
Assessing functional form in Cox regression

the `stfform` command

Daniele Spinelli¹ and Rino Bellocco¹

¹Dept. of Statistics and Quantitative Methods - University of Milano-Bicocca

Setting

Time-to-event (survival) data with a single spell, possibly with right censoring. For subject i with censoring time C_i and failure time T_i , let

- The follow-up time (`_t` in Stata)

$$Z_i = \min(T_i, C_i)$$

- The binary failure indicator (`_d` in Stata)

$$\delta_i = 1(T_i \leq C_i)$$

- Baseline covariates $\mathbf{X}_i$

The Cox Model

The hazard at time t is:

$$\lambda(t, \mathbf{X}) = \lambda_0(t) \exp(\beta \mathbf{X})$$

Cox model martingale residuals

Under the Cox model, the martingale residual process is

$$\widehat{M}_i(t) = N_i(t) - \int_0^t 1(Z_i \geq u) \exp(\widehat{\beta}^\top \mathbf{X}_i) d\widehat{\Lambda}_0(u).$$

where $N_i(t)$ is the observed number of events for individual i at time t

Interpretation

It is the observed minus the expected number of events up to time t under the fitted model.

Cumulative martingale residuals

For covariate X_j , order the residual contribution by the observed value of X_{ij} :

$$W_j(x) = \sum_{i=1}^n 1(X_{ij} \leq x) \widehat{M}_i, \quad \widehat{M}_i = \widehat{M}_i(\infty).$$

Correct specification

- ▶ $W_j(x)$ fluctuates around zero
- ▶ the observed path is compatible with reference paths

Potential misspecification

- ▶ the observed path is unusually extreme
- ▶ its shape may suggest a different transformation

The same construction can assess the exponential link by replacing X_{ij} with $\exp(\widehat{\beta}^\top \mathbf{X}_i)$.

Why another diagnostic?

Current Stata approaches

- ▶ LOWESS smoothing of martingale residuals vs. the covariates
- ▶ Grouping a covariate into quartiles and refitting the model (see `lincheck`)
- ▶ Visual assessment of the resulting pattern

Main limitation

The conclusion can depend on the smoothing method, its parameters, or the selected cut points.

`stfform` combines a cumulative-residual plot with a simulation-based p -value.

Lin, Wei, and Ying (1993); Hosmer, Lemeshow, and May (2013); Stata Survival Analysis Reference Manual.

Plots and supremum test

Graphical assessment

- For covariate j plot the observed process $w_j(x)$ against x
- Overlay simulated realizations from the the Gaussian reference process $\widehat{W}_j(x)$
- Check whether the observed path is unusually extreme

Numerical assessment

$$s_j = \sup_x |w_j(x)|$$

$$p \approx \Pr(\widehat{S}_j \geq s_j), \quad \widehat{S}_j = \sup_x |\widehat{W}_j(x)|.$$

Independent standard-normal draws generate the reference processes; the default is 1,000 simulations.

Use the two outputs together

The p -value indicates how unusual the observed process is; the graph shows the nature of the departure.

Basic syntax and prerequisites

Command

```
stfform [varlist, options]
```

- ▶ Requires Stata 16 or later.
- ▶ Run after `stset` and `stcox`.
- ▶ Not intended for multiple-spell data.
- ▶ Without `varlist`, all numerical covariates from the last `stcox` call are tested.
- ▶ With `varlist`, only variables included in the previous Cox model are retained.

Options

Option	Purpose
<code>nsimulation(#)</code>	Number of simulated processes; default: 1,000.
<code>nplot(#)</code>	Number shown in each graph; default: 20.
<code>nograph</code>	Suppress all graphs.
<code>noxb</code>	Do not assess the exponential link.
<code>novars</code>	Do not assess covariate functional forms.
<code>nolog</code>	Suppress the execution progress bar.

Options

Option	Purpose
<code>saving(filename[, replace])</code>	Save a long-format .dta file containing var, X, M, and simulated processes W*.
<code>graphnames(stub)</code>	Set a prefix for graph names.
<code>obscolor(string)</code>	Set the line color of the observed process.
<code>simcolor(string)</code>	Set the line color of simulated processes.

If all simulations cannot be saved within the current `maxvar` setting, the command saves a subset and issues a warning.

Workflow

1. Declare the survival-time data with `stset`.
2. Fit the Cox model with `stcox`.
3. Run the diagnostic

The command returns `r(test)`, a matrix containing the test statistics and associated p -values.

Practice - Setup and initial Cox model

```
ssc install stiform  
webuse brcancer  
describe
```

Contains data from <https://www.stata-press.com/data/r19/brcancer.dta>
 Observations: 686 German breast cancer data
 Variables: 14 26 Aug 2021 06:54

Variable name	Storage type	Display format	Value label	Variable label
id	int	%8.0g		ID
hormon	byte	%8.0g		Hormonal therapy
x1	byte	%8.0g		Age (years)
x2	byte	%8.0g		Menopausal status
x3	int	%8.0g		Tumour size, mm
x4	byte	%8.0g		Tumour grade
x5	byte	%8.0g		Number of positive nodes
x6	int	%8.0g		Progesterone receptor, fmol
x7	int	%8.0g		Estrogen receptor, fmol
rectime	int	%8.0g		Recurrence free survival time
censrec	byte	%8.0g		Censoring indicator
x4a	byte	%8.0g		Tumour grade $\geq$ 2
x4b	byte	%8.0g		Tumour grade $=$ 3
x5e	float	%9.0g		$\exp(-0.12 \cdot x5)$

Practice - Setup and initial Cox model

```
stset rectime, failure(censrec)
stcox x1 i.x2 x3 x4 x5 x6 x7
set seed 123456789
stfform
```

Cox regression with Breslow method for ties

No. of subjects = 686

Number of obs = 686

No. of failures = 299

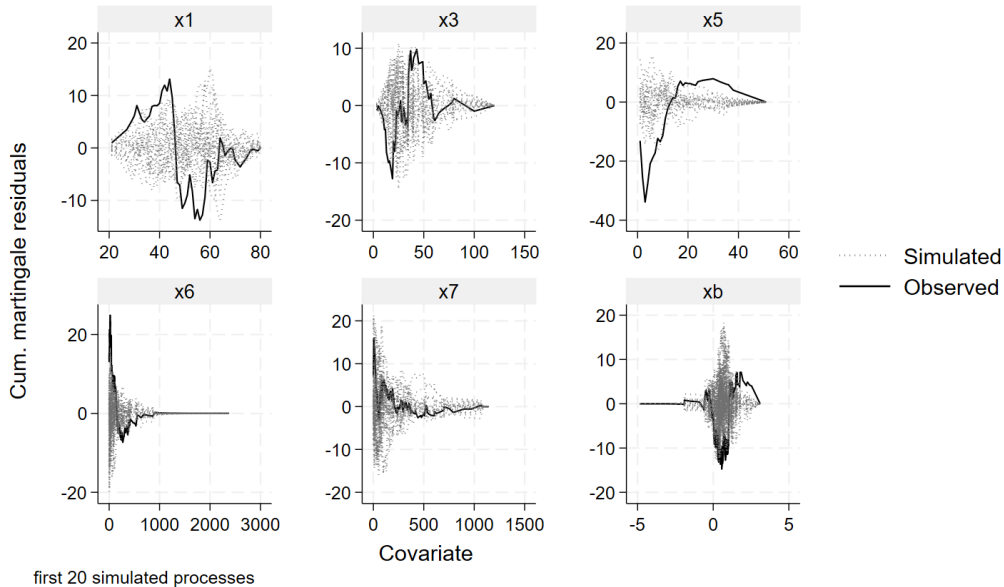
Time at risk = 771,400

LR chi2(7) = 94.77

Log likelihood = -1740.7903

Prob > chi2 = 0.0000

_t	Haz. ratio	Std. err.	z	P> z	[95% conf. interval]	
x1	0.989	0.009	-1.172	0.241	0.971	1.007
2.x2	1.251	0.230	1.216	0.224	0.872	1.793
x3	1.007	0.004	1.908	0.056	1.000	1.015
x4	1.356	0.144	2.875	0.004	1.102	1.668
x5	1.052	0.008	6.737	0.000	1.037	1.068
x6	0.998	0.001	-3.888	0.000	0.997	0.999
x7	1.000	0.000	0.237	0.813	0.999	1.001



Functional form test based on cumulative martingale residuals - 1000 replications

	s	$p(S \geq s)$
x1	13.761	0.072
x3	12.770	0.233
x5	33.869	0.001
x6	24.908	0.002
x7	16.002	0.131
xb	14.725	0.129

$P(S \geq s)$: p value under the null hypothesis that the functional form is correctly specified

xb: test for the link function

Considerations

The graphs suggest that

- ▶ Overestimate hazards when x_5 is low and underestimate hazards when x_5 is large.
- ▶ The highly negative residual for x_5 leads to the low p -value.
- ▶ Underestimate hazards when x_6 is low and overestimate hazards when x_5 is large.
- ▶ The hazards appear well estimated when x_b is low,
- ▶ For medium to large values of x_b the estimates become unstable: first overestimating hazards, then underestimating.
- ▶ Extreme values of x_b are outside the simulated process.

Conclusion

- ▶ The functional forms of x_5 and x_6 are inappropriate.
- ▶ log transformation for x_5 and square-root transformation for x_6

Practice - Alternative Cox model

```
gen lx5=log(x5)
gen x6sqrt=sqrt(x6)
stcox x1 i.x2 x3 x4 lx5 x6sqrt x7
stform
```

Cox regression with Breslow method for ties

No. of subjects = 686

Number of obs = 686

No. of failures = 299

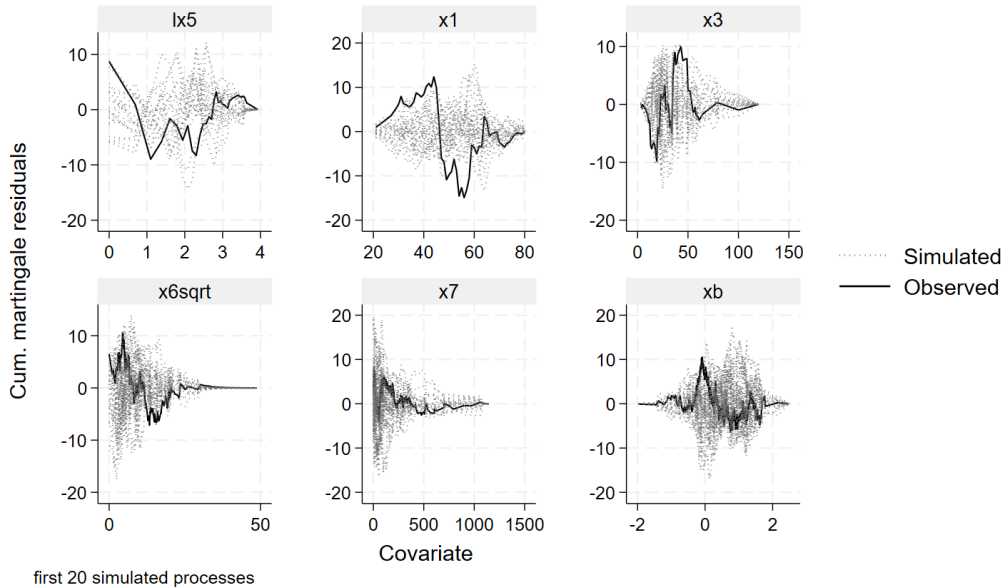
Time at risk = 771,400

LR chi2(7) = 123.75

Log likelihood = -1726.3002

Prob > chi2 = 0.0000

_t	Haz. ratio	Std. err.	z	P> z	[95% conf. interval]	
x1	0.990	0.009	-1.071	0.284	0.972	1.008
2.x2	1.157	0.212	0.796	0.426	0.808	1.658
x3	1.005	0.004	1.249	0.212	0.997	1.013
x4	1.237	0.134	1.964	0.050	1.000	1.529
1x5	1.627	0.108	7.332	0.000	1.428	1.853
x6sqrt	0.940	0.011	-5.289	0.000	0.919	0.962
x7	1.001	0.000	1.194	0.232	1.000	1.001



Functional form test based on cumulative martingale residuals - 1000 replications

	s	p($S \geq s$)
x1	14.919	0.039
x3	9.952	0.518
lx5	8.963	0.398
x6sqrt	10.550	0.461
x7	8.418	0.833
xb	10.551	0.491

$P(S \geq s)$: p value under the null hypothesis that the functional form is correctly specified

xb: test for the link function

Considerations

- ▶ more complicated parameter interpretation
- ▶ all p -values are substantially larger, except for the one associated with x_1 .
- ▶ the observed cumulative martingale residual processes for the log of x_5 ($\log x_5$) and the square-root of x_6 ($\sqrt{x_6}$) lie within the range of simulated values,
- ▶ this model represents an improvement over the previous one in terms of the overall adequacy of the exponential link function.

What the command adds

- ▶ A cumulative martingale-residual diagnostic for continuous covariates in Cox regression.
- ▶ Reference paths that avoid choosing a smoother or arbitrary quantile groups.
- ▶ A simulation-based supremum test that complements the graph.
- ▶ Separate assessment of covariate functional forms and the Cox exponential link.
- ▶ Reproducible saved data, named graphs, and returned test results.

Conclusions and future developments

- ▶ `stfform` is a new Stata command for assessing the functional form of the Cox model using cumulative martingale residuals.
- ▶ It combines graphical diagnostics with a simulation-based statistical test.
- ▶ Possible extensions include:
 - an omnibus test for overall model specification;
 - cumulative-residual tests for generalized linear models;
 - functional-form diagnostics for competing-risks models.