Minnesota Population Center
Training and Development

IPUMS – HigherEd Extraction and Analysis

Exercise 2 - 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 student loan debt and the relatedness between highest degree and work activities.

6/2/2015

IPUMS – Higher Ed Training and Development

Research Questions

What are the most common work activities for scientists and engineers in the US?
What factors are most important when predicting remaining student loan debt for college graduates?

Objectives

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

IPUMS Variables

- GENDER: Respondent's gender
- WAPRI: Primary work activity
- AGEP: Age
- MINRTY: Minority indicator
- GOVSUP: Indicator of government support for work
- DGRDG: Degree type of respondent's highest degree
- BAFLN: Financial support during bachelor's degree included loans
- UGBALPB: Balance left on undergraduate student loans
- UGLOANPB: Total amount of undergraduate student loans

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 login screen, enter email address and password and submit it!

Step 1

Make an Extract

- Go back to homepage and go to Select Data
- Click the Select Samples box on the SESTAT tab
- Check the box labeled Select all samples – this will select all years of SESTAT samples. Click on Submit sample selections
- Using the drop down menu or search feature, select the following variables:

GENDER: Gender

MINRTY: Minority indicator

WAPRI: Primary work activity

AGEP: Age

GOVSUP: Indicator of government support for work

DGRDG: Degree type of respondent's highest degree

BAFLN: Loans as financial support for bachelor's degree

UGBALPB: Balance of undergraduate student loans

UGLOANPB: Total amount of undergraduate student loans

WEIGHT: SESTAT sample weight variable

•••

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 DGRDG variable and write down each code value, and what category each code represents. _____

B) Describe the kind of information WAPRI provides.

_____

C) What were the top three most common work activities for SESTAT respondents in 2010? _____

```
tab wapri if year == 2010 & wapri < 90, sort
```

D) What were the top three most common work activities for SESTAT respondents in 2010 whose highest degree was a doctorate? _____

• • •

Step 2

Weighting the Data

```
tab wapri if year == 2010 & wapri < 90 & dgrdg == 3, sort
```

Using weights (WEIGHT)

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. The variable WEIGHT is specifically adjusted to analyze the entire SESTAT sample in a particular year.

A) How many doctorate recipients nationally had an occupation in which the primary work activity was teaching?

_____

B) What proportion of doctorate recipients nationally had an occupation in which the primary work activity was basic research?

_____

```
replace weight = round(weight)
tab wapri if year == 2010 & wapri < 90 & dgrdg == 3
[w=weight], sort
```

Analyze the Sample – Part II Relationships in the Data

Section 1

Crosstabs

A) What is the universe for GOVSUP?

_____

B) What percentage of the science and engineering workforce in the United States received funding from the federal government for their work? Note that we are excluding respondents who were out of universe for GOVSUP. _____

```
tab year govsup if govsup < 90 [w=weight], row
```

C) Do doctorate or bachelor's degree holders tend to receive federal funding for their work? _____

```
tab dgrdg govsup if govsup < 90 & year == 2010 [w=weight], row
```

D) How many individuals in the science and engineering workforce in 1999 used loans from lending institutions to finance their bachelor's degree?

_____

```
tab year bafln if bafln < 90 [w=weight], row
```

Analyze the Sample – Part II Relationships in the Data

Section 2

Means

A) What is the average amount of remaining undergraduate loans for those who used loans to finance their undergraduate degree in 1999? _____

```
mean ugbalpb if ugbalpb < 999990 & year == 1999
[fw=weight]
```

B) What is the average amount of total undergraduate loans borrowed for those who used loans to finance their undergraduate degree in 1999? _____

```
mean ugloanpb if ugloanpb > 0 & ugloanpb < 999990 & year
== 1999 [fw=weight]
```

Note: We exclude zeros from original loans, but not from the remaining balance. We want to include the people who paid off the loans, not the ones who never took out loans.

Regression

C) Now let's use some common variables to predict the value of the remaining undergraduate student loans of the science and engineering workforce in 1999. _____

```
reg ugbalpb agep gender minrty [fw=weight] if ugbalpb <
999990 & year == 1999
```

• • •

Complete! Check your Answers!

ANSWERS - Analyze the Sample – Part I Frequencies

Step 1

Analyze the Data

A) On the website, find the codes page for the DGRDG variable and write down each code value, and what category each code represents. _1 = Bachelor's 2 = Master's 3 = Doctorate 4 = Professional 5 = Other_____

B) Describe the kind of information WAPRI provides. _The work activity the respondent performs for the majority of his/her job. _

C) What were the top three most common work activities for SESTAT respondents in 2010? _ Management and Administration, Professional Services, Teaching ___

```
tab wapri if year == 2010 & wapri < 90, sort
```

• • •

Step 2

Weighting the Data

D) What were the top three most common work activities for SESTAT respondents in 2010 whose highest degree was a doctorate? _ Teaching, Basic research, Management and Administration__

```
tab wapri if year == 2010 & wapri < 90 & dgrdg == 3, sort
```

Using weights (WEIGHT)

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. The variable WEIGHT is specifically adjusted to analyze the entire SESTAT sample in a particular year.

A) How many doctorate recipients nationally had an occupation in which the primary work activity was teaching? 234,630

B) What proportion of doctorate recipients nationally had an occupation in which the primary work activity was basic research? _17.7%_

```
replace weight = round(weight)
tab wapri if year == 2010 & wapri < 90 & dgrdg == 3
[w=weight], sort
```

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

Section 1

Crosstabs

A) What is the universe for GOVSUP in 2010? ____Worked during calendar year 2009_____

B) What percentage of the science and engineering workforce in the United States received funding from the federal government for their work in 2010? Note that we are excluding respondents who were out of universe for GOVSUP. __16.15%_____

```
tab year govsup if govsup < 90 [w=weight], row
```

C) Do doctorate or bachelor's degree holders tend to receive federal funding for their work? ___Doctorates 30% (vs Bachelors, of which only 14.1% receive funding)_____

```
tab dgrdg govsup if govsup < 90 & year == 2010 [w=weight], row
```

D) How many individuals in the science and engineering workforce in 1999 had used loans from lending institutions to finance their bachelor's degree? _____48.95%_____

```
tab year bafln if bafln < 90 [w=weight], row
```

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

Section 2

Means

A) What is the average amount of remaining undergraduate loans for those who used loans to finance their undergraduate degree in 1999? _____$7088.90_____

```
mean ugbalpb if ugbalpb < 999990 & year == 1999
[fw=weight]
```

B) What is the average amount of total undergraduate loans borrowed for those who used loans to finance their undergraduate degree in 1999? _____$8948.18_____

```
mean ugloanpb if ugloanpb > 0 & ugloanpb < 999990 & year
== 1999 [fw=weight]
```

Note: We exclude zeros from original loans, but not from the remaining balance. We want to include the people who paid off the loans, not the ones who never took out loans.

Regression

C) Now let's use some common variables to predict the value of the remaining undergraduate student loans of the science and engineering workforce in 1999. _(Coefficients below)___

```
reg ugbalpb agep gender minrty [fw=weight] if ugbalpb <
999990 & year == 1999
```

Variable	Coef	t-stat
Age	-134	-127
Gender	-568	-52
Minority	563	39
Constant	12008	319