Choi, Asgm2

Yeonsil Choi

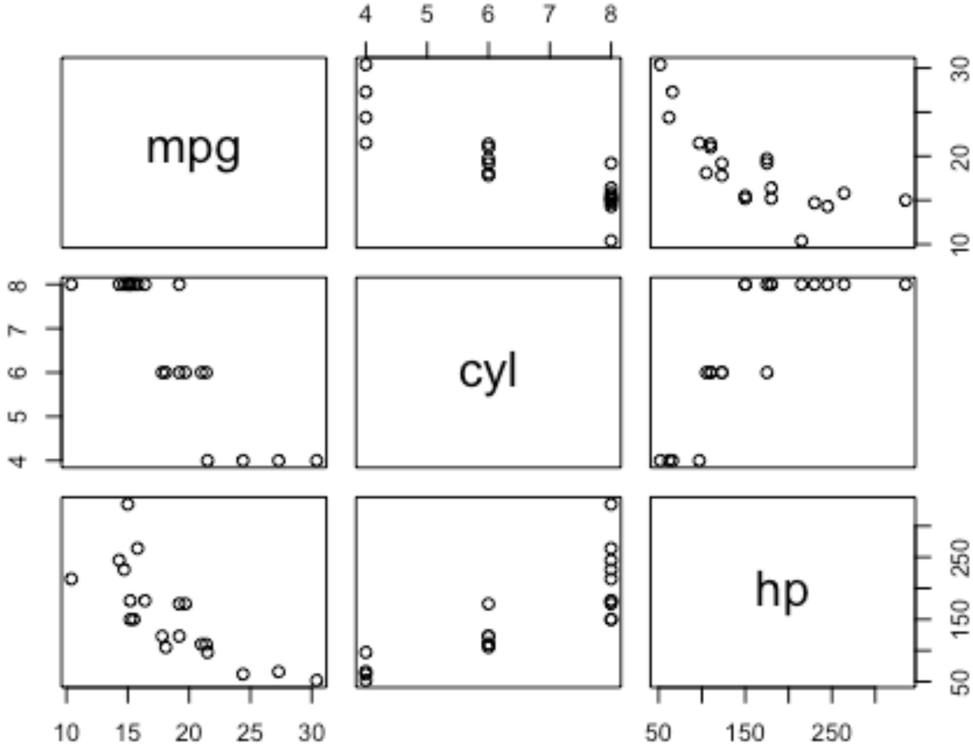
23/10/2021

Q1 For problem 1, you will use subsets of the mtcars data. (Run "?mtcars" for detailed information about the data.) Run the following R codes and use the mtcars2 dataset to answer the questions (a)-(j).

```
# check detailed information about data
?mtcars

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

##                      mpg cyl  hp
## Merc 280            19.2   6 123
## AMC Javelin         15.2   8 150
## Valiant             18.1   6 105
## Lincoln Continental 10.4   8 215
## Honda Civic         30.4   4  52
## Pontiac Firebird    19.2   8 175
```

For the moment, consider a multiple linear regression model that predicts the fuel consumption (Y = mpg) from number of cylinders (X1 = cyl) and horsepower (X2 = hp).

```
##(a) Plot a scatterplot matrix and briefly discuss the relationships (e.g.,
linear, nonlinear or independent) between all pairs of the variables.

# Plot a scatterplot matrix
pairs(~.,data=mtcars2)
```



Between mpg and cyl, there is a negative linear relationship. Also, between mpg and hp, there is a negative linear relationship.

```
##(b) Obtain the fitted model of the regression. What percentage of the
variation in fuel consumption is explained by your fitted model?

# obtain the fitted model of the regression
fitted_model = lm(mpg ~ cyl + hp, data = mtcars2)
summary(fitted_model)

##
## 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
```

```
# y_hat = 34.90165 + (-2.20816 * x_i1) + (-0.01082 * x_i2)

# print R^2 from the lm object
summary(fitted_model)$r.squared
```

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

77.72 percentage of the variation in fuel consumption is explained by my fitted model.

```
##(c) Construct a 90% confidence interval for cyl
confint(fitted_model, level = 0.9)
```

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

From the output, a 90% confidence interval for β_{cyl} is (-3.21120343, -1.20512178).

```
##(d) Predict the fuel efficiency of new cars A, B and C. Use the fitted
model to obtain the predicted fuel efficiencies (point estimates).
predict(fitted_model, newdata=data.frame(cyl = 4, hp = 90))
```

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

```
predict(fitted_model, newdata=data.frame(cyl = 6, hp = 150))
```

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

```
predict(fitted_model, newdata=data.frame(cyl = 8, hp = 210))
```

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

```
##(e) Based on the fitted model, is it likely that the actual fuel efficiency
of car C is near 5 miles/gal? You may consider a prediction interval to
support your answer.
predict(fitted_model, newdata=data.frame(cyl = 8, hp = 210),
interval="prediction", level=0.95)
```

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

It is not likely that the actual fuel efficiency of car C is near 5 miles/gal since the prediction interval ranges from 9.72 to 20.21.

```
##(f) Fill in the following ANOVA table.
null_mpg_model = lm(mpg ~ 1, data = mtcars2)
full_mpg_model = lm(mpg ~ cyl + hp, data = mtcars2)

# Significance of regression test
anovatable = anova(null_mpg_model, full_mpg_model) # RSS equals to SSE
anovatable

## Analysis of Variance Table
##
## Model 1: mpg ~ 1
## Model 2: mpg ~ cyl + hp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     19 432.70
## 2     17  96.42  2    336.28 29.646 2.869e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

ssr = anovatable$RSS[1] - anovatable$RSS[2]
ssr

## [1] 336.2815

sse = anovatable$RSS[2]
sse

## [1] 96.41603

sst = anovatable$RSS[1]
sst

## [1] 432.6975

n = nrow(mtcars2)
n

## [1] 20

p = 3 # num of beta under H1
p

## [1] 3

q = 1 # num of beta under H0 (intercept)
q

## [1] 1

msd = ssr/(p-q)
msd
```

```
## [1] 168.1407

mse = sse/(n-p)
mse

## [1] 5.671531

f = msd/mse
f

## [1] 29.64644

mat1 = c(ssr, p-q, msd, f)
mat2 = c(sse, n-p, mse, "N/A")
mat3 = c(sst, n-q, "N/A", "N/A")

mat4 = rbind(mat1, mat2, mat3)
colnames(mat4) <- c("sum of sqaures", "df", "mean squares", "f" )
rownames(mat4) <-c("regression", "error", "total")

mat4

##            sum of sqaures     df   mean squares       f
## regression "336.281469330663" "2"  "168.140734665332" "29.6464443668462"
## error      "96.4160306693369" "17" "5.67153121584335" "N/A"
## total      "432.6975"         "19" "N/A"              "N/A"

##(g) Test the statement "None of the two predictors has a significant linear relationship with the response."
summary(full_mpg_model)

##
## 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
```

```r
p_val = 1-pf(f,p-q,n-p)
p_val

## [1] 2.868807e-06

p_val < 0.05

## [1] TRUE
```

Therefore, H0 is rejected. At least one of the 2 predictors is still important.

```r
##(h) Test H0: beta_hp = 0 vs H1: beta_hp != 0 at alpha = 0.05
summary(full_mpg_model)

##
## 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

0.40114 > 0.05

## [1] TRUE
```

We fail to reject null hypothesis. The predictor hp is not needed given that the other predictors are already used.

```r
##(i) Fit another regression model: horsepower is the only predictor. Test
H0: beta_hp = 0 vs H1: beta_hp != 0 at alpha = 0.05
fitted_model2 = lm(mpg ~ hp, data = mtcars2)
summary(fitted_model2)

##
## 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
```

```
8.58e-05 < 0.05
```

```
## [1] TRUE
```

We reject null hypothesis. Therefore, predictor hp has a significant linear relationship with mpg.

(j) Were your conclusions from (h) and (i) consistent? If not, how can the contradictory results be explained?

Conclusions from (h) and (i) are not consistent because the predictors are highly correlated. The result is not contradictory because predictors are correlated.

```
set.seed (2)
sub_index = sample(nrow(mtcars),27,replace=FALSE)
mtcars3 = mtcars[sub_index,c(1:4, 10)]

##(k) The summary output of the fitted model shows that the individual p-
values for cyl, disp, hp and gear are all larger than 0.05. Does this mean
that none of the predictors is linearly related with the response at alpha =
0.05?

# The p-values were high for all predictors.
# This means that "given the others are in the model",
# the predictor of interest is not important.
# F-statistic = 22.8 and its corresponding p-value = 1.471e-07 < 0.05.
# It shows that at least one of them is important.
fitted_model3 = lm(mpg ~ ., data = mtcars3)
summary(fitted_model3)
```

```
##
## Call:
## lm(formula = mpg ~ ., 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
```

```
##(l) At alpha = 0.05, test that H0: beta_disp = beta_hp = beta_gear = 0 vs
H1: At least one of beta_j != 0 (j = disp, hp, gear).

null_mpg_model = lm(mpg ~ cyl, data = mtcars3)
full_mpg_model = lm(mpg ~ ., data = mtcars3)
anova(null_mpg_model, full_mpg_model)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl
## Model 2: mpg ~ cyl + disp + hp + gear
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     25 203.08
## 2     22 153.14  3    49.947 2.3918 0.09596 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-val is 0.09596 > 0.05. Therefore, we fail to reject null hypothesis. The predictors disp, hp, gear may be dropped.

Q2 For problem 2, run the following R codes and use the mtcars2 dataset to answer the questions (a)-(d).

```
set.seed(30)
idx <- sample(32,25,replace=FALSE)
mtcars2 <- mtcars[idx, ]
mtcars2$cyl <- as.factor(mtcars2$cyl)
mtcars2$cyl
```

```
##  [1] 6 8 4 8 8 8 4 4 4 6 8 8 4 4 8 8 4 6 8 4 4 4 8 8 8
## Levels: 4 6 8
```

```
#use the mlr model where xi1 is weight, wi1 is 1 if cyl = 6 and 0 otherwise,
and wi2 is 1 if cyl = 8 and 0 otherwise
mpg_wt_cyl = lm(mpg ~ wt + cyl, data = mtcars2)
summary(mpg_wt_cyl)
```

```
##
## Call:
## lm(formula = mpg ~ wt + cyl, data = mtcars2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.686 -1.775 -0.348  1.268  5.666
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.1991     2.1962  15.572 5.22e-13 ***
## wt           -3.2506     0.9106  -3.570  0.00181 **
## cyl6         -3.6159     2.0446  -1.769  0.09150 .
## cyl8         -5.8197     2.0992  -2.772  0.01141 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.85 on 21 degrees of freedom
## Multiple R-squared:  0.8373, Adjusted R-squared:  0.8141
## F-statistic: 36.03 on 3 and 21 DF,  p-value: 1.828e-08
```

The mtcars2 data set contains 25 observations. Consider a regression model where the response is mpg and two predictors are weight and cylinder. Note that this time, we treat the cylinder predictor as a categorical variable with three categories (4,6,8).

```
##(a) Obtain the fitted value of mpg at weight = 3, cylinder = 6.
summary(mpg_wt_cyl)
```

```
##
## Call:
## lm(formula = mpg ~ wt + cyl, data = mtcars2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.686 -1.775 -0.348  1.268  5.666
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.1991     2.1962  15.572 5.22e-13 ***
## wt           -3.2506     0.9106  -3.570  0.00181 **
## cyl6         -3.6159     2.0446  -1.769  0.09150 .
## cyl8         -5.8197     2.0992  -2.772  0.01141 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.85 on 21 degrees of freedom
## Multiple R-squared:  0.8373, Adjusted R-squared:  0.8141
## F-statistic: 36.03 on 3 and 21 DF,  p-value: 1.828e-08
```

```
34.1991 + (-3.2506)*3 + (-3.6159)
```

```
## [1] 20.8314

##(b) Is cyl an important predictor given that wt is used as a predictor?

# Test H0: beta_cyl = 0 vs H1: beta_cyl != 0
# Low p-value --> Reject H0 --> Using two fitted lines gives a much better
fit.
summary(mpg_wt_cyl)

##
## Call:
## lm(formula = mpg ~ wt + cyl, data = mtcars2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.686 -1.775 -0.348  1.268  5.666
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.1991     2.1962  15.572 5.22e-13 ***
## wt           -3.2506     0.9106  -3.570  0.00181 **
## cyl6         -3.6159     2.0446  -1.769  0.09150 .
## cyl8         -5.8197     2.0992  -2.772  0.01141 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.85 on 21 degrees of freedom
## Multiple R-squared:  0.8373, Adjusted R-squared:  0.8141
## F-statistic: 36.03 on 3 and 21 DF,  p-value: 1.828e-08

# Compare two models using R^2.
# Adding a single predictor (cyl) increases the goodness-of-fit a lot.
mpg_wt = lm(mpg ~ wt, data = mtcars2)

summary(mpg_wt)$r.squared

## [1] 0.7765233

summary(mpg_wt_cyl)$r.squared

## [1] 0.8373285

# Same test using F-test
reduced = lm(mpg ~ wt, data = mtcars2)
full = lm(mpg ~ wt + cyl, data = mtcars2)

anova(reduced,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     23 234.36
## 2     21 170.59  2    63.765 3.9248 0.03563 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Therefore, cyl is an important predictor given that wt is used as a
predictor.
```

Suppose we wonder if there is a significant interaction between the weight and cylinder predictors.

```
##(c) Obtain the fitted value of mpg at weight = 3, cylinder = 8.
mpg_wt_cyl2 = lm(mpg ~ wt + cyl + wt:cyl, data = mtcars2)
summary(mpg_wt_cyl2)

##
## Call:
## lm(formula = mpg ~ wt + cyl + wt:cyl, data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9435 -1.4773 -0.7729  1.3495  5.3542
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   40.347      4.004  10.078 4.64e-09 ***
## wt            -6.046      1.776  -3.404  0.00298 **
## cyl6         -14.839     14.903  -0.996  0.33189
## cyl8         -15.560      5.896  -2.639  0.01618 *
## wt:cyl6        4.437      4.945   0.897  0.38075
## wt:cyl8        3.677      2.060   1.784  0.09032 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.765 on 19 degrees of freedom
## Multiple R-squared:  0.8615, Adjusted R-squared:  0.8251
## F-statistic: 23.64 on 5 and 19 DF,  p-value: 1.513e-07

40.347 + (-6.046 * 3) + (-14.839 * 0) + (-15.560) + (4.437 * 3 * 0) + (3.677
* 3)

## [1] 17.68

##(d) Test the null hypothesis: "There is no significant interaction effect
between two predictors."
without_interaction = lm(mpg ~ wt + cyl, data = mtcars2)
with_interaction = lm(mpg ~ wt + cyl + wt:cyl, data = mtcars2)
anova(without_interaction,with_interaction)
```

```
## 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     21 170.59
## 2     19 145.24  2    25.347 1.6579 0.2169
```

Since p-val is 0.2169 > 0.05, we fail to reject null hypothesis. Therefore, the interaction between two predictors is not significant.

Q3 For problem 3, import the data in R from https://raw.githubusercontent.com/hgweon2/ss3859/master/hw2-data-1.csv The data set contains 100 observations with 4 variables: y (response), x1, x2 and x3.

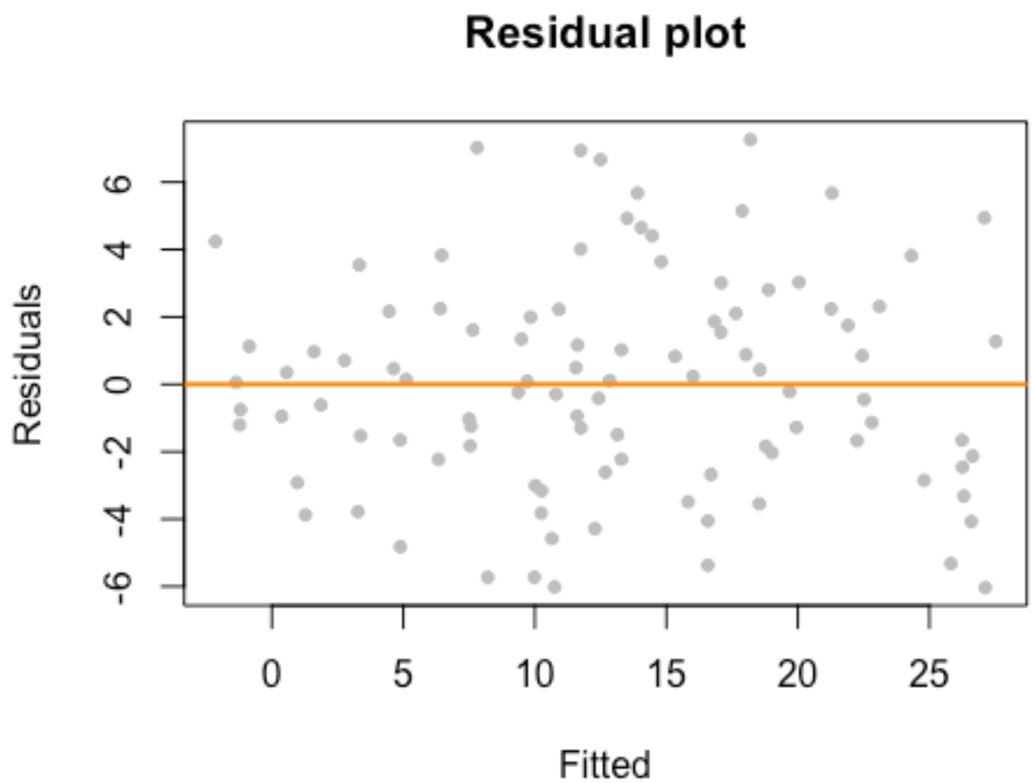
```
# import the data
dataset =
read.table("https://raw.githubusercontent.com/hgweon2/ss3859/master/hw2-data-
1.csv",
           header = TRUE,
           sep = ",",
           quote = "\"",
           comment.char = "",
           stringsAsFactors = FALSE)


##(a)  Given x2 = 50 and x3 = 7, one unit increase in x1 increases the
estimated mean of y by A units. Find A.

# fit_model = beta0_hat + (beta1_hat + beta4_hat*xi2 + beta5_hat*xi3 +
beta7_hat*xi2xi3)xi1 + beta2_hat*xi2 + beta3_hat*xi3 + beta6_hat*xi2xi3

fit_model_dataset = lm(y ~ x1 * x2 * x3, data = dataset)
b = coef(fit_model_dataset)
summary(fit_model_dataset)

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

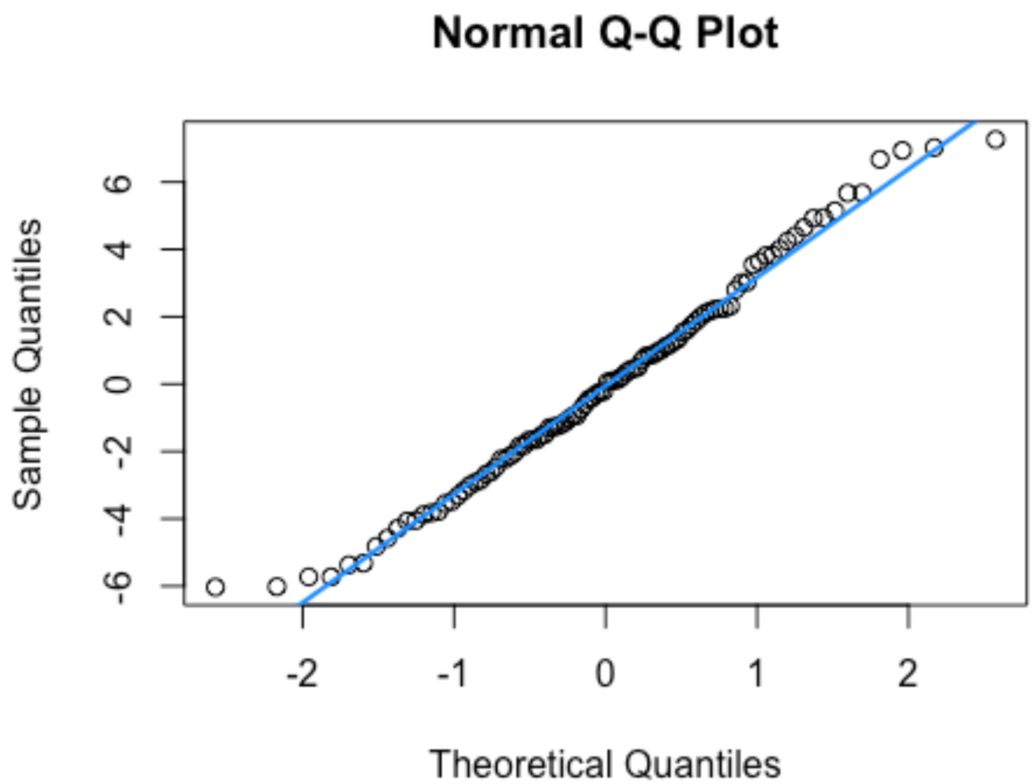
```
# A unit
b[2] + b[5]*50 + b[6]*7 + b[8]*50*7
```

```
##       x1
## 3.995269
```

```
##(b) Obtain the residual plot and normal QQ plot. Check the linearity, equal
variance and normality assumptions.

# Residual plot (fitted vs resid)
plot(fitted(fit_model_dataset), resid(fit_model_dataset), col = "grey", pch =
20,
     xlab = "Fitted", ylab = "Residuals", main = "Residual plot")
abline(h = 0, col = "darkorange", lwd = 2)
```




```
# normal QQ plot
qqnorm(resid(fit_model_dataset))
qqline(resid(fit_model_dataset), col = "dodgerblue", lwd = 2)
```



From the residual plot, it shows that at any area of Y_hat, the spread(variance) of e is roughly the same. Equal variance holds. (no violation) Also, at any of Y_hat, the mean of e is roughly zero. Therefore, the linear assumption holds. From the normal QQ plot, because tails are above the line, normality assumption may be violated.

```
##(c) Check the equal variance and the normality assumptions using
appropriate statistical tests
#install.packages("lmtest")
library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

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

bptest(fit_model_dataset)

##
##  studentized Breusch-Pagan test
##
```

```
## data:  fit_model_dataset
## BP = 6.4252, df = 7, p-value = 0.4911

# We can simply put the lm object as input
# What we are testing: H0: constant variance for all i
# Reject H0 if p_value is small

0.4911>0.05

## [1] TRUE

# We fail to reject null hypothesis.
# Therefore, constant variance for all i is constant. Equal variance
assumption holds.


# Now our interest is
# if the residuals of the reg models follow normal.
# All we need is to use the residuals as input of the function
shapiro.test(resid(fit_model_dataset))

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

0.2875>0.05

## [1] TRUE

# Large p-value -> no evidence against the normal assumption.
```

Therefore, equal variance assumption and normal assumption hold.

```
##(d) Was the three-way interaction term needed?
no_three_way = lm(y ~ x1*x2 + x2*x3 + x1*x3, data = dataset)
summary(no_three_way)

##
## Call:
## lm(formula = y ~ x1 * x2 + x2 * x3 + x1 * x3, data = dataset)
##
## 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 **
## x2:x3       -0.006757   0.002149  -3.145 0.002231 **
## x1:x3        0.042163   0.042737   0.987 0.326411
## ---
## 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(fit_model_dataset)
```

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

```
anova(no_three_way , fit_model_dataset)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 * x2 + x2 * x3 + x1 * x3
## Model 2: y ~ 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
```

```
0.4385 > 0.05
```

```
## [1] TRUE
```

Since p-val is 0.4385 > 0.05, we fail to reject null hypothesis. Therefore, the tree-way interaction term is not needed.

```
##(e) Test the statement “there are no interaction effects between the predictors”
no_intrc_model = lm(y ~ x1 + x2 + x3, data = dataset)
summary(no_intrc_model)

##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7397 -3.0566  0.1003  2.5123  8.4542
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.44998    1.17160   3.798 0.000256 ***
## x1           4.18143    0.25392  16.468  < 2e-16 ***
## x2          -0.14143    0.01367 -10.350  < 2e-16 ***
## x3           0.53417    0.06308   8.468 2.87e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.595 on 96 degrees of freedom
## Multiple R-squared:  0.8272, Adjusted R-squared:  0.8218
## F-statistic: 153.2 on 3 and 96 DF,  p-value: < 2.2e-16

summary(fit_model_dataset)

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

```
anova(no_intrc_model, fit_model_dataset)
```

```
## 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
```

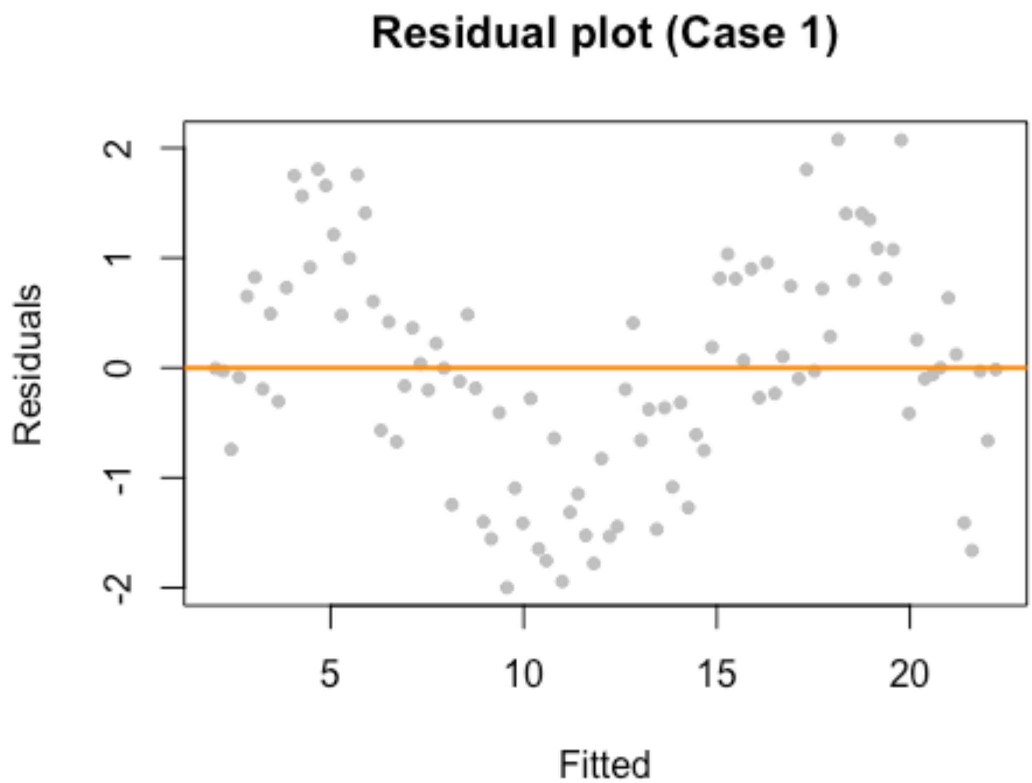
```
0.001297<0.05
```

```
## [1] TRUE
```

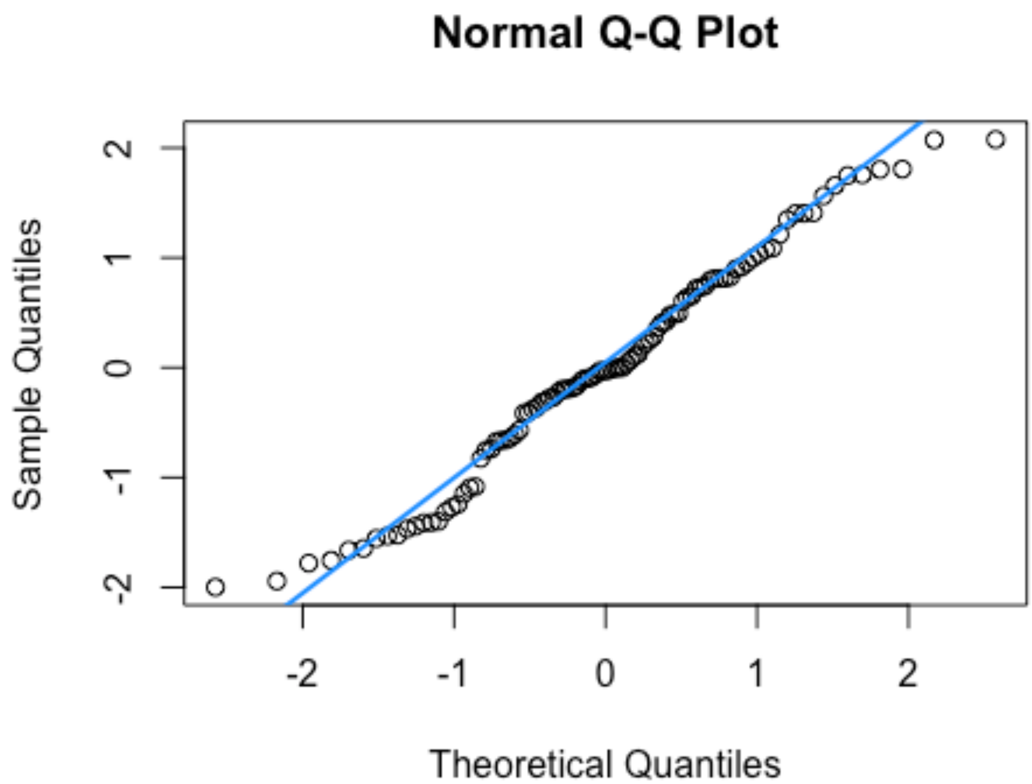
Therefore, we reject null hypothesis. There are interaction effects between the predictors.

##Q4 4. For problem 4, import the data in R from: https://raw.githubusercontent.com/hgweon2/ss3859/master/hw2-data-2.csv The data set contains 100 observations with 2 variables: y (response) and x (predictor). Obtain the fitted model. Using the residual plot and the normal QQ plot, check the linearity, normality and equal variance assumptions. (Justify your answer). Do the Breusch-Pagan test and Shapiro-Wilks test.

```
# import data
dataset2 =
read.table("https://raw.githubusercontent.com/hgweon2/ss3859/master/hw2-data-
2.csv",
                     header = TRUE,
                     sep = ",",
                     quote = "\"",
                     comment.char = "",
                     stringsAsFactors = FALSE)


# obtain fitted model
fit_1 = lm(y ~ x, data = dataset2)

# Residual plot (fitted vs resid)
plot(fitted(fit_1), resid(fit_1), col = "grey", pch = 20,
     xlab = "Fitted", ylab = "Residuals", main = "Residual plot (Case 1)")
abline(h = 0, col = "darkorange", lwd = 2)
```




```r
# normal QQ plot
qqnorm(resid(fit_1))
qqline(resid(fit_1), col = "dodgerblue", lwd = 2)
```




```
#bptest
#install.packages("lmtest")
library(lmtest)
bptest(fit_1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit_1
## BP = 0.0090726, df = 1, p-value = 0.9241
```

```
0.9241 > 0.05
```

```
## [1] TRUE
```

```
#shapiro test
shapiro.test(resid(fit_1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit_1)
## W = 0.97905, p-value = 0.1121
```

```r
0.1121 > 0.05

## [1] TRUE
```

From residual plot, the mean of e varies symptomatically. Therefore, the linear assumption is violated. But the error variance looks pretty same for all data points. Therefore, equal variance may hold.

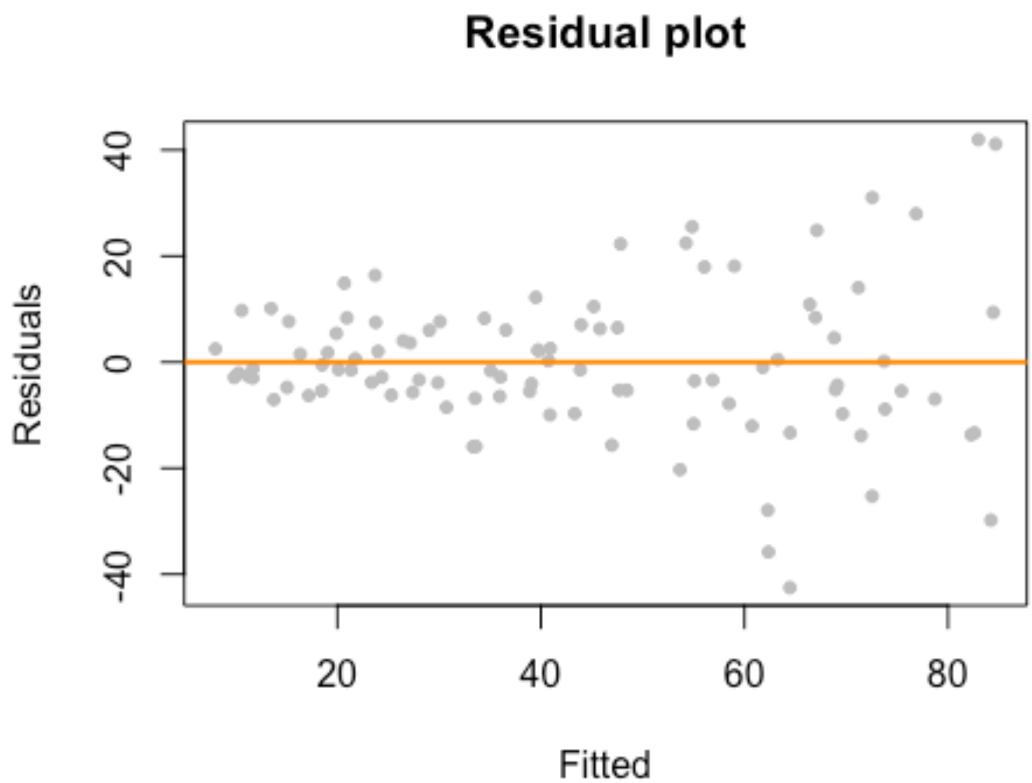
From normal QQ plot, some part of residuals are moved away from the line. Therefore, normality may be violated.

From bptest, p-val = 0.9241 > 0.05. We fail to reject null hypothesis. Therefore, constant variance for all i is constant. Equal variance assumption holds.

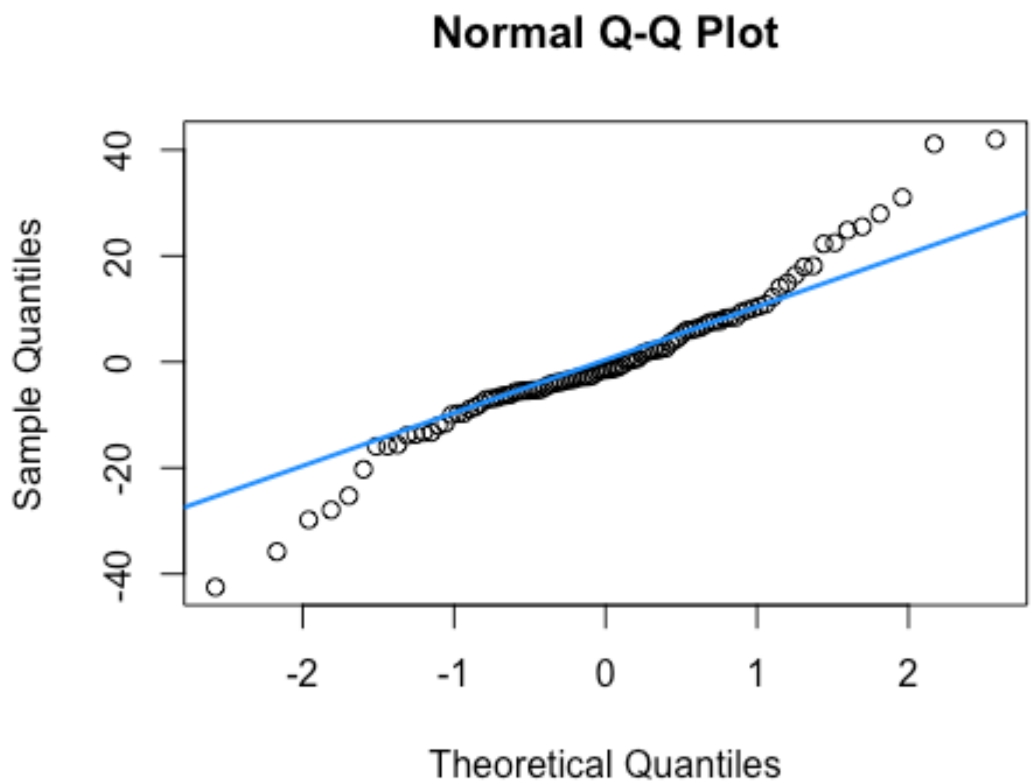
From shapiro test, p-val = 0.1121 > 0.05. Large p-value -> no evidence against the normality assumption.

##Q5 Repeat problem 4. This time, import the data from: https://raw.githubusercontent.com/hgweon2/ss3859/master/hw2-data-3.csv.

```r
# import data
dataset3 =
read.table("https://raw.githubusercontent.com/hgweon2/ss3859/master/hw2-data-
3.csv",
                      header = TRUE,
                      sep = ",",
                      quote = "\"",
                      comment.char = "",
                      stringsAsFactors = FALSE)

# obtain fitted model
fit_2 = lm(y ~ x, data = dataset3)

# Residual plot (fitted vs resid)
plot(fitted(fit_2), resid(fit_2), col = "grey", pch = 20,
     xlab = "Fitted", ylab = "Residuals", main = "Residual plot")
abline(h = 0, col = "darkorange", lwd = 2)
```




```
# normal QQ plot
qqnorm(resid(fit_2))
qqline(resid(fit_2), col = "dodgerblue", lwd = 2)
```




```
#bptest
#install.packages("lmtest")
library(lmtest)
bptest(fit_2)

##
##  studentized Breusch-Pagan test
##
## data:  fit_2
## BP = 22.542, df = 1, p-value = 2.056e-06

#shapiro test
shapiro.test(resid(fit_2))

##
##  Shapiro-Wilk normality test
##
## data:  resid(fit_2)
## W = 0.95913, p-value = 0.003487
```

From residual plot, the mean of e is roughly zero at any area of y_hat. Therefore, the linear assumption holds. However, the spread of e is not constant. Therefore, the equal variance assumption is violated.

From normal QQ plot, both tails are heavy. Normality is violated.

The p-value from the bptest: 2.056e-06 p-val < 0.05 Reject H0 if p_value is small. Therefore, equal variance assumption is violated.

From shapiro test, Small p-value -> the normal assumption is violated