



Applied Quantitative Methods Network

Kinky Graphs – Getting a bit more out of Stata graphics

Professor Vernon Gayle, University of Stirling
vernon.gayle@stir.ac.uk

20th June 2011 University of St Andrews

© Copyright

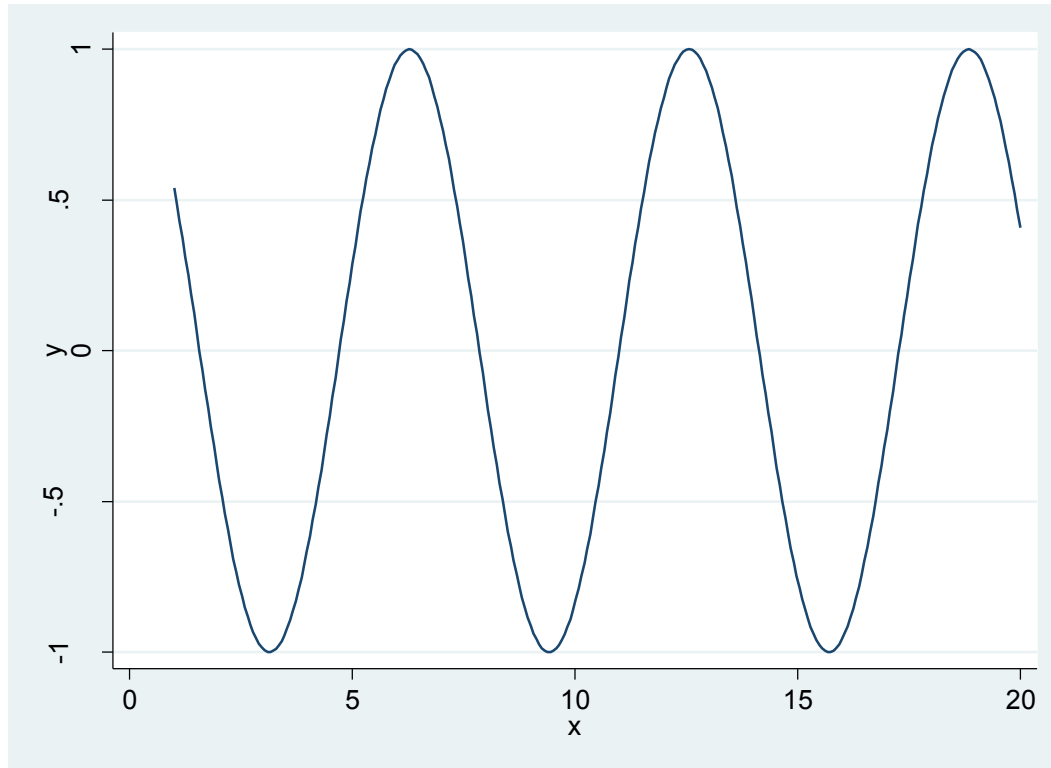
Abstract

- A picture paints a thousand words!
- In my experience, with the exception of weighting, graphing data is one of the most troublesome aspects of data analysis
- In this presentation I will demonstrate how social scientists can get more from the graphical and presentational features of Stata

OED

kinky, adj. and n.

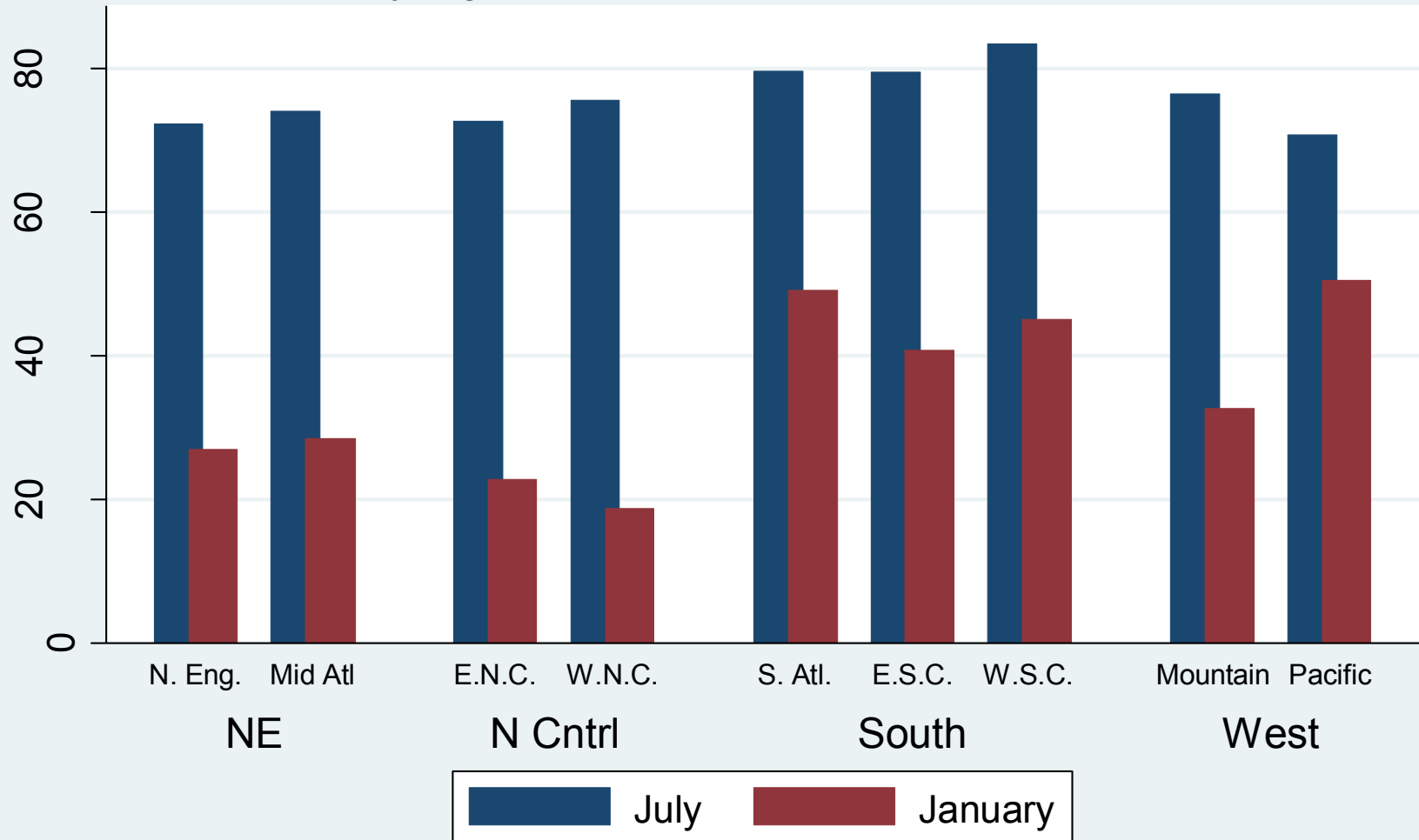
1. Having, or full of, kinks; closely curled or twisted



Graphs displaying data

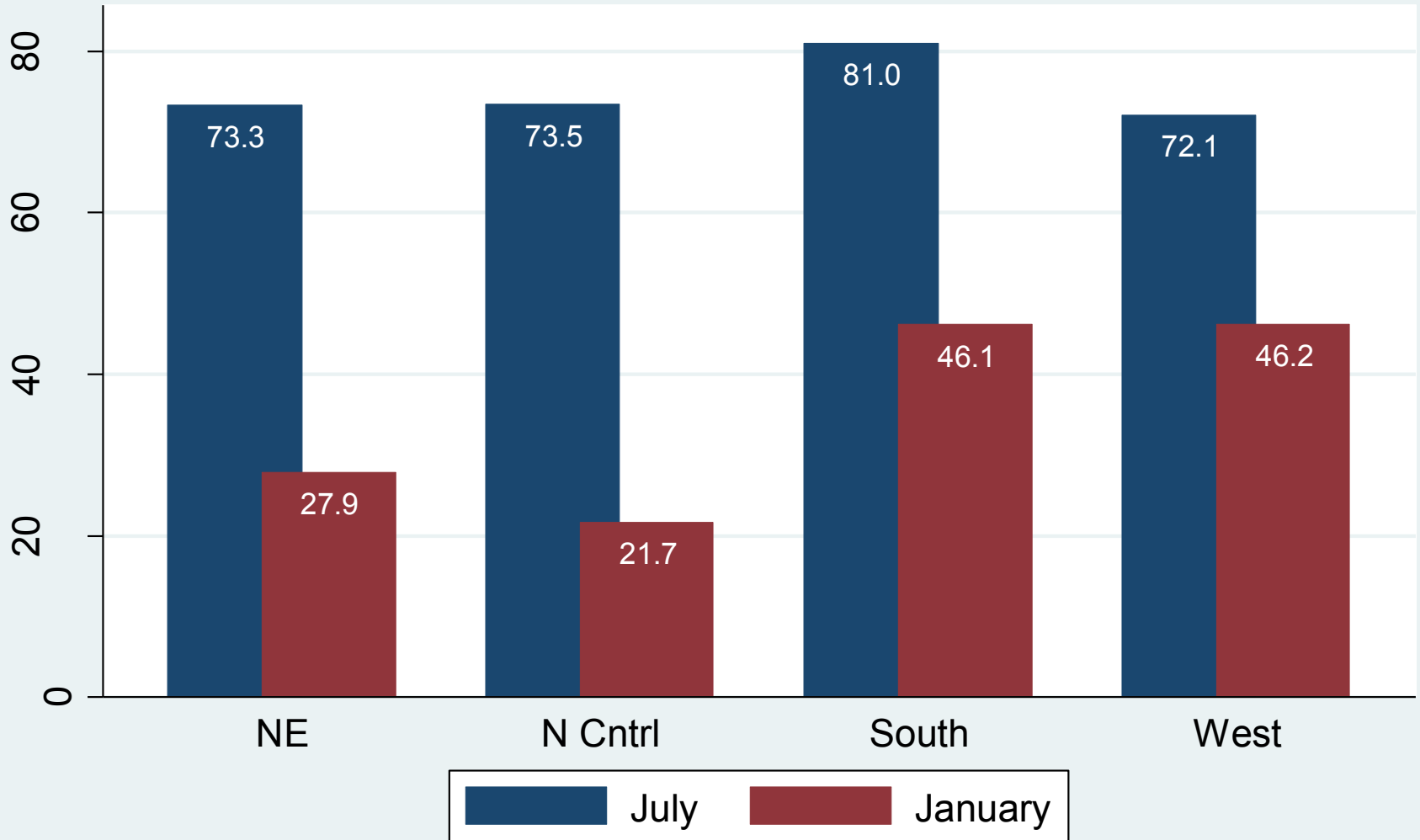
These are usually descriptive...

Average July and January temperatures by region and division of the United States



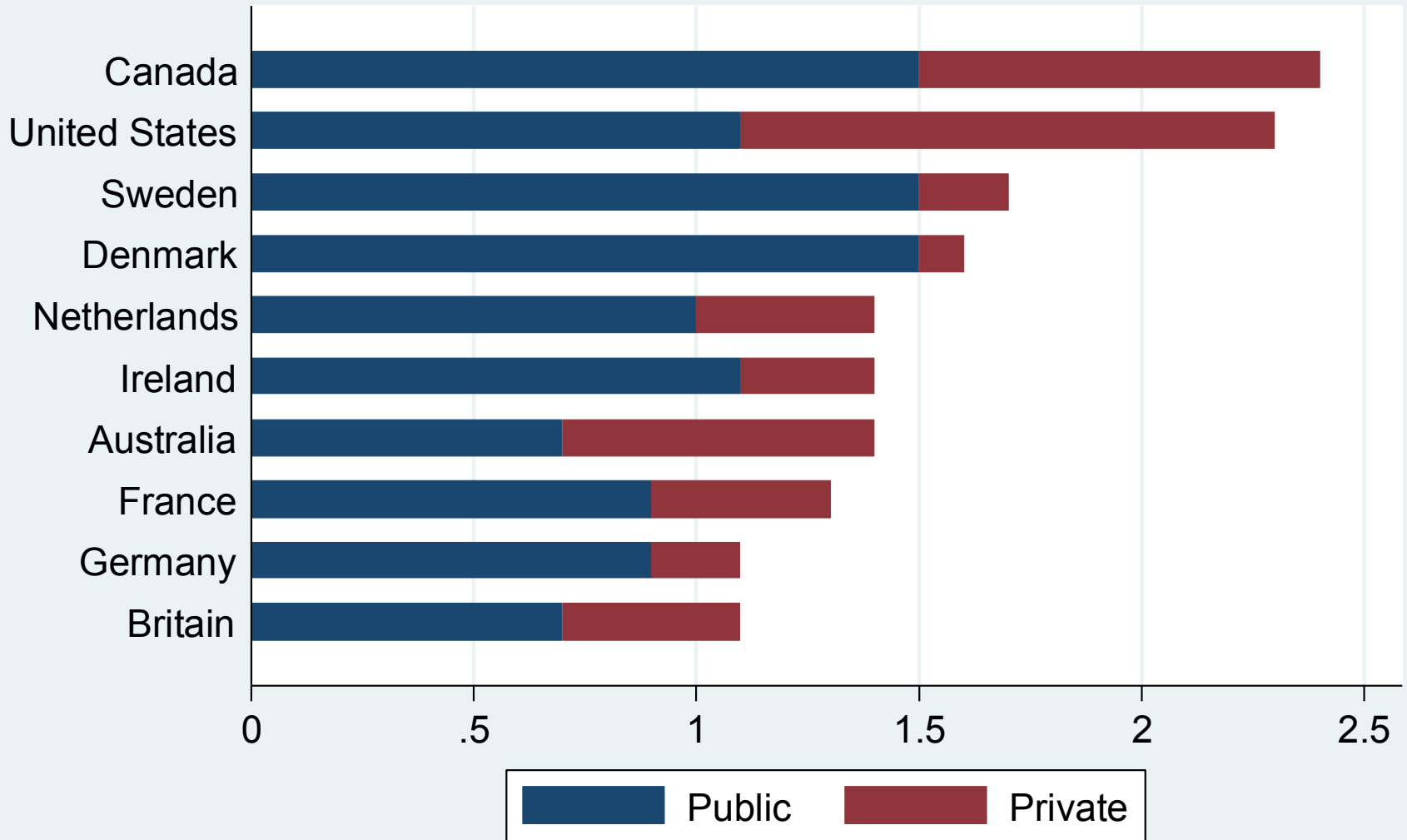
Source: U.S. Census Bureau, U.S. Dept. of Commerce

Average July and January temperatures by regions of the United States



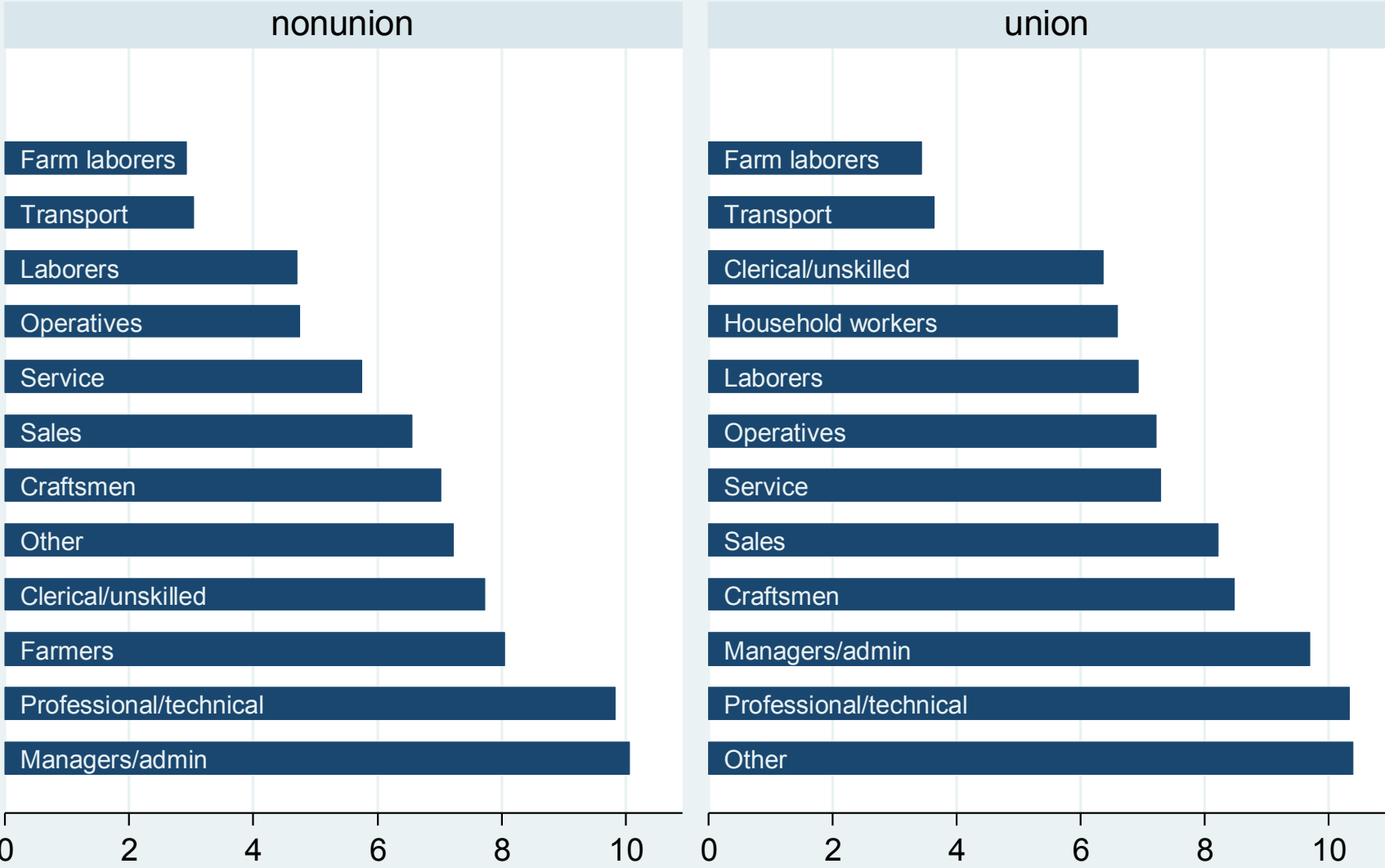
Source: U.S. Census Bureau, U.S. Dept. of Commerce

Spending on tertiary education as % of GDP, 1999



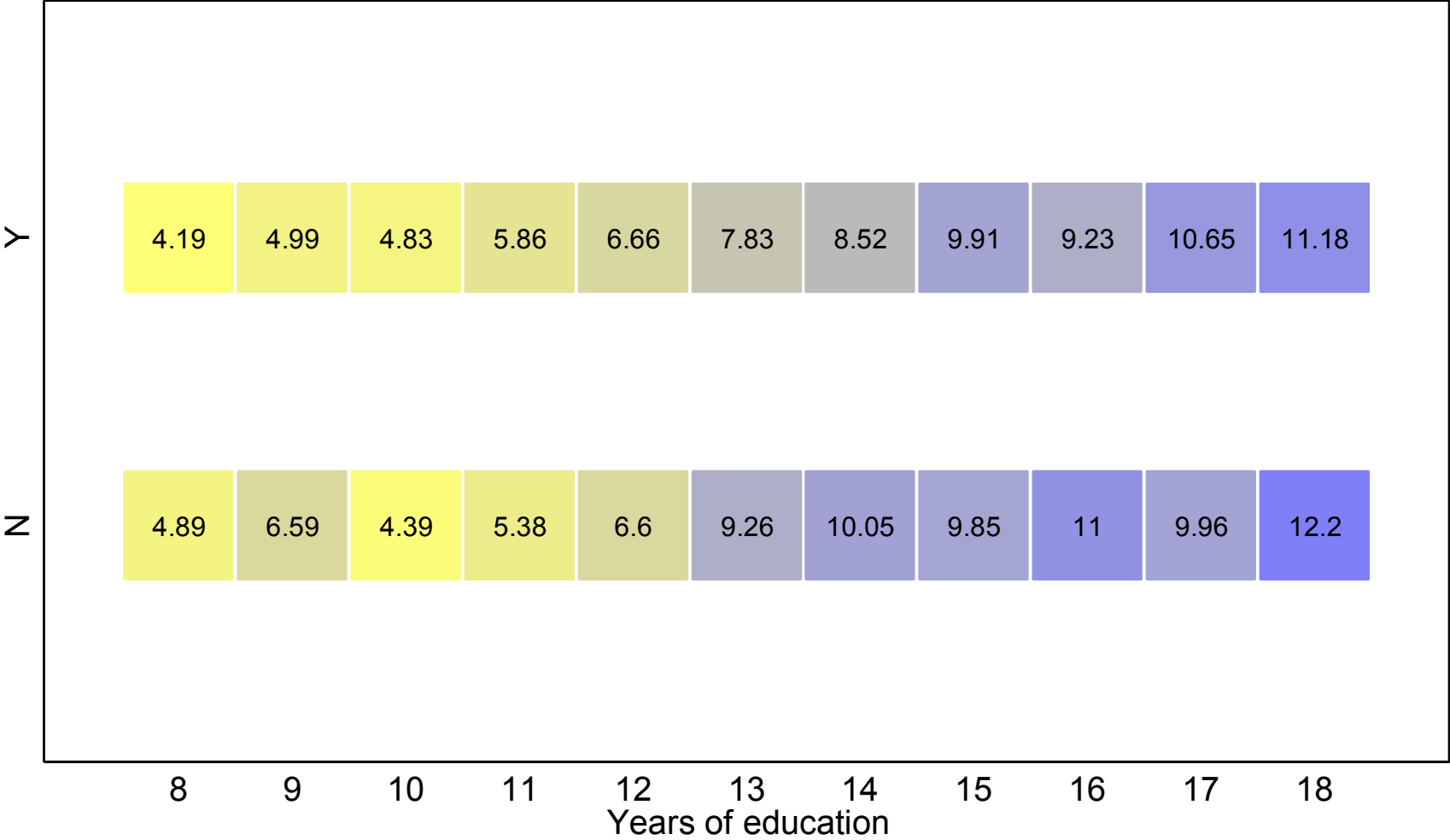
Source: OECD, Education at a Glance 2002

Average Hourly Wage, 1988, Women Aged 34-46

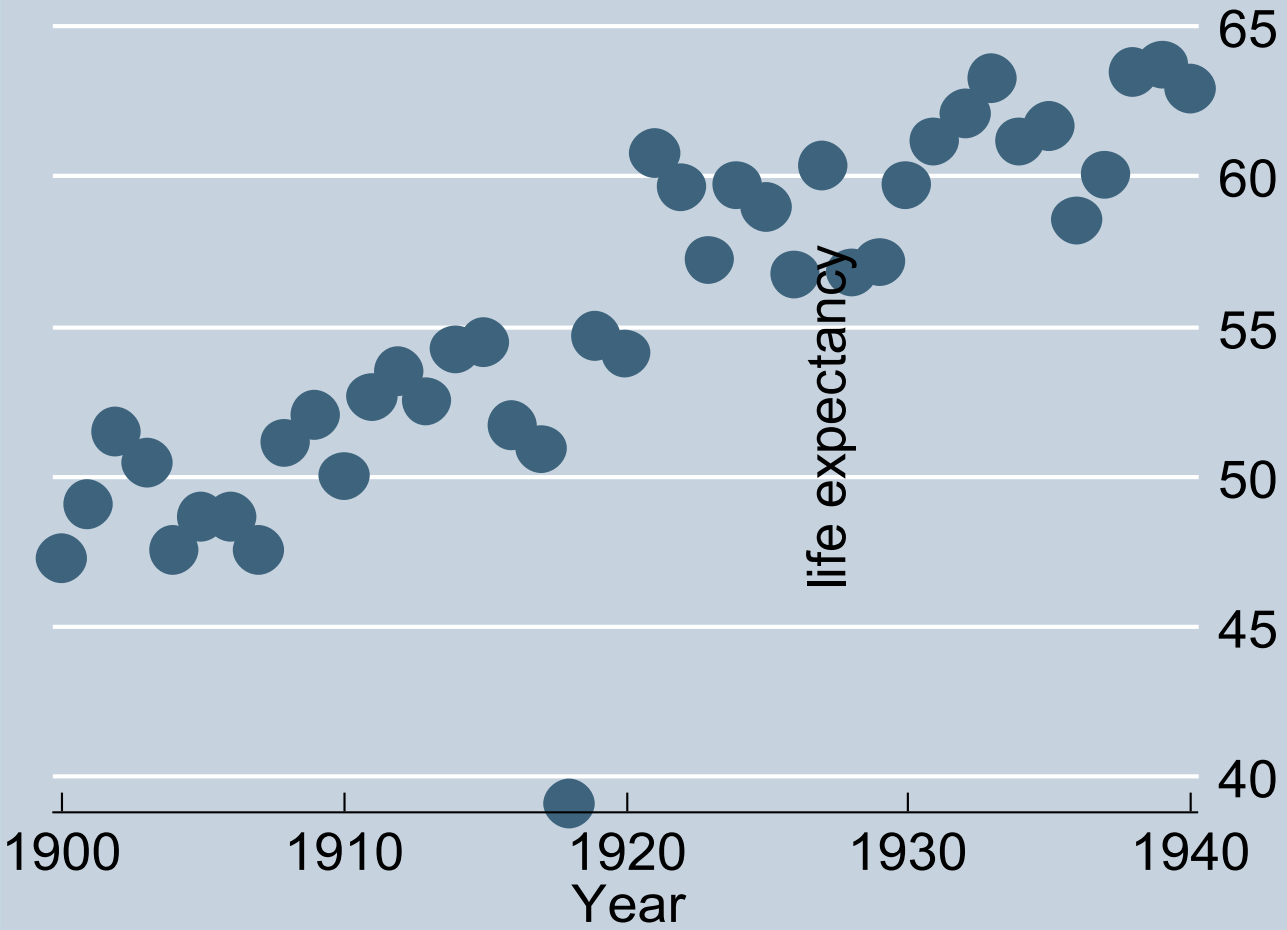


Source: 1988 data from NLS, U.S. Dept. of Labor, Bureau of Labor Statistics

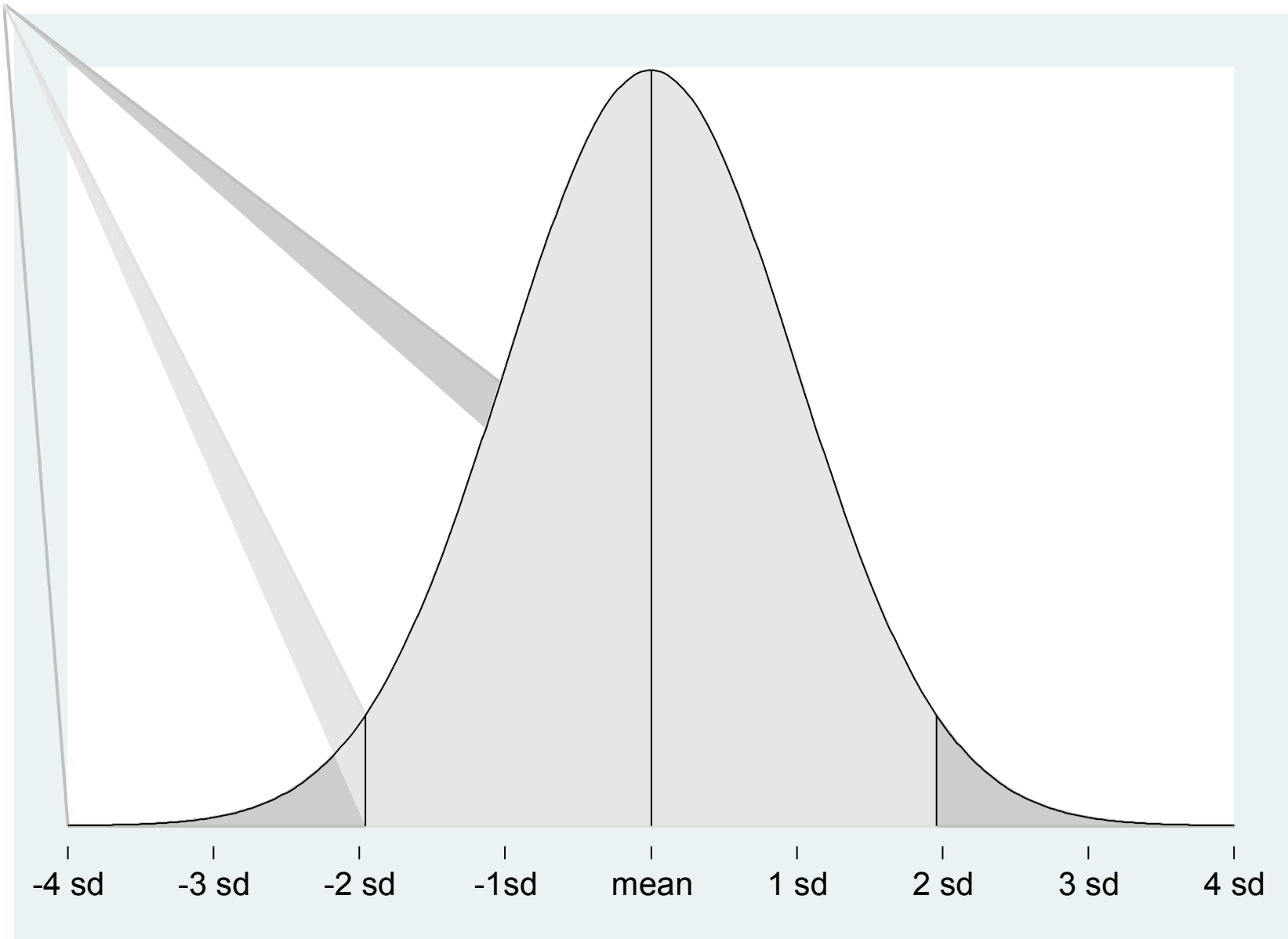
Wage heat map



Life expectancy at birth, U.S.



Source: National Vital Statistics Report, Vol. 50 No. 6

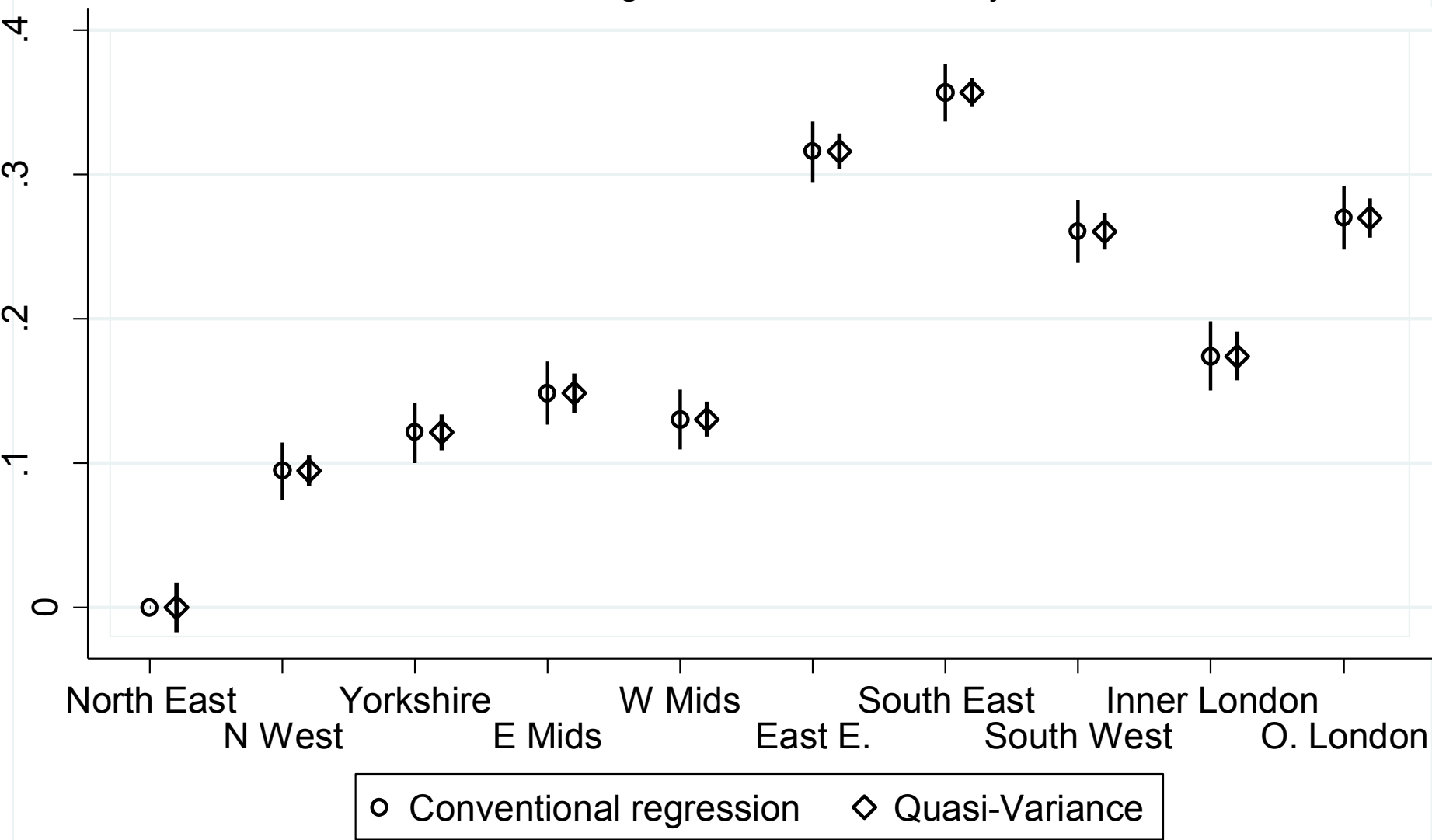


Graphs displaying results

These might include the results from a statistical model...

Predictions of Good Health, by Government Office Region

Confidence intervals of regression coefficients, by estimation method



Source: UK Census 2001 SARS for England, n=1099294.

Model 1: Logistic regression predicting 'Good Health'. Other controls for education and gender

Some preliminary points

I must use Stata syntax
I must use Stata syntax
I must use Stata syntax
I must use Stata syntax
I must use Stata syntax
I must use Stata syntax
I must use Stata syntax
I must use Stata syntax
I must use Stata syntax
I must use Stata syntax
I must use Stata syntax
I must use Stata syntax

I must use Stata syntax
I must use Stata syntax
I must use Stata syntax
I must use Stata syntax
I must use Stata syntax
I must use Stata syntax
I must use Stata syntax
I must use Stata syntax
I must use Stata syntax
I must use Stata syntax
I must use Stata syntax
I must use Stata syntax



Some preliminary points

- Audit trails
 - tracing what you did, why, when, how
- Reproducibility
 - same results every time whoever or wherever
 - supervisors, P.I. etc.
 - Editing, rewriting thesis or re-submission of papers
- Accuracy
 - minimising information loss and errors in analyses and output
- Efficiency
 - automation, using only one software
 - referees comments considerable time-lag returning to work

Some preliminary points

- Graphs are more a matter of taste than tables
- Stata graphics commands are all **programmable** therefore locate within the **audit trail** philosophy
 - Excel does not
- Large number of extension programmes created by users

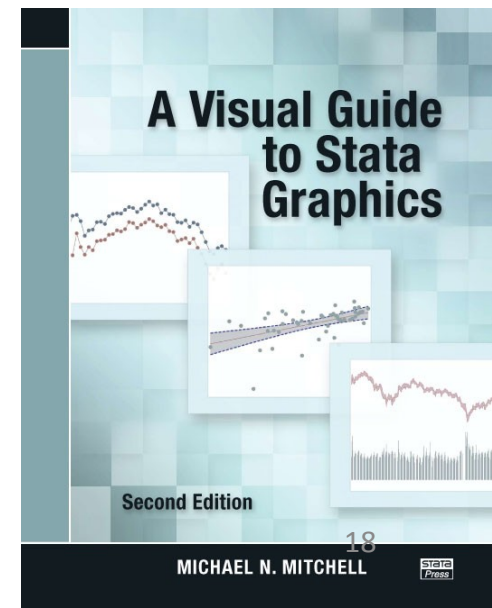
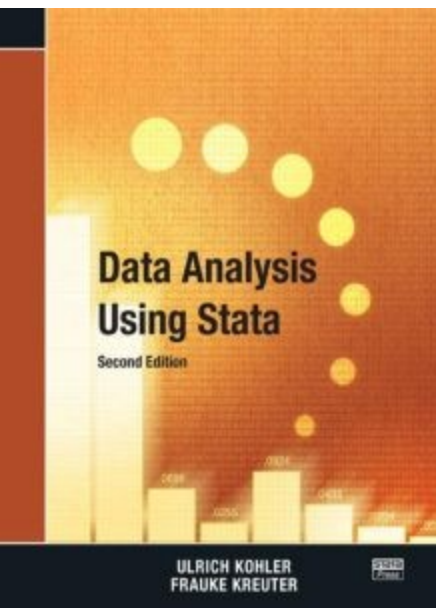
Some preliminary points

- Aim for a top quality output that can be dropped into your thesis, paper or report
- Aim for a 'stand-alone' graph (or table)
 - the reader should be able to understand the output in isolation!

Key references

Kohler, U. and Kreuter, F. (2009) *Data Analysis Using Stata*, Stata Press. ISBN 978-1-59-718-046-7

Mitchell, H. (2008) *A Visual Guide to Stata Graphics* (2nd edition), Stata Press. ISBN 1-59718-039-4

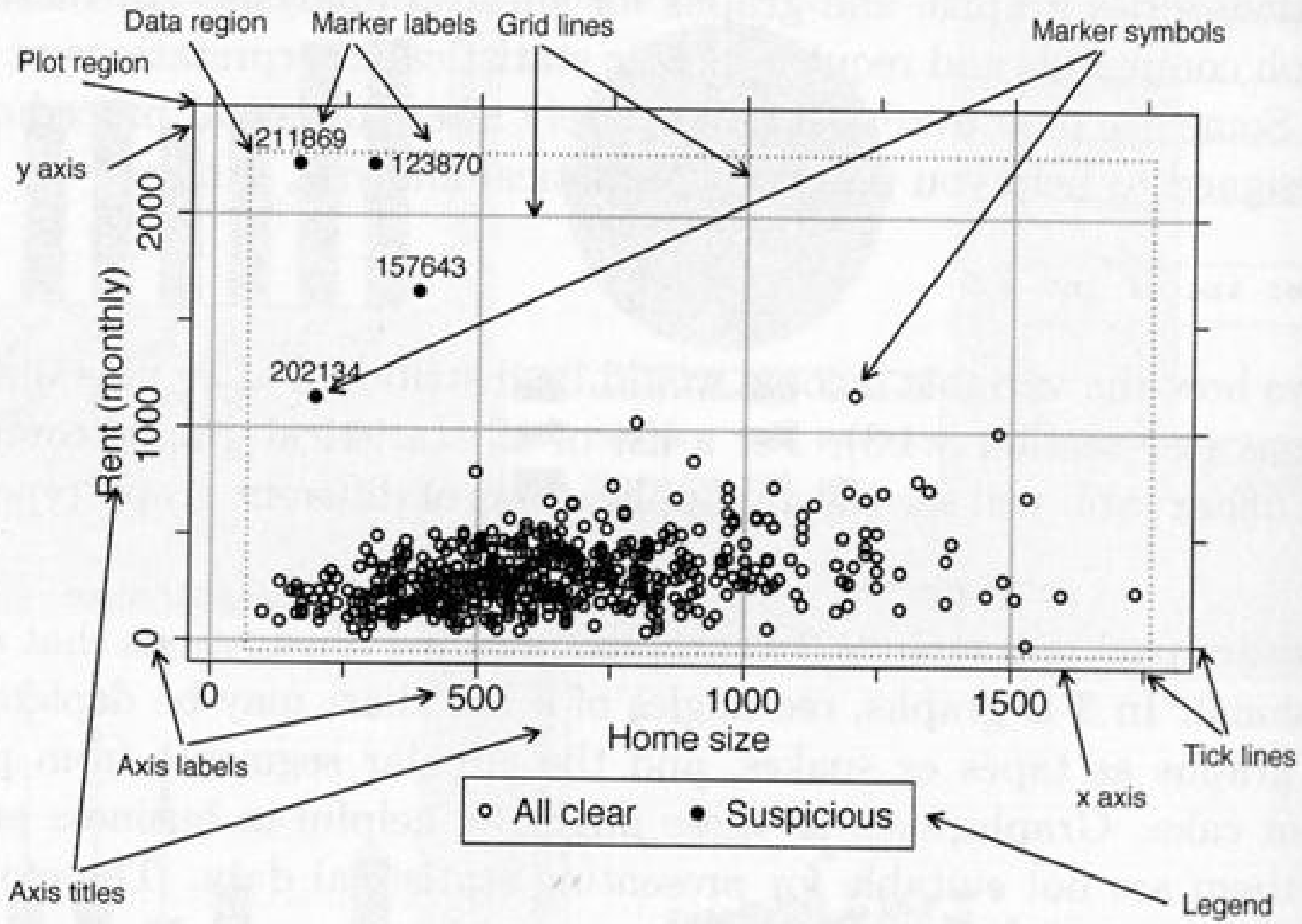


A little Stata graph theory

The appearance of graphs is defined by a series of elements

1. Elements that control the display of the data, including the shape, colour, and size of the “marker symbols”, as well as lines, bars and other ways to display data
2. Elements that control the size and shape of the graph
3. Elements which convey additional information within the graph region
e.g. marker symbol labels
4. Information outside the plot region e.g. axis labels

see Kohler and Kreuter (2009 pp.107-108).



Kohler and Kreuter (209 p.108)

/**

AQMeN Applied Quantitative Methods Network

Graphs and Graphical Representation - |
Getting a bit more out of Stata graphics (Summer 2011)

An AQMeN Stirling Local Group Activity

Professor Vernon Gayle, School of Applied Social Science, University of Stirling

```
*****  
* IT IS IMPORTANT THAT YOU READ THIS HANDOUT *  
* AND FOLLOW THE STATA.DO FILE LINE BY LINE! *  
*****
```

A picture paints a thousand words!

In my experience, with the exception of weighting, graphing data is one of the most troublesome aspects of data analysis.

Here is a brief introduction to graphs in Stata and getting a little more from your graphical outputs.

Please be patient. Computers often go wrong.

```
*****
*
*           Setting Up Stata
*
*****

* Clear the memory *

clear

/** More causes Stata to display --more-- and pause until any key is pressed.
    It is usually more convenient to have this function switched off **/

set more off

* change the amount of memory allocated to Stata *

set memory 64M

* keep a log file containing your output *

* close any log files already running *

capture log close

* change working directory *

cd c:\temp
pwd

* start a log file *

capture log using c:\temp\graphs_log_v1.txt, replace text

* specify version of Stata *

* version 10
```

Some very simple graphs

- Box graph
- Pie chart
- Bar chart
- Histogram

graph box

* the famous auto dataset *

```
sysuse auto.dta, clear
```

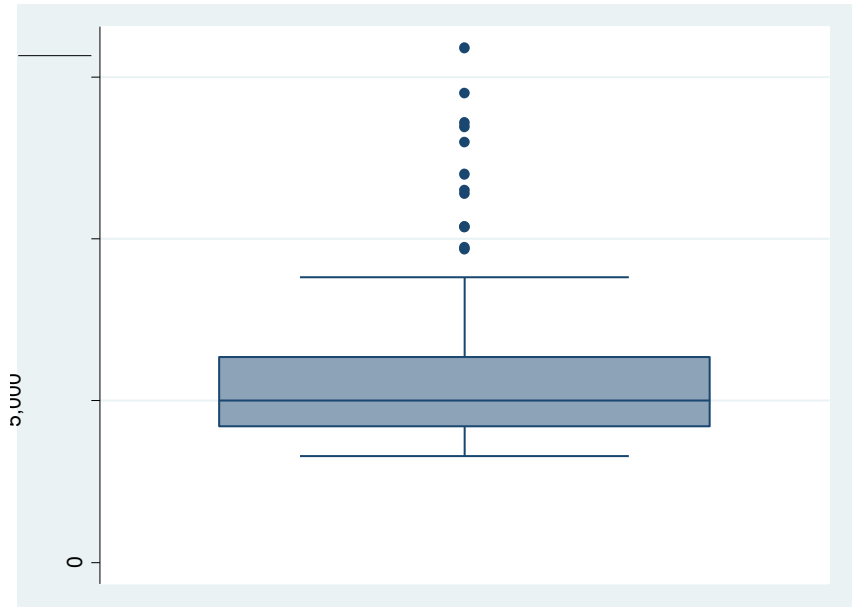
* add number labels *

```
numlabel _all, add
```

```
summarize
```

* A simple box graph *

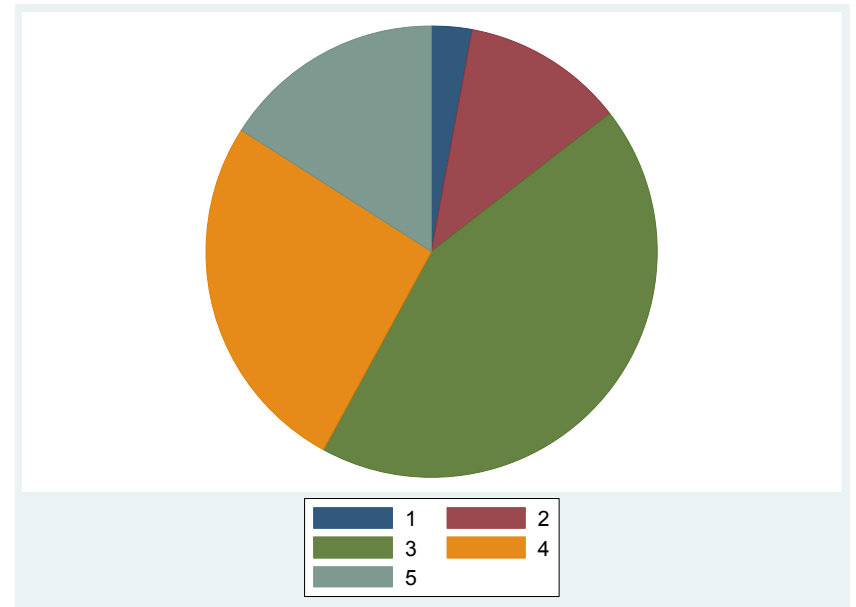
```
graph box price
```



graph pie

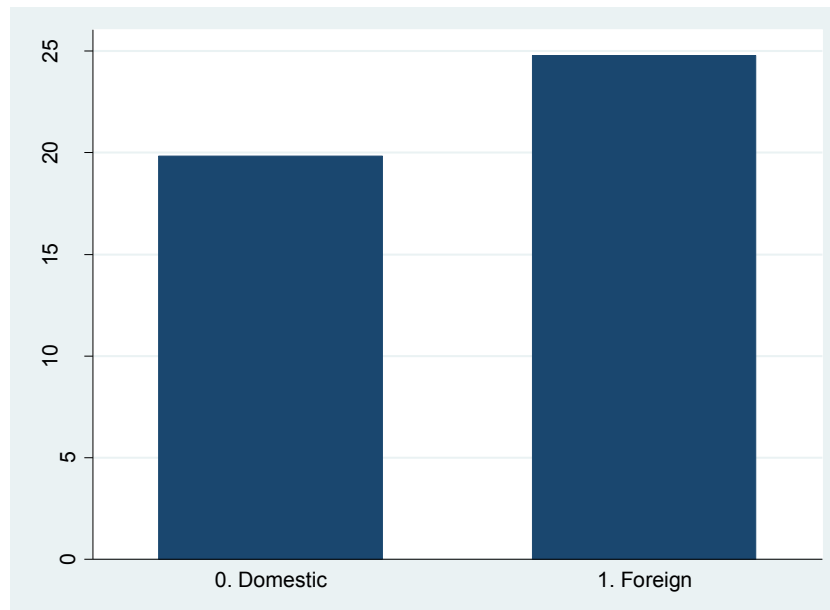
* A simple pie chart *

graph pie, over(rep78)



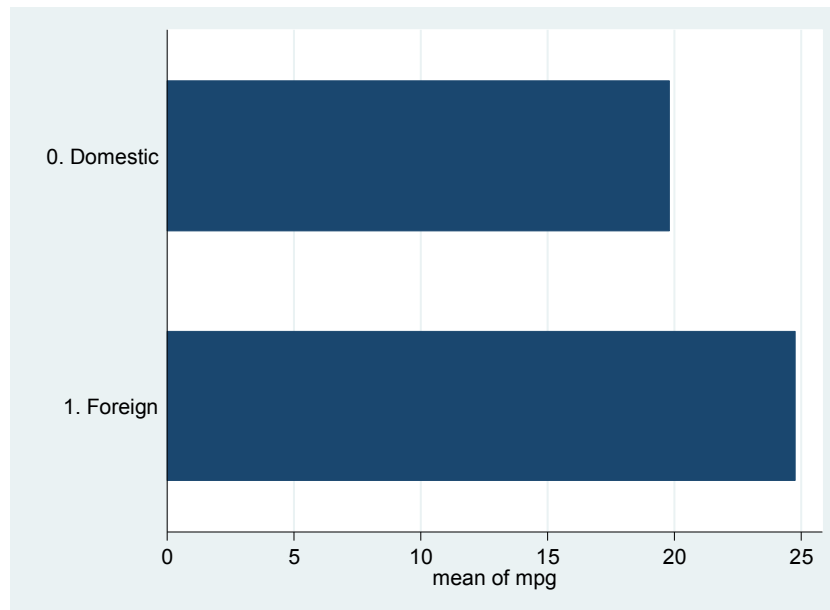
graph bar

graph bar (mean)mpg, over(foreign)



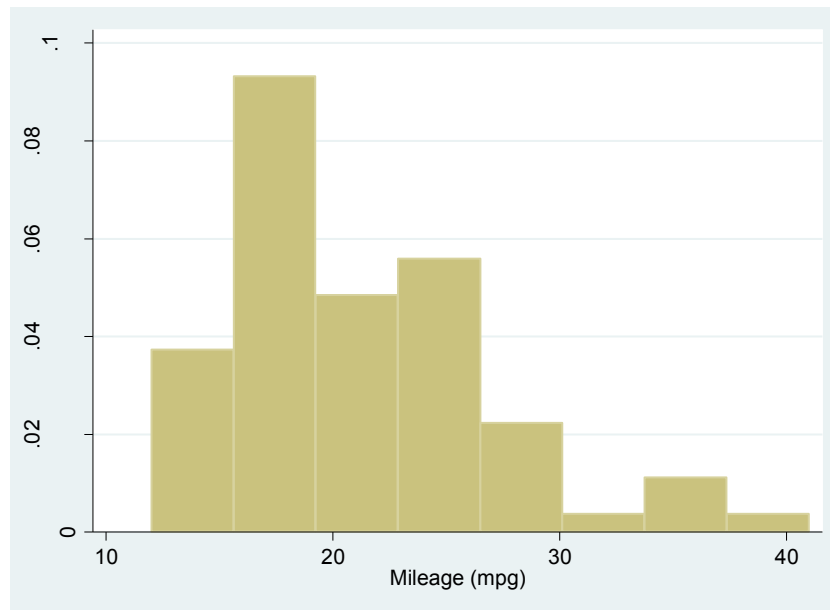
graph hbar

graph hbar (mean)mpg, over(foreign)



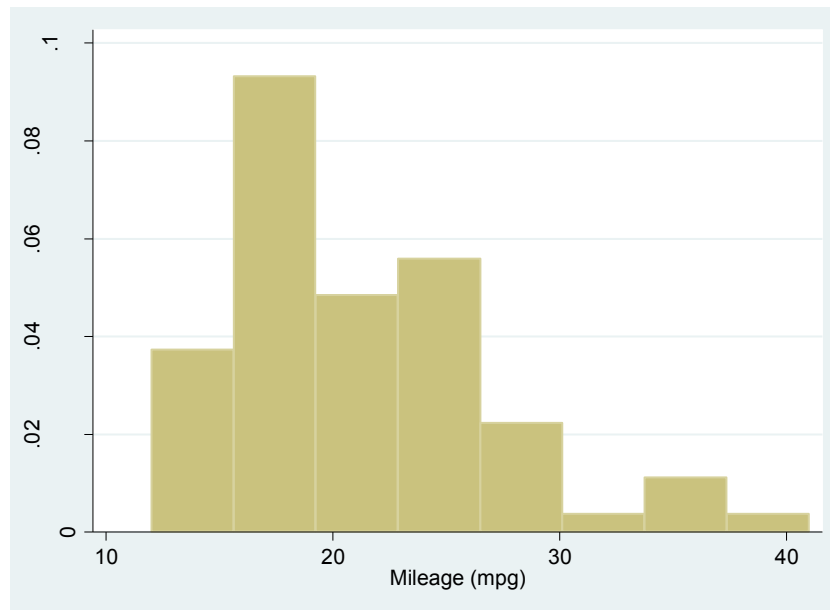
histogram

histogram mpg



histogram

histogram mpg



Naming and combining graphs

* naming graphs *

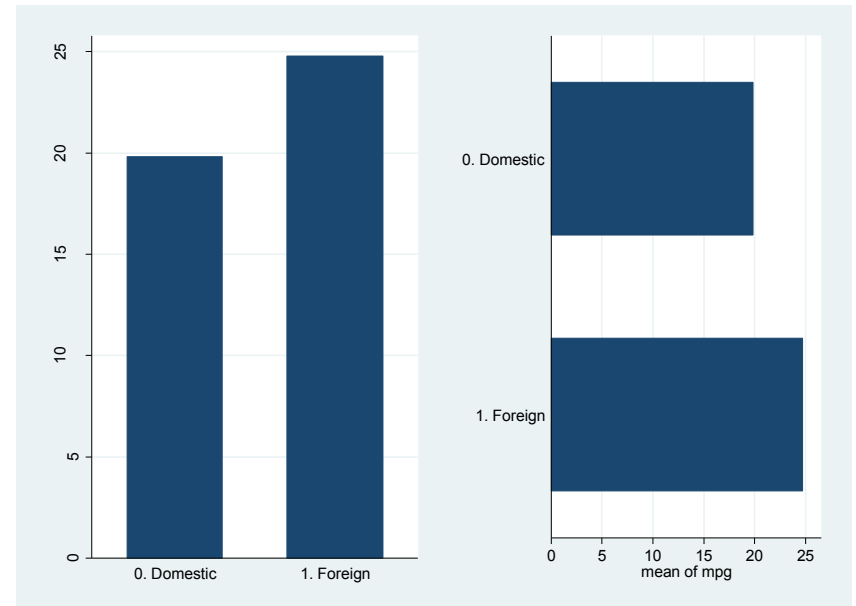
* graph g1 and graph g2 *

```
graph bar (mean)mpg,  
  over(foreign) name(g1, replace)
```

```
graph hbar (mean)mpg,  
  over(foreign) name(g2, replace)
```

* combining graphs in outputs *

```
graph combine g1 g2
```



Saving graphs

Stata keeps track of the last graph you have drawn, which is stored in memory, and calls it "Graph"

* displaying a graph that is in memory *

graph display g1

* saving and retrieving a graph *

graph save "c:\temp\g1", replace

graph use "c:\temp\g1"

Some more useful graphs

The Program Effort Data

Here are the famous program effort data from Mauldin and Berelson

This extract consists of observations on an index of social setting, an index of family planning effort, and the percent decline in the crude birth rate (CBR) between 1965 and 1975, for 20 countries in Latin America

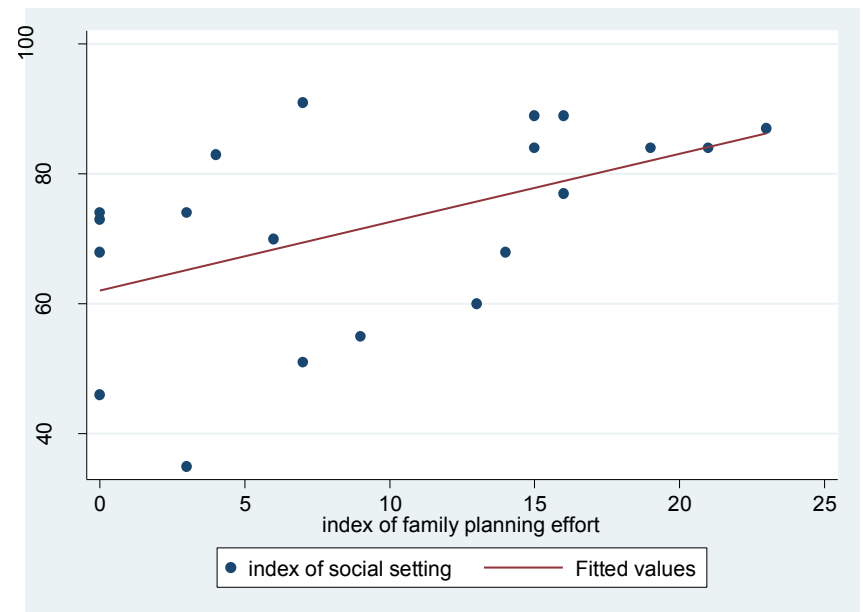
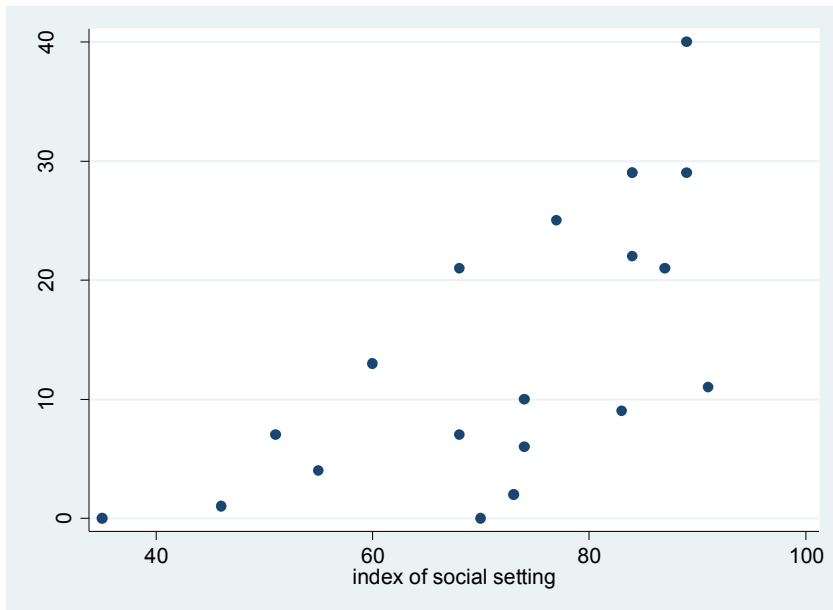
Reference:

P.W. Mauldin and B. Berelson (1978) Conditions of fertility decline in developing countries, 1965-75 *Studies in Family Planning*, 9:89-147 JSTOR: <http://www.jstor.org/stable/1965523>

German Rodriguez's site <http://data.princeton.edu/stata/graphics.html>

graph twoway scatter

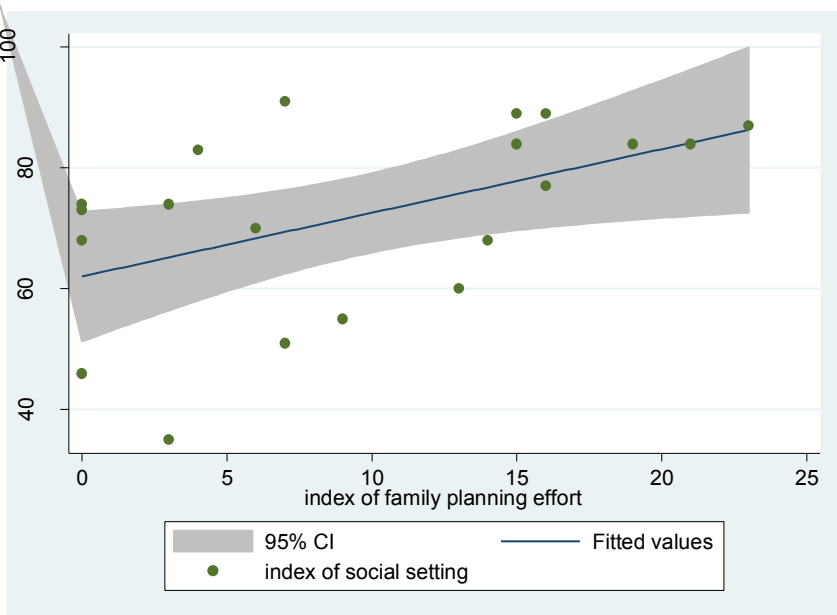
graph twoway scatter change setting



graph twoway (scatter setting effort) ///
(lfit setting effort)

graph twoway scatter

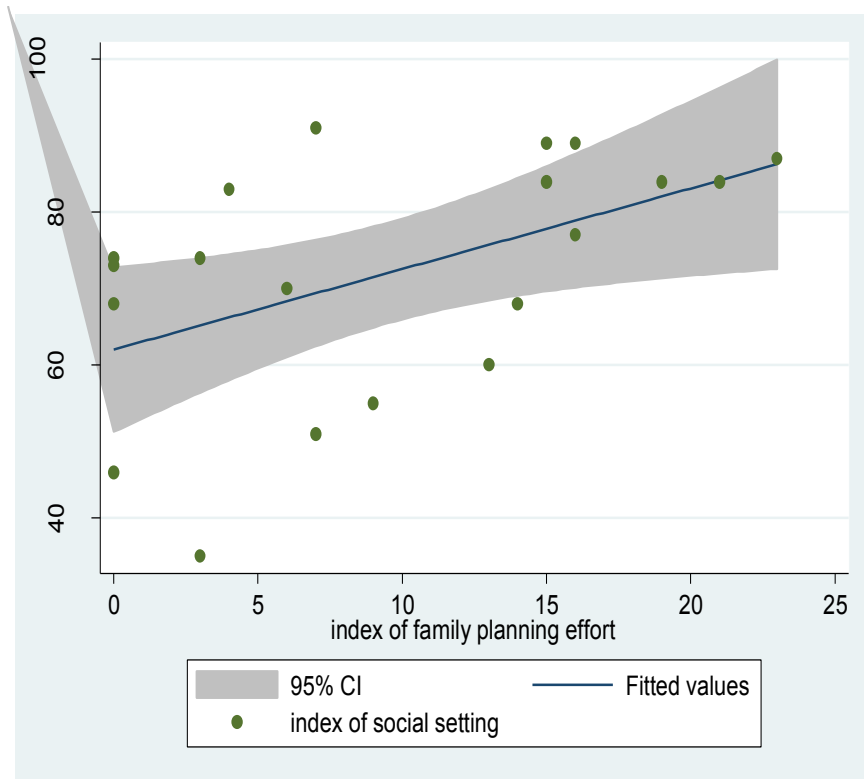
graph twoway (lfitci setting effort) ///
(scatter setting effort)



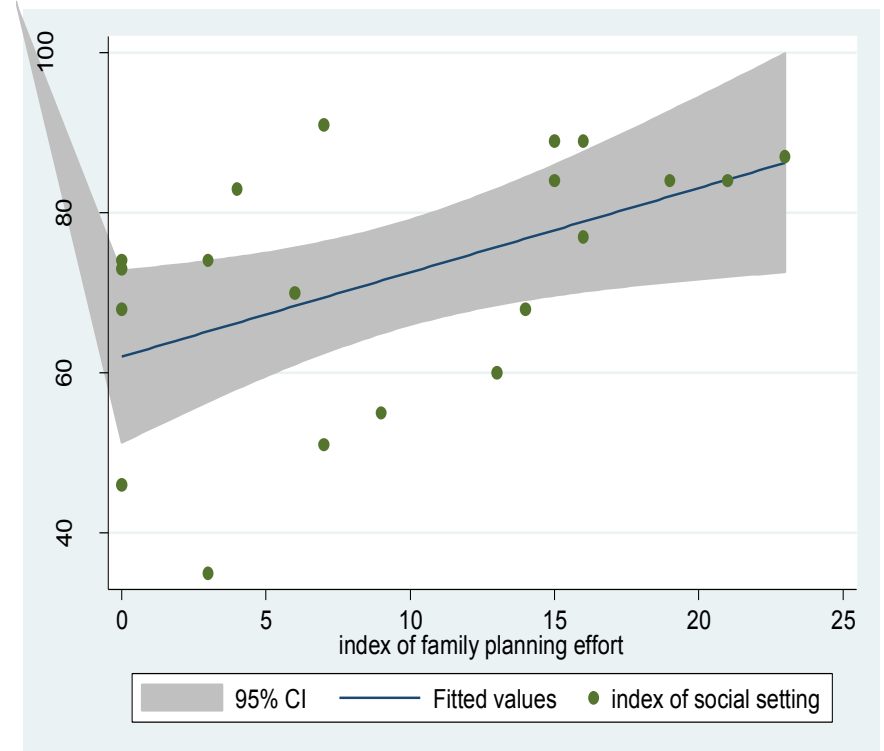
confidence region fitted as a gray band because the confidence band can obscure some points
we draw the region first and the points later

Axis and legends

```
graph twoway (lfitci setting effort) ///  
  (scatter setting effort) ///  
  , ytitle("Fertility Decline") legend(on)
```



```
graph twoway (lfitci setting effort) ///  
  (scatter setting effort) ///  
  , ytitle("Fertility Decline") legend(row(1))
```

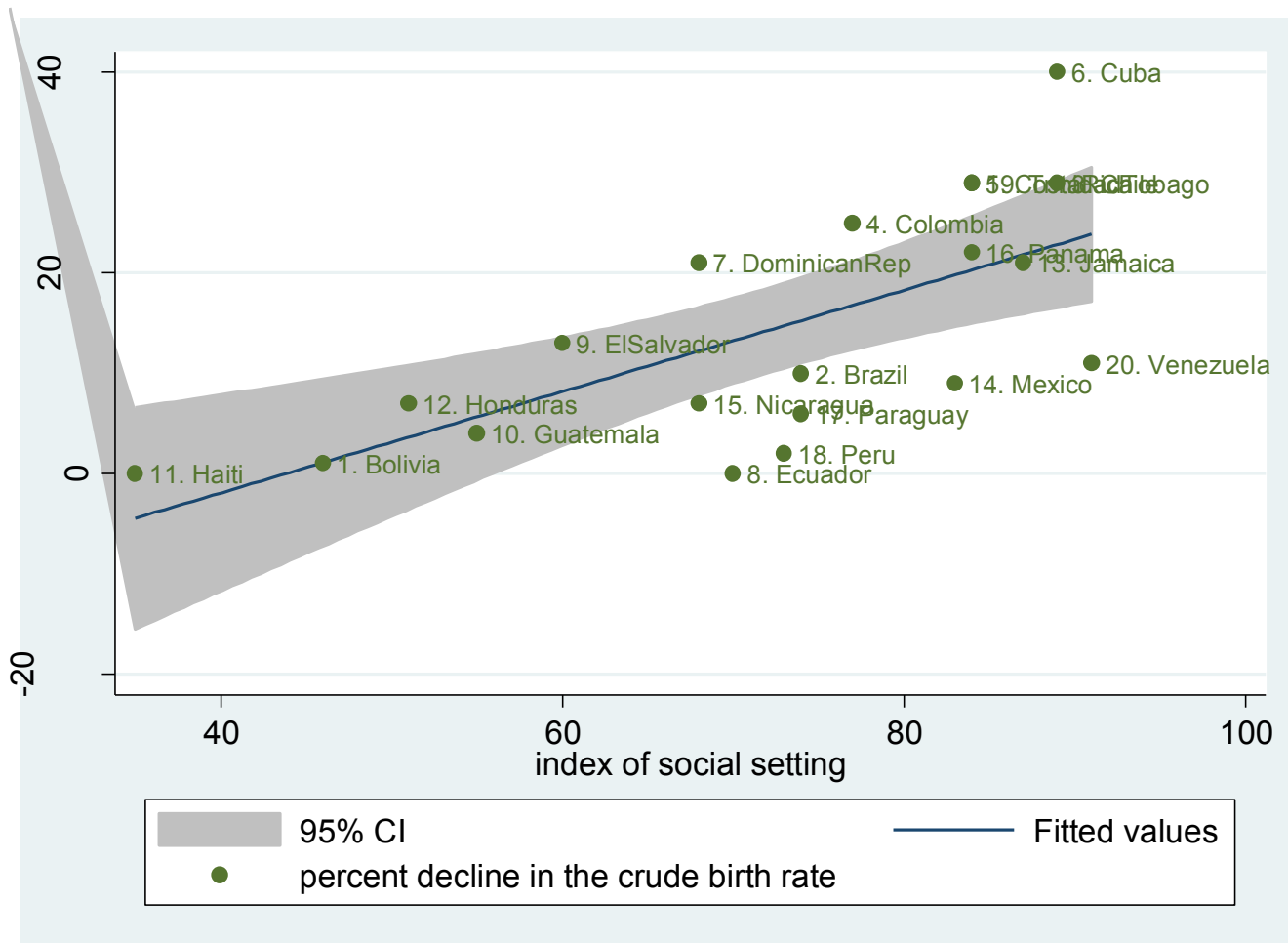


Labelling points

* this graph is too cluttered *

graph twoway (lfitci change setting) ///

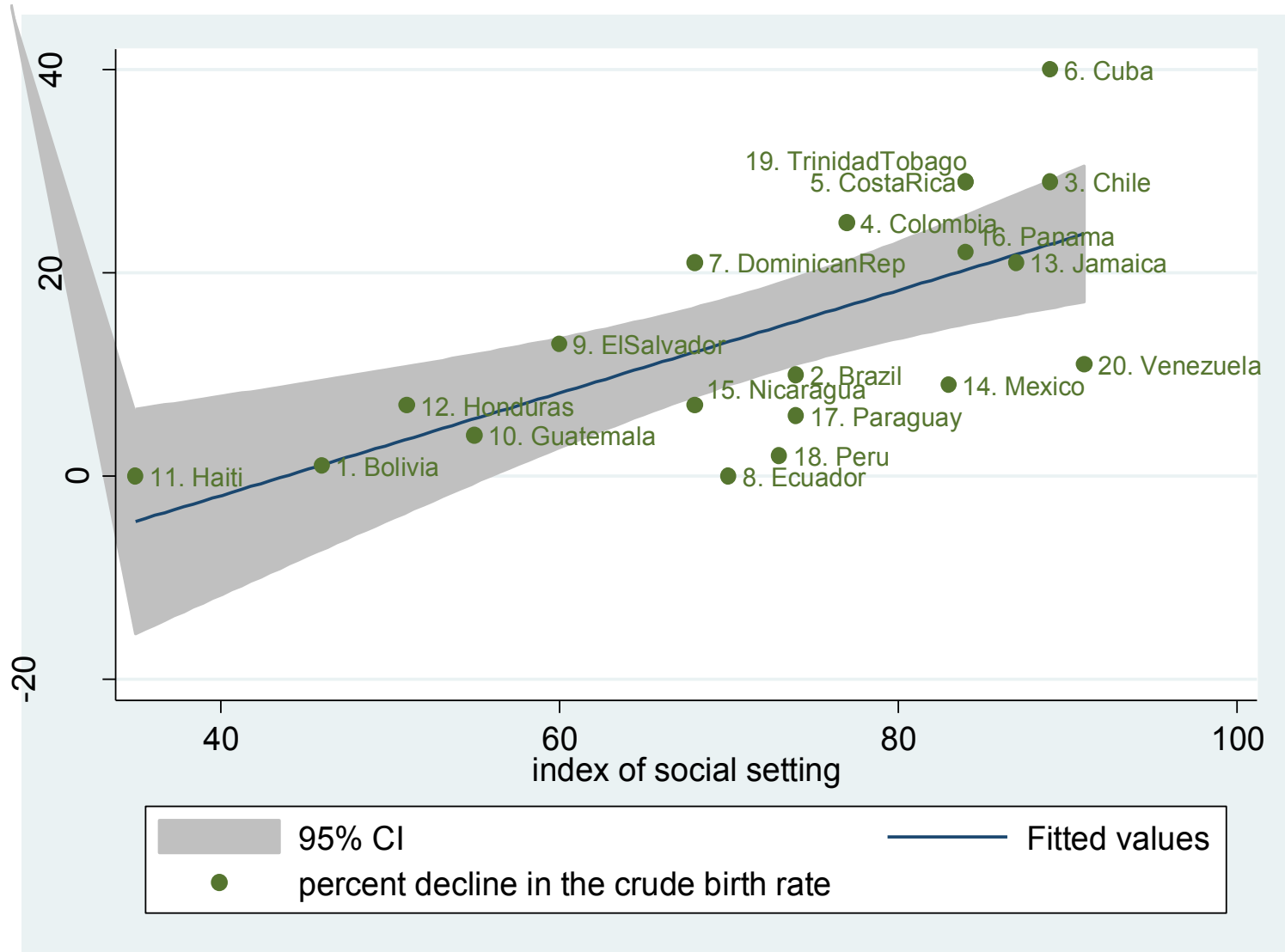
(scatter change setting, mlabel(country))



* changed position of some points with mlabv and a new variable position (clock) *

graph twoway (lfitci change setting) ///

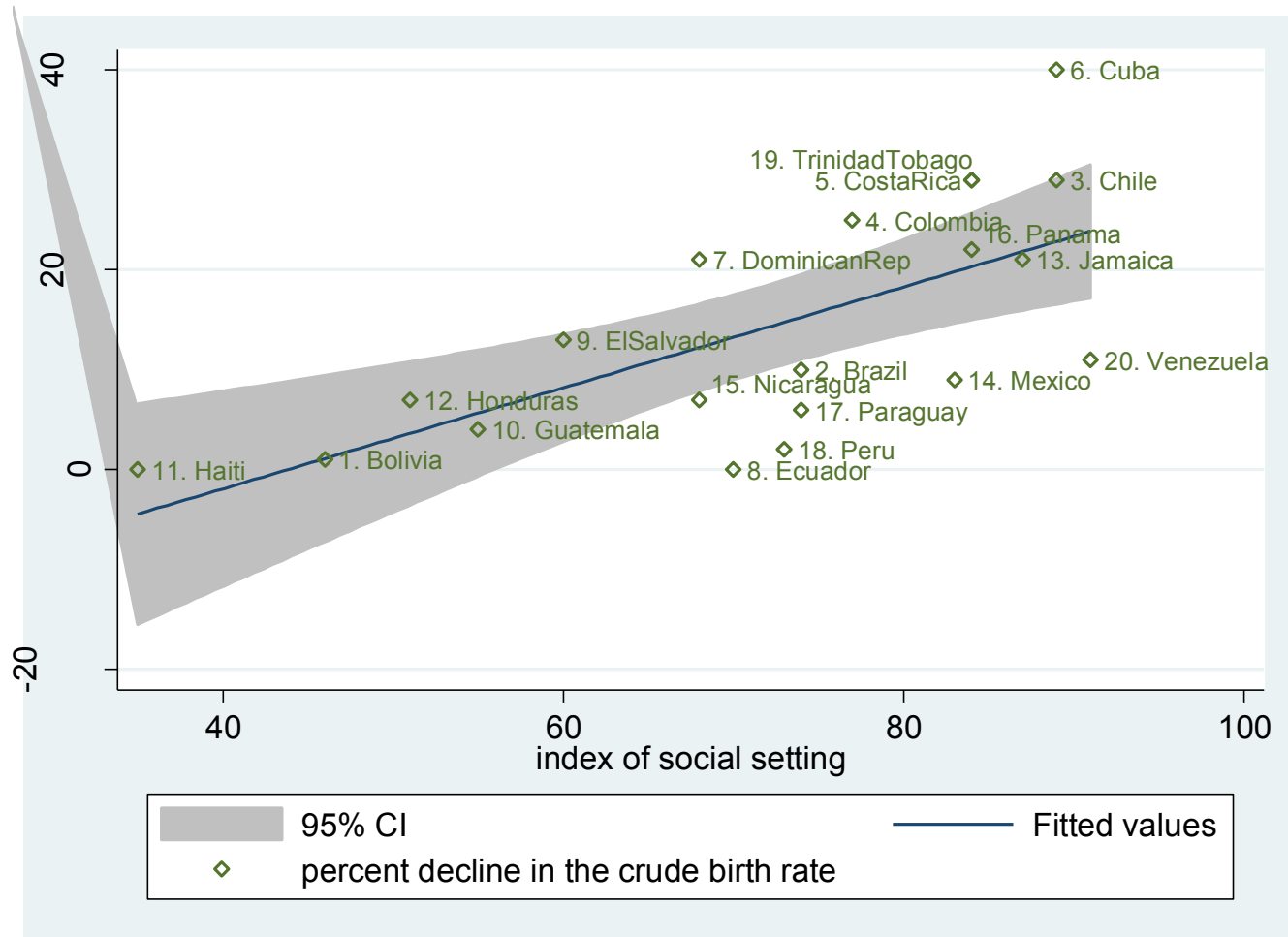
(scatter change setting, mlabel(country) mlabv(pos))



Graph symbol styles

graph twoway (lfitci change setting) ///

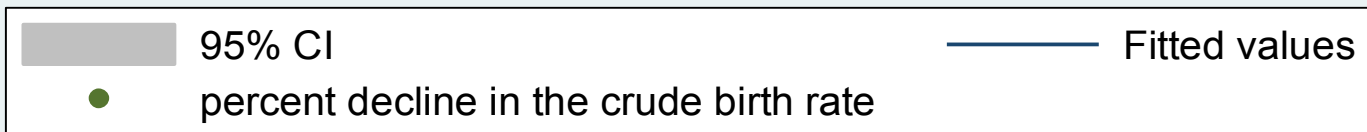
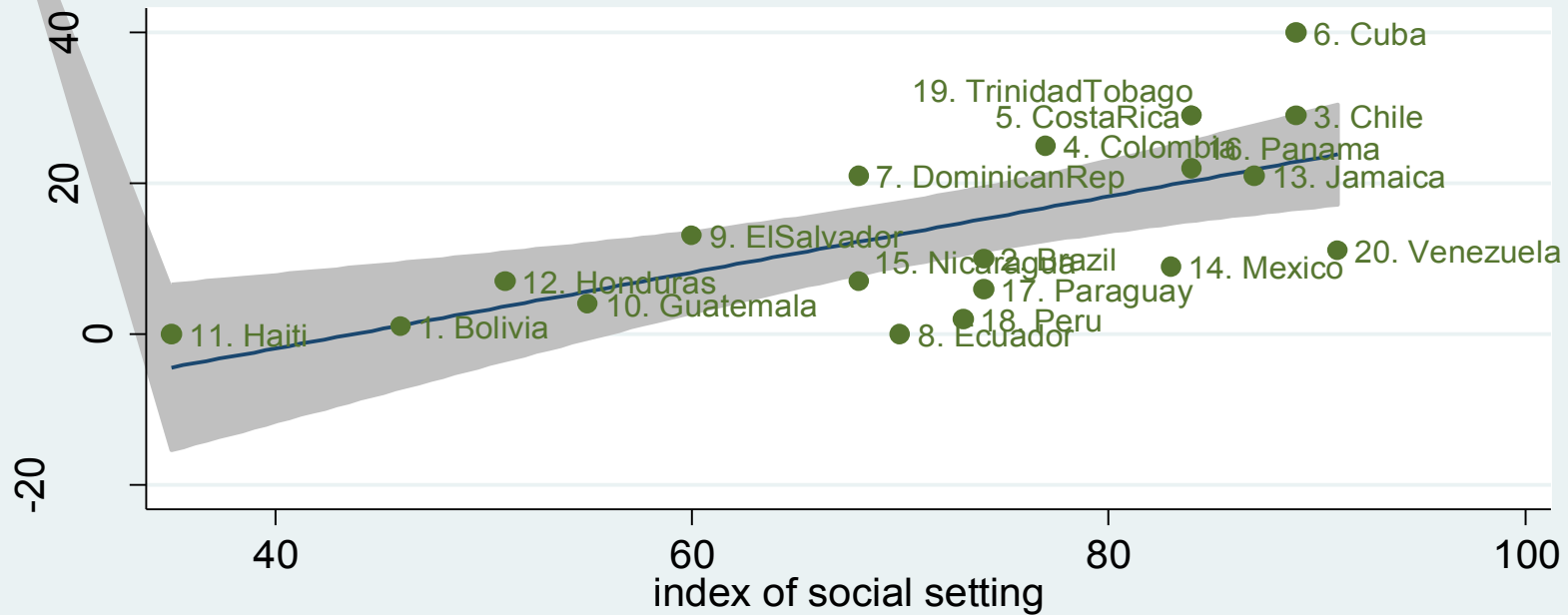
(scatter change setting, msymbol(smdiamond_hollow) mlabel(country) mlabv(pos))



Titles and captions

Fertility Decline by Social Setting

SUB TITLE

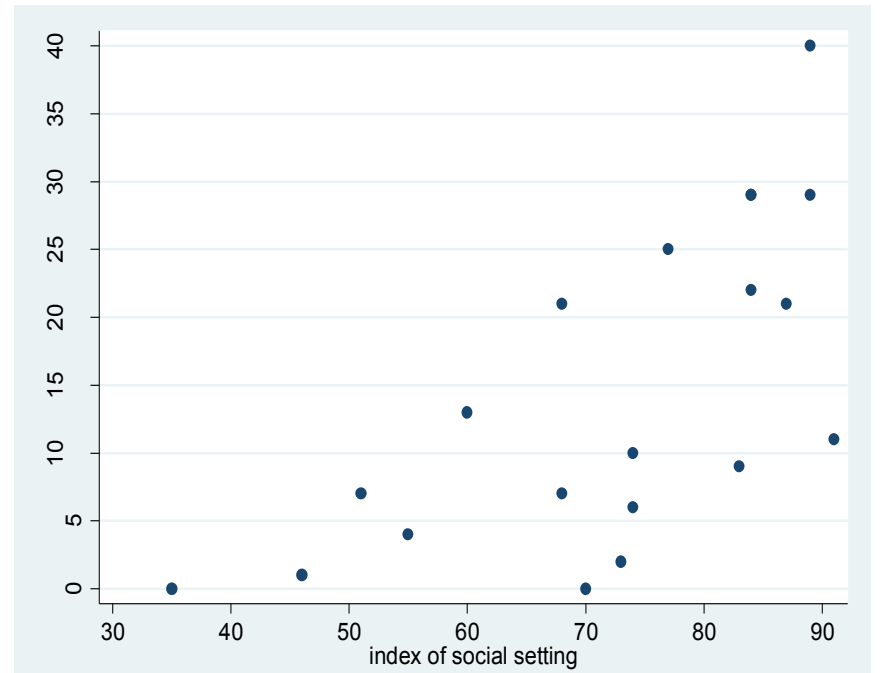
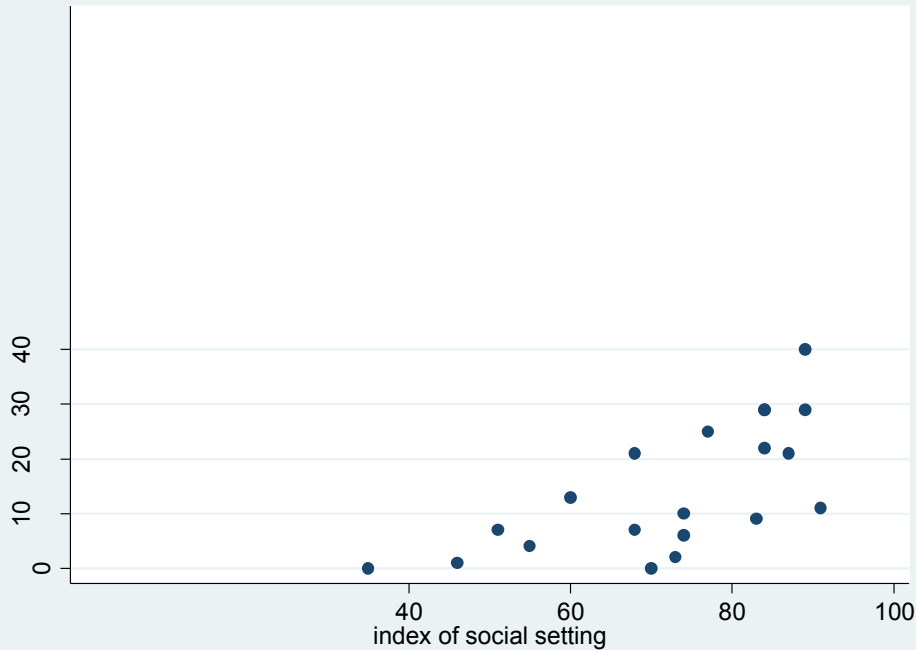
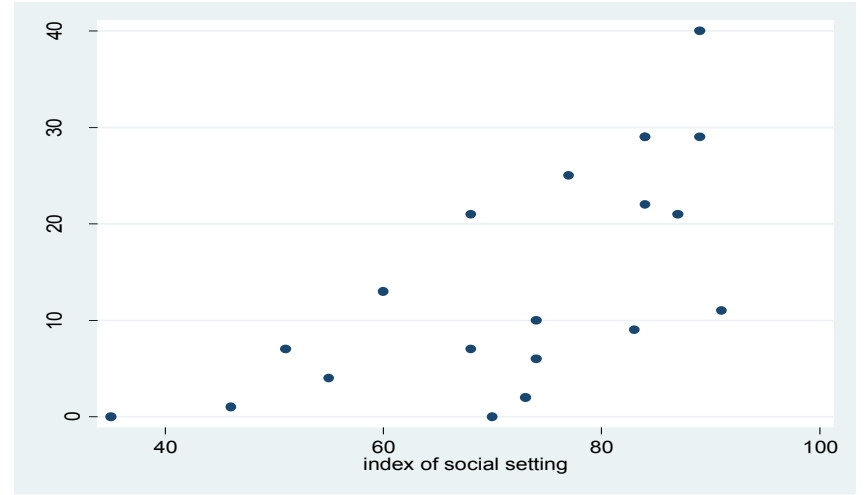
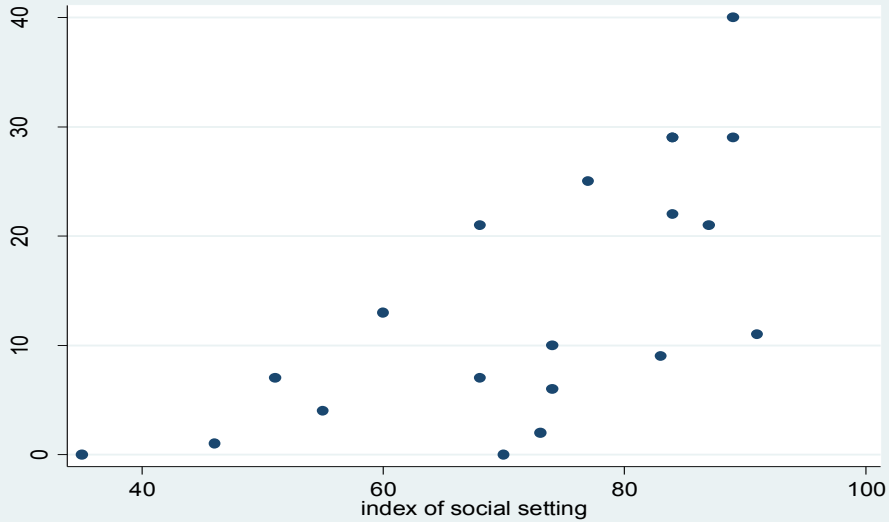


NOTE ABOUT GRAPH
caption goes here

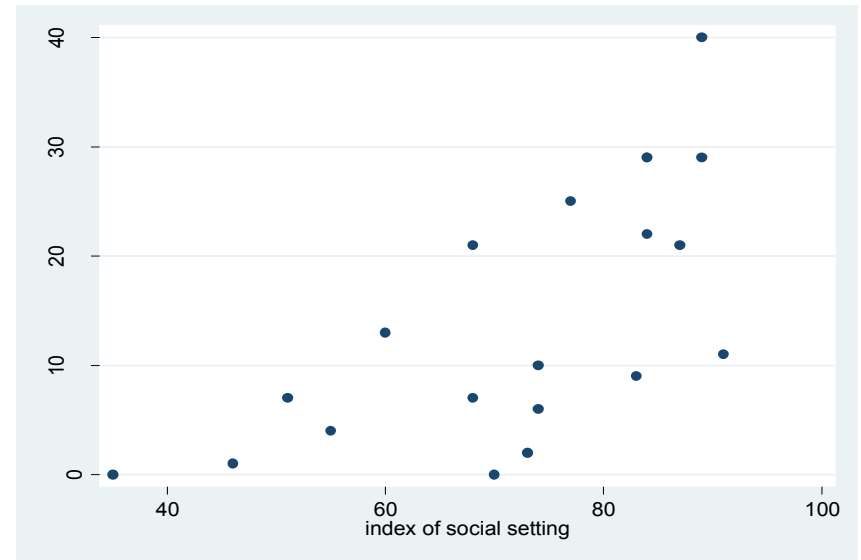
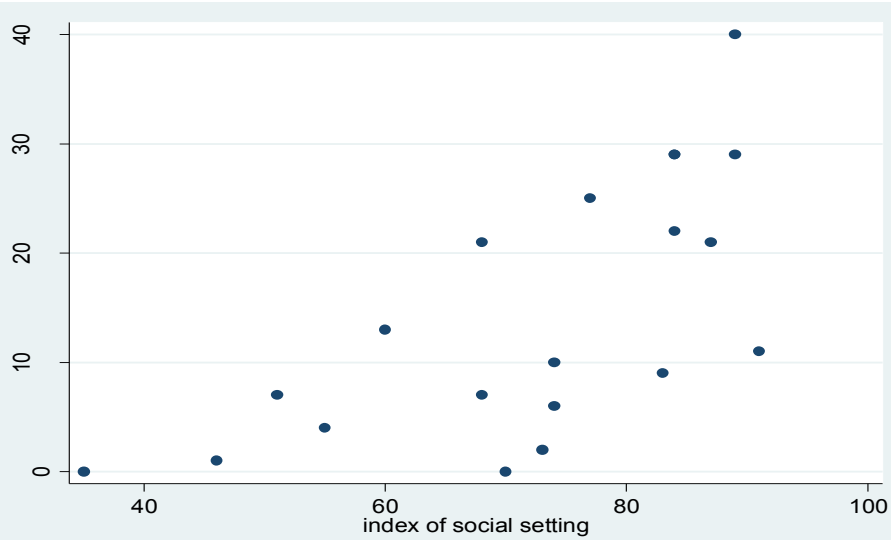
Titles and captions

```
graph twoway (lfitci change setting) ///
    (scatter change setting, mlabel(country)
    mlabv(pos) ) ///
    , title("Fertility Decline by Social Setting" " ")
    ///
    ytitle("Fertility Decline") ///
    subtitle("SUB TITLE") ///
    note("NOTE ABOUT GRAPH") ///
    caption("caption goes here")
```


More on axis

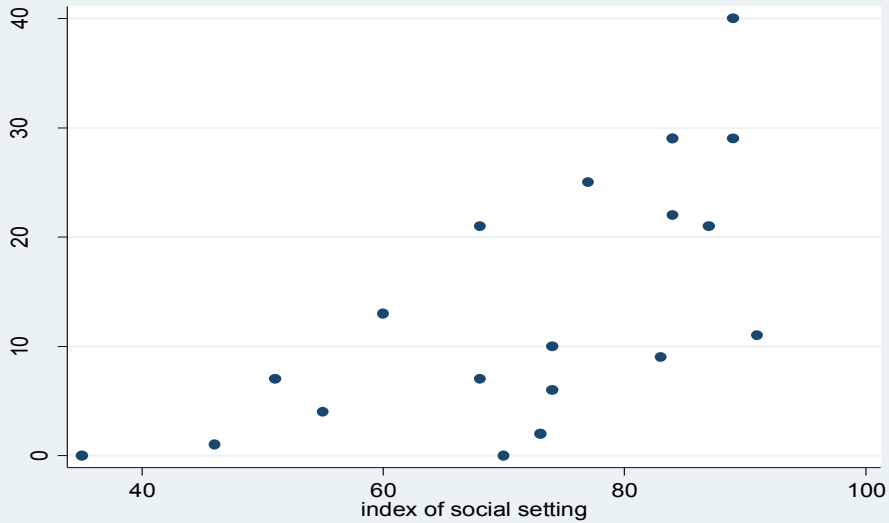


More on axis

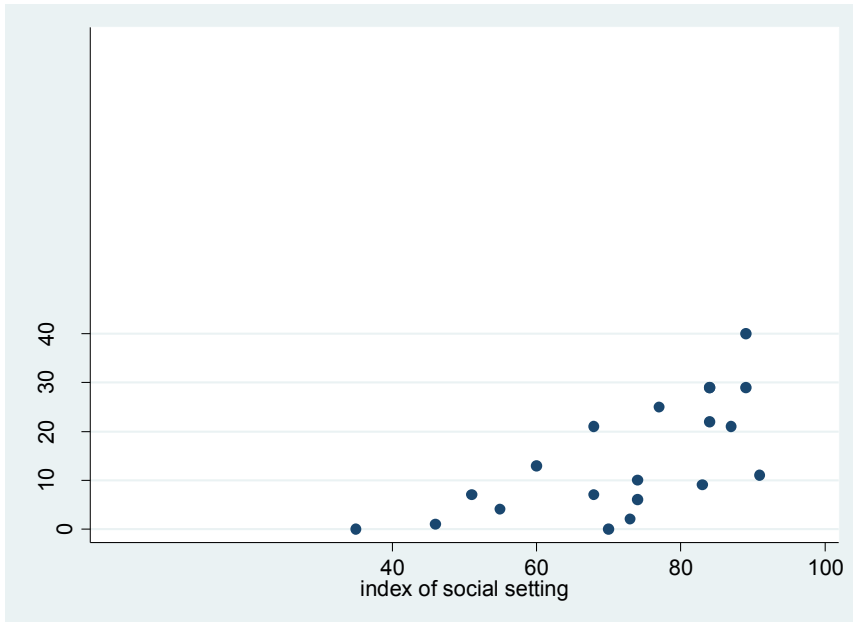


scatter change setting, yscale(noline) xscale(noline)

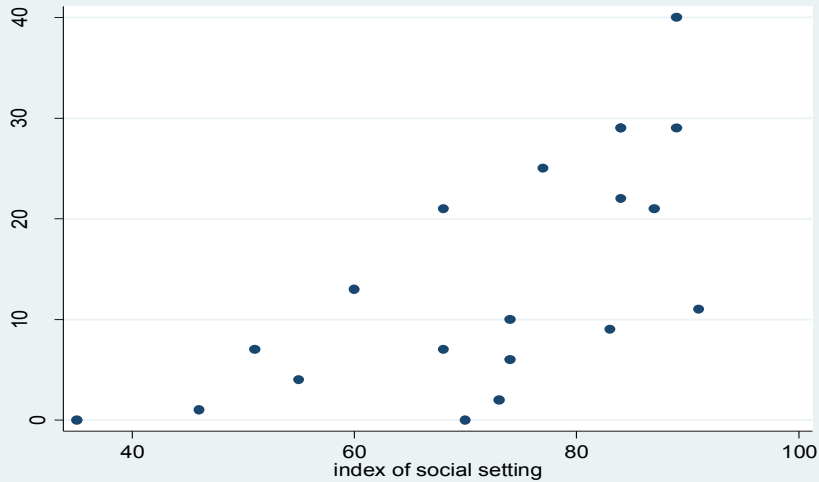
More on axis



scatter change setting, yscale(r(0 100)) xscale(r(0 100))

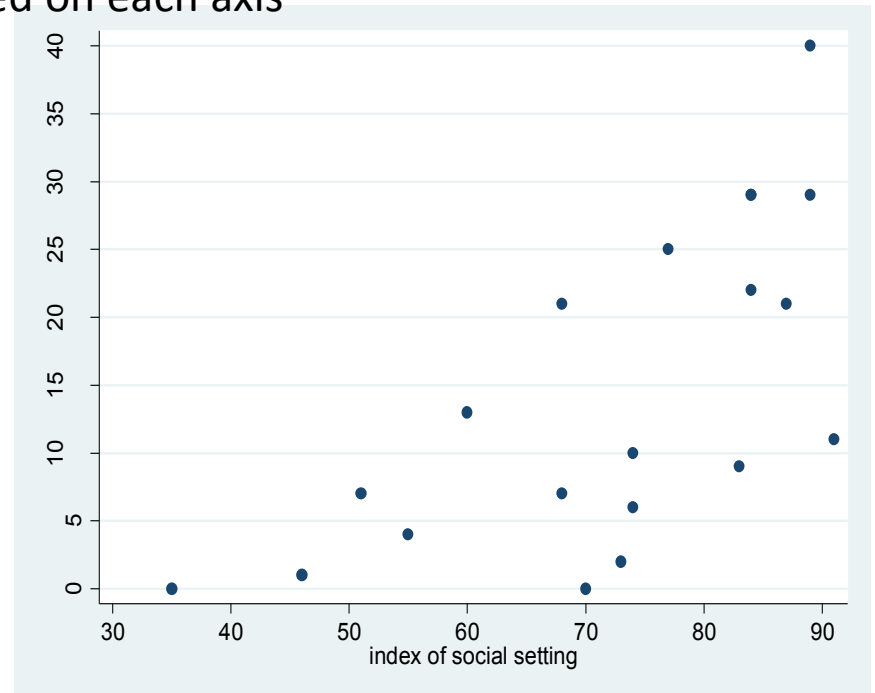


More on axis



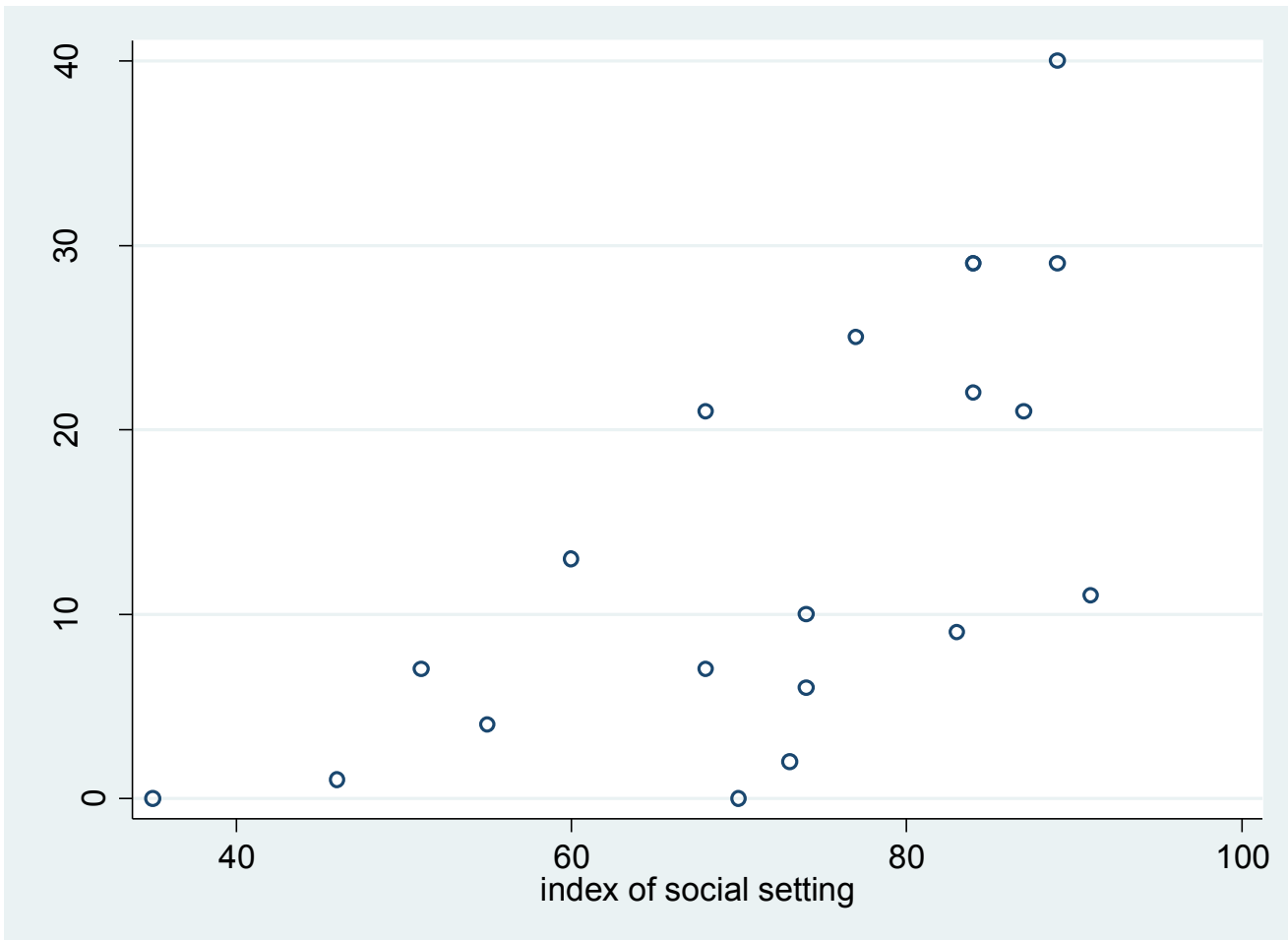
* the number of values that are labeled and ticked on each axis *

scatter change setting, ylabel(#10) xlabel(#10)



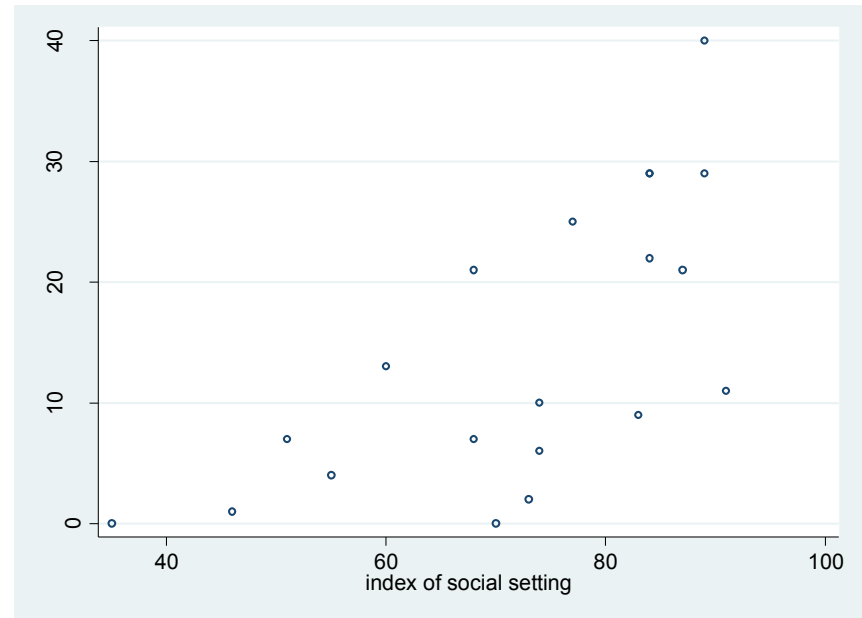
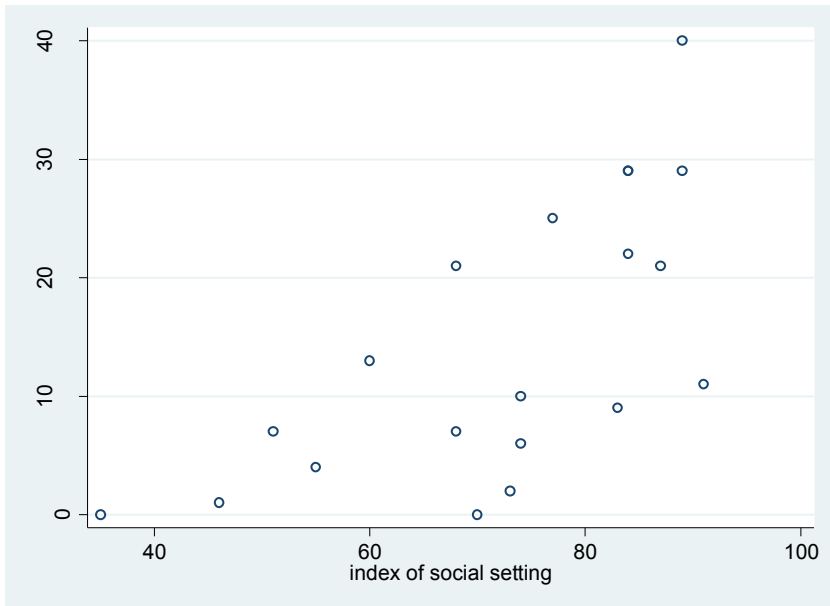
Markers

scatter change setting, msymbol(circle)



Markers

scatter change setting, `msymbol(smcircle_hollow)`



scatter change setting, `msymbol(circle_hollow)`

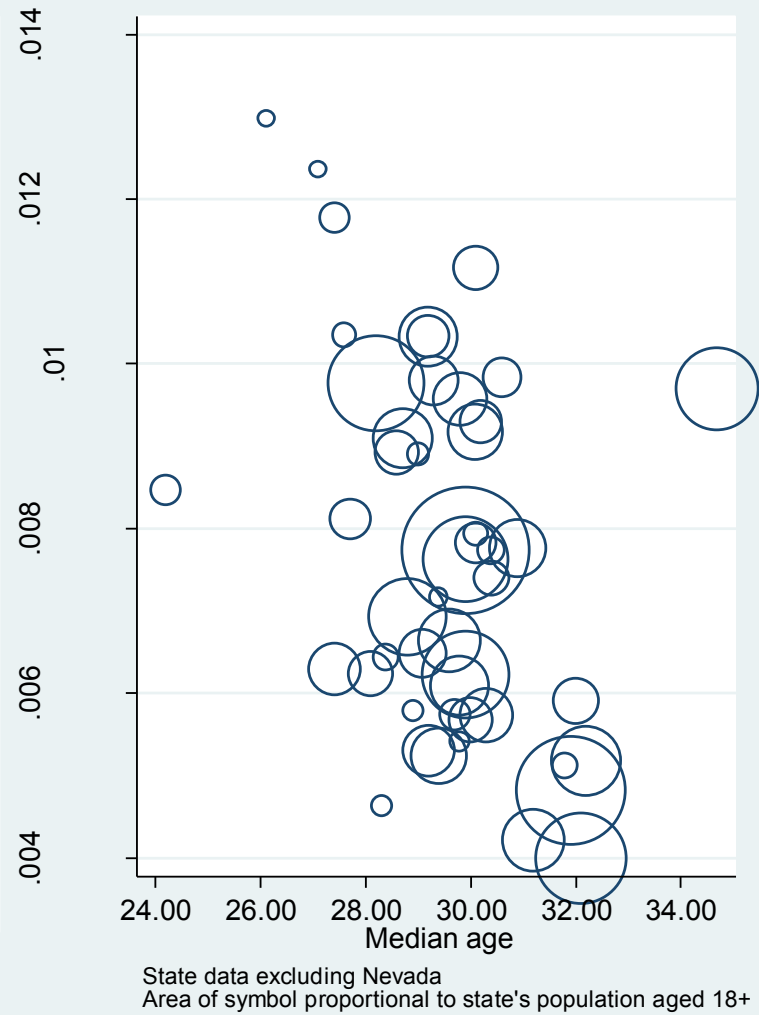
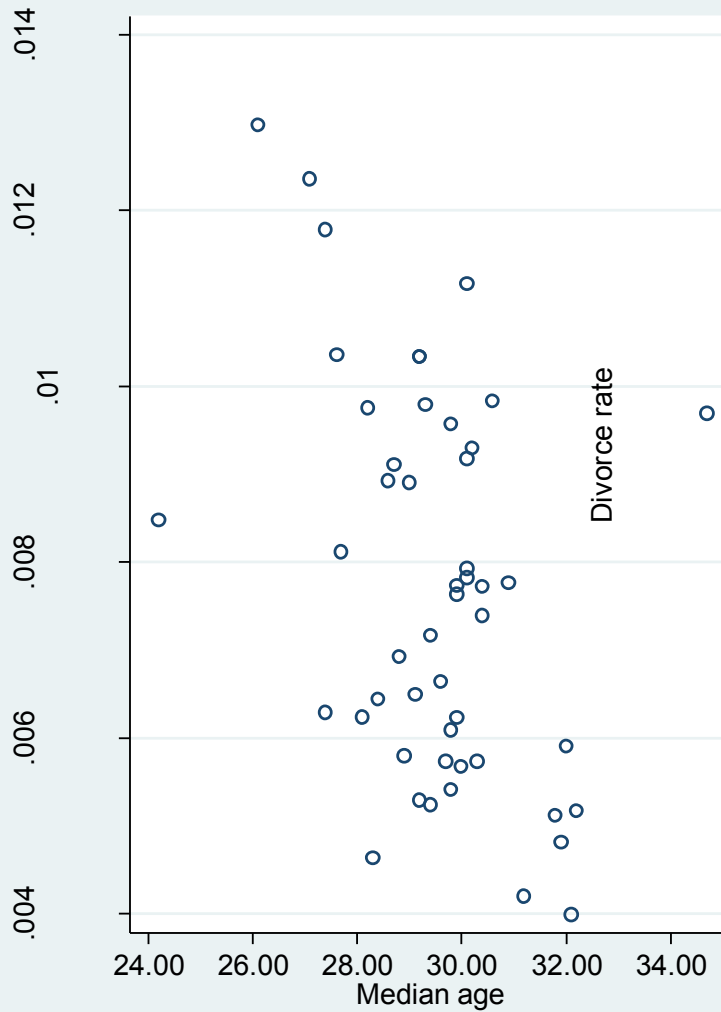
Markers

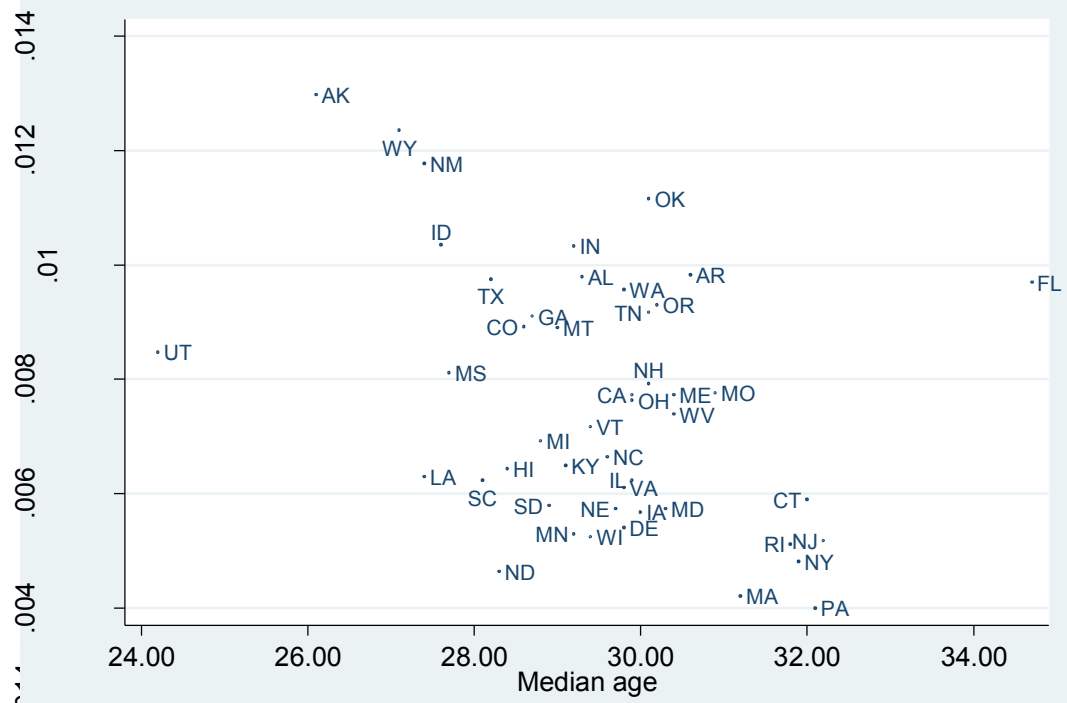
Weighted scatter plot

Example of US Census data

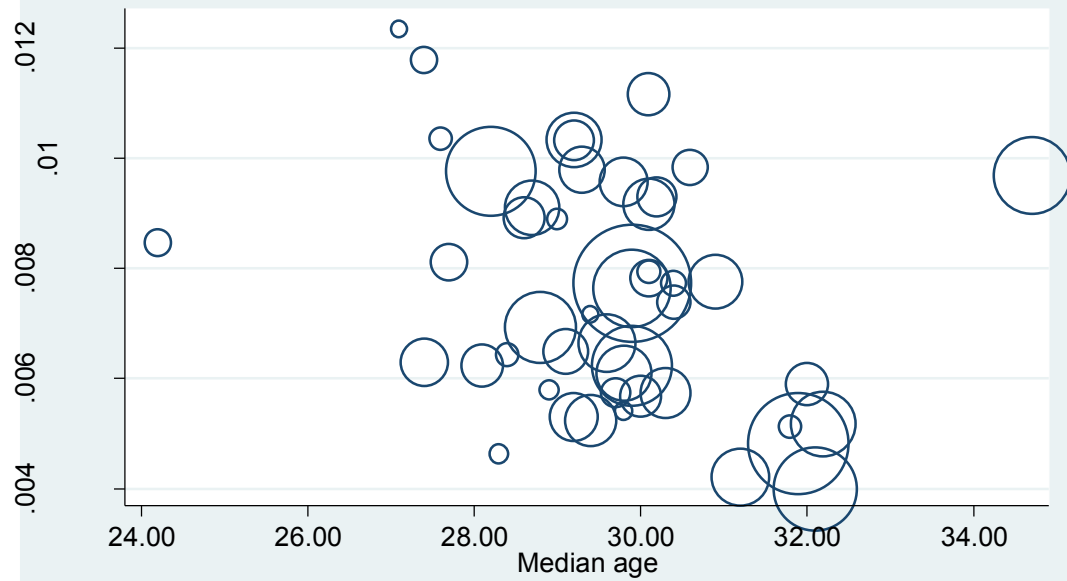
States divorce rate by mean age (except Nevada)

Markers





State data excluding Nevada



State data excluding Nevada
Area of symbol proportional to state's population aged 18+

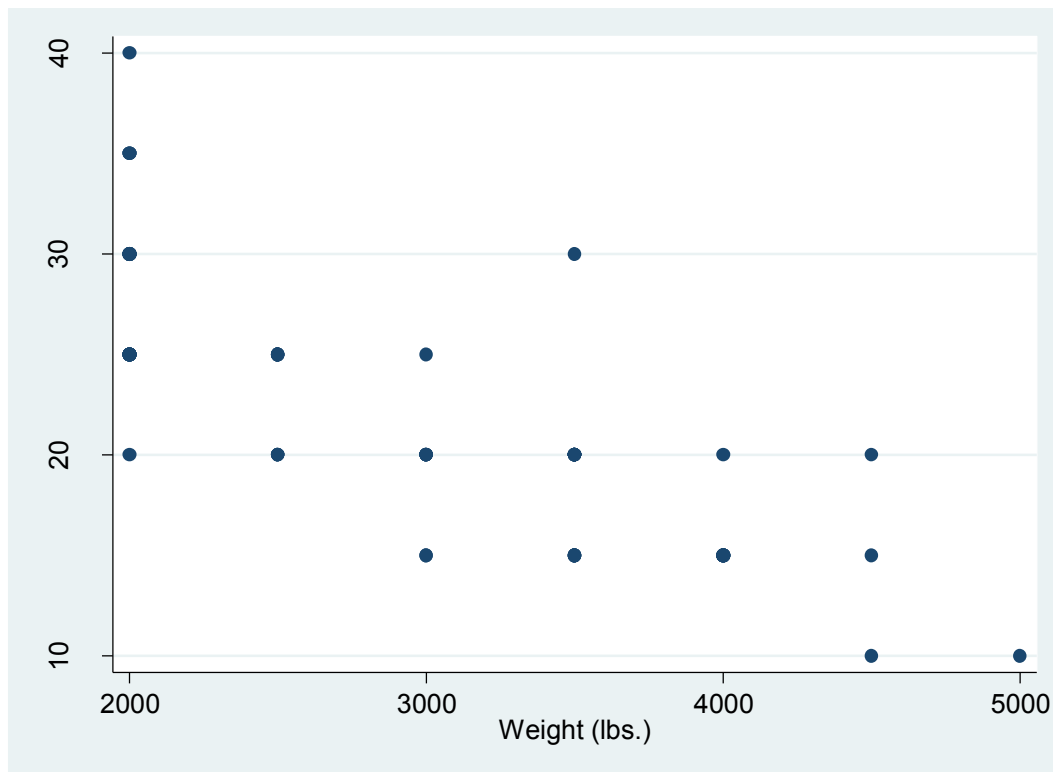
This is clever I will explain it a little more at the workshop

The graph is a little cluttered so I have cunningly dropped a couple of state labels and moved the position of others

```
graph twoway ///  
(scatter drate medage if pos!=., msymbol(hollow_circle)  
  mlabel(state2) mlabv(pos) msize(vtiny)) , ///  
note("State data excluding Nevada" ///  
"" ) ///  
  name(divorce5, replace)
```

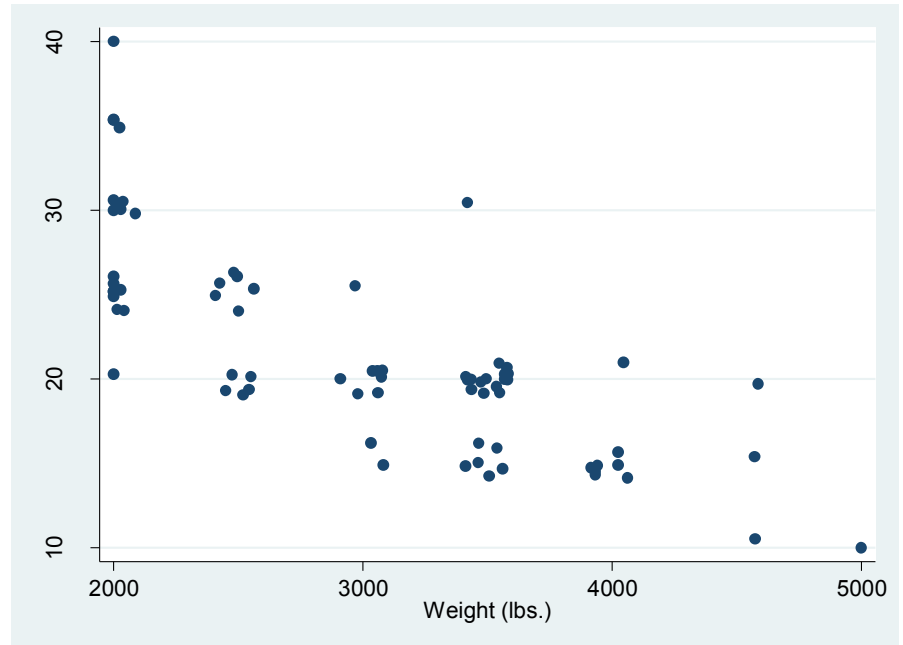
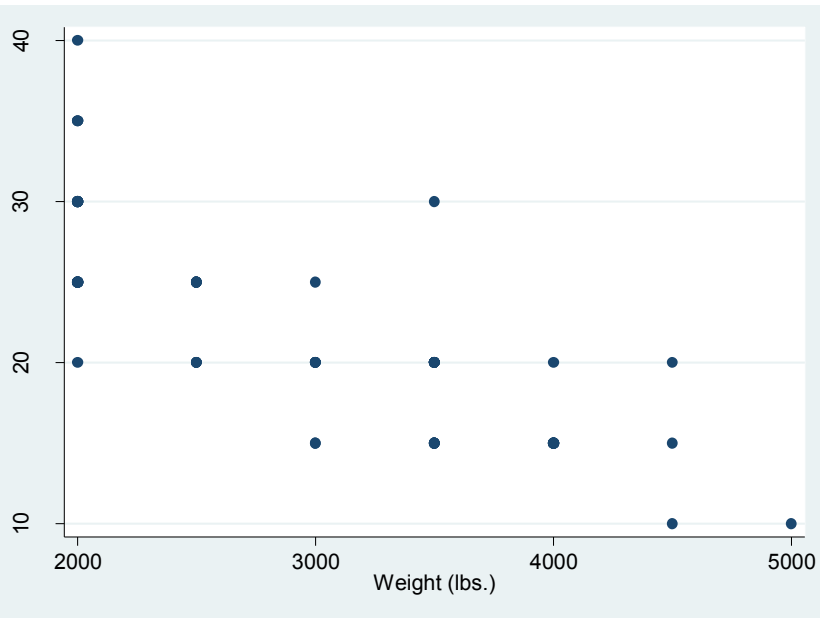
Jitter

There are 74 points in the graph, even though it appears because of overprinting as if there are only 19 - jittering solves this problem



Jitter

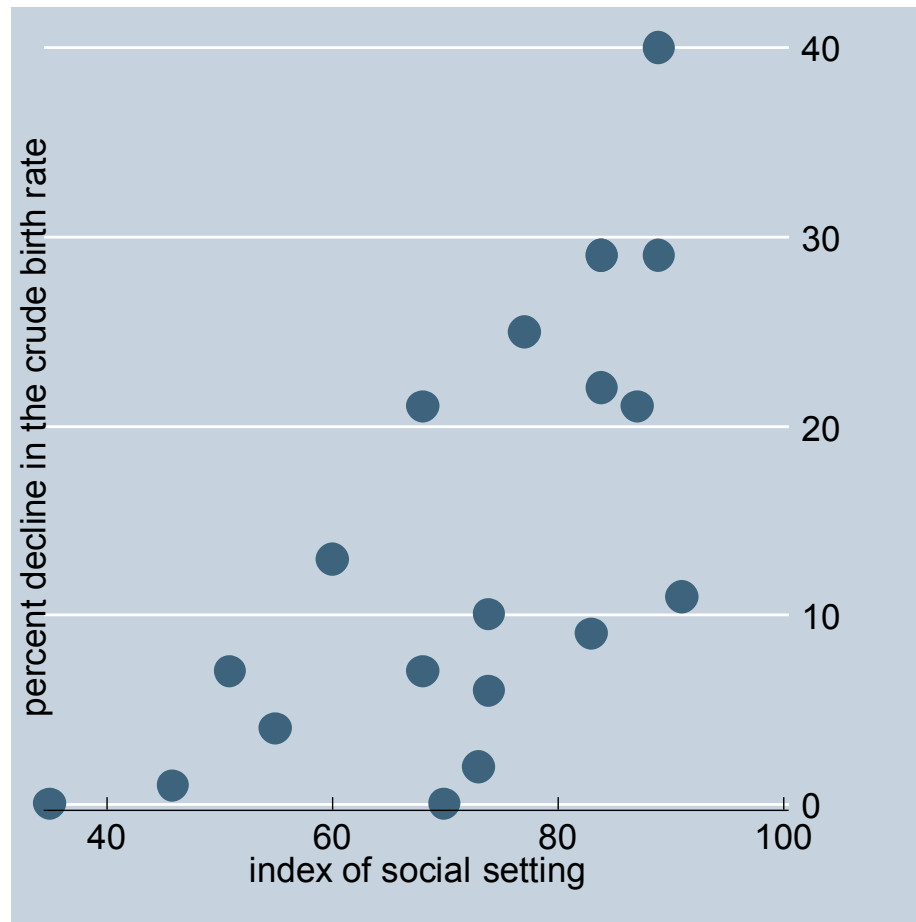
scatter mpg weight, jitter(7)



Spherical noise is added to the graph – more on this at the workshop

Graph schemes

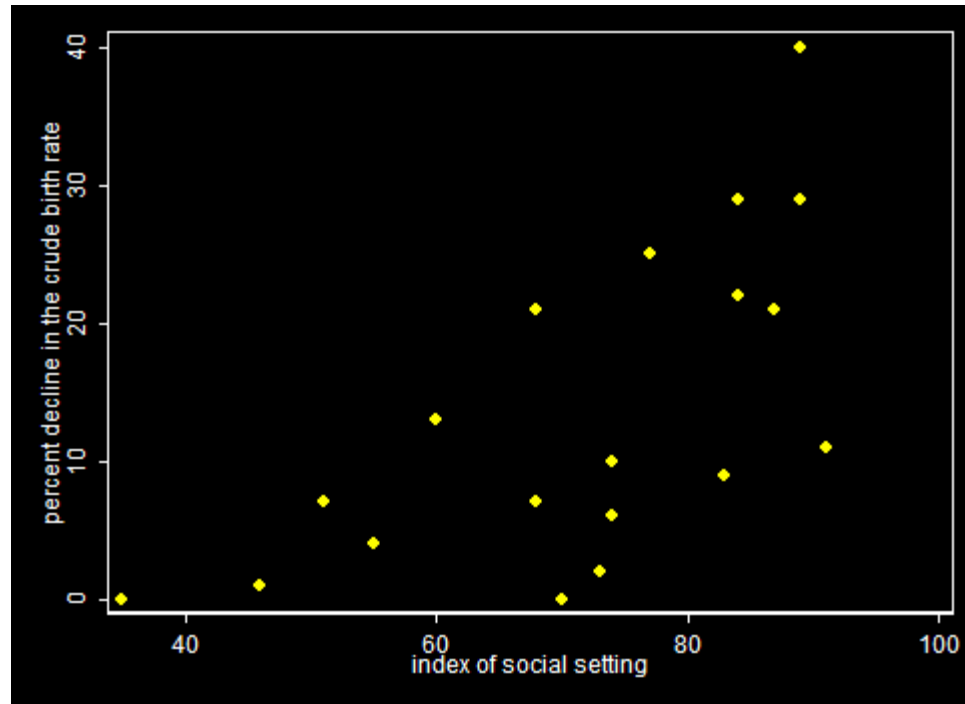
scatter change setting, scheme(economist)



Graph schemes

A slightly kinky scheme?

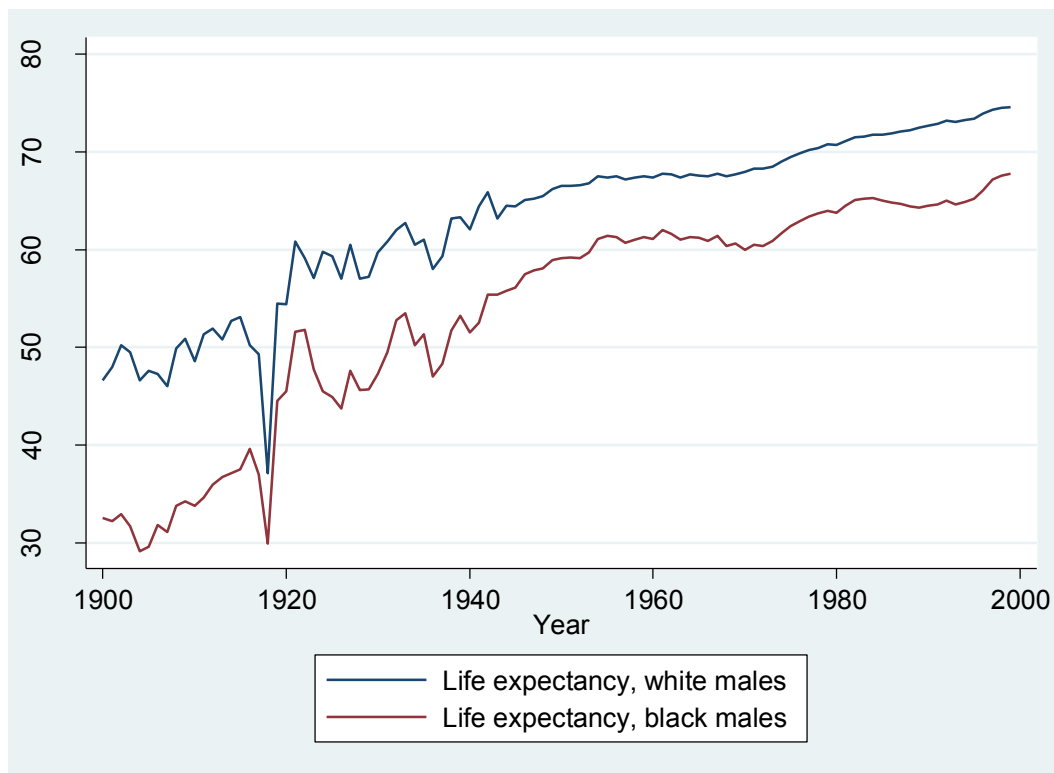
scatter change setting, scheme(s1rcolor)



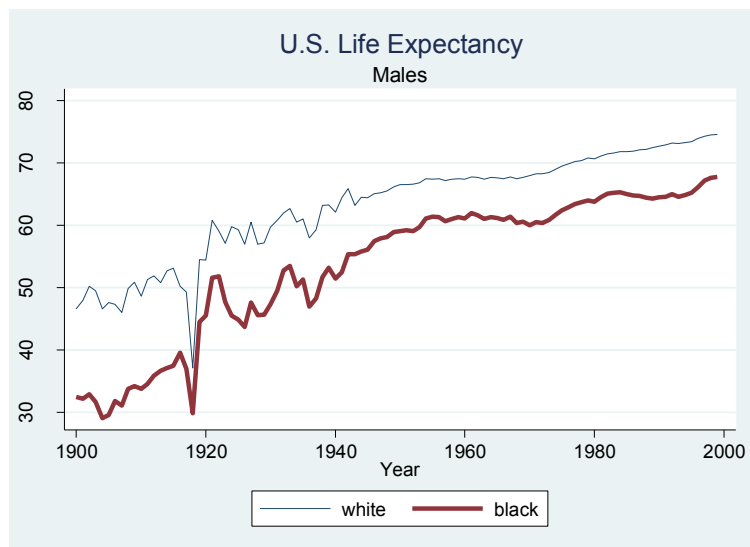
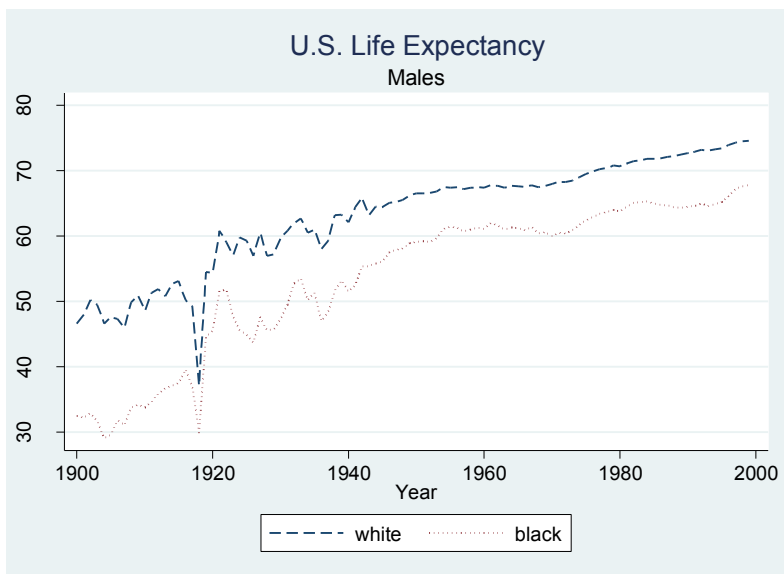
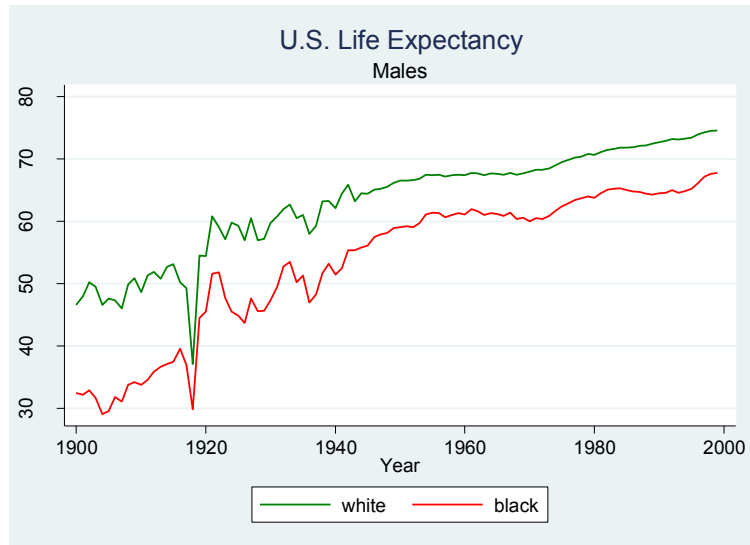
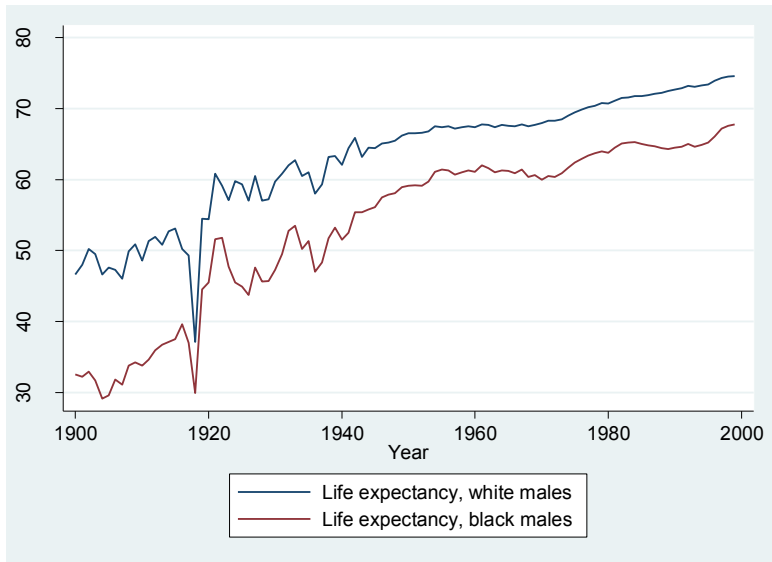
Line plots

```
graph twoway line le_wmale le_bmale year,  
legend(col(1) pos(6))
```

* If you are puzzled by the dip prior to 1920 just Google "US life expectancy 1918" *

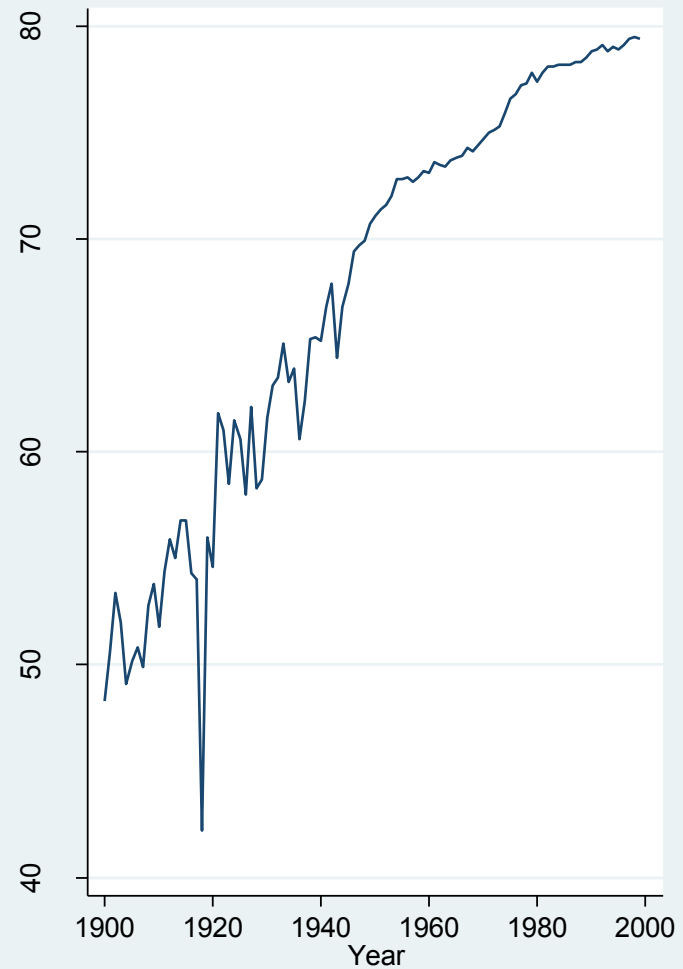
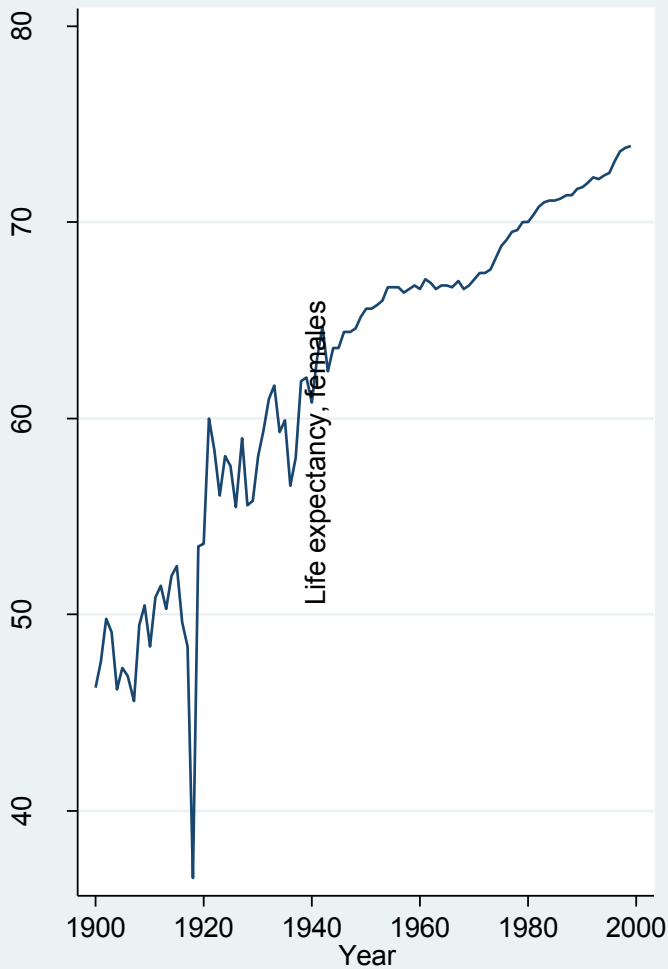


Line plots



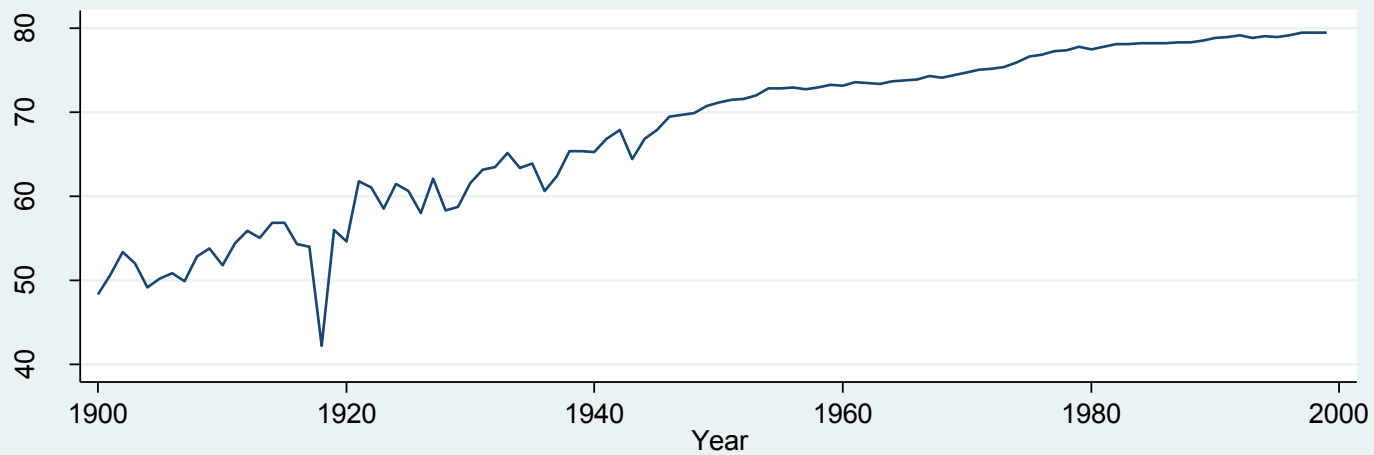
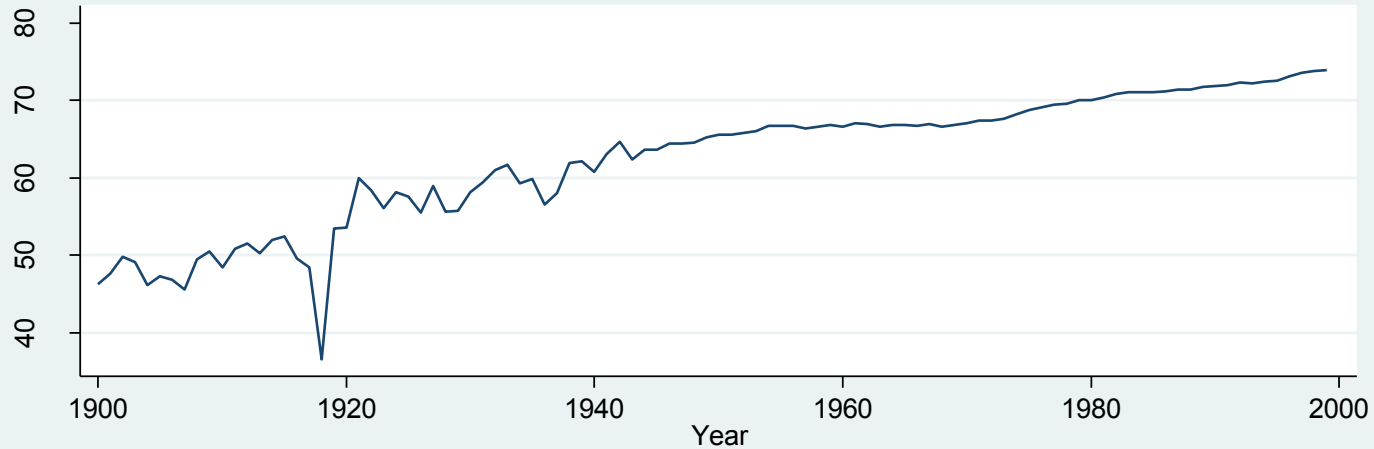
Line plots

Males and females side by side



Line plots

Males and females on top of each other



Graphing results

Collection results

(much more of this at the workshop)

Examples include

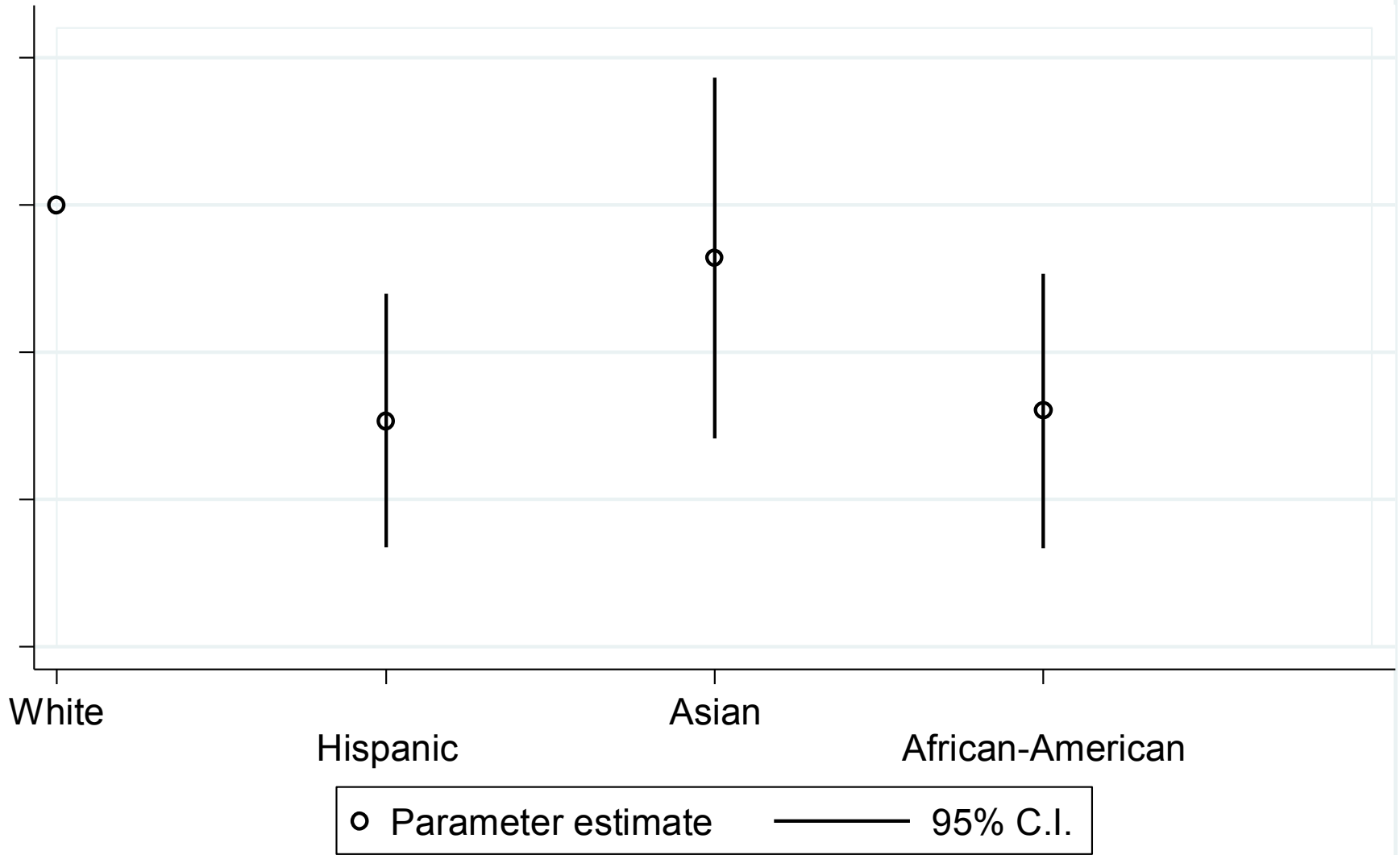
Statsby

Collapse

Modelling results

Standardised Reading Score, by Gender and Ethnicity

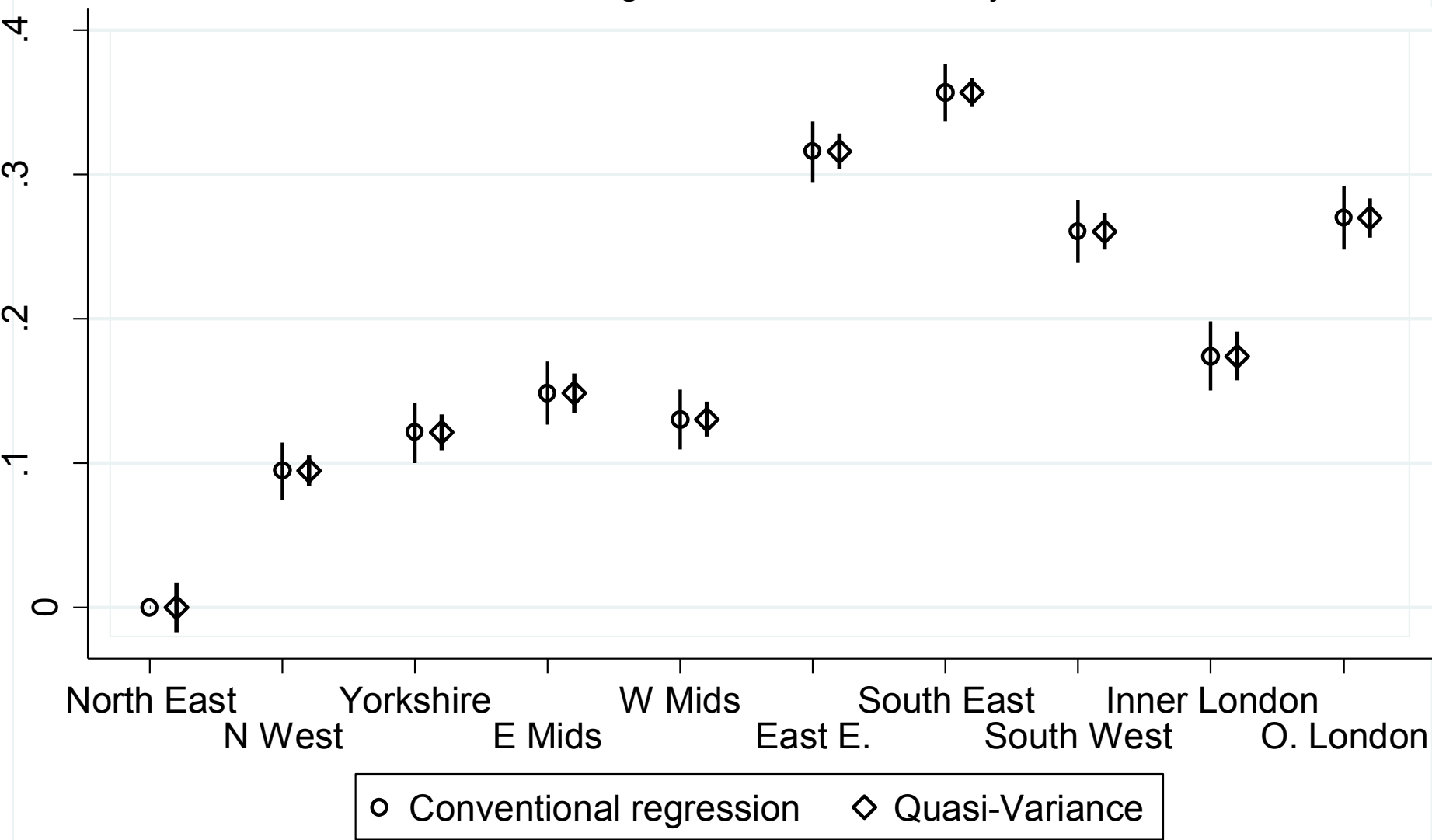
Regression coefficients and confidence intervals



Source: High School and Beyond, n=200.
Model 1: Regression model 'Reading score'.

Predictions of Good Health, by Government Office Region

Confidence intervals of regression coefficients, by estimation method



Source: UK Census 2001 SARS for England, n=1099294.

Model 1: Logistic regression predicting 'Good Health'. Other controls for education and gender

Some general points

- Aim for 'black and white' most of the time
- Some information (e.g. Figure 1) might be best controlled in your master document rather than within the graph
- Avoid graph clutter
 - clean, clear looks
 - avoid pseudo 3-D etc
 - limit the amount of greying (Excel problem)

References and useful links

Kohler,U. and Kreuter, F. (2009) Data Analysis Using Stata, Stata Press.
ISBN 978-1-59-718-046-7

Mitchell, H. (2008) A Visual Guide to Stata Graphics(2nd edition), Stata Press.
ISBN 1-59718-039-4

Professor Stephen Jenkins' web page <http://www.iser.essex.ac.uk/people/stephenj>

The UCLA Academic Technology Service website
<http://www.ats.ucla.edu/stat/stata/library/GraphExamples/default.htm>

Stata graphics visual overview
<http://www.stata.com/support/faqs/graphics/gph/statagraphs.html>

Survey Design and Analysis Services Pty Ltd website
<http://www.survey-design.com.au/Stata%20Graphs.html>



Applied Quantitative Methods Network

© *Copyright*

Vernon Gayle, University of Stirling.

This file has been produced for AQMeN by Vernon Gayle.

Any material in this file must not be reproduced, published or used for teaching without permission from Professor Gayle.

Many of the advanced ideas and examples presented in this document draw heavily on previous work developed by Professor Stephen Jenkins (s.jenkins@lse.ac.uk).

Over the last decade much of the Stata materials that Professor Gayle has developed have been in close collaboration with Dr Paul Lambert, Stirling University.

However, Professor Gayle is responsible for any errors in this file.

Professor Vernon Gayle (vernon.gayle@stir.ac.uk)