HW1-OLS_Regression

2023-10-19

$$SSE = \sum (y_i - \hat{y_i})^2 = \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki})]^2$$

Introduction

In this analysis we use linear regression to explain variation in median home sales prices by census block group in Philadelphia.

We use a multiple variable linear regression model to explain the correlation between median home value for owner occupied housing units in a census tracts and four predicative variables. The predictive variables are: proportion of residents in a block group with at least a bachelor's degree, proportion of housing units that are vacant, percent of housing units that are detached single family houses, and the number of households with incomes below 100% poverty level (i.e., number of households living in poverty).

The proportion of housing units that are detached single family homes is included as a predictor, because detached single family homes tend to be larger and there is typically a correlation between home size and property value.

We include the number of households with income below 100% poverty levels, because low income households are less likely to be able to afford homes in expensive areas. We include the proportion of residents in a block group with at least a bachelor degree because previous research has shown that there is a correlation between income and educational attainment as explained by Sean Reardon[1]. This correlation results in households without university having less wealth and are thus likely to only be able to purchase homes in areas with lower median home values.

We include the percentage of lots which are vacant as a predictor because previous research has shown correlation between vacanct lots and median home sales prices in Philadelphia. Gravin er all note that there are 40,000 vacant parcels in Philadelphia and most of these are concentrated in low income areas[2].

Methods

Data Cleaning

The data set used in our analysis contains information from the 2000 US Census for Philadelphia, with neighborhood characteristic variables included for 1,720 block groups. Our analysis incorporates the following variables:

- POLY_ID: Census Block Group ID
- MEDHVAL: Median value of all owner occupied housing units
- PCBACHMORE: Proportion of residents in Block Group with at least a bachelor's degree
- PCTVACANT: Proportion of housing units that are vacant
- PCTSINGLES: Percent of housing units that are detached single family houses

[1] Sean Reardon, The Widening Academic Achievement Gap Between the Rich and the Poor: New Evidence and Possible Explanations, Stanford University, https://cepa.stanford.edu/sites/default/files/reardon%20whither%20opportunity%20-%20chapter%205.pdf

[2] Eugenia Garvin et. all, More Than Just An Eyesore: Local Insights And Solutions on Vacant Land And Urban Health, Journal of Urban Health,https://link.springer.com/article/10.1007/s11524-012-9782-7

- NBELPOV100: Number of households with incomes below 100% poverty level (i.e., number of households living in poverty)
- MEDHHINC: Median household income

The original data set had 1,816 block groups and was cleaned using the following methods, which reduced to the total number of observations to 1,720:

- Block groups where population < 40
- Block groups where there are no housing units
- Block groups where the median house value is lower than $10,000
- One North Philadelphia block group which had a very high median house value (over $800,000) and a very low median household income (less than $8,000)

Exploratory Data Analysis

We will examine the summary statistics and distributions of the data set's variables, including the mean and standard deviation of our dependent variable and predictor variables.

As part of our exploratory data analysis, we will examine the Pearson correlations between the predictors. A Pearson correlation, denoted by "r", is a standardized measurement of the strength and direction of the linear relationship between two variables. The correlation between two variables is calculated using the following equation:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - x}{s_x} \right) \left(\frac{y_i - y}{s_x} \right)$$

The Pearson correlation value ranges between -1 to 1, with no units of measurement attached, and the observed variables are interchangeable between the x axis and y axis. A value of -1 represents a perfect negative linear relationship and a value of 1 represents a perfect positive linear relationship - in either case, points on a graph would appear in a straight line with either a negative or positive slope, respectively.

A Pearson correlation value of 0 indicates that there is no linear relationship between two variables. However, a different type of relationship can exist, such as an exponential or quadratic relationship, that the Pearson correlation does not measure.

Multiple Regression Analysis

Regression or Ordinary Least Square (OLS) regression is a statistical method used to examine the relationship between a variable of interest (dependent variable) and one or more explanatory variable (predictors). Regression tests the strength of the relationship, the direction of the relationship (positive. Negative, or 0) and goodness of model fit (how well a model will predict a future set of observations). Regression also allows the ability to calculate the amount by which your dependent variable changes when a predictor variable changes by one unit (holding all other predictors constant). Although, if an explanatory variable is a significant predictor of the dependent variable, it does not imply causation.

When we are dealing with more than one predictor, we run a multiple regression. In Multiple Regression Analysis we have $K > 1$ predictors (independent variable), so rather than getting a line in 2 dimensions from a linear regression, we get a surface in $K + 1$ dimensions ($+1$ accounts for the dependent variable). Here, each independent variable will have its own slope coefficient which indicates the relationship of that particular predictor with the dependent variable, controlling for all other independent variables in the regression.

To examine the relationship between median house values and several neighborhood characteristics we will need to run a multiple regression. Using Philadelphia data at the Census block group level we regressed Natural Log of Median

value of all owner occupied housing units (LNMEDHVAL) on the proportion of housing units that are vacant (PCTVACANT), the 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 with incomes below 100% poverty level (LNNBELPOV100). The equation is stated as:

$$LNMEDHHINC = \beta_0 + \beta_1 PCTVACANT + \beta_2 PCTSINGLES + \beta_3 PCTBACHMOR + \beta_4 LNNBELPOV100 + \varepsilon$$

The Beta coefficient β_i of each predictor may be interpreted as the amount by which the dependent variable changes as the independent variable increases by one unit (holding all other variables constant). The sign indicates whether the relationship between the dependent variable and the independent variables is positive (direct) or negative (inverse). It is important to look at the sign and value of β_i when the coefficient is statistically significant (significantly different from zero). The variable ε commonly referred to as the residual term or random error term in the model. The residual term ε allows the regression line to fall above (ε > 0) or below (ε < 0). The actual data points. ε is the difference between observed values of y and the values of y predicted by the regression model (denoted by yˆ).

When using regression to model the relationships between a dependent variable and predictors certain conditions must be met, these conditions are referred to as assumptions. Prior to making conclusions about the model estimates or using the model for predictions, certain assumptions must be met. These assumptions include Linearity, Independence of Observations, Normality of Residuals, Homoscedasticity, and No Multicollinearity.

Linearity assumes that there is a linear relationship between the dependent variable Y and each of the predictors X. Linearity can be checked by creating scatterplots between y and each of the predictor x. Possible fixes if the Linearity assumption is not met are to transform variables (e.g., log) or run a non-linear (e.g., polynomial) model.

Independence of Observations assumes that there should be no spatial, temporal, or other forms of dependence in the data. This means that each observation from the data must be independent of the others. In order to test for Independence of Observations one can look at the Moran's I of the residual, or the values of y to examine whether regression residuals, or the dependent variable itself, are spatially autocorrelated. If this assumption is not met one must run a spatial regression e.g., spatial lag, spatial error, geographically weighted regression) instead of OLS regression.

Normality of Residuals is violated because either the dependent variable and/or independent variables are themselves non-normal, and/or the linearity assumption is violated. This assumption is not ass important as the 3 previously stated assumptions, especially when dealing with a large sample size. In this context, a large sample size is generally defined as having 30+ observations with 10 additional observations for every additional predictor after the first one. One can test for this assumption by looking at the histogram of residuals to see if they are normal. If this assumption is violated possible fixes include removing outliers and transforming variables (e.g., log).

Homoscedasticity refers to the variance of the residuals ε being constant regardless of the values of each x (or the values of yˆ, i.e., values of y predicted by the model). When this assumption is violated, Heteroscedasticity is present meaning that the residuals ε differs across all values of the independent variable. This assumption can be checked by looking at scatterplots of standardized residuals against each predictor to see if variance of residuals remains the same for different values of each predictor. Violation of this assumption implies Heteroscedasticity which often means that that there is systematic under-or over-prediction happening in the model. Including additional predictors, running a spatial regression, transforming variables, and removing outliers may help reduce Heteroscedasticity.

No Multicollinearity which only applies to multiple regression, occurs when predictor variables are not strongly correlated with each other. Multicollinearity is when two or more predictors are very strongly correlated with each other: r > 0.8 or r < -0.8. If Multicollinearity is present in a model, it will become difficult for the model to estimate the relationship between each independent variable and the dependent variable independently making it difficult to identify significant predictors. One can check for this assumption by reviewing a correlation matrix of the predictors and check if r > 0.8 or r < -0.8. If two or more predictors are strongly correlated, include only one of them in the regression.

Performing multiple regression requires one to estimate the values for a critical set of parameters. These parameters include σ^2 which determines the amount of variability inherent in a regression model, a regression constant β_0 and one regression coefficient β_i for each independent variable in the model. The regression constant or the Intercept β_0 represents the mean value of the dependent variable when all the independent variables are equal to 0. The regression

coefficient, as stated above, is interpreted as the amount by which the dependent variable changes as the independent variable increases by one unit (holding all other variables constant).

In Multiple regression σ^2 is found with the following equation $\sigma^2 = SSE/(n-(k+1)) = MSE$, where k = # predictors and n = # observations. Here, MSE stands for mean squared error.

R^2, often referred to as the Coefficient of Multiple Determination, is the proportion of observed variance in the dependent variable y that is explained by the model by all k predictors. Higher values of R^2 are indicative of a better model. R^2 is calculated as $R^2 = 1 - SSE/SST$ Where SSE is the sum of squared residuals and SST is the total variability in the dependent variable. Adjusted R2 is the R2 adjusted for the number of predictors in the model. Larger values for Adjusted R^2, like R^2, are also indicative of a better model. The equation for Adjusted R^2 is:

$$R^2adj = \frac{(n-l)R^2 - k}{n-(k+1)}$$

The Beta coefficient β_i of each predictor may be interpreted as the amount by which the dependent variable changes as the independent variable increases by one unit (holding all other variables constant). The sign indicates whether the relationship between the dependent variable and the independent variables is positive (direct) or negative (inverse). It is important to look at the sign and value of β_i when the coefficient is statistically significant (significantly different from zero). The variable ε commonly referred to as the residual term or random error term in the model. The residual term ε allows the regression line to fall above ($\varepsilon > 0$) or below ($\varepsilon < 0$). The actual data points. ε is the difference between observed values of y and the values of y predicted by the regression model (denoted by $\widehat{y}$).

Regression Assumptions

Prior to making conclusions about the model estimates or using the model for predictions, certain assumptions must be met. These assumptions include Linearity, Independence of Observations, Normality of Residuals, Homoscedasticity, and No Multicollinearity.

Linearity assumes that there is a linear relationship between the dependent variable y and each of the predictors x. Linearity can be checked by creating scatterplots between y and each of the predictor x. Possible fixes if the Linearity assumption is not met are to transform variables (e.g., log) or run a non-linear (e.g., polynomial) model.

Independence of Observations assumes that there should be no spatial, temporal, or other forms of dependence in the data. This means that each observation from the data must be independent of the others. In order to test for Independence of Observations one can look at the Moran's I of the residual, or the values of y to examine whether regression residuals, or the dependent variable itself, are spatially autocorrelated. If this assumption is not met one must run a spatial regression e.g., spatial lag, spatial error, geographically weighted regression) instead of OLS regression.

Normality of Residuals is violated when either the dependent variable and/or independent variables are themselves non-normal, and/or the linearity assumption is violated. This assumption is not as important as the 3 previously stated assumptions, especially when dealing with a large sample size. In this context, a large sample size is generally defined as having 30+ observations with 10 additional observations for every additional predictor after the first one. One can test for this assumption by looking at the histogram of residuals to see if they are normal. If this assumption is violated possible fixes include removing outliers and transforming variables (e.g., log).

No Homoscedasticity which refers to the variance of the residuals ε being constant regardless of the values of each x (or the values of $\widehat{y}$, i.e., values of y predicted by the model). When this assumption is violated, Heteroscedasticity is present. When Heteroscedasticity occurs the residuals ε differ across all values of the independent variables, meaning that there is systematic under-or over-predictions happening in the model. This assumption can be checked by looking at scatterplots of standardized residuals against each predictor to see if variance of residuals remains the same for different values of each predictor. Including additional predictors, running a spatial regression, transforming variables, and removing outliers may help reduce Heteroscedasticity.

No Multicollinearity which only applies to multiple regression, occurs when predictor variables are not strongly correlated with each other. Multicollinearity is when two or more predictors are very strongly correlated with each other: $r > 0.9$ or $r < -0.9$. If Multicollinearity is present in a model, it will become difficult for the model

to estimate the relationship between each independent variable and the dependent variable independently making it difficult to identify significant predictors. One can check for this assumption by reviewing a correlation matrix of the predictors and check if $r > 0.9$ or $r < -0.9$. If two or more predictors are strongly correlated, include only one of them in the regression.

Multiple Regression Parameters & Estimation

Performing multiple regression requires one to estimate the values for a critical set of parameters. These parameters include σ^2 which determines the amount of variability inherent in a regression model, a regression constant β_0 and one regression coefficient β_i for each independent variable in the model. The regression constant or the Intercept β_0 represents the mean value of the dependent variable when all the independent variables are equal to 0. The regression coefficient, as stated above, is interpreted as the amount by which the dependent variable changes as the independent variable increases by one unit (holding all other variables constant). In multiple regression we estimate these parameters by finding the values β_0 and β_i that minimize the Sum of Squared Errors (SSE) of prediction. Meaning, the differences between a observation's actual score on the dependent value and the score that's predicted for them using the actual scores on the independent variables. SSE will produce the Least Square estimates $\widehat{\beta_0}$ & $\widehat{\beta_k}$. The equation for SSE is:

$$SSE = \sum (y_i - \hat{y_i})^2 = \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki})]^2$$

In Multiple regression σ^2 is found with the following equation $\sigma^2 = \frac{SSE}{n-(k+1)} = MSE$, where k = # predictors and n = # observations. Here, MSE stands for mean squared error.

Coefficient of Multiple Determination R^2

R^2, often referred to as the Coefficient of Multiple Determination, is the proportion of observed variance in the dependent variable y that is explained by the model by all k predictors. Higher values of R2 are indicative of a better model. R^2 is calculated as $R^2 = 1 - \frac{SSE}{SST}$ Where SSE is the sum of squared residuals and SST is the total variability in the dependent variable. Adjusted R^2 is the R^2 adjusted for the number of predictors in the model. Larger values for Adjusted R^2, like R^2, are also indicative of a better model. The equation for Adjusted R^2,

$$R^2_{adj} = \frac{(n-1)R^2 - k}{n-(k+1)}$$

Hypothesis Testing

When stating the hypothesis we first look at F-Ratio which tests the overall significance of the model by testing the Null Hypothesis $H_0 : \beta_i = 0$ that none of the independent variables in the model is a significant predictor of the dependent variable (LNMEDHVAL) against the Alternative Hypothesis $H_a : \beta_i \neq 0$ that at least one of the independent variables is a significant predictor of the dependent variable (LNMEDHVAL). Second, we look at the T-Test which will test the Null Hypothesis against the Alternative Hypothesis for each of the chosen predictors $\beta_1 PCTVACANT$, $\beta_2 PCTSINGLES$, $\beta_3 PCTBACHMOR$, $\beta_4 LNNBELPOV100$.

PCTVACANT: $H_0 : \beta_1 = 0$: Implies that the variable PCTVACANT is not a significant predictor of the dependent variable (LNMEDHVAL). $H_a : \beta_1 \neq 0$: Implies that the variable PCTVACANT is a significant predictor of the dependent variable (LNMEDHVAL).

PCTSINGLES: $H_0 : \beta_2 = 0$: Implies that the variable PCTSINGLES is not a significant predictor of the dependent variable (LNMEDHVAL). $H_a : \beta_2 \neq 0$: Implies that the variable PCTSINGLES is a significant predictor of the dependent variable (LNMEDHVAL).

PCTBACHMOR: $H_0 : \beta_3 = 0$: Implies that the variable PCTBACHMOR is not a significant predictor of the dependent variable (LNMEDHVAL). $H_a : \beta_3 \neq 0$: Implies that the variable PCTBACHMOR is a significant predictor of the dependent variable (LNMEDHVAL).

LNNBELPOV100: $H_0 : \beta_4 = 0$: Implies that the variable LNNBELPOV100 is not a significant predictor of the dependent variable (LNMEDHVAL). $H_a : \beta_4 \neq 0$: Implies that the variable LNNBELPOV100 is a significant predictor of the dependent variable (LNMEDHVAL).

Stepwise Regression

Stepwise regression is a data mining method which selects predictors based on the following criteria. 1) P-values are below a certain threshold (variables where the P-value < 0.1) and 2) the smallest value of the Akaike Information Criterion (AIC), which measures the relative quality of statistical models. There are a number of limitations when using Stepwise regression which include: 1) the final model is not guaranteed to be optimal in any specified sense. 2) The stepwise procedure produces a single final model, although there are often several equally good models. 3) it does not consider a researcher's knowledge of the predictors. 4) Although the order in which the variables are removed or added can provide valuable information, it's important not to over interpret the order. 5) one should not conclude that all the important variables for predicting y have been identified or that all unimportant variables have been eliminated.

K-Fold Cross-Validation

The k-Fold Cross-Validation approach involves randomly dividing the set of observations into k groups (or folds), of approximately equal size . In practice it is typical to use $k = 5$ or $k = 10$. The first fold is treated as a validation set and the model is fitted on the remaining $k-1$ folds (training data set). The mean squared error (MSE) is then computed on the observations in the held-out fold. The procedure is repeated k times, each time, a different fold is treated as the validation set. This process results in k estimates of MSE, The k-fold MSE estimate is computed by averaging the MSEs across the k folds. The k-fold Root Mean Squared Error (RMSE) is computed as the square root of the MSE $RMSE = \sqrt{\frac{\sum_{i=1}^{n} \epsilon_i^2}{n}}$. RMSE values can be compared across different models, the model with the smallest RMSE will be the best model.

Additional Analyses

Software

This report used the open source software R to conduct statistical analyses.

Results

Exploratory Results

Summary Statistics

This table below shows the mean and standard deviation for our independent variable (Median house value) and our four dependent variables. The average median home value for census block groups in Philadelphia is 66,287.73 USD, and the standard deviation is 60,006 USD. The large standard deviation indicates a large amount of variability in average home sale prices across the different census block groups in Philadelphia.

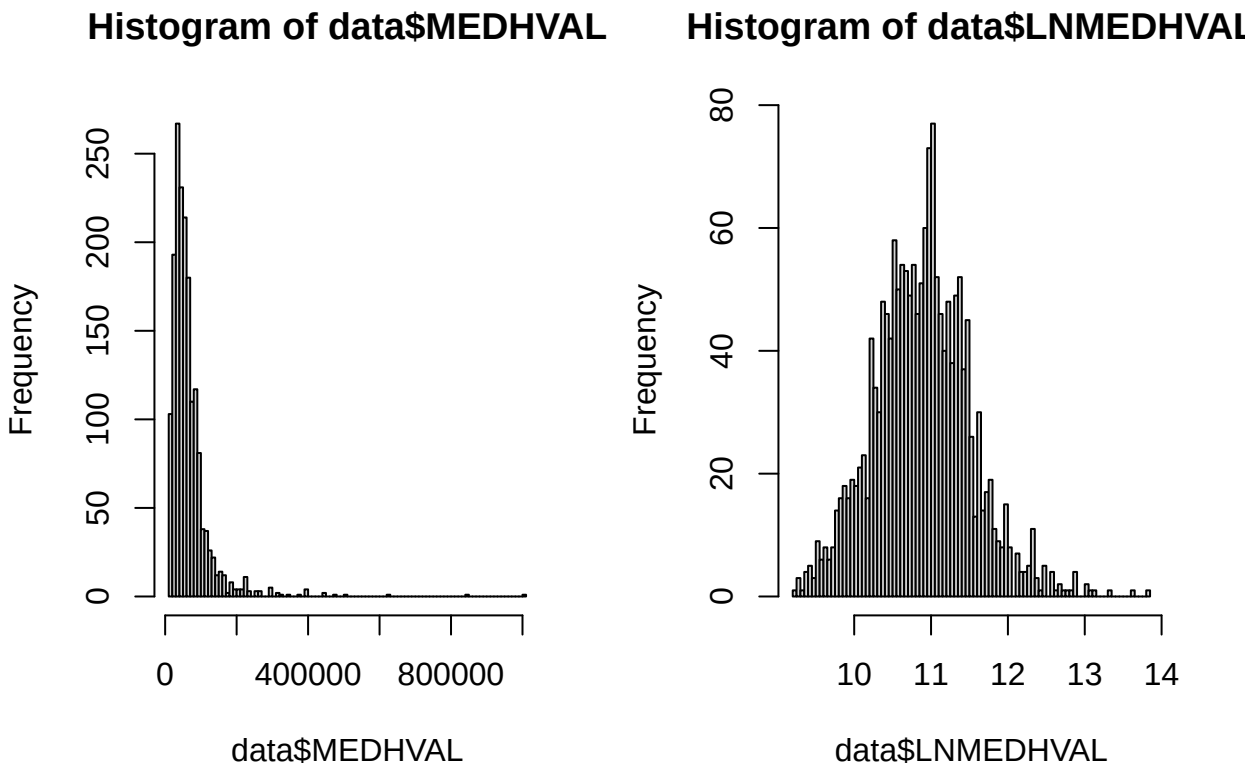
Variable	Mean	Standard Deviation
Median Houme Value of all occupied housing units	66287.733139	60006.075990
% of Individuals with Bachelor Degrees or Higher	16.081372	17.769558
# Households Living in Poverty	189.770930	164.318480
% of Vacant Houses	11.288529	9.628472
% of Single House Units	9.226473	13.249250

quick code that checks which variables have 0 values for logarithmic transformation

Histograms

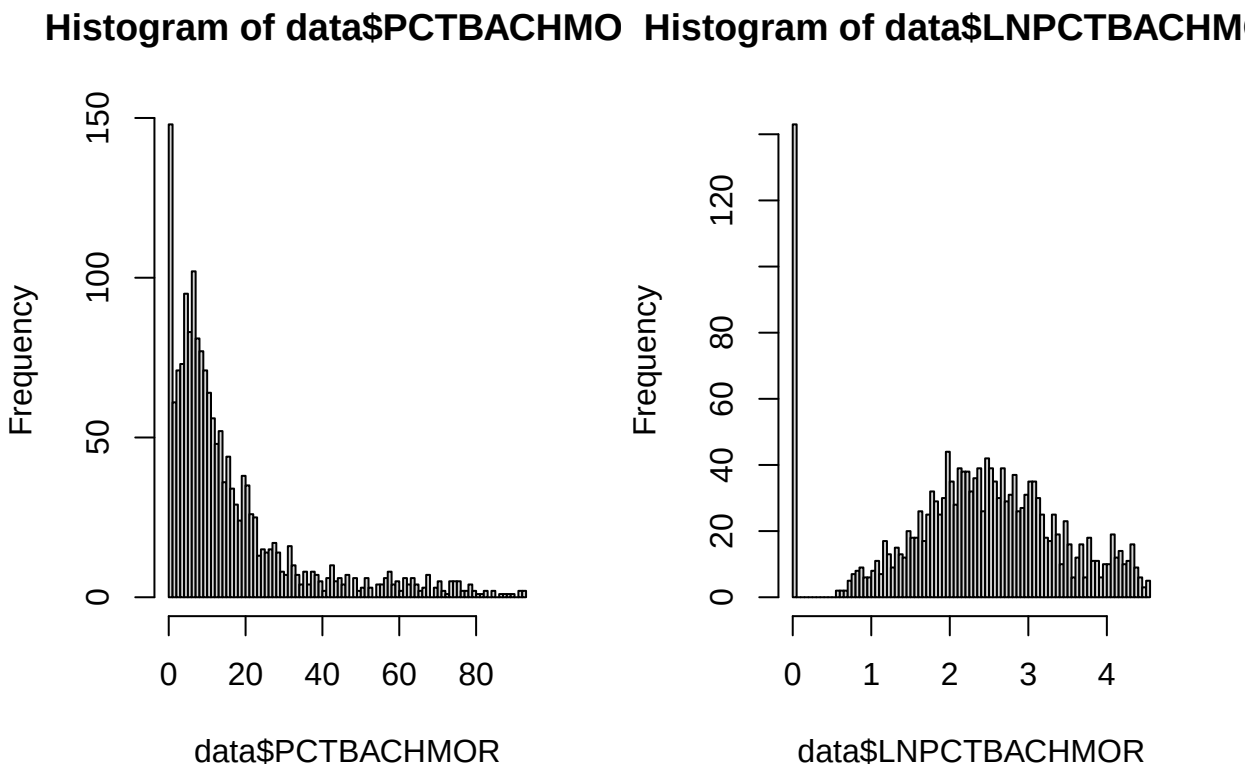
Median Home Value of owner occupied housing units The two histograms below show the distribution of our dependent variable, median home values of owner occupied housing units by census tract before and after applying a natural log transformation. The histogram without the natural log transformation peaks around 75,000 USD and is right skewed - it does not have a normal distribution. After applying a natural log transformation, the mean home value has a near normal distribution with a peak around 11.

Our analysis will use the natural log transformation of median home sales value as the independent variable because it is best practice to use a variable with a normal distribution when conducting a linear regression analysis. Using a normally distributed variable can help mitigate issues resulting from a non-linear relationship between our independent and dependent variable. The natural log transformation is most widely used with positive-skewed (i.e: right skewed) data. As mentioned above, our dependent variable has a right skewed distribution.



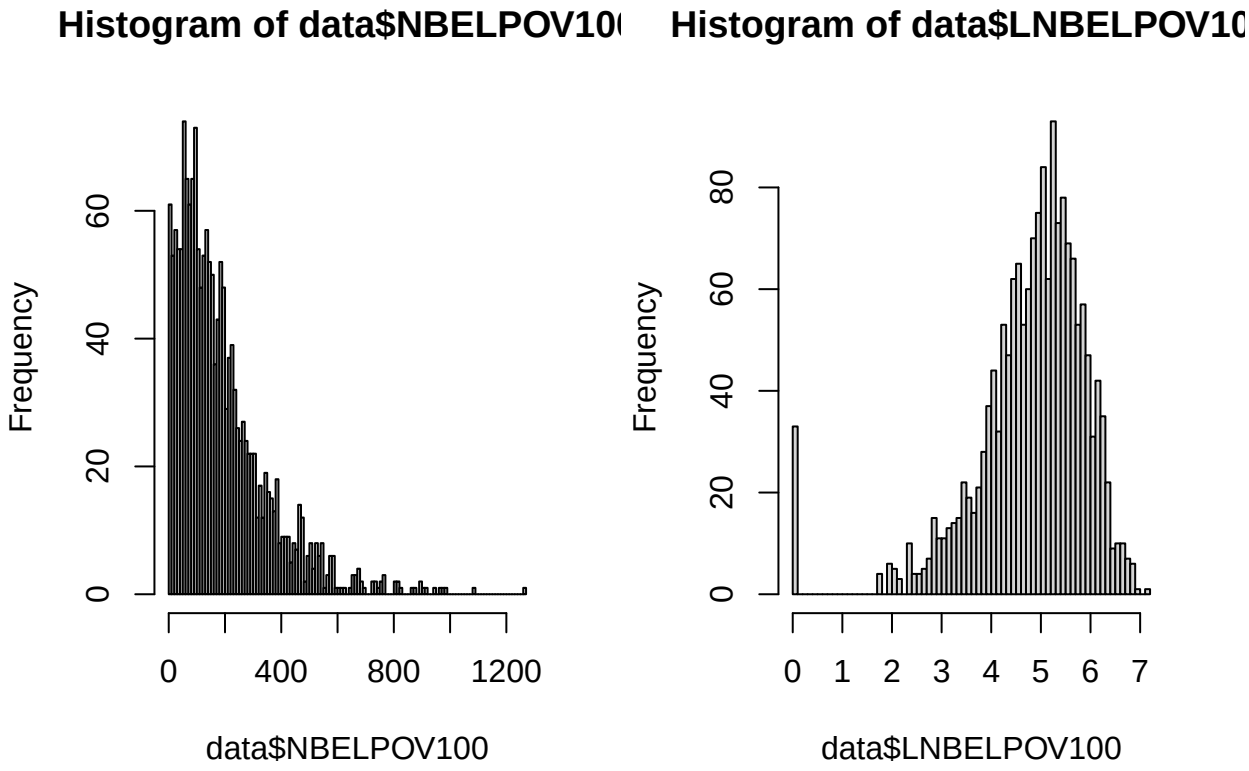
Percent of Population with a Bachelor Degree: These histograms show the distribution of the percent of the population with a bachelor's degree by census tract before and after applying a natural log transformation. The histogram without the natural log transformation is right skewed, and there are 143 census blocks where 0% of the population has a bachelor degree.

After applying a natural log transformation, the 143 census blocks which had a value of 0% continue to have a value of 0. The distribution with the natural log transformation applied is not normal and has a zero infrared distribution. Because both the variable with and without the natural log transformation applied are both not normal we will use the variable without the natural log transformation (PCTBACHMORE) for the regression.



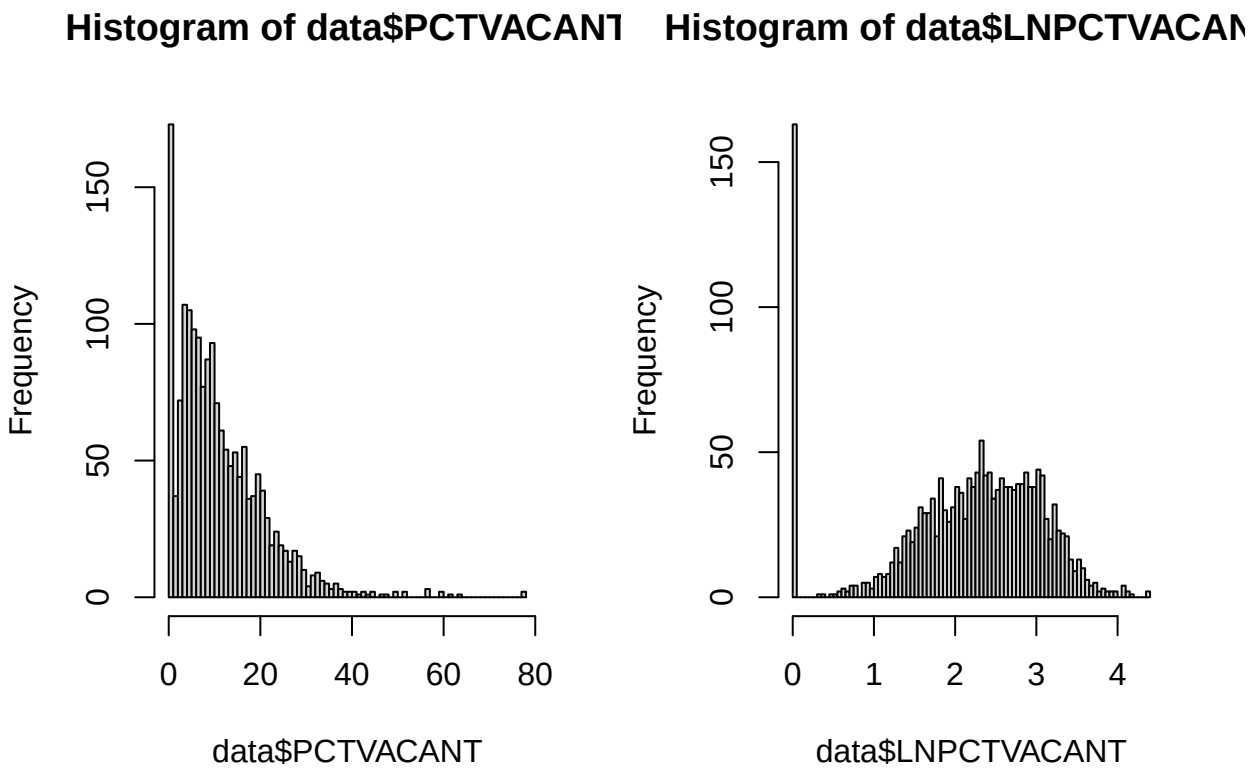
Population Below the Poverty Line The histograms below show the distribution of the population living below the poverty line in each census block with and without the natural log transformation applied. The histogram without the natural log transformation, is again right skewed and peaks around 100 households. There is a very long tail to the right, and multiple outliers are present. Notably, the maximum value is 1,267 households below the poverty line in one census tract - which is more than six times larger than the mean value.

After applying a natural log transformation, the variable displays a distribution which is closer to a normal distribution. There is a clear peak around 5.5, but the data is slightly skewed to the left and is zero inflated, but is closer to a normal distribution than the non natural log transformed variable. Because the natural log transformed variable is closer to a normal distribution we use the natural log transformation of the Population Below the Poverty Line (LNBELPOV100) in our regression. Additionally, the large positive skew in the non natural log transformed data supports the usage of the natural log transformed variable.



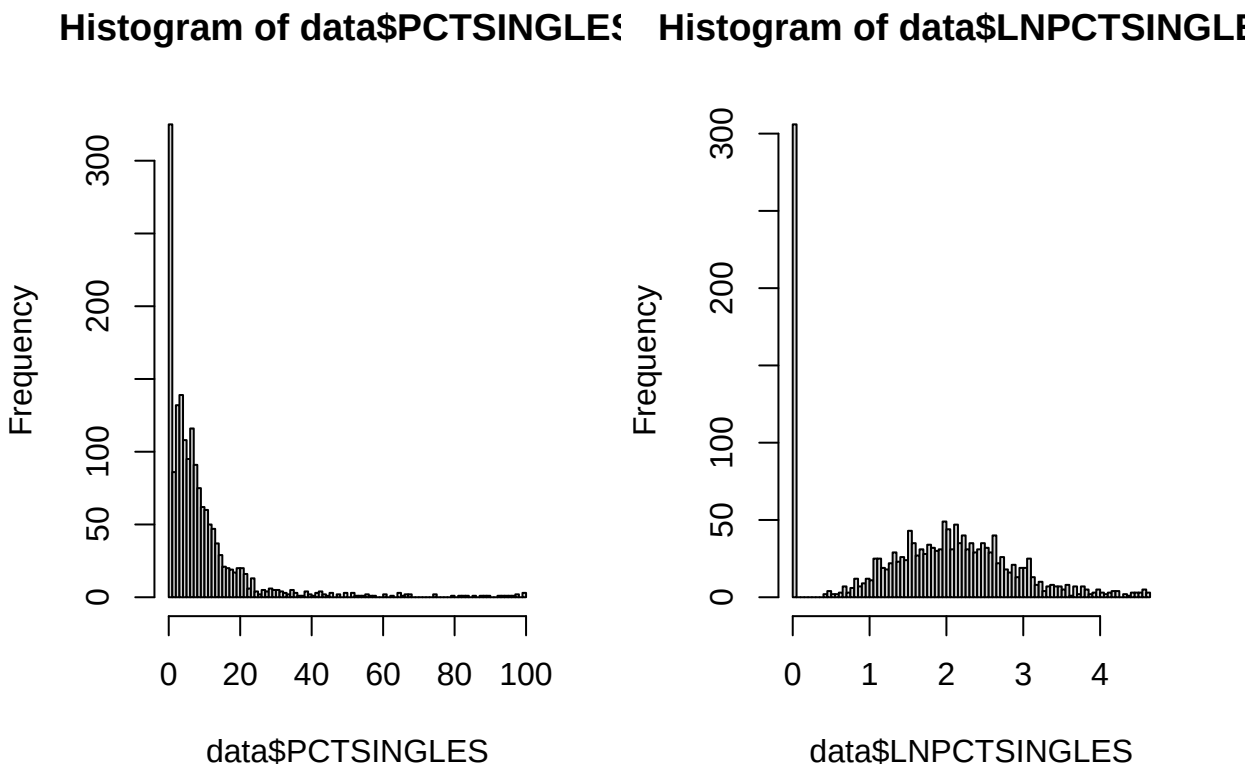
Percent of Households units which are vacant The histograms below show the distribution for the percent of housing units in a census tract which are vacant with and without the natural log transformation. The histogram without the natural log transformation is right skewed and has a long tail, with multiple outliers present. Additionally, there are 163 census block groups where 0% of the housing units are vacant.

After applying the natural log transformation, the 163 census block groups which have a value of 0% still have a value of 0, the presence of the large number of census block groups with a value of zero prevents the distribution from being considered normal and results in a zero-inflated distribution. Because neither distribution is normal we use the variable without the natural log transformation in our regression analysis (PCTVACANT).



Percent of housing units that are detached single family houses The histograms below show the distribution for the percent of housing units that are detached single family homes by census block group with and without the natural log transformation. The histogram without the natural log transformation is very right skewed, and the vast majority of census block groups (i.e: 1,548) have a percentage less than 20%, and there are 306 block groups where the percentage is 0%. There are 172 census block groups which have percentages above 20% including three homes with values of 100%. The extreme outliers are likely are a result of the inclusion of sub-urban census block groups in Northeast and Northwest where most homes are detached single family homes.

After applying the log transformation the 306 block groups where the percentage of detached single family homes is 0%, continue to have a value of 0 resulting in a zero inflated distribution. Because both the natural log transformed and non natural log transformed variable do not have a normal distribution we use the non natural log transformed variable in our regression analysis.



Maps

This section includes coloropleth maps of our dependent and four independent variables.

Map of Dependent Variable This map shows our dependent variable, which is the median house value of owner occupied units by census tract with a natural log transformation. We observe that the census tracts with the highest median home values are primarily clustered in center city and northwest Philadelphia. The census tracts with the lowest median home values are located North of center city and in areas of West Philadelphia located west of University City.

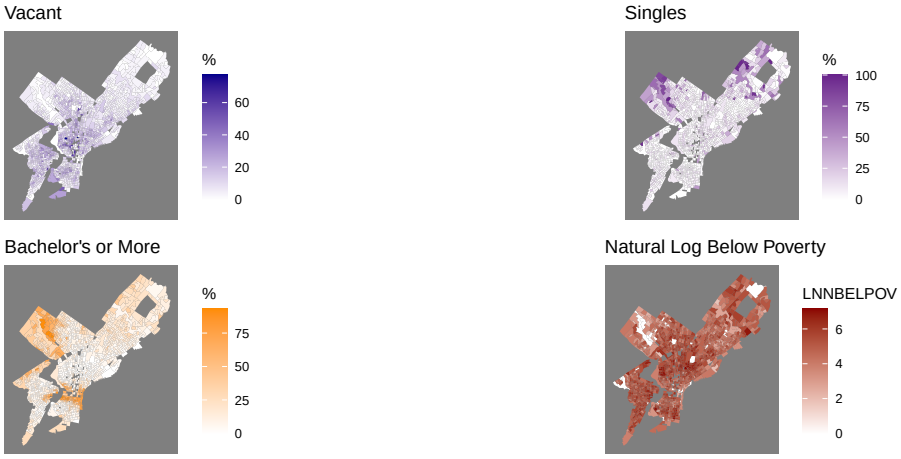
Natural Log Median Home Value

Maps of Independant Variables The maps below show the spatial patterns on our four independent variables: PCTBACHMOR, PCTVACANT, PCTSINGLES, and LNBELPOV100.

Based on a review of the maps, the PCTBACHMOR variable appears to be the independent variable with the strongest correlation with our dependent variable. Like our dependent variable, the percent of residents living in the census tract with a bachelor degree is highest in center city and in Northwest Philadelphia. The areas with the lowest percentage of residents with a bachelor degree are located in West Philadelphia west of University City and north of Center City - these are the same areas where the natural log transformed median home values are lowest.

Conversely, the PCTSINGLES variable appears to be less correlated with our dependent variable. This is because the percent of housing units that are detached single family homes tends to be low in center city, and high in Northwest Philadelphia which are both neighborhoods with high median home prices values.

The independent variables PCTBACHMOR and PCTVACANT appear to have a strong negative correlation, as areas with a high PCTVACANT rate also have a low PCTBACHMOR rate. Conversely, areas with a high PCTBACHMOR rate have a low PCTVACANT rate. This negative correlation indicates that there may be multicollinearity between these two variables. We will check the strength of this correlation using Pearsons correlation to determine if this multicollinearity could be an issue in our regression.

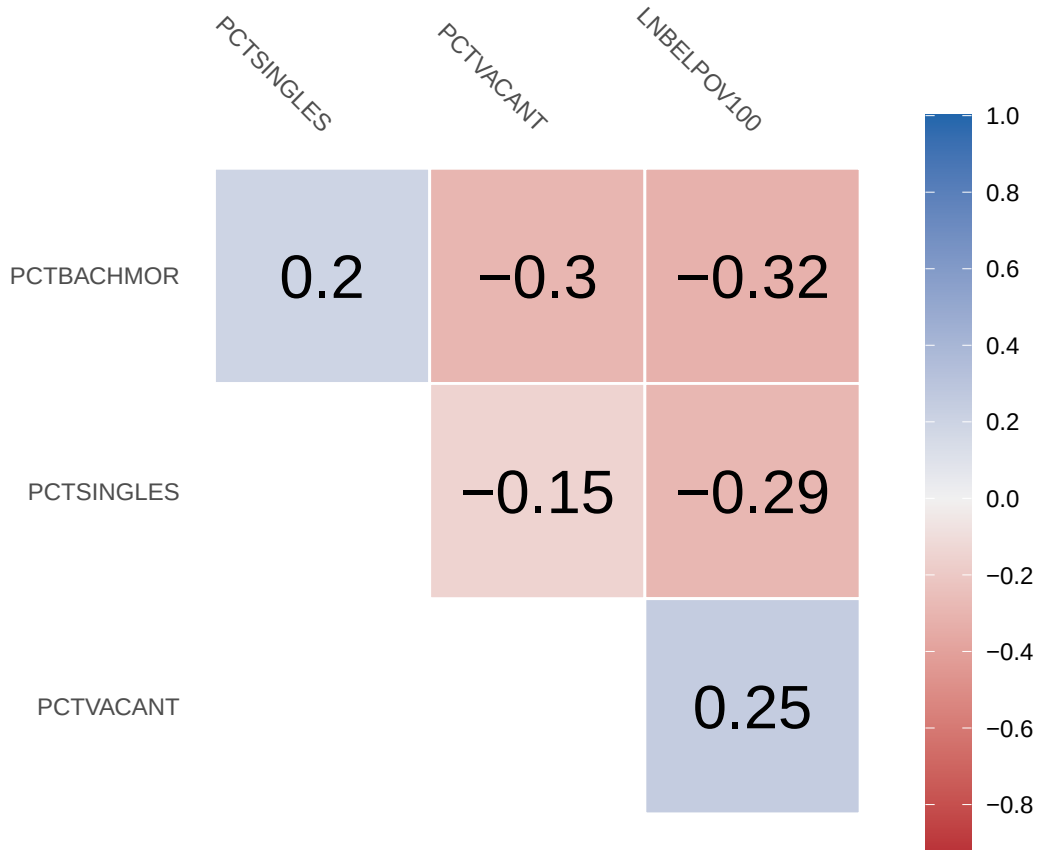


Pearson correlations

The correlation matrix shows the Pearson correlation between each our dependent variables. For example, PCTBACHMOR is negatively correlated with LNBELPOV100 and PCTVACANT is positively correlated with LNBELPOV100.

Despite the correlations, we can conclude that there is not severe multicollinearity. This is because the Pearson correlation values are all between 0.8 and -0.8. When Pearson correlation values are within this range we generally do not need to be concerned about multicollinearity.

The Pearson correlation value for the relationship between PCTBACHMOR and PCTVACANT is -0.3 supporting our previous conclusion that there is a negative correlation between the variables. However, because the Pearson correlation is within the range -0.8 to 0.8 we do not need to be concerned about severe multicollinearity.



Regression Analysis

When we review our regression analysis we start by examining the f-ratio. The f-ratio is 840.9, and the p-value associated with the f-ratio is less than 0.0001. Thus, we can reject the null hypothesis that all β coefficients are zero.

After reviewing the f-ratio, we proceed to reviewing the β coefficients, standard errors, t statistics, and p values for our four dependent variables. All four dependent variables have a p-score which indicates that the β coefficients are statistically significant (<0.05) and we can reject the null hypotheses that any of our β coefficients are equal to 0.

Because we rejected the null hypotheses, we can conclude that:

- There is a statistically significant negative relationship between the natural log of the median home value and proportion of homes which are vacant.
- There is a statistically significant negative relationship between the natural log of the median home value and the natural log of the number of households living in poverty.
- There is a statistically significant positive relationship between the natural log of the median home value and the proportion of homes which are single family homes.
- There is a statistically significant positive relationship between the natural log of the median home value and the percent of individuals with a bachelors degree.

When reviewing our beta coefficients it is important to remember that our dependent variable is natural log transformed. For the variables proportion of homes which are standalone single family homes, proportion of homes which are vacant, and percent of individuals with a bachelor degree only the dependent variable is natural log transformed. The Beta coefficients are also less than 0.3. Thus, we can conclude that as our independent variables go up by 1 unit, the expected change in median home value is approximately $100\beta_1$. Thus, as the proportion of homes which are vacant goes up by

1% the median home value will decrease by approximately 1.916%. As the proportion of homes which are standalone single family homes goes up by 1% the median home value will increase by approximately 0.298%. As the percent of individuals who have a bachelors degree goes up by 1% the median home value will increase by approximately 2.091%.

For the population below the poverty line both our independent variable and our dependent variable are both natural log transformed. Thus, as the population below the poverty line increases by 1 the median home value will change by approximately $(1.01_1^{\beta} - 1) \bullet 100$, i.e: the median home value will decrease by approximately 0.07848%.

The R-squared value is 0.6623, indicating that 66.23% of the variance in our dependent variable is explained by our four independent variables. 33.77% of the variance is not explained by our four independent variables. Our adjusted R-squared value is 0.6615, indicating that 66.15% of the variance in our dependent variable is explained by our four independent variables after adjusting the r-squared to account for the model including more than one independent variable.

```
##
## Call:
## lm(formula = LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR +
##     LNBELPOV100, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.25825 -0.20391  0.03822  0.21744  2.24347
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 11.1137661  0.0465330 238.836 < 0.0000000000000002 ***
## PCTVACANT   -0.0191569  0.0009779 -19.590 < 0.0000000000000002 ***
## PCTSINGLES   0.0029769  0.0007032   4.234            0.0000242 ***
## PCTBACHMOR   0.0209098  0.0005432  38.494 < 0.0000000000000002 ***
## LNBELPOV100 -0.0789054  0.0084569  -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
```

The table below shows an analysis of the variance table for our linear regression model. The Sum of Square Errors (SSE) for our model is 230.44. The Regression Sum of Squares (SSR) is equal to 451.745, and the Total Sum of Squares (SST) is equal to 672.185. We can calculate the R^2 for our model by dividing the SSR by the SST, i.e: 451.745 / 682.089 which equals 0.6623.

```
## Analysis of Variance Table
##
## Response: LNMEDHVAL
##               Df  Sum Sq Mean Sq  F value                Pr(>F)
## PCTVACANT      1 180.392 180.392 1343.087 < 0.00000000000000022 ***
## PCTSINGLES     1  24.543  24.543  182.734 < 0.00000000000000022 ***
## PCTBACHMOR     1 235.118 235.118 1750.551 < 0.00000000000000022 ***
## LNBELPOV100    1  11.692  11.692   87.054 < 0.00000000000000022 ***
## Residuals   1715 230.344   0.134
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression Assumptions Checks

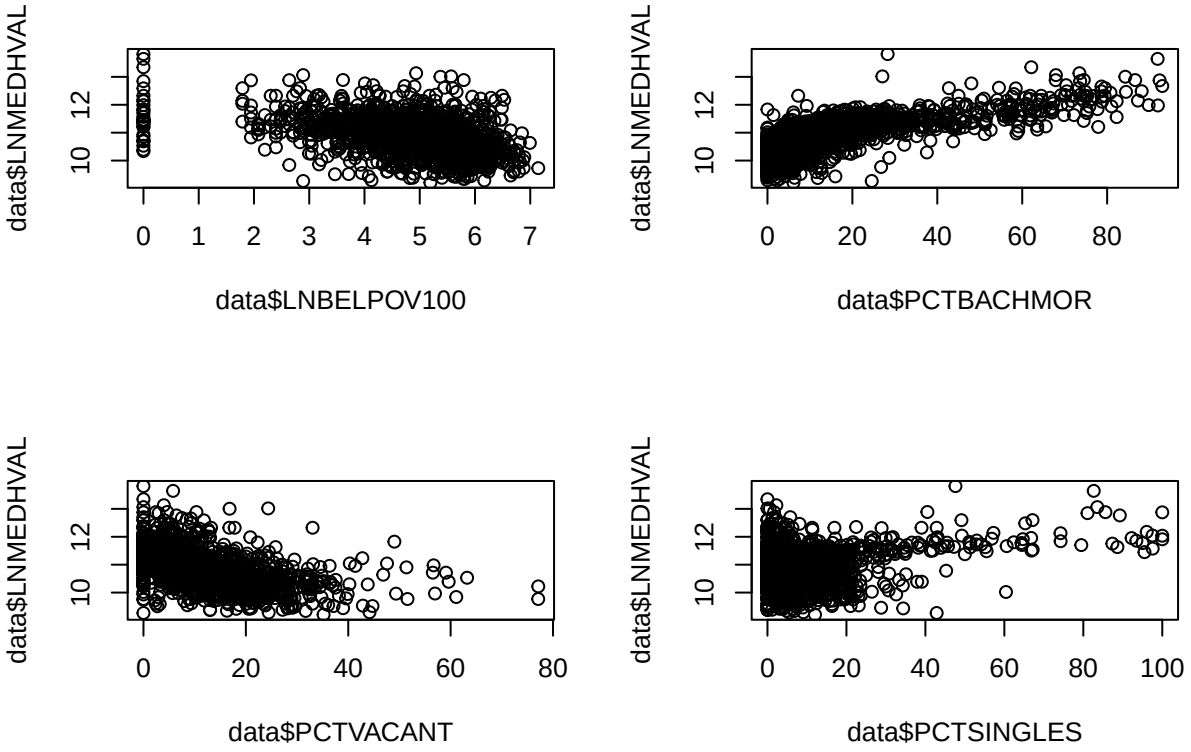
In this section, we will discuss testing model assumptions. We have already examined the variable distributions in a prior section.

Scatter Plots - Linear Relationships Between Variables

When running linear regressions, a core assumption is that there is a linear relationship between the dependent variable and each of the predictor variables. To check this assumption, we plot the dependent variable with each of the predictor variables in a scatter plot.

In cases where this assumption is not met, log transformations are often used. Based on the results of our variable distributions, we have already conducted log transformations for the dependent variable for median home value, now LNMEDHVAL, and for the predictor variable for number of households below poverty, now LNNBELPOV100.

The following scatter plots show the relationship between the dependent variable, LNMEDHVAL, and each the predictor variables LNNBELPOV100, PCTBACHMOR, PCTVACANT, and PCTSINGLES. There does not appear to be a clear linear relationship between LNMEDHVAL and LNNBELPOV100 or PCTSINGLES. The other two variables appear to show a more linear relationship, but PCTVACANT does not appear to show a strong linear relationship.

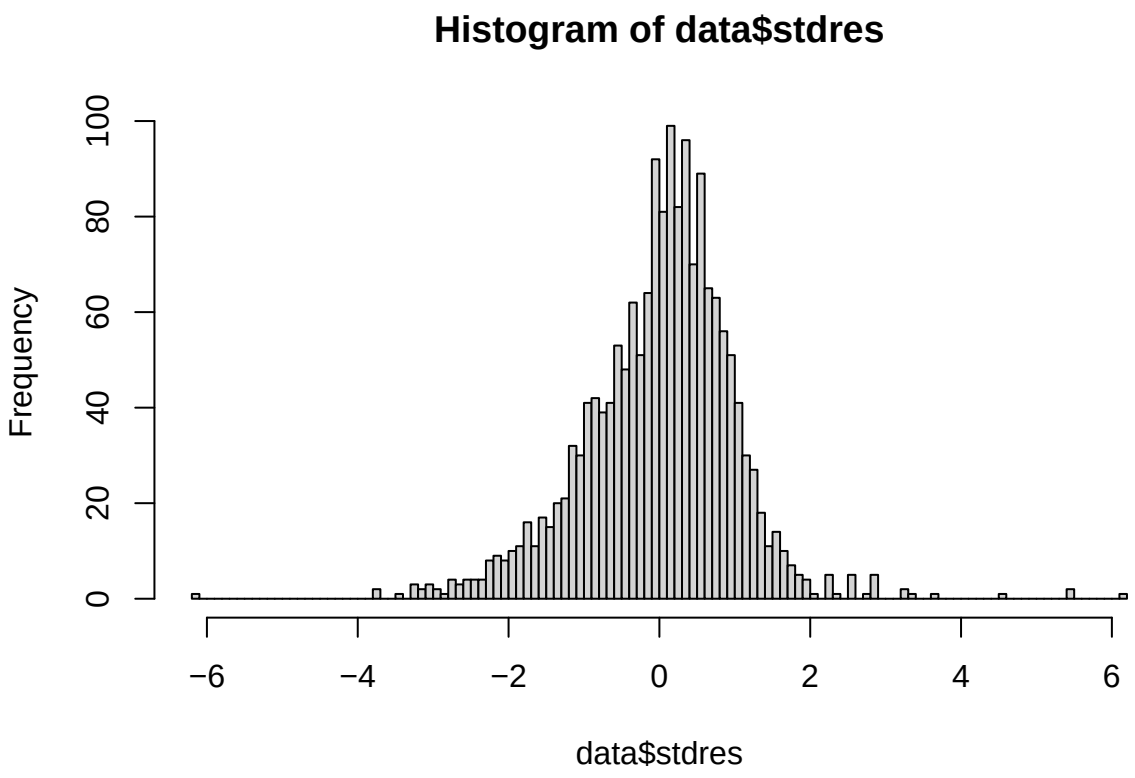


Histogram of the standardized residuals

Another assumption when running linear regression is that regression residuals are distributed normally. However, this assumption of normality is not considered critical in a regression, especially for data sets with a large number of observations.

In order to compare residuals for different observations, we standardize the residuals through dividing a residual by its standard error. Standardizing allows us to observe how many standard deviations a residual is from our model's estimate

The following histogram of standardized residuals shows that residuals appear normally distributed.



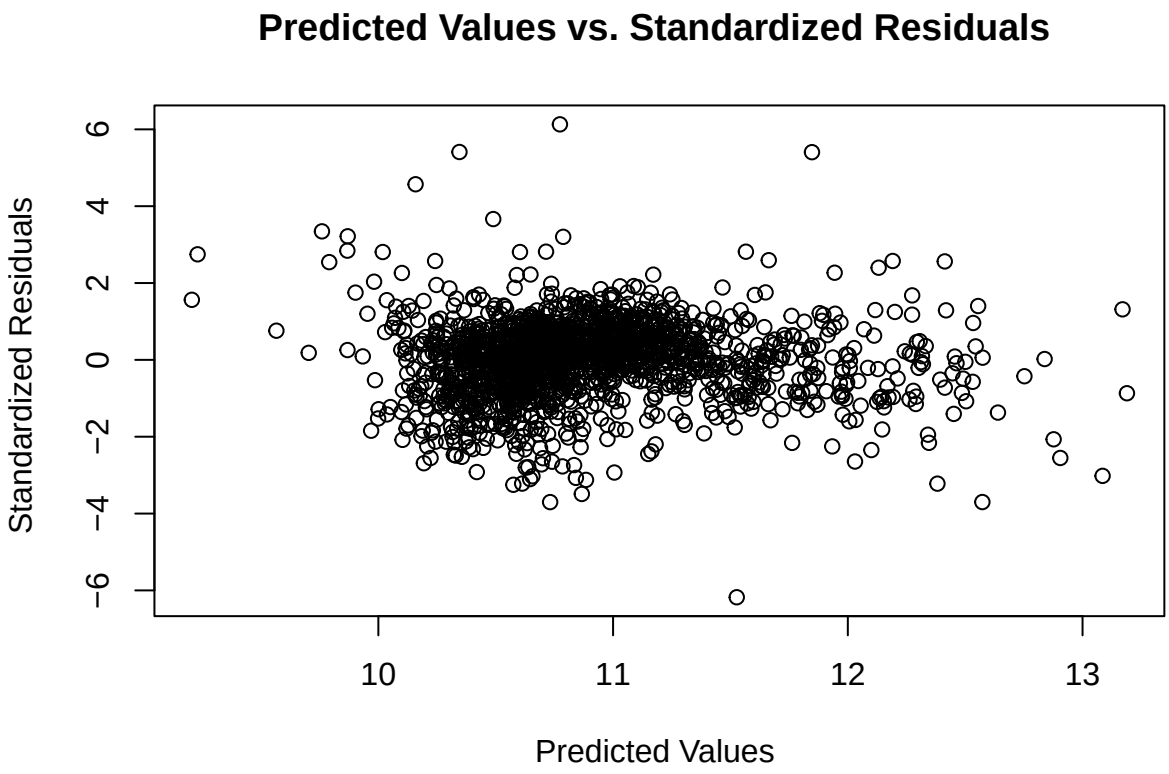
Scatter Plot - Standardized Residual by Predicted Value

An additional core assumption of linear regression is that there is constant variance in residuals compared to the predicted values of the model - this relationship is referred to as homoscedastic. If non-constant variance is observed, the relationship is heteroscedastic.

Given that there multiple predictors, we can plot the standardized residuals of the model by our predicted values of LNMEDHVAL. The scatter plot of standardized residuals appears to show a slight heteroscedastic relationship, based on a small "bow-tie" shape present around the predicted value of 11.5.

Outliers also appear to be present based on our scatter plot - there are several positive standardized residuals above 4 standard deviations above 0 and at least one standardized residual beyond -6 standard deviation below 0.

The standardized residuals also appear to be heavily clustered around between the predicted values of about 10.5 to 11.5.



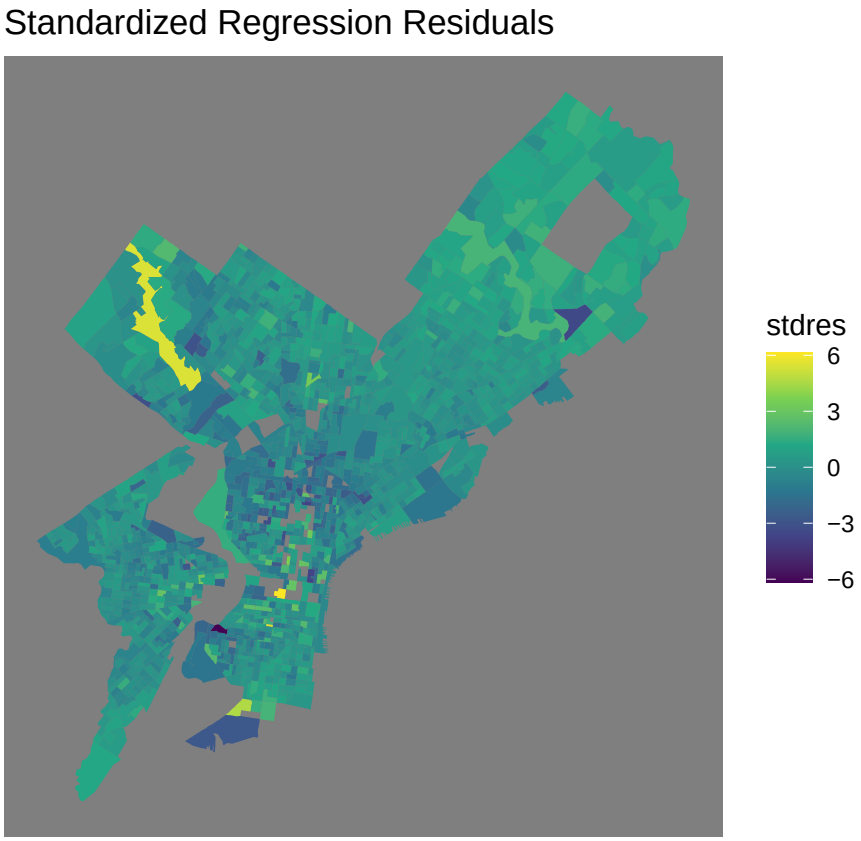
Spatial Autocorrelation of Variables

A critical assumption when building a regression model is that each observation of a variable is independent from other observations of that variable which we examine in our variables as spatial dependence, otherwise known as spatial autocorrelation. Based on the maps of the dependent variable LNMEDHVAL and the predictor variables, we can estimate whether observations of each variable appear to show spatial autocorrelation, which would appear as block groups clustering with similar values:

- The log of median home values appears to show spatial autocorrelation. High values appear to cluster in the downtown area and northwest Philadelphia, while low values appear to cluster around North Philadelphia and southwest Philadelphia, among other areas.
- Vacancy percentages appear to cluster in similar areas as the low value clusters of the log of median home values - North Philadelphia and southwest Philadelphia.
- Singles percentages appear to cluster in the northwest and northeast regions of the city.
- Bachelor's degree or more percentages appear to cluster in similar areas to the high values of the log of median home values - the downtown area and northwest Philadelphia.
- The log of the number of households in poverty does not appear to show spatial autocorrelation.

Based on these maps, each of the variables in our regression model appear to show spatial autocorrelation.

Choropleth map of the standardized regression residuals



Additional Models

Stepwise Regression

Inputting the variables of our model into a stepwise regression test results in a final model that retains all four of the original predictor variables. Therefore, all four predictors had sufficiently low p-values and removing predictors did not lower the value of the AIC.

```
## Start:  AIC=-3448.07
## LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNBELPOV100
##
##               Df Sum of Sq    RSS     AIC
## <none>                     230.34 -3448.1
## - PCTSINGLES   1     2.407 232.75 -3432.2
## - LNBELPOV100  1    11.692 242.04 -3364.9
## - PCTVACANT    1    51.546 281.89 -3102.7
## - PCTBACHMOR   1   199.020 429.36 -2379.0

## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
```

```
## LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNBELPOV100
##
## Final Model:
## LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNBELPOV100
##
##
##   Step Df Deviance Resid. Df Resid. Dev       AIC
## 1                       1715   230.3435 -3448.073
```

Cross-Validation

The following table provides the results of k-fold cross validation, when the number of folds (k) is set to 5. The root mean square error of the original model is represented by rmse1, which we compare to a second regression model's root mean square error, rmse2. This second model only includes the percent of vacancy and median household income as predictors for median home value. Given that rmse1 is lower than rmse2, our original model is considered more generalizable when using different folds of the data and is thus a better fit for the model.

Discussion and Limitations

Based on our model we can conclude that there is a statistically significant relationship between our dependent variable, the natural log of the median home value and our four dependent variables. We can conclude that areas of Philadelphia where a larger number of residents live in poverty, residents do not have higher education degrees, and a large number of vacant homes are present are likely to also have a low median home value.

This conclusion was not surprising, as poorer households are likely to only be able to afford homes in areas with lower median home values. Additionally, redlining during the mid 1900s in Philadelphia identified undesirable neighborhoods. Redlining has a lasting impact on the Philadelphia housing landscape. Neighborhoods which were identified as undesirable continue to have higher vacancy rates and house low income households.

Our model does a good job explaining the variation in median home value across the city of Philadelphia. The results of the F-test indicate that all our dependent variables are statistically significant predictors of median home value. Our R^2 value of 0.6623 indicates there is a strong linear correlation between our independent variables and the natural log of the median home sales value. We consider a R^2 value above 0.5 to be a indicative of a strong correlation, the 0.5 threshold core a strong correlation is acceptable in the social sciences domain.

Our stepwise regression model included all four of the original predictor variables. Therefore, all four variables are considered statistically significant predictors of our dependent variable, the natural log of median home values, and our model performs best with all four predictors included as opposed to some subset of the independent variables.

```
Several core assumptions of OLS regressions are violated by our model. Our predictor variables do not al
```

The predictor variable NBELPOV100 is the only independent variable that is provided as a raw number as opposed to a percentage that normalizes the values to compare across block groups. Block groups with smaller counts of population would experience an outsized impact in the regression model by increasing the count by one household, whereas block groups with larger populations would not experience the same level of impact. Normalizing to percentages allows us to compare 1 percent increases, representing proportional impacts to different block groups.

Our cross-validation results show that the root mean square error for the four predictor model was lower than that of the two predictor model, identifying the former model as better and more generalizable for new data.

Ridge Regression is a method that offers solutions to issues that can arise in OLS Regression such as: allowing for a large number of predictors relative to the number of observations, allows for multicollinearity, and deals with overfitting by shrinking the coefficients of variables towards 0 (which can significantly reduce their variance and the RMSE in the validation set). Ridge regression functions by minimizing the SSE subject to a found (i.e., constraint) on the on

the quantity called L2 norm (the square root of the sum of the squared β coefficients) where in OLS we minimize the SSE. Lasso Regression (least absolute shrinkage & selection operator), is like ridge regression except that it will set the values of some coefficients to exactly 0 for different values of λ, minimizing SSE subject to a bound on the quantity called L1 norm (the sum of absolute values of the β coefficients). Both ridge and lasso regression would not be appropriate here because they drop the assumption of no multicollinearity. When this assumption is violated, we can get incorrect estimates of β_i as well as incorrect estimates of p-values of significance.