**Assignment #3: Linear Regression**

**Submit on eCampus for Assignment 3**

**Deadline: October 4 (Friday) @11:59pm**

**# YOUR NAME: Joel Dsouza**

**#**

**# RULES:**

**#**

**# 1) You can discuss the homework with each other in general terms,**

**# but you must write your own solutions and not copy from anyone.**

**#**

**# 2) Edit your answers into this file following each problem.**

**#**

**# 3) Post your solutions to e-campus**

**#**

**# The filename should have this format: LastName-FirstName-hw03.doc**

**# An example would be: Eksin-Ceyhun-hw03.doc**

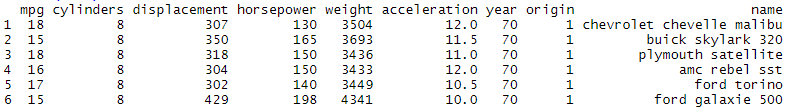
**Problem 1 (5pt)**

Use the Auto data set to answer the following questions:

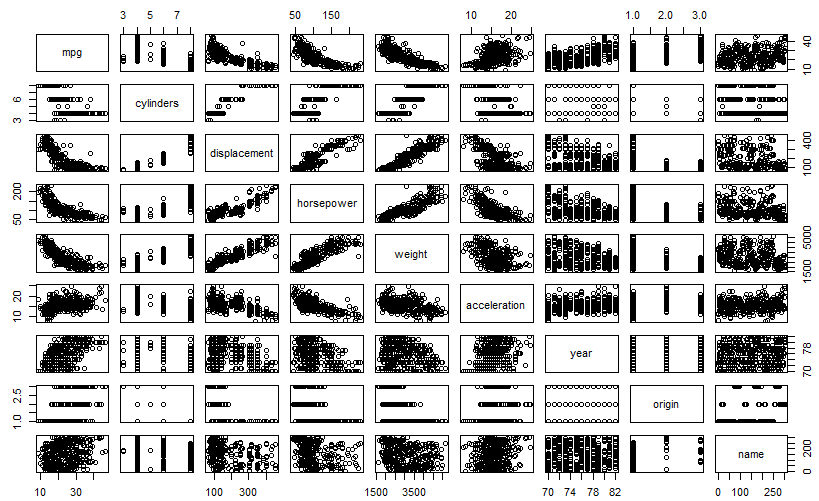
**(a) Produce a scatterplot matrix which includes all of the variables in the data set. Which predictors appear to have an association with the response?**

library(ISLR)

head(Auto)



plot(Auto)



Assumption: Response variable is mpg as given in the below questions.

Looking at the scatterplot above, we see a clear relationship between the following predictors with the response variable mpg:

1. Displacement (Negative slope)
2. Horsepower (Negative slope)
3. Weight (Negative slope)

Thus all three variables are negatively correlated to the **response variable mpg** and thus an increase in any of these variables results in a decrease of mpg

**(b) Compute the matrix of correlations between the variables (using the function cor()). You will need to exclude the name variable, which is qualitative.**

x<-Auto[,-9]

mpg cylinders displacement horsepower weight acceleration year origin

1 18 8 307.0 130 3504 12.0 70 1

2 15 8 350.0 165 3693 11.5 70 1

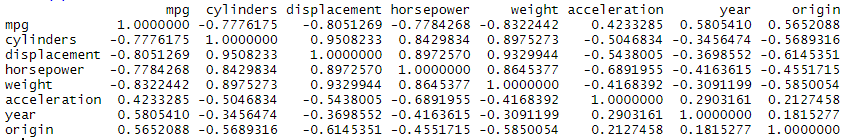
3 18 8 318.0 150 3436 11.0 70 1

4 16 8 304.0 150 3433 12.0 70 1

5 17 8 302.0 140 3449 10.5 70 1

The correlation matrix is as below:

cor(x)



(c) Perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Comment on the output. For example,

linreg = lm(mpg~., data=x)

summary(linreg)

Call:

lm(formula = mpg ~ ., data = x)

Residuals:

Min 1Q Median 3Q Max

-9.5903 -2.1565 -0.1169 1.8690 13.0604

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -17.218435 4.644294 -3.707 0.00024 \*\*\*

cylinders -0.493376 0.323282 -1.526 0.12780

displacement 0.019896 0.007515 2.647 0.00844 \*\*

horsepower -0.016951 0.013787 -1.230 0.21963

weight -0.006474 0.000652 -9.929 < 2e-16 \*\*\*

acceleration 0.080576 0.098845 0.815 0.41548

year 0.750773 0.050973 14.729 < 2e-16 \*\*\*

origin 1.426141 0.278136 5.127 4.67e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.328 on 384 degrees of freedom

Multiple R-squared: 0.8215, Adjusted R-squared: 0.8182

F-statistic: 252.4 on 7 and 384 DF, p-value: < 2.2e-16

**i. Is there a relationship between the predictors and the response?**

Using the F-statistic of 252.4>>1 and p-value of the F-statistic < 2.2e-16, we determine that there is a clear relationship between the predictor and response variable. This can be determined with the large value of F-statistic and extremely low value of its probability which is much lower that 0.05 which is the cutoff p value

**ii. Which predictors have a statistically significant relationship to the response?**

The following predictors have statistically significant relationship to the response:

Predictor t value Pr(>|t|)

1. displacement 2.647 0.00844 \*\*
2. weight -9.929 < 2e-16 \*\*\*
3. year 14.729 < 2e-16 \*\*\*
4. origin 5.127 4.67e-07 \*\*\*

From the above table, we can infer that the mentioned variables have a high t-value and a corresponding low p-value. With this information, we can reject the null hypothesis as t>>1 and p<<0.05 which is the assumed confidence interval of 95%

**iii. What does the coefficient for the year variable suggest?**

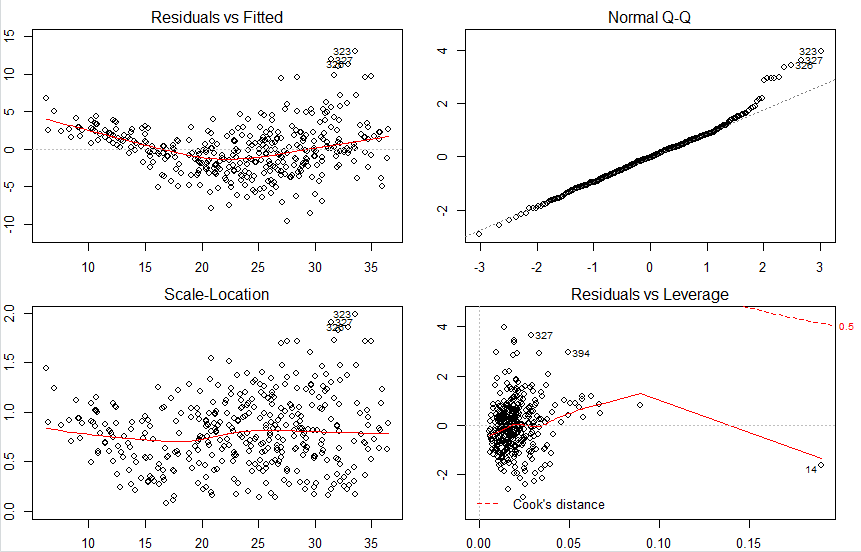
The co-efficient of the year variable is **0.750773** which shows that the year predictor is positively correlated to the response mpg. It means that if we increase the year, it means there is improved mpg. We can infer that better and newer models of cars are more efficient with mpg. When every other predictor held constant, the mpg value increases with each year that passes. Specifically, mpg increase by 1.43 each year.

**(d) Produce diagnostic plots of the linear regression fit. Comment on each plot.**

par(mfrow=c(2,2))

par(mar = rep(2, 4))

plot(linreg)



Residual vs Fitted plot:

There is a slight pattern in this plot which implies that the data is slightly non-linear. We can correct this non linearity by transforming the predictors with either log(x) or root(x)

Normal Q-Q:

This plot has certain values outside the straight line which indicates that certain values are outliers and high leverage points. We need to address these values accordingly in order to improve the accuracy and prediction of our model. The graph shows that the residuals are normally distributed and right skewed

Scale-Location:

This graph shows that the constant variance of error assumption is not true for this model.

Residual vs Levarage:

This graph shows that there is an observation that stands out as a potential leverage point (labeled 14 on the graph)

**(e) Is there serious collinearity problem in the model? Which predictors are collinear?**

install.packages('faraway')

library(faraway)

vif(linreg)



We know that VIF greater than 5 means there is a problem of collinearity. In the above table, we see cylinder, displacement, horsepower, weight have values greater than 5. Hence there is a serious concern of collinearity in our model. Here cylinder, displacement, horsepower, and weight have values of VIF greater than 5 and thus are a problem with high collinearity.

**(f) Fit linear regression models with interactions. Are any interactions statistically significant?**

> model = lm(mpg ~.-name+displacement:weight, data = Auto)

> summary(model)

Call:

lm(formula = mpg ~ . - name + displacement:weight, data = Auto)

Residuals:

Min 1Q Median 3Q Max

-9.9027 -1.8092 -0.0946 1.5549 12.1687

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.389e+00 4.301e+00 -1.253 0.2109

cylinders 1.175e-01 2.943e-01 0.399 0.6899

displacement -6.837e-02 1.104e-02 -6.193 1.52e-09 \*\*\*

horsepower -3.280e-02 1.238e-02 -2.649 0.0084 \*\*

weight -1.064e-02 7.136e-04 -14.915 < 2e-16 \*\*\*

acceleration 6.724e-02 8.805e-02 0.764 0.4455

year 7.852e-01 4.553e-02 17.246 < 2e-16 \*\*\*

origin 5.610e-01 2.622e-01 2.139 0.0331 \*

displacement:weight 2.269e-05 2.257e-06 10.054 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.964 on 383 degrees of freedom

Multiple R-squared: 0.8588, Adjusted R-squared: 0.8558

F-statistic: 291.1 on 8 and 383 DF, p-value: < 2.2e-16

> model = lm(mpg ~.-name+displacement:cylinders+displacement:weight+acceleration:horsepower, data=Auto)

> summary(model)

Call:

lm(formula = mpg ~ . - name + displacement:cylinders + displacement:weight +

acceleration:horsepower, data = Auto)

Residuals:

Min 1Q Median 3Q Max

-9.3344 -1.6333 0.0188 1.4740 11.9723

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.725e+01 5.328e+00 -3.237 0.00131 \*\*

cylinders 6.354e-01 6.106e-01 1.041 0.29870

displacement -6.805e-02 1.337e-02 -5.088 5.68e-07 \*\*\*

horsepower 6.026e-02 2.601e-02 2.317 0.02105 \*

weight -8.864e-03 1.097e-03 -8.084 8.43e-15 \*\*\*

acceleration 6.257e-01 1.592e-01 3.931 0.00010 \*\*\*

year 7.845e-01 4.470e-02 17.549 < 2e-16 \*\*\*

origin 4.668e-01 2.595e-01 1.799 0.07284 .

cylinders:displacement -1.337e-03 2.726e-03 -0.490 0.62415

displacement:weight 2.071e-05 3.638e-06 5.694 2.49e-08 \*\*\*

horsepower:acceleration -7.467e-03 1.784e-03 -4.185 3.55e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.905 on 381 degrees of freedom

Multiple R-squared: 0.865, Adjusted R-squared: 0.8615

F-statistic: 244.2 on 10 and 381 DF, p-value: < 2.2e-16

> model = lm(mpg ~.-name+displacement:cylinders+displacement:weight+year:origin+acceleration:horsepower, data=Auto)

> summary(model)

Call:

lm(formula = mpg ~ . - name + displacement:cylinders + displacement:weight +

year:origin + acceleration:horsepower, data = Auto)

Residuals:

Min 1Q Median 3Q Max

-8.6504 -1.6476 0.0381 1.4254 12.7893

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.287e+00 9.074e+00 0.583 0.560429

cylinders 4.249e-01 6.079e-01 0.699 0.485011

displacement -7.322e-02 1.334e-02 -5.490 7.38e-08 \*\*\*

horsepower 5.252e-02 2.586e-02 2.031 0.042913 \*

weight -8.689e-03 1.086e-03 -7.998 1.54e-14 \*\*\*

acceleration 5.796e-01 1.582e-01 3.665 0.000283 \*\*\*

year 5.116e-01 9.976e-02 5.129 4.66e-07 \*\*\*

origin -1.220e+01 4.161e+00 -2.933 0.003560 \*\*

cylinders:displacement -4.368e-04 2.712e-03 -0.161 0.872156

displacement:weight 1.992e-05 3.608e-06 5.522 6.21e-08 \*\*\*

year:origin 1.630e-01 5.341e-02 3.051 0.002440 \*\*

horsepower:acceleration -6.735e-03 1.781e-03 -3.781 0.000181 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.874 on 380 degrees of freedom

Multiple R-squared: 0.8683, Adjusted R-squared: 0.8644

F-statistic: 227.7 on 11 and 380 DF, p-value: < 2.2e-16

> model = lm(mpg ~.-name-cylinders-acceleration+year:origin+displacement:weight+

+ displacement:weight+acceleration:horsepower+acceleration:weight, data=Auto)

> summary(model)

Call:

lm(formula = mpg ~ . - name - cylinders - acceleration + year:origin +

displacement:weight + displacement:weight + acceleration:horsepower +

acceleration:weight, data = Auto)

Residuals:

Min 1Q Median 3Q Max

-9.5074 -1.6324 0.0599 1.4577 12.7376

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.868e+01 7.796e+00 2.396 0.017051 \*

displacement -7.794e-02 9.026e-03 -8.636 < 2e-16 \*\*\*

horsepower 8.719e-02 3.167e-02 2.753 0.006183 \*\*

weight -1.350e-02 1.287e-03 -10.490 < 2e-16 \*\*\*

year 4.911e-01 9.825e-02 4.998 8.83e-07 \*\*\*

origin -1.262e+01 4.109e+00 -3.071 0.002288 \*\*

year:origin 1.686e-01 5.277e-02 3.195 0.001516 \*\*

displacement:weight 2.253e-05 2.184e-06 10.312 < 2e-16 \*\*\*

horsepower:acceleration -9.164e-03 2.222e-03 -4.125 4.56e-05 \*\*\*

weight:acceleration 2.784e-04 7.087e-05 3.929 0.000101 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.861 on 382 degrees of freedom

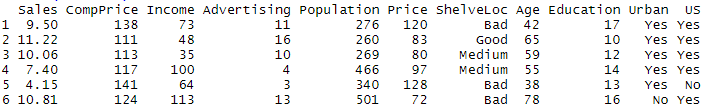
Multiple R-squared: 0.8687, Adjusted R-squared: 0.8656

F-statistic: 280.8 on 9 and 382 DF, p-value: < 2.2e-16

From all the 4 models, the last model is the only one with all variables being significant. And, based on results from a few trials not show here, it is very likely that it is the best combination of predictors and interaction terms. The R-squared statistics estimates that 87% of the changes in the response can be explained by this particular set of predictors ( single and interaction.) A higher value was not obtained from the trials. We also notice that the cylinders:displacement interaction term is not significant

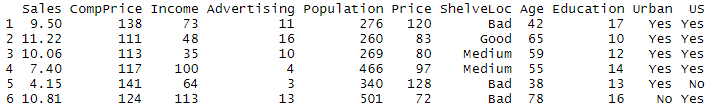
**Problem 2 (5pt)**

Use the Carseats data set to answer the following questions:



**(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.**

head(Carseats)



lmcar = lm(Sales~Price+Urban+US, data=Carseats)

summary(lmcar)

Call:

lm(formula = Sales ~ Price + Urban + US, data = Carseats)

Residuals:

Min 1Q Median 3Q Max

-6.9206 -1.6220 -0.0564 1.5786 7.0581

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 13.043469 0.651012 20.036 < 2e-16 \*\*\*

Price -0.054459 0.005242 -10.389 < 2e-16 \*\*\*

UrbanYes -0.021916 0.271650 -0.081 0.936

USYes 1.200573 0.259042 4.635 4.86e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.472 on 396 degrees of freedom

Multiple R-squared: 0.2393, Adjusted R-squared: 0.2335

F-statistic: 41.52 on 3 and 396 DF, p-value: < 2.2e-16

**(b) Provide an interpretation of each coefficient in the model (note: some of the variables are qualitative).**

1. Price: The coefficient of the “Price” variable may be interpreted by saying that the average effect of a price increase of 1 dollar is a decrease of 54.4588492 units in sales all other predictors remaining fixed.
2. Urban: The coefficient of the “Urban” variable may be interpreted by saying that on average the unit sales in urban location are 21.9161508 units less than in rural location all other predictors remaining fixed.
3. US: The coefficient of the “US” variable may be interpreted by saying that on average the unit sales in a US store are 1200.5726978 units more than in a non US store all other predictors remaining fixed.

**(c) Write out the model in equation form.**

*The model may be written as*

Sales =13.0434689 + (−0.0544588) × Price + (−0.0219162) × Urban + (1.2005727) × US + Error

*with*Urban=1 *if the store is in an urban location and*0*if not, and*US=1 *if the store is in the US and*0*if not.*

**(d) For which of the predictors can you reject the null hypothesis ?**

Based on the summary of the model,

Estimate Std. Error t value Pr(>|t|)

Price -0.054459 0.005242 -10.389 < 2e-16 \*\*\*

UrbanYes -0.021916 0.271650 -0.081 0.936

USYes 1.200573 0.259042 4.635 4.86e-06 \*\*\*

We can thus reject the null hypothesis for Price and US as their p-values are lower than the assumed confidence interval of 95%. The p-value of Urban is 0.936 which is very high and thus the predictor is insignificant in the prediction of the model and can be removed from the model.

**(e) On the basis of your answer to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the response.**

fit2 <- lm(Sales ~ Price + US, data = Carseats)

summary(fit2)

Call:

lm(formula = Sales ~ Price + US, data = Carseats)

Residuals:

Min 1Q Median 3Q Max

-6.9269 -1.6286 -0.0574 1.5766 7.0515

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 13.03079 0.63098 20.652 < 2e-16 \*\*\*

Price -0.05448 0.00523 -10.416 < 2e-16 \*\*\*

USYes 1.19964 0.25846 4.641 4.71e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.469 on 397 degrees of freedom

Multiple R-squared: 0.2393, Adjusted R-squared: 0.2354

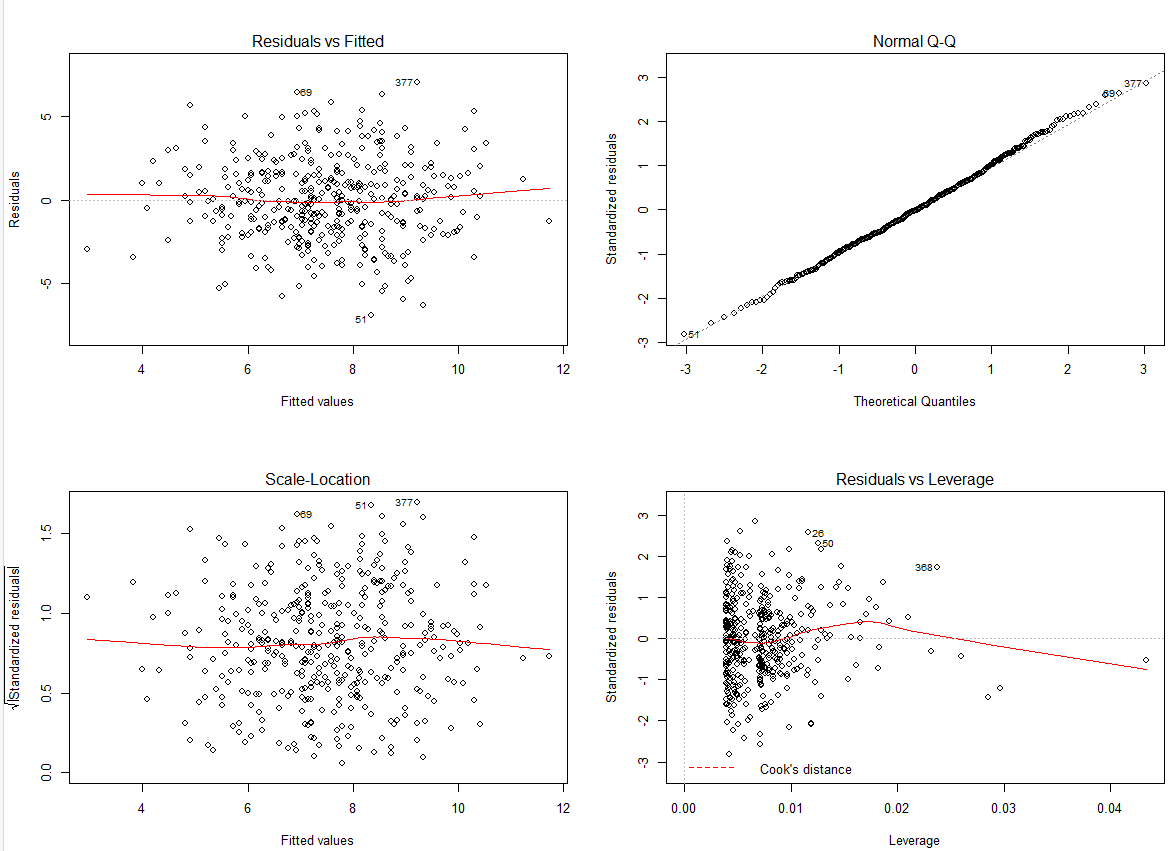
F-statistic: 62.43 on 2 and 397 DF, p-value: < 2.2e-16

**(f) How well do the models in (a) and (e) fit the data?**

The R2 for the smaller model is marginally better than for the bigger model. Essentially about 23.9262888% of the variability is explained by the model. Based on their respective R-square values(in summary tables), these two models are mediocre (only 24% change in response explained).

Thus even removing the insignificant variable did not improve the performance of the model to a very great extent and this hints that the data needs to be realized by a more flexible model than linear regression.

**(g) Is there evidence of outliers or high leverage observations in the model from (e)?**



Based on the Normal Q-Q plot and the Residuals vs Leverage plot, there are a few outliers and some leverage points that may be problematic. However, the Residual vs Leverage plot does not explicitly mention any high leverage points that we need to be alarmed about.

**Problem 3 (10pt)**

It is 1970 and you are hired by a real estate company in Boston as a data consultant. The company would like to make price estimates based on different preferences that a customer can have, so that they can assess if a customer’s budget is realistic or not. Your task is to build a model, and write an executive summary that outlines important features that affect the price of a house.

Attach “Boston” data included in the MASS library. The data contains information from 504 geographic areas. There are 14 attributes in each area of the dataset. You may find the description of each attribute using help(Boston) code in R.

Your model should have the log transformation median house value (log(medv)) as the output variable. When you are searching for a meaningful regression equation, consider the following attributes:

1. Structural properties of a house (age, number of rooms, lot size allowed by zoning laws in the area)
2. Accessibility (distance to major employment centers and closeness to highways)
3. Neighborhood (crime rate, education quality, whether it is by the Charles river or not)

You may include other terms in your model but you should discuss a model that contains above terms as needed. When you discuss the effect of an attribute on price, you need to describe the meaning of the attribute and the ranges of values associated with the attribute. For example, if there is an effect of crime rate on house prices, it would be necessary to report not only the coefficient associated with the age but what is the worst crime rate and best crime rate across different areas. Decide on one model and use it to write an executive summary.

**Executive Summary**

Summarize your findings to the president of the real estate company at a level that a non-technical person (president or a realtor) can understand. The executive summary is **at most one pages**, **single spaced and in 12-point type.** Do not include any graphs or statistical concepts. Use the following page to write your executive summary.

In the executive summary,

1. Discuss important attributes that significantly affect the value of a house.
2. Introduce a common baseline scenario such as relatively young houses near decent schools with great accessibility, and discuss its estimated price and price range.
3. Describe the attributes that yield highest house values (houses that are in top 5% in value)
4. Discuss certain bargains, such as if you sacrifice a certain neighborhood attribute you can afford to live at a bigger house for a budget that is at the mean of house values.
5. Use rounded numbers that are memorable.

**Technical Summary**

In the technical summary you speak to your peers. You show them that you performed a reasonable analysis and that you interpreted the results competently. This should not be a step by step report of what you did in R, but a summary of the most important steps in logical, not chronological, order. Even in a technical summary it is not of interest to hear, for example, how you used R; it is simply assumed that you know how to execute with available software.

The technical narrative should explain what values were used in the executive summary and how they were rounded. In addition, it should explain what contributions to the model were neglected because their effect on house value is too small.

The technical summary should, among other things, explain the final fitted model, term by term and estimate by estimate. It should mention model diagnostics that were performed and their outcomes, possibly accompanied by plots. Report data points (geographical areas) that you may have removed, if any, and why you did so.

Technical summary should at most be **five pages including any figures/tables you choose to include.** It should follow your executive summary page.

# Executive Summary

**To:** President X

**From:** Data Consultant Y

**Subject:** Analysis of Boston Suburban Housing Values, Executive Summary’

The Boston housing dataset is a dataset that has median value of the house along with 13 other parameters that could potentially be related to housing prices. These are the factors such as structural properties of the house, accessibility, and neighborhood of the house. The aim of this project is to build a linear regression model to estimate the median price of owner-occupied homes in Boston and provide insights on the parameters and how they influence the price.

The parameters considered for influence on the house prices are age, number of rooms, and lot size allowed by zoning laws in the area for the structural properties and distance to major employment centers and closeness to highways for accessibility, and crime rate, education quality, and whether it is by the Charles river or not for the neighborhood.

The most significant attributes that affect the price of the house are the age, number of rooms, distance to employment centers, crime rate, education quality, and proximity to the Charles river. Whereas, the lot size does not have a great effect on the house price. The proximity of highways has a slight effect on the price of the house as mostly everyone owns a car and this parameter is affected by people who use public transportation.

From our findings, we observe an increase in number of rooms has the most positive effect on the house price. This may be due to the fact that more rooms results in more space in house which is more desirable by people. Other factors that positively affect the house price is the proximity to the Charles river that may be because of the view it offers. The crime rate negatively affects the house price. The lower pupil to teacher ratio results in higher house price as it results in better schools. Also, houses that are closer to the highway result in higher price as it results in better accessibility to workplaces and other city centers. A notable finding was houses that are further away from employment centers have higher prices. From this observation, we infer that employment centers have a lot of noise pollution that is not very desirable by our clients that seek more peaceful and serene neighborhoods to raise a family. However, people who do not have a family prefer houses that are closer to employment centers once we model the data we see that the distance from employment centers lowers the house prices.

We observe that the majority of the clients buying houses are family oriented people because education quality and number of rooms along with lower crime rates are affecting the house prices to a great extent.

We notice that if someone is willing to live a little away from the Charles river, the person can afford a larger house with the same price range as the houses near the river. Also if someone has a car and is fine with living a little away from the highway, the person can get cheaper houses with the same number of rooms. Thus we notice a few compromises can help clients achieve their dream house at a lower price.

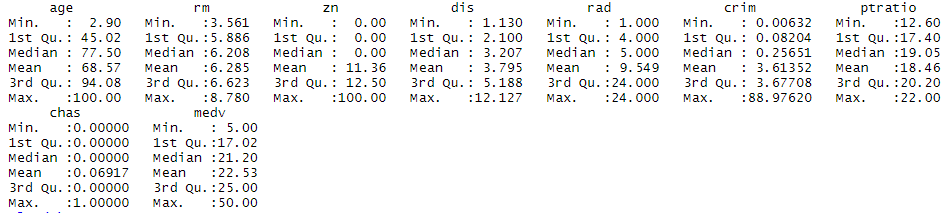
# Technical Summary

The Boston housing dataset is used for this project and we consider certain parameters out of the dataset that affect the structural properties such as the following:

1. Structural properties of a house (age, number of rooms, lot size allowed by zoning laws in the area)
2. Accessibility (distance to major employment centers and closeness to highways)
3. Neighborhood (crime rate, education quality, whether it is by the Charles river or not)

There are 506 observations in the data for 9 variables including the median price of house in Boston. There are 8 numerical variables in our dataset and 1 categorical variable. The aim of this project is to understand the data and build a linear regression model estimate the median price of owner-occupied homes in Boston.

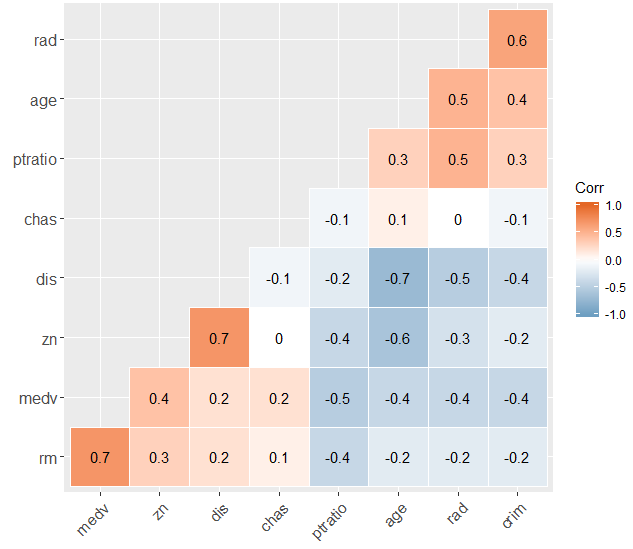
The summary statistics of the dataset were first observed to understand the range of values of the variables in the dataset.



From the summary we observe the following information:

1. Age : Around 69% of the houses we built before 1940 which indicates that this is a very outdated dataset and the preferences of clients may have changed now
2. Rm : Houses have an average of 6 rooms per dwelling with a range from 3 rooms to 9 rooms. This indicates that the neighborhoods had relatively large houses and cater for more family oriented clientele
3. Zn : Around 11% of the houses are in neighborhoods have lots greater than 25000 sqft. This indicates that the neighborhoods were relatively small and more focused towards smaller communities
4. Dis : The weighted mean of distance to 5 Boston employment centers is around 4
5. Rad : The accessibility to highways is averaged at around 10 and ranging from 1 to 24 that indicates that the houses are fairly accessible to the highways
6. Crim : The per capita crime rate is averaged at around 4 with a maximum of 89. However the 3rd quantile statistics indicate that the neighborhoods are relatively crime free
7. Ptratio : The mean pupil to teacher ratio is around 19 meaning that there is 1 teacher for every 19 students. This indicates good quality of education in the Boston area.
8. Chas : 69% of the houses are close to the Charles river.
9. Medv : The house prices rance from $5000 to $50000 with average at $23000

The correlation matrix was then observed and we notice that housing value has a strong positive correlation with rm(number of rooms). It is expected, as a spacious house with more rooms would have a higher valuation.



From the matrix, we observe certain variables are highly correlated among eachother that can create problems in our model. Distance from employment is highly correlated with crime rates that indicates that neighborhoods that are away from employment centers have higher crime rates. Neighborhoods away from highways also have higher crime rates.

We now observe all parameters that are positively correlated to the response medv. These are rm, zn, dis, chas and the ones negatively correlated to the medv are age, rad, crim, ptratio. This indicates that houses with more rooms, that have a larger neighborhood, that are closer to the city center, and that are in proximity to the Charles river have higher median prices. Houses that have a higher crime rate, a higher pupil to teacher ratio, more age, more distance from the highway reduce the median house price.

Multivariate regression analysis was then performed without any standardization to generate results that acted as a benchmark. These results were later compared with the results generated from other regression techniques to understand the accuracy of the predictions.

Call:

lm(formula = log(medv) ~ ., data = x)

Residuals:

Min 1Q Median 3Q Max

-0.90411 -0.09235 -0.00237 0.09958 1.40785

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.4471267 0.1837368 13.319 < 2e-16 \*\*\*

age -0.0038778 0.0005999 -6.464 2.44e-10 \*\*\*

rm 0.2574523 0.0171603 15.003 < 2e-16 \*\*\*

zn 0.0003593 0.0006895 0.521 0.602491

dis -0.0261205 0.0092381 -2.827 0.004881 \*\*

rad -0.0031750 0.0018620 -1.705 0.088793 .

crim -0.0136077 0.0016557 -8.219 1.80e-15 \*\*\*

ptratio -0.0325271 0.0063319 -5.137 4.02e-07 \*\*\*

chas 0.1468136 0.0438925 3.345 0.000886 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2449 on 497 degrees of freedom

Multiple R-squared: 0.6468, Adjusted R-squared: 0.6411

F-statistic: 113.7 on 8 and 497 DF, p-value: < 2.2e-16

We notice that the variables zn is not affecting the model prediction in any way as it is clearly failing the t test with high p value of 0.6. Also the parameter rad with p value 0.08 is in the borderline and is not a strong parameter.

We now drop these two variables and recreate the model to compare the accuracy.

Call:

lm(formula = log(medv) ~ . - zn - rad, data = x)

Residuals:

Min 1Q Median 3Q Max

-0.92600 -0.09895 0.00607 0.10194 1.38227

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.5020776 0.1811080 13.815 < 2e-16 \*\*\*

age -0.0039890 0.0005945 -6.710 5.29e-11 \*\*\*

rm 0.2569318 0.0169400 15.167 < 2e-16 \*\*\*

dis -0.0211954 0.0079667 -2.660 0.00805 \*\*

crim -0.0149082 0.0014199 -10.499 < 2e-16 \*\*\*

ptratio -0.0370843 0.0056251 -6.593 1.10e-10 \*\*\*

chas 0.1442719 0.0439110 3.286 0.00109 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

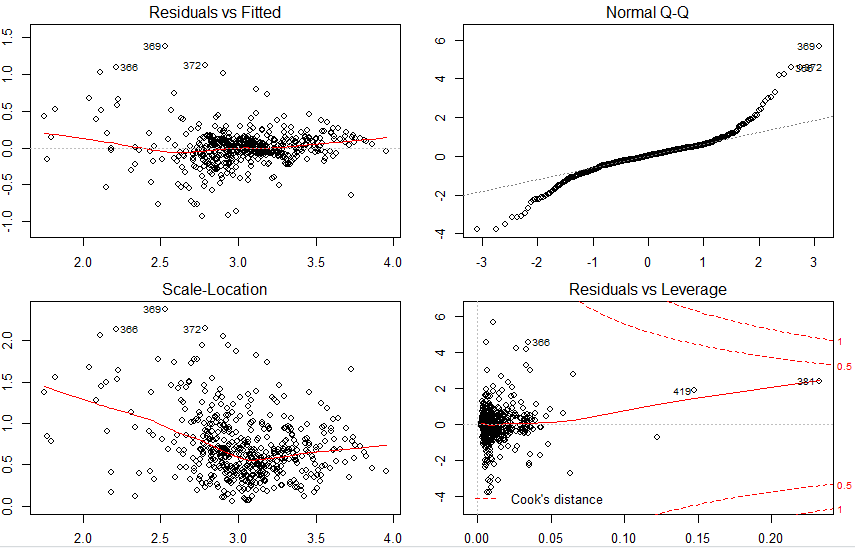
Residual standard error: 0.2451 on 499 degrees of freedom

Multiple R-squared: 0.6446, Adjusted R-squared: 0.6404

F-statistic: 150.9 on 6 and 499 DF, p-value: < 2.2e-16

Dropping those parameters did not improve the R2 value to a very great extent indicating that the model may not be very linear and thus the multiple linear regression is not able to understand the data to a very great extent.

Finally we run the diagnostic plots to find any problems in the model and seek for strategies to address them.



These plots give us the following insights:

1. The data has a lot of non-linearity in it with a clear pattern in residual vs fitted
2. The data has a lot of outliers with the Q-Q plot having a curve at the head and tail
3. The data has a lot of outliers and high leverage points as seen from residual vs leverage plot.

We notice that even with the log transformation of the response variable medv, although it helped solve some of the problems, it still lead to a lot of non-linearity and outliers and these can be improved by addressing the outliers that is out of the scope of this report.