M.Sc. in Data Science

Athens University of Economics & Business

ΟΠΑ AUEB

Statistics and Probability Assignment 2 Solutions Report

Spyros Mastrodimitris Gounaropoulos

January 2023

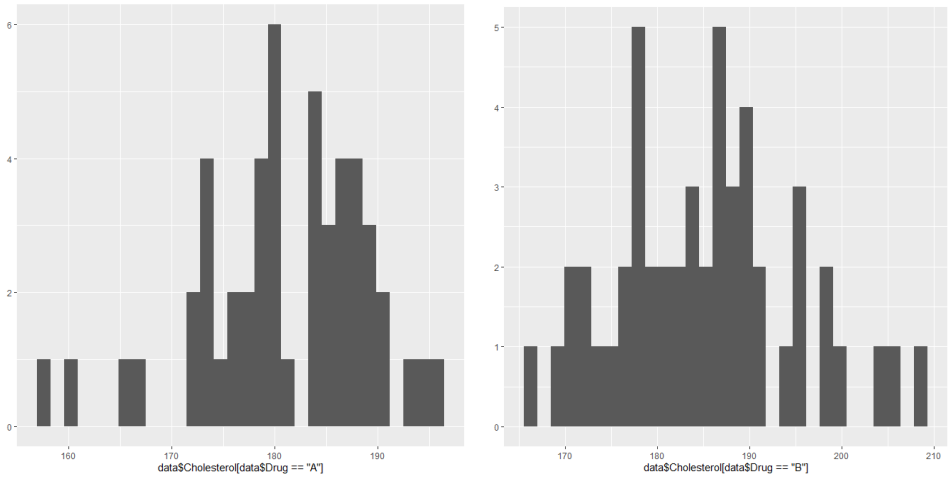
Assignment 1

a) After the data insertion we proceed using t.test(), in order to find the 99% confidence interval.

```
> t.test(data$Cholesterol,conf.level=0.99)

        One Sample t-test

data:  data$Cholesterol
t = 198.77, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 180.6816 185.5204
sample estimates:
mean of x
  183.101
```

So, for confidence interval of 99% cholesterol values are between 180.68 and 185.52.

b) At this point we visualize the cholesterol distributions.



Using the same technique:

```
> t.test(data$Cholesterol[data$Drug=="A"],conf.level=0.95)

        One Sample t-test

data:  data$Cholesterol[data$Drug == "A"]
t = 154.57, df = 49, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 178.6626 183.3694
sample estimates:
mean of x
  181.016
```

We find that the confidence interval of 95% for the volunteer cholesterol who tried the drug A is 178.66-183.37.

For drug B:

```
> t.test(data$Cholesterol[data$Drug=="B"],conf.level=0.95)

        One Sample t-test

data:  data$Cholesterol[data$Drug == "B"]
t = 135.06, df = 49, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 182.4306 187.9414
sample estimates:
mean of x
  185.186
```

From the 2 confidence intervals we observe a significant difference between the cholesterol values before and after the medication A. While drug B seems ineffective.

c) To be more precise we try to quantify the true difference.

```
> t.test(data$Cholesterol[data$Drug=="A"],data$Cholesterol[data$Drug=="B"],conf.level=0.9)

        Welch Two Sample t-test

data:  data$Cholesterol[data$Drug == "A"] and data$Cholesterol[data$Drug == "B"]
t = -2.3126, df = 95.66, p-value = 0.02289
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -7.164971 -1.175029
sample estimates:
mean of x mean of y
  181.016   185.186
```

We have proven than for 90% confidence interval the cholesterol of the volunteers who took drug A would drop 1.175-7.165 more than if they took drug B.

d) The next step is to reject the hypothesis that the two cholesterol means after the medication are equal.

```
> t.test(data$Cholesterol[data$Drug=="A"],data$Cholesterol[data$Drug=="B"],alternative = "less",p.value=0.05)

        Welch Two Sample t-test

data:  data$Cholesterol[data$Drug == "A"] and data$Cholesterol[data$Drug == "B"]
t = -2.3126, df = 95.66, p-value = 0.01144
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -1.175029
sample estimates:
mean of x mean of y
  181.016   185.186
```

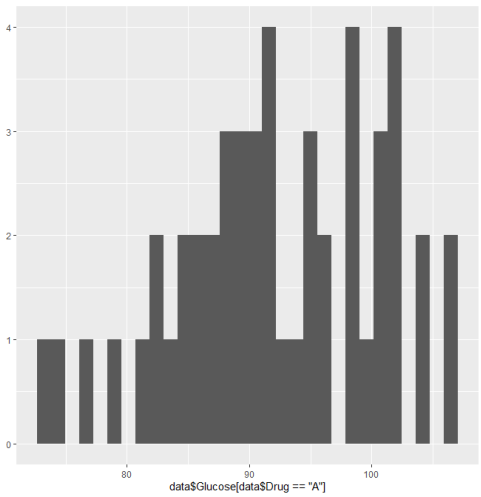
Using one-tailed hypothesis testing we find p_value < a, meaning we can reject the hypothesis of the equal means.

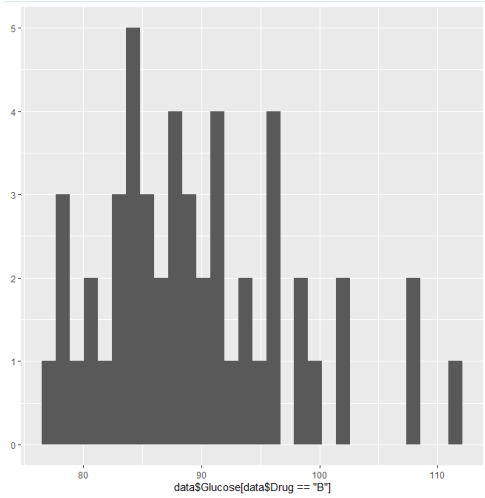
e) For Glucose fluctuation we will use var.test()

```
> var.test(data$Glucose[data$Drug=="A"],data$Glucose[data$Drug=="B"],alternative = "two.sided",p.value=0.01)

        F test to compare two variances

data:  data$Glucose[data$Drug == "A"] and data$Glucose[data$Drug == "B"]
F = 1.0484, num df = 49, denom df = 49, p-value = 0.8694
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5949146 1.8473933
sample estimates:
ratio of variances
          1.048352
```

p_value > α so we reject the Ho hypothesis, and any variance differences are statistically unimportant.

f) In the last question we looked for variation differences but what is the relation of the means?





From the plots a slight difference is spotted as drug B shows a longer tail to the right. The confidence intervals for 95% showed [89.768, 94.467] for A and [87.323, 91.912] for B.

Hypothesis testing for Ho mean == mean:

```
> t.test(data$Glucose[data$Drug=="A"],data$Glucose[data$Drug=="B"],alternative = "two.sided",var.equal = TRUE,p.value=0.95)

        Two Sample t-test

data:  data$Glucose[data$Drug == "A"] and data$Glucose[data$Drug == "B"]
t = 1.5297, df = 98, p-value = 0.1293
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.7431391  5.7431391
sample estimates:
mean of x mean of y
   92.118    89.618
```

We see that drug A has a higher mean.

g) Hypothesis testing and confidence interval 95%:

```
> prop.test(7,100,conf.level=0.95)

        1-sample proportions test with continuity correction

data:  7 out of 100, null probability 0.5
X-squared = 72.25, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.03101985 0.14376573
sample estimates:
   p
0.07
```

Volunteers suffering from myalgia are equal to 7 and the 95% confidence interval shows that by percentage are between 3.1% and 14.4% of all volunteers.

h)

```
> prop.test(7,100, p = 0.05, alternative = "greater",conf.level=0.05)

        1-sample proportions test with continuity correction

data:  7 out of 100, null probability 0.05
X-squared = 0.47368, df = 1, p-value = 0.2456
alternative hypothesis: true p is greater than 0.05
5 percent confidence interval:
 0.1180799 1.0000000
sample estimates:
   p
0.07
```

Indeed, it seems that for 5% of cases, patients suffering from myalgia represent more than 5% of all patients.

i) Checking Independence:

```
> chisq.test(myalgiaBalance)

        Pearson's Chi-squared test with Yates' continuity correction

data:  myalgiaBalance
X-squared = 2.4578, df = 1, p-value = 0.1169
```

p_value > α so we accept Ho, meaning that drug effects and myalgia symptoms are independent for a=0.05.

```
> chisq.test(myalgiaBalance2)

        Pearson's Chi-squared test with Yates' continuity correction

data:  myalgiaBalance2
X-squared = 4.3291, df = 1, p-value = 0.03747


> chisq.test(myalgiaBalance3)

        Pearson's Chi-squared test with Yates' continuity correction

data:  myalgiaBalance3
X-squared = 4.3956, df = 1, p-value = 0.03603
```

Although if the myalgia patient was absent from drug B group test or 2 more of the volunteers for drug A where suffering from myalgia then the result would be different.

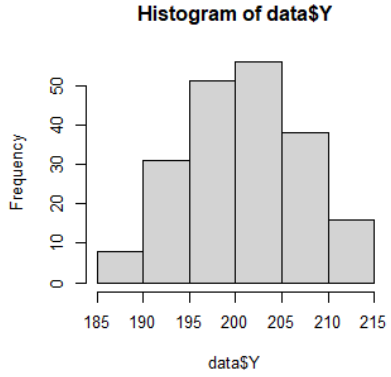
j) The two subsets have the same variance:

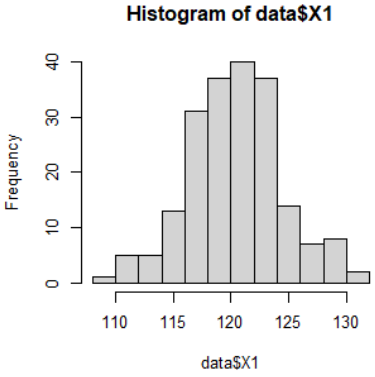
```
> var.test(symptoms,noSymptoms,conf.level=0.95)

        F test to compare two variances

data:  symptoms and noSymptoms
F = 1.328, num df = 6, denom df = 92, p-value = 0.5054
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5209815 6.5350403
sample estimates:
ratio of variances
          1.327961

> t.test(symptoms,noSymptoms,var.equal = TRUE,conf.level=0.95)

        Two Sample t-test

data:  symptoms and noSymptoms
t = -1.2264, df = 98, p-value = 0.223
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.326163   2.437991
sample estimates:
mean of x mean of y
 87.20000  91.14409
```

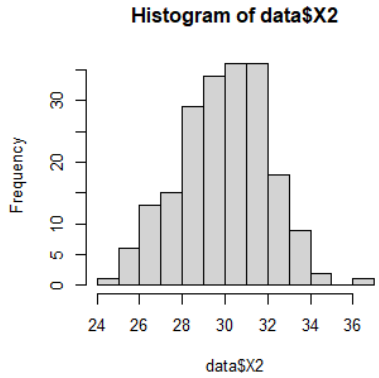
95% confidence interval for difference between mean for glucose between healthy and unhealthy volunteers is between -10.326 and 2.438 measuring units.

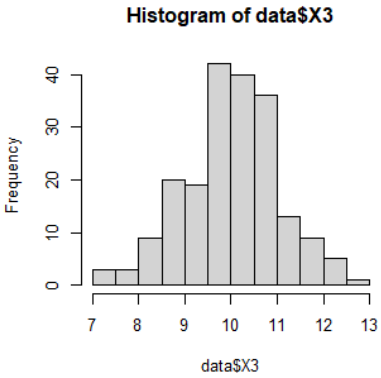
Assignment 2

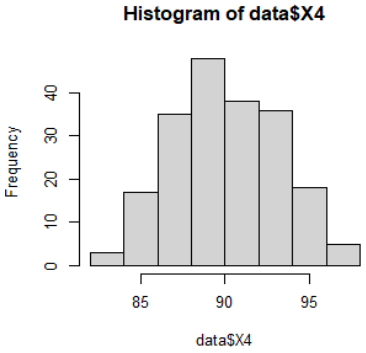
a)

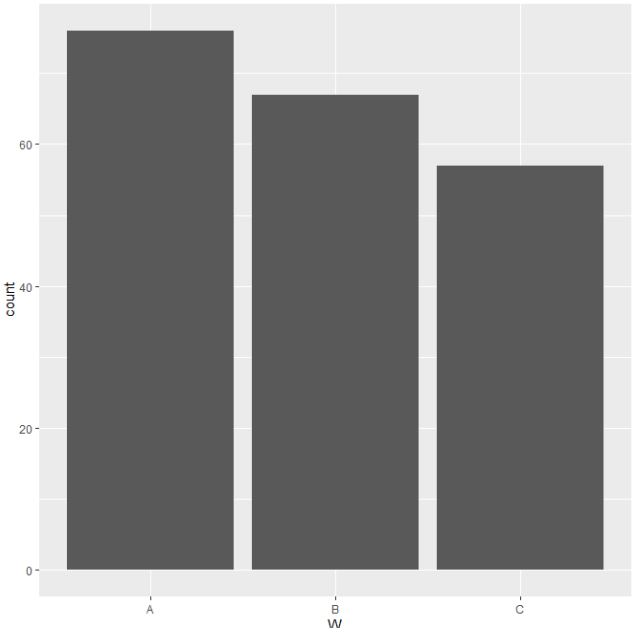


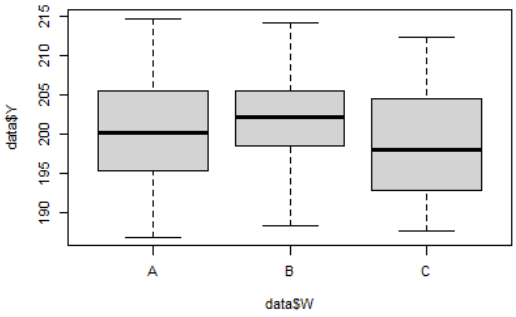


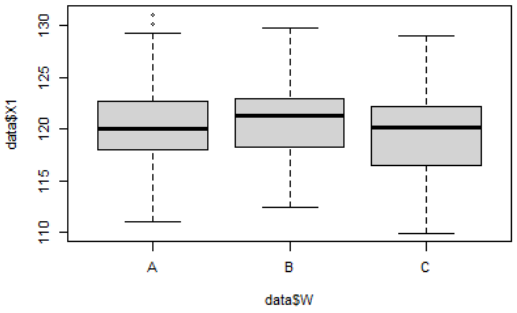


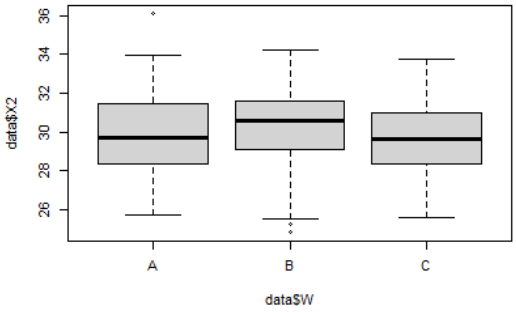


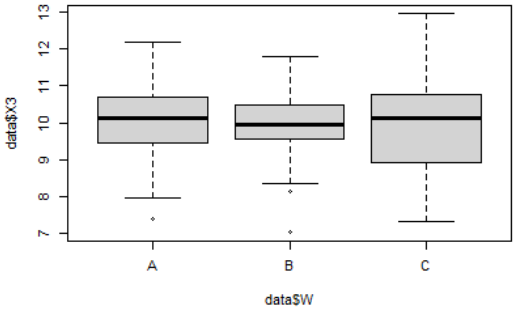


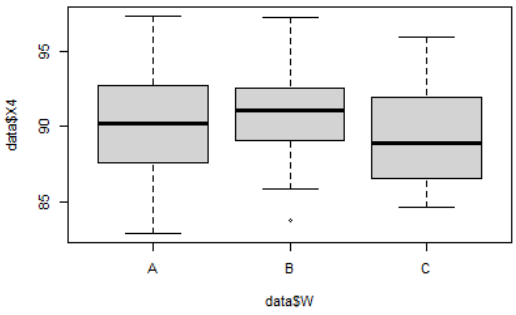


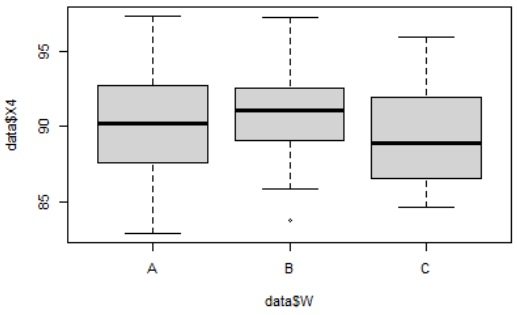










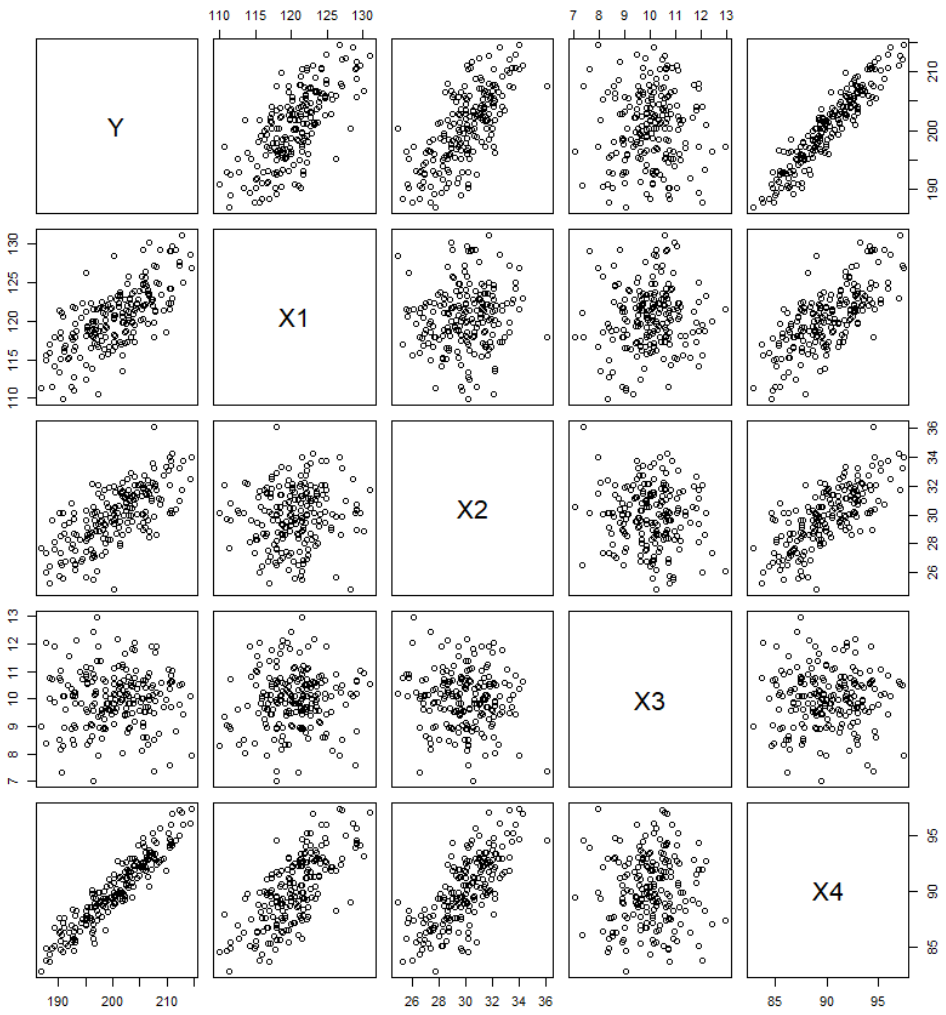


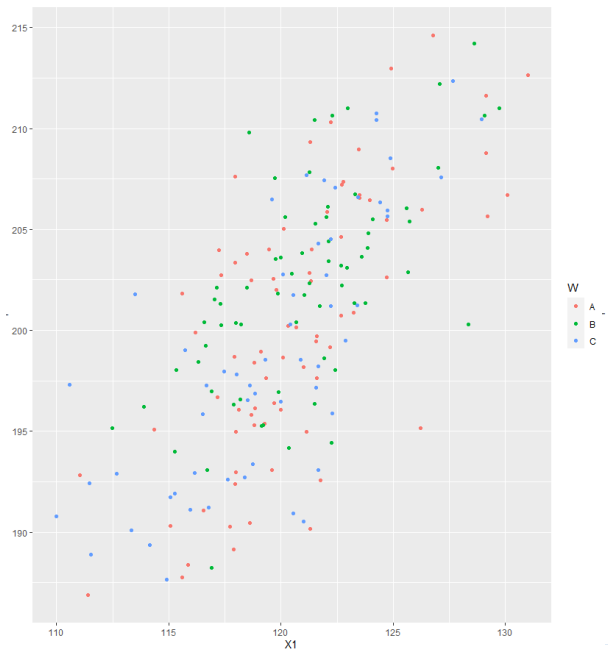

```
> fit = aov(XY~f*xy)
>
> summary(fit)
             Df  Sum Sq Mean Sq   F value   Pr(>F)
f             2     300     150    11.000 1.89e-05 ***
xy            4 4630611 1157653 84770.112  < 2e-16 ***
f:xy          8     202      25     1.852   0.0642 .
Residuals   985   13452      14
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> TukeyHSD(fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = XY ~ f * xy)

$f
          diff           lwr         upr     p adj
2-1  0.6495145 -0.0005656801  1.29959475 0.0502578
3-1 -0.7472281 -1.4269397338 -0.06751641 0.0270312
3-2 -1.3967426 -2.0957457631 -0.69773944 0.0000093

$xy
          diff        lwr        upr p adj
2-1  -80.09825  -81.10816  -79.08834     0
3-1 -170.61835 -171.62826 -169.60844     0
4-1 -190.52785 -191.53776 -189.51794     0
5-1 -110.41990 -111.42981 -109.40999     0
```

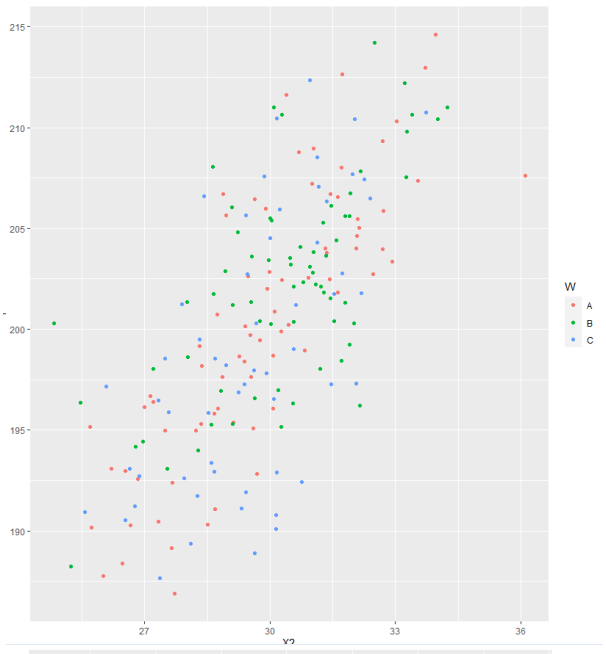
Analysis of Variance shows p_value < a meaning unequal variance between the variables.

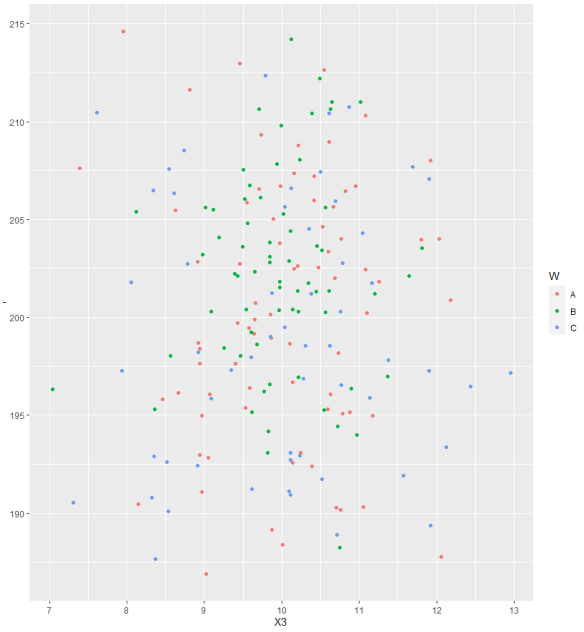
A more inquisitive look reveals that the variance of c is the odd one. While variance of A and B are equivalent.

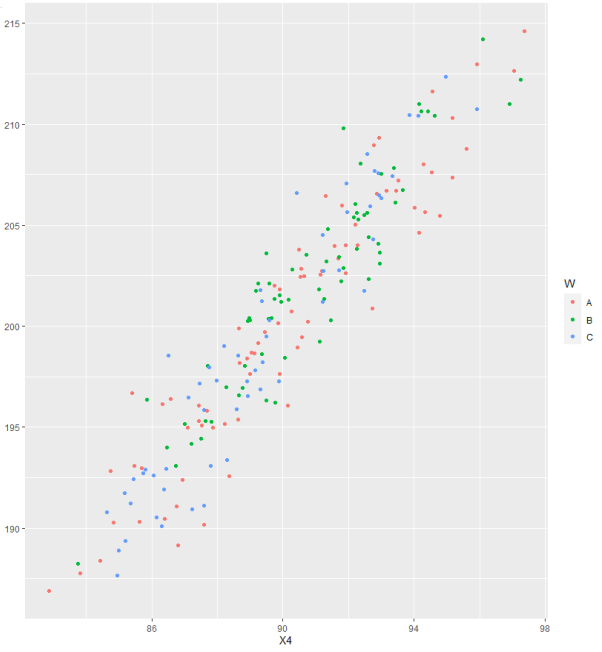
b)




215
210
205
200
195
190
110
115
120
125
130
X1
W
A
B
C


215
210
205
200
195
190
27
30
33
36
W
A
B
C


215
210
205
200
195
190
7
8
9
10
11
12
13
X3
W
A
B
C


215
210
205
200
195
190
86
90
94
98
X4
W
A
B
C


Y
X1
X2
X3
X4
W
Corr: 0.703***
A: 0.683***
B: 0.641***
C: 0.763***
Corr: 0.710***
A: 0.812***
B: 0.636***
C: 0.614***
Corr: 0.028
A: 0.032
B: 0.059
C: 0.025
Corr: 0.941***
A: 0.945***
B: 0.926***
C: 0.949***
Corr: 0.101
A: 0.231*
B: -0.087
C: 0.057
Corr: 0.078
A: 0.076
B: 0.066
C: 0.099
Corr: 0.718***
A: 0.745***
B: 0.598***
C: 0.776***
Corr: -0.094
A: -0.065
B: -0.112
C: -0.113
Corr: 0.746***
A: 0.808***
B: 0.724***
C: 0.649***
Corr: 0.019
A: 0.036
B: -0.030
C: 0.049
A
B
C

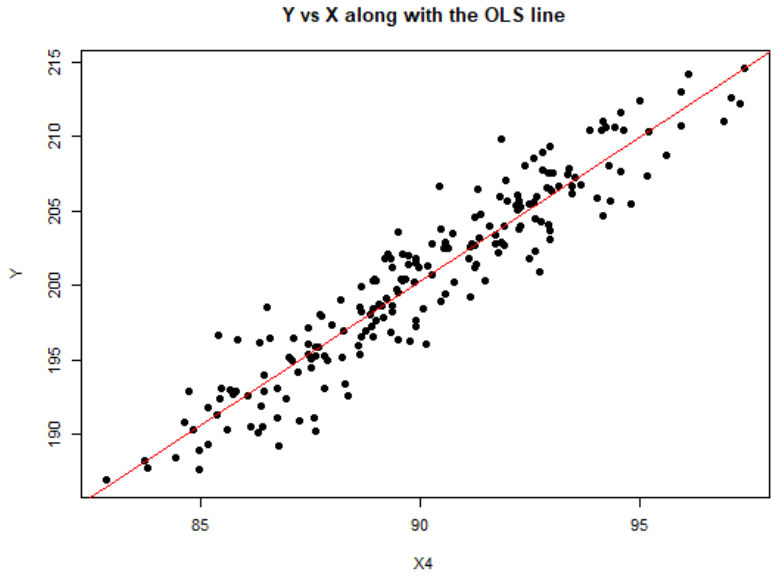
c) Before applying any regression, I tried to find if any correlation exists between the variables.

```
> cor(df$Y,df$X4)
[1] 0.9413168
> cor.test(df$Y,df$X4)

        Pearson's product-moment correlation

data:  df$Y and df$X4
t = 39.243, df = 198, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9231368 0.9552968
sample estimates:
      cor
0.9413168
```

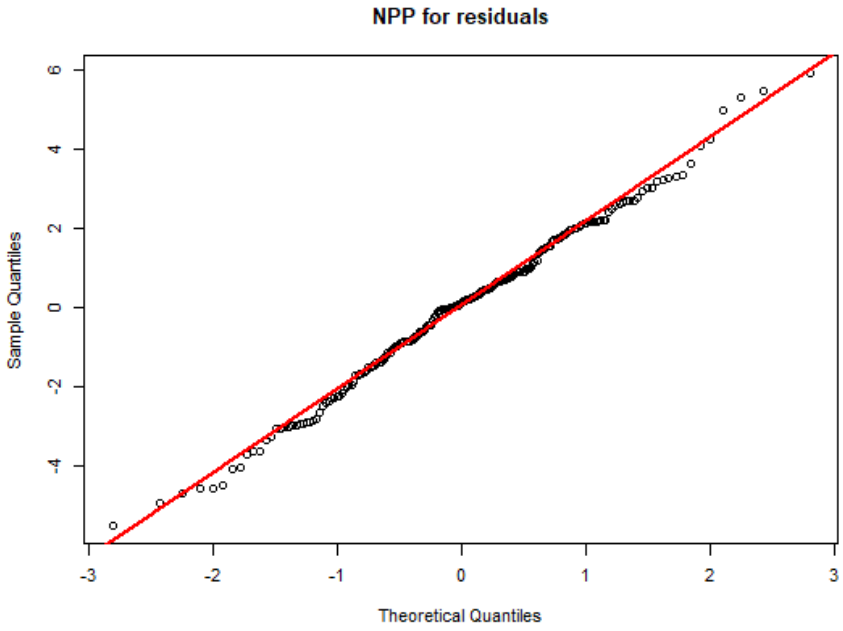
```
> fit<-lm(df$Y ~ df$X4)
> summary(fit)

Call:
lm(formula = df$Y ~ df$X4)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5133 -1.3818  0.1039  1.4803  5.9044

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.1973     4.4449   5.894  1.6e-08 ***
df$X4         1.9347     0.0493  39.243  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.129 on 198 degrees of freedom
Multiple R-squared:  0.8861,    Adjusted R-squared:  0.8855
F-statistic:  1540 on 1 and 198 DF,  p-value: < 2.2e-16
```

The regression function is y = 1.93*X4 +26.2

p_value < 0.05 so using X4 for our model is a good start since its impact is statistically important in describing y. In fact, it can describe about 88% of y.



Linearity and strong correlation are observed



The residuals seem to have normal distribution this is verified by the Shapiro Wilk test.

```
> shapiro.test(fit$residuals)

        Shapiro-Wilk normality test

data:  fit$residuals
W = 0.99416, p-value = 0.6247
```

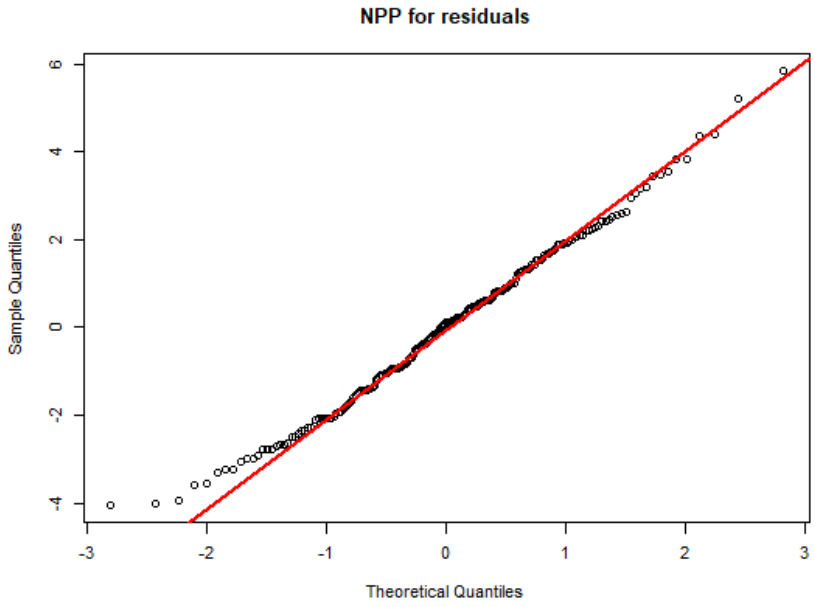
Not bad…

d) The model so far:

```
> fit<-lm(df$Y ~ df$X1+df$X2+df$X3+df$X4+W)
> summary(fit)

Call:
lm(formula = df$Y ~ df$X1 + df$X2 + df$X3 + df$X4 + W)

Residuals:
   Min     1Q Median     3Q    Max
-4.049 -1.430  0.115  1.321  5.832

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.87133    4.67653   3.608 0.000393 ***
df$X1        0.96269    0.14487   6.645 2.98e-10 ***
df$X2        1.93024    0.30071   6.419 1.03e-09 ***
df$X3        0.23562    0.13817   1.705 0.089735 .
df$X4        0.08043    0.28433   0.283 0.777583
W            0.19876    0.17001   1.169 0.243807
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.932 on 194 degrees of freedom
Multiple R-squared:  0.9081,    Adjusted R-squared:  0.9057
F-statistic: 383.3 on 5 and 194 DF,  p-value: < 2.2e-16
```

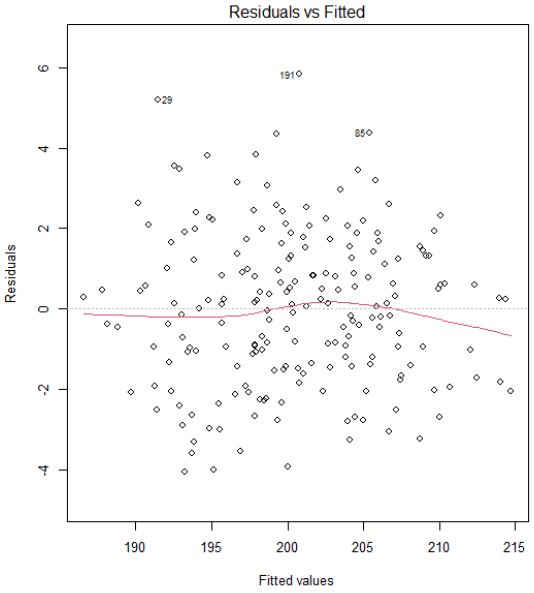
Our model can describe about 90.6% and X1, X2, are statistically important as α=0,05 and X3 for α=0,1. The resulting:

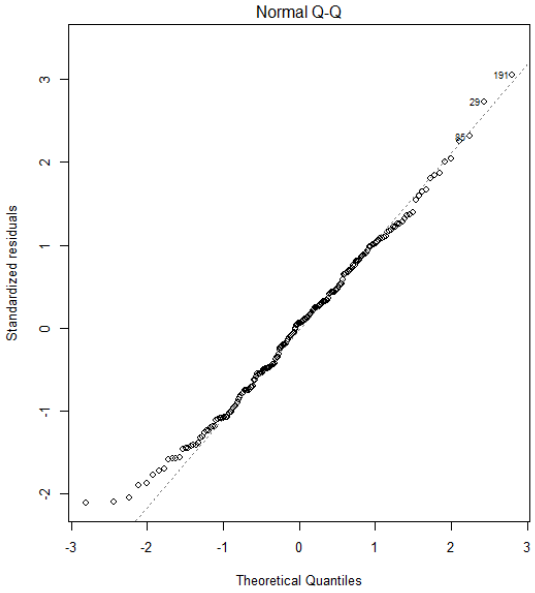
Y = 0.96269*X1 + 1.93024*X2 + 0.23562*X3 + 0.08043*X4 + 0.19876*W + 16.87133

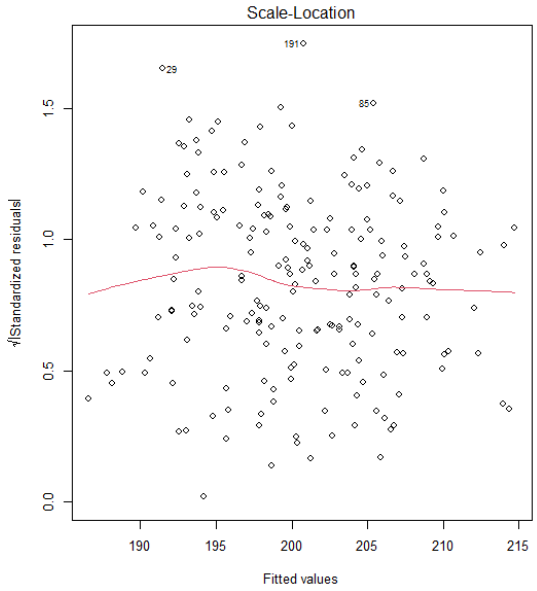


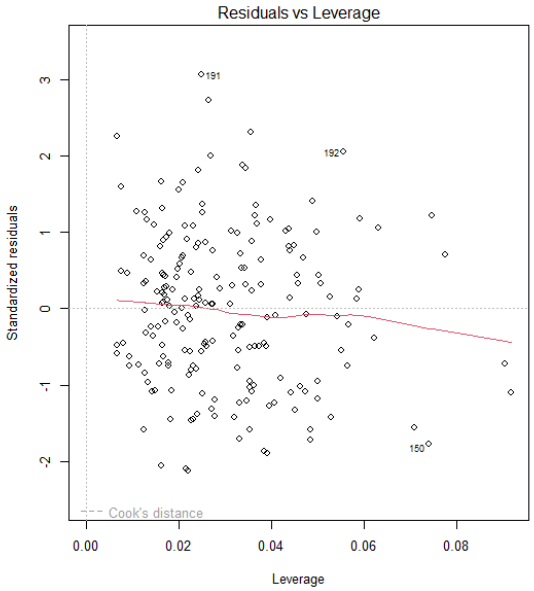
It is unclear if the residuals violate normal distribution.

```
> shapiro.test(fit$residuals)

        Shapiro-Wilk normality test

data:  fit$residuals
W = 0.99214, p-value = 0.3573

> lillie.test(fit$residuals)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  fit$residuals
D = 0.041548, p-value = 0.5455
```









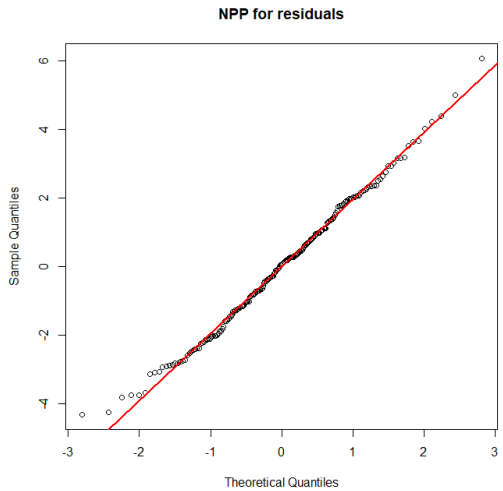
From the residual analysis we observe no information load that we failed to include toour model.

e)

```
> fit<-lm(df$Y ~ df$X1+df$X2)
> summary(fit)

Call:
lm(formula = df$Y ~ df$X1 + df$X2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3319 -1.3238  0.0474  1.3152  6.0823

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.88520    4.44928   4.469 1.32e-05 ***
df$X1        1.00369    0.03458  29.022  < 2e-16 ***
df$X2        1.99812    0.06806  29.360  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.94 on 197 degrees of freedom
Multiple R-squared:  0.9059,    Adjusted R-squared:  0.9049
F-statistic:   948 on 2 and 197 DF,  p-value: < 2.2e-16
```

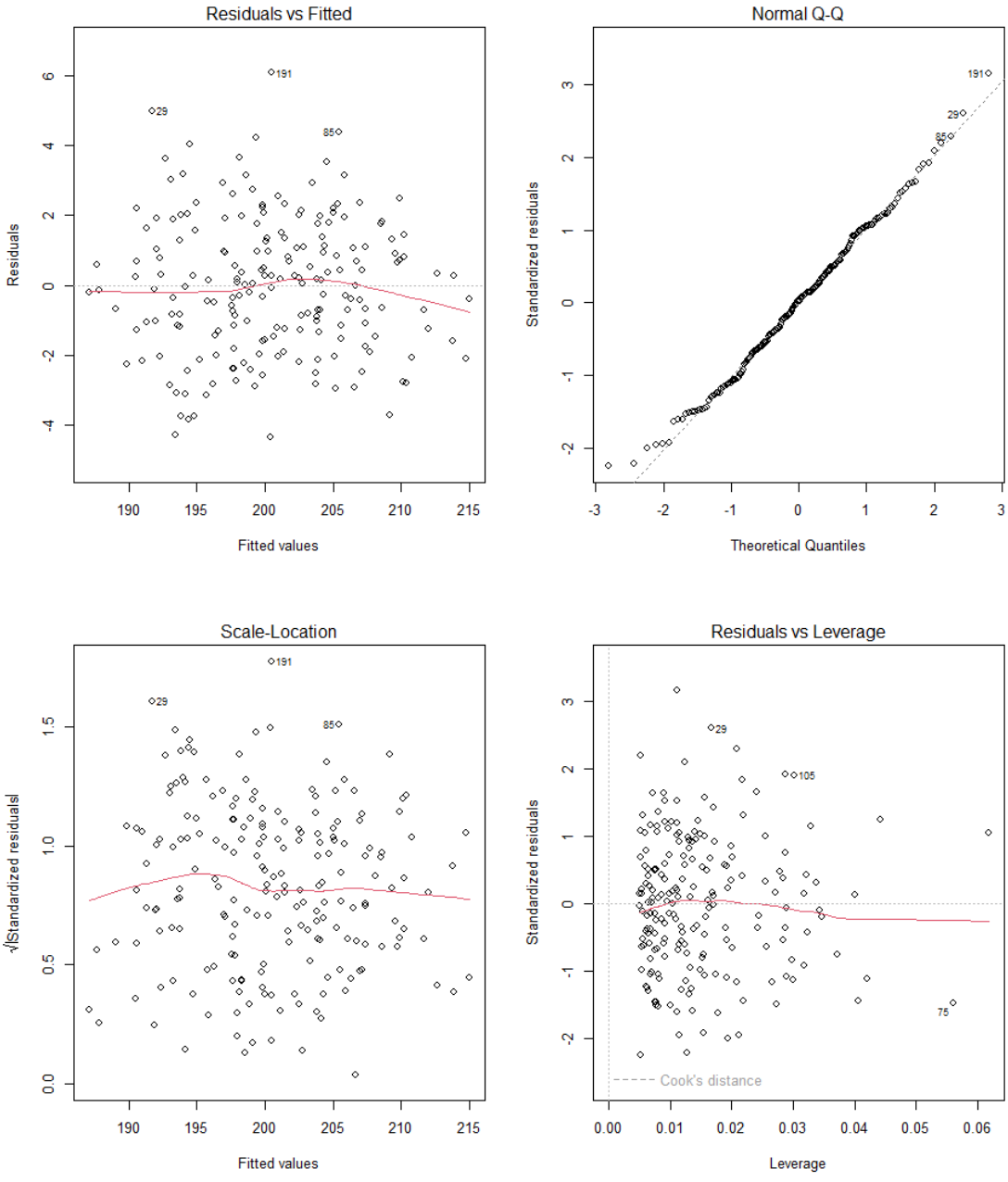
I selected the least possible variables achiving a 90%

y = 1.00369*X1 + 1.99812*X2 + 19.88520 .



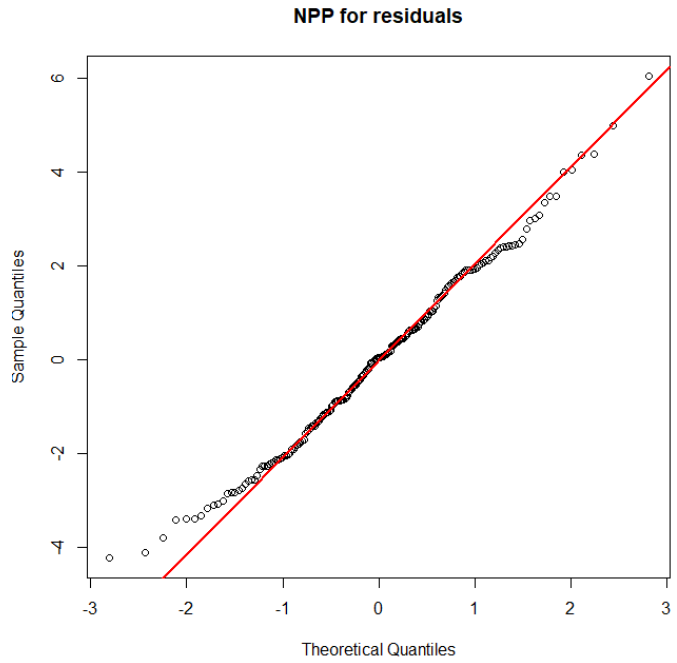
```
> shapiro.test(fit$residuals)

        Shapiro-Wilk normality test

data:  fit$residuals
W = 0.99375, p-value = 0.5641

>
> #install.packages("nortest")
> library(nortest)
> lillie.test(fit$residuals)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  fit$residuals
D = 0.036828, p-value = 0.7306
```

No violation of normality from the residuals.

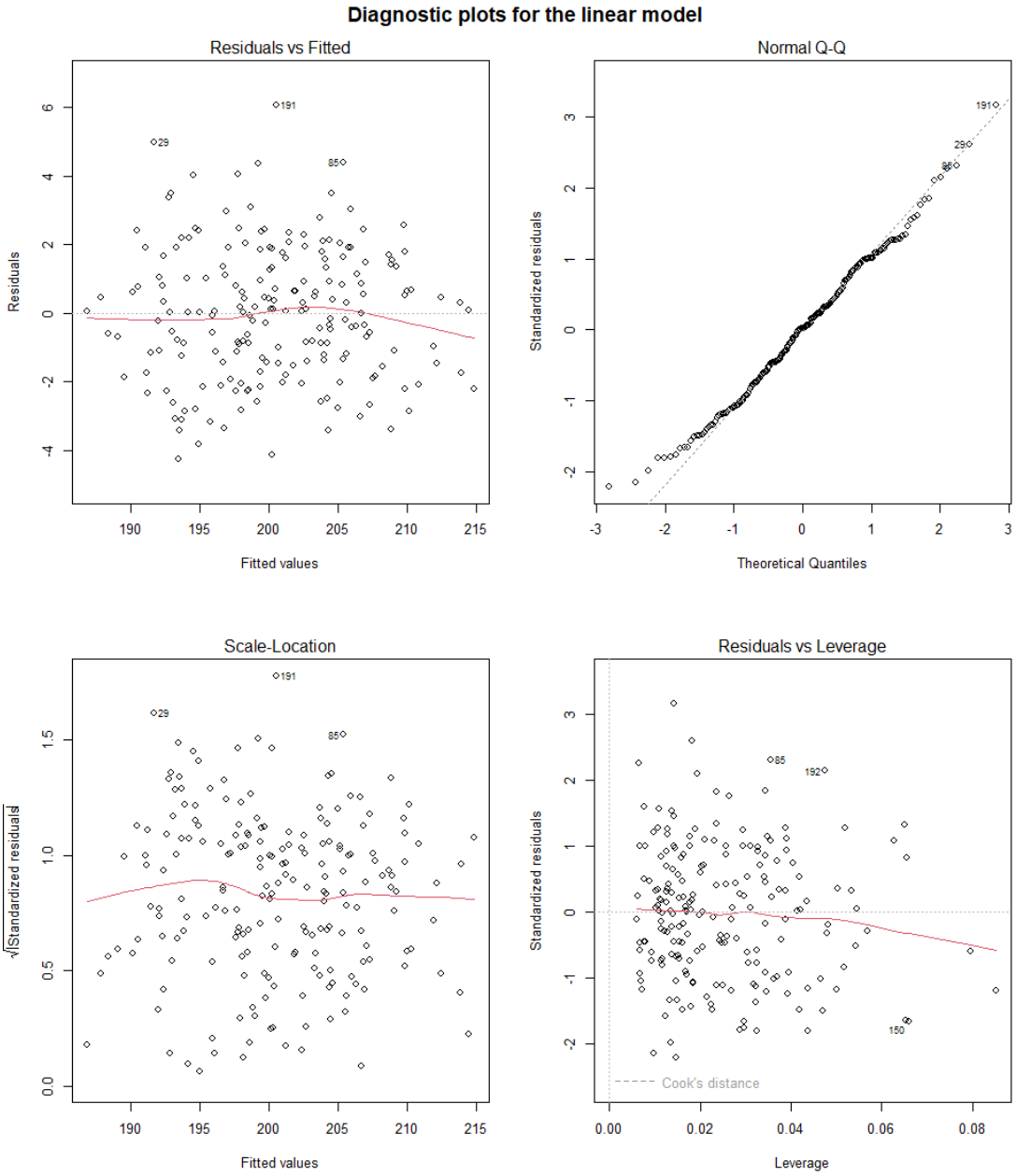


f) For step wise I chose forward selection.

```
> forward$anova
  Step Df   Deviance Resid. Df Resid. Dev      AIC
1      NA         NA       199  7879.7378 736.7465
2 + X4 -1 6982.05740       198   897.6804 304.2994
3 + X1 -1   11.57984       197   886.1006 303.7026
4 + X2 -1  145.69640       196   740.4042 269.7758
5 + X3 -1   10.96424       195   729.4400 268.7919
>
> summary(forward)

Call:
lm(formula = Y ~ X4 + X1 + X2 + X3, data = dffs)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2411 -1.4183  0.0376  1.3783  6.0649

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.82622    4.60897   3.868  0.00015 ***
X4           0.06437    0.28426   0.226  0.82109
X1           0.96631    0.14497   6.666 2.64e-10 ***
X2           1.94436    0.30075   6.465 7.94e-10 ***
X3           0.23677    0.13829   1.712  0.08848 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.934 on 195 degrees of freedom
Multiple R-squared:  0.9074,    Adjusted R-squared:  0.9055
F-statistic: 477.9 on 4 and 195 DF,  p-value: < 2.2e-16
```




```
> shapiro.test(forward$residuals)

        Shapiro-Wilk normality test

data:  forward$residuals
W = 0.99187, p-value = 0.3289

>
> lillie.test(forward$residuals)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  forward$residuals
D = 0.035262, p-value = 0.788
```



g)

```
> prediction.lm = lm(Y ~ X1*X2*X3*X4)
>
> test = data.frame (
+   X1 = 120,
+   X2 = 30,
+   X3 = 10,
+   X4 = 90
+ )
>
> predict(prediction.lm, test, interval="predict")
       fit     lwr      upr
1 200.2812 196.511 204.0515
```

```
> test = data.frame (
+   X1 = 120,
+   X2 = 30,
+   X3 = 10,
+   X4 = 90,
+   W = 2
+ )
>
> predict(prediction.lm, test, interval="predict")
       fit      lwr      upr
1 200.1521 196.4887 203.8155
```

h)

```
   Q1 Q2 Q3 Q4
A 21 14 21 20
B  9 19 21 18
C 20 17  8 12
```

```
    A  B  C
Q1 21  9 20
Q2 14 19 17
Q3 21 21  8
Q4 20 18 12
```

```
> fit = aov(y~z)
> summary(fit)
             Df Sum Sq Mean Sq F value Pr(>F)
z             3  112.7   37.56     398 <2e-16 ***
Residuals   196   18.5    0.09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> TukeyHSD(fit)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = y ~ z)

$z
    diff       lwr       upr p adj
2-1 0.48 0.3207828 0.6392172     0
3-1 1.14 0.9807828 1.2992172     0
4-1 2.00 1.8407828 2.1592172     0
3-2 0.66 0.5007828 0.8192172     0
4-2 1.52 1.3607828 1.6792172     0
4-3 0.86 0.7007828 1.0192172     0
```