Homework 1:

Using OLS Regression to Predict Median House Values in Philadelphia

Yu Wang, Yuhao Jia, Zile Wu

1. Introduction

As an important part of the social economy, housing prices reflect the development of society. Therefore, the forecast of house prices can help people better judge the development of the economy. In this article, based on the Proportion of residents in Block Group with at least a bachelor degree; Proportion of housing units that are vacant; Percent of housing units that are detached single family houses and Number of households with incomes below 100% poverty level, we are going to predict the Median value of all owner-occupied housing units.

Based on past experience with the U.S. housing situation. Highly educated people and low-density residential areas tend to have higher housing prices, while areas with a high concentration of poor people, and areas with high housing vacancy rates, tend to have lower housing prices due to lack of maintenance of infrastructure and poor security. Therefore, we choose the above four data as our predictor.

2. Methods

a) Data Cleaning

The Original Philadelphia block group consists of 1816 observations in total, we removed 96 observations that don’t fit the qualifications as suitable data for our model. We removed these observations based on the following standards, whether the block group population is greater than 40, whether there are housing units associated with the block group, median house value less than $10,000, and two outlier groups that one block has extreme high median house value, $800,000 and a very low median household income group with less than $8,000. Therefore, the final dataset that we used to fit our model consists of 1720 observations.

b) Exploratory Data Analysis

We will complete summary statistics check for each of the independent variables and dependent variable to have a basic idea of the average value and deviation of each variable, and this information will help us to see the distribution of each variable better. It is important to check the distributions of each variable that be contained in our regression model since the default distribution of each parameter should be normally distributed. If there is any parameter that has a skewed distribution with a spike near 0 and a tail on the right of the x-axis, then we will need to convert the non-normally distributed variable back to a normally distributed variable. However, there is also the case when simply applying log transformation to the variable does not help, we might need to stick with the non-normally distributed variables.

Also state that as part of your exploratory data analysis, you will examine the correlations between the predictors. Explain what a correlation is, and provide the formula for the sample correlation coefficient r. Also mention the possible range of r values, and what correlation of 0 means.

One useful measurement for us to check the presence of multicollinearity in our model is checking the correlation between each predictor. Correlation is a unitless measurement that quantifies the relationship between two variables. The equation that will be used to calculate correlation is:

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

And the possible results of the correlation are between -1 to 1. A result equals to -1 means the two variables are perfectly negatively related, and vice versa for a result that equals to 1, we can also have two unrelated variables if we have a result that equals to 0. In Multiple regression, we should avoid multicollinearity from happening, thus checking the correlation coefficient between each potential predictors that we will be using in our model is critical. Multicollinearity exists when there is more than one predictor that are strongly correlated with each other, consequently, when one predictor changes, the other correlated predictor will also be affected.

c) Multiple Regression Analysis

In this project, OLS regression is used, which is a statistical method used for finding the beta coefficients for each of the independent variables in the regression model. OLS stands for Ordinary Least Squares, and it is efficient in minimizing the sum of square differences between the observed data and predicted value.

The OLS regression we will produce in the next step is going to be in the form as

$$\begin{aligned} LNMEDHVAL &= \beta_0 + \beta_1 \times PCTVANT + \beta_2 \times PCTSINGLES \\ &\quad + \beta_3 \times PCTBACHMOR + \beta_4 \times LNNBELPOV100 + \mathcal{E} \end{aligned}$$

Once we have the output of the model, based on the output of the regression model, we can see the relationships between the dependent variable (LNMEDHVAL) and the independent variables (LNNBELPOV100, PCBACHMORE, PCTVACANT, PCTSINGLES), then get a meaningful result by interpreting the effect of each independent variable on LNMEDHVAL in the content of Philadelphia neighborhood characteristics. β_1, β_2, β_3, and β_4 are the coefficients of each predictor. β_0 is the predicted LNMEDHVAL value when all the predictors are equal to 0. Ɛ is the difference between the actual LNMEDHVAL value and the predicted value.

When fitting the model, the model is constructed based on the following assumptions, linear relationship between LNMEDHVAL and the independent predictors, the homoscedastic residuals, Ɛ are random and normally distributed.

The parameter that will be estimated are β_0, β_1, β_2, β_3, and β_4, and once we find the values of these parameters, these parameters will provide insightful information about their relationships with the dependent value. We can use these parameters to interpret how will the dependent value change if we increase 1 unit of a specific parameter while holding the other parameters constant. Besides the parameters of each predictor, we are also interested in the variance of the model since it determines the deviation of the dependent values when we take the square-root of the variance. As a result, we can find out an expectation of how much deviation that the dependent values might vary from the regression line predicted by the model.

To get the predicted parameters of our independent variables, the process of minimizing the sum of squared vertical distance is required. Basically, we are trying to predict a linear trend line that will return the smallest errors, the residual is calculate as this:

$$\mathcal{E}_i = y_i - \hat{y}_i$$

y_i is the actual LNMEDHVAL value and $\hat{y}_i$ is the predicted value.

The equation used to find the independent variable parameters as a matrix B is

$$B = (X'X)^{-1}X'Y$$

In this equation, $X'X$ creates the total value of the squared X, and $X'Y$ creates the total value of the product from X and Y.

When we fit the regress model, the model output will show the values of R^2 and adjusted R^2. The equation used to calculate R^2 is

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

The equation can be expanded to the form:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$$

The SSE in the equation is the sum of all the squared errors between the actual LNMEDHVAL value and the predicted LNMEDHVAL value for each observation, and SST is the sum of squares total, it is the sum of errors between the actual LNMEDHVAL value and the average value of LNMEDHVAL. By using the equation mentioned above, a R^2 value of the model is produced, and it is a measurement of how much variance in the regression is represented by the four independent variables that we used in the regression. Adjusted R^2 is similar to R^2, but it will decrease as more insignificant independent variables are included in the regression as a punishment. The value for Adjusted R^2 is calculated from the equation:

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

If we plug in the total number of observations (n), and the number of independent variables (k) into the equation, the Adjusted R2value can be calculated as

$$R^2_{Adj} = \frac{(1720-1)R^2 - 4}{1720-(4+1)}$$

Now we have the regression model that can predict the LNMEDHVAL in Philadelphia, a F-ratio test will be needed to test the Null hypothesis and Alternative hypothesis. The Null hypothesis states that all the independent predictors in our regression are not significant in predicting LNMEDHVAL, H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, and on the other hand, the Alternative hypothesis states there is at least one significant variable among the four that is useful in predicting the LNMEDHVAL. If we complete the F-ratio test and got a result that rejects the Null, we can proceed further to complete a t-test on each individual variable, and the Null hypothesis is that the variable being tested is not significant, H_0: $\beta_i = 0$, and the Alternativa hypothesis is H_A: $\beta_i \neq 0$. Moreover, the purpose of completing the F-ratio test and t-tests on our regression is proving the model that we built is meaningful and effective in predicting.

d) Additional Analyses

Stepwise regression is a method that we can use to select necessary predictors to include in our model by evaluating the p-values of t-tests and Akaike Information Criterion (AIC). The final model returned by stepwise regression will include predictors that reject the null hypothesis of t-test only, which each of them should have a p-value less than 0.1 in the t- test to be qualified. However, we should be cautious with the final regression returned because there are limitations of this method. The major issue is that we can’t conclude that all the significant predictors are captured by stepwise regression and all the irrelevant predictors are omitted without additional tests to support the results from stepwise regression. The first reason is that when t-tests are being conducted for each predictor, there is a high chance that unwanted factors might be kept, or critical factors being excluded. Moreover, stepwise regression doesn’t know the significance of specific factors especially when fitting a regression model for read world problems, since some factors conveys special meanings that corresponded to the context of the issue that we are trying to predict by regression model.

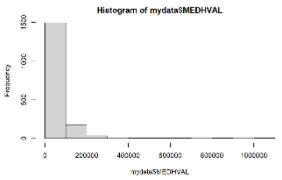
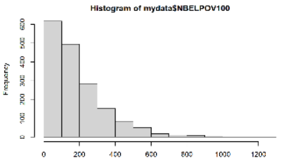
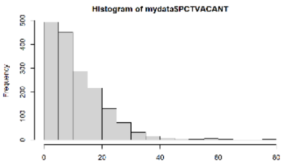
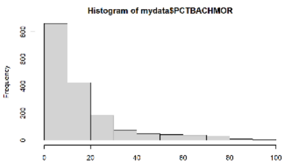
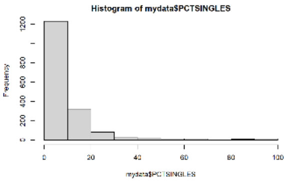
Another important test we will be using to test our model is the k-fold cross-validation test, and the squared rooted mean squared error returned by the test will be a critical aspect that we evaluate our regression model. In this assignment, we chose k to be equal to 5, which means the data will be randomly divided into 5 groups. The test will be repeated 5 times, and each time one group will be picked as the validation set while the remaining models will be used for fitting the model. After each round, a squared mean error will be calculated, then we will take the average of the five squared mean errors and take the squared root of the average (RMSE) to derive the result. Among different models, the model with lowest RMSE is going to be the one we are looking for.

In the next step, R will be used for calculating for AIC and RMSE of our regression, and ArcGIS will be used to create map layers to help visualize the depend and independent variables.

3. Results

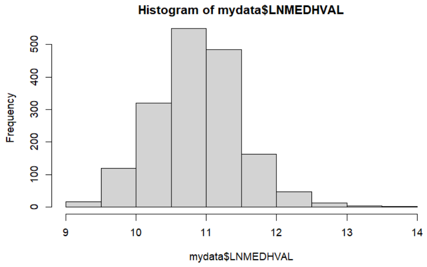
a) Exploratory Results

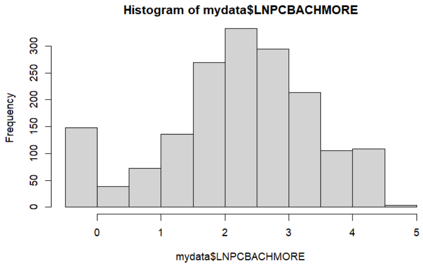
Below is the summary statistic of the data. The dependent variable *Median House Value* has a mean $66,287.73 with a standard deviation $60,006.08. For the predictors, Number of Households Living in Poverty has a mean 189.77 with a standard deviation 164.32; Percent of Vacant Houses has a mean 11.29% with a standard deviation 9.63%; Percent of Bachelor's Degree has a mean 16.08% with a standard deviation 17.77%; Percent of Single House Units has a mean 9.23% with a standard deviation 13.25%.

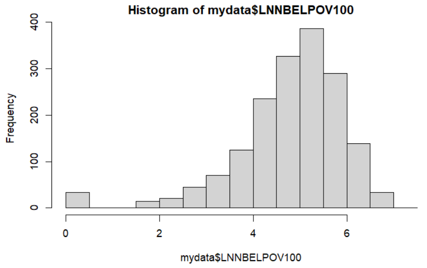
Variable Name	**Mean**	**SD**	**Distribution**
Dependent variable: Median House Value	$66,287.73	$60,006.08	
Predictors: Number of Households Living in Poverty	189.77	164.32	
Percent of Vacant Houses	11.29	9.63	
Percent of Bachelor's Degree	16.08	17.77	
Percent of Single House Units	9.23	13.25	

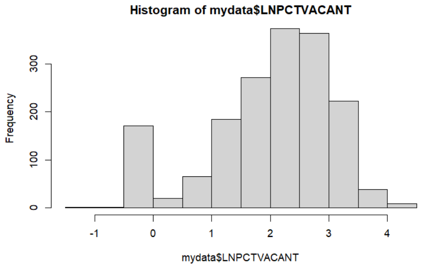
As is showed in the distribution histograms above, both the dependent variable and all of the predictors are not normal. After the logarithmic transformation, the *Median*

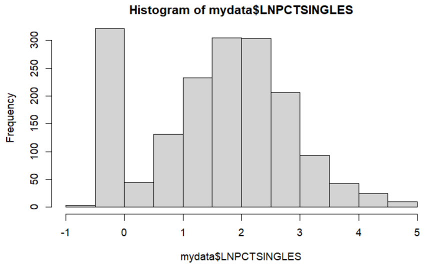
House Value looks normal, hence LNMEDHVAL will be used as the dependent variable in the regression analysis. For all the predictors, only the nature log of *Number of Households Living in Poverty* appears normal while others show zero-inflated distribution, so only LNNBELPOV100 will be used in the subsequent analyses. The original, untransformed PCBACHMORE, PCTVACANT, and PCTSINGLES variables will be kept and used in the regression.



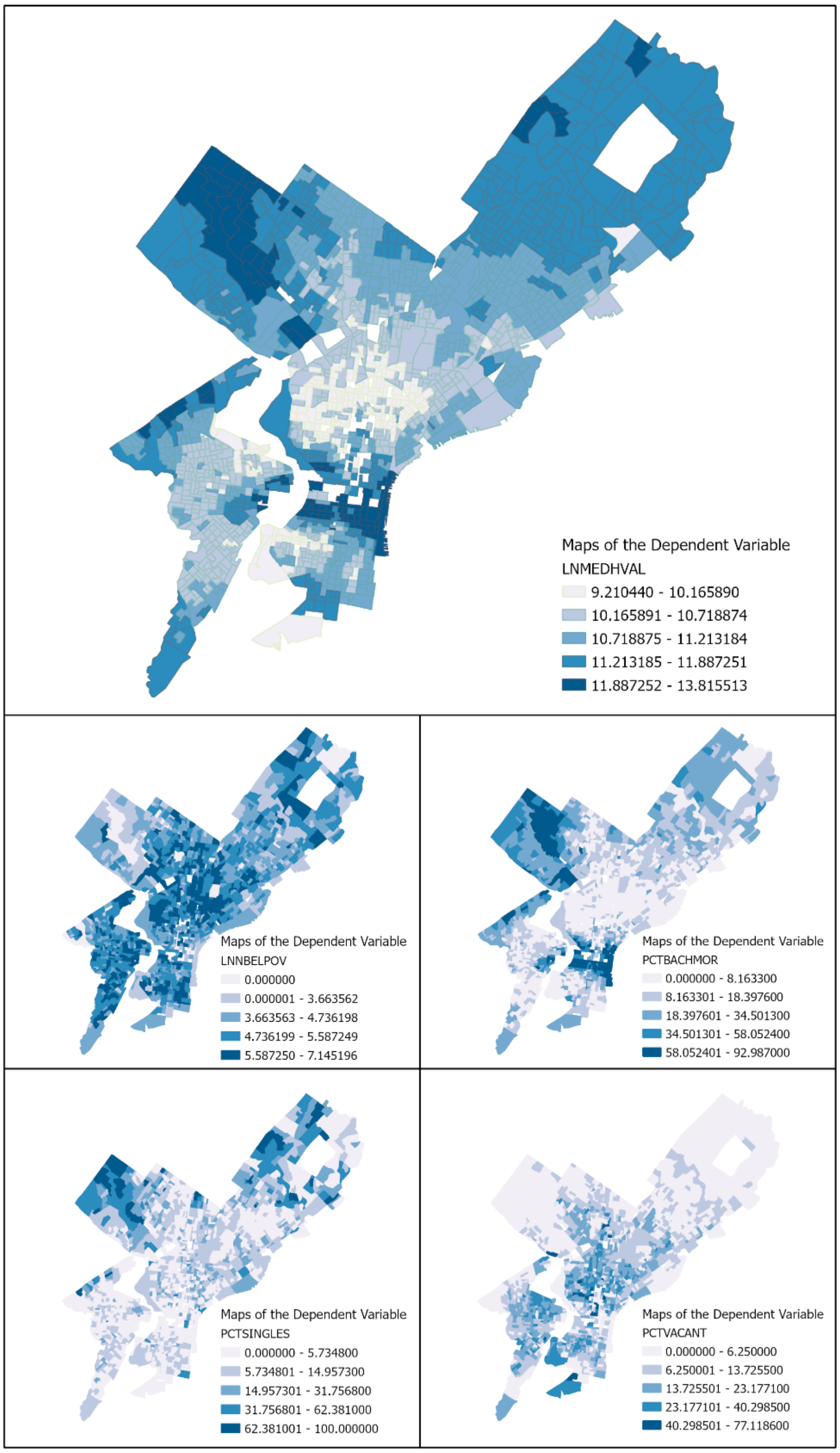




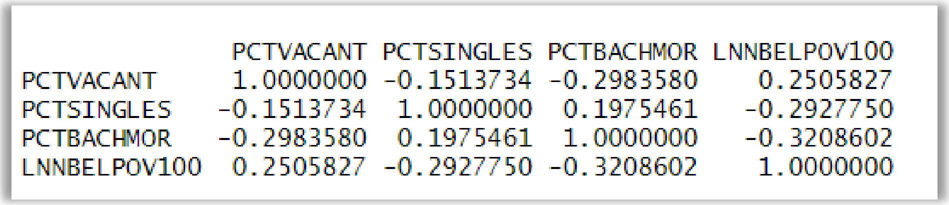




As is showed in the choropleth maps of the dependent variable below， the map of LNMEDHVAL, PCTVACANT and PCTBACHMOR looks similar. In the northern and center city areas of Philadelphia, PCBACHMORE showed a strong positive correlation with LNMEDHVAL results, while PCTVACANT showed a strong negative correlation with LNMEDHVAL. LNNBELPOV100 and PCTSINGLE are different compared to LNMEDHVAL, and both are more evenly distributed spatially compared to LNMEDHVAL. As a result, PCBACHMORE and PCTVACANT are expect be strongly associated with the dependent variable based on the visualization. At the same time, the higher the percentage of people with a bachelor's degree, the lower the percentage of Households Living in Poverty in that area. Therefore, PCBACHMORE and LNNBELPOV100 seem to be inter-correlated. However, because there are still a large number of areas in the northern part of the city and the central city that show different correlations, we do not expect severe multicollinearity to be an issue here.



Multicollinearity is when two or more predictors are very strongly correlated with each other: r>0.9 (or r<-0.9). According to the correlation matrix below, none of the correlations in the table have a value greater than 0.9 or less than -0.9, so there is no severe multicollinearity between any of two predictors. This is consistent with the conclusion of our previous visualization.



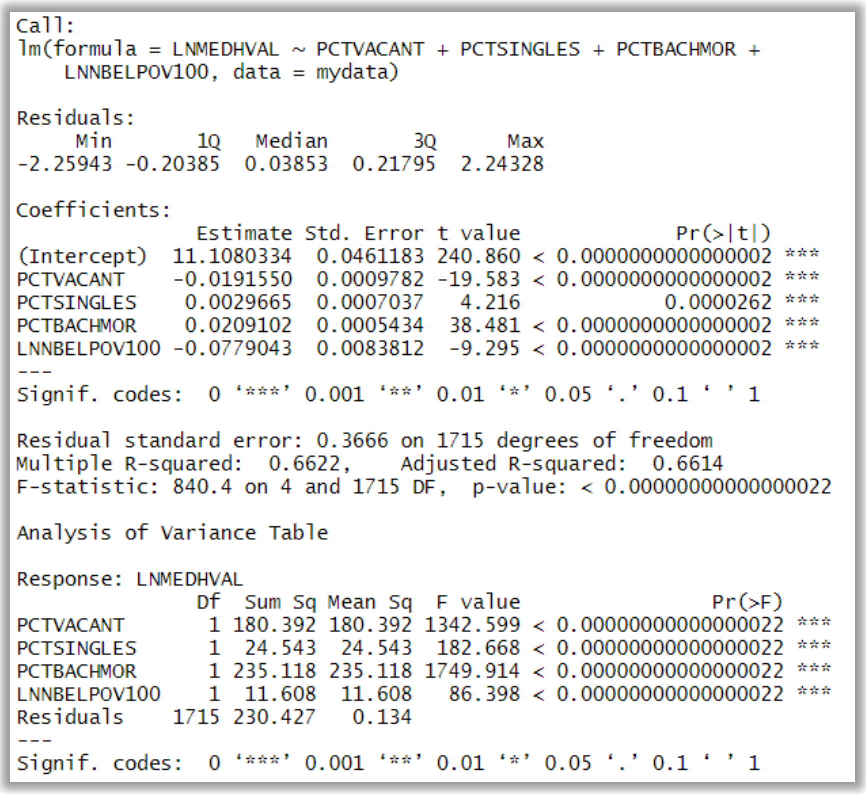
```
             PCTVACANT PCTSINGLES PCTBACHMOR LNNBELPOV100
PCTVACANT    1.0000000 -0.1513734 -0.2983580    0.2505827
PCTSINGLES  -0.1513734  1.0000000  0.1975461   -0.2927750
PCTBACHMOR  -0.2983580  0.1975461  1.0000000   -0.3208602
LNNBELPOV100 0.2505827 -0.2927750 -0.3208602    1.0000000
```

b) Regression Results

The nature log of median house value (LNMEDHVAL) was regressed on the % of vacant houses (PCTVACANT), % of single house units (PCTSINGLES), % of bachelor's degree (PCTBACHMOR), and Nature log of number of households living in poverty (LNNBELPOV100). The regression equation is:

$$\text{LNMEDHVAL} = \beta_0 + \beta_1 \text{PCTVACANT} + \beta_2 \text{PCTSINGLES} + \beta_3 \text{PCTBACHMOR} + \beta_4 \text{LNNBELPOV100} + \varepsilon$$

Below is the output of our regression model from R.



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

Residuals:
     Min       1Q   Median       3Q      Max 
-2.25943 -0.20385  0.03853  0.21795  2.24328 

Coefficients:
               Estimate Std. Error t value             Pr(>|t|)    
(Intercept)  11.1080334  0.0461183 240.860 < 0.0000000000000002 ***
PCTVACANT    -0.0191550  0.0009782 -19.583 < 0.0000000000000002 ***
PCTSINGLES    0.0029665  0.0007037   4.216            0.0000262 ***
PCTBACHMOR    0.0209102  0.0005434  38.481 < 0.0000000000000002 ***
LNNBELPOV100 -0.0779043  0.0083812  -9.295 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3666 on 1715 degrees of freedom
Multiple R-squared:  0.6622,	Adjusted R-squared:  0.6614 
F-statistic: 840.4 on 4 and 1715 DF,  p-value: < 0.00000000000000022

Analysis of Variance Table

Response: LNMEDHVAL
               Df  Sum Sq Mean Sq  F value                Pr(>F)    
PCTVACANT       1 180.392 180.392 1342.599 < 0.00000000000000022 ***
PCTSINGLES      1  24.543  24.543  182.668 < 0.00000000000000022 ***
PCTBACHMOR      1 235.118 235.118 1749.914 < 0.00000000000000022 ***
LNNBELPOV100    1  11.608  11.608   86.398 < 0.00000000000000022 ***
Residuals    1715 230.427   0.134                                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The regression output shows that PCTSINGLES and PCTBACHMOR are highly significant and positively

associated with LNMEDHVAL, while PCTVACANT and LNNBELPOV100 are highly significant and negatively associated with LNMEDHVAL (p<0.0001 for all four variables).

A one unit (i.e., percentage point) increase of % of vacant house units is associated with a $(e^{\beta_1} - 1) * 100\% = (e^{0.019} - 1) * 100\%$ = 1.91% decrease in median house value. The p-value of less than 0.0001 for PCTVACANT shows that if there is no relationship between PCTVACANT and the dependent variable LNMEDHVAL (i.e., if the null hypothesis that β_1 =0 is true), then the probability of getting a β_1 coefficient estimate of - 0.019 is less than 0.0001. These low probabilities indicate that we can safely rejected H_0: $\beta_1 = 0$ for Ha: $\beta_1 \neq 0$.

A one unit (i.e., percentage point) increase of % of single house units is associated with a $(e^{\beta_2} - 1) * 100\% = (e^{0.003} - 1) * 100\%$ = 0.30% increase in median house value. The p-value of less than 0.0001 for PCTSINGLES shows that if there is no relationship between PCTSINGLES and the dependent variable LNMEDHVAL (i.e., if the null hypothesis that β_2=0 is true), then the probability of getting a β_2 coefficient estimate of 0.003 is less than 0.0001. These low probabilities indicate that we can safely rejected H0: $\beta_2 = 0$ for Ha: $\beta_2 \neq 0$.

A one unit (i.e., percentage point) increase of % of bachelor's degree is associated with a $(e^{\beta_3} - 1) * 100\% = (e^{0.021} - 1) * 100\%$ = 2.12% increase in median house value. The p-value of less than 0.0001 for PCTBACHMOR shows that if there is no relationship between PCTBACHMOR and the dependent variable LNMEDHVAL (i.e., if the null hypothesis that $\beta_3 = 0$ is true), then the probability of getting a β_3 coefficient estimate of 0.021 is less than 0.0001. These low probabilities indicate that we can safely rejected H_0: $\beta_3 = 0$ for Ha: $\beta_3 \neq 0$.

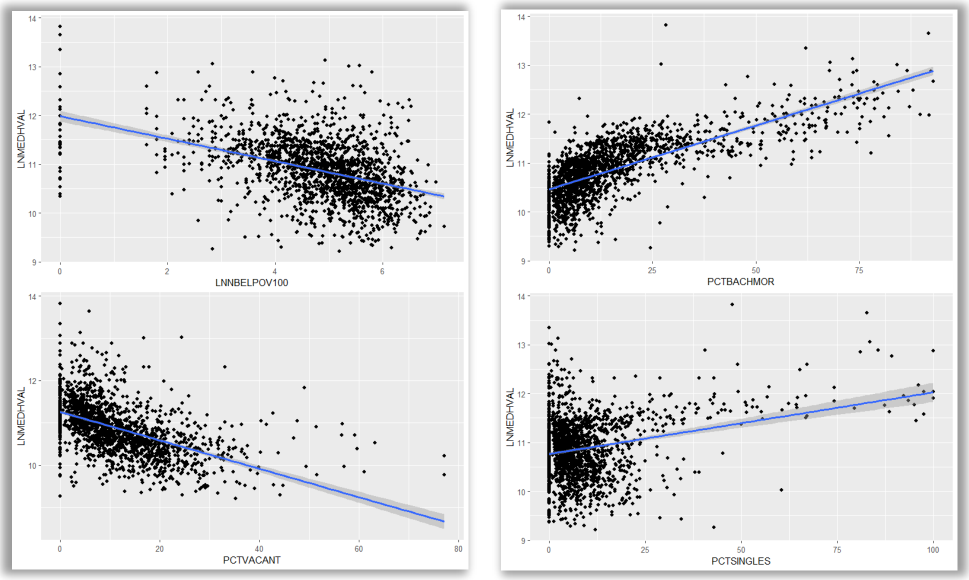
A one unit (i.e., percentage point) increase of % of households living in poverty is associated with a $(1.01^{\beta_4} - 1) * 100\% = (1.01^{0.078} - 1) * 100\% = 0.08\%$ decrease in median house value. The p-value of less than 0.0001 for LNNBELPOV100 shows that if there is no relationship between LNNBELPOV100 and the dependent variable LNMEDHVAL (i.e., if the null hypothesis that $\beta_4 = 0$ is true), then the probability of getting a β_4 coefficient estimate of - 0.078 is less than 0.0001. These low probabilities indicate that we can safely rejected H_0: $\beta_4 = 0$ for Ha: $\beta_4 \neq 0$.

A little less than two-thirds of the variance in the dependent variable is explained by the model (R^2 and adjusted R^2 are 0.662 and 0.661, respectively). The low p-value associated with the F-ratio shows that we can reject the null hypothesis that all coefficients in the model are 0.

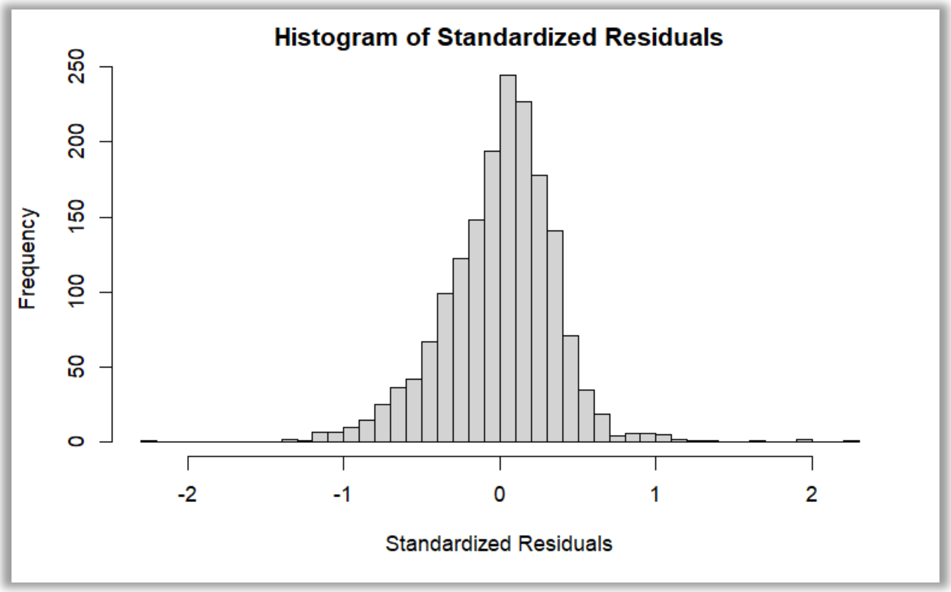
c) Regression Assumption Checks

In this section, we will talk about testing model assumptions and aptness. We have already looked at the variable distributions earlier.

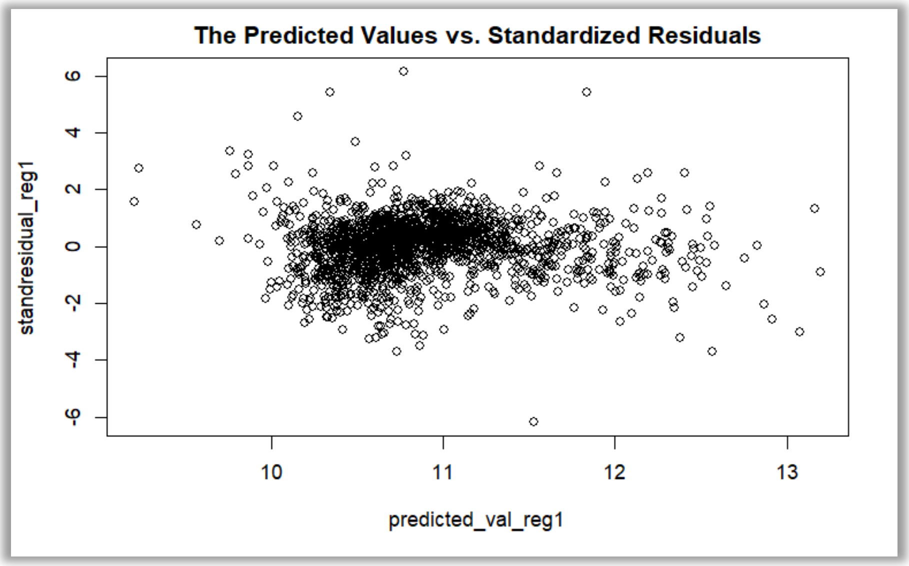
Below are the scatter plots of dependent variable and each of the predictors. None of them look linear, that is, the linearity assumption is violated.



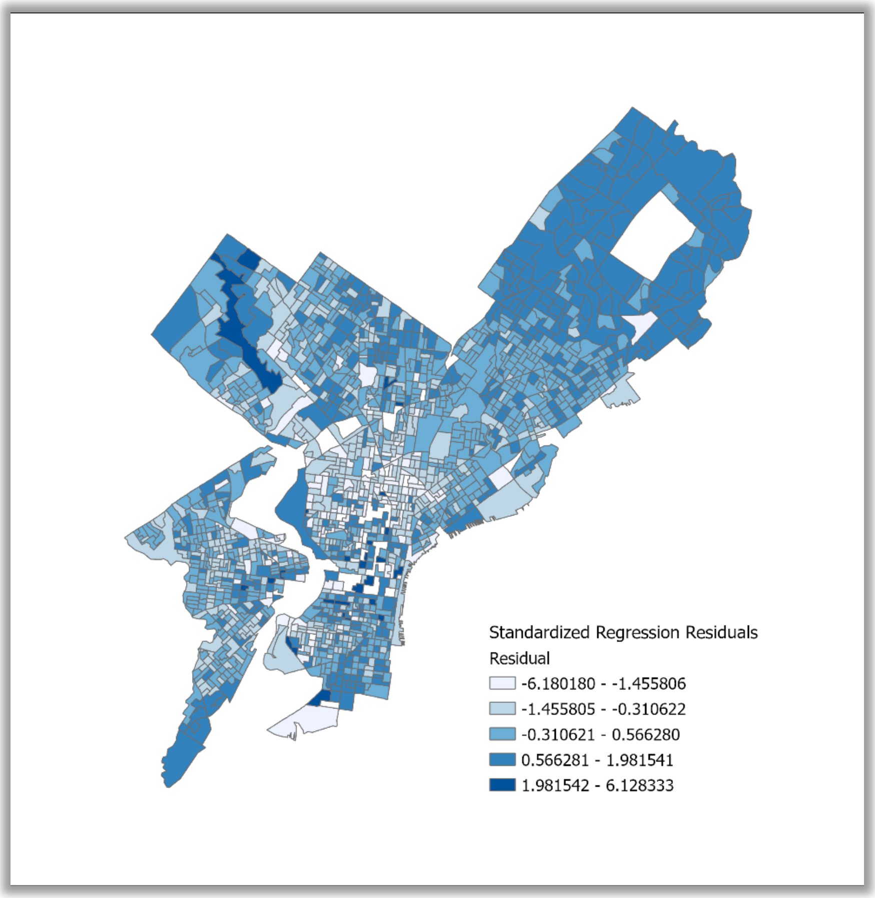
Below is the histogram of the standardized residuals. The residuals are not normal, that is, the assumption of normality of residuals is violated.



Standardized residuals are residuals divided by their standard error. They are used to compare residuals for different observations to each other to check if the model has heteroscedasticity. Below is the scatter plot of standardized residuals by predicted value. It shows heteroscedasticity, that is, when the predicted value increases, the standardized residuals tend to decrease. There are also some outliers with standardized residuals around 6 or -6.



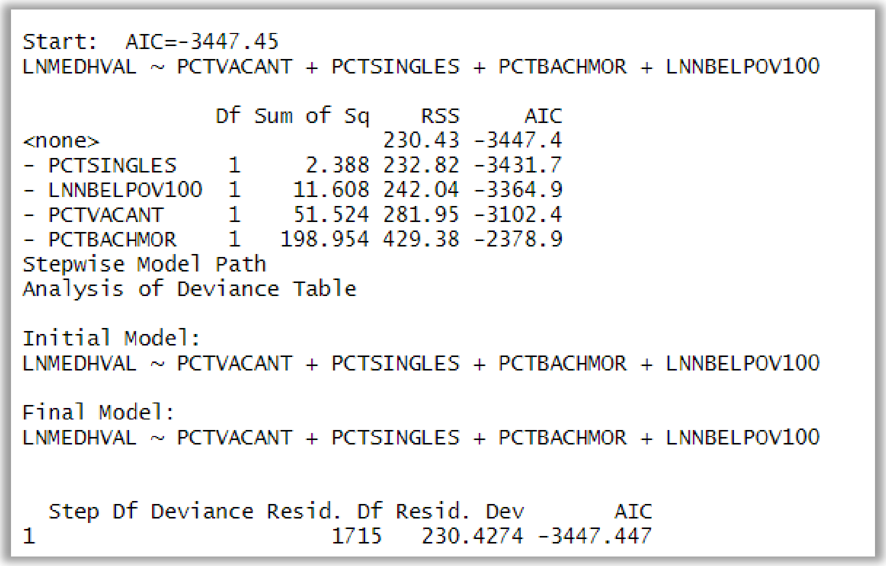
Among all 4 variables, PCTBACHMOR, PCTSINGLES and PCTVACANT all shows strong spatial autocorrelation. For PCTBACHMOR, there are clustering in the north-east, the north-west and downtown area. For PCTSINGLES, there are clustering in the north-east and north-west area. For PCTVACANT, there are clustering in the downtown, west and south area. The log-transformed poverty LNNBELPOV100 seems pretty much random distribution.



Above is the choropleth map of the standardized regression residuals. It can be seen that the forecast results are generally low in the Northwest region, while the forecast results are generally high in parts of North Philadelphia. Therefore, the results seem to be spatially autocorrelated.

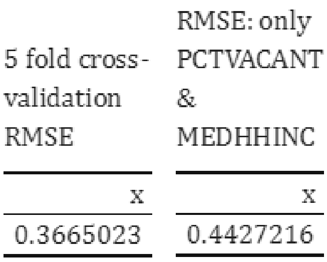
d) Additional Models

Below is the output from R for the stepwise regression model, all the 4 predictors were kept in the final model.



```
Start:  AIC=-3447.45
LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNNBELPOV100

               Df Sum of Sq    RSS     AIC
<none>                      230.43 -3447.4
- PCTSINGLES    1     2.388 232.82 -3431.7
- LNNBELPOV100  1    11.608 242.04 -3364.9
- PCTVACANT     1    51.524 281.95 -3102.4
- PCTBACHMOR    1   198.954 429.38 -2378.9
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   230.4274 -3447.447
```

Below are charts showing the cross-validation results of two models. One is the original model and the other is the model only includes PCTVACANT and MEDHHINC as predictors. The original model has lower RMSE than the model with 2 predictors.



5 fold cross-validation RMSE	RMSE: only PCTVACANT & MEDHHINC
x	x
0.3665023	0.4427216

4. Discussion and Limitations (~1 page)

a) Conclusions and Discussions

In this assignment, we examined the suitability of the available variables in the data, then picked two log transformed variables (LNMEDHVAL, LNNELPOV100) and other variables in their original form (PCTVACANT, PCTSINGLES, PCTBACHMOR) based on their distributions. Once we decided what variables we are going to pick, a linear regression was fitted using the lm command in R. Based on the results from ANOVA table and summary of the regression, we learned that the predictors used in this regression are significant. Finally, the last two testes that we applied to get our final regression model were stepwise regression and cross-validation. The final model returned by stepwise regression was the same as the regression model we fitted initially. Same result was concluded from cross-validation process that our model performed better than the other model that consisted PCTVACANT and MEDHHINC in terms of root mean square error.

The regression result shows that all four variables are highly significant and associated with median house value. Percent of single house units and percent of bachelor’s degree are positively associated with median house value, while percent of vacant houses and percent of households living in poverty are negatively associated with median house value. This result was not surprising as it is consistent with our logical perception of whether these four factors have positive or negative impact on housing quality

b) Quality of the Model

The result of F test shows that at least one of the coefficients is not 0, which also means that we can trust the value of R-squared. Both the R-squared and adjusted R-squared show that all the independent variables have only explained about 66% of the dependent variable. So, this model is not that good overall. There are a lot of other features we can consider including in to improve the performance of our model. One kind is urban facilities nearby like schools, parks, transport stations, crime (as a special facility) and so on. Another kind is the houses’ own features like quality, amenities, scenery and so on. As the model still have spatial autocorrelation, spatial process like neighborhood, racial segregation can be included in the model to avoid the effect of spatial autocorrelation.

The final model of stepwise regression includes all 4 predictors. This result tells us that when including all 4 predictors, the model has smallest value of the AIC, which means a relative high quality of the model.

The original model with 4 predictors has lower RMSE than the model with 2 predictors, which means the original model is better.

c) Limitations of the Model

In our model, the linearity is violated, which makes the coefficients estimates, the intercept estimates and the fitted values biased.

The normality of residuals is also violated, which does not affect the model as we have more than 30 observations due to Central Limit Theorem, but the normality is essential when we want to use the model to predict future median house value.

There is heterogeneity in our model, which makes the coefficient estimates less precise.

There are also several outliers, which makes the model poor fitted.

In our model, we log transformed the percent of households living in poverty as a dependent variable. As the coefficient’s absolute value of this variable is less than 0.3, 1% change of the log of percent of households living in poverty can be interpreted as 1% change in households living in poverty, which is associated with a 0.08% decrease in median house value. If we use the raw number of households living in poverty, we cannot interpret the coefficient in this way.