REGRESSION ASSIGNMENT 2

Narges Goudarzi, 250993028

1)

A)

```
mpg_wt_cyl = lm(mpg ~ wt + cyl, data = mtcars2)
summary(mpg_wt_cyl)
int_4cyl = coef(mpg_wt_cyl)[1]
int_6cyl = coef(mpg_wt_cyl)[1] + coef(mpg_wt_cyl)[3]
int_8cyl = coef(mpg_wt_cyl)[1] + coef(mpg_wt_cyl)[4]
slope_all_cyl = coef(mpg_wt_cyl)[2]
fit=(3*slope_all_cyl+int_6cyl)
fit
```

wt

19.95467

B)

We need to test $H_0 : \beta_{cyl} = 0$

```
null=lm(mpg~wt,data=mtcars2)
full=lm(mpg~wt+cyl,data=mtcars2)
anova(null,full)
```

```
Analysis of Variance Table

Model 1: mpg ~ wt
Model 2: mpg ~ wt + cyl
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     23 200.88
2     21 139.62  2    61.253 4.6063 0.02194 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-value is smaller than 0.05 ➔hypothesis is rejected because➔ the reduced model is rejected, fu ll model is accepted. ➔ cyl should be in the model and it is significant at 0.05.

C)

```
mpg_wt_cyl = lm(mpg ~ wt + cyl + wt:cyl, data = mtcars2)
summary(mpg_wt_cyl)
int_4cyl = coef(mpg_wt_cyl)[1]
int_6cyl = coef(mpg_wt_cyl)[1] + coef(mpg_wt_cyl)[3]
int_8cyl = coef(mpg_wt_cyl)[1] + coef(mpg_wt_cyl)[4]
slope_4cyl = coef(mpg_wt_cyl)[2] # slope 1
slope_6cyl = coef(mpg_wt_cyl)[2] + coef(mpg_wt_cyl)[5] # slope 2
```

```
slope_8cyl = coef(mpg_wt_cyl)[2] + coef(mpg_wt_cyl)[6] # slope 3
fit1=(int_8cyl+3*slope_8cyl)
fit1
```

```
Call:
lm(formula = mpg ~ wt + cyl + wt:cyl, data = mtcars2)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.6507 -1.1242 -0.5088  1.4086  5.2918 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  38.6787     3.7624  10.280 3.37e-09 ***
wt           -5.4880     1.5419  -3.559  0.00209 ** 
cyl6         -4.3800    16.9168  -0.259  0.79849    
cyl8        -16.2269     5.7241  -2.835  0.01059 *  
wt:cyl6       0.8649     5.2116   0.166  0.86995    
wt:cyl8       3.7042     1.8856   1.964  0.06427 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.466 on 19 degrees of freedom
Multiple R-squared:  0.8452,	Adjusted R-squared:  0.8045 
F-statistic: 20.75 on 5 and 19 DF,  p-value: 4.241e-07
(Intercept) 

  17.10022
```

D)

Null hypothesis: "There is no significant interaction effect between two predictors."
In the part (C) P-value for $\beta_{wt:cyl6}$ and $\beta_{wt:cyl8}$ is larger than 0.05 ➔ there not significant interaction effect between two predictors.

By the way we can test $H_0 : \beta_{wt:cyl6} = \beta_{wt:cyl8} = 0$ as follows

```
nullmpg_wt_cyl = lm(mpg ~ wt + cyl, data = mtcars2)
fullmpg_wt_cyl = lm(mpg ~ wt + cyl + wt:cyl, data = mtcars2)
anova(nullmpg_wt_cyl,fullmpg_wt_cyl)
```

```
Analysis of Variance Table

Model 1: mpg ~ wt + cyl
Model 2: mpg ~ wt + cyl + wt:cyl
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
```

1	21	139.62			
2	19	115.54	2	24.086	1.9804 0.1655

we fail to reject null hypothesis is rejected ➔ the reduced model can be accepted ➔there is no significant interaction effect between two predictors.

2-

A)

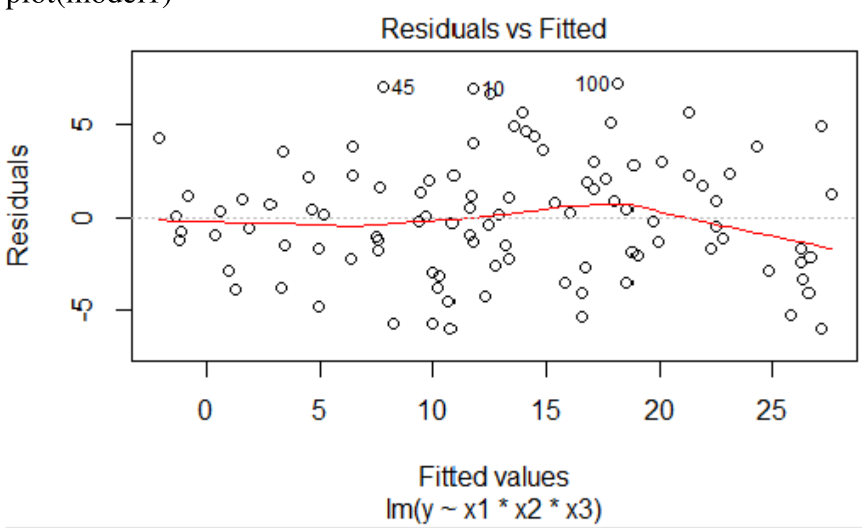
model1 = lm(y ~ x1*x2*x3 , data = hw2_data_1)
summary(model1)

$$E(y \mid x_2 = 50, x_3 = 7) = 7.327393 + 1.709184x_1 \text{-} 0.166497x_2 + 0.561826x_3$$
$$+0.038134x_1x_2 + 0.1217x_1x_3 - 0.003239x_2x_3 - 0.001350x_1x_2x_3 =$$
$$7.327393 + 1.709184x_1 \text{-} 8.32485 + 3.932782 + 1.9067x_1 + 0.8519x_1$$
$$-1.13365 - 0.4725x_1 = 1.801675 + 3.995284x_1$$

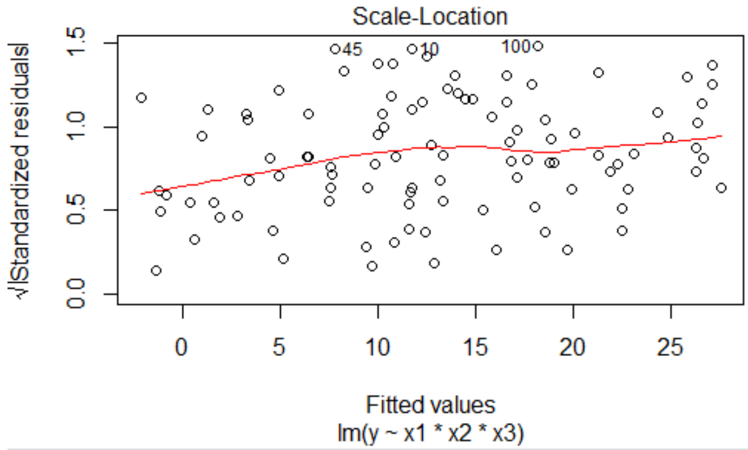
x2 = 50 and x3 = 7 ➔ one unit increase in x_1 increases the estimated mean of y by 3.995284 units.

B)

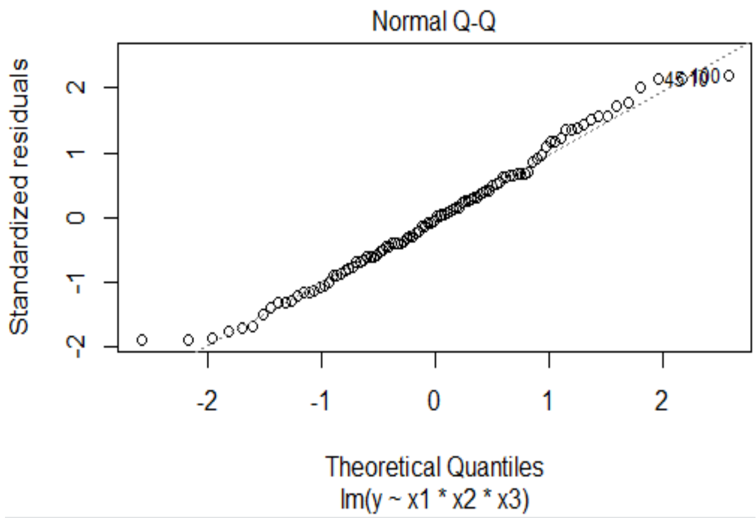
plot(model1)



The red line should be horizontal at zero ➔ there is no fitted pattern. It means we can accept linearity assumption.



residuals are spread equally along the ranges of predictors, the variances of the residual points is fairly constant ➔ assume equality of variance.



we can accept the normality assumption.

C)

```
datahw2 = lm(y ~ x1*x2*x3 , data = hw2_data_1)
summary(datahw2)
```

```
Call:
lm(formula = y ~ x1 * x2 * x3, data = hw2_data_1)

Residuals:
   Min     1Q Median     3Q    Max
-6.034 -2.224 -0.081  2.121  7.264
```

Coefficients:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.327393   3.559242   2.059   0.0424 *
x1           1.709184   1.251519   1.366   0.1754
x2          -0.166497   0.059186  -2.813   0.0060 **
x3           0.561826   0.312254   1.799   0.0753 .
x1:x2        0.038134   0.020579   1.853   0.0671 .
x1:x3        0.121700   0.110824   1.098   0.2750
x2:x3       -0.003239   0.005007  -0.647   0.5193
x1:x2:x3    -0.001350   0.001735  -0.778   0.4385
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.336 on 92 degrees of freedom
Multiple R-squared:  0.8574,	Adjusted R-squared:  0.8466
F-statistic: 79.04 on 7 and 92 DF,  p-value: < 2.2e-16
```

$H_0 : \beta_{x_1 x_2 x_3} = 0$ P-value is more than 0.05 ➔ we fail to reject H_0➔three way interaction is not significant in the model.

D)

We are going to test $H_0 : \beta_4 = \beta_5 = B_6 = \beta_7 = 0$.

```
null_datahw2= lm(y ~ x1+x2+x3, data =  hw2_data_1)
full_datahw2 = lm(y ~ x1*x2*x3 , data = hw2_data_1)
anova(null_datahw2,full_datahw2)
```

```
Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3
Model 2: y ~ x1 * x2 * x3
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     96 1240.8
2     92 1023.6  4    217.16 4.8795 0.001297 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value is less than 0.05➔ H_0 is rejected ➔at least there is at least one of the

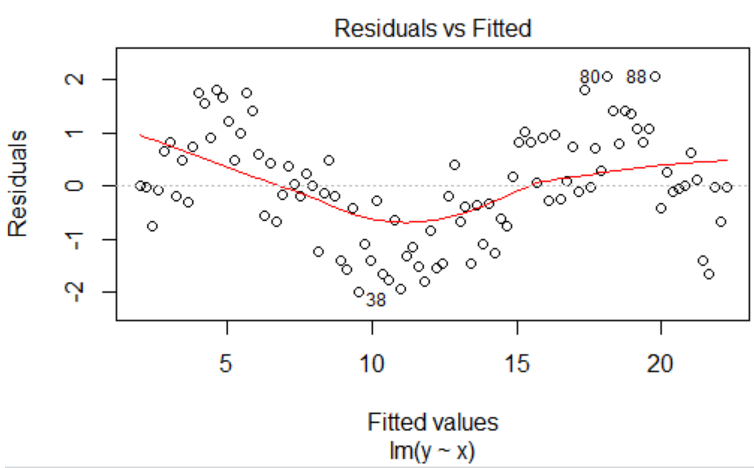
$\beta_4, \beta_5, B_6, \beta_7$

which is not zero and it is not significant at 0.05.

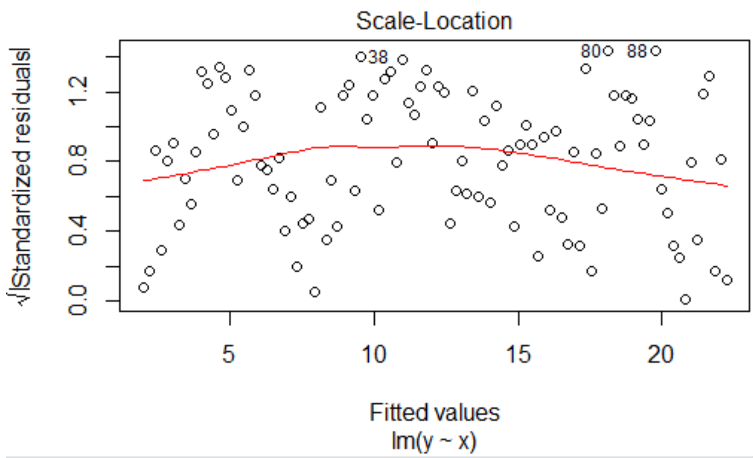
3-

```
model2= lm(y ~ x, data = hw2_data_2)
              summary(model2)

Call:
lm(formula = y ~ x, data = hw2_data_2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.99865 -0.65994 -0.02829  0.75697  2.07659

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.01828    0.20149   10.02   <2e-16 ***
x            2.02023    0.03481   58.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.015 on 98 degrees of freedom
Multiple R-squared:  0.9717,           Adjusted R-squared:  0.9714
F-statistic:  3368 on 1 and 98 DF,  p-value: < 2.2e-16
```

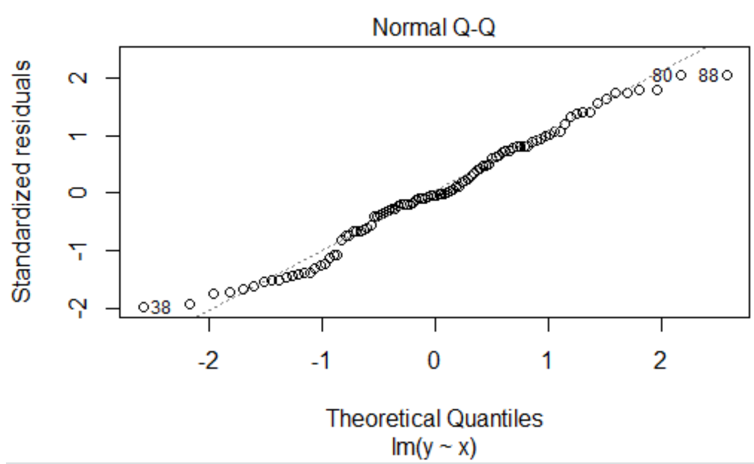
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 2.01828 + 2.02023(x)$$



Residuals vs Fitted: linearity assumption is not valid.



Equality of variance: the distribution of residuals has the same variance→ level of error in the model is approximately the same regardless of the value of the predictor variable



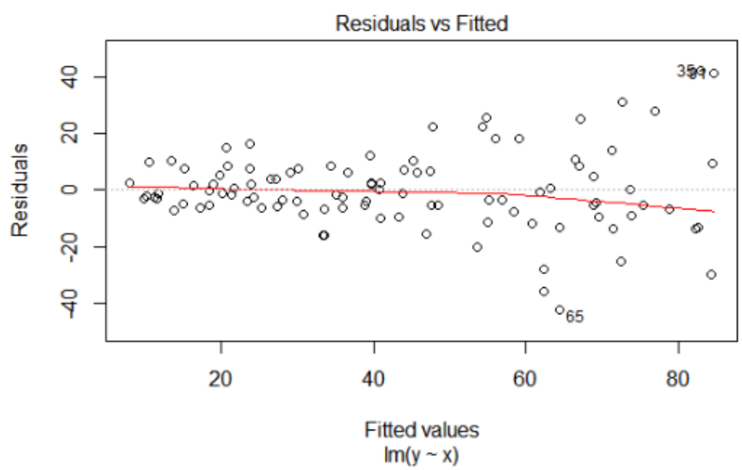
Normal Q-Q:

all the points fall approximately along the reference line, so we → normality is hold and normality assumption is accepted.

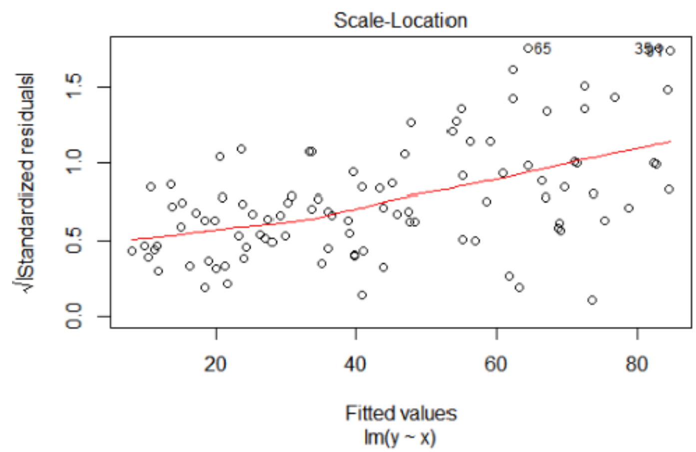
4-

```
model4= lm(y ~ x, data = hw2_data_3)
           summary(model4)

Call:
lm(formula = y ~ x, data = hw2_data_3)

Residuals:
    Min      1Q  Median      3Q     Max
-42.505  -6.346  -1.484   7.156  41.962

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.5459     3.3358  -1.363    0.176
x             3.8815     0.2443  15.887   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.88 on 98 degrees of freedom
Multiple R-squared:  0.7203,	Adjusted R-squared:  0.7175
F-statistic: 252.4 on 1 and 98 DF,  p-value: < 2.2e-16
```

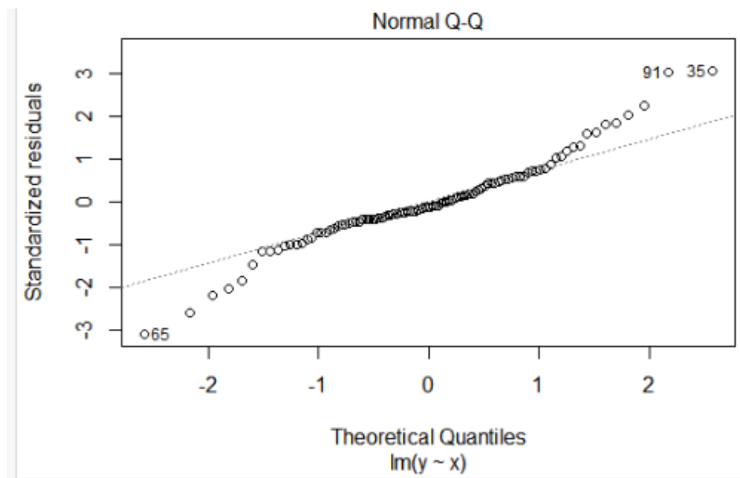
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = -4.5459 + 3.8815(x)$$



Residuals vs Fitted: we can **accept** linearity assumption.



Equality of variance: distribution of residuals does not have the same variance ➔assumption of equality of variances is rejected.



Normal Q-Q:

The pattern is non-linear, so plot give evidence against assume normality ➔Normality assumption is rejected.

5-

a)

```
lev_fit = lm(y ~ x, data = mydata)
leverages = hatvalues(lev_fit)
hatvalues(lev_fit) > 2 * mean(hatvalues(lev_fit))
```

```
    1     2     3     4     5     6     7     8     9    10 
FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
```

It means C and E have a high leverage.

b)

```
lev_fit1 = lm(y ~ x, data = newmydata)
> leverages = hatvalues(lev_fit1)
> leverages
        1         2         3         4         5         6         7         8         9        10 
0.1609862 0.1261784 0.4655547 0.1011603 0.5525743 0.1026106 0.1087745 0.1319797 0.1490
210 0.1011603
```

It means it has not changed.

C)

```
> leverages.df <- as.data.frame(t(leverages))
> leverages.df [, c (2,3,5,8)]
         2          3         5          8
1 2.021882 -0.6087305 -2.093985 -0.9718407
```

d)

```
cooks.distance(lev_fit)
cooks.distance.df <- as.data.frame(t(cooks.distance(lev_fit)))
cooks.distance.df [, c (2,3,5,8)]
          2         3        5          8
1 0.2951508 0.1613941 2.707616 0.07180215
```

```
> newdata <- cooks.distance.df [, c (2,3,5,8)]
newdata > 4 / length(newdata)
         2     3    5     8
[1,] FALSE FALSE TRUE FALSE
```

Only E is an influential point.