Development Data Boot Camp: Graphs

Ge Sun

University of Notre Dame

May 22, 2023

Outline

Outline

What can graphs do?

▶ Visualize data and demonstrate important features of the data
▶ Present and transfer information effectively and quickly.
  * *"One graph is better than 10 tables"*
▶ Identify outliers and help with debugging
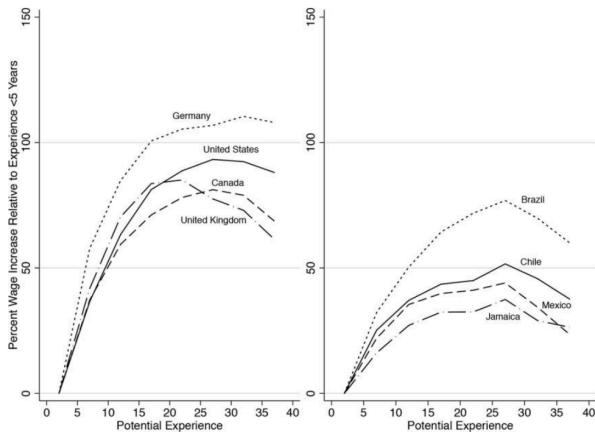
What can graphs do?

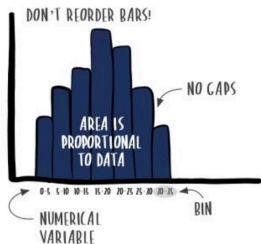Figure: Cross-sectional Experience-wage Profiles from Lagakos et al.(2018)

Types of graphs

- ▶ Different types of data require distinct graph formats to bring out their features.
- ▶ Types of graph we are going to make in Stata
    - Scatter and line plots
    - Bar graphs
    - Pie charts
    - Distribution plots
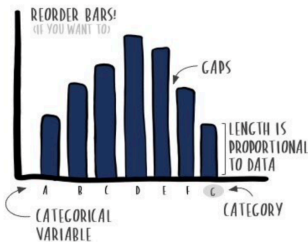   - Regression and fit plots
    - Range and area plots

Question: What is the difference between bar graphs and distribution graphs?

Digression: Bar graphs v.s. Histogram

This is a **histogram**...

DON'T REORDER BARS!

NO GAPS

AREA IS PROPORTIONAL TO DATA

0-5 5-10 10-15 15-20 20-25 25-30 30-35   BIN

NUMERICAL VARIABLE

This is a **bar chart**...

REORDER BARS!
(IF YOU WANT TO)

GAPS

LENGTH IS PROPORTIONAL TO DATA

A   B   C   D   E   F   G   CATEGORY

CATEGORICAL VARIABLE

- ▶ Histograms present the underlying distribution of non-discrete variables; Bar graph shows the comparison across categories.
- ▶ The size of bars in histograms carries meaning (like pdf), while in bar graphs, only the height of the bars matters.
- ▶ Conventional standard: In a histogram, bars should be touching, whereas in a bar chart, there is space between the bars.

Outline

Digressions: the GSS dataset

For five decades, the General Social Survey (GSS) has studied the growing complexity of American society. It is the only full-probability, personal-interview survey designed to monitor changes in both social characteristics and attitudes currently being conducted in the United States.
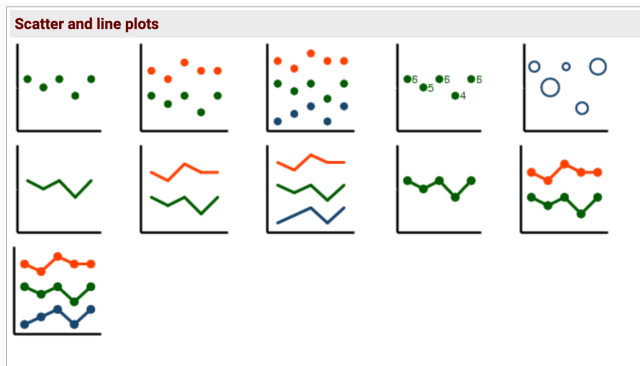
▶ GSS official website

Stata graph commands: Basics

The Stata graph commands begin with the word *"graph"* (optional) followed by the graph plottype, usually *"twoway"*. Then, we specify the graph types, *"scatter"*, *"line"*, or *"connected"*, and variables, we want to make picture about: $y = [y_1, y_2, ...], x$.

```
. graph twoway scatter y x
    . twoway scatter y x
```

P.S. *"twoway" is a family of plots, all of which fit on numeric y and x scales.*

The "twoway" family

Scatter and line plots

- ▶ *scatter*: scatterplot
- ▶ *line*: line plot
- ▶ *connected*: connected-line plot

Stata graph commands: Basics

- ▶ Import GSS2022.dta into Stata. Prepare the dateset for graphs by destring variables: weight, height, age
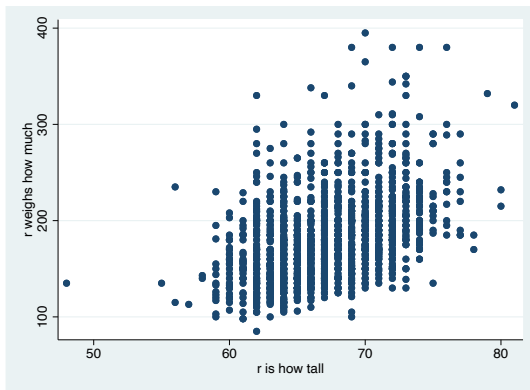
Stata graph commands: Basics

- ▶ Import GSS2022.dta into Stata. Prepare the dateset for graphs by destring variables: weight, height, age
- ▶ Type

  twoway scatter weight height

Stata graph commands: Basics

▶ Import GSS2022.dta into Stata. Prepare the dateset for graphs by destring variables: weight, height, age

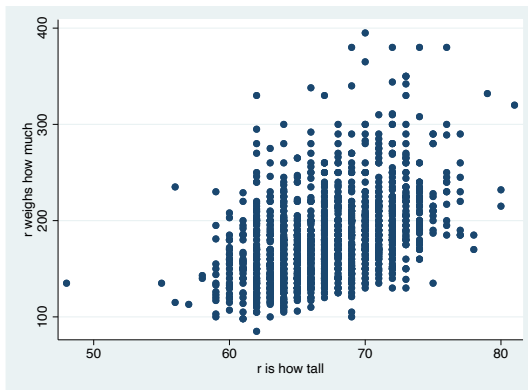▶ Type

twoway scatter weight height

Stata graph commands: Basics

▶ Import GSS2022.dta into Stata. Prepare the dateset for graphs by destring variables: weight, height, age

▶ Type

twoway scatter weight height

▶ Where do those weird axis titles come from?
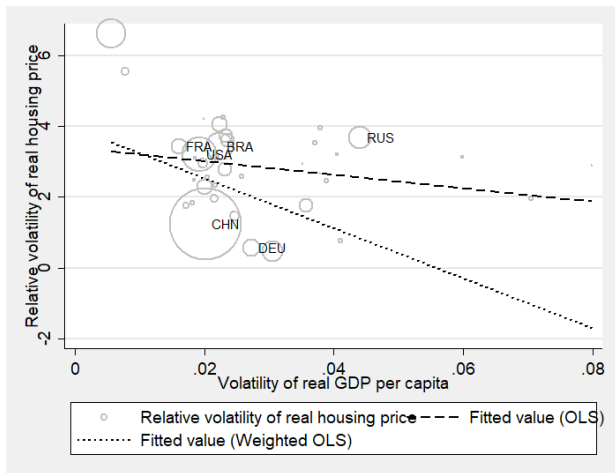
Stata graph commands: Advanced

Complete Syntax:

> [twoway] scatter *varlist* [if] [in] [weight] [,options]

▶ *[if]*: restricts the scope of a command to those observations
for which the value of the expression is true.

> i.e. *if region == 1*

▶ *[in]*: restricts the scope of the command to a specific
observation range.

> i.e. *in 5/1005*: only applies to observations 5-1005

▶ *[weight]*: indicates the weight to be attached to each
observation.

> i.e. the relationship between the volatility of GDP growth rate
> and the volatility of housing price, weighted by the country's
> population

Weighted

▶ the relationship between the volatility of GDP growth rate and the volatility of housing price, weighted by population

Overlaying Graphs

Add one more graph to its original picture

- ▶ Basic syntax

    twoway (plottype varlist...) (plottype varlist...)

    or

    twoway plottype varlist ... ‖ plottype varlist ...

- ▶ *["lfit"]*: linear prediction plot
- ▶ Example:
    - . scatter weight height ‖ lfit weight height
    - . twoway (scatter weight height) (lfit weight height)

Outline

Guidelines for graph usage in an economics paper

- ▶ Include a legend to make your graph self-explanatory.
- ▶ Clearly label all axes: Ensure that both the horizontal and vertical axes are clearly labeled with appropriate units of measurement.
- ▶ Use a consistent scale: Maintain a consistent scale on both axes to ensure the accurate representation of data.
- ▶ Incorporate appropriate sources.

Stata graph commands: Advanced

Complete Syntax:

[twoway] scatter *varlist* [if] [in] [weight] [,options]

► [,options]

options	Description
marker_options	change look of markers (color, size, etc.)
marker_label_options	add marker labels; change look or position
connect_options	change look of lines or connecting method
composite_style_option	overall style of the plot
jitter_options	jitter marker positions using random noise
axis_choice_options	associate plot with alternative axis
twoway_options	titles, legends, axes, added lines and text, by, regions, name, aspect ratio, etc.

Stata twoway manual

legend and axis options

▶ legend options: options for specifying legends

  (plottype varlist ..., legend(contents/locations))

  * order(): which keys appear and their order
  * label():override text for a key
  * off or on: suppress or force display of legend
  * position(clockposstyle): where agents appear

▶ axis options: options for specifying numeric axes

  * axis title options

  graph_command ..., ... ytitle("My Y Title")

  * axis label options

  graph_command ..., ... ylabel(#5)
  graph_command ..., ... xlabel(10(10)50)
  graph_command ..., ... ylabel(, angle(0))

  * axis minor label options

  graph_command ..., ... ymtick(#9)

legend and axis options

Exercise:

1. Generate **one** scatter graph between weight and height for girls and boys separately. (new variable *sex*)
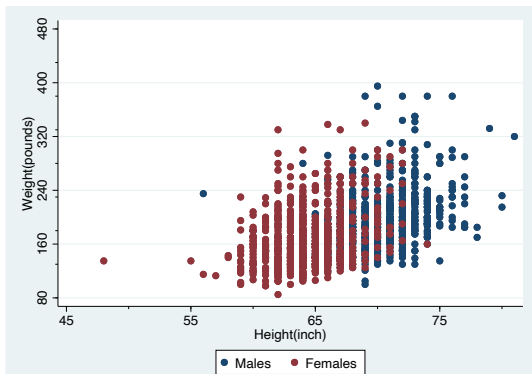
legend and axis options

1. Generate **one** scatter graph between weight and height for girls and boys separately. (new variable *sex*)

   scatter weight height if sex==1 || scatter weight height if sex==2

2. Unfortunately, the default legend at the bottom is completely useless, so you'll need to specify what it should say.

   scatter weight height if sex==1 || scatter weight height if sex==2 , legend(order(1 "Males" 2 "Females"))

3. Turn off the legend by the option, , legend(off)

legend and axis options

4. Now, let's try with axis options:

    scatter weight height if sex==1 || scatter weight height if
    sex==2, legend(order(1 "Males" 2 "Females"))
    ylabel(80(80)480) ymtick(80(50)480)
    ytitle("Weight(pounds)") xlabel(45(10)80) xmtick(45(5)80)
    xtitle("Height(inch)")

marker options and scheme

- ▶ marker options: options for specifying markers
  - * msymbol(): shape of marker
  - * mcolor(): color of marker, inside and out
  - * msize(): size of marker
- ▶ You can use "showmarkers" to see the markers' options.

  ssc install showmarkers

  showmarkers, over(msymbol)

- ▶ scheme: A scheme specifies the overall look of the graph.
  - * possible choice: s2color, s2mono...
  - * help scheme

marker options and scheme

- ▶ marker options: options for specifying markers
  - * msymbol(): shape of marker
  - * mcolor(): color of marker, inside and out
  - * msize(): size of marker
- ▶ You can use "showmarkers" to see the markers' options.

  ssc install showmarkers

  showmarkers, over(msymbol)

- ▶ scheme: A scheme specifies the overall look of the graph.
  - * possible choice: s2color, s2mono...
  - * help scheme

A life-saver: Graphics

Outline

Bar graphs

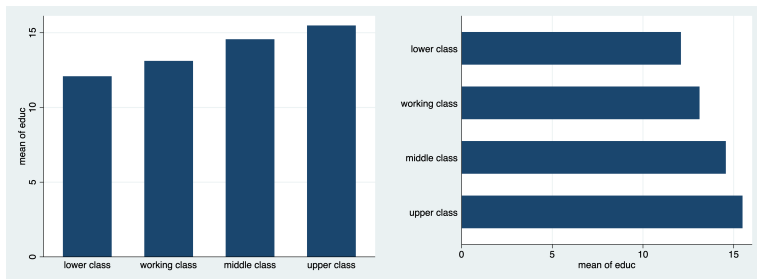▶ Mean of a Quantitative Variable Across a Categorical Variable

Figure: graph bar edu, over(class)

Figure: graph hbar edu, over(class)

Bar graphs

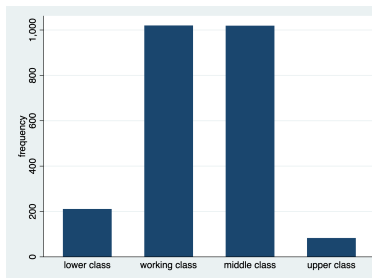▶ Frequencies of a Categorical Variable
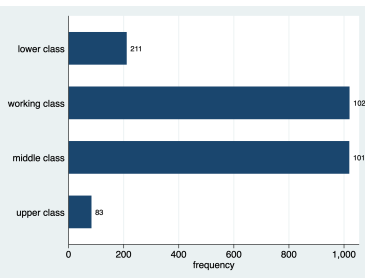

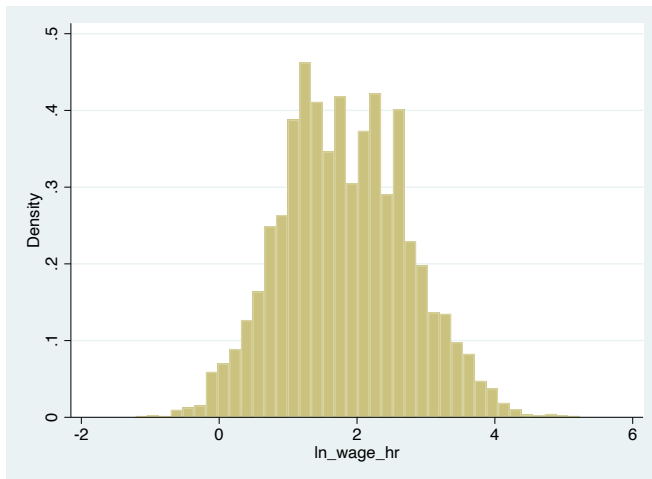Figure: graph bar (count), over(class)

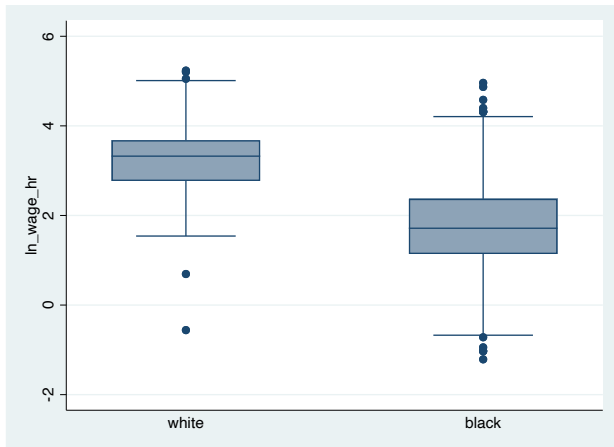Figure: graph hbar (count), over(class) blabel(bar)

Histogram

- ▶ hist ln_wage_hr (example from your previous exercise)

Box Plots

▶ Box plots visually show the distribution of numerical data and skewness through displaying the data quartiles (or percentiles) and averages.
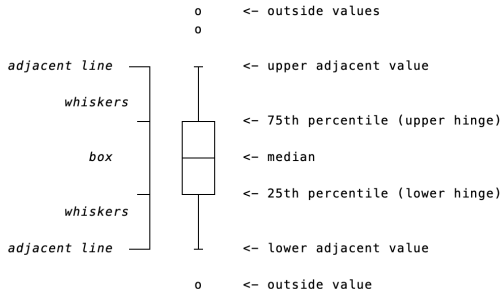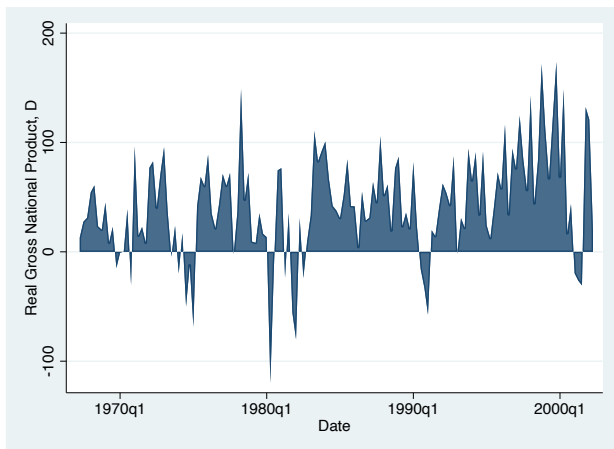
Box Plots

▶ How to understand Box Plots:

help graph box

The encoding and the words used to describe the encoding are

```
                                        o     <- outside values
                                        o

adjacent line ──────┐                   ┬     <- upper adjacent value
                    │
       whiskers     │
                    ├─────            ┌─────┐  <- 75th percentile (upper hinge)
                    │                 │     │
           box      │                 ├─────┤  <- median
                    │                 │     │
                    ├─────            └─────┘  <- 25th percentile (lower hinge)
       whiskers     │
                    │
adjacent line ──────┘                   ┴     <- lower adjacent value

                                        o     <- outside value
```

Area Plots

- ▶ The change in GDP:

Useful links

- Stata Graphics
- Visual overview for creating graphs