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

Yiming Cao, Sujan Kakumanu, Angel Sanaa Rutherford

2025-10-15

1 Introduction

In this analysis, we use a multiple linear regression model to predict median house values in Philadelphia. Drawing on Philadelphia's tract-level census data, we examine the impact of our four predictors on our response variable median house value: percentage with at least a bachelor's degree, percentage of vacant spaces, number living below the poverty line, and percentage of single family housing units.

Prior theoretical knowledge of the relationships between housing markets and socioeconomic factors has led us to hypothesize a relationship between these four predictors and median house value. High rates of educational attainment and single-family homes are likely positively associated with house values as they may indicate neighborhood stability by signaling higher earning potential and long-term residency. Conversely, high rates of vacancy and poverty levels are likely negatively associated with house values as they may indicate neighborhood instability through a lack of high earning residents and occupants overall. In this analysis, we aim to assess the explanatory power of these predictors and briefly explore if these relationships possess any spatial patterns.

2 Methods

2.1 Data Cleaning

2.2 Exploratory Data Analysis

2.3 Multiple Regression Analysis

2.4 Additional Analysis

2.4.1

Using the stepAIC() and step$anova command, we applied bidirectional stepwise regression to analyze the fit of our linear model. Stepwise regression determines the minimum number of predictors that yield the best model. Stepwise regression automatically selects or eliminates predictors, either forwards, backwards, or bidirectionally, based on some type of criteria that measures the goodness of fit. In this case, we are attempting to determine the predictor or combination of predictors that minimize the Akaike Information Criterion (AIC).

Stepwise regression, however, poses many limitations as it does not consider theoretical relevance of the predictors, may overlook alternative valid models, and runs the risk of excluding important predictors and including unimportant predictors, especially due to the numerous t-tests measuring whether the null hypothesis, $\beta_k = 0$, is true.

2.4.2

To perform cross-validation, we used the trainControl() function with the method parameter set to "cv" (cross-validation) and the train() function with the method parameter set to "lm" (linear model). Cross-validation is a technique that measures model performance unbiasedly by training the model on a select subgroup of observations and seeing how well it estimates deliberately excluded observations. K-fold cross validation where k=5, specifically, divides data sets into five non-overlapping folds and repeatedly uses four folds for training the model and one fold for validating the model so that each fold trains the model multiple times and validates the model once. This method ensures a model's generalizability to new data and minimizes distortion by avoiding omitting and duplicating data in its measure of fit. After all five iterations are complete, the Root Mean Squared Error (RMSE) of the model is returned as a measurement of the average magnitude of predicted residuals or errors between predicted values, $\hat{y}_i$, of observation i, estimated by the model's β coefficient, and the actual value, y_i, of the validation set. The complete formula for RMSE is as follows:

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

After performing k-fold cross-validation on two or more models, the RMSE of the models can be compared to determine which model has the best performance. A smaller RMSE indicates that the model's predictions are, on average, closer to the actual values, and thus more representative of the data.

2.5 Software Used

2.5.1

All data analysis was conducted using R. Within R, the following packages were used to perform data preparation, exploratory analysis, regression modeling, and visualization: ggplot, dplyr, sf, patchwork, MASS, and caret.

3 Results

3.1 Exploratory Results

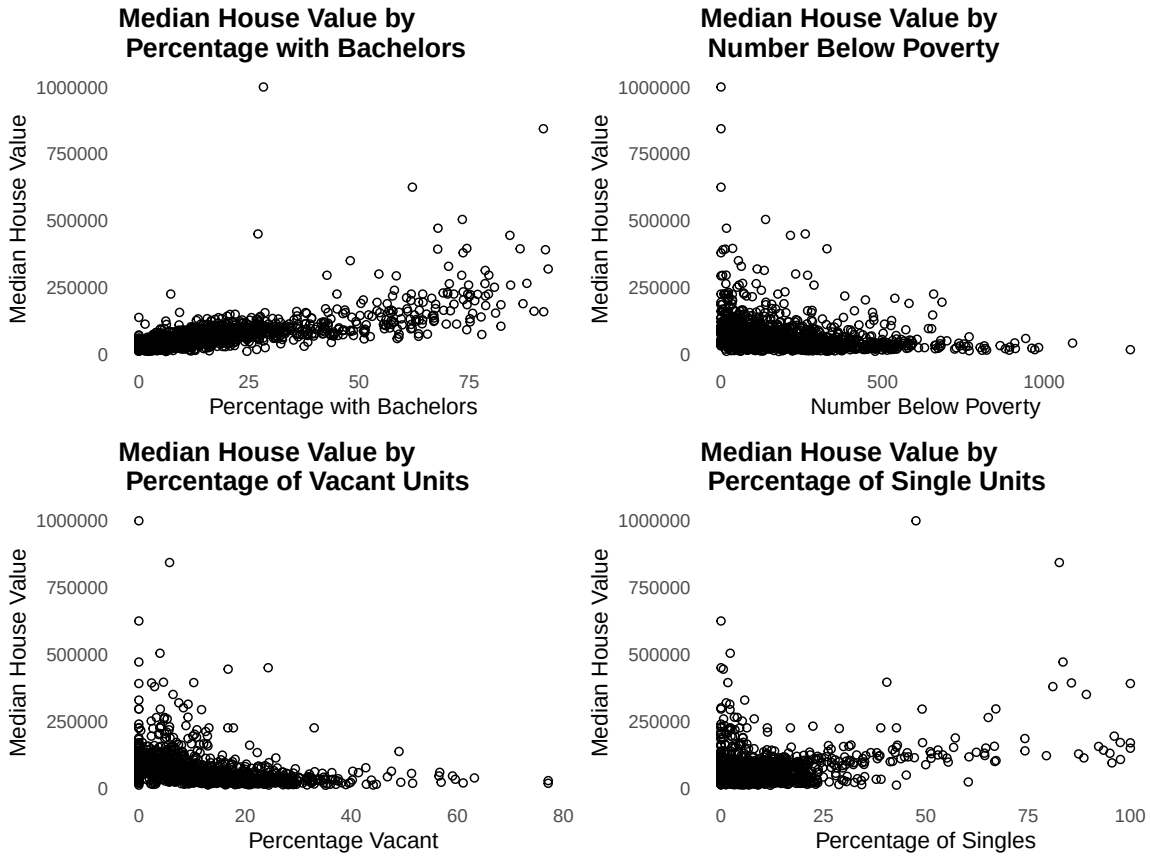
3.2 Regression Results

3.3 Regression Assumption Checks

3.3.1

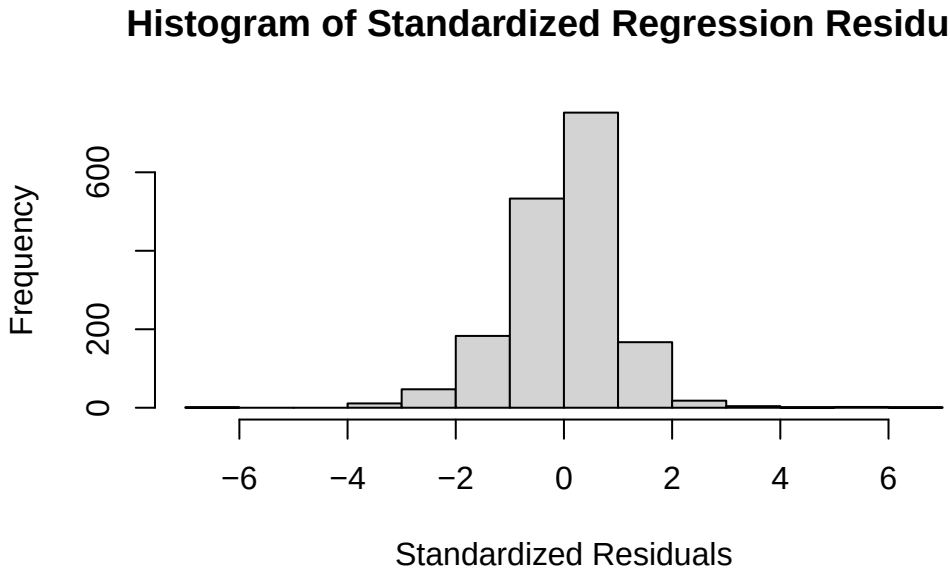
In this section, we assess whether our multiple regression model meets key assumptions and take the necessary steps to address any violations of these assumptions. Early visualizations of the distribution of the predictors PCTBACHMOR, NBELOWPOV100, PCTVACANT, and PCTSINGLES and the dependent variable MEDHVAL were presented by histograms which all showed positively-skewed distributions for all predictors. While multiple regression assumptions require the normality of residuals and not predictor values, non-normal distribution of predictors values can indicate violations of the assumptions of non-normal residuals and a lack of linearity.

3.3.2



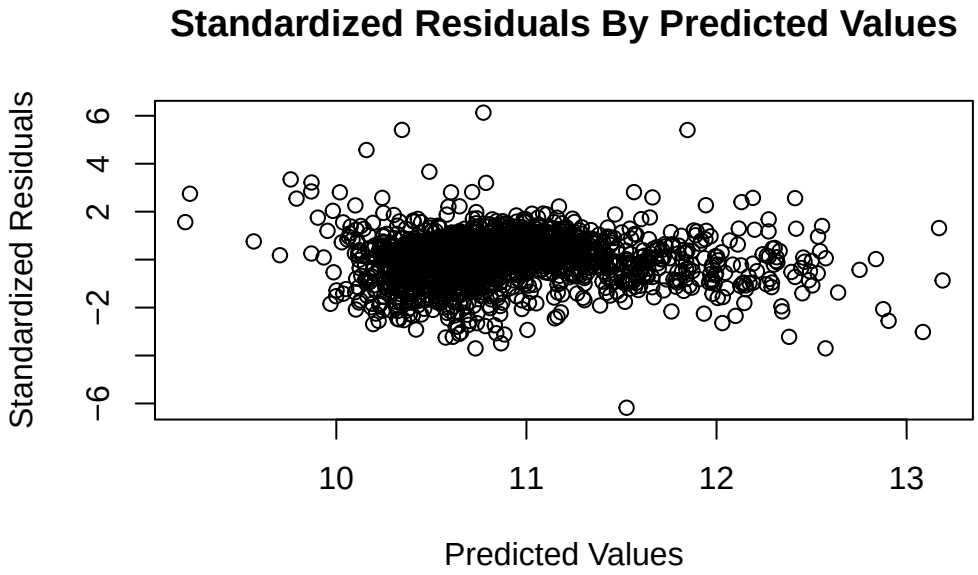
The skewedness of histograms of each predictor is reflected in the scatter plots of the predictors by the dependent variable median house value, MEDHVAL, as, as the predictor values increase, y values cluster towards lower values. The lack of linearity between the predictor and MEDHVAL and the skewed distribution of predictor values suggest that some type of nonlinear transformation may need to occur in order to normally distribute values and achieve linearity. We performed log transformations which are commonly used to correct the positively skewed distributions evident in our variables.

3.3.3



We applied logarithmic transformations to all predictors and the dependent variable to see whether the transformations would improve distribution of their values and subsequently allow us to assume linearity and residual normality. We only substituted the log-transformed MEDHVAL (LNMEDHVAL) and log-transformed NBELOPOV (LNNBELOPOV) for the rest of our analysis. The log-transformed predictors PCTVACANT, PCTSINGLES, and PCTBACHMOR did not return an improvement and instead produced zero-inflated distributions. We proceeded to calculate our standardized residuals with the new model of original predictors PCTVACANT, PCTSINGLES, and PCTBACHMOR and with our log-transformed predictor LNNBELPOV by our log transformed dependent variable LNMEDHVAL. The histogram of the standardized residuals show the normality in residuals needed per our assumption and support the need for the logarithmic transformations.

3.3.4



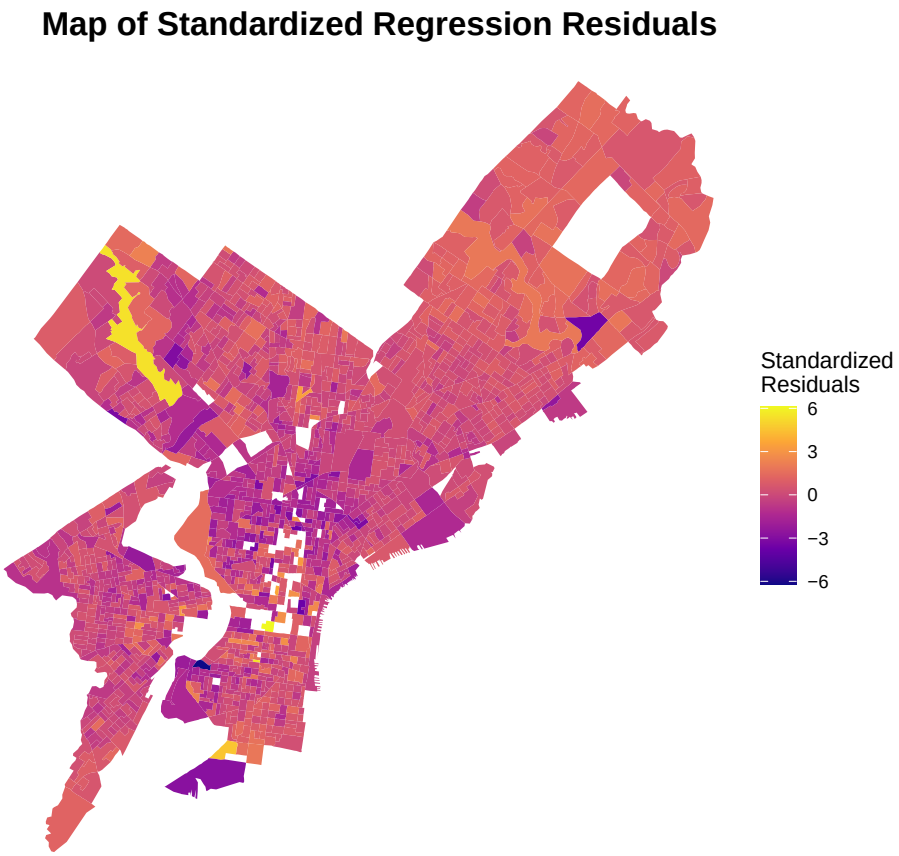
Standardized residuals are residuals divided by their standard deviation as a means to prime residuals across different observations for comparison through normalization. The scatter plot of our standardized residuals shows general homoscedasticity or consistent variance of residuals. There is general uniformity of the standardized residuals as most lie between -2 and positive 2. There are some outliers that extend past -4 and 4 but they do not dominate the overall pattern. There is also no funneling affect or any other pattern of non-constant variance. Thus, our model satisfies the assumption of homoscedasticity of residuals.

3.3.5

Initial spatial visualizations of the dependent and predictor variables suggest that there may be some spatial autocorrelation between their respective measurements. The choropleth map of the logged dependent variable LNMEDHVAL shows that lower values seem to be concentrated in parts of North, Southwest, and West Philadelphia while higher values were clustered in Upper North Philadelphia. The choropleth map of the predictor PCTSINGLES shows higher percentages in parts of Upper North and Northeast Philadelphia. The choropleth map of the predictor PCTBACHMOR shows higher percentages in parts of Upper North and Center Philadelphia. The choropleth map of the logged predictor LNNBELPOV showed lower values in parts of Upper North, Northeast, and Center Philadelphia. The choropleth map of the predictor PCTVACANT shows higher percentages in parts of North, West, Southwest, South,

and Center Philadelphia.This visual inspection suggests that block groups might not be entirely independent of each other and could require further spatial assessment.

3.3.6



The choropleth of standardized residuals suggests possible spatial autocorrelation as there seems to be a concentration of lower values in the southern half of Philadelphia. Visually, there seems to be a gradient effect stemming outward from North Philadelphia into West, Southwest, and South Philadelphia. This indicates that their could be additional factors producing systematic under prediction in the southern half of Philadelphia, especially in North Philadelphia.

3.4 Additional Models

3.4.1

Stepwise Regression ANOVA table

```
Start:  AIC=-3448.16
LNMEDHVAL ~ PCTVACANT + PCTSINGLES + PCTBACHMOR + LNNBELPOV

             Df Sum of Sq    RSS     AIC
<none>                     230.33 -3448.2
- PCTSINGLES  1     2.407 232.74 -3432.3
- LNNBELPOV   1    11.692 242.02 -3365.0
- PCTVACANT   1    51.543 281.87 -3102.8
- PCTBACHMOR  1   199.014 429.35 -2379.0


Stepwise Model Path
Analysis of Deviance Table

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

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


  Step Df Deviance Resid. Df Resid. Dev       AIC
1                       1715   230.3317 -3448.162
```

Our initial model before performing stepwise regression:

$$\text{LNMEDHVAL} \sim \text{PCTVACANT} + \text{PCTSINGLES} + \text{PCTBACHMOR} + \text{LNNBELPOV}$$

As mentioned earlier, stepwise regression based on AIC evaluates whether a predictor improves the model fit by reducing the AIC. Our initial model had an AIC of -3448.162. When PCTSINGLES was removed, the AIC increased to –3432.3. When LNNBELPOV was removed, the AIC increased to –3365.0. When PCTVACANT was removed, the AIC increased to –3102.8. When our last predictor PCTBACHMOR was removed, the AIC increased drastically to –2379.0. Since the removal of each predictor resulted in a higher AIC, all four initial predictors were retained in the final model. This suggests that the initial model was selected by stepwise regression as being a model that balances explanatory power and complexity.

3.4.2

K-fold Cross-validation Table

```
Linear Regression

1720 samples
   5 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 1376, 1376, 1376, 1376, 1376
Resampling results:

  RMSE       Rsquared   MAE
  0.3665376  0.6616172  0.2726298

Tuning parameter 'intercept' was held constant at a value of TRUE


Linear Regression

1720 samples
   3 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 1376, 1376, 1376, 1376, 1376
Resampling results:

  RMSE      Rsquared  MAE
  0.442709  0.506841  0.3182664

Tuning parameter 'intercept' was held constant at a value of TRUE
```

We performed 5 fold cross-validation on two models, the first model including all of our original predictors and the second model being a reduced set of predictors that alternatively included MEDHHINC as a predictor. The second model is as follows:

$$\text{LNMEDHVAL} \sim \text{PCTVACANT} + \text{MEDHHINC}$$

The original model yielded a RMSE of 0.368 while the reduced model yielded a RMSE of 0.443, signaling that the additional predictors in the full model had better predictive power compared to PCTVACANT and MEDHHINC alone.

4 Discussion & Limitations