PartA

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()
```

```
library(dplyr)
library(broom)
library(reghelper)
```

```
## Warning: package 'reghelper' was built under R version 3.6.2

##
## Attaching package: 'reghelper'

## The following object is masked from 'package:base':
##
##     beta
```

```
library(latexpdf)
library(psych) ### for skewness and kurtosis
```

```
##
## Attaching package: 'psych'

## The following object is masked from 'package:reghelper':
##
##     ICC

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

A1

Reading dataset

```r
#importing data and store it to variable "data1"  then vewing structure  of these data
data <- read.csv(file ="C:/Users/Ghars/Documents/data/sheet1_partA.csv", sep = ";")

str(data)
```

```
## 'data.frame':    66 obs. of  9 variables:
##  $ vpnr       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ age        : int  23 34 26 21 26 24 27 19 31 27 ...
##  $ sex        : int  1 2 2 1 1 1 1 2 1 2 ...
##  $ school     : num  2.1 1.9 2.4 2 2.6 2.8 1.5 3 2.9 2.9 ...
##  $ bachelor   : num  2 2 3 2.8 1.2 2.7 1.5 2.8 3.1 2.6 ...
##  $ abroad     : int  1 12 0 8 1 2 11 7 1 0 ...
##  $ internships: int  2 4 5 2 4 2 7 4 3 2 ...
##  $ interview  : num  3.21 3.86 3.61 2.67 3.85 ...
##  $ performance: num  3.34 4.17 3.84 3.77 3.74 ...
```

Mostly there are some wrong data in the data set , thus Data preperation firstly maintained before the analysis

The vairable "SCHOOL" and "BACHELOR" have missing values , dropping observations is adopted here.

In addition there are some negative valuable of bachelor and school , it doesnt make sense of someone who have -5 of bachelor . Omitting these observation will be better for our analysis

```r
#quantify how many missing values first

data[data<0] = NA  # changing negative values to NA as missing values
colSums(is.na(data))
```

```
##        vpnr         age         sex      school    bachelor      abroad
##           0           0           0           3           4           1
## internships   interview performance
##           1           0           0
```

```r
data1 <- na.omit(data) #no missing values
summary(data1$bachelor)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.500   2.393   2.800   3.700
```

```r
summary(data1$school)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.400   2.000   2.500   2.448   2.900   3.800
```

FOr the gender we will assume male = 1 and female = 2 in the data set.

A2

Descriptive statistics of variables "PERFORMANCE" and "SCHOOL"

Performance

```
round(describeBy(data1$performance),3)
```

```
## Warning in describeBy(data1$performance): no grouping variable requested

##    vars  n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
## X1    1 61 3.98 0.54   4.03       4 0.64 2.67 4.84  2.17 -0.23    -0.83 0.07
```

From the figure we see performance over all for both genders is between min = 2.674 and max= 4.840 , 25%-quantile = 3.555 , 50%-quantile = 4.029 , 75%-quantile = 4.433 , mean = 3.982,median = 4.029 and Standard deviation = 0.54

According to the gender:

```
describeBy(data1$performance , data1$sex)
```

```
##
##  Descriptive statistics by group
## group: 1
##    vars  n mean   sd median trimmed mad  min  max range  skew kurtosis  se
## X1    1 30 3.92 0.57   3.93    3.95 0.7 2.67 4.82  2.15 -0.35    -0.83 0.1
## ------------------------------------------------------------
## group: 2
##    vars  n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
## X1    1 31 4.04 0.52    4.1    4.04 0.68 3.19 4.84  1.65 -0.02    -1.26 0.09
```

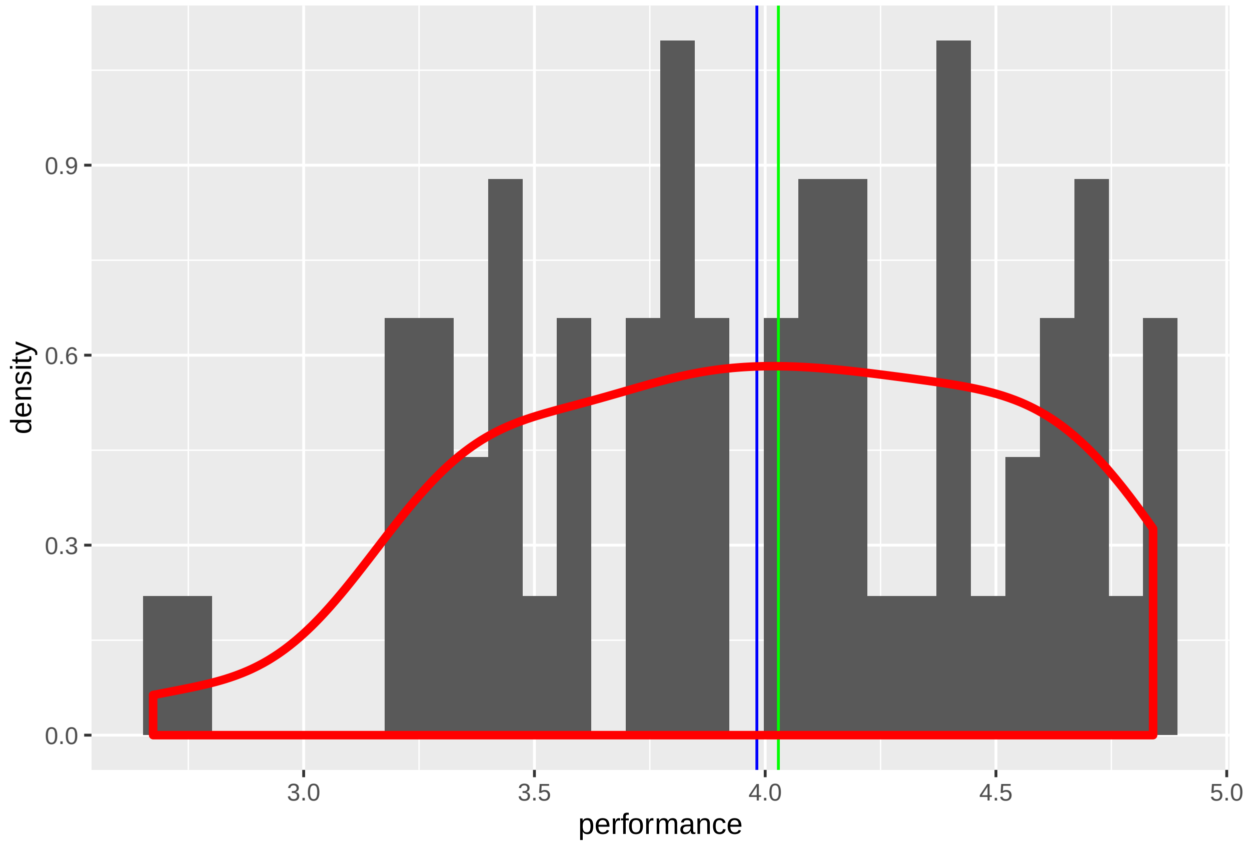
Females scored average performance better (mean= 4.04) than the males (mean =3.92), but the variabilty of score for males is higher (sd = 0.57 & Performance of [2.67,4.82]) , while females achieved (sd = 0.52 & Performance of [3.19,4.84])

Regarding "Performance" we cannot conclude preceisly from standard deviation only and above descriptive statistics that the distribution is normally distributed and shaped like a bell curve .

Graphing the variables will give us more sight on the distribution :

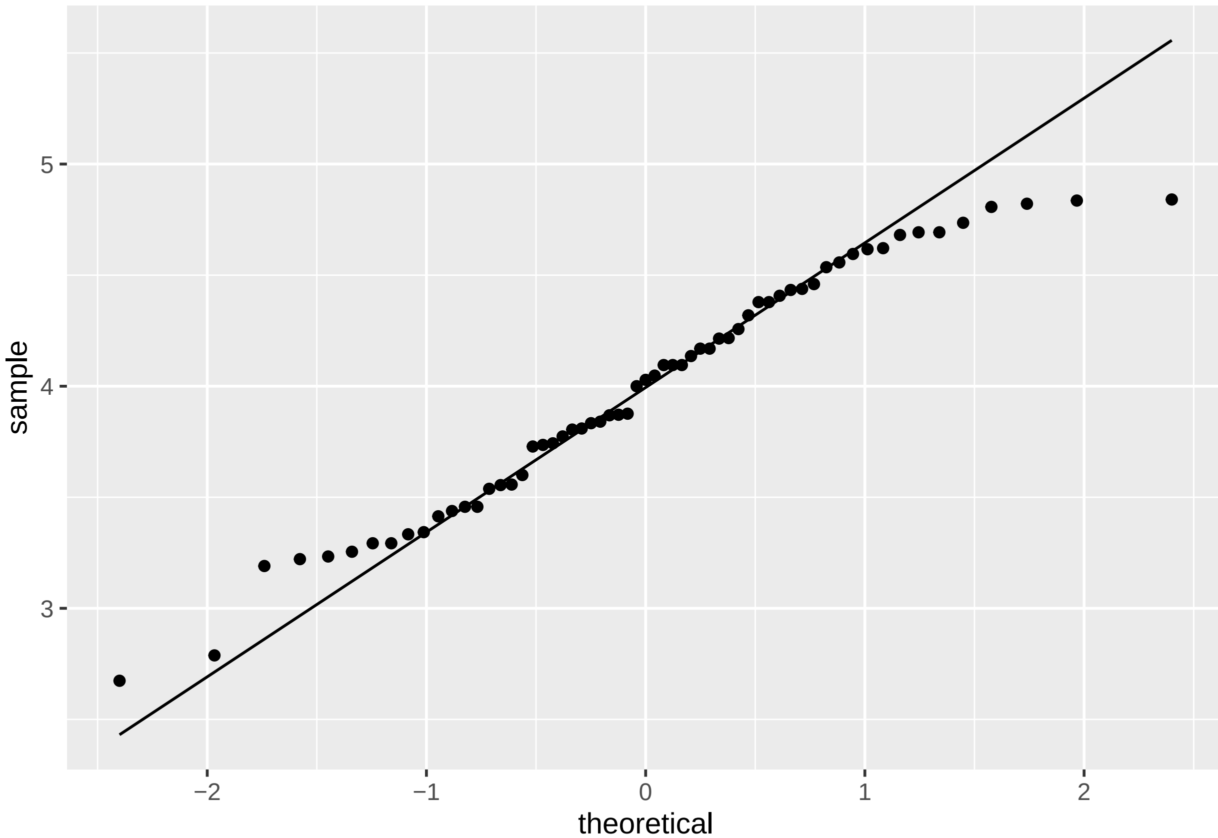
A histogram of Performance :

```
ggplot (data1, aes (x = performance, y = ..density..)) +
geom_histogram(bins = 30) +
geom_vline(xintercept=mean(data1$performance), color="blue") + geom_vline(xintercept=median(data1$perfo
```

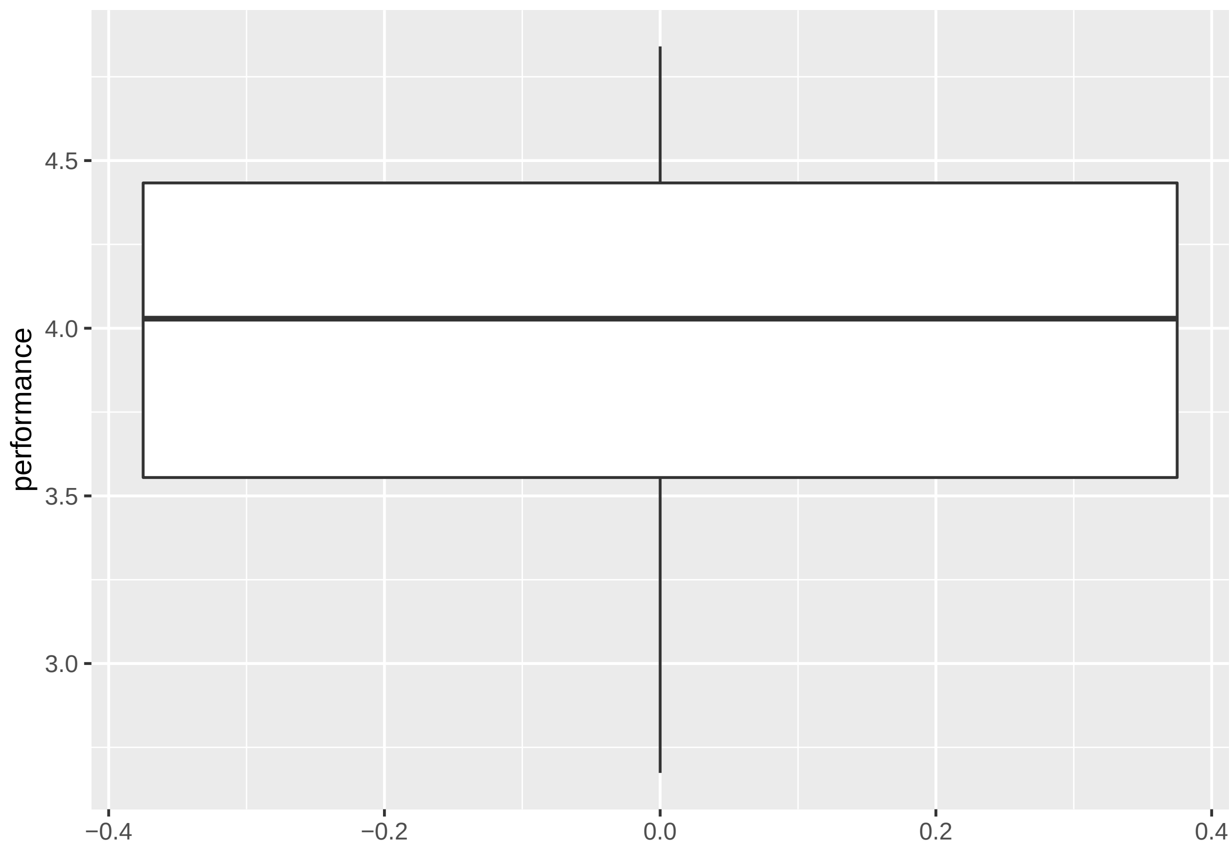



```r
hist(data1$performance)
abline(v = mean(data1$performance), col = "blue", lwd = 3)
```


```r
ggplot(data1 , aes(sample = performance)) + stat_qq() + stat_qq_line()
```




```
ggplot(data1 , aes(y=performance)) + geom_boxplot()
```



Here median = 4.029 ~ mean = 3.982 (median slightly larger than the mean) this implies a little left skewness . This also confirmed by the histogram . Also the box plot shows slight left skewness.

To confirm our conclusion we will see skew value using describeBy from psych library.

As concluded skew is negative = -0.23 refers to skewness towards the left due to existance of some outliers at the left tail

Regarding kurtosis=-0.83 ,it apepars the data are more located around the mean or there is a mass density under the mean and median as there are fewer outliers in the left tail of distribution.

A negative kurtosis refers also to a flatter curve than a perfect normal distribution as shown from the density function at the histogram , because kurtosis is a measurement of taildness in the curve and we cannot infer only from its value that the curve has flatter distribution .

Moreover from the QQ-plot its seen that the distribution has light tails.

An apropriate normal distribution has the 3 sigma rule (68% , 95% and 99.7% lies within 3 standard deviations). By looking at the histogram and distribution this rule doesnt apply.

Applying the rule in R :

```r
3.98 + 1*0.54
```

```
## [1] 4.52
```

The quantile after 1 sd = 4.52 hence the max value 4.84 , this means the assumptions failed and the performance is flat in the middle with light tails

An apropriate measure is the T statistics with respect to standard error (sd/n) and df , since we dont know the population standard deviation of all candidates also.

School

```
round(describeBy(data1$school),3)
```

```
## Warning in describeBy(data1$school): no grouping variable requested
```

```
##    vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 61 2.45 0.53    2.5    2.46 0.59 1.4 3.8   2.4 -0.08    -0.64 0.07
```

From the figure we see School grade lies between min = 1.4 and max= 3.8 , 25%-quantile = 2 , 50%-quantile = 2.5 , 75%-quantile = 2.9 , mean = 2.45 and median = 2.5 and Standard deviation = 0.53.

According to the gender:

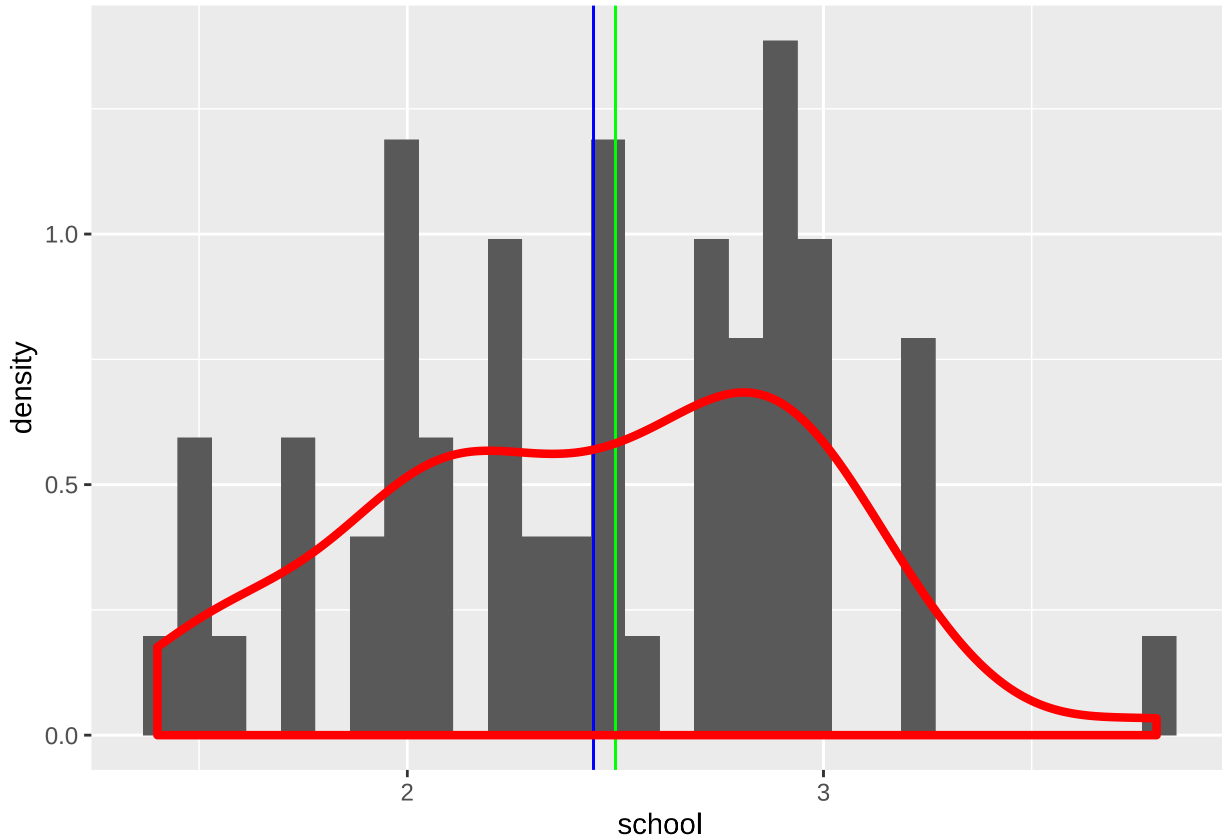
```
describeBy(data1$school , data1$sex)
```

```
##
##  Descriptive statistics by group
## group: 1
##    vars  n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 30 2.32 0.59   2.25    2.31 0.74 1.4 3.8   2.4 0.26     -0.6 0.11
## ------------------------------------------------------------
## group: 2
##    vars  n mean   sd median trimmed  mad min max range  skew kurtosis   se
## X1    1 31 2.57 0.43    2.7    2.58 0.44 1.7 3.2   1.5 -0.21    -1.19 0.08
```

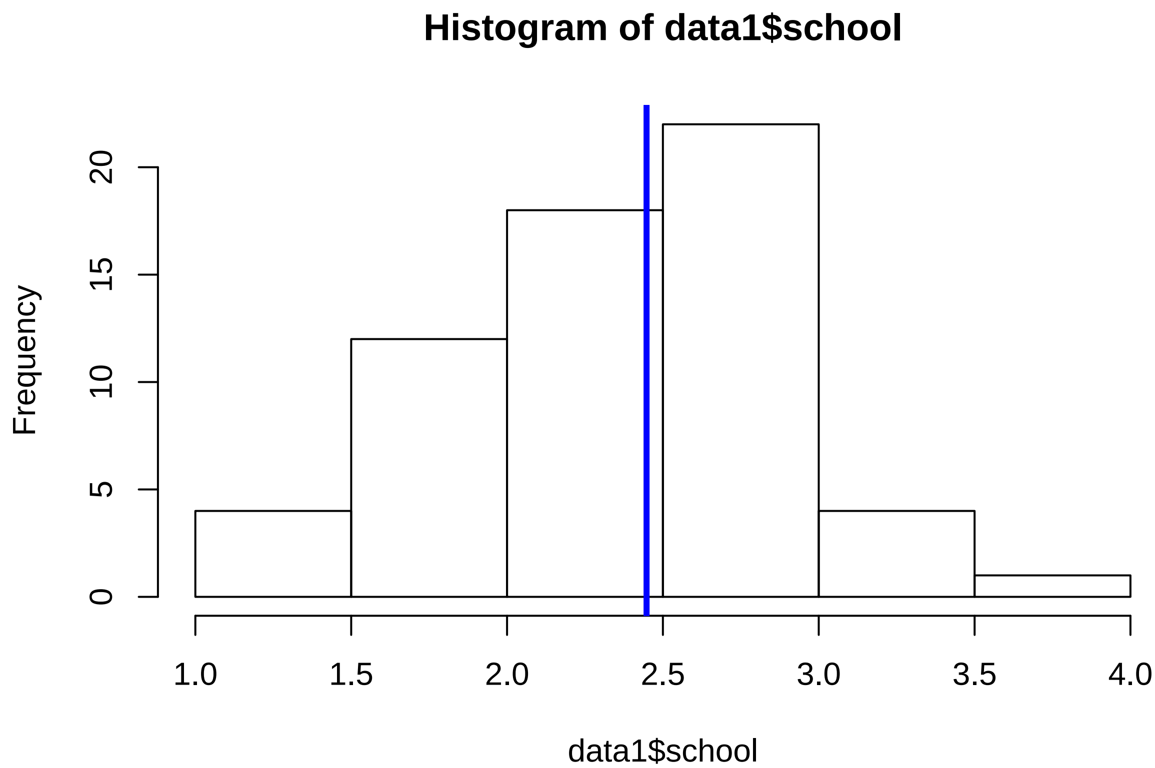
Similarly to Performance females also school grade slightly better (mean= 2.57) than the male (mean =2.32), but the variabilty of score for males is higher (sd = 0.59 & grade of [1.4,3.8]) , while females achieved (sd = 0.43 & grade of [1.7,3.2])

A histogram of School variable :

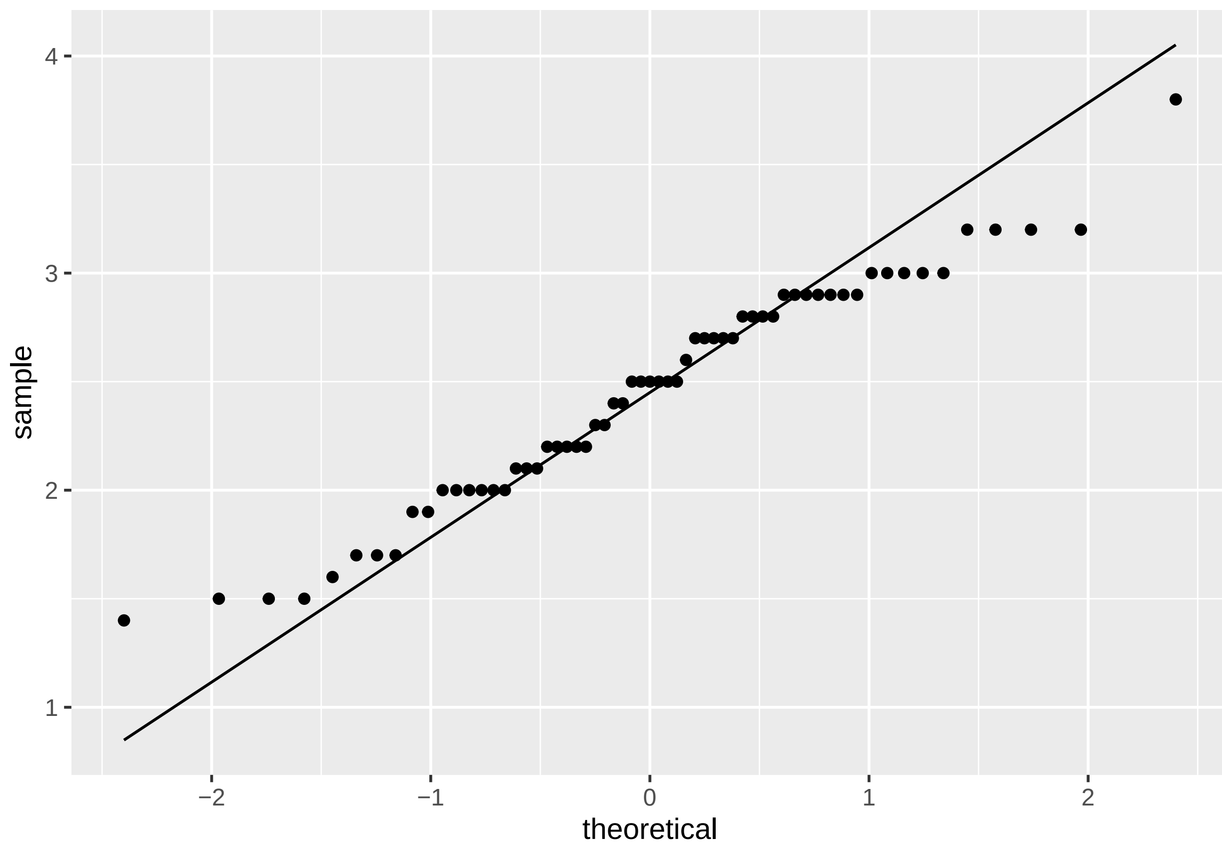
```
ggplot (data1, aes (x = school, y = ..density..)) +
geom_histogram(bins = 30) +
geom_vline(xintercept=mean(data1$school), color="blue") +
geom_vline(xintercept=median(data1$school), color="green")+
geom_density(size = 1.5, color = "red")
```



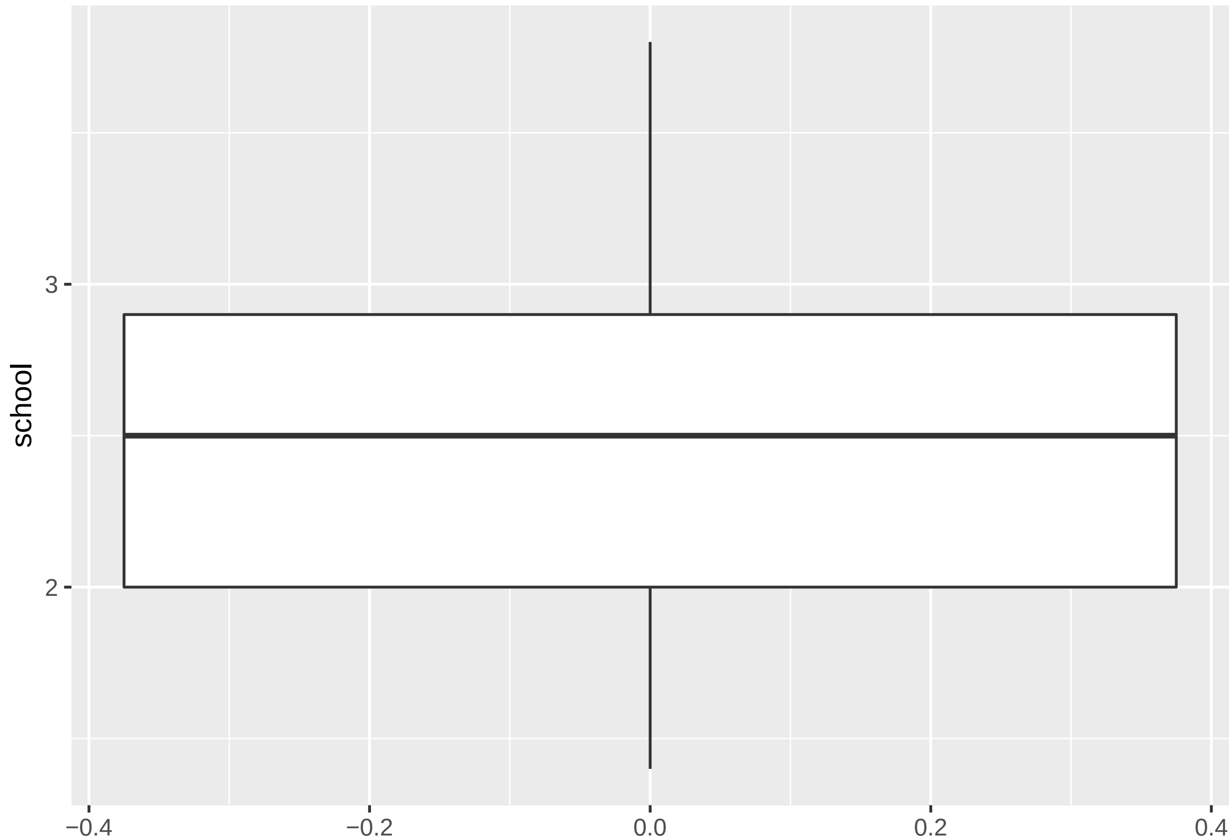

```r
hist(data1$school)
abline(v = mean(data1$school), col = "blue", lwd = 3)
```




```r
ggplot(data1 , aes(sample = school)) + stat_qq() + stat_qq_line()
```




```r
ggplot(data1 , aes(y=school)) + geom_boxplot()
```



It appears School variable data is almost simliar to performance distribution with light tails too. however,its a little more symmetric around the mean with bimodal distribution and little more distributed than performance variable . The skewness is less of -0.08 and also kurtosis comparing with performance

A3

IV = Bachelor grade , DV = Performance

Using Pearson Correlation test :

Correlation value between two variables lies [-1 & 1] ,

H0 : there is no association between bachelor grade and performance (R= zero)

HA : there is an association between bachelor grade and performance (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 = 59 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(data1$bachelor, data1$performance)
```

```
## [1] 0.1583035
```

```
cor.test(data1$bachelor, data1$performance, method="pearson")
```

```
##
```

```
##  Pearson's product-moment correlation
##
## data:  data1$bachelor and data1$performance
## t = 1.2315, df = 59, p-value = 0.223
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.09739997  0.39440180
## sample estimates:
##       cor
## 0.1583035
```

p-value = 0.223 , H0 failed to be rejected and there is no correlation between bachelor grade and performance.

Using Regression and ß1 coefficient :

H0 : there is no association between bachelor grade and performance (ß1= zero)

HA : there is an association between bachelor grade and performance ((ß1 not equal to zero))

```
est <- lm(performance~bachelor , data = data1)
summary (est)
```

```
##
## Call:
## lm(formula = performance ~ bachelor, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.18487 -0.45180  0.02964  0.51150  0.92136
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.6518     0.2769  13.187   <2e-16 ***
## bachelor      0.1379     0.1120   1.231    0.223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5418 on 59 degrees of freedom
## Multiple R-squared:  0.02506,    Adjusted R-squared:  0.008536
## F-statistic: 1.517 on 1 and 59 DF,  p-value: 0.223
```

ß1 = 0.1379 & p-value = 0.223 , H0 failed to be rejected and there is no correlation between bachelor grade and performance at p-value =0.223

Different IV : Performance ~ Age

IV = age, DV = Performance

Similarly as before , Using Pearson Correlation: test :

H0 : there is no association between candidate age and performance (R= zero)

HA : there is an association between candidate age and performance ((R not equal to zero))

```r
cor(data1$age, data1$performance)
```

```
## [1] 0.07954215
```

```r
cor.test(data1$age, data1$performance, method="pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  data1$age and data1$performance
## t = 0.61292, df = 59, p-value = 0.5423
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1757999  0.3248559
## sample estimates:
##        cor
## 0.07954215
```

p-value = 0.5423 , H0 failed to be rejected and there is no correlation between candidate age and performance.

Using Regression and ß1 coefficient :

H0 : there is no association between candidate age and performance (ß1= zero)

HA : there is an association between candidate age and performance ((ß1 not equal to zero))

```r
est0 <- lm(performance~age , data = data1)
summary (est0)
```

```
##
## Call:
## lm(formula = performance ~ age, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.36613 -0.42125  0.03845  0.44413  0.92107
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.59373    0.63722   5.640 5.08e-07 ***
## age          0.01539    0.02510   0.613    0.542
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.547 on 59 degrees of freedom
## Multiple R-squared:  0.006327,   Adjusted R-squared:  -0.01051
## F-statistic: 0.3757 on 1 and 59 DF,  p-value: 0.5423
```

ß1 = 0.01539 & p-value = 0.542

H0 failed to be rejected and there is no correlation between candidate age and performance.

A4

Correlation table :

```
cor_table <- cor(data1)
round(cor_table , 2)
```

```
##              vpnr   age   sex school bachelor abroad internships interview
## vpnr         1.00 -0.25  0.11  -0.01    -0.01   0.06       -0.02     -0.09
## age         -0.25  1.00  0.15   0.02    -0.06  -0.03        0.23      0.05
## sex          0.11  0.15  1.00   0.24     0.28   0.04        0.24     -0.02
## school      -0.01  0.02  0.24   1.00     0.54  -0.38       -0.23     -0.25
## bachelor    -0.01 -0.06  0.28   0.54     1.00  -0.22       -0.07      0.03
## abroad       0.06 -0.03  0.04  -0.38    -0.22   1.00        0.26      0.33
## internships -0.02  0.23  0.24  -0.23    -0.07   0.26        1.00      0.46
## interview   -0.09  0.05 -0.02  -0.25     0.03   0.33        0.46      1.00
## performance -0.01  0.08  0.11  -0.18     0.16   0.29        0.33      0.48
##             performance
## vpnr              -0.01
## age                0.08
## sex                0.11
## school            -0.18
## bachelor           0.16
## abroad             0.29
## internships        0.33
## interview          0.48
## performance        1.00
```

From the table , r = [-1,1] when R is close to 1 this means high positive correlation , while R is close to -1 perfect negative correlation between variables.

Collinearity may affect efficieny of effect size ßi and not necessarily leads to bias estimation.Since Bachelor&School = 0.54,internships&interview = 0.46 are the highest among other positive and negative correlations. These correlations and not too high or close to one . Thus its not a big problem for the next multiple regression models

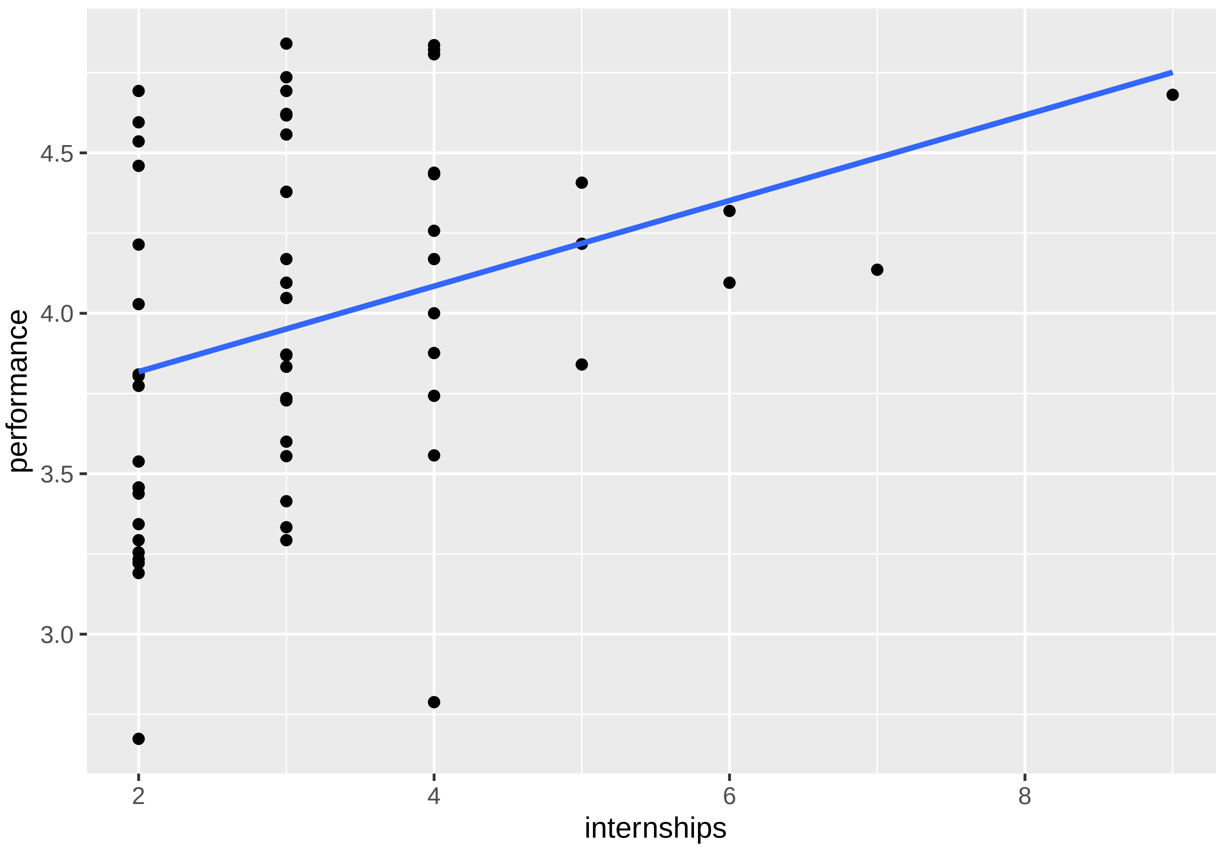
Hypothesis test using regression :

```
est1 <- lm(performance~internships , data = data1)
summary (est1)
```

```
##
## Call:
## lm(formula = performance ~ internships, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29652 -0.36092 -0.04426  0.39622  0.88914
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.55152    0.17204  20.643  < 2e-16 ***
## internships  0.13327    0.04916   2.711  0.00878 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5175 on 59 degrees of freedom
## Multiple R-squared:  0.1108, Adjusted R-squared:  0.09569
## F-statistic: 7.349 on 1 and 59 DF,  p-value: 0.008775
```

```r
ggplot(data1,aes(internships,performance)) + geom_point()+ geom_smooth(method='lm' , se = FALSE)
```



Since p-value = 0.00878 , we reject the null hypothesis stating that there is no association between performance and number of previous internships.At this value we have a strong evidence against the null hypothesis and its accepted that there is association between performance and number of internships made by a candidate

However , R-squared or correlation coefficient = 0.1108 indicates that the model is not agood fit and only 10% of performance variability described by this model and there are more variables need to be considered.

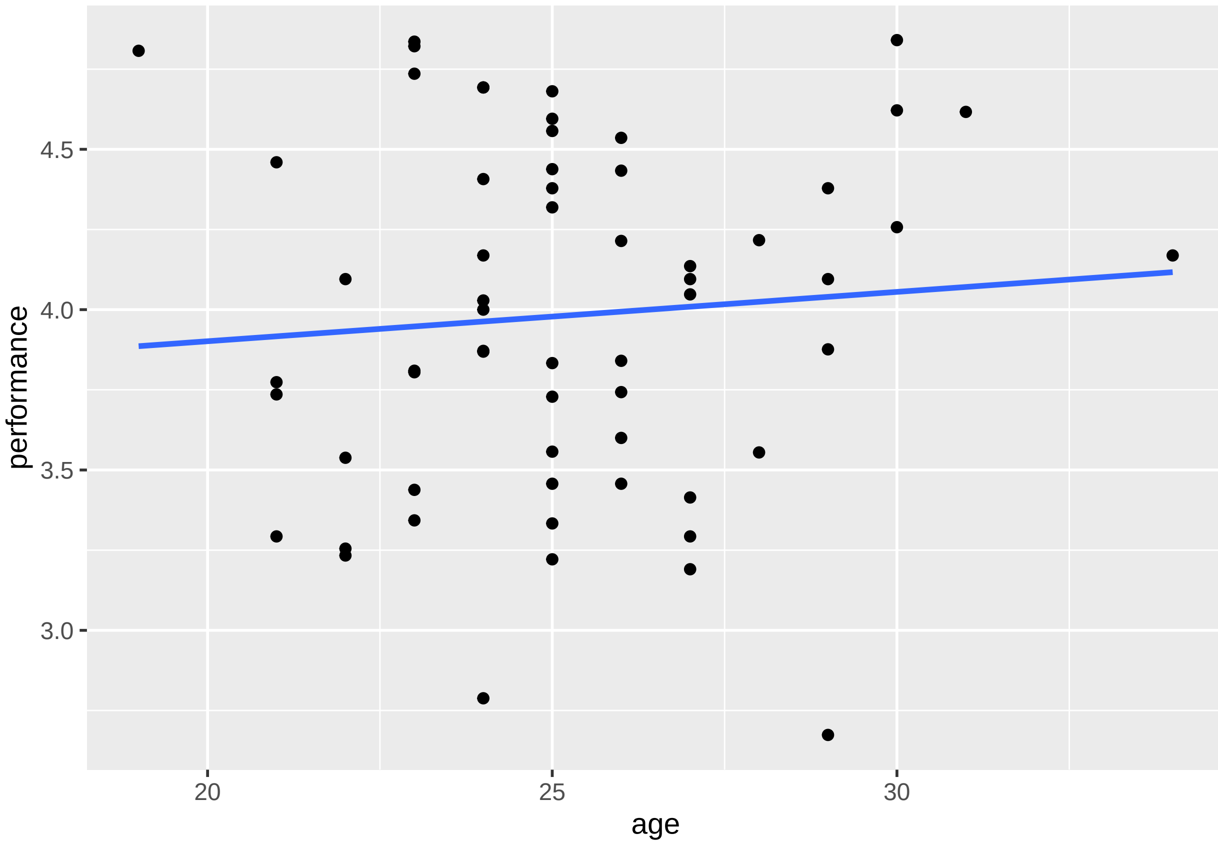
Step 3: "Regression Analysis"

In all models, PERFORMANCE is the dependent variable.

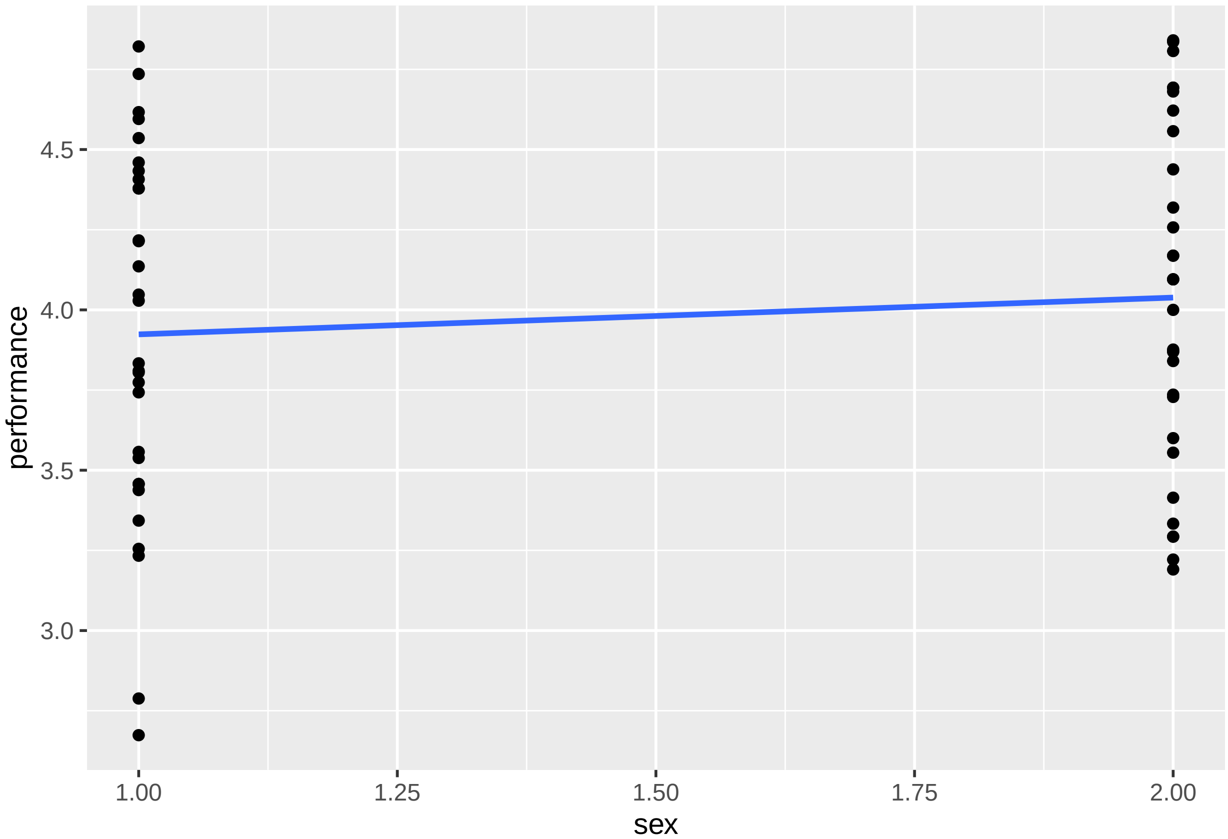
Since R-squared is a biased estimator for multiple regression analysis and increases when variable is added , the adjusted-R squared is the reference for a good model.

Assumptions : ## 1) Independent variables linearity with Performance : Overall linearity exists between explanatory variables and Performance.

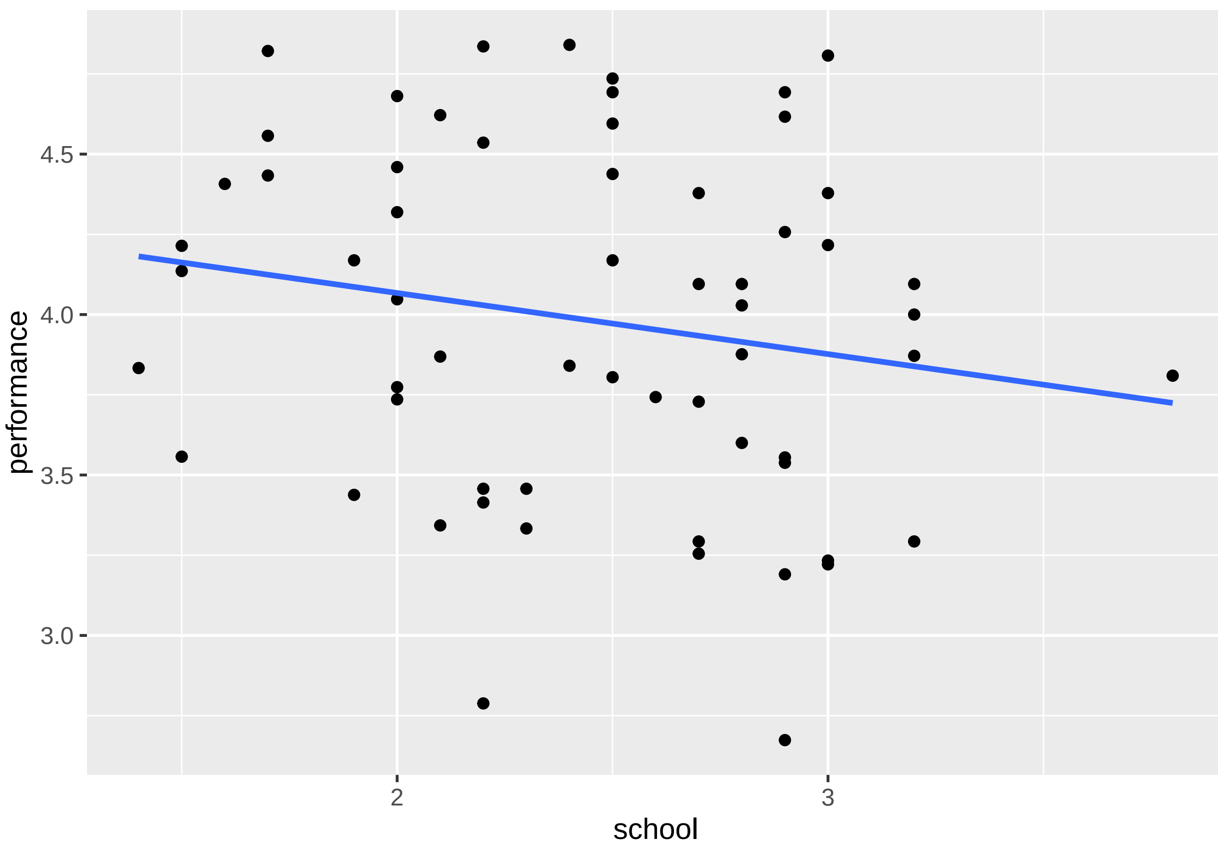
```
ggplot(data1 , aes(age,performance)) + geom_point() + geom_smooth(method = "lm" , se = FALSE) #Performa
```



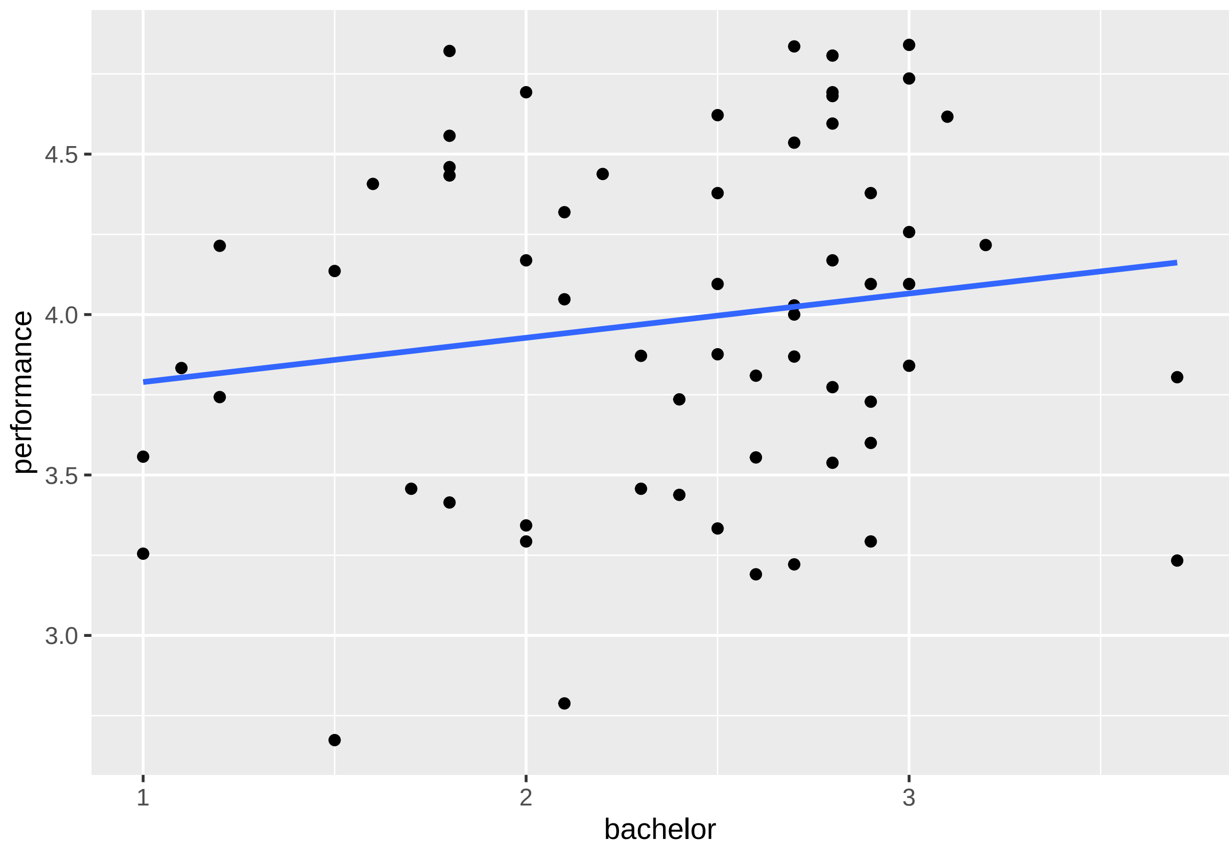

```
ggplot(data1 , aes(sex,performance)) + geom_point() + geom_smooth(method = "lm" , se = FALSE) #Performa
```



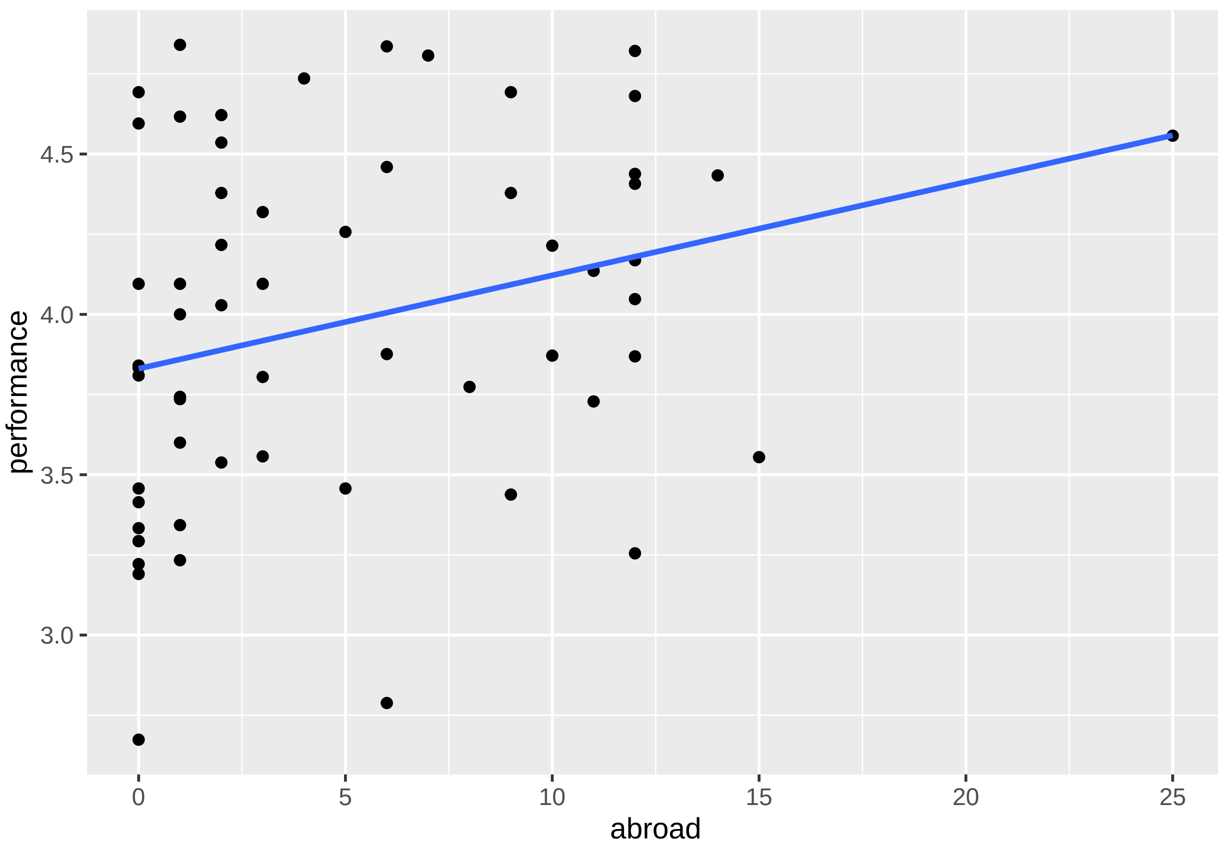

```
ggplot(data1 , aes(school,performance)) + geom_point() + geom_smooth(method = "lm" , se = FALSE) #Perfo
```



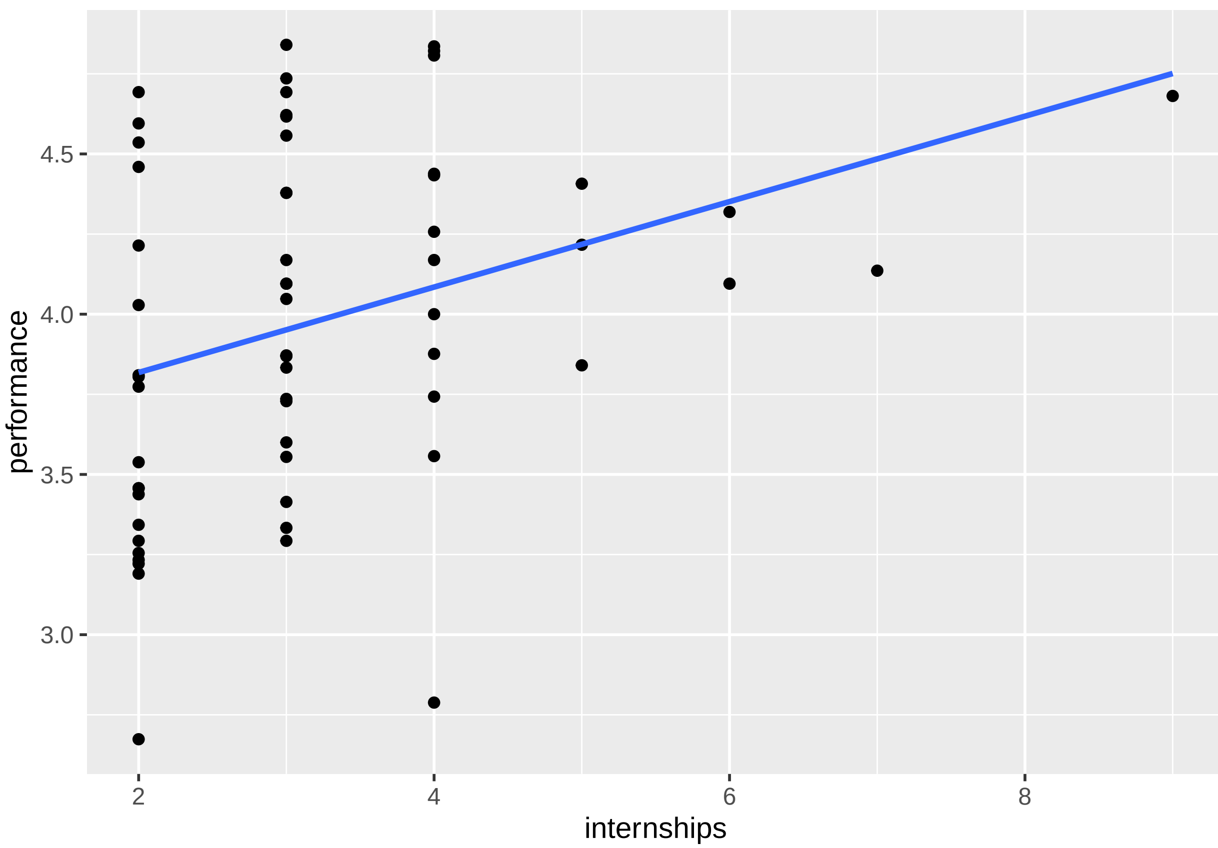

```
ggplot(data1 , aes(bachelor,performance)) + geom_point() + geom_smooth(method = "lm" , se = FALSE) #Per
```



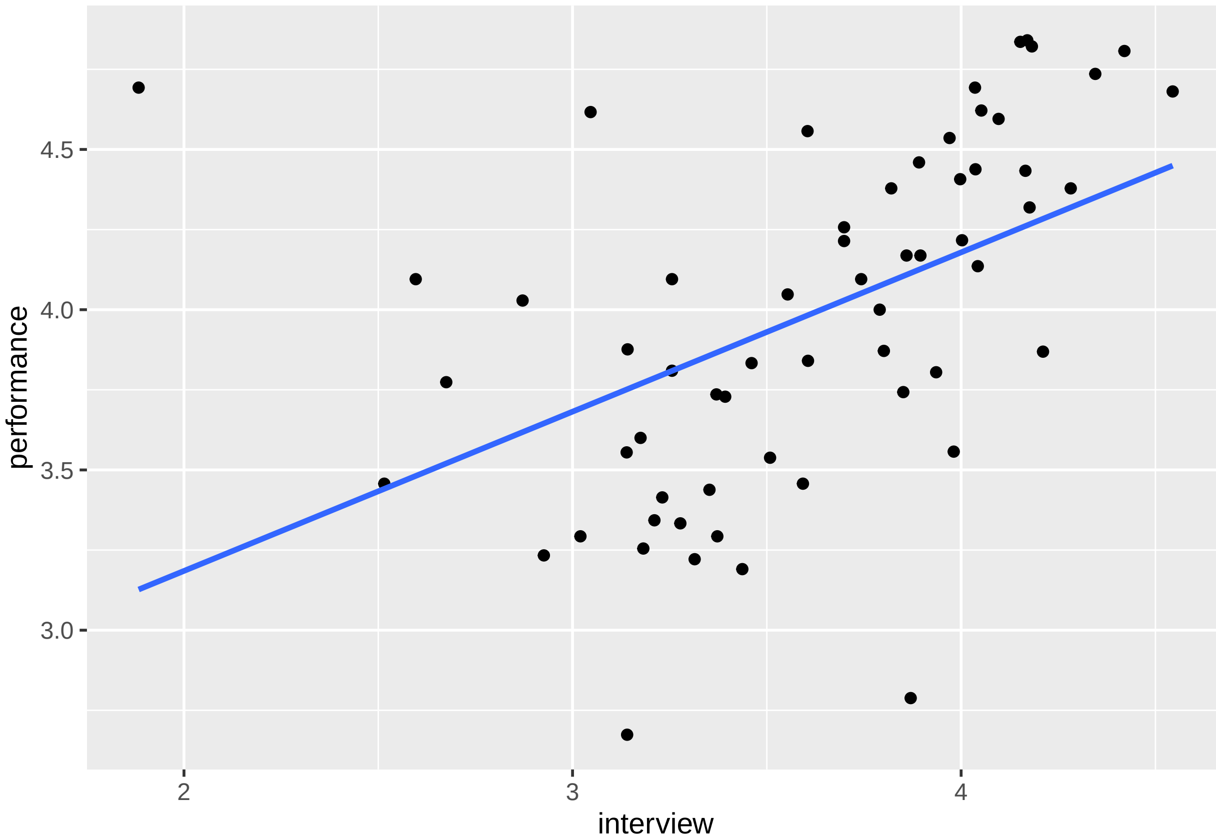

```
ggplot(data1 , aes(abroad,performance)) + geom_point() + geom_smooth(method = "lm" , se = FALSE) #Perfo
```




```
ggplot(data1 , aes(internships,performance)) + geom_point() + geom_smooth(method = "lm" , se = FALSE) #
```




```
ggplot(data1 , aes(interview,performance)) + geom_point() + geom_smooth(method = "lm" , se = FALSE) #Pe
```



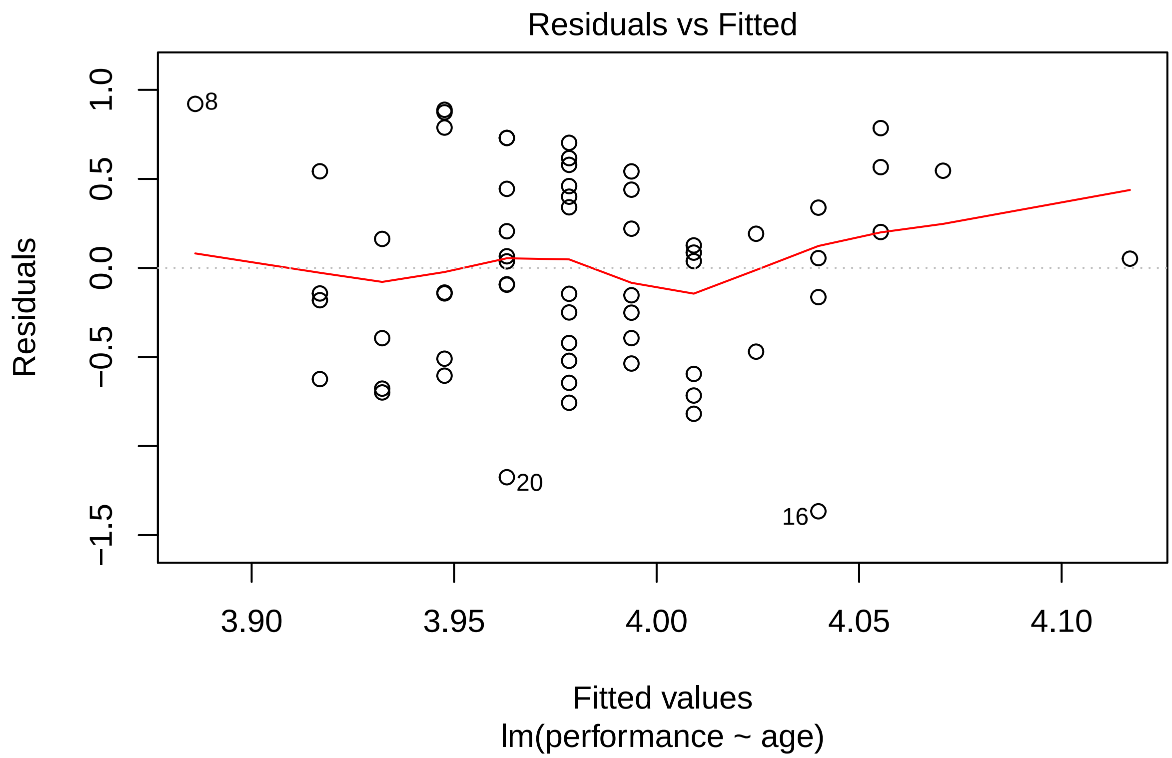
2)

Checking exogeneity assumption : this assumption is accepted here also , however there is a spurious correlation among internships , interviews and abroad as all of them are correlated the most with performance , but the correlation is not significant and thus its not addmissble to omit any of them

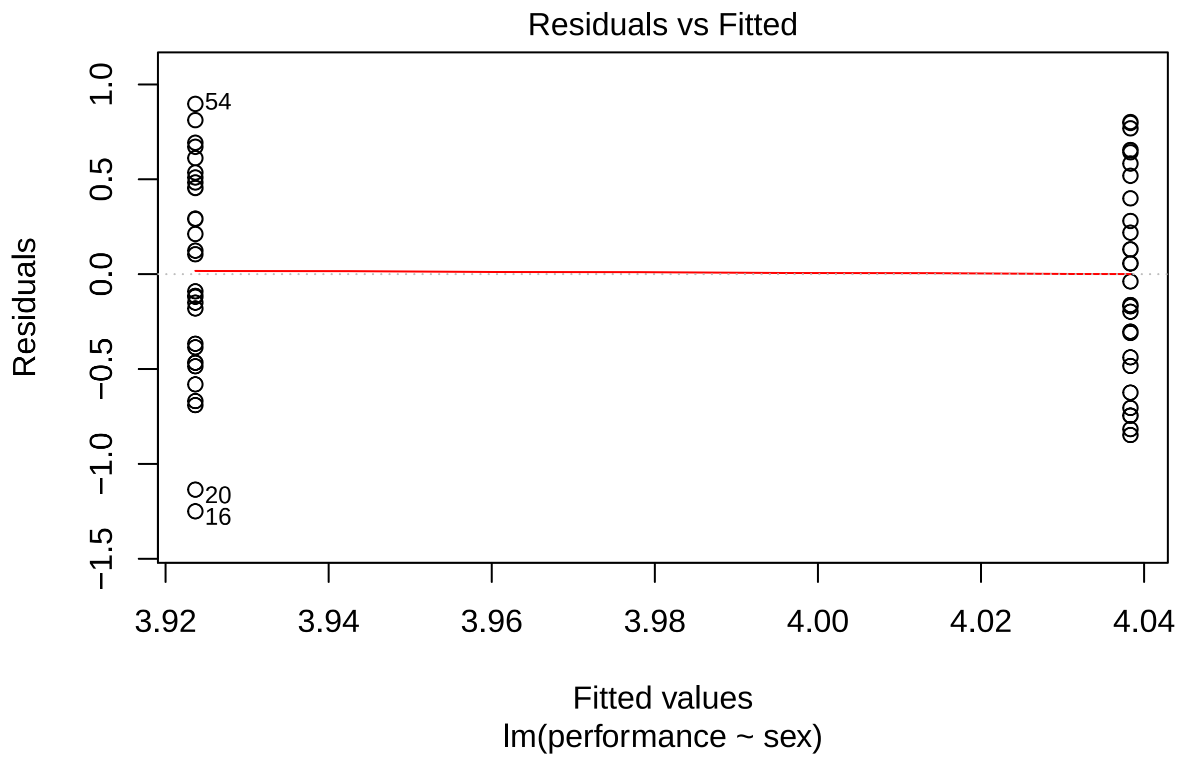
3)

homoscedasticity assumption can be seen from residual plots

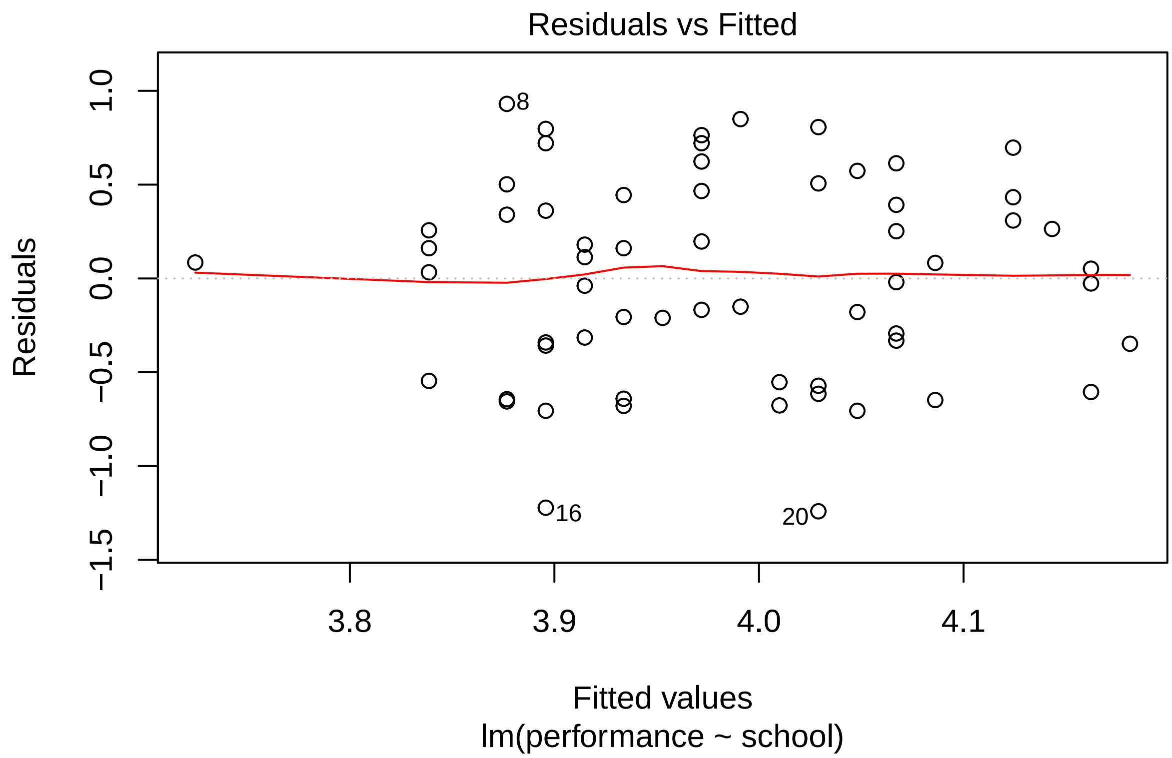
```
plot(lm(performance~age , data = data1) , which = 1) #performance~age
```



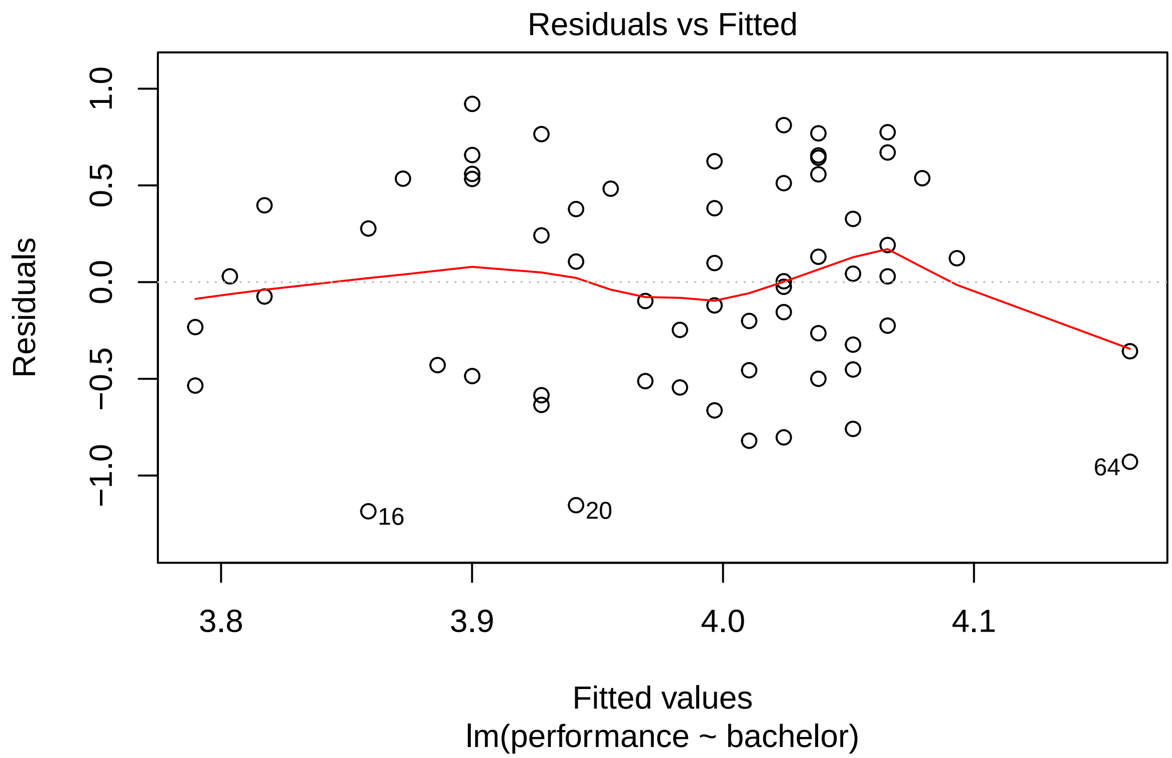

```
plot(lm(performance~sex , data = data1) , which = 1) #performance~sex
```



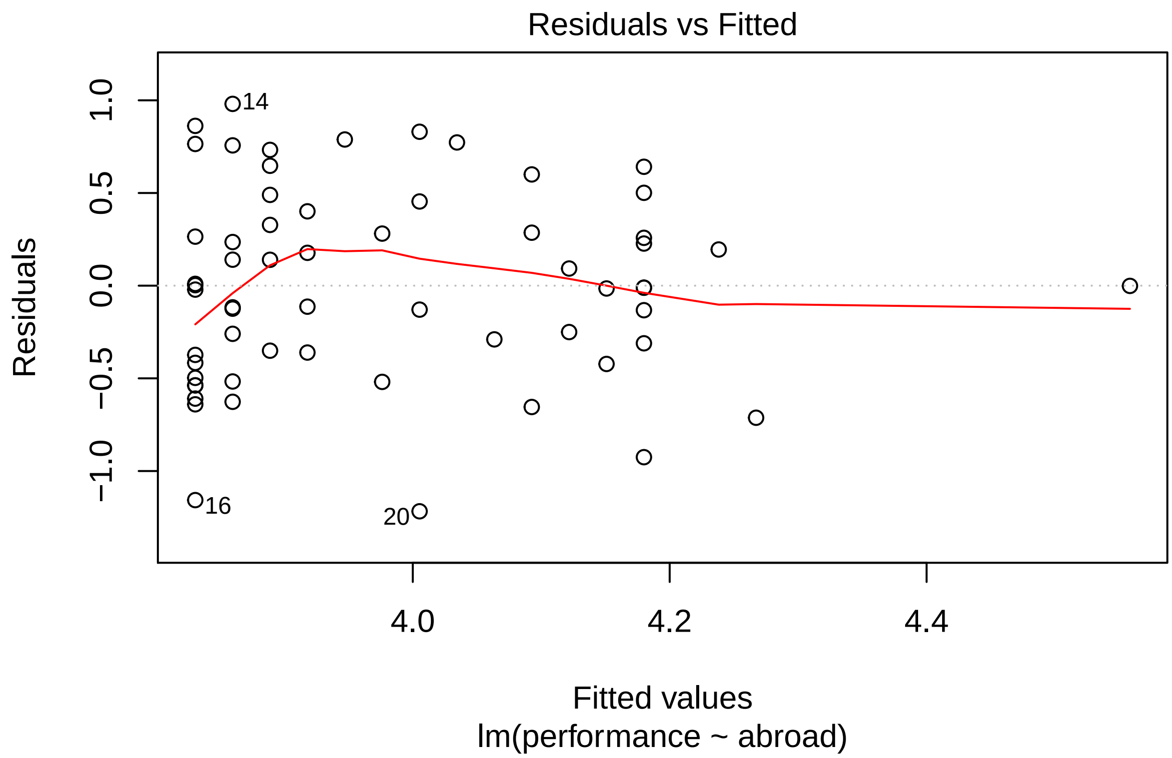

```
plot(lm(performance~school , data = data1) , which = 1) #performance~school
```



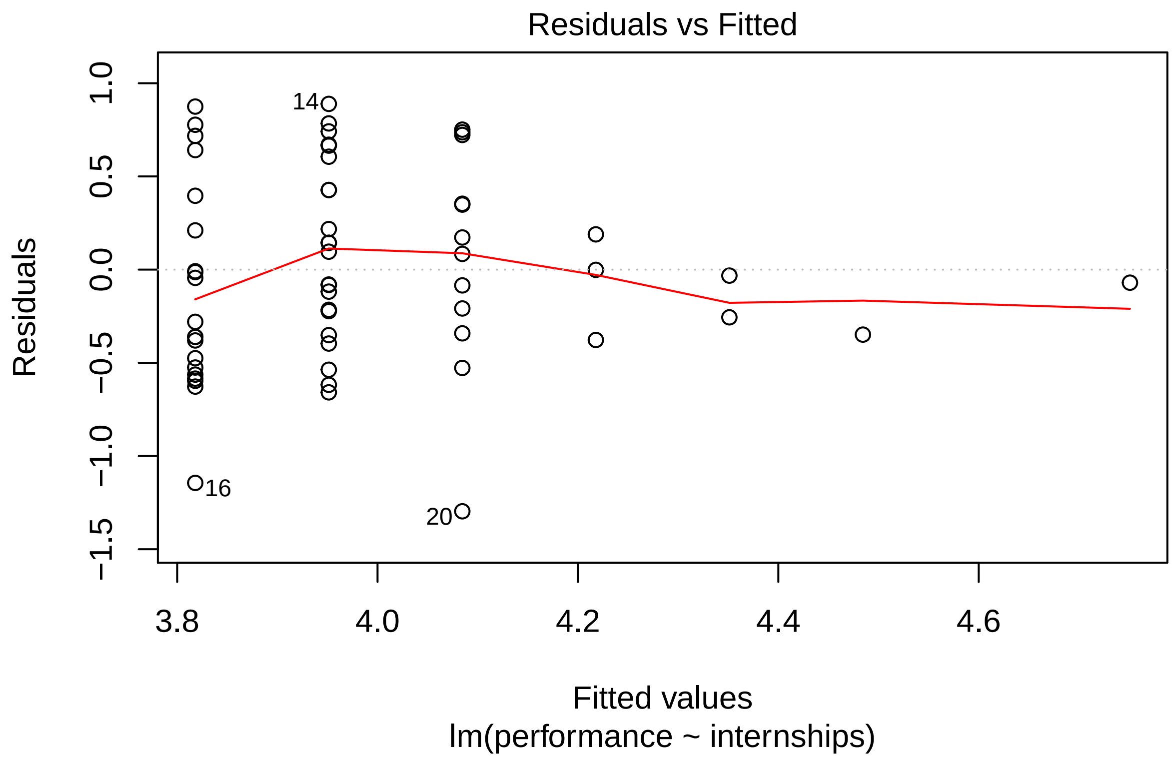

```
plot(lm(performance~bachelor , data = data1) , which = 1) #performance~bachelor
```



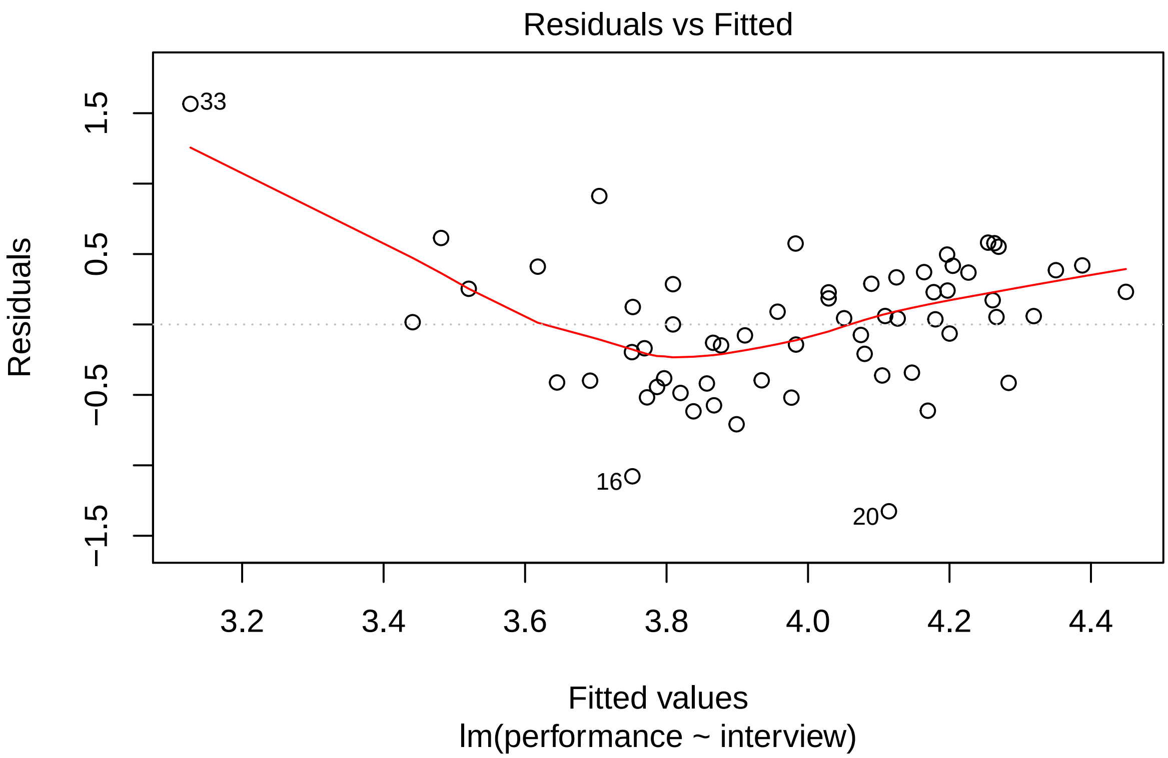

```
plot(lm(performance~abroad , data = data1) , which = 1) #performance~abroad
```




```
plot(lm(performance~internships , data = data1) , which = 1) #performance~internships
```




```r
plot(lm(performance~interview , data = data1) , which = 1) #performance~internships
```



Overall variabilty is accepted even though there are some outliers in the models which is normal.

Multicorrinealty , variations already exists and auto correlation assumptions are assumed.

A5

The R-squared value depicts the DV explained variability realised in each model and tell us about the model is agood fit or not.

ßi(effect size) tell us how much response variable "Performance" on average increase/decrease at each additional unit of every independent variables holding others as constants at multiple regression

F-value is a test that all ßi for every Independent variable are zero.Null hypothesis such that no association between performance and all independent variables.

hypothesis test for all models :

H0 : there is no association between performance and all explantory variables included in the model (all ßi or effect size of independent variables = zero)

HA : There is an association between Performance and at least one of the explantory variables included in the model

Model 1

Model1: Age in years (AGE) and sex (SEX)

```r
model1 <- lm(performance ~ age + sex, data=data1)
summary(model1)
```

```
##
## Call:
## lm(formula = performance ~ age + sex, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30260 -0.44278  0.01477  0.49354  0.92039
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.50804    0.65039   5.394 1.33e-06 ***
## age          0.01256    0.02550   0.493    0.624
## sex          0.10406    0.14229   0.731    0.468
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5492 on 58 degrees of freedom
## Multiple R-squared:  0.01541,    Adjusted R-squared:  -0.01855
## F-statistic: 0.4538 on 2 and 58 DF,  p-value: 0.6375
```

R-squared Interpretation : the performace variability explained by the model is 0.01541 (1.5%) still a bad fit overall

F-statistic: 0.4538 , p-value: 0.6375

Interpretation : F value is large and thus we dont reject the null hypothesis which states all correlation coefficient are zero at p-value: 0.6375

Explanatory variable : since the model doesnt fit there is no explanatory variable in this model

Model 2

Model 1 + school leaving grade (SCHOOL) and bachelor grade (BACHELOR)

```r
model2 <- lm(performance ~ age + sex + school + bachelor, data=data1)
summary(model2)
```

```
##
## Call:
## lm(formula = performance ~ age + sex + school + bachelor, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.13311 -0.40257  0.07331  0.39238  0.99678
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.63607    0.69306   5.246 2.46e-06 ***
## age          0.01868    0.02431   0.769  0.44542
## sex          0.09696    0.14147   0.685  0.49592
```

```
## school      -0.41483    0.15294  -2.712  0.00886 **
## bachelor     0.31067    0.13085   2.374  0.02104 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5201 on 56 degrees of freedom
## Multiple R-squared:  0.1475, Adjusted R-squared:  0.08662
## F-statistic: 2.422 on 4 and 56 DF,  p-value: 0.05881
```

R-squared = 0.1475

Interpretation : the performace variability explained by the model becomes 0.1475 (almost 15%).

F-statistic: 2.422 , p-value: 0.05881

Interpretation : F value is lower = 2.422 , However the model is better than before but still larger than 5% signficance level as p-value = 0.05881 so we fail to reject the null hypothesis

Explanatory variable : bachelor grade at p-value=0.02104 & School grade at p-value = 0.02104 both are lower than significant level.

Model 3

Model3: Model 2 + total duration of previous stays abroad in months (ABROAD) and number of previously completed internships at home and abroad (INTERNSHIPS)

```r
model3 <- lm(performance ~ age + sex + school + bachelor + abroad + internships, data=data1)
summary(model3)
```

```
##
## Call:
## lm(formula = performance ~ age + sex + school + bachelor + abroad +
##     internships, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24252 -0.34898 -0.03896  0.31324  1.09607
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.136773   0.696550   4.503 3.61e-05 ***
## age          0.011877   0.023901   0.497   0.6213
## sex         -0.008806   0.141745  -0.062   0.9507
## school      -0.247422   0.159359  -1.553   0.1264
## bachelor     0.312876   0.125628   2.490   0.0159 *
## abroad       0.022351   0.013124   1.703   0.0943 .
## internships  0.092698   0.053196   1.743   0.0871 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4978 on 54 degrees of freedom
## Multiple R-squared:  0.2468, Adjusted R-squared:  0.1631
## F-statistic: 2.949 on 6 and 54 DF,  p-value: 0.01457
```

R-squared = 0.2468

Interpretation : the performace variability explained by the model increases to 0.2468 (almost 25%) . Adjusted R-squared increased to 0.1631.

F-statistic: 2.949 , p-value: 0.01457

Interpretation : at p-value = 0.01457 which is lower than the significance level of 5 % , we reject the null hypothesis that states that there is no association between all independent variables and performance of candidates.

Explanatory variable : bachelor at p-value = 0.0159 , abroad at p-value = 0.0943 and internships at p-value = 0.0871.

Now when h0 is rejected , its seen that bachelor grade , semesters abroad and number of internships has association with performance and each has effect size ßi (0.312876 ,0.022351 and 0.092698)respectively holding all equal.

Model 4

Model4: Model 3 + performance in the interview (INTERVIEW)

```r
model4 <- lm(performance ~ age + sex + school + bachelor + abroad + internships + interview , data=data1
summary(model4)
```

```
##
## Call:
## lm(formula = performance ~ age + sex + school + bachelor + abroad +
##     internships + interview, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26292 -0.31620 -0.00085  0.29548  1.51859
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.07342    0.80389   2.579   0.0127 *
## age          0.01191    0.02293   0.520   0.6055
## sex          0.04660    0.13797   0.338   0.7369
## school      -0.19204    0.15465  -1.242   0.2198
## bachelor     0.24380    0.12398   1.966   0.0545 .
## abroad       0.01532    0.01293   1.185   0.2413
## internships  0.03704    0.05614   0.660   0.5122
## interview    0.33989    0.14282   2.380   0.0209 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4776 on 53 degrees of freedom
## Multiple R-squared:  0.3195, Adjusted R-squared:  0.2297
## F-statistic: 3.555 on 7 and 53 DF,  p-value: 0.003321
```

R-squared = 0.3195

Interpretation : the performace variability explained by the model increases to 0.3195 (almost 32%).

Adjusted R-squared:0.2297 is the maximum so far and it indicates the model is not a perfect fit . only 22% variability of performance is explained in this model.

F-statistic: 3.555 , p-value: 0.003321

Interpretation : at p-value = 0.003321 which very low comparing with the significance level of 5%. There is a very strong evidence against the null hypothesis.Therefore , this model is confident against the H0 and provides a strong evidence to accept the HA that there is association between performance and the explantory variables specially the Interview performance and bachelor grade in this model.

Explanatory variable : this time a notable effect size of interview performance at the first place at p-value = 0.0209 . then the bachelor grade which decreased alittle bit comparing with model3 at p-value= 0.0545 it outstand one of disadvantage of p-value appraoch as its very close to the significant level and it should be rejected.

Step 4: "Interpretation of the Regression Analysis Results"

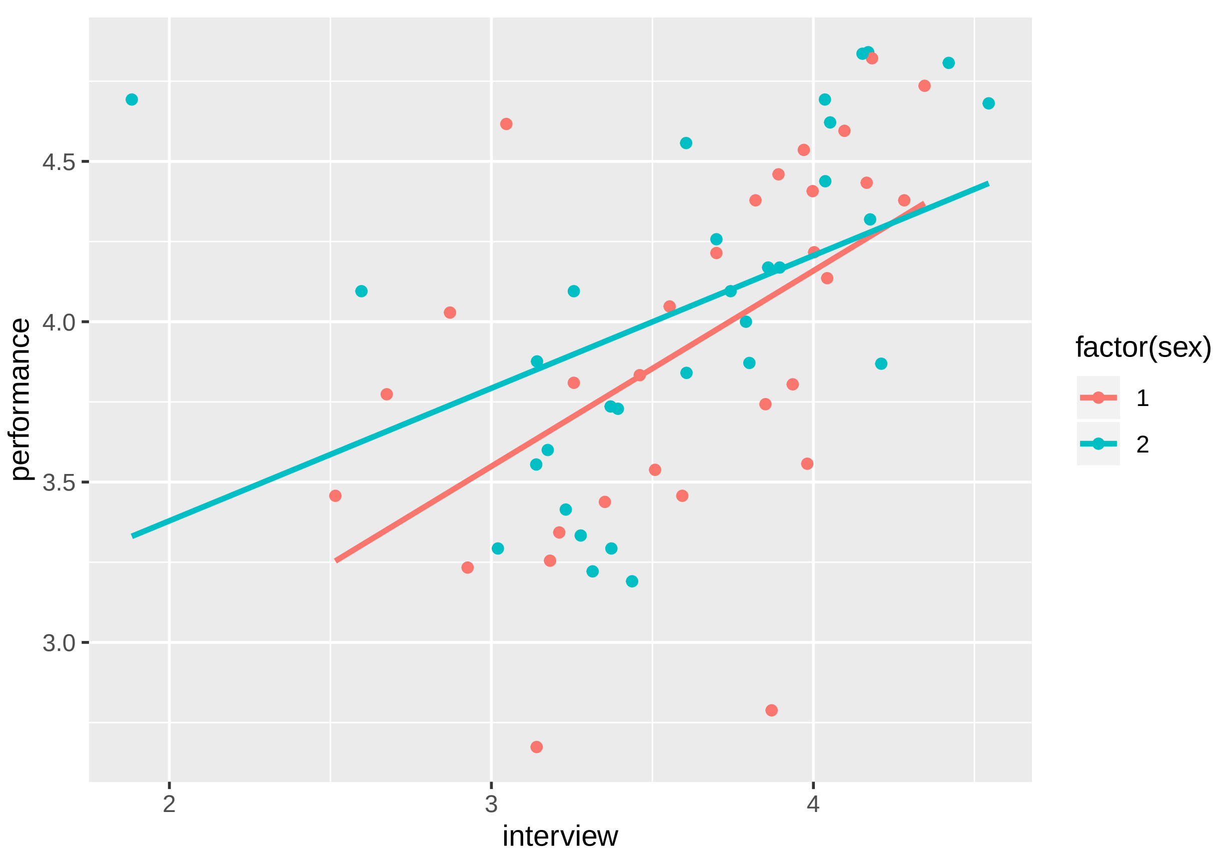
##A6

model4 is non-standardized : ßi of variables or effect size

age-coefficient : holding all equal performance score increase on average by 0.01191 point when the candidate age increase one year.

sex-coefficient : holding all equal performance score of one of the both sex is 0.04660 point higher than the other.

```
ggplot(data1, aes(x = interview, y = performance, colour = factor(sex))) +  geom_point() + geom_smooth(
```



```r
# visuallising gender performance with respect to male and females with thier respective interview scor
# now its obvious that sex 2 (females) are the reference variables and they have higher average perform
```

School-coefficient: performance score decreases by -0.19204 when school grade increase by one unit holding others as constants.

Bachelor-coefficient: holding all equal performance increases on average 0.24380 point at each additional bachelor grade point.

Abroad-coefficient: when number of semester increase by one candidate is expected to achieve more on 0.01532 performance score.

Internships-coefficient : its expected that candidate will achieve 0.03704 point more score at each additional internships.

Interview-coefficient : at interview when the candidate achive one point this will increase the over all performance score by 0.33989 point.

Strongest effect here is interview-coefficient ß=0.33989.

standardized coefficients:

```r
# standardizing ß coeifficents with beta function from reghelper package
model4_z <- beta(model4)
model4_z
```

```
##
## Call:
## lm(formula = c("performance.z ~ age.z + sex.z + school.z + bachelor.z + abroad.z + internships.z + "
## "    interview.z"), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.32089 -0.58108 -0.00156  0.54300  2.79073
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.366e-16  1.124e-01   0.000   1.0000
## age.z          6.159e-02  1.185e-01   0.520   0.6055
## sex.z          4.317e-02  1.278e-01   0.338   0.7369
## school.z      -1.857e-01  1.495e-01  -1.242   0.2198
## bachelor.z     2.798e-01  1.423e-01   1.966   0.0545 .
## abroad.z       1.530e-01  1.292e-01   1.185   0.2413
## internships.z  9.250e-02  1.402e-01   0.660   0.5122
## interview.z    3.270e-01  1.374e-01   2.380   0.0209 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8777 on 53 degrees of freedom
## Multiple R-squared:  0.3195, Adjusted R-squared:  0.2297
## F-statistic: 3.555 on 7 and 53 DF,  p-value: 0.003321
```

Strongest standardized effect here is interview-coefficient ß= 0.0327.

age-Zcoefficient : for an increase of 1 standard deviation of age and holding all equal performance score increase on average by 0.06159

sex-Zcoefficient : for an increase of 1 standard deviation of females data and holding all equal thier performance score increase on average by 0.04317 more than the males

School-Zcoefficient: for an increase of 1 standard deviation of school data and holding all equal performance score decrease on average by -0.1857.

Bachelor-Zcoefficient: for an increase of 1 standard deviation of bachelor and holding all equal performance score increase on average by 0.2798

Abroad-Zcoefficient: for an increase of 1 standard deviation of abroad data and holding all equal performance score increase on average by 0.1530

Internships-Zcoefficient : for an increase of 1 standard deviation of internships data and holding all equal performance score increase on average by 0.09250

Interview-Zcoefficient : for an increase of 1 standard deviation of interview score data and holding all equal performance score increase on average by 0.3270

Hypothesis at 5% and 10% significance level :

F-statistic: 3.555 & p-value: 0.003321 :

The p-value is lower than both signficance and provides a strong evidence against the null hypothesis. This confirm our alternative hypothesis that there is an association between explantaory variables in model 4 and performance.

Decesion on the interview :

Based on the Adjusted R-squared=0.2297 (almost 23%) variability of performance is described by the model , it tells us that the model is not a perfect fit over all.

However from the empirical analysis interview must be proceeded as it has the most effect size on overall performance . This is justified over all part A questions.