Zhenzhao Xu, Hanpu Yao, and Ben Aiken
Statistical and Data Mining Methods - MUSA 500
Professor Eugene Brusilovskiy
Assignment 1

Using OLS Regression to Predict Median House Values in Philadelphia

Introduction

This investigation compares the median house value of Philadelphia census tracts with several variables including: education levels, as measured by the percent of residents that hold at least a bachelor's degree, vacancy percentage, single family homes, poverty levels and median household income. This relationship is explored by creating a predictive model using OLS regression. In order to create the most accurate model, variables are transformed, correlations are determined to detect multicollinearity, and a k-fold cross validation is completed.

The predictor variables are all closely related to median house value. Household income and poverty are likely strong predictors of house value since an individual or family's financial situation determines what house they can afford. Furthermore, education level is likely a worthwhile predictor since there is significant research that education level and income are positively related.[1] Vacancy of surrounding lots contributes to the house value as density is a priority to home buyers and often an indicator of long-term value.[2] While these predictors are certainly not exhaustive, together they may create a relevant predictive model.

Methods

a) Data Cleaning

In order to remove outliers and block groups with few residents, the original dataset, which contained 1816 observations was edited by removing block groups using the following criteria:

1) Block groups with population less than 40
2) Block groups without any housing units
3) Block groups with exceptionally low median house value: less than $10,000

Additionally, there was a single block group in North Philadelphia with a high median house value of over $800,000, yet a low median household income of less than $8,000. This block group was not included in the analysis. After removing these observations, the data set had 1720 block groups total.

[1] Orley Ashenfelter and Cecilia Rouse, "Schooling, Intelligence, and Income in America: Cracks in the Bell Curve," NBER (National Bureau of Economic Research, January 1, 1999), https://www.nber.org/papers/w6902.
[2] Whitaker, Stephan & Fitzpatrick, Thomas. (2011). The Impact of Vacant, Tax-Delinquent, and Foreclosed Property on Sales Prices of Neighboring Homes.

b) Exploratory Data Analysis

To understand the variables, summary statistics including the mean and standard deviation were examined. Histograms were also created to understand the distribution, potential skews, and possible transformations of each variable.

Additionally, the correlations between predictors were investigated. Correlation is any statistical relationship between two variables and shows the degree and direction to which a pair of variables are related. One of the most common measures of correlation is *Pearson's correlation coefficient,* which was used for this investigation.

When the standard deviation of the two variables is not zero, the correlation coefficient is defined. If a linear relationship between the two variables exists, the correlation coefficient ranges from -1 to 1. When both variables increase simultaneously, there is a positive correlation, and the correlation coefficient is greater than zero. When one variable increases and another variable decreases, there is negative correlation, and the correlation coefficient is less than zero. When the two variables are linearly independent, the correlation coefficient is 0.

Following formula is used to determine the correlation coefficient r between two random variables X and Y:

$$r = \frac{cov(X,Y)}{\sigma_X, \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X, \sigma_Y} \quad (1)$$

Commented [EB1]: But what's the formula for the sample correlation coefficient?

In this equation, E is the expected value operator, μ is the expected value of X or Y, σ is the standard deviation of X or Y, and cov means covariance.

c) Multiple Regression Analysis

The method of regression used in this investigation is Ordinary Least Squares (OLS) regression. OLS regression is a statistical method used to examine the relationship between a variable of interest (dependent variable) and one or more explanatory variables (predictors). It enables us to calculate the amount by which the dependent variable changes when a predictor variable changes. In this regression, there are four predictors, so it is also a multiple regression.

Our regression models LNMEDHVAL regressed on PCTVACANT, PCTSINGLES, PCTBACHMOR and LNNBELPOV100 using the equation below:

$$LNMEDHVAL = \beta_0 + \beta_1 PCTVACANT + \beta_2 PCTSINGLES + \beta_3 PCTBACHMOR + \beta_4 LNBELOWPOV100 + \varepsilon \quad (2)$$

In this equation, β_i is the coefficient variables of x_i, (for example, $\beta_1 PCTVACANT$) and it shows the true average change in the dependent variable y ($LNMEDHVAL$) associated with a 1-unit

increase in the predictor x_i (holding all other variables constant). β_0 is the y-intercept, which means that it is the value of the dependent variable when all predictors are 0. β_i changes in different samples, so it is a random variable. The variable ε in the model is referred to as the residual, which is the difference between y_i and $\hat{y}_i$, here $\hat{y}_i$ refers to the estimated value of y_i.

There are six regression assumptions in this model:

1. The relationship between dependent variable, y, and every predictor, x_i, is linear.
2. Residuals, ε, are random and normally distributed.
3. The variance of the residuals is homoscedastic - constant regardless of the values of each x.
4. Observations are independent - no spatial, temporal, or other forms of dependence in the data.
5. No multicollinearity predictor variables. Specifically, correlation between any pair of predictors is less than 0.8.
6. More than 10 observations are used per predictor.

Commented [EB2]: And have a mean of 0

Commented [EB3]: Awkward phrasing: residuals are homoscedastic, meaning that their variance is constant. Saying that the variance is homoscedastic is a bit off.

There are several parameters that need to be estimated in the regression including σ^2 and β_0 through β_k. σ^2 is the variance of residuals ε. The estimator of σ^2 is s^2, and can be calculated in the equation below:

Commented [EB4]: +1. Excellent!

$$\sigma^2 = \ s^2 = \frac{SSE}{n-k+1} \tag{3}$$

Here, k is the number of predictors and n is the number of observations. Additionally, the estimators of β_i is $\hat{\beta}_i$, and they can be chosen simultaneously to minimize the expression for the Error Sum of Squares (SSE), which can be calculated in the equation below.

$$SSE = \sum_{i=1}^{n} \varepsilon^2 = \sum_{i=1}^{n} (y - \hat{y})^2 \tag{4}$$

There are two tests that will be used to check the validity of the model: the utility test (F-test) for all independent variables and hypothesis test (T-test) for each variable. Usually, a F-test is conducted before the T-test, and if the T-test is significant then the F-test is completed as well. The F-test is completed using the following formulas:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \tag{5}$$

$$H_1: \beta_1 \neq 0 \ or \ \beta_2 \neq 0 \ldots or \ \beta_k \neq 0 \tag{6}$$

H_0: All coefficients in the model are (jointly) zero. H_1: At least one of the coefficients is not zero. This determines if there is at least one significant predictor in the model.

On the other hand, the T-test, examines the p-value associated with each independent variable x_i.

$$H_0: \beta_i = 0,\ x_i \text{ is not a significant predictor of } y.$$
$$H_1: \beta_i \neq 0,\ x_i \text{ is a significant predictor of } y.$$

If the p-value for $\beta_i = 0$ is less than 0.05, the null hypothesis is rejected, determining that x_i is a significant predictor of y. When the null hypothesis cannot be rejected, the dependent variable is not related to the independent variable.

Coefficient of multiple determination, R^2, is the proportion of observed variation in the dependent variable y that was explained by the model. R^2 can be calculated by:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} \tag{7}$$

$$SSE = \sum_{i=1}^{n} \varepsilon^2 = \sum_{i=1}^{n} (y - \hat{y})^2 \tag{8}$$

$$SST = \sum_{i=1}^{n} (y - \bar{y})^2 \tag{9}$$

$$SSR = SST - SSE \tag{10}$$

Here, SSE is the error sum of squares, which is the sum of squared residuals. SST is the total sum of squares, which is the sum of squared deviations about the sample mean of the observed y values. SSR is the regression sum of squares, and it equals SST minus SSE. Because extra predictors will generally increase R^2, it is typically adjusted as ${R^2}_{adj}$ using the equation below:

$${R^2}_{adj} = \frac{(n-1)R^2 - k}{n - (k+1)} \tag{11}$$

d) Additional Analyses

Stepwise regression is a method to automatically select predictors based on the partial F-tests. Stepwise regression has several disadvantages: [3]

1. The final model may not be the best model among others.
2. There may be other possible models other than the final model.
3. Domain knowledge is not considered during the process of stepwise regression.

[3] 10.2 - Stepwise Regression | STAT 501". 2021. *Pennstate: Statistics Online Courses*. https://online.stat.psu.edu/stat501/lesson/10/10.2

4. The order of the variables in stepwise regression may not make sense.
5. There may be Type I or Type II error in stepwise regression.
6. Many t-tests are conducted, so the variables in stepwise regression may not be the most important ones.

K-fold cross-validation is used when the sample size is limited such that it cannot be tested on a separate testing set. It determines the generalizability of the model by testing new data. The procedure of a K-fold cross-validation is: first, partition the dataset into K equal part. Next, for each part, use the remaining K-1 parts to train the model and test it by calculating the MSE on each part. Then the average MSE and $RMSE$ are calculated of the whole model. In practice, k is usually set to 5 or 10, and for this investigation k=5 in the cross-validation.

$RMSE$ stands for root mean square error, which is an estimate of the magnitude of a typical residual. It is often used to compare different models. The model that has the smaller $RMSE$ is the better model.

$$RMSE = \sqrt{\overline{MSE}} = \sqrt{\frac{\sum_{i=1}^{n}(y - y)^2}{n}} \quad (12)$$

The R program was used for data analysis and ArcGIS was used to create the maps.

Results

a) Exploratory Results

The mean and standard deviation for the dependent variable, Median House Value, and the predictors, number of households living in poverty, percent of individuals with a bachelor's degree or higher, percent of vacant houses, and percent of single house units for the Philadelphia Census Tracts were calculated and displayed in the table below. In all cases, the variables have large standard deviations. This is hardly surprising given the size of Philadelphia and the variability of wealth, education, and vacancy; however, the large standard deviation indicates that the data is spread out, rather than clustered around the mean values.

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

The histograms below (Figure 1) show the distribution of the predictors without any transformation. None of the predictors are normally distributed. Instead, they are positively skewed. With that in mind, a log transformation will be applied to attempt to achieve normality.

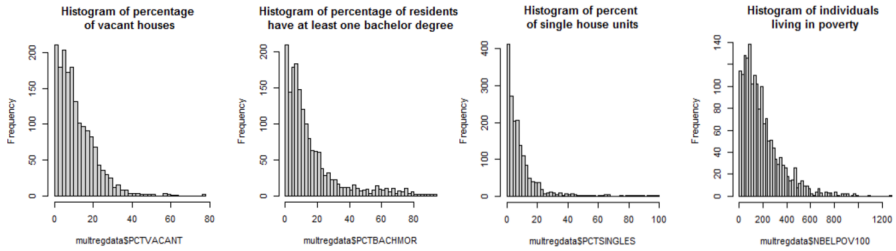

Figure 1 – Predictor Distributions

After a log transformation, displayed in the histograms below (Figure 2), it is apparent that all four of the transformed predictors have values of zero. While they look more normally distributed after the transformation, with values of zero, the log data cannot be used in the regression model.

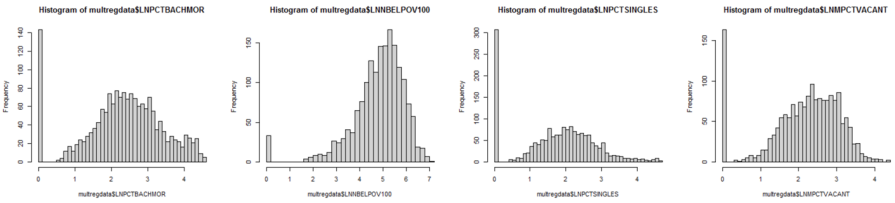

Figure 2 – Log Transformed Predictors

Figure 3 below contains a histogram of the dependent variable, Median House Value, on the left. Since the data is positively skewed, not normally distributed, a log transformation was applied, and a histogram of this transformed data is on the right. Unlike the predictors, this histogram looks normal after transformation, and does not contain any 0 values. Therefore, the Log of Median House Value will be used in the regression model.

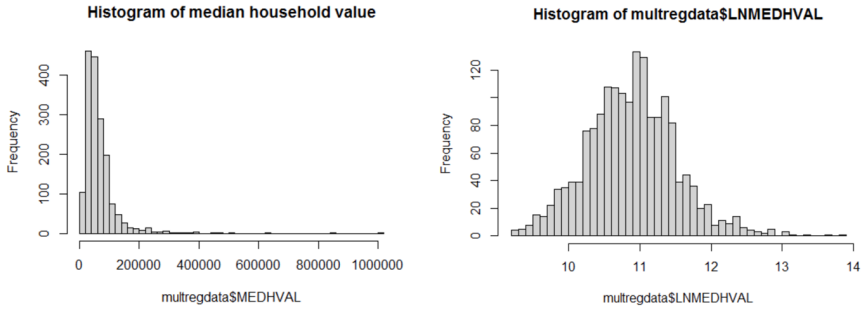

Figure 3 – Dependent Variable Histograms, log transformed on right

Although the predictors are not normal, the assumption that the residuals will be normal is investigated and discussed in the Regression Assumption Checks section of this report along with the other regression assumptions.

Figure 4 displays the predictors across Philadelphia with a set of choropleth maps. While there are some similarities, the clusters in the city are not consistent on different maps. For example, in both the PCTSINGLES and PCTMACHMOR maps, there are clusters of high percentages in Northeast Philadelphia, but low percentages in West Philadelphia. Given how similar these two predictors are, there is a chance that they are inter-correlated and could cause multicollinearity issues.

On the other hand, PCTVACANT is close to zero for much of Northeast Philadelphia, but much more clustered in West Philadelphia. This suggests, that in some parts of the city, the percentage of vacant homes are negatively correlated with high education and single-family homes. Furthermore, the LNNBELPOV is visually different from the other three maps as it has clusters throughout the city, making it hard to recognize clear spatial trends.

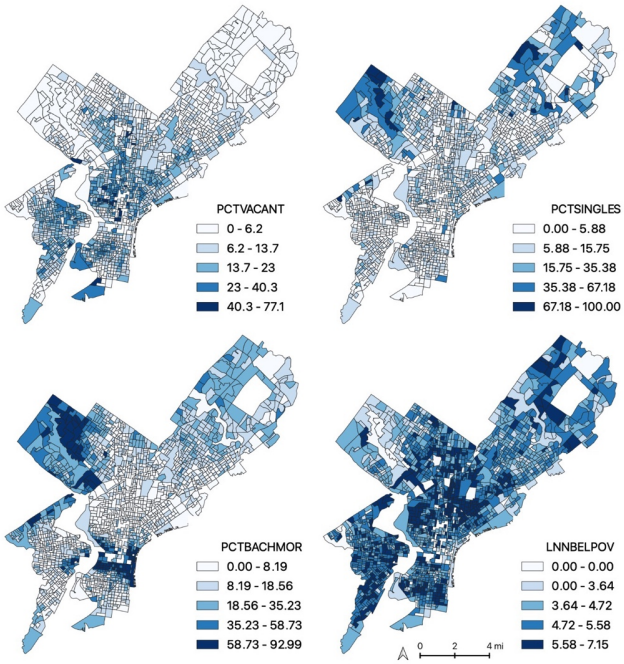

Figure 4 – Maps of Predictor Variables

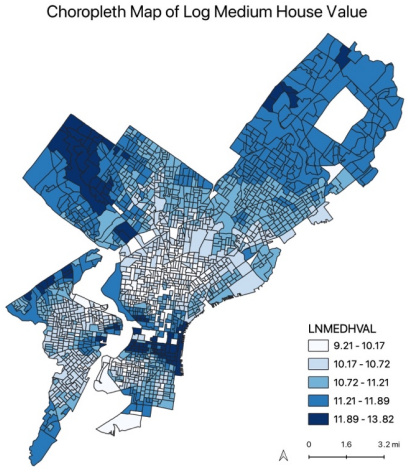

Figure 5 –Log of Median House Value

The map above (Figure 5) shows the dependent variable, Log of Median House Value. Compared with the predictor maps, this map has larger clusters and more consistent values. The dependent variable is most like PCTBACHMORE and PCTSINGLES, with high values located in the Northeast and Northwest parts of Philadelphia and lower values in West Philadelphia. This confirms some predictions stated in the introduction that higher education and single-family homes are positively correlated with house value. The dependent variable appears to be inversely related, or negatively correlated with PCTVACANT, especially in neighborhoods just north of Center City. Finally, LNNVELPOV and LNMEDHVAL do not exhibit clear negative or positive correlation.

	PCTVACANT	PCTSINGLES	PCTBACHMOR	LNNBELPOV100
PCTVACANT	1.000	-0.151	-0.298	0.250
PCTSINGLES	-0.151	1.000	0.198	-0.291
PCTBACHMOR	-0.298	0.198	1.000	-0.320
LNNBELPOV100	0.250	-0.291	-0.320	1.000

Figure 6 – Correlation Matrix

Figure 6 above shows the correlation of the predictors. The strongest correlation exists between LNNBELPOV100 and PCTBACHMOR, which is negative at -0.320. This was not apparent by simply looking at the map, but it is consistent with the prediction that higher education is related to higher income and therefore related to poverty. PCTBACHMOR and PCTSINGLES have a correlation of just 0.198, which is positive, but not a terribly strong correlation despite what the maps suggested. None of these correlations are significant enough to suggest an issue with multicollinearity.

Commented [EB5]: Don't use the term significant unless you test for statistical significance. The proper term to use is substantial.

b) Regression Results

Figures 7 and 8 below show the regression output from R sum of squares.

```
Call:
lm(formula = LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR +
    LNNBELPOV100, 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 ***
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 7 – R Summary of Regression

```
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 ***
LNNBELPOV100    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
```

Figure 8 – Regression and Sum of Squares

First, the F-statistic's p-value is less than 0.05 revealing that the null hypothesis of the F-Test is rejected. This indicates that at least one of the independent variables is a significant predictor. Next, the p-values of each independent variable are examined and since all variables have p-values less than 0.05 then the null hypothesis that each predictor is not significant can be rejected – the predictors are in fact significant. In fact, all four of them have p-values less than 0.0001 making them highly significant.

Given the negative coefficients of PCTVACANT and LNNBELPOV100, they are negatively associated with LNMEDHVAL. PCTSINGLES and PCTBACHMOR, on the other hand, are positively associated given their positive coefficients.

The log transformations of the dependent variable and one of the predictors make the interpretation of the coefficients more complicated than a one-unit increase. The first three predictors were not log transformed and are below 0.3 in absolute value. As the percentage of vacant homes increases by 1% the median house value changes by $100\beta\% = 100 * (-0.0192)\% = -1.9\%$, while holding the other predictors constant. Also, as the percentage of single-family homes increases by 1%, the median house value changes by 0.29%, while holding the other predictors constant. And finally, as the percentage of residents with at least a bachelor's degree increases by 1%, the median house value changes by 2.09%, while holding the other predictors constant.

Since NBELPOV100 was log transformed, the coefficient indicates that the number of households with incomes below the poverty level changes by 1%, the median house value changes by $(1.01^{-0.079} - 1) * 100\%$ or -0.078%, holding all other variables constant.

A great deal of the variance of the model, is explained. The R^2 value, or the coefficient of determination, is 0.6623, and the adjusted R^2, or ${R^2}_{adj}$, is 0.6615.

c) Regression Assumption Checks

This section is about testing model assumptions and aptness.

Assumption 1, linear relationship between dependent variable y and every predictor x_i, is violated. Below (Figure 9) are scatter plots of the dependent variable and each of the predictors. Though we assumed that every predictor x has a linear relationship with the dependent variable y, they are not linear.

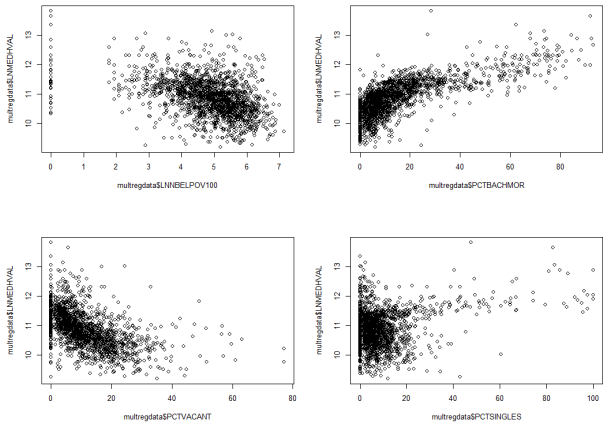

Figure 9 – Scatter Plots of LNMEDHVAL and Predictors

Assumption 2, residuals being random and normally distributed, is basically valid. *Figure 10* shows that residuals have a slightly positive-skewed distribution, which implies that there are many places that have been under-estimated by the model.

Commented [EB6]: -1. This check is for normality, not for randomness. Randomness implies that there's a lack of any systematic pattern in residuals. The assumption of randomness is violated if, for example, there is heteroscedasticity, spatial autocorrelation of residuals, or polynomial trends in the residual-by-predicted plot.

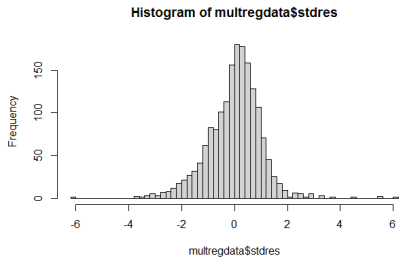

Figure 10 – Histogram of Residuals

Assumption 3, homoscedastic of variance of the residuals, is violated. The standardized residual is the residual divided by its standard deviation. While the residual measures the deviation of the observed values of an element from its predicted values, its unit is the same as

the independent variable. As a result, the standardization of residuals makes it unitless and easier to interpret.

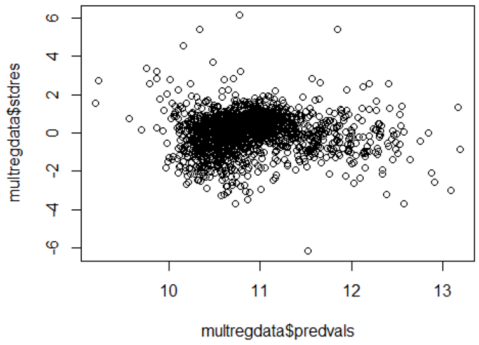

Figure 11 – Standard Residuals by Predicted Value

Figure 11 shows the relationship between standardized residual (on y-axis) and the predicted value (on x-axis). Though standardized residuals are around 0, it suggests that the relationship between the median house value and the predictors is not very linear. Note that the residuals depart from 0 in a systematic manner. The residuals are negative for small x values, positive for medium y values, and negative again for large x values. Clearly, this is heteroscedasticity, and a non-linear model would better describe the relationship among the variables.

Commented [EB7]: Basically, there may be a quadratic trend.

Also, there are many outliers. The y values in Figure 11 show how many standard deviations are above or below 0. Many outliers are 4-6 standard deviations away from 0. For data that are normally distributed, 95% of the measurements fall within 2 standard deviations of the mean.

Assumption 4, independence of observations, is violated.
Referring to Figure 4, there seems to be positive spatial autocorrelation in these predictors: % of housing units that are vacant (PCTVACANT), % of housing units that are detached single family houses (PCTSINGLES) and % of residents in block groups with at least a bachelor's degree (PCTBACHMORE). These 3 variables have some spatial clustering pattern in some certain areas. For example, the map of PCTSINGLES shows that the Northeast and Northwest parts of the city have high value clusters while the rest of the city is low. However, log transformation of the number of households with incomes below 100% poverty level (LNNBELPOV100) doesn't show obvious spatial autocorrelation. Compared to other variables, its values are more random spatially.
Referring to Figure 5, the dependent variable, log of median value of all owners occupied housing units (LNMEDHVAL) also shows a strong positive spatial autocorrelation. The spatial pattern is that the high value clustering in the northwest and central part of the city, and within the clusters the highest value in the core diffuses to its neighbors.

Figure 12 shows there is obvious spatial autocorrelation in the standardized residuals. The spatial pattern of studentized residuals is also clustering.

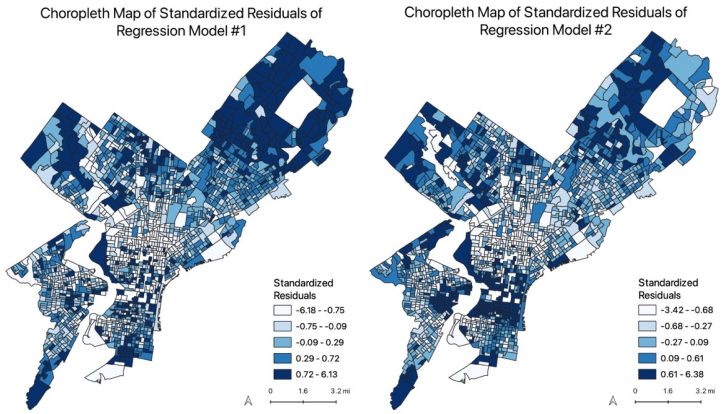

Figure 12 –Maps of Standardized Residuals of Regression Models

Assumption 5, no multicollinearity predictor variables is valid. Referring to Figure 6, the Correlation Matrix, the correlations of variables are all below 0.4, which is little multicollinearity.

Assumption 6 is valid. There are more than 10 observations per predictor.

Commented [EB8]: Holds (or is met).

d) Additional Models

In addition to the Ordinary Least Squares Regression model, a Stepwise Regression was completed in R. The results are displayed below in Figure 13. The stepwise regression model kept all four predictors of the original model.

```
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
```

Figure 13 – Stepwise Regression Results

Additionally, a k-fold cross validation was completed of the original model resulting in an $RMSE$ value of 0.366. Another model was created only using PCTVACANT and MEDHHINC as predictors. This model was once again put through a k-fold cross validation which produced an $RMSE$ value of 0.443. This suggests that the model with all four predictors is better and has less error than the model with just the vacancy and income variables.

Discussion and Limitations

In this report, we built a regression model using OLS regression to predict median house values with several variables. This relationship is explored by creating a predictive model using OLS regression.

In terms of each variable, the percentage of singles is not as important as others. Although the p-value of 0.00002 shows that it is significant, the coefficient of the percentage of singles is only 0.003, which means that as it goes up by 1%, the value of median house value goes up by only 0.29%, so it has little impact on the result. This is not surprising since the percentage of singles and median house value don't have a clear relationship. Other predictors are both important and significant.

The overall quality of the model is not strong. Although it is significant in the F-ratio test, R^2 is only 0.6615, which means that only 66% of variance in the model are explained by all 4 predictors. One reason may be that we didn't include enough related predictors. Additional predictors that could improve the model include the type of zoning code, average built year of houses, transportation coverage rate, amenities (parks, schools, hospitals) and crime rate.

Commented [EB9]: That's a pretty good R-square in the social sciences.

The final model of the stepwise regression includes all 4 predictors, which implies that we didn't include enough predictors for the stepwise regression, so that it cannot exclude any of them, and it also shows the high possibility that all 4 predictors are significant.

It is not surprising that the $RMSE$ is better for the 4-predictor model. The $RMSE$ values are listed below. This implies that the percentage of singles, the percentage of bachelor, the percentage of vacant and number of poverty variables have greater impact on the overall quality of the model than medium house income and the percentage of vacant alone.

$$RMSE_{4-predictor} = 0.366, RMSE_{2-predictor} = 0.443$$

There are several limitations of the model. First, the assumption of OLS that linear relationship between dependent variable and predictors is violated, which means that there is no clear linear relationship between them, and it will have a negative effect the model significance and accuracy. Second, the homoscedastic of residuals' variance is violated, which implies that there is systematic under or over- prediction happening in the model. Adding predictors or using spital regression may help to solve this problem. Third, the independence of observations is violated. There are clear clustering patterns in predictors, dependent variable, and residuals,

Commented [EB10]: homoscedasticity

which means that running a spatial regression like SE, SL or GWR instead of OLS regression is a better choice.

Another limitation is the spatial nonuniformity of predictor. In the model, predictor NBELPOV100 is biased, because the population of Philadelphia is not evenly distributed in space, and it is clustered in this case, so that the number of the poverty may also be clustered. Thus, the predictor NBELPOV100 cannot represent the true poverty level of an area. The problem can be solved if the percentage of individuals living in poverty is used instead.

Commented [EB11]: ?

If you have two block groups where there are 10 people living in poverty, and the total population of one block group is 20 and the total population of the other block group is 1000, then we have 50% vs. 1% of people living in poverty. So percentages may be better here.