Stata Review
and
a BRIEF introduction to matrices and matrix calculations

Stata Review

- If it has been awhile since you opened/used Stata, then it would be wise to walk through this brief review of some key features.

I. Reading Data:

- use

Read data that have been saved in Stata format.

- infile

Read ".raw" and ".data" data and "dictionary" files.

- insheet

Read spreadsheets saved as "CSV" files from a package such as Excel.

II. Do Files

- **What is a do file?**

A “do” file is a set of commands just as you would type them in one-by-one during a regular Stata session. Any command you use in Stata can be part of a do file. Do files are very useful, particularly when you have many commands to issue repeatedly, or to reproduce results with minor or no changes.

Example: lab1.do

```
*the path and name of the files are specific to your computer;
*change the directory to where you have saved the files for use in lab 1
cd "C:\Users\ejohnson\Documents\LDA2013"
log using "lab1.log"
insheet using "pigs.csv"
save "pigs.dta“
Etc…
```

You can edit a do file anywhere then save as a file with the extension “.do”. In Windows or Mac, you can type “doedit” in Stata to open and edit any do files.

- **Where to put a do file?**

Put the do file in the working directory of Stata.

- **How to run a do file?**

do mydofile

Example: do lab1

III. Ado files

- **What is an ado file?**

An ado file is just a Stata program. You can use it as a command.
A *.ado file usually contains a program called * in it.
For example, the first non-comment line "autocor.ado" is program define autocor

- **Where do I save ado files?**

Save the .ado files and the corresponding .hlp files in your personal Stata "ado" directory.

Use "**adopath**" to find out where Stata is looking for ado files.

Here is an example in a Windows PC (Ado directory may be different among different platforms).

```
. adopath
  [1]  (BASE)      "C:\Program Files (x86)\Stata13\ado\base/"
  [2]  (SITE)      "C:\Program Files (x86)\Stata13\ado\site/"
  [3]              "."
  [4]  (PERSONAL)  "c:\ado\personal/"
  [5]  (PLUS)      "c:\ado\plus/"
  [6]  (OLDPLACE)  "c:\ado/"
```

I would store my ado files in the "c:\ado\personal" directory.

NOTE: There is always a few students for which this does not work! If that is the case, then you will want to put a copy of the ado file in the directory where you will be working. Stata should see it and everything should be fine.

- **How do I run an ado file?**

Use the name of the program as a command as you use other default Stata commands.

For example:

. autocor outcome time id

IV. Convert data from wide to long or vice versa

• **Two forms of data: wide and long**
Longitudinal data is stored in one of two formats: wide or long.
It is important to know how to go back and forth between these two formats.

Example: Incomes of 3 individuals in 1980-1982
(wide format)

```
-i- ------- x_ij --------
id sex inc80 inc81 inc82
-------------------------------
1 0 5000 5500 6000
2 1 2000 2200 3300
3 0 3000 2000 1000
```

(long format)

```
-i- -j- -x_ij
id year sex inc
1 80 0 5000
1 81 0 5500
1 82 0 6000
2 80 1 2000
2 81 1 2200
2 82 1 3300
3 80 0 3000
3 81 0 2000
3 82 0 1000
```

• **Reshape converts data from one form to the other:**
• **From Wide to Long**
```
. reshape long inc, i(id) j(year)
```
• **From Long to Wide**
```
. reshape wide inc, i(id) j(year)
```

Example: Guinea Pigs Weight data

```
. use pigs.dta, clear

. * List the first two observations
. list in 1/2

     +------------------------------------------------------------------------------------------------+
     | weight1   weight2   weight3   weight4   weight5   weight6   weight7   weight8   weight9   id |
     |------------------------------------------------------------------------------------------------|
  1. |      24        32        39      42.5        48      54.5        61        65        72    1 |
  2. |    22.5      30.5      40.5        45        51      58.5        64        72        78    2 |
     +------------------------------------------------------------------------------------------------+

. * Reshape to a long format
. reshape long weight, i(id) j(time)
(note: j = 1 2 3 4 5 6 7 8 9)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                       48   ->     432
Number of variables                  10   ->       3
j variable (9 values)                     ->   time
xij variables:
            weight1 weight2 ... weight9   ->   weight
-----------------------------------------------------------------------------
```

```
. list in 1/5

     +--------------------+
     | id   time   weight |
     |--------------------|
  1. |  1      1       24 |
  2. |  1      2       32 |
  3. |  1      3       39 |
  4. |  1      4     42.5 |
  5. |  1      5       48 |
     +--------------------+

. * Reshape back to long format
. reshape wide weight, i(id) j(time)
(note: j = 1 2 3 4 5 6 7 8 9)

Data                                   long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                          432   ->      48
Number of variables                       3   ->      10
j variable (9 values)                  time   ->   (dropped)
xij variables:
                                     weight   ->   weight1 weight2 ... weight9
-----------------------------------------------------------------------------
```

Part B: Longitudinal data analysis in Stata

I. Convert an ordinary dataset into a longitudinal dataset: use xtset

- "xtset" declares ordinary data to be panel data,

Cross-sectional data: one panel

Longitudinal (cross-sectional time-series) data: multi-panel

Each observation in a cross-sectional time-series (xt) dataset is an observation of x for unit i (panel) at time t.

For this course, we use cross-sectional time-series data.

Syntax for "xtset" for cross-sectional time-series data:

```
. xtset panelid timevar
```

Some but not all of our analysis commands will require that you initialize the data as a longitudinal (or panel) dataset.

Example:

```
    . use "endoflifemedicarecosts20032007_long", clear
    *Check to see if stata recognizes data as longitudinal
    . xtset
panel variable not set; use xtset varname ...
r(459);


    *Set data as longitudinal using xtset command
    . xtset statenum year

panel variable:  statenum (strongly balanced)
 time variable:  year, 2003 to 2007
         delta:  1 unit


    * What does xtset know now?
    . xtset
panel variable:  statenum (strongly balanced)
 time variable:  year, 2003 to 2007
         delta:  1 unit
```

II. xt commands

The xt series of commands provide tools for analyzing cross-sectional time-series (panel) datasets:

- xtdes Describes pattern of xt data

```
. use "endoflifemedicarecosts20032007_long", clear

. xtset statenum year
       panel variable:  statenum (strongly balanced)
        time variable:  year, 2003 to 2007
                delta:  1 unit

. xtdes

statenum:  1, 2, ..., 53                                   n =         51
    year:  2003, 2004, ..., 2007                           T =          5
           Delta(year) = 1 unit
           Span(year)  = 5 periods
           (statenum*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         5       5       5         5         5       5       5

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
       51    100.00  100.00 |  11111
 ---------------------------+---------
       51    100.00         |  XXXXX
```

- **Other xt commands that we will use often:**

xtsum Summarize xt data
xttab Tabulate xt data
xtmixed We will predominantly use this command to fit our linear models
xtgee Population-averaged panel-data models using Generalized Estimating Equations

- **Stata offers many additional historical commands (I say historical because now most of what you want to do can be done within xtmixed and xtgee).**

xtreg Fixed-, between- and random-effects, and population-averaged linear models
xtlogit Fixed-effects, random-effects, & population-averaged logit models
xtprobit Random-effects and population-averaged probit models
xttobit Random-effects tobit models
xtpois Fixed-effects, random-effects, & population-averaged Poisson models
xtnbreg Fixed-effects, random-effects, & population-averaged negative binomial models
xtclog Random-effects and population-averaged cloglog models
xtintreg Random-effects interval data regression models
xtrchh Hildreth-Houck random coefficients models
xtgls Panel-data models using GLS

III. Graphs for longitudinal data (This is just the start, Lecture 2 will provide you with many options here and all the Stata code!)

- **xtgraph**

Download the xtgraph.ado file from course website.

Syntax:

xtgraph varname [if] [in] , group(groupvar) av(avtype) bar(bartype) graph options xt options

xtgraph , av(avtype)

The average types (avtype) are

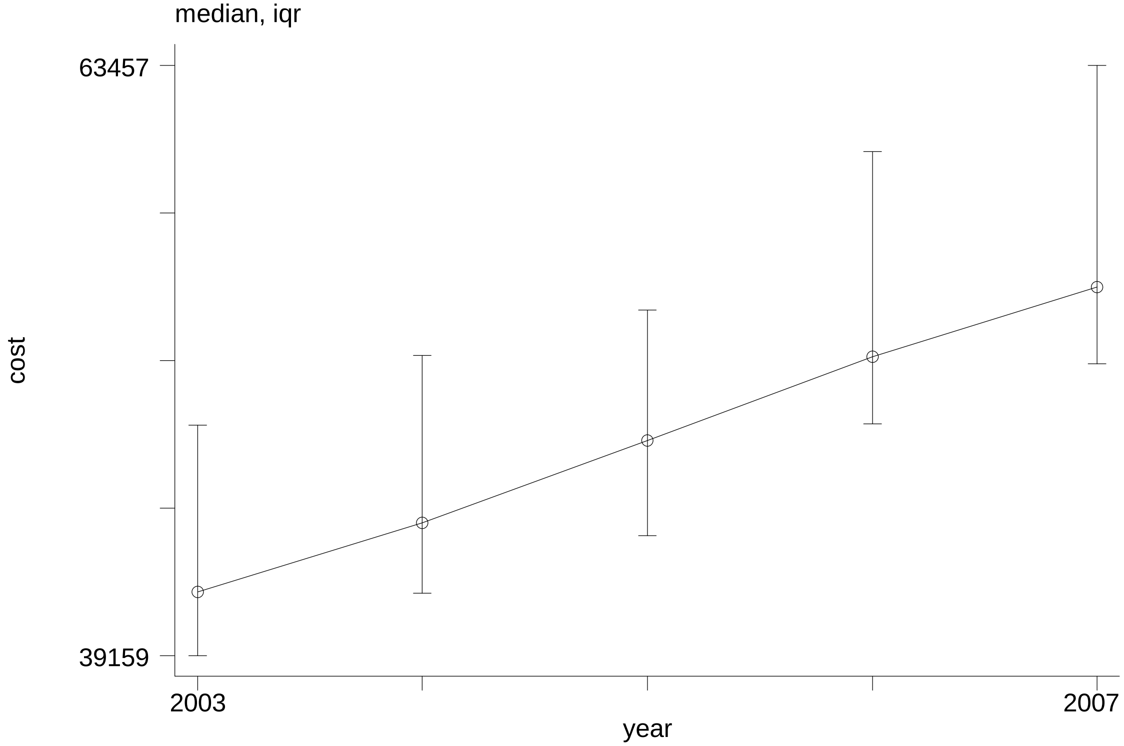
- am - arithmetic mean, the default
- gm - geometric mean
- hm - harmonic mean
- median - only with bars ci - default, iqr or rr.

The bar types (bartype) are

- ci - the default, significance set by level()
- se - standard error
- sd - standard deviation
- rr - reference range, level set by level()
- iqr -same as bar(rr) level(50)
- no - no bars

Examples (still using end of life medicare data)**:**

- xtgraph cost, av(median) bar(iqr) t1("median, iqr")



- NOTE: this graph is not "pretty". In Lecture 2 and throughout the course, we will provide you with important and useful options for making your graphs publishable quality.

Matrix Algebra

Definition: An $m \times n$ matrix, $\mathbf{A}_{m \times n}$, is a rectangular array of real numbers with m rows and n columns. Element in the i^{th} row and the j^{th} column is denoted by a_{ij}.

$$\mathbf{A}_{m \times n} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{21} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & \cdots & \cdots & a_{mn} \end{bmatrix}_{m \times n}$$

Definition: A vector $\mathbf{a}$ of length n is an $n \times 1$ matrix with each element denoted by a_i. The i^{th} element is called the i^{th} component of the vector and n is the dimensionality.

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

Matrix Operations

1. Two matrices **A** and **B** of the same dimensions can be added. The sum **A** + **B** has (i, j) entry $a_{ij} + b_{ij}$. So

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} \\ a_{21}+b_{21} & a_{22}+b_{22} \end{bmatrix}$$

$$\text{Example : } \mathbf{A} = \begin{bmatrix} 1 & 3 \\ 2 & 6 \\ -1 & -3 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 3 & 2 \\ 9 & 5 \\ 3 & 0 \end{bmatrix} \quad \mathbf{A}+\mathbf{B} = \begin{bmatrix} 4 & 5 \\ 11 & 11 \\ 2 & -3 \end{bmatrix}$$

- **A** + **B** = **B** + **A**
- (**A** + **B**) + **C** = **A** + (**B** + **C**)

2. A matrix may also be multiplied by a constant c. The product c**A** is the matrix that results from multiplying each element of **A** by c. Thus

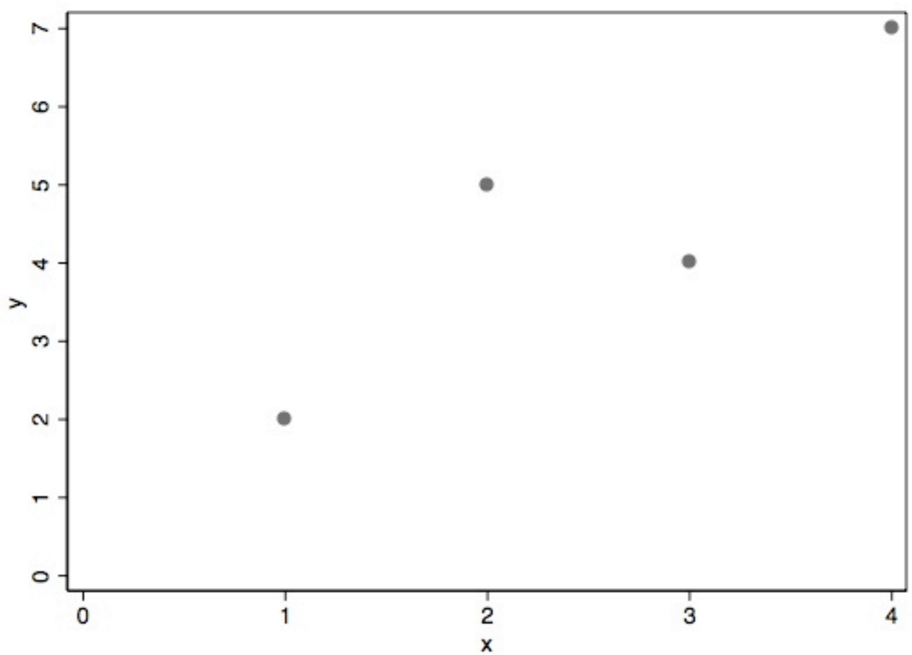
$$c\mathbf{A}_{m\times n} = \begin{bmatrix} ca_{11} & ca_{12} & \cdots & ca_{1n} \\ ca_{21} & ca_{21} & \cdots & ca_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ ca_{m1} & \cdots & \cdots & ca_{mn} \end{bmatrix}_{m\times n}$$

Illustrative Example of Simple Linear Regression

$X_1 = 1$ $Y_1 = 2$
$X_2 = 2$ $Y_2 = 5$
$X_3 = 3$ $Y_3 = 4$
$X_4 = 4$ $Y_4 = 7$

$$Y_i = E(Y_i | X_i) + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Illustrative Example of Simple Linear Regression

$X_1 = 1$ $Y_1 = 2$
$X_2 = 2$ $Y_2 = 5$
$X_3 = 3$ $Y_3 = 4$
$X_4 = 4$ $Y_4 = 7$

$Y_i = E(Y_i|X_i) + \varepsilon_i$

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Least Squares Solution

$$\overline{X} = 2.5 \qquad \overline{Y} = 4.5$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$\hat{\beta}_1 = \frac{(2-4.5)(1-2.5)+(5-4.5)(2-2.5)+(4-4.5)(3-2.5)+(7-4.5)(4-2.5)}{(1-2.5)^2+(2-2.5)^2+(3-2.5)^2+(4-2.5)^2} = 1.4$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X} = 4.5 - (2.5)(1.4) = 1$$

Matrix Representation of Simple Linear Regression

$Y_i = E(Y_i | X_i) + \varepsilon_i$

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$

$$\mathbf{Y} = \begin{bmatrix} 2 \\ 5 \\ 4 \\ 7 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

$$\begin{bmatrix} 2 \\ 5 \\ 4 \\ 7 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

Find βs that minimize $(\mathbf{Y}-\mathbf{X\beta})^T(\mathbf{Y}-\mathbf{X\beta})$

Solution: $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$

Matrix Representation of Simple Linear Regression

Solution: $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \quad \mathbf{X}^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 1(1)+1(1)+1(1)+1(1) & 1(1)+1(2)+1(3)+1(4) \\ 1(1)+1(2)+1(3)+1(4) & 1(1)+2(2)+3(3)+4(4) \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}$$

Matrix Representation of Simple Linear Regression

Matrix Property:

A square matrix that does not have a matrix inverse is called a **singular** matrix
The inverse of a 2×2 matrix is given by

$$\mathbf{A}_{2\times 2} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \mathbf{A}^{-1}_{2\times 2} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Remember -> Solution: $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{4(30) - 10(10)} \begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix} = \frac{1}{20} \begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \frac{1}{20} \begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix} \times \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix} = \frac{1}{20} \begin{bmatrix} 20 & 10 & 0 & -10 \\ -6 & -2 & 2 & 6 \end{bmatrix}$$

Matrix Representation of Simple Linear Regression

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \frac{1}{20}\begin{bmatrix} 20 & 10 & 0 & -10 \\ -6 & -2 & 2 & 6 \end{bmatrix}$$

And Finally….

Solution: $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \frac{1}{20}\begin{bmatrix} 20 & 10 & 0 & -10 \\ -6 & -2 & 2 & 6 \end{bmatrix}\begin{bmatrix} 2 \\ 5 \\ 4 \\ 7 \end{bmatrix} = \frac{1}{20}\begin{bmatrix} 20 \\ 28 \end{bmatrix} = \begin{bmatrix} 1 \\ 1.4 \end{bmatrix}$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1.4 \end{bmatrix}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = 1.4$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X} = 4.5 - (2.5)(1.4) = 1$$

Matrices: Multiple Linear Regression

$$\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times p+1}\,\boldsymbol{\beta}_{p+1\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$