DOI: 10.23977/tracam.2023.030103
ISSN 2616-1826 Vol. 3 Num. 1

Mathematical Analysis of the Relationship between College Entrance Exam Scores and Information and Computing Science Discipline Performance in a Chinese University

Bing Hao[*], Huan Luo, Fubin Chen

Oxbridge College, Kunming University of Science and Technology, Kunming, 650106, China
[*]Corresponding author

Keywords: Multiple regression analysis, canonical correlation analysis, correlation coefficient, college entrance examination score, discipline score

Abstract: This article uses multiple regression analysis methods and canonical correlation analysis, combined with the college entrance examination scores and partial discipline scores in the first three years of university for students majoring in Information and Computing Science in a certain university. We separately analyze the connection between the college entrance examination and university discipline scores, and compare the advantages and disadvantages of the two methods. We hope to quantitatively analyze the transmission and extension of subject knowledge at different stages from a mathematical perspective, and also promote the practical application value of mathematics, providing more people with inspiration and thought on mathematics.

1. Introduction

Mathematics is the foundation of the entire social development, it is the science of "quantity", and a science that comes from and guides the solution of practical problems [1]. Combining the discipline scores for the first three years of college and the entrance examination scores data for students majoring in Information and Computing Science at a certain private university, we use multiple regression analysis and canonical correlation analysis to analyze the relationship between the two scores [2]. We aim to expand the application capabilities of mathematics and hope to bring more inspiration and thought.

2. Multiple Regression Analysis and Canonical Correlation Analysis Theory and Data Description

2.1. Brief Introduction to Multiple Regression Analysis Theory

Assume $X=(x_1,x_2,......,x_n)^T$ is an n-dimensional random variable, and $Y=(y_1,y_2,......,y_m)$ is a m-dimensional random variable. There exists a correlation between X and Y, and it is assumed that there is no interaction between them [1,3].

Then the linear regression model between X and Y is $y_i = \beta_0 + \beta_{i1}x_1 + \beta_{i2}x_2 + ... + \beta_{in}x_n + \varepsilon$, where β_0 is the constant term, $\beta_{i1}, \beta_{i2}, ..., \beta_{in}$ is the regression coefficient, and ε is a random variable. It is also assumed that ε follows a normal distribution with an expectation of 0, and ε is called a random error. Through the linear regression equation, we can clearly observe the linear influence relationship between X and Y.

2.2. Brief Introduction to Canonical Correlation Analysis Theory

Canonical correlation analysis [4], as an important part of multivariate statistics, is a major content of correlation analysis research. The concept is developed based on the correlation of two variables, simplifying the complex correlation relationship between two sets of variables and reflecting the information between the two sets of variables with a few pairs of correlations, while ensuring that these pairs of variables are not related to each other.

Generally speaking, suppose X and Y are p-dimensional and q-dimensional random variables, respectively. At this time, there are p*q correlation coefficients, which are relatively complicated to analyze, and the analysis is difficult to grasp the essence of the matter due to the complex relationships between the components. In order to study the relationship between X and Y, we just need to find a linear combination of the components of X, and at the same time find a linear combination of the components of Y, and make the variables represented by these two linear combinations have the maximum correlation. We call this correlation canonical correlation, and call this pair of new random variables canonical variables; then, we can find a second pair of linear combinations from X and Y, which are uncorrelated with the first pair of linear combinations, and this pair of linear combinations has the maximum correlation, so we get the second pair of canonical variables. We continue this process until the correlation of X and Y is basically extracted, and then analyze the correlation of X and Y based on the canonical variables. This statistical method is called canonical correlation analysis.

2.3. Data Selection Description

We selected the college entrance examination subject scores of 23 students in the 2008 class of the Information and Computing Science major at a certain university (both the college entrance examination scores and university subject scores were provided by the academic affairs office of the school). The average score of the college entrance examination subjects is noted as $X = (x_1, x_2, x_3, x_4, x_5)$, where x_1, x_2 , x_3, x_4, x_5 represent the scores of Chinese, Mathematics, English, Physics, and Chemistry subjects respectively; the average score of the university subjects is noted as $Y = (y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9, y_{10})$, where y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, y_9, y_{10} represent the scores of subjects such as Advanced Mathematics, Mathematical Analysis, College English, Discrete Mathematics, Differential Equations, C++ Programming Design, Numerical Analysis, Computer Graphics, Optimization Methods, and Mathematical Modeling, respectively.

Based on these simple sample data, we hope to explore some connection between the college entrance examination and university professional discipline scores through the above two methods [5], and also provide a simple comparison and explanation of the two methods.

3. Specific Calculation and Analysis Based on Multiple Regression Analysis Method

3.1. The Calculation Based on Multiple Regression Analysis

According to the linear regression model between X and Y [6-9], which is $y_i = \beta_0 + \beta_{i1}x_1 + \beta_{i2}x_2 + ... + \beta_{in}x_n + \varepsilon$, we use $SPSS$ software to calculate and obtain the following results for four specific university disciplines (taking Advanced Algebra, C++, College English, and Mathematical Modeling as examples):

(1)The regression equation and analysis between Advanced Algebra and the scores of the five subjects in the college entrance examination.

The calculated regression equation is:

$$y_1 = 43.418 - 0.021x_1 + 0.291x_2 - 0.022x_3 + 0.457x_4 - 0.251x_5$$

Simultaneously, the determination coefficient $R^2 = 0.388$, indicating that the linear regression relationship between y_1 and x_1, x_2, x_3, x_4, x_5 is generally significant. This suggests that the linear relationship between Advanced Algebra and the scores of the five subjects in the college entrance examination is generally significant. Further, from the regression equation, we can see that Mathematics and Physics in the college entrance examination subjects have a greater impact on the university variable, namely Advanced Algebra.

(2)The regression equation and analysis between C++ and the scores of the five subjects in the college entrance examination.

The calculated regression equation is:

$$y_6 = 70.069 - 0.188x_1 + 0.060x_2 + 0.054x_3 + 0.058x_4 + 0.186x_5$$

Simultaneously, the determination coefficient $R^2 = 0.253$, indicating that the linear regression relationship between y_6 and x_1, x_2, x_3, x_4, x_5 is not very good, suggesting that the linear relationship between C++ and the scores of the five subjects in the college entrance examination is not very clear. From the regression equation, we can see that Mathematics, Chemistry, and Physics in the college entrance examination subjects have a relatively weak impact on the study of C++, which also indicates that students in this major have certain difficulties in learning C++ course.

(3)The regression equation and analysis between College English and the scores of the five subjects in the college entrance examination.

The calculated regression equation is:

$$y_3 = 38.593 - 0.022x_1 + 0.242x_2 + 0.133x_3 - 0.022x_4$$

Simultaneously, the determination coefficient $R^2 = 0.175$, indicating that the linear regression relationship between y_3 and x_1, x_2, x_3, x_4, x_5 is relatively weak, suggesting that the linear relationship between College English and the scores of the five subjects in the college entrance examination is poor. This also shows that the teaching outline and requirements of College English are no longer suitable for exam-oriented education, and there is a significant gap from high school English teaching. Relatively speaking, from the regression equation, we can see that the English subject in the college entrance examination has a greater impact on College English.

(4)The regression equation and analysis between Mathematical Modeling and the scores of the five subjects in the college entrance examination.

The calculated regression equation is:

$$y_{10} = 97.819 - 0.141x_1 + 0.064x_2 - 0.056x_3 + 0.058x_4 - 0.058x_5$$

Simultaneously, the determination coefficient $R^2 = 0.283$, indicating that the linear regression relationship between y_{10} and x_1, x_2, x_3, x_4, x_5 has improved compared to the previous equation. From the regression equation, we can see that Mathematics and Physics in the college entrance examination subjects have a greater impact on Mathematical Modeling. This also reflects that the discipline of Mathematical Modeling pays more attention to dealing with practical problems and applying theory to the solution of practical problems. This process is often not easy for students to accept and also reflects certain drawbacks of exam-oriented education.

3.2. Conclusion Analysis

Multiple regression analysis only studies the correlation between the scores of specific individual subjects and the scores of each subject in the college entrance examination [10]. Furthermore, the regression coefficient of the equation is relatively small, and the degree of fit is not ideal. This approach cannot fully demonstrate the connection between the subjects of the college entrance examination and the subjects of the information and computing science major. Thus, this method still lacks persuasive power to evaluate the relationship between the college entrance examination and university major subjects.

4. Specific Calculation and Analysis Based on Canonical Correlation Method

4.1 Correlation Analysis Calculation of College Entrance Examination Subjects and University Subjects Scores

The following information is obtained by processing data through Matlab:

Correlation coefficient matrix(Tan et al., 2020) between the scores of college entrance examination subjects:

$$A = \begin{pmatrix} 1 & 0.004 & 0.264 & -0.013 & -0.140 \\ 0.004 & 1 & 0.409 & 0.424 & 0.488 \\ 0.264 & 0.409 & 1 & 0.072 & 0.089 \\ -0.013 & 0.424 & 0.072 & 1 & 0.622 \\ -0.140 & 0.488 & 0.089 & 0.622 & 1 \end{pmatrix}$$

Correlation coefficient matrix between the scores of university subjects:

$$B = \begin{pmatrix} 1 & 0.604 & 0.340 & 0.517 & 0.414 & 0.432 & 0.570 & 0.184 & 0.294 & 0.186 \\ 0.604 & 1 & 0.315 & 0.541 & 0.471 & 0.519 & 0.534 & 0.395 & 0.192 & 0.417 \\ 0.340 & 0.315 & 1 & 0.101 & 0.591 & 0.486 & 0.491 & 0.115 & 0.440 & 0.139 \\ 0.517 & 0.541 & 0.101 & 1 & 0.496 & 0.613 & 0.542 & 0.530 & 0.476 & 0.288 \\ 0.414 & 0.471 & 0.591 & 0.496 & 1 & 0.599 & 0.687 & 0.416 & 0.571 & 0.070 \\ 0.432 & 0.519 & 0.486 & 0.613 & 0.599 & 1 & 0.621 & 0.445 & 0.297 & 0.192 \\ 0.570 & 0.534 & 0.491 & 0.542 & 0.687 & 0.621 & 1 & 0.517 & 0.396 & 0.229 \\ 0.184 & 0.395 & 0.115 & 0.530 & 0.416 & 0.445 & 0.517 & 1 & 0.409 & 0.542 \\ 0.294 & 0.192 & 0.440 & 0.476 & 0.571 & 0.297 & 0.386 & 0.409 & 1 & 0.001 \\ 0.186 & 0.417 & 0.139 & 0.288 & 0.070 & 0.192 & 0.229 & 0.542 & 0.001 & 1 \end{pmatrix}$$

Correlation matrix between the scores of college entrance examination subjects and university subjects:

$$C=\begin{pmatrix} 0.009 & -0.218 & 0.063 & -0.369 & -0.167 & -0.215 & -0.402 & -0.420 & 0.088 & -0.402 \\ 0.315 & 0.312 & 0.302 & 0.256 & 0.050 & 0.306 & 0.274 & -0.101 & -0.100 & -0.017 \\ 0.051 & 0.137 & 0.385 & -0.030 & 0.220 & 0.145 & 0.087 & -0.405 & 0.071 & -0.369 \\ 0.562 & 0.729 & 0.067 & 0.410 & 0.305 & 0.332 & 0.253 & 0.137 & 0.009 & 0.132 \\ 0.185 & 0.503 & 0.100 & 0.491 & 0.177 & 0.441 & 0.211 & 0.310 & 0.066 & 0.012 \end{pmatrix}$$

The canonical correlation matrix between the scores of the college entrance examination subjects and the scores of university subjects is calculated [4,11]:

$$D=B^{-1}C^{T}A^{-1}C=\begin{pmatrix} 0.70 & 0.77 & 0.62 & 0.47 & 0.43 & 0.44 & 0.06 & 0.12 & 0.96 & -0.62 \\ 0.51 & 0.01 & 0.66 & 0.14 & 0.37 & 0.21 & -0.04 & 0.35 & 0.23 & -0.42 \\ 0.40 & 0.64 & 0.59 & 0.30 & 0.96 & 0.43 & -0.05 & 0.14 & 0.63 & -0.35 \\ 0.56 & 0.73 & 0.58 & 0.45 & 0.23 & 0.42 & -0.01 & 0.11 & 0.85 & -0.57 \\ -0.36 & -0.63 & -0.53 & -0.34 & 0.78 & -0.42 & 0.08 & -0.10 & -0.65 & 0.45 \\ 0.74 & 0.27 & 0.92 & 0.15 & 0.86 & 0.39 & -0.01 & 0.42 & 0.54 & -0.54 \\ -0.52 & -0.80 & -0.68 & -0.43 & -0.26 & -0.51 & 0.07 & -0.15 & 0.84 & 0.51 \\ 0.98 & 0.22 & 0.01 & 0.77 & 0.19 & 0.71 & 0.02 & 0.22 & 0.40 & -0.87 \\ 0.79 & 0.87 & 0.68 & 0.57 & 0.63 & 0.48 & 0.07 & 0.13 & 0.10 & -0.75 \\ -0.69 & -0.27 & -0.83 & -0.81 & -0.65 & -0.86 & -0.16 & -0.44 & -0.61 & 0.04 \end{pmatrix}$$

4.2. Canonical Variable Correlation Analysis Calculation of College Entrance Examination Subjects and University Subjects Scores

The eigenvalues of the canonical correlation matrix D are calculated to be :

$\lambda_1^2=0.48,\lambda_2^2=0.28,\lambda_3^2=0.057,\lambda_4^2=0.013,......$, and according to Bartlett's test method, the first three eigenvalues are significantly larger than the later eigenvalues. Therefore, we only take the first three eigenvalues and analyze the three pairs of canonical variables.

Using the eigenvalues $\lambda_1=0.69,\lambda_2=0.53,\lambda_3=0.24$ of matrix D and its normalized eigenvectors $v_i\left(i=1,2,3\right)$, we obtain the coefficients of the canonical variables of the university subjects:

$$W_i=\begin{pmatrix} 0.09 & 0.53 & 0.20 & -0.37 & -0.32 & 0.35 & -0.14 & 0.56 & -0.44 & 0.50 \\ 0.34 & -0.18 & -0.26 & 0.06 & 0.17 & -0.33 & -0.13 & -0.26 & 0.78 & -0.09 \\ 0.06 & -0.27 & -0.12 & 0.58 & 0.21 & -0.08 & -0.12 & -0.05 & 0.11 & -0.16 \end{pmatrix}^{T}$$

Similarly, we can obtain the coefficients of the first three canonical variables of the college entrance examination subjects:

$$u_i=\begin{pmatrix} -0.05 & 0.64 & -0.25 & 0.33 & 0.25 \\ 0.67 & 0.10 & 0.38 & -0.21 & -0.25 \\ -0.05 & 0.21 & 0.11 & 0.51 & 0.48 \end{pmatrix}^{T}$$

4.3. Canonical Variable Results Analysis of College Entrance Examination Subjects and University Subjects Scores

(1)The analysis of the first pair of canonical variables is as follows:
University subject canonical variable equation:

$$w_1 = 0.09y_1 + 0.53y_2 + 0.2y_3 - 0.37y_4 - 0.32y_5 + 0.35y_6 - 0.14y_7 + 0.56y_8 - 0.44y_9 + 0.5y_{10}$$

College entrance examination subject canonical variable equation:

$$z_1 = -0.05x_1 + 0.64x_2 - 0.25x_3 + 0.33x_4 + 0.25x_5$$

The canonical correlation coefficient between w_1 and z_1 is 0.69, indicating a very close relationship between them. Through the coefficient analysis of each equation, we find that computer graphics, mathematical analysis, C++ programming design, and mathematical modeling occupy a larger proportion in the university subject canonical equation. Combining the characteristics of university subjects, we find that: the subject of computer graphics is a combination of mathematics and graphics, and it is easier for students to understand than the relatively dry mathematical analysis; compared with high school mathematics, mathematical analysis is the foundation of university mathematics, and its importance is self-evident; C++ programming is based on mathematics, and mathematical modeling reflects the comprehensive application ability of the subject major; the analysis also reflects the characteristics of this major: based on mathematics, comprehensively and proficiently use mathematical ability to solve practical problems, and improve the application ability of mathematics. On the contrary, the main factors in our college entrance examination subject variable equation are: mathematics, physics, chemistry; among which mathematics occupies the largest proportion, reaching 0.64, the good or bad of mathematics scores can directly affect the learning of related subjects in university; at the same time, the knowledge of elementary physics and chemistry also plays a good understanding and help role in other university subjects like mathematical modeling.

(2)The analysis of the second pair of canonical variables is as follows:
Canonical variable equation of university subject:

$$w_2 = 0.34y_1 - 0.18y_2 - 0.26y_3 + 0.06y_4 + 0.17y_5 - 0.33y_6 - 0.13y_7 - 0.26y_8 + 0.78y_9 - 0.09y_{10}$$

Canonical variable equation of college entrance examination subject:

$$z_2 = 0.67x_1 + 0.10x_2 + 0.38x_3 - 0.21x_4 - 0.25x_5$$

The canonical correlation coefficient between w_2 and z_2 is 0.53, which is slightly smaller than the previous coefficient. Continue to analyze the coefficients of each equation, we find that the optimal method in the university subject occupies the largest proportion in the university subject canonical variable equation. Combined with the characteristics of this subject, we know: the main research object of the optimization method is the management problem of various organized systems and their production and operation activities, mainly using mathematical methods to study the optimal path and plan of various systems, providing the basis for scientific decision-making for decision-makers; at the same time, good Chinese scores indicate that students have a deep understanding of literal sentences, with the aid of a certain mathematical foundation, can understand the mathematical background described in the optimal method subject, and then can choose the appropriate method for processing and decision-making.

(3)The analysis of the third pair of canonical variables is as follows:
Canonical variable equation of university subject:

$$w_3 = 0.06y_1 - 0.27y_2 - 0.12y_3 + 0.58y_4 + 0.21y_5 - 0.08y_6 - 0.12y_7 - 0.05y_8 + 0.11y_9 - 0.16y_{10}$$

Canonical variable equation of college entrance examination subject:

$$z_3 = -0.05x_1 + 0.21x_2 + 0.11x_3 + 0.51x_4 + 0.48x_5$$

The canonical correlation coefficient between w_3 and z_3 is 0.24, which is a little smaller than the second pair. Through the coefficient analysis of each equation, we find that discrete mathematics in the university subject occupies the largest proportion in the university subject canonical variable equation. Combined with the characteristics of this subject, we know: the course of discrete mathematics mainly introduces the basic concepts, basic theories and basic methods of various branches of discrete mathematics, and these concepts, theories and methods are widely used in professional courses such as digital circuits and compiler principles; at the same time, the main factors in the college entrance examination subject canonical variable equation are physics and chemistry, good scores in physics and chemistry indicate students' mastery of the principles of physical and chemical phenomena, combined with a certain mathematical logic basis, can understand the mathematical principles described in the subject of discrete mathematics, and then can master this subject.

4.4. Analysis

Through canonical correlation analysis of students' college entrance examination scores and university subject scores in the first three years, the connection between the subjects of college entrance examination and university has been deeply analysed [4,11]. The college entrance examination has strongly promoted the in-depth study of relevant professional knowledge in universities. At the same time, according to the professional settings of the subjects of advanced algebra, mathematical analysis, numerical analysis, discrete mathematics, mathematical modeling, etc., they are greatly influenced by the subjects of mathematics, physics, and chemistry in the college entrance examination and occupy a very important position in higher education. Therefore, when reforming the high school curriculum system and implementing the 3+X college entrance examination model, enough class hours should be left for mathematics, physics, and chemistry to ensure the teaching quality and enable high school education to send more college students with solid basic knowledge to higher education institutions.

5. Conclusion

Through multivariate linear regression analysis of the scores of college entrance examination and university subjects, the coefficient of unknowns in the equation is small, and the effect of regression equation is not very ideal, indicating that this method lacks more favorable support to quantify the connection between the two. However, through the canonical correlation analysis method, establish canonical related variables, determine three pairs of canonical related variables through the size of eigenvalues, and rationally analyze and explain the transmission and extension of subject knowledge at two stages relatively well, and indirectly explain the importance of mathematical method selection. Based on the limited data in this paper, it is necessary to explore more effective mathematical methods to achieve better quantitative support, and also hope to give more people thinking about mathematics.

Acknowledgements

The relevant work of this paper has been supported and funded by the Science Research Fund

Project of Yunnan Provincial Department of Education (2022J1097).

References

[1] Wei Z.S. (1993). Course of Probability Theory and Mathematical Statistics. Higher Education Press.
[2] Wu Z.H., Feng D.M., Wei Z.C. (2008). Education Administration. (2nd Edition). East China Normal University Press.
[3] Su B., Xie Y.Q. (2006). Application of Statistical Analysis in Student Performance Evaluation. Systems Engineering Theory and Practice .7. 134-140.
[4] Cheng G. W., Chen Q. S., Li W.F. (2002). Canonical Correlation Analysis Method for Researching and Analyzing Students' On-Campus Performance. Journal of Wuhan University of Science and Technology (Social Science Edition). 4 70-74.
[5] Tan J.B., Guo Q., Li M. L. (2020). Analysis of Students' Performances Based on R Language. Journal of Changchun Normal University. 8.6-12.
[6] Gao H. X. (2002). Practical Statistical Method and SPSS System. Peking University Press.
[7] Liu H.S. (2003). Application of Multivariate Statistical Analysis in Comprehensive Evaluation of Student Performance. Journal of North China University of Science and Technology.1. 77-79.
[8] Yang Y.Y. (2021). Application Research of Multivariate Regression Analysis in the Performance Analysis of Senior Two Arts Students. College Entrance Examination. 29. 109-110.
[9] Zhong T., Gu Q.Y. (2020). Research on the Application of Multivariate Statistical Analysis in University Student Performance Evaluation. China Education Technology Equipment, 4. 110-111+114.
[10] Sun Q., Ca Z.L. (2020). Regression Analysis of Influencing Factors of College Mathematics Scores. Journal of Hubei Normal University (Natural Science Edition). 40 (03) 80-83.
[11] Tong L., Yao J. (2003). Research on Canonical Correlation Analysis Method of Relationship between Undergraduate Basic Courses and Professional Courses. Journal of Shanghai University of Science and Technology (Social Science Edition). 2. 70-73.