STATA WORSHOP SERIES

Data Analysis Using Stata

Jin Chen

Indiana Statistical Consulting Center

Feb 11, 2012

Outline of the Workshop

- Review of Part I
- Statistics
  - Descriptive statistics
  - Inferential statistics
- Data visualization
- Automation (if time permits)
  - Macros
  - Loops

Review

- Data Management in Stata
  - File types: .dta , log-file, do-file, ado-file
  - Data loading : use , insheet, infile, import excel
  - Data file exploration: codebook, describe, summarize, list
  - Variable manipulation: gen, egen, recode, rename, replace
  - Notes and labels: note, label data, label var, label define, label values
  - Organize data and observations: order, sort, keep and drop
  - Getting help: help, search

Efficiency and Replication In Data Analysis

- Use do file instead of command window or menus
- Create or apply a do-file template that works for you
  - E.g. see the slides for Stata workshop section I
- Comment as much as possible
- Name your data files, do files and log files in a systematic way

Statistics in Stata

- Descriptive statistics
  - Summary statistics
    - codebook and summarize (see workshop part I)
  - Oneway tables of frequencies:
    - tabulate *varname [if] [in] [weight] [,missing nolabel generate summarize()…]*
    - tab1 *varlist [if] [in] [weight] [, options]*
  - Twoway tables of frequencies:
    - tabulate *varname1 varname2 [if] [in][wt] [, row col cell nofreq missing nolab…]*
    - tab2 *varlist [if] [in][wt] [,options]*

- Examples

```
sysuse auto, clear
codebook, comp
tabulate rep78
tabulate rep78, m
tabulate foreign
tabulate foreign, nolab
tabulate foreign, summarize(price)
tabulate foreign, gen(foreign_d)
tab rep78 foreign, chi2
tab rep78 foreign, col row
```

- Table of summary statistics
  - tabstat *varlist [if] [in] [wt] [, by () statistics() columns(variables) columns(statistics) nototal missing…]*
    - Displays summary statistics for a series of numeric variables in one table, can be broken down on another variable
    - A wide range of descriptive statistics are available including but not limited to mean, n, sum, max, min, range, sd, variance, skewness, kurtosis, percentiles
    - Flexible in ways of displaying the table
  - Examples
    - tabstat price weight, statistics(n mean sd)
    - tabstat price weight, by(foreign) stat(n mean sd)
    - tabstat price weight, by(foreign) stat(n mean sd) col(statistics)

- Covariance and Correlation
  - correlate [*varlist*] [*if*] [*in*] [weight] [, *options*]
    - Display correlation matrix or covariance matrix
  - pwcorr [*varlist] [if] [in] [weight] [, options]*
    - Display all pairwise correlation coefficients
  - Inspect [*varlist*] [*if*] [*in*]
    - Display simple summary of data attributes
  - Examples
    - correlate price mpg weight
    - correlate price mpg weight, cov
    - pwcorr price mpg weight, star(.05)

Inferential Statistics

- T-test (mean-comparison test)
  - One sample t-test

    ttest varname==# [if] [in] [, level(#)]
  - Two sample t-test

    ttest varname1==varname2 [if] [in], unpaired [unequal level(#)]
  - Paired-sample t-test

    ttest varname1==varname2[if] [in] [, level(#)]
  - Test group means

    ttest varname [if] [in], by(group) [options]

Examples

```
sysuse auto, clear
ttest mpg==20
*=================================================
webuse fuel1
ttest mpg1==mpg2
*=================================================
webuse fuel3
ttest mpg, by(treated)
```

- Linear Regression
  - regress depvar indvarlist [if] [in] [wt] [, options]
  - factor variables

    i.var

    i.var1#i.var2

    i.var1#c.var3
  - Save/recall results

    estimate store *name*

    estimates restore *name*

    Alternatively, you can use

    eststo *name*

    eststo *name*: regress depvar indvarlist

- Post-estimation: test coefficients

  test var1

  test varlist

  test var1

  test var2, accum

  test var1=var2

  test var1-var2=0

- Prediction

  predict newvar [, xb]

  predict newvar , stdp

  predict newvar , residual

- Export publication-quality regression tables

  estout [namelist] [using filename] [, options]
  esttab [namelist] [using filename] [, options]

  esttab is a wrapper for estout. It produces a pretty-looking publication-style regression table from stored estimates without much typing. The compiled table is displayed in the Stata results window or, optionally, written to a text file specified by using filename.

- Try commands # 7 in demo do-file

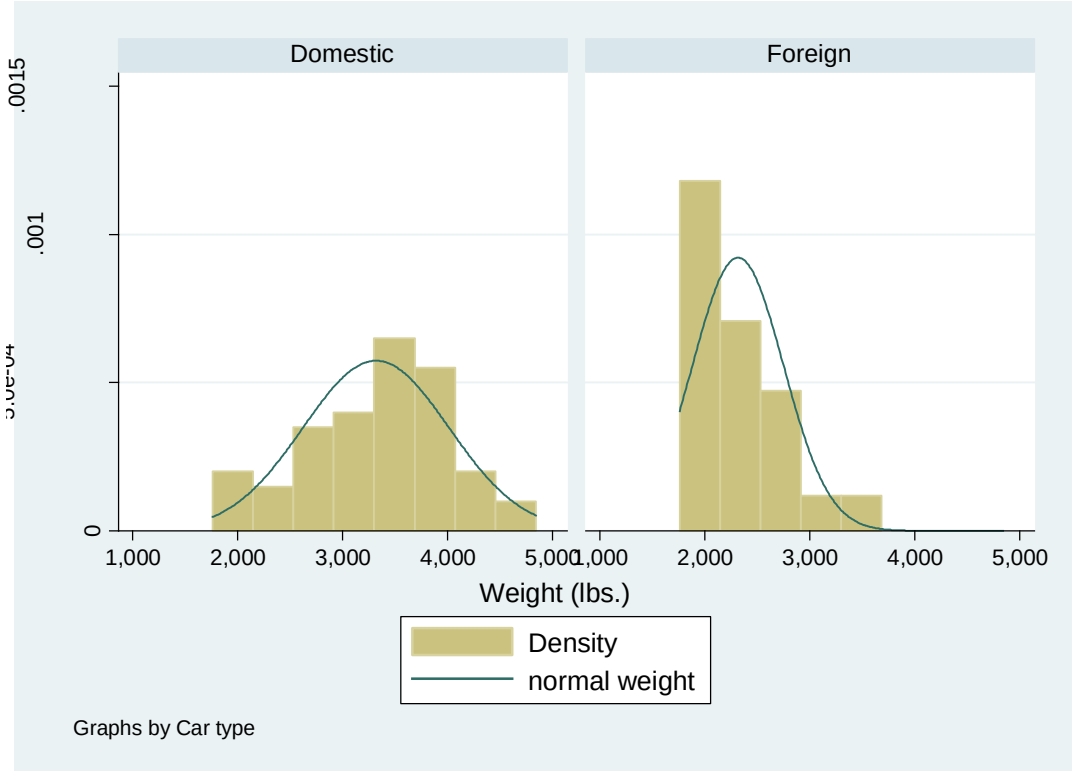
- More advanced models
  - logistic regression : logit
  - ordered logistic regression: ologit
  - multinomial logistic regression: mlogit
  - poisson regression : poisson
  - panel/longitudinal data modeling: xt-
  - complex survey designs: svy-
  - survival analysis: st-
  - time series analysis: ts-
- Read the Stata manual for more detailed information

Data Visualizations

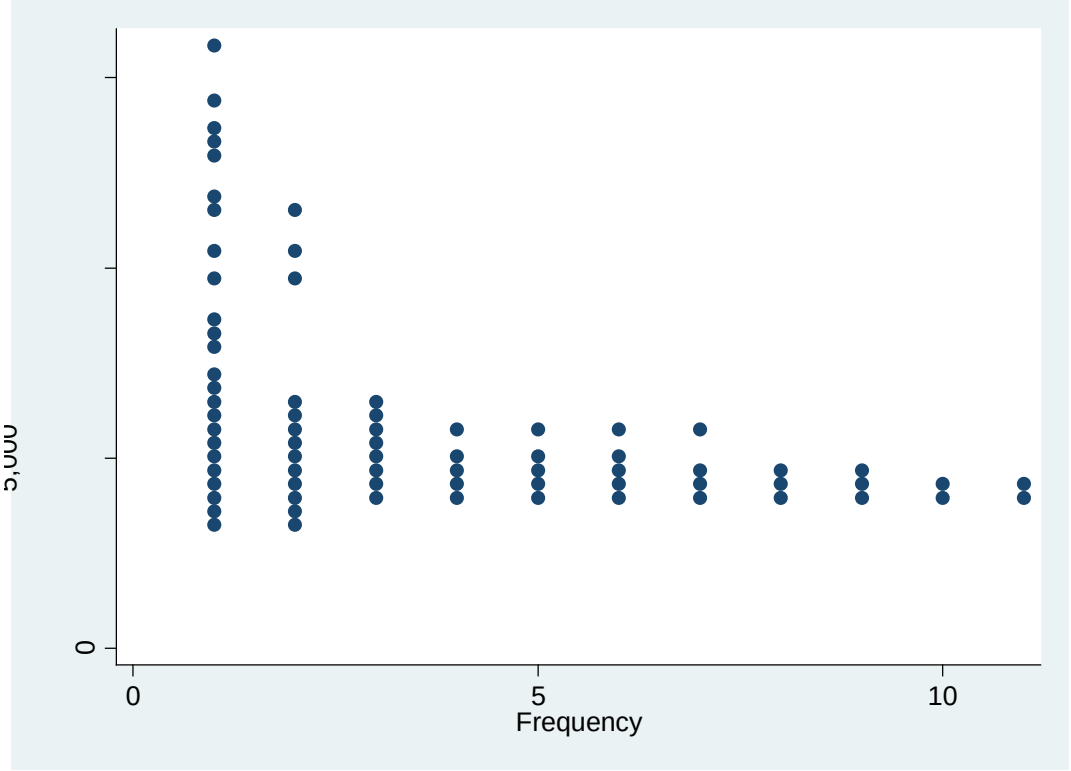
- In general, graphics menus allows complex specifications of the graphs, while syntax is easier for drawing convenient graphs
- For replication purposes, you can use menus to draw the desired graphs and save corresponding syntax in your do-file
- A good reference book: ***A Visual Guide to Stata Graphics, 3rd Edition,*** by Michael N. Mitchell

- Univariate Distribution (by group)

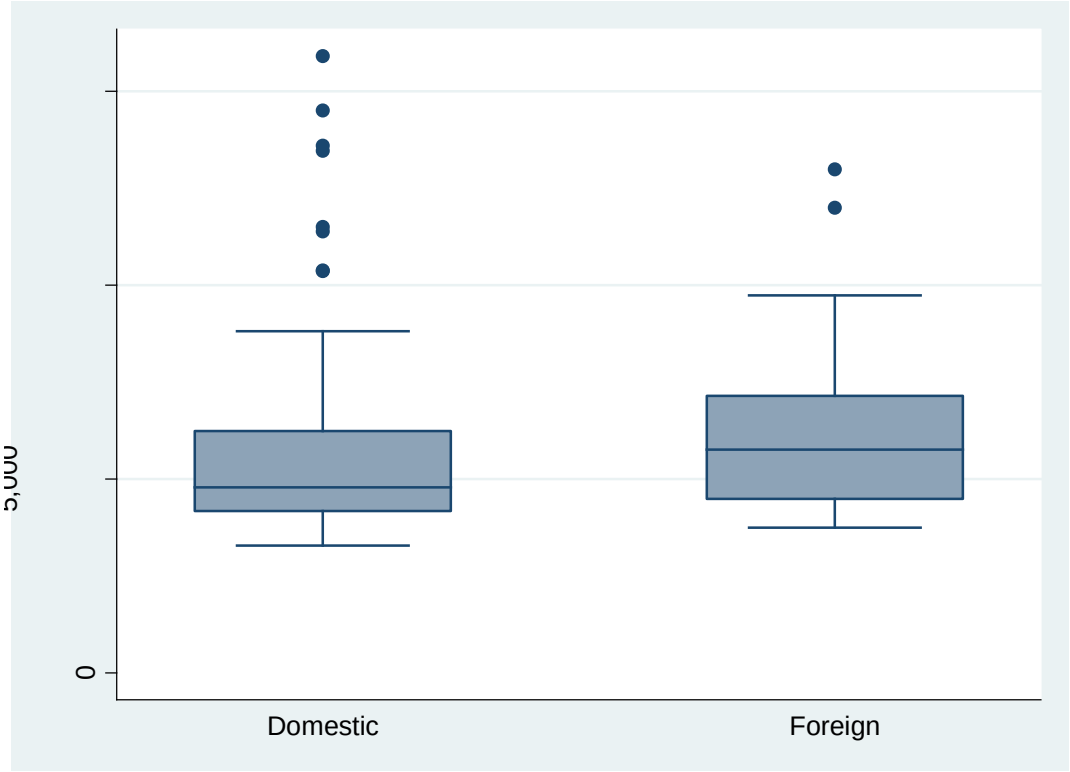
  histogram weight, normal by(foreign)



dotplot price

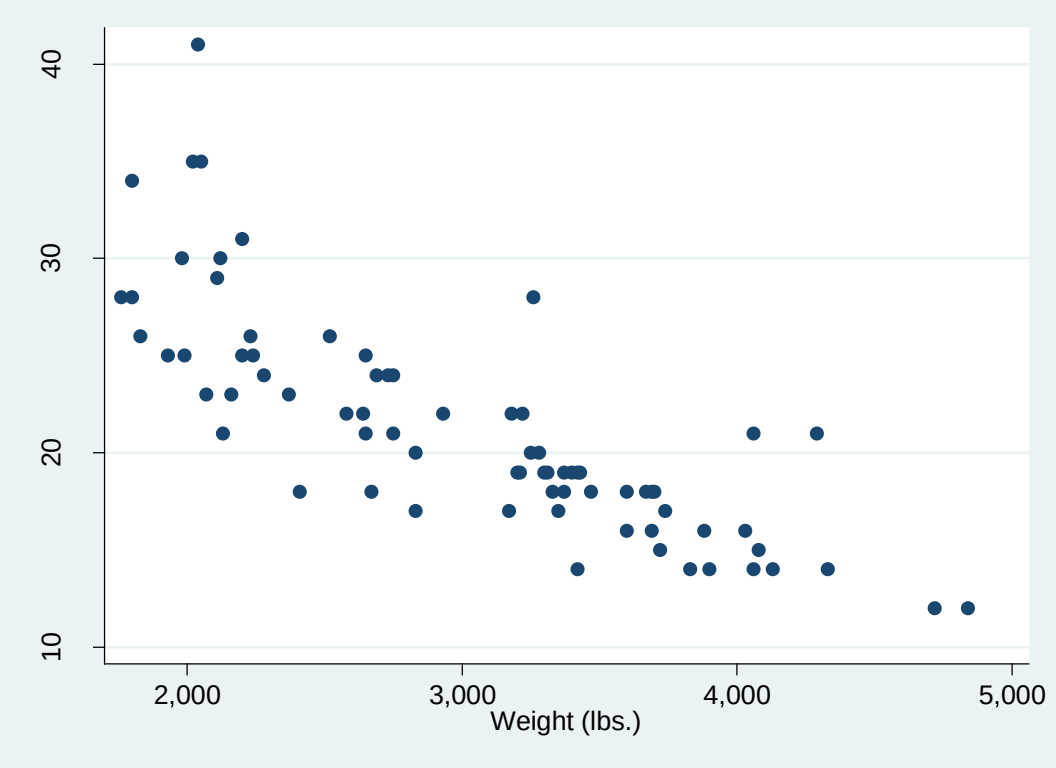


graph box price, over(foreign)



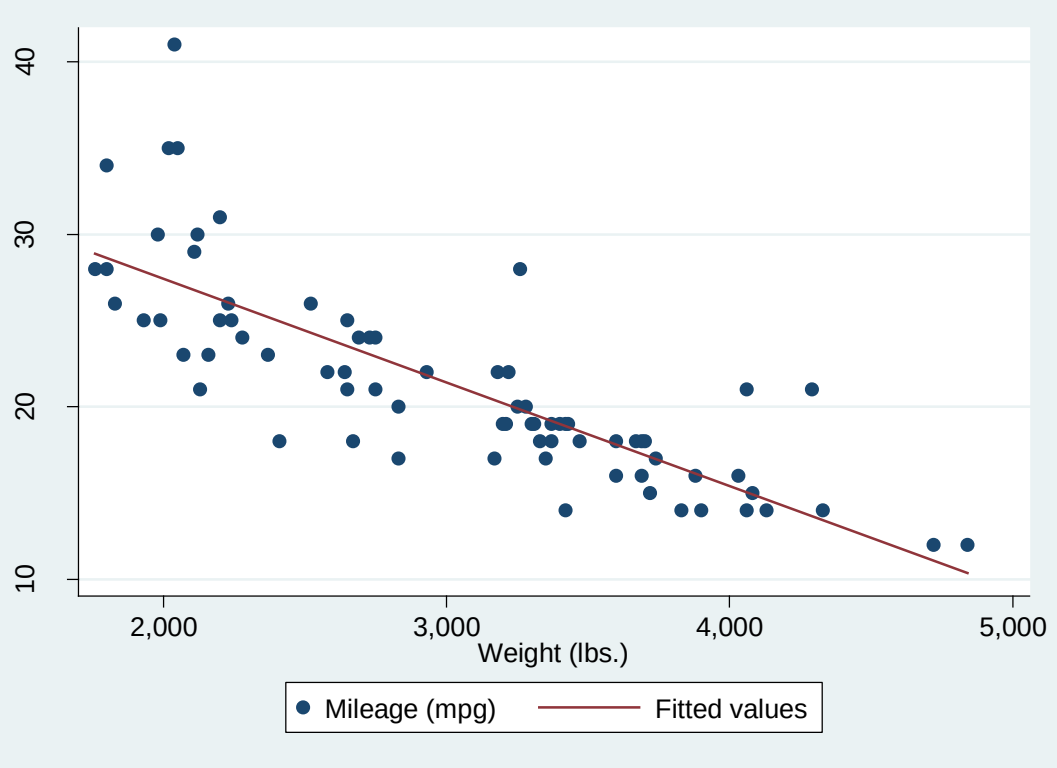
- Twoway Graphs

  scatter mpg weight



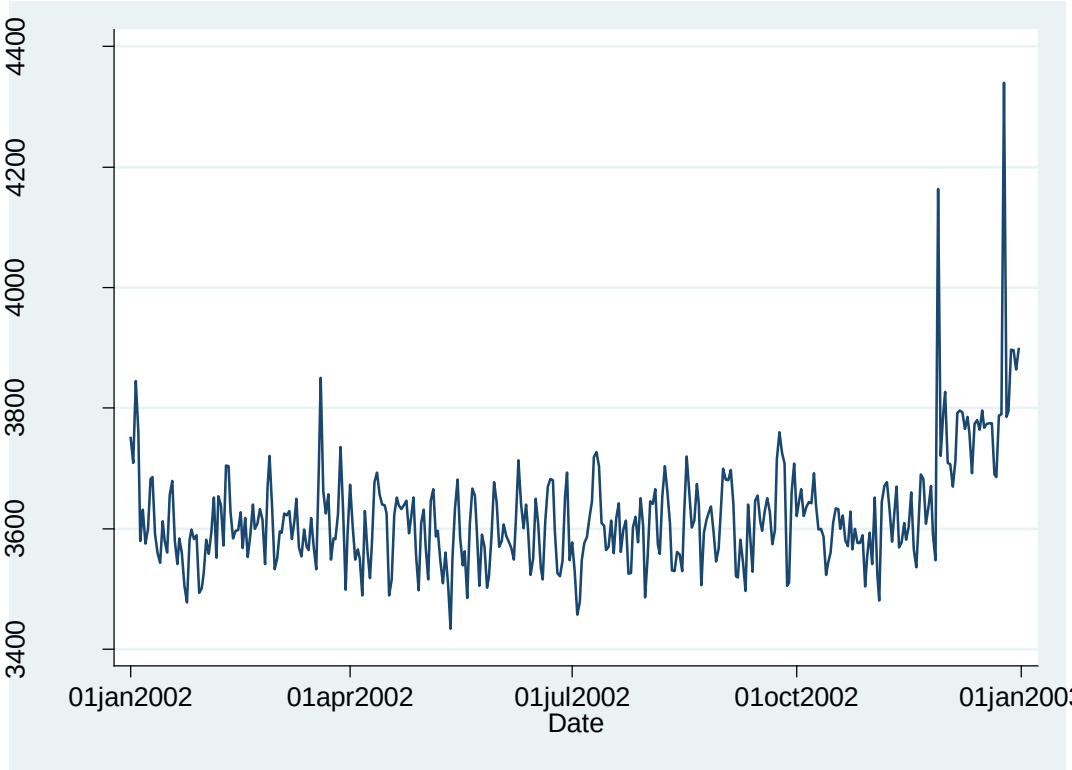
- Adding fitted line to twoway scatter plot

  scatter mpg weight || lfit mpg weight

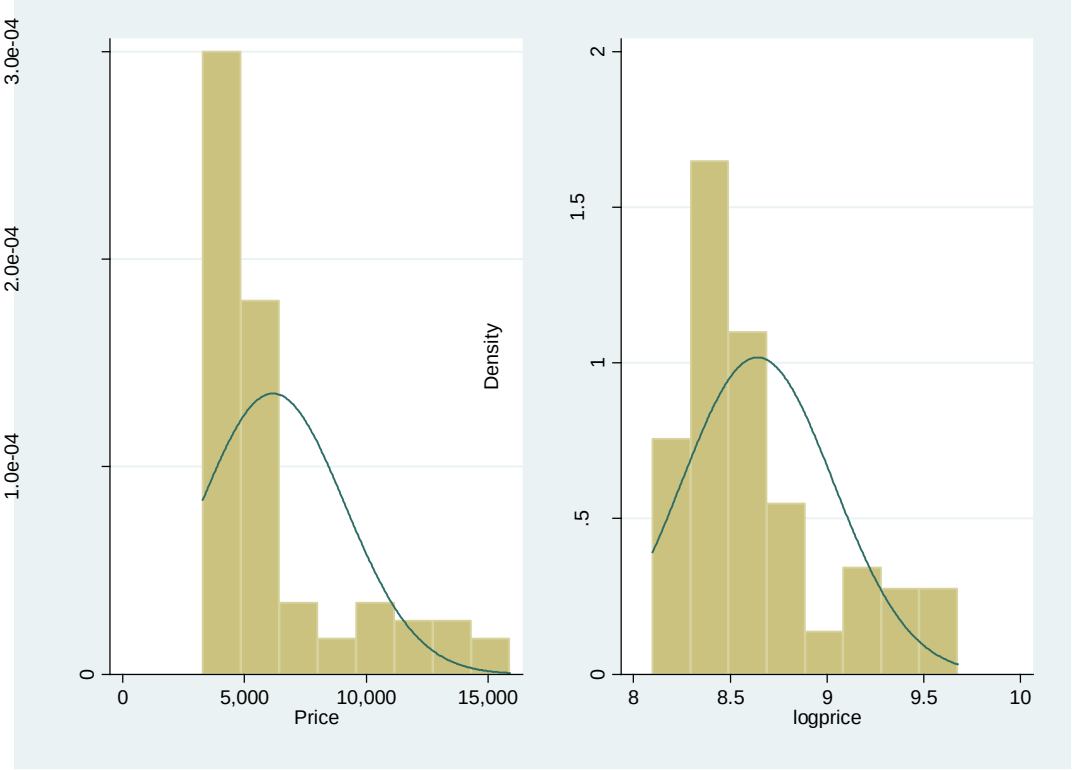


graph matrix price mpg weight

twoway (line calories day, sort)



```
histogram price, normal name(graph1)
gen logprice=log(price)
histogram logprice , normal name(graph2)
graph combine graph1 graph2, name(combined)
```



- Export graph from memory

  graph display *name*

  graph export *filename.suffix [, replace]*

  - *Suffix includes .ps .eps .wmf .emf .pdf .png .tif*

  e.g.

  graph export combined.png, replace

- Explore the graphics menus to produce more complex graphs
  - Add graph title, captions, and notes
  - Add labels for y-axis and x-axis
  - Define max, min values of the axes, etc.
- Use graph editor to refine your graphs

Automation: Towards Programming

- Macros

Assign names to represent a series of text, variables, numbers, etc.

- global: once defined are available anywhere in Stata (not recommended)
- local: exist only within the program or do-file where they are defined

e.g. local depvar "mpg foreign weight"

reg price `depvar'

- Loops

  repeat commands for each element in the list

  - foreach: loop over an arbitrary list of variables or numbers

  foreach *lname* in *list* {

  command referring to `lname'

  }

  e.g.

  foreach var in `depvar' {

  egen mn_`var'=mean(`var')

  }

- forvalues : loop over consecutive numbers

```
forvalues x=rang {
commands referring to `lname'
}
e.g.
forvalues i=1/5 {
sum price if rep78==`i’
}
```

Thank You! ☺

- Questions on and suggestions for this workshop??
- Need help for statistical analysis??
- Please contact us

  chen92@umail.iu.edu

  iscc@indiana.edu

References

- StataCorp. 2009. *Stata 11 Base Reference Manual*. College Station, TX: Stata Press.
- StataCorp. 2011. *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP.
- Long, J.S. 2009. *The Workflow of Data Analysis Using Stata*. College Station, TX: Stata Press.
- Wolfe, J.D. (2011, January 21). *Introduction to Stata*. Powerpoint lecture presented on the Indiana University campus. http://mypage.iu.edu/~jdwolfe/stintro_beamer_01-21-11.pdf