Term Project and Homework Assignments
Returns to Education

ECON 4400

1 Overview

Human Capital, defined as the skills, knowledge, and abilities that an individual possesses, has been a focal point of economic research in labor, development, and political economy—to name a few. Education and job training are important human capital investments, leading to higher earnings and non-pecuniary benefits. Your ECON 4400 project focuses on the former. You will quantify the returns to human capital, estimating 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.

As part of your project, you will explore a part of that debate by replicating (approximately, I have simplified the analysis to a degree) the results of Angrist and Krueger (1990) using the 2021 American Community Survey (ACS). I chose this approach to foster critical thinking and deepen econometric knowledge. Our analysis will also draw upon Bound, Jaeger, and Baker's (1995) critique of the instrumental variables approach used in Angrist and Kreuger (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 include the tables created for each assignment with your term project. A homework assignment will also ask you to introduce, discuss, and explain particular sections of your term project, e.g., data, regression analysis, results, and econometric methodologies. After I return the assignment, you should edit and expand the section following the outline below, addressing any notes or needed corrections.

You will analyze the returns to education and labor force participation at the state level. Refer to Table 1 to see your assigned state. To download your data file, log on to Carmen, go to Modules, scroll toward the bottom of the page, and download the state data file assigned to you.

2 Paper Requirements and Expectations

You will write a three to four 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 Tuesday, 12/05. The paper will 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 redundant or incorrect 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. For background, read the following papers:

   - "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) Cite and discuss the data used for the analysis

   (b) Discuss the sub-samples used for the 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 3.2)

   (c) Discuss the OLS and Logistic results

4. Returns to Education

   (a) Introduce and discuss the wage equation (see Section 3.3)

   (b) Discuss OLS return to education

   (c) Discuss why the OLS estimate for the return to education is biased

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

   (e) Discuss the instrumental variables (see Section 3.4), including the relevancy and validity requirements

   (f) Discuss the 2SLS return to education

(g) 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 - Chicago 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 (Chicago). When you cite an article or research paper, you must include a reference section with your paper.

Citation and reference examples:

In-text citation	**Reference list**
Author Year	First author's last name, first author's first name, second author's first and last names, third author's first and last name, . . . , and last author's first and last name. 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 - Parenthetical	
(Tesseur 2022)	Tesseur, W. 2022. "Translation as inclusion? An analysis of international NGOs' translation policy documents." *Language Problems and Language Planning*, 45(3): 261–283.
Example - Narrative	
Piketty and Saez (2003)	Piketty, Thomas, and Emmanuel Saez. 2003. "Income Inequality in the United States, 1913–1998." *The Quarterly Journal of Economics*, 118(1): 1–41.

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 - Labor Force Participation
...code here...
********************
********************
**Homework 3 - OLS Returns to Education
...code here...
********************
********************
**Homework 4 - 2SLS Returns to Education
...code here...
********************
```

2.3 Data Assignments

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

Name	State FIP	State
Azimi, Hassina	20	Kansas
Backlin, Will	6	California
Bian, Shuo	16	Idaho
Bodnar, Andrew	9	Connecticut
Brady, Alex	12	Florida
Carroll, Jack	4	Arizona
Cavanagh, Devin	15	Hawaii
Cervantes Lopez, Diego	39	Ohio
Chu, Shirley	55	Wisconsin
Cowan, Ben	25	Massachusetts
Dallin, Abdul	12	Florida
DeMarco, Jill	25	Massachusetts
Deng, Peter	31	Nebraska
Doumet, Daniel	27	Minnesota
Dudek, Alli	19	Iowa
Dunlap, Joe	51	Virginia
Earls, Claire	24	Maryland
Ellis, Sydney	27	Minnesota
Fang, Junjie	24	Maryland
Gardner, Luke	44	Rhode Island

To be Continued

Name	State FIP	State
Greenwalt, Jack	26	Michigan
Guo, Geyang	51	Virginia
Guzzo, Patrick	36	New York
Halkatti, Anirudh	4	Arizona
Healy, Caroline	44	Rhode Island
Hernandez, Angelina	42	Pennsylvania
Huang, Longji	9	Connecticut
Knapp, Jenna	17	Illinois
Koontz, Kyle	13	Georgia
Kramer, David	5	Arkansas
Kratt, Peter	45	South Carolina
Lang, Lionel	31	Nebraska
Liu, Chunlin	45	South Carolina
Maxamod, Caisha	39	Ohio
McNary, Ross	9	Connecticut
Meng, Kevin	19	Iowa
Miranda, Jack	36	New York
Morris, Dejanique	25	Massachusetts
Nakanishi, Sanenao	15	Hawaii
Palmquist, Zoey	5	Arkansas
Parker, Hannah	4	Arizona
Patel, Shaalin	45	South Carolina
Phillips, Natalie	42	Pennsylvania
Radous, Colin	32	Nevada
Reott, Hayden	38	North Dakota
Robinson, Rachael	24	Maryland
Rockwell, Aneesha	17	Illinois
Rupp, Ian	6	California
Sheese, Micah	35	New Mexico
Sides, Rachel	41	Oregon
Spurlock, Noah	42	Pennsylvania
Sundaram, Suriya	25	Massachusetts
Tang, Zhuangyu	44	Rhode Island
Taylor, Kendall	39	Ohio
Terrell, Jamari	19	Iowa
Thongkhot, Fiat	16	Idaho
Thotli, Varun	12	Florida
Wadsworth, Emma	51	Virginia
Wang, Jinze	6	California
Wang, Lin	55	Wisconsin
Wang, Mingyao	35	New Mexico
Wang, Shupeng	24	Maryland
Wang, Xinyu	38	North Dakota

To be Continued

Name	State FIP	State
Wentzel, John	13	Georgia
Werntz, Cole	41	Oregon
Yi, Eleanor	32	Nevada
Zhang, Chaoran	26	Michigan
Zhang, Peter	20	Kansas

3 Homework: Putting Together Your Analysis

3.1 Homework 1, Due Tuesday, 09/26

Overview of assignment and what you will submit: Generate a table reporting summary statistics of various samples. Write one to two paragraphs summarizing, characterizing, and noting the similarities or dissimilarities between the samples. 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 Tuesday, 09/26.

What to Submit - three items:

1. A write-up discussing the data source, the samples, and summary statistics
2. A table of summary statistics
3. Attach a printout of your do-file (the entire document)

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., $20{,}000 \leq \textit{incwage} \leq 40{,}000$, 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 indicates 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 (Note: If you copy the Stata code from this PDF, some characters may not correctly reproduce onto the do-file. If you receive an error message after executing your do-file, check whether the source is due to an incorrectly copied character.):

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

where *path* denotes the directory path where the data file 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 Asain 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 self-identifies other than White
    ```
    gen drace=(race!=1)
    ```

- Generate a dummy variable indicating whether a respondent works in a Metropolitan Statistical Area (MSA)
  ```
  gen msa=(pwtype==1 | pwtype==2 | pwtype==3 | pwtype==4 | pwtype==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_lbl` 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*, denoting highest completed schooling. The code for generating a variable reflecting years of schooling is
  ```
  gen grade=.
  replace grade=0 if educd>=0 & 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==60 | educd==62 | educd==63 | educd==64
  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>=80 & educd<=83
  replace grade=15.5 if educd==100
  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 (*occsoc*) 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=occsoc
  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,” but save it as sample1. To save, follow the command below.

```
save path/acs_2021_state_sample1.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 data file. The option `replace` allows you to overwrite the old file with the new changes.

Summary Statistics for the Sample 1 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*_sample1 data file. For example, in the data file, 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” data file

```
use path/acs_2021_state_sample1.dta, clear
```

After uploading the data file, 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 sample1: `use path/acs_2021_state_sample1.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 "Worked: Ages 19-65" (the term *worked* denotes respondents reporting an income in 2020) To save, follow the command below.
```
save path/acs_2021_state_worked1965.dta, replace
```

To generate summary statistics, call the "Employed Sample: Ages 19-65" data file
```
use path/acs_2021_state_worked1965.dta, clear
```
After uploading the data file, 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 "Worked: 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 an income in 2020. 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 "Worked: Ages 19-65" sample: `use path/acs_2021_state_worked1965.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 "Worked: Ages 29-40." To save, follow the command below:
```
save path/acs_2021_state_worked2940.dta, replace
```

We will use the subsample of 19 to 40-year-old respondents who are not in school and reported a 2020 income 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" data file
```
use path/acs_2021_state_worked2940.dta, clear
```

After uploading the data file, use the `sum` command as instructed above. For example, `sum grade`. Replicate Table 2 below and input the mean and standard deviation for each variable in each sample. Make sure to state standard deviations in parentheses. Please make note of the blank cells under the "Main Sample" column. We do not have an hourly wage variable for that sample. It includes respondents who did not report an income. The sample average wage would not accurately characterize the average worker income. Thus, we do not want to report it for the "Main Sample."

Table 2: Summary Statistics for Alabama

	Main Sample* (Ages 19-65)	Employed† (Ages 19-65)	Employed‡ (Ages 29-40)
Years of Schooling	13.3340	13.8052	13.9727
	(2.9637)	(2.7869)	(2.8296)
Age	43.5709	43.4261	34.6538
	(13.9893)	(12.6347)	(3.5066)
Male	0.4879	0.5260	0.5236
	(0.4999)	(0.4993)	(0.4995)
White	0.6748	0.7030	0.6973
	(0.4685)	(0.4570)	(0.4595)
Black	0.2222	0.1983	0.1921
	(0.4158)	(0.3988)	(0.3940)
Asian	0.0176	0.0173	0.0154
	(0.1317)	(0.1302)	(0.1233)
Hispanic	0.0397	0.0376	0.0456
	(0.1952)	(0.1902)	(0.2086)
Married	0.5261	0.5968	0.5902
	(0.4993)	(0.4906)	(0.4918)
Children[1]	0.6909	0.8088	1.2658
	(1.0541)	(1.0789)	(1.2243)
Works in MSA	0.3425	0.5091	0.5154
	(0.4746)	(0.4999)	(0.4998)
Employed	0.6558	0.9760	0.9743
	(0.4751)	(0.1530)	(0.1583)
Hourly Wage		29.4113	28.4412
		(48.3768)	(55.5631)
N	28,388	16,341	4,278

Standard deviation in parentheses

[1] Number of children living in the household

* Includes individuals between the age of 19 and 65

† includes individuals between the age of 19 and 65 and reported an income in 2020

‡ includes individuals between the age of 19 and 40 and reported and income in 2020

3.1.1 Example of Describing and Characterizing Samples

The below excerpt is from a published study on returns to source country human capital that I along with a coauthor wrote. I am including the excerpt for reference. You must submit your own work.

> "We use data from the New Immigrant Survey (NIS), 2003. The NIS provides extensive information on new immigrants to the U.S., who received their permanent residency (green card) in 2002-2003. 8573 adult immigrants completed interviews. If the spouse of a principal immigrant was also an immigrant, then the spouse was also interviewed. In all, 4915 spouses were interviewed. Therefore, the total number of foreign-born immigrants interviewed who received their green card between 2002 and 2003 was 13488. Our study design requires us to focus on immigrants who worked in their source countries before immigrating to the U.S. and worked in the U.S. after immigrating to the U.S. Out of 13488 immigrants, 6977 (59%) held at least one job in their source countries before moving to the U.S. We focus on only those immigrants who were between the ages of 16 and 65 (both included) at the time of immigrating to the U.S. This reduced the sample size to 5919 immigrants. We further restrict the sample based on the availability of covariates, those who report participating in the U.S. labor market after immigrating, and those who have an accurate work history in their source countries. The sample size after accounting for the availability of covariates and labor market participation is 2713. However, approximately nine percent of that sample have stated work histories where the start of their first source country job is after the completion of their last source country job or there is an overlap between the start and end dates of their first and last source country jobs. We omit these observations from our analysis, which leaves a sample of immigrants who participated in the U.S. labor market of 2460.
> ...
> Table 2 presents the summary statistics for all immigrants (column 1), stayers (columns 2-4), and switchers (columns 5-7). There are three samples of stayers (and switchers) because we define stayers (and switchers) based on the 1, 2, and 3-digit occupational coding. Among all workers 40.2% are female. The percentage of female workers varies between 37.6% and 39.5% among stayers and 40.5% and 42% among the switcher sample. The average number of years of education among all workers is 13.82 years. Stayers have slightly more education than switchers. Average education varies between 14.9 and 14.38 years among the stayer samples, and they vary between 13.44 and 13.51 years for the switcher samples. Stayers are more proficient in English than switchers are. In our data, between 53.3% and 61.3% of stayers have "very good" English speaking skills. On the other hand, between 32.7% and 35.2% of switchers have an English proficiency categorized as "very good." We use interviewer reported measure for fluency in spoken English, which is available for all interviewees. The measure takes a value of one through four with one for very good English to four for poor English. The English proficiency measure is given a value of five if the interview was conducted in a language other than English because the interviewees were not comfortable in answering questions in English. "

3.2 Homework 2, Due Tuesday, 10/10

Overview of assignment and what you will submit: To estimate the pecuniary return to education, our analysis uses only individuals who report an income. It is, therefore, of interest to understand the factors influencing the likelihood of participating in the labor force. For homework assignment 2, you will model and quantify an individual's probability of participating in the labor force using the ordinary least squares and logit estimators (see Chapter 13). Estimate each specification stated below for all individuals and females and males separately. Report your findings by replicating Table 3, and include the table in your term project. You must include a write-up of the results and discuss how the interaction between married and the number of children under the age of five in the household affected the estimates for each group. What is the economic significance of the results? Additionally, discuss whether the effect of years of completed school on labor force participation is robust to the inclusion of the interaction term. What inference, if any, do you draw from that result?

What to Submit - three items:

1. A write-up of the model and results
2. A results tables that includes your OLS and Logit estimates
3. Attach a printout of your do-file (the entire document)

Objective: We want to explain changes in the probability that an individual participates in the labor force. Use the **`acs_2021_`*`state`*`_sample1.dta`** data file, estimate the below labor force participation models using OLS (linear probability model) and logit estimators for all individuals, females only, and males only between the ages 19 and 65.

The linear probability model for each specification is:

$$lf_i = \beta_0 + \beta_1 grade_i + \beta_2 age_i + \beta_3 drace_i + \beta_4 married_i + \beta_5 male_i + \beta_6 nchlt5_i + \varepsilon_i$$

$$lf_i = \beta_0 + \beta_1 grade_i + \beta_2 age_i + \beta_3 drace_i + \beta_4 married_i + \beta_5 male_i + \beta_6 nchlt5_i + \beta_7 marchlt5_i + \varepsilon_i$$

and the standard documentation format for the logit equation for each specification is:

$$L:Pr(lf_i = 1) = \beta_0 + \beta_1 grade_i + \beta_2 age_i + +\beta_3 drace_i + \beta_4 married_i + \beta_5 male_i + \beta_6 nchlt5_i + \varepsilon_i$$

$$\begin{aligned} L:Pr(lf_i = 1) &= \beta_0 + \beta_1 grade_i + \beta_2 age_i + +\beta_3 drace_i + \beta_4 married_i + \beta_5 male_i + \beta_6 nchlt5_i \\ &+ \beta_7 marchlt5_i + \varepsilon_i \end{aligned}$$

The Stata commands to estimate β_k using OLS and Logit are:
OLS: `reg` $y\ x_1\ x_2\ x_3 \ldots x_K$
Logit: `logit` $y\ x_1\ x_2\ x_3 \ldots x_K$

When estimating the labor force participation equation, we want to exclude individuals attending school, which requires including a qualifier with the `reg` and `logit` commands. To do so, you will include an "if" statement after the last explanatory variable, limiting the analysis to only those observations not in school. 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)`

Estimating the regression models separately for females and males with the school attendance qualifier requires you to include the `&` between the two qualifying statements. For example, `reg y x if attending==0 & male==0, vce(robust).`

Note:

1. The option command `vce(robust)` informs Stata to estimate heteroskedasticity-corrected (HC) standard errors. We will learn more about the consequences and remedies for heteroskedasticity later in the term. In short, the presence of heteroskedasticity results in biased OLS standard errors. While HAC standard errors typically avoid the consequences of heteroskedasticity, the HAC standard error estimator is still biased, but generally more accurate than OLS standard errors.
2. When estimating the binary dependent model separately for female and male observations using OLS and logit, we must omit the explanatory variable `male` from the specification. Can you explain why?

Reporting Results

Replicate the below table (12 sets of estimates), report the results (standard errors in parentheses), sample size (*N*), and indicate significance levels using asterisks placed in the superscript of the estimate (*, **, *** indicates significance at the 10%, 5%, and 1% levels, respectively). Note that the odd-number columns exclude the married and number of children under the age of five in the household interaction term, while the even-numbered columns report the OLS and logit estimates with the interaction term included in the specification.

Table 3: Labor Force Participation: Alabama

	All				FeMales				Males			
	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	0.0333***	0.0335***	0.1736***	0.1747***	0.0391***	0.0394***	0.1934***	0.1953***	0.0266***	0.0268***	0.1569***	0.1570***
	(0.0009)	(0.0009)	(0.0058)	(0.0059)	(0.0013)	(0.0013)	(0.0084)	(0.0085)	(0.0012)	(0.0012)	(0.0083)	(0.0083)
Age	-0.0085***	-0.0085***	-0.0437***	-0.0439***	-0.0088***	-0.0088***	-0.0430***	-0.0431***	-0.0090***	-0.0091***	-0.0512***	-0.0514***
	(0.0002)	(0.0002)	(0.0013)	(0.0013)	(0.0003)	(0.0003)	(0.0018)	(0.0018)	(0.0003)	(0.0003)	(0.0019)	(0.0019)
Race†	-0.0206***	-0.0218***	-0.1101***	-0.1155***	0.0204**	0.0186**	0.0961**	0.0878**	-0.0705***	-0.0706***	-0.4228***	-0.4226***
	(0.0062)	(0.0062)	(0.0318)	(0.0318)	(0.0088)	(0.0088)	(0.0432)	(0.0433)	(0.0084)	(0.0084)	(0.0475)	(0.0476)
Male	0.1050***	0.1071***	0.5551***	0.5649***								
	(0.0054)	(0.0054)	(0.0290)	(0.0291)								
Married	0.1128***	0.1214***	0.5679***	0.6074***	0.0152*	0.0246***	0.0621	0.1068**	0.2174***	0.2266***	1.2345***	1.2490***
	(0.0060)	(0.0061)	(0.0306)	(0.0316)	(0.0085)	(0.0088)	(0.0412)	(0.0427)	(0.0081)	(0.0083)	(0.0478)	(0.0487)
Children Less Than 5 y.o.	-0.0442***	0.0217	-0.2252***	0.0617	-0.1076***	-0.0547***	-0.5420***	-0.3002***	0.0132**	0.1284***	0.5989***	0.8447***
	(0.0063)	(0.0134)	(0.0384)	(0.0732)	(0.0096)	(0.0165)	(0.0462)	(0.0785)	(0.0062)	(0.0179)	(0.1052)	(0.1985)
Married × Children		-0.0879***		-0.4077***		-0.0773***		-0.3536***		-0.1378***		-0.3655
		(0.0149)		(0.0828)		(0.0196)		(0.0911)		(0.0187)		(0.2300)
Constant	0.5305***	0.5235***	0.0602	0.0287	0.5207***	0.5120***	0.0246	-0.0232	0.6978***	0.6938***	0.9209***	0.9214***
	(0.0171)	(0.0172)	(0.1005)	(0.1010)	(0.0254)	(0.0255)	(0.1437)	(0.1448)	(0.0215)	(0.0215)	(0.1391)	(0.1389)
N	25,777	25,777	25,777	25,777	13,089	13,089	13,089	13,089	12,688	12,688	12,688	12,688

Robust standard errors are reported in parentheses

* 0.10, ** 0.05, and *** 0.01 denote significance levels

† Race indicates whether a respondent self-identified other than White

3.3 Homework 3, Due Tuesday, 10/31

Overview of assignment and what you will submit: State the below wage equation, define the variables, and discuss the model. Estimate the wage equation using OLS and discuss the estimated return to an additional year of schooling for all workers and separately for female and male workers.

Using the **acs_2021_*state*_employed2940.dta**, estimate the wage equation using OLS and report the *grade* coefficient (β_1) for all workers. Then separately estimate the return to an additional year of schooling for female and male workers. You only need to report the coefficient estimate for *grade* ($\hat{\beta}_1$), standard error

$SE(\hat{\beta}_1)$, significance level (denoted by asterisks on the coefficient estimate, see below), R^2, and sample size (N). Replicate Table 4 and report the OLS estimates and statistics for columns (1), (3), and (5). We will estimate the returns to education using Two-stage Least Squares (2SLS) for the last homework assignment.

What to Submit - three items:

1. A write-up of the model and results
2. Replicate Table 4 below and report only OLS estimates and statistics (you will report the 2SLS estimates as part of homework assignment 4, columns 2, 4, and 6)
3. Attach a printout of your do-file (the entire document)

Generating Variables Needed For The Wage Equation

Before we proceed with estimating the return to an additional year of schooling, we need to generate a few more variables.

1. The natural log of the hourly wage rate for person *i*

   ```
   gen lnw=ln(hwage)
   ```

2. Dummy variables that take on a value of one if person *i* was born in year *t*, zero otherwise. Note, the sample acs_2021_*state*_worked2940.dta includes workers 29 to 40 years of age. Therefore, we will generate 12 dummy variables for birth year.

   ```
   tab byear, gen(birthyear)
   ```

   The variable *byear* was generated in the first homework assignment. It reports a respondents birth year. The above command tabulates *byear* and generates a dummy variable for each birth year in the sample, e.g., *birthyear1, birthyear2, ... , birthyear11*.

The wage equation:

$$lnw_i = \beta_0 + \beta_1\, grade_i + \beta_2\, age_i + \beta_3\, male_i + \beta_4\, drace_i + \beta_5\, msa_i + \sum_{b=1}^{B} \beta_b\, birthyear_{bi} + \varepsilon_i \qquad (1)$$

I am providing the code for estimating equation (1) using acs_2021_*state*_worked940.dta for all workers. You will need to amend the code to separately estimate the wage equation for female and male workers (refer to assignment three for how to use a qualifier to estimate a subsample, e.g, female workers only). When estimating the wage equation for female and male workers, you must drop the explanatory variable *male* from the wage equation.

```
reg lnw grade age male drace msa birthyear1-birthyear11, vce(robust)
```

Notes

1. Do not report a F statistic with your OLS results

2. The dummy variable for *birthyear12* is omitted from the wage equation. Therefore, it is the reference category. Note: If $birthyear_b$ is perfectly collinear with another, Stata will drop it from the regression analysis.

Table 4: Returns to Education: Alabama

	All		**Females**		**Males**	
	OLS	2SLS	OLS	2SLS	OLS	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)
Education	0.0888***	0.0118	0.0997***	0.0865**	0.0797***	0.0529
	(0.0048)	(0.0450)	(0.0077)	(0.0388)	(0.0059)	(0.0359)
F[†]		1.08		1.49		1.44
R^2	0.16	0.09	0.15	0.15	0.14	0.13
N	4,278	4,278	2,038	2,038	2,240	2,240

Robust standard errors are reported in parentheses

* 0.10, ** 0.05, and *** 0.01 denote significance levels

Additional controls: age, sex, race, msa, birth year

[†] Test for weak instruments

Example of Model Introduction and Transition to Results Discussion

The below excerpt come from a research article of mine studying the returns to immigrant source country human capital. I am including the excerpt for reference. You must submit your own work.

> "We start by estimating the following wage regression (non-linear terms not shown in the equation to reduce clutter).
>
> $$w_i = \alpha + \beta GE_i + X_i\psi + \varepsilon_i \tag{2}$$
>
> The dependent variable, w_i, is the (log of) hourly wage rate of immigrant i in the first U.S. job. The vector of individual characteristics is denoted by X_i and includes the immigrant's years of education in source country, gender, proficiency in English, visa type, number of years between last source country job and first U.S. job, region of residence in the U.S., source country fixed effects, and U.S. entry year fixed effects. Since previous literature on transferability of source country work experience has only included returns to GE, we start with a specification where we include only GE. We use a flexible functional form with up to the fourth order work experience terms as suggested by Murphy (1990). They show that standard (quadratic) Mincer earnings function may understate the early career earnings growth and may overstate mid-career earnings growth. In our regressions, we divide squared, cubed, and quartic terms by 1000 to re-scale the coefficients.
>
> Table X presents the results based on the first U.S. wage regressions. We estimate this equation for all workers (column 1) and then separately for stayers (columns 2-4) and switchers (columns 5-7). The classification of who is a stayer (or switcher) depends on the granularity of occupational definitions. When we define occupation at the 1-digit level, we have 1006 stayers and 1454 switchers. At the 2-digit level, we have 694 stayers and 1766 switchers. At the 3-digit level, we have 551 stayers and 1909 switchers. "

3.4 Homework 4, Due Tuesday, 11/21

Overview of assignment and what you will submit: As we saw in chapter 6, OLS does not consistently estimate $\widehat{\beta}_k$ when $COV(x_k, \varepsilon) \neq 0$. The method of instrumental variables, specifically two-staged least squares (2SLS), provides a general solution to the endogeneity problem (chapter 14). Following Angrist and Kreuger (1990) (closely but not exactly) , we define a set of exogenous variables

$$\begin{aligned} \mathbf{z} &= (1, \text{age}, \text{male}, \text{drace}, \text{msa}, \text{birthyear1}, \text{birthyear2}, \ldots, \text{birthyear11}, \\ & \quad \text{qtr1} \times \text{birthyear1}, \text{qtr1} \times \text{birthyear2}, \ldots., \text{qtr3} \times \text{birthyear11}) \end{aligned}$$

to instrument for `grade`. We can write the reduced form equation for `grade` as

$$\begin{aligned} grade_i &= \alpha_0 + \alpha_1 age_i + \alpha_3 male_i + \alpha_4 drace_i + \alpha_5 msa_i + \sum_b \delta_b \, birthyear_{ib} \\ &+ \sum_b \sum_j \gamma_{bj} \, birthyear_{ib} \times qrt_{ij} + u_i \end{aligned}$$

where $grade_i$ is the education of the ith individual, age_i, $drace_i$, and msa_i are covariates (independent variables) that influence the hourly wage of the ith individuals, and $birthyear_{ib}$ is a binary variable indicating whether an individual as born in year b ($b = 1, 2, 3, \ldots, 11$). The interaction terms between birth year and quarter of birth ($qtr \times birthyear$) are the excluded instrumental variables. The reference categories for each excluded instrument variable is if an individual was born in the fourth quarter of a year and the year of birth for the youngest worker ($birthyear12$).

Using **acs_2021_*state*_employed2940.dta**, estimate the wage equation using 2SLS and report the estimated coefficient of *grade* for all workers. Then separately, estimate the return to an additional year of schooling for female and male workers. You only need to report the coefficient estimate for *grade* ($\widehat{\beta}_1$), standard error $SE(\widehat{\beta}_1)$, significance level (denoted by asterisks on the coefficient estimate, see below), R^2 (the reported Centered R2), sample size (N), and the test for weak instruments' F statistic. Report the estimates and statistics in columns (2), (4), and (6) in Table 4.

What to Submit - three items (with subitems)

1. A write-up of the 2SLS estimator, which includes:
   (a) How 2SLS is a remedy for endogeneity bias
   (b) Discussion of the excluded instrumental variables ($z_1, z_2, \ldots, z_M$), including whether the validity and relevancy conditions are satisfied. You will need to read Angrist and Kreuger (1990) for information regarding the authors' argument for why quarter of birth and interacting quarter of birth with birth year satisfy the validity ($COV(z_m, u) = 0$) and relevancy (the excluded instrumental variables $z_1, z_2, \ldots, z_M$ are strongly correlation with `grade`) conditions. Is the argument reasonable? (see Bound, Jaeger, and Baker, 1995)
   (c) Report and discuss 2SLS estimates of the return to education (for all, female, and male workers). How do 2SLS compare to OLS estimates? Did the estimates change as expected? Explain.
   (d) Conduct a test for Weak Instruments and discuss the implications of your conclusion to the test. You will derive an F-statistic for each estimate
2. Report the 2SLS estimates of the returns to education and statistics along with your OLS estimates in Table 4 that you estimated for the third homework assignment
3. Attach a printout of your do-file (the entire document)

Code For Generating Birth Quarter and Birther Year Interactions

- Before you estimate equation (1) using 2SLS, generate interaction terms for each pairwise combination of `qtr` and `birthyear`. Note: We generated the variable birth year (`byear`) and dummy variables for quarter of birth (`qtr`) and birth year (`birthyear`) in previous assignments. **Note: Be mindful of the characters below. When calling a local macro (deleted when program or do-file ends) use the grey accent (back tick) (`) and vertical apostrophe (').**

```
qui tab byear
local num=r(r)
forvalues k=1(1)`num'{
   forvalues y=1(1)4   {
      gen yearqtr`k'_`y'=birthyear`k'*qtr`y'
   }
}
```

Two-Stage Least Squares Using the Stata Command *ivreg2*

I am providing the code for estimating equation (1) using the two-stage least squares estimator. You will need to amend the code to separately estimate the wage equation for female and male workers. When estimating the wage equation for female and male workers, you must drop the explanatory variable *male* from the equation. Note, you will need to install the *ivreg2* package and a complementary package called ranktest to execute the two-stage least squares estimator in Stata. To install the packages, go to the results window, click in the command line and type *ssc install ivreg2* and hit the Enter/return key. Then, type *ssc install ranktest* in the command window, and hit the Enter/return key).

```
ivreg2 lnw (grade=yearqtr1_1 yearqtr1_2 /*
*/ yearqtr1_3 yearqtr2_1 yearqtr2_2 yearqtr2_3 yearqtr3_1 yearqtr3_2 /*
*/ yearqtr3_3 yearqtr4_1 yearqtr4_2 yearqtr4_3 yearqtr5_1 yearqtr5_2 /*
*/ yearqtr5_3 yearqtr6_1 yearqtr6_2 yearqtr6_3 yearqtr7_1 yearqtr7_2 /*
*/ yearqtr7_3 yearqtr8_1 yearqtr8_2 yearqtr8_3 yearqtr9_1 yearqtr9_2 /*
*/ yearqtr9_3 yearqtr10_1 yearqtr10_2 yearqtr10_3 yearqtr11_1 /*
*/ yearqtr11_2 yearqtr11_3) age male drace msa /*
*/ birthyear1-birthyear11, r
```

Note: The `r` after the comma denotes robust standard errors. We need to use the option `r` instead of `vce(robust)` with the `ivreg2` command.

Test For Weak Instruments

When conducting a test for weak instruments, the associated test statistic is an F-statistic. That is the F you will report along with your estimates in Table 4. On your do-file, you must hard code in the test, as shown in class. I will deduct points if you report a Wald F statistic using a Stata command other than what was presented in class. Recall that an F-test requires estimating an unrestricted and restricted model. Use your class notes and the posted Stata demonstration for two-stage least squares for reference.

4 Write-up Your Project

Take your results from Section 3 and write an analysis on the return to education and labor force participation following the outline in Section 2. Remember to attach your working do-file. You must submit your term project on **Tuesday, 12/05** at the beginning of the class period.

What to Submit

1. A three-page paper on the returns to education and labor force participation for state *X*, where *X* is a placeholder for the assigned state. You will also include three tables reporting summary statistics, returns to education, and coefficient estimates of the factors influencing labor force participation. The paper length is independent of the inclusion of the tables.

2. Attach your executable Stata do-file. Makes sure to run the entire do-file before printing to ensure that Stata can execute all commands without any error.