

streg — Parametric survival models

Description	Quick start	Menu	Syntax
Options	Remarks and examples	Stored results	Methods and formulas
References	Also see		

Description

`streg` performs maximum likelihood estimation for parametric regression survival-time models. `streg` can be used with single- or multiple-record or single- or multiple-failure `st` data. Survival models currently supported are exponential, Weibull, Gompertz, lognormal, loglogistic, and generalized gamma. Parametric frailty models and shared-frailty models are also fit using `streg`.

Also see [\[ST\] `stcox`](#) for proportional hazards models.

Quick start

Weibull survival model with covariates `x1` and `x2` using `stset` data

```
streg x1 x2, distribution(weibull)
```

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

```
streg x1 x2, distribution(weibull) time
```

Different intercepts and ancillary parameters for strata identified by `svar`

```
streg x1 x2, distribution(weibull) strata(svar)
```

Lognormal survival model

```
streg x1 x2, distribution(lognormal)
```

As above, but also model frailty using the gamma distribution

```
streg x1 x2, distribution(lognormal) frailty(gamma)
```

Specify shared frailty within groups identified by `gvar`

```
streg x1 x2, distribution(lognormal) frailty(gamma) shared(gvar)
```

Menu

Statistics > Survival analysis > Regression models > Parametric survival models

Syntax

`streg` [*varlist*] [*if*] [*in*] [, *options*]

options

Description

Model

<code>noconstant</code>	suppress constant term
<code>distribution(exponential)</code>	exponential survival distribution
<code>distribution(gompertz)</code>	Gompertz survival distribution
<code>distribution(loglogistic)</code>	loglogistic survival distribution
<code>distribution(llogistic)</code>	synonym for <code>distribution(loglogistic)</code>
<code>distribution(weibull)</code>	Weibull survival distribution
<code>distribution(lognormal)</code>	lognormal survival distribution
<code>distribution(lnormal)</code>	synonym for <code>distribution(lognormal)</code>
<code>distribution(ggamma)</code>	generalized gamma survival distribution
<code>frailty(gamma)</code>	gamma frailty distribution
<code>frailty(invgaussian)</code>	inverse-Gaussian distribution
<code>time</code>	use accelerated failure-time metric

Model 2

<code>strata(varname)</code>	strata ID variable
<code>offset(varname)</code>	include <i>varname</i> in model with coefficient constrained to 1
<code>shared(varname)</code>	shared frailty ID variable
<code>ancillary(varlist)</code>	use <i>varlist</i> to model the first ancillary parameter
<code>anc2(varlist)</code>	use <i>varlist</i> to model the second ancillary parameter
<code>constraints(constraints)</code>	apply specified linear constraints
<code>collinear</code>	keep collinear variables

SE/Robust

<code>vce(vcetype)</code>	<i>vcetype</i> may be <code>oim</code> , <code>robust</code> , <code>cluster clustvar</code> , <code>opg</code> , <code>bootstrap</code> , or <code>jackknife</code>
---------------------------	--

Reporting

<code>level(#)</code>	set confidence level; default is <code>level(95)</code>
<code>nohr</code>	do not report hazard ratios
<code>tr</code>	report time ratios
<code>noshow</code>	do not show st setting information
<code>noheader</code>	suppress header from coefficient table
<code>nolrtest</code>	do not perform likelihood-ratio test
<code>nocnsreport</code>	do not display constraints
<code>display_options</code>	control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling

Maximization

<code>maximize_options</code>	control the maximization process; seldom used
<code>coeflegend</code>	display legend instead of statistics

You must `stset` your data before using `streg`; see [ST] [stset](#).

`varlist` may contain factor variables; see [U] [11.4.3 Factor variables](#).

`bootstrap`, `by`, `fp`, `jackknife`, `mfp`, `mi estimate`, `nestreg`, `statsby`, `stepwise`, and `svy` are allowed; see [U] [11.1.10 Prefix commands](#).

`vce(bootstrap)` and `vce(jackknife)` are not allowed with the `mi estimate` prefix; see [MI] [mi estimate](#).

`shared()`, `vce()`, and `noheader` are not allowed with the `svy` prefix; see [SVY] [svy](#).

`fweights`, `iweights`, and `pweights` may be specified using `stset`; see [ST] [stset](#). However, weights may not be specified if you are using the `bootstrap` prefix with the `streg` command.

`coeflegend` does 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. A specified `distribution()` is remembered from one estimation to the next when `distribution()` is not specified.

For instance, typing `streg x1 x2, distribution(weibull)` fits a Weibull model. Subsequently, you do not need to specify `distribution(weibull)` to fit other Weibull regression models.

All Stata estimation commands, including `streg`, redisplay results when you type the command name without arguments. To fit a model with no explanatory variables, type `streg, distribution(distname)...`

`frailty(gamma | invgaussian)` specifies the assumed distribution of the frailty, or heterogeneity.

The estimation results, in addition to the standard parameter estimates, will contain an estimate of the variance of the frailties and a likelihood-ratio test of the null hypothesis that this variance is zero. When this null hypothesis is true, the model reduces to the model with `frailty(distname)` not specified.

A specified `frailty()` is remembered from one estimation to the next when `distribution()` is not specified. When you specify `distribution()`, the previously remembered specification of `frailty()` is forgotten.

`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 viewed in either metric; it is just a matter of changing the interpretation.

`time` must be specified at estimation.

Model 2

`strata(varname)` specifies the stratification ID variable. Observations with equal values of the variable are assumed to be in the same stratum. Stratified estimates (with equal coefficients across strata but intercepts and ancillary parameters unique to each stratum) are then obtained. This option is not available if `frailty(distname)` is specified.

`offset(varname)`; see [R] [estimation options](#).

`shared(varname)` is valid with `frailty()` and specifies a variable defining those groups over which the frailty is shared, analogous to a random-effects model for panel data where `varname` defines the panels. `frailty()` specified without `shared()` treats the frailties as occurring at the observation level.

A specified `shared()` is remembered from one estimation to the next when `distribution()` is not specified. When you specify `distribution()`, the previously remembered specification of `shared()` is forgotten.

`shared()` may not be used with `distribution(ggamma)`, `vce(robust)`, `vce(cluster clustvar)`, `vce(opg)`, the `svy` prefix, or in the presence of delayed entries or gaps.

If `shared()` is specified without `frailty()` and there is no remembered `frailty()` from the previous estimation, `frailty(gamma)` is assumed to provide behavior analogous to `stcox`; see [ST] [stcox](#).

`ancillary(varlist)` specifies that the ancillary parameter for the Weibull, lognormal, Gompertz, and loglogistic distributions and that the first ancillary parameter (sigma) of the generalized log-gamma distribution be estimated as a linear combination of `varlist`. This option may not be used with `frailty(distname)`.

When an ancillary parameter is constrained to be strictly positive, the logarithm of the ancillary parameter is modeled as a linear combination of `varlist`.

`anc2(varlist)` specifies that the second ancillary parameter (kappa) for the generalized log-gamma distribution be estimated as a linear combination of `varlist`. This option may not be used with `frailty(distname)`.

`constraints(constraints)`, `collinear`; see [R] [estimation options](#).

SE/Robust

`vce(vctype)` specifies the type of standard error reported, which includes types that are derived from asymptotic theory (`oim`, `opg`), 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 [R] [vce_option](#).

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 models with a natural proportional-hazards parameterization: exponential, Weibull, and Gompertz. These three models, by default, report hazard ratios (exponentiated coefficients).

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

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

`noshow` prevents `streg` from showing the key `st` variables. This option is rarely used because most people 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](#).

`noheader` suppresses the output header, either at estimation or upon replay.

`nolrtest` is valid only with frailty models, in which case it suppresses the likelihood-ratio test for significant frailty.

`nocnsreport`; see [R] [estimation options](#).

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

Maximization

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

Setting the optimization type to `technique(bhhh)` resets the default *vcetype* to `vce(opg)`.

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

`coeflegend`; see [R] [estimation options](#).

Remarks and examples

[stata.com](http://www.stata.com)

Remarks are presented under the following headings:

Introduction

Distributions

Weibull and exponential models

Gompertz model

Lognormal and loglogistic models

Generalized gamma model

Examples

Parameterization of ancillary parameters

Stratified estimation

(Unshared-) frailty models

Shared-frailty models

Introduction

What follows is a brief summary of what you can do with `streg`. For a complete tutorial, see [Cleves, Gould, and Marchenko \(2016\)](#), which devotes four chapters to this topic.

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, yielding the linear model

$$\log t_j = \mathbf{x}_j\boldsymbol{\beta} + z_j$$

where \mathbf{x}_j is a vector of covariates, $\boldsymbol{\beta}$ is a vector of regression coefficients, and z_j is the error with density $f()$. The distributional form of the error term determines the regression model. If we let $f()$ be the normal density, the lognormal regression model is obtained. Similarly, by letting $f()$ be the logistic density, the loglogistic regression is obtained. Setting $f()$ equal to the extreme-value density yields the exponential and the Weibull regression models.

The effect of the AFT model is to change the time scale by a factor of $\exp(-\mathbf{x}_j\boldsymbol{\beta})$. Depending on whether this factor is greater or less than 1, time is either accelerated or decelerated (degraded). That is, if a subject at baseline experiences a probability of survival past time t equal to $S(t)$, then a subject with covariates \mathbf{x}_j would have probability of survival past time t equal to $S()$ evaluated at

the point $\exp(-\mathbf{x}_j\beta)t$, instead. Thus accelerated failure time does not imply a positive acceleration of time with the increase of a covariate but instead implies a deceleration of time or, equivalently, an increase in the expected waiting time for failure.

In the PH model, the concomitant covariates have a multiplicative effect on the hazard function

$$h(t_j) = h_0(t)g(\mathbf{x}_j)$$

for some $h_0(t)$, and for $g(\mathbf{x}_j)$, a nonnegative function of the covariates. A popular choice, and the one adopted here, is to let $g(\mathbf{x}_j) = \exp(\mathbf{x}_j\beta)$. The function $h_0(t)$ may either be left unspecified, yielding the Cox proportional hazards model (see [ST] **stcox**), or take a specific parametric form. For the **streg** command, $h_0(t)$ is assumed to be parametric. Three regression models are currently implemented as PH models: the exponential, Weibull, and Gompertz models. The exponential and Weibull models are implemented as both AFT and PH models, and the Gompertz model is implemented only in the PH metric.

The above model allows for the presence of an intercept term, β_0 , within $\mathbf{x}_j\beta$. Thus what is commonly referred to as the *baseline hazard function*—the hazard when all covariates are zero—is actually equal to $h_0(t)\exp(\beta_0)$. That is, the intercept term serves to scale the baseline hazard. Of course, specifying **noconstant** suppresses the intercept or equivalently constrains β_0 to equal zero.

streg is suitable only for data that have been **stset**. By **stsetting** your data, you define the variables **_t0**, **_t**, and **_d**, which serve as the trivariate response variable (t_0, t, d) . Each response corresponds to a period under observation, $(t_0, t]$, resulting in either failure ($d = 1$) or right-censoring ($d = 0$) at time t . As a result, **streg** is appropriate for data exhibiting delayed entry, gaps, time-varying covariates, and even multiple-failure data.

Distributions

Six parametric survival distributions are currently supported by **streg**. The parameterization and ancillary parameters for each distribution are summarized in table 1:

Table 1. Parametric survival distributions supported by **streg**

Distribution	Metric	Survivor function	Parameterization	Ancillary parameters
Exponential	PH	$\exp(-\lambda_j t_j)$	$\lambda_j = \exp(\mathbf{x}_j\beta)$	
Exponential	AFT	$\exp(-\lambda_j t_j)$	$\lambda_j = \exp(-\mathbf{x}_j\beta)$	
Weibull	PH	$\exp(-\lambda_j t_j^p)$	$\lambda_j = \exp(\mathbf{x}_j\beta)$	p
Weibull	AFT	$\exp(-\lambda_j t_j^p)$	$\lambda_j = \exp(-p\mathbf{x}_j\beta)$	p
Gompertz	PH	$\exp\{-\lambda_j \gamma^{-1}(e^{\gamma t_j} - 1)\}$	$\lambda_j = \exp(\mathbf{x}_j\beta)$	γ
Lognormal	AFT	$1 - \Phi\left\{\frac{\log(t_j) - \mu_j}{\sigma}\right\}$	$\mu_j = \mathbf{x}_j\beta$	σ
Loglogistic	AFT	$\{1 + (\lambda_j t_j)^{1/\gamma}\}^{-1}$	$\lambda_j = \exp(-\mathbf{x}_j\beta)$	γ
Generalized gamma				
if $\kappa > 0$	AFT	$1 - I(\gamma, u)$	$\mu_j = \mathbf{x}_j\beta$	σ, κ
if $\kappa = 0$	AFT	$1 - \Phi(z)$	$\mu_j = \mathbf{x}_j\beta$	σ, κ
if $\kappa < 0$	AFT	$I(\gamma, u)$	$\mu_j = \mathbf{x}_j\beta$	σ, κ

where PH = proportional hazards, AFT = accelerated failure time, and $\Phi(z)$ is the standard normal cumulative distribution. For the generalized gamma, $\gamma = |\kappa|^{-2}$, $u = \gamma \exp(|\kappa|z)$, $I(a, x)$ is the incomplete gamma function, and $z = \text{sign}(\kappa) \{ \log(t_j) - \mu_j \} / \sigma$.

Plotted in figure 1 are example hazard functions for five of the six distributions. The exponential hazard (not separately plotted) is a special case of the Weibull hazard when the Weibull ancillary parameter $p = 1$. The generalized gamma (not plotted) is extremely flexible and therefore can take many shapes.

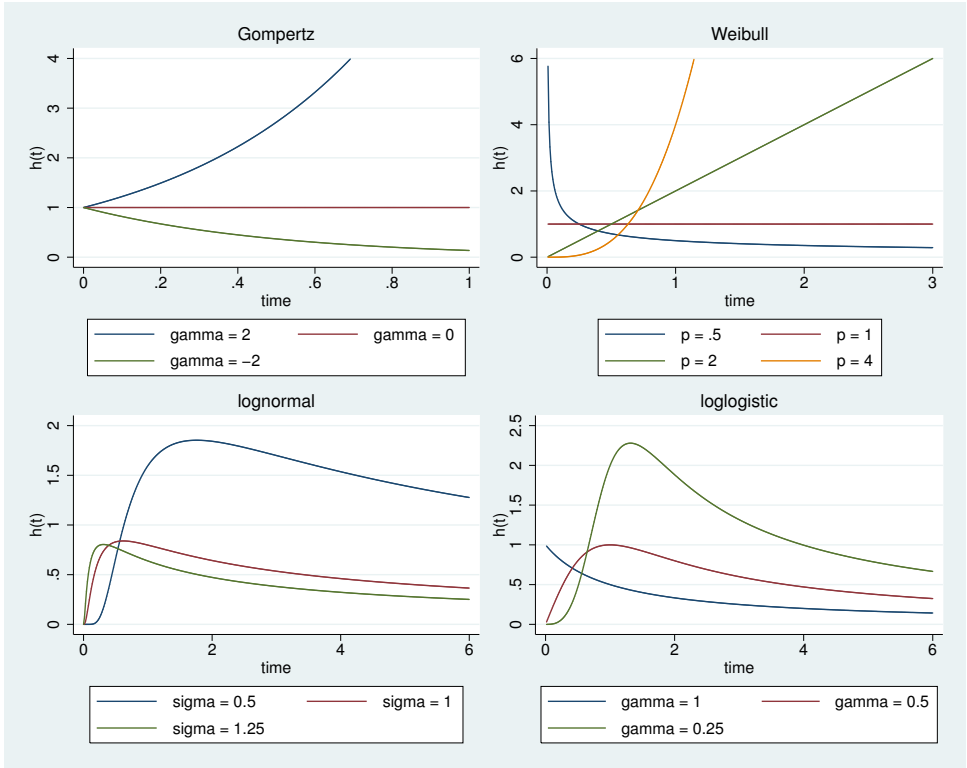


Figure 1. Example plots of hazard functions

Weibull and exponential models

The Weibull and exponential models are parameterized as both PH and AFT models. The Weibull distribution is suitable for modeling data with monotone hazard rates that either increase or decrease exponentially with time, whereas the exponential distribution is suitable for modeling data with constant hazard (see figure 1).

For the PH model, $h_0(t) = 1$ for exponential regression, and $h_0(t) = p t^{p-1}$ for Weibull regression, where p is the shape parameter to be estimated from the data. Some authors refer not to p but to $\sigma = 1/p$.

The AFT model is written as

$$\log(t_j) = \mathbf{x}_j\boldsymbol{\beta}^* + z_j$$

where z_j has an extreme-value distribution scaled by σ . Let $\boldsymbol{\beta}$ be the vector of regression coefficients derived from the PH model so that $\boldsymbol{\beta}^* = -\sigma\boldsymbol{\beta}$. This relationship holds only if the ancillary parameter, p , is a constant; it does not hold when the ancillary parameter is parameterized in terms of covariates.

`streg` uses, by default, for the exponential and Weibull models, the proportional-hazards metric simply because it eases comparison with those results produced by `stcox` (see [ST] `stcox`). You can, however, specify the `time` option to choose the accelerated failure-time parameterization.

The Weibull hazard and survivor functions are

$$h(t) = p\lambda t^{p-1}$$

$$S(t) = \exp(-\lambda t^p)$$

where λ is parameterized as described in table 1. If $p = 1$, these functions reduce to those of the exponential.

Gompertz model

The Gompertz regression is parameterized only as a PH model. First described in 1825, this model has been extensively used by medical researchers and biologists modeling mortality data. The Gompertz distribution implemented is the two-parameter function as described in Lee and Wang (2003), with the following hazard and survivor functions:

$$h(t) = \lambda \exp(\gamma t)$$

$$S(t) = \exp\{-\lambda\gamma^{-1}(e^{\gamma t} - 1)\}$$

The model is implemented by parameterizing $\lambda_j = \exp(\mathbf{x}_j\boldsymbol{\beta})$, implying that $h_0(t) = \exp(\gamma t)$, where γ is an ancillary parameter to be estimated from the data.

This distribution is suitable for modeling data with monotone hazard rates that either increase or decrease exponentially with time (see figure 1).

When γ is positive, the hazard function increases with time; when γ is negative, the hazard function decreases with time; and when γ is zero, the hazard function is equal to λ for all t , so the model reduces to an exponential.

Some recent survival analysis texts, such as Klein and Moeschberger (2003), restrict γ to be strictly positive. If $\gamma < 0$, then as t goes to infinity, the survivor function, $S(t)$, exponentially decreases to a nonzero constant, implying that there is a nonzero probability of never failing (living forever). That is, there is always a nonzero hazard rate, yet it decreases exponentially. By restricting γ to be positive, we know that the survivor function always goes to zero as t tends to infinity.

Although the above argument may be desirable from a mathematical perspective, in Stata's implementation, we took the more traditional approach of not restricting γ . We did this because, in survival studies, subjects are not monitored forever—there is a date when the study ends, and in many investigations, specifically in medical research, an exponentially decreasing hazard rate is clinically appealing.

Lognormal and loglogistic models

The lognormal and loglogistic models are implemented only in the AFT form. These two distributions are similar and tend to produce comparable results. For the lognormal distribution, the natural logarithm of time follows a normal distribution; for the loglogistic distribution, the natural logarithm of time follows a logistic distribution.

The lognormal survivor and density functions are

$$S(t) = 1 - \Phi \left\{ \frac{\log(t) - \mu}{\sigma} \right\}$$

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp \left[\frac{-1}{2\sigma^2} \left\{ \log(t) - \mu \right\}^2 \right]$$

where $\Phi(z)$ is the standard normal cumulative distribution function.

The lognormal regression is implemented by setting $\mu_j = \mathbf{x}_j\boldsymbol{\beta}$ and treating the standard deviation, σ , as an ancillary parameter to be estimated from the data.

The loglogistic regression is obtained if z_j has a logistic density. The loglogistic survivor and density functions are

$$S(t) = \{1 + (\lambda t)^{1/\gamma}\}^{-1}$$

$$f(t) = \frac{\lambda^{1/\gamma} t^{1/\gamma-1}}{\gamma \{1 + (\lambda t)^{1/\gamma}\}^2}$$

This model is implemented by parameterizing $\lambda_j = \exp(-\mathbf{x}_j\boldsymbol{\beta})$ and treating the scale parameter γ as an ancillary parameter to be estimated from the data.

Unlike the exponential, Weibull, and Gompertz distributions, the lognormal and the loglogistic distributions are indicated for data exhibiting nonmonotonic hazard rates, specifically initially increasing and then decreasing rates (figure 1).

Thus far we have considered the exponential, Weibull, lognormal, and loglogistic models. These models are sufficiently flexible for many datasets, but further flexibility can be obtained with the generalized gamma model, described below. Alternatively, you might consider using a Royston–Parmar model (Royston and Parmar 2002; Lambert and Royston 2009). Royston–Parmar models are highly flexible alternatives to the exponential, Weibull, lognormal, and loglogistic models that allow extension from proportional hazards to proportional odds and to scaled probit models. Additional flexibility can be obtained with restricted cubic spline functions as alternatives to the linear functions of log time considered in *Introduction*. See Royston and Lambert (2011) for a thorough treatment of this topic.

Generalized gamma model

The generalized gamma model is implemented only in the AFT form. The three-parameter generalized gamma survivor and density functions are

$$S(t) = \begin{cases} 1 - I(\gamma, u), & \text{if } \kappa > 0 \\ 1 - \Phi(z), & \text{if } \kappa = 0 \\ I(\gamma, u), & \text{if } \kappa < 0 \end{cases}$$

$$f(t) = \begin{cases} \frac{\gamma^\gamma}{\sigma t \sqrt{\gamma} \Gamma(\gamma)} \exp(z\sqrt{\gamma} - u), & \text{if } \kappa \neq 0 \\ \frac{1}{\sigma t \sqrt{2\pi}} \exp(-z^2/2), & \text{if } \kappa = 0 \end{cases}$$

where $\gamma = |\kappa|^{-2}$, $z = \text{sign}(\kappa)\{\log(t) - \mu\}/\sigma$, $u = \gamma \exp(|\kappa|z)$, $\Phi(z)$ is the standard normal cumulative distribution function, and $I(a, x)$ is the incomplete gamma function. See the [gammap\(a, x\)](#) entry in [\[FN\] Statistical functions](#) to see how the incomplete gamma function is implemented in Stata.

This model is implemented by parameterizing $\mu_j = \mathbf{x}_j\beta$ and treating the parameters κ and σ as ancillary parameters to be estimated from the data.

The hazard function of the generalized gamma distribution is extremely flexible, allowing for many possible shapes, including as special cases the Weibull distribution when $\kappa = 1$, the exponential when $\kappa = 1$ and $\sigma = 1$, and the lognormal distribution when $\kappa = 0$. The generalized gamma model is, therefore, commonly used for evaluating and selecting an appropriate parametric model for the data. The Wald or likelihood-ratio test can be used to test the hypotheses that $\kappa = 1$ or that $\kappa = 0$.

□ Technical note

Prior to Stata 14, `streg`'s option `distribution(gamma)` was used to fit generalized gamma models. As of Stata 14, the new option for fitting these models is `distribution(ggamma)`. The old option continues to work under version control. This option was renamed to avoid confusion with `mestreg`'s option `distribution(gamma)` for fitting mixed-effects survival gamma models; see [\[ME\] mestreg](#).

□

Examples

▷ Example 1

The Weibull distribution provides a good illustration of `streg` because this distribution is parameterized as both AFT and PH and serves to compare and contrast the two approaches.

We wish to analyze an experiment testing the ability of emergency generators with new-style bearings to withstand overloads. This dataset is described in [\[ST\] stcox](#). This time, we wish to fit a Weibull model:

```
. use http://www.stata-press.com/data/r14/kva
(Generator experiment)
. streg load bearings, distribution(weibull)
      failure _d: 1 (meaning all fail)
      analysis time _t: failtime
```

Fitting constant-only model:

```
Iteration 0: log likelihood = -13.666193
Iteration 1: log likelihood = -9.7427276
Iteration 2: log likelihood = -9.4421169
Iteration 3: log likelihood = -9.4408287
Iteration 4: log likelihood = -9.4408286
```

Fitting full model:

```
Iteration 0: log likelihood = -9.4408286
Iteration 1: log likelihood = -2.078323
Iteration 2: log likelihood = 5.2226016
Iteration 3: log likelihood = 5.6745808
Iteration 4: log likelihood = 5.6934031
Iteration 5: log likelihood = 5.6934189
Iteration 6: log likelihood = 5.6934189
```

Weibull regression -- log relative-hazard form

```
No. of subjects =          12          Number of obs   =          12
No. of failures =          12
Time at risk   =          896
Log likelihood =    5.6934189          LR chi2(2)       =          30.27
                                          Prob > chi2     =          0.0000
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
load	1.599315	.1883807	3.99	0.000	1.269616	2.014631
bearings	.1887995	.1312109	-2.40	0.016	.0483546	.7371644
_cons	2.51e-20	2.66e-19	-4.26	0.000	2.35e-29	2.68e-11
/ln_p	2.051552	.2317074	8.85	0.000	1.597414	2.505691
p	7.779969	1.802677			4.940241	12.25202
1/p	.1285352	.0297826			.0816192	.2024193

Because we did not specify otherwise, the estimation took place in the hazard metric, which is the default for `distribution(weibull)`. The estimates are directly comparable to those produced by `stcox`: `stcox` estimated a hazard ratio of 1.526 for load and 0.0636 for bearings.

However, we estimated the baseline hazard function as well, assuming that it is Weibull. The estimates are the full maximum-likelihood estimates. The shape parameter is fit as $\ln p$, but `streg` then reports p and $1/p = \sigma$ so that you can think about the parameter however you wish.

We find that p is greater than 1, which means that the hazard of failure increases with time and, here, increases dramatically. After 100 hours, the bearings are more than 1 million times more likely to fail per second than after 10 hours (or, to be precise, $(100/10)^{7.78-1}$). From our knowledge of generators, we would expect this; it is the accumulation of heat due to friction that causes bearings to expand and seize.

□ Technical note

Regression results are often presented in a metric other than the natural regression coefficients, that is, as hazard ratios, relative risk ratios, odds ratios, etc. In those cases, standard errors are calculated using the delta method.

However, the Z test and p -values given are calculated from the natural regression coefficients and standard errors. Although a test based on, say, a hazard ratio and its standard error would be asymptotically equivalent to that based on a regression coefficient, in real samples a hazard ratio will tend to have a more skewed distribution because it is an exponentiated regression coefficient. Also, it is more natural to think of these tests as testing whether a regression coefficient is nonzero, rather than testing whether a transformed regression coefficient is unequal to some nonzero value (one for a hazard ratio).

Finally, the confidence intervals given are obtained by transforming the endpoints of the corresponding confidence interval for the untransformed regression coefficient. This ensures that, say, strictly positive quantities such as hazard ratios have confidence intervals that do not overlap zero. □

▷ Example 2

The [previous estimation](#) took place in the PH metric, and exponentiated coefficients—hazard ratios—were reported. If we want to see the unexponentiated coefficients, we could redisplay results and specify the `nohr` option:

```
. streg, nohr
Weibull regression -- log relative-hazard form
No. of subjects =          12          Number of obs   =          12
No. of failures =          12
Time at risk   =          896
Log likelihood =    5.6934189          LR chi2(2)       =          30.27
                                          Prob > chi2      =          0.0000
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
load	.4695753	.1177884	3.99	0.000	.2387143	.7004363
bearings	-1.667069	.6949745	-2.40	0.016	-3.029194	-.3049443
_cons	-45.13191	10.60663	-4.26	0.000	-65.92053	-24.34329
/ln_p	2.051552	.2317074	8.85	0.000	1.597414	2.505691
p	7.779969	1.802677			4.940241	12.25202
1/p	.1285352	.0297826			.0816192	.2024193

▷ Example 3

We could just as well have fit this model in the AFT metric:

```
. streg load bearings, d(weibull) time nolog
      failure _d: 1 (meaning all fail)
      analysis time _t: failtime

Weibull regression -- accelerated failure-time form
No. of subjects =          12          Number of obs   =          12
No. of failures =          12
Time at risk   =          896
Log likelihood =    5.6934189          LR chi2(2)      =          30.27
                                          Prob > chi2    =          0.0000
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
load	-.060357	.0062214	-9.70	0.000	-.0725507	-.0481632
bearings	.2142771	.0746451	2.87	0.004	.0679753	.3605789
_cons	5.80104	.1752301	33.11	0.000	5.457595	6.144485
/ln_p	2.051552	.2317074	8.85	0.000	1.597414	2.505691
p	7.779969	1.802677			4.940241	12.25202
1/p	.1285352	.0297826			.0816192	.2024193

This is the same model we previously fit, but it is presented in a different metric. Calling the previous coefficients b , these coefficients are $-\sigma b = -b/p$. For instance, in the [previous example](#), the coefficient on load was reported as 0.4695753, and $-0.4695753/7.779969 = -0.06035696$. `d()` is a convenient shorthand for `distribution()`.

◀

▷ Example 4

`streg` may also be applied to more complicated data. Below we have multiple records per subject on a failure that can occur repeatedly:

```
. use http://www.stata-press.com/data/r14/mfail3
. stdescribe
```

Category	total	per subject			
		mean	min	median	max
no. of subjects	926				
no. of records	1734	1.87257	1	2	4
(first) entry time		0	0	0	0
(final) exit time		470.6857	1	477	960
subjects with gap	6				
time on gap if gap	411	68.5	16	57.5	133
time at risk	435444	470.2419	1	477	960
failures	808	.8725702	0	1	3

In this dataset, subjects have up to four records (most have two) and have up to three failures (most have one) and, although you cannot tell from the above output, the data have time-varying covariates, as well. There are even six subjects with gaps in their histories, meaning that, for a while, they went unobserved. Although we could estimate in the AFT metric, it is easier to interpret results in the PH metric (or the log relative-hazard metric, as it is also known):

```

. streg x1 x2, d(weibull) vce(robust)
Fitting constant-only model:
Iteration 0:   log pseudolikelihood = -1398.2504
Iteration 1:   log pseudolikelihood = -1382.8224
Iteration 2:   log pseudolikelihood = -1382.7457
Iteration 3:   log pseudolikelihood = -1382.7457
Fitting full model:
Iteration 0:   log pseudolikelihood = -1382.7457
Iteration 1:   log pseudolikelihood = -1328.4186
Iteration 2:   log pseudolikelihood = -1326.4483
Iteration 3:   log pseudolikelihood = -1326.4449
Iteration 4:   log pseudolikelihood = -1326.4449
Weibull regression -- log relative-hazard form
No. of subjects      =           926           Number of obs      =           1,734
No. of failures      =             808
Time at risk         =           435444
Log pseudolikelihood = -1326.4449           Wald chi2(2)       =           154.45
                                                    Prob > chi2        =           0.0000
                                                    (Std. Err. adjusted for 926 clusters in id)

```

_t	Robust		z	P> z	[95% Conf. Interval]	
	Haz. Ratio	Std. Err.				
x1	2.240069	.1812848	9.97	0.000	1.911504	2.625111
x2	.3206515	.0504626	-7.23	0.000	.2355458	.436507
_cons	.0006962	.0001792	-28.25	0.000	.0004204	.001153
/ln_p	.1771265	.0310111	5.71	0.000	.1163458	.2379071
p	1.193782	.0370205			1.123384	1.268591
1/p	.8376738	.0259772			.7882759	.8901674

A one-unit change in x_1 approximately doubles the hazard of failure, whereas a one-unit change in x_2 cuts the hazard to one-third its previous value. We also see that these data are close to being exponentially distributed; p is nearly 1.

Above we mentioned that interpreting results in the PH metric is easier, though regression coefficients are not difficult to interpret in the AFT metric. A positive coefficient means that time is decelerated by a unit increase in the covariate in question. This may seem awkward, but think of this instead as a unit increase in the covariate causing a delay in failure and thus *increasing* the expected time until failure.

The difficulty that arises with the AFT metric is merely that it places an emphasis on $\log(\text{time-to-failure})$ rather than risk (hazard) of failure. With this emphasis usually comes a desire to predict the time to failure, and therein lies the difficulty with complex survival data. Predicting the $\log(\text{time to failure})$ with `predict` assumes that the subject is at risk from time 0 until failure and has a fixed covariate pattern over this period. With these data, such assumptions produce predictions having little to do with the test subjects, who exhibit not only time-varying covariates but also multiple failures.

Predicting time to failure with complex survival data is difficult regardless of the metric under which estimation took place. Those who estimate in the PH metric are probably used to dealing with results from Cox regression, of which predicted time to failure is typically not the focus.

▷ Example 5

The multiple-failure data above are close enough to exponentially distributed that we will reestimate using exponential regression:

```
. streg x1 x2, d(exp) vce(robust)
Iteration 0:  log pseudolikelihood = -1398.2504
Iteration 1:  log pseudolikelihood = -1343.6083
Iteration 2:  log pseudolikelihood = -1341.5932
Iteration 3:  log pseudolikelihood = -1341.5893
Iteration 4:  log pseudolikelihood = -1341.5893
Exponential regression -- log relative-hazard form
No. of subjects      =           926           Number of obs      =           1,734
No. of failures      =           808
Time at risk        =          435444
Log pseudolikelihood = -1341.5893           Wald chi2(2)       =           166.92
                                                    Prob > chi2       =           0.0000
                                                    (Std. Err. adjusted for 926 clusters in id)
```

_t	Haz. Ratio	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
x1	2.19065	.1684399	10.20	0.000	1.884186	2.54696
x2	.3037259	.0462489	-7.83	0.000	.2253552	.4093511
_cons	.0024536	.0001535	-96.05	0.000	.0021704	.0027738

◀

□ Technical note

For our “complex” survival data, we specified `vce(robust)` when fitting the Weibull and exponential models. This was because these data were `stset` with an `id()` variable, and given the time-varying covariates and multiple failures, it is important not to assume that the observations within each subject are independent. When we specified `vce(robust)`, it was implicit that we were “clustering” on the groups defined by the `id()` variable.

You might sometimes have multiple observations per subject, which exist merely as a result of the data-organization mechanism and are not used to record gaps, time-varying covariates, or multiple failures. Such data could be collapsed into single-observation-per-subject data with no loss of information. In these cases, we refer to splitting the observations to form multiple observations per subject as *noninformative*. When the episode-splitting is noninformative, the model-based (nonrobust) standard errors produced will be the same as those produced when the data are collapsed into single records per subject. Thus, for these type of data, clustering of these multiple observations that results from specifying `vce(robust)` is not critical.

□

▷ Example 6

A reasonable question to ask is, “Given that we have several possible parametric models, how can we select one?” When parametric models are nested, the likelihood-ratio or Wald test can be used to discriminate between them. This can certainly be done for Weibull versus exponential or gamma versus Weibull or lognormal. When models are not nested, however, these tests are inappropriate, and the task of discriminating between models becomes more difficult. A common approach to this problem is to use the Akaike information criterion (AIC). Akaike (1974) proposed penalizing each log likelihood to reflect the number of parameters being estimated in a particular model and then comparing them. Here the AIC can be defined as

$$\text{AIC} = -2(\log \text{likelihood}) + 2(c + p + 1)$$

where c is the number of model covariates and p is the number of model-specific ancillary parameters listed in [table 1](#). Although the best-fitting model is the one with the largest log likelihood, the preferred model is the one with the smallest AIC value. The AIC value may be obtained by using the `estat ic` postestimation command; see [\[R\] estat ic](#).

Using `cancer.dta` distributed with Stata, let's first fit a generalized gamma model and test the hypothesis that $\kappa = 0$ (test for the appropriateness of the lognormal) and then test the hypothesis that $\kappa = 1$ (test for the appropriateness of the Weibull).

```
. use http://www.stata-press.com/data/r14/cancer
(Patient Survival in Drug Trial)
. stset studytime, failure(died)
  (output omitted)
. replace drug = drug==2 | drug==3 // 0, placebo : 1, nonplacebo
(48 real changes made)
. streg drug age, d(ggamma) nolog
      failure_d: died
      analysis time _t: studytime
Generalized gamma regression -- accelerated failure-time form
No. of subjects =          48                Number of obs   =          48
No. of failures =          31
Time at risk   =          744
Log likelihood = -42.452006                LR chi2(2)       =          36.07
                                                Prob > chi2     =          0.0000
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
drug	1.394658	.2557198	5.45	0.000	.893456 1.895859
age	-.0780416	.0227978	-3.42	0.001	-.1227245 -.0333587
_cons	6.456091	1.238457	5.21	0.000	4.02876 8.883421
/ln_sig	-.3793632	.183707	-2.07	0.039	-.7394222 -.0193041
/kappa	.4669252	.5419478	0.86	0.389	-.595273 1.529123
sigma	.684297	.1257101			.4773897 .980881

The Wald test of the hypothesis that $\kappa = 0$ (test for the appropriateness of the lognormal) is reported in the output above. The p -value is 0.389, suggesting that lognormal might be an adequate model for these data.

The Wald test for $\kappa = 1$ is

```
. test [kappa]_cons = 1
( 1) [kappa]_cons = 1
      chi2( 1) =    0.97
      Prob > chi2 =    0.3253
```

providing some support against rejecting the Weibull model.

We now fit the exponential, Weibull, loglogistic, and lognormal models separately. To directly compare coefficients, we will ask Stata to report the exponential and Weibull models in AFT form by specifying the `time` option. The output from fitting these models and the results from the generalized gamma model are summarized in [table 2](#).

Table 2. Results obtained from `streg`, using `cancer.dta` with `drug` as an indicator variable

Parameter	Exponential	Weibull	Lognormal	Loglogistic	Generalized gamma
Age	-0.0886715	-0.0714323	-0.0833996	-0.0803289	-0.078042
Drug	1.682625	1.305563	1.445838	1.420237	1.394658
Constant	7.146218	6.289679	6.580887	6.446711	6.456091
Ancillary		1.682751	0.751136	0.429276	0.684297
Kappa					0.466925
Log likelihood	-48.397094	-42.931335	-42.800864	-43.21698	-42.452006
AIC	102.7942	93.86267	93.60173	94.43396	94.90401

The largest log likelihood was obtained for the generalized gamma model; however, the lognormal model is preferred by the AIC.

◀

Parameterization of ancillary parameters

By default, all ancillary parameters are estimated as being constant. For example, the ancillary parameter, p , of the Weibull distribution is assumed to be a constant that is not dependent on any covariates. `streg`'s `ancillary()` and `anc2()` options allow for complete parameterization of parametric survival models. By specifying, for example,

```
. streg x1 x2, d(weibull) ancillary(x2 z1 z2)
```

both λ and the ancillary parameter, p , are parameterized in terms of covariates.

Ancillary parameters are usually restricted to be strictly positive, in which case the logarithm of the ancillary parameter is modeled using a linear predictor, which can assume any value on the real line.

▶ Example 7

Consider a dataset in which we model the time until hip fracture as Weibull for patients on the basis of age, sex, and whether the patient wears a hip-protective device (variable `protect`). We believe that the hazard is scaled according to sex and the presence of the device but believe the hazards for both sexes to be of different *shapes*.

```

. use http://www.stata-press.com/data/r14/hip3, clear
(hip fracture study)
. streg protect age, d(weibull) ancillary(male) nolog
      failure _d: fracture
      analysis time _t: time1
      id: id

Weibull regression -- log relative-hazard form
No. of subjects =          148          Number of obs   =          206
No. of failures =           37
Time at risk   =          1703
Log likelihood = -69.323532          LR chi2(2)       =          39.80
                                          Prob > chi2     =          0.0000

```

	_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
	_t						
	protect	-2.130058	.3567005	-5.97	0.000	-2.829178	-1.430938
	age	.0939131	.0341107	2.75	0.006	.0270573	.1607689
	_cons	-10.17575	2.551821	-3.99	0.000	-15.17722	-5.174269
	ln_p						
	male	-.4887189	.185608	-2.63	0.008	-.8525039	-.1249339
	_cons	.4540139	.1157915	3.92	0.000	.2270667	.6809611

From our estimation results, we see that $\ln(\hat{p}) = 0.454$ for females and $\ln(\hat{p}) = 0.454 - 0.489 = -0.035$ for males. Thus $\hat{p} = 1.57$ for females and $\hat{p} = 0.97$ for males. When we combine this with the main equation in the model, the estimated hazards are then

$$\hat{h}(t_j | \mathbf{x}_j) = \begin{cases} \exp(-10.18 - 2.13\text{protect}_j + 0.09\text{age}_j) 1.57t_j^{0.57} & \text{if female} \\ \exp(-10.18 - 2.13\text{protect}_j + 0.09\text{age}_j) 0.97t_j^{-0.03} & \text{if male} \end{cases}$$

If we believe this model, we would say that the hazard for males given age and protect is almost constant over time.

Contrast this with what we obtain if we type

```

. streg protect age if male, d(weibull)
. streg protect age if !male, d(weibull)

```

which is completely general, because not only will the shape parameter, p , differ over both sexes, but the regression coefficients will as well.

◀

The `anc2()` option is for use only with the gamma regression model, because it contains two ancillary parameters—`anc2()` is used to parametrize κ .

Stratified estimation

When we type

```

. streg xvars, d(distname) strata(varname)

```

we are asking that a completely stratified model be fit. By *completely stratified*, we mean that both the model's intercept and any ancillary parameters are allowed to vary for each level of the strata variable. That is, we are constraining the coefficients on the covariates to be the same across strata but allowing the intercept and ancillary parameters to vary.

▷ Example 8

We demonstrate this by fitting a stratified Weibull model to the cancer data, with the drug variable left in its original state: `drug==1` refers to the placebo, and `drug==2` and `drug==3` correspond to two alternative treatments.

```
. use http://www.stata-press.com/data/r14/cancer
(Patient Survival in Drug Trial)
. stset studytime, failure(died)
(output omitted)
. streg age, d(weibull) strata(drug) nolog
```

```
failure _d: died
analysis time _t: studytime
```

```
Weibull regression -- log relative-hazard form
```

```
No. of subjects =          48      Number of obs   =          48
No. of failures =           31
Time at risk   =          744
Log likelihood = -41.113074      LR chi2(3)     =          16.58
                                      Prob > chi2    =          0.0009
```

	_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_t	age	.1212332	.0367538	3.30	0.001	.049197	.1932694
	_Sdrug_2	-4.561178	2.339448	-1.95	0.051	-9.146411	.0240556
	_Sdrug_3	-3.715737	2.595986	-1.43	0.152	-8.803776	1.372302
	_cons	-10.36921	2.341022	-4.43	0.000	-14.95753	-5.780896
	ln_p						
ln_p	_Sdrug_2	.4872195	.332019	1.47	0.142	-.1635257	1.137965
	_Sdrug_3	.2194213	.4079989	0.54	0.591	-.5802418	1.019084
	_cons	.4541282	.1715663	2.65	0.008	.1178645	.7903919

◀

Completely stratified models are fit by first generating stratum-specific indicator variables (dummy variables) and then adding these as independent variables in the model and as covariates in the ancillary parameter. The `strata()` option is thus merely a shorthand method for generating the indicator variables from the drug categories and then placing these indicators in *both* the main equation and the ancillary equation(s).

We associate the term *stratification* with this process by noting that the intercept term of the main equation is a component of the baseline hazard (or baseline survivor) function. By allowing this intercept, as well as the ancillary shape parameter, to vary with respect to the strata, we allow the baseline functions to completely vary over the strata, analogous to a stratified Cox model.

▷ Example 9

We can produce a less-stratified model by specifying a factor variable in the `ancillary()` option.

```
. streg age, d(weibull) ancillary(i.drug) nolog
      failure _d: died
      analysis time _t: studytime
Weibull regression -- log relative-hazard form
No. of subjects =          48          Number of obs   =          48
No. of failures =          31
Time at risk   =          744
Log likelihood = -44.596379          LR chi2(1)       =          9.61
                                          Prob > chi2    =          0.0019
```

	_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
	_t						
	age	.1126419	.0362786	3.10	0.002	.0415373	.1837466
	_cons	-10.95772	2.308489	-4.75	0.000	-15.48227	-6.433162
	ln_p						
	drug						
	2	-.3279568	.11238	-2.92	0.004	-.5482176	-.107696
	3	-.4775351	.1091141	-4.38	0.000	-.6913948	-.2636755
	_cons	.6684086	.1327284	5.04	0.000	.4082657	.9285514

By doing this, we are restricting not only the coefficients on the covariates to be the same across “strata” but also the intercept, while allowing only the ancillary parameter to differ.

◀

By using `ancillary()` or `strata()`, we may thus consider a wide variety of models, depending on what we believe about the effect of the covariate(s) in question. For example, when fitting a Weibull PH model to the cancer data, we may choose from many models, depending on what we want to assume is the effect of the categorical variable `drug`. For all models considered below, we assume implicitly that the effect of `age` is proportional on the hazard function.

1. `drug` has no effect:

```
. streg age, d(weibull)
```

2. The effect of `drug` is proportional on the hazard (scale), and the effect of `age` is the same for each level of `drug`:

```
. streg age i.drug, d(weibull)
```

3. `drug` affects the shape of the hazard, and the effect of `age` is the same for each level of `drug`:

```
. streg age, d(weibull) ancillary(i.drug)
```

4. `drug` affects both the scale and shape of the hazard, and the effect of `age` is the same for each level of `drug`:

```
. streg age, d(weibull) strata(drug)
```

5. drug affects both the scale and shape of the hazard, and the effect of age is different for each level of drug:

```
. streg drug##c.age, d(weibull) strata(drug)
```

These models may be compared using Wald or likelihood-ratio tests when the models in question are nested (such as 3 nested within 4) or by using the AIC for nonnested models.

Everything we said regarding the modeling of ancillary parameters and stratification applies to AFT models as well, for which interpretations may be stated in terms of the baseline survivor function, that is, the unaccelerated probability of survival past time t .

□ Technical note

When fitting PH models, `streg` will, by default, display the exponentiated regression coefficients, labeled as hazard ratios. However, in our previous examples using `ancillary()` and `strata()`, the regression outputs displayed the untransformed coefficients instead. This change in behavior has to do with the modeling of the ancillary parameter. When we use one or more covariates from the main equation to model an ancillary parameter, hazard ratios (and time ratios for AFT models) lose their interpretation. `streg`, as a precaution, disallows the display of hazard/time ratios when `ancillary()`, `anc2()`, or `strata()` is specified.

Keep this in mind when comparing results across various model specifications. For example, when comparing a stratified Weibull PH model to a standard Weibull PH model, be sure that the latter is displayed using the `nohr` option. □

(Unshared-) frailty models

A frailty model is a survival model with unobservable heterogeneity, or *frailty*. At the observation level, frailty is introduced as an unobservable multiplicative effect, α , on the hazard function, such that

$$h(t|\alpha) = \alpha h(t)$$

where $h(t)$ is a nonfrailty hazard function, say, the hazard function of any of the six parametric models supported by `streg` described earlier in this entry. The frailty, α , is a random positive quantity and, for model identifiability, is assumed to have mean 1 and variance θ .

Exploiting the relationship between the cumulative hazard function and survivor function yields the expression for the survivor function, given the frailty

$$S(t|\alpha) = \exp \left\{ - \int_0^t h(u|\alpha) du \right\} = \exp \left\{ -\alpha \int_0^t \frac{f(u)}{S(u)} du \right\} = \{S(t)\}^\alpha$$

where $S(t)$ is the survivor function that corresponds to $h(t)$.

Because α is unobservable, it must be integrated out of $S(t|\alpha)$ to obtain the unconditional survivor function. Let $g(\alpha)$ be the probability density function of α , in which case an estimable form of our frailty model is achieved as

$$S_\theta(t) = \int_0^\infty S(t|\alpha) g(\alpha) d\alpha = \int_0^\infty \{S(t)\}^\alpha g(\alpha) d\alpha$$

Given the unconditional survivor function, we can obtain the unconditional hazard and density in the usual way:

$$f_{\theta}(t) = -\frac{d}{dt}S_{\theta}(t) \quad h_{\theta}(t) = \frac{f_{\theta}(t)}{S_{\theta}(t)}$$

Hence, an unshared-frailty model is merely a typical parametric survival model, with the additional estimation of an overdispersion parameter, θ . In a standard survival regression, the likelihood calculations are based on $S(t)$, $h(t)$, and $f(t)$. In an unshared-frailty model, the likelihood is based analogously on $S_{\theta}(t)$, $h_{\theta}(t)$, and $f_{\theta}(t)$.

At this stage, the only missing piece is the choice of frailty distribution, $g(\alpha)$. In theory, any continuous distribution supported on the positive numbers that has expectation 1 and finite variance θ is allowed here. For mathematical tractability, however, we limit the choice to either the gamma($1/\theta, \theta$) distribution or the inverse-Gaussian distribution with parameters 1 and $1/\theta$, denoted as IG($1, 1/\theta$). The gamma(a, b) distribution has probability density function

$$g(x) = \frac{x^{a-1}e^{-x/b}}{\Gamma(a)b^a}$$

and the IG(a, b) distribution has density

$$g(x) = \left(\frac{b}{2\pi x^3}\right)^{1/2} \exp\left\{-\frac{b}{2a}\left(\frac{x}{a} - 2 + \frac{a}{x}\right)\right\}$$

Therefore, performing the integrations described above will show that specifying `frailty(gamma)` will result in the frailty survival model (in terms of the nonfrailty survivor function, $S(t)$)

$$S_{\theta}(t) = [1 - \theta \log \{S(t)\}]^{-1/\theta}$$

Specifying `frailty(invgaussian)` will give

$$S_{\theta}(t) = \exp\left\{\frac{1}{\theta}\left(1 - [1 - 2\theta \log \{S(t)\}]^{1/2}\right)\right\}$$

Regardless of the choice of frailty distribution, $\lim_{\theta \rightarrow 0} S_{\theta}(t) = S(t)$, and thus the frailty model reduces to $S(t)$ when there is no heterogeneity present.

When using frailty models, distinguish between the hazard faced by the individual (subject), $\alpha h(t)$, and the “average” hazard for the population, $h_{\theta}(t)$. Similarly, an individual will have probability of survival past time t equal to $\{S(t)\}^{\alpha}$, whereas $S_{\theta}(t)$ will measure the proportion of the population surviving past time t . You specify $S(t)$ as before with `distribution(distname)`, and the list of possible parametric forms for $S(t)$ is given in [table 1](#). Thus when you specify `distribution()` you are specifying a model for an individual with frailty equal to one. Specifying `frailty(distname)` determines which of the two above forms for $S_{\theta}(t)$ is used.

The output of the estimation remains unchanged from the nonfrailty version, except for the additional estimation of θ and a likelihood-ratio test of $H_0: \theta = 0$. For more information on frailty models, [Hougaard \(1986\)](#) offers an excellent introduction. For a Stata-specific overview, see [Gutierrez \(2002\)](#).

► Example 10

Consider as an example a survival analysis of data on women with breast cancer. Our hypothetical dataset consists of analysis times on 80 women with covariates `age`, `smoking`, and `dietfat`, which measures the average weekly calories from fat ($\times 10^3$) in the patient's diet over the course of the study.

```
. use http://www.stata-press.com/data/r14/bc
. list in 1/12
```

	age	smoking	dietfat	t	dead
1.	30	1	4.919	14.2	0
2.	50	0	4.437	8.21	1
3.	47	0	5.85	5.64	1
4.	49	1	5.149	4.42	1
5.	52	1	4.363	2.81	1
6.	29	0	6.153	35	0
7.	49	1	3.82	4.57	1
8.	27	1	5.294	35	0
9.	47	0	6.102	3.74	1
10.	59	0	4.446	2.29	1
11.	35	0	6.203	15.3	0
12.	26	0	4.515	35	0

The data are well fit by a Weibull model for the distribution of survival time conditional on age, smoking, and dietary fat. By omitting the `dietfat` variable from the model, we hope to introduce unobserved heterogeneity.

```
. stset t, fail(dead)
  (output omitted)
. streg age smoking, d(weibull) frailty(gamma)
      failure _d:  dead
      analysis time _t:  t
```

Fitting Weibull model:

Fitting constant-only model:

```
Iteration 0:  log likelihood = -137.15363
Iteration 1:  log likelihood = -136.3927
Iteration 2:  log likelihood = -136.01557
Iteration 3:  log likelihood = -136.01202
Iteration 4:  log likelihood = -136.01201
```

Fitting full model:

```
Iteration 0:  log likelihood = -85.933969
Iteration 1:  log likelihood = -73.61173
Iteration 2:  log likelihood = -68.999447
Iteration 3:  log likelihood = -68.340858
Iteration 4:  log likelihood = -68.136187
Iteration 5:  log likelihood = -68.135804
Iteration 6:  log likelihood = -68.135804
```

```

Weibull regression -- log relative-hazard form
                    Gamma frailty
No. of subjects =      80                Number of obs   =      80
No. of failures =      58
Time at risk   =    1257.07
Log likelihood  =   -68.135804           LR chi2(2)       =    135.75
                                                Prob > chi2     =     0.0000

```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age	1.475948	.1379987	4.16	0.000	1.228811	1.772788
smoking	2.788548	1.457031	1.96	0.050	1.00143	7.764894
_cons	4.57e-11	2.38e-10	-4.57	0.000	1.70e-15	1.23e-06
/ln_p	1.087761	.222261	4.89	0.000	.6521376	1.523385
/ln_the	.3307466	.5250758	0.63	0.529	-.698383	1.359876
p	2.967622	.6595867			1.91964	4.587727
1/p	.3369701	.0748953			.2179729	.520931
theta	1.392007	.7309092			.4973889	3.895711

```

LR test of theta=0: chibar2(01) = 22.57                Prob >= chibar2 = 0.000

```

We could also use an inverse-Gaussian distribution to model the heterogeneity.

```

. streg age smoking, d(weibull) frailty(invgauss) nolog

```

```

    failure _d: dead
  analysis time _t: t

```

```

Weibull regression -- log relative-hazard form
                    Inverse-Gaussian frailty
No. of subjects =      80                Number of obs   =      80
No. of failures =      58
Time at risk   =    1257.07
Log likelihood  =   -73.838578           LR chi2(2)       =    125.44
                                                Prob > chi2     =     0.0000

```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age	1.284133	.0463256	6.93	0.000	1.196473	1.378217
smoking	2.905409	1.252785	2.47	0.013	1.247892	6.764528
_cons	1.11e-07	2.34e-07	-7.63	0.000	1.83e-09	6.79e-06
/ln_p	.7173904	.1434382	5.00	0.000	.4362567	.9985241
/ln_the	.2374778	.8568064	0.28	0.782	-1.441832	1.916788
p	2.049079	.2939162			1.546906	2.714273
1/p	.4880241	.0700013			.3684228	.6464518
theta	1.268047	1.086471			.2364941	6.799082

```

LR test of theta=0: chibar2(01) = 11.16                Prob >= chibar2 = 0.000

```

The results are similar with respect to the choice of frailty distribution, with the gamma frailty model producing a slightly higher likelihood. Both models show a statistically significant level of unobservable heterogeneity because the p -value for the likelihood-ratio (LR) test of $H_0: \theta = 0$ is virtually zero in both cases.

□ Technical note

With gamma-distributed or inverse-Gaussian-distributed frailty, hazard ratios decay over time in favor of the *frailty effect*, and thus the displayed “Haz. Ratio” in the above output is actually the hazard ratio only for $t = 0$. The degree of decay depends on θ . Should the estimated θ be close to zero, the hazard ratios regain their usual interpretation. The rate of decay and the limiting hazard ratio differ between the gamma and inverse-Gaussian models; see [Gutierrez \(2002\)](#) for details.

For this reason, many researchers prefer fitting frailty models in the AFT metric because the interpretation of regression coefficients is unchanged by the frailty—the factors in question serve to either accelerate or decelerate the survival experience. The only difference is that with frailty models, the unconditional probability of survival is described by $S_\theta(t)$ rather than $S(t)$.

□

□ Technical note

The LR test of $\theta = 0$ is a boundary test and thus requires careful consideration concerning the calculation of its p -value. In particular, the null distribution of the LR test statistic is not the usual χ_1^2 but rather is a 50:50 mixture of a χ_0^2 (point mass at zero) and a χ_1^2 , denoted as $\bar{\chi}_{01}^2$. See [Gutierrez, Carter, and Drukker \(2001\)](#) for more details.

□

To verify that the significant heterogeneity is caused by the omission of `dietfat`, we now refit the Weibull/inverse-Gaussian frailty model with `dietfat` included.

```
. streg age smoking dietfat, d(weibull) frailty(invgauss) nolog
      failure _d:  dead
      analysis time _t:  t
Weibull regression -- log relative-hazard form
                    Inverse-Gaussian frailty
No. of subjects =          80          Number of obs   =          80
No. of failures =          58
Time at risk   =       1257.07
Log likelihood =    -13.352142
LR chi2(3)     =          246.41
Prob > chi2    =          0.0000
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age	1.74928	.0985246	9.93	0.000	1.566453	1.953447
smoking	5.203552	1.704943	5.03	0.000	2.737814	9.889992
dietfat	9.229842	2.219331	9.24	0.000	5.761312	14.78656
_cons	1.07e-20	4.98e-20	-9.92	0.000	1.22e-24	9.45e-17
/ln_p	1.431742	.0978847	14.63	0.000	1.239892	1.623593
/ln_the	-14.29793	2673.364	-0.01	0.996	-5253.995	5225.399
p	4.185987	.4097439			3.45524	5.071278
1/p	.2388923	.0233839			.197189	.2894155
theta	6.17e-07	.0016502			0	.

```
LR test of theta=0:  chibar2(01) = 0.00          Prob >= chibar2 = 1.000
```

The estimate of the frailty variance component θ is near zero, and the p -value of the test of $H_0: \theta = 0$ equals one, indicating negligible heterogeneity. A regular Weibull model could be fit to these data (with `dietfat` included), producing almost identical estimates of the hazard ratios and ancillary parameter, p , so such an analysis is omitted here.

Also hazard ratios now regain their original interpretation. Thus an increase in weekly calories from fat of 1,000 would increase the risk of death by more than ninefold.

◀

Shared-frailty models

A generalization of the frailty models considered in the previous section is the *shared-frailty* model, where the frailty is assumed to be group specific; this is analogous to a panel-data regression model. For observation j from the i th group, the hazard is

$$h_{ij}(t|\alpha_i) = \alpha_i h_{ij}(t)$$

for $i = 1, \dots, n$ and $j = 1, \dots, n_i$, where by $h_{ij}(t)$ we mean $h(t|\mathbf{x}_{ij})$, which is the individual hazard given covariates \mathbf{x}_{ij} .

Shared-frailty models are appropriate when you wish to model the frailties as being specific to groups of subjects, such as subjects within families. Here a shared-frailty model may be used to model the degree of correlation within groups; that is, the subjects within a group are correlated because they share the same common frailty.

▶ Example 11

Consider the data from a study of 38 kidney dialysis patients, as described in [McGilchrist and Aisbett \(1991\)](#). The study is concerned with the prevalence of infection at the catheter-insertion point. Two recurrence times (in days) are measured for each patient, and each recorded time is the time from initial insertion (onset of risk) to infection or censoring.

```
. use http://www.stata-press.com/data/r14/catheter
(Kidney data, McGilchrist and Aisbett, Biometrics, 1991)
. list patient time infect age female in 1/10
```

	patient	time	infect	age	female
1.	1	16	1	28	0
2.	1	8	1	28	0
3.	2	13	0	48	1
4.	2	23	1	48	1
5.	3	22	1	32	0
6.	3	28	1	32	0
7.	4	318	1	31.5	1
8.	4	447	1	31.5	1
9.	5	30	1	10	0
10.	5	12	1	10	0

Each patient (`patient`) has two recurrence times (`time`) recorded, with each catheter insertion resulting in either infection (`infect==1`) or right-censoring (`infect==0`). Among the covariates measured are age and sex (`female==1` if female, `female==0` if male).

One subtlety to note concerns the use of the generic term *subjects*. In this example, the subjects are the individual catheter insertions, not the patients themselves. This is a function of how the data were recorded—the onset of risk occurs at catheter insertion (of which there are two for each patient) not, say, at the time of admission of the patient into the study. Thus we have two subjects (insertions) within each group (patient).

It is reasonable to assume independence of patients but unreasonable to assume that recurrence times within each patient are independent. One solution would be to fit a standard survival model, adjusting the standard errors of the parameter estimates to account for the possible correlation by specifying `vce(cluster patient)`.

We could also model the correlation by assuming that the correlation is the result of a latent patient-level effect, or frailty. That is, rather than fitting a standard model and specifying `vce(cluster patient)`, we fit a frailty model and specify `shared(patient)`. Assuming that the time to infection, given age and female, follows a Weibull distribution, and inverse-Gaussian distributed frailties, we get

```
. stset time, fail(infect)
  (output omitted)
. streg age female, d(weibull) frailty(invgauss) shared(patient) nolog
      failure _d: infect
      analysis time _t: time

Weibull regression --
      log relative-hazard form          Number of obs    =          76
      Inverse-Gaussian shared frailty    Number of groups  =          38
Group variable: patient

Obs per group:
No. of subjects =          76                min =          2
No. of failures =          58                avg  =          2
Time at risk    =         7424                max  =          2

LR chi2(2)      =          9.84
Log likelihood  = -99.093527                 Prob > chi2     =          0.0073
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age	1.006918	.013574	0.51	0.609	.9806623	1.033878
female	.2331376	.1046382	-3.24	0.001	.0967322	.5618928
_cons	.0110089	.0099266	-5.00	0.000	.0018803	.0644557
/ln_p	.1900625	.1315342	1.44	0.148	-.0677398	.4478649
/ln_the	.0357272	.7745362	0.05	0.963	-1.482336	1.55379
p	1.209325	.1590676			.9345036	1.564967
1/p	.8269074	.1087666			.638991	1.070087
theta	1.036373	.8027085			.2271066	4.729362

LR test of theta=0: $\text{chibar2}(01) = 8.70$

Prob >= $\text{chibar2} = 0.002$

Contrast this with what we obtain by assuming a subject-level lognormal model:

```
. streg age female, d(lnormal) frailty(invgauss) shared(patient) nolog
      failure _d: infect
      analysis time _t: time
Lognormal regression --
      accelerated failure-time form      Number of obs      =      76
      Inverse-Gaussian shared frailty   Number of groups   =      38
Group variable: patient

Obs per group:
      min =      2
      avg =      2
      max =      2
No. of subjects =      76
No. of failures =      58
Time at risk   =      7424
Log likelihood = -97.614583
LR chi2(2)     =      16.34
Prob > chi2    =      0.0003
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	-.0066762	.0099457	-0.67	0.502	-.0261694	.0128171
female	1.401719	.3334931	4.20	0.000	.7480844	2.055354
_cons	3.336709	.4972641	6.71	0.000	2.362089	4.311329
/ln_sig	.0625872	.1256185	0.50	0.618	-.1836205	.3087949
/ln_the	-1.606248	1.190775	-1.35	0.177	-3.940125	.7276282
sigma	1.064587	.1337318			.8322516	1.361783
theta	.2006389	.2389159			.0194458	2.070165

LR test of theta=0: $\text{chibar2}(01) = 1.53$ Prob >= $\text{chibar2} = 0.108$

The frailty effect is insignificant at the 10% level in the latter model yet highly significant in the former. We thus have two possible stories to tell concerning these data: If we believe the first model, we believe that the individual hazard of infection continually rises over time (Weibull), but there is a significant frailty effect causing the population hazard to begin falling after some time. If we believe the second model, we believe that the individual hazard first rises and then declines (lognormal), meaning that if a given insertion does not become infected initially, the chances that it will become infected begin to decrease after a certain point. Because the frailty effect is insignificant, the population hazard mirrors the individual hazard in the second model.

As a result, both models view the population hazard as rising initially and then falling past a certain point. The second version of our story corresponds to higher log likelihood, yet perhaps not significantly higher given the limited data. More investigation is required. One idea is to fit a more distribution-agnostic form of a frailty model, such as a piecewise exponential (Cleves, Gould, and Marchenko 2016, 345–348) or a Cox model with frailty; see [ST] **stcox**.

◀

Shared-frailty models are also appropriate when the frailties are subject specific yet there exist multiple records per subject. Here you would share frailties across the same `id()` variable previously `stset`. When you have subject-specific frailties and uninformative episode splitting, it makes no difference whether you fit a shared or an unshared frailty model. The estimation results will be the same.

Stored results

streg stores the following in `e()`:

Scalars

<code>e(N)</code>	number of observations
<code>e(N_sub)</code>	number of subjects
<code>e(N_fail)</code>	number of failures
<code>e(N_g)</code>	number of groups
<code>e(k)</code>	number of parameters
<code>e(k_eq)</code>	number of equations in <code>e(b)</code>
<code>e(k_eq_model)</code>	number of equations in overall model test
<code>e(k_aux)</code>	number of auxiliary parameters
<code>e(k_dv)</code>	number of dependent variables
<code>e(df_m)</code>	model degrees of freedom
<code>e(ll)</code>	log likelihood
<code>e(ll_0)</code>	log likelihood, constant-only model
<code>e(ll_c)</code>	log likelihood, comparison model
<code>e(N_clust)</code>	number of clusters
<code>e(chi2)</code>	χ^2
<code>e(chi2_c)</code>	χ^2 , comparison model
<code>e(risk)</code>	total time at risk
<code>e(g_min)</code>	smallest group size
<code>e(g_avg)</code>	average group size
<code>e(g_max)</code>	largest group size
<code>e(theta)</code>	frailty parameter
<code>e(aux_p)</code>	ancillary parameter (<code>weibull</code>)
<code>e(gamma)</code>	ancillary parameter (<code>gompertz</code> , <code>loglogistic</code>)
<code>e(sigma)</code>	ancillary parameter (<code>ggamma</code> , <code>lnormal</code>)
<code>e(kappa)</code>	ancillary parameter (<code>ggamma</code>)
<code>e(p)</code>	significance
<code>e(p_c)</code>	significance, comparison model
<code>e(rank)</code>	rank of <code>e(V)</code>
<code>e(rank0)</code>	rank of <code>e(V)</code> , constant-only model
<code>e(ic)</code>	number of iterations
<code>e(rc)</code>	return code
<code>e(converged)</code>	1 if converged, 0 otherwise

Macros

<code>e(cmd)</code>	model or regression name
<code>e(cmd2)</code>	<code>streg</code>
<code>e(cmdline)</code>	command as typed
<code>e(dead)</code>	<code>_d</code>
<code>e(depvar)</code>	<code>_t</code>
<code>e(title)</code>	title in estimation output
<code>e(title2)</code>	secondary title in estimation output
<code>e(clustvar)</code>	name of cluster variable
<code>e(shared)</code>	frailty grouping variable
<code>e(fr_title)</code>	title in output identifying frailty
<code>e(wtype)</code>	weight type
<code>e(wexp)</code>	weight expression
<code>e(t0)</code>	<code>_t0</code>
<code>e(vce)</code>	<i>vce</i> type specified in <code>vce()</code>
<code>e(vctype)</code>	title used to label Std. Err.
<code>e(frm2)</code>	hazard or time
<code>e(chi2type)</code>	Wald or LR; type of model χ^2 test
<code>e(offset1)</code>	offset for main equation
<code>e(stcurve)</code>	<code>stcurve</code>
<code>e(opt)</code>	type of optimization
<code>e(which)</code>	max or min; whether optimizer is to perform maximization or minimization
<code>e(ml_method)</code>	type of ml method
<code>e(user)</code>	name of likelihood-evaluator program
<code>e(technique)</code>	maximization technique
<code>e(properties)</code>	<code>b V</code>
<code>e(predict)</code>	program used to implement <code>predict</code>
<code>e(predict_sub)</code>	<code>predict</code> subprogram
<code>e(footnote)</code>	program used to implement the footnote display
<code>e(asbalanced)</code>	factor variables <code>fvset</code> as <code>asbalanced</code>
<code>e(asobserved)</code>	factor variables <code>fvset</code> as <code>asobserved</code>

Matrices

<code>e(b)</code>	coefficient vector
<code>e(Cns)</code>	constraints matrix
<code>e(ilog)</code>	iteration log (up to 20 iterations)
<code>e(gradient)</code>	gradient vector
<code>e(V)</code>	variance–covariance matrix of the estimators
<code>e(V_modelbased)</code>	model-based variance

Functions

<code>e(sample)</code>	marks estimation sample
------------------------	-------------------------

Methods and formulas

For an introduction to survival models, see [Cleves, Gould, and Marchenko \(2016\)](#). For an introduction to survival analysis directed at social scientists, see [Box-Steffensmeier and Jones \(2004\)](#).

Consider for $j = 1, \dots, n$ observations the trivariate response, (t_{0j}, t_j, d_j) , representing a period of observation, $(t_{0j}, t_j]$, ending in either failure ($d_j = 1$) or right-censoring ($d_j = 0$). This structure allows analysis of a wide variety of models and may be used to account for delayed entry, gaps, time-varying covariates, and multiple failures per subject. Regardless of the structure of the data, once they are `stset`, the data may be treated in a common manner by `streg`: the `stset`-created variable `_t0` holds the t_{0j} , `_t` holds the t_j , and `_d` holds the d_j .

For a given survivor function, $S(t)$, the density function is obtained as

$$f(t) = -\frac{d}{dt}S(t)$$

and the hazard function (the instantaneous rate of failure) is obtained as $h(t) = f(t)/S(t)$. Available forms for $S(t)$ are listed in [table 1](#). For a set of covariates from the j th observation, \mathbf{x}_j , define $S_j(t) = S(t|\mathbf{x} = \mathbf{x}_j)$, and similarly define $h_j(t)$ and $f_j(t)$. For example, in a Weibull PH model, $S_j(t) = \exp\{-\exp(\mathbf{x}_j\boldsymbol{\beta})t^p\}$.

Parameter estimation

In this command, $\boldsymbol{\beta}$ and the ancillary parameters are estimated via maximum likelihood. A subject known to fail at time t_j contributes to the likelihood function the value of the density at time t_j conditional on the entry time t_{0j} , $f_j(t_j)/S_j(t_{0j})$. A censored observation, known to survive only up to time t_j , contributes $S_j(t_j)/S_j(t_{0j})$, which is the probability of surviving beyond time t_j conditional on the entry time, t_{0j} . The log likelihood is thus given by

$$\log L = \sum_{j=1}^n \{d_j \log f_j(t_j) + (1 - d_j) \log S_j(t_j) - \log S_j(t_{0j})\}$$

Implicit in the above log-likelihood expression are the regression parameters, $\boldsymbol{\beta}$, and the ancillary parameters because both are components of the chosen $S_j(t)$ and its corresponding $f_j(t)$; see [table 1](#). `streg` reports maximum likelihood estimates of $\boldsymbol{\beta}$ and of the ancillary parameters (if any for the chosen model). The reported log-likelihood value is $\log L_r = \log L + T$, where $T = \sum \log(t_j)$ is summed over uncensored observations. The adjustment removes the time units from $\log L$. Whether the adjustment is made makes no difference to any test or result since such tests and results depend on differences in log-likelihood functions or their second derivatives, or both.

Specifying `ancillary()`, `anc2()`, or `strata()` will parameterize the ancillary parameter(s) by using the linear predictor, $\mathbf{z}_j\boldsymbol{\alpha}_z$, where the covariates, \mathbf{z}_j , need not be distinct from \mathbf{x}_j . Here `streg` will report estimates of $\boldsymbol{\alpha}_z$ in addition to estimates of $\boldsymbol{\beta}$. The log likelihood here is simply the log likelihood given above, with $\mathbf{z}_j\boldsymbol{\alpha}_z$ substituted for the ancillary parameter. If the ancillary parameter is constrained to be strictly positive, its logarithm is parameterized instead; that is, we substitute the linear predictor for the logarithm of the ancillary parameter in the above log likelihood. The gamma model has two ancillary parameters, σ and κ ; we parameterize σ by using `ancillary()` and κ by using `anc2()`, and the linear predictors used for each may be distinct. Specifying `strata()` creates indicator variables for the strata, places these indicators in the main equation, and uses the indicators to parameterize any ancillary parameters that exist for the chosen model.

Unshared-frailty models have a log likelihood of the above form, with $S_\theta(t)$ and $f_\theta(t)$ substituted for $S(t)$ and $f(t)$, respectively. Equivalently, for gamma-distributed frailties,

$$\log L = \sum_{j=1}^n [\theta^{-1} \log \{1 - \theta \log S_j(t_{0j})\} - (\theta^{-1} + d_j) \log \{1 - \theta \log S_j(t_j)\} + d_j \log h_j(t_j)]$$

and for inverse-Gaussian-distributed frailties,

$$\log L = \sum_{j=1}^n \left[\theta^{-1} \{1 - 2\theta \log S_j(t_{0j})\}^{1/2} - \theta^{-1} \{1 - 2\theta \log S_j(t_j)\}^{1/2} + d_j \log h_j(t_j) - \frac{1}{2} d_j \log \{1 - 2\theta \log S_j(t_j)\} \right]$$

In a shared-frailty model, the frailty is common to a group of observations. Thus, to form an unconditional likelihood, the frailties must be integrated out at the group level. The data are organized as $i = 1, \dots, n$ groups with the i th group comprising $j = 1, \dots, n_i$ observations. The log likelihood is the sum of the log-likelihood contributions for each group. Define $D_i = \sum_j d_{ij}$ as the number of failures in the i th group. For gamma frailties, the log-likelihood contribution for the i th group is

$$\log L_i = \sum_{j=1}^{n_i} d_{ij} \log h_{ij}(t_{ij}) - (1/\theta + D_i) \log \left\{ 1 - \theta \sum_{j=1}^{n_i} \log \frac{S_{ij}(t_{ij})}{S_{ij}(t_{0ij})} \right\} + D_i \log \theta + \log \Gamma(1/\theta + D_i) - \log \Gamma(1/\theta)$$

This formula corresponds to the log-likelihood contribution for multiple-record data. For single-record data, the denominator $S_{ij}(t_{0ij})$ is equal to 1. This formula is not applicable to data with delayed entries or gaps.

For inverse-Gaussian frailties, define

$$C_i = \left\{ 1 - 2\theta \sum_{j=1}^{n_i} \log \frac{S_{ij}(t_{ij})}{S_{ij}(t_{0ij})} \right\}^{-1/2}$$

The log-likelihood contribution for the i th group then becomes

$$\log L_i = \theta^{-1} (1 - C_i^{-1}) + B(\theta C_i, D_i) + \sum_{j=1}^{n_i} d_{ij} \{ \log h_{ij}(t_{ij}) + \log C_i \}$$

The function $B(a, b)$ is related to the modified Bessel function of the third kind, commonly known as the BesselK function; see [Wolfram \(1999, 767\)](#). In particular,

$$B(a, b) = a^{-1} + \frac{1}{2} \left\{ \log \left(\frac{2}{\pi} \right) - \log a \right\} + \log \text{BesselK} \left(\frac{1}{2}, b, a^{-1} \right)$$

For both unshared- and shared-frailty models, estimation of θ takes place jointly with the estimation of β and the ancillary parameters.

This command supports the Huber/White/sandwich estimator of the variance and its clustered version using `vce(robust)` and `vce(cluster clustvar)`, respectively. See [\[P\] `_robust`](#), particularly [Maximum likelihood estimators](#) and [Methods and formulas](#). If observations in the dataset represent repeated observations on the same subjects (that is, there are time-varying covariates), the assumption of independence of the observations is highly questionable, meaning that the conventional estimate of variance is not appropriate. We strongly advise that you use the `vce(robust)` and `vce(cluster clustvar)` options here. (`streg` knows to specify `vce(cluster clustvar)` if you specify `vce(robust)`.) `vce(robust)` and `vce(cluster clustvar)` do not apply in shared-frailty models, where the correlation within groups is instead modeled directly.

`streg` also supports estimation with survey data. For details on VCEs with survey data, see [\[SVY\] variance estimation](#).

Benjamin Gompertz (1779–1865) came from a Jewish family who left Holland and settled in England. Excluded from a university education, he was self-educated in mathematics. In 1819, his publications in mathematics earned him an invitation to join the Royal Society. In 1824, he was appointed as actuary and head clerk of the Alliance Assurance Company.

Gompertz carried out pioneering work on the application of differential calculus to actuarial questions, particularly the dependence of mortality on age. He is credited with introducing, in 1825, the concept that mortality is a continuous function over time. From this idea came the notion of a survival function, and ultimately, parametric survival-time analysis. Gompertz’s work also had a strong influence on the practice of demography, where it is used in the study of parity and fertility.

Aside from his work in actuarial sciences, Gompertz contributed to astronomy and the study of astronomical instruments. He was a member of the Astronomical Society nearly from its founding in 1820. The society’s memoirs recognize him as an important contributor to the study of the aberration of light. He also helped to develop the society’s catalog of the stars and make improvements to its instruments, including the convertible pendulum, transit instruments for studying the position of stars, and the differential sextant, his own invention.

Ernst Hjalmar Waloddi Weibull (1887–1979) was a Swedish applied physicist most famous for his work on the statistics of material properties. He worked in Germany and Sweden as an inventor and a consulting engineer, publishing his first paper on the propagation of explosive waves in 1914, thereafter becoming a full professor at the Royal Institute of Technology in 1924. Weibull wrote two important papers, “Investigations into strength properties of brittle materials” and “The phenomenon of rupture in solids”, which discussed his ideas about the statistical distributions of material strength. These articles came to the attention of engineers in the late 1930s.

References

- Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716–723.
- Bottai, M., and N. Orsini. 2013. [A command for Laplace regression](#). *Stata Journal* 13: 302–314.
- Box-Steffensmeier, J. M., and B. S. Jones. 2004. *Event History Modeling: A Guide for Social Scientists*. Cambridge: Cambridge University Press.
- Cleves, M. A., W. W. Gould, and Y. V. Marchenko. 2016. *An Introduction to Survival Analysis Using Stata*. Rev. 3 ed. College Station, TX: Stata Press.
- Cox, D. R., and D. Oakes. 1984. *Analysis of Survival Data*. London: Chapman & Hall/CRC.
- Crowder, M. J., A. C. Kimber, R. L. Smith, and T. J. Sweeting. 1991. *Statistical Analysis of Reliability Data*. London: Chapman & Hall/CRC.
- Crowther, M. J., K. R. Abrams, and P. C. Lambert. 2013. Joint modeling of longitudinal and survival data. *Stata Journal* 13: 165–184.
- Cui, J. 2005. [Buckley–James method for analyzing censored data, with an application to a cardiovascular disease and an HIV/AIDS study](#). *Stata Journal* 5: 517–526.
- Fisher, R. A., and L. H. C. Tippett. 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society* 24: 180–190.
- Gutierrez, R. G. 2002. [Parametric frailty and shared frailty survival models](#). *Stata Journal* 2: 22–44.
- 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.

- Hooker, P. F. 1965. Benjamin Gompertz. *Journal of the Institute of Actuaries* 91: 203–212.
- Hosmer, D. W., Jr., S. A. Lemeshow, and S. May. 2008. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. 2nd ed. New York: Wiley.
- Hougaard, P. 1986. Survival models for heterogeneous populations derived from stable distributions. *Biometrika* 73: 387–396.
- Kalbfleisch, J. D., and R. L. Prentice. 2002. *The Statistical Analysis of Failure Time Data*. 2nd ed. New York: Wiley.
- Klein, J. P., and M. L. Moeschberger. 2003. *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd ed. New York: Springer.
- Lambert, P. C., and P. Royston. 2009. Further development of flexible parametric models for survival analysis. *Stata Journal* 9: 265–290.
- Lee, E. T., and J. W. Wang. 2003. *Statistical Methods for Survival Data Analysis*. 3rd ed. New York: Wiley.
- McGilchrist, C. A., and C. W. Aisbett. 1991. Regression with frailty in survival analysis. *Biometrics* 47: 461–466.
- Olshansky, S. J., and B. A. Carnes. 1997. Ever since Gompertz. *Demography* 34: 1–15.
- Peto, R., and P. Lee. 1973. Weibull distributions for continuous-carcinogenesis experiments. *Biometrics* 29: 457–470.
- Pike, M. C. 1966. A method of analysis of a certain class of experiments in carcinogenesis. *Biometrics* 22: 142–161.
- Royston, P. 2006. Explained variation for survival models. *Stata Journal* 6: 83–96.
- Royston, P., and P. C. Lambert. 2011. *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*. College Station, TX: Stata Press.
- Royston, P., and M. K. B. Parmar. 2002. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* 21: 2175–2197.
- Schoenfeld, D. A. 1982. Partial residuals for the proportional hazards regression model. *Biometrika* 69: 239–241.
- Scotto, M. G., and A. Tobías. 1998. [sg83: Parameter estimation for the Gumbel distribution](#). *Stata Technical Bulletin* 43: 32–35. Reprinted in *Stata Technical Bulletin Reprints*, vol. 8, pp. 133–137. College Station, TX: Stata Press.
- . 2000. [sg146: Parameter estimation for the generalized extreme value distribution](#). *Stata Technical Bulletin* 56: 40–43. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 205–210. College Station, TX: Stata Press.
- Weibull, W. 1939. A statistical theory of the strength of materials. In *Ingeniörs Vetenskaps Akademien Handlingar*, vol. 151. Stockholm: Generalstabens Litografiska Anstalts Förlag.
- Wolfram, S. 1999. *The Mathematica Book*. 4th ed. Cambridge: Cambridge University Press.

Also see

- [ST] [streg postestimation](#) — Postestimation tools for streg
- [ST] [stcurve](#) — Plot survivor, hazard, cumulative hazard, or cumulative incidence function
- [ST] [stcox](#) — Cox proportional hazards model
- [ST] [stcrreg](#) — Competing-risks regression
- [ST] [sts](#) — Generate, graph, list, and test the survivor and cumulative hazard functions
- [ST] [stset](#) — Declare data to be survival-time data
- [ME] [mestreg](#) — Multilevel mixed-effects parametric survival models
- [MI] [estimation](#) — Estimation commands for use with mi estimate
- [PSS] [power exponential](#) — Power analysis for the exponential test
- [SVY] [svy estimation](#) — Estimation commands for survey data
- [TE] [stteffects](#) — Treatment-effects estimation for observational survival-time data
- [XT] [xtstreg](#) — Random-effects parametric survival models
- [U] [20 Estimation and postestimation commands](#)