Econometrics 140 Research Project

By:

Gurteg Singh

Abstract

The following research examines data collected from a Canadian university system with 3 separate campuses, all of which varied in student demographics, acceptance rates, and other criteria. I will evaluate, specifically, the disparate graduation rates across these three campuses as a dependent variable and explore potential causal factors using measures of regression, all of which were calculated using Stata programming.

Introduction and Methodology

In this project, I regressed and analyzed several variables that appear to accurately predict a student's graduation rate. To examine these, I formulated two guiding research questions (found below), followed by my data collection and analyses wherein potential relationships are pinpointed. During my analyses, I will also explore causal mechanisms that we may attribute the aforementioned relationships to before summarizing these findings in my conclusion. All data and tables used may be found in the Appendix section, followed by the utilized Stata commands.

Research Questions:

1. First, amongst an amalgam of factors between men and women, who are more likely to graduate in years 4, 5, and 6, given the same range of second-year GPA's (GPA_year2) values?
2. Using the same GPA_year2 increments, can we differentiate a student's probability of graduating from any of the college campuses in any given year?

To answer these questions, I analyzed the different types of data we were given, created an overview, and scanned for variables or values that may have outliers, tabling the values in 'Summary Statistics'. I then regressed the factors that affect the graduation rate between men and women, as well as campuses, tabling the results by the year of graduation. Afterward, I disaggregated GPA_year2 into 0.5 increments and plotted the marginal effects of women and campuses on the probability of graduating along with these values. Using various techniques in my regression including OLS binary (lp) models, Logit models and its marginal effects (dy/dx) - as seen in the attached dofile -- I was able to quantify, table, graph images, and predict across a spectrum of GPAs where they were further analyzed.

Data

First, I started with assorting the data by creating binary variables for the string terms 'sex', 'campus', and 'birthplace'. Note that here, it is important to have a base value to go off of, otherwise defining all possible sub-variables in each term will lead to collinearity. Afterward, I imported a package called Estout, a command for publication-styled regression tables that display nicely in Stata's results window or, optionally, can be exported to various formats such as CSV, RTF, HTML, or LaTeX. Here I converted all the following results into RTFs and .pngs (for the graphs) and created a 'Summary Statistics' report, with all the variables that I would regress with. The resulting table includes the number of observations, means, minimums, and maximums - all of which were important for cleaning missing data and interpreting average values.

With this, I created an equation to regress graduation rates with an LPM and Logit model, while marginally predicting at the means of each variable in years 4, 5 and 6. My results were placed in tables in the appendix. The reason I included the Logit models are that unlike the OLS models, Logit models deal with probability much better in terms of keeping the resulting probabilities between 0 and 1, as well as not having to deal with the homoskedastic assumptions being violated (which we can fix by simply adding a robust command at the end). I repeated the models for years 5 and 6.

With the tables being completed, I moved onto the graphs, placing our dependent variable, probabilities of graduation on the Y-axis, and GPA_year2 values on the X-axis. In these equations I used the Logit models and created interaction terms with GPA_year2 and also squared it with respect to females distinguishing any effect GPAs could have on a woman's chance at graduating. Using the same set of GPAs from 0 to 4.3 with 0.5 increments I combined the three graphs so we could analyze the marginal effects of women compared to men at each interval altogether. The tables in the Appendix were labeled '4-Year Graduation Regressions', '5-Year Graduation Regressions', and '6-Year Graduation Regressions', and the graphs under the title 'Probability of Graduating Male vs Female' were labeled '4-Year', '5-Year', and '6-Year', for years 4, 5, and 6 respectively.

For the second research question, I again used Logit models, regressing the graduation rates while creating interaction terms this time with GPA_year2, and GPA_year2 squared on campus which I put in as a dummy. Here we can emphasize the marginal effects campuses have at different GPA_year2 values for each year. Again, using the same Y-axis and X-axis, I plotted out the three campuses' marginal effects (as a dummy variable 'campus') while keeping the other variables at their mean value for each GPA_year2 increment, under the title ‘Probability of Graduating Across Campuses’, labeled ‘4-Year’, ‘5-Year’, and ‘6-Year’.

Analysis & Discussion

Reading these tables has to be done carefully, as some variables are binary and have X values of 0 and 1. When a binary variable’s value is 0, its alternative value is automatically inputted in the constant, other variables such as GPA are non-binary and take values between 0 and 4.3. Based on table 1, ‘Summary Statistics’ table, we can derive a myriad amounts of information. Most notably, it reveals 22,262 observations for each variable (indicating no missing data), and the GPA range will lie between 0 and 4.3. Additional findings include: graduation rate tends to increase from 54% to 78% to 85%, in years 4, 5, and 6. GPAs rise from an average of 2.52 in the first year, to 2.56 in the second year, 62.5% of overall student population female, and Campus 2 makes up about 17.1% of the student body while Campus 3 only makes up 12.4%.

Table 2 then accounts for the first regression that I conduct on the graduation rates of students in 4 years as shown below.

$$\begin{aligned} Graduation_i = {} & \beta_0 + \beta_1 HighSchool\_pct_i + \beta_2 GPA\_year1_i + \beta_3 GPA\_year2_i + \beta_4\, Female_i \\ & + \beta_5 Language_i + \beta_6 Age\_at\_entry_i + \beta_7 Probation\_year1 + \beta_8 Probation\_year2_i \\ & + \beta_9 Suspended\_year1_i + \beta_{10} Suspended\_year2_i + B_{11} Campus2_i + \beta_{12} Campus3_i \\ & + \beta_{13} Canada \end{aligned}$$

The betas are percentages as regressed on a binary variable, on the probability that graduation equals 1. Furthermore, both GPA’s of year 1 and 2 have betas of 4.45% and 11.9% which increase the probability of graduating for the OLS model. The Logit model demonstrates the marginal effect of GPA with other variables constant at the mean values-- this results in betas of 5.02% and 13.6% respectively. Notably, the Betas associated with Logit values, are combinations of the beta and it's partial effect as they depend on the x values associated with the betas. Other things we can note from this table is that identification as a female drastically improves a student's chance of graduating in comparison to a man, where women have a 10% to 11.9% increase in the likelihood of graduating in 4 years between the OLS and Logit models. Also, enrolling on campus 2 and 3 decreases the likelihood of graduation in 4 years by 2.76% and 8.74% respectively in comparison to campus 1 based solely on the Logit model's marginal effects.

Table 3 produces a similar regression for graduation rates of students in their 5th year. Here we can see a similar trend with women and GPA_year2 having positive marginal effects towards graduating in 5 years, however, GPA_year1 and enrolling in campus 2 are no longer significant and their betas seem to differ between the models, unlike all the previous betas that were significant at least at a 1% level. Furthermore, students enrolled in Campus 3 illustrate a 2.3% marginally increased chance of graduating in 5 years versus students in Campus 1.

Table 4 regresses graduation rates for students in their 6th year. Again, we can see that GPA_year2 and being a woman both have positive effects on this particular graduation rate, albeit lower than years 4 and 5. Also, Campuses 2 and 3 both have lowered levels of significance and their betas are approaching 0 in comparison to campus 1. The decreasing gap between the variables over time could be a result of students dropping out early with relatively low GPAs while students in their 4th, 5th, or 6 are more determined to graduate.

The probability graphs comparing men to women also reflect what's happening in the betas of variables I have described above. In determining the equation for this regression, I created a dummy variable for interactions with women and GPA_year2, and also women and

GPA_year2 squared, to help emphasize any irregularities or differences women face at different GPA_year2 values. Using the Logit model, I regressed the given variables on the probability of graduating and produced three graphs titled ‘Male vs Female’ where we can see the marginal effects of women compared to men at each GPA_year2 value. Evidenced by the graphs, women have an increased chance of graduating in *any* particular year at college compared to men, although the gap between them recedes as the years progress. Men and women with higher GPAs tend to end with slightly lower chances of graduating in a given year, but both max out near 60%, 80%, and 85% in years 4, 5, and 6 respectively. This could allude to the fact that students with higher GPAs take fewer units to better their grades, and in doing so, increase the amount of time needed to complete their majors.

The second set of graphs titled “Probability of Graduating Between Campuses,” is based on a regression similar to the first except for the change in interaction terms where now we interact GPA_year2 and GPA_year2 squared on the dummy variable campus so we can emphasize the effect of GPA at different campuses on a student’s of graduating in different years. Under ‘Probability of Graduating Across Campuses’ in the ‘4-Year’ graph, we can see that the probability of graduating is lower at a low GPA value and increases till a GPA of around 3.25 before decreasing slightly thereafter, for all campuses. Campus 1 and 2 also have noticeably lower graduation probabilities when compared to Campus 3 which maintains an increased gap in graduation until 2.0 where all the campuses intersect. This could be due to the fact that Campus 1 has a lower acceptance rate-- given in the prompt-- leading to a higher academic rigor and standard, where lower grades are penalized harsher. At GPA values greater than 2.0, campuses 1 and 2 have a higher chance of probability of graduating peaking near 60% with a GPA near 3.25, whereas Campus 3 struggles behind peaking near 50% at a GPA of 3.0. As the years progress, we can see that the Campuses' graduation rates become nearly identical, peaking near 80% in year 5, and 90% in year 6.

Conclusion

Given these results, I can conclude that women have an overall higher chance of graduating in any given year than men with the same GPAs, although the gap lessens through years 5 and 6. Across campuses, we see that Campus 1 & 2 tend to have an overall higher probability at least for students with GPA values > 2 graduating within 4 years. Nonetheless, in years 5 and 6 graduation rates become almost identical with all 3 campuses having similar graduation probabilities. These trends can likely be attributed to a combination of different lifestyles between male and female students across each campus in addition to features such as part-time/full-time statuses, major choices, and drop-outs.

Appendix

Table 1. *Summary Statistics*

	count	mean	sd	min	max
gradin4	22262	.5444255	.4980337	0	1
gradin5	22262	.7853293	.410603	0	1
gradin6	22262	.8527536	.3543594	0	1
hsgrade_pct	22262	50.59397	28.68976	1	100
GPA_year1	22262	2.521158	.8295679	0	4.3
GPA_year2	22262	2.56817	.8257197	0	4.3
female	22262	.6255053	.484003	0	1
age_at_entry	22262	18.73938	.6896549	17	21
probation_year1	22262	.1202947	.3253131	0	1
probation_year2	22262	.0423143	.2013097	0	1
suspended_year1	22262	.0001348	.011608	0	1
suspended_year2	22262	.0489174	.2157003	0	1
campus2	22262	.1712335	.3767213	0	1
campus3	22262	.1941874	.3955827	0	1
canada	22262	.8794358	.3256274	0	1

Table 2 *4-Year Graduation Regressions*

	(1) OLS	(2) Logit	(3) MFX
main			
hsgrade_pct	0.000111	0.000266	0.0000663
	(0.000145)	(0.000683)	(0.000171)
GPA_year1	0.0445^{***}	0.201^{***}	0.0502^{***}
	(0.00691)	(0.0329)	(0.00822)
GPA_year2	0.119^{***}	0.546^{***}	0.136^{***}
	(0.00591)	(0.0285)	(0.00711)
female	0.100^{***}	0.476^{***}	0.119^{***}
	(0.00638)	(0.0305)	(0.00762)
English	0	0	0
	(.)	(.)	(.)
French	-0.0653	-0.310	-0.0770
	(0.0432)	(0.202)	(0.0495)
Other	0.0344^{***}	0.167^{***}	0.0416^{***}
	(0.00766)	(0.0370)	(0.00918)
age_at_entry	-0.0128^{**}	-0.0654^{**}	-0.0163^{**}
	(0.00451)	(0.0219)	(0.00546)
probation_year1	-0.0866^{***}	-0.369^{***}	-0.0920^{***}
	(0.0149)	(0.0716)	(0.0179)
probation_year2	-0.163^{***}	-0.944^{***}	-0.236^{***}
	(0.0174)	(0.102)	(0.0256)
suspended_year1	-0.135	0	0
	(0.264)	(.)	(.)
suspended_year2	-0.181^{***}	-2.991^{***}	-0.747^{***}
	(0.0200)	(0.268)	(0.0673)
campus2	-0.0238^{**}	-0.111^{**}	-0.0276^{**}
	(0.00880)	(0.0418)	(0.0104)
campus3	-0.0734^{***}	-0.350^{***}	-0.0874^{***}
	(0.00842)	(0.0403)	(0.0101)
canada	-0.0126	-0.0663	-0.0166
	(0.0103)	(0.0500)	(0.0125)
Constant	0.345^{***}	-0.589	
	(0.0882)	(0.426)	
Observations	22262	22259	22259

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3 *5-Year Graduation Regressions*

	(1) OLS	(2) Logit	(3) MFX
main			
hsgrade_pct	-0.000346^{**}	-0.00237^{**}	-0.000363^{**}
	(0.000115)	(0.000865)	(0.000132)
GPA_year1	-0.00743	0.00291	0.000445
	(0.00549)	(0.0418)	(0.00640)
GPA_year2	0.0978^{***}	0.679^{***}	0.104^{***}
	(0.00470)	(0.0339)	(0.00511)
female	0.0529^{***}	0.385^{***}	0.0589^{***}
	(0.00507)	(0.0377)	(0.00574)
English	0	0	0
	(.)	(.)	(.)
French	0.0459	0.363	0.0514
	(0.0343)	(0.284)	(0.0355)
Other	0.0285^{***}	0.220^{***}	0.0326^{***}
	(0.00609)	(0.0472)	(0.00677)
age_at_entry	-0.0123^{***}	-0.0992^{***}	-0.0152^{***}
	(0.00358)	(0.0275)	(0.00420)
probation_year1	-0.0481^{***}	-0.324^{***}	-0.0496^{***}
	(0.0118)	(0.0808)	(0.0124)
probation_year2	-0.294^{***}	-1.120^{***}	-0.171^{***}
	(0.0138)	(0.0831)	(0.0129)
suspended_year1	-0.333	0	0
	(0.210)	(.)	(.)
suspended_year2	-0.540^{***}	-2.563^{***}	-0.392^{***}
	(0.0159)	(0.126)	(0.0206)
campus2	-0.00344	-0.0204	-0.00313
	(0.00699)	(0.0525)	(0.00803)
campus3	0.0211^{**}	0.151^{**}	0.0230^{**}
	(0.00669)	(0.0511)	(0.00782)
canada	-0.0272^{***}	-0.228^{***}	-0.0349^{***}
	(0.00818)	(0.0653)	(0.00999)
Constant	0.825^{***}	1.774^{***}	
	(0.0701)	(0.537)	
Observations	22262	22259	22259

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4 *6-Year Graduation Regressions*

	(1) OLS	(2) Logit	(3) MFX
main			
hsgrade_pct	-0.000228^{*}	-0.00184	-0.000182
	(0.0000971)	(0.00103)	(0.000102)
GPA_year1	-0.0166^{***}	-0.0702	-0.00694
	(0.00463)	(0.0501)	(0.00495)
GPA_year2	0.0807^{***}	0.765^{***}	0.0756^{***}

	(0.00396)	(0.0393)	(0.00379)
female	0.0386***	0.388***	0.0384***
	(0.00428)	(0.0445)	(0.00438)
English	0	0	0
	(.)	(.)	(.)
French	0.0402	0.472	0.0407
	(0.0289)	(0.357)	(0.0253)
Other	0.0207***	0.220***	0.0209***
	(0.00513)	(0.0560)	(0.00511)
age_at_entry	-0.0109***	-0.122***	-0.0120***
	(0.00302)	(0.0323)	(0.00319)
probation_year1	-0.0290**	-0.331***	-0.0327***
	(0.00998)	(0.0940)	(0.00928)
probation_year2	-0.286***	-1.218***	-0.120***
	(0.0116)	(0.0878)	(0.00894)
suspended_year1	-0.411*	0	0
	(0.177)	(.)	(.)
suspended_year2	-0.581***	-2.398***	-0.237***
	(0.0134)	(0.118)	(0.0128)
campus2	-0.00214	-0.0157	-0.00155
	(0.00589)	(0.0625)	(0.00618)
campus3	0.0152**	0.139*	0.0138*
	(0.00564)	(0.0596)	(0.00589)
canada	-0.0134	-0.157*	-0.0155*
	(0.00690)	(0.0762)	(0.00753)
Constant	0.927***	2.684***	
	(0.0591)	(0.633)	
Observations	22262	22259	22259

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$