

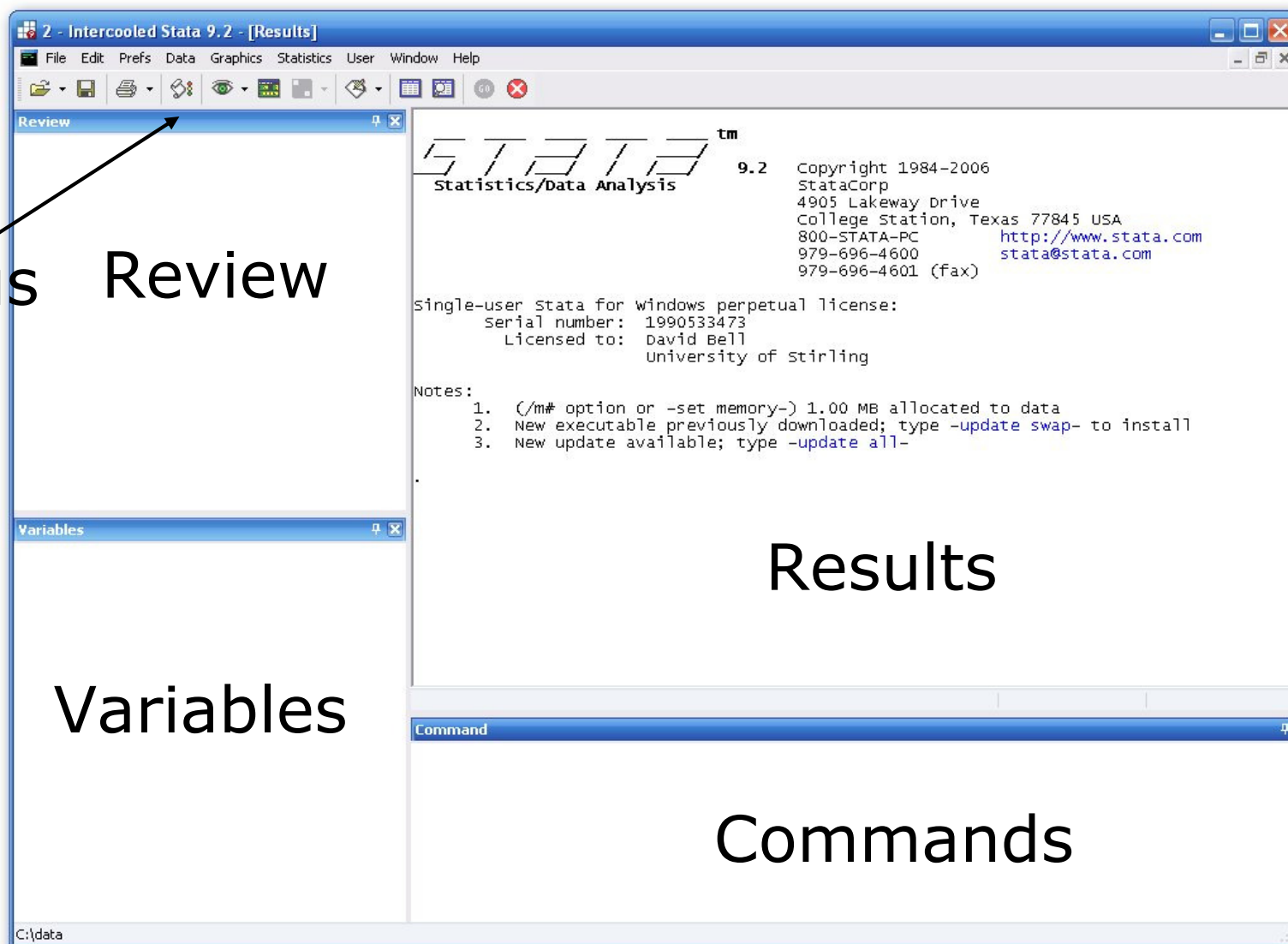
Introduction to Stata

David Bell

Department of Economics
University of Stirling

The Stata Interface

Menus Review



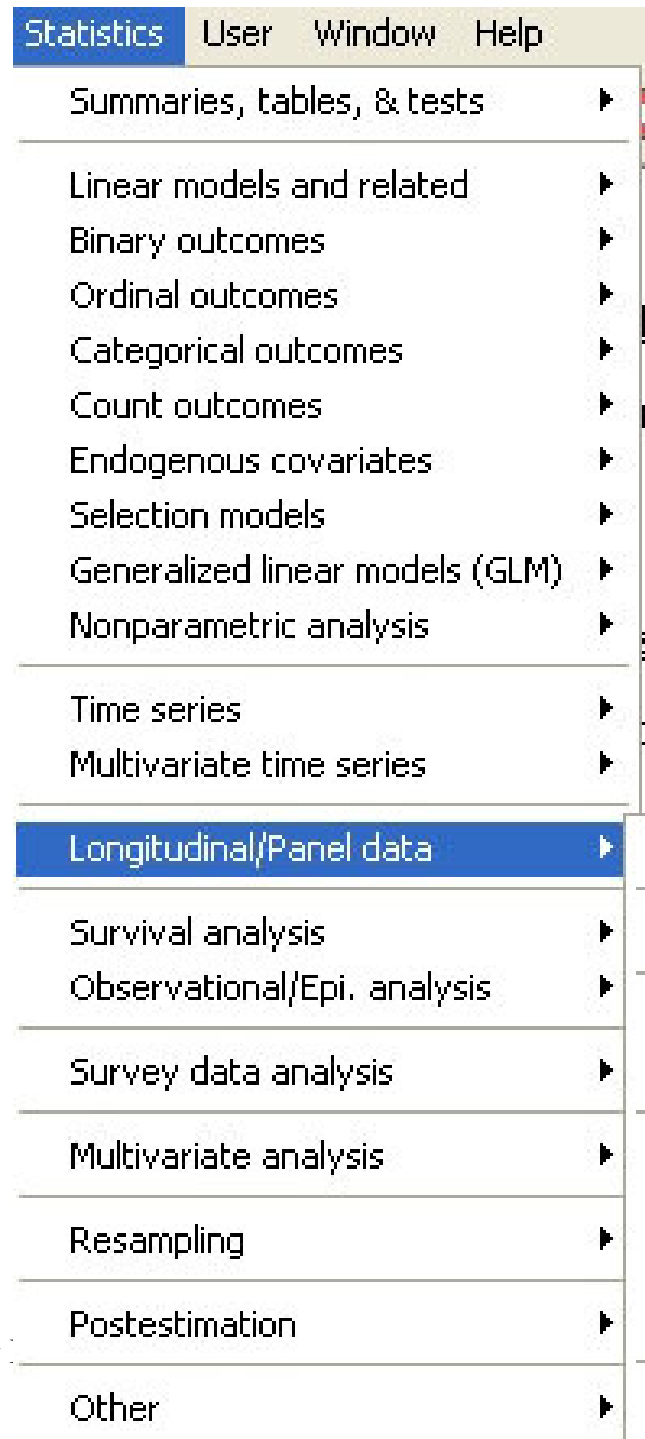
Variables

Results

Commands

Introduction to Stata

What can
Stata do?



What can Stata do?

Linear regression

Regression diagnostics

ANOVA

Box-Cox regression

Errors-in-variables regression

Frontier models

Truncated regression

Constrained linear regression

Multiple equation models

Censored regression

Fractional polynomials

Other

What can Stata do?

Setup & utilities ▶

ARIMA models

ARCH/GARCH ▶

Prais-Winsten regression

Regression with Newey-West std. errors

Rolling windows estimation

Smoothers/univariate forecasters ▶

Tests ▶

Graphs ▶

What can Stata do?

Setup & utilities ▶

Vector autoregression (VAR)

Basic VAR

Structural vector autoregression (SVAR)

Vector error-correction model (VECM)

Cointegrating rank of a VECM

VAR diagnostics and tests ▶

VEC diagnostics and tests ▶

Dynamic forecasts ▶

IRF & FEVD analysis ▶

Manage IRF results and files ▶

What can Stata do?

Setup & utilities ▶

Linear models ▶

Multilevel mixed-effects linear regression
Random coefficients regression by GLS

Endogenous covariates ▶

Dynamic panel data ▶

Contemporaneous correlation ▶

Frontier models

Binary outcomes ▶

Count outcomes ▶

Censored outcomes ▶

Generalized estimating equations (GEE) ▶

Line plots

Getting Help

- Known commands

`help` tabulate from the Viewer command line:

- You need not know the command name. Get information about nonparametric tests:

`search nonparametric` from the Viewer command line

- To search Stata and the net for information on goodness-of-fit tests with panel estimators:

`findit panel goodness`

Files in Stata

- .dta - Stata dataset
 - .ado - Stata program
 - .do - Stata command file
 - .smcl, .log - Stata output files
 - .gph - Stata graph file
-
- Note also that Stata can read and write from/to .csv files (spreadsheets)

do and ado files

- *do setofcommands.do*
 - executes a set of commands stored in the text file *setofcommands.do*
- *program*
 - executes the *program* stored in the text file *program.ado*
 - almost all Stata commands are themselves ado files
 - this structure contributes hugely to the extensibility of Stata
 - *update regularly* to get the most recent versions of ado files and executables

log and smcl files

- Results are immediately available in the results window and can be copied and pasted into a spreadsheet
- log and smcl files are used when a lot of output material is being generated
 - log using mylog - opens smcl file
 - log using mylog.log - opens log (text) file
 - log using mylog, replace -
 - replaces previous version of mylog
 - log using mylog, append
 - appends to previous version of mylog
 - log close
 - pause/restart logs using log on and log off

Starting Point: A Rectangular Dataset

N Observations \rightarrow

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1K} \\ x_{21} & x_{22} & \dots & x_{2K} \\ \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{NK} \end{pmatrix}$$

\uparrow
 K Variables

Loading dataset, saving, clearing

Allocate memory to Stata dataset

- `set mem 100m`

Load Stata-format dataset

- `use [varlist] [if] [in] using filename [, clear nolabel]`

Save data in memory to file

- `save [filename] [, save_options]`

Clearing dataset from memory

- `clear`

Variable Names

- Variable names can be 1-32 characters, but Stata often abbreviates long variable names in output
- The letters **a-z**, the numbers **0-9** and **_** (underscore) are valid characters.
- Names must start with a letter (or an underscore, but because many Stata-generated variables also start with an underscore).
- These are valid variable names: **q17 q_17 pregnant sex**

Stata is case-sensitive!

- Variable names may include lowercase and uppercase letters, but Stata is case-sensitive: **sex** and **Sex** are two different variable names. Best stick to lowercase.

Numeric Types

- Most often you don't need worry about numeric types, but they can be relevant, particularly if you run out of memory. Standard types are:

| Type | Bytes | Precision (digits) | | Range (approx.) |
|----------------|---------------|--------------------|----|---------------------|
| Integer | byte | 1 | 2 | ± 100 |
| | int | 2 | 4 | $\pm 32,000$ |
| | long | 4 | 9 | $\pm 2 \times 10^9$ |
| Floating point | | | | |
| | float | 4 | 7 | $\pm 10^36$ |
| | double | 8 | 16 | $\pm 10^{308}$ |

- Sometimes your data will be in string format and you will have to use the [encode](#) command to convert it to numeric
- [compress](#) can reduce the size of your dataset considerably by finding the most economical way of storage.

Stata command syntax

[prefix :] command [varlist] [=exp] [if] [in] [weight]
[using filename] [, options]

[] implies that the enclosed arguments are *optional*
The **weight** option *requires* square brackets

A **varlist** is a list of variable names with blanks in between. There are a number of shorthand conventions to reduce the amount of typing. e.g.

| | |
|-----------------------|-------------------------------|
| myvar | just one variable |
| myvar thisvar thatvar | three variables |
| myvar* | variables starting with myvar |
| *var | variables ending with var |
| my1-my4 | my1, my2, my3 and my4 |

if

Syntax:

- command if exp
- exp in the syntax diagram means an expression

Examples:

- list make mpg if mpg>25
- list make mpg if mpg>25 & mpg<30
- list make mpg if mpg>25 | mpg<10
- regress mpg weight displ if foreign==1

Relational expressions (used in **if** expressions)

| | |
|--------------------|-------------------------------|
| <code>==</code> | - is equal to |
| <code><=</code> | - is less than or equal to |
| <code><</code> | - is less than |
| <code>></code> | - is greater than |
| <code>>=</code> | - is greater than or equal to |
| <code>~=</code> | - is not equal to |
| <code>&</code> | - logical and |
| <code> </code> | - logical or |

in

Syntax:

- command in range
- Range in the expression means a *range* of observations

Examples:

- list in 10
- list in 10/20
- list in 20/l (lowercase l at end of range)
- list in 1/10 (numeric 1 in beginning of range)

weight

Most Stata commands can deal with weighted data. Stata allows four kinds of weights:

- `fweights`, or frequency weights indicate the number of duplicated observations.
- `pweights`, or sampling weights denote the inverse of the probability that the observation is included due to the sampling design.
- `awweights`, or analytic weights, are inversely proportional to the variance of an observation; i.e., the variance of the j -th observation is assumed to be σ^2/w_j , where w_j are the weights. Typically, the observations represent averages and the weights are the number of elements that gave rise to the average.
- `iweights`, or importance weights, are weights that indicate the "importance" of the observation in some vague sense.

numlist

Some commands also require a numlist. This is sometimes shown in syntax diagrams as #list. is a list of numbers with blanks or commas in between. Conventions to reduce their size include:

1/3 three numbers, 1, 2, 3

3/1 the same three numbers in
 reverse order

1 6 8/12 seven numbers 1 6 8 9 10 11 12

1 2 to 4 four numbers, 1, 2, 3, 4

1(2)9 five numbers, 1, 3, 5, 7, 9

Describing the dataset

```
. des
```

```
Contains data from C:\Documents and Settings\All Users\Documents\Data\lf
> q2\qlfsjm06.dta
  obs:      124,223
 vars:       676
size:  114,285,160 (45.5% of memory free)
```

| variable name | storage type | display format | value label | variable label |
|-----------------|-----------------|-------------------|----------------|--|
| caseno | str14 | %14s | | case identifier |
| quota | int | %8.0g | | stint number where intervi took place |
| week | byte | %8.0g | week | week number when interview place |
| w1yr | byte | %8.0g | w1yr | year that address first en survey |
| qrtr | byte | %8.0g | qrtr | quarter that address first entered survey |
| variabl0 | byte | %8.0g | variabl0 | address number on intervie address list |
| wavfnd | byte | %8.0g | wavfnd | wave at which household wa first found |
| hhld | byte | %8.0g | hhld | household reference |

Editing the dataset

Command is
edit

or

Window
Data editor

Dataset is
the LFS

Data Editor

Preserve Restore Sort << >> Hide Delete...

caseno [3] = 101520110102

| | caseno | persno | recno | sex | dobm | doby | age | hallres | marsta | marchk | liwvth | hrpid | xr |
|----|--------------|--------|-------|--------|------|------|-----|---------|----------|--------|--------|-------|------|
| 1 | 101510110101 | 1 | 1 | female | 3 | 1933 | 72 | no | widowed | . | . | yes | |
| 2 | 101520110101 | 1 | 1 | male | 10 | 1938 | 67 | no | married, | yes | . | yes | |
| 3 | 101520110102 | 2 | 2 | female | 9 | 1940 | 65 | no | married, | yes | . | yes | sp |
| 4 | 101520210101 | 1 | 1 | male | 12 | 1929 | 76 | no | single, | . | . | yes | |
| 5 | 101530120101 | 1 | 1 | male | 10 | 1965 | 40 | no | married, | yes | . | yes | |
| 6 | 101530120102 | 2 | 2 | female | 11 | 1956 | 49 | no | married, | yes | . | yes | sp |
| 7 | 101530210101 | 1 | 1 | male | 7 | 1957 | 48 | no | married, | yes | . | yes | |
| 8 | 101530210102 | 2 | 2 | female | 2 | 1973 | 33 | no | married, | yes | . | no | sp |
| 9 | 101530310101 | 1 | 1 | male | 11 | 1944 | 61 | no | married, | yes | . | yes | |
| 10 | 101530310102 | 2 | 2 | female | 5 | 1950 | 55 | no | married, | yes | . | yes | sp |
| 11 | 101530310103 | 3 | 3 | male | 4 | 1981 | 24 | no | single, | . | no | no | natu |
| 12 | 101540110101 | 1 | 1 | female | 10 | 1953 | 52 | no | divorced | . | no | yes | |
| 13 | 101540110102 | 2 | 2 | male | 12 | 1986 | 19 | no | single, | . | no | no | natu |
| 14 | 101540110103 | 3 | 3 | female | 11 | 1988 | 17 | no | single, | . | no | no | natu |
| 15 | 101540310101 | 1 | 1 | female | 7 | 1953 | 52 | no | single, | . | . | yes | |
| 16 | 101540410101 | 1 | 1 | female | 4 | 1967 | 38 | no | married, | . | no | yes | |
| 17 | 101540410102 | 2 | 2 | male | 2 | 2002 | 4 | . | single, | . | . | . | natu |
| 18 | 101540410103 | 3 | 3 | male | 12 | 1987 | 18 | no | single, | . | no | no | natu |
| 19 | 101540410104 | 4 | 4 | male | 7 | 1990 | 15 | . | single, | . | . | . | natu |
| 20 | 101610110101 | 1 | 1 | male | 10 | 1968 | 37 | no | married, | yes | . | yes | |
| 21 | 101610110102 | 2 | 2 | female | 11 | 1961 | 44 | no | married, | yes | . | yes | sp |
| 22 | 101610110103 | 3 | 3 | female | 9 | 1989 | 16 | no | single, | . | no | yes | natu |
| 23 | 102510310101 | 1 | 1 | male | 8 | 1921 | 84 | no | married, | yes | . | yes | |
| 24 | 102510310102 | 2 | 2 | female | 3 | 1920 | 85 | no | married, | yes | . | yes | sp |
| 25 | 102510410101 | 1 | 1 | male | 8 | 1944 | 61 | no | married, | yes | . | yes | |
| 26 | 102510410102 | 2 | 2 | female | 8 | 1948 | 57 | no | married, | yes | . | yes | sp |
| 27 | 102510410103 | 3 | 3 | male | 4 | 1972 | 33 | no | single, | . | no | no | natu |
| 28 | 102520210101 | 1 | 1 | male | 1 | 1939 | 67 | no | married, | yes | . | yes | |
| 29 | 102520210102 | 2 | 2 | female | 11 | 1940 | 65 | no | married, | yes | . | yes | sp |
| 30 | 102520330101 | 1 | 1 | male | 4 | 1978 | 27 | no | single, | . | yes | no | |
| 31 | 102520330102 | 2 | 2 | female | 9 | 1977 | 28 | no | single, | . | yes | yes | coha |
| 32 | 102520330103 | 3 | 3 | female | 5 | 1979 | 26 | no | single, | . | no | yes | othe |

start 2 Micros... 2 Micros... Stata Course Adobe Rea... 2 Stata Adobe Pho... EN 11:04

Describing variables

Describe data in memory or in file

- describe [varlist] [,describe_m_options]

Note: commands can be shortened to the *underlined* letters

Examples:

- des party_member
- des party_member, det
- des party_member, nol

Tabulating data

Creating one-way and two-way tables

- **tabulate varname [if] [in] [weight] [, tabulate1_options]**
- **tabulate varname1 varname2 [if] [in] [weight] [, options]**

Example: one-way tables

- `tabulate foreign`
- `tabulate region [aweight=pop]`

Example: two-way tables

- `tabulate foreign rep78`
- `tabulate region citysize [aweight=pop]`

Tabulate option: summarize

Report summary statistics for one variable using the categories of another

Example: one-way tables

- `tabulate foreign, summarize(mpg)`
- `tabulate region [aweight=pop], summarize(age)`

Example: two-way tables

- `tabulate foreign rep78, sum(mpg)`
- `tabulate region citysize [aw=pop], sum(age)`

Control tabular output with tab options

Show percentages by column

- tabulate foreign rep78, col

Show percentages by cell

- tabulate region citysize [aw=pop], cell

Show percentages by row and do not
show frequencies

- tab region citysize, row nofreq

Prefix option: by

by makes a command operate on subgroups of the data. Data must be pre-sorted e.g.

sort sex

by sex: summarize age height weight

or, in one line:

bysort sex: summarize age height weight

Create or change contents of variable

Create new variable

- `generate [type] newvar[:lblname] =exp [if]
[in]`

Change contents of existing variable

- `replace oldvar =exp [if] [in] [,nopromote]`

```
gen agesq = age^2
```

```
replace tenure = tenure + 4 if age < 42
```

Functions to use with generate, replace

- man function

| Type of function | see help |
|---|--|
| Mathematical functions Probability distributions and density functions Random-number functions String functions Programming functions Date functions Time-series functions Matrix functions | <code>math functions</code> <code>density functions</code> <code>random-number functions</code> <code>string functions</code> <code>programming functions</code> <code>date functions</code> <code>time-series functions</code> <code>matrix functions</code> |

- `gen p = min(y)`
- `gen f = normal(z)`

Generating dummy variables with tabulate

- One-way tabulate with gen option
- e.g. `tab sex, gen(gender)`

Produces indicator (dummy) variables for each category of sex. These will be named `gender1` and `gender2`

Recoding data

- **recode** changes the values of a variable – to produce new groupings or to transform a continuous variable into dummy variables.

```
recode age (55/max=3)(15/55=2)(min/15=1) ,  
gen(agegr)
```

- Value labels for the new variable may be included at once:

```
recode age (55/max=3 "55+")(15/55=3 "15-54") ///  
– (min/15=1 "0-14") , gen(agegr)
```

- The **generate** option creates a new recoded variable; without **generate** the original information in age will be destroyed.

Missing Values

- Missing values are omitted from calculations.
- The *system missing value* is shown as a . (period). It is created in input when a numeric field is empty; by invalid calculations, e.g. division by 0, or calculations involving a missing value.
- Unfortunately no data entry program accepts . in a numeric field. You might choose the code **-9** and ask Stata to recode them:
- ***recode _all (-9=.)***
- Missing values are large positive numbers – this is very important for calculating conditions.

Regression

- `regress depvar [indepvars] [if] [in] [weight] [, options]`
- Options
 - `noconstant`
 - `robust`
 - `cluster(var)`

More Regression Commands

| | |
|---------------------|--|
| areg | an easier way to fit regressions with many dummy variables |
| arch | regression models with ARCH errors |
| arma | ARIMA models |
| boxcox | Box-Cox regression models |
| cnreg | censored-normal regression |
| cnsreg | constrained linear regression |
| eivreg | errors-in-variables regression |
| frontier | stochastic frontier models |
| heckman | Heckman selection model |
| intreg | interval regression |
| ivreg | instrumental variables (2SLS) regression |
| ivtobit | tobit regression with endogenous variables |
| newey | regression with Newey-west standard errors |
| qreg | quantile (including median) regression |
| reg3 | three-stage least-squares (3SLS) regression |
| rreg | a type of robust regression |
| sureg | seemingly unrelated regression |
| svy: heckman | Heckman selection model with survey data |

Even More Regression Commands!

| | |
|---------------------|---|
| svy: heckman | Heckman selection model with survey data |
| svy: intreg | interval regression with survey data |
| svy: ivreg | instrumental variables regression with survey data |
| svy: regress | linear regression with survey data |
| tobit | tobit regression |
| treatreg | treatment-effects model |
| truncreg | truncated regression |
| xtabond | Arellano-Bond linear, dynamic panel-data estimator |
| xtfrontier | panel-data stochastic frontier model |
| xtgls | panel-data GLS models |
| xhtaylor | Hausman-Taylor estimator for error-components models |
| xtintreg | panel-data interval regression models |
| xtivreg | panel-data instrumental variables (2SLS) regression |
| xtpcse | OLS or Prais-Winsten models with panel-corrected standard errors |
| xtreg | fixed- and random-effects linear models |
| xtregar | fixed- and random-effects linear models with an AR(1) disturbance |
| xttobit | panel-data tobit models |

And so to panel data ...