Final Project

Keith Sheridan

Due: May 12, 2023

Introduction

The Advanced Placement (AP) Calculus AB exam (herein, the AP exam) is a comprehensive test administered by the College Board to all students enrolled in AP Calculus AB who opt to complete the examination. The scores range from 1-5 on a discrete scale (see Table 1 below). The purpose of the exam is to measure the mastery of the procedural and conceptual learning objectives of the course. In addition to student achievement, these scores are one measure of teaching efficacy for AP teachers.

The purpose of this analysis is to explore the relationship between obtaining a specified score on the AP Calculus AB exam and certain predictors (to be discussed). Specifically, I will attempt to build a model for predicting AP scores based on the aforementioned predictors. As a result, I will also be able to assess my secondary goal of checking the compatibility between a grade earned in the course with the score earned on the AP exam. For example, if a student receives a 95 in the course, they should receive a 5 on the exam. By examining this relationship, I will be able to determine, in part, the efficacy of my teaching.

Table 1: AP Score Scale Table

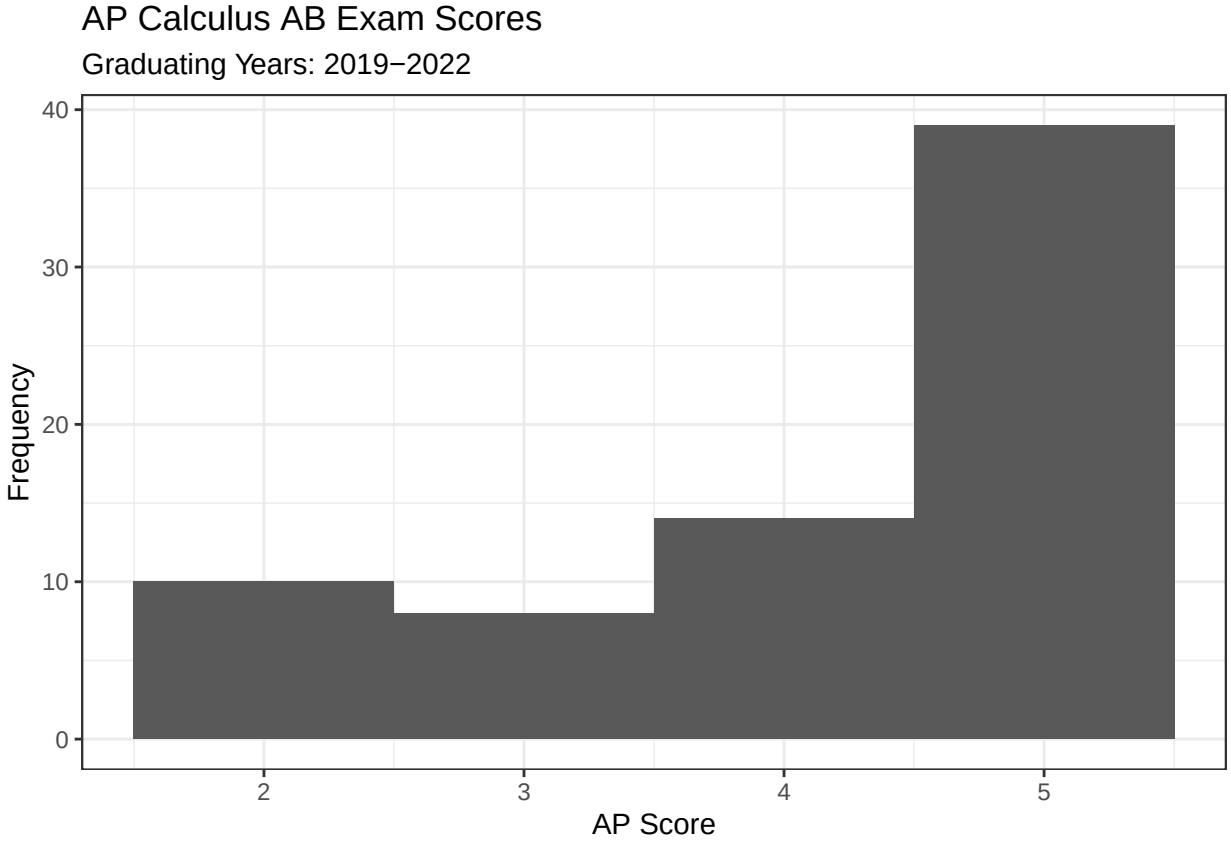
AP Exam Score	Recommendation	College Course Grade Equivalent
5	Extremely Well Qualified	A+ or A
4	Very Well Qualified	A-, B+, or B
3	Qualified	B-, C+, or C
2	Possibly Qualified	—
1	No Recommendation	—

Statistical Summaries

For this analysis, the response variable will be the score obtained on the AP exam and the predictors of interest will be: final course grade, PSAT scores, gender, relative GPA, and relative rank. The following will provide statistical summaries and graphics of all relevant variables. Note: 5 observations were dropped due to missing PSAT scores. The values are missing for the students who transferred to our school for their senior year. Thus, I did not have access to their PSAT scores (which occur during junior year).

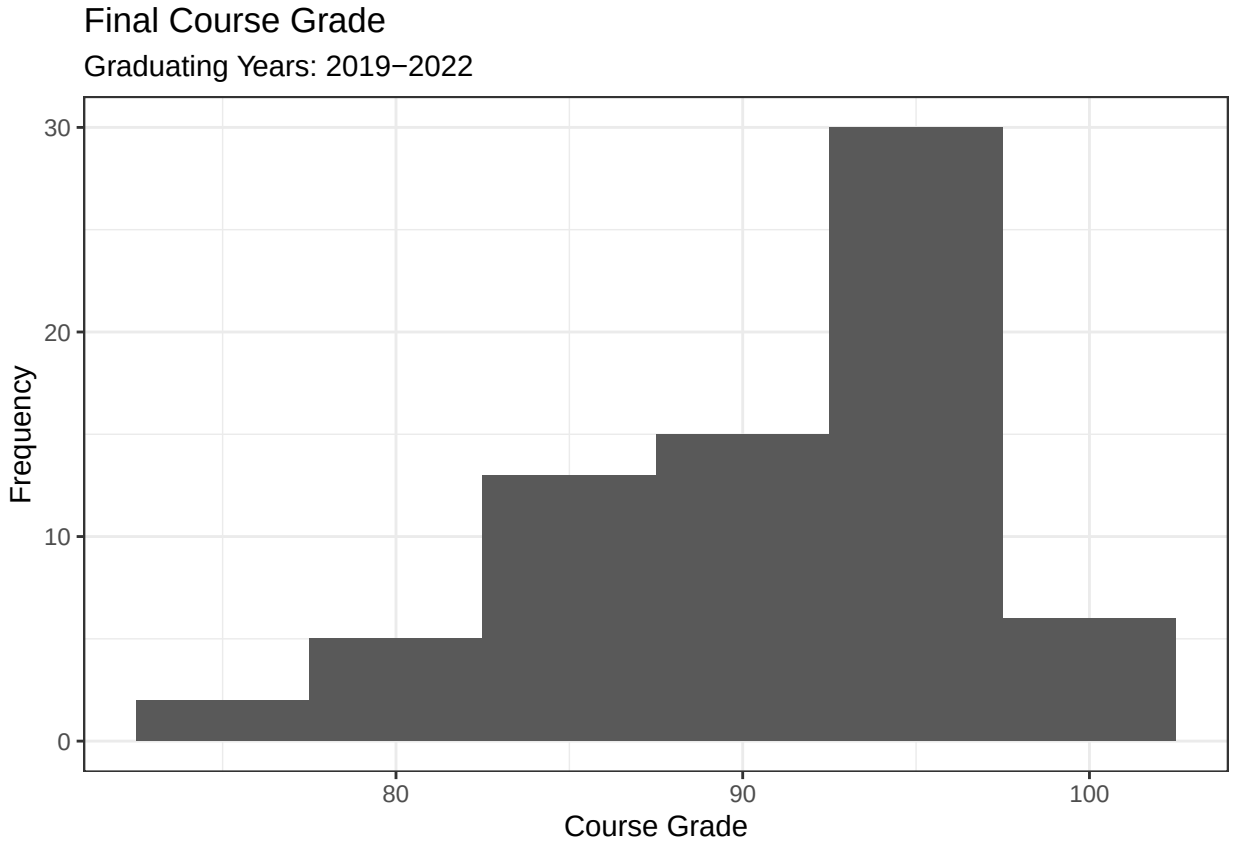
AP Calculus AB Exam Scores (Response Variable)

The plot below displays a histogram of the AP exam scores. The data is centered at approximately 5 and skewed slightly left with no visible outliers.



Final Course Grade (Predictor Variable)

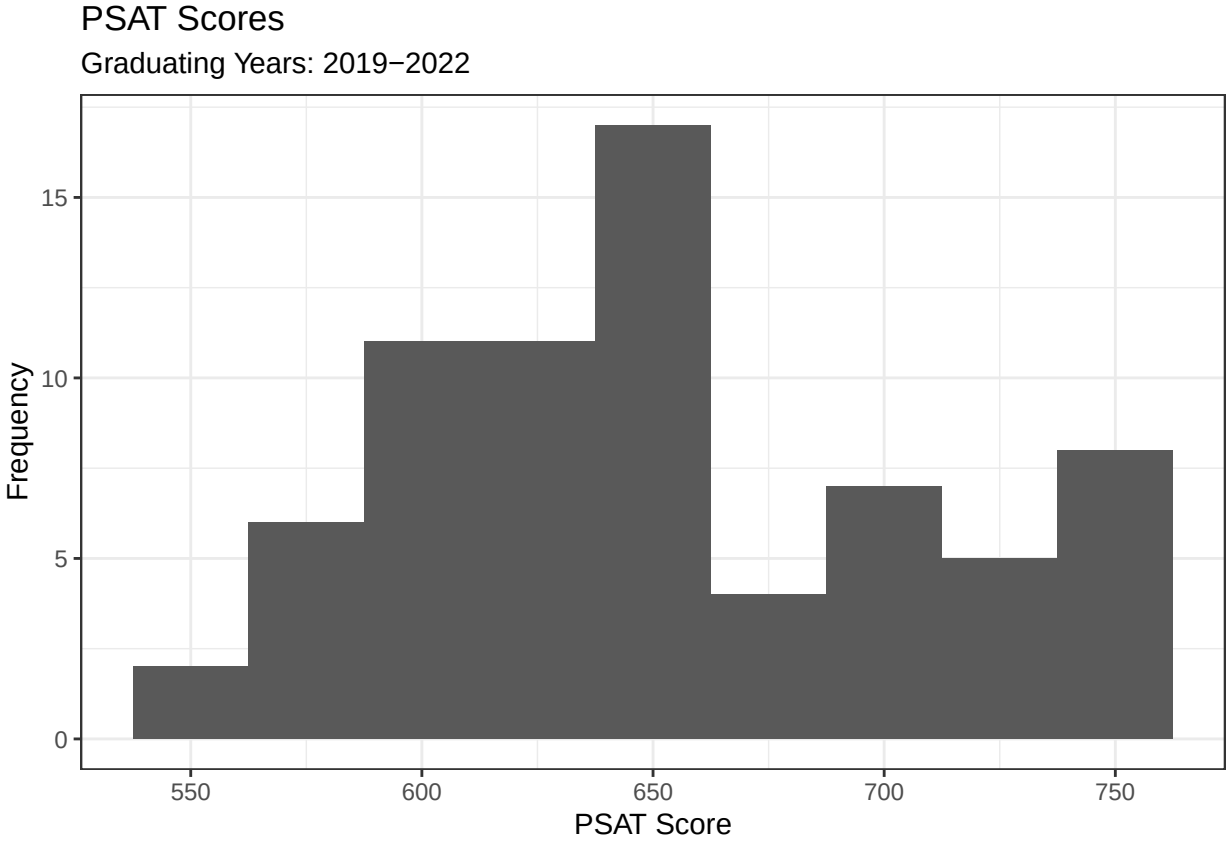
The plot below displays a histogram of the final course grades. The data is centered at approximately 93 and skewed slightly left with no visible outliers.



PSAT Scores (Predictor Variable)

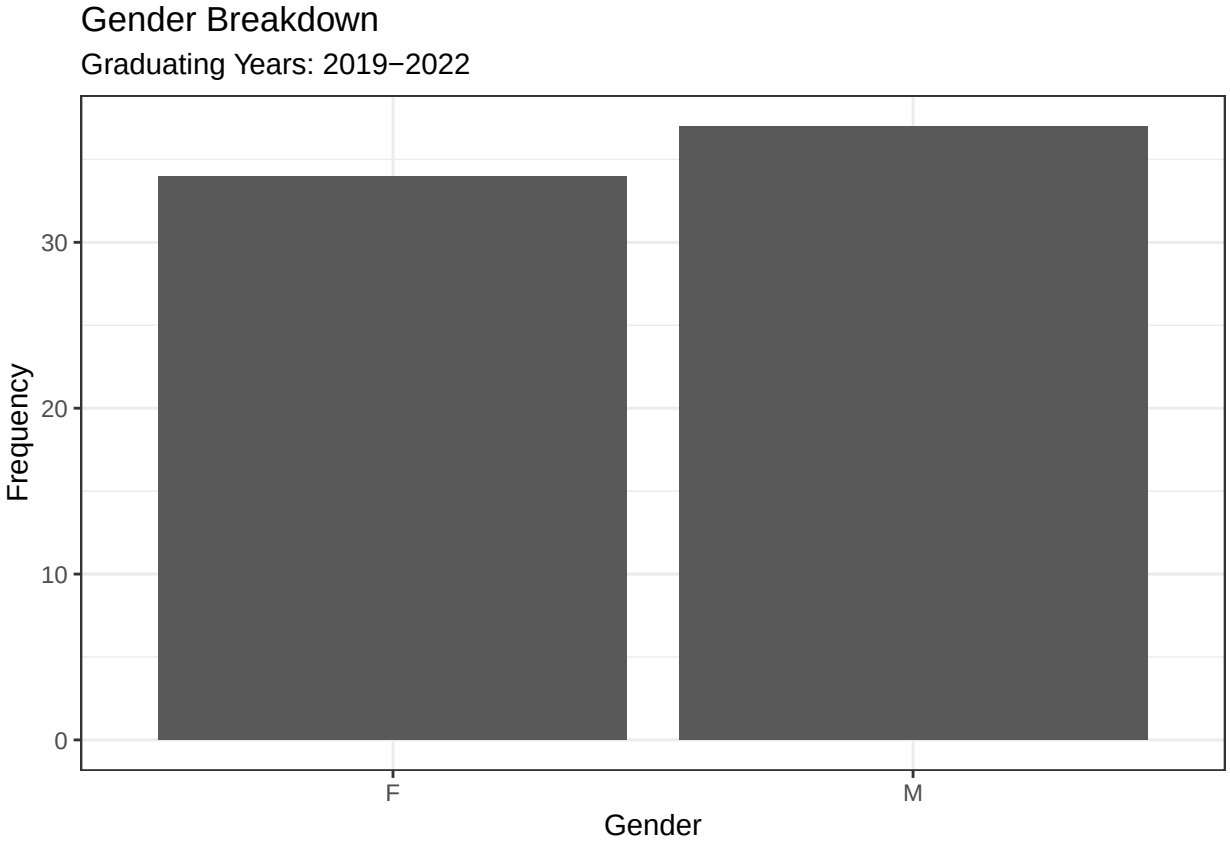
The plot below displays a histogram of the PSAT scores. The data is centered at approximately 650 and skewed slightly right with no visible outliers. The reason for the unique shape in the upper half of the data comes from the scaled scores of the PSAT, which vary from test to test. The raw score of 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.

For example, for the exam administered on October 13, 2021, for raw scores 43 - 48, each unit decrease resulted in a 10 point reduction in scaled score (i.e. 710 - 760). However, for raw scores 40 - 42, each unit decrease resulted in a 20 point reduction in scaled score (i.e. 650 - 690). As a result, the scores of 660 and 680 were impossible to achieve. This phenomena typically occurs at the lower and higher end of the conversion scale and varies from test to test.



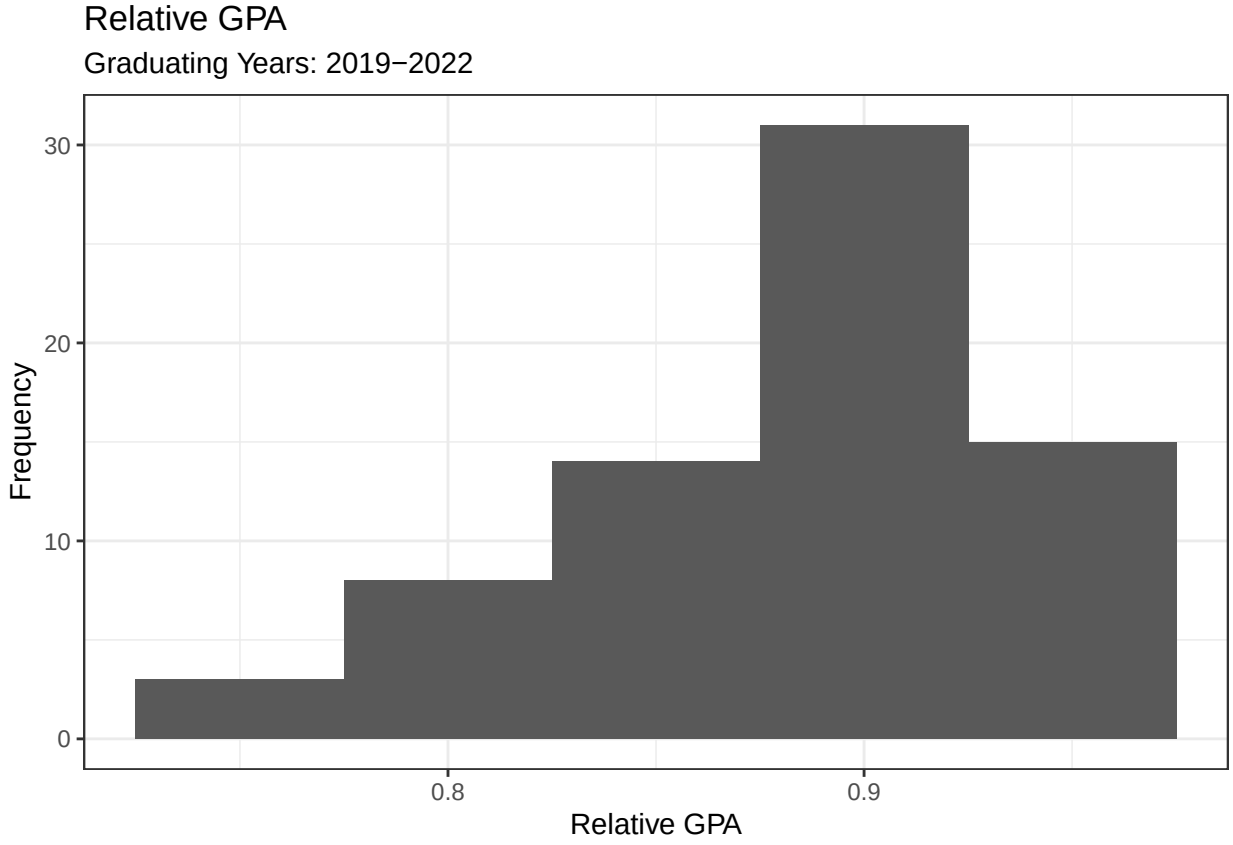
Gender (Predictor Variable)

The plot below displays a bar chart of gender. We can see there is a good balance between genders in the AP course.



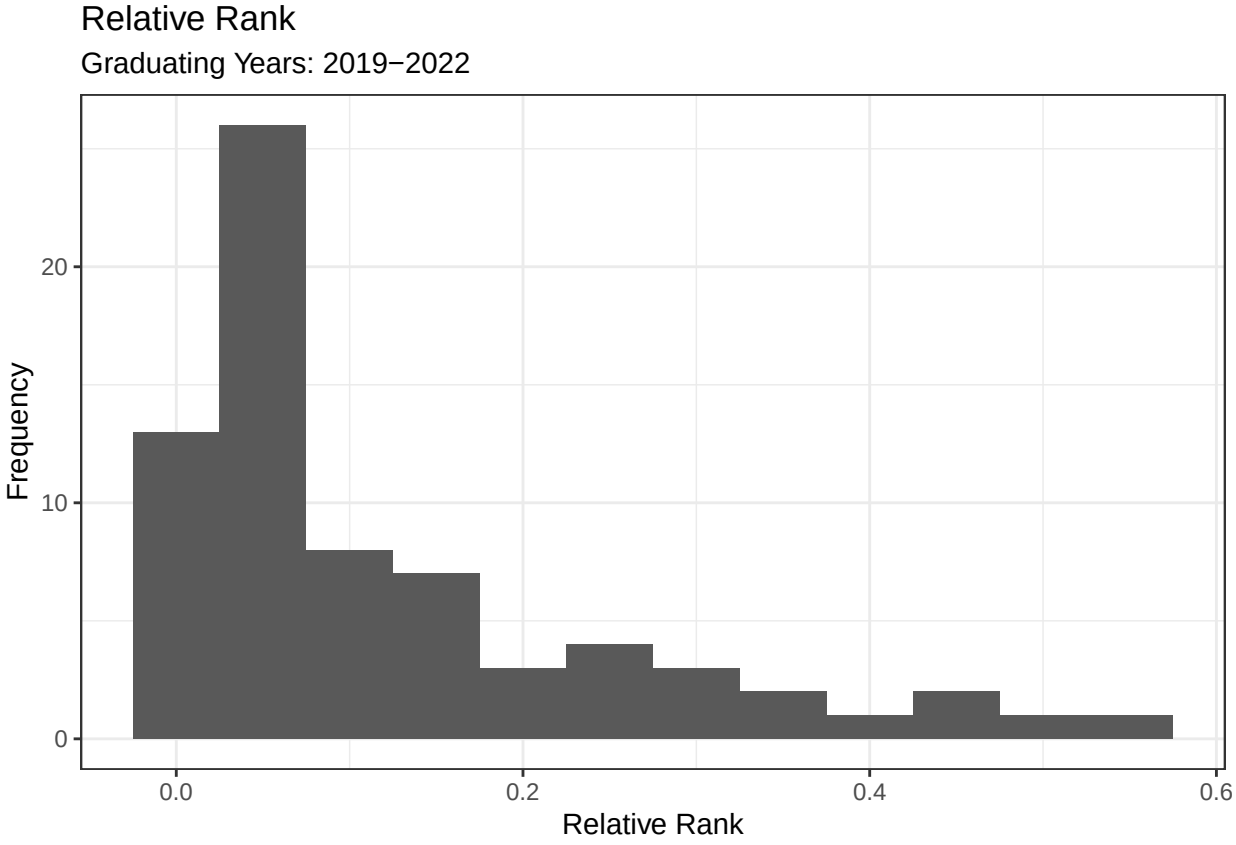
Relative GPA (Predictor Variable)

The plot below displays a histogram of relative GPA. The data is centered at approximately 0.9 and skewed slightly left with no visible outliers. I computed a "relative GPA" because the GPA scale changed after the 2018-2019 school year. Since all maximum GPAs were not the same, I decided the "relative" metric would be appropriate. Discussion regarding this change will occur later in the analysis.



Relative Rank (Predictor Variable)

The plot below displays a histogram of relative rank. The data is centered at approximately 0.06 and skewed significantly right with potential for outliers. I computed a "relative rank" because student rank is subject to class size. On average, a rank of 1 out of 300 is more significant than 1 out of 10. Since the class sizes vary from year to year, I decided the relative metric would be appropriate.



Initial Analysis – Multiple Linear Regression

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

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 I(male) + \beta_4 x_4 + \beta_5 x_5 + \epsilon$ where $\epsilon \overset{iid}{\sim} N(0, \sigma^2)$ and

y = AP exam score, x_1 = course grade, x_2 = PSAT score, $x_3 = 1$ (if male, 0 otherwise), x_4 = relative GPA, x_5 = relative rank.

The estimate results are displayed in the Table 2 below.

Table 2: Estimates for MLR Full Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-14.4337542	3.5073785	-4.115254	0.0001111
grade	0.1009711	0.0207692	4.861575	0.0000077
psat	0.0040129	0.0018722	2.143423	0.0358265
sexM	0.4329066	0.1796370	2.409897	0.0187968
relative_gpa	7.2348077	2.9700193	2.435946	0.0176031
relative_rank	1.4380614	1.4034878	1.024634	0.3093341

Thus, our estimated full model is $\hat{y} = -14.434 + 0.101x_1 + 0.004x_2 + 0.433I(male) + 7.235x_4 + 1.438x_5$

Additionally, we see from the associated p-values for each estimate that relative rank is not statistically significant (at the 0.05 level) conditional on the other predictors in the model. However, it's worth noting relative rank was marginally significant with a negative slope coefficient (analysis not shown). Additionally, the slope estimate for the relative rank predictor is positive, which is illogical. An increase in relative rank (which indicates a decrease in academic achievement) would be associated with an increase in the student's AP score. As a result, we may not want to include relative rank as a predictor. The relative rank predictor will be discussed further in subsequent sections.

Model Selection

To select the optimal model, I conducted a best subset regression.

```
##                 Best Subsets Regression
## ---------------------------------------------------------
## Model Index    Predictors
## ---------------------------------------------------------
##      1         grade
##      2         grade psat
##      3         grade sex relative_gpa
##      4         grade psat sex relative_gpa
##      5         grade psat sex relative_gpa relative_rank
## ---------------------------------------------------------
##
##                                                  Subsets Regression Summary
## ------------------------------------------------------------------------------------------------------------
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square     C(p)        AIC         SBIC        SBC         MSEP
## ------------------------------------------------------------------------------------------------------------
##   1        0.4459      0.4379      0.4193    24.4165    178.5887    -23.8937    185.3767    48.6287
##   2        0.5128      0.4985      0.4709    15.3848    171.4580    -30.7590    180.5088    43.3987
##   3        0.5670      0.5477      0.5115     8.4361    165.0784    -36.4597    176.3918    39.1517
##   4        0.5997      0.5754      0.5311     5.0499    161.5126    -39.2385    175.0886    36.7566
##   5        0.6061      0.5757      0.5214     6.0000    162.3749    -38.0237    178.2137    36.7376
## ------------------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

Based on the best subset regression output (via most selection criteria), a model with four predictors, namely course grade, PSAT, gender, and relative GPA, should be used. It is worth noting that the adjusted R-squared value is slightly higher for the full model (with all five predictors) in addition to a lower MSEP. However, I chose to select the model with the four aforementioned parameters.

The estimate results are displayed in Table 3 below.

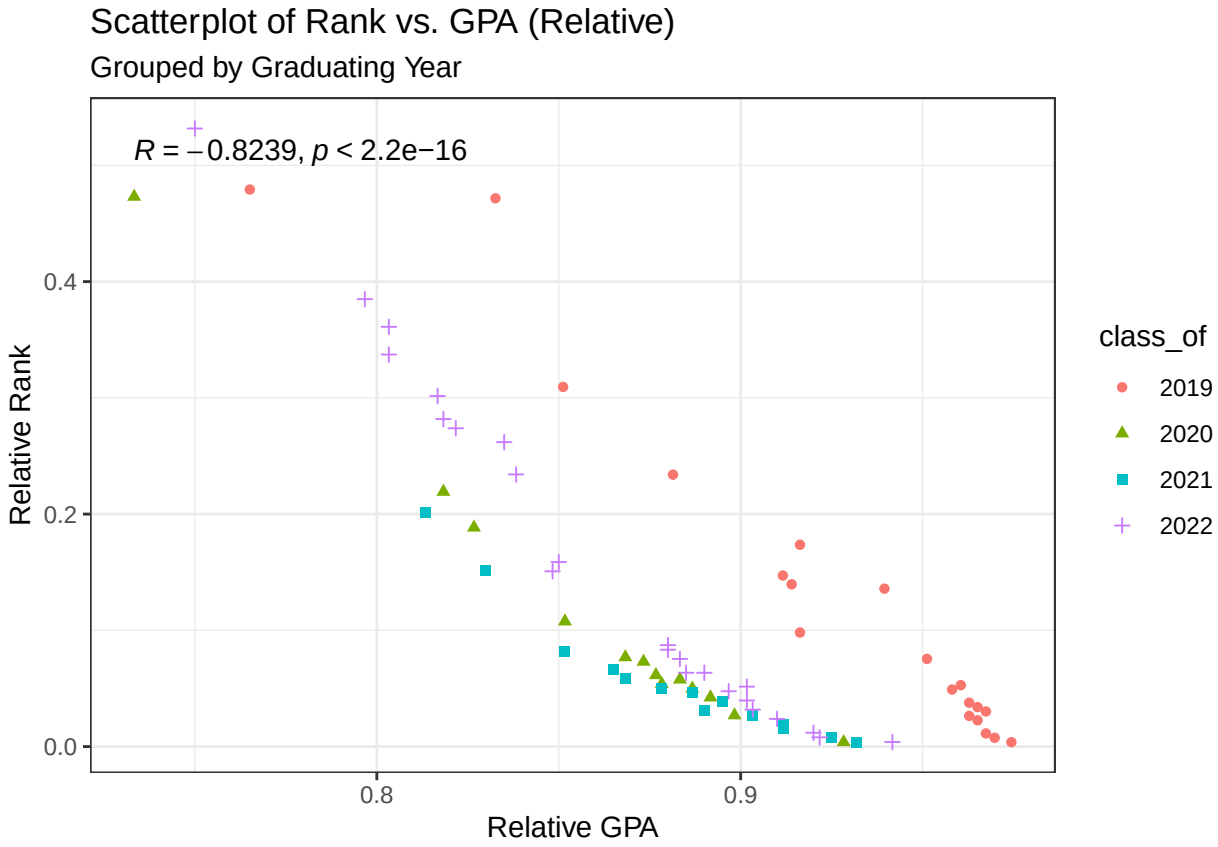
Table 3: Estimate for MLR Optimal Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-11.2659909	1.6570333	-6.798892	0.0000000
grade	0.0897663	0.0176639	5.081917	0.0000033
psat	0.0042971	0.0018523	2.319940	0.0234464
sexM	0.4830882	0.1728972	2.794077	0.0068061
relative_gpa	4.7636800	1.7340430	2.747152	0.0077419

Thus, our estimated optimal model is $\hat{y} = -11.266 + 0.090x_1 + 0.004x_2 + 0.483I(male) + 4.764x_4$. We note that all p-values are significant at the 0.05 level and have decreased compared to the full model.

Variance Inflation Investigation

Due to the correlative nature of GPA and rank, I created a scatterplot of relative rank vs. relative GPA to assess the correlation between the variables.



The correlation between relative rank and relative GPA appears high. As a result, I calculated the variance inflation factors of the full and optimal models. We can see from Table 4 below the VIFs for the full model are higher than that of the reduced (optimal) model. Due to this fact coupled with the relative rank predictor producing nonsensical behavior, I believe the earlier decision to exclude relative rank as a predictor in the model is valid.

Table 4: VIF (Full Model v. Optimal Model)

	VIF
grade	1.994765
psat	1.455474
sex	1.106012
relative_gpa	3.607268
relative_rank	4.671642

	VIF
grade	1.441766
psat	1.423534
sex	1.023803
relative_gpa	1.228716

Influence Diagnostics

Potential Outliers

Table 5: Standardized Residuals (Optimal Model)

	Studentized.Res	R.student
12	2.098	2.155
47	-2.068	-2.122
69	-2.232	-2.303

Table 5 above shows the observations whose standardized residual value is greater than a magnitude of 2. Given our data set has 71 observations, we would expect approximately 3 or 4 values to have standardized residual values greater than a magnitude of 2. Thus, none of these points should be considered a significant outlier. Influential points will be analyzed in the next section.

Influential Points

Table 6: Influence Measures

	Cooks	DFFITS	COVRATIO
3	0.088	-0.670	1.119
4	0.009	0.214	1.303
13	0.061	0.559	0.992
25	0.067	0.591	0.912
33	0.004	0.134	1.298
47	0.055	-0.540	0.822
62	0.068	-0.595	0.900
67	0.061	0.553	1.123
69	0.110	-0.765	0.810

Table 6 above shows the observations which violate the threshold value of at least one influence metric listed. There are 6 observations which violate two of the three metrics, while no observation violated all three metrics. The cutoff values for each metric can be seen in Table 7 below.

Table 7: Influence Measure Cutoff Values

Metric	Cutoff Value
Cook's	0.056
DFFITS	0.531
COVRATIO	< 0.789 or > 1.211

While the given metrics have flagged these observations as influential, they are not severely atypical or worthy of removal from the data set. When looking at the observations in question, the students tend to fall in one of three categories:

(1) Students who do not complete organizational and duty-oriented tasks (i.e. homework, submitting assignments on time, classwork, etc.). As a result, their GPA and course grade are negatively impacted. However, some of these students (at least as it pertains to AP Calculus) are naturally gifted at math and are able to achieve high AP scores as a result.

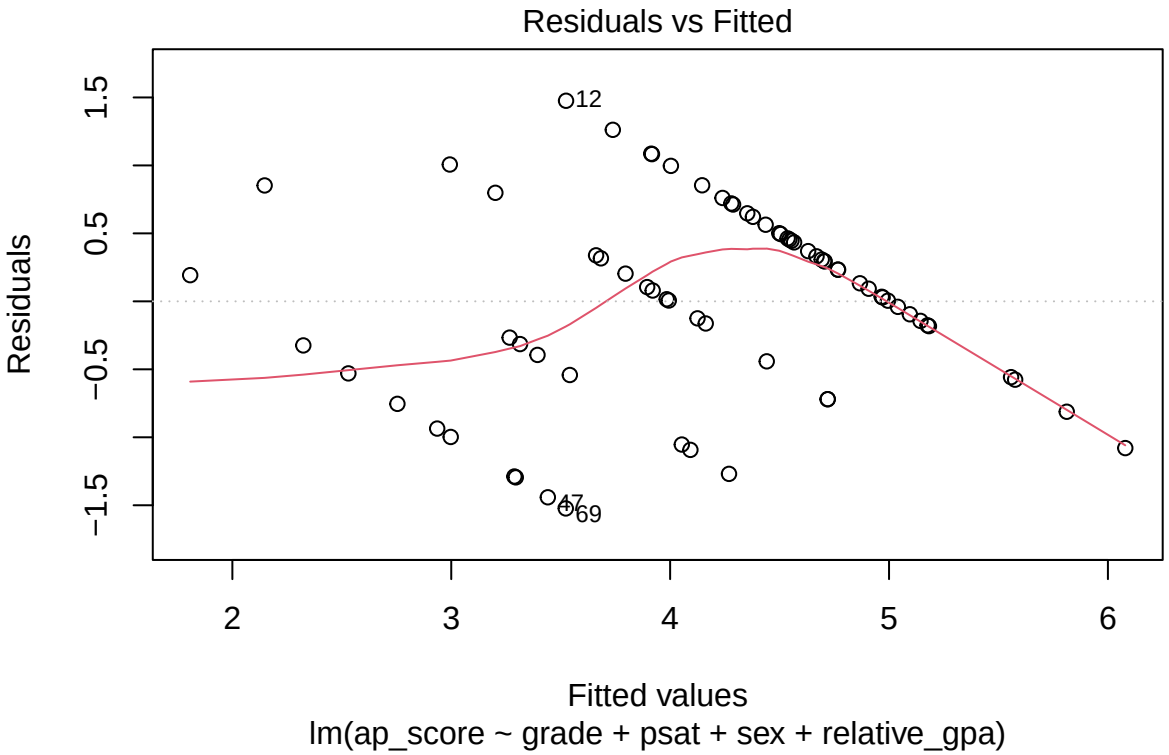
(2) Exchange students whose primary language is not English. With these limitations, it's often more difficult for these students to communicate effectively, more so in reading- and writing-based courses such as English, Literature, and History.

(3) Students who are challenging themselves with the AP class. If a student is interested in challenging themselves or needs the course for future study, we will often allow them to enroll even if the prerequisite requirements are not fully met.

As these observations are representative of the future student population, I do not believe there is a valid reason to remove them.

Model Assumptions

Linearity and Equal Variance



The residuals versus fitted plot appears problematic for the linearity and equal variance assumptions of multiple linear regression. The points do not appear to be centered at zero and it is difficult to evaluate if the variance is constant throughout.

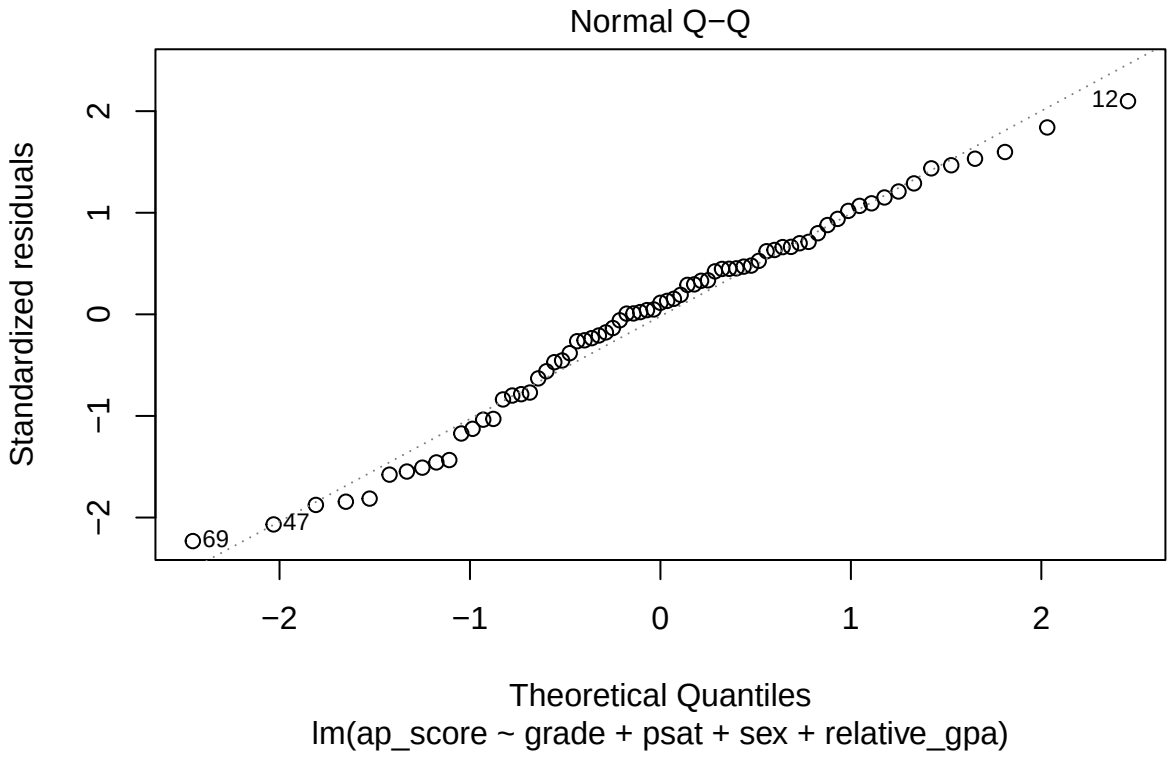
Breusch-Pagan Test I decided to perform the Breusch-Page Test to determine if heteroscedasticity is present. The hypotheses of the test are

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

```
##
##  studentized Breusch-Pagan test
##
## data:  mlr_ap_model_best
## BP = 6.4341, df = 4, p-value = 0.169
```

Given $p = 0.169 > 0.05$, we fail to reject H_0 at the 0.05 level. We do not have enough evidence to show heteroscedasticity is present. While this result is positive, this test does not confirm or disprove a linear association.

Normality



The Normal Q-Q plot above shows strong evidence of normality of the residuals.

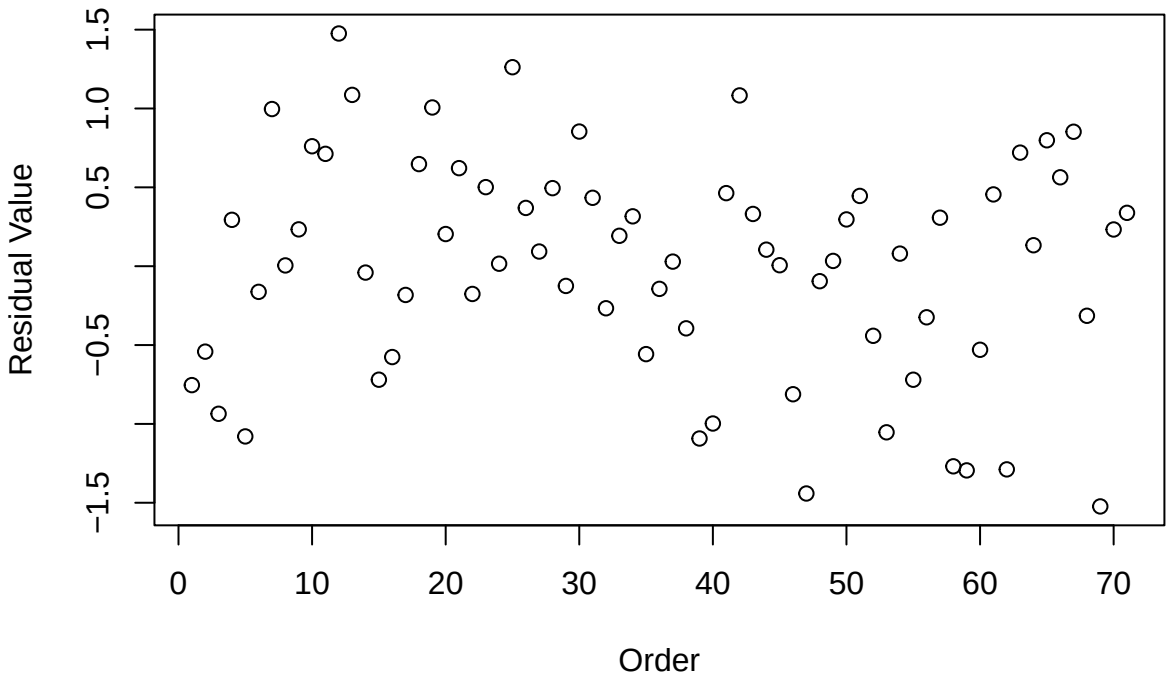
Shapiro-Wilk Test I decided to perform the Shapiro-Wilk Test to determine if the residual values are normally distributed. The hypotheses of the test are

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

```
##
##  Shapiro-Wilk normality test
##
## data:  mlr_ap_model_best$residuals
## W = 0.98021, p-value = 0.3249
```

Given $p = 0.3249 > 0.05$, we fail to reject H_0 at the 0.05 level. We do not have enough evidence to show the residuals are non-normal.

Independence



A plot of the residuals vs. order is shown above. There does not appear to be any significant patterns, which is evidence of independence. I believe independence is a reasonable assumption for this data.

Final Thoughts - Multiple Linear Regression

While there could be some value to this model, due to the discrete nature of my response variable, I decided to investigate ordinal logistic regression to achieve my goal of predicting AP scores. I believe this model may be better suited for my data structure.

Additional Analysis – Ordinal Logistic Regression

Mathematical Formulation

The ordinal logistic regression model is given as $logit[P(Y \leq j)] = \zeta_j - \sum \beta_i x_i$ where $j \in \{1, 2, 3\}$ and $i \in \{1, 2, 3, 4\}$. In this model j = the level of an ordered category, i = predictor variable, and ζ_j = intercept of category level j. Given none of the AP scores in the data have a value of 1, this results in three cutoff values: 2 to 3, 3 to 4, and 4 to 5. Thus, there are 3 associated intercepts.

The OLR formulation gives the log odds of being in category j or lower. After fitting, probabilities can be extracted for each category. First, we fit a full model with all predictor variables. The estimate results are displayed in Table 8 below.

Table 8: Estimates for OLR Full Model

	Value	Std. Error	t value	p-value
grade	0.3272438	0.0399715	8.186938	0.0000
psat	0.0190135	0.0058091	3.273077	0.0011
sexM	1.5273859	0.6359913	2.401583	0.0163

	Value	Std. Error	t value	p-value
relative_gpa	26.9508813	0.3305178	81.541379	0.0000
relative_rank	7.0549413	2.5529507	2.763446	0.0057
2|3	63.7526425	0.4697556	135.714503	0.0000
3|4	65.3447755	0.7273798	89.835842	0.0000
4|5	67.3050663	0.8681696	77.525254	0.0000

Thus, our estimated full model is

$logit[P(Y \leq 2)] = 63.753 - (0.327x_1 + 0.019x_2 + 1.527I(male) + 26.951x_4 + 7.055x_5)$

$logit[P(Y \leq 3)] = 65.345 - (0.327x_1 + 0.019x_2 + 1.527I(male) + 26.951x_4 + 7.055x_5)$

$logit[P(Y \leq 4)] = 67.305 - (0.327x_1 + 0.019x_2 + 1.527I(male) + 26.951x_4 + 7.055x_5).$

Upon viewing the estimated results, we see all predictors are significant at the 0.05 level, including relative rank. However, the estimate for the slope coefficient is still positive, which produces illogical results. Thus, I decided to explore other models via best subset selection with AIC criterion.

Model Selection

I conducted a brute-force best subset regression (analysis not shown). The two lowest AIC values were for a model with the four predictors: course grade, PSAT, gender, and relative GPA (AIC = 117.3327) and the full model (AIC = 117.0995). While the full model has a slightly lower AIC (difference of 0.2332), I decided to choose the model with four predictors because the slope coefficient of the relative rank variable was nonsensical.

The estimate results for the optimal model are displayed in Table 9 below.

Table 9: Estimates for OLR Optimal Model

	Value	Std. Error	t value	p-value
grade	0.2593785	0.0392276	6.612140	0.0000
psat	0.0185371	0.0056663	3.271474	0.0011
sexM	1.7862837	0.5753878	3.104487	0.0019
relative_gpa	15.4101355	0.0048768	3159.869786	0.0000
2|3	46.5662137	0.0194775	2390.763987	0.0000
3|4	48.0888841	0.4691523	102.501645	0.0000
4|5	49.9687507	0.5537845	90.231399	0.0000

Thus, our estimated optimal model is

$logit[P(Y \leq 2)] = 46.566 - (0.259x_1 + 0.019x_2 + 1.786I(male) + 15.410x_4)$

$logit[P(Y \leq 2)] = 48.089 - (0.259x_1 + 0.019x_2 + 1.786I(male) + 15.410x_4)$

$logit[P(Y \leq 2)] = 49.969 - (0.259x_1 + 0.019x_2 + 1.786I(male) + 15.410x_4)$

Model Assumptions

No Multicollinearity

Comparing, again, the VIFs of the full model vs. the optimal model (see Table 4 on page 10). I believe the model chosen during the selection period is appropriate as the VIFs are lower for all predictors in the model

without relative rank. While the VIFs for the full model do not exceed 5, I believe having lower VIFs with the ordinal logistic regression is desirable.

Proportional Odds

The proportional odds assumption says the coefficients that describe the odds of being in the lowest category vs. all higher categories of the response variable are the same as those that describe the odds between the second lowest category and all higher categories, etc. In other words, the slope coefficients must the same for each category. I ran the Brant test to determine if the proportional odds assumption is met. The hypotheses of the test are

H_0 : Proportional odds holds vs. H_a : Proportional odds is violated.

```
## --------------------------------------------
## Test for X2  df  probability
## --------------------------------------------
## Omnibus      1.14    8   1
## grade        0.02    2   0.99
## psat     0.02    2   0.99
## sexM     0.17    2   0.92
## relative_gpa 0.79    2   0.67
## --------------------------------------------
##
## H0: Parallel Regression Assumption holds
```

From the code output above, we see the overall p-value of the test is 1. Thus, we fail to reject H_0 at the 0.05 level and conclude the proportional odds assumption is reasonable.

Prediction

Unfortunately, I do not have current data to test the model due to 2023 AP scores being released in July. However, I was able to gather some data from a few volunteer students to test the model. For now, I am using my knowledge of the AP exam as well as the individual students to formulate a "teacher prediction". The student data is given in Table 10 below.

Table 10: Student Data

Student	Course Grade	PSAT	Gender	Relative GPA
Student 1	85	570	M	0.8733333
Student 2	87	580	F	0.8916667
Student 3	98	650	M	0.9233333

For each student, the probability of obtaining a particular score is shown in Table 11.

(1) Student #1 is most likely to score a 3.

(2) Student #2 is most likely to score a 1 or 2.

(3) Student #3 is most likely to score a 5.

Given my experience with AP score outcomes and knowledge of the students, I would predict a 4, 3, and 5 for students 1, 2, and 3, respectively. I am very interested to compare the model predictions with the

AP scores upon their release in July. The only model prediction to match my prediction is for student #3. However, P(score=4) is very close to P(score=3) for student #1 (similar circumstances for student #2), so I believe my prediction could still be correct.

Table 11: Predicted Outcome

	P(score=1or2)	P(score=3)	P(score=4)	P(score=5)
Student1	0.2155844	0.3419309	0.3344477	0.1080370
Student2	0.3794343	0.3576212	0.2113118	0.0516327
Student3	0.0009898	0.0035320	0.0243822	0.9710960

Cross Validation Model - Nascent Stage

I wanted to use cross validation to test the predictive power of my model. However, I was only able to perform some basic analysis. Note: There will be some raw code output. I created a model using 10-fold cross validation using the same four predictors from my optimal MLR and OLR models. The model output is shown below

```
## Ordered Logistic or Probit Regression
##
## 71 samples
##  4 predictor
##  4 classes: '2', '3', '4', '5'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 64, 64, 63, 63, 64, 65, ...
## Resampling results across tuning parameters:
##
##   method    Accuracy   Kappa
##   cauchit   0.6803571  0.4886917
##   cloglog   0.6619048  0.4318943
##   logistic  0.6970238  0.5015304
##   loglog    0.6946429  0.4953053
##   probit    0.7113095  0.5183822
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was method = probit.
```

When calling the train function, I passed it the value of "polr" for the "method =" parameter. Since the polr() function also has a "method =" parameter (which I believe can be passed the values "cauchit", "cloglog", "logistic" [default], "loglog", or "probit"), I believe the train() function is testing the accuracy of the model using all of the different "links" (I believe this is what they are called). The link with the highest accuracy is selected; specifically in this case, the "probit" link, with an accuracy of approximately 71%.

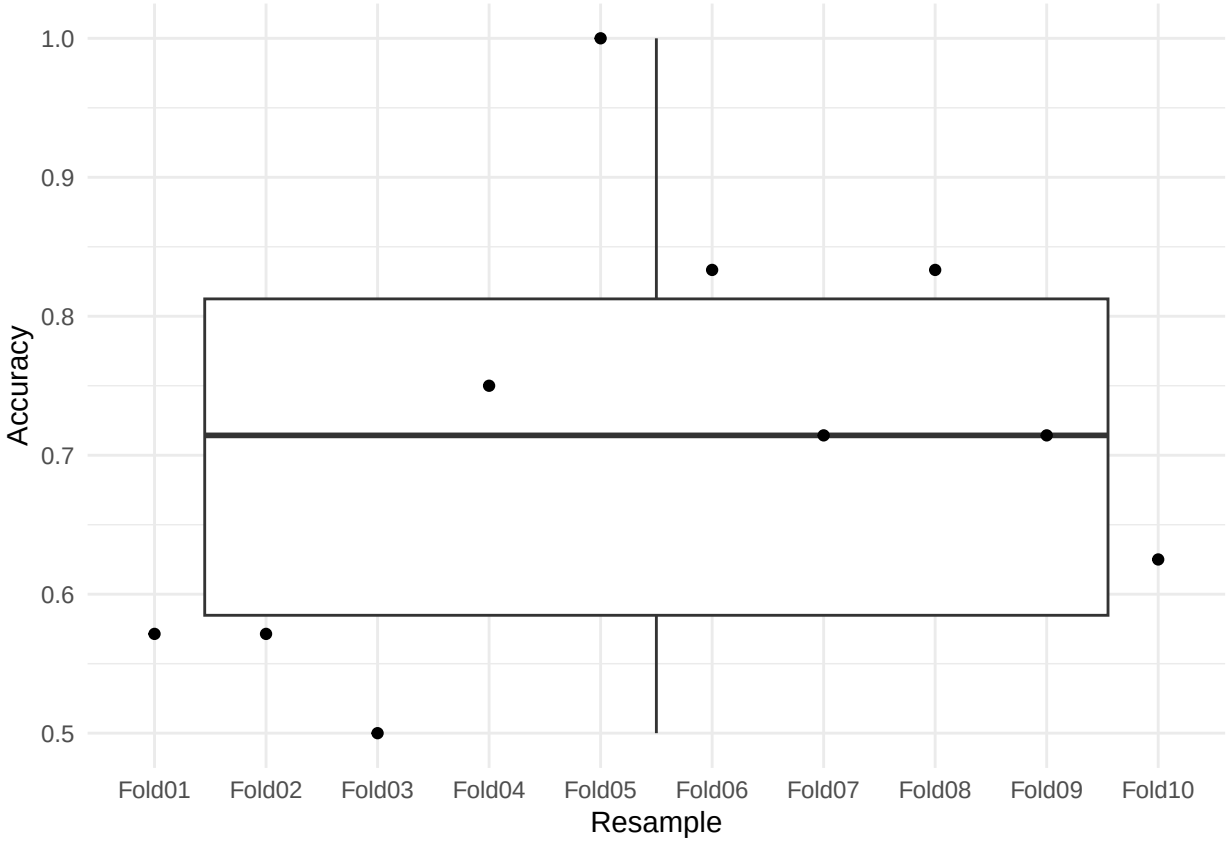
We can also get a summary of the estimates (shown below).

```
##                    Value  Std. Error     t value p-value
## grade         0.15123331 0.022856241    6.616718  0.0000
## psat          0.01043548 0.003270137    3.191146  0.0014
## sexM          1.06293290 0.327379031    3.246796  0.0012
```

```
## relative_gpa  8.99036306 0.001806751 4975.982150  0.0000
## 2|3          26.91465416 0.013069827 2059.296901  0.0000
## 3|4          27.80277163 0.267230586  104.040380  0.0000
## 4|5          28.89873314 0.310006210   93.219852  0.0000
```

We can see all of the predictors are significant at the 0.05 level. Also, the AIC of the model is 116.3092 (not shown), which is lower than the optimal model from the MLR and OLR. However, since the probit method was used, I am unsure how to formulate the model. Furthermore, if another method is used (other than logistic), does the model formulation change?

Lastly, I plotted the accuracy of test data for each of the ten folds (shown below).



We can see the accuracy ranges from 50% to 100%, with the median being approximately 71% (mentioned earlier). Given the validity of my formulation, I am hopeful this model will be able to help predict future AP scores of my students.

Final Thoughts & Future Considerations

Overall, I feel I was able to create a solid foundation for building a working model. There is still much work and fine-tuning to accomplish, but overall I am pleased with the results. The true test will come when I compare the predictions of the model to the 2023 AP scores.

For future analyses, I would like consider the following:

(1) Create a new model for academic years in which the maximum GPA is 6.0. After viewing the scatterplot of relative rank vs. relative GPA, it was clear that higher relative GPAs were more common under the

former GPA system (max of 4.3). Thus, I would like to see the prediction accuracy of a model with one GPA scale.

(2) Additional relevant predictors. Some suggestions by classmates include: a categorical variable for race/ethnicity, a numerical variable for math GPA, or a variable which describes amount of effort on the exam (could be categorical or numerical). Perhaps a variable which captures the number of extracurricular activities or sport commitments would be appropriate to account for the amount of time they have to spend on academic pursuits.

(3) Continue learning more about the cross validation models and how they are formulated.

(4) Consider multiple imputations for missing PSAT values.