Project 1

Problem 1: Boston Housing Data

Descriptive Summary

This data set represents a sample of 506 entries representing about 14 features for homes from various suburbs in Boston, MA during 1978. For this analysis, I have elected to focus on four specific variables: *medv*, *lstat*, *rm*, and *crim*. *medv* is the median value of owner-occupied homes. It has a mean of $22.53k, a median of $21.20k, and a range of $45.00k. *lstat* represents the lower status of the population, as a percentage. It has a mean of 12.65%, a median of 11.36%, and a range of 36.24%. *rm* represents the average number of rooms per dwelling. It has a mean of 6.25, a median of 6.208, and a range of 5.219. *crim* is the per capita crime rate by town. It has a mean of 3.61352, a median of 0.25651, and a range of 88.96988. For more information regarding the summary statistics, please reference Figure 1.

medv	lstat	rm	crim
Min. : 5.00	Min. : 1.73	Min. :3.561	Min. : 0.00632
1st Qu.:17.02	1st Qu.: 6.95	1st Qu.:5.886	1st Qu.: 0.08205
Median :21.20	Median :11.36	Median :6.208	Median : 0.25651
Mean :22.53	Mean :12.65	Mean :6.285	Mean : 3.61352
3rd Qu.:25.00	3rd Qu.:16.95	3rd Qu.:6.623	3rd Qu.: 3.67708
Max. :50.00	Max. :37.97	Max. :8.780	Max. :88.97620

Figure 1: Summary Statistics

Figures 2 & 3 depict the distributions of the four variables of interest using boxplots and histograms. *medv* is nearly symmetric with 38 upper outliers. *lstat* is slightly positively (right) skewed with 7 outliers. *rm* is approximately symmetric with 8 lower outliers and 22 upper outliers. *crim* is extremely positively (right) skewed with 66 upper outliers. All four variables are unimodal.

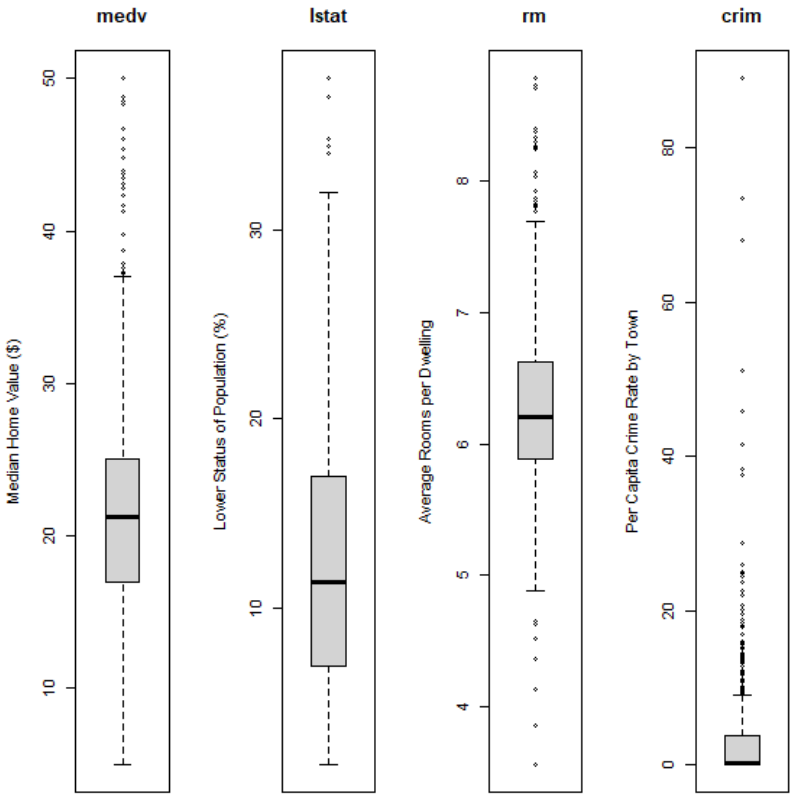

Figure 2: Boxplots

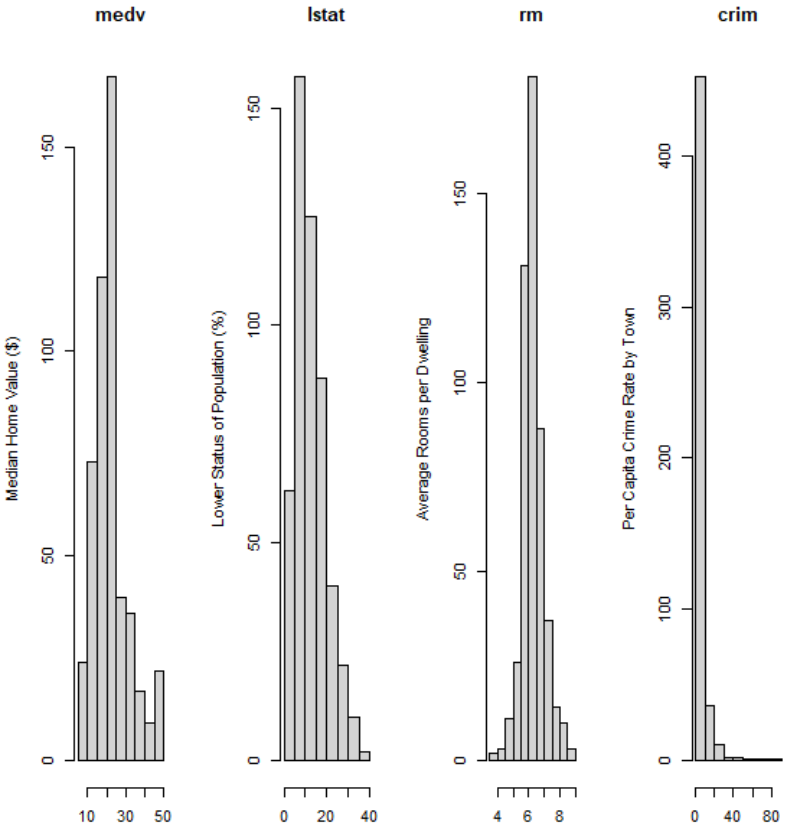

Figure 3: Histograms

Inferential Analysis

Looking at the best model selection in Figure 4, we are presented with three options for predicting *medv*. These options provide the best model based on the number of predictors. For a model using only one predictor, *lstat* will provide the best model (Option 1). For a model using two predictors, *lstat* and *rm* will be the best options (Option 2). The third and final option would be to use all three predictors (Option 3). For these models' inferential analyses, I decided to conduct both confidence intervals and hypothesis tests. Both the confidence intervals and hypothesis tests use a confidence level of 95% ($\alpha = 0.05$).

```
Subset selection object
Call: regsubsets.formula(medv ~ ., data = data)
3 variables  (and intercept)
      Forced in Forced out
lstat     FALSE      FALSE
rm        FALSE      FALSE
crim      FALSE      FALSE
1 subsets of each size up to 3
Selection Algorithm: exhaustive
         lstat rm  crim
1  ( 1 ) "*"   " " " "
2  ( 1 ) "*"   "*" " "
3  ( 1 ) "*"   "*" "*"
```

Figure 4: Best Model Selection

For Option 1, I created a simple linear regression (Figure 7) which yielded a regression equation of *medv* = 34.55384 - 0.95005 x *lstat*. The confidence interval for this regression was (23.20955, 24.31302). This means there is a 95% chance that the median home value is between 23.20955 and 24.31302. The null hypothesis was $H_0 : \beta_{lstat} = 0$, the alternative hypothesis was $H_A : \beta_{lstat} \neq 0$, the F-statistic was 601.6, and the p-value was 2.2×10^{-16}. Based on the results of the test, we reject the null hypothesis because the p-value of 2.2×10^{-16} is less than the alpha of 0.05. There is statistically significant evidence that suggests β_{lstat} is different from 0. When *lstat* takes its median value, *medv* has a value of 23.76128 (Figure 6), which deviates significantly from the sample median of 21.20 (See Figure 1). Remember that *lstat* was considered to be the best option for a model using only one predictor. This indicates that none of the predictors alone are useful in predicting the response. In order to get a better prediction, additional predictors will have to be added.

```
     fit      lwr      upr
23.76128 23.20955 24.31302
```

Figure 5: Confidence Interval for Option 1

```
     fit      lwr      upr
23.76128 11.53683 35.98573
```

Figure 6: Prediction Interval for Option 1

```
Call:
lm(formula = medv ~ lstat, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-15.168  -3.990  -1.318   2.034  24.500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,    Adjusted R-squared:  0.5432
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

Figure 7: Summary of the Simple Linear Regression in Option 1

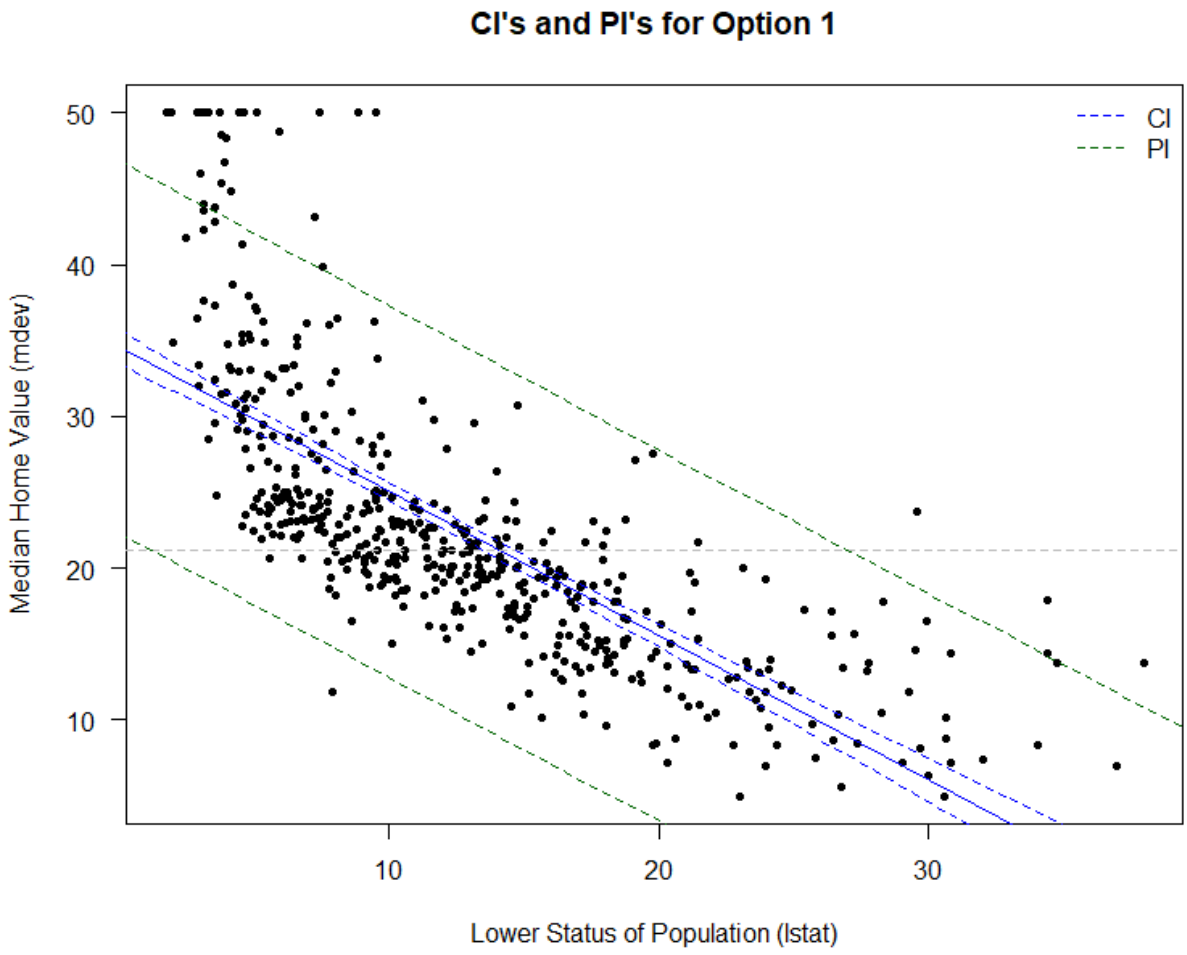

Figure 8: Graphed Confidence Intervals and Prediction Intervals for Option 1

For Option 2, I created a multiple linear regression (Figure 10) which yielded a regression equation of *medv* = -1.35827 - 0.64236 x *lstat* + 5.09479 x *rm*. The confidence interval for this regression was (22.46564, 23.48541). This means there is a 95% chance that the median home value is between 22.46564 and 23.48541. The null hypothesis was H_0 : $\beta_{lstat} = \beta_{rm} = 0$, the alternative hypothesis was H_A : at least one of the regression coefficients $\neq 0$, the F-statistic was 444.3, and the p-value was 2.2×10^{-16}. Based on the results of the test, we reject the null hypothesis because the p-value of 2.2×10^{-16} is less than the alpha of 0.05. There is statistically significant evidence that suggests at least one of β_{lstat} , β_{rm} is different from 0. When *lstat* and *rm* take their median values, *medv* has a value of 22.97553, which (compared to Option 1) deviates a little less from the sample median of 21.20.

```
       fit      lwr      upr
22.97553 22.46564 23.48541
```

Figure 9: Confidence Interval for Option 2

```
Call:
lm(formula = medv ~ lstat + rm, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-18.076  -3.516  -1.010   1.909  28.131

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.35827    3.17283  -0.428    0.669
lstat       -0.64236    0.04373 -14.689   <2e-16 ***
rm           5.09479    0.44447  11.463   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.54 on 503 degrees of freedom
Multiple R-squared:  0.6386,    Adjusted R-squared:  0.6371
F-statistic: 444.3 on 2 and 503 DF,  p-value: < 2.2e-16
```

Figure 10: Summary of the Multiple Linear Regression in Option 2

For Option 3, I created a multiple linear regression (Figure 13) which yielded a regression equation of *medv* = -2.56225 - 0.57849 x *lstat* + 5.21695 x *rm* - 0.10294 x *crim*. The confidence interval for this regression was (22.70073, 23.75768). This means there is a 95% chance that the median home value is between 22.70073 and 23.75768. The null hypothesis was H_0 : $\beta_{lstat} = \beta_{rm} = \beta_{crim} = 0$, the alternative hypothesis was H_A : at least one of the regression coefficient $\neq 0$, the F-statistic was 305.2, and the p-value was 2.2×10^{-16}. Based on the results of the test, we reject the null hypothesis because the p-value of 2.2×10^{-16} is less than the alpha of 0.05. There is statistically significant evidence that suggests at least one of β_{lstat} , β_{rm} , β_{crim} is different from 0. When *lstat*, *rm*, and *crim* take their median values, *medv* has a value of 23.22921, which

significantly deviates from the sample median of 21.20. Therefore, adding a third predictor into the model did not prove to make the prediction any more accurate.

```
     fit      lwr      upr
23.22921 22.70073 23.75768
```

Figure 11: Confidence Interval for Option 3

```
     fit      lwr      upr
23.22921 12.43093 34.02749
```

Figure 12: Prediction Interval for Option 3

```
Call:
lm(formula = medv ~ ., data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-17.925  -3.566  -1.157   1.906  29.024

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.56225    3.16602  -0.809  0.41873
lstat       -0.57849    0.04767 -12.135  < 2e-16 ***
rm           5.21695    0.44203  11.802  < 2e-16 ***
crim        -0.10294    0.03202  -3.215  0.00139 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.49 on 502 degrees of freedom
Multiple R-squared:  0.6459,    Adjusted R-squared:  0.6437
F-statistic: 305.2 on 3 and 502 DF,  p-value: < 2.2e-16
```

Figure 13: Summary of the Multiple Linear Regression in Option 3

Conclusion

As was demonstrated in the boxplots and histograms, we saw that the predictor and response variables exhibited a combination of approximately symmetric and right-skewed distributions. There were also quite a bit of outliers in each variables, most of which were upper outliers. These outliers gave the data an inherent, positive (right) bias, which caused me to adopt median as the preferred measure of central tendency when doing any calculations, as it is not as heavily influenced by outliers. I noticed a few things regarding the inferential analysis. First, there was no lone predictor that was useful in predicting the response. Therefore, a simple linear regression is not an appropriate model to use in this data set. The integration of all predictors into a model also were not useful in predicting the response. Many people think that adding more variables to a model makes the model more accurate. However, the results from Option 3 do not support this notion. Believe it or not, the best prediction model was actually Option 2, which used two predictors: *lstat* and *rm*. Overall, Option 2 did a reasonable job at fitting the data. The prediction interval tightened up using this model. On the other hand, there was still a little more variation in the prediction interval than I would have liked to see. Something else worth noting is that Option 2 had an R-squared of 0.6371. In general, an R-squared of this size is considered to be moderate. However, it’s really impressive when you think about it because the model only used two predictors.

Looking at the normal quantile plot in Figure 16, the patterns of the plots run roughly close to that of a straight line along the trend line. Also, the histogram of residuals in Figure 15 is

approximately normally distributed. These two factors make it reasonable to assume normality in the data. The residuals in Figure 14 appear to be randomly distributed, making the assumption of equal variances valid. Figure 17 displays a variance inflation factor (VIF) calculation to assess whether multicollinearity exists between the three predictor variables. Given the calculated values of 1.941883, 1.616468, and 1.271372 for *lstat*, *rm*, *crim*, respectively, we can conclude that there does not appear to be multicollinearity. The correlation between these three variables is not severe enough to warrant any corrective measures.

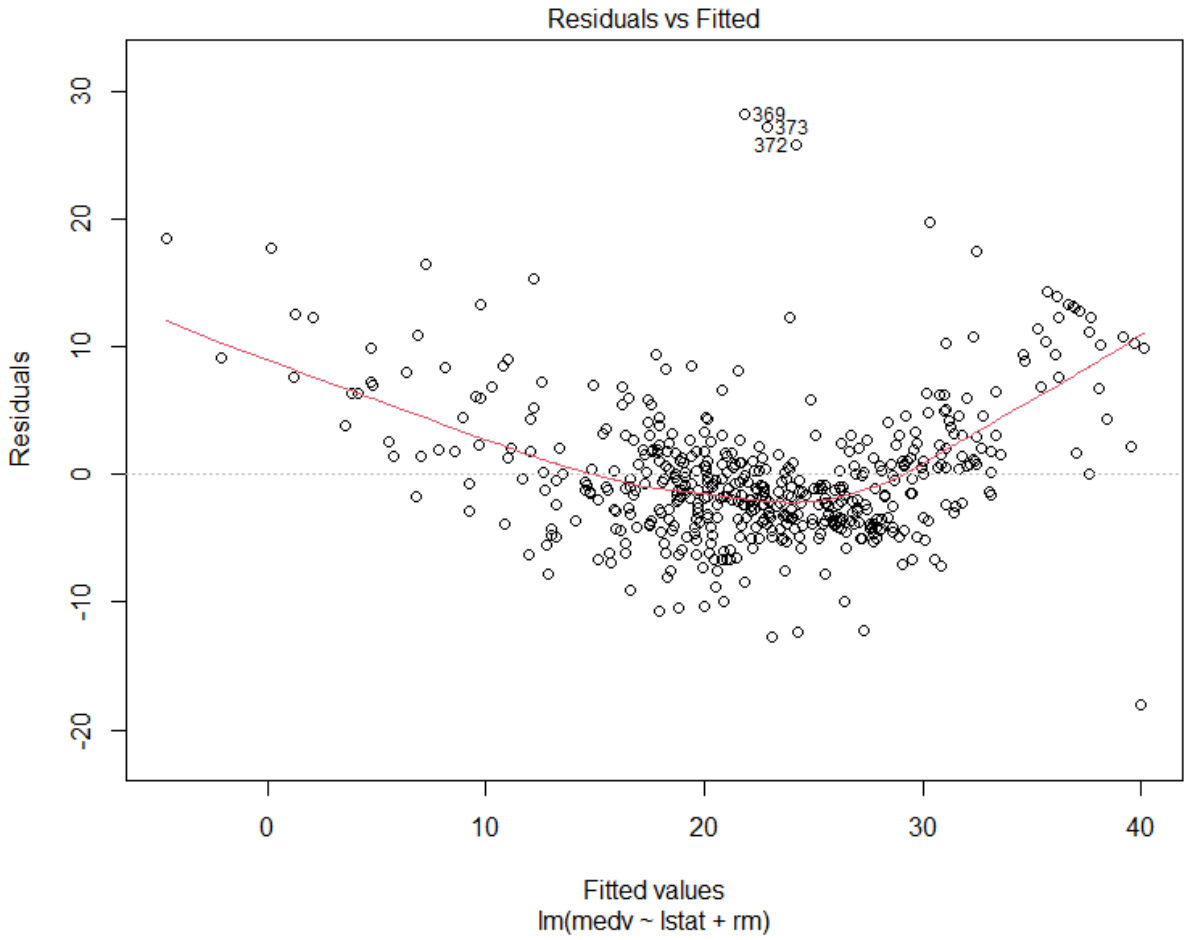

Figure 14: Residual Plot

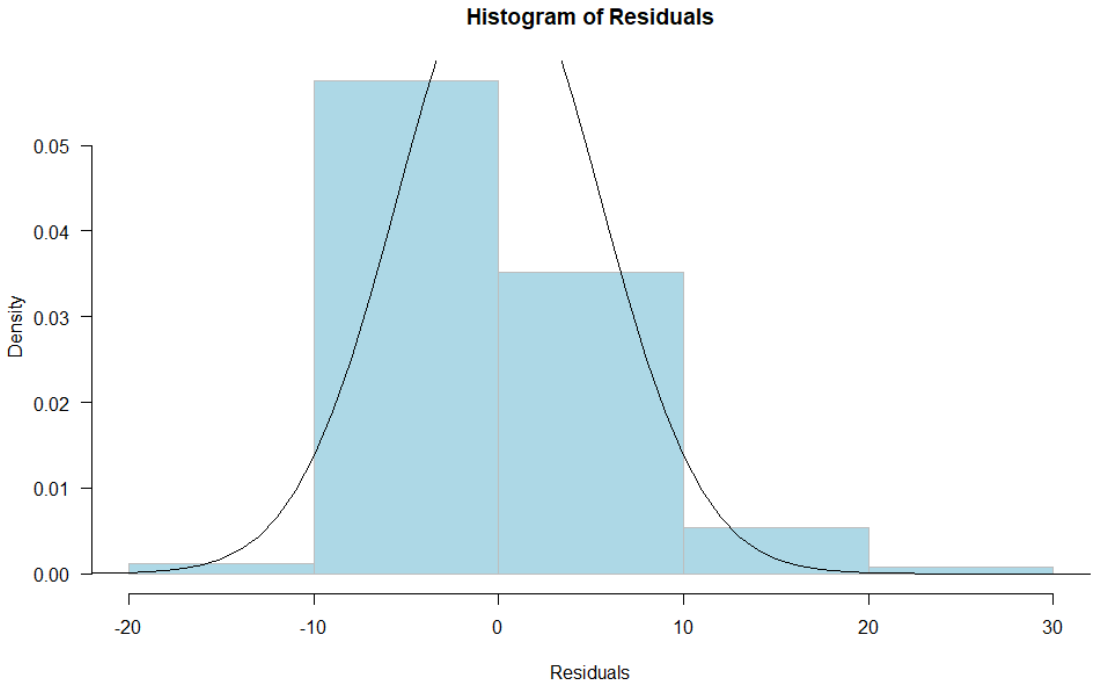

Figure 15: Histogram of Residuals

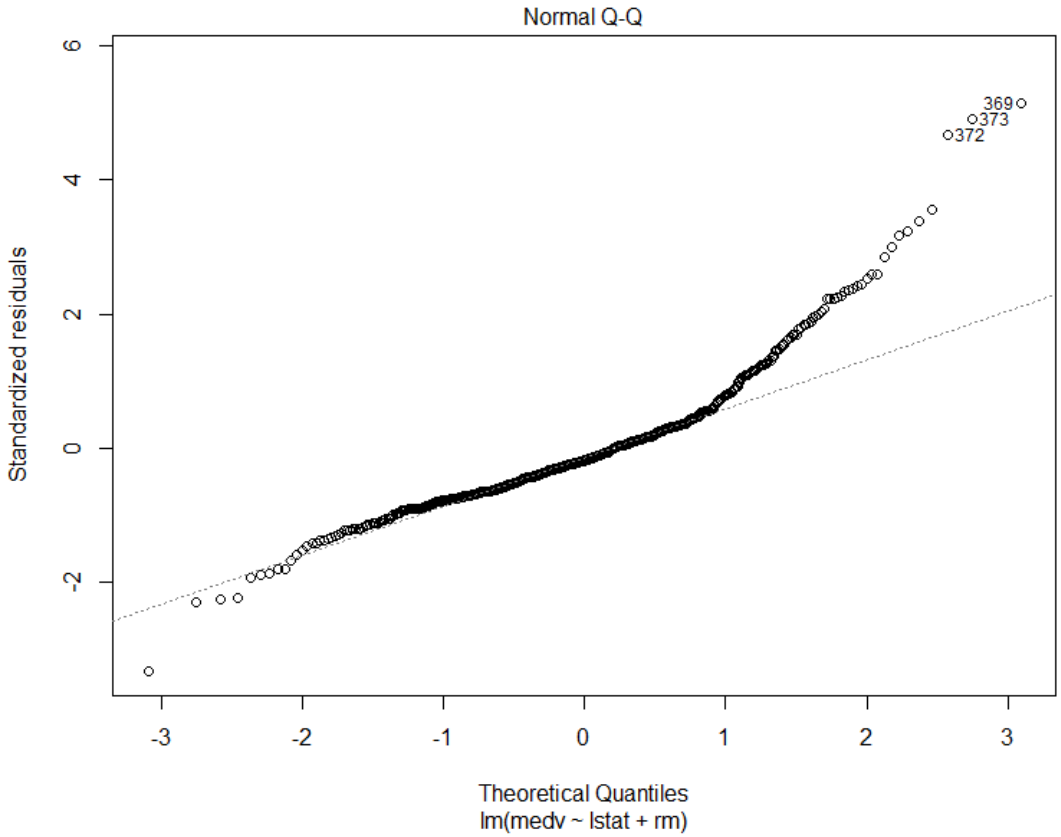

Figure 16: Normal Q-Q Plot

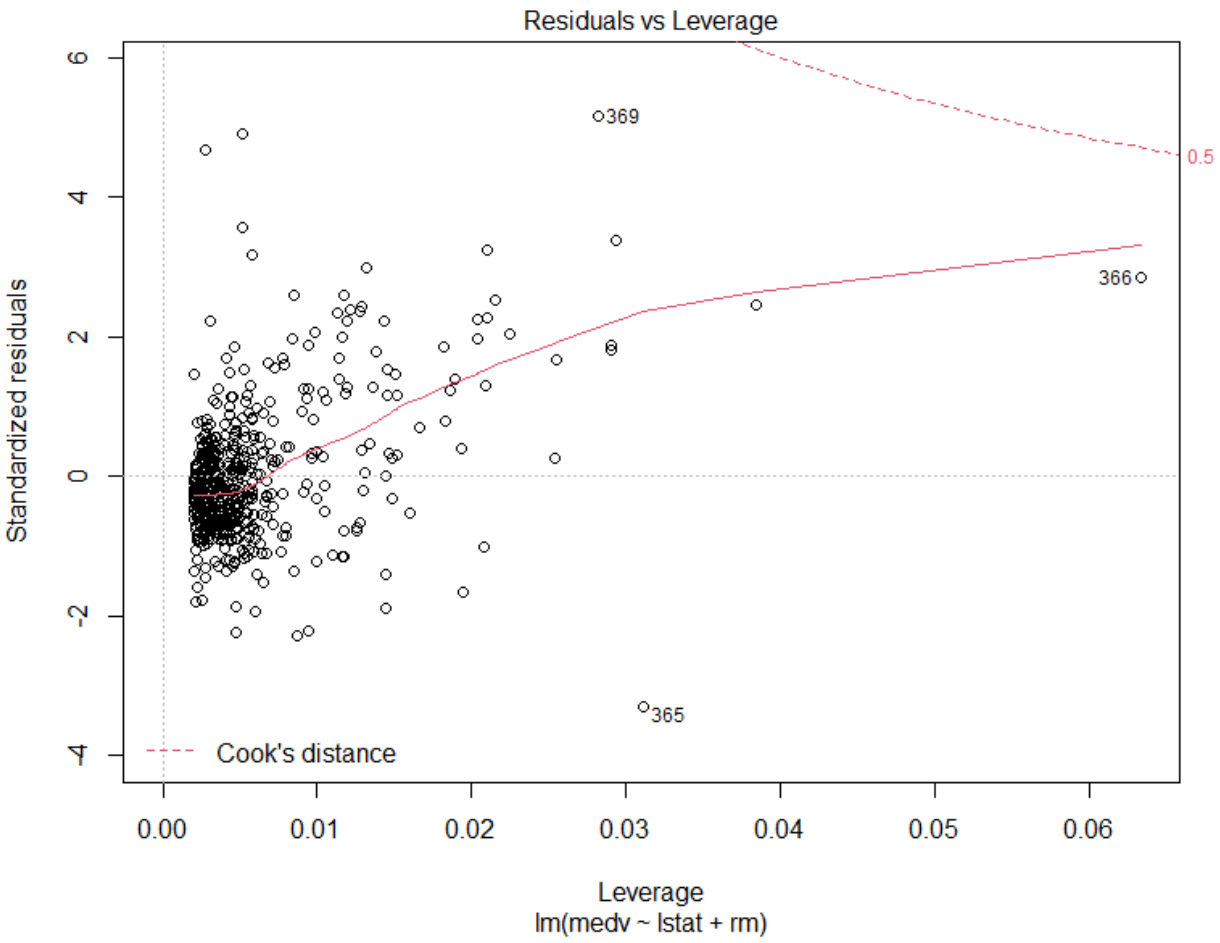

Figure 16: Leverage

```
> vif(mlr)
   lstat       rm     crim
1.941883 1.616468 1.271372
```

Figure 17: VIF for Multicollinearity

Problem 2: Salaries Data Set

Descriptive Summary

This data set represents a sample of nine-month academic salaries for professors at a U.S. college. The data was collected as part of an ongoing effort to monitor salary differences between male and female faculty members. This analysis focuses on six specific variables: *rank*, *discipline*, *yrs.since.phd*, *yrs.service*, *sex*, and *salary*. *rank* represents the tenure of the professor. They can be an Assistant Professor ("AsstProf"), an Associate Professor ("AssocProf"), or a Full Professor ("Prof"). There are 67 Assistant Professors, 64 Associate Professors, and 266 Full Professors. *discipline* represents the professor's department. It has a value of either "A" for theoretical departments or "B" for applied departments. There are 181 "A" professors and 216 "B" professors. *yrs.since.phd* represents the number of years since the professor received their PhD. It has a mean of 22.31 years, a median of 21.00 years, and a range of 55.00 years. *yrs.service* represents the number of years the professor has taught at the college. It has a mean of 17.61 years, a median of 16.00 years, and a range of 60.00 years. *sex* represents the gender of the professor, which can either be "Male" or "Female." There are 39 female professors and 358 male professors. Finally, *salary* represents the nine-month salary of the professor. It has a mean of $113,706, a median of $107,300, and a range of $173,745. For more information regarding the summary statistics and their distributions, please reference Figures 18, 21, 22, and 23.

```
> summary(Salaries)
        rank     discipline yrs.since.phd    yrs.service        sex          salary      
 AsstProf : 67   A:181      Min.   : 1.00   Min.   : 0.00   Female: 39   Min.   : 57800  
 AssocProf: 64   B:216      1st Qu.:12.00   1st Qu.: 7.00   Male  :358   1st Qu.: 91000  
 Prof     :266              Median :21.00   Median :16.00               Median :107300  
                            Mean   :22.31   Mean   :17.61               Mean   :113706  
                            3rd Qu.:32.00   3rd Qu.:27.00               3rd Qu.:134185  
                            Max.   :56.00   Max.   :60.00               Max.   :231545
```

Figure 18: Summary Statistics

The boxplots in Figure 19 provide a side-by-side comparison of *sex* and *salary*. The Female boxplot is slightly positively (right) skewed with no known outliers. The Male boxplot is also slightly positively (right) skewed with three upper outliers. The boxplots in Figure 20 show an enhanced side-by-side comparison of *sex* and *salary*, this time breaking it down by *discipline*. The A.Female boxplot is extremely positively (right) skewed with no known outliers. The B.Female boxplot is slightly positively skewed with no known outliers. The A.Male boxplot is slightly positively (right) skewed with three upper outliers. The B.Male boxplot is slightly positively (right) skewed with one upper outlier.

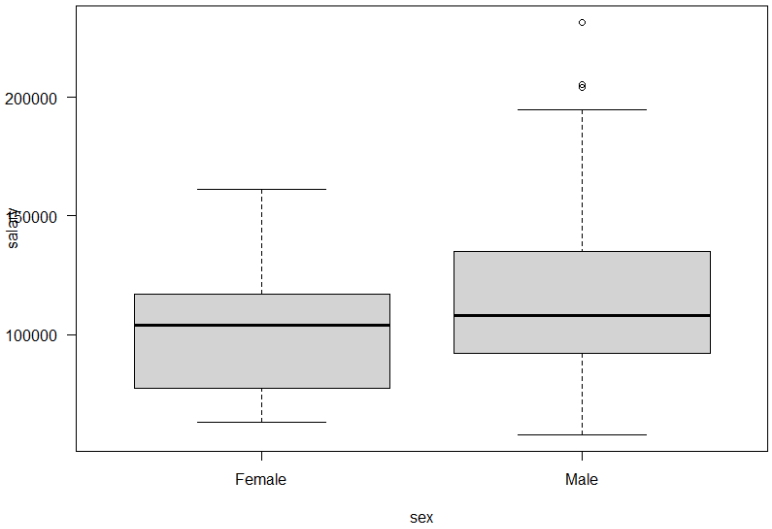

Figure 19: Boxplot of sex and salary

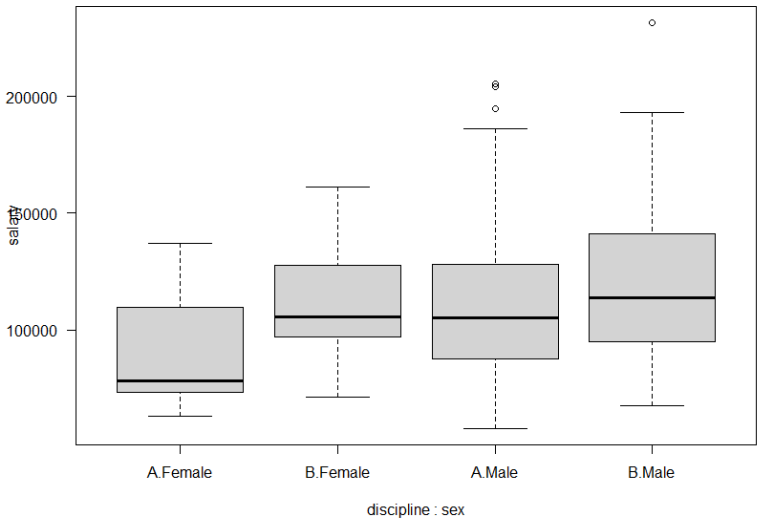

Figure 20: Boxplots of sex and salary Broken down by discipline

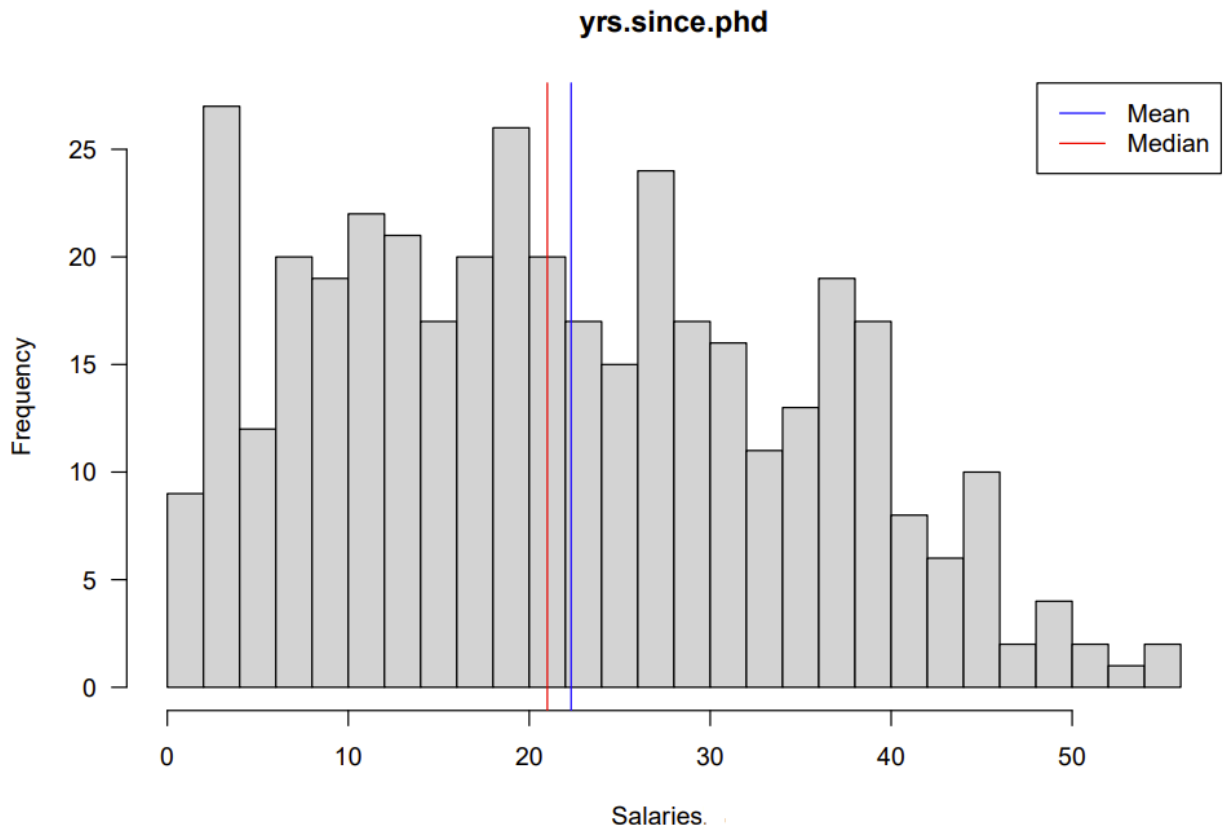

Figure 21: Histogram of yrs.since.phd

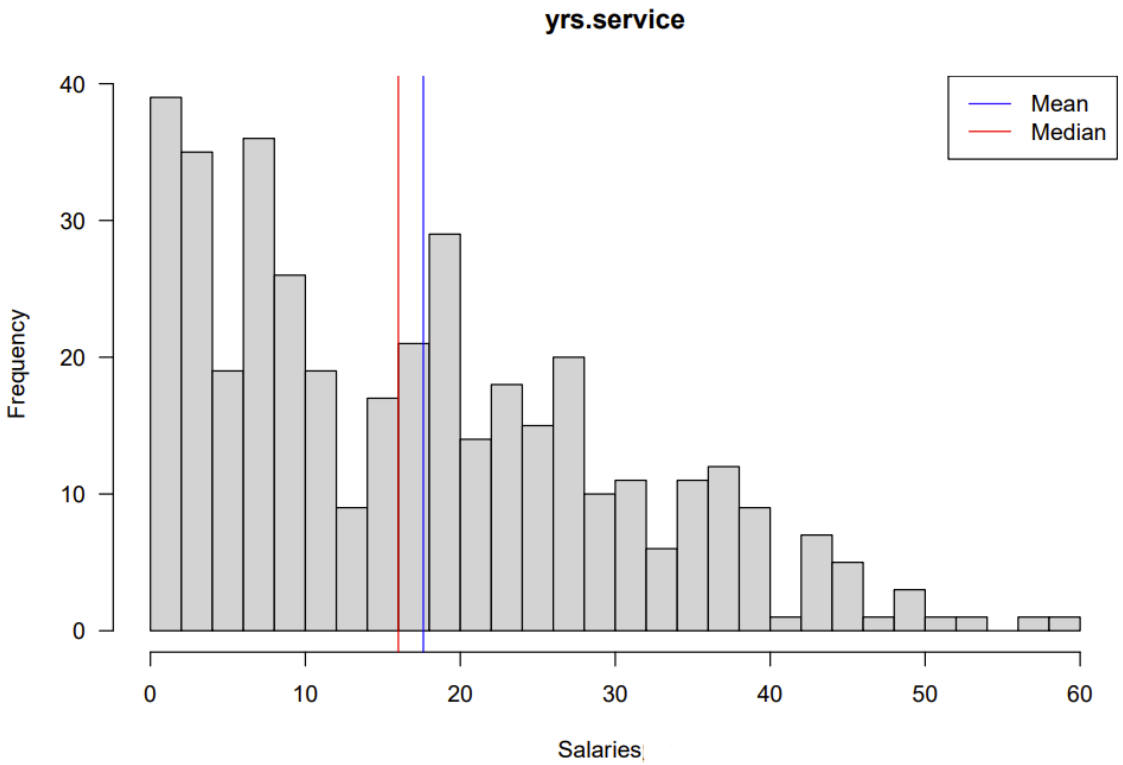

Figure 22: Histogram yrs.service

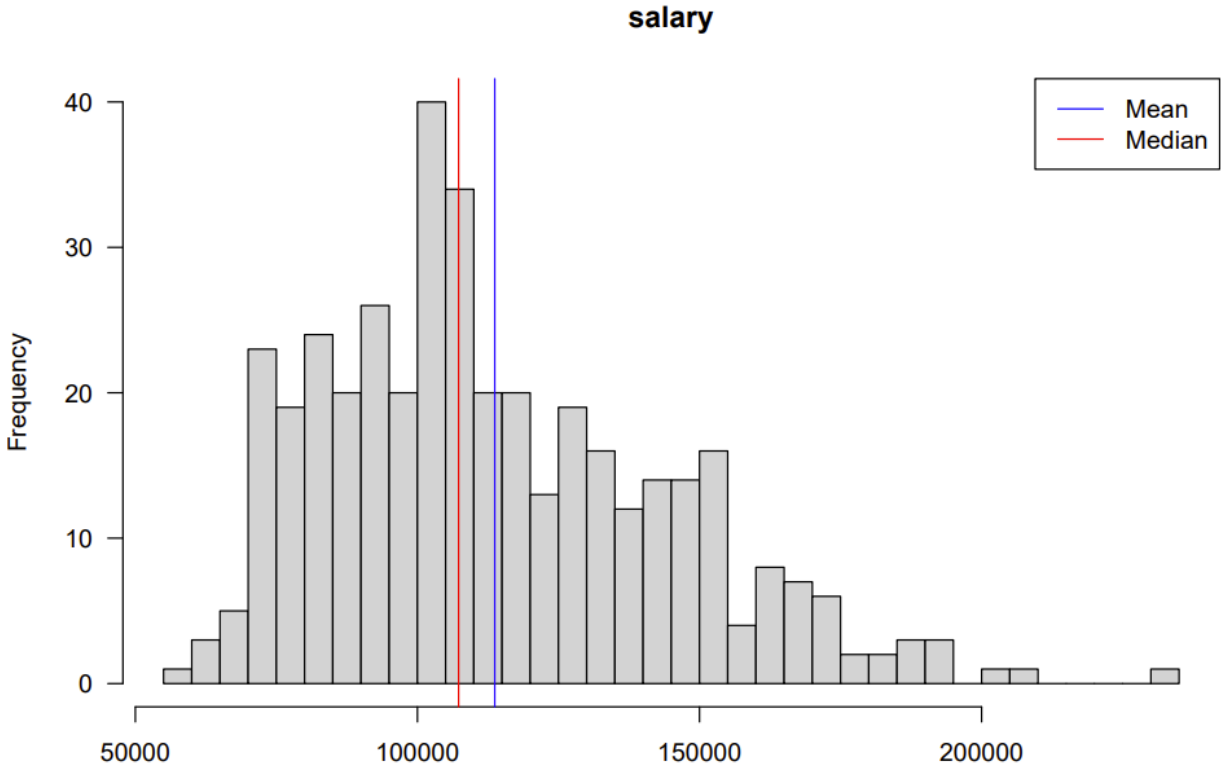

Figure 23: Histogram of salary

Inferential Analysis

I first decided to create a simple linear regression using sex as the lone predictor. This yielded a regression equation of *salary* = 101002 + 14088 x *sexMale*. 101002 is the y-intercept (β_0). It represents the salary of a professor when β_1 is equal to zero. *sexMale* is a dummy variable. According to this regression, if the professor is a male, then they earn $14,088 more. Obviously, this cannot be the only factor that contributes to salary. It takes more than one variable to predict, as is evidenced by the extremely low adjusted R-squared of 0.01673, which indicates that only 1.673% of the variation in salary can be explained by the professor's sex. More variables will have to be added for a better regression. A simple linear regression simply does not work for predicting a response like salary.

```
Call:
lm(formula = salary ~ sex, data = Salaries)

Residuals:
   Min     1Q Median     3Q    Max
-57290 -23502  -6828  19710 116455

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   101002       4809  21.001  < 2e-16 ***
sexMale        14088       5065   2.782  0.00567 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30030 on 395 degrees of freedom
Multiple R-squared:  0.01921,   Adjusted R-squared:  0.01673
F-statistic: 7.738 on 1 and 395 DF,  p-value: 0.005667
```

Figure 24: Simple Linear Regression

Next, I made a multiple linear regression using *discipline*, *rank*, *sex*, and *yrs.service* as the predictors. This yielded a regression equation of *salary* = 68351.67 + 13473.38 x *disciplineB* + 14560.40 x *rankAssocProf* + 49159.64 x *rankProf* + 4771.25 x *sexMale* – 88.78 x *yrs.service*. 68351.67 is the y-intercept (β_0). *disciplineB* is a dummy variable that gives a $13,473.38 increase in salary to professors teaching in the applied departments. *rankAssocProf* is a dummy variable that gives a $14,560.40 increase in salary to professor either having or having had the title of Associate Professor. *rankProf* is a dummy variable that gives a $49,159.64 increase in salary to the professor for having the title of Full Professor. *sexMale* is a dummy variable that adds $4,771.25 to the salary of professors that are males. Finally, *yrs.service* is a predictor that can be interpreted as a $88.78 reduction in salary for each additional year that the professor has been at the college. The null hypothesis was H_0 : $\beta_{disciplineB} = \beta_{rankAssocProf} = \beta_{rankProf} = \beta_{sexMale} = \beta_{yrs.service} = 0$, the alternative hypothesis was
H_A: at least one of the regression coefficients $\neq 0$, the confidence level was 95% ($\alpha = 0.05$), the F-statistic was 63.41, and the p-value was 2.2×10^{-16}. Based on the results of the test, we reject the null hypothesis because the p-value of 2.2×10^{-16} is less than the alpha of 0.05. There is statistically significant evidence that suggests $\beta_{disciplineB}$, $\beta_{rankAssocProf}$,

$\beta_{rankProf}$, $\beta_{sexMale}$, $\beta_{yrs.service}$ is different from 0. Looking at the adjusted R-squared, this regression is a better predictor of salary, as 44.07% of the variation in salary can be explained by discipline, rank, sex, yrs.service.

```
Call:
lm(formula = salary ~ discipline + rank + sex + yrs.service,
    data = Salaries)

Residuals:
   Min     1Q Median     3Q    Max
-64202 -14255  -1533  10571  99163

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   68351.67    4482.20  15.250  < 2e-16 ***
disciplineB   13473.38    2315.50   5.819 1.24e-08 ***
rankAssocProf 14560.40    4098.32   3.553 0.000428 ***
rankProf      49159.64    3834.49  12.820  < 2e-16 ***
sexMale        4771.25    3878.00   1.230 0.219311
yrs.service     -88.78     111.64  -0.795 0.426958
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22650 on 391 degrees of freedom
Multiple R-squared:  0.4478,    Adjusted R-squared:  0.4407
F-statistic: 63.41 on 5 and 391 DF,  p-value: < 2.2e-16
```

Figure 25: Multiple Linear Regression

Conclusion

As was demonstrated in the boxplots and histograms, all numeric variables exhibited some form of positive (right) skew. Outliers were not as prominent as in the Boston data set, so there is little cause for concern of extreme positive biases in the data. Regarding the inferential analysis, a key takeaway is that, once again, a simple linear regression is not an appropriate way to model the data in the Salaries data set. Adding more predictors into the regression proved to contribute more to the total variation. Something that I noticed regarding the relationship between sex and salary from the multiple linear regression was that the gender of the professor did not have as much of an impact on the salary, as did other variables like rank and discipline. By being a male, the professor only made $4,771.25 more, compared to them making an additional $13,473.38 for teaching in the "B" discipline.

Looking at the normal quantile plot in Figure _, the patterns of the plots run roughly close to that of a straight line along the trend line, making it reasonable to assume normality in the data. The residuals in Figure _ appear to be randomly distributed, making the assumption of equal variances valid. Figure _ displays a variance inflation factor (VIF) calculation to assess whether multicollinearity exists between the three predictor variables. Given the calculated values of 1.029040, 1.757308, 2.515233, 1.030803, and 1.627110 for *disciplineB, rankAssocProf, rankProf, sexMale,* and *yrs.service,* respectively, we can conclude that there does not appear to

be multicollinearity. The correlation between these five variables is not severe enough to warrant any corrective measures.

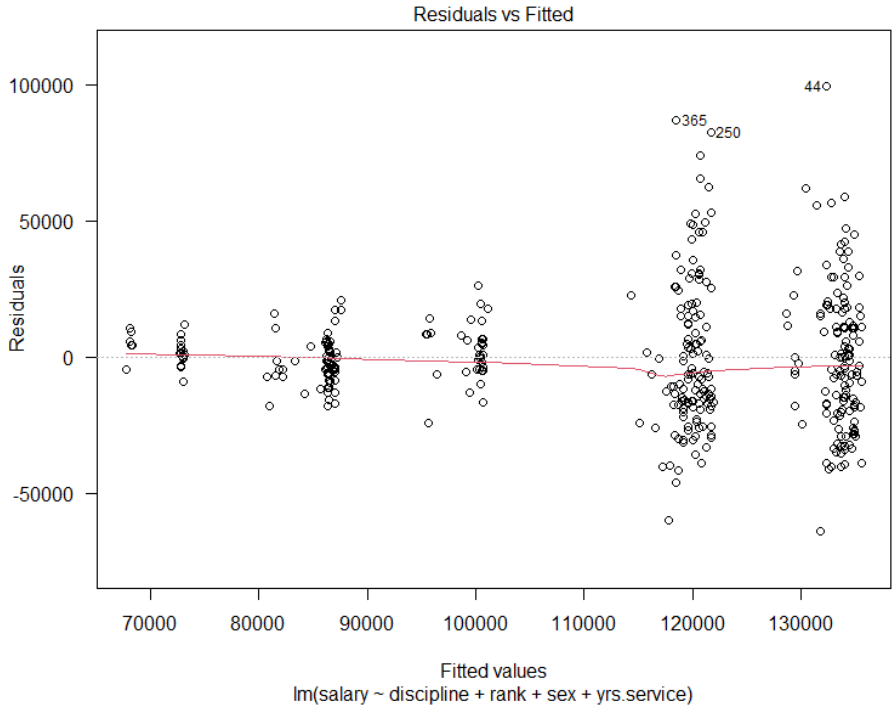

Figure 26: Residual Plot

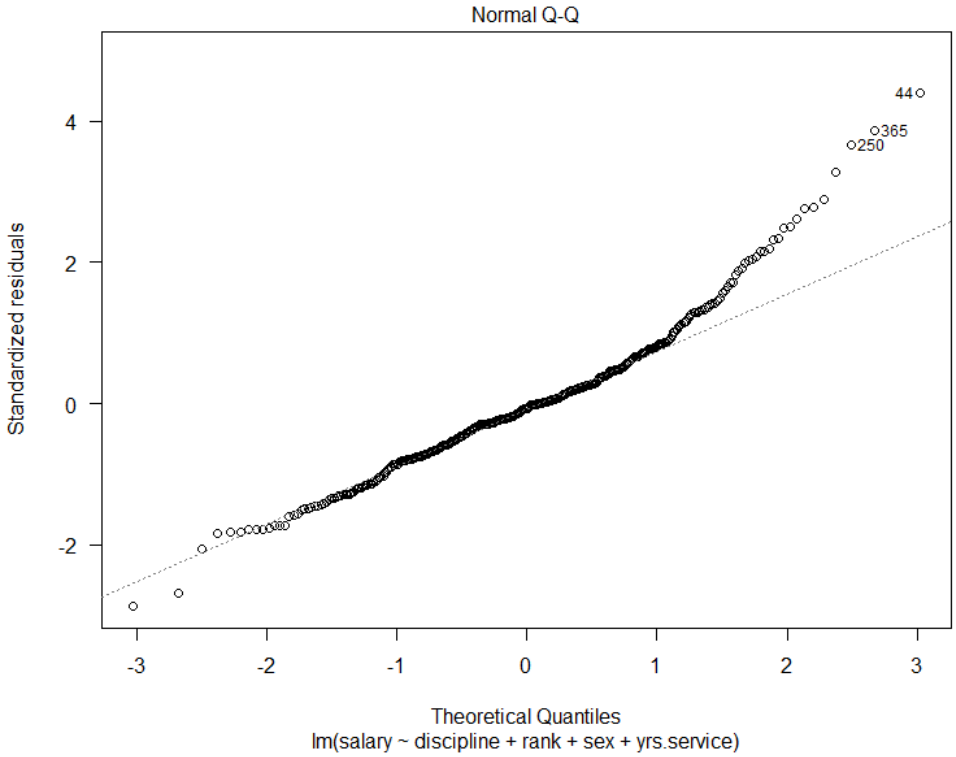

Figure 27: Normal Q-Q Plot

```
vif(lmod)
disciplineB rankAssocProf      rankProf       sexMale   yrs.service
   1.029040      1.757308      2.515233      1.030803      1.627110
```

Figure 28: VIF for Multicollinearity