**xtdata — Faster specification searches with xt data**

### Syntax

\[
\texttt{xtdata} \ [\textit{varlist}] \ [\textit{if}] \ [\textit{in}] \ [\ , \textit{options}] 
\]

### Options

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

A panel variable must be specified; use \textit{xtset}; see \texttt{[XT] xtset}.

### Menu

Statistics > Longitudinal/panel data > Setup and utilities > Faster specification searches with xt data

### Description

\texttt{xtdata} produces a transformed dataset of the variables specified in \textit{varlist} or of all the variables in the data. Once the data are transformed, Stata’s \texttt{regress} command may be used to perform specification searches more quickly than \texttt{xtreg}; see \texttt{[R] regress} and \texttt{[XT] xtreg}. Using \texttt{xtdata}, \texttt{re} also creates a variable named \texttt{constant}. When using \texttt{regress} after \texttt{xtdata}, \texttt{re}, specify \texttt{noconstant} and include \texttt{constant} in the regression. After \texttt{xtdata}, \texttt{be} and \texttt{xtdata, fe}, you need not include \texttt{constant} or specify \texttt{regress}’s \texttt{noconstant} option.

### Options

- \texttt{re} specifies that the data are to be converted into a form suitable for random-effects estimation. \texttt{re} is the default if \texttt{be}, \texttt{fe}, or \texttt{re} is not specified. \texttt{ratio()} must also be specified.
- \texttt{ratio(#)} (use with \texttt{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.
- \texttt{be} specifies that the data are to be converted into a form suitable for between estimation.
- \texttt{fe} specifies that the data are to 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.

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.

clear specifies that the data may be converted even though the dataset has changed since it was last saved on disk.

**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:

```stata
. use http://www.stata-press.com/data/r13/nlswork
   (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
. xtdata ln_w grade age* ttl_exp* tenure* black not_smsa south, be
. regress ln_w grade age* ttl_exp* tenure* black not_smsa south
```

```
Source |       SS    df  MS
----------|-----------|---------|---------|----------|
Model    | 415.021613   10 41.5021613
Residual | 431.954995 4686 .092179896
         |         |         |         | R-squared = 0.4900
         |         |         |         | Adj R-squared = 0.4889
         |         |         |         | Root MSE = .30361
Total    | 846.976608 4696 .180361288

    ln_wage  Coef.  Std. Err.    t  P>|t|     [95% Conf. Interval]
----------|--------|-----------|------|---------|------------------|
grade     |  .0607602 |  .0020006 | 30.37 |  0.000 |   .0568382    .0646822
age       |  .0323158 |  .0087251 |  3.70 |  0.000 |   .0152105    .0494211
age2      | -.0005997 |  .0001429 | -4.20 |  0.000 |  -.0008799   -.0003194
          |          |          |       |         |                |
south     |  -.0993378 |  .010136 |   -9.80  |  0.000 |  -.1192091  -.0794665
_cons     |   .3339113 |  .1210434 |   2.76  |  0.006 |   .0966093    .5712133
```

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 http://www.stata-press.com/data/r13/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.

```
. xtdata, 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

Source | SS       | df | MS
-------------|----------|----|------------------
Model    | 412.443881 | 8  | 51.554852
Residual | 1976.12232 28082 | 0.070369714
          |          |    |                  
Total    | 2388.5662 28090 | 0.085032617
          |          |    |                  
        |          |    |                  
ln_wage | Coef.    | Std. Err. | t  | P>|t| | [95% Conf. Interval]
-------------|----------|------------|----|-------|----------------------
        | grade 0  | (omitted) |    |       |                      
    age | .0359987 | .0030903   | 11.65 | 0.000 | .0299415 | .0420558
age2  | -.000723 | .0005486   | -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 | .0000447
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 http://www.stata-press.com/data/r13/xtdatasmpl, clear
(National Longitudinal Survey. Young Women 14-26 years of age in 1968)
.xtdata, 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
```

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 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`, `be`, 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 http://www.stata-press.com/data/r13/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)
```
. use http://www.stata-press.com/data/r13/xtdatasmpl, clear
(National Longitudinal Survey. Young Women 14-26 years of age in 1968)
. xtda, fe
. scatter ln_wage age, title(Within data) msymbol(o) msize(tiny)

. use http://www.stata-press.com/data/r13/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

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

_xtdata_, _re_ requires the user to specify _r_ as an estimate of \( \sigma_v / \sigma_\epsilon \). \( \theta_i \) is calculated from

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

Also see

[XT] xtsum — Summarize xt data