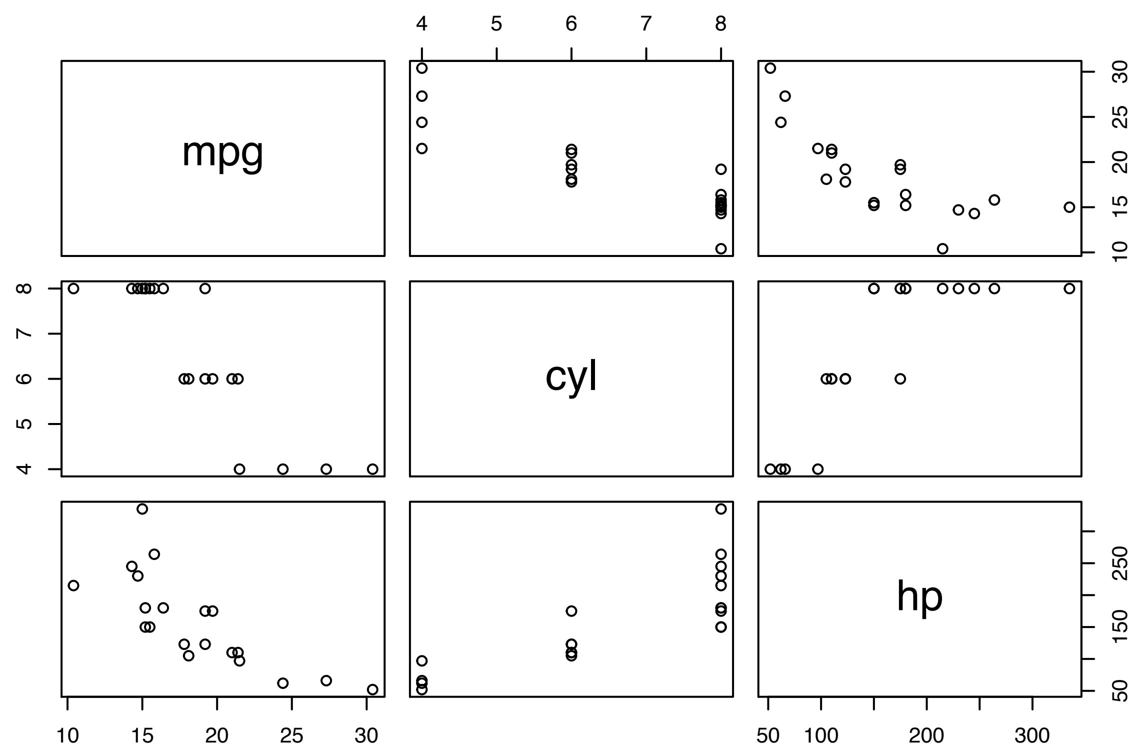
SS3859A Assignment 2

Bradley Assaly-Nesrallah

23/10/2020

Question 1

```r
set.seed(100)
sub_index=sample(nrow(mtcars),20,replace=FALSE)
mtcars2=mtcars[sub_index,c(1,2,4)]

##a.We plot a scatterplot matrix briefly discuss the relationships between, mph, cyl,
##and hp pairwise :

pairs(mtcars2)
```



```r
##mpg and cyl appear to have nonlinear relationship because cyl is categorical
##mpg and hp appear to have a linear relationship
##cyl and hp appear to have a nonlinear relationship because cyl is categorical

##b. We obtain the fitted model and determine the percentage variation in the fitted model:

mpg_cyl_hp = lm(mpg ~ cyl + hp, data=mtcars2)
summary(mpg_cyl_hp)$r.squared
```

```
## [1] 0.7771745
```

```r
## Thus 77.71745% of the variation in fuel consumption is explained by the fitted model

##c. We construct a 90% CI for Bcyl as follows:
confint(mpg_cyl_hp,level=0.90)
```

```
##                    5 %        95 %
## (Intercept) 30.45355849 39.34974881
## cyl         -3.21120343 -1.20512178
## hp          -0.03268241  0.01103919
```

```r
## The 90% CI for Bcyl is (-3.21120343,-1.20512178)

##d.We predict the fuel efficiency of three cars A,B,C as follows;
##predict for Car A, 4cyl 90hp -
predict(mpg_cyl_hp,newdata = data.frame(cyl=4,hp=90))
```

```
##        1
## 25.09506
```

```r
##Thus we predict A has a mpg of 25.09506
##predict for Car B, 6cyl 150hp -
predict(mpg_cyl_hp,newdata = data.frame(cyl=6,hp=150))
```

```
##        1
## 20.02944
```

```r
##Thus we predict B has a mpg of 20.02944
##predict for Car C, 8cyl 210hp -
predict(mpg_cyl_hp,newdata = data.frame(cyl=8,hp=210))
```

```
##        1
## 14.96381
```

```r
##Thus we predict B has a mpg of 14.96381

##e.We determine if based on the fitted model it is likely the fuel efficiency
##of car C is 3mpg by using a PI
predict(mpg_cyl_hp,newdata=data.frame(cyl=8,hp=210),interval="prediction",level=0.95)
```

```
##        fit      lwr      upr
## 1 14.96381 9.717618 20.21001
```

```
##Thus the PI at alpha=0.95 that the actual fuel efficeiency of car C
##is (9.717618 20.21001), thus 3 is not an element of the PI, so it
##is very unlikely

##f.We fill in the anova table as follows
anova(mpg_cyl_hp)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cyl        1 332.08  332.08 58.5513 6.653e-07 ***
## hp         1   4.21    4.21  0.7416    0.4011
## Residuals 17  96.42    5.67
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SSE = sum((fitted(mpg_cyl_hp)-mtcars2$mpg)^2)
SSR = sum((fitted(mpg_cyl_hp)-mean(mtcars2$mpg))^2)
SSR
```

```
## [1] 336.2815
```

```
SSE
```

```
## [1] 96.41603
```

```
SST=SSR+SSE
SST
```

```
## [1] 432.6975
```

```
MSE =SSE/17
MSR =SSR/2
MSE
```

```
## [1] 5.671531
```

```
MSR
```

```
## [1] 168.1407
```

```
F=MSR/MSE
F
```

```
## [1] 29.64644
```

```
## Source | Sum of Squares | df | Mean Squares | F
## Regres | 336.2815       |  2 | 168.1407     | 29.64644
## Error  |  96.42         | 17 | 5.672
## Total  | 432.6975       | 19

##g.We test H0:Bcyl=Bhp=0 vs Ha:Bcyl=\=Bhp=\=0 by F test at a=0.95
anova(mpg_cyl_hp)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cyl        1 332.08  332.08 58.5513 6.653e-07 ***
## hp         1   4.21    4.21  0.7416    0.4011
## Residuals 17  96.42    5.67
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
qf(0.95,2,17)
```

```
## [1] 3.591531
```

```
## Thus 29.6>3.6=F(0.95,2,17) thus we reject H0 and conclude at least one of the predictors
##has a significant linear relationship at alpha=0.05 (reject original statement)

##h. We test H0:Bhp=0 vs Ha:Bhp=/=0 at alpha=0.05
summary(mpg_cyl_hp)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp, data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5097 -1.0290 -0.0737  1.1809  4.8937
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.90165    2.55695  13.650 1.37e-10 ***
## cyl         -2.20816    0.57659  -3.830  0.00134 **
## hp          -0.01082    0.01257  -0.861  0.40114
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.381 on 17 degrees of freedom
## Multiple R-squared:  0.7772, Adjusted R-squared:  0.751
## F-statistic: 29.65 on 2 and 17 DF,  p-value: 2.869e-06
```

```
## Thus the P-value for Bhp=0 is 0.4>0.05 thus we accept H0,and
##conclude that Bhp=0 at the alpha=0.05 level
```

```r
##i.We fit another regression model Yi=B_0+B_hpxi2+eps_i and
##test H0:Bhp=0 vs Ha:Bhp=/=0 at alpha=0.05
mpg_hp = lm(mpg ~ hp, data=mtcars2)
summary(mpg_hp)
```

```
##
## Call:
## lm(formula = mpg ~ hp, data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3636 -2.3581 -0.0478  1.5760  6.5460
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.434985   1.703951  15.514 7.33e-12 ***
## hp          -0.049634   0.009855  -5.037 8.58e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.159 on 18 degrees of freedom
## Multiple R-squared:  0.5849, Adjusted R-squared:  0.5619
## F-statistic: 25.37 on 1 and 18 DF,  p-value: 8.579e-05
```

```r
##Thus we reject H0 and conclude that Ha:Bhp=/= at alpha=0.5

##j. The conclusions from h and j were different, in the first model Bhp is
##equal to zero vs the second model where it is not. This can be explained by
##considering the relationship between cylinders and hp, the amount of cylinders
##correlates with the amount of hp, where in a model with both predictors, in
##the model with both hp is insignifigant whereas in the model with just hp it
##has a statistically signifigant effect in determining mpg

set.seed(2)
sub_index=sample(nrow(mtcars),27,replace=FALSE)
mtcars3=mtcars[sub_index,c(1:4,10)]
mph_full =lm(mpg~cyl+disp+hp+gear,data=mtcars3)
summary(mph_full)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + hp + gear, data = mtcars3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7621 -1.8497 -0.5353  1.4011  6.6236
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.29649    5.62682   4.673 0.000117 ***
## cyl         -0.81743    0.77101  -1.060 0.300555
## disp        -0.01348    0.01131  -1.192 0.245971
```

```
## hp          -0.02423    0.02196  -1.103 0.281782
## gear         1.35239    1.07202   1.262 0.220327
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.638 on 22 degrees of freedom
## Multiple R-squared:  0.8057, Adjusted R-squared:  0.7704
## F-statistic:  22.8 on 4 and 22 DF,  p-value: 1.471e-07
```

```r
##k.The summary output shows the individual p values of each Bcyl,disp,hp,gear
##are larger than 0.05, this does not mean that none of the predictors are
##linearly related to the response at 0.05. It just means that the variables are
##highly correled, in fact you can see that many of the variables are linearly
##related to mpg, they are just correlated with each other

##l. we test H0:Bdisp=Bhp=Bgear=0 vs Ha at least on of Bj=/=0
null_mpg_model = lm(mpg ~ cyl, data = mtcars3)
full_mpg_model = lm(mpg ~disp+hp+gear,data=mtcars3)
anova(null_mpg_model,full_mpg_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl
## Model 2: mpg ~ disp + hp + gear
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     25 203.08
## 2     23 160.96  2    42.123 3.0095 0.06903 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(full_mpg_model)
```

```
##
## Call:
## lm(formula = mpg ~ disp + hp + gear, data = mtcars3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3015 -1.6681 -0.5983  1.2605  6.6554
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.40307    4.27470   5.241 2.57e-05 ***
## disp        -0.01729    0.01075  -1.609   0.1213
## hp          -0.03694    0.01845  -2.003   0.0571 .
## gear         1.78517    0.99393   1.796   0.0856 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.645 on 23 degrees of freedom
## Multiple R-squared:  0.7958, Adjusted R-squared:  0.7691
## F-statistic: 29.87 on 3 and 23 DF,  p-value: 4.149e-08
```

```r
#Thus the p value of the f test is less than 0.1 so we reject H0 and claim that
##at least one of them is statistically different from 0 so we are done
```

Question 2

```r
set.seed(50)
idx <- sample(32,25,replace=FALSE)
mtcars2<-mtcars[idx,]
mtcars2$cyl<-as.factor(mtcars2$cyl)
summary(mtcars2)
```

```
##       mpg         cyl         disp              hp             drat
##  Min.   :10.40   4: 9   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
##  1st Qu.:15.20   6: 4   1st Qu.:121.0   1st Qu.: 95.0   1st Qu.:3.150
##  Median :18.70   8:12   Median :167.6   Median :150.0   Median :3.700
##  Mean   :19.62          Mean   :234.6   Mean   :147.2   Mean   :3.635
##  3rd Qu.:22.80          3rd Qu.:350.0   3rd Qu.:180.0   3rd Qu.:3.920
##  Max.   :33.90          Max.   :472.0   Max.   :264.0   Max.   :4.930
##        wt             qsec             vs            am             gear
##  Min.   :1.615   Min.   :14.50   Min.   :0.0   Min.   :0.00   Min.   :3.00
##  1st Qu.:2.770   1st Qu.:17.02   1st Qu.:0.0   1st Qu.:0.00   1st Qu.:3.00
##  Median :3.435   Median :17.82   Median :0.0   Median :0.00   Median :4.00
##  Mean   :3.312   Mean   :17.90   Mean   :0.4   Mean   :0.36   Mean   :3.64
##  3rd Qu.:3.730   3rd Qu.:18.61   3rd Qu.:1.0   3rd Qu.:1.00   3rd Qu.:4.00
##  Max.   :5.424   Max.   :22.90   Max.   :1.0   Max.   :1.00   Max.   :5.00
##       carb
##  Min.   :1.00
##  1st Qu.:2.00
##  Median :3.00
##  Mean   :2.84
##  3rd Qu.:4.00
##  Max.   :6.00
```

```r
##We obtain the fitted value of mpg at weight=3,cyl=6 as follows

mpg_wt_cyl=lm(mpg~wt+cyl,data=mtcars2)

predict(mpg_wt_cyl,newdata = data.frame(wt=3,cyl='6'))
```

```
##        1
## 19.83571
```

```r
##We obtain the fitted value that mpg is 19.83571 when wt=3 and cyl=6

##b. We determine that cyl is an important predictor given that wt is used as a
##predictor at the alpha=0.05 level
summary(mpg_wt_cyl)
```

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

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5074 -1.4844 -0.6048  1.5388  5.9212
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.7210     2.3004  14.659 1.68e-12 ***
## wt           -3.1292     0.8741  -3.580  0.00177 **
## cyl6         -4.4975     1.7790  -2.528  0.01955 *
## cyl8         -6.2877     1.8811  -3.342  0.00309 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.769 on 21 degrees of freedom
## Multiple R-squared:  0.8267, Adjusted R-squared:  0.802
## F-statistic:  33.4 on 3 and 21 DF,  p-value: 3.523e-08
```

```r
##thus since the p value of cyl are less than 0.02,0.003<.05=alpha
##so they are important given that wt is used

##c.We obtain the fitted value of mpg at wt=3,cyl=8 as follows

mpg_wt_cyl_int=lm(mpg~wt+cyl+wt:cyl,data=mtcars2)
predict(mpg_wt_cyl,newdata = data.frame(wt=3,cyl='8'))
```

```
##       1
## 18.0456
```

```r
##We predict the fitted value of mpg at wt=3 cyl=8 is 18.0456

##d. We use the f test at alpha=0.05 totest H0 there is significant interaction
##between two predictors
anova(mpg_wt_cyl,mpg_wt_cyl_int)
```

```
## 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     21 160.96
## 2     19 132.26  2    28.698 2.0613 0.1548
```

```r
#$As P value =0.15>0.05 we do not reject H0, thus there is no interaction
##between two predictors
```

Question 3

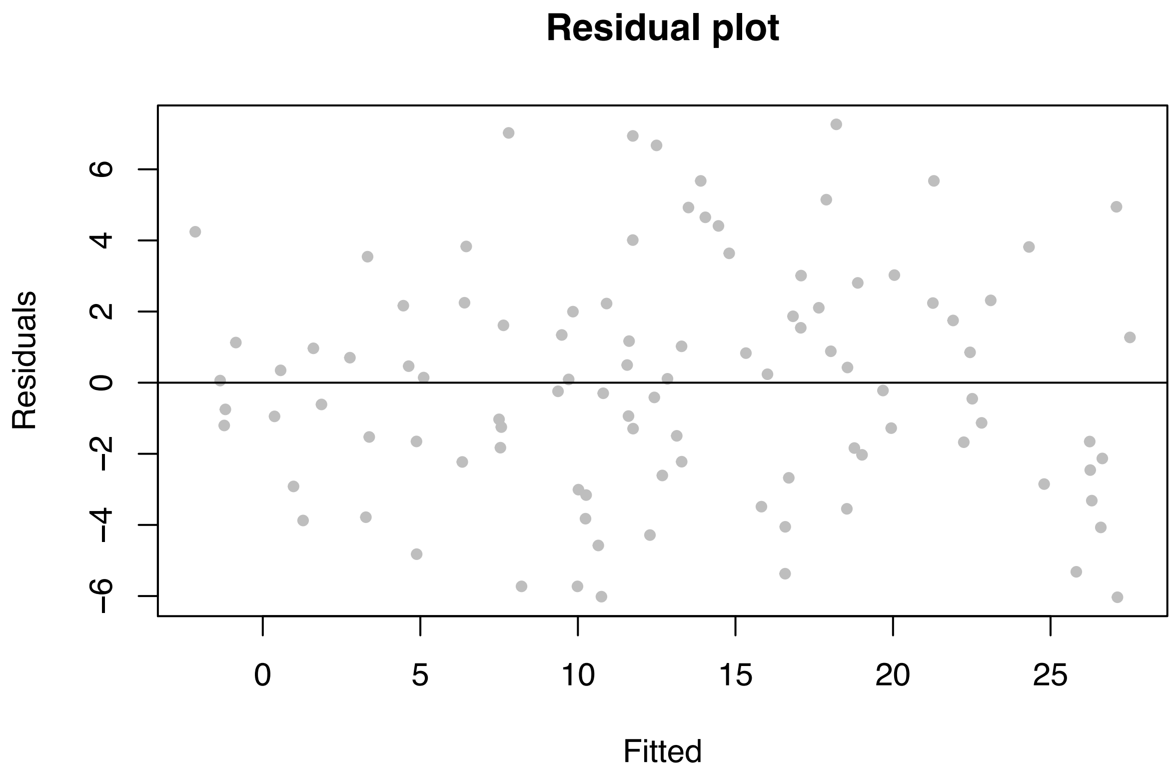
```r
data_1 = read.csv("https://raw.githubusercontent.com/hgweon2/ss3859/master/hw2-data-1.csv")
lm_x123=lm(y~x1+x2+x3+x1:x2+x1:x3+x2:x3+x1:x2:x3,data=data_1)

##a.We compute the increase in Yi for one unit of x1 given x2=50,x3=7 as follows
summary(lm_x123)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3 + x1:x2:x3,
##     data = 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
```

```r
##We sum all of the terms with x1, given the x2,x3 values
A=1.709*1+0.03813*1*50+0.1217*1*7-0.00135*1*50*7
A
```

```
## [1] 3.9949
```

```r
##Therefore A is 3.9949 so done
##b.We plot the residuals and normal qq plot of the lm to check the linearity,
##equal variance and normality assumtion
plot(fitted(lm_x123),resid(lm_x123), col = "grey", pch = 20,
     xlab = "Fitted", ylab = "Residuals", main = "Residual plot")
abline(h=0)
```

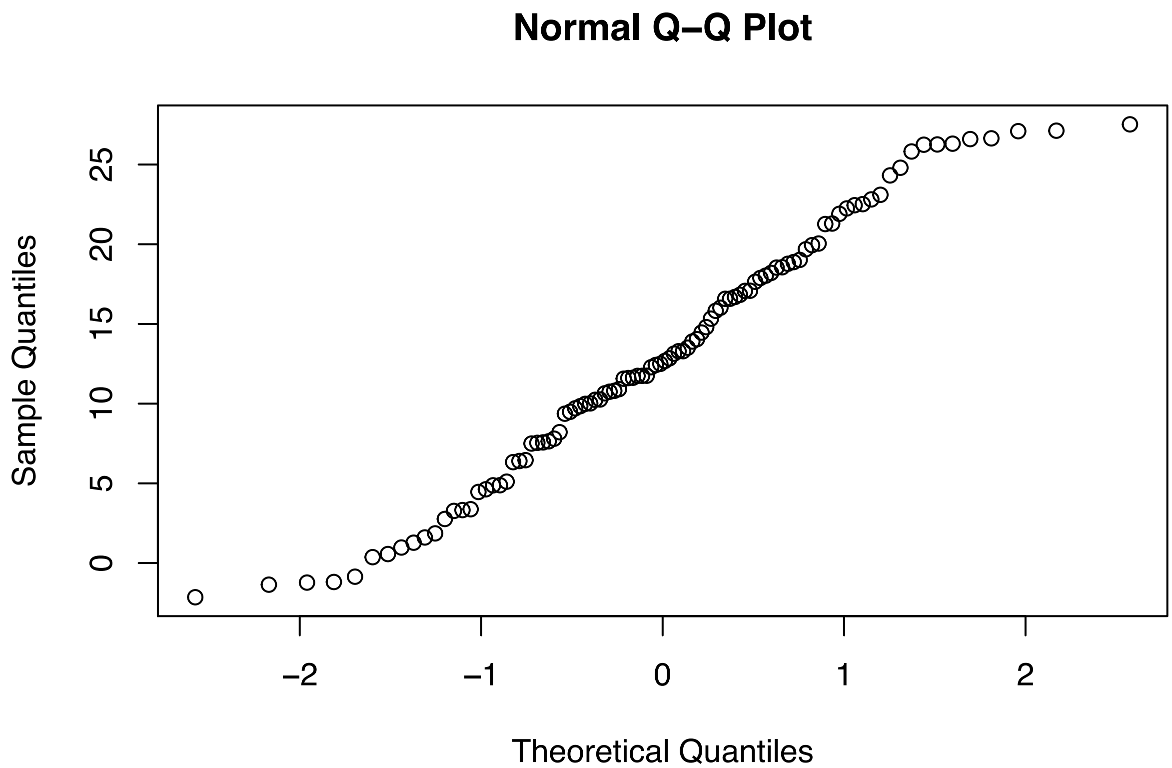



```
qqnorm(fitted(lm_x123))
##Thus we can see that the residuals are approximately linear, equal variance
##and normal from inspecting the qqplot and residual plot as the qqplot
##appears normal and the resid plot is approx evenly distributed

##c.We check the equal variance and normality assumtion using statistical tests
##at alpha=-0.05 level
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```




```r
bptest(lm_x123)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm_x123
## BP = 6.4252, df = 7, p-value = 0.4911
```

```r
shapiro.test(resid(lm_x123))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(lm_x123)
## W = 0.98441, p-value = 0.2875
```

```r
##in both tests the p-value is greater than 0.05 so we do not reject the equal
##variance and normality assumption

##d.We check if the three way interaction term was needed
lm_x12=lm(y~x1+x2+x3+x1:x2+x1:x3+x2:x3,data=data_1)
anova(lm_x12,lm_x123)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3
## Model 2: y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3 + x1:x2:x3
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     93 1030.3
## 2     92 1023.6  1     6.737 0.6055 0.4385
```

```
##thus we conclude that the three way term is not needed since the p value 0.43>0.1
##e. We test the statement "there are no interaction effects between
##the predictors" at alpha=0.05
summary(lm_x12)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3, data = data_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8732 -2.2382  0.0436  2.1369  7.2053
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.154811   2.202770   2.340 0.021415 *
## x1           2.538743   0.654183   3.881 0.000194 ***
## x2          -0.128616   0.033589  -3.829 0.000233 ***
## x3           0.767712   0.165470   4.640 1.14e-05 ***
## x1:x2        0.023718   0.008941   2.653 0.009386 **
## x1:x3        0.042163   0.042737   0.987 0.326411
## x2:x3       -0.006757   0.002149  -3.145 0.002231 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.329 on 93 degrees of freedom
## Multiple R-squared:  0.8565, Adjusted R-squared:  0.8472
## F-statistic: 92.51 on 6 and 93 DF,  p-value: < 2.2e-16
```

```
summary(lm_x123)
```

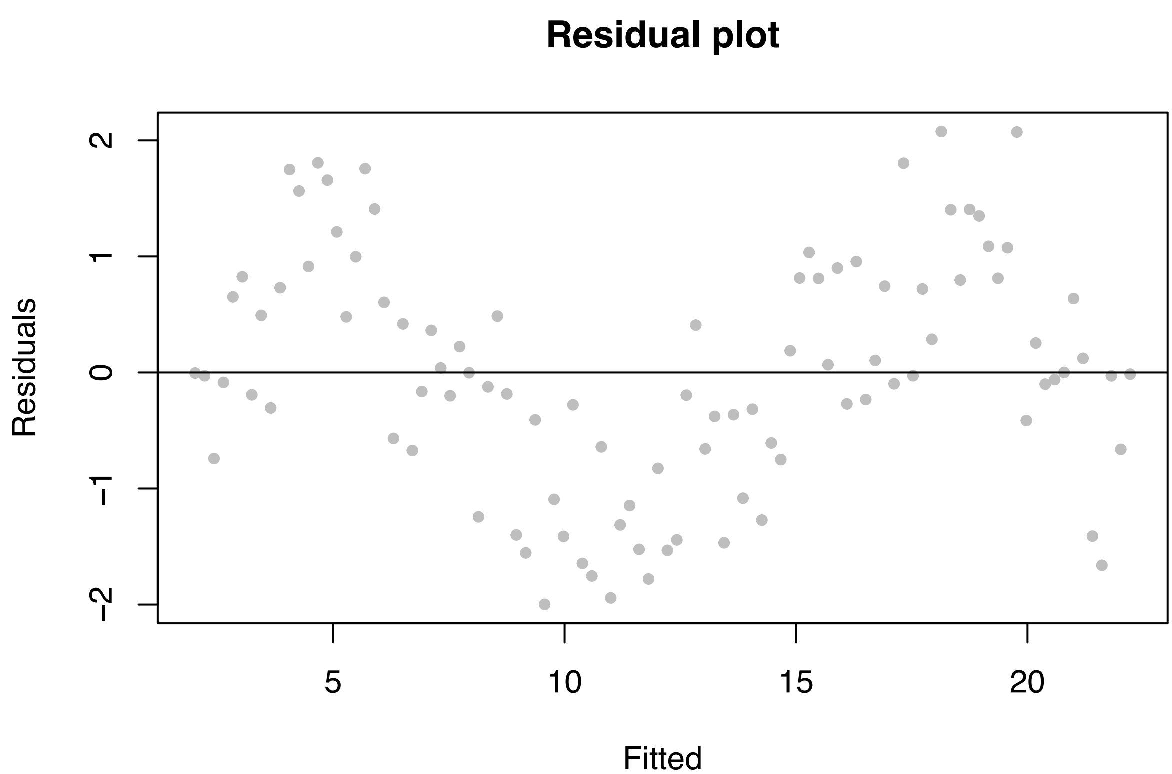
```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x1:x2 + x1:x3 + x2:x3 + x1:x2:x3,
##     data = data_1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.034 -2.224 -0.081  2.121  7.264
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.327393   3.559242   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
```

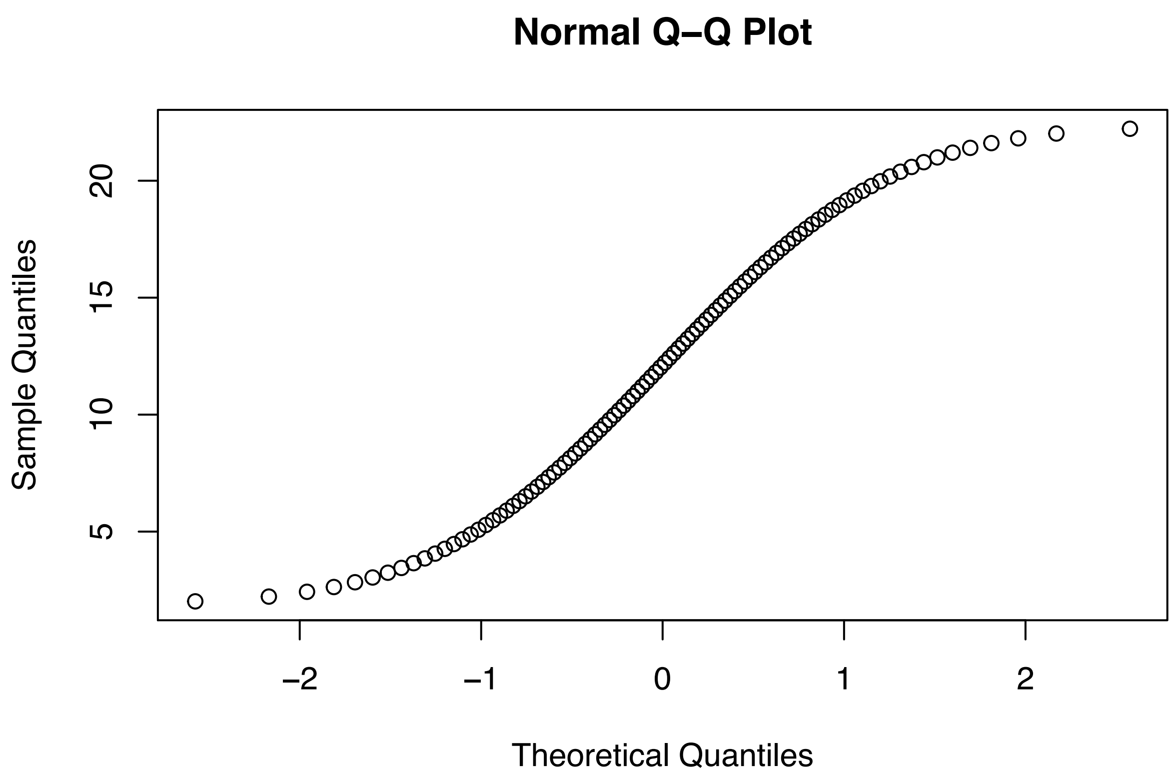
```
##Thus there are no interaction effects betwen the predictors at alpha=0.05
```

Question 4

```
data_2 = read.csv("https://raw.githubusercontent.com/hgweon2/ss3859/master/hw2-data-2.csv")

lm_q4=lm(y~x,data=data_2)
plot(fitted(lm_q4),resid(lm_q4), col = "grey", pch = 20,
     xlab = "Fitted", ylab = "Residuals", main = "Residual plot")
abline(h=0)
```




```
qqnorm(fitted(lm_q4))
```




```r
##it appears the equal variance and normality assumptions hold but linearity
##fails when we inspect the redidual plot, as the resid are pos at certain
##regions and negative in others with equal variance, meanwhile the qq plot
##looks approximately normal
##we perform the bp and shapiro test at alpha=0.05
bptest(lm_q4)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm_q4
## BP = 0.0090726, df = 1, p-value = 0.9241
```

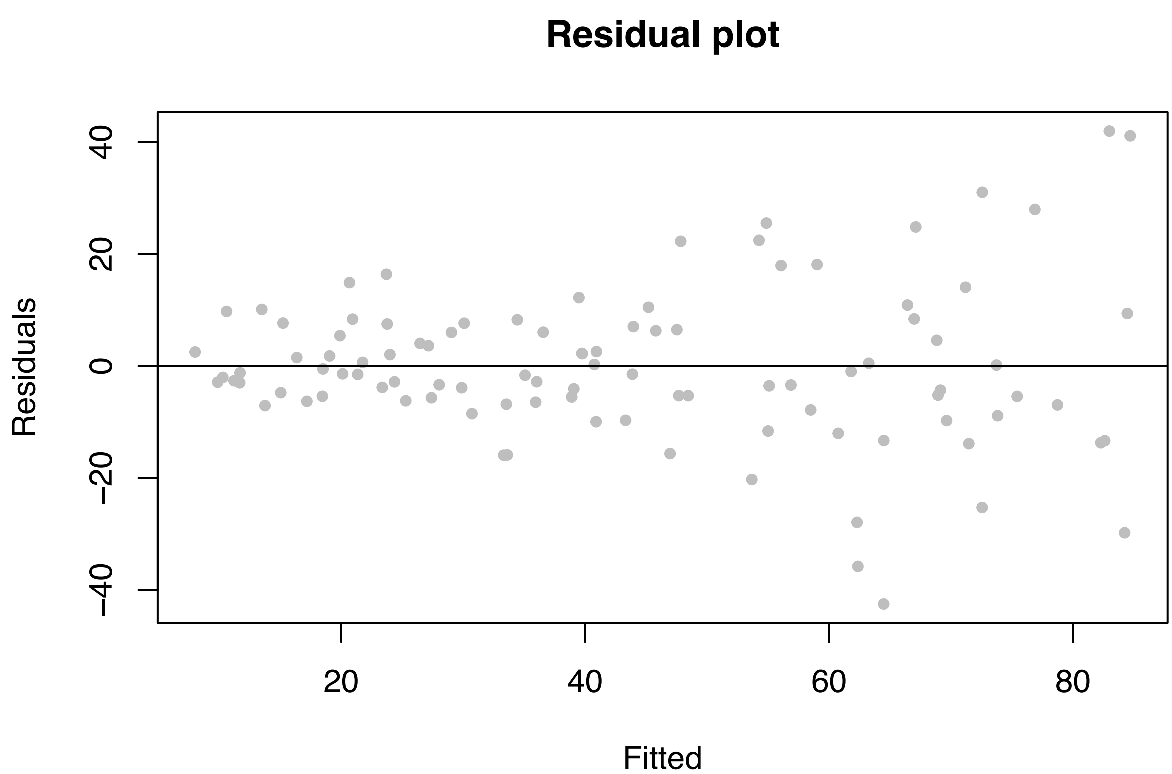
```r
shapiro.test(resid(lm_q4))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(lm_q4)
## W = 0.97905, p-value = 0.1121
```

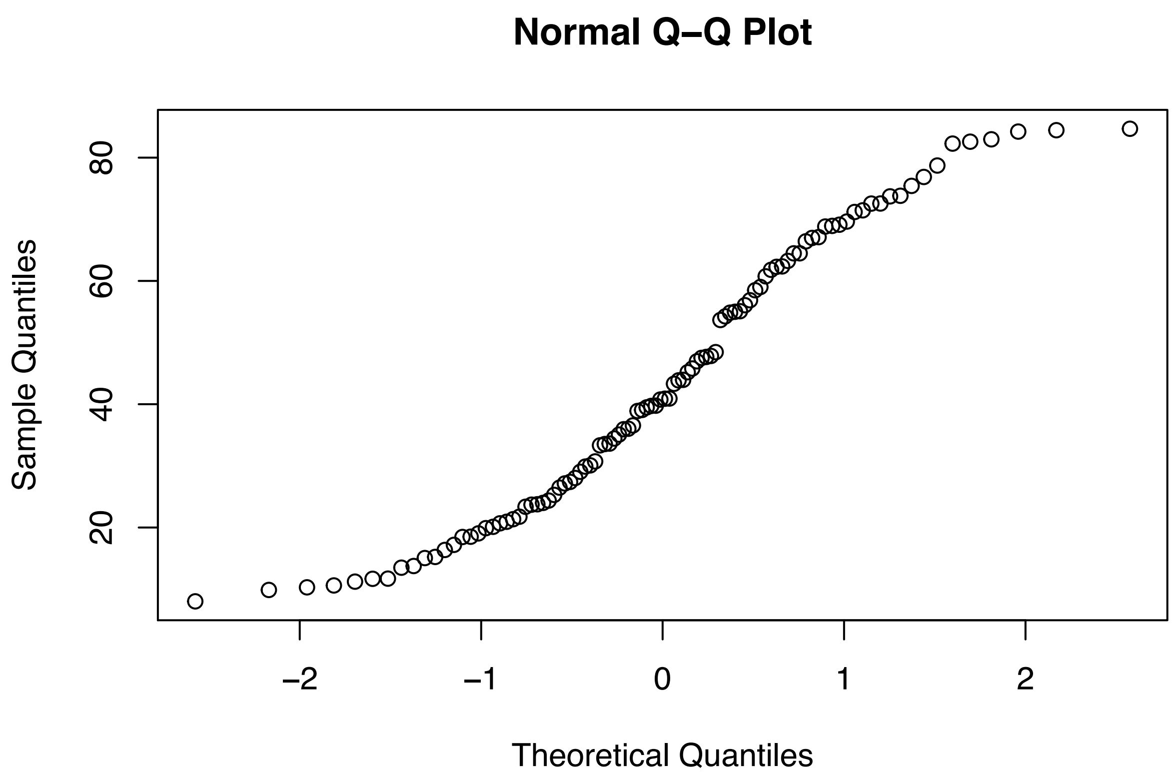
```r
#thus both tests the p value is greater than alpha=0.05 so we do
##not reject the normality assumption nor the equal variance assumption
```

Question 5

```r
data_3 = read.csv("https://raw.githubusercontent.com/hgweon2/ss3859/master/hw2-data-3.csv")

lm_q5=lm(y~x,data=data_3)
plot(fitted(lm_q5),resid(lm_q5), col = "grey", pch = 20,
     xlab = "Fitted", ylab = "Residuals", main = "Residual plot")
abline(h=0)
```




```r
qqnorm(fitted(lm_q5))
```




```r
##it appears the linearity hold but equal variance and normality
##fails as the resid plot is linear with variance increassing as fitted
##value increased meanwhile the norm qq plot appears to not be normal
##we perform the bp and shapiro test at alpha=0.05
bptest(lm_q5)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lm_q5
## BP = 22.542, df = 1, p-value = 2.056e-06
```

```r
shapiro.test(resid(lm_q5))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(lm_q5)
## W = 0.95913, p-value = 0.003487
```

```r
#thus both tests the p value is less than alpha=0.05 so we do reject the
##normality assumption and the equal variance assumption
```