STAT3655 Survival Analysis

Yu Gu, PhD
Assistant Professor

Department of Statistics & Actuarial Science
The University of Hong Kong

Table of Contents

Table of Contents

Table of Contents

Regression Modeling

- In many applications, it is interesting to study the associations between the failure time and the covariates/risk factors.
    - Does smoking increase the risk of lung cancer?
    - Are COVID-19 vaccines effective against infection/hospitalization/death?
    - Do people with type 2 diabetes have higher risk of Alzheimer's disease?
- This kind of questions can be formulated through regression models, where the covariates are denoted by a p-dimensional vector $X = (x_1, x_2, \ldots, x_p)^{\mathrm{T}}$, with each element representing a covariate.
- For example, $x_1 =$ age, $x_2 =$ gender, $x_3 =$ smoking status, etc.

Regression Models for Survival Data

- In survival analysis, the most common regression models take the form

$$\lambda(t; X) = \lambda_0(t) \exp\{\beta^{\mathsf{T}} X\} \tag{1}$$

  - $\lambda(t; X)$: covariate-specific hazard function
  - $\lambda_0(t)$: unknown baseline hazard function
  - β: p-dimensional unknown regression parameters

- Special cases of parametric models:
  - If $\lambda_0(t) \equiv \lambda$, model (1) becomes exponential regression model.
  - If $\lambda_0(t) = \lambda r(\lambda t)^{r-1}$, model (1) becomes Weibull regression model.

- In this chapter, we do NOT make any parametric assumptions on $\lambda_0(t)$. Then model (1) is the Cox proportional hazards (PH) model.

Proportional Hazards

- For two subjects with covariates X_1 and X_2, their hazard ratio over time is

$$HR(t; X_1, X_2) = \frac{\lambda(t; X_1)}{\lambda(t; X_2)} = \frac{\lambda_0(t)\exp\{\beta^\top X_1\}}{\lambda_0(t)\exp\{\beta^\top X_2\}} = \exp\{\beta^\top (X_1 - X_2)\},$$

  which is constant over t. This property is called proportional hazards.

- For the jth covariate, e^{β_j} is hazard ratio and β_j is log hazard ratio.

- Generalizations of Cox PH model:
  - Time-dependent covariates $X(t)$: blood pressure, air pollution, vaccination status, number of tumor relapse
  - Time-varying coefficient $\beta(t)$: useful for evaluating long-term treatment effects (e.g., COVID-19 vaccine efficacy)
  - Stratification $\lambda_{0s}(t)$: stratum s is determined by some covariates such as age, gender, and treatment arm

  The proportional hazards property no longer holds.

Cox PH Model Versus Logistic Model

- The Cox PH model is closely related to the logistic regression model. To see this, we discretize the continuous failure time T by defining

$$T^* = s_l \quad \text{if} \quad s_l \le T < s_{l+1},$$

where $\{s_l : l = 0, 1, 2, \dots\}$ is an arbitrary partition of $[0, \infty)$.

- For the discrete variable T^*, its conditional hazard function given the covariates X is given by

$$\begin{aligned}\lambda^*(s_l; X) &= \Pr(T^* = s_l \mid T^* \ge s_l, X) \\ &= 1 - \exp\left\{-\int_{s_l}^{s_{l+1}} \lambda(u; X) du\right\}\end{aligned}$$

- Conditional on $T^* \ge s_l$, we specify a logistic regression model for the binary outcome $I(T^* = s_l)$:

$$\log \frac{\lambda^*(s_l; X)}{1 - \lambda^*(s_l; X)} = \alpha_l + \beta^\mathrm{T} X, \qquad \text{for } l = 0, 1, 2, \dots$$

Cox PH Model Versus Logistic Model

- Define $\lambda_0(t) = \lambda(t; X = 0)$. It can be easily observed that

$$\frac{\lambda^*(s_l; X)}{1 - \lambda^*(s_l; X)} = \frac{\lambda^*(s_l; X = 0)}{1 - \lambda^*(s_l; X = 0)} e^{\beta^\top X}$$

$$\Rightarrow \frac{1 - \exp\left\{-\int_{s_l}^{s_{l+1}} \lambda(u; X) du\right\}}{\exp\left\{-\int_{s_l}^{s_{l+1}} \lambda(u; X) du\right\}} = \frac{1 - \exp\left\{-\int_{s_l}^{s_{l+1}} \lambda_0(u) du\right\}}{\exp\left\{-\int_{s_l}^{s_{l+1}} \lambda_0(u) du\right\}} e^{\beta^\top X}$$

- Since the above proportionality holds for any partition, it implies that

$$\frac{1 - \exp\left\{-\int_t^{t+\Delta t} \lambda(u; X) du\right\}}{1 - \exp\left\{-\int_t^{t+\Delta t} \lambda_0(u) du\right\}} = \frac{\exp\left\{-\int_t^{t+\Delta t} \lambda(u; X) du\right\}}{\exp\left\{-\int_t^{t+\Delta t} \lambda_0(u) du\right\}} e^{\beta^\top X}$$

- Letting $\Delta t \downarrow 0$ and applying L'Hôpital's rule to the left-hand side yields

$$\frac{\lambda(t; X)}{\lambda_0(t)} = e^{\beta^\top X} \Rightarrow \lambda(t; X) = \lambda_0(t) e^{\beta^\top X} \quad \text{(Cox PH model)}$$

Estimation for Cox Model

- Cox model is a semiparametric model in that it contains both finite-dimensional parameter β and infinite-dimensional parameter $\lambda_0(t)$.

- The primary interest usually lies in the estimation of β, such that $\lambda_0(t)$ is regarded as nuisance parameter and ideally should be eliminated from the estimation procedure.

- In some cases, however, the estimation of $\lambda_0(t)$ is also useful. For example, to predict a patient's survival outcome, both β and $\lambda_0(t)$ need to be estimated.

Table of Contents

Marginal and Conditional Likelihood

- General notation:
  - Z: vector of observations with density $f_Z(z;\theta)$
  - θ: vector of parameters, $\theta = (\beta, \lambda)$
  - β: parameter of interest (finite-dimensional)
  - λ: nuisance parameter (infinite-dimensional)

- If $Z = (V^{\mathsf{T}}, W^{\mathsf{T}})^{\mathsf{T}}$, the likelihood for θ can be written as

$$f_Z(z;\theta) = \underbrace{f_{W|V}(w|v;\theta)}_{\text{conditional likelihood}} \times \underbrace{f_V(v;\theta)}_{\text{marginal likelihood}} \tag{2}$$

- Even in complex models, one of the conditional and marginal likelihoods above may not involve λ, and can be used directly for inference on β.

- The gain in avoiding the estimation of λ may compensate for any loss in efficiency by using only part of the likelihood in (2).

Partial Likelihood: A Generalization

- Now suppose that Z can be transformed into a sequence of pairs $(V_1, W_1, V_2, W_2, \ldots, V_K, W_K)$. The likelihood for θ can be written as

$$\begin{aligned}
f_Z(z;\theta) &= f_{V_1,W_1,V_2,W_2,\ldots,V_K,W_K}(v_1, w_1, v_2, w_2, \ldots, v_K, w_K; \theta) \\
&= \prod_{k=1}^{K} \Big\{ f_{W_k|V_1,W_1,\ldots,V_{k-1},W_{k-1},V_k}(w_k|v_1, w_1, \ldots, v_{k-1}, w_{k-1}, v_k; \theta) \\
&\qquad\qquad \times f_{V_k|V_1,W_1,\ldots,V_{k-1},W_{k-1}}(v_k|v_1, w_1, \ldots, v_{k-1}, w_{k-1}; \theta) \Big\} \\
&= \underbrace{\left\{ \prod_{k=1}^{K} f_{W_k|Q_k}(w_k|q_k;\theta) \right\}}_{\text{partial likelihood: free of } \lambda} \times \left\{ \prod_{k=1}^{K} f_{V_k|P_k}(v_k|p_k;\theta) \right\},
\end{aligned} \tag{3}$$

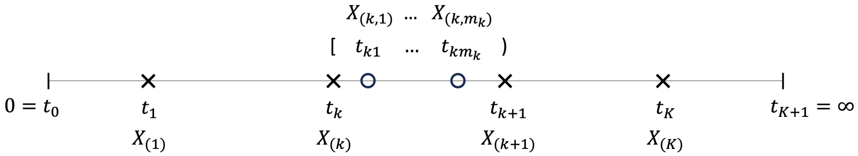
where $P_1 = \emptyset$, $Q_1 = V_1$, and for $k = 2, \ldots, K$, $P_k = (V_1, W_1, \ldots, V_{k-1}, W_{k-1})$ and $Q_k = (V_1, W_1, \ldots, V_{k-1}, W_{k-1}, V_k)$.

- Cox (1975)[1] suggests using partial likelihood for inference on β.

[1] *Cox, D. R. (1975). Partial likelihood. Biometrika, 62(2), 269-276.*

Partial Likelihood for Cox PH Model

- Original data: (Y_i, δ_i, X_i) $(i = 1, \dots, n)$
  - Assume no ties among observed failure times
  - Independent censoring assumption: $T_i \perp\!\!\!\perp C_i \mid X_i$
- Transformed data:
  - $t_1 < \cdots < t_k < \cdots < t_K$: observed failure times
  - $(1), \dots, (k), \dots, (K)$: labels for failing subjects $(T_{(k)} = t_k)$
  - $X_{(1)}, \dots, X_{(k)}, \dots, X_{(K)}$: covariates for failing subjects
  - Data from m_k subjects censored within $[t_k, t_{k+1})$ $(k = 0, \dots, K)$:
    - $t_{k1}, \dots, t_{km_k}$: observed censoring times
    - $(k, 1), \dots, (k, m_k)$: labels for censored subjects
    - $X_{(k,1)}, \dots, X_{(k,m_k)}$: covariates for censored subjects



Partial Likelihood for Cox PH Model (Cont.)

- Conditional on the covariates $\{X_i : i = 1, \ldots, n\}$, we construct $(V_1, W_1, \ldots, V_K, W_K)$ as follows: for $k = 1, \ldots, K$,

$$
\begin{aligned}
V_k &= \left[t_k, \left\{t_{k-1,l}, (k-1, l) : \; l = 1 \ldots, m_{k-1}\right\}\right] \\
&= \text{one failure at } t_k + \text{times and labels of all censorings in } [t_{k-1}, t_k), \\
W_k &= \{(k)\} \\
&= \text{label for the failing subject at } t_k.
\end{aligned}
$$

- Thus,

$$
\begin{aligned}
P_k &= (V_1, W_1, \ldots, V_{k-1}, W_{k-1}) \\
&= \text{times and labels of all censorings in } [0, t_{k-1}) \\
&\qquad + \text{times and labels of all failures in } [0, t_{k-1}], \\
Q_k &= (V_1, W_1, \ldots, V_{k-1}, W_{k-1}, V_k) \\
&= P_k + V_k \\
&= \text{failure and censoring history up to } t_k^- + \text{one failure at } t_k.
\end{aligned}
$$

Partial Likelihood for Cox PH Model (Cont.)

- By the definition of partial likelihood in (3), we only need to derive the conditional distribution of W_k given Q_k.

- Define $\mathcal{R}_k = \{i : \; Y_i \geq t_k\}$ to be the risk set at t_k. Then

$$
\begin{aligned}
& \Pr\{W_k = (k) \mid Q_k\} \\
= & \Pr\{\text{subject } (k) \text{ fails at } t_k \mid \mathcal{R}_k, \text{ one failure at } t_k\} \\
= & \frac{\Pr\{T_{(k)} \in [t_k, t_k + dt) \mid T_{(k)} \geq t_k\} \prod_{j \in \mathcal{R}_k \setminus \{(k)\}} \Pr\{T_j \notin [t_k, t_k + dt) \mid T_j \geq t_k\}}{\sum_{i \in \mathcal{R}_k} \left[\Pr\{T_i \in [t_k, t_k + dt) \mid T_i \geq t_k\} \prod_{j \in \mathcal{R}_k \setminus \{i\}} \Pr\{T_j \notin [t_k, t_k + dt) \mid T_j \geq t_k\} \right]} \\
= & \frac{\lambda\{t_k; X_{(k)}\} dt \prod_{j \in \mathcal{R}_k \setminus \{(k)\}} \{1 - \lambda(t_k; X_j) dt\}}{\sum_{i \in \mathcal{R}_k} \left[\lambda(t_k; X_i) dt \prod_{j \in \mathcal{R}_k \setminus \{i\}} \{1 - \lambda(t_k; X_j) dt\} \right]} \\
\approx & \frac{\lambda\{t_k; X_{(k)}\}}{\sum_{i \in \mathcal{R}_k} \lambda(t_k; X_i)} \\
= & \frac{\exp\{\beta^{\mathsf{T}} X_{(k)}\}}{\sum_{i \in \mathcal{R}_k} \exp(\beta^{\mathsf{T}} X_i)}
\end{aligned}
$$

Partial Likelihood for Cox PH Model (Cont.)

- Thus, the partial likelihood for the Cox PH model is given by

$$
\begin{aligned}
L(\beta) &= \prod_{k=1}^{K} \Pr\{W_k = (k) \mid Q_k\} \\
&= \prod_{k=1}^{K} \frac{\exp\{\beta^{\mathsf{T}} X_{(k)}\}}{\sum_{i \in \mathcal{R}_k} \exp(\beta^{\mathsf{T}} X_i)}
\end{aligned} \tag{4}
$$

- If we further assume noninformative censoring, that is,

$$\Pr\{\text{subjects censored in } [t, t+dt) \mid \text{risk set at } t, \text{subjects failing at } t\}$$

  does not depend on β, then the second term $\prod_{k=1}^{K} f_{V_k|P_k}(v_k|p_k;\theta)$ in (3) contains little or no information about β.

- Therefore, loss in efficiency arising from the use of partial likelihood for inference on β is negligible.

Partial Likelihood = Marginal Likelihood

- Interestingly, partial likelihood can also be derived as a marginal likelihood for the ranks of $\{T_i : \; i = 1, \ldots, n\}$.

- We first consider the simple setting without censoring. The marginal likelihood of the ranks is given by

$$\Pr\{T_{(1)} < T_{(2)} < \cdots < T_{(n)}\} = \int_0^\infty \int_{t_1}^\infty \cdots \int_{t_{n-1}}^\infty \prod_{i=1}^n f\{t_i; X_{(i)}\} dt_n \cdots dt_1.$$

- Let $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$ be the cumulative baseline hazard function. First,

$$\begin{aligned} &\int_{t_{n-1}}^\infty f\{t_n; X_{(n)}\} dt_n \\ &= S\{t_{n-1}; X_{(n)}\} = \exp\left[-\Lambda_0(t_{n-1}) \exp\{\beta^\mathsf{T} X_{(n)}\}\right] \\ &= \left[\prod_{i \geq n} \frac{\exp\{\beta^\mathsf{T} X_{(i)}\}}{\sum_{j \geq i} \exp\{\beta^\mathsf{T} X_{(j)}\}}\right] \exp\left[-\Lambda_0(t_{n-1}) \sum_{j \geq n} \exp\{\beta^\mathsf{T} X_{(j)}\}\right]. \end{aligned}$$

Partial Likelihood = Marginal Likelihood (Cont.)

- Next,

$$\int_{t_{n-2}}^{\infty} f\{t_{n-1}; X_{(n-1)}\} \left[\prod_{i \geq n} \frac{\exp\{\beta^{\mathsf{T}} X_{(i)}\}}{\sum_{j \geq i} \exp\{\beta^{\mathsf{T}} X_{(j)}\}}\right] \exp\left[-\Lambda_0(t_{n-1}) \sum_{j \geq n} \exp\{\beta^{\mathsf{T}} X_{(j)}\}\right] dt_{n-1}$$

$$= \left[\prod_{i \geq n} \frac{\exp\{\beta^{\mathsf{T}} X_{(i)}\}}{\sum_{j \geq i} \exp\{\beta^{\mathsf{T}} X_{(j)}\}}\right] \int_{t_{n-2}}^{\infty} \lambda\{t_{n-1}; X_{(n-1)}\} \exp\left[-\Lambda_0(t_{n-1}) \sum_{j \geq n-1} \exp\{\beta^{\mathsf{T}} X_{(j)}\}\right] dt_{n-1}$$

$$= \left[\prod_{i \geq n-1} \frac{\exp\{\beta^{\mathsf{T}} X_{(i)}\}}{\sum_{j \geq i} \exp\{\beta^{\mathsf{T}} X_{(j)}\}}\right] \exp\left[-\Lambda_0(t_{n-2}) \sum_{j \geq n-1} \exp\{\beta^{\mathsf{T}} X_{(j)}\}\right].$$

- Recursive calculation yields

$$\Pr\{T_{(1)} < T_{(2)} < \cdots < T_{(n)}\} = \prod_{i=1}^{n} \frac{\exp\{\beta^{\mathsf{T}} X_{(i)}\}}{\sum_{j \geq i} \exp\{\beta^{\mathsf{T}} X_{(j)}\}},$$

which is equal to the partial likelihood in (4) for uncensored data.

Partial Likelihood = Marginal Likelihood (Cont.)

- When there are censored subjects, the marginal likelihood is the sum of $\Pr\{T_{(1)} < \cdots < T_{(n)}\}$ over all ranks of $\{T_i : i = 1, \ldots, n\}$ that are consistent with the observed data.

- For example, suppose $(Y_1, Y_2, Y_3, Y_4) = (28, 15, 17, 6)$ and $(\delta_1, \delta_2, \delta_3, \delta_4) = (0, 0, 1, 1)$. Then all possible ranks are $(4, 2, 3, 1)$, $(4, 3, 1, 2)$, and $(4, 3, 2, 1)$.

- Using the original labels $(1), \ldots, (K)$ for failing subjects and $(k, 1), \ldots, (k, m_k)$ for censored subjects in $[t_k, t_{k+1})$, the marginal likelihood can be written as

$$\Pr\Big[T_{(1)} < \cdots < T_{(K)},\ \big\{T_{(k,l)} > T_{(k)} :\ k = 1, \ldots, K;\ l = 1, \ldots, m_k\big\}\Big].$$

Partial Likelihood = Marginal Likelihood (Cont.)

- Conditional on $T_{(k)} = t_k$, the likelihood contribution from the m_k censored subjects is

$$g(t_k) = \prod_{l=1}^{m_k} \Pr\{T_{(k,l)} > t_k\} = \exp\left[-\Lambda_0(t_k) \sum_{l=1}^{m_k} \exp\{\beta^{\mathsf{T}} X_{(k,l)}\}\right]$$

- Thus, the marginal likelihood reduces to

$$\int_0^\infty \int_{t_1}^\infty \cdots \int_{t_{K-1}}^\infty \prod_{k=1}^{K} f\{t_k; X_{(k)}\} g(t_k) dt_K \cdots dt_1$$
$$= \prod_{k=1}^{K} \frac{\exp\{\beta^{\mathsf{T}} X_{(k)}\}}{\sum_{i \in \mathcal{R}_k} \exp\{\beta^{\mathsf{T}} X_i\}},$$

which is exactly the partial likelihood in (4).

- The equivalence of the partial and marginal likelihoods suggests that inferences based on partial likelihood are efficient.

Partial Likelihood for Tied Data

- Although failure time is a continuous random variable, ties in observed failure times are still possible in practice (e.g., when failure times are measured in integer days).

- Notation for tied data:
  - $t_1 < \cdots < t_k < \cdots < t_K$: distinct observed failure times
  - $\mathcal{R}_k = \{i : \; Y_i \geq t_k\}$: risk set at t_k
  - $\mathcal{D}_k = \{i : \; Y_i = t_k, \delta_i = 1\}$: set of all subjects failing at t_k
  - $d_k = |\mathcal{D}_k|$: number of subjects failing at t_k

- We can follow a similar procedure to derive the partial likelihood for tied data. The conditional probability of W_k given Q_k now becomes

$$\Pr\{W_k = \mathcal{D}_k \mid Q_k\} = \Pr\{\text{subjects in } \mathcal{D}_k \text{ fail at } t_k \mid \mathcal{R}_k,\, d_k \text{ failures at } t_k\}. \tag{5}$$

Partial Likelihood for Tied Data (Cont.)

- Recall that Cox PH model can be approximated by logistic model:

$$\log \frac{\lambda(t;X)dt}{1-\lambda(t;X)dt} = \alpha(t) + \beta^{\mathsf{T}} X$$

$$\Rightarrow \Pr\{\text{subject fails at } t_k \mid \text{at risk at } t_k,\, X\} = \frac{\exp(\alpha_k + \beta^{\mathsf{T}} X)}{1+\exp(\alpha_k + \beta^{\mathsf{T}} X)}$$

- Under the logistic model, the conditional probability in (5) is given by

$$\frac{\prod_{i\in\mathcal{D}_k} \frac{\exp(\alpha_k+\beta^{\mathsf{T}} X_i)}{1+\exp(\alpha_k+\beta^{\mathsf{T}} X_i)} \prod_{i\in\mathcal{R}_k\setminus\mathcal{D}_k} \frac{1}{1+\exp(\alpha_k+\beta^{\mathsf{T}} X_i)}}{\sum_{\mathcal{D}\in\mathcal{C}(\mathcal{R}_k,d_k)} \prod_{i\in\mathcal{D}} \frac{\exp(\alpha_k+\beta^{\mathsf{T}} X_i)}{1+\exp(\alpha_k+\beta^{\mathsf{T}} X_i)} \prod_{i\in\mathcal{R}_k\setminus\mathcal{D}} \frac{1}{1+\exp(\alpha_k+\beta^{\mathsf{T}} X_i)}}$$

$$= \frac{\exp(\beta^{\mathsf{T}} S_{\mathcal{D}_k})}{\sum_{\mathcal{D}\in\mathcal{C}(\mathcal{R}_k,d_k)} \exp(\beta^{\mathsf{T}} S_{\mathcal{D}})},$$

where $\mathcal{C}(\mathcal{R}_k, d_k)$ is the collection of all sets of d_k failing subjects chosen from $\mathcal{R}_k$, and $S_{\mathcal{D}} = \sum_{i\in\mathcal{D}} X_i$.

Partial Likelihood for Tied Data (Cont.)

- The partial likelihood for tied data then follows:

$$L(\beta) = \prod_{k=1}^{K} \frac{\exp(\beta^{\mathsf{T}} S_{\mathcal{D}_k})}{\sum_{\mathcal{D} \in \mathcal{C}(\mathcal{R}_k, d_k)} \exp(\beta^{\mathsf{T}} S_{\mathcal{D}})}$$

- The computation of the above partial likelihood can be very intensive since the denominator requires enumeration of all possible failing set.

- Some approximation methods have been proposed to simplify the computation, including the Breslow and Efron approximations.

Breslow and Efron Approximations

- Breslow approximation[2] is the easiest method for handling ties. It suggests the partial likelihood

$$L(\beta) = \prod_{k=1}^{K} \frac{\exp(\beta^{\mathsf{T}} S_k)}{\left\{\sum_{i \in \mathcal{R}_k} \exp(\beta^{\mathsf{T}} X_i)\right\}^{d_k}},$$

  where $S_k = \sum_{i \in \mathcal{D}_k} X_i$.

- Alternatively, Efron approximation[3] suggests the partial likelihood

$$L(\beta) = \prod_{k=1}^{K} \frac{\exp(\beta^{\mathsf{T}} S_k)}{\prod_{i=1}^{d_k} \left\{\sum_{j \in \mathcal{R}_k} \exp(\beta^{\mathsf{T}} X_j) - \frac{i-1}{d_k} \sum_{j \in \mathcal{D}_k} \exp(\beta^{\mathsf{T}} X_j)\right\}}$$

[2] *Breslow, N. (1974). Covariance analysis of censored survival data. Biometrics, 89-99.*

[3] *Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. Journal of the American Statistical Association, 72(359), 557-565.*

Breslow and Efron Approximations (Cont.)

- When there are a large number of ties, Efron approximation is more accurate than Breslow approximation.
- When the number of ties is small, there is typically little difference between the two approaches.
- Many software implement the Breslow approach for its simplicity, but the "survival" package uses Efron approximation as the default.
- For simplicity, the remaining sections will be based on Breslow's likelihood.

Table of Contents

Maximum Partial Likelihood Estimation

Breslow's partial likelihood:

$$L_n(\beta) = \prod_{k=1}^{K} \frac{\exp(\beta^{\mathsf{T}} S_k)}{\left\{\sum_{i \in \mathcal{R}_k} \exp(\beta^{\mathsf{T}} X_i)\right\}^{d_k}} = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta^{\mathsf{T}} X_i)}{\sum_{j=1}^{n} R_j(Y_i) \exp(\beta^{\mathsf{T}} X_j)} \right\}^{\delta_i},$$

where $R_j(t) = I(Y_j \geq t)$ is the at-risk indicator at time t for the jth subject.

Log partial likelihood:

$$\ell_n(\beta) = \sum_{i=1}^{n} \delta_i \left[\beta^{\mathsf{T}} X_i - \log\left\{ \sum_{j=1}^{n} R_j(Y_i) e^{\beta^{\mathsf{T}} X_j} \right\} \right]$$

Maximum partial likelihood estimator:

$$\hat{\beta} = \arg\max_{\beta} \ell_n(\beta)$$

Newton-Raphson Algorithm

Score function:

$$U_n(\beta) = \dot{\ell}_n(\beta) = \sum_{i=1}^{n} \delta_i \left[X_i - \frac{\sum_{j=1}^{n} R_j(Y_i) e^{\beta^\top X_j} X_j}{\sum_{j=1}^{n} R_j(Y_i) e^{\beta^\top X_j}} \right]$$

Information matrix:

$$\mathcal{I}_n(\beta) = -\ddot{\ell}_n(\beta) = \sum_{i=1}^{n} \delta_i \left[\frac{\sum_{j=1}^{n} R_j(Y_i) e^{\beta^\top X_j} X_j^{\otimes 2}}{\sum_{j=1}^{n} R_j(Y_i) e^{\beta^\top X_j}} - \frac{\left\{ \sum_{j=1}^{n} R_j(Y_i) e^{\beta^\top X_j} X_j \right\}^{\otimes 2}}{\left\{ \sum_{j=1}^{n} R_j(Y_i) e^{\beta^\top X_j} \right\}^{2}} \right]$$

Updating formula:

$$\hat{\beta}^{(\text{new})} = \hat{\beta}^{(\text{old})} + \left[\mathcal{I}_n\{\hat{\beta}^{(\text{old})}\} \right]^{-1} U_n\{\hat{\beta}^{(\text{old})}\}$$

Large-Sample Theory for $\hat{\beta}$

Theorem (Consistency and limiting distribution)

Under certain regularity conditions, the following are true:

(i) $\hat{\beta} \xrightarrow{p} \beta$.

(ii) $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Sigma^{-1})$, *where* $\Sigma = \lim_{n\to\infty} n^{-1}\mathcal{I}_n(\beta)$.

(iii) $n^{-1/2}U_n(\beta) \xrightarrow{d} N(0, \Sigma)$.

Regularity conditions:

(i) Subjects are i.i.d.

(ii) $X_i(\cdot)$ are bounded.

(iii) $\int_0^\tau \lambda_0(t)dt < \infty$, where $\tau = \sup_{i=1}^n Y_i$.

(iv) For any $t \in [0, \tau]$, $\Pr\{R_i(t) = 1\} > 0$.

(v) Σ is positive definite.

Hypothesis Testing for β

- $H_0 : \beta = \beta^*$
- Wald test:

$$W_n = (\hat{\beta} - \beta^*)^{\mathsf{T}} \mathcal{I}_n(\hat{\beta})(\hat{\beta} - \beta^*) \xrightarrow{d} \chi_p^2 \quad \text{under } H_0$$

- Score test:

$$SC_n = U_n(\beta^*)^{\mathsf{T}} \{\mathcal{I}_n(\beta^*)\}^{-1} U_n(\beta^*) \xrightarrow{d} \chi_p^2 \quad \text{under } H_0$$

- Likelihood ratio test:

$$LR_n = 2\{\ell_n(\hat{\beta}) - \ell_n(\beta^*)\} \xrightarrow{d} \chi_p^2 \quad \text{under } H_0$$

Test A Subset of Parameters

- Partition of parameters and statistics:

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}, \quad U_n = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix},$$

$$\mathcal{I}_n = \begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{pmatrix}, \quad \mathcal{I}_n^{-1} = \begin{pmatrix} \mathcal{I}^{11} & \mathcal{I}^{12} \\ \mathcal{I}^{21} & \mathcal{I}^{22} \end{pmatrix},$$

where $\mathcal{I}^{11} = (\mathcal{I}_{11} - \mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21})^{-1}$.

- $H_0 : \beta_1 = \beta_1^*$, where β_1 is a q-dimensional subvector of β.

- Wald test:

$$W_n = (\hat{\beta}_1 - \beta_1^*)^{\mathsf{T}} \{\mathcal{I}^{11}(\hat{\beta})\}^{-1} (\hat{\beta}_1 - \beta_1^*) \xrightarrow{d} \chi_q^2 \quad \text{under } H_0$$

Test A Subset of Parameters (Cont.)

- Score test:

$$SC_n = U_1(\beta_1^*, \tilde{\beta}_2)^{\mathsf{T}} \{\mathcal{I}^{11}(\beta_1^*, \tilde{\beta}_2)\} U_1(\beta_1^*, \tilde{\beta}_2) \xrightarrow{d} \chi_q^2 \quad \text{under } H_0,$$

where $\tilde{\beta}_2 = \arg\max_{\beta_2} \ell_n(\beta_1^*, \beta_2)$ is the restricted MLE under $\beta_1 = \beta_1^*$.

$$\begin{aligned}[\text{Hint: } & U_1(\beta_1^*, \tilde{\beta}_2) = U_1(\beta_1^*, \beta_2) - \mathcal{I}_{12}(\beta_1^*, \beta_2)(\tilde{\beta}_2 - \beta_2) + o(1) \\ &= U_1(\beta_1^*, \beta_2) - \mathcal{I}_{12}(\beta_1^*, \beta_2)\{\mathcal{I}_{22}(\beta_1^*, \beta_2)\}^{-1} U_2(\beta_1^*, \beta_2) + o(1)]\end{aligned}$$

- Likelihood ratio test:

$$LR_n = 2\{\ell_n(\hat{\beta}) - \ell_n(\beta_1^*, \tilde{\beta}_2)\} \xrightarrow{d} \chi_q^2 \quad \text{under } H_0$$

Test on Linear Combination of Parameters

- $H_0 : C\beta = C\beta^*$, where C is a $q \times p$ matrix of full rank q $(q \leq p)$.
- Wald test:

$$W_n = (C\hat{\beta} - C\beta^*)^{\mathsf{T}} \left[C\{\mathcal{I}_n(\hat{\beta})\}^{-1} C^{\mathsf{T}} \right]^{-1} (C\hat{\beta} - C\beta^*) \xrightarrow{d} \chi_q^2 \quad \text{under } H_0$$

- In clinical trials, this kind of tests are useful for comparing effects of different treatments.
- For example, suppose that x_1 and x_2 are the binary indicators for treatments 1 and 2, respectively. To test the difference between the two treatments, we can let $C = (1, -1)$.

Table of Contents

Estimation of Λ_0

- Several methods have been proposed to estimate infinite-dimensional parameters related to λ_0. One appealing estimator of $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ is Breslow estimator:

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \frac{I(Y_i \leq t)\delta_i}{\sum_{j=1}^{n} R_j(Y_i)\exp(\hat{\beta}^\mathsf{T} X_j)} = \sum_{k=1}^{K} \frac{I(t_k \leq t)d_k}{\sum_{i \in \mathcal{R}_k} \exp(\hat{\beta}^\mathsf{T} X_i)}$$

- Breslow estimator is a natural generalization of the Nelson-Aalen estimator for homogeneous samples. When there are no covariates, $\hat{\Lambda}_0$ reduces to the NA estimator.

- The rationale behind is that one subject in the risk set failing at rate $\lambda_0(t)e^{\hat{\beta}^\mathsf{T} X_i}$ produces the same expected number of failures as $e^{\hat{\beta}^\mathsf{T} X_i}$ subjects, each failing with rate $\lambda_0(t)$.

Weak Convergence of $\hat{\Lambda}_0$

Theorem (Limiting distribution of $\hat{\Lambda}_0$)

Under certain regularity conditions, the stochastic process $G(t) = \sqrt{n}\{\hat{\Lambda}_0(t) - \Lambda_0(t)\}$ converges weakly to a mean-zero Gaussian process whose covariance function can be consistently estimated by

$$\widehat{Cov}\{G(s), G(t)\} = \left\{\int_0^s E(\hat{\beta}, u)d\hat{\Lambda}_0(u)\right\}^T \{\mathcal{I}_n(\hat{\beta})/n\}^{-1} \left\{\int_0^t E(\hat{\beta}, u)d\hat{\Lambda}_0(u)\right)\right\} + \int_0^{s\wedge t} \frac{d\hat{\Lambda}_0(u)}{S^{(0)}(\hat{\beta}, u)},$$

where

$$S^{(r)}(\beta, t) = n^{-1}\sum_{i=1}^{n} R_i(t)e^{\beta^T X_i} X_i^{\otimes r} \qquad \text{for } r = 0, 1, 2;$$

$$E(\beta, t) = \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)}.$$

Estimation of λ_0

Estimation of $\lambda_0(t)$ can be done by applying the kernel smoothing method to the Breslow estimator, as we did based on the NA estimator in Chapter 3.

Estimation of Survival Function

- The covariate-specific survival function is

$$S(t;X) = \exp\left\{-\Lambda_0(t)e^{\beta^\top X}\right\} = S_0(t)^{\exp(\beta^\top X)}$$

- We simply plug in $\hat{\beta}$ and $\hat{\Lambda}_0$ to estimate $S(t;X)$:

$$\hat{S}(t;X) = \big[\underbrace{\exp\{-\hat{\Lambda}_0(t)\}}_{\hat{S}_0(t)}\big]^{\exp(\hat{\beta}^\top X)}$$

Table of Contents

Model Misspecification

Cox model relies on the following assumptions:

- Proportional hazards (PH): hazard ratio is constant over time
- Functional forms of covariates: e.g., age or log(age)?
- Link function: $\psi(\beta^\mathsf{T} X) = \beta^\mathsf{T} X$

When at least one of these assumptions does not hold, the model is misspecified, resulting in loss of power for testing covariate effects and even biased regression parameter estimates.

Asymptotic Properties Under Misspecified Models

Let β^* be the limit of the solution to the score equation $U_n(\beta) = 0$.

(i) $U_n(\beta^*) \dot{\sim} N(0, B)$, where $B = \sum_{i=1}^n W_i^{\otimes 2}$ and

$$W_i = \delta_i \left\{ X_i - \frac{S^{(1)}(Y_i; \widehat{\beta})}{S^{(0)}(Y_i; \widehat{\beta})} \right\} - \sum_{j=1}^n \frac{\delta_j R_i(Y_j) \exp\{\widehat{\beta}^{\mathrm{T}} X_i\}}{n S^{(0)}(Y_j; \widehat{\beta})} \left\{ X_i - \frac{S^{(1)}(Y_j; \widehat{\beta})}{S^{(0)}(Y_j; \widehat{\beta})} \right\}$$

(ii) $\widehat{\beta} \dot{\sim} N(\beta^*, D)$, where $D = \mathcal{I}_n^{-1}(\widehat{\beta}) B \mathcal{I}_n^{-1}(\widehat{\beta})$.

D is called the robust variance estimator, which is always valid. In contrast, the model-based variance estimator $\mathcal{I}_n^{-1}(\widehat{\beta})$ may not be valid when the model is misspecified.

To compute robust variance estimator in R, simply specify “robust = TRUE” in coxph().

Stratified Cox Model

Setup: G strata, n_g subjects in the gth stratum

- T_{gi}: failure time of the ith subject in the gth stratum
- C_{gi}: censoring time of the ith subject in the gth stratum
- $Y_{gi} = \min(T_{gi}, C_{gi})$: observation time of the ith subject in the gth stratum
- $\delta_{gi} = I(T_{gi} \leq C_{gi})$: failure indicator of the ith subject in the gth stratum
- X_{gi}: covariates of the ith subject in the gth stratum

Observed data: $(Y_{gi}, \delta_{gi}, X_{gi})$, for $g = 1, \ldots, G$ and $i = 1, \ldots, n_g$

Independent censoring: $T_{gi} \perp\!\!\!\perp C_{gi}$ given X_{gi} within each stratum g

Stratified Cox Model (Cont.)

The stratified Cox model is given by

$$\lambda(t; X_{gi}) = \lambda_{0g}(t) e^{\beta^\top X_{gi}}, \quad \text{for } g = 1, \ldots, G; \ i = 1, \ldots, n_g$$

- $\lambda_{0g}(t)$: stratum-specific baseline hazard function
- β: common regression parameters across all strata

Notation:

- $R_{gi}(t) = I(Y_{gi} \geq t)$: at-risk indicator
- For $r = 0, 1, 2$, define

$$S_g^{(r)}(t; \beta) = \sum_{i=1}^{n_g} R_{gi}(t) e^{\beta^\top X_{gi}} X_{gi}^{\otimes r}$$

Maximum Partial Likelihood Estimation

$$L_n(\beta) = \prod_{g=1}^{G} \prod_{i=1}^{n_g} \left\{ \frac{e^{\beta^\top X_{gi}}}{S_g^{(0)}(Y_{gi};\beta)} \right\}^{\delta_{gi}}$$

$$U_n(\beta) = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \delta_{gi} \left\{ X_{gi} - \frac{S_g^{(1)}(Y_{gi};\beta)}{S_g^{(0)}(Y_{gi};\beta)} \right\}$$

$$\mathcal{I}_n(\beta) = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \delta_{gi} \left[\frac{S_g^{(2)}(Y_{gi};\beta)}{S_g^{(0)}(Y_{gi};\beta)} - \frac{\left\{ S_g^{(1)}(Y_{gi};\beta) \right\}^{\otimes 2}}{\left\{ S_g^{(0)}(Y_{gi};\beta) \right\}^2} \right]$$

You can easily verify that when $G = 1$, the above formulas reduce to those on Slides 28–29.

The asymptotic properties of $U_n(\beta)$ and $\widehat{\beta}$ are the same as the unstratified case (see Slide 30).

Fit Stratified Cox Model in R

```
# stratification on variable "vstrata"
coxph(Surv(time, status) ~ covariates + strata(vstrata),
      ties = "breslow")
```

Table of Contents

Accelerated Failure Time Model

- Another class of semiparametric survival models is the accelerated failure time model (AFT model), which specifies that the covariate effect is multiplicative on the time scale:

$$\log T = \alpha + \beta^{\mathsf{T}} X + \sigma W,$$

  where W is an error variable with unspecified density f of standard form.

- Under the AFT model, the role of the covariates X is to accelerate or decelerate the failure time T.

- The survival and hazard functions for T take the form

$$S(t; X) = \exp\left\{-\Lambda_0(te^{-\alpha-\beta^{\mathsf{T}} X})\right\},$$
$$\lambda(t; X) = \lambda_0(te^{-\alpha-\beta^{\mathsf{T}} X}) \exp(-\alpha - \beta^{\mathsf{T}} X),$$

  where $\lambda_0(t)$ is some unknown baseline hazard function and $\Lambda_0(t) = \int_0^t \lambda_0(u) du$.

Examples of AFT Model

- If $W \sim N(0, 1)$, then $\log T \sim N(\beta^{\mathsf{T}} X, \sigma^2)$. That is, T follows a log-normal distribution.

- If W follows the extreme value distribution with density and survival functions

  $$f_W(w) = \exp\left(w - e^w\right), \quad S_W(w) = \exp\left(-e^w\right),$$

  then T has a Weibull distribution with survival function

  $$S_T(t) = \exp\left(-\theta t^{\alpha}\right),$$

  where $\theta = \exp\left[-\left(\alpha + x'\beta\right)/\sigma\right]$ and $\alpha = 1/\sigma$. This is a parametric version of the class of proportional hazards models.

Examples of AFT Model (Cont.)

- If W follows the standard logistic distribution with density function

$$f(w) = \frac{e^w}{(1+e^w)^2},$$

then T follows a log-logistic distribution with survival function

$$S_T(t) = [1 + \theta t^\alpha]^{-1},$$

where $\theta = \exp[-(\alpha + x'\beta)/\sigma]$ and $\alpha = 1/\sigma$. This is a parametric version of the class of proportional odds models.