Transactions on Comparative Education (2021) Vol. 3: 27-30
Clausius Scientific Press, Canada
DOI: 10.23977/trance.2021.030405
ISSN 2523-5818

Correlation Analysis and Model Establishment of English Achievement

Chunxia Wu, Xinhong Liu*, Shichao You

Beijing Institute of Petrochemical Technology, Beijing, 102617, China

*Corresponding author: liuxinhong@bipt.edu.cn

Keywords: linear regression, English achievement, regression mode.

Abstract: Using the method of regression analysis in mathematical statistics, this paper makes a correlation analysis on the English achievements of some college students in our university, establishes a regression prediction model between CET4 and English academic achievements, and forecasts the achievements of some students.

1. Introduction

CET-4 is a national teaching test presided over by the Department of higher education of the Ministry of education. It has a certain degree of recognition in the society. Our college attaches great importance to CET4, Over the years, teachers and teaching management departments have been studying and exploring teaching reform and process management, and achieved more research results. Students' English level has also been greatly improved. This paper collects, arranges and analyzes the English test scores of some professional students in our University (the data is up to September 2019), analyzes the correlation between English scores and CET-4 scores through sampling, establishes a multiple regression prediction model between CET-4 scores and relevant English scores, and forecasts the scores of some students.

2. Correlation analysis of English achievement

Some professional students from 2015 to 2016 were selected to collect and sort out their English academic achievements, namely, college English viewing,listening and speaking(1) X_1 ,college English viewing, listening and speaking(II) X_2 ,college English reading, writing and translation(1) X_3, college English reading, writing and translation(II) X_4, college English comprehensive training(1) X_5, college English comprehensive training(II) X_6, college entrance examination scores X_7 and CET-4 scores Y . Suppose that the English scores obey the normal distribution and are independent of each other. Through the correlation analysis of the score data by using the statistical software SAS, the correlation coefficient between CET4 Y and each English score $X_k (k=1,2,\cdots,7)$ and the P value of hypothesis test for the correlation coefficient are obtained.

Table 1: Correlation analysis between CET-4 and English scores

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
correlation coefficient	0.6279	0.4549	0.664	0.599	0.592	0.564	0.719
P value	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

The results of correlation analysis show that there is a positive correlation between CET-4 and English scores, and both have statistical significance. That is, the English foundation before entering school and the English academic achievement during college ultimately affect the score of CET-4, but the correlation coefficient between the score of college entrance examination and CET-4 is the largest, indicating that it is more favorable for students with good English foundation to pass CET-4 in the first semester of freshman year.

3. Regression prediction model

In order to better study the relationship between English scores and CET-4 scores and provide better suggestions and help for students, this paper makes a regression analysis on the selected students' English scores. Multiple regression analysis is an effective mathematical method to deal with the interdependence between multiple variables. Using this method, the regression prediction model is established with the help of statistical software SAS, and the rationality of the model and the significance of regression coefficient are tested.

3.1 Multiple linear regression model

In practical problems, random variables are often related to multiple ordinary variables $x_1, x_2, \cdots, x_n$, and the multiple linear regression model is

$$Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_n x_n + \varepsilon, \quad \varepsilon \sim N(0,1).$$

The maximum likelihood estimation and least square method are used to obtain the estimated values of the coefficients of the regression model $\hat{b}_i (i = 1, 2, \cdots n)$, Get regression equation

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \hat{b}_3 x_3 + \cdots + \hat{b}_n x_n$$ [1]

3.2 Regression prediction model between CET-4 and English academic achievement

The CET-4 score is set as a random variable Y, and the common variables are college English viewing, listening and speaking(1) x_1, college English viewing, listening and speaking(II) x_2, college English reading, writing and translation(1) x_3, college English reading, writing and translation(II) x_4, college English comprehensive training(1) x_5, college English comprehensive training(II) x_6, college entrance examination scores x_7. Regression analysis is carried out on the English scores of students majoring in chemical engineering and Technology (cet-16). Using the statistical software SAS and the stepwise regression method [2], the final variables x_1, x_5, x_6, x_7 enter the model at the selection level $\alpha = 0.15$ and elimination level $\beta = 0.15$. The results of analysis of variance and parameter estimation are as follows:

Table 2: Stepwise regression analysis of variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob > F
Model	4	60819	15205	56.07	< .0001
Error	22	5966.2	271.19		
U Total	26	66785			

Table 3: Stepwise regression parameter estimation and significance test

Variable	Parameter Estimate	Standard Error	F Value	Pr≥F
Intercept	141.10	25.57	30.44	<.0001
x_1	3.441	0.43	63.26	<.0001
x_5	1.130	0.58	3.73	0.0666
x_6	1.280	0.56	5.24	0.0321
x_7	2.128	0.24	76.93	<.0001

The analysis of variance in Table 2 shows that $F = 56.07$, $P < 0.0001$, It shows that the regression model of CET-4 is statistically significant, and the determination coefficient of the regression equation is $R^2 = 60819 / 66785 = 0.91$, indicating that the fitting effect is good and the regression effect is significant; Table 3 shows the estimation of regression coefficient by stepwise regression, which shows that college English viewing, listening and speaking (1) x_1 ,college English comprehensive training (1) x_5 , college English comprehensive training (II) x_6 , college entrance examination scores x_7 have a significant impact on CET4 scores, and the regression effect is good. The expression of the regression equation is:

$$\hat{y}_1 = 141.10 + 3.441x_1 + 1.1300x_5 + 1.28x_6 + 2.12x_7$$

This equation can be used to predict the CET-4 scores of students who will take part in CET-4.

3.3 Prediction results

Take the scores of the first five people from the sample data, and use the above regression prediction model to calculate their actual test scores and prediction scores of CET4 and the prediction interval with confidence level of 0.95. The results are shown in Table 4.

Table 4: Four grade test scores and forecast

scores of CET-4	Prediction score	Predicted value interval
449	451.17	(421.7, 489.6)
281	299.54	(260.4, 338.7)
431	425.73	(389.6, 461.8)
434	435.27	(397.0, 473.5)
498	505.2	(468.3, 542.0)

Through the actual calculation and comparison of some data, all the predicted values are within the prediction interval, and the error between the predicted value and the actual value is not large,

indicating that the established regression prediction model has a certain reference value.

4. Conclusion

(1) Through the correlation analysis and model establishment and prediction of students' CET-4 scores and English scores, the analysis results are in line with the expectations, that is, students who pass CET-4 in the first semester of freshman year tend to have a good English foundation, and their English scores in the college entrance examination are relatively high. After admission, they naturally have a good score in the course college English viewing, listening and speaking(1); For students with relatively weak foundation in senior high school, through the study of a series of courses, especially the study and intensive training of college English comprehensive training(1) and college English comprehensive training(II), their English level has been improved, which is very helpful for them to pass CET4. On the contrary, college English viewing, listening and speaking(II), college English reading, writing and translation(1), college English reading, writing and translation(II) are not selected into the model, which theoretically shows that students do not pay enough attention to these courses.

(2) The data analysis reflects the learning status of students. Surprise learning is common. It is suggested that students should treat the learning of all courses step by step and solidly, especially strengthen the learning of reading, writing and translation, so as to really improve their English level.

(3) Using the stepwise regression method in regression analysis, this paper establishes a multiple regression prediction model between CET-4 and relevant English scores on the sampling data, and tests the hypothesis of the model and regression coefficient, but the model only considers the relevant English scores and does not consider factors other than scores, such as the time of learning a foreign language and the degree of hard study, The model needs to be further improved.

Acknowledgements

The author is very grateful to the teachers of the Academic Affairs Office for their strong support and help.

References

[1] Ju Sheng, Shiqian Xie, et al. Probability theory and mathematical statistics [M]. Beijing, higher education press, 2009

[2] Ying Zhang, Yixiong Lei. Practical course of SAS software [M]. Beijing, Science Press, 2009