Title

> **xtstreg —** Random-effects parametric survival models

Description

xtstreg fits random-effects parametric survival-time models. The conditional distribution of the response given the random effects is assumed to be an exponential, a Weibull, a lognormal, a loglogistic, or a gamma distribution. xtstreg can be used with single- or multiple-record st data.

Quick start

Declare pvar as the panel variable using xtset, and declare data to be survival-time data using stset

    xtset pvar
    stset tvar, failure(fail)

Random-effects Weibull survival model with covariates x1 and x2

    xtstreg x1 x2, distribution(weibull)

Use accelerated failure-time metric instead of proportional-hazards metric

    xtstreg x1 x2, distribution(weibull) time

Same as above, but with panel-level probability weights wvar

    xtstreg x1 x2 [pw=wvar], distribution(weibull) time

Random-effects model with cluster–robust standard errors for panels nested within cvar

    xtstreg x1 x2, distribution(weibull) vce(cluster cvar)

Menu

Statistics > Longitudinal/panel data > Survival models > Parametric survival models (RE)

Syntax

> xtstreg [*indepvars*] [*if*] [*in*] [*weight*], <u>distribution</u>(*distname*) [*options*]

options	Description
[Model]	
<u>noc</u>onstant	suppress constant term
* <u>dist</u>ribution(*distname*)	specify survival distribution
time	use accelerated failure-time metric
<u>off</u>set(*varname*)	include *varname* in model with coefficient constrained to 1
constraints(*constraints*)	apply specified linear constraints
[SE/Robust]	
vce(*vcetype*)	*vcetype* may be oim, <u>r</u>obust, <u>cl</u>uster *clustvar*, <u>boot</u>strap, or <u>jack</u>knife
[Reporting]	
level(*#*)	set confidence level; default is level(95)
nohr	do not report hazard ratios
noshow	do not show st setting information
lrmodel	perform the likelihood-ratio model test instead of the default Wald test
nocnsreport	do not display constraints
<u>tr</u>atio	report time ratios
display_options	control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling
[Integration]	
<u>intm</u>ethod(*intmethod*)	integration method; *intmethod* may be <u>mvag</u>hermite (the default) or <u>g</u>hermite
intpoints(*#*)	use *#* quadrature points; default is intpoints(12)
[Maximization]	
maximize_options	control the maximization process; seldom used
startgrid(*numlist*)	improve starting value of the random-intercept parameter by performing a grid search
nodisplay	suppress display of header and coefficients
collinear	keep collinear variables
coeflegend	display legend instead of statistics

* distribution(*distname*) is required.

distname	Description
exponential	exponential survival distribution
loglogistic	loglogistic survival distribution
llogistic	synonym for loglogistic
weibull	Weibull survival distribution
lognormal	lognormal survival distribution
lnormal	synonym for lognormal
gamma	gamma survival distribution

You must stset your data before using xtstreg; see [ST] **stset**.

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

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

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

by, collect, fp, and statsby are allowed; see [U] **11.1.10 Prefix commands**.

fweights, iweights, and pweights are allowed; see [U] **11.1.6 weight**. Weights must be constant within panel.

startgrid(), nodisplay, collinear, and coeflegend do not appear in the dialog box.

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

Options

> Model

noconstant; see [R] **Estimation options**.

distribution(*distname*) specifies the survival model to be fit. *distname* is one of the following: exponential, loglogistic, llogistic, weibull, lognormal, lnormal, or gamma. This option is required.

time specifies that the model be fit in the accelerated failure-time metric rather than in the log relative-hazard metric. This option is valid only for the exponential and Weibull models because these are the only models that have both a proportional-hazards and an accelerated failure-time parameterization. Regardless of metric, the likelihood function is the same, and models are equally appropriate in either metric; it is just a matter of changing interpretation.

> time must be specified at estimation.

offset(*varname*) specifies that *varname* be included in the fixed-effects portion of the model with the coefficient constrained to be 1.

constraints(*constraints*); see [R] **Estimation options**.

> SE/Robust

vce(*vcetype*) specifies the type of standard error reported, which includes types that are derived from asymptotic theory (oim), that are robust to some kinds of misspecification (robust), that allow for intragroup correlation (cluster *clustvar*), and that use bootstrap or jackknife methods (bootstrap, jackknife); see [XT] ***vce_options***.

> Specifying vce(robust) is equivalent to specifying vce(cluster *panelvar*); see *xtstreg and the robust VCE estimator* in *Methods and formulas*.

⎡ Reporting ⎤

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

`nohr`, which may be specified at estimation or upon redisplaying results, specifies that coefficients rather than exponentiated coefficients be displayed, that is, that coefficients rather than hazard ratios be displayed. This option affects only how coefficients are displayed, not how they are estimated.

This option is valid only for the exponential and Weibull models because they have a natural proportional-hazards parameterization. These two models, by default, report hazards ratios (exponentiated coefficients).

`noshow` prevents `xtstreg` from showing the key st variables. This option is rarely used because most users type `stset, show` or `stset, noshow` to set once and for all whether they want to see these variables mentioned at the top of the output of every st command; see [ST] **stset**.

`lrmodel`, `nocnsreport`; see [R] **Estimation options**.

`tratio` specifies that exponentiated coefficients, which are interpreted as time ratios, be displayed. `tratio` is appropriate only for the loglogistic, lognormal, and gamma models or for the exponential and Weibull models when fit in the accelerated failure-time metric.

`tratio` may be specified at estimation or upon replay.

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

⎡ Integration ⎤

`intmethod(intmethod)`, `intpoints(#)`; see [R] **Estimation options**.

⎡ Maximization ⎤

maximize_options: `difficult`, `technique(algorithm_spec)`, `iterate(#)`, [no]`log`, `trace`, `gradient`, `showstep`, `hessian`, `showtolerance`, `tolerance(#)`, `ltolerance(#)`, `nrtolerance(#)`, `nonrtolerance`, and `from(init_specs)`; see [R] **Maximize**. These options are seldom used.

The following options are available with `xtstreg` but are not shown in the dialog box:

`startgrid(numlist)` performs a grid search to improve the starting value of the random-intercept parameter. No grid search is performed by default unless the starting value is found to be not feasible, in which case `xtstreg` runs `startgrid(0.1 1 10)` and chooses the value that works best. You may already be using a default form of `startgrid()` without knowing it. If you see `xtstreg` displaying Grid node 1, Grid node 2, . . . following Grid node 0 in the iteration log, that is `xtstreg` doing a default search because the original starting value was not feasible.

`nodisplay` is for programmers. It suppresses the display of the header and the coefficients.

`collinear`, `coeflegend`; see [R] **Estimation options**.

Remarks and examples

`xtstreg` fits random-effects parametric survival-time models. The following discussion relies on the jargon of survival analysis; if you are not familiar with survival-time data or time-to-event data, see *Introduction* in [ST] **Survival analysis** for an overview of survival analysis in Stata and [ST] **Glossary** for a glossary of technical terms.

xtstreg is suitable only for data that have been [stset](#) as well as xtset. stset weights are not used; instead, weights must be specified at estimation. xtstreg requires "subjects", as defined by option id() in stset, to be nested within panels identified by the *panelvar* specified in xtset. xtstreg can be used with multiple-record data, but it does not allow delayed entry or gaps.

For a Stata-specific introduction to survival analysis, see Cleves, Gould, and Marchenko (2016). Rabe-Hesketh and Skrondal (2022, chap. 14–15) is a good introductory read on applied longitudinal modeling of survival data.

▷ Example 1

In example 11 of [ST] **streg**, we fit a Weibull model with an inverse-Gaussian shared frailty to the recurrence times for catheter-insertion point infection for 38 kidney dialysis patients. In this example, the subjects are the catheter insertions, not the patients themselves. This is a function of how the data were recorded—the onset of risk occurs at the time the catheter is inserted and not, say, at the time of admission of the patient into the study. Thus we have two subjects (insertions) within each panel (patient). Each catheter insertion results in either infection (infect==1) or right-censoring (infect==0). Here we stset the data:

```
. use https://www.stata-press.com/data/r18/catheter
(Kidney data, McGilchrist and Aisbett, Biometrics, 1991)

. stset
-> stset time, failure(infect)

Survival-time data settings

         Failure event: infect!=0 & infect<.
Observed time interval: (0, time]
    Exit on or before: failure

_____

         76  total observations
          0  exclusions
_____

         76  observations remaining, representing
         58  failures in single-record/single-failure data
      7,424  total analysis time at risk and under observation
                                            At risk from t =         0
                                 Earliest observed entry t =         0
                                    Last observed exit t =       562
```

While it is reasonable to assume independence of patients, we would not want to assume that recurrence times within each patient are independent. The model used in [ST] **streg** allowed us to model the correlation by assuming that it was the result of a latent patient-level effect, or frailty.

Here we use `xtstreg` to fit a random-effects Weibull model with normally distributed random effects at the patient level. This is done by defining our panels as patients in the dataset.

```
. xtset patient
Panel variable: patient (balanced)
. xtstreg age female, distribution(weibull)
        Failure _d: infect
  Analysis time _t: time
Fitting comparison model:
Iteration 0:  Log likelihood = -1700989.9
Iteration 1:  Log likelihood =  -440.1998
Iteration 2:  Log likelihood = -336.62162
Iteration 3:  Log likelihood = -334.64937
Iteration 4:  Log likelihood = -334.57959
Iteration 5:  Log likelihood = -334.57944
Iteration 6:  Log likelihood = -334.57944
Refining starting values:
Grid node 0:  Log likelihood = -336.03903
Fitting full model:
Iteration 0:  Log likelihood = -336.03903  (not concave)
Iteration 1:  Log likelihood = -333.12188
Iteration 2:  Log likelihood = -330.38629
Iteration 3:  Log likelihood = -329.89525
Iteration 4:  Log likelihood = -329.87946
Iteration 5:  Log likelihood = -329.87938
Iteration 6:  Log likelihood = -329.87938
```

```
Random-effects Weibull PH regression          Number of obs     =          76
Group variable: patient                       Number of groups  =          38
                                              Obs per group:
                                                          min =           2
                                                          avg =         2.0
                                                          max =           2
Integration method: mvaghermite                Integration pts.  =          12
                                              Wald chi2(2)      =       10.17
Log likelihood = -329.87938                    Prob > chi2       =      0.0062
```

_t	Haz. ratio	Std. err.	z	P>|z|	[95% conf. interval]	
age	1.007329	.0137828	0.53	0.594	.9806742	1.034708
female	.1910581	.0999004	-3.17	0.002	.0685629	.5324042
_cons	.0073346	.0072307	-4.99	0.000	.0010623	.0506427
/ln_p	.222825	.1386296			-.0488841	.494534
/sigma2_u	.8234583	.4812598			.2619194	2.588902

```
Note: Estimates are transformed only in the first equation to hazard ratios.
Note: _cons estimates baseline hazard (conditional on zero random effects).
LR test vs. Weibull model: chibar2(01) = 9.40        Prob >= chibar2 = 0.0011
```

The results are similar to those in [ST] **streg**. The likelihood-ratio test compares the random-effects model with a survival model with fixed effects only. The results support the random-effects model.

For this model, `xtstreg` displays exponentiated coefficients, labeled as hazard ratios. These hazard ratios should be interpreted as "conditional hazard ratios", that is, conditional on the random effects.

For example, the hazard ratio for `age` is 1.01. This means that according to the model, for a given patient, the hazard would increase 1% with each year of age. However, at the population level,

marginal hazards corresponding to different levels of the covariates are not necessarily proportional; see example 5 in [ME] **mestreg postestimation** for further discussion.

The exponentiated coefficients of covariates that usually remain constant within a panel do not have a natural interpretation as conditional hazard ratios. However, the magnitude of the exponentiated coefficients always gives an idea of the effect of the covariates. In this example, female is constant within the panel. The estimated hazard ratio for female is 0.19, which indicates that hazard functions for females tend to be smaller than hazard functions for males.

◁

❑ Technical note

The random-effects model is calculated using quadrature, which is an approximation whose accuracy depends partially on the number of integration points used. We can use the quadchk command to see whether changing the number of integration points affects the results. If the results change, the quadrature approximation is not accurate given the number of integration points. Try increasing the number of integration points using the intpoints() option, and run quadchk again. Do not attempt to interpret the results of estimates when the coefficients reported by quadchk differ substantially. See [XT] **quadchk** for details and [XT] **xtprobit** for an example.

Because the xtstreg likelihood function is calculated by Gauss–Hermite quadrature, the computations on large problems can be slow. Computation time is roughly proportional to the number of points used for the quadrature.

❑

Stored results

xtstreg stores the following in e():

Scalars
e(N)	number of observations
e(N_g)	number of groups
e(k)	number of parameters
e(k_eq)	number of equations in e(b)
e(k_eq_model)	number of equations in overall model test
e(k_dv)	number of dependent variables
e(df_m)	model degrees of freedom
e(ll)	log likelihood
e(N_clust)	number of clusters
e(chi2)	χ^2
e(p)	p-value for model test
e(ll_c)	log likelihood, comparison model
e(chi2_c)	χ^2, comparison model
e(sigma_u)	panel-level standard deviation
e(n_quad)	number of quadrature points
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(ic)	number of iterations
e(rc)	return code
e(converged)	1 if converged, 0 otherwise

Macros
e(cmd)	gsem
e(cmd2)	xtstreg
e(cmdline)	command as typed

e(depvar)	_t
e(wtype)	weight type
e(wexp)	weight expression (first-level weights)
e(covariates)	list of covariates
e(ivar)	variable denoting groups
e(model)	model name
e(title)	title in estimation output
e(distribution)	distribution
e(clustvar)	name of cluster variable
e(offset)	offset
e(intmethod)	integration method
e(chi2type)	Wald; type of model χ^2
e(vce)	*vcetype* specified in vce()
e(vcetype)	title used to label Std. err.
e(frm2)	hazard or time
e(opt)	type of optimization
e(which)	max or min; whether optimizer is to perform maximization or minimization
e(ml_method)	type of ml method
e(user)	name of likelihood-evaluator program
e(technique)	maximization technique
e(properties)	b V
e(predict)	program used to implement predict
e(marginsok)	predictions allowed by margins
e(marginsnotok)	predictions disallowed by margins
e(marginswtype)	weight type for margins
e(marginswexp)	weight expression for margins
e(marginsdefault)	default predict() specification for margins
e(asbalanced)	factor variables fvset as asbalanced
e(asobserved)	factor variables fvset as asobserved

Matrices
e(b)	coefficient vector
e(Cns)	constraints matrix
e(ilog)	iteration log (up to 20 iterations)
e(gradient)	gradient vector
e(V)	variance–covariance matrix of the estimator
e(V_modelbased)	model-based variance

Functions
e(sample)	marks estimation sample

In addition to the above, the following is stored in r():

Matrices
r(table)	matrix containing the coefficients with their standard errors, test statistics, *p*-values, and confidence intervals

Note that results stored in r() are updated when the command is replayed and will be replaced when any r-class command is run after the estimation command.

Methods and formulas

Methods and formulas are presented under the following headings:

Survival models

Survival models have a trivariate response (t_0, t, d):

t_0 is the starting time under observation $t_0 \geq 0$;

t is the ending time under observation $t \geq t_0$; and

d is an indicator for failure $d \in \{0, 1\}$.

You must stset your data before using xtstreg; see [ST] stset. By stsetting your data, you define the variables _t0, _t, and _d, which serve as the trivariate response variable.

Let $i = 1, \ldots, n$ panels, $j = 1, \ldots, n_i$, and ν_is be unobservable panel-level random effects that are independent and identically distributed $N(0, \sigma_\nu^2)$.

Two often-used models for adjusting survivor functions for the effects of covariates are the accelerated failure-time (AFT) model and the multiplicative or proportional hazards (PH) model.

In the AFT model, the natural logarithm of the survival time, $\log t$, is expressed as a linear function of the covariates; when we incorporate the random effects, this yields the model

$$\log t_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \nu_i + \epsilon_{ij}$$

The $1 \times p$ row vector $\mathbf{x}_{ij}$ contains the covariates for the fixed effects, with regression coefficients (fixed effects) $\boldsymbol{\beta}$.

The ϵ_{ij} are the observation-level errors with density $\varphi(\cdot)$. The distributional form of the error term determines the regression model. Five regression models are implemented in xtstreg using the AFT parameterization: exponential, gamma, loglogistic, lognormal, and Weibull. The lognormal regression model is obtained by letting $\varphi(\cdot)$ be the normal density. Similarly, by letting $\varphi(\cdot)$ be the logistic density, one obtains the loglogistic regression. Setting $\varphi(\cdot)$ equal to the extreme-value density yields the exponential and the Weibull regression models.

In the PH models fit by xtstreg, the covariates have a multiplicative effect on the hazard function

$$h(t_{ij}) = h_0(t_{ij}) \exp(\mathbf{x}_{ij}\boldsymbol{\beta} + \nu_i)$$

for some baseline hazard function $h_0(t)$. For the xtstreg command, $h_0(t)$ is assumed to be parametric. The exponential and Weibull models are implemented in xtstreg for the PH parameterization. These two models are implemented using both the AFT and PH parameterizations.

The survivor function for a given family is the complement of the cumulative distribution function, $S(t) = 1 - F(t)$. The density for a failure at time t is given by

$$g(t) = \frac{\partial F(t)}{\partial t} = -\frac{\partial S(t)}{\partial t}$$

Some distributions contain ancillary parameters that are not denoted here.

The conditional density for a failure at time t is

$$g(t|t \geq t_0, d = 1) = g(t)/S(t_0)$$

and the conditional probability of survival without failure up to time t is

$$P(T \geq t|t \geq t_0, d = 0) = S(t)/S(t_0)$$

The conditional likelihood is given by

$$L(t_0, t, d) = \left\{ \frac{g(t)}{S(t_0)} \right\}^d \left\{ \frac{S(t)}{S(t_0)} \right\}^{1-d}$$

See *Survival distributions* in [SEM] **Methods and formulas for gsem** for the specific density function corresponding to each distribution.

Given the panel-level random effect $\boldsymbol{\nu}_i$, the conditional distribution of $\mathbf{t}_i = (t_{i1}, \ldots, t_{in_i})'$ given $\boldsymbol{\eta}_i = (\mathbf{x}_{i1}\boldsymbol{\beta} + \nu_i, \ldots, \mathbf{x}_{in_i}\boldsymbol{\beta} + \nu_i)$ for panel i is

$$f(\mathbf{t}_i | \boldsymbol{\eta}_i) = \prod_{j=1}^{n_i} f(t_{ij} | \eta_{ij})$$

where $f(t_{ij} | \eta_{ij})$ is the contribution to the likelihood from observation ij; that is,

$$f(t_{ij} | \eta_{ij}) = \left\{ \frac{g(t_{ij} | \mathbf{x}_{ij}\boldsymbol{\beta} + \nu_i)}{S(t_{0ij} | \mathbf{x}_{ij}\boldsymbol{\beta} + \nu_i)} \right\}^{d_{ij}} \left\{ \frac{S(t_{ij} | \mathbf{x}_{ij}\boldsymbol{\beta} + \nu_i)}{S(t_{0ij} | \mathbf{x}_{ij}\boldsymbol{\beta} + \nu_i)} \right\}^{1-d_{ij}}$$

where $g(t|\eta)$ and $S(t|\eta)$ are, respectively, the density and the survivor function conditional on the linear prediction η.

As mentioned above, xtstreg does not allow delayed entry or gaps. Therefore, the first observation for a given subject will have a value of $t_0 = 0$, and subsequent spells for the subject must start at the end of the previous spell. That is, if observations ij and $i, j + 1$ belong to the same subject, then $t_{0i,j+1} = t_{ij}$.

Given a normal distribution $N(0, \sigma_\nu^2)$ for the random effect ν_i, the panel-level likelihood $\mathcal{L}_i$ is given by

$$\mathcal{L}_i(\boldsymbol{\beta}, \sigma_\nu) = \int_{-\infty}^{\infty} \frac{e^{-\nu_i^2/2\sigma_\nu^2}}{\sqrt{2\pi}\sigma_\nu} \left\{ \prod_{j=1}^{n_i} f(t_{ij} | \mathbf{x}_{ij}\boldsymbol{\beta} + \nu_i) \right\} d\nu_i \tag{1}$$

The integration in (1) has no closed form and thus must be approximated. xtstreg offers two approximation methods: mean–variance adaptive Gauss–Hermite quadrature (default) and nonadaptive Gauss–Hermite quadrature. For details on these integration methods, see [SEM] **Methods and formulas for gsem**. To see how panel-level weights are incorporated into the likelihood function, see *Survey data* in [SEM] **Methods and formulas for gsem**.

The log likelihood for the entire dataset is simply the sum of the contributions of the n individual panels; namely, $\mathcal{L}(\boldsymbol{\beta}, \sigma_\nu) = \sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\beta}, \sigma_\nu)$.

Maximization of $\mathcal{L}(\boldsymbol{\beta}, \sigma_\nu)$ is performed with respect to $(\boldsymbol{\beta}, \sigma_\nu{}^2)$. Parameter estimates are stored in e(b) as $(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}_\nu^2)$, with the corresponding variance–covariance matrix stored in e(V).

xtstreg and the robust VCE estimator

Specifying vce(robust) or vce(cluster *clustvar*) causes the Huber/White/sandwich VCE estimator to be calculated for the coefficients estimated in this regression. See [P] **_robust**, particularly *Introduction* and *Methods and formulas*. Wooldridge (2020) and Arellano (2003) discuss this application of the Huber/White/sandwich VCE estimator. As discussed by Wooldridge (2020), Stock and Watson (2008), and Arellano (2003), specifying vce(robust) is equivalent to specifying vce(cluster *panelvar*), where *panelvar* is the variable that identifies the panels.

Clustering on the panel variable produces a consistent VCE estimator when the disturbances are not identically distributed over the panels or there is serial correlation in ϵ_{it}.

The cluster–robust VCE estimator requires that there be many clusters and that the disturbances be uncorrelated across the clusters. The panel variable must be nested within the cluster variable because of the within-panel correlation that is generally induced by the random-effects transform when there is heteroskedasticity or within-panel serial correlation in the idiosyncratic errors.

References

Arellano, M. 2003. *Panel Data Econometrics*. Oxford: Oxford University Press.

Cleves, M. A., W. W. Gould, and Y. V. Marchenko. 2016. *An Introduction to Survival Analysis Using Stata*. Rev. 3rd ed. College Station, TX: Stata Press.

Philips, A. Q. 2020. An easy way to create duration variables in binary cross-sectional time-series data. *Stata Journal* 20: 916–930.

Rabe-Hesketh, S., and A. Skrondal. 2022. *Multilevel and Longitudinal Modeling Using Stata*. 4th ed. College Station, TX: Stata Press.

Stock, J. H., and M. W. Watson. 2008. Heteroskedasticity-robust standard errors for fixed effects panel data regression. *Econometrica* 76: 155–174. https://doi.org/10.1111/j.0012-9682.2008.00821.x.

Wooldridge, J. M. 2020. *Introductory Econometrics: A Modern Approach*. 7th ed. Boston: Cengage.

Also see

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

[XT] **quadchk** — Check sensitivity of quadrature approximation

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

[ME] **mestreg** — Multilevel mixed-effects parametric survival models

[ST] **streg** — Parametric survival models

[ST] **stset** — Declare data to be survival-time data

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