International Journal of u- and e- Service, Science and Technology
Vol.9, No. 1 (2016), pp.117-122
http://dx.doi.org/10.14257/ijunesst.2016.9.1.13

Statistical Data Analysis and Prediction Model for Learning Assessment in Korean High Schools Based on EduData

Heeyoul Choi[1*] and Yunhee Kang[2]

[1]*Samsung Advanced Institute of Technology, Suwon, Korea*
[2]*Baekseok University, Cheonan, Korea*
[1]*heeyoul@gmail.com*, [2]*yunh.kang@gmail.com*

Abstract

Most analyses in pedagogy have been based on surveys, while in many other research areas like cognitive science and psychology, data-driven research has made significant progress based on large-scale data automatically generated and archived. Recently in pedagogy, learning achievement data has been archived, and EduData is one of such data sets provided by Korean ministry of education. Many data driven analysis algorithms can be applied to such data. As a first data-driven analysis to EduData, we applied the linear regression model to check which factors are effective to Korean student's learning achievement. Finally, we proposed a model to predict degree of achievement. Experimental results show the performance of our models.

Keywords*: EduData, Statistical Data Analysis, Causal Relation*

1. Introduction

Recently, in many research areas like biology, sociology and psychology, large-scale data has been generated and archived [1, 2]. Based on such data, corresponding research has made significant progresses in developing theories and discovering new knowledge. Most significantly, in our previous work in cognitive science, data-driven analysis has found new understanding and confirmed the previous hypotheses [2]. In pedagogy, however, most analysis approaches are still based on a small set of questionnaires rather than big data analysis [3].

There are many potential causal factors to learning achievement in school, including the number of classrooms, the number of students per classroom, the number of teachers, the hours of after-school-study, the number of library books, and so on. Such factors are supposed to have effects on learning performance in different ways on different subjects. Since there are too many factors to consider, it is hard to analyze their relationships to the performance manually. So, automatic analysis based on data is essential to understand the potential causes to the learning achievement.

This paper applies a statistical data analysis technique to EduData recently provided by Korean ministry of education [4]. The data includes general information about schools and exam scores. Our method shows which factors are important and how they are important in predicting exam scores, and predicts exam scores based on the factors. These results are helpful to understand learning achievement in school and can be exploited to work out pedagogical strategy for investment more efficiently.

The rest of the paper is organized as follows: Section 2 briefly reviews regression models and section 3 describes the dataset and its preprocessing. Section 4 shows how to handle missing values in the dataset. In section 5, we propose a prediction model. In section 6, experiment results and observations are reported. Finally, conclusions follow in section 7.

* Correspondence Author

ISSN: 2005-4246 IJUNESST

2. Previous Work

To understand relationships between variables (or factors), in general, regression analysis has been applied [5]. Although there are many variants in regression analysis, the general model is summarized as follows:

Let the independent variable be ***X***, the dependent variable ***Y***, and the parameter ***W***. The general regression model is given by

$$Y = f(X, W), \quad (1)$$

where the function *f* must be specified. According to *f*, the regression model can be linear or nonlinear. One of the most popular models is linear regression, which uses a linear function *f*.

3. Data Set

3.1. EduData

EduData includes many factors from different systems like KEDI, KICE and ministry of education, each including many factors like the numbers of students, native speakers for language classes, foreign students, buildings, computers, classrooms, energy consumption, water consumption, the area of school, exam scores and so on, from 2010 to 2012. Note that each data sample includes many empty cells.

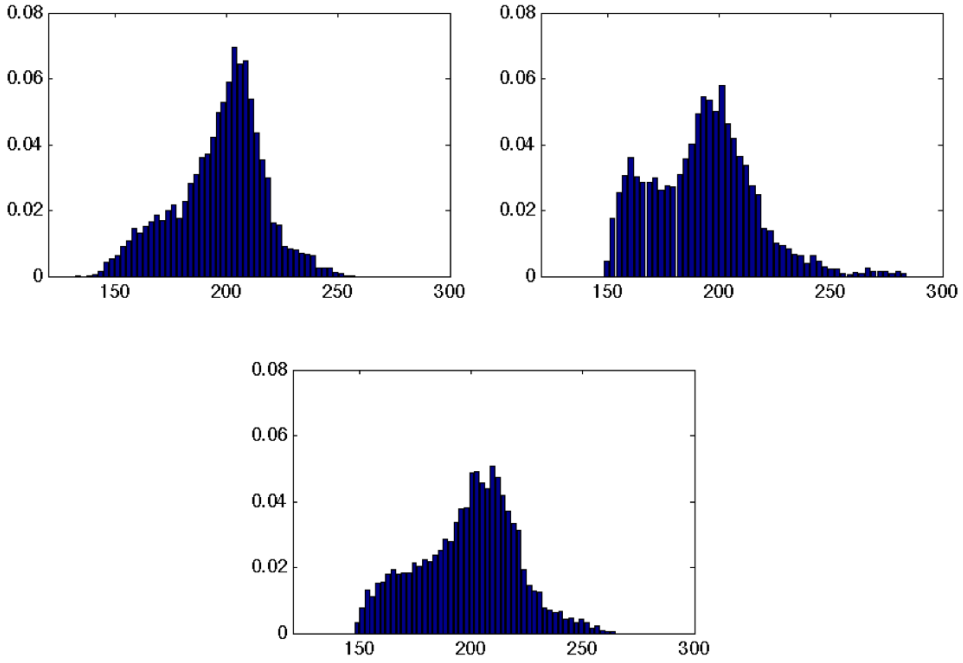


Figure 1. The Distributions of the Scores for Korean language (Top Left), Mathematics (Top Right) and English Language (Bottom). The Vertical and Horizontal Axes Indicate the Density [0, 0.08] and the Scores [120, 300], Respectively

3.2. Preprocessing

Although the EduData includes many scores, in this paper, we focused on 3 exam scores from high school data: Korean language, Mathematics, English language, Y=[Y_1, Y_2, Y_3]. The distributions of the scores are presented in Figure 1. They have different distribution shapes. Math has the most skewed distribution (see the kurtosis in Table 1),

and the average and highest scores are lower and higher than the other subjects, respectively. See Table 1 for the details of statistics.

Table 1. Statistics for the Scores

	Korean	Math	English
Mean	197.1	192.5	198.5
Standard Deviation	19.8	23.5	22.1
Kurtosis	3.0	3.6	2.7

Originally, there are 157 factors: $X=[X_1, X_2, \dots, X_{157}]$. After data cleansing to remove samples with too many empty cells, the number of factors is 107, as shown in Figure 2. To evaluate our model, we used randomly selected 90% of 5,593 samples to learn the parameters of the model and the rest to test the model.

4. Handling Missing Values

Since EduData includes many empty cells (or missing values), researchers need to deal with them. It is not desirable to set them simply to zero, since zero values have their own meaning in data. Before applying a prediction model, we filled the missing values by applying a trick with the same sprit of the Gibbs sampling method and the mean-field approximation.

Let X_{ij} be the *j*-th factor of the *i*-th sample. First, we filled them with random values. After finding the most similar samples based on the other factors except *j*-th one, the average value of the *j*-th factors from the *k* most similar samples replace the current value X_{ij}. This iteration repeats until all the missing values converge.

Then, we removed the factors that have zero variance, because they are constantly whose information amount is zero and not able to predict any other values including exam scores. Also, we normalized the factors so that their mean and standard deviation become 0 and 1, respectively.

5. Prediction Model

As a prediction model, we take a linear regression. Given 3 target scores $Y=[Y_1, Y_2, Y_3]$ and 107 factors $X=[X_1, X_2, \dots X_{107}]$ where all the cells are filled by the preprocessing step described above, the prediction model is defined as follows [5].

$$Y_i = \sum_j W_{ij} X_j \quad , \tag{2}$$

where $\boldsymbol{Y_i}$ is the *i*-th subject, $\boldsymbol{X_j}$ is the *j*-th factor, and $\boldsymbol{W_{ij}}$ is the regression weight of $\boldsymbol{X_j}$ to $\boldsymbol{Y_i}$. That is, $\boldsymbol{Y_i}$ is estimated by a weighted sum of all the factors. Given training data ($\boldsymbol{X}$, $\boldsymbol{Y}$), the weight matrix $\boldsymbol{W}$ is obtained by

$$W = YX^T (XX^T)^{-1} . \tag{3}$$

To test the model with test data ($\boldsymbol{x}$, $\boldsymbol{y}$), equation (2) with $\boldsymbol{W}$ from equation (3) is applied to estimate the score. Then to check the performance, the root mean squared error (RMSE) between the estimated score $\hat{y}_i$ and the true score y_i is measured by

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \quad . \tag{4}$$

6. Results

First, we checked how much effective the individual factors are to the exam scores based on the weight matrix $\boldsymbol{W}$. We expected that the scores would be influenced differently by the set of factors, which could lead to different $\boldsymbol{W}$ rows. The weights after training are presented in Figure 2, where we can see that the number of teachers and the number of students per classroom are important for all the 3 subject scores (The factor names which are originally Korean are not presented for clear presentation). That is, the more teachers, the higher the scores are expected. Also, the more students or the more classes the teachers teach, the lower the scores are expected. Interestingly, we could see that the factors about after-school are more important for English rather than Korean and Mathematics.

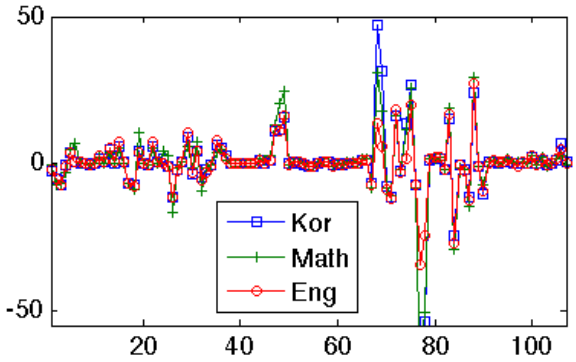


Figure 2. Weight Matrix for the 3 subjects. The High Values Mean They Make Positive Effects on the Corresponding Score

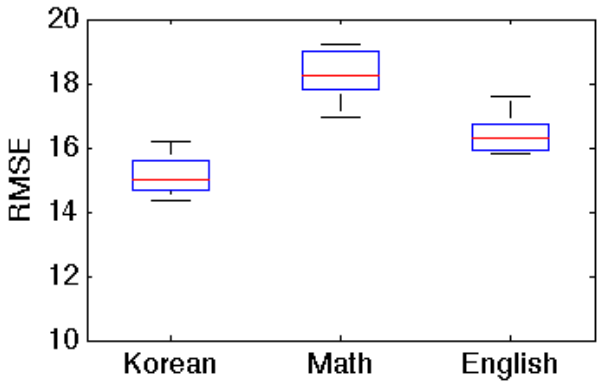


Figure 3. Boxplot of RMSEs for the 3 Subjects. Math is Harder to Predict Rather Than the Others

Then, given test data, we predicted exam scores with our model and measured the RMSE to evaluate the model performance. With random selections for training and test data sets, we conducted 10 experiments. The average RMSEs are 15.13, 18.49 and 16.35, for Korean, Mathematics and English, respectively. Figure 3 shows RMSEs for 3 subjects. As shown in Figure 3, it is hard to predict the score in Mathematics, while Korean and English are relatively easy. This could mean that the language learning achievement can be more improved by changes in school profile.

The correlation coefficients between the estimated and true scores for the test data are 0.65, 0.62, and 0.66 for the 3 subjects, respectively, as shown in Figure 4.

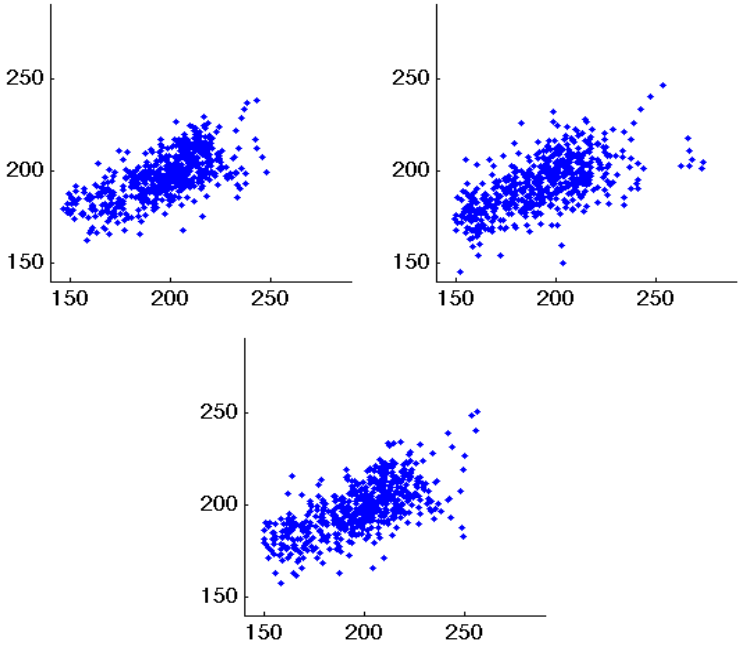


Figure 4. Score Predictions for Korean Language (top left), Mathematics (Top Right) and English (Bottom). In the 3 Graphs, the Vertical and Horizontal Axes Indicate the True and Estimated Scores, Respectively

7. Conclusions

In this paper, we proposed to apply statistical analysis to EduData. Our method showed how important the school profiles are to exam scores or learning achievement in school. Also, our model predicted the exam scores. This approach can be exploited to work out pedagogical strategy for more efficient investment to increase learning performance. Most significantly, regional analysis can provide a way for balanced educational environment.

In future research work, we can apply more complicated models (i.g. Granger causality or principal component regression) to make the prediction model more accurate. Also, to better understand the results, more pedagogical interpretations are required.

Acknowledgments

This work has been based on EDSS provided by Korean ministry of education. This work is partly based on the previous conference paper [6].

References

[1] Y. Kang and H. Choi, "An Empirical Study for Handling Scientific Datasets", International Journal of Grid and Distributed Computing, vol.5, no. 3 **(2012)**

[2] H. Choi, C. Yu, O. Sporns, and L. B. Smith, "From Data Streams to Information Flow: Information Exchange in Child-Parent Interaction", Proceedings of the Annual Meeting of The Cognitive Science Society, **(2011)**; Boston, MA.

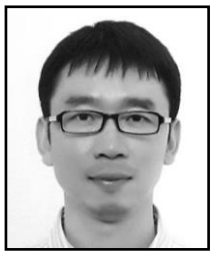
[3] C. Shon and K. Shon, "Effect of Study Motivation and Strategy in English Study", The Journal of Educational Research, vol. 9, no. 2, **(2011)**

[4] EDSS data: http://edss.moe.go.kr/

[5] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification", Wiley-Interscience; 2nd ed. New York, **(2001)**.

[6] H. Choi and Y. Kang, "Statistical Analysis for School Assessment Data", Proceedings of the International Conference on Platform Technology and Service, **(2014)**; Jeju, Korea.

Authors



Heeyoul Choi, He received his B.S and M.S. degrees in Computer Science from Pohang University of Science and Technology, Pohang, Korea, in 2002 and 2005, and his Ph.D. degree from Texas A&M University in Computer Science 2010. He was a post-doc researcher in Indiana University, Indiana from 2010 to 2011. He is currently a research staff member at Samsung Advanced Institute of Technology, Samsung Electronics, since 2011. His research interests cover deep learning, machine learning, pattern recognition, computational neuroscience, and cognitive science.

Yunhee Kang, He received his B.E. and M.S. degrees from Dongguk University, Seoul, Korea, in 1991, and 1993, respectively, and his Ph.D. degree from Korea University, Seoul, Korea, in 2002, all in computer engineering. He is currently an Assistant Professor with the Division of Information and Commutation, Baekseok University, Cheonan, Korea. His primary research interests lie broadly in distributed systems including fault-tolerance system, cloud computing and grid computing.