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

Yiming Cao, Sujan Kakumanu, Angel Sanaa Rutherford

2025-10-15

1 Introduction

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

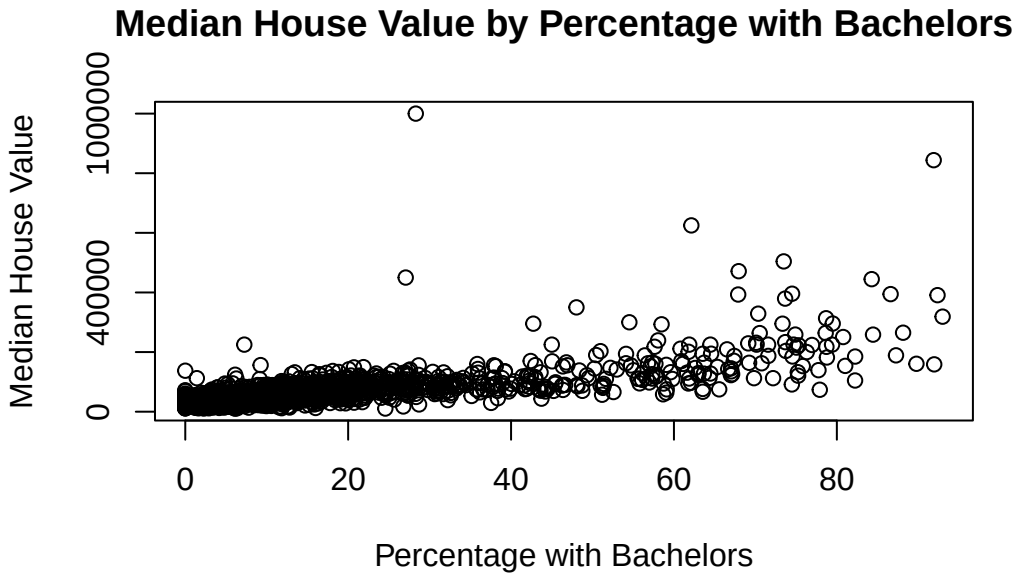
3.2 Regression Results

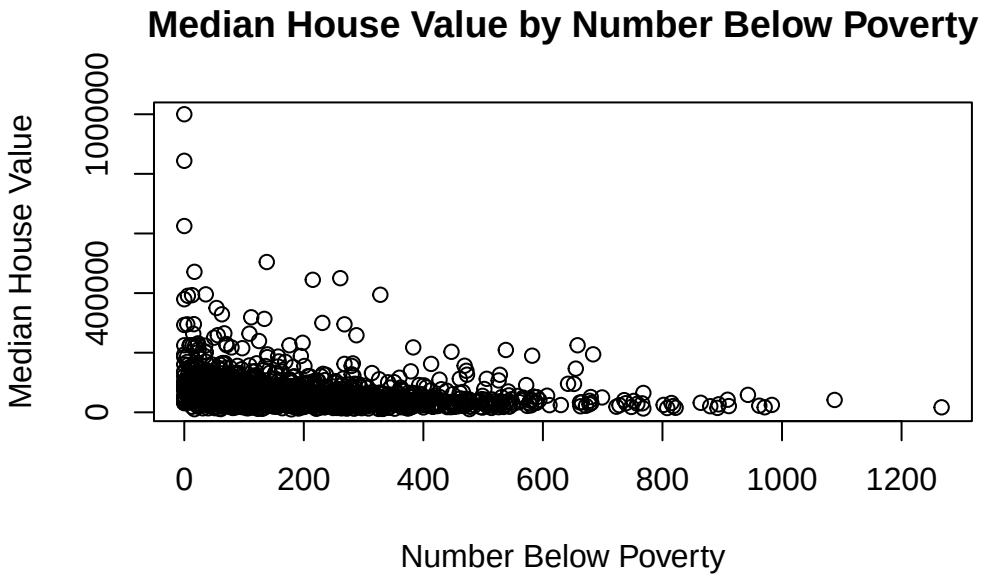
3.3 Regression Assumption Checks

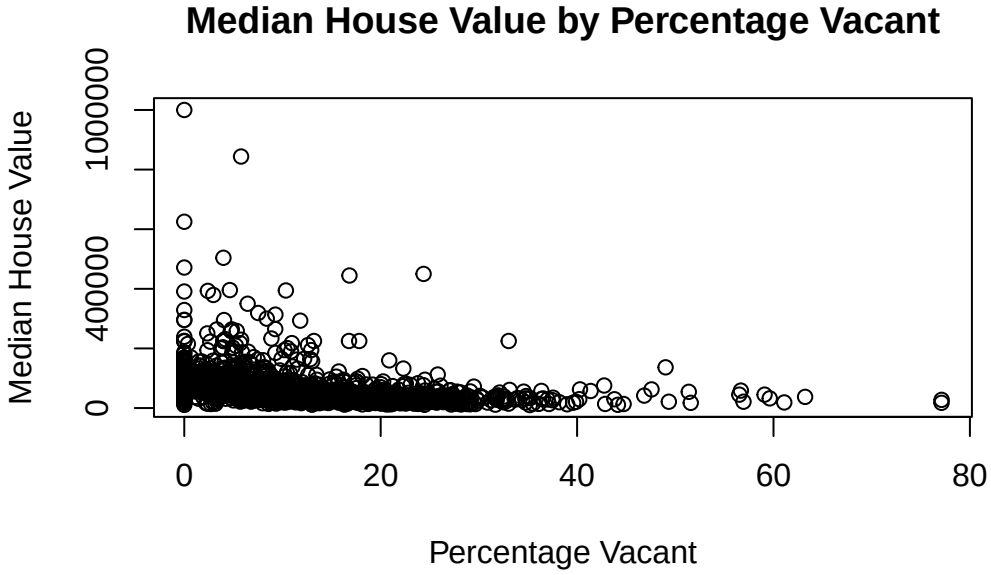
3.3.1

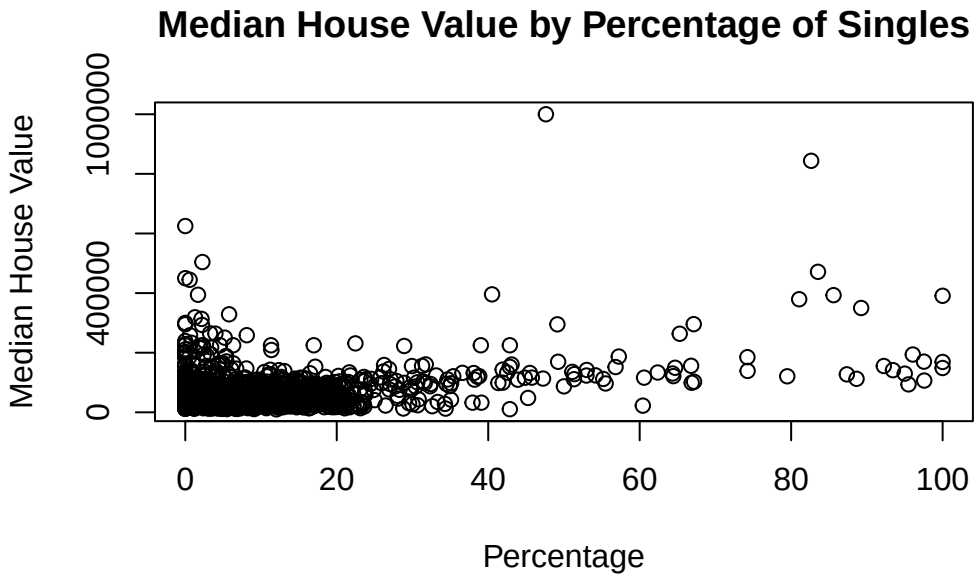
When assessing the relationship between various predictors and a dependent variable, we test whether the assumptions of multiple regressions are met to determine whether a linear model is appropriate for the data. The assumptions used to evaluate data include: a roughly linear relationship between each predictor and the outcome, observations that are independent and do not influence one another, homoscedasticity or consistent spread of residuals, the normality of residuals, especially for smaller samples, and no multicollinearity or linearity between predictors.

3.3.2

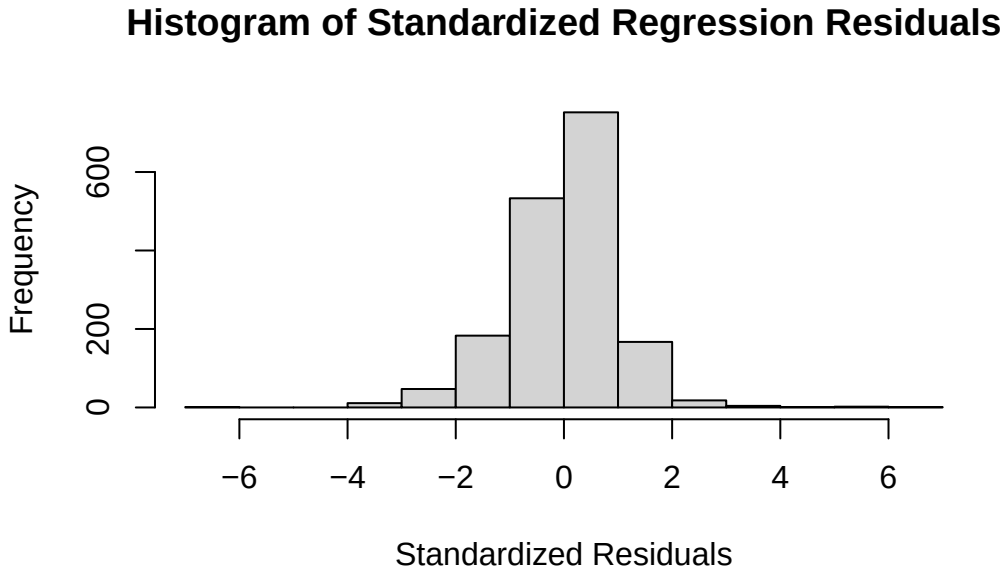



Median House Value by Number Below Poverty
Median House Value
0
400000
1000000
0
200
400
600
800
1000
1200
Number Below Poverty

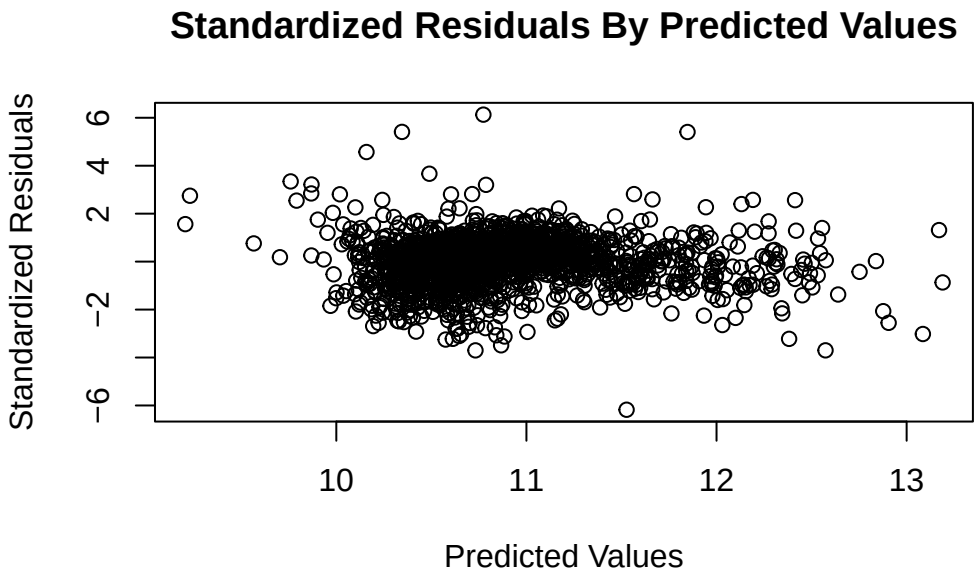

Median House Value by Percentage Vacant
Median House Value
0
400000
1000000
0
20
40
60
80
Percentage Vacant


Median House Value by Percentage of Singles
Median House Value
0
400000
1000000
0
20
40
60
80
100
Percentage

3.3.3

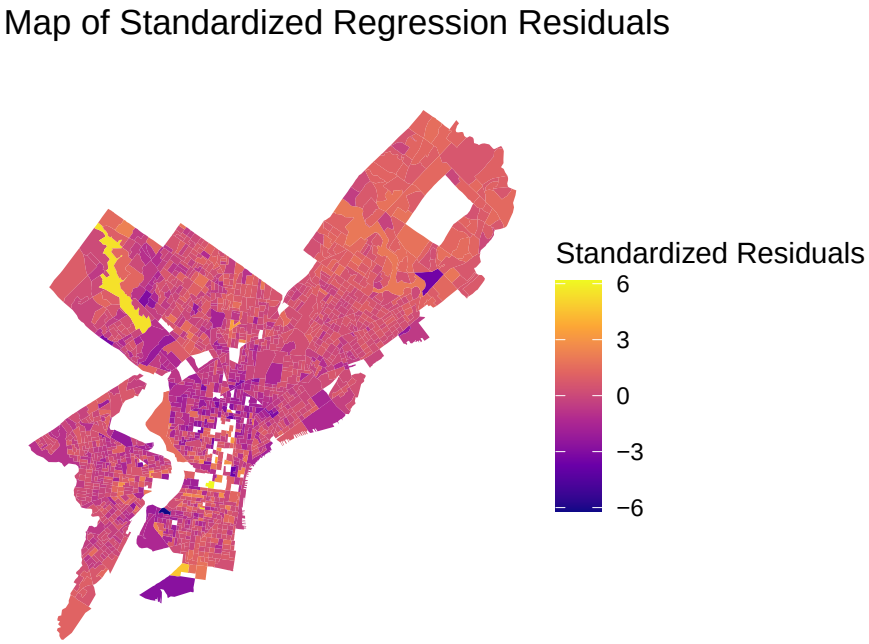

Histogram of Standardized Regression Residuals
Frequency
0
200
600
−6
−4
−2
0
2
4
6
Standardized Residuals

3.3.4


Standardized Residuals By Predicted Values
Standardized Residuals
6
4
2
−2
−6
10
11
12
13
Predicted Values

3.3.5

3.3.6



3.4 Additional Models

3.4.1

```
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

```
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
  0.367946  0.6619869  0.2737567

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
  0.4432052  0.5093347  0.3183268

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