What factors influence the overall rating of soccer players in the FIFA video game*

Yizhuo Liu, Leo Cheng, Haobo Ren

October-02-2024

Contributions

Liu Yizhuo: Data cleaning, introduction
Leo Cheng: Preliminary result
Haobo Ren: Data description

1 Introduction

In the long history of soccer, there have been many ways of evaluating the performance of soccer players. Whether by the number of goals they score or the number of trophies they earn, but these are very unilateral methods that only favor a certain kind of player, not everyone, especially when it comes to unknown players. Certainly, the video game franchise FIFA comes to mind. Every year FIFA evaluates professional soccer player's previous year's performance and gives them a rating from 0 to 100. From 1993 to now, there has been a new FIFA game every single year, yet no one knows what factors influence the rating from FIFA that the majority of the people agree on. Therefore, the main objective of this report and the research question is going to be "What factors influence the overall rating of soccer players in the FIFA video game". If we can accomplish this objective, then we can use it to quantify players' performance. Therefore, use it to predict further performance and other stuff related to performance. With our research question in mind, we proposed a hypothesis: factors such as shooting, wage, international reputation, passing, and physics will increase with the response variable - FIFA rating. To back up our hypothesis, the article "Predict the Value of Football Players Using FIFA Video Game Data and Machine Learning Techniques" (Al-Asadi and Tasdemır, 2022), which is on using FIFA to get a data-driven approach to player valuation, also uses very similar predictors that we have chosen, and the base model they end up using is also linear regression. In addition, in "PlayeRank: Data-driven Performance Evaluation and Player Ranking in Soccer via a Machine Learning Approach" (Pappalardo et al., 2019) they use real-world data to quantify players' performance with a reasonable success. Last but not least, the article "Predicting the Future Performance of Soccer Players" (Arndt and Brefeld, 2016), uses a combination of linear regression and multitask regression, to evaluate current player performance and predict the future to predict the outcome of a soccer game. Two out of the three articles have chosen linear regression, therefore it shows that linear regression is commonly used in this area. Also, by looking at the scatter plot for our response variables and predictions, there is a linear relationship. By using linear regression method we are looking to get an accuracy prediction of FIFA rating.

2 Data Description

The data of players in FIFA20 was originally found on Kaggle. The data of the players was collected by EA throughout the year based the players' performance in game.

*Code and data are available at: https://github.com/HaoboRrrr/FIFA20_Player_Potential_Rating_Analysis/tree/main

The column selected is overall, shooting, passing, physic, wage and international_reputation. We eliminated goal keepers and randomly selected 30% of observations, and replaced 0 salary with 1 to analyse.

Responsive Varible

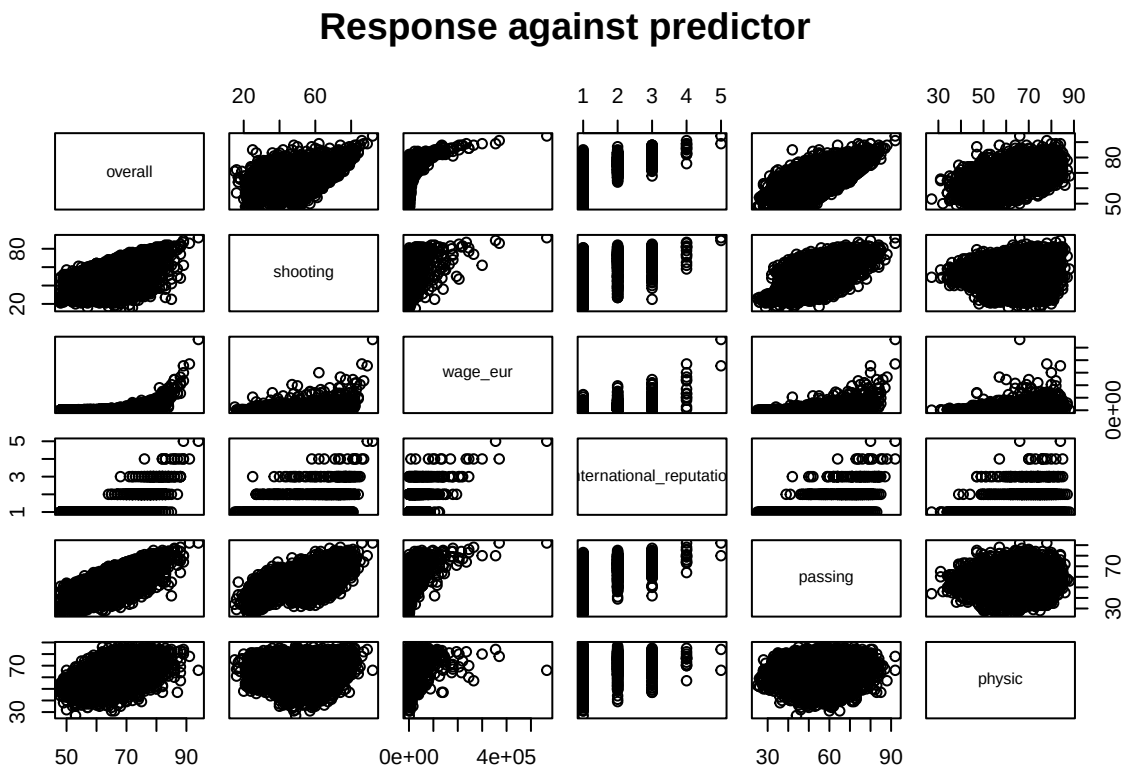
The response used in this research was "overall", it is the overall attribute of one player. This variable measures the players overall rating in the game, ranging from 0 to 100.

Table 1: Descriptive Statistic of Response Variable

min	max	mean	median	standard_deviation
48	94	66.43309	66	6.819944

The mean of the overall is 66.43

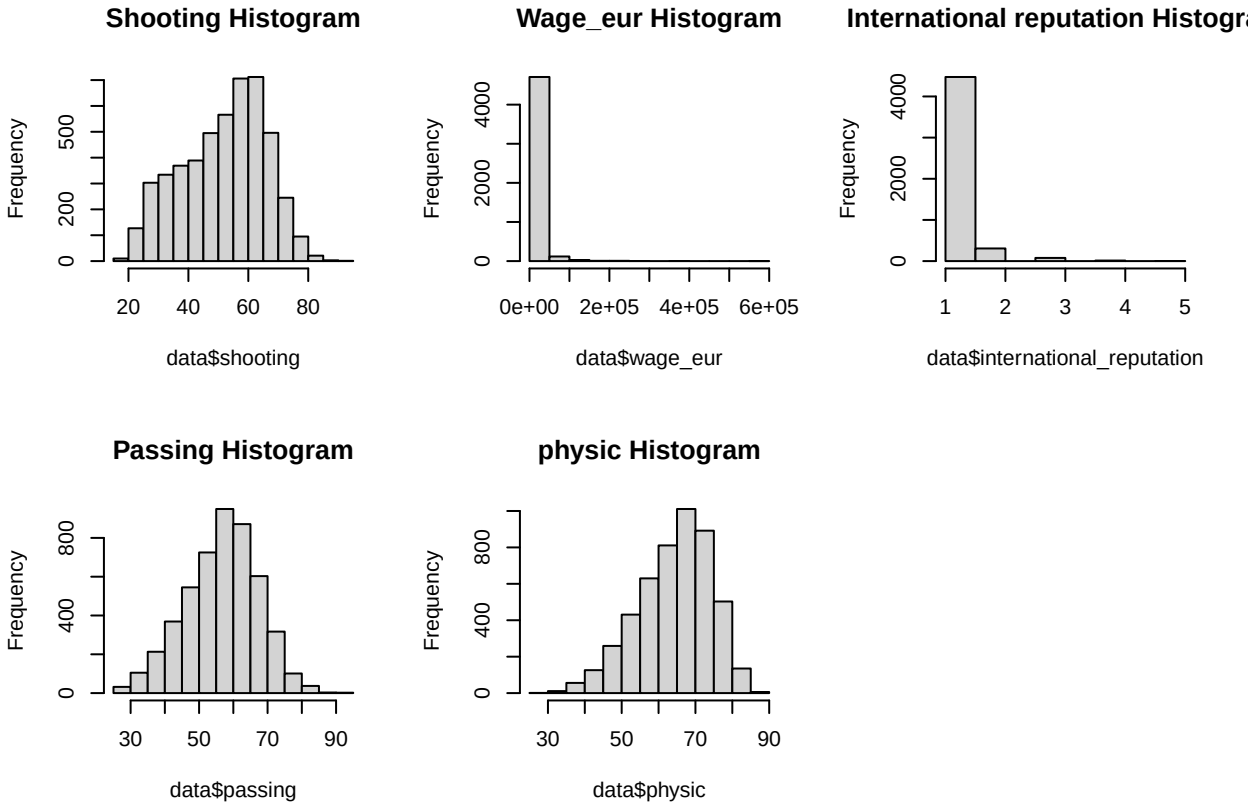
We believe the overall of a player is influenced by their attributes which we selected. As our understanding to soccer,player's overall rating should be higher if the player gets higher score in these attribute. We think the changes in predictors should result in proportional changes in the response variable. Therefore, we take overall as the responsive variable.



The response variable has an approximately linear relationship with some of the predictor variables. Meaning that as the predictors change, the response variable changes in a consistent, straight-line fashion. The international_reputation was originally a continuous predictor ranging from 1 to 5, we changed it to categorical predictor with 4 levels. ## Predictors

Table 2: Descriptive Statistic of Predictors

predictors	min	max	mean	median	standard_deviation
shooting	15	92	52.176108	54	1.399594e+01
wage_eur	0	565000	9769.088670	3000	2.238750e+04
international_reputation	1	5	1.104269	1	3.820632e-01
passing	25	92	57.128284	58	1.055442e+01
physic	27	88	64.792693	66	9.850584e+00



Attributes(predictor) of overall: * shooting: shooting ability, include strength, accuracy etc. * wage_eur: weekly salary in Euro. * international_reputation: popularity in the world. * passing: passing ability, include strength, accuracy etc. * physic: physical condition, include stamina, jumping etc.

Through the graph we can see that the distribution of passing is approximately normal, physic and shooting are slightly left skewed, international reputation and wage are extremely right skewed.

Each of these predictor contributes to the construction of the overall rating, a higher score means the player is better at this aspect, which should make the overall rating higher.

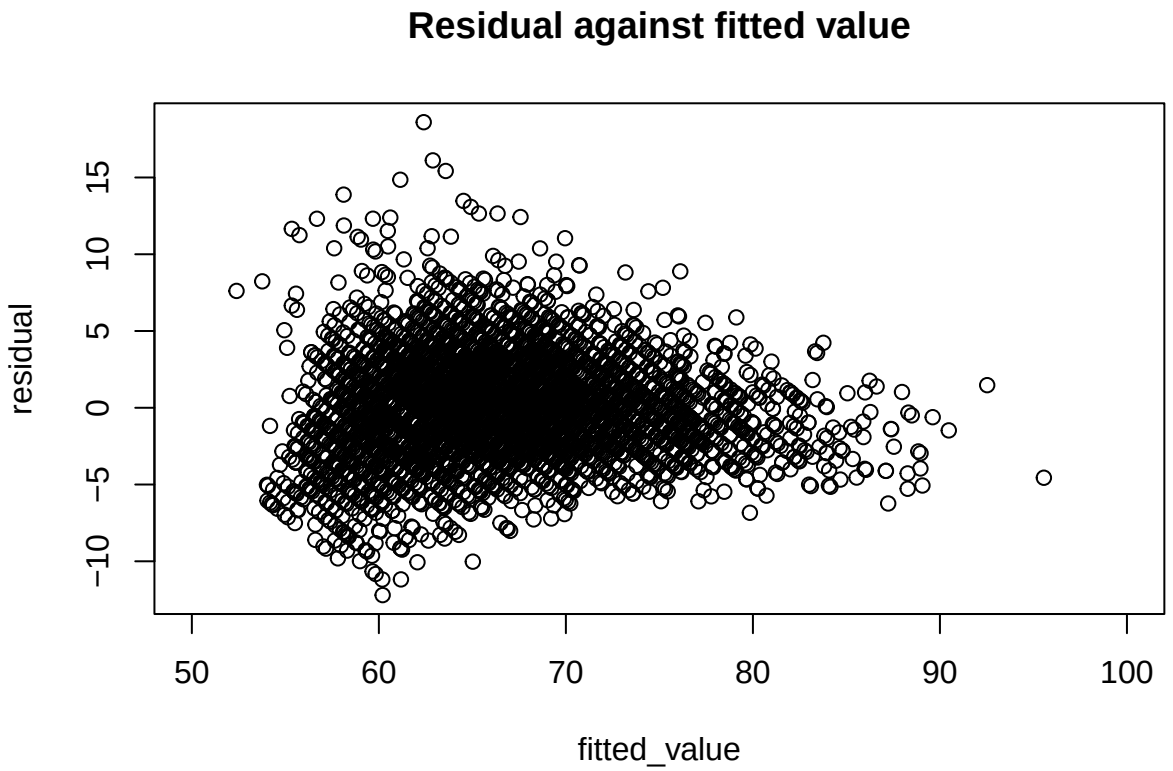
Preliminary Results

```
## bcPower Transformations to Multinormality
##                         Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## overall                    2.4109        2.41       2.2453       2.5765
## shooting                   1.3688        1.37       1.2734       1.4642
## wage_eur                   0.2149        0.21       0.2049       0.2248
```

```
## international_reputation  -15.2877      -15.29     -15.7175     -14.8579
## passing                      1.9380        2.00       1.8260       2.0501
## physic                       2.2949        2.29       2.1432       2.4465
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                       LRT df       pval
## LR test, lambda = (0 0 0 0 0 0) 20116.34  6 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                       LRT df       pval
## LR test, lambda = (1 1 1 1 1 1) 34743.88  6 < 2.22e-16
```

Inspired by (Al-Asadi and Tasdemır, 2022) and our understanding of soccer, we chose 5 predictors: shooting, wage, international reputation, passing and physics to fit a preliminary linear regression model. As usual, we produce several graphs on residuals, fitted values and QQ plots. First, the residuals and the above data visualization shows that there are 3 predictors that are inherently not linear, namely wage, physics and passing. We utilize the Box-Cox method to determine the transformation on the three predictors: exponential to 0.25, 2, 2 respectively.

Model Assumptions




```
##
## Call:
## lm(formula = overall ~ shooting + wage_eur_transformed + international_reputation +
##     passing_transformed + physic_transformed, data = transform_data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.202  -2.118  -0.043   2.015  18.601
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.223e+01  2.540e-01 166.240  < 2e-16 ***
## shooting                    3.694e-02  4.446e-03   8.309  < 2e-16 ***
## wage_eur_transformed        7.816e-01  2.141e-02  36.499  < 2e-16 ***
## international_reputation2   1.463e+00  2.125e-01   6.885 6.51e-12 ***
## international_reputation3   2.575e+00  3.999e-01   6.440 1.31e-10 ***
## international_reputation4   3.028e-01  9.283e-01   0.326    0.744
## international_reputation5  -1.853e+00  2.326e+00  -0.797    0.426
## passing_transformed         2.261e-03  5.772e-05  39.178  < 2e-16 ***
## physic_transformed          1.880e-03  4.050e-05  46.413  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.254 on 4863 degrees of freedom
## Multiple R-squared:  0.7728, Adjusted R-squared:  0.7724
## F-statistic:  2067 on 8 and 4863 DF,  p-value: < 2.2e-16
```

The first graph is the residual graph. Overall, the plot does not have a cluster, showing that the uncorrelation assumption between errors is satisfied. Also, the plot does not have a linear trend, which implies that the linearity assumption is fulfilled. However, there is a little bit of a shrinking trend along the fitted value, showing that the variance may not be necessarily constant, which is a problem we shall deal with.

The response v.s. fitted value ($\hat{y}$ v.s. y) graph should show a nearly linear relation (i.e. $y = x$ line on the graph). In our case, the points are accumulating around the line, showing a clear trend, so the linear assumption is preserved.

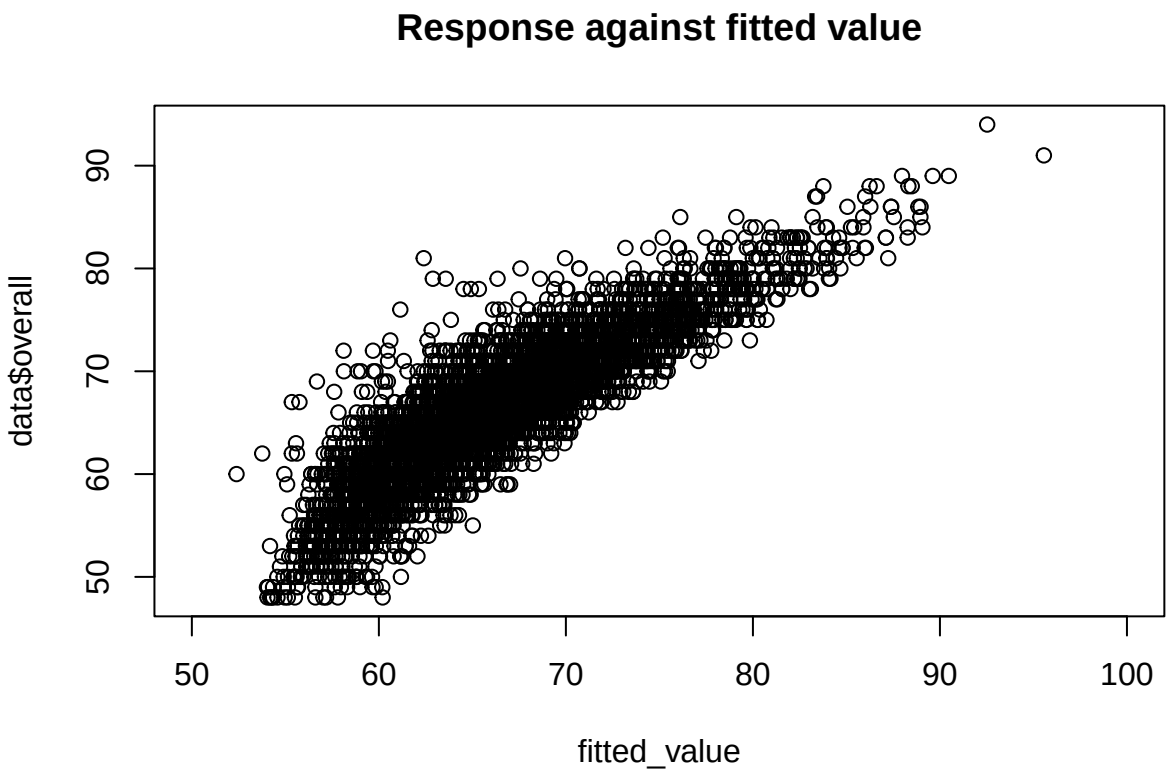
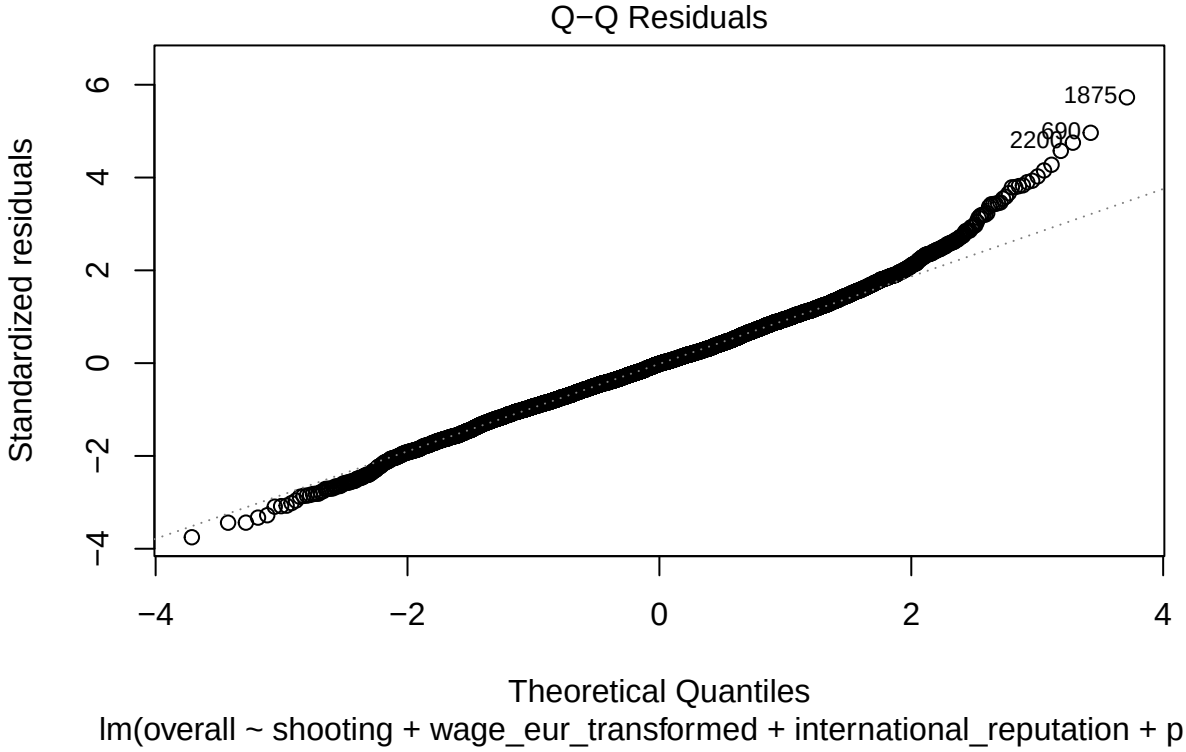


Figure 1: Response against fitted value test



QQ plot is used to determine whether the errors are normally distributed. The QQ plot shows that after transformations mentioned above, the errors are normal, since the points are showing approximately a straight line passing through 0.

To summarize, most of the assumptions are satisfied in our settings, so linear regression is a suitable model to use in this task. In (Al-Asadi and Tasdemır, 2022), the author mentions a correlation with different predictors and ranking and its preliminary result on multi-linear regression using market value as a response. The robust result from our fitting corresponds to the relatively high accuracy of the linear regression model in the paper.

References