Regression_Assignment

Venu

2022-11-11

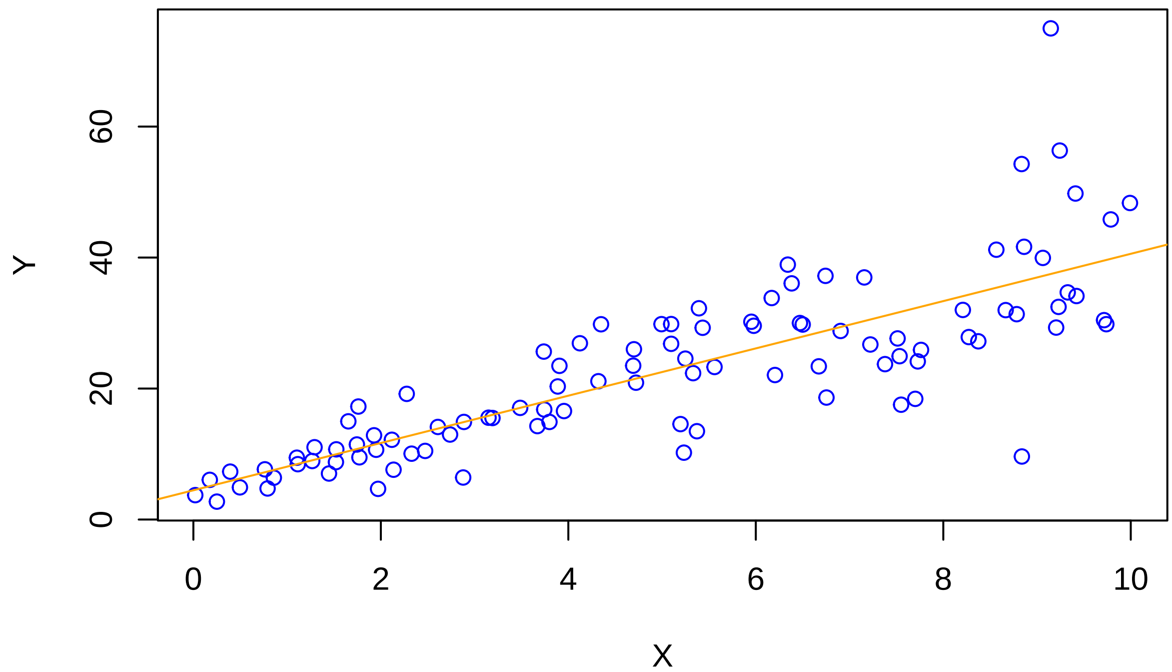
---

1. Run the following code in R-studio to create two variables X and Y. set.seed(2017) X=runif(100)*10* *Y=X*4+3.45 Y=rnorm(100)*0.29*Y+Y

```
set.seed(2017)
X=runif(100)*10
Y=X*4+3.45
Y=rnorm(100)*0.29*Y+Y
```

a) Plot Y against X. Include a screenshot of the plot in your submission. Using the File menu you can save the graph as a picture on your computer. Based on the plot do you think we can fit a linear model to explain Y based on X? (5 Marks)

```
plot(Y~X,xlab='X',ylab='Y',col='blue')
abline(lsfit(X, Y),col = "orange")
```



b)Construct a simple linear model of Y based on X. Write the equation that explains Y based on X. What is the accuracy of this model? (5 Marks)

Y=4.4655+3.6108*X Accuracy is 0.6517 or 65%

```
fit <- lm(Y ~ X)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.755  -3.846  -0.387   4.318  37.503
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4655     1.5537   2.874  0.00497 **
## X             3.6108     0.2666  13.542  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.756 on 98 degrees of freedom
## Multiple R-squared:  0.6517, Adjusted R-squared:  0.6482
## F-statistic: 183.4 on 1 and 98 DF,  p-value: < 2.2e-16
```

c)How the Coefficient of Determination, R2, of the model above is related to the correlation coefficient of X and Y? (5 marks)

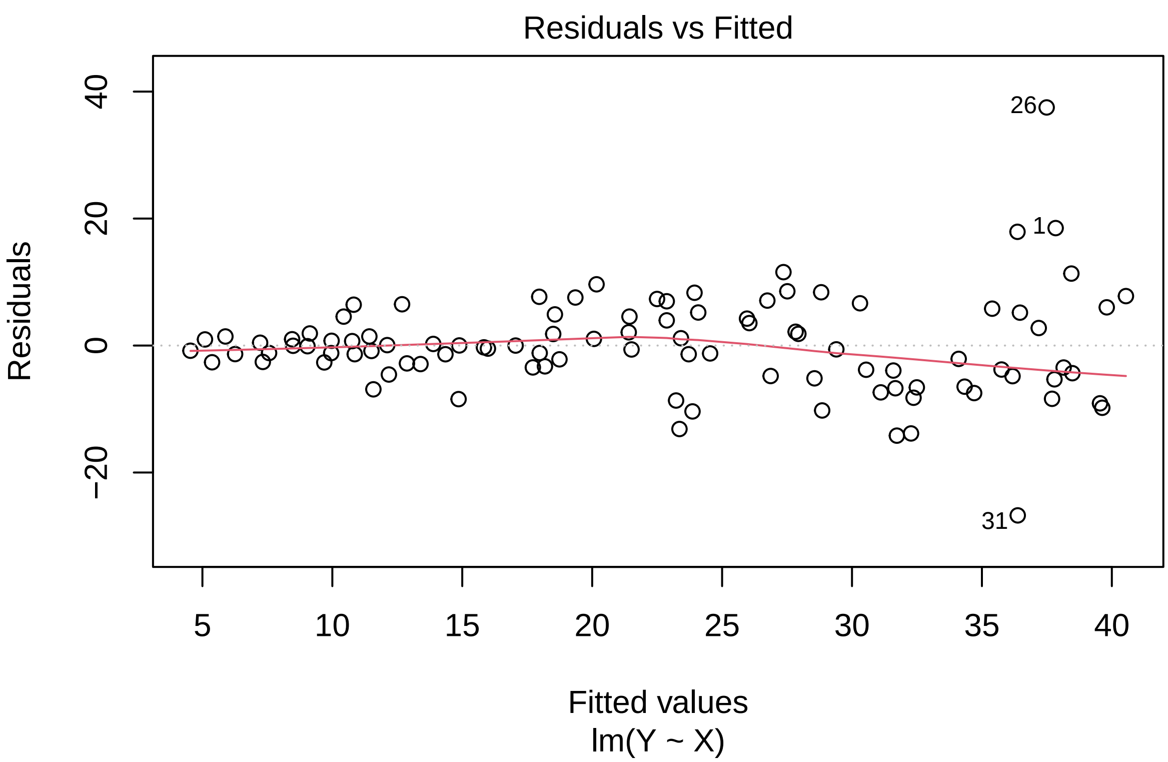
```
cor(X,Y)^2
```

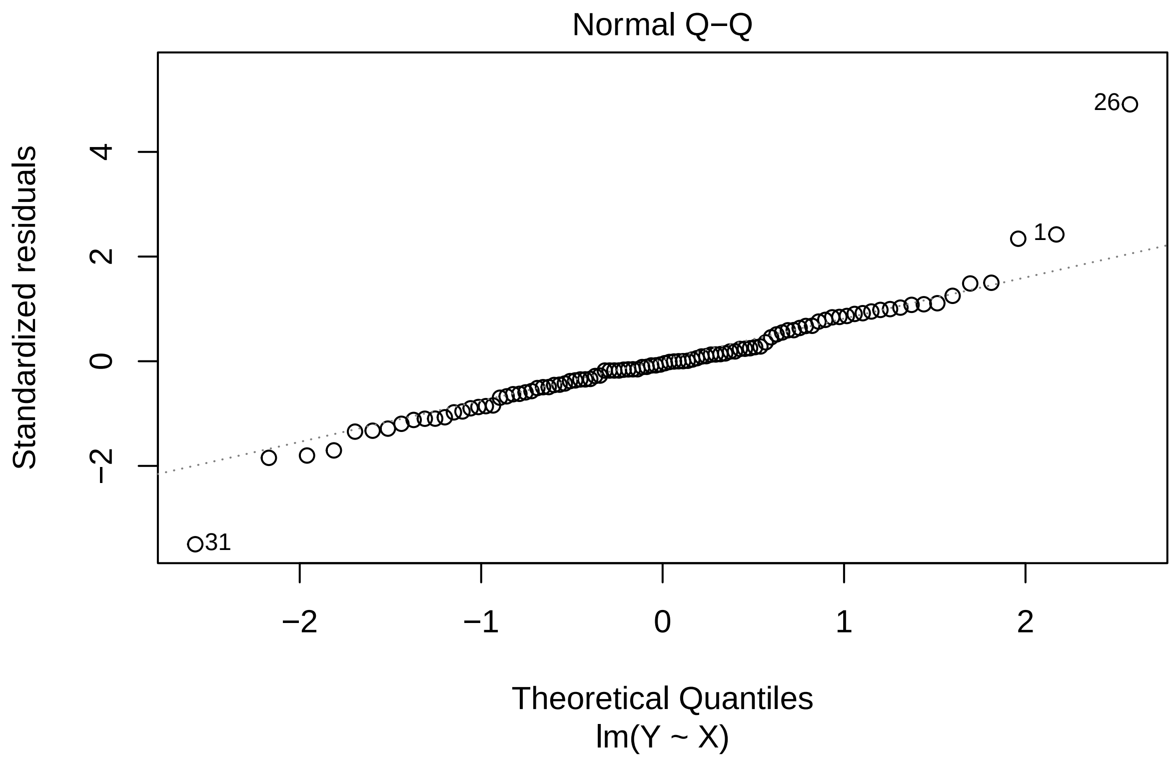
```
## [1] 0.6517187
```

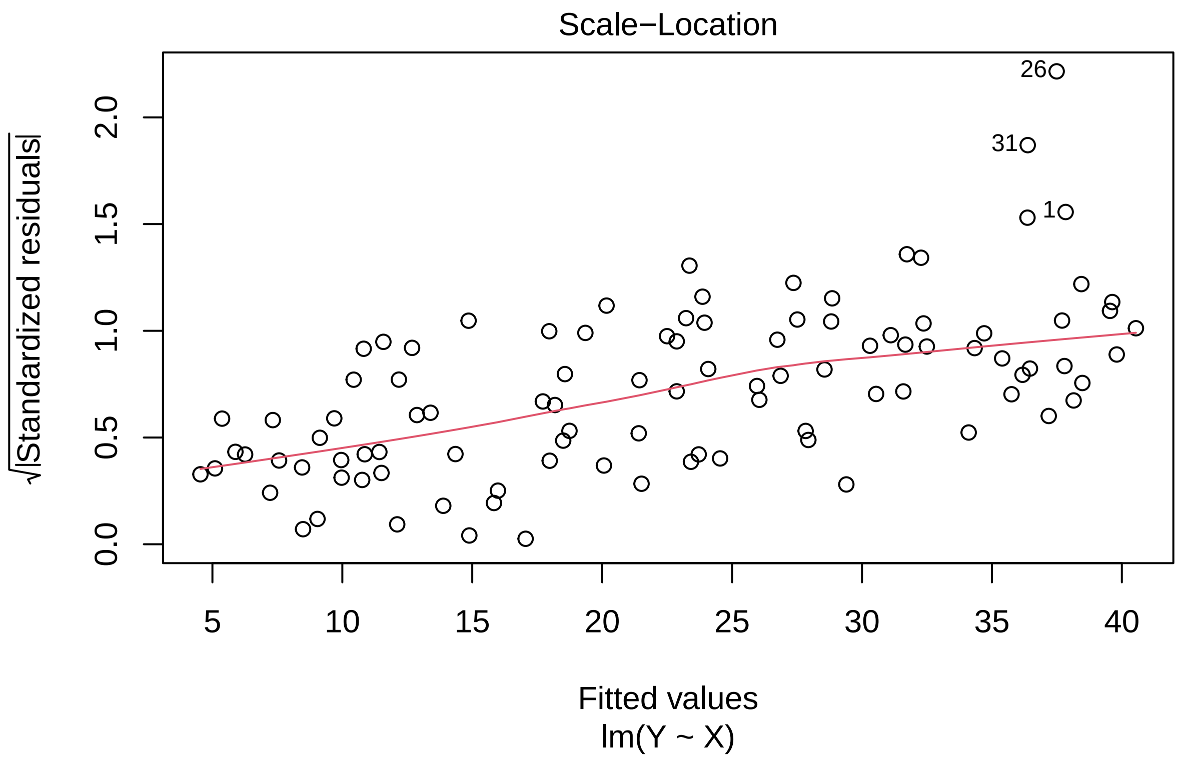
Solution: The square of correlation coefficient is same as coefficient of determination 65.17% #Coefficient of Determination= (Correlation Coefficient)2

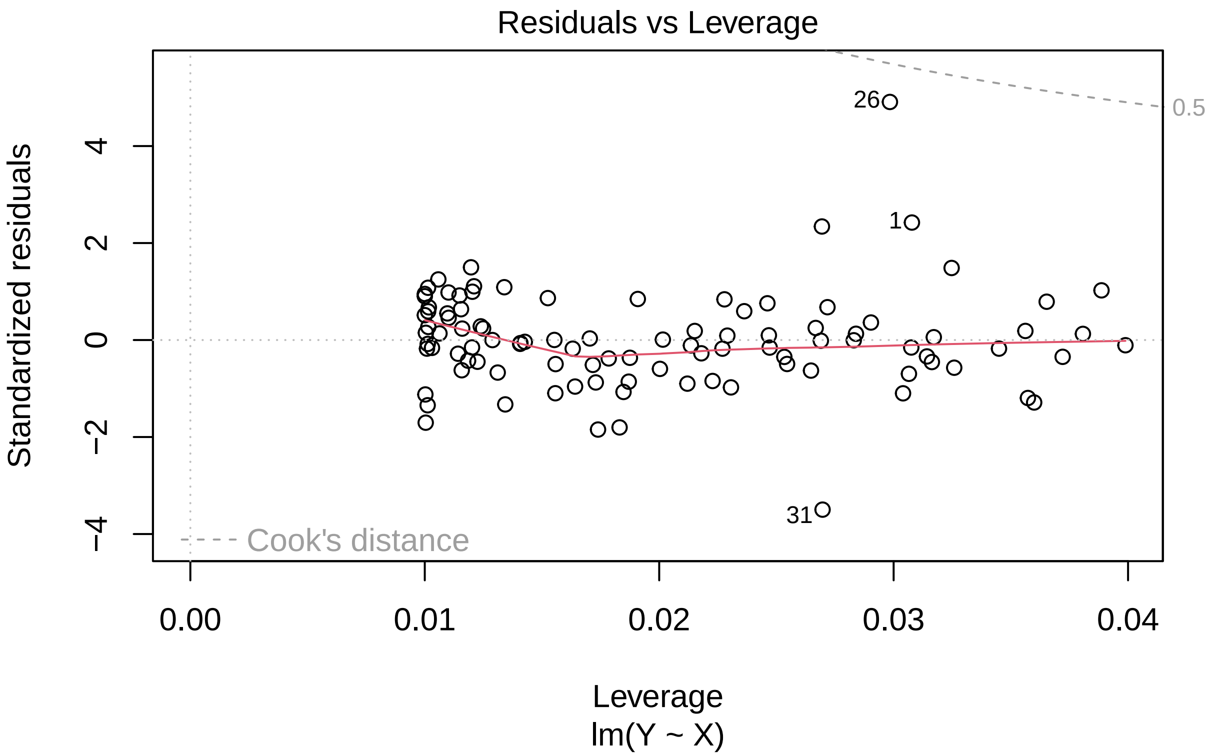
d)Investigate the appropriateness of using linear regression for this case (10 Marks). You may also find the story here relevant. More useful hints: #residual analysis, #pattern of residuals, #normality of residuals.

```
plot(fit)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Y ~ X)
26
1
31
4
2
0
−2
−2
−1
0
1
2


Scale−Location
26
31
1
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
5
10
15
20
25
30
35
40
Fitted values
lm(Y ~ X)



Residuals vs Fitted Plot

When using the plot() function, the first plot is the Residuals versus Fitted plot, which shows whether there are non-linear patterns in the residuals and can be used to look for underlying patterns that could indicate a problem with the model. This will determine whether the need for linear data, which is necessary for a valid linear regression, is satisfied.

Normal Q–Q (quantile-quantile) Plot

The Q-Q Plot will demonstrate that residuals are expected to have a normal distribution. It is a good sign that residuals are normally distributed if they plot as nearly a straight line. The Q-Q plot for our model exhibits strong alignment to the line, with a few slightly off-center dots at the top. A reasonable alignment and most likely not noteworthy.

Scale-Location

The homoscedasticity (equal variance) assumption of linear regression, which states that the residuals have equal variance along the regression line, is tested by this graphic. Alternatively, it is known as the Spread-Location plot. The residuals for our model are rather evenly distributed above and below a pleasing horizontal line; however, the beginning of the line has fewer points and so slightly less variation. ***

2. For this inquiry, the'mtcars' dataset will be used. Your R distribution already contains the dataset. Some of the characteristics of several cars are displayed in the dataset. The first six rows of the dataset are shown in the following as small samples. You can get a description of the dataset here.

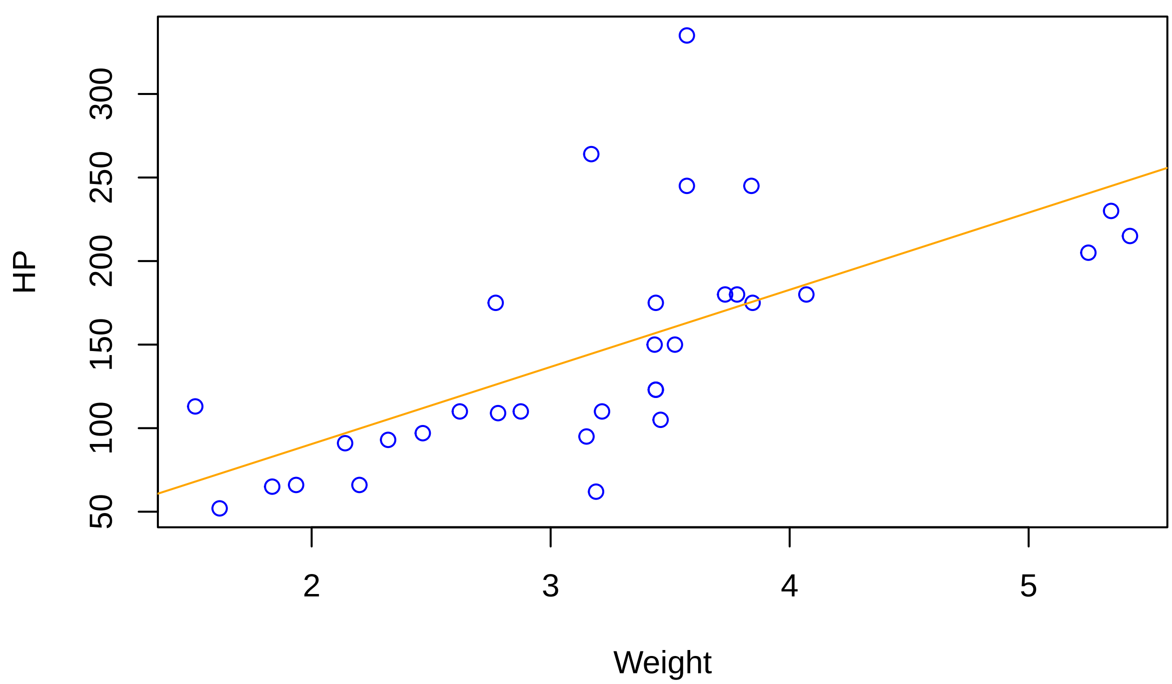
```
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

a) James wants to buy a car. He and his friend, Chris, have different opinions about the Horse Power (hp) of cars. James think the weight of a car (wt) can be used to estimate the Horse Power of the car while Chris thinks the fuel consumption expressed in Mile Per Gallon (mpg), is a better estimator of the (hp). Who do you think is right? Construct simple linear models using mtcars data to answer the question. (10 marks)

Building a model based on James estimation:

```
plot(mtcars$hp~mtcars$wt,xlab='Weight',ylab='HP',col='blue')
abline(lsfit(mtcars$wt,mtcars$hp),col = "orange")
```



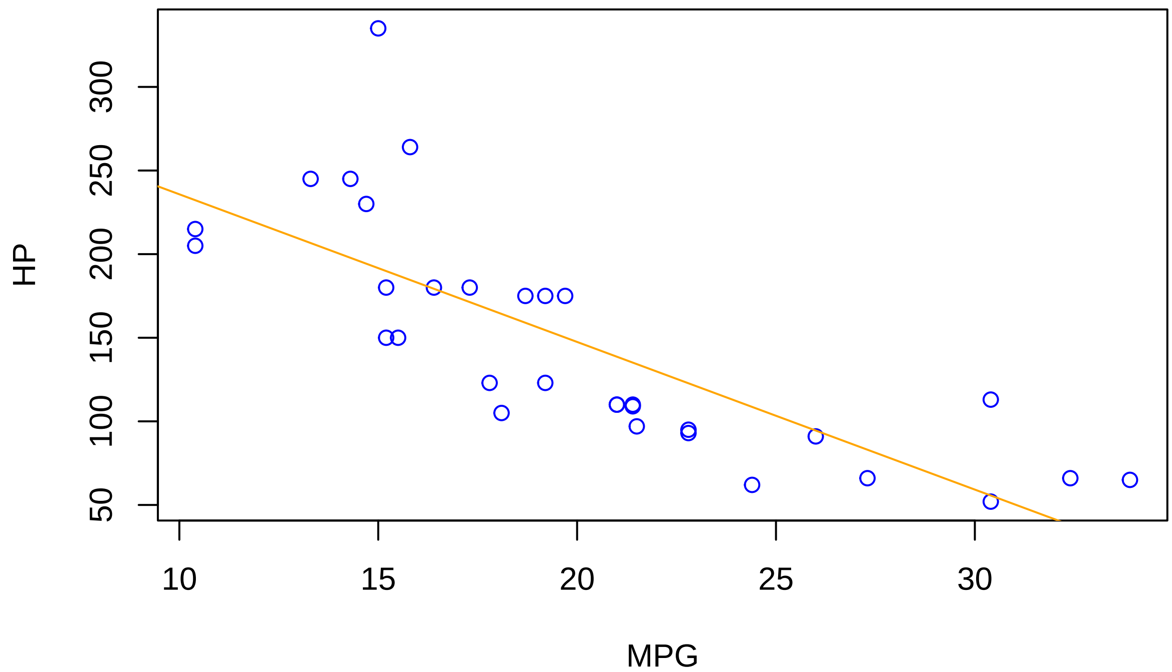
```r
Model1<-lm(formula =hp~wt, data = mtcars )
summary(Model1)
```

```
##
## Call:
## lm(formula = hp ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.430 -33.596 -13.587   7.913 172.030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.821     32.325  -0.056    0.955
## wt            46.160      9.625   4.796 4.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.44 on 30 degrees of freedom
## Multiple R-squared:  0.4339, Adjusted R-squared:  0.4151
## F-statistic:    23 on 1 and 30 DF,  p-value: 4.146e-05
```

Accuracy of Model1 is 0.4339.

Building a model based on Chris estimation:

```r
plot(mtcars$hp~mtcars$mpg,xlab='MPG',ylab='HP',col='blue')
abline(lsfit(mtcars$mpg, mtcars$hp),col = "orange")
```




```
Model2<-lm(formula =hp~mpg, data = mtcars )
summary(Model2)
```

```
##
## Call:
## lm(formula = hp ~ mpg, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59.26 -28.93 -13.45  25.65 143.36
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   324.08      27.43  11.813 8.25e-13 ***
## mpg            -8.83       1.31  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.95 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

Accuracy of the model2 is 0.6024

Conclusion: Chris Estimation is fairly accurate enough. Hence, Chris is right.

b) Build a model that uses the number of cylinders (cyl) and the mile per gallon (mpg) values of a car to predict the car Horse Power (hp).

I. Using this model, what is the estimated Horse Power of a car with 4 calendar and mpg of 22? (5 mark)
II. Construct an 85% confidence interval of your answer in the above question. Hint: use the predict function (5 mark)

```
Model3<-lm(hp~cyl+mpg,data = mtcars)
summary(Model3)
```

```
##
## Call:
## lm(formula = hp ~ cyl + mpg, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.72 -22.18 -10.13  14.47 130.73
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.067     86.093   0.628  0.53492
## cyl           23.979      7.346   3.264  0.00281 **
## mpg           -2.775      2.177  -1.275  0.21253
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.22 on 29 degrees of freedom
## Multiple R-squared:  0.7093, Adjusted R-squared:  0.6892
## F-statistic: 35.37 on 2 and 29 DF,  p-value: 1.663e-08
```

```
estimated_HP<-predict(Model3,data.frame(cyl=4,mpg=22))
estimated_HP
```

```
##        1
## 88.93618
```

```
predict(Model3,data.frame(cyl=4,mpg=22),interval = "prediction",level = 0.85)
```

```
##        fit      lwr      upr
## 1 88.93618 28.53849 149.3339
```

3. For this question, we are going to use BostonHousing dataset. The dataset is in 'mlbench' package, so we first need to install the package, call the library and load the dataset using the following commands

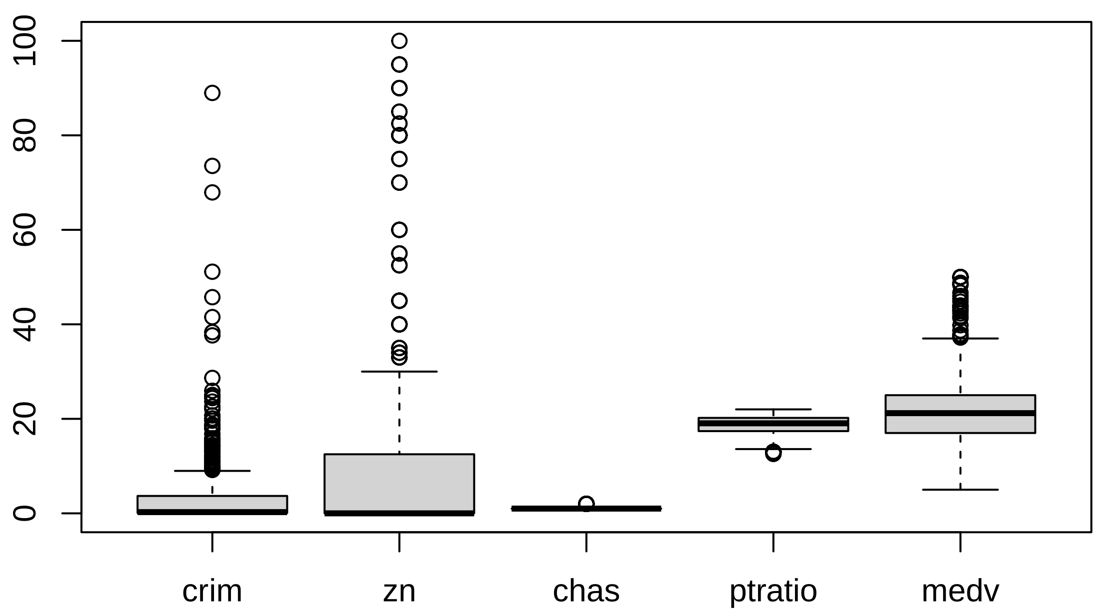
```
#install.packages('mlbench')
library(mlbench)
```

```
## Warning: package 'mlbench' was built under R version 4.2.2
```

```r
data(BostonHousing)
str(BostonHousing)
```

```
## 'data.frame':    506 obs. of  14 variables:
##  $ 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    : num  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  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 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

```r
#Let's look at the variation in the values of various variables present in the dataset. This is achieve
boxplot(BostonHousing[,c(1,2,4,11,14)])
```



a) Build a model to estimate the median value of owner-occupied homes (medv)based on the following variables: crime crate (crim), proportion of residential land zoned for lots over 25,000 sq.ft (zn), the local pupil-teacher ratio (ptratio) and weather the whether the tract bounds Chas River(chas). Is this an accurate model? (Hint check R2) (5 marks)

```
set.seed(123)
Model4<-lm(medv~crim+zn+ptratio+chas,data = BostonHousing)
summary(Model4)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + ptratio + chas, data = BostonHousing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.282  -4.505  -0.986   2.650  32.656
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49.91868    3.23497  15.431  < 2e-16 ***
## crim        -0.26018    0.04015  -6.480 2.20e-10 ***
## zn           0.07073    0.01548   4.570 6.14e-06 ***
## ptratio     -1.49367    0.17144  -8.712  < 2e-16 ***
## chas1        4.58393    1.31108   3.496 0.000514 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.388 on 501 degrees of freedom
## Multiple R-squared:  0.3599, Adjusted R-squared:  0.3547
## F-statistic: 70.41 on 4 and 501 DF,  p-value: < 2.2e-16
```

The Model Accuracy is 0.3599. The Model is not Accurate enough.

b) Use the estimated coefficient to answer these questions?

I. Imagine two houses that are identical in all aspects but one bounds the Chas River and the other does not. Which one is more expensive and by how much? (5 marks)

Answer: Chas is a two-level (levels "0" and "1") factor. One who is confined by the Chas River is denoted by "1," whereas those who are not are denoted by "0." The data description states that the median value of owner-occupied dwellings is $1,000 and that the chas1 coefficient is 4.58393. The pricey outcome of coefficient multiplication is 4583.93 dollars.

II. Imagine two houses that are identical in all aspects but in the neighborhood of one of them the pupil-teacher ratio is 15 and in the other one is 18. Which one is more expensive and by how much? (Golden Question: 10 extra marks if you answer)

Answer: House prices decline by 1.49367, or 1493.67, for every unit rise in the ptratio (in thousands). If ptratio is 15, a decrease of 15 * 1493.67 will result in 22405.05. The decrease will be equal to 18 * 1493.67, or 26886.06, if the ptratio is 18. As a result, if pt ratio is 15, it is $4481.01 more expensive than pt ratio 18.

c) Which of the variables are statistically important (i.e. related to the house price)? Hint: use the p-values of the coefficients to answer.(5 mark)

Answer: Yes, none of the variables' p-values are equal to zero, which allows us to confidently reject the null hypothesis that there is no correlation between the house price and the other factors in the model. Therefore, each variable has statistical significance.

d) Use the anova analysis and determine the order of importance of these four variables.(5 marks)

```
anova(Model4)
```

```
## Analysis of Variance Table
##
## Response: medv
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## crim        1  6440.8  6440.8 118.007 < 2.2e-16 ***
## zn          1  3554.3  3554.3  65.122 5.253e-15 ***
## ptratio     1  4709.5  4709.5  86.287 < 2.2e-16 ***
## chas        1   667.2   667.2  12.224 0.0005137 ***
## Residuals 501 27344.5    54.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: As we can see, the crim variable considerably explains more variability (sum squared) than any other variable. This could be explained by introducing the crim, which greatly enhanced the model. Even still, residuals demonstrate that a sizable fraction of the variability is unaccounted for.

The order of importance is crim, ptratio,zn, chas