Module 15: Multilevel Modelling of Repeated Measures Data

Stata Practical

Fiona Steele
London School of Economics and Political Science

Pre-requisites

- Stata practicals for Modules 3 and 5

If you find this module helpful and wish to cite it in your research, please use the following citation:

Steele, F. (2014). Multilevel Modelling of Repeated Measures Data: Stata Practical. LEMMA VLE Module 15, 1-61.
(http://www.bristol.ac.uk/cmm/learning/course.html).

Contents

P15.1 Repeated Measures Data

P15.1.1 Introduction to physical health functioning dataset

In the first part of this practical we will fit growth curve models to data on health functioning from a study of British civil servants called the Whitehall II study (also known as the Stress & Health Study).[1] Health functioning was assessed by the SF-36, a 36 item instrument that comprises eight subscales covering physical, psychological and social functioning. These eight scales can be summarised into physical and mental health components. These are scaled using general US population norms to have mean values of 50 and low scores imply poor functioning. We will study change in physical health functioning which was measured on up to six occasions for each respondent.

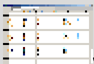
The data are in wide form, i.e. with one record per individual and six variables for health functioning at the six measurement occasions. The dataset also includes information on the respondent's age at each occasion, their employment grade at the first occasion, and their gender. The analysis file contains the following variables for 4427 individuals:

Variable	Description and codes
id	Individual identifier (coded 1, 2, . . ., 8815)
female	Gender (1=female, 0=male)
grade	Employment grade at baseline (1=high, 2=intermediate, 3=low)
age1	Age at occasion 1 (years)
phf1	Physical health functioning score at occasion 1
...	...
age6	Age at occasion 6 (years)
phf6	Physical health functioning score at occasion 6

[1]The data used in the practical were provided by Jenny Head (University College London). See http://www.bristol.ac.uk/cmm/learning/repeated-measures.pdf for an earlier analysis of these data. For further information on the Whitehall II study go to http://www.ucl.ac.uk/whitehallII

Load “15.1.dta” into memory and open the do-file “15.1.do” for this lesson

From within the LEMMA Learning Environment

- Go to **Module 15: Multilevel Modelling of Repeated Measures Data**, and scroll down to  ***Stata* Datasets and Do-files**
- Click “15.1.dta**”** to open the dataset

Use the `describe` command to produce a summary of the dataset:

```
 . describe

Contains data from 15.1.dta
  obs:         4,427
 vars:            15                          30 Jun 2014 13:46
 size:       203,642
-----------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-----------------------------------------------------------------------------------
--------
id              int     %9.0g                 Individual identifier
female          byte    %9.0g
grade           byte    %12.0g     grade      employment grade
age1            byte    %9.0g
phf1            float   %9.0g
age2            byte    %9.0g
phf2            float   %9.0g
age3            float   %9.0g
phf3            float   %9.0g
age4            float   %9.0g
phf4            float   %9.0g
age5            float   %9.0g
phf5            float   %9.0g
age6            float   %9.0g
phf6            float   %9.0g
-----------------------------------------------------------------------------------
Sorted by:
```

We will view selected variables for the first 5 individuals using the `list` command.

```
 . list id female grade age1 phf1 age6 phf6 in 1/5

      +-----------------------------------------------------------------+
      | id   female          grade   age1       phf1       age6       phf6 |
      |-----------------------------------------------------------------|
   1. |  1        1   intermediate     55   39.59072          .          . |
   2. |  2        1            low     56   29.31145   70.95313   22.99401 |
   3. |  3        1            low     53   46.74854          .          . |
   4. |  4        0           high     51   45.78801   64.70313   43.56795 |
   5. |  5        0           high     55   46.77223    69.3125   41.97438 |
      +-----------------------------------------------------------------+
```

We can see that individuals 1 and 3 have missing data for occasion 6. We will obtain a summary of missing data patterns before fitting any models.

P15.1.2 Restructuring data from wide to long form

Our first task is to convert the data from wide form to long form. This can be achieved using the `reshape` command.

```
. reshape long phf age, i(id) j(occ)
(note: j = 1 2 3 4 5 6)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                     4427   ->   26562
Number of variables                  15   ->       6
j variable (6 values)                     ->   occ
xij variables:
                      phf1 phf2 ... phf6   ->   phf
                      age1 age2 ... age6   ->   age
-----------------------------------------------------------------------------
```

The variable list includes the ‘stubnames’ of the time-varying variables: **phf** and **age** (the prefixes of **phf1**, **age1**, . . ., **phf6**, **age6**). After the comma, we specify the person identifier (**id**) and a new variable which will index the repeated measures in the long data file (**occ**, coded 1, 2, . . .6).

The restructured file contains 6 × 4427 = 26562 records and six variables: **id**, **occ**, **female**, **grade**, **age** and **phf**. We label the three new variables:

```
 . label var occ "measurement occasion"

 . label var phf "physical health functioning (from SF-36)"

 . label var age "age (years)"
```

And list records for the first 2 individuals, with a separator line after each individual’s records:

```
. list in 1/12, separator(6)

     +---------------------------------------------------------+
     | id   occ   female          grade         age        phf |
     |---------------------------------------------------------|
  1. |  1     1        1   intermediate          55   39.59072 |
  2. |  1     2        1   intermediate          57   38.61111 |
  3. |  1     3        1   intermediate    60.41615   39.92796 |
  4. |  1     4        1   intermediate    63.64844    21.9101 |
  5. |  1     5        1   intermediate    66.62012   25.65662 |
  6. |  1     6        1   intermediate           .          . |
     |---------------------------------------------------------|
  7. |  2     1        1            low          56   29.31145 |
  8. |  2     2        1            low          60   22.69034 |
  9. |  2     3        1            low    62.13279   24.19915 |
 10. |  2     4        1            low    65.85938   21.40272 |
 11. |  2     5        1            low    67.56468   17.21988 |
 12. |  2     6        1            low    70.95313   22.99401 |
     +---------------------------------------------------------+
```

Two features of the data are immediately apparent:

- There is individual variation in the timing of measurements. For example, individual 1 is age 55 at occasion 1, while individual 2 is age 56.

- The length of time between measurements is not fixed and varies between individuals. For example, for individual 1, there is 2 years between occasions 1 and 2 and 3.42 years between occasions 2 and 3. The corresponding gaps for individual 2 are 4 and 2.13 years.

Later we will obtain summary statistics for the distributions of age and time between measurements across all individuals.

P15.1.3 Summarising longitudinal data

Before proceeding with growth curve analysis, we look at the extent of missing data and the distribution of the time between occasions.

Missing data patterns

We begin with a simple frequency table showing the number of valid (non-missing) values for our response **phf** at each occasion, using the `tabulate` command.

```
 . tabulate occ if phf~=.

measurement |
   occasion |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      4,176       19.77       19.77
          2 |      3,768       17.84       37.60
          3 |      3,403       16.11       53.71
          4 |      3,281       15.53       69.24
          5 |      3,251       15.39       84.63
          6 |      3,246       15.37      100.00
------------+-----------------------------------
      Total |     21,125      100.00
```

The total number of individuals in the dataset is 4427, of whom 4176 (94%) were present at occasion 1, falling to 3246 (73%) at occasion 6.

Next we obtain the number of non-missing observations per individual. The following command counts the number of non-missing values of **phf** for each individual and stores the result in a variable called **numocc**. As we shall see, the `by id` prefix is very useful for deriving variables from longitudinal data in long form.

```
. by id: egen numocc = count(phf)
```

The values of **numocc** are constant across individuals.

```
 . list id occ numocc in 1/12, separator(6)

     +-------------------+
     | id   occ   numocc |
     |-------------------|
  1. |  1     1        5 |
  2. |  1     2        5 |
  3. |  1     3        5 |
  4. |  1     4        5 |
  5. |  1     5        5 |
```

```
   6. |  1     6        5 |
      |-------------------|
   7. |  2     1        6 |
   8. |  2     2        6 |
   9. |  2     3        6 |
  10. |  2     4        6 |
  11. |  2     5        6 |
  12. |  2     6        6 |
      +-------------------+
```

The distribution of **numocc** across individuals must be based on an individual-level file. One way to do this is to create an indicator variable which equals 1 for one of an individual's records and 0 for the others. We will create a variable called **pickone** which equals 1 for the first record (occasion 1) and 0 for the others. We then tabulate **numocc** when **pickone** equals 1 to obtain an individual-level frequency table.

```
 . egen pickone = tag(id)

 . tabulate numocc if pickone ==1

     numocc |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         73        1.65        1.65
          1 |        244        5.51        7.16
          2 |        316        7.14       14.30
          3 |        308        6.96       21.26
          4 |        415        9.37       30.63
          5 |        761       17.19       47.82
          6 |      2,310       52.18      100.00
------------+-----------------------------------
      Total |      4,427      100.00
```

Note that the total number of observations in the tabulation is 4427, which is the number of individuals in the data file. We see that a total of 317 individuals have either completely missing data or only 1 valid response. The individuals with **numocc**=0 will automatically be deleted from any analysis, but we will also delete those with only 1 record because they contribute no information about change in the response.

```
. drop if numocc==0 | numocc==1
(1902 observations deleted)
```

A total of 6 × 317 = 1902 records are dropped from the full dataset.

More detailed information about missing data patterns can be obtained using the `xt` commands,[2] developed for longitudinal analysis. Before issuing an `xt` command, the longitudinal structure of the data must be specified using `xtset`. In its simplest form, the cluster (individual) identifier is declared. We also declare a variable which indexes time (**occ** here). The `xtdescribe` command can then be used to summarise the pattern of non-missing values of **phf** across the six occasions.

```
 . xtset id occ
       panel variable:  id (strongly balanced)
```

[2] `xt` is Stata shorthand for 'cross-sectional time series', a term sometimes used for panel data.