One sample problem

- \( T_1, \cdots, T_n \sim 1 - S(\cdot), C_1, \cdots, C_n \sim G(\cdot) \) and \( T_i \perp C_i \)
- Observations: \((U_i, \delta_i), i = 1, \cdots, n\)
- Objective: Estimate \( H(t) \) and \( S(t) \)
One sample problem

- \( N_i(t) = I(U_i \leq t)\delta_i \),
- \( Y_i(t) = I(U_i \geq t) \)
- \( M_i(t) = N_i(t) - \int_0^t Y_i(s)h(s)ds \)
- \( N(t) = \sum_{i=1}^n N_i(t), Y(t) = \sum_{i=1}^n Y_i(t), \) and \( M(t) = \sum_{i=1}^n M_i(t). \)
One sample problem

- The NA estimator

\[ \hat{H}(t) = \int_0^t \frac{I(Y(s) > 0)}{Y(s)} dN(s) \]

- Consider

\[ Q(t) = \int_0^t \frac{I(Y(s) > 0)}{Y(s)} dM(s) = \hat{H}(t) - H(t) + D(t) \]

where

\[ D(t) = \int_0^t h(s)I(Y(s) = 0) ds \]
The Bias of NA estimator:

\[
E(\hat{H}(t) - H(t)) = E(Q(t)) - E(D(t)) = - \int_0^t h(s)[1 - S(s)\{1 - G(s)\}]^n ds
\]

Suppose that \( \tau > 0 \) such that \( S(\tau) > 0 \) and \( 1 - G(\tau) > 0 \), then

\[
0 < E\{D(t)\} < H(t)[1 - S(\tau)\{1 - G(\tau)\}]^n \to 0.
\]
Asymptotical Distribution

- For $t \in [0, \tau]$

\[ n^{1/2} \{ \hat{H}(t) - H(t) \} = n^{1/2} Q(t) - n^{1/2} D(t) = \sum_{i=1}^{n} \int_{0}^{t} \frac{n^{1/2} I(Y(s) > 0)}{Y(s)} dM_i(s) + o_p(1) \]

- Check the conditions of MCLT
  1. Condition (a)

\[ \langle \sum_{i} U_{in}, \sum_{i} U_{in} \rangle (t) = \sum_{i=1}^{n} \int_{0}^{t} \frac{nI(Y(s) > 0)}{Y^2(s)} Y_i(s) h(s) ds \to \int_{0}^{t} \frac{h(s)}{\pi(s)} ds \]

  2. Condition (b)

\[ \langle \sum_{i} U_{in,\epsilon}, \sum_{i} U_{in,\epsilon} \rangle (t) = \int_{0}^{t} \frac{I(Y(s) > 0)}{n^{-1}Y(s)} I \left( n^{1/2} I(Y(s) > 0)/Y(s) \geq \epsilon \right) h(s) ds \to 0 \]

since

\[ P(\sup_{[0,\tau]} \{ n^{1/2} I(Y(s) > 0)/Y(s) \} \geq \epsilon) \leq P(n^{-1/2} \epsilon^{-1} \geq n^{-1} Y(\tau)) \to 0. \]
By MCLT and Slucky theorem $n^{1/2}\{\hat{H}(t) - H(t)\}$ converges weakly to a Gaussian process $X(t)$, $t \in [0, \tau]$ with independent increment and

$$\text{var}(X(t)) = \int_0^t \frac{h(s)}{\pi(s)} ds,$$

where $\pi(t) = S(t)\{1 - G(t)\}$.

$\sigma^2(t)$ can be estimated by

$$\int_0^t \frac{d\hat{H}(t)}{\hat{\pi}(t)} = n \int_0^t \frac{I(Y(s) > 0)}{Y(s)^2} dN(s) = n \sum_{\tau_j \leq t} \frac{d_j}{Y(\tau_j)^2}.$$
KM estimator is equivalent to $e^{-\hat{H}(t)}$

$\hat{S}(t)$ converges to $S(t)$ uniformly in $[0, \tau]$

$n^{1/2}\{\hat{S}(t) - S(t)\}$ converges weakly to a Gaussian process with a variance function of

$$S(t)^2 \sigma^2(t) = S(t)^2 \int_0^t \frac{h(s)}{\pi(s)} ds,$$

which can be estimated by

$$\hat{S}^2(t) \int_0^t \frac{d\hat{\Lambda}(s)}{n^{-1} Y(s)},$$

similar but not identical to the Greenwood formulae.
Two sample problem

- \((T_i, C_i, Z_i), i = 1, \cdots, n\)
- \(P(T_i > t | Z_i = j) = 1 - F_j(t) = S_j(t)\)
- \(P(C_i > t | Z_i = j) = 1 - G_j(t) \) (\(G_1(t)\) could be different from \(G_0(t)\).)
- \(n_j\) the size of group \(j\).
- Objective: test \(H_0 : S_0(\cdot) = S_1(\cdot)\)
\( N_j(t) = \sum_{i=1}^{n} I(U_i \leq t, \delta_i = 1, Z_i = j), j = 0, 1. \)

\( Y_j(t) = \sum_{i=1}^{n} I(U_i \geq t, Z_i = j), j = 0, 1. \)

\( M_j(t) = N_j(t) - \int_{0}^{t} Y_j(s) h_j(s) ds. \)

\( p_j = n_j/n, j = 0, 1. \)

\( \pi_j(s) = P(U_i \geq t|Z_i = j), j = 0, 1. \)
Weighted logrank test statistics

Two sample problem

\[ Q_w = \left( \frac{n}{n_0n_1} \right)^{1/2} U_w(0) = \left\{ \int_0^\infty H_{1w}(s)dN_1(s) - \int_0^\infty H_{0w}(s)dN_0(s) \right\} \]

where

\[ H_{jw}(s) = \left( \frac{n}{n_0n_1} \right)^{1/2} \frac{Y_{1-j}(s)}{Y(s)} w_n(s), j = 0, 1. \]
Two sample problem

- Martingale representation

\[ Q_w(t) = \int_0^t H_1w(s)dM_1(s) - \int_0^t H_0w(s)dM_0(s) + R_n(t) \]

where

\[ R_n(t) = \int_0^t H_1w(s)dA_1(s) - \int_0^t H_0w(s)dA_0(s) \]

\[ = \left( \frac{n}{n_0n_1} \right)^{1/2} \int_0^t \frac{Y_0(s)Y_1(s)w_n(s)}{Y(s)}(h_1(s) - h_0(s))ds. \]
Two sample problem

- Apply MCLT

\[ \int_0^t H_{jw}(s) \, dM_j(s) \to W(\sigma_j^2(t)) \text{ weakly,} \]

where

\[ \sigma_{jw}^2(t) = p_{1-j} \int_0^t \pi_{1-j}(s) \times \frac{\pi_0(s)\pi_1(s)}{\pi(s)^2} w^2(s) h_j(s) \, ds. \]

- \( \int_0^t H_{1w}(s) \, dM_1(s) - \int_0^t H_{0w}(s) \, dM_0(s) \) converges to \( N\{0, \sigma_{1w}^2(t) + \sigma_{0w}^2(t)\} \) in distribution.

- The result can be generalized to \( t = +\infty \).
Null Variance

- Under the null,

\[
Q_w(\infty) \sim N(0, \sigma_w^2)
\]

where

\[
\sigma_w^2 = \int_0^\infty \frac{f_0(s)(1 - G_0(s))(1 - G_1(s))w^2(s)}{p_0(1 - G_0(s)) + p_1(1 - G_1(s))} ds.
\]

- \(\sigma_w^2\) can be estimated by

\[
\hat{\sigma}_w^2 = \int_0^\infty \frac{Y_0(s)Y_1(s)n_0^{-1}n_1^{-1}}{n^{-1}Y(s)} w_n^2(s)d\hat{H}_0(s)
\]

\[
= \left( \frac{n}{n_0n_1} \right) \int_0^\infty \frac{Y_0(s)Y_1(s)}{Y(s)^2} w_n^2(s)d\{N_1(s) + N_0(s)\}
\]
The Null Distribution

- Under the null

\[ Z_w = \frac{Q_w}{\hat{\sigma}_w} \sim N(0, 1) \text{ under } H_0. \]

- For the logrank test

\[ Z = \frac{\sum_j (O_j - E_j)}{\sqrt{\sum_j V_j \times \frac{Y(\tau_j) - d_j}{Y(\tau_j) - 1}}}. \]
Under Alternative

- Assume that under $H_1$

$$\log \left( \frac{h_1^{(n)}(t)}{h_0(t)} \right) = g(t) / \sqrt{n}.$$

- Under $H_1 : R_n \rightarrow \xi$.

- Under $H_1 : Q_w \rightarrow N(\xi, \sigma_w^2)$ under alternatives.
The shift under alternatives

- Under the contiguous alternatives:

  \[ R_n \approx \sqrt{p_0 p_1} \int_0^\infty \frac{\pi_0(s) \pi_1(s)}{\pi(s)} w(s) \sqrt{n} (h_1^{(n)}(s) - h_0(s)) ds \]
  \[ \approx \sqrt{p_0 p_1} \int_0^\infty \frac{f_0(s)(1 - G_0(s))(1 - G_1(s))w(s)g(s)}{p_0(1 - G_0(s)) + p_1(1 - G_1(s))} ds = \xi \]

- The noncentrality parameter \( \theta = \xi / \sigma_w \) determines the power of the test.
The shift under alternatives

- The noncentrailty parameter

\[
|\theta| \propto \left| \int_0^\infty m(s)g(s)w(s)ds \right| \leq \sqrt{\int_0^\infty m(s)w^2(s)ds} = \sqrt{\int_0^\infty m(s)g^2(s)ds}
\]

- The maximum value is achieved by letting

\[
w(t) \propto g(t).
\]

- In practice \( g(\cdot) \) is unknown, so we may choose \( w(t) \) based on our best guess.
Cox regression

- \((T_i, C_i, Z_i), i = 1, \cdots, n\)
- PH assumption: \(h(t \mid Z_i) = h_0(t)e^{\beta_i'Z_i}\)
- Noninformative censoring: \(T_i \perp C_i \mid Z_i\).
- Exist \(\tau\) such that \(C_i < \tau\) and \(P(T_i \geq \tau) > 0\).
- \(Z_i\) are i.i.d bounded random variables.
\( S^{(j)}(\beta, s) = \sum_{i=1}^{n} I(U_i \geq s) e^{\beta' Z_i} Z_i^{\otimes j}, j = 0, 1, 2. \)

- **log-PL function:**

\[
\log\{PL(\beta)\} = \sum_{i=1}^{n} \int_{0}^{\tau} \left[ \beta' Z_i - \log \left\{ S^{(0)}(\beta, s) \right\} \right] dN_i(t)
\]

- **The score function**

\[
S(\beta) = \sum_{i=1}^{n} \int_{0}^{\tau} \left[ Z_i - \frac{S^{(1)}(\beta, s)}{S^{(0)}(\beta, s)} \right] dN_i(s)
\]

- **The information matrix**

\[
I(\beta) = \sum_{i=1}^{n} \int_{0}^{\tau} \left[ \frac{S^{(2)}(\beta, s)}{S^{(0)}(\beta, s)} - \frac{S^{(1)}(\beta, s)^{\otimes 2}}{S^{(0)}(\beta, s)^2} \right] dN_i(s).
\]
0 = S(\hat{\beta}) = S(\beta_0) - \hat{I}(\beta^*)(\hat{\beta} - \beta_0)

n^{1/2}(\hat{\beta} - \beta) = \left\{n^{-1}\hat{I}(\beta^*)\right\}^{-1} n^{-1/2}S(\beta_0)

1. $n^{-1}\hat{I}(\beta^*) - i(\beta_0) = o_p(1)$
2. $n^{-1/2}S(\beta_0) \rightarrow N\{0, i(\beta_0)\}$ as $n \rightarrow \infty$.

where

$$i(\beta_0) = \int_0^\tau \left[ s^{(2)}(\beta_0, s) - \frac{s^{(1)}(\beta_0, s) \otimes s}{s^{(0)}(\beta_0, s)} \right] h_0(s)ds.$$
Consider the expansion $n^{-1} \hat{I}(\beta^*) - i(\beta_0) = M_n(\beta^*) + A_n(\beta^*) - A(\beta^*) + A(\beta^*) - i(\beta_0)$

where

$$M_n(\beta) = n^{-1} \sum_{i=1}^{n} \int_0^\tau \left[ \frac{S^{(2)}(\beta, s)}{S^{(0)}(\beta, s)} - \frac{S^{(1)}(\beta, s)^2}{S^{(0)}(\beta, s)^2} \right] dM_i(s)$$

$$A_n(\beta) = n^{-1} \int_0^\tau \left[ \frac{S^{(2)}(\beta, s)}{S^{(0)}(\beta, s)} - \frac{S^{(1)}(\beta, s)^2}{S^{(0)}(\beta, s)^2} \right] S^{(0)}(\beta_0, s)h_0(s) ds$$

$$A(\beta) = \int_0^\tau \left[ \frac{s^{(2)}(\beta, s)}{s^{(0)}(\beta, s)} - \frac{s^{(1)}(\beta, s)^2}{s^{(0)}(\beta, s)^2} \right] s^{(0)}(\beta_0, s)h_0(s) ds$$

$M_n(\beta^*) = o_p(1)$ (MCLT), $A_n(\beta^*) - A(\beta^*) = o_p(1)$ (LLN) and $A(\beta^*) - i(\beta_0) = o_p(1)$ (Consistency of $\beta^*$ and continuity).
Scores

- Score functions
  \[ n^{-1/2} S(\beta_0) = n^{-1/2} \sum_{i=1}^{n} \left[ Z_i - \frac{S^{(1)}(\beta_0, s)}{S^{(0)}(\beta_0, s)} \right] dM_i(s) \]

- By MCLT \( n^{-1/2} S(\beta_0) \) converges weakly to \( N(0, \Sigma_0) \), where
  \[ \Sigma_0 = \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} \int_0^{\tau} \left[ Z_i - \frac{S^{(1)}(\beta_0, s)}{S^{(0)}(\beta_0, s)} \right]^{\otimes 2} Y_i(s)e^{\beta_0'}Z_i h_0(s) ds = i(\beta_0) \]

the information equality!
Asymptotical Normality

- Consistency: \( \hat{\beta} - \beta_0 = o_p(1) \).
- Normality:
  \[
  n^{1/2}(\hat{\beta} - \beta_0) \to N(0, i(\beta_0)^{-1}).
  \]
- \( i(\beta_0) \) can be consistently estimated \( n^{-1} I(\hat{\beta}) \).
Asymptotical Properties of Breslow estimator

- The cumulative baseline hazard function can be estimated as

\[ \hat{H}_0(t) = \sum_{i=1}^{n} \int_0^t \left\{ \sum_{i=1}^{n} I(U_i \geq s)e^{\beta'Z_i} \right\}^{-1} dN_i(s). \]

- Consider the expansion

\[
n^{1/2} \left\{ \hat{H}_0(t) - H_0(t) \right\} = n^{1/2} \int_0^t \left[ \left\{ \sum_{i=1}^{n} I(U_i \geq s)e^{\beta'Z_i} \right\}^{-1} - \left\{ \sum_{i=1}^{n} I(U_i \geq s)e^{\beta'_0Z_i} \right\}^{-1} \right] dN(s) \\
+ n^{1/2} \left[ \int_0^t \left\{ \sum_{i=1}^{n} I(U_i \geq s)e^{\beta'_0Z_i} \right\}^{-1} dN(s) - H_0(s) \right] + o_p(1)
\]
Asymptotical Properties of Breslow estimator

- Continue the expansion for $n^{1/2}\{\hat{H}_0(t) - H_0(t)\}$

$$= - \left[ \int_0^t s^{(1)}(\beta_0, s)\{s^{(0)}(\beta_0, s)\}^{-1} h_0(s)ds \right]' n^{1/2}(\hat{\beta} - \beta_0)$$

$$+ n^{1/2} \sum_{i=1}^n \int_0^t \left\{ \sum_{i=1}^n I(U_i \geq s)e^{\beta_0'Z_i} \right\}^{-1} dM_i(s) + o_p(1)$$

converging to a mean zero Gaussian process
Asymptotical Properties of Breslow estimator

- The variance of $n^{1/2}\{\hat{H}_0(t) - H_0(t)\}$ can be estimated as the empirical variance of

$$-n^{-1/2} \left[ \sum_{i=1}^{n} \int_{0}^{t} \frac{S^{(1)}(\beta_0, s)}{S^{(0)}(\beta_0, s)} d\hat{H}_0(s) \right]' \hat{I}(\beta_0) \int_{0}^{\tau} \left[ Z_i - \frac{S^{(1)}(\beta_0, s)}{S^{(0)}(\beta_0, s)} \right] dN_i(s) G_i$$

$$+n^{-1/2} \sum_{i=1}^{n} \int_{0}^{t} \left\{ n^{-1} \sum_{i=1}^{n} I(U_i \geq s) e^{\beta_0 Z_i} \right\}^{-1} dN_i(s) G_i,$$

where $G_i \sim N(0, 1), i = 1, 2, \ldots, n.$
The Consequence of model mis-specification

- If \( h(t|Z_i) = h(t, Z_i) \neq h_0(t)e^{\beta_0'Z_i} \), then

- \( \hat{\beta} \) converges to \( \beta^* \) which is the solution of the estimating equation

\[
\int_0^\infty \left[ s^{(1)}(t) - \frac{s^{(1)}(\beta, t)}{s^{(0)}(\beta, t)} s^{(0)}(t) \right] dt = 0
\]

where

\[
s^{(j)}(t) = \lim n^{-1} \sum_{i=1}^n Y_i(t)h(t, Z_i)Z_i^j.
\]
The Consequence of model mis-specification

- $n^{1/2}(\hat{\beta} - \beta^*)$ still converges to a normal distribution with mean zero and variance $A^{-1}BA^{-1}$.

  where

  \[
  A = \int_0^\infty \left[ \frac{s^{(2)}(\beta^*, s)}{s^{(0)}(\beta^*, s)} - \frac{s^{(1)}(\beta^*, s)}{s^{(0)}(\beta^*, s)^2} \right] s^{(0)}(s) ds.
  \]

  \[
  B = E \left[ \int_0^\infty \left\{ Z_i - \frac{s^{(1)}(\beta^*, s)}{s^{(0)}(\beta^*, s)} \right\} dN_i(t) - \int_0^\infty \frac{Y_i(t) e^{Z_i'\beta^*}}{s^{(0)}(\beta^*, t)} \left\{ Z_i - \frac{s^{(1)}(\beta^*, t)}{s^{(0)}(\beta^*, t)} \right\} dE\{N_i(t)\} \right]^2.
  \]

- $\beta^*$ may depend on the censoring distribution!

- Robust variance estimator for $\hat{\beta}$. 

29