

## Title

**xt** — Introduction to xt commands

## Syntax

`xtcmd ...`

## Description

The xt series of commands provide 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  $\mathbf{x}_{it}$ , where  $\mathbf{x}_{it}$  is a vector of observations for unit  $i$  and time  $t$ . The particular commands (such as `xtdescribe`, `xtsum`, and `xtreg`) are documented in the entries that follow this entry. This entry deals with 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.

If your interest is in general time-series analysis, see [U] **26.14 Models with time-series data** and the *Stata Time-Series Reference Manual*.

## Remarks

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 (2005), Greene (2003), Hsiao (2003), or Wooldridge (2002).

### ▷ 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
```

	pid	yr_visit	fev	age	sex	height	smokes
1.	1071	1991	1.21	25	1	69	0
2.	1071	1992	1.52	26	1	69	0
3.	1071	1993	1.32	28	1	68	0
4.	1072	1991	1.33	18	1	71	1
5.	1072	1992	1.18	20	1	71	1
6.	1072	1993	1.19	21	1	71	0

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 reshape; see [D] reshape.



### ▷ Example 2

Data for some of the periods might be missing. That is, we have panel data on  $i = 1, \dots, n$  and  $t = 1, \dots, T$ , 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
```

	pid	yr_visit	fev	age	sex	height	smokes
1.	1071	1991	1.21	25	1	69	0
2.	1071	1992	1.52	26	1	69	0
3.	1071	1993	1.32	28	1	68	0
4.	1072	1991	1.33	18	1	71	1
5.	1072	1993	1.19	21	1	71	0
6.	1073	1991	1.47	24	0	64	0

Patient ID 1072 is not observed in 1992. The xt commands are robust to this problem.



### □ Technical Note

In many of the [XT] xt entries, 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: nlswork.dta and union.dta.

For 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/r10/nlswork
(National Longitudinal Survey. Young Women 14-26 years of age in 1968)
```

```
. describe
Contains data from http://www.stata-press.com/data/r10/nlswork.dta
obs:      28,534                               National Longitudinal Survey.
                                                Young Women 14-26 years of age
                                                in 1968
vars:      21                                  7 Dec 2006 17:02
size:      1,055,758 (89.9% of memory free)
```

variable name	storage type	display format	value 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		1=white, 2=black, 3=other
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

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
idcode	28534	2601.284	1487.359	1	5159
year	28534	77.95865	6.383879	68	88
birth_yr	28534	48.08509	3.012837	41	54
age	28510	29.04511	6.700584	14	46
race	28534	1.303392	.4822773	1	3
msp	28518	.6029175	.4893019	0	1
nev_mar	28518	.2296795	.4206341	0	1
grade	28532	12.53259	2.323905	0	18
collgrad	28534	.1680451	.3739129	0	1
not_smsa	28526	.2824441	.4501961	0	1
c_city	28526	.357218	.4791882	0	1
south	28526	.4095562	.4917605	0	1
ind_code	28193	7.692973	2.994025	1	12
occ_code	28413	4.777672	3.065435	1	13
union	19238	.2344319	.4236542	0	1
wks_ue	22830	2.548095	7.294463	0	76
ttl_exp	28534	6.215316	4.652117	0	28.88461
tenure	28101	3.123836	3.751409	0	25.91667
hours	28467	36.55956	9.869623	1	168
wks_work	27831	53.98933	29.03232	0	104
ln_wage	28534	1.674907	.4780935	0	5.263916

For `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 `age` (16–46), `grade` (years of schooling completed, ranging from 0 to 18), `not_smsa` (28% of the person-time was spent living outside an SMSA—standard metropolitan statistical area), `south` (41% of the person-time was in the South), and `southXt` (`south` interacted with year, treating 1970 as year 0). The dataset also has variable `union`. Overall, 22% of the person-time is marked as time under union membership, and 44% of these women have belonged to a union.

```
. use http://www.stata-press.com/data/r10/union
(NLS Women 14-24 in 1968)

. describe

Contains data from http://www.stata-press.com/data/r10/union.dta
obs:      26,200                NLS Women 14-24 in 1968
vars:      10                  27 Oct 2006 13:51
size:      393,000 (96.3% of memory free)
```

variable name	storage type	display format	value label	variable label
<code>idcode</code>	int	%8.0g		NLS ID
<code>year</code>	byte	%8.0g		interview year
<code>age</code>	byte	%8.0g		age in current year
<code>grade</code>	byte	%8.0g		current grade completed
<code>not_smsa</code>	byte	%8.0g		1 if not SMSA
<code>south</code>	byte	%8.0g		1 if south
<code>union</code>	byte	%8.0g		1 if union
<code>t0</code>	byte	%9.0g		
<code>southXt</code>	byte	%9.0g		
<code>black</code>	byte	%8.0g		race black

Sorted by:

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
<code>idcode</code>	26200	2611.582	1484.994	1	5159
<code>year</code>	26200	79.47137	5.965499	70	88
<code>age</code>	26200	30.43221	6.489056	16	46
<code>grade</code>	26200	12.76145	2.411715	0	18
<code>not_smsa</code>	26200	.2837023	.4508027	0	1
<code>south</code>	26200	.4130153	.4923849	0	1
<code>union</code>	26200	.2217939	.4154611	0	1
<code>t0</code>	26200	9.471374	5.965499	0	18
<code>southXt</code>	26200	3.96874	6.057208	0	18
<code>black</code>	26200	.274542	.4462917	0	1

With both datasets, we have typed

```
. xtset idcode year
```

□

## □ Technical Note

The `xtset` command sets the  $t$  and  $i$  index for `xt` data by declaring them as characteristics of the data; see [P] **char**. The panel variable is stored in `__dta[iis]` and the time variable is stored in `__dta[tis]`.

□

**□ Technical Note**

`xtmixed`, `xtmelogit`, and `xtmepoisson` do not use the information pertaining to  $i$  and  $t$  that is stored by `xtset`. Unlike the other **xt** commands, these can handle multiple nested levels of groups and thus use their own syntax for specifying the group structure of the data. □

**□ Technical Note**

Throughout the **xt** entries, when random-effects models are fitted, 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  $\chi_1^2$  but is rather a 50:50 mixture of a  $\chi_0^2$  (point mass at zero) and a  $\chi_1^2$ , denoted as  $\bar{\chi}_{01}^2$ . See Gutierrez, Carter, and Drukker (2001) for a full discussion, and see [XT] **xtmixed** for a generalization of the concept as applied to variance-component estimation in mixed models. □

**References**

- Baltagi, B. H. 2005. *Econometric Analysis of Panel Data*. 3rd ed. New York: Wiley.
- Center for Human Resource Research. 1989. *National Longitudinal Survey of Labor Market Experience, Young Women 14–26 years of age in 1968*. Columbus, OH: Ohio State University Press.
- Greene, W. H. 2003. *Econometric Analysis*. 5th ed. Upper Saddle River, NJ: Prentice Hall.
- Gutierrez, R. G., S. L. Carter, and D. M. Drukker. 2001. sg160: On boundary-value likelihood-ratio tests. *Stata Technical Bulletin* 60: 15–18. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 269–273.
- Hsiao, C. 2003. *Analysis of Panel Data*. 2nd ed. New York: Cambridge University Press.
- Wooldridge, J. M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

**Also See**

[XT] **xtset** — Declare data to be panel data