EC 2228 Problem Set: Is College Worth It?

Instructor: Suraj Kumar

Due Date: TBA

Total: 100 points

Instructions

1. This homework has three sections: student debt (questions 1 to 3, 15 points), the college premium (questions 4 to 18, 75 points), and your memo to the President (question 19, 10 points). You will be awarded a maximum of 100 points for this assignment.

2. Create a do-file with your code, an automated log-file of your answers from that code, and fill out the provided Word template (`College_answer_template.docx` on Canvas) with your final answers as indicated.

3. Submit all three files (do-file, log-file, and Word document) as one zip file online via the Canvas homework tool. Failure to submit 1 out of 3 files will result in a 33.33% penalty on this homework grade. Failure to submit 2 out of 3 files will result in a 66.66% penalty on this homework grade.

4. You are encouraged to work with up to two partners (partners may be from other sessions, so in total a group may have a maximum of three students), but each student needs to submit solutions on Canvas. If you do work with partners, you should clearly state your partners' names in your Word file.

5. Name your files with your full name, beginning with your last name. For example, for John Smith the files would be named: `smith_john.do`, `smith_john.log`, and `smith_john.docx`, all inside `smith_john.zip`.

6. For any questions related to this assignment or any concerns related to grading, please contact Suraj (Raj) at `suraj.kumar@bc.edu`.

Welcome to the CEA



Figure 1: The Eisenhower Executive Office Building, next to the White House. Most of the White House staff work here, including the Council of Economic Advisers. Photo: Abovfold, Wikimedia Commons, CC BY-SA 4.0.

We hope you are having fun using Stata. Stata is indeed one of the quickest ways to get started in data management without learning to code. Yes, I do agree Stata has its weird quirks, but still it's a good way to start.

This week, we are going to analyze the relationship between education and wages. Yes, you read that correctly! Sorry once again! But instead of boring you with `bcuse wage1.dta`, Wooldridge's version of the wage equation from Jacob Mincer's "Schooling, Experience, and Earnings" (1974), we are going to work on "cool" real-world data from NBER. I hope you will find it interesting.

Imagine you have been hired at the White House Council of Economic Advisers (CEA) as a labor economist. You are excited about a new job! Finally, the years of training in economics and stressing with Stata seem to have paid off. Suddenly, the office of the President calls your office, asking for the CEA's input on drafting a new policy related to student debt relief and the college premium.

The CEA Chair finds that since you recently graduated from BC (*yes, you will, believe me, no matter how hard it seems now*), your inputs are highly valuable. She calls you to ask about the college premium. First, she wants to know the recent trends in college debt statistics, which have increased a lot over the years. She wants to know whether student debt has flatlined or is flattening, and how the COVID years affected it. Since the Congress committee and the office of the President are looking at making policy about debt relief, they want to be sure that supporting college education is indeed worth it.[1] To establish this claim, you want to estimate the college premium. Finally, you will put the two together in a short memo to the President.

[1] I hope you like your BC education. I always try to remind myself that BC and Boston are such fantastic places to be!

1 Is student debt still climbing?

> **From: CEA Chair**
>
> Welcome aboard! First things first: the President keeps hearing that student debt is out of control. Is it still climbing, or has it levelled off? And what did COVID do to it? Numbers, please.

You have received the *Federal Student Aid* data from the U.S. Department of Education, called `fsa_student_debt.dta`. This dataset covers federal student loans only, which are about 92% of all U.S. student debt.[2] Private loans are not included here. These are loans as of September 30 of each fiscal year, and they are in nominal terms, meaning not adjusted for inflation. Use `fsa_student_debt.dta` to answer the questions.

1. What was the average federal student debt per borrower in 2024? **Hint:** `debt_per_borrower` is total federal student debt divided by the number of borrowers. **[5 points]**

2. How has the debt evolved over the years? Plot a time series. **Hint:** The data are already `tsset` on `year`, so `tsline debt_per_borrower` plots the debt over time. **[5 points]**

3. What is the growth rate of debt? **Hint:** To derive the growth rate, do $g_{\text{debt}} = 100\left(\frac{\text{debt}_t}{\text{debt}_{t-1}} - 1\right)$, or simply $g_{\text{debt}} = 100 \log\left(\frac{\text{debt}_t}{\text{debt}_{t-1}}\right)$. In Stata, `L.debt_per_borrower` is last year's value, so `gen g = 100*(debt_per_borrower/L.debt_per_borrower - 1)` gives the growth rate for every year. Is the debt growing faster in the post-COVID period compared to the pre-COVID period? **Hint:** Compare `summarize g if year <= 2019` with `summarize g if year >= 2020`. **[5 points]**

2 Is college worth it?

> **From: CEA Chair**
>
> Next: Congress wants to know if a degree actually pays. You have the CPS. Find out! Report!

You, being a government economist, have access to all kinds of data! It's like living in a candy store for kids. Data is candy for economists. You have been given access to the Current Population Survey (CPS), a monthly household survey run by the U.S. Census Bureau for the Bureau of Labor Statistics. It is the source of the official U.S. unemployment rate. Each household is interviewed for four consecutive months, is out of the survey for the next eight months, and is then interviewed for the same four calendar months a year later. Each person in a household is numbered by "Line No". Usually one respondent answers the questions for the whole household. Earnings and hours questions are asked only in the fourth and eighth interviews, which are exactly one year apart. Each household (identified by a combination of `hhid` and `hrhhid2`) therefore appears once in a year's file, in one month, with

[2] Author's calculation: on September 30, 2024, federal student debt was $1,639 billion (Federal Student Aid, Portfolio Summary), and all U.S. student debt, federal and private, was $1,773 billion (Federal Reserve, G.19 Consumer Credit, FRED series SLOAS). $1{,}639/1{,}773 \approx 0.92$.

one row per household member (identified by `lineno`). It can reappear in the next year's file, in the same calendar month. For our purpose here, we will only consider the fourth interview and the eighth interview. The National Bureau of Economic Research (NBER) combines these surveys into one annual file called the Merged Outgoing Rotation Groups (MORG). It is the standard source of microdata on U.S. hourly wages.

Table 1: Candidate ID variables in the 2024 MORG file, `morg24_main.dta`

Variable	Label in Stata	Description
`hhid`	Household ID	First part of the household identifier, a 15-character string assigned by the Census Bureau.
`hrhhid2`	Household ID (partII)	Second part of the household identifier, a 5-character string.
`intmonth`	Interview Month	Calendar month of the interview, from 1 (January) to 12 (December).
`lineno`	Person's Line Number	The person's number on the household roster: 1 for the first person listed, 2 for the second, and so on.
`hhnum`	Household number	Counts the households that have lived at the sampled address. It is 1 for most rows and goes up when a new household moves in.

Notes: "Label in Stata" is the variable label shown by `describe`. The 2024 file has 244,144 rows, one per person aged 16 or over in a household in its fourth or eighth interview. Source: NBER CPS Merged Outgoing Rotation Groups, 2024, `https://data.nber.org/morg/annual/`.

Before you touch the data, write down your guess: how much more does a Bachelor's graduate earn per year than a high school graduate?

4. Let's begin with the hard work of an economist! Open the 2024 MORG data file, `morg24_main.dta`, in Stata. Let's first try to identify the ID of this dataset. By that I mean what each row in this dataset represents. To find that, we have to look at the combination of columns which helps to uniquely identify each row. There are several potential choices for you. Look at Table 1. **Hint:** Remember you can test the ID of the data in Stata by typing `isid varA varB`, which checks if var$A$ × var$B$ uniquely identifies the dataset. **[7 points]**

5. Is this dataset cross-sectional data or time-series data? How many observations does this dataset have? **Hint:** Remember to always run `describe` after loading any dataset. To understand whether a dataset is a cross-section or a time series, always ask what the unit of the cross-section is, and whether this unit is interviewed multiple times in the data at hand. **[4 points]**

6. How many individuals are in their fourth interview round, or in their eighth interview round? **Hint:** `minsamp` denotes the month in the sample. `tabulate varA` counts how many rows take each value. **[3 points]**

7. Now that you have gained some sense of the data, you want to consider the college premium. First, you want to just look at the people who are wage earners. So, you filter out the people who are too young (age 17 or lower) or too old (age 65 or higher). How many people are left afterwards? **Hint:** To filter observations in Stata, we follow

a syntax such as `keep if varA >= (some low number) & varA <= (some high number)`. [**4 points**]

8. You wish to analyze the wages from this dataset, but you don't have an hourly wage, only the weekly earnings, `earnwke` (ignore `earnhre` for the purpose of this assignment), and the hours worked in the week, `uhourse`. You wish to create an hourly wage, just like Wooldridge's `wage1.dta`. As always, data is messy, and real-world data is even messier. So first you plot the histogram of `uhourse`. Do you find something odd in the data? **Hint:** Try `codebook uhourse if uhourse <= 0`. [**4 points**]

9. You correct for this by dropping observations where `uhourse <= 0`. How many observations are dropped? You create two indicator variables to identify where (1) the weekly wage, `earnwke`, is present, as `weekly_wage_present`, and where (2) the weekly hours, `uhourse`, are present, as `weekly_hours_present`. **Hint:** Start the dummy at zero with `gen varA = 0`, then switch it on with `replace varA = 1 if varB != .` [**4 points**]

10. Tabulate both these newly created variables in Stata using `tabulate weekly_wage_present weekly_hours_present`. What do you observe? How many observations have both of these present? How many have the weekly wage but not the weekly hours? And how many have the weekly hours but not the weekly wage? [**3 points**]

11. Now you just keep the observations which have both the weekly wage and weekly hours. **Hint:** In Stata, `&` means "and": `keep if varA > 0 & varB > 0` keeps only rows where both hold. Now you generate the hourly wage as $wage = \frac{earnwke}{uhourse}$. Look at the newly created variable with `summ wage, detail` and also by plotting the histogram. What do you find unique about this wage distribution? Is it very lopsided on one end? What are the min, max, 5th percentile, median, and 95th percentile of this data? Look at the max: who earns that much an hour? Do you believe it? And what about the min? [**8 points**]

12. Now you want to look at education but, crap, you don't have it here. It's because the subagency that measures the wage in the U.S. Census Bureau doesn't talk to the agency that measures education. Typical government bureaucracy! 😕 (Okay, fine: your TAs split the file so you could practice `merge`. 😉) So you have to combine the education data, `morg24_educ.dta`, with the data that you have been working with so far. Let's call it the master data, and the education data the using data. How will you merge these two datasets together? **Hint:** Both files share the same key (ID), so the command looks like `merge 1:1 varA varB varC using morg24_educ.dta`. Did you lose any observations during merging? What are the `_merge == 2` rows, and why are they there? Then keep only the matched rows with `keep if _merge == 3`. [**8 points**]

13. In this newly combined data, `grade92` denotes the highest level of school a person has completed.[3] Table 2 shows all the codes. The data also has `ihigrdc`, NBER's

[3]The "92" in the name is the year the coding started. Before 1992, the CPS asked how many years of school a person had completed. From 1992, it asks for the highest degree instead.

estimate of years of schooling completed. What are the top three most populated categories of `grade92` in this dataset? [**3 points**]

Table 2: Codes of `grade92`, the highest level of school completed

Code	Highest level completed
31	Less than 1st grade
32	1st, 2nd, 3rd, or 4th grade
33	5th or 6th grade
34	7th or 8th grade
35	9th grade
36	10th grade
37	11th grade
38	12th grade, no diploma
39	High school graduate (diploma or GED)
40	Some college, no degree
41	Associate degree, occupational or vocational program
42	Associate degree, academic program
43	Bachelor's degree (BA, AB, BS)
44	Master's degree (MA, MS, MEng, MEd, MSW, MBA)
45	Professional school degree (MD, DDS, DVM, LLB, JD)
46	Doctorate degree (PhD, EdD)

Notes: Codes 31–38 record the last grade attended without a high school diploma; codes 39–46 record the highest diploma or degree received. A GED counts as a high school diploma (code 39). The codes are categories, not years of schooling. Source: U.S. Census Bureau, Current Population Survey, variable PEEDUCA; NBER CPS Merged Outgoing Rotation Groups.

14. Now you wish to analyze `ihigrdc` across all the different categories of `grade92`. You also want to analyze if `ihigrdc` is filled for all the rows for which `grade92` is reported. In order to do that, you create an indicator variable by `gen filled_educ_years = !missing(ihigrdc)`. **Hint:** `table grade92, statistic(frequency) statistic(count ihigrdc) statistic(mean filled_educ_years)` You will notice something odd. What is that? **Hint:** Look at `grade92` 43 and 44. [**4 points**]

15. Annoyed by seeing these missing values of `ihigrdc` for some grades, you decide to take control of the situation and decide to assign the schooling yourself. Create a new variable called `educ_nber` by `generate educ_nber = ihigrdc`, and then replace it with `educ_nber = 16` when `grade92 == 43 & missing(ihigrdc)`, and replace it with 18 when `grade92 == 44 & missing(ihigrdc)`. What is the overall correlation between `wage` and `educ_nber`? [**8 points**]

16. Now compare the mean wage for high school (**Hint:** `grade92 == 39`), Bachelor's college (**Hint:** `grade92 == 43`), and Bachelor's or more (**Hint:** `grade92 >= 43`). What is the mean hourly wage when the person has just high school, when the person has just college, and when the person has done college and more? [**4 points**]

17. Let's make a scatter plot of `wage` and `educ_nber` and also fit the best-fit line using Stata's `lfit`. See how the plot looks. **Hint:** `twoway (scatter A B) (lfit A B)` It may not look very sensible because of very high and low values in wages. So let's trim the

plot by only plotting wages which are higher than the 1st percentile but lower than the 99th percentile. **Hint:** Run `summ wage, detail`, store `p1` and `p99` with `scalar p1 = r(p1)` and `scalar p99 = r(p99)`, and then while plotting use `twoway (scatter A B if wage > p1 & wage < p99) (lfit A B if wage > p1 & wage < p99)`. **[3 points]**

18. Assuming a person works, say, 40 hours per week for 52 weeks in a year, what is the total annual salary of the high school graduate vs. Bachelor's only vs. BA and more? We define the college premium as the difference between the annual earnings of a Bachelor's graduate and a high school graduate:

$$\text{College premium} = \text{Annual earnings}_{\text{Bachelor's}} - \text{Annual earnings}_{\text{High school}}$$

What is the college premium in 2024? How close was your guess? **[8 points]**

> **From: CEA Chair**
>
> Great work! This is what the real work of an economist looks like. Most days at the CEA, we are either wrestling with messy data or trying to make sense of it with simple plots, correlations, or regressions. Today you did the first part: you cleaned the data, plotted it, and found the numbers. Next, we will need regressions to back them up. But first, the President has a follow-up question: can students actually pay off their debt with this premium?

3 Your memo to the President

> **From: CEA Chair**
>
> Great work. The President needs your memo by Friday. (Sound familiar? 😉) One paragraph, one number, one recommendation.

19. Now that you know the average level of debt in 2024 for a borrower (Section 1), and the college premium you estimated in Section 2, how do you feel about the overall situation? How many years of work will it require for the debt to be paid off now in 2024? Write a memo of 150 words or less to the President. Lead with one headline number. **Hint:** Divide the 2024 debt per borrower by your college premium. Then ask what this simple division leaves out: taxes, living costs, interest, and who the borrowers are (graduate students, and people who never finished a degree). **[10 points]**