Chapter 1

Survival Analysis

1.1 Introduction to Survival Analysis

Survival analysis focuses on time-to-event data. Let T denote a random variable that is the time that an event occurs. In order to define T we need:

1. an unambiguous **time origin** (e.g. randomization time in a clinical trial, treatment initiation, hospital admission)
2. a **time scale** (e.g. days, years)
3. a definition of the **event** (e.g. death, clinical definition of a diabetes diagnosis, dementia onset, or myocardial infarction)

Event time variables are always **non-negative** i.e., $T \geq 0$.

T can be either **continuous** (defined on $(0,\infty)$) or **discrete** (taking a finite set of values e.g. $a_1, a_2, ..., a_n$).

In practice, T is usually subject to censoring. Let X denote a random variable that is the **censored event time**: $X = min(T, U)$ where U is a non-negative random variable denoting the censoring time.

If there is no censoring, non-survival statistical methods can easily be used. However, note that the selected method must still account for the fact that $T \geq 0$ e.g., traditional linear regression is likely not appropriate.

Throughout, we will use the subscript i to denote the variable for individual i e.g. T_i and X_i.

Types of censoring

- Right-censoring: instead of T_i, we observe $X_i = \min(T_i, U_i)$ because of loss to follow-up, drop-out, or study termination. This is called right censoring because the true unobserved event time is to the right of the censoring time. All we know is that the event has not yet happened at the censoring time.

  In addition to observing X_i, we also observe the **event indicator**:

  $$\delta_i = \begin{cases} 1 & \text{if} \quad T_i \leq U_i \\ 0 & \text{if} \quad T_i > U_i \end{cases}$$

  Right-censoring is the most common type of censoring assumption we will deal with in survival analysis.

- Left-censoring: instead of T_i, we observe $Y_i = \max(T_i, U_i)$ and the event indicator:

  $$\epsilon_i = \begin{cases} 1 & \text{if} \quad U_i \leq T_i \\ 0 & \text{if} \quad U_i > T_i \end{cases}$$

  For example, in studies of time to HIV seroconversion, some of the enrolled subjects have already seroconverted at entry into the study - they are left-censored.

- Interval-censoring: Interval-censored event times most commonly arise when the exact time of the event is 'silent' but the interval in which it occurred is known. In a clinical trial examining an intervention for diabetes prevention, participants may have their fasting plasma glucose measured every 6 months. A diabetes diagnosis is defined by fasting plasma glucose. If a participant is diagnosed with diabetes at some time R_i, their actual time to diabetes is actually within the interval (L_i, R_i) where L_i is their previous testing time.

Noninformative vs. informative censoring

Censoring is **noninformative** if U_i is independent of T_i. Unless otherwise stated, methods to analyze event times generally assume that censoring is noninformative. Noninformative censoring implies that an individual censored at U is representative of all subjects who survived to U.

Censoring is considered **informative** if the distribution of U_i contains any information about the parameters characterizing the distribution of T.

Examples: If U_i is the planned end of the study e.g. 2 years after randomization, then it is usually independent of the event times. This is often referred to as administrative censoring. If U_i is the time that a patient drops out of the study because they've gotten much sicker and can't come to the follow-up visit, then U_i and T_i are probably not independent.

ID	X_i (days)	δ_i
1	32	1
2	5	0
3	200	1
4	250	0
5	181	0
6	95	1

Class exercise:

(a) Illustrate this dataset by drawing a timeline for each individual, using an X for an event and an open circle for censoring.
(b) Now assume there is administrative censoring at a time of 180 days. Recreate the table and figure.

1.2 Distribution of T

There are several ways to characterize the distribution of an event time random variable. Some of these should be familiar; others are specific to the event time setting. We will use the following terms:

- $f(t)$, the probability density function (pdf)
- $F(t)$, the cumulative distribution function (cdf)
- $S(t)$, the survival function
- $\lambda(t)$, the hazard function
- $\Lambda(t)$, the cumulative hazard function

The probability density function is defined as

$$f(t) = \lim_{\Delta \to 0} \frac{1}{\Delta} P(t \leq T < t + \Delta).$$

The cumulative distribution function and survival function are defined as

$$F(t) = P(T \leq t) = \int_0^t f(x)dx, \text{ and}$$

$$S(t) = 1 - F(t) = P(T > t).$$

That is, $S(t)$ is the probability that the event occurs after time t.

The hazard function and cumulative hazard function are defined as

$$\lambda(t) = \lim_{\Delta \to 0} \frac{1}{\Delta} P(t \leq T < t + \Delta | T \geq t), \text{ and}$$

$$\Lambda(t) = \int_0^t \lambda(u)du.$$

This assumes $f(t)$ is continuous; things are more complicated when this is not true.

Facts about these functions:

$$S(t) = \int_t^\infty f(u)du,$$

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

$$S(t) = e^{-\Lambda(t)}.$$

Measuring Central Tendency in Survival

Mean Survival, μ

$$\mu = \int_0^\infty x f(x) dx$$

Median survival, τ

$$S(\tau) = 0.5$$

1.3 Parametric Estimation of Survival

We can estimate the survival (or hazard function) two different ways:

1. Specifying a parametric model
2. Developing an empirical estimate i.e., nonparametric estimation

Common parametric distributions used in survival analysis

- **Exponential** distribution, parameter $\lambda > 0$

$$\begin{aligned} f(t) &= \lambda e^{-\lambda t} \text{ for } t \geq 0 \\ S(t) &= \int_t^\infty f(u)du = e^{-\lambda t} \\ \lambda(t) &= \frac{f(t)}{S(t)} = \lambda \text{ (constant hazard!)} \\ \Lambda(t) &= \int_0^t \lambda(u)du = \int_0^t \lambda du = \lambda t \end{aligned}$$

  The log-transformation of the Exponential distribution is the extreme value distribution.

- **Weibull** distribution, scale parameter $\lambda > 0$ and shape parameter $\kappa > 0$

$$\begin{aligned} f(t) &= \kappa\lambda(\lambda t)^{\kappa-1}e^{-(\lambda t)^\kappa} \text{ for } t \geq 0 \\ S(t) &= e^{-(\lambda t)^\kappa} \\ \lambda(t) &= \kappa\lambda(\lambda t)^{\kappa-1} \\ \Lambda(t) &= (\lambda t)^\kappa \end{aligned}$$

  The Weibull distribution is nice because it is more flexible than the exponential, but still has relatively simple expressions for the different functions. It covers several hazard shapes:

  $\kappa = 1 \rightarrow$ constant hazard, equivalent to Exponential w/ parameter λ.

  $0 < \kappa < 1 \rightarrow$ decreasing hazard

  $\kappa > 1 \rightarrow$ increasing hazard

  Be careful with the scale and shape parameters; different software uses different parametrization.

- **Gamma** distribution, scale parameter $\theta > 0$ and shape parameter $\kappa > 0$

$$f(t) = \frac{1}{\Gamma(\kappa)\theta^\kappa}t^{\kappa-1}e^{-\frac{t}{\theta}}$$

where Γ is the Gamma function; for an integer κ, $\Gamma(\kappa) = (\kappa - 1)!$.

There is no nice form for the survival or hazard functions. When $\kappa = 1$ and $\theta = \frac{1}{\lambda}$ then the Gamma distribution is equivalent to the Exponential distribution with parameter λ.

Be very careful with the scale and shape parameters; different software uses different parametrization.

- **Log-normal** distribution, parameters $\mu \in (-\infty, \infty)$ and $\sigma > 0$

  If $N \sim \text{Normal}(\mu, \sigma^2)$, then $T = \exp(N)$ is log-normal, denoted Log-normal(μ, σ^2).

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log(t) - \mu)^2}{2\sigma^2}\right\}$$

  The mean is $e^{\mu+\sigma^2/2}$. Often considered intuitive and convenient.

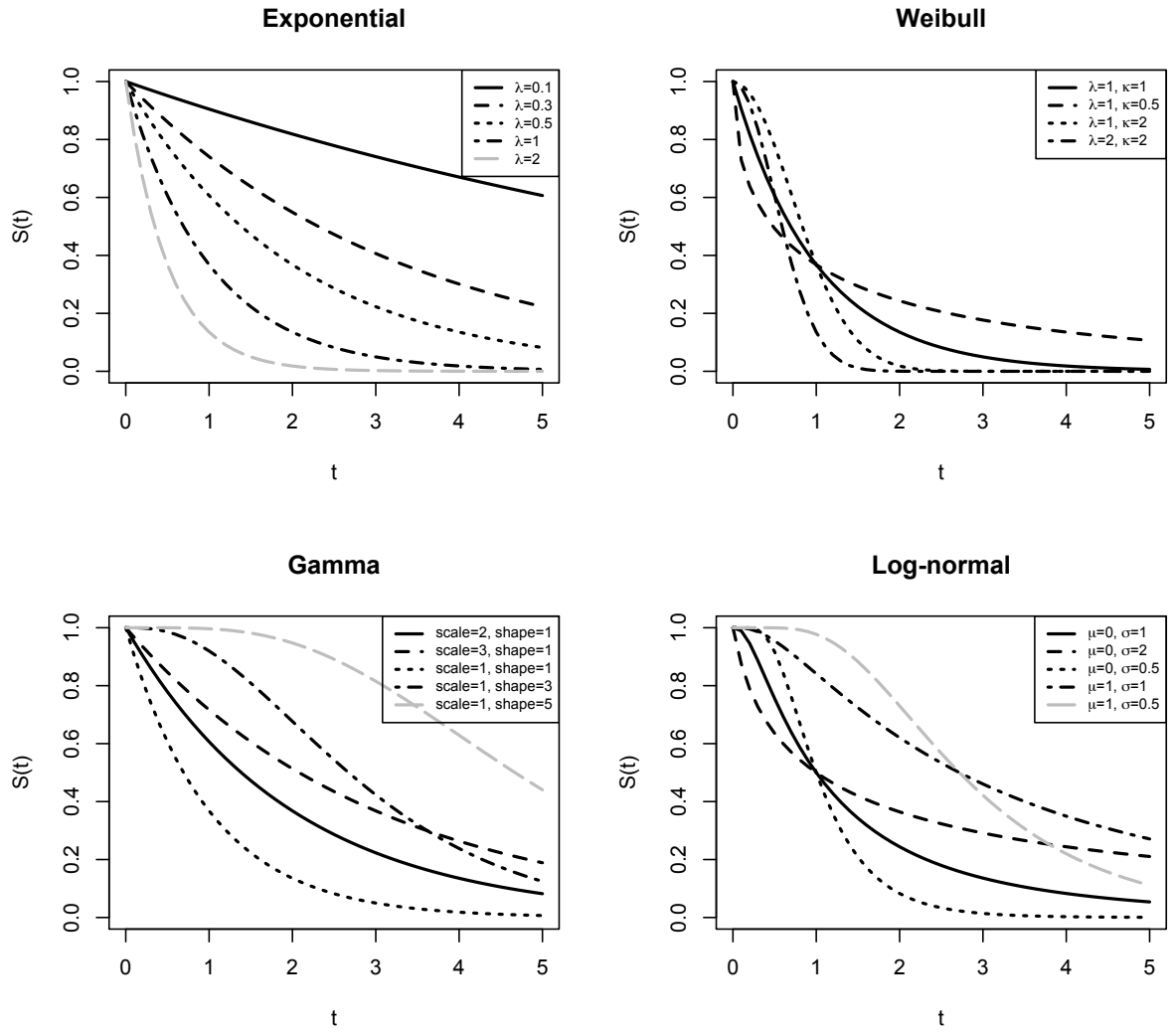


Figure 1.1:Parametric Survival Functions

Parametric estimation $\rightarrow$ consider the likelihood

Let $f_\lambda(t)$ denote the probability density function of T where λ denotes any parameters of the function, and $S_\lambda(t)$ is the corresponding survival function. The observed censored data is denoted as $\{(x_i, \delta_i), i = 1, ..., n\}$.

In survival analysis, under noninformative censoring, the likelihood is

$$L(\lambda) = \prod_{i=1}^{n} f_\lambda(x_i)^{\delta_i} S_\lambda(x_i)^{(1-\delta_i)}.$$

The maximum likelihood estimate (MLE) of λ is the value which maximizes $L(\lambda)$.

For the Exponential distribution, it can be shown that the MLE of λ is

$$\widehat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} x_i}.$$

1.4 Nonparametric Estimation of Survival

Now suppose we do not want to assume anything about the distribution ⇒ nonparametric estimation.

When there is no censoring, one can simply use the empirical distribution function to estimate $S(t)$:

$$\widehat{S}(t) = 1 - \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq t)$$

With censoring, one could consider the following extreme options:

1. Assume the event occurred at $x_i \Rightarrow$ use $\widehat{S}(t)$
2. Assume anyone who was censored never has the event i.e. $t_i > \tau$ where τ is the end of the study
3. Use only complete data i.e., drop all censored observations

None of these are good options.

The **Kaplan-Meier** estimate of survival, the most commonly used nonparametric estimate of $S(t)$, makes use of all the information available in the sample. The Kaplan-Meier estimate is based on the concept of conditional probability. Recall that:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

$$P(A,B) = P(A|B)P(B)$$

$$\text{If } A \perp B, \text{ then } P(A,B) = P(A)P(B)$$

Let $t_{(1)} < t_{(2)} < t_{(3)} < ... < t_{(K)}$ be the ordered unique event times.

- There are K unique event times
- $t_{(K)}$ is the final observed event time
- Partition the time scale based on the ordered event times:

$$[0, t_{(1)}) \cup [t_{(1)}, t_{(2)}) \cup [t_{(2)}, t_{(3)}) \cup ... \cup [t_{(K-1)}, t_{(K)}) \cup [t_{(K)}, \infty)$$

- $K + 1$ mutually exclusive intervals

To see how the Kaplan-Meier estimate uses conditional probability consider $P(T > t_{(j)})$.

$$\begin{aligned}
P(T > t_{(j)}) &= P(T > t_{(j)}, T > t_{(j-1)}, ..., T > t_{(1)}) \\
&= P(T > t_{(j)}, T > t_{(j-1)}, ..., T > t_{(2)} | T > t_{(1)}) P(T > t_{(1)}) \\
&= P(T > t_{(j)}, T > t_{(j-1)}, ..., T > t_{(3)} | T > t_{(2)}) P(T > t_{(2)} | T > t_{(1)}) P(T > t_{(1)}) \\
&= ... \\
&= .P(T > t_{(j)} | T > t_{(j-1)}) P(T > t_{(j-1)} | T > t_{(j-2)}) ... P(T > t_{(2)} | T > t_{(1)}) P(T > t_{(1)}) \\
&= \prod_{k=1}^{j} P(T > t_{(k)} | T > t_{(k-1)}) \\
&= \prod_{k=1}^{j} \left\{ 1 - P(T \leq t_{(k)} | T > t_{(k-1)}) \right\}
\end{aligned}$$

where $t_0 = 0$ and $P(T > t_0) = 1$.

Define the **risk set** $\mathcal{R}_k$ to be the set of individuals who were at risk for the event at time $t_{(k)}$.

For risk set $\mathcal{R}_k$ let

- n_k denote the number of individuals at risk at time $t_{(k)}$
- d_k denote the number of events at time $t_{(k)}$

The **Kaplan-Meier estimate of** $S(t)$ is:

$$\widehat{S}_{KM}(t) = \prod_{k:t_{(k)} \leq t} 1 - \frac{d_k}{n_k}$$

- An estimate of the variance of $\widehat{S}_{KM}(t)$ can be obtained as

$$\widehat{\text{var}}(\widehat{S}_{KM}(t)) = [\widehat{S}_{KM}(t)]^2 \sum_{k:t_{(k)} \leq t} \frac{d_k}{(n_k - d_k) n_k}$$

- This is known as Greenwood's formula and its derivation involves a log transformation and use of the delta method. Due to proven asymptotic properties of $\widehat{S}_{KM}(t)$, one could calculate a 95% confidence interval (CI) for $S(t)$ as

$$\left(\widehat{S}_{KM}(t) - 1.96 \widehat{\sigma}_{KM}(t), \widehat{S}_{KM}(t) + 1.96 \widehat{\sigma}_{KM}(t) \right)$$

  where $\widehat{\sigma}_{KM}(t) = \sqrt{\widehat{\text{var}}(\widehat{S}_{KM}(t)}$. However, this may result in a CI with bounds below 0 or above 1.

- An alternative it to calculate the 95% CI for a transformation of $S(t)$ and then transform back to the $S(t)$ scale. For example, it can be shown that one can estimate the variance of $\log\{\widehat{S}_{KM}(t)\}$ as

$$\widehat{\text{var}}[\log\{\widehat{S}_{KM}(t)\}] = \sum_{k:t_{(k)}\leq t} \frac{d_k}{(n_k - d_k)n_k}$$

  and thus, a 95% CI for $\log\{S(t)\}$ is (A, B), defined as:

$$\left(\log\{\widehat{S}_{KM}(t)\} - 1.96\sqrt{\{\widehat{\text{var}}[\log\{\widehat{S}_{KM}(t)\}]\}}, \log\{\widehat{S}_{KM}(t)\} + 1.96\sqrt{\{\widehat{\text{var}}[\log\{\widehat{S}_{KM}(t)\}]\}}\right).$$

- A 95% CI for $S(t)$ can be calculated as (e^A, e^B). The lower bound of this CI is bounded by 0, but the upper bound can still be greater than 1. Other transformations can be considered to force an upper bound. `R` **will truncate the bounds of the CI at 0 and 1.**

$\Rightarrow$ Be careful: different software does different things for the variance and CI. This class will follow the approach used by `R`; `R` uses the log transformation for the confidence interval by default.

Although the Kaplan-Meier estimate of survival is the most commonly used nonparametric estimate of $S(t)$, it is also worth knowing about the **Nelson-Aalen** estimator. The Nelson-Aalen estimator is a nonparametric estimate of the cumulative hazard function, but recall that we can get survival from the cumulative hazard as $S(t) = e^{-\Lambda(t)}$.

Remember that $\Lambda(t) = \int_0^t \lambda(u)du$. Similar to the Kaplan-Meier estimate, consider dividing up time into intervals, let's say the finest interval we can e.g. days for our working example.

The **Nelson-Aalen estimator** uses the concept of numerical integration to approximate $\Lambda(t)$ with a sum:

$$\widehat{\Lambda}_{NA}(t) = \sum_{k:t_k\leq t} \lambda(k)\Delta_k = \sum_{k:t_k\leq t} d_k/n_k$$

where for example Δ_k is the width of the time interval e.g. 1 day.

- The corresponding estimate of survival is $\widehat{S}_{FH}(t) = e^{-\widehat{\Lambda}_{NA}(t)}$ where FH stands for Fleming-Harrington.
- This survival estimate will often be close to but not exactly equal to $\widehat{S}_{KM}(t)$.

Why nonparametric estimation?

- If you use parametric estimation and your assumed model is incorrect, there is a risk that you will obtain incorrect estimates and make incorrect conclusions.
- If you use nonparametric estimation and a parametric model is correct, you will lose efficiency compared to if you had used the correct parametric estimation. That is, the variance of your estimates will be larger. This is especially problematic if you have a small sample size.
- Ask: does this distribution make sense substantively? For example, if your outcome is death following hernia surgery, ask a clinician - is it reasonable to assume that the risk of surviving from one day to the next (informal interpretation of hazard) is constant over time (this would be an exponential distribution)? No, probably not.
- Look at your data.

1.5 Comparing Survival Curves

The previous section focused on estimating the survival function, $S(t)$, over time for a single sample of individuals. Our data consisted of X_i and δ_i for each individual i. Now we assume that we observe $(X_i, \delta_i, \mathbf{Z}_i)$ where

- X_i is a censored event time random variable
- δ_i is the event indicator
- $\mathbf{Z}_i$ is a set of covariates or characteristics

Note that $\mathbf{Z}_i$ might be scalar (a single covariate, say treatment or gender) or may be a (p x 1) vector (representing several different covariates).

These covariates might be:

- continuous
- discrete
- time-varying/time-dependent

If $\mathbf{Z}_i$ is scalar and binary, then we are comparing the survival of two groups, like in the smoking example. More generally though, it is useful to build a model that characterizes the relationship between survival and all of the covariates of interest.

We will proceed as follows:

- Two-group comparison
- Multigroup and stratified comparisons
- Survival regression models e.g. Cox proportional hazards model

Two-group comparison

One approach to formally evaluating differences between groups is to compare their time-to-event experience at a particular time point, say t^*. Let $S_0(\cdot)$ and $S_1(\cdot)$ be the survivor functions for the two groups e.g. 1 is treatment and 0 is control.

We can consider testing the following hypothesis:

$$H_0 : S_1(t^*) = S_0(t^*)$$

vs

$$H_A : S_1(t^*) \neq S_0(t^*)$$

This is equivalent to considering:

$$H_0 : S_1(t^*) - S_0(t^*) = 0$$

vs

$$H_A : S_1(t^*) - S_0(t^*) \neq 0$$

We can construct a 95% confidence interval for this difference as:

$$\{\widehat{S}_1(t^*) - \widehat{S}_0(t^*)\} \pm 1.96\sqrt{\{\widehat{\sigma}_1^2(t^*) + \widehat{\sigma}_0^2(t^*)\}}$$

where here we are using the Kaplan-Meier estimate of survival and Greenwood's formula for the variance estimates. Note that this assumes independence between the two groups. We would then reject H_0 if the interval does not include 0.

Many clinical studies focus on estimating survival at a single time point t^* and testing for a difference at this time point. Using a single time point is also often used in personalized medicine when the goal is to provide a patient with an estimate of their likelihood of survival to t^*.

However, there are some limitations with focusing on a single time point:

- Why t^*? How to select a single or optimal t^*?
- Potential for abuse when applied post-hoc i.e. "fishing"
- Not looking at all the other information available. What about the other times? $\rightarrow$ inefficient use of the available information

Can we compare survivor functions across the entire observed time frame (see Figure 1.2)? Yes, using the **log-rank test**.

We now consider the hypothesis:

$$H_0 : S_1(t) = S_0(t) \text{ for all } t > 0$$

vs

$$H_1 : S_1(t) \neq S_0(t) \text{ for some } t > 0$$

Similar to the intuition behind the construction of the Kaplan-Meier estimate, recall that the Kaplan-Meier survival function only changes at the event times. Here, we will pool the data across both treatment groups and let

$$t_{(1)} < t_{(2)} < ... < t_{(K)}$$

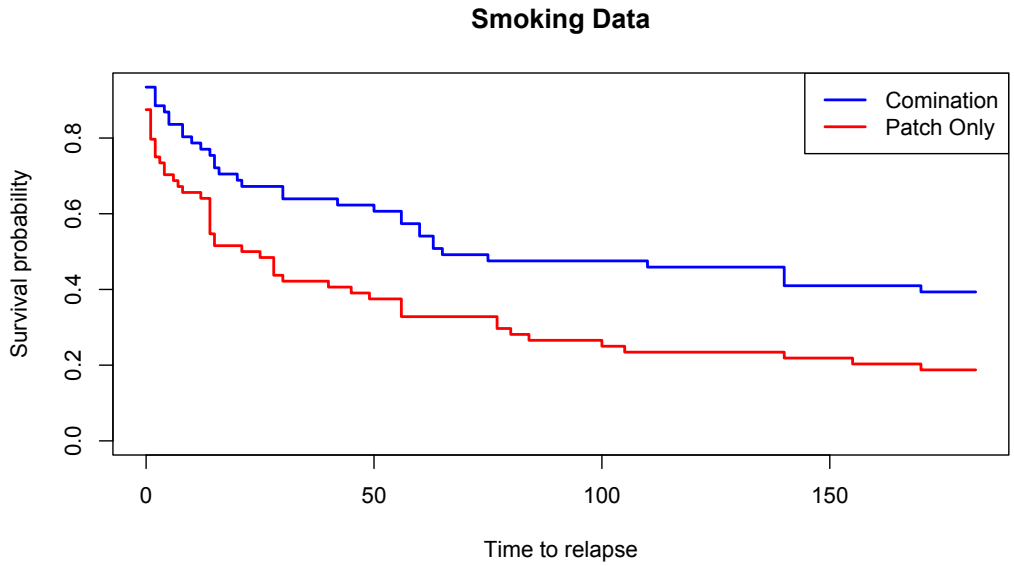


Figure 1.2:Example: Time to Relapse in Smoking Study

denote the ordered, observed event times, where there are K distinct event times in the entire sample.

The **log-rank test** compares the survival functions $S_1(t)$ and $S_0(t)$ across the observed time frame by considering differences between the observed number of events in Group 1 and the corresponding expected number of events across all K unique event times.

At the k^{th} event time, consider the 2x2 table:

	No event	**Event**	
Group 0	$n_{0k} - d_{0k}$	d_{0k}	n_{0k}
Group 1	$n_{1k} - d_{1k}$	d_{1k}	n_{1k}
	$n_k - d_k$	d_k	n_k

- n_k is the number of individuals at risk at time $t_{(k)}$
- d_k is the number of individuals who experience the event at time $t_{(k)}$

Under H_0, there is no association between group membership and the outcome, and, conditional on the margins of the 2 x 2 table, the distribution of the number of events in Group 1 is

$$D_{1K} \sim \text{Hypergeometric}(n_k, n_{1k}, d_k)$$

Based on the properties of a Hypergeometric distribution,

$$E[D_{1k}] = \frac{n_{1k}d_k}{n_k}$$

Now consider the difference between the observed and expected number of events in Group 1 at the k^{th} event time:

$$U_k = d_{1k} - E[D_{1k}] = d_{1k} - \frac{n_{1k}d_k}{n_k}$$

By construction, under H_0,

$$E[U_k] = 0$$

and it can be shown that

$$Var[U_k] = \frac{n_{1k}n_{0k}(n_k - d_k)d_k}{n_k^2(n_k - 1)}$$

Let's denote this as V_k. Now for each 2 x 2 table constructed at time t_k, we have U_k and V_k. The log-rank test statistic pools the U_k contributions across the K unique event times and under H_0, this pooled statistic follows a standard normal distribution:

$$\frac{\sum_{k=1}^{K} U_k}{\sqrt{\sum_{k=1}^{K} V_k}} \sim N(0,1)$$

Equivalently, the square of this pooled statistic has a chi-square distribution with 1 degree of freedom under H_0:

$$T_{LR} = \frac{[\sum_{k=1}^{K} U_k]^2}{\sum_{k=1}^{K} V_k} \sim \chi_1^2$$

T_{LR} is the log-rank test statistic $\rightarrow$ use χ_1^2 to compute a p-value.

- The log-rank test statistic assigns equal weight to each risk set
  - the numerator simply adds up the K U_k contributions
- A consequence of this is that it can be sensitive to differences in the tails of the survivor function
  - where the risk sets are smallest
  - where there is the least amount of information
- Big differences in the latter part of the time scale may have a large influence on the conclusions drawn regarding the entire time scale

Multigroup comparison

Suppose we have J groups and we would like to assess differences across them. Specifically, we are interested in testing:

$$H_0 : S_1(t) = ... = S_J(t) \text{ for all } t > 0$$

vs

$$H_1 : S_j(t) \neq S_{j'}(t) \text{ for some } t > 0 \text{ for at least some } j \neq j'$$

As with the two-group log-rank test, we first pool the data across the J treatment groups to obtain the ordered, observed event times:

$$t_{(1)} < t_{(2)} < ... < t_{(K)}$$

The test statistic for the J group setting is a straightforward generalization of the two-group statistic. At the k^{th} event time, consider the J X 2 table:

	No event	**Event**	
Group 1	$n_{1k} - d_{1k}$	d_{1k}	n_{1k}
Group 2	$n_{2k} - d_{2k}$	d_{2k}	n_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Group J	$n_{Jk} - d_{Jk}$	d_{Jk}	n_{Jk}
	$n_k - d_k$	d_k	n_k

- n_{jk} is the number of individuals at risk in the j^{th} group at time $t_{(k)}$
- d_{jk} is the number of individuals who experience the event in the j^{th} group at time $t_{(k)}$

For this table, let

$$\mathbf{O}_k = (d_{1k}, d_{2k}, ..., d_{(J-1)k})$$

denote the observed number of events in groups 1 to (J-1) at time $t_{(k)}$. Conditional on margins of the J X 2 table, under H_0, the expected number of events is

$$\mathbf{E}_k = \left(\frac{n_{1k}d_k}{n_k}, \frac{n_{2k}d_k}{n_k}, ..., \frac{n_{(J-1)k}d_k}{n_k} \right)$$

Pooling across the K unique event times gives the statistic:

$$T_{LR} = (\mathbf{O} - \mathbf{E})^T \mathbf{V}^{-1} (\mathbf{O} - \mathbf{E}),$$

where $\mathbf{O} = \sum_k \mathbf{O}_k$, $\mathbf{E} = \sum_k \mathbf{E}_k$, and $\mathbf{V} = \sum_k \mathbf{V}_k$. Here, $\mathbf{V}_k$ denotes the (J-1) X (J-1) variance-covariance matrix of $\mathbf{O}_k - \mathbf{E}_k$.

Under H_0 of no differences across the J treatment groups, T_{LR} has a χ^2 distribution with J-1 degrees of freedom.

Stratified comparison

Stratification is an important tool. In observational studies, stratification is one approach to control for potential confounding bias. In randomized trials, stratification is often used to ensure balance in the treatment allocation across levels of an important prognostic factor e.g. "Randomization was stratified according to the predominant site of distance metastasis(visceral or other), age ($\leq$ 42 years or $>$ 42 years), and estrogen-receptor status (positive or negative)."

The stratified log-rank test is one way to adjust for a single categorical variable when testing for a difference between the two treatment groups.

Suppose there are M strata. For the smoking study, $M = 2$ because our stratification variable is gender (see Figure 1.3). Consider testing the hypothesis:

$$H_0 : S_{1,m}(t) = S_{0,m}(t)$$

for all m and all $t > 0$ where

- $S_{1,m}(t)$ denotes the survivor function in group 1 within stratum m
- $S_{0,m}(t)$ denotes the survivor function in group 0 within stratum m

Under H_0, the survival function is the same between the two treatment groups, across all M strata. Again, pool the data across both treatment groups and all M strata to obtain the ordered, observed event times:

$$t_{(1)} < t_{(2)} < ... < t_{(K)}$$

At the k^{th} event time, consider the 2 x 2 table in stratum m:

	No event	**Event**	
Group 0	$n_{0k,m} - d_{0k,m}$	$d_{0k,m}$	$n_{0k,m}$
Group 1	$n_{1k,m} - d_{1k,m}$	$d_{1k,m}$	$n_{1k,m}$
	$n_{k,m} - d_{k,m}$	$d_{k,m}$	$n_{k,m}$

- $n_{1k,m}$ is the number of individuals at risk in treatment group 1 at time $t_{(k)}$ from stratum m
- $d_{1k,m}$ is the number of individuals who experience the event in treatment group 1 at time $t_{(k)}$ from stratum m

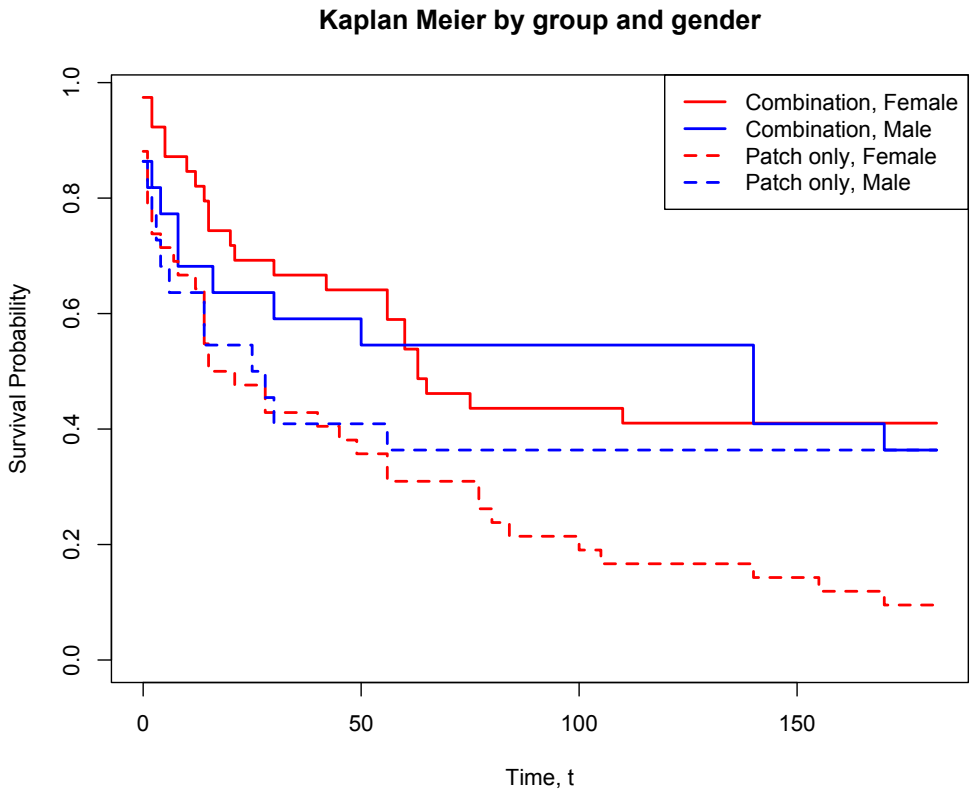


Figure 1.3: Example: Time to Relapse in Smoking Study, Stratified by Gender

Note that there are M such 2 x 2 tables at each unique event time.

Under H_0, and conditional on the stratum margins, the expected number of events in group 1 at time $t_{(k)}$ from stratum m is

$$E[D_{1k,m}] = \frac{n_{1k,m}d_{k,m}}{n_{k,m}}$$

Let

$$U_{k,m} = d_{1k,m} - \frac{n_{1k,m}d_{k,m}}{n_{k,m}}$$

be the difference between the observed and expected number of events in group 1 at time $t_{(k)}$ from stratum m.

Under H_0, $E[U_{k,m}] = 0$ and

$$Var[U_{k,m}] = V_{k,m} = \frac{n_{1k,m}n_{0k,m}(n_{k,m} - d_{k,m})d_{k,m}}{n_{k,m}^2(n_{k,m} - 1)}$$

The stratified log-rank test statistic pools across the M strata and K unique event times:

$$T_S = \frac{[\sum_{m=1}^{M}\sum_{k=1}^{K} U_{k,m}]^2}{\sum_{m=1}^{M}\sum_{k=1}^{K} V_{k,m}}$$

Under H_0, $T_S \sim \chi_1^2 \rightarrow$ use χ_1^2 to compute a p-value.

The stratified log-rank test generally has good power when the direction of the effect is the same within each stratum. Implicitly it is taken that the goal is to estimate a common effect across strata.

If you are interest in estimating and evaluating stratum-specific effects (i.e., interaction or effect modification), then you should not use the stratified log-rank test.

Weighted log-rank tests

Recall that the log-rank test statistic assigns equal weight to each risk set. In certain settings, you may want to give different weighting to different risk sets i.e., time points. Let w_k be the weight for the risk set at time t_k, then the weighted log-rank test statistic is

$$T_W = \frac{[\sum_{k=1}^{K} w_k U_k]^2}{\sum_{k=1}^{K} w_k^2 V_k}$$

The strategy is to select weights, w_k, that emphasize parts of the time scale where there are differences and/or de-emphasize parts where there are no differences.

For obvious reasons, it is important to pre-specify these weights, otherwise there is strong potential for post-hoc abuse. For randomized trials, these weights often have to be pre-registered in the analysis plan.

Many tests have been proposed with various weights:

Test statistic	**Weight**, w_k
Log-rank	1
Gehan-Breslow	n_k
Peto/Prentice	$n\widehat{S}(t_{(k)})$
Generalized Wilcoxon	$\widehat{S}(t_{(k)})$
Fleming-Harrington	$[\widehat{S}(t_{(k)})]^{\rho}[1-\widehat{S}(t_{(k)})]^{\gamma}$

For example, the Gehan-Breslow test weights the contribution from the k^{th} risk set by the number of subjects at risk, n_k, meaning there is larger weight given to earlier risk sets.

1.6 Modeling Survival Data: Regression

Recall that we observe $(X_i, \delta_i, \mathbf{Z}_i)$ where

- X_i is a censored event time random variable
- δ_i is the event indicator
- $\mathbf{Z}_i$ is a set of covariates or characteristics

When $\mathbf{Z}_i$ is a single variable that is either binary or categorical, we can use the log-rank test. But when $\mathbf{Z}_i$ is continuous or represents multiple variables, we must instead consider an approach to model the distribution of $T|\mathbf{Z}$.

Outside of survival analysis, this is where one would consider linear regression to model $E[T|\mathbf{Z}]$. However, we cannot simply do linear regression with time-to-event outcome data because:

- $T > 0$ and such a model could result in negative predictions for T
- time-to-event outcome distributions tend to be right-skewed
- the mean may not be a good measure of 'centrality'

Parametric Regression

Let's take a look at extending the parametric estimation we discussed earlier to a setting where we want to incorporate $\mathbf{Z}$. Suppose that

$$T_i \sim \text{Exponential}(\lambda_i)$$

Notice now that what we used to have as λ is specific to each study unit (e.g. person): λ_i. To incorporate $\mathbf{Z}_i$, and remembering that the parameter for the exponential itself must be > 0 we propose:

$$\log(\lambda_i) = \beta^T \mathbf{Z}_i$$

or alternatively,

$$\lambda_i = \exp\{\beta^T \mathbf{Z}_i\}$$

where β is a vector of regression coefficients. This is parametric regression assuming an exponential distribution for T. This can written as an **accelerated failure time (AFT) model**.

The general representation of an AFT model is:

$$\log(T_i) = \beta^T \mathbf{Z}_i + \epsilon_i$$

coupled with some specification of the distribution of ϵ_e,

The choice of the distribution of ϵ is an important one in that it:

- Defines the distribution of the time-to-event outcome T_i
- Dictates the flexibility with which the AFT model can represent complex data

Cox Proportional Hazards Model

An alternative to parametric regression in survival analysis is the **Cox Proportional Hazards model** which is the most common model used for survival data. The Cox proportional hazards model specifies the following model for the hazard function:

$$\lambda(t|\mathbf{Z}) = \lambda_0(t)\exp\{\beta^T\mathbf{Z}\}$$

- Note that now we are talking about modeling the hazard, which is not as straightforward to interpret
- $\lambda_0(t)$ is referred to as the baseline hazard function $\rightarrow$ it is the hazard for the population with $\mathbf{Z} = \mathbf{0}$
- $\lambda_0(t)$ is left unspecified $\rightarrow$ that is, we do not assume anything about it
- For this reason, the Cox proportional hazards model is a semiparametric model i.e. "half" parametric, "half" nonparametric

To interpret the components of β, consider two values for the covariates $\mathbf{Z}$:

$$\mathbf{z} = (z_1, ..., z_j, ..., z_p)$$

$$\mathbf{z}' = (z_1, ..., z_j + 1, ..., z_p)$$

Let's compare the hazard with $\mathbf{z}$ vs. the hazard with $\mathbf{z}'$ by looking at the ratio:

$$\begin{aligned}\frac{\lambda(t|\mathbf{Z}=\mathbf{z}')}{\lambda(t|\mathbf{Z}=\mathbf{z})} &= \frac{\lambda_0(t)\exp\{\beta^T\mathbf{z}'\}}{\lambda_0(t)\exp\{\beta^T\mathbf{z}\}} \\ &= \frac{\exp\{\beta^T\mathbf{z}'\}}{\exp\{\beta^T\mathbf{z}\}} \\ &= \exp\{\beta_j\}\end{aligned}$$

- $\exp\{\beta_j\}$ is the relative change in the hazard at time t associated with a unit change in Z_j, holding everything else constant
- $\exp\{\beta_j\}$ is the **hazard ratio**

Assuming the outcome is "bad":

- If $\beta_j < 0$ then $\exp\{\beta_j\} < 1$, and the impact of covariate Z_j is to **decrease** the magnitude of the hazard
    - intuitively, risk is **lower** and the impact is one of **benefit**
- If $\beta_j > 0$ then $\exp\{\beta_j\} > 1$, and the impact of covariate Z_j is to **increase** the magnitude of the hazard
    - intuitively, risk is **higher** and the impact is one of **harm**

The Cox model is a proportional hazards model in the sense that the hazard ratio does not depend on time. Regardless of the values of t, we have that:

$$\frac{\lambda(t|\mathbf{Z} = \mathbf{z}')}{\lambda(t|\mathbf{Z} = \mathbf{z})} = \exp\{\beta_j\}$$

where $\exp\{\beta_j\}$ has no t involved.

The fact that this model characterizes associations in this way implies that:

- we only need to report a single number regarding the impact of a covariates
- we don't need to invoke time when interpreting the results from a study

However, whether or not this assumption of proportional hazards **is true** for a study is important to think about; we will come back to this.

To estimate β_j, Cox (1972) proposed the idea of a **partial likelihood** (this is different from the usual specification of the likelihood). The idea incorporates some of the same concepts we have already discussed:

1. Contributions to the likelihood are only made when there is an event (not at censored times, and not at times in between event times)
2. Consider the "risk set": who is everyone at risk at a certain time?

At a particular event time t_j, the contribution to the partial likelihood is:

$$\begin{aligned} L_j(\beta) &= \frac{\lambda(t|\mathbf{Z}_j)}{\sum_{l \in \mathcal{R}_j} \lambda(t|\mathbf{Z}_l)} \\ &= \frac{\lambda_0(t_j)\exp\{\beta^T \mathbf{Z}_j\}}{\sum_{l \in \mathcal{R}_j} \lambda_0(t_j)\exp\{\beta^T \mathbf{Z}_l\}} \\ &= \frac{\exp\{\beta^T \mathbf{Z}_j\}}{\sum_{l \in \mathcal{R}_j} \exp\{\beta^T \mathbf{Z}_l\}} \end{aligned}$$

where $\mathcal{R}_j$ is the risk set at time t_j and $\mathbf{Z}_j$ denotes the covariates for the individual who had the event at time t_j; notice that the baseline hazard $\lambda_0(t_j)$ drops out.

The partial likelihood is the product of these contributions over the K event times:

$$L^{partial}(\beta) = \prod_{j=1}^{K} \frac{\exp\{\beta^T \mathbf{Z}_j\}}{\sum_{l \in \mathcal{R}_j} \exp\{\beta^T \mathbf{Z}_l\}}$$

Cox (1972) argues that you can now treat this like a regular likelihood i.e., take the derivative and set it equal to zero and solve to obtain the maximum partial likelihood estimates. When you have multiple covariates, this is not so easy and usually involves the Newton Raphson method.

The output you will get will look a lot like regular regression. In R, there is an option called `robust` and the default is `robust=TRUE`; this obtains robust variance estimates for $\widehat{\beta}$ and is ideal. For more information see: Lin, D. Y., & Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. Journal of the American Statistical Association, 84(408), 1074-1078.

Interpretation

Let's take a look at the estimates obtained from fitting a Cox model.

Here, I am using the smoking study with the covariates: age (continuous, 22-86), gender (male vs. female), and treatment group (grp; Combination or Patch Only):

Covariate	β	$\exp(\beta)$	$se(\beta)$	z	$P(>\|z\|)$
age	-0.02178	0.97845	0.00972	-2.241	0.02501 *
genderMale	-0.12930	0.87871	0.23315	-0.555	0.57918
grppatchOnly	0.56295	1.75584	0.21696	2.595	0.00947 *

Note that R has used Female as the reference group and Combination as the reference group. Recall from above that the column of most interest for the same of interpretability is going to be $\exp(\beta)$ because this is the hazard ratio (HR).

You have to be very careful with how you interpret these coefficients. With this output, do NOT interpret results in terms of the survival probability or risk, you can only talk about the hazard and as we have discussed, the hazard is not that intuitive. Below is a paragraph reflecting what you might say in a paper to summarize the results above. In the next section, we will discuss taking this model and using it to estimate $S(t)$ which is more interpretable.

We investigated the association between time to relapse and age, gender, and treatment group using a Cox proportional hazards model. Results showed that age (hazard ratio [HR] = 0.978,

p<0.05) and treatment group(HR [reference is combination group] = 1.756, p<0.01) were significantly associated with time to relapse. Holding treatment group and gender constant, older age was associated with a decreased hazard of relapse. Holding age and gender constant, the estimated hazard of relapse in the patch only treatment group was 1.76 times the hazard of relapse in the combination group. Gender was not associated with time to relapse.

Estimating the Baseline Hazard

A major advantage of the Cox model is that we can ignore the baseline hazard and still get estimate of the hazard ratios for our covariates of interest. However, if you want to say anything about survival, not just hazard, then you do need an estimate of the baseline hazard. Recall that

$$S(t) = e^{-\Lambda(t)}$$

and if we are using the Cox proportional hazards model:

$$\begin{aligned} S(t|\mathbf{Z}) &= e^{-\Lambda(t|\mathbf{Z})} = \exp\left\{-\int_0^t \lambda(u|\mathbf{Z})du\right\} \\ &= \exp\left\{-\int_0^t \lambda(u|\mathbf{Z})du\right\} \\ &= \exp\left\{-\int_0^t \lambda_0(u)\exp\{\beta^T\mathbf{Z}\}du\right\} \\ &= \exp(-\Lambda_0(t)\exp\{\beta^T\mathbf{Z}\}) \end{aligned}$$

Breslow(1972) proposed to estimate the baseline hazard as follows:

- The baseline hazard at time t_k can be estimated as

$$\widehat{\lambda}_0(t_k) = \frac{d_k}{\sum_{l\in\mathcal{R}_j}\exp\{\beta^T\mathbf{Z}_l\}}$$

- The cumulative baseline hazard hazard at time t can then be estimated as

$$\widehat{\Lambda}_0(t) = \sum_{k:t_k\leq t}\frac{d_k}{\sum_{l\in\mathcal{R}_j}\exp\{\beta^T\mathbf{Z}_l\}}$$

Note that if you consider the single sample case with no covariates, i.e., set $\mathbf{Z} = 0$, this is equivalent to the nonparametric Nelson-Aalen estimator that we discussed previously.

We can now get an estimate of $S(t|\mathbf{Z})$ as

$$\widehat{S}(t|\mathbf{Z}) = \exp(-\widehat{\Lambda}_0(t)\exp\{\widehat{\beta}^T\mathbf{Z}\})$$

These are personalized estimates of survival.

1.7 Model Diagnostics

[Reference: Applied Survival Analysis Using R, Moore, D. F. (2016), Chapter 7.]

Residuals are often used for model checking in linear regression where they are typically plotted versus some quantity, e.g. a covariate value, and the observed pattern is used to diagnose possible problems with the fitted model. Some residuals have the additional property of not only indicating problems but also suggesting remedies. That is, the pattern of the plotted residuals may suggest an alternative model that fits the data better.

Many of these residuals have been generalized to survival analysis. In addition, the fact that survival data evolves over time, and requires special assumptions such as proportional hazards for the Cox model, additional diagnostic residual methods are necessary. Here, we focus on model diagnostics within the Cox proportional hazards model framework.

Martingale and Deviance Residuals

An important tool for assessing the goodness of fit of a model is to compare the censoring indicator (0 for censored, 1 for event) for each subject to the expected value of that indicator under the Cox model. If there are no time dependent covariates (we will discuss later) and if the event times are right-censored, the **martingale residual** is:

$$m_i = \delta_i - \widehat{\Lambda}_0(x_i) \exp\{\widehat{\beta}^T \mathbf{Z}_i\}$$

- These residuals originate from the counting process theory underlying the Cox model
- These residuals range in value from $-\infty$ to a maximum of 1, and each has an expected value of 0
- They represent the discrepancy between the observed value of a subject's failure indicator and its expected value, integrated over the time for which that patient was at risk
- Positive values mean that the patient died sooner than expected (according to the model); negative values mean that the patient lived longer than expected (or were censored)

Unlike traditional residuals, the sum of squares of martingale residuals cannot be used as a measure of goodness of fit. Instead, the **deviance residual** does have this property and can be defined in terms of the martingale residual:

$$d_i = \text{sign}(m_i)\{-2[m_i + \delta_i \log(\delta_i - m_i)]\}^{1/2}$$

- Symmetric version of martingale residuals
- If the fitted model is correct, these residuals are symmetrically distributed with expected value 0

Checking the Proportional Hazards Assumption

The proportional hazards assumption is key to the construction of the partial likelihood, since it is this property that allows one to cancel out the baseline hazard function from the partial likelihood factors. If one has a binary predictor variable, such as experimental vs. standard treatment, what this assumption means is that the hazard functions are proportional, and hence that the log-hazards are separated by a constant at all time points. Similarly, a categorical variable with many levels will result in parallel log hazard functions. This assumption is at best an approximation in practice, and minor violations are unlikely to have major effects on inferences on model parameters. For this reason, formal hypothesis tests of the proportional hazards assumption are often of limited value. Still, it is useful to assess, in a particular data set, if this assumption is reasonable, and what one can do if it is not. Here we will examine some commonly used assessment methods.

If we are comparing survival times between two groups, there is a simple plot that can help us assess the proportional hazards assumption. Let's say the two groups are coded as $Z = 1$ vs. $Z = 0$ and thus e^{β} is the hazard ratio comparing $Z = 1$ vs. $Z = 0$. Under the Cox model, we have

$$\begin{aligned} & \frac{\lambda_1(t)}{\lambda_0(t)} = e^{\beta} \\ \Rightarrow \quad & \log\{-\log\{S_1(t)\}\} = \beta + \log\{-\log\{S_0(t)\}\} \end{aligned}$$

The function $g(u) = \log\{-\log\{u\}\}$ is called a **complementary log-log** transformation. Based on these derivations, a plot of $g(S_1(t)$ and $g(S_0(t))$ versus t or $\log(t)$ should yield two parallel curves separated by a constant if the proportional hazards assumption is correct.

Scaled Schoenfeld residuals are another way to assess the proportional hazards assumption. Schoenfeld residuals are based on the score function.

When they are scaled by a function of the variance of $\widehat{\beta}$, the scaled residuals essentially provide an estimate of β as a function of t. If the proportional hazards assumption holds, then $\beta(t)$ as a function of t should be constant. Thus, one can formally test the proportional hazards assumption by testing whether the slope of $\beta(t)$ is 0.

Options when the proportional hazards assumption does not hold:

- Piecewise constant baseline hazard function
- Interaction with t or $\log(t)$ (now a time-varying covariate)
- Basis methods e.g. cubic spline
- Nonparametric kernel smoothing e.g. smoothed covariate
- Stratify the baseline hazard

Modeling Decisions

When deciding upon a model in regression with survival data, many of the same principles and tools that are used in linear regression can be used. For example, variable selection methods such as stepwise regression or regularized regression, machine learning methods such as trees and boosted models, and model fit criteria, which are discussed below.

When comparing two **nested models**, we can use the (partial) likelihood ratio test. Let Model A denote the larger model with degrees of freedom a and log likelihood l_a and let Model B denote the smaller nested model with degrees of freedom b and log likelihood l_b. Since

$$2(l_a - l_b) \sim \chi^2_{a-b},$$

we can compare $2(l_a - l_b)$ to the χ^2_{a-b} to obtain a p-value, p. If $p < 0.05$, then Model A is the superior model.

When comparing two **non-nested models**, we can use the **Akaike Information Criterion**, or AIC. For Model A, this quantity is $AIC = -2l_a + 2a$. The value of the AIC balances two quantities which are properties of a model. The first is goodness of fit, $-2l_a$, which is smaller for models that fit the data well. The second quantity, the number of parameters, is a measure of complexity. This enters the AIC as a penalty term. Thus, a "good" model is one that fits the data well (small value of $-2l_a$) with few parameters ($2a$), so that smaller values of AIC should in theory indicate better models.

An alternative to the AIC is the **Bayesian Information Criterion**, $BIC = -2l_a + a\log(n)$. The key difference is that the BIC penalizes the number of parameters by a factor of $\log(n)$ rather than by a factor of 2 as in the AIC. As a result, using the BIC in model selection will tend to result in models with fewer parameters as compared to AIC.

1.8 Prediction

Prediction is very difficult, especially if it's about the future.

-Niels Bohr

I shall go further and say that even if an examination of the past could lead to any valid prediction concerning man's future, that prediction would be the contrary of reassuring.

-Julien Benda

Estimation involves finding the optimal value of a parameter using historical data i.e., the data have already been collected. **Prediction** uses historical data to compute the random value of future unseen data. Forecasting lives within prediction and implies a temporal aspect for example, with time series data.

In the previous sections we have examined how to estimate overall survival and how to calculate predictions for survival for individuals. Here, we will focus on how to evaluate individual-level predictions.

1.8.1 Evaluating Predictions

First, consider a simple regression framework where we predict some outcome Y using a model or method and denote it as $\widehat{Y}$. Generally, the prediction accuracy is quantified by deciding upon a **"loss" function**:

$$\frac{1}{n}\sum_{i=1}^{n}(Y-\widehat{Y})^2$$

is the mean squared error (L2 loss), while

$$\frac{1}{n}\sum_{i=1}^{n}|Y-\widehat{Y}|$$

is the mean absolute error (L1 loss). For both of these, smaller values imply more accurate predictions.

Now consider a binary outcome, $Y = 0$ or 1. Let's say the "prediction" is also in binary form, $\widehat{Y} = 0$ or 1. A typical example of this setting is **diagnostic testing** where Y is whether has a disease and $\widehat{Y}$ is whether a diagnostic test indicates that someone has a disease.

To be specific, let's consider a COVID test. Define the following:

- D^+ means that an individual is "disease positive" i.e., has COVID
- D^- means that an individual is "disease negative" i.e., does not have COVID
- T^+ means that a COVID test was positive i.e., the test indicates the person has COVID
- T^- mean that a COVID test was negative i.e, the test indicates the person does not have COVID

If a test is T^+ and the person is D^+, then the classification is correct; this is a **true positive**. Similarly, if a test is T^- and the person is D^-, then the classification is correct; this is **true negative**. Any other combination results in a mis-classification or error. If a test is T^+ and the person is D^-, this is a **false positive**. If a test is T^- and the person is D^+, this is a **false negative**.

In this binary case, the L1 and L2 loss functions above are equivalent and they quantify the overall mis-classification rate where again, smaller values imply more accurate predictions. However, it is often the case that the "costs" of false negative vs. false positive are different.

For example, a false positive COVID test may result in a self-quarantine for X number of days, with possible loss in wages. A false negative COVID test may result in the person continuing to socialize, spreading COVID, making others sick, and possibly result in hospitalization or death of the person or someone else (in an extreme case).

As another example, in breast cancer screening, a false positive may result in a subsequent unnecessary biopsy. A false negative may result in delayed treatment, metastasis, and death that could have been prevented with earlier treatment.

In such cases, you would want to know the error rate for these two types of errors *separately*.

	Disease, D^+	Disease, D^-	
Test, T^+	a	b	a+b
Test, T^-	c	d	c+d
	a+c	b+d	

The following quantities achieve this goal, using the table above,

- The **sensitivity** is the probability that a test will be T^+ given that a person is D^+: $P(T^+|D^+)$, estimated as $\frac{a}{a+c}$
- The **specificity** is the probability that a test will be T^- given that a person is D^-: $P(T^-|D^-)$ estimated as $\frac{d}{b+d}$
- The **positive predictive value (PPV)** is the probability that a person is D^+ given the test is T^+: $P(D^+|T^+)$, estimated as $\frac{a}{a+b}$

- The **negative predictive value (NPV)** is the probability that a person is D^- given the test is T^-: $P(D^-|T^-)$, estimated as $\frac{d}{c+d}$
- The **false positive rate (FPR)** is the probability that a person will be T^+ given that a person is D^-: $P(T^+|D^-)$, estimated as $\frac{b}{b+d}$
- The **false negative rate (FNR)** is the probability that a person will be T^- given that a person is D^+: $P(T^-|D^+)$, estimated as $\frac{c}{a+c}$

The sensitivity and specificity are characteristics of the test. Note that sensitivity + FNR = 1 and specificity + FPR = 1.

PPV and NPV for a particular type of test depend upon the prevalence of a disease in a population.

Class exercise

For the table below, calculate the estimated sensitivity, specificity, PPV, NPV, FPR, and FNR.

	Disease, D^+	Disease, D^-
Test, T^+	94	7
Test, T^-	15	89

In most settings, the test is not truly positive or negative. Instead, there is often some **threshold** θ that is used to define a positive vs. negative test. Let R **be a measurement obtained from the test**, for example number of proteins, number of antibodies, a composite score from a psychological test, or in the case of a COVID test, the ratio between test and control bands on the lateral immunochromatography test.

The threshold θ is used as follows (note that the inequalities can easily be flipped):

- If $R \leq \theta$, then the test is T^+
- If $R > \theta$, then the test is T^-

Now, there is a potential to estimate each of the six quantities above at every possible θ.

For example, suppose R is a score from a cognitive functioning test, which ranges from 0 to 100, and is used to determine whether someone has mild cognitive impairment.

- If $R \leq \theta$, then T^+ i.e., the test indicates that the person has mild cognitive impairment.

- If $R > \theta$, then T^- i.e., the test indicates that the person does not have mild cognitive impairment.
- Here, D^+ means the person truly has mild cognitive impairment (determined by brain image) and D^- means they do not.

Focusing on sensitivity and specificity for now, with this threshold θ:

- **Sensitivity** $= P(R \leq \theta | D^+)$
- **Specificity** $= P(R > \theta | D^-)$

You can essentially make the same 2 x 2 table and calculate sensitivity and specificity as above:

	Disease, D^+	Disease, D^-	
Test, $R \leq \theta$	a	b	a+b
Test, $R > \theta$	c	d	c+d
	a+c	b+d	

In practice, you might be considering different possible θ values. Each θ value can result in different sensitivity and specificity estimates.

Class exercise

The table below shows the true disease status and score result, R, where each row is one person.

1. Fill in the last two columns of the table where each cell is either T^+ or T^- and is determine by the θ value. That is, if $R \leq \theta$ then the cell is T^+; the first row is completed as an example.
2. For each θ, construct the 2 x 2 table.
3. For each θ, calculate sensitivity and specificity.
4. Recall that R ranges from 0 to 100. Calculate the sensitivity and specificity when $\theta = 0$ and when $\theta = 100$.

Disease	R	Test, $\theta = 10$	Test, $\theta = 50$
D^+	2	T^+	T^+
D^+	5		
D^+	7		
D^+	9		
D^+	10		
D^+	7		
D^+	4		
D^+	15		
D^+	35		
D^+	51		
D^-	9		
D^-	7		
D^-	12		
D^-	26		
D^-	49		
D^-	87		
D^-	99		
D^-	60		

1.8.2 The ROC Curve

It is rare to have a test that achieves both high sensitivity and high specificity.

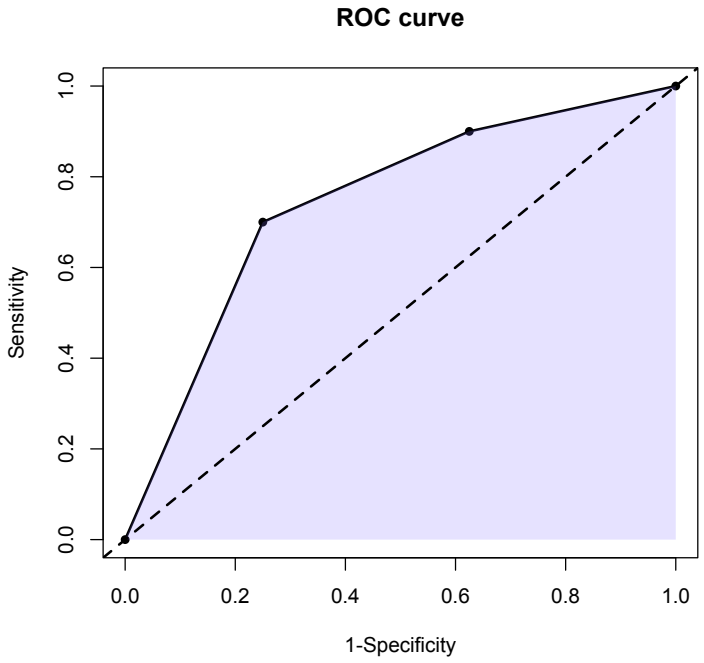
It is often the case that one must strike a balance.

The **receiver operating characteristic (ROC) curve** allows us to visually examine this tradeoff.

This curve is constructed by plotting 1-specificity vs. sensitivity for each potential value of θ. A diagonal line through the origin with slope 1 reflects a "null" or useless test which would be equivalent to using a coin flip to classify someone as T^+ or T^-.

For our data above, the ROC curve would look as shown in Figure 1.4. Some argue that the "optimal" θ should be selected as the θ that lies closest to the upper left of the curve. In truth, the optimal value depends heavily on the costs of each type of error.

Figure 1.4: ROC curve for example



The **area under the ROC curve (AUC)** is shaded in Figure 1.4.

- The AUC provides an overall single number summary of the accuracy of the test across all possible θ values.
- It quantifies the degree or measure of separability and the extent to which the test is capable of distinguishing between groups (D^+ vs. D^-).
- An AUC of 1 is perfect; higher AUC values indicate better performance.
- The AUC for the diagonal line is 0.5; this reflects a useless test (equivalent to a coin flip).
- The AUC is equivalent to the C-statistic.

Now let's move on to consider a setting where you do not necessarily have a diagnostic test, but instead, you have a **model that estimates the probability of D^+**.

For example, we have data where we have fit a logistic regression model predicting D^+ from some set of covariates, $\mathbf{Z}$, where the goal is to estimate $p = P(D^+|\mathbf{Z})$. We can use this model to estimate $p_i = P(D^+|\mathbf{Z}_i)$ for each person.

The process for constructing an ROC curve and calculating the AUC for this model is exactly the same as above, except that here the potential values of θ are probability cutoffs.

$\Rightarrow$ For example, if $\theta = 0.5$, this means that person i is classified as T^+ if $p_i \geq 0.5$ and as T^- if $p_i < 0.5$. We can technically consider any $\theta \in [0, 1]$.

Formally, the ROC is defined as

$$\text{ROC}(u) = \text{Sens}\{\text{Spec}^{-1}(1-u)\}$$

where $\text{Spec}(\theta) = P(p < \theta|D^-)$ and $\text{Sens}(\theta) = P(p \geq \theta|D^+)$, and

$$\text{AUC} = \int \text{ROC}(u)du.$$

To discuss estimation, we will use the following notation:

- m is the number of D^+ individuals
- n be the number of D^- individuals
- $N = m + n$
- $X_i, i = 1, 2, ...m$ is the value of the variable that the test is based on (e.g., p in the probability model case) among the D^+ individuals
- $Y_j, j = 1, 2, ..., n$ is this value among the D^- individuals

By construction, higher p indicates a higher likelihood of being D^+.

Sensitivity and specificity are estimated as:

$$\widehat{\text{sens}}(\theta) = \frac{1}{m}\sum_{i=1}^{m} I(X_i \geq \theta)$$

$$\widehat{\text{spec}}(\theta) = \frac{1}{n}\sum_{j=1}^{n} I(Y_j < \theta)$$

where $I(c)$ is the indicator function which is 1 if c is true, and 0 otherwise.

The AUC can be expressed as:

$$AUC = P(X > Y)$$

This expression shows that **the AUC has a useful interpretation.** It can be interpreted as the probability that a randomly chosen D^+ subject is ranked as more likely to be diseased than a randomly chosen D^- subject. This interpretation is based on the nonparametric Mann-Whitney U-statistic that is used in calculating AUC.

AUC can be estimated nonparametrically as:

$$\widehat{AUC} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} I(X_i > Y_j)$$

If there are ties, then this is more complicated:

$$\widehat{AUC} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} \psi(X_i, Y_j),$$

$$\text{where} \quad \psi(X,Y) = \begin{cases} 1, & \text{if} \quad X > Y \\ \frac{1}{2}, & \text{if} \quad X = Y \\ 0, & \text{if} \quad X < Y. \end{cases}$$

Note that the estimate involves assessing each pairwise combination of X and Y.

There are several alternative ways to estimate the AUC including parametric models. For example, the binormal approach assumes that X and Y are both normally distributed. Here, we will focus on the nonparametric empirical estimate defined above.

Another measure of prediction accuracy is the **Brier score**. For this definition only, let A be the disease status where $A = 1$ if D^+ and $A = 0$ for D^- and let $p = P(D^+|\mathbf{Z})$. The Brier Score is defined as:

$$BS = E\{(p_i - A_i)^2\}$$

and can be estimated as

$$\widehat{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - A_i)^2.$$

The Brier Score parallels the L2 loss described previously; it captures the mean squared error.

Net reclassification improvement: If you are classifying people into more than 2 categories, you can consider using the net reclassification improvement (NRI) to look at incremental value.

- The NRI quantifies how well a new model (with the new predictor) reclassifies subjects - either appropriately or inappropriately - as compared to an old model (without the new predictor).
- It can be used for two categories and is not exactly the same as the AUC.
- Some have argued it should be used for two categories but in practice, it is not used as much as the AUC.
- For more details, see: Pencina, M. J., D'Agostino Sr, R. B., D'Agostino Jr, R. B., & Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Statistics in medicine, 27(2), 157-172.

1.8.3 Incremental Value

The **incremental value of a predictor** can be quantified by looking at the difference in the AUC for a model that includes the predictor vs. a model that does **not** include the predictor.

- Let Z_b be the predictor of interest.
- Let $p(Z_b)$ denote a prediction p that uses Z_b information, while $p(-Z_b)$ denotes a prediction that does not use Z_b information.
- For example, $p(Z_b)$ can be obtained from a logistic regression model using all available predictors while $p(-Z_b)$ can be obtained from a logistic regression model that uses all available predictors except for Z_b.
- The incremental value of Z_b in terms of prediction accuracy is:

$$IV(Z_b) = AUC(p(Z_b)) - AUC(p(-Z_b))$$

  where $AUC(p)$ is the AUC of the predictions p.
- The quantity $IV(Z_b)$ can be estimated using the estimated corresponding AUCs.

Why would we care about the incremental value of a predictor? What do we even mean by incremental value?

Generally, Z_b is something that is **costly or invasive** to obtain, often a biomarker of some kind. Therefore, we want to know whether it is worth going through trouble of measuring it. If Z_b is not useful when added to predictors you already have (i.e. incremental value), then it is likely not worth the cost and/or burden to obtain it.

Examples of "high cost" predictors include:

- Genetic score
- Stool DNA test
- Breast biopsy
- Data from wearables/devices

Examples of "low cost" predictors include:

- Age
- Gender
- Race/ethnicity
- Employment status
- Routinely obtained lab measurements e.g. blood pressure, cholesterol

Importantly, it is possible for Z_b to be **significantly associated** with the outcome in, for example, a logistic regression model, but **not** have much incremental value. For more, see: Pepe, M. S., Janes, H., Longton, G., Leisenring, W., & Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. American Journal of Epidemiology, 159(9), 882-890.

If you are trying to make the case that a new predictor is useful, you must show that it has incremental value.

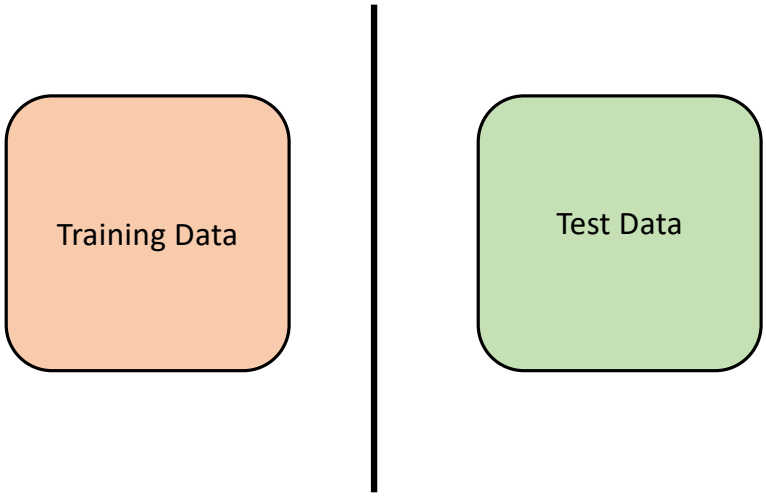
1.8.4 Cross-Validation

[Reference: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. New York: Springer.]

So far, we have been discussing building a **prediction** model and **evaluating** the model within the same dataset, the dataset that you have.

In truth, that is inappropriate.

- You should not evaluate a prediction tool using the same data you used to build the prediction tool.
- Essentially, this is kind of like "cheating". Of course your prediction tool is going to perform reasonably well in these data, because these are the data you used to make the tool in the first place.



- Ideally, you will build your prediction tool/model using one dataset - we refer to this as the **training data**, and then you evaluate your prediction tool in a totally separate dataset - we refer to this as the **validation data** or **test data**.

Let the **statistical learning method** refer to the model/method/tool you used to generate predictions.

The **test error rate** is the average error that results from using a statistical learning method to predict the response on a new observation— that is, a measurement that was not used in training the method.

The use of a particular statistical learning method is warranted if it results in a low test error. This is the error rate that you are truly interested in.

The **training error rate** can be calculated by applying the statistical learning method to the observations used in its training. The training error rate often is quite different from the test error rate, and in particular it can **dramatically underestimate** the test error rate.

Usually, we don't have the luxury of having two separate datasets. Instead, we will consider a class of methods that estimate the test error rate by **holding out a subset** of the training observations from the fitting process, and then applying the statistical learning method to those held out observations.

Single holdout procedure:

- We can randomly divide the available dataset into two parts, a training set and a test set (see Figure 1.5.
- The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the test set.

- The resulting test set error rate or performance assessment —for example as quantified by the AUC —provides an estimate of the test error rate.

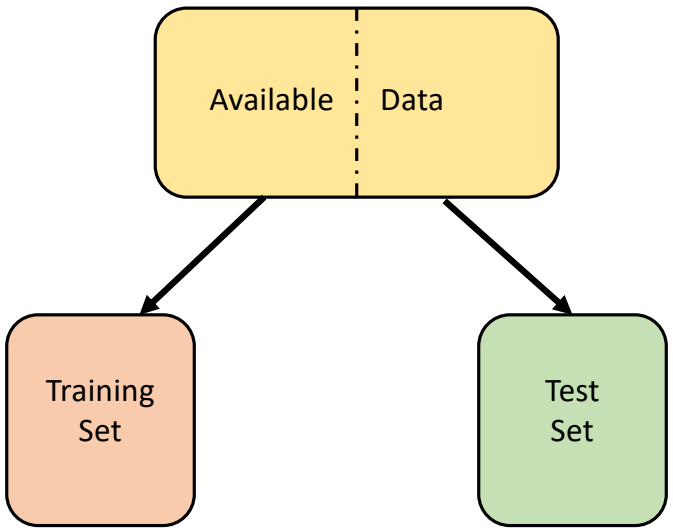


Figure 1.5: Single Split

This is conceptually simple and easy to implement, but there are two drawbacks:

1. The estimate of the test set error rate will be **highly variable**. If we do this once and get an estimate, and then do it again, there is no guarantee that those estimates will be close.

2. With this approach, you only use a **subset** of the data to train the model. It is known that statistical methods tend to perform worse when they are trained on less data. Thus, your estimate may actually over-estimate the test set error rate!

Cross-validation addresses these two issues.

The **Leave-one-out cross-validation** (LOOCV) procedure is as follows:

- Split the available data into two parts, just as we did above. However, instead of creating two subsets of comparable size, **a single observation** (Z_1, Y_1) is used for the validation set, and the remaining observations make up the training set.

- The statistical learning method is fit on the $N-1$ training observations, and a prediction $\widehat{Y}_1$ is made for the excluded observation using its Z_1 value. Since (Z_1, Y_1) was not used in the fitting process, the error rate based on comparing Y_1 vs. $\widehat{Y}_1$ provides an approximately unbiased estimate for the test error. Let's denote this error rate as TE_1, which could be the AUC or the L1 or L2, etc.

- But even though it is unbiased for the test error, it is actually a poor estimate because it is highly variable, since it is based upon a single observation (Z_1, Y_1).
- We can **repeat the procedure** by selecting (Z_2, Y_2) for the test set, training the statistical learning procedure on the now remaining $N-1$ observations and comparing Y_2 vs. $\widehat{Y}_2$ (see Figure 1.6). Repeating this approach N times produces N test error rate estimates, $TE_1, TE_2, ..., TE_N$.
- The LOOCV estimate for the test error rate is the **average of these N test error estimates**:

$$\frac{1}{N}\sum_{i=1}^{N} TE_i$$

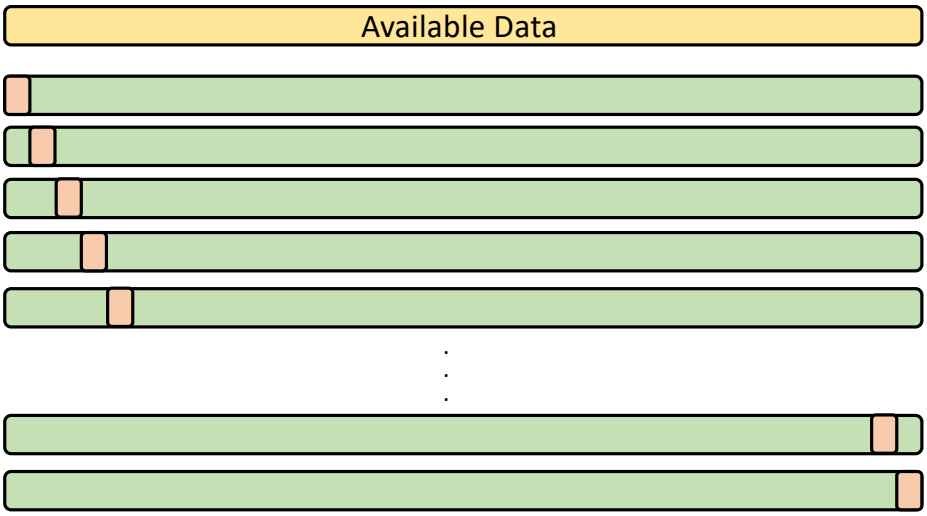


Figure 1.6: Leave-one-out Cross-validation

LOOCV has a couple of major advantages over the single split approach.

First, it has far **less bias**. In LOOCV, we repeatedly fit the statistical learning method using training sets that contain $N-1$ observations, almost as many as are in the entire data set. This is in contrast to the single split approach, in which the training set is typically around half the size of the original data set. Consequently, the LOOCV approach tends not to overestimate the test error rate as much as the single split approach does.

Second, in contrast to the single split approach which will yield different results when applied repeatedly due to randomness in the training/test set splits, performing LOOCV multiple times will **always yield the same results**: there is no randomness in the training/test set splits.

However, LOOCV can be **expensive to implement**, computationally, since the model has to be fit N times. This can be very time consuming if N is large, and if each individual model is slow to fit.

LOOCV is a very general method, and can be used with any kind of predictive modeling.

k-fold Cross-validation

- An alternative to LOOCV is k-fold CV. This approach involves randomly dividing the set of observations into k groups, or folds, of approximately equal size. The first fold is treated as a test set, and the method is fit on the remaining $k-1$ folds.
- The test error rate is then computed on the observations in the held-out fold; let's denote this as TE_1. This procedure is repeated k times; each time, a different group of observations is treated as a test set (see Figure 1.7.
- This process results in k estimates of the test error, $TE_1, TE_2, ..., TE_k$. The k-fold CV estimate is computed by averaging these values,

$$\frac{1}{k}\sum_{j=1}^{k} TE_j$$

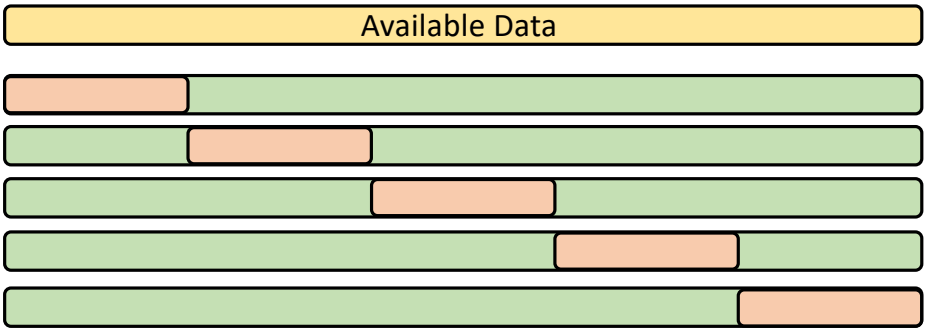


Figure 1.7: 5-fold Cross-validation

LOOCV is a special case of k-fold CV in which k is set to equal N. In practice, one typically performs k-fold CV using $k = 5$ or $k = 10$.

What is the advantage of using $k = 5$ or $k = 10$ rather than $k = N$? The most obvious advantage is **computational**. LOOCV requires fitting the statistical learning method N times. This has the potential to be computationally expensive. But cross-validation is a very general approach that can be applied to almost any statistical learning method. Some statistical learning methods have computationally intensive fitting procedures, and so performing LOOCV may pose computational problems, especially if N is extremely large.

In contrast, performing 10-fold CV requires fitting the learning procedure only ten times, which may be much more feasible.

Generalized Cross-validation

- Generalized cross-validation is a general form of cross-validation. As with the single split approach, this approach involves randomly dividing the set of observations into two groups, the training set and the test set. These two groups do not necessary have to be of similar size.
- Let h reflect the proportion of the data chosen as training data i.e., if $h = 0.5$, then half the data is randomly selected as training data, and half is selected as test data. The method is fit on the randomly selected training data.
- The test error rate is then computed on the test data; let's denote this as TE_1. This procedure is repeated M times where M is large such as $M = 100$ or $M = 500$. For each of the M times, a different random subset is used as the test data, but these subsets are not mutually exclusive.
- This process results in M estimates of the test error, $TE_1, TE_2, ..., TE_M$. The generalized CV estimate is computed by averaging these values,

$$\frac{1}{M}\sum_{m=1}^{M} TE_m$$

Generalized cross-validation offers the benefit of optimizing the amount of data used for training vs. test data and will generally result in less bias than k-fold CV. However, it is computationally more expensive than k-fold CV, though usually less expensive than LOOCV.

When we examine real data, we do not know the true test MSE, and so it is difficult to determine the accuracy of the cross-validation estimate. However, if we examine simulated data, then we can compute the true test error, and can thereby evaluate the accuracy of our cross-validation results.

When we perform cross-validation, our goal might be to determine how well a given statistical learning procedure can be expected to perform on independent data; in this case, the actual estimate of the test MSE is of interest. But at other times we are interested only in the **location of the minimum point** in the estimated test error curve. This is because we might be performing cross-validation on a number of statistical learning methods, or on a single method using different levels of flexibility, in order to identify the method that **results in the lowest test error**. For this purpose, the location of the minimum point in the estimated test error curve is important, but the actual value of the estimated test error is not.

Here, we have focused on a setting where the outcome is binary. But **cross-validation is much more general** and can be used when the outcome is continuous, or when the outcome

is qualitative e.g., categories of some kind. In this latter case, the test error rate is some measure of the mis-classification rate.

A Note on Variance Estimation

Sometimes it is of interest to report a standard error for the estimated AUC and/or give a confidence interval for the AUC.

- Variance estimation and asymptotic properties that allow for construction of a confidence interval have been proposed in: DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 837-845.
- The `ci.auc` function of the `pROC` package can calculate a confidence interval for the AUC using either the Delong et al. 1988 approach or bootstrapping.
- However, this CI calculation does not account for the fact that p is, in fact, an estimate itself, and has variability. To **account for the variation in** p as well, you must bootstrap the entire procedure of (a) building the model and (2) evaluating the model.
- This is not implemented in `ci.auc` and must be coded separately.

1.8.5 Prediction of Time-to-event Outcomes

Finally, we can return to time-to-event outcomes. We now want to consider how we take a statistical learning model, the Cox proportional hazards model, and evaluate the resulting predictions when the outcome is not 0/1, but is instead a time-to-event outcome.

Now, p **is the probability that the event occurred before** t **i.e.** $p = P(T \leq t)$ and our estimate, $\widehat{p}$ for an individual i comes from our Cox proportional hazards model (or an alternative survival model).

This is achieved using **time dependent** versions of sensitivity, specificity, and the AUC.

Recall from earlier that the (not time-dependent versions of the) sensitivity, specificity, ROC, and AUC are defined as

$$\begin{aligned} \text{Spec}(\theta) &= P(p < \theta | D^-) \\ \text{Sens}(\theta) &= P(p \geq \theta | D^+) \\ \text{ROC}(u) &= \text{Sens}\{\text{Spec}^{-1}(1-u)\} \\ \text{AUC} &= \int \text{ROC}(u) du \end{aligned}$$

For the time-dependent versions, there are three types:

- **Cumulative/dynamic definitions**

$$\begin{aligned} \text{Spec}(\theta, t) &= P(p < \theta | T > t) \\ \text{Sens}(\theta, t) &= P(p \geq \theta | T \leq t) \end{aligned}$$

where T is the time of the event.

Using these definitions, at any fixed time t the entire population is classified as either a case or a control on the basis of their status at time t. Also, each individual i plays the role of a control for times $t < T_i$, but then contributes as a case for later times, $t \geq T_i$.

In Figure 1.8, individuals colored in green - person D, F, G - are considered cases or D^+. The individuals in pink are considered controls or D^-.

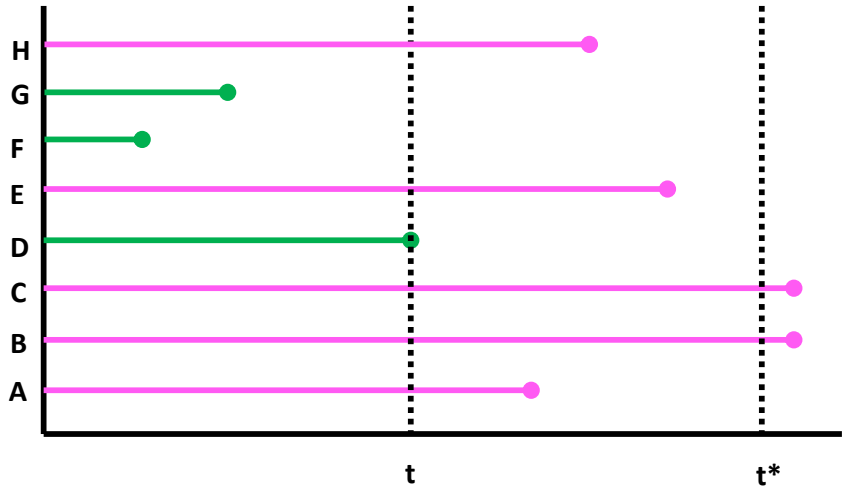


Figure 1.8: Cumulative/dynamic classifications

Cumulative/dynamic accuracy summaries are most appropriate when a specific time t (or a collection of times) is important and scientific interest lies in discriminating between subjects who die prior to a given time t and those that survive beyond t.

The ROC and AUC are

$$\begin{aligned} \text{ROC}(u, t) &= \text{Sens}\{\text{Spec}^{-1}(1-u), t\} \\ \text{AUC}(t) &= \int \text{ROC}(u, t) du \end{aligned}$$

- **Incident/dynamic definitions**

$$\begin{aligned} \text{Spec}^{ID}(\theta, t) &= P(p < \theta | T > t) \\ \text{Sens}^{ID}(\theta, t) &= P(p \geq \theta | T = t) \end{aligned}$$

where ID indicates this is the incident/dynamic definition.

Using this approach a subject can play the role of a control for an early time, $t < T_i$, but then play the role of case when $t = T_i$. This dynamic status parallels the multiple contributions that a subject can make to the partial likelihood function.

In Figure 1.8, with this definition, only person D is considered a case or D^+. The individuals in pink are considered controls or D^-.

Here sensitivity measures the expected fraction of subjects with a p greater than or equal to θ among the subpopulation of individuals who die at time t, while specificity measures the fraction of subjects with p less than θ among those who survive beyond time t.

Incident sensitivity and dynamic specificity are defined by dichotomizing the risk set at time t into those observed to die (cases) and those observed to survive (controls).

- **Incident/static definitions**

$$\begin{aligned} \text{Spec}^{IS}(\theta, t) &= P(p < \theta | T > t^*) \\ \text{Sens}^{IS}(\theta, t) &= P(p \geq \theta | T = t) \end{aligned}$$

where IS indicates this is the incident/dynamic definition and t^* is some pre-defined time point, usually long-term.

Using this definition, each subject does not change disease status and is treated as either a case or a control. Cases are stratified according to the time at which the event occurs (incident) and controls are defined as those subjects who are event free through a fixed follow-up period, $(0, t^*)$ (static).

In Figure 1.8, with this definition, only person D is considered a case or D^+ and only individuals B and C are considered controls or D^-.

We will use the cumulative/dynamic definitions. It is often argued that this definition has more clinical relevance.

Referring back to our discussion about how the AUC has a nice interpretation, it is also the case here. Recall that

- $X_i, i = 1, 2, ...m$ is the value of p among the D^+ individuals (cases)
- $Y_j, j = 1, 2, ..., n$ is this value of p among the D^- individuals (controls)

The AUC now can be expressed as $P(X_i > Y_j)$ or:

$$AUC(t) = P(p_i > p_j | T_i \leq t, T_j > t), i \neq j$$

That is, it can be interpreted as the probability that given two randomly chosen subjects, one who died before t and one who died after t, the value p will correctly rank the one who died as higher risk than the one who did not.

Estimation for Time-to-event Outcomes

If there was no censoring before t, one could estimate these quantities just as we did previously, with empirical estimates. That is:

$$\widehat{\text{sens}}(\theta, t) = \frac{\sum_{i=1}^{N} I(p_i \geq \theta) I(T_i \leq t)}{\sum_{i=1}^{N} I(T_i \leq t)}$$

$$\widehat{\text{spec}}(\theta, t) = \frac{\sum_{i=1}^{N} I(p_i < \theta) I(T_i > t)}{\sum_{i=1}^{N} I(T_i > t)}$$

$$\widehat{AUC}(t) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} I(p_i > p_j) I(T_i \leq t) I(T_j > t)}{\sum_{i=1}^{N} \sum_{j=1}^{N} I(T_i \leq t) I(T_j > t)}$$

However, in practice, there is often censoring before t. This makes it difficult to classify people as cases vs. controls because we need to know whether $T \leq t$ or $T > t$. Instead of observing T, we instead observe X.

Importantly, we know some information:

- If someone has $X_i > t$, we know that $T_i > t$.
- If someone has $X_i \leq t$ and $\delta_i = 1$, then we know that $T_i \leq t$.
- It is only for the people with $X_i \leq t$ and $\delta_i = 0$ that we are not sure.

We don't want to ignore these latter individuals. At best, ignoring them would be losing information. At worst, ignoring them would induce bias.

There are many ways to estimate these quantities in the presence of censoring. We will focus on the **inverse probability of censoring weighting (IPCW)** approach.

IPCW estimators correct for censored subjects by giving extra weight to subjects who are not censored. Each subject i is weighted by the inverse of an estimate of the conditional probability of having remained uncensored until time t.

The IPCW estimates are:

$$\widehat{\text{sens}}(\theta, t) = \frac{\sum_{i=1}^{N} \widehat{W}_i I(p_i \geq \theta) I(X_i \leq t)}{\sum_{i=1}^{N} \widehat{W}_i I(X_i \leq t)}$$

$$\widehat{\text{spec}}(\theta, t) = \frac{\sum_{i=1}^{N} \widehat{W}_i I(p_i < \theta) I(X_i > t)}{\sum_{i=1}^{N} \widehat{W}_i I(X_i > t)}$$

$$\widehat{AUC}(t) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} \widehat{W}_i \widehat{W}_j I(p_i > p_j) I(X_i \leq t) I(X_j > t)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \widehat{W}_i \widehat{W}_j I(X_i \leq t) I(X_j > t)}$$

where

$$\widehat{W}_i = \frac{I(X_i > t)}{\widehat{G}(t)} + \frac{I(X_i \leq t)\delta_i}{\widehat{G}(X_i)}$$

where $\widehat{G}(\cdot)$ is the Kaplan Meier estimator of $G(t) = P(C > t)$.

Let's break this down.

- The estimates may look overwhelming but we simply took the empirical estimates above and 1) added a weight $\widehat{W}_i$ for each person within each sum and 2) replaced the T's with X's, since we observe X.
- **What is $G(t)$?** G(t) is the survival distribution for censoring. This may see strange, but it's essentially saying - what if what I cared about was the time until someone was censored? I would characterize the distribution of that censoring time as $G(t) = P(C > t)$. We can estimate $G(\cdot)$ with the Kaplan-Meier estimator for time to censoring.
- **What is $\widehat{W}_i$ doing exactly?** For anyone with $X_i > t$, we know that $T_i > t$ and $C_i > t$. We give these people "extra" weight. The numerical value of that weight is $1/\widehat{G}(t)$. Informally, they get extra credit for making it past t without being censored.
- For anyone with $X_i \leq t$ and $\delta_i = 1$, they also get extra weight and the value is $1/\widehat{G}(X_i)$. We know that this person had the event at X_i and that their hypothetical censoring time would be after X_i. Notice then that $G(X_i)$ is the probability that they made it past X_i without being censored.

- For anyone with $X_i \leq t$ and $\delta_i = 0$, their $\widehat{W}_i = 0$ meaning they are not included in the estimate. But didn't I say earlier we should not ignore them? We aren't ignoring them because we use them to estimate $\widehat{G}(\cdot)$; we use everyone to estimate $\widehat{G}(\cdot)$.

All previous points on quantifying the incremental value of a variable and cross-validation apply in the time-to-event setting as well.

In some cases, an analytic variance has been derived for these estimates. In practice, bootstrapping or some type of resampling (wild bootstrap) is generally used to obtain confidence intervals.

References:

Heagerty, P. J., & Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1), 92-105.

Heagerty, P. J., Lumley, T., & Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2), 337-344.

Uno, H., Cai, T., Tian, L., & Wei, L. J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478), 527-537.

1.9 Time-varying Covariates

So far we have considered models for the hazard that are of the form:

$$\lambda(t|\mathbf{Z}) = \lambda_0(t)\exp\{\beta^T\mathbf{Z}\}$$

so that, implicitly, we have only considered covariates that are time-invariant i.e. **the components of Z are measured at baseline.**

In some settings, covariates of interest may **vary over time.**

Examples include:

- Pollution exposure
- Blood glucose
- Medication adherence
- Alcohol consumption
- Whether a patient has had a heart transplant (0/1); they may be a 0 at the beginning of the study, but then receive a transplant during the study

These are **time-varying or time-dependent** covariates.

We are going to use a dataset from a heart transplant study. The primary outcome was time to death and available covariates at baseline include age and prior bypass surgery. However, we also have the time of transplant, for those who received a transplant.

That is, at the start of the study, no one has had a transplant. During the study follow-up, some patients receive a transplant. Therefore, this is a variable whose value changes during follow-up. Table 1.1 shows some patients from the study.

- Person 1 had a transplant at 11 days and died at 57 days.
- Person 2 had a transplant at 25 days and was censored at 152 days.
- Person 3 died at 7 days and did not have a transplant.

We are interested in investigating the association between receiving a transplant and mortality. To do this, we need additional notation. We define the **time-varying covariate**:

$$Z_1(t) = \begin{cases} 0 & \text{if no transplant by time } t \\ 1 & \text{if transplant by time } t \end{cases}$$

Patient ID	Time	Status	Age	Prior	Transplant	Wait Time
1	57	1	42.50240	0	1	11
2	152	0	47.98084	0	1	25
3	7	1	53.19370	0	0	NA
4	80	1	54.57358	0	1	16
5	1386	0	54.01232	0	1	36
6	1	1	53.81520	1	0	NA
7	307	1	49.44832	0	1	27
8	35	1	20.33128	0	0	NA
9	42	0	56.84873	0	1	19
10	36	1	59.12389	0	0	NA
11	27	1	55.27995	0	1	17

Table 1.1: Patients from the heart transplant study

The value of the covariate starts at 0 and "jumps" to 1 when the patient received a transplant.

Using this definition, we can now write the model for the hazard as:

$$\lambda(t|\mathbf{Z}) = \lambda_0(t)\exp\{\beta_1 Z_1(t) + \beta_2 Z_2 + \beta_3 Z_3 + ...\}$$

where Z_2, Z_3, etc. are additional time-invariant covariates included in the model.

More generically we can write :

$$\lambda(t|\mathbf{Z}) = \lambda_0(t)\exp\{\beta^T \mathbf{Z}(t)\}$$

with the understanding that some of the components of $\mathbf{Z}(t)$ are time varying while others are time-invariant.

Notice that this is still a proportional hazards model:

- β does not depend on time
- the covariates effects are constant over time

The **interpretation** of covariate effects can follow the same rubric as the one we discussed for the standard Cox model.

Operationally, although incorporating time-varying covariates seems to result in a more complex model, estimation and inference can nevertheless proceed via the same partial likelihood as before.

Importantly, the values of the time-varying covariates are updated when they are being compared in the risk set:

$$L^{partial}(\beta) = \prod_{j=1}^{K} \frac{\exp\{\beta^T \mathbf{z}_j(t_{(j)})\}}{\sum_{l\in\mathcal{R}_j}\exp\{\beta^T \mathbf{z}_l(t_{(j)})\}}$$

- As a patient "moves" through time, each time their covariates are compared to those of the individual who experienced the event, we have to make sure that the value of $z_l(t_{(j)})$ is the "current" one.
- Operationally, this involves splitting person-time by creating multiple records for patients who receive a transplant.
- Let's take patient 1 above. Their time-varying transplant covariate can be represented as follows:

$$Z_1(t) = \begin{cases} 0 & t < 11 \\ 1 & t \geq 11 \end{cases}$$

- Consequently, for this patient, we are going to split their record into two parts to reflect person-time when they did not have a transplant and person-time after they had a transplant.

Patient ID	Time	Status	Age	Prior	Transplant	Start	End
1	57	1	42.50240	0	0	0	11
1	57	1	42.50240	0	1	11	57

- Note that all time invariant covariates stay the same.
- Once you have the dataset in this form, you can use standard software to fit the model without any further modification.

One common question with this data setup is whether **we need to worry about correlated data**, since a given subject has multiple observations. The answer is no, we do not. The reason is that this representation is simply a programming trick. The likelihood equations at any time point use only one copy of any subject, the estimation procedure uses the correct row of data at each time according to the current t.

There two exceptions to this rule:

- When subjects have multiple events, then the rows for the events are correlated within subject and a cluster variance is needed.
- When a subject appears in overlapping intervals. This however is almost always a data error, since it corresponds to two copies of the subject being present in the same strata at the same time.

Once you have an estimated $\widehat{\beta}$ for the time-varying covariate, you can interpret it just as you did before.

For example, let's say $e^{\widehat{\beta}}$ is greater than 1. We could say: the hazard for mortality for patients who received a transplant is estimated to be $e^{\widehat{\beta}}$ times higher than the hazard for individuals who did not receive a transplant, holding all other covariates constant.

Most importantly, PREDICTION IS BROKEN.

The rule is clear: **we cannot predict survival using covariate values from the future.** You must be very careful.

Results from this heart transplant study were published in the Annals of Internal Medicine in 1971.

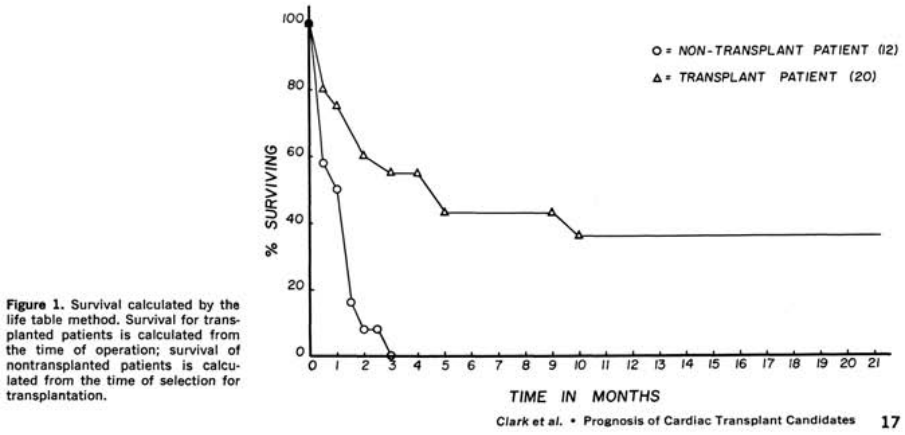


Figure 1.9: Survival curves from Clark et al. (1971)

They concluded that patients who received heart transplants lived significantly longer than those who did not. They essentially plotted the Kaplan-Meier estimate of survival among those who received a transplant versus those who did not which appeared to indicate that transplants are extremely effective in increasing the lifespan of the recipients (see Figure 1.9).

The problem here is that receipt of a transplant is a time-dependent covariate; patients who received a transplant **had to live long enough to receive that transplant.** Essentially, the analysis only shows that patients who live longer (i.e. long enough to receive a transplant) have longer lives than patients who don't live as long, which of course is a tautology.

There were **two main problems** with this analysis:

- First, the transplant variable was **not modeled correctly.** It is not a binary covariates available at baseline. It must be coded as a time-varying covariate and modeled accordingly.
- Second, while you **can** talk about the hazard, you **cannot** talk about the survival because this is equivalent to prediction of the future using data from the future.

For the first problem, you now know how to appropriately code a time-varying covariate (described above) and interpret the resulting hazard.

For the second problem, if your goal is prediction and/or to make a statement about "survival," you can alternatively consider **landmark prediction or dynamic prediction.**

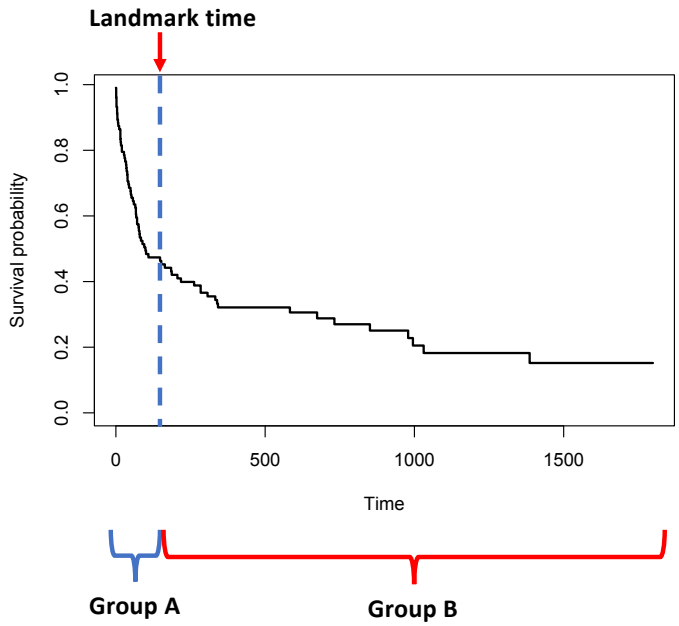


Figure 1.10: Landmark Prediction, split patients

- Landmark prediction involves specifying a "landmark time" denoted by t_0 which divides patients into two groups: Group A is those that have the event (e.g. die) before t_0 i.e., $T \leq t_0$ and Group B is those that survive past t_0 without experiencing the event, i.e. $T > t_0$.
- Modeling or building your prediction tool is done only among Group B. In effect, you move the baseline to t_0 and predict survival after t_0. Because of this, you can now consider the time-varying covariate information up to t_0 as essentially, a baseline covariate.
- For example, for the heart transplant data, we can set a landmark at 30 days.
  - We first select those patients who lived at least 30 days: 79 of the 103 patients lived this long.
  - Of these 79 patients, 33 had a transplant within 30 days, and 46 did not.
  - Of these 46, 30 subsequently had a heart transplant **after 30 days**, but we still count them in the "no transplant within 30 days" group.

  - In this way we have created a variable (we shall call it "transplant30") which has a fixed value (that is, it does not change over time) for all patients in our set of 30-day survivors.

- You **can now do prediction** but prediction is conditional on $T > t_0$.

- All prediction accuracy measures are also now conditional on $T > t_0$. For example,

$$\begin{aligned} \text{Spec}(\theta, t, t_0) &= P(p < \theta | T > t, T > t_0) = P(p < \theta | T > t) \\ \text{Sens}(\theta, t, t_0) &= P(p \geq \theta | T \leq t, T > t_0) \end{aligned}$$

  where p is the probability that the event occurred before t **given it did not occur before** t_0, i.e. $p = P(T \leq t | T > t_0)$

- Two disadvantages: 1) how do you choose a landmark time? 2) You can't say anything about those with $T \leq t_0$.

References:

Clark, D. A., Stinson, E. B., Griepp, R. B., Schroeder, J. S., Shumway, N. E., & Harrison, D. C. (1971). Cardiac transplantation in man: VI. prognosis of patients selected for cardiac transplantation. Annals of Internal Medicine, 75(1), 15-21.

Gail, M. H. (1972). Does cardiac transplantation prolong life? A reassessment. Annals of Internal Medicine, 76(5), 815-817.

van Houwelingen, H., & Putter, H. (2011). Dynamic prediction in clinical survival analysis. CRC Press.

Parast, L., & Cai, T. (2013). Landmark risk prediction of residual life for breast cancer survival. Statistics in Medicine, 32(20), 3459-3471.

1.10 Additional Topics

Competing Risks

Until now we have considered survival times with a single, well-defined outcome, such as death or some other event. In some applications, however, a patient may potentially experience multiple events, only the first-occurring of which can be observed. For example, we may be interested in time from diagnosis with prostate cancer until death from that disease (Cause 1) or death from some other cause (Cause 2), but for a particular patient we can only observe the time to the first event. Of course, a patient may also be censored if he is still alive at the last follow-up time.

If interest centers on a particular outcome, time to prostate cancer death, for example, a simplistic analysis method would be to treat death from other causes as a type of censoring. This approach has the advantage that implementing it is straightforward using the survival analysis methods we have already discussed. However, a key assumption about censoring is that it is independent of the event in question. In most competing risk applications, this assumption may be questionable, and in some cases may be quite unrealistic. Consequently, interpretation of survival analyses in the presence of competing risks will always be subject to at least some ambiguity due to uncertainty about the degree of dependence among the competing outcomes.

The more appropriate analysis is to consider the **cause-specific hazard and cumulative incidence function.** Instead of $\delta_i \in \{0, 1\}$, we now have $\delta_i \in \{0, 1, 2, ...K\}$ where K is the number of different types of causes of death (or events).

The cause-specific hazard for event k is:

$$\lambda_k(t) = \lim_{\Delta \to 0} \frac{1}{\Delta} P(t \leq T < t + \Delta | T \geq t, \delta = k)$$

for event k, for $k = 1, ...K$. If we add up all the cause-specific hazards we get the previous hazard

$$\lambda(t) = \sum_{k=1}^{K} \lambda_k(t).$$

The cumulative incidence function, also called the sub-distribution function, is the cumulative probability that an individual dies from that cause k by time t:

$$F_k(t) = P(T \leq t, C = k).$$

For more details, see:

Applied Survival Analysis Using R, Moore, D. F. (2016), Chapter 9, Section 9.2.

Semi-competing Risks, Fine, J. P., Jiang, H., & Chappell, R. (2001). On semi-competing risks data. *Biometrika,* 88(4), 907-919.

Frailty Models

So far, the survival times of each case have been assumed to be independent. Methods for analyzing such survival data will not be sufficient if **cases are not independent or if the event is something that can occur repeatedly**. An example of the first type would be clustered data. For instance, one might be interested in survival times of individuals that are in the same family or in the same unit, such as a town or school.

In this case, genetic or environmental factors mean that survival times within a cluster are more similar to each other than to those from other clusters, so that the independence assumption no longer holds. In the second case, if the event of interest is, for example, the occurrence of a seizure, the event may repeat indefinitely. Then we would have multiple times per person.

One way to accommodate such structure in the data is to assign each individual in a cluster a common factor known as a **frailty** or, alternatively, as a random effect. We denote the frailty for all individuals in the ith cluster by ω_i. Then we may express the hazard function for the jth subject in the ith cluster as follows:

$$\lambda_{ij}(t_{ij}) = \lambda_0(t_{ij})\omega_i e^{\beta z_{ij}}.$$

For more details, see:
Applied Survival Analysis Using R, Moore, D. F. (2016), Chapter 9, Section 9.1.

Cure Models

When death is the outcome of interest, we can be sure that everyone will occur some day in the future. However, for other outcomes, it is not necessarily the case that every subject will hypothetically experience the event at some point. Models to accommodate this are referred to as **cure models** because, in effect, some subset of subjects are "cured" and will never experience the event. This is a common analysis approach in analyses of cancer clinical trial data.

Cure models can be a useful alternative to the standard Cox proportional hazards models for such data. First, the assumption of proportional hazards can fail when survival curves have plateaus at their tails. Second, survival plots with long plateaus may indicate heterogeneity within a patient population that can be useful to describe explicitly.

Cure models allow us to investigate what covariates are associated with either short-term or long-term effects. For example, cure models can allow us to evaluate

whether a new therapy is associated with an increase or decrease in the probability of being a long-term survivor or an improvement or detriment in survival for those who are not long-term survivors.

There are 2 major classes of cure models, mixture and nonmixture models.

- **Mixture cure models**, as the name suggests, explicitly model survival as a mixture of 2 types of patients: those who are cured and those who are not cured. Typically, the probability a patient is cured is modeled with logistic regression. The second component of the model is a survival model for patients who are not cured.
- A benefit of the mixture cure model is that it allows covariates to have different influence on cured patients and on patients who are not cured. For example, a therapy may increase the proportion of patients who are cured (evidenced by a significant hazard ratio) but not affect survival for patients who are not cured (evidenced by a nonsignificant hazard ratio). A mixture cure model allows us to tease out that relationship.
- **Nonmixture models**, instead, specify that the probability of being alive at time t is a function of the probability of being cured and the overall survival at time t. In nonmixture models, covariates can be incorporated both in the model for the probability of being cured and the survival probability.
- The interpretation of covariates is different with the nonmixture cure model than with the mixture model. Covariates included in the survival distribution characterize a "short-term" effect, but the covariates do not describe the survival for those who are not cured because the nonmixture model does not directly model a mixture population.

For more details, see:

Othus, M., Barlogie, B., LeBlanc, M. L., & Crowley, J. J. (2012). Cure Models as a Useful Statistical Tool for Analyzing SurvivalCure Models and Multiple Myeloma. *Clinical Cancer Research*, 18(14), 3731-3736.