Regression Analysis

Purpose of this Project:

The purpose of this Project is to build a Linear Regression Model using 'mtcars' and 'BostonHousing' dataset to answer relevant questions based on a given scenario.

About mtcars dataset: *The mtcars dataset is a built-in dataset in R. It comprises 11 features of 32 automobiles from the 1974 Motor Trend US magazine. .*

Scenario 1:

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.

```
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
```

```
#constructing linear regression model to determine hp based on weight of the car:
linear_model1<- lm(hp~wt,data=mtcars)
summary(linear_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
```

```r
#constructing linear regression model to determine hp based on Mile per Gallon(mpg) of the car:
linear_model2<-lm(hp~mpg,data=mtcars)
summary(linear_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
```

To best determine which variable can be used to estimate the horse power of a car, we are considering R square value as it implies the proportion of variability of the dependent variable accounted for the independent variable.

R square value to estimate horse power based on weight is 43.39 percent whereas the R square value to estimate horse power based on miles per Gallon is 60.24 percent.

Therefore, it is clear to say that the horse power can be best estimated with the value of mpg and not based on the weight of the car.

Hence, Chris is right about estimating the horse power of the car

Constructing a model to predict the car horse power based on number of cylinders and miles per Gallon:

```r
linear_model3<- lm(hp~cyl+mpg,data = mtcars)
summary(linear_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
```

Linear equation:

$$hp = 54.067 + 23.979 * X_1 - 2.775 * X_2$$

$$where X_1 = cyl, X_2 = mpg$$

Estimated horsepower of a car with 4 cylinders and mpg of 22:

```
predicted_hp_value<-predict(linear_model3,data.frame(cyl=c(4),mpg=c(22)))

predicted_hp_value
```

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

The estimated horse power of a car with 4 cylinders and 22 mpg is 88.93618

About BostonHousing Dataset *The Boston Housing Dataset is a derived from information collected by the U.S. Census Service concerning housing in the area of Boston MA.Each of the 506 rows in the dataset describes a Boston suburb or town, and it has 14 columns with information such as average number of rooms per dwelling, pupil-teacher ratio, and per capita crime rate.*

Building a model to estimate the median value of owner occupied homes based on crime rate, proportion of residential land zoned for lots over 25,000 sq.ft,the local pupil-teacher ratio (ptratio) and weather the whether the tract bounds Chas River:

```
library(mlbench)
data(BostonHousing)

linear_model4<-lm(medv~crim+zn+ptratio+chas,data=BostonHousing)

summary(linear_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 above model has R square value of 35.99 percent which is a relatively low.R square is the coefficient of determination used in Regression Model as a performance of measure to explain the amount of variability between dependent and independent variables. Since R square is relatively low, we don't consider this as a good model

Identifying which of the two identical houses is more expensive:

To identify which home is expensive comparing the one that bounds the Chas river and the one's do not, we consider the coefficient of the chas value in the above linear model. The coefficient is 4.58393, indicates that the houses that bounds by the Chas river are 4.58393 times more expensive than the houses which do not bounds by the river.

Moreover, in the dataset, the values of chas river are 1 or 0 which means the houses which bounds by the river are assigned a value of 1, otherwise 0. So for the houses which do not bounds by the river are going to have 0 times change in their value

Finding which of the variables are statistically important:

All the variables including crime rate,proportion of residential land zoned for lots over 25,000 sq.ft,the local pupil-teacher ratio,the tract bounds Chas River are statistically important as all of them has very low P value

(d)Determining the order of importance of the 4 variables using ANOVA analysis:

```
anova_lm<-anova(linear_model4)
anova_lm
```

```
## 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
```

The importance of variables can be determined by their Sum of Squares value. Higher the Sum of squares, the more important is the variable in estimating the value of a dependent variable

Order of importance of variables:

crim-per capita crime rate by town

ptratio-pupil-teacher ratio by town.

zn-proportion of residential land zoned for lots over 25,000 sq.ft.

Chas-Charles River dummy variable