Survival analysis

– or reliability analysis.

Two characteristics:

- Survival distributions are skewed distributions on the positive half line. It is the shape rather than the location of the distribution that reflects interesting differences between populations.
- There is almost always a censoring mechanism, and certain aspects of the data are consequently missing. We need to deal with this in the modeling.

Example I

In engineering we want to estimate the life time of an electrical component. We record whenever a component is put to work and whenever it fails. At a given time, all working components that have not yet failed are censored.

To estimate the life time based on the observed life times for the components that have failed up to this time will give a too pessimistic, biased result.

Example II

A "real" survival application.

Patients are enrolled in a study whenever they are diagnosed with a given (serious, life threatening) disease. Data on the subjects are collected.

At a planned calendar time the statistical analysis is done, and patients alive at this time are censored.

Many questions are of interest, e.g. how different covariates or treatments are associated with the survival after diagnosis for this particular disease.

The setup

We consider n individuals, $T_1^*, \ldots, T_n^*$ independent, positive random variables (survival times). We observe

$$T_i = \min\{T_i^*, C_i\}$$

with censoring times $C_1, \ldots, C_n$. With

$$e_i = 1(T_i^* \leq C_i)$$

we observe the pairs

$$(T_1, e_1), \ldots, (T_n, e_n).$$

The process of individuals at risk

$$Y(t) = \sum_{i=1}^{n} 1(t \leq T_i).$$

The Kaplan-Meier estimator

The distribution function is $F(t) = P(T_1^* \leq t)$ and the survival function is

$$S(t) = 1 - F(t) = P(T_1^* > t).$$

Based on the censored survival observations (T_i, e_i), the process, $Y(s)$, of at risk individuals, and the (ordered) observed survival times t_i for $i = 1, \ldots, k$ up to time t the Kaplan-Meier estimator is

$$\begin{aligned} \hat{S}(t) &= \left(1 - \frac{1}{Y(t_1)}\right)\left(1 - \frac{1}{Y(t_2)}\right)\ldots\left(1 - \frac{1}{Y(t_k)}\right) \\ &= \prod_{i:t_i \leq t}\left(1 - \frac{1}{Y(t_i)}\right). \end{aligned}$$

This estimator is the survival analysis version of the empirical distribution function.

The hazard rate

If F is continuously differentiable with derivative f (the density for the survival distribution), we introduce the hazard rate

$$\lambda(t) = \frac{f(t)}{S(t)}.$$

Observe that

$$\begin{aligned} \lambda(t) &= \lim_{\varepsilon \to 0+} \frac{1}{\varepsilon} \frac{F(t+\varepsilon) - F(t)}{S(t)} \\ &= \lim_{\varepsilon \to 0+} \frac{1}{\varepsilon} P(T_i^* \in (t, t+\varepsilon] \mid T_i^* > t). \end{aligned}$$

Thus $\lambda(t)$ is the instantaneous rate of death at time t.

The Weibull distribution has hazard rate

$$\lambda(t) = \alpha\gamma t^{\gamma-1}$$

for $\alpha, \gamma > 0$.

The cumulative hazard function

Note that

$$\lambda(t) = -(\log S(t))'$$

hence

$$\Lambda(t) := \int_0^t \lambda(s)\mathrm{d}s = -\log S(t),$$

which is called the cumulative hazard function.

Observe that

$$S(t) = \exp\left(-\Lambda(t)\right).$$

Setup

Assume that T^* is a positive random variable with density f and survival function S, C is a positive random variable with density g and survival function H.

We define

$$T = \min\{T^*, C\} \quad \text{and} \quad e = 1(T^* \leq C).$$

Theorem

If T^* and C are independent the joint distribution of (T, e) has density

$$f(t)^e S(t)^{1-e} g(t)^{1-e} H(t)^e$$

w.r.t. the product measure $m \otimes \tau$ (the Lebesgue measure times the counting measure).

The full likelihood

With $(T_1, e_1), \ldots, (T_n, e_n)$ i.i.d. with the same distribution as (T, e) the full likelihood is

$$L = \prod_{i=1}^{n} f(T_i)^{e_i} S(T_i)^{1-e_i} g(T_i)^{1-e_i} H(T_i)^{e_i}.$$

We assume that $f = f_\beta$ is parametrized by β and that the distribution, given by g, of the censoring mechanism holds no information about β. This implies that

$$L(\beta) = \prod_{i=1}^{n} f_\beta(T_i)^{e_i} S_\beta(T_i)^{1-e_i} K_i$$

with K_i depending on the observations but not the parameter β.

The likelihood

From hereon the likelihood to consider is

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} f_\beta(T_i)^{e_i} S_\beta(T_i)^{1-e_i} \\ &= \prod_{i=1}^{n} \lambda_\beta(T_i)^{e_i} S_\beta(T_i) \end{aligned}$$

recalling the definition of the hazard rate

$$\lambda_\beta(t) = \frac{f_\beta(t)}{S_\beta(t)}.$$

The log-likelihood is

$$\ell(\beta) = \sum_{i=1}^{n} e_i \log \lambda_\beta(T_i) - \sum_{i=1}^{n} \Lambda_\beta(T_i)$$

recalling that the cumulative hazard function is

$$\Lambda_\beta(t) = -\log S_\beta(t).$$

MLE for the censored exponential

If the survival distribution is the exponential with parameter λ being the rate, the MLE is

$$\hat{\lambda} = \frac{n_u}{\sum_{i=1}^n T_i}$$

with n_u the number of failures/deaths (the number of uncensored observations).

- If we ignore censoring the MLE is

$$\frac{n}{\sum_{i=1}^n T_i} > \hat{\lambda},$$

  which will overestimate the rate.

- If we discard censored observations the MLE is

$$\frac{n_u}{\sum_{i=1}^n e_i T_i} > \hat{\lambda},$$

  which will overestimate the rate.

Accelerated failure time models

Definition

An AFT model has survival function given as

$$S_\eta(t) = 1 - G((\log t - \eta)/\sigma)$$

with η the linear predictor, G a distribution function (on $\mathbb{R}$), and $\sigma > 0$ the scale parameter.

A unit change of X_j increases the failure time by a factor e^{β_j}.

The combined effect of η is a scale transformation by e^η of the baseline distribution with survival function

$$S_0(t) = 1 - G(\log t/\sigma).$$

Proportional hazards models

Definition

The proportional hazards model has hazard rate

$$\lambda(t) = \lambda_0(t)e^{\eta}$$

with η the linear predictor and λ_0 the baseline hazard rate.

It follows that for the cumulative hazard function

$$\Lambda(t) = \Lambda_0(t)e^{\eta}$$

the proportionality holds too.

The factor e^{β_j} is the hazard ratio between two models corresponding to a unit change of x_j.

Weibull example

The Weibull baseline hazard rate and cumulative hazard function are

$$\lambda_0(t) = \gamma t^{\gamma-1} \qquad \Lambda_0(t) = t^{\gamma}.$$

The log-likelihood is

$$\begin{aligned} \ell &= \sum_{i=1}^{n} e_i \log(\gamma T_i^{\gamma-1} e^{\eta_i}) - T_i^{\gamma} e^{\eta_i} \\ &= \underbrace{\sum_{i=1}^{n} e_i \log(T_i^{\gamma} e^{\eta_i}) - T_i^{\gamma} e^{\eta_i}}_{\text{Poisson log-likelihood}} + \sum_{i=1}^{n} e_i \log(\gamma T_i^{-1}). \end{aligned}$$

This is (surprisingly) up to a constant the log-likelihood for a Poisson model of the e_i's with log link and mean value $T_i^{\gamma} e^{\eta_i}$ for fixed γ.

The glm-framework can be used to fit the model (for fixed γ) with the survival times entering as an offset term.

Weibull example

The survival function for the proportional hazards model with a Weibull baseline is

$$\exp\left(-e^{\eta}t^{\gamma}\right) = \exp\left(-e^{\gamma(\log t + \eta/\gamma)}\right).$$

We recognize this as an AFT-model with scale parameter $\sigma = 1/\gamma$ and linear predictor

$$\eta' = -\eta/\gamma.$$

We can fit such models using the `survreg` function from the survival package in R. The parameters in the proportional hazards parametrization are obtained from the linear transformation

$$\hat{\beta}_j = -\hat{\beta}'_j/\sigma.$$