Minnesota Population Center

Training and Development

IPUMS – HigherEd Extraction and Analysis

Exercise 1 - Stata

OBJECTIVE: Gain an understanding of how an IPUMS dataset is structured and how it can be leveraged to explore your research interests. This exercise will use the IPUMS dataset to explore the factors that affect doctorate recipient's salaries and the relatedness between doctorate recipients' field of degree, employer sector, and gender.

6/2/2015

IPUMS – Higher Ed Training and Development

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?

Objectives

- Explore a sample of variables from IPUMS-HIGHER ED
- Analyze the data using example code

IPUMS Variables

- GENDER: Respondent's gender
- SALARP: Annual salary
- AGEP: Age
- EMSECPB: Employer sector
- NDGMEDP: Field of degree category
- CTZUSIN: US citizenship
- OCEDRLP: Degree to which respondent's job related to highest degree

Stata Code to Review

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

Review Answer Key (page 6)

Common Mistakes to Avoid

1 Not changing the working directory to the folder where your data is stored

2 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.

3 Forgetting to put [weight=*weightvar*] into square brackets

Registering with IPUMS

Go to http://highered.ipums.org, click on "Register to Use IPUMS-HIGHER ED" and apply for access. On the login screen, enter email address and password and submit it!

Step 1

Make an Extract

- Go back to the homepage and go to Select Data
- Click 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
- Using the drop down menu or search feature, select the following variables:

  GENDER: Gender

  AGEP: Age

  MINRTY: Minority background indicator

  SALARP: Annual salary

  LFSTAT: Employment status

  EMSECPB: Employer sector

  HRSWKP: Hours typically worked per week

  CTZUSIN: US citizenship indicator

  OCEDRLP: Degree to which respondent's job related to highest degree

  NDGMEDP: Field of major for highest degree

  NDGMEMG: Field of major for highest degree (6 groups)

  WTSURVY: Full SDR weight variable

  SUPWK: Work includes supervisory role

…

Step 2

Request the Data

- Click the green VIEW CART button under your data cart
- Review variable selection. Click the green Create Data Extract button
- Review the 'Extract Request Summary' screen, describe your extract and click Submit Extract
- You will get an email when the data is available to download
- To get to the page to download the data, follow the link in the email, or follow the Download and Revise Extracts link on the homepage

Getting the data into your statistics software

The following instructions are for Stata. If you would like to use a different stats package, see: http://highered.ipums.org/highered/extract_instructions.shtml

Step 1

Download the Data

- Go to http://highered.ipums.org and click on Download or Revise Extracts
- Right-click on the data link next to extract you created
- Choose "Save Target As..." (or "Save Link As...")
- Save into "Documents" (that should pop up as the default location)
- Do the same thing for the Stata link next to the extract

Step 2

Decompress the Data

- Find the "Documents" folder under the Start menu
- Right click on the ".dat" file
- Use your decompression software to extract here
- Double-check that the Documents folder contains three files starting "highered_000…"
- Free decompression software is available at http://www.irnis.net/soft/wingzip/

Step 3

Read in the Data

- Open Stata from the Start menu
- In "File" menu, choose "Change working directory…"

  Select "Documents", click "OK"
- In "File" menu, choose "Do…"

  Select the *.do file
- You will see "end of do-file" when Stata has finished reading in the data

Analyze the Sample – Part I Frequencies

Step 1

Analyze the Data

A) On the website, find the codes page for the OCEDRLP variable and write down each code value, and what category each code represents. ____

B) What is the universe for OCEDRLP?

____

C) How many doctorate recipients were employed in an occupation closely related to their field in 2013 in the SDR survey? __________

D) 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
```

• • •

Step 2

Weighting the Data

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.

A) How many doctorate recipients nationally had an occupation closely related to their field of degree in 2013?

____

```
replace wtsurvy = round(wtsurvy)
tabulate year ocedrlp if ocedrlp != 98 [fw=wtsurvy], row
```

B) What proportion of doctorate recipients nationally had an occupation closely related to their field of degree in 2013?

____

C) How many doctorate recipients were working in the United States in 2013? ____

Analyze the Sample – Part II Relationships in the Data

Section 1

Crosstabs and Means

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

____

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

B) What is the difference in the mean salary between employed female and male doctorate recipients? ____________________

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

C) What are the average salaries for doctorate recipients by employer sector in 2013? ______________________________

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

Regression

D) Use OLS regression to predict reported salaries of doctorate recipients for 2013. __________________________________

```
tab emsecpb, gen(sector)

tab ndgmemg, gen(field)

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

•••

Complete! Check your Answers!

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.

ANSWERS - Analyze the Sample – Part I Frequencies

Step 1

Analyze the Data

•••

Step 2

Weighting the Data

A) On the website, find the codes page for the OCEDRLP variable and 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 *sample reference period.*

C) How many doctorate recipients were employed in an occupation closely related to their field in 2013 in the SDR survey? 17,696

D) What proportion of doctorate recipients in the SDR survey were employed in an occupation closely related to their field in 2013? 66.33%

```
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.

A) How many doctorate recipients nationally had an occupation closely related to their field of degree in 2013? 474,761

```
replace wtsurvy = round(wtsurvy)
tabulate year ocedrlp if ocedrlp != 98 [fw=wtsurvy], row
```

B) What proportion of doctorate recipients nationally had an occupation closely related to their field of degree in 2013? 65.88%

C) How many doctorate recipients were working in the United States in 2013? 720,626

ANSWERS - Analyze the Sample – Part II Relationships in the Data

Section 1

Crosstabs and Means

A) Which doctorate fields were dominated by women in 2013?

Psychology, Sociology/Anthropology, Health, Non-science

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

B) What is the difference in the mean salary between employed female and male doctorate recipients?

Women: $84,402.91; Men:$102,621.40; Difference: $18,218.49

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

C) What are the average salaries for doctorate recipients by employer sector in 2013?

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

Sector	Mean Salary ($)
2 Year College	63,192.15
4 Year College	86,859.80
Government	104,100.10
Business/industry	106,823.80

ANSWERS - Analyze the Sample – Part II Relationships in the Data

Section 1

Regression

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

```
tab emsecpb, gen(sector)
tab ndgmemg, gen(field)
reg salarp gender agep ctzusin minrty supwk sector2 sector3
sector4 field1-field6 if salarp <=150000 & lfstat == 1 & year ==
2013 [fw=wt]
```

Variable	Coefficient	t-statistic
Gender	10020	103.31
Age	331	83.69
US citizen	6996	52.5
Minority	-4806	-31.37
Supervisory Work	23178	269.95
4 year college	17111	73.57
Government	34322	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	8152	3.71
Engineering	25490	11.61
S&E related fields	17322	7.87
Constant	5059	2.28