UCDAVIS
COLLEGE OF LETTERS AND SCIENCE

Aman Singh
Econometrics
Empirical Project

6/3/2020

Introduction:

As a college student at the University of California, Davis, I have always wondered what really limits a students abilities and how it is truly related to a students academic performance. Was it their gender? Was it their native language? Birth Place? Age? There are a lot of factors in life that can truly make people very different from each other which can result in a different type of academic performance. There are also other factors that we cannot even account for such as intellectual curiosity, whether they truly like their major (most people forcefully do stem to make more money), or even their mental intellect. This part in fact is represented in the following example of a multiple linear regression of 2 variables:

$$\hat{Y} = B_0 + B_1X_1 + B_2X_2 + U_t$$

Where U_t is shown as the unsystematic component that we cannot account for.

Literature Review:

Student performance has been a very hotly talked about topic for an extremely long time. There have been many research papers done on trying to best understand how a students performance can vary based on their data. In fact in an research article, it is stated that "In 2002, it was reported that more than 30% of first-year students did not return for their second year of college, and only 40% are reported to actually complete their degree and graduate."(Goodreads) In realization, each college is different and each student is different. It is almost impossible to be extremely accurate in trying to depict a students performance based on data. There is even data that most people will not be willing or skew to share such as how much studying they do per week, their sleep patterns and maybe even if they have a proper breakfast, lunch and dinner everyday. There are just too many factors that are out of our control in order to have an accurate depiction of a students performance. Florida State University in fact recently came out with a way to figure out student performance. (L.,Theresa) However, it is almost guaranteed to not be extremely accurate as well because it is based on a student's self-reporting instrument that then uses the information to measure academic success in college students. This will result in some biases and thus incorrectness. However, this seems to be the best tool on the internet that can best visualize a students intellect because it also evaluates many constructs previously obtained only by administering numerous individual measures.

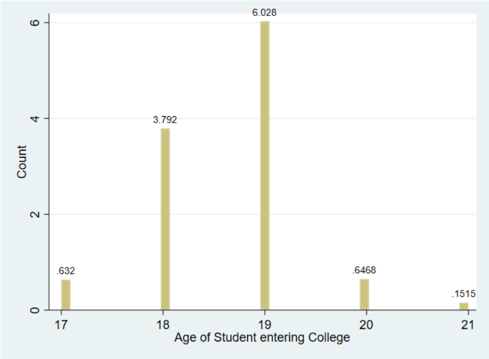
Analyzing the dataset:

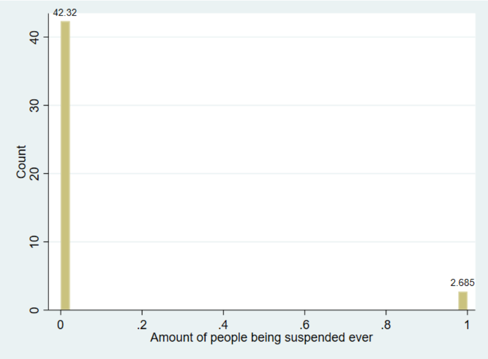
After wondering for a while, I decided to use Stata to carry out my observations. I obtained a cross-sectional dataset from online about a large Canadian University which has 3 main campuses. Campus 1 is the main campus and Campus 2 and 3 are the side campuses. The variables in this dataset are listed below:

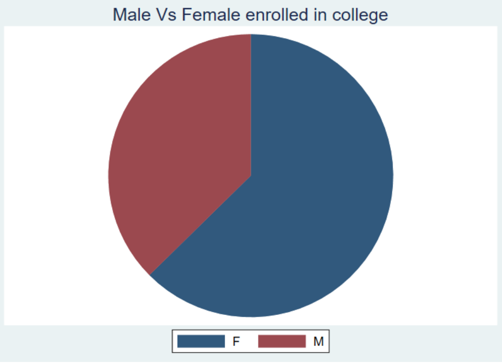
Variable:	**Explanation:**
campus1	If the student is in campus 1
campus2	If the student is in campus 2
campus3	If the student is in campus 3
hsgrade_pct	HS grade percentile ranking
birthplace	Where the student was born
GPA_year1	First year GPA of the student
GPA_year2	Second year GPA of the student
totcredits_year1	Total credits attempted in year 1

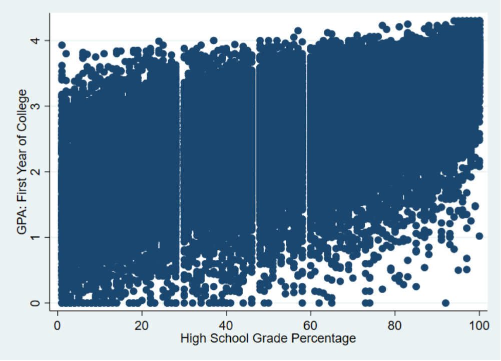
totcredits_year2	Total credits attempted in year 2
sex	Gender
mtongue	Mother tongue
age_at_entry	Age of student when entering university
gradin4	If student graduates in 4 years
gradin5	If student graduates in 5 years
gradin6	If student graduates in 6 years
gpacutoff	The cut-off gpa for the student
probation_year1	If the student was placed on probation after the 1st year
probation_year2	If the student was placed on probation after the 2nd year

After conducting some basic visual analysis and coming up with a couple of graphs:









In these graphs, I noticed that there seems to be a positive correlation with High School grade Percentage with GPA of the first year of college. I also noticed that a majority of people who also ended up being enrolled in the college were mostly Female. The first two histograms also showed that the average age of students who entered college was at age 19, which shows that it is not a normal distribution and is skewed. The other histogram also shows a majority of people were never suspended. I also took a look at the basic summary of the whole dataset and saw that most of the regressions were binary logistics with some also being categorical logistics as well such as birthplace. I also noticed that the maximum a student took credits was 9 which is considered as a part time student year which made me think that they have a completely different system than what I have experienced in the US. I also noticed that the highest GPA they can get is higher than a 4.0 where here in America the highest is a 4.0. After doing such simple analysis I decided to come up with a question to analyze the dataset with:

How do students' GPA in their first year vary with their demographic characteristics?

I believe this question is in fact very essential to many people. It potentially can be used by the college to help them with admissions to figure out who they should accept. It can also be used by other colleges to maybe have a more vivid description on how they can have their acceptances. It can mostly also be used by students themselves to use as a tool to have a higher chance of getting into college. However, once again there are always downsides of having a regression figuring out conclusions. *The Fires: How a Computer Formula, Big Ideas, and the Best of Intentions Burned Down New York City--and Determined the Future of Cities* by Joe Flood in fact very explicitly explains how a regression is just not enough to come to conclusions about a certain result. We cannot calculate that a trend will continue based on past data, however we can try our best to use a regression for advice, not follow it word for word. After looking at all the variables in the data, I decided to start the regression by adding all the terms I believe will accurately describe students' performance with their demographic characteristics. I first noticed that in order to include some variables in my regression I have to first change some variables such as mother tongue, country, language, from a categorical variable to individual variables for ease of regression analysis. I then proceeded to rename the variables so it can be easier for me to know which variable represents what in the regression. I then dropped the original variables to not be confused between them. In my regression as well, I personally chose not to use second year variables as since we are trying to only estimate the gpa of the 1st year, it makes sense to only include variables that we can account for in the first year. My first regression, I decided to do the following:

$$GP\hat{A}year1 = \beta_0 + GPA.year1X_1 + totcredits.year1X_2 + age.at.17X_3 + age.at.18X_4 + age.at.19X_5 + age.at.20X_6 + age.at.21X_7 + gradin4X_8 + gradin5X_9 + gradin6X_{10} + probation.year1X_{11} + suspended.year1X_{12} + campus1X_{13} + campus2X_{14} + campus3X_{15} + femaleX_{16} + maleX_{17} + englishX_{18} + frenchX_{19} + other.languageX_{20} + americaX_{21} + asiaX_{22} + canadaX_{23} + other.countryX_{24} + hs.gradepctX_{25}$$

Looking at the Standard Errors of the variables, it seems that a majority of the variables are extremely statistically significant. Looking at the R^2, we see that the regression explains approximately 61% of the data and the $Adj.R^2$ explaining approximately 61% of the data as well. I also decided to remove some variables due to multicollinearity such as age.at.17, campus1, male,other.language, and America. Before conducting any tests, I decided to check for Heteroskedastic errors and what I did was conduct a 2way scatter plot of the estimated residuals vs the fitted variables and then against each of the X variables. Looking at the Fitted vs the Residuals, I saw that there seems to be a decreasing trend in the graph as the fitted values increase.(These figures are located at the index to save space.) This shows a classic type of "coning" heteroskedasticity that is seen in the graphs. Thus, I redid the regression but using the "robust" standard error. I also in this graph do not have to worry about any significant outliers as well.

When I looked at the statistical significance, I realized that a couple of variables such as French, other countries, and Asia do not seem statistically significant. I decided to then conduct an F-Test. The reasoning behind is that I want to get rid of variables that are not important to my regression. I conducted the F-test based on the fact that the null hypothesis being:

$$H_0 = \beta_{french} = \beta_{asia} = \beta_{other.country} = \beta_{age.at.21} = \beta_{canada} = \beta_{english} = \beta_{america} = \beta_{gradin6} = \beta_{age.at.17} = \beta_{other.language} = 0$$

.

Upon conducting the F-test, I saw that the p-value is .4538 thus, we fail to reject the null hypothesis, meaning that they are useless to the regression and not worth adding to our regression. I thus proceed to drop them from the regression. After dropping the variables, I ran the linear regression again and noticed this time that all the variables were statistically significant!

Prediction using updated Multiple Linear Regression:

Looking at a random observation in the dataset, I proceeded to fit all my variables based on the following to predict the Students GPA in their first year. The observation I chose to compare to was observation 12. This was the final regression which I than fit in the data:

$$GP\hat{A}year1 = \beta_0 + totcredits.year1X_1 + age.at.18X_2 + age.at.19X_3 + age.at.20X_4 + gradin4X_5 + gradin5X_6 + probation.year1X_7 + suspended.year1X_8 + campus2X_9 + campus3X_{10} + femaleX_{11} + hsgrade.pctX_{12} + suspended.everX_{13}$$

When I fit the data according to the observation, I saw that my prediction was 2.365379 and the actual students GPA was a 2.3. (I have attached a screenshot of the observation in the index). I also decided to try another random observation, observation 2560. My prediction came out to be 2.642396 while the actual students GPA for the first year was a 3.02. While the first prediction was pretty close, the second prediction was off by a significant amount. I tried to predict another random observation with it being 30,000 and saw that my prediction came out to be 2.328587 while the actual students GPA was a 2.51.

Interpreting the Final Regression

______Looking at the regression, we see that all of the variables are now statistically significant from zero. The first thing I usually do is to analyze the coefficients and instantly I notice that the GPA of the students in their first year in fact goes down if their age is either 18 or 19 or 20. The variables for probation_year1 and suspended_year 1 have major implications to the student GPA. This in fact makes perfect sense because a student is normally on probation if they have a really low GPA and have been continuously failing classes. A student also normally gets suspended if they do something that offends the University, or if they keep being on probation, so in fact it seems like they have a very high correlation. What I also noticed is that if you are a female, you will be more likely to have a worse GPA. I wonder what attributes to this, as I feel like females are typically smarter than men in general. I also noticed that High School grade percentage in fact has a huge effect in finding a students first year GPA. The higher the HS percentage, the higher the GPA becomes for the student.

StepWise Regression:

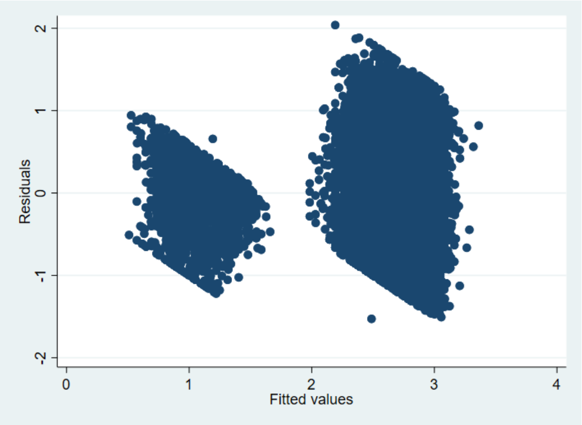
Upon finding my final regression, I wanted to test using a machine learning method called stepwise regression. Forward Stepwise regression being starting from an empty model and adding variables to it to make the AIC,BIC & AICC the lowest possible. Backward Stepwise regression is alo the same except we start from a full model and move backwards going to an empty model. After running the code, I found that they came to the same conclusion as I had upon running the model and both the forward and backward stepwise model removed French, Asia, and other variables which I also had removed using the F-Test to come to conclusions and to remove the non statistically significant variables which leads to a presumably better model fit.

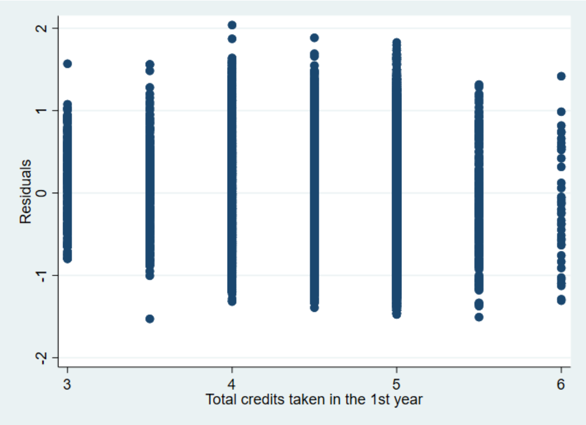
Conclusion:

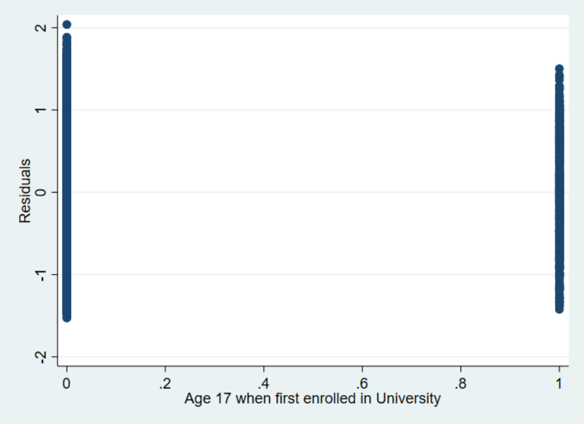
After coming to a conclusion about my regression and trying my best to fit the regression of trying to estimate the GPA of the students first year in college and comparing it to other variables, my regression, while accounting for almost 61% of the data, was still unfortunately off. It was accurate in some observations but not accurate in some other observations. The difference between my prediction could be due to omitted variable bias, however, I believe it can also be due to different circumstances. Such as right now, we are dealing through an epidemic that we can not record in a dataset. Each individual is different and there can be many factors into why someone's GPA can be higher or lower than my prediction. My prediction is only based on the data given to me. However, it does not account for many variables such as parents income, whether they had a tutor, and other such variables that can help better make the students GPA. Every individual is different and we just cannot assume that a computer can accurately predict something. It is important to take in the regression factors no doubt, but to use it as an advising type of measurement and to not come to conclusions immediately based on what a regression produces. A perfect example was in fact in New York in the 1970's where they had a group of scientists develop a regression tool to come to conclusions about where they should build firehouses in the city. This backfired massively as they assumed the data would also incorporate intellectual thought and as a result it put New York in an even worse state than before.

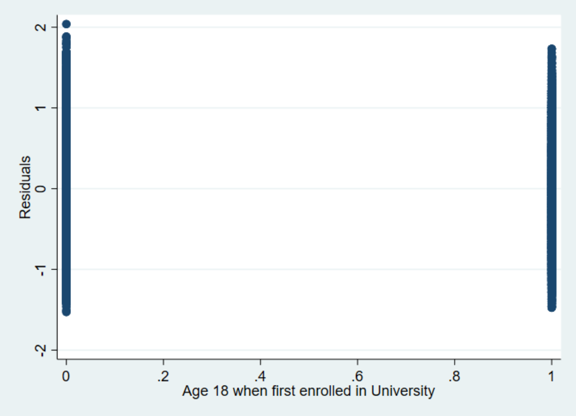
Index:

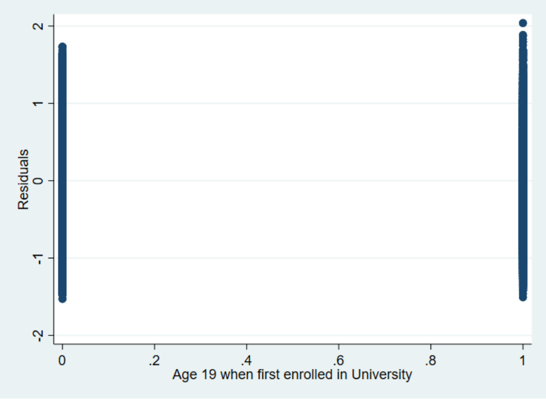
Test for Heteroskedasticity:

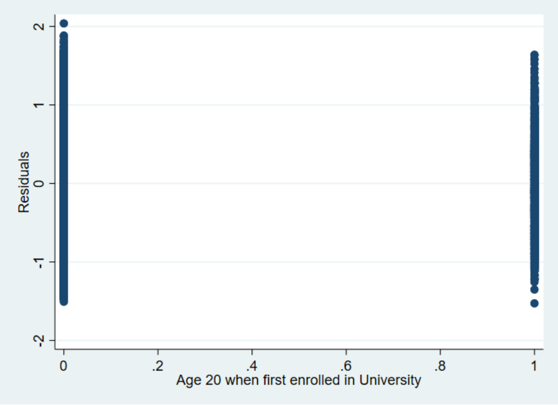


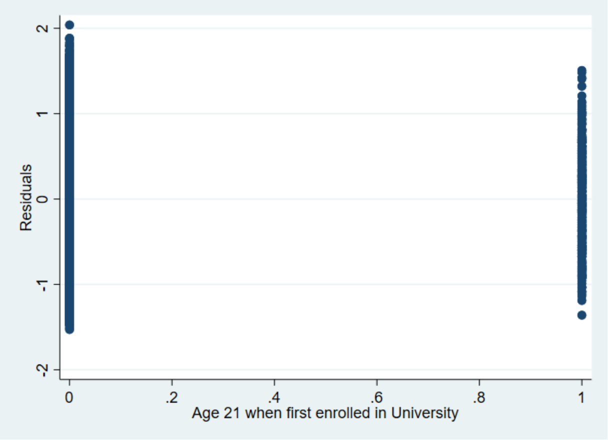


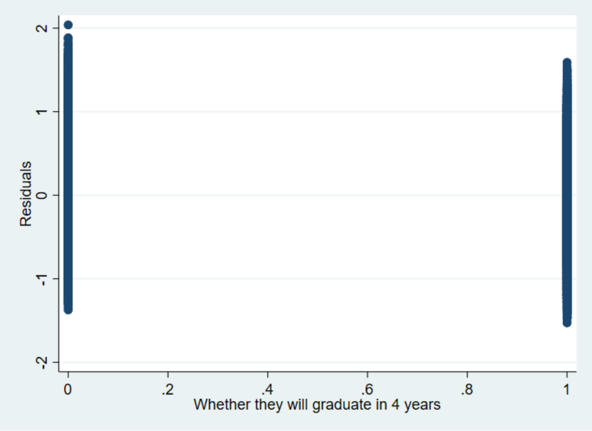


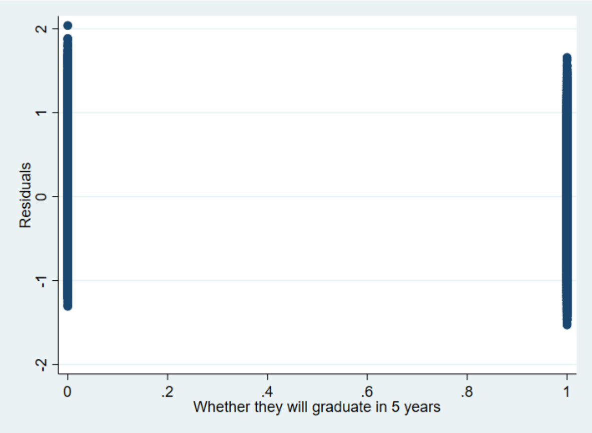


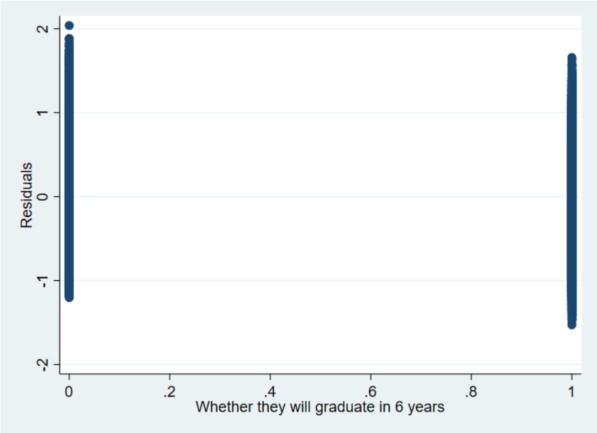


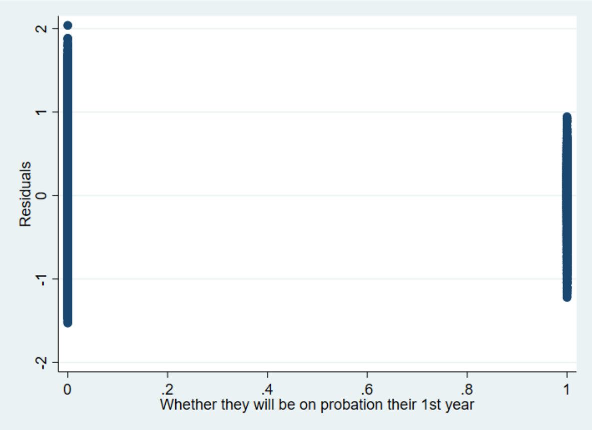


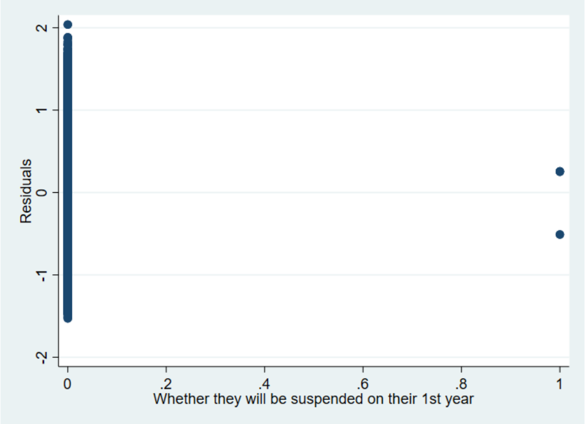

Residuals
2
1
0
-1
-2
0
.2
.4
.6
.8
1
Age 21 when first enrolled in University

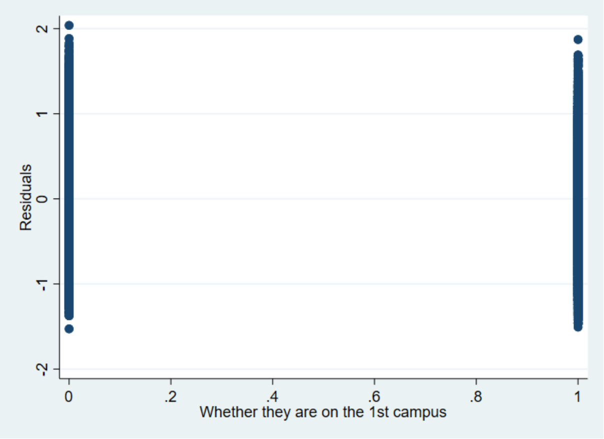

Residuals
2
1
0
-1
-2
0
.2
.4
.6
.8
1
Whether they will graduate in 4 years

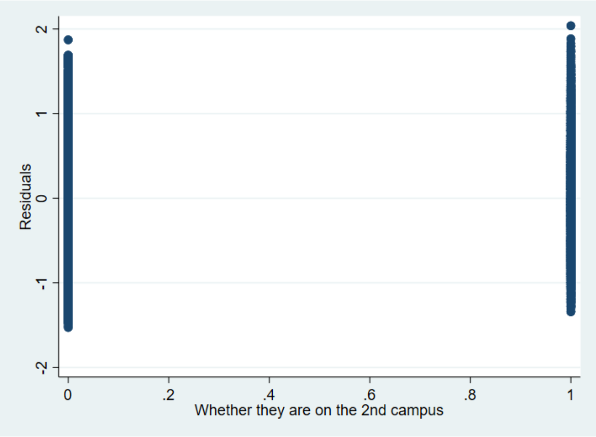

Residuals
2
1
0
-1
-2
0
.2
.4
.6
.8
1
Whether they will graduate in 5 years

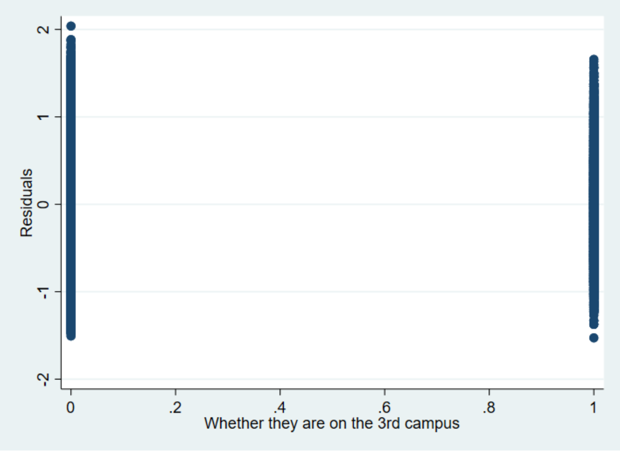

Residuals
2
1
0
-1
-2
0
.2
.4
.6
.8
1
Whether they will graduate in 6 years

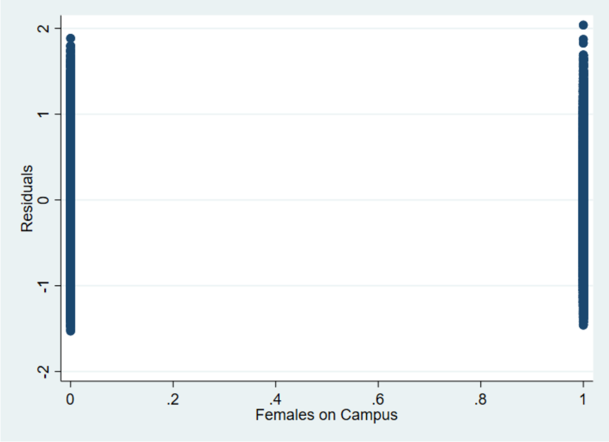

Residuals
2
1
0
-1
-2
0
.2
.4
.6
.8
1
Whether they will be on probation their 1st year

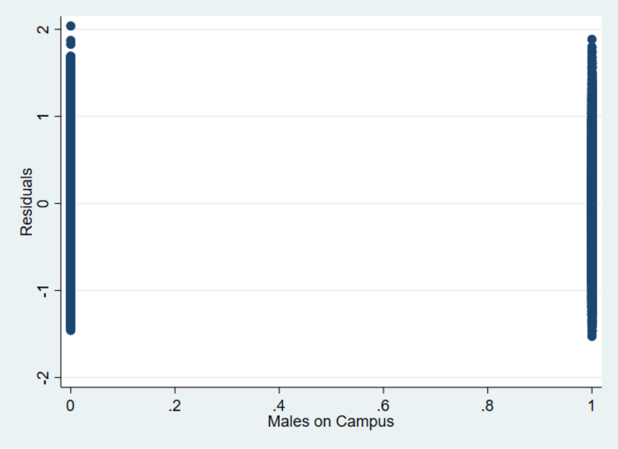

Residuals
2
1
0
-1
-2
0
.2
.4
.6
.8
1
Whether they will be suspended on their 1st year

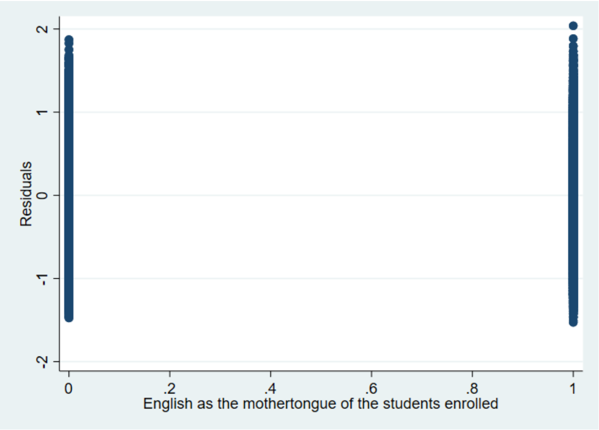

Residuals
2
1
0
-1
-2
0
.2
.4
.6
.8
1
Whether they are on the 1st campus

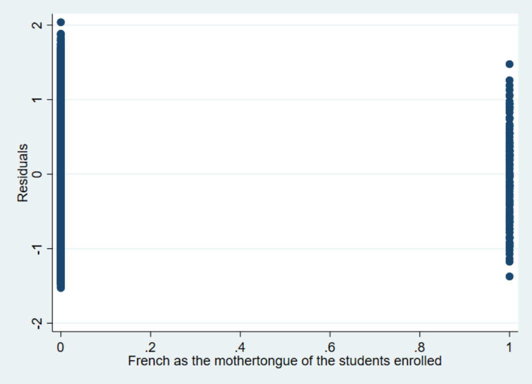

Residuals
2
1
0
-1
-2
0
.2
.4
.6
.8
1
Whether they are on the 2nd campus

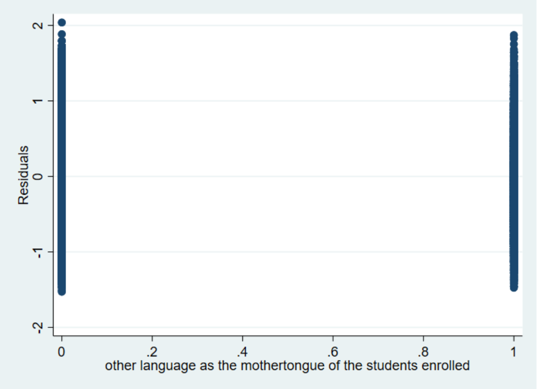

Residuals
2
1
0
-1
-2
0
.2
.4
.6
.8
1
Whether they are on the 3rd campus

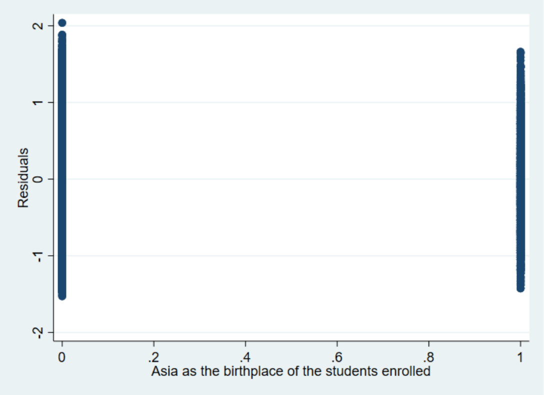

Residuals
2
1
0
-1
-2
0
.2
.4
.6
.8
1
Females on Campus

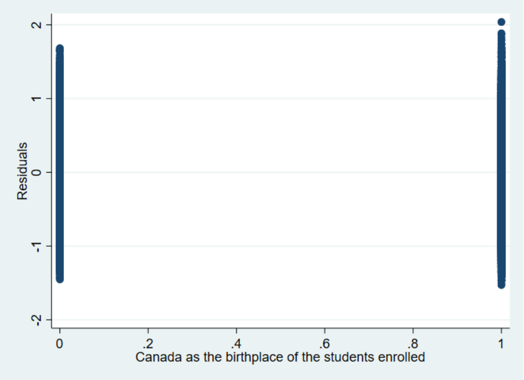

Residuals
2
1
0
-1
-2
0
.2
.4
.6
.8
1
Males on Campus

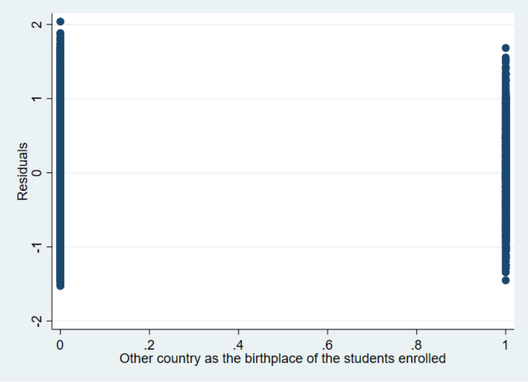

Residuals
2
1
0
-1
-2
0
.2
.4
.6
.8
1
English as the mothertongue of the students enrolled


Residuals
2
1
0
-1
-2
0
.2
.4
.6
.8
1
French as the mothertongue of the students enrolled


Residuals
2
1
0
-1
-2
0
.2
.4
.6
.8
1
other language as the mothertongue of the students enrolled


Residuals
2
1
0
-1
-2
0
.2
.4
.6
.8
1
Asia as the birthplace of the students enrolled


Residuals
2
1
0
-1
-2
0
.2
.4
.6
.8
1
Canada as the birthplace of the students enrolled


Residuals
2
1
0
-1
-2
0
.2
.4
.6
.8
1
Other country as the birthplace of the students enrolled

Code Appendix:

```
clear all

 use "C:\Users\aman2220\Desktop\canadiancollege.dta"

 . summarize

 histogram age_at_entry, ytitle(count) xtitle(Age of student entering college)
(bin=45, start=17, width=.08888889)


graph pie, over(sex) title(Male vs Female enrolled in college)

scatter GPA_year1 hsgrade_pct, ytitle(GPA: First Year of College) xtitle(High School
Grade Percentage)


histogram suspended_ever, addlabel ytitle(Count) xtitle(Amount of people being
suspended ever)
(bin=45, start=0, width=.02222222)


tab sex, gen(gender)
tab mtongue, gen(mothertongue)
tab birthplace, gen(birthplace1)
tab age_at_entry, gen(ageatentry)

rename gender1 female
rename gender2 male
rename mothertongue1 english
rename mothertongue2 french
rename mothertongue3 other_language
rename birthplace11 america
rename birthplace12 asia
rename birthplace13 canada
rename birthplace14 other_country
rename ageatentry1 age_at_17
rename ageatentry2 age_at_18
rename ageatentry3 age_at_19
rename ageatentry4 age_at_20
rename ageatentry5 age_at_21

drop birthplace mtongue sex

regress GPA_year1 totcredits_year1 age_at_17 age_at_18 age_at_19 age_at_20
age_at_21 gradin4 gradin5 gradin6 probation_year1 suspended_year1 campus1
campus2 campus3 female male english french other_language america asia canada
other_country hsgrade_pct

#taking out variables due to multicollinearity
regress GPA_year1 totcredits_year1 age_at_18 age_at_19 age_at_20 age_at_21
gradin4 gradin5 gradin6 probation_year1 suspended_year1 campus2 campus3 female
english french asia canada other_country hsgrade_pct

drop birthplace mtongue sex



predict yhat
predict uhat, resid
twoway scatter uhat yhat
twoway scatter uhat totcredits_year1, xtitle(Total credits taken in the 1st year)
twoway scatter uhat age_at_17, xtitle(Age 17 when first enrolled in University)
twoway scatter uhat age_at_18, xtitle(Age 18 when first enrolled in University)
twoway scatter uhat age_at_19, xtitle(Age 19 when first enrolled in University)
twoway scatter uhat age_at_20, xtitle(Age 20 when first enrolled in University)
twoway scatter uhat age_at_21, xtitle(Age 21 when first enrolled in University)
twoway scatter uhat gradin4, xtitle(Whether they will graduate in 4 years)
twoway scatter uhat gradin5, xtitle(Whether they will graduate in 5 years)
twoway scatter uhat gradin6, xtitle(Whether they will graduate in 6 years)
twoway scatter uhat probation_year1, xtitle(Whether they will be on probation their 1st
year)
twoway scatter uhat suspended_year1, xtitle(Whether they will be suspended on their
1st year)
twoway scatter uhat campus1, xtitle(Whether they are on the 1st campus)
twoway scatter uhat campus2, xtitle(Whether they are on the 2nd campus)
twoway scatter uhat campus3, xtitle(Whether they are on the 3rd campus)
twoway scatter uhat female, xtitle(Females on Campus)
twoway scatter uhat male, xtitle(Males on Campus)
twoway scatter uhat english, xtitle(English as the mothertongue of the students enrolled)
twoway scatter uhat french, xtitle(French as the mothertongue of the students enrolled)
twoway scatter uhat other_language, xtitle(other language as the mothertongue of the
students enrolled)
twoway scatter uhat asia, xtitle(Asia as the birthplace of the students enrolled)
twoway scatter uhat canada, xtitle(Canada as the birthplace of the students enrolled)
twoway scatter uhat other_country, xtitle(Other country as the birthplace of the students
enrolled)


regress GPA_year1 totcredits_year1 age_at_17 age_at_18 age_at_19 age_at_20
age_at_21 gradin4 gradin5 gradin6 probation_year1 suspended_year1 campus2
campus3 male female english french other_language america asia canada
other_country hsgrade_pct gpacutoff suspended_ever,robust

test french other_country asia age_at_21 canada english america gradin6 age_at_17
other_language

#final regression which I interpreted
regress GPA_year1 totcredits_year1 age_at_18 age_at_19 age_at_20 gradin4 gradin5
probation_year1 suspended_year1 campus2 campus3 female hsgrade_pct
suspended_ever,robust


lincom _cons + totcredits_year1*4.5 + age_at_20 + gradin5 + hsgrade_pct*35 + female

lincom _cons + totcredits_year1*4.5 + age_at_19 + gradin4 + gradin5 + hsgrade_pct*51
+ female

lincom _cons + totcredits_year1*4 + age_at_19 + hsgrade_pct*42 + female



#stepwise regression
stepwise,pr(.05): regress GPA_year1 totcredits_year1 age_at_18 age_at_19 age_at_20
age_at_21 gradin4 gradin5 gradin6 probation_year1 suspended_year1 campus2
campus3 female english french america asia canada hsgrade_pct
suspended_ever,robust
```

```
stepwise,pe(.05): regress GPA_year1 totcredits_year1 age_at_18 age_at_19
age_at_20 age_at_21 gradin4 gradin5 gradin6 probation_year1 suspended_year1
campus2 campus3 female english french america asia canada hsgrade_pct
suspended_ever,robust
```

Works Cited

“The Fires: How a Computer Formula, Big Ideas, and the Best of Intentions Burned Down New York City-and Determined the Future of Cities by Joe Flood.” *Goodreads*, Goodreads, 27 May 2010, www.goodreads.com/book/show/7906964-the-fires.

L., Theresa, and Theresa Lopez. “Analysis of the Academic Success Inventory for College Students: Construct Validity and Factor Scale Invariance.” *Fsu.digital.flvc.org*, 1 Jan. 1970, fsu.digital.flvc.org/islandora/object/fsu:175747.

Mansfield, Phylis M, et al. “College Students and Academic Performance: A Case of Taking Control.” *Journal of Student Affairs Research and Practice*, vol. 46, no. 3, 2009, doi:10.2202/1949-6605.5023.