Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.

https://simfit.org.uk
https://simfit.silverfrost.com

Cox regression can be used with data sets containing strata with covariates where it is not convenient to use a nonparametric method or a fully defined statistical model. Rather it makes the somewhat restrictive assumption of proportional hazards.

From the main SIMFIT menu choose [Statistics], [Time series and survival], then [Cox regression], and study the default test file `cox.tf4` which has three covariates and three strata as shown next.

x_1	x_2	x_3	y	t	s
0	0	0	1	1.0072	2
1	1	0	0	0.0209	1
0	1	0	0	0.7954	1
1	1	1	0	0.0582	2
0	0	1	1	0.0611	2
1	0	0	0	0.1750	2
1	0	1	0	0.2593	2
0	1	0	0	0.9463	1
0	1	1	1	0.0898	3
1	0	1	0	0.0787	2
0	0	1	0	0.1378	3
0	0	0	1	0.6303	2
1	0	0	1	0.2115	1
1	1	1	1	0.1085	3
0	0	0	1	0.5227	2
0	0	1	0	0.0164	3
1	1	0	0	0.6804	1
0	1	0	0	1.1091	2
0	0	0	0	0.0154	1
1	1	1	0	0.0816	2
1	0	1	0	0.4498	3
0	0	0	1	0.0847	2
1	1	1	0	1.0198	1
1	1	1	0	0.0607	2
0	0	0	0	0.0968	2
1	0	1	1	0.2083	2
0	0	0	1	5.0050	1
0	0	1	0	0.0243	2
1	0	0	1	1.0054	3
1	0	0	1	0.1810	1
0	1	1	1	0.0512	3
0	1	0	1	0.2579	2
1	0	0	1	0.5309	2
0	0	1	0	0.2753	3
0	1	1	0	0.1252	3
1	0	1	0	0.0664	3

Continued on next page

x_1	x_2	x_3	y	t	s
0	1	0	1	0.6108	2
0	0	1	0	0.0187	1
0	0	1	0	0.1026	1
1	0	1	0	0.1016	2
0	0	1	0	0.4314	1
1	0	1	0	0.7174	1
0	0	1	1	0.2297	1
1	1	0	1	0.0346	2
0	1	1	1	0.0478	2
0	1	0	0	0.1261	2
0	1	0	1	0.0827	3
1	0	0	0	0.8903	1
0	0	1	0	0.2746	1
1	0	0	1	0.9722	1

Note that Cox regression data files for m covariates must be formatted as follows $x_1, x_2, ..., x_m, y, t, s$. Here, for example, with test file `cox.tf4` we have the following format.

- column 1, 2, and 3: covariates x_1, x_2, and x_3
- column 4: $y = 0$ (failure) or $y = 1$ (right censored)
- column 5: time t
- column 6: strata s

The results table is shown next.

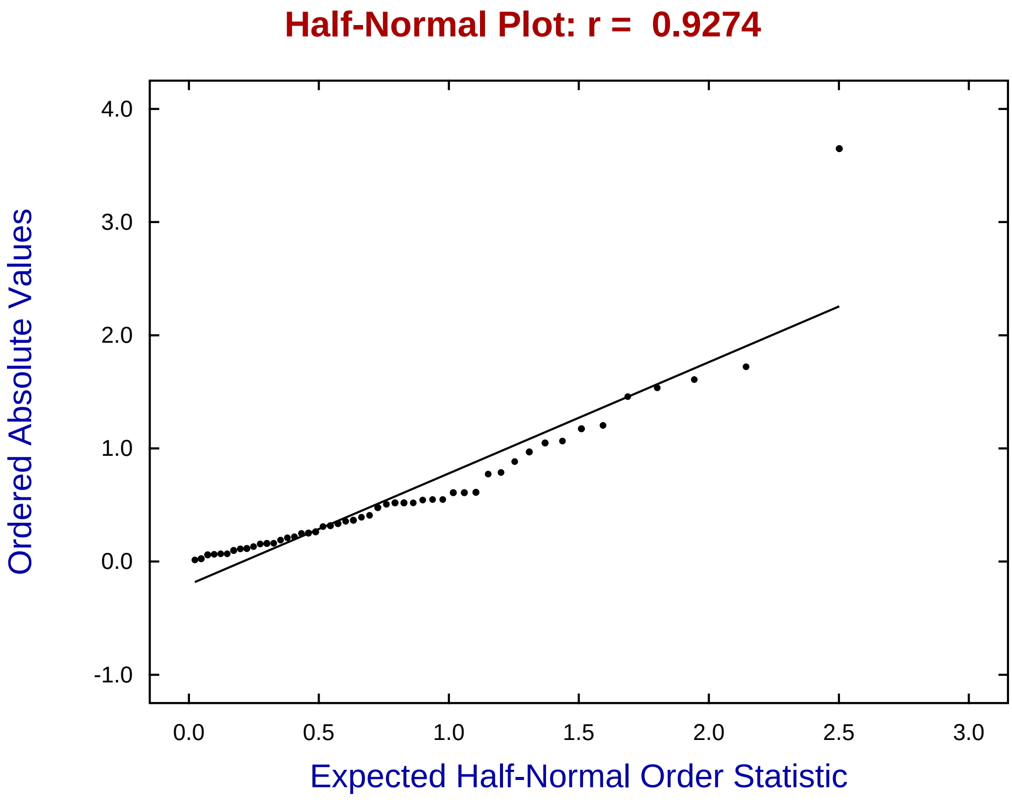
Deviance = 109.25, Number of time points = 50

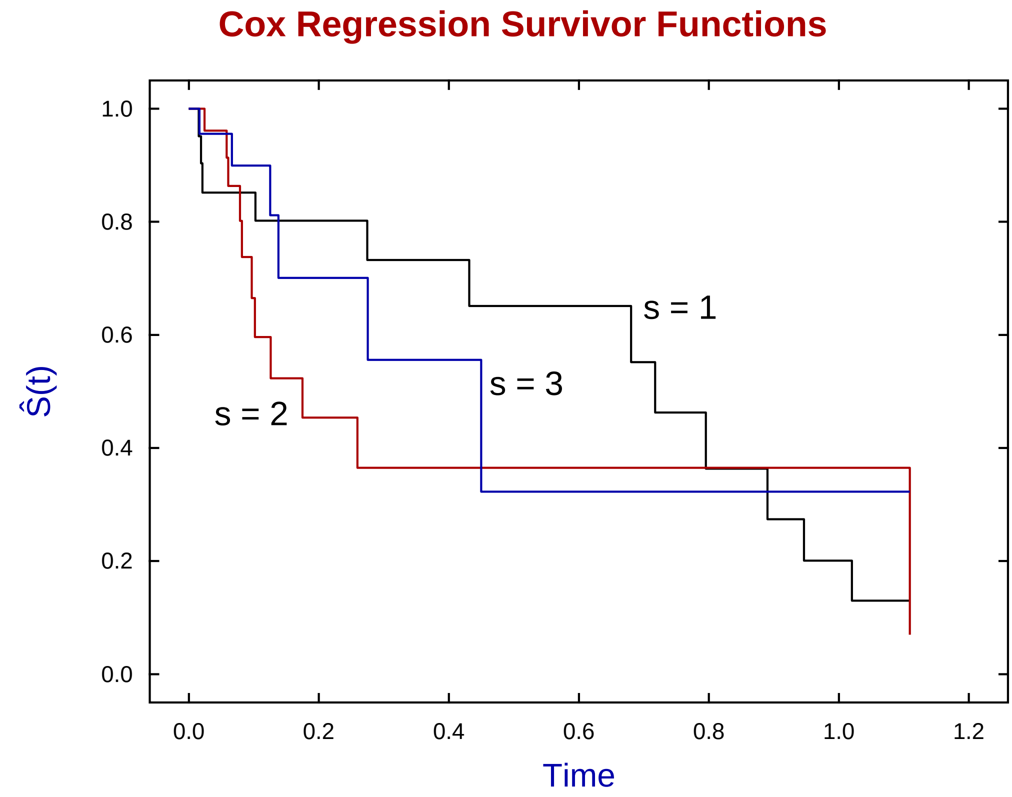
$B(i)$	Estimate	Score	Lower95%cl	Upper95%cl	Std.error	p	
1	-0.4893	8.156E-05	-1.423	0.445	0.464	0.2973	***
2	0.1609	-2.865E-05	-0.724	1.046	0.440	0.7162	***
3	1.5749	2.992E-04	0.562	2.588	0.504	0.0030	

Before proceeding to demonstrate further features of the SimFIT Cox regression procedure it is necessary to caution about the fact that analyzing the same data using different software packages may give differing results. This is inevitable with all nonlinear iterative methods and is because of several factors.

1. The Cox regression procedure does not completely specify a unique statistical model.
2. The solution point found is not unique but will depend on the method used and the starting values used.
3. The results obtained will depend on the technique used to deal with ties.
4. It is fairly common to find that some of the parameters are not well defined, i.e. not statistically different from zero as shown by stars in the last column.
5. The scores are derived from the partial derivatives estimated at the solution point so any values much less than about 10^{-6} can be regarded as effectively undefined due to rounding errors.

Probably the easiest way to check the goodness of fit is to inspect the half-normal residuals plot, while to compare strata the collected survivor function estimates should be viewed. These two plots for the current analysis are shown next.


Half-Normal Plot: r = 0.9274
Ordered Absolute Values
4.0
3.0
2.0
1.0
0.0
-1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Expected Half-Normal Order Statistic


Cox Regression Survivor Functions
$\hat{S}(t)$
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Time
s = 1
s = 3
s = 2

Theory

It should be pointed out that parameter estimates using the comprehensive Cox procedure may be slightly different from parameter estimates obtained by the GLM procedure if there are ties in the data, as the Breslow approximation for ties may sometimes be used by the comprehensive procedure, unlike the Cox exact method which is employed by the GLM procedures.

Another advantage of the comprehensive procedure is that experienced users can input a vector of offsets, as the assumed model is actually

$$\lambda(t, x) = \lambda_0(t) \exp(\beta^T x + \omega)$$

for parameters β, covariates x and offset ω.

Then the maximum likelihood estimates for β are obtained by maximizing the Kalbfleisch and Prentice approximate marginal likelihood

$$L = \prod_{i=1}^{n_d} \frac{\exp(\beta^T s_i + \omega_i)}{[\sum_{l \in R(t_{(i)})} \exp(\beta^T x_l + \omega_l)]^{d_i}}$$

where, n_d is the number of distinct failure times, s_i is the sum of the covariates of individuals observed to fail at $t_{(i)}$, and $R(t_{(i)})$ is the set of individuals at risk just prior to $t_{(i)}$.

In the case of multiple strata, the likelihood function is taken to be the product of such expressions, one for each stratum. For example, with ν strata, the marginal likelihood will be

$$L = \prod_{k=1}^{\nu} L_k.$$

Once parameters have been estimated the survivor function $\exp(-\hat{H}(t_{(i)})$ and residuals $r(t_l)$ are then calculated using

$$\hat{H}(t_{(i)}) = \sum_{t_j \le t_i} \left(\frac{d_j}{\sum_{l \in R(t_{(j)})} \exp(\hat{\beta}^T x_l + \omega_l)} \right)$$
$$r(t_l) = \hat{H}(t_l) \exp(\hat{\beta}^T x_l + \omega_l),$$

where there are d_j failures at t_j.

Note that the deviance is minus twice the log of marginal likelihood and the significance of nested models with different parameters contributing can be assessed by chi-square tests, as an alternative to the two-tailed t test given in the results table. Also, stratum differences (i.e. differences between groups) can be examined using the log-ranks test.