# Biostat 537: Survival Analysis

## TA Session 5

Ethan Ashby

February 13, 2024

# Review from Last Time

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 $\beta$.

# Presentation Overview

# Overview of Cox Regression

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

*Cox PH regression*: assumes covariates modify the underlying baseline hazard proportionally (baseline hazard treated as a *nuisance*).

$$h(t|X) = h_0(t) \cdot \exp(\beta X)$$

# Partial Likelihood

Suppose we want to estimate the survival difference between two groups ($z = 0, 1$) using a Cox model: assuming $h_1(t) = \psi h_0(t)$ with $\psi = e^{\beta z}$.

Suppose we have a set of $n$ in the risk set $R_1$. Suppose participant $i$ failed at the first failure time $t_1$. The probability that this happened 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.

# Cox Model

The Cox model assumes the conditional hazard satisifies

$$h(t|X) = h_0(t)\exp(\beta^T X)$$

$X \in \mathbb{R}^p$ is a vector of $p$ explanatory variables (covariates) and $\beta$ is the vector of log hazard ratios.

The Cox partial likelihood provides the basis for estimation and inference on $\beta$.

$$L(\beta) = \prod_{j=1}^{D} \frac{h_0(t_j)\exp(\beta^T X_j)}{\sum_{k\in R_j} h_0(t_j)\exp(\beta^T X_k)} = \prod_{j=1}^{D} \frac{\exp(\beta^T X_j)}{\sum_{k\in R_j} \exp(\beta^T X_k)}$$

# Check your understanding

Suppose we are planning a randomized trial to evaluate whether a vaccine is better than placebo in preventing flu infection. Brainstorm with your neighbors regarding a reasonable method to analyze the data that would come out from the trial.

1. Estimation versus testing?
2. Validity of PH assumption?
3. Missingness assumptions?
4. Adjustment for other explanatory factors?

# Switching between survival quantities w/ Cox Model

Recall we can estimate the baseline hazard under a Cox model using a Nelson-Aaelen-type estimator

$$\hat{h}_0(t_i) = \frac{d_i}{\sum_{j \in R_j} \exp(x_j \hat{\beta})}$$

Where $d_i$ is the number who experienced the event at time $t_i$ and the denominator is the number at risk while controlling for covariates $x_j$.

The baseline and conditional survivor curve can be estimated as

$$\hat{S}_0(t) = \exp\left(-\int_0^t \hat{h}_0(u)du\right)$$

$$\hat{S}(t|x) = [\hat{S}_0(t)]^{\exp(\hat{\beta}^T X)}$$

# Ties

Recall that the Cox partial likelihood we discussed so far requires a *unique order* of event times.

However, in practice, there may exist ties in event times due to (a) coarse measurement of a continuous time process or (b) truly discrete event times.

# Method 1 of Tie Handling: Continuous

Idea: if event times are truly continuous, average over all possible unique orderings of tied failure times.

E.g., suppose that at the first event time $t_1$, there are $A$ treated ($X = 1$) subjects and $B$ control ($X = 0$) subjects at risk. One treated and one control participant experience the event tied at $t_1$. The factor in the partial likelihood will be

$$p_1 := \underbrace{\left(\frac{1}{Ae^{\beta} + B}\right)\left(\frac{e^{\beta}}{Ae^{\beta} + B - 1}\right)}_{\text{Control First}} + \underbrace{\left(\frac{e^{\beta}}{Ae^{\beta} + B}\right)\left(\frac{1}{(A-1)e^{\beta} + B}\right)}_{\text{Treated First}}$$

# Method 1 of Tie Handling: Discrete

Idea: if event times are truly discrete, compare observed hazard to sum of hazards over all possible tie combinations.

E.g., suppose that at the first event time $t_1$, there are $A$ treated ($X = 1$) subjects and $B$ control ($X = 0$) subjects at risk. One treated and one control participant experience the event tied at $t_1$. The factor in the partial likelihood will be

$$p_1 := \frac{e^{\beta} \cdot 1}{\sum_{i \in R_{t_1}} \sum_{j \in R_{t_1} : j > i} \psi_i \cdot \psi_j}$$

# Need for Alternatives

When there are many ties, enumerating all combinations of event times becomes computationally prohibitively expensive. Shortcuts are needed.

*Breslow approximation*: adjusts both terms in the marginal method such that they have the same denominator.

$$p_1 := \underbrace{\left(\frac{1}{Ae^\beta + B}\right)\left(\frac{e^\beta}{Ae^\beta + B - 1}\right)}_{\text{Control First}} + \underbrace{\left(\frac{e^\beta}{Ae^\beta + B}\right)\left(\frac{1}{(A-1)e^\beta + B}\right)}_{\text{Treated First}}$$

$$\approx \left(\frac{e^\beta}{(Ae^\beta + B)^2}\right) + \left(\frac{e^\beta}{(Ae^\beta + B)^2}\right) = \left(\frac{2e^\beta}{(Ae^\beta + B)^2}\right)$$

# Need for Alternatives

*Efron method*: each participant has an equal probability of being at risk in second denominator

$$p_1 = \left(\frac{1}{Ae^{\beta} + B}\right)\left(\frac{e^{\beta}}{(A - 0.5)e^{\beta} + B - 0.5}\right)$$

This is the tie-handling method used by default in R.

# Importance of Model Selection

The Cox model represents a powerful tool for estimating and inferring the association between a covariate and survival time.

In practice, we may collect data on several covariates, which may/may not be explanatory of the survival experience. How do we develop a model that is both

1. parsimonious enough to preserve power to detect true associations
2. Complex enough to be meaningful and eliminate obvious sources of confounding

# Partial Likelihood Ratio Test

Suppose we are interested in testing whether a covariate $X_1$ belongs in the Cox model. This is the case of *nested models*. We can calculate the log partial likelihood of the data for a Cox model with the covariate $X_1$ (full model) and without the covariate (reduced model).

$$2\left[\ell_{\text{Full}} - \ell_{\text{Reduced}}\right] \sim \chi^2_{\dim(X_1)}$$

# AIC for non-nested models

Suppose we are interested in sorting through a large collection of covariates to obtain a parsimonious model.

We can pursue a stepwise procedure that adds/subtracts variables one-by-one with the goal of optimizing a criterion function such as the Akaike Information Criterion (AIC).

$$\text{AIC} = -2\ell(\hat{\beta}) + 2k$$

where $\ell(\hat{\beta})$ is the log partial likelihood at the MPLE and $k$ is the number of parameters in your model.

AIC balances the model fit and model complexity, identifying a "good-fitting" model with few parameters.

# Stepwise selection in R

```
1  modelAll.coxph <- coxph(Surv(ttr, relapse) ~ grp +
       gender + race +employment+yearsSmoking+
       levelSmoking+ ageGroup4 + priorAttempts +
       longestNoSmoke)
2  result.step <- step(modelAll.coxph, scope=list(upper
       =~ grp + gender + race + employment +
       yearsSmoking + levelSmoking + ageGroup4 +
       priorAttempts + longestNoSmoke, lower=~grp) )
```

# Handling Nonlinearity

Recall that including continuous covariate $X$ in a Cox model implies the hazard grows linearly in $X$.

In many cases, we may wish to model the effect of a continuous covariate $X$ that may affect the hazard nonlinearly.

# Handling Nonlinearity

Recall that including continuous covariate $X$ in a Cox model implies the hazard grows linearly in $X$.

In many cases, we may wish to model the effect of a continuous covariate $X$ that may affect the hazard nonlinearly.

An appealing solution are *splines*, piecewise polynomials stitched together at support points called knots.

# Handling Nonlinearity

Splines will improve the fit of the model by introducing additional flexibility. We turn to a *penalized partial likelihood* approach for estimation and inference, where a penalty is paid for a more complex spline curve.

```
1  modelAll.coxph <- coxph(Surv(ttr, relapse) ~ grp +
       employment+pspline(age, df=4))
```

# Handling Nonlinearity