BA64036_Assignment3

Ruthvick Bulagakula

2023-10-25

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

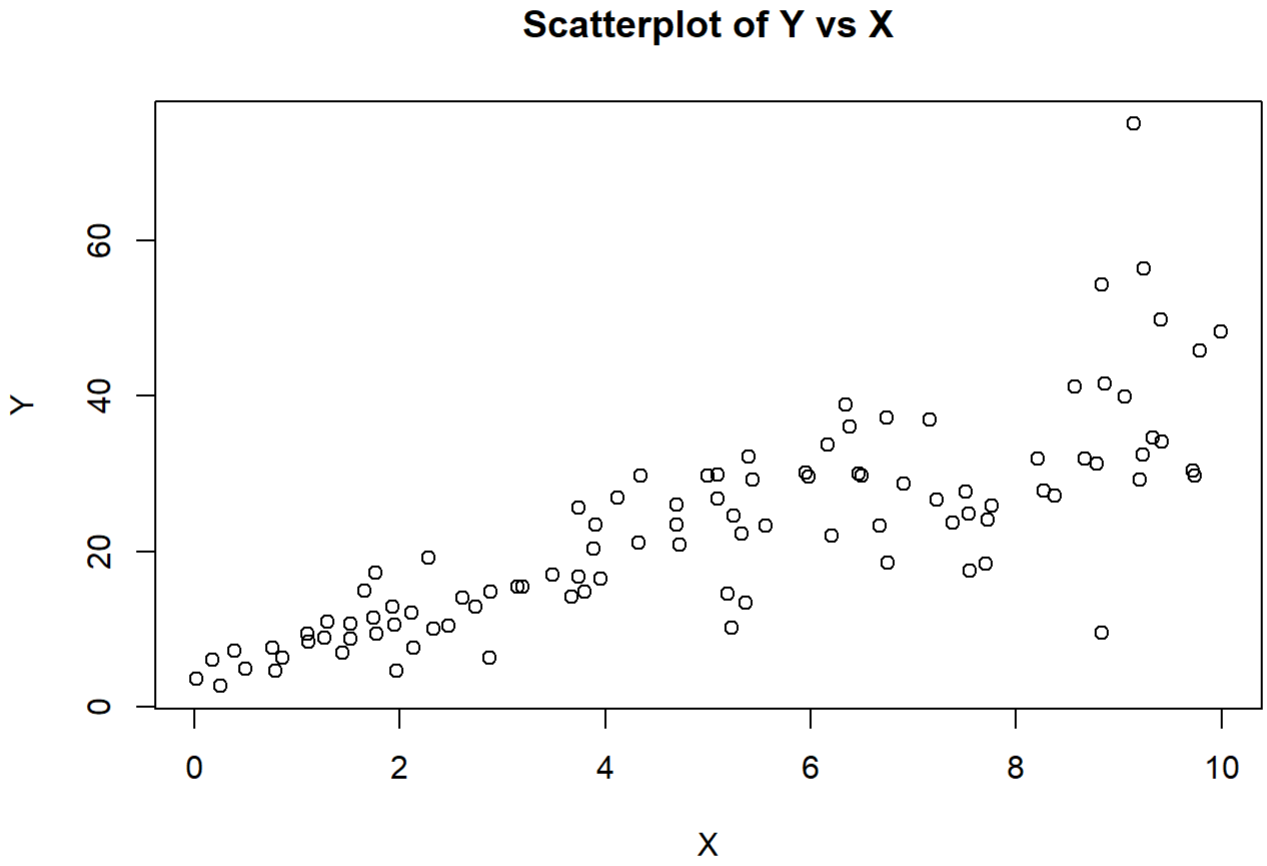
Y=rnorm(100)*0.29*Y+Y

```
set.seed(2017)

X = runif(100) * 10

Y = X * 4 + 3.45

Y = rnorm(100) * 0.29 * Y + Y
```

1a)

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? (8% of total points)

```
plot(X, Y, main="Scatterplot of Y vs X", xlab="X", ylab="Y")
```



Based on above plot i think linear model is applicable to explain Y based on X. Because, this plot shows linear pattern between X and Y. When X increases, Y also increases.

1b)

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? (8% of total points)

```
lm_model = lm(Y ~ X)

summary(lm_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
```

```
coefficients(lm_model)
```

```
## (Intercept)           X 
##    4.465490    3.610759
```

4.4655 is the estimated intercept 3.6108 is the estimated coefficient for X

Accuracy:

The residual standard error (7.756) which distance between observed Y and predicted Y values.

The R-squared value is 0.6517 which indicates that approximately 65.17% of the variance in Y can be explained by the linear relationship with X

1c)

How the Coefficient of Determination, R 2, of the model above is related to the correlation coefficient of X and Y? (8% of total points)

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

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

```
summary(lm_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
```

As you can see R^2 = 0.6517 and r = 0.807291. In linear regression model R^2 should equal to r^2. So, let’s check R^2 = r^2 for the above model.

r^2 = (0.807291)^2 = 0.6571

If you look at r^2 is equal to R^2. Based on above problem square of correlation and r square is equal.

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.

```
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
```

2a)

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. (17% of total points)

```
model_wt = lm(hp ~ wt, data = mtcars)
model_mpg = lm(hp ~ mpg, data = mtcars)

summary(model_wt)
```

```
##
## 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
```

```
summary(model_mpg)
```

```
##
## 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
```

By looking at hp ~ wt LM model summary, we can tell p-value for wt has three ’*’ which tells it’s highly related with hp and for R-squared it’s 0.4339 which means 43.39% variance in hp can be explained by wt.

By looking at hp ~ mpg LM model summary, we can tell p-value for mpg has three ’*’ which tells it’s highly related with hp and for R-squared it’s 0.6024 which means 60.24% variance in hp can be explained by mpg.

By comparing both analysis, we can tell hp ~ mpg is more significant than hp ~ wt. So, i would support chris.

2b)

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). Using this model, what is the estimated Horse Power of a car with 4 calendar and mpg of 22? (17% of total points)

```
model = lm(hp ~ cyl + mpg, data = mtcars)

intercept = coef(model)["(Intercept)"]

coef_cyl = coef(model)["cyl"]

coef_mpg = coef(model)["mpg"]

new_cyl = 4

new_mpg = 22

estimated_hp = intercept + coef_cyl * new_cyl + coef_mpg * new_mpg

cat("Estimated horse power for 4 cylinders and 22 MPG is", estimated_hp)
```

```
## Estimated horse power for 4 cylinders and 22 MPG is 88.93618
```

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

install.packages(‘mlbench’)

library(mlbench)

data(BostonHousing)

```
library(mlbench)

data(BostonHousing)

BostonHousing$chas = factor(BostonHousing$chas, levels = c(0, 1))
```

3a)

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 R 2) (8% of total points)

```
model_medv = lm(medv ~ crim + zn + ptratio + chas, data = BostonHousing)

summary(model_medv)
```

```
##
## 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
```

By looking at the model summary, we can tell R-Square value is 0.3599 which is 35.99% of variance in medv explained by crim, zn, ptratio, chas1.

By looking at the model summary, we can tell p-values for all dependent variable is highly significant.

By comparing both analysis, variance in medv by all variables is slighty lower and p-values tells all variable highly significant. By looking at p-values is accurate enough, but r-squared perspective is not highly accurate enough.

3b1)

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? (8% of total points)

```
model = lm(medv ~ crim + zn + ptratio + chas, data = BostonHousing)

house1 = data.frame(crim = 0.5, zn = 0, ptratio = 20, chas = factor(0, levels = c(0, 1)))
house2 = data.frame(crim = 0.5, zn = 0, ptratio = 20, chas = factor(1, levels = c(0, 1)))

coefficients_model = coef(model)

medv_house1 = coefficients_model[1] + coefficients_model[2] * house1$crim + coefficients_model
[3] * house1$zn + coefficients_model[4] * house1$ptratio + coefficients_model[5] * as.numeric(ho
use1$chas)

cat("House one price is",medv_house1*1000,"dollars")
```

```
## House one price is 24499.07 dollars
```

```
medv_house2 = coefficients_model[1] + coefficients_model[2] * house2$crim + coefficients_model
[3] * house2$zn + coefficients_model[4] * house2$ptratio + coefficients_model[5] * as.numeric(ho
use2$chas)

cat("\nHouse two price is",medv_house2*1000,"dollars")
```

```
##
## House two price is 29083 dollars
```

```
price_difference_chas = medv_house2 - medv_house1

cat("\nPrice difference between house one and house two is",price_difference_chas*1000, "Dollar
s")
```

```
##
## Price difference between house one and house two is 4583.926 Dollars
```

```
model = lm(medv ~ chas, data = BostonHousing)

house1 = data.frame(chas = factor(0, levels = c(0, 1)))
house2 = data.frame(chas = factor(1, levels = c(0, 1)))

coefficients_model = coef(model)

medv_house1 = coefficients_model[1] + coefficients_model[2] * as.numeric(house1$chas)

cat("House one price is",medv_house1*1000,"dollars")
```

```
## House one price is 28440 dollars
```

```
medv_house2 = coefficients_model[1] + coefficients_model[2] * as.numeric(house2$chas)

cat("\nHouse two price is",medv_house2*1000,"dollars")
```

```
##
## House two price is 34786.16 dollars
```

```
price_difference_chas = medv_house2 - medv_house1

cat("\nPrice difference between house one and house 2 is",price_difference_chas*1000, "Dollars")
```

```
##
## Price difference between house one and house 2 is 6346.157 Dollars
```

By looking at the first model, where medv is dependent on crim, zn, ptratio, and chas. We get House two is more expensive by $4583.926.

By looking at the second model. where medv is dependent on chas. We get House two is more expensive by 6346.157.

By looking at both analysis, we can tell Price of the house will get increased based on how many dependent variables we are cinsidering.

3b2)

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 moren expensive and by how much? (Golden Question: 4% extra)

```
model = lm(medv ~ ptratio, data = BostonHousing)

house1 = data.frame(ptratio = 15)
house2 = data.frame(ptratio = 18)

coefficients_model = coef(model)

medv_house1 = coefficients_model[1] + coefficients_model[2] * house1$ptratio

cat("House one price is",medv_house1*1000,"dollars")
```

```
## House one price is 29987 dollars
```

```
medv_house2 = coefficients_model[1] + coefficients_model[2] * house2$ptratio

cat("\nHouse two price is",medv_house2*1000,"dollars")
```

```
##
## House two price is 23515.47 dollars
```

```
price_difference_chas = medv_house1 - medv_house2

cat("\nPrice difference between house one and house two is",price_difference_chas*1000, "Dollar
s")
```

```
##
## Price difference between house one and house two is 6471.526 Dollars
```

By looking at the above model, we can tell House one is more expensive by $6471.526 if we consider ptratio 15 and 18 for two houses.

```
model = lm(medv ~ crim + zn + ptratio + chas, data = BostonHousing)

house1 = data.frame(crim = 0.5, zn = 0, ptratio = 15, chas = factor(0, levels = c(0, 1)))
house2 = data.frame(crim = 0.5, zn = 0, ptratio = 18, chas = factor(0, levels = c(0, 1)))

coefficients_model = coef(model)

medv_house1 = coefficients_model[1] + coefficients_model[2] * house1$crim + coefficients_model
[3] * house1$zn + coefficients_model[4] * house1$ptratio + coefficients_model[5] * as.numeric(ho
use1$chas)

cat("House one price is",medv_house1*1000,"dollars")
```

```
## House one price is 31967.43 dollars
```

```
medv_house2 = coefficients_model[1] + coefficients_model[2] * house2$crim + coefficients_model
[3] * house2$zn + coefficients_model[4] * house2$ptratio + coefficients_model[5] * as.numeric(ho
use2$chas)

cat("\nHouse two price is",medv_house2*1000,"dollars")
```

```
##
## House two price is 27486.42 dollars
```

```
price_difference_chas = medv_house1 - medv_house2

cat("\nPrice difference between house one and house two is",price_difference_chas*1000, "Dollar
s")
```

```
##
## Price difference between house one and house two is 4481.018 Dollars
```

By looking at the above first model, we can tell House one(ptratio 15) is more expensive by $6471.526 if we consider ptratio 15 and 18 for two houses.

By looking at the second model, we can tell House one(ptratio 15) is more expensive by $4481.018 if we consider ptratio 15 and 18 and other variables for two houses.

3c)

Which of the variables are statistically important (i.e. related to the house price)? Hint: use the p-values of the coefficients to answer. (8% of total points)

```
summary(model_medv)
```

```
##
## 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
```

By looking at above model summary, for crim p-value is 2.20e-10 which is extremely low, for zn p-value is 6.14e-06 which is very low, for ptratio p-value is < 2e-16 which is extremely low, for chas1 p-value is 0.000514 which is low. By looking all variables are stastisically important since changing or removing value can make huge difference as we seen in above House expensive question.

3d)

Use the anova analysis and determine the order of importance of these four variables. (18% of total points)

To determine the order of importance of the four variables, you can use an ANOVA analysis:

```
anova(model_medv)
```

```
## 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
```

By looking at above anova analysis, for crim F-value is 118.07 and p-value is < 2.2e-16 which is highly significant, for zn F-value is 65.122 and p-value is 5.253e-15 which is highly significant, for ptratio F-value is 86.287 and p-value is < 2.2e-16 which is highly significant, for chas F-value is 12.224 and p-value is 0.0005137 is significant.

Order of significance:

crim

ptratio

zn

chas