The Prediction Model Research on Objectives and Results of Football Game Based on Regression Method

Yonggan Wang

Institute of physical education, Langfang Teacher's College, Langfang, China

w_angyonggan@126.com

Keywords: Football game; Objective results; Regression analysis; Method factor; Regression model; Fitting degree

Abstract. With China's rising international status,the development of sports undertakings has been paid more and more attention,especially football sports are highly expected.The objectives and results of the football game prediction can effectively improve the level of Chinese football training and competition skills , narrowing the gap between the technical and tactical Chinese football and world powers. 10 important technical and tactical datum are studied and analyzed by studying impact factor of the 17 World Cup soccer game and using regression analysis method.Relevant factor regression model is built,the objectives and results are scientifically predicted,the degree of fitting between predication and actual observed value is studied,the feasible theoretical basis of the technical and tactical training and prediction of competition results before game is provided.

Introduction

Chinese football competitive sports which is in the transformation pried is still at the initial stage now, the emphasis for the sports is rising.For decades,China's sports undertakings has made a spurt of progress, ranking the best in the world sports competitions[1].The level of sports competition is mainly refected in the result of match,so people attach more and more significance to sports.The effective way of predicting the objectives and results is increasingly paid attention by people.Many experts or physical education teaching workers has done a lot of theoretical analysis in the field.

Regression analysis method

A. The outline of regression analysis method

Regression analysis is a statistical method and skills solving the relationship between variable x and variable y.The outcome of relationship between variables is certain,but the outcome of Y is not so in study,so probability distribution can be used to express.Here,we define conditional mathematical expectation of X and Y to be stochastic variable Y being as average regression function of X.

$$f(x) = E(y/x) \tag{1}$$

Formula 1 displays statistical law of variable x and variable y according to the average sense .We define x as to be independent variable and define y as to be dependent variable in practice use.If we use x to predict y,so the prediction outcome is required,which means sample observations(x_1, y_1),(x_2, y_2),$\cdots$,(x_n, y_n),when the outcome of x is appointed ,bring the outcome into formula,the outcome of y can be obtained.The outcome is the prediction outcome of y.

Regression analysis model is mainly discovering the quantitative relationship between correlated variable between things.As shown in Figure 1.

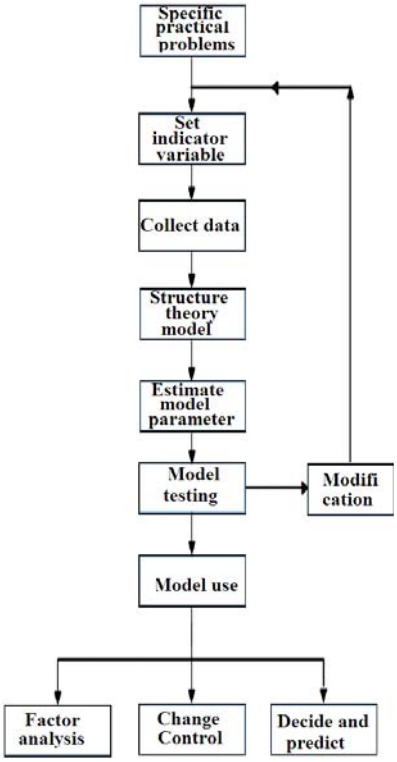

Figure 1. The setup procedure of regression model

B. Multiple linear regression analysis method

*1)*The general form of the multivariate linear regression model

The general form of the multivariate linear regression model is as follows[2]:

$$\eta(u) = \beta_1\varphi_1(u) + \beta_2\varphi_2(u) + \cdots + \beta_m\varphi_m(u) \quad (2)$$

$$y = \beta_1\varphi_1(u) + \beta_2\varphi_2(u) + \cdots + \beta_m\varphi_m(u) + \varepsilon \quad (3)$$

ε is random error,as well as $\varepsilon \sim N\left(0,\sigma^2\right), \varphi_i\left(u\right), i = 1, 2, \cdots, m$,all of them are explaining variable in practical problem,they are known function.

Assume that n experiences have been done,the prediction data is got.

$$\begin{bmatrix} u_1 & y_1 \\ \vdots & \vdots \\ u_{n-1} & y_{n-1} \\ u_n & y_n \end{bmatrix}$$

Substitute them into (9.3),so

$$y_i = \beta_1\varphi_1(u_i) + \beta_2\varphi_2(u_i) + \cdots + \beta_m\varphi_m(u_i) + \varepsilon_i, i = 1, 2, \cdots, n \quad (4)$$

ε_i are the random errors happened in the i times experiment and they are independent with each other $\varepsilon_i \sim N\left(0,\sigma^2\right)$.

The model concerning regression coefficient $\beta_1, \beta_2, \cdots, \beta_m$ is linear.u is ofen the vector quantity.For convenience,bring it into matrix notation.

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} \varphi_1(u_1) & \varphi_2(u_1) & \cdots & \varphi_m(u_1) \\ \varphi_1(u_2) & \varphi_2(u_2) & \cdots & \varphi_m(u_2) \\ \vdots & \vdots & & \vdots \\ \varphi_1(u_n) & \varphi_2(u_n) & \cdots & \varphi_m(u_n) \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

X is called as model design matrix,it's constant matrix,y and ε are random vector,also in this formula[3] :

$Y \sim N_n\left(X \cdot \beta, \sigma^2 I\right), \varepsilon \sim N_n\left(0, \sigma^2 I\right), I$ is unit matrix, ε is unobservable vector quantity of random error, β is vector made up of regression coefficient,it is unknown and undetermined constant vector.

The next content tells the problem of how to estimate regression coefficient β and test conspicuousness and the fitting degree about model.

*2)*The least estimation of regression coefficient

Choose one estimated data of β ,regard it as $\hat{\beta}$ .Make square sum of random error ε get minimum,namely,

$$\min_{\beta} \varepsilon^T \cdot \varepsilon = \min_{\beta} (Y - X \cdot \beta)^T (Y - X \cdot \beta)$$
$$= \left(Y - X \cdot \hat{\beta}\right)^T \left(Y - X \cdot \hat{\beta}\right) \triangleq Q\left(\hat{\beta}\right) \qquad (5)$$

Write it as component form.

Notice that $Q(\beta_1, \beta_2, \cdots, \beta_m)$ is nonnegative secondary type.It is differentiable.According to necessary condition multivariate that multivariate function should take extreme, $\frac{\partial Q}{\partial \beta_j} = 0,\ j = 1, 2, \cdots, m$ is got,that is

is the formula

$$\sum_{i=1}^{n} [y_i - \hat{\beta}_1 \varphi_1(u_i) - \hat{\beta}_2 \varphi_2(u_i) - \cdots - \hat{\beta}_m \varphi_m(u_i)] \varphi_j(u_i) = 0,\ j = 1, \cdots, m \qquad (6)$$

Arrange it as

$$\begin{cases} [\sum_{i=1}^{n} \varphi_1^2(u_i)]^2 \hat{\beta}_1 + [\sum_{i=1}^{n} \varphi_1(u_i)\varphi_2(u_i)]\hat{\beta}_2 + \cdots + [\sum_{i=1}^{n} \varphi_1(u_i)\varphi_m(u_i)]\hat{\beta}_m = \sum_{i=1}^{n} \varphi_1(u_i) y_i, \\ \quad \cdots\cdots \quad \cdots\cdots \\ [\sum_{i=1}^{n} \varphi_1(u_i)\varphi_m(u_i)]\hat{\beta}_1 + [\sum_{i=1}^{n} \varphi_2(u_i)\varphi_m(u_i)]\hat{\beta}_2 + \cdots + [\sum_{i=1}^{n} \varphi_m^2(u_i)]\hat{\beta}_m = \sum_{i=1}^{n} \varphi_m(u_i) y_i, \end{cases}$$

Oρ

$$X^T \cdot X \cdot \hat{\beta} = X^T \cdot Y$$

It's called normal equations.Record $A = X^T \cdot X$ as to be coefficient matrix and $B = X^T \cdot Y$ to be constant matrix. If A^{-1} exist,it's called as correlation matrix.It can be proved that normal equations systems are always having result even if x and y are given randomly.Although the result is not unique when x is not full rank, $\hat{\beta}$ makes residual sum of squares least for any group result.

Especially ,when x is full rank, $r\left(X\right) = r(X^T \cdot X) = m$, the solution of normal equations system is $\hat{\beta} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$,which is estimation of regression coefficient.

Because $\hat{\beta} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$ and $\hat{\beta}$ is also a random vector and its prediction is

$$E(\hat{\beta}) = E((X^T \cdot X)^{-1} \cdot X^T \cdot Y) = (X^T \cdot X)^{-1} X^T \cdot E(Y)$$
$$= (X^T \cdot X)^{-1} X^T X \beta = \beta$$

In a similar way,the variance is $D(\hat{\beta}) = \sigma^2 (X^T \cdot X)^{-1}$,that is to say, $\hat{\beta}$ is a unbiased estimation of β.

The estimation about model can be got when bring $\hat{\beta}$ into model $\eta(u)$: $\hat{Y} = X^T \hat{\beta}$,which is the unbiased estimation of model $\eta(u)$.

Namely

$E(\hat{Y}) = E(X^T \hat{\beta}) = X^T E(\hat{\beta}) = X^T \beta = \eta$ and among them

$X = (\varphi_1(u), \varphi_2(u), \cdots, \varphi_m(u))^T$.

The objective and result of football game

A. Set up estimation of objective and result of football game

Mainly test whether the model be certainly closely related to the explanatory variables , and whether has the form of formula (2).Assume that η is independent of u ,which means $\eta = \beta_0$ is constant.The same situation as that,record the mean value of experiment as $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$,total deviation sum of squares is SS_T ,that is[4]

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$
$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$
$$\triangleq SS_E + SS_R$$

The sum of residual method is [5]

$$SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \left(Y - X \cdot \hat{\beta}\right)^T \left(Y - X \cdot \hat{\beta}\right) = Y^T Y - Y^T X \cdot \hat{\beta}$$

Regression sum of squares is $SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

Now regression sum of squares should be primarily taken into account .Define multiple correlation coefficient as $R = \frac{SS_R}{SS_T}$,using it to test the evaluation of the effectiveness of the model.If the bigger the data of R is,the closer the relationship between regressor variable and response. If not ,the result is totally different.

A F statistical magnitude need to be setted in order to investigate R.At first,degree of freedom is worked out.Degree of freedom of total deviation sum of squares $f_T = n-1$,and degree of freedom of regression sum of squares $f_R = m-1$,the degree of freedom of sum of squared residuals $f_E = f_T - f_R = n-m$,so the relevant quadratic mean is[6]

$$MS_R = \frac{1}{m-1}SS_R, MS_E = \frac{1}{n-m}SS_E$$

It can be proved that if $\eta = \beta_0$,due to $y_i \sim N\left(0, \sigma^2\right)$,so

$$E(MS_R) = E(\frac{1}{m-1}SS_R) = \sigma^2, E(MS_E) = E(\frac{1}{n-m}SS_E) = \sigma^2$$

It shows that MS_E is the unbiased estimation of σ^2 ,namely $\frac{SS_E}{\sigma^2} \sim \chi^2(n-m), \frac{SS_R}{\sigma^2} \sim \chi^2(m-1)$,as well as SS_R and SS_E are mutually independent[7].

$$F = \frac{MS_R}{MS_E} = \frac{SS_R / m-1}{SS_E / n-m} \sim F\left(f_R, f_E\right) = F\left(m-1, n-m\right) \tag{8}$$

Take a remarkable level,refer to the Figure $F_\alpha\left(m-1, n-m\right)$ and work out $F\left(m-1, n-m\right)$ and then compare it with $F_\alpha\left(m-1, n-m\right)$.

When $F\left(m-1, n-m\right) > F_\alpha\left(m-1, n-m\right)$,the model is thought as remarkable,so $\eta = \beta_0$ will be established.There are obvious function relationships between η and u .

When $F\left(m-1, n-m\right) < F_\alpha\left(m-1, n-m\right)$,he model is thought as unremarkable,so $\eta = \beta_0$ will be established.There are no obvious function relationships between η and u .

When 4 common factors are substituted into formula 7,the result is as follows: the Brazil team is top1; the Germany team is top2; the Turkey team is top 3 ; the Korea team is top4; the Spain team is top5;the England team is top6; the Senegal team is top7 ;the U.S team is top 8; the Japan team is top 9; the Denmark team is top 10; the Mexico team is top11;the Ireland team is top12;the Sweden team

is top13;the Belgium team is top14; the Italy team is top15; the Paraguay team is top 16; the South Africa team is top17;the Argentina team is top18; the Costa Rica team is top19;the Cameroon team is top20.The goodness of fit is nearly towards 1,which shows that descriptions of common factor and regression model is fitting very well.Figure 2 shows that determination coefficient $R = 0.987$,which indicates that the fitting degree between regression model and actual value is extremely high, the model feasibility and adaptability are extremely strong, further proving that the degree of fitting of the team's results and 4 common factor is extremely high.

TABLE I. Model summary

Model	R^2	R^2	**Adjust** R^2	**Standard error of estimate**
1	0.987	0.961	0.925	0.176

Conclusion

The important technical and tactical datum are studied and analyzed, multiple regression method is used to set up a model of predicting the result of match.According to the factors affecting the variable, fitting degree can be got $R = 0.987$,which tests that the fitting degree between regression model and actual value is extremely high .Multiple regression prediction model can solve complicated hierarchy problem, and make the problem become clear, can scientifically forecast sports results.The prediction of the 4 common factor for the football game provides a theoretical basis for technical and tactical training.

Reference

[1] Wang kai,Lv Wei,He Jiangchuan.Regression analysis of football players' tactical quality level and winning factor . Journal of Capital Institute of Physical Education, 2012(3):77-81.

[2] Zhu Wenfu.Sports results forecasting model study.Journal of Chongqing Technology and Business University(Social Science Edition,2011(3):27-31.

[3] Ma Yongkai,Tang Xiaowo.Multi-objective optimization combination forecast model research.Journal of Southwest University,2012(7):35-38.

[4] WangLiming,ChenYing,YangNan.Application of the regression analysis.Fudan University Press, 2008:253-261..

[5] Jin Chuanjiang. Comprehensive evaluation on the seventeenth session of the World Cup teams of technical and tactical ability by using Q cluster . Journal of Beijing Sport University, 2011 (4):56-59.

[6] Gong Mingbo.The scale space hierarchical clustering in the football team of technical and Tactical Ability .Sports science,2009(12):101-105..

[7] JiangXixi.Delta Cup football offensive statistical analysis . Journal of Beijing Sports University, 2011 (10):98-102.