STAT 757 Assignment 6

DUE 4/08/2018 11:59PM

AG Schissler

2/14/2018

Instructions [20 points]

Modify this file to provide responses to the Ch.6 Exercises in Sheather (2009). You can find some helpful code here: http://www.stat. tamu.edu/~sheather/book/docs/rcode/Chapter6NewMarch2011.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-Assignment6.Rmd and SURNAME-FIRSTNAME-Assignment6.pdf.

Exercise 6.7.5 [60 points]

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

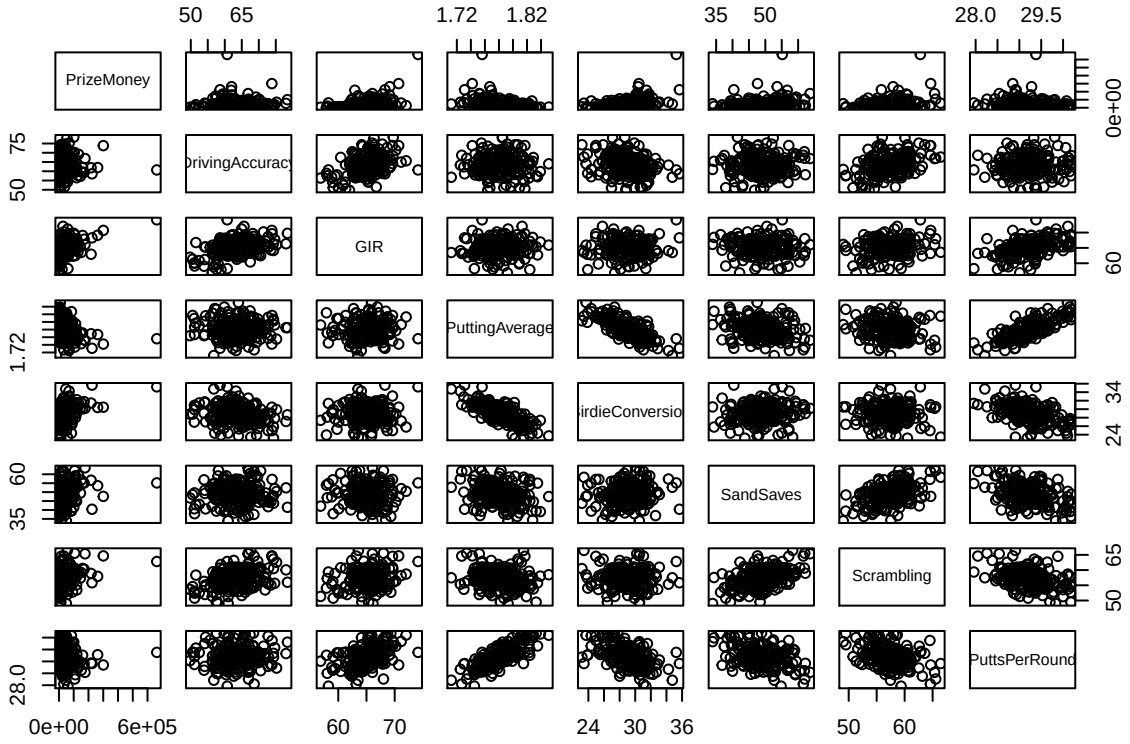
```
## 'data.frame':    196 obs. of  12 variables:
##  $ Name              : Factor w/ 196 levels "Aaron Baddeley",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ TigerWoods        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PrizeMoney        : int  60661 262045 3635 17516 16683 107294 50620 57273 86782 23396 ...
##  $ 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
```

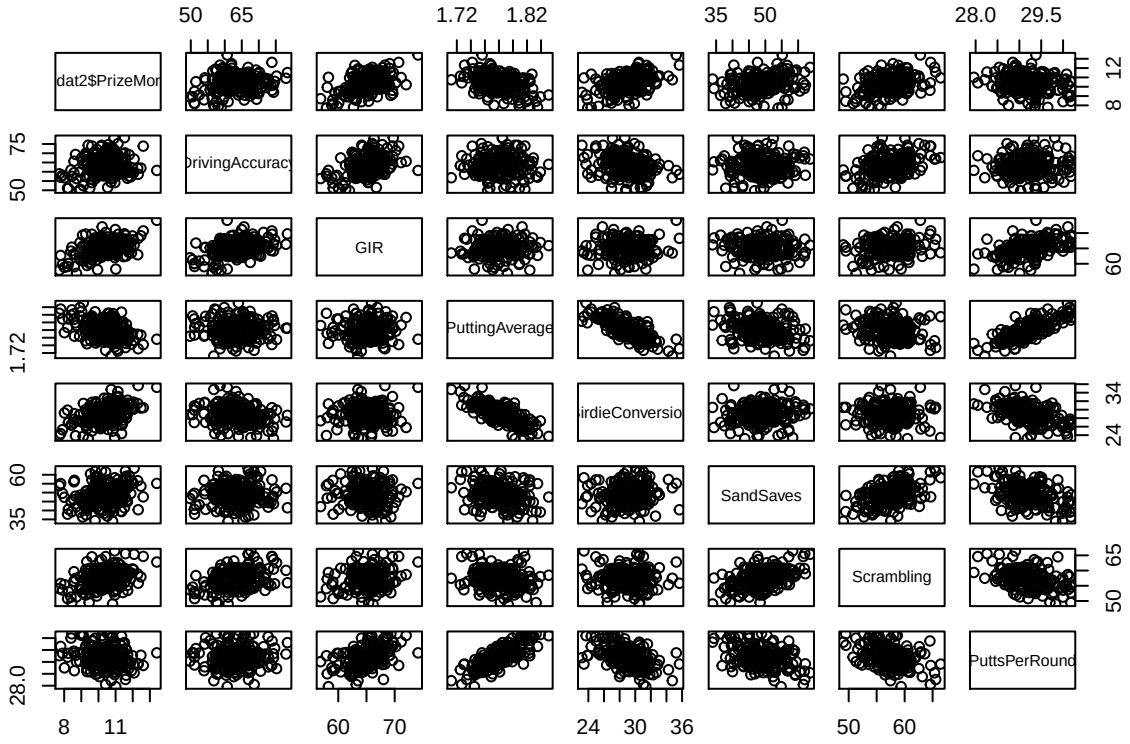
Part A

Based solely on the scatterplots, a log(Y) transformation greatly reduces the skew in Y. All pairs appear Gaussian and so the transformation will likely lead to a good fit. A residual analysis post-fit must be completed to further confirm this approach's validity.

```
pairs(dat2)
```



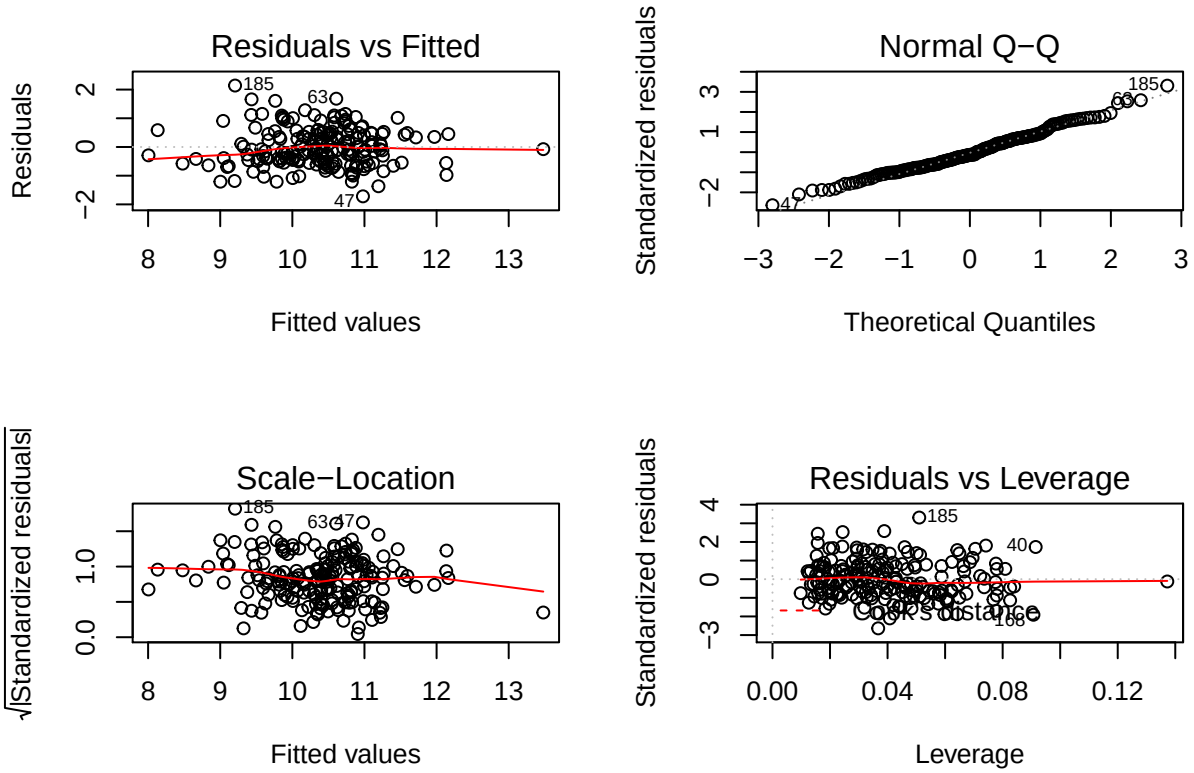

```
pairs(cbind(log(dat2$PrizeMoney), dat2[,-1]))
```



Part B

The fit appears adequate, while errors approximately normally distributed with 0 mean and constant variance.

```
m1 <- lm(log(PrizeMoney) ~ DrivingAccuracy + GIR +
             PuttingAverage + BirdieConversion + SandSaves +
             Scrambling + PuttsPerRound, data = dat2)
par(mfrow = c(2,2))
plot(m1)
```



Part C

No observation has a large Cook's distance based on the Residual vs Leverage plot. So there are no "bad" leverage points. However, row 185 has a standardized residual of 3.3090 which is slightly unusual for data set with 196 observations. The next largest residual, corresponding to row 47, is large (2.6) but arises with the expected probability for this data set. Row 178 inhibits high leverage and corresponds to Tiger Woods (the best golfer during this time). It may be interesting to see how the parameter estimates vary if this point was removed.

```
## standardized residuals
head(sort(abs(rstandard(m1)), decreasing = T), 10)
```

```
##    185     47     63    180      9    122     30    168    101    128
## 3.3090 2.6389 2.5841 2.5306 2.4402 2.1035 1.9448 1.9093 1.8821 1.8791
```

```
1 - pnorm(3.3090)
```

```
## [1] 0.00046815
```

```
1/196
```

```
## [1] 0.005102
```

```
1 - pnorm(2.6389)
```

```
## [1] 0.0041588
```

```
1/196
```

```
## [1] 0.005102
```

```
## leverage
head(sort(hatvalues(m1), decreasing = T), 10)
```

```
##      178       40      168       77       70       16      120      132
## 0.137225 0.091473 0.090696 0.083993 0.082613 0.082597 0.080911 0.078117
##      172      142
## 0.077966 0.077956
```

```
dat[178,]
```

```
##             Name TigerWoods PrizeMoney AveDrivingDistance DrivingAccuracy
## 178 Tiger Woods          1     662771              306.4           60.71
##       GIR PuttingAverage BirdieConversion SandSaves Scrambling BounceBack
## 178 74.15          1.756            35.26     55.17      62.81      24.77
##     PuttsPerRound
## 178         29.38
```

```
## high leverage cutoff
(2*8)/196
```

```
## [1] 0.081633
```

Part D

Examining the model summary below, we see that overall the model is significant with $F = 33.9$ with a p-value essentially zero. However, only two of the seven predictors are significant. Variable selection (Ch.7) will help rememdy this situation.

```
summary(m1)
```

```
##
## Call:
## lm(formula = log(PrizeMoney) ~ DrivingAccuracy + GIR + PuttingAverage +
##     BirdieConversion + SandSaves + Scrambling + PuttsPerRound,
##     data = dat2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7195 -0.4861 -0.0917  0.4456  2.1401
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.19430    7.77713    0.02  0.98009
## DrivingAccuracy   -0.00353    0.01177   -0.30  0.76464
## GIR                0.19931    0.04382    4.55  9.7e-06
## PuttingAverage    -0.46630    6.90570   -0.07  0.94624
## BirdieConversion   0.15734    0.04038    3.90  0.00014
## SandSaves          0.01517    0.00986    1.54  0.12555
## Scrambling         0.05151    0.03179    1.62  0.10679
## PuttsPerRound     -0.34313    0.47355   -0.72  0.46960
##
```

```
## Residual standard error: 0.664 on 188 degrees of freedom
## Multiple R-squared:  0.558,  Adjusted R-squared:  0.541
## F-statistic: 33.9 on 7 and 188 DF,  p-value: <2e-16
```

Part E

Removing all the non-significant predictors at once is a poor idea. Correlations among the predictors could mask relationships between $PrizeMoney$ and other predictors. Later, we'll see that correlation between predictors inflates the variance of regression estimates, leading to poor confidence intervals/hypothesis test results.

Project milestones [20 points]

1. Prepare a data analysis plan.

- What model(s) will you use?
- How will you fit this model (code)?
- How will you generate fake data from this model?
- What model diagnostics will you use?
- How will you refine the model? Or select from competing models?

References

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