Biostat 537: Survival Analaysis

TA Session 4

Ethan Ashby

January 29, 2024

Review of Last Time

1. Parametric survival methods offer convenient estimation but often suffer from unmet assumptions.
2. Nonparametric survival methods are often preferred because they enable estimation and inference without making unnecessary assumptions.
3. The Kaplan-Meier estimator is the most common estimator of the survival curve.
4. Nonparametric estimators of the the hazard function require smoothing to account for noisy data.
5. The Logrank test and its variants are nonparametric tests of the equality of survival between groups.

Presentation Overview

The Logrank Test: A Review

The logrank test is a test of the equality of survivor curves across two groups $H_0 : S_0(t) = S_1(t)$.

Formally, the test is based on the logrank statistic which depends on the sum over all the unique failure times of the observed minus expected failures under H_0.

$$\text{Logrank Statistic} = \frac{\left(\sum_{t_{(i)}} (d_{1i} - e_{1i})\right)^2}{\text{Var}(d_{1i} - e_{1i})} \overset{H_0}{\sim} \chi^2_1$$

Motivating the Stratified Logrank Test

Suppose we wish to pursue a study to test whether smoking has a *causal effect* on lung-cancer free survival.

We sample a cohort of smokers and non-smokers without lung cancer from a registry, and we follow them to their lung cancer diagnosis or end of study.

Suppose we wish to test $H_0 : S_{\text{Smoke}}(t) = S_{\text{No Smoke}}(t)$ using a logrank test.

What are some potential limitations of this analysis?

Introducing the Stratified Logrank Test

Solution: calculate observed minus expected event counts *separately* within groups of participants with the same confounder (e.g., alcohol consumption)

$$(O - E)_{\text{No Alcohol}} = \left(\sum_{t_{(i)}} (d_{1i}^{\text{No Alc}} - e_{1i}^{\text{No Alc}}) \right)$$

$$(O - E)_{\text{Alcohol}} = \left(\sum_{t_{(i)}} (d_{1i}^{\text{Alc}} - e_{1i}^{\text{Alc}}) \right)$$

We then pool observed minus expected event counts across levels of the confounder

$$\frac{(O - E)}{\text{Var}(O - E)} = \frac{\sum_{s \in \mathcal{S}} (O - E)_s}{\sum_{s \in \mathcal{S}} \text{Var}(O - E)_s} \overset{H_0}{\sim} \chi^2_{|\mathcal{S}|-1}$$

Stratified Logrank Test: in R

```
library(survival)
survdiff(Surv(tt,delta)~smoke+strata(alcohol), rho=0)
```

Introducing Stratified Logrank Test

The stratified logrank test is a good idea if

1. The exposure of interest (e.g., smoking) is not randomly assigned and is likely entangled with other explanatory variables called confounders (e.g., alcohol consumption) which may affect the outcome.
2. There exist a small number of discrete variables which are believed to contribute all/most of the confounding.

The stratified logrank test is a bad idea if

1. The exposure of interest is randomly assigned.
2. There exist many/high-dimensional/continuous variables believed to be confounders.
3. You have a relatively limited sample size.

Roadmap

1. More on the Logrank Test
2. The Cox Regression Model
3. Cox Modelling in R

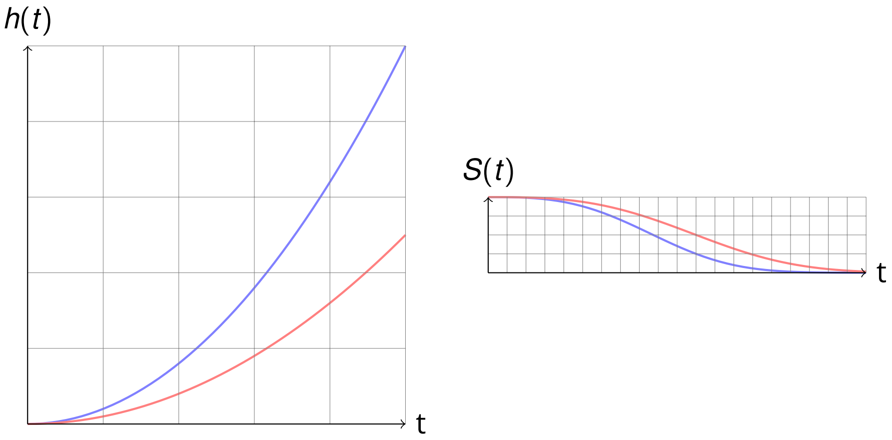
Proportional Hazards

The Logrank test is a test of the null hypothesis $H_0 : S_0(t) = S_1(t)$.

The logrank test is designed to distinguish H_0 from $H_A : [S_0(t)]^{\psi} = S_1(t)$ for $\psi \neq 1$.

The alternative hypothesis is equivalent to $H_A : h_1(t) = \psi h_0(t)$, which represents the *proportional hazards assumption*.

What does proportional hazards look like?



Group exercise: are the hazards proportional?

Consider the following covariates and opine whether the proportional hazards assumption will be satisfied.

1. Effect of placebo versus a prevention drug with short half-life on time-to-influenza.
2. Effect of helmetless versus helmeted cycling on the time to head injury.
3. Effect of each additional $100 monthly income on time until someone declares they are happy with their life.

Proportional Hazards

One way to compare survival distributions is to *assume* the hazards are proportional, $h_1(t) = \psi h_0(t)$, and test whether $H_0 : \psi = 1$ or $H_A : \psi \neq 1$.

Key idea: we can incorporate covariates X in the hazard modifier: $\psi = \exp(\beta X)$! Hence, $H_0 : \psi = 1 \iff H_0 : \beta = 0$.

This sets up a very useful framework for regression modelling of survival data!

Overview of Regression and a Challenges

Goal of regression: develop and estimate a meaningful model relating a set of explanatory variables (covariates) X and an outcome.

Challenge in Survival Setting:

1. If we adopt a parametric approach: estimation is possible, but model may not reflect reality.
2. If we adopt a nonparametric approach: how do we perform estimation and inference esp w/ censored data and without a likelihood?

Partial Likelihood

Suppose we want to estimate the survival difference between two groups ($z = 0, 1$) assuming $h_1(t) = \psi h_0(t)$ with $\psi = e^{\beta z}$. Hence $\psi = 1$ for $z = 0$ and e^{β} for $z = 1$.

Suppose we have a set of n in the risk set R_1. Suppose we go to the first failure time t_1 which was when participant i failed. The probability that participant i failed at time t_1 is given by

$$p_1 := \frac{h_i(t_1)}{\sum_{k \in R_1} h_k(t_1)}$$

$$= \frac{\psi_i h_0(t_1)}{\sum_{k \in R_1} \psi_k h_0(t_1)} = \frac{\psi_i}{\sum_{k \in R_1} \psi_k}$$

Partial Likelihood

$$p_1 := \frac{h_i(t_1)}{\sum_{k \in R_1} h_k(t_1)} = \frac{\psi_i h_0(t_1)}{\sum_{k \in R_1} \psi_k h_0(t_1)} = \frac{\psi_i}{\sum_{k \in R_1} \psi_k}$$

The baseline hazard cancels out in the above expression.

At second event time t_2, there are $n - 1$ people in the risk set, R_2. Suppose person j fails. The probability this occurred was

$$p_2 := \frac{h_j(t_2)}{\sum_{k \in R_2} h_k(t_2)} = \frac{\psi_j}{\sum_{k \in R_2} \psi_k}$$

Partial Likelihood

We can calculate $p_1, p_2, \ldots, p_T$ for all the T event times. Then the partial likelihood of the observed data is the product $L(\psi) := p_1 \cdot, p_2 \ldots p_T$.

In the partial likelihood, the baseline hazard $h_0(t)$, which describes the potential of experiencing the event in group $z = 0$, is treated as a *nuisance* – a statistical quantity not of direct interest.

Example

Patient	Survtime	Censor	Group (z)	ψ
1	6	1	0	1
2	7	0	0	1
3	10	1	1	$\exp(\beta)$
4	15	1	0	1
5	19	0	1	$\exp(\beta)$
6	25	1	1	$\exp(\beta)$

$$p_1^{t_1=6} = \frac{1 \cdot h_0(t_1)}{3h_0(t_1) + 3\psi h_0(t_1)} = \frac{1}{3\psi + 3} \qquad p_2^{t_2=10} = \frac{\psi}{3\psi + 1}$$

$$p_3^{t_3=15} = \frac{1}{2\psi + 1} \qquad p_4^{t_4=25} = 1$$

Example

The partial likelihood $L(\psi)$ takes the form

$$L(\psi) := \frac{\psi}{(3\psi + 3)(3\psi + 1)(2\psi + 1)(1)}$$

Recalling $\psi = e^{\beta}$, the log-partial likelihood takes the form

$$\ell(\beta) = \beta - \log(3e^{\beta} + 3) - \log(3e^{\beta} + 1) - \log(2e^{\beta} + 1)$$

The *maximum partial likelihood estimator* can be solved by finding the value of β which maximizes the partial likelihood score equation.

$$\frac{\partial \ell}{\partial \beta} = 0$$

This step is often done w/ a computer: yields $\hat{\beta} = -1.326 \implies \hat{\psi} = 0.265!$

Example

Recall our null hypothesis of "no group effect": $H_0 : \beta = 0$.

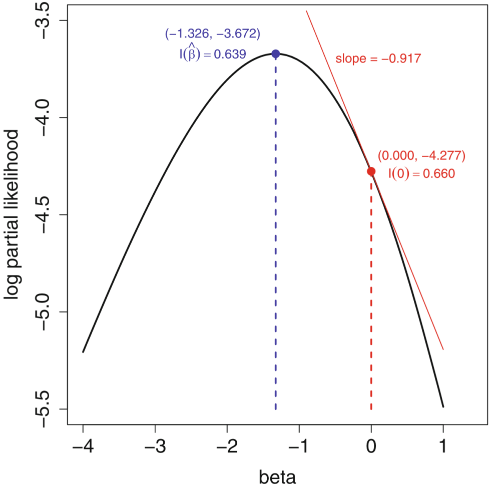
One way of testing H_0 is to calculate the maximum partial likelihood estimate $\hat{\beta}$ and compare it to the null value β_0, scaled by the standard error. This is a *Wald test*.

$$Z = \frac{(\hat{\beta}_{\text{MPLE}} - \beta_0)}{\sqrt{I(\hat{\beta})}} \qquad I(\hat{\beta}) = \frac{\partial^2}{\partial \beta^2} \log(L(\beta))\Big|_{\beta=\hat{\beta}}$$

Another way of testing H_0 is to evaluate the derivative/slope of the log partial likelihood function at the null value $\beta = 0$ and see if it is close to 0 (meaning we are near the maximum). This is a *Score test*.

$$Z_s = \frac{S(\beta = 0)}{\sqrt{I(\beta = 0)}} \qquad S(\beta = 0) = \frac{\partial}{\partial \beta} \log(L(\beta))\Big|_{\beta=0}$$

Example


(−1.326, −3.672)
I(β̂) = 0.639
slope = −0.917
(0.000, −4.277)
I(0) = 0.660
log partial likelihood
−3.5
−4.0
−4.5
−5.0
−5.5
−4
−3
−2
−1
0
1
beta

Amazing facts about the partial likelihood

1. Amazingly, the slope of the partial likelihood function at $\beta = 0$ is *equivalent* to the value of the logrank statistic!
2. Unlike the logrank test, the Cox partial likelihood can accommodate X as discrete or continuous variables.
3. The partial likelihood does not account for the particular *values* of the failure times – only their orders.
4. The Cox model only assumes $h(t|X) = h_0(t)\exp(\beta X)$. Such a model is a semiparametric model.

Roadmap

In R

```
result.cox<-coxph(Surv(tt,status)~grp)
summary(result.cox)
```

```
Call: coxph(formula = Surv(tt, status) ~ grp)

  n= 6, number of events= 4

       coef exp(coef) se(coef)     z Pr(>|z|)
grp -1.3261    0.2655   1.2509 -1.06    0.289

    exp(coef) exp(-coef) lower .95 upper .95
grp    0.2655      3.766   0.02287     3.082

Concordance= 0.7  (se = 0.187 )
Rsquare= 0.183   (max possible= 0.76 )
Likelihood ratio test= 1.21  on 1 df,   p=0.2715
Wald test            = 1.12  on 1 df,   p=0.2891
Score (logrank) test = 1.27  on 1 df,   p=0.2591
```

Summary

1. The Logrank test: $H_0 : S_0(t) = S_1(t)$ without making parametric assumptions. Stratified variants enables control of a few discrete confounders.
2. Regression modelling of the effects of covariates, X, on the survival experience can be done under the assumption of proportional hazards $h(t|X) = h_0(t)\exp(\beta X)$.
3. The *Cox partial likelihood* is the basis for estimation and inference on β.