Title

stata.com

estat gofplot — Goodness-of-fit plots after streg, stcox, stintreg, or stintcox

Description

`estat gofplot` plots the estimated cumulative hazard function for the Cox–Snell residuals versus the residuals themselves to assess the goodness of fit of the model visually after `streg`, `stcox`, `stintreg`, or `stintcox`.

Quick start

Plot the default cumulative hazard function for the Cox–Snell residuals versus the residuals themselves

```
estat gofplot
```

Plot the minus log of the Kaplan–Meier estimator for the Cox–Snell residuals versus the residuals themselves

```
estat gofplot, km
```

Plot separate cumulative hazard functions for each group of `x` on the same graph

```
estat gofplot, by(x)
```

Plot separate cumulative hazard functions for each group of `x` on different graphs

```
estat gofplot, by(x) separate
```

For the stratified model, plot separate cumulative hazard functions for each stratum

```
estat gofplot, stratify
```

Menu for estat

Statistics > Postestimation

Syntax

```
estat gofplot [ , options ]
```

options	Description
`na`	calculate the cumulative hazard function of the Cox–Snell residuals using the Nelson–Aalen estimator; the default for and available only after `streg` and `stcox`
`km`	calculate the cumulative hazard function of the Cox–Snell residuals using the minus log of the Kaplan–Meier estimator; available only after `streg` and `stcox`
`turnbull`	calculate the cumulative hazard function of the Cox–Snell-like residuals using the Turnbull estimator; the estimator for `stintreg` and `stintcox`
`by(`*varlist*`)`	estimate and graph separate functions for each group formed by *varlist*
`stratify`	estimate and graph separate functions for each stratum defined by `strata()` in estimation
`separate`	show curves on separate graphs; default is to show curves one on top of another
`outfile(`*filename* `[ , replace ])`	save values used to plot the goodness-of-fit graph
Plot	
connect_options	affect rendition of plotted cumulative hazard function
Reference line	
`rlopts(`*cline_options*`)`	affect rendition of the reference line
Add plots	
`addplot(`*plot*`)`	add other plots to the generated graph
Y axis, X axis, Titles, Legend, Overall	
twoway_options	any options other than `by()` documented in [G-3] ***twoway_options***

`estat gofplot` is not appropriate with `svy` estimation results and is not supported after estimation with `stcox, tvc()`.

Options

`na` specifies that the cumulative hazard function of the Cox–Snell residuals be calculated using the Nelson–Aalen estimator, which is the default after `streg` and `stcox`. `na` is not available after `stintreg` and `stintcox`.

`km` specifies that the cumulative hazard function of the Cox–Snell residuals be calculated using the minus log of the Kaplan–Meier estimator instead of the default Nelson–Aalen estimator after `streg` and `stcox`. `km` is not available after `stintreg` and `stintcox`.

`turnbull` specifies that the cumulative hazard function of the Cox–Snell-like residuals be calculated using the Turnbull estimator. `turnbull` is the only estimator after `stintreg` and `stintcox`, and it is not available after `streg` and `stcox`.

`by(`*varlist*`)` estimates a separate function for each by-group and plots all the functions on one graph. By-groups are identified by equal values of the variables in *varlist*. Up to five variables are allowed. `by()` may not be combined with `stratify`.

`stratify` requires that a stratified model has been previously fit using the `strata()` option; it estimates a separate function for each stratum and plots all the functions on one graph. `stratify` may not be combined with `by()`.

`separate` is meaningful only with `by()` or `stratify`; it requests that each group be placed on its own graph rather than one on top of the other.

`outfile(`*filename*`[ , replace ])` saves in *filename*`.dta` the values used to plot the goodness-of-fit graph.

Plot

connect_options affect the rendition of the plotted cumulative hazard function; see [G-3] ***connect_options***.

Reference line

`rlopts(`*cline_options*`)` affects the rendition of the reference line; see [G-3] ***cline_options***.

Add plots

`addplot(`*plot*`)` provides a way to add other plots to the generated graph; see [G-3] ***addplot_option***.

Y axis, X axis, Titles, Legend, Overall

twoway_options are any of the options documented in [G-3] ***twoway_options***, excluding `by()`. These include options for titling the graph (see [G-3] ***title_options***) and for saving the graph to disk (see [G-3] ***saving_option***).

Remarks and examples

stata.com

To assess the overall model fit, we can use the Cox–Snell residuals. If the survival regression model fits the data, these residuals should have a censored standard exponential distribution for right-censored data. Therefore, when we consider these residuals as failure (or censoring) times together with the original censoring variable, the hazard function should be constant and equal to 1, and the cumulative hazard should be a straight line with slope 1. `estat gofplot` allows us to verify the model's fit visually by calculating an empirical estimate of such a cumulative hazard function, which is based on either the Nelson–Aalen estimator or the Kaplan–Meier estimator for `streg` or `stcox` or is based on the Turnbull estimator for `stintreg` or `stintcox`. If the model fits the data, a plot of the cumulative hazard versus the residuals themselves should approximate a straight line with slope 1. See example 2 in [ST] **streg postestimation**, example 4 in [ST] **stcox postestimation**, example 4 in [ST] **stintreg postestimation**, and example 2 in [ST] **stintcox postestimation** for more detailed discussions. For interval-censored data, the Cox–Snell-like residuals are defined, and, under the correct model assumption, they are expected to approximate an interval-censored sample from the standard exponential distribution; see [ST] **stintcox postestimation**.

Example 1: Goodness-of-fit plots for stratified model

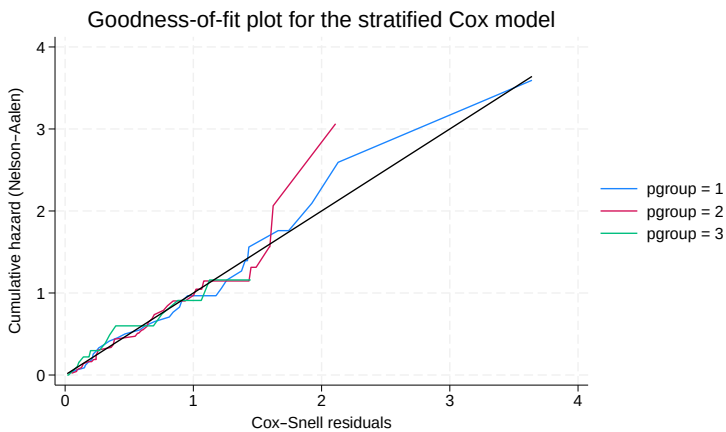
Returning to the Stanford heart experiment data from example 8 in [ST] **stcox**, we refit our model stratified by year of diagnosis group.

```
. use https://www.stata-press.com/data/r18/stan3
(Heart transplant data)
. generate pgroup = year
. recode pgroup min/69=1 70/72=2 73/max=3
(172 changes made to pgroup)
. stcox age posttran surg year, strata(pgroup) nolog
        Failure _d: died
  Analysis time _t: t1
       ID variable: id
Stratified Cox regression with Breslow method for ties
Strata variable: pgroup
No. of subjects =      103                     Number of obs =    172
No. of failures =       75
Time at risk    = 31,938.1
                                               LR chi2(4)    =  20.67
Log likelihood = -213.35033                    Prob > chi2   = 0.0004
```

_t	Haz. ratio	Std. err.	z	P>|z|	[95% conf.	interval]
age	1.027406	.0150188	1.85	0.064	.9983874	1.057268
posttran	1.075476	.3354669	0.23	0.816	.583567	1.982034
surgery	.2222415	.1218386	-2.74	0.006	.0758882	.6508429
year	.5523966	.1132688	-2.89	0.004	.3695832	.825638

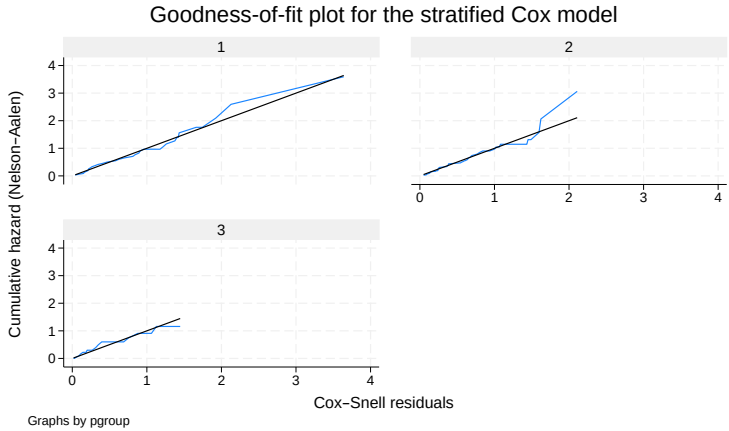
We can visually explore the goodness of fit for the stratified model by using `estat gofplot`. Here we will use the `stratify` option to plot the residuals separately for each stratum.

```
. estat gofplot, stratify
```



Even with perfect (artificially simulated) data, we expect to see departures from the diagonal at the right end of the curve, where values are based on only a few observations and greater fluctuation is observed. The above plot indicates the model fits well in all strata. To aid visual inspection of the plot, we can also add the `separate` option to produce separate graphs for each stratum.

```
. estat gofplot, stratify separate
```



Methods and formulas

The Cox and Snell (1968) residual for the jth observation at time t_j is defined as the estimated cumulative hazard function, $\widehat{H}_j(t_j) = -\log\widehat{S}_j(t_j)$, from the fitted model (Klein and Moeschberger 2003). Cox and Snell argued that if the correct model has been fit to the data, these residuals are n observations from a censored standard exponential distribution for right-censored data. Thus, a plot of the cumulative hazard rate of the residuals against the residuals themselves should result in a straight line of slope 1. Cox–Snell residuals can never be negative and therefore are not symmetric about 0. In practice, we can calculate an empirical estimate of the cumulative hazard rate of the residuals.

The default method of calculating the cumulative hazard rate of the residuals after `streg` and `stcox` is to use the Nelson–Aalen estimator (Nelson 1972; Aalen 1978). Alternatively, we may use the minus log of the Kaplan and Meier (1958) estimator by specifying the `km` option. For multiple-record data, the overall Cox–Snell residual is used, and hence, the cumulative hazard function is evaluated at the subject level defined by `id()` in the `stset` command.

For interval-censored data, Cox–Snell-like residuals are intervals themselves for single-record-per-subject dataset. Farrington (2000) proposed to calculate the Cox–Snell-like residuals for both lower and upper endpoints of the time intervals, then to use those predicted Cox–Snell-like residual intervals as the new time intervals and to compute the cumulative hazard function using the Turnbull estimator (Turnbull 1976). For multiple-record-per-subject data fit by `stintcox`, the overall Cox–Snell-like residual is used, and the cumulative hazard function is evaluated at the subject level defined by the `id()` option of `stintcox`.

References

Aalen, O. O. 1978. Nonparametric inference for a family of counting processes. *Annals of Statistics* 6: 701–726. https://doi.org/10.1214/aos/1176344247.

Cox, D. R., and E. J. Snell. 1968. A general definition of residuals (with discussion). *Journal of the Royal Statistical Society*, Series B 30: 248–275. https://doi.org/10.1111/j.2517-6161.1968.tb00724.x.

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.

Kaplan, E. L., and P. Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53: 457–481. https://doi.org/10.2307/2281868.

Klein, J. P., and M. L. Moeschberger. 2003. *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd ed. New York: Springer.

Nelson, W. 1972. Theory and applications of hazard plotting for censored failure data. *Technometrics* 14: 945–966. https://doi.org/10.2307/1267144.

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

Also see

[ST] **stcox postestimation** — Postestimation tools for stcox

[ST] **stintcox postestimation** — Postestimation tools for stintcox

[ST] **stintreg postestimation** — Postestimation tools for stintreg

[ST] **streg postestimation** — Postestimation tools for streg

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