Introduction to Social Data Analytics

Class 10

Arushi Kaushik

arkaushi@ucsd.edu

23rd April 2019

Today: Plotting in Stata

By the end of today's lecture, you should be able to:

- Create the following plots in Stata: scatter, line, bar, box, histogram
- Add elements to plots: titles, legends, fitted-lines, etc.
- Interpret elements of plots after creating them (e.g. quartiles in box plots)
- Demonstrate how to overlay multiple plots

Open class10.do if you haven't already.

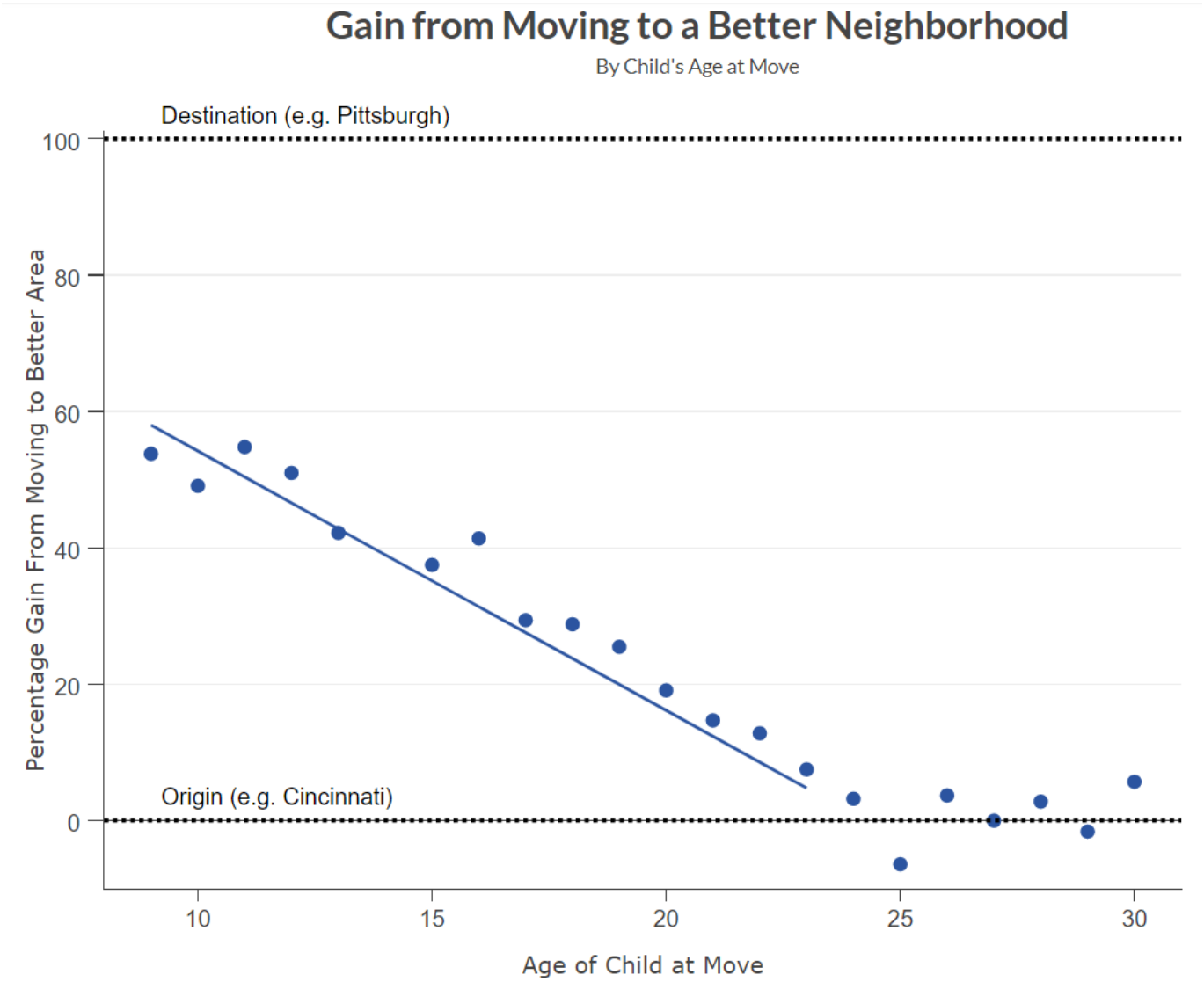
Why do we present data visually?

- The purpose of plotting is to demonstrate relationships between variables in your data.

Why do we present data visually?

- The purpose of plotting is to demonstrate relationships between variables in your data.
- Visual evidence is often more convincing than summary statistics by themselves.

Why do we present data visually?



Plotting syntax

The general syntax is always:

```
graph twoway plottype varlist, [options]
```

For example, you can create a scatter plot with:

```
graph twoway scatter y x
```

Plotting syntax

The general syntax is always:

```
graph twoway plottype varlist, [options]
```

For example, you can create a scatter plot with:

```
graph twoway scatter y x
```

There are *a lot* of different graphs to choose from in Stata.

You don't always need `graph twoway` unless you are presenting different types of graphs on the same plot, but you *always* need the graph type.

Some plot types

Today will we cover the following plot types (though there are several others):

- line
- scatter
- histogram
- bar
- box

Line and Scatter Plots

Syntax: `graph twoway line y1 [y2] [y3] ... x, [options]`

- Use `line` to show the change in a variable (y) over time (x)
- Use `scatter` to show the relationship between two variables where the points are not ordered (e.g. not ordered by time)

Line and Scatter Plots

Syntax: `graph twoway line y1 [y2] [y3] ... x, [options]`

- Use `line` to show the change in a variable (y) over time (x)
- Use `scatter` to show the relationship between two variables where the points are not ordered (e.g. not ordered by time)

Some helpful options:

- `lcolor(color)`
- `legend(order(1 "Label 1" 2 "Label 2" ...))`

Line and Scatter Plots

Syntax: `graph twoway line y1 [y2] [y3] ... x, [options]`

- Use `line` to show the change in a variable (y) over time (x)
- Use `scatter` to show the relationship between two variables where the points are not ordered (e.g. not ordered by time)

Some helpful options:

- `lcolor(color)`
- `legend(order(1 "Label 1" 2 "Label 2" ...))`

Try it:

```
sysuse uslifeexp, clear
graph twoway line le_male le_female year, ///
lcolor(red blue) legend(order(1 "Men" 2 "Women"))
```

Options that apply to all plots

The order of options doesn't matter, but they all come after the comma:

- `title("text")`
- `xtitle("text")`
- `ytitle("text")`
- `name(`*`plotname`*`, replace)`

Options that apply to all plots

The order of options doesn't matter, but they all come after the comma:

- `title("text")`
- `xtitle("text")`
- `ytitle("text")`
- `name(`*`plotname`*`, replace)`

Try adding titles and names to your line plot:

```
graph twoway line le_male le_female year, ///
lcolor(red blue) legend(order(1 "Men" 2 "Women")) ///
title("Life Expectancy for Men vs Women") ///
xtitle("Year") ///
ytitle("Life Expectancy") ///
name(lifeexpmw, replace)
```

Histograms

Syntax: `histogram x`

- Recall from Class 5: histograms display the distribution of the data
- Histograms require only one variable (the y-axis is the frequency)

Histograms

Syntax: `histogram x`

- Recall from Class 5: histograms display the distribution of the data
- Histograms require only one variable (the y-axis is the frequency)

Some helpful options:

- `bin(# of bins)`
- `width(size)`
- `frequency`
- `kdensity`

Histograms

Syntax: `histogram x`

- Recall from Class 5: histograms display the distribution of the data
- Histograms require only one variable (the y-axis is the frequency)

Some helpful options:

- `bin(# of bins)`
- `width(size)`
- `frequency`
- `kdensity`

Try it:

```
sysuse citytemp, clear
histogram tempjan, width(5) frequency kdensity
```

Bar Plots

Syntax: `graph bar (function) x1 [x2]..., [over(`*varname*`)]`

Example: `graph bar (mean) tempjan tempjul, over(region)`

- This creates a barplot depicting the mean temperatures in January and July by region
- `function` can be mean, median, sum, count, percent, etc.

Bar Plots

Syntax: `graph bar (function) x1 [x2]..., [over(`*varname*`)]`

Example: `graph bar (mean) tempjan tempjul, over(region)`

- This creates a barplot depicting the mean temperatures in January and July by region
- `function` can be mean, median, sum, count, percent, etc.

Some helpful options:

- `bar(1, color(`*color1*`)), bar(2, color(`*color2*`))`

Bar Plots

Syntax: `graph bar (function) x1 [x2]..., [over(`*varname*`)]`

Example: `graph bar (mean) tempjan tempjul, over(region)`

- This creates a barplot depicting the mean temperatures in January and July by region
- `function` can be mean, median, sum, count, percent, etc.

Some helpful options:

- `bar(1, color(`*color1*`)), bar(2, color(`*color2*`))`

Try it:

```
graph bar (mean) tempjan tempjul, over(region) ///
legend(order(1 "January" 2 "July")) ///
bar(1, color(green)) bar(2, color(purple))
```

Box Plots

Syntax: `graph box x1 [x2]..., [over(`*varname*`)]`

- Box plots describe statistical properties of data
- Usually used to compare across different groups

Box Plots

Syntax: `graph box x1 [x2]..., [over(`*varname*`)]`

- Box plots describe statistical properties of data
- Usually used to compare across different groups
- Definition: *quartile* - what is it?

Box Plots

Syntax: `graph box x1 [x2]..., [over(`*varname*`)]`

- Box plots describe statistical properties of data
- Usually used to compare across different groups
- Definition: *quartile* - what is it?
- The line inside of the box is the median; the line at the top of the box is the third quartile, the line at the bottom of the box is the first quartile
- The *interquartile range* (IQR) is: IQR = Q3 - Q1

Box Plots

Syntax: `graph box x1 [x2]..., [over(`*varname*`)]`

- Box plots describe statistical properties of data
- Usually used to compare across different groups
- Definition: *quartile* - what is it?
- The line inside of the box is the median; the line at the top of the box is the third quartile, the line at the bottom of the box is the first quartile
- The *interquartile range* (IQR) is: IQR = Q3 - Q1

Try it:

```
graph box tempjan tempjul, over(region) ///
legend(order(1 "January" 2 "July")) ///
box(1, color(green)) box(2, color(purple))
```

Finish up class10.do in your small groups.

Here are the commands/operators we covered today:

- `graph twoway line`
- `graph twoway scatter`
- `histogram`
- `graph bar`
- `graph box`

- `title, xtitle, ytitle`
- `name`
- `legend(order(1 "label1" 2 "label2"))`
- `lcolor`
- `bin, width, frequency, kdensity`
- `bar(1, color(color))`
- `box(1, color(color))`