

Assignment 2 Business Analytics

1. Run the following code in R-studio to create two variables X and Y.

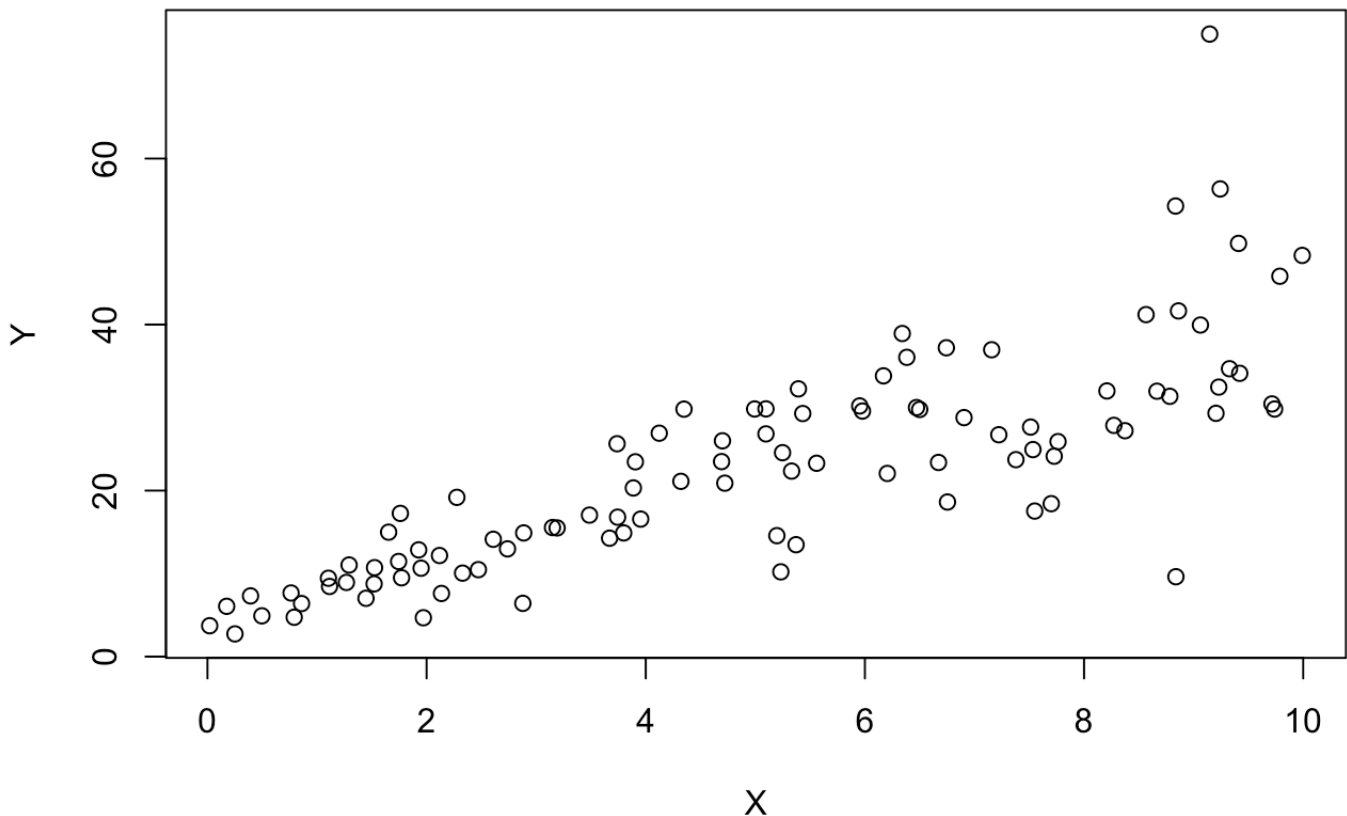
```
set.seed(2017)
X=runif(100)*10

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)
```



- 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?

```
M_XY=lm(Y~X)
```

```
summary(M_XY)
```

```
##
## 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
```

```
M_XY$coefficients
```

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

The equation that explains Y in terms of X is

$Y = (3.610759 \cdot X) + 4.465490$

The model's accuracy is 0.6517, which is quite good.

- c. How the Coefficient of Determination, R^2 , of the model above is related to the correlation coefficient of X and Y? # Because the regression is only dependent on one variable, R^2 is just the squared value of the correlation coefficient between x and y.

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

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

we can see that the R-Squared value of 0.6517 is the same as before.

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.

- a. James wants to buy a car. He and his friend, Chris, have different opinions about the Horse Power (hp) of cars. James thinks 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
```

```
J_model<-lm(mtcars$hp~mtcars$mpg,data=mtcars)
```

```
summary(J_model)
```

```
##
## Call:
## lm(formula = mtcars$hp ~ mtcars$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 ***
## mtcars$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
```

```
Ch_model<-lm(mtcars$hp~mtcars$mpg,data=mtcars)
summary(Ch_model)
```

```
##
## Call:
## lm(formula = mtcars$hp ~ mtcars$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 ***
## mtcars$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
```

chris is correct based on the linear model of

the mtcars dataset.

The accuracy of chris' model is 0.6024, which is significantly higher than j(James) is (0.4339)

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

```
M_HP<-lm(mtcars$hp~mtcars$mpg+mtcars$cyl,data=mtcars)
```

```
summary(M_HP)
```

```
##
## Call:
## lm(formula = mtcars$hp ~ mtcars$mpg + mtcars$cyl, data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -53.72 -22.18 -10.13  14.47 130.73
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.067      86.093   0.628  0.53492
## mtcars$mpg    -2.775       2.177  -1.275  0.21253
## mtcars$cyl    23.979       7.346   3.264  0.00281 **
## ---
## 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
```

```
M_HP$coefficients
```

```
## (Intercept) mtcars$mpg mtcars$cyl
##    54.066600  -2.774769  23.978626
```

```
predict_hp<-(M_HP$coefficients[2]*22)+(M_HP$coefficients[3]*4)+M_HP$coefficients[1]

print(paste('The predicted Horse power of a car with FOUR calenders and a 22-mpg fuel
Economy is ',predict_hp))
```

```
## [1] "The predicted Horse power of a car with FOUR calenders and a 22-mpg fuel Econ
omy is 88.9361796789223"
```

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)` You should have a dataframe with the name of BostonHousing in your Global environment now.
 - 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?

```
library(mlbench)

data(BostonHousing)

M_Boston<-lm(formula = BostonHousing$medv~BostonHousing$crim+BostonHousing$zn+BostonH
ousing$ptratio+BostonHousing$chas,data=BostonHousing)

summary(M_Boston)
```

```
##
## Call:
## lm(formula = BostonHousing$medv ~ BostonHousing$crim + BostonHousing$zn +
##      BostonHousing$lstat + BostonHousing$chas, data = BostonHousing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.282  -4.505  -0.986   2.650  32.656
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    49.91868     3.23497   15.431 < 2e-16 ***
## BostonHousing$crim    -0.26018     0.04015   -6.480 2.20e-10 ***
## BostonHousing$zn      0.07073     0.01548    4.570 6.14e-06 ***
## BostonHousing$lstat  -1.49367     0.17144   -8.712 < 2e-16 ***
## BostonHousing$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
```

This model is not very precise because the r square value is 0.3599 which is extremely Low.

b. Use the estimated coefficient to answer these questions?

1. 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?

```
M_Boston1<-lm(formula=BostonHousing$medv~BostonHousing$chas, data=BostonHousing)
```

```
M_Boston1
```

```
##
## Call:
## lm(formula = BostonHousing$medv ~ BostonHousing$chas, data = BostonHousing)
##
## Coefficients:
##      (Intercept)  BostonHousing$chas1
##           22.094             6.346
```

```
# we can calculate the values of both houses by utilizing the coeff. from the given model.
```

```
# we are consider Two houses as H0 ,H1
# H 0 has no chas while H 1 has chas.
```

```
H0<-M_Boston1$coefficients[1]+M_Boston1$coefficients[2]*0
H1<-M_Boston1$coefficients[1]+M_Boston1$coefficients[2]*1
print(paste('The House with the chas and is more expensive,',H1-H0))
```

```
## [1] "The House with the chas and is more expensive, 6.34615711252662"
```

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?

```
M_Boston2<-lm(formula=BostonHousing$medv~BostonHousing$ptratio,data=BostonHousing)
```

```
M_Boston2
```

```
##
## Call:
## lm(formula = BostonHousing$medv ~ BostonHousing$ptratio, data = BostonHousing)
##
## Coefficients:
##          (Intercept)  BostonHousing$ptratio
##          62.345          -2.157
```

```
# Because of the coeff.is negative ,we may deduce that increased ptratio will result in lower pricing.
# we may calculate the values of both houses using the coeff. of the given model.
# H 0 has no chas while H1 has chas.
```

```
H3<-M_Boston2$coefficients[1]+M_Boston2$coefficients[2]*15
H4<-M_Boston2$coefficients[1]+M_Boston2$coefficients[2]*18
print(paste('Houses with a pupil-teacher ratio of 15, are more expensive and the most expensive was ',H3-H4 ))
```



```
## [1] "Houses with a pupil-teacher ratio of 15, are more expensive and the most expensive was 6.47152588818295"
```

- 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

```
summary(M_Boston)
```

```
##
## Call:
## lm(formula = BostonHousing$medv ~ BostonHousing$scrim + BostonHousing$zn +
##      BostonHousing$ptratio + BostonHousing$chas, data = BostonHousing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.282  -4.505  -0.986   2.650  32.656
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    49.91868    3.23497   15.431 < 2e-16 ***
## BostonHousing$scrim  -0.26018    0.04015   -6.480 2.20e-10 ***
## BostonHousing$zn      0.07073    0.01548    4.570 6.14e-06 ***
## BostonHousing$ptratio -1.49367    0.17144   -8.712 < 2e-16 ***
## BostonHousing$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
```

we may conclude that none of the variables are statistically significant because the p-value for the model is 0.05

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

```
anova(M_Boston)
```

```
## Analysis of Variance Table
##
## Response: BostonHousing$medv
##              Df Sum Sq Mean Sq F value    Pr(>F)
## BostonHousing$crim      1  6440.8   6440.8  118.007 < 2.2e-16 ***
## BostonHousing$zn        1  3554.3   3554.3   65.122 5.253e-15 ***
## BostonHousing$ptratio    1  4709.5   4709.5   86.287 < 2.2e-16 ***
## BostonHousing$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
```

According to the model, the order is as follows in Desc order.

1.crim

2.ptratio

3.zn

4.chas