ISSN:2790-1688 DOI: 10.56028/aetr.4.1.246.2023

Player score prediction based on multiple linear regression model

Siyuan Wang[1, *, &], Jinming Chen[1, &] and Haolong Chen[1, &]

[1]Faculty of Science and Technology , BNU-HKBU United International College, Zhuhai, China

*Corresponding author: q030024266@mail.uic.edu.cn

[&] These authors contributed equally to this work

Abstract. The recent World Cup in Qatar has just come to an end and the performance of a player is key to winning the tournament, so predicting a player's score during the season based on various metrics and performance largely determines whether or not he or she will play. Our main work is to establish how far the linear regression model is based on the discovery of the linear relationship between the data set. First, we filtered the variables with correlations greater than 0.9 by Pearson's correlation coefficient to eliminate the co-linearity problem and identified the 11 variables we used. Then a random sample was extracted, the data set was cut and the null was removed. We then screened the variables again by Forward selection to build the first regression model with an R2 of 0.67. Since some of the data had some nonlinearity, we compared the transformation of the global data (Box-Cox method) with the transformation of the local data (ln transformation of x_4) and found that adding the latter was better. After rejecting significant variables, we conducted regression again and obtained our final model with an R2 of 0.97+. Then we carried out a model diagnosis and proved that our model was indeed consistent with the linear regression model through five hypothesis tests and collinearity tests. Finally, we ran our results using the test set and found that the results were better on the test set.

Keywords: Multiple linear regression model; Box-Cox method; hypothesis testing; FIFA.

1. Introduction

Football is one of the most popular sports and the recent Qatar World Cup took the game to a new high. Our study is based on the use of multiple linear regression models in the FIFA dataset. On the one hand, our goal is to predict the score of each player in the season more accurately so as to provide reference for the actual situation. On the other hand, we hope to explore the relationship between various indicators and performance of a player and his final comprehensive score and find important influencing factors.

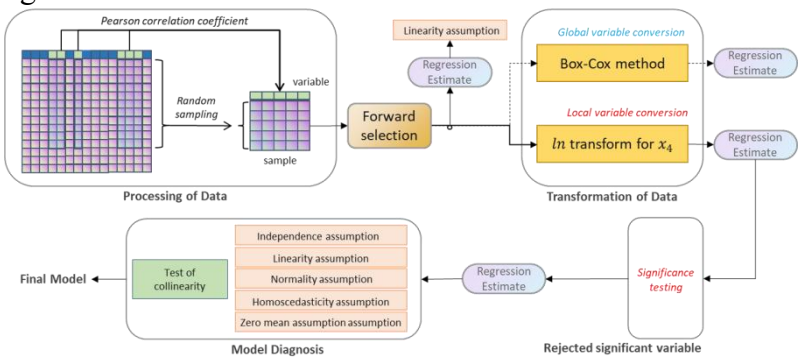


Fig. 1 The whole process of our multiple linear regression modeling

First, we eliminated collinear problems by Pearson correlation analysis and identified 11 variables. Then the first regression model is established after data processing. Due to some nonlinearity of some data, R2 is low, so we transform local data (ln transform for x4). Significance test was conducted and regression was performed again, and the final model with R2 of 0.97+ was obtained.

ISSN:2790-1688 DOI: 10.56028/aetr.4.1.246.2023

2. Data Processing

2.1 DataSet

Our dataset is from FIFA (https://data.world/raghav333/fifa-players), we use random sampling method to obtain the data of 2500 players, each data tuple (player) corresponds to 50 items information which covers the basic information of the player (such as state, age, height, etc.) and his scores in various performances during the season (such as finishing, dribbling crossing, heading accuracy, etc.). Label is the overall rating, we hope to predict the player's comprehensive score in the season through these data. After getting access to the FIFA dataset of 2500 players, we separate them into training dataset and testing datasets on a ratio of 0.75:0.25, then drop the row where the missing value is located.

2.2 Selection of Variables

To reduce the useless information caused by multicollinearity, we discarded variables with a correlation higher than 0.9 through the Pearson correlation between them, the correlation matrix and finally selected 12 items as the dependent variables of our model in combination with the common sense of the soccer tournament, whcih shown in Table 1. z_1 and z_2 are categorical variables. If the player's body type is neither lean nor normal then he is strong, and the others are numerical variables.

Fig. 2 Heatmap of the key variable

Table 1. Variables and assumptions

Symbol	Meaning
x_1	age
x_2	height(cm)
x_3	weight(kg)
x_4	player's value in euro($)
x_5	finishing: the ability to terminate the game(0-100)
x_6	dribbling: players' ability to bribble the ball(0-100)
x_7	volleys: players' volley the ability(0-100)
x_8	acceleration:players' accelerating ability(0-100)
x_9	crossing: players' croissing ability(0-100)
x_{10}	heading accuracy: players' heading accuracy(0-100)
z_1	body type:lean(boolean:1 if lean, 0 otherwise)
z_2	body type:normal(boolean:1 if normal, 0 otherwise)
y	overall rating of a football player(0-100)

3. Multiple linear regression model

3.1 Object

The regression analysis method that establishes the quantitative relationship between indicators and multiple influencing factors is called multiple regression analysis. Among multiple regressions, multiple linear regression is the most basic method[1].

Although the relationship between regressors and response is still ambiguous, we can observe that there exists some nonlinearity between x_4 and y, some adjustment might be needed in the subsequent discussion. We construct a multiple regression for the regressors in Function 1 that we set in Table 1. And then need to use the least square estimation which aims to get the parameter β during regression estimation in Function 1.

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 3 + \cdots + \beta_{10} x_{10} + \beta_{11} z_1 + \beta_{12} z_2$$

$$\widehat{\beta} = (X^T X)^{-1} X^T y \quad (1)$$

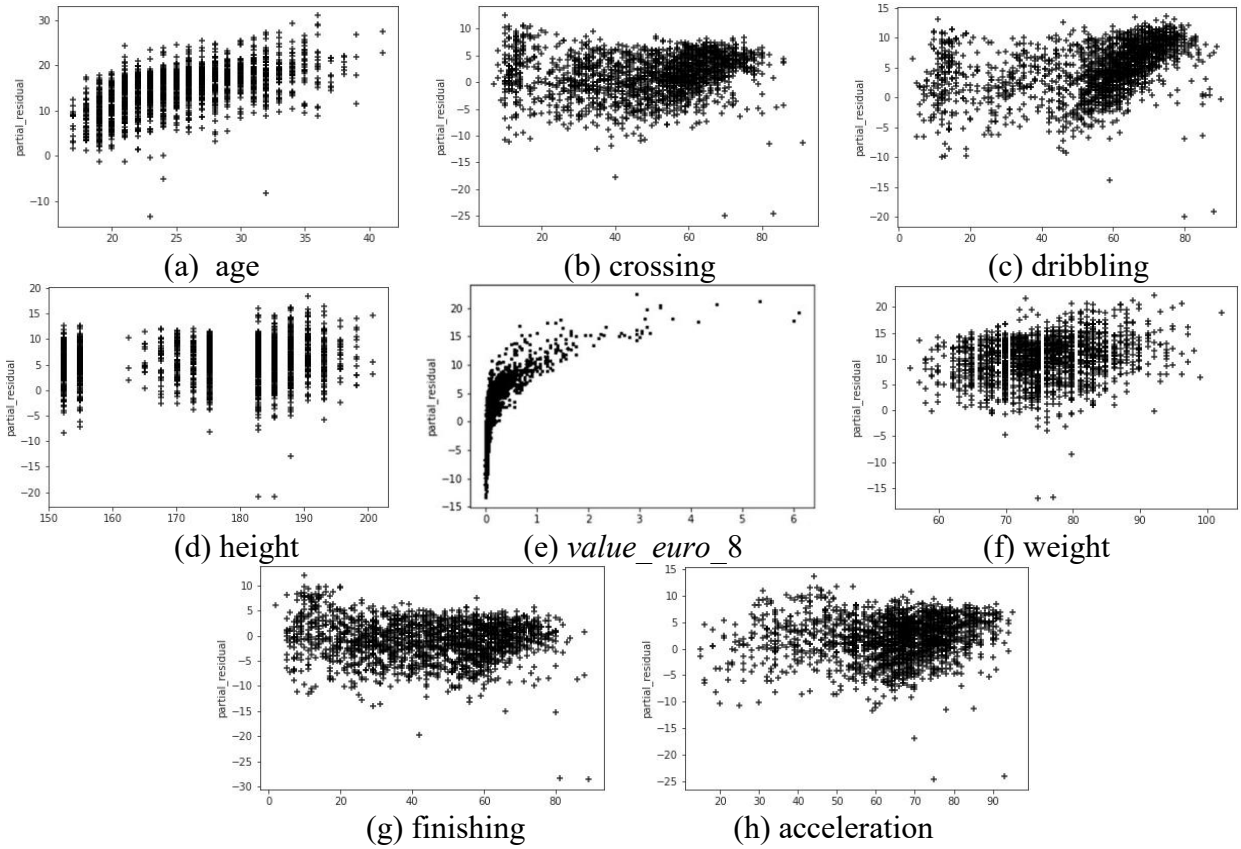
ISSN:2790-1688 DOI: 10.56028/aetr.4.1.246.2023

3.2 Forward Selection

Based on Function 1 we apply a forward approach to pick up the potentially useful subset of independent random variables. For the forward approach, each time we add an x_j to the model, then do the hypothesis testing(H_0:β_j= 0, H_1: $\beta_j \neq 0$), if F-test-value > $F_{\alpha_{entry}}$ or p-value <α_{entry}, we add x_j to our model, if the converse result appears, we don't add it. We can observe that during the process, x_7: volleys, x_{10}: *heading accuracy* and body type dummy variables(z_1, z_2)'s p-value is greater than our threshold α_{entry}, therefore do not add these variables to the potential subset. We don't include intersection terms between x_i and z_i because the body type has no extra contribution to the model, they were all rejected by the forward choice method. We get a linear equation with 8 factors, which R^2 is 0.6715. The function is shown as follows:

$$y = 26.71724 + 0.57941x_1 + 0.03236x_2 + 0.13034x_3 + 7.531802 \times 10^{-8}x_4 - 0.01927x_5 + 0.08684x_6 + 0.03048x_8 + 0.02696x_9 \quad (2)$$

We analyzed the partial residual of the regression equation of Function 2 and found that the overall data may have some non-linearity, polynomial regression, or logarithmic relationship.



(a) age (b) crossing (c) dribbling

(d) height (e) *value_euro_*8 (f) weight

(g) finishing (h) acceleration

Fig. 3 Partial Residual in Function 2

4. Data Transform

Since there are some non-linear trends in function 2. Therefore, we adopt two methods of Box-Cox conversion for the whole (Function 3) and ln conversion for the local variable x_4 (Function 4). By comparing R^2, we find that the latter has an obvious effect because the former is 0.1. Local correction is more effective than overall deformation. Finally, we choose the method of ln conversion for the local variable x_4 to carry out regression again. Here, we follow the

ISSN:2790-1688 DOI: 10.56028/aetr.4.1.246.2023
significance test of t distribution, the confidence level is 0.05, and after rejecting the significant variables, we get a four-factor final model in Function 5.

4.1 Box-Cox Method

Since e is an unobservable random error vector, it generally does not satisfy the four basic conditions of linearity, error independence, error variance chi-square, and normality of error distribution. To perform least squares estimation, it is necessary to take certain measures to "integrate" the processed data and seek some kind of transformation to reduce the complex nonlinear problem to the proposed linear form, and the Box-Cox conversion is a good method.

When such non-linearity occurs, we try the Box-Cox method of the data transformation to find the best estimate of λ and transformed value Y^*. Box-Cox method function is as follows[2]:

$$Y^* = \begin{cases} log(Y), & if\ \lambda = 0, \\ \frac{Y^{\lambda}-1}{\lambda}, & otherwise. \end{cases} \tag{3}$$

As shown in Figure 4, we get the output λ =5.96 and it is convergent when CI is 0.95. Based on Function 2, we get a new model after the Box-Cox method, whose R^2 is 0.82303 and the function as:

$$\hat{y}^{5.96} = -1.98108 * 10^8 + 543592582x_1 + 19062678x_2 + 1110322277x_3 + 1342.08791x_4 - 18387255x_5 + 66037987x_6 + 10794795x_8 + 2684332x_9 \tag{4}$$

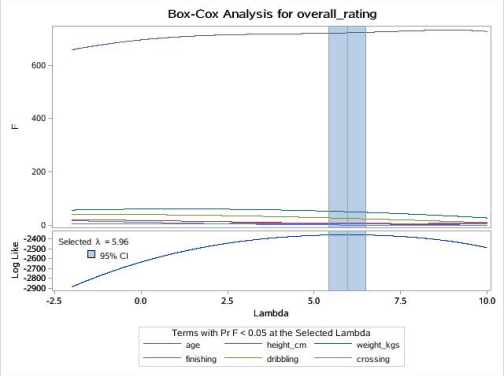


Fig. 4 Box-Cox method based on Function 3, it is convergent, when CI = 0.95

4.2 *ln* Transform for x_4

As shown in Figure 3, according to the partial residual of x_4 in the scatter map and the 8-factor model from Function 2, Through simple fitting of x_4 through Python, the shape is approximately ln(x) function. Therefore, we believe that x4 exists in logarithmic form and transform it to make the original x_4 become $ln(x_4)$. Again, by making the regression prediction, we get a new model, which R^2 is 0.9760 and the function as:

$$y = -6.36684 + 0.43885x_1 + 0.00026926x_2 + 0.00642x_3 + 4.50456ln(x_4) - 0.02848x_5 + 0.00485x_6 + 0.00169x_8 + 0.00696x_9 \tag{5}$$

We choose our confidence level as 0.05, bilateral inspection, if p is greater than 0.05, is not significant. After t distribution test, the insignificant variables are rejected, and then the regression prediction is made again, we get a new 4-factor model, whose R^2 is 0.9727 is shown as follows:

ISSN:2790-1688 DOI: 10.56028/aetr.4.1.246.2023

$$y = -5.83719 + 0.43832x_1 + 4.51819ln(x_4) - 0.02625x_5 + 0.0092x_9 \quad (6)$$

Since R^2 only decreased by 0.004 but helped us to filter out 4 items, we believe that its rejection of those variables that are insignificant and not strongly correlated helps us to avoid over-fitting and enhance the generalization of the model. Compared with the Box-Cox method, in terms of MSE and $R^2_{adjusted}$ we can tell that our model modification performs better. Therefore, we believe that the simplified equation(6) is the final model.

5. Model Diagnosis

In this section we will check if the 5 assumptions for the least square method stand to modify the model, they are independence, linearity, normality, homoscedasticity and zero means. In the end, we do a multiple collinearity test to confirm our model.

5.1 Independence assumption

The error terms are independent. $H_0: \varepsilon_t$ are independent (not auto-correlated) *vs.* $H_1: \varepsilon_t$ are1st order positively auto-correlated. The test statistic for detecting 1st order auto-correlation is:

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \quad (7)$$

where 0 ≤ DW ≤ 4. From the result we can see that the DW test value is very close to 2(1.999) and self-correlation is 0, we can thereby claim that the error terms are independent and assumption 1 holds[3].

5.2 Linearity Assumption

The true relationship between the mean of the response variables and explanatory variables is linear. The linearity assumption stands is equivalent to verify the partial residual plots are linear, which is shown in Figure 5-8. The partial residual between $e + \beta_i x_i$ and x_i are linear.

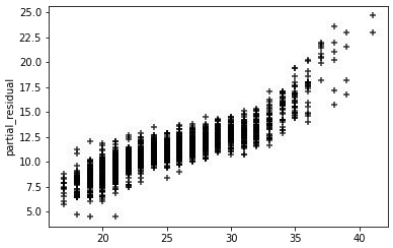


Fig. 5 Aage partial residual

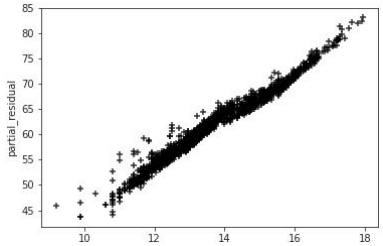


Fig. 6 ln value europartial residual

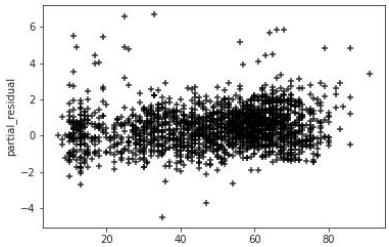


Fig. 7 Crossing partial residual

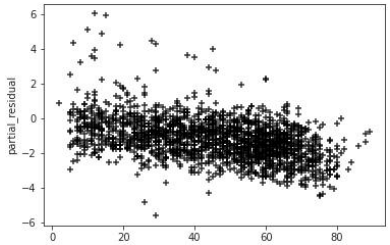


Fig. 8 Finishing partial residual

5.3 Normality assumption

The error terms are normally distributed. Here we use the Q-Q plot to verify this assumption. If the normal probability plot shows a straight line, which indicates the observed sample's quantile

ISSN:2790-1688 DOI: 10.56028/aetr.4.1.246.2023

is identical to the corresponding standard normal quantile. It is reasonable to assume the observed value comes from a normal distribution. From the Q-Q plot shown in Figure 9, most of the points lie on a straight line, therefore we can say that the normality assumption holds[4].

5.4 Homoscedasticity assumption

The error terms all have the same variance σ^2. Before we verify this assumption, we first introduce studentized residual. Studendize residual is defined as follows:

$$r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}} \tag{8}$$

Where $h_{ii} = X_i^T(X^TX)x_i$. Plot the studentized residual which shown in Figure 9, we can judge that the studentized residual indeed carter to the requirements, which is around [-2,2], so the assumption is verified that is the error terms all have the same variance σ^2.

5.5 Zero mean assumption

The error terms have a mean zero. Same to the homoscedasticity assumption, we use studentized residual (Function 5) and Figure 10. Instead of verifying the error terms, we verify if the studentized residual lies around zero and within[-2,2]. We can judge that the studentized residual indeed carter to the requirements. Hence we conclude that the homoscedasticity assumption and 0 mean assumption holds.

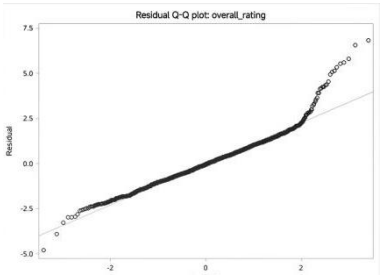


Fig. 9 Q-Q plot of final mode

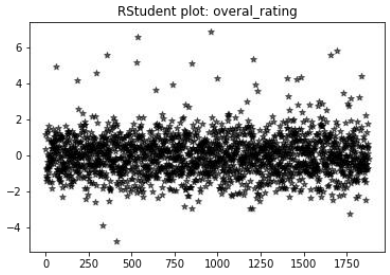


Fig.10 Studentized residual plot of final mode

5.6 Multiple collinearity

Additionally, we examine the noncollinearity. Since the intercept of our model doesn't have significant meaning, we choose to run a collinearity test with no corrected intercept. The collinearity test conditional index larger than 10 usually indicates the existence of collinearity, our model's conditional index is far lower than 10. The results show that we do not have collinearity problems in Table 2.

Table 2. The results on train and test

No.	Eigenvalue	Conditional index	age	Deviation ratio ln(value)	finishing	crossing
1	2.02499	1.00000	0.01701	0.09894	0.09548	0.09747
2	0.908871	1.43112	0.87210	0.01297	0.03503	0.01440
3	0.63750	1.78225	0.09670	0.88665	0.07463	0.10770
4	0.34880	2.40948	0.01418	0.00143	0.79486	0.78043

ISSN:2790-1688 DOI: 10.56028/aetr.4.1.246.2023

6. Conclusion

Table 3. The results on train and test

Model	Train		Test	
	R2	RMSE	R2	RMSE
8-factor model	0.6729	3.96979	0.584883	4.705862
8-Box-Cox model	0.86988	2.87461	0.42396	13.32367
8-ln-value model	0.9760	1.07452	0.976369	1.122765
4-factor-$ln(x_4)$ model	0.9726	1.14428	0.97646	1.120401

First of all, we found that the linear regression model is very good using the FIFA data set. Secondly, we find that the euro value after ln transform has a crucial impact on the forecast. Therefore, the euro value not only has an obvious fitting trend but also has a direct impact on the overall rating. Finally, from the results in Table 3, the improvement trend of our Test results is consistent with that of Train, and the scores are slightly higher than those of Train. It can also be shown that the simplification of variables in the optimization process of the linear regression model helps to avoid overfitting, thus improving the generalization of the model and saving the calculation time.

In this paper, our main contributions are: We established a multiple linear regression model applicable to the FIFA data set, and achieved excellent fitting results in training and testing. Then we found the core variable affecting this data set, namely the Euro value after ln transform, which is highly correlated with the prediction accuracy of the overall rating. The experiment proves that simplifying multiple linear regression equation is helpful to avoid overfitting and improve model generalization.

References

[1] Shan YZ, Xu HC, Shan WZ, et al. Multiple linear regression prediction method for submunition drop point dispersion[J]. Journal of Nanjing University of Science and Technology (Natural Science Edition),2013,37(5):720-724.

[2] Qiu B, Wang LJ, Zhu JJ, et al. Box-Cox transform and its application in ground settlement analysis caused by pipe jacking construction[J]. Engineering Survey,2009,37(7):55-58.

[3] Xu Gongwei. Exploration of D.W test in regression analysis[J]. Mathematical Statistics and Management,2001,20(3):46-48.

[4] Bao, Zhonghua, Gong, Shenguang, Ma, Ke, et al. Testing the normality of the measured marine environmental electric field data[J]. Journal of Wuhan University of Technology (Transportation Science and Engineering Edition),2010,34(4):776-779.