Minnesota Population Center

Training and Development

ATUS Extraction and Analysis

Exercise 1

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 2011.

11/13/2017

Research Questions

Is educational attainment associated with participation in religious or government service activities? Does time spent participating in sports differ by employment status or by day of the week?

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_SPORTS: Sports, exercise, and recreation
- BLS_SOCIAL_RELIG: Religious and spiritual activities
- EDUC: Highest level of education attained
- DAY: Day of the week of interview
- SEX: Sex
- EMPSTAT: Employment status
- WT06: Statistical weight

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

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 Select Samples, select the year 2011, keep the default "ATUS respondents" and click Submit sample selections.
- Click on the "Time Use" dropdown menu, and select "Activity coding structure". Click on the plus sign next to ACT_SPORTS to select the variable and add it to the cart.
- Click again on the "Time Use" dropdown menu, and select "BLS published tables". Select the variable BLS_SOCIAL_RELIG (close to the bottom of the page) to add to our data cart by clicking the plus sign next to the variable.
- Select the Person dropdown menu, click on Technical Person, and select the DAY variable.
- Select the variables SEX (under Person ïf core demographic), EDUC (Person ïf Education) and EMPSTAT (Person ïf Work status).

Step 2: Request the Data

- Click on the orange "View Cart" button at the top right.
- Choose "Create data extract".
- Describe your extract , and click on "Submit extract"
  (If you are not signed in, you will be asked to sign in at this point)
- 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 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 Generate a New Variable

Section 1: Create a Variable

A) On the website, go to Build an Extract, which you'll find on the left hand side of the homepage. Click on Education in the Person dropdown menu under Select Variables. Click on EDUC and find the codes for educational attainment.

B) Create a variable that combines the codes into four categories.

```
ipums_val_labels(data$EDUC)

data <- data %>%
  mutate(
    EDUC_CAT = EDUC %>%
      lbl_na_if(~.lbl == "NIU (Not in universe)") %>%
      lbl_relabel(
        lbl(1, "Less than HS") ~ .val %in% 10:19,
        lbl(2, "HS Degree") ~ .val %in% 20:29,
        lbl(3, "Some college") ~ .val %in% 30:39,
        lbl(4, "College degree +") ~ .val %in% 40:49
      )
  )
```

C) Is there a difference in the average number of minutes spent doing religious activities reported by individuals of different education levels in this sample?

____

```
data %>%
  group_by(EDUC_CAT = as_factor(EDUC_CAT)) %>%
  summarize(BLS_SOCIAL_RELIG = mean(BLS_SOCIAL_RELIG))
```

Section 2: Using weights (WT06)

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.

D) Now answer question C using weights. ____________________

```
data %>%
  group_by(EDUC_CAT = as_factor(EDUC_CAT)) %>%
  summarize(BLS_SOCIAL_RELIG = weighted.mean(BLS_SOCIAL_RELIG, WT06))
```

Analyze the Sample â€“ Part II Relationships in the Data

Section 1: Analyze the data

A) Go to the homepage and choose " Time Use Variables" on the left column under Data. Click ACT_SPORTS and then the link to the description. Is physical exercise the only thing that falls under this category? _________________________

B) What is the average number of minutes spent doing activities under the ACT_SPORTS category for each day of the week? ___________________________

```
data %>%
  group_by(DAY = as_factor(DAY)) %>%
  summarize(ACT_SPORTS = weighted.mean(ACT_SPORTS, WT06))
```

C) What is the average number of minutes spent on these activities for each day of the week if the time was more than zero? Do you have a theory why weekends might differ generally from weekdays? Do these averages seem reasonable?

____

```
data %>%
  filter(ACT_SPORTS > 0) %>%
  group_by(DAY = as_factor(DAY)) %>%
  summarize(ACT_SPORTS_IF_ANY = weighted.mean(ACT_SPORTS, WT06))
```

D) How many people reported exercise or watching sports on the day of their interview? ____

```
data %>%
  group_by(ANY_SPORTS = ACT_SPORTS > 0) %>%
  summarize(ACT_SPORTS_ANY_NUM = n()) %>%
  mutate(ACT_SPORTS_ANY_PCT = ACT_SPORTS_ANY_NUM / sum(ACT_SPORTS_ANY_NUM))
```

E) How many observations does this year's sample have for ACT_SPORTS? What percentage of people reported time spent doing ACT_SPORTS activities?

____

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

Section 1: Analyze the data

A) What is the percent of people employed in the sample?

____

```
data %>%
  group_by(EMPSTAT = as_factor(EMPSTAT)) %>%
  summarize(n = n()) %>%
  mutate(pct = n / sum(n))
```

B) Find the average amount of time participating in sports according to employment status for women. ________________

```
data %>%
  filter(SEX == 2) %>%
  group_by(EMPSTAT = as_factor(EMPSTAT)) %>%
  summarize(ACT_SPORTS = weighted.mean(ACT_SPORTS, WT06))
```

C) Find the average amount of time participating in sports according to employment status for men. What differences do you see?

____

```
data %>%
  filter(SEX == 1) %>%
  group_by(EMPSTAT = as_factor(EMPSTAT)) %>%
  summarize(ACT_SPORTS = weighted.mean(ACT_SPORTS, WT06))
```

ANSWERS Analyze the Sample â€“ Part I Generate a New Variable

Section 1: Create a Variable

A) On the website, go to Build an Extract, which you'll find on the left hand side of the homepage. Click on Education in the Person dropdown menu under Select Variables. Click on EDUC and find the codes for educational attainment.

B) Create a variable that combines the codes into four categories.

```
ipums_val_labels(data$EDUC)
#> # A tibble: 18 x 2
#>      val                                         lbl
#>    <dbl>                                       <chr>
#>  1    10                        Less than 1st grade
#>  2    11                1st, 2nd, 3rd, or 4th grade
#>  3    12                           5th or 6th grade
#>  4    13                           7th or 8th grade
#>  5    14                                  9th grade
#>  6    15                                 10th grade
#>  7    16                                 11th grade
#>  8    17                    12th grade - no diploma
#>  9    20                  High school graduate - GED
#> 10    21              High school graduate - diploma
#> 11    30                  Some college but no degree
#> 12    31      Associate degree - occupational vocational
#> 13    32          Associate degree - academic program
#> 14    40            Bachelor's degree (BA, AB, BS, etc.)
#> 15    41  Master's degree (MA, MS, MEng, MEd, MSW, etc.)
#> 16    42 Professional school degree (MD, DDS, DVM, etc.)
#> 17    43                  Doctoral degree (PhD, EdD, etc.)
#> 18   999                            NIU (Not in universe)

data <- data %>%
  mutate(
    EDUC_CAT = EDUC %>%
      lbl_na_if(~.lbl == "NIU (Not in universe)") %>%
      lbl_relabel(
        lbl(1, "Less than HS") ~ .val %in% 10:19,
        lbl(2, "HS Degree") ~ .val %in% 20:29,
        lbl(3, "Some college") ~ .val %in% 30:39,
        lbl(4, "College degree +") ~ .val %in% 40:49
      )
  )
```

C) Is there a difference in the average number of minutes spent doing religious activities reported by individuals of different education levels in this sample?
Less than HS diploma: 17.16 min; HS Degree: 14.69; Some College: 12.72; College Degree +: 12.17

```
data %>%
  group_by(EDUC_CAT = as_factor(EDUC_CAT)) %>%
  summarize(BLS_SOCIAL_RELIG = mean(BLS_SOCIAL_RELIG))
#> # A tibble: 4 x 2
#>           EDUC_CAT BLS_SOCIAL_RELIG
#>             <fctr>            <dbl>
#> 1     Less than HS         17.16347
#> 2        HS Degree         14.68923
#> 3     Some college         12.72771
#> 4 College degree +         12.17303
```

Section 2: Using weights (WT06)

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.

D) Now answer question C using weights.
Less than HS diploma: 12.26 min; HS Diploma: 9.35; Some College: 8.58; College Degree +: 8.08

```
data %>%
  group_by(EDUC_CAT = as_factor(EDUC_CAT)) %>%
  summarize(BLS_SOCIAL_RELIG = weighted.mean(BLS_SOCIAL_RELIG, WT06))
```

```
#> # A tibble: 4 x 2
#>           EDUC_CAT BLS_SOCIAL_RELIG
#>             <fctr>            <dbl>
#> 1     Less than HS        12.261173
#> 2        HS Degree         9.350110
#> 3     Some college         8.583664
#> 4 College degree +         8.081551
```

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

Section 1: Analyze the data

A) Go to the homepage and choose " Time Use Variables" on the left column under Data. Click ACT_SPORTS and then the link to the description. Is physical exercise the only thing that falls under this category?

No, ACT_SPORTS includes physical exercise, sports recreation, and watching sports. To find only time spent on physical exercise, you would need to Create your own time variable. See Exercise 2 for instructions for creating new variables in the extract builder.

B) What is the average number of minutes spent doing activities under the ACT_SPORTS category for each day of the week?

Sunday 22; Monday 18; Tuesday 19; Wednesday 20; Thursday 17; Friday 17; Saturday 28

```
data %>%
  group_by(DAY = as_factor(DAY)) %>%
  summarize(ACT_SPORTS = weighted.mean(ACT_SPORTS, WT06))
#> # A tibble: 7 x 2
#>         DAY ACT_SPORTS
#>      <fctr>      <dbl>
#> 1    Sunday   22.10527
#> 2    Monday   17.79482
#> 3   Tuesday   19.01565
#> 4 Wednesday   19.97011
#> 5  Thursday   17.30908
#> 6    Friday   17.06546
#> 7  Saturday   28.43936
```

C) What is the average number of minutes spent on these activities for each day of the week if the time was more than zero? Do you have a theory why weekends might differ generally from weekdays? Do these averages seem reasonable?

Sunday 128; Monday 89; Tuesday 93; Wednesday 99; Thursday 84; Friday 104; Saturday 136; Weekend days might have greater time spent exercising because the person does not have work, or watching professional sports games that are traditionally held on weekends.

```
data %>%
  filter(ACT_SPORTS > 0) %>%
  group_by(DAY = as_factor(DAY)) %>%
  summarize(ACT_SPORTS_IF_ANY = weighted.mean(ACT_SPORTS, WT06))
#> # A tibble: 7 x 2
#>         DAY ACT_SPORTS_IF_ANY
#>      <fctr>             <dbl>
#> 1    Sunday         127.73953
#> 2    Monday          89.49505
#> 3   Tuesday          92.50202
#> 4 Wednesday          98.51066
#> 5  Thursday          84.29281
#> 6    Friday         103.92199
#> 7  Saturday         136.25373
```

D) How many people reported exercise or watching sports on the day of their interview?

2,319 people

```
data %>%
  group_by(ANY_SPORTS = ACT_SPORTS > 0) %>%
  summarize(ACT_SPORTS_ANY_NUM = n()) %>%
  mutate(ACT_SPORTS_ANY_PCT = ACT_SPORTS_ANY_NUM / sum(ACT_SPORTS_ANY_NUM))
#> # A tibble: 2 x 3
#>   ANY_SPORTS ACT_SPORTS_ANY_NUM ACT_SPORTS_ANY_PCT
#>        <lgl>              <int>              <dbl>
#> 1      FALSE              10160          0.8141678
#> 2       TRUE               2319          0.1858322
```

E) How many observations does this year's sample have for ACT_SPORTS? What percentage of people reported time spent doing ACT_SPORTS activities?

18.6%

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

Section 1: Analyze the data

A) What is the percent of people employed in the sample?

57.62%

```
data %>%
  group_by(EMPSTAT = as_factor(EMPSTAT)) %>%
  summarize(n = n()) %>%
  mutate(pct = n / sum(n))
#> # A tibble: 5 x 3
#>                  EMPSTAT     n         pct
#>                   <fctr> <int>       <dbl>
#> 1     Employed - at work  7191 0.576248097
#> 2      Employed - absent   297 0.023799984
#> 3 Unemployed - on layoff    66 0.005288885
#> 4   Unemployed - looking   750 0.060100970
#> 5     Not in labor force  4175 0.334562064
```

B) Find the average amount of time participating in sports according to employment status for women.

See table below

```
data %>%
  filter(SEX == 2) %>%
  group_by(EMPSTAT = as_factor(EMPSTAT)) %>%
  summarize(ACT_SPORTS = weighted.mean(ACT_SPORTS, WT06))
#> # A tibble: 5 x 2
#>                  EMPSTAT ACT_SPORTS
#>                   <fctr>      <dbl>
#> 1     Employed - at work   15.02719
#> 2      Employed - absent   19.67868
#> 3 Unemployed - on layoff   19.26337
#> 4   Unemployed - looking   19.00277
#> 5     Not in labor force   13.75864
```

C) Find the average amount of time participating in sports according to employment status for men. What differences do you see?

The average time spent participating in sports doesn't seem to differ across employment status for women, but it does for men.

```
data %>%
  filter(SEX == 1) %>%
  group_by(EMPSTAT = as_factor(EMPSTAT)) %>%
  summarize(ACT_SPORTS = weighted.mean(ACT_SPORTS, WT06))
#> # A tibble: 5 x 2
#>                  EMPSTAT ACT_SPORTS
#>                   <fctr>      <dbl>
#> 1     Employed - at work   21.66151
#> 2      Employed - absent   19.64504
#> 3 Unemployed - on layoff   24.35737
#> 4   Unemployed - looking   49.23854
#> 5     Not in labor force   31.12702
```