1) Introduction

The purpose of this report is to predict median house values in Philadelphia at census block group level, with several neighborhood characteristics, including households living in poverty, percentage of individuals with bachelor's degrees or higher, percentage of vacant houses, percentage of single house units.

From experience, all four predictors we are using might be related to the median house value. First, the number of households living in poverty represents a pay level for each tract, which affects the level of local house prices. Also, the pay level is associated with the percentage of high education. In addition to that, the rate of vacant houses can reflect supply and demand, that says how desirable or popular houses are in the market, and tracts with more desirable houses usually have higher house value. At last, places with high house values are typically occupied by more single house units, while there are more multi-family assets and townhouses in the places with lower values.

2) Methods

a) Data *Cleaning*

Philadelphia County Census block group data were obtained from the United States Census Bureau in the form of a shapefile and csv. Census tables for block group ID, median value of owner-occupied units, proportion of residents in the block group with at least a bachelor's degree, proportion of vacant housing units, percentage of detached single family housing units, number of households living in poverty, and median household income were all included in the data retrieved from the Census.

There were 1816 block group records, however block groups in which the population was less than 40, did not contain any housing units, or median house value less than $10,000 were removed from the study. Additionally, one block group in Northern Philadelphia was removed because it had a very high median house value (over $800,000) but had a very low median household income (less than $8,000).

b) Exploratory *Data Analysis*

To prepare the data for use in the OLS regression, the data were first explored using summary statistics and plotted the data in histograms. These summary statistics are useful for quickly identifying the number observations and means, medians, and standard deviations within those observations. Further, by plotting the histograms of the dependent and predictor variables, we can determine if any of variables do not meet the assumption of normality for the OLS regression and take appropriate action, such as logistically transforming those data.

Additionally, correlation coefficients (r) were calculated to measure the strength of the relationship between the variables. The formula below is used to calculate *r,* where n is the number of observations, $\overline{x}$ *and* $\overline{y}$ are the mean of those variables, and s_x *and* s_y are the standard deviations of those variables.

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

The possible values of correlation coefficients are between -1 and 1, with higher positive values indicating stronger positive correlation, lower negative values indicating negative correlation, and values around zero indicate no linear correlation. While non-linear relationships can have correlation coefficients, they are unintelligible because Pearson's r assumes a linear relationship.

c) Multiple Regression Analysis

Multiple regression is a statistical technique that refers to Ordinary Least Squares (OLS) regression and can be used to analyze the relationship between a single dependent variable and several independent variables. The objective of multiple regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value. Each predictor value is weighed, the weights denoting their relative contribution to the overall prediction, controlling for all other independent variables in the regression.

i) The LN(MEDHVAL) is regressed on LN(NBELPOV100), PCTBACHMOR, PCTVACANT and PCTSINGLES, the equation for this multiple regression is as follows:
***LNMEDHVAL=*β_0*+*β_1*PCBACHMORE+*β_2*LNNBELPOV100+*β_3*PCTVACANT+*β_4*PCTSINGLES+*ε**
In this equation, β_0 is the Y intercept, which is the mean value of LNMEDHVAL when holding all four predictors constant at zero. β_1, β_2, β_3, β_4 are coefficients of variables PCTBACHMOR, LNNBELPOV100, PCTVACANT, PCTSINGLES. The coefficient β_i of each predictor is the amount by which the dependent variable changes as the independent variable increases by one unit (holding all other variables constant). ε is the residual (A random variable, which is assumed to be normally distributed, with $E(\varepsilon) = 0$.

ii) The assumptions to use OLS regression are as follow. As for assumptions between the dependent variable and predictors, the mathematical relationship between each predictor in the equation and the dependent variable is linear. That says the relationship can be described as a linear function, whose graph lies on a straight line, and which can be described by giving the slope and y intercept. To validate the linearity, a scatter plot should be created to see whether the relationship is indeed linear.

As for assumptions among predictor variables, there should not be multicollinearity among different independent predictors. Collinear independent variables are related in some fashion, although the relationship may or may not be casual. For example, past performance might be related to market capitalization, as stocks that have performed well in the past will have increasing market values. At the same time, every observation (i.e., every observation of PCTBACHMORE in different census tract) is independent and have no effect on other observation, which means the occurrence of one observation provides no information about the occurrence of the other observation.

As for assumptions for residuals, they are random, independent, and normally distributed, with $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$. At the same time, residuals should have homoscedasticity, which means variance of residuals remain constant at any value of independent predictors.

iii) In the multiple regression, parameter β_0, β_1,..., β_k, and σ^2 should be estimated. σ^2 is the variance of residual, determining the amount of variability inherent in the regression model. If σ^2 is small, then the variance of ε is small, meaning a less variability. If σ^2 is large, then the variance of ε is large, meaning a bigger variability in the regression model.

iv) To describe the relationship between predictors and dependent variables at a best way, the parameters β_0, β_1,..., β_k in the equation should be estimated in the multiple regression simultaneously, using 'Least Square Estimate'. It is expressed as follows.

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

'Least Square Estimate' is that given n observations on y, and k predictors x_1... x_k , the estimates $\hat{\beta}_0$, $\hat{\beta}_1$,..., $\hat{\beta}_k$ are chosen simultaneously to minimize the expression for the Error Sum of Squares (SSE).

As for σ^2, in Multiple Regression, it should be estimated as follows. k is the number of predictors and n is the number of observations, and MSE stands for mean squared error

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

v) To quantify effect of the multiple regression, R^2 is introduced. It is the coefficient of multiple determination, or the proportion of variance in the model explained by all k predictors. It can be calculated as follows.

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

In this equation, SSE is the Error Sum of Squares and can be calculated as indicated in *v.* SST is the Total Sum of Squares. It can be calculated as follows. $SST = \sum(y_i - \overline{y})^2$. Different from SSE, which is the sum of square of every actual value of the dependent variable minus corresponding predicted values, SST calculates the sum of square of every actual values of the dependent variables minus the mean value of actual values of the dependent variable. However, in multiple regression, due to possible collinearities among predictors, extra predictors will generally increase R^2. To eliminate this effect, R^2 is typically adjusted for the number of predictors using the follow formula, where n is the number of observations and k is the number of predictors. In this way, we can obtain the adjusted R-Squared.

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

vi) Besides using R^2 to describe the effect of predictors, a F-Test for all predictors and Hypothesis Tests(T-Test) for every predictor should be used to describe the effect. F-Test is the model utility test, measuring the goodness of fit for the regression. Its null hypothesis is that all coefficients in the model are (jointly) zero, which means none of the independent variable is a significant predictor. Its alternative hypothesis is that at least one of the coefficients is not zero.

Null Hypothesis: $\boldsymbol{H_0\colon \beta_1 = \beta_2 = \cdots = \beta_k = 0}$

Alternative Hypothesis: $\boldsymbol{H_a\colon \beta_1 \neq 0 \; or \; \beta_2 \neq 0 \; or \ldots or \; \beta_k \neq 0}$

If we can reject the null hypothesis for the alternative hypothesis, that means at least one predictor is significant in the regression. Then we need to do hypothesis tests for every predictor. For every predictor, the Null Hypothesis is always that the predictor *i* is not associated with the dependent variable. The Alternative Hypothesis is always that that the predictor *i* is associated with the dependent variable.

Null Hypothesis: $\boldsymbol{H_0\colon \beta_i = 0}$

Alternative Hypothesis: $\boldsymbol{H_a\colon \beta_i \neq 0}$

To do so, we can use p-value. The p-value is the probability of observing a value that is at least as different from 0 (the value stated in H0) as the given estimated value. If this probability is small enough (generally, *p<0.05*), we reject the null hypothesis of $\beta_i = 0$ for an alternative hypothesis of $\beta_i \neq 0$.The rejection of a null hypothesis indicates that the independent variable is a statistically significant predictor of the dependent variable. If this probability is not small enough (generally, *p>0.05*), we fail to reject the null hypothesis of $\beta_i = 0$, which means the dependent variable is not related to the independent variable.

d) Additional *Analysis*

i) Stepwise regression, a data mining method that selects predictors automatically based on several criteria, wil be applied here after the OLS regression. There are two main filter criteria: variables with p-value less than 0.1; models with the smallest value of Akaike Information Criterion.

However, selecting predictors in this way is not an ideal method. First of all, the final model resulting from stepwise regression is not optimal in any specified sense. Secondly, while there may be many equally fitted models, it can only yield one single final model. Thirdly, this procedure does not include any researchers' expertise which is indispensable when reserving significant variables that are not eligible in stepwise models. Fourthly, Type I and Type II errors are still problems in final models, especially when many t-tests for testing βk = 0 are conducted, so we cannot jump to absolute conclusions of variables' significance. At last, although the order in which variables are removed or added can provide valuable information about the quality of the predictors, we should be careful about over interpretation of the order.

ii) To estimate the skill of the final model on new data, k-fold cross-validation in which k=5 will be used. K-fold cross-validation is a resampling procedure. First, shuffle the dataset randomly and split it into 5 groups. Then, for each unique group, take the group as a hold out or test data set while taking the remaining groups as a training data set, fit a model on the training set and evaluate it on the test set, and retain the evaluation score and discard the model. Lastly, summarize the skill of the model using the sample of model evaluation scores, which is RMSE in this case.

RMSE is the square root of the average of squared errors across the 5 folds, which allows us to measure how far predicted values are from observed values in regression models.

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

where Σ is a fancy symbol that means "sum", $\hat{y}_i$ is the predicted value for the *i*th observation in the dataset, y_i is the observed value for the *i*th observation in the dataset, ε_i is the difference between the predicted value and observed value for the *i*th observation, *n* is the sample size.

e) Software

In this report, R will be used for all data analysis and histogram graphics, and ArcGIS will be used to create maps.

3) Results

a) Exploratory Results

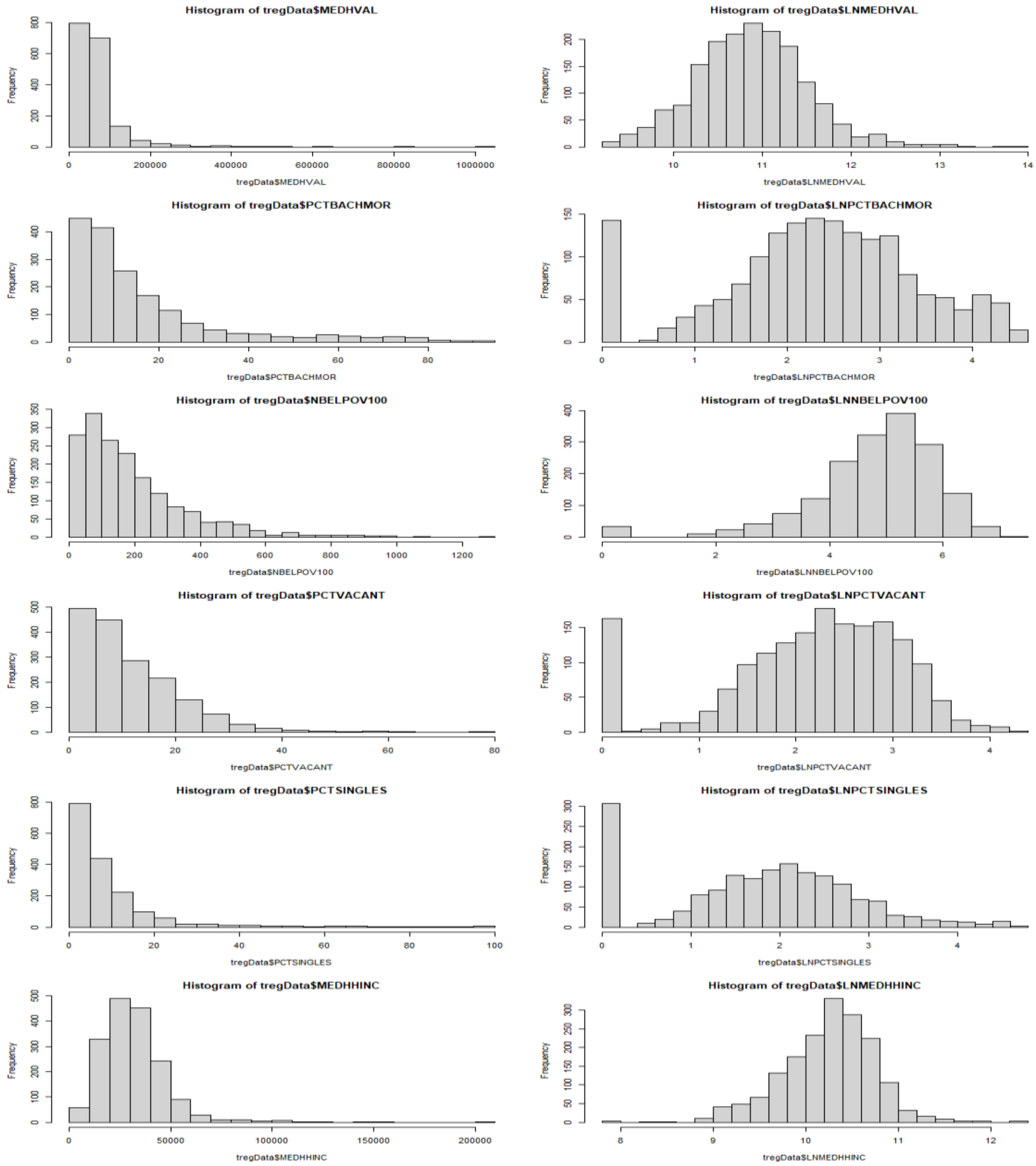
i) Summary Statistics

The distributions of variables are listed below. First and foremost, the distributions of variables are listed below. The dependent variable, median house value, has a mean of 66288, and its standard deviation is 60006, which is relatively huge compared with its mean value. And the average number of households living in poverty is around 190 by census tract, while its standard deviation is 164. For the percent of individuals with bachelor's degrees or higher, its mean value is approximately 16, with a standard deviation of above 17. The average percent of vacant houses is slightly higher than 11, and the standard deviation is 9.6. At last, the mean and standard deviation of the percent of single house units is around 9 and 13 respectively.

Variable	**Mean**	**SD**
Dependent Variable		
Median House Value	66288	60006.08
Predictors		
# Households Living in Poverty	189.8	164.31
% of Individuals with Bachelor's Degrees or Higher	16.08	17.77
% of Vacant Houses	11.29	9.63
% of Single House Unit	9.23	13.25

ii) Histogram Distribution and Logarithmic Transformation

At the same time, observations from the histograms of above variables tells us that none of the variables looks normal. This being the case, to meet the prerequisite of OLS regression that predictors should be normally distributed, we examine whether a logarithmic transformation of the variable helps achieve a normal distribution.

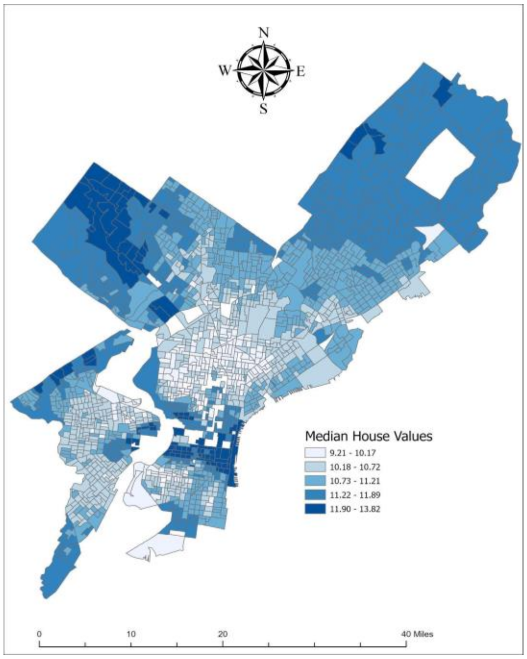


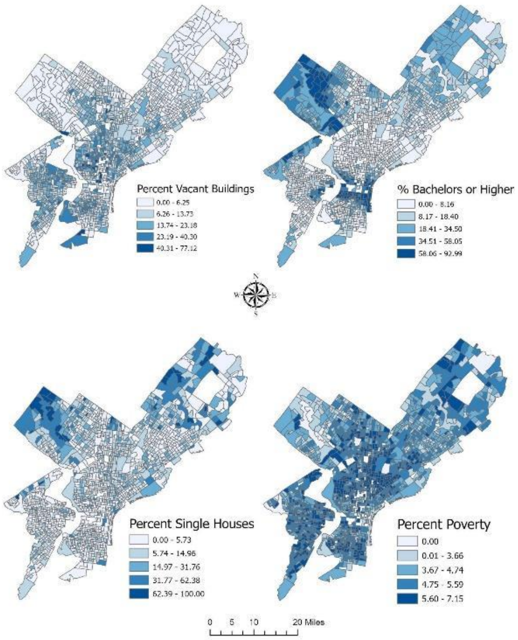
The dependent variable does look more or less normal after the transformation – hence, LNMEDHVAL will be used as the dependent variable in the regression analysis. Also for the predictors, the logarithmic transformation only helps normalize the NBELPOV100 variable (so we will use LNNBELPOV100 in the subsequent analyses). The other variables have a large spike at zero after the transformations, so we will use the original, untransformed PCBACHMORE, PCTVACANT, and PCTSINGLES variables in the regression.1.

Other regression assumptions, such as linear relationship with the dependent variable, homoscedastic of residuals, and independence of observations and residuals will be examined in the assumption check section below.

iii) Choropleth maps of the variables

The Choropleth maps of the dependent variable and the predictors are visualized as follow. From the maps, we can tell that the choropleth of the median house values is similar with the choropleth of Percentage of Bachelor or higher, displaying a highly positive relationship between these two variables. Also, the choropleth of median house values is different with the choropleth of vacant buildings, displaying a negative relationship between these two variables. Besides, looking through the maps, we can find out a negative intercorrelated relationships between PCTBACHMOR and the LNNBELPOV100 , but whether there is a highly multicollinearity between these two predictors. We should look at the correlation matrix in the next section.





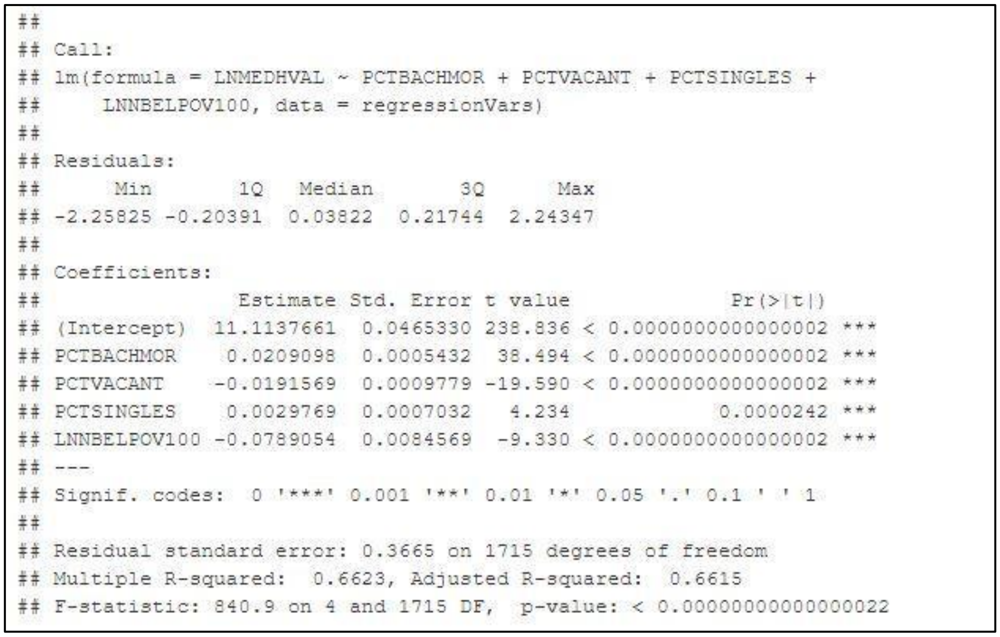
iv) Correlations Matrix

The Correlations Matrix is presented as follows. From this table, there is not severe multicollinearity among these predictors (because there are no correlations where r>.8 or r<.8) Besides, the correlation matrix corresponds with conclusions in the previous section, and also tells a moderate multicollinearity between the PCTBACHMOR and the LNNBELPOV100. This means we don’t need to remove one from these two predictors.

	PCTBACHMOR	PCTVACANT	PCTSINGLES	LNMEDHVAL	LNNBELPOV100
PCTBACHMOR	1	-0.298	0.198	0.736	-0.320
PCTVACANT	-0.298	1	-0.151	-0.514	0.250
PCTSINGLES	0.198	-0.151	1	0.265	-0.291
LNMEDHVAL	0.736	-0.514	0.265	1	-0.424
LNNBELPOV100	-0.320	0.250	-0.291	-0.424	1

b) Regression Results

We regressed the median house value (LNMEDHVAL) on the % of individuals with a bachelor’s degree or more (PCTBACHMOR), the % of vacant dwellings in the census tract (PCTVACANT), the % of single family dwellings in the census tract (PCTSINGLES) and the number of households living below the 100% poverty level (LNNBELPOV100). The regression output indicates that all predictor variables (PCTBACHMOR, PCTVACANT, PCTSINGLES, AND LNNBELPOV100) are highly significant and are positively associated with median house value (p<0.0001 for all variables). A one-unit increase (i.e., percentage point) in the % of people with a



```
##
## Call:
## lm(formula = LNMEDHVAL ~ PCTBACHMOR + PCTVACANT + PCTSINGLES +
##     LNNBELPOV100, data = regressionVars)
##
## 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 ***
## PCTBACHMOR    0.0209098  0.0005432  38.494 < 0.0000000000000002 ***
## PCTVACANT    -0.0191569  0.0009779 -19.590 < 0.0000000000000002 ***
## PCTSINGLES    0.0029769  0.0007032   4.234            0.0000242 ***
## 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
```

bachelor's degree or more is associated with an 2.11% increase in median house value in a particular census tract. Similarly, a one-unit increase (1%) in the % of single-family dwellings in a census tract is associate with a 0.298% increase in median house value. Additionally, a one-unit increase (1%) in the % of vacant dwellings in a census tract is associate with a 1.89% decrease in median house value. Finally, a one-percent increase in the number of households below the 100% poverty line in a census tract is associate with a 7.85% decrease in median house value.

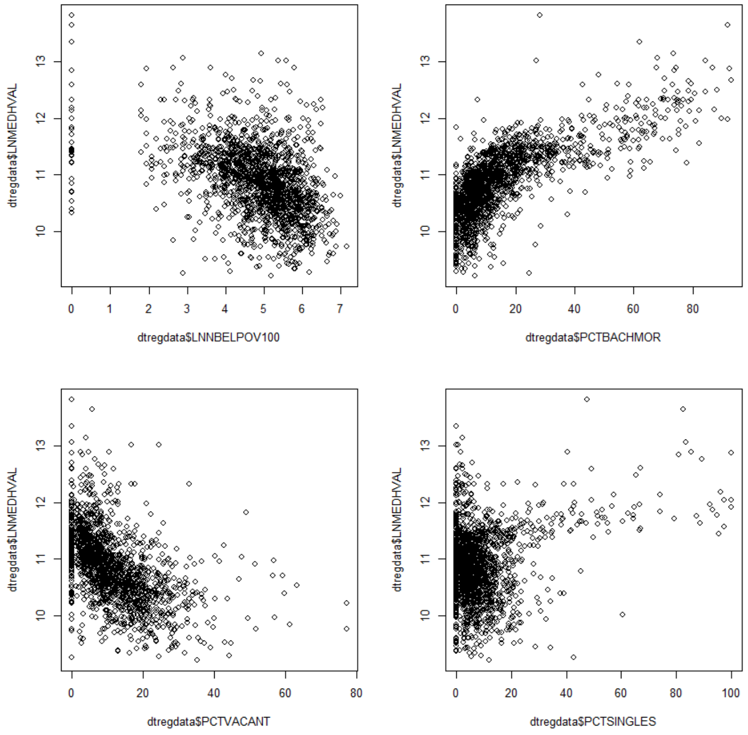
The p-value of less than 0.0001 for PCTBACHMOR indicates that *if* there is actually no relation between PCTBACHMOR and the dependent variable, MEDHVAL, thus indicating that if the null hypothesis that $\beta_1=0$ is actually true, then the probability of getting a β1 coefficient estimate of 0.0209 is less than 0.0001. Similarly, the p-value of less than 0.0001 for PCTSINGLES indicates that *if* there is no actual relationship between PCTSINGLES and the dependent variable, MEDHVAL, (if the null hypothesis that $\beta_4=0$ is actually true), then the probability of getting a β4 coefficient estimate of 0.00297 is less than 0.0001. Additionally, the p-value of less than 0.0001 for PCTVACANT indicates that *if* there is no actual relationship between PCTVACANT and MEDHVAL, (if the null hypothesis that $\beta_3=0$ is actually true), then the probability of getting a β3 coefficient estimate of -0.01916 is less than 0.0001. Finally, the p-value of less than 0.0001 for LMNBELPOV100 indicates that *if* there is no actual relationship between LMNBELPOV100 and MEDHVAL, (if the null hypothesis that $\beta_2=0$ is actually true), then the probability of getting a β2 coefficient estimate of -0.0789 is less than 0.0001. Therefore, because these probabilities are low, we can safely reject *H0:* $\beta_1 = 0$ *for Ha:* $\beta_1 \neq 0$, *H0:* $\beta_2 = 0$ *for Ha:* $\beta_2 \neq 0$, *H0:* $\beta_3 = 0$ *for Ha:* $\beta_3 \neq 0$, and *H0:* $\beta_4 = 0$ *for Ha:* $\beta_4 \neq 0$ *(at most reasonable levels of* $\alpha = P(\text{Type I error})$*).*

More than 60% 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 ($p<0.0001$) associated with the F-ratio indicates that we can reject the null hypothesis that all model coefficients are 0.

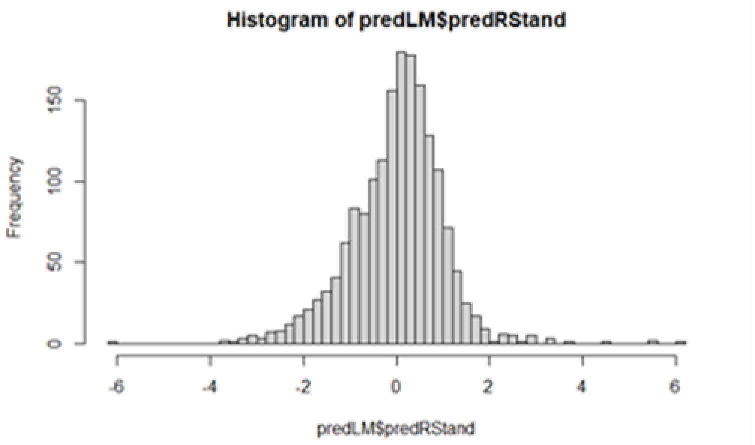
c) Regression Assumption *Checks*

i) In this section, the model assumptions will be tested. As we have already checked the variable distribution earlier in section 3(a), for variables MEDHVAL and BELPOV100 which have a better normal distribution after logarithmic transformation, the variable transformations have been made, here we directly looked at the scatter plots of the dependent variable and each of the predictors.

ii) To examine whether the relationships between dependent variable and each of the predictors are linear. The scatter plots are made as follows. From the graphs, a linear relationship cannot be observed, which means the assumption of linear relationship cannot be met and there will undoubtedly result in some bias in the estimate of $\hat{y}$.



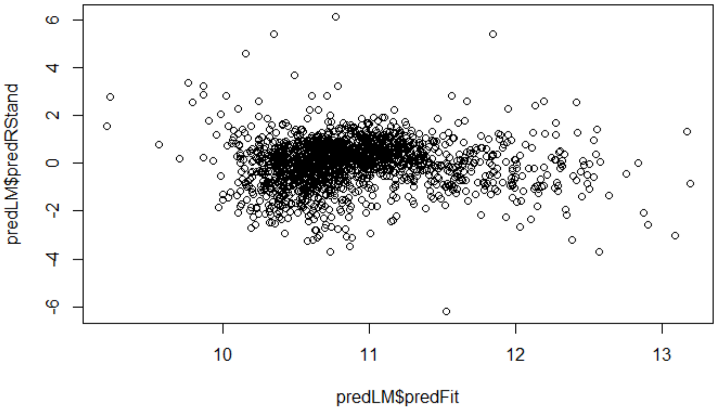
iii) To examine whether the residuals are normally distributed. A histogram of the standardized residuals is presented as follows. From the graph, we can tell the Standard Residuals are close to normal distribution, meaning the assumption of normality of residuals is not violated. Normality is essential for all sample sizes to predict future values of the dependent variable (Here is the Median House Value).



iv) To examine the homoscedasticity of residuals, a scatter plot of standardized residual by predicted value has been made. The reason we use standardized residuals is that we need to compare residuals for different observations to each other, standardized residuals can help us achieve comparisons. Standard Residuals are the raw residuals divided by the standard error of estimate. It can be calculated through the following formula.

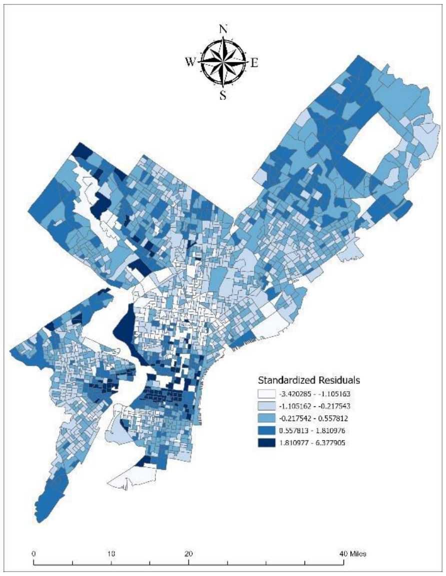
$$e_i^* \approx \frac{\varepsilon_i}{s} \approx \frac{\varepsilon_i}{\sqrt{\frac{SSE}{n-2}}}$$

From this graph, conclusions are drawn as follows. Heteroscedasticity exists in residuals, violating the assumption about residual homoscedasticity. And there are some outliers in the graph especially in the low and high interval in predicted house value. However, the model has a better prediction effect in the middle interval of the predicted house value, because in this interval, the standardized residuals are concentrated around zero with few outliers.



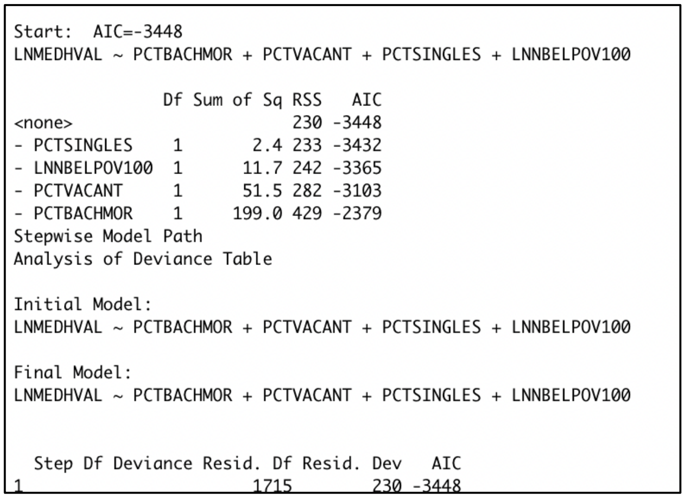
v) We can draw conclusions about independence of observations referencing the maps of dependent variable and predictors in 3-a section. From those maps, spatial autocorrelation can be easily observed, similar observations tend to cluster in space. Just as Waldo Tobler (1970) said, “Everything is related to everything else, but near things are more related than distant things.” Median House Value is a variable highly related to space, for which spatial autocorrelation is inevitably. This makes assumption of independent observation invalid. And also, this spatial autocorrelation may account for the heteroscedasticity of standardized residuals, because they are not account in the multiple regression, their effect therefore exists in residuals.

vi) The Choropleth map of the standardized regression residuals is displayed below. The spatial patterns of standardized residuals are noticeable. The median house values are underpredict in the city center, and are over predict in the north and west Philadelphia, which means further special operation is needed to account for this spatial clustering.



d) Additional *Models*

i) In the result of stepwise regression, all four predictors in the original model are kept in the final model, which means the original model is the best model with lowest AIC. Here is the output of the procedure.



```
Start:  AIC=-3448
LNMEDHVAL ~ PCTBACHMOR + PCTVACANT + PCTSINGLES + LNNBELPOV100

               Df Sum of Sq RSS   AIC
<none>                      230 -3448
- PCTSINGLES    1       2.4 233 -3432
- LNNBELPOV100  1      11.7 242 -3365
- PCTVACANT     1      51.5 282 -3103
- PCTBACHMOR    1     199.0 429 -2379
Stepwise Model Path
Analysis of Deviance Table

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

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


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

ii) The RMSE of the original model that includes all four predictors is 0.366, and the RMSE of the model that only includes PCTVACANT and MEDHHINC as predictors is 0.443. Therefore, the original four-predictor model is best, considering that the lower RMSE one model has, the better it is.

4) Discussion and Limitations

We used one form of Ordinary Least Squares Regression (Multiple Regression) to regress median household value on the four predictors: the percentage of individuals with a bachelor's degree or more, the percentage of vacant houses in a census tract, the percentage of single-family detached homes in a census tract, and the number of households with incomes below the 100% poverty level. We first log-transformed the median house value and the number of households living the below 100% poverty line because they were not originally normally distributed.

We found that all four predictor variables were significantly correlated with median household value. These findings are not particularly surprising because neighborhoods with people that are more educated tend to have higher salaries over their lifetime, which means they can live in places where homes are more expensive. Further, we would expect a lower number of people living below 100% poverty in neighborhoods where those households have higher incomes. Additionally, single-family houses tend to occupy more land area than townhomes, condos, or apartment dwellings, and thus would be associated with higher home values. Moreover, we would expect more desirable neighborhoods to have higher occupancy within the neighborhood (percentage of vacant dwellings).

The R^2 value of our model was 0.66, which means the model is able to predict 66% of variation within median household value across our sample tracts. The p-value of the F-ratio test was statistically significant (<0.001), which indicates that at least one of the predictor variables was significantly correlated with median house value. However, we believe that there are still predictors of median house value which are not accounted for in this model. For example, the model could account for spatial processes such as neighborhood characteristics or amenities such as schools, hospitals, parks, and open space. Additionally, internal characteristics of the houses could be accounted for by averaging the square footage, number of rooms, and house age across the census tract.

As mentioned in the results section, the assumption of linearity between the predictors and dependent variables was violated. Additionally, the residuals were heteroskedastic and violated the assumption of homoskedasticity. Further, using the raw count of number of households living under the 100% poverty line could be subject to bias because it does not indicate the relative frequency of the households in poverty compared to the total number of households in a tract. For example, if one tract (tract A) has 1000 households and 5 of those households are under the 100% poverty line, and another tract (tract B) has 100 households and 5 of those households are under the 100% poverty line, they will have the same effect in the regression model if we use the raw count value, whereas if we use the relative frequency, tract A is generally less impoverished than tract B.