

PGA Tour Statistics

Data Munging

```
all <- read.csv("./data/PGA Stats.csv", header = FALSE, stringsAsFactors = F)
header <- all[c(8, 13, 18, 23, 33, 38, 104, 56, 61), 1]
header <- append(c("Name", "Year"), header)
```

Example of creating the player data from the csv

```
#### Tiger####
temp1 <- all[c(1, 2, 9, 14, 19, 24, 34, 39, 105, 57, 62), 1]
temp2 <- all[c(1, 2, 9, 14, 19, 24, 34, 39, 110, 57, 67), 2]
temp3 <- all[c(1, 2, 9, 14, 19, 24, 34, 39, 106, 53, 63), 3]
temp <- rbind(temp1, temp2, temp3)
temp4 <- rbind(header, temp)
temp4 <- as.data.frame(temp, stringsAsFactors = F)
colnames(temp4) <- header
# remove $ , % from data
for (i in grep("%", temp4)) temp4[, i] <- gsub("%", "", temp4[, i])
for (i in grep("\\$", temp4)) temp4[, i] <- gsub("\\$", "", temp4[, i])
for (i in grep(",", temp4)) temp4[, i] <- gsub(",", "", temp4[, i])
for (i in 2:11) temp4[, i] <- as.numeric(temp4[, i])
tiger <- temp4
```

##	Name	Year	Driving Distance	Driving Accuracy	Percentage
## temp1	Tiger Woods	2013	291.6		61.85
## temp2	Tiger Woods	2012	297.4		63.93
## temp3	Tiger Woods	2011	293.7		48.90
##	Greens in Regulation Percentage Strokes Gained - Putting				
## temp1			67.55		0.835
## temp2			67.58		0.332
## temp3			67.74		0.258
##	Birdie Average Scoring Average Scoring Average (Actual)				
## temp1		4.00	68.65		70.09
## temp2		3.97	68.90		69.78
## temp3		3.92	70.46		70.77
##	FedExCup Season Points Money Leaders				
## temp1		3059	7687119		
## temp2		2269	6133158		
## temp3		318	660238		

Set Testing Data

```
allPlayers <- rbind(baddeley, bradely, clark, crane, duval, fowler, furyk,  
  kuchar,  
    ohair, tiger, watson)  
testData <- allPlayers[, -c(1, 2, 8, 9, 10)]  
colnames(testData) <- c("drivingDistance", "drivingAccuracyPercentage",  
  "greensRegulationPercentage",  
    "Putting", "birdieAverage", "Money")
```

Test Regression Models

AIC

```
aicFormula <- step(lm1.1)
```

```
## Start:  AIC=968.3
## Money ~ drivingDistance + drivingAccuracyPercentage +
greensRegulationPercentage +
##      Putting + birdieAverage
##
##              Df Sum of Sq      RSS AIC
## - greensRegulationPercentage  1  9.82e+10 5.59e+13 966
## <none>                        5.58e+13 968
## - birdieAverage              1  6.48e+12 6.23e+13 970
## - drivingAccuracyPercentage  1  7.21e+12 6.30e+13 970
## - drivingDistance           1  1.14e+13 6.72e+13 973
## - Putting                   1  1.87e+13 7.45e+13 976
##
## Step:  AIC=966.4
## Money ~ drivingDistance + drivingAccuracyPercentage + Putting +
##      birdieAverage
##
##              Df Sum of Sq      RSS AIC
## <none>                        5.59e+13 966
## - birdieAverage              1  7.12e+12 6.30e+13 968
## - drivingAccuracyPercentage  1  1.47e+13 7.06e+13 972
## - drivingDistance           1  1.82e+13 7.41e+13 974
## - Putting                   1  1.86e+13 7.45e+13 974
```

BIC

```
Posterior probabilities(%):
  drivingDistance drivingAccuracyPercentage greensRegulationPercentage
Putting          birdieAverage
94.7             84.6             59.0             75.0             28.3
```

```
Coefficient posterior expected values:
(Intercept) drivingDistance drivingAccuracyPercentage
greensRegulationPercentage Putting birdieAverage
-24235511      69202      85475
17339      1898624      -20312
```

Compare Different Inputs

Test model differences for different inputs

```
lm2.1 <- lm(Money ~ drivingDistance + drivingAccuracyPercentage +
greensRegulationPercentage +
  Putting + birdieAverage, data = testData)
lm3.1 <- lm(Money ~ drivingDistance + drivingAccuracyPercentage + Putting, data
= testData)
lm4.1 <- lm(Money ~ drivingDistance + drivingAccuracyPercentage +
greensRegulationPercentage +
  Putting, data = testData)
lm5.1 <- lm(Money ~ drivingDistance + drivingAccuracyPercentage + Putting +
  birdieAverage, data = testData)
```

Example of output from first model

```
##  
## Call:  
## lm(formula = Money ~ drivingDistance + drivingAccuracyPercentage +  
##     greensRegulationPercentage + Putting + birdieAverage, data = testData)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -1670499  -798762  -261860   627231  3639490   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)   -28322090    9875466  -2.87   0.0078 **   
## drivingDistance      84349     35235    2.39   0.0236 *    
## drivingAccuracyPercentage  120383     63298    1.90   0.0675 .    
## greensRegulationPercentage -19456     87617   -0.22   0.8259      
## Putting        2108450     688569    3.06   0.0048 **   
## birdieAverage    -29622     16432   -1.80   0.0822 .    
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 1410000 on 28 degrees of freedom  
## Multiple R-squared:  0.498, Adjusted R-squared:  0.408   
## F-statistic: 5.55 on 5 and 28 DF, p-value: 0.00113
```

Test Regression Models without Duval's Rows

Remove David Duval's results because they are poor and skew the models

AIC

```
aicFormula <- step(lm1.2)
```

```
## Start:  AIC=872.5
## Money ~ drivingDistance + drivingAccuracyPercentage +
## greensRegulationPercentage +
## Putting + birdieAverage
##
##              Df Sum of Sq      RSS AIC
## <none>                3.52e+13  873
## - greensRegulationPercentage  1  5.69e+12  4.09e+13  875
## - drivingAccuracyPercentage  1  1.04e+13  4.56e+13  879
## - drivingDistance            1  1.15e+13  4.68e+13  879
## - birdieAverage              1  1.57e+13  5.09e+13  882
## - Putting                    1  3.18e+13  6.71e+13  890
```

```
##
## Call:
## lm(formula = Money ~ drivingDistance + drivingAccuracyPercentage +
## greensRegulationPercentage + Putting + birdieAverage, data = testNoDuval)
##
## Coefficients:
##              (Intercept)              drivingDistance
##              -47597856                93227
## drivingAccuracyPercentage greensRegulationPercentage
##              156075                199216
##              Putting              birdieAverage
##              3544187              -82081
```

BIC

Posterior probabilities(%):				
drivingDistance	drivingAccuracyPercentage	greensRegulationPercentage		
Putting	birdieAverage			
	82.1		80.1	64.8
100.0	95.2			
Coefficient posterior expected values:				
(Intercept)	drivingDistance	drivingAccuracyPercentage		
greensRegulationPercentage	Putting	birdieAverage		
-42365071	85704		140407	167625
3350533	-73739			

Compare Different Inputs

Test model differences for different inputs

```
lm2.2 <- lm(Money ~ drivingDistance + drivingAccuracyPercentage +  
greensRegulationPercentage +  
    Putting + birdieAverage, data = testNoDuval)  
lm3.2 <- lm(Money ~ drivingDistance + drivingAccuracyPercentage + Putting, data  
= testNoDuval)  
lm4.2 <- lm(Money ~ drivingDistance + drivingAccuracyPercentage +  
greensRegulationPercentage +  
    Putting, data = testNoDuval)  
lm5.2 <- lm(Money ~ drivingDistance + drivingAccuracyPercentage + Putting +  
    birdieAverage, data = testNoDuval)
```

Example of output from first model

```
##  
## Call:  
## lm(formula = Money ~ drivingDistance + drivingAccuracyPercentage +  
##     greensRegulationPercentage + Putting + birdieAverage, data = testNoDuval)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -2053184  -719272  -136336   467516  2434317  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)   -47597856   10151532  -4.69   8.3e-05 ***  
## drivingDistance     93227     32592    2.86   0.0084 **  
## drivingAccuracyPercentage  156075     57577    2.71   0.0120 *  
## greensRegulationPercentage  199216     99124    2.01   0.0554 .  
## Putting        3544187     745787    4.75   7.1e-05 ***  
## birdieAverage    -82080      24584   -3.34   0.0026 **  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 1190000 on 25 degrees of freedom  
## Multiple R-squared:  0.63,    Adjusted R-squared:  0.556  
## F-statistic: 8.53 on 5 and 25 DF,  p-value: 8.1e-05
```