FifaAnalysisABR

Ahmad, Beka, and Rafael

10/27/2017

## Introduction

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. 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 in order 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 “Prefered Positions”, we excluded every value (in case of two or more prefered 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.

### APPENDIX A

* 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 atlhetic 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

### 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

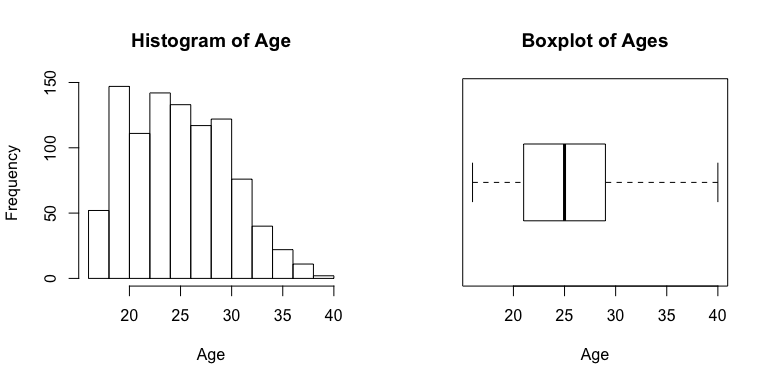
## Exploratory Analysis

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.

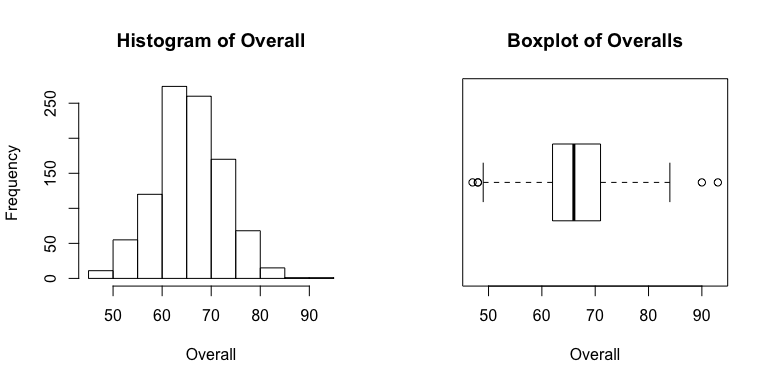
#Either use:  
 #file.choose()  
 #/Users/Beka/Downloads/Analysis/fifa18.csv  
 #/Users/ahmadosman-sw/Downloads/Stats/Project #1/First Draft/fifa18.csv  
fifa <- read.csv("/Users/ahmadosman-sw/Downloads/Stats/Project #1/First Draft/fifa18.csv", header=TRUE, sep=",")  
  
attach (fifa)  
head(fifa, n=10)

## 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. DÃ­az 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

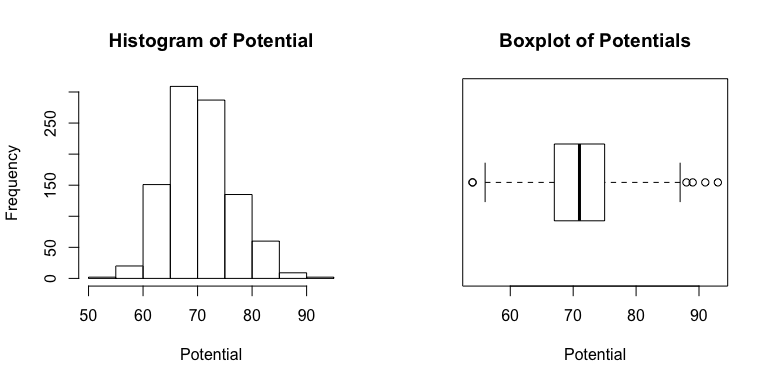
par (mfrow = c(1, 2))  
hist(Age)  
boxplot(Age, horizontal = T, xlab="Age", main="Boxplot of Ages")



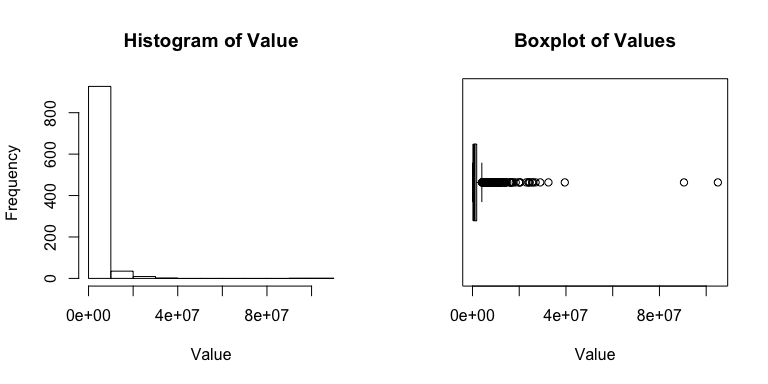
hist(Overall)  
boxplot(Overall, horizontal = T, xlab="Overall", main="Boxplot of Overalls")



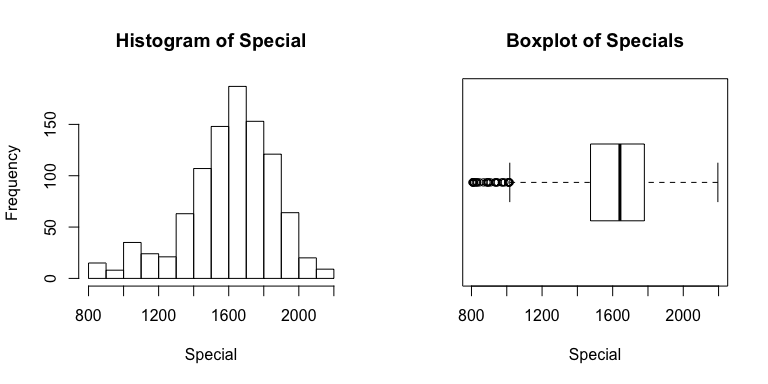
hist(Potential)  
boxplot(Potential, horizontal = T, xlab="Potential", main="Boxplot of Potentials")



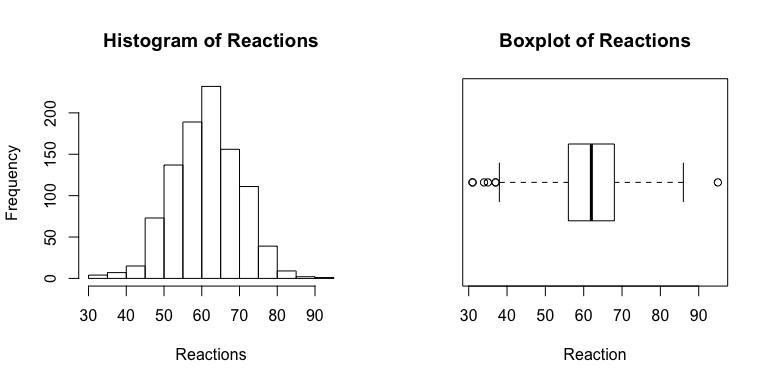
hist(Value)  
boxplot(Value, horizontal = T, xlab="Value", main="Boxplot of Values")



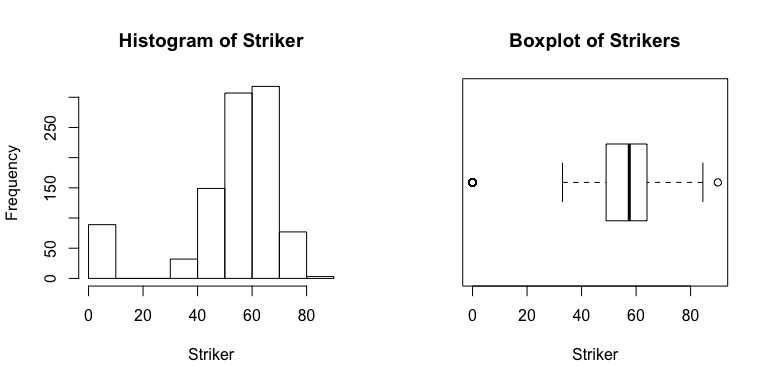
hist(Special)  
boxplot(Special, horizontal = T, xlab="Special", main="Boxplot of Specials")



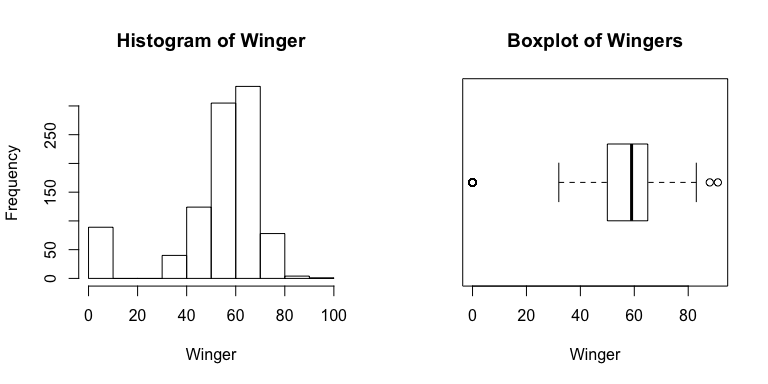
hist(Reactions)  
boxplot(Reactions, horizontal = T, xlab="Reaction", main="Boxplot of Reactions")



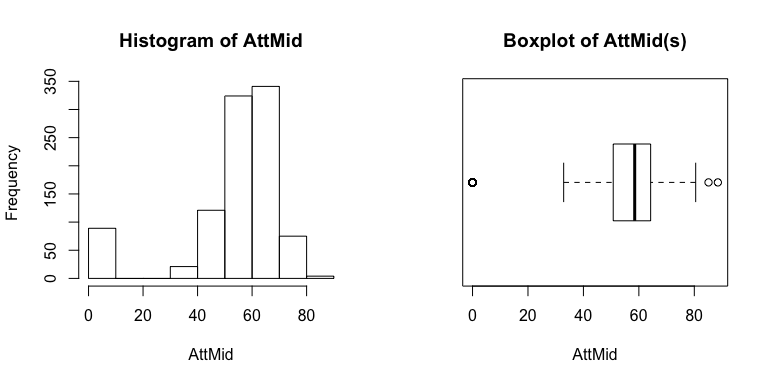
hist(Striker)  
boxplot(Striker, horizontal = T, xlab="Striker", main="Boxplot of Strikers")



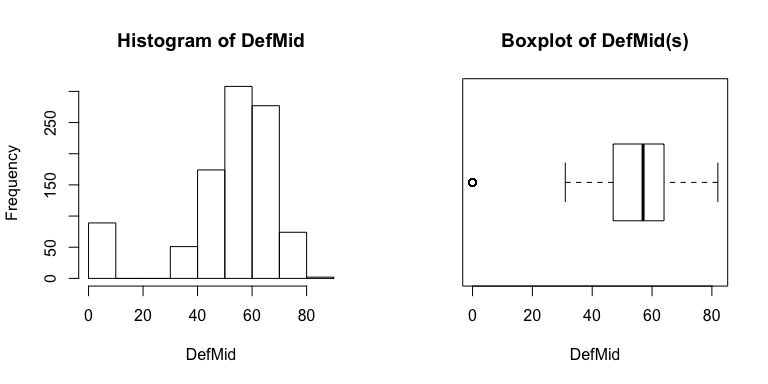
hist(Winger)  
boxplot(Winger, horizontal = T, xlab="Winger", main="Boxplot of Wingers")



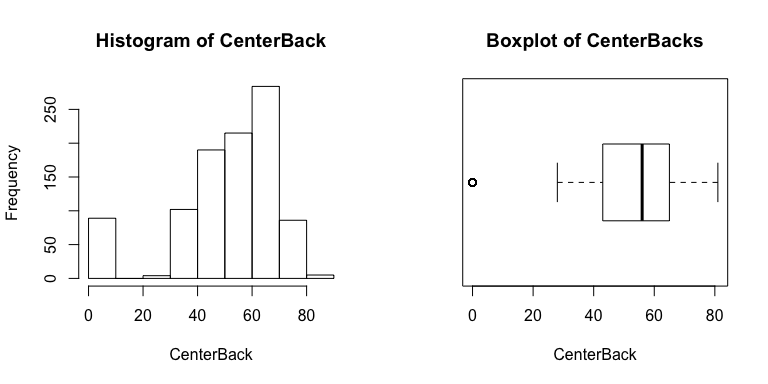
hist(AttMid)  
boxplot(AttMid, horizontal = T, xlab="AttMid", main="Boxplot of AttMid(s)")



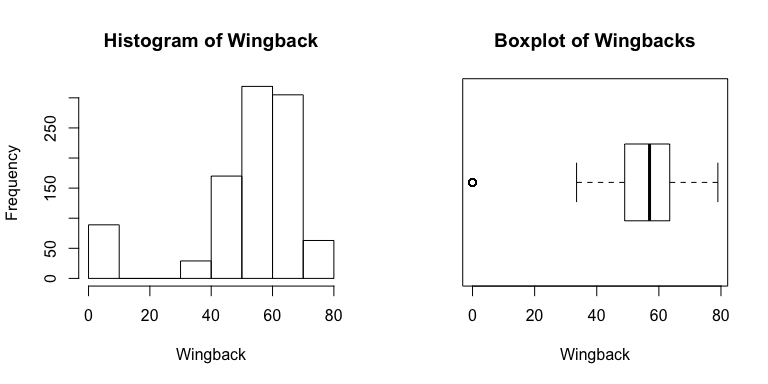
hist(DefMid)  
boxplot(DefMid, horizontal = T, xlab="DefMid", main="Boxplot of DefMid(s)")



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

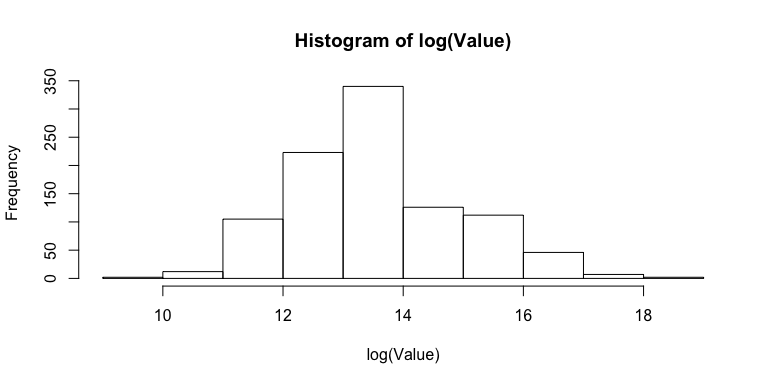


hist(Wingback)  
boxplot(Wingback, horizontal = T, xlab="Wingback", main="Boxplot of Wingbacks")



Anlayzing all the distributions we believe it is suitable to do a log transformation on the Value variable.

hist(log(Value))

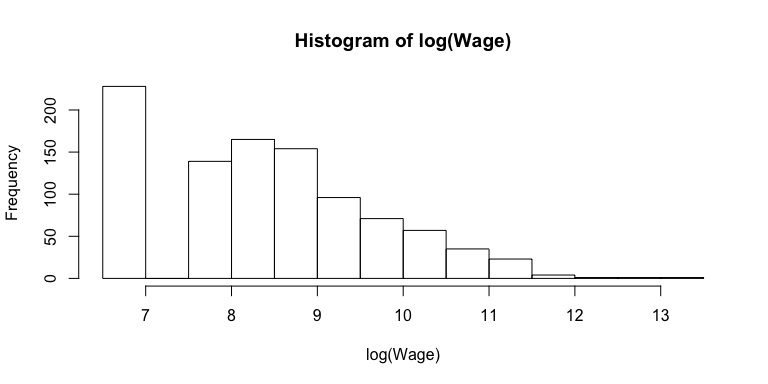


par (mfrow = c(1, 2))  
hist (Wage)  
boxplot (Wage, horizontal = T, xlab="Wage", main="Boxplot of Wages")

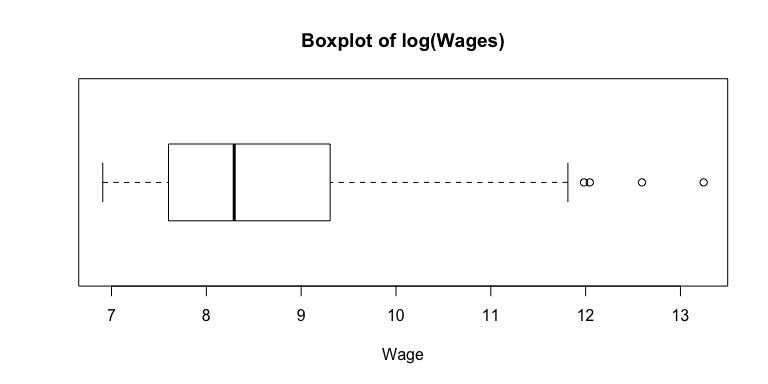


The distribution of Wage is extremely right skewed, therefore we believe a log transformation on such variable would be an appropriate approach as shown below.

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 professioanl team and do not receive 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.

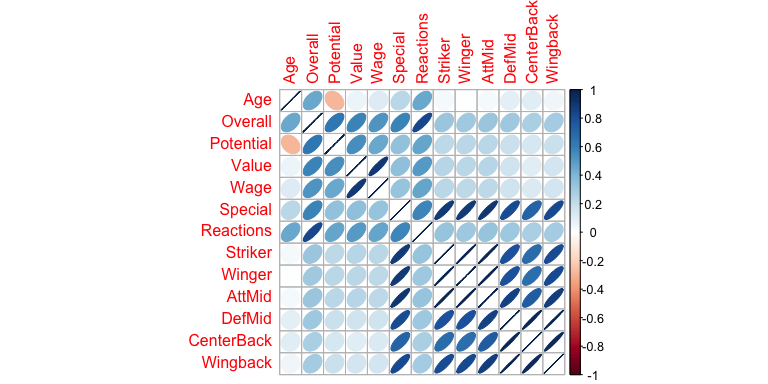
cormat = cor(fifa[,c(3,4,5,6,7,8,9,11,12,13,14,15,16)])  
round(cormat, 2)

## 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 pairs plot shows only non-linear relationship between Age and Potential. A log transformation may or may not resolve that. This will be addressed later on in this report. 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.

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

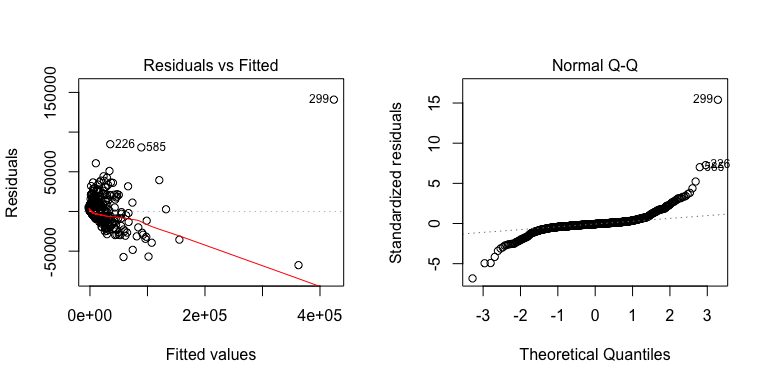
##   
## Call:  
## lm(formula = Wage ~ Age + Overall + Potential + Value + Special +   
## Reactions + Preferred.Positions + Striker + Winger + AttMid +   
## DefMid + CenterBack + Wingback)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -67642 -3440 -816 2027 140866   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.286e+04 7.863e+03 -1.635 0.102325   
## Age 6.312e+02 1.674e+02 3.770 0.000173 \*\*\*  
## Overall -3.902e+02 1.844e+02 -2.117 0.034553 \*   
## Potential 2.583e+02 1.521e+02 1.698 0.089773 .   
## Value 4.006e-03 8.544e-05 46.889 < 2e-16 \*\*\*  
## Special -3.099e+00 7.756e+00 -0.400 0.689577   
## Reactions -4.270e+01 8.334e+01 -0.512 0.608528   
## Preferred.PositionsCenterBack -2.678e+03 1.772e+03 -1.511 0.131062   
## Preferred.PositionsDefMid -8.634e+02 1.573e+03 -0.549 0.583160   
## Preferred.PositionsGK 1.001e+04 8.442e+03 1.185 0.236144   
## Preferred.PositionsStriker -3.664e+00 1.615e+03 -0.002 0.998190   
## Preferred.PositionsWingback 1.751e+03 1.470e+03 1.191 0.233862   
## Preferred.PositionsWinger 3.964e+03 2.048e+03 1.935 0.053268 .   
## Striker 3.241e+02 3.437e+02 0.943 0.345946   
## Winger -6.155e+02 5.648e+02 -1.090 0.276081   
## AttMid 6.124e+02 6.545e+02 0.936 0.349626   
## DefMid -6.803e+02 5.867e+02 -1.160 0.246503   
## CenterBack 4.948e+02 3.432e+02 1.442 0.149629   
## Wingback 1.171e+02 3.625e+02 0.323 0.746800   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11750 on 956 degrees of freedom  
## Multiple R-squared: 0.8042, Adjusted R-squared: 0.8005   
## F-statistic: 218.1 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 67.6948 6.203e-16 \*\*\*  
## Overall 1 1.9262e+11 1.9262e+11 1395.2544 < 2.2e-16 \*\*\*  
## Potential 1 4.3975e+09 4.3975e+09 31.8539 2.190e-08 \*\*\*  
## Value 1 3.3241e+11 3.3241e+11 2407.8597 < 2.2e-16 \*\*\*  
## Special 1 4.1438e+08 4.1438e+08 3.0016 0.08350 .   
## Reactions 1 2.4768e+07 2.4768e+07 0.1794 0.67197   
## Preferred.Positions 6 1.2950e+09 2.1583e+08 1.5634 0.15461   
## Striker 1 8.8083e+04 8.8083e+04 0.0006 0.97985   
## Winger 1 8.5960e+08 8.5960e+08 6.2267 0.01275 \*   
## AttMid 1 4.5544e+07 4.5544e+07 0.3299 0.56585   
## DefMid 1 2.5060e+08 2.5060e+08 1.8152 0.17820   
## CenterBack 1 3.2911e+08 3.2911e+08 2.3839 0.12292   
## Wingback 1 1.4399e+07 1.4399e+07 0.1043 0.74680   
## Residuals 956 1.3198e+11 1.3805e+08   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

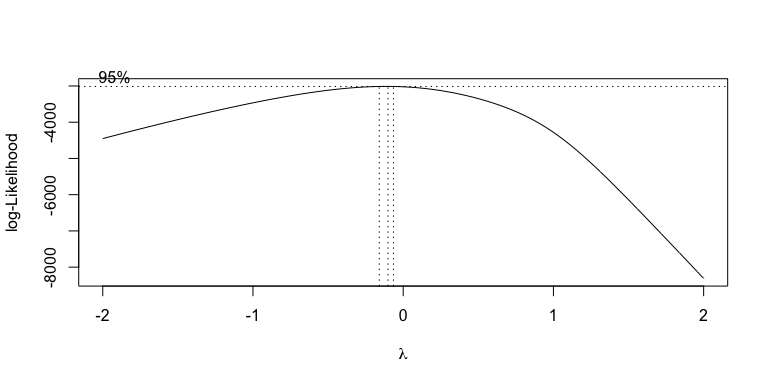
The analysis of the ANOVA table suggests that Age, Overall, Potential, Value and Winger are significant predictors of a player's Wage. The coefficients tests show that only Age, Overall and Value significant for this model. The R-squared is 0.8042, with R-squared = 0.8005, which indicates that most of the variability in Wage is being explained by this model. The residual standard error is 11750, 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))



The residual analysis of our first order linear model is nearly unreadable when analyzing the residuals vs Fitted plot. It shows an extreme right skewness with a non-linear relationship. This non-linear relationship was not explained by the model and was left out in the residuals. We will do a Box-Cox analysis to verify the necessary transformation(s).

library("MASS")  
boxcox(model1)



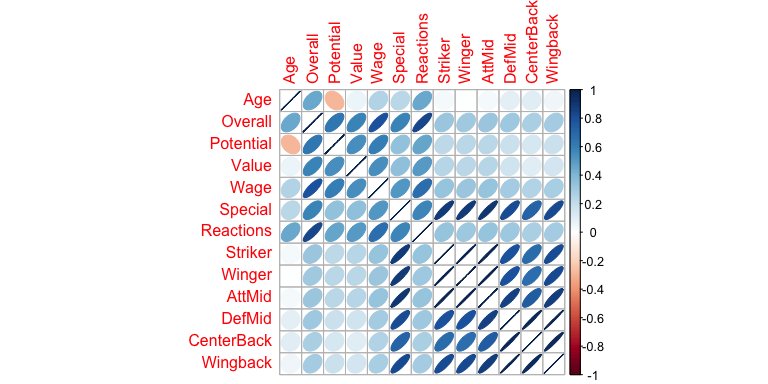
The Box-Cox analysis suggests an inverse power transformation, with λ between -0.1 and -0.2. The value zero, is just outside the 95% confidence interval, but we will try a log transformation first.

# 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  
cormat2 = cor(new\_df)  
round(cormat2, 2)

## Age Overall Potential Value Wage Special Reactions Striker  
## Age 1.00 0.45 -0.28 0.06 0.25 0.23 0.45 0.03  
## Overall 0.45 1.00 0.63 0.58 0.80 0.59 0.84 0.32  
## Potential -0.28 0.63 1.00 0.54 0.61 0.34 0.47 0.23  
## Value 0.06 0.58 0.54 1.00 0.55 0.35 0.50 0.24  
## Wage 0.25 0.80 0.61 0.55 1.00 0.52 0.67 0.34  
## Special 0.23 0.59 0.34 0.35 0.52 1.00 0.57 0.88  
## Reactions 0.45 0.84 0.47 0.50 0.67 0.57 1.00 0.33  
## Striker 0.03 0.32 0.23 0.24 0.34 0.88 0.33 1.00  
## Winger 0.01 0.31 0.23 0.24 0.32 0.89 0.32 0.99  
## AttMid 0.03 0.33 0.23 0.24 0.34 0.90 0.34 0.99  
## DefMid 0.10 0.31 0.18 0.16 0.30 0.82 0.32 0.79  
## CenterBack 0.11 0.28 0.14 0.11 0.26 0.71 0.27 0.68  
## Wingback 0.06 0.29 0.18 0.15 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 0.24 0.24 0.16 0.11 0.15  
## 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 coefficient relationship table shows a positive linear relationship between all variables except Age and Potential. The log transformation did not resolve this issue. It is then most likely the case that they are not correlated. There are still masking relationships between Attacking role positions and Defense role positions in this log-Transformed model.

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

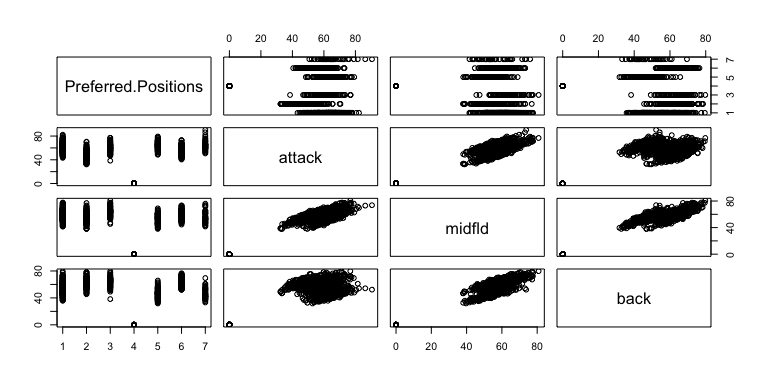
##   
## Call:  
## lm(formula = logWage ~ Age + Overall + Potential + Value + Special +   
## Reactions + Preferred.Positions + Striker + Winger + AttMid +   
## DefMid + CenterBack + Wingback)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.64182 -0.19236 -0.00063 0.18770 0.88748   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.777e-01 2.086e-01 -3.249 0.00120 \*\*   
## Age -5.564e-03 4.441e-03 -1.253 0.21058   
## Overall 5.190e-02 4.891e-03 10.612 < 2e-16 \*\*\*  
## Potential 9.178e-03 4.034e-03 2.275 0.02313 \*   
## Value 6.738e-09 2.266e-09 2.973 0.00302 \*\*   
## Special -3.353e-04 2.057e-04 -1.630 0.10351   
## Reactions -2.400e-04 2.211e-03 -0.109 0.91359   
## Preferred.PositionsCenterBack 2.767e-02 4.701e-02 0.589 0.55625   
## Preferred.PositionsDefMid -8.274e-03 4.172e-02 -0.198 0.84284   
## Preferred.PositionsGK 6.325e-01 2.239e-01 2.824 0.00484 \*\*   
## Preferred.PositionsStriker -1.442e-03 4.284e-02 -0.034 0.97315   
## Preferred.PositionsWingback 2.242e-02 3.899e-02 0.575 0.56545   
## Preferred.PositionsWinger -3.019e-02 5.434e-02 -0.556 0.57862   
## Striker 5.154e-03 9.119e-03 0.565 0.57205   
## Winger 3.018e-03 1.498e-02 0.201 0.84039   
## AttMid 1.229e-02 1.736e-02 0.708 0.47911   
## DefMid -1.535e-02 1.556e-02 -0.986 0.32418   
## CenterBack 2.142e-02 9.103e-03 2.353 0.01884 \*   
## Wingback -9.960e-03 9.617e-03 -1.036 0.30060   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3117 on 956 degrees of freedom  
## Multiple R-squared: 0.6731, Adjusted R-squared: 0.667   
## F-statistic: 109.4 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 18.083 186.1319 < 2.2e-16 \*\*\*  
## Overall 1 167.884 167.884 1728.0600 < 2.2e-16 \*\*\*  
## Potential 1 0.523 0.523 5.3792 0.0205885 \*   
## Value 1 1.101 1.101 11.3316 0.0007923 \*\*\*  
## Special 1 0.930 0.930 9.5710 0.0020342 \*\*   
## Reactions 1 0.001 0.001 0.0062 0.9374271   
## Preferred.Positions 6 0.678 0.113 1.1631 0.3238505   
## Striker 1 0.845 0.845 8.6988 0.0032618 \*\*   
## Winger 1 0.329 0.329 3.3895 0.0659215 .   
## AttMid 1 0.045 0.045 0.4635 0.4961363   
## DefMid 1 0.270 0.270 2.7778 0.0959040 .   
## CenterBack 1 0.463 0.463 4.7705 0.0291932 \*   
## Wingback 1 0.104 0.104 1.0727 0.3005994   
## Residuals 956 92.877 0.097   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The log-transformed model ANOVA table suggests that Age, Potential, Overall, Value, Special, Striker and CenterBack are significant predictors of player's Wages. The coefficient tests suggest that Potential, Overall, Value Preferred.PositionsGK and CenterBack are significant predictors. The R-Squared is 0.6731, with adjusted R-squared = 0.667, which indicate that most of the variability in the Wage is explained by this model.

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



We chose to summarize player's positions into three variables as we proved that the previous variables were masking each other.

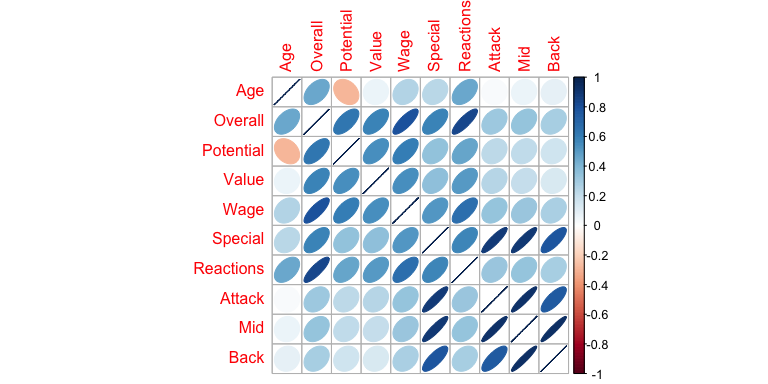
new\_df2 = new\_df[,1:7]  
new\_df2[8] = attack  
new\_df2[9] = midfld  
new\_df2[10] = back  
colnames(new\_df2)[8] <- "Attack"  
colnames(new\_df2)[9] <- "Mid"  
colnames(new\_df2)[10] <- "Back"  
  
head(new\_df2, n= 10)

## Age Overall Potential Value Wage Special Reactions Attack Mid  
## 1 25 64 67 550000 3.301030 1749 59 60.50 62.0000  
## 2 27 70 70 1900000 3.954243 1901 62 69.25 64.0625  
## 3 33 68 68 575000 4.176091 1715 65 65.00 52.3125  
## 4 35 62 62 130000 3.000000 1701 68 59.00 59.0625  
## 5 19 67 81 1600000 4.623249 1675 63 67.50 56.1250  
## 6 28 71 71 2400000 3.954243 1824 68 68.00 58.1875  
## 7 18 55 68 160000 3.000000 1504 55 52.00 51.6250  
## 8 37 69 69 325000 3.000000 1916 70 65.25 67.4375  
## 9 23 67 73 1000000 3.778151 1576 65 65.25 51.3750  
## 10 30 76 76 7500000 4.929419 1897 77 73.50 61.1250  
## Back  
## 1 59.25  
## 2 61.25  
## 3 43.75  
## 4 55.50  
## 5 44.25  
## 6 55.25  
## 7 47.25  
## 8 66.75  
## 9 40.25  
## 10 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 0.06 0.25 0.23 0.45 0.02 0.07  
## Overall 0.45 1.00 0.63 0.58 0.80 0.59 0.84 0.31 0.33  
## Potential -0.28 0.63 1.00 0.54 0.61 0.34 0.47 0.23 0.22  
## Value 0.06 0.58 0.54 1.00 0.55 0.35 0.50 0.24 0.21  
## Wage 0.25 0.80 0.61 0.55 1.00 0.52 0.67 0.33 0.33  
## Special 0.23 0.59 0.34 0.35 0.52 1.00 0.57 0.89 0.89  
## Reactions 0.45 0.84 0.47 0.50 0.67 0.57 1.00 0.32 0.34  
## Attack 0.02 0.31 0.23 0.24 0.33 0.89 0.32 1.00 0.93  
## Mid 0.07 0.33 0.22 0.21 0.33 0.89 0.34 0.93 1.00  
## Back 0.08 0.29 0.16 0.13 0.27 0.78 0.29 0.75 0.94  
## Back  
## Age 0.08  
## Overall 0.29  
## Potential 0.16  
## Value 0.13  
## Wage 0.27  
## Special 0.78  
## Reactions 0.29  
## Attack 0.75  
## Mid 0.94  
## Back 1.00

corrplot(cormat3, method = "ellipse")

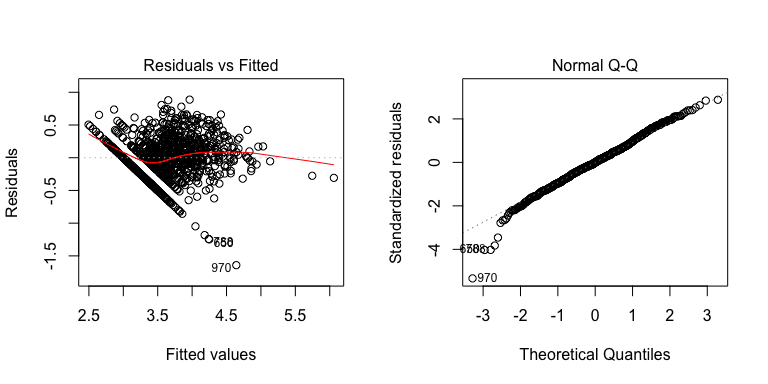


#pairs(cbind.data.frame(log(Wage), Overall, Potential, attack, midfld, back))

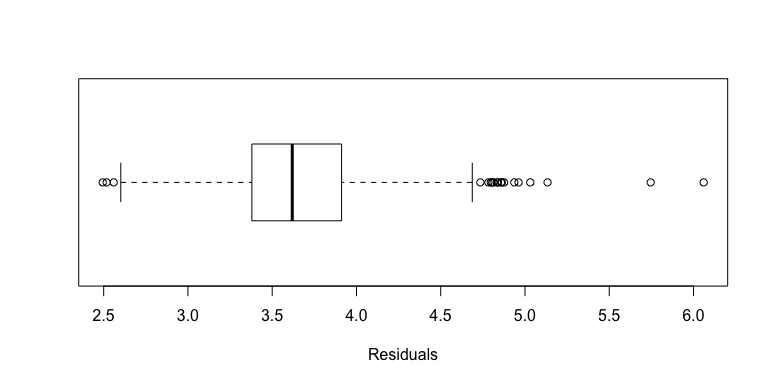
We then plotted the pairs to see their relation with log(Wage).

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

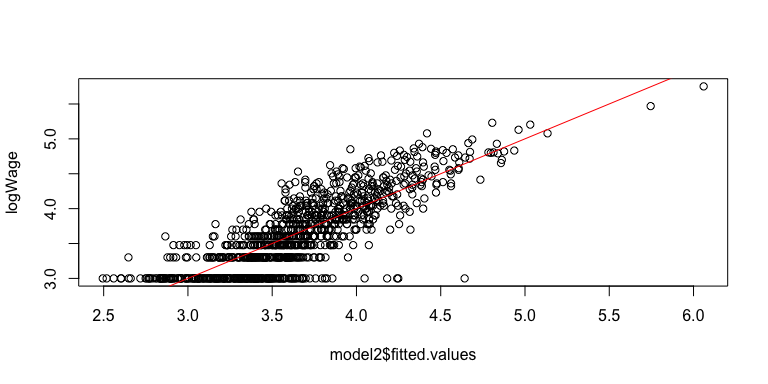
par (mfrow = c(1, 2))  
plot(model2,which = c(1,2))



boxplot(model2$fitted.values, horizontal = T, xlab="Residuals")



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



The residual analysis of the log-transformed model looks better than in the first order model. The abline is 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.

# Box-Cox Optimal Transformation.

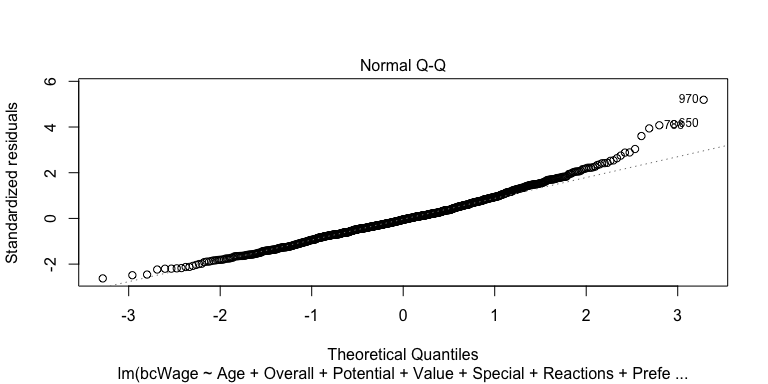
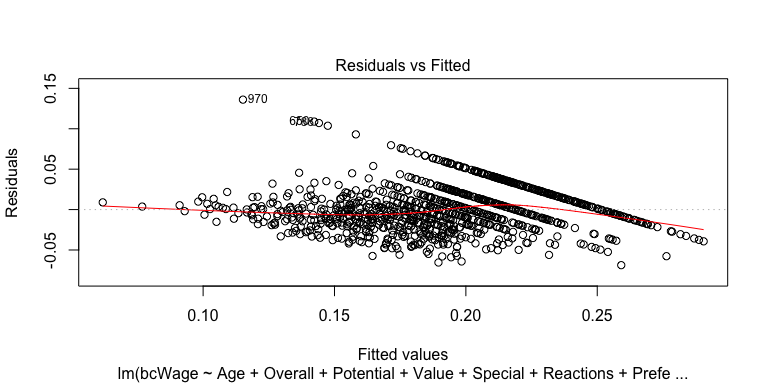
par (mfrow = c(1,2))  
bcWage = Wage^(-0.2)  
model3 = lm (bcWage~Age+Overall+Potential+Value+Special+Reactions+Preferred.Positions+Striker+Winger+AttMid+DefMid+CenterBack+Wingback)  
summary(model3)

##   
## Call:  
## lm(formula = bcWage ~ Age + Overall + Potential + Value + Special +   
## Reactions + Preferred.Positions + Striker + Winger + AttMid +   
## DefMid + CenterBack + Wingback)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.068769 -0.017055 -0.001406 0.015297 0.136033   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.640e-01 1.777e-02 31.748 <2e-16 \*\*\*  
## Age 5.620e-04 3.783e-04 1.486 0.1377   
## Overall -4.771e-03 4.166e-04 -11.452 <2e-16 \*\*\*  
## Potential -6.392e-04 3.436e-04 -1.860 0.0631 .   
## Value 1.988e-10 1.930e-10 1.030 0.3034   
## Special 2.728e-05 1.752e-05 1.557 0.1199   
## Reactions 4.383e-05 1.883e-04 0.233 0.8160   
## Preferred.PositionsCenterBack -3.443e-03 4.004e-03 -0.860 0.3901   
## Preferred.PositionsDefMid -4.672e-04 3.554e-03 -0.131 0.8954   
## Preferred.PositionsGK -4.817e-02 1.907e-02 -2.525 0.0117 \*   
## Preferred.PositionsStriker -4.575e-04 3.648e-03 -0.125 0.9002   
## Preferred.PositionsWingback -1.522e-03 3.321e-03 -0.458 0.6469   
## Preferred.PositionsWinger 2.519e-03 4.628e-03 0.544 0.5864   
## Striker -2.845e-05 7.767e-04 -0.037 0.9708   
## Winger -7.087e-04 1.276e-03 -0.555 0.5788   
## AttMid -1.007e-03 1.479e-03 -0.681 0.4961   
## DefMid 1.364e-03 1.326e-03 1.029 0.3038   
## CenterBack -1.938e-03 7.753e-04 -2.499 0.0126 \*   
## Wingback 1.019e-03 8.191e-04 1.244 0.2138   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.02655 on 956 degrees of freedom  
## Multiple R-squared: 0.6507, Adjusted R-squared: 0.6442   
## F-statistic: 98.96 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 193.8263 < 2.2e-16 \*\*\*  
## Overall 1 1.09041 1.09041 1547.2385 < 2.2e-16 \*\*\*  
## Potential 1 0.00170 0.00170 2.4124 0.120705   
## Value 1 0.00046 0.00046 0.6546 0.418658   
## Special 1 0.00666 0.00666 9.4544 0.002166 \*\*   
## Reactions 1 0.00002 0.00002 0.0271 0.869192   
## Preferred.Positions 6 0.00594 0.00099 1.4041 0.209876   
## Striker 1 0.00611 0.00611 8.6633 0.003325 \*\*   
## Winger 1 0.00107 0.00107 1.5131 0.218977   
## AttMid 1 0.00012 0.00012 0.1638 0.685798   
## DefMid 1 0.00146 0.00146 2.0672 0.150826   
## CenterBack 1 0.00369 0.00369 5.2406 0.022283 \*   
## Wingback 1 0.00109 0.00109 1.5477 0.213774   
## Residuals 956 0.67374 0.00070   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

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.

For the final report we plan on removing the goal keepers because they are causing outliers due to absency of in-line player's attributes, which skews the data. Also, we plan on enhancing our model by utilizing the log of Value.