Description

_xtdata_ produces a transformed dataset of the variables specified in _varlist_ or of all the variables in the data. Once the data are transformed, Stata’s _regress_ command may be used to perform specification searches more quickly than _xtreg_; see [R] _regress_ and [XT] _xtreg_. Using _xtdata_, _re_ also creates a variable named _constant_. When using _regress_ after _xtdata_, _re_, specify _noconstant_ and include _constant_ in the regression. After _xtdata_, _be_ and _xtdata_, _fe_, you need not include _constant_ or specify _regress_’s _noconstant_ option.

Quick start

Convert data to a form suitable for random-effects estimation using _xtset_ data

    xtdta, re

As above, but convert only variables _v1_, _v2_ and _v3_

    xtdta v1 v2 v3, re

Convert all variables beginning with _prefix_ to a form suitable for fixed-effects estimation

    xtdta prefix*, fe

Convert data for between estimation if the dataset has changed since last _save_

    xtdta, be clear

Menu

Statistics > Longitudinal/panel data > Setup and utilities > Faster specification searches with _xt data_
Syntax

xtdata [varlist] [if] [in] [, options]

<table>
<thead>
<tr>
<th>options</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main</td>
<td></td>
</tr>
<tr>
<td>re</td>
<td>convert data to a form suitable for random-effects estimation</td>
</tr>
<tr>
<td>ratio(#)</td>
<td>ratio of random effect to pure residual (standard deviations)</td>
</tr>
<tr>
<td>be</td>
<td>convert data to a form suitable for between estimation</td>
</tr>
<tr>
<td>fe</td>
<td>convert data to a form suitable for fixed-effects (within) estimation</td>
</tr>
<tr>
<td>nodouble</td>
<td>keep original variable type; default is to recast type as double</td>
</tr>
<tr>
<td>clear</td>
<td>overwrite current data in memory</td>
</tr>
</tbody>
</table>

A panel variable must be specified; use xtset; see [XT] xtset.

Options

- **Main**
  - **re** specifies that the data be converted into a form suitable for random-effects estimation. re is the default if be, fe, or re is not specified. ratio(#) must also be specified.
  - ratio(#) (use with xtdata, re only) specifies the ratio $\sigma_\nu/\sigma_\epsilon$, which is the ratio of the random effect to the pure residual. This is the ratio of the standard deviations, not the variances.
  - be specifies that the data be converted into a form suitable for between estimation.
  - fe specifies that the data be converted into a form suitable for fixed-effects (within) estimation.
  - nodouble specifies that transformed variables keep their original types, if possible. The default is to recast variables to double.
  - clear specifies that the data may be converted even though the dataset has changed since it was last saved on disk.

Remember that xtdata transforms variables to be differences from group means, pseudodifferences from group means, or group means. Specifying nodouble will decrease the size of the resulting dataset but may introduce roundoff errors in these calculations.

Remarks and examples

If you have not read [XT] xt and [XT] xtreg, please do so.

The formal estimation commands of xtreg—see [XT] xtreg—do not produce results instantaneously, especially with large datasets. Equations (2), (3), and (4) of [XT] xtreg describe the data necessary to fit each of the models with OLS. The idea here is to transform the data once to the appropriate form and then use regress to fit such models more quickly.

Example 1

We will use the example in [XT] xtreg demonstrating between-effects regression. Another way to estimate the between equation is to convert the data in memory to the between data:
The output is the same as that produced by `xtreg, be`; the reported $R^2$ is the $R^2$ between. Using `xtdata` followed by just one `regress` does not save time. Using `xtdata` is justified when you intend to explore the specification of the model by running many alternative regressions.

#### Technical note

When using `xtdata`, you must eliminate any variables that you do not intend to use and that have missing values. `xtdata` follows a casewise-deletion rule, which means that an observation is excluded from the conversion if it is missing on any of the variables. In the example above, we specified that the variables be converted on the command line. We could also drop the variables first, and it might even be useful to preserve our estimation sample:

```
use https://www.stata-press.com/data/r16/nlswork, clear
(National Longitudinal Survey. Young Women 14-26 years of age in 1968)
. generate age2=age^2
(24 missing values generated)
. generate ttl_exp2 = ttl_exp^2
. generate tenure2=tenure^2
(433 missing values generated)
. generate byte black = race==2
. keep id year ln_w grade age* ttl_exp* tenure* black not_smsa south
. save xtdatasmpl
file xtdatasmpl.dta saved
```
Example 2

xtdata with the `fe` option converts the data so that results are equivalent to those from estimating by using `xtreg` with the `fe` option.

```
. xtdta, fe
. regress ln_w grade age* ttl_exp* tenure* black not_smsa south
note: grade omitted because of collinearity
note: black omitted because of collinearity
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 28,091</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>412.443881</td>
<td>8</td>
<td>51.5554852</td>
<td>F(8, 28082) = 732.64</td>
</tr>
<tr>
<td>Residual</td>
<td>1976.12232</td>
<td>28,082</td>
<td>.070369714</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>2388.5662</td>
<td>28,090</td>
<td>.085032617</td>
<td>R-squared = 0.1727</td>
</tr>
</tbody>
</table>

| ln_wage | Coef. | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|---------|-------|-----------|-------|------|---------------------|
| grade   | 0 (omitted) | | | | |
| age     | .0359987 | .0030903 | 11.65 | 0.000 | .0299415 -.042058 |
| age2    | -.000723 | .0000486 | -14.88 | 0.000 | -.0008183 -.0006277 |
| ttl_exp | .0334668 | .0027061 | 12.37 | 0.000 | .0281627 .0387708 |
| ttl_exp2 | .0002163 | .0001166 | 1.86 | 0.064 | -.0000122 .0004447 |
| tenure  | .0357539 | .0016871 | 21.19 | 0.000 | .0324472 .0390606 |
| tenure2 | -.0019701 | .0001141 | -17.27 | 0.000 | -.0021937 -.0017465 |
| black   | 0 (omitted) | | | | |
| not_smsa | -.0890108 | .0086982 | -10.23 | 0.000 | -.1060597 -.0719619 |
| south   | -.0606309 | .0099761 | -6.08 | 0.000 | -.0801845 -.0410772 |
| _cons   | 1.03732 | .0443093 | 23.41 | 0.000 | .9504716 1.124168 |

The coefficients reported by `regress` after `xtdata, fe` are the same as those reported by `xtreg, fe`, but the standard errors are slightly smaller. This is because no adjustment has been made to the estimated covariance matrix for the estimation of the person means. The difference is small, however, and results are adequate for a specification search.

Example 3

To use `xtdata, re`, you must specify the ratio $\sigma_{\nu}/\sigma_\epsilon$, which is the ratio of the standard deviations of the random effect and pure residual. Merely to show the relationship of `regress` after `xtdata, re` to `xtreg, re`, we will specify this ratio as $0.25790526/0.29068923 = 0.88721987$, which is the number `xtreg` reports when the model is fit from the outset; see the random-effects example in [XT] xtreg. For specification searches, however, it is adequate to specify this number more crudely, and, when performing the specification search for this manual entry, we used `ratio(1)`.

```
. use https://www.stata-press.com/data/r16/xtdatasmpl, clear
   (National Longitudinal Survey. Young Women 14-26 years of age in 1968)
. xtdta, clear re ratio(.88721987)
```

<table>
<thead>
<tr>
<th>theta</th>
<th>min</th>
<th>5%</th>
<th>median</th>
<th>95%</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.2520</td>
<td>0.2520</td>
<td>0.5499</td>
<td>0.7016</td>
<td>0.7206</td>
</tr>
</tbody>
</table>

`xtdata` reports the distribution of $\theta$ based on the specified ratio. If these were balanced data, $\theta$ would have been constant.
When running regressions with these data, you must specify the `noconstant` option and include the variable `constant`:

```
.regress ln_w grade age* ttl_exp* tenure* black not_smsa south constant, > noconstant
```

```
<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 28,091</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>13271.7208</td>
<td>11</td>
<td>1206.52007</td>
<td>F(11, 28080) = 14302.56</td>
</tr>
<tr>
<td>Residual</td>
<td>2368.74223</td>
<td>28,080</td>
<td>.084356917</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>15640.463</td>
<td>28,091</td>
<td>.556778435</td>
<td>R-squared = 0.8486</td>
</tr>
</tbody>
</table>

```

Adj R-squared = 0.8485

Root MSE = .29044

```

| ln_wage | Coef. | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|---------|-------|-----------|-------|------|----------------------|
| grade   | .0646499 | .0017812 | 36.30 | 0.000 | .0611587 to .0681411 |
| age     | .0368059 | .0031195 | 11.80 | 0.000 | .0306915 to .0429203 |
| age2    | -.0007133 | .00005  | -14.27| 0.000 | -.0008113 to -.0006153 |
| south   | -.0868922 | .0073032 | -11.90| 0.000 | -.1012068 to -.0725775 |
| constant| .2387206 | .049469  | 4.83  | 0.000 | .141759 to .3356822 |

```

Results are the same coefficients and standard errors that `xtreg, re` estimated in example 4 of `[XT] xtreg`. The summaries at the top, however, should be ignored, as they are expressed in terms of (4) of `[XT] xtreg`, and, moreover, for a model without a constant.

💡 Technical note

Using `xtdata` requires some caution. The following guidelines may help:

1. `xtdata` is intended for use only during the specification search phase of analysis. Results should be estimated with `xtreg` on unconverted data.

2. After converting the data, you may use `regress` to obtain estimates of the coefficients and their standard errors. For `re regress after xtdata, fe`, the standard errors are too small, but only slightly.

3. You may loosely interpret the coefficient’s significance tests and confidence intervals. However, for results after `xtdata, fe` and `re`, an incorrect (but close to correct) distribution is assumed.

4. You should ignore the summary statistics reported at the top of `regress`’s output.

5. After converting the data, you may form linear, but not nonlinear, combinations of regressors; that is, if your data contained age, it would not be correct to convert the data and then form age squared. All nonlinear transformations should be done before conversion. (For `xtdata, fe`, you can get away with forming nonlinear combinations ex post, but the results will not be exact.)
Technical note

The `xtdata` command can be used to help you examine data, especially with `scatter`.

`. use https://www.stata-press.com/data/r16/xtdatasmpl, clear
(National Longitudinal Survey. Young Women 14-26 years of age in 1968)
`. xtdata, be
`. scatter ln_wage age, title(Between data) msymbol(o) msize(tiny)

![Between data](image)

`. use https://www.stata-press.com/data/r16/xtdatasmpl, clear
(National Longitudinal Survey. Young Women 14-26 years of age in 1968)
`. xtdata, fe
`. scatter ln_wage age, title(Within data) msymbol(o) msize(tiny)

![Within data](image)
. use https://www.stata-press.com/data/r16/xtdatasmpl, clear
(National Longitudinal Survey. Young Women 14–26 years of age in 1968)
. scatter ln_wage age, title(Overall data) msymbol(o) msize(tiny)

Methods and formulas

(This section is a continuation of the Methods and formulas of [XT] xtreg.)

_xtdata, be, fe, and re_ transform the data according to (2), (3), and (4), respectively, of [XT] xtreg, except that _xtdata, fe_ adds back in the overall mean, thus forming the transformation

\[ x_{it} - \overline{x}_i + \overline{x} \]

_xtdata, re_ requires the user to specify r as an estimate of \( \sigma_{\nu}/\sigma_{\epsilon} \). \( \theta_i \) is calculated from

\[ \theta_i = 1 - \frac{1}{\sqrt{T_ir^2 + 1}} \]

Also see

[XT] _xtsum_ — Summarize xt data