Quantifying the Effects of Demographic Background on Academic Success

Andrew Eeckman, 914834317

Abstract

Colleges are often considered the forefront of progressive thought and a place where equity and equality are considered paramount. However, looks can be deceiving and there may still exist biases in the level of academic success made available to students of different backgrounds. This paper explores the above relationship between demographics and college level GPA and how it may progress throughout an individual's first two years of higher education. Both gender and age may leave certain individuals at a greater advantage than others. It is important to mention, however, that the above demographics can only explain a small portion of the variation found in both years. Unobserved attributes such as IQ, talent, and motivation are left out, more than likely accounting for much of the variation described. The following paper will go more into depth.

Introduction

With college enrollment at an all-time high, it has become imperative to ensure that all students, regardless of their backgrounds, are treated fairly and equally. Because of this, any source of discrimination found to be acting unfairly on certain students must be eliminated. Demographic factors such as gender, age, birthplace, and even native language are four such characteristics that have faced discriminated in the past. Any potential discrimination may result students of different backgrounds performing worse academically, especially if it is the result of biases held by their professors and fellow classmates. That is why the primary goal of this paper is to answer the following two questions:

- How is the demographic background of a student related to their academic success?

- How do these influences change over time?

To help answer these questions, two separate regressions are performed and compared. The first regression regresses first year GPA on both academic and demographic information. The second regression does the same but with second year GPA instead. The results of both regressions are then compared. In both regressions, GPA is found to be largely determined by the number of credits taken that year and any other previous academic performance, however, there is evidence that demographic information does still play a role. In the first regression, gender is found to have a negative effect on first year GPA; students that identify as female upon entering college perform worse than their likewise counterparts. In the second regression, a few things change, female identify students as well as native English speakers appear to perform better than their counterparts, while Asia-borne students appear to perform slightly worse. In the following sections, these differences shall be examined and potential practices and policies shall be proposed to mitigate these effects.

Data

The dataset used in this paper comes from information collected on students at a large Canadian university. This dataset contains cross-sectional information detailing the academic performance and demographic background of the 34,881 students enrolled in the university. The school has three campuses, one with a low acceptance rate of roughly 55% and the other two with higher acceptance rates of around nearly 77%. These campuses will be called campus 1 and campus 2 and 3, respectively. Campus 2 and 3 are significantly smaller than campus 1 and are home to both more commuter and part-time students. The variables available in this dataset are presented in the tables below:

Table 1: Quantitative Variables Present

Variable Name	Description	Mean	Std. Dev.	(Min, Max)
GPA_year1	The GPA the student earns in their first year (4.0 scale).	2.54752	.8199082	(0, 4.3)
GPA_year2	The GPA the student earns in their second year (4.0 scale).	2.590055	.812557	(0, 4.3)
totcredits_year1	The total number of credits the student attempts in their first year.	4.608182	.4926066	(3, 6)
totcredits_year2	The total number of credits the student attempts in their second year.	4.344804	.8943842	(0, 9)
age_at_entry	The age of the student when they first enter the university.	18.63513	.734615	(17, 21)
gpacutoff	The GPA cutoff whereby earning any GPA lower than the number listed may result in probation.	1.521318	.0409561	(1.5, 1.6)
hsgrade_pct	The student's high school grade percentile ranking	52.4151	28.50681	(1, 100)

Table 2: Categorical/Dummy Variables Present

Variable Name / Group	Subsequent Dummies	Description	Mean
sex	female	=1 if the student is female	.6259568
	male	=1 if the student is male	.3740432
mtongue	l_english	=1 if native language is English	.7169233
	l_french	=1 if native language is French	.0050171
	other	=1 if native language is neither English nor French	.2780597
gradin_	gradin4	=1 if the student graduated in four years	.5456634
	gradin5	=1 if the student graduated in five years	.7864834
	gradin6	=1 if the student graduated in six years	.8539311
probation_year_	probation_year1	=1 if the student was placed on probation after the first year	.111694
	probation_year2	=1 if the student was placed on probation after the second year	.0396778
suspended_year_*	suspended_year1*	=1 if the student was suspended after the first year	.0000573
	suspended_year2*	=1 if the student was suspended after the second year	.0405665
	suspended_ever*	=1 if the student was suspended ever	.0584846
campus_	campus1	=1 if the student is attending campus 1	.6087555

	campus2	=1 if the student is attending campus 2	.1780626
	campus3	=1 if the student is attending campus 3	.213182
birthplace	bp_america	=1 if the student was born in America	.0018635
	bp_asia	=1 if the student was born in Asia	.0771193
	bp_canada	=1 if the student was born in Canada	.8784152
	bp_other	=1 if the student was not born in any of the above three areas	.042602

*suspension occurs when a student is on probation and they fail to bring their grade back up past the GPA cut off.

Regression Analysis

This paper will use two primary regressions to help determine the relationship between a student's demographic background and their predicted GPA and how that relationship changes from one year to the next. To do so, we will need variables that are statistically significant and could reasonably affect one's GPA in a given year.

Our first regression is acting on the GPA of students in their first year, GPA_year1. That being said, this means that our above dataset has quite a few explanatory variables that become relevant after a student has already obtained their first year GPA. For this reason, we will not be using the following variables in our first regression: probation_year1, probation_year2, suspension_year1, suspension_year2, suspension_ever, gradin4, gradin5, gradin6, and gpacutoff. Campus1, other, male, and l_other will also be left out as, if included, the subsequent model would fall into a dummy variable trap.

Since the remaining variables not being excluded could have an impact on the student's first year GPA, our initial model will include all of them and we will begin by weeding out those that are statistically insignificant. The cut off for any variable will be at the 90% confidence level or those with a p-value > 0.10. If several variables are found to be individually statistically insignificant, we will also conduct a joint F test. Variables that fail both tests will be dropped from our model. So, to begin, the population model for the first regression is shown below.

$$GPA\_year1 = \beta_0 + \beta_1 totcredits\_year1 + \beta_2 hsgrade\_pct + \beta_3 campus1 + \beta_4 campus3 + \beta_5 female + \beta_6 l\_english + \beta_7 l\_french + \beta_8 bp\_america + \beta_9 bp\_asia + \beta_{10} bp\_canada + \beta_{11} age\_at\_entry + u$$

Upon carrying out this regression, there are a few explanatory variables in the above model which appear to be statistically insignificant. These variables include l_english, l_french, bp_america, bp_asia, and bp_canada, all of which have p-values greater than 0.10. Further analysis shows that the aforementioned variables are also jointly statistically insignificant at the 90% confidence interval. Thus, these variables can be dropped from the above regression without negatively affecting the model's overall fit. Both interaction and quadratic terms should now be considered as the true model may express a great deal of non-linearity.

When deciding on which interaction terms to use for the above model, we should consider creating interaction terms between variables that are highly correlated or that we suspect may be influencing one another. The focus of this paper is to explain how a student's demographic background could have an effect on their GPA so special attention must be paid in regards to both female and age_at_entry. The goal here is to see if there are biases in the grading of likewise similar students, so we will need to primarily compare students that possess the same level of intellect that only vary by either their age or their gender. Unfortunately, the dataset does not include the IQ scores of the presented students nor any other explicit measure of intelligence. With this in mind, for first year students, we will use hsgrade_pct as a proxy measure to account for this. However, there are a few problems with doing so as not everyone has a similar high school experience and high school can vary greatly in the quality of education they provide. Regardless, hsgrade_pct has the advantage that it pairs students up against others who received the same training and education. With this in mind, we will create two interaction terms, one

with age_at_entry and the other with female, both of them interacting with hsgrade_pct. When added to the model, there are both statistically significant and so they are kept. Despite however the focus being on demographic variables, it is important to consider other terms interacting as well. Both hsgrade_pct and totcredits_year1 are highly correlated. An interaction term between them is added and passes significance testing; it will be included in the final model. Campus choice also seems to have a strong correlation with totcredits_year1 as campus 3 and campus 1 differ in the student type, this correlation makes logical sense. Two interaction terms are subsequently added and tested for statistical significance before being added to the final model, both interact either campus1 or campus3 with totcredits_year1. No other terms of statistical significance appear to have strong correlations with any other and so no more interaction terms are considered. However, quadratic terms still need to be investigated.

When considering quadratic terms, it is important consider terms that may either have a depreciating effect on GPA_year1 or an exponential effect. For this model, it is suspected that there are two main variables that potentially exhibit these individual behaviors. As a student attempts a greater number of credits in their first year, they may become overwhelmed and their GPA may start to suffer. Thus, the variable sq_totalcreds is created as a quadratic term for totcredits_year1 to account for this diminishing effect. In contrast, hsgrade_pct may exhibit the opposite behavior. Seeing as the average for high school percentile is 52.4151 and its standard deviation is 28.50681, it makes sense that, in an ideal world, universities would make courses challenging but still accommodating to this mean. That being said, student who are far above this level and have done remarkably better then their peers thus far may experience increasing returns to their GPA in college. The term sq_hsgrade_pct is added as a quadratic transformation of

hsgrade_pct to this model and is found to be statistically significant. With both quadratic and interaction terms added, our final OLS model for GPA_year1 becomes:

$$\begin{aligned} GPA\_year1 = {} & \beta_0 + \beta_1 totcredits\_year1 + \beta_2 hsgrade\_pct + \beta_3 campus1 + \beta_4 campus3 \\ & + \beta_5 female + \beta_6 age\_at\_entry + \beta_7 (female * hsgrade\_pct) \\ & + \beta_8 (age\_at\_entry * hsgrade\_pct) + \beta_9 (hsgrade\_pct * totcredits\_year1) \\ & + \beta_{10} (campus1 * totcredits\_year1) + \beta_{11} (campus3 * totcredits\_year1) \\ & + \beta_{12} sq\_totalcreds + \beta_{13} sq\_hsgrade\_pct + u \end{aligned}$$

The above regression has an R^2 of 0.3643, meaning that 36.43% of the variation in GPA_year1 can be explained by this model's regressors. Heteroskedastic robust standard errors were also used as heteroskedasticity was detected in the residual chart of hsgrade_pct. This chart can be seen in **Figure 1** in the appendix. Partial effects are also presented in **Table 3** in the appendix and are listed side-by-side the partial effects from the previously unaltered model. The variable with the largest partial effects are totcredits_year1, campus1, and campus3. For every additional credit a student attempts, their estimated GPA increases by .3105085 linearly. However, this variable does in fact experience significant diminishing returns as well as several interactions. For every additional credit, this variable's linear partial effect declines by -2(.0292152 * totcredits_year1). An increase in hsgrade_pct also results in an increase of .0011612, enrollment on campus 1 results in an increase of .0586667 and campus 3 results in an increase of .1217602 in first year GPA. Choice of campus also matters outright, students enrolled in campus 1 have an expected -.3555181 decline in their first year GPAs. Students enrolled in campus 3 experience a worse decline of -.495724 on their first year GPAs. These results worsen slightly based on the number of credits a student is taking, less is more in this case. Demographic characteristics also had statistically significant effects on first year GPA, albeit at a lesser degree

than the aforementioned variables. Identifying as female caused the student to suffer a decline of -.047622 to their first year GPAs. Female students also experienced slightly lower returns on their high school percentile than did males with a decline of -.0005296 for each 1 percentile increase. Older students actually performed better than their counterparts with a partial effect of .007343 as long as their high school percentile was roughly below 15.72. At numbers higher than this, older students performed worse than their younger counterparts as GPA declined by -.000467 for every high school percentile increase. The marginal effects of age and gender are summarized in **Figure 2** and **Figure 3** in the appendix. Now, it is important to note that these partial effects may also be subject to flaws within the OLS model used, so to account for this, probit and logit models have also been constructed, along with a new variable. Both probit and logit model utilize the same explanatory terms but use a new term, above_avg, as their dependent variable. The term above_avg is simply a dummy variable that equals 1 if students have an above average first year GPA and equals 0 if they do not. The average marginal effects of regressors used in the probit and logit regressions can be found on **Table 3** in the appendix. For both regressions, age_at_entry is no longer considered statistically significant at the 90% confidence level with a roughly a -0.3% decline in the probability that someone gets an above average GPA in their first year. What is statistically significant is that females are -5.29888% (on probit) and -5.51369% (on logit) less likely to gain an above average first year GPA. Now we will see how these influences may change in the student's second year at college.

In our second regression, we will be using all variables initially considered for our first-year model as well as ones that now apply to a student's second year. So, in addition, this model will include GPA_year1 and totcredits_year2. The initial model is shown below:

$$GPA\_year2 = \beta_0 + \beta_1 totcredits\_year1 + \beta_2 hsgrade\_pct + \beta_3 campus1 + \beta_4 campus3 + \beta_5 female + \beta_6 l\_english + \beta_7 l\_french + \beta_8 bp\_america + \beta_9 bp\_asia + \beta_{10} bp\_canada + \beta_{11} age\_at\_entry + \beta_{12} GPA\_year1 + \beta_{13} totcredits\_year2 + u$$

Again, statistical significance is tested with 90% being the cutoff for exclusion. The terms l_french, bp_america, and bp_canada fail both individual t tests and joint F tests and so they are dropped from our model. It is now important again to consider which interaction terms to add and which quadratic terms may be necessary

In order to keep this model comparable to the final model used in our first regression, we should use likewise interaction and quadratic terms, eliminating only those that are statistically insignificant. With that in mind, we will add interactions between GPA_year1 and both female and age_at_entry. We will also add interactions between both campuses and totcredits_year2. For quadratics, we will generate two new variables, the square of GPA_year1, sq_gpa, and the square of totcredits_year2, sq_totcreds_2. Any statistically insignificant terms will be removed. Our final model for year 2 is as follows:

$$GPA\_year2 = \beta_0 + \beta_1 totcredits\_year1 + \beta_2 hsgrade\_pct + \beta_3 campus1 + \beta_4 campus3 + \beta_5 female + \beta_6 l\_english + \beta_7 l\_french + \beta_8 bp\_america + \beta_9 bp\_asia + \beta_{10} bp\_canada + \beta_{11} age\_at\_entry + \beta_{12} GPA\_year1 + \beta_{13} totcredits\_year2 + \beta_{14}(hsgrade\_pct * age\_of\_entry) + \beta_{14}(GPA\_year1 * female) + \beta_{14}(GPA\_year1 * age\_of\_entry) + \beta_{14}(campus3 * totcredits\_year2) + sq\_totalcreds\_2 + sq\_gpa + u$$

The above regression has a R^2 of .5203 which means that 52.03% of the variation in GPA_year2 can be explained by the model's explanatory variables. Partial effects can be found

on **Table 4** in the appendix. The regressors and their effects in this model vary quite a bit from what they were in the first regression for year 1. Identifying as female in one's second year has a partial effect of 0.2273874 – (0.0557948 * GPA_year1). The higher GPA_year1 is, the lower GPA_year2 is expected to be. The expected benefit is actually completely negated if the student got above a 4.075 GPA in year 1. Age also plays a different role with a partial effect of -0.0300895 – (.0005757 * hsgrade_pct) + (.018611 * GPA_year1). In terms of hsgrade_pct, the greater the percentile, the more old age becomes harmful; the opposite is true for GPA_year1. The regressors bp_asia and l_english also play a significant role in this regression with partial effects -.0722697 and .0545931. Basically, students born in Asia are expected to have a lower second year GPA than their elsewhere born counterparts and native English-speaking students are expected to have a higher second year GPA. Subsequent probit and logit regressions confirm these finding and their results can be seen in **Table 4** in the appendix.

Conclusion

The demographic backgrounds of students enrolled in college are not supposed to have a direct effect on GPA; everyone should have a fair shot regardless of the things they cannot change. Gender in the first regression is seen to have a negative effect on GPA_year1, but by the second regression, it actually appears to be having a positive effect on GPA_year2. Age in both regressions appears to be having a negative effect on student's GPAs at low limits. Birthplace does not appear to affect regression 1 but in regression 2, being born in Asia may actually result in a lower second year GPA. Language is similar in this regard except that in regression 2, native English-speakers can actually benefit in terms of a higher second year GPA. With all this being said, demographic background certainly plays a role in determining a student's academic success but it is most certainly a minute one. Causality also could not be determined in this study and so

it is unclear whether or not discrimination is the reason for these effects. For example, non-native English speakers could be being discriminated against by their peers, but they could also simply be having a harder time understanding all the more complex wordings and structures found in upper-level college course. It is unclear, but regardless, potential discrimination can be reduced via racial bias testing for teachers, community get-togethers, and awareness campaigns.

Appendix

Table 4: First Regression (GPA_year1)

Variable Name	OLS (pop. model) GPA_year1	OLS (final model) GPA_year1	Probit above_avg	Logit above_avg
totcredits_year1	. 1776907 (.0081397)***	.4457217 – (-0.0303884* totcredits_year1) (.147488)***	.118142 (.1224558)	.1363707 (.1294841)
hsgrade_pct	.0168028 (.0001393)***	.0097886 + (0 .0000675* hsgrade_pct) (.0005481)***	.0050366 (.0003869)***	.0048813 (.000392)***
campus1	-.1085456 (.0100817)***	-.0868486 (.0101977)***	-.0459602 (.0081903)***	-.0478157 (.0085149)***
campus3	.0431882 (.0114566)***	.0647669 (.0117276)***	.0282209 (.0093272)***	.0288275 (.0096515)***
female	-.077633 (.0073838)***	-.0753478 (.0072599)***	-.0529888 (.0059197)***	-.0551369 (.0062034)***
l_english	.0002555 (.0088254)	-	-	-
l_french	. 0466231 (.0484195)	-	-	-
bp_america	.1274929 (.0785009)	-	-	-
bp_asia	-.00913 (.0224399)	-	-	-
bp_canada	-.014008 (.0187655)	-	-	-
age_at_entry	-.0146339 (.0050001)***	-.0164342 (.0164293)**	-.0031387 (.0039201)	-.003431 (.0040987)
_cons	1.235463 (.1040528)***	.9222972 (.3879579)***	.5187893 (.0023148)***	.5183682 (.0023231)***

- denotes a statistically insignificant variable not used in later models

* denotes significance at the 90% level

** denotes significance at the 95% level

*** denotes significance at the 99% level

Figure 1

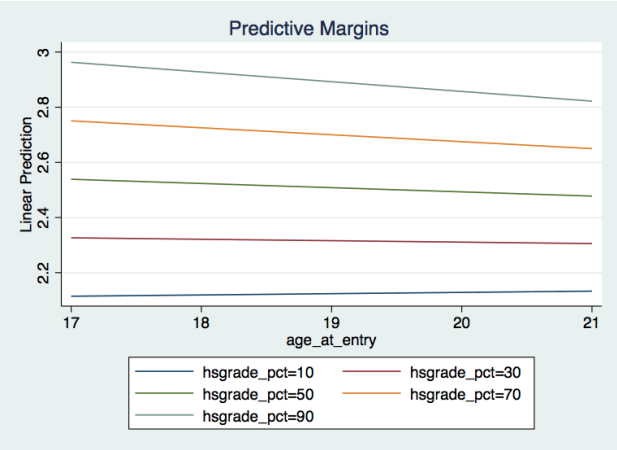
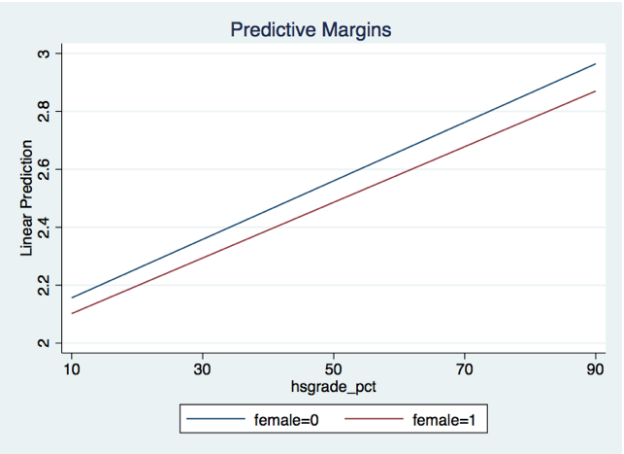


Figure 2



Linear Prediction in this case is equal to GPA_year1

Table 4: Second Regression

Variable Name	OLS (pop. model) GPA_year2	OLS (final model) GPA_year2	Probit above_avg2	Logit above_avg2
GPA_year1	.6805048 (.0058964)***	.232326 + (.0768891* GPA_year1) (.0365638)***	-.1149309 (.0308467)**	-.1263418 (.0319789)***
totcredits_year2	.0571172 (.0045961)***	-.0566617+ (.016048 * totcredits_year2) (.0175357)***	-.0588569 (.0095552)***	-.0566975 (.00846)***
totcredits_year1	.036017 (.0071623)***	.0343537 (.0071589)***	.0156927 (.0040552)***	. 0156467 (.0036319)***
hsgrade_pct	. 0034936 (.0001522)***	.0033503 (.0001535)***	.0014163 (.0001053)***	.0013697 (.000093)***
campus1	-.1381586 (.0085494)***	-.1357325 (.0084471)***	-.0448721 (.0057145)***	-.0446971 (.0051107)***
campus3	-.1160687 (.0097366)***	-.1089405 (.0095056)***	-.0452821 (.0064026)***	-.044852 (.0057003)***
female	.0805965 (.0064421)***	.0854909 (.0064608)***	.0382773 (.0041494)***	.0375219 (.0037208)***
l_english	.055503 (.0074814)***	. 0545931 (.0072086)***	.0329558 (.0046865)***	.0333768 (.0042577)***
l_french	. 0452989 (.0397043)	-	-	-
bp_america	-.0185003 (.0620508)	-	-	-
bp_asia	-.0489209 (.0192932)***	- 0722697 (.0123441)***	-.0420598 (.0078793)***	-.0426854 (.0072269)***
bp_canada	.0237914 (.0162875)	-	-	-
age_at_entry	-.0129279 (.0042325)***	-.0128925 (.0041984)**	-.0038523 (.0025536)	-.0040568 (.0022985)
_cons	.4671131 (.0905616)***	1.498637 (.2993027)***	.5375208 (.0021045)***	.5368078 (.0021082)***

Do File Below:

```
clear all

log using mylog, text replace

cd "/Users/andreweeckman/Desktop/"
use "/Users/andreweeckman/Desktop/canadiancollege.dta"

tab sex, gen(gender)
rename gender1 female
rename gender2 male

tab mtongue, gen(mtongue)
rename mtongue1 l_english
rename mtongue2 l_french
rename mtongue3 other

tab birthplace, gen(bp)
rename bp1 bp_america
rename bp2 bp_asia
rename bp3 bp_canada
rename bp4 bp_other

sum

hist GPA_year1
graph export "GPA_year1_dist.png", replace

hist GPA_year2
graph export "GPA_year2_dist.png", replace

corr totcredits_year1 hsgrade_pct campus1 campus3 female l_english l_french bp_america
bp_asia bp_canada age_at_entry

* How do externalities affect the GPA of students in the first year at college?

* OLS (Year 1):
```

```
***** OLS (1st Model)
*************************************************************************
*******
regress GPA_year1 totcredits_year1 hsgrade_pct campus1 campus3 i.female l_english l_french
bp_america bp_asia bp_canada age_at_entry, robust

test l_english l_french bp_america bp_asia bp_canada
*************************************************************************
**********************************


***** OLS (Final Model)
*************************************************************************
*******
gen sq_hsgrade_pct = hsgrade_pct * hsgrade_pct
gen sq_totalcreds = totcredits_year1 * totcredits_year1

corr totcredits_year1 hsgrade_pct campus1 campus3 female age_at_entry

regress GPA_year1 totcredits_year1 hsgrade_pct i.campus1 i.campus3 i.female age_at_entry
i.female#c.hsgrade_pct c.hsgrade_pct#c.totcredits_year1 c.age_at_entry#c.hsgrade_pct
i.campus1#c.totcredits_year1 i.campus3#c.totcredits_year1 sq_totalcreds sq_hsgrade_pct, robust

margins, dydx(*)

predict gpa_hat, residuals
scatter gpa_hat hsgrade_pct
graph export "hsgrade_residuals.png", replace

margins, dydx(age_at_entry) at(hsgrade_pct=(10(20)90))
margins, at(age_at_entry=(17 21) hsgrade_pct=(10(20)90)) vsquish
marginsplot, noci x(age_at_entry) recast(line) xlabel(17(1)21)
graph export "age_hs.png", replace

margins, dydx(hsgrade_pct) at(totcredits_year1=(3(1)6))
margins, at(hsgrade_pct=(10(20)90) totcredits_year1=(3(1)6)) vsquish
marginsplot, noci x(hsgrade_pct) recast(line) xlabel(10(20)90)
graph export "hs_creds.png", replace

margins, dydx(hsgrade_pct) at(female=(0 1))
```

```
margins, at(hsgrade_pct=(10(20)90) female=(0 1)) vsquish
marginsplot, noci x(hsgrade_pct) recast(line) xlabel(10(20)90)
graph export "effect_of_gender_hs.png", replace

margins, dydx(totcredits_year1) at(campus1=(0 1))
margins, at(totcredits_year1=(3(1)6) campus1=(0 1)) vsquish
marginsplot, noci x(totcredits_year1) recast(line) xlabel(3(1)6)
graph export "effect_of_campus1_creditsyear1.png", replace

margins, dydx(totcredits_year1) at(campus3=(0 1))
margins, at(totcredits_year1=(3(1)6) campus3=(0 1)) vsquish
marginsplot, noci x(totcredits_year1) recast(line) xlabel(3(1)6)
graph export "effect_of_campus3_creditsyear1.png", replace

*************************************************************************************
***********************************

***** GENERATE DUMMY VARIABLE ABOVE_AVG
*************************************************************************
egen meanGPA = mean(GPA_year1)
gen above_avg = 0
replace above_avg = 1 if GPA_year1 > meanGPA
replace above_avg = 0 if GPA_year1 < meanGPA

sum above_avg
*************************************************************************************
***********************************

***** PROBIT REGRESSION
*************************************************************************************
*************
probit above_avg totcredits_year1 hsgrade_pct i.campus1 i.campus3 i.female age_at_entry
i.female#c.hsgrade_pct c.hsgrade_pct#c.totcredits_year1 c.age_at_entry#c.hsgrade_pct
i.campus1#c.totcredits_year1 i.campus3#c.totcredits_year1 sq_totalcreds sq_hsgrade_pct, robust

margins, dydx(hsgrade_pct) at(totcredits_year1 campus1 campus3 female age_at_entry)
margins, dydx(campus1) at(totcredits_year1 campus3 female age_at_entry hsgrade_pct)
margins, dydx(campus3) at(totcredits_year1 female age_at_entry hsgrade_pct campus1)
margins, dydx(female) at(totcredits_year1 age_at_entry hsgrade_pct campus1 campus3)
margins, dydx(age_at_entry) at(totcredits_year1 hsgrade_pct campus1 campus3 female)
```

```
margins, dydx(totcredits_year1) at(hsgrade_pct campus1 campus3 female age_at_entry)

margins, dydx(_cons)

predict above_avg_prohat
hist above_avg_prohat
graph export "above_avg_probit.png", replace
*************************************************************************************
***********************************

***** LOGIT REGRESSION
*************************************************************************************
*************
logit above_avg totcredits_year1 hsgrade_pct i.campus1 i.campus3 i.female age_at_entry
i.female#c.hsgrade_pct c.hsgrade_pct#c.totcredits_year1 c.age_at_entry#c.hsgrade_pct
i.campus1#c.totcredits_year1 i.campus3#c.totcredits_year1 sq_totalcreds sq_hsgrade_pct, robust

margins, dydx(hsgrade_pct) at(totcredits_year1 campus1 campus3 female age_at_entry)
margins, dydx(campus1) at(totcredits_year1 campus3 female age_at_entry hsgrade_pct)
margins, dydx(campus3) at(totcredits_year1 female age_at_entry hsgrade_pct campus1)
margins, dydx(female) at(totcredits_year1 age_at_entry hsgrade_pct campus1 campus3)
margins, dydx(age_at_entry) at(totcredits_year1 hsgrade_pct campus1 campus3 female)
margins, dydx(totcredits_year1) at(hsgrade_pct campus1 campus3 female age_at_entry)

margins, dydx(_cons)

predict above_avg_lohat
hist above_avg_lohat
graph export "above_avg_logit.png", replace
*************************************************************************************
***********************************


* The effect of having a higher high school gpa is different if a person is male or female. The
slopes of the regression lines between GPA and hsgrade_pct are different for different genders.

***** OLS Year 2 (1st Model)
*************************************************************************************
regress GPA_year2 GPA_year1 totcredits_year2 totcredits_year1 hsgrade_pct campus1 campus3
i.female l_english l_french bp_america bp_asia bp_canada age_at_entry, robust
```

```
test l_english l_french bp_america bp_asia bp_canada          //Test all dummy variables with
insignificant p values
test l_english bp_asia bp_canada                                                    //Test the exact value
of dummy variables with sig. p values
*************************************************************************************
***********************************

***** OLS Year 2(Final Model)
*************************************************************************************
*


gen sq_totalcreds_2 = totcredits_year2 * totcredits_year2
gen sq_gpa = GPA_year1 * GPA_year1

corr GPA_year1 totcredits_year2 totcredits_year1 hsgrade_pct campus1 campus3 female
l_english bp_asia age_at_entry

regress GPA_year2 GPA_year1 totcredits_year2 totcredits_year1 hsgrade_pct i.campus1
i.campus3 i.female i.l_english i.bp_asia age_at_entry c.hsgrade_pct#c.age_at_entry
c.GPA_year1#i.female c.GPA_year1#c.age_at_entry i.campus3#c.totcredits_year2
sq_totalcreds_2 sq_gpa, robust

margins, dydx(*)

predict gpa_hat2, residuals

scatter gpa_hat2 GPA_year1
graph export "gpa1_residuals_inyear2.png", replace

margins, dydx(age_at_entry) at(hsgrade_pct=(10(20)90))
margins, at(age_at_entry=(17 21) hsgrade_pct=(10(20)90)) vsquish
marginsplot, noci x(age_at_entry) recast(line) xlabel(17(1)21)
graph export "age_hs_inyear2.png", replace

margins, dydx(GPA_year1) at(female=(0(1)1))
margins, at(GPA_year1=(0(.5)4.5) female=(0(1)1)) vsquish
marginsplot, noci x(GPA_year1) recast(line) xlabel(0(.5)4.5)
```

```
graph export "effect_of_gender_gpa_inyear2.png", replace

margins, dydx(GPA_year1) at(age_at_entry=(17(1)21))
margins, at(GPA_year1=(0(.5)4.5) age_at_entry=(17(1)21)) vsquish
marginsplot, noci x(GPA_year1) recast(line) xlabel(0(.5)4.5)
graph export "effect_of_age_gpa_inyear2.png", replace

margins, dydx(totcredits_year2) at(campus3=(0 1))
margins, at(totcredits_year2=(3(1)6) campus3=(0 1)) vsquish
marginsplot, noci x(totcredits_year2) recast(line) xlabel(3(1)6)
graph export "effect_of_campus3_creditsyear2.png", replace

*************************************************************************************
***********************************

***** GENERATE DUMMY VARIABLE ABOVE_AVG2
*************************************************************************
egen meanGPA2 = mean(GPA_year2)
gen above_avg2 = 0
replace above_avg2 = 1 if GPA_year2 > meanGPA2
replace above_avg2 = 0 if GPA_year2 < meanGPA2

sum above_avg2
*************************************************************************************
***********************************

***** PROBIT REGRESSION YEAR 2
*************************************************************************************
******
probit above_avg2 GPA_year1 totcredits_year2 totcredits_year1 hsgrade_pct i.campus1
i.campus3 i.female i.l_english i.bp_asia age_at_entry c.hsgrade_pct#c.age_at_entry
c.GPA_year1#i.female c.GPA_year1#c.age_at_entry i.campus3#c.totcredits_year2
sq_totalcreds_2 sq_gpa, robust

margins, dydx(GPA_year1) at(totcredits_year2 totcredits_year1 hsgrade_pct campus1 campus3
female l_english bp_asia  age_at_entry)
margins, dydx(totcredits_year2) at(totcredits_year1 hsgrade_pct campus1 campus3 female
l_english bp_asia  age_at_entry GPA_year1)
margins, dydx(totcredits_year1) at(hsgrade_pct campus1 campus3 female l_english bp_asia
age_at_entry GPA_year1 totcredits_year2)
```

```
margins, dydx(hsgrade_pct) at(campus1 campus3 female l_english bp_asia  age_at_entry
GPA_year1 totcredits_year2 totcredits_year1)
margins, dydx(campus1) at(campus3 female l_english bp_asia  age_at_entry GPA_year1
totcredits_year2 totcredits_year1 hsgrade_pct)
margins, dydx(campus3) at(female l_english bp_asia age_at_entry GPA_year1 totcredits_year2
totcredits_year1 hsgrade_pct campus1)
margins, dydx(female) at(l_english bp_asia age_at_entry GPA_year1 totcredits_year2
totcredits_year1 hsgrade_pct campus1 campus3)
margins, dydx(l_english) at(bp_asia age_at_entry GPA_year1 totcredits_year2 totcredits_year1
hsgrade_pct campus1 campus3 female)
margins, dydx(bp_asia) at(age_at_entry GPA_year1 totcredits_year2 totcredits_year1
hsgrade_pct campus1 campus3 female l_english)
margins, dydx(age_at_entry) at(GPA_year1 totcredits_year2 totcredits_year1 hsgrade_pct
campus1 campus3 female l_english bp_asia)

margins, dydx(_cons)


predict above_avg_prohat2
hist above_avg_prohat2
graph export "above_avg_probit2_inyear2.png", replace
**********************************************************************************
***********************************

***** LOGIT REGRESSION YEAR 2
**********************************************************************************
*******
logit above_avg2 GPA_year1 totcredits_year2 totcredits_year1 hsgrade_pct i.campus1
i.campus3 i.female i.l_english i.bp_asia age_at_entry c.hsgrade_pct#c.age_at_entry
c.GPA_year1#i.female c.GPA_year1#c.age_at_entry i.campus3#c.totcredits_year2
sq_totalcreds_2 sq_gpa, robust

margins, dydx(GPA_year1) at(totcredits_year2 totcredits_year1 hsgrade_pct campus1 campus3
female l_english bp_asia  age_at_entry)
margins, dydx(totcredits_year2) at(totcredits_year1 hsgrade_pct campus1 campus3 female
l_english bp_asia  age_at_entry GPA_year1)
margins, dydx(totcredits_year1) at(hsgrade_pct campus1 campus3 female l_english bp_asia
age_at_entry GPA_year1 totcredits_year2)
margins, dydx(hsgrade_pct) at(campus1 campus3 female l_english bp_asia  age_at_entry
GPA_year1 totcredits_year2 totcredits_year1)
```

```
margins, dydx(campus1) at(campus3 female l_english bp_asia  age_at_entry GPA_year1 totcredits_year2 totcredits_year1 hsgrade_pct)
margins, dydx(campus3) at(female l_english bp_asia age_at_entry GPA_year1 totcredits_year2 totcredits_year1 hsgrade_pct campus1)
margins, dydx(female) at(l_english bp_asia age_at_entry GPA_year1 totcredits_year2 totcredits_year1 hsgrade_pct campus1 campus3)
margins, dydx(l_english) at(bp_asia age_at_entry GPA_year1 totcredits_year2 totcredits_year1 hsgrade_pct campus1 campus3 female)
margins, dydx(bp_asia) at(age_at_entry GPA_year1 totcredits_year2 totcredits_year1 hsgrade_pct campus1 campus3 female l_english)
margins, dydx(age_at_entry) at(GPA_year1 totcredits_year2 totcredits_year1 hsgrade_pct campus1 campus3 female l_english bp_asia)

margins, dydx(_cons)

predict above_avg_lohat2
hist above_avg_lohat2
graph export "above_avg_logit2_inyear2.png", replace


log close
```