# PartB

## Libraries

```
#loading all required libraries here first :
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------ tidyverse 1.2.1 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts --------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

## Reading dataset

```
#importing data and store it to variable "data1"  then vewing structure  of these data
data <- read.csv(file ="C:/Users/Ghars/Documents/data/sheet1_partB.csv", sep = ";")

str(data)
```

```
## 'data.frame':    100 obs. of  3 variables:
##  $ case: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ x   : int  7 12 8 9 7 5 9 5 12 6 ...
##  $ y   : num  17 17.8 19.4 21.1 16.1 ...
```

## B1

IV = ECTS credits in subject X obtained , DV = interest of students in this subject.

Using Pearson Correlation test:

H0 : there is no association between subject ECTS-credits and interest of students in this subject(R= zero)

HA : there is an association between subject ECTS-credits and interest of students in this subject(R not equal to zero)

Significance level or proability of type 1 error = 5 %

For this case we will use T statistics with df = n-2 = 98 we will reject the null hypothesis of the P-value of the T statistics resulted a p-value less than 0.05 using double side test assuming H0 is true .

```
cor(data$x, data$y)
```

```
## [1] 0.7461826
```

```r
cor.test(data$x, data$y, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$x and data$y
## t = 11.096, df = 98, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6441776 0.8221093
## sample estimates:
##       cor
## 0.7461826
```

p-value = 2.2e-16 is significantly smaller than 5% significant level , this provide avery strong evidence against the null hypothesis. There is a strong relation between subjects ECTS and interest of students.

Using Regression and ß1 coefficient :

H0 : there is no association between subject ECTS-credits and interest of students in this subject(ß1= zero)

HA : there is an association between subject ECTS-credits and interest of students in this subject(ß1 not equal to zero)

```r
est <- lm(y~x , data = data)
summary (est)
```

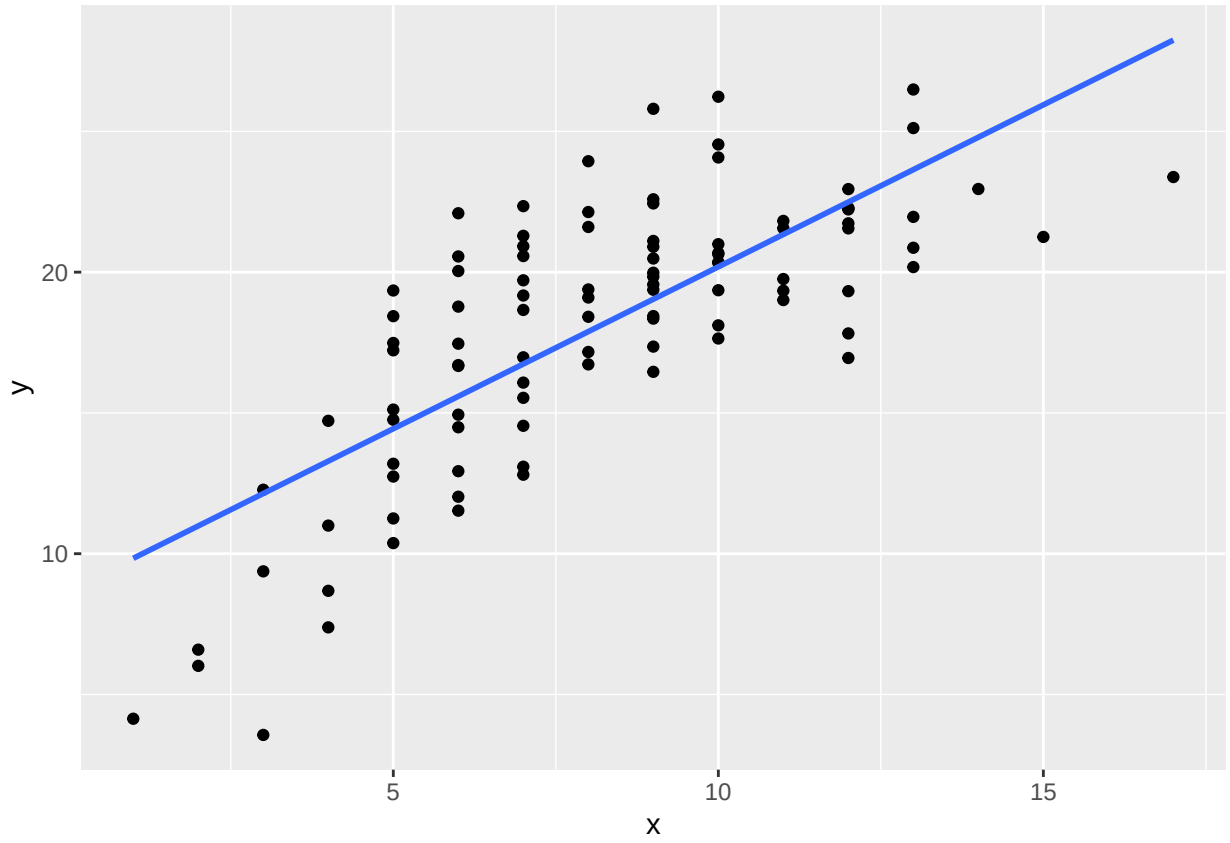
```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5741 -2.2192  0.1546  1.9569  6.7624
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.6867     0.9051   9.598 9.03e-16 ***
## x             1.1503     0.1037  11.096  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.189 on 98 degrees of freedom
## Multiple R-squared:  0.5568, Adjusted R-squared:  0.5523
## F-statistic: 123.1 on 1 and 98 DF,  p-value: < 2.2e-16
```

Same conclusion as before at p-value = 2.2e-16 and there is a strong relation between subjects ECTS and interest of students.

## B2

scatter plot:

```
ggplot(data , aes(x , y)) + geom_point() + geom_smooth(method = "lm" , se=FALSE)
```




The plot shows a postive linear association between response variable which is interest of students and explanatory variable the subject ECTS.This is justified by steep "lm" line in the plot.Therefore it confirms the alternative hypothesis from part B1.

ß1 or effect size of subjects ECTS credits increase by one credit will on average increase the interest of students by 1.1503.

## B3

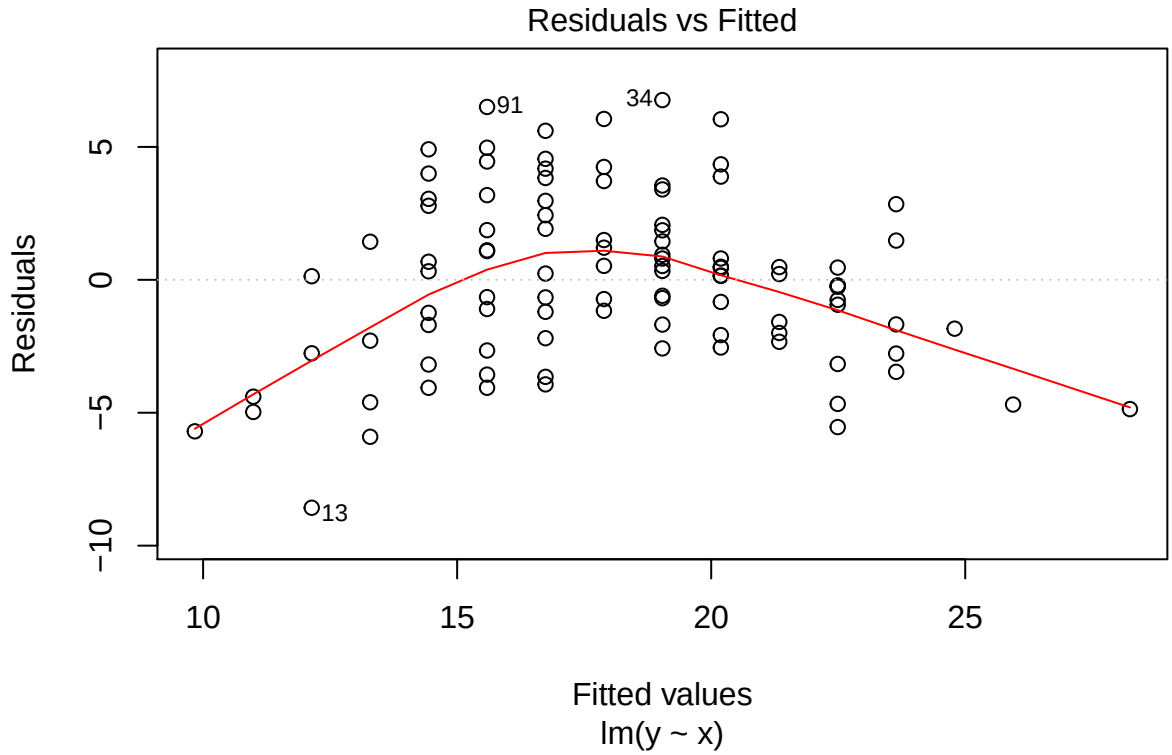
simple linear regression analysis:

```
smodel <- lm(y~x , data = data)
summary(smodel)
```

```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5741 -2.2192  0.1546  1.9569  6.7624
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.6867     0.9051   9.598 9.03e-16 ***
## x             1.1503     0.1037  11.096  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.189 on 98 degrees of freedom
## Multiple R-squared:  0.5568, Adjusted R-squared:  0.5523
## F-statistic: 123.1 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
plot(smodel , which = 1)
```




R-squared= 0.5568 (around 55%) variability of student interest are explained by this model which is plausible fit hence its only one explanatory variable.

ß-coefficient is the same From part B2 , its also known that the null hypothesis is rejected.By looking at F-value=123.1 & DF=98 and respective p-value =2.2e-16 the same as its still only ECTS credits only as explantory variables.

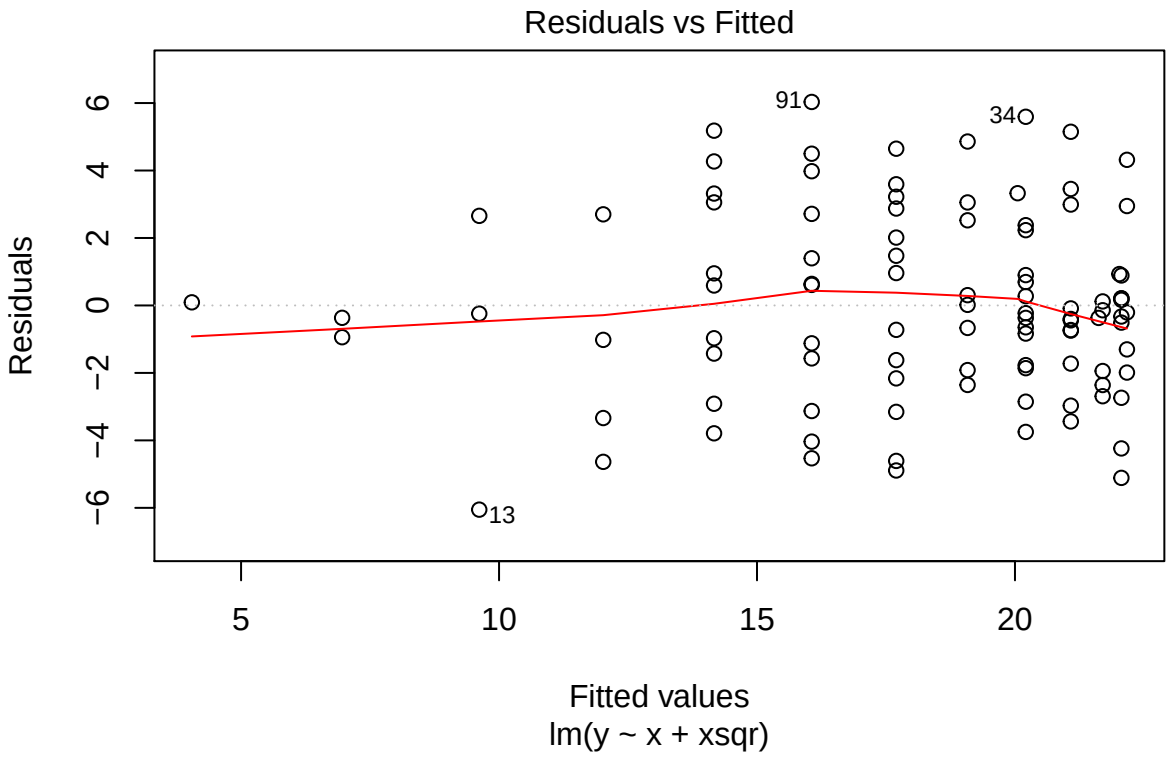
In addition , it seems the linearity of x by looking at the residuals plot doesnt fit here.It follows a parabolic shape.

### Including X and X-squared to the model:

```r
# interest = b0+b1*x +b2*xsqr+e
ndata <- data %>% mutate(xsqr= x^2) # adding the new variable to ndata
nmodel <- lm(y~ x + xsqr , data = ndata)
summary(nmodel)
```

```
##
## Call:
## lm(formula = y ~ x + xsqr, data = ndata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.053 -1.874 -0.236  2.268  6.032
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.87736    1.57119   0.558    0.578
## x            3.29564    0.38443   8.573 1.60e-13 ***
## xsqr        -0.12751    0.02221  -5.740 1.08e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.769 on 97 degrees of freedom
## Multiple R-squared:  0.6692, Adjusted R-squared:  0.6623
## F-statistic:  98.1 on 2 and 97 DF,  p-value: < 2.2e-16
```

```r
plot(nmodel , which = 1)
```




The residual plot shows that the linearity of independent variables enhanced alot with thier relation with response variable the interest of students.

### B4

Comparsion :

p-value didnt change between simple model and the new model = 2.2e-16 , hence the null hypothesis is rejected.

F-value changed from 123.1 to 98.1 in the model

ß1 of subjects ECTS raised from 1.150 to 3.29564 in the new model.Now the average interest of student increase by this value when there is additional ECTS is added to the subject.

ß2 of xsqr has negative coefficient in the new model which altered the parabolic shape of the simple model.It depicts the interest of student decreases on average by 0.12751 when the number of squared ECTS increases by one ECTS.

## Variability :

In the simple model R-squared was 0.5568 (around 55%) , it increased to 0.6692 which approximately 11% more variability explanation in the new model or in another meaning 11% more fit to the new model.

However R-squared is a bias estimator , hence it increases at each additional variable added to the multiple regression model thats why adjusted R-squared is reference for model evaluation.

adjusted R-squared was 0.5523 and becomes now 0.6623 in the new model which approximatly the same like the R-squared or coefficient of determination.Therefore the interpertation is the same.Regarding the fact that adjusted R-squared will always be slightly smaller than R-squared at any model.