First Name: Susheela

Last Name: Polepalli

M-ID: M10727836

Homework 4

Generate a report to answer the following questions.

1. Read PGA data into R (PGA.csv). Below is the description of variables.

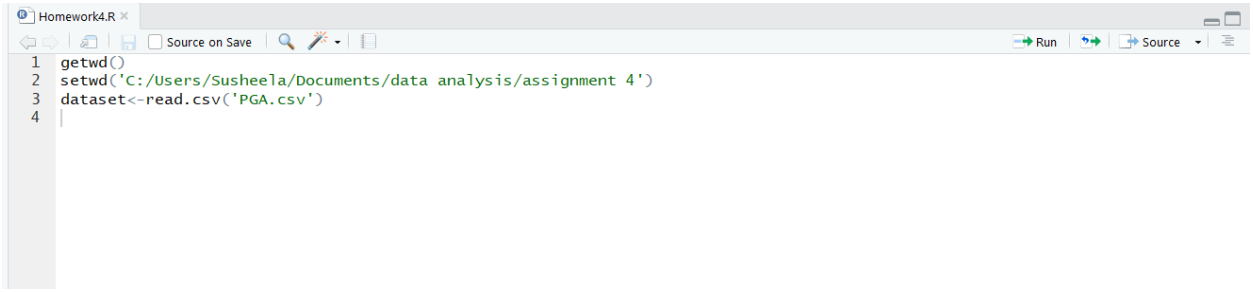
Source: sportsillustrated.cnn.com

Description: Performance statistics and winnings for 196 PGA participants during, 2004 season.

Variable: Name, Age, Average Drive (Yards), Driving accuracy (percent), Greens on regulation (%), Average # of putts, Save Percent, Money Rank, # Events, Total Winnings ($), Average winnings ($).

Solution:

Code:



2. Visualize the data using scatter plot and histogram.

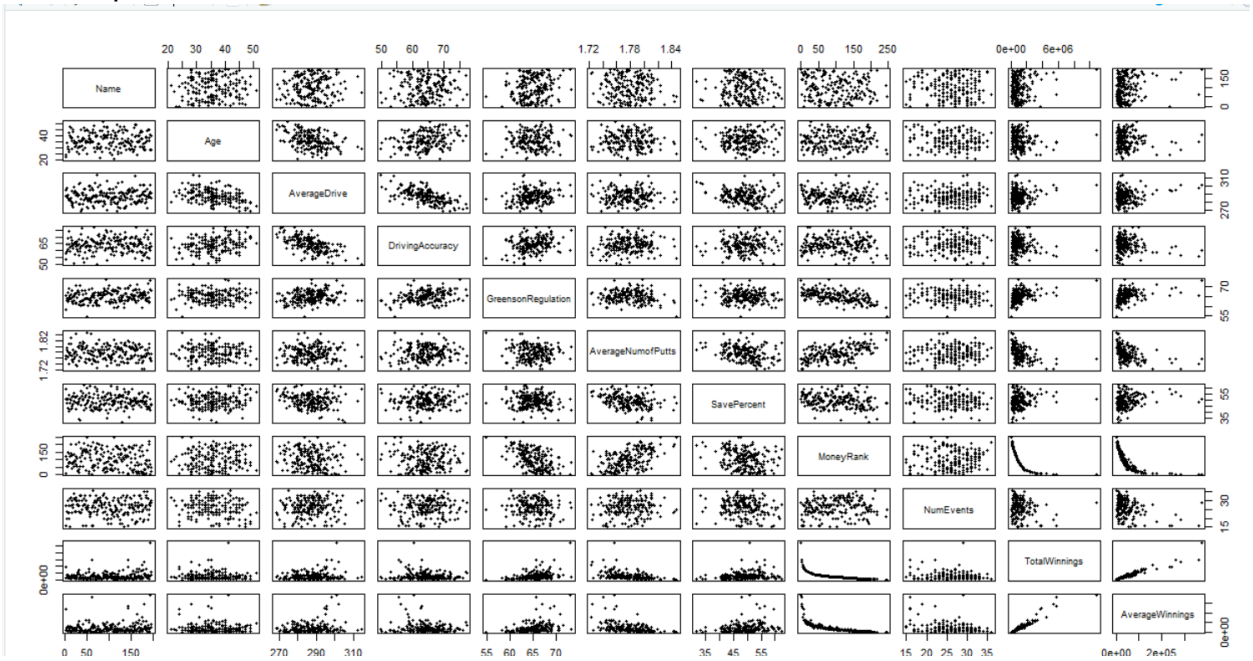
Solution:

Code:

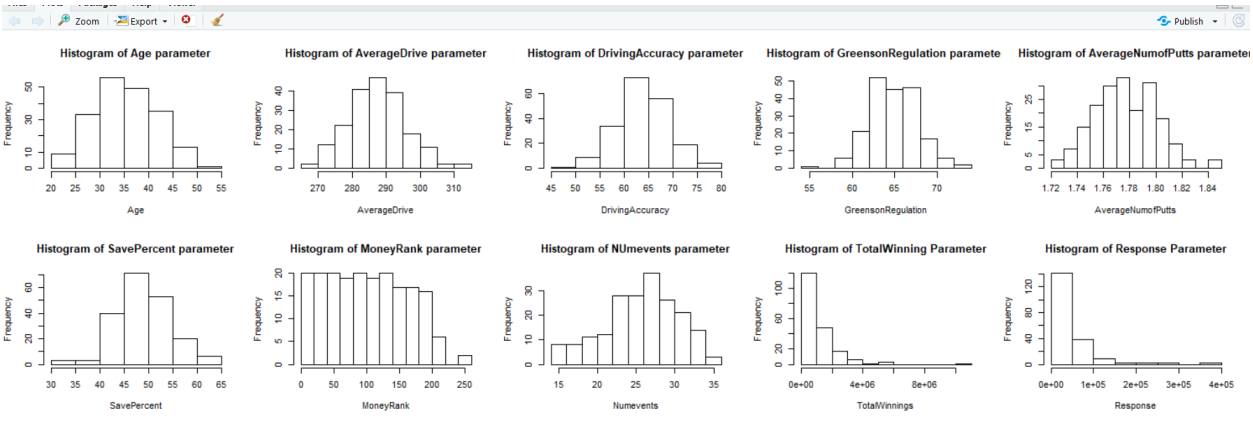
```r
getwd()
setwd('C:/Users/Susheela/Documents/data analysis/assignment 4')
dataset<-read.csv('PGA.csv',header=TRUE)
names(dataset)
pairs(dataset,pch=20)  #scatter plot
par(mfrow=c(2,5))
hist(as.numeric(dataset$Age),xlab="Age",main="Histogram of Age parameter")
hist(as.numeric(dataset$AverageDrive),xlab="AverageDrive",main="Histogram of AverageDrive parameter")
hist(as.numeric(dataset$DrivingAccuracy),xlab="DrivingAccuracy",main="Histogram of DrivingAccuracy parameter")
hist(as.numeric(dataset$GreensonRegulation),xlab="GreensonRegulation",main="Histogram of GreensonRegulation parameter")
hist(as.numeric(dataset$AverageNumofPutts),xlab="AverageNumofPutts",main="Histogram of AverageNumofPutts parameter")
hist(as.numeric(dataset$SavePercent),xlab="SavePercent",main="Histogram of SavePercent parameter")
hist(as.numeric(dataset$MoneyRank),xlab="MoneyRank",main="Histogram of MoneyRank parameter")
hist(as.numeric(dataset$Numevents),xlab="Numevents",main="Histogram of NUmevents parameter")
hist(as.numeric(dataset$TotalWinnings),xlab="TotalWinnings",main="Histogram of TotalWinning Parameter")
hist(as.numeric(dataset$Response),xlab="Response",main="Histogram of Response Parameter")
```

Output:

Scatter plot:



Histogram:



3. Build a linear regression using Average winnings as response variable and using Age, Average Drive (Yards), Driving accuracy (percent), Greens on regulation (%), Average # of putts, Save Percent, and # Events as covariates.

Solution:

Code:

```
getwd()
setwd('C:/Users/Susheela/Documents/data analysis/assignment 4')
dataset<-read.csv('PGA.csv',header=TRUE)
names(dataset)
pairs(dataset,pch=20)
names(dataset)[1]="Name"
names(dataset)[2]="Age"
names(dataset)[3]="AverageDrive"
names(dataset)[4]="DrivingAccuracy"
names(dataset)[5]="GreensonRegulation"
names(dataset)[6]="AverageNumofPutts"
names(dataset)[7]="SavePercent"
names(dataset)[8]="MoneyRank"
names(dataset)[9]="Numevents"
names(dataset)[10]="TotalWinnings"
names(dataset)[11]="Response"
attach(dataset)
model1<-lm(Response~ Age+AverageDrive+DrivingAccuracy+GreensonRegulation+AverageNumofPutts+SavePercent+Numevents,data=dataset)
```

4. Perform t tests for these coefficient estimates. Obtain t statistics and p values, interpret the results, make a conclusion (i.e. reject or not reject) and explain why. Note: please explain what the null hypothesis is.

Solution:

Code:

```
model1<-lm(Response~ Age+AverageDrive+DrivingAccuracy+GreensonRegulation+AverageNumofPutts+SavePercent+Numevents,data=dataset)
summary(model1)$coef[,3]   # t values
summary(model1)$coef[,4]   # p values
```

Output:

```
> summary(model1)$coef[,3]
      (Intercept)               Age       AverageDrive    DrivingAccuracy GreensonRegulation  AverageNumofPutts        SavePercent
        3.0912760        -1.1305567          -0.1670039        -2.7640660          6.4930148         -5.0249465          2.3754460
        Numevents
       -4.9037814
> summary(model1)$coef[,4]
      (Intercept)               Age       AverageDrive    DrivingAccuracy GreensonRegulation  AverageNumofPutts        SavePercent
     2.296050e-03      2.596820e-01       8.675464e-01       6.276772e-03       7.300592e-10       1.167423e-06       1.853368e-02
        Numevents
     2.026906e-06
```

The t-value for intercept is 3.091 and p-value is 2.296050e-03. This means slope is different from 0. Thus, the null hypothesis H0: β1=0 is rejected.

5. Use F test to test the significance of the regression. Obtain the F statistic and p value, interpret the results and make a conclusion.

Code:

```
summary(model1)
```

Output:

```
> summary(model1)

Call:
lm(formula = Response ~ Age + AverageDrive + DrivingAccuracy +
    GreensonRegulation + AverageNumofPutts + SavePercent + Numevents,
    data = dataset)

Residuals:
   Min     1Q Median     3Q    Max
-71690 -22176  -6735  17147 247928

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         945579.88  305886.59   3.091  0.00230 **
Age                   -587.13     519.32  -1.131  0.25968
AverageDrive           -94.76     567.42  -0.167  0.86755
DrivingAccuracy      -2360.57     854.02  -2.764  0.00628 **
GreensonRegulation    8466.04    1303.87   6.493 7.30e-10 ***
AverageNumofPutts  -694226.49  138155.99  -5.025 1.17e-06 ***
SavePercent           1395.67     587.54   2.375  0.01853 *
Numevents            -3159.22     644.24  -4.904 2.03e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41430 on 188 degrees of freedom
Multiple R-squared:  0.4527,    Adjusted R-squared:  0.4323
F-statistic: 22.21 on 7 and 188 DF,  p-value: < 2.2e-16
```

The F-value is 22.21383 and p value is < 2.2e – 16. The p value is less than 0.05. Thus, the null hypothesis is H0: β1=0 is rejected.

6. Use a partial F test to test for two variables Age and Average Drive (Yards) together. According to your results, what do you conclude? Similarly, use the partial F test to test for three variables Age, Average Drive (Yards), and Save Percent together, what do you conclude?

Solution:

Code:

```
22 model2<-lm(Response~ Age+AverageDrive,data=dataset)
23 anova(model1,model2)
24 model3<-lm(Response ~ Age + AverageDrive + SavePercent,data=dataset)
25 anova(model1,model3)
26
```

Output:

```
> anova(model1,model2)
Analysis of Variance Table

Model 1: Response ~ Age + AverageDrive + DrivingAccuracy + GreensonRegulation +
    AverageNumofPutts + SavePercent + Numevents
Model 2: Response ~ Age + AverageDrive
  Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
1    188 3.2273e+11
2    193 5.6610e+11 -5 -2.4336e+11 28.353 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(model1,model3)
Analysis of Variance Table

Model 1: Response ~ Age + AverageDrive + DrivingAccuracy + GreensonRegulation +
    AverageNumofPutts + SavePercent + Numevents
Model 2: Response ~ Age + AverageDrive + SavePercent
  Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
1    188 3.2273e+11
2    192 5.2941e+11 -4 -2.0668e+11 30.099 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

7. Obtain the interval estimation for all the intercept and slope coefficients.

Solution:

```
> confint(model1,level=0.95)
                           2.5 %       97.5 %
(Intercept)          342168.8016 1548990.9491
Age                   -1611.5765     437.3259
AverageDrive          -1214.0883    1024.5657
DrivingAccuracy       -4045.2641    -675.8748
GreensonRegulation     5893.9435   11038.1279
AverageNumofPutts   -966761.6620 -421691.3083
SavePercent            236.6508    2554.6825
Numevents             -4430.0971   -1888.3510
```

8. Using the regression in question 3, make a prediction for the case of:

Age = 35,

AverageDrive = 287,

DrivingAccuracy = 64,

GreensonRegulation = 64.9,

AverageNumofPutts = 1.778,

SavePercent = 48,

NumEvents = 26,

The prediction should include fitted value and interval estimation.

Solution:

Code:

```
predict(model1,level=.95,list(Age = 35,
                               AverageDrive = 287,
                               DrivingAccuracy = 64,
                               GreensonRegulation = 64.9,
                               AverageNumofPutts = 1.778,
                               SavePercent = 48,
                               Numevents = 26),interval="confidence")
```

Output:

```
> predict(model1,level=.95,list(Age = 35,
+                                AverageDrive = 287,
+                                DrivingAccuracy = 64,
+                                GreensonRegulation = 64.9,
+                                AverageNumofPutts = 1.778,
+                                SavePercent = 48,
+                                Numevents = 26),interval="confidence")
       fit     lwr      upr
1 46720.76 40657.8 52783.72
```

9. Similarly, make another prediction for the case of

Age = 42,

AverageDrive = 295,

DrivingAccuracy = 69,

GreensonRegulation = 67.7,

AverageNumofPutts = 1.80,

SavePercent = 54,

NumEvents = 30,

The prediction should again include the fitted value and interval estimation. Compare the interval from question 8, what do you observe? For example, which interval is wider? And why?

Solution:

Code:

```
predict(model1,level=.95,list(Age = 42,
                               AverageDrive = 295,
                               DrivingAccuracy = 69,
                               GreensonRegulation = 67.7,
                               AverageNumofPutts = 1.80,
                               SavePercent = 54,
                               Numevents = 30),interval="confidence")
```

Output:

```
> predict(model1,level=.95,list(Age = 42,
+                               AverageDrive = 295,
+                               DrivingAccuracy = 69,
+                               GreensonRegulation = 67.7,
+                               AverageNumofPutts = 1.80,
+                               SavePercent = 54,
+                               Numevents = 30),interval="confidence")
       fit      lwr      upr
1 34218.97 14565.55 53872.39
>
```

It is observed that the interval has become wider than the interval obtained in question 8.

10. Obtain the standardized regression coefficients and compare the influence of all variables.

Solution:

Code:

```
41  model1
42
```

Output:

```
> model1

Call:
lm(formula = Response ~ Age + AverageDrive + DrivingAccuracy +
    GreensonRegulation + AverageNumofPutts + SavePercent + Numevents,
    data = dataset)

Coefficients:
       (Intercept)                Age        AverageDrive     DrivingAccuracy  GreensonRegulation   AverageNumofPutts         SavePercent
         945579.88             -587.13              -94.76            -2360.57              8466.04          -694226.49             1395.67
         Numevents
          -3159.22

>
```