Labor Economics with STATA

Liyousew G. Borga

October 19, 2015

Data Management

Missing Data, Descriptive Statistics, and Graphics

Outline

1. Handling Missing Data

2. Descriptive Information and Statistics

3. Graphs in Stata

Evaluate and Understand Missing Data

The problem of missing data

- Missing data always cause some loss of information which cannot be recovered
- Missing data can pose major problems when estimating econometric models since it is generally unlikely that missing values are completely random
- More seriously, missing data can introduce bias into our estimates
- However, statistical methods can often help us make best use of the data which has been observed

Evaluate and Understand Missing Data

General Steps for Analysis with Missing Data

- Identify patterns/reasons for missing and recode correctly
- Understand distribution of missing data
- Decide on best method of analysis

Evaluate and Understand Missing Data

Step One: Understand your data

- Attrition due to social/natural processes

  Example: School graduation, dropout, death
- Skip pattern in survey

  Example: Certain questions only asked to respondents who indicate they are married/employed
- Intentional missing as part of data collection process
- Random data collection issues
- Respondent refusal/Non-response

Evaluate and Understand Missing Data

Step One: Understand your data

- Attrition due to social/natural processes

  Example: School graduation, dropout, death
- Skip pattern in survey

  Example: Certain questions only asked to respondents who indicate they are married/employed
- Intentional missing as part of data collection process
- Random data collection issues
- Respondent refusal/Non-response

Starting point: Find information from survey (codebook, questionnaire)

- Identify skip patterns and/or sampling strategy from documentation

Evaluate and Understand Missing Data

Step Two: Missing data Mechanism (or probability distribution of “missingness”)

- Consider the probability of “missingness”
  - Are certain groups more likely to have missing values?
    Example: Respondents in service occupations less likely to report income
  - Are certain responses more likely to be missing?
    Example: Respondents with high income less likely to report income
- Certain analysis methods assume a certain probability distribution

Evaluate and Understand Missing Data

Missing Data Mechanisms:

- Missing Completely at Random (MCAR)

  Missing value (y) neither depends on x nor y

  Example: some survey questions asked of a simple random sample of original sample
- Missing at Random (MAR)

  Missing value (y) depends on x, but not y

  Example: Respondents in service occupations less likely to report income
- Missing not at Random (NMAR)

  The probability of a missing value depends on the variable that is missing

  Example: Respondents with high income less likely to report income

Evaluate and Understand Missing Data

Step 3: Deal with missing data

- Use what you know about
    - Why data is missing
    - Distribution of missing data
- Decide on the best analysis strategy to yield the least biased estimates
    - Deletion Methods: List-wise deletion, pairwise deletion
    - Single Imputation Methods: Mean/mode substitution, dummy variable method, single regression
    - Model-Based Methods: Maximum Likelihood, Multiple imputation

How STATA handles missing data in STATA procedures

- Patterns of missing values:

```
// Create a small dataset
cd "D:\Study\CERGE-EI\TA\Labor_economics\ES\ES2\Datasets"
import excel using Missing_values.xlsx, firstrow clear
describe
// Examine the dataset for missing values; i.e., determine which
    variables have a lot of missing values
// A user written Stata program called "mdesc" counts the number of
    missing values in both numeric and character variables: download it
    by typing "findit mdesc"
mdesc
// Now we know the number of missing values in each variable
```

How STATA handles missing data in STATA procedures

- Obtaining the number of missing values per observation:

```
egen nmis=rmiss2(landval improval totval salepric saltoapr)
// creates a variable called "nmis" that gives the number of missing
    values for each observation
// You can download rmiss2() over the internet from within Stata by
    typing "findit rmiss2"
// It counts the number of missing values in the varlist
tab nmis
```

- When all the variables of interest are numeric:

```
\item
mvpatterns
//produces output for all variables in the dataset, for missing data
    patterns across a subset of variables
misschk landval improval totval salepric saltoapr, gen(miss)
// The output for misschk consists of three tables
// The first table lists the number of missing values, as well as
    percent missing for each variable
// The second table shows the distribution of missing values
// The third table shows the distribution of the number of missing
    values per case
```

Summary Statistics and Tables

Getting an overview of your file

- The **describe** command shows you basic information about a Stata data file

```
sysuse auto // to load the ''auto'' data file that was shipped with
    Stata
describe
// It tells us the number of observations in the file, the number of
    variables, the names of the variables, and more
```

- The **codebook** command is a great tool for getting a quick overview of the variables in the data file

```
codebook
// It produces a kind of electronic codebook from the data file
```

- Another command for getting a quick overview of a data file is **inspect**

```
inspect
```

- The **list** command is useful for viewing all or a range of observations

```
list make price mpg rep78 foreign in 1/10
```

Summary Statistics and Tables

Generating summary statistics

- For summary statistics, we can use the **summarize** command
- We can use the **detail** option of the **summarize** command to get more detailed summary statistics
- The **tabstat** command provides a more flexible alternative to summarize

```
summarize mpg
sum mpg, detail
bysort foreign: summarize mpg
tabulate foreign, summarize(mpg)
tabstat mpg, stats(n, mean, sd, min, max)
tabstat mpg, stats(n, mean, sd, min, max) by(foreign)
```

Summary Statistics and Tables

Frequency Tables and Two-Way Cross-Tabulations

- The summary statistics described above apply mainly to measurement variables
- Categorical variables require different approaches, often starting with simple one- or two-way tables

```
use Granite2011_6.dta, clear
tabulate trackus
// tabulate can produce frequency distributions for variables that have
    thousands of values
tabulate educ trackus
// tabulate followed by two variable names creates a two-way cross-
    tabulation
```

- Multiple Tables and Multi-Way Cross-Tabulations

```
tab1 tparty obama trackus
tab1 tparty-trackus
tab2 tparty obama trackus
tab obama college, col nof
bysort sex: tab obama college, col nof
```

Summary Statistics and Tables

Frequency Tables and Two-Way Cross-Tabulations

- Alternative way to produce multi-way tables is through Stata's general table-making command, **table**

```
table obama college, contents(freq)
table obama college sex, contents(freq)
table obama college sex, contents(freq) by(married)
table obama college sex, contents(mean age) by(married)
```

Creating Publication-Quality Tables in Stata

Stata users have written programs that create publication-quality tables

- Tables of summary statistics

```
ssc install estout

sysuse auto
estpost summarize price mpg rep78 foreign, listwise
esttab, cells("mean sd min max") nomtitle nonumber
// Summary statistics can also be posted by estpost tabstat:
estpost tabstat price mpg rep78, listwise statistics(mean sd)
// Type columns(statistics) to print statistics in columns:
esttab, cells("price mpg rep78") nomtitle nonumber
by foreign: eststo: quietly estpost summarize price mpg rep78, listwise
```

- Post summary statistics by subgroups (summarize):

```
esttab, cells("mean sd") label nodepvar
eststo clear
// Alternative way to post summary statistics by subgroups:
estpost tabstat price mpg rep78, by(foreign) statistics(mean sd) columns
    (statistics) listwise
esttab, main(mean) aux(sd) nostar unstack noobs nonote nomtitle nonumber
```

Creating Publication-Quality Tables in Stata

- Post results from two-sample mean-comparison tests (ttest):

```
estpost ttest price mpg headroom trunk, by(foreign)
esttab, wide nonumber mtitle("diff.")
```

- Post a one-way frequency table (tabulate)

```
estpost tabulate foreign
esttab, cells("b(label(freq)) pct(fmt(2)) cumpct(fmt(2))") varlabels(,
   blist(Total "{hline @width}{break}")) nonumber nomtitle noobs
```

- Post a two-way frequency table (tabulate):

```
estpost tabulate rep78 foreign
esttab, cell(colpct(fmt(2))) unstack noobs
```

- Post correlation coefficients (correlate):

```
estpost correlate price turn foreign rep78
esttab, cell("rho p count") noobs
estpost correlate price turn foreign rep78, matrix listwise
esttab, unstack not noobs compress
```

Creating Publication-Quality Tables in Stata

- Using Word, Excel and LaTeX

```
esttab using example.rtf
(output written to example.rtf)
esttab using example.csv, cells("mean sd count")
esttab using example.tex, label nostar title(Results Table\label{tab1})
(output written to example.tex)
```

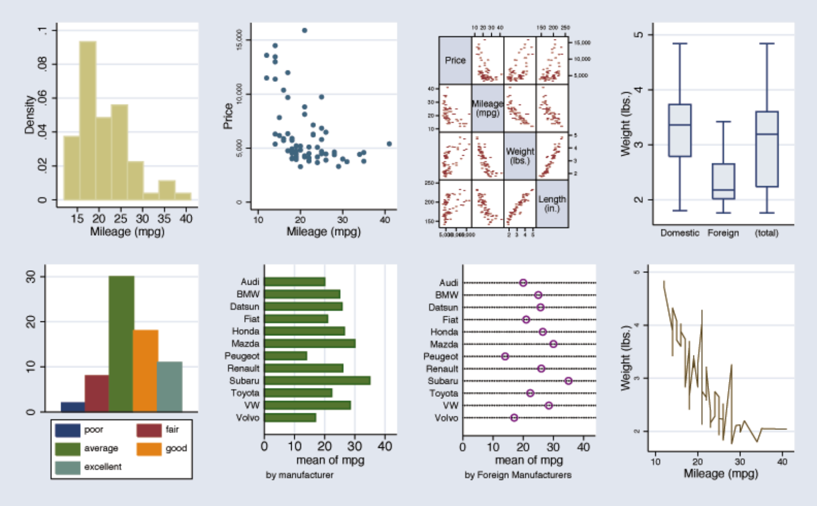
- Using the “tabout” command

```
ssc install tabout
tabout [ varlist ] [ if exp ] [ in range ] [ weight = exp ] using
    filename [ , options ]

help tabout
// The command produces publication quality tables for export to a text
    file (in word, excel, and latex compilers) - many functionalities
    worth checking out
```

Graph styles

Bar charts; Box plot; Histogram; Spike plots; Pie charts; Scatterplot matrix; Dot chart; Line charts; Area charts; Two-way scatterplot



Graph styles

Histograms:

- The command **histogram** displays the distribution of measurement variables

```
use Nations2.dta, clear
describe
histogram adfert, percent
```

- Options can be listed in any order following the comma in a graph command

```
histogram adfert, frequency start(0) width(10) xlabel(0(20)200) xtick
    (10(20)210) ylabel(0(5)35, grid gmax) title("Adolescent fertility
    rate in 194 nations") note(a)
histogram adfert, percent start(0) width(10) by(region, total)
```

Graph styles

Box Plots:

- Box plots convey information about center, spread, symmetry and outliers at a glance

```
graph box adfert
graph box adfert, marker(1, mlabel(country)) ytitle("Births per 1,000
    females 15-19")
graph hbox adfert, marker(1, mlabel(country)) yline(39.2) over(region)
```

- Box plots can have a horizontal orientation instead of vertical, via the “graph hbox command”

```
graph hbox co2, over(region) ///
note("note: {bf:Statistics with Stata}, version 12") ///
caption("caption: United Nations Human Development Report 2011") ///
title("title: {it:Example of horizontal box plots}") ///
ytitle("ytitle: Tons of CO{subscript:2} emitted per capita")
```

Graph styles

Scatter-plots and Overlays:

- Scatterplots belong to a broad family called "twoway graphs"

```
graph twoway scatter life school
graph twoway scatter life school [fweight=pop], msymbol(Oh)
```

- We can overlay two or more graphs to build more complex images
- Simple regression lines (lfit) are a different twoway type. However, we can overlay the scatterplot and regression line together

```
graph twoway lfit life school, lwidth(medthick)
graph (twoway scatter life school [fweight=pop]), msymbol(Oh) || lfit
    life school, lwidth(medthick)
graph twoway scatter life school, msymbol(Oh) || lfit life school,
    lwidth(medthick) || , ylabel(45(5)85) xlabel(2(2)12) xtick(1(2)13)
    legend(col(1) ring(0) position(11))
```

Graph styles

Scatterplot matrices:

- Scatterplot matrices are not twoway plot types and cannot be overlaid with other graphs
- However, they involve multiple scatterplots that follow the same marker symbol conventions
- A scatterplot matrix is the visual counterpart to a correlation matrix, which can be useful in multivariate analysis

```
graph matrix gdp school adfert chldmort life, msymbol(oh)
graph matrix gdp school adfert chldmort life, half msymbol(oh)
```

Graph styles

Line Plots and Connected-Line Plots:

- Connected-line plots (graph twoway connect) are just scatterplots in which the points are connected by line segments

```
use Arctic9.dta, clear
describe

graph twoway line area year, title("Arctic sea ice, Sept. 1979-2011")

graph twoway line area extent year, xlabel(1980(5)2010) xtitle("")
    lwidth(medium medthick) lpattern(solid dash) legend(row(2) ring(0)
    position(9) label(1 "Area") label(2 "Extent") order(2 1)) ylabel(0(1)
    8, grid gmin gmax) ytitle("Million km{superscript:2}") title("Arctic
    sea ice, September 1979`=char(150)'2011")
```

Graph styles

Bar Charts and Pie Charts:

- The **graph bar** command provides clear visualizations of relationships involving many categories and two or more variables

```
use Nations2.dta, clear
describe region gdp pop

generate gdp1000 = gdp/1000
summarize gdp gdp1000

graph bar (mean) gdp1000 (median) gdp1000, over(region) ytitle("Per
    capita GDP, thousands of 2005 US dollars") blabel(bar, format(%3.1f))
     bar(1, color(blue)) bar(2, color(orange)) legend(ring(0) position
    (11) col(2) label(1 "Mean") label(2 "Median") symxsize(*.5))
```

- Pie charts rarely clarify the analysis but are popular for some public presentations

```
gen popmil = pop/1000000
summarize pop popmil
graph pie popmil, over(region) pie(2, explode) plabel(_all sum, format
    (%4.0f)) title("World population in millions, by region") legend(col
    (1) position(9))
```

Graph styles

Managing Graphs:

- Stata keeps track of the last graph you have drawn, which is stored in memory, and calls it "Graph"
- You can actually keep more than one graph in memory if you use the `name()` option to name the graph when you create it
- To save the current graph on disk

```
graph save Graph graph_bar.gph, replace
graph export graph_bar.png, as(png) replace
graph export graph_bar.eps, as(eps) replace
graph export graph_bar.emf, replace /* can insert in Word*/
```

- Retrieving and Combining Graphs are also possible by the **graph use** command and the **graph combine** command

```
graph use graph_bar.gph
graph combine graph_bar.gph graph_pie.gph, rows(1) altshrink title("
    Combining Figures", size(medium))
graph save Graph graph_combined.gph, replace
graph export graph_combined.emf, as(emf) replace
```