stintreg postestimation — Postestimation tools for stintreg

Postestimation commands

The following postestimation commands are of special interest after `stintreg`:

Command	Description
* `estat gofplot`	produce goodness-of-fit plot
`stcurve`	plot the survivor, failure, hazard, or cumulative hazard function

*`estat gofplot` is not appropriate with `svy` estimation results.

The following standard postestimation commands are also available:

Command	Description
`contrast`	contrasts and ANOVA-style joint tests of parameters
`estat ic`	Akaike's, consistent Akaike's, corrected Akaike's, and Schwarz's Bayesian information criteria (AIC, CAIC, AICc, and BIC, respectively)
`estat summarize`	summary statistics for the estimation sample
`estat vce`	variance–covariance matrix of the estimators (VCE)
`estat` (svy)	postestimation statistics for survey data
`estimates`	cataloging estimation results
`etable`	table of estimation results
* `hausman`	Hausman's specification test
`lincom`	point estimates, standard errors, testing, and inference for linear combinations of parameters
* `lrtest`	likelihood-ratio test
`margins`	marginal means, predictive margins, marginal effects, and average marginal effects
`marginsplot`	graph the results from margins (profile plots, interaction plots, etc.)
`nlcom`	point estimates, standard errors, testing, and inference for nonlinear combinations of parameters
`predict`	hazard ratios, survivor functions, influence statistics, residuals, etc.
`predictnl`	point estimates, standard errors, testing, and inference for generalized predictions
`pwcompare`	pairwise comparisons of parameters
`suest`	seemingly unrelated estimation
`test`	Wald tests of simple and composite linear hypotheses
`testnl`	Wald tests of nonlinear hypotheses

*`hausman` and `lrtest` are not appropriate with `svy` estimation results.

predict

Description for predict

predict creates a new variable containing predictions such as median and mean survival times, hazards, hazard ratios, linear predictions, standard errors, probabilities, and Cox–Snell-like and martingale-like residuals.

Menu for predict

Statistics > Postestimation

Syntax for predict

predict [*type*] *newvar* [*if*] [*in*] [, *statistic options*]

predict [*type*] $newvar_l$ $newvar_u$ [*if*] [*in*] , *statistic2* [*options*]

predict [*type*] *stub** [*if*] [*in*], scores

statistic	Description
Main	
median time	median survival time; the default
median lntime	median ln(survival time)
mean time	mean survival time
mean lntime	mean ln(survival time)
hr	hazard ratio, also known as the relative hazard
xb	linear prediction $\mathbf{x}_j\widehat{\boldsymbol{\beta}}$
stdp	standard error of the linear prediction; $\mathrm{SE}(\mathbf{x}_j\widehat{\boldsymbol{\beta}})$
* mgale	martingale-like residuals

statistic2	Description
Main	
hazard	hazard for interval endpoints t_l and t_u
surv	survivor probability for interval endpoints t_l and t_u
* csnell	Cox–Snell-like residuals for interval endpoints t_l and t_u

options	Description
Main	
nooffset	ignore the offset() variable specified in stintreg
oos	make *statistic* and *statistic2* available in and out of sample

Unstarred statistics are available both in and out of sample; type `predict ... if e(sample) ...` if wanted only for the estimation sample. Starred statistics are calculated for the estimation sample by default, but the `oos` option makes them available both in and out of sample.

The predicted hazard ratio, `hr`, is available only for the exponential, Weibull, and Gompertz models. `mean time` and `mean lntime` are not available for the Gompertz model.

`csnell` and `mgale` are not allowed with `svy` estimation results.

Options for predict

Main

`median time` calculates the predicted median survival time in analysis-time units. When no options are specified with `predict`, the predicted median survival time is calculated for all models.

`median lntime` calculates the natural logarithm of what `median time` produces.

`mean time` calculates the predicted mean survival time in analysis-time units. This option is not available for Gompertz regression.

`mean lntime` predicts the mean of the natural logarithm of `time`. This option is not available for Gompertz regression.

`hazard` calculates the predicted hazard for both the lower endpoint t_l and the upper endpoint t_u of the time interval.

`hr` calculates the hazard ratio. This option is valid only for models having a proportional-hazards parameterization.

`xb` calculates the linear prediction from the fitted model. That is, you fit the model by estimating a set of parameters, $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$, and the linear prediction is $\widehat{y}_j = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1j} + \widehat{\beta}_2 x_{2j} + \cdots + \widehat{\beta}_k x_{kj}$, often written in matrix notation as $\widehat{y}_j = \mathbf{x}_j \widehat{\boldsymbol{\beta}}$.

The $x_{1j}, x_{2j}, \ldots, x_{kj}$ used in the calculation are obtained from the data currently in memory and need not correspond to the data on the independent variables used in estimating $\boldsymbol{\beta}$.

`stdp` calculates the standard error of the linear prediction, that is, the standard error of $\widehat{y}_j$.

`surv` calculates each observation's predicted survivor probabilities for both the lower endpoint t_l and the upper endpoint t_u of the time interval.

`csnell` calculates the Cox–Snell-like residuals for both the lower endpoint t_l and the upper endpoint t_u of the time interval.

`mgale` calculates interval-censored martingale-like residuals, which are an interval-censored version of martingale residuals for right-censored data.

`nooffset` is relevant only if you specified `offset(`*varname*`)` with `stintreg`. It modifies the calculations made by `predict` so that they ignore the offset variable; the linear prediction is treated as $\mathbf{x}\boldsymbol{\beta}$ rather than $\mathbf{x}\boldsymbol{\beta}$ + offset.

`oos` makes `csnell` and `mgale` available both in and out of sample. `oos` also dictates that summations and other accumulations take place over the sample as defined by `if` and `in`. By default, the summations are taken over the estimation sample, with `if` and `in` merely determining which values of *newvar*, *newvar*$_l$, and *newvar*$_u$ are to be filled in once the calculation is finished.

`scores` calculates equation-level score variables. The number of score variables created depends upon the chosen distribution.

The first new variable will always contain $\partial \ln L / \partial(\mathbf{x}_j \boldsymbol{\beta})$.

The subsequent new variables will contain the partial derivative of the log likelihood with respect to the ancillary parameters.

margins

Description for margins

`margins` estimates margins of response for median and mean survival times, hazard ratios, and linear predictions.

Menu for margins

Statistics > Postestimation

Syntax for margins

`margins` [*marginlist*] [, *options*]

`margins` [*marginlist*] , `predict(`*statistic* ...`)` [`predict(`*statistic* ...`)` ...] [*options*]

statistic	Description
`median time`	median survival time; the default
`median lntime`	median ln(survival time)
`mean time`	mean survival time
`mean lntime`	mean ln(survival time)
`hr`	hazard ratio, also known as the relative hazard
`xb`	linear prediction $\mathbf{x}_j\widehat{\boldsymbol{\beta}}$
`stdp`	not allowed with `margins`
`hazard`	not allowed with `margins`
`surv`	not allowed with `margins`
`csnell`	not allowed with `margins`
`mgale`	not allowed with `margins`

Hazard estimation is not allowed because it produces interval estimates.

Statistics not allowed with `margins` are functions of stochastic quantities other than `e(b)`.

For the full syntax, see [R] **margins**.

Remarks and examples

Remarks are presented under the following headings:

Predicted values
Residuals and diagnostic measures

Predicted values

`predict` after `stintreg` is used to generate a new variable or variables containing predicted values or residuals.

Regardless of the metric used, `predict` can generate predicted median survival times and median log survival-times for all models and predicted mean times and mean log survival-times where available. Predicted survival, hazard, and residuals are also available for all models. The predicted hazard ratio can be calculated only for models with a proportional-hazards parameterization, that is, the Weibull, exponential, and Gompertz models. However, the estimation need not take place in the log-hazard metric. You can perform, for example, a Weibull regression specifying the `time` option and then ask that hazard ratios be predicted.

▷ Example 1: Obtaining predictions

Continuing with example 1 of [ST] **stintreg**, we refit a proportional-hazards Weibull model for the effect of treatment on breast retraction for breast cancer patients:

```
. use https://www.stata-press.com/data/r19/cosmesis
(Cosmetic deterioration of breast cancer patients)

. stintreg i.treat, interval(ltime rtime) distribution(weibull)
  (output omitted)
```

We can predict, for example, the median survival time and the log-median survival time for each observation by specifying the `median time` and `median lntime` options, respectively.

```
. predict time, median time

. predict lntime, median lntime

. tabulate treat, summarize(time) means freq

             | Summary of Predicted
             |      median for
             |     (ltime,rtime]
   Treatment |        Mean        Freq.
-------------+-------------------------
       Radio |   39.332397          46
   Radio+Che |   22.300791          48
-------------+-------------------------
       Total |   30.635407          94
```

```
. tabulate treat, summarize(lntime) means freq

              Summary of Predicted
                 median log for
                  (ltime,rtime]
  Treatment           Mean        Freq.

      Radio     3.6720486          46
  Radio+Che     3.1046221          48

      Total     3.3822989          94
```

From the `tabulate` command, the expected mean of the predicted median survival time for patients with radiotherapy only is approximately 39 months, and the expected mean of the predicted median survival time for patients with both radiotherapy and chemotherapy is 22 months. We can also obtain the same results by using `margins`.

```
. margins treat, predict(median time)

Adjusted predictions                                    Number of obs = 94
Model VCE: OIM

Expression: Predicted median for (ltime,rtime], predict(median time)

                          Delta-method
                Margin    std. err.      z    P>|z|     [95% conf. interval]

       treat
       Radio   39.3324    5.342493     7.36   0.000      28.8613    49.80349
 Radio+Chemo  22.30079    2.436642     9.15   0.000     17.52506    27.07652

. margins treat, predict(median lntime)

Adjusted predictions                                    Number of obs = 94
Model VCE: OIM

Expression: Predicted median log for (ltime,rtime], predict(median lntime)

                          Delta-method
                Margin    std. err.      z    P>|z|     [95% conf. interval]

       treat
       Radio  3.672049    .1358293    27.03   0.000     3.405828    3.938269
 Radio+Chemo  3.104622    .1092626    28.41   0.000     2.890471    3.318773
```

Because the `median` option is the default, we could have omitted it in the above specifications of `predict` and `margins`.

◁

▷ Example 2: Obtaining survivor probabilities

Continuing with the example above, we can compute observation-specific survivor probabilities. As with `predict` after [ST] **streg**, we will use `predict`'s `surv` option. For interval-censored data, however, estimates of survivor probabilities, as well as hazard estimates and Cox–Snell-like residuals, are intervals. So, to compute these statistics, we must specify two new variable names with `predict` instead of one; one variable will contain statistics computed using the lower time endpoint, and the other will contain statistics computed using the upper time endpoint.

```
. predict surv_l surv_u, surv
(38 missing values generated)
. list ltime rtime treat surv_l surv_u in 1/10
```

	ltime	rtime	treat	surv_l	surv_u
1.	0	7	Radio	1	.95814
2.	0	8	Radio	1	.948338
3.	0	5	Radio	1	.9754614
4.	4	11	Radio	.9828176	.9151379
5.	5	12	Radio	.9754614	.9029849
6.	5	11	Radio	.9754614	.9151379
7.	6	10	Radio	.967206	.9267811
8.	7	16	Radio	.95814	.8501493
9.	7	14	Radio	.95814	.8773297
10.	11	15	Radio	.9151379	.8639108

Listed above are the survivor probabilities, `surv_l` and `surv_u`, evaluated at the lower and upper time endpoints `ltime` and `rtime`, for the first 10 subjects, all of whom happen to be in the radiotherapy-only group.

◁

Residuals and diagnostic measures

For uncensored or right-censored data, several types of residuals have been introduced to assess the appropriateness of the fitted parametric survival models; see *Remarks and examples* in [ST] **streg postestimation** for details. Farrington (2000) proposed extensions of those residuals, including Cox–Snell-like residuals and martingale-like residuals, to interval-censored data; see the reference for applications and a discussion of limitations of the residuals for interval-censored data.

Cox–Snell-like residuals are used with interval-censored event-time data in assessing the overall model fit. If the model fits the data, those residuals should have the standard exponential distribution. To use them for checking the goodness of fit, we can estimate the cumulative hazard function corresponding to these residuals and plot them against the values at which the hazard is evaluated. If the model fits the data, the plot should be a straight line with a slope of 1 through the origin.

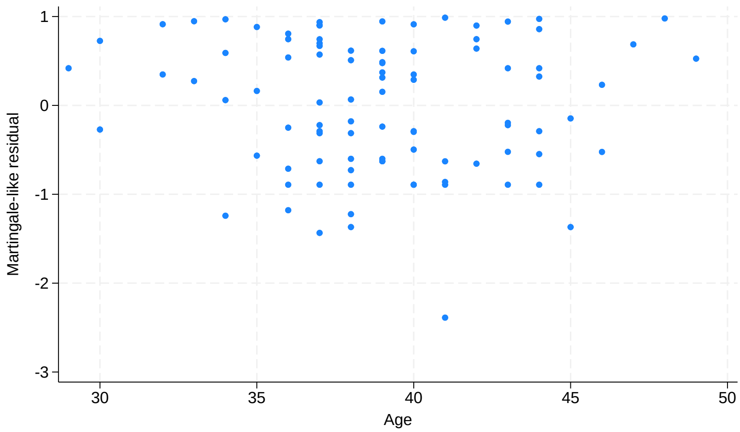
As with right-censored data, martingale-like residuals for interval-censored data do not arise naturally from martingale theory for parametric survival models as they do for the Cox proportional hazards model. For right-censored data, martingale residuals are defined using Cox–Snell residuals. For interval-censored data, Cox–Snell-like residuals are intervals themselves. So Farrington (2000) proposed a single measure, called adjusted Cox–Snell residuals, which are expectations of the interval residuals under the standard exponential distribution. Then, following Lagakos's (1981) definition of martingale residuals for right-censored data, an interval-censored version of martingale residuals is defined as one minus the adjusted Cox–Snell residuals. These martingale-like residuals are commonly used to examine the functional form of covariates. You could also use them to assess whether some covariates are needed in the model. Or you could plot them against observation numbers to identify outliers.

▷ Example 3: Check whether additional covariates should be included in the model

Martingale-like residuals may be used as a diagnostic tool to assess the need of including some other covariates in the model. If the model fits well without the covariate of interest, the plot of martingale residuals against that covariate should not show any trend.

Continuing with example 1, suppose that we want to check whether the patient's age (age) should be included in our model. We can specify the mgale option with predict to obtain the martingale-like residuals from the current model and store them in the mg variable. We then produce a scatterplot of mg against age.

```
. predict mg, mgale
. scatter mg age
```



The figure does not show any systematic trend, suggesting that age is not needed in the model. In fact, if we included age in our Weibull model in the first place, we would have found that age is not statistically significant. You can verify this by typing

```
. stintreg i.treat age, interval(ltime rtime) distribution(weibull)
```
(*output omitted*)

We can produce scatterplots of mg against other variables of interest to identify potential omitted predictors.

◁

▷ Example 4: Assess overall model fit

Returning to example 1, suppose that we instead want to fit the model with an exponential distribution and visually assess the overall model fit. We type

```
. quietly stintreg i.treat, interval(ltime rtime) distribution(exponential)
. estat gofplot
```

estat gofplot plots the estimated cumulative hazards for Cox–Snell-like residuals against the residuals themselves. The estimated cumulative hazards are calculated using the algorithm proposed by Turnbull (1976). The Cox–Snell-like residuals plotted against themselves form the 45° reference line. If the model fits the data well, the estimated cumulative hazards plotted against the Cox–Snell-like residuals should be close to the reference line. Comparing the jagged line with the reference line in figure 1, we observe that the estimated cumulative hazards deviate from the reference line. So the exponential model does not appear to fit these data well.

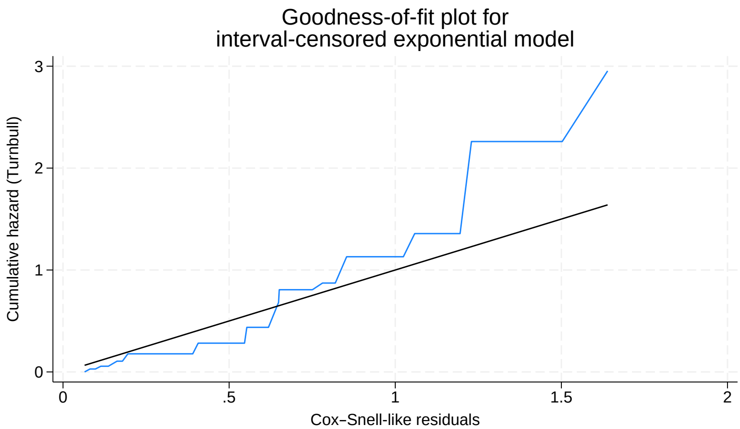


Figure 1. Goodness-of-fit plot for the exponential model

Let's refit this model using our original Weibull distribution and obtain the goodness-of-fit plot.

```
. quietly stintreg i.treat, interval(ltime rtime) distribution(weibull)
. estat gofplot
```

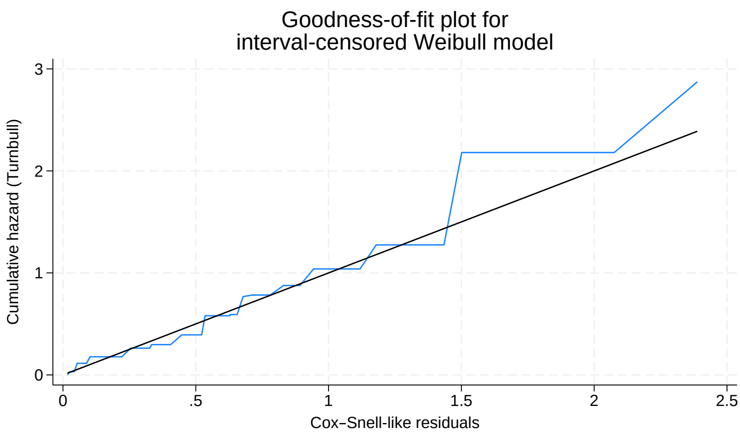


Figure 2. Goodness-of-fit plot for the Weibull model

The goodness-of-fit plot above shows that the jagged line stays very close to the 45° reference line. Therefore, we conclude that the Weibull model fits the data better than the exponential model.

◁

Methods and formulas

`predict` *newvar*, *statistic* may be used after `stintreg` to predict various quantities, according to the following *statistic*:

`median time`:

$$newvar_j = \{t : \widehat{S}_j(t) = 1/2\}$$

where $\widehat{S}_j(t)$ is $S_j(t)$ for observation j with the parameter estimates "plugged in" and $S_j(t)$ is defined in table 1 of [ST] **stintreg**.

`median lntime`:

$$newvar_j = \left\{y : \widehat{S}_j(e^y) = 1/2\right\}$$

`mean time`:

$$newvar_j = \int_0^\infty \widehat{S}_j(t)dt$$

`mean lntime`:

$$newvar_j = \int_{-\infty}^\infty ye^y \widehat{f}_j(e^y)dy$$

where $\widehat{f}_j(t)$ is $f_j(t)$ with the parameter estimates plugged in and $f_j(t) = -(d/dt)S_j(t)$.

`hr` (proportional hazards models only):

$$newvar_j = \exp(\mathbf{x}_j^* \widehat{\boldsymbol{\beta}^*})$$

where $\widehat{\boldsymbol{\beta}^*}$ does not contain the constant and $\mathbf{x}_j^*$ does not contain the coefficient of 1 corresponding to the constant.

`xb`:

$$newvar_j = \mathbf{x}_j\widehat{\boldsymbol{\beta}}$$

`stdp`:

$$newvar_j = \widehat{\mathrm{se}}(\mathbf{x}_j\widehat{\boldsymbol{\beta}})$$

`mgale`:

$$newvar_j = \frac{\widehat{S}_j(t_{lj}) \log \widehat{S}_j(t_{lj}) - \widehat{S}_j(t_{uj}) \log \widehat{S}_j(t_{uj})}{\widehat{S}_j(t_{lj}) - \widehat{S}_j(t_{uj})}$$

For right-censored data, martingale residuals can be defined as the scores of the regression parameters. This property can carry over to the interval-censored data to define martingale-like residuals. Therefore, these residuals are expected to have mean zero and are uncorrelated asymptotically. Furthermore, these residuals are orthogonal to variables included in the model. Thus, we can use it to assess the need of including some other covariates in the model. See Farrington (2000) for details.

These residuals take values between $-\infty$ and 1 and have an expected value of 0, although like the Cox–Snell-like residuals, they are not symmetric about 0, making them difficult to interpret.

`predict` *newvar*$_l$ *newvar*$_u$, *statistic2* may be used after `stintreg` to predict a pair of quantities for each observation for both the lower and upper endpoints of the time interval (t_{lj}, t_{uj}), according to the following *statistic2*:

`hazard`:

$$newvar_{lj} = \widehat{f}_j(t_{lj})/\widehat{S}_j(t_{lj})$$

$$newvar_{uj} = \widehat{f}_j(t_{uj})/\widehat{S}_j(t_{uj})$$

`surv`:

$$newvar_{lj} = \widehat{S}_j(t_{lj})$$

$$newvar_{uj} = \widehat{S}_j(t_{uj})$$

`csnell`:

$$newvar_{lj} = -\log \widehat{S}_j(t_{lj})$$

$$newvar_{uj} = -\log \widehat{S}_j(t_{uj})$$

The Cox–Snell-like residuals are the estimates of the cumulative hazard function obtained from the fitted model. They are computed separately for each of the two interval endpoints. Under the correct model assumption, the Cox–Snell-like residuals are expected to approximate an interval-censored sample from the standard exponential distribution. Therefore, they can be used for checking the overall model fit. Cox–Snell-like residuals can never be negative and therefore are not symmetric about zero. See Farrington (2000) for details.

References

Farrington, C. P. 2000. Residuals for proportional hazards models with interval-censored survival data. *Biometrics* 56: 473–482. https://doi.org/10.1111/j.0006-341X.2000.00473.x.

Lagakos, S. W. 1981. The graphical evaluation of explanatory variables in proportional hazard regression models. *Biometrika* 68: 93–98. https://doi.org/10.2307/2335809.

Turnbull, B. W. 1976. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society*, B ser., 38: 290–295. https://doi.org/10.1111/j.2517-6161.1976.tb01597.x.

Also see

[ST] **stintreg** — Parametric models for interval-censored survival-time data

[ST] **estat gofplot** — Goodness-of-fit plots after streg, stcox, stintreg, stintcox, or stmgintcox

[ST] **stcurve** — Plot the survivor or related function after streg, stcox, and more

[U] 20 Estimation and postestimation commands

For suggested citations, see the FAQ on citing Stata documentation.