Hazard-based parametric regression models

Francisco Javier Rubio

University College London
Department of Statistical Science.

Lecture Aims

- To discuss the importance of the hazard function in the analysis of survival data.
- To introduce a general hazard regression model.
- To discuss a real data example and available software.

Survival analysis

- Survival analysis methods have been applied in a number of areas, including medicine, epidemiology, genetics, engineering, and biology, to name but a few.

Survival analysis

- Survival analysis methods have been applied in a number of areas, including medicine, epidemiology, genetics, engineering, and biology, to name but a few.
- The **survival function** and the **hazard function** represent two quantities of interest in this area.

Survival analysis

- Survival analysis methods have been applied in a number of areas, including medicine, epidemiology, genetics, engineering, and biology, to name but a few.
- The **survival function** and the **hazard function** represent two quantities of interest in this area.
- The survival function provides information about the probability that an individual or population will survive beyond a certain time point: $S(t) = P(T > t)$.

Hazard function

- We are exposed to many forces of mortality: aging, illnesses, natural disasters, accidents, crime, COVID, and etcetera.

Hazard function

- We are exposed to many forces of mortality: aging, illnesses, natural disasters, accidents, crime, COVID, and etcetera.
- Intuitively, the hazard function quantifies these forces of mortality at each time point $t > 0$.

Hazard function

- We are exposed to many forces of mortality: aging, illnesses, natural disasters, accidents, crime, COVID, and etcetera.
- Intuitively, the hazard function quantifies these forces of mortality at each time point $t > 0$.
- Mathematically [Rinne, 2014], this is interpreted in terms of the hazard function (homework)

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

Hazard function

- We are exposed to many forces of mortality: aging, illnesses, natural disasters, accidents, crime, COVID, and etcetera.
- Intuitively, the hazard function quantifies these forces of mortality at each time point $t > 0$.
- Mathematically [Rinne, 2014], this is interpreted in terms of the hazard function (homework)

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

where $S_T(t) = P(T > t)$, and $f_T(t)$ is the probability density function of T.

The typical data set

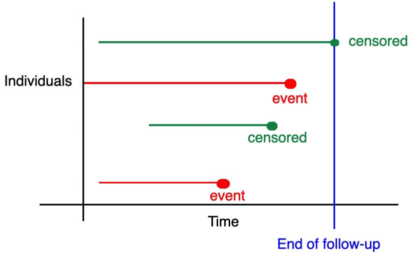
- Sample of **times to event** (possibly right-censored) $(t_1, \ldots, t_n)$ from a group of individuals.

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).

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.



The typical data set

- 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, $\ldots$).

The typical data set

- In some cases, we may know some additional characteristics about the individuals, meaning we have access to **covariates** $\mathbf{x}_i = (x_{i1}, \dots, x_ip)^\top$, (age, sex, ...).
- How do we incorporate information on covariates to the hazard function?

General/Extended hazard (GH) structure

- Etezadi-Amoli and Ciampi [1987] and Chen and Jewell [2001] proposed a very natural unifying hazard structure. The corresponding hazard and cumulative hazard functions are:

$$
\begin{aligned}
h(t; \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= h_0\left(t \exp\left\{\tilde{\mathbf{x}}^\top \boldsymbol{\alpha}\right\}\right) \exp\left\{\mathbf{x}^\top \boldsymbol{\beta}\right\}, \\
H(t; \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= H_0\left(t \exp\left\{\tilde{\mathbf{x}}^\top \boldsymbol{\alpha}\right\}\right) \exp\left\{\mathbf{x}^\top \boldsymbol{\beta} - \tilde{\mathbf{x}}^\top \boldsymbol{\alpha}\right\},
\end{aligned}
$$

$\tilde{\mathbf{x}} \subseteq \mathbf{x}$. It has been used in many applications [Rubio et al., 2019].

General/Extended hazard (GH) structure

- Etezadi-Amoli and Ciampi [1987] and Chen and Jewell [2001] proposed a very natural unifying hazard structure. The corresponding hazard and cumulative hazard functions are:

$$
\begin{aligned}
h(t; \mathbf{x}, \alpha, \beta) &= h_0\left(t \exp\left\{\tilde{\mathbf{x}}^\top \alpha\right\}\right) \exp\left\{\mathbf{x}^\top \beta\right\}, \\
H(t; \mathbf{x}, \alpha, \beta) &= H_0\left(t \exp\left\{\tilde{\mathbf{x}}^\top \alpha\right\}\right) \exp\left\{\mathbf{x}^\top \beta - \tilde{\mathbf{x}}^\top \alpha\right\},
\end{aligned}
$$

  $\tilde{\mathbf{x}} \subseteq \mathbf{x}$. It has been used in many applications [Rubio et al., 2019].
- The structure includes covariates that act at the time-level ($\tilde{\mathbf{x}}$) and covariates that have an effect at the hazard level ($\mathbf{x}$).

General/Extended hazard (GH) structure

- Etezadi-Amoli and Ciampi [1987] and Chen and Jewell [2001] proposed a very natural unifying hazard structure. The corresponding hazard and cumulative hazard functions are:

$$\begin{aligned} h(t;\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta}) &= h_0\left(t\exp\left\{\tilde{\mathbf{x}}^\top\boldsymbol{\alpha}\right\}\right)\exp\left\{\mathbf{x}^\top\boldsymbol{\beta}\right\}, \\ H(t;\mathbf{x},\boldsymbol{\alpha},\boldsymbol{\beta}) &= H_0\left(t\exp\left\{\tilde{\mathbf{x}}^\top\boldsymbol{\alpha}\right\}\right)\exp\left\{\mathbf{x}^\top\boldsymbol{\beta}-\tilde{\mathbf{x}}^\top\boldsymbol{\alpha}\right\}, \end{aligned}$$

  $\tilde{\mathbf{x}} \subseteq \mathbf{x}$. It has been used in many applications [Rubio et al., 2019].
- The structure includes covariates that act at the time-level ($\tilde{\mathbf{x}}$) and covariates that have an effect at the hazard level ($\mathbf{x}$).
- When $\boldsymbol{\alpha} = \mathbf{0}$, we recover the PH model. For $\boldsymbol{\beta} = \mathbf{0}$ we obtain the AH model. The AFT model is obtained for $\boldsymbol{\alpha} = \boldsymbol{\beta}$.

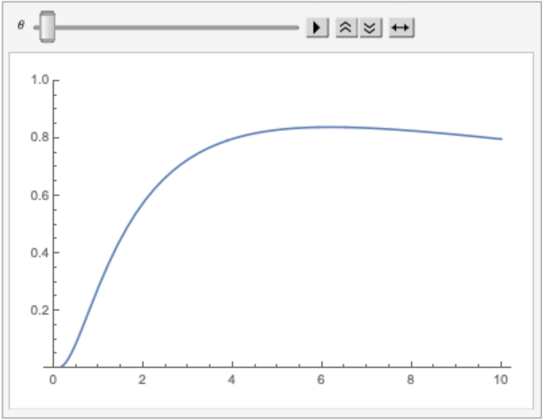
General/Extended hazard (GH) structure

- Etezadi-Amoli and Ciampi [1987] and Chen and Jewell [2001] proposed a very natural unifying hazard structure. The corresponding hazard and cumulative hazard functions are:

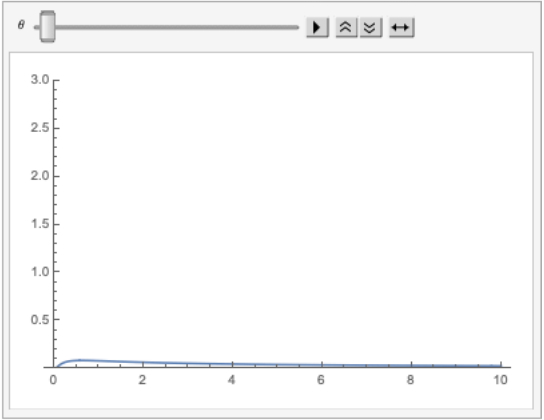
$$
\begin{aligned}
h(t; \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= h_0\left(t \exp\left\{\tilde{\mathbf{x}}^\top \boldsymbol{\alpha}\right\}\right) \exp\left\{\mathbf{x}^\top \boldsymbol{\beta}\right\}, \\
H(t; \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= H_0\left(t \exp\left\{\tilde{\mathbf{x}}^\top \boldsymbol{\alpha}\right\}\right) \exp\left\{\mathbf{x}^\top \boldsymbol{\beta} - \tilde{\mathbf{x}}^\top \boldsymbol{\alpha}\right\},
\end{aligned}
$$

  $\tilde{\mathbf{x}} \subseteq \mathbf{x}$. It has been used in many applications [Rubio et al., 2019].
- The structure includes covariates that act at the time-level ($\tilde{\mathbf{x}}$) and covariates that have an effect at the hazard level ($\mathbf{x}$).
- When $\boldsymbol{\alpha} = \mathbf{0}$, we recover the PH model. For $\boldsymbol{\beta} = \mathbf{0}$ we obtain the AH model. The AFT model is obtained for $\boldsymbol{\alpha} = \boldsymbol{\beta}$.
- The GH model is **identifiable** provided that the baseline hazard is not a member of the Weibull family of distributions (when PH = AFT = AH). [Chen and Jewell, 2001]

- The GH structure also includes time-level effects through α ($\tilde{\mathbf{x}}$).



- The GH structure also includes hazard-level effects through $\beta(\mathbf{x})$.



Parametric GH Models and Inference

- We will focus on the case where we model the baseline hazard function through a parametric distribution $h_0(\cdot;\theta)$.

Parametric GH Models and Inference

- We will focus on the case where we model the baseline hazard function through a parametric distribution $h_0(\cdot;\theta)$.
- Once we choose a parametric baseline hazard (e.g. LogNormal, LogLogistic, Gamma, Generalised Gamma, Power Generalised Weibull, ...), we can estimate the parameters using maximum likelihood inference. (`HazReg` R and Julia packages)

Brief catalogue of parametric distributions

- [Gamma].
- [Weibull].
- [Lognormal].
- [Loglogistic].
- [Generalised Gamma].
- Among many many others. [PGW], [EW].

Warning:

Different distributions can capture different shapes of the hazard function:

Warning:

Different distributions can capture different shapes of the hazard function:

- Weibull: increasing, decreasing, flat.

Warning:

Different distributions can capture different shapes of the hazard function:

- Weibull: increasing, decreasing, flat.
- Lognormal: unimodal (up then down).

Warning:

Different distributions can capture different shapes of the hazard function:

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

Warning:

Different distributions can capture different shapes of the hazard function:

- Weibull: increasing, decreasing, flat.
- Lognormal: unimodal (up then down).
- Generalised gamma, PGW, EW: 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 (AIC, BIC).

Real data example

- In this example, we analyse the LeukSurv data set from the R package spBayesSurv. This data set contains information about the survival of acute myeloid leukemia in 1,043 patients.

Real data example

- In this example, we analyse the LeukSurv data set from the R package spBayesSurv. This data set contains information about the survival of acute myeloid leukemia in 1,043 patients.
- We will fit several models (GH, PH, AFT, AH) with different baseline hazards, and select the best model using AIC and BIC.

Real data example

- In this example, we analyse the LeukSurv data set from the R package spBayesSurv. This data set contains information about the survival of acute myeloid leukemia in 1,043 patients.
- We will fit several models (GH, PH, AFT, AH) with different baseline hazards, and select the best model using AIC and BIC.
- We summarise the best selected model with the available tools in the `HazReg` R package.

[HazReg]

Y.Q. Chen and N.P. Jewell. On a general class of semiparametric hazards regression models. *Biometrika*, 88(3):687–702, 2001.

J. Etezadi-Amoli and A. Ciampi. Extended hazard regression for censored survival data with covariates: a spline approximation for the baseline hazard function. *Biometrics*, 43:181–192, 1987.

H. Rinne. *The Hazard rate: Theory and inference (with supplementary MATLAB-Programs)*. Justus-Leibig-University, Giessen, Germany, 2014.

F.J. Rubio, L. Remontet, N.P. Jewell, and A. Belot. On a general structure for hazard-based regression models: an application to population-based cancer research. *Statistical Methods in Medical Research*, 28:2404–2417, 2019.