Survival Analysis: Parametric Models

Francisco Javier Rubio

@FJavierRubio1

Lectures I and II

1. Parametric Survival Analysis in the Overall Survival Framework.
2. Parametric Survival Analysis in the Relative Survival Framework.

Lecture Aims

1. To describe the aims of survival analysis (overall survival).
2. To describe parametric approaches to estimate the survival function.
3. To describe two regression models: PH and AFT.
4. To present software tools for survival analysis and a real data example.
5. To discuss other parametric, semiparametric and non parametric alternatives.

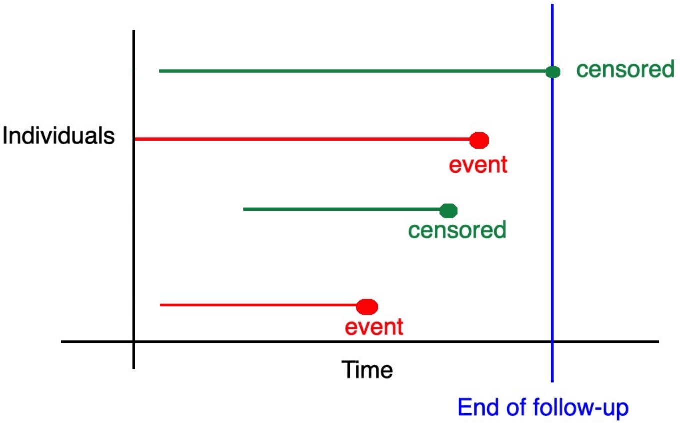
Survival analysis in practice

- In many areas such as medicine, biology, and engineering (reliability), scientists have access to the survival times of a group of individuals or items.
- Example 1. The survival of cancer patients after a diagnosis of cancer. (*)
- Example 2. The lifespan of a product/device. (x)
- These two scenarios are identical from a theoretical perspective. However, in practice, the former has ethical considerations.

The typical data set

- Sample of **times to event** (possibly right-censored) $(t_1, \ldots, t_n)$ from a group of individuals.
- Vital status (or **censoring** indicators) $(\delta_1, \ldots, \delta_n)$. ($\delta_i = 1$: death, $\delta_i = 0$, right-censored/alive). Censoring may be due to random drop-out, lost to follow-up, or administrative censoring.
- In some cases, we may know some additional characteristics about the individuals, meaning we have access to **covariates** $\mathbf{x}_i = (x_{i1}, \ldots, x_i p)^\top$, (age, sex, deprivation level, comorbidities, tumour stage, $\ldots$).
- Overall survival: in this framework, we know the times to event, but we do not know consider the cause of death.

Censoring


censored
Individuals
event
censored
event
Time
End of follow-up

Aims of Survival Analysis

- One of the aims of Survival Analysis is to quantify the survival of the group of individuals. The typical quantities of interests are:
  1. The **survival function** $S(t) = 1 - F(t) = \mathbb{P}(T > t)$, where T is a positive random variable representing the survival time with CDF F.
  2. The **hazard function** is defined as the instantaneous risk:

$$h(t) = \lim_{dt \to 0} \frac{P[t \leq T < t + dt \mid T \geq t]}{dt} \overset{homework}{=} \frac{f(t)}{S(t)}.$$

Survival and Hazard functions

- The hazard function and the survival function are linked through the relationship:

$$S(t) \overset{homework}{=} \exp\left\{-\int_0^t h(s)ds\right\}.$$

- The function $H(t) = \int_0^t h(s)ds$ is known as the **cumulative hazard** function.

The importance of the hazard function

- We are exposed to many forces of mortality: ageing, illnesses, natural disasters, accidents, crime, and etcetera.
- In Biostatistics (and epidemiology), the concept of the hazard plays a key role as it reflects these forces of mortality.

Maximum Likelihood Estimation: No covariates

- Suppose that $(t_1, \ldots, t_n)$ are independent and identically distributed.
- Suppose that we are interested in estimating the survival function using a parametric distribution $F(\cdot \mid \theta)$.
- Which distribution?

Brief catalogue of parametric distributions

- [Gamma].
- [Weibull].
- [Lognormal].
- [Loglogistic].
- [Generalised Gamma].
- Among many many others. [see]

Warning:

Different distributions can capture different shapes of the hazard function:

- Weibull: increasing, decreasing and bathtub (down then up).
- Lognormal: unimodal (up then down).
- Generalised gamma: increasing, decreasing, bathtub, and unimodal.

By selecting a parametric from the catalogue of distributions, we are making assumptions about the possible hazard rates of the true distribution. Selecting the best model using formal tools is usually recommended.

Maximum Likelihood Estimation: No covariates

- We need to take the difference in the contribution of the observed and right-censored times:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i \mid \boldsymbol{\theta})^{\delta_i} S(t_i \mid \boldsymbol{\theta})^{1-\delta_i}.$$

- The MLE $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is the value that maximises the likelihood function (numerical methods).
- The fitted survival and hazard functions are $S(t \mid \widehat{\boldsymbol{\theta}})$ and $h(t \mid \widehat{\boldsymbol{\theta}})$.

Software

- Since these are standard methods and distributions, many have already been implemented in R.
- Fitting a number of distributions to survival data using the [flexsurv] R package.

Regression models: using covariates

- Spoiler: there is not unique way of including covariates to model survival times.
- Biostatisticians and epidemiologists tend to think in terms of the hazard, based on the previous interpretation. Thus, unsurprisingly, most approaches to include covariates in the survival model consist of hazard-based regression models.
- The most popular models are: the proportional hazards (PH) model, and the accelerated failure time (AFT) model.

Proportional hazards models

- The PH model postulates that the covariates affect a “baseline hazard”, by either increasing it or decreasing it. This is,

$$h_{PH}(t \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta}) = h_0(t \mid \boldsymbol{\theta}) \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}\right\}.$$

- The corresponding cumulative hazard function is:

$$H_{PH}(t \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta}) \overset{\textit{homework}}{=} H_0(t \mid \boldsymbol{\theta}) \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}\right\}.$$

Proportional hazards: interpretation

- $h_0(t \mid \boldsymbol{\theta})$ is the hazard associated to $\mathbf{x}_i = \mathbf{0}$. This represents a sort of "reference point". This can be the hazard associated to the distributions discussed previously.
- If $\mathbf{x}_i^\top \boldsymbol{\beta} > 0$,
$$h_{PH}(t \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta}) > h_0(t \mid \boldsymbol{\theta}).$$
This means that the combination of characteristics of the individual, for a fixed value of $\boldsymbol{\beta}$, lead to an increase of the hazard compared to the baseline hazard.
- If $\mathbf{x}_i^\top \boldsymbol{\beta} < 0$,
$$h_{PH}(t \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta}) < h_0(t \mid \boldsymbol{\theta}).$$
This means that the combination of characteristics of the individual, for a fixed value of $\boldsymbol{\beta}$, lead to an decrease of the hazard compared to the baseline hazard.

Proportional hazards: hazard ratios

- Let $\mathbf{x}_i$ denote the characteristics of the *i*th patient.
- For a specific value of $k \in \{1, \ldots, p\}$, let $\tilde{\mathbf{x}}_i$ be defined as

$$\tilde{x}_{ij} = \begin{cases} x_{ij} & j \neq k, \\ x_{ij} + 1 & j = k. \end{cases}$$

- Then,

$$\frac{h_{PH}(t \mid \tilde{\mathbf{x}}_i, \boldsymbol{\theta}, \boldsymbol{\beta})}{h_{PH}(t \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta})} \overset{homework}{=} \exp\{\beta_k\}.$$

  This is, the exponential of β_k represents the increase in hazard due to an increase of one unit of the covariate x_{ik}.

Accelerated Failure Time model

- The AFT postulates that covariates affect simultaneously the time scale and the hazard scale:

$$h_{AFT}(t \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta}) = h_0\left(t \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}\right\} \mid \boldsymbol{\theta}\right) \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}\right\}.$$

- The corresponding cumulative hazard function is:

$$H_{AFT}(t \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta}) \overset{homework}{=} H_0\left(t \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}\right\} \mid \boldsymbol{\theta}\right).$$

AFT model: interpretation

- The interpretation of the AFT model is easier if we transform the model as follows. The survival function is:

$$
\begin{aligned}
S_{AFT}(t_i \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta}) &\overset{homework}{=} \exp\left\{-H_0\left(t_i \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}\right\} \mid \boldsymbol{\theta}\right)\right\} \\
&= S_0\left(t_i \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}\right\} \mid \boldsymbol{\theta}\right) \\
&= 1 - F_0\left(t_i \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}\right\} \mid \boldsymbol{\theta}\right) \\
&= \mathbb{P}\left(T_i > t_i \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}\right\} \mid \boldsymbol{\theta}\right) \\
&= \mathbb{P}\left(\log T_i > \log(t_i) + \mathbf{x}_i^\top \boldsymbol{\beta} \mid \boldsymbol{\theta}\right) \\
&= \mathbb{P}\left(\log T_i > \log(t_i) - \mathbf{x}_i^\top \boldsymbol{\alpha} \mid \boldsymbol{\theta}\right),
\end{aligned}
$$

where $\boldsymbol{\alpha} = -\boldsymbol{\beta}$.

AFT model: interpretation

- Let $Y_i = \log(T_i)$ and $Y_i \mid \mathbf{x}_i^\top \boldsymbol{\alpha} \sim G_0(\cdot \mid \boldsymbol{\theta})$. Let g_0 be the corresponding pdf.
- Then,

$$S_{AFT}(t_i \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta}) \leftrightarrow 1 - G_0\left(y_i - \mathbf{x}_i^\top \boldsymbol{\alpha} \mid \boldsymbol{\theta}\right).$$

- Consequently, the density function is

$$f_{AFT}(t_i \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta}) \leftrightarrow g_0\left(y_i - \mathbf{x}_i^\top \boldsymbol{\alpha} \mid \boldsymbol{\theta}\right).$$

- Finally, we can identify this pdf with the pdf associated to a log-linear model:

$$y_i = \log(t_i) = \mathbf{x}_i^\top \boldsymbol{\alpha} + \epsilon_i, \quad i = 1, \ldots, n,$$

  where $\epsilon_i \overset{iid}{\sim} G_0(\cdot \mid \boldsymbol{\theta})$.
- Consequently, the covariates have a direct effect on the log survival time.

Regression models: the likelihood function

- The likelihood function (for PH and AFT models) is:

$$
\begin{aligned}
L(\boldsymbol{\beta},\boldsymbol{\theta}) \quad &= \quad \prod_{i=1}^{n} f_j(t_i \mid \mathbf{x}_i,\boldsymbol{\theta},\boldsymbol{\beta})^{\delta_i} S_j(t_i \mid \mathbf{x}_i,\boldsymbol{\theta},\boldsymbol{\beta})^{1-\delta_i} \\
&\overset{homework}{=} \quad \prod_{i=1}^{n} h_j(t_i \mid \mathbf{x}_i,\boldsymbol{\theta},\boldsymbol{\beta})^{\delta_i} \exp\left\{-H_j(t_i \mid \mathbf{x}_i,\boldsymbol{\theta},\boldsymbol{\beta})\right\}, \quad (*)
\end{aligned}
$$

$j = PH, AFT$.

- This also shows that the likelihood can be characterised using the hazard function.

Example: lung cancer data

- Let t_i, $i = 1, \ldots, 228$ denote the survival times associated to patients with advanced lung cancer from the North Central Cancer Treatment Group.
- 165 patients died within the follow-up period (1022 days).

No covariates

- In the first scenario, we compare the survival functions associated to three parametric models: Weibull, Lognormal, and Generalised Gamma. R code.

Regression models

- We consider two covariates, “age” and “sex” . We first compare the PH and AFT models using a lognormal baseline hazard, and later a Weibull baseline hazard (where PH=AFT). R code.

Discussion

- Parametric survival models are useful tools in several applied areas.
- Their use is not automatic as the user needs to select the parametric model(s).
- PH and AFT models are the most popular hazard-based regression models.
- Confidence intervals for the parameters and confidence regions for the estimate of the survival function.

Extensions: Parametric approaches

- The General Hazard (GH) model represents an extension of the PH and AFT models.
- The GH model is discussed in Lecture II.
- The GH model is implemented in the R package [HazReg]
- Other model structures include the Proportional Odds model and related extensions.

Extensions: Semiparametric and Nonparametric approaches

- The Kaplan-Meier estimator is a NP estimator of the survival function.
- The Nelson-Aalen estimator is a NP estimator of the cumulative hazard function.
- The Cox model is a semiparametric version of the PH model.

https://rpubs.com/FJRubio/CPHM

The Cox PH Model and the Partial Likelihood function

- The PH model:

$$h_{PH}(t \mid \mathbf{x}_i, \boldsymbol{\beta}) = h_0(t) \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}\right\},$$

- In order to avoid misspecification of the baseline hazard (wrong model), it is often preferred to estimate it non-parametrically, while the coefficients $\boldsymbol{\beta}$ are estimated using the log partial likelihood function Cox [1972]:

$$\ell_p(\boldsymbol{\beta}) = \sum_{\delta_i=1} \mathbf{x}_i^\top \boldsymbol{\beta} - \sum_{\delta_i=1} \log\left(\sum_{k \in \mathcal{R}(t_i)} \exp\left\{\mathbf{x}_k^\top \boldsymbol{\beta}\right\}\right),$$

where t_i, $i = 1, \ldots, n$, are the survival times, $\mathcal{R}(t) = \{i : t_i \geq t\}$ denotes the risk set at time t.

D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B*, 34(2):187–220, 1972.

F.E. Harrell. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

J.D. Kalbfleisch and R.L. Prentice. *The Statistical Analysis of Failure Time Data*, volume 360. John Wiley & Sons, 2011.

J.P. Klein, H.C. Van-Houwelingen, J.G. Ibrahim, and T.H. Scheike. *Handbook of Survival Analysis*. CRC Press, 2016.