Lab 2: Student Perfomance

https://github.com/mids-w203/lab2_StudentPerformance

Theo Abraham, Yiwen Hou, Rohan Kapur, Carla Tapia

April 18, 2025

Introduction and Context

Despite the significant investment in time and financial resources that is incurred by most students who start an undergraduate program, the overall graduation rate for students attending undergraduate courses in the United States was only 64% in 2020. Still, a four-year college degree is the most important indicator of economic success, and a lot of research has been conducted to determine why some students are successful in college while others fail to earn a degree. We pose the question:

> *Is academic performance in high school a good predictor of success in college?*

We propose that students who have higher GPAs in high school do well in college, either because they develop good study habits in high school or because they are exposed to more preparatory material in the form of honors and AP courses. For the purpose of this research, we will compare students' GPAs from high school (X random variable) with their cumulative GPAs in college (Y variable) to build a linear regression model, which we will use to test for a consistent correlation.

Data and Methodology

To answer our research question we utilize the Student Performance Metrics Dataset from mendeley.com, which is a survey of undergraduate students from a university in Malaysia. This dataset includes attributes such as Gender, High School GPA, Family Income, Computer Usage, proficiency in English, Class Attendance, Study time outside class, Extra Curricular activities, and Overall College GPA. The survey was based on a structured questionnaire that was filled out by students who volunteered to participate in the study. The dataset is relatively clean and robust with 443 individual records. This report will be focus on and analyze students in the Computer Science and Engineering department.

To better understand the data and the relationships inherent in the variables, we built two linear models based on the Mendeley dataset described above. The first model is a simple credible model in which we build a relationship between students' High School GPAs, captured in the 'HSC' column of the dataset, and their overall College GPAs, captured in the 'Overall' column. We call this the "simple model" and propose the following hypothesis:

> **H_0**: There is no relationship between highschool GPA and acamedic performance in college (college GPAs).

The second model, which we call the "complex model," includes two additional X variables and compares students' High School GPA, their college class attendance, captured in the 'Attendance' column, and student gender, captured in the 'Gender' column, against their performance in college. The purpose of this model is to go beyond the distribution presented by high school and college GPAs and to discover the more complex relationships in our data. The decision to use Gender and Attendance was decided based on external research that uncovered attendance positively affecting class grade and overall GPA (The New School) and females tending to outperform males academically in higher education (Verbree et al., 2022). Our null hypothesis for this models is as follows:

> **H_0**: There is no relationship between, on one hand, highschool GPA, college class attendence, and student gender, and, on the other hand, student performance in college.

To explore this complex model further, the scatter plot below demonstrates students' College GPAs vs. High School GPAs separated by Gender and color-coded by Attendance level. We noticed both male and female students with high attendance levels had high high school and college GPAs which follows our initial proposal. The opposite occurs for male and female students with low attendance levels, they tend to have lower high school and college GPAs. Overall, there is an observed trend that students with higher high school GPAs tend to also have higher college GPAs.

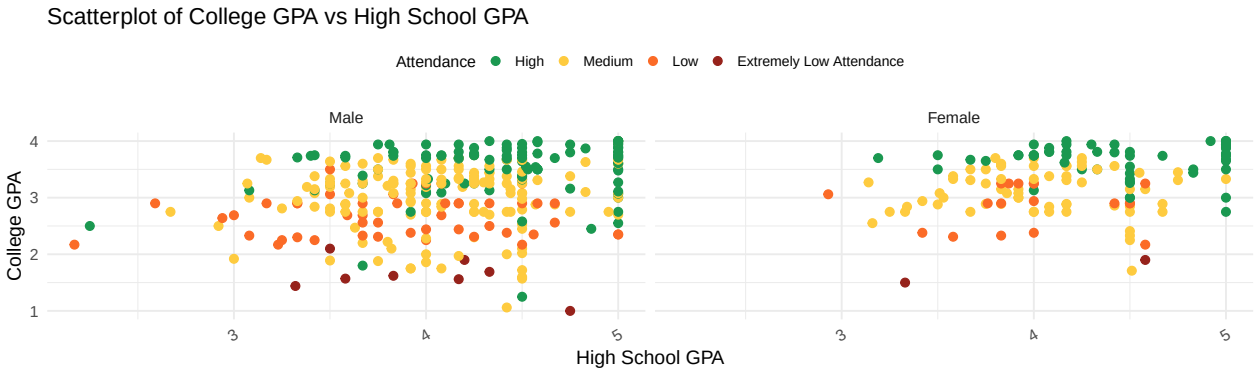

Figure 1: College vs. High School GPA by Gender & Attendance

Model Specification & Assumptions

The following assumptions were assessed to ensure the dataset was viable for our proposed linear regression use.

1. I.I.D. data (Independent and Identically Distributed): The IID assumption is not fully met as the data is collected only from students at the University Malaya in Malaysia which introduces geographical bias. We acknowledge that there might be conditions specific to the undergraduate experience in Malaysia that are not relevant to students in other parts of the world. Additionally, the data resulted from survey responses, so there may be some bias as some students may have decided to not respond or opt-ed out. Lastly, we are observing data from students only in the Computer Science department. We do believe the sample is still a good representation of the Computer Science department and the CS department gender distribution. Though we acknowledge these discrepancies, the dataset contains a large amount of data that would be necessary and useful to answering our research question.

2. Linear Conditional Expectation: The linear conditional expectation assumption is met for both models. The predicted and residual values for these models, and each individual variable, were plotted onto a scatter plot and the average of the residuals stays centered at 0. These plots are shown in the Appendix. Given this assumption we proceed to use a simple linear function to model the relationship between academic performance in high school (High School GPA) and performance in college (College GPA), in the case of the simple model. And a multivariate linear function to model the relationship between academic performance in high school (High School GPA) and performance in college (College GPA), Gender, and Class Attendance, in the case of the complex model.

3. No Perfect Collinearity: A correlation matrix was developed to assess if perfect collinearity between the variables exists. Through this evaluation, none of the variables presented perfect collinearity which is shown in the Appendix. Though this assumption is satisfied, we did note that Attendance and High School GPA had a correlation of 0.339, indicating a moderate positive relationship between these two variables.

4. Homoskedastic Errors: Through the use of the Breusch-Pagen test, the simple model produced a p-value of 0.2725, which meant the model exhibited homoskedastic errors as we fail to reject the null hypothesis that the error variances are homoskedastic. The complex model produced a p-value of 0.007165, exhibiting heteroskedastic errors as we reject the null. To correct this, robust standard errors were utilized in our estimates shown in the regression coefficient table.

5. Normally Distributed Errors: A histogram and QQ plot of the model residuals were plotted to ensure normality. Both models exhibited a relatively normal distribution with a minor left skew which was verified with the QQ plot. These plots can be observed in the Appendix.

Results & Analysis

After ensuring assumption compliance, the below stargazer table exhibits the output of each specified model: simple & complex.

	Dependent variable:	
	Overall College GPA	
	Simple model	Complex model
	(1)	(2)
High School GPA	0.29*** (0.05)	0.05 (0.05)
Female		0.13*** (0.04)
Low Attendance		1.09*** (0.11)
Medium Attendance		1.38*** (0.11)
High Attendance		1.91*** (0.11)
Constant	1.97*** (0.20)	1.41*** (0.20)
Observations	443	443
R^2	0.06	0.44
Adjusted R^2	0.06	0.43
Residual Std. Error	0.58 (df = 441)	0.45 (df = 437)
F Statistic	29.97*** (df = 1; 441)	68.21*** (df = 5; 437)
Note:		*p<0.1; **p<0.05; ***p<0.01

Simple linear model: College GPA = 1.97 + 0.29 High School GPA

Complex linear model: College GPA = 1.41 + 0.05 High School GPA + 0.13 Female + 1.09 Low Attendance + 1.38 Mid Attendance + 1.91 High Attendance

For the simple mode, High School GPA accounts for about 6% of College GPA, based on R^2 value. After adding Gender and Attendance variables, the model's explanatory power improved significantly — from 6% to 43% based on adjusted R^2 value. The model showed that female students had, on average, 0.13 higher GPA compared to male students, holding other variables constant. From all X variables, Attendance is the biggest contributor for College GPA. All Attendance-related variables are statistically significant, and had a relatively high coefficient number. Students in the low attendance group had 1.09 higher GPA compared to student who fell in the extremely low attendance group, holding other variables constant. The data clearly reflects a real-world pattern — higher levels of attendance are associated with much higher college GPA.

While Attendance alone accounts for only about 10% of the variation in College GPA, the substantial increase in R^2 when it is included in the model suggests that Attendance may be capturing the influence of other underlying factors. These could include knowledge proficiency, time spent studying, or overall student engagement — all of which are likely associated with both higher attendance and better academic performance. This can be useful for educators and those in education policy to understand the importance of attendance and how to improve attendance for struggling students.

```
## Analysis of Variance Table
##
## Model 1: Overall ~ HSC
## Model 2: Overall ~ HSC + Gender + Attendance
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    441 149.113
## 2    437  89.444  4    59.668 72.881 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To test whether the additional variables in our complex model significantly improve the model's explanatory power compared to the simple model, we conducted an analysis of variance using an F-test. The null hypothesis is that the simple model, with only High School GPA, as a predictor fits the data just as well as the model with gender and attendance as additional predictors. The F statistic of 72.88 is large and the associated P value is 2.2e-16 which is well below the significance threshold of 0.05. These two values give us more than enough evidence to reject the null hypothesis. Therefore the complex model which includes more variables provides a significantly better it to the data. To conclude, the variables Gender and Attendance improves our ability to predict the college GPA.

Conclusion

Our simple linear regression model returned an adjusted R^2 value of 0.062 and the complex model returned an R^2 value of 0.432, indicating minimal correlation in the simple model and moderate correlation in the complex model. The relatively low value for R^2 in the simple model suggests that variation in college GPAs are not adequately explained by academic outcomes in high school alone. This requires that we consider other factors, other than high school GPA, which are more likely to contribute to academic success in college. The complex model, which included student Gender and Attendance had a higher R^2 value which suggests that these two factors are more likely to explain variability in college GPAs. However, we also noticed some collinearity between high school GPA and attendance levels in college (see Appendix for collinearity tests). This alone is not surprising and simply suggests that students who did better in high school are more likely to attend classes regularly in college. The substantial increase in R^2 when it is included in the model suggests that Attendance may be capturing the influence of other underlying factors, knowledge proficiency, time spending on study etc. The remaining variable in the complex model, student Gender, likely affects variability in college GPAs to a greater extent and further research on this dataset will likely definitively show that students of a certain gender are more likely to have higher GPAs in college.

Appendix

References: https://guidetoteaching.newschool.org/why-attendance-matters/

http://pmc.ncbi.nlm.nih.gov/articles/PMC9379878/

Data Source: https://data.mendeley.com/datasets/5b82ytz489/1

https://github.com/mids-w203/lab2_StudentPerformance/blob/main/data/model/clean_student_performance.csv

The data wrangling R file can be found in the src folder.

Additional Models: We developed a two additional models prior to deciding on the two used in this report. The first model included HS GPA and Gender as our X variables with college GPA as our Y variable. When doing analysis, we noted the complex model used in this report performed better so we ultimately went with that model. The other model included Attendance as the X Variable and college GPA as our Y, which was a stronger model than the first one but our complex model still performed better.

IID Assumption: There are 302 males and 141 females in the dataset.

Linear Conditional Assumption:

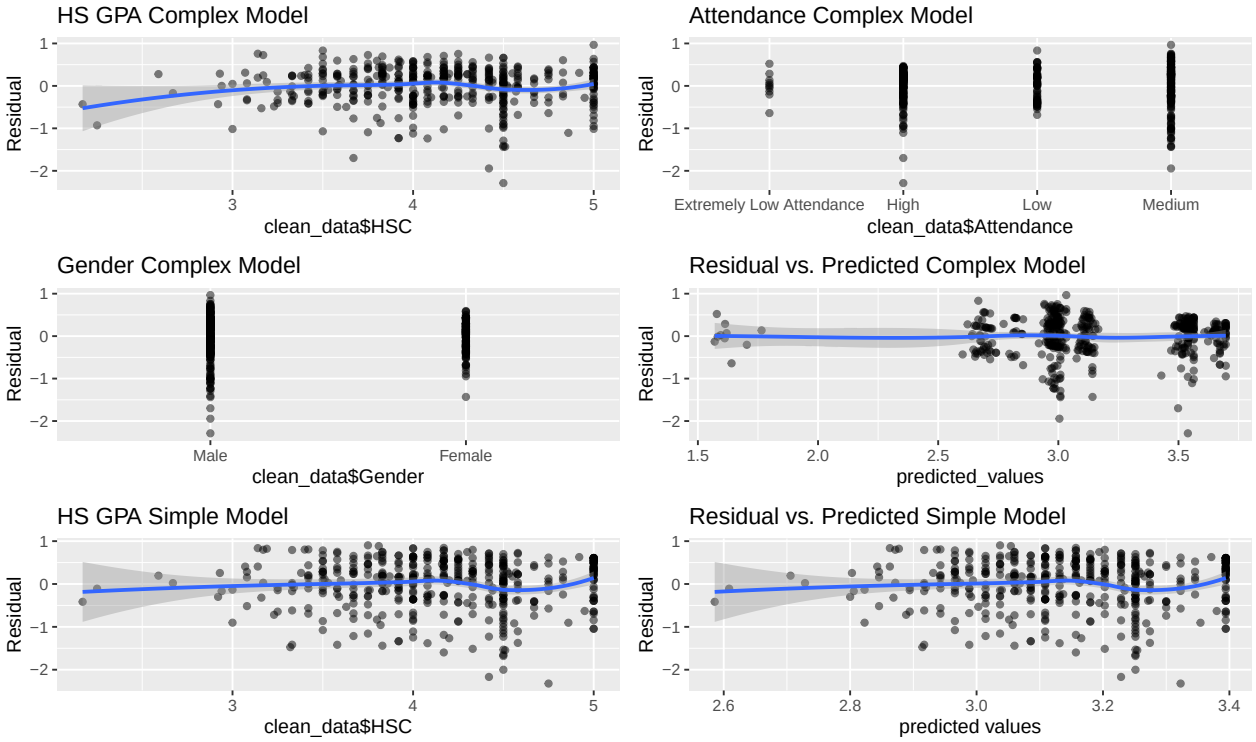

Figure 2: Linear Conditional Expectation Assumption Plots

No Perfect Collinearity Assumption:

```
##                  High School GPA     Gender Attendance
## High School GPA      1.00000000 0.06006344 0.33994832
## Gender               0.06006344 1.00000000 0.07360205
## Attendance           0.33994832 0.07360205 1.00000000
```

Homoskedastic Assumption:

Breusch-Pagen test:

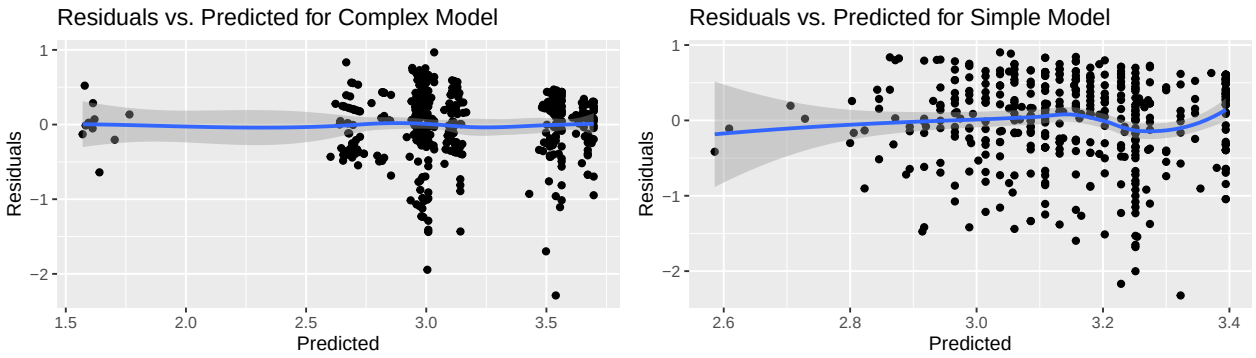

Figure 3: Homoskedastic Assumption Check: Scatterplot of Residuals vs. Predicted Values for Models

```
##
##  studentized Breusch-Pagan test
##
## data:  complex
## BP = 15.89, df = 5, p-value = 0.007165

##
##  studentized Breusch-Pagan test
##
## data:  simple
## BP = 1.2041, df = 1, p-value = 0.2725
```

Normally Distributed Errors Assumption:

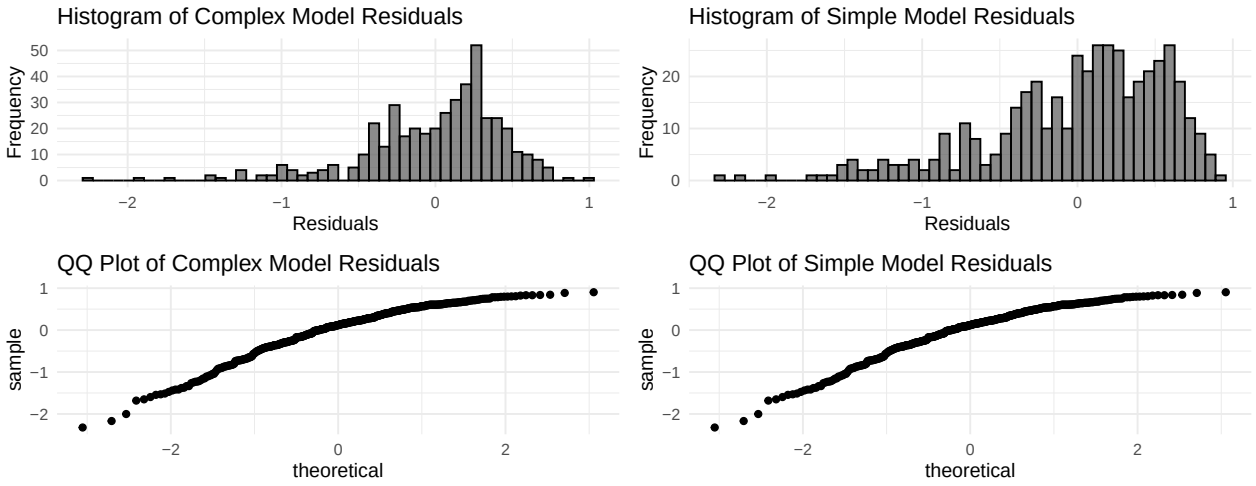

Figure 4: Normally Distributed Error Assumption: Histograms & QQ Plot of Residuals