Quiz1

Nam Jun Lee

09/22/2021

a. Import the data set in R.

```r
# dataset into dat
dat <- read.csv("Quiz1data-2.csv")
```

Read in the Quiz 1 data csv file from working directory.

b. The data contains the variables y, X1, X2 and X3. The objective of this problem is to predict the response y based on X1, X2 and X3 and to determine which variables are significantly associated with the response. Perform a multiple regression to answer this question. Provide a prediction at X1 = 0.25, X2 = 0.5, X3 = 0 and compute the corresponding confidence and prediction intervals.

```r
# multiple regression
lm.fit <- lm(y ~ X1 + X2 + X3, data = dat)
# summary multiple regression
summary(lm.fit)
```

```
##
## Call:
## lm(formula = y ~ X1 + X2 + X3, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0418  -0.8509  -0.2402   0.5012  21.6247
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.2251     0.2082   5.884 2.63e-08 ***
## X1            1.2972     0.6970   1.861   0.0647 .
## X2           -1.1943     0.7249  -1.647   0.1016
## X3            0.7232     0.1753   4.125 6.21e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.525 on 146 degrees of freedom
## Multiple R-squared:  0.1792, Adjusted R-squared:  0.1623
## F-statistic: 10.62 on 3 and 146 DF,  p-value: 2.322e-06
```

```r
# prediction interval at X1 = 0.25, X2 = 0.5, X3 = 0
predict(lm.fit, data.frame(X1 = c(0.25), X2 = c(0.5), X3 = c(0)), interval = "prediction")
```

```
##         fit       lwr      upr
## 1 0.9522809 -4.072724 5.977286
```

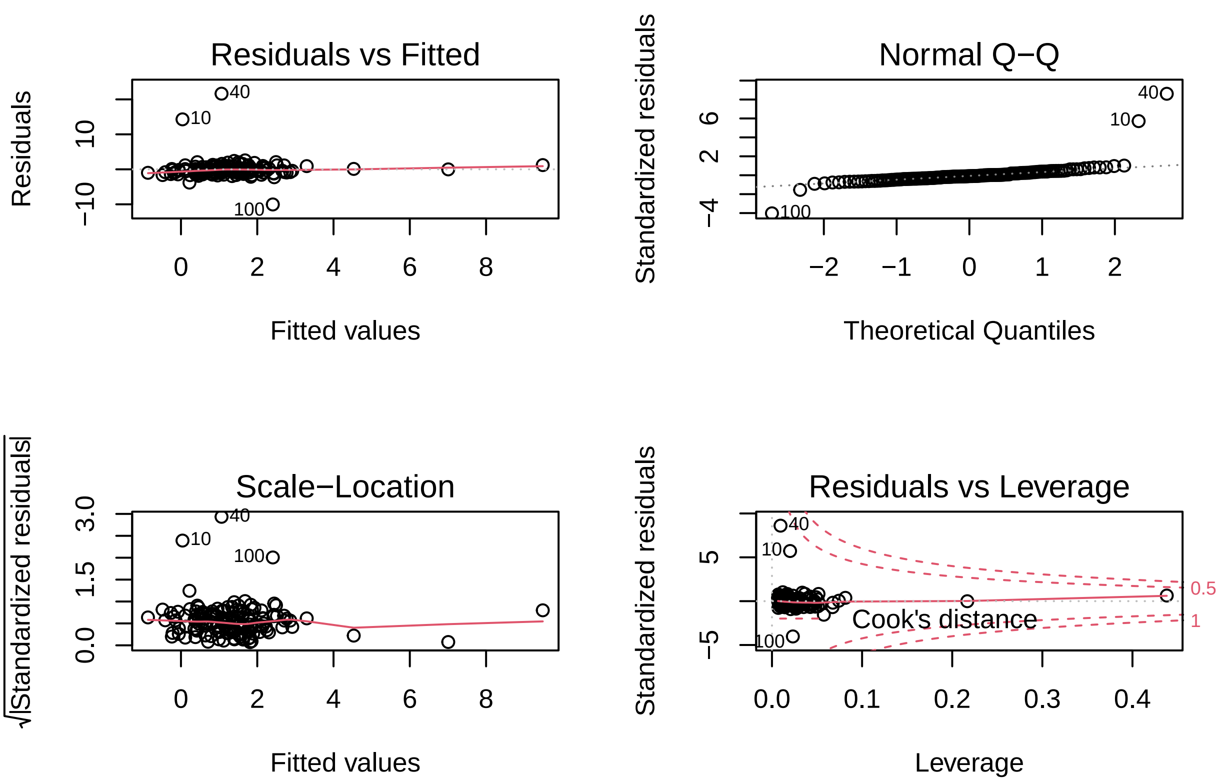
```r
# confidence interval at X1 = 0.25, X2 = 0.5, X3 = 0
predict(lm.fit, data.frame(X1 = c(0.25), X2 = c(0.5), X3 = c(0)), interval = "confidence")
```

```
##         fit       lwr      upr
## 1 0.9522809 0.3543294 1.550232
```

It may reject the null hypothesis for **X3** at the 0.001 level. each **X1**, **X2** predictor are not statistically significant to the response variable. Also RSE is **2.525** and Adjusted R-squared is **16.23%**. The distance between the prediction interval and the confidence interval is wide.

c. Analyse the residuals to detect potential problems with your analysis in part (b).

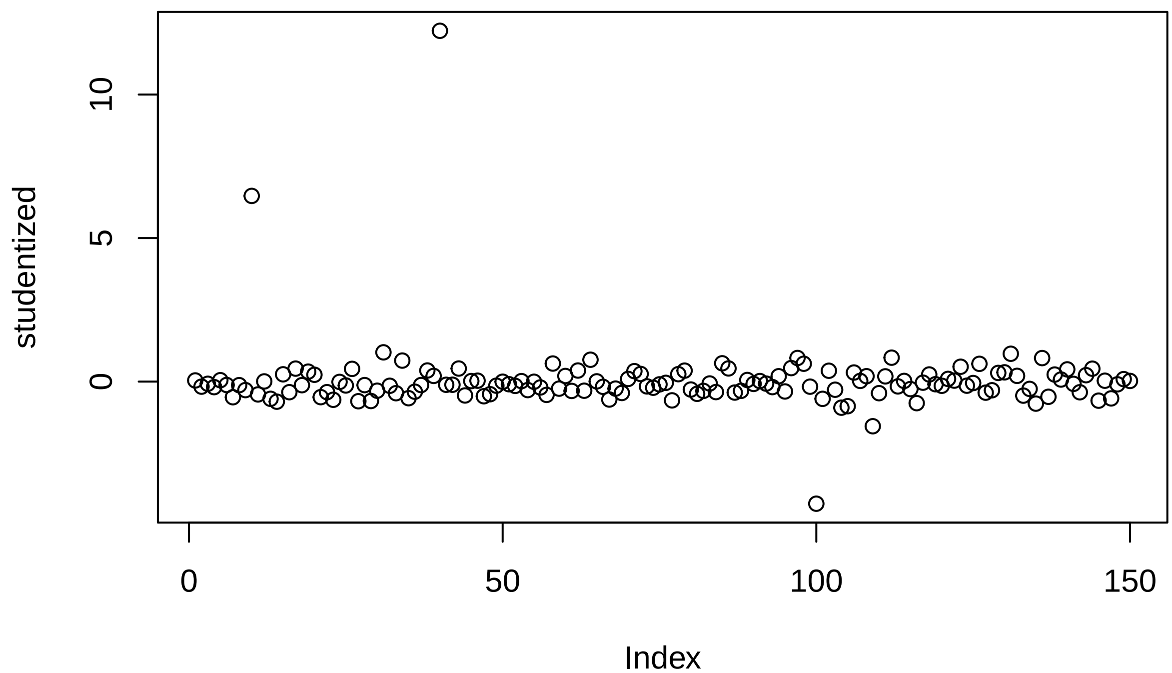
```r
# diagnostic plots of the least squares regression
par(mfrow = c(2, 2))
plot(lm.fit)
```



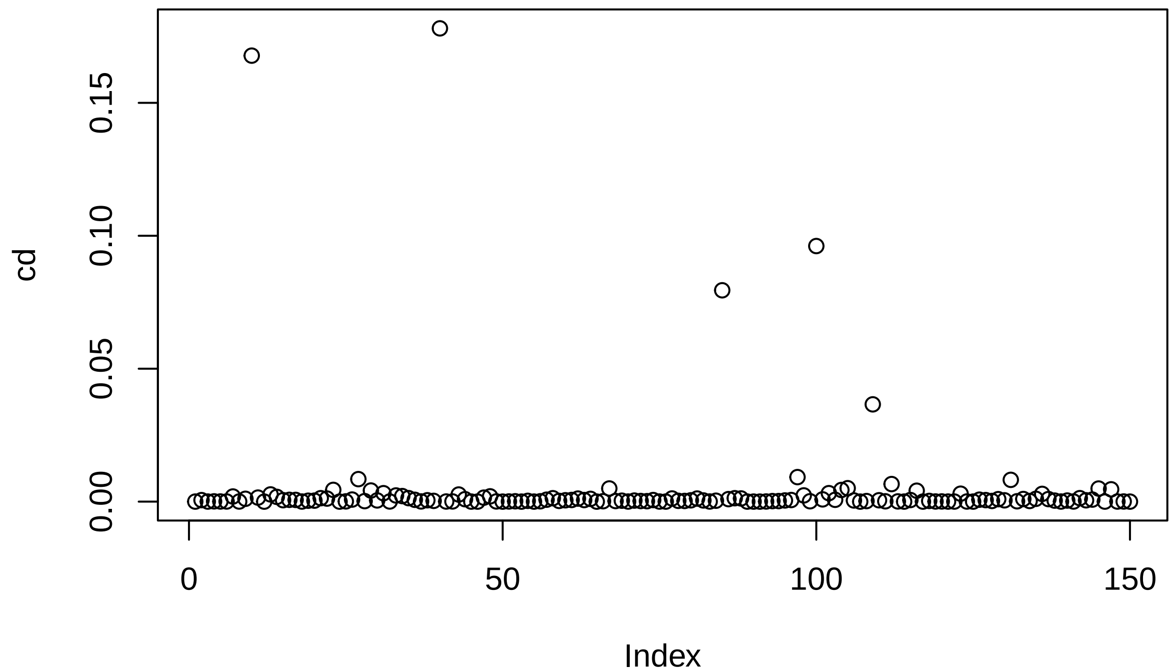
The **residual versus fitted** plot does not follow a normal distribution with constant variance. **Normal Q-Q** plot shows that the residuals follow the standard deviation well, but there are several outliers. Also, there are several outliers in the **scale location** plot, and the distribution is skewed to one side. The plot of the **residuals versus leverage** shows that the residual exceeds cook's distance.

d. Propose solutions to the problems that you detect in part (c) and implement them on the data set. [Hint: studentized residual can be computed using the function studres(), cooks distance using cooks.distance() and the variance inflation factor using vif(). The function vif() is part of the R package car which you may need to install and unpack].

```r
# find outliers
studentized <- studres(lm.fit)
plot(studentized)
```




```r
# find high leverage points
cd <- cooks.distance(lm.fit)
plot(cd)
```




```
# check variance inflation factor
vif(lm.fit)
```

```
##        X1        X2        X3
## 22.190320 23.046548  1.478556
```

```
# outlier (> 3)
outlier_stu <- which(abs(studentized) > 3)
# total row in dat
n <- nrow(dat)
# high leverage (>4/n)
leverage_high <- which(cd > 4/n)

# union outliers and high leverage points
remove_out <- union(outlier_stu, leverage_high)
# show outliers and high leverage points
remove_out
```

```
## [1]  10  40 100  85 109
```

```
# remove outliers and high leverage points into newdata
newdata <- dat[-remove_out, ]
```

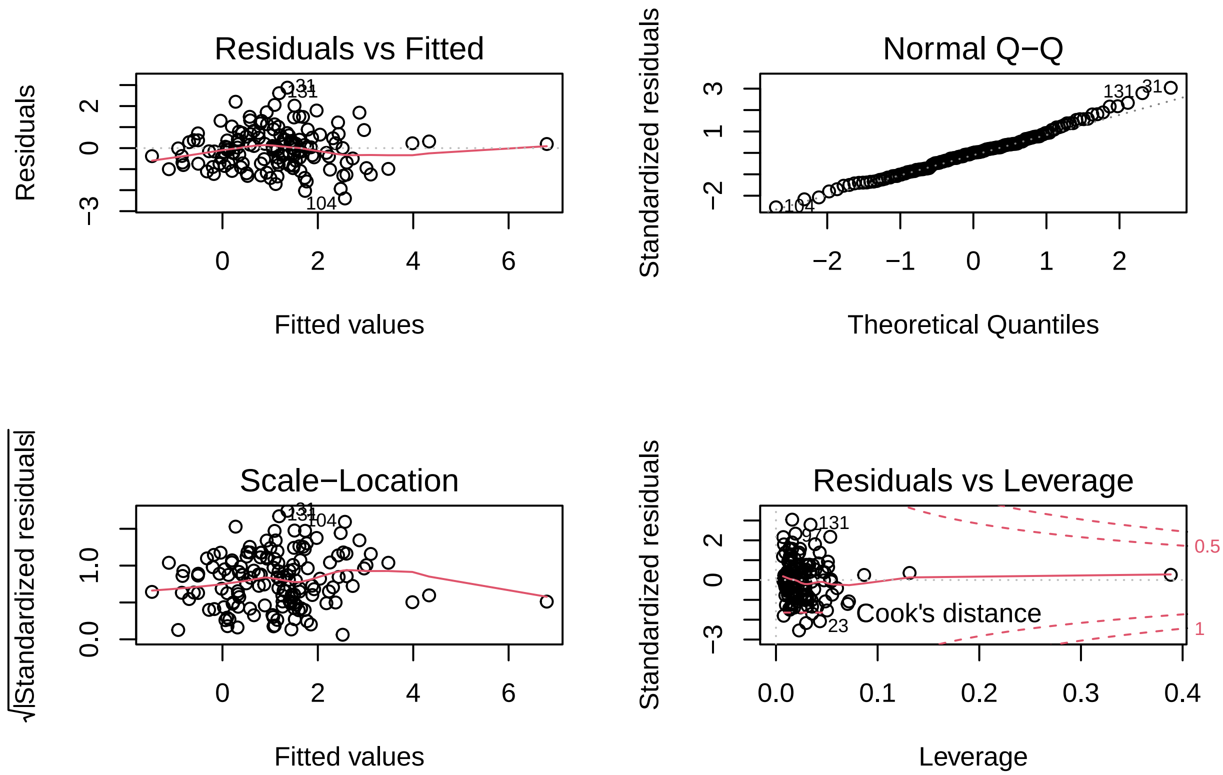
As a result of checking the multicollinearity, it can be confirmed that X1 and X2 were high. first, the outliers and high leverage detected in part c are remove, the data set is to be implemented again. Using the studres() function, remove residuals exceeding -3 and 3 and using the cooks.distance function to remove leverage exceeding the distance. And entered it in the new data, 'newdata'.

e. Rerun your analysis of part (a) on the data that you obtain from part (d)

```
# data that obtain from part (d)
lm.fit1 <- lm(y ~ X1 + X2 + X3, data = newdata)
summary(lm.fit1)
```

```
##
## Call:
## lm(formula = y ~ X1 + X2 + X3, data = newdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.39489 -0.68839  0.00361  0.49400  2.87541
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.04362    0.07994  13.055  < 2e-16 ***
## X1           1.65116    0.26698   6.185 6.35e-09 ***
## X2          -1.74576    0.28140  -6.204 5.77e-09 ***
## X3           0.91696    0.07222  12.697  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9532 on 141 degrees of freedom
## Multiple R-squared:  0.5874, Adjusted R-squared:  0.5786
## F-statistic: 66.91 on 3 and 141 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(lm.fit1)
```



It may reject the null hypothesis for all **X1**, **X2**, and **X3** at the 0.001 level. As such, it can be seen that the predictor is statistically significant for the response variable. Also RSE is **0.9532** and Adjusted R-squared is **57.86%**.

f. Provide a prediction at X1 = 0.25, X2 = 0.5, X3 = 0 and compute the corresponding confidence and prediction intervals. Compare with the prediction in part (a) and comment on which you think is more believable.

```
# prediction interval new data
predict(lm.fit1, data.frame(X1 = c(0.25), X2 = c(0.5), X3 = c(0)), interval = "prediction")
```

```
##         fit       lwr      upr
## 1 0.5835313 -1.315572 2.482635
```

```
# confidence interval new data
predict(lm.fit1, data.frame(X1 = c(0.25), X2 = c(0.5), X3 = c(0)), interval = "confidence")
```

```
##         fit       lwr       upr
## 1 0.5835313 0.3478518 0.8192108
```

Compared to the prediction in part (a), **I think the newly constructed model (model 2) is more reliable**. Because in the newly constructed model, all predictors are statistically significant to the response variable. And the RSE was significantly lower and the adjusted R-squared increased. This is also because the distance between the prediction interval and the confidence interval is much narrower.