# Poli 5D Social Science Data Analytics

## More on Stata

Shane Xinyang Xuan
ShaneXuan.com

February 1, 2017

# Contact Information

Shane Xinyang Xuan

xxuan@ucsd.edu

The teaching staff is a team!

| | | |
|---|---|---|
| Professor Roberts | M | 1600-1800 (SSB 299) |
| Jason Bigenho | Th | 1000-1200 (Econ 116) |
| Shane Xuan | M | 1100-1150 (SSB 332) |
| | TH | 1200-1250 (SSB 332) |

Supplemental Materials

UCLA STATA starter kit

`http://www.ats.ucla.edu/stat/stata/sk/`

Princeton data analysis

`http://dss.princeton.edu/training/`

# Road map

Some quick notes before we start today's section:

- Make sure that you pass around the attendance sheet
- Open a .do file
- Import your data ("h1_fams_data.xlsx")
- I will be using my slides, and you will need to type the code in your .do file

# Announcement

I have changed my office hours to

- Monday 11-11:50 am
- Thursday 12-12:50 pm

in order to accommodate as many students as possible.

# Data management

- You should have the data imported before the section starts:
  - cd "/Users/Shane/Dropbox/Poli5D/psets/"
  - import excel "h1_fams_data.xlsx", sheet("Families") firstrow clear

# Data management

- You should have the data imported before the section starts:
  - cd "/Users/Shane/Dropbox/Poli5D/psets/"
  - import excel "h1_fams_data.xlsx", sheet("Families") firstrow clear
- We want to generate a new variable (age_dad2)

# Data management

- You should have the data imported before the section starts:
  - cd "/Users/Shane/Dropbox/Poli5D/psets/"
  - import excel "h1_fams_data.xlsx", sheet("Families") firstrow clear
- We want to generate a new variable (age_dad2)

  generate age_dad2 = age_dad + 1

# Data management

- You should have the data imported before the section starts:
  - cd "/Users/Shane/Dropbox/Poli5D/psets/"
  - import excel "h1_fams_data.xlsx", sheet("Families") firstrow clear
- We want to generate a new variable (age_dad2)
  generate age_dad2 = age_dad + 1
- We want to replace a value in variable race_mom

# Data management

- You should have the data imported before the section starts:
  - cd "/Users/Shane/Dropbox/Poli5D/psets/"
  - import excel "h1_fams_data.xlsx", sheet("Families") firstrow clear
- We want to generate a new variable (age_dad2)
  generate age_dad2 = age_dad + 1
- We want to replace a value in variable race_mom
  replace race_mom = "Black" if race_mom == "Blck"

# Label your variables

- Create a mapping (mom_older_names)

# Label your variables

- Create a mapping (mom_older_names)
  label define mom_older_names 1 "Yes" 0 "No"

# Label your variables

- Create a mapping (mom_older_names)
  label define mom_older_names 1 "Yes" 0 "No"
- Associate the mapping with a variable

# Label your variables

- Create a mapping (mom_older_names)
  label define mom_older_names 1 "Yes" 0 "No"
- Associate the mapping with a variable
  label values mom_older mom_older_names

# Label your variables

- Create a mapping (mom_older_names)
  label define mom_older_names 1 "Yes" 0 "No"
- Associate the mapping with a variable
  label values mom_older mom_older_names
- Assign label

# Label your variables

- Create a mapping (mom_older_names)
  label define mom_older_names 1 "Yes" 0 "No"
- Associate the mapping with a variable
  label values mom_older mom_older_names
- Assign label
  label variable mom_older "Whether mom is older"

# Label your variables

- Create a mapping (mom_older_names)
  label define mom_older_names 1 "Yes" 0 "No"
- Associate the mapping with a variable
  label values mom_older mom_older_names
- Assign label
  label variable mom_older "Whether mom is older"
- Tabulate your results

# Label your variables

- Create a mapping (mom_older_names)
  label define mom_older_names 1 "Yes" 0 "No"
- Associate the mapping with a variable
  label values mom_older mom_older_names
- Assign label
  label variable mom_older "Whether mom is older"
- Tabulate your results
  tab mom_older

# Deal with missingness

- Generate missing

# Deal with missingness

- Generate missing

  generate dadmiss = missing(age_dad)

# Deal with missingness

- Generate missing

  generate dadmiss = missing(age_dad)
- Tabulate your results

# Deal with missingness

- Generate missing

  generate dadmiss = missing(age_dad)
- Tabulate your results

  tab dadmiss

# Deal with missingness

- Generate missing

  generate dadmiss = missing(age_dad)
- Tabulate your results

  tab dadmiss
- lookup functions

# Deal with missingness

- Generate missing

  generate dadmiss = missing(age_dad)
- Tabulate your results

  tab dadmiss
- lookup functions

  list if dadmiss == 1

# Create some "bins"

Scenario: We want to recode interval variables into ordinal variables.

- recode functions

# Create some "bins"

Scenario: We want to recode interval variables into ordinal variables.

- recode functions
  recode age_dad (15/25=1) (26/35=2) (36/55=3), gen(age_dad3)

# Create some "bins"

Scenario: We want to recode interval variables into ordinal variables.

- recode functions
  recode age_dad (15/25=1) (26/35=2) (36/55=3), gen(age_dad3)
- Create a mapping

# Create some "bins"

Scenario: We want to recode interval variables into ordinal variables.

- recode functions
  recode age_dad (15/25=1) (26/35=2) (36/55=3), gen(age_dad3)
- Create a mapping
  label define agenames 1 "young" 2 "middle" 3 "older"

# Create some "bins"

Scenario: We want to recode interval variables into ordinal variables.

- recode functions
  recode age_dad (15/25=1) (26/35=2) (36/55=3), gen(age_dad3)
- Create a mapping
  label define agenames 1 "young" 2 "middle" 3 "older"
- Apply the mapping

# Create some "bins"

Scenario: We want to recode interval variables into ordinal variables.

- recode functions
  recode age_dad (15/25=1) (26/35=2) (36/55=3), gen(age_dad3)
- Create a mapping
  label define agenames 1 "young" 2 "middle" 3 "older"
- Apply the mapping
  label values age_dad3 agenames

# Create some "bins"

Scenario: We want to recode interval variables into ordinal variables.

- recode functions
  recode age_dad (15/25=1) (26/35=2) (36/55=3), gen(age_dad3)
- Create a mapping
  label define agenames 1 "young" 2 "middle" 3 "older"
- Apply the mapping
  label values age_dad3 agenames
- Tabulate results, calculate by row

# Create some "bins"

Scenario: We want to recode interval variables into ordinal variables.

- recode functions
  recode age_dad (15/25=1) (26/35=2) (36/55=3), gen(age_dad3)
- Create a mapping
  label define agenames 1 "young" 2 "middle" 3 "older"
- Apply the mapping
  label values age_dad3 agenames
- Tabulate results, calculate by row
  tab age_dad3 welfare, row

# Visualization in Stata

- Histogram
  - histogram age_mom
  - histogram age_mom, frequency
  - histogram age_mom, percent

# Visualization in Stata

- Histogram
  - histogram age_mom
  - histogram age_mom, frequency
  - histogram age_mom, percent
- Scatterplot
  - twoway (scatter age_mom age_dad, mlabel(idnum) mlabsize(tiny) msize(tiny))

# Visualization in Stata (2)

- Boxplot

# Visualization in Stata (2)

- Boxplot
  - graph box age_mom

# Visualization in Stata (2)

- Boxplot
  - graph box age_mom
  - graph box age_mom, scheme(s1manual)

# Visualization in Stata (2)

- Boxplot
  - graph box age_mom
  - graph box age_mom, scheme(s1manual)
- Barplot

# Visualization in Stata (2)

- Boxplot
  - graph box age_mom
  - graph box age_mom, scheme(s1manual)
- Barplot
  - Code race_mom into numeric variable

# Visualization in Stata (2)

- Boxplot
  - graph box age_mom
  - graph box age_mom, scheme(s1manual)
- Barplot
  - Code race_mom into numeric variable

    encode race_mom, generate(race_mom2)

# Visualization in Stata (2)

- Boxplot
  - graph box age_mom
  - graph box age_mom, scheme(s1manual)
- Barplot
  - Code race_mom into numeric variable
    encode race_mom, generate(race_mom2)
  - install -catplot-

# Visualization in Stata (2)

- Boxplot
  - graph box age_mom
  - graph box age_mom, scheme(s1manual)
- Barplot
  - Code race_mom into numeric variable
    encode race_mom, generate(race_mom2)
  - install -catplot-
    ssc inst catplot

# Visualization in Stata (2)

- Boxplot
  - graph box age_mom
  - graph box age_mom, scheme(s1manual)
- Barplot
  - Code race_mom into numeric variable
    encode race_mom, generate(race_mom2)
  - install -catplot-
    ssc inst catplot
  - Plot

# Visualization in Stata (2)

- Boxplot
  - graph box age_mom
  - graph box age_mom, scheme(s1manual)
- Barplot
  - Code race_mom into numeric variable
    encode race_mom, generate(race_mom2)
  - install -catplot-
    ssc inst catplot
  - Plot
    catplot race_mom2

# Visualization in Stata (3)

## Histogram across units

# Visualization in Stata (3)

## Histogram across units

- histogram age_mom if race_mom==“Black”

# Visualization in Stata (3)

## Histogram across units

- histogram age_mom if race_mom==“Black”
- histogram age_mom if race_mom==“White”

# Visualization in Stata (3)

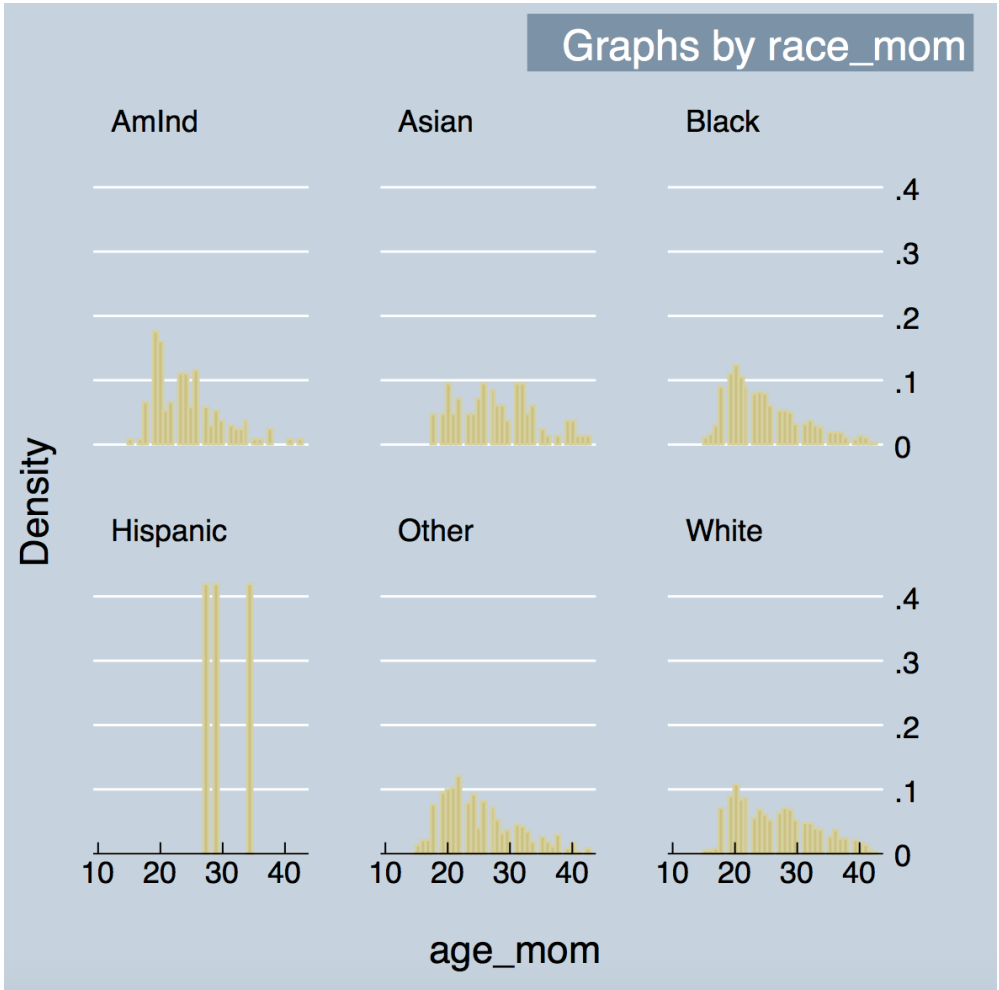
## Histogram across units

- histogram age_mom if race_mom==“Black”
- histogram age_mom if race_mom==“White”
- histogram age_mom, by(race_mom)

# Visualization in Stata (3)

## Histogram across units

- histogram age_mom if race_mom==“Black”
- histogram age_mom if race_mom==“White”
- histogram age_mom, by(race_mom)



# Midterm Review

Please ask questions.