IPUMS Data Training Exercise:

An introduction to IPUMS NHIS

(Exercise 2 for R)

IPUMS HEALTH SURVEYS

Learning goals

- Create and download an NHIS data extract
- Decompress data file and read data in R
- Analyze the data using sample code

Summary

In this exercise, you will gain an understanding of how the NHIS dataset is structured and how it can be leveraged to explore your research interests. This exercise will use the NHIS dataset to explore associations among BMI, poverty, health status, sleep, and frequency of exercise. You will create a data extract that includes the variables AGE, SEX, POORYN, HEALTH, BMI, HRSLEEP, and VIG10FWK; then you will use the sample code to analyze these data.

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:

Code	Purpose
%>%	The pipe operator 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 provided for IPUMS data into a factor variable for R
summarize	Summarize a dataset's 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
ggplot	Make graphs using ggplot2
weighted.mean	Get the weighted mean of the variable

Common Mistakes to Avoid

- Not changing the working directory to the folder where your data is stored.
- 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 NHIS

Go to https://nhis.ipums.org/nhis/, click on Log in and login if you are a registered user. If you are a first time user, click on Create an account, enter an email address and password, and then submit your user information so you can create NHIS data extracts.

Make a Data Extract

- Return to the homepage and click on Browse and Select Data.

Select Samples

- Click the Select Samples button, and check the box for the 2004 through 2010 samples. Click the submit sample selections button.

Select Variables

- The variable drop-down menus allow you to explore variables by topic. For example, you might expect to find variables about sleep under the "Health Behaviors" 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 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. If you are reviewing variable-specific information, you may click on the "Add to Cart" button near the top of the screen to add this variable to your data cart.

- Using the drop down menu or search feature, select the following variables and add them to your data cart using the plus symbol to the left of the variables:
    - AGE: Age
    - SEX: Sex
    - POORYN: Above or below poverty threshold
    - HEALTH: Health status
    - BMI: Body Mass Index
    - HRSLEEP: Usual hours of sleep per day
    - VIG10FWK: Frequency of vigorous activity (10+ min) per week

Review and submit your extract

- Click the green VIEW CART button under your data cart.
- Review variable selection. Note that additional variables are in your data cart. The data extract system automatically supplies variables that indicate the sample (YEAR), are needed for variance estimation (SERIAL, PERNUM), and are used for weighting the variables and years selected. Click the green Create Data Extract button.
- Review the 'Extract Request Summary' screen, describe your extract, and click Submit Extract.
- You will receive an email when the data is available to download.
- To access the page to download the data, follow the link in the email, or click on the Download or 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: https://ipums.org/support/exercises

Download the data

- Go to https://nhis.ipums.org/nhis/ and click on Download or Revise Extracts.

- Right-click on the Data link next to the extract you created.
- Choose "Save Target As..." (or "Save Link As...").
- Save into "Documents" (Documents 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.

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")
```

Read 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("~/")
```

Note: "~/" 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 have already made):

```
library(ipumsr)
ddi <- read_ipums_ddi("nhis_00001.xml")
data <- read_ipums_micro(ddi)
```

Or, if you downloaded the R script, the following is equivalent: source("nhis_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)
library(ggplot2)
```

- 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-labels vignette in the R package. From R run command:

```
vignette("value-labels", package = "ipumsr")
```

Analyze the Sample

Part 1: Group Means

1. On the website, find the codes page for the HRSLEEP and HEALTH variables. What code values for HRSLEEP should be excluded to avoid skewing the average number of hours slept? How would you restrict the code values for HEALTH to eliminate unknown responses? _____

   _____

   _____

2. Suppose you wanted to study the relationship between hours of sleep and health status. Determine the average reported hours of sleep per night by health status. On average, how many hours does an individual with excellent health in this sample sleep per night? _____

```
data %>%
     filter(HEALTH < 6 & HRSLEEP > 0 & HRSLEEP < 25) %>%
     group_by(HEALTH = as_factor(HEALTH)) %>%
     summarize(HRSLEEP = mean(HRSLEEP))
```

3. Is there a noticeable trend between health status and hours of sleep using this sample? _____

   _____

4. Does the trend change for people under 60 in this sample? _____

```
data %>%
     filter(HEALTH < 6 & HRSLEEP > 0 & HRSLEEP < 25 & AGE < 60)
%>%
     group_by(HEALTH = as_factor(HEALTH)) %>%
     summarize(HRSLEEP = mean(HRSLEEP))
```


Part 2: Weighting the Data

To get a more accurate estimation of demographic patterns from the sample, you will have to use the person weight.

5. Without weights, what proportion of people in this sample was below the poverty threshold in 2010? _____

```
data %>%
     group_by(YEAR, POORYN = as_factor(POORYN)) %>%
     summarize(n = n()) %>%
     mutate(pct = n / sum(n)) %>%
     filter(YEAR == 2010)
```

6. Using weights, what proportion of the population was below the poverty threshold in 2010? _____

```
data %>%
     group_by(YEAR, POORYN = as_factor(POORYN)) %>%
     summarize(n = sum(PERWEIGHT)) %>%
     mutate(pct = n / sum(n)) %>%
     filter(YEAR == 2010)
```

7. Using the household weight (and you must exclude all but one individual from a household), what proportion of households was below the poverty threshold in 2010? _____

```
data %>%
     filter(PERNUM == 1) %>%
     group_by(YEAR, POORYN = as_factor(POORYN)) %>%
     summarize(n = sum(HHWEIGHT)) %>%
     mutate(pct = n / sum(n)) %>%
     filter(YEAR == 2010)
```

Part 3: Generating Variables

Generate a variable that is 0 when an individual exercises less than 3 times a week, and 1 when an individual exercises 3 or more times a week. ***NOTE: Graphing procedures (in Part 4 of this exercise) across statistical packages do not consistently allow for all weight and variance estimation options. Because this section generates variables used in the graphing section of this exercise, neither Part 3 nor Part 4 of this exercise use weights.***

8. Check the output of the do file in the Log window to find the codes for VIG10FWK. Which code means "Never"? _____
   Note: You'll have to exclude codes above 28 when defining when exer3 is greater than 3 times a week.
9. What is the average difference in BMI for an individual in this sample who exercises at least 3 times a week compared to someone who exercises fewer than 3 times per week? _____
   Remember to restrict the codes for BMI so unknown and missing codes are excluded.

```
data <- data %>%
     mutate(EXER3 = VIG10FWK >= 3 & VIG10FWK <= 28)
data %>%
```

```
    filter(BMI > 0 & BMI < 99) %>%
    group_by(EXER3) %>%
    summarize(BMI =  mean(BMI))
```

10. What percent of more frequent exercisers report excellent health? _____
11. What percent of less frequent exercisers report excellent health? _____

```
data %>%
    filter(HEALTH < 6) %>%
    group_by(EXER3, HEALTH = as_factor(HEALTH)) %>%
    summarize(n = n()) %>%
    mutate(pct = n / sum(n))
```

Part 4: Graphing

Create a graph to show the mean BMI over age for males and females.

12. How does the universe for BMI appear on this graph? _____
13. Approximately at what age does BMI peak:

    for women? _____ for men? _____

```
data_summary <- data %>%
    filter(BMI > 0 & BMI < 99) %>%
    group_by(SEX = as_factor(SEX), AGE) %>%
    summarize(BMI = mean(BMI))


ggplot(data_summary, aes(x = AGE, y = BMI, color = SEX)) +
    geom_line()+
    scale_color_manual(values = c(Male = "#7570b3", Female =
"#e6ab02"))
```

Introduce the variable POORYN

14. Create a graph to show how an associated effect of poverty status on BMI differs with gender, controlling for frequent exercise. Comment on three apparent trends.

    _____

```
data_summary <- data %>%
     filter(BMI > 0 & BMI < 99 & POORYN != 9) %>%
     group_by(EXER3, SEX = as_factor(SEX), POORYN =
as_factor(POORYN)) %>%
     summarize(BMI = mean(BMI))


ggplot(data_summary, aes(x = interaction(EXER3, SEX), y = BMI,
fill = poorYN)) +
     geom_col(position = "dodge") +
     scale_fill_manual(values = c("#7570b3", "#e6ab02")) +
     theme(legend.position = "bottom")
```

Answers

Part 1: Group Means

1. On the website, find the codes page for the HRSLEEP and HEALTH variables. What code values for HRSLEEP should be excluded to avoid skewing the average number of hours slept? How would you restrict the code values for HEALTH to eliminate unknown responses? HRSLEEP: 00 NIU; 25 Less than 1 hour; 97 Unknown-refused; 98 Unknown-not ascertained; 99 Unknown-don't know; HEALTH: 7 Unknown-refused; 8 Unknown-not ascertained; 9 Unknown-don't know

2. Suppose you wanted to study the relationship between hours of sleep and health status. Determine the average reported hours of sleep per night by health status. On average, how many hours does an individual with excellent health in this sample sleep per night? 7.2 hours

3. Is there a noticeable trend between health status and hours of sleep using this sample? There seems to be no trend at all, except perhaps Excellent and Poor health have slightly higher averages, which could indicate people in poor health sleep to improve health and people with excellent health are associated with getting more sleep.

4. Does the trend change for people under 60 in this sample? When excluding the older population (perhaps with a higher incidence of poor health), better health is associated with more hours of sleep, though the differences between averages is small.

Part 2: Weighting the Data

5. Without weights, what proportion of people in this sample was below the poverty threshold in 2010? 16.48% of the sample

6. Using weights, what proportion of the population was below the poverty threshold in 2010? 13.76% of the sample

7. Using the household weight (and you must exclude all but one individual from a household), what proportion of households was below the poverty threshold in 2010? 12.91% of the sample

Part 3: Generating Variables

8. Check the output of the do file in the Log window to find the codes for VIG10FWK. Which code means "Never"? 95 Never

9. What is the average difference in BMI for an individual in this sample who exercises at least 3 times a week compared to someone who exercises fewer than 3 times per week? 1.2 BMI (27.7-26.5)

10. What percent of more frequent exercisers report excellent health? 41.37%

11. What percent of less frequent exercisers report excellent health 34.19%

Part 4: Graphing

12. How does the universe for BMI appear on this graph? There appears to be no BMI for individuals below 18, because the universe for BMI is only for adults older than 18.

13. Approximately at what age does BMI peak for

women? ~ 61 years old men? ~50 years old

```
data_summary <- data %>%
 filter(BMI > 0 & BMI < 99) %>%
 group_by(SEX = as_factor(SEX), AGE) %>%
 summarize(BMI = mean(BMI))
ggplot(data_summary, aes(x = AGE, y = BMI, color = SEX)) +
 geom_line() +
 scale_color_manual(values = c(Male = "#7570b3", Female =
     "#e6ab02"))
```

Introduce the variable POORYN

14. Create a graph to show how an associated effect of poverty status on BMI differs with gender, controlling for frequent exercise. Comment on three apparent trends.

    Women under the poverty threshold are more likely to have a higher BMI on average whether or not they exercise. Frequent exercisers have lower BMI's on average in each category. Men under the poverty threshold seem to have a lower BMI on average controlling for exercise.

```
data_summary <- data %>%
     filter(BMI > 0 & BMI < 99 & POORYN != 9) %>%
     group_by( EXER3, SEX = as_factor(SEX), POORYN =
as_factor(POORYN)) %>%
     summarize(BMI = mean(BMI))

ggplot(data_summary, aes(x = interaction(EXER3, SEX), y = BMI,
fill = POORYN)) +
     geom_col(position = "dodge") +
     scale_fill_manual(values = c("#7570b3", "#e6ab02")) +
     theme(legend.position = "bottom")
```