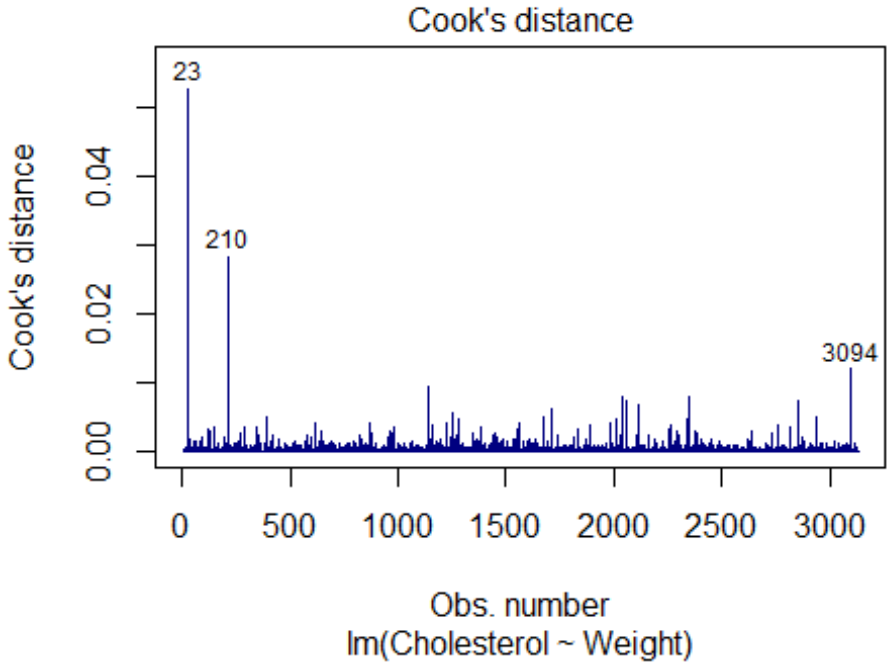
# STA 6443 - HW3 solution

Yeonjoo Park

Exercise 1.

(a) Since a few observations with of Cook's distance greater than 0.015 are detected, we refit the model without them. Specifically they are observations 23 and 210.

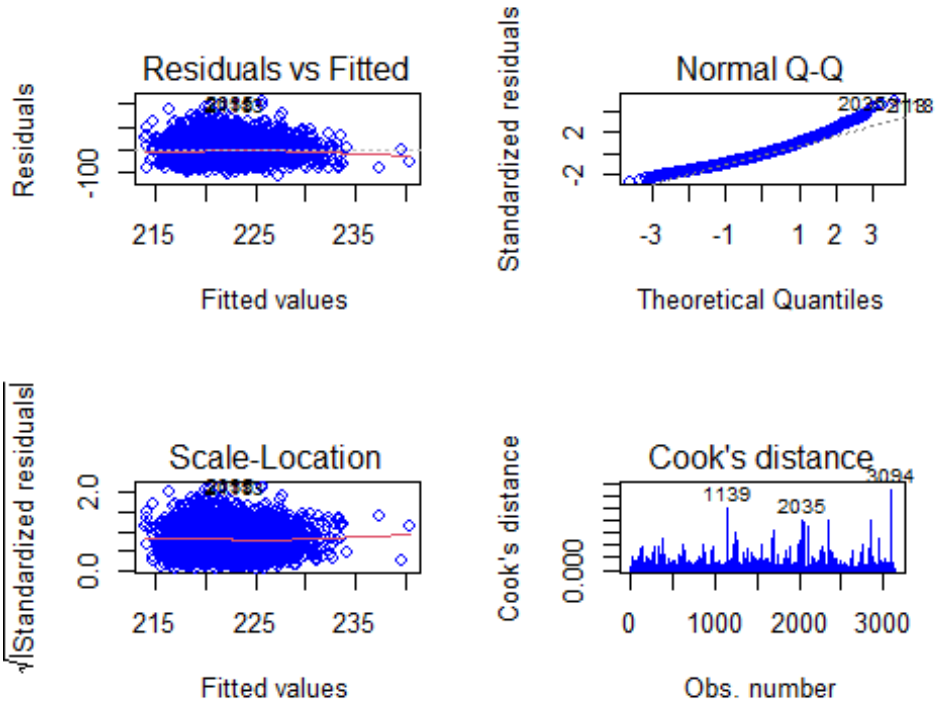



(b) The final model is significant with very small p-value from F-test. In other words, the coefficient of weight is non-zero. The slight non-normal behavior is observed in QQ-plot but there are no obvious remaining diagnostic. No more observations with Cook's D larger than 0.015. The model tells us that an increase of weight is associated with an increase of cholesterol. On average, the cholesterol is predicted to have an increase of 0.122 units when weight increases by 1 unit.
Only around 0.63% of the variation of cholesterol is explained by weight, implying that it is not be a good-fit model for cholesterol.

```
inf.id=which(cooks.distance(lm.heart)>0.015)
lm.heart.refit=lm(Cholesterol~Weight, data=heart[-inf.id,])

summary(lm.heart.refit)

##
## Call:
## lm(formula = Cholesterol ~ Weight, data = heart[-inf.id, ])
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -112.369  -29.395   -4.482   23.672  209.348
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 203.57605    4.18543  48.639  < 2e-16 ***
## Weight        0.12264    0.02745   4.469 8.16e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.92 on 3130 degrees of freedom
## Multiple R-squared:  0.006339,   Adjusted R-squared:  0.006022
## F-statistic: 19.97 on 1 and 3130 DF,  p-value: 8.155e-06
```

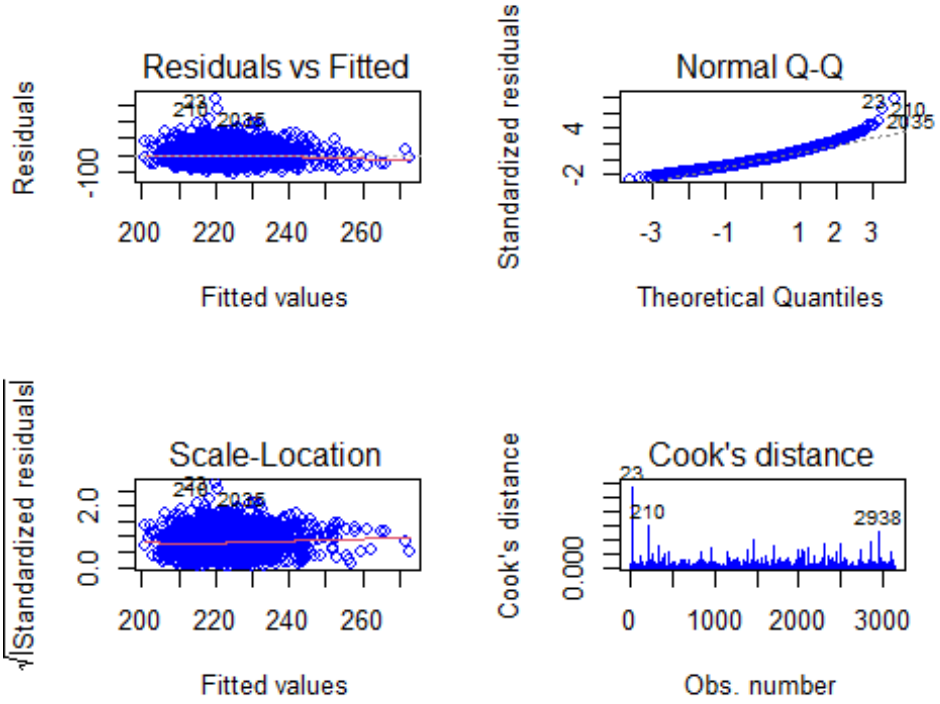
```
par(mfrow=c(2,2))
plot(lm.heart.refit, which=1:4,  col="blue")
```




Exercise 2.

(a) We again observe two observations (same data points in Exercise 1) with cook's D greater than 0.015. The model is refitted without them. In terms of diagnostics plots, QQ-plot does not show a perfect line but it does not look like serious abnormality. For remedy of skewness, we can try transformation of Y, although it is not addressed here.

```
lm.heart2 <- lm(Cholesterol~Weight+Diastolic+Systolic, data=heart)

par(mfrow=c(2,2))
plot(lm.heart2, which=1:4, col="blue")
```




```
inf.id=which(cooks.distance(lm.heart2)>0.015)
lm.heart2.refit=lm(Cholesterol~Weight+Diastolic+Systolic, data=heart[-inf.id,])
```

(b) The weight is not significant but diastolic and systolic are both significant with p-values less than 0.05 in t-test. If the bottom number in the blood pressure (Diastolic) and weight are fixed, the increase of 1 unit in top number of blood pressure (Systolic ) would predict an increase of 0.30 units of cholesterol on average. Similarly, if the top number (Systolic) and weight are unchanged, 1 unit increase of bottom number (Diastolic) predicts an increase of 0.25 units in cholesterol on average. There is no multicollinearity issue among the predictors since the VIFs are all less than 10 and the scatterplot matrix does not indicate very strongly correlated predictors.

The R-square is 3.77%, meaning that only 3.77% variance of cholesterol is explained, which is pretty low. So it is not a good model for predicting cholesterol levels.

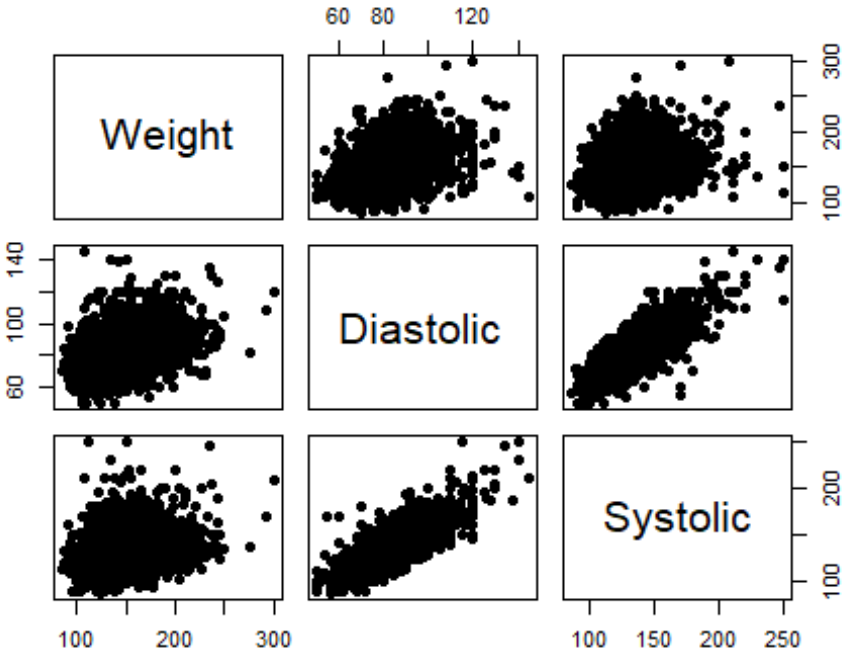
```
summary(lm.heart2.refit)

##
## Call:
## lm(formula = Cholesterol ~ Weight + Diastolic + Systolic, data = heart[-inf.id,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -110.617  -29.371   -4.476   23.755  216.041
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.32618    6.27153  24.926  < 2e-16 ***
## Weight        0.03671    0.02860   1.284   0.1994
## Diastolic     0.24922    0.10665   2.337   0.0195 *
## Systolic      0.30073    0.06340   4.743  2.2e-06 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.26 on 3128 degrees of freedom
## Multiple R-squared:  0.03767,    Adjusted R-squared:  0.03675
## F-statistic: 40.81 on 3 and 3128 DF,  p-value: < 2.2e-16
```

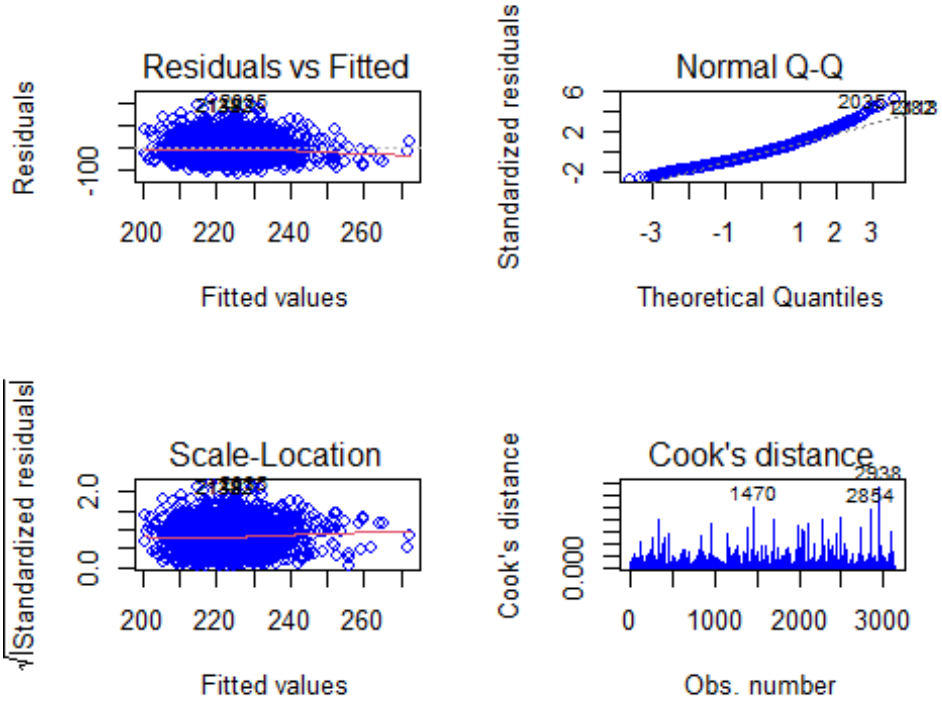
```
vif(lm.heart2.refit)
```

```
##    Weight Diastolic  Systolic
##  1.120631  2.558914  2.454207
```

```
pairs(heart[,c(1:3)], pch = 19)
```




```
par(mfrow=c(2,2))
plot(lm.heart2.refit, which=1:4,  col="blue")
```




Exercise 3.

(a) The stepwise selection is performed and the final model includes Systolic and Diastolic. Since all the observations are of Cook's distance less than 0.015, there is not any highly influential point. Similarly, a bit skewed histogram and non-linear QQ plot are detected, but they don't seem to be serious.

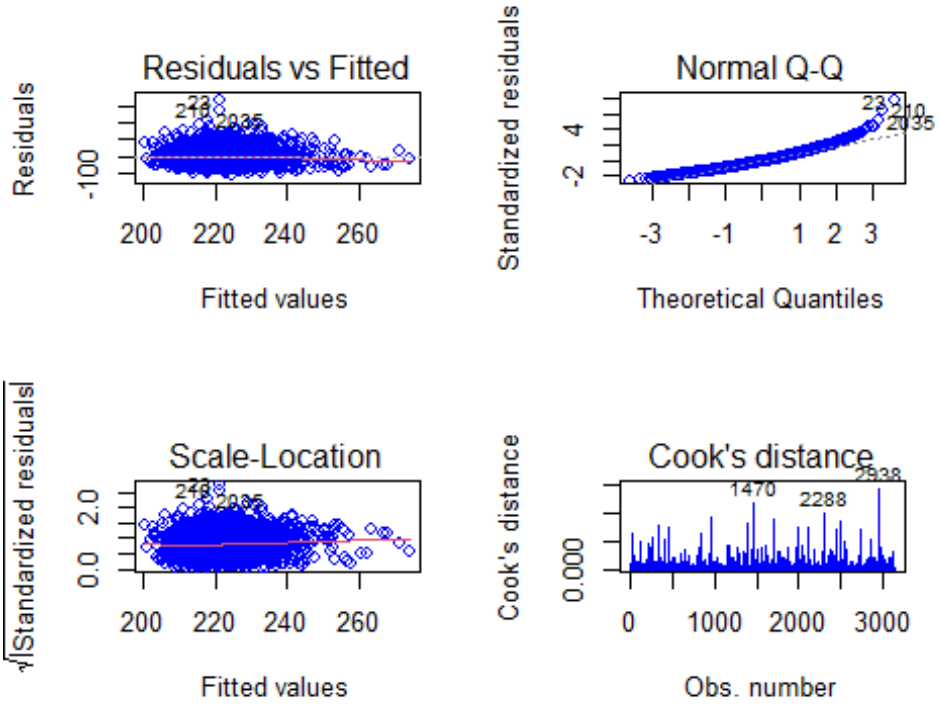
```
lm.full=lm(Cholesterol~Weight+Diastolic+Systolic, data=heart[-inf.id, ])
model.stepwise<-ols_step_both_p(lm.full, pent = 0.05, prem = 0.05, details = F)
model.stepwise
```

```
##
##                               Stepwise Selection Summary
## ----------------------------------------------------------------------------------------
##                         Added/                   Adj.
## Step    Variable       Removed     R-Square    R-Square     C(p)        AIC         RMSE
## ----------------------------------------------------------------------------------------
##    1    Systolic       addition       0.035       0.035    8.6850    32349.7666    42.3013
##    2    Diastolic      addition       0.037       0.037    3.6480    32344.7321    42.2606
## ----------------------------------------------------------------------------------------
```

```
lm.final.step <- lm(Cholesterol~Systolic+Diastolic, data=heart)
summary(lm.final.step)
```

```
##
## Call:
## lm(formula = Cholesterol ~ Systolic + Diastolic, data = heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -109.52  -29.58   -4.57   23.79  328.47
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 159.63995    5.91244  27.001  < 2e-16 ***
## Systolic      0.30193    0.06442   4.687 2.89e-06 ***
```

```
## Diastolic     0.27609    0.10612   2.602  0.00932 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.94 on 3131 degrees of freedom
## Multiple R-squared:  0.03589,    Adjusted R-squared:  0.03527
## F-statistic: 58.27 on 2 and 3131 DF,  p-value: < 2.2e-16

#diagnostics plots
par(mfrow=c(2,2))
plot(lm.final.step, which=1:4 , col= "blue")
```




(b) An increase of 1 unit in Systopic predicts an increase of 0.3 units of cholesterol on average, when Diastolic is fixed. An increase of 1 unit in Diastolic predicts an increase of 0.28 units of cholesterol on average, when Systolic is fixed. The R-square is 3.50%, meaning that only 3.50% of the variation of the cholesterol variable is explained by this model, which is pretty low. It is not a good-fit model for cholesterol.

This model's percentage of variation explained is much larger than that of Exercise 1, but smaller than the model in Exercise 2 with less number of predictors.

Exercise 4.

(a) Based on adjusted-R square criteria, model 3 is the final model with the largest measure of 0.0367. It is the model with all three variables, Weight, Diastolic, and Systolic.

(b) Based on AIC criteria, model 2 is the final model with the smallest AIC of 32344.73. It is the model with two variables, Diastolic and Systolic.

(c) Final models from adjusted-R square and AIC criterion are different. The model from adjusted-R square criteria is same for the final model based on stepwise selection.

(extra comments - if we need to choose one final model in practice, it can be either model 2 or model 3 based on your justification. But I would prefer model 2 as adding one more variable, weight, in model 3 does not have a clear improvement, but just 0.0005 increase in R-square. Adding one more variable did not increase prediction power that much so I would rather choose the simpler model, model 2)

```
lm.full=lm(Cholesterol~Weight+Diastolic+Systolic, data=heart[-inf.id, ])
model.best.subset<-ols_step_best_subset(lm.full)
model.best.subset
```

```
##         Best Subsets Regression
## ----------------------------------------
## Model Index    Predictors
## ----------------------------------------
##      1         Systolic
##      2         Diastolic Systolic
##      3         Weight Diastolic Systolic
## ----------------------------------------
##
##                                                                Subsets Regression Summary
## ----------------------------------------------------------------------------------------------------------------------
----------------------------------
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square     C(p)        AIC           SBIC          SBC            MSEP
FPE        HSP       APC
## ----------------------------------------------------------------------------------------------------------------------
----------------------------------
##   1        0.0350      0.0347      0.0337    8.6847    32349.7666    23461.5297    32367.9149    5604396.2
122    1790.5412    0.5719    0.9662
##   2        0.0372      0.0365      0.0352    3.6475    32344.7321    23456.5056    32368.9298    5593610.3
978    1787.6653    0.5710    0.9647
##   3        0.0377      0.0367      0.0351    4.0000    32345.0829    23456.8621    32375.3300    5592453.6
261    1787.8655    0.5710    0.9648
## ----------------------------------------------------------------------------------------------------------------------
----------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```