Bayesian Statistics - Final Project

Keith Sheridan

Due: 8/2/2023

Introduction

Calculus is often regarded as a challenging subject, demanding rigorous analytical thinking and an aptitude for problem-solving. While classes are typically inhabited by those seeking a career in mathematics and/or science, Calculus offers a broad range of advantages to all who enroll, including but not limited to: engaging with practical applications, enhancing problem-solving skills, boosting college acceptance probability, widening the scope of career opportunities, and stimulating intellectual growth. As a result, the decision to enroll in a Calculus class while in high school should be well-informed for all interested students.

The main objective of this analysis is to be able to predict future Calculus grades from a set of five explanatory variables, namely, Math PSAT score, GPA, Precalculus grade, Precalculus teacher, and level of Calculus. In addition, two minor objectives will be explored: (1) Does the Precalculus teacher affect the student's performance in Calculus, as measured by their final course grade? and (2) which numerical predictor is most important for determining a student's Calculus grade? Overall, this analysis aims to examine the nuances of calculus grades, exploring factors that may impact student performance, identifying patterns, and generalizing potential outcomes.

Statistical Summaries

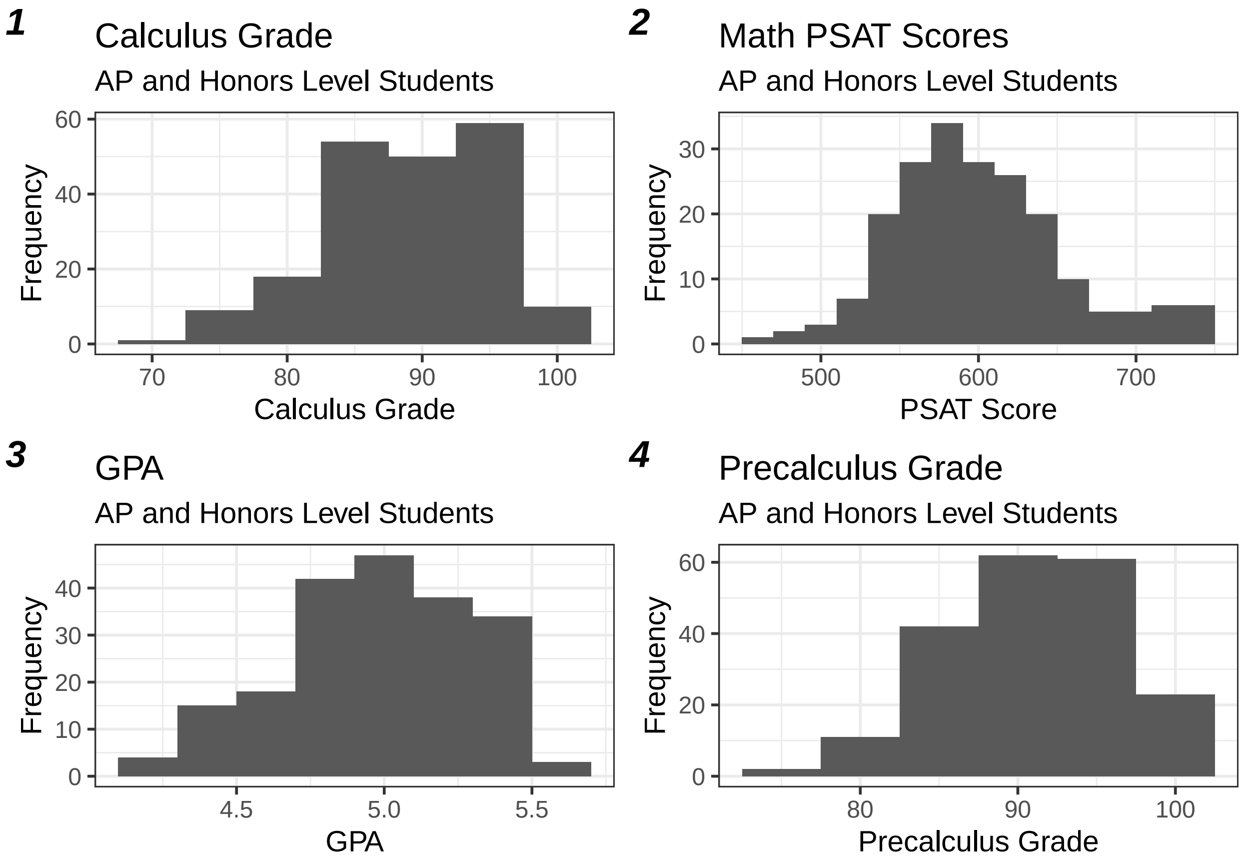
For this analysis, the response variable will be the score obtained in a Calculus course and the predictors of interest will be: Math PSAT score, GPA, Precalculus grade, Precalculus teacher, and level of Calculus (Honors or Advanced Placement). The following will provide statistical summaries and graphics of all relevant variables.

Note: 34 observations were dropped due to missing Precalculus grades and/or Math PSAT scores. The values are missing for the students who transferred to our school for their senior year, enrolled in Precalculus at a separate institution, or completed Precalculus at a level lower than Honors. As this analysis is intended for those enrolled in the Honors level of Precalculus, those enrolled in lower levels would not be considered. Lastly, while those who are missing either a Precalculus grade or Math PSAT score (18 observations total) should be investigated, those concerns will not be addressed in this paper.

Table 1: Frequency counts for teacher and Calculus level (1=Honors, 2=AP), respectively.

Teacher	Count
1	96
2	14
3	16
4	72
5	3

Level	Count
1	127
2	74



(1) Plot 1 above displays a histogram of the Calculus grades (the response variable). The data is centered at approximately 90 and skewed slightly left with no visible outliers.

(2) Plot 2 above displays a histogram of the Math PSAT scores. The data is centered at approximately 600 and skewed slightly right with no visible outliers. The reason for the unique shape in the upper half of the data comes from the scaled scores of the PSAT, which vary from test to test. The raw score of the math section of the PSAT ranges from 0 - 48, with the scaled scores ranging from 160 - 760. Each 1 unit decrease in raw score can correspond to a 0, 10, or 20 point decrease in scaled score.

(3) Plot 3 above displays a histogram of GPA. The data is centered at approximately 5 and skewed slightly left with no visible outliers.

(4) Plot 4 above displays a histogram of Precalculus grades. The data is centered at approximately 91 and skewed slightly left with no visible outliers.

While graphics for the categorical variables will not be shown. Please see frequency Table 1 above for a breakdown of the counts.

Initial Analysis - Multiple Linear Regression

I began my analysis by fitting a multiple linear regression model with all predictors:

$y = \beta_0 + \sum_{i=1}^{9} \beta_i x_i + \epsilon$ where $\epsilon \overset{iid}{\sim} N(0, \sigma^2)$, with

- y = Calculus Grade
- x_1 = Math PSAT z-score, x_2 = GPA z-score, x_3 = Precalculus grade z-score,
- $x_4 = 1$ (if Teacher 1, 0 otherwise), $x_5 = 1$ (if Teacher 2, 0 otherwise), $x_6 = 1$ (if Teacher 3, 0 otherwise)
- $x_7 = 1$ (if Teacher 4, 0 otherwise) , $x_8 = 1$ (if Teacher 5, 0 otherwise), $x_9 = 1$ (if AP-level, 0 otherwise).

The purpose of the least squares model was to obtain starting values for the parameters in the Markov chain of the Bayesian analysis. Output from the model is shown in the table below.

Table 2: Output from the Least Squares Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3328053	0.0804466	4.1369729	0.0000526
scale(psat)	0.2723235	0.0583592	4.6663324	0.0000057
scale(gpa)	0.3169231	0.0729226	4.3460194	0.0000225
scale(grade_htpc)	0.4601823	0.0852961	5.3951151	0.0000002
ap_yn2	-0.4460061	0.1223291	-3.6459535	0.0003430
teach_num2	-0.3688017	0.1942968	-1.8981352	0.0591784
teach_num3	-1.0755548	0.2166771	-4.9638608	0.0000015
teach_num4	-0.1639619	0.1122710	-1.4604118	0.1458113
teach_num5	0.0959775	0.4030851	0.2381073	0.8120518

Model Assumptions

Linearity and Equal Variance

The residuals versus fitted plot (not shown) appears problematic for the equal variance assumption of multiple linear regression. From the plot, it appears possible the variance of the residuals for the upper end of the fitted values spectrum is different from the rest of the data. However, the residual values seem to be centered at zero, suggesting a linear relationship.

Breusch-Pagan Test I decided to perform the Breusch-Pagan Test to determine if heteroscedasticity is present. The hypotheses of the test are

H_0 : Heteroscedasticity is not present vs. H_a : Heteroscedasticity is present.

Since $p = 0.030 < 0.05$, we reject H_0 at the 0.05 level. There is evidence to show heteroscedasticity is present.

Normality

The Normal Q-Q plot (not shown) shows some concerning behavior in the tails of the distribution. It is possible the error terms are not normally distributed.

Shapiro-Wilk Test I decided to perform the Shapiro-Wilk Test to determine if the residual values are normally distributed. The hypotheses of the test are

H_0 : Residuals are normal vs. H_a : Residuals are non-normal.

Since $p = 0.017 < 0.05$, we reject H_0 at the 0.05 level. There is evidence the residual values are not normally distributed.

Independence

A plot of the residuals vs. order (not shown) was constructed. The variance of the residual values seems lower for approximately the first 50 students. I believe it is worth noting the first 50 students graduated in the year 2020. As a result, they had a Precalculus experience which was unaffected by the COVID-19 pandemic. Moreover, the nation-wide shutdown (which spanned from March 2020 until the end of the school year) may have contributed to the lesser variance.

Bayesian Analysis

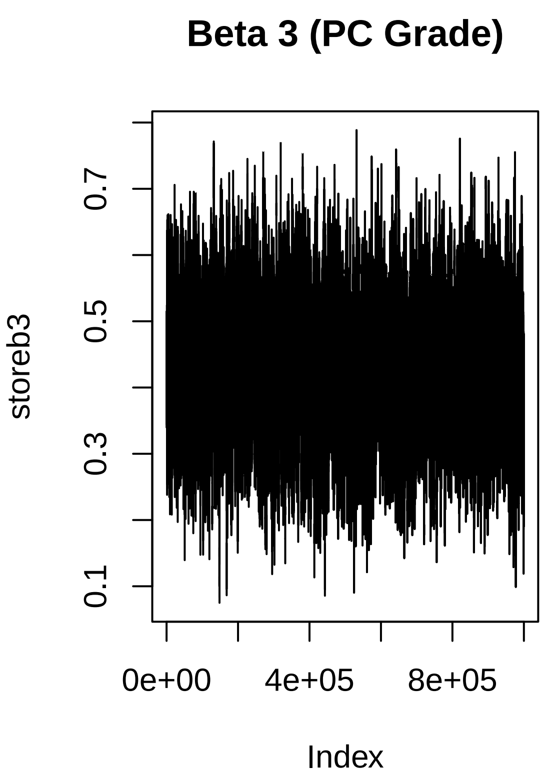
Prior: $\beta_i \sim N(0,1)$ for $i \in \{0,1,2,3\}$, $\beta_j \sim N(0,\sigma_t^2)$ for $j \in \{4,5,6,7,8\}$, $\beta_9 \sim N(-1,1), \sigma \sim Exp(1)$, and $\sigma_t \sim Exp(1)$, where all parameters are mutually independent.

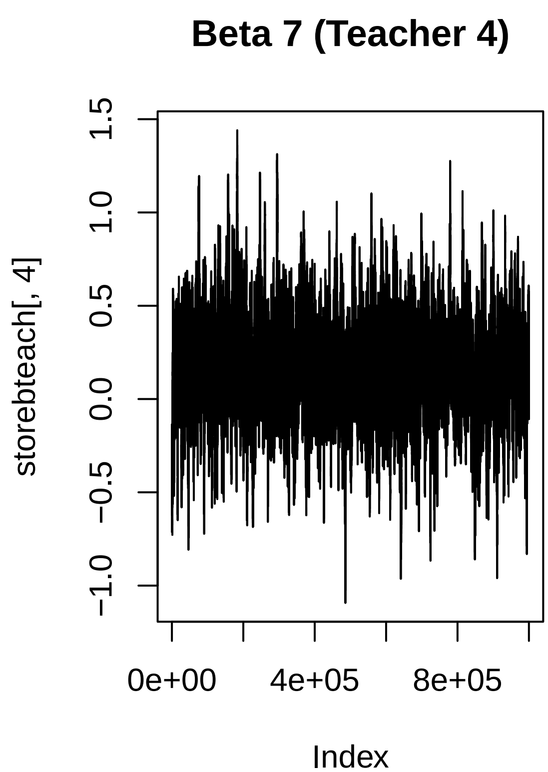
Likelihood: $y \sim N\left(\beta_0 + \sum_{i=1}^{9} \beta_i x_i, \sigma^2\right)$.

Results

Time Series Plots and Effective Sample Sizes

The time series plots below are the results of the Markov chain for the parameters β_3 and β_7. In the interest of layout and brevity, effective sample sizes will be reported instead of all time series plots.





	Beta 0	Beta 1	Beta 2	Beta 3	Beta 4	Beta 5
Effective Sample Size	803	6568	4323	3087	966	1254

	Beta 6	Beta 7	Beta 8	Beta 9	Sigma	Sigma_t
Effective Sample Size	1145	997	1009	2162	15017	3788

Point Estimates and Credible Intervals

The table below gives a point estimate for each parameter as well as a 95% equal-tailed credible interval. We can see that none of the credible intervals for β_1, β_2, and β_3 contain zero, providing evidence that each of these associated predictors is beneficial in the model. Moreover, we see the interval for β_6 (Teacher 3) does not contain zero rather only negative values, providing evidence that students are expected to have lower Calculus grades if taught Precalculus by Teacher 3.

Table 5: Estimates and Credible Intervals

Parameter	Point Estimate	95% CI (LL)	95% CI (UL)
Beta 0 (Intercept)	0.0132	-0.4980	0.5269
Beta 1 (PSAT)	0.2897	0.1767	0.4046
Beta 2 (GPA)	0.3306	0.1858	0.4757
Beta 3 (PC Grade)	0.4249	0.2505	0.5935
Beta 4 (Teacher 1)	0.3026	-0.2144	0.8334
Beta 5 (Teacher 2)	-0.0408	-0.6034	0.5150
Beta 6 (Teacher 3)	-0.6046	-1.2709	-0.0714
Beta 7 (Teacher 4)	0.1578	-0.3722	0.6683
Beta 8 (Teacher 5)	0.2269	-0.4414	1.0223
Beta 9 (Level)	-0.4520	-0.6881	-0.2088
Sigma	0.6741	0.6108	0.7467
Sigma (Teacher)	0.5196	0.1933	1.2077

Hypothetical Student Predictions

The table below displays the predictions of outcomes from three hypothetical students. Specifically, their data is as follows:

(1) Student 1: Math PSAT Score = 680, GPA = 5.3, Precalculus Grade = 98, Teacher = 4, Level of Calculus = 1 (AP)
(2) Student 2: Math PSAT Score = 600, GPA = 5.0, Precalculus Grade = 90, Teacher = 4, Level of Calculus = 1 (AP)
(3) Student 3: Math PSAT Score = 560, GPA = 4.7, Precalculus Grade = 88, Teacher = 4, Level of Calculus = 1 (AP)

The data above was chosen to simulate an above-average, average, and below-average student, respectively. As there is typically more concern about enrolling in the AP level of Calculus, I wanted to check the model performance on students of varying skill level with the same teacher. The prediction function calculates a predicted calculus grade, constructs a 95% equal-tailed credible interval, and determines the probability of achieving each possible letter grade.

Table 6: Hypothetical Predictions for Calculus Grade

Student	Calculus Grade	95% CI (LL)	95% CI (UL)	Probability A	Probability B	Probability C	Probability F
1	95	87	103	0.886	0.113	0.000	0.000
2	87	78	95	0.207	0.732	0.061	0.000
3	82	74	91	0.035	0.668	0.295	0.002

Teacher "Effectiveness"

The table below displays the rank of each teacher by overall proportion of highest beta value. For the teacher betas, higher values are associated with higher Calculus grades. For each run of the simulation (1,000,000 runs), the teacher with the largest beta value was recorded. Afterwards, we find the sum for each teacher and divide by the total number of runs to arrive at the proportions given in the table.

Table 7: Rank of Teacher by Beta Proportion

Teacher	Highest Beta Proportion
1	0.543
5	0.381
4	0.059
2	0.017
3	0.000

In addition to ranking the teachers, I created a table (below) of pairwise comparisons. By doing so, each teacher is able to be compared with all other teachers. The value in each cell of the table is the probability of the row teacher producing better Calculus grades than the column teacher. For example, the value in row 1, column 2 is 0.970. This means that in 97% of the runs of the simulation the beta value for Teacher 1 was larger than that of Teacher 2. Conversely, in row 2, column 1, the value 0.030 indicates that in 3% of the runs of the simulation the beta value for Teacher 2 was larger than that of Teacher 1.

Table 8: Pairwise Comparison of 'Teacher Effectiveness'

	Teacher 1	Teacher 2	Teacher 3	Teacher 4	Teacher 5
Teacher 1	NA	0.970	1.000	0.902	0.610
Teacher 2	0.030	NA	0.987	0.145	0.222
Teacher 3	0.000	0.013	NA	0.000	0.014
Teacher 4	0.098	0.855	1.000	NA	0.433
Teacher 5	0.390	0.778	0.986	0.567	NA

Predictor Importance

The table below displays the rank of each numerical predictor by overall proportion of highest beta value. For each run of the simulation (1,000,000 runs), the largest numeric beta value was recorded. Afterwards, we find the sum for each beta and divide by the total number of runs to arrive at the proportions given in the table. Since the beta values are indicating the expected increase in Calculus grade z-score per unit increase in the respective predictor, given all other variables in the model, larger beta values are associated with larger Calculus grades.

Table 9: Rank of Predictor by Beta Proportion

Predictor	Highest Beta Proportion
Precalculus Grade	0.719
GPA	0.240
PSAT	0.041