Nkweteyim-FinalProject

Raisa Nkweteyim

5/12/2021

Analysis of Boston Housing Dataset to Determine Factors Affecting Full-Value Property-Tax Rate

Introduction

To carry out this data analysis, I will be using an inbuilt dataset in R called 'BostonHousing2'. The data was curated for a paper which investigated housing market data in Boston to measure the willingness to pay for clean air. The findings were published in the Journal of Environmental Economics and Management. The dataset has 506 observations and 19 variables which include:

crim - per capita crime rate by town zn - proportion of residential land zoned for lots over 25,000 sq.ft indus - proportion of non-retail business acres per town chas - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) nox - nitric oxides concentration (parts per 10 million) rm - average number of rooms per dwelling age - proportion of owner-occupied units built prior to 1940 dis - weighted distances to five Boston employment centres rad - index of accessibility to radial highways tax - full-value property-tax rate per USD 10,000 ptratio - pupil-teacher ratio by town b - 1000(B - 0.63)^2 where B is the proportion of blacks by town lstat - percentage of lower status of the population medv - median value of owner-occupied homes in USD 1000's cmedv - corrected median value of owner-occupied homes in USD 1000's town - name of town tract - census tract lon - longitude of census tract lat - latitude of census tract

I will be investigating the possible effects the above variables may have on tax (full-value property-tax rate per USD 10,000).

I chose tax as my response variable because it would be interesting to find out if the value of a home, its number of rooms, its age and its location (distance to employment centers) has an effect on the full-value property-tax rate paid by homeowners who occupy their homes. By measuring property tax rate, one can predict the likelihood of someone investing in real estate in a particular town. The lower the property tax rate, the greater the incentive to invest in a particular town.

1. Exploratory Data Analysis

Reading The Data

```
# Reading the data
library(mlbench)
data(BostonHousing2)
dim(BostonHousing2)
```

```
## [1] 506  19
```

Missing Data and Predictor Characteristics

```
# Check for missing data
anyNA(BostonHousing2)
```

```
## [1] FALSE
```

```
# Predictor characteristics
str(BostonHousing2)
```

```
## 'data.frame':    506 obs. of  19 variables:
##  $ town   : Factor w/ 92 levels "Arlington","Ashland",..: 54 77 77 46 46 46 69 69 69 69 ...
##  $ tract  : int  2011 2021 2022 2031 2032 2033 2041 2042 2043 2044 ...
##  $ lon    : num  -71 -71 -70.9 -70.9 -70.9 ...
##  $ lat    : num  42.3 42.3 42.3 42.3 42.3 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
##  $ cmedv  : num  24 21.6 34.7 33.4 36.2 28.7 22.9 22.1 16.5 18.9 ...
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : int  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ b      : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
```

There was no missing data and almost all the variables have continuous values except the response variable (tax), tract, b, and rad which have integer values and town and chase which have categorical values.

First, I will remove the categorical variables to ease the exploration of the dataset.

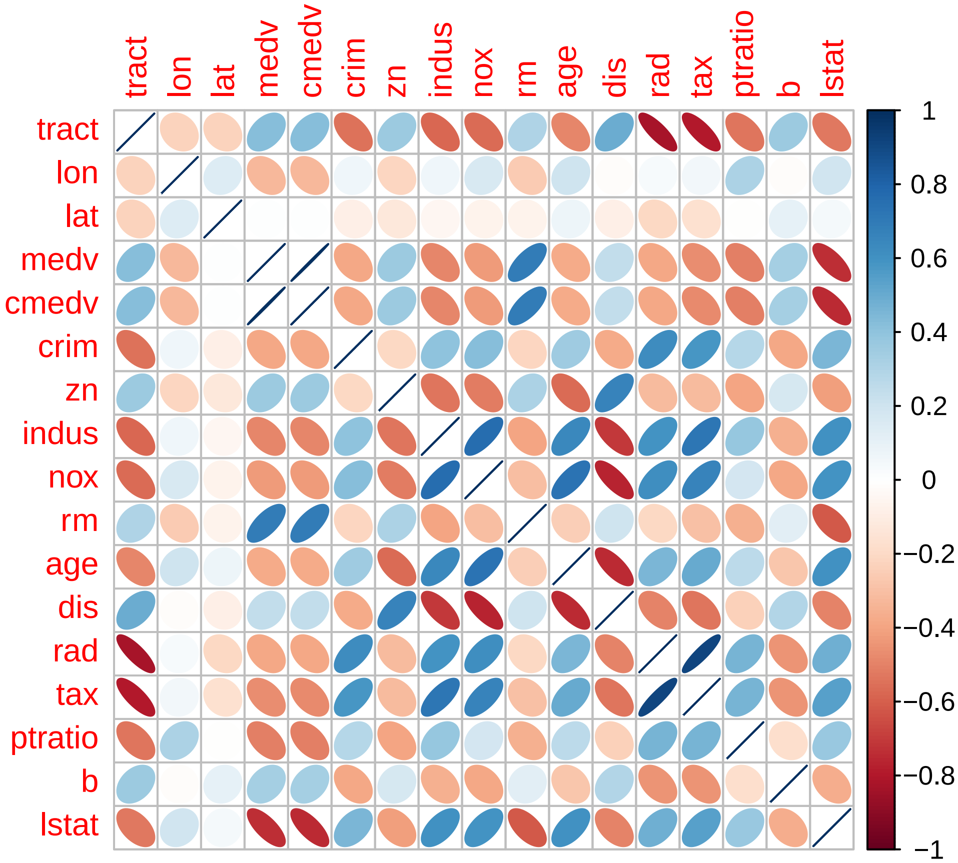
```
boston <- subset(BostonHousing2, select = -c(town, chas))
```

Correlation between predictors and outcome variable Next, I will check the correlation between the response (tax) and the predictors by plotting a correlation matrix.

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

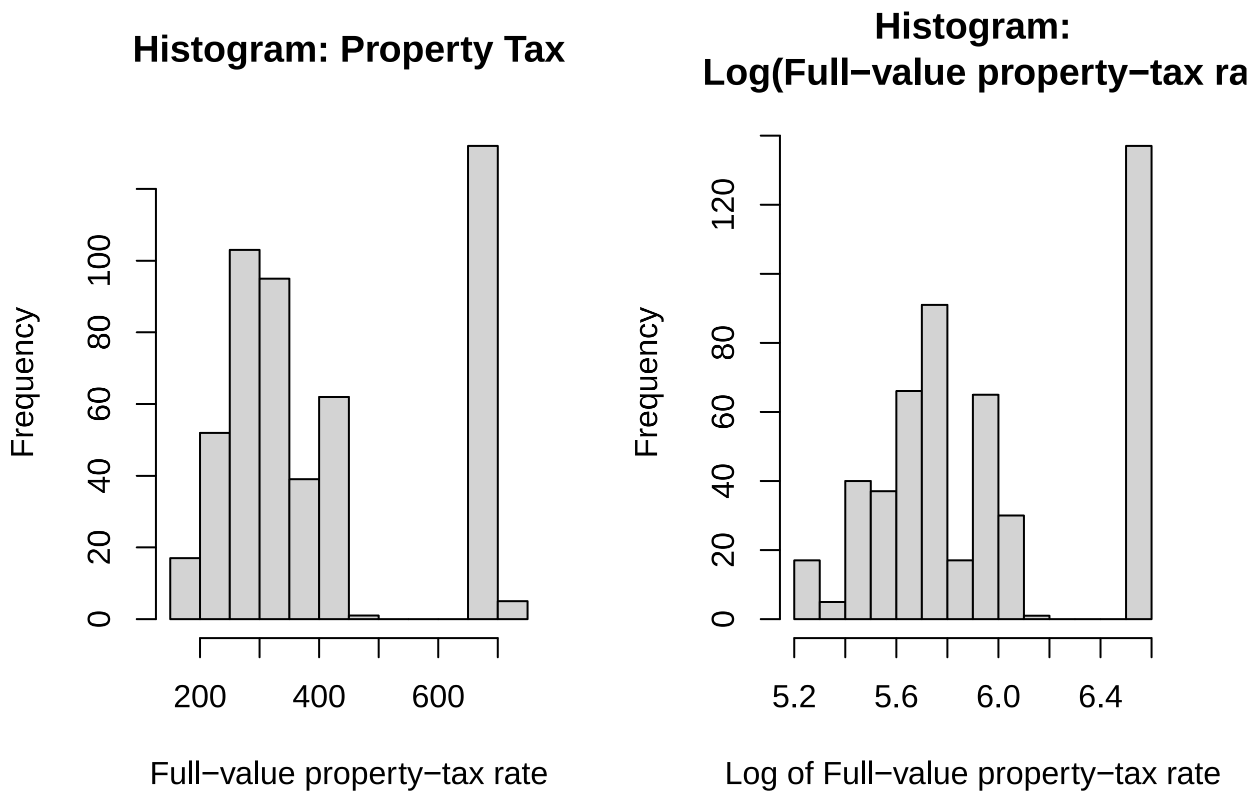
```
COR <- cor(boston, use="everything", method ="pearson")
corrplot(COR, method="ellipse")
```



It is seen that there is a significant positive correlation between rm and cmedv in addition to tax and age, a weak positive correlation between cmedv and dis, a weak negative correlations between tax and rm, dis and cmedv. Thus, as rm (average number of rooms per dwelling), cmedv (corrected median value of owner-occupied homes in USD 1000's) or dis (weighted distances to five Boston employment centres) is increasing, the tax is decreasing and as age is increasing, the tax of the property is increasing as well. As rm (average number of rooms per dwelling) increases, cmedv (corrected median value of owner-occupied homes in USD 1000's) increases as well.

Data Distribution Now, I will check the distribution of the response variable.

```
par(mfrow=c(1,2))
hist(boston$tax, xlab="Full-value property-tax rate", main="Histogram: Property Tax")
log_boston_tax <- log(boston$tax)
hist(log_boston_tax, xlab="Log of Full-value property-tax rate", main="Histogram:
     Log(Full-value property-tax rate)")
```



There was no significant improvement in the shape of the histograms of the response variable that had undergone a log transformation and when there was no transformation. Hence, the data was left as is. Even though the histograms are not entirely bell-shaped, since the sample size is large (506 observations), normality can be inferred.

2. Linear Regression and Variable Selection

(a) Partitioning data randomly with a 2:1 ratio

```
library(caTools)
set.seed(123)
split = sample.split(boston, SplitRatio = 2/3)
train = subset(boston, split == TRUE)
test = subset(boston, split == FALSE)
dim(train)
```

```
## [1] 327  17
```

```
dim(test) #number of rows of train and test sums up to 506 like original dataset
```

```
## [1] 179  17
```

Fitting Full Model

```
# Full model
fit.full.train <- lm(train$tax ~ ., data = train)
summary(fit.full.train)
```

```
##
## Call:
## lm(formula = train$tax ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -234.715  -17.501   -6.052   16.217  219.356
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.880e+03  4.778e+03   0.812 0.417354
## tract       -1.566e-02  6.618e-03  -2.366 0.018578 *
## lon         -2.700e+01  5.071e+01  -0.532 0.594849
## lat         -1.300e+02  7.873e+01  -1.651 0.099655 .
## medv         5.420e+00  6.116e+00   0.886 0.376216
## cmedv       -7.493e+00  6.166e+00  -1.215 0.225227
## crim        -3.340e-01  4.612e-01  -0.724 0.469470
## zn           7.989e-01  2.089e-01   3.824 0.000159 ***
## indus        8.162e+00  8.462e-01   9.646  < 2e-16 ***
## nox         -1.351e+01  6.038e+01  -0.224 0.823134
## rm           2.941e+00  7.145e+00   0.412 0.680919
## age         -3.074e-02  1.991e-01  -0.154 0.877386
## dis         -2.684e+00  3.273e+00  -0.820 0.412772
## rad          1.168e+01  1.038e+00  11.262  < 2e-16 ***
## ptratio     -5.553e-01  2.212e+00  -0.251 0.801992
## b           -5.009e-05  4.221e-02  -0.001 0.999054
## lstat       -7.646e-01  8.525e-01  -0.897 0.370446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.35 on 310 degrees of freedom
## Multiple R-squared:  0.894,  Adjusted R-squared:  0.8885
## F-statistic: 163.4 on 16 and 310 DF,  p-value: < 2.2e-16
```

Stepwise variable selection method

```
library(MASS)
fit.step <- stepAIC(fit.full.train, direction="both", k=log(nrow(train)), trace = FALSE)
fit.step$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## train$tax ~ tract + lon + lat + medv + cmedv + crim + zn + indus +
##     nox + rm + age + dis + rad + ptratio + b + lstat
##
## Final Model:
```

```
## train$tax ~ cmedv + zn + indus + rad
##
##
##         Step Df     Deviance Resid. Df Resid. Dev      AIC
## 1                                  310   984251.2 2717.593
## 2        - b  1 4.470815e-03       311   984251.2 2711.804
## 3      - age  1 7.579798e+01       312   984327.0 2706.039
## 4      - nox  1 2.127714e+02       313   984539.8 2700.319
## 5  - ptratio  1 9.040842e+01       314   984630.2 2694.560
## 6       - rm  1 4.928184e+02       315   985123.0 2688.933
## 7     - crim  1 1.436080e+03       316   986559.1 2683.620
## 8      - lon  1 1.561220e+03       317   988120.3 2678.347
## 9     - medv  1 2.320275e+03       318   990440.6 2673.324
## 10     - dis  1 2.301307e+03       319   992741.9 2668.293
## 11   - lstat  1 2.726715e+03       320   995468.6 2663.400
## 12     - lat  1 6.815179e+03       321  1002283.8 2659.841
## 13   - tract  1 7.585656e+03       322  1009869.4 2656.516
```

```r
summary(fit.step)
```

```
##
## Call:
## lm(formula = train$tax ~ cmedv + zn + indus + rad, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -239.21  -19.76   -5.44   13.09  240.92
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 210.6137    13.1415   16.03  < 2e-16 ***
## cmedv        -1.4652     0.3826   -3.83 0.000154 ***
## zn            0.7783     0.1563    4.98 1.04e-06 ***
## indus         8.4616     0.6535   12.95  < 2e-16 ***
## rad          13.5487     0.4480   30.24  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56 on 322 degrees of freedom
## Multiple R-squared:  0.8912, Adjusted R-squared:  0.8899
## F-statistic: 659.6 on 4 and 322 DF,  p-value: < 2.2e-16
```

Final Model: train$tax ~ cmedv + zn + indus + rad The adjusted R-squared value is 0.8899 implying that 88.99% of the variation in property taxes can be accounted for by the cmdev (median value of owner-occupied homes), zn (proportion of residential land zoned for lots over 25,000 sq.ft), indus (proportion of non-retail business acres per town) and rad (index of accessibility to radial highways). The p-values of these 4 variables are all also less than 0.001 implying that these predictors are highly significantly related to the response, tax.

Applying Model to Test Data

```r
predict_step <- predict(fit.step, newdata = test)
# sum of squared prediction error (SSPE)
sum((test$tax - predict_step)**2)
```

```
## [1] 598078.4
```

Refitting Final Model to Entire Dataset

```
fit.final <- lm(boston$tax ~ boston$cmedv + boston$zn + boston$indus + boston$rad, data=boston)
summary(fit.final)
```

```
## 
## Call:
## lm(formula = boston$tax ~ boston$cmedv + boston$zn + boston$indus + 
##     boston$rad, data = boston)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -222.257  -20.532   -4.244   14.439  263.571 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  215.9482    10.9240  19.768  < 2e-16 ***
## boston$cmedv  -1.5035     0.3182  -4.725 2.99e-06 ***
## boston$zn      0.7510     0.1284   5.851 8.85e-09 ***
## boston$indus   7.3908     0.5286  13.982  < 2e-16 ***
## boston$rad    14.1703     0.3618  39.169  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 56.33 on 501 degrees of freedom
## Multiple R-squared:  0.8892, Adjusted R-squared:  0.8883 
## F-statistic:  1005 on 4 and 501 DF,  p-value: < 2.2e-16
```

The adjusted R-squared value is 0.8883 implying that 88.83% of the variation in property taxes can be accounted for by the cmdev, zn, indus, and rad. The p-values of these 4 variables are all also less than 0.001 implying that these predictors are highly significantly related to the response, tax. 88.83% is not very different from 88.85% (adjusted R-squared value of full model). However, the number of variables decreased from 17 to 4. Generally, the less complicated the model is, the better such as to prevent overfitting.

3. Perform model diagnostics on the final model.

(a) Check Normality

```
library(car)
```

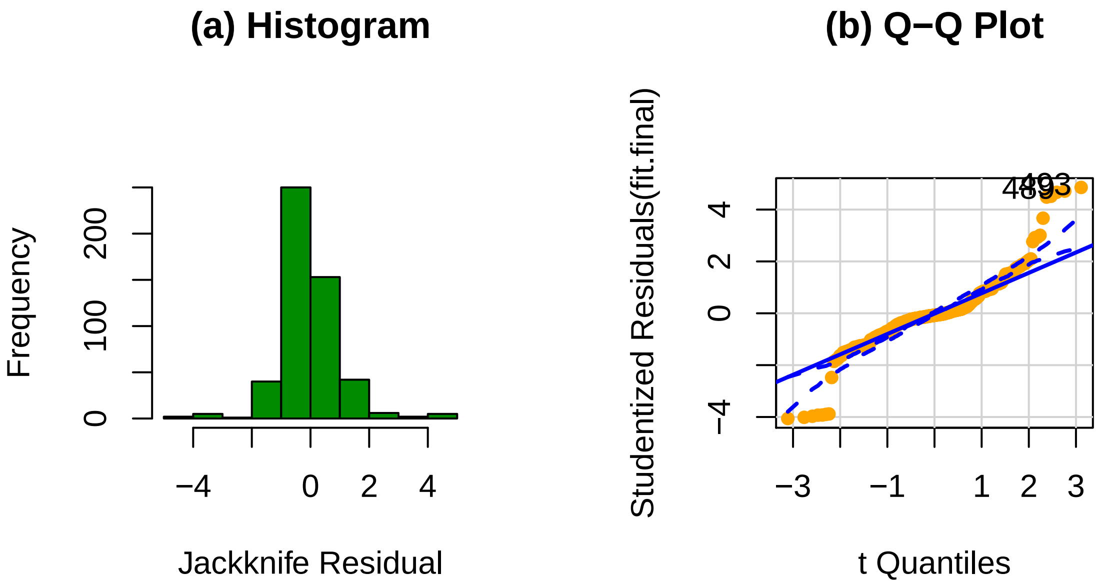
```
## Warning: package 'car' was built under R version 4.0.4
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.0.3
```

```r
#Studentized jackkinfe residuals
r.jack <- rstudent(fit.final)

# Graphs
par(mfrow=c(1,2),mar=c(8,4,8,4))
hist(r.jack, xlab="Jackknife Residual", col="green4", main="(a) Histogram")
qqPlot(fit.final, pch=19, cex=.8, col="orange", main="(b) Q-Q Plot")
```



```
## [1] 489 493
```

```r
# THE SHAPIRO-WILKS NORMALITY TEST
shapiro.test(r.jack)
```

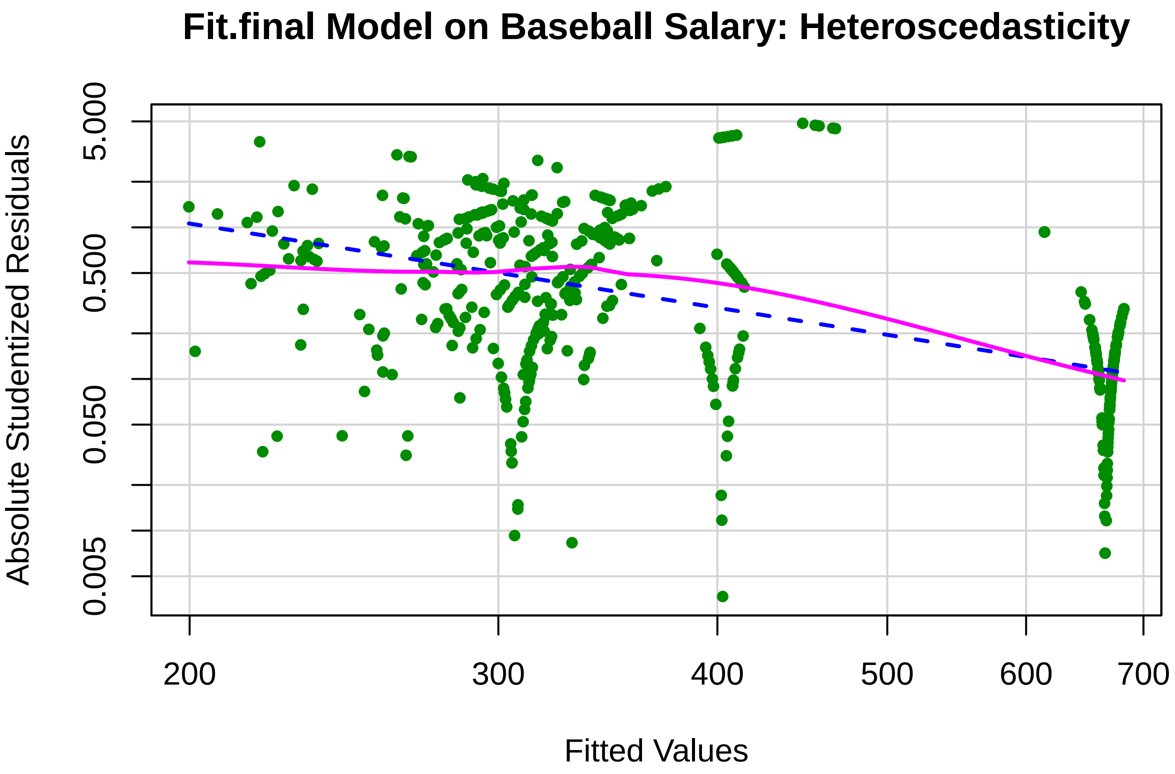
```
##
##  Shapiro-Wilk normality test
##
## data:  r.jack
## W = 0.866, p-value < 2.2e-16
```

The histogram has a fairly bell-shaped structure with a a couple of outliers while in the qqplot, majority of the points fall on the solid blue line with a couple of outliers again. The Shapiro-Wilk's test gave a p-value of 2.359e-10. All of these results suggest that the normality assumption has been violated. However, from the Central Limit Theorem, it can be concluded that since the data is large, the entire data tends towards a normal distribution *(b) Check Homoscedasticity*

```
ncvTest(fit.final)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 24.36302, Df = 1, p = 7.9786e-07
```

```
# Plot of Absolute Jackknife Residuals vs. Fitted values
spreadLevelPlot(fit.final, pch=20, cex=0.5, col="green4",
    main="Fit.final Model on Baseball Salary: Heteroscedasticity")
```



```
##
## Suggested power transformation:  2.843588
```

```
# IF THE LINES ARE FLAT, THEN EQUAL VARIANCE IS JUSTIFIED.
```

In the plot, the solid purple and dashed blue lines are not flat. Again, there appears to be a trend in the data points in such a way that they form a defined pattern. These results imply that the equal variance assumption has been violated. Again, the Breusch-Pagan Test produced a significant p-value of 7.9786e-07 (<0.05). This result confirms that the data indeed does not have equal variances and the assumption has been violated. Thus, a power transformation of 2.843588 has been suggested.

```r
response <- boston$tax^(2.843588)
fit.final <- lm(boston$tax ~ boston$cmedv + boston$zn + boston$indus + boston$rad, data=boston)
```

(c) Check Independence

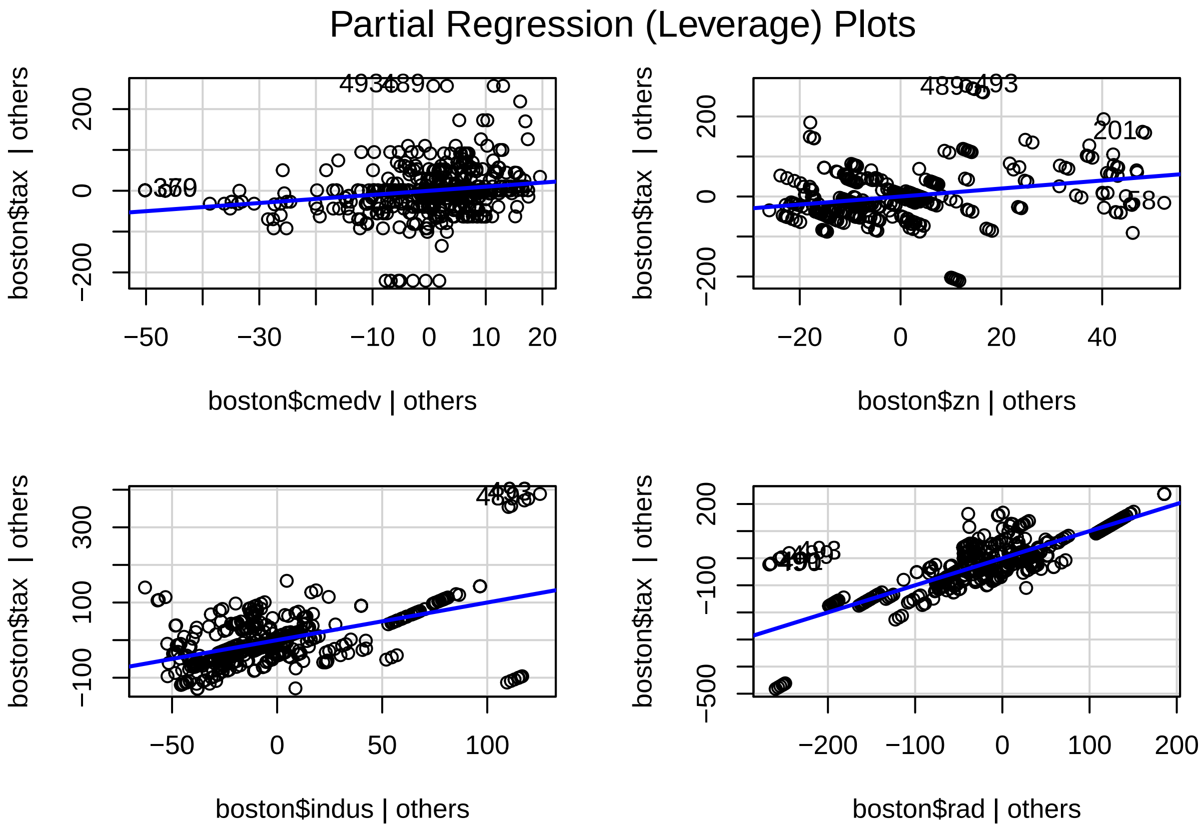
```r
# Durbin-Watson Test for Autocorrelated Errors
durbinWatsonTest(fit.final)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.7343293     0.5279609       0
##  Alternative hypothesis: rho != 0
```

The Durbin-Watson test produced a p-value < 0.05. This implies that the data is in fact independent.

(d) Check Linearity

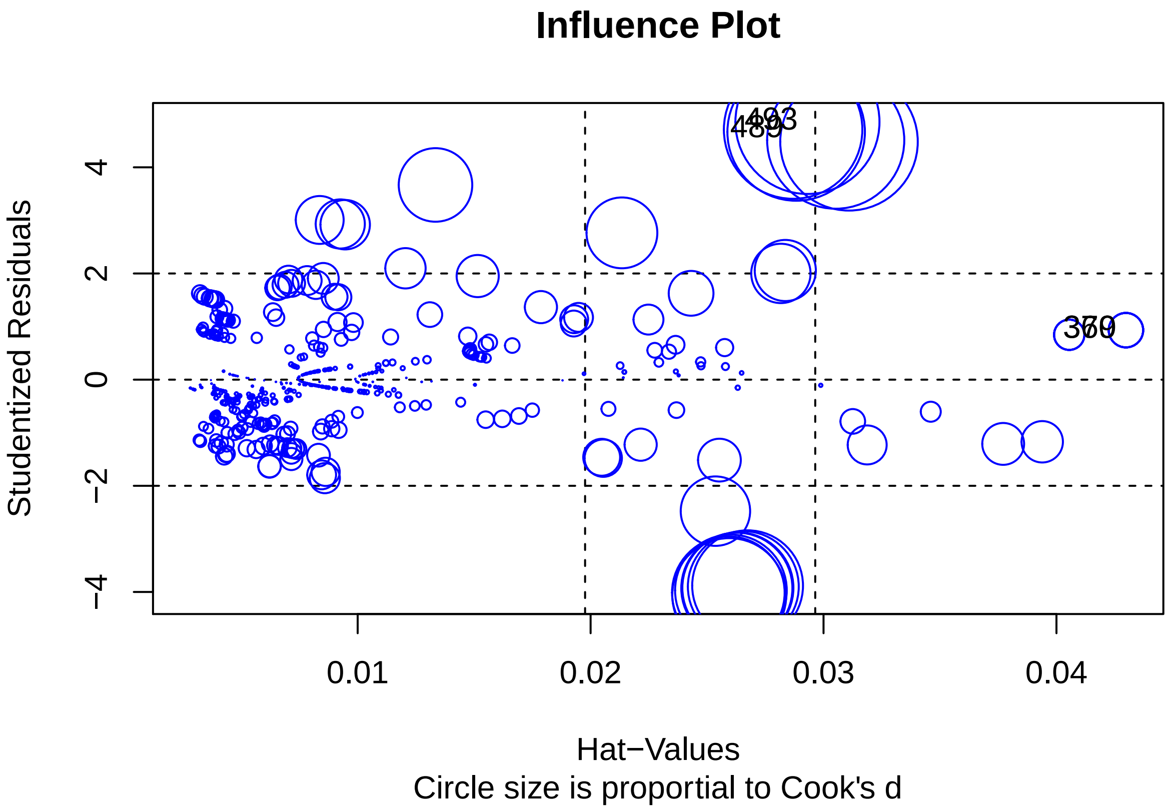
```r
# leverage plots or partial regression plot
leveragePlots(fit.final, main="Partial Regression (Leverage) Plots")
```



The blue lines on the 4 leverage plots follow a straight line. This implies that all the predictors are linear.

(e) Outlier Detection

```r
influencePlot(fit.final, id.method="identify",
    col="blue",
    main="Influence Plot",
    sub="Circle size is proportial to Cook's d")
```



```
##       StudRes        Hat       CookD
## 369 0.932056 0.04298333 0.007805638
## 370 0.932056 0.04298333 0.007805638
## 489 4.711564 0.02869456 0.125836261
## 493 4.854948 0.02930951 0.136203927
```

(f) Multicollinearity

```
# CONDITION NUMBER (> 100?)
kappa(lm(boston$tax ~ boston$cmedv + boston$zn + boston$indus + boston$rad, data=boston -1, x=TRUE)$x);
```

```
## [1] 105.6268
```

```
# COMPUTE VIF USING FUNCTION vif (> 10?)
vif(lm(boston$tax ~ boston$cmedv + boston$zn + boston$indus + boston$rad, data=boston -1, x=TRUE))
```

```
## boston$cmedv    boston$zn boston$indus    boston$rad
##     1.358740     1.426556     2.093012     1.579241
```

When disregarding the intercept, kappa = 105.6268 which is only slightly greater than 100 implying that multicollinearity might not be a big issue. Using vif function across all the predictors, the resulting values are all below 10 implying that multicollinearity should be due to the intercept term.

Conclusion

To conclude, the biggest indicators of the full-value property-tax rate per USD 10,000 is the cmedv - corrected median value of owner-occupied homes in USD 1000's, zn - proportion of residential land zoned for lots over 25,000 sq.ft, indus - proportion of non-retail business acres per town and rad - index of accessibility to radial highways. The variation in tax is accounted for by these variables by 88.99%. So, the prediction rate is not very high.

In the future, I would observe if an increase in variable 'indus' (proportion of non-retail business acres per town) causes an increase in 'nox' (nitric oxides concentration (parts per 10 million) from town to town. I could also check if an increase in 'nox' is related to whether of not the town is bounded by the Charles River (chas - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)). If such a relationship exists, it might be an indication that waste from industries are being dumped in the river.

References

Harrison, D. and Rubinfeld, D.L. (1978). Hedonic prices and the demand for clean air. Journal of Environmental Economics and Management, 5, 81–102.

Gilley, O.W., and R. Kelley Pace (1996). On the Harrison and Rubinfeld Data. Journal of Environmental Economics and Management, 31, 403–405. [Provided corrections and examined censoring.]

Newman, D.J. & Hettich, S. & Blake, C.L. & Merz, C.J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

Pace, R. Kelley, and O.W. Gilley (1997). Using the Spatial Configuration of the Data to Improve Estimation. Journal of the Real Estate Finance and Economics, 14, 333–340. [Added georeferencing and spatial estimation.]