Based on the Principal Component Analysis and Ridge Regression Equation to Explore the Influencing Factors of English Ability in Chengdu Technological University, China

Ziqian Liu[1], Feilong Qin[1,4(✉)], Zheng Zeng[2(✉)], Shilin Wang[3], Jie He[1], Ke Wang[1], and Liuli Lu[1]

[1] School of Big Data and Artificial Intelligence, Chengdu Technological University, Chengdu 611730, China
lida_112@163.com

[2] School of Foreign Languages and International Education, Chengdu Technological University, Chengdu 611730, China
zzheng2@cdtu.edu.cn

[3] School of Network and Communication Engineering, Chengdu Technological University, Chengdu 611730, China

[4] College of Mathematics and Science, University of Electronic Science and Technology of China, Chengdu 611731, China

Abstract. In order to explore the evaluation model of college students' English ability, the questionnaire survey in this paper data statistics on interest in English reading, grammatical analysis ability, reading skills, etc. Then this paper uses Principal Component Analysis (PCA) to statistical modeling and analyzes the English ability of students in Chengdu Technological University, China. Using Ridge Regression to explore and analyze the factors that affect English grade, and establish a prediction model of variables and performance. This article aims to provide new ideas and methods for teachers teaching in different levels of classes and for non-native English-speaking students in the process of learning English.

Keywords: Evaluation model · Statistical modeling · English ability · Principal component analysis · Ridge regression

1 Introduction

With the acceleration of globalization, English is playing an increasingly important role. Many industries gradually have cross-relationships with English. For example, China's Education Modernization 2035 [1] pointed out that it is necessary to realize the modernization of education and step into an educational power. In addition, the progress and application of science and technology are closely related to English. At present, many universities and educational services are gradually paying more attention to the cultivation of students' English ability.

V. Balakrishnan et al. (Eds.): ICMEIM 2022, AHSSEH 6, pp. 729–739, 2022.
https://doi.org/10.2991/978-94-6463-044-2_92

In the current related research, the analysis and research on the influencing factors of English reading ability are relatively scattered. In particular, there are few studies on the teaching policy planning of classes with different English proficiency and the related influencing factors of English achievement. And most of the research is biased towards the theoretical aspect, and there is no mathematical statistical method or other formula to prove the authority of the theory. Secondly, some authors such as Yujun Ren and Yue Yang use cluster analysis to study the relevant factors what will affect English proficiency [2], but this method is difficult to obtain clustering conclusions when the sample size is too large, and it is difficult to obtain accurate data conclusions when the sample size is too small. In addition, since the similarity coefficient is based on the reflection of the subjects to establish an index reflecting the internal connection between the subjects, but in practice, although the data obtained from the subjects' reflections sometimes find that there is a close relationship between them, there is no relationship between things. At this time, it is obviously inappropriate to obtain cluster analysis results based on distance or similarity coefficients, but the cluster analysis model itself cannot identify such errors. Therefore, this paper uses the principal component analysis method to extract the main components to express the comprehensive English ability. And finally uses the ridge regression to avoid the collinearity of the variables to obtain the prediction model between the English grades and the variables. On this basis, combining the model to predict English grades is conducive to the rapid improvement of students' English ability, and has important relevant teaching application value.

2 Objects and Methods

2.1 Data Collection

The subjects of this survey are the 2020 students of Chengdu technological university, China (2 basic classes of college English courses (average score in final assessment: 70.28), 2 extended classes of college English courses (average score in final assessment: 74.53), a total of four classes). During the class, the teacher sent students to answer in the form of questionnaires and test questions, and a total of 166 questionnaires were distributed (numbers 1–76 for the extended class, 77–166 for the basic class). The data collected include vocabulary size, grammatical structure, time spent reading extracurricular books per day, average number of English articles read in a month, time required to read an English article with a CET-4 difficulty level, time to capture key information, Attitude towards extracurricular reading, how many foreigners you have communicated with in English, and final English grades. Among them, the grammatical structure and extracurricular reading attitude were converted by the expert scoring method [4], and solicit the opinions of many teachers in the foreign language college of our school to discuss the score. In order to obtain objective and real feedback on teaching effects, and further improve the quality of course teaching, the collected materials meet the following standards: (1) Collect certain samples in each of the basic class and the extended class; (2) Fill in the data completely; (3) Samples The total number of data is more than 10 times greater than the total number of variables; (4) All students who participated in the questionnaire are known.

The test of Cronbach's α coefficient of data as shown in Table 1.

Table 1. Test of Cronbach's α coefficient table

Cronbach's α coefficient	Standardized Cronbach's α coefficient	number of variables	Number of samples
0.841	0.901	8	166

Table 2. KMO test and Bartlett test

KMO value		0.907
Bartlett sphericity test	approximate chi-square	1597.233
	df	28
	p	0.000***

Note: ***, **, * represent the significance levels of 1%, 5%, and 10%, respectively

From Table 1, the Cronbach's α coefficient value of the model is 0.841, indicating that the reliability of the questionnaire is good.

2.2 Investigation Method

The form of the questionnaire adopts the self-filling method, so that the students participating in the questionnaire are not affected by external factors and can express their thoughts truly. Secondly, In the self-administered questionnaire, since the questions are raised using standard vocabulary, and everyone sees the same questions, so there is no subjective arbitrariness and inducement in the interpretation of the investigators. And effectively improve the validity and reliability of the questionnaire [5].

The test of KMO test and Bartlett's as shown in Table 2.

From Table 2, the data collected by the questionnaire were tested by KMO and Bartlett, and the obtained KMO value reached $0.907 > 0.9$. The results of Bartlett's sphericity test show that the significant P value is $0.000*** < 0.05$, which is significant at the level, rejecting the null hypothesis, and there is a correlation between the variables. Prove that there is a high degree of correlation between independent variables, so it is very suitable to use the principal component analysis method to achieve dimensionality reduction analysis and statistical modeling.

3 Analysis of the Influencing Factors of College English Course Learning

This research uses R language for data sorting, principal component analysis, ridge regression analysis and model testing. And the following X_7 and X_8 variables are processed as negative indicators, and the remaining variables are processed as positive indicators.

Table 3. Correlation coefficient matrix R of variables

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1	0.268	0.913	0.882	0.927	0.017	0.962	0.905
X_2	0.268	1	0.298	0.196	0.264	0.093	0.239	0.232
X_3	0.913	0.298	1	0.796	0.836	0.070	0.869	0.827
X_4	0.882	0.196	0.796	1	0.848	0.011	0.851	0.781
X_5	0.927	0.264	0.836	0.848	1	0.086	0.909	0.876
X_6	0.017	0.093	0.070	0.011	0.086	1	0.052	0.039
X_7	0.962	0.239	0.869	0.851	0.909	0.052	1	0.875
X_8	0.905	0.232	0.827	0.781	0.876	0.039	0.875	1

Table 4. Principal component load coefficient

	F1	F_2	F_3	F_4	F_5	F_6	F_7	F_8
Feature vector	0.422	−0.062	−0.016	−0.038	−0.071	−0.144	0.199	−0.866
	0.137	0.517	0.840	0.053	0.066	0.018	0.037	0.013
	0.397	0.018	0.007	−0.203	−0.823	0.127	−0.270	0.183
	0.388	−0.098	−0.071	0.817	0.054	0.392	0.030	0.101
	0.408	0.001	−0.057	0.006	0.374	−0.448	−0.694	0.082
	0.028	0.846	−0.529	0.030	0.002	0.032	0.040	−0.038
	0.413	−0.044	−0.062	−0.047	0.024	−0.492	0.627	0.430
	0.397	−0.050	−0.050	−0.532	0.414	0.605	0.078	0.108
Eigenvalues	5.444	1.051	0.873	0.223	0.181	0.115	0.088	0.025
Contribution rate	0.6806	0.1313	0.1091	0.0278	0.0226	0.0144	0.0109	0.0031

3.1 Establishment of Principal Component Analysis and the Influencing Factors Analysis

In the following table, X_1 represents vocabulary number, X_2 represents grammatical structure, X_3 represents the time spent reading extra-curricular books every day, X_4 represents the number of English articles read in a month, X_5 represents the attitude of extra-curricular reading, X_6 represents the number of foreigners who have communicated in English, X_7 represents Average minutes of reading a 300 words English article, and X_8 represents the time to capture a key information.

Eight variables that may affect English ability are designed in the questionnaire. After obtaining the correlation matrix (shown in Table 3) by the formula XTX, using principal component analysis to processing the correlation matrix, and obtain variable eigenvalues and eigenvectors (shown in Table 4).

3.1.1 Modeling by Principal Component Analysis

Sorting the eigenvalues, namely: 5.444 > 1.051 > 0.873 > 0.223 > 0.181 > 0.115 > 0.088 > 0.025, and the cumulative contribution rate of the first four eigenvalues is about 81% > 80%. Therefore, the first two principal components can be used to approximately describe the English reading level of students (the principal components are named F1 and F2 below).

The principal component loading matrix is composed of the eigenvectors of the eigenvalues of the correlation matrix, which can reflect the correlation between each principal component and the variable. And the two principal component expressions can be obtained from the loading matrix as follows:

$$F_1 = 0.422X_1 + 0.137X_2 + 0.397X_3 + 0.388X_4 + 0.408X_5 + 0.028X_6 + 0.413X_7 + 0.397X_8$$

$$F_2 = -0.062X_1 + 0.517X_2 + 0.018X_3 - 0.098X_4 + 0.001X_5 + 0.846X_6 - 0.044X_7 - 0.050X_8$$

Note: ZX_i represents the individual ability contribution of the i indicator; F_i represents the i principal component.

According to the formula:

F_1 has larger load values at X_1, X_3, X_4, X_5, X_7, and X_8 and is positively correlated. This principal component is mainly from vocabulary, time spent reading extracurricular books every day, the average number of English articles read in one month, the attitude of extracurricular reading, the average number of minutes to read an English article with a difficulty of CET-4, capture the length of a key message in these six aspects reflects students' interest in learning and the basic level of English and reading skills strategies.

F_2 has a larger load value and a positive correlation at the X_2 and X_6 variables. This principal component mainly reflects the students' English communication ability from two aspects: grammatical analysis ability and how many foreigners have communicated in English, and is named as communication ability.

The next step is to analyze the English ability of students in the basic class and the expansion class through the scores of each main component, and compare the level of the two classes.

3.1.2 The Influencing Factors Analysis

Bring the standardization data into the first 2 principal component expressions to get the scores of the first 2 principal components. But because there are too many samples, so only the top ten students ranked in each principal component score are counted (shown in Table 5 and Table 6).

Among the top 10 students with the F_1 score of the principal component, seven students are from the extended class, and the remaining three students are from the basic class. This shows that the students in the extended class are better than the students in the basic class in terms of their learning interest, basic English level and reading skills.

Among the top 10 students with the F_2 score of the principal component (Table 6), two students are from the extended class, and the remaining eight students are from the basic class. This shows that the level of communication ability of the students in the basic class is higher than that of the students in the extended class.

Table 5. Principal component score table (F_1 score in descending order)

Sample number	F_1 principal component score
53	2.452
44	2.363
43	2.323
62	2.317
107	2.251
83	2.231
92	2.226
55	2.219
68	2.198
57	2.180

Table 6. Principal component score table (F_2 score in descending order)

Sample number	F_2 principal component score
166	1.255
14	0.506
121	0.503
159	0.489
126	0.480
130	0.479
142	0.473
31	0.469
136	0.467
109	0.467

3.1.3 The Suggestions of Data Analysis

From the above principal component score ranking, we can know the comprehensive level of the students in the extended class and the basic class. From the descending order of the scores of principal component 1 and principal component 2, it can be seen that the students' interest in learning and the basic level of English in the extended class are higher than those in the basic class, but the level of speaking ability is lower than that of the basic class. It can be seen that the teaching content of the teachers in the extension class is more inclined to cultivate students' ability to take exams. Teachers and students spend most of their time preparing for English exams that they don't spend much time in the classroom to improve students' oral English ability, which will cause students to

become “dumb English”. English is a language with both practicality and humanity 6, and the cultivation of comprehensive ability is very important. Therefore, based on the current situation, the teaching strategies in the basic and extended classes can be referred to as follows:

(1) Teachers should encourage or guide students to establish reading habits according to their areas of interest in the course of teaching basic classes [7], and expand students’ knowledge dimension and develop reading habits. Secondly, encouraging teachers use appropriate time to evaluate each student’s reading ability, so that students’ reading training should match their own ability.
(2) Teachers should encourage students in the development class to speak freely and bravely in the class. Secondly, Combining extracurricular resources and British and American culture to build a real and effective language environment [7], such as setting up flexible and diverse teaching modes, situational dialogue and role play to stimulate students’ thirst for knowledge and curiosity in the process of learning English, and also stimulate students’ initiative in English learning [9], helping students regain their confidence in learning English.
(3) The teacher should spend more time to solidify the English foundation of the students in the basic class, such as popularizing knowledge of topics in reading materials, understanding knowledge of foreign social culture and knowledge of article structure. In this way, we can better and faster master reading skills such as “quick reading”, “skimming”, “skip reading” and “timed reading”, so as to grasp and understand the structure and context of the article in a short time and understand the article accurately [9].
(4) Using the online and offline mixed teaching mode [11], let students complete the learning of the basic knowledge of the course before class (online), and discuss with the teacher the problems left during online learning during the class (offline) time Consolidate knowledge and complete after-school (online) homework or tasks after class. The English teaching model based on the mixture of online and offline can increase students’ interest in English learning, improve teachers’ teaching quality and improve students’ learning efficiency, and transform “teaching” as the center into “student-centered”.
(5) Cultivate teachers’ professional level and practical ability, and use multimedia to assist teaching in the process of classroom teaching [12]. Because of demonstration method allows students to immerse themselves in the situation and understand a thing more intuitively. In addition, using multimedia teaching can let teachers save the time of writing lesson plans to answer questions for students or complete other things.

At present, the influence of relevant factors on English performance assessment is unknown, so the following is to explore the correlation between English performance and 8 variables by establishing a Ridge Regression model.

Table 7. The results of ridge regression analysis

K = 0.143	Unstandardized coefficient		Standardized coefficient	*t*	*p*	R^2	Adjustment R^2	*F*
	B	Standard error	Beta					
Constant	0.132	0.009	−0.009	14.774	0.000***	0.979	0.978	930.81(0.000***)
X_3	0.3	0.016	0.252	18.291	0.000***			
X_7	0.127	0.01	0.157	12.166	0.000***			
X_8	0.104	0.013	0.114	8.266	0.000***			
X_6	−0.021	0.023	−0.009	−0.911	0.363			
X_1	0.228	0.009	0.256	24.807	0.000***			
X_5	0.071	0.009	0.111	8.102	0.000***			
X_4	0.11	0.013	0.117	8.508	0.000***			
X_2	0.006	0.004	0.016	1.545	0.090			
Dependent variable: English final grade								

Note: ***, **, * represent the significance levels of 1%, 5%, and 10%, respectively

3.2 Ridge Regression Model Establishment and Test

According to the determined variables, establish a regression model: $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8$. Using dimension reduction method to select feature, and adding the regular term on the basis of the squared error that can get the regression coefficient: $\omega = (X^T X + \lambda I)^{-1} X^T y$. The following uses Y (English grade) to establish a regression model for 8 variables.

Through the establishment of ridge regression model, the analysis results are shown in Table 7.

The results of ridge regression show that based on X_3, X_7, X_8, X_1, X_5, X_4 and X_2 regression models, the significance *p* value is 0.000 * * *, which is significant at the level, and the original hypothesis is rejected, indicating that there is a regression relationship between independent variables and dependent variables. At the same time, the goodness of fit of the model R^2 is 0.979, and the model is relatively excellent, so the model basically meets the requirements. However, the p value of partial regression coefficient b_6 is greater than 0.1, indicating that the independent variable X_6 has no significant impact on the model. Secondly, combined with the current evaluation of English in China, oral communication ability has not been forcibly included in the evaluation range of English performance, so X_6 is excluded here. The formula of the prediction model is:

$$y = 0.132 + 0.3X_3 + 0.228X_1 + 0.127X_7 + 0.11X_4 + 0.104X_8 + 0.071X_5 + 0.006X_2$$

The original data graph and model fitting value of this model in a visual form shown in Fig. 1.

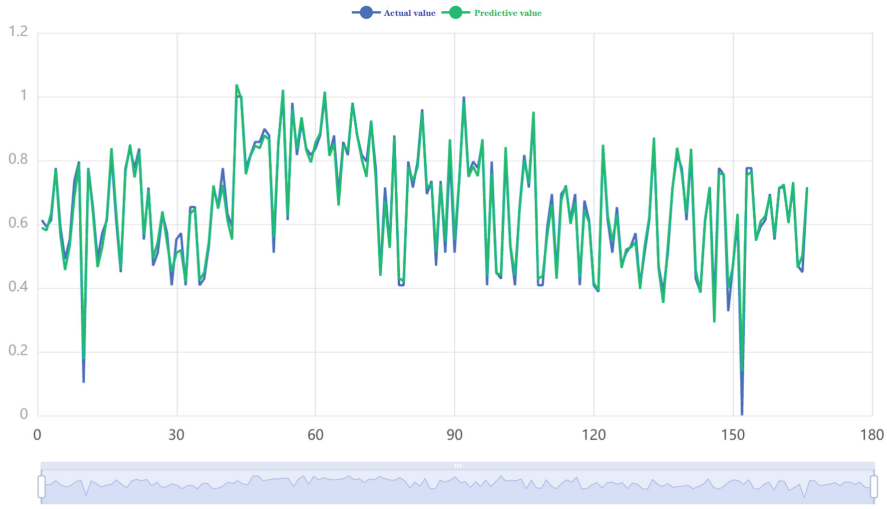

Fig. 1. Fitting effect diagram

3.3 Methods of Learning English

It can be seen from the model formula that Y (English score) is related to X_1 (vocabulary), X_2 (grammatical structure), X_3 (the time spent reading extracurricular books every day), X_4 (the average number of English articles read in one month), X_5 (Attitude towards extracurricular reading), X_7 (how many minutes to read an English article with a CET-4 difficulty on average), and X_8 (the time it takes to capture a key message) are all positively correlated. Among them, the variables X_3, X_4, and X_5 have a particularly obvious impact on English scores. It can be seen that the students themselves like English and their attitudes are very important. Therefore, in the early stage of cultivating interest, you can choose your favorite readings to expand your knowledge [13], and then choose "boring" articles in textbooks to challenge in the middle and later stages, forming a gradual adaptation process. Secondly, the influence of variables X_1 and X_2 on English performance ranks second. It can be seen that the basic level of students' English is very important. Using more fragmented time to memorize words, learning to summarize and use divergent thinking to construct a mind map to aid memory. Secondly, you should cheer yourself up more in your heart, tell yourself that you are in the golden age of memory, and treat learning with hope [14]. The influence of variables X_7 and X_8 on English performance ranks third, and effective reading strategies or skills can make students' English reading more effective. However, English is a language subject, and everyone's reading methods are different. Therefore, in the early stage, "deep reading" should be strengthened in the process of reading articles to form own reading system and method [15].

First of all, it is necessary to cultivate the interest in learning and have a positive attitude towards the subject of English. In the process of learning English, if you memorizing words and master the training reading skills, you will get twice the result with half the effort. Secondly, you must master enough vocabulary to ensure that you can

understand more than 70% of the English article, and then use reading methods and strategies such as guessing the words up and down to assist yourself in completing the understanding of the entire article.

4 Conclusion

This paper uses the principal component analysis method to analyze the factors that affect English ability, and provides new ideas and strategies for teachers to teach English to students with different English levels. Then this paper uses Ridge Regression to analyze the 8 factors affecting English performance assessment, providing intuitive and clear factors affecting English performance for English learners, and provides solutions for learners to effectively improve their English proficiency. However, due to the limited collection of data samples and the limited scope of knowledge involved in the questionnaire, there are more subdivided factors that have not been considered, so further detailed research is needed.

Acknowledgements. We are thankful to the First Geological Brigade of the Hubei Geological Bureau for providing the data. This study was supported by the Program of Science & Technology Department of Sichuan Province (No. 2021YJ0360, 2021YFG0170), the scientific research project of Chengdu Technological University (No. 2021ZR019, No. 20211103), the Key research base of Humanities and social sciences of Sichuan Provincial Department of Education (XJYX2020B15) and the project of Sichuan education and scientific research (SCJG20A022).

Bibliography

1. Zhang W., Education Modernization: Concept, System, System, Content, Method and Governance: Based on the Objectives and Tasks of "China's Education Modernization 2035" [J]. Journal of Jilin Normal University (Humanities and Social Sciences Edition), 2022,50 (01):51-58.
2. Ren Y J, Gao J Y. Research on the Comprehensive Evaluation of English Writing Scores of Independent College Students Based on Cluster Analysis: Taking Non-English Majors in the School of Information of Huaibei Normal University as an example [J]. Journal of Huaibei Normal University (Philosophy and Social Sciences Edition), 2018, 39(05): 99-103.
3. Yang Y., The application of cluster analysis method in the analysis of students' English learning level [J]. Education Science, 2008, 24(06): 50-53.
4. Ping W W., Research progress of Delphi method and its application in medicine [J]. Journal of Disease Control, 2003(03): 243-246.
5. Jia X Y., On the cultivation strategies of students' interest in English reading [J]. The Road to Talent, 2021(21):43-45.
6. Wu B J., Research on the Reading Characteristics of College Students in the New Era and the Reading Promotion Strategy of University Libraries—Taking the Medical College Library of Shaoguan University as an Example [J]. Heilongjiang Archives, 2022(01):315-317.
7. Sun S M., Strategies for Cultivating College Students' English Reading Habits in the New Era [J]. Heilongjiang Science, 2021,12(21):132-133.
8. Jia F., Problems and countermeasures of spoken English teaching in higher vocational colleges [J]. Popular Literature and Art, 2022(05):146-148.

9. Cheng L., Problem Analysis and Improvement Countermeasures in Higher Vocational English Teaching [J]. Exam Weekly, 2016(81): 82.
10. Xiong Y., Talking about the cultivation of college students' English reading skills [J]. Campus English, 2014(23):28-2
11. Gao M J., Analysis of the online and offline blended college English "golden class" teaching mode - Commentary on "Contemporary College English Teaching and Blended Learning Mode Exploration" [J]. Journal of Tropical Crops, 2021, 42(09): 2750.
12. Fang Y X., Explore English teaching strategies in multimedia environment from the perspective of teaching elements [J]. Motherland, 2017(20):1.
13. Wang J., Using classic reading materials to cultivate junior high school students' interest in English reading [J]. Campus English, 2020(12):183.
14. Sun M., Talking about how to effectively expand the English vocabulary of vocational students [J]. Campus English, 2021(12):61-62.
15. Yan H., The use of English reading strategies and the cultivation of ability [J]. Navigation of Arts and Sciences (Early Ten Days), 2022(04):31-33.