Regression in survival analysis: The Cox model

Motivation

Aim: study, say, survival conditional on covariates X,

$$S(t \mid X = x).$$

Cox model

The Cox proportional hazards model parameterizes the hazard function as

$$\alpha(t \mid \mathbf{X}_i) = \alpha_0(t)\, e^{\beta^T \mathbf{X}_i(t)},$$

where β is a vector of coefficients and $\mathbf{X}_i(t) = (X_{i1}(t), X_{i2}(t), \ldots, X_{ip}(t))$ are fixed or time-varying, predictable covariates.

Remember:

A statistical model $\mathcal{P} = \{ P_\eta : \eta \in \Gamma \}$.

Parametric model $\mathcal{P} = \{ P_\theta : \theta \in \Theta \}$ $\Theta \subseteq \mathbb{R}^k$ for a positive finite k.

Cox-model is semi-parametric because

- $\alpha_0(t)$ is non parametric, and
- $e^{\beta^T x_i(t)}$ is parametric.

Implications of the Cox model:

Def: the hazard ratio "HR" of two hazards $\alpha_1(t)$ and $\alpha_2(t)$ is $\frac{\alpha_1(t)}{\alpha_2(t)}$ $\left(\neq \frac{1-S_1(t)}{1-S_2(t)} \neq \frac{S_1(t)}{S_2(t)} \right)$

The HR in a Cox model is

$$\frac{\alpha(t|x_1)}{\alpha(t|x_2)} = e^{\beta^T(\mathbf{X}_1(t) - \mathbf{X}_2(t))}$$

Often assumed that $\mathbf{X}(t) = \mathbf{X}$.

Suppose two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ satisfy $x_{1,j} - x_{2,j} = 1$, $x_{1,i} = x_{2,i}$ $i \neq j$, then the HR is e^{β_j}.

Lecture 9

- Cox regression model
- Interpretation of regression coefficients, hazard ratios
- Collapsibility.

Partial likelihood

AIM (intuitively) get rid of $\alpha_0(t)$ and only consider β.

Construction:

$$\lambda_i(t) = Z_i(t)\,\alpha_0(t)\,e^{\beta^T \mathbf{x}_i(t)} = \alpha_0(t)\left[Z_i(t)\,e^{\beta^T \mathbf{x}_i(t)}\right]$$

$\mathbf{x}_i(t)$ is predictable.

$$N_\bullet(t) = \sum_{\ell=1}^{n} N_\ell(t), \quad \lambda_\bullet(t) = \sum_{\ell=1}^{n} \lambda_\ell(t) = \sum_{\ell=1}^{n} Z_\ell(t)\,\alpha_0(t)\,e^{\beta^T \mathbf{x}_\ell(t)}$$

Now, $\lambda_i(t) = \lambda_\bullet(t)\,\pi(i|t)$, where

$$\pi(i|t) = \frac{\lambda_i(t)}{\lambda_\bullet(t)} = \frac{Z_i(t)\,\cancel{\alpha_0(t)}\,e^{\beta^T \mathbf{x}_i(t)}}{\sum_{\ell=1}^{n} Z_\ell(t)\,\cancel{\alpha_0(t)}\,e^{\beta^T \mathbf{x}_\ell(t)}}$$

Def partial likelihood

Consider event times $T_1 < T_2 < T_3 \dots$ and let i_j be the index of the individual who has an event at T_j. The partial likelihood is:

$$\mathcal{L}(\beta) = \prod_{T_j} \pi(i_j \mid T_j) = \prod_{T_j} \left(\frac{Z_{i_j}(T_j)\, e^{\beta^T \mathbf{x}_{i_j}(T_j)}}{\sum_{l=1}^{n} Z_l(T_j)\, e^{\beta^T \mathbf{x}_l(T_j)}} \right) = \prod_{T_j} \frac{e^{\beta^T \mathbf{x}_{i_j}(T_j)}}{\sum_{l \in R_j} e^{\beta^T \mathbf{x}_l(T_j)}}$$

$$R_j = \{ l : Z_l(T_j) = 1 \}.$$

Result:

Let $\hat{\beta}$ be the value of β that maximizes $\mathcal{L}(\beta)$, called the "maximum partial likelihood estimator".

In large samples: $\hat{\beta} \sim N\left(\beta_0, \underline{I}(\beta_0)^{-1}\right)$, $\underline{I} = -\frac{\partial^2}{\partial \beta_h \, \partial \beta_l} \log L(\beta)$

"partial information matrix".

proof technique:
Show that $U(\beta) = \frac{\partial \log L(\beta)}{\partial \beta}$ is a martingale.

From a Cox model to survival:

$$\lambda_{\bullet}(t) = \sum_{\ell=1}^{n} \lambda_{\ell}(t) = \sum_{\ell=1}^{n} \underline{Z_\ell(t)}\, \underline{\alpha_0(t)}\, \underline{e^{\beta^T x_\ell(t)}}$$

Survival $N_{\bullet}(t)$ has intensity $\lambda_{\bullet}(t)$.

If β were known, we could consider a N-A "type" of estimator:

$$\hat{H}(t;\beta) = \int_0^t \frac{dN_{\bullet}(u)}{\sum_{\ell=1}^{n} Z_i(u) e^{\beta^T x_i(u)}}$$

Breslow estimator:

$$\hat{H}_0(t;\hat{\beta}) = \int_0^t \frac{dN_{\bullet}(u)}{\sum_{\ell=1}^{n} Z_i(u) e^{\hat{\beta}^T x_i(u)}} = \sum_{T_j \le t} \frac{1}{\sum_{\ell \in R_j} e^{\hat{\beta}^T x_i(T_j)}}$$

Thus, $\hat{H}(t;x_0) = \hat{H}_0(t) e^{\hat{\beta}^T x_0}$,

$$S(t|x_0) = \prod_{u \le t} \left(1 - dH(u|x_0)\right), \qquad \hat{S}(t|x_0) = \prod_{T_j \le t} \left(1 - \Delta\hat{H}(T_j|x_0)\right)$$

Model checking

Consider a cox model with fixed covariates.

$$\alpha(t \mid \mathbf{x}) = \alpha_0(t)\, e^{\beta^T \mathbf{x}}$$

$$-\log\left(S(t \mid \mathbf{x})\right) = \int_0^t \alpha(s \mid \mathbf{x})\, ds = \int_0^t \alpha_0(s)\, e^{\beta^T \mathbf{x}}\, ds$$

$$\log\left(-\log\left(S(t \mid \mathbf{x})\right)\right) = \log\left\{ \int_0^t \alpha_0(s)\, ds \right\} + \beta^T \mathbf{x}.$$

$$\log\left(-\log\left(S(t \mid \mathbf{x}_1)\right)\right) - \log\left(-\log\left(S(t \mid \mathbf{x}_2)\right)\right) = \beta^T \mathbf{x}_1 - \beta^T \mathbf{x}_2.$$

Ex: Collapsibility

Suppose I have estimates of $P(T^{a=1} > t \mid V=v)$ and $P(T^{a=0} > t \mid V=v)$ for a covariate vector V.

$$P(T^{a=1} > t) - P(T^{a=0} > t) = \sum_{v} \left[P(T^{a=1} > t \mid V=v) - P(T^{a=0} > t \mid V=v) \right] P(V=v)$$

Survival difference is collapsible.

$$\sum_{v} P(V=v) = 1.$$

Hazard ratio and collapsibility

$A \in \{0,1\}$, $Z \in [0,\infty)$ baseline covariate.

$r > 0$, $\alpha(t) > 0 \quad \forall t > 0$.

Laplace transform: $L(c) = \mathbb{E}\left(e^{-cZ}\right)$ for $c \in \mathbb{C}$.

$$\alpha(t \mid A=0, Z) = Z\alpha(t)$$

$$\alpha(t \mid A=1, Z) = rZ\alpha(t)$$

$$S(t \mid A=0) = \mathcal{L}\left(H(t)\right), \quad H(t) = \int_0^t \alpha(s)\,ds.$$

$$\alpha(t \mid A=0) = -\alpha(t)\,\frac{\mathcal{L}'(H(t))}{\mathcal{L}(H(t))}.$$

$$S(t \mid A=1) = \mathcal{L}\left(rH(t)\right)$$

Z is gamma distributed with mean 1 and variance δ. Then,

$$L(c) = \{1 + \delta c\}^{\frac{-1}{\delta}},$$

$$S(t) := \{1 + \delta H(t)\}^{\frac{-1}{\delta}}$$

$$\alpha(t \mid A=0) = \frac{\alpha(t)}{1 + \delta H(t)}$$

$$\alpha(t \mid A=1) = \frac{r\alpha(t)}{1 + r\delta H(t)}$$

Suppose $\delta = 1$.

$$\frac{\alpha(t \mid A=1)}{\alpha(t \mid A=0)} = r\,\frac{1 + H(t)}{1 + rH(t)} \neq r$$

for $r \neq 1$ at all $t > 0$.