US Open Men 2013 Analysis

April 16, 2017

Contents

1 Summary

- Predicting the results of sporting matches is complicated business. Factors that cannot be captured by data such as crowd enthusiasm, home field advantage, and rivalries do exist and have an impact on the outcome.
- We began our analysis of the 2013 US Men's Open with 42 variables broken into one response variables, 18 un-useful variables, 1 un-useful categorical variable and 23 possible covariates. Breaking the data 70/30 into training and testing data, we developed a prediction model utilizing 22 covariates to predict the Final Games Won for Player 1. This model had an R^2 of 88.15% and an Adjusted R^2 of 84.62%–there was someone work to be done in slimming the number of predictor variables.
- Through Partial F Testing, we eliminated 17 variables from our model to be left with BCP.1(break points created by player 1), BPW.1(break points won by player 1), TPW.1(total points won by player 1), TPW.2(total points won by player 2).
- Due to the nature of our data, we assumed early on that multicollinearity would be an issue. We found, using Variance Inflation Factors (VIF), that TPW.1 and TPW.2 were correlated. Outside of the data, this makes sense as the more points player one wins, the more points player two will no win. Also, both final points for the players will increase due to the nature of a tennis game. To adjust for this multicollinearity, we removed TPW.1 from the model.
- Our final model, **FNL1= 0.6722 BPC.1 + 0.1892 TPW.1** , suggests that As the number of break points created by player 1 increases by 1, the final number of games won by player 1 0.6722 As player 2's total number of points increases by 1, the final number of games won by player 1 0.1892
- After our final model was built, we applied it to the testing data. The testing data was originally partitioned as 30% of the original data prior to the model building. The data was tested using the three iterations of the model we created. The Mean Squared Error of the predicted values versus the actual values of the testing data was determined. As we refined our model to the final two variables of BPC.1(break points created for player one) and TPW.1(total points won by player one) the MSE reached a low point.
- In conclusion, predicting the outcome of a tennis match based on the provided variables will be highly uncertain. We would suggest collecting more data about player's statistics (such as player 1 age, titles, etc) in an attempt to further refine a model for prediction.

2 Business Understanding

This is a public, multivariate dataset concerning the US Men's Tennis Open in 2013. Our data is sourced from UCI Machine Learning Repository (Jauhari, Morankar, Fokoue)

(https://archive.ics.uci.edu/ml/datasets/Tennis+Major+Tournament+Match+Statistics).

It has 42 variables in all. Each observation represents a singles tennis match.

In tennis, a match is composed of two to three sets. To win a match, a player must win two sets. A set is composed of at least six games. To win a set, one must win at least six games, with a two game advantage over the games won by the opposing player. The scoring of games in tennis is a little odd. The score begins 0-0 or "love." After the first point is scored, the score is "love"-15. The scoring continues sequentially, "love"(0 points), 15(2 points), 30(3 points), 40(4 points). The game must be won by two points. If the score is 40-40, that is called 40-all or "deuce," the one player must score an additional two points consecutively to win the game.

The response variable which we will be attempting to predict will by FNL.1 (the final number of games won by player 1). In predicting this variable, some other variables become obsolete such as Result and STX.Y.

Four other variables that we will not include in our analysis are WNR.X and UFE.X as no data populates those fields. Below is the complete list of variables and their uses.

2.1 Variables in the data set

Use	Variable	Type	Description
None	Player 1		Name of player 1
None	Player 2		Name of player 2
None	Result	Ordinal (0/1)	Referenced on player 1. If player 1 wins- result=1 (FNL.1>FNL.2)
Covariate	FSP.1	Real Number	First serve percentage for player 1
Covariate	FSW.1	Real Number	First serve won by player 1
Covariate	SSP.1	Real Number	Second serve percentage for player 1
Covariate	SSW.1	Real Number	Second serve won by player 1
Covariate	ACE.1	Numeric-Integer	Aces won by player 1
Covariate	DBF.1	Numeric-Integer	Double faults committed by player 1
None	WNR.1	Numeric	Winners earned by player 1
None	UFE.1	Numeric	Unforced errors committed by player 1
Covariate	BPC.1	Numeric	Break points created by player 1
Covariate	BPW.1	Numeric	Break points won by player 1
Covariate	NPA.1	Numeric	Net points attempted by player 1
Covariate	NPW.1	Numeric	Net points won by player 1
Covariate	TPW.1	Numeric	Total points won by player 1
None	ST1.1	Numeric-Integer	Set 1 Result for player 1
None	ST2.1	Numeric-Integer	Set 2 Result for player 1
None	ST3.1	Numeric-Integer	Set 3 Result for player 1
None	ST4.1	Numeric-Integer	Set 4 Result for player 1
None	ST5.1	Numeric-Integer	Set 5 Result for player 1
Response	FNL.1	Numeric-Integer	Final number of games won by player 1
Covariate	FSP.1	Real Number	First serve percentage for player 2
Covariate	FSW.2	Real Number	First serve won by player 2
Covariate	SSP.2	Real Number	Second servce percentage for player 2
Covariate	SSW.2	Real Number	Second serve won by player 2
Covariate	ACE.2	Numeric-Integer	Aces won by player 2
Covariate	DBF.2	Numeric-Integer	Double faults committed by player 2
None	WNR.2	Numeric	Winners earned by player 2
None	UFE.2	Numeric	Unforced errors committed by player 2
Covariate	BPC.2	Numeric	Break points created by player 2
Covariate	BPW.2	Numeric	Break points won by player 2
Covariate	NPA.2	Numeric	Net points attempted by player 2
Covariate	NPW.2	Numeric	Net points won by player 2
Covariate	TPW.2	Numeric	Total points won by player 2
None	ST2.1	Numeric-Integer	Set 1 Result for player 2
None	ST2.2	Numeric-Integer	Set 2 Result for player 2
None	ST2.3	Numeric-Integer	Set 3 Result for player 2
None	ST2.4	Numeric-Integer	Set 4 Result for player 2
None	ST2.5	Numeric-Integer	Set 5 Result for player 2
None	FNL.2	Numeric-Integer	Final number of games won by player 2
Covariate	Round	Numeric-Integer	Round of the tournament at which the game is played

3 Data Understanding

- We will use `dim` function to get the dimensions of the data
- `str` function to get the variable information
- `head` function to read first few observations
- `summary` function to get a summary of the data set

3.1 Data set summary

```
usopen <- read.csv("USOpen-men-2013.csv")
dim(usopen)
```

```
## [1] 126  42
```

```
str(usopen)
head(usopen)
```

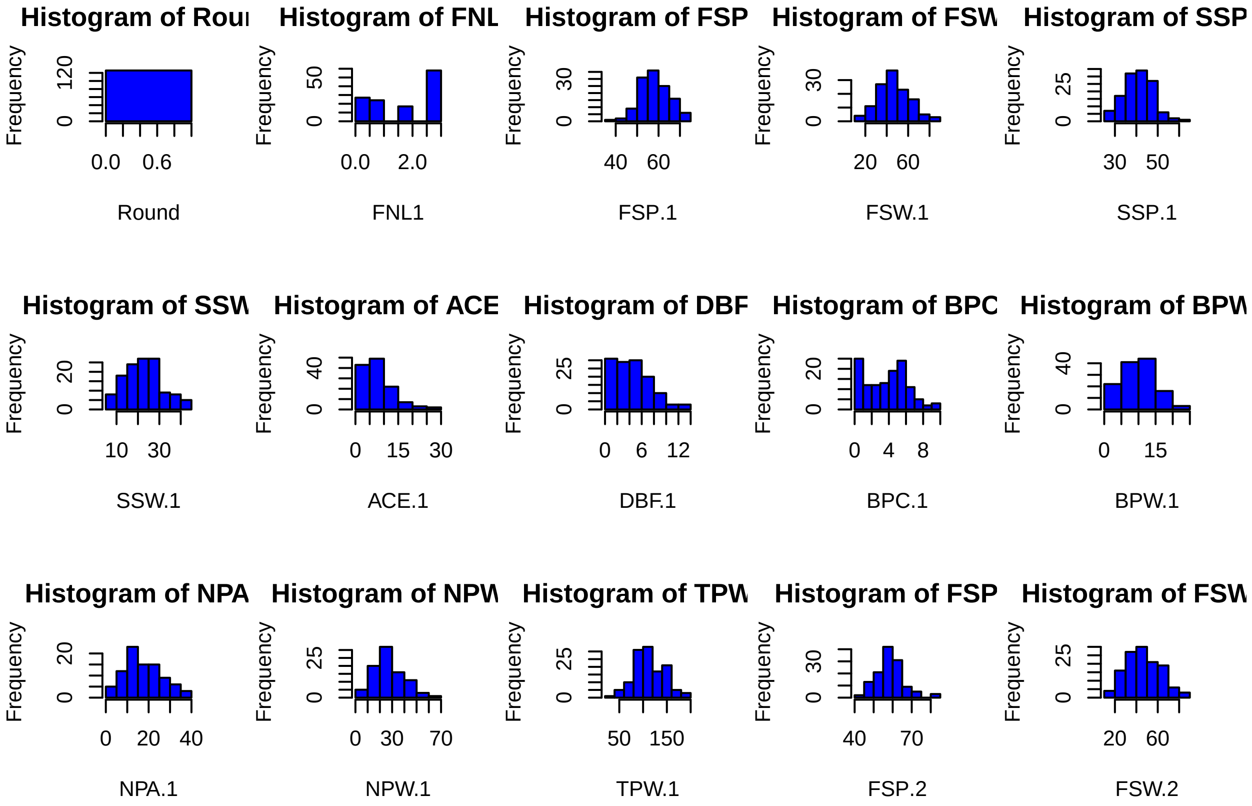
```
summary(usopen)
```

```
##                Player1                  Player2       Round       Result
##  Novak Djokovic    : 6   Rafael Nadal   : 7   Min.   :1   Min.   :0.0000
##  Richard Gasquet   : 6   David Ferrer   : 5   1st Qu.:1   1st Qu.:0.0000
##  Andy Murray       : 5   Mikhail Youzhny: 4   Median :1   Median :0.0000
##  Roger Federer     : 4   Milos Raonic   : 4   Mean   :1   Mean   :0.4683
##  Stanislas Wawrinka: 4   Tomas Berdych  : 4   3rd Qu.:1   3rd Qu.:1.0000
##  Tommy Robredo     : 4   Denis Istomin  : 3   Max.   :1   Max.   :1.0000
##  (Other)           :97   (Other)        :99
##       FNL1            FNL2           FSP.1           FSW.1
##  Min.   :0.000   Min.   :0.000   Min.   :38.00   Min.   :14.00
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:53.00   1st Qu.:38.00
##  Median :2.000   Median :3.000   Median :59.00   Median :47.00
##  Mean   :1.841   Mean   :1.881   Mean   :58.65   Mean   :47.44
##  3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:63.00   3rd Qu.:57.00
##  Max.   :3.000   Max.   :3.000   Max.   :75.00   Max.   :87.00
##
##      SSP.1           SSW.1           ACE.1            DBF.1
##  Min.   :25.00   Min.   : 8.00   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:37.00   1st Qu.:17.00   1st Qu.: 5.000   1st Qu.: 3.000
##  Median :41.00   Median :23.00   Median : 8.000   Median : 5.000
##  Mean   :41.35   Mean   :23.38   Mean   : 8.508   Mean   : 4.952
##  3rd Qu.:47.00   3rd Qu.:29.00   3rd Qu.:11.000   3rd Qu.: 7.000
##  Max.   :62.00   Max.   :45.00   Max.   :29.000   Max.   :14.000
##
##   WNR.1          UFE.1             BPC.1            BPW.1
##  Mode:logical   Mode:logical   Min.   : 0.000   Min.   : 0.00
##  NA's:126       NA's:126       1st Qu.: 2.000   1st Qu.: 7.00
##                                Median : 5.000   Median :10.50
##                                Mean   : 4.198   Mean   :10.26
##                                3rd Qu.: 6.000   3rd Qu.:14.00
##                                Max.   :10.000   Max.   :23.00
##
##      NPA.1           NPW.1           TPW.1           ST1.1
##  Min.   : 4.00   Min.   : 4.00   Min.   : 38.0   Min.   :0.000
##  1st Qu.:12.00   1st Qu.:19.75   1st Qu.: 91.0   1st Qu.:4.000
##  Median :17.00   Median :25.50   Median :111.5   Median :6.000
##  Mean   :18.28   Mean   :27.86   Mean   :112.9   Mean   :4.968
##  3rd Qu.:25.00   3rd Qu.:36.25   3rd Qu.:137.0   3rd Qu.:6.000
##  Max.   :39.00   Max.   :63.00   Max.   :195.0   Max.   :7.000
##  NA's   :38      NA's   :38
##      ST2.1           ST3.1           ST4.1           ST5.1
##  Min.   :1.000   Min.   :0.000   Min.   :0.000   Min.   :0.00
##  1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:4.00
##  Median :6.000   Median :6.000   Median :6.000   Median :4.00
##  Mean   :4.889   Mean   :4.696   Mean   :4.821   Mean   :4.52
##  3rd Qu.:6.000   3rd Qu.:6.000   3rd Qu.:6.000   3rd Qu.:6.00
##  Max.   :7.000   Max.   :7.000   Max.   :7.000   Max.   :7.00
##                  NA's   :1       NA's   :59      NA's   :101
##      FSP.2           FSW.2           SSP.2           SSW.2
##  Min.   :44.00   Min.   :10.00   Min.   :16.00   Min.   : 2.00
##  1st Qu.:55.00   1st Qu.:34.25   1st Qu.:37.00   1st Qu.:16.00
##  Median :58.00   Median :45.50   Median :42.00   Median :23.00
##  Mean   :58.92   Mean   :46.94   Mean   :41.08   Mean   :23.13
##  3rd Qu.:63.00   3rd Qu.:59.50   3rd Qu.:45.00   3rd Qu.:28.00
##  Max.   :84.00   Max.   :90.00   Max.   :56.00   Max.   :48.00
##
##      ACE.2            DBF.2          WNR.2          UFE.2
##  Min.   : 0.000   Min.   : 0.000   Mode:logical   Mode:logical
##  1st Qu.: 4.250   1st Qu.: 3.000   NA's:126       NA's:126
##  Median : 8.000   Median : 4.000
##  Mean   : 9.262   Mean   : 4.595
##  3rd Qu.:11.000   3rd Qu.: 6.000
##  Max.   :39.000   Max.   :15.000
##
##      BPC.2            BPW.2            NPA.2           NPW.2
##  Min.   : 0.000   Min.   : 0.00   Min.   : 4.00   Min.   : 6.00
##  1st Qu.: 2.000   1st Qu.: 7.00   1st Qu.:12.00   1st Qu.:19.00
##  Median : 4.000   Median :10.00   Median :18.50   Median :27.50
##  Mean   : 4.087   Mean   :10.25   Mean   :19.84   Mean   :31.17
##  3rd Qu.: 6.000   3rd Qu.:13.75   3rd Qu.:26.00   3rd Qu.:41.00
##  Max.   :11.000   Max.   :26.00   Max.   :48.00   Max.   :81.00
##                                   NA's   :38      NA's   :38
##      TPW.2            ST1.2            ST2.2           ST3.2
##  Min.   : 45.00   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.: 87.25   1st Qu.:3.250   1st Qu.:3.000   1st Qu.:3.000
##  Median :114.00   Median :6.000   Median :6.000   Median :6.000
##  Mean   :113.18   Mean   :5.016   Mean   :4.516   Mean   :4.616
##  3rd Qu.:137.00   3rd Qu.:6.000   3rd Qu.:6.000   3rd Qu.:6.000
##  Max.   :207.00   Max.   :7.000   Max.   :7.000   Max.   :7.000
##                                                   NA's   :1
##      ST4.2           ST5.2
##  Min.   :0.000   Min.   :1.00
##  1st Qu.:4.000   1st Qu.:5.00
##  Median :6.000   Median :6.00
##  Mean   :5.015   Mean   :5.32
##  3rd Qu.:6.000   3rd Qu.:6.00
##  Max.   :7.000   Max.   :7.00
##  NA's   :59      NA's   :101
```

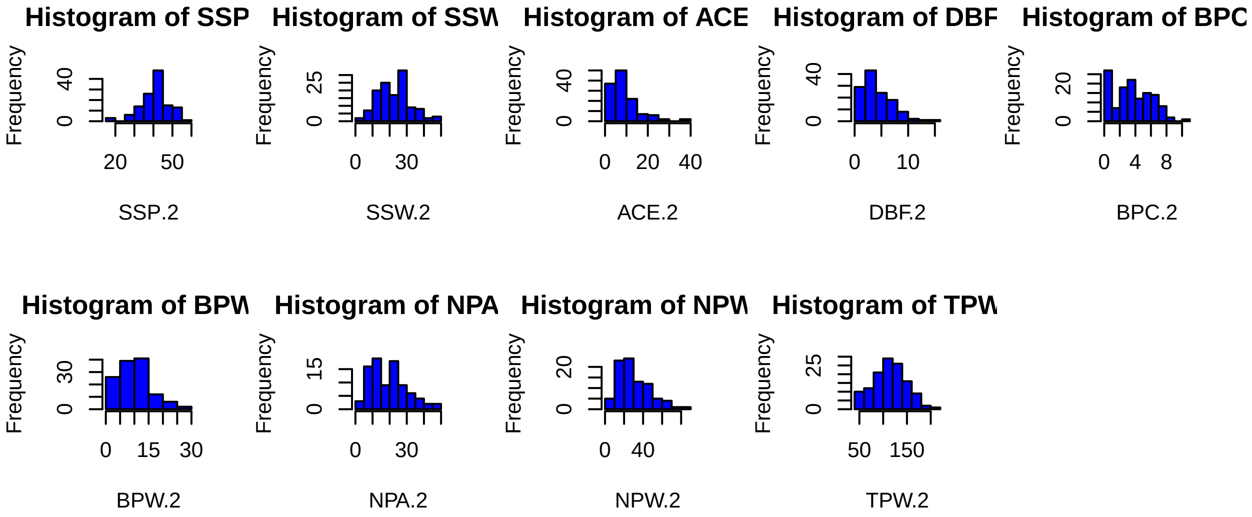
- Summary statistics
  - There are 126 observations with 42 varaibles
  - We read first few observations from the data set
  - WNR.1, WNR.2, UFE.2 and UFE.1 variables have no data
  - There are missing observations for ST4.1, ST5.1, NPA.2, NPW.2, ST3.2, ST4.2 and ST5.2 variables
- We use `hist` function to plot the histograms
- We use `plot` function to plot the density function
- As the varaibles are larger in number we avoid scatter plots at this moment

3.2 Plots

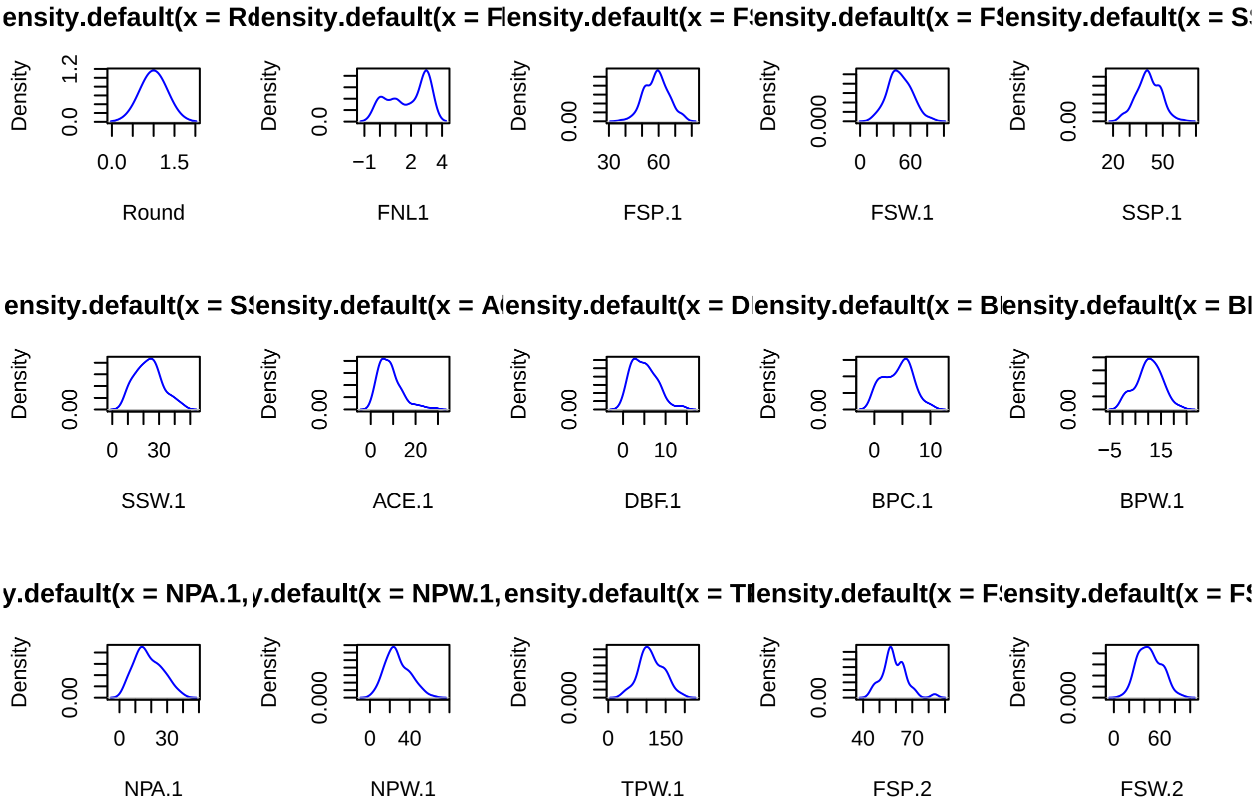
```
#Histograms
par(mfrow=c(3,5))
hist(Round, col="blue",xlab="Round")
hist(FNL1, col="blue",xlab="FNL1")
hist(FSP.1, col="blue",xlab="FSP.1")
hist(FSW.1, col="blue",xlab="FSW.1")
hist(SSP.1, col="blue",xlab="SSP.1")
hist(SSW.1, col="blue",xlab="SSW.1")
hist(ACE.1, col="blue",xlab="ACE.1")
hist(DBF.1, col="blue",xlab="DBF.1")
hist(BPC.1, col="blue",xlab="BPC.1")
hist(BPW.1, col="blue",xlab="BPW.1")
hist(NPA.1, col="blue",xlab="NPA.1")
hist(NPW.1, col="blue",xlab="NPW.1")
hist(TPW.1, col="blue",xlab="TPW.1")
hist(FSP.2, col="blue",xlab="FSP.2")
hist(FSW.2, col="blue",xlab="FSW.2")
```



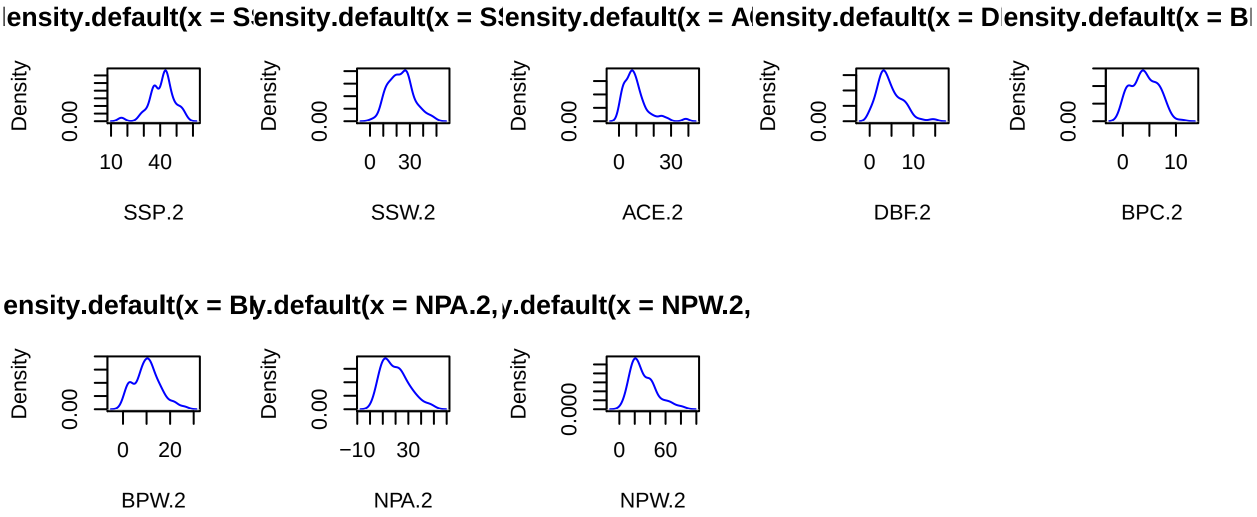
```r
par(mfrow=c(3,5))
hist(SSP.2, col="blue",xlab="SSP.2")
hist(SSW.2, col="blue",xlab="SSW.2")
hist(ACE.2, col="blue",xlab="ACE.2")
hist(DBF.2, col="blue",xlab="DBF.2")
hist(BPC.2, col="blue",xlab="BPC.2")
hist(BPW.2, col="blue",xlab="BPW.2")
hist(NPA.2, col="blue",xlab="NPA.2")
hist(NPW.2, col="blue",xlab="NPW.2")
hist(TPW.2, col="blue",xlab="TPW.2")

#Density Plot
par(mfrow=c(3,5))
```



```
plot(density(Round), col="blue",xlab="Round")
plot(density(FNL1), col="blue",xlab="FNL1")
plot(density(FSP.1), col="blue",xlab="FSP.1")
plot(density(FSW.1), col="blue",xlab="FSW.1")
plot(density(SSP.1), col="blue",xlab="SSP.1")
plot(density(SSW.1), col="blue",xlab="SSW.1")
plot(density(ACE.1), col="blue",xlab="ACE.1")
plot(density(DBF.1), col="blue",xlab="DBF.1")
plot(density(BPC.1), col="blue",xlab="BPC.1")
plot(density(BPW.1), col="blue",xlab="BPW.1")
plot(density(NPA.1,na.rm=T), col="blue",xlab="NPA.1")
plot(density(NPW.1,na.rm=T), col="blue",xlab="NPW.1")
plot(density(TPW.1), col="blue",xlab="TPW.1")
plot(density(FSP.2), col="blue",xlab="FSP.2")
plot(density(FSW.2), col="blue",xlab="FSW.2")
```



```r
par(mfrow=c(3,5))
plot(density(SSP.2), col="blue",xlab="SSP.2")
plot(density(SSW.2), col="blue",xlab="SSW.2")
plot(density(ACE.2), col="blue",xlab="ACE.2")
plot(density(DBF.2), col="blue",xlab="DBF.2")
plot(density(BPC.2), col="blue",xlab="BPC.2")
plot(density(BPW.2), col="blue",xlab="BPW.2")
plot(density(NPA.2,na.rm=T), col="blue",xlab="NPA.2")
plot(density(NPW.2,na.rm=T), col="blue",xlab="NPW.2")
```



- Most of the data is numeric with little or no cleaning required. We can replace the missing values with zero (or mean value) to simplify the data modeling process.

4 Data Preparation

- We will not consider the variable Round as it is a constant and is not impacting the response or regressor variables as evident from the scatter plots
- We have chosen the response variable as FNL1 - Final number of games won by player 1
- Covariates for consideration - FSP.1, FSW.1, SSP.1, SSW.1, ACE.1, DBF.1, BPC.1, BPW.1, NPA.1, NPW.1, TPW.1, FSP.2, FSW.2, SSP.2, SSW.2, ACE.2, DBF.2, BPC.2, BPW.2, NPA.2, NPW.2, TPW.2
- We observe NA values in NPA and NPW variables and replace them with 0

```
usopen$NPA.1[is.na(usopen$NPA.1)] <- 0
usopen$NPW.1[is.na(usopen$NPW.1)] <- 0
usopen$NPW.2[is.na(usopen$NPW.2)] <- 0
usopen$NPA.2[is.na(usopen$NPA.2)] <- 0
```

- Rest of the data looks pretty clean

5 Modeling

- We will utilize multiple linear regression method for this model.
- Based on the correlation matrix and partial f tests we will decide on the final list of covariates and final number of observations

5.1 Correlation martix for selected covariates

```
cor(usopen[, c(  "FSP.1",  "FSW.1",  "SSP.1",  "SSW.1",  "ACE.1",  "DBF.1",  "BPC.1",  "BPW.1",
  "NPA.1",  "NPW.1",  "TPW.1",  "FSP.2",  "FSW.2",  "SSP.2",  "SSW.2",  "ACE.2",  "DBF.2",  "BPC.2",
  "BPW.2",  "NPA.2",  "NPW.2",  "TPW.2"
)])
```

- We observe high correlation between NPA.1 and NPW.1, TPW.1 and FSW.1,NPA.2 and NPW.2, FSW.2 and TPW.2
- Other combinations of variables are also correlated
- There is a good probability that we may experience multicollinearity in our model
- We split the data set into testing (30%) and training data (70%)

5.2 Data split

```
# setting the seed to make the partition reproductible
set.seed(999)
index <-
  sample(seq_len(nrow(usopen)), size = floor(0.70 * nrow(usopen)))

usopen_train <- usopen[index,]
usopen_test <- usopen[-index, ]
```

- We now create model based on training data
- `summary` function is used to get the model details

5.3 Model details

```
model_usopen <-
  lm(
    FNL1 ~ FSP.1 + FSW.1 + SSP.1 + SSW.1 + ACE.1 + DBF.1 + BPC.1 + BPW.1 + NPA.1 +
      NPW.1 + TPW.1 + FSP.2 + FSW.2 + SSP.2 + SSW.2 + ACE.2 + DBF.2 + BPC.2 +
      BPW.2 + NPA.2 + NPW.2 + TPW.2,
    data = usopen_train
  )
summary(model_usopen)
```

```
##
## Call:
## lm(formula = FNL1 ~ FSP.1 + FSW.1 + SSP.1 + SSW.1 + ACE.1 + DBF.1 +
##     BPC.1 + BPW.1 + NPA.1 + NPW.1 + TPW.1 + FSP.2 + FSW.2 + SSP.2 +
##     SSW.2 + ACE.2 + DBF.2 + BPC.2 + BPW.2 + NPA.2 + NPW.2 + TPW.2,
##     data = usopen_train)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97603 -0.30841  0.03371  0.33374  1.05906
##
## Coefficients: (2 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.262626   1.376240  -0.917  0.36220
## FSP.1        0.002551   0.015872   0.161  0.87280
## FSW.1       -0.002686   0.018716  -0.144  0.88630
## SSP.1              NA         NA      NA       NA
## SSW.1        0.007278   0.025829   0.282  0.77899
## ACE.1        0.005061   0.013210   0.383  0.70284
## DBF.1        0.024640   0.025376   0.971  0.33504
## BPC.1        0.116067   0.063603   1.825  0.07248 .
## BPW.1       -0.049449   0.026547  -1.863  0.06689 .
## NPA.1       -0.019090   0.030332  -0.629  0.53125
## NPW.1        0.007268   0.020210   0.360  0.72028
## TPW.1        0.061255   0.014343   4.271 6.29e-05 ***
## FSP.2        0.024471   0.015574   1.571  0.12083
## FSW.2       -0.009003   0.020432  -0.441  0.66090
## SSP.2              NA         NA      NA       NA
## SSW.2        0.027966   0.023308   1.200  0.23443
## ACE.2        0.021018   0.010839   1.939  0.05671 .
## DBF.2       -0.004163   0.029013  -0.143  0.88633
## BPC.2        0.086542   0.073086   1.184  0.24055
## BPW.2       -0.025209   0.026448  -0.953  0.34393
## NPA.2        0.041423   0.025749   1.609  0.11238
## NPW.2       -0.023605   0.017060  -1.384  0.17105
## TPW.2       -0.053608   0.016279  -3.293  0.00158 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4957 on 67 degrees of freedom
## Multiple R-squared:  0.8815, Adjusted R-squared:  0.8462
## F-statistic: 24.93 on 20 and 67 DF,  p-value: < 2.2e-16
```

- The null hypothesis is $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$
- That is, there is not a single predictor which can be considered statistically significant
- The alternate hypothesis is H_a : At least one β_j is not zero
- That is, there is at least predictor which can explain the change in resultant variable
- We reject null hypothesis when the p value is < 0.05
- From the model summary we observe that there are only 2 statisticallly significant variables (null hypothesis is rejected for these)
  - TPW.1
  - TPW.2
- The F-statistic is 24.93 and the corresponding p-value is significantly lower than 0.05 so we can conclude to reject that null hypothesis that no predictor varaible explains the variability in the response variable
- The R^2 value is 0.8815
- The model explains 88.15% of the varaibility in FNL1

5.4 Partial F tests

- We now do a partial f-test for the variables FSP.1, FSW.1, SSP.1, SSW.1, ACE.1, DBF.1, NPA.1, NPW.1, FSP.2, FSW.2, SSP.2, SSW.2, ACE.2, DBF.2, NPA.2, NPW.2 and BPC.2

```r
model_usopen_p <-
  lm(FNL1 ~  BPC.1 + BPW.1 + TPW.1 + BPW.2 + TPW.2,
     data = usopen_train)
anova(model_usopen, model_usopen_p)
```

```
## Analysis of Variance Table
##
## Model 1: FNL1 ~ FSP.1 + FSW.1 + SSP.1 + SSW.1 + ACE.1 + DBF.1 + BPC.1 +
##     BPW.1 + NPA.1 + NPW.1 + TPW.1 + FSP.2 + FSW.2 + SSP.2 + SSW.2 +
##     ACE.2 + DBF.2 + BPC.2 + BPW.2 + NPA.2 + NPW.2 + TPW.2
## Model 2: FNL1 ~ BPC.1 + BPW.1 + TPW.1 + BPW.2 + TPW.2
##   Res.Df    RSS  Df Sum of Sq     F Pr(>F)
## 1     67 16.465
## 2     82 20.008 -15   -3.5425 0.961 0.5046
```

- We observe that p value is > 0.05 and therefore all these variables are not statistically significant
- We can now exclude these variables from our analysis
- We summarize our current model

```r
summary(model_usopen_p)
```

```
##
## Call:
## lm(formula = FNL1 ~ BPC.1 + BPW.1 + TPW.1 + BPW.2 + TPW.2, data = usopen_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94966 -0.29081  0.04009  0.31831  1.20486
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.409225   0.201126   2.035  0.04512 *
## BPC.1        0.163649   0.043440   3.767  0.00031 ***
## BPW.1       -0.055868   0.017387  -3.213  0.00188 **
## TPW.1        0.048078   0.005446   8.828 1.58e-13 ***
## BPW.2       -0.024529   0.015967  -1.536  0.12834
## TPW.2       -0.034274   0.005362  -6.392 9.40e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.494 on 82 degrees of freedom
## Multiple R-squared:  0.856,  Adjusted R-squared:  0.8473
## F-statistic: 97.53 on 5 and 82 DF,  p-value: < 2.2e-16
```

- We run another partial f-test for the variable BPW.2

```r
model_usopen_p_2 <-
  lm(FNL1 ~ BPC.1 +  BPW.1 + TPW.1 +  TPW.2,
     data = usopen_train)
anova(model_usopen_p, model_usopen_p_2)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: FNL1 ~ BPC.1 + BPW.1 + TPW.1 + BPW.2 + TPW.2
## Model 2: FNL1 ~ BPC.1 + BPW.1 + TPW.1 + TPW.2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     82 20.008
## 2     83 20.584 -1  -0.57582 2.3599 0.1283
```

- We again get a p-value > 0.05
- Hence BPW.2 is also not statistically significant
- We summarize our current model

```
summary(model_usopen_p_2)
```

```
##
## Call:
## lm(formula = FNL1 ~ BPC.1 + BPW.1 + TPW.1 + TPW.2, data = usopen_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04081 -0.40456  0.08174  0.36555  1.07170
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.397609   0.202623   1.962 0.053078 .
## BPC.1        0.132087   0.038586   3.423 0.000964 ***
## BPW.1       -0.053149   0.017437  -3.048 0.003090 **
## TPW.1        0.053237   0.004322  12.317  < 2e-16 ***
## TPW.2       -0.040555   0.003496 -11.599  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.498 on 83 degrees of freedom
## Multiple R-squared:  0.8519, Adjusted R-squared:  0.8448
## F-statistic: 119.4 on 4 and 83 DF,  p-value: < 2.2e-16
```

5.5 Standardize the covariate coefficients

- We now standardize the regression coefficients using unit normal scaling

```
usopen_train_standard = as.data.frame(apply(usopen_train[, c("FNL1", "BPC.1", "BPW.1", "TPW.1", "TPW.2"
  (x - mean(x)) / sd(x)
}))
```

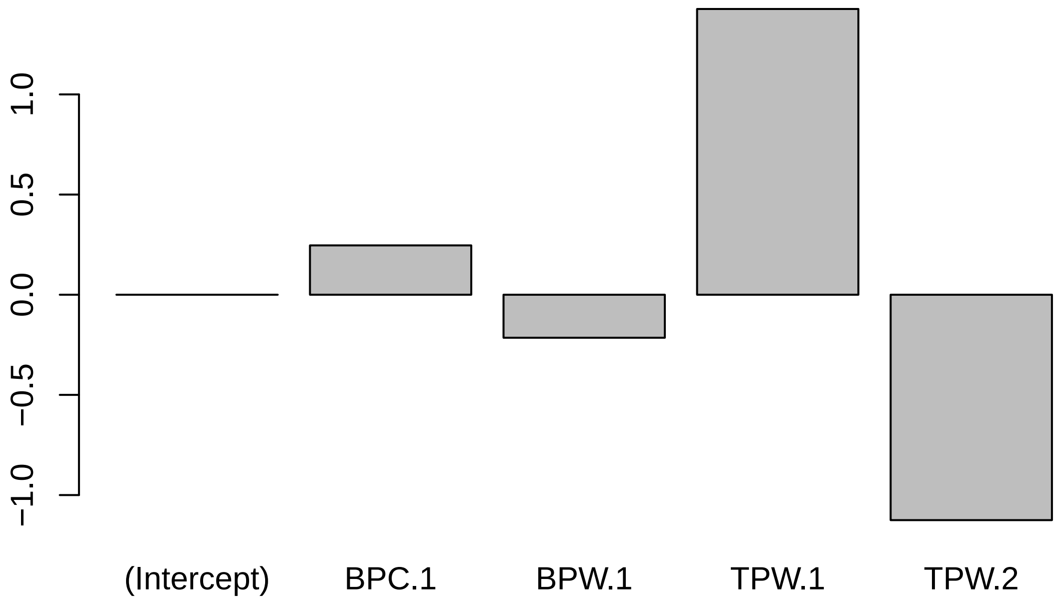
- We now create the new model using standardized values

```
model_usopen_std <-
  lm(FNL1 ~ BPC.1 +  BPW.1 + TPW.1 +  TPW.2,
     data = usopen_train_standard)
summary(model_usopen_std)
```

```
##
## Call:
## lm(formula = FNL1 ~ BPC.1 + BPW.1 + TPW.1 + TPW.2, data = usopen_train_standard)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82346 -0.32007  0.06467  0.28921  0.84790
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.475e-16  4.200e-02   0.000 1.000000
## BPC.1        2.462e-01  7.192e-02   3.423 0.000964 ***
## BPW.1       -2.148e-01  7.047e-02  -3.048 0.003090 **
## TPW.1        1.426e+00  1.158e-01  12.317  < 2e-16 ***
## TPW.2       -1.125e+00  9.700e-02 -11.599  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.394 on 83 degrees of freedom
## Multiple R-squared:  0.8519, Adjusted R-squared:  0.8448
## F-statistic: 119.4 on 4 and 83 DF,  p-value: < 2.2e-16
```

- To better visualize the coefficients we plot the `barplot`

```
barplot(model_usopen_std$coefficients)
```



- We observe that the most statistically significant variables are TPW.1 and TPW.2

5.6 Multicollinearity checks

- We check our model for multicolllinearity
- We will the `vif` function from `car` library to examine multicollinearity

```r
library(car)
vif(model_usopen_std)
```

```
##    BPC.1    BPW.1    TPW.1    TPW.2
## 2.899012 2.782959 7.514762 5.273466
```

- We observe that there is multicollinearity in our model (as was expected)
- We find the colleration matrix of the variables

```r
cor(usopen_train_standard[, c("FNL1", "BPC.1", "BPW.1", "TPW.1", "TPW.2")])
```

```
##              FNL1       BPC.1     BPW.1     TPW.1       TPW.2
## FNL1   1.00000000  0.75807021 0.5302779 0.4943759 -0.06581077
## BPC.1  0.75807021  1.00000000 0.6804967 0.4540390 -0.00928395
## BPW.1  0.53027788  0.68049669 1.0000000 0.6824489  0.35180657
## TPW.1  0.49437585  0.45403905 0.6824489 1.0000000  0.79731644
## TPW.2 -0.06581077 -0.00928395 0.3518066 0.7973164  1.00000000
```

- Since TPW.1 and TPW.2 are highly correlated, we keep only TPW.1 and recreate the model

```r
model_usopen_std_2 <-
  lm(FNL1 ~ BPC.1 +  BPW.1 + TPW.1 ,
     data = usopen_train_standard)
summary(model_usopen_std_2)
```

```
## 
## Call:
## lm(formula = FNL1 ~ BPC.1 + BPW.1 + TPW.1, data = usopen_train_standard)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.32305 -0.46140 -0.00508  0.45556  1.20729 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  7.619e-17  6.759e-02   0.000  1.00000    
## BPC.1        7.449e-01  9.279e-02   8.028 5.34e-12 ***
## BPW.1       -1.557e-01  1.131e-01  -1.377  0.17229    
## TPW.1        2.624e-01  9.302e-02   2.821  0.00597 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.6341 on 84 degrees of freedom
## Multiple R-squared:  0.6118, Adjusted R-squared:  0.598 
## F-statistic: 44.14 on 3 and 84 DF,  p-value: < 2.2e-16
```

- We now calculate the VIF to check for multicollinearity

```r
library(car)
vif(model_usopen_std_2)
```

```
##    BPC.1    BPW.1    TPW.1 
## 1.863158 2.768419 1.872437
```

- There is no multicolinearity on this model

- As there is one statistically insignificant variable in this model, we run partial F test again for the variable BPW.1

```
model_usopen_std_2p <-
  lm(FNL1 ~ BPC.1 +   TPW.1 ,
     data = usopen_train_standard)
summary(model_usopen_std_2p)
```

```
##
## Call:
## lm(formula = FNL1 ~ BPC.1 + TPW.1, data = usopen_train_standard)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41846 -0.44510  0.05935  0.48807  1.28546
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.064e-16  6.795e-02   0.000   1.0000
## BPC.1       6.722e-01  7.670e-02   8.764 1.62e-13 ***
## TPW.1       1.892e-01  7.670e-02   2.467   0.0156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6374 on 85 degrees of freedom
## Multiple R-squared:  0.6031, Adjusted R-squared:  0.5937
## F-statistic: 64.58 on 2 and 85 DF,  p-value: < 2.2e-16
```

```
library(car)
vif(model_usopen_std_2p)
```

```
##    BPC.1    TPW.1
## 1.259686 1.259686
```

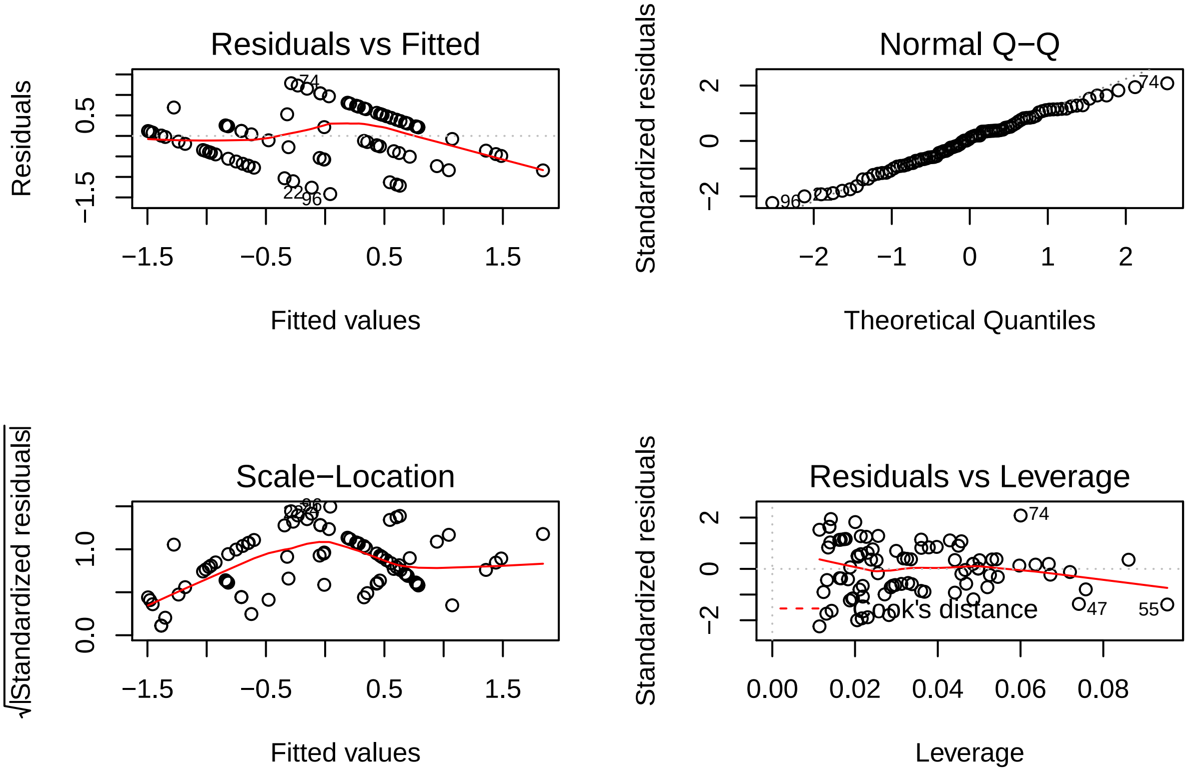
- This model does not have multicollinearity

```
anova(model_usopen_std_2, model_usopen_std_2p)
```

```
## Analysis of Variance Table
##
## Model 1: FNL1 ~ BPC.1 + BPW.1 + TPW.1
## Model 2: FNL1 ~ BPC.1 + TPW.1
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     84 33.770
## 2     85 34.532 -1  -0.76185 1.895 0.1723
```

- Using `anova` function we arrive to the conclusion that the vairable BPW.1 is not statistically significant
- We now do the residual plots

```
par(mfrow = c(2, 2))
plot(model_usopen_std_2p)
```



6 Evaluation

- We calculate the mean square error for the models created

```
t <- predict(model_usopen_std, usopen_test)
mean((t - usopen_test[, c("FNL1")]) ** 2)
```

```
## [1] 1852.487
```

```
t <- predict(model_usopen_std_2, usopen_test)
mean((t - usopen_test[, c("FNL1")]) ** 2)
```

```
## [1] 969.4526
```

```
t <- predict(model_usopen_std_2p, usopen_test)
mean((t - usopen_test[, c("FNL1")]) ** 2)
```

```
## [1] 564.1643
```

- We observe that as we proceeded to refine our model the MSE kept on decreasing
- And we have the least MSE in our final model