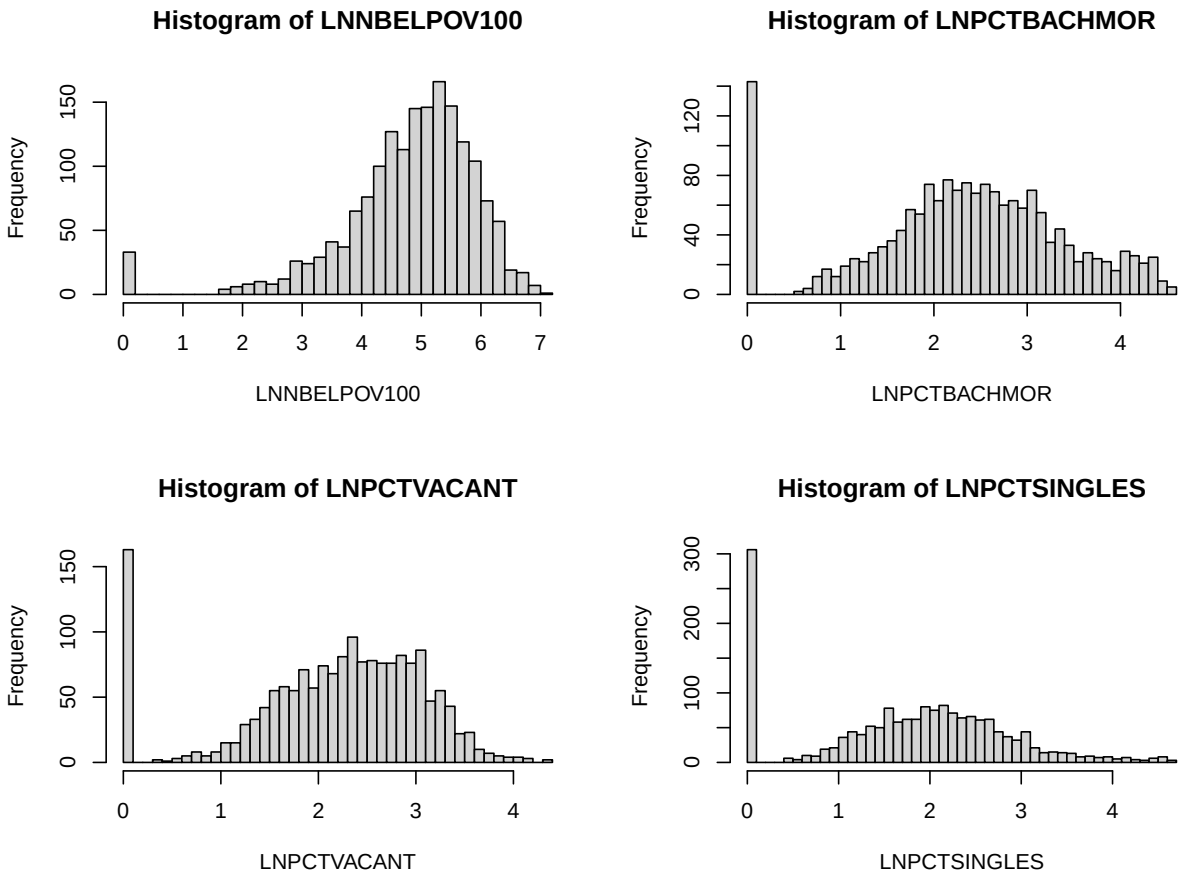




From the new histograms, we can see that LNPCTBACHMOR, LNPCTVACANT, LNPCTSINGLES all have zero inflation, which means that there are very high frequency of zero values in the histograms after log transformation. Keeping that in mind, we will only use the log Median House Value presented as LNMEDHVAL (dependent variable), and log Number of Household living in Poverty LNNBELPV100 (one predictor) for the following regression analysis, while keeping the other three variables original.

Other assumptions for linear regression including checking the linear relationship between dependent variable y and each of the predictors x, homoscedasticity of the variance of residuals, independence of observations, and multicollinearity will also be examined in the following section 3.3


Histogram of LNMEDHVAL
Frequency
0
20
40
60
80
100
120
10
11
12
13
14
LNMEDHVAL



Choropleth maps

Choropleth maps of each variable, LNMEDHVAL, LNNBELPOV, PCTVACANT, PCTSINGLES, and PCTBACHMOR are presented below. We used 5 quantile breaks as the map representation method. From the maps, we can see that there are some clear overlaps between LNMEDHVAL map(Figure 1) and PCTBACHMOR map(Figure 5), showing a strong correlation of the predictor % of Bachelor's Degree to the dependent variable Median House Value. Whereas the LNNBELPOV and PCTVACANT maps have completely different patterns from the MEDHVAL map, illustrating very weak correlations. PCTSINGLES however, is presenting a partially similar pattern to the MEDHVAL, which may have some extent of correlation to the dependent variable.

Among the predictors, there are no obvious similarities between the maps, while LNNBELPOV does show a little overlap with the PCTVACANT, we do not expect there to be severe multicollinearity between the predictors.

LN Median House Value in Philadelphia

Date Source: U.S. Census

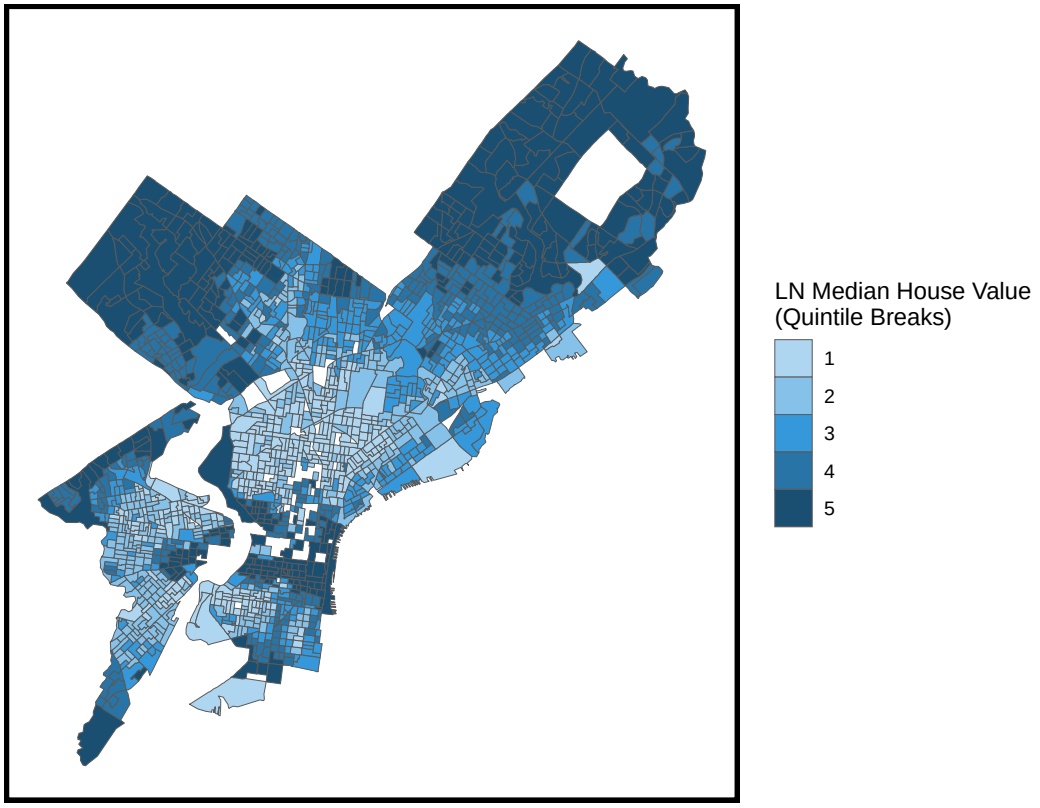


Figure 1

LN number of households living in poverty in Philadelphia

Date Source: U.S. Census

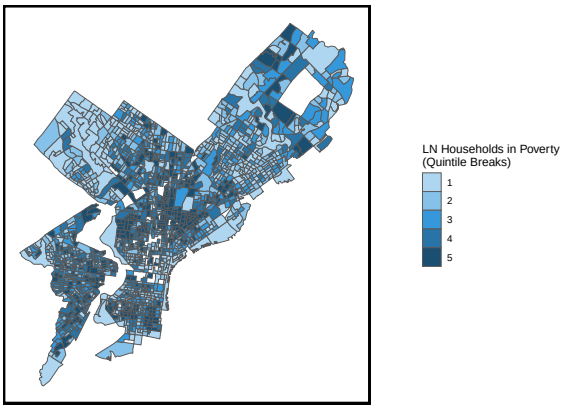


Figure 2

Proportion of housing units that are vacant in Philadelphia

Date Source: U.S. Census

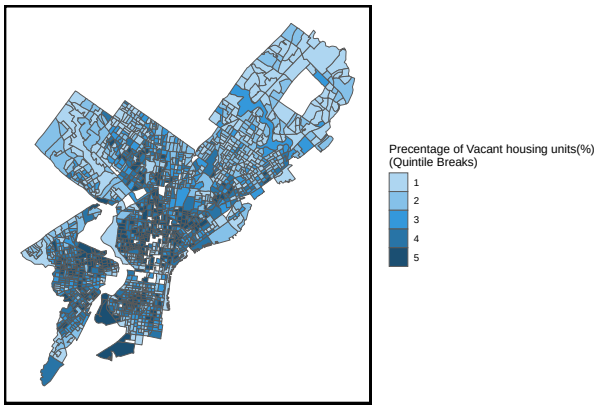


Figure 3

Percent of housing units that are detached single family houses in Philadelphia

Date Source: U.S. Census

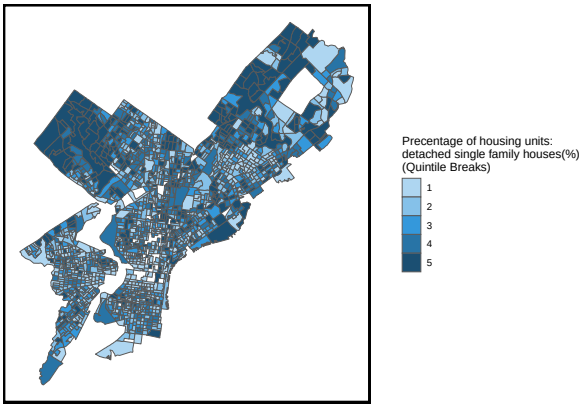


Figure 4

Proportion of residents in Block Group with at least a bachelor's degree in Phila

Date Source: U.S. Census

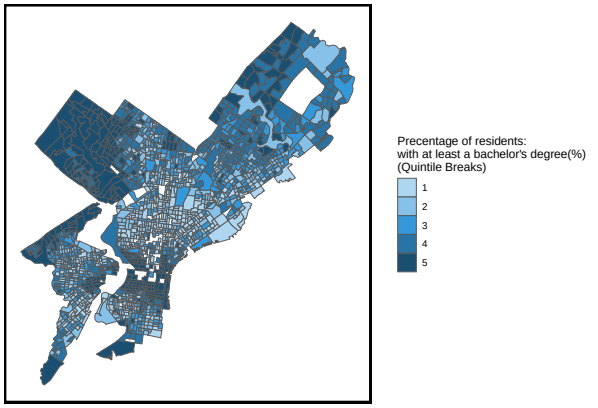
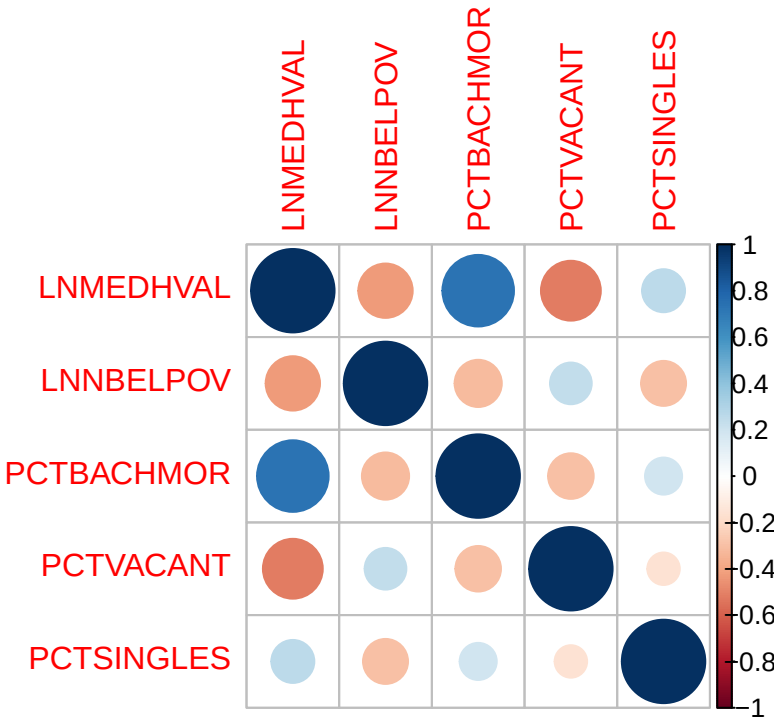


Figure 5

Correlation Matrix

The correlation matrix in the graph and table is shown below. It is used to test multicollinearity between the predictors. The table shows that the greatest correlation coefficients between predictors are -0.31, 0.24, and -0.29, which are not strongly correlated. Therefore, no severe multicollinearity between the predictors has been observed, which aligns with the expectations from previous map observations.



Look at multicollinearity (exclude the dependent variable in the correlation matrix table)

```
##              LNNBELPOV PCTBACHMOR  PCTVACANT PCTSINGLES
## LNNBELPOV   1.0000000 -0.3197668  0.2495470 -0.2905159
## PCTBACHMOR -0.3197668  1.0000000 -0.2983580  0.1975461
## PCTVACANT   0.2495470 -0.2983580  1.0000000 -0.1513734
## PCTSINGLES -0.2905159  0.1975461 -0.1513734  1.0000000
```

Regression Results

In our analysis of median house values in Philadelphia, we used a linear regression model to investigate the relationships between several predictor variables and the natural logarithm of median house values(LNMEDHVAL).

The final equation is as follows:

$$ln(y) = LNMEDHVAL = \beta_0 + \beta_1 PCTVACANT + \beta_2 PCTSINGLES + \beta_3 PCTBACHMOR + \beta_4 LNBELPOV + \epsilon$$

From the statistical summary table, there are several key findings. Firstly, the F-statistic is high and the P-value is much smaller than 0.05. Therefore, we can reject the null hypothesis that none of the independent variables in the model is a significant predictor of the dependent variable. Secondly, the coefficients for each predictor variable also provide some insights.

Notably, a higher percentage of vacant housing units(PCTVACANT) is associated with a significant decrease in median house values, indicating the negative impact of housing vacancy on property values. That is to say, a 1% additional proportion of vacant housing units is associated with a $19 decrease in median house values. In addition, a higher number of households with incomes below 100% of the poverty level(LNNBELPOV) is associated with a significant decrease in median house values as well. As the number of households

in poverty changes by 1%, the expected value of median house values changes by $(1.01^{\beta_1} - 1) * 100 = (1.01^{-.079} - 1) * 100 = -0.0786\%$ Conversely, the percentage of housing units that are detached single-family houses(PCTSINGLES) has a strong positive relationship with house values. 1% additional percentage of housing units that are detached single-family houses is associated with a $3 increase in median house values. Also, a higher proportion of residents with at least a bachelor's degree(PCTBACHMOR) exhibits a strong positive relationship with house values, showing that areas with a well-educated population tend to have higher property values. Specifically, the expected change in median house values associated with 1 additional percentage of residents who has at least a bachelor's degree is $(1.01^{\beta_1} - 1) * 100 * (\$1000) = (1.01^{.021} - 1) * 100 * (\$1000) = \$20.9$ Finally, the multiple R-square(0.6623) indicates that approximately 66.23% of the variance in median house values can be explained by the model. The adjusted R-squared further takes into account the number of predictors in the model, which is 66.15% in this case.

```
# run the 'lm' function
lm1 <- lm(LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNNBELPOV,
          data=data_cor)
```

```
# print out the statistical summary table
summary(lm1)
```

```
##
## Call:
## lm(formula = LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR +
##     LNNBELPOV, data = data_cor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.25817 -0.20391  0.03822  0.21743  2.24345
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 11.1137781  0.0465318 238.843 < 0.0000000000000002 ***
## PCTVACANT   -0.0191563  0.0009779 -19.590 < 0.0000000000000002 ***
## PCTSINGLES   0.0029770  0.0007032   4.234            0.0000242 ***
## PCTBACHMOR   0.0209095  0.0005432  38.494 < 0.0000000000000002 ***
## LNNBELPOV   -0.0789035  0.0084567  -9.330 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3665 on 1715 degrees of freedom
## Multiple R-squared:  0.6623, Adjusted R-squared:  0.6615
## F-statistic: 840.9 on 4 and 1715 DF,  p-value: < 0.00000000000000022
```

```
# check the anova result
anova(lm1)
```

```
## Analysis of Variance Table
##
## Response: LNMEDHVAL
##              Df  Sum Sq Mean Sq  F value                Pr(>F)
## PCTVACANT     1 180.383 180.383 1343.093 < 0.00000000000000022 ***
## PCTSINGLES    1  24.543  24.543  182.741 < 0.00000000000000022 ***
## PCTBACHMOR    1 235.111 235.111 1750.586 < 0.00000000000000022 ***
## LNNBELPOV     1  11.692  11.692   87.054 < 0.00000000000000022 ***
```

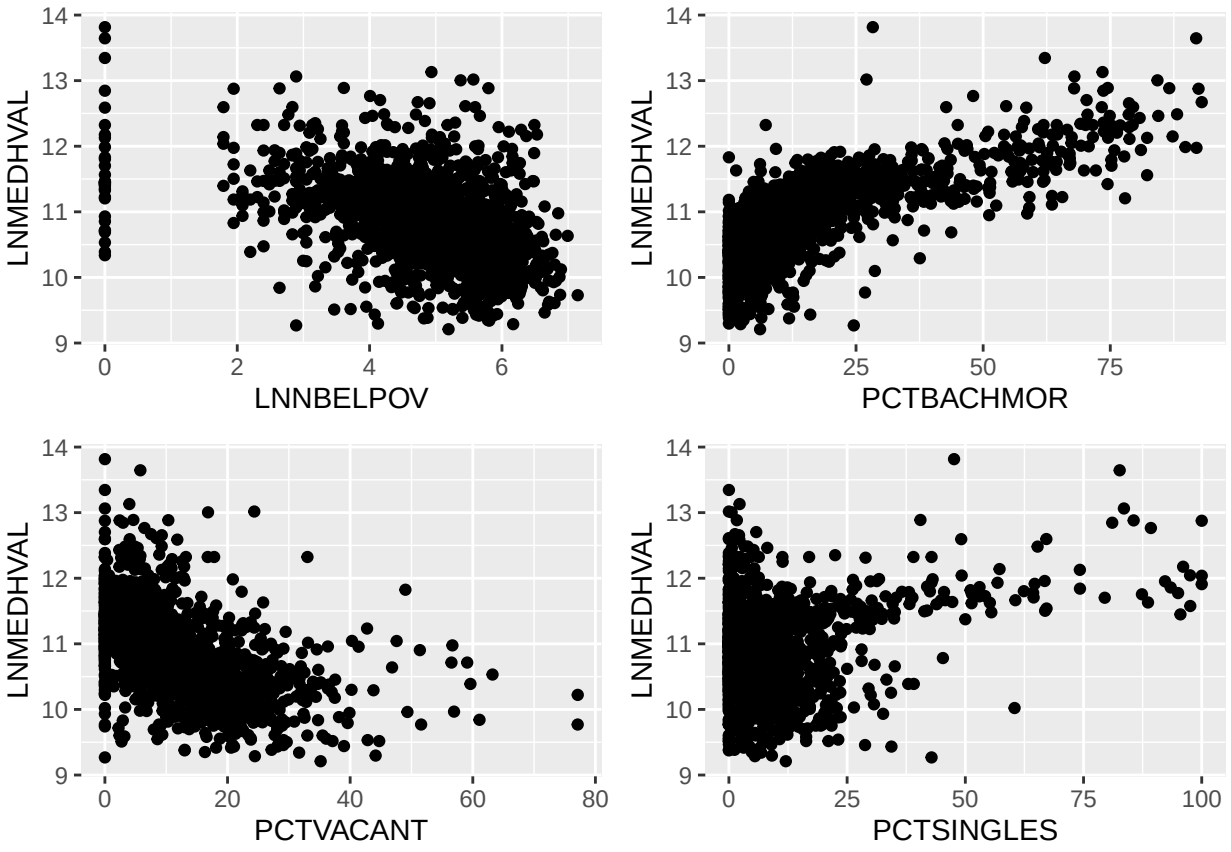
```
## Residuals  1715 230.332   0.134
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression Assumption Checks

In this section, we'll be talking about testing model assumptions. Through the exploratory analysis above, we already have a general understanding of the distribution of the variables through histograms. In this part, we will further test whether the assumption of the linear relationship between the dependent variable (LNMEDHVAL) and each predictor is valid.

Model Assumptions: Linearity

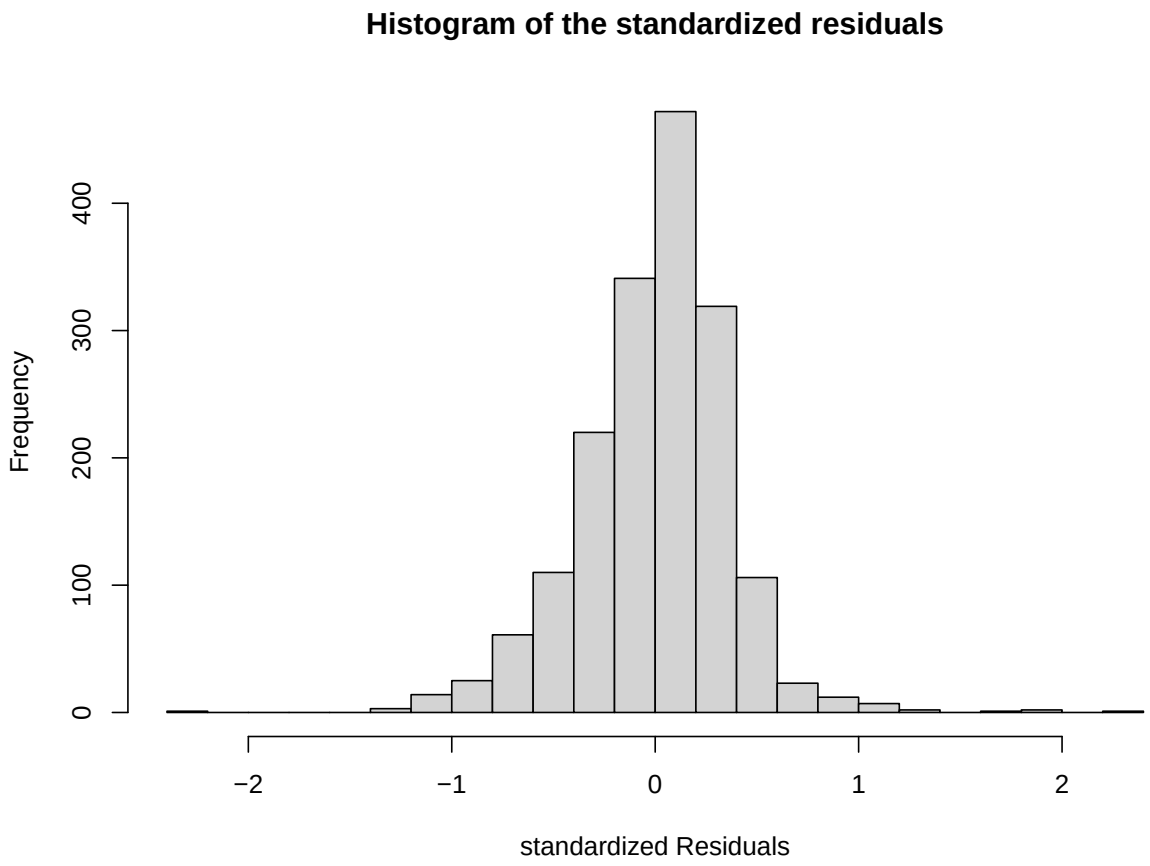
As we can see from the scatter plots, none of the relationships between the dependent variable (LNMEDHVAL) and each of the predictors appear to be strictly linear. Except for the relationship between median home value (LNMEDHVAL) and the percentage of residents with at least a bachelor's degree (PCTBACHMOR) which seems to be the most linear. The other scatter plots show data points either concentrated in the center or the lower left corner. Thus, with the exception of the relationship between LNMEDHVAL and PCTBACHMOR, the relationships between the dependent variable and most of the predictors deviate significantly from the assumption of strict linearity.



Normality of residuals

The normality of residuals is important for point estimation, confidence intervals, and hypothesis tests only for small samples due to the central limit theorem. In our model, the number of observations reaches more

than 1,400. Meanwhile, it's easy to notice from the histogram that most of the residuals are clustered around 0 and the trend seems to be normal.



Additional Checks: Homoscedasticity

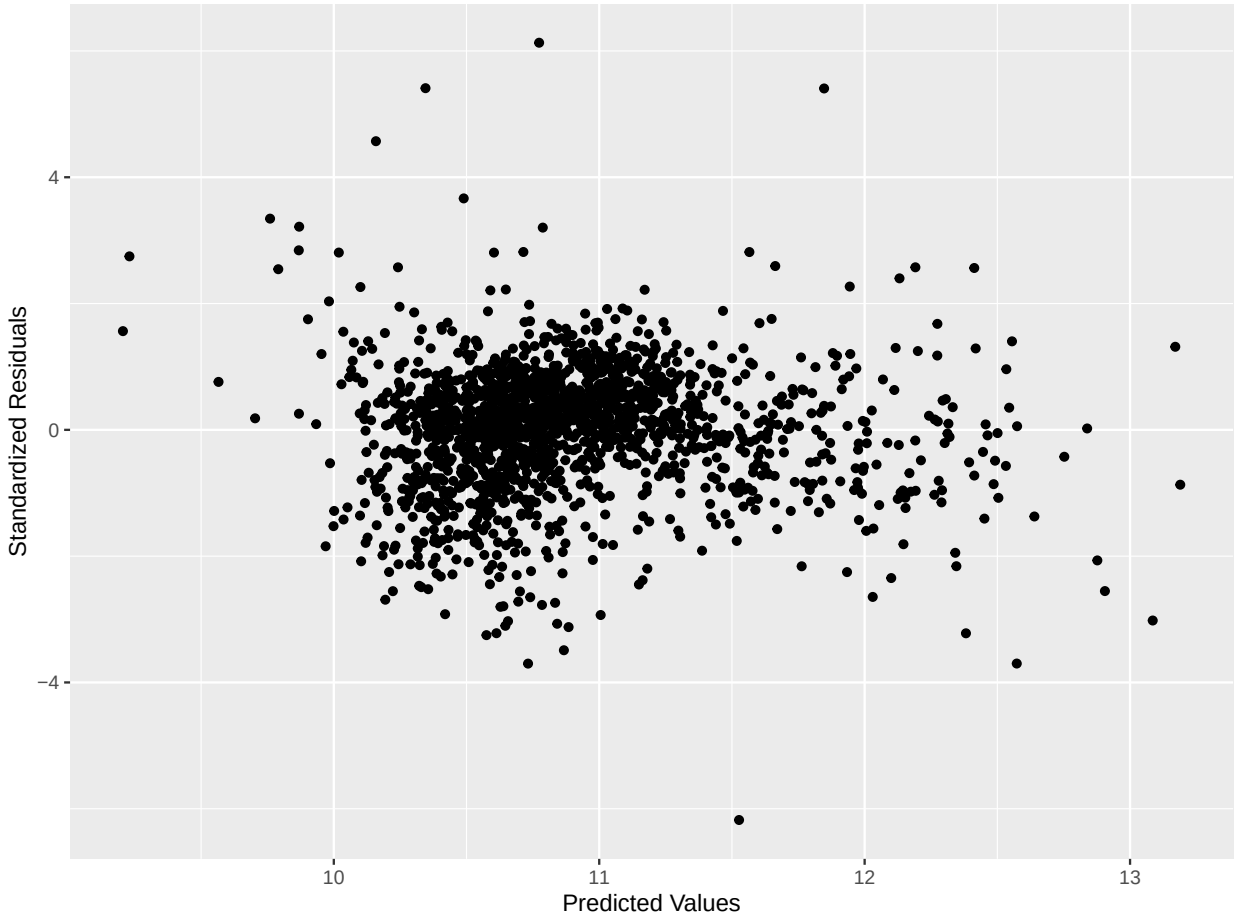
Standardized residuals are residuals divided by their standard error.

$$e_i^* \approx \frac{\epsilon_i}{s} \approx \frac{\epsilon_i}{\sqrt{\frac{SSE}{n-2}}}$$

They are used to compare residuals for different observations to each other. If a particular standardized residual is 2, then the residual itself is 2 (estimated) standard deviations larger than what would be expected from fitting the "correct" model.

By examining the 'Standardized Residual by Predicted Value' scatter plot, the goal is to discern the presence of heteroscedasticity — a scenario where the variance of residuals differs for various fitted values. A clear pattern or funnel shape in this scatter plot would indicate heteroscedasticity, suggesting systematic under-predictions or over-predictions by the model for certain ranges of fitted values. Upon analysis, the scatter plot demonstrates a relatively consistent spread of residuals across the range of fitted values, pointing towards homoscedasticity.

Additionally, some points lie further from the dense cluster, potentially indicating outliers. These extremely standardized residuals can influence model estimates and might warrant further investigation.



Spatial Autocorrelation

Observing the maps of the dependent variable and the predictors, there's a discernible spatial autocorrelation between the median house values and the percentage of residents holding at least a bachelor's degree. Prominent clusters of higher values can be identified in the northwest, northeast, center city, and university city regions of Philadelphia. This suggests that block groups nearby tend to exhibit similar values, challenging the notion that these observations are spatially independent.

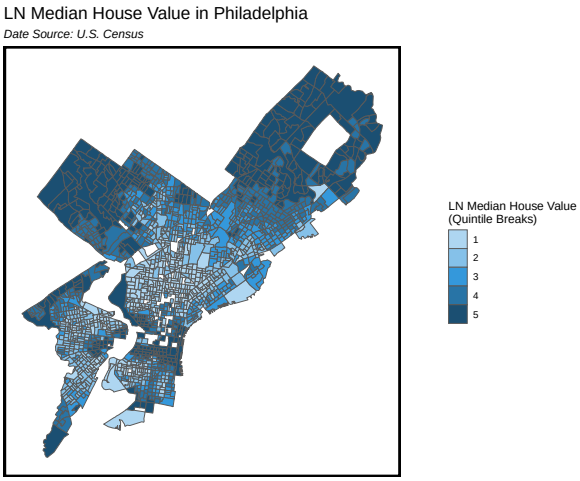


Figure 1

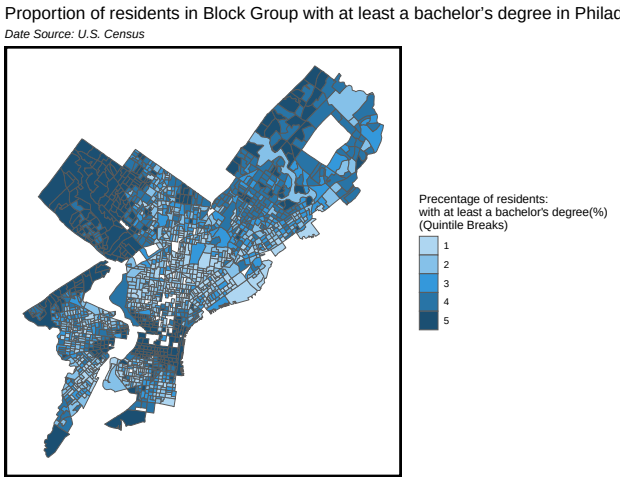
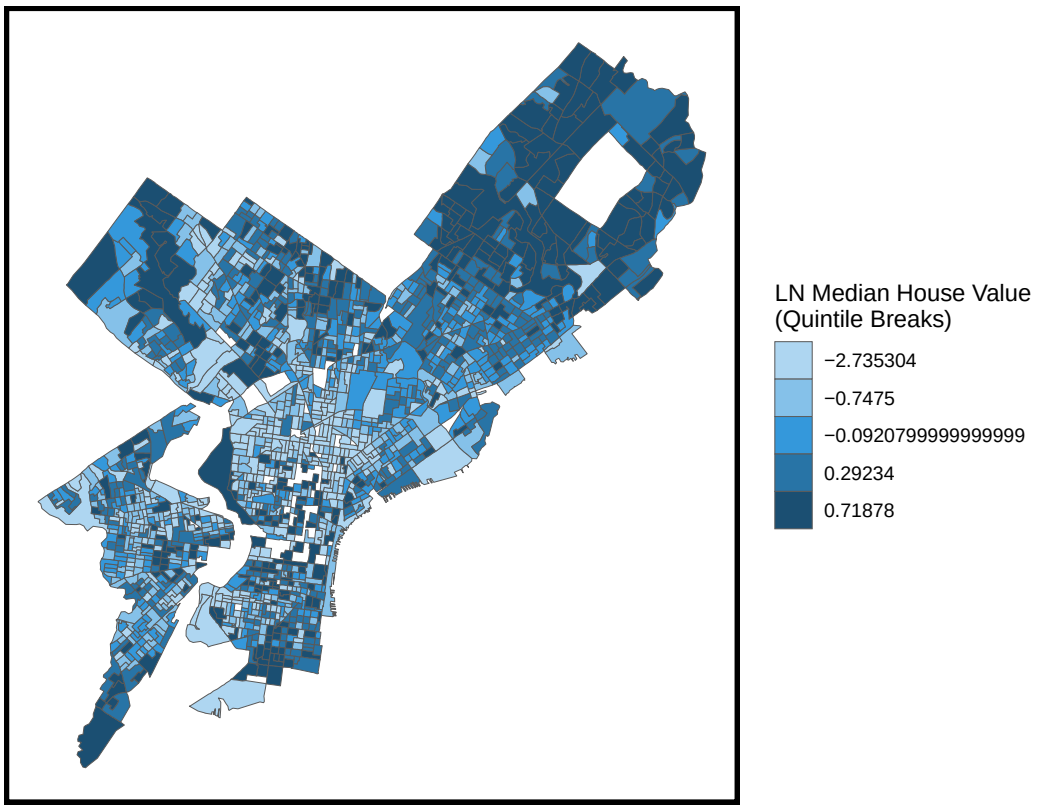


Figure 5

Standardized regression residuals map

From the map of the standardized regression residuals, there appear to be clusters of similar color. This suggests that there might be certain areas where the residuals are consistently high or low, which means the model might have systematically overestimated or underestimated the house values. In the map, given the clustering of similar colors in certain regions, there appears to be some degree of positive spatial autocorrelation in the residuals.



Map

Additional Models

using stepwise regression and determine the best model

As is depicted in the result of the stepwise model, all 4 predictors in the original model are retained in the final model. To be more specific, compared with other models with some of the variables dropped, the original has the smallest AIC, -3448.16, indicating that the original model does the best prediction.

```
best_model <- lm(LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNNBELPOV100, data=data_cor)
step <- stepAIC(best_model, direction="both")
```

```
## Start:  AIC=-3448.16
## LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNNBELPOV100
##
##                Df Sum of Sq    RSS     AIC
## <none>                      230.33 -3448.2
## - PCTSINGLES    1     2.407 232.74 -3432.3
## - LNNBELPOV100  1    11.692 242.02 -3365.0
## - PCTVACANT     1    51.543 281.87 -3102.8
## - PCTBACHMOR    1   199.014 429.35 -2379.0
```

```r
# stepwise regression - Analysis of Variance (ANOVA)
step$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNNBELPOV100
##
## Final Model:
## LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNNBELPOV100
##
##
##   Step Df Deviance Resid. Df Resid. Dev       AIC
## 1                       1715   230.3317 -3448.162
```

K-fold model

After performing cross-validation on the models, we obtained the following results: the RMSE for the original regression model stands at 0.366, while the secondary model has an RMSE of 0.443.

```r
rmse1
```

```
## [1] 0.3664306
```

```r
rmse2
```

```
## [1] 0.442712
```

Discussion and Limitations

Conclusion

In summary, we examined the relationship between median house values and several neighborhood characteristics using Philadelphia data at the Census block group level. More specifically, linear regression was done between the dependent variable MEDHVAL (Median House Values) and the predictors PCTBACHMOR (Percentage of Bachelor's Degree or Higher), NBELPOV100 (Number of Households living in Poverty), PCTVACANT (Percentage of Vacant Houses), and PCTSINGLES (Percentage of Single House Units). Because all variables are positively skewed, we applied log transformation to each variable, and decided to use log MEDHVAL, and log NBELPOV100 in our model, because although all variables are normalized after log transformation, the other three variables all have zero inflations, while NBELPOV100 and MEDHVAL do not have/have negligible frequency of zero values.

After that, regression assumptions were checked: multicollinearity, linear relationship between dependent variables and predictors, homoscedasticity of variance of residuals, and normality of residuals, where all the assumption requirements are successfully met in this model. The result of linear regression presented that all four predictors are significant with p values far less than 0.05, which means that the null hypothesis of the beta coefficient equal to zero can be rejected, and all four predictors are significantly correlated with the dependent variable LNMEDHVAL. Within those, predictor PCTBACHMOR demonstrates the most significant association with the LNMEDHVAL.

Finally, we applied stepwise regression and k-fold cross-validation to further validate our result. Overall, all four predictors are kept in the stepwise regression, and the rmse value (root mean squared error) in our model is significantly lower than the rmse value in the model that only has PCTVACANT and MEDHHINC (Median household income) as predictors. Therefore, it validates that our model performs better.

Model Quality

The conducted analysis indicates that the regression model is a good one overall. Firstly, based on the regression results, the high F-statistic means that we can reject the null hypothesis that none of the independent variables in the model is a significant predictor of the dependent variable. Also, the multiple R-square (0.6623) indicates that approximately 66.23% of the variance in median house values can be explained by the model. The result of the stepwise regression further supports the strength of the model. Specifically, all predictors in the original model are retained in the final model, which means that the original model is the best. All predictors have a statistically strong relationship with the dependent variable. Moreover, a cross-validation comparison reveals the original model's superiority. When considering all four variables, the model achieves a lower RMSE of 0.366, while a model based solely on 'housing vacancy' and 'median household income' has a higher RMSE of 0.443. This lower RMSE implies that the comprehensive model offers better predictive accuracy and alignment with actual values.

This analysis confirms a robust relationship between median household value and factors including residents' educational level, housing vacancy, the proportion of detached single-family houses, and number of households living in poverty. Specifically, housing vacancy rates and the percentage of detached single-family homes can be seen as reflections of housing quality and price trends. Meanwhile, poverty and education levels provide insight into the economic standing and purchasing power of potential homeowners.

While the current model is insightful, there is potential to enhance the model's comprehensiveness by introducing additional variables. For instance, the age of the housing stock exerts a profound influence on housing prices, and the number of bedrooms within a property can also significantly impact its market value. These considerations underscore the opportunity for enriching the model with a more comprehensive set of predictors.

Limitations

In terms of the model's limitations, it's important to note that the relationships between predictors and the dependent variable are not strictly linear. Only the relationship between PCTBACHMOR and MEDHVAL appears to be relatively linear. This violation of linearity assumptions could potentially introduce bias into parameter estimates and result in inaccurate outcomes. While attempts were made to address this by transforming some predictors using logarithmic transformations, certain predictors still contained significant zero values, so they were retained in their original form. Additionally, some predictors are interrelated with each other; for instance, the percentage of residents with at least a bachelor's degree is negatively correlated with the poverty status. Furthermore, an examination of the residuals map reveals the presence of spatial autocorrelation, which could lead to inefficient parameter estimates. As such, addressing nonlinear relationships, correlated observations, and spatial autocorrelation is crucial when modeling this data.

Also, for the NBELPOV100 variable, we use raw numbers instead of percentages, which might make it difficult to compare across different geographical regions or time periods, as it lacks contexts or normalization. This may further lead to misleading interpretations of the predictor's effect. On the other hand, it's also challenging to explain the practical implications of changes in the number of households in poverty without considering the total population or percentage of poverty.

Ridge and LASSO Regression

Ridge and LASSO regression are alternative regression techniques similar to each other that allow for multicollinearity, allow for a larger number of predictors than observations, and deal with overfitting by shrinking

the coefficient of variables to 0. However, both Ridge and LASSO regression will result in biased predicted values while the variance becomes lower, and they will increase the complexity of the model and the way of interpretation. The problem of Ridge regression is that all k predictors will be included, that is, it cannot perform variable selection. LASSO regression on the other hand can do variable selection but still has other limitations. Normally, ridge/LASSO regression will be applied when there is severe multicollinearity, few observations relative to the number of predictors, or we would like a better fit for unseen data than with OLS regression. In this case, as our model does not have the problem of multicollinearity and our number of observations is larger than the number of predictors, we assume that it is unnecessary to perform those two regression methods.

References

Economy League. (2023). Philadelphia's Housing Cost-burden: A Pre- and Post-Pandemic Comparison. https://www.economyleague.org/resources/philadelphias-housing-cost-burden-pre-and-post-pandemic-comparison