STAT 757 Assignment 7

DUE 4/15/2018 11:59PM

AG Schissler

2/14/2018

Instructions [20 points]

Modify this file to provide responses to the Ch.7 Exercises in Sheather (2009). You can find some helpful code here: http://www.stat. tamu.edu/~sheather/book/docs/rcode/Chapter7.R. Also address the project milestones indicated below. Please email **both** your .Rmd (or roxygen .R) and one of the following either .HTML, .PDF, or .DOCX using the format SURNAME-FIRSTNAME-Assignment7.Rmd and SURNAME-FIRSTNAME-Assignment7.pdf.

Exercise 7.5.3 [60 points]

Setup and fit model from 6.7.5

```
library(leaps)
library(car)
myDir <- "~/OneDrive - University of Nevada, Reno/Teaching/STAT_757/Sheather_data/Data/"
dat <- read.delim(file.path(myDir,"pgatour2006.csv"), sep = ",")
str(dat)
```

```
## 'data.frame':    196 obs. of  12 variables:
##  $ Name              : Factor w/ 196 levels "Aaron Baddeley",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ TigerWoods        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PrizeMoney        : int  60661 262045 3635 17516 16683 107294 50620 57273 86782 23396 ...
##  $ AveDrivingDistance: num  288 301 303 289 288 ...
##  $ DrivingAccuracy   : num  60.7 62 51.1 66.4 63.2 ...
##  $ GIR               : num  58.3 69.1 59.1 67.7 64 ...
##  $ PuttingAverage    : num  1.75 1.77 1.79 1.78 1.76 ...
##  $ BirdieConversion  : num  31.4 30.4 29.9 29.3 29.3 ...
##  $ SandSaves         : num  54.8 53.6 37.9 45.1 52.4 ...
##  $ Scrambling        : num  59.4 57.9 50.8 54.8 57.1 ...
##  $ BounceBack        : num  19.3 19.4 16.8 17.1 18.2 ...
##  $ PuttsPerRound     : num  28 29.3 29.2 29.5 28.9 ...
```

```
## subset to only the Y and seven predictors of interest
dat2 <- dat[,c("PrizeMoney", "DrivingAccuracy", "GIR", "PuttingAverage",
               "BirdieConversion", "SandSaves", "Scrambling", "PuttsPerRound")]
```

```
m1 <- lm(log(PrizeMoney) ~ DrivingAccuracy + GIR +
             PuttingAverage + BirdieConversion + SandSaves +
             Scrambling + PuttsPerRound, data = dat2)
```

Part A

```
best_subsets_dat <- regsubsets(as.matrix(dat2[,-1]),log(dat2$PrizeMoney),
                               method = "exhaustive")
sum_best <- summary(best_subsets_dat)
sum_best
```

```
## Subset selection object
## 7 Variables  (and intercept)
##                   Forced in Forced out
## DrivingAccuracy       FALSE      FALSE
## GIR                   FALSE      FALSE
## PuttingAverage        FALSE      FALSE
## BirdieConversion      FALSE      FALSE
## SandSaves             FALSE      FALSE
## Scrambling            FALSE      FALSE
## PuttsPerRound         FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: exhaustive
##          DrivingAccuracy GIR PuttingAverage BirdieConversion SandSaves
## 1  ( 1 ) " "             "*" " "            " "               " "
## 2  ( 1 ) " "             "*" " "            " "               " "
## 3  ( 1 ) " "             "*" " "            "*"               " "
## 4  ( 1 ) " "             "*" " "            "*"               "*"
## 5  ( 1 ) " "             "*" " "            "*"               "*"
## 6  ( 1 ) "*"             "*" " "            "*"               "*"
## 7  ( 1 ) "*"             "*" "*"            "*"               "*"
##          Scrambling PuttsPerRound
## 1  ( 1 ) " "        " "
## 2  ( 1 ) " "        "*"
## 3  ( 1 ) "*"        " "
## 4  ( 1 ) "*"        " "
## 5  ( 1 ) "*"        "*"
## 6  ( 1 ) "*"        "*"
## 7  ( 1 ) "*"        "*"
```

```
## The code below is to get AIC, AICc.
## store models in a list

all_mod <- vector(mode = "list", length = 7)
all_mod[[1]] <- lm(log(PrizeMoney)~GIR, data = dat2)
all_mod[[2]] <- lm(log(PrizeMoney)~GIR + PuttsPerRound, data = dat2)
all_mod[[3]] <- lm(log(PrizeMoney)~GIR + BirdieConversion + Scrambling, data = dat2)
all_mod[[4]] <- lm(log(PrizeMoney)~GIR + Scrambling + BirdieConversion + SandSaves,
                   data = dat2)
all_mod[[5]] <- lm(log(PrizeMoney)~GIR + PuttsPerRound + BirdieConversion +
                       SandSaves + Scrambling, data = dat2)
all_mod[[6]] <- lm(log(PrizeMoney)~GIR + PuttsPerRound + BirdieConversion +
                       SandSaves + Scrambling + DrivingAccuracy, data = dat2)
all_mod[[7]] <- lm(log(PrizeMoney)~., data = dat2)

## calculate AICc, AIC for each model
n <- nrow(dat2)
parta_mat <- do.call("rbind", lapply(all_mod, function(tmp_mod){
```

```
    ## number of parameters (p + 1 coefficients + sigma)
    npar <- length(tmp_mod$coefficients) + 1
    ## Calculate AIC
    AIC <- extractAIC(tmp_mod,k=2)[2]
    ## Calculate AICc
    AICc <- (extractAIC(tmp_mod,k=2)+2*npar*(npar+1)/(n-npar-1))[2]
    ## Calculate BIC to check later
    BIC <- extractAIC(tmp_mod,k=log(n))[2]
    return(c(AIC = AIC, AICc = AICc, BIC = BIC))
}))

## BIC and R^2_adj are included in the summary
xtable(cbind(adjr2 = sum_best$adjr2, parta_mat))
```

% latex table generated in R 3.4.2 by xtable 1.8-2 package % Mon Apr 16 14:35:51 2018

	adjr2	AIC	AICc	BIC
1	0.25	-62.52	-62.39	-55.96
2	0.49	-135.22	-135.01	-125.39
3	0.54	-155.31	-154.99	-142.20
4	0.54	-156.29	-155.85	-139.90
5	0.55	-156.64	-156.05	-136.97
6	0.54	-154.73	-153.96	-131.78
7	0.54	-152.74	-151.77	-126.51

The optimal model based on BIC is

$log(PrizeMoney) = \beta_0 + \beta_1(GIR) + \beta_2(BirdieConversion) + \beta_3(Scrambling) + e.$

The other criterions agree that the best model is

$log(PrizeMoney) = \beta_0 + \beta_1(GIR) + \beta_2(BirdieConversion) + \beta_3(Scrambling) + \beta_4(PuttsPerRound) + \beta_5(SandSaves) + e.$

Part B

```
## fit the full model to step backwards from
om7 <- lm( log(PrizeMoney) ~ ., data = dat2)
backAIC <- step(om7, direction="backward", data=dat2)
```

```
## Start:  AIC=-152.74
## log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion +
##     SandSaves + Scrambling + PuttsPerRound
##
##                     Df Sum of Sq  RSS  AIC
## - PuttingAverage     1      0.00 82.9 -155
## - DrivingAccuracy    1      0.04 82.9 -155
## - PuttsPerRound      1      0.23 83.1 -154
## <none>                           82.9 -153
## - SandSaves          1      1.04 83.9 -152
## - Scrambling         1      1.16 84.0 -152
## - BirdieConversion   1      6.69 89.6 -140
## - GIR                1      9.12 92.0 -134
##
```

```
## Step:  AIC=-154.73
## log(PrizeMoney) ~ DrivingAccuracy + GIR + BirdieConversion +
##     SandSaves + Scrambling + PuttsPerRound
##
##                    Df Sum of Sq  RSS  AIC
## - DrivingAccuracy   1      0.04 82.9 -157
## <none>                         82.9 -155
## - PuttsPerRound     1      1.03 83.9 -154
## - SandSaves         1      1.05 83.9 -154
## - Scrambling        1      1.79 84.7 -153
## - BirdieConversion  1      8.67 91.5 -137
## - GIR               1     17.05 99.9 -120
##
## Step:  AIC=-156.64
## log(PrizeMoney) ~ GIR + BirdieConversion + SandSaves + Scrambling +
##     PuttsPerRound
##
##                    Df Sum of Sq   RSS  AIC
## <none>                          82.9 -157
## - PuttsPerRound     1      1.00  83.9 -156
## - SandSaves         1      1.11  84.0 -156
## - Scrambling        1      1.76  84.7 -154
## - BirdieConversion  1     10.83  93.7 -135
## - GIR               1     20.55 103.5 -115
```

The final model by backwards selection in the last model described above:

$log(PrizeMoney) = \beta_0 + \beta_1(GIR) + \beta_2(PuttsPerRound) + \beta_3(BirdieConversion) + \beta_4(Scrambling) + \beta_5(SandSaves) + e.$

Removing any fifth predictor in the model only increases AIC.

```
## adjust the penalty term by log(n)
backBIC <- step(om7, direction="backward", data=dat2, k=log(n))
```

```
## Start:  AIC=-126.51
## log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage + BirdieConversion +
##     SandSaves + Scrambling + PuttsPerRound
##
##                    Df Sum of Sq  RSS  AIC
## - PuttingAverage    1      0.00 82.9 -132
## - DrivingAccuracy   1      0.04 82.9 -132
## - PuttsPerRound     1      0.23 83.1 -131
## - SandSaves         1      1.04 83.9 -129
## - Scrambling        1      1.16 84.0 -129
## <none>                         82.9 -126
## - BirdieConversion  1      6.69 89.6 -117
## - GIR               1      9.12 92.0 -111
##
## Step:  AIC=-131.78
## log(PrizeMoney) ~ DrivingAccuracy + GIR + BirdieConversion +
##     SandSaves + Scrambling + PuttsPerRound
##
##                    Df Sum of Sq  RSS  AIC
## - DrivingAccuracy   1      0.04 82.9 -137
## - PuttsPerRound     1      1.03 83.9 -135
```

```
## - SandSaves          1      1.05 83.9 -135
## - Scrambling         1      1.79 84.7 -133
## <none>                           82.9 -132
## - BirdieConversion   1      8.67 91.5 -118
## - GIR                1     17.05 99.9 -100
##
## Step:  AIC=-136.97
## log(PrizeMoney) ~ GIR + BirdieConversion + SandSaves + Scrambling +
##     PuttsPerRound
##
##                     Df Sum of Sq   RSS    AIC
## - PuttsPerRound      1      1.00  83.9 -139.9
## - SandSaves          1      1.11  84.0 -139.6
## - Scrambling         1      1.76  84.7 -138.1
## <none>                           82.9 -137.0
## - BirdieConversion   1     10.83  93.7 -118.2
## - GIR                1     20.55 103.5  -98.9
##
## Step:  AIC=-139.9
## log(PrizeMoney) ~ GIR + BirdieConversion + SandSaves + Scrambling
##
##                     Df Sum of Sq   RSS    AIC
## - SandSaves          1       1.3  85.2 -142.2
## <none>                           83.9 -139.9
## - Scrambling         1       7.6  91.5 -128.2
## - GIR                1      35.3 119.2  -76.3
## - BirdieConversion   1      36.6 120.5  -74.3
##
## Step:  AIC=-142.2
## log(PrizeMoney) ~ GIR + BirdieConversion + Scrambling
##
##                     Df Sum of Sq   RSS    AIC
## <none>                           85.2 -142.2
## - Scrambling         1      15.8 101.0 -114.2
## - GIR                1      34.1 119.2  -81.6
## - BirdieConversion   1      40.3 125.5  -71.5
```

The final model by backwards selection in the last model described above:

$log(PrizeMoney) = \beta_0 + \beta_1(GIR) + \beta_2(BirdieConversion) + \beta_3(Scrambling) + e.$

Removing any third predictor in the model only increases BIC.

Part C

```
mint <- lm( log(PrizeMoney) ~ 1, data=dat2 )
## AIC
forwardAIC <- step(mint,
                   scope=list(lower= ~ 1,
                              upper= ~ GIR + PuttsPerRound + BirdieConversion + SandSaves +
                                  Scrambling + DrivingAccuracy + PuttingAverage),
                              direction="forward", data=dat2)
```

```
## Start:  AIC=-6.84
## log(PrizeMoney) ~ 1
```

```
##
##                     Df Sum of Sq RSS   AIC
## + GIR                1      47.8 140 -62.5
## + BirdieConversion   1      40.9 146 -53.2
## + PuttingAverage     1      34.7 153 -44.9
## + Scrambling         1      25.3 162 -33.2
## + SandSaves          1      10.9 176 -16.6
## + PuttsPerRound      1       6.3 181 -11.5
## + DrivingAccuracy    1       6.2 181 -11.4
## <none>                           187  -6.8
##
## Step:  AIC=-62.52
## log(PrizeMoney) ~ GIR
##
##                     Df Sum of Sq   RSS    AIC
## + PuttsPerRound      1      44.2  95.4 -135.2
## + PuttingAverage     1      39.7  99.8 -126.2
## + BirdieConversion   1      38.6 101.0 -124.0
## + SandSaves          1      15.0 124.6  -82.9
## + Scrambling         1      14.1 125.5  -81.4
## <none>                           139.6  -62.5
## + DrivingAccuracy    1       0.2 139.4  -60.8
##
## Step:  AIC=-135.22
## log(PrizeMoney) ~ GIR + PuttsPerRound
##
##                     Df Sum of Sq  RSS  AIC
## + BirdieConversion   1      8.17 87.2 -151
## + DrivingAccuracy    1      2.63 92.7 -139
## + SandSaves          1      1.17 94.2 -136
## + PuttingAverage     1      1.06 94.3 -135
## <none>                           95.4 -135
## + Scrambling         1      0.05 95.3 -133
##
## Step:  AIC=-150.78
## log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion
##
##                    Df Sum of Sq  RSS  AIC
## + Scrambling        1      3.17 84.0 -156
## + SandSaves         1      2.52 84.7 -154
## + PuttingAverage    1      1.26 85.9 -152
## <none>                          87.2 -151
## + DrivingAccuracy   1      0.06 87.1 -149
##
## Step:  AIC=-156.04
## log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion + Scrambling
##
##                    Df Sum of Sq  RSS  AIC
## + SandSaves         1     1.108 82.9 -157
## <none>                          84.0 -156
## + DrivingAccuracy   1     0.099 83.9 -154
## + PuttingAverage    1     0.000 84.0 -154
##
## Step:  AIC=-156.64
```

```
## log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion + Scrambling +
##     SandSaves
##
##                    Df Sum of Sq  RSS  AIC
## <none>                          82.9 -157
## + DrivingAccuracy  1    0.0377 82.9 -155
## + PuttingAverage   1    0.0001 82.9 -155
```

The final model selected in the last model described above:

$log(PrizeMoney) = \beta_0 + \beta_1(GIR) + \beta_2(PuttsPerRound) + \beta_3(BirdieConversion) + \beta_4(Scrambling) + \beta_5(SandSaves) + e.$

Adding any predictor as a sixth predictor in the model only increases AIC.

```
## BIC
forwardBIC <- step(mint,
                   scope=list(lower= ~ 1,
                              upper= ~ GIR + PuttsPerRound + BirdieConversion + SandSaves +
                                  Scrambling + DrivingAccuracy + PuttingAverage),
                              direction="forward", data=dat2, k=log(n))
```

```
## Start:  AIC=-3.56
## log(PrizeMoney) ~ 1
##
##                     Df Sum of Sq RSS   AIC
## + GIR                1      47.8 140 -56.0
## + BirdieConversion   1      40.9 146 -46.6
## + PuttingAverage     1      34.7 153 -38.4
## + Scrambling         1      25.3 162 -26.7
## + SandSaves          1      10.9 176 -10.1
## + PuttsPerRound      1       6.3 181  -5.0
## + DrivingAccuracy    1       6.2 181  -4.9
## <none>                           187  -3.6
##
## Step:  AIC=-55.96
## log(PrizeMoney) ~ GIR
##
##                     Df Sum of Sq   RSS    AIC
## + PuttsPerRound      1      44.2  95.4 -125.4
## + PuttingAverage     1      39.7  99.8 -116.4
## + BirdieConversion   1      38.6 101.0 -114.2
## + SandSaves          1      15.0 124.6  -73.0
## + Scrambling         1      14.1 125.5  -71.5
## <none>                           139.6  -56.0
## + DrivingAccuracy    1       0.2 139.4  -50.9
##
## Step:  AIC=-125.39
## log(PrizeMoney) ~ GIR + PuttsPerRound
##
##                     Df Sum of Sq  RSS  AIC
## + BirdieConversion   1      8.17 87.2 -138
## + DrivingAccuracy    1      2.63 92.7 -126
## <none>                           95.4 -125
## + SandSaves          1      1.17 94.2 -122
## + PuttingAverage     1      1.06 94.3 -122
```

```
## + Scrambling        1      0.05 95.3 -120
##
## Step:  AIC=-137.67
## log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion
##
##                   Df Sum of Sq  RSS  AIC
## + Scrambling       1      3.17 84.0 -140
## + SandSaves        1      2.52 84.7 -138
## <none>                         87.2 -138
## + PuttingAverage   1      1.26 85.9 -135
## + DrivingAccuracy  1      0.06 87.1 -132
##
## Step:  AIC=-139.65
## log(PrizeMoney) ~ GIR + PuttsPerRound + BirdieConversion + Scrambling
##
##                   Df Sum of Sq  RSS  AIC
## <none>                         84.0 -140
## + SandSaves        1     1.108 82.9 -137
## + DrivingAccuracy  1     0.099 83.9 -135
## + PuttingAverage   1     0.000 84.0 -134
```

The final model selected in the last model described above:

$log(PrizeMoney) = \beta_0 + \beta_1(GIR) + \beta_2(PuttsPerRound) + \beta_3(BirdieConversion) + \beta_4(Scrambling) + e.$

Adding any predictor as a fifth predictor in the model only increases BIC.

Part C

Both forward and backward selection are *approximate* methods to find the "best" model. Neither method is guaranteed to find the optimal subset. On the other hand, an exhaustive search (best subsets) will find the optimal predictors given the data and a class of models. From an empirical/philosophical standpoint, the forward method starts with less information (only one predictor, then two predictors, etc) while backwards has all the information to consider before the first selection is made. This would seem to be a favorable situation for the backward selection process if computationally feasible. Then it does not surprise me that backwards and best subsets algorithms agree while forward selection does not (notice that the minimum BIC obtained by forward selection is greater than the backwards selection).

Part D

I'd recommend a final model based on my perspective of the research goal. In Ch.6 (p. 224), the research question was stated as "what is the relative importance of each different aspect of the game on average prize money in professional golf"? As such, I prefer a fuller model to a more simpler model so that parameter estimates corresponding to the different aspects can be compared. With that in mind, I'd recommend the best subsets model with 5 predictors, namely

$log(PrizeMoney) = \beta_0 + \beta_1(GIR) + \beta_2(BirdieConversion) + \beta_3(Scrambling) + \beta_4(PuttsPerRound) + \beta_5(SandSaves) + e.$

If the goal was to *predict* prize money, then I would naively (or perform more sophicated predictive evaluation) go with the BIC model with 3 predictors. In general, it's best to use a model from best subsets as discussed above.

Part E

```
final_model <- lm( log(PrizeMoney) ~ GIR + BirdieConversion + Scrambling +
                     PuttsPerRound + SandSaves, data=dat2 )
summary(final_model)
```

```
##
## Call:
## lm(formula = log(PrizeMoney) ~ GIR + BirdieConversion + Scrambling +
##     PuttsPerRound + SandSaves, data = dat2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.713 -0.482 -0.091  0.448  2.158
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.58318    7.15872   -0.08    0.935
## GIR               0.19702    0.02871    6.86  9.3e-11
## BirdieConversion  0.16275    0.03267    4.98  1.4e-06
## Scrambling        0.04963    0.02474    2.01    0.046
## PuttsPerRound    -0.34974    0.23100   -1.51    0.132
## SandSaves         0.01552    0.00974    1.59    0.113
##
## Residual standard error: 0.661 on 190 degrees of freedom
## Multiple R-squared:  0.557,  Adjusted R-squared:  0.546
## F-statistic: 47.9 on 5 and 190 DF,  p-value: <2e-16
```

Both greens in regulation and birdie conversion are highly statistical significance (slopes are different from zero) and larger values correspond to more prize money. In light of the fact that the model is a result of data-driven variable selection, I'd say that the other three predictors are not different from zero at the 5% significance level. That being said, there is a trend towards better scrambling/sand saves and fewer putts correspond to more prize money.

Project milestones [20 points]

1. Conduct your data analysis plan.

- Apply your model to fake data and ensure a proper fit.
- Apply your model to real data.
- Decide whether model is valid for the real data.

2. Refine your model as needed until you are satisfied with the fit.

- Don't make decisions based on p-values or other inferential devices!
- Only consider the fit and whether your model addresses your research hypothesis.

References

Sheather, Simon. 2009. *A Modern Approach to Regression with R*. Springer Science & Business Media.