ISSN : 0974 - 7435 Volume 8 Issue 9

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 8(9), 2013 [1199-1204]

Sports competition results predict a linear regression model applied research

Jun Hao[1], Yuhao Zhang[2]

[1]Institute of Physical Education, East China Jiaotong University, Nanchang 330013, (CHINA)

[1]Email:37679055@qq.com

[2]Sport School, Shandong Sport University, Jinan 250101, (CHINA)

[2]Email:39310257@qq.com

ABSTRACT

This paper uses two-factor variance analysis, studies the top three achievements of men's 110m hurdles in the 15th to 30th Summer Olympic Games, establishes the optimal quaternary linear regression prediction model by optimizing and comparing, respectively builds the achievements prediction equation of the third place, runner-up and champion performance, predicts the top three scores of men's 110 meters hurdles in the 31th Summer Olympic Games. Studies have shown that achievement of the third place in the 31th Olympic 100-meter hurdles is 13.11, achievement of the runner up is 13.09, and achievement of the champions is 12.91. The model and results well descript the overall level variation law of the recent 16 Summer Olympic Games men's 110-meter hurdles, predict the achievements development, and provide a new way for the best combination methods to play a role in sports competitions and research.

KEYWORDS

Regression model;
The 110-meter hurdles;
Prediction equation.

INTRODUCTION

Although our technical level of the men's 110-meter hurdles belongs to the first level in Asia, it is still relatively below in the world. By observing the list of top three athletes in the previous summer Olympics, there are few Asian athletes, and the name of the Chinese athletes only appears one time. We must face up to our own defects, compensate through technical improvements and reasonable training, and may enter into the advanced level in the world. For the determination of athletes training strategy, first training objectives need to be determined. At present to carry through correct and reasonable prediction on the achievements of the top three athletes with the international advanced level in the 2016 Summer Games is a must.

This paper first analyzed the achievements data of the top three athletes of the 15th to the 30th Summer Olympic Games men's 110m hurdles through the two-way ANOVA method; then it established a multi-combination type linear regression model and conducted statistical tests on the model and the results, which optimized the model; using the optimized linear regression model to describe the sports level of men's 110-meter hurdles in international; Through the prediction and understanding on the development direction of this project, it finally arrived at the prediction results on the top three scores of the 31th Olympic Games men's 110-meter hurdles, which provides target value for the athletes selection and training strategies determination of this project, as well as provides methods for the data prediction of similar sports project that has correlation over time.

RESEARCH OBJECT AND METHOD

Research object

This paper takes the champion, runner-up and third place achievement data of the recent 16 Summer Olympic men's 110 meters hurdles as the research object; the statistical data is from the official website of the Chinese Olympic Committee http://www.olympic.cn/games/list2.html, and the data is reliable.

Research method

1) Document literature

Through ISI/EI/CNKI and other data base, download related research papers, mathematical statistics tutorials and matrix calculations related tutorial, which makes adequate preparation for the correctness of the theory and calculation.

2) Mathematical statistics

Use SPSS software, EXCEL functions and mathematical integrated computing system-FORLAB to conduct processing and computing on the statistical data.

THE PREDICTION MODEL ESTABLISHMENT OF MULTIPLE LINEAR REGRESSIONS

Multiple linear regressions

Here we establish multiple linear equations: If the random variable y is related to $p(p \geq 2)$ numbers of ordinary variable $x_1, x_2, x_3, \cdots, x_p$, and satisfies equation (1), equation (1) is the mathematical description way of the multiple linear regressions.

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p + \varepsilon \\ E(\varepsilon) = 0, Var(\varepsilon) = \delta^2 < +\infty \end{cases} \quad (1)$$

Wherein $\beta_0, \beta_1, \beta_2, \beta_3, \cdots, \beta_p, \delta^2$ are the unknown parameters that have nothing to do with $x_1, x_2, x_3, \cdots, x_p$, and ε is the unobserved random variables. Call equation (1) p meta-theory linear regression model; $\beta_0, \beta_1, \beta_2, \beta_3, \cdots, \beta_p$ are the regression coefficients; are called the regression factors or design factor, factors for short. Parameters reflect the influence of factors to the observation, sois also known as the effects of factor.

Suppose there are groups of sample observations which are not all the same, obtaining formula (2) by formula (1):

$$\begin{cases} y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + \varepsilon_i \\ E(\varepsilon_i) = 0, Var(\varepsilon_i) = \delta^2 < +\infty \end{cases} \quad (2)$$

Wherein $i = 1,2,3,\cdots,n$, moreover $\varepsilon_1, \varepsilon_2, \varepsilon_3, \cdots, \varepsilon_n$ is independent to each other.

The matrix expression of the formula (2) is shown in formula (3):

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \begin{cases} Y = X\beta + \varepsilon \\ E(\varepsilon) = 0, Var(\varepsilon) = \delta^2 I_n \end{cases} \quad (3)$$

Wherein I_n is the n-order unit matrix, 1_n is the-dimensional column vector that the elements are all 1. $\boldsymbol{Y}$ is called the observation vector of random variables, $\boldsymbol{\beta}$ is the unknown parameter vector, $\boldsymbol{X}$ is design matrix, $\boldsymbol{\varepsilon}$ is the $\boldsymbol{n}$-dimensional random error vector. In the regression analysis, generally assume $\boldsymbol{rank(X) = p+1}$, namely require that $\boldsymbol{X}$ is the column satisfaction. So it has $\boldsymbol{E(Y) = X\beta, Var(\varepsilon) = \delta^2 I_n}$. Generally take $\boldsymbol{\varepsilon \sim N(0, \delta^2)}$, so the expression form of formula (3) is shown as formula (4) below:

$$\begin{cases} Y = X\beta + \varepsilon \\ \varepsilon \sim N(0, \delta^2 I_n) \end{cases} \quad (4)$$

Regression coefficient calculation method

Using the least squares method to solve the estimated vector $\hat{\beta}$ of the regression coefficients column vector β in model (4), the regression coefficients error sum of squares is shown in equation (5):

$$Q(\beta_0, \beta_1, \beta_2, \cdots, \beta_p) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_0 x_{i2} - \cdots - \beta_p x_{ip})^2 \quad (5)$$

Based on the idea of the least squares method, choose the group with the smallest error sum of square as the parameters estimation of the regression coefficients, that is to say, if $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p$ exists and it satisfies the equation (6):

$$Q\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p\right) = \min \left\{Q(\beta_0, \beta_1, \beta_2, \cdots, \beta_p)\right\} \quad (6)$$

$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p$ Is the least squares estimation of parameters $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$.

Prediction method on sports results

The regression equation obtained by achievements sample $(x_{i1}, x_{i2}, x_{i3}, \cdots, x_{ip}; y_i)(i = 1,2,\cdots,n)$ is shown in formula (7) below:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \cdots + \hat{\beta}_p x_p \quad (7)$$

After test the regression effect and regression coefficient are both significant. When given a set of values $(x_{01}, x_{02}, x_{03}, \cdots, x_{0p})$, it is reasonable to take $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \hat{\beta}_3 x_{03} + \cdots + \hat{\beta}_p x_{0p}$ as the point estimation of $E(y_0)$, because according to formula (4), $\hat{y}$ is the unbiased estimation of $E(y_0)$.

RESULT ANALYSIS

Previous achievement data trends

According to the data in Table 1, we make use of EXCEL formula to draw scattered trend chart, as shown in Figure 1.

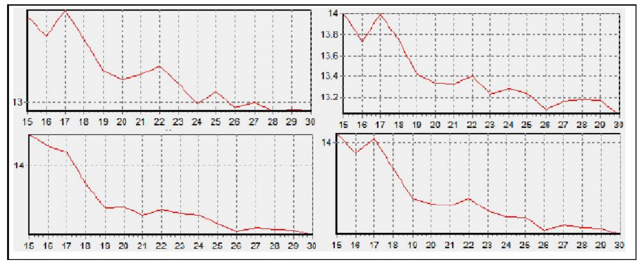

Figure 1 : The line chart of Champion Runner-up third place and average achievements of this session

The upper left of Figure 1 is the championship achievements trend chart; the upper right is the runner-up achievements trend chart; the lower left is the third place achievements trend chart; left lower is the average scores trend chart of all previous champion runner-up and third place. As can be seen from Figure 1, the achievement trend of four categories change with the increase of the session number; the change pace of the scores is getting slower. Obviously simple linear regression cannot be well fitting for the above four categories of curves. According to the change trend, we can carry through logarithmic transformation and power transformation to the Olympic session number. Then conduct the free combination, and choose the best fitting curve.

TABLE 1 : List of data after processing

X_1 Session number	X_2 Square	X_3 Natural logarithm	Y_1 Championship scores	Y_2 Runner-up scores	Y_3 Third place scores	Y_4 average scores
15	225	2.7080502011	13.91	14.00	14.40	14.1033
16	256	2.7725887222	13.70	13.73	14.25	13.8933
17	289	2.8332133441	13.98	13.99	14.17	14.0467
18	324	2.8903717579	13.67	13.74	13.78	13.73
19	361	2.9444389792	13.33	13.42	13.46	13.4033
20	400	2.9957322736	13.24	13.34	13.48	13.3533
21	441	3.0445224377	13.30	13.33	13.38	13.3367
22	484	3.0910424534	13.39	13.40	13.44	13.41
23	529	3.1354942159	13.20	13.23	13.40	13.2767
24	576	3.1780538303	12.98	13.28	13.38	13.2133
25	625	3.2188758249	13.12	13.24	13.26	13.2067
26	676	3.258096538	12.95	13.09	13.17	13.07
27	729	3.295836866	13.00	13.16	13.22	13.1267
28	784	3.3322045102	12.91	13.18	13.20	13.0967
29	841	3.36729583	12.93	13.17	13.18	13.0933
30	900	3.4011973817	12.92	13.04	13.12	13.0267

FULL PAPER

Data processing

Through the above study, conduct data processing on the Olympic session number$(15,16,17,\cdots,30)$. Respectively calculate its squares and natural logarithm, and use vector $X_1=(x_{1,1} \quad x_{1,2} \quad \ldots \quad x_{1,16})$ to represent session number; use vector $X_2=(x_{2,1} \quad x_{2,2} \quad \cdots \quad x_{2,16})$ to represent the square of the session number; use vector $X_3=(x_{3,1} \quad x_{3,2} \quad \cdots \quad x_{3,16})$ to present the natural logarithm values of the session number; vector $Y_1=(y_{1,1} \quad y_{1,2} \quad \cdots \quad y_{1,16})$ represents championship results; vector $Y_2=(y_{2,1} \quad y_{2,2} \quad \cdots \quad y_{2,16})$ represents the runner-up results; vector $Y_3=(y_{3,1} \quad y_{3,2} \quad \cdots \quad y_{3,16})$ represents the third place results; vector $Y_4=(y_{4,1} \quad y_{4,2} \quad \cdots \quad y_{4,16})$ represents the average results; the calculation results are shown in Table 1 below:

Respectively using X_1, X_2, X_3 or its combinations $(X_1, X_2),(X_1, X_3),(X_2, X_3),(X_1, X_2, X_3)$ to do unitary linear regression or multiple linear regressions on Y_1, Y_2, Y_3, Y_4. This kind of combinations have a total of 28 groups, expressed as (1,1), (1,2), (1,3), (1,4),(2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4), (1, 2,1), (1,2,2), (1,2,3), (1,2,4), (1,3,1), (1,3,2),(1,3,3), (1,3,4), (2,3,1), (2,3,2), (2,3,3), (2,3,4), (1,2,3, 1), (1,2,3,2), (1,2,3,3), (1,2,3,4). It represents the combination of X_i and Y_i, the corresponding relationship is one-to-one or many-to-one. The parameters of regression equation are shown in TABLE 2:

TABLE 2 shows the method of selecting combination. According to the test method in 3.3, select the minimum residual sum of squares, the minimum variance and the maximum multiple correlation coefficient. The champion combination is (1, 2, 3, 1) with the minimum residual sum of squares and (1, 2, 1) with the minimum variance and maximum multiple correlation coefficient. The runner-up combination is (1, 2, 3, 2) with the minimum residual sum of squares and maximum multiple correlation coefficient, and (1, 3, 2) with the maximum multiple correlation coefficient and minimum variance. The third place combination is (1, 2, 3, 3) with the minimum residual sum of squares and maximum multiple correlation coefficient, and (1, 3, 3) with the minimum variance. The average combination is (1, 2, 3, 4) with the minimum residual sum of squares, and (1, 3, 4) with the maximum multiple correlation coefficient and minimum variance. But the goodness of fitting reflects the maximum multiple correlation coefficient, the final combination of championship is (1,2,1), the final combination of runner-up is the mean value of (1,2,3,2) and (1,3,2), the final combination of the third place is (1,2,3,3), the final average combination is (1,3,4).

Establish the regression equation of achievements predictions

According to the least squares method, ultimately select combinations to carry through optimum multiple linear regressions fitting in accordance with TABLE 2.

In accordance with the combination (1, 2, 1) the regression equation of championship achievements is shown as equation (8) below:

$$\hat{Y}_1 = 16.8776 - 0.2582\,\hat{X}_1 + 0.0042\,\hat{X}_2 \quad (8)$$

In accordance with the combination (1, 2, 3, 2) and (1, 3, 2) the regression equation of runner-up average results is shown as equation (9) below:

$$\begin{cases} \hat{Y}_2' = 25.7107 + 0.2512\,\hat{X}_1 - 0.0012\,\hat{X}_2 - 5.6126\,\hat{X}_3 \\ \hat{Y}_2'' = 24.0055 + 0.1486\,\hat{X}_1 - 4.5133\,\hat{X}_3 \\ \hat{Y}_2 = \dfrac{\hat{Y}_2' + \hat{Y}_2''}{2} \\ \hat{Y}_2 = 24.8581 + 0.1999\,\hat{X}_1 - 0.0006\,\hat{X}_2 - 5.06295\,\hat{X}_3 \end{cases} \quad (9)$$

$$\hat{Y}_3 = 56.9289 + 1.8098\,\hat{X}_1 - 0.0173\,\hat{X}_2 - 24.2559\,\hat{X}_3 \quad (10)$$

In accordance with the combination (1, 3, 4) the regression equation of average results is shown as equation (11) below:

$$\hat{Y}_4 = 26.1633 + 0.1807\,\hat{X}_1 - 5.4436\,\hat{X}_3 \quad (11)$$

Predict the 31th Olympic 110 meters hurdles results

Substitute the data in TABLE 2 into formula (8), (9), (10) and (11), $E(y_{1,31})$, $E(y_{2,31})$, $E(y_{3,31})$ and $E(y_{4,31})$ can be obtained. The calculation formula and calculation results are shown in formula (12) below:

$$\begin{cases} E(y_{1,31}) = 16.8776 - 0.2582\times 31 + 0.0042\times 31^2 = 12.91 \\ E(y_{2,31}) = 24.8581 + 0.1999\times 31 - 0.0006\times 31^2 - 5.06295\times \ln(31) = 13.09 \\ E(y_{3,31}) = 56.9289 + 1.8098\times 31 - 0.0173\times 31^2 - 24.2559\times \ln(31) = 13.11 \\ E(y_{4,31}) = 26.1633 + 0.1807\times 31 - 5.4436\times 31^2 = 13.07 \end{cases} \quad (12)$$

TABLE 2 : List of regression equation parameters

Combination mode	The residual sum of squares	The sum of deviation squares	Multiple correlation coefficient	Variance
(1,1)	0.301	1.9301	0.9187	0.301
(2,1)	0.3868	1.9301	0.8942	0.3868
(3,1)	0.2403	1.9301	0.9357	0.2403
(1,2,1)	0.2003	1.9301	0.9467	0.2003
(1,3,1)	0.2045	1.9301	0.9455	0.2045
(2,3,1)	0.2029	1.9301	0.946	0.2029
(1,2,3,1)	0.199	1.9301	0.947	0.4461
The combination selected by the champion	**(1,2,3,1)**	/	**(1,2,1)**	**(1,2,1)**
(1,2)	0.2798	1.3948	0.8941	0.2798
(2,2)	0.3557	1.3948	0.8631	0.3557
(3,2)	0.2176	1.3948	0.9187	0.2176
(1,2,2)	0.1539	1.3948	0.9432	0.1539
(1,3,2)	0.1511	1.3948	0.9443	0.1511
(2,3,2)	0.1516	1.3948	0.9441	0.1516
(1,2,3,2)	0.1509	1.3948	0.9443	0.3885
The combination selected by the runner-up	**(1,2,3,2)**	/	**(1,3,2) (1,2,3,2)**	**(1,3,2)**
(1,3)	0.5332	2.5206	0.888	0.5332
(2,3)	0.7006	2.5206	0.8497	0.7006
(3,3)	0.3811	2.5206	0.9213	0.3811
(1,2,3)	0.1674	2.5206	0.9662	0.1674
(1,3,3)	0.1376	2.5206	0.9723	0.1376
(2,3,3)	0.1479	2.5206	0.9702	0.1479
(1,2,3,3)	0.1121	2.5206	0.9775	0.3349
The combination selected by the third place	**(1,2,3,3)**	/	**(1,2,3,3)**	**(1,3,3)**
(1,4)	0.312	1.8678	0.9127	0.312
(2,4)	0.4193	1.8678	0.8806	0.4193
(3,4)	0.223	1.8678	0.9384	0.223
(1,2,4)	0.1308	1.8678	0.9643	0.1308
(1,3,4)	0.1247	1.8678	0.966	0.1247
(2,3,4)	0.1263	1.8678	0.9656	0.1263
(1,2,3,4)	0.1237	1.8678	0.9663	0.3517
Selected combination averagely	(1,2,3,4)	/	(1,3,4)	(1,3,4)

In summary, we extracted the performance law that changes over time from the data trend of the 15th-the 30th Summer Olympic men's 110-meter hurdles, and well predicted the performance of the top three in the 31th Olympic Games.

On the whole the time segment that scores change the fastest is the 15th to the 24th session. Although there are fluctuations, the scores variation amplitude is relatively large and develops toward better results. The scores increase relatively slow from the 24th to the 30th Olympic Games. As can be seen from the predicted results, the Olympics scores is growing over time, but the 31th Olympic Games in 2016 still cannot reach the 12.88 records by Liu Xiang in World Championship and there is hope to refresh his 12.91 Olympic records.

CONCLUSIONS

The linear regression model of optimal combination well explains the achievements variation trends and laws of the Summer Olympic Games men's 110-meter hurdles. It predicted the top three scores in the 31th

Summer Olympic Games of this project, forecast achievements of the 31th Olympic champions is 12.91, forecast achievements of the runner up is 13.09, forecast achievements of the third place is 13.11, and the results can be used as the target value of training decisions; We can use the variables combination method to conduct regression analysis, and then optimize the portfolio based on its statistical parameters. This optimization model can be closer to the actual problem.

The model can be applied to other sports data processing that changes over time. With the development of science the determined variable quantity will continue to increase. In order to make better use of new scientific results and make it play a role in sports competitions and research, to use the optimal combination method is a correct way.

REFERENCES

[1] Baoren Che, Hongjiang Li, Chongxi Tai; Track and field. Beijing: Higher Education Press, 118- 189 **(2000)**.

[2] Bing Zhang, Yan Feng; The Special Quality Evaluation of the Triple Jump and the Differential Equation Model of Long Jump Mechanics Based on Gray Correlation Analysis. International Journal of Applied Mathematics and Statistics, **40(10)**, 136-143 **(2013)**.

[3] Chungen Lu, Ziqing Liu; 110m hurdle technical characteristics of Shi Dong-peng. Shandong Sports Science & Technology, **30(1)**, 41- 43 **(2008)**.

[4] Haifeng Yan; A Comparative Study on Technical Parameters of China and World Outstanding 110m Hurdlers. Journal of Hubei Sports Science, **04**, **(2004)**.

[5] Weilan Huang; Structure and Arrangement Analysis of Specific Abillity on Field Evente in AthleticsÿJournal of Guangzhou Physical Education Institue, **25(2)**, 95-97 **(2001)**.

[6] Weishou Zhang, Lihua Wang, Xiaofeng Xu; Changes of whole speed of elite men's 110-meter hurdle runners in China. Journal of Shandong Institute of Physical Education and Sports, **23(5)**, 101-103 **(2007)**.

[7] Yunling Deng, Guanglin Zhang, Zhongbing Wang; Building on male 110m hurdler's special ability index system and performance predicted modelÿ Journal of Northwest Normal University (Natural Science), **43(2)**, 98-101 **(2007)**.