Description   Remarks and examples   References   Also see

Description

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 $\mathbf{x}_{it}$, 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 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.

If your interest is in general time-series analysis, see **[U] 27.14 Time-series models** and the *Time-Series Reference Manual*. If your interest is in multilevel mixed-effects models, see **[U] 27.16 Multilevel mixed-effects models** and the *Multilevel Mixed-Effects Reference Manual*. If you are interested in SAR (spatial autoregressive or simultaneously autoregressive) models for panel data, see [SP] **spxtregress**. If you are interested in extended panel-data regression models that address endogenous covariates, nonrandom treatment assignment, and endogenous sample selection, see the *Extended Regression Models Reference Manual*. If you are interested in the mixed logit choice model for panel data, see [CM] **cmxtmixlogit**.

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

Linear regression estimators

xtreg	Linear models for panel data
xtregar	Fixed- and random-effects linear models with an AR(1) disturbance
xtgls	GLS linear model with heteroskedastic and correlated errors
xtpcse	Linear regression with panel-corrected standard errors
xthtaylor	Hausman–Taylor estimator for error-components model
xtfrontier	Stochastic frontier models for panel data
xtrc	Random-coefficients model
xtivreg	Instrumental variables and two-stage least squares for panel-data models
xtheckman	Random-effects regression with sample selection
xtdidregress	Fixed-effects difference in differences
xthdidregress	Heterogeneous difference in differences for panel data
xteregress	Random-effects models with endogenous covariates, treatment, and sample selection

Unit-root tests

xtunitroot	Panel-data unit-root tests

Cointegration tests

xtcointtest	Panel-data cointegration 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
xtvar	Panel-data vector autoregressive models

Censored-outcome estimators

xttobit	Random-effects tobit model
xtintreg	Random-effects interval-data regression model
xteintreg	Random-effects interval-data regression models with endogenous covariates, treatment, and sample selection

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
xteprobit	Random-effects probit models with endogenous covariates, treatment, and sample selection

Ordinal-outcome estimators

xtologit	Random-effects ordered logistic model
xtoprobit	Random-effects ordered probit model
xteoprobit	Random-effects ordered probit models with endogenous covariates, treatment, and sample selection

Categorical-outcome estimators

xtmlogit	Fixed-effects and random-effects multinomial logit models
cmxtmixlogit	Panel-data mixed logit choice model

Count-data estimators

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

Survival-time estimators

xtstreg	Random-effects parametric survival models

Generalized estimating equations estimator

xtgee	GEE population-averaged panel-data models

Spatial autoregressive or simultaneously autoregressive estimator

spxtregress	Spatial autoregressive models for panel data

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 (2018, chap. 11), Hsiao (2014), and Wooldridge (2010). Cameron and Trivedi (2022) illustrate many of Stata's panel-data estimators.

For an introduction to linear, nonlinear, and dynamic panel-data analysis in Stata, we offer NetCourse 471, *Introduction to Panel Data Using Stata*; see https://www.stata.com/netcourse/panel-data-intro-nc471/.

▷ 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 $i = 1, \ldots, n$ 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
```

	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 entries in [XT], we will use data from a subsample of the NLSY data (Center for Human Resource Research 1989) on young women aged 14–24 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 https://www.stata-press.com/data/r19/nlswork, clear
(National Longitudinal Survey of Young Women, 14-24 years old in 1968)

. describe

Contains data from https://www.stata-press.com/data/r19/nlswork.dta
 Observations:        28,534                  National Longitudinal Survey of
                                                Young Women, 14-24 years old in
                                                1968
    Variables:            21                  27 Nov 2024 08:14
                                              (_dta has notes)
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    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
```

```
. summarize
```

Variable	Obs	Mean	Std. dev.	Min	Max
idcode	28,534	2601.284	1487.359	1	5159
year	28,534	77.95865	6.383879	68	88
birth_yr	28,534	48.08509	3.012837	41	54
age	28,510	29.04511	6.700584	14	46
race	28,534	1.303392	.4822773	1	3
msp	28,518	.6029175	.4893019	0	1
nev_mar	28,518	.2296795	.4206341	0	1
grade	28,532	12.53259	2.323905	0	18
collgrad	28,534	.1680451	.3739129	0	1
not_smsa	28,526	.2824441	.4501961	0	1
c_city	28,526	.357218	.4791882	0	1
south	28,526	.4095562	.4917605	0	1
ind_code	28,193	7.692973	2.994025	1	12
occ_code	28,413	4.777672	3.065435	1	13
union	19,238	.2344319	.4236542	0	1
wks_ue	22,830	2.548095	7.294463	0	76
ttl_exp	28,534	6.215316	4.652117	0	28.88461
tenure	28,101	3.123836	3.751409	0	25.91667
hours	28,467	36.55956	9.869623	1	168
wks_work	27,831	53.98933	29.03232	0	104
ln_wage	28,534	1.674907	.4780935	0	5.263916

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 age and c.age#c.age are related. After fitting this model, the marginal effect reported by margins includes the full effect of age on the log of income, including the contribution of both coefficients.

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

There are other reasons for preferring factor variables; see [R] **margins** for examples.

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 a standard metropolitan statistical area (SMSA), and south (41% of the person-time was in the South). 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 https://www.stata-press.com/data/r19/union
(NLS Women 14-24 in 1968)

. describe

Contains data from https://www.stata-press.com/data/r19/union.dta
 Observations:        26,200                  NLS Women 14-24 in 1968
    Variables:             8                  4 May 2024 13:54
                                              (_dta has notes)
------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    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

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
      idcode |     26,200    2611.582    1484.994          1       5159
        year |     26,200    79.47137    5.965499         70         88
         age |     26,200    30.43221    6.489056         16         46
       grade |     26,200    12.76145    2.411715          0         18
    not_smsa |     26,200    .2837023    .4508027          0          1
-------------+---------------------------------------------------------
       south |     26,200    .4130153    .4923849          0          1
       union |     26,200    .2217939    .4154611          0          1
       black |     26,200     .274542    .4462917          0          1
```

In many of the examples where the union dataset is used, we also include an interaction between the year variable and the south variable—south#c.year. This interaction is created using factor-variables notation; see **[U] 11.4.3 Factor variables**.

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

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 χ^2_1 but is rather a 50:50 mixture of a χ^2_0 (point mass at zero) and a χ^2_1, denoted as $\overline{\chi}^2_{01}$. See Gutierrez, Carter, and Drukker (2001) for a full discussion.

❑

References

Andreß, H.-J., K. Golsch, and A. W. Schmidt. 2013. *Applied Panel Data Analysis for Economic and Social Surveys*. Berlin: Springer.

Baltagi, B. H. 2013. *Econometric Analysis of Panel Data*. 5th ed. Chichester, UK: Wiley.

Cameron, A. C., and P. K. Trivedi. 2022. *Microeconometrics Using Stata*. 2nd ed. College Station, TX: Stata Press.

Center for Human Resource Research. 1989. *National Longitudinal Survey of Labor Market Experience, Young Women 14–24 years of age in 1968*. Columbus, OH: Ohio State University Press.

Frees, E. W. 2004. *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge: Cambridge University Press.

Greene, W. H. 2018. *Econometric Analysis*. 8th ed. New York: Pearson.

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. College Station, TX: Stata Press.

Hsiao, C. 2014. *Analysis of Panel Data*. 3rd ed. New York: Cambridge University Press.

Sun, Y., and G. Dhaene. 2019. xtspj: A command for split-panel jackknife estimation. *Stata Journal* 19: 335–374.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.

Also see

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