5. Cox Regression

- This lecture's topics:
  - proportional hazards model
  - interpretation of parameters
  - partial likelihood
  - example
- Text: TG Chapter 3; FH Chapter 8; KP Chapter 4

Proportional Hazards Model

- Proposed by Cox (1972, JRSS-B), primarily to model the relationship between hazard function and covariates
  - most cited paper in statistics ($\approx 41,000$ as of April 2016)
  - one of the most cited in science
- Several extensions to more complex data structures
  - clustered failure time data
  - recurrent event data

Data Structure

- i = subject
- T_i = potential event time
- C_i = potential censoring time
- $X_i = T_i \wedge C_i$ = observed time

- $\Delta_i = I(T_i < C_i)$
- $N_i(t) = I(X_i \leq t, \Delta_i = 1)$
- $\mathbf{Z}_i(t)$ = covariate vector (possibly time-dependent)
- Observed data: $\{X_i, \Delta_i, \mathbf{Z}_i(\cdot)\} \sim i.i.d.$

Cox PH Model

- Cox model:

$$\lambda_i(t) = \lambda(t|\mathbf{Z}_i) = \lambda_0(t)\exp\{\boldsymbol{\beta}'\mathbf{Z}_i\}$$

- semiparametric model:
  - $\exp\{\boldsymbol{\beta}'\mathbf{Z}_i\}$, parametric assumption on covariate effects
  - multiplicative model
  - $\boldsymbol{\beta}$: $p \times 1$ vector, $p < \infty$
  - $\lambda_0(t)$, nonparametric; is ∞ dimensional
  - shape of hazard function is unspecified

- Due to nonparametric component, standard maximum likelihood theory does not apply

- Let Z_{ij} be the jth element of $\mathbf{Z}_i$
  - β_j = difference in log hazards
  - $\exp\{\beta_j\}$ = ratio of hazards; assumed constant for all t

- $\lambda_0(t)$: baseline hazard; common to all subjects,

$$\lambda_0(t) = \lambda_i(t|\mathbf{Z}_i = \mathbf{0}),$$

  where $\mathbf{0}$ is a vector of 0's

- The hazard ratio $\exp\{\beta_j\}$ is sometimes referred to as a *relative risk*
  - risk = probability, not a rate
  - hazard is a *rate*, not a probability
  - in ratio of hazards, time dimension cancels out

- Direction of effect:
  - $\beta_j > 0$: $\uparrow\ \lambda_i(t)$, $\downarrow\ S_i(t)$
  - $\beta_j < 0$: $\downarrow\ \lambda_i(t)$, $\uparrow\ S_i(t)$

- Magnitude of effect is easy to interpret w.r.t. $\lambda_i(t)$

- Cumulative hazard function,

$$\begin{aligned}\lambda_i(t) &= \lambda_0(t)\exp\{\boldsymbol{\beta}'\mathbf{Z}_i\} \\ \Lambda_i(t) &= \int_0^t \lambda_0(s)\exp\{\boldsymbol{\beta}'\mathbf{Z}_i\}ds \\ &= \Lambda_0(t)\exp\{\boldsymbol{\beta}'\mathbf{Z}_i\}\end{aligned}$$

- Survival function,

$$\begin{aligned}S_i(t) &= \exp\{-\Lambda_i(t)\} \\ &= \exp\{-\Lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i)\} \\ &= S_0(t)^{\exp\{\boldsymbol{\beta}'\mathbf{Z}_i\}}\end{aligned}$$

- By fitting a Cox model, one can readily interpret the multiplicative effect on the hazard
  - ex) randomized trial: treatment (Z_i=1) versus placebo (Z_i=0); $\widehat{\beta} = 0.405$ ($\exp(\widehat{\beta})$=1.5)
  - $\lambda_i(t)$ for treated patients is 50% more of that of the controls
  - irrespective of $\lambda_0(t)$

- Nevertheless, $\Lambda_0(t)$ is required in order to determine Z_i's effect on $S_i(t)$

- e.g., $S(t|Z_i = 0)$=0.95 vs. $S(t|Z_i = 1)$=0.93
- e.g., $S(t|Z_i = 0)$=0.70 vs. $S(t|Z_i = 1)$=0.59

Cox Model: Independent Censoring

- Independent censoring assumption is less stringent than in nonparametric estimation
- Assumption is often written as $T_i \perp C_i|\mathbf{Z}_i$:

$$\begin{aligned} & \lim_{\delta\to 0} \delta^{-1} P(t \le T_i < t+\delta | T_i \ge t, C_i \ge t, \mathbf{Z}_i) \\ = & \lim_{\delta\to 0} \delta^{-1} P(t \le T_i < t+\delta | T_i \ge t, \mathbf{Z}_i) \end{aligned}$$

- **Note**: C_i is allowed to depend on $\mathbf{Z}_i$.

Semiparametric PH Model: General

- General expression for multiplicative proportional hazards model:

$$\lambda_i(t) = \lambda_0(t) g(\boldsymbol{\beta}'\mathbf{Z}_i)$$

  - $g(x)$: link function, specified
  - $g(x) \ge 0$ for all x; twice differentiable
  - $g(x) = \exp(x)$, special case

- Other choices for link function (e.g., Self & Prentice, 1983):

$$\begin{aligned} g(x) &= 1+x \\ g(x) &= (1+x)^{-1} \\ g(x) &= \log(1+x) \end{aligned}$$

- **Notes**:
  - not all choices of $g(x)$ lead to clear interpretation of β_j
  - certain choices of $g(x)$ lead to numerical issues;

    e.g., likelihood is flat; local maxima, etc
  - $g(x) \neq \exp(x)$ has received little attention in the literature

Multiplicative Model

- Cox model is a *multiplicative* model
  - covariates assumed to affect survival probability by multiplying the baseline hazard
- Additive models have also been proposed
  - e.g., Lin & Ying (1994; *Bka*),

$$\lambda_i(t) = \lambda_0(t) + \boldsymbol{\beta}'\mathbf{Z}_i$$

  - e.g., Aalen (1989; *SIM*):

$$\lambda_i(t) = \beta_0(t) + \boldsymbol{\beta}_1'(t)\mathbf{Z}_i$$

  - less commonly used

Proportional Hazards Regression and Multiplicative Intensity Model

- Recall Counting process - martingale representation

$$N(t) = I(X \le t, \Delta = 1)$$

$$Y(t) = I(X \ge t)$$

$$M(t) = N(t) - \underbrace{\int_0^t \overbrace{Y(u)\lambda_0(u)e^{\beta' Z}}^{\text{intensity } l(u)} du}_{\text{cumulative intensity } A(t)}$$

$$\mathcal{F}_t = \sigma\{N(u), Y(u+), Z : 0 \le u \le t\}$$

- **Multiplicative Intensity Model**:

$$l(t) = Y(t)\lambda_0(t)e^{\beta' Z(t)}$$

- **Counting process**: $N(t)$ = Number of events of a specified type that have occurred by time t
  - $N(t)$ may take more than one jump
  - multiple infections, repeated breakdowns, hospital admissions
  - $EN(t) < \infty$

- **At-risk process**:

$$\begin{aligned} Y(t) &= \text{left-cont. process} \\ &= \begin{cases} 1 & \text{if failure can be observed at time } t \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

  - $Y(t)$ can be used to represent situation in which a subject enter and exit risk sets several times
  - $Y(t)$ may be 1 even after an observed failure

- **Covariate process**: $Z(t)$ = (bounded) predictable process
  - time-dependent treatment, risk factors
  - model checking and relaxing PH assumption
- **Baseline hazard function**: $\lambda_0(\cdot)$ = an arbitrary deterministic function
- **Filtration**:

$$\mathcal{F}_t = \sigma\{N(u), Y(u+), Z(u) : 0 \leq u \leq t)$$

- **Martingale**:

$$M(t) = N(t) - \int_0^t l(u)du.$$

- **Intensity function**:

$$E\{dN(t)|\mathcal{F}_{t-}\} = l(t)dt.$$

- **Data**:
  n indep. observations on $\{N(\cdot), Y(\cdot), Z(\cdot)\}$

Likelihood; conditional, marginal and partial likelihoods

- X = vector of observations; $f_X(x, \theta)$ = density of X
- θ = vector parameter; $\theta = (\beta', \phi')'$
- β = parameter of interest; ϕ = nuisance parameter
- *likelihood:* $f_X(x, \theta) = f_{W|V}(w|v, \theta) f_V(v, \theta)$
  - $X = (V', W')'$;
  - infinite-dimensional ϕ;
  - $f_{W|V}(w|v, \theta)$ does not involve $\phi \Rightarrow$ use $f_{W|V}(w|v, \beta)$ (*conditional likelihood*);
  - $f_V(v, \theta)$ does not involve $\phi \Rightarrow$ use $f_V(v, \beta)$ (*marginal likelihood*);

$$X = (V_1, W_1, V_2, W_2, \dots, V_K, W_K)$$

$$\begin{aligned}
f_X(x, \theta) &= f_{V_1, W_1, \dots, V_K, W_K}(v_1, w_1, \dots, v_K, w_K; \theta) \\
&= f_{V_1}(v_1; \theta) f_{W_1|V_1}(w_1|v_1; \theta) f_{V_2|V_1, W_1}(v_2|v_1, w_1; \theta) \\
&\quad \times f_{W_2|V_1, W_1, V_2}(w_2|v_1, w_1, v_2; \theta) \dots \\
&= \left\{ \prod_{i=1}^{K} f_{W_i|Q_i}(w_i|q_i; \theta) \right\} \left\{ \prod_{i=1}^{K} f_{V_i|P_i}(v_i|p_i; \theta) \right\}
\end{aligned}$$

- $P_1 = \phi, P_i = (V_1, W_1, \dots, V_{i-1}, W_{i-1})$
- $Q_1 = V_1, Q_i = (V_1, W_1, \dots, W_{i-1}, V_i)$
- $\prod_{i=1}^{K} f_{W_i|Q_i}(w_i|q_i; \theta)$ is free of $\phi \Rightarrow$ use $\prod_{i=1}^{K} f_{W_i|Q_i}(w_i|q_i; \beta)$ (*partial likelihood*)

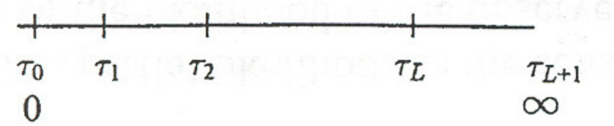
Partial & Marginal Likelihoods

- Focus on Proportional Hazards Model:
  i.e., (X_i, δ_i, Z_i) $i = 1, \cdots, n$ (n independent triplets)

$$\lambda(t|Z) = \underbrace{\lambda_0(t)}_{\text{unspecified}} e^{\beta' Z}, \qquad S(t|Z) = \{S_0(t)\}^{e^{\beta' Z}} \tag{1}$$

- Partial Likelihood: assume no ties, absolutely continuous failure distribution

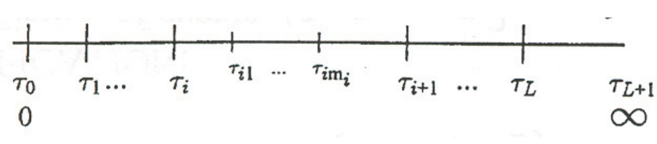
- Suppose there are L observed failures
  at $\tau_1 < \cdots < \tau_L$ (set $\tau_0 \equiv 0$ & $\tau_{L+1} \equiv \infty$)



- Let (i) be the label for individual failing at τ_i
  (set $(L+1) \equiv n+1$). Note $t_{(i)} = \tau_i$

- Covariates for L failures: $(Z_{(1)}, Z_{(2)}, \cdots, Z_{(L)})$. (Hereafter, condition on $\{Z_i : i = 1, \cdots, n\}$)

- Censorship times in $[\tau_i, \tau_{i+1})$:
  $(\tau_{i1}, \cdots, \tau_{i,m_i})$ with covariates $(Z_{(i,1)}, \cdots, Z_{(i,m_i)})$
  i.e., (i, j) is label for item censored at τ_{ij}



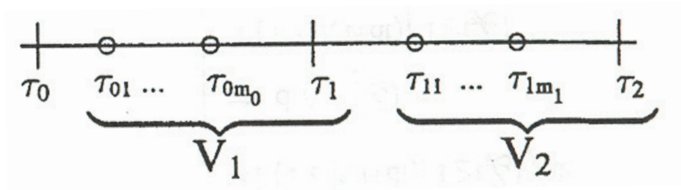
- The data can be divided into sets

$$(V_1, W_1, V_2, W_2, \cdots, V_{L+1}, W_{L+1})$$

where, for $i = 1, \cdots, L, L+1$

$$V_i = \{\tau_i, \tau_{i-1,j}, \quad (i-1, j) : j = 1, \cdots, m_{i-1}\}$$

$$\& \quad W_i = \{(i)\}$$



- Example:

id	a	b	c	d	e	f	g	h
x	1	2	5	3	6	10	11	∞
δ	0	1	1	0	0	1	0	

$\tau_1 = 2$	$\tau_2 = 5$	$\tau_3 = 10$	$\tau_4 \equiv \infty$
(1)=b	(2)=c	(3)=f	(4)=h
$V_1 = \{2, 1\}$	$V_2 = \{5, 3\}$	$V_3 = \{10, 6\}$	$V_4 = \{\infty, 11\}$
$W_1 = \{b\}$	$W_2 = \{c\}$	$W_3 = \{f\}$	$W_4 = \{h\}$

where $L = 3$

- GOAL: Build a likelihood on a subset of the full data set
  - carrying most of the information about β
  - carrying no information on nuisance parameters $\{\lambda_0(t) : t \geq 0\}$

- PROPOSAL: Generate likelihood of $\{W_1, \cdots, W_L\}$

- JUSTIFICATION:
  - Timing of events $\{\tau_1, \tau_2, \cdots, \tau_L\}$ can be explained by $\lambda_0(\cdot)$.
  - Censoring times and labels can be ignored if we assume non-informative censorship (independent censoring).

- So this is a partial likelihood in the sense that it is only part of the likelihood of the observed data.
  - If $Q_i \equiv (V_1, W_1, \cdots, V_{i-1}, W_{i-1}, V_i)$ and $\mathcal{F}\tau_i \equiv (Q_i, Z)$, the partial likelihood is:

$$\prod_{i=1}^{L} P(W_i = (i)|\mathcal{F}\tau_i) \quad \left\{ \begin{array}{l} \text{i.e., given the risk set at } \tau_i, \\ \text{and given event occurs at } \tau_i. \end{array} \right.$$

- Denote $R_i \equiv \{j : X_j \geq \tau_i\} \cdots$ risk set at τ_i
  Then, by the assumption of independent censoring,
  $P(W_i = (i)|\mathcal{F}\tau_i) =$

$$= \frac{P\{t_{(i)} \in [\tau_i, \tau_i + d\tau)|\mathcal{F}\tau_i\} \prod_{j \in R_i - (i)} \overbrace{P\{t_j \notin [\tau_i, \tau_i + d\tau)|\mathcal{F}\tau_i\}}^{=1-P\{t_j \in [\tau_i, \tau_i + d\tau)|\mathcal{F}\tau_i\}}}{\sum_{l \in R_i} \left[P\{t_l \in [\tau_i, \tau_i + d\tau)|\mathcal{F}\tau_i\} \prod_{j \in R_i - l} P\{t_j \notin [\tau_i, \tau_i + d\tau)|\mathcal{F}\tau_i\} \right]}$$

$$= \frac{d\Lambda(\tau_i|Z_{(i)}) \prod_{j \in R_i - (i)} \{1 - \overbrace{d\Lambda(\tau_i|Z_j)}^{=0}\}}{\sum_{l \in R_i} \left[d\Lambda(\tau_i|Z_l) \prod_{j \in R_i - l} \{1 - \underbrace{d\Lambda(\tau_i|Z_j)}_{=0}\} \right]} \tag{2}$$

$$= \frac{\lambda(\tau_i|Z_{(i)})}{\sum_{l \in R_i} \lambda(\tau_i|Z_l)} \qquad \begin{array}{l} |d\Lambda(t|Z)/dt = \lambda(t|Z) \\ | = P\{T \in [t, t + dt)|T \geq t, Z\}/dt \end{array}$$

$$\overset{(1)}{=} \frac{\exp(\beta' Z_{(i)})}{\sum_{l \in R_i} \exp(\beta' Z_l)} \qquad \begin{array}{l} | \Rightarrow d\Lambda(t|Z) = \\ |P\{T \in [t, t + dt)|T \geq t, Z\} \end{array}$$

Thus, the Partial Likelihood is:

$$\prod_{i=1}^{L} \frac{\exp(\beta' Z_{(i)})}{\sum_{l \in R_i} \exp(\beta' Z_l)} \tag{3}$$

- In summary, *Partial likelihood:*:

$$L(\beta) = \prod_{i=1}^{L} \frac{\exp\{\beta' Z_{(i)}\}}{\sum_{l \in R_i} \exp\{\beta' Z_l\}}$$

Note: unspecified $\lambda_0(\cdot)$ + noninformative censoring $\Rightarrow$
$\prod_{i=1}^{L} f_{V_i|P_i}(v_i|p_i;\theta)$ contains little or no information about β.

- *Counting process notation:*

$$L(\beta) = \prod_{i=1}^{n} \prod_{t \geq 0} \left\{ \frac{\exp(\beta' Z_i)}{\sum_{j=1}^{n} Y_j(t) \exp(\beta' Z_j)} \right\}^{dN_i(t)}$$

$$\text{where } \; dN_i(t) = \begin{cases} 1 & \text{if } N_i(t) - N_i(t-) = 1 \\ 0 & \text{otherwise} \end{cases}$$

- *Maximum partial likelihood estimator (MPLE):* $L(\hat{\beta}) = \max_\beta L(\beta)$ (using Newton-Raphson (NR) algorithm)

  – Specifically, the log partial likelihood is then

$$l(\beta) = \sum_{i=1}^{n} \int_0^\infty \left[Y_i(t) Z_i \beta - \log \left(\sum_{j=1}^{n} Y_j(t) \exp(\beta' Z_j) \right) \right] dN_i(t)$$

  – The score vector, $U(\beta)$, can be obtained by differentiating $l(\beta)$ w.r.t. β:

$$U(\beta) = \sum_{i=1}^{n} \int_0^\infty \left\{ Z_i - \bar{Z}(\beta, t) \right\} dN_i(t)$$

  where $\bar{Z}(\beta, t)$ is a weighted mean of Z over those observations still at risk at time t,

$$\bar{Z}(\beta, t) = \frac{\sum_{i=1}^{n} Y_i(t) Z_i \exp(\beta' Z_i)}{\sum_{i=1}^{n} Y_i(t) \exp(\beta' Z_i)}.$$

– The information matrix, $\mathcal{I}(\beta)$, is the negative second derivative where

$$\mathcal{I}(\beta) = \sum_{i=1}^{n} \int_0^\infty V(\beta, t) dN_i(s),$$

and

$$V(\beta, t) = \frac{\sum_{i=1}^{n} Y_i(t) \exp(\beta' Z_i)\{Z_i - \bar{Z}(\beta, t)\}'\{Z_i - \bar{Z}(\beta, t)\}}{\sum_{i=1}^{n} Y_i(t) \exp(\beta' Z_i)},$$

a weighted variance of Z at time t.

– Then, the MPLE, $\hat{\beta}$, is found by solving the partial likelihood equation:

$$U(\hat{\beta}) = 0.$$

– Under some regularity conditions, $\hat{\beta}$ is consistent and asymptotically normally distributed with mean β and variance $\mathrm{E}\{\mathcal{I}(\beta)\}^{-1}$ (will be shown later...)

– The NR algorithm to solve the partial likelihood equation: Compute iteratively

$$\hat{\beta}^{(n+1)} = \hat{\beta}^{(n)} + \mathcal{I}^{-1}(\hat{\beta}^{(n)}) U(\hat{\beta}^{(n)})$$

until convergence (requires an initial value $\hat{\beta}^{(0)}$).

– **Note**:

1. (incredibly) Robust algorithm!
2. $\hat{\beta}^{(0)} = 0$ usually works.

Cox Proportional Hazards Model

- Cox model:

$$\lambda_i(t) = \lambda(t|\mathbf{Z}_i) = \lambda_0(t)\exp\{\boldsymbol{\beta}'\mathbf{Z}_i\} = \lambda_0(t)\exp(\beta_1 Z_{i1} + \cdots + \beta_k Z_{ik})$$

  Equivalently:

$$\begin{aligned}\log\lambda(t|\mathbf{Z}_i) &= \log(\lambda_0(t) + \beta_1 Z_{i1} + \cdots + \beta_k Z_{ik})\\ S(t|\mathbf{Z}_i) &= [S_0(t)]^{\exp(\beta_1 Z_{i1}+\cdots+\beta_k Z_{ik})}\end{aligned}$$

- **Note**:
  - $\lambda_0(t) = \lambda(t|Z_1 = Z_2 = \cdots = Z_k = 0)$
  - $\exp(\beta_1 Z_1 + \cdots + \beta_k Z_k) = \text{RR} = \underbrace{\dfrac{\lambda(t|Z_1, \cdots Z_k)}{\lambda(t|Z_1 = \cdots = Z_k = 0)}}$

    relative risk of hazard of death comparing covariates values $Z_1, \cdots Z_k$ to $Z_1 = \cdots = Z_k = 0$

- Interpreting Cox Model Coefficients: β_k is the log RR (hazard ratio) for a unit change in Z_k, given all other covariates remain constant.

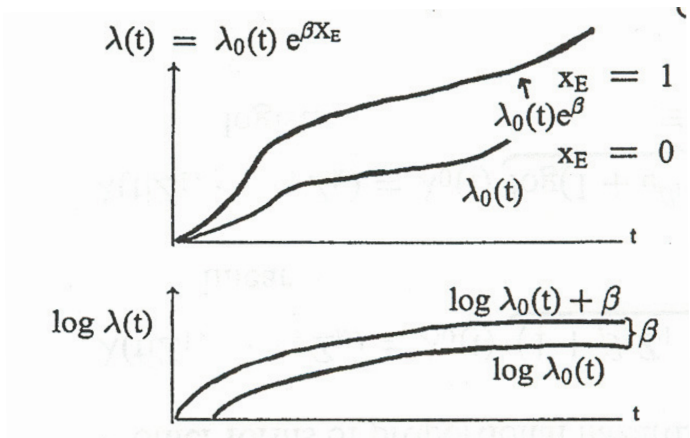
i.e.
$$\begin{aligned}\frac{\lambda(t|Z_1, \cdots, Z_{k'} + 1, \cdots, Z_k)}{\lambda(t|Z_1, \cdots, Z_{k'}, \cdots, Z_k)} &= \exp(\beta_1 \cdot 0 + \cdots + \beta_{k'} \cdot (Z_{k'} + 1 - Z_{k'}) + \cdots + \beta_k \cdot 0)\\ &= \exp(\beta_{k'})\end{aligned}$$

- The RR comparing 2 sets of values for the covariates $(Z_1, Z_2, \cdots Z_k)$ vs $(Z'_1, Z'_2, \cdots Z'_k)$

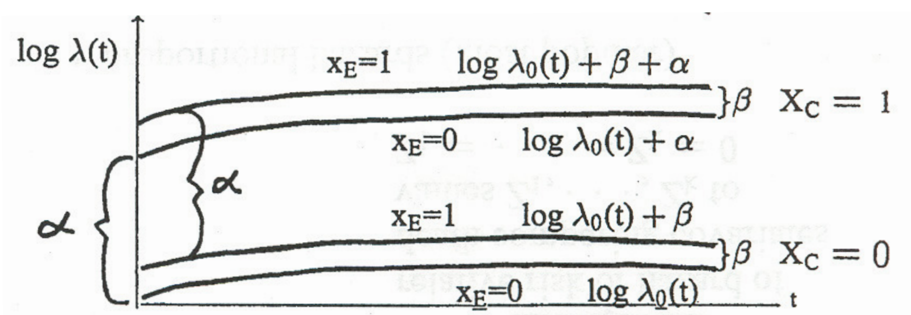
$$\text{RR} = \frac{\lambda(t|Z_1, \cdots, Z_k)}{\lambda(t|Z'_1, \cdots, Z'_k)} = \exp\{\beta_1(Z_1 - Z'_1) + \cdots + \beta_k(Z_k - Z'_k)\}$$

Cox Model Examples

1. One dichotomous covariate $X_E = \begin{cases} 1 & \text{exposed} \\ 0 & \text{not} \end{cases}$



2. Two covariates: Dichotomous Exposure X_E as above and Dichotomous Confounder $X_C = \begin{cases} 1 & \text{level 2} \\ 0 & \text{level 1} \end{cases}$



$$\lambda(t) = \lambda_0(t) \exp(\alpha X_C + \beta X_E)$$

$$\log \lambda(t) = \log \lambda_0(t) + \alpha X_C + \beta X_E$$

Comparison of Nested Models

- Nested models:

$$\begin{cases} \underline{\text{Full Model:}} \quad \lambda(t) = \lambda_0(t)\exp(\beta_1 Z_1 + \cdots + \beta_p Z_p + \beta_{p+1} Z_{p+1} + \cdots + \beta_k Z_k) \\ \underline{\text{Reduced Model:}} \quad \lambda(t) = \lambda_0(t)\exp(\beta_1 Z_1 + \cdots + \beta_p Z_p) \end{cases}$$

- To test:

H_0 : Reduced Model $\Leftrightarrow H_0 : \beta_{p+1} = \cdots = \beta_k = 0$ vs.

H_A : Full Model $\Leftrightarrow H_A :\neq$ somewhere

Use the **partial likelihood ratio statistic:**

$$X^2_{Cox} = -2[\log\ PL(\text{Reduced model}) - \log\ PL(\text{Full Model})]$$

Under H_0 : Reduced model and when n is large

$$X^2_{Cox} \sim \chi^2_{\underline{k-p}},\ k-p \text{ is \# parameters set to zero by } H_0.$$

Estimating Survival and Cumulative Hazard Functions

- Given a combination of covariate values $Z_1, \cdots Z_k$ and coefficient estimates $\widehat{\beta}_1, \cdots, \widehat{\beta}_k$ (t_i are ordered failure times in the data set),

$$\widehat{\Lambda}_0(t;\widehat{\boldsymbol{\beta}}) = \int_0^t \frac{\sum_{i=1}^n dN_i(s)}{\sum_{i=1}^n Y_i(s)\exp(\widehat{\boldsymbol{\beta}}' Z_i)} = \sum_{i:t_i \le t} \frac{D_i}{\sum\limits_{j \in R_i} \exp(\widehat{\beta}_1 Z_{j1} + \cdots + \widehat{\beta}_k Z_{jk})}$$

$$\begin{aligned} \widehat{S}_0(t) &= e^{-\widehat{\Lambda}_0(t)} \\ \widehat{\Lambda}(t|Z_1,\cdots,Z_k) &= \widehat{\Lambda}_0(t) e^{\widehat{\beta}_1 Z_1 + \cdots + \widehat{\beta}_k Z_k} \\ \widehat{S}(t|Z_1,\cdots,Z_k) &= [\widehat{S}_0(t)]^{e^{\widehat{\beta}_1 Z_1 + \cdots + \widehat{\beta}_k Z_k}} \end{aligned}$$

Example: PBC Data

- Fitting a Cox model with 5 covariates:
  - Death (`status==2`) is the event of interest and liver transplantation (`status==1` will be treated as censoring).
  - 5 covariates: age, edema, bilirubin (log scale), protime (log scale), and albumin (log scale)
  - R code and result

```
# Loading the 'survival' package;
library(survival)

# The data set 'pbc' is now ready to use;
data(pbc)

fit.pbc <- coxph(Surv(time,status==2) ~ age + edema + log(bili) + log(protime)
+ log(albumin), data=pbc)
print(fit.pbc)

---------------------------------------------------------------------------


                coef exp(coef) se(coef)     z       p
age           0.0396    1.0404  0.00767  5.16 2.4e-07
edema         0.8963    2.4505  0.27141  3.30 9.6e-04
log(bili)     0.8636    2.3716  0.08294 10.41 0.0e+00
log(protime)  2.3868   10.8791  0.76851  3.11 1.9e-03
log(albumin) -2.5069    0.0815  0.65292 -3.84 1.2e-04

Likelihood ratio test=231  on 5 df, p=0  n= 416, number of events= 160
   (2 observations deleted due to missingness)
```

  - In `Surv(time, status==2)`, `status==2` specifies the *event* code.
  - Interpretation:

bilirubin : the most important variable ($z = 10.41$ with $p = 0.0e+00$); each 1 point change in log(bilirubin) is associated with a 2.4 fold increase in a patient's risk.

age : Each additional year of age is associated with an estimated 4% increase in risk (Q: How about an additional decade?)

edema : The estimated risk of death in patients with severe edema (`edema==1`) is 2.45 times that of patients with no edema (Q: How about the effect of moderate edema (`edema==0.5`)?)

prothrombin time : The risk of accruing for each unit in log(prothrombin time) is the greatest, 10.9, but the overall impact on study patient is much less than bilirubin or age

albumin : The estimated coefficient is negative (in log scale).

– SAS code and output

```
PROC PHREG DATA=SURV.PBC;
MODEL time*status(0 1) = age edema log_bili log_prot log_albumin / TIES=efron;

log_bili = log(bili);
log_prot = log(protime);
log_albumin = log(albumin);
RUN;

--------------------------------------------------------------------------
```

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
age	1	0.03961	0.00767	26.6549	<.0001	1.040
edema	1	0.89631	0.27141	10.9061	0.0010	2.451
log_bili	1	0.86355	0.08294	108.4020	<.0001	2.372
log_prot	1	2.38684	0.76851	9.6460	0.0019	10.879
log_albumin	1	-2.50693	0.65292	14.7424	0.0001	0.082

Stratification

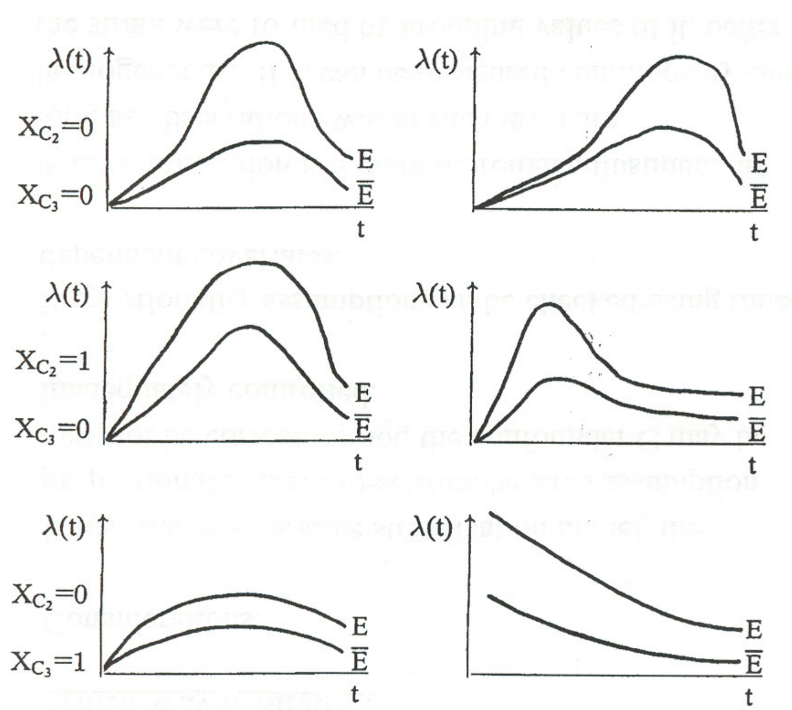
- <u>Two Ways to Stratify</u>

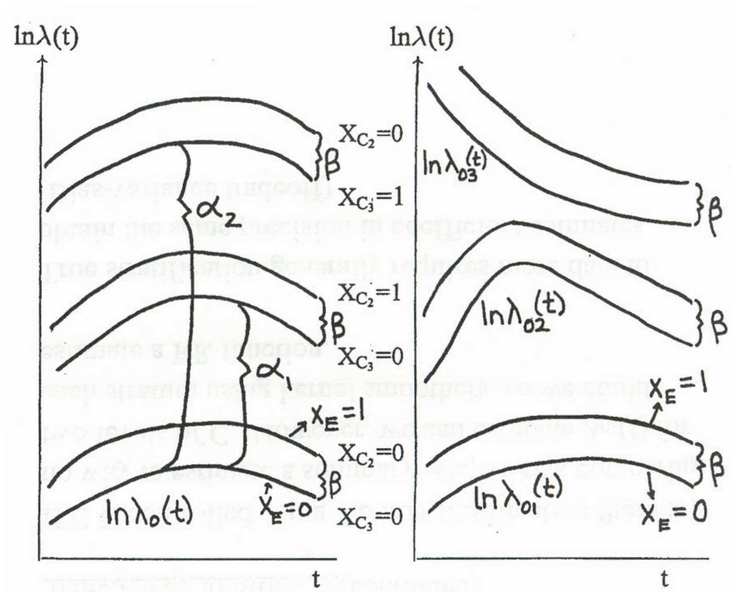
  Suppose a confounder C has 3 levels on which we would like to stratify when comparing $\lambda(t|E)$ and $\lambda(t|\bar{E})$.

  How? $X_E = \begin{cases} 1 & E \text{ (exposed)} \\ 0 & \bar{E} \text{ (not exposed)} \end{cases}$

- <u>Cox Model Stratification</u>

Dummy Variable:	'True' Stratification:
$\lambda(t) = \lambda_0(t)\exp(\alpha_1 X_{C_2} + \alpha_2 X_{C_3} + \beta X_E)$	$\lambda(t) = \lambda_{0i}(t)\exp(\beta X_E) \quad i = 1, 2, 3$
$\log\lambda(t) =$ $\log\lambda_0(t) + \alpha_1 X_{C_2} + \alpha_2 X_{C_3} + \beta X_E$	$\log\lambda(t) = \log\lambda_{0i}(t) + \beta X_E$





- Which Way to Stratify?

1. Under dummy variable stratification model, the proportional stratum-to-stratum hazards assumption may not be correct. If not, the confounder C may be inadequately controlled.

2. Proportionality assumption can be checked using time-dependent covariates.

3. True stratification is a more thorough adjustment, as long as observations within each level are homogeneous. If C can be measured continuously and the strata were formed by grouping values of it, better control for C might be achieved with continuous (could be time-dependent) covariate adjustment.

4. If C is controlled using the true stratification there is no way to estimate one summary relative risk comparing two levels of C. However, we can estimate $\lambda_{0i}(t)$ for each stratum then we can estimate a RR function.

5. True stratification generally requires more data to obtain the same precision in coefficient estimates.

- "True" stratification:

$\lambda(t) = \lambda_{0i}(t)\exp(\beta_E X_E)$

level 1 of C $\quad \lambda(t) = \lambda_{01}(t)\exp(\beta_E)$
level 2 of C $\quad \lambda(t) = \lambda_{02}(t)\exp(\beta_E)$

$RR(t) = \lambda_{01}(t)/\lambda_{02}(t) \qquad \widehat{RR}(t) = \widehat{\lambda_{01}}(t)/\widehat{\lambda_{02}}(t)$

- Dummy stratification:

$\lambda(t) = \lambda_0(t)\exp(\beta_2 X_{C_2} + \beta_3 X_{C_3} + \beta_E X_E)$

level 1 of C, exposed: $\quad \lambda(t) = \lambda_0(t)\exp(\beta_E)$
level 2 of C, exposed: $\quad \lambda(t) = \lambda_0(t)\exp(\beta_2 + \beta_E)$

$RR = \exp(-\beta_2)$

- Likelihood function?

- Advantage and disadvantage of stratification:

Advantage : the most general adjustment for a confounding variable

Disadvantage : no direct estimate of the importance of the strata effect

- Example 1 (R): true stratification based on the presence/absence of ascites

```
coxph(Surv(time, status==2) ~ age + edema + log(bili) + log(protime)
+ log(albumin) + strata (ascites), data=pbc)

-------------------------------------------------------------------------

                  coef exp(coef) se(coef)     z       p
```

```
age            0.0311     1.032  0.00907  3.43 0.00061
edema          0.6020     1.826  0.32060  1.88 0.06000
log(bili)      0.8683     2.383  0.10060  8.63 0.00000
log(protime)  3.0277    20.650  1.03933  2.91 0.00360
log(albumin) -2.9766     0.051  0.78093 -3.81 0.00014

Likelihood ratio test=146  on 5 df, p=0  n= 312, number of events= 125
   (106 observations deleted due to missingness)
```

Example 1 (SAS):

```
PROC PHREG DATA=SURV.PBC;
MODEL time*status(0 1) = age edema log_bili log_prot log_albumin / TIES=efron;
STRATA ascites;

log_bili = log(bili);
log_prot = log(protime);
log_albumin = log(albumin);
RUN;

---------------------------------------------------------------------------
```

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
age	1	0.03107	0.00907	11.7367	0.0006	1.032
edema	1	0.60245	0.32058	3.5316	0.0602	1.827
log_bili	1	0.86828	0.10060	74.4931	<.0001	2.383
log_prot	1	3.02834	1.03931	8.4902	0.0036	20.663
log_albumin	1	-2.97644	0.78095	14.5260	0.0001	0.051

- Strata by covariate interactions: it is not always reasonable to assume that the effect of every covariate is constant across strata

  Example : In a placebo-controlled drug trial, stratified by institution, results yield a coefficient for treatment of -0.22, showing a 20% overall improvement in death rate for the new therapy. Are we willing to assume that treatment is giving a uniform 20% improvement to each?

- If all of the covariate by strata interaction terms are added to a model, what will happen?

- Example 2 (R): Testing an age by strata interaction for the PBC model stratified on edema

```
coxph(Surv(time, status==2) ~ log(bili) + age*strata (edema), data=pbc)

---------------------------------------------------------------------------

                             coef exp(coef) se(coef)      z       p
log(bili)                  0.9632      2.62   0.0849 11.345 0.0e+00
age                        0.0355      1.04   0.0088  4.038 5.4e-05
age:strata(edema)edema=0.5 0.0215      1.02   0.0235  0.917 3.6e-01
age:strata(edema)edema=1   0.0727      1.08   0.0321  2.265 2.4e-02

Likelihood ratio test=148  on 4 df, p=0  n= 418, number of events= 161
```

- Example 2 (SAS):

```
PROC PHREG DATA=SURV.PBC;
MODEL time*status(0 1) =  log_bili age1 age2 age3 / TIES=efron;
STRATA edema;
INTERACT: test age1 = age2, age1 = age3;

log_bili = log(bili);
age1 = age*(edema=0);
age2 = age*(edema=0.5);
age3 = age*(edema=1);

RUN;

---------------------------------------------------------------------------
                        Analysis of Maximum Likelihood Estimates

                       Parameter    Standard                                Hazard
   Parameter   DF      Estimate        Error    Chi-Square    Pr > ChiSq     Ratio

   log_bili     1       0.96313      0.08490      128.6986        <.0001     2.620
   age1         1       0.03552      0.00880       16.3077        <.0001     1.036
   age2         1       0.05707      0.02175        6.8840        0.0087     1.059
   age3         1       0.10818      0.03085       12.2974        0.0005     1.114

                            Linear Hypotheses Testing Results

                                         Wald
                       Label        Chi-Square        DF     Pr > ChiSq

                       INTERACT         5.4858         2         0.0644
```

Test statistics

- The standard asymptotic likelihood inference tests, Wald, score, and likelihood ratio (LR), still can be applied for the Cox partial likelihood.

- Test statistics for testing $H_0 : \beta = \beta^{(0)}$:

  1. The likelihood ratio test (LRT) is

     $$2\{l(\hat{\beta}) - l(\beta^{(0)}\}$$

     where $l(\beta)$ is the log partial likelihood and $\hat{\beta}$ is the partial likelihood estimator of β.

  2. The Wald test is

     $$(\hat{\beta} - \beta^{(0)})'\mathcal{I}(\hat{\beta})(\hat{\beta} - \beta^{(0)}).$$

     Note: For $p = 1$, this reduces to $\hat{\beta}/SE(\hat{\beta})$.

  3. The score test statistic

     $$U(\beta^{(0)})'\mathcal{I}(\beta^{(0)})^{-1}U(\beta^{(0)}).$$

- **Note**: The score test statistic can be approximated by the first iteration of the NR algorithm:

  $$\hat{\beta}^1 - \beta^{(0)} = -U'(\beta^{(0)})\mathcal{I}(\beta^{(0)})^{-1}.$$

- Under H_0, all three tests $\sim \chi^2(p)$ asymptotically.

- Their finite sample properties may differ; in general, the LRT is the most reliable, the Wald test is the least...

- Simulation results comparing the 3 test statistics: $n = 100$ with a single binary covariate

β	$\hat{\beta}$	LRT	Score	Wald	Max/Min
0.00	-0.23	0.79	0.79	0.78	1.004
0.25	0.26	1.24	1.25	1.25	1.009
0.50	0.79	8.52	8.52	8.12	1.049
1.00	1.14	20.87	21.93	19.88	1.103
2.00	2.79	83.44	89.19	60.85	1.466

  - Three tests are similar for smaller values of β;
  - No statistically different conclusion even at the largest value;

- When $p = 1$ and the single covariate is categorical, the score test is identical to the log-rank test.

- PBC example: Recall the example of fitting a Cox model with 5 covariates. Below is a R code for a global test ($H_0 : \beta_1 = \ldots = \beta_5 = 0$):

```
fit.pbc <- coxph(Surv(time,status==2) ~ age + edema + log(bili) + log(protime)
+ log(albumin), data=pbc)
summary(fit.pbc)

---------------------------------------------------------------------------

                  coef exp(coef)  se(coef)      z Pr(>|z|)
age           0.039609  1.040404  0.007672  5.163 2.43e-07 ***
edema         0.896311  2.450547  0.271410  3.302 0.000959 ***
log(bili)     0.863551  2.371566  0.082941 10.412  < 2e-16 ***
log(protime)  2.386839 10.879054  0.768509  3.106 0.001898 **
log(albumin) -2.506923  0.081519  0.652916 -3.840 0.000123 ***
---
Likelihood ratio test= 231  on 5 df,   p=0
Wald test            = 234.2  on 5 df,   p=0
Score (logrank) test = 301.8  on 5 df,   p=0
```

Check below for a related SAS code and output:

```
PROC PHREG DATA=SURV.PBC;
MODEL time*status(0 1) = age edema log_bili log_prot log_albumin / TIES=efron;
STRATA ascites;
```

```
log_bili = log(bili);
log_prot = log(protime);
log_albumin = log(albumin);
RUN;

--------------------------------------------------------------------------

                        Testing Global Null Hypothesis: BETA=0

                 Test                 Chi-Square       DF     Pr > ChiSq

                 Likelihood Ratio       230.9751        5         <.0001
                 Score                  301.8424        5         <.0001
                 Wald                   234.1454        5         <.0001
```

For testing an individual variable, say bilirubin,

- the Wald test is simple; the result is already printed (10.412 with $p < 0.001$)

```
log(bili)     0.863551  2.371566  0.082941 10.412  < 2e-16 ***
```

- A LRT for bilirubin requires the data set be refit without `log(bili)`.

```
fit.pbc.nobili <- coxph(Surv(time,status==2) ~ age + edema + log(protime)
                    + log(albumin), data=pbc)
summary(fit.pbc.nobili)

--------------------------------------------------------------------------

  n= 416, number of events= 160
   (2 observations deleted due to missingness)

                  coef exp(coef)  se(coef)      z Pr(>|z|)
age           0.028713  1.029129  0.007793  3.685 0.000229 ***
edema         1.330363  3.782417  0.272586  4.881 1.06e-06 ***
log(protime)  3.073509 21.617630  0.685675  4.482 7.38e-06 ***
log(albumin) -3.491136  0.030466  0.599770 -5.821 5.86e-09 ***
---
Likelihood ratio test= 127.1  on 4 df,   p=0
Wald test            = 153.1  on 4 df,   p=0
Score (logrank) test = 180.1  on 4 df,   p=0
```

The LRT statistic = $(231 - 127.1) = 103.9$ on 1 degree of freedom.

– For the score test, first fit a model without `log(bili)` to obtain an appropriate initial values. Then, fit a model with all 5 covariates while specifying the appropriate initial values and setting the number of iteration to 0.

```
fit.pbc.bili0 <- coxph(Surv(time,status==2) ~ log(bili) + age + edema
      + log(protime) + log(albumin), data=pbc,
      init=c(0,coef(fit.pbc.nobili)), iter=0)
fit.pbc.bili0$score
```

```
----------------------------------------------------------------------------

[1] 116.6826
```

Note: `PROC PHREG` in SAS does not support initial values, so this test is not directly available in SAS.

Handling ties

- Real data sets often contain tied event times.
- When do we have ties?
  1. Continuous event times are grouped into intervals.
  2. Event time scale is discrete.
- Four commonly used ways of handling ties: 1) Breslow approximation, 2) Efron approximation, 3) Exact partial likelihood, and 4) Averaged likelihood
- When the underlying time is continuous but ties are generated due to a grouping, the contribution to the partial likelihood for the ith event at time t_i is

$$\frac{\exp(\beta' Z_i)}{\sum_{j \in R_i} Y_j(t_i) \exp(\beta' Z_j)}$$

  Two commonly used methods are

  1. **Breslow approximation**
  2. **Efron approximation**

  – Example: Assume 5 subjects are at risk of dying at time t and two die at the same time t (because of grouping of time) If the time data had been more precise, then the first two terms in the likelihood would be either

$$\left(\frac{\exp(\beta' Z_1)}{\sum_{j=1}^{5}\exp(\beta' Z_j)}\right)\left(\frac{\exp(\beta' Z_2)}{\sum_{j=2}^{5}\exp(\beta' Z_j)}\right)$$

or

$$\left(\frac{\exp(\beta' Z_2)}{\sum_{j=1}^{5}\exp(\beta' Z_j)}\right)\left(\frac{\exp(\beta' Z_1)}{\sum_{j=1,j\neq 2}^{5}\exp(\beta' Z_j)}\right)$$

but do we know which one is the case?

– Strategy: use the average of the terms or some approximation to the average

Note: The product of the numerators remain constant.

– **Breslow approximation**:

1. Breslow (1972, JRSS-C), Peto (1972, JRSS-C);
2. Simplest so easy to program;
3. Least accurate but fast;
4. Default option in most statistical software;
5. Using the complete sum $(\sum_{j=1}^{5}\exp(\beta' Z_j))$ for both denominators;
6. Counting failed individuals more than once in the denominator;
7. Producing β estimates too close to 0 (conservative bias).

– **Efron approximation**:

1. Quite accurate unless the # of tied events / size of risk set is extremely large;
2. As fast as the Breslow method;
3. Default option in `R`;

4. Using the average denominator $(0.5\exp(\beta' Z_1) + 0.5\exp(\beta' Z_2) + \sum_{j=3}^{5} \exp(\beta' Z_j))$ in the second term
5. **Q**: what if there were 3 tied deaths out of n subjects?

- For genuinely discrete data,
  - the contribution to the likelihood from d subjects with tied events out of n individual at risk is

$$\frac{r_1 r_2 \cdots r_d}{\sum_{S(d,n)} r_{k_1} r_{k_2} \cdots r_{k_d}},$$

    where $r_i = \exp(\beta' Z_i)$ and $S(d,n)$ denotes the set of all possible selections.

  - For the example above (2 events with 5 at risk), there are 10 unique pairs and the likelihood term is

$$\frac{r_1 r_2}{r_1 r_2 + r_1 r_3 + \cdots + r_4 r_5}.$$

  - Cox model for discrete-time data (proportional odds model):

$$\begin{aligned} \frac{\lambda_i(t)}{1-\lambda_i(t)} &= \frac{\lambda_0(t)}{1-\lambda_0(t)} \exp(\beta' Z_i) \\ \Leftrightarrow \log \frac{\lambda_i(t)}{1-\lambda_i(t)} &= \alpha_t + \beta' Z_i. \end{aligned}$$

    where $\lambda_i(t) = \Pr(T_i = t | T_i \geq t)$ and α_t is baseline log-odds of $\lambda_i(t)$.

– **Exact partial likelihood**

1. Exhaustive enumeration of the possible risk sets at each tied death time (not an approximation);

2. could be very time-consuming;

- Example for handling ties

  – The PBC data set has very few tied death times (5 tied pairs).
  – Assume that the times were recorded quarterly instead of in days.
  – R code and result:

```
# Creating tied observed times by recording time quarterly instead of in days.

# time2: observed time recorded quarterly
pbc$time2 = round(pbc$time/91.25)
# printing out the frequencies of times at each quarter
table(pbc$time2)

 0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
 3  8 10  4  7  3  5  8 11 14  9 13 15 18 14 18 19 11 14 12  9 10  9  8 13 12 10 11 14

 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 53
 8  7  10  2  6  7  7  4  4  8  7  2  2  5  2  5  3  6  1  1  6  2  1

# Using Efron method: The Efron method is default in R.
fit.pbc.efron <- coxph(Surv(time2, status==2) ~ trt + age + sex + edema + bili,
   ties="efron", data=pbc)
summary(fit.pbc.efron)

--------------------------------------------------------------------------------

  n= 312, number of events= 125
   (106 observations deleted due to missingness)
           coef exp(coef)  se(coef)      z Pr(>|z|)
treat -0.042191  0.958686  0.188174 -0.224   0.8226
age    0.035726  1.036371  0.009158  3.901 9.59e-05 ***
sex   -0.541119  0.582096  0.245414 -2.205   0.0275 *
edema  1.658801  5.253006  0.297226  5.581 2.39e-08 ***
bili   0.130509  1.139408  0.014525  8.985  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
      exp(coef) exp(-coef) lower .95 upper .95
treat    0.9587     1.0431    0.6630    1.3863
age      1.0364     0.9649    1.0179    1.0551
sex      0.5821     1.7179    0.3598    0.9417
edema    5.2530     0.1904    2.9337    9.4060
bili     1.1394     0.8776    1.1074    1.1723

Concordance= 0.803  (se = 0.029 )
Rsquare= 0.351   (max possible= 0.984 )
Likelihood ratio test= 135.1  on 5 df,   p=0
Wald test            = 171.1  on 5 df,   p=0
Score (logrank) test = 264.4  on 5 df,   p=0

# Using Breslow method
fit.pbc.breslow <- coxph(Surv(time2, status==2) ~ trt + age + sex + edema + bili,
   ties="breslow", data=pbc)
summary(fit.pbc.breslow)

--------------------------------------------------------------------------------

  n= 312, number of events= 125
   (106 observations deleted due to missingness)

           coef exp(coef)  se(coef)      z Pr(>|z|)
treat -0.039970  0.960818  0.188000 -0.213 0.831633
age    0.035191  1.035818  0.009166  3.839 0.000123 ***
sex   -0.533023  0.586828  0.245302 -2.173 0.029786 *
edema  1.601095  4.958457  0.295940  5.410  6.3e-08 ***
bili   0.127727  1.136243  0.014510  8.803  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

      exp(coef) exp(-coef) lower .95 upper .95
treat    0.9608     1.0408    0.6647    1.3889
age      1.0358     0.9654    1.0174    1.0546
sex      0.5868     1.7041    0.3628    0.9491
edema    4.9585     0.2017    2.7761    8.8563
bili     1.1362     0.8801    1.1044    1.1690

Concordance= 0.803  (se = 0.029 )
Rsquare= 0.341   (max possible= 0.984 )
Likelihood ratio test= 130.2  on 5 df,   p=0
Wald test            = 164  on 5 df,   p=0
Score (logrank) test = 249.3  on 5 df,   p=0
```

```
# Using Exact method
fit.pbc.exact <- coxph(Surv(time2, status==2) ~ trt + age + sex + edema + bili,
   ties="exact", data=pbc)
summary(fit.pbc.exact)

--------------------------------------------------------------------------------

  n= 312, number of events= 125
   (106 observations deleted due to missingness)

           coef exp(coef)  se(coef)      z Pr(>|z|)
treat -0.022272  0.977974  0.192379 -0.116 0.907835
age    0.036098  1.036757  0.009339  3.865 0.000111 ***
sex   -0.541694  0.581762  0.249720 -2.169 0.030067 *
edema  1.696709  5.455963  0.308832  5.494 3.93e-08 ***
bili   0.141415  1.151903  0.016182  8.739  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

      exp(coef) exp(-coef) lower .95 upper .95
treat    0.9780     1.0225    0.6708    1.4259
age      1.0368     0.9645    1.0180    1.0559
sex      0.5818     1.7189    0.3566    0.9491
edema    5.4560     0.1833    2.9785    9.9942
bili     1.1519     0.8681    1.1159    1.1890

Rsquare= 0.355   (max possible= 0.97 )
Likelihood ratio test= 137  on 5 df,   p=0
Wald test            = 155.3  on 5 df,   p=0
Score (logrank) test = 253.6  on 5 df,   p=0
```

– SAS code and output:

```
**************************************************;
DATA PBC;
   SET SURV.PBC;

   time2 = ROUND(time/91.25);
RUN;

PROC FREQ DATA=PBC;
   TABLES time2;
RUN;

**************************************************;
PROC PHREG DATA=PBC;
   MODEL time2*status(0 1) = trt age sex edema bili / TIES=efron;
RUN;

**************************************************;
PROC PHREG DATA=PBC;
   MODEL time2*status(0 1) = trt age sex edema bili / TIES= breslow;
RUN;

**************************************************;
PROC PHREG DATA=PBC;
   MODEL time2*status(0 1) = trt age sex edema bili / TIES= discrete;
RUN;

**************************************************;
PROC PHREG DATA=PBC;
   MODEL time2*status(0 1) = trt age sex edema bili / TIES= exact;
RUN;
```

	Coefficients				
	Treatment	Age	Sex	Edema	Bililubin
Efron	-0.042	0.036	-0.541	1.659	0.131
Breslow	-0.040	0.035	-0.533	1.601	0.128
Exact	-0.022	0.036	-0.542	1.697	0.141

Results

- The Efron and exact results are close to the original data (Try with the original data)
- The Breslow approximation attenuates the coefficients.

Cox Model: Properties of Estimators

- Topics of interest:
  - cumulative baseline hazard estimator
  - asymptotic distribution of score function
  - consistency of MPLE
  - asymptotic normality of MPLE
  - asymptotic behavior of cumulative hazard estimator
- Text: FH Chapter 8; KP Chapter 5

Cox Model: Counting Processes

- We define the following filtrations,

$$\begin{aligned}\mathcal{F}_i(t) &= \sigma\{N_i(s), Y_i(s+);\ s \in (0,t]\} \\ \mathcal{F}(t) &= \sigma\{N_i(s), Y_i(s+);\ s \in (0,t];\ i = 1, \ldots, n\}\end{aligned}$$

- Intensity process,

$$\begin{aligned}A_i(t;\boldsymbol{\beta}) &= \int_0^t Y_i(s) d\Lambda_i(s;\boldsymbol{\beta}) \\ dA_i(s;\boldsymbol{\beta}) &= Y_i(s)\exp(\boldsymbol{\beta}'\boldsymbol{Z}_i) d\Lambda_0(s)\end{aligned}$$

- We assume independent censoring, as described before

- Martingale increment,

$$dM_i(t;\boldsymbol{\beta}) = dN_i(t) - dA_i(t;\boldsymbol{\beta})$$

  i.e., $M_i(t;\boldsymbol{\beta}_0)$ is an $\mathcal{F}_i(t)$ martingale

Note: $\boldsymbol{\beta}_0$ = true value of regression parameter

Cumulative Hazard Estimator

- The quantity, $M_i(t;\boldsymbol{\beta}_0)$ is an $\mathcal{F}_i(t)$ martingale
- Therefore, $M(t;\boldsymbol{\beta})$ can serve as an unbiased estimating function, since

$$E[M_i(t;\boldsymbol{\beta}_0)] = 0$$

- As in GEE, set up the estimating function, then solve for $\Lambda_0(t)$

$$\begin{aligned} dN_i(s) &= Y_i(s)\exp(\boldsymbol{\beta}'\boldsymbol{Z}_i)d\Lambda_0(s) \\ \sum_{i=1}^n dN_i(s) &= \sum_{i=1}^n Y_i(s)\exp(\boldsymbol{\beta}'\boldsymbol{Z}_i)d\Lambda_0(s) \\ d\widehat{\Lambda}_0(s;\boldsymbol{\beta}) &= \frac{\sum_{i=1}^n dN_i(s)}{\sum_{i=1}^n Y_i(s)\exp(\boldsymbol{\beta}'\boldsymbol{Z}_i)} \end{aligned}$$

- Replacing $\boldsymbol{\beta}$ with a consistent estimator of $\boldsymbol{\beta}_0$,

$$\begin{aligned} d\widehat{\Lambda}_0(s;\widehat{\boldsymbol{\beta}}) &= \frac{\sum_{i=1}^n dN_i(s)}{\sum_{i=1}^n Y_i(s)\exp(\widehat{\boldsymbol{\beta}}'\boldsymbol{Z}_i)} \\ \widehat{\Lambda}_0(t;\widehat{\boldsymbol{\beta}}) &= \int_0^t \frac{\sum_{i=1}^n dN_i(s)}{\sum_{i=1}^n Y_i(s)\exp(\widehat{\boldsymbol{\beta}}'\boldsymbol{Z}_i)} \end{aligned}$$

Note: $\widehat{\Lambda}_0(t;\mathbf{0})$ returns the Nelson-Aalen estimator

- We can write,

$$\begin{aligned} \widehat{\Lambda}_0(t;\widehat{\boldsymbol{\beta}}) &= n^{-1}\sum_{i=1}^n \int_0^t S^{(0)}(s;\widehat{\boldsymbol{\beta}})^{-1}dN_i(s) \\ &= \int_0^t S^{(0)}(s;\widehat{\boldsymbol{\beta}})^{-1}d\overline{N}(s) \end{aligned}$$

- For convenience, we define the following

$$\begin{aligned} S^{(0)}(t;\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^{n} Y_i(t)\exp\{\boldsymbol{\beta}'\boldsymbol{Z}_i\} \\ S^{(1)}(t;\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^{n} \boldsymbol{Z}_i Y_i(t)\exp\{\boldsymbol{\beta}'\boldsymbol{Z}_i\} \\ S^{(2)}(t;\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^{n} \boldsymbol{Z}_i^{\otimes 2} Y_i(t)\exp\{\boldsymbol{\beta}'\boldsymbol{Z}_i\} \end{aligned}$$

  where $\boldsymbol{Z}_i^{\otimes 2} = \boldsymbol{Z}_i\boldsymbol{Z}_i'$

- Also, define

$$\begin{aligned} \overline{\boldsymbol{Z}}(t;\boldsymbol{\beta}) &= \frac{S^{(1)}(t;\boldsymbol{\beta})}{S^{(0)}(t;\boldsymbol{\beta})} \\ V(t;\boldsymbol{\beta}) &= \frac{S^{(2)}(t;\boldsymbol{\beta})}{S^{(0)}(t;\boldsymbol{\beta})} - \overline{\boldsymbol{Z}}(t;\boldsymbol{\beta})^{\otimes 2} \end{aligned}$$

  risk weighted mean and variance

- Note that

$$\begin{aligned} S^{(1)}(t;\boldsymbol{\beta}) &= \frac{\partial}{\partial\boldsymbol{\beta}} S^{(0)}(t;\boldsymbol{\beta}) \\ S^{(2)}(t;\boldsymbol{\beta}) &= \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} S^{(0)}(t;\boldsymbol{\beta}) \\ &= \frac{\partial}{\partial\boldsymbol{\beta}'} S^{(1)}(t;\boldsymbol{\beta}) \end{aligned}$$

Assumed Conditions

- We assume the following regularity conditions

  (i) $\{Y_i(\cdot), N_i(\cdot), \boldsymbol{Z}_i\}$ are independent and identically distributed

(ii) $\boldsymbol{Z}_i$ is bounded w.p. 1

(iii) $\Lambda_0(\tau) < \infty$

(iv) $P\{Y_i(t) = 1\} > 0$ for all $t \in (0, \tau]$

(v) $I_1(\boldsymbol{\beta}_0)$ is positive-definite

- Conditions (i) to (iv) can be relaxed, at the expense of increased technical complexity

Asymptotic Properties

- We will now derive some major asymptotic properties for the Cox model:

$$n^{-1/2}U(\boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0}, I_1(\boldsymbol{\beta}_0))$$

where $I_1(\boldsymbol{\beta}) = \int_0^\tau v(t;\boldsymbol{\beta})s^{(0)}(t;\boldsymbol{\beta})\lambda_0(t)dt$

$$\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$$
$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(0, I_1(\boldsymbol{\beta}_0)^{-1})$$

- We define the following limiting values

$$\begin{aligned}
s^{(0)}(t;\boldsymbol{\beta}) &= E[Y_i(t)\exp\{\boldsymbol{\beta}'\boldsymbol{Z}_i\}] \\
s^{(1)}(t;\boldsymbol{\beta}) &= E[\boldsymbol{Z}_i Y_i(t)\exp\{\boldsymbol{\beta}'\boldsymbol{Z}_i\}] \\
s^{(2)}(t;\boldsymbol{\beta}) &= E[\boldsymbol{Z}_i^{\otimes 2} Y_i(t)\exp\{\boldsymbol{\beta}'\boldsymbol{Z}_i\}] \\
\lim_{n\to\infty} V(t;\boldsymbol{\beta}) &\equiv v(t;\boldsymbol{\beta}) \\
v(t;\boldsymbol{\beta}) &= \frac{s^{(2)}(t;\boldsymbol{\beta})}{s^{(0)}(t;\boldsymbol{\beta})} - \overline{\boldsymbol{z}}(t;\boldsymbol{\beta})^{\otimes 2} \\
\overline{\boldsymbol{z}}(t;\boldsymbol{\beta}) &= \lim_{n\to\infty} \overline{\boldsymbol{Z}}(t;\boldsymbol{\beta})
\end{aligned}$$

- Under our assumed conditions, $S^{(d)}(t;\boldsymbol{\beta}) \xrightarrow{p} s^{(d)}(t;\boldsymbol{\beta})$ for $d = 0, 1, 2$, by the WLLN

Normality of Score Function

- *Theorem*: The scaled score function, $n^{-1/2}U(\boldsymbol{\beta}_0)$ converges in distribution to a zero-mean Normal with a variance that can be consistently estimated by $\int_0^\tau V(t;\widehat{\boldsymbol{\beta}})d\overline{N}(t)$

  (Sketch) of Proof:

  – To begin, we define the score process,

  $$U(t;\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n}\int_0^t \{\mathbf{Z}_i - \overline{\boldsymbol{Z}}(s;\boldsymbol{\beta})\}dN_i(s)$$

  where $U(\boldsymbol{\beta}) \equiv U(\tau;\boldsymbol{\beta})$

  – We can re-express the score process as follows,

  $$U(t;\boldsymbol{\beta}) \;=\; \sum_{i=1}^{n}\int_0^t \{\mathbf{Z}_i - \overline{\boldsymbol{Z}}(s;\boldsymbol{\beta})\}dM_i(s;\boldsymbol{\beta})$$

  since

  $$\begin{aligned} U(t;\boldsymbol{\beta}) &= \sum_{i=1}^{n}\int_0^t \{\mathbf{Z}_i - \overline{\boldsymbol{Z}}(s;\boldsymbol{\beta})\}dN_i(s) \\ &= \sum_{i=1}^{n}\int_0^t \{\mathbf{Z}_i - \overline{\boldsymbol{Z}}(s;\boldsymbol{\beta})\}dM_i(s;\boldsymbol{\beta}) \\ &\quad + \sum_{i=1}^{n}\int_0^t \{\mathbf{Z}_i - \overline{\boldsymbol{Z}}(s;\boldsymbol{\beta})\}Y_i(s)\exp\{\boldsymbol{\beta}'\boldsymbol{Z}_i\}\lambda_0(s)ds \end{aligned}$$

  since the second term can be shown to equal zero.

– That is,

$$\begin{aligned} & \sum_{i=1}^{n} \{\mathbf{Z}_i - \overline{\boldsymbol{Z}}(s;\boldsymbol{\beta})\} Y_i(s) \exp\{\boldsymbol{\beta}'\boldsymbol{Z}_i\} \\ = & \sum_{i=1}^{n} \boldsymbol{Z}_i Y_i(s) \exp\{\boldsymbol{\beta}'\boldsymbol{Z}_i\} - \overline{\boldsymbol{Z}}(s;\boldsymbol{\beta}) \sum_{i=1}^{n} Y_i(s) \exp\{\boldsymbol{\beta}'\boldsymbol{Z}_i\} \\ = & \; nS^{(1)}(s;\boldsymbol{\beta}) - nS^{(0)}(s;\boldsymbol{\beta})\overline{\boldsymbol{Z}}(s;\boldsymbol{\beta}) \\ = & \; \mathbf{0} \end{aligned}$$

– We can express the normalized score process, evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, as follows,

$$n^{-1/2} U(t;\boldsymbol{\beta}_0) = n^{-1/2} \sum_{i=1}^{n} \int_0^t \{\mathbf{Z}_i - \overline{\boldsymbol{Z}}(s;\boldsymbol{\beta}_0)\} dM_i(s;\boldsymbol{\beta}_0)$$

- a martingale transform
- i.e., $n^{-1/2}\{\mathbf{Z}_i - \overline{\boldsymbol{Z}}(s;\boldsymbol{\beta}_0)\}$ is $\mathcal{F}$ predictable

– Since, $n^{-1/2}U(t;\boldsymbol{\beta}_0)$ is a martingale process,

(i) $E[n^{-1/2}U(t;\boldsymbol{\beta}_0)] = \mathbf{0}$

(ii) $E[n^{-1/2}U(s;\boldsymbol{\beta}_0)\{n^{-1/2}U(t;\boldsymbol{\beta}_0) - n^{-1/2}U(s;\boldsymbol{\beta}_0)\}] = \mathbf{0}$, for $s < t$

(iii) variance,

$$\begin{aligned} & V\{n^{-1/2}U(t;\boldsymbol{\beta}_0)\} \\ = & \; E\left[\langle n^{-1/2}U(t;\boldsymbol{\beta}_0)\rangle\right] \\ = & \; E\left[n^{-1}\sum_{i=1}^{n}\int_0^t \{\mathbf{Z}_i - \overline{\boldsymbol{Z}}(s;\boldsymbol{\beta}_0)\}^{\otimes 2} Y_i(s)\exp\{\boldsymbol{\beta}_0'\boldsymbol{Z}_i\}\lambda_0(s)ds\right] \\ & \vdots \\ = & \; E\left[\int_0^t V(s;\boldsymbol{\beta}_0)S^{(0)}(s;\boldsymbol{\beta}_0)\lambda_0(s)ds\right] \end{aligned}$$

(iv) also, by the MCLT,

$$
\begin{aligned}
& V\{n^{-1/2}U(t;\boldsymbol{\beta}_0)\} \\
= & \lim_{n\to\infty} \langle n^{-1/2}U(t;\boldsymbol{\beta}_0)\rangle \\
= & \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\int_0^t \{\mathbf{Z}_i - \overline{\boldsymbol{Z}}(s;\boldsymbol{\beta}_0)\}^{\otimes 2} Y_i(s)\exp\{\boldsymbol{\beta}_0'\boldsymbol{Z}_i\}\lambda_0(s)ds \\
= & \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\int_0^t \{\mathbf{Z}_i - \overline{\boldsymbol{z}}(s;\boldsymbol{\beta}_0)\}^{\otimes 2} Y_i(s)\exp\{\boldsymbol{\beta}_0'\boldsymbol{Z}_i\}\lambda_0(s)ds \\
= & E\left[\int_0^t \{\mathbf{Z}_i - \overline{\boldsymbol{z}}(s;\boldsymbol{\beta}_0)\}^{\otimes 2} Y_i(s)\exp\{\boldsymbol{\beta}_0'\boldsymbol{Z}_i\}\lambda_0(s)ds\right]
\end{aligned}
$$

by the WLLN

- We can obtain a consistent variance estimator via (iii),
  - need to replace $\Lambda_0(t)$ with estimator
  - everything else in the mean in (iii) is observed
  - suggests the following estimator:

$$
\begin{aligned}
\widehat{V}\{n^{-1/2}U(t;\boldsymbol{\beta}_0)\} &= \int_0^t V(s;\boldsymbol{\beta}_0)S^{(0)}(s;\boldsymbol{\beta}_0)d\widehat{\Lambda}_0(s;\widehat{\boldsymbol{\beta}}) \\
&= \int_0^t V(s;\boldsymbol{\beta}_0)d\overline{N}(s)
\end{aligned}
$$

which completes the proof

- **Note**: $\int_0^t V(s;\boldsymbol{\beta})dN(s)$ equals $I(\boldsymbol{\beta})$, the observed information matrix corresponding to $PL(\boldsymbol{\beta})$.

Consistency of MPLE

- Claim: The MPLE, $\widehat{\boldsymbol{\beta}}$ converges in probability to $\boldsymbol{\beta}_0$.

(Sketch) of Proof:

To begin, we note the following result

- *Lemma*: Let $f_1, f_2, \ldots$ be a sequence of random concave functions, and let f be a deterministic function. If $\sup_x |f_n(x) - f(x)| \xrightarrow{p} 0$, then:
  - the function f is also concave
  - if f_n has unique maximum at x_n^* and if f has a unique maximum at x^*, then $x_n^* \xrightarrow{p} x^*$ as $n \to \infty$
  - Setting $\ell = \log PL$, we work with the following process,

$$\begin{aligned} X_n(t;\boldsymbol{\beta}) &= n^{-1}\{\ell(t;\boldsymbol{\beta}) - \ell(t;\boldsymbol{\beta}_0)\} \\ &= n^{-1}\sum_{i=1}^{n}\int_0^t (\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{Z}_i dN_i(s) \\ &\quad - \int_0^t \log\left\{\frac{S^{(0)}(s;\boldsymbol{\beta})}{S^{(0)}(s;\boldsymbol{\beta}_0)}\right\} dN(s) \end{aligned}$$

  - We can argue that this is a submartingale, with compensator $B_n(t;\boldsymbol{\beta})$, where

$$\begin{aligned} B_n(t;\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^{n}\int_0^t (\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{Z}_i Y_i(s)\lambda_i(s;\boldsymbol{\beta}_0)ds \\ &\quad - \sum_{i=1}^{n}\int_0^t \log\left\{\frac{S^{(0)}(s;\boldsymbol{\beta})}{S^{(0)}(s;\boldsymbol{\beta}_0)}\right\} Y_i(s)\lambda_i(s;\boldsymbol{\beta}_0)ds \end{aligned}$$

  - We then have the martingale, $X_n(t;\boldsymbol{\beta}) - B_n(t;\boldsymbol{\beta})$, which equals

$$n^{-1}\sum_{i=1}^{n}\int_0^t \left[(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{Z}_i - \log\left\{\frac{S^{(0)}(s;\boldsymbol{\beta})}{S^{(0)}(s;\boldsymbol{\beta}_0)}\right\}\right] dM_i(s;\boldsymbol{\beta}_0)$$

- The martingale $X_n(t;\boldsymbol{\beta}) - B_n(t;\boldsymbol{\beta})$ has predictable variable variation process,

$$\begin{aligned}&\langle X_n(t;\boldsymbol{\beta}) - B_n(t;\boldsymbol{\beta})\rangle \\ = \;& n^{-2}\sum_{i=1}^{n}\int_0^t \left[(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{Z}_i - \log\left\{\frac{S^{(0)}(s;\boldsymbol{\beta})}{S^{(0)}(s;\boldsymbol{\beta}_0)}\right\}\right]^2 dA_i(s;\boldsymbol{\beta})\end{aligned}$$

  - ∘ it can be shown that $n\langle X_n(t;\boldsymbol{\beta}) - B_n(t;\boldsymbol{\beta})\rangle$ converges in probability to a constant function of t
  - ∘ therefore, $\langle X_n(t;\boldsymbol{\beta}) - B_n(t;\boldsymbol{\beta})\rangle$ converges in probability to 0
  - ∘ implies that $X_n(\tau;\boldsymbol{\beta}) - B_n(\tau;\boldsymbol{\beta}) \xrightarrow{p} 0$

- The compensator, $B_n(t;\boldsymbol{\beta})$ converges in probability to

$$\begin{aligned}B(t;\boldsymbol{\beta}) \;=\; & \int_0^t (\boldsymbol{\beta}-\boldsymbol{\beta}_0)'s^{(1)}(s;\boldsymbol{\beta}_0)\lambda_0(s)ds \\ & - \int_0^t \log\left\{\frac{s^{(0)}(s;\boldsymbol{\beta})}{s^{(0)}(s;\boldsymbol{\beta}_0)}\right\} s^{(0)}(s;\boldsymbol{\beta}_0)\lambda_0(s)ds\end{aligned}$$

- Hence, $X_n(\tau;\boldsymbol{\beta}) \xrightarrow{p} B(\tau;\boldsymbol{\beta})$
- It can be shown that $X_n(\tau;\boldsymbol{\beta})$ is a concave function with unique maximum (at $\widehat{\boldsymbol{\beta}}$)
- In addition, it can be shown that $B(\tau;\boldsymbol{\beta})$ is concave with unique maximizer $\boldsymbol{\beta}_0$
- Applying the previous lemma, it follows that $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$

Asymptotic Normality of MPLE

- We now derive the limiting distribution of the MPLE, $\widehat{\boldsymbol{\beta}}$

- Using a first-order Taylor Series expansion,

$$U(\widehat{\boldsymbol{\beta}}) - U(\boldsymbol{\beta}_0) \;=\; \frac{\partial}{\partial \boldsymbol{\beta}'} U(\boldsymbol{\beta}) \bigg|_{\boldsymbol{\beta}_*} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

  where $\boldsymbol{\beta}_*$ lies on the line segment connecting $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$

- Recalling that $\widehat{\boldsymbol{\beta}}$ solves $U(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$, we have

$$U(\boldsymbol{\beta}_0) \;=\; I(\boldsymbol{\beta}_*)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

  such that

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \;=\; \{n^{-1} I(\boldsymbol{\beta}_*)\}^{-1} n^{-1/2} U(\boldsymbol{\beta}_0)$$

- Now, since $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ and since $||\boldsymbol{\beta}_* - \boldsymbol{\beta}_0|| \leq ||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||$, $\boldsymbol{\beta}_* \xrightarrow{p} \boldsymbol{\beta}_0$, meaning that

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \;=\; \{n^{-1} I(\boldsymbol{\beta}_0)\}^{-1} n^{-1/2} U(\boldsymbol{\beta}_0) + o_p(1)$$

  where the $o_p(1)$ term converges in probability to 0 as $n \to \infty$

- First, we demonstrate that $n^{-1/2} U(\boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0},\, \mathcal{I}_1(\boldsymbol{\beta}_0))$

- Recall that $n^{-1/2} U(\boldsymbol{\beta}_0)$ can be written as

$$n^{-1/2} U(\boldsymbol{\beta}_0) \;=\; \sum_{i=1}^{n} \int_0^{\tau} \{\mathbf{Z}_i - \overline{\boldsymbol{Z}}(s; \boldsymbol{\beta}_0)\} dM_i(s; \boldsymbol{\beta}_0)$$

  which is a martingale transform

- Recall also that, by the MCLT, $n^{-1/2} U(\boldsymbol{\beta}_0)$ converges to a zero-mean Normal

- Variance is given by limit of predictable variation process:

$$
\begin{aligned}
& \langle n^{-1/2} U(t;\boldsymbol{\beta}_0) \rangle \\
= \quad & n^{-1} \sum_{i=1}^{n} \int_0^t \{\mathbf{Z}_i - \overline{\boldsymbol{Z}}(s;\boldsymbol{\beta}_0)\}^{\otimes 2} Y_i(s) \exp\{\boldsymbol{\beta}_0' \boldsymbol{Z}_i\} \lambda_0(s) ds \\
\vdots \quad & \\
= \quad & \int_0^t V(s;\boldsymbol{\beta}_0) S^{(0)}(s;\boldsymbol{\beta}_0) \lambda_0(s) ds \\
\xrightarrow{p} \quad & \int_0^\tau v(s;\boldsymbol{\beta}) s^{(0)}(s;\boldsymbol{\beta}_0) \lambda_0(s) ds \\
\equiv \quad & \mathcal{I}_1(\boldsymbol{\beta}_0)
\end{aligned}
$$

- Therefore, by Rebolledo's Theorem,

$$
n^{-1/2} U(\boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0},\, \mathcal{I}_1(\boldsymbol{\beta}_0))
$$

- We now work on the information matrix component

- To begin, we use a familiar technique,

$$
\begin{aligned}
n^{-1} I(\boldsymbol{\beta}_0) &= n^{-1} \sum_{i=1}^{n} \int_0^\tau V(t;\boldsymbol{\beta}_0) dN_i(t) \\
&= n^{-1} \sum_{i=1}^{n} \int_0^\tau V(t;\boldsymbol{\beta}_0) dM_i(t;\boldsymbol{\beta}_0) \\
&\quad + n^{-1} \sum_{i=1}^{n} \int_0^\tau V(t;\boldsymbol{\beta}_0) Y_i(t) \exp\{\boldsymbol{\beta}_0' \boldsymbol{Z}_i\} \lambda_0(t) dt
\end{aligned}
$$

- Now, the first term is the realization of a martingale process (at $t = \tau$) since $V(t;\boldsymbol{\beta}_0)$ is $\mathcal{F}_t$ predictable
  - therefore, this quantity will have mean 0 if its predictable variation process $\xrightarrow{p} \mathbf{0}$ (Lenglart's Inequality)

- Examining the predictable variation process,

$$\begin{aligned}
&\left\langle n^{-1}\sum_{i=1}^{n}\int_0^\tau V(t;\boldsymbol{\beta}_0)dM_i(t;\boldsymbol{\beta}_0)\right\rangle \\
&= n^{-2}\sum_{i=1}^{n}\int_0^\tau V(t;\boldsymbol{\beta}_0)\otimes V(t;\boldsymbol{\beta}_0)Y_i(t)\exp\{\boldsymbol{\beta}_0'\boldsymbol{Z}_i\}\lambda_0(t)dt \\
&= n^{-1}\int_0^\tau V(t;\boldsymbol{\beta}_0)\otimes V(t;\boldsymbol{\beta}_0)S^{(0)}(t;\boldsymbol{\beta}_0)\lambda_0(t)dt
\end{aligned}$$

which is asmyptotically equivalent to

$$n^{-1}\int_0^\tau v(t;\boldsymbol{\beta}_0)\otimes v(t;\boldsymbol{\beta}_0)s^{(0)}(t;\boldsymbol{\beta}_0)\lambda_0(t)dt \xrightarrow{p} \mathbf{0}$$

since, under the assumed regularity conditions, integral will be bounded

- Therefore, ignoring $o_p(1)$ terms, we obtain

$$\begin{aligned}
n^{-1}I(\boldsymbol{\beta}_0) &= n^{-1}\sum_{i=1}^{n}\int_0^\tau V(t;\boldsymbol{\beta}_0)Y_i(t)\exp\{\boldsymbol{\beta}_0'\boldsymbol{Z}_i\}\lambda_0(t)dt \\
&= \int_0^\tau V(t;\boldsymbol{\beta}_0)S^{(0)}(t;\boldsymbol{\beta}_0)\lambda_0(t)dt \\
&\xrightarrow{p} \int_0^\tau v(t;\boldsymbol{\beta}_0)s^{(0)}(t;\boldsymbol{\beta}_0)\lambda_0(t)dt \\
&\equiv \mathcal{I}_1(\boldsymbol{\beta}_0)
\end{aligned}$$

- Note that $\mathcal{I}_1(\boldsymbol{\beta}_0)$ is assumed to be positive definite, such that

$$n^{-1}I(\boldsymbol{\beta}_0) \xrightarrow{p} \mathcal{I}_1(\boldsymbol{\beta}_0) \implies \{n^{-1}I(\boldsymbol{\beta}_0)\}^{-1} \xrightarrow{p} \mathcal{I}_1(\boldsymbol{\beta}_0)^{-1}$$

- In summary, we have argued that

$$\begin{aligned}
\{n^{-1}I(\boldsymbol{\beta}_*)\}^{-1} &\xrightarrow{p} \mathcal{I}_1(\boldsymbol{\beta}_0)^{-1} \\
n^{-1/2}U(\boldsymbol{\beta}_0) &\xrightarrow{D} N(\mathbf{0},\ \mathcal{I}_1(\boldsymbol{\beta}_0))
\end{aligned}$$

- Then, by Slutsky's Theorem,

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0},\ \mathcal{I}_1(\boldsymbol{\beta}_0)^{-1})$$

- That is, since the variance is given by

$$\begin{aligned} V\{n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\} &= V\{\mathcal{I}_1(\boldsymbol{\beta}_0)^{-1} n^{-1/2} U(\boldsymbol{\beta}_0)\} \\ &= \mathcal{I}_1(\boldsymbol{\beta}_0)^{-1} V\{n^{-1/2} U(\boldsymbol{\beta}_0)\} \mathcal{I}_1(\boldsymbol{\beta}_0)^{-1} \\ &= \mathcal{I}_1(\boldsymbol{\beta}_0)^{-1} \mathcal{I}_1(\boldsymbol{\beta}_0) \mathcal{I}_1(\boldsymbol{\beta}_0)^{-1} \end{aligned}$$

- We can estimate $\mathcal{I}_1(\boldsymbol{\beta}_0)$ consistently by $n^{-1}I(\widehat{\boldsymbol{\beta}})$, using the facts that $n^{-1}I(\boldsymbol{\beta}) \xrightarrow{p} \mathcal{I}_1(\boldsymbol{\beta})$ (WLLN, continuity) and $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$

- The above results are the basis for almost all interval estimation and hypothesis testing for the Cox model, with respect to covariate effects

Digression: Lenglart's Inequality

- Theorem: (Lenglart's Inequality) If N is a counting process adapted to $\mathcal{F}$ with intensity A, and if H is predictable, then for any $\mathcal{E}, \delta > 0$,

$$P\left(\sup_{t\in(0,\tau]} N(t) > \mathcal{E}\right) \le \frac{\delta}{\mathcal{E}} + P\left(A(\tau) > \delta\right)$$

- i.e., written in this form: the probability that the sub-martingale takes on large values at time t is bounded by the probability its compensator take on large values at time τ.

- Lenglart's Inequality is useful in establishing asymptotic results (e.g., consistency, asymptotic normality)

- *Corollary*: If N is a counting process adapted to $\mathcal{F}$ with intensity A, H is predictable, M is a square integrable martingale and $Z = \int H dM$, then for any $\mathcal{E}, \delta > 0$,

$$P\left(\sup_{t\in(0,\tau]} Z(t)^2 > \mathcal{E}\right) \leq \frac{\delta}{\mathcal{E}} + P\left\{\langle Z\rangle(\tau) > \delta\right\}$$

- Relating this corollary to Lenglart's Inequality, more generally, we have

$$P\left(\sup_{t\in(0,\tau]} |M(t)| > \sqrt{\mathcal{E}}\right) \ \leq \ \frac{\delta}{\mathcal{E}} + P\left\{\langle M\rangle(\tau) > \delta\right\}$$

- Applications of Lenglart's Inequality: set $Z_n = \int H_n dM_n$,
  - if $\langle Z_n\rangle(t)$ is non-decreasing in t for all n, and if $\langle Z_n\rangle(\tau) \xrightarrow{p} 0$, then $\sup_{t\in(0,\tau]} |Z_n(t)| \xrightarrow{p} 0$
    - i.e., a martingale converges in probability to 0 uniformly if its predictable variation process converges to 0 at τ
  - in addition to $\langle Z_n\rangle(t)$ being non-decreasing and converging in probability to 0 uniformly, if $\int_{(0,\tau]} H_n dA_n \xrightarrow{p} c$, then $\int_{(0,\tau]} H_n dN_n \xrightarrow{p} c$
    - if a submartingale's compensator converges in probability to a constant, the submartingale converges to that same constant
  - Recall: by the Martingale CLT, we have $Z_n = \int H_n dM_n \xrightarrow{D} N(0, \sigma_Z^2(t))$
  - For this result to be of practical value, we need a consistent estimator of $\sigma_Z^2(t)$; we previously suggested

$$\widehat{\sigma}_Z^2(t) \ = \ \int_0^t H_n^2(s) dN_n(s)$$

– In order to prove that $\widehat{\sigma}_Z^2(t) \xrightarrow{p} \sigma_Z^2(t)$, consider the following martingale,

$$\int_0^t H_n^2(s)dM_n(s) \;=\; \int_0^t H_n^2(s)\{dN_n(s) - dA_n(s)\}$$

– That is,

$$\int_0^t H_n^2(s)dM_n(s) \;=\; \widehat{\sigma}_Z^2(t) - \int_0^t H_n^2(s)dA_n(s)$$

– To apply the MCLT, we needed to confirm that

$$\int_0^t H_n^2(s)dA_n(s) \;\xrightarrow{p}\; \sigma_Z^2(t)$$

– Therefore, $\widehat{\sigma}_Z^2(t)$ will be consistent for $\sigma_Z^2(t)$ if we can show that

$$\widehat{\sigma}_Z^2(t) - \int_0^t H_n^2(s)dA_n(s) \;=\; \int_0^t H_n^2(s)dM_n(s)$$

converges in probability to 0

– By Lenglart's inequality, it suffices to show that

$$\left\langle \int H_n^2 dM_n \right\rangle \;=\; \int_0^t H_n^4(s)dA_n(s) \xrightarrow{p} 0$$

– In fact, we can demonstrate that

$$\sup_{t\in(0,\tau]} \left| \widehat{\sigma}_Z^2(t) - \int_0^t H_n^2(s)dA_n(s) \right| \;\xrightarrow{p}\; 0$$

by showing that

$$\int_0^{\tau} H_n^4 dA_n \ \xrightarrow{p} \ 0$$

- ◦ this last result is often not difficult to verify
- ◦ note the changes in the range of integration

Predicted Survival

- Hereafter, we treat $\boldsymbol{z}_i$ as a specified value of $\boldsymbol{Z}_i$
- Although covariate effects (quantified through $\widehat{\boldsymbol{\beta}}$) are of primary interest, investigators are often interested in the survival function
  - fitted survival probabilities
  - e.g., predicted 1-, 3- and 5-year survival for a 50-year old male African American diabetic patient who receives a deceased-donor kidney transplant after spending 3 years on dialysis
  - point and interval estimates
- Since point and interval estimators for $S(t|\boldsymbol{z}_i)$ can be obtained through $\widehat{\Lambda}(t|\boldsymbol{z}_i)$, we focus on the latter
- We now derive the limiting distribution of

$$n^{1/2}\{\widehat{\Lambda}(t|\boldsymbol{z}_i) - \Lambda(t|\boldsymbol{z}_i)\} = \widehat{\Delta}_{i1}(t) + \widehat{\Delta}_{i2}(t) \text{ where}$$

$$\widehat{\Delta}_{i1}(t) = n^{1/2}\left\{\widehat{\Lambda}(t|\boldsymbol{z}_i;\widehat{\boldsymbol{\beta}}) - \widehat{\Lambda}(t|\boldsymbol{z}_i;\boldsymbol{\beta}_0)\right\}$$
$$\widehat{\Delta}_{i2}(t) = n^{1/2}\left\{\widehat{\Lambda}(t|\boldsymbol{z}_i;\boldsymbol{\beta}_0) - \Lambda(t|\boldsymbol{z}_i)\right\}$$

where we define

$$\begin{aligned}\widehat{\Lambda}(t|\boldsymbol{z}_i;\boldsymbol{\beta}) &= \exp(\boldsymbol{\beta}'\boldsymbol{Z}_i)\widehat{\Lambda}_0(t;\boldsymbol{\beta}) \\ \widehat{\Lambda}_0(t;\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^{n}\int_0^t S^{(0)}(s;\boldsymbol{\beta})^{-1}dN_i(s)\end{aligned}$$

- We can re-write $\widehat{\Delta}_{i1}(t)$ as follows,

$$\widehat{\Delta}_{i1}(t) = n^{1/2}\left\{\exp(\widehat{\boldsymbol{\beta}}'\boldsymbol{Z}_i)\widehat{\Lambda}_0(t;\widehat{\boldsymbol{\beta}}) - \exp(\boldsymbol{\beta}_0'\boldsymbol{Z}_i)\widehat{\Lambda}_0(t;\boldsymbol{\beta}_0)\right\}$$

- Based on a first-order Taylor series expansion,

$$\widehat{\Delta}_{i1}(t) = \frac{\partial}{\partial\boldsymbol{\beta}'}\exp(\boldsymbol{\beta}'\boldsymbol{Z}_i)\widehat{\Lambda}_0(t;\boldsymbol{\beta})\Big|_{\boldsymbol{\beta}_*} n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

with $\boldsymbol{\beta}_*$ lying between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$, as previously defined

- Computing the key derivative,

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\beta}}\widehat{\Lambda}_0(t;\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^{n}\int_0^t \frac{\partial}{\partial\boldsymbol{\beta}}S^{(0)}(s;\boldsymbol{\beta})^{-1}dN_i(s) \\ &= -\,n^{-1}\sum_{i=1}^{n}\int_0^t \frac{S^{(1)}(s;\boldsymbol{\beta})}{S^{(0)}(s;\boldsymbol{\beta})^2}dN_i(s) \\ &= -\int_0^t \overline{\boldsymbol{Z}}(s;\boldsymbol{\beta})d\widehat{\Lambda}_0(s;\boldsymbol{\beta})\end{aligned}$$

- Then, completing derivative, we have

$$\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\beta}} \left\{ \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i) \widehat{\Lambda}_0(t; \boldsymbol{\beta}) \right\} \\
= & - \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i) \int_0^t \overline{\boldsymbol{Z}}(s; \boldsymbol{\beta}) d\widehat{\Lambda}_0(s; \boldsymbol{\beta}) + \boldsymbol{z}_i \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i) \widehat{\Lambda}_0(t; \boldsymbol{\beta}) \\
= & \exp(\boldsymbol{\beta}' \boldsymbol{Z}_i) \int_0^t \{ \boldsymbol{z}_i - \overline{\boldsymbol{Z}}(s; \boldsymbol{\beta}) \} d\widehat{\Lambda}_0(s; \boldsymbol{\beta}) \\
\equiv & \hat{\boldsymbol{k}}_i(t; \boldsymbol{\beta})
\end{aligned}$$

- Then, completing the Taylor approximation, we obtain

$$\widehat{\Delta}_{i1}(t) = \hat{\boldsymbol{k}}_i'(t; \boldsymbol{\beta}_*) \, n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

- As argued previously, $\boldsymbol{\beta}_* \xrightarrow{p} \boldsymbol{\beta}_0$, such that

$$||\hat{\boldsymbol{k}}_i(t; \boldsymbol{\beta}_*) - \hat{\boldsymbol{k}}_i(t; \boldsymbol{\beta}_0)|| \xrightarrow{p} \boldsymbol{0}$$

- Moreover, using the facts that $\overline{\boldsymbol{Z}}(s; \boldsymbol{\beta}) \xrightarrow{p} \overline{\boldsymbol{z}}(s; \boldsymbol{\beta})$ and $\widehat{\Lambda}_0(t; \boldsymbol{\beta}_0) \xrightarrow{p} \Lambda_0(t)$,

$$\begin{aligned}
\hat{\boldsymbol{k}}_i(t; \boldsymbol{\beta}_0) & \xrightarrow{p} \boldsymbol{k}_i(t; \boldsymbol{\beta}_0) \\
& \equiv \exp(\boldsymbol{\beta}_0' \boldsymbol{Z}_i) \int_0^t \{ \boldsymbol{z}_i - \overline{\boldsymbol{z}}(s; \boldsymbol{\beta}_0) \} d\Lambda_0(t)
\end{aligned}$$

  We can show that $\widehat{\Delta}_{i1}(t)$ is asymptotically equivalent to $\boldsymbol{k}_i'(t; \boldsymbol{\beta}_0) \, n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$.

- With respect to the term involving the MPLE, we recall that,

$$n^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \mathcal{I}_1(\boldsymbol{\beta}_0)^{-1} n^{-1/2} \sum_{1=1}^{n} \int_0^\tau \{ \boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(s; \boldsymbol{\beta}_0) \} dM_i(s; \boldsymbol{\beta}_0)$$

asymptotically.

- Using this fact, we have

$$\widehat{\Delta}_{i1}(t) = \boldsymbol{k}_i'(t;\boldsymbol{\beta}_0)\,\mathcal{I}_1(\boldsymbol{\beta}_0)^{-1}n^{-1/2}\sum_{1=1}^{n}\int_0^{\tau}\{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(s;\boldsymbol{\beta}_0)\}dM_i(s;\boldsymbol{\beta}_0)$$

asymptotically.

- We have already shown that

$$n^{-1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \;\xrightarrow{D}\; N(\boldsymbol{0},\, \mathcal{I}_1(\boldsymbol{\beta}_0)^{-1})$$

- This being the case, then

$$\widehat{\Delta}_{i1}(t) \;\xrightarrow{D}\; N(\boldsymbol{0},\, \boldsymbol{k}_i'(t;\boldsymbol{\beta}_0)\mathcal{I}_1(\boldsymbol{\beta}_0)^{-1}\boldsymbol{k}_i(t;\boldsymbol{\beta}_0))$$

as a linear combination of an asymptotic normal variate.

- We now consider the second term,

$$\begin{aligned}\widehat{\Delta}_{i2}(t) &= n^{1/2}\left\{\widehat{\Lambda}(t|\boldsymbol{z}_i;\boldsymbol{\beta}_0) - \Lambda(t|\boldsymbol{z}_i)\right\}\\ &= \exp(\boldsymbol{\beta}_0'\boldsymbol{Z}_i)n^{1/2}\{\widehat{\Lambda}_0(t;\boldsymbol{\beta}_0) - \Lambda_0(t)\}\end{aligned}$$

- Working on the baseline component,

$$
\begin{aligned}
& n^{1/2}\{\widehat{\Lambda}_0(t;\boldsymbol{\beta}_0) - \Lambda_0(t)\} \\
= \;& n^{1/2}\left\{ n^{-1}\sum_{i=1}^{n}\int_0^t S^{(0)}(s;\boldsymbol{\beta}_0)^{-1} dN_i(s) - \Lambda_0(t) \right\} \\
= \;& n^{-1/2}\sum_{i=1}^{n}\int_0^t S^{(0)}(s;\boldsymbol{\beta}_0)^{-1} dN_i(s) - n^{1/2}\int_0^t \frac{S^{(0)}(s;\boldsymbol{\beta}_0)}{S^{(0)}(s;\boldsymbol{\beta}_0)} d\Lambda_0(s) \\
= \;& n^{-1/2}\sum_{i=1}^{n}\left\{ \int_0^t \frac{dN_i(s)}{S^{(0)}(s;\boldsymbol{\beta}_0)} - \int_0^t \frac{Y_i(s)e^{\boldsymbol{\beta}_0'\boldsymbol{Z}_i}}{S^{(0)}(s;\boldsymbol{\beta}_0)} d\Lambda_0(s) \right\} \\
= \;& n^{-1/2}\sum_{i=1}^{n}\int_0^t S^{(0)}(s;\boldsymbol{\beta}_0)^{-1} dM_i(s;\boldsymbol{\beta}_0)
\end{aligned}
$$

which is a martingale transform since $S^{(0)}(s;\boldsymbol{\beta})^{-1}$ is predictable

- Therefore, by the MCLT, $n^{1/2}\{\widehat{\Lambda}_0(t;\boldsymbol{\beta}_0) - \Lambda_0(t)\}$ converges to a Normal distribution with mean 0 and variance given by the limit of the predictable variation process:

$$
\begin{aligned}
& \left\langle n^{1/2}\{\widehat{\Lambda}_0(t;\boldsymbol{\beta}_0) - \Lambda_0(t)\} \right\rangle \\
= \;& n^{-1}\sum_{i=1}^{n}\int_0^t S^{(0)}(s;\boldsymbol{\beta}_0)^{-2} Y_i(s) e^{\boldsymbol{\beta}_0'\boldsymbol{Z}_i}\lambda_0(s)ds \\
= \;& \int_0^t S^{(0)}(s;\boldsymbol{\beta}_0)^{-1}\lambda_0(s)ds \\
\xrightarrow{p} \;& \int_0^t s^{(0)}(s;\boldsymbol{\beta}_0)^{-1}\lambda_0(s)ds \\
\equiv \;& g(t;\boldsymbol{\beta}_0)
\end{aligned}
$$

- Combining the results for $\widehat{\Delta}_{i2}(t)$, we obtain

$$
\widehat{\Delta}_{i2}(t) \;\xrightarrow{D}\; \exp(\boldsymbol{\beta}_0'\boldsymbol{Z}_i) \times N(0,\; g(t;\boldsymbol{\beta}_0))
$$

as $n \to \infty$

- Therefore, as a linear combination of a Normal,

$$\widehat{\Delta}_{i2}(t) \xrightarrow{D} N(0,\ \exp(2\boldsymbol{\beta}_0'\boldsymbol{Z}_i)g(t;\boldsymbol{\beta}_0))$$

Orthogonal Martingales

- Having derived the asymptotic variance for $\widehat{\Delta}_{i1}(t)$ and $\widehat{\Delta}_{i2}(t)$, we now consider the covariance
  - we will now demonstrate that $\text{cov}(\widehat{\Delta}_{i1}(t), \widehat{\Delta}_{i2}(t)) = 0$
  - in doing so, we can omit the constant terms
- Recall that for two transforms

$$G_1(t) = \int_0^t H_1(s)dM(s) \qquad G_2(t) = \int_0^t H_2(s)dM(s)$$

which are martingales with respect to the same filtration, the predictable covariation process is given by

$$\langle G_1(t), G_2(t)\rangle \;=\; \int_0^t H_1(s)H_2(s)dA(s).$$

- Applying this idea, we need the predictable covariation process for the following two martingale transforms

$$n^{-1/2}\sum_{1=1}^{n}\int_0^{\tau}\{\boldsymbol{Z}_i - \overline{\boldsymbol{Z}}(s;\boldsymbol{\beta}_0)\}dM_i(s;\boldsymbol{\beta}_0)$$
$$n^{-1/2}\sum_{i=1}^{n}\int_0^{t} S^{(0)}(s;\boldsymbol{\beta}_0)^{-1}dM_i(s;\boldsymbol{\beta}_0)$$

- This predictable covariation process is given by

$$
\begin{aligned}
& n^{-1}\sum_{1=1}^{n}\int_0^t \{\boldsymbol{Z}_i-\overline{\boldsymbol{Z}}(s;\boldsymbol{\beta}_0)\}S^{(0)}(s;\boldsymbol{\beta}_0)^{-1}Y_i(s)\exp(\boldsymbol{\beta}_0'\boldsymbol{Z}_i)\lambda_0(s)ds \\
= & \int_0^t S^{(1)}(s;\boldsymbol{\beta}_0)S^{(0)}(s;\boldsymbol{\beta})^{-1}\lambda_0(s)ds - \int_0^t \overline{\boldsymbol{Z}}(s;\boldsymbol{\beta}_0)\lambda_0(s)ds \\
= & \ \mathbf{0}.
\end{aligned}
$$

- Hence, with the covariance equaling 0, we have

$$
\begin{aligned}
& n^{1/2}\{\widehat{\Lambda}(t|\boldsymbol{z}_i)-\Lambda(t|\boldsymbol{z}_i)\} = \widehat{\Delta}_{i1}(t)+\widehat{\Delta}_{i2}(t) \\
\xrightarrow{D} & \ N(0,\sigma^2(t|\boldsymbol{z}_i))
\end{aligned}
$$

with liming variance function

$$
\sigma^2(t|\boldsymbol{z}_i) \ = \ \boldsymbol{k}_i'(t;\boldsymbol{\beta}_0)\mathcal{I}_1(\boldsymbol{\beta}_0)^{-1}\boldsymbol{k}_i(t;\boldsymbol{\beta}_0) \ + \ \exp(2\boldsymbol{\beta}_0'\boldsymbol{Z}_i)g(t;\boldsymbol{\beta}_0)
$$

with all quantities as previously defined

- This completes the proof.