Predicting Number of Shots on Target for All Time Premier League Player - Final Draft

Emma Tran, Tracy Bui, Duy Nguyen

2022-10-31

Table of Contents

# Abstract

This report analyzes a data set on all time Premier League soccer players’ statistics with the purpose of developing a linear regression model that can predict the number of shots on target from various features of Premier League Players. The features include position, age, number of times hitting woodwork, number of big chances missed, number of interception, number of recoveries, number of aerial battle won, number of assists, and number of passes per match. We found that we can predict the square root of shots on target with the adjusted R^2 = 0.9398, and the residual standard error is 0.7751 square root shots on target. The significant predictors are Age, Hit woodwork, Big chances missed, Interceptions, Recoveries, Assists, and Shots on target.

We discover that older soccer player with more number of hit woodwork, big chances missed, recoveries, assists, and less number of interceptions tends to have a higher number of shots on target. We also found 8 interaction effects, being: one between age and big chances missed, one between age and interceptions, one between age and recoveries, one between hit woodwork and big chances missed, one between hit woodwork and interceptions, oen between big chances missed and interceptions, one between interceptions and assists, and one between recoveries and assists.

# Introduction

This data set is from Kaggle naming “All Time Premier League Player Statistics”. (Source: <https://www.kaggle.com/datasets/rishikeshkanabar/premier-league-player-statistics-updated-daily?fbclid=IwAR0gFPjRhFVvAEMeDZSqjkQMlvxlVeisrNh2cRnxVwyybIZYT_MJiSi3qtI>). This data set contains the statistical figure of 308 players in Premier League. The variables are: ● Shots on Target ->Response variable ● Position (Midfielder or Forward) ● Age (Age of the player) ● Hit woodwork ● Big chances missed ● Interception ● Recoveries ● Aerial battle won ● Assists ● Passes per match

The response variable is the number of shots on target. We are interested in which variables can best predict and how they influence the number of shot on target. We will begin by exploring the data numerically and graphically. We will choose, fit, and assess a series of multiple linear regression models following a strategy that begins with a first-order linear model. Then we will investigate possible interaction effects. We will finish with a thorough diagnostic assessment of the final model and show some example predictions.

# Data Collection (as needed) and Data Characteristics

## Read in data

install.packages(“data.table”)

library("data.table")   
  
shotsOnTarget <- fread("dataset - 2020-09-24.csv",  
 select = c("Position", "Age", "Hit woodwork", "Big chances missed",  
 "Interceptions", "Recoveries", "Aerial battles won", "Assists",  
 "Passes per match", "Shots on target"), fill = TRUE)  
  
head (shotsOnTarget)

## Position Age Hit woodwork Big chances missed Interceptions Recoveries  
## 1: Goalkeeper 28 NA NA NA NA  
## 2: Goalkeeper 26 NA NA NA NA  
## 3: Goalkeeper 25 NA NA NA NA  
## 4: Defender 25 3 NA 208 732  
## 5: Defender 23 0 NA 12 63  
## 6: Defender 19 0 NA 0 0  
## Aerial battles won Assists Passes per match Shots on target  
## 1: NA 0 27.86 NA  
## 2: NA 0 0.00 NA  
## 3: NA 0 0.00 NA  
## 4: 161 18 44.53 NA  
## 5: 10 1 32.44 NA  
## 6: 0 0 0.00 NA

## Exploratory Analysis

### Check distributions of the predictor variables

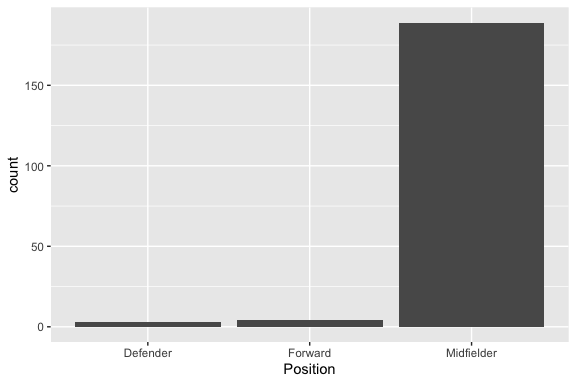
head(is.na(shotsOnTarget))

## Position Age Hit woodwork Big chances missed Interceptions Recoveries  
## [1,] FALSE FALSE TRUE TRUE TRUE TRUE  
## [2,] FALSE FALSE TRUE TRUE TRUE TRUE  
## [3,] FALSE FALSE TRUE TRUE TRUE TRUE  
## [4,] FALSE FALSE FALSE TRUE FALSE FALSE  
## [5,] FALSE FALSE FALSE TRUE FALSE FALSE  
## [6,] FALSE FALSE FALSE TRUE FALSE FALSE  
## Aerial battles won Assists Passes per match Shots on target  
## [1,] TRUE FALSE FALSE TRUE  
## [2,] TRUE FALSE FALSE TRUE  
## [3,] TRUE FALSE FALSE TRUE  
## [4,] FALSE FALSE FALSE TRUE  
## [5,] FALSE FALSE FALSE TRUE  
## [6,] FALSE FALSE FALSE TRUE

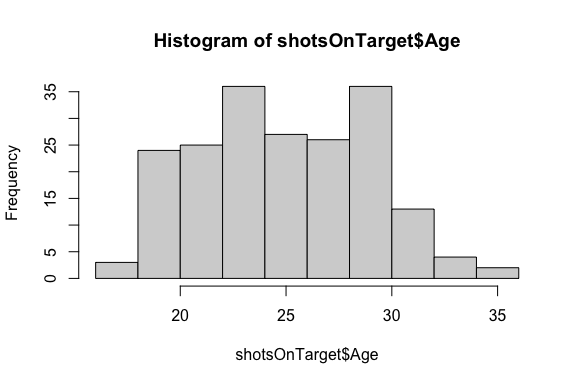
shotsOnTarget = na.omit(shotsOnTarget)

Since we find that there are null values in our data set, we decide to omit all of them. Because our omit function deletes all rows with null values, it only keeps data of Midfielder and Forward. We suspect it is due to these positions creating shots on target while the other positions do not.

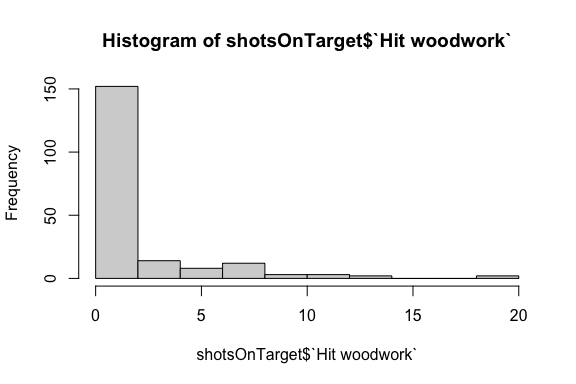
library (tidyr)  
library (ggplot2)  
  
ggplot(data.frame(shotsOnTarget), aes(x=Position)) +  
 geom\_bar()



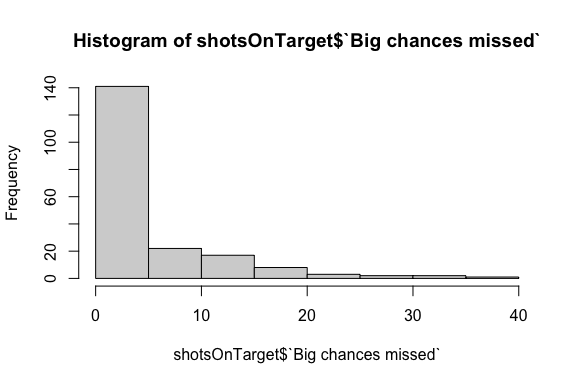
hist(shotsOnTarget$Age)



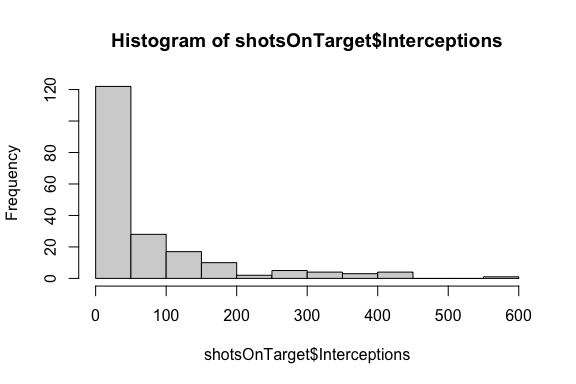
hist(shotsOnTarget$`Hit woodwork`)



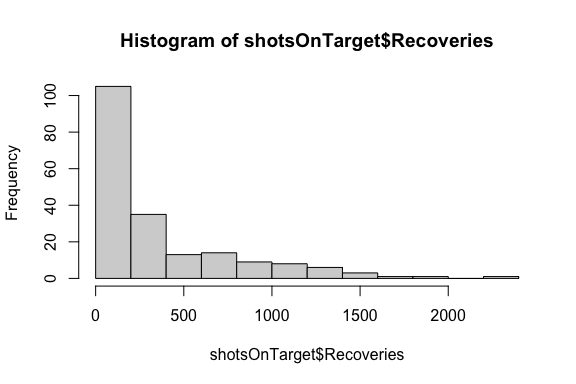
hist(shotsOnTarget$`Big chances missed`)



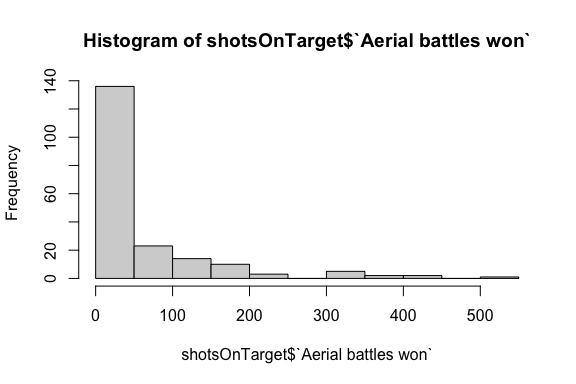
hist(shotsOnTarget$Interceptions)



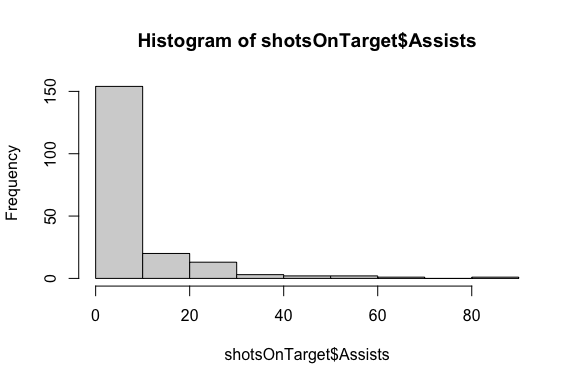
hist(shotsOnTarget$Recoveries)



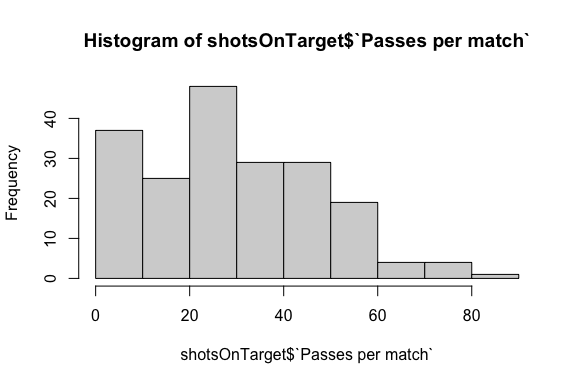
hist(shotsOnTarget$`Aerial battles won`)



hist(shotsOnTarget$`Assists`)



hist(shotsOnTarget$`Passes per match`)

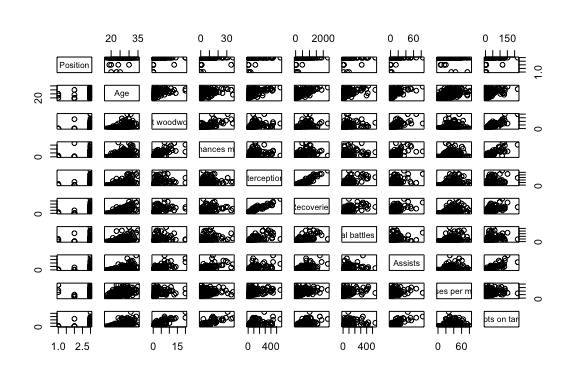


The total count is around 200, which means that about 100 rows were removed by the na.omit function.

According to the histograms, Age is quite normally distributed while all other predictor are right-skewed. Therefore, we will consider the log transformation for all right-skewed predictors.

### Pairwise correlations

plot(shotsOnTarget)



The number of shots on target per soccer player is most correlated with Age, Hit woodwork, Big chances missed, and Assists. There are almost no correlation between any predictor except Interception and Recoveries.

Category predictor variable is Postion (Midfielder, Forward).

# Results

## Model 1: The first model will include all individual predictors.

### First-order model with all predictors

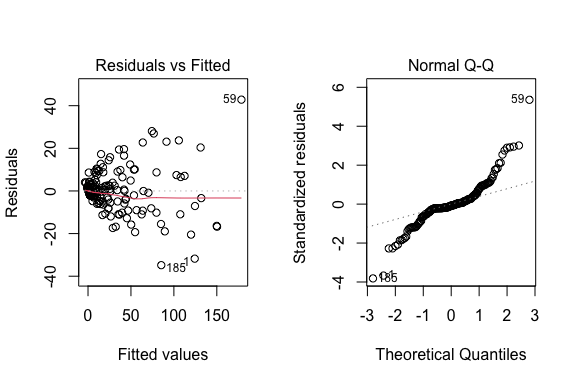
shotsOnTarget$y.offset = shotsOnTarget$`Shots on target`+0.5  
fit1 = lm (y.offset ~ as.factor(Position) + Age + `Hit woodwork` + `Big chances missed` + Interceptions + Recoveries + `Aerial battles won` + `Assists` + `Passes per match`, data=shotsOnTarget)  
  
summary (fit1)

##   
## Call:  
## lm(formula = y.offset ~ as.factor(Position) + Age + `Hit woodwork` +   
## `Big chances missed` + Interceptions + Recoveries + `Aerial battles won` +   
## Assists + `Passes per match`, data = shotsOnTarget)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -34.814 -2.474 -0.746 2.427 42.836   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.5931515 7.3009958 0.492 0.6232   
## as.factor(Position)Forward -4.8455934 7.5489612 -0.642 0.5217   
## as.factor(Position)Midfielder -1.8102556 5.7294505 -0.316 0.7524   
## Age 0.0333013 0.2376444 0.140 0.8887   
## `Hit woodwork` 3.4570800 0.3526216 9.804 < 2e-16 \*\*\*  
## `Big chances missed` 2.0501029 0.1766878 11.603 < 2e-16 \*\*\*  
## Interceptions -0.0009537 0.0258198 -0.037 0.9706   
## Recoveries 0.0179508 0.0070791 2.536 0.0120 \*   
## `Aerial battles won` -0.0253485 0.0121481 -2.087 0.0383 \*   
## Assists 0.6563874 0.1105974 5.935 1.42e-08 \*\*\*  
## `Passes per match` -0.0770271 0.0473442 -1.627 0.1054   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.477 on 185 degrees of freedom  
## Multiple R-squared: 0.935, Adjusted R-squared: 0.9314   
## F-statistic: 265.9 on 10 and 185 DF, p-value: < 2.2e-16

Using all predictor variables, the model explains 93.14% of the variation in the number of shots on target (Adjusted R-square). The residual standard error is 9.477 (shots on target per player). The most significant predictors are Hit woodwork, Big chances missed, Assists.

### Residual Analysis - First-order model

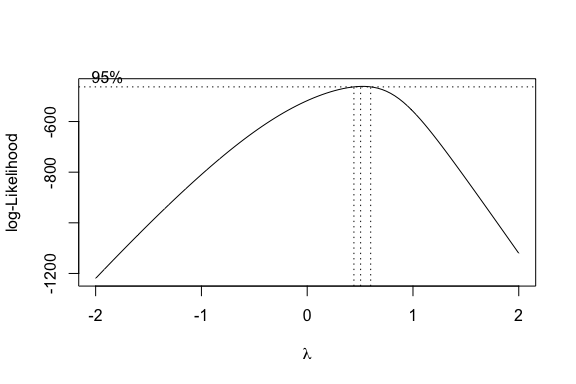
par (mfrow=c(1,2))  
plot (fit1, which=c(1,2))



1. The residuals vs fitted plot show increasing residual variance with increasing fitted values. This violates the assumption of constant residual variance.
2. The normal Q-Q plot shows that the residuals do not follow a normal distribution.

Given these concerns, we will not interpret the parameter estimates in this model. Instead, we will run a Box-Cox analysis to find a power transformation that may address these issues with the residuals.

library(MASS)  
boxcox (fit1)



The Box-Cox Analysis indicates that an optimal transformation would have a power just above zero. Since 95% confident interval for the optimal lambda λ is close to 0.5, so the best transformation is Y^0.5.

## Model 2 - Square root Shots on target vs. all predictors

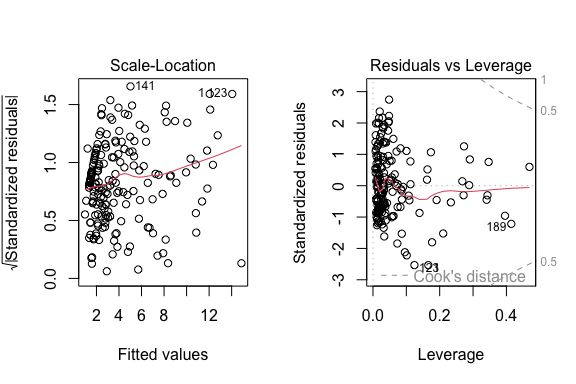
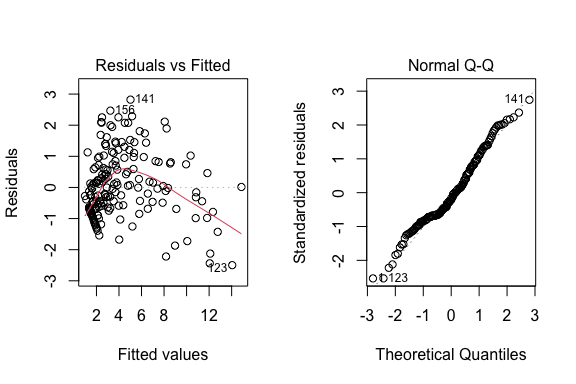
fit2 = lm ((y.offset)^0.5 ~ as.factor(Position) + Age + `Hit woodwork` + `Big chances missed` + Interceptions + Recoveries + `Aerial battles won` + `Assists` + `Passes per match`, data=shotsOnTarget)  
  
summary (fit2)

##   
## Call:  
## lm(formula = (y.offset)^0.5 ~ as.factor(Position) + Age + `Hit woodwork` +   
## `Big chances missed` + Interceptions + Recoveries + `Aerial battles won` +   
## Assists + `Passes per match`, data = shotsOnTarget)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.4974 -0.7120 -0.2018 0.6599 2.8212   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.1087123 0.8118101 0.134 0.8936   
## as.factor(Position)Forward 0.0512004 0.8393818 0.061 0.9514   
## as.factor(Position)Midfielder 0.5489023 0.6370673 0.862 0.3900   
## Age 0.0380742 0.0264241 1.441 0.1513   
## `Hit woodwork` 0.2025319 0.0392086 5.165 6.17e-07 \*\*\*  
## `Big chances missed` 0.1856123 0.0196462 9.448 < 2e-16 \*\*\*  
## Interceptions -0.0054564 0.0028709 -1.901 0.0589 .   
## Recoveries 0.0034413 0.0007871 4.372 2.05e-05 \*\*\*  
## `Aerial battles won` -0.0001251 0.0013508 -0.093 0.9263   
## Assists 0.0250359 0.0122975 2.036 0.0432 \*   
## `Passes per match` 0.0058902 0.0052643 1.119 0.2646   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.054 on 185 degrees of freedom  
## Multiple R-squared: 0.8944, Adjusted R-squared: 0.8887   
## F-statistic: 156.6 on 10 and 185 DF, p-value: < 2.2e-16

This model explains 88.87% of the variation in Square-root(Shots on target). The residual standard error is 1.054 Square-root(Shots on target), which is 1.110916 (shots on target per player).

### Residual Analysis - Second-order model

par (mfrow=c(1,2))  
plot (fit2)



The residuals for model 2 look much better compared to model 1. No obvious issues are seen in any of the plots. Next we will interpret the parameters for model 2.

##Qualitative interpretations of estimates

confint (fit2)

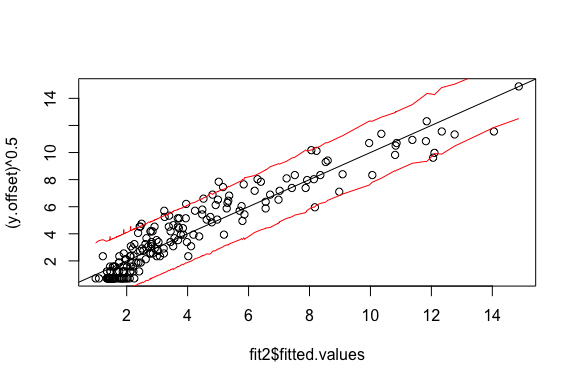
## 2.5 % 97.5 %  
## (Intercept) -1.4928835442 1.710308060  
## as.factor(Position)Forward -1.6047908212 1.707191528  
## as.factor(Position)Midfielder -0.7079486962 1.805753291  
## Age -0.0140571058 0.090205478  
## `Hit woodwork` 0.1251784226 0.279885322  
## `Big chances missed` 0.1468529102 0.224371761  
## Interceptions -0.0111203605 0.000207650  
## Recoveries 0.0018883641 0.004994213  
## `Aerial battles won` -0.0027899661 0.002539827  
## Assists 0.0007744938 0.049297297  
## `Passes per match` -0.0044955203 0.016275961

Qualitative interpretations of estimates, each one accounting for the other predictors in the model (i.e., preface each statement with, “All other things being the same, …”): - Position (whether a forward or a midfielder) does not have a significant association with the number of shots on target. - Age does not have a significant association with the number of shots on target. - The more times hitting woodwork will increase the number of shots on target. - The higher number of big chances misses will increase the number of shots on target. - Interceptions does not have a significant association with the number of shots on target. - The higher number of recoveries will increase the number of shots on target. - Aerial battles won does not have a significant association with the number of shots on target. - The higher number of Assists will increase the number of shots on target. - Passes per a match does not have a significant association with the number of shots on target.

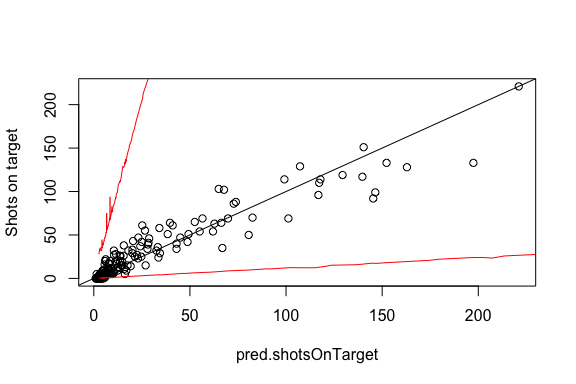
#Plot shotsOnTarget vs predicted for fit 2  
  
plot ((y.offset)^0.5 ~ fit2$fitted.values, data=shotsOnTarget)  
abline (0, 1)  
  
# Add prediction limits   
  
fit2.pred = as.data.frame (predict (fit2, interval="prediction"))

## Warning in predict.lm(fit2, interval = "prediction"): predictions on current data refer to \_future\_ responses

order2 = order (fit2.pred$fit)  
lines (fit2.pred$fit [order2], fit2.pred$lwr[order2], col='red')  
lines (fit2.pred$fit [order2], fit2.pred$upr[order2], col='red')



# Back-transform fitted Shots on target and plot  
  
pred.shotsOnTarget = (fit2$fitted.values)^2  
plot (`Shots on target` ~ pred.shotsOnTarget, data=shotsOnTarget)  
abline (0, 1)  
  
# Add prediction limits to previous plot  
  
lines (exp (fit2.pred$fit [order2]), exp (fit2.pred$lwr[order2]), col='red')  
lines (exp (fit2.pred$fit [order2]), exp (fit2.pred$upr[order2]), col='red')



The plot above shows the observed number of Shots on Target vs the fitted Shots on Target, back-transformed from Model 2 to the original scale (number of shots on target). Note that on the original scale of shots on target, predicted number of shot on targets are less precise.

### Stepwise Regression:

Next, we apply step-wise regression to the full square root (Shots on Targets) model using the AIC criterion.

fit2aic = step (fit2, direction='both')

## Start: AIC=31.2  
## (y.offset)^0.5 ~ as.factor(Position) + Age + `Hit woodwork` +   
## `Big chances missed` + Interceptions + Recoveries + `Aerial battles won` +   
## Assists + `Passes per match`  
##   
## Df Sum of Sq RSS AIC  
## - as.factor(Position) 2 1.712 207.13 28.829  
## - `Aerial battles won` 1 0.010 205.43 29.212  
## - `Passes per match` 1 1.390 206.81 30.525  
## <none> 205.42 31.203  
## - Age 1 2.305 207.73 31.390  
## - Interceptions 1 4.011 209.43 32.993  
## - Assists 1 4.602 210.03 33.546  
## - Recoveries 1 21.223 226.65 48.474  
## - `Hit woodwork` 1 29.628 235.05 55.610  
## - `Big chances missed` 1 99.113 304.54 106.372  
##   
## Step: AIC=28.83  
## (y.offset)^0.5 ~ Age + `Hit woodwork` + `Big chances missed` +   
## Interceptions + Recoveries + `Aerial battles won` + Assists +   
## `Passes per match`  
##   
## Df Sum of Sq RSS AIC  
## - `Aerial battles won` 1 0.052 207.19 26.879  
## - `Passes per match` 1 1.138 208.27 27.903  
## <none> 207.13 28.829  
## - Age 1 2.881 210.01 29.537  
## - Interceptions 1 4.409 211.54 30.957  
## + as.factor(Position) 2 1.712 205.42 31.203  
## - Assists 1 4.679 211.81 31.207  
## - Recoveries 1 23.136 230.27 47.583  
## - `Hit woodwork` 1 29.468 236.60 52.900  
## - `Big chances missed` 1 98.430 305.56 103.033  
##   
## Step: AIC=26.88  
## (y.offset)^0.5 ~ Age + `Hit woodwork` + `Big chances missed` +   
## Interceptions + Recoveries + Assists + `Passes per match`  
##   
## Df Sum of Sq RSS AIC  
## - `Passes per match` 1 1.104 208.29 25.920  
## <none> 207.19 26.879  
## - Age 1 2.873 210.06 27.578  
## + `Aerial battles won` 1 0.052 207.13 28.829  
## + as.factor(Position) 2 1.754 205.43 29.212  
## - Assists 1 5.233 212.42 29.768  
## - Interceptions 1 5.319 212.51 29.847  
## - Recoveries 1 23.172 230.36 45.658  
## - `Hit woodwork` 1 29.452 236.64 50.930  
## - `Big chances missed` 1 105.988 313.17 105.854  
##   
## Step: AIC=25.92  
## (y.offset)^0.5 ~ Age + `Hit woodwork` + `Big chances missed` +   
## Interceptions + Recoveries + Assists  
##   
## Df Sum of Sq RSS AIC  
## <none> 208.29 25.920  
## + `Passes per match` 1 1.104 207.19 26.879  
## + `Aerial battles won` 1 0.018 208.27 27.903  
## - Assists 1 4.609 212.90 28.210  
## + as.factor(Position) 2 1.477 206.81 28.525  
## - Age 1 5.177 213.47 28.732  
## - Interceptions 1 6.676 214.97 30.104  
## - Recoveries 1 28.704 236.99 49.225  
## - `Hit woodwork` 1 29.540 237.83 49.914  
## - `Big chances missed` 1 104.970 313.26 103.908

Position, number of aerial battles won, number of passes per match were removed from the model. Some of the remaining predictors are only marginally significant.

summary (fit2aic)

##   
## Call:  
## lm(formula = (y.offset)^0.5 ~ Age + `Hit woodwork` + `Big chances missed` +   
## Interceptions + Recoveries + Assists, data = shotsOnTarget)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.5021 -0.7664 -0.1608 0.6380 2.8362   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.4327397 0.5732191 0.755 0.4512   
## Age 0.0520376 0.0240103 2.167 0.0315 \*   
## `Hit woodwork` 0.2018266 0.0389834 5.177 5.73e-07 \*\*\*  
## `Big chances missed` 0.1806190 0.0185069 9.760 < 2e-16 \*\*\*  
## Interceptions -0.0064771 0.0026315 -2.461 0.0147 \*   
## Recoveries 0.0037951 0.0007436 5.104 8.09e-07 \*\*\*  
## Assists 0.0238731 0.0116737 2.045 0.0422 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.05 on 189 degrees of freedom  
## Multiple R-squared: 0.8929, Adjusted R-squared: 0.8895   
## F-statistic: 262.6 on 6 and 189 DF, p-value: < 2.2e-16

The AIC stepwise model has an adjusted r-squared of 0.8895 and a residual standard error of 1.05.

### Model 2 with Interactions

Next, we well add all possible two-way interaction effects to the AIC step-wise regression model obtained above.

mycenter = function (x) x - mean (x)  
  
shotsOnTarget$Age.c = mycenter (shotsOnTarget$Age)  
shotsOnTarget$Hit\_woodwork.c = mycenter (shotsOnTarget$`Hit woodwork`)  
shotsOnTarget$Big\_chances\_missed.c = mycenter (shotsOnTarget$`Big chances missed`)  
shotsOnTarget$Interceptions.c = mycenter (shotsOnTarget$Interceptions)  
shotsOnTarget$Recoveries.c = mycenter(shotsOnTarget$Recoveries)  
shotsOnTarget$Assists.c = mycenter(shotsOnTarget$Assists)  
  
fit2aic.int = lm ((y.offset)^0.5 ~ (Age.c + Hit\_woodwork.c + Big\_chances\_missed.c + Interceptions.c + Recoveries.c + Assists.c)^2, data=shotsOnTarget)  
  
summary (fit2aic.int)

##   
## Call:  
## lm(formula = (y.offset)^0.5 ~ (Age.c + Hit\_woodwork.c + Big\_chances\_missed.c +   
## Interceptions.c + Recoveries.c + Assists.c)^2, data = shotsOnTarget)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.9360 -0.4693 -0.1754 0.4541 2.3244   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.662e+00 9.549e-02 48.823 < 2e-16 \*\*\*  
## Age.c -2.515e-03 2.184e-02 -0.115 0.908437   
## Hit\_woodwork.c 1.064e-01 7.073e-02 1.504 0.134387   
## Big\_chances\_missed.c 1.962e-01 2.564e-02 7.651 1.30e-12 \*\*\*  
## Interceptions.c -1.246e-02 4.117e-03 -3.026 0.002852 \*\*   
## Recoveries.c 5.283e-03 1.054e-03 5.014 1.31e-06 \*\*\*  
## Assists.c 1.399e-01 2.252e-02 6.212 3.75e-09 \*\*\*  
## Age.c:Hit\_woodwork.c 3.902e-03 2.120e-02 0.184 0.854143   
## Age.c:Big\_chances\_missed.c 1.679e-02 7.478e-03 2.245 0.026011 \*   
## Age.c:Interceptions.c 3.665e-03 1.080e-03 3.392 0.000858 \*\*\*  
## Age.c:Recoveries.c -9.998e-04 2.905e-04 -3.442 0.000723 \*\*\*  
## Age.c:Assists.c -5.864e-03 4.745e-03 -1.236 0.218201   
## Hit\_woodwork.c:Big\_chances\_missed.c -7.136e-03 3.725e-03 -1.916 0.057065 .   
## Hit\_woodwork.c:Interceptions.c 1.223e-03 7.521e-04 1.626 0.105798   
## Hit\_woodwork.c:Recoveries.c -8.572e-05 2.036e-04 -0.421 0.674207   
## Hit\_woodwork.c:Assists.c 1.806e-03 1.878e-03 0.961 0.337737   
## Big\_chances\_missed.c:Interceptions.c -1.385e-03 5.310e-04 -2.608 0.009904 \*\*   
## Big\_chances\_missed.c:Recoveries.c 1.501e-04 1.276e-04 1.176 0.241191   
## Big\_chances\_missed.c:Assists.c -2.904e-03 1.583e-03 -1.835 0.068231 .   
## Interceptions.c:Recoveries.c -2.605e-06 2.228e-06 -1.169 0.244087   
## Interceptions.c:Assists.c 7.041e-04 3.117e-04 2.259 0.025135 \*   
## Recoveries.c:Assists.c -1.937e-04 8.389e-05 -2.309 0.022097 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.773 on 174 degrees of freedom  
## Multiple R-squared: 0.9465, Adjusted R-squared: 0.9401   
## F-statistic: 146.7 on 21 and 174 DF, p-value: < 2.2e-16

The model above has a slightly higher adjusted R^2 (0.9401 vs. 0.8887) and lower residual standard error (0.773 vs. 1.054).

##Interaction model step-wise regression

A step-wise regression will be used to get a more streamlined model.

nrows = fit2aic.int$rank + fit2aic.int$df.residual  
fit2.int.aic = step (fit2aic.int, direction='both', k=log(nrows))

summary (fit2.int.aic)

##   
## Call:  
## lm(formula = (y.offset)^0.5 ~ Age.c + Hit\_woodwork.c + Big\_chances\_missed.c +   
## Interceptions.c + Recoveries.c + Assists.c + Age.c:Big\_chances\_missed.c +   
## Age.c:Interceptions.c + Age.c:Recoveries.c + Hit\_woodwork.c:Big\_chances\_missed.c +   
## Hit\_woodwork.c:Interceptions.c + Big\_chances\_missed.c:Interceptions.c +   
## Interceptions.c:Assists.c + Recoveries.c:Assists.c, data = shotsOnTarget)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.02161 -0.43792 -0.09205 0.44816 2.23294   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.611e+00 8.463e-02 54.478 < 2e-16 \*\*\*  
## Age.c 3.150e-03 1.990e-02 0.158 0.874441   
## Hit\_woodwork.c 1.820e-01 4.505e-02 4.040 7.88e-05 \*\*\*  
## Big\_chances\_missed.c 1.977e-01 1.862e-02 10.617 < 2e-16 \*\*\*  
## Interceptions.c -1.640e-02 3.471e-03 -4.724 4.62e-06 \*\*\*  
## Recoveries.c 6.029e-03 8.868e-04 6.798 1.49e-10 \*\*\*  
## Assists.c 1.145e-01 1.515e-02 7.557 1.98e-12 \*\*\*  
## Age.c:Big\_chances\_missed.c 1.237e-02 4.192e-03 2.951 0.003585 \*\*   
## Age.c:Interceptions.c 3.419e-03 9.921e-04 3.447 0.000706 \*\*\*  
## Age.c:Recoveries.c -9.764e-04 2.560e-04 -3.813 0.000188 \*\*\*  
## Hit\_woodwork.c:Big\_chances\_missed.c -1.017e-02 2.723e-03 -3.733 0.000253 \*\*\*  
## Hit\_woodwork.c:Interceptions.c 7.781e-04 2.972e-04 2.618 0.009598 \*\*   
## Big\_chances\_missed.c:Interceptions.c -7.379e-04 1.709e-04 -4.318 2.59e-05 \*\*\*  
## Interceptions.c:Assists.c 6.819e-04 1.469e-04 4.643 6.57e-06 \*\*\*  
## Recoveries.c:Assists.c -2.194e-04 3.935e-05 -5.576 8.83e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7751 on 181 degrees of freedom  
## Multiple R-squared: 0.9441, Adjusted R-squared: 0.9398   
## F-statistic: 218.3 on 14 and 181 DF, p-value: < 2.2e-16

Eight interaction effects were retained using the AIC criterion, i.e, age vs. big chances missed and age vs. interceptions. This means that the slope between the squared root shots on target and age is different from different number of big chances missed and for differenct number of interceptions.

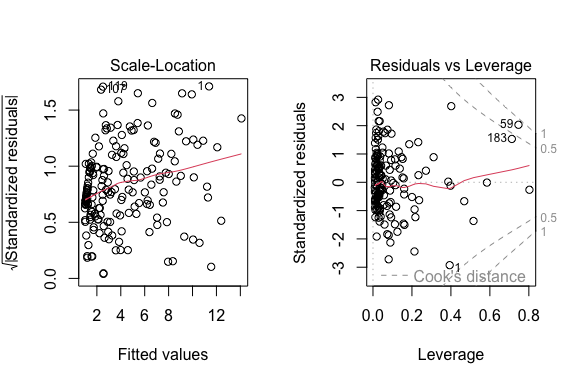
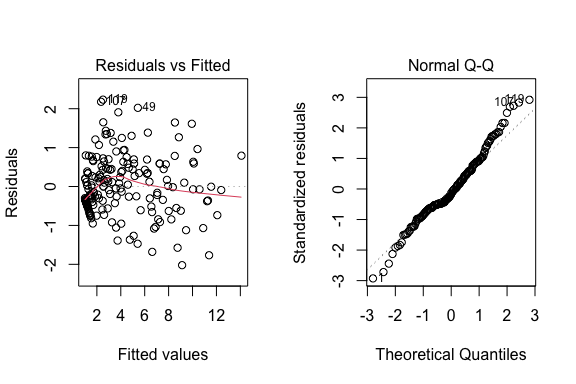
##Final Model

We will use the last model above as our final model.

### Residual and Influence Analysis

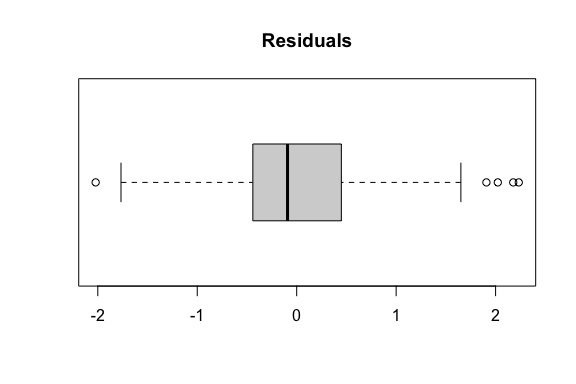
Next we will look at the residual and influence analysis for this model.

par (mfrow = c(1,2))  
plot (fit2.int.aic)



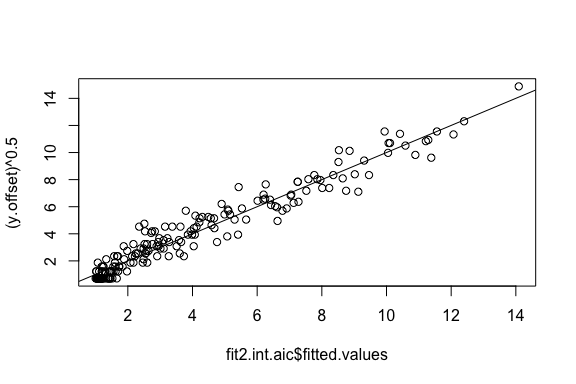
For the residual plots, the model 2 is better than model 1 in terms of constant variance, linearity and normality. No obvious issues are seen in any of the plots.

boxplot (fit2.int.aic$residuals, horizontal=T, main="Residuals")



The box plot of the residuals looks fine. The distribution is symmetric with no obvious outliers.

plot ((y.offset)^0.5 ~ fit2.int.aic$fitted.values, data=shotsOnTarget)  
abline (0, 1)



The plot of square-root (Shots on Target) vs.  the fitted square-root (Shots on Target) shows a good fit of the model to the data.

install.packages(“car”, dependencies=TRUE)

library(car)

## Loading required package: carData

car::vif (fit2.int.aic, type = 'predictor')

## GVIFs computed for predictors

## GVIF Df GVIF^(1/(2\*Df))  
## Age.c 84.29321 8 1.319356  
## Hit\_woodwork.c 742.16560 6 1.734636  
## Big\_chances\_missed.c 7006.35329 9 1.635454  
## Interceptions.c 23425.61864 11 1.579874  
## Recoveries.c 42495.63206 5 2.902916  
## Assists.c 224.17601 5 1.718141  
## Interacts With  
## Age.c Big\_chances\_missed.c, Interceptions.c, Recoveries.c  
## Hit\_woodwork.c Big\_chances\_missed.c, Interceptions.c  
## Big\_chances\_missed.c Age.c, Hit\_woodwork.c, Interceptions.c  
## Interceptions.c Age.c, Hit\_woodwork.c, Big\_chances\_missed.c, Assists.c  
## Recoveries.c Age.c, Assists.c  
## Assists.c Interceptions.c, Recoveries.c  
## Other Predictors  
## Age.c Hit\_woodwork.c, Assists.c  
## Hit\_woodwork.c Age.c, Recoveries.c, Assists.c  
## Big\_chances\_missed.c Recoveries.c, Assists.c  
## Interceptions.c Recoveries.c  
## Recoveries.c Hit\_woodwork.c, Big\_chances\_missed.c, Interceptions.c  
## Assists.c Age.c, Hit\_woodwork.c, Big\_chances\_missed.c

All of the VIF values are less than 5, which is good for the collinearity assessment.

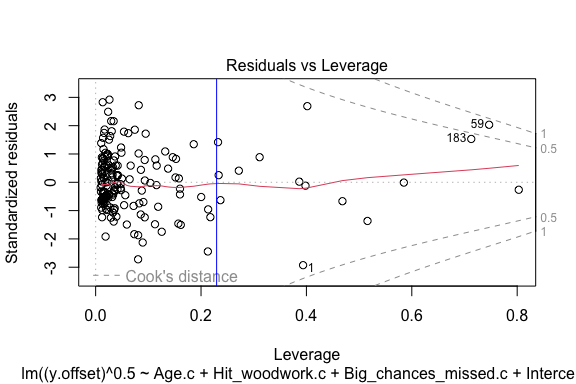
Leverage cutoff:

(levg.cutoff = 3\*fit2.int.aic$rank / nrows)

## [1] 0.2295918

The leverage cutoff is 0.2295918.

plot (fit2.int.aic, which=5)  
abline (v=levg.cutoff, col='blue')



hatvals = hatvalues (fit2.int.aic) > levg.cutoff  
sum (hatvals)

## [1] 15

sum (hatvals) / nrows

## [1] 0.07653061

There are 15 rows (4.9% of the sample size) with leverage values above the cutoff. We expect about 5%.

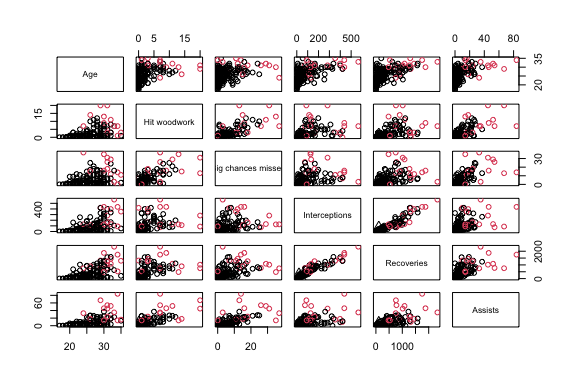
summary (cooks.distance(fit2.int.aic))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000000 0.0002868 0.0005907 0.0149279 0.0040491 0.8115658

There are no Cook’s Distance values above the cutoff of 0.5.&&&% #xem laiiii

Scatterplot matrix of variables in the final model:

shotsOnTarget$y.square\_root = (shotsOnTarget$y.offset)^0.5  
plot (shotsOnTarget [,c("Age", "Hit woodwork", "Big chances missed", "Interceptions", "Recoveries", "Assists")],   
 col= (hatvals > levg.cutoff) + 1, cex=ifelse (1:nrows==11, 2.5, 1),   
 pch=ifelse (1:nrows==11, 2, 1))



The plot above shows that the high leverage values (red points) do not mostlt appear around the edges of the two-dimensional scatter plots. This explains our small and positive residual.

No further remedial measures are needed.

### Interaction Plot:

Next, we plot the interaction effects.

par (mfrow=c(1,1))  
# Function to categorize a continuous variable into its quartiles  
categorize = function (x) {  
 quartiles = summary (x) [c(2, 3, 5)]  
 result = rep ("Q1", length (x))  
 result [(quartiles[1] < x) & (x <= quartiles [2])] = "Q2"  
 result [(quartiles[2] < x) & (x <= quartiles [3])] = "Q3"  
 result [quartiles[3] < x] = "Q4"  
 return (result)  
}  
  
# Interaction plots using the ggplot and dplyr packages  
  
library (ggplot2)  
library (dplyr)

##   
## Attaching package: 'dplyr'

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

## The following object is masked from 'package:MASS':  
##   
## select

## The following objects are masked from 'package:data.table':  
##   
## between, first, last

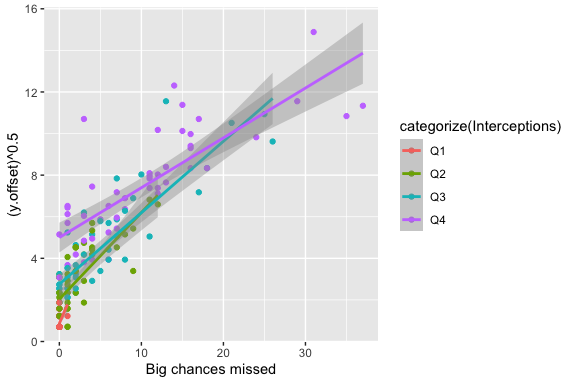
## The following objects are masked from 'package:stats':  
##   
## filter, lag

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

categorize = function (x) {  
 quartiles = summary (x) [c(2, 3, 5)]  
 result = rep ("Q1", length (x))  
 result [(quartiles[1] < x) & (x <= quartiles [2])] = "Q2"  
 result [(quartiles[2] < x) & (x <= quartiles [3])] = "Q3"  
 result [quartiles[3] < x] = "Q4"  
 return (result)  
}

# Plot square root of Shots on Target vs Big chances missed in accordance with Interceptions  
# Note the use of the "with" function to avoid having to specify  
# each column name with "shotsOnTarget$" in front of it  
  
with (shotsOnTarget,  
 qplot (`Big chances missed`, y=(y.offset)^0.5, color = categorize(Interceptions)) + geom\_smooth (method="lm"))

## `geom\_smooth()` using formula 'y ~ x'

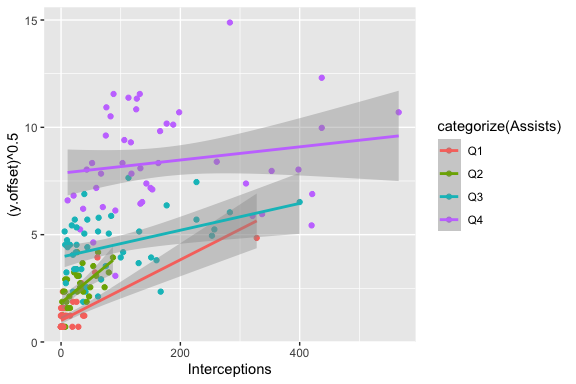


< Sth here >

Big\_chances\_missed.c:Interceptions.c 2.59e-05  ***Interceptions.c:Assists.c 6.57e-06***  Recoveries.c:Assists.c 8.83e-08 \*\*\*

# Plot square root of Shots on Target vs Interceptions in accordance with Assists  
# Note the use of the "with" function to avoid having to specify  
# each column name with "shotsOnTarget$" in front of it  
  
with (shotsOnTarget,  
 qplot (Interceptions, y=(y.offset)^0.5, color = categorize(Assists)) + geom\_smooth (method="lm"))

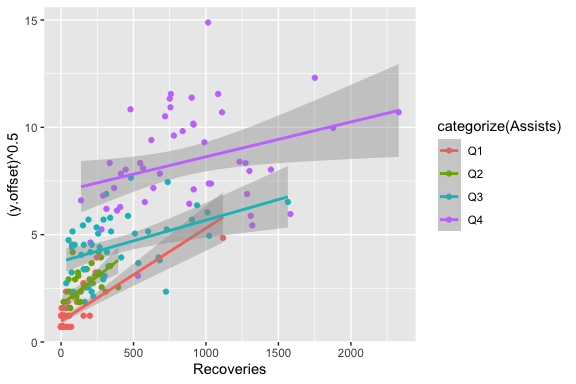
## `geom\_smooth()` using formula 'y ~ x'



The interaction plot above shows that the relationship between square-root(shots on targets) and Interception is stronger (steeper) for lower number of assists than it is for higher number of assists.%%%%% xem lai

# Plot square root of Shots on Target vs Recoveries in accordance with Assists  
# Note the use of the "with" function to avoid having to specify  
# each column name with "shotsOnTarget$" in front of it  
  
with (shotsOnTarget,  
 qplot (Recoveries, y=(y.offset)^0.5, color = categorize(Assists)) + geom\_smooth (method="lm"))

## `geom\_smooth()` using formula 'y ~ x'



The interaction plot above shows that the relationship between square-root(shots on targets) and Recoveries is stronger (steeper) for lower number of assists than it is for higher number of assists.%%%%% xem lai

### Example predictions

To illustrate use of the model for prediction, we show some example soccer players and their predicted shots on target:

# save the predictions and pred. limits for every player in the data frame  
# converting back to the actual scale  
  
preds = predict (fit2.int.aic, interval='prediction')

## Warning in predict.lm(fit2.int.aic, interval = "prediction"): predictions on current data refer to \_future\_ responses

shotsOnTarget$pred.shotsOnTarget = (preds [,1])^2  
shotsOnTarget$pred.lower = (preds [,2])^2  
shotsOnTarget$pred.upper = (preds [,3])^2

shotsOnTarget [c(120, 108, 25, 164, 165, 101), c("Age", "Hit woodwork", "Big chances missed", "Interceptions", "Recoveries", "Assists", "Shots on target", "pred.shotsOnTarget", "pred.lower", "pred.upper")]

## Age Hit woodwork Big chances missed Interceptions Recoveries Assists  
## 1: 18 0 0 0 1 0  
## 2: 17 0 0 0 4 0  
## 3: 29 1 3 37 313 15  
## 4: 20 1 9 22 156 6  
## 5: 24 7 37 127 750 33  
## 6: 34 7 14 437 1751 84  
## Shots on target pred.shotsOnTarget pred.lower pred.upper  
## 1: 1 1.03916 0.2854788 6.620771  
## 2: 1 1.06476 0.2787273 6.716861  
## 3: 38 24.02896 11.0620479 41.963239  
## 4: 11 22.64673 10.0771791 40.236920  
## 5: 128 145.73788 103.8208166 194.745892  
## 6: 151 153.74245 107.0357830 208.882700

# Count how many observations have prediction intervals that contain the  
# observed response value  
  
shotsOnTarget$in.interval = ifelse (shotsOnTarget$pred.lower <= shotsOnTarget$`Shots on target` &  
 shotsOnTarget$`Shots on target` <= shotsOnTarget$pred.upper,  
 1, 0)  
  
mean (shotsOnTarget$in.interval)

## [1] 0.7346939

Among the six example houses whose predictions are shown above, the players 18,17,34 are predicted having the number of shots on target with a high precision. On the other hand, the players 20,24 are predicted to have higher number of shots on target while the player 29 is predicted to have lower number of shots on target compared to the observed data. However, the predicted numbers of shots on target for players 20,24 and 29 lie in 95% confidence interval of our predictions; therefore, there is no obvious concern about this difference.

Among all of the players in the data set, 73.5% have prediction intervals that contain the observed shots on target. We would expect 95%.

# Conclusions

The final model has the adjusted R^2 = 0.9398, which means that 94% of the variation in square root shots on target is explained by the model. The residual standard error is 0.7751 square root shots on target. Based on our residual analysis, this model is consistent with the conditions for doing linear regression.

Because of the two interaction effects, we cannot numerically interpret the slope estimates for Shots on Target vs Interceptions, vs Recoveries. Instead, we note from the interaction plots above that the relationship between square-root(shots on targets) and Interception is stronger (steeper) for lower number of assists than it is for higher number of assists. And the same patterns registered from the relationship between square-root(shots on targets) and Recoveries.