MA 575 HW 8

Benjamin Draves

November 14

### Exerices 8.1

#### (a)

#read in data   
pga = read.csv("~/Desktop/Courses/MA 575/book\_data/pgatour2006.csv")  
  
#take a look   
head(pga)

## Name TigerWoods PrizeMoney AveDrivingDistance  
## 1 Aaron Baddeley 0 60661 288.3  
## 2 Adam Scott 0 262045 301.1  
## 3 Alex Aragon 0 3635 302.6  
## 4 Alex Cejka 0 17516 288.8  
## 5 Arjun Atwal 0 16683 287.7  
## 6 Arron Oberholser 0 107294 285.0  
## DrivingAccuracy GIR PuttingAverage BirdieConversion SandSaves  
## 1 60.73 58.26 1.745 31.36 54.80  
## 2 62.00 69.12 1.767 30.39 53.61  
## 3 51.12 59.11 1.787 29.89 37.93  
## 4 66.40 67.70 1.777 29.33 45.13  
## 5 63.24 64.04 1.761 29.32 52.44  
## 6 62.53 69.27 1.775 29.20 47.20  
## Scrambling BounceBack PuttsPerRound  
## 1 59.37 19.30 27.96  
## 2 57.94 19.35 29.28  
## 3 50.78 16.80 29.20  
## 4 54.82 17.05 29.46  
## 5 57.07 18.21 28.93  
## 6 57.67 20.00 29.56

#build model matrix   
mdat = pga[,c(3,5:10,12)]  
mdat[,1] = log(mdat[,1])  
  
#Set up data structures   
best = matrix(0, ncol = 5, nrow = 2^7)  
colnames(best) = c("Model", "R^2", "AIC", "AICc", "BIC")  
n = nrow(pga)  
  
#build aicc function   
AICC = function(model,K, n){  
 return(AIC(model) + 2\*(K)\*(K+1)/(n - K-1))  
}  
  
#build list of all possible subsets   
subsets = lapply(0:7, function(x) combn(7,x))  
  
#fix boundary case   
m = lm(mdat$PrizeMoney ~ 1)  
best[1,] = as.numeric(c(0, summary(m)$adj.r.squared, AIC(m), AICC(m, 1, n), BIC(m)))

To this point we've only prepared to search over all possible subsets of the seven input variables.

#loop over all possible subsets  
counter = 2   
#loop over subsets of size 1, 2,..., 7  
for(i in 2:length(subsets)){   
 #loop over subsets of size i  
 for(j in 1:length(subsets[[i]][1,])){  
   
 #get variable indices we'll be using  
 variable\_indices = subsets[[i]][,j] + 1  
 #build the model  
 m = lm(PrizeMoney~. , data = mdat[,c(1, variable\_indices)])  
   
 #get statistics   
 name = as.numeric(paste(variable\_indices, collapse = ""))  
 r2 = summary(m)$adj.r.squared   
 aic = AIC(m)  
 aicc = AICC(m, 1 + length(variable\_indices),n)  
 bic = BIC(m)  
 best[counter,] = c(name, r2, aic, aicc, bic)  
 counter = counter + 1  
 }  
}  
  
#identify best models  
R2m = best[which.max(best[,2]),1]  
AICm = best[which.min(best[,3]),1]  
AICcm = best[which.min(best[,4]),1]  
BICm = best[which.min(best[,5]),1]  
  
c(R2m, AICm, AICcm, BICm)

## Model Model Model Model   
## 35678 35678 35678 357

From here, we see that , , and all choose the same model referring to

while chooses a more parsimonious model

#### (b)

library(MASS)  
initialM = lm(PrizeMoney ~., data = mdat)  
lowerM = formula(Prize ~1)  
  
AICm = stepAIC(initialM,scope = c(lower = lowerM), direction = "backward")

## Start: AIC=-152.74  
## 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  
## 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  
## 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

BICm = stepAIC(initialM,scope = c(lower = lowerM), direction = "backward", k = log(n))

## Start: AIC=-126.51  
## 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  
## 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  
## 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  
## 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  
## 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

summary(AICm)

##   
## Call:  
## lm(formula = PrizeMoney ~ GIR + BirdieConversion + SandSaves +   
## Scrambling + PuttsPerRound, data = mdat)  
##   
## 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 \*\*\*  
## BirdieConversion 0.162752 0.032672 4.981 1.41e-06 \*\*\*  
## SandSaves 0.015524 0.009743 1.593 0.1127   
## Scrambling 0.049635 0.024738 2.006 0.0462 \*   
## PuttsPerRound -0.349738 0.230995 -1.514 0.1317   
## ---  
## 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

summary(BICm)

##   
## Call:  
## lm(formula = PrizeMoney ~ GIR + BirdieConversion + Scrambling,   
## data = mdat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.71081 -0.50717 -0.06683 0.41975 2.04147   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11.08314 1.45712 -7.606 1.23e-12 \*\*\*  
## GIR 0.15658 0.01787 8.761 1.01e-15 \*\*\*  
## BirdieConversion 0.20625 0.02164 9.531 < 2e-16 \*\*\*  
## Scrambling 0.09178 0.01539 5.965 1.16e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6661 on 192 degrees of freedom  
## Multiple R-squared: 0.5453, Adjusted R-squared: 0.5382   
## F-statistic: 76.75 on 3 and 192 DF, p-value: < 2.2e-16

Here we see that the backwards stepwise and the backwards stepwise approach recover the *best* model as identified in part . That is we see that this process finds that the optimal model given by is

and chooses a more parsimonious model

#### (c)

initialM = lm(PrizeMoney ~1, data = mdat)  
topM = formula(PrizeMoney~DrivingAccuracy +GIR +PuttingAverage +BirdieConversion +SandSaves +Scrambling +PuttsPerRound)  
  
  
AICm = stepAIC(initialM,scope = c(upper = topM), direction = "forward")

## Start: AIC=-6.84  
## 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  
## 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  
## 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  
## 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  
## 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  
## 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

BICm = stepAIC(initialM,scope = c(upper = topM),direction = "forward", k = log(n))

## Start: AIC=-3.56  
## 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  
## 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  
## 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  
## 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  
## 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

summary(AICm)

##   
## Call:  
## lm(formula = PrizeMoney ~ GIR + PuttsPerRound + BirdieConversion +   
## Scrambling + SandSaves, data = mdat)  
##   
## 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

summary(BICm)

##   
## Call:  
## lm(formula = PrizeMoney ~ GIR + PuttsPerRound + BirdieConversion +   
## Scrambling, data = mdat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.68884 -0.49753 -0.07461 0.43648 2.08504   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.39320 7.16112 0.055 0.95627   
## GIR 0.19352 0.02874 6.733 1.89e-10 \*\*\*  
## PuttsPerRound -0.37840 0.23122 -1.637 0.10338   
## BirdieConversion 0.16589 0.03274 5.066 9.52e-07 \*\*\*  
## Scrambling 0.06282 0.02341 2.684 0.00792 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6632 on 191 degrees of freedom  
## Multiple R-squared: 0.5516, Adjusted R-squared: 0.5422   
## F-statistic: 58.74 on 4 and 191 DF, p-value: < 2.2e-16

Here we see that again AIC choose the same model as above but BIC chooses a slightly different model. That is BIC chooses

#### (d)

By the way the forward/backward method proceeds answers why these processes choose different models. The backwards method removes the unneeded variables but has all variables in the model that are significant. For this reason, it always possess the significant variables involved. For this reason, the exhaustive and the backwards models are the same. In the forwards selection we see that Putts Per Round is included in the BIC model in the second iteration of the algorithm. In the next two iterations of the algorithm, we add in two correlated variables that actually make the Putts Per Round variable insignificant. But because we can never remove a variable in forwards selection, we must leave it in the model.

#### (e)

By working with this dataset before, we have high colinearity amongst the variables. This redundancy can lead to inflated varianes of coefficents. For this reason, I choose the most simple model, chosen by BIC, as

where , and .

#### (f)

final\_model = lm(PrizeMoney ~ GIR + BirdieConversion + Scrambling, data = mdat)  
summary(final\_model)

##   
## Call:  
## lm(formula = PrizeMoney ~ GIR + BirdieConversion + Scrambling,   
## data = mdat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.71081 -0.50717 -0.06683 0.41975 2.04147   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11.08314 1.45712 -7.606 1.23e-12 \*\*\*  
## GIR 0.15658 0.01787 8.761 1.01e-15 \*\*\*  
## BirdieConversion 0.20625 0.02164 9.531 < 2e-16 \*\*\*  
## Scrambling 0.09178 0.01539 5.965 1.16e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6661 on 192 degrees of freedom  
## Multiple R-squared: 0.5453, Adjusted R-squared: 0.5382   
## F-statistic: 76.75 on 3 and 192 DF, p-value: < 2.2e-16

Here we see that a one unit increase in GIR correlates to a increase in prize money, a one unit increase in BirdieConversion correlates to a increase in prize money, and a one unit increase in scrambling correlates to a increase in prize money.

We should be cautious. We chose the optimal model using the data we are using to report these values. Therefore, some of the significance in these variables may simply be noise - and we have not corrected for that. If we wanted to be certain of our inferential statements, we would need to have a complete training, validation, and testing set.

### Exercise 8.2

For ease of notation, let , . That is, no subscript implies we are considering the full candidate set, and refers to removing the ith entry. Before, we show our result, we will verify three equation useful in our derivations.

1. . To see this consider the matrices given below
2. X\_{(i)}^{T} Y\_{(i)} = X^{T}Y - x\_iy\_i. To see this we have the following
3. . Notice that . Therefore, corresponds to .

Having these facts, look to show that

Noting that , we can rearrange terms to show this is equivalent to

Using this, consider the following

Now, by the Sherman Morrison formula, with , , and , we see

This implies that

and