A Comparative Study of Bayesian Estimators for the Prediction of Time Spent on Online Class during COVID-19

Songjie GUO

Abstract

This report aims to use Bayesian methods to predict the time spent on online class of students with different age distributions. A simple regression model is tried first, followed by hierarchical models and change point analysis. Bayesian estimators introduce a parameter of adaptability of students of different age groups, thus exhibiting greater interpretability than a non-Bayesian one. The change point analysis further explore more reasonable divisions of age groups (7-12, 13-37, 38-50) to improve the hierarchical model. While the performance of the error metrics (RMSE, MAE) is roughly indistinguishable among Bayesian estimators and a non-Bayesian one, the Bayesian approach infers more intrinsic conclusions.
github: (https://github.com/songjie-guo/bayes)

KEYWORDS: Bayesian estimators, Hierarchical models, Change point analysis.

1. INTRODUCTION

The outbreak of COVID-19 has forced people to quarantine their homes to prevent the spread of virus. Many students shifted from face-to-face classes to online education. Such change in the form of study requires students' adaptability, which is a great concern for both teachers and institutions. The aim of this report is to explore the adaptability level of students of different ages, as well as to make predictions about the time of online study per day using Bayesian estimators. These Bayesian estimators will also be compared to a frequentist one.

1.1 Related Work

During the COVID-19, the closure of educational institutions required for a rapid transition to digital learning [3]. This rapid evolution of online class at a large scale has influenced students of all age groups [2]. The study [1] investigated the possible consequences of COVID-19 on the life of students of different age groups including the time spent on online class, mental situation. It [1] finds that the average time spent on online courses was significantly higher for students in the age group of 7-17 years (3.69 hrs/day), while the average time spent on online courses was lower for students in the age group of 18-22 years (2.98 hrs/day) and students in the age group of 23-59 years (2.66 hrs/day) ($P < 0.0001^*$).

1.2 Contribution

A change point analysis was used to try to rationalize the age groupings. Through the new clustering and the updated hierarchical modeling, it was concluded that there may be a more pronounced difference in the adaptability level of students between the ages of 7-12, 13-37, and 38-59. Comparisons were also made between Bayesian methods and a non-Bayesian one for the performance on the prediction of the time spent on online class.

2. OBJECTIVES

2.1 Dataset

The dataset used is from a web-based survey conducted to students through Google online platforms from July 13 to July 17, 2020 [1]. The dataset contains participants' demographics such as age, the region of residence. The time spent for online class daily is an important indicator of how well students are learning online. This dataset also contains other information including assessment of the experience of online learning, assessment of health issues and so on. In general, the dataset investigates the impact of the COVID-19 pandemic on the education, health and lifestyle of students of different age groups [1].

Since ... In this report, I focus on the relationship of two variables, age (`Age_of_Subject`) and the time spent for online study daily (`Time_spent_on_Online_Class`).

2.2 Problem Formulation

The `Time_spent_on_Online_Class` is an integer variable with values between 0 and 24. It is the number of hours a student attends online classes on a daily basis.

The (`Age_of_Subject`) is also an integer variable with values between 7 and 59. However, there are two ways to treat this variable. One way is to convert them into a categorical variable with three groups (7-17, 18-22 and 23-59),

which is used in the study [1] to draw conclusion in a frequenist way. Another way is to treat them as indices. Both ways will be used in the report for analyses.

The problem is to create a Bayesian estimator to predict the time spent on online class (Y_i) of different ages or age groups (i).

3. METHODS

I tried several different Bayesian estimators to predict the time spent on online class of students of different ages. First, a simple regression model was considered to predict a mean value for each age. Then, I categorized the age into groups to follow the idea in previous study [1]. Since there is more than one set of cut-off points for age groups, I tried to use change point analysis to find out other possible ways of dividing ages. Finally, I updated the previous hierarchical model with the new prior choice.

After these estimators were built, I conducted posterior checks and correctness tests for evaluation of the posterior approximation. Calibration using leave-one-out method, synthetic datasets were also included for different estimators.

4. ANALYSIS

4.1 A Simple Regression Model

A simple way of thinking is to predict the mean (rounded to integer) of the time spent on online class (Y_i) with different ages (i). As shown in Figure 1 , the age variable i is on the x-axis and the dependent variable is Y_i.

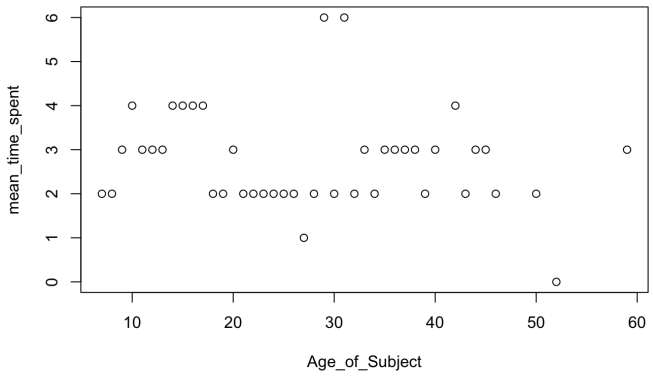

Figure 1: Plot of "average time spent on online class" and "age"

Here, I form a simple regression model. For each Y_i, I consider it to follow a Binomial distribution with $n = 24$, which is the maximum number of hours per day. p_i is a parameter reflecting a student's adaptability level. The assumption here is that the higher p_i, the higher probability that a student is spending time for online class in a given hour. The mathematical expression of the regression model is

$$\text{slope} \sim \mathcal{N}(0, 1) \tag{4.1}$$
$$\text{intercept} \sim \mathcal{N}(0, 1) \tag{4.2}$$
$$p_i = \text{logistic}(\text{slope} \cdot i + \text{intercept}) \tag{4.3}$$
$$Y_i \sim \text{Binom}(24, p_i) \tag{4.4}$$

The sampling result under SimPLLe is that intercept $= -0.03198$, slope $= -0.03547$. The small values of slope and intercept indicate that the model may not be well defined. To evaluate the performance of this simple regression model, I draw a plot of posterior distributions.

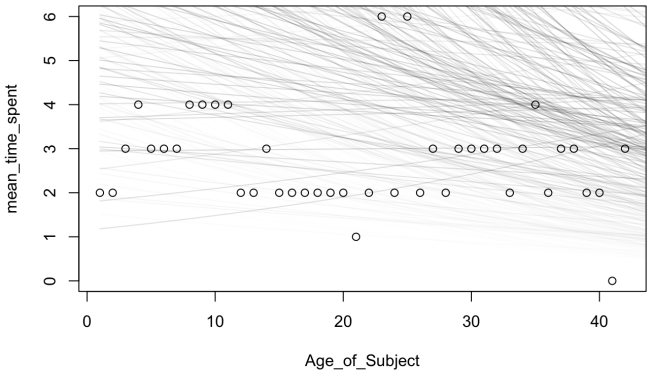

Figure 2: Posterior distributions after regression

The posterior distributions obtained from sampling in Figure 2 is very heterogeneous and do not fit well with the data points, suggesting that the regression is not satisfactory.

4.2 A Hierarchical Model

In the study [1], age is not considered on a stand-alone basis. Instead, it divide ages into three groups (7-17, 18-22 and 23-59). Following this line of thinking, I can categorize people according to the above groups and solve the problem with a hierarchical model. The densities in Figure 3 have some differences between each groups, but also some shapes in common. Therefore, a hierarchical model is considered.

For each Y_i, it still follows a Binomial distribution as in section 4.1. μ and λ are population parameter shared across three age groups. p_i is the "nuisance" parameters for each age group ($i \in \{1, 2, 3\}$) to reflect the adaptability level of online class of that age group.

The graph model of the hierarchical model is shown in Figure 4. The mathematical expression of the hierarchical model is

$$\mu \sim \text{BetaMP}(0.1, 10) \tag{4.5}$$
$$\lambda \sim \text{Exp}(0.001) \tag{4.6}$$

Figure 3: Density plots of "time spent on online class" under different "age groups"

$$p_i \mid \mu, \lambda \sim \text{BetaMP}(\mu, \lambda), \quad i \in \{1, 2, 3\} \tag{4.7}$$

$$Y_i \sim \text{Binom}(24, p_i) \tag{4.8}$$

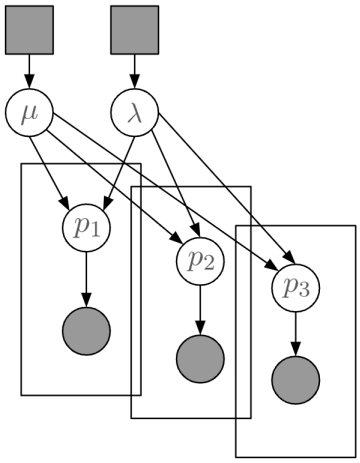

Figure 4: Graph model for the hierarchical model (4.2)

After implementing the above model into Stan, I get the result in Table 1, where the superscripts 1-3 stand for age group of 7-17, 18-22 and 23-59).

To evaluate the goodness of the fit, I use the leave-one-out technique to check the calibration on the prediction Y_i. The result shows that the prediction is calibrated with the actual coverage at 81.05% for the nominal coverage 80%. Therefore, the model is well-specified.

However, in Figure 5, the distribution of p of age group 18-22 is similar to that of age group 23-59. This indicates that the division of age groups may have been problematic. It fails to differentiate between students with different adaptability levels, which may have led to inaccurate results in the prediction of Y_i.

Table 1. Stan output of hierarchical model (4.2)

Rhat	mean
μ	0.10
λ	980.26
p_1	0.18
p_2	0.12
p_3	0.12
Y_1	4.38
Y_2	2.79
Y_3	2.71

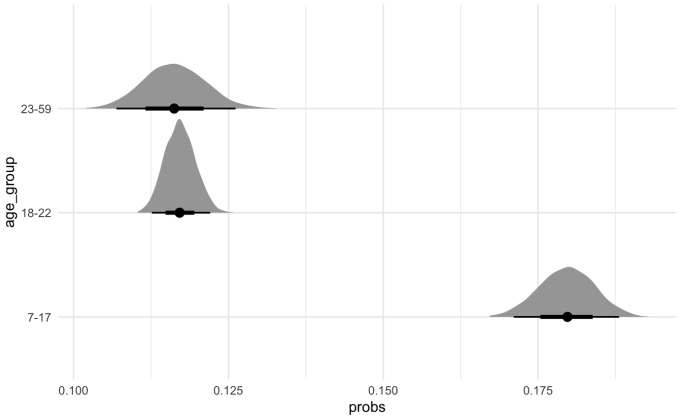

Figure 5: Posterior distributions of p_i (4.2)

4.3 Change Point Analysis

From the above analysis, it is important to derive a more reasonable division of age groups with statistical analysis. Therefore, I conduct change point analysis to explore possible cutting points for ages.

From Figure 5, the p_i has two distinct clusters (7-17 and 18-59). Thus, I form the model of a single change point. C is the parameter of the change point with a Uniform distribution between 7 and 59. The change point C separates the time spend on online class (Y_{age}) into two clusters or age groups ($i \in \{1, 2\}$). The mathematical expression of the change point model is

$$p_i \sim \text{BetaMP}(0.1, 10), \text{ for } i \in \{1, 2\} \tag{4.9}$$

$$C \sim \text{Unif}\{7, 8, \ldots 59\} \tag{4.10}$$

$$Y_{\text{age}} \sim \text{Binom}(24, p_{\mathbb{1}[\text{age}<C]+1}), \text{ for age} \in \{7, 8, \ldots, 59\} \tag{4.11}$$

The method is to use a custom MCMC sampler with alternation of two kernels K_1 and $K2$

$$K := K_1 \circ K_2,$$

where K_1 only modifies $\mathbf{p} = [p_1, p_2, p_3]$ and K_2 only modifies the change point C. For the $\mathbf{p}$, I use a trivariate standard

normal distribution centered at the current **p**. For updating change point C, I use a centered binomial random variable as update

$$b \sim \text{Binom}(106, 0.5) \tag{4.12}$$
$$\text{proposed}C = C + b - 53 \tag{4.13}$$

After sampling the MCMC, I conduct an "exact invariance test" to validate its correctness. The p-value of Kolmogorov-Smirnov test is 0.3136 and the p-value of Welch Two Sample t-test is 0.1001, implying no significant issues found.

Figure 6 shows that change points around 10 and 40 appear frequently. The density plot (Figure 7) also suggests a change point at 12 or 37. Therefore, the new division of age groups is 7-12, 13-37 and 37-59.

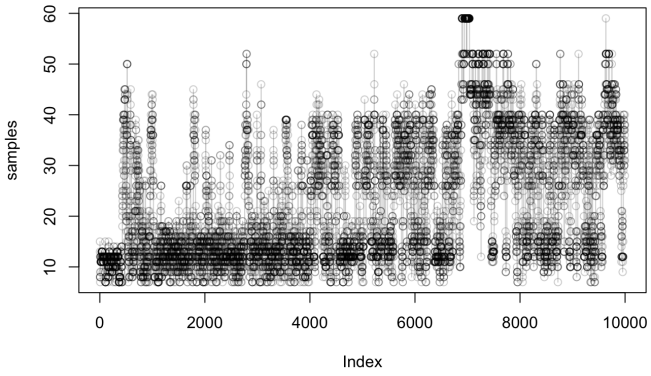

Figure 6: Change point trace plot

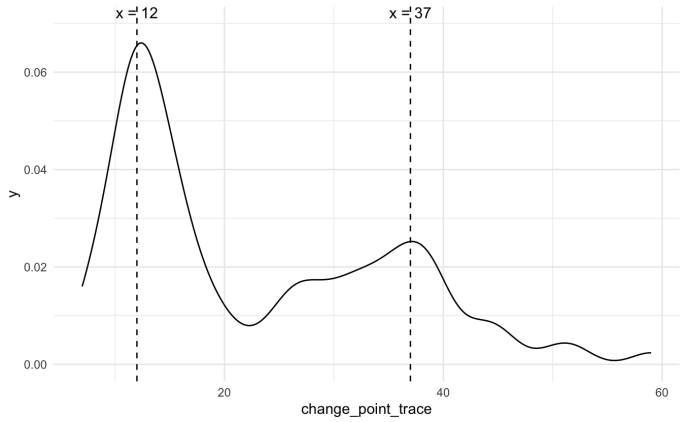

Figure 7: Density plot of change point trace

4.4 An Updated Hierarchical Model

With the new age groups, I replicate the previous analysis in section 4.2. The different distributions in Figure 8 are more discriminatory than those in Figure 3, suggesting that the new age divisions may better differentiate between students' adaptability level and their time spent on online class.

Figure 8: Density plots under new "age groups"

The mathematical expression of the updated hierarchical model is

$$\mu \sim \text{BetaMP}(0.1, 10) \tag{4.14}$$
$$\lambda \sim \text{Exp}(0.001) \tag{4.15}$$
$$p_i \mid \mu, \lambda \sim \text{BetaMP}(\mu, \lambda), \quad i \in \{1, 2, 3\} \tag{4.16}$$
$$Y_i \sim \text{Binom}(24, p_i) \tag{4.17}$$

Table 2 show the result of Stan.

Table 2. Stan output of hierarchical model (4.4)

Rhat	mean
μ	0.10
λ	1030.62
p_1	0.15
p_2	0.13
p_3	0.12
Y_1	3.58
Y_2	3.17
Y_3	2.91

For Figure 9 of posterior distributions of p_i, the updated distributions of p_i are more separate from each other. This indicate that the new division of age groups are more reasonable and succeeds to group students by their adaptability level.

5. RESULTS

5.1 Relationship of Age and Adapability

For the study [1], it uses a Kruskal Wallis test to produce a p-value to analyze the significance of difference between different age groups. The statistical test results in

Table 3. Results of Bayesian estimators and a non-Bayesian one

Estimator	Age Groups	p_1	p_2	p_3	Y_1	Y_2	Y_3	MAE	RMSE
1) Arithmetic Mean (Frequenist)	7-17, 18-22, 23-59	0.15	0.12	0.11	3.69	2.98	2.65	1.64	2.04
2) Posterior Mean of Model (4.2) (Bayesian)	7-17, 18-22, 23-59	0.18	0.12	0.12	4.38	2.79	2.71	1.62	2.01
3) Posterior Mean of Model (4.4) (Bayesian)	7-12, 13-37, 38-59	0.15	0.13	0.12	3.58	3.17	2.91	1.72	2.11

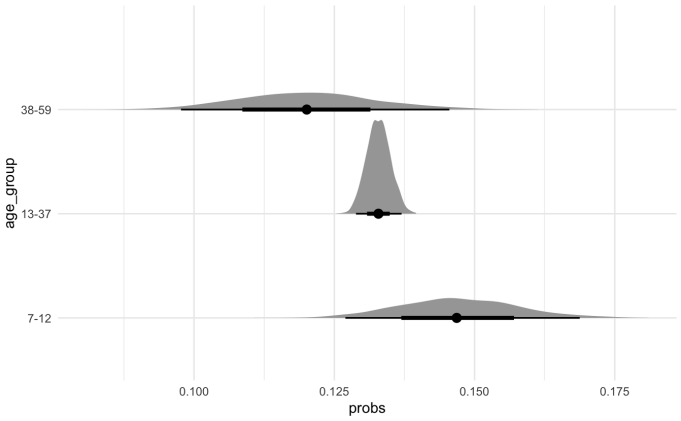

Figure 9: Posterior distributions of p_i (4.4)

$P < 0.0001^*$, indicating that there are significant difference in the time spent on online class between different age distributions.

In the Bayesian method, I introduce a parameter p to reflect the adaptability level of students for online class. If the posterior distributions of p_i $(i \in \{1, 2, 3\})$ have clear demarcation between each other, students of different age groups have different adaptive abilities to online class, which in turn essentially leads to a significant difference in the time spent on it.

From Figure 5, only p of age 7-17 is significantly higher than others. A new inference is drawn that the adaptability level of age 7-17 is higher, while the difference of time spent on online class between 18-22 and 23-59 may not be significant. After attempting to analyze the age divisions with change point analysis, the newly obtained posterior distributions of p have a more pronounced demarcation, as shown in Figure 9. The result indicates that the adaptability of students may significantly differentiate among age 7-12, 13-37 and 38-59.

5.2 Prediction of Time Spent on Online Class

The estimator of the study [1] is a method of point estimate, which calculates the mean values for each age groups. The two Bayesian estimators of hierarchical models are the posterior means of the time spent on online class. MAE and RMSE are used to compare the prediction performance of theses three estimators, and the results are shown in Table 3. For the three estimators, the MAE and RMSE are very close, with a slightly larger error for estimator 3). The overall performance of these estimators are similar.

6. DISCUSSION

In this report, two Bayesian estimators are formed through the idea of hierarchical modeling. They are under different priors through different divisions of ages. However, the performance of these more complicated Bayesian does not exceed the simple mean estimator for the prediction of time spent on online class.

One possibility is that in the Bayesian estimator I only considered the effect of a single variable age on adaptability. In reality, however, there may be a variety of other factors to inflect this parameter. Thus, despite the fact that the Bayesian model is much more sophisticated, it is unable to sample a better output under the limited information available.

Another limitation lies in the change point analysis. From the Figure 6, the trajectory of the change points do not converge after many iterations. It keeps fluctuating over a wide range, except that it stays longer at 12 and 37 (the good neighborhoods). This suggests that there may be multiple change points, or multiple local minima when simply performing change point analysis for two clusters. Thus, it may cause the model to not find the optimal solution well when searching for a single change point. Subsequent studies may try to optimize it with parallel tempering.

REFERENCES

[1] Chaturvedi, K., Vishwakarma, D. K. and Singh, N. (2020). COVID-19 and its impact on education, social life and mental health of students: A survey. *Children and Youth Services Review* **121** 105866–105866.

[2] Hasan, N. and Bao, Y. (2020). Impact of "e-Learning crack-up" perception on psychological distress among college students during COVID-19 pandemic: A mediating role of "fear of academic year loss". *Children and Youth Services Review* **118** 105355–105355.

[3] Kapasia, N., Paul, P., Roy, A., Saha, J., Zaveri, A., Mallick, R., Barman, B., Das, P. and Chouhan, P. (2020). Impact of lockdown on learning status of undergraduate and postgraduate students during COVID-19 pandemic in West Bengal, India. *Children and Youth Services Review* **116** 105194–105194.

Songjie Guo . The University of British Columbia, Canada.
E-mail address: guosj914@gmail.com