STAT 3215Q: Final Project

Eleanor Kirkland

05/09/2025

1. Introduction

This regression analysis sought to develop the best model for predicting the median house value in a particular area of Boston. To do so, a data set derived from information collected by the U.S. Census Service in 1970 concerning housing in the Boston, Massachusetts area was used. The information was originally for census tracts in the Boston Standard Metropolitan Statistical Area (SMSA). The data set contains 506 entries, each with 13 attributes. Its features are outlined below:

Response variable:

- `price`: median value of owner-occupied homes in the census tract in thousands of dollars

Structural variables:

- `rooms`: average number of rooms
- `age`: proportion of owner units built prior to 1940

Neighborhood variables:

- `status`: logarithmic proportion of population that is lower status
- `crime`: crime rate by town
- `zoned`: proportion of town's residential land zoned for lots greater than 25,000 square feet
- `ind_acres`: proportion nonretail business acres per town
- `tax_rate`: full value property tax rate
- `pt_ratio`: pupil-teacher ratio by town school district
- `river`: Charles River dummy (= 1 if census tract bounds Charles River)

Accessibility variables:

- `distance`: logarithm of weighted distances to five employment centers in Boston region
- `highway`: logarithm of index of accessibility to radial highways

Pollution variable:

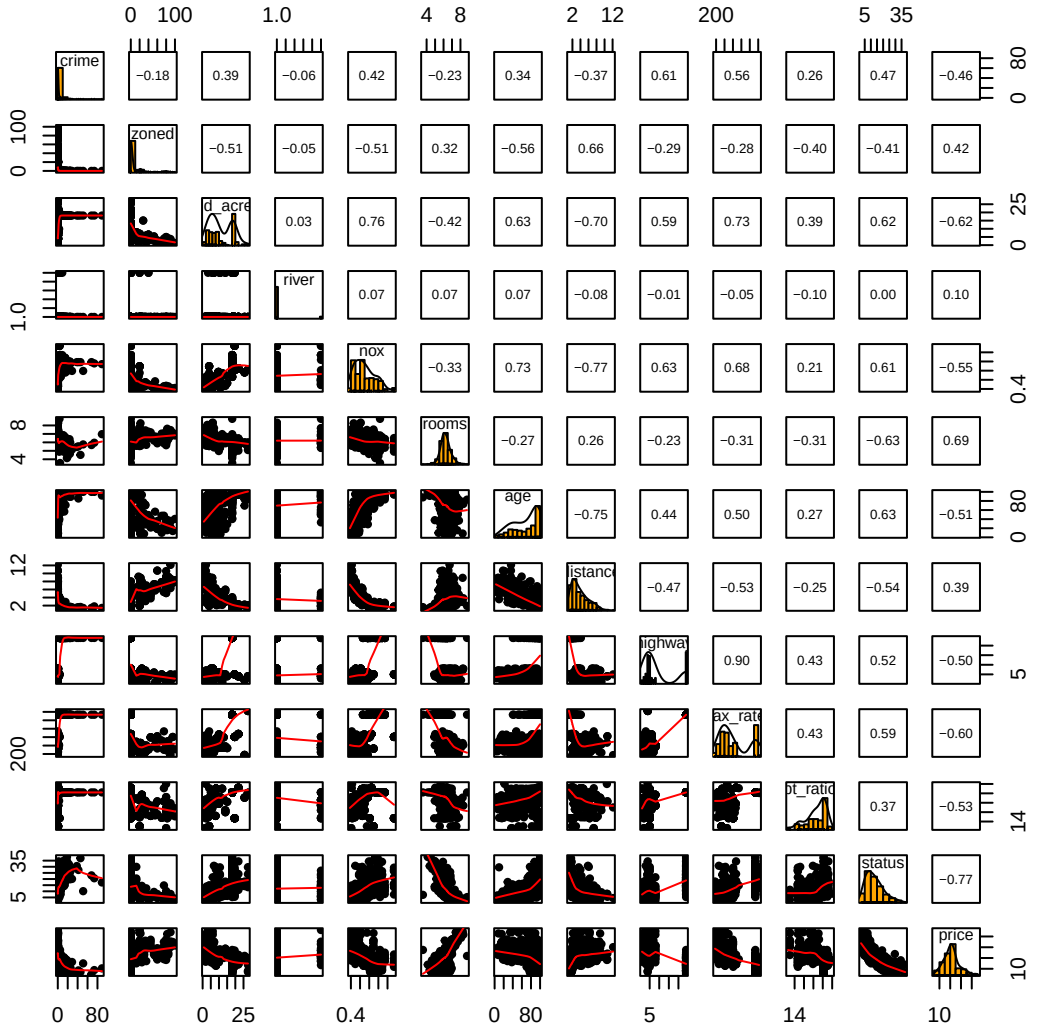
- `nox`: nitrogen oxide concentrations in pphm (annual average concentration in parts per hundred million)

2. Methodology

2A. Exploratory Analysis

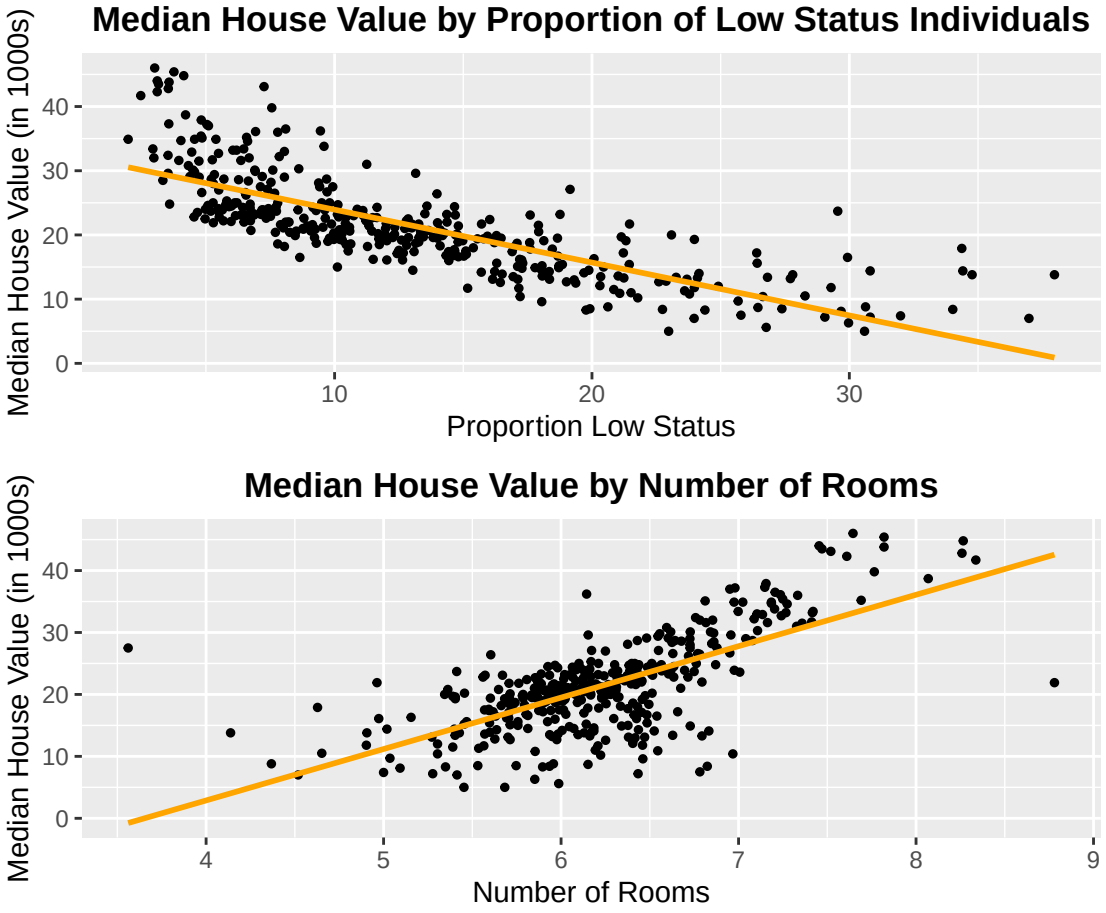
> Exploratory analysis gave insight into how the predictors are related to the response, thus guiding the more rigorous regression analysis.

The correlation matrix of the data set was first examined.



From this matrix, it was observed that the proportion of low status people (`status`) and the number of rooms (`rooms`) have the highest correlation with the median house value (`price`). This indicated that these two predictors may be most important in the prediction of median house value.

The marginal relationships between these two important predictors and median house value were be examined:



From these marginal plots, it appeared the proportion of low status individuals (`status`) was negatively correlated with the median house value (`price`), and it appeared the number of rooms (`rooms`) was positive correlated with the median house value (`price`). These relationships were explored further with regression analysis.

2B. Regression Methods

Simple Linear Regression

A simple linear regression model to predict median house value (`house_price`) was first constructed using proportion low status (`status`) as the single predictor.

$$E(\text{price}|X = x) = \beta_0 + \beta_1 \text{status}$$

The t-test on β_1 yielded a rejection of the null hypothesis $\beta_1 = 0$. Therefore, `status` provides useful information for predicting `price`. The R_2 value is 0.5677, meaning variability in `status` explains 57 percent of the variability in `price`.

A second simple linear regression model was fit using number of rooms (`rooms`) as the single predictor.

$$E(\text{price}|X = x) = \beta_0 + \beta_2 \text{rooms}$$

The t-test on β_2 yielded a rejection of the null hypothesis $\beta_2 = 0$. Therefore, the number of rooms (`rooms`) provides useful information for predicting the median house value (`house_value`). The R_2 value is 0.5654, meaning variability in `rooms` explains 57 percent of the variability in `price`.

In an effort to obtain a model that explains more of the variation in median house value (`price`) and accounts for interactions between predictors, a multiple linear regression model was constructed to predict `price`.

Multiple Linear Regression

A multiple linear regression model to predict `price` was constructed using `status`, `rooms`, `pt_ratio`, `tax_rate`, `highway`, `distance`, `age`, `nox`, `river`, `ind_acres`, `zoned`, and `crime` as predictors.

```
##
## Call:
## lm(formula = price ~ ., data = Housing.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8553 -2.3216 -0.3833  1.9067 12.8743
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  43.303858   5.157922   8.396 3.19e-15 ***
## crime        -0.098120   0.027837  -3.525 0.000502 ***
## zoned         0.039302   0.013862   2.835 0.004946 **
## ind_acres    -0.094760   0.064799  -1.462 0.144864
## river1        1.505136   0.949845   1.585 0.114289
## nox         -16.830959   4.152040  -4.054 6.69e-05 ***
## rooms         2.691245   0.469583   5.731 2.80e-08 ***
```

```
## age           0.003238   0.014588   0.222 0.824513
## distance     -1.254657   0.197745  -6.345 1.00e-09 ***
## highway       0.261272   0.063121   4.139 4.73e-05 ***
## tax_rate     -0.014167   0.003666  -3.865 0.000141 ***
## pt_ratio     -0.781460   0.135429  -5.770 2.28e-08 ***
## status       -0.467021   0.058451  -7.990 4.61e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.537 on 256 degrees of freedom
## Multiple R-squared:  0.802,  Adjusted R-squared:  0.7927
## F-statistic: 86.41 on 12 and 256 DF,  p-value: < 2.2e-16
```

The variability in the included predictors now accounted for 80 percent of the variability in `price`. This is a significant improvement from the initial simple linear regression models.

The **Overall F Test** for this multiple linear regression model was also considered. An Overall F Test determines if there is a regression relation between median house value (`price`) and the set of all regressors. In other words, it determines if the model is useful in predicting the median house value. The hypotheses for this test are as follows:

$$H_0 : \beta_1 = \cdots = \beta_{12} = 0$$

$$H_A : \text{at least one } \beta_j \neq 0 \text{ , j} = 1, \ldots, 12$$

The p-value associated with this Overall F Test was very small, so the null hypothesis (H_0) was rejected in favor of the alternative hypothesis (H_A). There is a regression relationship between median house value and the set of regressors.

Given that proportion of low status (`status`) and number of rooms (`rooms`) were most highly correlated with the median house value (`price`), a **General F Test** that tests the contribution of all remaining predictors to explaining the variability in median house value, given `status` and `rooms` are already included in the model was considered. The hypotheses for this test are as follows:

$$H_0 : \beta_3 = \cdots = \beta_{12} = 0$$

$$H_A : \text{at least one } \beta_j \neq 0 \text{ , j} = 3, \ldots, 12$$

```
## Analysis of Variance Table
##
## Model 1: price ~ status + rooms
## Model 2: price ~ crime + zoned + ind_acres + river + nox + rooms + age +
##     distance + highway + tax_rate + pt_ratio + status
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    266 5098.6
## 2    256 3201.9 10    1896.7 15.165 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value associated with this General F Test was very small, so the null hypothesis (H_0) was rejected in favor of the alternative hypothesis (H_A). At least one of the partial slopes corresponding to the remaining predictors is not equal to 0. Furthermore, the remaining predictors significantly improve the prediction of median house value achieved from proportion of low status and number of rooms.

The individual t-Tests indicate the non-significant regressors are `ind_acres` (proportion of non-retail business acres per town), `river` (proximity to Charles River), and `age`(proportion of owner units built preior to 1940). Another **General F Test** was conducted to test the contribution of `ind_acres`, `river`, and `age` to explaining the variability in median house value, given all other predictors are already included in the model. The hypotheses for this test are as follows[1]:

$$H_0 : \beta_3 = \beta_4 = \beta_7 = 0$$

$$H_A : \text{at least one } \beta_j \neq 0 \text{ , j} = 3, 4, 7$$

```
## Analysis of Variance Table
##
## Model 1: price ~ status + rooms + pt_ratio + tax_rate + highway + distance +
##     nox + zoned + crime
## Model 2: price ~ status + rooms + pt_ratio + tax_rate + highway + distance +
##     nox + zoned + crime + ind_acres + river + age
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    259 3255.1
## 2    256 3201.9  3    53.178 1.4172 0.2382
```

The p-value associated with this General F Test was quite large, so the null hypothesis (H_0) was not rejected. The partial slopes corresponding to `ind_acres`, `river`, and `age` do not differ from 0. In other words, `ind_acres`, `river`, and `age` do not significantly improve the prediction of median house value achieved from the predictors already included in the model. Thus, these three predictors were removed from the model and only 9 regressors were considered in the prediction of median house value.

[1]Note the order of the predictors was adjusted in order to conduct this General F Test.

2C. Regression Diagnostics

An **overall Outlier Test with Bonferroni Correction** was conducted in order to determine if the data set contains any outliers.

```
##     rstudent unadjusted p-value Bonferroni p
## 366 4.326737         2.1656e-05    0.0058254
```

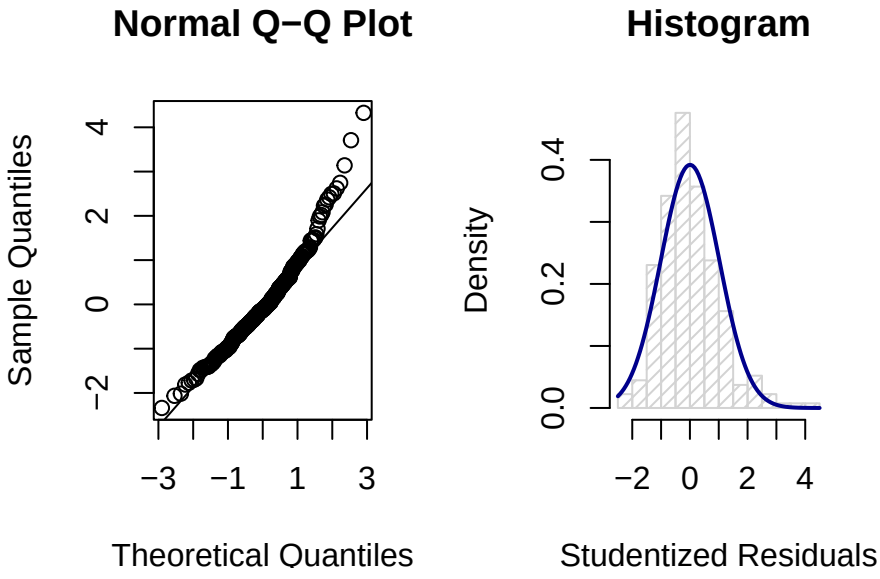
The Outlier Test identified one observation as an outlier. The multiple linear regression was conducted without this observation to determine if the observation significantly affected the previous findings. From examining the coefficients of the predictors and the residual standard errors, it was determined that the inclusion of the outlier does not affect the significance of individual regressors or the significance of the overall model. Therefore, the outlier remained in the data set.

Collinearity between regressors was investigated by examining the regressors' variance inflation factor (VIF) values.

```
##   status    rooms pt_ratio tax_rate  highway distance      nox    zoned
## 2.793875 1.785153 1.585291 6.413015 6.003095 3.278623 3.787645 2.218932
##    crime
## 1.646292
```

None of the predictors had a VIF value greater than or equal to 10. Therefore, none of the predictors are highly correlated with other predictor variables.

The **QQ-plot** and **histogram** were examined to investigate any non-normality of the errors.



From the QQ-plot, it was observed the residuals may have a right skew.

The **Shapiro-Wilk normality test** and **Anderson-Darling normality test** were conducted to determine if the normality of errors assumption holds.

```
##
##  Shapiro-Wilk normality test
##
## data:  ti
## W = 0.96346, p-value = 2.477e-06
```

```
##
##  Anderson-Darling normality test
##
## data:  ti
## A = 2.025, p-value = 3.632e-05
```

Since the p-value for each of these tests was very small, the null hypothesis was rejected; the errors do not follow a normal distribution.

The **Breuch-Pagan (BP) Test** was conducted in order to evaluate the assumption of the errors having equal variances.

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 2.322889, Df = 1, p = 0.12748
```

Since the p-value for this test was large, the null hypothesis was not rejected; the errors had a constant variance.

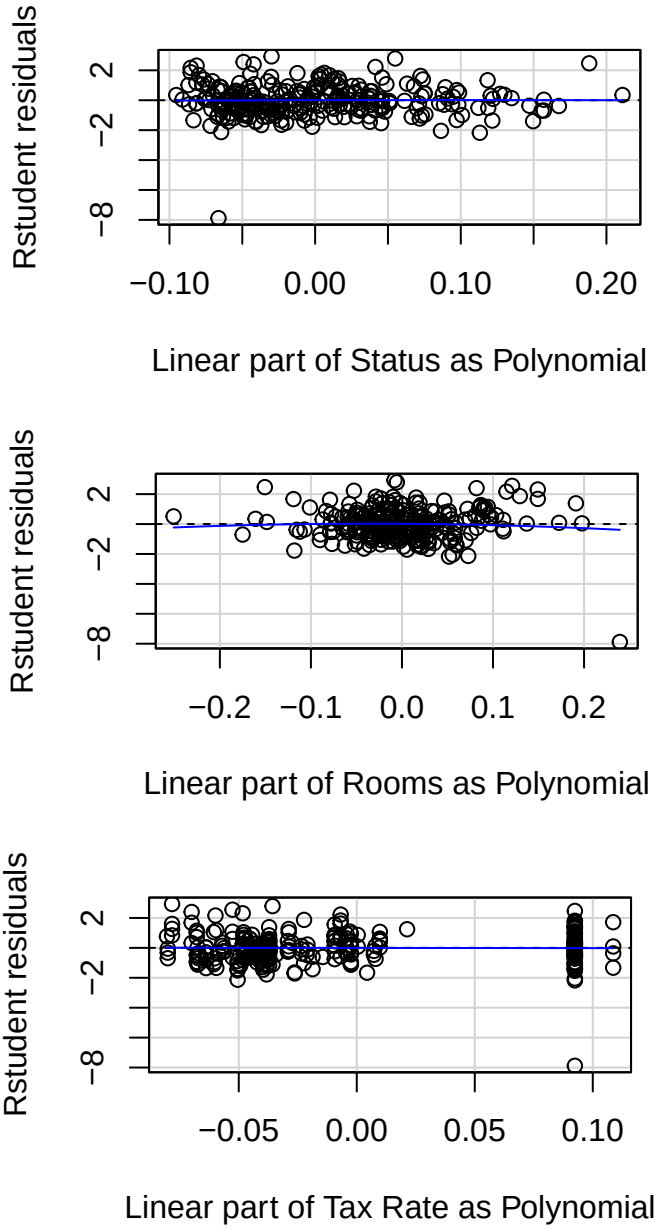
Tukey's Test for curvature/nonadditivity was conducted to determine if any of the regressors have a non-linear relationship with median house value (`price`).

```
##            Test stat Pr(>|Test stat|)
## status        5.4327        1.283e-07 ***
## rooms         8.3081        5.566e-15 ***
## pt_ratio      0.5436          0.58720
## tax_rate      2.4641          0.01439 *
## highway      -0.5218          0.60223
## distance      1.6704          0.09606 .
## nox          -0.0829          0.93399
## zoned         1.2220          0.22281
## crime         0.0329          0.97378
## Tukey test    8.9450        < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-values associated with the Tukey Tests for `status`, `rooms` and `tax_rate` were smaller than an $\alpha = 0.05$ significance level. Therefore, the relationship between each of these predictors and the median house value has a significant non-linear component. This suggested the inclusion of higher order terms of these regressors was warranted. Additionally, the p-value associated with the Tukey Test for nonadditivity was very small, which suggested the linear mean function specified in the multiple linear regression model is not sufficient.

2D. Regression Remedies

Including quadratic terms for `status`, `rooms`, and `tax_rate` was thought to remedy the linearity assumption. To do so, the multiple linear regression was fit with **orthogonal polynomials** that were functions of the original predictors.



The residual plots for `status`, `rooms`, and `tax_rate` now showed little curvature, indicating the linearity assumption seemed to be more satisfied.

In an attempt to address the non-normality of the errors, a **Box-Cox transformation** was applied to the `price`. However, this did not remedy the non-normality. A **bootstrap case resampling** was conducted to determine if the non-normality of the errors was severely affected the estimates of the regression coefficients. The confidence intervals of the bootstrap regression coeffients were much wider than those of the original model, so it appeared the analysis were affected by the non-normality.

Finally, it was observed that many observations had a median house value of exactly 50 and that these observations were skewing the distribution of the errors. These values were removed from the data set in an attempt to make the distribution more normal. While the null hypothesis of the Shapiro-Wilk normality test was still rejected, the p-value was less small and the QQ-plot showed a more normal distribution. This was deemed sufficient because this model was being constructed to *predict* median house values in Boston neighborhoods (rather than to make statistical inferences), so it was not imperative that median house values followed a normal distribution.

2E. Model/Variable Selection

The **full model** obtained from the previous sections was compared to candidate submodels. To clarify, the full model is `price ~ f(status) + f(rooms) + pt_ratio + f(tax_rate) + highway + distance + nox + river + zoned + crime`, where `f(status)`, `f(rooms)`, `f(pt_ratio)`, and `f(tax_rate)` are orthogonal functions of the corresponding original predictors.

To best assess validity of each candidate submodel, the data was partitioned into a training subset and a testing subset. These training and testing subsets were used in two variable selection techniques. The models obtained from these techniques were compared using an out-of sample metric.

First, an **Elastic Net regularization** was performed. Cross validation was performed to select the optimal tuning parameter λ.

```
## 14 x 1 sparse Matrix of class "dgCMatrix"
##                                                      s1
## (Intercept)                                  32.69986245
## poly(status, degree = 2, raw = FALSE)1      -51.85771103
## poly(status, degree = 2, raw = FALSE)2        4.82643754
## poly(rooms, degree = 2, raw = FALSE)1        31.88639814
## poly(rooms, degree = 2, raw = FALSE)2        19.00218662
## pt_ratio                                     -0.54681170
## poly(tax_rate, degree = 2, raw = FALSE)1 -15.68717425
## poly(tax_rate, degree = 2, raw = FALSE)2      4.30474147
## highway                                       .
## distance                                      .
## nox                                          -1.71077946
## river1                                        .
## zoned                                         .
## crime                                        -0.04628298
```

The Elastic Net regularization selected proportion low status (`status`), number of rooms (`rooms`), pupil to teacher ratio (`pt_ratio`), property tax rate (`tax_rate`), nitrogen oxide concetrations (`nox`), and crime rate by town (`crime`) as the set of active predictors. It selected the quadratic terms of `status`, `rooms`, and `tax_rate`.

Next, **SCAD** was implemented.

```
##                               (Intercept)
##                               45.57670515
##   poly(status, degree = 2, raw = FALSE)1
##                              -54.66773316
##   poly(status, degree = 2, raw = FALSE)2
##                                7.81815323
##    poly(rooms, degree = 2, raw = FALSE)1
##                               29.40320814
##    poly(rooms, degree = 2, raw = FALSE)2
##                               22.87721470
##                                  pt_ratio
##                               -0.73031964
## poly(tax_rate, degree = 2, raw = FALSE)1
##                              -33.90688314
## poly(tax_rate, degree = 2, raw = FALSE)2
##                                7.06592506
##                                   highway
##                                0.17949933
##                                  distance
##                               -0.96356996
##                                       nox
##                              -15.35840628
##                                    river1
##                                0.80037120
##                                     zoned
##                                0.02725878
##                                     crime
##                               -0.11625092
```

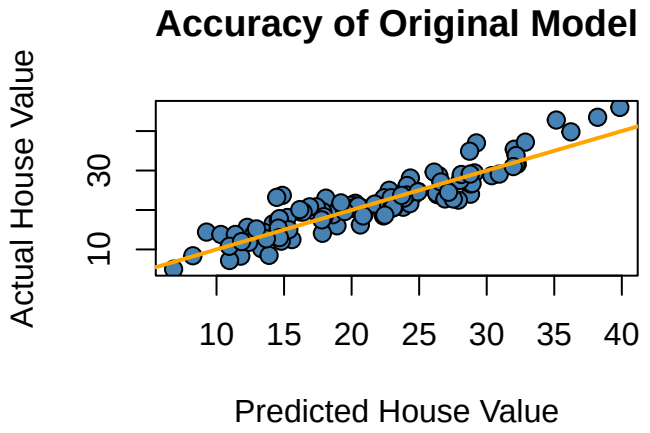
SCAD did not remove any predictors from the set of active predictors.

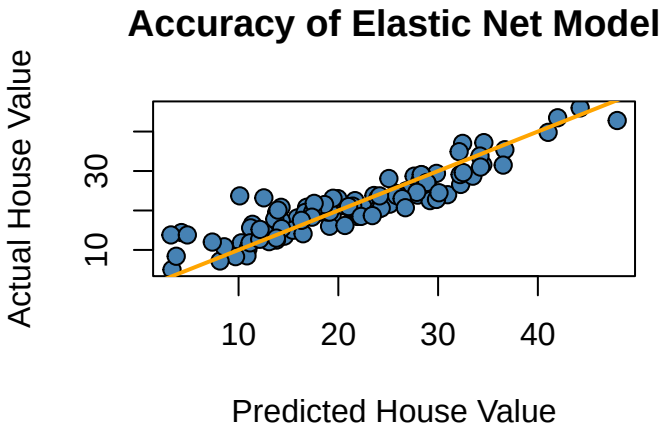
The **mean squared prediction error (MSE)** values of the full model, the Elastic Net model, and the SCAD model were compared in order to assess the predictive ability of each model.

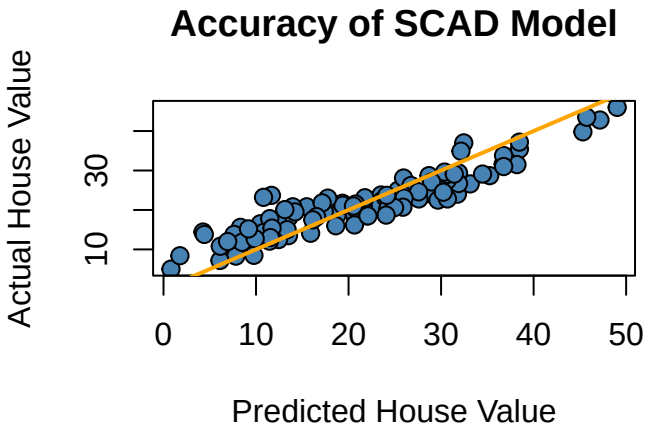
```
## Original      Net     SCAD
##  9.06955 14.31062 18.81622
```

The full model constructed before performing any variable selection had the lowest MSE value. This indicated the full model performs best at predicting median house value (`price`) based on new predictor values. Elastic Net and SCAD may have too aggressively shrunk important predictors to zero. Those regularization methods did provide more simple models, but given the purpose of this analysis is to develop a model with the highest prediction accuracy, the full model is favored over the simpler models.

Plots of each model's predictions of house value versus the actual house values were observed to confirm the selection of the original model.







It was observed the strongest linear relationship between predicted values and actual values was in the original model's plot. This confirmed the selection of the original model as the most suitable model for prediction of median house values.

2F. Statistical Tests

The statistical tests performed throughout the analysis are summarized again here for completeness.

Regression Methods

In the Simple Linear Regression section, an **individual t-test** on the coefficient of `status` was conducted. This test determined whether the coefficient of `status` was different from zero. The null hypothesis that the coefficient is zero was rejected, indicating `status` provides useful information for predicting `price`. Another individual t-test was conducted on the coefficient of `rooms`. The null hypothesis for this test was rejected, indicating `rooms` provides useful information for predicting `price`. These tests determined two predictors individually provide useful information for predicting median house value. In the Multiple Linear Regression section, an **Overall F Test** of the multiple linear regression model was conducted. The p-value associated with this test was very small, so the null hypothesis that all of the regression coefficients are zero was rejected. This test determined there was a regression relationship between median houes value and the set of regressors. Therefore, it determined the predictors included in the dataset provide useful information for the prediction of median house value. A **General F Test** was conducted to test the contribution of all remaining predictors to explaining the variability in `price`, given `status` and `rooms` are already included in the model. The null hypothesis was rejected, indicating at least one of the partial slopes corresponding to the remaining predictors was not equal to 0. This test demonstrated that the remaining predictors should be included in the model for predicting median house value. Another **General F Test** was conducted to test the contribution of `ind_acres` and `age` to explaining the variabiltiy in `price`, given all other predictors are already included in the model. The null hypothesis of this test was not rejected, indicating the partial slopes corresponding to `ind_acres` and `age` did not differ from 0. This test demonstrated that `ind_acres` and `age` were not important in the prediction of median house value.

Regression Diagnostics

An **overall Outlier Test with Bonferroni Correction** was conducted to determine if the data set contained any outliers. One data point was determined to be an outlier, as it had a very small Bonferroni p-values. This test revealed a point that could have influenced the building and analysis of the model. Two **normality tests** were conducted to determine if the normality of errors assumption held. The null hypothoses of these tests were rejected, indicating the errors did not follow a normal distribution. This test indicated an assumption of the model was violated. The **Breuch-Pagan (BP) Test** was conducted to evaluate the assumption of the errors having constant variance. The null hypothesis of this test was not rejected, indicating the errors had a constant variance. This test demonstrated an assumption of the model was violated. **Tukey's Test** was conducted for each predictor. The null hypotheses associated with `status` `rooms`, and `tax_rate` were rejected, indicating these predictors have a non-linear relationship with the response `price`. This test revealed a better prediction of median house value could be achieved by including higher-order terms.

3. Conclusion

This analysis determined that the following predictors are useful in the prediction of median house value: proportion low status, average number of rooms, pupil-teacher ratio, property tax rate, accessibility to radial highways, distance to employment centers, nitrogen oxide concentrations, proprtion of residential land zoned for large lots, and crime rate. This analysis also determined proportion low status, number of rooms, and property tax rate have a nonlinear relationship with median house value. This suggests median house value is impacted differently depending on the values of these predictors. This analysis captured these non-linear relationships by developing a linear model with higher-order terms. This model was able to accurately predict median house values given unseen predictor values. This model could be used by economists and policymakers to predict how proposed changes to any of the predictors would affect the median house value in a given area of Boston. For instance, if policymakers were to propose a new property tax rate for Dorchester, this tax rate could be fed to the model in order to determine how the median house value in Dorchester would be affected by this change in tax rate. In general, this linear regression model is a useful tool in predicting how changes to the makeup of Boston neighborhoods will affect the value of homes in those neighborhoods.