THE MYSTERY OF PLAYER VALUE

5/8/2023

ABSTRACT

Developing a soccer player is hard and developing a soccer player in FIFA 21 is also incredibly difficult. Are a player’s wage, release clause, pace, shooting, passing, and defending a good indicator of a player’s value.

Ikenna Atum

The Mystery of Player Value

A multiple linear regression analysis to determine if a player's wage, release clause, pace, shooting, passing, and defending are a good indicator of a player's value.

Introduction

FIFA 21 is a soccer video game based on its real-life counterpart. In the game players can take advantage of a variety of in game modes including training sessions, king of the hill, volta football, career mode etc. For the analysis of this paper, the players being used in the manager mode will be at the center of the discussion. In the career mode, players can choose to be either a player or a manager. Based on the player's choice, the game will simulate them from the perspective of the player or the manager, while giving the FIFA player control of all the in-game features. Playing FIFA in manager mode I would often run into the same issue, getting good players. The only two ways to get good players in the game is through scouts getting prospects from different sides of the world or buying players from the transfer market. With the transfer market being too much of a premium, the latter choice became the default option. With one problem addressed a new problem came to take its place. With an abundance of new prospective talent coming to my in-game club, the new question became "how do I make these guys good?" More specifically "what attributes can be used to increase the overall value (level) of the players that are being brought into the club?"

The goal of this paper is to assess whether the variables wage, release clause, pace, shooting, passing, and defending impact the value of a player in the game.

Data Description

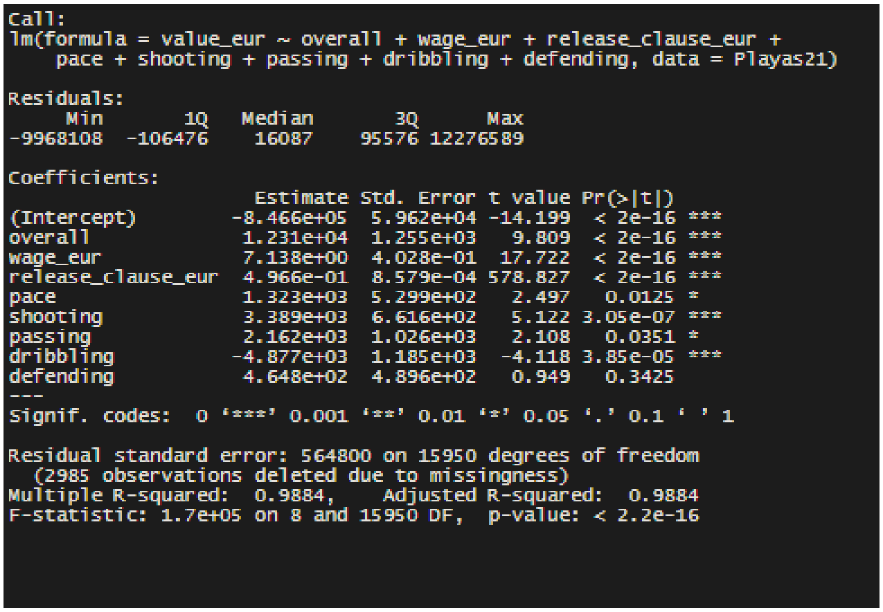
The data from this analysis originates from Kaggle. The data ranges from the games FIFA 15 all the way to game FIFA 21. For this analysis the dataset from FIFA 21 was the only one utilized. The dataset also contains 100+ attributes for all the players in the game. These attributes include attacking crossing, shooting dribbling, passing, physic, preferred foot etc. The dataset also contained a variety of non-skilled related attributes such as player value, wages, contract release clause, nationality, team position etc. Though there were many variables to choose from, only a few predictors were chosen for this analysis. The predictors chosen were,

1. Wage_eur – Which is the wage the player was making at the time of the game release in euros.
2. Release_Clause_eur – Which is the amount of money the club set for players contract to be released in euros.
3. Pace – The speed of a player at the time of the game being released.
4. Shooting – The shot quality of the player at the time of the game being released.
5. Passing – The passing ability of the player at the time that the game was released,
6. Defending – The players ability to defend at the time of the game being released.

The response variable for this analysis is the variable value_eur, which represents the value of the player in Euros at the time of the game release. The dimensions of the dataset were 134 columns by 189444 rows. Because of the number of variables and the size of the dataset, coming up with a reliable model was difficult. This will be discussed more in depth later in the paper.

Methods and Results

Finding the final model for this analysis was not a simple process. The first model chosen for this analysis was "(lm(value_eur ~ overall + wage_eur + release_clause_eur + pace + shooting + passing + dribbling + defending, data = Playas21)". The model seemed to be fine at first. All the variables were statistically significant apart from defending. This went against the intuition behind the project. Defense is a huge part of the game, so a player being a good defender could impact their value. Upon reviewing the residuals versus predictor graph, issues with non- constant variance and independence were discovered.



```
Call:
lm(formula = value_eur ~ overall + wage_eur + release_clause_eur +
    pace + shooting + passing + dribbling + defending, data = Playas21)

Residuals:
     Min       1Q   Median       3Q      Max
-9968108  -106476    16087    95576 12276589

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -8.466e+05  5.962e+04 -14.199  < 2e-16 ***
overall             1.231e+04  1.255e+03   9.809  < 2e-16 ***
wage_eur            7.138e+00  4.028e-01  17.722  < 2e-16 ***
release_clause_eur  4.966e-01  8.579e-04 578.827  < 2e-16 ***
pace                1.323e+03  5.299e+02   2.497   0.0125 *
shooting            3.389e+03  6.616e+02   5.122 3.05e-07 ***
passing             2.162e+03  1.026e+03   2.108   0.0351 *
dribbling          -4.877e+03  1.185e+03  -4.118 3.85e-05 ***
defending           4.648e+02  4.896e+02   0.949   0.3425
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 564800 on 15950 degrees of freedom
  (2985 observations deleted due to missingness)
Multiple R-squared:  0.9884,    Adjusted R-squared:  0.9884
F-statistic: 1.7e+05 on 8 and 15950 DF,  p-value: < 2.2e-16
```

Figure 1: This figure shows the results of the regression for the first iteration of the regression. The R-squared here is 0.98.

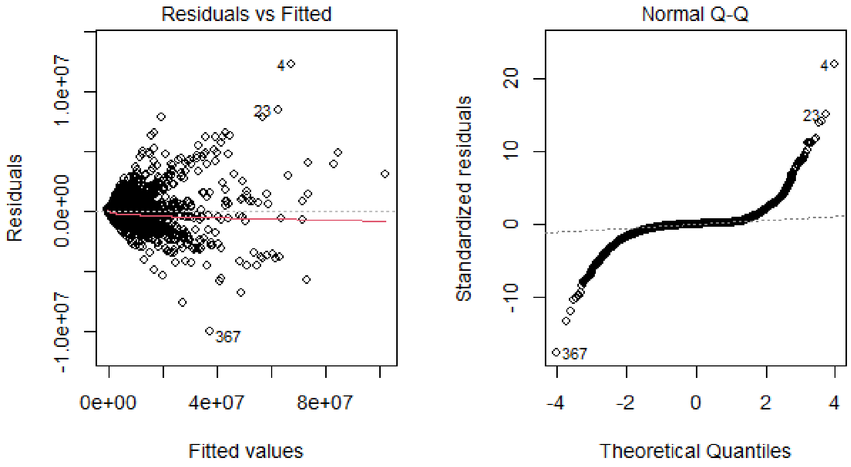
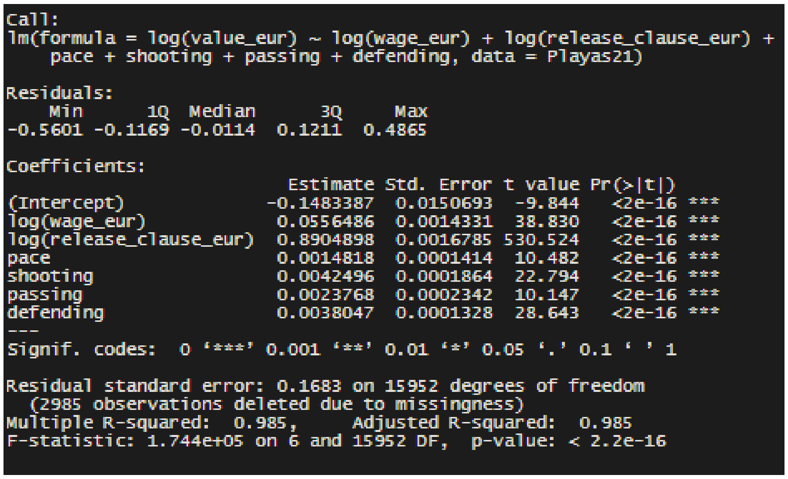


Figure 2: This figure shows the Residuals versus Fitted plot as well as the QQ plot for the regression in Figure 1. The Residuals versus Fitted plot is having issues with non- constant variance and independence.

The other issue with the original model was multicollinearity. A few of the predictors, namely dribbling and overall rating, were highly correlated with the other predictors. To correct these issues transformations were explored and predictors were dropped from the model. In the final model the variables dribbling and overall, where dropped, since they each had the highest VIFs. The response variable value_eur and the two predictors release clause and wage_eur underwent a log transformation to correct the normality, independence, and non-constant variance issues. Then after testing various models this became the final version “lm(log(value_eur) ~ log(wage_eur) + log(release_clause_eur) + pace + shooting + passing + defending, data = Playas21)”. The final estimated model became $\log\left(Value_{eur}\right)\sim -0.14483 + 0.05564 * \log\left(Wage_{eur}\right) + 0.8904898 * \log\left(Release_{Clause_{Eur}}\right) + 0.00014818 * Pace + 0.0042496 * Shooting + 0.0023768 * Passing + 0.0038047 * Defending + \varepsilon$. With a each of the coefficients being interpreted as follows:

1. For every one-unit change in log(wage_eur), log(value_eur) will change by 0.05564.
2. For every one-unit change in log(release_clause_eur), log(value_eur) will change by 0.8904898.
3. For every one-unit change in pace, log(value_eur) will change by 0.00014818.
4. For every one-unit change in shooting, log(value_eur) will change by 0.0042496.
5. For every one unit change in passing, log(value_eur) will change by 0.0023768.
6. For every one-unit change in defending, log(value_eur) will change by 0.0038047.



```
Call:
lm(formula = log(value_eur) ~ log(wage_eur) + log(release_clause_eur) +
    pace + shooting + passing + defending, data = Playas21)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5601 -0.1169 -0.0114  0.1211  0.4865 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)             -0.1483387  0.0150693  -9.844   <2e-16 ***
log(wage_eur)            0.0556486  0.0014331  38.830   <2e-16 ***
log(release_clause_eur)  0.8904898  0.0016785 530.524   <2e-16 ***
pace                     0.0014818  0.0001414  10.482   <2e-16 ***
shooting                 0.0042496  0.0001864  22.794   <2e-16 ***
passing                  0.0023768  0.0002342  10.147   <2e-16 ***
defending                0.0038047  0.0001328  28.643   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1683 on 15952 degrees of freedom
  (2985 observations deleted due to missingness)
Multiple R-squared:  0.985,	Adjusted R-squared:  0.985 
F-statistic: 1.744e+05 on 6 and 15952 DF,  p-value: < 2.2e-16
```

Figure 3: This figure shows the results of the final estimated regression. The R-squared for the regression is 0.985.

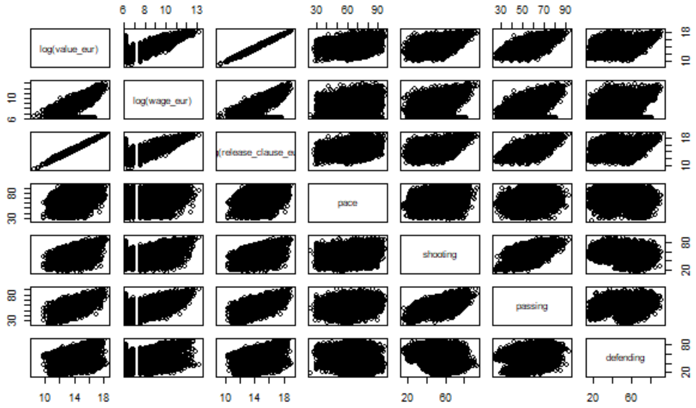


Figure 4: This figure shows the results of the scatter plot matrix. The matrix shows that all the predictors are positively correlated with the response. The figure also shows that all the predictors are loosely correlated with each other.

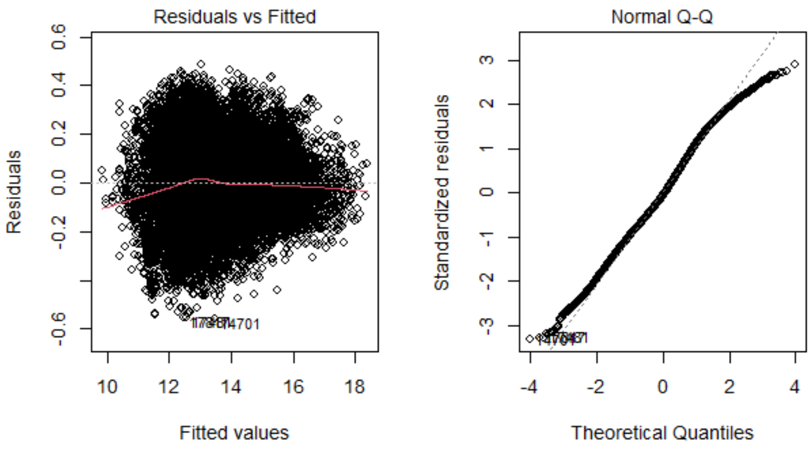


Figure 5: This figure shows the Residuals versus Fitted and QQ plot for the regression summary in Figure 3. There are no problems with constant variance, independence, and normality.

In the final model we can see that all the predictors are statistically significant at all three levels, with the contract release clause being the most significant. According to the scatter plot matrix all the predictors are positively correlated with the response. Also, according to the scatter plot matrix all the predictors are loosely correlated with each other. The residual versus fitted plot satisfies the assumptions of constant variance and independence. The QQ plot satisfies our assumption of normality.

To understand the power of this model, some predictions were run to see the difference between the observed value and the actual value of some of the players in the dataset. While undergoing the analysis a problem arose with the predictions. Players that were highly skilled and well paid had the same player value across the board. Meaning that a player such as Mo Salah who is a winger has the same value as Sadio Mane who is also a winger. This is despite the fact these players have different attributes, wages, and release clauses. This could be for a variety of reasons. One of the reasons could be after a certain value of predictors the model starts to quantify player impact and player value the same way across the board. It could also mean that more predictors are needed to differentiate between players to see which player is worth more and which are worth less.

Conclusion

In conclusion, we can reject the null hypothesis of this analysis that each of these predictors are not correlated with player value. And we can accept the alternative that the predictors wage_eur, release_clause_eur, pace, shooting, passing, and defending are correlated with a player's value. A person playing the game of FIFA on manager mode looking to develop their players can use this model as base to begin. There are limitations for this model. As previously discussed in the Data and Methods section, players that had high observed values all had the same predicted values when ran in the model. To improve upon this in the future, a model utilizing more attacking options and defending options can be created. Another improvement could be adding player positions to the model, to understand which players' positions are more valued than others. This could aid new potential managers in selecting new prospects to join their club. A final comment on improvement could come from analyzing the data itself. With FIFA being a big business, teams, players, advertising agencies etc. all have a collective interest in selling the game. Does this collectivist interest lead to inaccurate reporting of player attributes? And if that would be the case, how does this impact that value of the player in the game? To conclude, we reject the null hypothesis, and accept the alternative that these numbers are correlated with the response.

Appendix

Original Regression

```
library(readr)

Playas21 <- read_csv("C:/Users/Ikeat/Desktop/Playas21.csv")

View(Playas21)

lm1 <- lm(value_eur ~ overall + wage_eur + release_clause_eur + pace + shooting + passing + dribbling + defending, data = Playas21)



summary(lm1)
```

Residuals versus Fitted and QQ plot

```
par(mfrow = c(1,2))

plot(lm1, 1:2)
```

Final Model

```
lm1.1 <- lm(log(value_eur) ~ log(wage_eur) + log(release_clause_eur) + pace + shooting + passing + defending, data = Playas21)



summary(lm1.1)
```

Final Residuals versus Fitted and QQ plot

```
par(mfrow = c(1,2))

plot(lm1.1, 1:2)
```

Scatter Plot Matrix for Final Model

```
pairs(log(value_eur) ~ log(wage_eur) + log(release_clause_eur) + pace + shooting + passing + defending, data = Playas21)
```

Player Predictions

```
new_x <- data.frame(wage_eur = 250000   , release_clause_eur = 144300000, pace = 94, shooting = 85, passing = 80, defending = 44 , data = Playas21)



#Sadio Mane



predict(lm1.2, newdata = new_x, type = "response")

new_x8 <- data.frame(wage_eur = 250000 , release_clause_eur = 144300000, pace = 93, shooting = 86, passing = 81, defending = 45 , data = Playas21)



#Mo Salah



predict(lm1.2, newdata = new_x, type = "response")

new_x9 <- data.frame(wage_eur = 1950                    , release_clause_eur = 124400, pace = 77, shooting = 83, passing = 88, defending = 68 , data = Playas21)



#Bruno Fernades



predict(lm1.2, newdata = new_x, type = "response")
```

```
```