College GPA Statistical Analysis

(COMP3125 Individual Project)

Devon Lapierre
Wentworth Institute of Technology
School of Computing and Data Science

***Abstract*—**This report looks at the different variables that play a role in determining college GPA. Models such as Linear Regression, ANOVA, and Multiple Linear Regression are used in this analysis to gather information on which variables are significant in determining college GPA. The conclusions drawn from this report can help both students and educators make informed decisions on how they should handle their education to achieve the best academic standing possible, which in turn results in more career opportunities and further success.

Keywords—significance, regression, hypothesis testing, test statistics

I. Introduction

With college education becoming such an important factor in career success, it is essential to look in depth at the leading factors that contribute to the GPA of college students. It can be intriguing to analyze what categories of college students have a significant impact on their grades and overall measured success in their college education, since this measurement plays a key role in future life and career endeavors for most individuals. Some obvious categories that fit the general criteria for determining college GPA include high school GPA, major, study hours, attendance, and extracurricular activities, but how significant are these factors and others such as part-time jobs, library usage, sleep, and more in predicting a model for a given student's GPA? A study done at the University of Alaska [1] determined that high school GPA is the most significant predictor of college GPA, but can other factors within the dataset be just as significant or more? One interesting variable to consider would be having a part time job while being a college student, and another could dive into the different major classifications and how they could play a role in determining GPA. With all of the different categories of potential significant explanatory variables, building a predictive model of their relationships with college GPA can itself be significant in looking at career success.

II. Datasets

A. Kaggle.com

The source of the dataset is Kaggle, a public site with free access to thousands of datasets. The dataset that analyzed in this report is licensed by Massachusetts Institute of Technology, giving it credibility and integrity. However, minimal information on its generation and its author is provided.

B. Dataset Characteristics

The dataset [2] is a comma-separated values file just under 100 kB in size. The actual file contains 10 columns of nine explanatory variables and a response variable, with 2000 rows of data. The columns and their respective data types are as follows:

Column	Data Type
Study Hours per Week	Hours
Attendance Rate	Grade Scale (0-100)
Major	Categorical
High School GPA	Grade Scale (0-4
Extracurricular Activities	Integer Value
Part-Time Job	Binary (Yes/No)
Library Usage per Week	Hours
Online Coursework Engagement	Hours
Sleep Hours per Night	Hours
College GPA	Grade Scale (0-4)

While no conversions were made, the data was cleaned to be easily interpreted as a data frame in the python programming language. Empty rows in the Major column were changed to "Undeclared", as in undeclared in the dataset rather than undeclared majors. Empty rows in all other columns were dropped, as replacing them with 0 would inaccurately skew the data, and duplicates were dropped as well, resulting in 1473 rows of data. One additional column of Part-Time Job Indicator was created, with the binary data type of zeros and ones, correlating to No and Yes, respectively, in the Part-Time Job column.

III. Methodology

A. Pearson Correlation Coefficient

The first method used in the investigation is the Pearson Correlation Coefficient, a calculated value that shows the correlation between two quantitative variables. This method was used to test the correlation between high school GPA and college GPA. This method was chosen since it is a simple and efficient way to answer the question of whether there is significance in high school GPA determining college GPA. The python module applied to get the correlation coefficient and p value was the "pearsonr" function within the "scipy.stats" module.

B. Linear Regression Model

The second method used to answer the questions is a simple linear regression model, which analyzes the significance of the response variable based on a binary explanatory variable. In the case of the dataset, the impact of having a part time job on college GPA was studied. The equation for this model prediction is as follows:

979-8-3315-0993-4/25/$31.00 ©2025 IEEE

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad (1)$$

The assumptions of this model are that the explanatory variable observations must be independent of each other, the error term must be normally distributed with a mean of 0, and that the variance of the error term be constant. The advantage of this model is that the β values as well as the error term are directly calculated using the mean of the two classifications of the explanatory variable, making them highly accurate and an effective predictor for the model. I chose this model because it is a perfect fit for the explanatory variable of part-time job, as the data is either yes or no, which can be effectively converted to binary data. The python "t.cdf" function is used in this model within the "scipy.stats" module to gather a p value based on the test statistic t.

C. ANOVA Model

The third method in the analysis is an ANOVA model accompanied by an ANOVA table. This model analyzes the variance of different levels of an explanatory variable and how these levels impact the response variable. In the case of the data being investigated, the different classifications of Major are being studied to answer if there is a difference in GPA for different majors. There are multiple equations for this model involving Sum of Squares and Mean of Squares in order to get the desired test statistic.

$$SS_{group} = \sum_{i=1}^{k} n_i \alpha_i^2 \quad (2)$$

$$SS_{error} = \sum_{i=1}^{k} (n_i - 1) S_i^2 \quad (3)$$

$$SS_{total} = SS_{group} + SS_{error} \quad (4)$$

The degrees of freedom for each of the sum of squares equations are the number of groups – 1, the number of data entries – the number of groups, and the number of data entries – 1, respectively.

$$MS_{group} = SS_{group} / df_{group} \quad (5)$$

$$MS_{error} = SS_{error} / df_{error} \quad (6)$$

The assumptions of the ANOVA model are the same as those of the Linear Regression Model, with an independence of the data entries from one another, and a normally distributed error term with a mean of 0 and a constant variance. The advantage of using this model is being able to test the significance of a categorical explanatory variable such as Major classification with more than two distinct categories. The model was chosen for this section of analysis because there are five distinct classifications of major in the dataset, making ANOVA a great fit. The calculations in python for this model were almost all done without the use of built in functions or modules, however the "f.cdf" function within the "scipy.stats" module was used to obtain the p value from the f test statistic.

D. Multiple Linear Regression

The final method used in the investigation is the Multiple Linear Regression model for predicting the response variable based on multiple explanatory variables. In the case of the dataset, Study Hours, Attendance Rate, Extracurricular Activities, and Sleep Hours were all considered to predict college GPA.

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \quad (7)$$

The assumptions of this model are that the data points are independent from each other, the variance of the error terms are constant across all explanatory variables, and that the explanatory variables are not highly correlated with each other. The advantage of this model is that it can handle multiple predicting variables in order to give an accurate estimate of the response variable. This model was chosen because there is an area of interest in which additional variables can be impactful in predicting college GPA. In python, machine learning was used for this method, with the "train_test_split" and "LinearRegression" functions used from the "sklearn.model_selection" and "sklearn.linear_model" packages in python. In order to try to improve the model, a t test statistic was calculated for each of the explanatory variables in an attempt to see which variables were significant.

IV. Results

A. Pearson Correlation Coefficient Results

The Pearson Correlation Coefficient used to test the significance of high school GPA in determining college GPA was calculated to be 0.2958. This positive value indicates that as high school GPA increases, college GPA tends to increase. However, since the value is not close to 1, there is a lot of variability in the correlation meaning little consistency and predictability to college GPA increasing as high school GPA increases. The figure below can help to visualize these results.

FIGURE 1. GPA CORRELATION

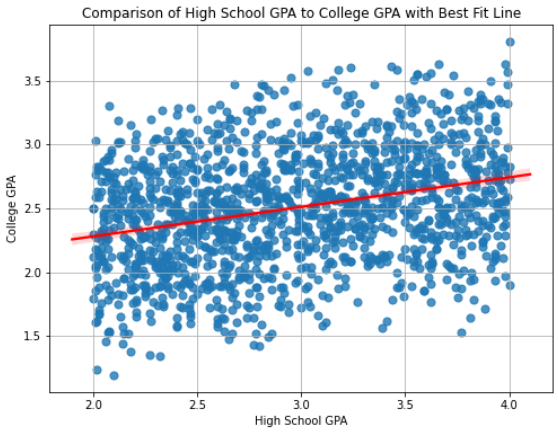

Fig. 1. The figure shows all the data points scatter with the relationship between the x axis variable High School GPA and y axis variable College GPA shown in the red line of best fit.

B. Linear Regression Results

The linear regression model for determining the significance of part-time job on college GPA first calculated the β values for Eq. (1) and obtaining the values of $\beta_0 = 2.518$ and $\beta_1 = -0.034$. The β_0 value indicates the mean college GPA of the no part-time job group, while the β_1 value indicates the difference between β_0 and the mean college GPA of the part-time job group. The error of each data point in the set was then calculated using these values, and the assumptions of the model were tested using the error term, finding a normal distribution of the error with a mean of 0 and a constant

variance. The t test statistic was then calculated and a p value of 0.15 was obtained, which is greater than the α value used of 0.05. This indicates that the test failed to reject the null hypothesis of $\beta_1 = 0$, meaning no significant evidence was found that college GPA of students with a part-time job is different than the GPA of students with no part-time job.

C. ANOVA Model Results

In the ANOVA model analyzing the significance of different major classifications on college GPA, a lot of calculations were involved using the equations shown in Section C of the Methodology Section. The assumptions of the model are the same as the linear regression model from Section B, so they have already proven to pass. First the grand mean, the mean GPA of the entire dataset, was calculated at 2.50. Next the data was broken into groups based on the major classifications of Science, Engineering, Arts, Business, and Undeclared, with the mean GPA of each group being calculated as 2.54, 2.53, 2.43, 2.50, and 2.52, respectively. Next the alpha values, the difference between the group mean and grand mean, were calculated to perform the calculations in the ANOVA table. The alpha values of 0.04, 0.03, -0.07, -0.0002, and 0.02 were found in the same order of classification as above. Finally, the calculations for the ANOVA table were performed as shown below.

TABLE I. ANOVA

ANOVA	**Source**				
	Degrees of Freedom (df)	***Sum of Squares (SS)***	***Mean Squares (MS)***	***Test Statistic (F)***	***P value***
Group	4	2.50	0.6244	3.1078	0.0147
Error	1468	294.92	0.2009	-	-
Total	1472	297.42	-	-	-

The test statistic for the ANOVA model is the F test statistic, so an F distribution was applied to obtain the p value of 0.0147. This p value is less than the alpha value of 0.05, meaning the null hypothesis that all alpha values are equal. This means that at least one of the alpha values is not equal, and when looking at the summary statistics and alpha values, the Arts major classification has a much lower mean GPA and alpha value than the others, signifying that this could be the classification that caused the ANOVA model to reject the null hypothesis, although there is no tested evidence to prove this.

D. Multiple Linear Regression Results

The multiple linear regression model used machine learning training and testing to create the proper model for determining college GPA based on Study Hours, Attendance Rate, Extracurricular Activities, and Sleep Hours. 80 percent of the data from the dataset was used to train the model, while the remaining 20 percent was the testing group. The model resulted in an R squared value of 0.4269, meaning that the model does not fit the data very well. With this result, the p values for each of the explanatory variables were obtained to determine which variables were significant in the model and which were not. Extracurricular Activities and Sleep Hours resulted in p values of 0.68 and 0.46, well over the alpha value of 0.05, meaning these variables were very insignificant in the model. The model was changed to only include the variables Study Hours and Attendance rate, and the process was done again with a resulting R squared value of 0.4286. This new R squared value was hardly higher than the old one, meaning that the absence of the two variables subtracted from the model did not change how well the model fits the data. These results were unsatisfactory overall, with none of the chosen explanatory variables being a highly significant fit in determining college GPA.

V. Discussion

The results from the Multiple Linear Regression model were unsatisfactory and ultimately could be greatly improved. In future work, the Multiple Linear Regression model could be run in stepwise fashion, adding or subtracting an explanatory variable each time to the model until a satisfactory R squared value is achieved. Another option would have been to simply choose different variables from the dataset, however for the purpose of this project the variables chosen were separate from variables already analyzed in different models, so in future work some of the other significant variables from other sections of the project could be involved.

VI. Conclusion

The important results found in the analysis of this project are that high school GPA is a significant indicator of college GPA, and that college GPA can differ between major classifications. Additionally, it was found that having a part-time job while being in college does not have a significant impact on GPA, and there is a lot of variability in Study Hours, Attendance, Extracurriculars, and Sleep Hours when determining college GPA. These results have a legitimate effect on the real world, as it is important to prove that successful academic standing in high school does translate to success in college, and it is important to take education seriously throughout the levels of education. It can also be said that when choosing a college major, some majors can be more rigorous or demanding, leading to variation in college GPA. Finally, having a part-time job as a full-time student does not have any significant effect on academics in terms of GPA, so students can use this information to decide whether they believe they can handle the responsibility of working and studying simultaneously.

References

[1] "How well does high school grade point average predict college performance by student urbanicity and timing of college entry? | IES," *Ed.gov*, 2017. https://ies.ed.gov/use-work/resource-library/report/descriptive-study/how-well-does-high-school-grade-point-average-predict-college-performance-student-urbanicity-and

[2] Peter Mushemi, "Dataset for predicting the College GPA of students," *Kaggle.com*, 2024. https://www.kaggle.com/datasets/petermushemi/dataset-for-predicting-the-college-gpa-of-students?select=Academic.csv (accessed Apr. 12, 2025).