Assignment 2 Business Analytics - Regression

Srushti Padade

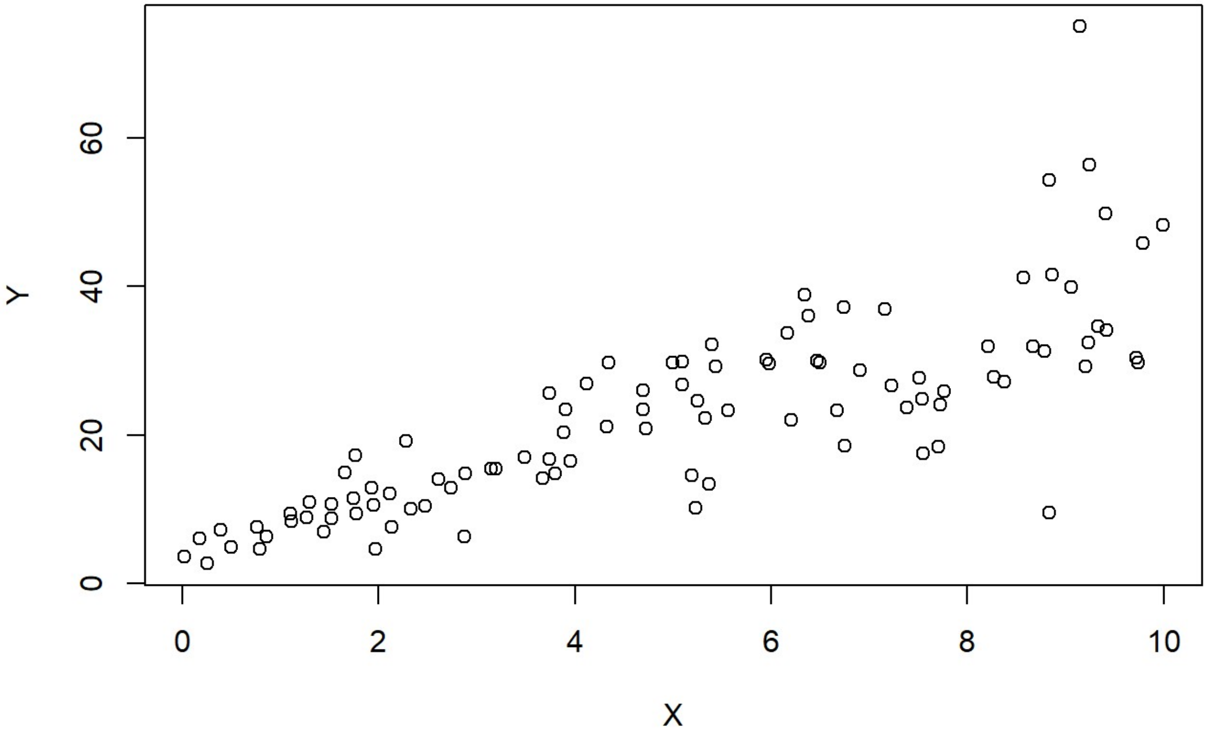
11/10/2019

Question 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
```

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?

```
plot(X,Y)
```



- On basis of the plotted graph we can see that the points of Y against X can form a linear line.
- Thus assuming this, we can fit the linear regression model to explain the Y based on X.

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?

```
Linear_Model <- lm(formula = Y~X)
summary(Linear_Model)
```

```
## 
## 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
```

- Equation : Y = 4.4655 + 3.6108 X

- Accuracy of the Model : 65.17 % (which tells that the model is relevently significant.)

c. How the Coefficient of Determination, R2, of the model above is related to the correlation coefficient of X and Y?

```
cor(Y,X)
```

```
## [1] 0.807291
```

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

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

- Solution:

- Co- efficient of Determination = Co-efficient of Correlation ^ 2
- Thus, here co-efficient of Determination is 0.6517187

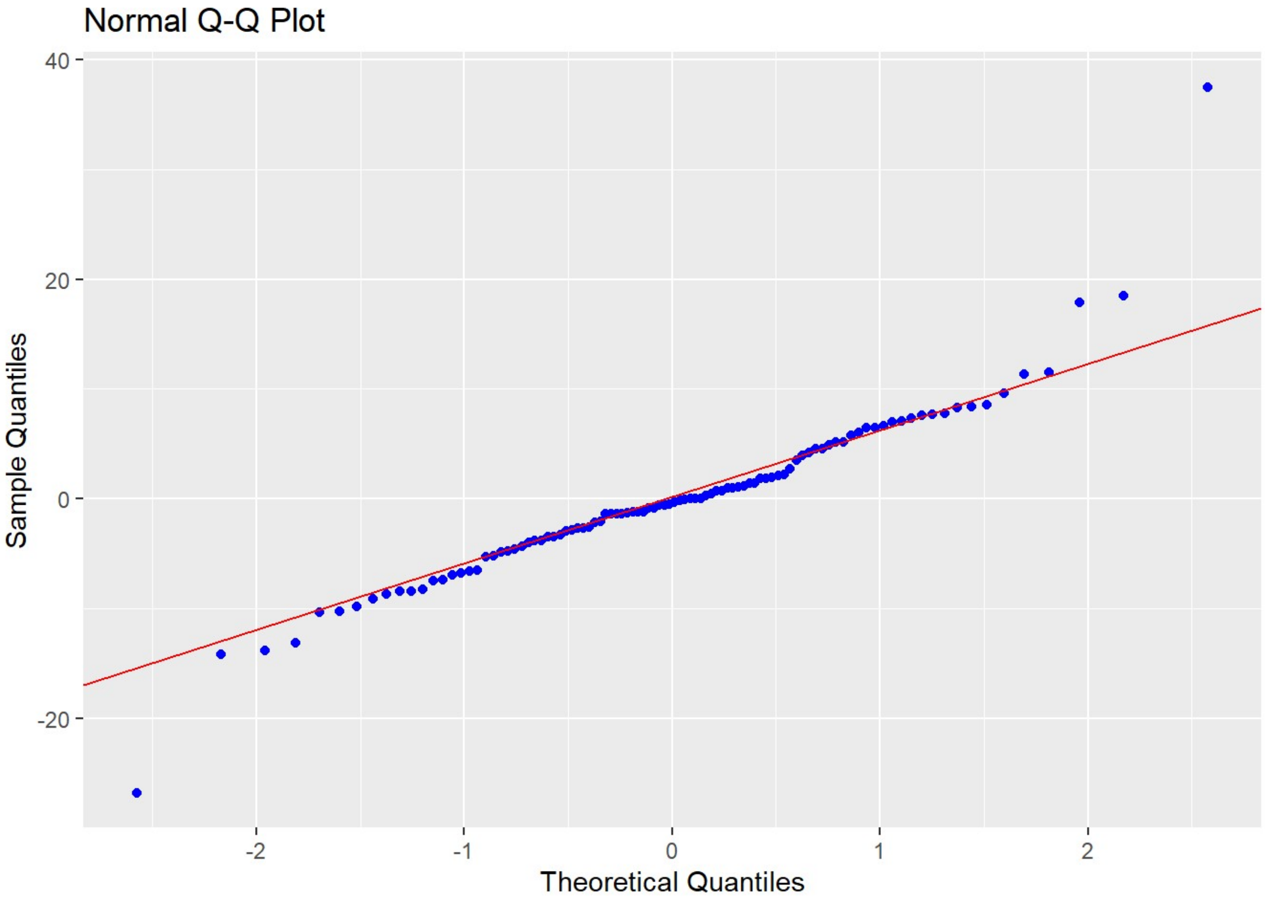
d. Investigate the appropriateness of using linear regression for this case (10 Marks). You may also find the story here relevant. More useful hints: #residual analysis, #pattern of residuals, #normality of residuals.

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

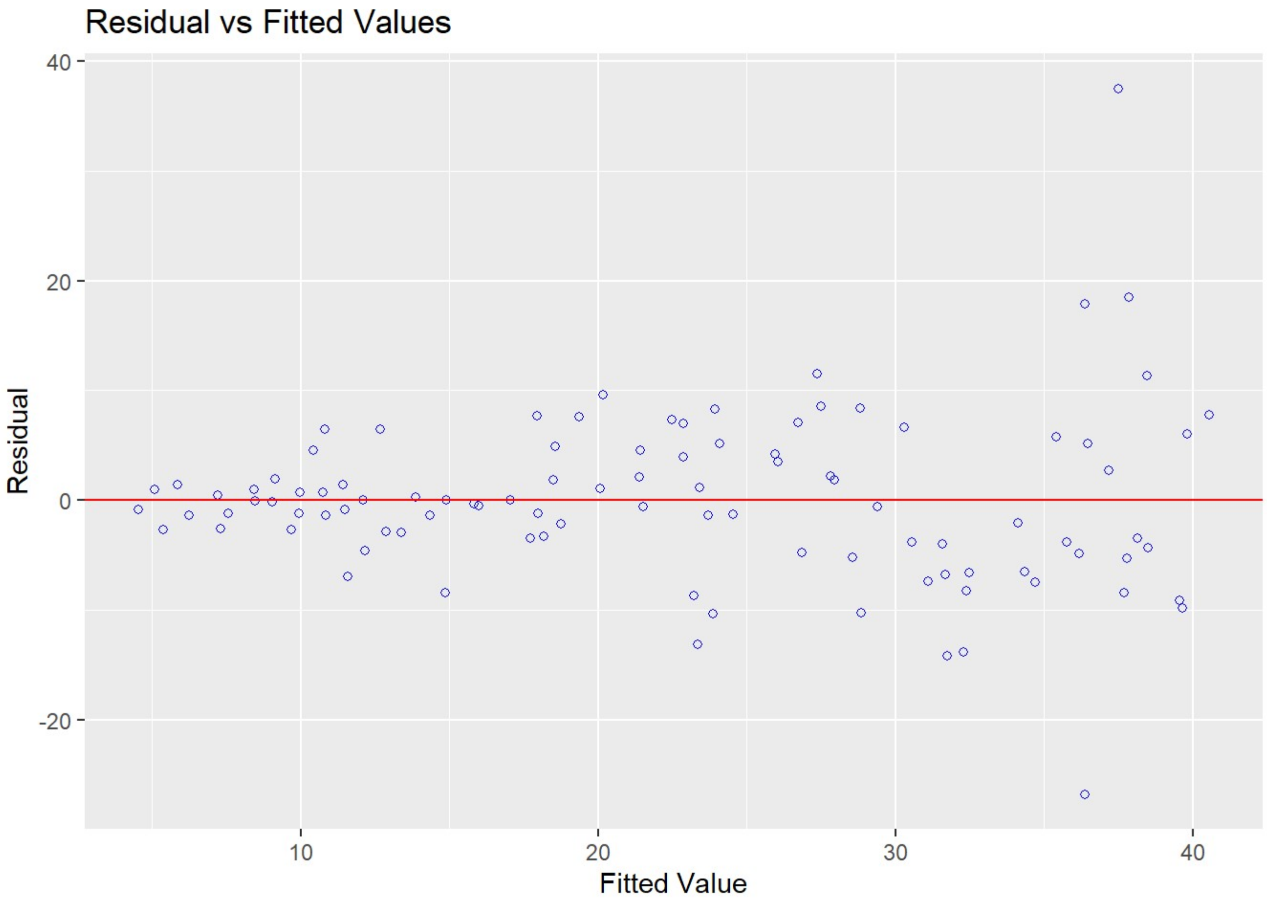
```
#Graph for detecting violation of normality assumption.
ols_plot_resid_qq(Linear_Model)
```



```
#Correlation between observed residuals and expected residuals under normality.
ols_test_correlation(Linear_Model)
```

```
## [1] 0.9512519
```

```
#Residual vs Fitted Values Plot
ols_plot_resid_fit(Linear_Model)
```



- The residuals are spread randomly around 0, which indicates the Linear relationship with a fair homogeneity in error variance.
- Negligible residual is visibly away from the random pattern of the residuals indicating that there are no outliers.

Thus, the linear model is appropriate for determining the relationship between Y and X.

Question 2:

We will use the ‘mtcars’ dataset for this question. The dataset is already included in your R distribution. The dataset shows some of the characteristics of different cars. The following shows few samples (i.e. the first 6 rows) of the dataset. The description of the dataset can be found 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
```

```
Linear_Model_mtcars <- lm(hp~mpg+wt, data = mtcars)
summary(Linear_Model_mtcars)
```

```
## 
## Call:
## lm(formula = hp ~ mpg + wt, data = mtcars)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -59.42 -30.75 -12.07  24.82 141.84 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)   
## (Intercept)  349.287    103.509   3.374  0.00212 **
## mpg           -9.417      2.676  -3.519  0.00145 **
## wt            -4.168     16.485  -0.253  0.80217   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 44.65 on 29 degrees of freedom
## Multiple R-squared:  0.6033, Adjusted R-squared:  0.576 
## F-statistic: 22.05 on 2 and 29 DF,  p-value: 1.505e-06
```

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.

- Solution:

- According to the hypothesis testing from the above linear model we can see that the probability of weight (wt) being 0 is high i.e. 80% whereas, the same for Fuel Consumption being 0 is almost 0%. Hence we can say that the Fuel Consumption is statistically significant to determine the Horse power of the Car.
- Thus, Chris is correct stating that the fuel consumption expressed in Mile Per Gallon (mpg), is a better estimator of the Horse Power(hp).

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).

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

```
##
## 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
```

I. Using this model, what is the estimated Horse Power of a car with 4 calendar and mpg of 22?

```
Predict_Value_1 <- predict(LM_mtcars, data.frame(cyl = 4, mpg = 22), interval = "predi
ction")
Predict_Value_1
```

```
##        fit      lwr      upr
## 1 88.93618 5.398783 172.4736
```

OR

- Equation: hp = 54.067+ 23.979 * cyl - 2.775 * mpg
- hp = 54.067+ 23.979 * 4 - 2.775 * 22 = 88.93

II. Construct an 85% confidence interval of your answer in the above question. Hint: use the predict function.

```
Predict_Value_2 <- predict(LM_mtcars, data.frame(cyl = 4, mpg = 22), interval = "confi
dence", level = 0.85)
Predict_Value_2
```

```
##        fit      lwr      upr
## 1 88.93618 67.64473 110.2276
```

Question 3:

For this question, we are going to use BostonHousing dataset. The dataset is in ‘mlbench’ package, so we first need to instal the package, call the library and the load the dataset using the following commands and You should have a dataframe with the name of BostonHousing in your Global environment now.

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

The dataset contains information about houses in different parts of Boston. Details of the dataset is explained here. Note the dataset is old, hence low house prices!

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)

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

```
##
## 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 R2 value i.e. Co-efficient of Determination is very low 35.99 %. Hence the model is not an accurate one.

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?

- Solution:

- Chas1 is the probability of having a Chase River (Yes) over not having it (No).
- The house bounded by the chase river is more expensive by the factor of 4.58393.

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)

- Solution:

- As the ptratio decrease by a unit there is a 1.49367 times increase in medv.

- Formula: Estimated Coefficient of ptratio * ptratio

- Considering all the other variables to be 1

- For ptratio 15:
- Housing_Price _15 = 49.91868 + (-0.26018) + 0.07073 + (-1.49367)*15 + 4.58393 = 31.90811

- For ptratio 15:
- Housing_Price _18 = 49.91868 + (-0.26018) + 0.07073 + (-1.49367)*18 + 4.58393 = 27.4271

- Cost_Difference = 31.90811 - 27.4271 = 4.48101
- The house having the pupil-teacher ratio 15 is more expensive by the factor of 4.48101 then that of the house having the pupil-teacher 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.

- Solution:
- All the variables are statistically important in determining the house price as per the summary probability of the Variables in the built Linear model.

d. Use the anova analysis and determine the order of importance of these four variables.

```
anova(LM_BostonHousing)
```

```
## 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 order of importance of varibale in determining the Housing price is:

1. Crime Rate (crim)
2. Pupil-teacher ratio (ptratio)
3. Proportion of residential land zoned for lots over 25,000 sq.ft (zn)
4. Weather the whether the tract bounds Chas River(chas).