Longitudinal Data Analysis in Stata

Heide Jackson
heidej@umd.edu

University of Maryland Population Research Center

August 2019

High Level Things to Know

- Focus will be on structuring data, so that it is suitable for a range of longitudinal analyses.
- Start simple and wide.
- Track time varying and non-time varying measures.
- Make sure time varying variables share a common suffix before moving to long format.

Wide and Long Format

- Two ways of specifying the data structure.
- Wide format has one observation per entity.
- Long format can have multiple observations per entity.
- Data doesn't have to be longitudinal and it can be multi-leveled.

Wide Long Format Examples

Household Person Data

- Wide format: one row per household, separate variables for each person in household.
- Long format: potentially multiple rows per household. Separate rows for each person in the household.

Person Time Data

- Wide format: one row per person, separate variables for each year observed
- Long format: multiple rows per person. Separate rows for each time period observed.

Start Wide

- I recommend starting analyses with the data in wide format.
- Why? Get to know the data at the unit of analysis.
- Make sure time varying variables have a common suffix before going long.
- Generate any time-invariant measures before going long.
- How to check if data is wide:

```
sort id
by id: gen obs=_N
tab obs if obs>1
/*if any observations appear,
data are not wide*/
```

Preparing Data for Long Format

- Restrict the data set to variables to be used in the analysis.
- Make sure time varying variables have a common suffix usually a number (i.e. income1 income2 income 3 if there are three time points).
- Make sure there is a unique id variable.
- Consider making a list of time varying and non-time varying measures like this:

```
local timevar timevar1 timevar2...
local notimevar demo birthyear
```

The Syntax of Reshape

Four key parts:

```
reshape 1. long 2.‘timevar’, 3. i(id) 4. j(time)
```

1. Reshape long takes data from wide to long format. Reshape wide does long to wide format.
2. All time varying variables should be specified. Non-time varying variables will be unchanged but if going from wide to long, they will be copied for all new rows created per entity.
3. i indicates entity id variable, should be unique if data are wide.
4. j will be created variable indicating number of observations per entity if going to long format. If going to wide format, j should reference a variable in the data set that will be converted to a suffix for time varying variables.

Working Example

```
webuse nlswork

keep age msp nev_mar grade birth_yr race
/*define variable types*/
local timevarying age msp nev_mar grade
local nottimevarying birth_yr race

reshape wide ‘timevarying’, i(idcode) j(year)

reshape long ‘timevarying’, i(idcode) j(year)
```

Data are Long, Now what?

- Check that the structure of the data is what you expect given what was observed when data were wide.
- Key to know, is the data balanced (all entities observed same number of times) or unbalanced. One way to do this is to count the number of times each entity is observed.

```
sort id
by id: gen obs=_N
tab _N
/*If more than one value, data are unbalanced*/
```

- Variety of statistical models (fixed effect, random effect, survival analysis, HLM) can be run with data in long format.