Cox's proportional hazards model and Cox's partial likelihood

Rasmus Waagepetersen

September 29, 2024

Non-parametric vs. parametric

Suppose we want to estimate unknown function, e.g. survival function.

Approaches:

- Non-parametric using Kaplan-Meier. Advantage: no assumption regarding type of distribution. Disadvantage: requires identically distributed observations (usually independence assumption too)
- Parametric model. Advantage: we only need to estimate a few parameters that completely characterize distribution (e.g. exponential or Weibull) - gives low variance of estimates. Can be extended to non-*iid* observations using regression on covariates. Disadvantage: assumed model class may be (or always is) incorrect leading to model error or in other words, bias.

Possible to combine the best of two approaches ?

Semi-parametric approach - Cox's proportional hazards model

Sir David Cox in a ground-breaking paper ('Regression models and life tables', 1972) suggested the following model for the hazard function given covariates $z \in \mathbb{R}^p$:

$$h(t; z) = h_0(t) \exp(z^\mathsf{T}\beta), \quad \beta \in \mathbb{R}^p.$$

Here $h_0(\cdot)$ completely unspecified function except that it must be non-negative.

Thus model combines great flexibility via non-parametric $h_0(\cdot)$ with the possibility of introducing covariate effects via exponential term $\exp(z^\mathsf{T}\beta)$

This model has become standard in medical statistics.

Some properties

Cumulative hazard:

$$H(t;z) = \exp(z^\mathsf{T}\beta)\int_0^t h_0(u)\mathrm{d}u = \exp(z^\mathsf{T}\beta)H_0(t)$$

Survival function

$$S(t;z) = S_0(t)^{\exp(z^\mathsf{T}\beta)} \quad S_0(t) = \exp(-H_0(t))$$

Proportional hazards:

$$\frac{h(t;z)}{h(t;z')} = \exp((z-z')^\mathsf{T}\beta)$$

i.e. constant hazard ratio for two different subjects - curves can not cross ! - this should be checked in any application.

Estimation - partial likelihood

Model useless if we can not estimate parameter β.

Problem: we can not use likelihood when $h_0(\cdot)$ unspecified.

Second break-through contribution of Cox: invention of *partial* likelikehood for estimating β.

Suppose we have observations (t_i, δ_i) as well as (fixed) covariates $z_1, \ldots, z_n$, $i = 1, \ldots, n$. We assume no ties (all t_i distinct) and define $D \subseteq \{1, \ldots, n\}$ as

$$D = \{l | \delta_l = 1\}$$

- i.e. the *index* set of death times.

For any $t \geq 0$ we further define the risk set

$$R(t) = \{l | t_l \geq t\}$$

i.e. the index set of subjects at risk at time t.

The partial likelihood

The partial likelihood is

$$L(\beta) = \prod_{l \in D} \frac{\exp(z_l^\mathsf{T} \beta)}{\sum_{k \in R(t_l)} \exp(z_k^\mathsf{T} \beta)}$$

Cox suggested to estimate β by maximizing $L(\beta)$.

- does not depend on h_0
- does not depend on actual death times - only their order
- censored observations only appear in risk set (as for Kaplan-Meier)

Cox's idea has proven to work very well - but why ? Lots of people have tried to make sense of this partial likelihood.

Cox's intuition

Consider for simplicity the case of no censoring and let $t_{(1)}, \ldots, t_{(n)}$ denote the set of ordered death times.

We can equivalently represent data as the set of inter-arrival times $v_i = t_{(i)} - t_{(i-1)}$ (taking $t_{(0)} = 0$) together with the information $r_1, r_2, \ldots, r_n$ about which subject died at each time of death - i.e. $r_i = l$ if subject l was the ith subject to die.

Cox then factored likelihood of $(v_1, \ldots, v_n, r_1, \ldots, r_n)$ as (using generic notation for densities and probabilities)

$$
\begin{aligned}
&f(v_1)p(r_1|v_1)f(v_2|v_1, r_1)p(r_2|v_1, v_2, r_1)\cdots \\
&\qquad f(v_n|v_1, \ldots, v_{n-1}, r_1, \ldots, r_{n-1})p(r_n|v_1, \ldots, v_n, r_1, \ldots, r_{n-1})
\end{aligned}
$$

Cox argued that terms $f(v_i|\ldots)$ could not contribute with information regarding β since the interarrival times can be fitted arbitrary well regardless of β when h_0 is unrestricted - we can essentially just choose h_0 to consist of 'spikes' at each death time.

Thus estimation of β should be based on remaining factors

$$L(\beta) = \prod_{i=1}^{n} p(r_i|H_i)$$

where $H_i = \{v_1, \ldots, v_i, r_1, \ldots, r_{i-1}\}$ history/previous observations.

Here $p(r_i|H_i)$ is the probability that subject r_i is the ith person to die given the previous observations.

More precisely, let R_i denote the random index of the ith subject that dies ($R_i = l$ means that T_l is the ith smallest death time, i.e. $T_{R_i} = T_{(i)} = T_l$).

Assume that $p(l|H_i)$ only depends on H_i through the knowledge that the ith death happens at time $t_{(i)}$ and that $R(t_{(i)})$ are the ones at risk at time $t_{(i)}$.

Thus

$$p(l|H_i) = P(R_i = l | T_{R_i} \in [t_{(i)}, t_{(i)} + dt[, R(t_{(i)}) = A)$$

This is the probability that l is the ith person to die given that the ith death happens at time $t_{(i)}$ and that the persons in A are at risk at time $t_{(i)}$ (thus probability is zero if $l \notin A$)

We now express the conditional probability in terms of the hazard function:

$$
\begin{aligned}
& P(R_i = l, T_{R_i} \in [t_{(i)}, t_{(i)} + dt[| R(t_{(i)}) = A) \\
= & P(T_l \in [t_{(i)}, t_{(i)} + dt[, T_k > T_l, k \in A \setminus \{l\} | R(t_{(i)}) = A) \\
`=' & h_0(t_{(i)}) \exp(z_l^\mathsf{T} \beta) \mathrm{d}t \prod_{k \in A \setminus \{l\}} (1 - h_0(t_{(i)}) \exp(z_k^\mathsf{T} \beta) \mathrm{d}t)
\end{aligned}
$$

Note '$=$' because we actually replace $T_k > T_l$ by $T_k > t_{(i)} + \mathrm{d}t$. This does not really matter since $\mathrm{d}t$ infinitesimal.

NB: if $R_i = l$ then $t_{(i)} = t_l$ so in the following we replace $t_{(i)}$ with t_l.

Finally,

$$
\begin{aligned}
& P(R_i = l | T_{R_i} \in [t_l, t_l + dt[, R(t_l) = A) \\
= & \frac{P(R_i = l, T_{R_i} \in [t_l, t_l + dt[| R(t_l) = A)}{P(T_{R_i} \in [t_l, t_l + dt[| R(t_l) = A)} \\
= & \frac{P(R_i = l, T_{R_i} \in [t_l, t_l + dt[| R(t_l = A))}{\sum_{j \in R(t_l)} P(R_i = j, T_{R_i} \in [t_l, t_l + dt[| R(t_l) = A)} \\
= & \frac{h_0(t_l) \exp(z_l^\mathsf{T} \beta) \mathrm{d}t \prod_{k \in R(t_l) \setminus \{l\}} (1 - h_0(t_l) \exp(z_k^\mathsf{T} \beta) \mathrm{d}t)}{\sum_{j \in R(t_l)} h_0(t_l) \exp(z_j^\mathsf{T} \beta) \mathrm{d}t \prod_{k \in R(t_l) \setminus \{j\}} (1 - h_0(t_l) \exp(z_k^\mathsf{T} \beta) \mathrm{d}t)} \\
= & \frac{\exp(z_l^\mathsf{T} \beta)}{\sum_{k \in R(t_l)} \exp(z_k^\mathsf{T} \beta)}
\end{aligned}
$$

Note: last $=$ follows after cancelling $h_0(t_l)\mathrm{d}t$ and noting that $(1 - h_0(t_l) \exp(z_k^\mathsf{T} \beta) \mathrm{d}t)$ tends to one when $\mathrm{d}t$ tends to zero.

NB: denominator is hazard for minimum of $T_k, k \in R(t_l)$ (exercise 18)

Conditional likelihood for matched case-control study

Cox's idea very closely related to conditional likelihood for matched case-control studies.

Let X denote a binary random variable (e.g. sick/healthy) for an individual in a population. We want to study the impact of a covariate z on X.

Assume that the population can be divided into homogeneous groups (strata) so that probability of being ill is given by a logistic regression

$$P(X = 1) = p_i(z) = \frac{\exp(\alpha_i + \beta z)}{1 + \exp(\alpha_i + \beta z)}$$

for an individual in the ith strata and with the covariate z.

Suppose $X_1 = 1$ with covariate z_1 is observed for a sick person in the ith stratum. In a matched case-control study this observation is paired with an observation $X_2 = 0$ with covariate z_2 for a randomly selected healthy person in the same stratum.

The conditional likelihood is now based on the conditional probabilities

$$P(X_1 = 1 | X_1 = 1, X_2 = 0 \text{ or } X_1 = 0, X_2 = 1) = \frac{p_i(z_1)(1 - p_i(z_2))}{p_i(z_1)(1 - p_i(z_2)) + (1 - p_i(z_1))p_i(z_2)}$$

This reduces to

$$\frac{\exp(\beta z_1)}{\exp(\beta z_1) + \exp(\beta z_2)}$$

which is free of the strata specific intercept α_i.

Note α_i is a nuisance parameter when we are just interested in β.

Invariance argument

Again consider the case of no censoring. Kalbfleisch and Prentice noticed that if one applies a strictly increasing differentiable function g to the survival times $T_1, \ldots, T_n$ then $\tilde{T}_i = g(T_i)$ again follows a proportional hazards model with a completely unspecified hazard function $\tilde{h}_0$ (exercise 17).

Hence estimation problem for β the same regardless of whether we consider T_i's or $\tilde{T}_i$'s.

They thus concluded that only the ordering (ranks) of the survival times and not the magnitudes of the survival times could matter for inference on β.

One can verify (exercise 23) that for the ranks R_i,

$$P(R_1 = r_1, \ldots, R_n = r_n) = P(T_{r_1} < T_{r_2} < \cdots < T_{r_n})$$

is precisely Cox's partial likelihood.

Profile likelihood

Cox's partial likelihood can also be derived as a profile likelihood.

Consider likelihood (assuming no ties)

$$\prod_{i=1}^{n}[h_0(t_i)\mathrm{d}t \exp(z_l^\mathsf{T}\beta)]^{\delta_i} \exp[-\exp(z_i^\mathsf{T}\beta)\int_0^{t_i} h_0(u)\mathrm{d}u].$$

Let's try to maximize wrt h_0. First, we need $h_0(t_l) > 0$ for $l \in D$. At the same time we should take $h_0(u) = 0$ between death times.

So we let $h_0(t)\mathrm{d}t = \alpha_l$ in very small intervals around death times, $[t_l, t_l + \mathrm{d}t[$, $l \in D$, and zero elsewhere. Note likelihood does not inform about $h_0(t)$ for t larger than $\max_i t_i$.

Then likelihood becomes

$$L(\alpha,\beta) = \left(\prod_{l \in D} \alpha_l \exp[z_l^{\mathsf{T}}\beta]\right) \exp(-\sum_{i=1}^{n} \exp(z_i^{\mathsf{T}}\beta) \sum_{l \in D: t_l \le t_i} \alpha_l)$$

$$= \left(\prod_{l \in D} \alpha_l \exp[z_l^{\mathsf{T}}\beta]\right) \exp(-\sum_{l \in D} \alpha_l \sum_{i \in R(t_l)} \exp(z_i^{\mathsf{T}}\beta))$$

Taking log and differentiating wrt α_l we obtain

$$\frac{\partial}{\partial \alpha_l} \log L(\alpha,\beta) = \frac{1}{\alpha_l} - \sum_{j \in R(t_l)} \exp(z_j^{\mathsf{T}}\beta)$$

Setting equal to zero and solving wrt α_l gives

$$\hat{\alpha}_l(\beta) = \frac{1}{\sum_{j \in R(t_l)} \exp(z_j^{\mathsf{T}}\beta)}$$

Plugging in $\hat{\alpha}_l(\beta)$ for α_l we finally obtain profile likelihood:

$$L_p(\beta) = L(\hat{\alpha}, \beta) = \left(\prod_{l \in D} \frac{\exp(z_l^{\mathsf{T}} \beta)}{\sum_{j \in R(t_l)} \exp(z_j^{\mathsf{T}} \beta)} \right) \exp(-|D|)$$

which is Cox's partial likelihood.

As a byproduct we obtain the Breslow estimate of H_0:

$$\hat{H}_0(t) = \sum_{\substack{l \in D: \\ t_l \leq t}} \frac{1}{\sum_{j \in R(t_l)} \exp(z_j^{\mathsf{T}} \beta)}$$

where we replace β by partial likelihood estimate $\hat{\beta}$.

This reduces to Nelson-Aalen estimator if $\beta = 0$.

Note $\hat{H}_0(t)$ is discontinuous in contrast to $H_0(t) = \int_0^t h_0(u) \mathrm{d}u$.

$\hat{H}_0(t)$ limiting case of H_0 with mass increasingly concentrated around death times.

Estimating function point of view

All previous derivations more or less heuristic.

However, not crucial to understand Cox's partial likelihood as a likelihood or as derived from a likelihood.

Just consider properties of associated estimating function.

Score of partial likelihood is an estimating function which (see next slide) is

- unbiased (each term mean zero)
- sum of uncorrelated terms (gives CLT)

- general theory for estimating functions suggests that partial likelihood estimates asymptotically consistent and normal.

Variance and mean heuristics - assuming no censoring

Score function

$$u(\beta) = \frac{\mathrm{d}}{\mathrm{d}\beta} \log L(\beta) = \sum_{i=1}^{n} u_i(\beta)$$

is sum of n terms

$$u_i(\beta) = z_{R_i} - \mathbb{E}[z_{R_i} | T_{R_i} \in [t_{(i)}, t_{(i)} + dt[, R(t_{(i)})].$$

Each term has mean zero:

$$\mathbb{E}[u_i(\beta)] = \mathbb{E}[\mathbb{E}[u_i(\beta)|H_i]] = 0$$

Moreover, terms are uncorrelated. For $i < j$:

$$\mathbb{E}[u_i(\beta)u_j(\beta)] = \mathbb{E}[u_i(\beta)\mathbb{E}[u_j(\beta)|H_j]] = 0$$

Thus good reason to believe that CLT works for score function.

Asymptotic properties of estimates and tests

The 'observed information' for the partial likelihood is

$$j(\beta) = -\frac{\mathrm{d}}{\mathrm{d}\beta^{\mathsf{T}}} u(\beta) = \sum_{i=1}^{n} \mathbb{V}\mathrm{ar}[z_{R_i} | T_{R_i} \in [t_{(i)}, t_{(i)} + dt[, R(t_{(i)})] =$$

$$\sum_{i=1}^{n} \mathbb{V}\mathrm{ar}[u_i(\beta) | H_i]$$

'Information' (see next slide for second '=')

$$i(\beta) = \mathbb{E} j(\beta) = \mathbb{V}\mathrm{ar}(u(\beta))$$

In analogy with usual asymptotic results we obtain for large n,

$$(\hat{\beta} - \beta) \approx N(0, i(\beta)^{-1})$$

In practice we estimate $i(\beta)$ by $j(\hat{\beta})$. This can be used for constructing confidence intervals in the usual way.

Moreover, we can construct Wald tests, score-tests and 'likelihood-ratio' tests in the usual way.

Second 'Bartlett identity'

Since $\mathbb{E}(u_i(\beta)|H_i) = 0$,

$$\mathbb{V}\mathrm{ar}\, u_i(\beta) = \mathbb{E}\mathbb{V}\mathrm{ar}(u_i(\beta)|H_i) + \mathbb{V}\mathrm{ar}\mathbb{E}(u_i(\beta)|H_i) = \mathbb{E}\mathbb{V}\mathrm{ar}(u_i(\beta)|H_i)$$

Moreover, since $u(\beta)$ is a sum of uncorrelated terms,

$$\mathbb{V}\mathrm{ar}\, u(\beta) = \sum_{i=1}^{n} \mathbb{V}\mathrm{ar}\, u_i(\beta)$$

Combining the above,

$$\mathbb{E}j(\beta) = \sum_{i=1}^{n} \mathbb{E}\mathbb{V}\mathrm{ar}(u_i(\beta)|H_i) = \mathbb{V}\mathrm{ar}\, u(\beta)$$

Asymptotic distribution - sketch

Let β^* denote 'true' value of regression parameter.

First order (multivariate) Taylor around $\hat{\beta}$

$$u(\beta^*) = u(\hat{\beta}) + \frac{\mathrm{d}}{\mathrm{d}\beta^{\mathsf{T}}} u(\beta)|_{\beta=\tilde{\beta}}(\beta^* - \hat{\beta}) = j(\tilde{\beta})(\hat{\beta} - \beta^*)$$

where $|\tilde{\beta} - \beta^*| \leq |\hat{\beta} - \beta^*|$ and we have used $u(\hat{\beta}) = 0$.

Thus

$$(\hat{\beta} - \beta^*) = j(\tilde{\beta})^{-1} u(\beta^*).$$

Moreover

$$i(\beta^*)^{1/2}(\hat{\beta} - \beta^*) = (i(\beta^*)^{-1/2} j(\tilde{\beta}) i(\beta^*)^{-1/2})^{-1} i(\beta^*)^{-1/2} u(\beta^*) \tag{1}$$

Assume now as n tends to infinity,

$$i(\beta^*)^{-1/2}u(\beta^*) \to N(0, I) \text{ (CLT)}$$

(convergence in distribution) and

$$i(\beta^*)^{-1/2}j(\tilde{\beta})i(\beta^*)^{-1/2} \to I$$

(convergence in probability).

Combining this with (1) on previous slide we obtain

$$i(\beta^*)^{1/2}(\hat{\beta} - \beta^*) \to N(0, I)$$

in distribution.

In other words

$$\hat{\beta} \approx N(\beta^*, i(\beta^*)^{-1})$$

Consider $H_0 : \beta = \beta_0$. Several possibilities under H_0:

- (Wald) $j(\beta_0)^{1/2}(\hat{\beta} - \beta_0) \approx N(0, I)$
- (Score test) $j(\beta_0)^{-1/2}u(\beta_0) \approx N(0, I)$
- ('likelihood ratio) $-2\log(L(\beta_0)/L(\hat{\beta})) \approx \chi^2(p)$

See KM 8.3 and 8.5 for further details.

NB: in the case of $z_i \in \{0, 1\}$ (two-group scenario), score-test for $H_0 : \beta = 0$ is equivalent with log-rank test (exercise 19).

Data with ties

Suppose we have tied death times

$$t^*_{11} = t^*_{12} = \cdots = t^*_{1d_1} < t^*_{21} = \cdots = t^*_{2d_2} < \cdots < t^*_{r1} = \cdots = t^*_{rd_r}$$

I.e. r distinct death times with d_l deaths at the l' distinct time. Let z^*_{lj} be the covariate for the individual with death time t^*_{lj} and let $z^*_{l\cdot} = \sum_{j=1}^{d_l} z^*_{lj}$.

Suppose we knew $t^*_{l1} < t^*_{l2} < \cdots < t^*_{ld_l}$, $l = 1, \ldots, r$ and let $B_{l(j-1)}$ consist of individuals who die at times $t^*_{l1}, \ldots, t^*_{l(j-1)}$.

Then Cox's partial likelihood is

$$\prod_{l=1}^{r} \prod_{j=1}^{d_l} \frac{\exp(\beta^{\mathsf{T}} z^*_{lj})}{\sum_{k \in R(t^*_{l1}) \setminus B_{l(j-1)}} \exp(z_k^{\mathsf{T}} \beta)}$$

$$= \prod_{l=1}^{r} \frac{\exp(\beta^{\mathsf{T}} z^*_{l\cdot})}{\prod_{j=1}^{d_l} [\sum_{k \in R(t_{l1})} \exp(z_k^{\mathsf{T}} \beta) - \sum_{k \in B_{l(j-1)}} \exp(z_k^{\mathsf{T}} \beta)]}$$

When we do not know the ordering of $t^*_{l1}, \ldots, t^*_{ld_l}$ we can not compute term $\sum_{k \in B_{l(j-1)}} \exp(z_k^\mathsf{T} \beta)$.

Breslow: simply ignore this sum. Resulting partial likelihood becomes

$$\prod_{l=1}^{r} \frac{\exp(\beta^\mathsf{T} z^*_{l\cdot})}{(\sum_{k \in R(t_{l1})} \exp(z_k^\mathsf{T} \beta))^{d_l}}$$

Efron: replace sum by $j-1$ times average, that is

$$\sum_{k \in B_{l(j-1)}} \exp(z_k^\mathsf{T} \beta) \approx (j-1) \frac{1}{d_l} \sum_{k=1}^{d_l} \exp(\beta^\mathsf{T} z^*_{lk})$$

Cox's discrete time proportional odds model

Reuse notation from actuarial estimate but introduce covariates:

$p_k(z) = P(\text{indiv. with covariates } z \text{ dies in } [u_{k-1}, u_k[\mid \text{alive at time } u_{k-1})$.

Cox proposed proportional odds model:

$$O_k(z) = \frac{p_k(z)}{1 - p_k(z)} = \frac{p_k(0)}{1 - p_k(0)} \exp(z^\mathsf{T}\beta) = O_k(0) \exp(z^\mathsf{T}\beta)$$

Let D_k be index set of d_k individuals who die in $[u_{k-1}, u_k[$.
Probability that precisely individuals in D_k die given risk set $R(u_{k-1})$ is

$$\prod_{l \in D_k} p_k(z_l) \prod_{l \in R(u_{k-1}) \setminus D_k} (1 - p_k(z_l)) = \prod_{l \in D_k} O_k(z_l) \prod_{l \in R(u_{k-1})} (1 - p_k(z_l))$$

Probability that d_k individuals die:

$$\sum_{\substack{A \subseteq R(u_{k-1}): \\ \#A = d_k}} \prod_{l \in A} O_k(z_l) \prod_{l \in R(u_{k-1})} (1 - p_k(z_l))$$

Discrete time partial likelihood

Partial likelihood based on probabilities that individuals in D_k die given d_k individuals die and given $R(u_{k-1})$.

Only consider intervals with $d_k > 0$

$$L(\beta) = \prod_{k:d_k>0} \frac{\exp(\sum_{l \in D_k} z_l^\mathsf{T} \beta)}{\sum_{\substack{A \subseteq R(u_{k-1}): \\ \#A = d_k}} \exp(\sum_{l \in A} z_l^\mathsf{T} \beta)}$$

Note: $O_k(0)$ plays the same role as $\exp(\alpha_i)$ in matched case control model.

Different approaches to handling ties vary regarding computational complexity. On modern computers all options usually feasible.