Player's Market Value in 'Big Five'

Andrew Shan, Ruoyuan Qian

Introduction

In this project, we aimed to explore the contributing factors of market values for the players in the five most successful football leagues in Europe (i.e., "Big Five") in the setting of FIFA 21. FIFA 21, as part of the FIFA series, is an association football simulation video game, which is developed and released annually by Electronic Arts under the EA Sports label. It is the 28th installment in the FIFA series, and was released on 9 October 2020 for Microsoft Windows, Nintendo Switch, PlayStation 4 and Xbox One. Enhanced versions for the PlayStation 5 and Xbox Series X and Series S were released on 3 December 2020, in addition to a version for Stadia in March 2021. With an official license from FIFA, the world governing body of football, the game provides comprehensive data featuring more than 30 official leagues, over 700 clubs, and over 17,000 players.

The 'Big Five' represents the five most successful football leagues in Europe, which are made up of the Premier League in England, La Liga in Spain, the Bundesliga in Germany, Serie A in Italy and Ligue 1 in France. Since 1955, the first edition of the European Cup took place during the 1955–56 season, which provided an unique competition opportunity for football clubs in Europe. In 1960, the association coefficient was introduced to rank the football associations in Europe, and thus determine the number of clubs from an association that will participate in the UEFA Champions League, the UEFA Europa League and the UEFA Europa Conference League. Since then, the Premier League, La Liga, the Bundesliga, and Serie A have won the majority of the titles, which was more than titles won by other associations combined. Though Ligue 1 in France has never won the title, it has been ranked among the top 5 ever since. The combined revenue of these five leagues, which each represents the highest tier football division in their countries, has more than doubled in the past decade, reaching a total of approximately 15.1 billion euros in 2019/20. As such, the 'Big Five' is consistently attracting talent players to joint the leagues.

In this study, we aimed to examine the contributing factors of market values for the players, which could potentially provide some insights of the key strategy in ability training for players to increase their market value. In addition, the analyses could examine the disparities in the market value across player's position and the optimal league they should join to maximize their market value.

Data

The data we used in this project was from the dataset which was originally scraped from the publicly available website `https://sofifa.com`. The original data contains 106 variables and 18,944 observations. Since our goal was to analyze contribution factors of players value in "Big Five", we only keep the players in "Spain Primera Division", "Italian Serie A", "German 1. Bundesliga", "French Ligue 1" and "English Premier League". In addition, goal keepers were removed due to the different attributes between them and other players and the subdiminision for each attribute was removed. The variable, BMI, was generated incorporating the information on player's weight and height. The team position variable was recoeded into "“Defender", "Midfielder", and "Forward." Therefore, the final sample contained a total of 2,756 observations with 16 variables including `Sofifa Id`, `Short name`, `Age`, `BMI`, `League name`, `Preferred foot`, `Team position`, `Overall`, `Pace`, `Shooting`, `Passing`, `Dribbling`, `Defending`, `Physic`, and `Dribbling`.

According to the density plot (`Appendix Figure 1`) of the response (value in Euro), the original data was heavily right skewed, while the distribution after log-transformation was more likely to be Normal. Thus, log transformed value in Euro was chosen as the response. The plot matrix for numeric variables was presneted

in `Figure 1`. The overall scores and log-value in Euro are highly correlated ($\rho = 0.972$), suggesting that overall scores was a very important predictors. Besides, the correlation between passing and dribbling was over 0.8, indicating that there might be some potential collinearity.

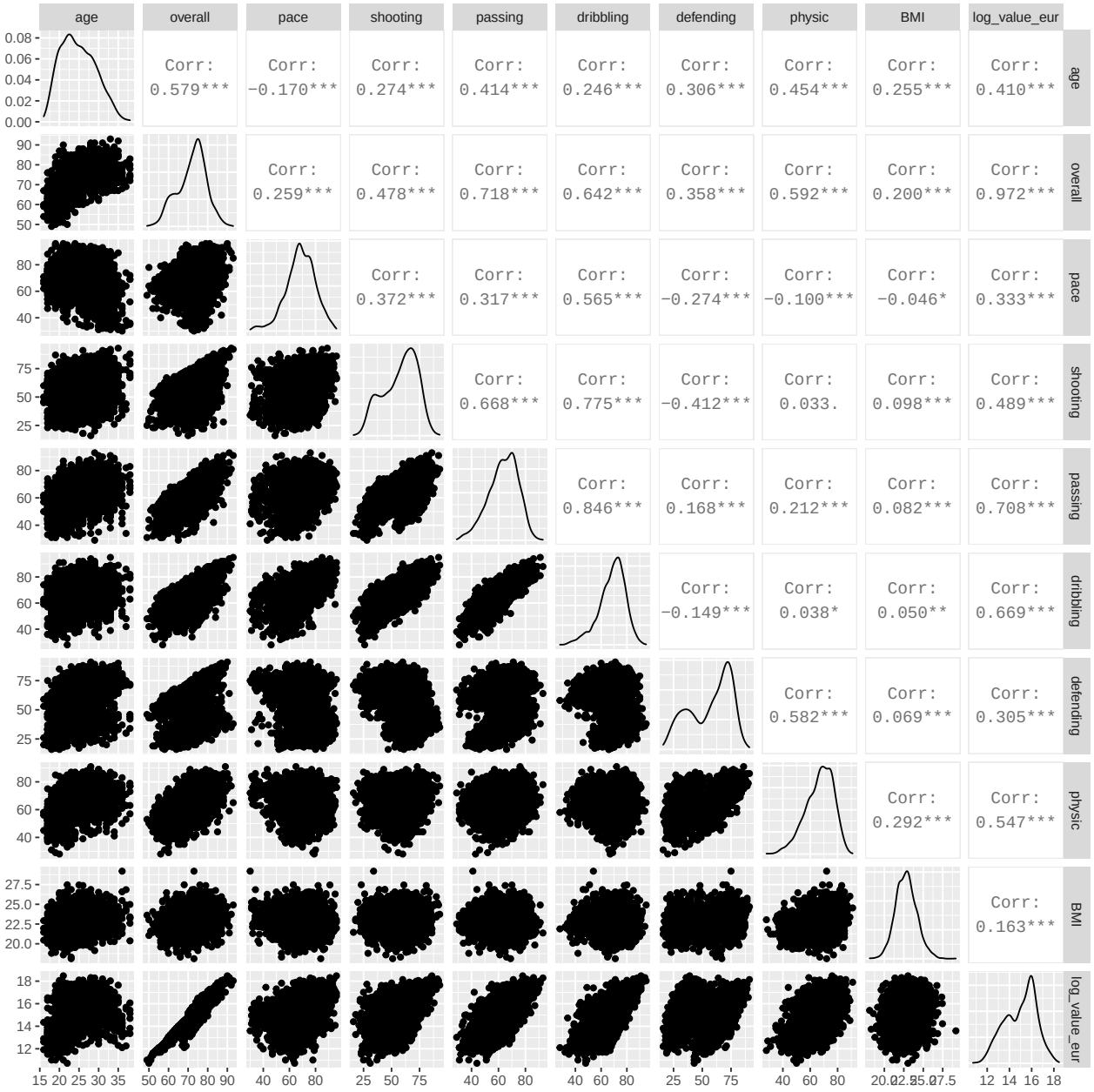

Figure 1: plot matrix for numeric variables

Methods and model building

Since the determinants of the market value was complex, a simple linear regression which regressed market value on the overall score was unlikely to solve the research problems. Thus, the methods using multiple linear regression with model selection was chosen. Specifically, the Bayesian information criterion was preferred in this case given it generally penalizes free parameters more strongly than the Akaike information criterion and penalizes the complexity of the model where complexity refers to the number of parameters in the model.

Since our sample size was much larger than the total number of potential parameters, the limitation that BIC was only valid for sample size n much larger than the number k of parameters in the model could be avoid.

The BIC is formally defined as:

$$BIC = kln(n) - 2ln(\hat{L})$$

- $\hat{L}$ is the maximized value of the likelihood function of the mode.
- x = the observed data;
- n = the number of data points in x, the number of observations, or equivalently, the sample size;
- k = the number of parameters estimated by the model.

Table 1: Summary statistics for Models

names	r.squared	adj.r.squared	AIC	BIC	sigma	df	df.residual
Null	0.944	0.944	1799.655	1817.420	0.335	1	2754
Full	0.980	0.980	-1023.252	-916.664	0.200	16	2739
BIC Forward/Stepwise	0.980	0.980	-995.085	-953.634	0.202	5	2750
BIC Backward	0.980	0.980	-1006.721	-953.427	0.201	7	2748

We used three methods, forward, backward, and stepwise, with BIC as the criterion for model selection. Given the high correlation between `market value` and `overall score` (`Figure 1`), we chose the model regressing the `log_value_eur` on `overall` as the null model for the forward selection method. The forward and stepwise methods yielded one same final model, which contains `overall`, `age`, `shooting`, and `team position`. The backward method yielded the model inlcuding `overall`, `age`, `team position`, `shooting`, `defending`, and `physic`. The summary statistics for two models were similar (`Table 1`) and two models had smaller AIC and BIC values compared to the null and full models. Around 98% of variability for market values was explained by the independent variables included in both model. Since we were unable to determine the best model based on the summary statistics, we used the 10-fold cross-validation method to evaluate two models. The root-mean-square error (RMSE) was used in evaluating the cross-validation results. The RMSE measured the differences between values predicted by a model or an estimator and the values observed. The model from backward selection was chosen due to its smaller median RMSE (0.204 vs. 0.198) (`Figure 2`).

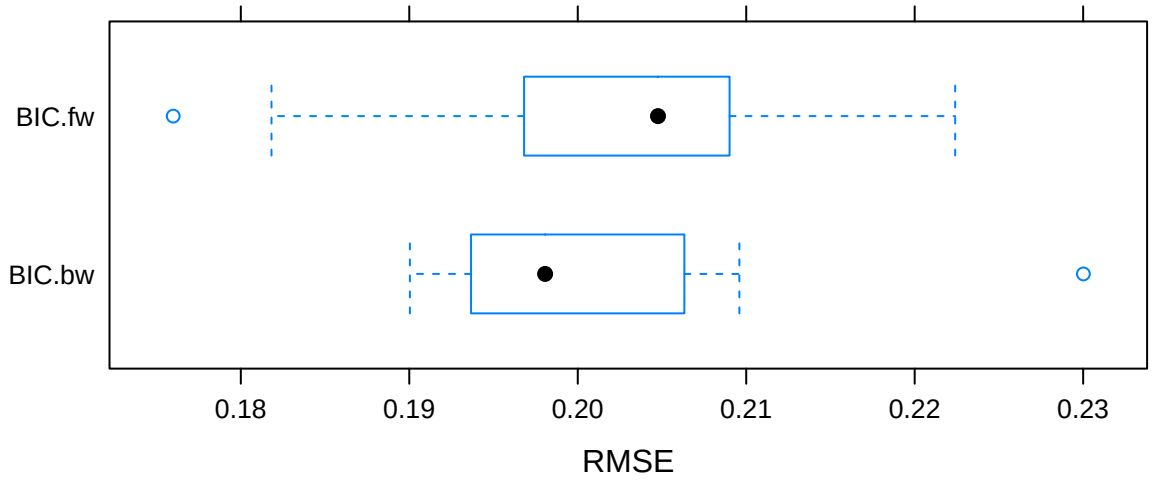

Figure 2: Distribtuion of RMSE

Therefore, the model we selected was:

$$\begin{aligned} E[Y|X] = \beta_0 &+ \beta_1 Age + \beta_2 * I(L = FR) + \beta_3 * I(L = DE) + \beta_4 * I(L = IT) + \beta_5 * I(L = SP) \\ &+ \beta_6 * overall + \beta_7 * I(P = FW) + \beta_8 * I(P = MF) + \beta_9 * shooting \\ &+ \beta_{10} * Dribbling + \beta_{11} * Defending + \beta_{12} * phsic \end{aligned}$$

Model diagnosis was processed by detecting whether there existed some high-leverage and highly influential points, as well as outliers. `Fugire 3` showed our data had relatively low leverage for all observations ($h_{ii} \leq 0.01$) and Cook's distances measuring effects of influential points were smaller than 0.03, which were much smaller than the threshold of 1. Studentized residuals were used in examining the potential outliers. Several observations were out of range of the 95% familywise confidence interval for the Studentized residuals (`Table 2`). All of them were aged over 33 years old, majority of them were defenders, and their market values were either in the lower tail or the upper tail of the distribution. In addition, their overall scores were either very high or relatively low. Given the large sample size (n=2756) and their values were not unusual, we decided to keep these data points assuming the impact of these observations was negligible. In the model diagnosis plots (`Figure 4`), we observed a left-skewed problem in the residual Q-Q plot, and the z shape was observed in the residual against fitted value plot and residual against overall score plot. By the assumption of multiple linear regression, residuals should randomly scatter around zero. Additionally, the scatterplots in `Figure 1` indicated a potential quadratic relationship between outcome variable and age. We did not observe any problems in other diagnosis plots. Thus, we added additional non-linear terms for `age` and `overall`.

Table 2: Obeservations with large absoluted studentized residuals

log_value_eu	age	league_name	overall	team_position	shooting	dribbling	defending	physic	.ti
18.02764	33	Spain Primera Division	93	Midfielder	92	95	38	65	-4.6854
17.64415	35	Italian Serie A	92	Forward	93	89	35	77	-4.6635
17.01418	34	Spain Primera Division	89	Defender	70	73	88	85	-4.7097
14.91412	36	English Premier League	80	Defender	61	61	81	83	-4.9339
13.26213	36	Spain Primera Division	72	Defender	60	69	73	65	-4.6603
13.07107	36	Italian Serie A	71	Defender	49	59	74	68	-4.4458
12.25486	37	Italian Serie A	69	Defender	39	42	71	71	-5.9963
12.10071	36	Italian Serie A	68	Defender	41	50	70	72	-6.1291

In the model including quadratic term for both `age` and `overall`, the signs of higher order terms of the variable `overall` were different from the linear term and coefficient estimate differed with 79.37 and -1.73 for the linear and quadratic, respectively. It indicated potential colinearity problems between linear and quadratic term for `overall`. Thus, we dropped the quadratic term for `overall` from the model.

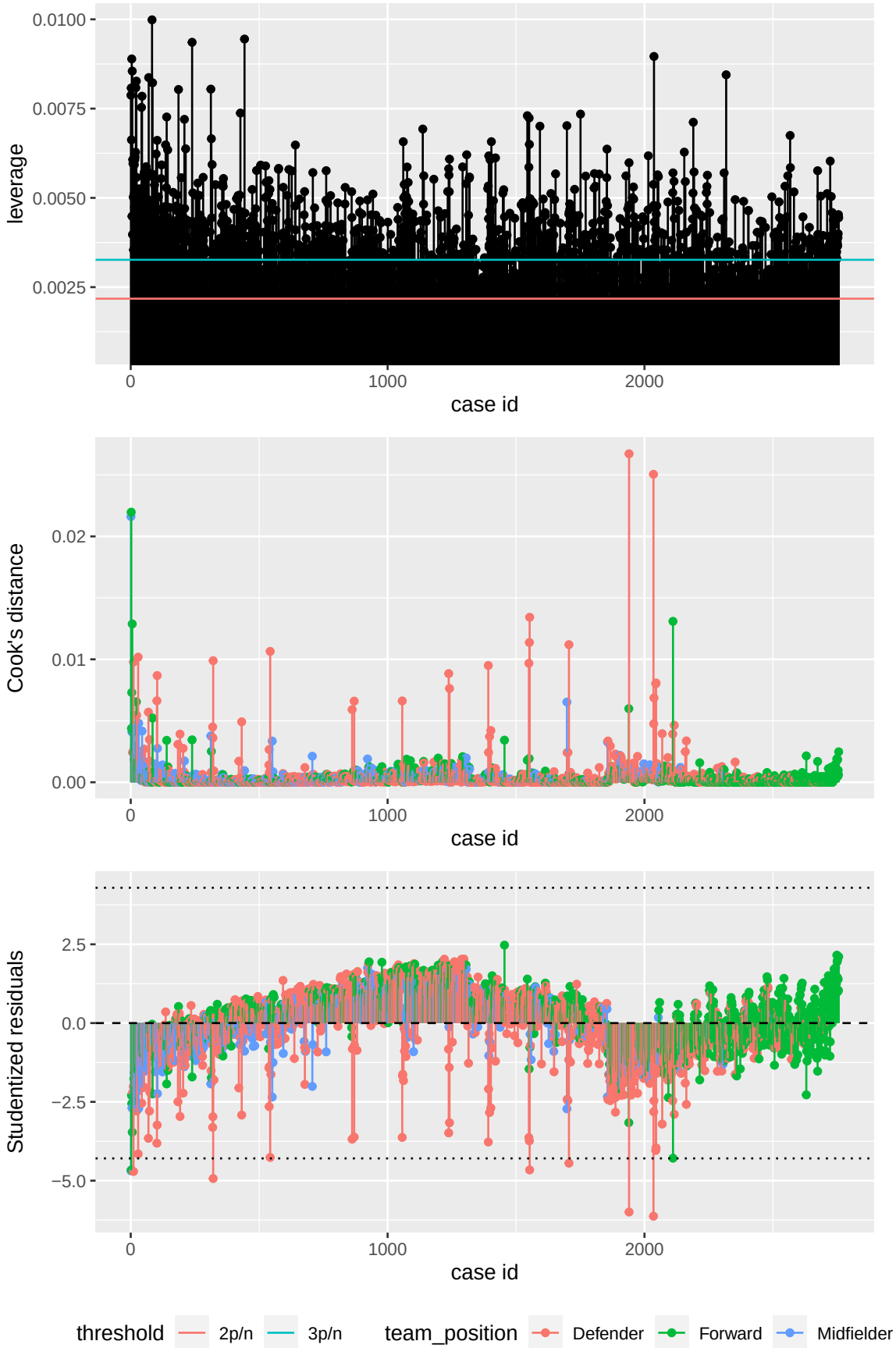

Figure 3: Model Diagnosis

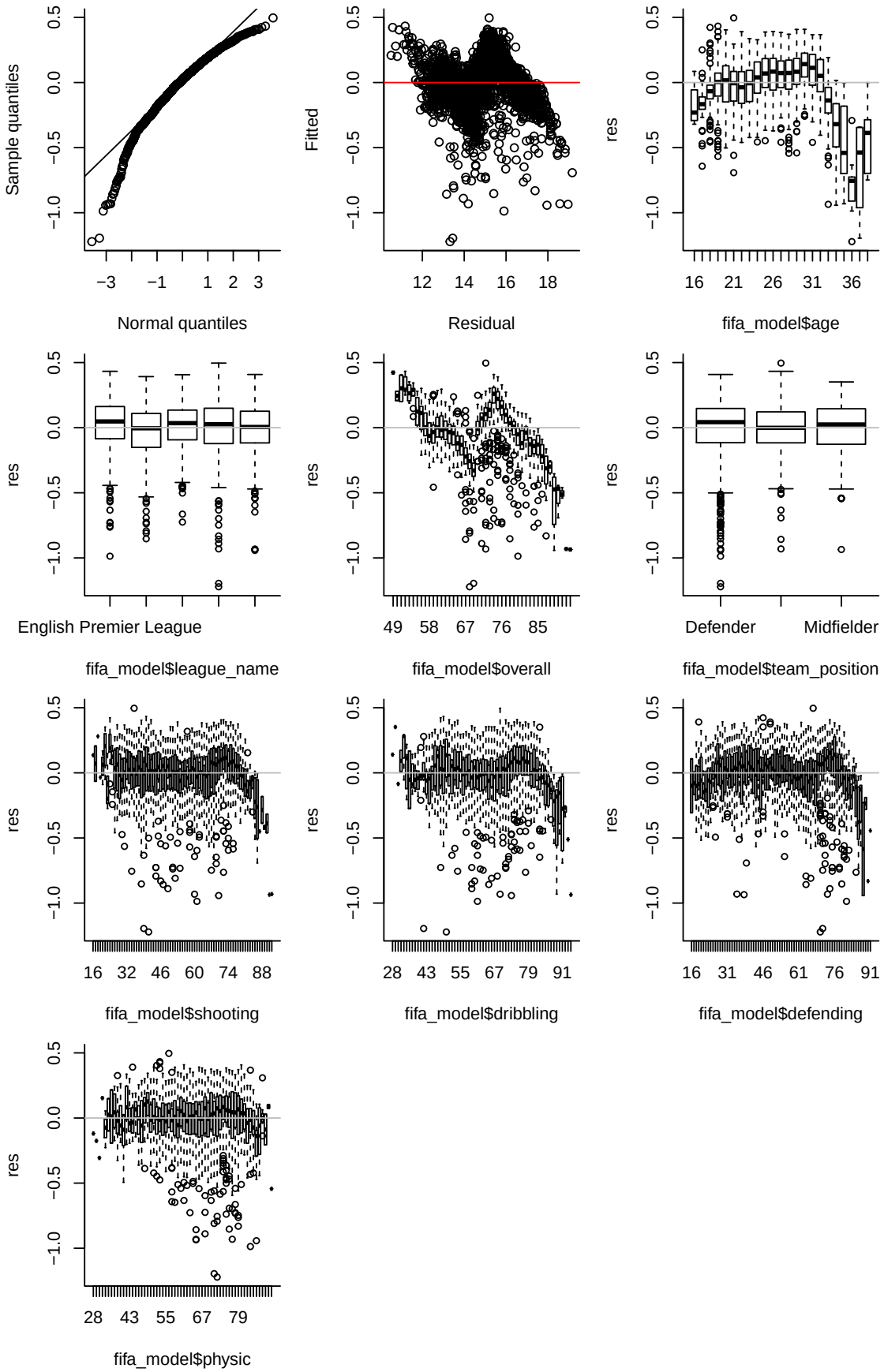

Figure 4: Model Diagnosis

Result

Final model

Let Y_i be the log-transformed player value in Euro, e_i be a R.V. and $e_i \overset{i.i.d.}{\sim} N(0, \sigma^2), i = 1, 2, ..., 2756$.

$$\begin{aligned} log(Y_i) = & 0.1587 - 15.66 * Age - 4.598 * Age^2 - 0.061 * I(L = FR) - 0.0146 * I(L = DE) \\ & - 0.0372 * I(L = IT) - 0.033 * I(L = SP) + 0.205 * Overall - 0.0254 * I(P = FW) \\ & + 0.0259 * I(P = MF) + 0.0023 * Shooting + 0.0004 * Dribbling - 0.0009 * Defending \\ & + 0.0002 * physic + e_i \end{aligned}$$

Where L denoted the categorical variable League Name, the reference was Premier League in English, FR, GE, IT, SP denoted the Ligue 1 in France, 1. Bundesliga in Germany, Serie A in Italy and Primera Division in Spain, respectively.

P denoted the categotical variable Team position, the reference was Defender, FW, MF denote the Forward and Midfielder, respectively.

Interpretation

According to the final model, all the predictors were significant except `dribbling` and `physic`. Adjusting league of the players in, overall scores, team position, shooting, dribbling, defending, physic abilities, `age` was negatively associated with the player value, which was reasonable in reality that the younger the player, the more potential he had. In addition, all the estimates of dummy variables about the league were negative, meaning that the player in Premier League in England was more valuable than the same player in other leagues. The sencond valuable league was Bundesliga in Germany, and the third one was Primera Division in Spain, the most unvaluable league in Big-Five was Ligue 1 in France. Additionally, the overall score was positively related to the player value, which followed the same pattern as the scatter plot in EDA. The midfielders had higher market values than the players in the defending and forward position. It might be because there were fewer top players who were good at playing at midfield than players at other two positions. Shooting, dribbling and physic scores were positive associated to the player value whereas the defending ability was negatively related to the response.

To sum up, our guidance for the player was that in order to increase their values, they should start playing as early as possible, join in Premier League in English, be trained as a midfielder and enhance their ability, particularly in shooting, dribbling and physic skills.

Discussion

According to the residual Q-Q plot (`Figure 5`), the normality was much better than the model without the quadratic term of age. However, as for the residual v.s. `overall` plot, there might be still a trend in overall score, and the shape of residual v.s. `overall` was very similar to residual v.s. fitted value plot, indicating the trend in overall score might affect the model residual a lot. We had tried several ways to fix the problem, adding polynomial terms of overall score and adding interaction terms for overall score and other predictors, but none of the methods were able to solve the trend problem. The trend might be explained by other potential predictors that we did not collect. Thus, based on the principle of easy interpretation, we decided to use the model without polynomial and interaction terms of overall scores.

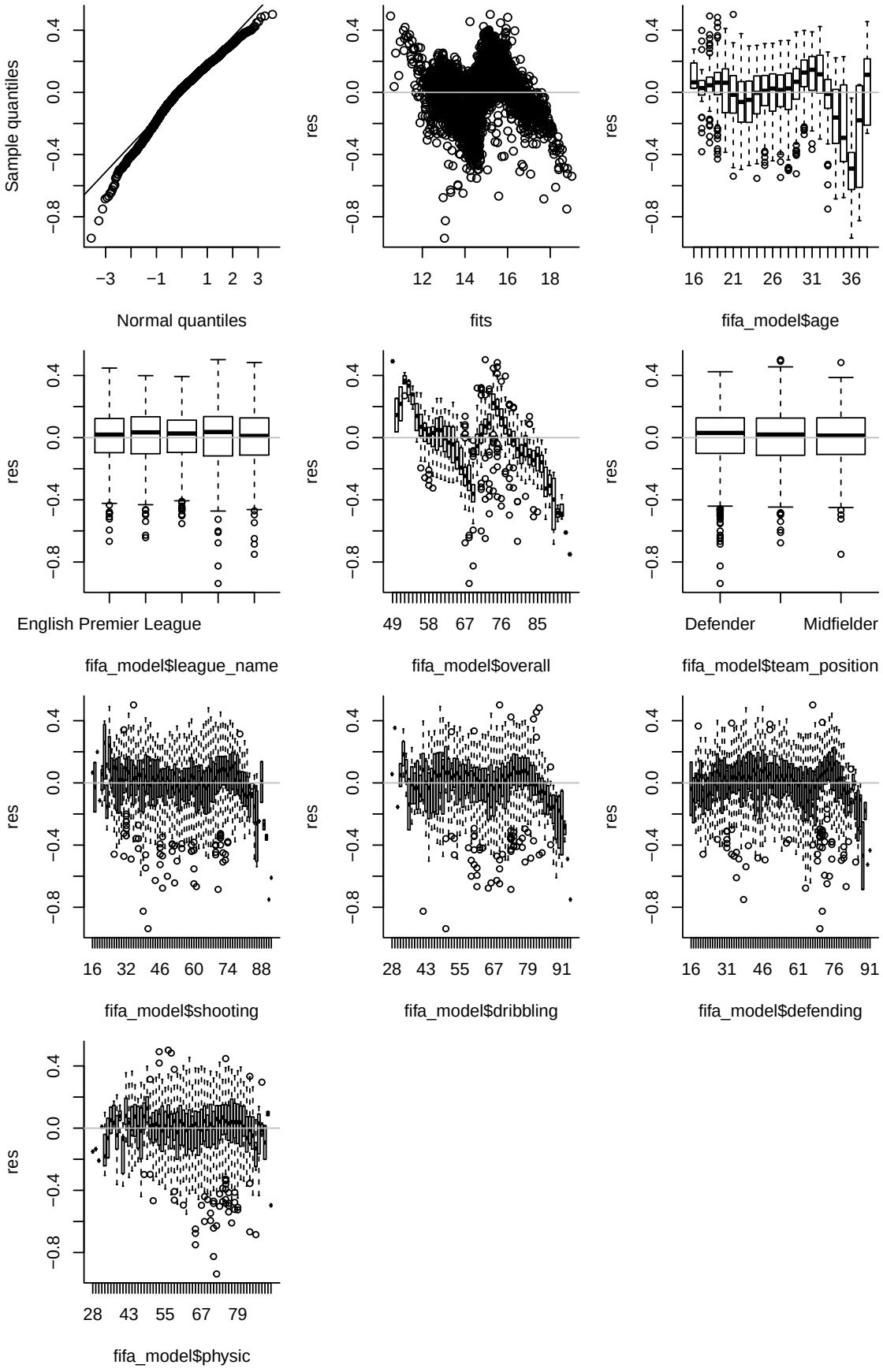

Figure 5: Final Model Diagnosis

Appendeix

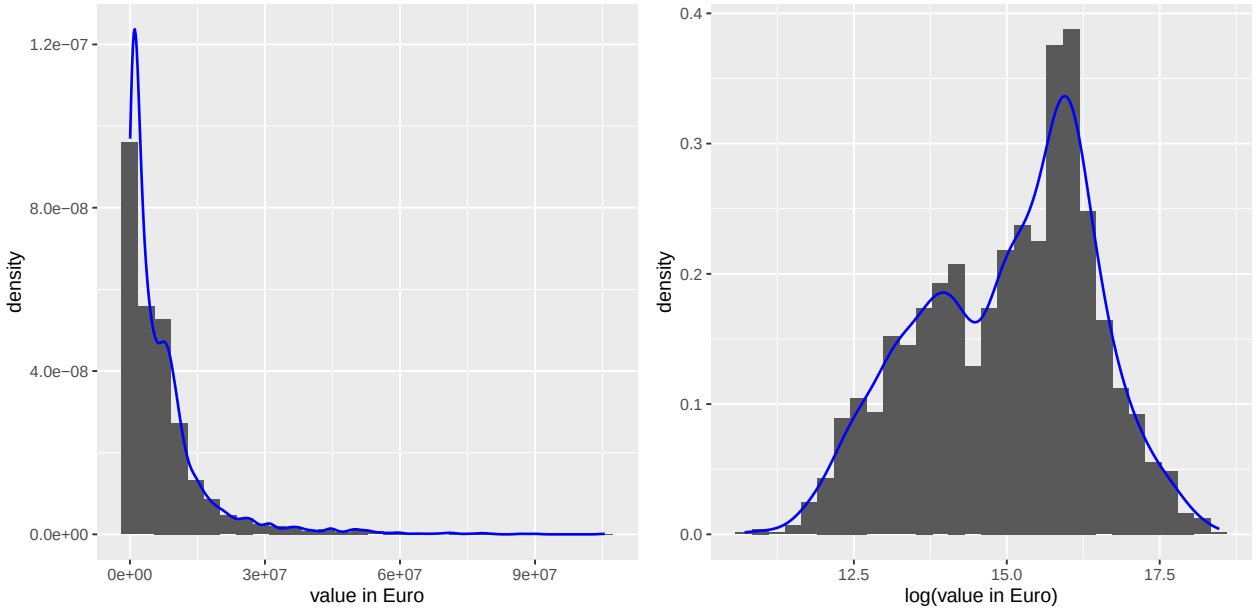

Figure 6: Appendeix 1: Distribution of response