Simfit

Tutorials and worked examples for simulation,
curve fitting, statistical analysis, and plotting.
https://simfit.org.uk
https://simfit.silverfrost.com

The general linear modeling technique (GLM) can be used to analyze survival data when there are covariates. It should be emphasized that GLM is a very powerful technique, but it must be used with great care as it requires more understanding from users than most analytical techniques. It defines an error type for the observations, and assumes that the distribution of mean values is described in a link function which is a linear combination of covariates. Further, additional model information in the form of data transformation, offsets, weights, and strata may be required. For this reason SIMFIT provides a simplified interface for fitting survival data which will now be described

From the main SIMFIT menu choose [Statistics], [Time series and survival], then [GLM], and study the default test file `cox.tf1` which has data from P. Feigel and M. Zelen Biometrics 21, 826-838 (1965) in the following format.

covariate x_1	covariate x_2	covariate x_3	observation y	time in weeks t	indicator s
0.8329	0	0	0	65.00	1
-0.2877	0	0	0	156.0	1
1.4586	0	0	0	100.0	1
0.9555	0	0	0	134.0	1
1.7918	0	0	0	16.00	1
2.3514	0	0	0	108.0	1
2.3026	0	0	0	121.0	1
2.8332	0	0	0	4.000	1
1.6864	0	0	0	39.00	1
1.9459	0	0	0	143.0	1
2.2407	0	0	0	56.00	1
3.4657	0	0	0	26.00	1
3.5553	0	0	0	22.00	1
4.6052	0	0	0	1.000	1
4.6052	0	0	0	1.000	1
3.9512	0	0	0	5.000	1
4.6052	0	0	0	65.00	1
1.4816	1	1.4816	0	56.00	1
1.0986	1	1.0986	0	65.00	1
1.3863	1	1.3863	0	17.00	1
0.4055	1	4.0547	0	7.000	1
2.1972	1	2.1972	0	16.00	1
1.6677	1	1.6677	0	22.00	1
2.3026	1	2.3026	0	3.000	1
2.9444	1	2.9444	0	4.000	1
3.2958	1	3.2958	0	2.000	1
3.3322	1	3.3322	0	3.000	1
3.4340	1	3.4340	0	8.000	1
3.2581	1	3.2581	0	4.000	1
3.0445	1	3.0445	0	3.000	1
4.3694	1	4.3694	0	30.00	1
4.6052	1	4.6052	0	4.000	1
4.6052	1	4.6052	0	43.00	1

The above data format, i.e. the meaning of these six columns of data for this example of GLM survival analysis with three covariates must be thoroughly understood as will be explained.

If there are m covariates the first m columns must be the covariates, then column $m + 1$ must be either 0 (failure) or 1 (right censoring), column $m + 2$ must be the nonnegative survival time, while column (m + 3) could be a default value of 1, or the weight for replicates or (in some case) the stratum indicator.

For these data the particular details are as follows.

- **Column 1:**
  covariate x_1 = log white blood cell count (in thousands)
- **Column 2:**
  covariate x_2 = AG-factor positive or negative (0 or 1)
- **Column 3:**
  covariate x_3 (in this special case $x_3 = x_1 x_2$ i.e. column 1 multiplied by column 2)
- **Column 4:**
  observation y (where $y = 0$ for failure, or $y = 1$ for censored)
- **Column 5:**
  t = survival time in weeks (t must be > 0)
- **Column 6:**
  $s = 1$ this should usually be 1. However, it could be interpreted as a weighting factor for replicates, except for the SimFIT advanced Cox regression procedure when it would be assumed to be the stratum indicator.

In order to fit survival data using generalized linear models (GLM) by maximum likelihood four components must be defined.

1. A random variable, say Y with mean $E(Y) = \mu$, and variance $V(Y)$
2. A set of covariates $x_1, x_2, \ldots, x_m$ recorded at the same time as Y
3. A link function $g(.)$ which is a function of μ
4. A linear predictor function of the covariates $\eta = \sum_{j=1}^{m} \beta_j x_i$

In addition it is supposed that the relationship between $E(Y)$ and η is

$$g(\mu) = \eta$$

and the fit is achieved by an iterative process.

As the GLM technique for fitting survival models is very complicated, requiring careful choices for the distribution of Y and the link function $g(.)$ as well as the calculation of offsets and use of data transformations, SimFIT supplies a simplified interface to handle the following four special cases.

- The exponential model
- The Weibull model
- The extreme distribution model
- The Cox model

The following table displays the results from analyzing the same test file `cox.tf1` using each of these models sequentially.

Model: exponential survival

No. parameters = 4, Rank = 4, No. points = 33, Deg. freedom = 29

Parameter	Value	Lower95%cl	Upper95%cl	Std.error	p	
$Constant$	-5.1498	-6.201	-4.098	0.5142	0.0000	
$B(1)$	0.4818	0.115	0.849	0.1795	0.0119	
$B(2)$	1.8705	0.374	3.367	0.7317	0.0161	
$B(3)$	-0.3278	-0.831	0.175	0.2460	0.1931	**

Deviance = 38.55, A = 1

Model: Weibull survival

No. parameters = 4, Rank = 4, No. points = 33, Deg. freedom = 29

Parameter	Value	Lower95%cl	Upper95%cl	Std.error	p	
$Constant$	-5.0405	-6.182	-3.899	0.5580	0.0000	
$B(1)$	0.4761	0.108	0.844	0.1800	0.0131	
$B(2)$	1.8413	0.338	3.344	0.7349	0.0181	
$B(3)$	-0.3244	-0.829	0.180	0.2465	0.1985	**
α	0.9777	0.889	1.066	0.0434	0.0000	

Deviance = 37.06

Deviance - $2n\log[alpha] = 38.55$

Model: Extreme value survival

No. parameters = 4, Rank = 4, No. points = 33, Deg. freedom = 29

Parameter	Value	Lower95%cl	Upper95%cl	Std.error	p
$Constant$	-5.2457	-6.502	-3.989	0.6143	0.0000
$B(1)$	0.9024	0.520	1.284	0.1868	0.0000
$B(2)$	3.8711	2.272	5.471	0.7821	0.0000
$B(3)$	-0.7195	-1.241	-0.198	0.2549	0.0085
α	0.0344	0.030	0.039	0.0020	0.0000

Deviance = 35.69

Deviance - $2n\log[alpha] = 258.1$

Model: Cox proportional hazards

Deviance = 131.48, Number of time points = 33

Parameter	Estimate	Score	Lower95%cl	Upper95%cl	Std.error	p	
$B(1)$	0.7325	5.138E-06	0.248	1.217	0.2371	0.0043	
$B(2)$	2.7557	1.886E-06	0.731	4.780	0.9913	0.0093	
$B(3)$	-0.5792	5.062E-06	-1.188	0.030	0.2981	0.0615	*

It is very difficult to check goodness of fit when using the simplified GLM procedure in a situation where, as in this case, the number of covariates is greater than zero, because only a limited number of techniques are available for checking the deviance residuals as the technique is not simply estimating the parameters of a theoretical equation for survival as a function of time. The most useful technique is probably to examine the half-normal residuals plot for apparent linearity. Another indication is the final deviance, and the pattern of convergence displayed during the iteration to find the minimum deviance. Again, the statistical significance of the parameter estimates should be taken into account. The p values reported in the above table refer to a, approximate two-tailed t test on the ratio of parameter estimate to the corresponding standard error in order to test the null hypothesis

$$H_0 : \text{ The population parameter is not significantly different from zero.}$$

In other words, a p value less than 0.05 suggests that the parameter estimate could be meaningful, i.e. the corresponding parameter has been estimated reasonably well and it seems to be significantly different from zero. However, when p values exceed 0.05 this is indicated by stars as in the above table, drawing attention to the fact that the 95% confidence region for that parameter includes zero.

Theory

Many survival models can be fitted to N_u uncensored together with N_r right censored survival times with associated explanatory variables using the GLM technique from SIMFIT programs **linfit**, **gcfit** in mode 4, or **simstat**.

For instance, the SIMFIT simplified GLM interface allows you to read in data for the covariates, x, the variable y which can be either 1 for right-censoring or 0 for failure, together with the times t in order to fit survival models. With a density $f(t)$, survivor function $S(t) = 1 - F(t)$ and hazard function $h(t) = f(t)/S(t)$ a proportional hazards model is assumed for $t \geq 0$ with

$$\begin{aligned}
h(t_i) &= \lambda(t_i)\exp(\sum_j \beta_j x_{ij}) \\
&= \lambda(t_i)\exp(\beta^T x_i) \\
\Lambda(t) &= \int_0^t \lambda(u)\,du \\
f(t) &= \lambda(t)\exp(\beta^T x - \Lambda(t)\exp(\beta^T x)) \\
S(t) &= \exp(-\Lambda(t)\exp(\beta^T x)).
\end{aligned}$$

The SIMFIT comprehensive GLM procedure allows almost any model to be fitted to survival data, but it requires that users must understand the numerous choices that have to be made concerning distributions to be assumed, starting estimates to provide, link functions required, offsets that have to be provided, etc.

For these reasons the SIMFIT simplified GLM interface can fit several survival models using the appropriate choices for error distribution, link function, offset, data transformation, etc. required, as long as data are provided in the format demonstrated for the SIMFIT test file `cox.tf1`.

The exponential survival model

The exponential model has constant hazard and is particularly easy to fit, since

$$\begin{aligned}
\eta &= \beta^T x \\
f(t) &= \exp(\eta - t\exp(\eta)) \\
F(t) &= 1 - \exp(-t\exp(\eta)) \\
\lambda(t) &= 1 \\
\Lambda(t) &= t \\
h(t) &= \exp(\eta) \\
\text{and } E(t) &= \exp(-\eta),
\end{aligned}$$

so this simply involves fitting a GLM model with Poisson error type, a log link, and a calculated offset of $\log(t)$.

The selection of a Poisson error type, the log link and the calculation of offsets are all done automatically by the simplified interface from the data provided, as will be appreciated on fitting the test file `cox.tf1.` It should be emphasized that the values for y in the simplified GLM procedure for survival analysis must be either $y = 0$ for failure or $y = 1$ for right censoring, and the actual time for failure t must be supplied paired with the y values.

Internally, the SIMFIT simplified GLM interface reverses the y values to define the Poisson variables and uses the t values to calculate offsets automatically. Users who wish to use the advanced GLM interface for survival analysis must be careful to declare the Poisson variables correctly and provide the appropriate offsets as offset vectors.

The Weibull survival model

Weibull survival is similarly easy to fit, but is much more versatile than the exponential model on account of the extra shape parameter α as in the following equations.

$$
\begin{aligned}
f(t) &= \alpha t^{\alpha-1} \exp(\eta - t^{\alpha} \exp(\eta))\\
F(t) &= 1 - \exp(-t \exp(\eta))\\
\lambda(t) &= \alpha t^{\alpha-1}\\
\Lambda(t) &= t^{\alpha}\\
h(t) &= \alpha t^{\alpha-1} \exp(\eta)\\
E(t) &= \Gamma(1 + 1/\alpha) \exp(-\eta/\alpha).
\end{aligned}
$$

However, this time, the offset is $\alpha \log(t)$, where α has to be estimated iteratively and the covariance matrix subsequently adjusted to allow for the extra parameter α that has been estimated. The iteration to estimate α and covariance matrix adjustments are done automatically by the SimFIT simplified GLM interface, and the deviance is also adjusted by a term $-2n \log \hat{\alpha}$.

The extreme value survival model

Extreme value survival is defined by

$$f(t) = \alpha \exp(\alpha t) \exp(\eta - \exp(\alpha t + \eta))$$

which is easily fitted, as it is transformed by $u = \exp(t)$ into Weibull form, and so can be fitted as a Weibull model using t instead of $\log(t)$ as offset. However it is not so useful as a model since the hazard increases exponentially and the density is skewed to the left.

The Cox proportional hazards model

This model assumes an arbitrary baseline hazard function $\lambda_0(t)$ so that the hazard function is

$$h(t) = \lambda_0(t) \exp(\eta).$$

It should first be noted that Cox regression techniques may often yield slightly different parameter estimates, as these will often depend on the starting estimates, and also since there are alternative procedures for allowing for ties in the data. In order to allow for Cox's exact treatment of ties in the data, i.e., more than one failure or censoring at each time point, this model is fitted by the SimFIT GLM techniques after first calculating the risk sets at failure times t_i, that is, the sets of subjects that fail or are censored at time t_i plus those who survive beyond time t_i. Then the model is fitted using the technique for conditional logistic analysis of stratified data. The model does not involve calculating an explicit constant as that is subsumed into the arbitrary baseline function. However, the model can accommodate strata in two ways. With just a few strata, dummy indicator variables can be defined as in test files `cox.tf2` and `cox.tf3` but, with large numbers of strata, data should be prepared as for `cox.tf4.`

As an example, consider the results shown in the previous table from fitting an exponential, Weibull, then Cox model to data in the test file `cox.tf1`. In this case there is little improvement from fitting a Weibull model after an exponential model, as shown by the deviances and half normal residuals plots. The deviances from the full models (exponential, Weibull, extreme value) can be compared for goodness of fit, but they can not be compared directly to the Cox deviance.