International Conference
NEW PERSPECTIVES in SCIENCE EDUCATION

# Analysis of the Influence of Learning State before University Admission to College Dropout Using Hierarchical Bayesian Model

**SHIRATORI Naruhiko (1), TAIRI Shintaro (2), OISHI Tetsuya (3), MORI Masao (4), MUROTA Masao (5)**

Kaetsu University, Japan (1)
Yokohama College of Commerce, Japan (2)
Tokyo Institute of Technology, Japan (3)
Tokyo Institute of Technology, Japan (4)
Tokyo Institute of Technology, Japan (5)

## Abstract

*In this research, we express the probability of college students dropping out using a hierarchical Bayesian model and derive to what extent the variables before admission influence college dropout. In Japan, about 5% of students entered university, but they drop out without graduation. Many studies have shown that the proportion of dropout students varies from university to university and correlates with university scale and deviation value. Many researches have been made to forecast dropouts in advance, but studies have not been made to express how the dropout probability dynamically changes depending on the situation before admission. It is assumed that the change in the dropout probability varies according to the variables before enrollment (number of absences at high school, type of high school etc.). Variables used in the model consist of pre-admission variables and post-admission variables, and the post-admission variables are the number of units per semester and GPA. We created a hierarchical Bayesian model using pre-admission variables and post-admission variables to derive the drop-out probability for that term. It was found that the probability of dropping out depends on the variables before entering the school, especially the type of high school and the number of absences. The probability of withdrawal is increased because the type of high school is communication system and the number of days absent at high school is large. Furthermore, we found that the change in the probability of dropout for each term differs for each variable before enrollment.*

**Keywords:** *Hierarchical Bayesian Model, Logistic Regression, Dropout;*

## 1. Introduction

The dropout rate in universities in Japan has been rising since the 1990s and is becoming a serious problem [1]. Tellingly, the student dropout rate has become a major point of interest in the annual survey conducted by the Yomiuri Shimbun and is increasingly used in university evaluations [2]. Overall, 8% of Japanese students drop out of college, with a lower rate at the national and public universities and a higher rate at private universities. Since the rate of graduation in Japan is higher than other OECD countries, there is not much to discuss dropouts in Japan. Especially at Kaetsu University where the author belongs, it shows a high dropout rate of 34.7% in the survey of 2014.

In this research, we express the probability of college students dropping out using a hierarchical Bayesian model and derive to what extent the variables before admission influence college dropout. To predict dropping out, you can do a regression using the results of the semester and the unit number. However, the score and the number of credits will change depending on the state before enrollment. In this research, we group state before enrollment and create a dropout prediction model with the group difference taken into account using hierarchical Bayesian model. After that, by examining the results of the dropout prediction model with the group difference taken into consideration and the results of the nonadditional model, we examine how much the group difference before entering the classroom is heard.

## 2. Related Works

Research on predicting student dropouts and establishing the underlying causes is widely conducted. It has been shown that there are a number of variables affecting the decision to dropout, including gender, teacher-student ratios, deviation values, university size, economic factors and variables related to dropping so far.

Tajiri analyses the relationship between educational history and dropping out at a business university and derives a set of micro variables for leaving students [3]. He found a low probability of female

students dropping out within four years, but found that the probability is reversed beyond the four-year period. He also identified factors such as academic ability tests, grades, and number of credits being taken as significant micro variables related to dropout.

According to Kondo [4], a dropout prediction can be made at the beginning of the third year by using prior data (sex, undergraduate, entrance examination class, attendance rate, etc.). He compares various methods for making dropout predictions, including a logistic regression model, a support vector machine, and the random forest algorithm. In the U.S., where retention rate (school enrolment rate) is often used rather than dropout rate, Bingham et al. [5] used a logistic regression model to show that enrolment rate differs according to the parents' educational background and ethnicity.

## 3. Methods

### 3.1 Data

The data in this study is from students who entered K University in Tokyo, Japan, from 2012 to 2014. The university is a college that teaches the social sciences. Transfer students and graduate students are excluded from this data. There are 657 students with no missing data, of which the number of dropouts is 194 people.

It is known that dropping out has a correlation with the results at the first and second semester in the university. For example, in the study of Kondo [4], at the end of the 5th week of the 1st spring semester, it is estimated that about 40% of students who drop out by the beginning of the third year can be predicted.

In the graph 1 below, it is a box plot chart showing the GPA of the 1st spring semester of a dropout student and a regular student. The median value of GPA of students who dropped out is 2.36, and the median value of GPA of regular students is 1.43. As can be seen from this figure, there is a clear difference between the dropout student and the other students after entering the university.

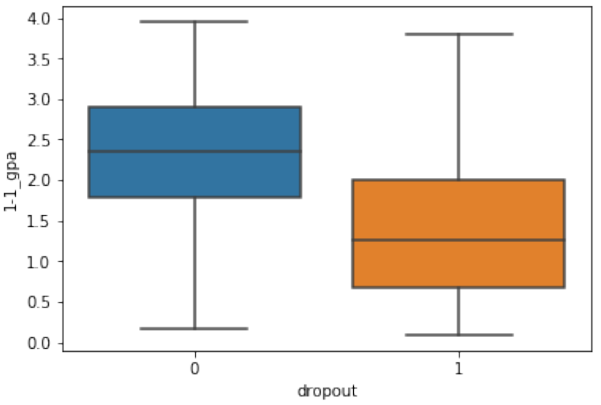


Fig. 1: GPA of 1st semester between dropout student or not

Next, it checks whether the number of absence days at high school which is a pre-admission variable has a relation with GPA which is a variable after admission. Figure 2 below is a box plot showing the relationship between absence days at high school and GPA of 1st semester. The X axis is a discretized number of absent days at high school. On the X axis 1 represents absent days from high school days 0 to 9, 2 from 10 to 19 days, 3 from 20 to 29 days, and 4 represents over 30 days. The Y axis is GPA of 1st semester. The median value of each is 2.25, 2.11, 1.82, 1.58. From this we can see that the variables after enrollment are influenced from the pre-admission variables.

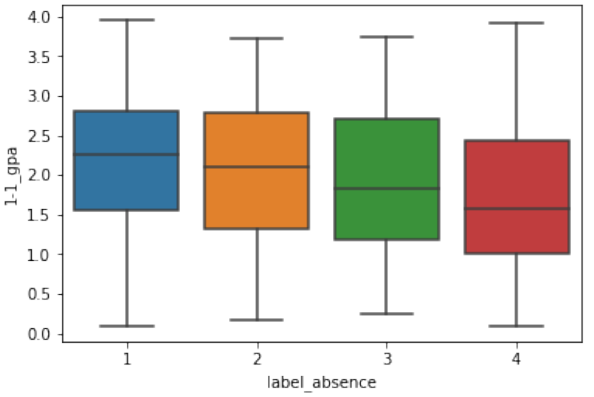


Fig. 2: GPA of 1st semester and high school absent days

### 3.2 Variables

Among the data, the variables to be used this time are the following five.

- 1. Fulltime Highschool or not: a student in fulltime high school or not (0,1)
- 2. Number of Absence in high school (1,2,3,4)
- 3. 1-1 GPA: GPA in the 1st semester (numerical, lower=0, upper=4)
- 4. 1-1 credits (numerical, lower=0, upper=24)
- 5. dropout (A student who dropout of the college or not): (0,1)

In this study, we use 1. high school type and 2. high school absent days as pre-admission variables. High school type represents whether it is full-time high school or not, high school absent days are discretized into 4 and used. 1 in high school absence days represents absent days from 0 to 9, 2 is from 10 to 19, 3 is from 20 to 29, and 4 is over 30 days. As the variable after enrollment, use 3.1st semester's GPA and 4.1st semester unit number. However, since the variables 3 and 4 after enrollment are highly correlated (corr = 0.81), only the GPA of 3 is used as a variable after admission. The relationship between the GPA of the variable 3 after admission and the unit number of the post-admission variable 4 is represented by the following scatter diagram, Fig.3.

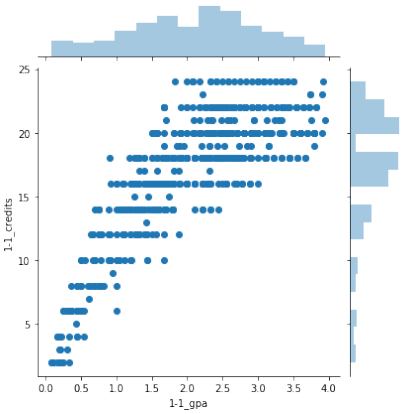


Fig. 3: scatter diagram of GPA & Credits

### 3.3 Model

Using pre - admission variables and post - admission variables, we can verify how much we can predict dropping out. For verification, the following three models are defined. In addition, logistic regression models are used for each model.

Model 1: using only post-admission variables

$$Y[n] \sim Bernoulli(\frac{1}{1+exp(-(a_1 GPA[n] + b_1))}$$

n = 1, 2, ..., N. N represents the number of students and n represents the index of the student. Y represents dropout/not. a - 1 is a GPA variable, and b is an intercept

Model 2: using post-admission variables for each group of pre-admission variables

$$Y[n] \sim Bernoulli(\frac{1}{1+exp(-(a_1[LAbsence[n]] GPA[n] + b_1[LAbsence[n]]))}$$

n = 1, 2, ..., N. A 1 and b for logit function change for each absence group. N represents the number of students and n represents the index of the student. Y represents dropout/not.

Model 3: using groups of pre-admission variables and post-admission variables, Bayesian hierarchical Model

$$Y[n] \sim Bernoulli(\frac{1}{1+exp(-(a[LAbsence[n]] GPA[n] + b[LAbsence[n]]))}$$

$$a[k] = a_{all} + a_{group}[k]$$
$$b[k] = b_{all} + b_{group}[k]$$
$$a_{group}[k] \sim Normal(0, \sigma_a)$$
$$b_{group}[k] \sim Normal(0, \sigma_b)$$

In model 3, a [k] of each group is divided into terms expressing a common overall average in all Absence groups and terms expressing each group difference. a_all and b_all are average values of the entire group, a_group, b_group are variables representing each group difference. Each group difference is assumed to follow the average 0 and standard deviation σ. k is 4 representing the number of absent groups in high school.
These three models were modeled by Stan and Bayesian estimation was performed using PyStan and MCMC (Markov Chain Monte Carlo methods). For model 1 and model 2, Iteration was used 2000 times, warmup was 500 times, chain number was 4. For model 3, Iteration was used 7000 times, warmup was 1000 times, chain number was 4

## 4. Results

The estimation result of the parameter of model 1 is as follows. The average value of a1 is -1.32, the average value of b is 1.63, and both of Rhat are 1.1 or less, and even if you look at the graph, it converges.

| | mean | se_mean | sd | 2.5% | 25% | 50% | 75% | 97.5% | n_eff | Rhat |
|---|---|---|---|---|---|---|---|---|---|---|
| a1 | -1.32 | 3.1e-3 | 0.12 | -1.55 | -1.4 | -1.32 | -1.24 | -1.09 | 1484 | 1.0 |
| b | 1.63 | 6.0e-3 | 0.23 | 1.18 | 1.47 | 1.62 | 1.78 | 2.08 | 1474 | 1.0 |
| lp__ | -324.2 | 0.02 | 0.98 | -326.7 | -324.5 | -323.9 | -323.5 | -323.3 | 1768 | 1.0 |

The estimation result of the parameters of Model 2 is as follows. The slope and intercept for each group converged because Rhat is less than 1.1.

| | mean | se_mean | sd | 2.5% | 25% | 50% | 75% | 97.5% | n_eff | Rhat |
|---|---|---|---|---|---|---|---|---|---|---|
| a[1] | -1.27 | 2.4e-3 | 0.16 | -1.58 | -1.37 | -1.26 | -1.16 | -0.97 | 4178 | 1.0 |
| a[2] | -1.67 | 6.2e-3 | 0.38 | -2.47 | -1.92 | -1.66 | -1.4 | -0.98 | 3805 | 1.0 |
| a[3] | -0.81 | 6.0e-3 | 0.38 | -1.62 | -1.05 | -0.8 | -0.54 | -0.1 | 4132 | 1.0 |
| a[4] | -1.67 | 6.5e-3 | 0.42 | -2.54 | -1.93 | -1.64 | -1.38 | -0.93 | 4138 | 1.0 |
| b[1] | 1.35 | 4.6e-3 | 0.3 | 0.77 | 1.14 | 1.34 | 1.55 | 1.96 | 4347 | 1.0 |
| b[2] | 2.14 | 0.01 | 0.69 | 0.85 | 1.66 | 2.12 | 2.59 | 3.56 | 3953 | 1.0 |
| b[3] | 1.9 | 0.01 | 0.86 | 0.3 | 1.31 | 1.85 | 2.44 | 3.75 | 4002 | 1.0 |
| b[4] | 2.59 | 0.01 | 0.71 | 1.3 | 2.11 | 2.55 | 3.05 | 4.04 | 4192 | 1.0 |
| lp__ | -317.7 | 0.04 | 2.05 | -322.6 | -318.9 | -317.4 | -316.2 | -314.8 | 2131 | 1.0 |

The estimation result of the parameters of Model 3 is as follows. The slope and intercept for each group converged because Rhat is less than 1.1.

| | mean | se_mean | sd | 2.5% | 25% | 50% | 75% | 97.5% | n_eff | Rhat |
|---|---|---|---|---|---|---|---|---|---|---|
| a0 | -1.27 | 0.01 | 0.4 | -2.01 | -1.43 | -1.26 | -1.1 | -0.52 | 1252 | 1.0 |
| b0 | 1.84 | 0.01 | 0.68 | 0.74 | 1.49 | 1.78 | 2.11 | 3.22 | 2868 | 1.0 |
| al[1] | -0.06 | 0.01 | 0.4 | -0.82 | -0.24 | -0.05 | 0.09 | 0.7 | 1174 | 1.0 |
| al[2] | -0.16 | 0.01 | 0.42 | -1.05 | -0.32 | -0.1 | 0.03 | 0.5 | 1279 | 1.0 |
| al[3] | 0.3 | 0.01 | 0.44 | -0.35 | 0.05 | 0.26 | 0.47 | 1.2 | 1391 | 1.0 |
| al[4] | -0.06 | 0.01 | 0.42 | -0.97 | -0.21 | -0.03 | 0.11 | 0.63 | 1239 | 1.0 |
| bl[1] | -0.37 | 0.01 | 0.69 | -1.82 | -0.64 | -0.26 | -4.6e-3 | 0.67 | 2677 | 1.0 |
| bl[2] | -0.08 | 0.01 | 0.7 | -1.5 | -0.33 | -0.03 | 0.18 | 1.21 | 3495 | 1.0 |
| bl[3] | 0.25 | 0.01 | 0.76 | -1.09 | -0.06 | 0.13 | 0.56 | 1.87 | 3287 | 1.0 |
| bl[4] | 0.18 | 0.01 | 0.69 | -1.06 | -0.07 | 0.1 | 0.43 | 1.6 | 3509 | 1.0 |
| s_a | 0.52 | 0.02 | 0.66 | 0.03 | 0.2 | 0.34 | 0.6 | 2.1 | 1572 | 1.0 |
| s_b | 0.85 | 0.03 | 1.18 | 0.02 | 0.28 | 0.57 | 1.02 | 3.46 | 2193 | 1.0 |
| a[1] | -1.32 | 1.7e-3 | 0.14 | -1.61 | -1.42 | -1.33 | -1.23 | -1.05 | 6784 | 1.0 |
| a[2] | -1.43 | 0.02 | 0.26 | -2.0 | -1.58 | -1.41 | -1.25 | -1.0 | 210 | 1.01 |
| a[3] | -0.96 | 0.01 | 0.3 | -1.52 | -1.18 | -0.96 | -0.74 | -0.42 | 481 | 1.0 |
| a[4] | -1.33 | 0.02 | 0.28 | -1.96 | -1.49 | -1.32 | -1.14 | -0.89 | 198 | 1.01 |
| b[1] | 1.47 | 2.2e-3 | 0.28 | 0.92 | 1.29 | 1.48 | 1.65 | 2.02 | 15941 | 1.0 |
| b[2] | 1.76 | 8.6e-3 | 0.45 | 0.94 | 1.47 | 1.72 | 2.02 | 2.75 | 2755 | 1.0 |
| b[3] | 2.09 | 0.03 | 0.62 | 1.05 | 1.62 | 2.01 | 2.49 | 3.43 | 613 | 1.0 |
| b[4] | 2.01 | 0.02 | 0.48 | 1.25 | 1.66 | 1.97 | 2.3 | 3.08 | 401 | 1.01 |
| lp__ | -315.1 | 0.29 | 4.06 | -323.4 | -317.7 | -315.0 | -312.4 | -307.2 | 197 | 1.01 |

## 5. Discussion

By comparing models 1, 2 and 3, it became possible to express common items and differences for each pre - admission variable group. The average of common coefficients was -1.27, the intercept was 1.84, and we could derive the parameter difference for each pre-entrance variable group. For example, look at the case of a student whose GPA is 1, the probability of dropout in model 1 is 0.576, and in case of model 2 it varies from 0.519 to 0.74 according to the number of absences before enrollment. Furthermore, in the case of Model 3, it can be represented as high precision of the system in the form of common variables and group differences.

## Acknowledgment

This work was supported by JSPS KAKENHI Grant-in-Aid for Scientific Research C 15K04380.

## References

[1] Hozawa Yasuo: Gakko Kihonchosa nimiru Chutai to Ryunen, IDE—Gentai no Kotokyoiku 546: 64-67, 2012.
[2] Yomiurishinbun Kyoiku Network Jimukyoku: Daigakuno Jitsuryoku 2016, Chuo Koronsinsha, 2015.
[3] Shintaro Tajiri et al. : Bijinesukei Daigaku ni okeru Gakushurireki to Katsudo Data wo Motiita Seizonjikan Bunseki, The Japan Society for Educational Sociology 65th Annual Report 120-121, 2013.
[4] Kondo Nobuhiko, Hatanaka Toshiharu: “Modelling of Students' Learning States Using Big Data of Students through the Baccalaureate Degree Program” Japanese Society for Information and Systems in Education 33(2), 94-103, 2016
[5] Bingham, M. A., & Solverson, N. W.: “Using Enrollment Data to Predict Retention Rate.” Journal of Student Affairs Research and Practice, 53(1), 51–64. 2016