SS 9859

Assignment #3

Narges Goudarzi

250993028

Question 1

A) TRUE. When we add predictors to a model, the value of R-squared never decreases. It is even possible that by adding useless predictors R-squared increase, and in fact it is one of the problems with coefficient of determination.

B) TRUE) Multicollinearity is a negative concept that happens when some variables in a regression model are correlated. This correlation can cause a problem because independent variables must be independent. If there is a correlation between independent variables, variance of the coefficient estimates can go up and make the estimates too sensitive to small changes in the regression model. In other wors, we may make a wrong conclusion about estimated parameters.

C) FALSE) We know $VIF_j = \frac{1}{1-R_j^2}$ and $R_j^2 = R^2$ from the regression of x_j on the other predictors.

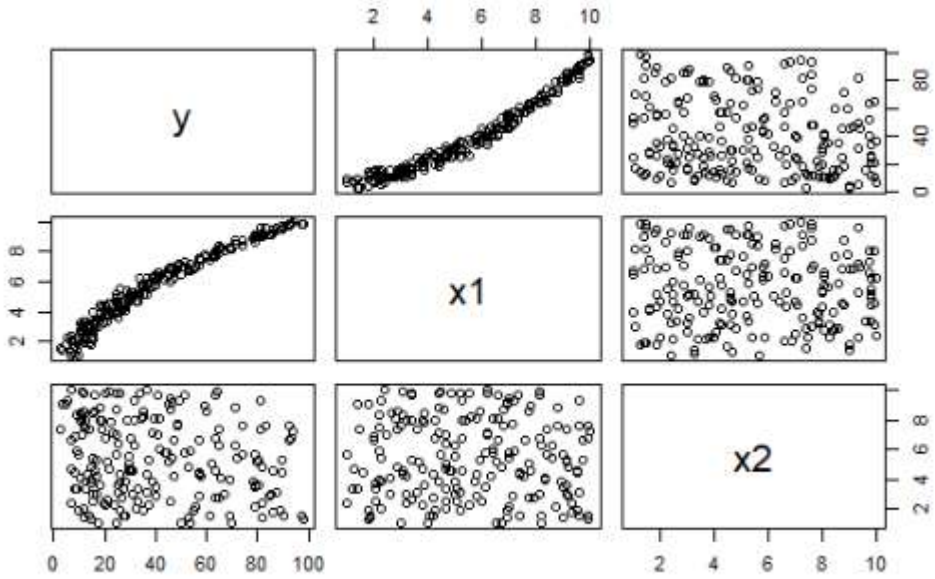
D) FALSE) An influential point usually have high leverage but a high leverage value is not always an influential point. A point is influential if both leverage and residual are high. In other words,The influence depend on both leverage and outlier. it is possible that a data point with a high leverage has small residual and finally it is not an influential point.

E) FALSE) it is not always true as these methods are different and the final results are depend on some properties of the variables. You may get different selection of predictors by applying different criteria.

2)

A)

```
pairs( y ~ x1+ x2, data = hw3_data,main="Scatterplot Matrix")
```



There is a nonlinear relationship between y and x1 .Pearson correlation is closed to +1 .It shows that maybe we don't have some assumption of regression model. Also there is a negative relationship between y and x2. There is no significant relationship between x1 and x2.

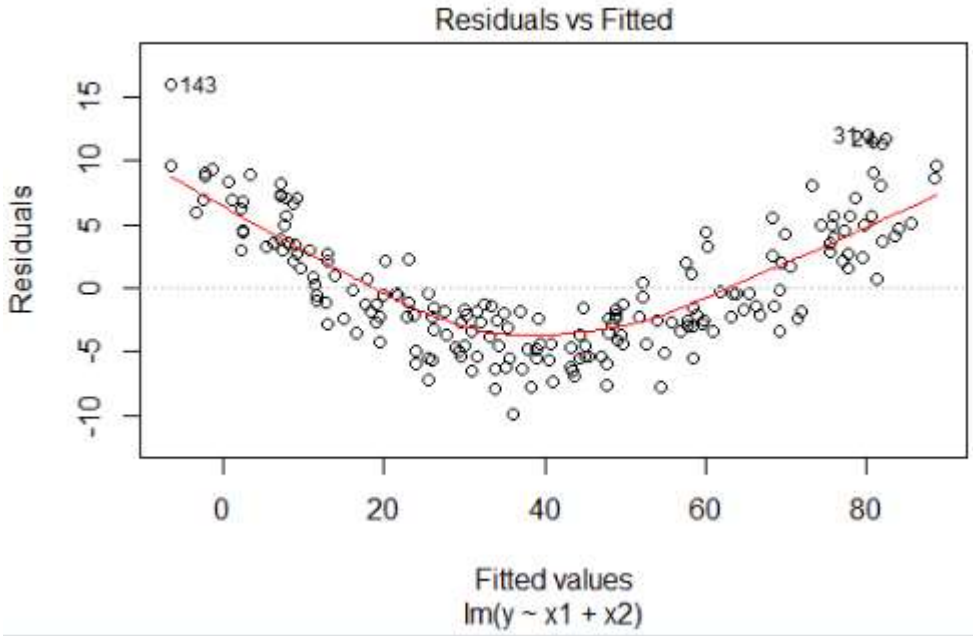
B)

```
model1= lm(y ~ x1+x2,data=hw3_data)
summary(model1)
plot(model1)
```

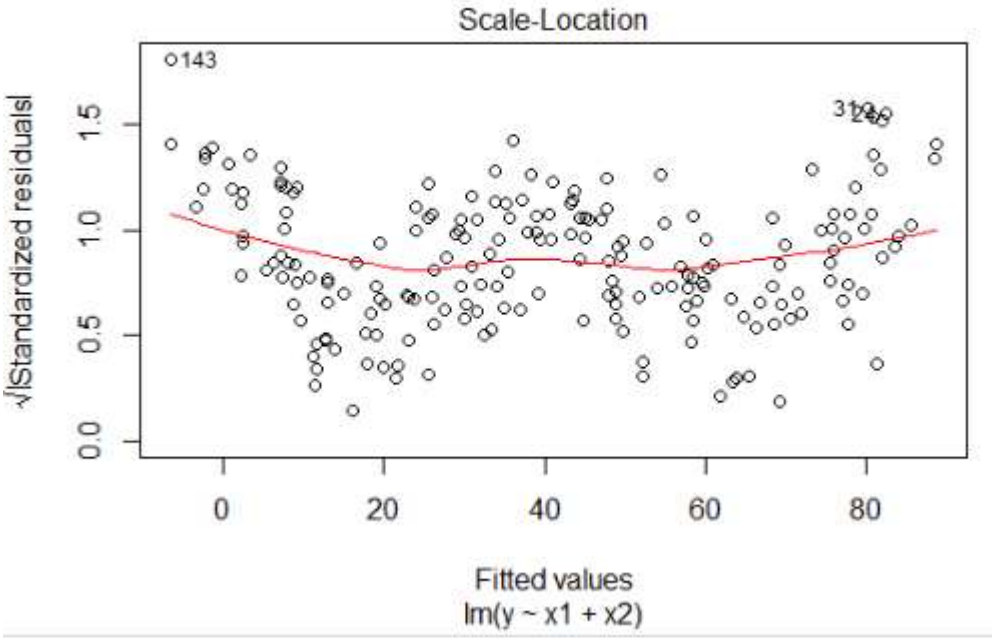
```
Call:
lm(formula = y ~ x1 + x2, data = hw3_data)

Residuals:
   Min     1Q Median     3Q    Max
-9.963 -3.503 -1.347  3.473 15.919
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -9.5112     1.1359  -8.374 1.03e-14 ***
x1           10.0947     0.1402  71.983  < 2e-16 ***
x2           -1.2387     0.1309  -9.461  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.927 on 197 degrees of freedom
Multiple R-squared:  0.9646,   Adjusted R-squared:  0.9642
F-statistic:  2681 on 2 and 197 DF,  p-value: < 2.2e-16
```

$$\hat{Y} = -9.5112 + 10.0947 X_1 - 1.2387 X_2$$

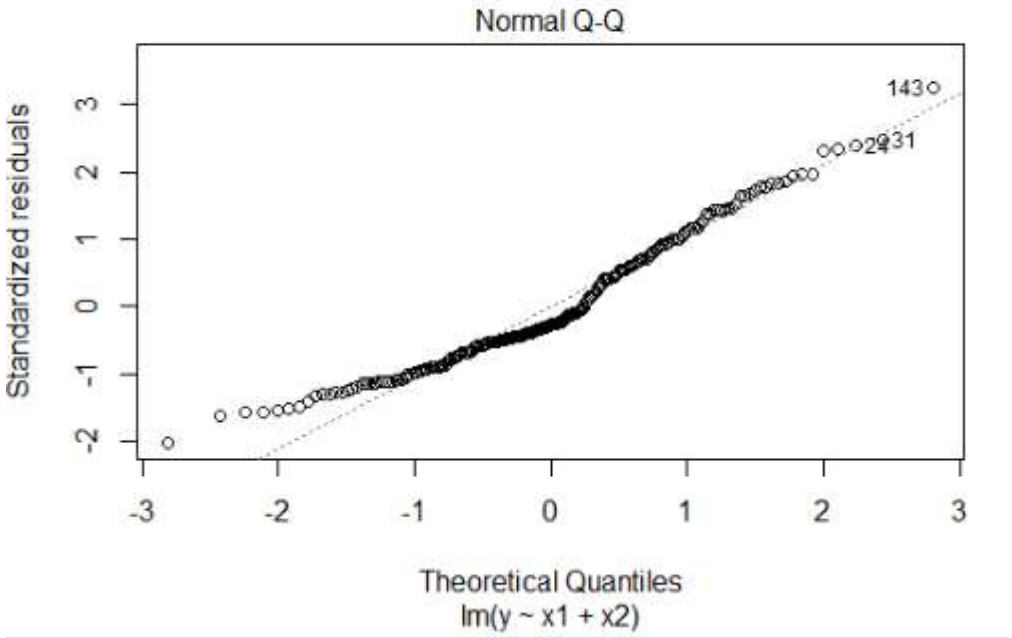


Residuals vs Fitted: Used to check the linear relationship assumptions. A horizontal line, without distinct patterns is an indication for a linear relationship. In this example we can see a pattern which it shows we **could not accept** linearity assumption.




```
 bptest(model1)
studentized Breusch-Pagan test

data:  model1
BP = 0.094601, df = 2, p-value = 0.9538
```

Equality of variance: for each value of X, the range of residuals has the same value. This means that the level of error in the model is approximately the same regardless of the value of the predictor variable. Also, we can use Breusch-Pagan Test that it also proves the assumption of equality of variances is accepted.




```
shapiro.test(residuals(model1))
```

```
Shapiro-Wilk normality test

data:  residuals(model1)
W = 0.95915, p-value = 1.603e-05
```

The QQ plot of residuals can be used to check the normality assumption. The normal probability plot of residuals should approximately follow a straight line. In this case, the pattern is non-linear, as a result, we can reject normality assumption. However, we can double check by using Shapiro test that it shows we can reject normality assumption. Therefore, **Normality assumption is rejected.**

C)

```
inf_i = which(cooks.distance(model1) > 4 / length(cooks.distance(model1)))
inf_i
```

```
  6  18  24  31  35  51  74  87 111 126 128 139 143 193
```

D)

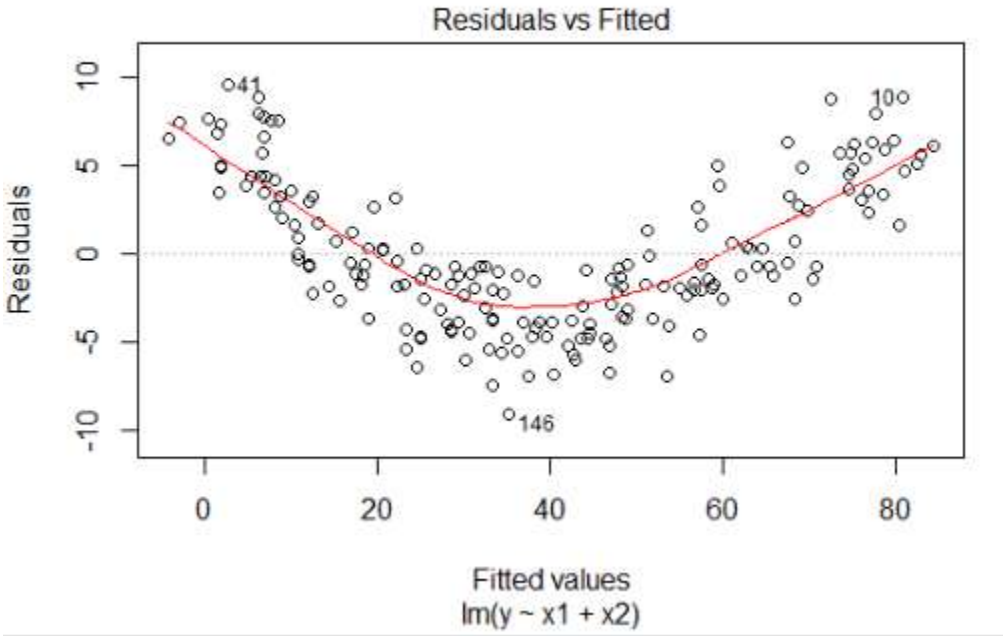
```
rstandard(model1) #standardized residuals
abs(rstandard(model1)) > 2 # Last observation is an outlier
out_i = which(abs(rstandard(model1)) > 2)
rstandard(model1)[out_i]
```

```
       24        31       139       143       159       193
 2.403397  2.471281  2.306463  3.267128 -2.029343  2.356038
```

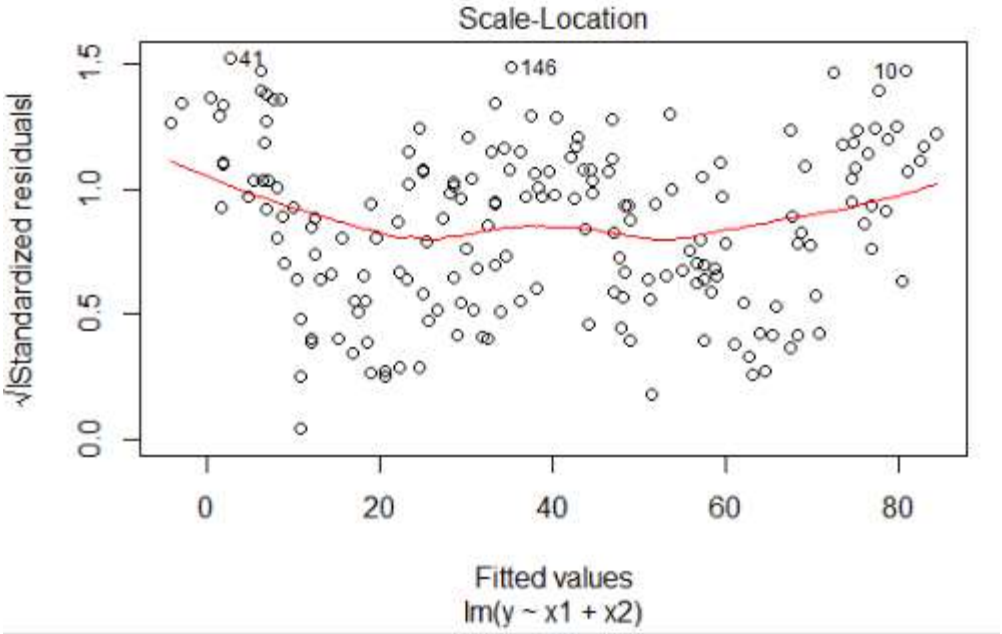
The result is 24, 31,139,143,193.It means among the influential points,5 of them are also considered outliers.

E)

```
hw3_data1<-hw3_data[-c(inf_i),]
model2= lm(y ~ x1+x2,data=hw3_data1)
Call:
lm(formula = y ~ x1 + x2, data = hw3_data1)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1238 -3.1455 -0.7818  3.2732  9.5680

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.4842     1.0544  -9.944   <2e-16 ***
x1           10.0788     0.1332  75.662   <2e-16 ***
x2           -1.1841     0.1146 -10.329   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.161 on 183 degrees of freedom
Multiple R-squared:  0.9704,   Adjusted R-squared:  0.9701
F-statistic:  3003 on 2 and 183 DF,  p-value: < 2.2e-16
```



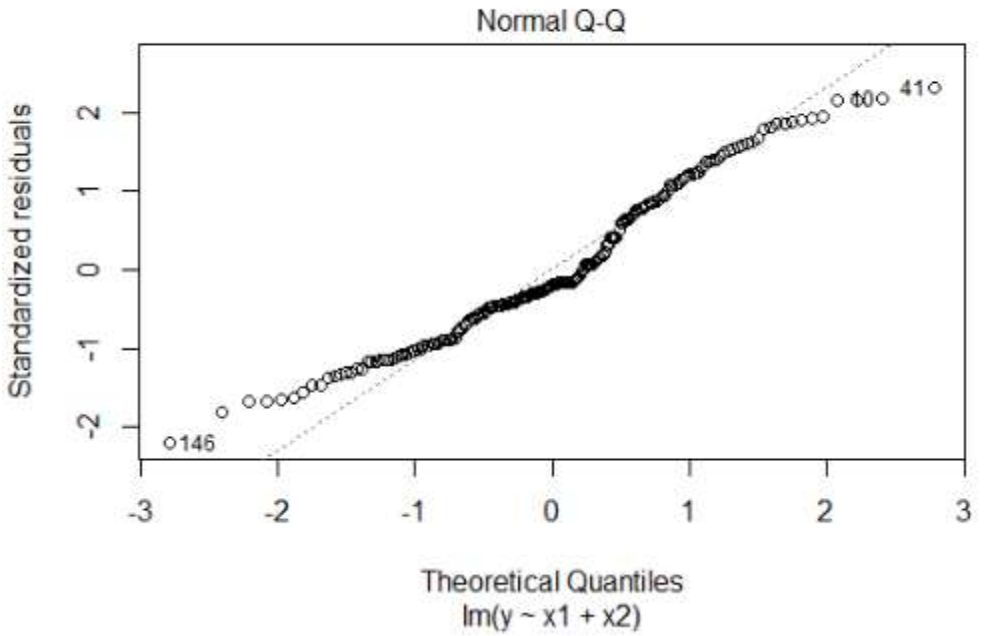
Residuals vs Fitted: A horizontal line, without distinct patterns is an indication for a linear relationship. In this example we can see a pattern that shows we can’t accept linearity assumption.




```
bptest(model2)
```

```
        studentized Breusch-Pagan test
data:  model2
BP = 0.78179, df = 2, p-value = 0.6764
```

Equality of variance: Regarding to above plots for each value of X, the residuals has the same range. This means that the level of error in the model is approximately the same regardless of the value of the predictor variable. Also, we can use Breusch-Pagan Test that **it also proves the assumption of equality of variances.**




```
shapiro.test(residuals(model2))
```

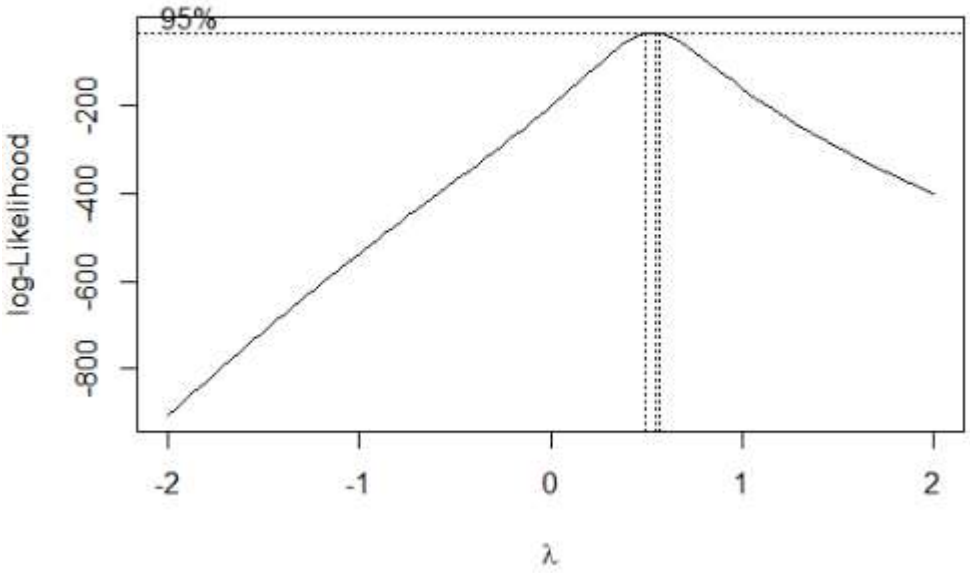
```
Shapiro-Wilk normality test
data:  residuals(model2)
W = 0.96638, p-value = 0.0001911
```

The QQ plot of residuals can be used to check the normality assumption. The normal probability plot of residuals should approximately follow a straight line. In this problem, the pattern is non-linear, so plot give evidence against normality assumption. It means we reject normality assumption. However, we can double check by using Shapiro test that it shows we can reject normality assumption. Therefore, **Normality assumption is rejected.**

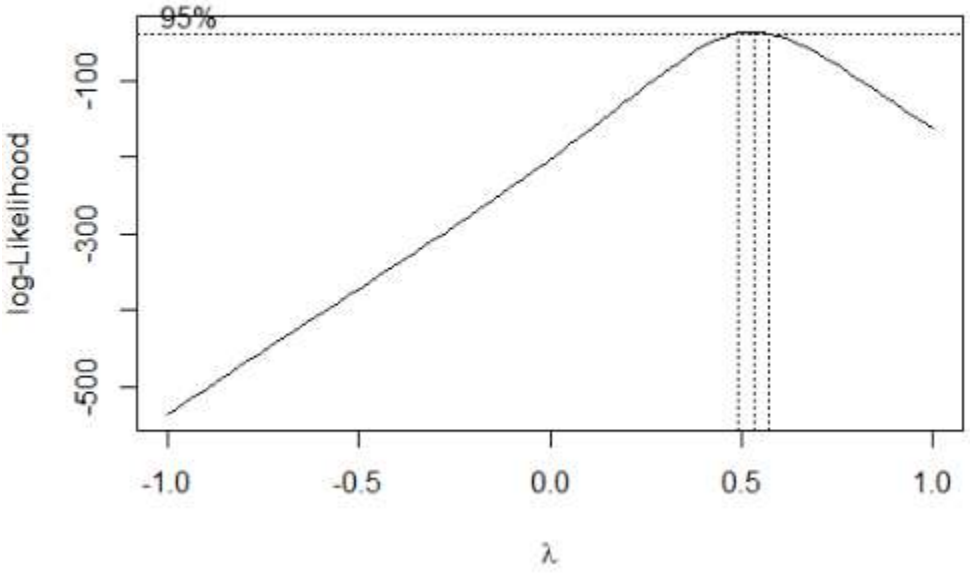
The removal of the influential points is not useful for correcting the model assumptions. It means we can't accept normality and linearity assumption.

F)

```
par(mfrow=c(1,1))
 boxcox(model1)
```

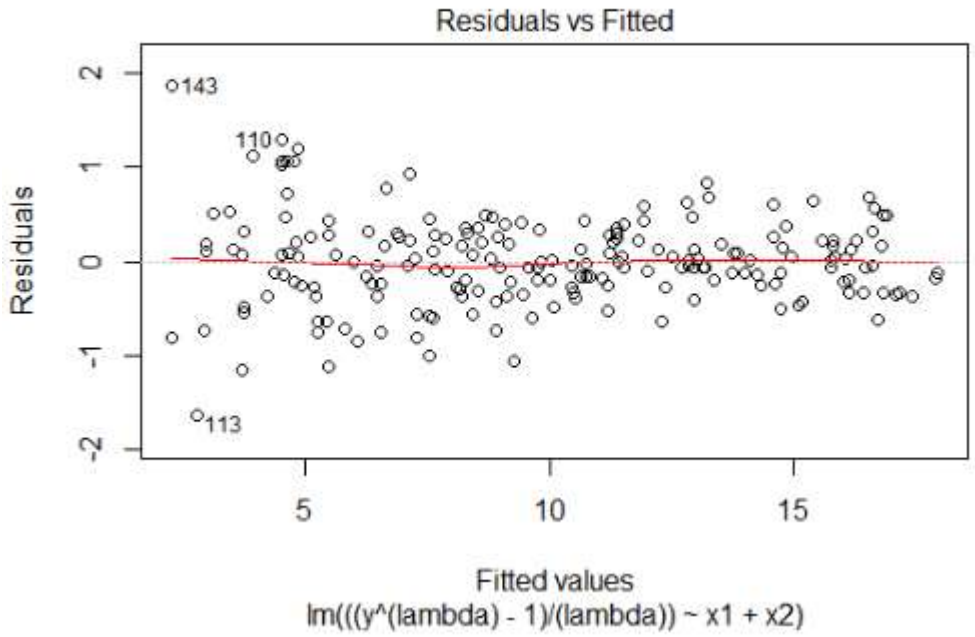



```
 boxcox(model1, lambda = seq(-1, 1, by = 0.05))
```

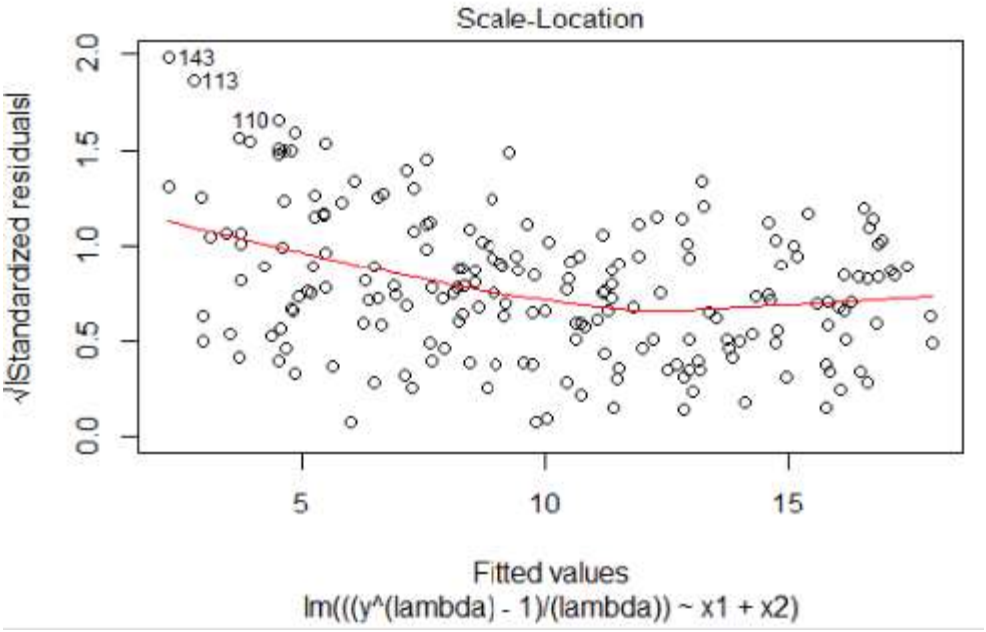


```
lambda = 0.5
model1_transf <- lm(((y^(lambda)-1)/(lambda))~x1+x2,data=hw3_data)
summary(model1_transf)

Call:
lm(formula = ((y^(lambda) - 1)/(lambda)) ~ x1 + x2, data = hw3_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.63948 -0.26509 -0.00651  0.26212  1.86093

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.78737    0.10981   16.28   <2e-16 ***
x1           1.65931    0.01356  122.40   <2e-16 ***
x2          -0.20386    0.01266  -16.11   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4763 on 197 degrees of freedom
Multiple R-squared:  0.9875,   Adjusted R-squared:  0.9873
F-statistic:  7751 on 2 and 197 DF,  p-value: < 2.2e-16
```

```
plot(model1_transf)
```



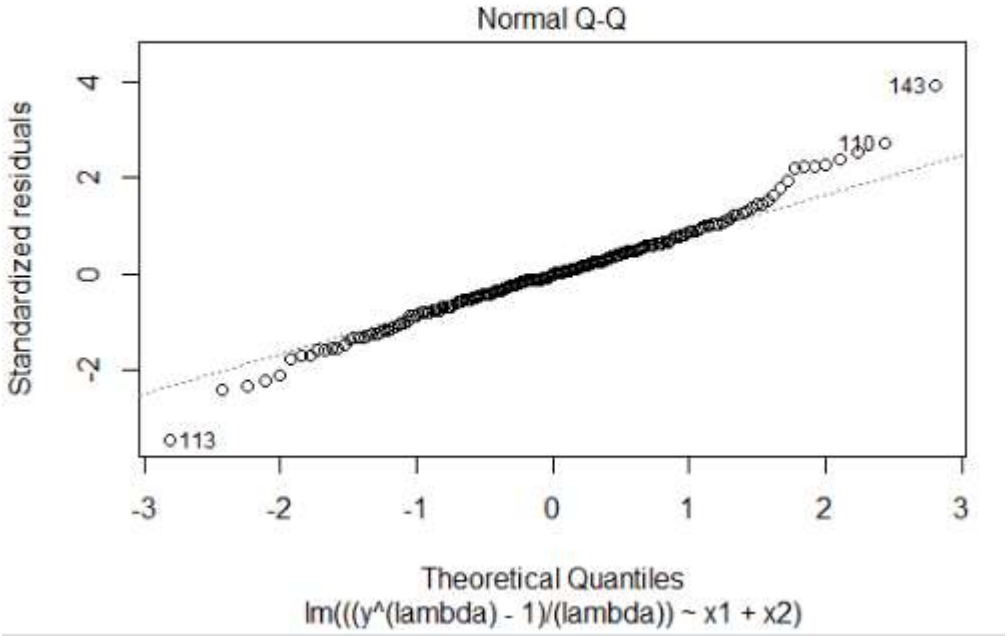
Residuals vs Fitted A horizontal line, without distinct patterns is an indication for a linear relationship. In this example mean of error is zero everywhere. **we can accept linearity assumption.**




```
bptest(model1_transf)
```

```
        studentized Breusch-Pagan test
data:  model1_transf
BP = 26.212, df = 2, p-value = 2.033e-06
```

Equality of variance: the distribution of residuals does not have the same range over different values of X. Also, we can use Breusch-Pagan which also **proves the assumption of equality of variances is rejected.**




```
shapiro.test(residuals(model1_transf))
```

```
        Shapiro-Wilk normality test
data:  residuals(model1_transf)
W = 0.9816, p-value = 0.01006
```

The QQ plot of residuals can be used to check the normality assumption. The normal probability plot of residuals should approximately follow a straight line. In this problem, the pattern is non-linear, so plot give evidence against assume normality. It means we reject normality assumption. However, we can double check by using Shapiro test that it shows we reject normality assumption. Therefore, **Normality assumption is rejected.**

This transformation is not helpful for correcting the model assumptions because Normality assumption is rejected. It helps only for linearity assumption. The assumption of equality of variances is rejected.

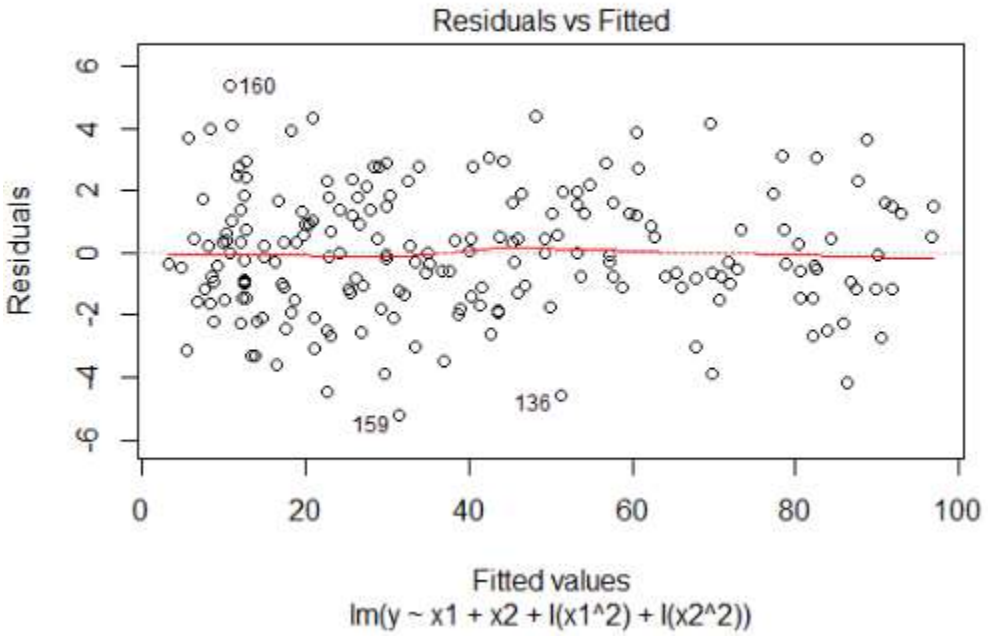
g)

```
lm_1 = lm(y ~ x1 +x2+ I(x1^2) + I(x2^2), data = hw3_data)
 summary(lm_1)
```

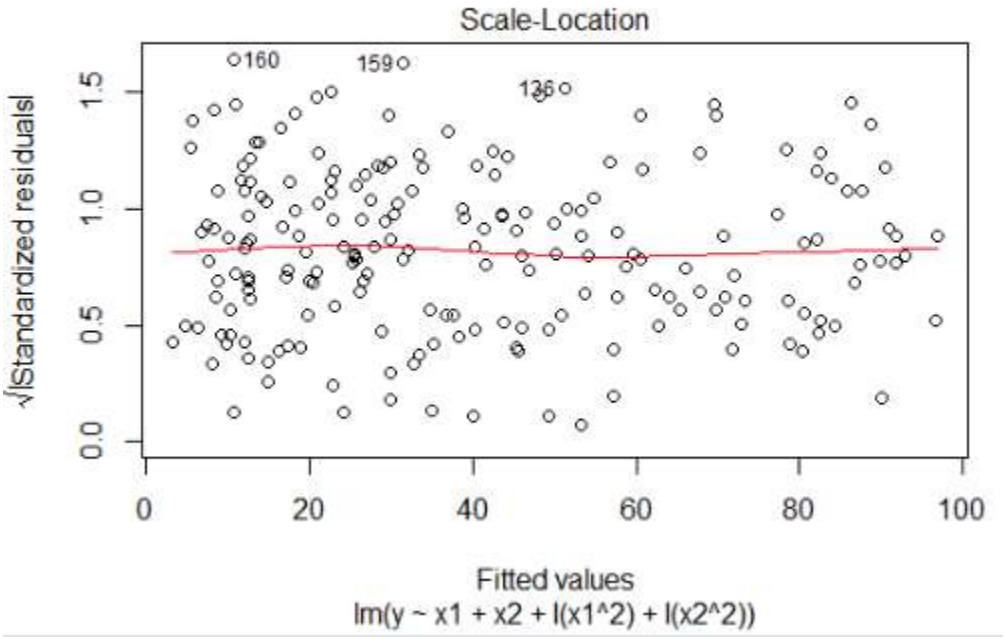
```
Call:
lm(formula = y ~ x1 + x2 + I(x1^2) + I(x2^2), data = hw3_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.2370 -1.2533 -0.0942  1.3701  5.3505
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.65216    0.93122   9.291  < 2e-16 ***
x1           1.30413    0.28367   4.597 7.68e-06 ***
x2          -0.72887    0.25617  -2.845  0.00491 **
I(x1^2)      0.77857    0.02463  31.614  < 2e-16 ***
I(x2^2)     -0.02560    0.02259  -1.133  0.25854
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.995 on 195 degrees of freedom
Multiple R-squared:  0.9942,   Adjusted R-squared:  0.9941
F-statistic:  8422 on 4 and 195 DF,  p-value: < 2.2e-16
```

```
plot(lm_1)
```



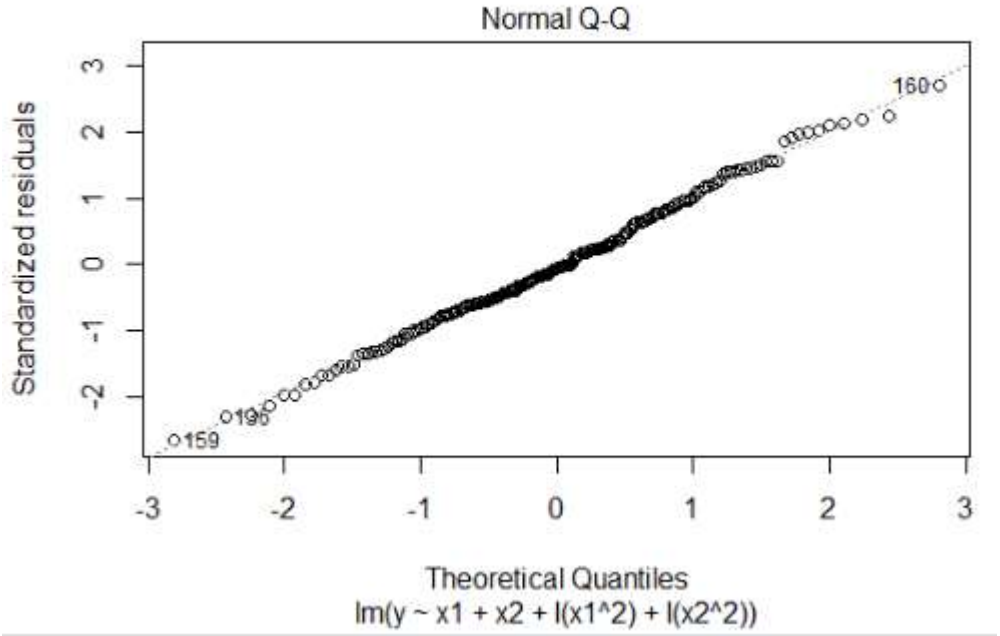
Residuals vs Fitted: A horizontal line, without distinct patterns is an indication for a linear relationship. In other words, mean of error is zero everywhere. In this example **we can accept linearity assumption.**




```
bptest(lm_1)
```

```
        studentized Breusch-Pagan test
data:  lm_1
BP = 2.6009, df = 4, p-value = 0.6267
```

Equality of variance: Regarding to above plots for each value of X, the residuals has the same range over different values of X. This means that the level of error in the model is approximately the same regardless of the value of the predictor variable. Also, we can use Breusch-Pagan Test which also **proves the assumption of equality of variances is accepted.**




```
shapiro.test(residuals(lm_1))
```

```
        Shapiro-Wilk normality test
data:  residuals(lm_1)
W = 0.9956, p-value = 0.8331
```

The QQ plot of residuals can be used to check the normality assumption. The normal probability plot of residuals should approximately follow a straight line. In this problem, all the points fall approximately along this reference line, so we can assume normality. However, we can double check by using Shapiro test that it shows we fail to reject normality assumption. **Therefore, Normality assumption is accepted.**

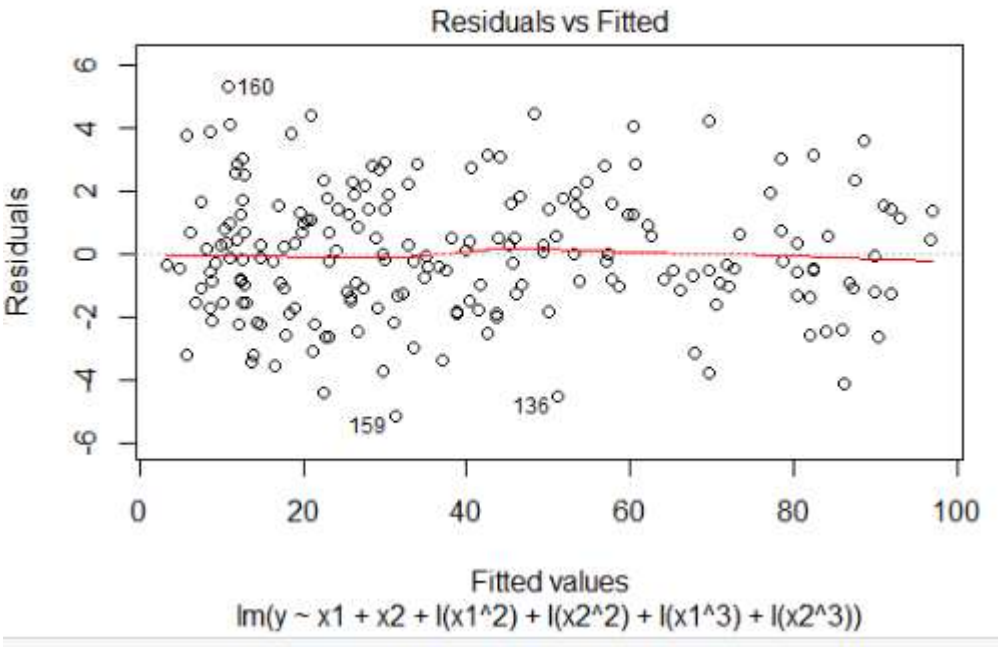
This polynomial model preferable to the resulting models in (b) and (f) because all assumption is accepted.

h)

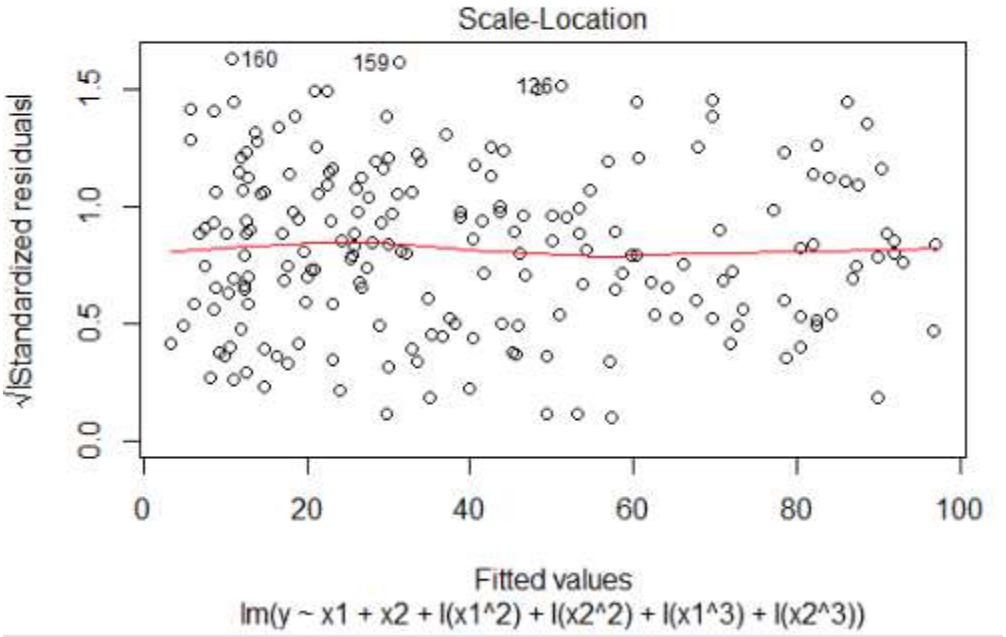
```
lm_2 = lm(y ~ x1 +x2+ I(x1^2) + I(x2^2)+ I(x1^3) + I(x2^3), data = hw3_data)
 summary(lm_2)
```

```
Call:
lm(formula = y ~ x1 + x2 + I(x1^2) + I(x2^2) + I(x1^3) + I(x2^3),
    data = hw3_data)

Residuals:
   Min     1Q Median     3Q    Max
-5.166 -1.281 -0.122  1.359  5.273

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.065491   1.810742   5.007 1.25e-06 ***
x1           1.420580   0.929225   1.529    0.128
x2          -1.182477   0.801651  -1.475    0.142
I(x1^2)      0.755965   0.182125   4.151 4.97e-05 ***
I(x2^2)      0.069683   0.161015   0.433    0.666
I(x1^3)      0.001279   0.010753   0.119    0.905
I(x2^3)     -0.005755   0.009623  -0.598    0.551
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.004 on 193 degrees of freedom
Multiple R-squared:  0.9943,   Adjusted R-squared:  0.9941
F-statistic:  5568 on 6 and 193 DF,  p-value: < 2.2e-16
```



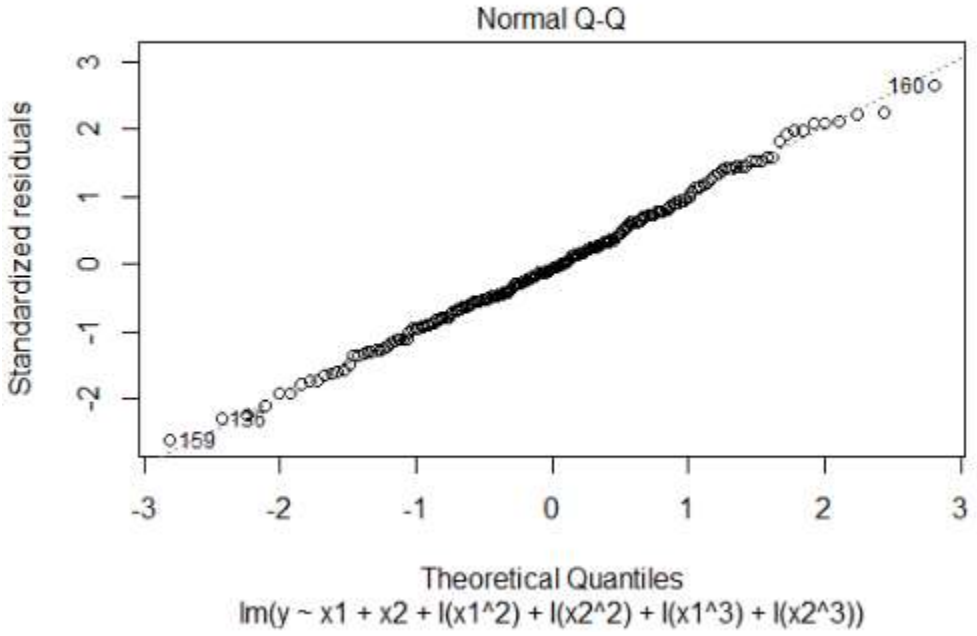
Residuals vs Fitted: Used to check the linear relationship assumptions. A horizontal line, without distinct patterns is an indication for a linear relationship. In other words, the mean of error is zero over different values of X.In this example **we can accept linearity assumption.**




```
bptest(lm_2)
```

```
        studentized Breusch-Pagan test
data:  lm_2
BP = 4.2839, df = 6, p-value = 0.6383
```

Equality of variance: Regarding to above plots for each value of X, the distribution of residuals has the same variance. This means that the level of error in the model is approximately the same regardless of the value of the predictor variable. Also, we can use Breusch-Pagan Test which also **proves the assumption of equality of variances is accepted.**




```
shapiro.test(residuals(lm_2))
```

```
        Shapiro-Wilk normality test
data:  residuals(lm_2)
W = 0.99579, p-value = 0.8581
```

The QQ plot of residuals can be used to check the normality assumption. The normal probability plot of residuals should approximately follow a straight line. In this problem, all the points fall approximately along this reference line, so we can assume normality. However, we can double check by using Shapiro test that it shows we fail to reject normality assumption. **Therefore, Normality assumption is accepted.**

The quadratic model be preferred to the cubic one. P-value and R-square are the same in both models but in cubic model there are two more coefficients which is a negative point in regression models. By the way, both of coefficient related to cubic term are insignificant. It means they are not important in the model.

Question 3

A)

```
model_a = lm(mpg ~ cyl+disp+ hp+wt+drat, data = mtcars)

summary(model_a)
```

```
Call:
lm(formula = mpg ~ cyl + disp + hp + wt + drat, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7014 -1.6850 -0.4226  1.1681  5.7263

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.00836    7.57144   4.756  6.4e-05 ***
cyl         -1.10749    0.71588  -1.547  0.13394
disp         0.01236    0.01190   1.039  0.30845
hp          -0.02402    0.01328  -1.809  0.08208 .
wt          -3.67329    1.05900  -3.469  0.00184 **
drat         0.95221    1.39085   0.685  0.49964
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.538 on 26 degrees of freedom
Multiple R-squared:  0.8513,   Adjusted R-squared:  0.8227
F-statistic: 29.77 on 5 and 26 DF,  p-value: 5.618e-10
```

```
vif(model_a)
```

```
      cyl      disp        hp        wt      drat
 7.869010 10.463957  3.990380  5.168795  2.662298
```

For disp VIF is larger than 10. It means there is collinearity.Also,
The p-value for each predictor is high. However, the p-value for the significance of regression test is low. This happens because there is collinerity. Collinearity affects in the regression analysis. In other words, the variance of disp is inflated. Variance of the coefficient estimates can go up and make the estimates too sensitive to small changes in the regression model. It can lead to unstable coefficient estimates and as a result the interpretation of regression model can be difficult.

B)

```
model_a1 = lm(mpg ~ cyl+hp+wt+drat, data = mtcars)
summary(model_a1)
```

```
Call:
lm(formula = mpg ~ cyl + hp + wt + drat, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6171 -1.5663 -0.6058  1.2612  5.8161
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.49588    7.44101   4.636  8.1e-05 ***
cyl         -0.76229    0.63502  -1.200  0.24040
hp          -0.02089    0.01295  -1.613  0.11845
wt          -2.97331    0.81818  -3.634  0.00116 **
drat         0.81771    1.38684   0.590  0.56034
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.541 on 27 degrees of freedom
Multiple R-squared:  0.8451,   Adjusted R-squared:  0.8222
F-statistic: 36.84 on 4 and 27 DF,  p-value: 1.438e-10
```

 # we want to obtain VIF without any built-in function in R

```
rsquare_cyl<-summary(lm(cyl~hp+wt+drat, mtcars))$r.squared
 vif_cyl<- (1/(1-rsquare_cyl))
 vif_cyl
[1] 6.17356

rsquare_hp<-summary(lm(hp~cyl+wt+drat, mtcars))$r.squared
>  vif_hp<- (1/(1-rsquare_hp))
>  vif_hp
[1] 3.78467

rsquare_wt<-summary(lm(wt~cyl+hp+drat, mtcars))$r.squared
> vif_wt<- (1/(1-rsquare_wt))
> vif_wt
[1] 3.076225

rsquare_drat<-summary(lm(drat~cyl+hp+wt, mtcars))$r.squared
> vif_drat<- (1/(1-rsquare_drat))
> vif_drat
[1] 2.639229
```

There is not any VIF higher than 10. It means there is not collinearity, however, in some books if VIF is higher than 5 shows there is collinearity.

C)

```
fit_null = lm(mpg ~ cyl+disp+ hp+wt+drat, data = mtcars)
fit_forw_aic = step(fit_null,
       scope = mpg ~ cyl+disp+ hp+wt+drat,
        direction = "forward")
Start:  AIC=64.95
mpg ~ cyl + disp + hp + wt + drat


fit_step_aic = step(fit_null,
+                     scope = mpg ~ cyl+disp+ hp+wt+drat,
+                     direction = "both")
Start:  AIC=64.95
mpg ~ cyl + disp + hp + wt + drat

       Df Sum of Sq    RSS    AIC
- drat  1     3.018 170.44 63.526
- disp  1     6.949 174.38 64.255
```

```
<none>              167.43 64.954
- cyl   1    15.411 182.84 65.772
- hp    1    21.066 188.49 66.746
- wt    1    77.476 244.90 75.124

Step:  AIC=63.53
mpg ~ cyl + disp + hp + wt

       Df Sum of Sq    RSS    AIC
- disp  1     6.176 176.62 62.665
<none>              170.44 63.526
- hp    1    18.048 188.49 64.746
+ drat  1     3.018 167.43 64.954
- cyl   1    24.546 194.99 65.831
- wt    1    90.925 261.37 75.206

Step:  AIC=62.66
mpg ~ cyl + hp + wt

       Df Sum of Sq    RSS    AIC
<none>              176.62 62.665
- hp    1    14.551 191.17 63.198
+ disp  1     6.176 170.44 63.526
- cyl   1    18.427 195.05 63.840
+ drat  1     2.245 174.38 64.255
- wt    1   115.354 291.98 76.750
```

Based on AIC, the stepwise selection approach introduces the best subset of predictors to predict mpg. mpg ~ cyl + hp + wt because The AIC is the smallest.

D)

```
n = nrow(mtcars)
  fit_back_bic = step(fit_null, direction = "backward", k=log(n))
Start:  AIC=73.75
mpg ~ cyl + disp + hp + wt + drat

       Df Sum of Sq    RSS    AIC
- drat  1     3.018 170.44 70.854
- disp  1     6.949 174.38 71.584
- cyl   1    15.411 182.84 73.100
<none>              167.43 73.748
- hp    1    21.066 188.49 74.075
- wt    1    77.476 244.90 82.453

Step:  AIC=70.85
mpg ~ cyl + disp + hp + wt

       Df Sum of Sq    RSS    AIC
- disp  1     6.176 176.62 68.528
- hp    1    18.048 188.49 70.609
<none>              170.44 70.854
- cyl   1    24.546 194.99 71.694
- wt    1    90.925 261.37 81.069

Step:  AIC=68.53
mpg ~ cyl + hp + wt

       Df Sum of Sq    RSS    AIC
```

```
- hp    1    14.551 191.17 67.595
- cyl   1    18.427 195.05 68.237
<none>               176.62 68.528
- wt    1   115.354 291.98 81.147

Step:  AIC=67.6
mpg ~ cyl + wt

       Df Sum of Sq    RSS    AIC
<none>              191.17 67.595
- cyl   1     87.15 278.32 76.149
- wt    1    117.16 308.33 79.426
```

Based on this method, the model is mpg ~ cyl + wt. We want to check "Is the resulting mod el is significantly different from the model obtained in (c)?We need to test $H_0 : \beta_{cyl} = 0$

```
null=lm( mpg ~ cyl + wt,data=mtcars)
full=lm(mpg ~ cyl + wt+hp,data=mtcars)
anova(null,full)
```

```
Analysis of Variance Table

Model 1: mpg ~ cyl + wt
Model 2: mpg ~ cyl + wt + hp
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     29 191.17
2     28 176.62  1    14.551 2.3069   0.14
```

It means we fail to reject null hypothesis. It shows there is not a significant difference between two models.

Question 4

A)

```
model1=lm(lpsa ~lcavol+lweight+svi, data = prostate)
summary(model1)
model2=lm(lpsa ~lcavol+lweight+svi+lbph, data = prostate)
summary(model2)
model3=lm(lpsa ~lcavol+lweight+svi+lbph+lcp+gleason, data = prostate)
summary(model3)
summary(model1)$adj.r.squared
```

```
AIC(model1,model2,model3)
       df      AIC
model1  5 216.5979
model2  6 215.9223
model3  8 218.9735
```

```
BIC(model1,model2,model3)
       df      BIC
model1  5 229.4714
model2  6 231.3705
model3  8 239.5712
```

```
summary(model1)$adj.r.squared
[1] 0.6143899
summary(model2)$adj.r.squared
[1] 0.6208036
summary(model3)$adj.r.squared
[1] 0.6161501
```

Based on AIC the best model is model2 (model B).
Based on BIC the best model is model1 (model A).
Based on adjusted r-squared the best model is model2 (model B).

B)

```
sqrt(sum((resid(model1)/(1-hatvalues(model1)))^2)/97)
[1] 0.7381178
> sqrt(sum((resid(model2)/(1-hatvalues(model2)))^2)/97)
[1] 0.7355329
> sqrt(sum((resid(model3)/(1-hatvalues(model3)))^2)/97)
[1] 0.7458586
```

Based on PRESS the best model is model2 (model B).

C)

```
summary(model1)$r.squared
[1] 0.6264403
> summary(model2)$r.squared
[1] 0.6366035
> summary(model3)$r.squared
[1] 0.6401407
```

1. **Based on r-squared the best model is model3 (model C).** When we add a predictor to a model, the R-squared increases. For model C the number of predictor has been increased and we cannot rely on R-squared. It is mentioned in question 1 part A When we add a predictor to a model, the R-squared never decreases. When we add a variable to our model, the value of its estimated coefficient can either be zero, in which case the proportion of explained variance stays fixed, or can be a non-zero (positive value) value it improves the quality of the fit. For solving this problem adjusted R-square has been defined. In other words, $R^2 = \frac{SSR}{SST}$. We can assume that in a regression model we use p variables and have a certain value of R-square. Now suppose that we add one more variable to the model. The total variance cannot change when we add a variable. The explained variance by the model cannot decrease. It can remain unchanged or it can Increase. Therefore, by adding a new variable R-square cannot decrease.