

Term Project and Homework Assignments

Returns to Education

ECON 4400

1 Overview

Human capital accumulation increases productivity and employment opportunities. One approach to quantifying the returns to human capital is to estimate the effect of a year of schooling on an individual's wage. While economics has developed sound theoretical foundations, empirical work on the return to human capital has been at the center of considerable debate.

For our term project, we will explore a part of that debate by estimating the returns to education replicating (approximately, I have simplified the analysis to a degree) the results of Angrist and Krueger (1990). I chose this approach to foster critical thinking and deepen econometric knowledge. Our analysis will also draw upon the Bound, Jaeger, and Baker's (1995) critique of the instrumental variables approach used in Angrist and Krueger (1990).

Throughout the term, you will complete parts of the analysis and submit each component as a homework assignment. In doing so, I can assist with your learning of econometrics in practice. Additionally, the homework assignments enable me to address issues with coding or analysis.

For each assignment, you need only to submit what is requested. You will save the project components completed as homework and compile each one into single document that you will submit at the end of the term. The compiled document, described below, is your term paper for ECON 4400.

We will use the 2021 American Community Survey (ACS) to estimate the returns to education and the probability of participating in the labor force. Per the U.S. Census Bureau:

“The American Community Survey (ACS) is an ongoing survey that provides vital information on a yearly basis about our nation and its people.”

You will conduct your analysis of the returns to education and labor force participation using a sample of individuals residing in a U.S. state. Table 1 lists the state assigned to each student. To download your assigned datafile, log on Carmen, go to Modules, scroll toward the bottom of the page, and download the state datafile assigned to you.

2 Paper Requirements and Expectations

You will write, at most, a three-page analysis (not including tables and can be longer if needed) of the returns to education (and of labor force participation) and submit it at the beginning of class on Friday, 04/21. The paper will also include three tables: a table of summary statistics, labor force participation estimates, and returns to education estimates (see Sections 3.1, 3.2, 3.3, and 3.4). You need to attach your do-file with the

paper. If you do not submit a working do-file, you will receive, at most, half credit for this assessment.

Your do-file needs to be cleaned of any incorrect commands, i.e., commands that do not produce output, generate an error, and any redundant or unneeded commands. The entire do-file needs to be executable. In other words, if you click the *execute* icon, Stata executes every command without error.

Your write-up of the analysis should follow the below general outline—the sub-items do not need to follow the stated order. At a minimum, you must address each enumerated item. Your writing needs to flow (does not read as an itemized list). Each paragraph must consist of one key idea and includes supportive statements (evidence, results, etc.) of that key idea. Additionally, you need to ensure your writing includes transitions between key ideas (paragraphs).

1. Introduction

(a) Discuss the importance and benefit of education in the context of earnings.

(b) For background, read

- “Economic returns to education: What We Know, What We Don’t Know, and Where We Are Going—Some Brief Pointers” by Dickson and Harmon (2011)
- “Does Compulsory School Attendance Affect Schooling and Earnings” by Angrist and Krueger (1990)
- “Educational Attainment and Quarter of Birth: A Cautionary Tale of LATE” by Barua and Lang (2008)
- “Problems With Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak” by Bound, Jaeger, and Baker (1995)

You can access the papers on Carmen Modules, *Articles for Term Project*—bottom of the Modules page

2. Data and Methodology

(a) Discuss the data used for analysis

(b) Discuss the subsamples used for analysis, referencing the summary statistics

3. Labor Force Participation

(a) State the objective of using regression analysis to explain labor force participation

(b) Include the labor force participation model (see Section 3.4)

(c) Discuss the OLS and Logistic results

4. Returns to Education

(a) Introduce and discuss the wage equation (see Section 3.2)

(b) Discuss OLS return to education

(c) Discuss how robust the estimated return to education is to the inclusion of occupational dummy variables

(d) Discuss Why OLS estimate for the return to education is biased

(e) Discuss Two Stage Least Squares (2SLS) estimator—how does it address the endogeneity problem?

- (f) Discuss the instrumental variables (see Section 3.3), including the relevancy and validity requirements
- (g) Discuss the 2SLS return to education
- (h) Discuss the local average treatment effect (LATE)
- (i) Discuss how robust the estimated return to education is to the inclusion of occupational dummy variables
- (j) Compare and discuss OLS versus 2SLS estimates. Do the result meet expectations? Explain (Hint: why is OLS biased?) Discuss the F-statistic from the test for weak instruments. What insights does the test provide regarding the results?

5. Discussion and Conclusion

2.1 Paper Formatting

- Font: 11pt Times New Roman font
- Margins: One-inch margins (top, bottom, left, and right)
- Line spacing: 1.5 lines
- Start of new paragraph: Indent (no additional spacing between paragraphs)
- Text Alignment: justified
- Make sure to include your first and last name on the paper

References and Citations - Harvard Style If you choose to support an argument by drawing on the work of other scholars, you need to follow the below citation and reference style (Harvard). When you cite an article or research paper, you must include a reference section with your paper.

Citation and reference examples:

In-text citation

Author Year

Example - Parenthetical

(Tesseur 2022)

Reference list

Author(s) surname(s), Initial(s). (Year of publications). Title of article. *Title of Journal*, volume number(issue/number, or date/month of publication if volume and issue are absent), page numbers (if any).

Example - Narrative

Mayombe (2021)

Mayombe, C. (2021). Partnership with stakeholders as innovative model of work-integrated learning for unemployed youths. *Higher Education, Skills and Work-Based Learning* [online], 12(2), pp.309-327.

2.2 Stata Do-File

You will generate one do-file for this project. Each assignment will have you add to your code document (do-file). You must save your do-file at each step of the project (I recommend saving it regularly when working on an assignment). Separate each part using asterisks. For example:

```
**ECON 4400 Project: Name - Assigned State
*****
**Homework 1 - Summary Statistics
...code here...
*****
**Homework 2 - Returns to Education OLS
...code here...
*****
*****
**Homework 3 - Returns to Education 2SLS
...code here...
*****
*****
**Labor Force Participation
...code here...
*****
```

2.3 Data Assignments

Table 1: Data Assignments for Term Project (and Homework Assignments)

Name	State FIP	State
Alexander, Regan	48	Texas
Andebrhan, Yonaton	29	Missouri
Ataman, Clinton	47	Tennessee
Bay, Brenton	39	Ohio
Bays, Jake	26	Michigan
Belsito, Devon	34	New Jersey
Berman, Louis	17	Illinois
Bodnar, Andrew	13	Georgia
Borg, Nathan	12	Florida
Bugala, Kristen	27	Minnesota
Cai, Edward	17	Illinois
Cariddi, Ross	34	New Jersey
Chang, Shannon	18	Indiana
Dai, Yutao	37	North Carolina
DeMichele, Amanda	29	Missouri
Dong, Amy	12	Florida
Dunn, Griffin	24	Maryland
Eavers, Matt	13	Georgia

To be Continued

Name	State FIP	State
Eldredge, Donte	8	Colorado
Fehlman, Josh	27	Minnesota
Ferguson, Daniel	42	Pennsylvania
Gu, Lin	6	California
Guan, Xijiangtong	51	Virginia
Hagler, Grant	29	Missouri
Huang, Ancheng	26	Michigan
Iskandarova, Madina	13	Georgia
Kaupila, Jacob	42	Pennsylvania
Kiener, Claire	53	Washington
Lagana, Joey	34	New Jersey
Li, Jun	18	Indiana
Luo, Qianrui	45	South Carolina
Luo, Ruisi	55	Wisconsin
Mell, Tyler	45	South Carolina
Modeste, Armani	27	Minnesota
Mohamud, Abdulkadir Mohamed	48	Texas
Montano, Juan	37	North Carolina
Nickison, Connor William	39	Ohio
Niemeyer, Robin	8	Colorado
Odeh, Jonathan	45	South Carolina
Omar, Noor	53	Washington
Papuga, Trew	24	Maryland
Parrett, Kaylee	25	Massachusetts
Radous, Colin	12	Florida
Rodriguez, Noah	51	Virginia
Rosselot, Sam	42	Pennsylvania
Segerman, Andy	55	Wisconsin
Sentelle, Nick	17	Illinois
Sipes, Rachel	36	New York
Sorba, Logan	36	New York
Ting, Jonathan	24	Maryland
Turner, Chandler	25	Massachusetts
Upperman, Kyle	6	California
Vogelpohl, Nicholas	6	California
Wang, Johnny	47	Tennessee
Wang, Zixuan	26	Michigan
Warthman, Tristan	37	North Carolina
Weislogel, Peter	36	New York
Westfall, Nolan	8	Colorado
Yang, Dorothy	18	Indiana
Yoesting, Noah	39	Ohio
Zhou, Jiayi	25	Massachusetts

3 Homework: Putting Together Your Analysis

3.1 Homework 1, Due Wednesday, 03/01

Overview of assignment and what you will submit: Generate a table reporting summary statistics of various samples. Write one to two paragraphs summarizing the samples using the reported summary statistics. You will submit a paper copy of your write-up with the summary statistics table and a print-out of your do-file at the beginning of class on Wednesday, 03/01.

We will generate three subsamples for our analysis. The first sample consists of all individuals between the ages of 19 and 65 who are not on active duty. The second sample consists of individuals in the labor force who reported a wage or salary in 2020 (the 2021 ACS reports income from the prior year). The third sample includes only individuals between 29 and 40 years old who reported a wage or salary and participated in the labor force. We will use the latter sample to estimate the returns to education.

Your first homework assignment will require you to complete a process known as data cleaning. Researchers often need to recode or generate new variables from survey data. The below commands will walk you through how to “clean ACS data” to estimate the returns to education and the probability that an individual participates in the labor force.

The task of data cleaning is often an arduous one. To cultivate skills in command-based coding and data analytics using Stata, I provide code enabling us to use the ACS data for regression analysis. There is one exception (see below), where I ask you to generate a dummy variable indicating whether a person is employed. All other variable recoding or generation processes are provided in this section.

In Stata to indicate a range, e.g., tabulate *incwage* between 20,000 and 40,000, i.e., $20,000 \leq incwage \leq 40,000$, the code is `tab incwage if incwage >= 20000 & incwage <= 40000`. Suppose you want a “or” statement, use `|`. For example, you want a count of respondents who are married: `count if marst==1 | marst==2`, where a value of one indicates a married person and two married but separated (for assigned values and designations regarding marital status: `label list marst_lbl`). The vertical line `|` denotes “or” and `&` denotes “and” in Stata.

It is best practice to describe (label) newly generated variables. It will describe the variable enabling you to determine what it represents or measures when referring back to it. I am leaving variable labeling to you. It is not something you need to do, but it may be helpful later in the term.

```
label var variable_name "Description"
```

To begin, upload your assigned data into Stata:

```
use path/acs_2021_X.dta, clear
```

where *path* denotes the directory path where the datafile is saved on your computer. The “X” is a place holder for the State FIP code, e.g., if assigned California, the State FIP code is 6.

Define sample: To estimate the returns to education and labor force participation, we need to define the appropriate subsamples for analysis.

Keep all observations between the ages of 19 and 65.

```
keep if age>=19 & age<=65
```

Generating variables for analysis:

- Generate a dummy variable indicating whether a respondent reports participating in the labor force

```
gen lf=(labforce==2)
```
- Generate a dummy variable indicating whether a respondent reports being enrolled in school

```
gen attending=(school==2)
```
- Generate a set of dummy variables indicating which quarter of the year they were born, e.g., 1st, 2nd, 3rd, or 4th. The below command will produce four dummy variables labeled *qtr1*, *qtr2*, *qtr3*, and *qtr4*.

```
tab birthqtr, gen(qtr)
```
- Generate a variable *byear* indicating a respondent's birth year. The variable will be used to generate dummy variables for birth year, capturing variation in wages by birth cohort (see Homework 2).

```
gen byear=year-age
```
- Generate a variable for the square of age

```
gen age2=age^2
```
- Generate a dummy variable indicating if a respondent is married

```
gen married=(marst==1 | marst==2)
```
- Generate an interaction term between the variable *married* and the number of children under the age of five in the household (*nchlt5*)

```
gen marchlt5=married*nchlt5
```
- Generate a dummy variable if respondent identified as male

```
gen male=(sex==1)
```
- Generate dummy variables for race and ethnicity.
 - Generate a dummy variable if respondent identified race as White non-Hispanic

```
gen white=(race==1 & hispan==0)
```
 - Generate a dummy variable if respondent identified race as Black

```
gen black=(race==2)
```
 - Generate a dummy variable if respondent identified race as Asian or Pacific Islander

```
gen asian=(race>=4 & race<=6)
```
 - Generate a dummy variable if respondent identified as Hispanic

```
gen hispan2=(hispan>=1 & hispan<=4)
```
- Generate a dummy variable indicating whether a respondent works in a Metropolitan Statistical Area (MSA)

```
gen msa=(pwttype==1 | pwttype==2 | pwttype==3 | pwttype==4 | pwttype==5)
```
- You try: Generate a dummy variable indicating whether a respondent reports being employed. You will create a variable labeled *employed* using the ACS variable *empstat*. To do so, type `label list empstat_1b1` on the Results Window command line. Stata will display labels and corresponding values associated with each employment category. Using that information, you will generate a binary variable that takes on the value of one if employed and zero otherwise.

- Generate a new variable for years of schooling. When using the ACS, researchers need to recode education attainment to properly reflect a respondent's years of schooling. To see why, in the Stata command line, type `label list educd_lbl`. We will name the new educational attainment variable *grade*. The code for generating a variable reflecting years of schooling is

```
gen grade=.
replace grade=0 if educd==0 | educd==11 | educd==12
replace grade=1 if educd==14
replace grade=2 if educd==15
replace grade=3 if educd==16
replace grade=4 if educd==17
replace grade=5 if educd==22
replace grade=6 if educd==23
replace grade=7 if educd==25
replace grade=8 if educd==26
replace grade=9 if educd==30
replace grade=10 if educd==40
replace grade=11 if educd==50 | educd==61
replace grade=12 if educd==62 | educd==63
replace grade=12.5 if educd==65
replace grade=13 if educd==70
replace grade=13.5 if educd==71
replace grade=14 if educd==81
replace grade=16 if educd==101
replace grade=18 if educd==114
replace grade=19 if educd==115
replace grade=20 if educd==116
```

- Generate dummy variables for reported occupation using two-digit SOC classifications. To “clean” the ACS variable indicating occupation (*occ_soc*) requires advanced coding skills. I am providing the code below—copy it into your do-file to generate the occupational dummy variables. Make sure the code that you copied into your do-file has all the same characters. If not, you may need to edit the copied content in your do-file.

```
gen occupation=occ_soc
replace occupation=subinstr(occupation,"X","0",.)
replace occupation=subinstr(occupation,"Y","0",.)
destring occupation, replace
replace occupation=floor(occupation/10000)
keep if occupation<55 //Keeping observations not on active duty
```

Save data with the above changes: We now have our first subsample of the 2021 ACS, which we will refer to as the “Main Sample.” To save, follow the command below.

```
save path/acs_2021_state_main1965.dta, replace
```

where *path* denotes the directory path (folder location) and *state* denotes your assigned state, e.g., Michigan. The option `replace` allows you to overwrite an existing file. Suppose you made changes to a previously saved datafile. The option `replace` allows you to overwrite the old file with the new changes.

Summary Statistics for the “Main Sample: Ages 19-65”

You will replicate Table 2 and report each stated variable's mean and standard deviation. **Important: The variable labels in the table below may differ from the variable labels in your `acs_2021_state_main1965` datafile. For example, in the datafile, the variable *grade* reports years of schooling, but we will label it "Education" in the table.**

In Stata use the `sum` command to report the mean and standard deviation of the variables listed in Table 2 for the Main Sample. You will report the standard deviation in parentheses below the mean (see the below examples).

To generate summary statistics, call the "Main Sample" datafile

```
use path/acs_2021_state_main1965.dta, clear
```

After uploading the datafile, use the `sum` command as instructed above. For example, `sum grade`.

Summary Statistics for the "Employed Sample: Ages 19-65"

We will now generate our second subsample, consisting of individuals who report participating in the labor force (employed or unemployed) and a wage or salary. Before we can obtain the mean and standard deviation for the variables listed in the sample, we need to clean the data further. First call in the main subsample:

```
use path/acs_2021_state_main1965.dta, clear
```

Next, keep observations that meet the following criteria.

- Dropping observations that not report an income
`drop if incwage==999999 | incwage==999998 | incwage==0 | incwage==.`
- Dropping observations not in the labor force
`drop if empstat==0 | empstat==3`
- Dropping observation that report a top or bottom code for typical hours worked per week (type: label list `uhrswork_lbl` for top and bottom codes)
`drop if uhrswork==99 | uhrswork==0`
- Drop observations that report bottom code for weeks worked in a year
`drop if wkswork1==0`
- Drop outlier observations for typical hours worked in a week
`drop if uhrswork<10`
- Drop observations that report attending school
`drop if school==2`

Generating an imputed hourly wage rate:

```
gen hwage=(incwage/wkswork1)/uhrswork
```

Save data with the above changes: We now have our second subsample of the 2021 ACS, which we will refer to as the "Employed Sample: Ages 19-65." To save, follow the command below.

```
save path/acs_2021_state_employed1965.dta, replace
```

To generate summary statistics, call the "Employed Sample: Ages 19-65" datafile

```
use path/acs_2021_state_employed1965.dta, clear
```

After uploading the datafile, use the `sum` command as instructed above. For example, `sum grade`. You will input each variable's mean and standard deviation under the column "Employed: Ages 19-65" in Table 2.

Summary Statistics for the "Employed Sample: Ages 29-40"

We will now generate our third subsample, consisting of individuals 29 to 40 years of age who report participating in the labor force (employed or unemployed) and a wage or salary. Before we can obtain the mean and standard deviation for the variables listed in the sample, we need to clean the data further. First call in the "Employed: Ages 19-65" sample: `use path/acs_2021_state_employed1965.dta, clear`

Next, keep observations that meet the following criteria:

```
keep if age>=29 & age<=40 & school!=2
```

Save data with the above changes: We now have our third subsample of the 2021 ACS, which we will refer to as the "Employed Sample: Ages 29-40." To save, follow the command below:

```
save path/acs_2021_state_employed2940.dta, replace
```

We will use the subsample of 19 to 40-year-old employed workers (not in school) to estimate the returns to education. The sample consists of individuals who likely completed intended educational pursuits and, more recently, finished schooling relative to older workers.

To generate summary statistics, call the "Employed Sample: Ages 29-40" datafile

```
use path/acs_2021_state_employed2940.dta, clear
```

After uploading the datafile, use the `sum` command as instructed above. For example, `sum grade`. You will input the variable means and standard deviations under column "Employed: Ages 29-40" in Table 2.

Table 2: Summary Statistics for State X

	Main Sample (Ages 19-65)	Employed (Ages 19-65)	Employed (Ages 29-40)
Education	13.6304 (2.3922)		
Age		33.1425 (2.5643)	
Male			0.5145 (0.0345)
White			
Black			
Asian			
Hispanic			
Married			
Children			
Work in MSA			
Employment			
Hourly Wage	Leave Blank Leave Blank		

Standard deviation in parentheses

3.2 Homework 2, Due Friday, 03/24

Will update after Homework 1 due date

3.3 Homework 3, Due Wednesday, 04/12

Will update after Homework 2 due date

3.4 Final Analysis: Labor Force Participation-Not Included As a Homework Assignment But Is A Part of The Term Project

Overview of assignment and what you will submit: Model and quantify the probability that an individual participates in the labor force. Use a linear probability model (LPM) and Logit estimator (see chapter 13). Replicate the results table below with your estimates. Include the table in your paper and discuss. You will not submit this project component as part of any homework assignment. However, it will be included in your final write-up (see Section 2).

We want to explain changes in the probability that an individual participates in the labor force. Using your `acs_2021_state_main1965` datafile, estimate the below labor force participation model using OLS

(linear probability model) and Logit estimators. You will generate a separate results table (see Table 3 below).

To call in the datafile:

```
use path/acs_2021_state_main1965.dta, clear
```

The labor force participation equation is specified below.

$$lfi = \beta_0 + \beta_1 \text{grade}_i + \beta_2 \text{age}_i + \beta_3 \text{age2}_i + \beta_4 \text{white}_i + \beta_5 \text{married}_i + \beta_6 \text{male}_i + \beta_7 \text{nchlt5}_i + \epsilon_i$$

When estimating the labor force participation equation, we want to exclude individuals attending school. This requires including a conditional statement with the `reg` and `logit` commands. You will include the “if” statement after the last explanatory variable. Additionally, we want to use Heteroskedasticity and autocorrelation consistent (HAC) standard errors (`vce(robust)`). An example with an “if” statement and specifying HAC standard errors:

```
reg y x if attending==0, vce(robust)
```

Estimate a second specification that includes the interaction term between married and the number of children under the age of five living in the household (*marchlt5*).

Repeat the above two specifications using OLS and Logit for males and females. Example code for females only: `logit y x if attending==0 & male==0, vce(robust)`.

You will replicate the below table, report the results (standard errors in parentheses), and indicate significance levels using asterisks (*, **, *** indicates significance at the 10%, 5%, and 1% levels, respectively). If needed, horizontally orientate the table. Note that the even number columns report the OLS and Logit estimates ($\hat{\beta}$) of the labor force participation model with the interaction term. The odd columns report the OLS and Logit estimates of the labor force participation model with the interaction term excluded from the specification (see the above equation). The *N* row reports the sample size or the number of observations used with each set of estimates.

Table 3: Estimates for Labor Force Participation (Values shown as Example)

	Full Sample				Males				Females			
	(OLS)	(OLS)	(Logit)	(Logit)	(OLS)	(OLS)	(Logit)	(Logit)	(OLS)	(OLS)	(Logit)	(Logit)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Education	0.4567***		1.1657***									
	(0.0023)		(0.0245)									
Age	0.1234**											
	(0.0919)											
Age Squared	-0.0234*											
	(0.0124)											
White												
Married												
Male												
# of Children<5	-0.6545***	-0.1456**										
	(0.0211)	(0.02567)*										
Married×Children<5		-0.6345***										
		(0.0012)										
<i>N</i>												