The xt series of commands provides tools for analyzing panel data (also known as longitudinal data or in some disciplines as cross-sectional time series when there is an explicit time component). Panel datasets have the form $x_{it}$, where $x_{it}$ is a vector of observations for unit $i$ and time $t$. The particular commands (such as xtdescribe, xtsum, and xtreg) are documented in alphabetical order in the entries that follow this entry. If you do not know the name of the command you need, try browsing the second part of this description section, which organizes the xt commands by topic. The next section, Remarks and examples, describes concepts that are common across commands.

The xtset command sets the panel variable and the time variable; see [XT] xtset. Most xt commands require that the panel variable be specified, and some require that the time variable also be specified. Once you xtset your data, you need not do it again. The xtset information is stored with your data.

If you have previously tsset your data by using both a panel and a time variable, these settings will be recognized by xtset, and you need not xtset your data.


Setup

**xtset**

Declare data to be panel data

Data management and exploration tools

**xtdescribe**

Describe pattern of xt data

**xtsum**

Summarize xt data

**xttab**

Tabulate xt data

**xtdata**

Faster specification searches with xt data

**xtline**

Panel-data line plots
Introduction to xt commands

Linear regression estimators

- `xtreg` Fixed-, between-, and random-effects, and population-averaged linear models
- `xtregar` Fixed- and random-effects linear models with an AR(1) disturbance
- `xtgls` Panel-data models by using GLS
- `xtpcse` Linear regression with panel-corrected standard errors
- `xthtaylor` Hausman–Taylor estimator for error-components models
- `xtfrontier` Stochastic frontier models for panel data
- `xtrc` Random-coefficients regression
- `xtivreg` Instrumental variables and two-stage least squares for panel-data models

Unit-root tests

- `xtunitroot` Panel-data unit-root tests

Dynamic panel-data estimators

- `xtabond` Arellano–Bond linear dynamic panel-data estimation
- `xtdpd` Linear dynamic panel-data estimation
- `xtdpdsys` Arellano–Bover/Blundell–Bond linear dynamic panel-data estimation

Censored-outcome estimators

- `xttobit` Random-effects tobit models
- `xtintreg` Random-effects interval-data regression models

Binary-outcome estimators

- `xtlogit` Fixed-effects, random-effects, and population-averaged logit models
- `xtprobit` Random-effects and population-averaged probit models
- `xtcloglog` Random-effects and population-averaged cloglog models

Ordinal-outcome estimators

- `xtologit` Random-effects ordered logistic models
- `xtoprobit` Random-effects ordered probit models

Count-data estimators

- `xtpoisson` Fixed-effects, random-effects, and population-averaged Poisson models
- `xtnbreg` Fixed-effects, random-effects, & population-averaged negative binomial models

Generalized estimating equations estimator

- `xtgee` Population-averaged panel-data models by using GEE

Utility

- `quadchk` Check sensitivity of quadrature approximation
Remarks and examples

Consider having data on \( n \) units—individuals, firms, countries, or whatever—over \( T \) periods. The data might be income and other characteristics of \( n \) persons surveyed each of \( T \) years, the output and costs of \( n \) firms collected over \( T \) months, or the health and behavioral characteristics of \( n \) patients collected over \( T \) years. In panel datasets, we write \( x_{it} \) for the value of \( x \) for unit \( i \) at time \( t \). The \( xt \) commands assume that such datasets are stored as a sequence of observations on \((i, t, x)\).

For a discussion of panel-data models, see Baltagi (2013), Greene (2012, chap. 11), Hsiao (2003), and Wooldridge (2010). Cameron and Trivedi (2010) illustrate many of Stata’s panel-data estimators.

> Example 1

If we had data on pulmonary function (measured by forced expiratory volume, or FEV) along with smoking behavior, age, sex, and height, a piece of the data might be

```
  list in 1/6, separator(0) divider
```

<table>
<thead>
<tr>
<th>pid</th>
<th>yr_visit</th>
<th>fev</th>
<th>age</th>
<th>sex</th>
<th>height</th>
<th>smokes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1071</td>
<td>1991</td>
<td>1.21</td>
<td>25</td>
<td>1</td>
<td>69</td>
</tr>
<tr>
<td>2</td>
<td>1071</td>
<td>1992</td>
<td>1.52</td>
<td>26</td>
<td>1</td>
<td>69</td>
</tr>
<tr>
<td>3</td>
<td>1071</td>
<td>1993</td>
<td>1.32</td>
<td>28</td>
<td>1</td>
<td>68</td>
</tr>
<tr>
<td>4</td>
<td>1072</td>
<td>1991</td>
<td>1.33</td>
<td>18</td>
<td>1</td>
<td>71</td>
</tr>
<tr>
<td>5</td>
<td>1072</td>
<td>1992</td>
<td>1.18</td>
<td>20</td>
<td>1</td>
<td>71</td>
</tr>
<tr>
<td>6</td>
<td>1072</td>
<td>1993</td>
<td>1.19</td>
<td>21</td>
<td>1</td>
<td>71</td>
</tr>
</tbody>
</table>

The \( xt \) commands need to know the identity of the variable identifying patient, and some of the \( xt \) commands also need to know the identity of the variable identifying time. With these data, we would type

```
  . xtset pid yr_visit
```

If we resaved the data, we need not respecify \( xtset \).

> Technical note

Panel data stored as shown above are said to be in the long form. Perhaps the data are in the wide form with 1 observation per unit and multiple variables for the value in each year. For instance, a piece of the pulmonary function data might be

```
  pid  sex  fev91  fev92  fev93  age91  age92  age93
  1071 1  1.21  1.52  1.32    25    26    28
  1072 1  1.33  1.18  1.19    18    20    21
```

Data in this form can be converted to the long form by using \texttt{reshape}; see [D] \texttt{reshape}.

> Example 2

Data for some of the periods might be missing. That is, we have panel data on \( i = 1, \ldots, n \) and \( t = 1, \ldots, T \), but only \( T_i \) of those observations are defined. With such missing periods—called unbalanced data—a piece of our pulmonary function data might be

```
  . list in 1/6, separator(0) divider
```
Patient ID 1072 is not observed in 1992. The xt commands are robust to this problem.

Technical note

In many of the entries in [XT], we will use data from a subsample of the NLSY data (Center for Human Resource Research 1989) on young women aged 14–26 years in 1968. Women were surveyed in each of the 21 years 1968–1988, except for the six years 1974, 1976, 1979, 1981, 1984, and 1986. We use two different subsets: \texttt{nlswork.dta} and \texttt{union.dta}.

For \texttt{nlswork.dta}, our subsample is of 4,711 women in years when employed, not enrolled in school and evidently having completed their education, and with wages in excess of $1/hour but less than $700/hour.
. use http://www.stata-press.com/data/r13/nlswork
(National Longitudinal Survey. Young Women 14-26 years of age in 1968)
. describe
Contains data from http://www.stata-press.com/data/r13/nlswork.dta
    obs:     28,534
    National Longitudinal Survey. Young Women 14-26 years of age in 1968
    vars:    21
    size: 941,622
    27 Nov 2012 08:14

    storage  display value
     variable name  type format label variable label

    idcode      int  %8.0g NLS ID
    year        byte  %8.0g interview year
    birth_yr    byte  %8.0g birth year
    age         byte  %8.0g age in current year
    race        byte  %8.0g racelbl race
    msp         byte  %8.0g 1 if married, spouse present
    nev_mar     byte  %8.0g 1 if never married
    grade       byte  %8.0g current grade completed
    collgrad    byte  %8.0g 1 if college graduate
    not_smsa    byte  %8.0g 1 if not SMSA
    c_city      byte  %8.0g 1 if central city
    south       byte  %8.0g 1 if south
    ind_code    byte  %8.0g industry of employment
    occ_code    byte  %8.0g occupation
    union       byte  %8.0g 1 if union
    wks_ue      byte  %8.0g weeks unemployed last year
    ttl_exp     float %9.0g total work experience
    tenure      float %9.0g job tenure, in years
    hours       int   %8.0g usual hours worked
    wks_work    int   %8.0g weeks worked last year
    ln_wage     float %9.0g ln(wage/GNP deflator)

Sorted by: idcode year
Many of the variables in the nlswork dataset are indicator variables, so we have used factor variables (see [U] 11.4.3 Factor variables) in many of the examples in this manual. You will see terms like c.age#c.age or 2.race in estimation commands. c.age#c.age is just age interacted with age, or age-squared, and 2.race is just an indicator variable for black (race = 2).

Instead of using factor variables, you could type

```
  . generate age2 = age*age
generate black = (race==2)
```

and substitute age2 and black in your estimation command for c.age#c.age and 2.race, respectively.

There are advantages, however, to using factor variables. First, you do not actually have to create new variables, so the number of variables in your dataset is less.

Second, by using factor variables, we are able to take better advantage of postestimation commands. For example, if we specify the simple model

```
  . xtreg ln_wage age age2, fe
```

then age and age2 are completely separate variables. Stata has no idea that they are related—that one is the square of the other. Consequently, if we compute the average marginal effect of age on the log of wages,

```
  . margins, dydx(age)
```

then the reported marginal effect is with respect to the age variable alone and not with respect to the true effect of age, which involves the coefficients on both age and age2.

If instead we fit our model using an interaction of age with itself for the square of age,

```
  . xtreg ln_wage age c.age#c.age, fe
```
then Stata has a deep understanding that the coefficients \texttt{age} and \texttt{c.age#c.age} are related. After fitting this model, the marginal effect reported by \texttt{margins} includes the full effect of age on the log of income, including the contribution of both coefficients.

\begin{verbatim}
. margins, dydx(age)
\end{verbatim}

There are other reasons for preferring factor variables; see \cite{margins} for examples.

For \texttt{union.dta}, our subset was sampled only from those with union membership information from 1970 to 1988. Our subsample is of 4,434 women. The important variables are \texttt{age} (16–46), \texttt{grade} (years of schooling completed, ranging from 0 to 18), \texttt{not_smsa} (28\% of the person-time was spent living outside a standard metropolitan statistical area (SMSA)), and \texttt{south} (41\% of the person-time was in the South). The dataset also has variable \texttt{union}. Overall, 22\% of the person-time is marked as time under union membership, and 44\% of these women have belonged to a union.

\begin{verbatim}
. use http://www.stata-press.com/data/r13/union
   (NLS Women 14–24 in 1968)
. describe
Contains data from http://www.stata-press.com/data/r13/union.dta
    obs:       26,200 NLS Women 14–24 in 1968
    vars:        8 4 May 2013 13:54
    size:    235,800

variable name type storage display value label
  variable label
idcode     int     %8.0g     NLS ID
year     byte     %8.0g     interview year
age     byte     %8.0g     age in current year
grade     byte     %8.0g     current grade completed
not_smsa     byte     %8.0g     1 if not SMSA
south     byte     %8.0g     1 if south
union     byte     %8.0g     1 if union
black     byte     %8.0g     race black

Sorted by: idcode year
. summarize

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>idcode</td>
<td>26200</td>
<td>2611.582</td>
<td>1484.994</td>
<td>1</td>
<td>5159</td>
</tr>
<tr>
<td>year</td>
<td>26200</td>
<td>79.47137</td>
<td>5.965499</td>
<td>70</td>
<td>88</td>
</tr>
<tr>
<td>age</td>
<td>26200</td>
<td>30.43221</td>
<td>6.489056</td>
<td>16</td>
<td>46</td>
</tr>
<tr>
<td>grade</td>
<td>26200</td>
<td>12.76145</td>
<td>2.411715</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td>not_smsa</td>
<td>26200</td>
<td>.2837023</td>
<td>.4508027</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

south    | 26200 | .4130153  | .4923849 | 0   | 1   |
union    | 26200 | .2217939  | .4154611 | 0   | 1   |
black    | 26200 | .274542   | .4462917 | 0   | 1   |
\end{verbatim}

In many of the examples where the \texttt{union} dataset is used, we also include an interaction between the \texttt{year} variable and the \texttt{south} variable—\texttt{south#c.year}. This interaction is created using factor-variables notation; see \cite{factor-variables}

With both datasets, we have typed

\begin{verbatim}
. xtset idcode year
\end{verbatim}
Technical note

The \texttt{xtset} command sets the $t$ and $i$ index for xt data by declaring them as characteristics of the data; see \cite{char}. The panel variable is stored in \_dta[iis] and the time variable is stored in \_dta[tis].

Technical note

Throughout the entries in [XT], when random-effects models are fit, a likelihood-ratio test that the variance of the random effects is zero is included. These tests occur on the boundary of the parameter space, invalidating the usual theory associated with such tests. However, these likelihood-ratio tests have been modified to be valid on the boundary. In particular, the null distribution of the likelihood-ratio test statistic is not the usual $\chi^2_1$ but is rather a 50:50 mixture of a $\chi^2_0$ (point mass at zero) and a $\chi^2_1$, denoted as $\chi^2_{01}$. See Gutierrez, Carter, and Drukker (2001) for a full discussion.

References

Cameron, A. C., and P. K. Trivedi. 2010. 	extit{Microeconometrics Using Stata}. Rev. ed. College Station, TX: Stata Press.

Also see

\[\text{[XT]} \texttt{xtset} \quad \text{— Declare data to be panel data}\]