

- Semiparametric proportional hazards model

$$\lambda(t | Z) = \lambda_0(t) \exp(\beta^T Z)$$

- Estimation of  $\beta$  (QPS approach)

- log-likelihood function as if  $\lambda_0(t)$  were known

$$l(\beta) = \sum_{i=1}^n \left[ \Delta_i \log \lambda_0(X_i) + \Delta_i \beta Z_i - \int_0^{X_i} \lambda_0(u) \exp(\beta Z_i) du \right]$$

- partial derivative w.r.t  $\beta$

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^n \int_0^{\infty} Z_i \left\{ dN_i(u) - Y_i(u) \exp(\beta^T Z_i) \lambda_0(u) du \right\}$$

- $\frac{\partial l}{\partial \beta}$  is not enough because  $\lambda_0(t)$  is unknown

- estimate  $\lambda_0(t)$  by replacing  $Z_i$  as if  $\beta$  were known

- Here is an estimating equation for  $\beta$ :

$$S_n(\beta) = \sum_{i=1}^n \int_0^{\infty} \{Z_i - \bar{Z}(u; \beta)\} dN_i(u)$$

- under the proportional hazards model, this is the partial score equation
- in general, this is called quasi partial score equation
- this way of getting appropriate estimating functions is called Quasi Partial Scoring (QPS).
- Estimator  $\hat{\beta}_n$ :  $S_n(\hat{\beta}_n) = 0$

- Primary inferential questions: for true value  $\beta = \beta_0$ 
  - consistent:  $\hat{\beta}_n \rightarrow_P \beta_0$
  - asymptotically normal:  $\sqrt{n}(\hat{\beta}_n - \beta_0) \rightarrow_D \mathcal{N}(0, \sigma^2)$
- Secondary inferential questions
  - optimality
  - asymptotic properties for Breslow estimator

- From estimating function to estimators:

$$S_n(\hat{\beta}_n) - S_n(\beta_0) = -S_n(\beta_0) \approx S'(\beta_0)(\hat{\beta}_n - \beta_0)$$

- $S_n(\beta_0)$ : for zero mean and variation
- $S'(\beta_0)$ : for efficiency

- Recall:  $\bar{Z}(u; \beta) = \frac{\sum_{i=1}^n Z_i Y_i(u) \exp(\beta Z_i)}{\sum_{i=1}^n Y_i(u) \exp(\beta Z_i)}$

- What is  $S_n(t; \beta_0)$ ?

$$\begin{aligned}
 S_n(t; \beta_0) &= \sum_{i=1}^n \int_0^t \{Z_i - \bar{Z}(u; \beta_0)\} dN_i(u) \\
 &= \sum_{i=1}^n \int_0^t \{Z_i - \bar{Z}(u; \beta_0)\} [dN_i(u) - Y_i(u) \lambda_0(u) \exp(\beta_0 Z_i) du] \\
 &\quad + \sum_{i=1}^n \int_0^t \{Z_i - \bar{Z}(u; \beta_0)\} Y_i(u) \exp(\beta_0 Z_i) \lambda_0(u) du \\
 &= \sum_{i=1}^n \int_0^t \{Z_i - \bar{Z}(u; \beta_0)\} dM_i(u; \beta_0) \\
 &\quad + \int_0^t \left\{ \sum_{i=1}^n Z_i Y_i(u) \exp(\beta_0 Z_i) - \sum_{i=1}^n Y_i(u) \exp(\beta_0 Z_i) \bar{Z}(u; \beta_0) \right\} \lambda_0(u) du \\
 &= \sum_{i=1}^n \int_0^t \{Z_i - \bar{Z}(u; \beta_0)\} dM_i(u; \beta_0)
 \end{aligned}$$

- Predictable variation of  $U_n(t) = n^{-1/2}S_n(t; \beta_0)$

$$\begin{aligned} \langle U_n, U_n \rangle (t) &= n^{-1} \sum_{i=1}^n \int_0^t \{Z_i - \bar{Z}(u)\}^2 Y_i(u) \exp(\beta_0 Z_i) \lambda_0(u) du \\ &\rightarrow \int_0^t E \left[ \{Z - \mu_Z(u)\}^2 Y(u) \exp(\beta_0 Z) \right] \lambda_0(u) du \end{aligned}$$

- Properties of  $S_n(t; \beta_0)$ :

1.  $ES_n(t; \beta_0) = 0$

2. independent increment for  $s \leq t$

$$\text{cov}[S_n(s; \beta_0) \{S_n(t; \beta_0) - S_n(s; \beta_0)\}] = 0$$

3.  $\text{var}\{n^{-1/2}S_n(t; \beta_0)\} = E \langle U_n, U_n \rangle (t)$

- Consistency (sketch of a proof)

- roughly,  $-S_n(\beta_0) \approx S'(\beta_0)(\hat{\beta}_n - \beta_0)$

- by WLLN,  $n^{-1}S_n(\beta_0) \rightarrow_P 0$

- for arbitrary  $\beta \neq \beta_0$

$$\begin{aligned} n^{-1}S_n(\beta) - n^{-1}S_n(\beta_0) &= -n^{-1} \sum_{i=1}^n \int_0^t \{\bar{Z}(u; \beta) - \bar{Z}(u; \beta_0)\} dN_i(u) \\ &\rightarrow_P \Gamma(\beta) = - \int_0^t \{\mu(u; \beta) - \mu(u; \beta_0)\} d\Pr\{X \leq u, \Delta = 1\} \end{aligned}$$

- we can verify

$$\frac{\partial \mu(u; \beta_0)}{\partial \beta} = \frac{E[\{Z - \mu(u, \beta_0)\}^2 \exp(\beta_0 Z) Y(u)]}{E[\exp(\beta_0 Z) Y(u)]} > 0$$

provided  $Z$  is not constant

– then for  $\Gamma(\beta)$

1.  $n^{-1}S_n(\beta) \rightarrow_P \Gamma(\beta)$

2.  $\Gamma(\beta)$  and  $n^{-1}S_n(\beta)$  are strictly decreasing

3.  $\Gamma(\beta_0) = 0$

– similar to the Glivenko-Cantelli Lemma  $\Rightarrow n^{-1}S_n(\beta) \rightarrow_P \Gamma(\beta)$  uniformly in  $\beta \in U(\beta_0)$

– for any  $\epsilon > 0$ , there exists sufficiently large  $n$ , s.t.

$$S_n(\beta_0 - \epsilon) > 0, S_n(\beta_0 + \epsilon) < 0$$

therefore,  $\hat{\beta}_n \in (\beta_0 - \epsilon, \beta_0 + \epsilon)$  with sufficiently large probability

–  $\hat{\beta}_n \rightarrow_P \beta_0$

- Asymptotic normality

- partial score equation

$$S_n(\beta) = \sum_{i=1}^n \int_0^\infty \{Z_i - \bar{Z}(u; \beta)\} dN_i(u)$$

- mean value expansion: for  $\beta_n^*$  lies between  $\hat{\beta}_n$  and  $\beta_0$

$$S_n(\hat{\beta}_n) = 0 = S_n(\beta_0) + \left. \frac{\partial S_n(\beta)}{\partial \beta} \right|_{\beta=\beta_n^*} (\hat{\beta}_n - \beta_0)$$

$$\Rightarrow n^{1/2}(\hat{\beta}_n - \beta_0) = \left\{ -n^{-1} \left. \frac{\partial S_n(\beta)}{\partial \beta} \right|_{\beta=\beta_n^*} \right\}^{-1} \times n^{-1/2} S_n(\beta_0)$$

- recall

$$n^{-1/2} S_n(\beta_0) \rightarrow_D \mathcal{N}(0, \sigma^2)$$

– slope:

$$\frac{\partial S_n(\beta)}{\partial \beta} = - \sum_{i=1}^n \int_0^\infty \frac{\sum_{j=1}^n \{Z_j - \bar{Z}(u)\}^2 \exp(\beta_0 Z) Y_j(u)}{\sum_{j=1}^n \exp(\beta_0 Z) Y_j(u)} dN_i(u)$$

– since  $\hat{\beta}_n^* \rightarrow \beta_0$ ,

$$\begin{aligned} & -n^{-1} \left. \frac{\partial S_n(\beta)}{\partial \beta} \right|_{\beta=\hat{\beta}_n^*} \\ & \rightarrow \int_0^\infty \frac{E[\{Z - \bar{Z}(u)\}^2 \exp(\beta_0 Z) Y(u)]}{E[\exp(\beta_0 Z) Y(u)]} d\Pr\{X \leq u, \Delta = 1\} = \sigma^2 \end{aligned}$$

– therefore

$$n^{1/2}(\hat{\beta}_n - \beta_0) \rightarrow_D \mathcal{N}(0, \sigma^{-2})$$

- How to estimate  $\sigma^{-2}$

$$\hat{\sigma}^{-2} = \left[ n^{-1} \sum_{i=1}^n \frac{\Delta_i \sum_{j=1}^n \{Z_j - \bar{Z}(X_i)\}^2 \exp(\hat{\beta}_n Z_j) Y_j(X_i)}{\sum_{j=1}^n \exp(\hat{\beta}_n Z_j) Y_j(X_i)} \right]^{-1}$$

- therefore, standardized version

$$\frac{\hat{\beta}_n - \beta_0}{\left[ \sum_{i=1}^n \int_0^\infty \frac{\sum_{j=1}^n \{Z_j - \bar{Z}(u)\}^2 \exp(\hat{\beta}_n Z) Y_j(u)}{\sum_{j=1}^n \exp(\hat{\beta}_n Z) Y_j(u)} dN_i(u) \right]^{-1/2}} \rightarrow_D \mathcal{N}(0, 1)$$

- what if  $\beta_0 = 0$  and  $Z = 0/1$ ?

- denote  $S_Z(u; \beta) = \frac{\sum_{j=1}^n \{Z_j - \bar{Z}(u)\}^2 \exp(\beta Z) Y_j(u)}{\sum_{j=1}^n \exp(\beta Z) Y_j(u)}$

- 95% confidence interval for  $\beta_0$  is

$$\hat{\beta}_n \pm 1.96 \times \left[ \sum_{i=1}^n \int_0^\infty S_Z(u; \hat{\beta}_n) dN_i(u) \right]^{-1/2}$$

- Hypothesis testing on  $H_0 : \beta = \beta_0$ : rejects  $H_0$  at 5% type-I error

- Wald's test:

$$\left| \frac{\hat{\beta}_n - \beta_0}{\left[ \sum_{i=1}^n \int_0^\infty S_Z(u; \hat{\beta}_n) dN_i(u) \right]^{-1/2}} \right| \geq 1.96$$

- score test:

$$\left| \frac{S_n(\beta_0)}{\widehat{\text{var}}[S_n(\beta_0)]} \right| \geq 1.96$$

i.e.,

$$\left| \frac{\sum_{i=1}^n \int_0^\infty \{Z_i - \bar{Z}(u; \beta_0)\} dN_i(u)}{\left[ \sum_{i=1}^n \int_0^\infty S_Z(u; \beta_0) dN_i(u) \right]^{1/2}} \right| \geq 1.96$$

- what if  $\beta = 0$  and  $Z = 0/1$  in score test:

$$\frac{\sum_{i=1}^n \int_0^\infty \{Z_i - \bar{Z}(u; \beta_0)\} dN_i(u)}{\left[ \sum_{i=1}^n \int_0^\infty S_Z(u; \beta_0) dN_i(u) \right]^{1/2}}$$

–

$$S_Z(u; \beta_0) = \frac{\sum_{j=1}^n \{Z_j - \bar{Z}(u)\}^2 \exp(\beta_0 Z) Y_j(u)}{\sum_{j=1}^n \exp(\beta_0 Z) Y_j(u)} = \bar{Z}(u) \{1 - \bar{Z}(u)\}$$

- score test becomes the Log-rank test

- Summary on QPS estimation approach
  - write down full likelihood function
  - partial derivative with respect to the finite-dimensional parameter of interest
  - use unit weight to derive an estimator for the infinite-dimensional nuisance parameter
  - solve the quasi-partial score equation
  - derive asymptotics
- What's the probability structure underlying QPS approach?

- Partial likelihood

- order the complete failure times by  $t_{(1)}, t_{(2)}, \dots, t_{(K)}$

- $Z_{(i)}$  is the covariate corresponding to  $t_{(i)}$

- let  $\Gamma_i$  denote all the information available up to time  $t_{(i)}$  and a failure occurred at  $t_{(i)}$ ,  $\Gamma_k \cup \{Z_{(k)}\} \subset \Gamma_{k+1}$

- \*  $\Gamma = \bigcup_{i=1}^K \Gamma_i = \{(X_j, \Delta_j, Z_j), j = 1, 2, \dots, n\}$

- \*  $\Gamma = \bigcup_{i=1}^K \{\Gamma_i \cup \{Z_{(i)}\}\} = \cup\{\Gamma_1, Z_{(1)}; \dots; \Gamma_K, Z_{(K)}; \Gamma\}$

- full likelihood is  $\Pr\{\Gamma = \gamma; \beta, \lambda_0(\cdot)\}$
- decomposed into product of conditional likelihood

$$\begin{aligned}
 & \Pr\{\Gamma_1 = \gamma_1\} \\
 & \times \Pr\{Z_{(1)} = z_{(1)} \mid \Gamma_1 = \gamma_1\} \\
 & \times \Pr\{\Gamma_2 = \gamma_2 \mid Z_{(1)} = z_{(1)}, \Gamma_1 = \gamma_1\} \\
 & \times \Pr\{Z_{(2)} = z_{(2)} \mid Z_{(1)} = z_{(1)}, \Gamma_1 = \gamma_1, \Gamma_2 = \gamma_2\} \\
 & \dots
 \end{aligned}$$

- for a general  $k$ ,

$$\begin{aligned}
 & \Pr\{Z_{(k)} = z_{(k)} \mid Z_{(k-1)} = z_{(k-1)}, \Gamma_{k-1} = \gamma_{k-1}, \Gamma_k = \gamma_k\} \\
 & = \Pr\{Z_{(k)} = z_{(k)} \mid \Gamma_k = \gamma_k\}
 \end{aligned}$$

- partial likelihood

$$PL = \prod_{k=1}^K \Pr\{Z_{(k)} = z_{(k)} \mid \Gamma_k = \gamma_k; \beta, \lambda_0(\cdot)\}$$

- what is  $\Pr\{Z_{(k)} = z_{(k)} \mid \Gamma_k = \gamma_k; \beta, \lambda_0(\cdot)\}$ 
  - conditional  $\Gamma_k$ ,  $n_k$  individuals are at-risk with covariate values of  $Z_{kj}$ ,  $j = 1, \dots, n_k$
  - assuming no-ties, this probability equals to  $Z_{(k)}$  failed given exactly one out of  $n_k$  subjects failed at  $t_{(k)}$

$$\Pr\{Z_{(k)} = z_{(k)} \mid \Gamma_k = \gamma_k; \beta, \lambda_0(\cdot)\} = \frac{\lambda(t_{(k)} \mid Z_{(k)})}{\sum_{j=1}^{n_k} \lambda(t_{(k)} \mid Z_{ij})}$$

- proportional hazards model:  $\lambda(t \mid Z_{ij}) = \lambda_0(t) \exp(\beta Z_{ij})$  and  $\lambda(t \mid Z_{(k)}) = \lambda_0(t) \exp(\beta Z_{(k)})$
- partial likelihood

$$PL = \prod_{k=1}^K \frac{\exp(\beta Z_{(k)})}{\sum_{j=1}^{n_k} \exp(\beta Z_{ij})} = \prod_{i=1}^n \left[ \frac{\exp(\beta Z_i)}{\sum_{j=1}^n \exp(\beta Z_j) I(X_j \geq X_i)} \right]^{\Delta_i}$$

- partial score function

- log partial-likelihood

$$l(\beta) = \sum_{i=1}^n \Delta_i \left[ \beta Z_i - \log \sum_{j=1}^n \exp(\beta Z_j) I(X_j \geq X_i) \right]$$

- partial score function

$$\begin{aligned} \frac{\partial l}{\partial \beta} &= \sum_{i=1}^n \Delta_i \left[ Z_i - \frac{\sum_{j=1}^n Z_j \exp(\beta Z_j) I(X_j \geq X_i)}{\sum_{j=1}^n \exp(\beta Z_j) I(X_j \geq X_i)} \right] \\ &= \sum_{i=1}^n \int_0^{\infty} \left[ Z_i - \frac{\sum_{j=1}^n Z_j \exp(\beta Z_j) I(X_j \geq u)}{\sum_{j=1}^n \exp(\beta Z_j) I(X_j \geq u)} \right] dN_i(u) \\ &= \sum_{i=1}^n \int_0^{\infty} \{Z_i - \bar{Z}(u; \beta)\} dN_i(u) \end{aligned}$$

- Proportional hazards model for multivariate covariates  $Z \in \mathcal{R}^p$

$$\lambda(t | Z) = \lambda_0(t) \exp(\beta^T Z)$$

- $\beta = (\beta_1, \dots, \beta_p)^T$  is  $p$ -vector
- partial score function

$$S_n(\beta) = \sum_{i=1}^n \int_0^\infty \{ Z_i^{p \times 1} - \bar{Z}^{p \times 1}(u, \beta) \} dN_i(u)$$

where

$$Z_i^{p \times 1} = \begin{bmatrix} Z_{i1} \\ Z_{i2} \\ \dots \\ Z_{ip} \end{bmatrix}_{p \times 1}, \quad \bar{Z}^{p \times 1}(u, \beta) = \begin{bmatrix} \bar{Z}_1(u, \beta) \\ \bar{Z}_2(u, \beta) \\ \dots \\ \bar{Z}_p(u, \beta) \end{bmatrix}_{p \times 1}$$

–  $n^{-1/2}S_n(\beta_0) \rightarrow_D \mathcal{N}(0, \Sigma(\beta_0))$ , where  $\Sigma(\beta_0)$  is the limit of

$$n^{-1} \sum_{i=1}^n \int_0^\infty \frac{\sum_{i=1}^n \{Z_i - \bar{Z}\}_{p \times 1} \{Z_i - \bar{Z}\}_{1 \times p}^T \exp(\beta Z_j) Y_i(u)}{\sum_{i=1}^n \exp(\beta Z_j) Y_i(u)} dN_i(u)$$

$$= n^{-1} \sum_{i=1}^n \int_0^\infty \frac{\sum_{i=1}^n \{Z_i - \bar{Z}\}_{p \times p}^{\otimes 2} \exp(\beta Z_j) Y_i(u)}{\sum_{i=1}^n \exp(\beta Z_j) Y_i(u)} dN_i(u)$$

where  $v^{\otimes 0} = 1$ ,  $v^{\otimes 1} = v$  and  $v^{\otimes 2} = vv^T$ .

– partial derivative

$$-\frac{\partial S_n(\beta_n^*)}{\partial \beta} \xrightarrow{p \times p} \Sigma(\beta_0)$$

–

$$n^{1/2}(\hat{\beta}_n - \beta_0) \rightarrow_D \mathcal{N}(0, \Sigma^{-1}(\beta_0))$$

- Estimation of baseline hazard function (Breslow estimator)

$$\hat{\Lambda}_0(t; \hat{\beta}_n) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n Y_i(u) \exp(\hat{\beta}_n^T Z_i)}$$

– consistent

– asymptotic normality:  $n^{1/2}[\hat{\Lambda}_0(t; \hat{\beta}_n) - \Lambda_0(t)]$  equals

$$\begin{aligned} & n^{1/2}[\hat{\Lambda}_0(t; \beta_0) - \Lambda_0(t)] + n^{1/2}[\hat{\Lambda}_0(t; \hat{\beta}_n) - \hat{\Lambda}_0(t; \beta_0)] \\ & = \text{Term I} + \text{Term II} \end{aligned}$$

– Term I:

$$n^{1/2} \sum_{i=1}^n \int_0^t \frac{dM_i(t; \beta_0)}{\sum_{i=1}^n \exp(\beta_0^T Z_i) Y_i(u)}$$

– sum of martingale residuals with predictable variation

$$n \sum_{i=1}^n \int_0^t \frac{Y_i(u) \exp(\beta_0^T Z_i) \lambda_0(u) du}{[\sum_{i=1}^n \exp(\beta_0^T Z_i) Y_i(u)]^2} \rightarrow \int_0^t \frac{\lambda_0(u) du}{E[\exp(\beta_0^T Z) Y(u)]} = \sigma_1^2(t)$$

– Term II: mean value expansion

$$\begin{aligned}
 & n^{1/2}[\widehat{\Lambda}_0(t; \widehat{\beta}_n) - \widehat{\Lambda}_0(t; \beta_0)] \\
 &= n^{1/2}(\widehat{\beta}_n - \beta_0) \int_0^t \frac{\sum_{i=1}^n \exp(\beta_n^{*T} Z_i) Z_i Y_i(u) dN(u)}{[\sum_{i=1}^n \exp(\beta_n^{*T} Z_i) Y_i(u)]^2}
 \end{aligned}$$

–

$$\int_0^t \frac{\sum_{i=1}^n \exp(\beta_n^{*T} Z_i) Z_i Y_i(u) dN(u)}{[\sum_{i=1}^n \exp(\beta_n^{*T} Z_i) Y_i(u)]^2} \rightarrow \mu(t)$$

–

$$n^{1/2}(\widehat{\beta}_n - \beta_0) \simeq \frac{n^{-1/2} \sum_{i=1}^n \int_0^\infty \{Z_i - \bar{Z}\} dM_i(u; \beta_0)}{\int_0^\infty S_Z^2(u; \beta_0) dN(u)} \rightarrow_D \mathcal{N}(0, \sigma^{-2})$$

- covariance between Term I and Term II

$$\sum_{i=1}^n \int_0^t \frac{dM_i(t; \beta_0)}{\sum_{i=1}^n \exp(\beta_0^T Z_i) Y_i(u)}, \quad \sum_{i=1}^n \int_0^\infty \{Z_i - \bar{Z}(u)\} dM_i(u; \beta_0)$$

- predictable covariation

$$\sum_{i=1}^n \int_0^t \frac{\sum_{i=1}^n \{Z_i - \bar{Z}(u)\} \exp(\beta_0^T Z_i) Y_i(u) \lambda_0(u) du}{\sum_{i=1}^n \exp(\beta_0^T Z_i) Y_i(u)} = 0$$

- Term I and II are uncorrelated

–

$$n^{1/2} [\hat{\Lambda}_0(t; \hat{\beta}_n) - \Lambda_0(t)] \rightarrow \mathcal{N} \left( 0, \sigma_1^2(t) + \frac{\mu(t)^2}{\sigma^2} \right)$$