Minnesota Population Center

Training and Development

IPUMS â€“ USA Extraction and Analysis

Exercise 1

OBJECTIVE: Gain an understanding of how the IPUMS dataset is structured and how it can be leveraged to explore your research interests. This exercise will use the IPUMS to explore farm ownership and veteran status in the United States.

11/13/2017

Research Questions

What proportion of the U.S. population lives on farms? Is there an association between veteran status and labor-force participation? What is the trend in carpooling over time by metropolitan area status?

Objectives

- Create and download an IPUMS data extract
- Decompress data file and read data into R
- Analyze the data using sample code
- Validate data analysis work using answer key

IPUMS Variables

- FARM: Household Farm Status
- EMPSTAT: Employment Status
- VETSTAT: Veteran Status
- METRO: Metropolitan Status
- CARPOOL: Mode of carpooling

R Code to Review

This tutorial's sample code and answers use the so-called "tidyverse" style, but R has the blessing (and curse) that there are many different ways to do almost everything. If you prefer another programming style, please feel free to use it. But, for your reference, these are some quick explanations for commands that this tutorial will use:

- **%>%** - The pipe operator which helps make code with nested function calls easier to read. When reading code, it can be read as "and then". The pipe makes it so that code like `ingredients %>% stir() %>% cook()` is equivalent to `cook(stir(ingredients))` (read as "take *ingredients* and then *stir* and then *cook*").
- **as_factor** - Converts the value labels provide for IPUMS data into a factor variable for R
- **summarize** - Summarize a datasets observations to one or more groups
- **group_by** - Set the groups for the summarize function to group by
- **filter** - Filter the dataset so that it only contains these values
- **mutate** - Add on a new variable to a dataset

- **weighted.mean** - Get the weighted mean of the a variable

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 "<-" (or "="). Use "==" to test for equality.

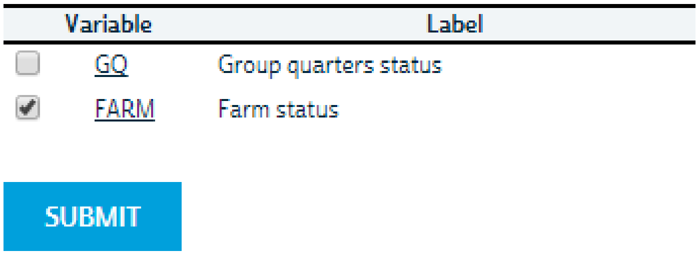
Note: In this exercise, for simplicity we will use "weighted.mean". For analysis where variance estimates are needed, use the survey or srvyr package instead.

Registering with IPUMS

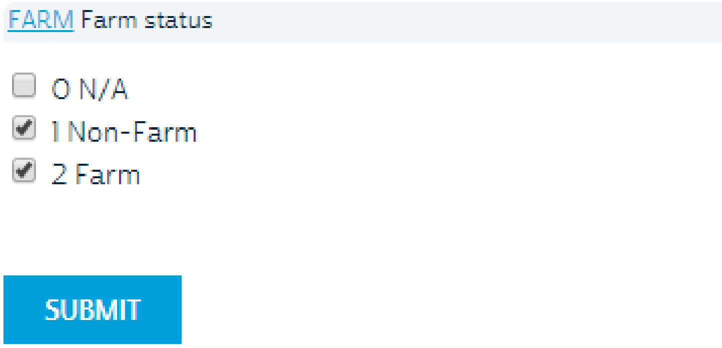
Go to http://usa.ipums.org, click on IPUMS Registration and Login and Apply for access. On login screen, enter email address and password and submit it!

Step 1: Make an Extract

- Go back to homepage and go to Select Data
- Click the Select Samples box, check the boxes for the 1860, 1940, and 1960 1% samples, then click Submit Sample Selections
- Using the drop down menu or search feature, select the following variables:
  - FARM: Household Farm Status
- Click the blue VIEW CART button under your data cart
- Review variable selection. Click the blue Create Data Extract button
- Click â€œSelect Casesâ€•, then select FARM. Then choose only â€œFarmâ€• or â€œNon-Farmâ€• and Submit



	Variable	Label
☐	GQ	Group quarters status
☑	FARM	Farm status



FARM Farm status

☐ 0 N/A
☑ 1 Non-Farm
☑ 2 Farm

Step 2: Request the Data

- Review the â€˜Extract Request Summaryâ€™ screen, describe your extract and click Submit Extract
- You will get an email when the data is available to download
- To get to 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 R. If you would like to use a different stats package, see: http://usa.ipums.org/usa/extract_instructions.shtml

Step 1: Download the Data

- Go to http://usa.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 DDI link next to the extract
- (Optional) Do the same thing for the R script

- You do not need to decompress the data to use it in R

Step 2: Install the ipumsr package

- Open R from the Start menu
- If you haven't already installed the ipumsr package, in the command prompt, type the following command:

```
install.packages("ipumsr")
```

Step 3: Read in the data

- Set your working directory to where you saved the data above by adapting the following command (Rstudio users can also use the "Project" feature to set the working directory. In the menubar, select File -> New Project -> Existing Directory and then navigate to the folder):

```
setwd("~/") # "~/" goes to your Documents directory on most computers
```

- Run the following command from the console, adapting it so it refers to the extract you just created (note the number may not be the same depending on how many extracts you've already made):

```
library(ipumsr)
ddi <- read_ipums_ddi("usa_00001.xml")
data <- read_ipums_micro(ddi)

# Or, if you downloaded the R script, the following is equivalent:
#    source("usa_00001.R")
```

- This tutorial will also rely on the dplyr package, so if you want to run the same code, run the following command (but if you know other ways better, feel free to use them):

```
library(dplyr)
```

- To stay consistent with the exercises for other statistical packages, this exercise does not spend much time on the helpers to allow for translation of the way IPUMS uses labelled values to the way base R does. You can learn more about these in the value-labes vignette in the R package. From R run command: `vignette("value-labels", package = "ipumsr")`

Analyze the Sample â€“ Part I Frequencies

Section 1: Analyze the Variables

Get a basic frequency of the FARM variable for selected historical years.

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

____

B) How many people lived on farms in the US in 1860?

____

C) What proportion of the population lived on a farm in 1860? 1960?

____

```
data %>%
  group_by(YEAR, FARM = as_factor(FARM, level = "both")) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n))
```

Section 2: Using Weights

Using household weights (HHWT)

Suppose you were interested not in the number of people living farms, but in the number of households that were farms. To get this statistic you would need to use the household weight.

In order to use household weight, you should be careful to select only one person from each household to represent that household's characteristics. And you will need to apply the household weight (HHWT).

D) What proportion of households in the sample lived on farms in 1940? (Hint: donâ€™t use the weight quite yet) ______________________

E) How many households were farms in 1940? ___________________

F) What proportion of households were farms in 1940? Does the sample over or under-represent farm households? ________________

____

```
data %>%
  filter(PERNUM == 1 & YEAR == 1940) %>%
  group_by(FARM = as_factor(FARM)) %>%
  summarize(n = sum(HHWT)) %>%
  mutate(pct = n / sum(n))
```

Analyze the Sample â€“ Part II Frequencies

Section 1: Analyze the Data

Create an extract with the variables VETSTAT and EMPSTAT for the years 1980 (5% state) and 2000 (1%) using the instructions above. Run command `source()` *for the new file.*

A) What is the universe for EMPSTAT for this sample, and what are the codes for this variable? ______________________________

____

B) Using the variable description for VETSTAT, describe the issue a researcher would face if they had a research question regarding women serving in the armed forces from World War II until the present.

____

____

C) What percent of veterans and non-veterans were:

i. Employed in 1980? ______________________________

ii. Not part of the labor force in 1980? _____________

D) What percent of veterans and non-veterans were:

i. Employed in 2000? ______________________________

ii. Not part of the labor force in 2000? _____________

```r
data %>%
  filter(YEAR == 1980) %>%
  group_by(
    VETSTAT = as_factor(VETSTAT),
    EMPSTAT = as_factor(EMPSTAT)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n))

data %>%
  filter(YEAR == 2000) %>%
  group_by(
    VETSTAT = as_factor(VETSTAT),
    EMPSTAT = as_factor(EMPSTAT)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n))
```

E) What could explain the difference in relative labor force participation in veterans versus non-veterans between 1980 and 2000?

____

F) How do relative employment rates change when non-labor force participants are excluded in 2000? ______________________

____

```r
data %>%
  filter(YEAR == 2000 & EMPSTAT != 3) %>%
  group_by(
    VETSTAT = as_factor(VETSTAT),
    EMPSTAT = as_factor(EMPSTAT)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n))
```

Analyze the Sample - Part III Advanced Exercises

Section 1: Analyze the Data

Create an extract for 2010 ACS and 1980 5% state with the variables METRO and CARPOOL and read the data into R using the instructions above.

A) What are the codes for METRO and CARPOOL? ____________

____

____

____

____

What might be a limitation of CARPOOL if we are using 2010 and 1980? How could the limitation be fixed? _____________________

____

____

____

Section 2: Weighting explanation

B) What are the proportion of carpoolers and lone drivers not in the metro area, in the central city, and outside the central city in 1980? First, weâ€™ll need to define a new variable from CARPOOL. Letâ€™s name it â€œcarâ€•. If car is 0, it indicates a lone driver, if 1, itâ€™s any form of carpooling. If 2, driving to work is not applicable.

```
data <- data %>%
  mutate(CAR = lbl_relabel(
    CARPOOL,
    lbl(2, "Any form of carpooling") ~ .val %in% c(2, 3, 4, 5)
  ))

data %>%
  filter(YEAR == 1980 & METRO %in% c(1, 2, 3)) %>%
  group_by(METRO = as_factor(METRO), CAR = as_factor(CAR)) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n))
```

Section 3: Analyze the Data

METRO	% drive alone	% carpooler
Not in Metro area	___________	___________

Central city ____________ ____________

Outside central city ____________ ____________

C) Does this make sense? ______________________________________

____

____

____

____

____

D) Do the same for 2010. What does this indicate for the trend in carpooling/driving alone over time in the US? ____________________

____

____

ANSWERS Analyze the Sample â€“ Part I Frequencies

Section 1: Analyze the Variables

Get a basic frequency of the FARM variable for selected historical years.

A) On the website, find the codes page for the FARM variable and write down the code value, and what category each code represents.
0 N/A; 1 Non-Farm; 2 Farm

B) How many people lived on farms in the US in 1860? 1960? *14,393,596 in 1860; 15,880,955 in 1960*

C) What proportion of the population lived on a farm in 1860? 1960?
47.36% in 1860; 8.89% in 1960

```
data %>%
  group_by(YEAR, FARM = as_factor(FARM, level = "both")) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n))
#> # A tibble: 6 x 4
#> # Groups:   YEAR [3]
#>    YEAR         FARM         n        pct
#>   <int>       <fctr>     <dbl>      <dbl>
#> 1  1860 [1] Non-Farm  14391448 0.52632652
#> 2  1860     [2] Farm  12951746 0.47367348
#> 3  1940 [1] Non-Farm 100394864 0.77021942
#> 4  1940     [2] Farm  29950932 0.22978058
#> 5  1960 [1] Non-Farm 163412097 0.91142459
#> 6  1960     [2] Farm  15880955 0.08857541
```

Section 2: Using Weights

Using household weights (HHWT)

Suppose you were interested not in the number of people living farms, but in the number of households that were farms. To get this statistic you would need to use the household weight.

In order to use household weight, you should be careful to select only one person from each household to represent that household's characteristics. And you will need to apply the household weight (HHWT).

D) What proportion of households in the sample lived on farms in 1940? (Hint: donâ€™t use the weight quite yet)
18.61% of households

E) How many households were farms in 1940?
7,075,918 households

F) What proportion of households were farms in 1940? Does the sample over or under-represent farm households?
18.32% of households, sample over-represents farm households

```
data %>%
  filter(PERNUM == 1 & YEAR == 1940) %>%
  group_by(FARM = as_factor(FARM)) %>%
  summarize(n = sum(HHWT)) %>%
  mutate(pct = n / sum(n))
#> # A tibble: 2 x 3
#>       FARM        n     pct
#>     <fctr>    <dbl>   <dbl>
#> 1 Non-Farm 31539654 0.81676
#> 2     Farm  7075918 0.18324
```

ANSWERS Analyze the Sample â€“ Part II Frequencies

Section 1: Analyze the Data

Create an extract with the variables VETSTAT and EMPSTAT for the years 1980 (5% state) and 2000 (1%) using the instructions above. Run command `source()` *for the new file.*

A) What is the universe for EMPSTAT for this sample, and what are the codes for this variable?
Persons age 16+; not available for Puerto Rico & 0 N/A, 1 Employed, 2 Unemployed, 3 Not in labor force

B) Using the variable description for VETSTAT, describe the issue a researcher would face if they had a research question regarding women serving in the armed forces from World War II until the present. s *Women were not counted in VETSTAT until the 1980 Census.*

C) What percent of veterans and non-veterans were:

i. Employed in 1980? *Non-veterans: 54.32%, Veterans: 76.06%*

ii. Not part of the labor force in 1980? *Non-veterans: 20.09%, Veterans: 41.70%*

D) What percent of veterans and non-veterans were:

i. Employed in 2000? *Non-veterans 61.82%, Veterans 54.5%*

ii. Not part of the labor force in 2000? *Non-veterans 34.42%, Veterans 43.11%*

```
data %>%
  filter(YEAR == 1980) %>%
  group_by(
    VETSTAT = as_factor(VETSTAT),
    EMPSTAT = as_factor(EMPSTAT)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n))
#> # A tibble: 7 x 4
#> # Groups:   VETSTAT [3]
#>         VETSTAT            EMPSTAT        n        pct
#>          <fctr>             <fctr>    <dbl>      <dbl>
#> 1           N/A                N/A 56014640 1.00000000
#> 2 Not a veteran           Employed 77354620 0.54319883
#> 3 Not a veteran         Unemployed  5667700 0.03979967
#> 4 Not a veteran Not in labor force 59383400 0.41700151
#> 5       Veteran           Employed 21632520 0.76058257
#> 6       Veteran         Unemployed  1094580 0.03848458
#> 7       Veteran Not in labor force  5714940 0.20093284

data %>%
  filter(YEAR == 2000) %>%
  group_by(
    VETSTAT = as_factor(VETSTAT),
    EMPSTAT = as_factor(EMPSTAT)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n))
#> # A tibble: 10 x 4
#> # Groups:   VETSTAT [3]
#>          VETSTAT            EMPSTAT         n         pct
#>           <fctr>             <fctr>     <dbl>       <dbl>
#>  1           N/A                N/A  64287527 0.942815781
#>  2           N/A           Employed    957167 0.014037438
#>  3           N/A         Unemployed    265760 0.003897532
#>  4           N/A Not in labor force   2676278 0.039249249
#>  5 Not a veteran           Employed 115511618 0.618198136
#>  6 Not a veteran         Unemployed   7009758 0.037515008
#>  7 Not a veteran Not in labor force  64330721 0.344286856
#>  8        Veteran           Employed  14378741 0.544998637
#>  9        Veteran         Unemployed    631334 0.023929506
#> 10        Veteran Not in labor force  11373002 0.431071857
```

E) What could explain the difference in relative labor force participation in veterans versus non-veterans between 1980 and 2000?
Either a growing number of aging veterans or an uptick in PTSD diagnoses in veterans.

F) How do relative employment rates change when non-labor force participants are excluded in 2000?
Veterans have a higher employment rate than non-veterans. (95.8% vs 93.4% employment).

```
data %>%
  filter(YEAR == 2000 & EMPSTAT != 3) %>%
  group_by(
    VETSTAT = as_factor(VETSTAT),
    EMPSTAT = as_factor(EMPSTAT)
  ) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n))
#> # A tibble: 7 x 4
#> # Groups:   VETSTAT [3]
#>         VETSTAT    EMPSTAT         n         pct
#>          <fctr>     <fctr>     <dbl>       <dbl>
#> 1           N/A        N/A  64287527 0.981332338
#> 2           N/A   Employed    957167 0.014610905
#> 3           N/A Unemployed    265760 0.004056757
#> 4 Not a veteran   Employed 115511618 0.942787469
#> 5 Not a veteran Unemployed   7009758 0.057212531
#> 6       Veteran   Employed  14378741 0.957939317
#> 7       Veteran Unemployed    631334 0.042060683
```

ANSWERS Analyze the Sample - Part III Advanced Exercises

Section 1: Analyze the Data

Create an extract for 2010 ACS and 1980 5% state with the variables METRO and CARPOOL and read the data into R using the instructions above.

A) What are the codes for METRO and CARPOOL?
CARPOOL: 0 N/A; 1 Drives alone; 2 Carpool; 3 Shares driving; 4 Drives others only; 5 Passenger only; METRO: 0 Not identifiable; 1 Not in metro area; 2 Central city; 3 Outside central city; 4 Central city status unknown
What might be a limitation of CARPOOL if we are using 2010 and 1980? How could the limitation be fixed?

The code 2 for CARPOOL was taken for the 2010 sample, but 3, 4, and 5 are taken for the 1980 sample. A new variable could be defined to combine these codes.

Section 2: Weighting explanation

B) What are the proportion of carpoolers and lone drivers not in the metro area, in the central city, and outside the central city in 1980? First, weâ€™ll need to define a new variable from CARPOOL. Letâ€™s name it â€œcarâ€•. If car is 1, it indicates a lone driver, if 2, itâ€™s any form of carpooling. If 0, driving to work is not applicable.

```
data <- data %>%
  mutate(CAR = lbl_relabel(
    CARPOOL,
    lbl(2, "Any form of carpooling") ~ .val %in% c(2, 3, 4, 5)
  ))

data %>%
  filter(YEAR == 1980 & METRO %in% c(1, 2, 3)) %>%
  group_by(METRO = as_factor(METRO), CAR = as_factor(CAR)) %>%
  summarize(n = sum(PERWT)) %>%
  mutate(pct = n / sum(n))
#> # A tibble: 9 x 4
#> # Groups:   METRO [3]
#>                                              METRO                    CAR
#>                                             <fctr>                 <fctr>
#> 1                             Not in metro area                    N/A
#> 2                             Not in metro area           Drives alone
#> 3                             Not in metro area Any form of carpooling
#> 4         In metro area, central / principal city                    N/A
#> 5         In metro area, central / principal city           Drives alone
#> 6         In metro area, central / principal city Any form of carpooling
#> 7 In metro area, outside central / principal city                    N/A
#> 8 In metro area, outside central / principal city           Drives alone
#> 9 In metro area, outside central / principal city Any form of carpooling
#> # ... with 2 more variables: n <dbl>, pct <dbl>
```

Section 3: Analyze the Data

METRO	% drive alone	% carpooler
Not in Metro area	24.64%	8.52%
Central city	22.68%	7.05%
Outside central city	32.30%	8.70%

C) Does this make sense?

Yes, commuters outside the metro area or central city are more likely to drive than those in the central city, for whom carpooling is not applicable because they could use public transportation. Commuters outside the central city might be more likely to carpool than *those outside the metro area because they are likely to work within the central city and may live close to others who work in the same concentrated urban area.*

D) Do the same for 2010. What does this indicate for the trend in carpooling/driving alone over time in the US?
In 2010, a greater proportion of the population drove alone and a smaller proportion carpooled.