ASEE 2022 ANNUAL CONFERENCE
Excellence Through Diversity
MINNEAPOLIS, MINNESOTA, JUNE 26TH-29TH, 2022 ASEE
Paper ID #38338

Linkage between Students' Study Habits and their grades analyzed through Bayesian statistics

Muhammad Dawood (Dr.)

Melissa J. Guynn

Patti Wojahn (Professor)

Linkage between Students' Study Habits and their grades analyzed through Bayesian statistics

Abstract: It is well documented that particular study habits and strategies, among other factors, can influence students' grades and contribute to college success, independent learning, and retention. To further explore such connections, we studied and analyzed distinctions between two groups, those reporting use of "good" study habits vs. "not so good" study habits. Results establish a link between the two groups and Exam 1 grades, coded as Pass/Fail. Logistic regression, Bayesian statistics, and Matlab programming language were utilized for analysis, indicating that study habits depicted in the group reporting use of study habits research indicates as more powerful are indeed more likely to obtain a passing grade compared with those who did not report employing the better study strategies. because Bayesian statistics are not commonly employed to analyze these effects, a brief is provided on Bayesian statistics to facilitate researchers in this field who might also wish to use Bayesian statistics in their work.

Introduction

Students' performance in a majority of STEM courses is recognized in terms of their grades. There is a growing body of literature suggesting positive correlation between students' study habits, their grades, independent learning, and retention in STEM courses (Walck-Shannon et al, 2021 [1], Kaur and Singh, 2020 [2], Veenstra et al, 2009, [3], Scalise et al, 2000 [4], Besterfield et al, 1997 [5], Shuman et al, (2003) [6], Blumner and Richards, (197) [7], Dey, (2014) [8]). Moreover, Blumner and Richards [7] report that more meaningful study habits could improve overall academic performance. Fortunately, as Bandura (Bandura, 1993 [9]) and Dweck (Dweck, 2006 [10]) indicate, study habits and skills can be developed through practice and positive reinforcement, inculcating self-efficacy.

Most students enter college directly from high schools, where many of them were successful, but high school study skills tend to prove unrelated to academic success in college (Matt et al, 1991 [11], Balduf, 2009 [12]). As a result, struggling students in hardcore college STEM courses often wonder what they may do differently to earn better grades. Responses from a majority of instructors tend to boil down to "spend more time on the subject," "complete the homework," and so on. Such advice may satisfy some students but many are already spending greater amount of time outside the class for reading and doing the assigned homework, without the success they desire.

Confronted with the above dilemma, the first author attended a workshop by Dr. McGuire on keys to focus on learning, work that is now available in book form (McGuire, 2015 [13]). Table 11.8 (page 147 [13]) from McGuire's book, with some modifications, was selected for intervention in various engineering classes. The author left the workshop with a goal to suggest particular proven, more effective yet simple study strategies for students to monitor and improve their own learning.

Motivation

As briefly summarized above, good study habits and strategies, among other factors, influence students' grades and contribute to college success. Since study habits vary across many dimensions, what kind of study habits may correspond to better grades? It is of particular interest to quantify, if possible, any relationship between students' grades and their reported study habits. The first objective in the study reported on here was to determine "*To what extent do students' self-reported study practices predict their academic performance, defined in terms of exam grades?"*

Second, although Bayesian statistics seems a promising candidate for such analysis, a vast majority of education research relies on what is termed "frequentists statistics." The use of Bayesian statistics in education research in general and STEM in particular is minimal (Konig and Schoot, 2017 [14]). Therefore, a second objective is to explore the use Bayesian statistics to answer the first objective and, if seeming effective, providing the field a Bayesian framework for such analysis.

Hypothesis

We hypothesize that students' self-reported study habits can be linked to their exam scores when coded as *failure* (less than 70%) vs. *success* (equal to or greater than 70%). More specifically, we hypothesize that students scoring high on study habits conducive and better suited to learning (designated as group 2) vs. those practicing habits not-so conducive to learning (designated as group 1) are more likely to pass the exam.

Theoretical Framework

A. Logistic Regression

In this study we are interested to examine the linkage between students' self-reported study habits (shown in Table 1) and the probability that they obtain 70% or greater score on Exam 1 (E). The total of responses to five prompts about individual study habits determine which of the two groups students are placed into. We label this student response data as *'x.'* The outcome Exam 1 score is converted into a categorical variable, 0 (E score less than or equal to 69, failure) or 1 (E score greater 69, success). Given this outcome, logistic regression is deemed a preferred candidate. For logistic regression analyses, the relative log odds corresponding to both outcomes can be presented as:

$$\ln\left(\frac{\Pr(E\leq 69)}{\Pr(E>69)}\right) = \beta_0 + \beta_1 x \tag{1}$$

Where, β_0 and β_1 are the coefficients to be estimated from the self-reported data *'x'* in G1 and G2. β_0 is the intercept, and β_1 called slope indicates the relative risk of failing versus passing the course.

We can further show that probabilities, $\Pr(E \leq 69)$ or $\Pr(E > 69)$ are given as:

$$\Pr(E \leq 69) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{2}$$

$$\Pr(E > 69) = 1 - \Pr(E \leq 69) = \frac{1}{1+e^{\beta_0+\beta_1 x}} \quad (3)$$

The mathematical model in equation (2) and (3) is used during the data analysis process explained below.

B. Bayesian Approach

A Bayesian approach shows promise in updating the prior degree of belief in an event after considering new data. This updated belief or probability is termed the posterior probability. In our example, we must determine the posterior probability of two parameters β_0 and β_1 (represented in terms of θ), given the observed data '*x*': $p_{post}(\theta|x)$. Assuming we have a prior belief about the probability distribution of the variable θ, $p_{prior}(\theta)$, and the observations '*x*' having a sampling density $p_{samp}(x|\theta)$, then $p_{post}(\theta|x)$ is given by the Bayes' Theorem:

$$p_{post}(\theta|x) = \frac{p_{samp}(x|\theta)}{p(x)} \times p_{prior}(\theta) \quad (4)$$

Where, $p(x)$ is a normalizing constant to make the posterior probability not to exceed 1.

The posterior probability is proportional to the product of likelihood function $\mathcal{L}(x|\theta)$, comprising of the sampling density, and the priors:

$$p_{post}(\theta|x) \propto \mathcal{L}(x|\theta) \times p_{prior}(\theta) \quad (5)$$

The likelihood function $\mathcal{L}(x|\theta)$ is estimated from the observed data *'x'* and the corresponding Exam1 score (E, coded either 0 or 1) as a product of densities, p_i, for each data point since it is independent of each other:

$$\mathcal{L}(x|\theta) = \prod_{i=1}^{N} p_i \,(E,\, n,\, P_r(failure\ \ OR\ \ success)) \quad (6)$$

Where *N* is the total number of participating students, *E* either 0 or 1, *n* is the number of trials (*n* = 1 for present example), and P_r is either equation (2) or (3).

Since there are only two outcomes (0 or 1, failure or success) pertaining to each student, the density p_i for each data point can be estimated using the Binomial distribution:

$$p_i = \frac{n!}{(n-E)!E!} P_r^E (1 - P_r)^{n-E} \quad (7)$$

If we know some prior distribution of the parameter θ, it can be used to generate values for β_0 and β_1. In this case since we don't have any prior knowledge about the distribution of the parameters β_0 and β_1, we model θ as normally distributed with mean 0 and the standard deviation σ: $\theta \sim N(0, \sigma)$. On average, these priors will yield $\Pr(E \leq 69) = \Pr(E > 69) = 0.5$, i.e., equally likely, indicating self-reported study habits are equally likely to predict both outcomes.

Methods

A. Participants

The participants in this study were students enrolled in one of four classes, Electrical Engineering (EE)351, EE310, Engineering Technology (ET)240, and Engineering (ENGR)100, spread over 1, 7, 3, or 1 semester(s), respectively, for each class. A total of 281 students consented to participate in the study.

B. Data Collection Instrument and Data Preparation

Modifying information from McGuire [13], we call our survey instrument shown in Table 1 the "Self-Evaluation Study Strategy Instrument" (SESSI). This survey was administered to each class at the start of the semester just after Exam 1, about 6 weeks into the semester. Each survey item (10 in all) was Likert scale coded in accordance with the following statements: *1=Strongly disagree with the statement; 2 = Disagree with the statement; 3 = Neither agree nor disagree with the statement; 4 = Agree with the statement; 5 = Strongly agree with the statement.*

Table 1: Self-Evaluation Study Strategy Instrument (SESSI) Prompts

Group 1 (G1)	**Your response**	**Group 2 (G2)**	**Your response**
I did not spend enough time on the material		I did preview-review for every class	
I started the homework too late		I did a little of the homework at a time	
I didn't memorize the needed information		I made flashcards to prepare for the exam	
I did not use the book		I used the book and did the suggested problems	
I assumed I understood information that I had read and re-read but not applied		I practiced explaining the information to others	
Total for G1, ranging from 5 to 25	*x*	Total for G2, ranging from 5 to 25	*x*

Modified from McGuire [13] Table 11.8, p. 147.

The data matrix *'x'* consisted of two columns pertaining to total G1 and G2 scores, and N=281 rows, one for each student. Similarly, the outcome vector consisted of Exam 1 scores, converted to either 0 (less than 70) or 1 otherwise. Table 2 exemplifies a few rows of the data to highlight the structure and coding of the data in accordance with the procedure explained above.

Table 2: Few rows of data structure and coding

Student #	**Total Score, *'x', raw and normalized***				**Exam 1 (*E*)**	
	G1 score		**G2 score**			
	Raw	*Normalized*	*Raw*	*Normalized*	*Raw*	*Coded*
12	9	-1.1108	6	-2.1207	65	0
13	16	0.7315	11	-0.6958	27	0
14	11	-0.5844	11	-0.6958	100	1
15	13	-0.0581	13	-0.1258	83	1

C. Data Analysis and Results

We analyzed the data in Matlab since it is a standard if not default computing language in most engineering classes. After uploading the data in Matlab, the *'x'* data is centered and normalized, and the Exam1 data is coded as shown in Table 2 and 3. The Matlab program is modeled in accordance with the example shown in Matlab documentation, [15]. In Bayesian data analysis, Monte Carlo methods are often used for summarizing various parameters of the posterior distribution, $p_{post}(\theta|x)$. Since many posterior distributions cannot be computed analytically,

generating random samples from the underlying distribution can be sufficient to compute various posterior statistics, such as mean and median. To this end, Matlab provides a "slicesample" command to generate a Markovian sequence having stationary distribution equivalent to the target or underlying distribution. In our example, posterior distribution in equation (5) is the target distribution. (Various parameters in "slicesample" are explained in Matlab documentation [16]).

Tabulated in Table 3 are some simple statistics pertaining to both *'x'* and *'E'* data. We see that both G1 and G2 means are almost identical, indicating no preference for G1 or G2.

Table 3: Some statistics pertaining to the data

Data Type	***Mean*, μ**	***Std*, σ**	***Min***	***Max***	***Normalizing/ Coding***
G1	13.2206	3.7996	5	24	$x_{norm} = \frac{x - \mu}{\sigma}$
G2	13.4413	3.5088	4	25	
E	73.3676	17.7636	12	100	Either 0 or 1

The Matlab program is executed for 5000 iterations for both G1 and G2. Tabulated in Table 4 and shown in Figures 1 – 3 are the salient output of the program.

Table 4: Estimated statistics for intercept $\widehat{\boldsymbol{\beta_0}}$ and slope $\widehat{\boldsymbol{\beta_1}}$

Data Type	***Estimated Mean***		***Estimated Std***		***95% Credible Interval - CI95 range***	
	$\widehat{\boldsymbol{\beta_0}}$	$\widehat{\boldsymbol{\beta_1}}$	σ_{β_0}	σ_{β_1}	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$
G1	-0.6042	0.7306	0.1269	0.1373	-0.8530 to -0.3554	0.4616 to 0.9997
G2	-0.5548	-0.3085	0.1223	0.1282	-0.7945 to -0.3152	-0.5597 to -0.0573

Figure 1 depicts how the mean values evolved with number of samples for both intercept and the slope. These attain steady state values after about 100 iterations. The steady state mean values are tabulated in Table 4.

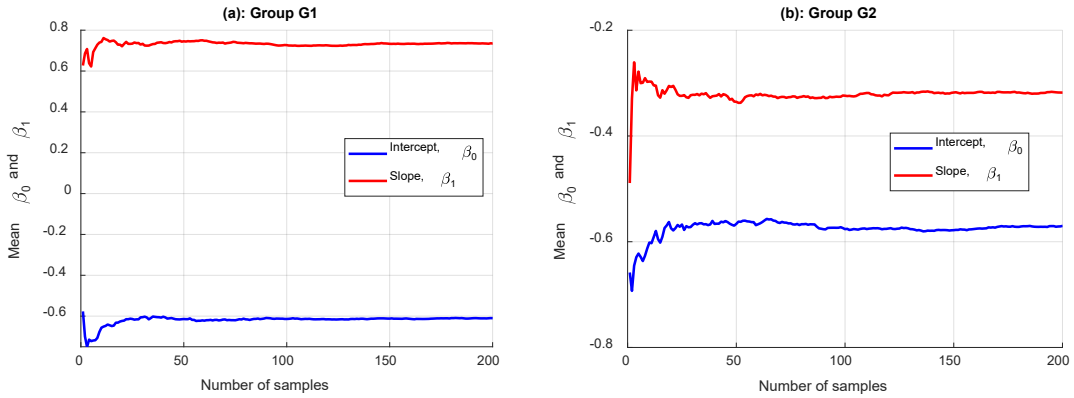

Figure 1: Evolution of the mean β_0 and β_1 for both groups, (a) – G1, (b) – G2, indicating transients and the steady states after about 100 iterations. Final values are given in Table 4.

Since both intercept β_0 and slope β_1 are statistical parameters, their estimated probability density functions (pdf) are shown in Figure 2 for both groups, G1 and G2. These estimated pdfs are almost identical to the normal pdf $N(\mu, \sigma)$ with parameters μ and σ. These pdfs are also plotted

in Figure 1, using the estimated means and the standard deviations shown in table 4 for both groups.

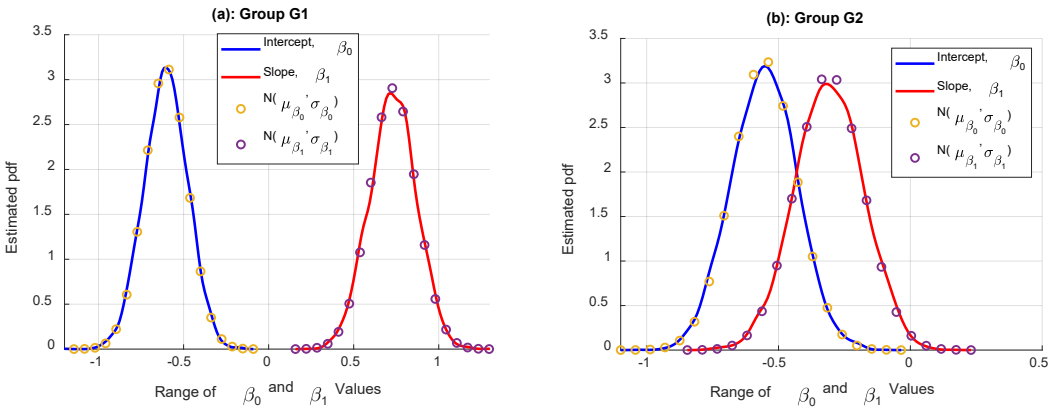

Figure 2: Estimated probability density functions (pdf) for intercept and slope parameters pertaining to both groups (a) – G1 and (b) – G2. Also shown are normal pdfs, $N(\widehat{\beta_0}, \sigma_{\beta_0})$ and $N(\widehat{\beta_1}, \sigma_{\beta_1})$ for both intercept and the slope. Values for the variables are depicted in Table 4.

The SESSI prompts under G1, Table 1 are identified as not conducive to learning vs. the prompts in G2. Although there could be many such prompts or variants thereof, the reported prompts in Table 1 are simple and a few to consider, and based on results from McGuire's work [14], the power of these simple prompts with regard to students' success was expected. Our study provides statistical analysis pertaining to these prompts linking them with students' grades (pass/fail in this example).

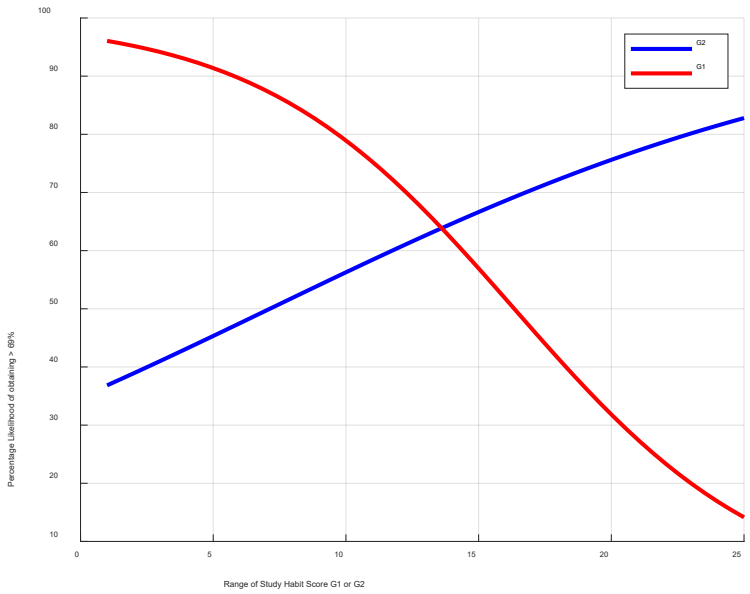

Figure 3: Plots of equation (3) showing the percentage likelihood of obtaining greater than 69% on Exam 1 vs. self-reported score on G1 and G2.

Figure 3 shows the results of equation (3) using the estimated values for both intercept β_0 and the slope β_1 for both groups. It depicts the percentage likelihood of obtaining greater than 69% on Exam 1 score vs. the individual's self-reported study habit scores under G1 and G2. The probabilistic distributions underpinning the logistic regression coefficients indicate these values being confined to a narrow range. The 95% Credible Interval (*CI 95%*) provides 95% probability that regression parameters will fall within this range. This range points in the direction of varying students' G1 and/or G2 cores, indicating a broader range around 14 points. Further, the program converged to the final values rapidly within a few hundred iterations, indicating suitability of Bayesian approach for such problems.

The results in Figure 3 indicate that a student is strongly likely on average to have a passing grade if he/she has self-reported a score of 14 or better under group G2 and less than 14 under group G1, the lower the better. The outcome in Figure 3 is a positive indicator in favor of urging students to consider adopting habits outlined under G2 and avoid the ones in G1.

Conclusions and Future Work

The focus of our effort reported here was twofold: identify (i) whether students' self-reported study habits can be linked to their Exam1 scores, and (ii) whether Bayesian analysis can productively be used to provide statistical analysis underpinning any links. This paper presents findings on both aspects utilizing the data from 4 classes, spread over 12 semesters. Based on the data analyses and the discussion above, we conclude that there is statistically significant positive correlation and linkage between the students' study habit scores and their passing or failing the Exam1. On average, a student is strongly likely to have a passing grade if he/she has self-reported a score of 14 or better under group G2 and less than 14 under group G1, the lower the better. Further, the prompts in Table 1 are few, which are likely to be adopted by students with relative ease. The first author shared aggregated results with students in each semester, resulting in overall improved passing grades.

This work is not complete yet. We intend to compare the statistical findings using Bayesian approach with those of using the Frequentists' approach. After completing those analyses, we will report our findings to make further contributions to the field in this area.

Acknowledgements

This material is based upon work partly supported by the National Science Foundation (NSF) under grant No. 1612445. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of NSF.

References

[1] Walck-Shannon, E. M., Rowelli, S. F., and Freys, R. F., "To What Extent Do Study Habits Relate to Performance?" CBE—Life Sciences Education, 20:ar6, 1–15, Spring 2021

[2] Kaur, J., and Singh. P., "Study Habits And Academic Performance: A Comparative Analysis," European Journal of Molecular & Clinical Medicine, Volume 07, Issue 07, 2020

[3] Veenstra, C.P, Dey, E.L., and Herrin, G. D., "A model for freshmen engineering retention", Advances in Engineering Education, ASEE, Winter 2009, pp. 1-33.

[4] Scalise, a., M. Besterfield-sacre, L. shuman, and H. Wolfe (2000), "First term Probation: Models for identifying High risk students." in 2000 Proceedings of 30th ASEE/IEE Frontiers in Education Conference, Session F1F. [Online]. Available: http://fie-conference.org/fie2000/papers/1276.pdf

[5] Besterfield-sacre, M., C.J. atman, and L.J. shuman. (1997), "Characteristics of Freshman engineering students: Models for determining student attrition in engineering." Journal of Engineering Education 86, no. 2 (1997): 139–49. [Online]. Available: http://www.asee.org/publications/jee/PaPers/display.cfm?pdf=54.pdf

[6] Shuman, L., M. Besterfield-sacre, d. Budny, d, s. Larpkiattaworn, O. Muogboh, s. Provezis, and H. Wolfe. (2003), "What do we know about our entering students and how does it impact upon performance?" Proceedings of the 2003 American Society for Engineering Education Annual Conference and Exposition, Session 3553.

[7] Blumner, H. N, and Richards, H. C. (1997), "Study Habits and Academic Achievement of Engineering students," Journal of Engineering Education, 86(2), pp. 125-132.

[8] Dey, Chandana (2014), "Effect of Study Habit on Academic Achievement," International Journal of Research in Humanities and Social Sciences, Vol. 2, Issue 5, June 2014

[9] Bandura, A. (1993), "Perceived Self-efficacy in Cognitive Development and Functioning," Educational Psychologist, 28(2), 117-148, 1993.

[10] Dweck, C., "Mindset: The New Pschology of Success," Ballantine Books, 2006.

[11] G. E. Matt, B. Pechersky, and C. Cervantes, " High School Study Habits and Early College Achievemnets," Psychological Reports, 1991, 69, pp. 91-96.

[12] M. Balduf, " Underachievemnt Among College Students," Journal of Advanced Academics, Vol. 20, n. 2, 2009, pp. 274-294.

[13] McGuire, S. Y, and McGuire, S., "Teach Students How to Learn," Stylus Publishing, 2015.

[14] Konig, C. and Schoot, R. V. D., "Bayesian statistics in educational research: a look at the current state of affairs," Educational Review, 70(4), July 2017

[15] Matlab Documentation, [Online]. Available: https://www.mathworks.com/help/stats/bayesian-analysis-for-a-logistic-regression-model.html

[16] Matlab Documentation, [Online]. Available: https://www.mathworks.com/help/stats/slicesample.html