**FIFA 18: Predicting Player’s Wages**

Ahmad, Beka, and Rafael

11/13/2017

## Introduction

In the past two decades, escalating payouts from domestic and international television agreements, kit deals, sponsors, Champions League, and an influx of wealthy ownership groups have resulted in surging soccer wages around the world. With money being a dominating factor in soccer player’s transfers, new contracts and resigning clauses, the ability to evaluate the factors involved in such transactions is indeed valuable to the two parties involved. The top 20 highest-paid footballers collectively made 413 million in salary and bonus and another 162 million in endorsements for total earnings of 575 million. English Premier League players number the most followed by La Liga players. Among the most popular soccer leagues in the world, these clubs also have the top two highest average salaries per first team player. Per Sporting Intelligence’s annual Global Sports Salaries Survey 2016, EPL salaries average 3.2 million and La Liga salaries average $1.6 million. The chart below shows the average salaries in the world’s major football leagues.



This report attempts to analyze what attributes are most predictive of player’s wages, by generating simple and multiple linear regression models on FIFA 18 data obtained from Kaggle.com. Multiple variables are available to evaluation to find the best model to achieve our goal stated above. Based on that, we chose to examine Age, Nationality, Club, Overall rating, Potential Overall, Special, Value (Euros), Reactions and Preferred Positions. These are the most significant variables in predicting player’s wages in FIFA18. There are around 1000 observations in the sample data set.

## Data Cleaning

As part of cleaning up the data, we started by deleting all the columns except for ID, Name, Age, Nationality, Overall, Potential, Club, Value (in Euros), Wage (in Euros), Special, Reactions and Preferred Positions. We also deleted any rows that had missing values. We decided to look at the athletic variables that are the heavier influencers on the Overall and Potential variables’ values for the players, leading us to conclude that Reactions has the biggest influence on the Overall and Potential variables, thus we decided to maintain Reactions and delete the rest of the player attributes. Following, we converted the Value and Wages variables which ended either in “M” and “K” for example, 95M and 150K to 95,000,000 and 150,000. In the fifa18.csv file, there are players who have empty cells for Wage although are in a club, which might be a case of outliers. We came to a consensus to leave those in. Our group decided to delete players who have empty cell values for clubs as it would not be helpful in our analysis where we are looking at the relationship between wages and clubs. Their value will be zero and will lead to outliers with no valuable explanation. Furthermore, we decided to combine the 27 positions (represented by variables e.g. CAM, CM, RW, LW) into six different positions: Striker, Winger, Att-Mid, Def-Mid, Wingback, Center-Back and GK- each of these variables represent a group of variables (e.g. Winger= RW and LW) by taking the mean of those variables that we want to combine (Table 1). For the categorical variable “Preferred Positions”, we excluded every value (in case of two or more preferred positions) except the first one. We also switched the remaining preferred position to the corresponding one of the six summarized positions. We then randomized the rows and selected the first one thousand rows. We left GKs in to prove they are explainable outliers. They are fundamentally different than other players on the field.

### Table 1

* Striker -> RS, LS, ST, CF, LF, and RF
* Winger -> LW and RW
* AttMid -> CAM, LAM, RAM, RCM, LCM, LM, RM, and CM
* DefMid -> CDM, LDM, and RDM
* CenterBack -> CB, LCB, and RCB
* Wingback -> LWB, RWB, LB, and RB
* GK -> GK

### 

# Model Selection and Interpretation.

To begin our study, we constructed a series of multiple regression models, looking at each group of variables independently. For each of these models, we test to see which variables are significant in their own categories. Based on the result, Potential, Reactions, GKs, Winger and AttMid hold significance. The poor statistical results obtained through such model served as means for exploring the data, with a goal of improving our model to the best of our ability.

##Residual standard error: 21340 on 956 degrees of freedom  
## Multiple R-squared: 0.3539, Adjusted R-squared: 0.3417   
## F-statistic: 29.09 on 18 and 956 DF, p-value: < 2.2e-16

Fig.1

The residual analysis of the first order model as shown in Fig.2, suggested some problems with the model as well. The residuals are not symmetrically distributed showing non-constant variance. The presence of outliers might also be skewing the model. The Residuals vs Fitted plot suggests “Heteroscedasticity”. In other words, as the residuals get larger as predictions moves from small to large. The presence of a curve pattern in the Residuals vs Fitted also suggests room for improvement in the model.

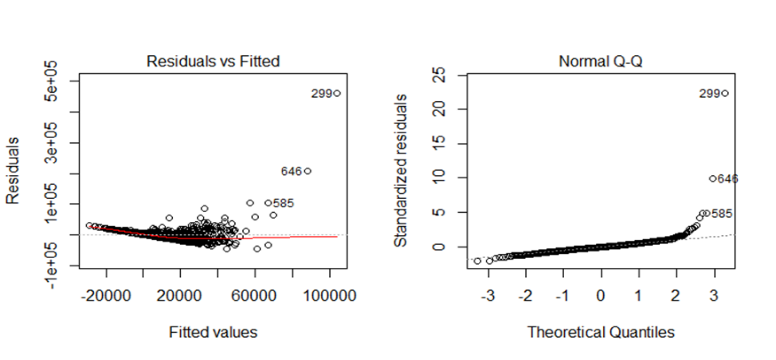
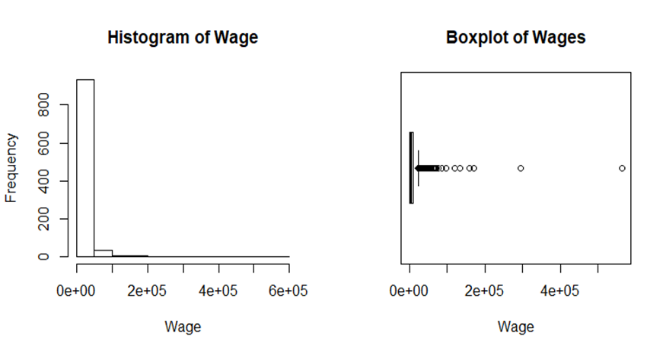


Fig. 2

We next turned to more inclusive and in-depth multiple linear regression models, combining predictors variables from all factor groups to form the models. First, we construct a model using the log of our response variable since its distribution is extremely right skewed, suggesting the need for a log transformation.



The result of such model showed some great improvement when compared to the first order model. Note here that the residual standard error is much smaller due to the fact that is in the log scale.

## Residual standard error: 0.309 on 956 degrees of freedom  
## Multiple R-squared: 0.6788, Adjusted R-squared: 0.6728   
## F-statistic: 112.3 on 18 and 956 DF, p-value: < 2.2e-16

Fig.3

Next, we decided to group players by areas of the field, as we observed that different positions could be masking each other in terms of statistical significance. Figure 4 displays a pair plot of the now, three different areas of the field and their correlation amongst each other. It is a good point to stop and analyze figure 4. We can see that there is a stronger correlation between attack/midfld and back/midfld but a weaker correlation between attack and back. The latter roles require completely different set of skills than each of them when compared to midfld.

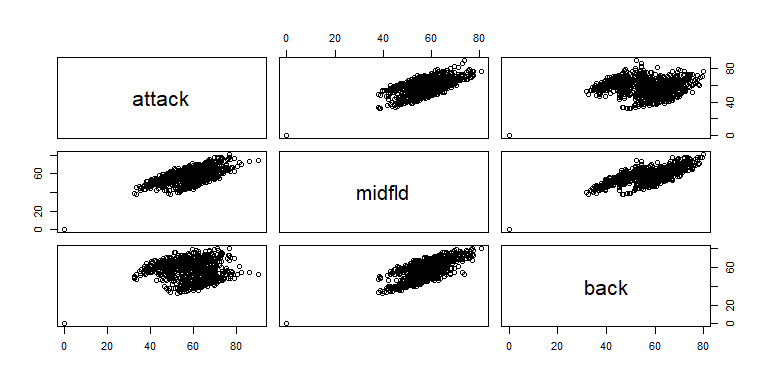


Fig.4

The residual analysis of the log transformed model shows some great improvements from the first order model as displayed below in Figure 5. The normal Q-Q plot resembles the normal Q-Q plot from the first order model although the residual vs fitted plot suggests some significant improvement. Note here the diagonal straight lines in the first plot in Figure 5. These lines indicate spreads of players that receive the same salary. We will come back to that later.

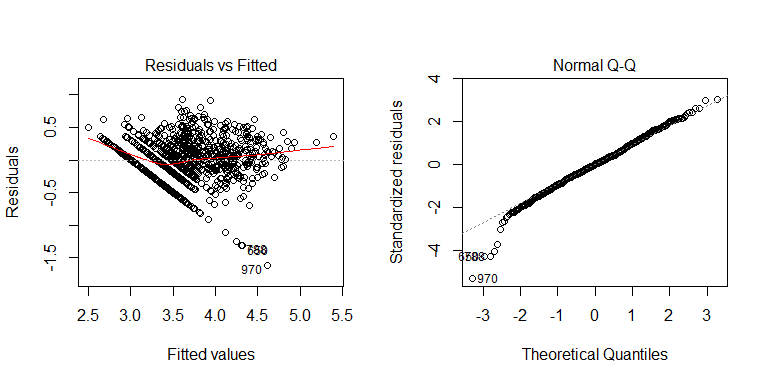


Fig.5

Next, we fit a model using a lambda transformation on the response variable raised to the power of -0.2. Figure six shown below displays some of the statistical results obtained. Based on these results we decided to continue with the log transformed model.

## Residual standard error: 0.02631 on 956 degrees of freedom  
## Multiple R-squared: 0.6571, Adjusted R-squared: 0.6506   
## F-statistic: 101.8 on 18 and 956 DF, p-value: < 2.2e-16

Fig.6

Next, we eliminated Goal Keepers from our data set, given the fact that Goal Keepers have their skill level = 0 and were skewing the data, as shown in each positions histogram shown in our Exploratory Analysis report. We have not attached the histograms here due to a need for a large raw text space.

Our fourth model is a combination of fitting the log transformed model with the positions grouped by areas of the field. The models variables and results are shown in Figure 7.

## Call:  
## lm(formula = logWage ~ Age + Overall + Potential + logValue +   
## Special + Reactions + Preferred.Positions + attack + midfld +   
## back, data = as.data.frame(fifa2))

## Residual standard error: 0.31 on 871 degrees of freedom  
## Multiple R-squared: 0.6778, Adjusted R-squared: 0.6726   
## F-statistic: 130.9 on 14 and 871 DF, p-value: < 2.2e-16

Fig.7

The results are indeed like our log transformed model without positions grouped by areas of the field. Therefore, we decided to continue with the model described in Figure 7. Although the analysis of the ANOVA table suggests that positions are not significant anymore after they have been grouped by respective areas of the field. Those results can be visualized in Figure 8.

## Analysis of Variance Table  
##   
## Response: logWage  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Age 1 21.337 21.337 221.9951 < 2.2e-16 \*\*\*  
## Overall 1 150.935 150.935 1570.3793 < 2.2e-16 \*\*\*  
## Potential 1 1.200 1.200 12.4856 0.0004317 \*\*\*  
## logValue 1 2.335 2.335 24.2890 9.924e-07 \*\*\*  
## Special 1 0.009 0.009 0.0952 0.7577185   
## Reactions 1 0.034 0.034 0.3523 0.5529502   
## Preferred.Positions 5 0.119 0.024 0.2470 0.9413687   
## attack 1 0.094 0.094 0.9810 0.3222246   
## midfld 1 0.027 0.027 0.2762 0.5993612   
## back 1 0.018 0.018 0.1895 0.6634572   
## Residuals 871 83.715 0.096   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Fig.8

Our next task was to produce a model doing a stepwise backward elimination using both directions using model 4 as described in Figure 7. Based on the AIC criterion we obtained the model as proposed in Figure 9.

## Step: AIC=-2058.96  
## logWage ~ Age + Potential + logValue

## Residual standard error: 0.3088 on 882 degrees of freedom  
## Multiple R-squared: 0.6763, Adjusted R-squared: 0.6752   
## F-statistic: 614.1 on 3 and 882 DF, p-value: < 2.2e-16

Fig.9

Next, we fit a model the model above using the interactions between each of the variables in pairs. Figure 10 displays the model and results. The model showed some minimal improvement. So far, this our best model.

## Call:  
## lm(formula = logWage ~ Age.c + Potential.c + logValue.c + Age.c \*   
## Potential.c + Age.c \* logValue.c + Potential.c \* logValue.c)

## Residual standard error: 0.3056 on 879 degrees of freedom  
## Multiple R-squared: 0.684, Adjusted R-squared: 0.6818   
## F-statistic: 317.1 on 6 and 879 DF, p-value: < 2.2e-16

Fig.10

Analyzing the ANOVA table for the model above, we noticed a weaker significance in two variables, which we experimentally decided to remove to compare results with the model above. Figure 11 displays the results of the experiment. We notice a decrease in R-square and adjusted R-squared and increased in residual standard error.

## Residual standard error: 0.3067 on 881 degrees of freedom  
## Multiple R-squared: 0.6811, Adjusted R-squared: 0.6797   
## F-statistic: 470.5 on 4 and 881 DF, p-value: < 2.2e-16

Fig.11

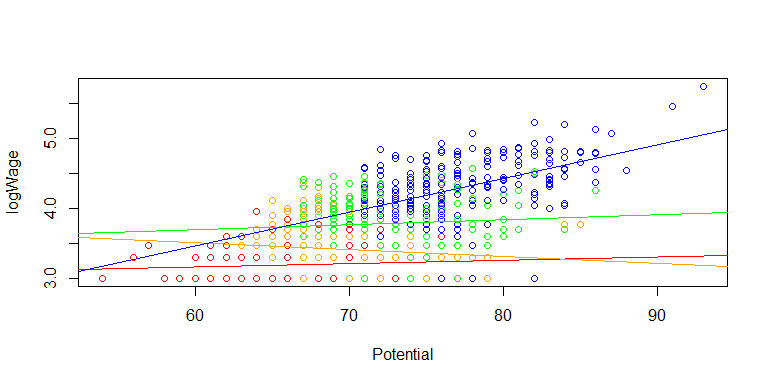
Based on that we refit the model with the highest least significant variable added back to it. Figure 12 displays the results, of what it is our best model.

## Residual standard error: 0.3056 on 880 degrees of freedom  
## Multiple R-squared: 0.6838, Adjusted R-squared: 0.682   
## F-statistic: 380.6 on 5 and 880 DF, p-value: < 2.2e-16

Fig.12

We tried a stepwise backward elimination on our best model to verify no other possible changes to it. Indeed, no further changes were needed.

With the best model established for estimating a player’s wage, we can look more in depth at the parts and ramifications of the model. We notice a spread of players that receive minimum wage in the model and some outliers that are not above the cut-off for Cook’s distance. We also notice there is a higher correlation between high value players’ wages and potential than players with medium and low values. This can be shown in Figure 13, where players are colored by value. Blue are the high value players.



**Concluding Remarks**

Establishing players’ wages is at the forefront of transactions and contracts in the soccer world. With salaries increasing worldwide, as what is being named as a “soccer bubble” problem, efforts in trying to establish the correct wage amount for each player is recommended. This model provides an opportunity to mildly obtain such Wage value. The information from this model can be somewhat helpful to help decide the correct amount a player should be payed, therefore preventing an exponential inflation of the “soccer bubble”.

To go into further detail, one could perform similar testing on a larger sample size. With the data set we used, we found what we believe to be the most descriptive model with the least amount of predictor variables.

## APPENDIX

* Variable -> Definitions
* Age -> Age of player in years
* Overall -> Overall rating of player from 0-99
* Potential -> Highest Overall rating that player can achieve from 0-99
* Value -> Estimated value of player in the market in Euros
* Wage -> Estimated current salary of player in Euros
* Special -> Sum of athletic attributes minus Composure (psychological)
* Reactions -> Estimated value of player's reaction to a play. From 0-99.
* Preferred Position -> Position on the pitch where the players performs best.
* Striker -> ST, CF, RF, LF
* Winger -> RW, LW
* AttMid -> CAM, CM, RM, LM, RAM, LAM
* DefMid -> CDM, RDM, LDM
* Centerback -> CB, RCB, LCB
* Wingback -> LWB, RWB, LB, RB

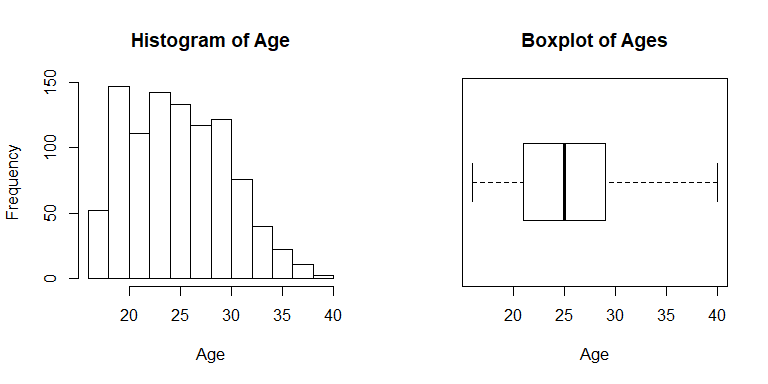
## Exploratory Analysis

We begin our investigation of by creating a histogram for every predictor variable except for Preferred.Positions as the latter is a categorical variable. Following, for player's wages by creating a histogram and boxplot of their distribution in order to gain a better overview of trends and characteristics of the data set.

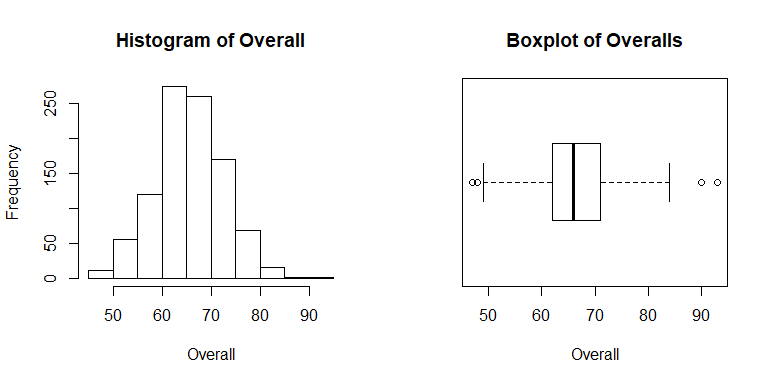
fifa <- read.csv("C:\\Users\\rafae\\Desktop\\Project 1 Stats\\fifa18.csv", header=TRUE, sep=",")  
  
attach (fifa)  
head(fifa, n=10)

## ï..ID Name Age Overall Potential Value Wage Special  
## 1 11019 J. Nolan 25 64 67 550000 2000 1749  
## 2 5715 Han Kyo Won 27 70 70 1900000 9000 1901  
## 3 7281 Cris Laranjeiros 33 68 68 575000 15000 1715  
## 4 13261 A. Hyodo 35 62 62 130000 1000 1701  
## 5 7766 S. DÃƒÂ­az 19 67 81 1600000 42000 1675  
## 6 4224 J. Vaughan 28 71 71 2400000 9000 1824  
## 7 16726 P. Pannier 18 55 68 160000 1000 1504  
## 8 6554 F. Vargas 37 69 69 325000 1000 1916  
## 9 7707 G. van Velzen 23 67 73 1000000 6000 1576  
## 10 1217 S. Long 30 76 76 7500000 85000 1897  
## Reactions Preferred.Positions Striker Winger AttMid DefMid CenterBack  
## 1 59 DefMid 59.0 62 63.000 61 57  
## 2 62 AttMid 68.5 70 67.125 61 59  
## 3 65 AttMid 66.0 64 59.625 45 40  
## 4 68 DefMid 60.0 58 60.125 58 56  
## 5 63 Striker 67.0 68 65.250 47 39  
## 6 68 Striker 69.0 67 62.375 54 55  
## 7 55 AttMid 51.0 53 54.250 49 45  
## 8 70 DefMid 65.5 65 66.875 68 69  
## 9 65 AttMid 63.5 67 60.750 42 34  
## 10 77 Striker 74.0 73 68.250 54 53  
## Wingback  
## 1 61.5  
## 2 63.5  
## 3 47.5  
## 4 55.0  
## 5 49.5  
## 6 55.5  
## 7 49.5  
## 8 64.5  
## 9 46.5  
## 10 57.0

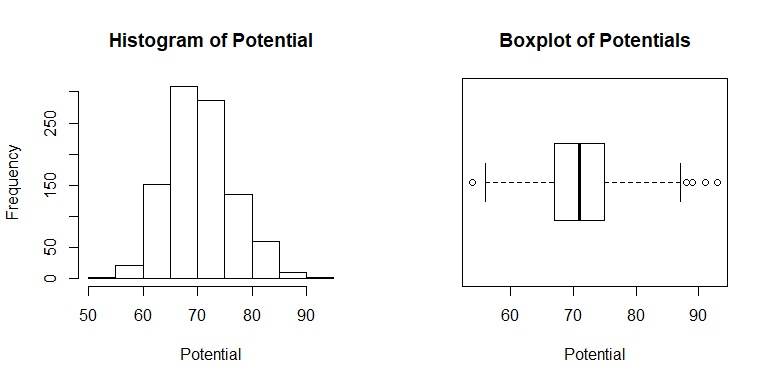
par (mfrow = c(1, 2))  
hist(Age)  
boxplot(Age, horizontal = T, xlab="Age", main="Boxplot of Ages")



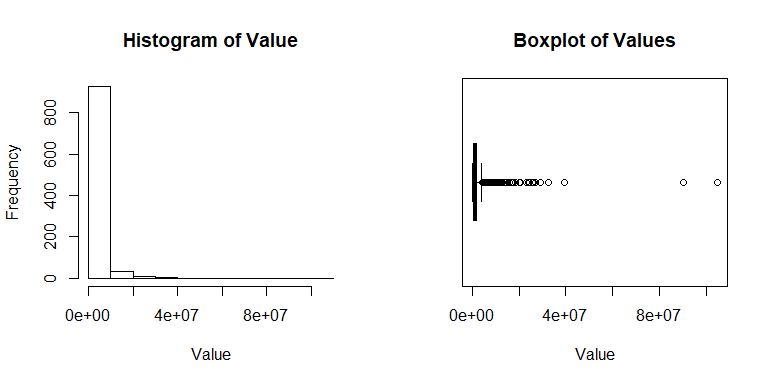
hist(Overall)  
boxplot(Overall, horizontal = T, xlab="Overall", main="Boxplot of Overalls")



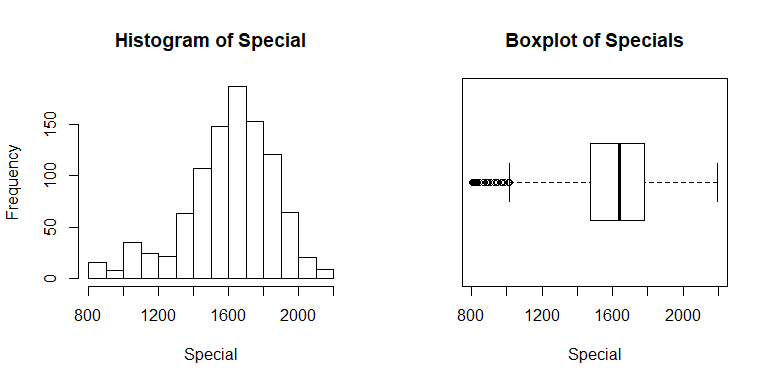
hist(Potential)  
boxplot(Potential, horizontal = T, xlab="Potential", main="Boxplot of Potentials")



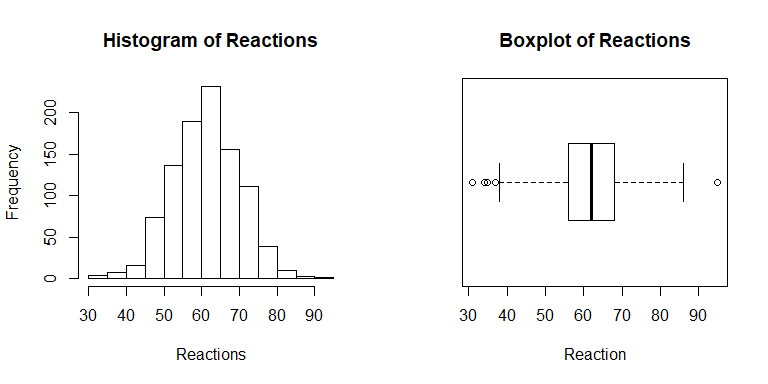
hist(Value)  
boxplot(Value, horizontal = T, xlab="Value", main="Boxplot of Values")



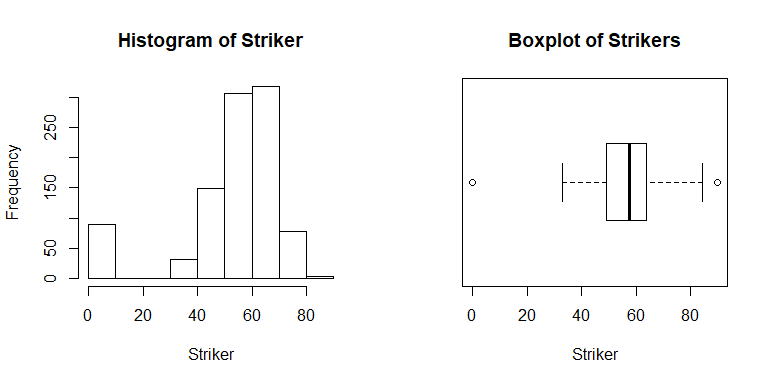
hist(Special)  
boxplot(Special, horizontal = T, xlab="Special", main="Boxplot of Specials")



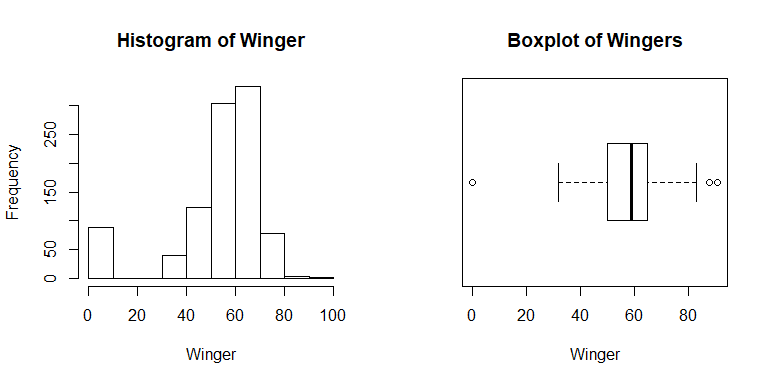
hist(Reactions)  
boxplot(Reactions, horizontal = T, xlab="Reaction", main="Boxplot of Reactions")



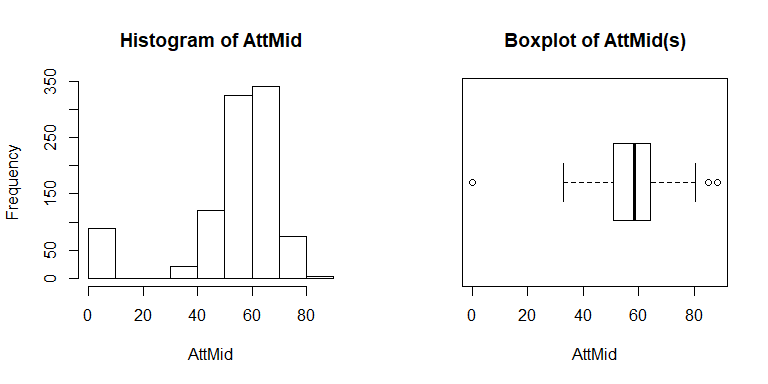
hist(Striker)  
boxplot(Striker, horizontal = T, xlab="Striker", main="Boxplot of Strikers")



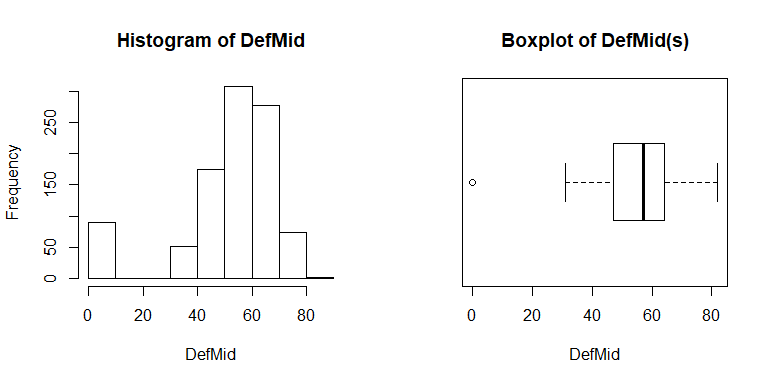
hist(Winger)  
boxplot(Winger, horizontal = T, xlab="Winger", main="Boxplot of Wingers")



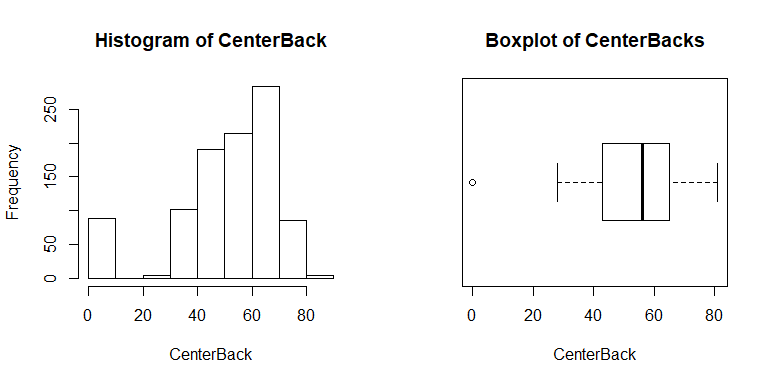
hist(AttMid)  
boxplot(AttMid, horizontal = T, xlab="AttMid", main="Boxplot of AttMid(s)")



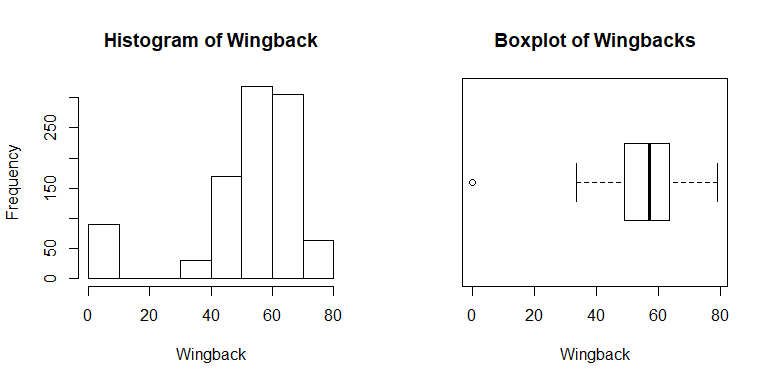
hist(DefMid)  
boxplot(DefMid, horizontal = T, xlab="DefMid", main="Boxplot of DefMid(s)")



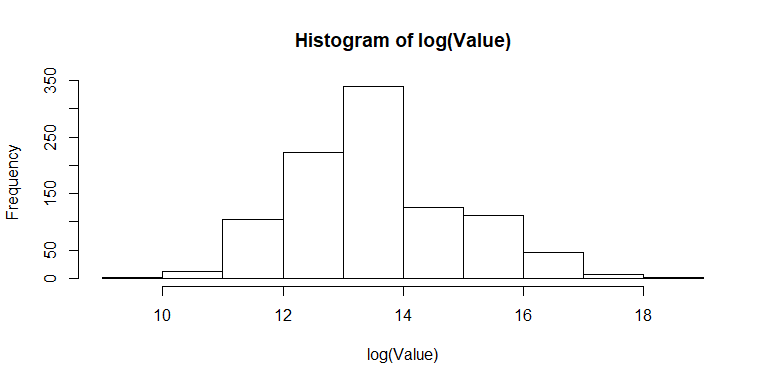
hist(CenterBack)  
boxplot(CenterBack, horizontal = T, xlab="CenterBack", main="Boxplot of CenterBacks")



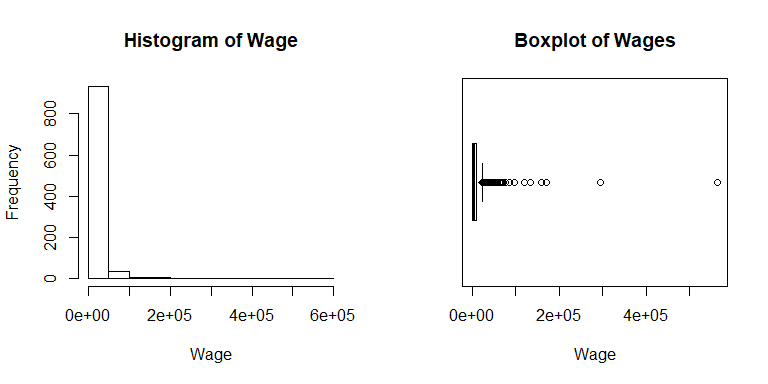
hist(Wingback)  
boxplot(Wingback, horizontal = T, xlab="Wingback", main="Boxplot of Wingbacks")



hist(log(Value))

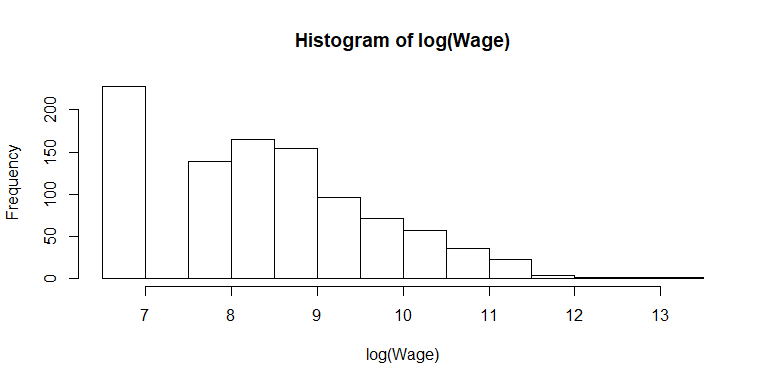


par (mfrow = c(1, 2))  
hist (Wage)  
boxplot (Wage, horizontal = T, xlab="Wage", main="Boxplot of Wages")

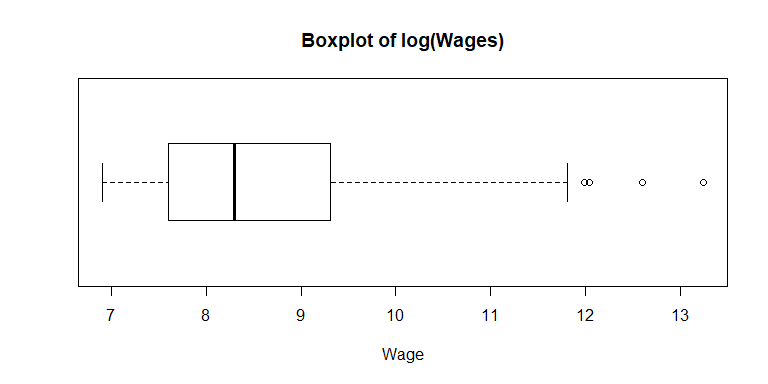


Analyzing all the distributions we believe it is suitable to do a log transformation on the Value variable, since it is extremely right skewed. This may be explained because in the soccer world there are very few players that are worthy great amounts when compared to the universe of players. The histogram below shows how a log transformation makes the distribution of Value more normal. From here on, we will proceed with log(Value).

hist(log(Wage))



boxplot (log(Wage), horizontal = T, xlab="Wage", main="Boxplot of log(Wages)")



The distribution of log(Wage) is more symmetric. There is a high frequency bin on the left of the histogram log, that may be explained by the players that have just moved up from the academy teams into the professional team and receive minimum wage. Maybe even players that played once or twice for the professional team and had to be included in this FIFA data set, but went back to the academy teams (e.g. U-21, U-19 teams). We will start by modelling wage, plot the residuals and then using Box-Cox analysis to determine the most appropriate transformation.

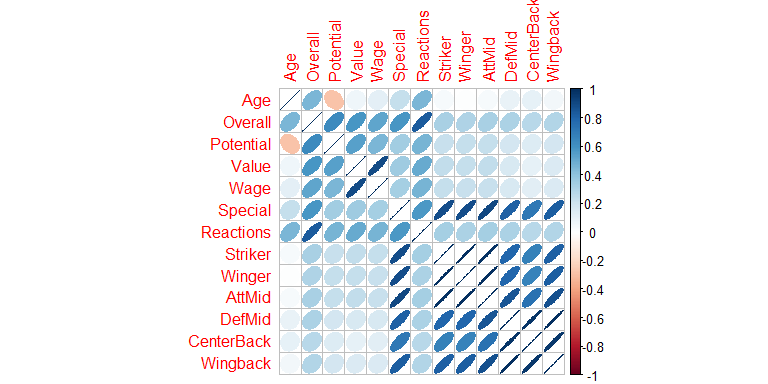
cormat = cor(fifa[,c(3,4,5,6,7,8,9,11,12,13,14,15,16)])  
round(cormat, 2)

## Age Overall Potential Value Wage Special Reactions Striker  
## Age 1.00 0.45 -0.28 0.06 0.12 0.23 0.45 0.03  
## Overall 0.45 1.00 0.63 0.58 0.53 0.59 0.84 0.32  
## Potential -0.28 0.63 1.00 0.54 0.46 0.34 0.47 0.23  
## Value 0.06 0.58 0.54 1.00 0.89 0.35 0.50 0.24  
## Wage 0.12 0.53 0.46 0.89 1.00 0.33 0.46 0.23  
## Special 0.23 0.59 0.34 0.35 0.33 1.00 0.57 0.88  
## Reactions 0.45 0.84 0.47 0.50 0.46 0.57 1.00 0.33  
## Striker 0.03 0.32 0.23 0.24 0.23 0.88 0.33 1.00  
## Winger 0.01 0.31 0.23 0.24 0.22 0.89 0.32 0.99  
## AttMid 0.03 0.33 0.23 0.24 0.23 0.90 0.34 0.99  
## DefMid 0.10 0.31 0.18 0.16 0.16 0.82 0.32 0.79  
## CenterBack 0.11 0.28 0.14 0.11 0.12 0.71 0.27 0.68  
## Wingback 0.06 0.29 0.18 0.15 0.15 0.82 0.29 0.82  
## Winger AttMid DefMid CenterBack Wingback  
## Age 0.01 0.03 0.10 0.11 0.06  
## Overall 0.31 0.33 0.31 0.28 0.29  
## Potential 0.23 0.23 0.18 0.14 0.18  
## Value 0.24 0.24 0.16 0.11 0.15  
## Wage 0.22 0.23 0.16 0.12 0.15  
## Special 0.89 0.90 0.82 0.71 0.82  
## Reactions 0.32 0.34 0.32 0.27 0.29  
## Striker 0.99 0.99 0.79 0.68 0.82  
## Winger 1.00 0.99 0.79 0.67 0.82  
## AttMid 0.99 1.00 0.86 0.74 0.88  
## DefMid 0.79 0.86 1.00 0.98 0.99  
## CenterBack 0.67 0.74 0.98 1.00 0.96  
## Wingback 0.82 0.88 0.99 0.96 1.00

library(corrplot)

## corrplot 0.84 loaded

corrplot(cormat, method = "ellipse")



The ellipse matrix plot shows only negative linear relationship between Age and Potential. In the soccer world, generally very few players have high Potential values at a low age. Exceptions are "golden boys" such as Neymar, Mmbape, Gabriel Jesus, Deli Alli, Gotze etc. These are players that have were scouted for several years prior to player professional soccer. Therefore, when entering the soccer globe they already have high potential values. On the other hand, it is more common for players to raise their potential value as they get older and prove themselves on the pitch. Based on the coefficient correlation table, there are some predictors that are extremely correlated and that may be indeed masking each other, for example Striker, Winger and AttMid, since in soccer these could be all categorized as Attacking role positions. Another example is the masking between DefMid, CenterBack and Wingback. These are all Defense role positions.

## First Order Model

Here, we fit a first-order linear model with all predictors.

logValue = log10(Value)  
  
model1 = lm(Wage~Age+Overall+Potential+logValue+Special+Reactions+Preferred.Positions+Striker+Winger+AttMid+DefMid+CenterBack+Wingback)  
  
summary(model1, correlation=FALSE)

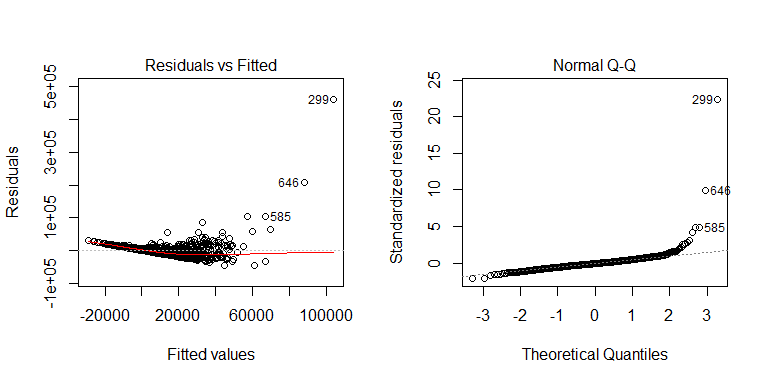
##   
## Call:  
## lm(formula = Wage ~ Age + Overall + Potential + logValue + Special +   
## Reactions + Preferred.Positions + Striker + Winger + AttMid +   
## DefMid + CenterBack + Wingback)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -43942 -8230 -1377 5953 460960   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -155191.43 14137.80 -10.977 < 2e-16 \*\*\*  
## Age 120.46 424.61 0.284 0.776709   
## Overall 909.67 673.37 1.351 0.177037   
## Potential 931.03 275.95 3.374 0.000771 \*\*\*  
## logValue 1023.78 6286.77 0.163 0.870674   
## Special -23.57 14.08 -1.674 0.094493 .   
## Reactions 298.49 151.17 1.975 0.048610 \*   
## Preferred.PositionsCenterBack 4307.22 3210.42 1.342 0.180032   
## Preferred.PositionsDefMid 1445.09 2860.45 0.505 0.613538   
## Preferred.PositionsGK 35380.51 15305.85 2.312 0.021013 \*   
## Preferred.PositionsStriker -4251.88 2931.87 -1.450 0.147324   
## Preferred.PositionsWingback 5001.60 2673.55 1.871 0.061683 .   
## Preferred.PositionsWinger 19099.42 3675.22 5.197 2.48e-07 \*\*\*  
## Striker 912.83 626.85 1.456 0.145660   
## Winger -2571.71 1025.21 -2.508 0.012290 \*   
## AttMid 3128.43 1185.06 2.640 0.008428 \*\*   
## DefMid -2032.55 1064.54 -1.909 0.056519 .   
## CenterBack 649.44 623.31 1.042 0.297705   
## Wingback 760.16 658.24 1.155 0.248442   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 21340 on 956 degrees of freedom  
## Multiple R-squared: 0.3539, Adjusted R-squared: 0.3417   
## F-statistic: 29.09 on 18 and 956 DF, p-value: < 2.2e-16

anova(model1)

## Analysis of Variance Table  
##   
## Response: Wage  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Age 1 9.3454e+09 9.3454e+09 20.5157 6.661e-06 \*\*\*  
## Overall 1 1.9262e+11 1.9262e+11 422.8479 < 2.2e-16 \*\*\*  
## Potential 1 4.3975e+09 4.3975e+09 9.6537 0.0019455 \*\*   
## logValue 1 7.6975e+08 7.6975e+08 1.6898 0.1939383   
## Special 1 3.8666e+08 3.8666e+08 0.8488 0.3571190   
## Reactions 1 1.9087e+09 1.9087e+09 4.1900 0.0409356 \*   
## Preferred.Positions 6 1.8574e+10 3.0956e+09 6.7957 4.531e-07 \*\*\*  
## Striker 1 5.9838e+09 5.9838e+09 13.1361 0.0003049 \*\*\*  
## Winger 1 5.8585e+08 5.8585e+08 1.2861 0.2570500   
## AttMid 1 2.0109e+09 2.0109e+09 4.4144 0.0358973 \*   
## DefMid 1 5.2618e+08 5.2618e+08 1.1551 0.2827548   
## CenterBack 1 7.8644e+08 7.8644e+08 1.7265 0.1891783   
## Wingback 1 6.0752e+08 6.0752e+08 1.3337 0.2484420   
## Residuals 956 4.3548e+11 4.5552e+08   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

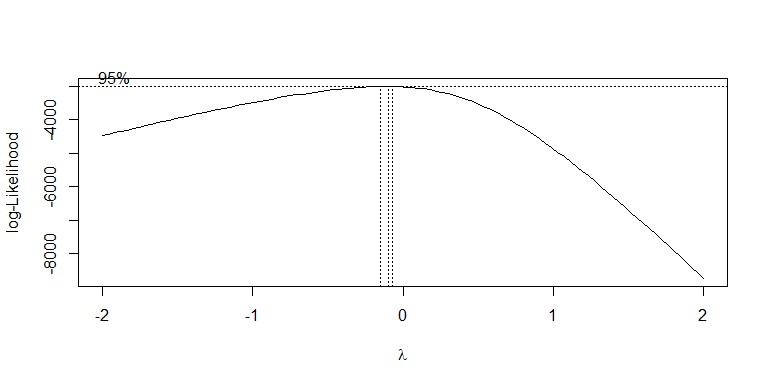
The analysis of the ANOVA table suggests that Age, Overall, Potential, Reactions, Preferred.Positions, Striker and AttMid are significant predictors of a player's Wage. The coefficients tests show that only Age, Overall and Value significant for this model. The R-squared is 0.3539, with adjusted R-squared = 0.3417, which indicates that little of the variability in Wage is being explained by this model. This is because we have not done a log transformation on the response variable. The residual standard error is 21340, which is small relative to the range of Wage values. We will analyze the residuals in order to verify the correctness of linearity of this model.

par (mfrow = c(1, 2))  
plot(model1,which = c(1,2))



Residual analysis suggests that there are outliers in this model. The residual analysis also suggest that the assumption of a linear relationship is reasonable. The Residuals vs Fitted plot suggest a non-constant variance. We will do a Box-Cox analysis to verify the necessary transformation(s).

library("MASS")  
boxcox(model1)



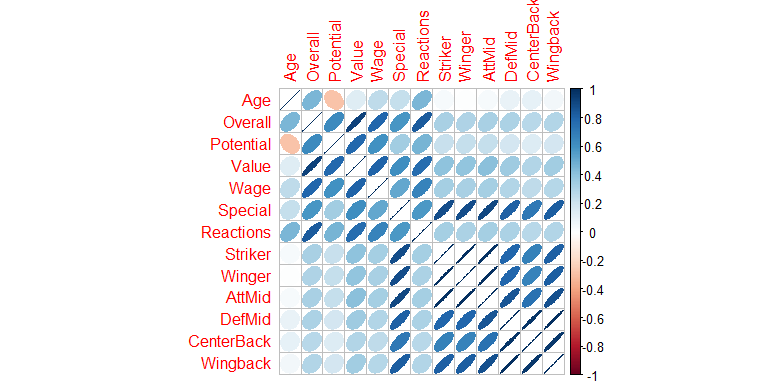
The Box-Cox analysis suggests an inverse power transformation, with λ between -0.1 and -0.2. The value zero, is just outside the 95% confidence interval, but we will try a log transformation first.

# Log-Transformed model.

logWage = log10(Wage)  
  
new\_df = fifa[,c(3,4,5,6,7,8,9,11,12,13,14,15,16)]  
new\_df[5] = logWage  
new\_df[4] = logValue  
cormat2 = cor(new\_df)  
round(cormat2, 2)

## Age Overall Potential Value Wage Special Reactions Striker  
## Age 1.00 0.45 -0.28 0.14 0.25 0.23 0.45 0.03  
## Overall 0.45 1.00 0.63 0.93 0.80 0.59 0.84 0.32  
## Potential -0.28 0.63 1.00 0.78 0.61 0.34 0.47 0.23  
## Value 0.14 0.93 0.78 1.00 0.80 0.61 0.76 0.41  
## Wage 0.25 0.80 0.61 0.80 1.00 0.52 0.67 0.34  
## Special 0.23 0.59 0.34 0.61 0.52 1.00 0.57 0.88  
## Reactions 0.45 0.84 0.47 0.76 0.67 0.57 1.00 0.33  
## Striker 0.03 0.32 0.23 0.41 0.34 0.88 0.33 1.00  
## Winger 0.01 0.31 0.23 0.40 0.32 0.89 0.32 0.99  
## AttMid 0.03 0.33 0.23 0.41 0.34 0.90 0.34 0.99  
## DefMid 0.10 0.31 0.18 0.35 0.30 0.82 0.32 0.79  
## CenterBack 0.11 0.28 0.14 0.29 0.26 0.71 0.27 0.68  
## Wingback 0.06 0.29 0.18 0.34 0.28 0.82 0.29 0.82  
## Winger AttMid DefMid CenterBack Wingback  
## Age 0.01 0.03 0.10 0.11 0.06  
## Overall 0.31 0.33 0.31 0.28 0.29  
## Potential 0.23 0.23 0.18 0.14 0.18  
## Value 0.40 0.41 0.35 0.29 0.34  
## Wage 0.32 0.34 0.30 0.26 0.28  
## Special 0.89 0.90 0.82 0.71 0.82  
## Reactions 0.32 0.34 0.32 0.27 0.29  
## Striker 0.99 0.99 0.79 0.68 0.82  
## Winger 1.00 0.99 0.79 0.67 0.82  
## AttMid 0.99 1.00 0.86 0.74 0.88  
## DefMid 0.79 0.86 1.00 0.98 0.99  
## CenterBack 0.67 0.74 0.98 1.00 0.96  
## Wingback 0.82 0.88 0.99 0.96 1.00

corrplot(cormat2, method = "ellipse")



The ellipse matrix plot shows a positive linear relationship between all variables except Age and Potential. It is similar to the first order model. There are still masking relationships between Attacking role positions and Defense role positions in this log-Transformed model.

model2 = lm(logWage~Age+Overall+Potential+logValue+Special+Reactions+Preferred.Positions+Striker+Winger+AttMid+DefMid+CenterBack+Wingback)  
  
summary(model2, correlation=FALSE)

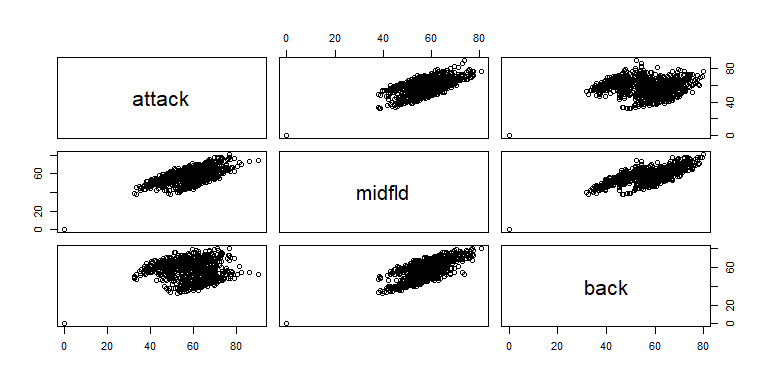
##   
## Call:  
## lm(formula = logWage ~ Age + Overall + Potential + logValue +   
## Special + Reactions + Preferred.Positions + Striker + Winger +   
## AttMid + DefMid + CenterBack + Wingback)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.61387 -0.18433 0.00786 0.18846 0.92719   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.2915367 0.2046564 -6.311 4.24e-10 \*\*\*  
## Age 0.0154229 0.0061465 2.509 0.01227 \*   
## Overall 0.0109538 0.0097475 1.124 0.26140   
## Potential 0.0119783 0.0039946 2.999 0.00278 \*\*   
## logValue 0.4639722 0.0910062 5.098 4.13e-07 \*\*\*  
## Special -0.0004157 0.0002038 -2.039 0.04168 \*   
## Reactions -0.0004428 0.0021883 -0.202 0.83971   
## Preferred.PositionsCenterBack 0.0491382 0.0464735 1.057 0.29062   
## Preferred.PositionsDefMid 0.0078906 0.0414074 0.191 0.84891   
## Preferred.PositionsGK 0.6540498 0.2215648 2.952 0.00323 \*\*   
## Preferred.PositionsStriker -0.0187023 0.0424413 -0.441 0.65956   
## Preferred.PositionsWingback 0.0414225 0.0387019 1.070 0.28476   
## Preferred.PositionsWinger 0.0012561 0.0532018 0.024 0.98117   
## Striker 0.0017426 0.0090742 0.192 0.84776   
## Winger 0.0046347 0.0148408 0.312 0.75489   
## AttMid 0.0147640 0.0171547 0.861 0.38965   
## DefMid -0.0162567 0.0154101 -1.055 0.29172   
## CenterBack 0.0217901 0.0090229 2.415 0.01592 \*   
## Wingback -0.0100773 0.0095285 -1.058 0.29051   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.309 on 956 degrees of freedom  
## Multiple R-squared: 0.6788, Adjusted R-squared: 0.6728   
## F-statistic: 112.3 on 18 and 956 DF, p-value: < 2.2e-16

anova(model2)

## Analysis of Variance Table  
##   
## Response: logWage  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Age 1 18.083 18.083 189.4413 < 2.2e-16 \*\*\*  
## Overall 1 167.884 167.884 1758.7844 < 2.2e-16 \*\*\*  
## Potential 1 0.523 0.523 5.4748 0.019497 \*   
## logValue 1 4.128 4.128 43.2498 7.918e-11 \*\*\*  
## Special 1 0.106 0.106 1.1101 0.292334   
## Reactions 1 0.001 0.001 0.0139 0.906069   
## Preferred.Positions 6 0.266 0.044 0.4636 0.835520   
## Striker 1 0.840 0.840 8.8006 0.003086 \*\*   
## Winger 1 0.160 0.160 1.6717 0.196347   
## AttMid 1 0.072 0.072 0.7532 0.385675   
## DefMid 1 0.229 0.229 2.4004 0.121634   
## CenterBack 1 0.480 0.480 5.0302 0.025137 \*   
## Wingback 1 0.107 0.107 1.1185 0.290506   
## Residuals 956 91.255 0.095   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The log-transformed model ANOVA table suggests that Age, Potential, Overall, logValue, Striker and CenterBack are significant predictors of players' Wages. The coefficient tests suggest that Age, Potential, logValue, Special and CenterBack are significant predictors. The logValue significance code shows that it is a predictor of players' Wages - the player's value is first decided, generally in real life, before the player's wage is decided. The R-Squared is 0.6788, with adjusted R-squared = 0.6728, which indicates an improvement from the first order model.

attack = rowMeans(cbind(Striker, Winger))  
midfld = rowMeans(cbind(AttMid, DefMid))  
back = rowMeans(cbind(CenterBack, Wingback))  
pairs(cbind.data.frame(attack, midfld, back))



We chose to summarize player's positions into three variables as we proved that the previous variables were masking each other. Also, in soccer players can be grouped into offensive, midfielder and defensive players. Players in the same group perform similar functions in the field and have similar attributes.

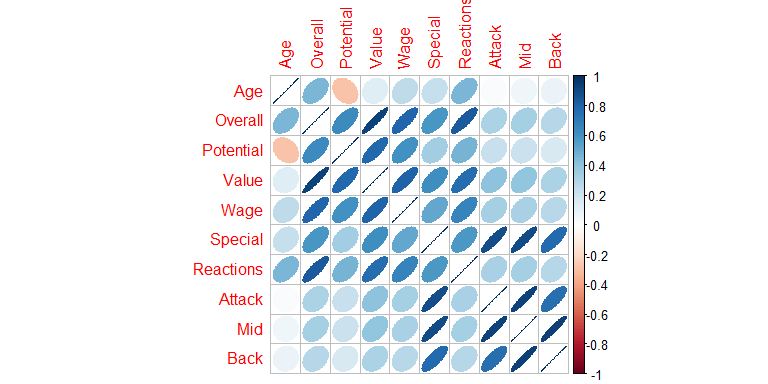
new\_df2 = new\_df[,1:7]  
new\_df2[8] = attack  
new\_df2[9] = midfld  
new\_df2[10] = back  
colnames(new\_df2)[8] <- "Attack"  
colnames(new\_df2)[9] <- "Mid"  
colnames(new\_df2)[10] <- "Back"  
  
head(new\_df2, n= 10)

## Age Overall Potential Value Wage Special Reactions Attack  
## 1 25 64 67 5.740363 3.301030 1749 59 60.50  
## 2 27 70 70 6.278754 3.954243 1901 62 69.25  
## 3 33 68 68 5.759668 4.176091 1715 65 65.00  
## 4 35 62 62 5.113943 3.000000 1701 68 59.00  
## 5 19 67 81 6.204120 4.623249 1675 63 67.50  
## 6 28 71 71 6.380211 3.954243 1824 68 68.00  
## 7 18 55 68 5.204120 3.000000 1504 55 52.00  
## 8 37 69 69 5.511883 3.000000 1916 70 65.25  
## 9 23 67 73 6.000000 3.778151 1576 65 65.25  
## 10 30 76 76 6.875061 4.929419 1897 77 73.50  
## Mid Back  
## 1 62.0000 59.25  
## 2 64.0625 61.25  
## 3 52.3125 43.75  
## 4 59.0625 55.50  
## 5 56.1250 44.25  
## 6 58.1875 55.25  
## 7 51.6250 47.25  
## 8 67.4375 66.75  
## 9 51.3750 40.25  
## 10 61.1250 55.00

cormat3 = cor(new\_df2)  
round(cormat3, 2)

## Age Overall Potential Value Wage Special Reactions Attack Mid  
## Age 1.00 0.45 -0.28 0.14 0.25 0.23 0.45 0.02 0.07  
## Overall 0.45 1.00 0.63 0.93 0.80 0.59 0.84 0.31 0.33  
## Potential -0.28 0.63 1.00 0.78 0.61 0.34 0.47 0.23 0.22  
## Value 0.14 0.93 0.78 1.00 0.80 0.61 0.76 0.41 0.39  
## Wage 0.25 0.80 0.61 0.80 1.00 0.52 0.67 0.33 0.33  
## Special 0.23 0.59 0.34 0.61 0.52 1.00 0.57 0.89 0.89  
## Reactions 0.45 0.84 0.47 0.76 0.67 0.57 1.00 0.32 0.34  
## Attack 0.02 0.31 0.23 0.41 0.33 0.89 0.32 1.00 0.93  
## Mid 0.07 0.33 0.22 0.39 0.33 0.89 0.34 0.93 1.00  
## Back 0.08 0.29 0.16 0.32 0.27 0.78 0.29 0.75 0.94  
## Back  
## Age 0.08  
## Overall 0.29  
## Potential 0.16  
## Value 0.32  
## Wage 0.27  
## Special 0.78  
## Reactions 0.29  
## Attack 0.75  
## Mid 0.94  
## Back 1.00

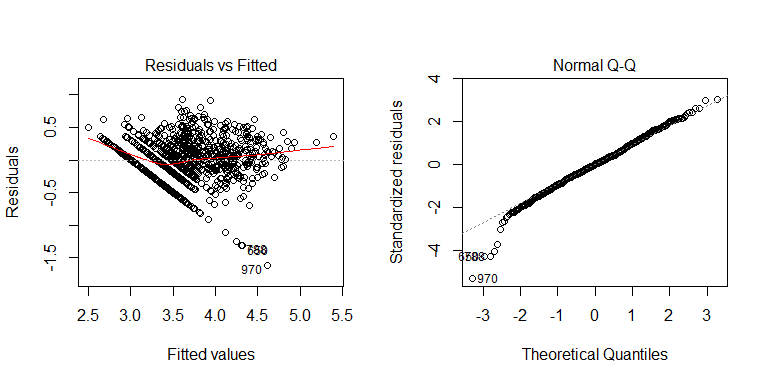
corrplot(cormat3, method = "ellipse")



Here we plotted the ellipse matrix to see their relation with log(Wage) with positions grouped by their respective areas on the field.

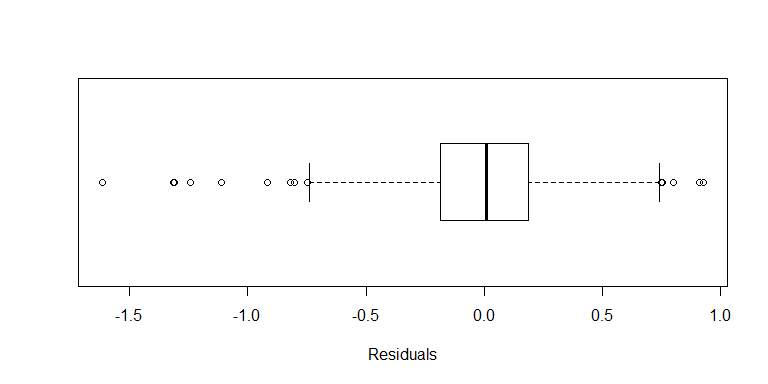
# Residual Analysis of the Log-Transformed first-order model.

par (mfrow = c(1, 2))  
plot(model2,which = c(1,2))

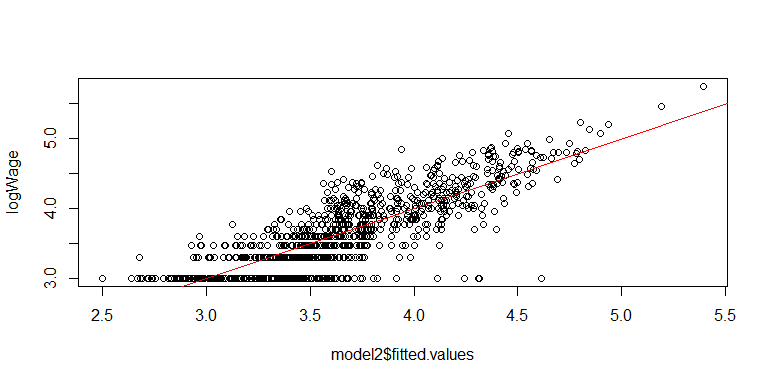


The Residuals vs Fitted plot shows a smooth fit line. The model also shows under prediction for both the smallest wages and the highest wages. In addition to that, the variability of the residuals is highest in the middle.

boxplot(model2$residuals, horizontal = T, xlab="Residuals")



plot(model2$fitted.values, logWage)  
abline(0, 1, col="red")



The residual analysis of the log-transformed model looks a little better than in the first order model. Although, the abline is not closer to 0. There are residuals at both ends of the scale that are somewhat more extreme when compared to a normal distribution. The plot of observed vs fitted Wage looks good. The value of 3.0 on the y-axis is the minimum value of the logWage showing that there is a wide range of players that receive minimum wage. There is spread on the x-axis also for players with logWage = 3.3, logWage = 3.5, logWage = 3.7. For all these the model is underpredicting and over predicting certain players.

# Box-Cox Optimal Transformation.

par (mfrow = c(1,2))  
bcWage = Wage^(-0.2)  
model3 = lm (bcWage~Age+Overall+Potential+logValue+Special+Reactions+Preferred.Positions+Striker+Winger+AttMid+DefMid+CenterBack+Wingback)  
summary(model3)

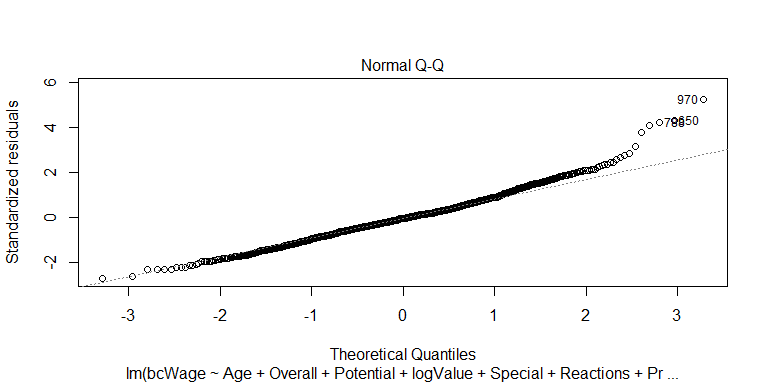
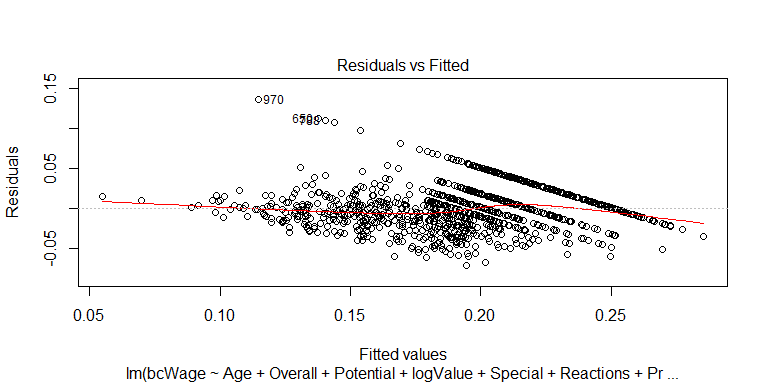
##   
## Call:  
## lm(formula = bcWage ~ Age + Overall + Potential + logValue +   
## Special + Reactions + Preferred.Positions + Striker + Winger +   
## AttMid + DefMid + CenterBack + Wingback)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.070387 -0.016063 -0.000915 0.014481 0.136308   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.842e-01 1.742e-02 33.524 < 2e-16 \*\*\*  
## Age -1.051e-03 5.233e-04 -2.008 0.0449 \*   
## Overall -1.571e-03 8.299e-04 -1.893 0.0587 .   
## Potential -7.272e-04 3.401e-04 -2.138 0.0328 \*   
## logValue -3.354e-02 7.748e-03 -4.329 1.65e-05 \*\*\*  
## Special 2.960e-05 1.735e-05 1.706 0.0884 .   
## Reactions 1.172e-04 1.863e-04 0.629 0.5295   
## Preferred.PositionsCenterBack -3.802e-03 3.957e-03 -0.961 0.3368   
## Preferred.PositionsDefMid -1.245e-03 3.525e-03 -0.353 0.7240   
## Preferred.PositionsGK -4.537e-02 1.886e-02 -2.405 0.0164 \*   
## Preferred.PositionsStriker 6.684e-05 3.614e-03 0.018 0.9852   
## Preferred.PositionsWingback -2.344e-03 3.295e-03 -0.711 0.4770   
## Preferred.PositionsWinger 2.835e-03 4.530e-03 0.626 0.5316   
## Striker 3.207e-04 7.726e-04 0.415 0.6782   
## Winger -1.162e-03 1.264e-03 -0.920 0.3579   
## AttMid -7.542e-04 1.461e-03 -0.516 0.6057   
## DefMid 1.197e-03 1.312e-03 0.913 0.3617   
## CenterBack -1.938e-03 7.682e-04 -2.523 0.0118 \*   
## Wingback 1.138e-03 8.113e-04 1.403 0.1610   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.02631 on 956 degrees of freedom  
## Multiple R-squared: 0.6571, Adjusted R-squared: 0.6506   
## F-statistic: 101.8 on 18 and 956 DF, p-value: < 2.2e-16

anova(model3)

## Analysis of Variance Table  
##   
## Response: bcWage  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Age 1 0.13660 0.13660 197.4071 < 2.2e-16 \*\*\*  
## Overall 1 1.09041 1.09041 1575.8227 < 2.2e-16 \*\*\*  
## Potential 1 0.00170 0.00170 2.4570 0.11733   
## logValue 1 0.02272 0.02272 32.8333 1.345e-08 \*\*\*  
## Special 1 0.00112 0.00112 1.6214 0.20320   
## Reactions 1 0.00030 0.00030 0.4333 0.51053   
## Preferred.Positions 6 0.00330 0.00055 0.7937 0.57494   
## Striker 1 0.00418 0.00418 6.0400 0.01416 \*   
## Winger 1 0.00044 0.00044 0.6320 0.42684   
## AttMid 1 0.00007 0.00007 0.0980 0.75436   
## DefMid 1 0.00175 0.00175 2.5341 0.11174   
## CenterBack 1 0.00359 0.00359 5.1941 0.02288 \*   
## Wingback 1 0.00136 0.00136 1.9677 0.16102   
## Residuals 956 0.66151 0.00069   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The coefficient analysis of the Box-Cox optimal transformed model suggests that Age, Overall, logValue, Striker and CenterBack are significant predictor variables.

plot(model3, which = c(1,2))



The model above fairly well. To interpret it, we first note that the response variable Wage has been transformed by raising to the power of, -0.2. This reverses the direction of the relationships between Wage and its predictors. Each of the following statements is made in the context of the other predictors being held at fixed values.

As expected the distribution of player's values is just as right skewed as their Wage. In the soccer world the number of players with large values is relatively small when compared to the whole. This causes the skewness of the variable.

# Eliminating Goal Keepers

Removing goal keepers from the model.

fifa2 = fifa[!Preferred.Positions== "GK",]  
head(fifa2, n=10)

## ï..ID Name Age Overall Potential Value Wage Special  
## 1 11019 J. Nolan 25 64 67 550000 2000 1749  
## 2 5715 Han Kyo Won 27 70 70 1900000 9000 1901  
## 3 7281 Cris Laranjeiros 33 68 68 575000 15000 1715  
## 4 13261 A. Hyodo 35 62 62 130000 1000 1701  
## 5 7766 S. DÃƒÂ­az 19 67 81 1600000 42000 1675  
## 6 4224 J. Vaughan 28 71 71 2400000 9000 1824  
## 7 16726 P. Pannier 18 55 68 160000 1000 1504  
## 8 6554 F. Vargas 37 69 69 325000 1000 1916  
## 9 7707 G. van Velzen 23 67 73 1000000 6000 1576  
## 10 1217 S. Long 30 76 76 7500000 85000 1897  
## Reactions Preferred.Positions Striker Winger AttMid DefMid CenterBack  
## 1 59 DefMid 59.0 62 63.000 61 57  
## 2 62 AttMid 68.5 70 67.125 61 59  
## 3 65 AttMid 66.0 64 59.625 45 40  
## 4 68 DefMid 60.0 58 60.125 58 56  
## 5 63 Striker 67.0 68 65.250 47 39  
## 6 68 Striker 69.0 67 62.375 54 55  
## 7 55 AttMid 51.0 53 54.250 49 45  
## 8 70 DefMid 65.5 65 66.875 68 69  
## 9 65 AttMid 63.5 67 60.750 42 34  
## 10 77 Striker 74.0 73 68.250 54 53  
## Wingback  
## 1 61.5  
## 2 63.5  
## 3 47.5  
## 4 55.0  
## 5 49.5  
## 6 55.5  
## 7 49.5  
## 8 64.5  
## 9 46.5  
## 10 57.0

attach(fifa2)

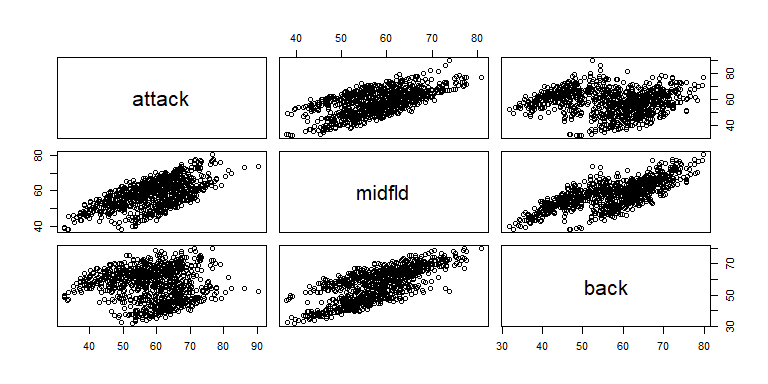
## The following objects are masked from fifa:  
##   
## Age, AttMid, CenterBack, DefMid, ï..ID, Name, Overall,  
## Potential, Preferred.Positions, Reactions, Special, Striker,  
## Value, Wage, Wingback, Winger

rm(fifa)

logWage = log10(Wage)  
logValue = log10(Value)  
bcWage = Wage^(-0.2)

attack = rowMeans(cbind(Striker, Winger))  
midfld = rowMeans(cbind(AttMid, DefMid))  
back = rowMeans(cbind(CenterBack, Wingback))

pairs(cbind.data.frame(attack, midfld, back))



The pairs plot after eliminating goal keepers increased the spread between players when grouped by sections of the field.

The pairs plots show a positive relation between all sections in the field although showing a more spread out relation between attack and back. In soccer, attacking role positions require a different set of skills than defensive role positions. This would explain why those two areas of the field show lower correlation than each of those compared to midfld.

model4 = lm(logWage~Age+Overall+Potential+logValue+Special+Reactions+Preferred.Positions+attack+midfld+back, data = as.data.frame(fifa2))  
summary(model4)

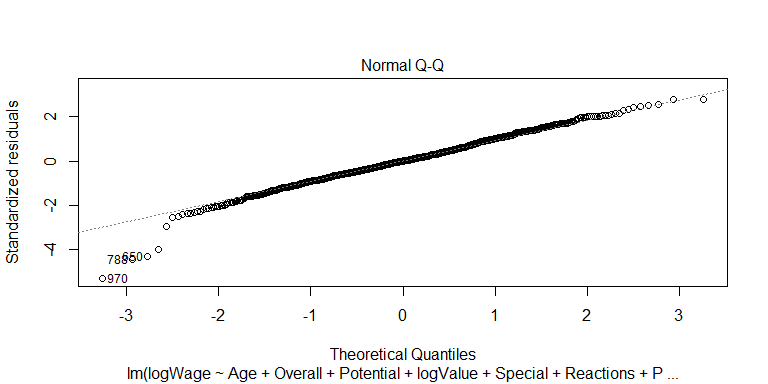
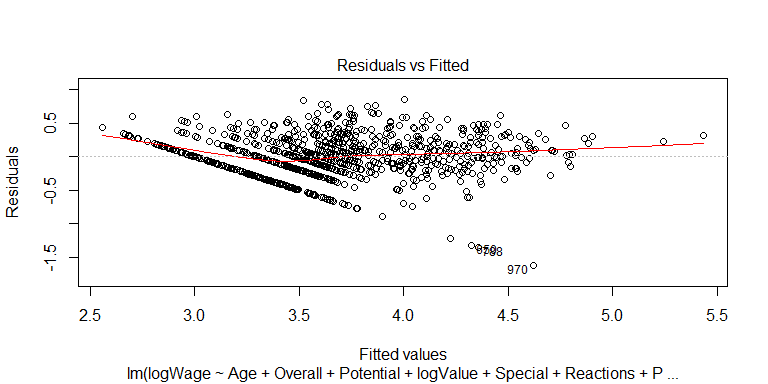
##   
## Call:  
## lm(formula = logWage ~ Age + Overall + Potential + logValue +   
## Special + Reactions + Preferred.Positions + attack + midfld +   
## back, data = as.data.frame(fifa2))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.61971 -0.18764 -0.00107 0.19288 0.84896   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.4788151 0.2128051 -6.949 7.20e-12 \*\*\*  
## Age 0.0200646 0.0064449 3.113 0.001911 \*\*   
## Overall 0.0104481 0.0110779 0.943 0.345864   
## Potential 0.0150083 0.0042530 3.529 0.000439 \*\*\*  
## logValue 0.4716517 0.1065939 4.425 1.09e-05 \*\*\*  
## Special -0.0000867 0.0002783 -0.312 0.755411   
## Reactions 0.0011808 0.0023306 0.507 0.612511   
## Preferred.PositionsCenterBack 0.0373155 0.0479477 0.778 0.436630   
## Preferred.PositionsDefMid 0.0189547 0.0412535 0.459 0.646011   
## Preferred.PositionsStriker 0.0123591 0.0396708 0.312 0.755464   
## Preferred.PositionsWingback 0.0302081 0.0379485 0.796 0.426232   
## Preferred.PositionsWinger 0.0055662 0.0533742 0.104 0.916967   
## attack 0.0050248 0.0061873 0.812 0.416950   
## midfld -0.0045384 0.0067086 -0.677 0.498898   
## back 0.0021038 0.0048331 0.435 0.663457   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.31 on 871 degrees of freedom  
## Multiple R-squared: 0.6778, Adjusted R-squared: 0.6726   
## F-statistic: 130.9 on 14 and 871 DF, p-value: < 2.2e-16

anova(model4)

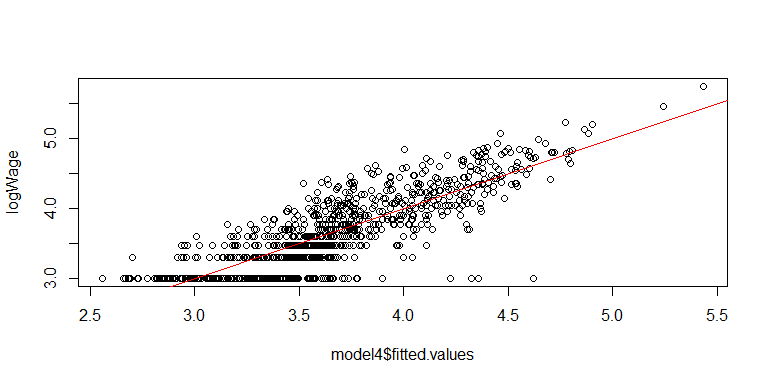
## Analysis of Variance Table  
##   
## Response: logWage  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Age 1 21.337 21.337 221.9951 < 2.2e-16 \*\*\*  
## Overall 1 150.935 150.935 1570.3793 < 2.2e-16 \*\*\*  
## Potential 1 1.200 1.200 12.4856 0.0004317 \*\*\*  
## logValue 1 2.335 2.335 24.2890 9.924e-07 \*\*\*  
## Special 1 0.009 0.009 0.0952 0.7577185   
## Reactions 1 0.034 0.034 0.3523 0.5529502   
## Preferred.Positions 5 0.119 0.024 0.2470 0.9413687   
## attack 1 0.094 0.094 0.9810 0.3222246   
## midfld 1 0.027 0.027 0.2762 0.5993612   
## back 1 0.018 0.018 0.1895 0.6634572   
## Residuals 871 83.715 0.096   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The analysis of the ANOVA table shows that Age, Overall, Potential, logValue are significant predictors of a player's wage for this model. Model4 suggest an improvement from model3 with a greater R-squared. Model4 suggests the same results from model2, therefore grouping players by areas of the field did not improve the model, neither did deleting goalkeepers.

plot(model4, which = c(1,2))



plot(model4$fitted.values, logWage)  
abline(0, 1, col="red")



The residual analysis of model4 suggests a better variance than the first order model. The abline is close to 0. There are still some outliers.

## Interaction Analysis

Here we do a stepwise backward elimination using both directions on model 4.

step.1st = step(model4, direction="both", k=log(nrow(fifa2)))

## Start: AIC=-1988.54  
## logWage ~ Age + Overall + Potential + logValue + Special + Reactions +   
## Preferred.Positions + attack + midfld + back  
##   
## Df Sum of Sq RSS AIC  
## - Preferred.Positions 5 0.08419 83.799 -2021.6  
## - Special 1 0.00933 83.725 -1995.2  
## - back 1 0.01821 83.733 -1995.1  
## - Reactions 1 0.02467 83.740 -1995.1  
## - midfld 1 0.04399 83.759 -1994.9  
## - attack 1 0.06339 83.779 -1994.7  
## - Overall 1 0.08550 83.801 -1994.4  
## <none> 83.715 -1988.5  
## - Age 1 0.93156 84.647 -1985.5  
## - Potential 1 1.19693 84.912 -1982.7  
## - logValue 1 1.88176 85.597 -1975.6  
##   
## Step: AIC=-2021.58  
## logWage ~ Age + Overall + Potential + logValue + Special + Reactions +   
## attack + midfld + back  
##   
## Df Sum of Sq RSS AIC  
## - Special 1 0.01129 83.811 -2028.2  
## - Reactions 1 0.03381 83.833 -2028.0  
## - back 1 0.05501 83.854 -2027.8  
## - attack 1 0.06945 83.869 -2027.6  
## - midfld 1 0.09662 83.896 -2027.3  
## - Overall 1 0.11725 83.917 -2027.1  
## <none> 83.799 -2021.6  
## - Age 1 0.94009 84.740 -2018.5  
## - Potential 1 1.21097 85.010 -2015.7  
## - logValue 1 1.85567 85.655 -2009.0  
## + Preferred.Positions 5 0.08419 83.715 -1988.5  
##   
## Step: AIC=-2028.25  
## logWage ~ Age + Overall + Potential + logValue + Reactions +   
## attack + midfld + back  
##   
## Df Sum of Sq RSS AIC  
## - Reactions 1 0.03653 83.847 -2034.7  
## - back 1 0.04391 83.855 -2034.6  
## - attack 1 0.07722 83.888 -2034.2  
## - midfld 1 0.10661 83.917 -2033.9  
## - Overall 1 0.12388 83.935 -2033.7  
## <none> 83.811 -2028.2  
## - Age 1 0.93113 84.742 -2025.2  
## - Potential 1 1.21305 85.024 -2022.3  
## + Special 1 0.01129 83.799 -2021.6  
## - logValue 1 1.84441 85.655 -2015.8  
## + Preferred.Positions 5 0.08615 83.725 -1995.2  
##   
## Step: AIC=-2034.65  
## logWage ~ Age + Overall + Potential + logValue + attack + midfld +   
## back  
##   
## Df Sum of Sq RSS AIC  
## - back 1 0.04830 83.896 -2040.9  
## - attack 1 0.08794 83.935 -2040.5  
## - midfld 1 0.10386 83.951 -2040.3  
## - Overall 1 0.13583 83.983 -2040.0  
## <none> 83.847 -2034.7  
## - Age 1 0.99905 84.846 -2030.9  
## - Potential 1 1.21100 85.058 -2028.7  
## + Reactions 1 0.03653 83.811 -2028.2  
## + Special 1 0.01400 83.833 -2028.0  
## - logValue 1 1.90732 85.755 -2021.5  
## + Preferred.Positions 5 0.09574 83.752 -2001.7  
##   
## Step: AIC=-2040.92  
## logWage ~ Age + Overall + Potential + logValue + attack + midfld  
##   
## Df Sum of Sq RSS AIC  
## - attack 1 0.05344 83.949 -2047.2  
## - midfld 1 0.10392 83.999 -2046.6  
## - Overall 1 0.15932 84.055 -2046.0  
## <none> 83.896 -2040.9  
## - Age 1 0.98079 84.876 -2037.4  
## - Potential 1 1.18131 85.077 -2035.3  
## + back 1 0.04830 83.847 -2034.7  
## + Reactions 1 0.04092 83.855 -2034.6  
## + Special 1 0.00006 83.895 -2034.1  
## - logValue 1 1.89339 85.789 -2027.9  
## + Preferred.Positions 5 0.13306 83.762 -2008.4  
##   
## Step: AIC=-2047.15  
## logWage ~ Age + Overall + Potential + logValue + midfld  
##   
## Df Sum of Sq RSS AIC  
## - midfld 1 0.06700 84.016 -2053.2  
## - Overall 1 0.12939 84.078 -2052.6  
## <none> 83.949 -2047.2  
## - Age 1 1.06038 85.009 -2042.8  
## - Potential 1 1.15816 85.107 -2041.8  
## + attack 1 0.05344 83.896 -2040.9  
## + Reactions 1 0.05080 83.898 -2040.9  
## + Special 1 0.03425 83.915 -2040.7  
## + back 1 0.01380 83.935 -2040.5  
## - logValue 1 2.32071 86.270 -2029.8  
## + Preferred.Positions 5 0.08445 83.865 -2014.1  
##   
## Step: AIC=-2053.23  
## logWage ~ Age + Overall + Potential + logValue  
##   
## Df Sum of Sq RSS AIC  
## - Overall 1 0.10016 84.116 -2059.0  
## <none> 84.016 -2053.2  
## - Age 1 1.05618 85.072 -2048.9  
## - Potential 1 1.17577 85.192 -2047.7  
## + midfld 1 0.06700 83.949 -2047.2  
## + back 1 0.06460 83.951 -2047.1  
## + Reactions 1 0.02483 83.991 -2046.7  
## + attack 1 0.01652 83.999 -2046.6  
## + Special 1 0.00915 84.007 -2046.5  
## - logValue 1 2.33451 86.351 -2035.7  
## + Preferred.Positions 5 0.13484 83.881 -2020.7  
##   
## Step: AIC=-2058.96  
## logWage ~ Age + Potential + logValue  
##   
## Df Sum of Sq RSS AIC  
## <none> 84.116 -2059.0  
## + Overall 1 0.1002 84.016 -2053.2  
## + midfld 1 0.0378 84.078 -2052.6  
## + Reactions 1 0.0368 84.079 -2052.6  
## + back 1 0.0219 84.094 -2052.4  
## + attack 1 0.0074 84.109 -2052.2  
## + Special 1 0.0063 84.110 -2052.2  
## - Potential 1 1.6072 85.723 -2049.0  
## + Preferred.Positions 5 0.1340 83.982 -2026.4  
## - Age 1 6.1077 90.224 -2003.6  
## - logValue 1 22.1013 106.218 -1859.0

summary(step.1st)

##   
## Call:  
## lm(formula = logWage ~ Age + Potential + logValue, data = as.data.frame(fifa2))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.61977 -0.19149 0.00077 0.19828 0.85420   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.574834 0.176702 -8.912 < 2e-16 \*\*\*  
## Age 0.025595 0.003198 8.003 3.82e-15 \*\*\*  
## Potential 0.016175 0.003940 4.105 4.42e-05 \*\*\*  
## logValue 0.583290 0.038316 15.223 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3088 on 882 degrees of freedom  
## Multiple R-squared: 0.6763, Adjusted R-squared: 0.6752   
## F-statistic: 614.1 on 3 and 882 DF, p-value: < 2.2e-16

anova(step.1st)

## Analysis of Variance Table  
##   
## Response: logWage  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Age 1 21.337 21.337 223.73 < 2.2e-16 \*\*\*  
## Potential 1 132.268 132.268 1386.90 < 2.2e-16 \*\*\*  
## logValue 1 22.101 22.101 231.74 < 2.2e-16 \*\*\*  
## Residuals 882 84.116 0.095   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The stepwise procedure using the AIC criterion suggested this to be our best model: logWage ~ Age + Potential + logValue. This shows that the most significant predictor variables are Age, Potential and logValue, when trying to predict logWage.

The next step would be to check the model with interactions.

fifa2$Age.c = Age - mean(Age)  
fifa2$Potential.c = Potential - mean(Potential)  
fifa2$logValue.c = logValue - mean(logValue)  
attach(fifa2)

## The following objects are masked from fifa2 (pos = 3):  
##   
## Age, AttMid, CenterBack, DefMid, ï..ID, Name, Overall,  
## Potential, Preferred.Positions, Reactions, Special, Striker,  
## Value, Wage, Wingback, Winger

## The following objects are masked from fifa:  
##   
## Age, AttMid, CenterBack, DefMid, ï..ID, Name, Overall,  
## Potential, Preferred.Positions, Reactions, Special, Striker,  
## Value, Wage, Wingback, Winger

full.int = lm(logWage ~ Age.c + Potential.c + logValue.c + Age.c \* Potential.c + Age.c \* logValue.c + Potential.c \* logValue.c)  
summary (full.int)

##   
## Call:  
## lm(formula = logWage ~ Age.c + Potential.c + logValue.c + Age.c \*   
## Potential.c + Age.c \* logValue.c + Potential.c \* logValue.c)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.71821 -0.18212 -0.01316 0.18744 0.89446   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.6524432 0.0160247 227.925 < 2e-16 \*\*\*  
## Age.c 0.0291926 0.0034116 8.557 < 2e-16 \*\*\*  
## Potential.c 0.0200165 0.0045375 4.411 1.15e-05 \*\*\*  
## logValue.c 0.5187733 0.0444950 11.659 < 2e-16 \*\*\*  
## Age.c:Potential.c 0.0007302 0.0007673 0.952 0.341533   
## Age.c:logValue.c 0.0052448 0.0071214 0.736 0.461626   
## Potential.c:logValue.c 0.0092843 0.0024086 3.855 0.000124 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3056 on 879 degrees of freedom  
## Multiple R-squared: 0.684, Adjusted R-squared: 0.6818   
## F-statistic: 317.1 on 6 and 879 DF, p-value: < 2.2e-16

The model with interactions presented above (R squared = 0.684) does better than the models 1 through 4.

anova (full.int)

## Analysis of Variance Table  
##   
## Response: logWage  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Age.c 1 21.337 21.337 228.4065 < 2.2e-16 \*\*\*  
## Potential.c 1 132.268 132.268 1415.9074 < 2.2e-16 \*\*\*  
## logValue.c 1 22.101 22.101 236.5906 < 2.2e-16 \*\*\*  
## Age.c:Potential.c 1 0.614 0.614 6.5715 0.0105276 \*   
## Age.c:logValue.c 1 0.002 0.002 0.0168 0.8969854   
## Potential.c:logValue.c 1 1.388 1.388 14.8587 0.0001243 \*\*\*  
## Residuals 879 82.113 0.093   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The ANOVA table of the stepwise procedure best model with all interactions shows that Age.c:Potential.c and Age.c:logValue.c are not as significant as the other interactions. Therefore, we decided to remove such interactions.

reduced.int = lm(logWage ~ Age.c + Potential.c + logValue.c + Potential.c \* logValue.c)  
summary (reduced.int)

##   
## Call:  
## lm(formula = logWage ~ Age.c + Potential.c + logValue.c + Potential.c \*   
## logValue.c)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.68423 -0.18301 -0.01065 0.18574 0.88612   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.650955 0.012041 303.209 < 2e-16 \*\*\*  
## Age.c 0.025630 0.003176 8.070 2.29e-15 \*\*\*  
## Potential.c 0.016496 0.003914 4.215 2.76e-05 \*\*\*  
## logValue.c 0.556697 0.038730 14.374 < 2e-16 \*\*\*  
## Potential.c:logValue.c 0.008646 0.002354 3.674 0.000254 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3067 on 881 degrees of freedom  
## Multiple R-squared: 0.6811, Adjusted R-squared: 0.6797   
## F-statistic: 470.5 on 4 and 881 DF, p-value: < 2.2e-16

anova(reduced.int)

## Analysis of Variance Table  
##   
## Response: logWage  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Age.c 1 21.337 21.337 226.897 < 2.2e-16 \*\*\*  
## Potential.c 1 132.268 132.268 1406.548 < 2.2e-16 \*\*\*  
## logValue.c 1 22.101 22.101 235.027 < 2.2e-16 \*\*\*  
## Potential.c:logValue.c 1 1.269 1.269 13.495 0.0002537 \*\*\*  
## Residuals 881 82.847 0.094   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Removing the interactions listed above decreased our R-squared value but suggested that all the predictor variables are significant now.

reduced.int = lm(logWage ~ Age.c + Potential.c + logValue.c + Age.c \* Potential.c + Potential.c \* logValue.c)  
summary (reduced.int)

##   
## Call:  
## lm(formula = logWage ~ Age.c + Potential.c + logValue.c + Age.c \*   
## Potential.c + Potential.c \* logValue.c)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.71734 -0.17936 -0.01552 0.18487 0.89303   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.6598773 0.0124428 294.137 < 2e-16 \*\*\*  
## Age.c 0.0288986 0.0033872 8.532 < 2e-16 \*\*\*  
## Potential.c 0.0211583 0.0042633 4.963 8.34e-07 \*\*\*  
## logValue.c 0.5089426 0.0424346 11.994 < 2e-16 \*\*\*  
## Age.c:Potential.c 0.0011927 0.0004407 2.706 0.006936 \*\*   
## Potential.c:logValue.c 0.0088873 0.0023469 3.787 0.000163 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3056 on 880 degrees of freedom  
## Multiple R-squared: 0.6838, Adjusted R-squared: 0.682   
## F-statistic: 380.6 on 5 and 880 DF, p-value: < 2.2e-16

anova(reduced.int)

## Analysis of Variance Table  
##   
## Response: logWage  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Age.c 1 21.337 21.337 228.525 < 2.2e-16 \*\*\*  
## Potential.c 1 132.268 132.268 1416.644 < 2.2e-16 \*\*\*  
## logValue.c 1 22.101 22.101 236.714 < 2.2e-16 \*\*\*  
## Age.c:Potential.c 1 0.614 0.614 6.575 0.0105073 \*   
## Potential.c:logValue.c 1 1.339 1.339 14.341 0.0001629 \*\*\*  
## Residuals 880 82.163 0.093   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Since taking out both variables made our model worse we decided to take the least significant out (Potential.c:logValue.c ) and maintain the other (Age.c:Potential.c). By doing so we have reached our best model so far with R-squared = 0.6838 and lower residual error = 0.3056.

anova (step.1st, reduced.int)

## Analysis of Variance Table  
##   
## Model 1: logWage ~ Age + Potential + logValue  
## Model 2: logWage ~ Age.c + Potential.c + logValue.c + Age.c \* Potential.c +   
## Potential.c \* logValue.c  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 882 84.116   
## 2 880 82.163 2 1.9528 10.458 3.246e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Comparing the model suggest by the stepwise backward elimination with the reduced.int model we can show that the latter is better. A smaller p value for reduced.int shows that is statistically more significant than the model suggested through AIC criterion.

Below we do one more stepwise to ensure that there are no more insignificant variables to remove.

reduced.step = step(reduced.int, direction="both", k=log(nrow(fifa2)))

## Start: AIC=-2066.19  
## logWage ~ Age.c + Potential.c + logValue.c + Age.c \* Potential.c +   
## Potential.c \* logValue.c  
##   
## Df Sum of Sq RSS AIC  
## <none> 82.163 -2066.2  
## - Age.c:Potential.c 1 0.68381 82.847 -2065.6  
## - Potential.c:logValue.c 1 1.33894 83.502 -2058.7

summary(reduced.step)

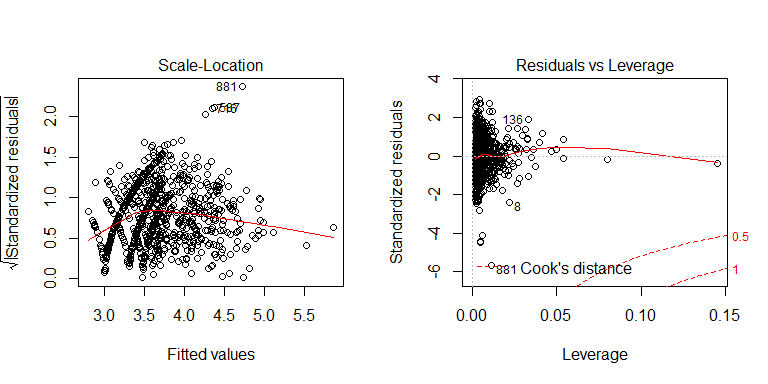
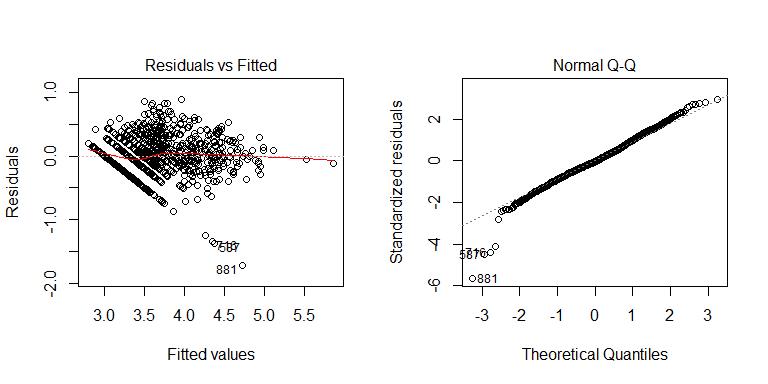
##   
## Call:  
## lm(formula = logWage ~ Age.c + Potential.c + logValue.c + Age.c \*   
## Potential.c + Potential.c \* logValue.c)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.71734 -0.17936 -0.01552 0.18487 0.89303   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.6598773 0.0124428 294.137 < 2e-16 \*\*\*  
## Age.c 0.0288986 0.0033872 8.532 < 2e-16 \*\*\*  
## Potential.c 0.0211583 0.0042633 4.963 8.34e-07 \*\*\*  
## logValue.c 0.5089426 0.0424346 11.994 < 2e-16 \*\*\*  
## Age.c:Potential.c 0.0011927 0.0004407 2.706 0.006936 \*\*   
## Potential.c:logValue.c 0.0088873 0.0023469 3.787 0.000163 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3056 on 880 degrees of freedom  
## Multiple R-squared: 0.6838, Adjusted R-squared: 0.682   
## F-statistic: 380.6 on 5 and 880 DF, p-value: < 2.2e-16

anova(reduced.step)

## Analysis of Variance Table  
##   
## Response: logWage  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Age.c 1 21.337 21.337 228.525 < 2.2e-16 \*\*\*  
## Potential.c 1 132.268 132.268 1416.644 < 2.2e-16 \*\*\*  
## logValue.c 1 22.101 22.101 236.714 < 2.2e-16 \*\*\*  
## Age.c:Potential.c 1 0.614 0.614 6.575 0.0105073 \*   
## Potential.c:logValue.c 1 1.339 1.339 14.341 0.0001629 \*\*\*  
## Residuals 880 82.163 0.093   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Based on our second attempt to do a stepwise backward there is no other interactions to remove. All of them are statistically significant.

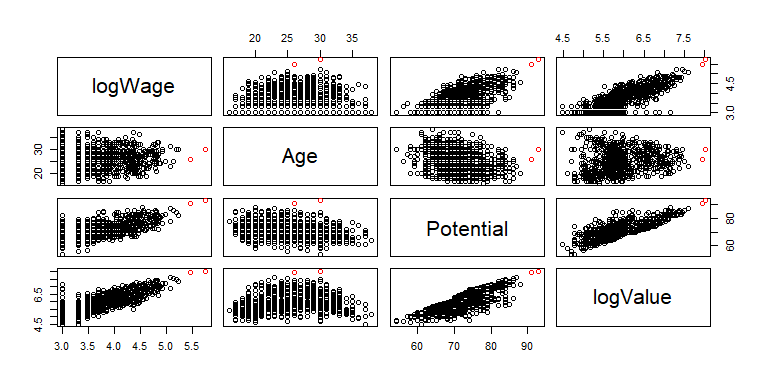
par(mfrow=c(1,2))  
plot(reduced.int)

 The Residual analysis of reduced.int show a more constant variance still being a little right skewed with some outliers. The Normal Q-Q plot shows the majority of observations are on the line with a few of them greater than the absolute value of 4.

hatvals = hatvalues(reduced.int)  
fifa2[hatvals>0.06,]

## ï..ID Name Age Overall Potential Value Wage Special  
## 299 1 L. Messi 30 93 93 105000000 565000 2154  
## 646 7 E. Hazard 26 90 91 90500000 295000 2096  
## Reactions Preferred.Positions Striker Winger AttMid DefMid CenterBack  
## 299 95 Winger 90.0 91 88.500 59 45  
## 646 85 Winger 84.5 88 85.125 61 47  
## Wingback Age.c Potential.c logValue.c  
## 299 59.5 4.8069977 21.89052 2.100216  
## 646 61.5 0.8069977 19.89052 2.035675

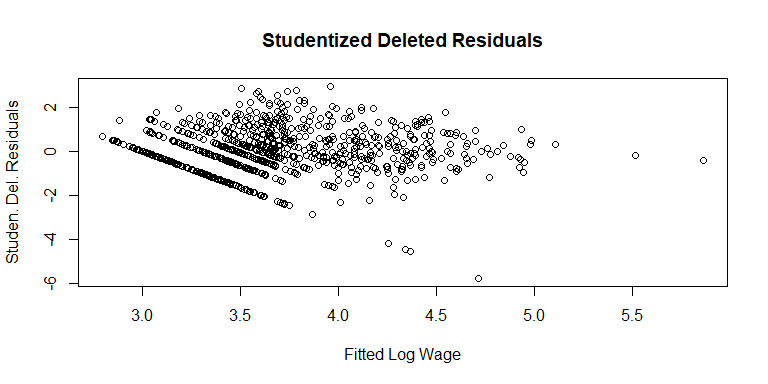
hat.colors = ifelse(hatvals > 0.06, 'red', 'black')  
pairs(cbind.data.frame(logWage, Age, Potential, logValue), col=hat.colors)



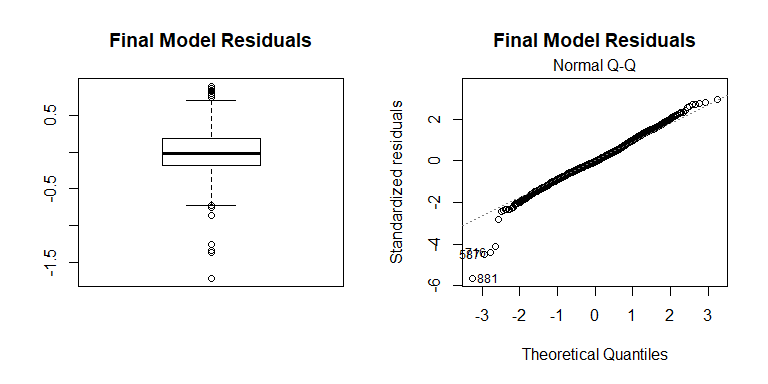
We observed from the Residuals VS Leverage plot that two players had greater hat values than the rest of the group which raised a concern whether their values could be higher than the cut-off value. Therefore, we used leverage = 2\*(sqrt(5/886)) in which 5 is the number of parameters and 886 is the number of observations in our data set. The result is: 0.1502443. Therefore, the two largest hat values are below the cut-off value for the Residuals VS Leverage plot.

# Studentizied residuals plot against the fitted values.

plot(reduced.int$fitted.values, rstudent(reduced.int), main = "Studentized Deleted Residuals", xlab="Fitted Log Wage", ylab= "Studen. Del. Residuals")

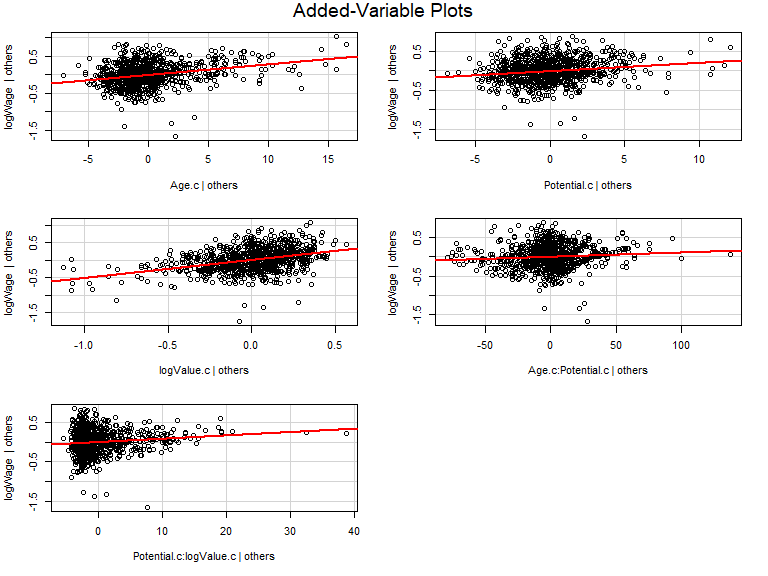


par (mfrow = c(1,2))  
boxplot (reduced.int$residuals, main="Final Model Residuals")  
plot (reduced.int, which=2, main="Final Model Residuals")



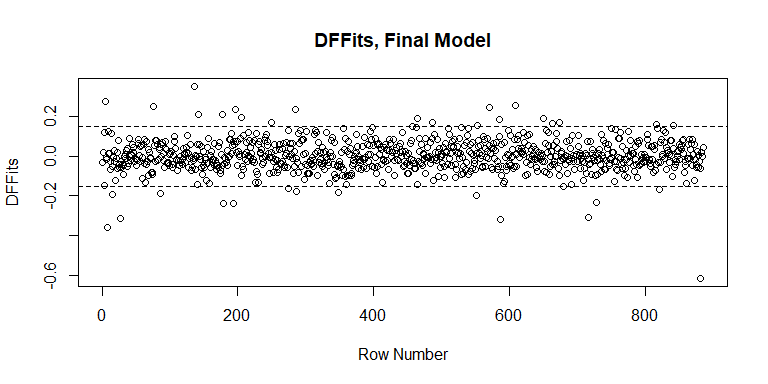
The box plot, along with the Q-Q plot of the final model residuals shows a few points in each tail that are wider than we would expect from a normal distribution, and some of the standardized residuals have an absolute value greater than 4.

par (mfrow = c(1,2))  
library (car)  
avPlots(reduced.int)



The added variable plots show a trend for each predictor in the model, given all of the other predictors. The directions of these trends correspond to the parameter estimates.

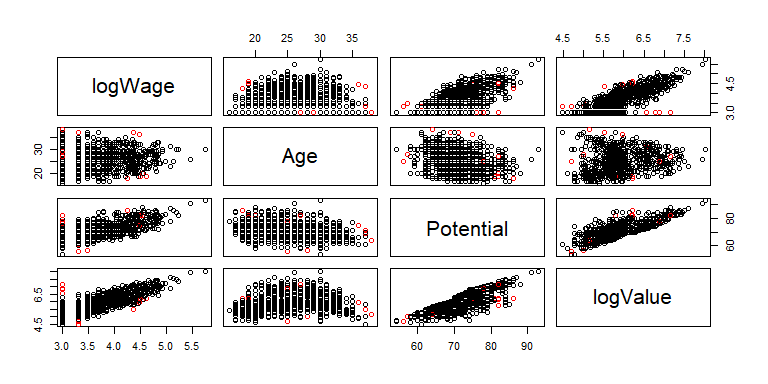
par (mfrow = c(1,1))  
fifa2$dffits = dffits (reduced.int)  
fifa2$Wage.fit = 10^reduced.int$fitted.values  
plot (seq (1, length (fifa2$dffits)), fifa2$dffits,   
 main="DFFits, Final Model",  
 xlab="Row Number", ylab="DFFits")  
abline (0.1502443, 0, lty=2)  
abline (-0.1502443, 0, lty=2)



fifa2 [abs(fifa2$dffits) > 0.2, c(2,3,5,7,21)]

## Name Age Potential Wage Wage.fit  
## 5 S. DÃƒÂ­az 19 81 42000 6111.024  
## 8 F. Vargas 37 69 1000 5323.651  
## 28 LÃƒÂ©o Matos 31 74 1000 17984.500  
## 85 M. Al Mutairi 25 56 2000 759.016  
## 152 C. Salcido 37 71 23000 6157.711  
## 158 M. Harzan 28 57 3000 1101.510  
## 196 M. Nelson 37 61 2000 1095.735  
## 197 K. JÃƒÂ³Ã…Âºwiak 19 82 1000 3502.517  
## 212 R. Kongolo 19 82 1000 3502.517  
## 216 Han Kwang Song 18 86 18000 6620.598  
## 311 M. Holgate 20 82 34000 6719.117  
## 629 P. Cannavaro 36 75 30000 13265.709  
## 650 J. VukoviÃ„â¡ 29 76 1000 23392.990  
## 675 N. VlaÃ…Â¡iÃ„â¡ 19 82 32000 6344.572  
## 788 S. Araujo 25 78 1000 22004.336  
## 802 E. Soligo 38 64 1000 2664.194  
## 970 Y. Rakitskyi 27 82 1000 52160.054

my.colors = ifelse(abs(fifa2$dffits) > 0.2, 'red', 'black')  
pairs(cbind.data.frame(logWage, Age, Potential, logValue), col=my.colors)



DFFITS is a diagnostic meant to show how influential a point is in a statistical regression. The table above shows players that have an absolute dffits value greater than 0.2.

# Interpretations

The final model has the following parameter estimates:

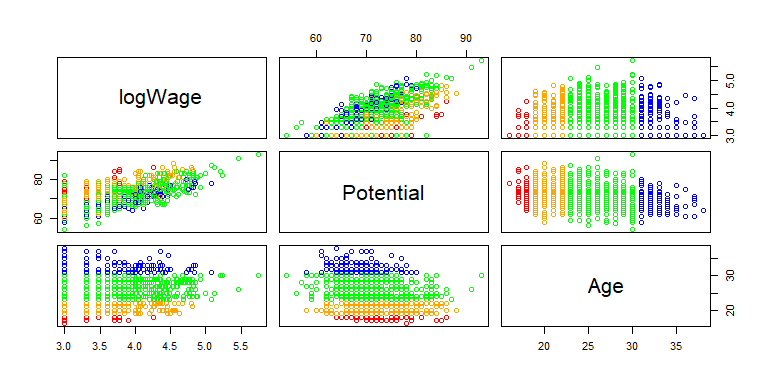
summary(reduced.int)$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 3.659877349 0.0124427832 294.136553 0.000000e+00  
## Age.c 0.028898575 0.0033872393 8.531601 6.255322e-17  
## Potential.c 0.021158341 0.0042633243 4.962874 8.338744e-07  
## logValue.c 0.508942561 0.0424346196 11.993570 8.355480e-31  
## Age.c:Potential.c 0.001192702 0.0004407185 2.706267 6.936135e-03  
## Potential.c:logValue.c 0.008887321 0.0023468645 3.786892 1.629022e-04

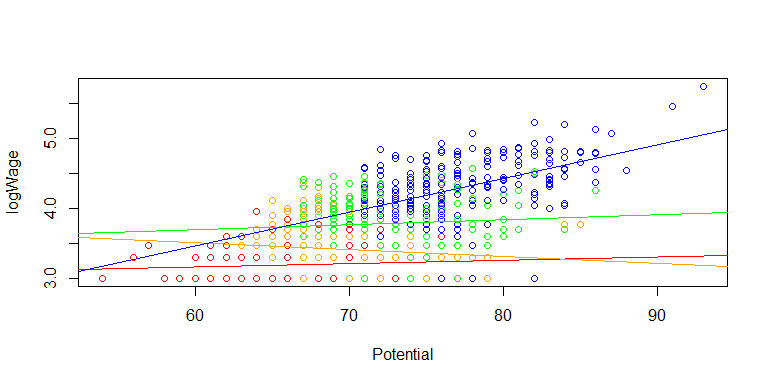
Since the response variable, Wage, was transformed to a log scale, we only interpret the direction of these effects.

The interaction effects are interpreted graphically. For example, plotting logWage vs Potential and coloring by Age, shows that the slope between logWage and Potential increases as Age increases.

my.colors = rep ("red", length (logWage))  
my.colors [Age > 18] = "orange"  
my.colors [Age > 22] = "green"  
my.colors [Age > 30] = "blue"  
pairs(cbind.data.frame (logWage, Potential, Age), col=my.colors)

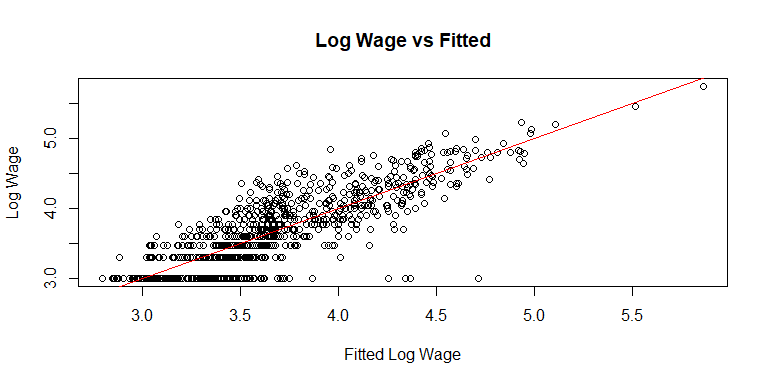


my.colors = rep ("red", length (logWage))  
my.colors [logValue > 5.54] = "orange"  
my.colors [logValue > 5.84] = "green"  
my.colors [logValue > 6.26] = "blue"  
plot(logWage ~ Potential, col=my.colors)  
abline(lm(logWage[logValue<5.54] ~ Potential[logValue < 5.54]),col='red')  
abline(lm(logWage[logValue>5.54&logValue<5.84] ~ Potential[logValue > 5.54&logValue<5.84]),col='orange')  
abline(lm(logWage[logValue>5.84&logValue<6.26] ~ Potential[logValue > 5.84&logValue<6.26]),col='green')  
abline(lm(logWage[logValue>6.26] ~ Potential[logValue > 6.26]),col='blue')



The plot above shows that players with the highest logValue (>6.26) have the steepest linear relationship between logWage and Potential. The other players have almost no linear relationship to negative linear relationships between logWage and Potential.

plot (logWage ~ reduced.int$fitted.values, main="Log Wage vs Fitted", xlab="Fitted Log Wage", ylab="Log Wage", col='black')  
abline(0, 1, col="red")



Above we examined LOG Wage vs Fitted Log Wage Values.

pred.all = as.data.frame (10^predict (model4, interval="prediction"))

## Warning in predict.lm(model4, interval = "prediction"): predictions on current data refer to \_future\_ responses

pred.all2 = round (pred.all, 1)  
pred.table = cbind.data.frame (Name, Wage, Age, Potential, Value, pred.all2$fit, pred.all2$lwr, pred.all2$upr)  
names (pred.table) [6:8] = c("Fit", "Lower", "Upper")  
pred.table$MOE = (pred.table$Upper / pred.table$Fit)  
pred.table [c(274,584,646,299,461,650),]

## Name Wage Age Potential Value Fit Lower  
## 274 L. Messi 565000 30 93 105000000 272101.1 64047.0  
## 584 E. Hazard 295000 26 91 90500000 174527.2 41545.5  
## 646 ÃƒÂlex MenÃƒÂ©ndez 5000 25 71 925000 5284.5 1282.4  
## 299 T. Hiraoka 1000 21 67 200000 1341.6 323.1  
## 461 Vitolo 68000 27 83 27000000 64287.0 15606.1  
## 650 Douglas 71000 26 72 2600000 10053.1 2448.7  
## Upper MOE  
## 274 1156010.0 4.248458  
## 584 733165.7 4.200868  
## 646 21777.3 4.120976  
## 299 5571.3 4.152728  
## 461 264820.5 4.119348  
## 650 41273.4 4.105540

The table above shows predictions for 6 players with Wage rescaled to its original value using a 10^ transformation. Analyzing the table, the model is under predicting for high wage players, roughly correctly predicting for low and medium wage players. The odd observation in this table is Douglas. Although there might be an explanation for such abnormality. Douglas is a Brazilian player that has always played in High-Payroll clubs in the world, despite being of medium Potential. A predictor variable that is not accounted in our data is which club’s players have played in their lives. Players in High-Payroll clubs tend to have a higher wage regardless of their skill level. The spread of margin of error is relatively the same for all 6 predictions.

pred.all = as.data.frame (10^predict (reduced.int, interval="prediction"))

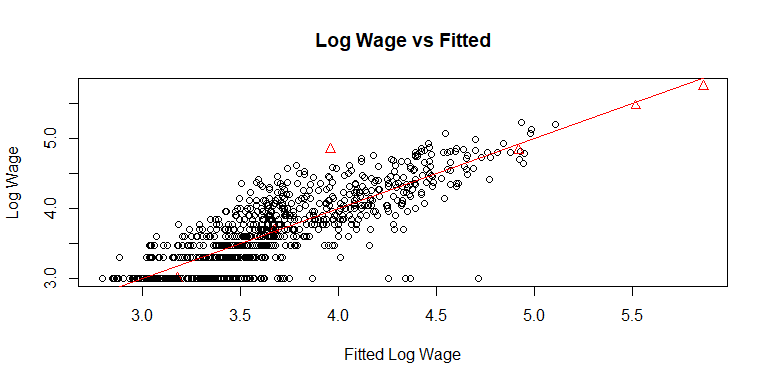
## Warning in predict.lm(reduced.int, interval = "prediction"): predictions on current data refer to \_future\_ responses

pred.all2 = round (pred.all, 1)  
pred.table = cbind.data.frame (Name, Wage, Age, Potential, Value, pred.all2$fit, pred.all2$lwr, pred.all2$upr)  
names (pred.table) [6:8] = c("Fit", "Lower", "Upper")  
pred.table$MOE = (pred.table$Upper / pred.table$Fit)  
pred.table [c(274,584,646,299,461,650),]

## Name Wage Age Potential Value Fit Lower  
## 274 L. Messi 565000 30 93 105000000 732735.6 167140.4  
## 584 E. Hazard 295000 26 91 90500000 330437.0 78685.8  
## 646 ÃƒÂlex MenÃƒÂ©ndez 5000 25 71 925000 4731.0 1187.8  
## 299 T. Hiraoka 1000 21 67 200000 1511.6 378.8  
## 461 Vitolo 68000 27 83 27000000 82721.7 20549.7  
## 650 Douglas 71000 26 72 2600000 9082.9 2276.5  
## Upper MOE  
## 274 3212278.5 4.383953  
## 584 1387652.6 4.199447  
## 646 18844.2 3.983133  
## 299 6032.1 3.990540  
## 461 332991.5 4.025443  
## 650 36239.7 3.989882

When we do the same predictions for our reduced.int model we get different results. It seems like now it is over estimating high Wage players and more or less correctly estimating low Wage players. For model4 predictions Messi and Hazard were being under predicted. Also, the intervals or mean squared error for the predictions is narrower which shows an improvement from model4. It suggests that the upper limit of the prediction is four times bigger than the fitted value.

my.colors = rep ("black", length(logWage))  
my.colors[c(274,584,646,299,461,650)] = 'red'  
symb = rep (1, length(logWage))  
symb[c(274,584,646,299,461,650)] = 2  
plot (logWage ~ reduced.int$fitted.values, main="Log Wage vs Fitted", xlab="Fitted Log Wage", ylab="Log Wage", col=my.colors, pch=symb)  
abline(0, 1, col="red")



The plot above shows were each of the players using in the predictions fall in the plot, using triangles. Douglas would be the triangle farthest away from the abline. His fitted logWage (3.95) is much lower than his actual LogWage(4.8).

anova(model4, reduced.int)

## Analysis of Variance Table  
##   
## Model 1: logWage ~ Age + Overall + Potential + logValue + Special + Reactions +   
## Preferred.Positions + attack + midfld + back  
## Model 2: logWage ~ Age.c + Potential.c + logValue.c + Age.c \* Potential.c +   
## Potential.c \* logValue.c  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 871 83.715   
## 2 880 82.163 -9 1.5519

# The ANOVA comparison above demonstrates once again how our reduced.int model does better against our previous experimented models. In this case, the log transformed model with stepwise backward elimination using the AIC criterion.

# References

<https://www.forbes.com/sites/christinasettimi/2017/05/26/the-worlds-highest-paid-soccer-players-2017-cristiano-ronaldo-lionel-messi-lead-the-list/#74543c5e210e>

<https://deadspin.com/chart-the-average-player-salaries-in-soccer-leagues-ar-1658856283>