Assignment 2

Yizhou Tang

October 23, 2019

Question 1

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

#a) Obtain the fitted value of mpg at weight = 3, cylinder = 6. (1 pt)

mpgModel = lm(mpg ~ wt + cyl, data = mtcars2)
newdata = data.frame(wt = 3,cyl=as.factor(6))
predict(mpgModel,newdata)
```

```
##        1
## 19.95467
```

```
#Predicted value: 19.95467

#b) Is cyl an important predictor given that wt is used as a predictor?
Answer by conducting an appropriate test at ?? = 0.05. (1 pt)

# Test H0: beta_am = 0 vs H1: beta_am != 0
summary(mpgModel)
```

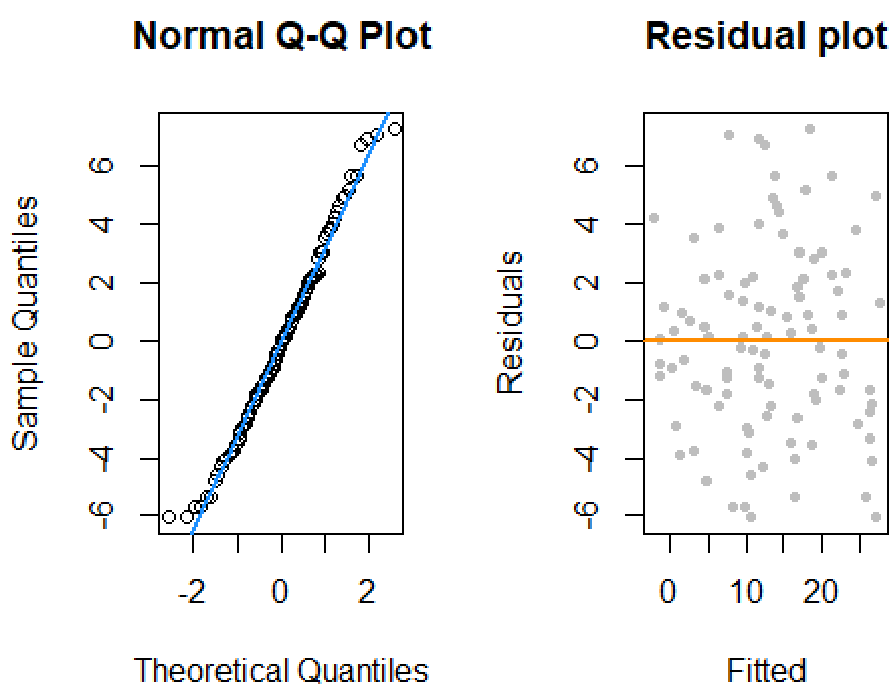
```
##
## Call:
## lm(formula = mpg ~ wt + cyl, data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8544 -1.7440 -0.4468  1.2646  6.6174
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  32.8988     2.3390  14.065  3.7e-12 ***
## wt           -3.0606     0.9136  -3.350  0.00303 **
## cyl6         -3.7623     1.7639  -2.133  0.04490 *
## cyl8         -5.4415     1.8085  -3.009  0.00668 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.579 on 21 degrees of freedom
```

```
## Multiple R-squared:  0.8129, Adjusted R-squared:  0.7862
## F-statistic: 30.42 on 3 and 21 DF,  p-value: 7.818e-08

# Since we have low p-values for both beta (<0.05) --> Reject H0 --> Using
two fitted lines gives a much better fit. Hence cyl is an important predictor

#c) Obtain the fitted value of mpg at weight = 3, cylinder = 8. (1 pt)

mpgModel2 = lm(mpg ~ wt + cyl + cyl:wt, data = mtcars2)
newdata = data.frame(wt = 3,cyl=as.factor(8))
predict(mpgModel2,newdata)

##        1
## 17.10022

#Predicted value: 18.27539

#(d) Test the null hypothesis: "There is no significant interaction effect
between two predictors." Use the significance level ?? = 0.05. (1 pt)

# Include am (dummy) without interaction
summary(mpgModel2)

##
## Call:
## lm(formula = mpg ~ wt + cyl + cyl:wt, 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

# The interaction effect is not significant since the p-value> alpha, null
hypothesis is not rejected
```

Question 2

```r
h2data =
read.csv("https://raw.githubusercontent.com/hgweon2/ss3859/master/hw2-data-
1.csv")

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

#model = lm(y ~ x1 + x2 + + x3 + x1:x2 + x1:x3 + x2:x3 + x1*x2*x3, data =
h2data)
model = lm(y ~ x1*x2*x3, data = h2data)
summary(model)

##
## Call:
## lm(formula = y ~ x1 * x2 * x3, data = h2data)
##
## 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

#Retrieve the coefficients
b0 = summary(model)$coefficients[1, 1]
b1 = summary(model)$coefficients[2, 1]
b2 = summary(model)$coefficients[3, 1]
b3 = summary(model)$coefficients[4, 1]
b4 = summary(model)$coefficients[5, 1]
b5 = summary(model)$coefficients[6, 1]
b6 = summary(model)$coefficients[7, 1]
b7 =summary(model)$coefficients[8, 1]
x2 = 50
x3 = 7
```

```
A = b1 + b4*x2 + b5*x3 + b7*x2*x3
A

## [1] 3.995269

#A is 3.995269

#(b) Obtain the residual plot and normal QQ plot. Check the linearity, equal
variance and normality assumptions. (1 pt)

#QQ norm
par(mfrow=c(1,2)) # Combining plots
qqnorm(resid(model))
qqline(resid(model), col = "dodgerblue", lwd = 2)

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



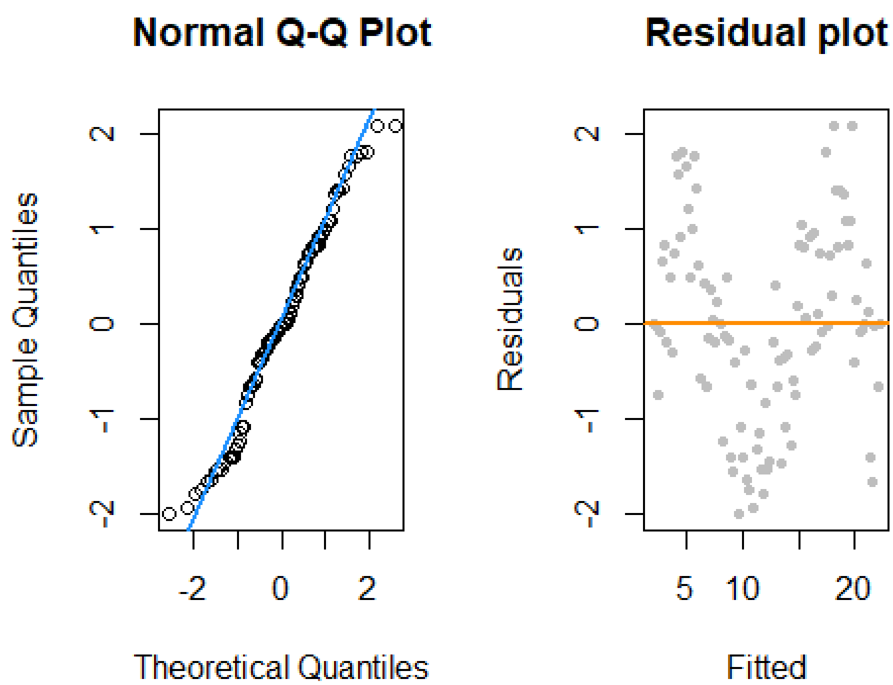

```
#Normality is not vioated: the observations follow very close to the normal
distribution, according to the Normal QQ plot.(However, there is a slight
difference in the tails, which should be kept in mind when working with the
model)

#Linearity is not violated, because residual plot shows mean of e does not
varies systematically, it is also roughly at 0.
```

```r
#Equal variance is not violated, because the spread of e does appear to be constant

#(d) Was the three-way interaction term needed? Why/why not? (1 pt)

summary(model)

##
## Call:
## lm(formula = y ~ x1 * x2 * x3, data = h2data)
##
## 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

#The three way interaction term was not needed, because at alpha = 0.05, we can see that the p-value is high, 0.4385.

#(e) Test the null hypothesis: ??4 = ??5 = ??6 = ??7 = 0 at ?? = 0.05. (2 pt)

# Calculate reduced model, compare to the full model we already have
reducedModel = lm(y ~ x1+x2+x3, data = h2data)

anova(reducedModel, model)

## 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

# Since p-value is small (<0.05), null hypothesis is rejected.
```

Question 3:

```
q3data =
read.csv("https://raw.githubusercontent.com/hgweon2/ss3859/master/hw2-data-
2.csv")

#Obtain fitted model:
SLRModel = lm(y ~ x, data = q3data)

#QQ norm
par(mfrow=c(1,2)) # Combining plots
qqnorm(resid(SLRModel))
qqline(resid(SLRModel), col = "dodgerblue", lwd = 2)

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

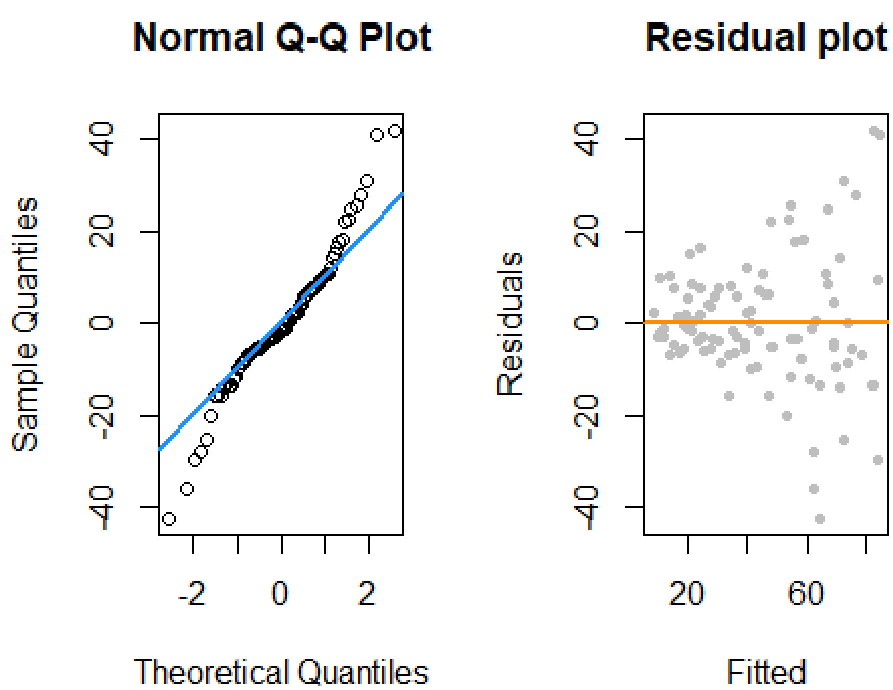



```
#Normality is vioated: the tails of the distribution clearly differs from the
normal distribution, according to the Normal QQ plot. The observations also
does not appear to be a perfect straight line.
```

```
#Linearity is  violated, mean of e varies systematically.

#Equal variance is not violated, because the spread of e does appear to be constant, the residual plot is showing a v shape.
```

Question 4:

```
q4data = read.csv("https://raw.githubusercontent.com/hgweon2/ss3859/master/hw2-data-3.csv")

#Obtain fitted model:
SLRModel = lm(y ~ x, data = q4data)

#QQ norm
par(mfrow=c(1,2)) # Combining plots
qqnorm(resid(SLRModel))
qqline(resid(SLRModel), col = "dodgerblue", lwd = 2)

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




```
#Normality is vioated: The observed values is not a straight line, the tails of the distribution differs from the normal distribution, according to the
```

```
Normal QQ plot.

#Linearity is not violated, the mean doet not vary systematically, according
to the residual plot.

#Equal variance is violated, because the spread of e does not appear to be
constant, according to the residual plot.
```

Question 5

```
xobs  = c(25,23,5,20,35,18,17,15,14,20)
yobs = c(85,120,20,64,50,84,50,26,36,60)
resi = c(14.49,53.29,-12.55,2.98,-39.49,26.78,-5.32,-25.53,-13.63,-1.02)
leverages= c(0.16,0.13,0.47,0.10,0.55,0.10,0.11,0.13,0.15,0.10)# = diag(H)
p = sum(leverages) # equals to p
n = 10
#build a df based on the observed values
dframe = data.frame(y = yobs,x=xobs)

#(a) Is there any observation that has a high leverage (higher than 2p/n)? If
so, what are they? (1 pt)


#Check if any obs with high leverage
leverages > 2 * p/n

##  [1] FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE

# Yes, there exists observations with high leverage. The observations are
0.47 and 0.55.

#b)
#If Y for observation B changes to 50, the leverage stays 0.13

#c)

lev_fit = lm(y~.,data = dframe)
# checking outliers
rstandard(lev_fit)[c(2,3,5,8)] #standardized residuals for B,C,E,H

##          2          3          5          8
##  2.0218823 -0.6087305 -2.0939853 -0.9718407

#d)
# Cook's distance
temp = cooks.distance(lev_fit)[c(2,3,5,8)]
temp > 4 /n

##     2     3     5     8
## FALSE FALSE  TRUE FALSE
```

```
# E is an influential point
```