HW6

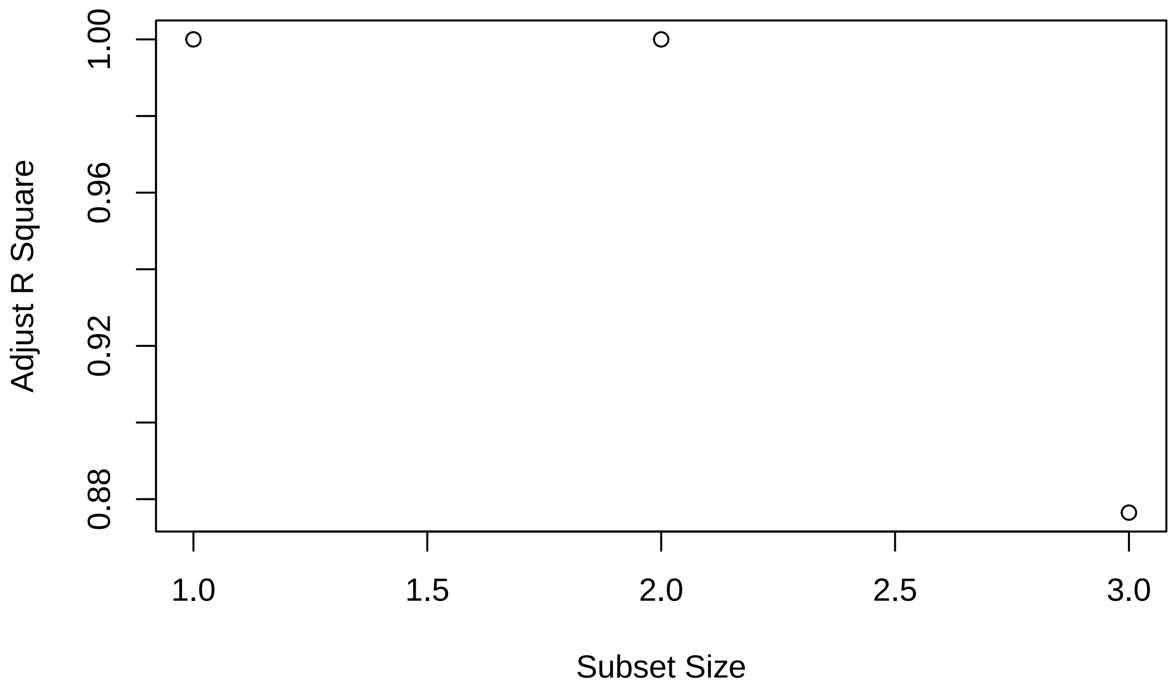
Getong Zhong

2022-12-02

Question 7.1

(a)

```
Case <- c(1:5)
Y <- c(5,6,8,9,11)
X1 <- c(1,200,-50,909,506)
X2 <- c(1004,806,1058,100,505)
X3 <- c(6,7.3,11,13,13.1)
data1 <- as.data.frame(cbind(Case, Y, X1, X2, X3))
mod1 <- lm(Y ~ X1 + X2 + X3, data = data1)
mod2 <- lm(Y ~ X1 + X2, data = data1)
mod3 <- lm(Y ~ X3, data = data1)
adjr2 <- c(summary(mod1)$adj.r.squared, summary(mod2)$adj.r.squared, summary(mod3)$adj.r.squared)
subset_size <- c(1,2,3)
plot(adjr2 ~ subset_size, xlab = "Subset Size", ylab = "Adjust R Square")
```



```
Predictors <- c("X1, X2, X3", "X1, X2", "X3")
AIC <- c(AIC(mod1, k=2), AIC(mod2, k=2), AIC(mod3, k=2))
BIC <- c(AIC(mod1, k=log(nrow(data1))), AIC(mod2, k=log(nrow(data1))), AIC(mod3, k=log(nrow(data1))))

results <- cbind(subset_size, Predictors, adjr2, AIC, BIC)
results
```

```
##      subset_size Predictors   adjr2               AIC
## [1,] "1"         "X1, X2, X3" "1"                 "-269.578999741786"
## [2,] "2"         "X1, X2"     "1"                 "-271.559992520695"
## [3,] "3"         "X3"         "0.876484701848241" "15.8806367928623"
##      BIC
## [1,] "-271.531810179616"
## [2,] "-273.122240870959"
## [3,] "14.7089505301646"
```

From the chart we can tell that the model 2 is the best model, since it has the highest adjusted r square value and lowest AIC and BIC values

(b)

AIC forward selection

```
AIC_selection <- step(lm(Y ~ 1), Y ~ X1 + X2 + X3, direction="forward")
```

```
## Start:  AIC=9.59
## Y ~ 1
##
##        Df Sum of Sq     RSS     AIC
## + X3    1   20.6879  2.1121 -0.3087
## + X1    1    8.6112 14.1888  9.2151
## + X2    1    8.5064 14.2936  9.2519
## <none>               22.8000  9.5866
##
## Step:  AIC=-0.31
## Y ~ X3
##
##        Df Sum of Sq    RSS      AIC
## <none>               2.1121 -0.30875
## + X2    1  0.066328 2.0458  1.53172
## + X1    1  0.064522 2.0476  1.53613
```

```
AIC_selection
```

```
##
## Call:
## lm(formula = Y ~ X3)
##
## Coefficients:
## (Intercept)           X3
##      0.7975       0.6947
```

BIC forward selection

```
BIC_forward<- step(lm(Y ~ 1), Y ~ X1 + X2 + X3, direction="forward", k = log(nrow(data1)))
```

```
## Start:  AIC=9.2
## Y ~ 1
##
##        Df Sum of Sq     RSS     AIC
## + X3    1   20.6879  2.1121 -1.0899
## + X1    1    8.6112 14.1888  8.4339
## + X2    1    8.5064 14.2936  8.4707
## <none>               22.8000  9.1961
##
## Step:  AIC=-1.09
## Y ~ X3
##
##        Df Sum of Sq    RSS      AIC
## <none>               2.1121 -1.08987
## + X2    1  0.066328 2.0458  0.36003
## + X1    1  0.064522 2.0476  0.36444
```

```
BIC_forward
```

```
##
## Call:
## lm(formula = Y ~ X3)
##
## Coefficients:
## (Intercept)           X3
##      0.7975       0.6947
```

(c)

Since the stepwise regression method choose the predictors one by one, there are many situations that the model might be over-fitted that is the p -values obtained after variable selection are much smaller than their true values, therefore there have some difference of outcome between (a) and (b). Therefore, the result of (a) and (b) are not same.

(d)

I would suggest the model from (b), since x1 and x2 has a really high correlation, that might affects the result of adjusted r square value, and the accuracy of the model, and model from (b) doesn't include either X1 or X2.

Problem 7.2

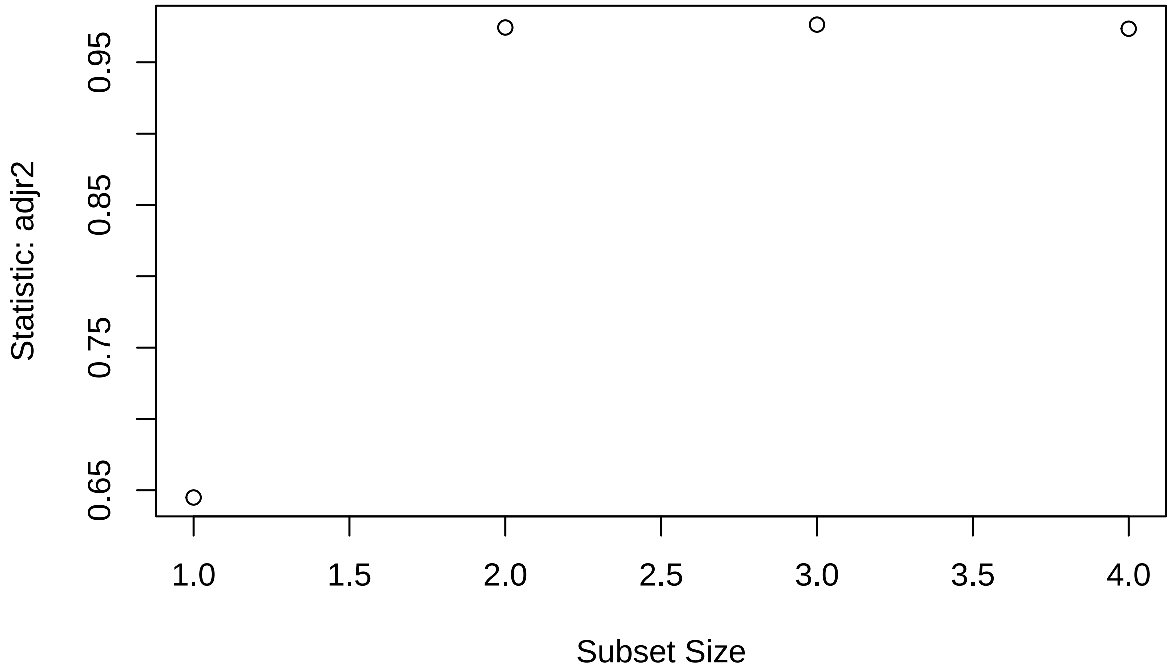
(a)

```
Y <- c(78.5,74.3,104.3,87.6,95.9,109.2,102.7,72.5,93.1,115.9,83.8,113.3,109.4)
x1 <- c(7,1,11,11,7,11,3,1,2,21,1,11,10)
x2 <- c(26,29,56,31,52,55,71,31,54,47,40,66,68)
x3 <- c(6,15,8,8,6,9,17,22,18,4,23,9,8)
x4 <- c(60,52,20,47,33,22,6,44,22,26,34,12,12)

data2 <- as.data.frame(c(Y, x1, x2, x3, x4))
mod1 <- lm(Y ~ x4, data2)
mod2 <- lm(Y ~ x1 + x2, data2)
mod3 <- lm(Y ~ x1 + x2 + x4, data2)
mod4 <- lm(Y ~ x1 + x2 + x3 + x4, data2)

adjr2 <- c(summary(mod1)$adj.r.squared, summary(mod2)$adj.r.squared, summary(mod3)$adj.r.squared, summa
subset_size <- c(1,2,3,4)

plot(adjr2 ~ subset_size, xlab = "Subset Size", ylab = "Statistic: adjr2")
```



```
Predictors <- c("X4", "X1, X2", "X1, X2, X4", "X1, X2, X3, X4")
AIC_col <- c(AIC(mod1, k=2), AIC(mod2, k=2), AIC(mod3, k=2), AIC(mod4, k=2))
BIC_col <- c(AIC(mod1, k=log(nrow(data2))), AIC(mod2, k=log(nrow(data2))), AIC(mod3, k=log(nrow(data2)))

allsubsets <- cbind(subset_size, Predictors, adjr2, AIC_col, BIC_col)
allsubsets
```

```
##      subset_size Predictors       adjr2               AIC_col
## [1,] "1"         "X4"             "0.644954869961756" "97.7440447788562"
## [2,] "2"         "X1, X2"         "0.974414049442758" "64.3123927621906"
## [3,] "3"         "X1, X2, X4"     "0.976447268267236" "63.8662854718626"
## [4,] "4"         "X1, X2, X3, X4" "0.97356343061152"  "65.8366897916517"
##      BIC_col
## [1,] "104.267206588543"
## [2,] "73.0099418417732"
## [3,] "74.7382218213408"
## [4,] "78.8830134110255"
```

Based on the results, I think model 2 and model 3 are both good, since they have relatively high adjusted r square and low AIC/BIC compare to other models.

(b)

AIC forward

```
attach(data2)
AIC_forward <- step(lm(Y ~ 1), Y ~ x1 + x2 + x3 + x4, direction="forward")
```

```
## Start:  AIC=71.44
## Y ~ 1
## 
##        Df Sum of Sq     RSS    AIC
## + x4    1   1831.90  883.87 58.852
## + x2    1   1809.43  906.34 59.178
## + x1    1   1450.08 1265.69 63.519
## + x3    1    776.36 1939.40 69.067
## <none>              2715.76 71.444
## 
## Step:  AIC=58.85
## Y ~ x4
## 
##        Df Sum of Sq    RSS    AIC
## + x1    1    809.10  74.76 28.742
## + x3    1    708.13 175.74 39.853
## <none>              883.87 58.852
## + x2    1     14.99 868.88 60.629
## 
## Step:  AIC=28.74
## Y ~ x4 + x1
## 
##        Df Sum of Sq    RSS    AIC
## + x2    1    26.789 47.973 24.974
## + x3    1    23.926 50.836 25.728
## <none>              74.762 28.742
## 
## Step:  AIC=24.97
## Y ~ x4 + x1 + x2
## 
##        Df Sum of Sq    RSS    AIC
## <none>              47.973 24.974
## + x3    1   0.10909 47.864 26.944
```

```
AIC_forward
```

```
## 
## Call:
## lm(formula = Y ~ x4 + x1 + x2)
## 
## Coefficients:
## (Intercept)           x4           x1           x2  
##     71.6483      -0.2365       1.4519       0.4161
```

BIC forward

```
BIC_forward <- step(lm(Y ~ 1), Y ~ x1 + x2 + x3 + x4, direction="forward", k = log(nrow(data2)))
```

```
## Start:  AIC=73.62
## Y ~ 1
##
##        Df Sum of Sq     RSS    AIC
## + x4    1   1831.90  883.87 63.200
## + x2    1   1809.43  906.34 63.527
## + x1    1   1450.08 1265.69 67.868
## + x3    1    776.36 1939.40 73.416
## <none>              2715.76 73.619
##
## Step:  AIC=63.2
## Y ~ x4
##
##        Df Sum of Sq    RSS    AIC
## + x1    1    809.10  74.76 35.265
## + x3    1    708.13 175.74 46.376
## <none>              883.87 63.200
## + x2    1     14.99 868.88 67.152
##
## Step:  AIC=35.26
## Y ~ x4 + x1
##
##        Df Sum of Sq    RSS    AIC
## + x2    1    26.789 47.973 33.671
## + x3    1    23.926 50.836 34.425
## <none>              74.762 35.265
##
## Step:  AIC=33.67
## Y ~ x4 + x1 + x2
##
##        Df Sum of Sq    RSS    AIC
## <none>              47.973 33.671
## + x3    1   0.10909 47.864 37.816
```

```
BIC_forward
```

```
##
## Call:
## lm(formula = Y ~ x4 + x1 + x2)
##
## Coefficients:
## (Intercept)           x4           x1           x2
##     71.6483      -0.2365       1.4519       0.4161
```

```
detach(data2)
```

Based on the results, the three predictors model works the best

(c)

AIC backward

```
AIC_backward <- step(lm(Y ~ x1 + x2 + x3 + x4), Y ~ x1 + x2 + x3 + x4, direction="backward")
```

```
## Start:  AIC=26.94
## Y ~ x1 + x2 + x3 + x4
##
##        Df Sum of Sq    RSS    AIC
## - x3    1    0.1091 47.973 24.974
## - x4    1    0.2470 48.111 25.011
## - x2    1    2.9725 50.836 25.728
## <none>              47.864 26.944
## - x1    1   25.9509 73.815 30.576
##
## Step:  AIC=24.97
## Y ~ x1 + x2 + x4
##
##        Df Sum of Sq    RSS    AIC
## <none>               47.97 24.974
## - x4    1      9.93  57.90 25.420
## - x2    1     26.79  74.76 28.742
## - x1    1    820.91 868.88 60.629
```

```
AIC_backward
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2 + x4)
##
## Coefficients:
## (Intercept)           x1           x2           x4
##     71.6483       1.4519       0.4161      -0.2365
```

BIC backward

```
BIC_backward <- step(lm(Y ~ x1 + x2 + x3 + x4), Y ~ x1 + x2 + x3 + x4, direction="backward", k = log(nr
```

```
## Start:  AIC=37.82
## Y ~ x1 + x2 + x3 + x4
##
##        Df Sum of Sq    RSS    AIC
## - x3    1    0.1091 47.973 33.671
## - x4    1    0.2470 48.111 33.709
## - x2    1    2.9725 50.836 34.425
## <none>              47.864 37.816
## - x1    1   25.9509 73.815 39.273
##
## Step:  AIC=33.67
## Y ~ x1 + x2 + x4
##
##        Df Sum of Sq    RSS    AIC
## - x4    1      9.93  57.90 31.943
## <none>               47.97 33.671
## - x2    1     26.79  74.76 35.265
## - x1    1    820.91 868.88 67.152
##
## Step:  AIC=31.94
## Y ~ x1 + x2
##
##        Df Sum of Sq     RSS    AIC
## <none>                57.90 31.943
## - x1    1    848.43  906.34 63.527
## - x2    1   1207.78 1265.69 67.868
```

```
BIC_backward
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2)
##
## Coefficients:
## (Intercept)           x1           x2
##     52.5773       1.4683       0.6623
```

Based on the results, the two or three predictors model works the best

(d)

Since the stepwise regression method choose the predictors one by one, there are many situations that the model might be over-fitted that is the p -values obtained after variable selection are much smaller than their true values, therefore there have some difference of outcome between (a) and (b)/(c). For (b) and (c), there is not gaurentee that the backward and forward will have the same results.

(e)

I would choose the two predictor models, since X2 and X4 has a really high correlationship, I'm afraid that might affect the accuracy of the model.

Question 7.3

(a)

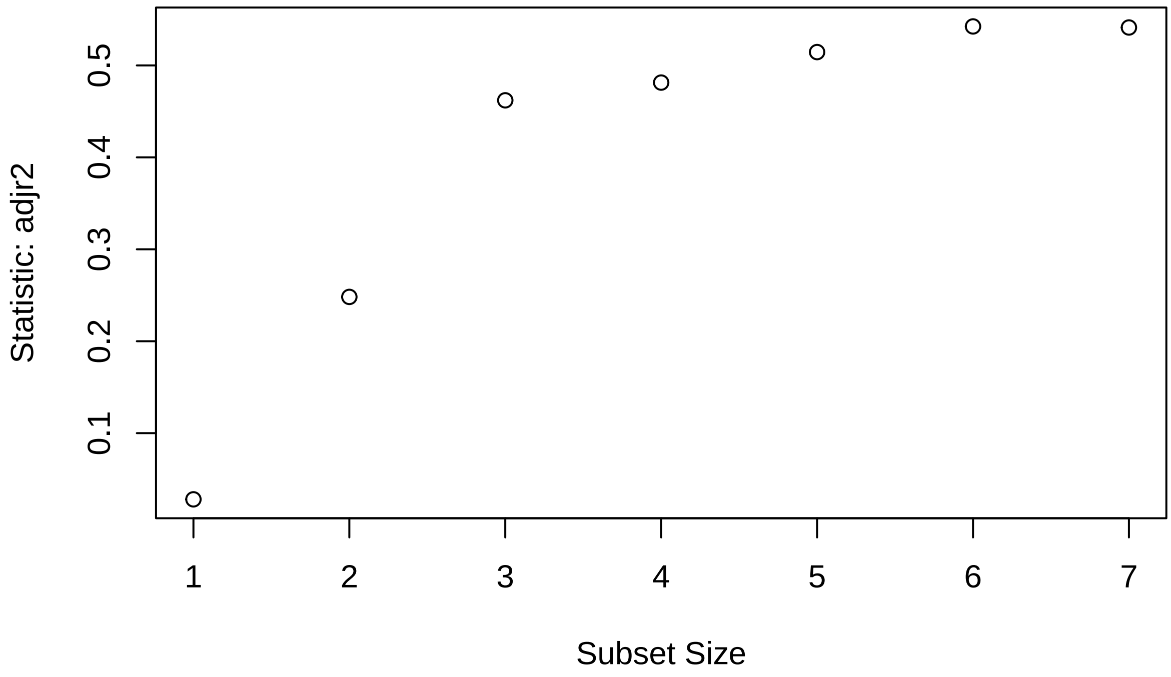
```
library(readr)
data3 <- read.csv("C:/Users/tonyg/Desktop/Academic/Grad/HUDM 5126/pgatour2006.csv")

mod1 <- lm(log(PrizeMoney) ~ DrivingAccuracy, data = data3)
mod2 <- lm(log(PrizeMoney) ~ DrivingAccuracy + GIR , data = data3)
mod3 <- lm(log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage, data = data3)
mod4 <- lm(log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion, data = data3)
mod5 <- lm(log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion + SandSaves, dat
mod6 <- lm(log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion + SandSaves + Sc
mod7 <- lm(log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion + SandSaves + Sc

adjr2 <- c(summary(mod1)$adj.r.squared, summary(mod2)$adj.r.squared, summary(mod3)$adj.r.squared, summa
```

```r
subset_size <- c(1,2,3,4,5,6,7)

plot(adjr2 ~ subset_size, xlab = "Subset Size", ylab = "Statistic: adjr2")
```



```r
Predictors <- c("X1", "X1, X2", "X1, X2, X3", "X1, X2, X3, X4", "X1, X2, X3, X4, X5", "X1, X2, X3, X4,
AIC_col <- c(AIC(mod1, k=2), AIC(mod2, k=2), AIC(mod3, k=2), AIC(mod4, k=2), AIC(mod5, k=2), AIC(mod6,
BIC_col <- c(AIC(mod1, k=log(nrow(data3))), AIC(mod2, k=log(nrow(data3))), AIC(mod3, k=log(nrow(data3))

allsubsets <- cbind(subset_size, Predictors, adjr2, AIC_col, BIC_col)
allsubsets
```

```
##      subset_size Predictors                   adjr2
## [1,] "1"         "X1"                         "0.0280205362863096"
## [2,] "2"         "X1, X2"                     "0.248191772696449"
## [3,] "3"         "X1, X2, X3"                 "0.462029443701811"
## [4,] "4"         "X1, X2, X3, X4"             "0.481328054159818"
## [5,] "5"         "X1, X2, X3, X4, X5"         "0.514497472883682"
## [6,] "6"         "X1, X2, X3, X4, X5, X6"     "0.542393277749667"
## [7,] "7"         "X1, X2, X3, X4, X5, X6, X7" "0.541240402006151"
##      AIC_col            BIC_col
## [1,] "546.804455187927" "556.638799165618"
## [2,] "497.448266278831" "510.560724915753"
## [3,] "432.833303156706" "449.223876452859"
## [4,] "426.649483837215" "446.318171792599"
## [5,] "414.66753819487"  "437.614340809484"
## [6,] "404.035059911805" "430.259977185649"
## [7,] "405.488443114175" "434.991475047249"
```

From Adjusted R2 ,AIC, AICC and BIC we can see that the model with 6 or 7 parameters all possess a relatively high values and indicate the model of subset 6 and 7 to be the best of the possible models.

(b)

AIC backward

```
AIC_backward <- step(lm(log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion + Sa
```

```
## Start:  AIC=-152.74
## log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion + 
##     SandSaves + Scrambling + PuttsPerRound
## 
##                    Df Sum of Sq    RSS     AIC
## - PuttingAverage    1    0.0020 82.868 -154.73
## - DrivingAccuracy   1    0.0396 82.905 -154.64
## - PuttsPerRound     1    0.2314 83.097 -154.19
## <none>                          82.866 -152.74
## - SandSaves         1    1.0436 83.909 -152.28
## - Scrambling        1    1.1576 84.023 -152.02
## - BirdieConversion  1    6.6928 89.558 -139.51
## - GIR               1    9.1200 91.986 -134.27
## 
## Step:  AIC=-154.73
## log(PrizeMoney) ~ DrivingAccuracy + GIR + BirdieConversion + 
##     SandSaves + Scrambling + PuttsPerRound
## 
##                    Df Sum of Sq    RSS     AIC
## - DrivingAccuracy   1    0.0377 82.905 -156.64
## <none>                          82.868 -154.73
## - PuttsPerRound     1    1.0263 83.894 -154.32
## - SandSaves         1    1.0461 83.914 -154.27
## - Scrambling        1    1.7855 84.653 -152.55
## - BirdieConversion  1    8.6663 91.534 -137.24
## - GIR               1   17.0549 99.922 -120.05
## 
## Step:  AIC=-156.64
## log(PrizeMoney) ~ GIR + BirdieConversion + SandSaves + Scrambling + 
##     PuttsPerRound
## 
##                    Df Sum of Sq     RSS     AIC
## <none>                           82.905 -156.64
## - PuttsPerRound     1    1.0003  83.905 -156.29
## - SandSaves         1    1.1078  84.013 -156.04
## - Scrambling        1    1.7566  84.662 -154.53
## - BirdieConversion  1   10.8275  93.733 -134.58
## - GIR               1   20.5479 103.453 -115.24
```

```
AIC_backward
```

```
##
## Call:
## lm(formula = log(PrizeMoney) ~ GIR + BirdieConversion + SandSaves +
##     Scrambling + PuttsPerRound, data = data3)
##
## Coefficients:
##      (Intercept)               GIR  BirdieConversion         SandSaves
##         -0.58318           0.19702           0.16275           0.01552
##       Scrambling     PuttsPerRound
##          0.04963          -0.34974
```

BIC backward

```
BIC_backward <- step(lm(log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion + Sa
```

```
## Start:  AIC=-126.51
## log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion +
##     SandSaves + Scrambling + PuttsPerRound
##
##                    Df Sum of Sq    RSS     AIC
## - PuttingAverage    1    0.0020 82.868 -131.78
## - DrivingAccuracy   1    0.0396 82.905 -131.69
## - PuttsPerRound     1    0.2314 83.097 -131.24
## - SandSaves         1    1.0436 83.909 -129.34
## - Scrambling        1    1.1576 84.023 -129.07
## <none>                          82.866 -126.51
## - BirdieConversion  1    6.6928 89.558 -116.56
## - GIR               1    9.1200 91.986 -111.32
##
## Step:  AIC=-131.78
## log(PrizeMoney) ~ DrivingAccuracy + GIR + BirdieConversion +
##     SandSaves + Scrambling + PuttsPerRound
##
##                    Df Sum of Sq    RSS     AIC
## - DrivingAccuracy   1    0.0377 82.905 -136.97
## - PuttsPerRound     1    1.0263 83.894 -134.65
## - SandSaves         1    1.0461 83.914 -134.60
## - Scrambling        1    1.7855 84.653 -132.88
## <none>                          82.868 -131.78
## - BirdieConversion  1    8.6663 91.534 -117.57
## - GIR               1   17.0549 99.922 -100.38
##
## Step:  AIC=-136.97
## log(PrizeMoney) ~ GIR + BirdieConversion + SandSaves + Scrambling +
##     PuttsPerRound
##
##                    Df Sum of Sq     RSS      AIC
## - PuttsPerRound     1    1.0003  83.905 -139.900
## - SandSaves         1    1.1078  84.013 -139.649
## - Scrambling        1    1.7566  84.662 -138.141
## <none>                            82.905 -136.973
## - BirdieConversion  1   10.8275  93.733 -118.192
## - GIR               1   20.5479 103.453  -98.853
##
## Step:  AIC=-139.9
## log(PrizeMoney) ~ GIR + BirdieConversion + SandSaves + Scrambling
##
##                    Df Sum of Sq     RSS      AIC
## - SandSaves         1     1.286  85.191 -142.198
## <none>                           83.905 -139.900
## - Scrambling        1     7.595  91.501 -128.194
## - GIR               1    35.317 119.222  -76.324
## - BirdieConversion  1    36.555 120.460  -74.299
##
## Step:  AIC=-142.2
## log(PrizeMoney) ~ GIR + BirdieConversion + Scrambling
##
##                    Df Sum of Sq     RSS      AIC
## <none>                           85.191 -142.198
## - Scrambling        1    15.786 100.977 -114.157
## - GIR               1    34.057 119.248  -81.560
## - BirdieConversion  1    40.308 125.499  -71.545
```

```
BIC_backward
```

```
##
## Call:
## lm(formula = log(PrizeMoney) ~ GIR + BirdieConversion + Scrambling,
##     data = data3)
##
## Coefficients:
##      (Intercept)               GIR  BirdieConversion        Scrambling
##        -11.08314           0.15658           0.20625           0.09178
```

model log(PrizeMoney) ~ DrivingAccuracy + GIR + BirdieConversion + SandSaves + Scrambling + PuttsPerRound is the best

(c)

AIC forward

```
AIC_forward <- step(lm(log(PrizeMoney) ~ 1, data = data3), log(PrizeMoney) ~ DrivingAccuracy + GIR + Pu
```

```
## Start:  AIC=-6.84
## log(PrizeMoney) ~ 1
##
##                    Df Sum of Sq    RSS     AIC
## + GIR               1    47.760 139.59 -62.516
## + BirdieConversion  1    40.930 146.43 -53.154
## + PuttingAverage    1    34.660 152.69 -44.936
## + Scrambling        1    25.260 162.09 -33.227
## + SandSaves         1    10.926 176.43 -16.618
## + PuttsPerRound     1     6.295 181.06 -11.540
## + DrivingAccuracy   1     6.184 181.17 -11.419
## <none>                         187.35  -6.841
##
## Step:  AIC=-62.52
## log(PrizeMoney) ~ GIR
##
##                    Df Sum of Sq     RSS      AIC
## + PuttsPerRound     1    44.240  95.355 -135.220
## + PuttingAverage    1    39.748  99.847 -126.197
## + BirdieConversion  1    38.618 100.977 -123.991
## + SandSaves         1    15.043 124.552  -82.864
## + Scrambling        1    14.096 125.499  -81.380
## <none>                          139.595  -62.516
## + DrivingAccuracy   1     0.185 139.410  -60.776
##
## Step:  AIC=-135.22
## log(PrizeMoney) ~ GIR + PuttsPerRound
##
##                    Df Sum of Sq    RSS     AIC
## + BirdieConversion  1    8.1732 87.181 -150.78
## + DrivingAccuracy   1    2.6309 92.724 -138.70
## + SandSaves         1    1.1746 94.180 -135.65
## + PuttingAverage    1    1.0592 94.295 -135.41
## <none>                         95.355 -135.22
## + Scrambling        1    0.0510 95.304 -133.32
##
## Step:  AIC=-150.78
## log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion
##
##                   Df Sum of Sq    RSS     AIC
## + Scrambling       1    3.1684 84.013 -156.04
## + SandSaves        1    2.5196 84.662 -154.53
## + PuttingAverage   1    1.2574 85.924 -151.63
## <none>                        87.181 -150.78
## + DrivingAccuracy  1    0.0611 87.120 -148.92
##
## Step:  AIC=-156.04
## log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion + Scrambling
##
##                   Df Sum of Sq    RSS     AIC
## + SandSaves        1   1.10778 82.905 -156.64
## <none>                        84.013 -156.04
## + DrivingAccuracy  1   0.09937 83.914 -154.27
## + PuttingAverage   1   0.00033 84.013 -154.04
##
## Step:  AIC=-156.64
## log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion + Scrambling +
##     SandSaves
##
##                   Df Sum of Sq    RSS     AIC
## <none>                        82.905 -156.64
## + DrivingAccuracy  1  0.037678 82.868 -154.73
## + PuttingAverage   1  0.000062 82.905 -154.64
```

```
AIC_forward
```

```
## 
## Call:
## lm(formula = log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion + 
##     Scrambling + SandSaves, data = data3)
## 
## Coefficients:
##      (Intercept)               GIR     PuttsPerRound  BirdieConversion  
##         -0.58318           0.19702          -0.34974           0.16275  
##       Scrambling         SandSaves  
##          0.04963           0.01552  
```

BIC forward

```
BIC_forward <- step(lm(log(PrizeMoney) ~ 1, data = data3), log(PrizeMoney) ~ DrivingAccuracy + GIR + Pu
```

```
## Start:  AIC=-3.56
## log(PrizeMoney) ~ 1
## 
##                    Df Sum of Sq    RSS     AIC
## + GIR               1    47.760 139.59 -55.960
## + BirdieConversion  1    40.930 146.43 -46.597
## + PuttingAverage    1    34.660 152.69 -38.379
## + Scrambling        1    25.260 162.09 -26.671
## + SandSaves         1    10.926 176.43 -10.062
## + PuttsPerRound     1     6.295 181.06  -4.983
## + DrivingAccuracy   1     6.184 181.17  -4.863
## <none>                         187.35  -3.563
## 
## Step:  AIC=-55.96
## log(PrizeMoney) ~ GIR
## 
##                    Df Sum of Sq     RSS      AIC
## + PuttsPerRound     1    44.240  95.355 -125.386
## + PuttingAverage    1    39.748  99.847 -116.362
## + BirdieConversion  1    38.618 100.977 -114.157
## + SandSaves         1    15.043 124.552  -73.030
## + Scrambling        1    14.096 125.499  -71.545
## <none>                         139.595  -55.960
## + DrivingAccuracy   1     0.185 139.410  -50.941
## 
## Step:  AIC=-125.39
## log(PrizeMoney) ~ GIR + PuttsPerRound
## 
##                    Df Sum of Sq    RSS     AIC
## + BirdieConversion  1    8.1732 87.181 -137.67
## + DrivingAccuracy   1    2.6309 92.724 -125.59
## <none>                         95.355 -125.39
## + SandSaves         1    1.1746 94.180 -122.54
## + PuttingAverage    1    1.0592 94.295 -122.30
## + Scrambling        1    0.0510 95.304 -120.21
```

```
##
## Step:  AIC=-137.67
## log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion
##
##                    Df Sum of Sq    RSS     AIC
## + Scrambling       1    3.1684 84.013 -139.65
## + SandSaves        1    2.5196 84.662 -138.14
## <none>                          87.181 -137.67
## + PuttingAverage   1    1.2574 85.924 -135.24
## + DrivingAccuracy  1    0.0611 87.120 -132.53
##
## Step:  AIC=-139.65
## log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion + Scrambling
##
##                    Df Sum of Sq    RSS     AIC
## <none>                          84.013 -139.65
## + SandSaves        1   1.10778 82.905 -136.97
## + DrivingAccuracy  1   0.09937 83.914 -134.60
## + PuttingAverage   1   0.00033 84.013 -134.37
```

```
BIC_forward
```

```
##
## Call:
## lm(formula = log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion +
##     Scrambling, data = data3)
##
## Coefficients:
##      (Intercept)               GIR     PuttsPerRound  BirdieConversion
##          0.39320           0.19352          -0.37840           0.16589
##       Scrambling
##          0.06282
```

model of log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion + Scrambling + SandSaves is the best

(d)

For results of (a), (b) and (c), the results seems oppsite from one and other is because as a model takes away variables while the other adds variables. There is no significant difference between both as they yield multivariate models of similar magnitude. All the models are both from the same of and has the AIC of -156.64

(e)

Considering the similarity of the results between the backward and forward approaches, we can tell that the 5 variable model seems to be the best as it has a higher overall AIC result. While the 7 variable model contains a significant AIC is higher, but this might boost by the ulticollinearity and correlation that might be present in the 7 variable model.

(f)

```
summary(lm(log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion + Scrambling + SandSaves, data = da
```

```
##
## Call:
## lm(formula = log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion +
##     Scrambling + SandSaves, data = data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71291 -0.48168 -0.09097  0.44843  2.15763
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.583181   7.158721  -0.081   0.9352
## GIR               0.197022   0.028711   6.862 9.31e-11 ***
## PuttsPerRound    -0.349738   0.230995  -1.514   0.1317
## BirdieConversion  0.162752   0.032672   4.981 1.41e-06 ***
## Scrambling        0.049635   0.024738   2.006   0.0462 *
## SandSaves         0.015524   0.009743   1.593   0.1127
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6606 on 190 degrees of freedom
## Multiple R-squared:  0.5575, Adjusted R-squared:  0.5459
## F-statistic: 47.88 on 5 and 190 DF,  p-value: < 2.2e-16
```

When all the predictors are zero, the average value of Prize is e to the power of -0.583181 A one unit increase in GIR results on a e to the power of 0.197022 average percentage change in Prize A one unit increase in PuttsPerRound results on a e to the power of -0.349738 average percentage change in Prize A one unit increase in BirdieConversion results on a e to the power of 0.162752 average percentage change in Prize A one unit increase in Scrambling results on a e to the power of 0.049635 average percentage change in Prize A one unit increase in SandSaves results on a e to the power of 0.015524 average percentage change in Prize This model has a low adjusted r square value of 0.5459, which means about 46 percent of variance has not explained by the model. Besides, we need to aware of that we did not account for multicolinearity and correlation between variables, this model is not yet a perfect model, ideed it has a lot of space of improvement.