Minnesota Population Center

Training and Development

ATUS Extraction and Analysis

Exercise 2

OBJECTIVE: Gain an understanding of how the ATUS dataset is structured and how it can be leveraged to explore your research interests. This exercise will use the ATUS dataset to explore patterns in time use of Americans in 2009 and 2011.

11/13/2017

Research Questions

What are the trends in time spent on consumer purchases in American households? Does time allocated to food preparation differ across income groups? What characteristics affect the amount of time spent caring for own children?

Objectives

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

ATUS Variables

- ACT_PURCH: Consumer purchases
- REGION: Major region of the United States
- FAMINCOME: Yearly family income
- AGE: Age
- SEX: Sex
- FOODPREP: Created variable for time spent preparing food
- CHILDCARE: Created variable for time with childcare as a secondary activity

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

- **ggplot** - Make graphs using ggplot2
- **weighted.mean** - Get the weighted mean of the a variable

Review Answer Key (At End)

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 ATUS

- Go to http://www.ATUSdata.org/, click on Login at the top, and apply for access. On login screen, enter email address and password and submit it!

Step 1: Make an Extract

- On the homepage, go to Build an Extract (on the left column)
- Click on the "Change Samples" box, and select years 2009 and 2011. Keep the defaults "ATUS respondents" and select "Submit sample selections".
- Under the "Time Use" dropdown menu, select "Activity coding structure". Click on the plus sign next to the variable ACT_PURCH to select the variable and add it to our data cart.
- Click on the "Create time use variable" box at the top. Select "Load" next to ACT_HHACT, then the diamond sign next to Household Activities to expand the category.
  - Unselect all subcategories except for "Food and Drink Preparation, Presentation, and Clean-up", and click "Save time use variable" at the bottom.
  - Name your new variable "foodprep" and select "Save time use variable". This selects the time use variable we just created, and adds it to our data cart.
- Click on the "Create time use variable" box again, and this time select the box at the top "Create variable from scratch".
  - Select the box next to All, then click on the "Secondary Activity" box at the top.

  - Under Secondary Activity, select "Duration of time spent during activity on secondary child care of all children" and then "Save time use variable".
  - Name this new variable "childcare" and label it "Secondary childcare". Then select "Save time use variable" again to select it and add it to the cart.
- Under the Household dropdown menu, click on Geographic and then select the variable REGION.
- Similarly, select the variables FAMINCOME (Household ïƒ Economic), AGE and SEX (both under Person ïƒ Core demographic),

Step 2: Request the Data

- Choose the orange "View Cart" at the top.
- Click on the orange "Create data 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/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://atus.ipums.org/atus/extract_instructions.shtml

Step 1: Download the Data

- Go to http://atus.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("atus_00001.xml")
data <- read_ipums_micro(ddi)

# Or, if you downloaded the R script, the following is equivalent:
#    source("atus_00001.R")
```

- This tutorial will also rely on the dplyr, tidyr and ggplot2 packages, 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)
library(tidyr)
```

- 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 Relationships in the Data

Section 1: Create a New Variable

A) Create a variable that distinguishes individuals who reported consumer purchases on the day of their interview.

```
data <- data %>%
  mutate(PURCHASE = ACT_PURCH > 0)
```

B) Find a frequency for reported consumer purchases for the sample for each year.

____

```
data %>%
  group_by(YEAR) %>%
  summarize(PURCHASE = mean(PURCHASE))
```

C) Is there a difference in incidence of consumer purchasing between men and women in 2009? ______________________________

```
data %>%
  group_by(YEAR, SEX = as_factor(SEX)) %>%
  summarize(PURCHASE = mean(PURCHASE))
```

D) In the sample, when consumer purchases are greater than zero, what is the average amount of time spent on purchases each year? Does it appear that the recession had any effect? ____

```
data %>%
  group_by(YEAR) %>%
  summarize(ACT_PURCH = mean(ACT_PURCH))
```

Section 2: Using Weights

The ATUS sample design requires use of weights to provide and accurate representation at the national level. Half of the interview days in the sample are weekdays, while the other half are weekends. The weight WT06 adjusts for the disproportional number of weekend days, and should be used to weight time use variables. More specifically, WT06 gives the number of person-days in the calendar quarter represented by each survey response. Also keep in mind that the "Eating and Health", "Well-Being", and "Employee Leave" Modules have weights unique to them.

E) Using weights, what is the mean value of time spent on purchases?

____

```
data %>%
  group_by(YEAR) %>%
  summarize(ACT_PURCH = weighted.mean(ACT_PURCH, WT06))
```

Part II: Relationships in the Data

A) Go to the ATUS homepage and choose Demographic Variables. What is the range of values for FAMINCOME? What values indicate family incomes of $35,000 and higher? ____________________________________

B) What is the average time spent in food preparation across income groups? Is there a trend? ____

```
data %>%
  group_by(FAMINCOME = as_factor(FAMINCOME, level = "both")) %>%
  summarize(foodprep = weighted.mean(foodprep, WT06))
```

C) Does the pattern change when you separate the analysis by year?

____

```
data %>%
  group_by(YEAR, FAMINCOME = as_factor(FAMINCOME, level = "both")) %>%
  summarize(foodprep = weighted.mean(foodprep, WT06)) %>%
  spread(YEAR, foodprep)
```

D) What could be an explanation for the result in parts B and C?

____

E) Graph the results from C.

```
data_summary <- data %>%
  group_by(YEAR, FAMINCOME = as_factor(FAMINCOME)) %>%
  summarize(foodprep = weighted.mean(foodprep, WT06))

ggplot(data_summary, aes(x = FAMINCOME, y = foodprep, fill = factor(YEAR))) +
  geom_col(position = "dodge") +
  scale_fill_manual(values = c("#7570b3", "#e6ab02")) +
  theme(
    axis.text.x = element_text(angle = 20, hjust = 1),
    legend.position = "bottom"
  )
```

Analyze the Sample â€“ Part III Frequencies in the Data

A) The way the variable CHILDCARE is constructed, what activities will it include?

____

B) What are the codes for REGION? Find it under Demographic Variables.

____

C) What is the average amount of time for adults to be taking care of children as a secondary activity? ________________________

```
data %>%
  filter(AGE >= 18) %>%
  summarize(childcare = weighted.mean(childcare, WT06))
```

D) Are there differences in means across regions in 2011 in time spent in secondary child care? What about between metropolitan status? Or between men and women? ____

```
data %>%
  filter(AGE >= 18) %>%
  group_by(REGION = as_factor(REGION)) %>%
  summarize(childcare = weighted.mean(childcare, WT06))

data %>%
  filter(AGE >= 18) %>%
  group_by(SEX = as_factor(SEX)) %>%
  summarize(childcare = weighted.mean(childcare, WT06))
```

ANSWERS Analyze the Sample â€“ Part I Relationships in the Data

Section 1: Create a New Variable

A) Create a variable that distinguishes individuals who reported consumer purchases on the day of their interview.

```
data <- data %>%
  mutate(PURCHASE = ACT_PURCH > 0)
```

B) Find a frequency for reported consumer purchases for the sample for each year.
2009: 41.58%; 2011: 40.82

```
data %>%
  group_by(YEAR) %>%
  summarize(PURCHASE = mean(PURCHASE))
#> # A tibble: 2 x 2
#>    YEAR  PURCHASE
#>   <dbl>     <dbl>
#> 1  2009 0.4158227
#> 2  2011 0.4082058
```

C) Is there a difference in incidence of consumer purchasing between men and women in 2009?
Women: 44.97; Men: 37.08

```
data %>%
  group_by(YEAR, SEX = as_factor(SEX)) %>%
  summarize(PURCHASE = mean(PURCHASE))
#> # A tibble: 4 x 3
#> # Groups:   YEAR [?]
#>    YEAR    SEX  PURCHASE
#>   <dbl> <fctr>     <dbl>
#> 1  2009   Male 0.3707905
#> 2  2009 Female 0.4497397
#> 3  2011   Male 0.3765568
#> 4  2011 Female 0.4328252
```

D) In the sample, when consumer purchases are greater than zero, what is the average amount of time spent on purchases each year? Does it appear that the recession had any effect?
2009: 25 minutes; 2011: 24.7 minutes; There appears to be no significant difference between the two years.

```
data %>%
  group_by(YEAR) %>%
  summarize(ACT_PURCH = mean(ACT_PURCH))
#> # A tibble: 2 x 2
#>    YEAR ACT_PURCH
#>   <dbl>     <dbl>
#> 1  2009  24.99520
#> 2  2011  24.65991
```

Section 2: Using Weights

The ATUS sample design requires use of weights to provide and accurate representation at the national level. Half of the interview days in the sample are weekdays, while the other half are weekends. The weight WT06 adjusts for the disproportional number of weekend days, and should be used to weight time use variables. More specifically, WT06 gives the number of person-days in the calendar quarter represented by each survey response. Also keep in mind that the "Eating and Health", "Well-Being", and "Employee Leave" Modules have weights unique to them.

E) Using weights, what is the mean value of time spent on purchases?
2009: 22.7 minutes; 2011: 22.2 minutes.

```
data %>%
  group_by(YEAR) %>%
  summarize(ACT_PURCH = weighted.mean(ACT_PURCH, WT06))
#> # A tibble: 2 x 2
#>    YEAR ACT_PURCH
#>   <dbl>     <dbl>
#> 1  2009  22.69660
#> 2  2011  22.22086
```

ANSWERS Part II: Relationships in the Data

A) Go to the ATUS homepage and choose Demographic Variables. What is the range of values for FAMINCOME? What values indicate family incomes of $35,000 and higher?
Codes 10 through 16.

B) What is the average time spent in food preparation across income groups? Is there a trend?
There appears to be a small peak in income groups 5 through 7, then a slight decline.

```
data %>%
  group_by(FAMINCOME = as_factor(FAMINCOME, level = "both")) %>%
  summarize(foodprep = weighted.mean(foodprep, WT06))
#> # A tibble: 19 x 2
#>                   FAMINCOME foodprep
#>                      <fctr>    <dbl>
#>  1      [1] Less than $5,000 38.24506
#>  2      [2] $5,000 to $7,499 36.21258
#>  3      [3] $7,500 to $9,999 41.97693
#>  4    [4] $10,000 to $12,499 37.75474
#>  5    [5] $12,500 to $14,999 36.10799
#>  6    [6] $15,000 to $19,999 37.59587
#>  7    [7] $20,000 to $24,999 39.15006
#>  8    [8] $25,000 to $29,999 39.60904
#>  9    [9] $30,000 to $34,999 31.34620
#> 10   [10] $35,000 to $39,999 31.97002
#> 11   [11] $40,000 to $49,999 33.02110
#> 12   [12] $50,000 to $59,999 30.21820
#> 13   [13] $60,000 to $74,999 31.57720
#> 14   [14] $75,000 to $99,999 30.02267
#> 15 [15] $100,000 to $149,999 29.49585
#> 16    [16] $150,000 and over 28.82982
#> 17             [996] Refused 32.94190
#> 18          [997] Don't know 33.12206
#> 19               [998] Blank 23.98826
```

C) Does the pattern change when you separate the analysis by year?

Not significantly

```
data %>%
  group_by(YEAR, FAMINCOME = as_factor(FAMINCOME, level = "both")) %>%
  summarize(foodprep = weighted.mean(foodprep, WT06)) %>%
  spread(YEAR, foodprep)
#> # A tibble: 19 x 3
#>                     FAMINCOME   `2009`   `2011`
#>  *                     <fctr>    <dbl>    <dbl>
#>  1      [1] Less than $5,000 38.15744 38.29645
#>  2      [2] $5,000 to $7,499 32.44303 38.65378
#>  3      [3] $7,500 to $9,999 32.81896 49.01534
#>  4    [4] $10,000 to $12,499 37.90156 37.65091
#>  5    [5] $12,500 to $14,999 34.96077 36.99698
#>  6    [6] $15,000 to $19,999 34.17832 40.33390
#>  7    [7] $20,000 to $24,999 39.46194 38.91364
#>  8    [8] $25,000 to $29,999 35.99867 42.83262
#>  9    [9] $30,000 to $34,999 32.33431 30.55648
#> 10   [10] $35,000 to $39,999 32.08862 31.87346
#> 11   [11] $40,000 to $49,999 34.08800 32.03447
#> 12   [12] $50,000 to $59,999 30.10791 30.31296
#> 13   [13] $60,000 to $74,999 31.75090 31.42826
#> 14   [14] $75,000 to $99,999 29.91233 30.12525
#> 15 [15] $100,000 to $149,999 29.48464 29.50493
#> 16    [16] $150,000 and over 28.01417 29.56650
#> 17             [996] Refused 32.94190       NA
#> 18          [997] Don't know 33.12206       NA
#> 19               [998] Blank 23.98826       NA
```

D) What could be an explanation for the result in parts B and C?

The lowest income group may have slightly lower food prep time because they may work multiple jobs or be single parents with not enough time to dedicate to food preparation, while on the other hand, high paying jobs such as lawyers may have a high opportunity cost of time and also work long hours.

E) Graph the results from C.

```
data_summary <- data %>%
  group_by(YEAR, FAMINCOME = as_factor(FAMINCOME)) %>%
  summarize(foodprep = weighted.mean(foodprep, WT06))

ggplot(data_summary, aes(x = FAMINCOME, y = foodprep, fill = factor(YEAR))) +
  geom_col(position = "dodge") +
  scale_fill_manual(values = c("#7570b3", "#e6ab02")) +
  theme(
    axis.text.x = element_text(angle = 20, hjust = 1),
    legend.position = "bottom"
  )
```

ANSWERS Analyze the Sample â€“ Part III Frequencies in the Data

A) The way the variable CHILDCARE is constructed, what activities will it include?
CHILDCARE should include time in any activity in which the respondent also reported child care at the same time.

B) What are the codes for REGION? Find it under Demographic Variables.
1: Northeast; 2: Midwest; 3: South; 4: West

C) What is the average amount of time for adults to be taking care of children as a secondary activity?
101.24 minutes a day

```
data %>%
  filter(AGE >= 18) %>%
  summarize(childcare = weighted.mean(childcare, WT06))
#> # A tibble: 1 x 1
#>   childcare
#>       <dbl>
#> 1  101.2388
```

D) Are there differences in means across regions in 2011 in time spent in secondary child care? What about between metropolitan status? Or between men and women?
The Northeast has the lowest average, while the South has the highest average. Women are much more likely to be incorporating childcare into other activities (74.9 minutes for men, 125.5 for women).

```
data %>%
  filter(AGE >= 18) %>%
  group_by(REGION = as_factor(REGION)) %>%
  summarize(childcare = weighted.mean(childcare, WT06))
```

```
#> # A tibble: 4 x 2
#>      REGION childcare
#>      <fctr>     <dbl>
#> 1 Northeast  93.09563
#> 2   Midwest 100.51188
#> 3     South 102.57567
#> 4      West 106.57059

data %>%
  filter(AGE >= 18) %>%
  group_by(SEX = as_factor(SEX)) %>%
  summarize(childcare = weighted.mean(childcare, WT06))
#> # A tibble: 2 x 2
#>        SEX childcare
#>     <fctr>     <dbl>
#> 1     Male  76.23817
#> 2 Female 124.65051
```