



## IPUMS Data Training Exercise: An introduction to IPUMS HigherEd (Exercise 1 for Stata)



### Learning goals

- Understand how IPUMS HigherEd dataset is structured
- Create and download an IPUMS data extract
- Decompress the data file and read the data into a statistical package
- Analyze the doctorate recipients' occupation and salary using sample code

### Summary

In this exercise, you will gain basic familiarity with the IPUMS HigherEd data exploration and extract system to answer the following research questions: "How many doctorate recipients are working in an occupation related to his/her highest degree? What factors are most important in determining a doctorate recipient's salary?" You will create a data extract that includes the variables GENDER, SALARY, AGE, EMSEC, NDGMED, CTZUSIN, and OCEDRLP; then you will use the sample code to analyze these data. After completing this exercise, you will have experience navigating the IPUMS HigherEd website and should be able to leverage these data to explore your own research interests.

## Stata Code to Review

Code	Purpose
<u>generate</u>	Creates a new variable, "replace" specifies a value according to cases
mean	Displays a simple tabulation and frequency of one variable
<u>tabulate</u>	Displays a cross-tabulation for up to 2 variables
<u>regress</u>	OLS regression

## Common Mistakes to Avoid

- Not fully decompressing the data
- Giving the wrong filepath to indicate the dataset
- Mixing up = and == ; To assign a value in generating a variable, use "=". Use "==" to specify a case when a variable is a desired value using an *if* statement.
- Forgetting to put [weight=*weightvar*] into square brackets

## Register for an IPUMS Account

Go to <http://highered.ipums.org>, click on User Registration and Login and Apply for access. On the login screen, enter email address and password and submit your application. Registration is automatic.

## Make a Data Extract

- Navigate to the IPUMS HigherEd homepage and click on "Browse Data."

### Select Samples

- Click on the "Select Samples" button to choose the census samples to include in your extract
- Check the Select Samples box and go to the Full SDR tab
- Check the very first check box labeled SDR – this will select all years of full SDR samples. Click on Submit sample selections

### Select Variables

- The variable drop-down menus allow you to explore variables by topic. For example, you might expect to find variables about a person's degree(s) under the "Education" group.



- The search tool allows you to search for variables. Observe the options for limiting your search results by variable characteristics or variable type.
- You may add a variable to your cart by clicking on the plus sign in the "Add to Cart" column of the topical variable list, or list of search results.
- You may view more information about the variable by clicking on the variable name, and navigating through the tabs that include a description of the variable, codes and value labels, the universe of persons asked the question, and information on the comparability of the variable among other pieces of information.
- Use either the drop down menu or the search feature to browse these variables, and add them to your data cart.
  - GENDER: Gender
  - AGE: Age
  - MINRTY: Minority indicator
  - CTZUSIN: U.S. citizenship status, binary
  - LFSTAT: Labor force Status
  - EMSEC: Employer sector
  - SUPWK: Supervisory Work/ supervised other employees as a part of his/her principal job.
  - SALARY Annual salary
  - NDGMED: Field of major for highest degree
  - NDGMEMG: Field of major for highest degree (7 groups)
  - OCEDRLP: Degree to which respondent's job related to highest degree
  - WTSURVY: Full SDR weight variable

## Review and submit your extract

- Click on the "View Cart" button underneath your data cart.
- Review your variable and sample selection to ensure your extract is complete.
  - You may notice a number of additional variables you did not select are in your cart; IPUMS preselects a number of key technical variables, which are automatically included in your data extract.
- Add additional variables or samples if they are missing from your extract, or click the "Create Data Extract" button.
- Review the Extract Request screen that summarizes your extract; add a description of your extract (e.g., "Doctorate Recipients: Occupation and Salary") and click "Submit Extract".
- You will receive an email when your data extract is available to download.



# Getting the data into your statistics software

The following instructions are for Stata. If you would like instructions for a different stats package, see <https://ipums.org/support/exercises>.

## Download the data

- Follow the link in the email notifying you that your extract is ready, or by clicking on the "Download and Revise Extracts" link on the left-hand side of the IPUMS HigherEd homepage.
- Right-click on the data link next to the extract you created.
- Choose "Save Target As..." or "Save Link As..."
- Save into your preferred working directory. This tutorial assumes you will save the file into "Documents" (which should pop up as the default location).
- Do the same thing to save the Stata command file.

## Decompress the data

- All IPUMS extracts are compressed. We recommend [7zip](#) for Windows users. Macs can open these types of files without additional software.
- Find the "Documents" folder under the Start menu.
- Double click on the ".dat" file.
- In the window that pops up, press the "Extract" button.
- After the extract has completed, confirm that the Documents folder contains three files that begin with "highered\_000##".

## Read in the data

- Open the "highered\_000##.do" file.
- In the syntax editing window, change the first line from "libname IPUMS '." to "libname IPUMS '\\Documents...;" using the file directory where you saved your data files.
- After "filename ASCIIIDAT", enter the full file location, ending with highered\_###.dat;
- Choose "Submit" under the Run file menu.



# Analyze the Data

## Part 1: Frequencies

For each variable below, search through the tabbed sections of the variable description to answer each question.

1. On the website find the codes page for the OCEDRLP variable:
  - a. Write down each code value, and what category each code represents.

\_\_\_\_\_

- b. What is the universe for OCEDRLP?

\_\_\_\_\_

2. How many doctorate recipients were employed in an occupation closely related to their field in 2013 in the SDR survey?

\_\_\_\_\_

**\*Note: the SDR only includes doctorate recipients.**

3. What proportion of doctorate recipients in the SDR survey were employed in an occupation closely related to their field in 2013?

\_\_\_\_\_

```
tabulate year ocedrlp if ocedrlp != 98, row
```

## Using Weights (WTSURVY)

In order to find a nationally representative estimate of doctorate recipients, we need to use a frequency weight. The above analysis finds frequencies specific to the survey samples, but a weight adjusts the analysis to be representative of the US target population for each year.

4. Using frequency weights, how many doctorate recipients nationally had an occupation closely related to their field of degree in 2013?

\_\_\_\_\_

5. Using frequency weights, what proportion of doctorate recipients nationally had an occupation closely related to their field of degree in 2013? \_\_\_\_\_

```
replace wtsurvey = round(wtsurvey)
```

```
tab year ocedrlp if ocedrlp != 98 [fw=wtsurvey], row
```



6. Using weight frequencies, how many doctorate recipients were working (employed) in the United States in 2013?
- 

```
tab year lfstat [fw=wtsurvey], row
```

## Part 2: Relationships in the Data

7. Which doctorate fields were dominated by women in 2013 (women comprised more than 50 percent)?
- 

```
tab ndgmed gender if year == 2013 [fw=wt], row
```

8. Using weight frequencies, what is the difference in the mean salary between employed female and male doctorate recipients in 2013?
- 

```
mean salary if year == 2013 & salary <=150000 [fw=wt],  
over(gender)
```

9. Using weight frequencies, what are the average salaries for doctorate recipients by employer sector in 2013?
- 

```
mean salary if year == 2013 & salary <=150000 [fw=wt],  
over(emsec)
```

10. Use OLS regression to predict reported salaries of doctorate recipients for 2013. Fill in the table below with the coefficients and t-statistics.

*Note: Because SDR respondents are interviewed every 2 to 3 years, regressing over multiple years will lead to over-counting individuals and standard errors that are too small.*

*The tab statements generate indicator variables for each value of the employee sector and field of degree.*



```
tab emsec, gen(sector)
```

```
tab ndgmemg, gen(field)
```

```
reg salarp gender age ctzusin minrty supwk sector2 sector3  
sector4 field1-field6 if salary <=150000 & lfstat == 1 & year ==  
2013 [fw=wt]
```

<u>Variable</u>	<u>Coefficient</u>	<u>t-statistic</u>
Gender		
Age		
US citizen		
Minority		
Supervisory Work		
4-year college		
Government		
Business and Industry		
Computer and Math Sciences		
Biological Sciences		
Physical Sciences		
Social Sciences		
Engineering		
S&E related fields		
Constant		



# Answers

## Part 1: Frequencies

1. On the website find the codes page for the OCEDRLP variable:
  - a. Write down each code value, and what category each code represents.  
1 Closely related; 2 Somewhat related; 3 Not related; 98 Logical Skip
  - b. What is the universe for OCEDRLP?  
Working during the week of reference period.
2. How many doctorate recipients were employed in an occupation closely related to their field in 2013 in the SDR survey? 17,696
3. What proportion of doctorate recipients in the SDR survey were employed in an occupation closely related to their field in 2013? 66.33%

## Using Weights (WTSURVY)

4. How many doctorate recipients nationally had an occupation closely related to their field of degree in 2013? 474,761
5. What proportion of doctorate recipients nationally had an occupation closely related to their field of degree in 2013? 65.88%
6. How many doctorate recipients were working (employed) in the United States in 2013? 720,626

## Part 2: Relationships in the Data

7. Which doctorate fields were dominated by women in 2013 (women comprised more than 50 percent)? Psychology, Sociology/Anthropology, Health, Non-science
8. What is the difference in the mean salary between employed female and male doctorate recipients in 2013? Women: \$ 83,511.42; Men: \$ 101,263.9;  
Difference: \$ 17,752.48



9. What are the average salaries for doctorate recipients by employer sector in 2013?

<u>Sector</u>	<u>Mean Salary (\$)</u>
2 Year College	62,839.32
4 Year College	86,621.32
Government	102,334.50
Business/industry	105,961.80

10. Use OLS regression to predict reported salaries of doctorate recipients for 2013.

<u>Variable</u>	<u>Coefficient</u>	<u>t-statistic</u>
Gender	5059	103.3
Age	331	83.69
US citizen	6995	52.50
Minority	-4806	-31.37
Supervisory Work	23177	269.95
4-year college	17110	73.57
Government	34321	130.47
Business and Industry	34617	148.64
Computer and Math Sciences	24866	11.31
Biological Sciences	12003	5.47
Physical Sciences	15632	7.12
Social Sciences	8151	3.71
Engineering	25489	11.61
S&E related fields	17322	7.87
Constant	5172	2.28

