**Homework 5**

Jiaqi Li

**A.**

From the summary statistics of the full model, we can obtain that predictor “Over45” has the largest p-value = 0.88046 > alpha-to-remove = 0.15. Thus, we remove “Over45” from the full model at the first step.

Then, we check the summary statistics again and find out that “KneeC” has the largest p-value = 0.7738 > alpha-to-remove = 0.15. Thus, we remove “KneeC” from the model.

Then, we check the summary statistics again and find out that “ThighC” has the largest p-value = 0.52389 > alpha-to-remove = 0.15. Thus, we remove “ThighC” from the model.

Then, we check the summary statistics again and find out that “Weight” has the largest p-value = 0.575914 > alpha-to-remove = 0.15. Thus, we remove “Weight” from the model.

Then, we check the summary statistics again and find out that “ForearmC” has the largest p-value = 0.31165 > alpha-to-remove = 0.15. Thus, we remove “ForearmC” from the model.

Then, we check the summary statistics again and find out that “AnkleC” has the largest p-value = 0.27225 > alpha-to-remove = 0.15. Thus, we remove “AnkleC” from the model.

Then, we check the summary statistics again and find no other predictors that have p-value larger than alpha-to-remove = 0.15.

**Now, we obtain the model from backward elimination:**

bodyfat.percentage = 21.0251 -0.3758\*Height -0.3672\*NeckC -0.1856\*ChestC + 0.9978\*AbdomenC -0.1920\*HipC + 0.3319\*BicepsC -1.4628\*WristC

**B.**

First, we set up an empty model with no predictor.

Then, we check the p-value for each predictor individually:

> pvalue.for1

Over45 Weight Height NeckC ChestC AbdomenC HipC ThighC KneeC

3.229072e-02 1.207304e-25 4.447621e-01 3.680961e-15 1.947170e-35 1.121969e-59 1.296390e-27 2.713443e-20 1.630623e-15

AnkleC BicepsC ForearmC WristC

1.132067e-04 2.046233e-15 9.947490e-09 3.477863e-07

We add “AbdomenC” into the model since it has the smallest p-value = 1.121969e-59 < alpha-to-enter = 0.15.

> pvalue.for2

Over45 Weight Height NeckC ChestC HipC ThighC KneeC AnkleC

4.836650e-01 2.162394e-10 1.704152e-08 1.329576e-06 1.840764e-04 1.382336e-05 2.033388e-02 9.819230e-05 8.641085e-03

BicepsC ForearmC WristC

3.766466e-02 1.370515e-02 1.643497e-09

We add “Weight” into the model since it has the smallest p-value = 2.162394e-10 < alpha-to-enter = 0.15.

> pvalue.for3

Over45 Height NeckC ChestC HipC ThighC KneeC AnkleC BicepsC

0.0840644225 0.1115683993 0.0487609414 0.3363419540 0.7685653134 0.0288935406 0.8650559575 0.5466249851 0.0536610258

ForearmC WristC

0.3019235997 0.0005950752

We add “WristC” into the model since it has the smallest p-value = 0.0005950752 < alpha-to-enter = 0.15.

> pvalue.for4

Over45 Height NeckC ChestC HipC ThighC KneeC AnkleC BicepsC ForearmC

0.60363794 0.08931123 0.42711420 0.44242323 0.73203069 0.12522783 0.57081871 0.17464573 0.02035225 0.07923609

We add “BicepsC” into the model since it has the smallest p-value = 0.02035225 < alpha-to-enter = 0.15.

> pvalue.for5

Over45 Height NeckC ChestC HipC ThighC KneeC AnkleC ForearmC

0.6954545 0.2729983 0.2073643 0.3112601 0.7398100 0.3270078 0.4995375 0.1292198 0.3027216

We add “AnkleC” into the model since it has the smallest p-value = 0.1292198 < alpha-to-enter = 0.15.

> pvalue.for6

Over45 Height NeckC ChestC HipC ThighC KneeC ForearmC

0.7918609 0.3002820 0.2875951 0.3555573 0.7006217 0.3853034 0.6820736 0.2924295

Now, no other predictor has p-vlaue < alpha-to-enter = 0.15, so we stop.

**Thus, we obtain the model from forward selection:**

bodyfat.percent = -34.9774 + 1.0133\*AbdomenC -0.1490\*Weight -1.8015\*WristC + 0.3766\*BicepsC + 0.3323\*AnkleC

This model is different with the model obtained in part A.

**C.**

The following are the best model of each size according to the 𝑅𝑆𝑆:

> sum.subset$which

(Intercept) Over45 Weight Height NeckC ChestC AbdomenC HipC ThighC KneeC AnkleC BicepsC ForearmC WristC

1 TRUE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

2 TRUE FALSE TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

3 TRUE FALSE TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE TRUE

4 TRUE FALSE TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE TRUE

5 TRUE FALSE TRUE FALSE FALSE FALSE TRUE FALSE FALSE FALSE TRUE TRUE FALSE TRUE

6 TRUE FALSE FALSE TRUE FALSE TRUE TRUE TRUE FALSE FALSE FALSE TRUE FALSE TRUE

7 TRUE FALSE FALSE TRUE TRUE TRUE TRUE TRUE FALSE FALSE FALSE TRUE FALSE TRUE

8 TRUE FALSE FALSE TRUE TRUE TRUE TRUE TRUE FALSE FALSE TRUE TRUE FALSE TRUE

9 TRUE FALSE FALSE TRUE TRUE TRUE TRUE TRUE FALSE FALSE TRUE TRUE TRUE TRUE

10 TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE FALSE FALSE TRUE TRUE TRUE TRUE

11 TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE

12 TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

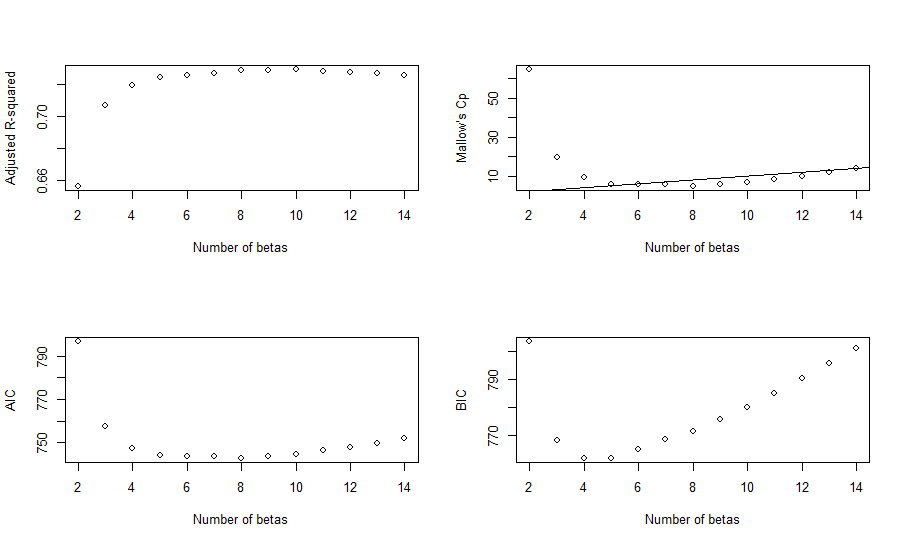
13 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

**The following are the values of RSS of the model for each size:**

> RSS.p

[1] 5952.557 5054.503 4817.476 4712.603 4668.264 4631.512 4580.936 4558.034 4538.593 4532.643 4524.897 4523.317

[13] 4522.882

**The following are adjusted R squares:**

> R2.adj

[1] 0.6563888 0.7070477 0.7196504 0.7246340 0.7261069 0.7271450 0.7290094 0.7292453 0.7292768 0.7285005 0.7278256

[12] 0.7267727 0.7256413

We choose the model with 10 betas based on the adjusted R square and the predictors are Height, NeckC, ChestC, AbdomenC, HipC, AnkleC, BicepsC, ForearmC, WristC.

**The following are values of Mallow’s Cp:**

> C.p

[1] 64.599152 19.739499 9.371655 5.899460 5.585886 5.668192 5.029166 5.834179 6.819792 8.509313 10.105127

[12] 12.022667 14.000000

We choose the model with 5 betas based on the Mallow’s Cp and the predictors are Weight, AbdomenC, BicepsC, WristC.

**The following are values of AIC:**

> aic.p

[1] 796.5288 757.6435 747.6361 744.1337 743.7704 743.7944 743.0494 743.7964 744.7279 746.3999 747.9723 749.8850 751.8610

We choose the model with 8 betas based on the Mallow’s Cp and the predictors are Height, NeckC, ChestC, AbdomenC, HipC, BicepsC, WristC.

**The following are values of BIC:**

> bic.p

[1] 803.5717 768.2079 761.7220 761.7410 764.8992 768.4447 771.2211 775.4896 779.9425 785.1360 790.2298 795.6640 801.1614

We choose the model with 4 betas based on the Mallow’s Cp and the predictors are Weight, AbdomenC, WristC.

The model obtained based on the adjusted R square is most complicated model and the model obtained based on the BIC is the simplest model. All 4 models are different with each other.

**D.**

> vif(lm.backward)

Height NeckC ChestC AbdomenC HipC BicepsC WristC

1.456085 3.482255 7.188742 8.272218 5.599539 2.618959 2.419499

> vif(lm.forward)

AbdomenC Weight WristC BicepsC AnkleC

4.570160 8.655984 2.245620 2.668126 1.689372

> vif(lm.R2.adj)

Height NeckC ChestC AbdomenC HipC AnkleC BicepsC ForearmC WristC

1.487721 3.634850 7.282090 8.518866 5.860135 1.627480 2.974710 2.322658 2.658725

> vif(lm.C.p)

Weight AbdomenC BicepsC WristC

7.529088 4.298628 2.657965 2.146810

> vif(lm.aic.p)

Height NeckC ChestC AbdomenC HipC BicepsC WristC

1.456085 3.482255 7.188742 8.272218 5.599539 2.618959 2.419499

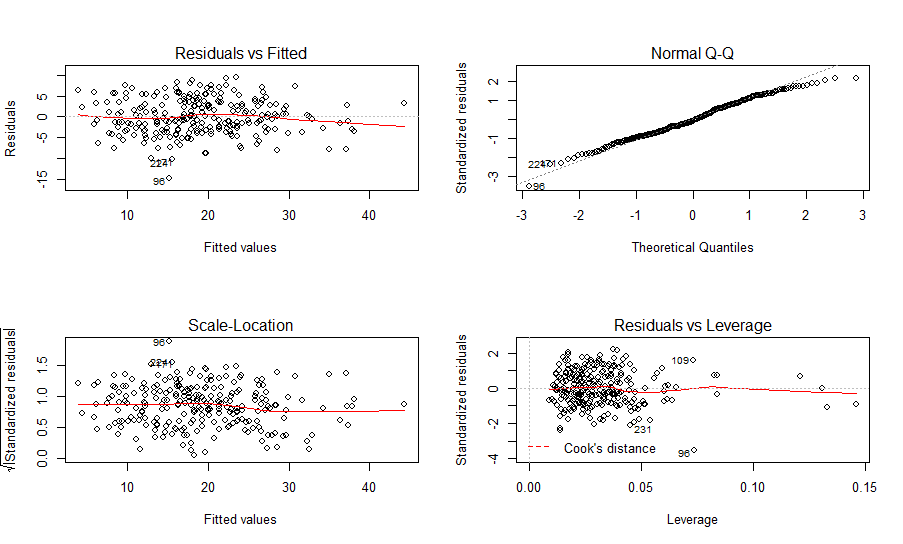
> vif(lm.bic.p)

Weight AbdomenC WristC

5.729221 4.263829 2.127194

By checking the VIF, there is no significant collinearity between predictors in each model. Thus, we do will not remove any predictors from each model. Only further investigation is needed for predictors that has VIF ≥ 4. In addition, we observe that model obtained from AIC is the same as the model obtained from backward elimination.

**The following is the model obtained from backward elimination:**



> shapiro.test(lm.backward$residuals)

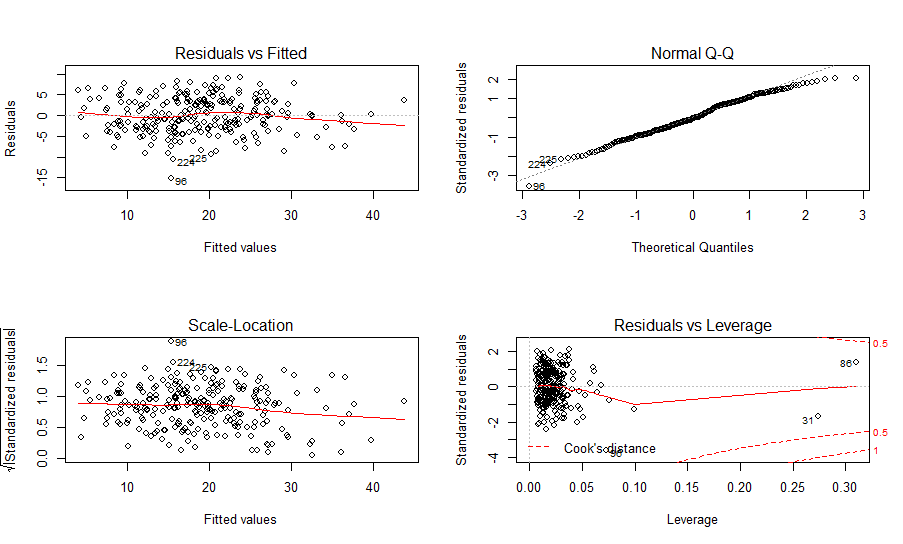
Shapiro-Wilk normality test

data: lm.backward$residuals

W = 0.98949, p-value = 0.06669

The model obtained from backward elimination satisfies the linearity and constant variance. The Q-Q plot and the Shapiro-Wilk test indicates normality. Thus, all assumptions are satisfied. This model has R-square=0.7366.

**The following is the model obtained from forward selection:**



> shapiro.test(lm.forward$residuals)

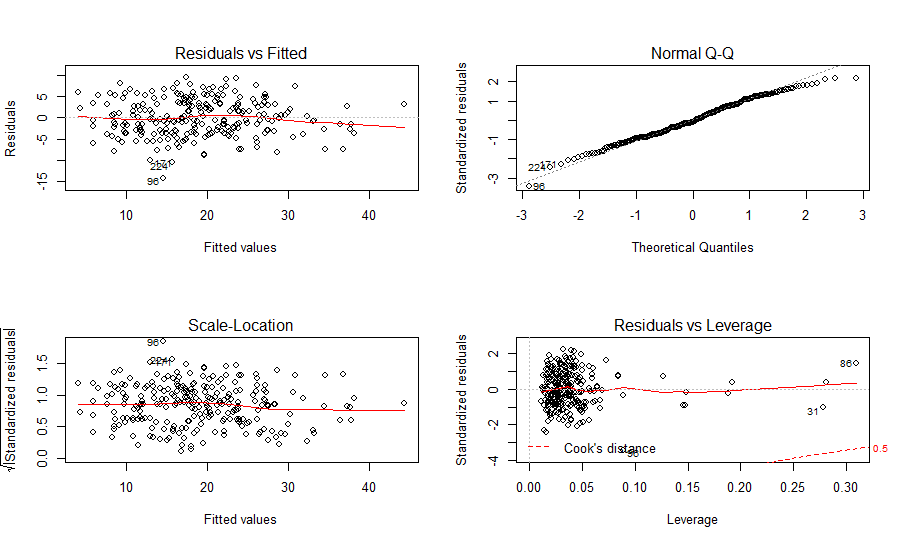
Shapiro-Wilk normality test

data: lm.forward$residuals

W = 0.9887, p-value = 0.04754

The model obtained from backward elimination satisfies the linearity and constant variance. The Q-Q plot and the Shapiro-Wilk test (p-value=0.04754, reject null hypothesis) does not indicate normality. Thus, one assumption (normality) is not satisfied. If we use this model, we need to be careful with the result since the result may be inaccurate. This model has R-square = 0.7316.

**The following is the model obtained from adjusted R-square:**



> shapiro.test(lm.R2.adj$residuals)

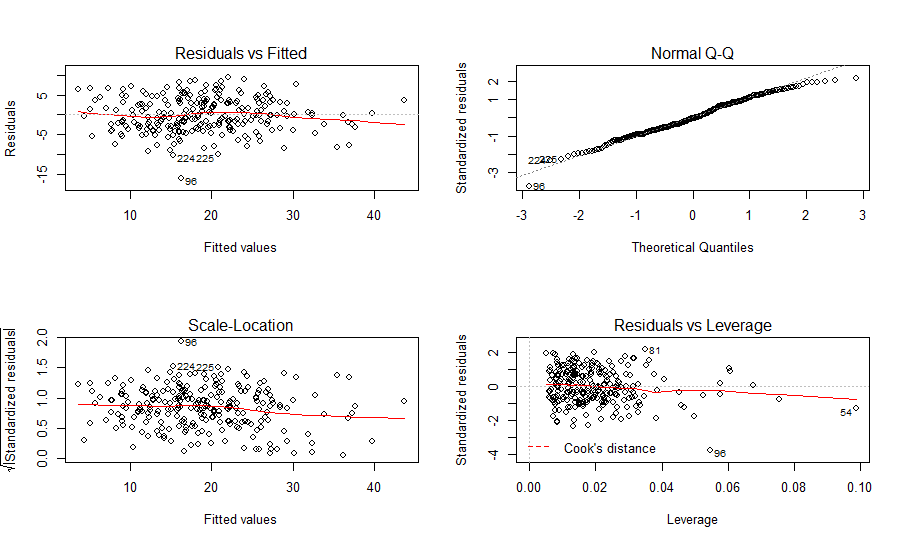
Shapiro-Wilk normality test

data: lm.R2.adj$residuals

W = 0.99067, p-value = 0.1105

The model obtained from backward elimination satisfies the linearity and constant variance. The Q-Q plot and the Shapiro-Wilk test indicates normality. Thus, all assumptions are satisfied. This model has R-square = 0.7391.

**The following is the model obtained from Mallow’s Cp:**



> shapiro.test(lm.C.p$residuals)

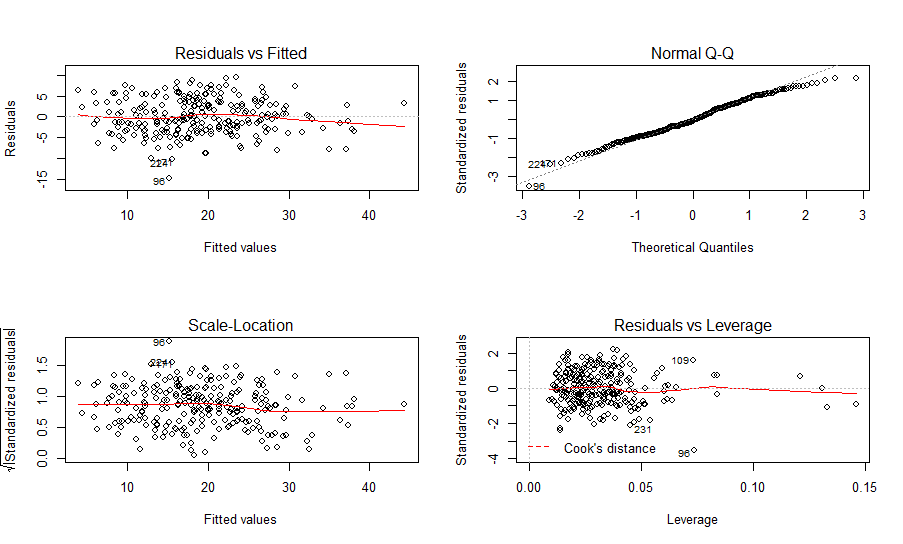
Shapiro-Wilk normality test

data: lm.C.p$residuals

W = 0.98836, p-value = 0.04113

The model obtained from backward elimination satisfies the linearity and constant variance. The Q-Q plot and the Shapiro-Wilk test (p-value=0.04113, reject null hypothesis) does not indicate normality. Thus, one assumption (normality) is not satisfied. If we use this model, we need to be careful with the result since the result may be inaccurate. This model has R-square = 0.7291.

**The following is the model obtained from AIC:**



> shapiro.test(lm.aic.p$residuals)

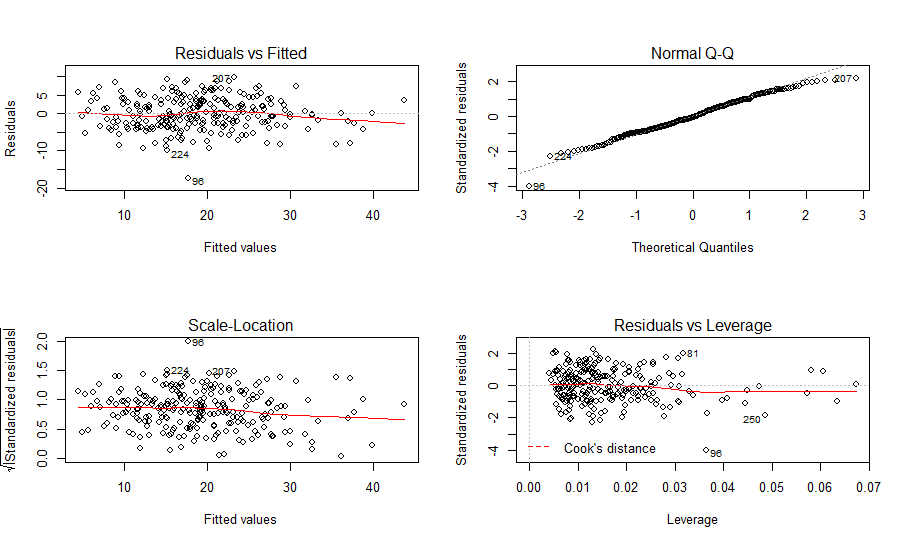
Shapiro-Wilk normality test

data: lm.aic.p$residuals

W = 0.98949, p-value = 0.06669

The model obtained from backward elimination satisfies the linearity and constant variance. The Q-Q plot and the Shapiro-Wilk test indicates normality. Thus, all assumptions are satisfied. This model has R-square = 0.7366.

**The following is the model obtained from BIC:**



> shapiro.test(lm.bic.p$residuals)

Shapiro-Wilk normality test

data: lm.bic.p$residuals

W = 0.9889, p-value = 0.0517

The model obtained from backward elimination satisfies the linearity and constant variance. The Q-Q plot and the Shapiro-Wilk test indicates normality. Thus, all assumptions are satisfied. This model has R-square = 0.723.

Based on the information above, we can tell that each model has similar R-square and variability and only models obtained from forward election and Mallow’s Cp do not satisfy normality. **Thus, we choose the model obtained from BIC**, which is simpler than other models and is useful just like other models, because all models are similar by looking at the diagnostic graphs and we want to choose the model that is both simple and useful.

**Interpretation:**

Holding other variables constant, the body fat percentage decrease 0.09985 when weight increase 1 unit.

Holding other variables constant, the body fat percentage increase 0.97919 when AbdomenC increase 1 unit.

Holding other variables constant, the body fat percentage decrease 1.55701 when WristC increase 1 unit.

The intercept of the model has no real meaning in the real life, because the body fat percentage cannot be negative, and the weight cannot be 0.

**For 95% prediction interval:**

> x0=data.frame(Weight=mean(bodyfat$Weight),AbdomenC=mean(bodyfat$AbdomenC),WristC=mean(bodyfat$WristC))

> PI=predict(lm.bic.p,new=x0,interval='prediction',level=0.95)

> PI

fit lwr upr

1 18.9852 10.25148 27.71892

The prediction interval for the average men is **[10.25148, 27.71892].**