

Statistics for Biology and Health


Chapter 1 Basic Quantities and Models

Qi Guo

July 18,2019



THE UNIVERSITY OF TEXAS AT DALLAS
School of Natural Sciences and Mathematics

- 
1. Introduction
 2. Basic Quantities
 3. Regression Models for Survival Data
 4. Models for Competing Risks

Introduction

- A common feature of survival data sets is they contain either censored or truncated observations.
- Censored data arises when an individual's life length is known to occur only in a certain period of time.
 1. Right censoring is known that the individual is still alive at a given time.
 2. Left censoring when all that is known is that the individual has experienced the event of interest prior to the start of the study.
 3. Interval censoring where the only information is that the event occurs within some intervals.

- Left truncation occurs when the subjects have been at risk before entering the study (for example: life insurance policy holders where the study starts on a fixed date, event of interest is age at death).
- Right truncation occurs when the entire study population has already experienced the event of interest (for example: a historical survey of patients on a cancer registry).
- Generally we deal with right censoring & sometimes left truncation.

Basic Quantities

- Let X be the time until some specified event, and it's a non-negative random variable from a homogeneous population.
 1. The survival function: The probability of an individual surviving to time x .
 2. The hazard rate function: The chance an individual of age x experiences the event in the next instant in time.
 3. The probability density function: The unconditional probability of the event's occurring at time x .
 4. The mean residual life: The mean time to the event of the interest, given the event has not occurred at x .
- If we know any one of these four functions, then the other three can be uniquely determined.

The survival function

- The probability of an individual surviving beyond time x (experiencing the event after time x)

$$S(x) = \Pr(X > x) = 1 - F(x) = \int_x^{\infty} f(t)dt$$

- For a discrete random variable X .

$$S(x) = \Pr(X > x) = \sum_{x_j > x} P(x_j)$$

- For the Weibull distribution, $S(x) = \exp(-\lambda x^\alpha)$, $\lambda > 0, \alpha > 0$, when $\alpha = 1$ is the exponential distribution.

The hazard function

- The hazard rate is defined by:

$$h(x) = \lim_{\Delta x \rightarrow 0} \frac{P(x \leq X < x + \Delta x | X \geq x)}{\Delta x}$$

- If X is a continuous random variable, $h(x) = \frac{f(x)}{S(x)} = -\frac{d \ln(S(x))}{dx}$

$$H(x) = \int_0^x h(\mu) d\mu = -\ln[S(x)]$$

$$S(x) = \exp[-H(x)] = \exp\left[-\int_0^x h(\mu) d\mu\right]$$

- For the Weibull distribution, $\alpha > 1$, $h(x)$ increasing, $\alpha < 1$, $h(x)$ decreasing.

The hazard function

- If X is a discrete random variable,

$$h(x_j) = \Pr(X = x_j | X \geq x_j) = \frac{P(x_j)}{S(x_{j-1})}, j = 1, 2, \dots$$

where $S(x_0) = 1$. $P(x_j) = S(x_{j-1}) - S(x_j)$ and
 $h(x_j) = 1 - S(x_j)/S(x_{j-1})$, $j = 1, 2, \dots$

- So the survival function is:

$$S(x) = \prod_{x_j \leq x} S(x_j)/S(x_{j-1}) = \prod_{x_j \leq x} [1 - h(x_j)]$$

The mean residual life function and median life

- The mean residual life function measures their expected remaining lifetime (area under survival curve to the right of x divided by $S(x)$)

$$mrl(x) = E(X - x | X > x)$$

$\mu = mrl(0)$ is the mean life, total area under the survival curve.

- For a continuous random variable,

$$mrl(x) = \frac{\int_x^\infty (t - x)f(t)dt}{S(x)} = \frac{\int_x^\infty S(t)dt}{S(x)}$$

$$\mu = E(x) = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt$$

- The median lifetime for a continuous random variable x is the $x_{0.5}$, so that $S(x_{0.5}) = 0.5$

Regression Models for Survival Data

- Consider a failure time $X > 0$, a vector $Z' = (Z_1, \dots, Z_p)$ of explanatory variables associated with failure time x , Z' may include quantitative variables (such as blood pressure, temperature, age, and weight), qualitative variables (such as gender, race, treatment, and disease status) and/or time-dependent variables, $Z'(x) = [Z_1(x), \dots, Z_p(x)]$.
- Two approaches to the modeling of covariate effects on survival.

- 1. The first approach is analogous to the classical linear regression approach, $Y = \ln(X)$. A linear model is assumed for Y , namely,

$$Y = \mu + \gamma'z + \delta w$$

where $\gamma' = (\gamma_1, \dots, \gamma_p)$ is a vector of regression coefficients and w is the error distribution.

If $W \sim N(0, 1)$, $Y \sim \log$ normal regression model.

If $W \sim \exp(w - e^w)$, $-\infty < w < \infty$, $Y \sim \text{Weibull}$ regression model.

This model is called the accelerated failure-time(AFT) model.

The accelerated failure time model

- Let $S_0(x)$ denote the survival function of $X = e^Y$ when Z is 0, that is $S_0(x)$ is the survival function of $\exp(\mu + \delta w)$, and

$$Pr(X > x|z) = S_0[x \exp(-\gamma'z)]$$

- So the hazard rate of an individual with a covariate value Z for the model is related to a baseline hazard rate h_0 by

$$h(x|z) = h_0[x \exp(-\gamma'z)] \exp(-\gamma'z)$$

- Although the accelerated failure-time model provides a direct extension of the classical linear model's construction for explanatory variables for conventional data, for survival data, its use is restricted by the error distributions one can assume.

Conditional hazard model

- 2. The major approach to modeling the effects of covariates on survival is to model the conditional hazard rate as a function of the covariates.
- Two general classes of models have been used to relate covariate effects to survival, the family of multiplicative hazard models and the family of additive hazard rate models.
- For the family of multiplicative hazard rate models the conditional hazard rate of an individual with covariate vector z is a product of a baseline hazard rate $h_0(x)$ and a non-negative function of the covariates, $c(\beta', z)$

$$h(x|z) = h_0(x)c(\beta'z)$$

$h_0(x)$ may have a specified parametric form or it may be left as an arbitrary non-negative function.

Multiplicative hazards models

- Most applications use the Cox model with $c(\beta'z) = \exp(\beta'z)$, which is chosen for its simplicity and for the fact it is positive for any value of $\beta'z$.
- A key feature of multiplicative hazards models is that, when all the covariates are fixed at time 0, the hazard rates of two individuals with distinct values of z are proportional.

$$\frac{h(x|z_1)}{h(x|z_2)} = \frac{h_0(x)c(\beta'z_1)}{h_0(x)c(\beta'z_2)} = \frac{c(\beta'z_1)}{c(\beta'z_2)}$$

which is a constant independent of time.

- So it can be expressed in terms of a baseline survival function $S_0(x)$ as

$$S(x|z) = S_0(x)^{c(\beta'z)}$$

- The conditional hazard function is:

$$h(x|z) = h_0(x) + \sum_{j=1}^p z_j(x)\beta_j(x)$$

- Estimation for additive models is typically made by non-parametric (weighted) least-squares methods.

Models for Competing Risks

Competing Risks

- When each subject may fail due to one of K ($K \geq 2$) causes, called competing risks.
- Occurrence of one of these events precludes us from observing the other event on this patient.
- Another classical example of competing risks is cause-specific mortality, such as death from heart disease, death from cancer, death from other causes, etc.
- Let X_i , $i = 1, \dots, K$ be the potential unobservable time to occurrence of the i th competing risk. What we observe for each patient is the time at which the subject fails from any cause, $T = \min(X_1, \dots, X_p)$ and an indicator δ which tells which of the K risks caused the patient to fail, that is, $\delta = i$ if $T = X_i$

Competing Risks

- The basic competing risks parameter is the cause-specific hazard rate for risk i defined by:

$$\begin{aligned}h_i(t) &= \lim_{\Delta t \rightarrow 0} \frac{P[t \leq T < t + \Delta t, \delta = i | T \geq t]}{\Delta t} \\&= \lim_{\Delta t \rightarrow 0} \frac{P[t \leq T < t + \Delta t, \delta = i | x_j \geq t, j = 1, 2, \dots, k]}{\Delta t}\end{aligned}$$

$h_i(t)$ tells us the rate at which subjects who have yet to experience any of the competing risks are experiencing the i th competing cause of failure.

- The overall hazard rate of the time to failure, T , is the sum of these K cause-specific hazard rates; that is

$$h_T(t) = \sum_{i=1}^K h_i(t)$$

Competing Risks

- The cause-specific hazard rate can be derived from the joint survival function of the K competing risks. Let $S(t_1, \dots, t_K) = \Pr[x_1 > t_1, \dots, x_K > t_K]$.

$$h_i(t) = \frac{-\partial S(t_1, \dots, t_K) / \partial t_i |_{t_1=\dots=t_K=t}}{S(t, \dots, t)}$$

- In competing risks problems we are often interested not in the hazard rate but rather in some probability which summarizes our knowledge about the likelihood of the occurrence of a particular competing risk. Three probabilities are computed, crude, net, and partial crude probabilities.

Competing Risks

- The crude probability is the probability of death from a particular cause in the real world where all other risks are acting on the individual (e.g: $P(\text{death from heart disease}) \rightarrow P(\text{die from heart disease prior to age 50})$)
- The net probability is the probability of death in a hypothetical world where the specific risk is the only risk acting on the population. (e.g: $P(\text{die from heart disease in the counterfactual world where men can only die from heart disease})$)
- The partial crude probabilities are the probability of death in a hypothetical world where some risks of death have been eliminated. (e.g: $P(\text{dies from heart disease in a world where cancer has been cured})$).

Competing Risks

- The Crude probabilities are typically expressed by the cause-specific sub-distribution function, also known as the cumulative incidence function, is defined as:

$$F_i(t) = P[T \leq t, \delta = i] = \int_0^t h_i(\mu) \exp\{-H_T(\mu)\} d\mu$$

Here $H_T(t) = \sum_{j=1}^K \int_0^t h_j(\mu) d\mu$ is the cumulative hazard rate of T .

- The net survival function, $S_i(t)$, is the marginal survival function found from the joint survival function by taking $t_j = 0$ for all $j \neq i$. When the competing risks are independent then the net survival function is related to the crude probabilities by:

$$S_T(t) \leq S_i(t) \leq 1 - F_i(t)$$

Competing Risks

- For partial crude probabilities, let J be the set of causes that an individual can fail from and J^C the set of causes which are eliminated from consideration.
- Let $T^J = \min(x_i, i \in J)$ then define the partial crude sub-distribution function by $F_i^J(t) = \Pr[T^J \leq t, \delta = i], i \in J$
- Define a partial crude hazard rate:

$$\lambda_i^J(t) = \frac{-\partial S(t_1, \dots, t_K) / \partial t_i |_{t_j=t, t_j \in J; t_j=0, t_j \in J^C}}{S(t_1, \dots, t_p) |_{t_j=t, t_j \in J; t_j=0, t_j \in J^C}}$$