Gabriel Hernandez, Timothy Oliver, Ziyi Tang

Professor Eugene Brusilovskiy

MUSA 5000/CPLN 6710

19 October 2023

Assignment 1: OLS Regression

1. Introduction

Market research was professionalized in the 1920s and has grown into a ubiquitous and targeted effort which includes demographic analysis focusing on an area of study (Booker). This extends to simple examination of buying preferences for certain goods and at other times more complex relationships like those between housing values and neighborhood characteristics. A housing market is based on the agglomeration of individual decisions that take into account some form of analysis. Moreover, the success of real-estate planning and development requires a constant understanding of the community and the local built environment that is facilitated with data. This analysis aims to examine and discuss the relationship between median house values and characteristics of Philadelphia's neighborhoods through the use of multiple linear regression models of 2000 census data.

Prior research has shown associations between housing values and various socioeconomic characteristics of neighborhoods. Higher educational attainment, measured as the proportion of residents holding at least a bachelor's degree, has been linked with higher home values, potentially indicative of those residents' higher earning potential (Ioannides & Zabel 2008). Greater prevalence of poverty and vacant housing units tend to correlate with lower home values and neighborhood instability (Mikelbank 2008). The percentage of single-family homes within an area has mixed impacts, either improving values in distressed areas or potentially reducing values in already stable neighborhoods (Ioannides & Zabel

2008). While not exhaustive, these studies provide context for the predictive power of our chosen indicators in relation to median housing values in Philadelphia. Our analysis aims to describe these relationships and evaluate the degree to which certain neighborhood characteristics explain the variation observed in Philadelphia's housing market.

2. Methods

2.1 Data Cleaning

The original Philadelphia housing characteristics dataset that we use in our analysis was retrieved from the American Census and it provides aggregated data for 1,816 observations at the block group level. We removed block groups with a population under 40, those with no housing units or a median house value below $10,000, and one outlier block group in North Philadelphia with both a disproportionately high median house value above $800,000 and a low median household income less than $8,000. The result of this data cleaning left us with 1,720 observations. In addition to the geometry for each included block group in our dataset, we also have the following variables:

1) PCTBACHMOR - the proportion of residents in the block group with a bachelor’s degree or higher education
2) PCTVACANT - the proportion of housing units that are vacant
3) PCTSINGLES - the percent of housing units that are detached single family houses
4) NBELPOV100 - the number of households with incomes below 100% of the poverty level–denoting living in poverty
5) MEDHHINC - the median household income

6) MEDHVAL - the median value of all owner-occupied housing units in the block group and the dependent variable for the regression models used.

Further description of the dataset is value-specific and discovered through exploratory data analysis.

2.2 Exploratory Data Analysis

2.2.1 Summary Statistics and Distribution Analysis

To begin, we delve into the summary statistics of each variable and view the distributions of values for each. Using version 4.3.1 of the R programming language software, we retrieve the metrics of mean, standard deviation, median, and quartile ranges with the former two displayed in Figure 1. R provides robust functionality for both the primary linear regression and additional analyses of stepwise regression and k-fold cross-validation to ensure accurate and reproducible results alongside plots and maps for visualizing data in the different stages of analysis. For instance, we generate histograms to visualize the distributions of the data's variables in order to discern whether the distribution of each approximates a normal distribution. The initially intended regression method–multiple Ordinary Least Squares (OLS) regression–contains an underlying assumption of normally distributed residuals, and viewing a normal distribution in values for each variable can quickly affirm such. Given the potential pitfalls of non-normality in linear regression, we will also examine if a logarithmic transformation of any of the variables results in a more normal distribution, thereby enhancing the robustness of our subsequent analyses. While this might normally be done for only those variables observed to be not normally distributed, we do this for all independent variables resulting in the following additional variables and observed distributions.

2.2.2 Correlation Analysis

Next, we evaluate the pairwise correlations amongst the predictors. Understanding the degree to which our predictors are correlated can give insights into potential multicollinearity, which can impact the reliability of regression coefficients. Correlation quantifies the linear relationship between two quantitative variables. A positive correlation indicates that as one variable increases, the other does too, and vice versa for a negative correlation.

The degree of linear association is quantified using the sample correlation coefficient denoted by r and represented through the equation

$$r = \frac{\sum_{i=1}^{n} (X_i - \underline{X})(Y_i - \underline{Y})}{\sqrt{\sum_{I=1}^{n} (X_i - \underline{X})^2 \sum_{i=1}^{n} (Y_i - \underline{Y})^2}}$$

Where:

X_i and Y_i are individual data points.

$\underline{X}$ and $\underline{Y}$ represent the mean values of X and Y respectively.

The value of r can lie between -1 and 1. An r value of -1 signifies a perfect negative linear relationship where one data point changes inversely to another often at an equal or multiple rate. Similarly, a value of 1 denotes a perfect positive linear relationship with the change in one variable being an increase as the other increases also. Importantly, an r value of 0 implies no linear correlation between the variables, meaning changes in one variable do not predict changes in the other in any consistent direction.

These correlations in each of the predictor variables are found using the *cor* command in R to compute Pearson correlations. It is observed if the correlation between each predictor and the dependent variable, LNMEDHVAL, is relatively high both numerically by the r

value and visually with scatter plots. Multicollinearity is also measured looking at the numerical r value between pairs of predictor variables.

2.3 Multiple Regression Analysis

Multiple linear regression is used over simple linear regression to model the relationship between median home values and multiple predictor variables. In that sense, the utility of the created linear model is assessed through both the F-ratio or F-test and individual significance tests for each predictor following parameterization. For clarification, the F-test is a hypothesis test challenging a null hypothesis where all the predictor variables are jointly insignificant in predicting the dependent variable. In this model, LNMEDHVAL is the continuous dependent variable representing the natural log of median home values. The predictors are PCTVACANT, PCTSINGLES, PCTBACHMOR, and LNNBELPOV100, representing the percentage of vacant houses, the percentage of single unit houses, the percentage of respondents with bachelor' degrees or higher, and natural log of individuals below the poverty line, respectively. The Ordinary Least Squares regression model will yield an equation with the dependent variable as a function of each of the predictors as follows.

$$LNMEDHVAL =$$

$$\beta_0 + \beta_1 PCTVACANT + \beta_2 PCTSINGLES + \beta_3 PCTBACHMOR + \beta_4 LNNBELPOV + \varepsilon$$

Where

- β_0 represents the constant or intercept term for the dependent variable when each independent variable is 0 or unobserved in the model.

- $β_1, β_2, β_3, β_4$: represents the coefficients that quantify the relationship between each independent variable and the dependent variable.
- ε is the error term capturing the variation in LNMEDHVAL not explained by the independent variables.

The use of multiple linear regression requires fulfillment of five assumptions.

1) Linearity: The relationship between independent and dependent variables should be linear.
2) Independence of Observations: Each observation in the dataset is independent of other observations.
3) Normality of Residuals: The residuals (or errors) of the regression should be normally distributed.
4) Homoscedasticity: The variance of the errors should remain consistent across all levels of the independent variables.
5) No Multicollinearity: Independent variables should not be too highly correlated with each other. Here, a correlation r value at or above the 0.8 threshold is considered highly correlated.

Aside from the parameters of the linear regression equation, the parameter $σ^2$, representing the variance of the errors or residuals will also be estimated using the common Least Squares method, which minimizes the sum of squared residuals or the sum of squared differences between observed and predicted values. Therefore, the coefficients are calculated as:

$$β = (X^T X)^{-1} X^T Y$$

where goodness of fit will be measured by finding the coefficient of multiple determination, R^2, its adjusted value, and conducting the F-test on the overall model. R^2 represents the proportion of variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating a better fit. It is found using the equation

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

Where SSE is the sum of squared errors found via

$$\text{SSE} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

and SST is the total sum of squares found via

$$\text{SST} = \sum_{i=1}^{n} \left(y_i - \bar{y}_i\right)^2$$

The adjusted R^2 is based on the number of predictors in the model, providing a more accurate measure when multiple predictors are involved due to R^2's general increase as the number of predictors grows. It is calculated using

$$Adjusted\ R^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

where *n* is the number of observations and *k* is the number of predictors.

Lastly, the F-test used to measure goodness of fit tests the overall significance of the regression model under the following null and alternative hypothesis.

H_0 : All regression coefficients are equal to zero (i.e., no relationship).

H_1 : At least one regression coefficient is not zero.

Individual predictor variables also underwent individual t-test to determine significance and p-values describing such. For instance, the predictor variable PCTVACANT with coefficient β_1 would undergo a t-test with the following hypotheses:

$H_0 : \beta_1 = 0$, suggests PCTVACANT has no effect on LNMEDHVAL.

$H_1 :\ \beta_1 \neq 0$, indicates PCTVACANT does influence LNMEDHVAL.

These hypotheses are tested for each coefficient in the model to determine the significance of individual predictors.

2.4 Additional Analyses

2.4.1 Stepwise Regression

Stepwise regression is a method used to select the most important predictors in a dataset. It involves adding or removing predictors sequentially based on specific criteria, such as having the model minimize the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), until the most optimal model is identified. It suffers from a set of limitations including a risk of overfitting when too many predictors are desired, being a greedy algorithm that might not optimize the created model globally but at local regions of data, and that it assumes the list of candidate predictor variables is comprehensive. These limitations might allow the stepwise regression model to perform well on training data but poorly on new data. However, the predictors selected from this model or resulting fit can be useful in determining impact and goodness of fit for the multiple linear regression model.

2.4.2 K-Fold Cross Validation

K-Fold cross-validation is a resampling procedure used to evaluate machine learning models on a limited dataset. The process involves three steps:

1) Dividing the dataset into k subsets (in our case, k = 5).
2) For each fold, using *k-1* subsets to train the model and the remaining subset to test it.
3) Repeating this process k times, each time using a different subset as the test set.

The primary purpose of k-fold cross-validation is to provide a more robust estimate of the model's performance on unseen data. The results from all k tests are averaged to provide an overall performance metric.

In this case, we use the Root Mean Square Error (RMSE) to measure the differences between predicted values by the model and the actual values. A lower RMSE indicates a better fit to the data, and it is particularly useful for comparing the performance of different models.

3. Results

3.1 Exploratory Results

Variable	**Mean**	**Standard Deviation**
Dependent Variable		
Median House Value	66287.7331395	60006.0759895
Predictors		
# Households Living in Poverty	189.7709302	164.3184803
% of Individuals with Bachelor's Degrees or Higher	16.0813724	17.7695578
% of Vacant Houses	11.2885295	9.6284719
% of Single House Units	9.2264728	13.2492502

Figure 1**:** *Summary Statistics of Variables*

Looking at the mean and standard deviation of each variable, we can observed a fair likelihood that the majority of values are low while those that are high are much higher than the mean or other observation values. In many cases, the standard deviation is very close to the mean and even surpasses the mean for PCTBACHMOR and PCTSINGLES. Viewing histograms of the observations will give a clearer view into their distributions.

3.1.1 Histograms of original and log-transformed variables

Figures 2-1, 2-2, 2-3, 2-4, and 2-5 present histograms of the original variables paired with discussions on their relevant log transforms. Generally, one can observe from the histograms that none of the variables looks normal as given with many appearing zero-inflated. For our dependent variable, Median House Value, preliminary examination of the data indicates a right-skewed, nearly zero-inflated distribution. Increasing the number of histogram bins to 50 reveals that the 20K to 40K range is the most inflated part of the distribution.

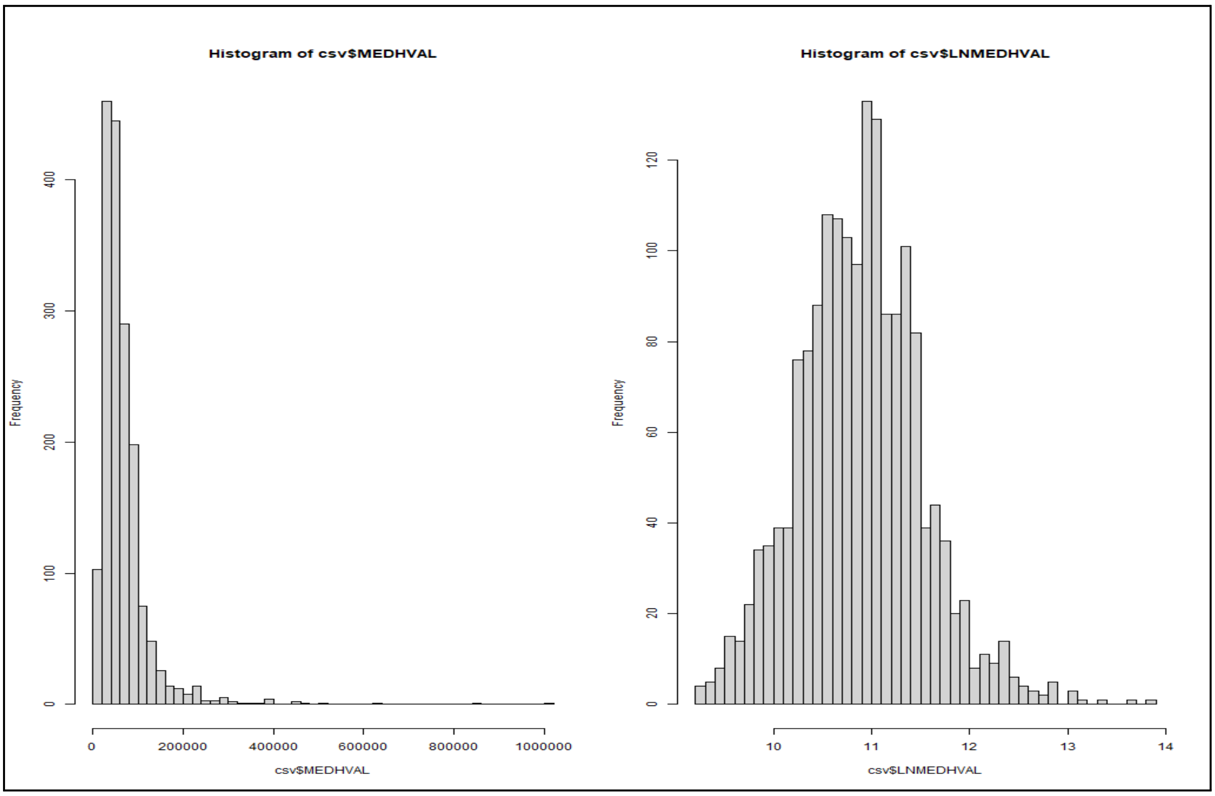


Figure 2-1**:** *Median House Value Histograms (As-Given & Transformed)*

For the number of households living in poverty (NBELPOV100), initial analysis shows the data is right-skewed but not zero-inflated. This suggests a log transformation may be beneficial for normalizing the data which is observed to be true with the transformed variable's histogram appearing much more normal.

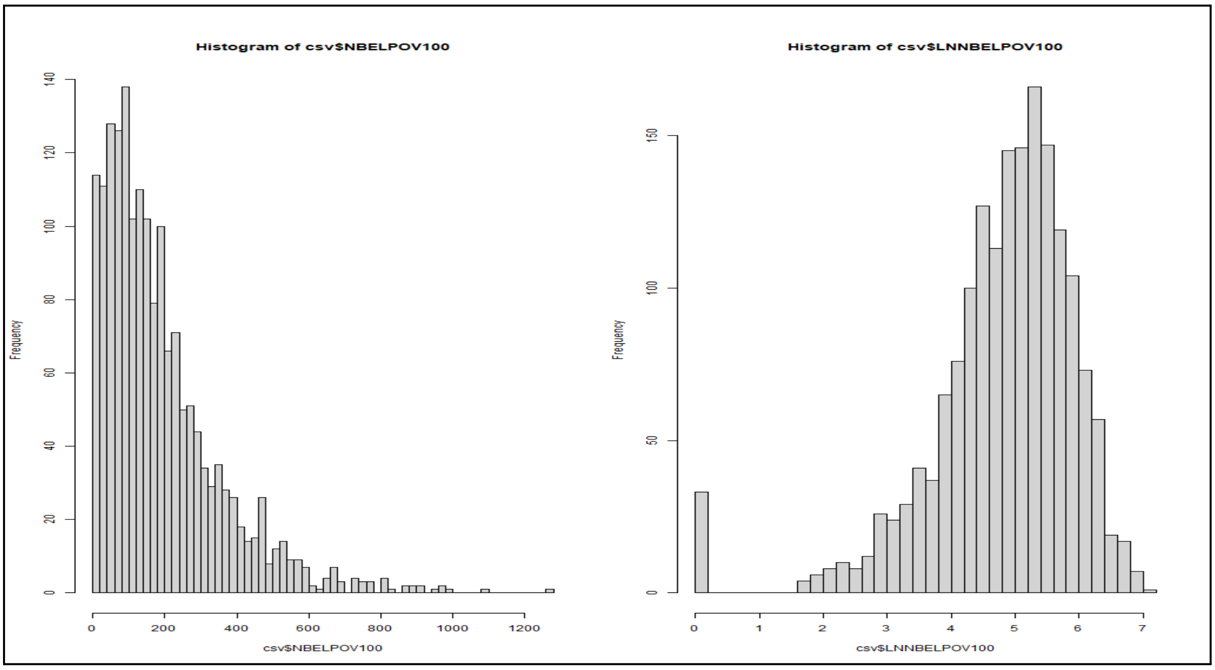
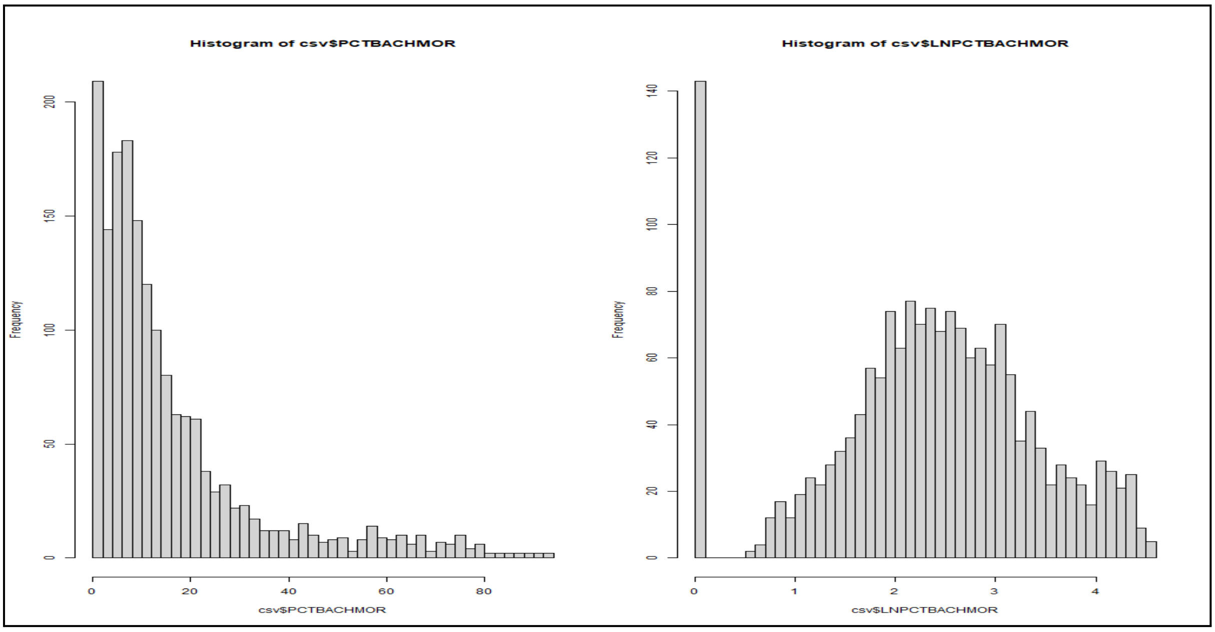


Figure 2-2: *Number of Households Living In Poverty Histograms (As-Given & Transformed)*

The histograms for the educational attainment variable PCTBACHMOR appears to have a distribution that is right-skewed and zero-inflated despite changes to the histogram such as increasing the break count to 50. Given the heavily skewed and zero-inflated nature, a log transformation is unlikely to normalize the data distribution which is shown to be true with zero-inflation being retained in the right histogram below.



<u>Figure 2-3</u>**:** *Percent with Bachelor's Degrees or Higher Histograms*

(As-Given & Transformed)

Initial examination of the percentage of vacant houses histogram shows the data has a right-skewed, zero-inflated distribution, similar to the variable PCTBACHMOR previously analyzed. Modifying the histogram parameters does not change the skewed, zero-inflated shape. Given the heavily skewed and zero-inflated nature, a log transformation is unlikely to normalize the data distribution, just as it was ineffective for PCTBACHMOR.

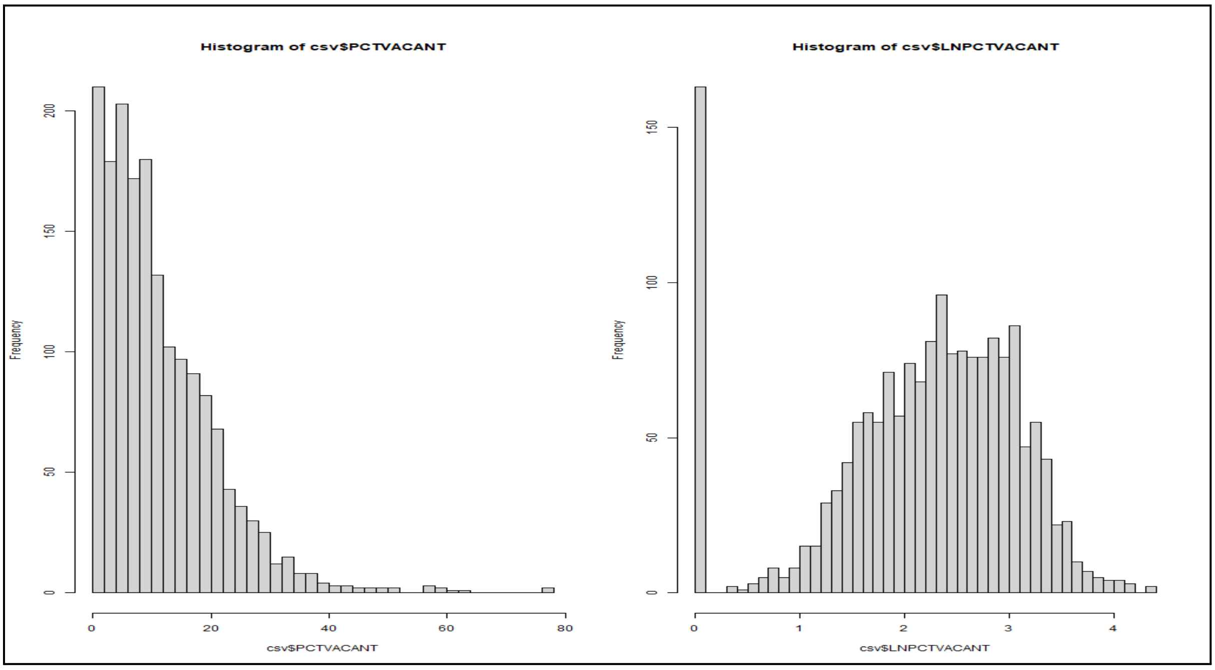


Figure 2-4**:** *Percent of Vacant Housing Units Histograms (As-Given & Transformed)*

Initial examination shows the data has a right-skewed, zero-inflated distribution, similar to the variable PCTBACHMOR and PCTVACANT previously analyzed. Modifying the histogram parameters does not change the skewed, zero-inflated shape. Given the heavily skewed and zero-inflated nature, a log transformation is expected to be unlikely in normalizing the data distribution, just as it was ineffective for PCTBACHMOR and PCTVACANT. Showing the histogram for the log-transformed variable once again shows this to be true.

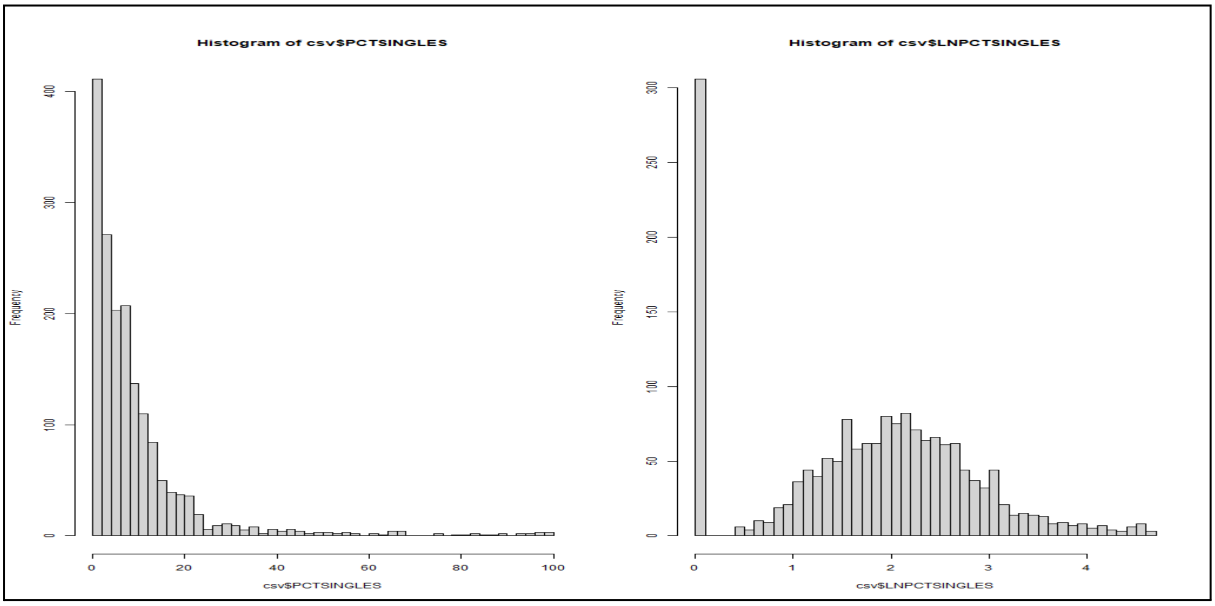
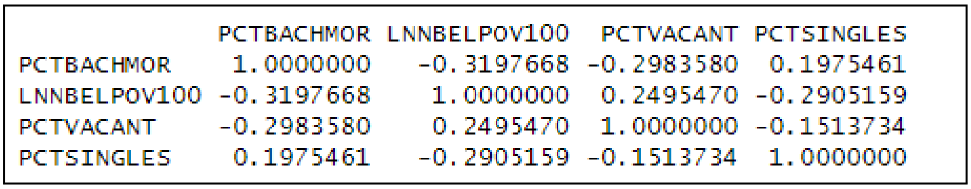


Figure 2-5: *Percent of Single Housing Units Histograms (As-Given and Transformed)*

Afterwards, we choose the log-transformed dependent variable, LNMEDHVAL, and predictor variable, LNNBELPOV100, alongside the other variables as they were received for our continued exploratory analysis and use in regression later on. This is done to increase the likelihood that the multiple linear regression assumption of normally distributed residuals is held. Other assumptions for linear regression will be examined in *Regression Assumption Checks*.

3.1.2 Pearson correlations

In order to gauge multicollinearity, the correlation matrix for each predictor variable seen in Figure 4 was yielded.The percentage of people with bachelor's degrees or higher education does not have strong correlation with any other predictor with the strongest correlation being a coefficient of -0.3198 with the natural logarithm of those below the poverty level. The natural log of those below 100% of the poverty level is not co-linear with any other predictor. Nor is the percent of household vacancies or the percent of single unit households.



```
             PCTBACHMOR LNNBELPOV100  PCTVACANT PCTSINGLES
PCTBACHMOR    1.0000000   -0.3197668 -0.2983580  0.1975461
LNNBELPOV100 -0.3197668    1.0000000  0.2495470 -0.2905159
PCTVACANT    -0.2983580    0.2495470  1.0000000 -0.1513734
PCTSINGLES    0.1975461   -0.2905159 -0.1513734  1.0000000
```

Figure 3: *Correlation Matrix of Predictor Variables*

With these results, no two predictors are shown to be correlated to another with a correlation coefficient of 0.8 or higher leading to no removal of variables from our regression model.

3.1.3 Choropleth Maps

Upon closer inspection of the maps, striking similarities can be observed between the LNMEDHVAL Map and the PCBACHMORE Map, indicating a robust association between the predictor—percentage of individuals with bachelor's degrees—and the dependent variable, which is the natural logarithm of median home values. For instance, the highest variable values are observed in North Philadelphia and somewhat near Center City though there are widely different groups of census tracts in the center of the county. A similar relationship is evident between other predictors such as the percentage of single houses, the natural logarithm of the number of people below the poverty line, and the natural logarithm of median home values. In stark contrast, the predictor representing the percentage of vacant houses deviates significantly from the others, as it lacks the clear and consistent patterns observed in other maps trending instead towards low percentage values where the other variables were high. This discrepancy suggests that the percentage of vacant houses may not exhibit a strong correlation with the dependent variable, natural logarithm of median home values.

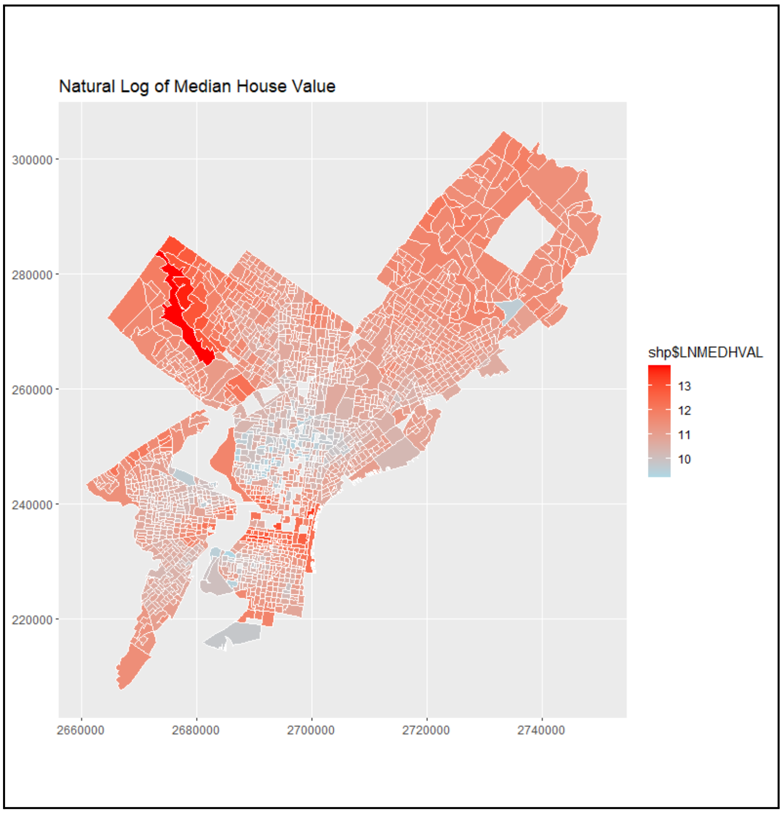


Figure 4-1: *LNMEDHVAL in Philadelphia County Census Tracts, 2000*

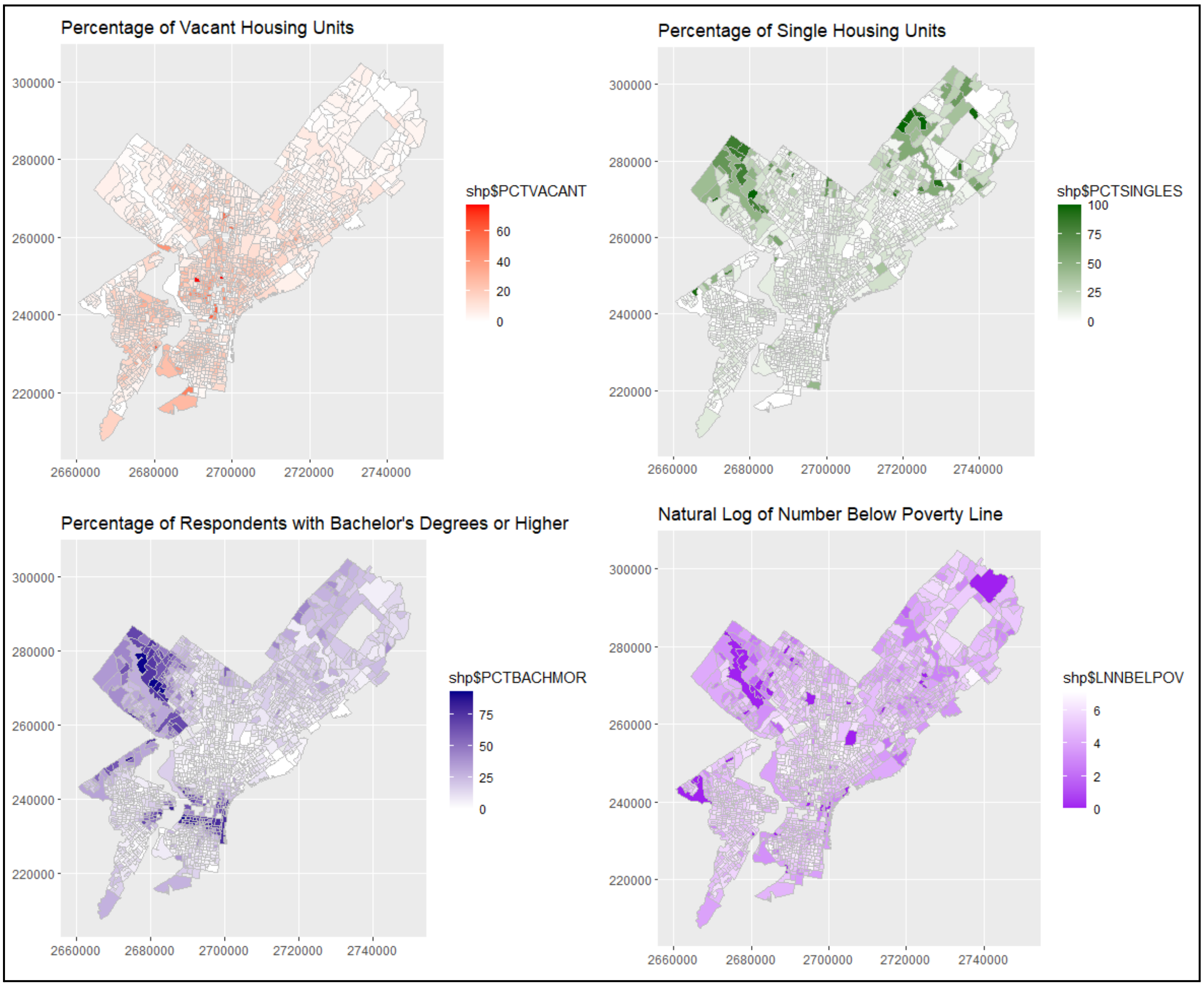

Percentage of Vacant Housing Units
shp$PCTVACANT
60
40
20
0
Percentage of Single Housing Units
shp$PCTSINGLES
100
75
50
25
0
Percentage of Respondents with Bachelor's Degrees or Higher
shp$PCTBACHMOR
75
50
25
0
Natural Log of Number Below Poverty Line
shp$LNNBELPOV
6
4
2
0
300000
280000
260000
240000
220000
2660000
2680000
2700000
2720000
2740000

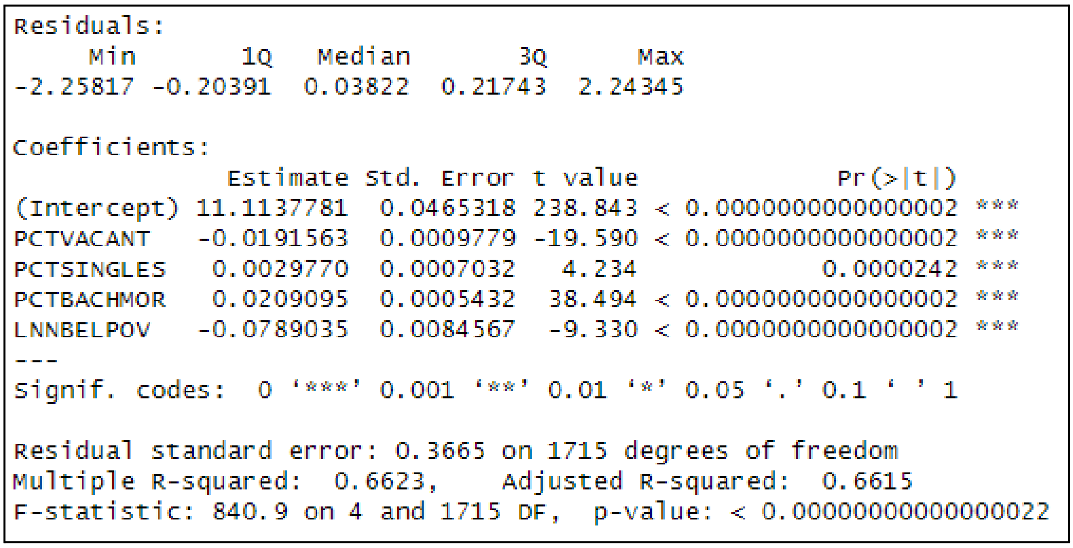
Figure 4-2: *Predictor Variables in Philadelphia County Census Tracts, 2000*

On the whole, the census tracts appear to be largely similar in value to the tracts nearby it in each map. While there are some stark transitions from high to low values across tracts, a slight spatial component may be present that the current predictor variables might not account for.

3.2 Regression Results

3.2.1 Summary table

Using the linear model command, *lm*, in R, the multiple linear regression model was created with the dependent and predictor variables as indicated above and resulted in the following summary output.



```
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
```

Figure 5: *Linear Regression Output*

To enumerate, we regressed the natural log of median home values (LNMEDHVAL) on the percentage of vacant houses (PCTVACANT), percentage of single houses (PCTSINGLES), percentage with bachelor's degrees (PCTBACHMOR), and the natural log of the number below the poverty line (LNNBELPOV). The regression output reveals the following:

First, the association between the percentage of Vacant Houses (PCTVACANT) and LNMEDHVAL is highly significant and negative as displayed by a p-value less than 0.001 and a coefficient β_1 that approximately equals -0.0192. Acknowledging our log-transformation of the dependent variable, this coefficient yields the interpretation that a one unit increase in Vacant Houses is associated with a $(e^{-0.0192} - 1) * 100\%$ change or 1.90% decrease in median home values, provided the other three predictors are held constant.

The percentage of Single Houses (PCTSINGLES) has a significant and positive association with LNMEDHVAL where the p-value is lower than 0.001 and β_2 approximately equals 0.002977. Specifically, an increase of one unit in Single Houses corresponds to a 0.30% increase in median home values, holding the other three predictors constant..

The relationship between the percentage with Bachelor's Degrees or higher (PCTBACHMOR) and LNMEDHVAL is highly significant and positive with a p-value lower than 0.001 and coefficient β_3 that equals 0.0209095. This is interpretable as a one percentage increase in Bachelor's Degrees being associated with a 2.11% increase in median home values, holding the other three predictors constant.

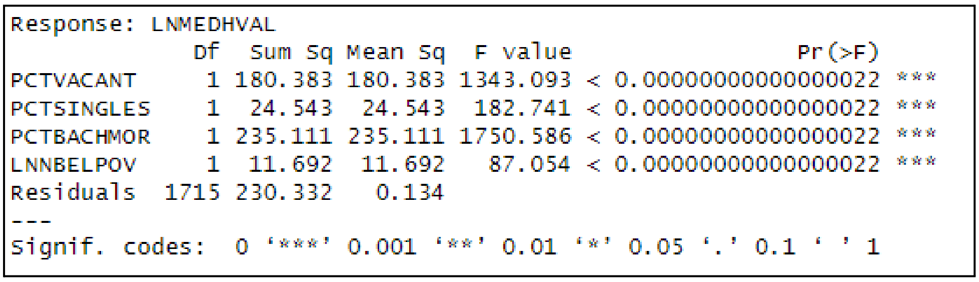
There is a highly significant and negative relationship between the natural log of individuals below the poverty line (LNNBELPOV100) and LNMEDHVAL with a p-value lower than 0.001 and coefficient β_4 that equals 0.0789035. Considering that both predictor and dependent variable are log-transformed, a one percent increase in LNNBELPOV100 is associated with a $(1.01^{-0.0789035} - 1) * 100\%$ change or a 0.078% decrease in median home values, holding the other three predictors constant.

These p-values demonstrate that should the null hypothesis be true and there is truly no relationship between the respective predictors and LNMEDHVAL (i.e., the null hypothesis for each coefficient being equal to 0), then the likelihood of observing these coefficient estimates is extremely low (less than 0.0001) in the case of each predictor. Consequently, we can confidently reject the null hypothesis for each predictor, suggesting that each variable does have a significant relationship with LNMEDHVAL.

The model explains about two-thirds (66.23%) of the variance in LNMEDHVAL, as evidenced by the R^2 value of 0.6623 and the Adjusted R^2 value of 0.6615. The latter communicates that the larger number of predictor variables was not especially detrimental to our model. Additionally, this is considerably higher than the provided example. The extremely low p-value associated with the F-ratio, which is $2.2E^{-16}$, indicates that we can reject the null hypothesis that all coefficients in the model are 0, affirming that the predictors in the model collectively exhibit a significant relationship with LNMEDHVAL.

3.2.2 ANOVA table

The sum of squares for residuals, which is 230.332, represents the total variance that the model could not explain and is also referred to as the error sum of squares. For this model specifically, it is the summation of the squared differences between our model's predicted LNMEDHVAL values, **y-estimates**, and the observed values, y_i. This indicates the unexplained variance in the dependent variable LMEDHVAL after accounting for the variance explained by the predictors in the model with each predictor variable's individual explained sum of squares or sum of squares due to regression (SSR) shown below in Figure 6.



```
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
```

Figure 6: *ANOVA Output*

3.3 Regression Assumption Checks

Following the completion of the aforementioned methods, each of the five linear model assumptions were tested by observing plots of the variables and their residuals. While the distributions of each variable was discussed earlier in *Histograms of original and log-transformed variables*, the remaining checks follow.

3.3.1 Scatter plots (dependent variable and predictors)

Considering the relationship between the predictor variables and the dependent variable of LNMEDHVAL was most easily done through relational scatter plots as seen in Figure 7.

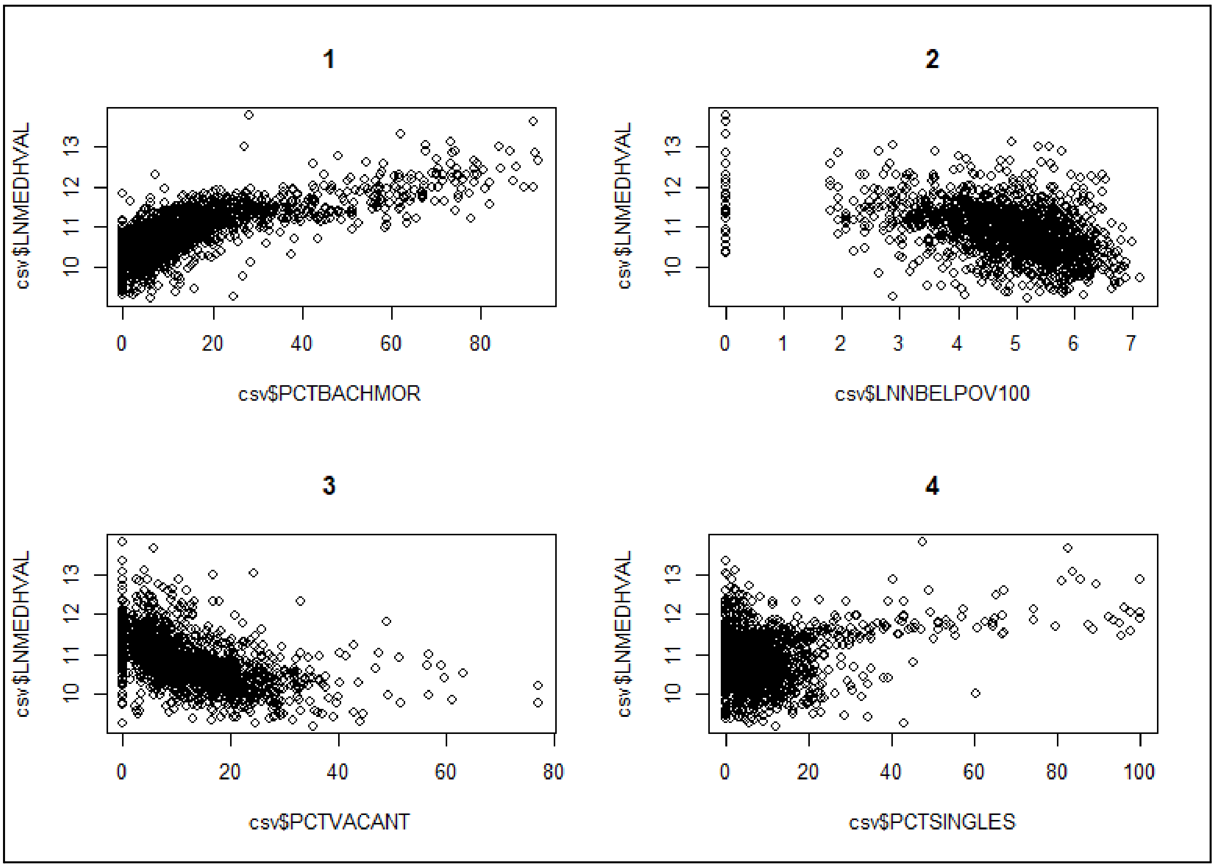


Figure 7: *LNMEDHVAL as a Function of Predictors*

Scatter Plot 1 in Figure 3 shows the relationship between the percentage with bachelor's degrees (PCTBACHMOR) on the x-axis and the natural log of median home values (LN_MEDHVAL) on the y-axis. It appears that as the percentage with bachelor's degrees increases, the natural log of median home values also tends to increase. However, the relationship does not seem to be perfectly linear, especially given the leveling off at higher percentages. It might be more appropriately described by a curve, possibly logarithmic.

Scatter Plot 2 (top-right of Figure 7) displays the relationship between the natural log of the number below the poverty line per 100 (LNBLPOV100) and the natural log of median home values. The points are densely packed without a clear upward or downward trend. However, this relationship doesn't seem to be linear. The dense clustering of data points suggests little variation in the dependent variable with changes in the predictor.

Scatter Plot 3 (bottom-left of Figure 7) illustrates the relationship between the percentage of vacant houses (PCTVACANT) and the natural log of median home values. Initially, there seems to be a decrease in the median home values as the percentage of vacant houses increases, but then there's an erratic spread of points without a clear trend. The relationship does not seem to be linear. The initial decline followed by the dispersed spread of data points suggests a more complex relationship.

Scatter Plot 4 (bottom-right of Figure 7) shows the relationship between the percentage of single houses (PCTSINGLES) and the natural log of median home values. Most of the data points cluster towards the left, suggesting that the majority of areas have a higher percentage of single houses with similar median home values. However, the relationship is not linear. The dense concentration on the left and the scattered data points to the right indicate that there isn't a consistent linear trend between the two variables though attempting to fit a line to the observed points would likely yield a line of best fit with a slightly positive slope. Overall, we anticipate a low relationship between these variables and the natural log of median house value with that for PCTSINGLES and PCTBACHMOR being positive and LNNBELPOV100 and PCTVACANT having negative relationships with the dependent variable.

3.3.2 Histogram of standardized regression residuals

The histogram in Figure 8 depicts the distribution of standardized residuals (shp$stand_resid) in the linear regression model. For clarity, the standardized residuals are the residuals for each observation-prediction pair divided by the standard error so that residuals for different observations can be compared. The x-axis represents the standardized residual value, ranging from -6.177958 to 6.130070, while the y-axis denotes frequency, with values reaching slightly beyond 600. The shape of the histogram resembles a bell curve, which is characteristic of a normal distribution. There are no evident long tails or sharp peaks, and both sides of the histogram appear to mirror each other quite closely. Therefore, the standardized residuals seem to follow a reasonably normal distribution based on the histogram.

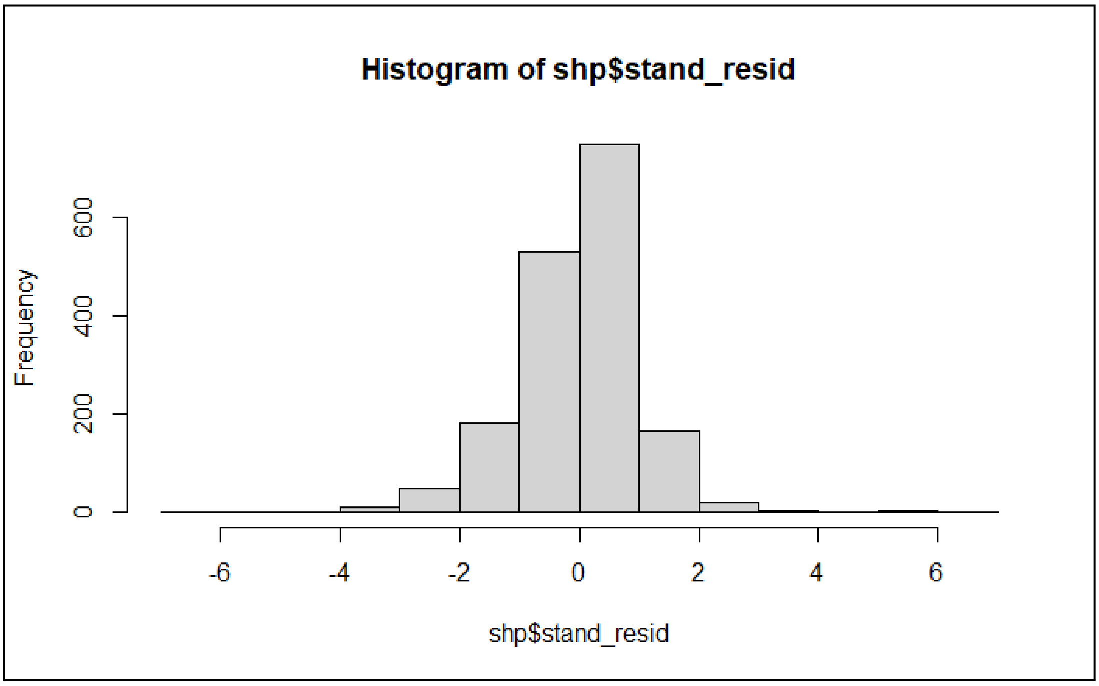


Figure 8: *Histogram of Standardized Residuals*

3.3.3 Scatter plot of standardized residual by predicted value

The scatter plot shown in Figure 9 displays predicted values on the x-axis ranging approximately from 10 to 13, and standardized residuals on the y-axis ranging from around -6 to 6. The densest concentration of data points is centered around a predicted value of roughly 11 and standardized residuals close to 0. This concentration suggests that for many observations, the predicted values closely align with the actual observed values. Heteroscedasticity refers to the situation where the variance of the residuals is not constant across all levels of the independent variables. In the plot, there's a slight fanning out or spread of residuals as the predicted values increase, especially noticeable towards the higher end of the predicted values. This suggests potential heteroscedasticity, indicating that the model might not be capturing all the variability in the data uniformly across the range of predicted values.

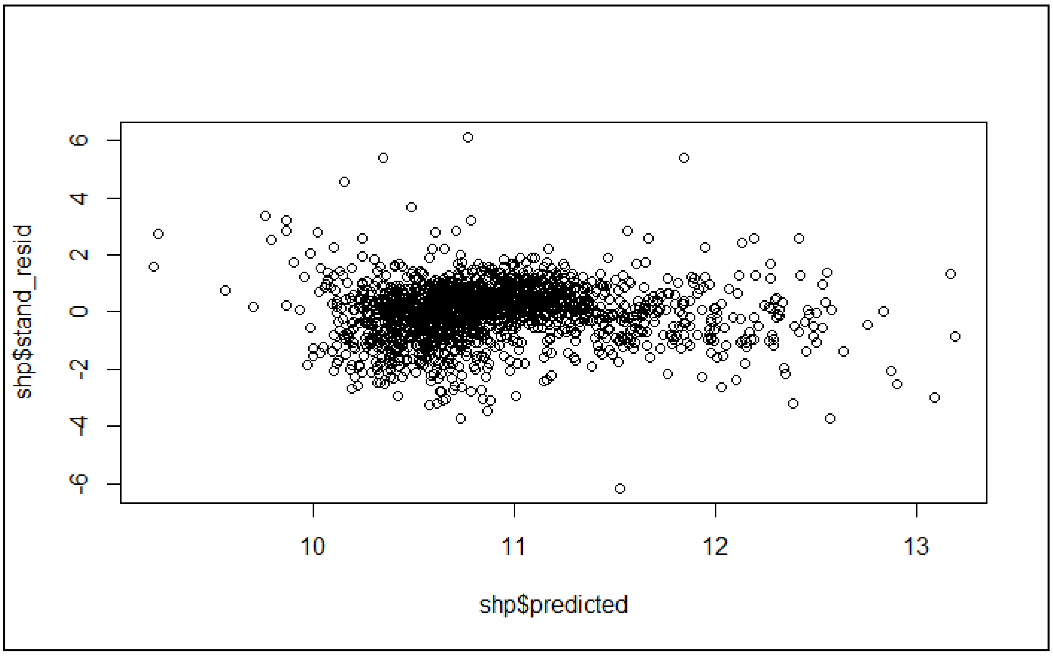


Figure 9: *Scatter plot of standardized residual by predicted value*

While the majority of data points are clustered around the center, there are some data points that lie farther away from the main concentration, especially on the top and bottom extremes of the standardized residuals. These could be potential outliers or influential points that could affect the robustness of the model. There also does not seem to be any distinct or systematic trend in the residuals, which is a good sign. A clear pattern or curve would indicate that the model is missing a relationship, but the random spread suggests that such systematic biases are likely absent. The model appears to fit a majority of the data well, given the dense clustering around the center. However, potential heteroscedasticity is a concern, suggesting that the model may not capture variability uniformly across the range of predicted values. Additionally, potential outliers need to be investigated further to understand their influence on the model.

3.3.4 Choropleth maps

To reiterate observations about the choropleth maps for each variable, the census block groups appear to be largely similar in value to the block groups nearby it in each map leading to the conclusion that there is spatial autocorrelation in the variables and that block group observations are not independent of each other. Further testing is needed for a definite conclusion, but this pattern of spatial autocorrelation also continues in the choropleth map shown in Figure 10. It displays the standardized regression residuals across different block groups and enables us to discern how the observed values differ from the values predicted by our regression model in a spatial context.

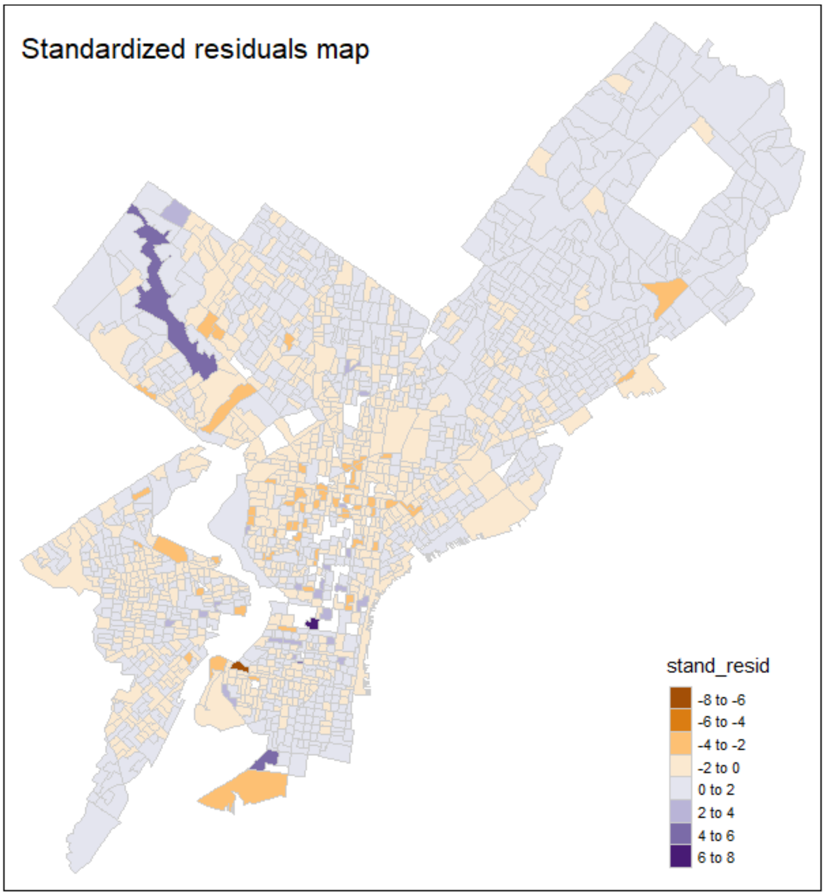


Figure 10: *Choropleth Map of Standardized Regression Residuals*

From the gradient scale provided, regions shaded in lighter colors represent areas with residuals closer to zero, implying that the observed and predicted values in these regions are relatively close. Conversely, areas colored in darker shades, either on the positive or negative end of the scale, indicate larger discrepancies between the observed and predicted values.

Upon close inspection, there are evident spatial patterns in the residuals. Regions with deep blue hues, signifying standardized residuals between 6 to 8, are not isolated but tend to cluster in certain areas (e.g. central and north Philly), suggesting a potential underprediction by our model in these localities. Similarly, the pockets of orange, indicating standardized residuals between -8 to -6, hint at an overprediction in these zones.

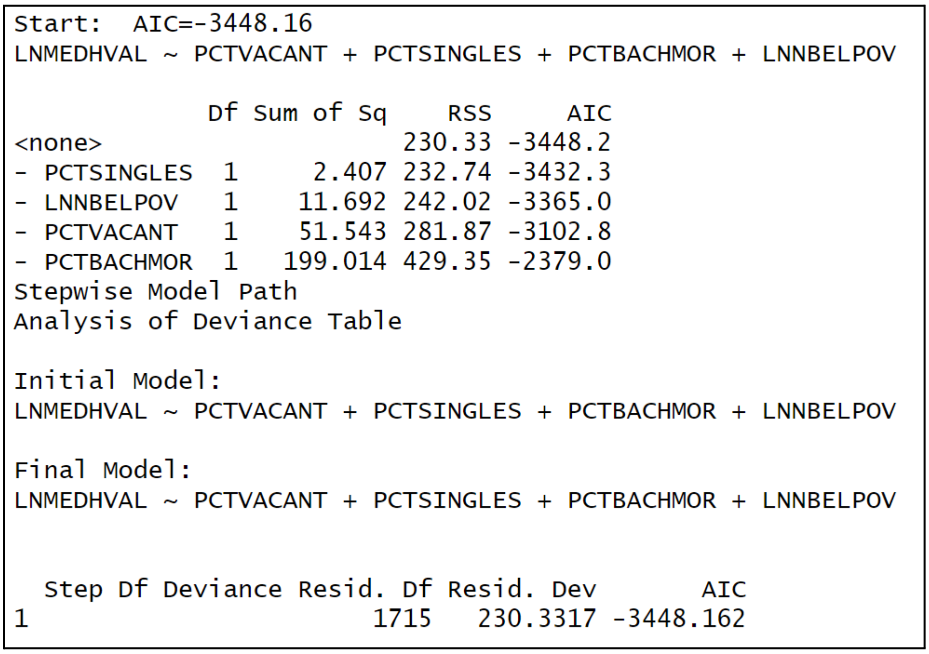
The presence of such clusters is indicative of spatial autocorrelation in the residuals, which implies that nearby locations are not independent of each other. The spatial patterns observed suggest that there could be underlying spatial processes or omitted variables that influence the dependent variable in a manner not captured by the regression model.

Overall, the data appears to meet only two of the five assumptions of linear regression. The assumptions of normality of residuals and no multicollinearity are met while linearity, independence of observations, and homoscedasticity were shown to not be followed through discussions on the scatter plots of LNMEDHVAL versus the predictors, the observed spatial pattern in the choropleth maps, and the observed heteroscedasticity in the standardized residuals scatter plot, respectively. The failed assumptions imply that the linear regression model can yield incorrect or misleading results and that some other form of regression might be better suited to this dataset.

3.4 Additional Models

3.4.1 Stepwise Regression

We run stepwise regression and determine the best model based on the Akaike Information Criterion. The table presented in Figure 11 describes the results of a stepwise regression.



```
Start:  AIC=-3448.16
LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNNBELPOV

             Df Sum of Sq    RSS     AIC
<none>                    230.33 -3448.2
- PCTSINGLES  1     2.407 232.74 -3432.3
- LNNBELPOV   1    11.692 242.02 -3365.0
- PCTVACANT   1    51.543 281.87 -3102.8
- PCTBACHMOR  1   199.014 429.35 -2379.0
Stepwise Model Path 
Analysis of Deviance Table

Initial Model:
LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNNBELPOV

Final Model:
LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNNBELPOV


  Step Df Deviance Resid. Df Resid. Dev       AIC
1                       1715   230.3317 -3448.162
```

Figure 11: *Stepwise Regression Output*

The regression started with a model including all four predictor variables (PCTVACANT, PCTSINGLES, PCTBACHMOR, LNNBELPOV) to predict the dependent variable, LNMEDHVAL. The Akaike Information Criterion (AIC) for this model is -3448.16. AIC is a measure used to compare different models, with a lower AIC indicating a better fit given the number of predictors. The next section of the table indicates the effect of removing each predictor on the model. For instance, the Sum of Squares increases by 2.407, the Residual Sum of Squares (RSS) becomes 232.74, and the AIC is -3432.3 when

PCTSINGLES is removed from the model. This AIC is higher (worse) than the original AIC, indicating that removing PCTSINGLES might not be a good idea.

The same logic applies to the other variables. In this table, removing PCTBACHMOR has the most substantial negative impact on the model, as evidenced by the dramatic increase in the AIC. The initial model includes all four predictor variables, and the final model is the same as the initial model, suggesting that the stepwise procedure did not find it beneficial to remove any of the predictors.

3.4.2 K-fold cross-validation

Lastly, we employed the k-fold cross-validation technique, setting k to 5. By segmenting the dataset into five equal parts and iterating through each as a testing set, we gain a more holistic view of a model's potential performance on unseen data.

The original regression model was first subjected to this k-fold cross-validation method. The outcome of this evaluation provided a root mean square error (RMSE) of 0.3664. This metric signifies that the model's predictions, on average, deviated from the actual values by a margin of 0.3664 units. This value serves as our benchmark for comparison when evaluating alternative or modified models. We continued our analysis by re-running the k-fold cross validation on a model with only two predictors: PCTVACANT and MEDHHINC. The resulting RMSE for this model was observed to be 0.4427. This indicates that when the model relied solely on these two predictors, its predictions, on average, were off by about 0.4427 units from the true values.

Juxtaposing the two models, the original regression model with an RMSE of 0.3664 demonstrated superior performance in comparison to the comparative model that relied solely on PCTVACANT and MEDHHINC as predictors with a higher RMSE value of 0.4427. This suggests that the original model might capture the intricacies of the data more efficiently.

However, it's also essential to consider the trade-offs: while the revised model may not perform as well, its simplicity might offer more straightforward interpretability in certain contexts.

4. Discussion and Limitations

This analysis examined the relationship between median house values and several Philadelphia neighborhood characteristics including educational attainment, poverty, housing vacancy, and the presence of single-family homes. A multiple linear regression model was developed using 2000 census data at the block group level to analyze their relationships. The results demonstrated statistically significant associations between median home values and all four neighborhood predictors. As expected, higher educational attainment and prevalence of single-family homes were positively associated with home values, while greater poverty and housing vacancy correlated with depressed home values. The model explained about 66% of the variation in median home values and the F-statistic p-value of $2.2E^{-16}$ indicates a reasonably good model fit and with a high predictor significance.

However, we want to acknowledge some of the limitations in our analysis. First, only two of the five assumptions of linear regression were met in this analysis. The assumptions of linearity, independence of observations, and homoscedasticity were shown to not be followed. The assumption of normality of residuals was observed but could also be considered violated given the non-normal distributions of the un-transformed variables and the suggestion of a spatial pattern in the choropleth map of the standardized residuals. Transformations only partially normalized the data. This is a problem because non-normal residuals can bias coefficient estimates and significance tests. Heteroscedasticity is indicated in the residual plots, which implies non-constant variance. Lastly, using the raw number of households experiencing poverty, rather than a percentage, also has drawbacks. That is because the scale differs greatly across block groups of varying populations, resulting in concentrations of poverty that are obscured in more populous block groups.

The stepwise regression retained all four predictors, affirming their utility in predicting housing values. However, the model may be improved by incorporating additional variables like crime rates or public school quality which can influence housing demand. It could also be improved if we accounted for some of the spatial autocorrelation that we observed in the choropleth map of our residuals. Ridge and LASSO regression are regularization techniques used to prevent overfitting when many weakly predictive variables exist, especially when more predictor variables exist than observations, by constraining coefficients to reduce model complexity. These methods are inappropriate for this study since there are only four predictors which is much lower than the 1,720 observation count.

Overall, the model provides meaningful insights into the components of housing values that we observe. Unfortunately, our analysis has limitations related to the failure of statistical assumptions and its scope of predictors. Further statistical methods could examine non-linear relationships, interacted effects, and spatial autocorrelation between neighborhoods.

Works Cited

Booker, Bel. “The 100-Year History of Market Research - 1920 to 2020.” *Attest*, 7 Apr. 2023, www.askattest.com/blog/articles/history-of-market-research.

Ioannides, Yannis M. and Jeffrey E. Zabel, “Interactions, neighborhood selection and housing demand.” *Journal of Urban Economics*, Volume 63, Issue 1, 2008, pp. 229-252, ISSN 0094-1190, https://doi.org/10.1016/j.jue.2007.01.010.

Mikelbank, Brian A. “Spatial Analysis of the Impact of Vacant, Abandoned, and Foreclosed Properties.” *Federal Reserve Bank of Cleveland, Community Development Reports*, 2008, https://www.clevelandfed.org/publications/cd-reports/2008/sr-20081101-spatial-analysis-of-impact-of-vacant-abandoned-foreclosed-properties.

Ryan, Bill, et al. “Demographics & Lifestyle Analysis.” *Community Economic Development*, economicdevelopment.extension.wisc.edu/articles/demographics-lifestyle-analysis/. Accessed 13 Oct. 2023.