Kyle McCarthy
Lauren Nolan
Dustin Renniger

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

INTRODUCTION

Philadelphia is a diverse city earning the title of the "city of neighborhoods." Although neighborhood boundaries are often subjective, there are over fifty neighborhoods within the city limits. Each neighborhood has its distinct characteristics, consisting of significant differences in demographics, infrastructure, and land use. The spatial variability across the city creates a challenge for spatial modeling. Our goal is to evaluate median house values and their relationship to numerous census variables for this project. These variables include the proportion of residents with a bachelor's degree, the proportion of vacant units, percent of housing units that are detached single-family houses, number of households living in poverty, and the median household income. These variables are then used to predict median house values in Philadelphia, and statistical analysis will be performed to determine the model's effectiveness.

Each of the variables chosen is expected to have a relationship with house prices. For example, housing units that are detached from single-family houses are expected to be of higher value. Additionally, residents with a bachelor's degree are likely to have more money to spend on a home. Much of the data used in this study has been previously utilized to create numerous home prediction models. Kenstens et al. (2006) developed a similar model, focusing on housing-type, age, educational attainment, and educational attainment. Their findings demonstrated a significant effect of income on the spatial distribution of rent prices, and a positive relationship between educational attainment and rent prices. Zhou et al. (2010) concluded similar results in Atlanta, indicating that housing prices are sensitive to regions with lower mean incomes. Vacant properties were used in this study to estimate the value of a location. Poverty rates are implied to correlate to median income directly; the higher the poverty rate, the lower the median income. It is important to note that the variables used in this model are not the only variables that can be used to predict housing prices. Distance from infrastructure, house square footage, age of the house, distance to transportation, and racial demographics have been shown to have relationships with housing values and have been used in other prediction studies. However, these variables will not be included in order to better evaluate the performance of the chosen variables.

METHODS

Data Cleaning

The census data used in this analysis originally had 1816 observations. The data was cleaned to achieve a dataset with 1720 observations. The data was cleaned by removing the following block groups:

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

These blocks were removed because they can be considered outliers and would significantly impact the statistical results.

Exploratory Data Analysis

We began with an exploratory data analysis in order to prepare the data for regression analysis. Specifically, we examined the summary statistics and distribution of each variable. We also examined the relationship--or correlation-- between variables. The correlation, notated by the variable *r,* is a measure of the strength of the relationship between variables. The value of *r* ranges from -1 to 1. The value of *r* does not depend on which variable is labeled x and y. The value is also independent of the units of measurement of x and y. When *r* = 1, all (x_i,y_i) pairs lie on a straight line with a positive slope. A positive slope is when the x variable increases as the y variable also increases. When *r* = -1, all of (x_i,y_i) pairs lie on a straight line with a negative slope. A negative slope is when the x variable increases, as the y variable decreases. When *r* = 0, there is no relationship between the two variables. An R^2 value is representative of a strong relationship that varies by field and study. In social sciences, an R^2 value above 0.6 is often considered to be a strong relationship. The formula for R^2 is presented in the “Multiple Regression Analysis” section below.

Multiple Regression Analysis

Our analysis uses multiple regression to examine the relationship between variables and create a model to predict median house value. A regression is a statistical relationship between a variable of interest, known as the dependent variable, and one or more explanatory variables. It allows for the calculation of the amount by which the dependent variable changes when a predictor variable changes by one unit. Just like correlation, if an explanatory variable is a significant predictor of the dependent variable, it does not imply that the explanatory variable is a cause for the dependent variable.For example, when comparing the number of firemen at a fire and the damage from the fire, there is a direct relationship; as the number of firemen increases, the damage from the fire also increases. However, an increase in the number of firemen is not the cause of the increased damage. Instead, there is an external factor that is not represented in this relationship. OLS regression minimizes the sum of squared errors by creating a line of best, explained in more depth later in this section. Multiple Regression can be defined by the following formula:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon.$$

In this formula, β_0 is the intercept and $\beta_1 \ldots \beta_k$ are the slopes of the individual regressors. The slope is the true average of a change in the dependent variable y associated with a one-unit increase in the predictor x. The slope of the least squares regression line, β $\hat{}_k$ is a point estimator of β_k . β $\hat{}_k$ varies from sample to sample because it is a random variable. β $\hat{}_k$ can be standardized by subtracting its mean and dividing the difference by the standard deviation. Each coefficient has a β value, representing the value associated with a one unit increase in the predictor. The ε refers to the residual or random error in the model. Without ε, any observed pair (x,y) would fall exactly on the true regression line.

Our analysis uses the following regression equation:

$$lnMedHVal = B_0 + \beta_1 PCBACHMORE + \beta_2 PCTVACANT + \beta_3 PCTSINGLES + \beta_4 lnNBELPOV100 + \varepsilon$$

lnMedHVal refers to the log of median house value in a block group; PCBCHMORE refers to the percentage of residents with a bachelor's degree or higher; PCTVACANT refers to the percentage of vacant units; PCTSINGLES refers to the percentage of single family detached homes; and lnNBELPOV100 refers to the log of the number of persons with incomes below 100% of the poverty level.

Linear regression relies on several assumptions: (1) that the relationship between the dependent variable y and the predictors (x) is linear; (2) that there is no multicollinearity; (3) that observations are independent of one another; (4) and that residuals are homoskedastic. Linearity refers to a constant slope (either positive or negative). Multicollinearity is when variables have a direct correlation with one another, meaning that their *r* value is close to either -1 or 1. However, it is important to keep in mind that linear dependence could exist among three or more variables. The model assumes this is not the case. An *r* value greater than .8 or less than -.8 is typically indicative of multicollinearity, but that value is dependent on the goals of the project. For the purposes of this assignment, a relationship with an *r* value of greater than .8 or less than -.8 is considered too high. Normality of residuals means that the model assumes a normal distribution of the residual values (ε). Violations of the normality of residuals may occur if the dependent and/or independent variables are not normally distributed, or the linearity assumption is violated (e.g., the variables have a logarithmic or sinodal relationship). The final assumption of homoscedasticity assumes that there is no pattern in the residuals. Patterns may include logarithmic relationships or heteroscedasticity. Heteroscedasticity implies non-constant variance in residuals. As x increases, so does the variance of residuals. If this occurs, the assumption of homoscedasticity is violated.

In addition to the parameters $\beta_{1...}$ β_k, and β_0 , which have already been discussed, the variance--notated by σ^2--is also essential to the regression model. The variance (σ^2) determines the amount of variability in the regression model. If σ^2 is small, then the pairs (x_i, y_i) will fall very close to the regression model's best fit line. If the σ^2 value is large, then the pairs (x_i, y_i) will fall farther away from the regressions best fit line. Therefore, a large σ^2 value correlates with a larger mean average residual value (ε). β_k and β_0 are minimizing the sum of vertical distances (ε) between the points and the estimated regression line. Given *n* observations on *y* and *k* predictors ($x_1...x_k$), the estimates β_0, β_1,....β_k,are chosen to minimize the Error Sums of Squares (SSE), expressed in the following formula:

$$SSE = \sum_{i=1}^{n} \varepsilon^2 = \sum_{i=1}^{n} (y - \hat{y})^2 =$$

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_k x_{ki})^2$$

The sum of squares attempts to minimize the distance between points and the estimated regression line. Note that the SSE is simply the sum of the residuals squared. The estimate of σ^2 is defined by the following formula:

$$\sigma^2 = \frac{SSE}{n-(k+1)} = MSE$$

SSE is the numerator and the denominator represents the degrees of freedom for residuals. Here, *n* represents the number of observations and *k* represents the number of predictors. For this particular study, there are 1720 observations and 4 predictors, so the denominator is equal to 1720 - (4+1), representing 1715 degrees of freedom. The MSE in this formula is a function of both bias and variance. Bias refers to how accurate the results are, while variance refers to the variability in the data. Specifically, the MSE is the sum of the variance y^, square bias of y^ and the variance of ε. The OLS model works to minimize the MSE value because it yields unbiased estimates of y^ with the smallest variance. Note that although OLS estimators are unbiased, they can still have greater variance in comparison to other models such as Ridge or Lasso regression. Taking the square root of the MSE yields the root mean squared error (RMSE), which is an estimate of the magnitude of a typical residual. A model with a smaller RMSE is considered to be a better model.

In contrast to SSE, the total sum of squares (SST) is calculated by subtracting the mean value of y from each y_i and then squaring each difference and summing all the differences. The formula can be written out in statistical notation as follows:

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

In this formula, y-bar represents the mean and y_i represents the y value of a singular point. The SST and SSE are both essential in calculating the R^2 using the formula :

In this formula, SSR is equal to 1- SSE, representing the regression sum of squares. R^2 is known as the coefficient of determination. It is the proportion of observed variation in the dependent variable y that is explained by the model. The higher the R^2 value, the greater the proportion of variation explained by the model. We also calculate an adjusted-R^2 for our model. Extra predictors in a model will generally inflate the value of R^2. The adjusted-R^2 adjusts for the number of predictors using the following formula:

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

where R^2 is the coefficient of determination, *n* is the number of observations, and *k* is the number of parameters.

A model utility test, known as the F-ratio is also utilized in this study. It tests the null hypothesis that all beta coefficients are equal to zero. Conversely, the alternative hypothesis is that at least one of the coefficients is not equal to 0. For a model to be successful, the null hypothesis should be rejected. The null hypothesis is rejected when the p-value is greater than 0.05. The p-value is the probability of observing a value that is at least as different from the stated value of H_0 as the given estimated value. If we fail to reject the null hypothesis, none of the independent variables are predictors of dependent variables. The following formulas state the hypothesis and the null hypothesis in statistical notation:

$$H_0\text{: } \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_a\text{: } \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or … or } \beta_k \neq 0$$

The null hypothesis (H_0) is when all the beta coefficients have a value of 0, while the alternative hypothesis (H_a) is when at least one of the beta coefficient values is not equal to 0. In addition to the hypothesis test, a t-test is completed for each predictor *i*. Similar to the model utility test, the null hypothesis is typically rejected when the p value is greater than 0.05 in favor of the alternative. For each variable being tested, the null hypothesis is such that the independent variable is not related to the dependent variable (H_0: β = 0). The alternative hypothesis is such that the dependent variable (H_0: β != 0) is not equal to 0. Consequently, when the two variables are related with a p value greater than 0.05, the null hypothesis is rejected in favor of the alternative. The null hypothesis is written out in statistical notation below .

$$H_o\text{: } \beta = 0$$

$$H_{a:} \beta \neq 0$$

Additional Analysis

Stepwise analysis is a data mining method that selects predictors only on the basis of specific criteria. An example of stepwise analysis is performing the analysis only on predictors below a specific threshold. Another example is taking the smallest value of the Akaike Information Criterion, which is a measure of relative quality of statistical models. However, there are quite a few problems associated with stepwise regression. Since many predictors are excluded from the model on the basis of failing to meet a specific threshold, the results of the final model are not guaranteed to be accurate. Stepwise regression yields a singular model, when often there are several models yielding satisfactory results, which allows the researcher to decide the best model for a particular data set. In data science, variables are often picked as a function of a researcher's knowledge of the topic. For example, in this particular study, the percent of people with a bachelor's degree was included because it is known that those with a bachelor degree often have different spending habits compared to those who do not. When a variable is excluded from the data due to not meeting a specific threshold, the researcher's knowledge is no longer considered in the model. Consequently, there is a higher probability that some important predictors were included in the model, or some important predictors were excluded from the model.

Next, leave out cross validation is a method that holds the first observation out, fitting the model on the remaining observations. If *n* equals the number of observations, then the remaining *n* -1 observations are the training set. Using the parameter estimates from each observation in the training set, the residual value (ε) is calculated for the single hold out in the observation. This process is then iterated over the rest of the values in the training set. Then the RMSE value is calculated using the following formula (as discussed above);

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}\varepsilon_i^2}{n}}$$

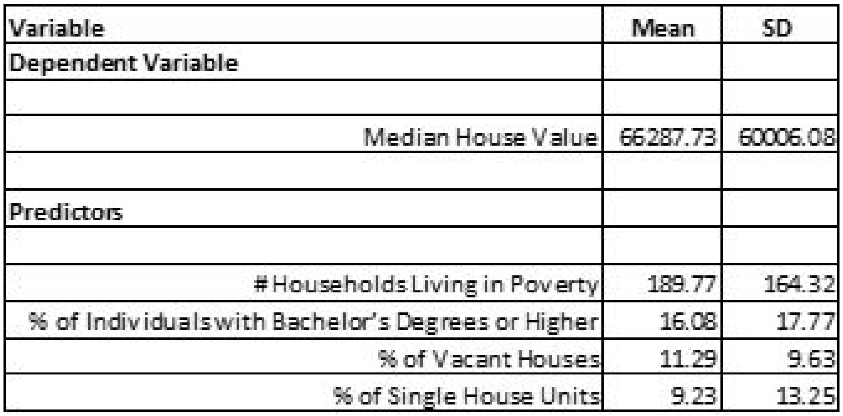
K-fold validation is very similar to leave out cross validation. K-fold validation is the method used in this study. It involves dividing the dataset into k groups of equal size. Each group is known as a fold. The first fold is the validation data set and the model is fitted on the remaining k - 1 folds. The MSE is then computed for the fold validation. The process is then iterated over for each fold acting as the validation data set. The MSE for all the folds is then averaged to create an estimate for the MSE. The square root of the MSE can be taken to yield the RMSE value. This method makes more sense when there are a large number of observations like there are in this data set. Overfitting is a problem often associated with k-fold validation. Even though a linear model does well in predicting the data, a more complicated model that is not linear may not do as well in its predictions. When this occurs, the RMSE value is small for training data and large for validation data.

The statistical analysis was completed in R, and all graphics were made in R or ArcGIS software.

RESULTS

Exploratory Results

Table 1. Summary Statistics



Variable	Mean	SD
Dependent Variable		
Median House Value	66287.73	60006.08
Predictors		
# Households Living in Poverty	189.77	164.32
% of Individuals with Bachelor’s Degrees or Higher	16.08	17.77
% of Vacant Houses	11.29	9.63
% of Single House Units	9.23	13.25

Table 1 contains summary statistics for the following,

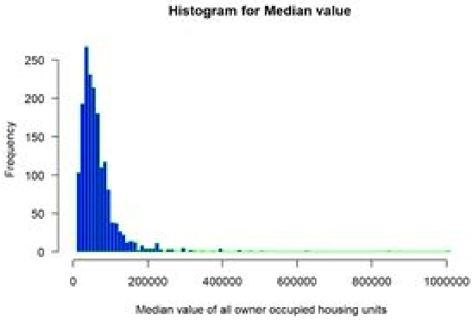
- Median House Value (MEDHVAL)
- Number of Households living in poverty (NBELPOV100)
- Percent of Individuals with a bachelor’s degree or higher (PCBACHMORE)
- Percent of vacant houses (PCTVACANT)
- Percent of single family detached units (PCTSINGLES)

The mean of median house value is $66,287 with a standard deviation of 66,288. The mean for percent of persons below 100% of the poverty line is 189.77 with a standard deviation of 164.32. Mean for percent of persons with a bachelor’s degree is 16.08% with a standard deviation of 17.77. The mean for percent of vacant homes is 11.29 with a standard deviation of 9.63, and the mean for percent of single family homes is 9.23 with a standard deviation of 13.25.

Histograms were created to explore each variable for variance and distribution patterns. Other assumptions will be evaluated in following sections below (see Regression Assumption Checks). Figure 1.1 is the histogram for MEDHVAL. The MEDHVAL histogram shows a left skewed data concentration. Since there is a right tailed distribution, due to high value outliers, we needed to look at the logarithmic transformation, x’ (x'=log$_{10}$(x), x=variable), to “normalize” or to get a log-normal distribution to be able to better fit the data for analysis (see Figure 1.2). Since the distribution in the log-transformed MEDHVAL (LNMEDHVAL) has a log-normal distribution, we will be using the LNMEDVAL for analysis.

Figure 1.1

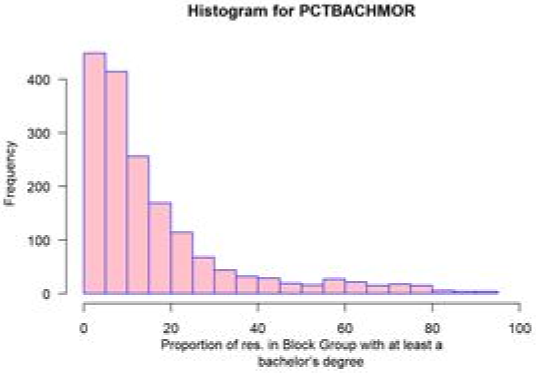
Figure 1.2



Looking at the predictors, PCBACHMORE, NBELPOV100, PCTVACANT, and PCTSINGLES, we see similar distributions in the histograms but for different reasons. Starting with PCBACHMORE, we also see a right tailed distribution in Figure 2.1. When the log-transformation, with a constant added to address the issue of 0 values, x’ ($x'=1+\log_{10}(x)$, x=variable) is applied in Figure 2.2 For the variable (LNPCBACHMORE), we can see a normal transformation take place but with a spike at 0% (or zero-inflated distribution). Since this is a negative response, we can disregard the tailing distribution on Figure 2.1 and assume it has a normal distribution. Hence, we will use PCBACHMORE for the analysis.

Figure 2.1

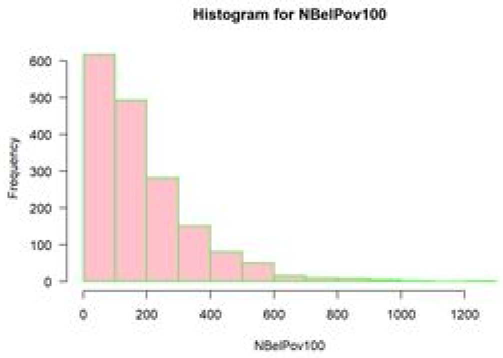
Figure 2.2

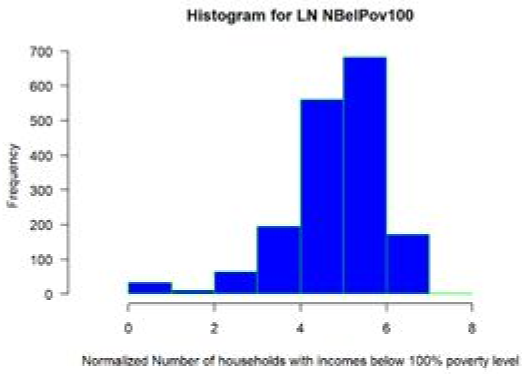


The NBELPOV100 histogram, Figure 3.1, also has a right tailed distribution with high variance, similar to MEDHVAL. Figure 3.2 represents the log-transformation, with a constant (1) applied to account for 0 values. The log-transformed data will be used for analysis.

Figure 3. 1

Figure 3.2





Since the histogram for PCTVACANT, Figure 4.1, is a percentile similar to PCBACHMORE, it has a similar issue with 0% response (zero-inflated distribution), skewing the curve ($x'=1+\log_{10}(x)$). Therefore, a similar method was employed to create Figure 4.2 LNPCTVACANT showing similar results. We will be using PCTVACANT for analysis.

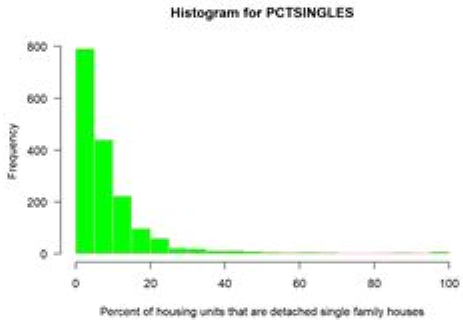
Figure 4. 1

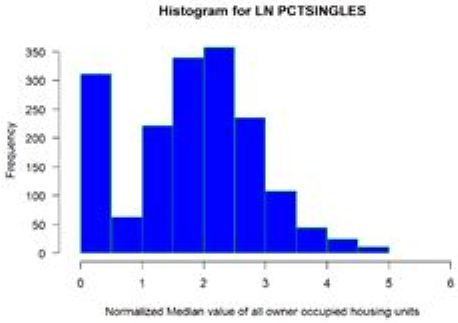
Figure 4.2

The same assumptions are also applied to the PCTSINGLES histogram, Figure 5.1, as PCTVACANT. As observed in Figure 5.2 (LNPCTSINGLES), the skew is caused by the negative response of 0% and can be disregarded. We will be using PCTSINGLES for analysis.

Figure 5. 1

Figure 5.2

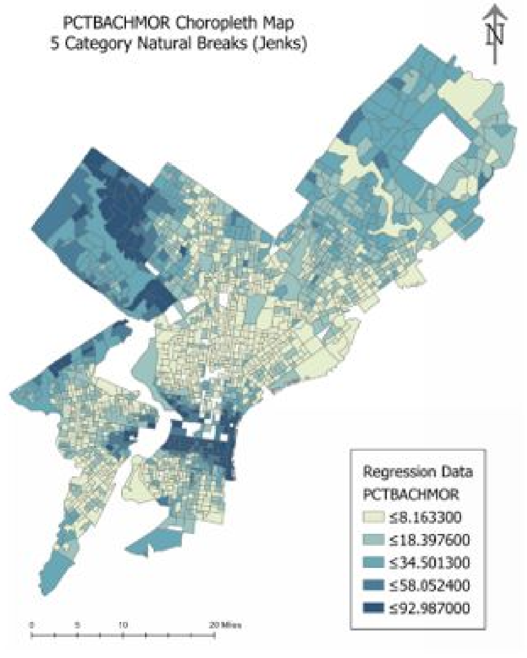


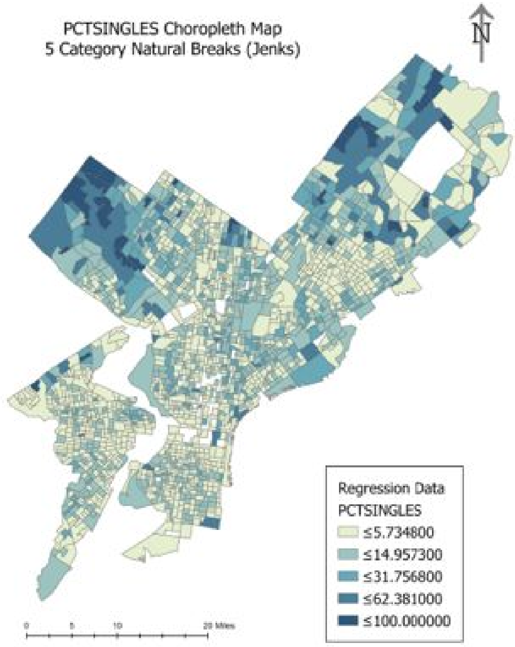


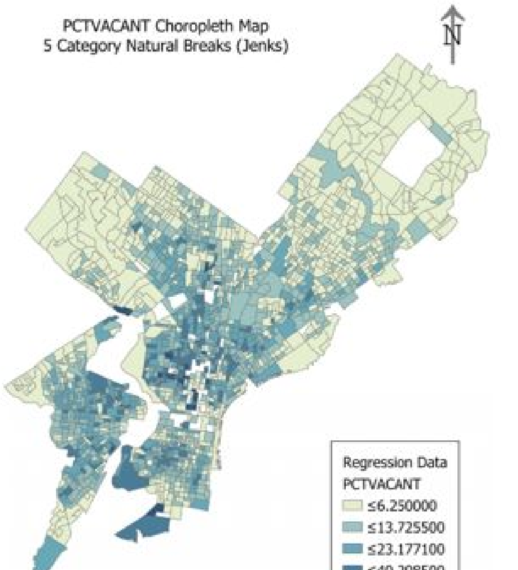
CHOROPLETH MAPS

Choropleth maps were created for the independent variables PCTVACANT, PCTSINGLES, LNNBELPOV10 (see Figure 7.1) and the dependent variable LNMEDHVAL (see Figure 7.2) to spatially visualize the data. Looking at the similarities of the maps, there is most likely a strong positive relationship between PCTSINGLES and LNMEDHVAL and PCTBACHMORE and LNMEDHVAL. There is also likely a strong negative relationship between PCTVACANT and LNMEDHVAL and LNNBELPOV and LNMEDHVAL. In terms of predictors that are inter-correlated, PCTBACHMORE and PCTVACANT are negatively correlated. Multicollinearity could be an issue. The maps also raise concerns about spatial autocorrelation.

Figure 7.1


PCTBACHMOR Choropleth Map
5 Category Natural Breaks (Jenks)
N
Regression Data
PCTBACHMOR
≤8.163300
≤18.397600
≤34.501300
≤58.052400
≤92.987000
0
5
10
20 Miles


PCTSINGLES Choropleth Map
5 Category Natural Breaks (Jenks)
N
Regression Data
PCTSINGLES
≤5.734800
≤14.957300
≤31.756800
≤62.381000
≤100.000000
0
5
10
20 Miles


PCTVACANT Choropleth Map
5 Category Natural Breaks (Jenks)
N
Regression Data
PCTVACANT
≤6.250000
≤13.725500
≤23.177100
≤40.298500
≤77.118600
0
5
10
20 Miles

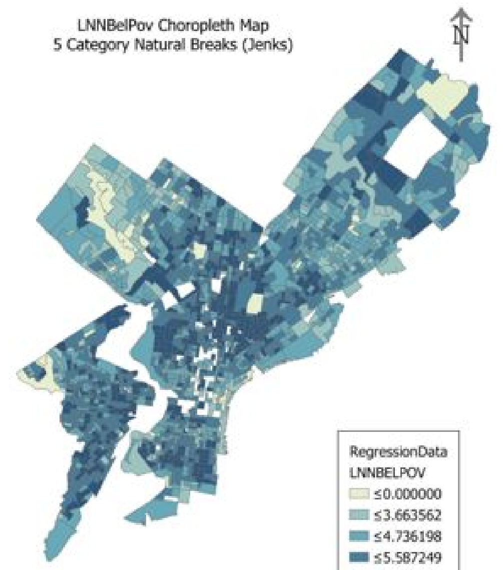

LNNBelPov Choropleth Map
5 Category Natural Breaks (Jenks)
N
RegressionData
LNNBELPOV
≤0.000000
≤3.663562
≤4.736198
≤5.587249
≤7.145196
0
5
10
20 Miles

Figure 7.2 (LNMEDVAL Choropleth Map)

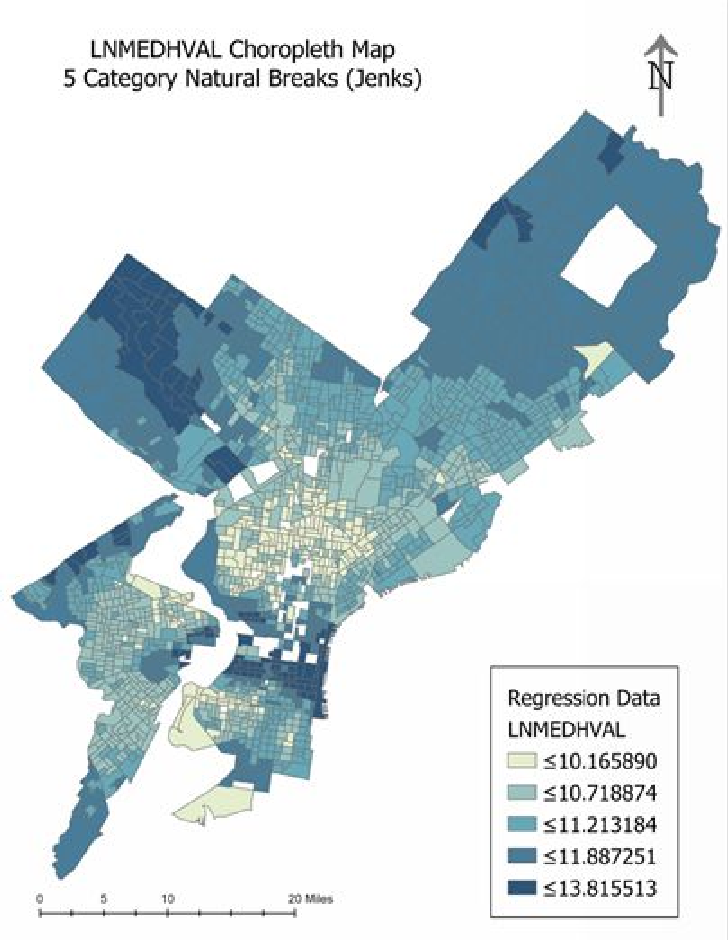


Table 2. Correlation Matrix

Table 2 presents the correlation matrix for predictors versus the dependent variable LNMEDHVAL. There is not evidence of severe multicollinearity, as there are no correlations where *r*>.8 or *r*<-.8. The correlation matrix supports the conclusions we drew from inspecting the predictor maps-- that PCTBACHMOR and PCTSINGLES are positively correlated with LNMEDHVAL, that PCTVACANT and LNNBELPOV are negatively correlated with LNMEDHVAL, and that PCTBACHMORE and PCTVACANT are negatively correlated.

Regression Results

We regressed the log of median house value (LNMEDHVAL) on the proportion of vacant unit (PCTVACANT), the percent of single family detached homes (PCTSINGLES), the proportion of residents with at least a bachelor's degree (PCTBACHMOR and the log of the number of households below 100% of the poverty line (LNNBELPOV100). Table 3 presents the summary output of the regression.

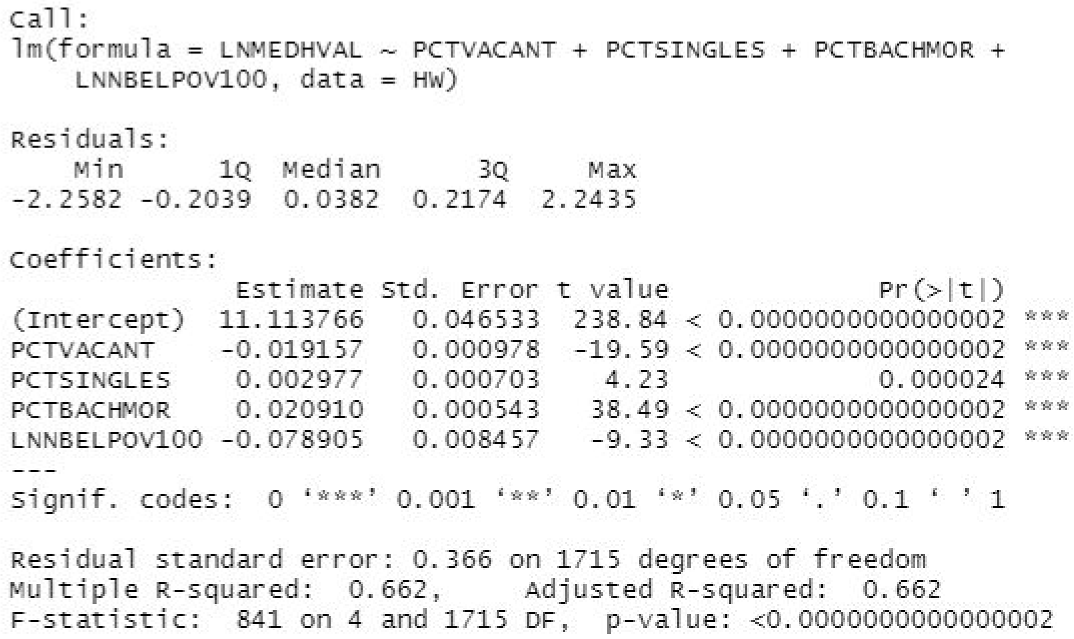
The model's r-squared value is 0.662, meaning 66.2% of the variation in the dependent variable (log of median house value) can be explained by the predictors. The adjusted r-squared value (which controls for the number of predictors in the model) is also 0.662, again indicating that 66.2% of the variation in the log of median house value is explained by the predictors.

The model's F-statistic is 841 with a corresponding p-value of less than 0.0001. This low p-value indicates that we can reject the null hypothesis that all coefficients in the model are 0 for the alternative hypothesis that at least one coefficient is not equal to zero.

The regression output (see below) indicates that the coefficients on all variables are statistically significant at the 1% level with p-values less than than 0.01. We know from examining the sign of the beta coefficients that PCTVACANT and LLNBELPOV100 are negatively associated with median house value. PCT SINGLES and PCTBACHMOR are positively associated with median house value. If we examine the coefficients displayed in the output above, we see that holding all other predictors constant, a one unit (i.e., 1%) increase in the percent of vacant homes (PCTVACANT) is associated with an approximate 1.92% decrease in median house value. (The true value of the change is (e^B1-1)*100%. However, when a beta coefficient is small (less than the absolute value of 0.3) as is the case with each coefficient in this model, (e^B1-1)*100 is approximately equal to 100%B1 percent.). Holding all other predictors constant, a one unit (i.e., 1%) increase in the percent of single family homes is associated with an approximate 0.30%

increase in median house value. Holding all other predictors constant, a one unit (i.e. 1%) increase in the share of residents with a bachelor's degree or higher is associated with an approximate 2.09% increase in median house value. Finally, holding all other predictors constant, a 1% increase in the number of persons below 100% of the poverty line is associated with a 7.89% decrease in median house value. As noted above, all of the coefficients had very low p-values (p<0.0001). Using PCTVACANT as an example, this tells us that if there is no relationship between PCTVACANT and the dependent variable LNMEDHINC, namely, if the null hypothesis that b1=0 is true, than the probability of obtaining a b1 coefficient estimate of -0.0192 is less than 0.0001. Therefore, we can reject the null hypothesis H_0: β_1=0 for the alternative H_a:β_1≠ 0, at the alpha=1% level. (The alpha value refers to the probability of a Type I error--incorrectly rejecting a true null hypothesis). This is the case for all predictors.

Table 3: Regression Summary



```
Call:
lm(formula = LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR +
    LNNBELPOV100, data = HW)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2582 -0.2039  0.0382  0.2174  2.2435

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)  11.113766   0.046533  238.84 < 0.0000000000000002 ***
PCTVACANT    -0.019157   0.000978  -19.59 < 0.0000000000000002 ***
PCTSINGLES    0.002977   0.000703    4.23             0.000024 ***
PCTBACHMOR    0.020910   0.000543   38.49 < 0.0000000000000002 ***
LNNBELPOV100 -0.078905   0.008457   -9.33 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.366 on 1715 degrees of freedom
Multiple R-squared:  0.662,     Adjusted R-squared:  0.662
F-statistic:  841 on 4 and 1715 DF,  p-value: <0.0000000000000002
```

Table 4 presents the ANOVA table for the regression. ANOVA refers to "Analysis of Variance." The ANOVA table includes the sum of squares and mean sum of squares for the model. The sum of squares of the error (SSE) is 230.3, which refers to the total variance in the dependent variable (LNMEDHVAL) unexplained by the model. If we add together the respective sum of squares for each predictor (180.4 + 24.5 +235.1 + 11.7= 451.7), we get the sum of squares of the regression-- the total variance in the dependent variable that is explained by the model.

Table 4: Regression ANOVA

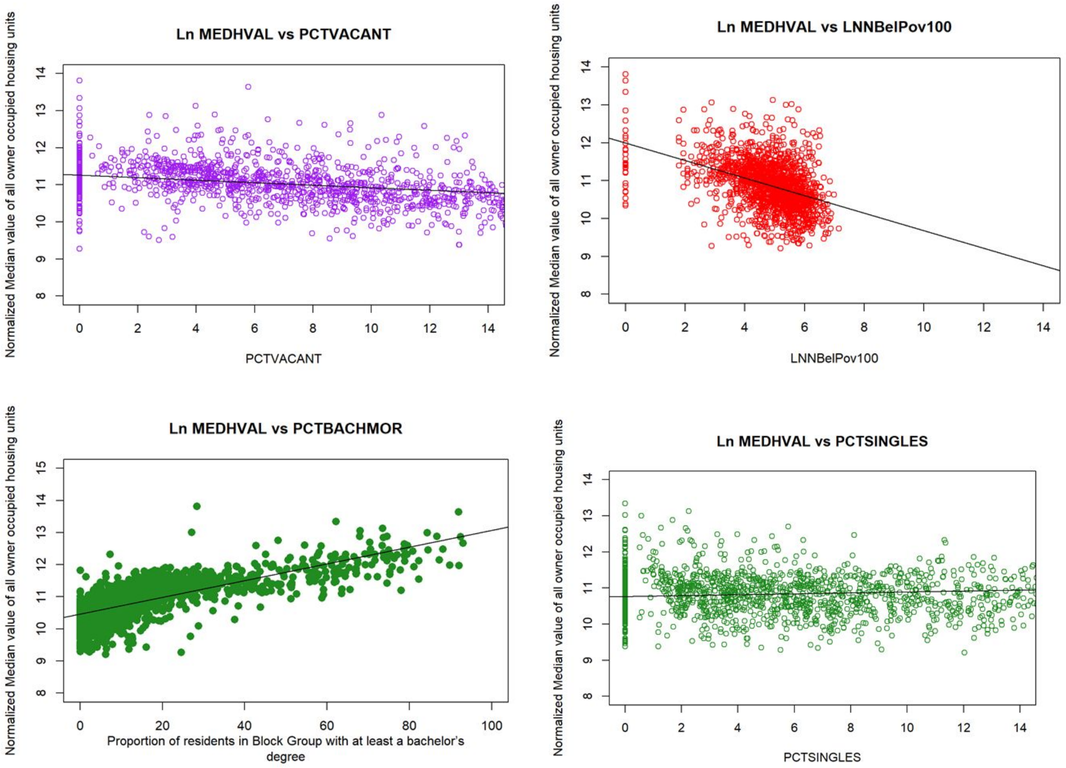
Regression Assumption Checks

In this section, we will discuss the results of our checks of the model assumptions and aptness. We have already discussed variable distribution (see above).

Linearity

The first assumption of OLS regression is linearity of relationship between the dependent variable and predictors. To check this assumption, we examined scatter plots of the dependent variable and each of the predictors (see Figure 8 below). Most of the relationships do not appear linear. The only predictor that seems truly linear is PCTSINGLES. The scatter plots are not truly linear although the r in some of the predictors seems strong, and suggest linearity. The patterns in the scatter plots are problematic due to bends (see PCTBACHMOR), and the different shapes that are created (see LNNBELPOV100). The observed curve in in PCTBACHMORE suggests a non-linear model may be appropriate. Using a linear model is not the appropriate way of evaluation and the curved data requires a different approach using curved lined methods. The only predictor that seems truly linear is PCTSINGLES.

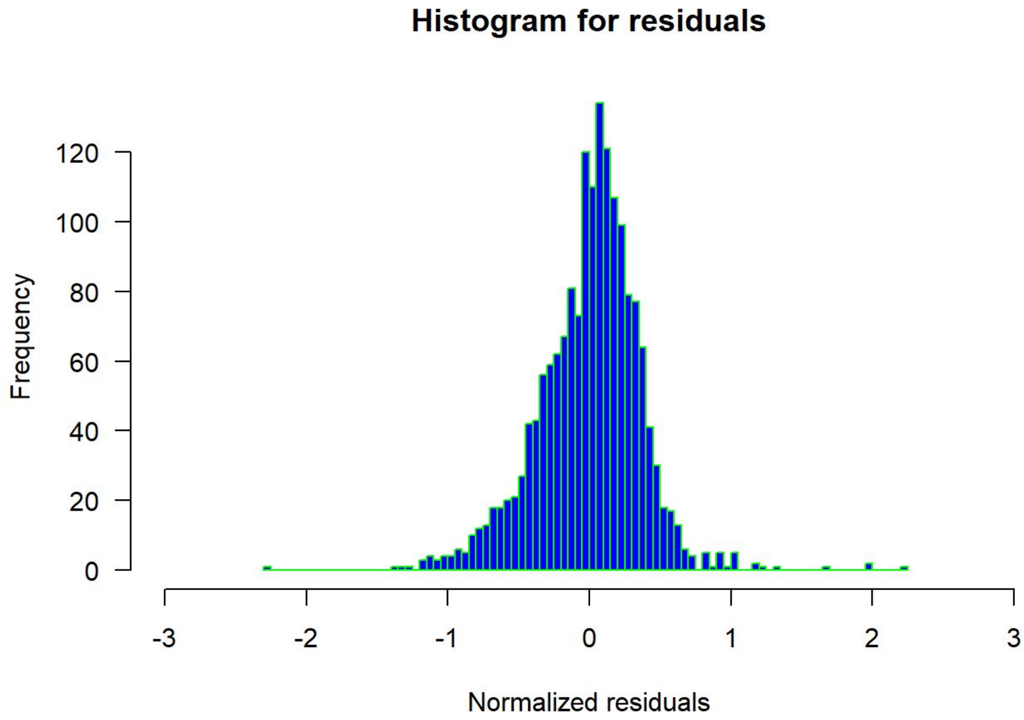
Figure 8.



Normality of Residuals

OLS regression assumes the model residuals are normally distributed. To test this assumption, we examined a histogram of the residuals for normality. The residuals do look mostly normal in the histogram for standardized residuals. They do slightly violate the normality assumption under the Pearson correlation coefficients, due to the slight lean to the right and sudden peak. The issue may be caused by the outliers or from the predictors not being linearly related to the dependent variable. This points to a possible non-linear relationship with the predictors. The removal of outliers may be necessary.

Figure 9.



Homoscedasticity
OLS regression assumes homoskedastic, meaning the variance of the model residuals is constant. To test this assumption we examined a scatterplot of the standardized residuals by predicted value (y-hat) (see Figure 10). The standardized residual is the residual divided by its standard deviation. Standardized residuals are a measure of the difference between observed and expected values.

The formula can also be expressed like the following:

$$Standardized\ Residual\ i = \frac{Residual\ i}{Standard\ Deviation\ of\ Residual\ i}$$

There does seem to be some heteroscedasticity in Figure 10, Predicted Values vs Standardized Residuals. This violates the assumption of Homoscedasticity and invalidates the assumption that the modeling errors all have the same variance. Additionally, the standardized residual plot indicates there may be outliers in the data. Outliers could be removed if they are deemed to be an error in the data. But, OLS should be abandoned, if removal of outliers is unacceptable for generalized least squares (GLS). This is due to OLS being inefficient or possibly leading to incorrect inferences when there is not a linear relationship.

Figure 10.

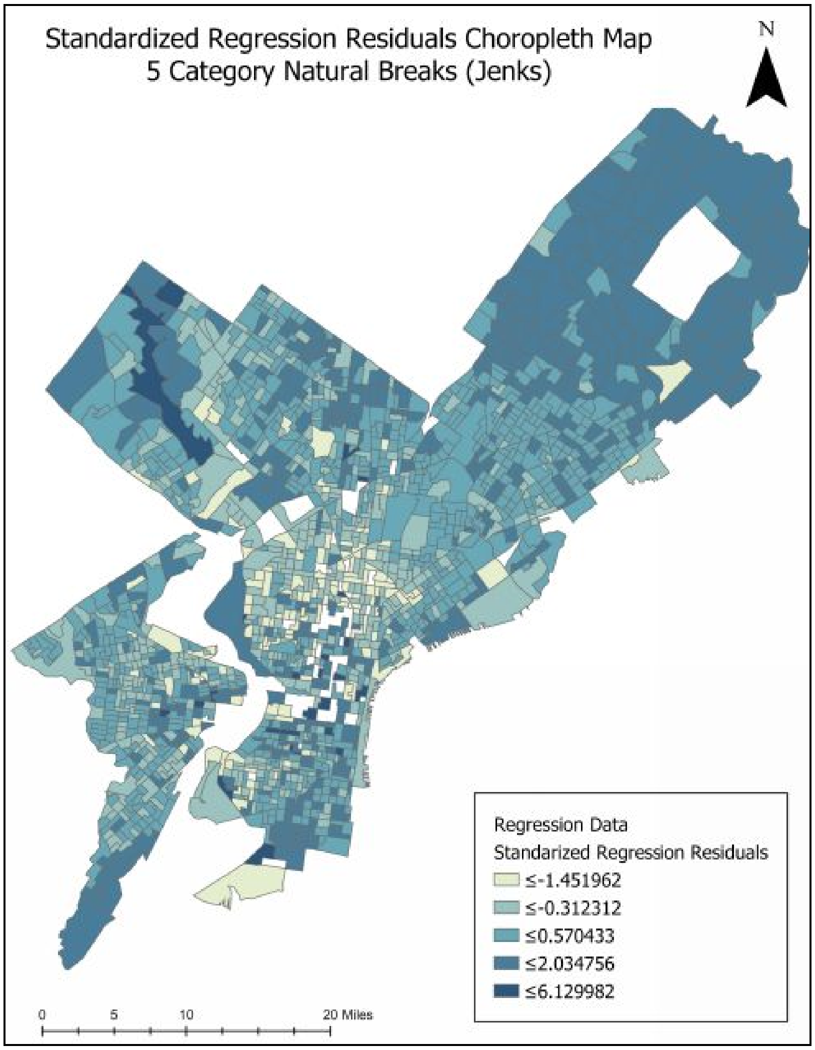
Spatial Autocorrelation

Spatial autocorrelation is a term used to describe the spatial variability in a variable. A positive spatial autocorrelation means that like values are closer together. The variables in this study depict maps with spatial autocorrelation. The median house values show a trend where the North Philadelphia city center has lower median house values. Farther away from the city center in larger parcel sizes is typically associated with higher median rent values.

When looking at the maps of the predictors, each of the predicting variables depict a spatial pattern across space. Some of the variables appear to have more spatial autocorrelation than others. For example, block groups with a higher percentage of residents with a bachelor' degree are typically located away from the city center. Similarly, there are also seemingly higher percentages of single family homes away from the city center. However, vacancies and the log of the number below poverty values have opposite trends, where values decrease away from the city center. Note that although these trends exist in the maps, there are exceptions, as many spatial patterns exist among differing spatial scales.

The standardized residual map depicts widespread spatial autocorrelation as a result of the model. Values close to 0 indicate that the model performed a more accurate prediction, meaning that the difference between the observed median house value and the expected median house value is smaller. In contrast, values farther away from 0 indicate a larger difference between the observed mean median house value and the expected median house value. Referencing figure 11, the map indicates that the model performs better in West Philadelphia and Northeast Philadelphia. These regions had residuals closer to 0. In the center, the residual values were more highly variable. It is clear that there are spatial clusterings of standardized residuals at different spatial scales, indicating spatial autocorrelation.

Figure 11

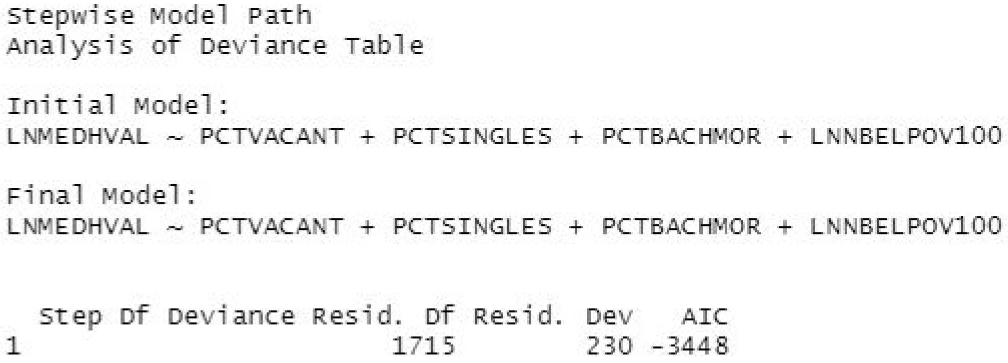

Standardized Regression Residuals Choropleth Map
5 Category Natural Breaks (Jenks)
N
Regression Data
Standarized Regression Residuals
≤-1.451962
≤-0.312312
≤0.570433
≤2.034756
≤6.129982
0
5
10
20 Miles

Additional Models

Stepwise Regression

We performed stepwise regression using the Akaike Information Criterion. The results of the stepwise regression indicate that under the optimal outcome, all four predictors (PCTVACANT, PTCSINGLES, PCTBACHMOR, and LNNBELPOV100) in the initial regression model should remain in the model. (See Table 5 below).

Table 5: Stepwise Regression ANOVA



```
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNNBELPOV100

Final Model:
LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNNBELPOV100


  Step Df Deviance Resid. Df Resid. Dev   AIC
1                       1715        230 -3448
```

Cross-Validation

We performed k-fold cross-validation on the original regression model and on a model that only uses PCTVACANT and MEDHHINC as predictors. Five folds were used (k=5). The root mean squared error (RMSE) for the full model is 0.366. The RSME for the revised model with two predictors is 0.443. The full model (four predictors) has the lower RSME and is therefore optimal.

DISCUSSION & LIMITATIONS

Summary of Analysis & Results

In this paper, we use ordinary least squares (OLS) regression to predict the median house value for block groups in Philadelphia. We first examined our data. After exploring and plotting our data, it was determined that it was appropriate to log transform our dependent variable and one independent variable: number of persons below 100% of the poverty line. We next checked the assumptions inherent to OLS. After noting any violations of those assumptions, we ran a linear regression model with the log of median house value (LNMEDHVAL) as the dependent variable and four variables-- percent vacant units (PCTVACANT), percent of residents with a bachelor's degree or higher PCTBACHMORE), log of number of persons below 100% of the poverty line (LNNBELPOV100), and percent single family detached homes (PCTSINGLES)-- as predictors. The results indicate that PCTVACANT and LNNBELPOV100 were negatively associated with the log of median house value and PCTBACHMORE and PCTSINGLE were positively associated with the log of median house value. All four predictors were statistically significant with p-values less than 0.0001. These findings are not particularly surprising, as there are intuitive connections between our predictors and median house value. Individuals living below the poverty line have limited incomes with which to secure housing and will therefore reside in lower-cost housing. Conversely, educational attainment is strongly related to income. Areas with a large share of persons with bachelor's degrees means incomes and subsequent home prices will be higher. Vacancy rates are a sign of neighbrohood distress and disinvestment, which is reflected in weak housing values. Finally, single family homes tend to be larger and more expensive than multifamily housing, particularly if they are detached.

Our results indicate that our model is of high quality. The r-squared and adjusted r-squared values were both 0.662, indicating 66.2% percent of the variation in LNMEDHVAL can be explained by our model. This is fairly high. Furthermore, the F-ratio test was highly statistically significant with a p-value less than 0.0001. Again, this is an indicator that our model is of high quality.

We next ran stepwise regression to determine the optimal number of predictors for the model, and k-fold cross-validation to compare our initial regression model with a simplified model that regressed LNMEDHVAL on PCTVACANT and a new variable-- median household income (MEDHINC). The results of our stepwise regression indicated that the optimal model includes all four of our initial predictors. This indicates that our initial full model is best as opposed to models that drop one or more predictors. The results of our cross-validation illustrate that our full model of four predictors had a lower RMSE than the two predictor model, again indicating our initial model is optimal.

There are a number of predictors that were not included in our model that are likely associated with our dependent variable. One predictor (median household income) is already included in our dataset. Most others are not. Such predictors could include crime rates, school quality, median square footage, median number of bedrooms, or proximity to parks or other amenities. These additional predictors have the potential to enhance the model. However, the data would need to be collected and examined first before making the decision to include additional predictors.

Limitations

The biggest limitation is a violation of the assumption of the independence of observations as a result of spatial autocorrelation. Other potential limitations include concerns about non-linearity of the relationship between the dependent variable and some predictors, and potential evidence of a non-normal distribution of residuals/heteroskedasticity. Violations of the normality assumption will result in bias in the coefficients. A non-normal distribution of residuals, heteroskedasticity, and spatial autocorrelation will reduce the precision of the estimates, meaning our hypothesis tests and statistical inferences will be affected.

Another limitation is the use of the number of people below 100 percent of the poverty line (NBELPOV100) as a predictor as opposed to the share of persons below the poverty line. Using the raw number as opposed to the share of residents in poverty does not control for the size of the block group and can skew our findings. A large block group could have a small share of persons in poverty but a large absolute number relative to other block groups. Conversely, a blog group with a small population could have a very large share of persons in poverty but a low absolute number relative to other block groups.

References

Kestens, Y., Thériault, M. & Des Rosiers, F. Heterogeneity in hedonic modelling of house prices: looking at buyers' household profiles. *J Geograph Syst* 8, 61–96 (2006). https://doi.org/10.1007/s10109-005-0011-8

Zhou, X.; Tong, W.; Li, D. Modeling Housing Rent in the Atlanta Metropolitan Area Using Textual

Information and Deep Learning. *ISPRS Int. J. Geo-Inf.* **(2010)**, *8*, 349.

James, Witten, Hastie & Tibshirani. (2013). An introduction to statistical learning with applications in R. New York [u.a.]: Springer.

Zhōnghuá Yùfáng-Yīxué Zázhì, Multiple linear regression models with natural logarithmic transformations of variables. (2020). 54(4), 451-456. doi:10.3760/cma.j.cn112150-20191030-00824