Title

xtgls — Fit panel-data models by using GLS

Syntax

`xtgls` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [, *options*]

options	Description
Model	
`noconstant`	suppress constant term
`panels(iid)`	use i.i.d. error structure
`panels(heteroskedastic)`	use heteroskedastic but uncorrelated error structure
`panels(correlated)`	use heteroskedastic and correlated error structure
`corr(independent)`	use independent autocorrelation structure
`corr(ar1)`	use AR1 autocorrelation structure
`corr(psar1)`	use panel-specific AR1 autocorrelation structure
`rhotype(`*calc*`)`	specify method to compute autocorrelation parameter; see *Options* for details; seldom used
`igls`	use iterated GLS estimator instead of two-step GLS estimator
`force`	estimate even if observations unequally spaced in time
SE	
`nmk`	normalize standard error by $N - k$ instead of N
Reporting	
`level(#)`	set confidence level; default is `level(95)`
display_options	control column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling
Optimization	
optimize_options	control the optimization process; seldom used
`coeflegend`	display legend instead of statistics

A panel variable must be specified. For correlation structures other than `independent`, a time variable must be specified. A time variable must also be specified if `panels(correlated)` is specified. Use `xtset`; see [XT] **xtset**.

indepvars may contain factor variables; see **[U] 11.4.3 Factor variables**.

depvar and *indepvars* may contain time-series operators; see **[U] 11.4.4 Time-series varlists**.

`by` and `statsby` are allowed; see **[U] 11.1.10 Prefix commands**.

`aweight`s are allowed; see **[U] 11.1.6 weight**.

`coeflegend` does not appear in the dialog box.

See **[U] 20 Estimation and postestimation commands** for more capabilities of estimation commands.

Menu

Statistics > Longitudinal/panel data > Contemporaneous correlation > GLS regression with correlated disturbances

Description

`xtgls` fits panel-data linear models by using feasible generalized least squares. This command allows estimation in the presence of AR(1) autocorrelation within panels and cross-sectional correlation and heteroskedasticity across panels.

Options

Model

`noconstant`; see [R] **estimation options**.

`panels(`*pdist*`)` specifies the error structure across panels.

`panels(iid)` specifies a homoskedastic error structure with no cross-sectional correlation. This is the default.

`panels(heteroskedastic)` specifies a heteroskedastic error structure with no cross-sectional correlation.

`panels(correlated)` specifies a heteroskedastic error structure with cross-sectional correlation. If `p(c)` is specified, you must also specify a time variable (use `xtset`). The results will be based on a generalized inverse of a singular matrix unless $T \geq m$ (the number of periods is greater than or equal to the number of panels).

`corr(`*corr*`)` specifies the assumed autocorrelation within panels.

`corr(independent)` specifies that there is no autocorrelation. This is the default.

`corr(ar1)` specifies that, within panels, there is AR(1) autocorrelation and that the coefficient of the AR(1) process is common to all the panels. If `c(ar1)` is specified, you must also specify a time variable (use `xtset`).

`corr(psar1)` specifies that, within panels, there is AR(1) autocorrelation and that the coefficient of the AR(1) process is specific to each panel. `psar1` stands for panel-specific AR(1). If `c(psar1)` is specified, a time variable must also be specified; use `xtset`.

`rhotype(`*calc*`)` specifies the method to be used to calculate the autocorrelation parameter:

`regress`	regression using lags; the default
`dw`	Durbin–Watson calculation
`freg`	regression using leads
`nagar`	Nagar calculation
`theil`	Theil calculation
`tscorr`	time-series autocorrelation calculation

All the calculations are asymptotically equivalent and consistent; this is a rarely used option.

`igls` requests an iterated GLS estimator instead of the two-step GLS estimator for a nonautocorrelated model or instead of the three-step GLS estimator for an autocorrelated model. The iterated GLS estimator converges to the MLE for the `corr(independent)` models but does not for the other `corr()` models.

force specifies that estimation be forced even though the time variable is not equally spaced. This is relevant only for correlation structures that require knowledge of the time variable. These correlation structures require that observations be equally spaced so that calculations based on lags correspond to a constant time change. If you specify a time variable indicating that observations are not equally spaced, the (time dependent) model will not be fit. If you also specify force, the model will be fit, and it will be assumed that the lags based on the data ordered by the time variable are appropriate.

SE

nmk specifies that standard errors be normalized by $N - k$, where k is the number of parameters estimated, rather than N, the number of observations. Different authors have used one or the other normalization. Greene (2012, 280) remarks that whether a degree-of-freedom correction improves the small-sample properties is an open question.

Reporting

level(#); see [R] **estimation options**.

display_options: noomitted, vsquish, noemptycells, baselevels, allbaselevels, nofvlabel, fvwrap(#), fvwrapon(*style*), cformat(%*fmt*), pformat(%*fmt*), sformat(%*fmt*), and nolstretch; see [R] **estimation options**.

Optimization

optimize_options control the iterative optimization process. These options are seldom used.

iterate(#) specifies the maximum number of iterations. When the number of iterations equals #, the optimization stops and presents the current results, even if convergence has not been reached. The default is iterate(100).

tolerance(#) specifies the tolerance for the coefficient vector. When the relative change in the coefficient vector from one iteration to the next is less than or equal to #, the optimization process is stopped. tolerance(1e-7) is the default.

nolog suppresses display of the iteration log.

The following option is available with xtgls but is not shown in the dialog box:

coeflegend; see [R] **estimation options**.

Remarks and examples

stata.com

Remarks are presented under the following headings:

Introduction

Information on GLS can be found in Greene (2012), Maddala and Lahiri (2006), Davidson and MacKinnon (1993), and Judge et al. (1985).

If you have many panels relative to periods, see [XT] **xtreg** and [XT] **xtgee**. `xtgee`, in particular, provides capabilities similar to those of `xtgls` but does not allow cross-sectional correlation. On the other hand, `xtgee` allows a richer description of the correlation within panels as long as the same correlations apply to all panels. `xtgls` provides two unique features:

1. Cross-sectional correlation may be modeled (`panels(correlated)`).
2. Within panels, the AR(1) correlation coefficient may be unique (`corr(psar1)`).

`xtgls` allows models with heteroskedasticity and no cross-sectional correlation, but, strictly speaking, `xtgee` does not. `xtgee` with the `vce(robust)` option relaxes the assumption of equal variances, at least as far as the standard error calculation is concerned.

Also, `xtgls, panels(iid) corr(independent) nmk` is equivalent to `regress`.

The `nmk` option uses $n-k$ rather than n to normalize the variance calculation.

To fit a model with autocorrelated errors (`corr(ar1)` or `corr(psar1)`), the data must be equally spaced in time. To fit a model with cross-sectional correlation (`panels(correlated)`), panels must have the same number of observations (be balanced).

The equation from which the models are developed is given by

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \epsilon_{it}$$

where $i = 1, \ldots, m$ is the number of units (or panels) and $t = 1, \ldots, T_i$ is the number of observations for panel i. This model can equally be written as

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_m \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_m \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_m \end{bmatrix}$$

The variance matrix of the disturbance terms can be written as

$$E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'] = \boldsymbol{\Omega} = \begin{bmatrix} \sigma_{1,1}\boldsymbol{\Omega}_{1,1} & \sigma_{1,2}\boldsymbol{\Omega}_{1,2} & \cdots & \sigma_{1,m}\boldsymbol{\Omega}_{1,m} \\ \sigma_{2,1}\boldsymbol{\Omega}_{2,1} & \sigma_{2,2}\boldsymbol{\Omega}_{2,2} & \cdots & \sigma_{2,m}\boldsymbol{\Omega}_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m,1}\boldsymbol{\Omega}_{m,1} & \sigma_{m,2}\boldsymbol{\Omega}_{m,2} & \cdots & \sigma_{m,m}\boldsymbol{\Omega}_{m,m} \end{bmatrix}$$

For the $\boldsymbol{\Omega}_{i,j}$ matrices to be parameterized to model cross-sectional correlation, they must be square (balanced panels).

In these models, we assume that the coefficient vector $\boldsymbol{\beta}$ is the same for all panels and consider a variety of models by changing the assumptions on the structure of $\boldsymbol{\Omega}$.

For the classic OLS regression model, we have

$$\begin{aligned} E[\epsilon_{i,t}] &= 0 \\ \mathrm{Var}[\epsilon_{i,t}] &= \sigma^2 \\ \mathrm{Cov}[\epsilon_{i,t}, \epsilon_{j,s}] &= 0 \qquad \text{if } t \neq s \text{ or } i \neq j \end{aligned}$$

This amounts to assuming that $\mathbf{\Omega}$ has the structure given by

$$\mathbf{\Omega} = \begin{bmatrix} \sigma^2\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \sigma^2\mathbf{I} \end{bmatrix}$$

whether or not the panels are balanced (the $\mathbf{0}$ matrices may be rectangular). The classic OLS assumptions are the default `panels(iid)` and `corr(independent)` options for this command.

Heteroskedasticity across panels

In many cross-sectional datasets, the variance for each of the panels differs. It is common to have data on countries, states, or other units that have variation of scale. The heteroskedastic model is specified by including the `panels(heteroskedastic)` option, which assumes that

$$\mathbf{\Omega} = \begin{bmatrix} \sigma_1^2\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sigma_2^2\mathbf{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \sigma_m^2\mathbf{I} \end{bmatrix}$$

▷ Example 1

Greene (2012, 1112) reprints data in a classic study of investment demand by Grunfeld and Griliches (1960). Below we allow the variances to differ for each of the five companies.

```
. use http://www.stata-press.com/data/r13/invest2
. xtgls invest market stock, panels(hetero)
Cross-sectional time-series FGLS regression

Coefficients:  generalized least squares
Panels:        heteroskedastic
Correlation:   no autocorrelation

Estimated covariances      =         5          Number of obs      =       100
Estimated autocorrelations =         0          Number of groups   =         5
Estimated coefficients     =         3          Time periods       =        20
                                                Wald chi2(2)       =    865.38
                                                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      invest |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      market |   .0949905    .007409    12.82   0.000     .0804692    .1095118
       stock |   .3378129   .0302254    11.18   0.000     .2785722    .3970535
       _cons |   -36.2537   6.124363    -5.92   0.000    -48.25723   -24.25017
------------------------------------------------------------------------------
```

◁

Correlation across panels (cross-sectional correlation)

We may wish to assume that the error terms of panels are correlated, in addition to having different scale variances. The variance structure is specified by including the `panels(correlated)` option and is given by

$$\mathbf{\Omega} = \begin{bmatrix} \sigma_1^2\mathbf{I} & \sigma_{1,2}\mathbf{I} & \cdots & \sigma_{1,m}\mathbf{I} \\ \sigma_{2,1}\mathbf{I} & \sigma_2^2\mathbf{I} & \cdots & \sigma_{2,m}\mathbf{I} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m,1}\mathbf{I} & \sigma_{m,2}\mathbf{I} & \cdots & \sigma_m^2\mathbf{I} \end{bmatrix}$$

Because we must estimate cross-sectional correlation in this model, the panels must be balanced (and $T \geq m$ for valid results). A time variable must also be specified so that `xtgls` knows how the observations within panels are ordered. `xtset` shows us that this is true.

▷ Example 2

```
. xtset
       panel variable:  company (strongly balanced)
        time variable:  time, 1 to 20
                delta:  1 unit
. xtgls invest market stock, panels(correlated)
Cross-sectional time-series FGLS regression

Coefficients:  generalized least squares
Panels:        heteroskedastic with cross-sectional correlation
Correlation:   no autocorrelation

Estimated covariances      =        15          Number of obs      =       100
Estimated autocorrelations =         0          Number of groups   =         5
Estimated coefficients     =         3          Time periods       =        20
                                                Wald chi2(2)       =   1285.19
                                                Prob > chi2        =    0.0000
```

invest	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
market	.0961894	.0054752	17.57	0.000	.0854583	.1069206
stock	.3095321	.0179851	17.21	0.000	.2742819	.3447822
_cons	-38.36128	5.344871	-7.18	0.000	-48.83703	-27.88552

The estimated cross-sectional covariances are stored in `e(Sigma)`.

```
. matrix list e(Sigma)
symmetric e(Sigma)[5,5]
             _ee         _ee2         _ee3         _ee4         _ee5
 _ee   9410.9061
_ee2  -168.04631    755.85077
_ee3  -1915.9538   -4163.3434     34288.49
_ee4  -1129.2896   -80.381742    2259.3242    633.42367
_ee5   258.50132     4035.872   -27898.235   -1170.6801    33455.511
```

◁

▷ Example 3

We can obtain the MLE results by specifying the igls option, which iterates the GLS estimation technique to convergence:

```
. xtgls invest market stock, panels(correlated) igls
Iteration 1: tolerance = .2127384
Iteration 2: tolerance = .22817
  (output omitted)
Iteration 1046: tolerance = 1.000e-07

Cross-sectional time-series FGLS regression

Coefficients:  generalized least squares
Panels:        heteroskedastic with cross-sectional correlation
Correlation:   no autocorrelation

Estimated covariances      =        15          Number of obs      =       100
Estimated autocorrelations =         0          Number of groups   =         5
Estimated coefficients     =         3          Time periods       =        20
                                                Wald chi2(2)       =    558.51
Log likelihood             = -515.4222          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      invest |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      market |    .023631    .004291     5.51   0.000     .0152207    .0320413
       stock |   .1709472   .0152526    11.21   0.000     .1410526    .2008417
       _cons |  -2.216508   1.958845    -1.13   0.258    -6.055774    1.622759
------------------------------------------------------------------------------
```

Here the log likelihood is reported in the header of the output.

◁

Autocorrelation within panels

The individual identity matrices along the diagonal of $\boldsymbol{\Omega}$ may be replaced with more general structures to allow for serial correlation. xtgls allows three options so that you may assume a structure with corr(independent) (no autocorrelation); corr(ar1) (serial correlation where the correlation parameter is common for all panels); or corr(psar1) (serial correlation where the correlation parameter is unique for each panel).

The restriction of a common autocorrelation parameter is reasonable when the individual correlations are nearly equal and the time series are short.

If the restriction of a common autocorrelation parameter is reasonable, this allows us to use more information in estimating the autocorrelation parameter to produce a more reasonable estimate of the regression coefficients.

When you specify corr(ar1) or corr(psar1), the iterated GLS estimator does not converge to the MLE.

▷ Example 4

If corr(ar1) is specified, each group is assumed to have errors that follow the same AR(1) process; that is, the autocorrelation parameter is the same for all groups.

```
. xtgls invest market stock, panels(hetero) corr(ar1)
Cross-sectional time-series FGLS regression

Coefficients:  generalized least squares
Panels:        heteroskedastic
Correlation:   common AR(1) coefficient for all panels  (0.8651)

Estimated covariances      =         5          Number of obs      =       100
Estimated autocorrelations =         1          Number of groups   =         5
Estimated coefficients     =         3          Time periods       =        20
                                                Wald chi2(2)       =    119.69
                                                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      invest |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      market |   .0744315   .0097937     7.60   0.000     .0552362    .0936268
       stock |   .2874294   .0475391     6.05   0.000     .1942545    .3806043
       _cons |  -18.96238   17.64943    -1.07   0.283    -53.55464    15.62987
------------------------------------------------------------------------------
```

◁

▷ Example 5

If corr(psar1) is specified, each group is assumed to have errors that follow a different AR(1) process.

```
. xtgls invest market stock, panels(iid) corr(psar1)
Cross-sectional time-series FGLS regression

Coefficients:  generalized least squares
Panels:        homoskedastic
Correlation:   panel-specific AR(1)

Estimated covariances      =         1          Number of obs      =       100
Estimated autocorrelations =         5          Number of groups   =         5
Estimated coefficients     =         3          Time periods       =        20
                                                Wald chi2(2)       =    252.93
                                                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      invest |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      market |   .0934343   .0097783     9.56   0.000     .0742693    .1125993
       stock |   .3838814   .0416775     9.21   0.000      .302195    .4655677
       _cons |   -10.1246   34.06675    -0.30   0.766     -76.8942    56.64499
------------------------------------------------------------------------------
```

◁

Stored results

xtgls stores the following in e():

Scalars

e(N)	number of observations
e(N_g)	number of groups
e(N_t)	number of periods
e(N_miss)	number of missing observations
e(n_cf)	number of estimated coefficients
e(n_cv)	number of estimated covariances
e(n_cr)	number of estimated correlations
e(df_pear)	degrees of freedom for Pearson χ^2
e(chi2)	χ^2
e(df)	degrees of freedom
e(g_min)	smallest group size
e(g_avg)	average group size
e(g_max)	largest group size
e(rank)	rank of e(V)
e(rc)	return code

Macros

e(cmd)	xtgls
e(cmdline)	command as typed
e(depvar)	name of dependent variable
e(ivar)	variable denoting groups
e(tvar)	variable denoting time within groups
e(coefftype)	estimation scheme
e(corr)	correlation structure
e(vt)	panel option
e(rhotype)	type of estimated correlation
e(wtype)	weight type
e(wexp)	weight expression
e(title)	title in estimation output
e(chi2type)	Wald; type of model χ^2 test
e(rho)	ρ
e(properties)	b V
e(predict)	program used to implement predict
e(asbalanced)	factor variables fvset as asbalanced
e(asobserved)	factor variables fvset as asobserved

Matrices

e(b)	coefficient vector
e(Sigma)	$\widehat{\mathbf{\Sigma}}$ matrix
e(V)	variance–covariance matrix of the estimators

Functions

e(sample)	marks estimation sample

Methods and formulas

The GLS results are given by

$$\widehat{\beta}_{\mathrm{GLS}} = (\mathbf{X}'\widehat{\mathbf{\Omega}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathbf{\Omega}}^{-1}\mathbf{y}$$

$$\widehat{\mathrm{Var}}(\widehat{\beta}_{\mathrm{GLS}}) = (\mathbf{X}'\widehat{\mathbf{\Omega}}^{-1}\mathbf{X})^{-1}$$

For all our models, the $\mathbf{\Omega}$ matrix may be written in terms of the Kronecker product:

$$\mathbf{\Omega} = \mathbf{\Sigma}_{m\times m} \otimes \mathbf{I}_{T_i \times T_i}$$

The estimated variance matrix is obtained by substituting the estimator $\widehat{\mathbf{\Sigma}}$ for $\mathbf{\Sigma}$, where

$$\widehat{\mathbf{\Sigma}}_{i,j} = \frac{\widehat{\boldsymbol{\epsilon}}_i{}' \, \widehat{\boldsymbol{\epsilon}}_j}{T}$$

The residuals used in estimating $\mathbf{\Sigma}$ are first obtained from OLS regression. If the estimation is iterated, residuals are obtained from the last fitted model.

Maximum likelihood estimates may be obtained by iterating the FGLS estimates to convergence for models with no autocorrelation, `corr(independent)`.

The GLS estimates and their associated standard errors are calculated using $\widehat{\mathbf{\Sigma}}^{-1}$. As Beck and Katz (1995) point out, the $\mathbf{\Sigma}$ matrix is of rank at most $\min(T, m)$ when you use the `panels(correlated)` option. For the GLS results to be valid (not based on a generalized inverse), T must be at least as large as m, as you need at least as many period observations as there are panels.

Beck and Katz (1995) suggest using OLS parameter estimates with asymptotic standard errors that are corrected for correlation between the panels. This estimation can be performed with the `xtpcse` command; see [XT] **xtpcse**.

References

Baum, C. F. 2001. Residual diagnostics for cross-section time series regression models. *Stata Journal* 1: 101–104.

Beck, N. L., and J. N. Katz. 1995. What to do (and not to do) with time-series cross-section data. *American Political Science Review* 89: 634–647.

Blackwell, J. L., III. 2005. Estimation and testing of fixed-effect panel-data systems. *Stata Journal* 5: 202–207.

Davidson, R., and J. G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. New York: Oxford University Press.

Greene, W. H. 2012. *Econometric Analysis*. 7th ed. Upper Saddle River, NJ: Prentice Hall.

Grunfeld, Y., and Z. Griliches. 1960. Is aggregation necessarily bad? *Review of Economics and Statistics* 42: 1–13.

Hoechle, D. 2007. Robust standard errors for panel regressions with cross-sectional dependence. *Stata Journal* 7: 281–312.

Judge, G. G., W. E. Griffiths, R. C. Hill, H. Lütkepohl, and T.-C. Lee. 1985. *The Theory and Practice of Econometrics*. 2nd ed. New York: Wiley.

Maddala, G. S., and K. Lahiri. 2006. *Introduction to Econometrics*. 4th ed. New York: Wiley.

Also see

[XT] **xtgls postestimation** — Postestimation tools for xtgls

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

[XT] **xtpcse** — Linear regression with panel-corrected standard errors

[XT] **xtreg** — Fixed-, between-, and random-effects and population-averaged linear models

[XT] **xtregar** — Fixed- and random-effects linear models with an AR(1) disturbance

[R] **regress** — Linear regression

[TS] **newey** — Regression with Newey–West standard errors

[TS] **prais** — Prais–Winsten and Cochrane–Orcutt regression

[U] **20 Estimation and postestimation commands**