OLS Regression to Predict Median House Values in Philadelphia

Fall 2025/MUSA 5000

Jun Luu

Alex Stauffer

Tessa Vu

Introduction

The purpose of this analysis is to examine the relationship between median house values (MEDHVAL) and several neighborhood characteristics using Philadelphia data at the Census block group level.

The dataset used for this study comes from the 2000 U.S. Census and includes 1,720 block groups after data cleaning to remove areas with little to no housing or extreme property values. Key variables in this analysis include the proportion of residents with at least a bachelor's degree (PCBACHMORE), the proportion of vacant housing units (PCTVACANT), the percentage of single-family detached homes (PCTSINGLES), the number of households living below the poverty line (NBELPOV100), and median household income (MEDHHINC).

Higher education levels (PCBACHMORE) have been shown to correlate with higher income levels (MEDHHINC), which can lead to greater homeownership rates and increased housing prices in certain areas (Wang et al., 2022). Urban blight, defined as the presence of deteriorating, substandard, vacant, or abandoned properties, is shown through higher proportion of vacant units (PCTVACANT). This contributes to declining housing values, higher crime rates, and overall neighborhood disinvestment and distress (Bieretz & Schilling, 2019).

By examining these predictor variables, this analysis seeks to better understand how neighborhood characteristics influence median house values across Philadelphia.

Methods

Data Cleaning

The 2000 Philadelphia Census block group level dataset RegressionData.csv contains, among other variables, the variables below:

1. **POLY_ID:** Census Block Group ID.
2. **MEDHVAL:** Median value of all owner occupied housing units.
3. **PCBACHMORE:** Proportion of residents in Block Group with at least a bachelor's degree.
4. **PCTVACANT:** Proportion of housing units that are vacant.
5. **PCTSINGLES:** Percent of housing units that are detached single family houses.
6. **NBELPOV100:** Number of households with incomes below 100% poverty level (i.e., number of households living in poverty).
7. **MEDHHINC:** Median household income.

The original Philadelphia block group dataset has 1,816 observations. The data was cleaned by removing the following block groups:

1. Block groups where population < 40.
2. Block groups where there are no housing units.
3. Block groups where the median house value is lower than $10,000.
4. 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).

Thereby achieving a data set with 1,720 observations.

Exploratory Data Analysis

This section examines the summary statistics and distributions of variables. To achieve this, histograms of each of the predictors are created to check for normal distribution and examine for significant outliers. As part of the exploratory data analysis, this section also examines the correlations between the predictors.

A correlation coefficient is a statistical measure that quantifies the strength and direction of the linear relationship between two or more variables. The sample correlation coefficient r is calculated as:

$$Corr(x,\ y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\left(\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)\left(\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\right)}$$

Where x_i and y_i represent individual observations of the sample data $\overline{x}$ and $\overline{y}$ are the sample means.

The value of correlation coefficient r ranges from -1 to 1. A value of $r = 1$ indicates a perfect positive linear relationship, in that all $(x_i,\ y_i)$ pairs lie on a line with a positive slope, and $r =$

-1 indicates a perfect negative linear relationship, in that all (x_i, y_i) pairs lie on a line with a negative slope. A value of $r = 0$ indicates a lack of linear relationship between the variables.

Multiple Regression Analysis

Method of Regression

Multiple linear regression is a method used to examine the relationship between a dependent, or response, variable and one or more independent, or predictor, variables. In other words, this method visualizes the relationship between dependent variable y and a set of predictors x_1, x_2, … , x_k. Regression is used to predict the dependent variable based on the predictors, to test hypotheses about relationships between variables, and to understand which predictors are important in assessing the dependent variable. Each independent variable has its own slope coefficient, which indicates the relationship of that particular predictor with the dependent variable, while controlling for all other predictor variables in the regression. However, it is important to note that if an independent variable is a significant predictor of the dependent variable, this *does not* imply causation.

Equation

The equation for this project is:

$$\ln(MEDHVAL) = \beta_0 + \beta_1(PCTVACANT) + \beta_2(PCTSINGLES) +$$

$$\beta_3(PCTBACHMOR) + \beta_4(\ln(NBELPOV100)) + \varepsilon$$

Each independent variable in the model was assigned a slope coefficient β_i, which represents the relationship between that specific predictor and the dependent variable, controlling for all other predictors. In this context, each β_i quantifies the expected change in the dependent variable associated with a one-unit increase in the corresponding independent variable, holding the remaining variables constant. The intercept β_0 represents the predicted value of the dependent variable when all predictors are equal to zero. The random error term or residual ε captures the unexplained variation in the dependent variable—the difference between the observed and the predicted values based on the regression equation.

Regression Assumptions

Multiple regression analysis relies on several key assumptions required for operation. The first assumption is that the relationship between the response variable and each of the predictor variables is linear. In other words, the change in a predictor variable maintains a constant and proportional relationship with the change in the dependent variable. By examining scatter plots of y in relation to each predictor x , this assumption can be verified. Should the relationship not be linear, variables can be logarithmically transformed, or a polynomial regression can be conducted.

The second assumption maintains that all observations and residuals are independent, in other words, they should not be influenced or related to the other observations—residuals should not be correlated nor predict the next observation. Violations, such as temporal or spatial autocorrelation, can bias standard errors and lead to incorrect statistical inferences.

The third assumption requires that the variance of the residuals is constant across all levels of the predictor variables, or homoscedasticity. This assumption is assessed visually using a scatter plot of standardized residuals by predicted values.

The fourth assumption relies on the normality of residuals, in that the errors are normally distributed with a mean of zero. According to the Central Limit Theorem, the sampling distribution of the regression coefficients approaches normality as the sample size increases, which provides theoretical justification for this assumption even when individual observations are not normally distributed.

The fifth assumption is that there is no multicollinearity among predictors, meaning the independent variables are not highly correlated with each other. To test multicollinearity, we can regress the predictor on all remaining predictors. If the correlation R^2 between predictors is $R^2 >$ 0.8, this indicates the existence of multicollinearity. Furthermore, we can test the R^2 from above to calculate the Variance Inflation Factor (VIF). The VIF for each predictor k is defined as follows:

$$VIF_k = \frac{1}{1 - R\frac{2}{k}}$$

In the formula above $R\frac{2}{k}$ is the R^2 from regressing the predictor k on the remaining predictors. A the VIF value of 1 means there is no correlation among k^{th} predictor and the remaining predictor variables. The general rule of thumb is that VIF values exceeding four warrant further investigation, while VIFs exceeding ten indicate serious multicollinearity, requiring correction.

Parameters to Estimate

The multiple regression model requires estimation of $k + 2$ parameters: the intercept β_0, the k regression coefficients $\beta_1, \beta_2, \ldots, \beta_k$, and the error variance σ^2. In this case, they are β_1, β_2, β_3, and β_4, or PCTVACANT, PCTSINGLES, PCTBACHMOR, and LNNBELPOV100, respectively. The parameter σ^2 determines the amount of variability within the regression model. If $\sigma^2$is small, the observed pairs (x_i, y_i) will fall close to the true regression line. If σ^2 is large, the observed pairs (x_i, y_i) will be spread out from the true regression line, reflecting greater unexplained variability.

Method of Estimating Parameters

The method of estimating the parameters in multiple regression is the sum of squared errors. The least squares estimators for the regression coefficients are obtained with the below equation:

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

$$\sum_{i=1}^{n}\left(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_{1i} - \widehat{\beta_2}x_{2i} - \cdots - \widehat{\beta_k}x_{ki}\right)^2$$

Given n observations on observed response value y , and number of predictors k and their observed values $x_1 \dots x_k$, the beta coefficient estimates $\widehat{\beta_0}$, $\widehat{\beta_1}$, $\widehat{\beta_2}$, … , $\widehat{\beta_k}$ are chosen simultaneously to minimize the expression for the Error Sum of Squares (SSE), given by the above equation.

In this case, $\widehat{\beta_1}$, $\widehat{\beta_2}$, $\widehat{\beta_3}$, and $\widehat{\beta_4}$ all represent the estimated change in y_i when the given predictor increases by a single unit.

Coefficient of Determination (R^2)

The coefficient of multiple determination R^2 is the proportion of variance explained by all k predictors in the regression model, and generally, the more predictor variables there are in a model, the larger R^2 will be. The coefficient of multiple determination's equation is shown below:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \left(\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}\right)$$

In the previous paragraph regarding an increase in the R^2 value being associated with an increase in the number of predictor variables, an adjustment to the R^2 value needs to be calculated, known as adjusted R^2, with the equation given below:

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

The resulting adjusted R^2 is the value reported; it is a more precise modification because it decreases when a predictor variable has a very marginal improvement to the model, meaning it penalizes any predictor variables deemed irrelevant.

SSE is the Sum Squared Errors, which is the summation of the squared difference between $\hat{y}$, which is the predicted value of y , from each y_i, which is the observed value. This value is the amount of variance in y that is unexplained by the regression model. SST is the Total Sum of Squares, which is the summation of the squared difference between $\overline{y}$, the mean value of y , from each y_i. This value is the total amount of variance in y . So, the value when dividing SSE by SST is the proportion of total variance that is unexplained by the model, and when that value is subtracted from 1, the value is then the proportion of observed variation in the dependent variable y explained by the model. The term n is the number of observations in the sample, in this case it is 1,720. The term k is the number of predictors in the model, in this case it is four.

Hypothesis Testing

Two types of inferences test the hypotheses below:

$$H_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_a = at\ least\ 1\ \beta_i \neq 0$$

H_0 (null hypothesis) is where all coefficients are equal to zero, indicating no relationship.

H_a (alternative hypothesis) is where one of the predictor variables β_i does not equal zero, indicating at least one relationship between the response variable and a predictor variable.

The F-Ratio, also known as the F-Test, is a preliminary step in hypothesis testing. The F-Ratio tests the H_0, where all coefficients in the model are all zero, versus the H_a, where at least one of the predictors is not zero. In other words, it compares the means among three or more groups, determining if at least one predictor is useful. A large F-Ratio is a strong case against the H_0, and determining if that value is statistically significant would come from the associated p-value being extremely small.

The t-Test is also conducted for each predictor i , and compares the means of the dependent and independent variables, that is, comparing their correlation and relationship with one another, and determining if they are statistically significant with the associated p-value. In other words, it compares the means between two groups, that of the predictor variable and the response variable.

How these significance tests play with one another is as follows: the F-Test explains whether at least one predictor in the multiple regression model is useful. However, it does not explain what predictor that is, nor how many predictors are useful. So, the t-Test is conducted on each of the predictors β_i to determine their significance with the response variable y . And again, an associated extremely small p-value would indicate a rejection of the H_0, where it is unlikely, but not impossible, that H_0 is true.

Additional Analyses

Stepwise Regression

Stepwise regression is a model selection procedure that automatically selects a subset of predictor variables to include in the final model. The method iteratively adds or removes predictors based on statistical criteria, such as p-values or information criteria, to balance model complexity and fit. Stepwise regression can be useful for exploratory data analysis and for identifying the most important predictors.

However, it has several limitations: it can be sensitive to multicollinearity, may select different models depending on the algorithm used (forward vs. backward selection), can result in biased parameter estimates and standard errors, and may capitalize on chance relationships in the data. Additionally, the statistical significance levels of individual coefficients in a stepwise model can be misleading.

k-Fold Cross-Validation

K-fold cross-validation is used to evaluate a model’s quality. For this model, a five-fold cross-validation ($k = 5$) procedure was used. In this approach, the dataset was randomly divided into five approximately equal-sized folds. For each iteration, four folds were used to train the model, and the remaining fold served as the validation set. The Mean Squared Error (MSE) was calculated for the validation set, and this process was repeated five times so that each fold served as the validation set once. The overall MSE estimate is calculated by averaging the MSE values across the five folds.

The Root Mean Squared Error (RMSE) was derived as the square root of this average MSE:

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

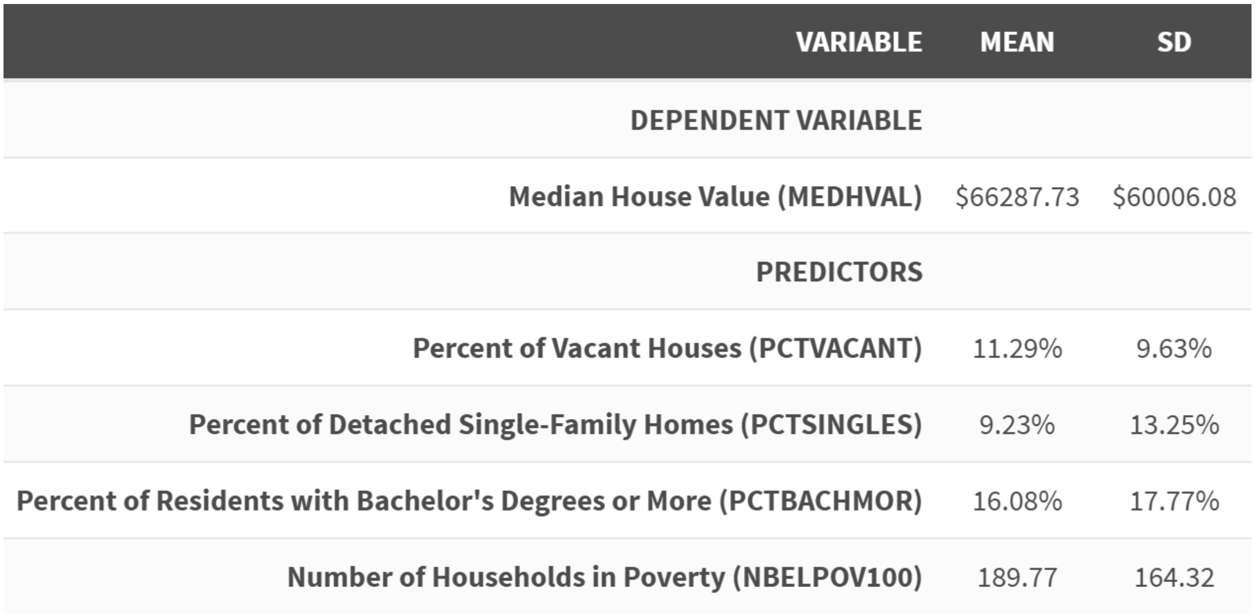
The RMSE measures a model's prediction error in the same units as the dependent variable. Models are compared based on their RMSE values, with the model yielding the lowest considered the best performing.

Software

All analyses were conducted in R Studio 2025.09.1+401 using R language version 4.5.1.

Results

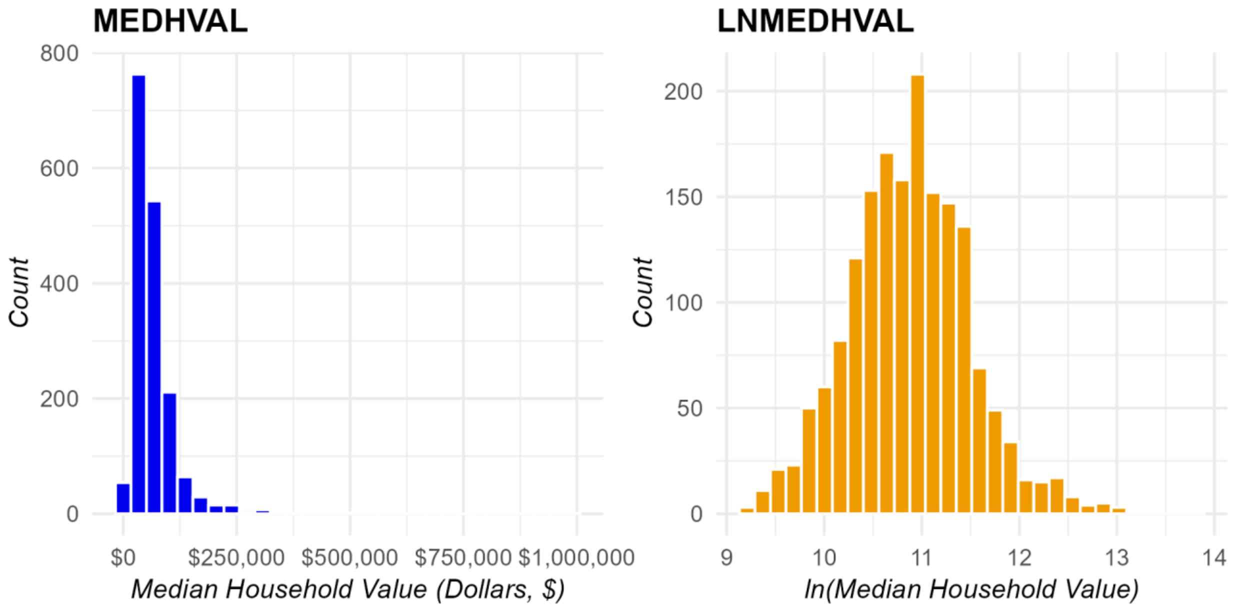
Exploratory Results



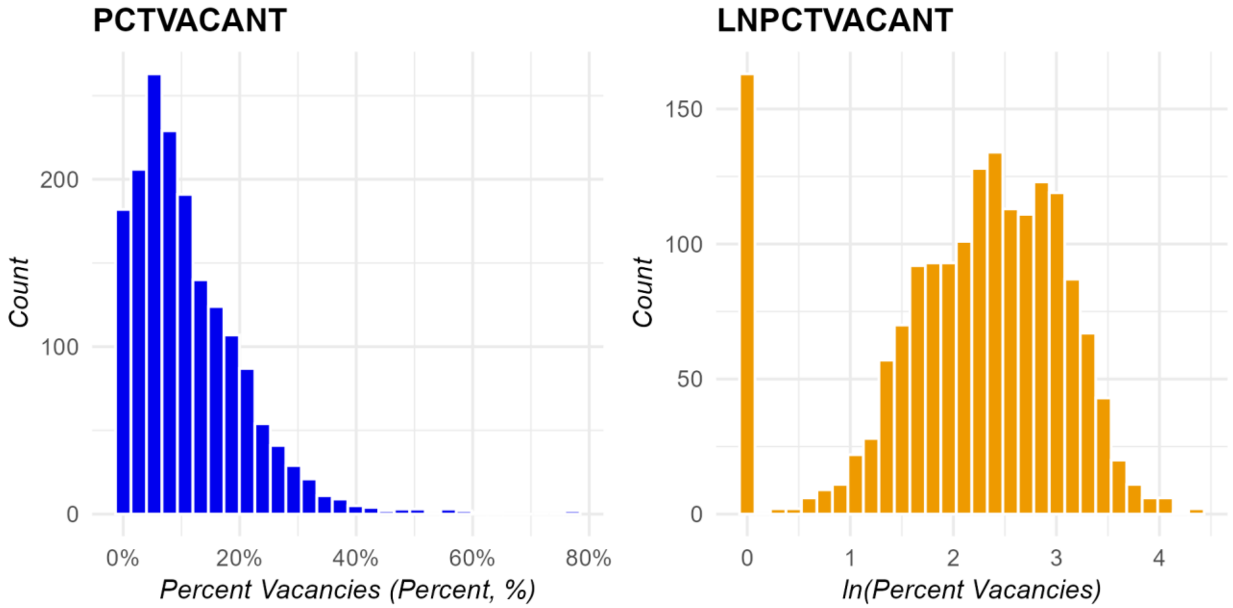
VARIABLE	MEAN	SD
DEPENDENT VARIABLE		
Median House Value (MEDHVAL)	$66287.73	$60006.08
PREDICTORS		
Percent of Vacant Houses (PCTVACANT)	11.29%	9.63%
Percent of Detached Single-Family Homes (PCTSINGLES)	9.23%	13.25%
Percent of Residents with Bachelor's Degrees or More (PCTBACHMOR)	16.08%	17.77%
Number of Households in Poverty (NBELPOV100)	189.77	164.32

Table 1 Summary Statistics

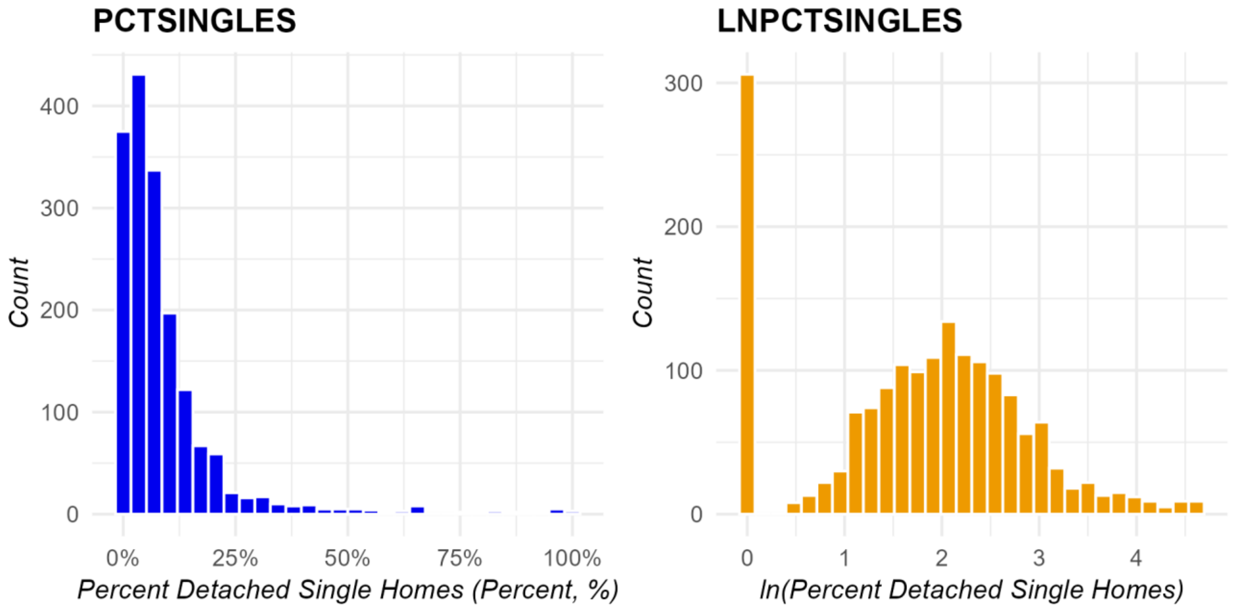
The Summary Statistics Table 1 shows the dependent variable, median house value (MEDHVAL), and the predictor variables. Overall, the means and standard deviations are relatively close across variables, though the bachelor’s degree variable (PCTBACHMOR) shows a standard deviation higher than its mean. This suggests considerable variability within the data. Since several variables have standard deviations as high as or higher than their means, the data may be quite inconsistent, indicating fluctuations and potential disparities in housing values across neighborhoods. The presence of large deviations also suggests that outliers may be influencing the averages, meaning the mean may not accurately represent the typical value. In this case, the median might better capture central tendencies, and the distribution may be skewed rather than normally distributed.



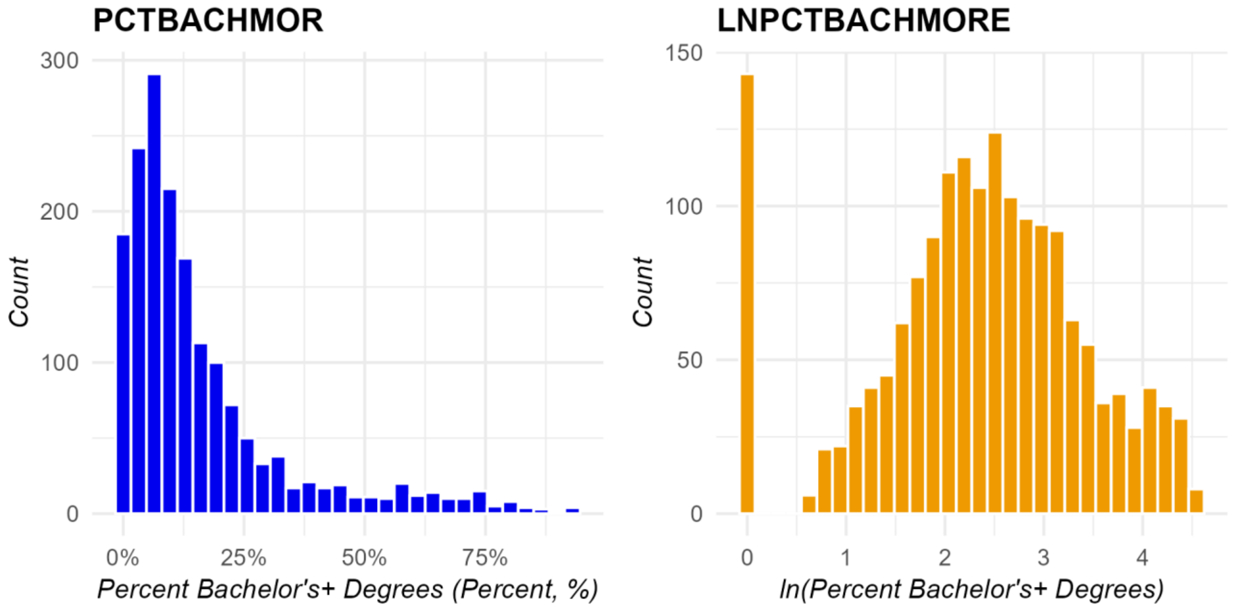
Graph 1.1 MEDHVAL and LNMEDHVAL Histograms



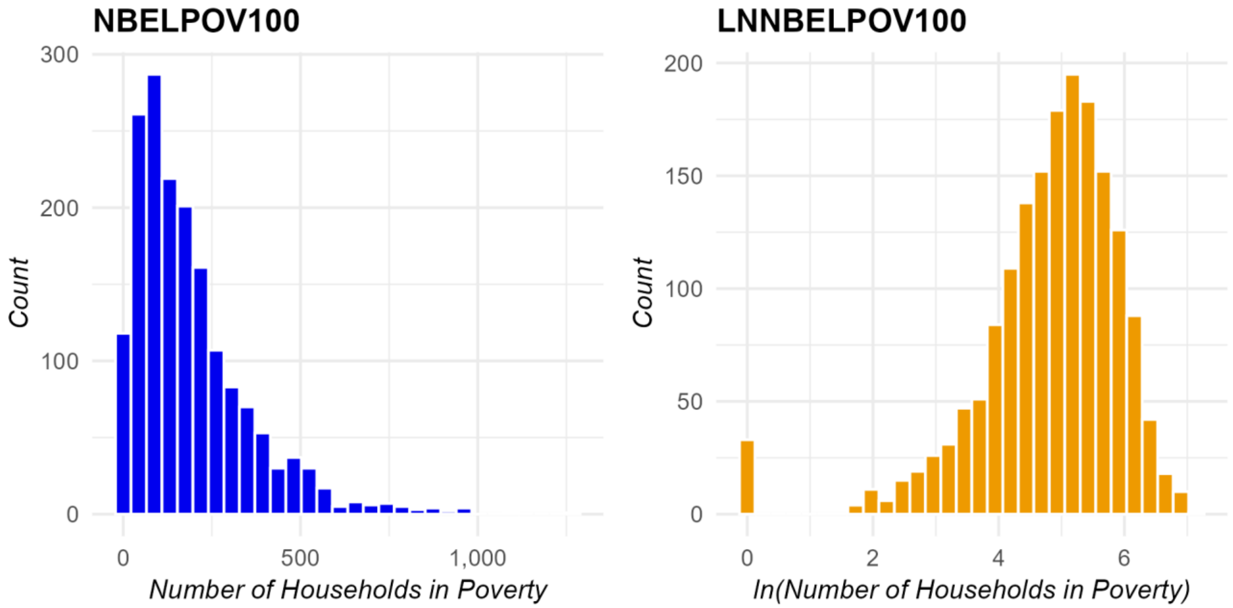
Graph 1.2 PCTVACANT and LNPCTVACANT Histograms



Graph 1.3 PCTSINGLES and LNPCTSINGLES Histograms



Graph 1.4 PCTMBACHMOR and LNPCTBACHMORE Histograms

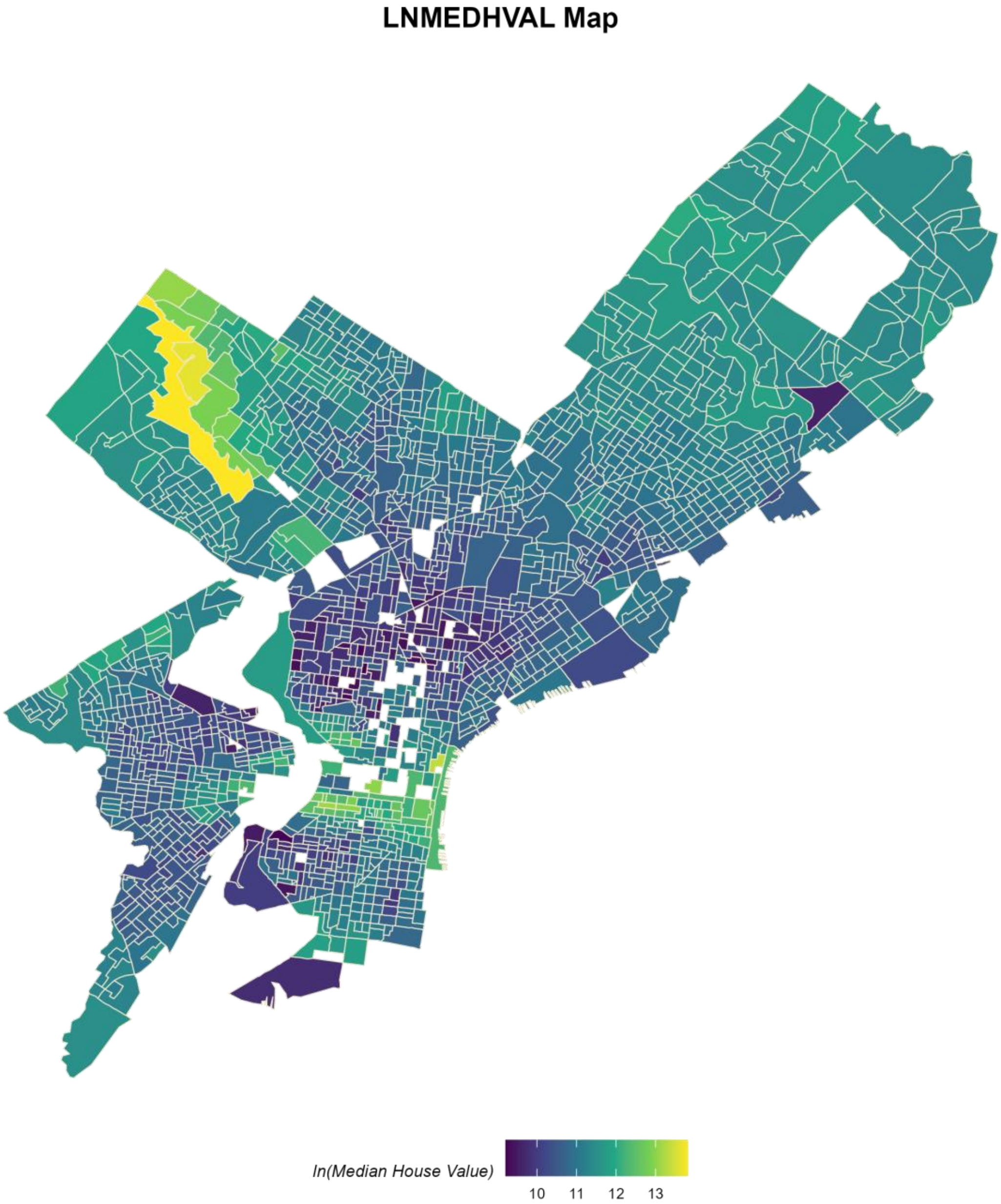


Graph 1.5 NBELPOV100 and LNNBELPOV100 Histograms

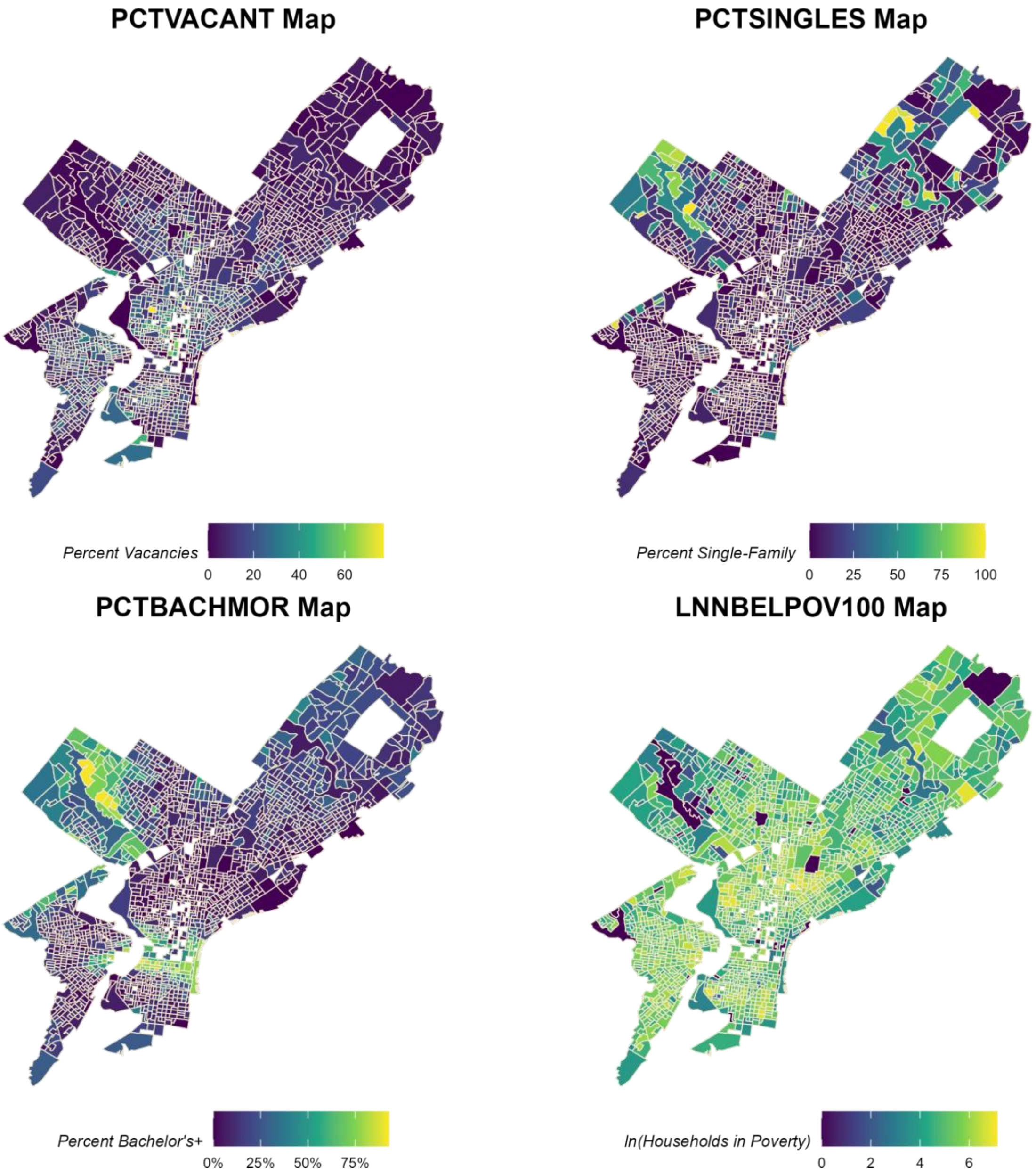
All variables were first graphed to review normality using histograms. None of the original variables appeared to follow a normal distribution, so all logarithmic transformations were applied to assess whether the distributions could be normalized.

Overall, the logarithmic transformation substantially improved normality for the dependent variable (MEDHVAL), resulting in a roughly symmetric distribution for LNMEDHVAL. The only predictor that appeared more normally distributed after transformation was NBELPOV100. The other predictors, PCTBACHMOR, PCTVACANT, and PCTSINGLES, were still skewed and showed zero-inflated distributions even after transformation.

An examination of the other regression assumptions will be discussed in the Regression Assumption Checks section below.



Graph 2.1 LNMEDHVAL Choropleth Map



Graph 2.2 PCTVACANT, PCTSINGLES, PCTBACHMOR, and LNBELPOV100 Choropleth Maps

Refer to Graph 2.1 for the spatial distribution of the dependent variable, median house value. All the predictor variables across Philadelphia's block groups are shown in Graph 2.2.

Visually, the maps for median house value, percentage of population with a bachelor's degree, and percentage of detached homes appear quite similar. There are higher values clustered towards the downtown area and decreasing as you move away from the city center, suggesting a strong positive relationship between income and housing value. Adversely, the percentage of vacant units and households below the poverty line have an inverse relationship. Based on these visual trends, income and education may be correlated with each other as well as with housing value, which raises the possibility of multicollinearity among these predictors.

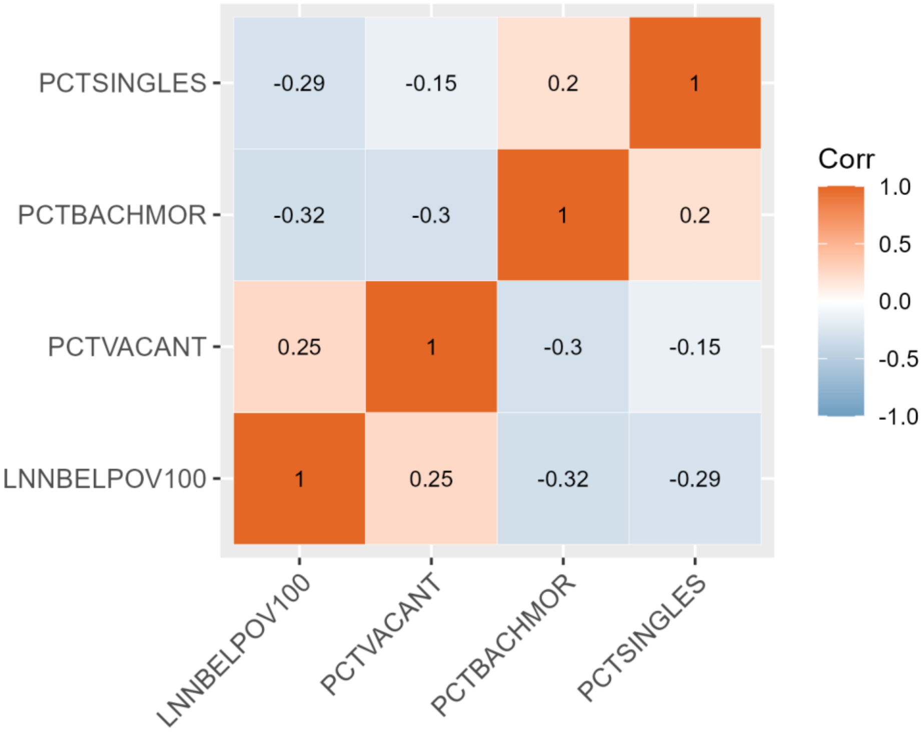
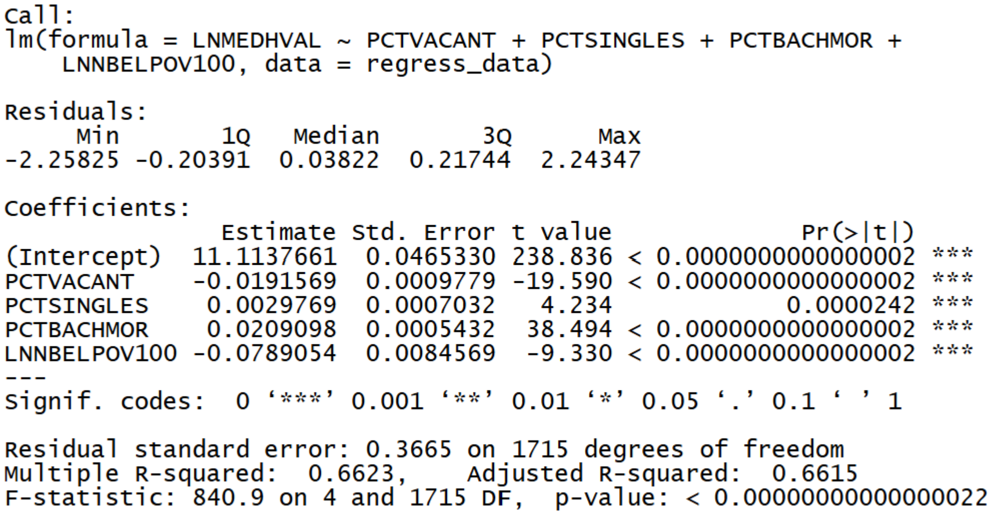


Figure 1 Correlation Matrix

The correlation matrix in Figure 1 does not show severe multicollinearity, all coefficients within the figure have absolute values of 0.32 or less, indicating that the predictor variables have very minimal correlation with each other. Multicollinearity was considered a possibility when observing the choropleth maps' spatial patterns and the pockets of inverted blocks between LNNBELPOV100 and predictor variables PCTBACHMOR and PCTSINGLES, but the correlation matrix indicates there is very little correlation among the predictor variables, and a VIF analysis could be conducted as another multicollinearity check.

Regression Results



```
Call:
lm(formula = LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR +
    LNNBELPOV100, data = regress_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 ***
LNNBELPOV100 -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
```

Figure 2 Summary Statistics

The log of median household value (LNMEDHVAL) was regressed on the percent of vacant housing units (PCTVACANT), percent of detached single-family houses (PCTSINGLES), percent of residents with at least a bachelor's degree (PCTBACHMOR), and the log of number of households in poverty (LNNBELPOV100). Considering the log-transformed median household value response variable (LNMEDHVAL) and poverty predictor variable (LNNBELPOV100), a 0.01 change in their values may be interpreted as a percent change in the original, pre-transformed value.

The regression output indicates that the percent of vacant housing units, percent of detached single-family homes, percent of residents with at least a bachelor's degree, and number of households in poverty are highly significant and are positively associated with median household value ($p < 0.0001$ for all variables).

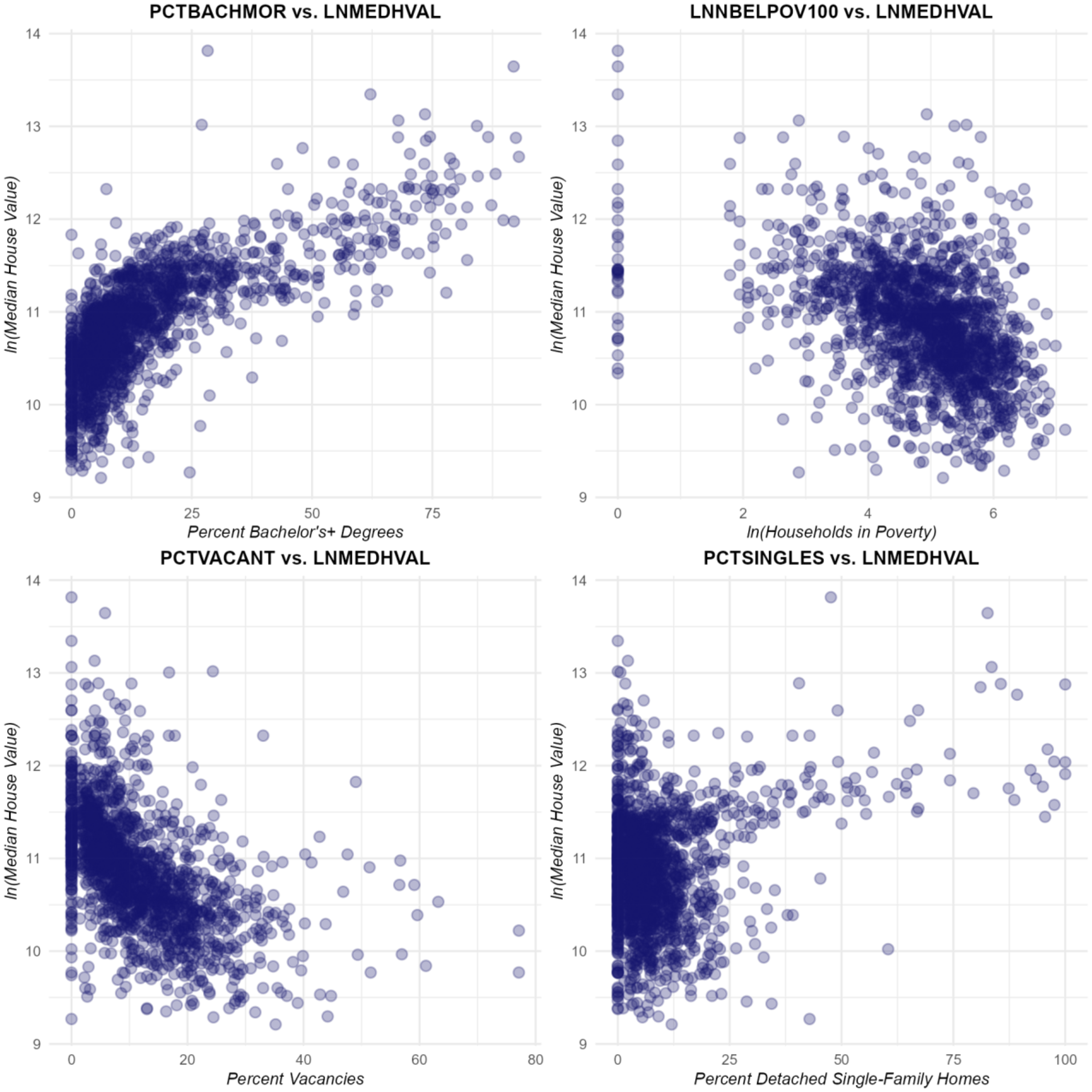
A one percent increase in the percentage of vacancies within the block group is associated with a rounded $\beta_1 = -0.0192$ change, an approximately 1.92% decrease in median household value. A one percent increase in the percentage of detached single-family homes within the block group is associated with a rounded $\beta_2 = 0.0030$ change, an approximately 0.03% marginal increase in median household value. A one percent increase in the percentage of residents with at least a bachelor's degree is associated with a $\beta_3 = 0.0209$ change, an approximately 2.09% increase in median household value. A one percent increase in the number of those in poverty is associated with a rounded $\beta_4 = -0.0800$ change, an approximately 0.08% decrease in median household value. And all given predictor beta coefficient changes are holding other predictors constant.

The p-value of less than 0.0001 for all predictor variables indicates that if there is actually no relationship between the predictor variables and the dependent variable LNMEDHVAL (i.e. if the null hypotheses that $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ are actually true), then the probability of getting a β_1 coefficient estimate of –0.0192, a β_2 coefficient estimate of 0.0030, a β_3 coefficient estimate of 0.0209, and a β_4 coefficient estimate of –0.08 are all less than 0.0001. These low probabilities indicate that we can safely reject $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ for $H_a: H_a = at\ least\ 1\ \beta_i \neq 0$ (at most reasonable levels of $\alpha = P(Type\ I\ error)$).

66.15% of the variance in the dependent variable is explained by the model (R^2 and adjusted R^2 are 0.6623 and 0.6615, respectively). The low p-value less than 0.0001 associated with a large F-Ratio 840.9 shows that it is safe to reject the null hypothesis that all coefficients in the model are 0.

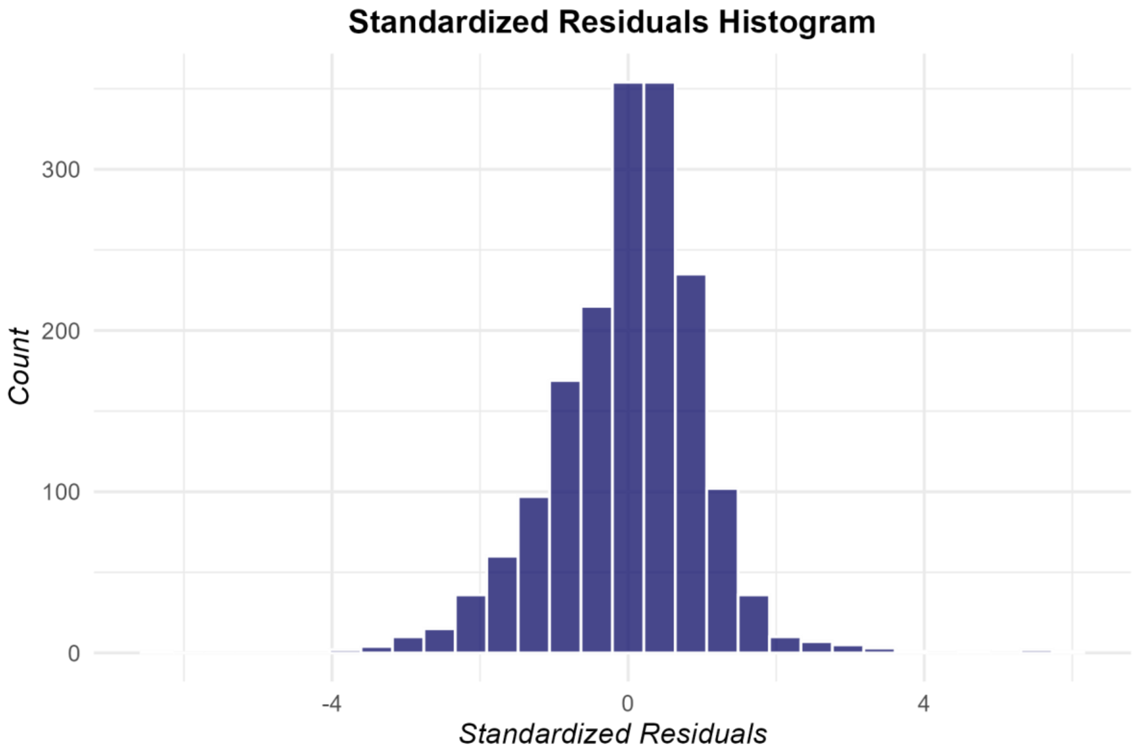
Regression Assumption Checks

This section discusses testing model assumptions, and variable distributions were analyzed earlier.



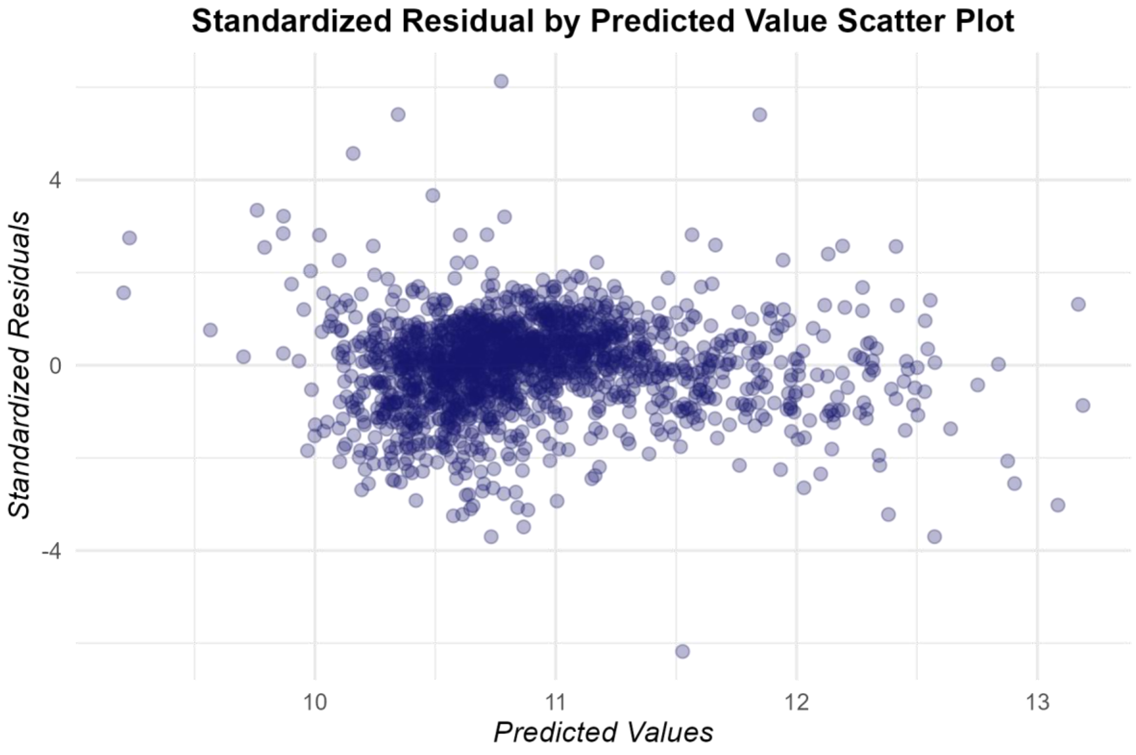
Graph 3.1 Scatter Plots of Predictor Variables vs log(Median House Values)

Graph 3.1 presents the scatter plots of each of the predictor variables' relationships with the dependent variable LNMEDHVAL. PCTBACHMOR has an almost concave downward curve that rises again as the percentage of those who hold at least a bachelor's increases, LNNBELPOV100 is very clustered to one area of the graph between values four to six on the x-axis without any clear linearity, PCTVACANT has a downward sloping curve, and PCTSINGLES has a concentrated form from zero to 25 on the x-axis with several observations aligned horizontally around when y is 11.5 to 12.5. The multivariate regression model assumes relationship linearity; however, it's clear that none of these predictors are linear.



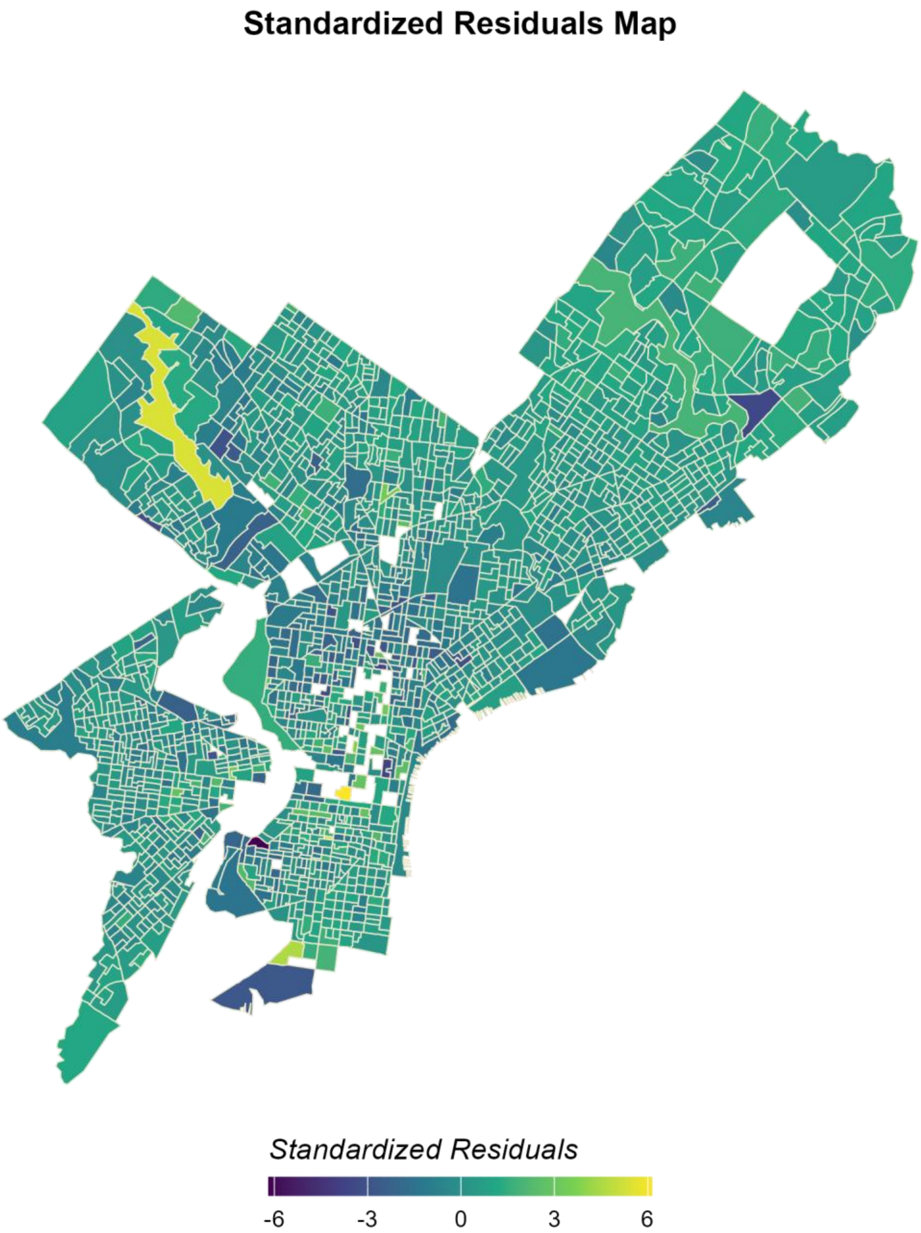
Graph 3.2 Standardized Residuals Histogram

Graph 3.2 above presents a histogram of the standardized residuals, which have a normal shape in accordance to one of the assumptions in OLS regression.



Graph 3.3 Standardized Residuals Scatter Plot

Graph 3.3 presents a scatter plot of the standardized residual by the predicted value. Standardized residuals are residuals divided by the standard error, and they are used for comparing residuals of different observations to each other. Visually, the plot is homoscedastic, which is in accordance with one of the assumptions in OLS regression. Another assumption is that there are no vast outliers, and while it seems like there are a few outlier observations, they might not have significant leverage, except for perhaps the observation at the bottom that lies on approximately (11, -6) that deviates the generally constant variance—further analysis in removing outlier points and comparing with the original would be needed to determine if that is the case.

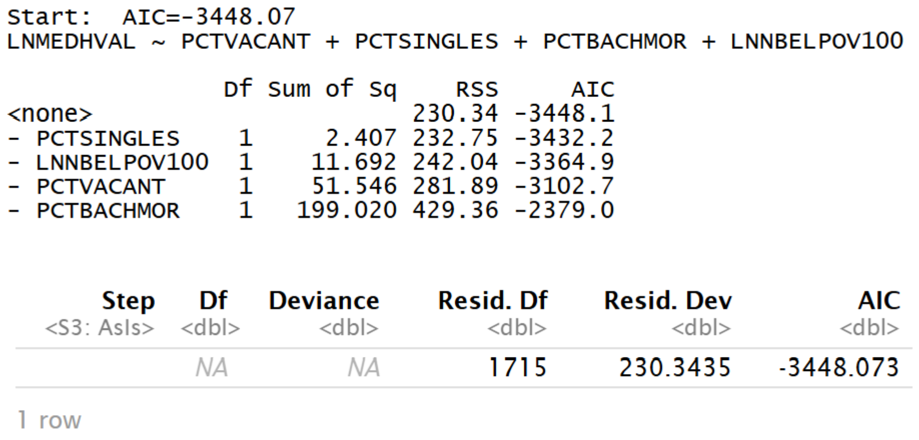


Graph 3.4 Standardized Residuals Choropleth Map

Referencing the maps of the dependent variable and the predictors presented earlier in Graph 2.2, there seems to be spatial autocorrelation in the variables as it seems that the block group observations are not independent of each other. There are many high-high house value and high-high percentages of individuals with at least a bachelor's degree in Center City and in the Wissahickon Valley Park area toward northwest Philadelphia on the east side of the Schuylkill River. The percentage of detached homes had high-high value clusters in the Wissahickon Valley Park area as well, in addition to the most northeastern part of the county, which is quite in contrast to the low-low values for individuals living in poverty in that same region. Lastly, the percentage of vacancies seemed to have high-high value clusters in North Philadelphia.

Perhaps not as stark as the other maps in Graph 2.2, but when observing Graph 3.4, the most distinct portion of the choropleth map is in the Wissahickon Valley Park region and there are low-low value clusters in North Philadelphia as well as South Philadelphia, so this choropleth map still violates the assumption that the block groups are independent of one another.

Additional Models



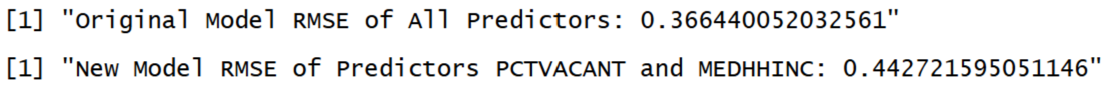
```
Start:  AIC=-3448.07
LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNNBELPOV100

               Df Sum of Sq    RSS     AIC
<none>                      230.34 -3448.1
- PCTSINGLES    1     2.407 232.75 -3432.2
- LNNBELPOV100  1    11.692 242.04 -3364.9
- PCTVACANT     1    51.546 281.89 -3102.7
- PCTBACHMOR    1   199.020 429.36 -2379.0
```

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
<S3: AsIs>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
	NA	NA	1715	230.3435	-3448.073

1 row

Figure 3.1 Stepwise Regression Model and ANOVA Output

Referring to figure 3.1, the stepwise regression determined that all four predictor variables were worth keeping because the <none> value, indicating that when none of the predictor variables are dropped, the AIC value was the lowest of all the predictors at –3448.073.



```
[1] "Original Model RMSE of All Predictors: 0.366440052032561"
[1] "New Model RMSE of Predictors PCTVACANT and MEDHHINC: 0.442721595051146"
```

Figure 3.2 RMSE Comparison of Original Model and New Model

The RMSE is the estimate of a typical residual's magnitude, and ideally the lowest value is best when comparing different models, and the RMSE value will always decrease with the addition of predictor variables. According to Figure 3.2, the original model with all four predictors is 0.3664 and the new model with percent vacancies (PCTVACANT) and median household income (MEDHHINC) is 0.4427, because the original model's RMSE is lower than the new model's RMSE, the former is the preferred model.

Discussion and Limitations

This paper conducted a multiple linear regression with y the natural log of the median house value (LNMEDHVAL) as the dependent, or response, variable and the following as the independent, or predictor, variables: β_1 percent vacancies (PCTVACANT), β_2 percent of detached single-family homes (PCTSINGLES), β_3 percent of residents with at least a bachelor's degree or more (PCTBACHMOR), and β_4 the natural log of the number of residents living in poverty (LNNBELPOV100). Concerning variable limitations, almost all predictors were measured in proportions, whereas the original NBELPOV100 measured raw count of individuals living in poverty, which would be a difficult statistic to compare to other block groups; the aggregation of fifty individuals in one small block group versus fifty individuals in a much larger block group is difficult to compare, but that small block group might only have 5% of residents living in poverty versus 50% in a larger block group.

Before determining the above variables to conduct multiple linear regression on, exploratory data analysis was performed to check for assumptions. First, the summary statistics were observed, namely, the mean and the standard deviation for all variables, where the standard deviations were almost as high or even higher than their associated means' values, which indicated high fluctuations in the observations. Second, histograms were plotted for all variables in their raw form and their log-transformed form to observe for normal distribution—this is where the variables were solidified for the regression model because while all of the variables were highly skewed right, only two log-transformations were maintained on median household value and on the number of individuals living in poverty. This is because when the other variables were log-transformed, they had very large frequencies at zero, which made them inappropriate for regression modeling. Third, choropleth maps were created for the final variables to visually observe for spatial autocorrelation and multicollinearity along with a correlation matrix, and it was determined that while the predictor maps' spatial patterns seemed inverted in certain areas, that multicollinearity was negligible. However, the maps had several high-high and low-low clusters, and the presence of spatial autocorrelation violated the assumption that the observations and residuals are independent. Fourth, a scatter plot and histogram of the standardized residuals were made, the former of which had a normal shape and the latter of which was homoscedastic, but the scatter plots of the predictor variables and the response variable presented a clearly non-linear shape. This violated the assumption that the response and given predictor variable relationship is linear, meaning that a linear regression model is not ideal to represent these variables' relationships, and could be rectified by using polynomial terms to explain the scatter plots' curvatures.

Despite some assumption violations described, this model was strong. The F-Ratio was 840.9 with a p-value less than 0.001, which was significant, indicating at least one of the predictors had a significant relationship with the response variable, and this was further confirmed with the t-tests being also having a very small p-value when testing the relationships between the individual predictors and the response. The adjusted R^2 of the model indicated that the independent variables explained 66.15% of the variance in the dependent variable. Then, when running stepwise regression, the results showed that the final model kept all predictors in as in the original model because it had the lowest AIC value at -3448.073, and comparing its RMSE to another model with just PCTVACANT and MEDHHINC as predictors, the original had the lowest value. While these indicate the model is strong, it could potentially be improved by adding crime rates, or spatial predictors like distances to parks or other nearby amenities.

Considering the use of Ridge or LASSO regression, neither are ideal for this situation as they're both regularization regressions that seek to reduce overfitting, prioritizing prediction by introducing a penalty term to stabilize a model (Murel & Kavlakoglu, n.d.). While both regressions shrink the models with their penalties, Ridge keeps all predictors because it reduces their coefficients to near-zero and LASSO can remove some predictors it deems irrelevant to the response because it reduces their coefficients to exactly zero, the latter regression thereby acting as a predictor selector and simplifying models—this is because Ridge's penalty is the squared sum of coefficients and LASSO's penalty is the absolute value of the sum of coefficients (Murel & Kavlakoglu, n.d.). Their use-cases differ in this regard; Ridge's performance excels when there are many predictors with roughly equivalent coefficients, and LASSO's performance excels when more predictors have very insignificant or zero coefficients. In this case, it makes no sense to implement these regressions, the current model is stable as all predictors are statistically significant to the response, there's no multicollinearity, no heteroscedasticity, and the cross-validated out-sample RMSE 0.3664 is very close to the in-sample residual standard error 0.3665, so the model generalizes well already and isn't overfit.

Citations

Bieretz, B., & Schilling, J. (2019, July). *Pay for success and blighted properties: Insights and opportunities for funding vacant property reclamation and neighborhood stabilization.* Urban Institute. https://www.urban.org/sites/default/files/publication/100464/pfs_and_blighted_properties_0.pdf

Murel, J., & Kavlakoglu, E. (n.d.). What is regularization? https://www.ibm.com/think/topics/regularization

Wang, H., Cheng, Z., Smyth, R., Sun, G., Li, J., & Wang, W. (2022). *University education, homeownership and housing wealth. China Economic Review, 71*, 101742. https://doi.org/10.1016/j.chieco.2021.101742