Using R effectively

13 October 2020
EARNConversations

Ben Zipperer
Economic Policy Institute

bzipperer@epi.org
@benzipperer

Materials for this class

https://economic.github.io/data_bootcamp/

Outline for today

1. Review last time
   - American Community Survey data
   - Low-wage workers in Virginia
2. Recoding new variables with “if” conditions
3. Combining and transforming datasets: stacking, joining, and reshaping
4. Complex analysis: county-level statistics using ACS
5. Project management in Rstudio

Last time: review, data cleaning

```
# Load the ACS data from IPUMS
acs <- read_dta(”/home/benzipperer/Downloads/acs_2018.dta”)

# Clean up the data
acs_clean <- acs %>%
  # keep only workers
  filter(incwage > 0 & incwage < 999998) %>%
  filter(uhrswork > 0) %>%
  # full-year workers only
  filter(wkswork2 == 6) %>%
  # restrict analysis to VA
  filter(statefip == 51) %>%
  # define wages and low-wage workers
  mutate(wage = incwage / (uhrswork * 51)) %>%
  mutate(low_wage = wage <= 15)
```

Last time: review, analysis

```
# Shares of low-wage workers, overall and by demographic cuts
acs_clean %>%
  summarize(weighted.mean(low_wage, perwt))

acs_clean %>%
  group_by(sex) %>%
  summarize(weighted.mean(low_wage, perwt))

acs_clean %>%
  group_by(race) %>%
  summarize(weighted.mean(low_wage, perwt))
```

Recoding new variables using “if” conditions

Specific tasks

1. Redefine race category to identify Hispanics
   a. define indicator for Hispanic ethnicity/origin
   b. redefine “race” to be more coarse and include Hispanic origin
2. Expand analysis to use all workers rather than just full-year

Examples

1. define 0-1 Hispanic ethnicity from detailed country of origin
2. define aggregated race variable from detailed race
3. define “average” weeks worked, based on binned weeks worked

Useful functions

- `ifelse()`
- `case_when()`

Recoding new variables with “if” conditions: review

Useful functions

- `ifelse(test, yes, no)` creates values=yes/no corresponding to test=true/false
- `case_when(test1 ~ value1, test2 ~ value2, ...)` assigns value if test true

For complex recoding, *always* double-check the results

- something like `count(oldvar, newvar)` can be very helpful

Combining datasets

Specific tasks

1. Create single summary dataset with race-specific *and* overall shares of low-wage workers
2. Add more summary statistics: population counts and sample sizes
3. How does VA compare to the US overall?
4. How does VA compare to nearby states?

Useful functions

- **`bind_rows()`**, **`is.na()`**
- **`summarize()`** summary functions **`sum()`** and **`n()`**
- **`rename()`** and **`full_join()`**
- multiple groups in **`group_by()`**
- **`pivot_wider()`**

County-level analysis using ACS data

ACS data contains one substate identifer

- *puma* = PUMA or Public-Use Microdata Area
- PUMAs are state-specific
- but can overlap several counties

Construct and join PUMA -> county mapping to ACS data

- Geocorr: *http://mcdc.missouri.edu/applications/geocorr2018.html*
- re-scale sample weights to account for PUMA -> county duplication

Project management

Do not automatically save/restart your workspace

R projects

Directories