Regression Analysis

Purpose of this Project:

The objective of this Project is to construct a Linear Regression Model utilizing the 'mtcars' and 'BostonHousing' datasets in order to address pertinent inquiries in accordance with a provided scenario.

About mtcars dataset: *The R programming language has the mtcars dataset, which contains information on 32 automobiles from the 1974 Motor Trend US magazine. It has 11 features about these vehicles.*

Scenario 1:

Sam is interested in purchasing a car, but he and his friend Dean have differing views on how to estimate the car's Horse Power (hp). Sam believes that the weight of the car (wt) can be used as an indicator of its Horse Power, while Dean argues that the fuel consumption, measured in Mile Per Gallon (mpg), is a more accurate estimator. To determine who is correct, simple linear models can be constructed using the mtcars data.

```
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
```

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

```r
#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 find out the most suitable variable for estimating a car's horsepower, we are looking at the R square value. This value indicates the proportion of variability in the dependent variable that can be explained by 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, Dean 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:

```
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 *1. The Boston Housing Dataset contains data from the U.S. Census Service about housing in Boston MA. It has 506 rows, each representing a suburb or town in Boston, with 14 columns including details like average number of rooms, pupil-teacher ratio, and crime rate per capita.*

Creating a model to predict the middle value of homes that are owned, considering factors like crime rate, the amount of residential land zoned for large lots, the pupil-teacher ratio, and whether the area is adjacent to the 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
```

1. The model above has an R square value of 35.99 percent, indicating a low level of performance. R square is a measure used in Regression Model to show how much variability exists between dependent and independent variables. With a relatively low R square value, this model is not considered good.

Identifying which of the two identical houses is more expensive:

1. To determine the price difference between homes near the Chas river and those that are not, we look at the coefficient of the Chas value in the linear model. With a coefficient of 4.58393, it means that houses near the Chas river are 4.58393 times pricier than those that are not near the river, The dataset shows that houses along the chas river are assigned a value of 1, while those not along the river are assigned a value of 0. Houses not by the river will not see a change in their value.

Finding which of the variables are statistically important:

All variables, including crime rate, residential land zoning, pupil-teacher ratio, and proximity to Chas River, are statistically significant due to their low P values.

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
```

Variables can be ranked by their Sum of Squares value, indicating their importance. The greater the Sum of Squares, the more influential the variable is in predicting the value of a dependent variable.

- 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