GPA PREDICTOR PROJECT

Dagmawi Zerihun, Khushi Mohapatra

Denison University, Department of Economics

ECON 307-01 - Intro to Econometrics

Dr. Ngwunui Azenui

May 5, 2023

I. Before estimation:

In an attempt to derive the "best" model, as our first guess, we choose the following variables, in addition to the SAT score when predicting one's college GPA. The table provides theoretical justification for each of the variables and indicates the expected influence (sign) that each one has on GPA.

As college students, we have insight through our personal journey at Denison that there is a distinct difference between the high school and college academic experience, which explains that there might be some differences in the factors that influence both levels' GPAs. In this project, we would like to explore how a student's distribution and management of time on campus influence their GPA, which can be captured by the following variables. Apart from the variables that have an obvious impact on COLLGPA (like SAT, HSGPA, DECISION, etc.), the rest of the variables (like HRSTD, DOUBLE, SLEEP, GREEK, FRIENDS, ATHLETE, etc.) indicate how a student chooses to distribute their time.

No.	Variable Name	Variable Definition and Type	Theoretical Justification	Expected influence (sign) on COLLGPA
	COLLGPA	Overall college grade point average Type: Continuous, Dependent variable		
1.	**SAT**	Aggregate math and critical reading SAT score - out of a maximum score of 1600 (note for those with only ACT scores multiply by 44.5) Type: Continuous, Independent variable	On average, students who score higher on standardized tests like the SAT are likely to have better academic skills and be better prepared for college coursework and therefore, SAT score will impact their GPA score.	**Positive** On average, If student has a higher SAT score, they are more likely to have better academic skills and attain a higher college GPA.
2.	**HSGPA**	Overall high school grade point average (leave as high school calculated) Type: Continuous, independent variable	High school GPA is a strong predictor of college GPA, as it reflects a student's academic performance and motivation.	**Positive** On average, if a student has a strong academic standing (high school GPA) in high school, they are likely to continue upholding their academic values and skills in college and attain a high GPA.
3.	**SAT_HSGPA**	This is an interaction term between SAT score and high school GPA. It helps capture the combined effect of these	The interaction between SAT and HSGPA allows for the possibility of synergy between these two variables.	**Positive** If the student has a higher SAT score and a higher

		two variables on college GPA, which means that the impact of each of these predictors on GPA is dependent on the level of the other predictor. Type: Interaction term, Independent variable	Students who excel in both areas may perform even better in college. This interaction term can help improve the accuracy of the model by accounting for potential nonlinearities in the relationship between the predictor variables and the response variable	high school GPA, they are more likely to have a higher COLLGPA as the individual variables have a positive impact.
4.	**HRSTD**	Average number of hours spent studying per week (0-168 hours) Type: Continuous, Independent variable	Spending more time studying has an obvious influence on the GPA obtained in college, as it provides more opportunities to learn and practice.	**Positive** On average, a student studying more hours will result in a higher grade on assignments which will result in a higher GPA.
5.	**HRSTD_SQ**	This is the interaction term between the average number of hours spent studying per week and itself (squared). It captures the non-linear relationship between study time and college GPA. Type: Interaction term, Independent variable	The relationship between studying time and GPA may not be linear, and there may be diminishing returns to additional study time. Squaring this variable allows for a curved relationship.	**Positive (possibly negative after an optimal level of study time)** On average, the effect of study time on GPA is expected to increase at a decreasing rate.
6.	**HREXTRA**	Average number of hours spent in extracurricular activities per week (clubs, meetings, job, athletics, community service, etc.) Type: Continuous, Independent variable	Participating in extracurricular activities can enhance a student's skills, knowledge, and social networks, which will most likely have an effect on academic performance, thus on the COLLGPA.	Unsure because participating in extracurricular activities can enhance a student's skills and can help the student’s learning, however, it can be argued that a student might perform badly in class due to focusing too much on extracurricular activities. Therefore, the direction of impact is uncertain.
7.	**DOUBLE**	Double Major (1=Yes/0=No)	Pursuing a double major indicates academic ambition and may indicate a student's	**Positive**

		Type: Dummy, Independent Variable Base group: Double = 0 (Student is not a double major)	ability to handle a heavier workload and showcase that the student prioritizes academics. Therefore, whether or not a student is a double major impact COLLGPA	On average, if a student is a double major, the student will be more motivated to tackle a heavier workload and perform better in classes because of skills like time management and motivation to study. Therefore, there is a positive impact on COLLGPA if a student is a double major, compared to if they only have one major.
8.	**FEMALE**	Gender is FEMALE (1=Female/0=Male) Type: Dummy Independent variable Base group: Gender is Male	This is a relevant variable because of the existence of gender-based discrimination, which can affect academic performance and opportunities, therefore, have an impact on COLLGPA	**Positive** On average, women tend to perform better academically compared to men, which may be due to a range of factors such as motivation, study habits, and social support, compared to students who identify as men.
9.	**SLEEP**	Average hours sleep per week (0-168 hours) Type: Continuous, Independent variable	This is a relevant variable because sleep is essential for cognitive function and academic performance and has an influence on the ability to learn effectively, so we can state that it would have an impact on COLLGPA.	Unsure because adequate sleep is essential for cognitive function and academic performance, and students who get more sleep may have better GPAs, it can be argued that sleeping too much might be a sign of neglecting studies and possibly signs of bad mental health. Therefore, the direction of impact on COLLGPA is uncertain.
10.	**INTERNATIONAL**	International Student (1=Yes/0=No) Type: Dummy Independent	International students face unique challenges in adjusting to a new academic and cultural environment,	**Negative** On average, being an international student

		variable Base group: Student is not an International student, but a domestic student.	which may have an impact on their ability to study and perform well in class, therefore this variable has an impact on the COLLGPA.	creates potential language barriers and cultural differences that may create additional challenges in adjusting to the academic environment and academic demands of the new country. Additionally, international students may face additional financial and social pressures that can impact their academic performance, causing a negative impact on COLLGPA, compared to domestic students.
11.	**INTERNATIONAL_DOUBLE**	This is the interaction between being an international student and having a double major. It captures the additional impact on GPA that having a double major has for international students, beyond the effect for domestic students. Type: Dummy interaction, independent variable Base: Domestic students with one major	International students who pursue a double major may face even greater challenges and workload because they have more barriers to overcome in terms of adjusting to a new culture and place, on top of the additional workload of being a double major.	Unsure because being an international student may bring unique challenges such as language barriers and cultural differences that can affect academic performance, but at the same time, being a double major may demonstrate a high level of commitment and drive, which can have a positive effect on GPA. Therefore, the net effect of international_double on COLLGPA is difficult to predict, compared to the students who are domestic with only one major.
12.	**SCHOLAR**	Have ever received scholarship money (1=Yes/0=No) Type: Dummy Independent variable Base group: Student has not	This variable is relevant because receiving scholarship money can reduce financial stress and allow students to focus more on their studies, which may have an impact on COLLGPA.	**Positive** On average, Receiving scholarship money can reduce financial stress and allow students to focus more on their studies, which may have

		received scholarship		a positive impact on COLLGPA. Additionally, scholarship programs often have academic requirements that students must meet to maintain their eligibility, which can serve as an additional incentive to perform well in their classes
13.	**GREEK**	Greek affiliation (1=Yes/0=No) Type: Dummy Independent variable Base group: Students who are not affiliated with Greek life.	This variable is relevant because greek affiliation can provide social support and opportunities for leadership that require time out of a student's schedule. Therefore, there are two ways it can impact COLLGPA.	Unsure because being part of a Greek organization can provide a supportive social network and access to academic resources causing a positive impact, however, it can also lead to distractions and time commitments that can negatively impact academic performance. Therefore, the overall impact of this variable on COLLGPA is uncertain and requires further analysis, compared to students not affiliated with Greek life.
14.	**DRINK**	Average number of alcoholic drinks per week (check the one that applies): 0 (Don't Drink) 1 (1-7) 8 (8-14) 15 (15-21) 22 (22-28(29 (29 and above) Type: Continuous, independent variable	Alcohol use can impair cognitive function and negatively impact academic performance, therefore have an impact on a student's COLLGPA	**Negative** On average, excessive alcohol consumption can have a detrimental effect on cognitive functioning, attention, and motivation, all of which are important for academic success. Therefore, this variable is expected to have a negative impact on the COLLGPA of a student.
15.	**GREEK_DRINK**	This is the interaction between being in a Greek	This variable is relevant and important for this model	Negative

		organization and the average number of alcoholic drinks consumed per week. It captures the additional impact on GPA that alcohol consumption has for Greek organization members, beyond the effect for non-members. Type: Interaction term, Independent variable	because the effects of alcohol consumption on academic performance may be different for students who are part of a Greek organization compared to those who are not. This can help to provide a more accurate prediction of COLLGPA.	On average, alcohol consumption can have negative effects on academic performance, and being part of a Greek organization can lead to social pressures and expectations that may also have a negative impact on COLLGPA. Therefore, the net impact is expected to be negative.
16.	**ATHLETE**	Intercollegiate Athlete (1=Yes/0=No) Type: Dummy Independent variable Base group: Student is not an athlete	Being an athlete changes a student's priorities and the amount of time they can spend studying so there is an influence of the COLLGPA	Unsure what direction the impact on COLLGPA would be because athletes may face unique challenges in balancing academic and athletic commitments in terms of time management, but may also benefit from discipline and time management skills, compared to students who are not athletes.
17.	**FRIEND**	The percent of the time while in college with a significant other Type: Continuous, independent variable	Having a significant other or a really close friend means the student chooses to spend a certain number of hours with them, indicating they choose to not spend a certain number of hours studying, therefore, there is an impact on the COLLGPA	Unsure what direction the impact on COLLGPA would be because having a significant other or really close friend can provide emotional support, but may also distract from academic responsibilities.
18.	**DECISION**	Apply to Denison using Early Decision (1=Yes/0=No) Type: Dummy Independent variable Base group: Student did not apply to Denison using Early Decision, but applied with Regular admission process	Applying to college through early decision indicates a stronger commitment to attending that school and may lead to a better fit between the student and the institution, which has an impact on COLLGPA, as the student is committed to the course.	**Positive** On average, by applying early, a student indicates that a university is their top choice and they are more likely to attend if accepted. This increased likelihood of attending can result in higher

				motivation and engagement, leading to a positive effect on COLLGPA, compared to students who did not apply with Early Decision, but applied with the Regular admission process.
19.	**DECISION_SCHOLAR**	This is the interaction between applying to Denison using Early Decision and having ever received scholarship money. It captures the additional impact on GPA that applying Early Decision has for students who have received scholarship money, beyond the effect for students who have not received scholarship money. Type: Interaction term, Independent variable	This interaction variable will capture the additional impact on COLLGPA that applying Early Decision has for students who have received scholarship money, beyond the effect for students who have not received scholarship money. This makes sense because the early decision applicant will indicate a stronger commitment to schoolwork and the same goes for a student who receives a scholarship (especially if the scholarship has minimum GPA requirements)	**Positive** On average if a student applied to college through early decision, they are highly committed to attending and, if the student has also received a scholarship, it may provide them with additional motivation to perform well academically in order to maintain the scholarship. Therefore, there will be a net positive impact on COLLGPA.
20	**HELP**	Average number of hours per week obtained help outside of class (tutors, professors, study sessions) Type: Continuous, independent variable	Obtaining help outside of class has an impact on a student's understanding of course material and therefore, has an impact on COLLGPA.	**Positive** On average, a student is expected to have a higher GPA in college if the student seeks additional hours of help to improve their understanding of the material.

Why the following variables were not chosen for this initial model:
1. **MAJOR1** and **MAJOR2**: These variables were not included because they seemed redundant with the DOUBLE variable, which indicates whether the student had a double major, for this model.

2. **HIGHSCH**: This variable was not included because the focus of the analysis was on college performance, rather than high school characteristics and there seemed to be a weak link between this independent variable and COLLGPA within the context of our project as we focus on the different activities and distribution of a student's time during college.

3. **PRESCHOOL**: This variable is not included because it is unlikely to have a significant effect on college GPA.

4. **ETHNIC**: This variable has not been included because it is difficult to interpret the effect of ethnicity on GPA without specific research or theory about how ethnicity would be related to GPA. There will be many biases in play if we focus on this relationship such as sampling bias.

5. **STATUS**: This variable has not been included because it is unclear how class status (sophomore, junior, senior) would be related to GPA, especially since the effect might be explained by other factors like age or credit hours completed.

6. **REGION**: This variable might not have been included because it is unclear how home region would be related to GPA, especially since the effect might be explained by other factors like college quality or availability of resources. Also there is a bias as the sample is from a university in the Midwest, so it is more likely that there are more people from this region than others.

7. **PAYMENT**: This variable has not been included because it is unclear how the percentage of college expenses contributed by the student would be related to GPA, especially since the effect might be confounded by other factors like financial need or ability to work while in school. The desired effect can be examined by looking at scholarship.

8. **JUDICIAL**: This variable has not been included because the sample for this variable is very limited and probably biased; a university might not be open to admitting students with a Judicial record.

II. Estimation:

MODEL 1:

COLLGPAi = B0 + B1SATi + B2HSGPAi + B3SAT*HSGPAi + B4HRSTDi + B5HRSTD^2i + B6HREXTRAi +B7 DOUBLEi +B8 FEMALEi + B9SLEEPi + B10INTERNATIONALi +B11INTERNATIONAL*DOUBLEi + B12GREEKi + B13DRINKi + B14GREEK*DRINKi +B15 ATHLETEi + B16FRIENDi + B18DECISIONi + ei

STATA OUTPUT:
Model 1; Regression Equation 1:

```
. reg COLLGPA SAT HSGPA SAT_HSGPA HRSTD HRSTD_SQ HREXTRA DOUBLE FEMALE SLEEP INTER
> NATIONAL INTERNATIONAL_DOUBLE SCHOLAR GREEK DRINK GREEK_DRINK ATHLETE FRIEND DEC
> ISION DECISION_SCHOLAR HELP
```

Source	SS	df	MS		
Model	20.683997	20	1.03419985	Number of obs	= 392
Residual	37.1067194	371	.100018111	F(20, 371)	= 10.34
				Prob > F	= 0.0000
Total	57.7907164	391	.147802344	R-squared	= 0.3579
				Adj R-squared	= 0.3233
				Root MSE	= .31626

COLLGPA	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
SAT	.0025322	.0006365	3.98	0.000	.0012807	.0037838
HSGPA	.8350137	.2513188	3.32	0.001	.3408257	1.329202
SAT_HSGPA	-.0004968	.0001796	-2.77	0.006	-.0008499	-.0001437
HRSTD	.0046991	.0027189	1.73	0.085	-.0006473	.0100454
HRSTD_SQ	-.0000465	.0000333	-1.40	0.164	-.000112	.000019
HREXTRA	-.0003422	.0020546	-0.17	0.868	-.0043823	.003698
DOUBLE	-.0095829	.0455201	-0.21	0.833	-.0990927	.0799268
FEMALE	.1021383	.0352661	2.90	0.004	.0327917	.1714849
SLEEP	-.0028144	.0015278	-1.84	0.066	-.0058185	.0001898
INTERNATIONAL	-.049735	.0490885	-1.01	0.312	-.1462616	.0467916
INTERNATIONAL~E	.194713	.0756949	2.57	0.010	.0458682	.3435578
SCHOLAR	.1363107	.0779102	1.75	0.081	-.0168903	.2895117
GREEK	.0624378	.052739	1.18	0.237	-.0412671	.1661427
DRINK	.0048101	.0030821	1.56	0.119	-.0012504	.0108706
GREEK_DRINK	-.0069606	.0041764	-1.67	0.096	-.0151731	.0012518
ATHLETE	-.0300607	.0397108	-0.76	0.450	-.1081471	.0480257
FRIEND	.0013746	.0005587	2.46	0.014	.000276	.0024733
DECISION	-.0294928	.1036012	-0.28	0.776	-.233212	.1742264
DECISION_SCHO~R	-.0374292	.109322	-0.34	0.732	-.2523976	.1775392
HELP	.0027131	.0056024	0.48	0.628	-.0083033	.0137295
_cons	-.715881	.8762236	-0.82	0.414	-2.438869	1.007107

Table: Model 1

```
. estout Model_1, cells(b(star fmt(3)) se(par fmt(3))) stats(r2 ar2 N, labels("R-squared" "Adjusted R-squared" "N") fmt(3))

                    Model_1
                       b/se
SAT                  0.003***
                    (0.001)
HSGPA                0.835***
                    (0.251)
SAT_HSGPA           -0.000**
                    (0.000)
HRSTD                0.005
                    (0.003)
HRSTD_SQ            -0.000
                    (0.000)
HREXTRA             -0.000
                    (0.002)
DOUBLE              -0.010
                    (0.046)
FEMALE               0.102**
                    (0.035)
SLEEP               -0.003
                    (0.002)
INTERNATIO~L        -0.050
                    (0.049)
INTERNATIO~E         0.195*
                    (0.076)
SCHOLAR              0.136
                    (0.078)
GREEK                0.062
                    (0.053)
DRINK                0.005
                    (0.004)
GREEK_DRINK         -0.007
                    (0.004)
ATHLETE             -0.030
                    (0.040)
FRIEND               0.001*
                    (0.001)
DECISION            -0.029
                    (0.104)
DECISION_S~R        -0.037
                    (0.109)
HELP                 0.003
                    (0.006)
_cons               -0.716
                    (0.876)
R-squared            0.358
Adjusted R~d         0.323
N                  392.000

.
```

III. Discussion of Results:

A. To determine whether a model is “good”, we can use several criteria of measurement. For this project, we look at the goodness of fit measures, the significance of individual coefficients, multicollinearity, heteroskedasticity, and omitted variable bias.

1. Goodness of fit

As can be seen in the above table, our regression model has an adjusted R-squared value of 0.323, which means that our regression model, adjusted for degrees of freedom, explains 32.3% of the variation in college GPA. While this is a decent first attempt, this model can be improved to better explain the variation in college GPA by exploring better relationships between the independent variables and College GPA. So, we can state that according to these criteria, the model is a somewhat acceptable fit but there is definitely room for improvement.

2. Significance of individual coefficients.

In the above model, we looked at the p-values and confidence intervals of our individual coefficients: SAT, GPA, HS_GPA, FEMALE, FRIEND, and INTERNATIONAL_DOUBLE, were statistically significant with a p-value < 0.05, and Confidence Intervals that did not include 0, whereas the remaining majority of the variables were not individually statistically significant. So, we can state that according to these criteria, there is much room for improvement.

3. Multicollinearity

We test for multicollinearity as the potential existence of multicollinearity between our model's independent variables can make it difficult to interpret the effects of our individual coefficients and may lead to unstable estimates.

vif

Variable	VIF	1/VIF
SAT_HSGPA	147.48	0.006780
HSGPA	83.34	0.012000
SAT	37.08	0.026972
ECISION_S~R	11.04	0.090606
DECISION	10.33	0.096791
HRSTD	7.88	0.126979
HRSTD_SQ	7.02	0.142366
GREEK_DRINK	4.01	0.249092
DRINK	2.68	0.373648
NTERNATIO~E	2.58	0.387034
GREEK	2.44	0.409213
SCHOLAR	2.18	0.458747
NTERNATIO~L	1.94	0.516086
DOUBLE	1.76	0.569380
ATHLETE	1.47	0.681543
HREXTRA	1.36	0.735983
HELP	1.27	0.787140
FEMALE	1.14	0.880205
FRIEND	1.09	0.914951
SLEEP	1.09	0.919875
Mean VIF	16.46	

As we can observe from the above output, if we take a VIF>10 as our benchmark, we can observe multicollinearity between SAT, GPA, and their interacted term SAT_HSGPA as well as DECISION and its interaction term DECISION_SCHOLAR. It is not uncommon to encounter high multicollinearity when interaction terms are involved, and it can be attributed to the relationship between the main variables and the interaction term rather than a model specification or a sample-related problem.

Since SAT, GPA, and SAT_HSGPA are statistically significant and are strongly founded on our theoretical model, we believe that keeping them is important. However, the p-values of DECISION and DECISION_SCHOLAR tell us that they are not statistically significant, so we will modify the equation by dropping the interaction term as a remedy.

4. Heteroskedasticity

We test our model for heteroskedasticity as it could potentially result in biased standard errors and incorrect interpretations. We use the Breusch-Pagan test and visually examine the residual plots to check for heteroskedasticity.

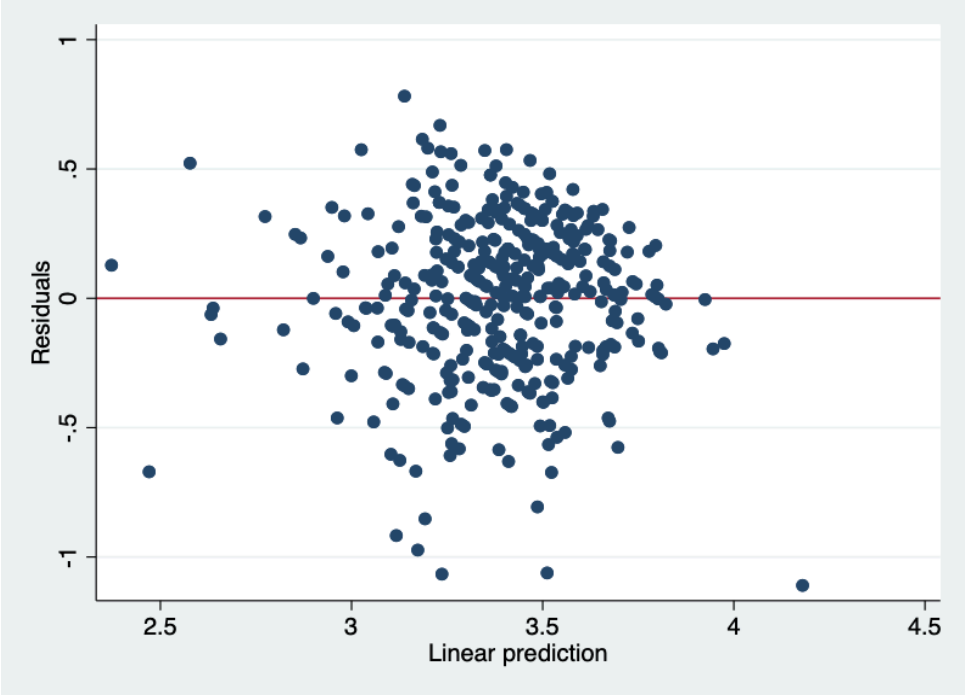
```
. hettest

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variable: Fitted values of COLLGPA

H0: Constant variance

    chi2(1) =   2.11
Prob > chi2 = 0.1462
```

Upon conducting the Breusch-Pagan test, we have
HO = Error terms have constant variance
HA = HO is not true.
P value = 0.1462
As $P > 0.05$, we fail to reject the null hypothesis as the P value is > 0.05, and heteroskedasticity does not exist. We also perform a visual inspection by plotting the residuals against the fitted values to confirm our results and present them as supporting evidence.



5. Omitted Variable Bias

We check if there are important variables that might have been omitted from the model among the remaining unused variables in the dataset. We perform the omitted variable test on our model.

HO: Model has no omitted variable
HA: HO is not true

```
. ovtest

Ramsey RESET test for omitted variables
Omitted: Powers of fitted values of COLLGPA

H0: Model has no omitted variables

F(3, 368) =    7.59
 Prob > F = 0.0001
```

As the p-value is less than the common significance level of 0.05, we reject the null hypothesis. This suggests that there might be omitted variables from our original model. To explore possible remedies, we look at potentially relevant variables that might have been excluded or alternative forms of existing variables.

6. Model Specification

Our model has interaction terms and non-linear relationships that we believe better capture the relationship between some variables like HRSTUDY and college GPA. Based on the individual significance of the

coefficients and our tests for heteroskedasticity, multicollinearity, and omitted variable bias, we observe that some variables like SAT_HSGPA have a strong significance whereas others like DECISION_SCHOLAR do not. In addition, we observe that HR_STD and HR_STDSQ are both not significant in our model. However, our theory is that the number of hours studied per week is an important predictor of college GPA. Thus, in our model modification, we will consider alternative functional forms of HR_STD and reassess the non-linear relationship between HR_STD and HR_STDSQ we specified in our theoretical model.
To examine if the omitted variable bias we observed through the Ramsey test is due to a variable that was omitted from our model, we also re-evaluate the exclusion of variables in our model that could be incorporated into our model modification. In addition, we understand that the relationship between many of the independent variables is complex and might not be captured by a linear relationship. Thus, we will reassess if there are variables that could improve the overall fit of our model if interacted.

To conclude, we think the following modifications on MODEL_1 will make the model better (with theoretical justification):

1. Dropping DECISION AND DECISION_SCHOLAR: as the VIF>10 and their p-value indicates that they are not statistically significant, we drop the variables and remedy the multicollinearity between the interaction term and DECISION.

Based on our initial theoretical justification, we hypothesized that students who applied to college through the early decision and received a scholarship would have a net positive impact on their COLLGPA due to their high commitment and motivation to perform well academically. However, after running the regression analysis, we found that the interaction term between DECISION and SCHOLAR has a high multicollinearity with the main effect variables, which can make it challenging to interpret the independent effects of each variable on COLLGPA accurately.

Furthermore, the p-value of the interaction term is above the 0.05 significance level, indicating that the interaction term is not statistically significant in explaining the variation in COLLGPA. This suggests that the interaction effect between early decision application and scholarship on COLLGPA may not be as strong as initially hypothesized.

Based on these findings, we decided to modify the model by dropping the interaction term between DECISION and SCHOLAR. This will help to reduce multicollinearity and provide a more accurate estimation of the independent effects of the main variables on COLLGPA. We will continue to explore other factors that could potentially influence COLLGPA and refine our model accordingly.

2. Logging the HRSTD variable: we believe that the relationship between a number of hours studied per week and college GPA can be captured better when the HRSTD variable is logged, which might improve the significance of the variable as well as the overall fit of the model.

In our original model, we included the HRSTD variable, which represents the average number of hours spent studying per week, as a predictor of COLLGPA. Theoretically, it is reasonable to expect that the more time a student spends studying, the better their academic performance would be, as reflected in their COLLGPA. However, the relationship between study hours and COLLGPA may not necessarily be linear, as diminishing returns could set in after a certain point.

Taking the natural logarithm of the HRSTD variable allows us to explore this potential non-linear relationship between study hours and COLLGPA. The log transformation can help capture the diminishing returns effect, where the initial increase in study hours has a larger impact on COLLGPA than subsequent increases. In other words, the log transformation allows us to model the idea that an additional hour of study has a greater impact on a student's performance when they are studying fewer hours per week than when they are already studying many hours per week.

By logging the HRSTD variable, we can better understand the relationship between study hours and COLLGPA, accounting for the possibility of diminishing returns. This transformation may provide a more accurate representation of the underlying relationship and lead to a better-fitting and more interpretable model.

3. Dropping the squared term on HRSTD
The initial model includes a squared term, HRSTD_SQ, which represents the square of the number of hours spent studying (HRSTD). The inclusion of a squared term is often used to account for potential non-linear relationships between the independent and dependent variables. In this case, the squared term is meant to capture the possibility that the effect of studying on college GPA (COLLGPA) is not constant but rather changes with the number of hours spent studying.

However, upon examining the statistical results of the initial model, we find that the HRSTD_SQ term is not statistically significant ($P>|t| = 0.164$). This indicates that the additional complexity introduced by including the squared term does not significantly improve the model's explanatory power. As a result, it is theoretically justified to drop the HRSTD_SQ term from the model in order to simplify the model and reduce the risk of overfitting.

By dropping the HRSTD_SQ term, we retain only the linear relationship between HRSTD and COLLGPA in the model. This implies that while the relationship between study hours and COLLGPA may not be perfectly linear, the linear term (HRSTD) is sufficient to capture the majority of the relationship, and the added complexity of the squared term is not warranted.
Thus, the theoretical justification for dropping the squared term (HRSTD_SQ) from the model is based on the lack of statistical significance of the term, which suggests that the added complexity does not contribute substantially to the model's explanatory power.

Therefore, theoretically, the following regression model would be a better fit:

COLLGPAi= B0 + B1SATi + B2HSGPAi + B3SAT_HSGPAi + B$LN_HRSTDi + B5HREXTRAi + B6DOUBLEi + B7FEMALEi + B8SLEEPi + B9INTERNATIONALi + B10INTERNATIONAL*DOUBLEi + B11SCHOLARi + B12GREEKi + B12DRINKi + B14GREEK*DRINKi + B15ATHLETEi + B16FRIENDi + B17DECISIONi + B18HELPi + ei

IV. Model Modifications and Discussions:

A. Here are the modifications that were suggested above:

STATA OUTPUT:

Regression MODEL 2:

```
.
.  reg COLLGPA SAT HSGPA SAT_HSGPA LN_HRSTD HREXTRA DOUBLE FEMALE SLEEP INTERNATIONAL INTERNATIONAL_DOUBLE SCHOLAR GREEK DRINK GREEK_DRINK ATHLETE FRIE
> ND DECISION HELP
```

Source	SS	df	MS		
Model	20.7139663	18	1.15077591	Number of obs	= 390
				F(18, 371)	= 11.62
Residual	36.7561982	371	.09907331	Prob > F	= 0.0000
				R-squared	= 0.3604
				Adj R-squared	= 0.3294
Total	57.4701645	389	.147738212	Root MSE	= .31476

COLLGPA	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
SAT	.0024613	.0006281	3.92	0.000	.0012262	.0036963
HSGPA	.8105809	.2482787	3.26	0.001	.322371	1.298791
SAT_HSGPA	-.0004804	.0001775	-2.71	0.007	-.0008294	-.0001314
LN_HRSTD	.0569015	.0235553	2.42	0.016	.0105828	.1032201
HREXTRA	-.0007619	.0020439	-0.37	0.710	-.004781	.0032572
DOUBLE	-.0077946	.0453195	-0.17	0.864	-.0969098	.0813206
FEMALE	.0985346	.0349658	2.82	0.005	.0297787	.1672905
SLEEP	-.0027117	.0015222	-1.78	0.076	-.0057049	.0002814
INTERNATIONAL	-.0540086	.0489261	-1.10	0.270	-.150216	.0421987
INTERNATIONAL_DOUBLE	.1979442	.0752072	2.63	0.009	.0500583	.3458301
SCHOLAR	.1165172	.0555896	2.10	0.037	.0072071	.2258274
GREEK	.0651148	.052361	1.24	0.214	-.0378468	.1680764
DRINK	.0055687	.0030368	1.83	0.067	-.0004028	.0115402
GREEK_DRINK	-.007325	.0041325	-1.77	0.077	-.0154511	.0008011
ATHLETE	-.0278672	.0393816	-0.71	0.480	-.1053064	.049572
FRIEND	.0012707	.0005552	2.29	0.023	.000179	.0023624
DECISION	-.0606421	.0339507	-1.79	0.075	-.1274021	.006118
HELP	.0018107	.0055569	0.33	0.745	-.0091163	.0127378
_cons	-.6855695	.8555096	-0.80	0.423	-2.367825	.9966864

1. Goodness of fit

As can be seen in the above table, our regression model has an adjusted R-squared value of 0.3294, which means that our regression model, adjusted for degrees of freedom, explains 32.94% of the variation in college GPA. This shows an improvement from our Model 1 in this aspect.

2. Significance of individual coefficients.

In Model 1, we looked at the p-values and confidence intervals of our individual coefficients: SAT, GPA, HS_GPA, FEMALE, FRIEND, and INTERNATIONAL*DOUBLE, were statistically significant with a p-value < 0.05, and Confidence Intervals that did not include 0, whereas the remaining majority of the variables were not individually statistically significant. In our Model 2, LN_HRSTD is statistically significant in addition to the aforementioned independent variables. This is due to the fact that a natural log relationship better captures the relationship between the number of hours studied and college GPA and the remedied multicollinearity as a result of dropping DECISION_SCHOLAR.

3. Multicollinearity

Variable	VIF	1/VIF
SAT_HSGPA	144.93	0.006900
HSGPA	81.87	0.012215
SAT	36.34	0.027518
GREEK_DRINK	3.94	0.253552
DRINK	2.61	0.382765
INTERNATIO~E	2.57	0.388671
GREEK	2.42	0.413820
INTERNATIO~L	1.94	0.515681
DOUBLE	1.75	0.570359
ATHLETE	1.45	0.690334
LN_HRSTD	1.38	0.723109
HREXTRA	1.35	0.740980
HELP	1.26	0.794943
FEMALE	1.12	0.889614
SCHOLAR	1.12	0.893114
DECISION	1.11	0.898795
SLEEP	1.08	0.925823
FRIEND	1.07	0.930319
Mean VIF	16.07	

As can be seen from the above output, dropping DECISION_SCHOLAR resolves the multicollinearity that we observed on DECISION in model 1.

4. Heteroskedasticity

As we did not include any additional variables from our dataset that were not in Model 1, we did not attempt to resolve the omitted variable bias at this stage of the model modification.

5. Model Specification

The modifications we made to our model have improved the overall fit and resolved the multicollinearity we observed as a result of the inclusion of an interaction term. By making these changes, we have also observed that more of our independent variables are statistically significant and our changes have not introduced heteroskedasticity. However, we believe that our model can still be improved in terms of overall fit.

```
. hettest

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variable: Fitted values of COLLGPA

H0: Constant variance

    chi2(1) =   1.82
Prob > chi2 = 0.1779
```

As can be seen from the above output, the p-value is greater than the common significance level of 0.05,so we fail to reject the null hypothesis. This means that we do not have enough evidence to conclude that heteroskedasticity is present in the model. This verifies that the changes we made to model 1 did not violate the Classical Assumptions of Homoskedastic error terms in our model.

To conclude, we think the following modifications on MODEL_2 will make the model better:

Logging HREXTRA:

The original theoretical justification for including HREXTRA in the model acknowledged the uncertain direction of its impact on COLLGPA. By logging HREXTRA, we aim to capture the potential diminishing returns associated with participating in extracurricular activities. This transformation implies that initially, an increase in hours spent on extracurricular activities may have a positive impact on a student's GPA, but at a certain point, the benefits of additional hours may decline, eventually leading to a potential negative impact on GPA as the student becomes more occupied with extracurricular activities at the expense of their academic performance. The logged transformation of HREXTRA allows for a more flexible and potentially non-linear relationship with COLLGPA, helping to better understand the association between the two variables.

Dropping HELP:

The initial theoretical justification for HELP suggested that seeking additional help should improve a student's understanding of the material, leading to a higher GPA. However, the results from the regression show that the coefficient for HELP is statistically insignificant (p-value > 0.05), indicating that the variable might not be contributing significantly to the explanatory power of the model. This could be due to various reasons, such as a confounding factor not considered in the model, or that the relationship between HELP and COLLGPA is not linear. Given the statistically insignificant relationship between HELP and COLLGPA, it is reasonable to consider dropping HELP from the model.

To conclude, we think the mentioned justifications for the modifications on MODEL_3 will make it our “best” model.

Theoretical Equation:

MODEL 3:
COLLGPAi = B0 + B1SATi + B3HSGPAi + B3SAT*HSGPAi + B4LN_HRSTDi + B5LN_HREXRTAi + B6DOUBLEi + B7FEMALEi + B8SLEEPi + B9INTERNATIONALi + B10INTERNATIONAL*DOUBLEi + B11SCHOLARi + B12GREEKi + B13DRINKi + B14GREEK*DRINKi + B15 ATHLETEi + B16FRIENDi + B17DECISIONi+ei

```
. reg COLLGPA SAT HSGPA SAT_HSGPA LN_HRSTD LN_HREXTRA DOUBLE FEMALE SLEEP INTERNATIONAL INTERNATIONAL_DOUBLE SCHOLAR GREEK DRINK GREEK_DRINK ATHLETE FRIEND DECISION
```

Source	SS	df	MS		
Model	20.7754855	17	1.22208738	Number of obs	= 390
Residual	36.694679	372	.09864161	F(17, 372)	= 12.39
				Prob > F	= 0.0000
Total	57.4701645	389	.147738212	R-squared	= 0.3615
				Adj R-squared	= 0.3323
				Root MSE	= .31407

COLLGPA	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
SAT	.0025018	.0006242	4.01	0.000	.0012744	.0037292
HSGPA	.8198369	.2472708	3.32	0.001	.333613	1.306061
SAT_HSGPA	-.000491	.0001765	-2.78	0.006	-.0008381	-.000144
LN_HRSTD	.0548129	.0224808	2.44	0.015	.0106075	.0990184
LN_HREXTRA	.0198842	.0216579	0.92	0.359	-.0227031	.0624714
DOUBLE	-.0121833	.0452548	-0.27	0.788	-.1011706	.076804
FEMALE	.1019097	.0345702	2.95	0.003	.0339321	.1698873
SLEEP	-.0026601	.0015133	-1.76	0.080	-.0056359	.0003156
INTERNATIONAL	-.0508645	.0487862	-1.04	0.298	-.1467959	.0450669
INTERNATIONAL_DOUBLE	.1987325	.0748729	2.65	0.008	.0515052	.3459598
SCHOLAR	.120359	.0552383	2.18	0.030	.0117405	.2289775
GREEK	.0698237	.0516228	1.35	0.177	-.0316854	.1713327
DRINK	.0057247	.0029986	1.91	0.057	-.0001716	.011621
GREEK_DRINK	-.0076471	.0040963	-1.87	0.063	-.0157019	.0004076
ATHLETE	-.0456188	.037846	-1.21	0.229	-.1200377	.0288002
FRIEND	.0012602	.0005541	2.27	0.024	.0001706	.0023497
DECISION	-.065547	.0337348	-1.94	0.053	-.1318819	.0007879
_cons	-.7710578	.8522503	-0.90	0.366	-2.44689	.9047745

1. The goodness of fit

As can be seen in the above table, our regression model has an adjusted R-squared value of 0.33223, which means that our regression model, adjusted for degrees of freedom, explains 33.23% of the variation in college GPA. This shows an improvement from our Model 1 in this aspect.

2. Significance of individual coefficients.

In Model 2, we looked at the p-values and confidence intervals of our individual coefficients: SAT, GPA, HS_GPA, FEMALE, FRIEND, INTERNATIONAL_DOUBLE, LN_HRSTD and DECISION were statistically significant with a p-value < 0.05 and Confidence Intervals that did not include 0, whereas the remaining majority of the variables were not individually statistically significant. In our Model 3, SLEEP was statistically significant in addition to the aforementioned independent variables.

3. Multicollinearity

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
   SAT_HSGPA |    143.95    0.006947
       HSGPA |     81.56    0.012261
         SAT |     36.05    0.027740
 GREEK_DRINK |      3.89    0.256938
INTERNATIO~E |      2.56    0.390441
       DRINK |      2.56    0.390878
       GREEK |      2.36    0.423885
INTERNATIO~L |      1.94    0.516383
      DOUBLE |      1.76    0.569498
     ATHLETE |      1.34    0.744234
    LN_HRSTD |      1.27    0.790424
  LN_HREXTRA |      1.21    0.828426
     SCHOLAR |      1.11    0.900567
      FEMALE |      1.10    0.906120
    DECISION |      1.10    0.906369
      FRIEND |      1.08    0.929857
       SLEEP |      1.07    0.932576
-------------+----------------------
    Mean VIF |     16.82
```

As can be seen from the above output, our model modification did not introduce any additional multicollinearity.

4. Heteroskedasticity:

```
. hettest

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variable: Fitted values of COLLGPA

H0: Constant variance

    chi2(1) =   1.55
Prob > chi2 = 0.2130
```

As can be seen from the above output, there is no significant evidence that indicates heteroskedasticity, so our model modifications do not violate the classical assumption of homoskedasticity.

5. Model Specification

The modifications we made to our model have improved the overall fit and resolved the multicollinearity we observed as a result of the inclusion of an interaction term. By making these changes, we have also observed that more of our independent variables are statistically significant than MODEL 1 and MODEL 2 and our changes have not introduced heteroskedasticity or additional multicollinearity.

To conclude, we decided that Model_3 is our "best" model as it has the highest overall fit among our three models and our independent variables are mostly statistically significant. In addition, apart from the possible omitted variable bias indicated by the Ramsey test, our model shows no evidence of heteroskedasticity. Through our model modifications, we have reassessed and revised some of our initial theoretical assumptions.

V. Post Estimation:

A. MODEL_3 predicts one's college GPA the best. Based on the adjusted R^2 value of MODEL 3, 33.23% of the variation in college GPA is explained by the independent variables. In addition, most of the independent variables are statistically significant. For the independent variables that are not, we have reassessed our initial theoretical model and used reasoning to decide to keep them. This is because the significance of our independent variables might be influenced by the sample and theory supports the inclusion of the variables. We have tested our model for potential violation of classical assumptions and have found no significant evidence of homoskedasticity or multicollinearity(apart from the interaction term between SAT and HSGPA.

B.

To test the hypothesis "An individual's combined math and critical reading SAT score significantly increase his/her college GPA," we will perform a one-sided t-test.

Null hypothesis (H0): The coefficient of the SAT variable is equal to or less than 0, meaning that the combined SAT score does not significantly increase college GPA.
Alternative hypothesis (H1): The coefficient of the SAT variable is greater than 0, meaning that the combined SAT score significantly increases college GPA.

From the regression output, we have:

Coefficient (SAT) = 0.0025018
Standard error (SAT) = 0.0006242
t-statistic (SAT) = 4.01
P>|t| (SAT) = 0.000

For a one-sided t-test at the 5% level, the critical value is 1.65. For a one-sided t-test at the 1% level, the critical value is 2.33.

Since the t-statistic (4.01) is greater than both critical values, we reject the null hypothesis at both the 1% and 5% levels of significance.

Hence, an individual's combined math and critical reading SAT score significantly increases his/her college GPA at both the 1% and 5% levels of significance.

C. Testing if the combined SAT score has the greatest impact on college GPAs or if another variable does:

In this best model, does the combined SAT score have the greatest impact on college GPAs? If so, why…. If not, what variable does the greatest impact? Please explain how you arrived at your answer.

```
. reg COLLGPA SAT HSGPA SAT_HSGPA LN_HRSTD LN_HREXTRA DOUBLE FEMALE SLEEP INTERNATIONAL INTERNATIONAL_DOUBLE SCHOLAR GREEK DRINK GREEK_DRINK ATHLETE FRIEND DECISION

      Source |       SS           df       MS      Number of obs   =       390
-------------+----------------------------------   F(17, 372)      =     12.39
       Model |  20.7754855        17  1.22208738   Prob > F        =    0.0000
    Residual |   36.694679       372   .09864161   R-squared       =    0.3615
-------------+----------------------------------   Adj R-squared   =    0.3323
       Total |  57.4701645       389  .147738212   Root MSE        =    .31407

------------------------------------------------------------------------------
     COLLGPA | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         SAT |   .0025018   .0006242     4.01   0.000     .0012744    .0037292
       HSGPA |   .8198369   .2472708     3.32   0.001      .333613    1.306061
    SAT_HSGPA |   -.000491   .0001765    -2.78   0.006    -.0008381    -.000144
    LN_HRSTD |   .0548129   .0224808     2.44   0.015     .0106075    .0990184
   LN_HREXTRA |   .0198842   .0216579     0.92   0.359    -.0227031    .0624714
       DOUBLE |  -.0121833   .0452548    -0.27   0.788    -.1011706     .076804
       FEMALE |   .1019097   .0345702     2.95   0.003     .0339321    .1698873
        SLEEP |  -.0026601   .0015133    -1.76   0.080    -.0056359    .0003156
INTERNATIONAL |  -.0508645   .0487862    -1.04   0.298    -.1467959    .0450669
INTERNATIONAL_DOUBLE |   .1987325   .0748729     2.65   0.008     .0515052    .3459598
      SCHOLAR |    .120359   .0552383     2.18   0.030     .0117405    .2289775
       GREEK |   .0698237   .0516228     1.35   0.177    -.0316854    .1713327
       DRINK |   .0057247   .0029986     1.91   0.057    -.0001716     .011621
GREEK_DRINK |  -.0076471   .0040963    -1.87   0.063    -.0157019    .0004076
     ATHLETE |  -.0456188    .037846    -1.21   0.229    -.1200377    .0288002
      FRIEND |   .0012602   .0005541     2.27   0.024     .0001706    .0023497
    DECISION |   -.065547   .0337348    -1.94   0.053    -.1318819    .0007879
       _cons |  -.7710578   .8522503    -0.90   0.366     -2.44689    .9047745
------------------------------------------------------------------------------
```

To see if the combined SAT score has the greatest impact on college GPA, we can compare the t-statistics for each variable. The higher the t-statistic (in absolute terms), the more significant the relationship between the independent variable and the dependent variable (college GPA).

From the given regression output, we can compare the t-statistics:

SAT: 4.01
HSGPA: 3.32
FEMALE: 2.95
INTERNATIONAL_DOUBLE: 2.65
SCHOLAR: 2.18
FRIEND: 2.27
LN_HRSTD: 2.44
The SAT variable has the highest t-statistic (4.01), which indicates that it has the most significant impact on college GPAs in this model.

D. One-sentence interpretations that each variable has on one's GPA:

1. SAT - On average, for every one-point increase in the aggregate math and critical reading SAT score out of a maximum score of 1600, the overall college grade point average for a student increases by 0.0025 points, holding all other independent variables constant.
2. HSGPA - On average, for every one-point increase in the overall high school grade point average of a student, the student's overall college grade point average increases by 0.8198 points, holding all other independent variables constant.
3. LN_HRSTD- A 1% increase in the average number of hours spent studying per week is associated with an increase of 0.0548 points in one's college GPA, holding all other independent variables constant.
4. LN_HREXTRA - On average, for every 1% increase in the average number of hours spent studying per week increases the overall college grade point average for a student increases by 0.0199 points, holding all other independent variables constant.
5. DOUBLE - On average, if a student is pursuing a double major, the student is expected to have 0.0122 points less in the overall college grade point average compared to a student who is pursuing a single major, holding all other independent variables constant.
6. FEMALE - On average, if a student identifies as female, the overall college grade point average for the student is expected to be 0.102 more than a student who identifies as male, holding all other independent variables constant.
7. SLEEP - On average, for every one-hour increase in the average hours of sleep a student gets per week, the overall grade point average for the student decreases by 0.00266, holding all other independent variables constant.
8. INTERNATIONAL - On average, if a student is an international student, the student is expected to have 0.05086 points less in the overall college grade point average compared to a student who is a domestic student, holding all other independent variables constant.
9. INTERNATIONAL_DOUBLE - On average, an International Student who is a double major is expected to have a 0.1987 points better college GPA compared to a student who is not an International Student and Double Major, holding all other independent variables constant.
10. SCHOLAR - On average, if a student has received a scholarship, the student is expected to have 0.1204 points more in the overall college grade point average compared to a student who did not receive a scholarship, holding all other independent variables constant.
11. GREEK - On average, if a student is affiliated with a Greek organization, the student is expected to have 0.0698 points more in the overall college grade point average compared to a student who is not affiliated with a Greek organization, holding all other independent variables constant.
12. DRINK - An increase of 1 unit in the average number of alcoholic drinks per week is associated with an increase of 0.0057 points in one's college GPA, holding all other independent variables constant.
13. GREEK_DRINK - For students affiliated with a Greek organization, an increase of 1 unit in the average number of alcoholic drinks per week is associated with a decrease of 0.0076 units in one's college GPA compared to those not affiliated with a Greek organization, holding all other independent variables constant.
14. ATHLETE - On average, if a student is an intercollegiate athlete, the student is expected to have .0456 points less in the overall college grade point average compared to a student who is not an intercollegiate athlete, holding all other independent variables constant.
15. FRIEND - On average, for a 1% increase in time spent, while in college, with a significant other, a student is expected to increase their overall college grade point average by 0.0013, holding all other independent variables constant.

16. DECISION - On average, if a student has applied to the college through Early Decision, the student is expected to have .0655 points less in the overall college grade point average compared to a student who did not apply to college with Early Decision, holding all other independent variables constant.

    E. What does your best model predict YOUR GPA to be? Please show me your work. Does it do a good job of predicting your GPA?

COLLGPAi= B0 + B1SATi + B2HSGPAi + B3SAT_HSGPAi + B$LN_HRSTDi + B5HREXTRAi + B6DOUBLEi + B7FEMALEi + B8SLEEPi + B9INTERNATIONALi + B10INTERNATIONAL*DOUBLEi + B11SCHOLARi + B12GREEKi + B12DRINKi + B14GREEK*DRINKi + B15ATHLETEi + B16FRIENDi + B17DECISIONi + B18HELPi + ei

COLLGPA = -0.7710578 + (0.0025018 * 1480) + (0.8198369 * 3.95) + (-0.000491 * 1480 * 3.95) + (0.0548129 * 2.9957) + (0.0198842 * 1.6094) + (-0.0121833 * 1) + (0.1019097 * 0) + (-0.0026601 * 34) + (-0.0508645 * 1) + (0.1987325 * 1) + (0.120359 * 1) + (0.0698237 * 0) + (0.0057247 * 1) + (-0.0076471 * 0) + (-0.0456188 * 0) + (0.0012602 * 1) + (-0.065547 * 0)

COLLGPA ≈ 3.485

The predicted GPA according to the model is approximately 3.485, which is not far off from my actual GPA of 3.64. Given the possibility of the existence of an omitted variable, the model does a good job of predicting my college GPA.

VI. My Model:

A. AZENUI MODEL

```
. reg COLLGPA SAT LN_HSGPA HRSTD HRSTD_SQ DECISION FEMALE SAT_FEMALE ECON

      Source |       SS           df       MS      Number of obs   =       396
-------------+----------------------------------   F(8, 387)       =     23.87
       Model |  19.2540775         8  2.40675969   Prob > F        =    0.0000
    Residual |  39.0268511       387  .100844577   R-squared       =    0.3304
-------------+----------------------------------   Adj R-squared   =    0.3165
       Total |  58.2809286       395  .147546655   Root MSE        =    .31756

------------------------------------------------------------------------------
     COLLGPA | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         SAT |    .000796   .0001455     5.47   0.000     .0005099     .001082
    LN_HSGPA |    .789321   .1270285     6.21   0.000     .5395687    1.039073
       HRSTD |    .005484   .0025784     2.13   0.034     .0004146    .0105534
    HRSTD_SQ |   -.0000472   .0000329    -1.44   0.152    -.0001118    .0000174
    DECISION |   -.0667332   .0329952    -2.02   0.044    -.1316056   -.0018609
      FEMALE |    .1350697   .2857187     0.47   0.637    -.4266855    .6968249
  SAT_FEMALE |   -.0000202   .0002169    -0.09   0.926    -.0004466    .0004062
        ECON |   -.0394685   .0361716    -1.09   0.276    -.1105859     .031649
       _cons |    1.247748    .209842     5.95   0.000     .8351746    1.660321
------------------------------------------------------------------------------
```

B. One-sentence interpretations that each variable has on one's GPA:

a. SAT score - SAT score has a positive and significant effect on college GPA, indicating that a one-point increase in SAT score is associated with a 0.0008 point increase in college GPA, holding all other variables constant.

b. High school GPA has a positive and significant effect on college GPA, indicating that a one-point increase in high school GPA is associated with a 0.79-point increase in college GPA, holding all other variables constant.

c. Hours studied. Note, since it is quadratic, provide the sentence at the mean hours of studied (rounding to the closest half-hour). (This is pretty typical; the means –averages - are often the point at which the slope is determined for quadratics and reciprocals for quadratics and reciprocals.)

At the mean hours studied of 20.74 hours, an increase of one hour in studying is associated with an increase of approximately 0.0048 in one's college GPA, holding all other variables constant; however, the effect decreases as hours studied increases due to the quadratic nature of the relationship.

C. Testing if an individual's combined math and reading SAT score significantly influences his/her college GPA, at the 5% level:

Null hypothesis (H0): There is no significant relationship between the SAT score and college GPA.
H0: β (coefficient of SAT) = 0

Alternative hypothesis (H1): There is a significant relationship between the SAT score and college GPA.
H1: β (coefficient of SAT) $\neq$ 0

To perform the two-sided t-test, we use the t-statistic and degrees of freedom from the provided regression output:

t-statistic: 5.47
Degrees of freedom (df): 387

For a two-tailed test with a 5% significance level and 387 degrees of freedom, the critical value is 1.96.

Since the t-statistic (5.47) is greater than the critical value (1.96), we reject the null hypothesis (H0) in favor of the alternative hypothesis (H1). This means that there is a significant relationship between an individual's combined math and reading SAT score and his/her college GPA at the 5% level.

D. Comparing the AZENUI model with Model_3:

Overall fit:
Based on the overall fit of the model, our best model(Model 3) is slightly better than your model. Our model explains 33.23% of the variations in college GPA, adjusted for degrees of freedom whereas your model explains 31.65%.

Significance of individual coefficients:

In both our best model and your model, most independent variables are individually statistically significant.

Multicollinearity

In your model, SAT and SAT_FEMALE have a VIF >10, which indicates multicollinearity. Even though both independent variables are not statistically significant, it does not necessarily indicate a better or worse model as long as your theory supports the inclusion of these variables. In our best model, SAT and SAT have a high VIF as these variables are strongly correlated. However, our theory suggests that these variables belong in the model and their p-values indicate that they are statistically significant.

Heteroskedasticity

Upon examining your model, we observed that there is evidence of heteroskedasticity.

```
. hettest

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: Normal error terms
Variable: Fitted values of COLLGPA

H0: Constant variance

    chi2(1) =   5.26
Prob > chi2 = 0.0218

. vif
```

As shown in Section 3, our final model shows no evidence of heteroskedasticity.

Omitted Variable Bias

Based on the Ramsey test for omitted variable bias, both our model and your model suggest that there might be evidence of omitted variable bias. In our model modification, we have made changes that we believed would possibly fix the omitted variable bias, but our final model still shows evidence of possible bias. It is difficult to account for omitted variable bias as the variables in the dataset are not the only predictors of college GPA.

VII. Using STATA you can make a nice looking regression table. One of STATA code examples is "**estout**" or "**outreg2".** These codes help you automate the creation of research paper looking tables. For a detailed description and examples with these codes, type "help estout OR help outreg2" in your STATA command window and explore the links provided. Include steps in your printed STATA results and hand them in with written answers. Do not forget to label all your answers based on questions' numbers.

	Model 1 b/se	Model 2 b/se	Model 3 b/se	Azenui b/se
SAT	0.003***	0.002***	0.003***	0.001***
	(0.001)	(0.001)	(0.001)	(0.000)
HSGPA	0.835***	0.811***	0.820***	
	(0.251)	(0.248)	(0.247)	
SAT_HSGPA	-0.000***	-0.000***	-0.000***	
	(0.000)	(0.000)	(0.000)	
HRSTD	0.005*			0.005**
	(0.003)			(0.003)
HRSTD_SQ	-0.000			-0.000
	(0.000)			(0.000)
HREXTRA	-0.000	-0.001		
	(0.002)	(0.002)		
DOUBLE	-0.010	-0.008	-0.012	
	(0.046)	(0.045)	(0.045)	
FEMALE	0.102***	0.099***	0.102***	0.135
	(0.035)	(0.035)	(0.035)	(0.286)
SLEEP	-0.003*	-0.003*	-0.003*	
	(0.002)	(0.002)	(0.002)	
ETHNIC==I	-0.050	-0.054	-0.051	
	(0.049)	(0.049)	(0.049)	
INTERNATIONAL_DOUBLE	0.195**	0.198***	0.199***	
	(0.076)	(0.075)	(0.075)	
SCHOLAR	0.136*	0.117**	0.120**	
	(0.078)	(0.056)	(0.055)	
GREEK	0.062	0.065	0.070	
	(0.053)	(0.052)	(0.052)	
DRINK	0.005	0.006*	0.006*	
	(0.003)	(0.003)	(0.003)	
GREEK_DRINK	-0.007*	-0.007*	-0.008*	
	(0.004)	(0.004)	(0.004)	
ATHLETE	-0.030	-0.028	-0.046	
	(0.040)	(0.039)	(0.038)	

FRIEND	0.001**	0.001**	0.001**	
	(0.001)	(0.001)	(0.001)	
DECISION	-0.029	-0.061*	-0.066*	-0.067**
	(0.104)	(0.034)	(0.034)	(0.033)
DECISION_SCHOLAR	-0.037			
	(0.109)			
HELP	0.003	0.002		
	(0.006)	(0.006)		
LN_HRSTD		0.057**	0.055**	
		(0.024)	(0.022)	
LN_HREXTRA			0.020	
			(0.022)	
LN_HSGPA				0.789***
				(0.127)
SAT_FEMALE				-0.000
				(0.000)
ECON				-0.039
				(0.036)
Observations	392	390	390	396
Adjusted R-squared	0.323	0.329	0.332	0.317