Analyses of Postgraduates' Entrance Examination Scores Based on Linear Regression with Dummy Variables

Ning Xiaojun,
Graduate School, University of Science and Technology Beijing, Beijing, China
Huang Ruocheng, Liang Xiaoyi, Ai Dongmei*
School of Mathematics and Physics, University of Science and Technology Beijing, Beijing , China
*__corresponding author__ :**Ai Dongmei** , *E-mail: aidongmei@ustb.edu.cn*

Abstract

Detecting the main factors influencing the students' score is an important part of the student evaluation system. Several factors that have significant influence on postgraduates' entrance examination scores including enrollment category, university category, age, gender, fresh graduate were studied by ANOVA in this paper. Quantitative analysis of the correlation between the discrete variables and postgraduates' entrance examination scores were performed by linear regression with dummy variables and 85% confidence prediction interval of postgraduates' entrance examination scores were obtained. Data support were provided by these results for graduate school enrollment work.

Keywords: Postgraduates' Score; ANOVA; Dummy Variables; Linear Regression

1. Introduction

A digital campus information management system is being promoted in universities recently. Especially, complete postgraduate information management systems are being used by graduate school in universities [1]. Thus, a large number of postgraduates' data have formed valuable information resources. However, these data are only used for simple query and statistics at present, and its inherent information was not properly mined and utilized [2].

In this paper, enrollment information of postgraduates within three years was extracted and analyzed by ANOVA and stepwise regression analysis with dummy variable, in order to explore the factors that affect the postgraduates' entrance examination scores and measure the quality of enrollment postgraduates. The attribute variables (e.g. gender, university category) were analyzed using ANOVA in order to judge if they have a significant impact on the postgraduates' entrance examination scores. Those significant attribute variables were then applied in stepwise regression and obtained regression equations. Prediction interval with 85% confidence of postgraduates' entrance examination scores were obtained and the forecasted scores for each college were compared.

2. Method

2.1 *Analysis of Variance*

The attribute variable which has a significant impact on the postgraduates' entrance examination scores were selected using one-way ANOVA. Let a factor (attribute variable) A with r levels: $A_1, A_2, \cdots A_r$, collecting data X_{ij} by n times observation under A level, where $j = 1, 2, ..., n_i$, $i = 1, 2, ..., r$. Let μ be the grand mean of all the samples, α_i is the significance level, ε_{ij} is the random error, and then comes the following ANOVA models [3]:

$$\begin{cases} X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad j = 1, ..., n_i, i = 1, ..., r \\ \varepsilon_{ij} \sim N(0, \sigma^2), \ i.i.d. \\ \sum_{i=1}^{r} n_i \alpha_i = 0 \end{cases}$$

whose null hypothesis is $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_r$. Then its test statistic can be constructed base on the Sum of Squares of Deviation: $F(r-1, n-r)$. If the null hypothesis is rejected, it shows that there are significant differences in different levels. The variance homogeneity assumption of variance analysis can be tested by Bartlette's test or Levene's test [4]

2.2 *Stepwise Regression with Dummy Variables*

The quantitative relation between each attribute variable and postgraduates' entrance examination scores can be analyzed using stepwise regression with dummy variables.

The basic idea of linear regression is as follows: let $Y = X\beta + \varepsilon$, where X is independent variables, Y is dependent variables, β is the vector of regression

3 Data analysis

3.1 *Analyses of Admission Scores by ANOVA*

The data are from the postgraduate registration and enrollment information database of a university for three years, which totally consit of 5384 records. 29 variable attributes are included, such as admission college, test method and test scores after the raw data are merged and standardized. The universities where postgraduates got bachelor degree were divided into seven major categories, include 985 project university, 211 project university with graduate school, 211 project university without graduate school, key university, ordinary university, secondary college and this university, which in total has seven attribute variables.

The relationship between the five factors of gender (with 2 attribute values of Male and Female respectively), admission category (with 4 attribute values of Directional training, Non Directional training, Self-fund Training, Training respectively), register age groups (with 3 attribute values of less than 22 years old, between the ages of 22 and 25, over 25 years olds respectively), this year's graduates (with 2 attribute values of Yes and No respectively), graduate university (with 7 attribute values shown above respectively) and the postgraduates' entrance examination scores were analyzed by ANOVA. If the postgraduates' entrance examination scores show significant differences under the different values of one attribute variable, then this variable is an important factor to influence the postgraduates' entrance examination scores.

Each factor's homogeneity of variance was tested by Bartlett test function in R, while the one-way ANOVA for postgraduates' entrance examination scores by aov function. The result showed that admission category, graduate university and register age groups are the important factors to influence the postgraduates' entrance examination scores, while gender is not a significant factor. Take admission category as an example. Its ANOVA result is shown in Table 1, where the p-value is 0.0005, which indicates that there are significant difference among distinct admission category values. The significant differences among values of admission category can be shown in Fig. 1.

Table 1 ANOVA for admission category

Source	F.D.	Sum of Sq.	p-value
Admission Category	3	11.8574	0.0005
Error	7	1.1616	

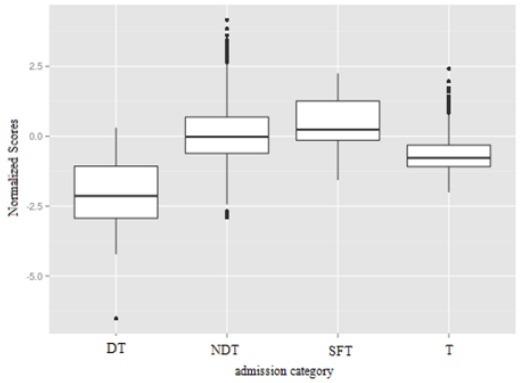

Fig. 1 normalized scores of admission category

3.2 *Analyses of Admission Scores by Regression*

The quantitative relationship between the postgraduates' entrance examination scores and the above three significant attribute variables was analyzed through regression analysis in this paper. We established an effective regression equation of independent variables with statistical significance by a stepwise regression method based on AIC information entropy. As an example of introducint dummy variables for a discrete attribute variable, let one's postgraduate admission score is Y which could take three different values, and thus

three dummy variables could be introduced for register age groups as follows:

$$C_{1i} = \begin{cases} 1 & 22-25 \text{ years old} \\ 0 & \text{otherwise} \end{cases}$$

$$C_{2i} = \begin{cases} 1 & \text{less than } 22 \text{ years old} \\ 0 & \text{otherwise} \end{cases}$$

$$C_{3i} = \begin{cases} 1 & \text{over } 25 \text{ years old} \\ 0 & \text{otherwise} \end{cases}$$

Likewise, four dummy variables of D were introduced for admission category, and seven dummy variables of E were introduced for graduate university, so the regression equation could be represented as

$$Y_i = \beta_0 + \beta_1 C_{1i} + \ldots + \beta_3 C_{3i} + \beta_4 D_{1i} + \ldots + \beta_7 D_{4i} + \beta_8 E_{1i} + \ldots + \beta_7 E_{7i} + \varepsilon_i$$

which contains 14 regression variables, and totally 15 regression coefficients to be determined. We extracted postgraduates' data of different department of graduate school, then analyzed these independent variables by a stepwise regression method based on AIC information entropy [9] respectively. The results from the School of Metallurgical and Ecological Engineering are shown in Table 2. As for postgraduate students in this School, the p-values of attribute variables except 985 project university are all less than 0.05 , which shows these attribute variables are significant factors to influence the postgraduates' entrance examination scores of postgraduates. And the obtained regression results for each school can be used to predict their students' scores, and the result are shown in Fig. 2, from which it can be seen that the School of Foreign Language, Computer and Community, Economy and Management and Humanity and Law have much higher predictions.

Table 2 Results of independent variables by a stepwise regression method

	estimated	Standard error	t-value	p-value
β_0	396.469	6.094	65.057	< 2e-16
985 project university	10.479	6.173	1.698	0.090
this university	13.116	4.571	2.869	0.004
211 university without graduate school	16.558	5.722	2.894	0.004
key university	16.681	4.927	3.386	0.001
ordinary university	11.789	4.323	2.727	0.007
Non Directional	10.946	3.353	3.265	0.001
less than 22 years old	15.903	5.663	2.808	0.005
between the ages of 22 and 25	11.683	4.055	2.881	0.004

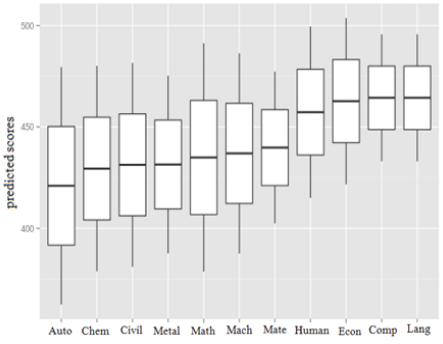

Fig. 2 Predicted scores from regression results

4. Conclusion

Enrollment data of postgraduates of one university are processed by ANOVA and stepwise regression analysis with dummy variable in this paper. A wealth of information was revealed, which can be important guide to cultivation for postgraduates.

Analyzing the factors affecting the postgraduates' entrance examination scores of postgraduates and prediction of the postgraduates' entrance examination scores of postgraduates, can provide a more scientific basis for the development of the graduate school and achievement clarifying objectives and scientific work.

Acknowledgements

This paper is supported by The Postgraduate Education and Development Grant of University of Science and Technology Beijing.

References

1. Hu Qintai, Zheng kai, Lin Nanhui (2014), Transformation: From ”Digital Campus” to “ Intelligent Campus”. China Audio Visual Education (1): 35-39.
2. Zhang Junchao (2014), Institutional Research and University Management in the Age of Big Data. Researches in Higher Education of Engineering (1): 023.
3. Chen Xiru (2009), Advanced Mathematical Statistics. University of Science and Technology China Press.
4. Marozzi, Marco (2011), Levene type tests for the ratio of two scales. Journal of Statistical Computation and Simulation 81 (7): 815-826.
5. He Xiaoqun, Liu Wenqing (2007), Applied Regression Analysis. Renmin University of China Press.
6. Fang Kaitai (1989), Practical Multivariate Statistical Analysis. East China Normal University Press.
7. Spiegelhalter, David J., et al (2014), The deviance information criterion: 12 years on. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76(3): 485-493.
8. Kleinbaum, et al (2014), Applied regression analysis and other multivariable methods. Cengage Learning.
9. Chowdry, Haroon, et al (2013), Widening participation in higher education: analysis using linked administrative data. Journal of the Royal Statistical Society: Series A (Statistics in Society), 176 (2): 431-457.