IPUMS– HigherEd Extraction and Analysis

Exercise 1 - SAS

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.

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

SAS Code to Review

Code	Purpose
proc freq;	Begins a frequency procedure
proc means;	Begins a means procedure, returns the mean value of a variable
tables	Required syntax to display frequencies
where	Selects only specified cases to include in a procedure

Review Answer Key (page 7)

Common Mistakes to Avoid

1 Not fully decompressing the data

2 Giving the wrong filepath to indicate the dataset

3 Forget to close a procedure with "run;"

4 Forget to terminate a command with a semicolon ";"

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 SAS. 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 SAS 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 the "highered_000##.sas" file.
- In the do file window, change the first line from "libname IPUMS '.'" to "libname IPUMS \\Documents...;" using the file directory where you saved your data files.
- After "filename ASCIIDAT", enter the full file location, ending with highered_000##.dat";
- Choose Submit under the Run file menu.

Analyze the Sample – Part I Frequencies

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?

```
proc freq;
        tables ocedrlp;
        by year;
        where ocedrlp<98; run;
```

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?

_____

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

_____

```
proc freq;
        tables ocedrlp;
        by year; where ocedrlp<98;
        weight wtsurvy; run;
```

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

_____

Section 1

Crosstabs and Means

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

```
proc freq;
        tables ndgmedp*gender;
        where (salarp <= 150000) and (year eq 2013);
        weight wtsurvy; run;
```

_____

B) What is the difference in the mean salary between employed female and male doctorate recipients in 2013? _____

```
proc means;
        class gender; var salarp;
        where (salarp <= 150000) and (year eq 2013);
        weight wtsurvy; run;
```


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

```
proc means;
        class emsecpb; var salarp;
        where (salarp <= 150000) and (year eq 2013);
        weight wtsurvy; run;
```

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

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.

```
proc reg;
        model salarp = gender agep ctzusin minrty supwk;
        where (salarp <= 150000) and (year eq 2013);
        weight wtsurvy;
run;
```

Step 1

Analyze the Data

A) Find the codes page for the OCEDRLP variable and write down each code value, and what category each code represents. <u>1 Closely related; 2 Somewhat related; 3 Not related; 98 Logical Skip</u>

B) What is the universe for OCEDRLP? <u>Working during the week of *reference period.*</u>

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

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

<u>66.33%</u>

```
proc freq;
        tables ocedrlp;
        by year;
        where ocedrlp<98; run;
```

• • •

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?
<u>474,761</u>

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

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

```
proc freq;
        tables ocedrlp;
        by year; where ocedrlp<98;
        weight wtsurvy; run;
```

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

```
proc freq;
        tables ndgmedp*gender;
        where (salarp <= 150000) and (year eq 2013);
        weight wtsurvv: run:
```

B) What is the difference in the mean salary between employed female and male doctorate recipients in 2013?
  Women: $84,409.26; Men:$102,628.63; Difference: $18,228.37

```
proc means;
        class gender; var salarp;
        where (salarp <= 150000) and (year eq 2013);
```

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

```
proc means;
        class emsecpb; var salarp;
        where (salarp <= 150000) and (year eq 2013);
        weight wtsurvy; run;
```

Sector	Mean Salary ($)
2 Year College	63,194.48
4 Year College	86,864.94
Government	104,107.12
Business/industry	106,827.68

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

Section 1

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.

Regression

```
proc reg;
        model salarp = gender agep ctzusin minrty supwk;
        where (salarp <= 150000) and (year eq 2013);
        weight wtsurvy;
run;
```

Variable	Coefficient	t-statistic
Gender	15474	30.85
Age	264	12.30
US citizen	3891	5.39
Minority	-8068	-9.63
Supervisory Work	23115	49.49
Constant	5172	33.44