J. COMBIN. MATH. COMBIN. COMPUT. 127a (2025) 5673--5693

Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

A linear regression modeling study of university language education and students' expressive skills

Meiying He[1], ✉

[1] *School of Humanities, Zhejiang Guangxia Vocational and Technical University of Construction, Dongyang, Zhejiang, 322100, China*

ABSTRACT

This study aims to investigate the influence of university language education on students' expressive ability, and uses a questionnaire to collect the relevant factors affecting the relationship between students' expressive ability and university language education. The key principal factors were extracted from many variables by principal component analysis to simplify the data structure and retain the main information. Subsequently, a multiple linear regression model was constructed and the least squares method was applied to estimate the model parameters in order to quantitatively analyze the linear relationship between each principal component and students' expressive ability. In this paper, four principal factors, namely, "language organization ability, communication ability, language use ability and intonation ability", were identified under the principal component analysis technique, and their total variance explained reached 56.326%. It is found that the average score of students' expression ability is in the middle normal level, but the extreme difference of score between different students is as high as 27, which shows that there is a big gap between students' expression ability. The correlation coefficient between students' expressive ability and university language education is 0.8947, and the correlation coefficients of the four sub-dimensions of the two sig values are less than 0.01, indicating that the stronger the university language education, the higher the level of students' expressive ability. And the regression equation of students' expression ability and university language education is obtained as Y=0.893X-15.874.

Keywords: Questionnaire survey, Principal component analysis, Multiple linear regression model, Least squares method, Language education

1. Introduction

Regression analysis is an important statistical inference method. In practical application, medical medicine [1], agricultural production [2], economic development [3], industrial production [4],

✉ Corresponding author.
E-mail address: HMY123Z504@163.com (M. He).
Received 15 January 2024; Revised 10 February 2024; Accepted 30 November 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-316

business management [5], education and teaching [6] and many other fields need to use this method with the development of science, thus promoting the rapid development of regression analysis. Linear regression belongs to one kind of regression analysis, which is a common predictive analysis method in statistics, used to establish a linear relationship model between the independent variable and the dependent variable [7]. The model can be used to predict the value of the dependent variable by fitting a straight line or hyperplane, and its basic principle is to determine the linear relationship between the independent and dependent variables by the method of least squares [8-9]. The goal of the method of least squares is to minimize the sum of squares of the residuals of the predicted values of the model and the true observed values, and through the method of least squares, the coefficients and intercepts of the model can be obtained, so as to establish a linear regression model [10].

Language is a comprehensive discipline, which is not only a tool for learning words and language, but also a comprehensive embodiment of knowledge about culture, history, society and other aspects. At the university level, the teaching purposes and contents of language need to be more in-depth and comprehensive in order to cultivate students' language expression ability, literary appreciation and humanistic literacy [11]. In language teaching, expression is the basic means of communication, through which the two-way process of teacher's imparting knowledge to students and students' giving feedback to the teacher can be accomplished. In the language classroom, changing the traditional mode of teaching, seizing the golden period of students' expression ability, and emphasizing the cultivation of students' communication and expression ability can improve students' learning ability and the quality of classroom teaching [12-13]. Thus, expression as a fundamental skill, exploring students' expression ability in language education has a very important significance.

In order to deeply analyze the intrinsic connection between university language education and students' expression ability, this study designed and developed a questionnaire. With the help of principal component analysis, the most representative principal components were selected from the questionnaire data. Then, with these principal components as the independent variables and students' expressive ability as the dependent variable, a multiple linear regression model was established, and the least squares method was used to fit the model and solve the formal system of equations to obtain the parameter estimates of the multiple linear equations, so as to quantitatively analyze the linear dependence relationship between the principal components and students' expressive ability. Finally, the correlation between university language education and students' expressive ability is explored through regression analysis.

2. Multiple linear regression modeling of students' ability to express themselves

2.1. Study design

2.1.1. Survey Objects and Methods. In this study, 550 (male: 241, female: 309) teacher-training students majoring in Chinese language in their first to fourth year of college in 2 universities in N province were surveyed, 550 questionnaires were distributed, and the invalid questionnaires were excluded, and the final valid questionnaires we got were 543, with an effective rate of 98.73%.

The research methods used in this study are literature method, questionnaire survey method, and statistical analysis method.

1) Literature method. Taking full advantage of the information age, we studied the journals, theses and writings related to students' expressive ability, language cognitive load and language learning efficiency through Zhi.com, Wanfang and libraries, etc., and understood and clarified their origins, characteristics, as well as the theoretical development and application, which is the prerequisite for the formal writing of this thesis.

2) Questionnaire Survey Method. Drawing on the relevant scales and starting from the actual situation of the research subjects, the Questionnaire on Expressive Ability of College Students and the

Questionnaire on College Students' College Language Education were compiled and tested for reliability and validity, and the questionnaire surveys were carried out accordingly.
3) Statistical analysis method. SPSS software was used to make descriptive statistics, difference analysis, Pearson correlation analysis, regression analysis, and mediation effect analysis of the final questionnaire data.

2.1.2. Questionnaires. Students' expression ability is a comprehensive ability system, and university Chinese education is an important core of students' expression. In this paper, the questionnaire "Students' Expressive Ability Scale" was compiled, which included two index layers: language expression and text comprehension. The specific contents of the survey on students' expression ability include 10 contents in four aspects: "language use ability, stylistic writing ability, language organization ability, voice and intonation ability, communication ability, organization, consistency, emotional expression and attitude expression" in four aspects: "ability, speech ability, logical thinking ability and emotional attitude". The comprehension ability of students in university Chinese education includes 10 contents: "grasping word comprehension, semantic comprehension, sentence structure comprehension, sentence meaning, structure comprehension, subject comprehension, writing skills being understood, literary imagery comprehension, literary language comprehension, literary theme and ideological comprehension" in four aspects: word comprehension, sentence comprehension, text comprehension and literary appreciation. The content of the survey was divided into four stages according to its impact on expression ability, namely: 0-15 stages (with little impact); 16-22 phases (average impact); 23-30 phases (greater impact); 31-36 (very significant). The Students' Expressive Ability Scale is shown in Table 1.

Table 1. Student expression ability scale

Target layer	Dimension one	Dimension two
Verbal ability	Written ability	Writing ability
		Language ability
		Style writing ability
	Verbal ability	Language organization ability
		Voice tone
		Communication skills
	The logical thinking ability of expression	Rationalization
		Consistency
	Emotional attitude	Affective expression
		Attitude expression
College Chinese education	Word comprehension	Word-sense understanding
		Semantic understanding
	Sentence comprehension	Sentence structure understanding
		Sentence meaning
	Textual comprehension	Structure understanding
		Subject comprehension
		The writing technique is understood
	Literary appreciation	Literary image understanding
		Literary language understanding
		Literary themes and ideological understanding

2.2. *Principal Component Analysis*

2.2.1. Fundamentals and basic ideas:
1) Basic idea. Principal component analysis [14] is to take the method of language dimensionality reduction, to find a few composite variables to replace the original multitude of variables, so that these composite variables can represent as much as possible the amount of information of the original variables, and are not correlated with each other. If the first linear combination selected, i.e., the first

composite variable, is denoted as F_1, the “information” is measured by the variance, and the larger the $Var(F_1)$, the more information F_1 contains. Therefore, among all the linear combinations, F_1 should be the one with the highest variance, so F_1 is called the first principal component. If the first principal component is not enough to represent all the information of the original p variables, then consider selecting F_2 which is the second linear combination, in order to effectively reflect the original information, the information already in F_1 does not need to appear in F_2, expressed in the language language is the requirement of $Cov(F_1, F_2) = 0$, i.e., F_2 is the second principal component, and so on can be constructed to the third, fourth... ...the p th principal component.

2) Basic principles. According to the definition of the language model for principal component analysis, in order to perform principal component analysis, it is necessary to derive the principal component coefficients in order to obtain the principal component model based on the raw data, as well as the requirements of the three conditions of the model. This is the problem that needs to be solved to derive the principal components.

The model is first required to fulfill the following conditions:

(1) F_i, F_j are not correlated with each other $(i \neq j, i, j = 1, 2, 3 \cdots p)$;

(2) The variance of F_1 is greater than the variance of F_2 is greater than the variance of F_3;

(3) The contributions of $F_1, F_2, \ldots F_p$ add up to 1.

According to the condition (1) of the principal component language model requires that the principal components are uncorrelated with each other, and for this reason the covariance array between the principal components should be a diagonal array. Namely:

$$F = AX \tag{1}$$

Its covariance array should be:

$$\begin{aligned} Var(F) &= Var(AX) = (AX) \cdot (AX) \\ &= AXXA = \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix} \end{aligned} \tag{2}$$

Let the covariance array of the original data be V. If the original data is normalized, then the covariance array is equal to the correlation matrix, i.e., there is:

$$V = R = XX' \tag{3}$$

Then by the principal components of the language model conditions and the nature of the orthogonal matrix, if the conditions can be met the best requirement A for the orthogonal matrix, that is, to meet the $AA = I$. Thus, the covariance of the original data is obtained by substituting into the covariance matrix formula of the principal components:

$$Var(F) = AXX'A' = ARA' = \Lambda \tag{4}$$

$$ARA' = \Lambda RA' = A'\Lambda \tag{5}$$

Expanding the above equation yields:

$$\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{12} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{1p} & r_{2p} & \cdots & r_{pp} \end{pmatrix} \cdot \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{2p} \\ a_{12} & a_{22} & \cdots & a_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ a_{1p} & a_{2p} & \cdots & a_{pp} \end{pmatrix}$$
$$\begin{pmatrix} a_{11} & a_{21} & \cdots & a_{2p} \\ a_{12} & a_{22} & \cdots & a_{p2} \\ \vdots & \vdots & \vdots & \vdots \\ a_{1p} & a_{2p} & \cdots & a_{pp} \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix} \tag{6}$$

Expanding the left and right sides of the equation, by the property of equality of matrices, the equation derived here from the first column only is:

$$\begin{cases} (r_{11} - \lambda_1) a_{11} + r_{12} a_{12} + \cdots + r_{1p} a_{1p} = 0 \\ r_{21} a_{11} + (r_{22} - \lambda_1) a_{12} + \cdots + r_{2p} a_{1p} = 0 \\ \cdots\cdots\cdots \\ r_{p1} a_{11} + r_{p2} a_{21} + \cdots + (r_{pp} - \lambda_2) = 0 \end{cases} \tag{7}$$

In order to obtain the solution of this chi-square equation, the coefficient matrix determinant is required to be 0, i.e.:

$$\begin{vmatrix} r_{11} - \lambda_1 & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} - \lambda & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} - \lambda_1 \end{vmatrix} = 0 \tag{8}$$

$$|R - \lambda_1 I| = 0 \tag{9}$$

Obviously, λ_1 is the eigenvalue of the correlation coefficient matrix, and $a_1 = (a_{11}, a_{12}, \cdots a_{1p})$ is the corresponding eigenvector. Similar equations can be obtained based on the second column, the third column, etc., so that λ_2 is the p root of the equation, λ_3 is the eigenroot of the characteristic equation, and a_j is the component of its eigenvector. Then:

$$|R - \lambda I| = 0 \tag{10}$$

Let the p eigenroots of the correlation coefficient matrix R be $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p$ and the corresponding eigenvectors be a_j, then we have:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} \tag{11}$$

The variance with respect to F_1 is:

$$Var(F_1) = a_1 XX'a_1' = a_1 R a_1' = \lambda_1 \tag{12}$$

Similarly there is: $Var(F_i) = \lambda_i$, i.e., the variances of the principal components are decreasing in order. And the covariance is:

$$Cov\left(a_i'X', a_jX\right) = a_i'Ra_j = a_i\left(\sum_{a=1}^{L}\lambda_a a_a a_a'\right)a_1$$
$$= \sum_{a=1}^{L}\lambda_s\left(a_i a_\sigma\right)\left(a_a a_j\right) = 0, i \neq j \tag{13}$$

In summary, according to the proof the principal component covariance in principal component analysis should be a diagonal matrix whose elements on the diagonal are exactly equal to the eigenvalues of the correlation matrix of the original data, while the elements of the principal component coefficient matrix A are the eigenvectors corresponding to the eigenvalues of the correlation matrix of the original data. Matrix A is an orthogonal matrix. Thus, variable $\left(x_1, x_2, \cdots x_p\right)$ is transformed to obtain the new composite variable:

$$\begin{cases} F_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\ F_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \\ \cdots \\ F_p = a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p \end{cases} \tag{14}$$

The new random variables are uncorrelated with each other and have decreasing variance in that order.

2.2.2. Steps for counting wealth in principal component analysis. The matrix of sample observations is:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pp} \end{pmatrix} \tag{15}$$

Step 1: Normalize the raw data:

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{\sqrt{\operatorname{var}\left(x_j\right)}}\left(i = 1, 2, \cdots, n; j = 1, 2, \cdots p\right) \tag{16}$$

Among them:

$$\bar{x}_j = \frac{1}{n}\sum_{m=1}^{n}x_i, \operatorname{var}\left(x_j\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2, \left(j = 1, 2, \cdots p\right) \tag{17}$$

Step 2: Calculate the sample correlation coefficient matrix:

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{pmatrix} \tag{18}$$

For convenience, the correlation coefficient of the normalized data is assumed to be X even after normalization of the original data:

$$r_{ij} = \frac{1}{n-1}\sum_{i=1}^{n}x_{ij}x_{ij}$$
$$\left(j = 1, 2, \cdots p\right) \tag{19}$$

Step 3: Find the eigenvalues $(\lambda_1, \lambda_2, \cdots \lambda_p)$ and corresponding eigenvectors $a_i = (a_n, a_{i2}, \cdots a_p)(j = 1, 2, \cdots p)$ of the correlation coefficient matrix R using Jacobi's method.

Step 4: Select the significant principal components and write the principal component expressions.

Principal component analysis can get p principal components, however, because the variance of each principal component is decreasing, the amount of information contained is also decreasing, so the actual analysis, generally not to select p principal components, but according to the cumulative contribution rate of each principal component to select the first k principal components, where the contribution rate refers to the proportion of the variance of a principal component to the proportion of all the variance, which is the proportion of an eigenvalue to all the eigenvalues combined. Here the contribution ratio means the proportion of variance of a principal component to all variances, which is actually the proportion of an eigenvalue to all eigenvalues. That is:

$$\text{Contributing to Yukon} = \frac{\lambda_i}{\sum_{i=1}^{i} \lambda_i} \tag{20}$$

The larger the contribution rate, the stronger the information of the original variable contained in the principal component. The selection of the number of principal components k is mainly based on the cumulative contribution rate of the principal components, i.e., the cumulative contribution rate is generally required to reach more than 85%, so as to ensure that the composite variable can include the vast majority of information of the original variable.

Step 5: Calculate the principal component score.

According to the standardized raw data, according to each sample, respectively, substituting into the principal component expression, you can get the new data of each sample under each principal component, i.e., the principal component score. The specific form is as:

$$\begin{pmatrix} F_{11} & F_{12} & \cdots & F_{u} \\ F_{21} & F_{2n} & \cdots & F_{2x} \\ \vdots & \vdots & \vdots & \vdots \\ F_{n1} & F_{n2} & \cdots & F_{n} \end{pmatrix} \tag{21}$$

Step 6: Based on the data from the principal component scores, then further statistical analysis can be performed.

2.3. Statistical methods

2.3.1. Multiple linear regression models. Multiple linear regression model [15] is interpreted as a linear relationship between variable Y and multiple explanatory variables $X_1, X_2, \cdots, X_k$. The assumption that there is a linear relationship between the explained variable Y and multiple explanatory variables $X_1, X_2, \cdots, X_k$ is a multivariate linear function of the explanatory variables and is known as a multivariate linear regression model. Namely:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \mu \tag{22}$$

where Y is the explained variable (dependent variable), $X_j\,(j = 1, 2, \cdots, k)$ is the k explanatory variable (independent variable), $\beta_j\,(j = 0, 1, 2, \cdots, k)$ is the $k+1$ unknown parameter, which is the regression coefficient, and μ is the random error, also known as the residual, which is usually assumed to be $\mu \sim N(0, \sigma^2)$.

The linear equation between the expected value of the explained variable Y and the explanatory variable $X_1, X_2, \cdots, X_k$ is:

$$E(Y)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k \tag{23}$$

It is called the multivariate overall linear regression equation, or overall regression equation for short.

For n set of observations $Y_i, X_{1i}, X_{2i}, \cdots, X_{ki}\ (i=1,2,\cdots,n)$, the set of equations takes the form:

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\cdots+\beta_kX_{ki}+\mu_j,(i=1,2,\cdots,n) \tag{24}$$

To wit:

$$\begin{cases} Y_1=\beta_0+\beta_1X_{11}+\beta_2X_{21}+\cdots+\beta_kX_{k1}+\mu_1 \\ Y_2=\beta_0+\beta_1X_{12}+\beta_2X_{22}+\cdots+\beta_kX_{k2}+\mu_2 \\ \cdots\cdots \\ Y_n=\beta_0+\beta_1X_{1n}+\beta_2X_{2n}+\cdots+\beta_kX_{kn}+\mu_n \end{cases} \tag{25}$$

Its matrix form is:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}=\begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}+\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \tag{26}$$

To wit:

$$Y=X\beta+\mu \tag{27}$$

where:

$Y_{n\times1}=\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$ is the vector of observations for the explanatory variables;

$X_{n\times(k+1)}=\begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}$ is the matrix of observations of the explanatory variables;

$\beta_{(k+1)\times1}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$ is the vector of overall regression parameters; $\mu_{n\times1}=\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$ is the vector of random error terms. The overall regression equation is expressed as:

$$E(Y)=X\beta \tag{28}$$

Multiple linear regression analysis is the process of estimating each parameter in a model based on a sample of observations, and performing statistical tests on the estimated parameters and regression equations so that the regression model can be used for size prediction and analysis. Since the parameters $\beta_0,\beta_1,\beta_2,\cdots,\beta_k$ are all unknown, they can be estimated using the sample observations

$(X_{1i}, X_{2i}, \cdots, X_{ki}; Y_i)$. If the calculated parameter estimates are $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$, the parameter estimates are used to replace the unknown parameter $\beta_0, \beta_1, \beta_2, \cdots, \beta_k$ of the overall regression function to obtain the multivariate linear sample regression equation:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{kn} \tag{29}$$

where $\hat{\beta}_j \left(j = 0,1,2,\cdots,k \right)$ is the parameter estimate, $\hat{Y}_i \left(i = 1,2,\cdots,n \right)$ is the sample regression value or sample fit value for Y_i, and the sample estimate.

Its matrix expression is of the form:

$$\hat{Y} = X\hat{\beta} \tag{30}$$

Where $\hat{Y}_{n\times 1} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}$ is the column vector of $n \times 1$ rd order fitted values of the vector of sample observations of the explanatory variables Y ; $X_{n\times(k+1)} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}$ is the $n \times (k+1)$ th order sample observation matrix of the explanatory variables X; and $\hat{\beta}_{(k+1)\times 1} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$ is the column vector of $(k+1) \times 1$ th order estimates of the vector of unknown parameters β.

The deviation between the estimated value $\hat{Y}_i$ of the explanatory variable obtained from the sample regression equation and the actual observed value Y_i is called the residual e_i. Then:

$$e_i = Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_{2i} + \cdots + \hat{\beta}_{ki} X_{ki} \right) \tag{31}$$

When selecting samples, the sample capacity used for calculation must meet the following conditions:

(1) Minimum sample capacity: that is, from the principle of least squares and the principle of maximum nature, to obtain the parameter estimates, regardless of its quality, the lower limit of the required sample capacity. It follows that the above sample size should be greater than $n + 1$.

(2) Sample size that meets the basic requirements. From the perspective of parameter estimation, the sample size should be greater than 5 times the number of explanatory variables. From the perspective of validity of the test it should be greater than 30.

(3) The good properties of the model can only be theoretically proven with a large sample.

2.3.2. Least squares method for solving parameter estimates. The following assumptions must be met when the multiple linear regression model utilizes ordinary least squares [16-17] (OLS) for parameter estimation:

Assumption 1: Zero mean assumption: $E\left(\mu_i\right) = 0$, $i = 1,2,\cdots,n$, i.e.:

$$E(\mu)=E\begin{bmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_n\end{bmatrix}=\begin{bmatrix}E(\mu_1)\\ E(\mu_2)\\ \vdots\\ E(\mu_n)\end{bmatrix}=0 \tag{32}$$

Assumption 2: Same variance assumption (the variance of μ is the same constant):

$$Var(\mu_i)=E(\mu_i^2)=\sigma^2,(i=1,2,\cdots,n) \tag{33}$$

Assumption 3: No autocorrelation:

$$\begin{aligned}
&Cov(\mu_i,\mu_j)=E(\mu_i\mu_j)=0,(i\neq j,i,j=1,2,\cdots,n)\\
&E(\mu\mu')=E\left[\begin{bmatrix}\mu_1\\ \mu_2\\ \vdots\\ \mu_n\end{bmatrix}(\mu_1,\mu_2,\cdots,\mu_n)\right]\\
&=E\begin{bmatrix}\mu_1^2 & \mu_1\mu_2 & \cdots & \mu_1\mu_n\\ \mu_2\mu_1 & \mu_2^2 & \cdots & \mu_2\mu_n\\ \vdots & \vdots & \vdots & \vdots\\ \mu_n\mu_1 & \mu_n\mu_2 & \cdots & \mu_n^2\end{bmatrix}\\
&=\begin{bmatrix}E(\mu_1^2) & E(\mu_1\mu_2) & \cdots & E(\mu_1\mu_n)\\ E(\mu_2\mu_1) & E(\mu_2^2) & \cdots & E(\mu_2\mu_n)\\ \vdots & \vdots & \vdots & \vdots\\ E(\mu_n\mu_1) & E(\mu_n\mu_2) & \cdots & E(\mu_n^2)\end{bmatrix}\\
&=\begin{bmatrix}\sigma_\mu^2 & 0 & \cdots & 0\\ 0 & \sigma_\mu^2 & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \cdots & \sigma_\mu^2\end{bmatrix}=\sigma_u^2 I_n
\end{aligned} \tag{34}$$

Assumption 4: Random error term μ is uncorrelated with explanatory variable X:

$$Cov(X_{ji},\mu_i)=0,(j=1,2,\cdots,k,i=1,2,\cdots,n) \tag{35}$$

Assumption 5: The random error term μ follows a normal distribution with mean zero and variance σ^2:

$$\mu_i\sim N(0,\sigma_\mu^2 I_n) \tag{36}$$

Assumption 6: There is no multicollinearity between explanatory variables:

$$rank(X)=k+1\leq n \tag{37}$$

That is, the sample observations of each explanatory variable are linearly independent of each other, and the rank of the matrix of sample observations of explanatory variables X is the number of parameters $k+1$, which ensures that the estimate of parameter $\beta_0,\beta,\beta_2,\cdots,\beta_k$ is unique.

For a multiple linear regression model with k explanatory variables:

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\cdots+\beta_kX_{ki}+\mu_i\,(i=1,2,\cdots,n) \tag{38}$$

Setting $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ as the estimator of parameter $\beta_0, \beta_1, \cdots, \beta_k$, respectively, yields the sample regression equation:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki} \tag{39}$$

The residuals e_i of observation Y_i and regression $\hat{Y}_i$ are:

$$e_i = Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_{2i} + \cdots + \hat{\beta}_{ki} X_{ki} \right) \tag{40}$$

By the method of least squares $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ should minimize the sum of the squares of the residuals e_i of all observations Y_i and regressions $\hat{Y}_i$, even though:

$$\begin{aligned} Q\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k \right) &= \sum e_i^2 = \sum \left(Y_i - \hat{Y}_i \right)^2 \\ &= \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki} \right)^2 \end{aligned} \tag{41}$$

Obtain the minimum value. According to the extremum principle for multivariate functions, Q takes the first order partial derivative of $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ respectively and makes it equal to zero, denoted as:

$$\frac{\partial Q}{\partial \hat{\beta}_j} = 0, (j = 1, 2, \cdots, k) \tag{42}$$

Unfolding as:

$$\begin{cases} \dfrac{\partial Q}{\partial \hat{\beta}_0} = 2 \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki} \right)(-1) = 0 \\ \dfrac{\partial Q}{\partial \hat{\beta}_1} = 2 \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki} \right)\left(-X_{1i} \right) = 0 \\ \cdots\cdots \\ \dfrac{\partial Q}{\partial \hat{\beta}_k} = \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \cdots - \hat{\beta}_k X_{ki} \right)\left(-X_{ki} \right) = 0 \end{cases} \tag{43}$$

Simplify to obtain the following system of equations:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \Sigma X_{1i} + \hat{\beta}_2 \Sigma X_{2i} + \cdots + \hat{\beta}_k \Sigma X_{ki} = \Sigma Y_i \\ \hat{\beta}_0 \Sigma X_{1i} + \hat{\beta}_1 \Sigma X_{1i}^2 + \hat{\beta}_2 \Sigma X_{2i} X_{1i} + \cdots + \hat{\beta}_k \Sigma X_{ki} X_{1i} = \Sigma X_{1i} Y_i \\ \cdots\cdots \\ \hat{\beta}_0 \Sigma X_{ki} + \hat{\beta}_1 \Sigma X_{1i} X_{ki} + \hat{\beta}_2 \Sigma X_{2i} X_{ki} + \cdots + \hat{\beta}_k \Sigma X_{ki}^2 = \Sigma X_{ki} Y_i \end{cases} \tag{44}$$

The above $(k+1)$ equation is called the regular equation and its matrix form is:

$$\begin{bmatrix} n & \Sigma X_{1i} & \Sigma X_{2i} & \cdots & \Sigma X_{ki} \\ \Sigma X_{1i} & \Sigma X_{1i}^2 & \Sigma X_{2i} X_{1i} & \cdots & \Sigma X_{ki} X_{1i} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \Sigma X_{ki} & \Sigma X_{1i} X_{ki} & \Sigma X_{2i} X_{ki} & \cdots & \Sigma X_{ki}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \Sigma Y_i \\ \Sigma X_{1i} Y_i \\ \vdots \\ \Sigma X_{ki} Y_i \end{bmatrix} \tag{45}$$

Because:

$$\begin{bmatrix} n & \sum X_{1i} & \sum X_{2i} & \cdots & \sum X_{ki} \\ \sum X_{1i} & \sum X_{1i}^2 & \sum X_{2i}X_{1i} & \cdots & \sum X_{ki}X_{1i} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum X_{ki} & \sum X_{1i}X_{ki} & \sum X_{2i}X_{ki} & \cdots & \sum X_{ki}^2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ X_{k1} & X_{k2} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix} \tag{46}$$

$$= X'X \begin{bmatrix} \sum Y_i \\ \sum X_{1i}Y_i \\ \vdots \\ \sum X_{ki}Y_i \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ X_{k1} & X_{k2} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = X'Y$$

Let $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$ be the vector of estimates and the sample regression model $Y = X\hat{\beta} + e$ with both sides multiplied by the transpose matrix X' of the matrix of sample observations X, then we have:

$$X'Y = X'X\hat{\beta} + X'e \tag{47}$$

to obtain a regular system of equations:

$$X'Y = X'X\hat{\beta} \tag{48}$$

By Assumption 6, $R(X) = k+1$, $X'X$ are square matrices of order $(k+1)$, so $X'X$ is full rank and the inverse matrix $(X'X)^{-1}$ of $X'X$ exists. Thus:

$$\hat{\beta} = (X'X)^{-1} X'Y \tag{49}$$

is the OLS estimate of vector β.

2.3.3. Criteria for parameters. The process of establishing the regression equation is the process of estimating the parameters in the regression model. At the same time, the advantages and disadvantages of multiple linear regression modeling have some statistical criteria that facilitate the testing of the validity of the model. These criteria are four as follows:

1) Determine the coefficient of R^2: i.e., the ratio of the regression squared to the sum of the squares of the total deviations. It can be used to quantitatively evaluate the ratio of the total variation in Y that can be explained by a linear regression equation built from the n X variables. Valid models generally require $R^2 \geq 0.75$.

2) Residual standard deviation S: Can be used to quantitatively evaluate the estimation error generated by the regression equation. Valid models generally require $S \leq 0.05$.

3) Test for covariance: assesses the independence of the samples. Valid models generally require that the degree of covariance between samples is less than 10.

4) Extreme Sample Test: Calculates the residual RStudent for each sample studentized, i.e., the standardized error for that sample obtained by removing one sample. This test is used to assess the presence of extreme samples in the sample. Valid models generally require $|RStudent| \leq 2.5$.

3. Results and analysis of university language education and students' ability to express themselves

3.1. Questionnaire Reliability Test

3.3.1. Questionnaire Reliability Tests. Cronbach's Alpha coefficient is a common method for Likert scale reliability analysis, the higher the reliability of the measured scale, the more stable the questionnaire is. Therefore, in order to understand the stability of this questionnaire, SPSS was used: the stability of the questionnaire was analyzed using Cronbach's Alph coefficient. The results of the reliability test of the questionnaire are shown in Table 2, and the Cronbach's Alph test was conducted on the overall and dimensions of the main comprehension skills of language expression skills, and it was determined that the questionnaire's Cronbach's Alph coefficient was 0.978, and the dimensions' Cronbach 's Alph coefficient is greater than 0.85, indicating that the questionnaire possesses good internal consistency in both the overall and main dimensions and its reliability is good.

Table 2. The questionnaire Cronbachs alpha test

Variable name	Cronbachs alpha
The influence factors of the ability of mathematical language expression	0.978
Policy requirement	0.922
Knowledge foundation	0.875
Linguistic ability	0.944
Nonintellectual factor	0.891
Completion of math assignments	0.855

3.3.2. Validity test of the questionnaire. Validity can test the correctness or reliability of the results of the questionnaire test, by testing the structural validity of the questionnaire to understand the reliability of the data measured by the questionnaire. SPSS was used for validity analysis and the results of KMO factor analysis are shown in Table 3, the KMO value of the questionnaire data is 0.9753, indicating that it is suitable for exploratory factor analysis.

Table 3. KMO factor analysis

Kmo and bartlett tests		Questionnaire
The cayser-meyer-olkin metric of the sampling is sufficient		0.9753
Approximate card		1843.215
Bartlett's spherical test	df	527
	Sig.	0.000

The exploratory factor analysis of the study, based on the eigenvalues greater than 1 way to extract the factors, to get the total variance of the explanation is shown in Table 4. The model obtained from the construction of this paper can reach 82.911% cumulatively, which is greater than the standard of 80%, and the current questionnaire dimensions can reflect 82.911% of the information of the original questionnaire.

Table 4. Total variance interpretation

Constituent	Initial eigenvalue			Extracting the load of the load			Rotational load squared		
	Total	Variance %	Cumulative %	Total	Variance %	Cumulative %	Total	Variance %	Cumulative %
4	10.302	24.754	24.754	10.302	24.754	24.754	5.302	10.084	10.084

6	6.021	11.821	36.575	6.035	11.821	36.575	5.021	10.725	20.809
2	5.843	10.474	47.049	5.918	10.474	47.049	4.843	9.243	30.052
5	5.343	9.277	56.326	5.313	9.277	56.326	4.343	9.79	39.842
19	2.153	5.786	62.112	2.16	5.786	62.112	3.153	8.581	48.423
13	1.931	5.282	67.394	1.925	5.282	67.394	3.931	8.841	57.264
1	1.533	4.241	71.635	1.566	4.241	71.635	3.533	7.753	65.017
9	1.485	3.935	75.57	1.468	3.935	75.57	2.485	6.527	71.544
11	1.299	3.92	79.49	1.31	3.92	79.49	2.299	6.298	77.842
3	1.251	3.421	82.911	1.251	3.421	82.911	2.251	5.069	82.911
20	1.201	2.841	85.752						
16	1.097	2.612	88.364						
12	0.986	2.553	90.917						
17	0.926	2.206	93.123						
8	0.756	1.818	94.941						
14	0.752	1.561	96.502						
18	0.622	1.433	97.935						
10	0.605	1.228	99.163						
7	0.523	0.745	99.908						
15	0.511	0.092	100						

The gravel plot is shown in Figure 1. Finally, based on the criterion that the eigenvalue is greater than 1, 4 factors are determined. It can be seen that the first four factors, “language organization, communication, language use and intonation”, show a clear steep slope, and from the fifth factor onwards, the steep slope flattens out, indicating that the first four factors can already explain 56.326% of the total variance. The principal component analysis was used to extract the common factors, the maximum variance method was chosen for the rotation, and the empirical value greater than 0.35 was used for the screening loadings to explore the addition and deletion of items to obtain the variances. The question items were deleted based on the specified principles to enhance the construct validity of the questionnaire. Observing the common factor loadings of the items, it was found that the common factor loadings of each item were greater than 0.45 in at least one dimension, so all items were considered to be retained.

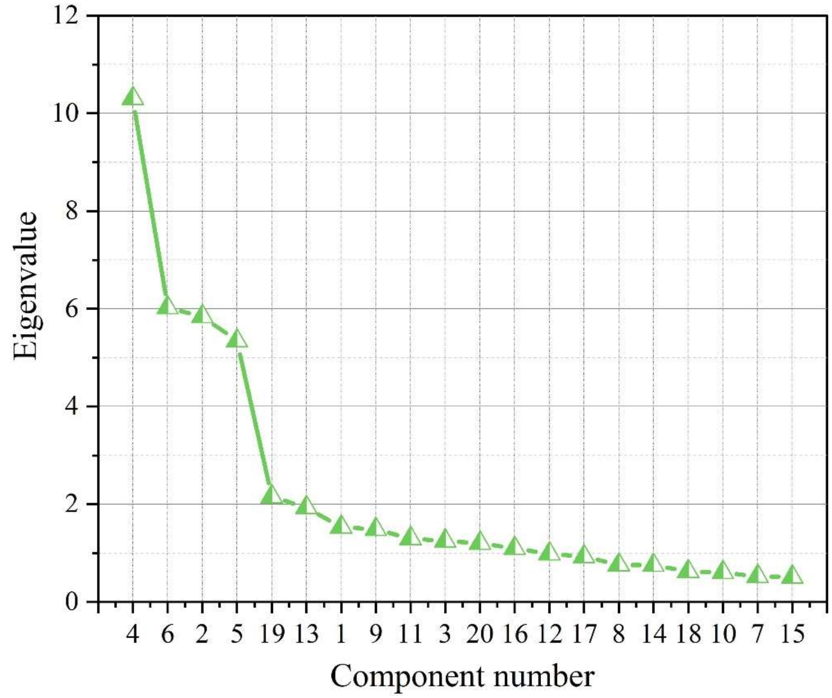

Fig. 1. Rubble map

3.2. Analysis of the results of the students' expression ability test

3.2.1. Overall situation. The test paper consisted of 10 questions with a total score of 36 points. Higher scores obtained by students indicate higher language expression ability. In order to explore the scores of university students in a more detailed and intuitive way, the descriptive statistics of the overall scores and the frequency distribution of the scores of university students were presented. The results of the descriptive statistics of students' expressive ability are shown in Table 5. The results show that the mean score of students' language expression ability is 27.43 with a standard deviation of 3.74, which indicates that the overall students' expression ability is at a normal level. The extreme difference of the test scores between students is 27, which indicates that there is a significant difference in the level of competence between students. The data in the frequency distribution graph of scores can be further verified by the fact that the scores between students cover a wide range. Only one person received the highest score of 35 and no student received scores of 35 and 36. The results of the distribution of students' expressive ability scores are shown in Figure 2. It can be seen that the overall level of expression ability of university students is not high, and most students are in the middle tier, and their ability needs to be strengthened. Accordingly, it can be inferred that the overall data of students' expression ability shows normal distribution.

Table 5. Descriptive statistical results of student expression ability

Content	Valid case number	Average score	Standard deviation	Maximum value	Minimum value	Degree of bias	kurtosis
Student expression	543	27.43	3.74	35	8	-0.238	-0.795

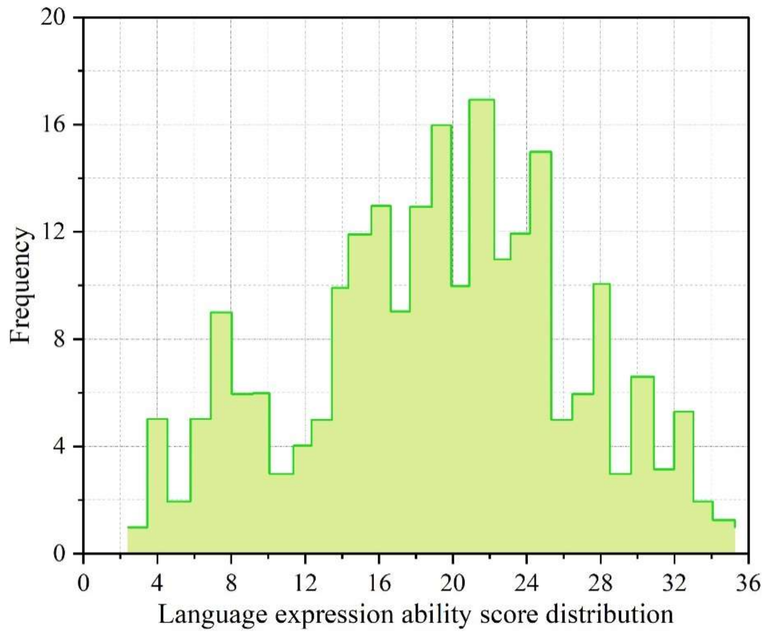

Fig. 2. Language expression ability score distribution

For the subsequent analysis of the study, it is still necessary to explore whether the data of students' scores satisfy the normal distribution and test the normality of their data. The normal Q-Q plot of students' language expression skills is shown in Figure 3. In the descriptive statistics, it shows that the skewness of the overall students' scores is -0.193 and the skewness is -0.562.To measure whether the data meets the approximate normal distribution, it can be judged from whether the skewness is satisfied in the interval [-3,+3] and the kurtosis is satisfied in the interval [-10,+10]. Also, in combination, the roughly linear distribution of the graph formed by the observations. Accordingly, it can be inferred that the overall data of students' expressive ability is close to normal distribution.

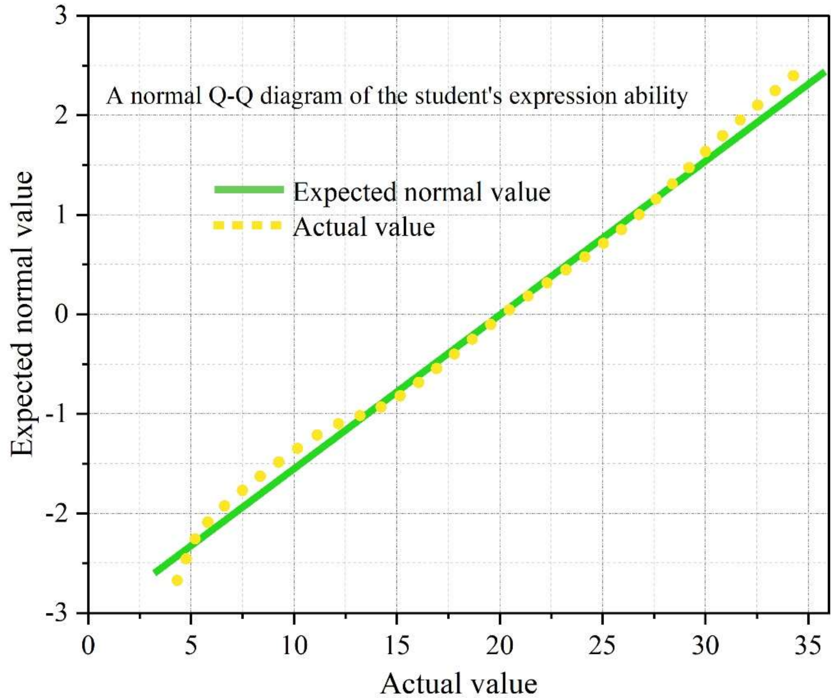

Fig. 3. The normal Q-Q diagram of the expression of the student's language

3.2.2. Results of the students' language expression test. In order to seek the expression of college students in different language and language types, this paper takes the form of descriptive statistics to analyze the performance of all college students in language and language expression, the total score of each sub-competency is 36 points, and then based on the scoring criteria for each sub-competency the students' scores on each sub-competency will be further segmented statistics, the results of the students' scores of expressive competence and language education are shown in Table 6. In the context of university language education, the highest number of students' expressive ability is in the score band of 16-22 (47.56%), and the lowest number of students is in the score band of 31-36 (10.68%). This shows that university students are more difficult to use written language to express their language knowledge clearly, and the accuracy and effectiveness of their written expression is related to the level of university language education. The scoring dimension of the university language education questionnaire shows that the lowest percentage of students is in the score band of 31-36, which is 13.26%, but it is higher than the percentage in the students' ability to express themselves. In addition, in students' expressive ability and university language education, the scoring results of the questionnaire showed that most of the students' scores were between 16-30, and their distribution accounted for 71.98% and 71.58% of the total number of students. In addition, this paper found that there is a certain correlation between the expression ability of college students and the level of university language education. In real life, schools with poor and very high educational levels are in the minority, and the vast majority of students' educational level is in the middle level. This is reflected in this paper by the fact that the proportion of students with scores of 0-15 and 31-36 is correspondingly small, while the number of students with scores of 16-22 and 23-30 is larger and corresponds to a similar level of education. This shows the significant relevance of the findings of this paper.

Table 6. Student expression ability and Chinese education score results

Fractional segment	Student expression		College Chinese education	
	Frequency	Percentage(%)	Frequency	Percentage(%)
0-15	94	17.34	82	15.16
16-22	258	47.56	142	26.24
23-30	133	24.42	246	45.34
31-36	58	10.68	72	13.26

3.3. Regression analysis of university language education and students' expressive ability

3.3.1. Relevance of university language education to students' expressive skills. The correlation between students' expressive ability and university language education is shown in Table 7. As can be seen from the table, the correlation coefficients of the four sub-dimensions of university language education and students' expressive ability are 0.589, 0.615, 0.497, 0.453 respectively, and the sig values are less than 0.01, which means that there is a significant positive correlation between university language education and students' expressive ability level. Especially, the correlation coefficient between mastering literary appreciation and students' expression ability in university language education is the highest, amounting to 0.618, which means that there is a close connection between mastering literary appreciation and students' expression ability. The reason for this is that the mastery and application of literary appreciation in the process of language learning can promote students' ability to understand and think in order to increase their logical expression ability. Teaching students' understanding of "word comprehension, sentence comprehension, text comprehension and literary appreciation" in university Chinese education can help students enhance their "writing ability, language ability, logical thinking ability and emotional attitude", and help students better express and understand in learning, which shows that there is a mutually reinforcing relationship between university Chinese education and students' expression ability.

Table 7. The correlation between student expression ability and Chinese education

Correlation analysis		Word comprehension	Sentence comprehension	Textual comprehension	Literary appreciation
Written ability	Pearson correlation	0.589**	0.579**	0.403**	0.618**
	Sig.2	0.000	0.000	0.000	0.000
Verbal ability	Pearson correlation	0.615**	0.509**	0.308*	0.609**
	Sig.2	0.000	0.000	0.06	0.000
The logical thinking ability of expression	Pearson correlation	0.497**	0.426**	0.314*	0.513**
	Sig.2	0.000	0.000	0.03	0.000
Emotional attitude	Pearson correlation	0.453**	0.418**	0.195*	0.519**
	Sig.2	0.000	0.000	0.015	0.000

3.3.2. Regression analysis of students' expressive ability and university language education. Above, through the correlation analysis of students' expression ability and university language education, it can be seen that the correlation coefficient between students' expression ability and university language education is 0.8947, which has a significant positive correlation.

In order to further explore the relationship between students' expression ability and university language education, one-way linear regression analysis was conducted with the total score of students' expression ability as the independent variable and the total score of university language education as the dependent variable, and the regression results of students' expression ability were obtained as shown in Table 8. a Predictor variable: (constant), total score of students' expression ability; b Dependent variable: total score of university language education. From the data in the table, it can be seen that the square of R is 0.884, and the value of the square of R is taken to be greater than 0.85, which indicates that the model is well fitted.

Table 8. Regression results of student expression ability

Model	R	R2	Adjusted R2	Standard error	Texbin Watson
	0.628a	0.884	0.883	12.695	1.962

Next, the regression model was tested. a Dependent variable: total college language education scores; b Predictor variable: (constant),total students' expressive skills. The results of the coefficients of the regression model of students' expressive ability are shown in Table 9 below. As can be seen from the table, F is equal to 131.073 and P=0.000<0.05, which indicates that the establishment of such regression model is statistically significant and the regression of the two groups of data is significant, so there is indeed a linear relationship between students' expressive ability and university language education.

Table 9. Regression model coefficient of student expression ability

Regression		Sum of squares	Mean square	F	Significance
Model	Regression	16480.219	16480.219	131.073	0.000b
	Residual error	21489.225	147.038		
	Total	37954.005			

The coefficients of the regression model for the dependent variables: total score of university language education, and students' expression ability are shown in Table 10. It is obtained that the regression model coefficient of students' expressive ability on university language education is 0.893 and t-test P=0.000<0.5, so students' expressive ability can have a significant effect on university language education. So the regression equation of students' expressive ability and university language education is Y=0.893X-15.874 (Y is students' expressive ability score and X is university language education score).

Table 10. Regression model coefficient of student expression ability

Model	Unnormalized coefficient		Normalization factor	t	Sig.
Constants	B	Standard error	Beta		
Total language	-15.874	6.056	0.737	-1.628	0.101
	0.893	0.075		11.122	0.000

Next, the residuals were tested for normality in P-P plots. The P-P plot of the residuals of the univariate linear regression is shown in Figure 4. It can be clearly seen that the residual analysis plot shows a normal distribution, which strongly proves that the regression equation is valid, this equation performs well in data fitting, and its predicted values are reliable and accurate.

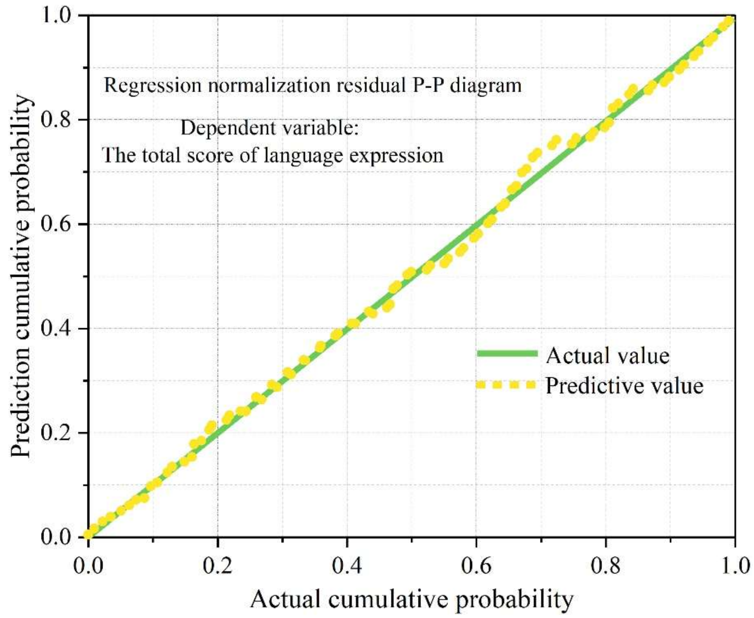

Fig. 4. One-dimensional linear regression residual P-P diagram

4. Conclusion

This paper takes the relationship between university language education and students' expressive ability as the research object, obtains the factors affecting students' expressive ability through the questionnaire survey, and then uses the principal component analysis to screen out the main factors, and then analyzes the correlation between students' language expressive ability and university language education with the help of multiple linear regression model under the university language education. The results show that:

The Cronbach's Alph coefficient of the questionnaire designed in this paper is 0.978, and the KMO value is 0.9753, indicating that the questionnaire has good reliability and is suitable for exploratory factor analysis. The total cumulative variance explained by the model in this paper reaches 82.911%, and the four main factors of "language organization ability, communication ability, language use ability and intonation ability" are identified by principal component analysis, which can explain 56.326% of the total variance.

The mean score of students' expression ability is 27.43, and the overall data shows a normal distribution, which indicates that the overall students' expression ability is at a normal level. However, the extreme difference in test scores between students is 27, which shows that there is a significant difference in the level of expressive ability between different students. In the dimension of textual language expression ability, students accounted for the largest number of students in the 16-22 score range and the smallest number of students in the 31-36 score range, indicating that college students are more difficult to use textual language to express their language knowledge clearly, and that the accuracy and effectiveness of their textual expression is directly related to the level of university language education.

References

[1] Schober, P., & Vetter, T. R. (2021). Linear regression in medical research. Anesthesia & Analgesia, 132(1), 108-109.

[2] Shastry, A., Sanjay, H. A., & Bhanusree, E. (2017). Prediction of crop yield using regression techniques. International Journal of Soft Computing, 12(2), 96-102.

[3] Stanley, T. D., Doucouliagos, H., & Steel, P. (2018). Does ICT generate economic growth? A meta-regression analysis. Journal of economic surveys, 32(3), 705-726.

[4] Yamaguchi, A., Arai, K., Aisnada, A. N. E., Lee, J. E., Kitadai, N., Nakamura, R., & Miyauchi, M. (2022). Multi-regression analysis of CO2 electroreduction activities on metal sulfides. The Journal of Physical Chemistry C, 126(5), 2772-2779.

[5] Shabbir, M. S., & Wisdom, O. (2020). The relationship between corporate social responsibility, environmental investments and financial performance: evidence from manufacturing companies. Environmental Science and Pollution Research, 27(32), 39946-39957.

[6] Wen, J., Wei, X., He, T., & Zhang, S. (2020). Regression Analysis on the Influencing Factors of the Acceptance of Online Education Platform among College Students. Ingénierie des Systèmes d'Information, 25(5).

[7] Shavit, Y., Edelman, B., & Axelrod, B. (2020, November). Causal strategic linear regression. In International Conference on Machine Learning (pp. 8676-8686). PMLR.

[8] Baždarić, K., Šverko, D., Salarić, I., Martinović, A., & Lucijanić, M. (2021). The ABC of linear regression analysis: What every author and editor should know. European science editing, 47.

[9] Lumley, T., & Scott, A. (2017). Fitting regression models to survey data. Statistical Science, 265-278.

[10] Liu, Y., & Zhang, S. (2017). Fast quantum algorithms for least squares regression and statistic leverage scores. Theoretical Computer Science, 657, 38-47.

[11] Gong, Y., Lyu, B., & Gao, X. (2018). Research on teaching Chinese as a second or foreign language in and outside mainland China: A bibliometric analysis. The Asia-Pacific Education Researcher, 27(4), 277-289.

[12] Xie, Y., Ryder, L., & Chen, Y. (2019). Using interactive virtual reality tools in an advanced Chinese language class: A case study. TechTrends, 63, 251-259.

[13] Xu, Q., & Peng, H. (2017). Investigating mobile-assisted oral feedback in teaching Chinese as a second language. Computer Assisted Language Learning, 30(3-4), 173-182.

[14] Jikui Wang,Cuihong Zhang,Wei Zhao,Xueyan Huang & Feiping Nie. (2025). Fast anchor graph optimized projections with principal component analysis and entropy regularization. Information Sciences121797-121797.

[15] Zhao Qi. (2022). Modeling and Analysis Method of National Fitness Big Data for Basketball Projects Based on a Multivariate Statistical Model. Security and Communication Networks

[16] Wenyan Bian,Wenzheng Dong,Qiang Zheng,Qian Gu,Shihui Bian & Yang Yang. (2024). Fast weighted least squares for detail and tone enhancement of medical images. DIGITAL HEALTH

[17] Hao Li,Xianchao Dai,Ligang Zhou,Qun Wu,Muhammet Deveci & Dragan Pamucar. (2024). A least-squares framework for developing interval type-2 fuzzy semantics. Applied Soft Computing(PA),112293-112293.