What Factors Contribute to a College Freshman's GPA?

The Dog Lovers: Betsy Blitch, Enrique Latoni, Alex Shen, Shruthi Kumar

November 16, 2020

Introduction/Data

Our data set looks at college freshmen GPA, and how various factors such as high school GPA, race, and/or gender play into a student's first-year college GPA. Our data set consists of 219 students that were randomly selected as a subset of the full data from the original collection. The original data set comes from a larger stat2data package, which was intended to be used for a statistics textbook. In 1996, a professor at a midwestern college collected this data set to explore how various factors, such as high school academic performance, race and other situational factors affected first year academic performance in college. We do not have any additional information on further details for exactly how this professor acquired this data set. Here is the link to where we found the data set:

from this larger data dump: https://vincentarelbundock.github.io/Rdatasets/datasets.html

this page describes our particular data: https://vincentarelbundock.github.io/Rdatasets/doc/Stat2Data/FirstYearGPA.html

Description of Data Set and Reasons for Choosing

The data we are analyzing includes 9 possible predictor variables and 1 response variable, which is the first year college GPA (We had this approved with Professor Tackett). There are 219 observations (first year college students) in our data set. The quantitative predictor variables are high school GPA (HSGPA), verbal SAT score (SATV), math SAT score (SATM), number of credit hours in high school for humanities (HU), and number of credit hours in high school for social sciences (SS). The categorical predictor variables are gender (Male), whether a student is a first generation college student (FirstGen), whether the student is white or not (White), and whether the student attended a high school where over 50% of the students were also going to college (CollegeBound).

We chose to work with this data because we are curious as to what characteristics and factors about a particular student are influential for their success in their freshman year, as measured by First Year GPA. We find this interesting because we have all been freshmen in a diverse college filled with many first generation students, a wide range of high school backgrounds, and different backgrounds in general. This is also relevant in today's world because we have variables related to SAT score, and there is a lot of discussion/debate surrounding whether standardized testing is an accurate predictor of whether a student will succeed academically in college. Furthermore, many admissions offices may find this data set interesting, as their job is to make sure that the students they admit will succeed in a college environment.

The general research question we want to explore is what factors have the biggest impact in a freshman's performance in college, which is measured (in this case) by their first year GPA.

Data Manipulation

To set ourselves up for later modeling, we did some data manipulation to aid in this analysis. In the original data set, all of the categorical responses had answers of either 1 or 0. To make things easier, we renamed the categorical responses with what the 1 and 0 represented as an answer choice. This allows everyone - ourselves and our readers - to better understand our data when working with it. The next thing we did was change our

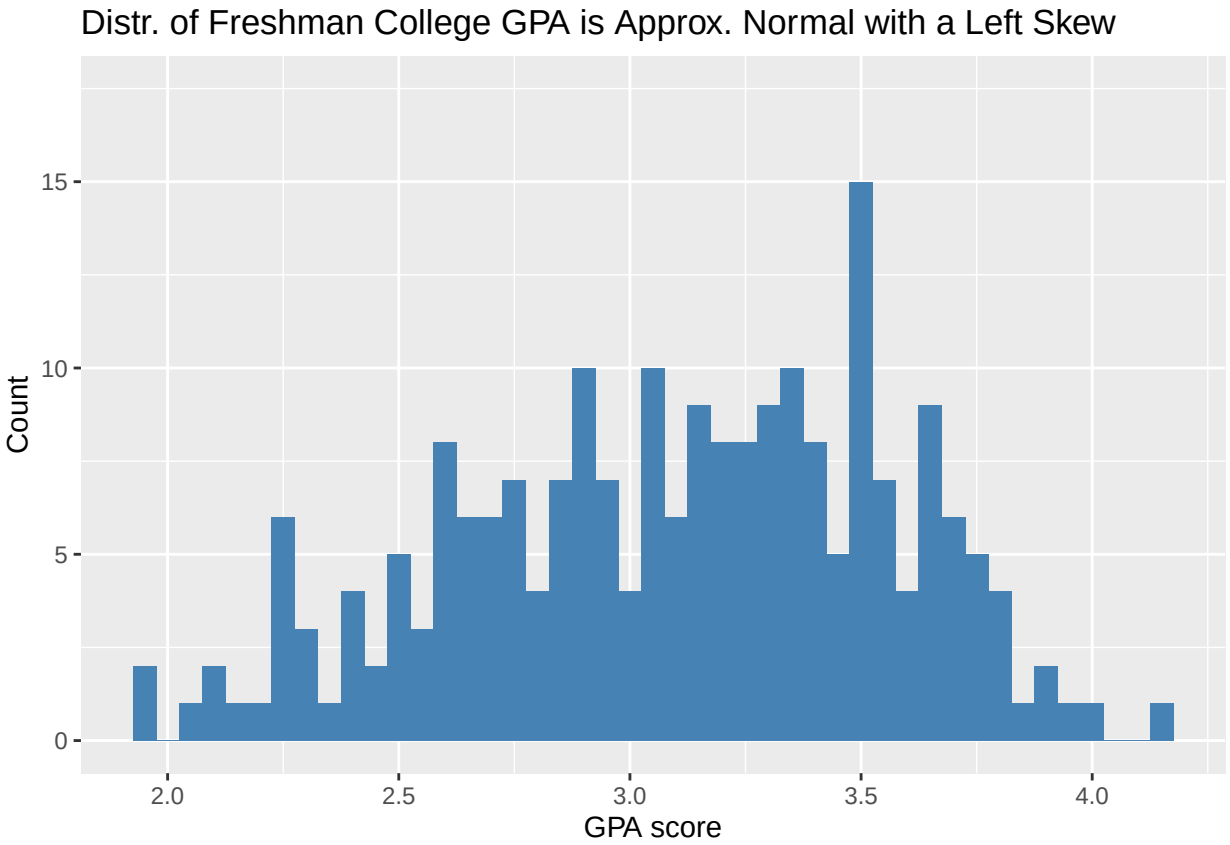
categorical predictor variable's data type to a factor for later modeling. And finally, we mean-centered our continuous predictor variables to allow for a more meaningful interpretation of the intercept in the model later on.

Hypotheses

We have several hypotheses for our data set. One of these is that there will be a linear, positive relationship between high school GPA and first-year college GPA. Furthermore, we hypothesize that SAT score - both verbal and math - will have positive, linear relationships with first-year college GPA. We also hypothesize that a student's race will have an effect on a first-year GPA, with the idea that White students will have higher freshman GPAs. Finally, we hypothesize that the relationship between humanities, as well as social science, credit hours in high school will have a positive, linear relationship with first-year college GPA because the distribution of credit hours might be skewed because of different high school requirements for graduation.

Exploratory Data Analysis

Let's start by looking at the distribution of College GPA scores.

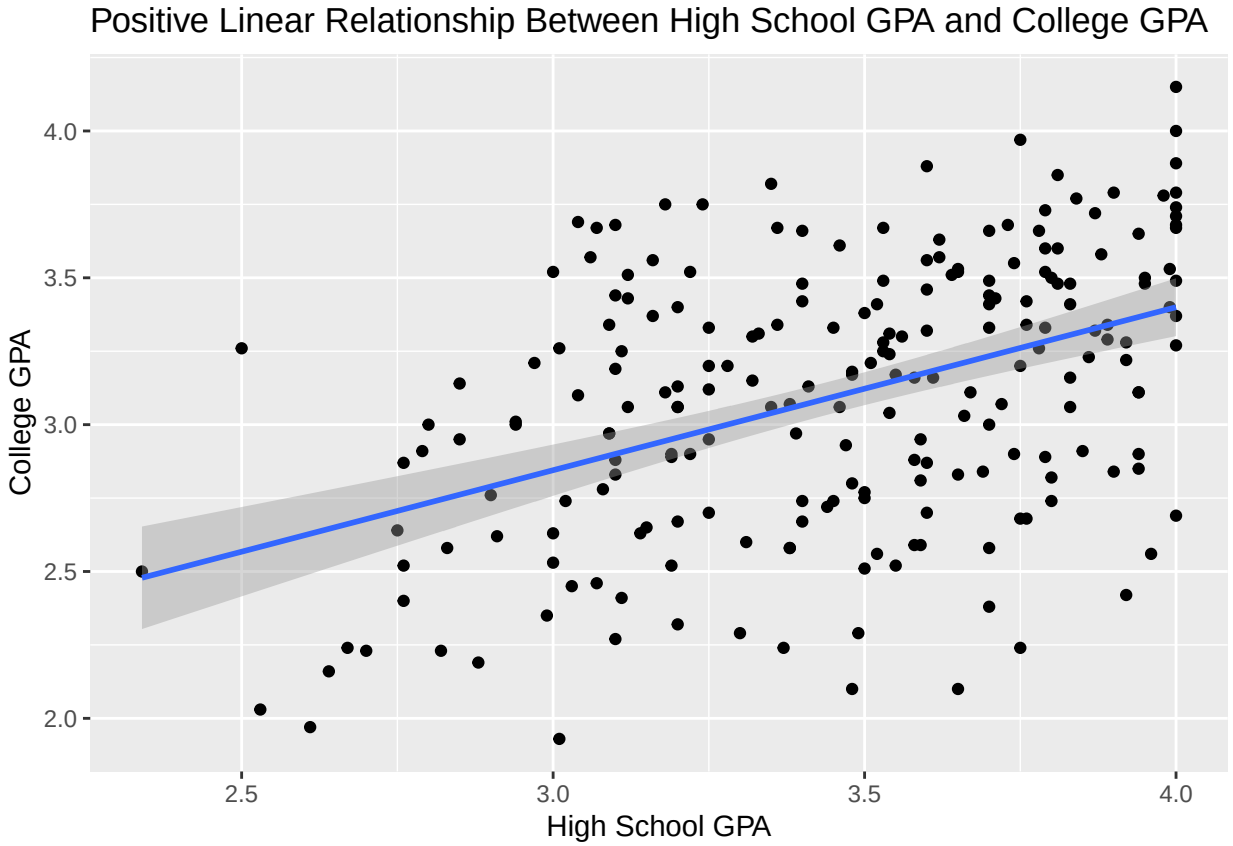


meangpa	mediangpa	sdgpa	q1gpa	q3gpa	iqrgpa	maxgpa	mingpa
3.096	3.15	0.465	2.745	3.48	0.735	4.15	1.93

From our visualization, we can see that the distribution of our response variable - GPA score - is slightly left skewed, but it still looks unimodal and relatively normal. We can confirm that it is left skewed because in our summary statistics, the mean GPA score of 3.096 is less than the median GPA score of 3.15. The mean and median are both measures of center, and even though they are relatively similar, it might be more appropriate to use the median because of this slight skew. The maximum value is 4.15 and the minimum

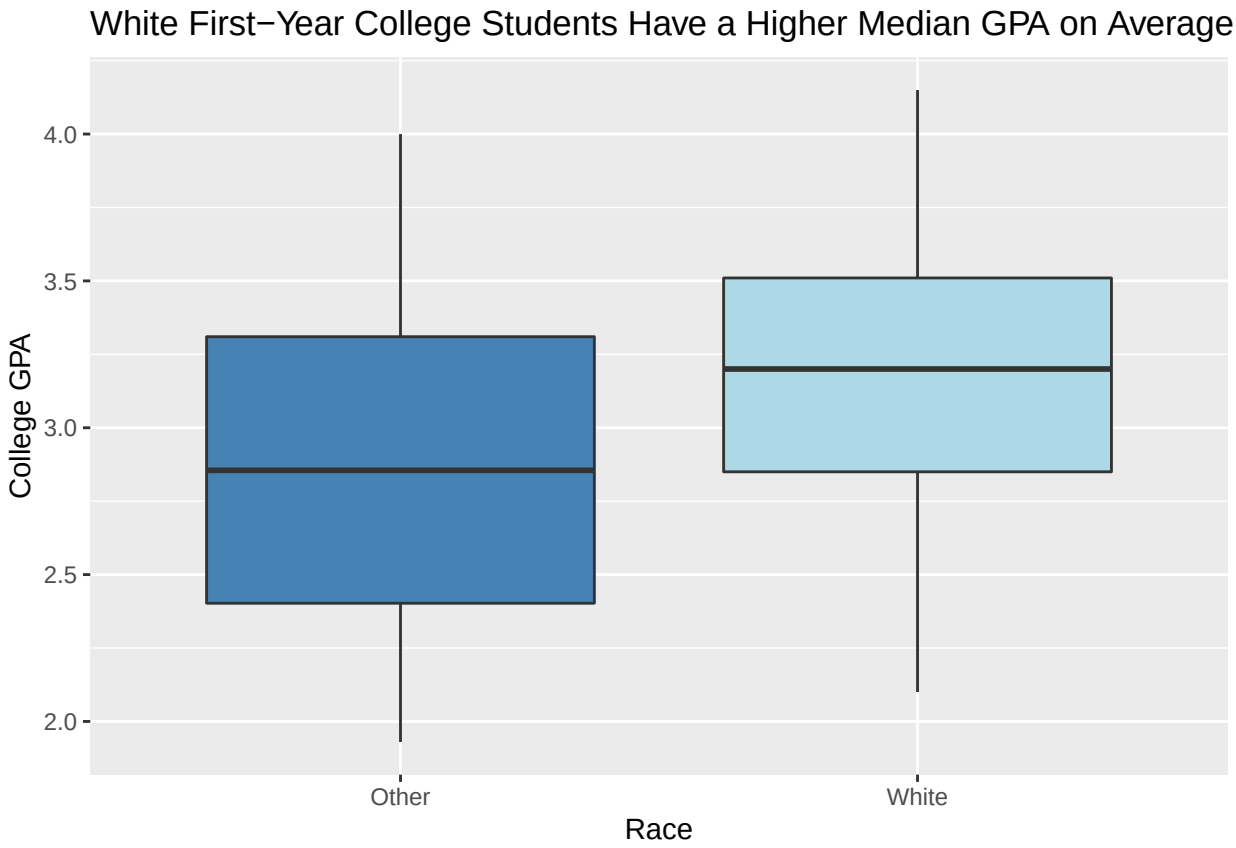
score is 1.93. It doesn't appear that there are any outliers. The standard deviation is 0.465 points of the GPA score, and the IQR is 0.735 (meaning that 50% of our data is between 2.745 and 3.48).

Since we hypothesized that High School GPA would probably be one of the most obvious predictors of College Freshman GPA, we decided to graph the 2 in relation to one another.



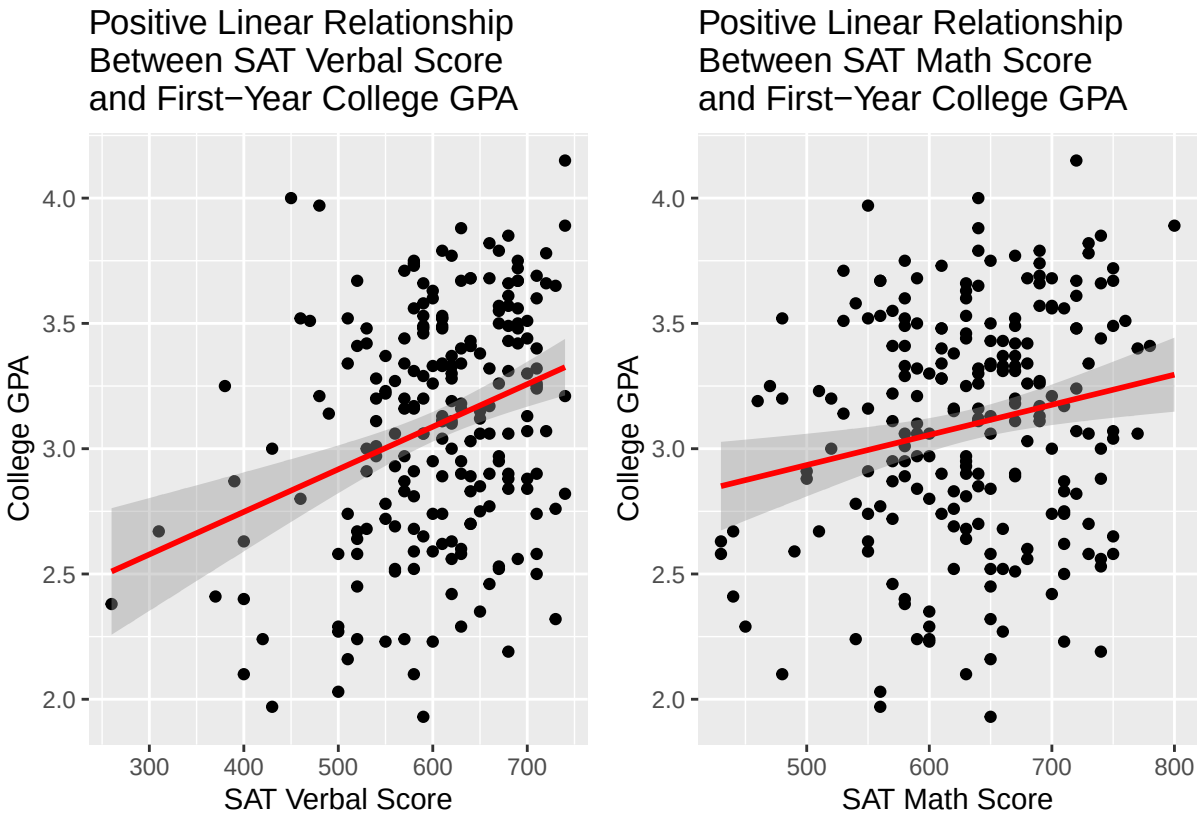
We can see from this initial scatterplot that as high school GPA increases, college freshman GPA increases as well. There seems to be a positive, linear relationship between the two.

We are also interested in whether race plays a role in College Freshman GPA score. Below is a boxplot that shows the relationship between the two.



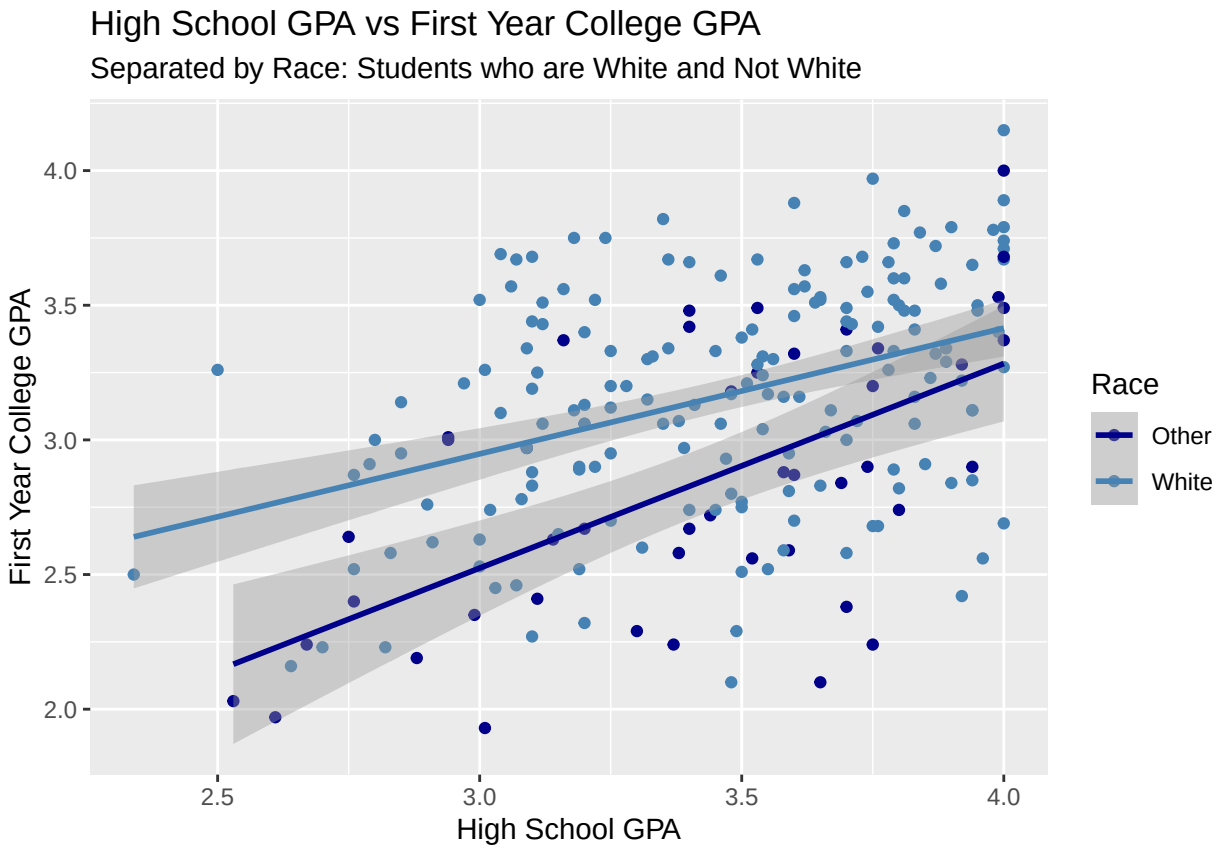
We can see from this boxplot that the median GPA, Q1 and Q3 scores for White students are higher than the median GPA score for non White students. Additionally, the max and min are higher for White students compared to non White students. However, there is a fair amount of overlap of the IQR between White and non White students' GPA.

Since SAT score is such a popular topic of discussion right now with Covid and many colleges are getting rid of this requirement for admission, let's explore the relationship between SAT score and Freshman GPA. We will create 2 plots - one for Verbal score and one for Math score.



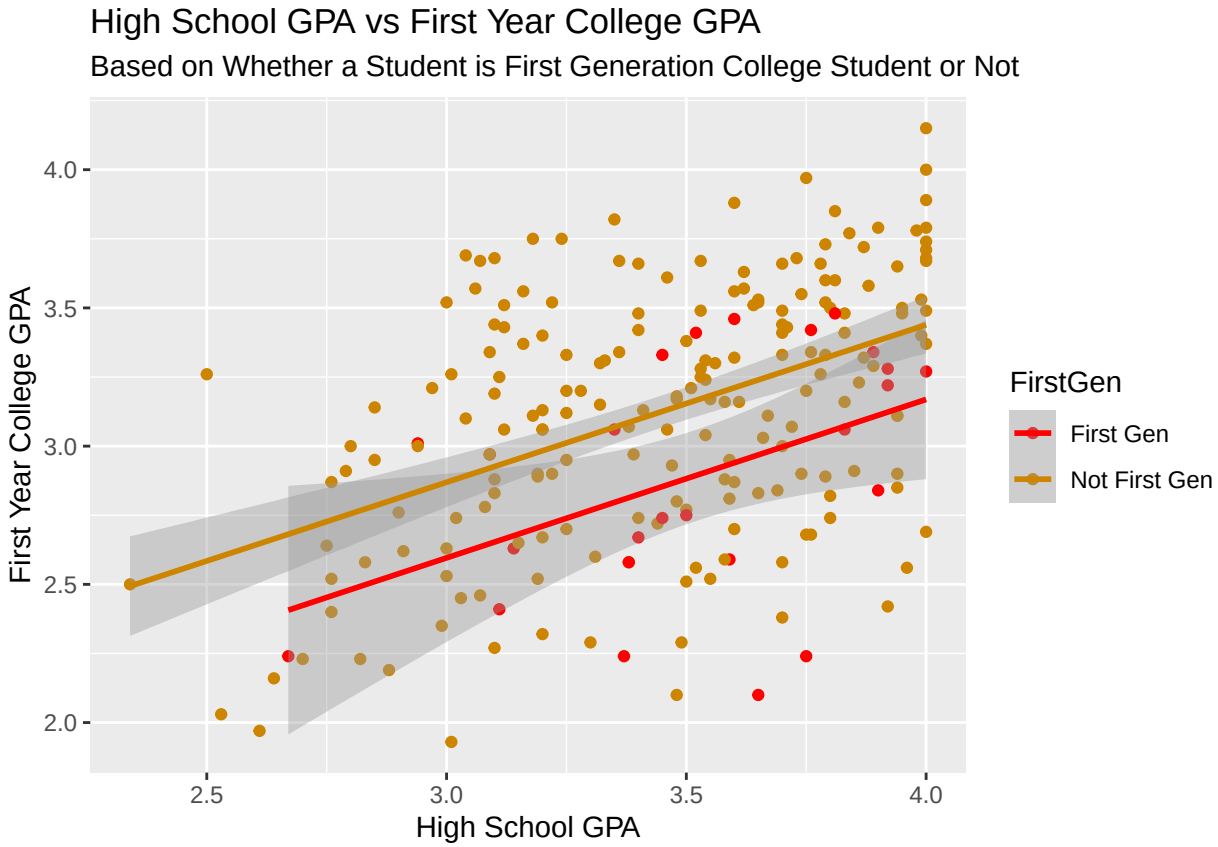
We can see from these scatter plots that both SAT scores have a positive linear relationship with the first year college GPA. However, it does appear that the line of best fit for the SAT verbal score has a larger slope than the line of best fit for the SAT math score.

Finally, to conclude our EDA, there are some interactions that we want to explore further. The first is whether race affects the relationship between high school GPA and first year college GPA. To do this, we will look at the regression lines for High School GPA and College GPA based on race.



We can see that the slope of the line of best fit is different for White vs. Other students. Therefore, we will include and assess this interaction term in our model selection process later.

The next interaction term we are interested in is whether being a first generation college student affects the relationship between high school GPA and first year college GPA. Again, we will look at the regression lines for High School GPA and College GPA based on race.



Interestingly, we can see that the slope of the line of best fit is approximately the same for First Generation vs. Not First Generation students. Therefore, we will not include and assess this interaction term in our model selection process later.

Methods

Since our research question focuses on what contributes to/influences First Year College GPA, we are going to create a regression model with College GPA as our response variable and the rest of our variables as predictors - both the quantitative and categorical included. We will choose multiple linear regression to do so since our response variable of Freshman GPA is quantitative and continuous. Additionally, as we mentioned earlier in the write-up, the quantitative predictor variables that we include in our full model are mean-centered to allow for a more meaningful interpretation.

Model Selection

To perform model selection, we created a full model with all possible predictors and an intercept only model. Then, using the step function, we will do backwards selection using AIC as the main criteria for selection. We want our final model to have the lowest AIC and BIC score, while also having the highest adjusted R squared value.

Here is our final selected model:

term	estimate	std.error	statistic	p.value
(Intercept)	2.933	0.060	48.889	0.000
SATV_cent	0.001	0.000	2.194	0.029
HSGPA_cent	0.474	0.071	6.682	0.000
HU_cent	0.017	0.004	4.385	0.000

term	estimate	std.error	statistic	p.value
SS_cent	0.008	0.005	1.424	0.156
WhiteWhite	0.206	0.069	3.004	0.003

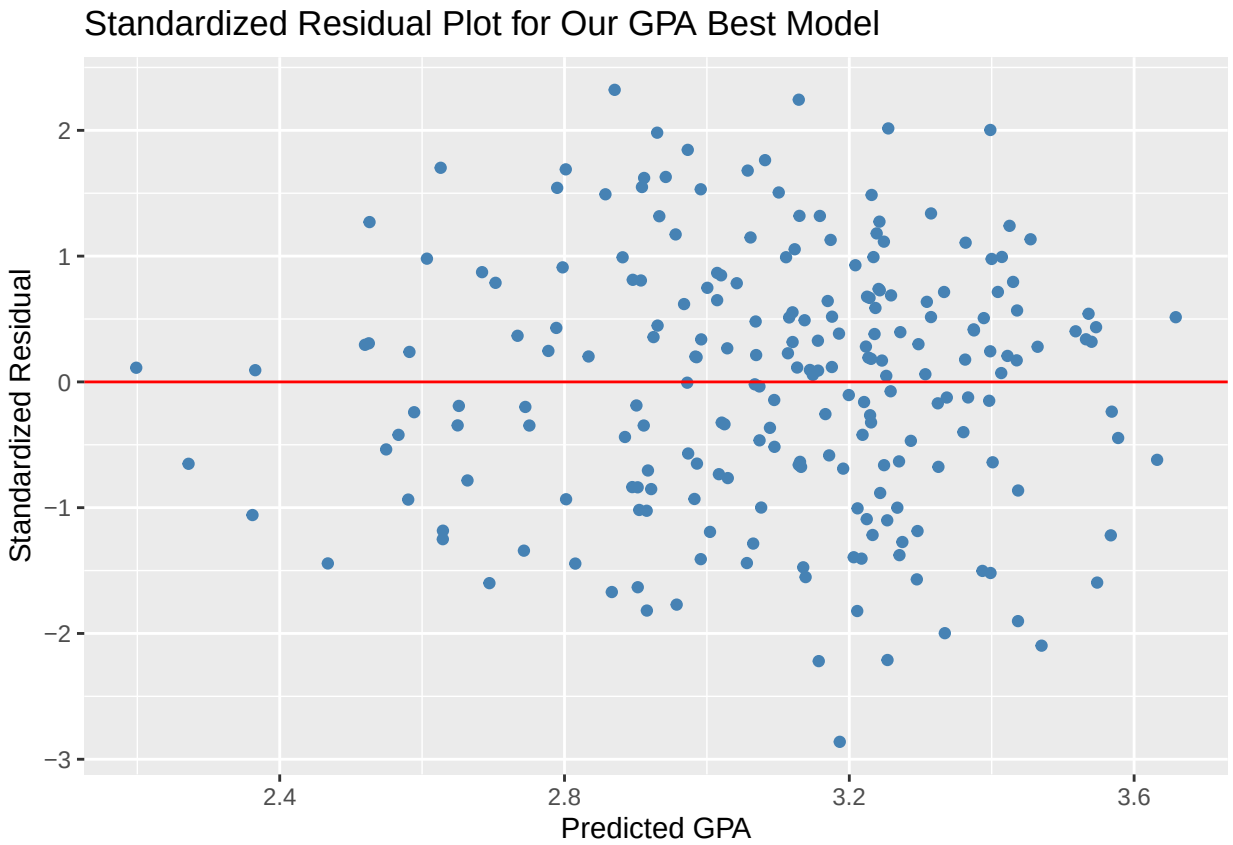
We will discuss the model further in the results section, but it is interesting that only SAT verbal score, and not the math score, is included. Furthermore, the interaction term between High school GPA and Race (connected to our EDA) is not included in our final model as well.

MLR Conditions

Now that we have our final model, we need to check for the conditions of multiple linear regression.

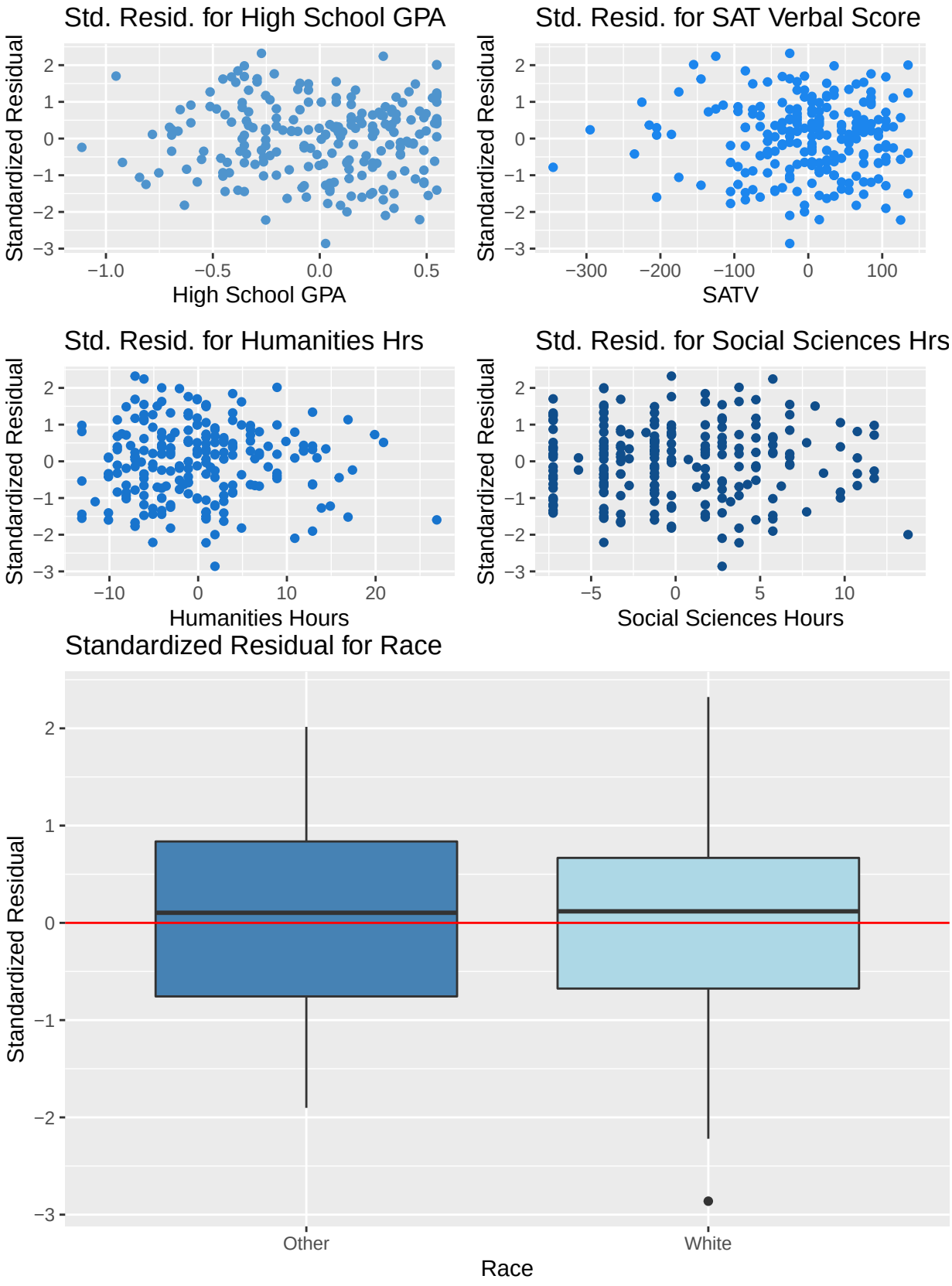
The first 2 conditions we will look at are linearity and constant variance. In order to check for linearity, we have to look at a graph of standardized residuals versus fitted values for the whole model, along with a graph of standardized residuals for all of the individual predictor variables. The first graph of standardized residuals for the whole model can also be used to look at constant variance.

Here is a scatterplot of our overall standardized residuals:



We can see from our standardized residual plot that there is no general pattern or trend in our residuals. Also, there seems to be about an equal number of residuals above and below 0. This confirms our linear model condition. Also, from this graph, we know that the constant variance condition is satisfied because the vertical spread of the residuals is relatively constant across the plot from left to right.

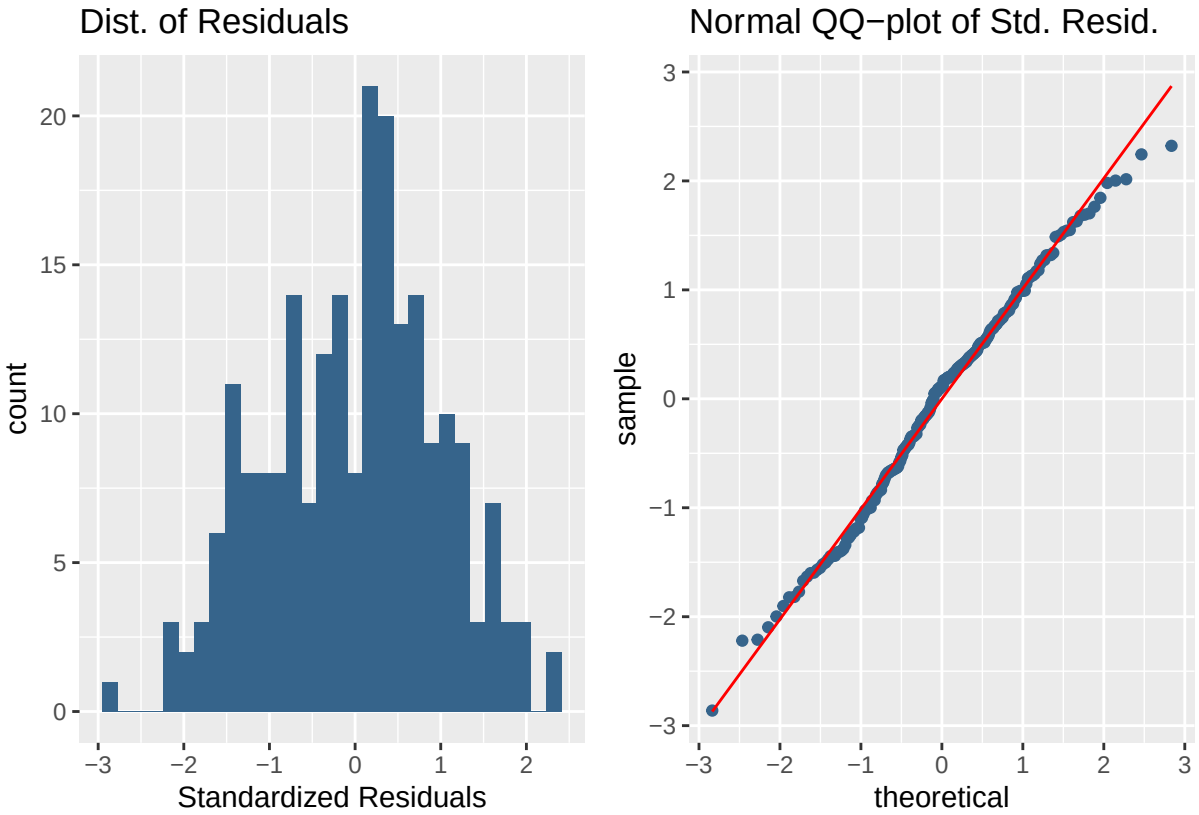
For the next part of linearity, we will check the standardized residual for each predictor variable:



As seen in the plots of the standardized residuals against each predictor variable, there are no distinguishable patterns. Because of this and the fact that there is no distinguishable pattern in the overall plot of the

standardized residuals and the predicted GPA, the linearity assumption is satisfied.

The next condition we will check is normality.



As seen in the graph of the standardized residuals on the left, the distribution is relatively normal. This is supported by the Normal QQ plot of standardized residuals, since the points generally follow the line. Therefore the normality assumption is satisfied.

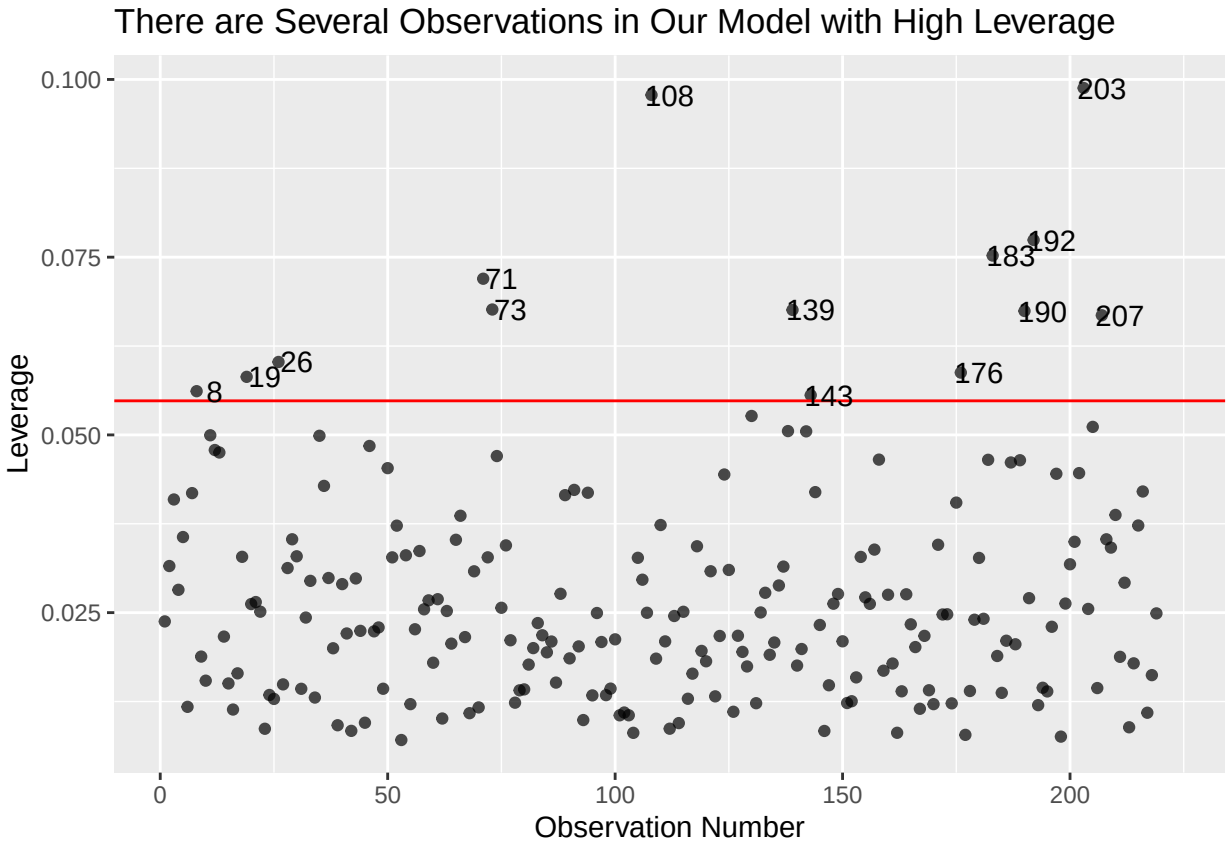
Finally, we will look at the independence condition. From what we know about the context of the data, there are no patterns or trends on how the data was collected. Additionally, our subset 219 students from the larger data set was random. Therefore, we conclude that the independence condition is satisfied.

MLR Diagnostics

Next, we will look further into our model to make sure we do not have any highly influential points in the model. Looking at leverage specifically:

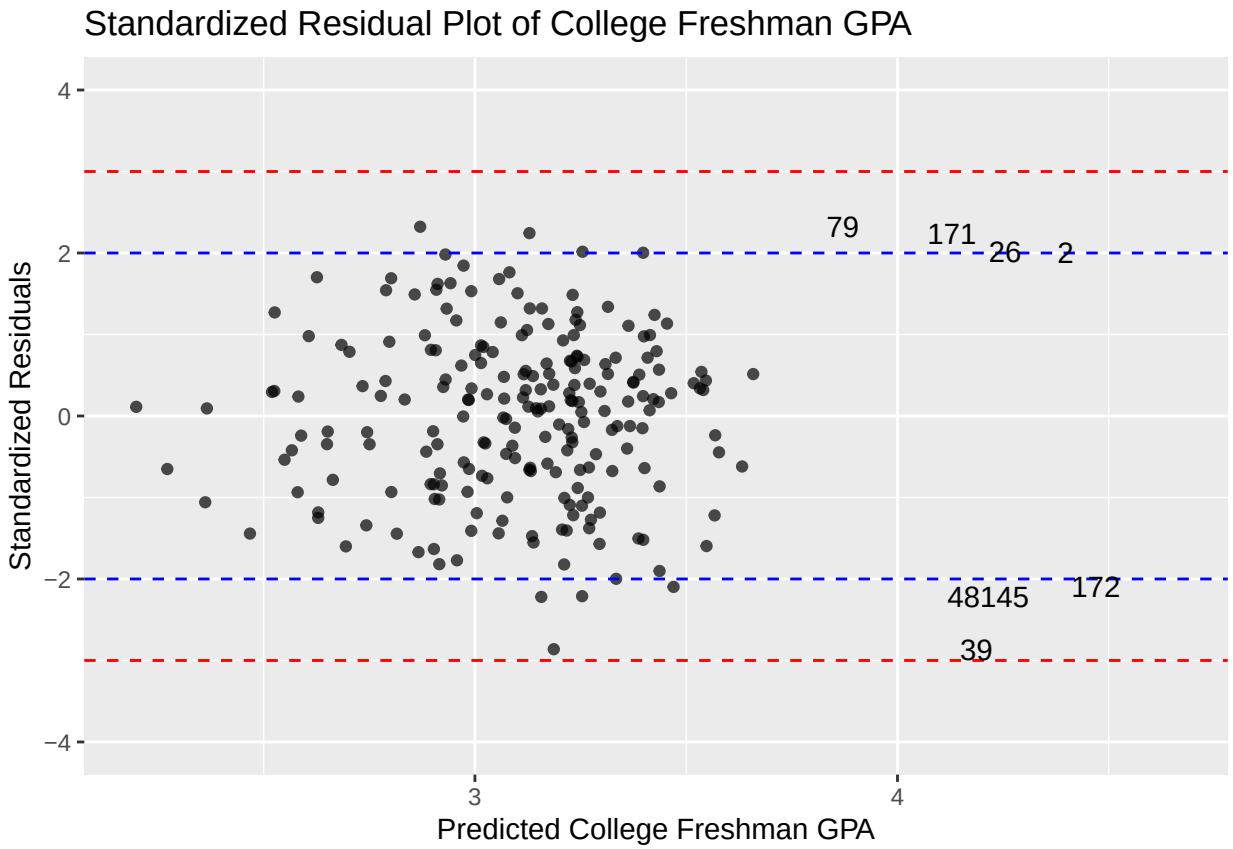
Our high leverage threshold is equal to 2 x (the number of predictors in the model + 1)/observations in the model.

```
## [1] 0.05479452
```



We can see from this graph that we have 14 observations that are considered to have high leverage.

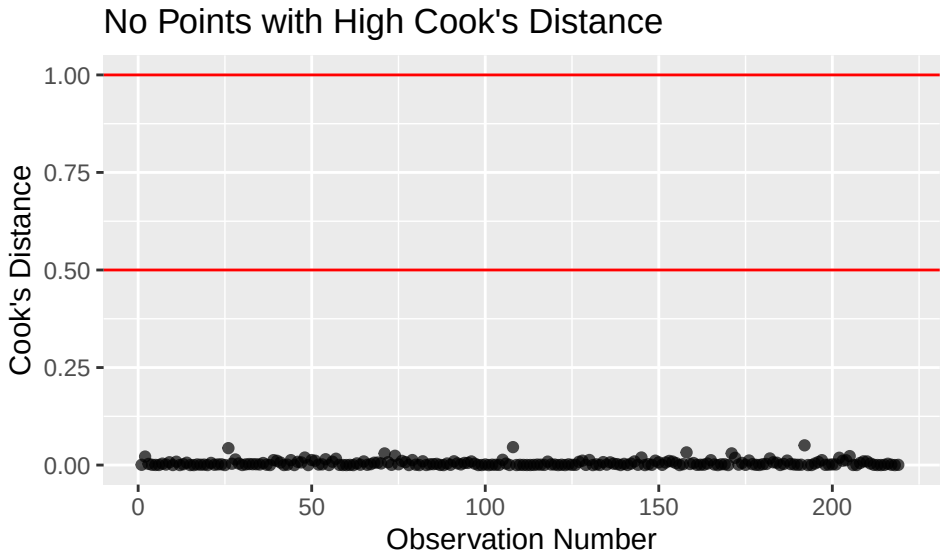
Continuing with our investigation of influential points in our model, here is a scatterplot of standard residuals versus predicted values:



GPA	.fitted	.resid	.std.resid	.hat
4.15	3.398	0.752	2.003	0.032
4.00	3.255	0.745	2.015	0.060
2.10	3.187	-1.087	-2.862	0.009
2.32	3.157	-0.837	-2.220	0.023
3.75	2.870	0.880	2.322	0.014
2.42	3.254	-0.834	-2.211	0.023
3.97	3.129	0.841	2.244	0.035
2.68	3.470	-0.790	-2.097	0.025

We can see from our graph that none of our observations have a standardized residual above absolute value of 3, but we have 8 observations that have standardized residuals above the value of 2. These are considered moderate outliers.

Finally, let's conclude our investigation of influential points by looking at a graph showing Cook's Distance.



This graph shows us that there are no observations in our model with a high Cook's Distance. Therefore, even though we did have some observations with high leverage and standardized residuals, we can conclude that we do not have any influential points in our model. As a result, we will keep our model as is - with all the observations included.

Detecting multicollinearity

Finally, we want to test for multicollinearity in our model. The reason we want to avoid multicollinearity in the model is because this would mean that we have highly correlated predictor variables in our model, thus reducing the accuracy of the coefficients in the model. We will use the VIF funtion to do so.

names	x
SATV_cent	1.211
HSGPA_cent	1.059
HU_cent	1.140
SS_cent	1.108
WhiteWhite	1.175

Since all of the values are below 10, there is no apparent multicollinearity between our predictor variables. This is great because it allows us to keep our model as is and we do not need to take any points out of our data set.

Results

As shown in the model output from the Methods section, the equation for our final model is GPA = 2.933 + 0.474 * HSGPA_cent + 0.001 * SATV_cent + 0.017 * HU_cent + 0.008 * SS_cent + 0.206 * WhiteWhite. In the Methods section, we also confirmed our conditions for our model by checking for linearity, normality, independence and constant variance. Then, we checked to make sure there were no influential points in our model by looking at leverage, standardized residuals, and Cook's distance. Based off of that analysis, we concluded that conditions and diagnostics are met for MLR.

Model Selection Decision

To come up with this model, we did backward selection with AIC as our selection criteria. Our original full model included all of our variable models along with the interaction between race and high school GPA

(mean-centered), but after model selection, our best model did not include the interaction term. Also, we noticed that the variable SS (number of credit hours for social science courses) had a p-value (0.156) that is relatively large, but we looked at the multicollinearity of our model and didn't find any problems. We also used AIC as the selection criteria, and the generated model included the SS variable as part of the output, so despite the large p-value, we will continue the analysis with this variable included in the model.

For our model, we used MLR because we have a quantitative response variable (Freshman College GPA) and we want to know about the relationships each of the different predictor variables have with Freshmen College GPA. After completing the multiple linear regression, the analysis is as follows:

Our intercept tells us that for a non-white student with the average high school GPA score (which is 3.45), the average verbal SAT score (which is 605.07), the average amount of credit hours for humanities courses (which is 13.11), and the average amount of credit hours for social science courses (which is 7.25), we expect the first-year GPA to be on average 2.933. This interpretation of the baseline intercept in our model is both interesting and relevant because this could be used in admissions offices to better understand how the "average" student (that is non-white) will perform in their freshman, if performance is measured by their GPA.

Analysis of Research Question and Hypotheses Using Model Results

Our main research question centered around which of the predictor variables have a statistically significant relationship with first-year GPA. Based off of our model, we can see that high school GPA (HSGPA_cent), SATV score (SATV_cent), Humanities credit hours (HU_cent), Social Science credit hours (SS_cent), and race (WhiteWhite) provide the best prediction of freshman GPA. Since the goal of our analysis is explanation, we will interpret a few of these variables that we think are most relevant to the reader and to the overall research question.

High School GPA

One of our hypotheses was that there will be a linear positive relationship between high school GPA and first-year college GPA. Our p-value for high school GPA is 0 so we can conclude that there is a relationship between high school GPA and first-year-college GPA. For every .1 point increase in high school GPA, we expect first-year college GPA for a student to increase by .0474, holding all other variables constant. This is important for admissions offices to understand because it shows that high school GPA is a realistic prediction for how a student may perform in their freshman year of college. If a school is wanting their freshman to have particularly high GPA scores, then they would want to pick students with higher high school GPA scores.

SAT Scores

We also hypothesized that verbal and math SAT scores will have positive, linear relationships with first-year college GPA. Again, the SAT Math score was not included in the final model, but based off our initial graphs in the introduction, there does appear to be a relationship. The p-value (0.029) for SAT Verbal, however, is low so we can conclude that there is a significant relationship between SAT verbal scores and first-year college GPA. For every 1 point increase in the SAT verbal score, we expect first-year college GPA for a student to increase by 0.001, holding all other variables constant. While the coefficient of 0.001 is very small, this makes sense because the SAT verbal section is scored out of 800, while GPA is scored out of 4.0. Another way to think about this is that for every 100 point increase in the SAT verbal section, a first year GPA score is expected to increase by 0.1. This is important to talk about because while there is a lot of pushback on submitting an SAT score right now for college admissions, especially with the difference in opportunity availability with Covid, our model does show that knowing a student's SAT verbal score can help in predicting how they will perform their freshman year. Admissions offices may want to look further into this relationship.

Race

We also wanted to look into race and how that plays a role in predicting Freshman GPA. We can see that p-value for WhiteWhite is low (0.003) so we can conclude that there is a significant relationship between race and first-year GPA. We expect the first-year GPA to be 0.206 higher for students that are white as compared to those who are not white. This is particularly interesting when applying our findings to questions of equality in education and how our backgrounds shape our success in different situations. If an admissions

office was reading this, they might want to better understand how they can support students who are not white to provide them an equal opportunity of success as a freshman.

Credit Hours

Finally, we hypothesized that the relationship between humanities, as well as social science, credit hours in high school will have a positive, linear relationship with first-year college GPA. The p-value for HU is low (0.000) so we can conclude that there is a significant relationship between humanities credit hours and first-year GPA. For every 1 credit hour increase in the number of humanities courses taken, we expect first-year college GPA for a student to increase by 0.017, holding all other variables constant. However, for SS (social sciences courses), the p-value is large, so we cannot conclude that there is a significant relationship between the number of social science credit hours and first-year GPA. However, if we look at the AIC, it is still considered a good predictor for first-year GPA. As discussed before, there are no issues with multicollinearity so we still include this in our final model.

Discussion

Based off of the data, the model that best predicts first-year college GPA is GPA = 2.933 + 0.474 * HSGPA_cent + 0.001 * SATV_cent + 0.017 * HU_cent + 0.008 * SS_cent + 0.206 * WhiteWhite. Thus, we can conclude that high school GPA, SAT verbal scores, number of humanities credit hours, number of social science credit hours, and having a race of white are the best variables to predict a student's first-year college GPA. However, we acknowledge that there are some limitations to our final model. Since our data is from over 20 years ago, the variables that affected GPA back then may not be exactly these same ones that affect first-year GPA now. So while the results of our analysis might have some impact today, they should also be looked at skeptically given that the data is old. The data that we had was also from one specific college in the Midwest so this may not have been the most representative sample for college across the US. Again, the results of the analysis might give us insight into how different factors influence first-year GPA, but given the limited sample space, further analysis may be necessary. We also do not know how the details of how the data was collected. This could make our data less valid because while we assumed that independence was satisfied, we don't know with 100% certainty that the samples were collected independently or randomly. Additionally, GPA distributions for different majors and different fields could be vastly different depending on the difficulty of the courses. For example, GPA for Pratt Engineering at Duke is usually lower than GPA for Trinity. Our data does not necessarily take this into consideration.

Future Work

If we could do the analysis differently, we would have picked a data set that is more representative of the status quo now. For example, we would first want to start with a data set that has newer data so our analysis is more relevant to the present day. We would also want data that is more representative of all colleges rather than just one midwestern college. This would entail our data set having samples/observations from multiple different colleges across the country. We also want to include some way to account for the difficulty of a class (maybe by adding a weighted term or adding a rating of how hard a class is).

In terms of the data analysis, if we could take this project further or even start over, we would definitely want to include more interaction terms. This would allow us to see whether or not certain variables like gender, being first generation, or going to a college bound high school have an effect on other predictor variables. Doing so would hopefully allow us to draw further conclusions about the data. Also, we would like to more thoroughly examine the model selection using different backwards selection criteria (BIC and adjusted R^2) and compare the outputs. We also might even want to try forward selection as well.

Lastly, our primary objective for this project was explanation, so we discussed the coefficients for the variables in our model. However, we could use the model for prediction, and it would be interesting to check the accuracy of our predictions of first year college GPA.