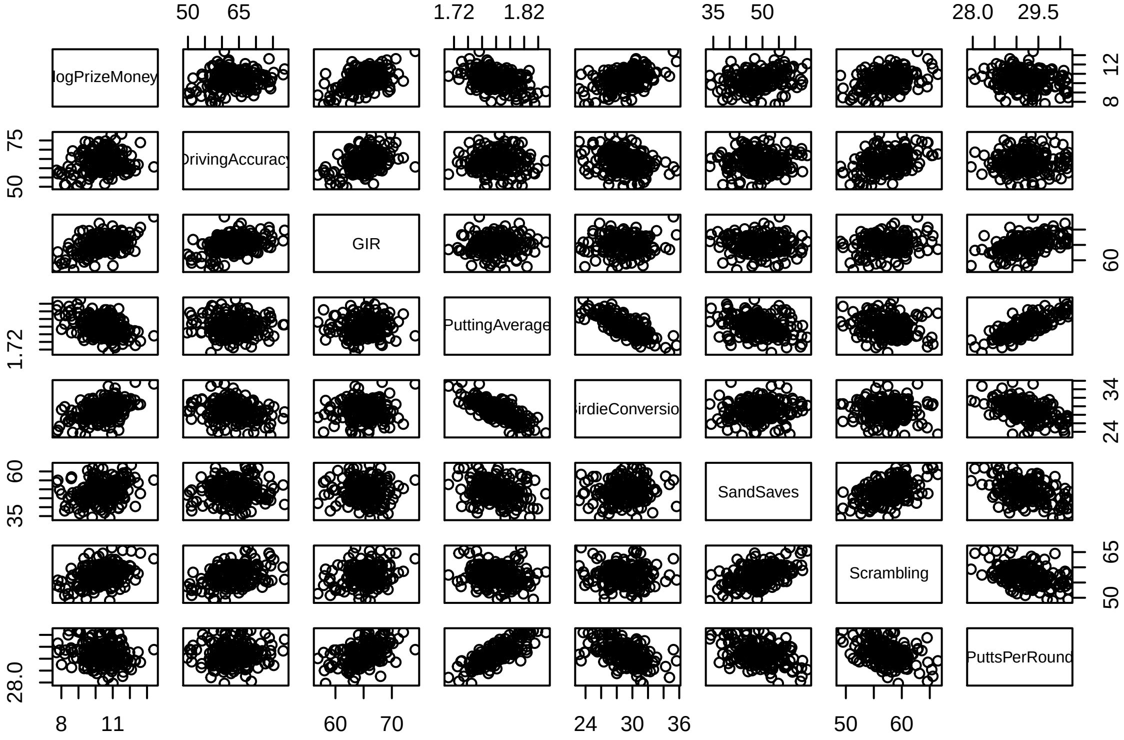
Homework 5

Due on Tuesday, Dec. 5th by 11:00 am

You are expected to solve all the four problems by yourselves. Discussions within your group are encouraged ***but you must write out your own answers in your own words****. Duplicate homework will not receive credit. Make sure you show all of your work and attach your R-script for full credit. The datasets for the problems can be downloaded from Canvas. Please turn in your homework report right before the lecture time.*

1. Recall Problem 3 (Sec. 6.7 Problem 5 in textbook) in homework4, which was trying to answer the following question: *what is the relative importance of each different aspect of the game on average prize money in professional golf?* The data file "pgatour2006.csv" is available on Canvas.

The scatter plot for the data with response variable $\log Y$ is given below.



The full model we'll consider is

$$\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \epsilon.$$

Given the full model above, find the "best" model applying the all possible subsets selection method and then do model validation.

We first randomly split the data to two groups. One is called "training data" for variable selection. The other one is called "test data" for model validation. The codes for splitting the data are given below.

```r
rm(list=ls())
pgatour <- read.csv("./data/pgatour2006.csv", header=TRUE)
pgatour$PrizeMoney <- log(pgatour$PrizeMoney)
```

```r
names(pgatour)[3] <- c("logPrizeMoney")
attach(pgatour)
#the number of predictors in full model
  m <- 7
#sample size
  n <- length(logPrizeMoney)
#split the data
set.seed(1)
train.indx <- sample(1:n, n/2)

#Select training data
y.train <- pgatour[train.indx,3]
x.train <- as.matrix(pgatour[train.indx,c(5:10,12)])

#Select the test data
y.test <- pgatour[-train.indx,3]
x.test <- as.matrix(pgatour[-train.indx,c(5:10,12)])
```

Please continue the analysis in R and answer the following questions.

a). (50 points) Use R to do all possible subsets selections and provide adjusted R^2, AIC, corrected AIC, and BIC. Based on your results, what will be your “best” model and why?

b). (30 points) Use R to do model diagnostics for your final model. Comment on the diagnostic plots for the final model. Based on your comments, is it a valid model?

c). (20 points) Re-estimate the model with the test data using R. Based on your results, are all the selected variables in part a) still significant? Are the fitted coefficients based on the test data consistent with those based on training data?