Appendix

Shivam K C

November 2, 2018

library(leaps)  
library(car)

## Loading required package: carData

library(mosaic); library(readr)

## Loading required package: dplyr

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':  
##   
## recode

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

## Loading required package: lattice

## Loading required package: ggformula

## Loading required package: ggplot2

## Loading required package: ggstance

##   
## Attaching package: 'ggstance'

## The following objects are masked from 'package:ggplot2':  
##   
## geom\_errorbarh, GeomErrorbarh

##   
## New to ggformula? Try the tutorials:   
## learnr::run\_tutorial("introduction", package = "ggformula")  
## learnr::run\_tutorial("refining", package = "ggformula")

## Loading required package: mosaicData

## Loading required package: Matrix

##   
## The 'mosaic' package masks several functions from core packages in order to add   
## additional features. The original behavior of these functions should not be affected by this.  
##   
## Note: If you use the Matrix package, be sure to load it BEFORE loading mosaic.

##   
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':  
##   
## mean

## The following object is masked from 'package:ggplot2':  
##   
## stat

## The following objects are masked from 'package:dplyr':  
##   
## count, do, tally

## The following objects are masked from 'package:car':  
##   
## deltaMethod, logit

## The following objects are masked from 'package:stats':  
##   
## binom.test, cor, cor.test, cov, fivenum, IQR, median,  
## prop.test, quantile, sd, t.test, var

## The following objects are masked from 'package:base':  
##   
## max, mean, min, prod, range, sample, sum

fpl <- read.csv("C:/Users/Shivam/Desktop/cleaned\_players.csv")

# Calculation of Standard Deviations Medain and Means of the variables

summary(fpl) #summary of the data

## first\_name second\_name goals\_scored assists   
## James : 10 Murphy : 3 Min. : 0.000 Min. : 0.000   
## Ben : 7 Smith : 3 1st Qu.: 0.000 1st Qu.: 0.000   
## Jack : 7 Sterling: 3 Median : 0.000 Median : 0.000   
## Kyle : 7 Taylor : 3 Mean : 1.527 Mean : 1.402   
## Sam : 7 Ward : 3 3rd Qu.: 2.000 3rd Qu.: 2.000   
## Aaron : 6 Ayew : 2 Max. :32.000 Max. :18.000   
## (Other):603 (Other) :630   
## total\_points minutes goals\_conceded creativity   
## Min. : 0.00 Min. : 0.0 Min. : 0.00 Min. : 0.0   
## 1st Qu.: 2.00 1st Qu.: 35.5 1st Qu.: 1.00 1st Qu.: 0.0   
## Median : 36.00 Median : 987.0 Median :14.00 Median : 64.8   
## Mean : 48.59 Mean :1158.4 Mean :17.31 Mean : 173.3   
## 3rd Qu.: 82.50 3rd Qu.:2039.0 3rd Qu.:30.00 3rd Qu.: 265.8   
## Max. :303.00 Max. :3420.0 Max. :63.00 Max. :1744.2   
##   
## influence threat bonus bps   
## Min. : 0.0 Min. : 0.0 Min. : 0.000 Min. : -1.0   
## 1st Qu.: 6.7 1st Qu.: 0.0 1st Qu.: 0.000 1st Qu.: 8.0   
## Median : 204.8 Median : 63.0 Median : 1.000 Median :171.0   
## Mean : 267.7 Mean : 184.6 Mean : 3.737 Mean :222.3   
## 3rd Qu.: 455.7 3rd Qu.: 239.0 3rd Qu.: 6.000 3rd Qu.:382.0   
## Max. :1496.2 Max. :2355.0 Max. :31.000 Max. :913.0   
##   
## ict\_index clean\_sheets red\_cards yellow\_cards   
## Min. : 0.00 Min. : 0.000 Min. :0.00000 Min. : 0.000   
## 1st Qu.: 1.85 1st Qu.: 0.000 1st Qu.:0.00000 1st Qu.: 0.000   
## Median : 39.30 Median : 3.000 Median :0.00000 Median : 1.000   
## Mean : 62.54 Mean : 3.949 Mean :0.06028 Mean : 1.796   
## 3rd Qu.:100.20 3rd Qu.: 7.000 3rd Qu.:0.00000 3rd Qu.: 3.000   
## Max. :454.40 Max. :19.000 Max. :2.00000 Max. :11.000   
##   
## selected\_by\_percent now\_cost   
## Min. : 0.000 Min. : 38.00   
## 1st Qu.: 0.100 1st Qu.: 44.00   
## Median : 0.400 Median : 47.00   
## Mean : 2.317 Mean : 51.26   
## 3rd Qu.: 2.200 3rd Qu.: 54.00   
## Max. :56.700 Max. :131.00   
##

sd(fpl$influence, data = fpl)

## [1] 280.839

sd(fpl$threat, data = fpl)

## [1] 292.5225

sd(fpl$bonus, data = fpl)

## [1] 5.535046

sd(fpl$selected\_by\_percent, data = fpl)

## [1] 5.07272

sd(fpl$total\_points, data = fpl)

## [1] 49.77803

sd(fpl$minutes, data = fpl)

## [1] 1089.496

sd(fpl$bps, data = fpl)

## [1] 222.9687

sd(fpl$ict\_index, data = fpl)

## [1] 72.35621

# Correlations Calculations

data <- data.frame (fpl$influence, fpl$threat, fpl$now\_cost, fpl$selected\_by\_percent, fpl$ict\_index , fpl$bonus, fpl$bps, fpl$creativity, fpl$minutes,fpl$total\_points)  
round(cor(data), 4)

## fpl.influence fpl.threat fpl.now\_cost  
## fpl.influence 1.0000 0.6260 0.4546  
## fpl.threat 0.6260 1.0000 0.7358  
## fpl.now\_cost 0.4546 0.7358 1.0000  
## fpl.selected\_by\_percent 0.6473 0.5987 0.5271  
## fpl.ict\_index 0.8642 0.8905 0.6722  
## fpl.bonus 0.8102 0.6952 0.5850  
## fpl.bps 0.9728 0.5550 0.4216  
## fpl.creativity 0.6437 0.7060 0.5749  
## fpl.minutes 0.9381 0.5400 0.3234  
## fpl.total\_points 0.9393 0.7442 0.5718  
## fpl.selected\_by\_percent fpl.ict\_index fpl.bonus  
## fpl.influence 0.6473 0.8642 0.8102  
## fpl.threat 0.5987 0.8905 0.6952  
## fpl.now\_cost 0.5271 0.6722 0.5850  
## fpl.selected\_by\_percent 1.0000 0.6528 0.7128  
## fpl.ict\_index 0.6528 1.0000 0.8173  
## fpl.bonus 0.7128 0.8173 1.0000  
## fpl.bps 0.6278 0.8267 0.7988  
## fpl.creativity 0.4607 0.8811 0.6414  
## fpl.minutes 0.5066 0.7980 0.6930  
## fpl.total\_points 0.7070 0.9192 0.8722  
## fpl.bps fpl.creativity fpl.minutes  
## fpl.influence 0.9728 0.6437 0.9381  
## fpl.threat 0.5550 0.7060 0.5400  
## fpl.now\_cost 0.4216 0.5749 0.3234  
## fpl.selected\_by\_percent 0.6278 0.4607 0.5066  
## fpl.ict\_index 0.8267 0.8811 0.7980  
## fpl.bonus 0.7988 0.6414 0.6930  
## fpl.bps 1.0000 0.6485 0.9551  
## fpl.creativity 0.6485 1.0000 0.6224  
## fpl.minutes 0.9551 0.6224 1.0000  
## fpl.total\_points 0.9468 0.7336 0.9023  
## fpl.total\_points  
## fpl.influence 0.9393  
## fpl.threat 0.7442  
## fpl.now\_cost 0.5718  
## fpl.selected\_by\_percent 0.7070  
## fpl.ict\_index 0.9192  
## fpl.bonus 0.8722  
## fpl.bps 0.9468  
## fpl.creativity 0.7336  
## fpl.minutes 0.9023  
## fpl.total\_points 1.0000

#highest to lowest correlation totalpoints has with = bps (0.95), influence (0.93),ict\_index (0.92), minutes(0.90), bonus (0.87), threat (0.74), creativity (0.73), selected by (0.70), now\_cost(0.57),

# Correlations for potential transformations

cor(fpl$total\_points,log(fpl$selected\_by\_percent+1))

## [1] 0.8257677

cor(log(fpl$total\_points+1),log(fpl$selected\_by\_percent+1))

## [1] 0.6245445

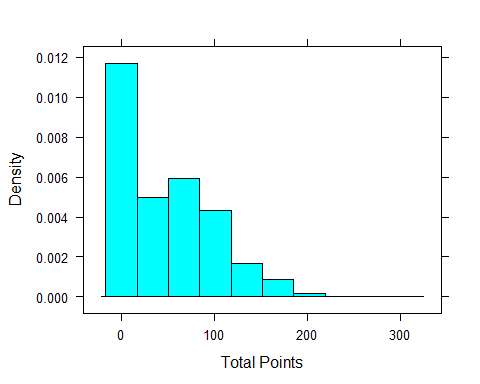
cor(log(fpl$total\_points+1),fpl$selected\_by\_percent)

## [1] 0.4259195

# Distributions of some variables

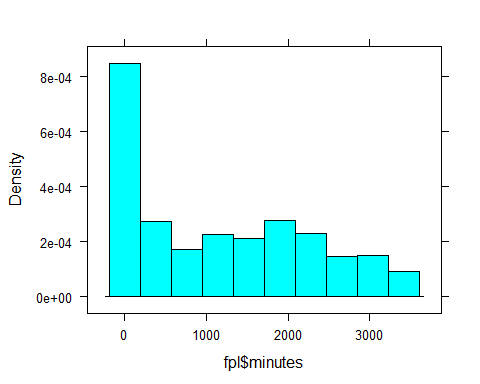
histogram(fpl$total\_points, xlab = 'Total Points', data=fpl) #so many players have zero points. the second most densed area in the histogram is between 50 and 90 pts.

## Warning in histogram.numeric(fpl$total\_points, xlab = "Total Points", data  
## = fpl): explicit 'data' specification ignored



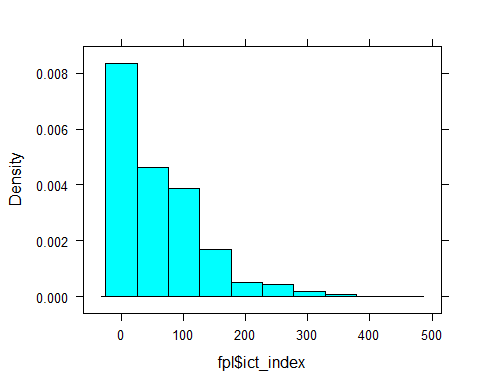
histogram(fpl$minutes, data=fpl) #so many players have zero minutes which answers why there are so many players with zero points. Shows how competitive is the premier league. There are more number of players without game time than there are players with game time. Even after one is part of the squad, it is still unsure if he gets any game time in the season.

## Warning in histogram.numeric(fpl$minutes, data = fpl): explicit 'data'  
## specification ignored



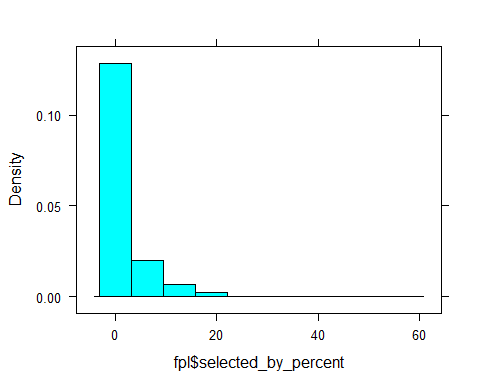
histogram(fpl$ict\_index, data=fpl)

## Warning in histogram.numeric(fpl$ict\_index, data = fpl): explicit 'data'  
## specification ignored



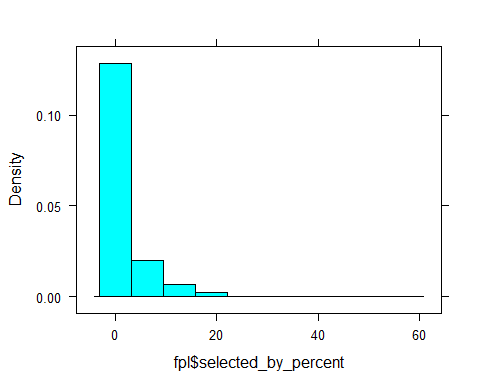
histogram(fpl$selected\_by\_percent, data=fpl) #it is very rare to see high percentage of fantasy players agreeing on selecting a particular soccer player for their team.

## Warning in histogram.numeric(fpl$selected\_by\_percent, data = fpl): explicit  
## 'data' specification ignored



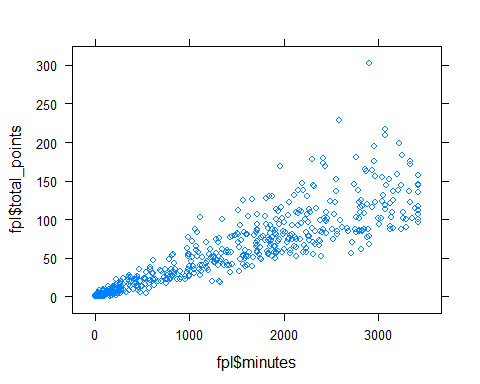
histogram(fpl$selected\_by\_percent, data=fpl)

## Warning in histogram.numeric(fpl$selected\_by\_percent, data = fpl): explicit  
## 'data' specification ignored

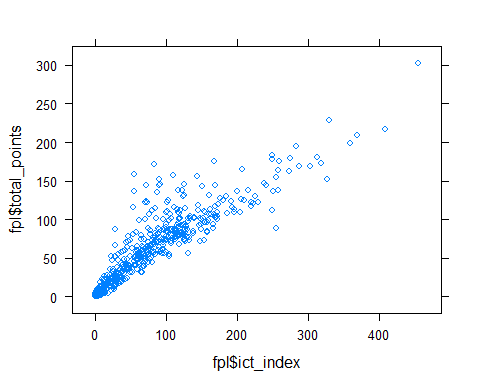


# Scatterplots of the predictors with Total Points

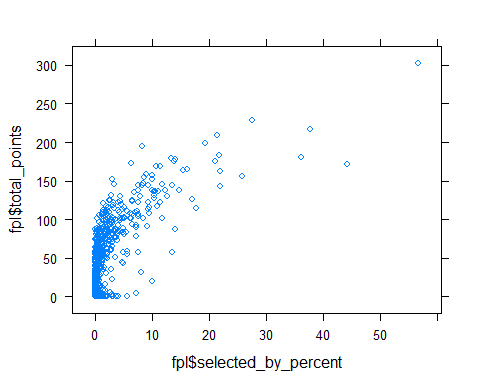
xyplot(fpl$total\_points~fpl$minutes, data=fpl) #linear but not that strong



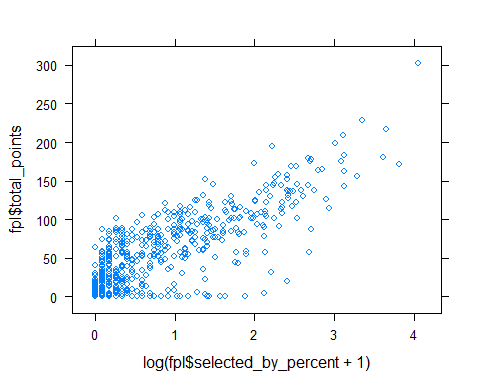
xyplot(fpl$total\_points~fpl$ict\_index, data=fpl) #linear and strong



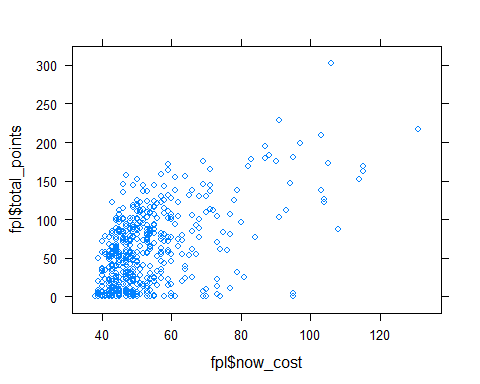
xyplot(fpl$total\_points~fpl$selected\_by\_percent, data=fpl) #not linear at all.might need log transformation. shows selected\_by\_percent can be deceiving. High selected\_by\_percent does not mean high total points. For instance, a soccer player who is selected by only 4 percent has more total points than a soccer player who is selected by 10 percent.



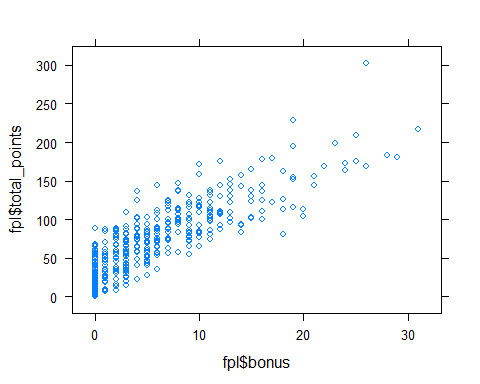
xyplot(fpl$total\_points~log(fpl$selected\_by\_percent+1), data=fpl)



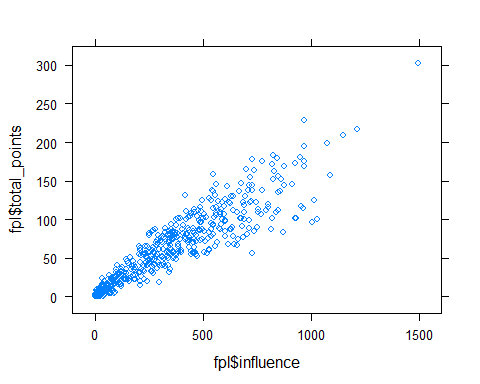
xyplot(fpl$total\_points~fpl$now\_cost, data=fpl) #not linear. maybe a transformation.



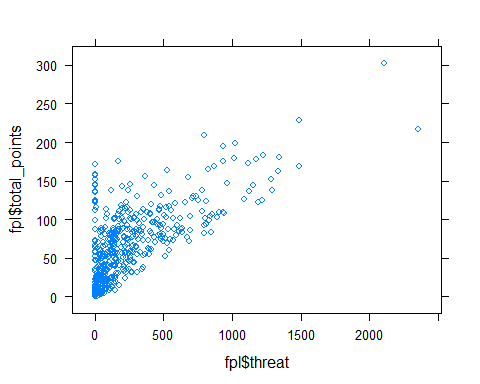
xyplot(fpl$total\_points~fpl$bonus, data=fpl) #kind of linear



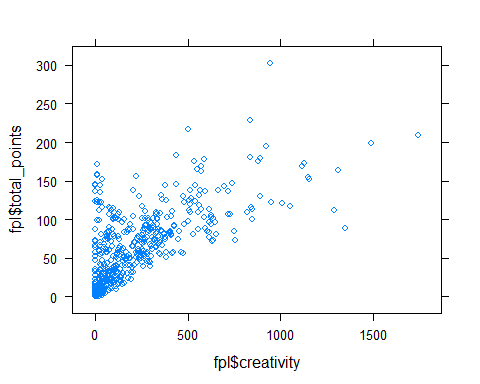
xyplot(fpl$total\_points~fpl$influence, data=fpl)# strongly linear



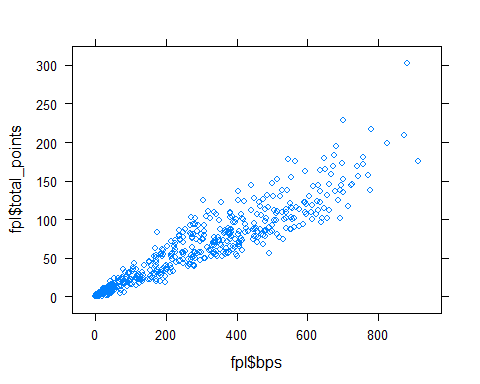
xyplot(fpl$total\_points~fpl$threat, data=fpl)# not that linear



xyplot(fpl$total\_points~fpl$creativity, data=fpl)#not that linear



xyplot(fpl$total\_points~fpl$bps, data=fpl)# strongly linear



# Experimenting with different models

ictinfluence.log <- lm(fpl$total\_points~fpl$ict\_index + fpl$influence ,data=fpl)  
summary(ictinfluence.log)

##   
## Call:  
## lm(formula = fpl$total\_points ~ fpl$ict\_index + fpl$influence,   
## data = fpl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -61.018 -3.807 -2.391 4.526 83.165   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.166226 0.728614 4.346 1.61e-05 \*\*\*  
## fpl$ict\_index 0.291877 0.014471 20.170 < 2e-16 \*\*\*  
## fpl$influence 0.101502 0.003728 27.225 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 13.39 on 644 degrees of freedom  
## Multiple R-squared: 0.9279, Adjusted R-squared: 0.9277   
## F-statistic: 4144 on 2 and 644 DF, p-value: < 2.2e-16

bonusbps.log <- lm(fpl$total\_points~fpl$bonus + fpl$bps ,data=fpl)  
summary(bonusbps.log)

##   
## Call:  
## lm(formula = fpl$total\_points ~ fpl$bonus + fpl$bps, data = fpl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -51.476 -3.850 -2.990 3.748 88.691   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.527503 0.721350 4.89 1.27e-06 \*\*\*  
## fpl$bonus 2.879665 0.151903 18.96 < 2e-16 \*\*\*  
## fpl$bps 0.154268 0.003771 40.91 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 12.86 on 644 degrees of freedom  
## Multiple R-squared: 0.9335, Adjusted R-squared: 0.9333   
## F-statistic: 4520 on 2 and 644 DF, p-value: < 2.2e-16

current.log <- lm(fpl$total\_points~fpl$bonus + fpl$bps + fpl$ict\_index + fpl$influence ,data=fpl)  
summary(current.log)

##   
## Call:  
## lm(formula = fpl$total\_points ~ fpl$bonus + fpl$bps + fpl$ict\_index +   
## fpl$influence, data = fpl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -48.115 -2.344 -1.557 3.287 58.703   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.713915 0.530412 3.231 0.0013 \*\*   
## fpl$bonus 1.547841 0.124699 12.413 < 2e-16 \*\*\*  
## fpl$bps 0.151769 0.007235 20.977 < 2e-16 \*\*\*  
## fpl$ict\_index 0.262899 0.011161 23.555 < 2e-16 \*\*\*  
## fpl$influence -0.033974 0.006365 -5.337 1.31e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.347 on 642 degrees of freedom  
## Multiple R-squared: 0.965, Adjusted R-squared: 0.9647   
## F-statistic: 4420 on 4 and 642 DF, p-value: < 2.2e-16

current1.log <- lm(fpl$total\_points~ fpl$bps + fpl$influence ,data=fpl)  
summary(current1.log)

##   
## Call:  
## lm(formula = fpl$total\_points ~ fpl$bps + fpl$influence, data = fpl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -57.239 -5.863 -1.880 5.496 89.685   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.87981 0.86513 2.173 0.0302 \*   
## fpl$bps 0.13723 0.01184 11.591 < 2e-16 \*\*\*  
## fpl$influence 0.06051 0.00940 6.438 2.37e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 15.56 on 644 degrees of freedom  
## Multiple R-squared: 0.9026, Adjusted R-squared: 0.9023   
## F-statistic: 2986 on 2 and 644 DF, p-value: < 2.2e-16

current2.log <- lm(fpl$total\_points~fpl$bonus + fpl$ict\_index ,data=fpl)  
summary(current2.log)

##   
## Call:  
## lm(formula = fpl$total\_points ~ fpl$bonus + fpl$ict\_index, data = fpl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -68.343 -9.610 -4.912 7.381 93.977   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.60981 0.86436 11.12 <2e-16 \*\*\*  
## fpl$bonus 3.27514 0.20496 15.98 <2e-16 \*\*\*  
## fpl$ict\_index 0.42760 0.01568 27.27 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 16.61 on 644 degrees of freedom  
## Multiple R-squared: 0.8889, Adjusted R-squared: 0.8886   
## F-statistic: 2577 on 2 and 644 DF, p-value: < 2.2e-16

current3.log <- lm(fpl$total\_points~ fpl$bps ,data=fpl)  
summary(current3.log)

##   
## Call:  
## lm(formula = fpl$total\_points ~ fpl$bps, data = fpl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -49.375 -6.299 -1.593 3.747 115.191   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.59335 0.89066 1.789 0.0741 .   
## fpl$bps 0.21137 0.00283 74.699 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 16.04 on 645 degrees of freedom  
## Multiple R-squared: 0.8964, Adjusted R-squared: 0.8962   
## F-statistic: 5580 on 1 and 645 DF, p-value: < 2.2e-16

current8.log <- lm(fpl$total\_points~ fpl$bps + fpl$threat + fpl$bonus ,data=fpl)  
summary(current8.log)

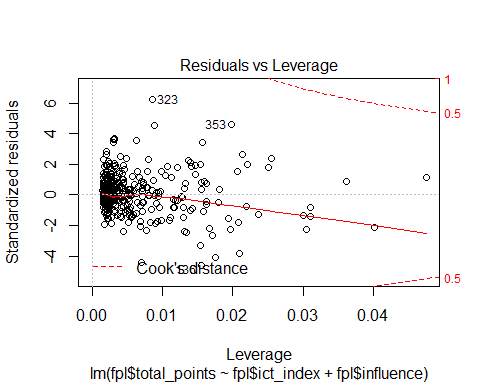
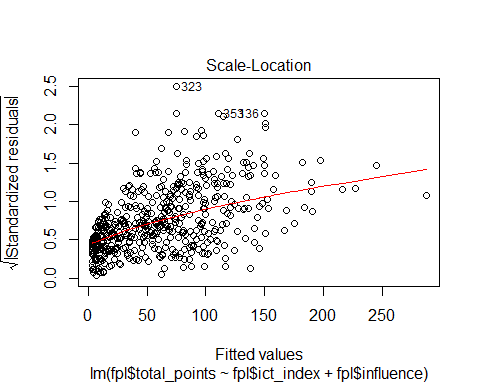
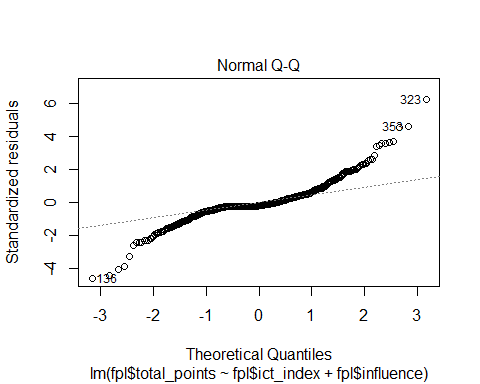
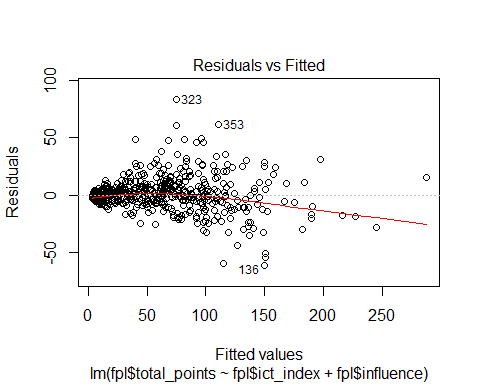
##   
## Call:  
## lm(formula = fpl$total\_points ~ fpl$bps + fpl$threat + fpl$bonus,   
## data = fpl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -49.439 -2.167 -1.231 3.163 41.790   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.368460 0.487326 2.808 0.00513 \*\*   
## fpl$bps 0.154320 0.002516 61.330 < 2e-16 \*\*\*  
## fpl$threat 0.045497 0.001605 28.344 < 2e-16 \*\*\*  
## fpl$bonus 1.206491 0.117297 10.286 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.579 on 643 degrees of freedom  
## Multiple R-squared: 0.9704, Adjusted R-squared: 0.9703   
## F-statistic: 7035 on 3 and 643 DF, p-value: < 2.2e-16

# Checking multicolinearity and conditions of inference for the above models

vif(ictinfluence.log) #checking multicolinearity between ict\_index and influence. not problematic.

## fpl$ict\_index fpl$influence   
## 3.951432 3.951432

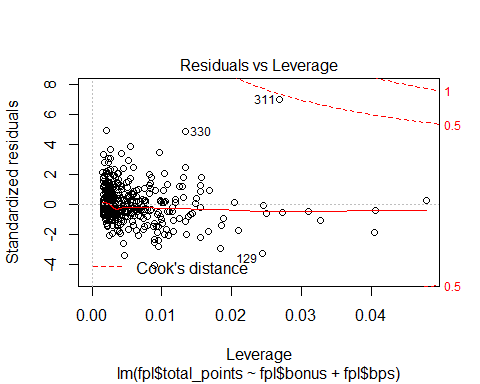
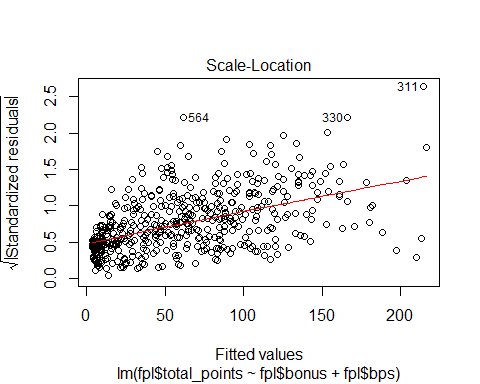
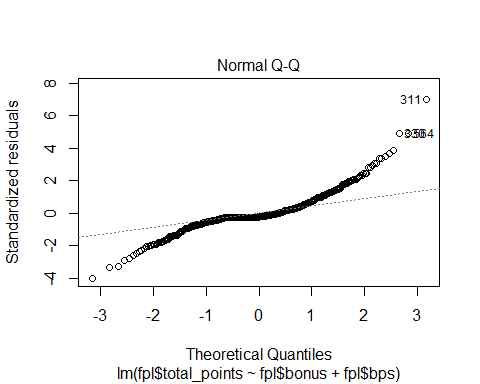
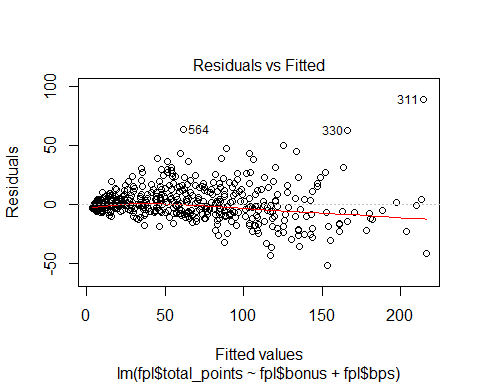
plot(ictinfluence.log) #normality is problematic for all. constant variance is ok.



vif(bonusbps.log) #checking multicolinearity between bonus and bps. not problematic.

## fpl$bonus fpl$bps   
## 2.762732 2.762732

plot(bonusbps.log) #constant variance is good



vif(current.log) #checking multicolinearity. bps and influence problematic.

## fpl$bonus fpl$bps fpl$ict\_index fpl$influence   
## 3.522437 19.242288 4.822074 23.628009

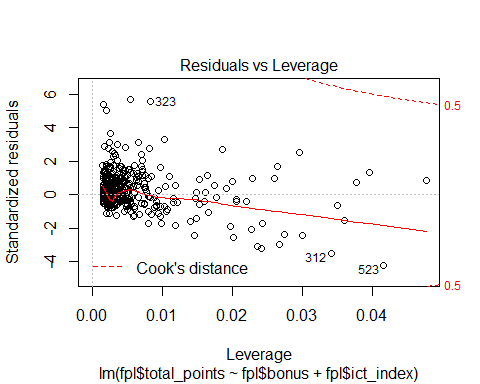
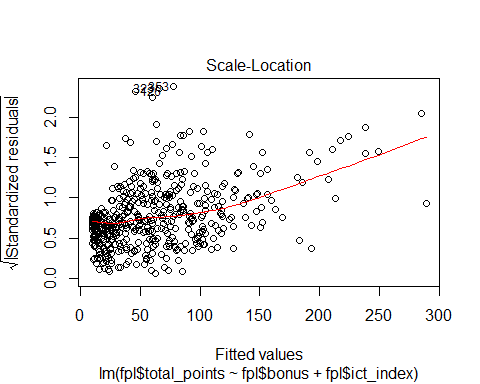
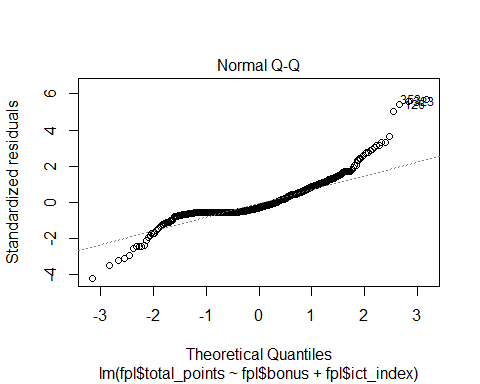
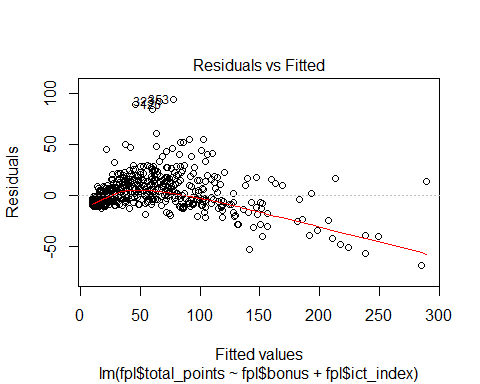
vif(current1.log) #checking multicolinearity. bps and influence problematic.

## fpl$bps fpl$influence   
## 18.60396 18.60396

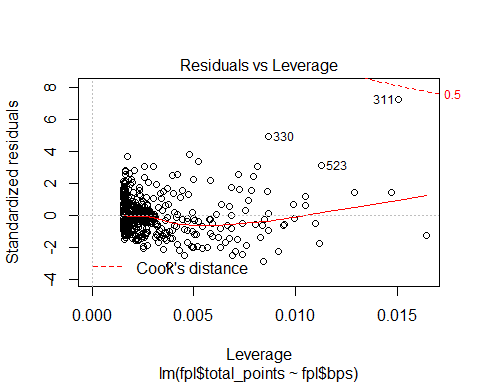
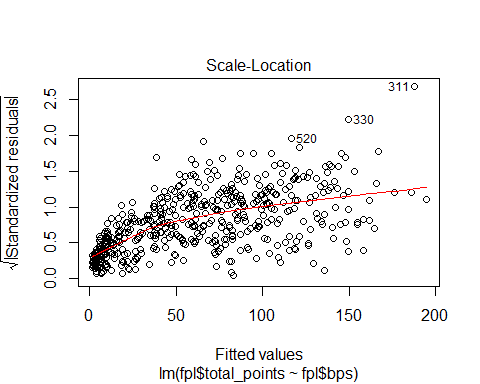
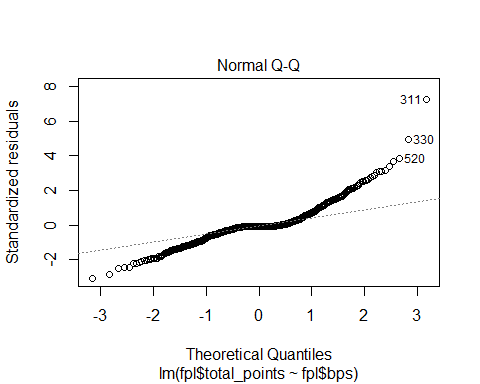
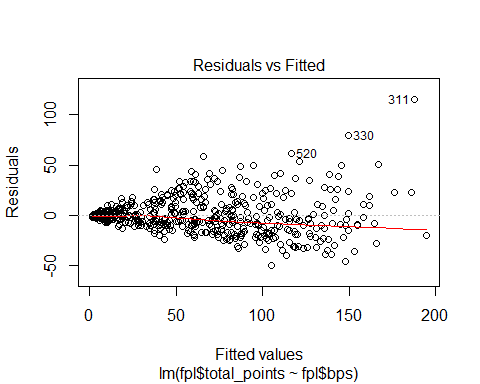
vif(current2.log) #checking multicolinearity. looks good.

## fpl$bonus fpl$ict\_index   
## 3.011688 3.011688

plot(current2.log)# constant variance is bad.



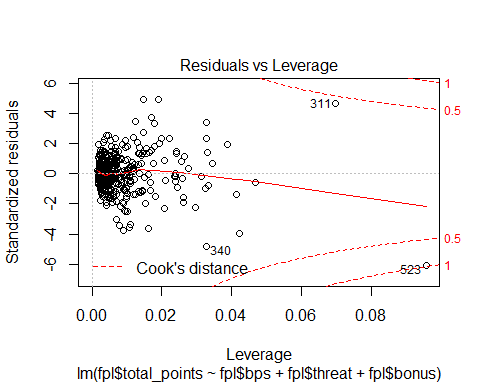
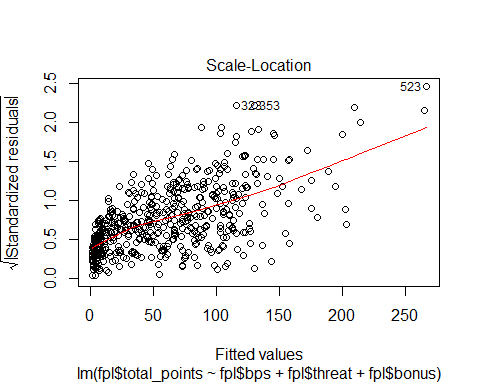
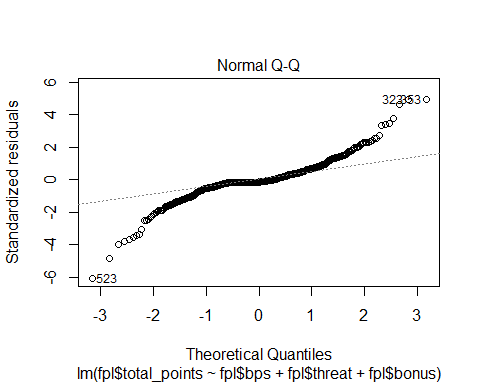
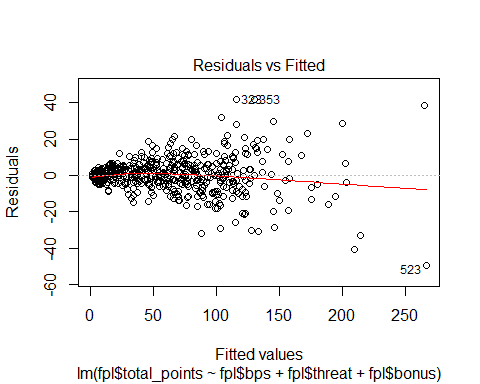
plot(current3.log)#constant variance and normality is problematic



vif(current8.log) # no problem of multicolinearity.

## fpl$bps fpl$threat fpl$bonus   
## 2.762733 1.935197 3.699787

plot(current8.log) # linearity is good, so is constant variance,normality is problematic, there is an influential point



# Fitting Potential Transformation models

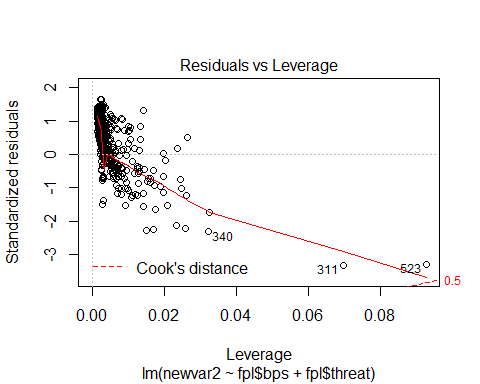
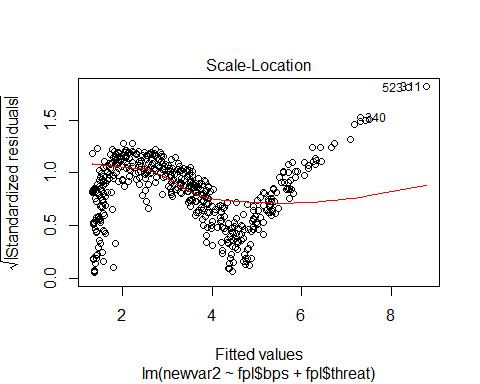
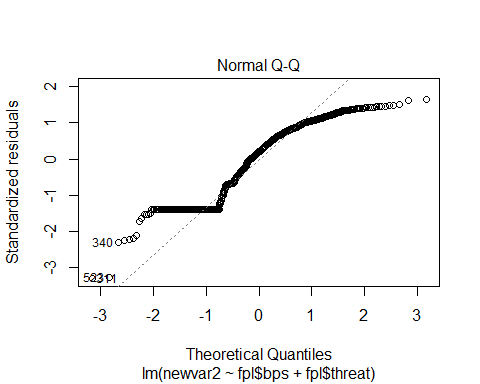
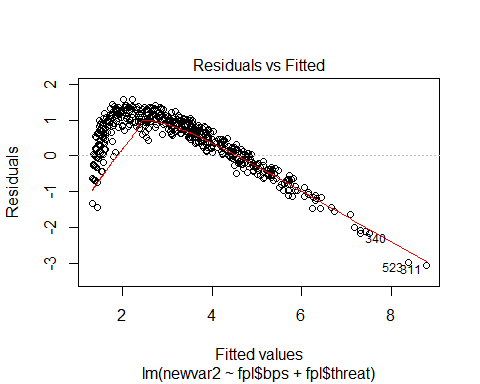
newvar2<-log(fpl$total\_points+1) # log transformation of Total Points is done because the conditions of inference for the final model are not met.  
current10.log <- lm(newvar2~ fpl$bps + fpl$threat ,data=fpl)  
summary(current10.log)

##   
## Call:  
## lm(formula = newvar2 ~ fpl$bps + fpl$threat, data = fpl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.0530 -0.8711 0.1898 0.8286 1.5692   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.3269156 0.0532441 24.921 < 2e-16 \*\*\*  
## fpl$bps 0.0062393 0.0002029 30.751 < 2e-16 \*\*\*  
## fpl$threat 0.0009228 0.0001547 5.967 3.99e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9565 on 644 degrees of freedom  
## Multiple R-squared: 0.7267, Adjusted R-squared: 0.7258   
## F-statistic: 856.1 on 2 and 644 DF, p-value: < 2.2e-16

vif(current10.log) # no problem of multicolinearity.

## fpl$bps fpl$threat   
## 1.445064 1.445064

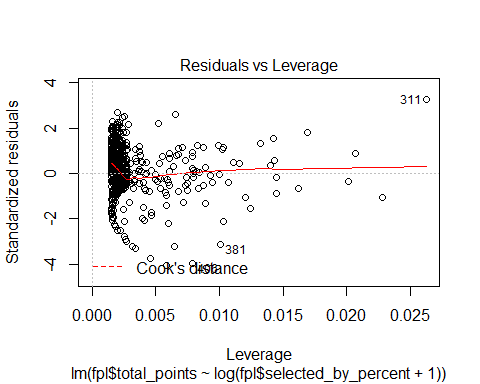
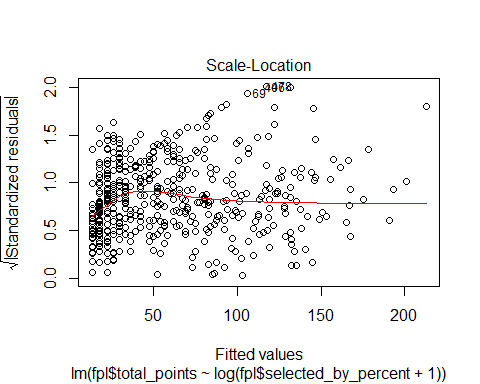
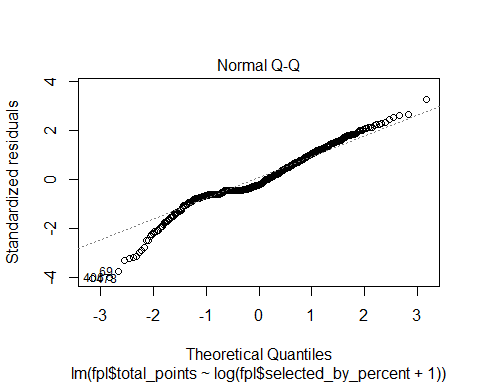
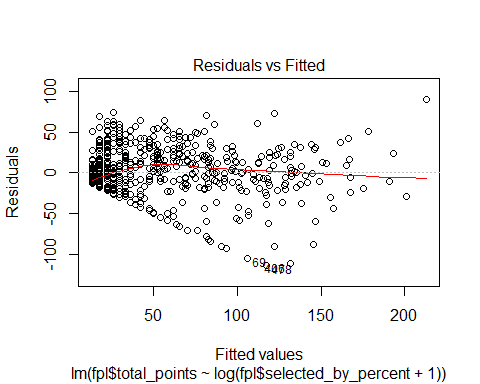
plot(current10.log)# no linear and constant variance, no normality, less adj r-squared of 0.73 compared to the final model of 0.96. So, log transformation of Total Points is not useful here.



#log transformation of selected by percent because the scatterplots of Total Points against Selected By Percent showed log relationship.  
current4.log <- lm(fpl$total\_points~ log(fpl$selected\_by\_percent+1) ,data=fpl)  
summary(current4.log)

##   
## Call:  
## lm(formula = fpl$total\_points ~ log(fpl$selected\_by\_percent +   
## 1), data = fpl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -113.337 -13.828 -6.123 18.195 90.178   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.123 1.459 8.992 <2e-16 \*\*\*  
## log(fpl$selected\_by\_percent + 1) 49.244 1.324 37.184 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 28.1 on 645 degrees of freedom  
## Multiple R-squared: 0.6819, Adjusted R-squared: 0.6814   
## F-statistic: 1383 on 1 and 645 DF, p-value: < 2.2e-16

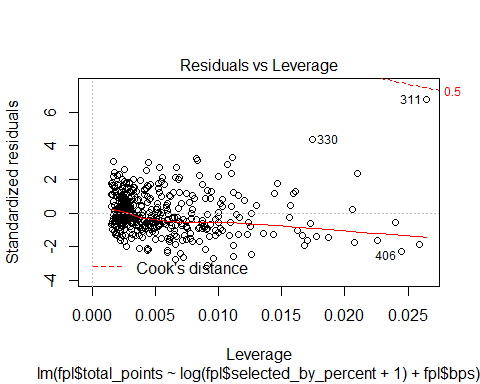
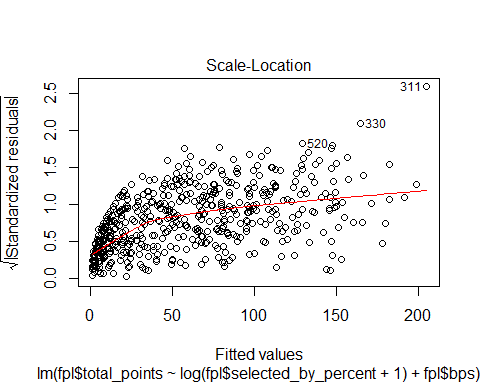
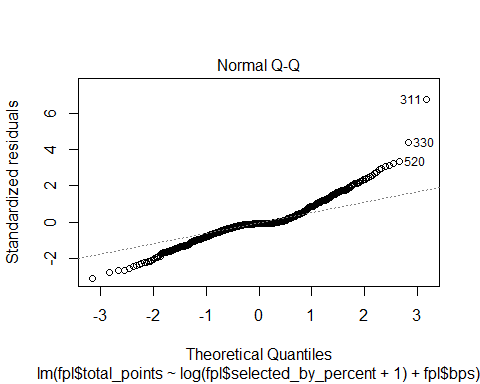
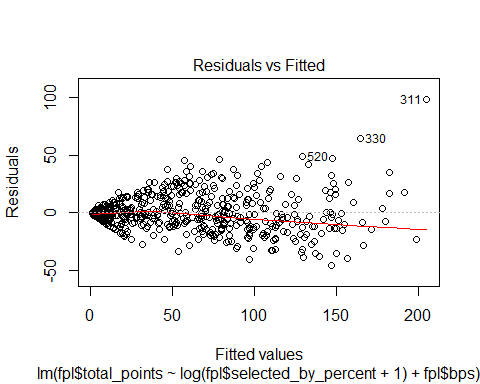
plot(current4.log)# linearity is good. Constant variance is questionable. best normality yet but still problematic. adjusted R-squared of 0.681 is pretty less compared to the final model of 0.96. Thus, the transformation is not useful here.



#Just experimenting with best single predictor and log of Selected By Percent  
current5.log <- lm(fpl$total\_points~ log(fpl$selected\_by\_percent+1) + fpl$bps ,data=fpl)  
summary(current5.log)

##   
## Call:  
## lm(formula = fpl$total\_points ~ log(fpl$selected\_by\_percent +   
## 1) + fpl$bps, data = fpl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -45.580 -6.561 -0.853 4.824 98.407   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.853159 0.825138 1.034 0.302   
## log(fpl$selected\_by\_percent + 1) 12.143799 1.142932 10.625 <2e-16 \*\*\*  
## fpl$bps 0.175361 0.004279 40.984 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 14.8 on 644 degrees of freedom  
## Multiple R-squared: 0.9118, Adjusted R-squared: 0.9116   
## F-statistic: 3330 on 2 and 644 DF, p-value: < 2.2e-16

plot(current5.log) #normality and constant variance is problematic. Linearity is ok. However, the normality is better than the final model. But adjusted R-squared of 91.2 is less than the final model (of 96.6%) of same number of variables. Thus, not considered.



# Stepwise Regression. Not helpful. Shows the best model to be the starting model with all the variables.

newvariable<-log(fpl$selected\_by\_percent+1)  
fullcat<-lm(fpl$total\_points~fpl$influence+fpl$threat+fpl$now\_cost+newvariable+ fpl$ict\_index+fpl$bonus+fpl$bps+fpl$creativity+fpl$minutes, data=fpl)  
  
MSE <- (summary(fullcat)$sigma)^2  
none <- lm(fpl$total\_points~1,data=fpl)  
step(none,scope=list(upper=fullcat),scale=MSE)

## Start: AIC=24606.21  
## fpl$total\_points ~ 1  
##   
## Df Sum of Sq RSS Cp  
## + fpl$bps 1 1434836 165857 1973.4  
## + fpl$influence 1 1412354 188339 2328.1  
## + fpl$ict\_index 1 1352426 248266 3273.4  
## + fpl$minutes 1 1303261 297431 4049.0  
## + fpl$bonus 1 1217589 383103 5400.5  
## + newvariable 1 1091500 509193 7389.6  
## + fpl$threat 1 886612 714081 10621.8  
## + fpl$creativity 1 861479 739214 11018.2  
## + fpl$now\_cost 1 523334 1077359 16352.5  
## <none> 1600692 24606.2  
##   
## Step: AIC=1973.42  
## fpl$total\_points ~ fpl$bps  
##   
## Df Sum of Sq RSS Cp  
## + fpl$threat 1 110746 55111 228.38  
## + fpl$ict\_index 1 94160 71696 490.02  
## + fpl$bonus 1 59405 106452 1038.30  
## + fpl$now\_cost 1 57998 107858 1060.49  
## + fpl$creativity 1 39524 126333 1351.93  
## + newvariable 1 24738 141119 1585.17  
## + fpl$influence 1 10028 155829 1817.23  
## <none> 165857 1973.42  
## + fpl$minutes 1 69 165787 1974.33  
## - fpl$bps 1 1434836 1600692 24606.22  
##   
## Step: AIC=228.38  
## fpl$total\_points ~ fpl$bps + fpl$threat  
##   
## Df Sum of Sq RSS Cp  
## + fpl$bonus 1 7787 47324 107.55  
## + fpl$influence 1 2930 52181 184.16  
## + newvariable 1 2925 52186 184.24  
## + fpl$now\_cost 1 678 54433 219.69  
## + fpl$minutes 1 473 54638 222.93  
## + fpl$creativity 1 407 54703 223.95  
## <none> 55111 228.38  
## + fpl$ict\_index 1 0 55111 230.38  
## - fpl$threat 1 110746 165857 1973.42  
## - fpl$bps 1 658970 714081 10621.75  
##   
## Step: AIC=107.55  
## fpl$total\_points ~ fpl$bps + fpl$threat + fpl$bonus  
##   
## Df Sum of Sq RSS Cp  
## + fpl$influence 1 3096 44228 60.708  
## + newvariable 1 796 46529 96.998  
## + fpl$minutes 1 541 46783 101.007  
## + fpl$creativity 1 435 46889 102.679  
## <none> 47324 107.548  
## + fpl$now\_cost 1 95 47229 108.041  
## + fpl$ict\_index 1 0 47324 109.547  
## - fpl$bonus 1 7787 55111 228.382  
## - fpl$threat 1 59128 106452 1038.301  
## - fpl$bps 1 276836 324160 4472.691  
##   
## Step: AIC=60.71  
## fpl$total\_points ~ fpl$bps + fpl$threat + fpl$bonus + fpl$influence  
##   
## Df Sum of Sq RSS Cp  
## + fpl$minutes 1 1037 43191 46.343  
## + newvariable 1 712 43516 51.479  
## <none> 44228 60.708  
## + fpl$creativity 1 70 44158 61.604  
## + fpl$ict\_index 1 59 44169 61.770  
## + fpl$now\_cost 1 6 44223 62.619  
## - fpl$influence 1 3096 47324 107.548  
## - fpl$bonus 1 7953 52181 184.160  
## - fpl$bps 1 57717 101945 969.199  
## - fpl$threat 1 60339 104568 1010.573  
##   
## Step: AIC=46.34  
## fpl$total\_points ~ fpl$bps + fpl$threat + fpl$bonus + fpl$influence +   
## fpl$minutes  
##   
## Df Sum of Sq RSS Cp  
## + newvariable 1 1162 42029 30.010  
## + fpl$now\_cost 1 255 42936 44.317  
## <none> 43191 46.343  
## + fpl$creativity 1 60 43131 47.402  
## + fpl$ict\_index 1 52 43139 47.530  
## - fpl$minutes 1 1037 44228 60.708  
## - fpl$influence 1 3592 46783 101.007  
## - fpl$bonus 1 8765 51956 182.615  
## - fpl$bps 1 28610 71801 495.668  
## - fpl$threat 1 53840 97031 893.682  
##   
## Step: AIC=30.01  
## fpl$total\_points ~ fpl$bps + fpl$threat + fpl$bonus + fpl$influence +   
## fpl$minutes + newvariable  
##   
## Df Sum of Sq RSS Cp  
## + fpl$now\_cost 1 357 41672 26.382  
## + fpl$creativity 1 163 41866 29.446  
## + fpl$ict\_index 1 149 41880 29.657  
## <none> 42029 30.010  
## - newvariable 1 1162 43191 46.343  
## - fpl$minutes 1 1488 43516 51.479  
## - fpl$influence 1 3611 45640 84.976  
## - fpl$bonus 1 7342 49371 143.837  
## - fpl$bps 1 23277 65306 395.215  
## - fpl$threat 1 49147 91176 803.317  
##   
## Step: AIC=26.38  
## fpl$total\_points ~ fpl$bps + fpl$threat + fpl$bonus + fpl$influence +   
## fpl$minutes + newvariable + fpl$now\_cost  
##   
## Df Sum of Sq RSS Cp  
## <none> 41672 26.382  
## + fpl$creativity 1 119.7 41552 26.494  
## + fpl$ict\_index 1 108.3 41564 26.673  
## - fpl$now\_cost 1 356.8 42029 30.010  
## - newvariable 1 1263.7 42936 44.317  
## - fpl$minutes 1 1839.3 43511 53.398  
## - fpl$influence 1 3410.3 45082 78.179  
## - fpl$bonus 1 7349.6 49021 140.323  
## - fpl$bps 1 19139.8 60812 326.316  
## - fpl$threat 1 25180.1 66852 421.603

##   
## Call:  
## lm(formula = fpl$total\_points ~ fpl$bps + fpl$threat + fpl$bonus +   
## fpl$influence + fpl$minutes + newvariable + fpl$now\_cost,   
## data = fpl)  
##   
## Coefficients:  
## (Intercept) fpl$bps fpl$threat fpl$bonus fpl$influence   
## -4.011454 0.158729 0.043721 1.363399 -0.040044   
## fpl$minutes newvariable fpl$now\_cost   
## 0.006697 3.124039 0.094657

# Backward Elimination ( not helpful like stepwise regression)

step(fullcat,scale=MSE,direction="backward")

## Start: AIC=10  
## fpl$total\_points ~ fpl$influence + fpl$threat + fpl$now\_cost +   
## newvariable + fpl$ict\_index + fpl$bonus + fpl$bps + fpl$creativity +   
## fpl$minutes  
##   
## Df Sum of Sq RSS Cp  
## <none> 40380 10.000  
## - fpl$now\_cost 1 271.8 40652 12.287  
## - fpl$influence 1 1019.2 41399 24.078  
## - fpl$minutes 1 1159.3 41539 26.288  
## - fpl$ict\_index 1 1172.4 41552 26.494  
## - fpl$creativity 1 1183.7 41564 26.673  
## - newvariable 1 1284.2 41664 28.259  
## - fpl$threat 1 1354.0 41734 29.359  
## - fpl$bonus 1 6410.3 46790 109.124  
## - fpl$bps 1 18095.0 58475 293.451

##   
## Call:  
## lm(formula = fpl$total\_points ~ fpl$influence + fpl$threat +   
## fpl$now\_cost + newvariable + fpl$ict\_index + fpl$bonus +   
## fpl$bps + fpl$creativity + fpl$minutes, data = fpl)  
##   
## Coefficients:  
## (Intercept) fpl$influence fpl$threat fpl$now\_cost   
## -3.733798 0.535411 0.609578 0.083072   
## newvariable fpl$ict\_index fpl$bonus fpl$bps   
## 3.179759 -5.696649 1.285205 0.160021   
## fpl$creativity fpl$minutes   
## 0.574026 0.005451

# Forward Selection (not helpful like stepwise regression and backeward elimination)

step(none,scope=list(upper=fullcat),scale=MSE,direction = "forward")

## Start: AIC=24606.21  
## fpl$total\_points ~ 1  
##   
## Df Sum of Sq RSS Cp  
## + fpl$bps 1 1434836 165857 1973.4  
## + fpl$influence 1 1412354 188339 2328.1  
## + fpl$ict\_index 1 1352426 248266 3273.4  
## + fpl$minutes 1 1303261 297431 4049.0  
## + fpl$bonus 1 1217589 383103 5400.5  
## + newvariable 1 1091500 509193 7389.6  
## + fpl$threat 1 886612 714081 10621.8  
## + fpl$creativity 1 861479 739214 11018.2  
## + fpl$now\_cost 1 523334 1077359 16352.5  
## <none> 1600692 24606.2  
##   
## Step: AIC=1973.42  
## fpl$total\_points ~ fpl$bps  
##   
## Df Sum of Sq RSS Cp  
## + fpl$threat 1 110746 55111 228.38  
## + fpl$ict\_index 1 94160 71696 490.02  
## + fpl$bonus 1 59405 106452 1038.30  
## + fpl$now\_cost 1 57998 107858 1060.49  
## + fpl$creativity 1 39524 126333 1351.93  
## + newvariable 1 24738 141119 1585.17  
## + fpl$influence 1 10028 155829 1817.23  
## <none> 165857 1973.42  
## + fpl$minutes 1 69 165787 1974.33  
##   
## Step: AIC=228.38  
## fpl$total\_points ~ fpl$bps + fpl$threat  
##   
## Df Sum of Sq RSS Cp  
## + fpl$bonus 1 7786.6 47324 107.55  
## + fpl$influence 1 2930.0 52181 184.16  
## + newvariable 1 2925.2 52186 184.24  
## + fpl$now\_cost 1 677.9 54433 219.69  
## + fpl$minutes 1 472.7 54638 222.93  
## + fpl$creativity 1 407.4 54703 223.95  
## <none> 55111 228.38  
## + fpl$ict\_index 1 0.1 55111 230.38  
##   
## Step: AIC=107.55  
## fpl$total\_points ~ fpl$bps + fpl$threat + fpl$bonus  
##   
## Df Sum of Sq RSS Cp  
## + fpl$influence 1 3096.00 44228 60.708  
## + newvariable 1 795.55 46529 96.998  
## + fpl$minutes 1 541.37 46783 101.007  
## + fpl$creativity 1 435.40 46889 102.679  
## <none> 47324 107.548  
## + fpl$now\_cost 1 95.49 47229 108.041  
## + fpl$ict\_index 1 0.03 47324 109.547  
##   
## Step: AIC=60.71  
## fpl$total\_points ~ fpl$bps + fpl$threat + fpl$bonus + fpl$influence  
##   
## Df Sum of Sq RSS Cp  
## + fpl$minutes 1 1037.40 43191 46.343  
## + newvariable 1 711.82 43516 51.479  
## <none> 44228 60.708  
## + fpl$creativity 1 69.95 44158 61.604  
## + fpl$ict\_index 1 59.43 44169 61.770  
## + fpl$now\_cost 1 5.63 44223 62.619  
##   
## Step: AIC=46.34  
## fpl$total\_points ~ fpl$bps + fpl$threat + fpl$bonus + fpl$influence +   
## fpl$minutes  
##   
## Df Sum of Sq RSS Cp  
## + newvariable 1 1162.08 42029 30.010  
## + fpl$now\_cost 1 255.18 42936 44.317  
## <none> 43191 46.343  
## + fpl$creativity 1 59.60 43131 47.402  
## + fpl$ict\_index 1 51.50 43139 47.530  
##   
## Step: AIC=30.01  
## fpl$total\_points ~ fpl$bps + fpl$threat + fpl$bonus + fpl$influence +   
## fpl$minutes + newvariable  
##   
## Df Sum of Sq RSS Cp  
## + fpl$now\_cost 1 356.80 41672 26.382  
## + fpl$creativity 1 162.58 41866 29.446  
## + fpl$ict\_index 1 149.17 41880 29.657  
## <none> 42029 30.010  
##   
## Step: AIC=26.38  
## fpl$total\_points ~ fpl$bps + fpl$threat + fpl$bonus + fpl$influence +   
## fpl$minutes + newvariable + fpl$now\_cost  
##   
## Df Sum of Sq RSS Cp  
## <none> 41672 26.382  
## + fpl$creativity 1 119.65 41552 26.494  
## + fpl$ict\_index 1 108.34 41564 26.673

##   
## Call:  
## lm(formula = fpl$total\_points ~ fpl$bps + fpl$threat + fpl$bonus +   
## fpl$influence + fpl$minutes + newvariable + fpl$now\_cost,   
## data = fpl)  
##   
## Coefficients:  
## (Intercept) fpl$bps fpl$threat fpl$bonus fpl$influence   
## -4.011454 0.158729 0.043721 1.363399 -0.040044   
## fpl$minutes newvariable fpl$now\_cost   
## 0.006697 3.124039 0.094657

# Summary Regression including log of selected by percent. Very helpful. Shows the best model to be: total points = threat + BPS because the model has high adj R-squared of 96.6% and is very simple with two predictors which are easy to interpret.

all<-regsubsets(fpl$total\_points~fpl$influence+fpl$threat+fpl$now\_cost+newvariable+ fpl$ict\_index+fpl$bonus+fpl$bps+fpl$creativity+fpl$minutes, data=fpl)  
#summary(all)  
cbind(as.data.frame(summary(all)$which), summary(all)$rsq, summary(all)$adjr2, summary(all)$cp)

## (Intercept) fpl$influence fpl$threat fpl$now\_cost newvariable  
## 1 TRUE FALSE FALSE FALSE FALSE  
## 2 TRUE FALSE TRUE FALSE FALSE  
## 3 TRUE FALSE TRUE FALSE FALSE  
## 4 TRUE TRUE TRUE FALSE FALSE  
## 5 TRUE TRUE TRUE FALSE FALSE  
## 6 TRUE TRUE TRUE FALSE TRUE  
## 7 TRUE TRUE TRUE FALSE TRUE  
## 8 TRUE TRUE TRUE FALSE TRUE  
## fpl$ict\_index fpl$bonus fpl$bps fpl$creativity fpl$minutes  
## 1 FALSE FALSE TRUE FALSE FALSE  
## 2 FALSE FALSE TRUE FALSE FALSE  
## 3 FALSE TRUE TRUE FALSE FALSE  
## 4 FALSE TRUE TRUE FALSE FALSE  
## 5 FALSE TRUE TRUE FALSE TRUE  
## 6 FALSE TRUE TRUE FALSE TRUE  
## 7 TRUE TRUE TRUE TRUE FALSE  
## 8 TRUE TRUE TRUE TRUE TRUE  
## summary(all)$rsq summary(all)$adjr2 summary(all)$cp  
## 1 0.8963844 0.8962238 1973.41951  
## 2 0.9655707 0.9654638 228.38225  
## 3 0.9704352 0.9702972 107.54766  
## 4 0.9723693 0.9721972 60.70766  
## 5 0.9730174 0.9728070 46.34247  
## 6 0.9737434 0.9734973 30.01038  
## 7 0.9740418 0.9737574 24.47678  
## 8 0.9746037 0.9742853 12.28729

# Checking conditions of inference for the best model

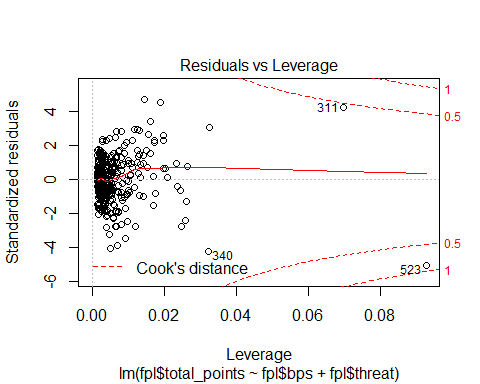
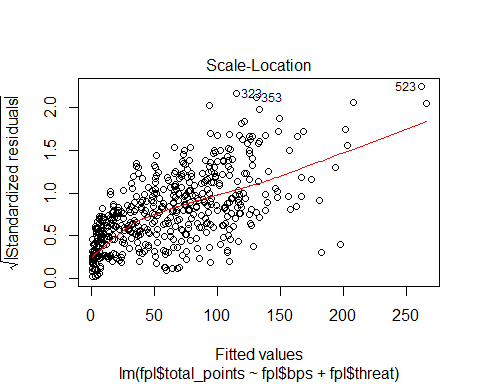
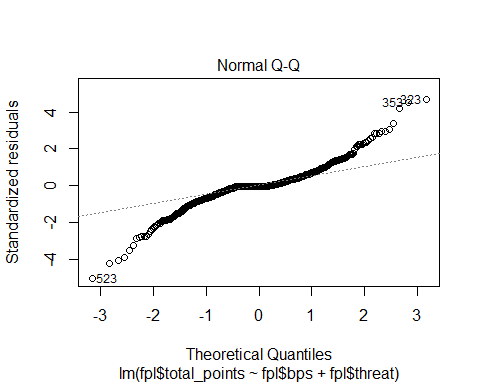
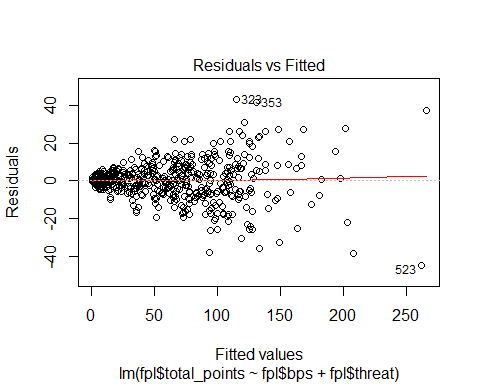
current6.log <- lm(fpl$total\_points~ fpl$bps + fpl$threat ,data=fpl)  
summary(current6.log) #the slopes and p-values of the coefficients requried for the t-test of the coefficients, and F-statistic are reported from this summary. The predictors are statistically significant because of their p-values being closer to zero. The model is useful because its F-statistic of 9030 is >> 1.

##   
## Call:  
## lm(formula = fpl$total\_points ~ fpl$bps + fpl$threat, data = fpl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -44.565 -2.902 -0.369 3.329 42.950   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.369046 0.514932 0.717 0.474   
## fpl$bps 0.172193 0.001962 87.752 <2e-16 \*\*\*  
## fpl$threat 0.053806 0.001496 35.974 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.251 on 644 degrees of freedom  
## Multiple R-squared: 0.9656, Adjusted R-squared: 0.9655   
## F-statistic: 9030 on 2 and 644 DF, p-value: < 2.2e-16

vif(current6.log) # no problem of multicolinearity.

## fpl$bps fpl$threat   
## 1.445064 1.445064

plot(current6.log) # constant variance is problematic. Normality is problematic. Linearity is good. Cook distance shows that 523 is an influential point. So, a new model is fitted below without the influential point to check whether anything improves.

 # The best model without influential point

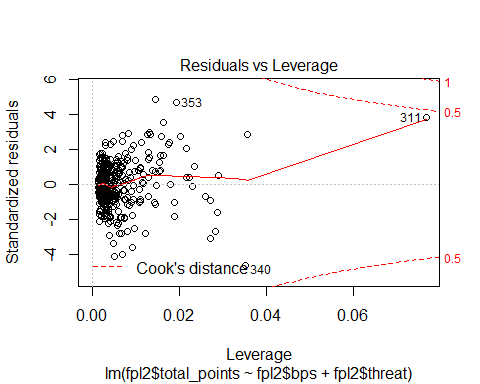
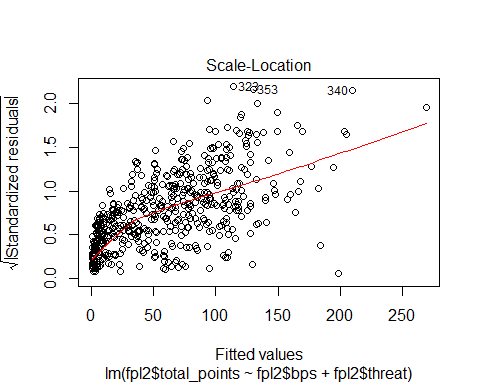
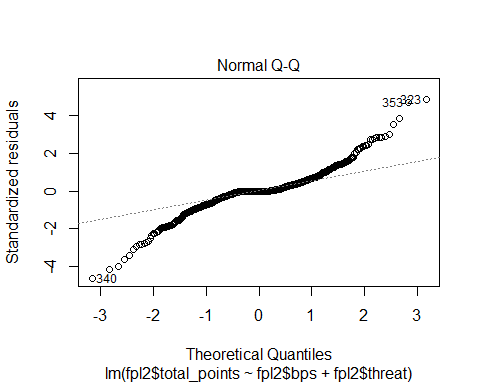
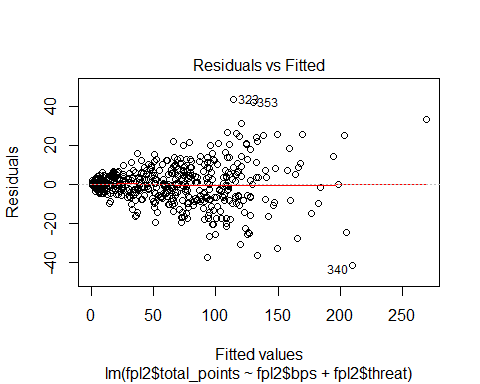
fpl2<- fpl[-523,] #removing influential point  
current7.log <- lm(fpl2$total\_points~ fpl2$bps + fpl2$threat ,data=fpl2)  
summary(current7.log) #adj r squared only slightly improves.

##   
## Call:  
## lm(formula = fpl2$total\_points ~ fpl2$bps + fpl2$threat, data = fpl2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -41.247 -2.878 -0.203 3.356 43.645   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.203482 0.506015 0.402 0.688   
## fpl2$bps 0.171399 0.001931 88.782 <2e-16 \*\*\*  
## fpl2$threat 0.056070 0.001531 36.620 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.072 on 643 degrees of freedom  
## Multiple R-squared: 0.9663, Adjusted R-squared: 0.9662   
## F-statistic: 9230 on 2 and 643 DF, p-value: < 2.2e-16

vif(current7.log) # no problem of multicolinearity.

## fpl2$bps fpl2$threat   
## 1.440167 1.440167

plot(current7.log) #linearity improves. constant variance still problematic.normality does not change still problematic.no influential point.Even though the best model without influential point is slightly improved from the model with influential point, I decided to include influential point in my final model because the improvement is not a significant one.



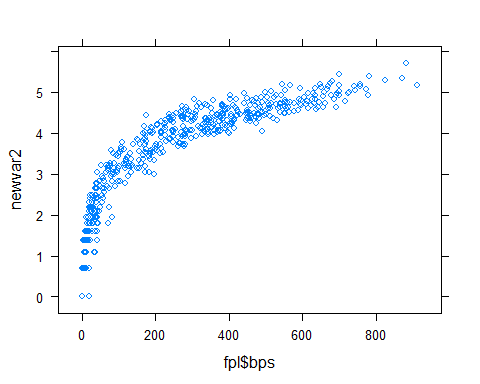
# Confidence intervals of the best model:

confint(current6.log)

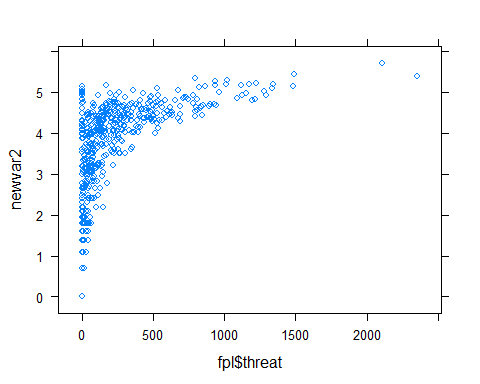
## 2.5 % 97.5 %  
## (Intercept) -0.64210328 1.38019463  
## fpl$bps 0.16834008 0.17604652  
## fpl$threat 0.05086912 0.05674318

# More Experiments (Extended from potential transformations) with other Models:

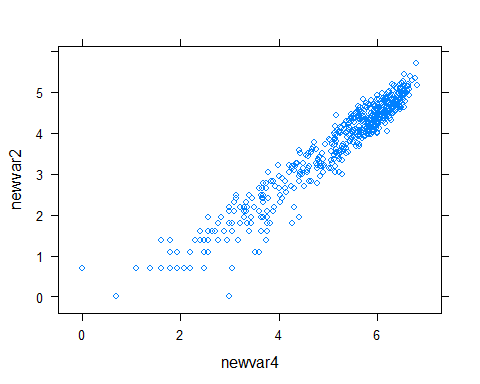
newvar2<-log(fpl$total\_points+1)  
newvar4<-log(fpl$bps+2)  
newvar5<-log(fpl$threat+1)  
xyplot(newvar2~fpl$bps, data=fpl)



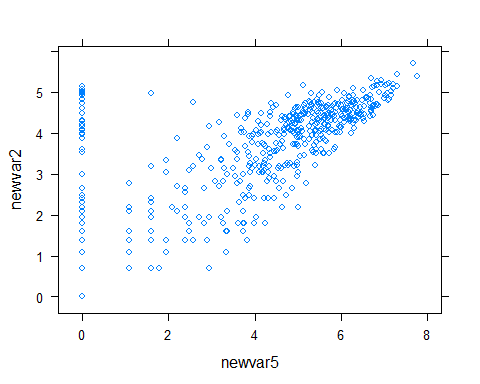
xyplot(newvar2~fpl$threat, data=fpl)



xyplot(newvar2~newvar4, data=fpl)



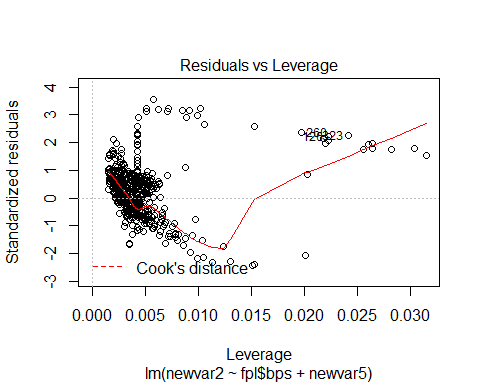
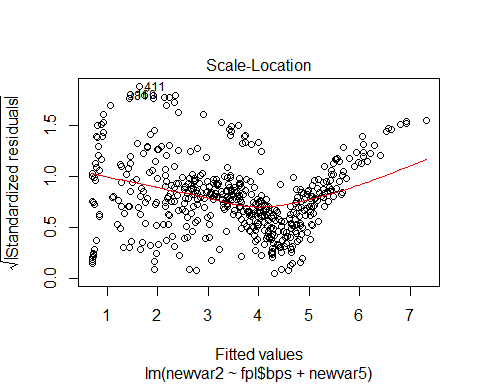
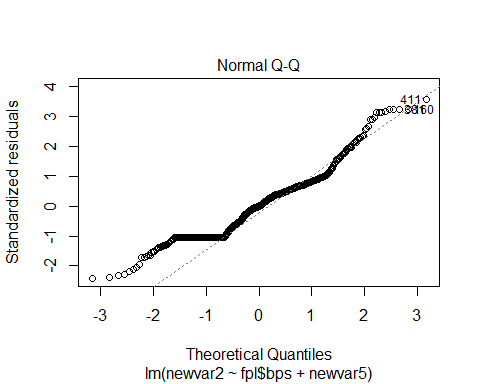
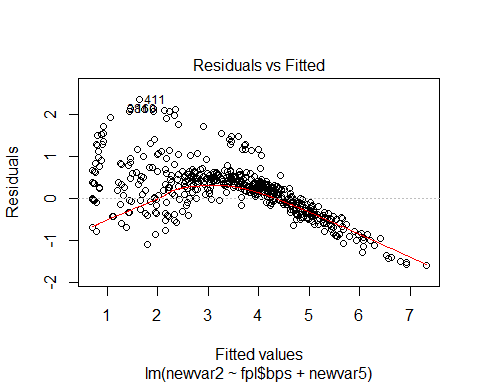
xyplot(newvar2~newvar5, data=fpl)



logthreat <- lm(newvar2~ fpl$bps + newvar5 ,data=fpl)  
summary(logthreat)

##   
## Call:  
## lm(formula = newvar2 ~ fpl$bps + newvar5, data = fpl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.59421 -0.70299 -0.01418 0.40023 2.35390   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.7029881 0.0436903 16.09 <2e-16 \*\*\*  
## fpl$bps 0.0043354 0.0001499 28.93 <2e-16 \*\*\*  
## newvar5 0.3643387 0.0131919 27.62 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6648 on 644 degrees of freedom  
## Multiple R-squared: 0.868, Adjusted R-squared: 0.8676   
## F-statistic: 2117 on 2 and 644 DF, p-value: < 2.2e-16

plot (logthreat) # conditions of inference got worse. So, rejected.



squareboth <- lm(newvar2~ fpl$bps + I(fpl$bps^2)+ I(fpl$threat^2)+ fpl$threat ,data=fpl)  
summary(squareboth)

##   
## Call:  
## lm(formula = newvar2 ~ fpl$bps + I(fpl$bps^2) + I(fpl$threat^2) +   
## fpl$threat, data = fpl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.0623 -0.5088 -0.1097 0.4827 2.3686   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.995e-01 4.080e-02 19.595 < 2e-16 \*\*\*  
## fpl$bps 1.485e-02 4.154e-04 35.760 < 2e-16 \*\*\*  
## I(fpl$bps^2) -1.409e-05 6.046e-07 -23.308 < 2e-16 \*\*\*  
## I(fpl$threat^2) -2.932e-07 1.598e-07 -1.834 0.0671 .   
## fpl$threat 1.155e-03 2.307e-04 5.008 7.11e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6471 on 642 degrees of freedom  
## Multiple R-squared: 0.8753, Adjusted R-squared: 0.8745   
## F-statistic: 1126 on 4 and 642 DF, p-value: < 2.2e-16

plot(squareboth) # conditions of inference got worse. So, rejected.

