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

Censoring Mechanisms

In this course, we always assume that the failure time T and the censoring time C are independent given the covariates X.

Here, we discuss some commonly seen censoring mechanisms in detail.

- Right censoring
  - Type I censoring
  - Type II censoring
  - Type III censoring
- Interval censoring
  - Case 1 interval censoring
  - Case 2 interval censoring
  - Case k interval censoring
  - Mixed case interval censoring

Type I Censoring

- Suppose that a study ends at time τ and that every subject is followed until failure or the end of the study, whichever occurs first.
- For those who have experienced failure before or at τ, their failure times are exactly observed; otherwise the failure time is right-censored at τ.
- Mathematically, $C_i = \tau$, which implies $Y_i = \min(T_i, \tau)$.
- This is the most commonly used censoring mechanism in experimental medical research (e.g., clinical trials for new treatments).

Type II censoring

- The choice of τ might be tricky: too small τ yields insufficient number of observed failures, while too large τ requires high costs or delays important clinical decisions.

- Another option is to terminate the study when a prespecified number of failures have been observed (e.g., 80 failures out of 100 subjects) and the remaining 20 subjects will be regarded as censored.

- Specifically, let $T_{(1)} < T_{(2)} < \cdots < T_{(r)} < T_{(r+1)} < \cdots < T_{(n)}$ be the order statistics of $T_1, \ldots, T_n$. We only observe the first r failure times and all other failures times are right-censored at $T_{(r)}$.

- This guarantees enough number of observed failures while controlling the cost and duration of a study.

Type III censoring

- Suppose that the study period is fixed, the entry and censoring times may differ across subjects. In other words, censoring time C is a random instead of fixed variable.
- A subject's failure time T_i is observed if $T_i \leq C_i$ and is right-censored at C_i otherwise.
- This corresponds to our general notation $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$.
- Type III censoring is more well-known as random censoring and is common in clinical trials or observational studies of chronic diseases.

Interval Censoring

- Interval-censored data arise when the failure of interest is known only to occur within a time interval.

- Such data are commonly encountered in medical research where it is too expensive or even impossible to determine the exact failure time.

- Examples:
  - Alzheimer's disease: no definitive symptoms; exact disease onset time can only be determined through periodic cognitive tests.
  - Breast cancer: daily screening is too expensive and harmful; exact time to occurrence/recurrence of tumor can only be determined through periodic clinic visits.

- In such cases, ascertainment of failure can take place only at a small number of monitoring times. The resulting data are called interval-censored data.

Case 1 Interval Censoring

- Case 1 interval censored data is also known as current status data.
- Each subject is examined at only one random monitoring time Z_i and is NOT under observation at any other times.
- Thus, the only observed information for the ith subject is whether he/she has experienced a failure by time Z_i.
- Observed data:
  - Z_i
  - $\delta_i = I(T_i \leq Z_i)$

Case 2 Interval Censoring

- There are two random monitoring times for each subject, say (U_{i1}, U_{i2}) with $\Pr(U_{i1} < U_{i2}) = 1$.
- The only observed information is whether the failure occurs before or at U_{i1} (left censoring), within $(U_{i1}, U_{i2}]$, or after U_{i2} (right censoring).
- Observed data:
  - (U_{i1}, U_{i2})
  - $\delta_{i1} = I(T_i \leq U_{i1})$ and $\delta_{i2} = I(U_{i1} < T_i \leq U_{i2})$

Case k Interval Censoring

- Similar to Case 1 and Case 2 interval censoring, now we suppose that there are a sequence of k random monitoring times for each subject, denoted by $U_{i1} < U_{i2} < \cdots < U_{ik}$.

- In this case, the only observed information is whether the failure occurs before or at the first monitoring time (left censoring), between any two successive monitoring times, or after the last monitoring time (right censoring).

- Observed data:
  - $(U_{i1}, U_{i2}, \ldots, U_{ik})$
  - $\delta_{il} = I(U_{il} < T_i \leq U_{i,l+1})$, for $l = 0, \ldots, k$, with $U_{i0} = 0$ and $U_{i,k+1} = \infty$

Mixed Case Interval Censoring

- On the basis of case k interval censoring, now we allow the number of monitoring times k to differ across subjects. In other words, k is now a random variable rather than a fixed number.

- For each individual subject, we change k to k_i. Everything else is the same as in case k interval censoring.

- Observed data:
  - $(U_{i1}, U_{i2}, \ldots, U_{ik_i})$
  - $\delta_{il} = I(U_{il} < T_i \leq U_{i,l+1})$, for $l = 0, \ldots, k_i$, with $U_{i0} = 0$ and $U_{i,k_i+1} = \infty$

- In fact, the above observed data can be simplified as (L_i, R_i), where $(L_i, R_i]$ is the smallest interval that brackets T_i.

Table of Contents

Data and Likelihood

Assume independent censoring always holds, i.e., $T \perp\!\!\!\perp C \mid X$.

Right-censored data (under random censoring):

- Data: (Y_i, δ_i, X_i), $i = 1, \ldots, n$
- Likelihood:

$$L_n(\theta) \propto \prod_{i=1}^{n} f(Y_i;\theta)^{\delta_i} S(Y_i;\theta)^{1-\delta_i} = \prod_{i=1}^{n} \lambda(Y_i;\theta)^{\delta_i} S(Y_i;\theta)$$

Interval-censored data (under mixed case censoring):

- Data: (L_i, R_i, X_i), $i = 1, \ldots, n$
- Likelihood:

$$L_n(\theta) \propto \prod_{i=1}^{n} \Big\{ S(L_i;\theta) - S(R_i;\theta) \Big\}$$

Maximum Likelihood Estimation

- Let $\ell_n(\theta) = \log L_n(\theta)$ be the log-likelihood function, with gradient $\dot{\ell}_n(\theta)$ and Hessian $\ddot{\ell}_n(\theta)$ with respect to θ. Write $\theta = (\theta_1, \ldots, \theta_r)$, then

$$\dot{\ell}_n(\theta) = \frac{\partial}{\partial \theta} \ell_n(\theta) = \left(\frac{\partial}{\partial \theta_1} \ell_n(\theta), \cdots, \frac{\partial}{\partial \theta_r} \ell_n(\theta) \right)^{\mathsf{T}}$$

$$\ddot{\ell}_n(\theta) = \frac{\partial^2}{\partial \theta \partial \theta^{\mathsf{T}}} \ell_n(\theta) = \begin{pmatrix} \frac{\partial^2}{\partial \theta_1^2} \ell_n(\theta) & \cdots & \frac{\partial^2}{\partial \theta_1 \partial \theta_r} \ell_n(\theta) \\ \vdots & \frac{\partial^2}{\partial \theta_j \partial \theta_k} \ell_n(\theta) & \vdots \\ \frac{\partial^2}{\partial \theta_r \partial \theta_1} \ell_n(\theta) & \cdots & \frac{\partial^2}{\partial \theta_r^2} \ell_n(\theta) \end{pmatrix}$$

- In the MLE setting, $\dot{\ell}_n(\theta)$ is called the score function, and $-\ddot{\ell}_n(\theta)$ is called the information matrix.

- To obtain the maximum likelihood estimator $\widehat{\theta}$, we solve the score equation $\dot{\ell}_n(\theta) = 0$.

Inference

Theorem (MLE theorem)

Under some mild regularity conditions,

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} N(0, \mathcal{I}(\theta)^{-1})$$

as $n \to \infty$, where $\mathcal{I}(\theta)$ is the Fisher information defined as

$$\mathcal{I}(\theta) = E\left\{\dot{\ell}_1(\theta)^{\otimes 2}\right\} = -E\left\{\ddot{\ell}_1(\theta)\right\}.$$

Thus, the covariance matrix of $\widehat{\theta}$ is

$$\text{Cov}(\widehat{\theta}) = \{n\mathcal{I}(\theta)\}^{-1} = \left[-E\{\ddot{\ell}_n(\theta)\}\right]^{-1},$$

which can be estimated by the inverse information matrix $\{-\ddot{\ell}_n(\theta)\}^{-1}$ or $\{-\ddot{\ell}_n(\widehat{\theta})\}^{-1}$ (when θ is unknown in practice).

Inference (Cont.)

Based on the limiting distribution of $\widehat{\theta}$, the $100(1-\alpha)\%$ confidence interval (CI) for θ_j can be constructed by

$$\left[\widehat{\theta}_j - z_{\alpha/2}\sqrt{\mathrm{Var}(\widehat{\theta}_j)},\ \widehat{\theta}_j + z_{\alpha/2}\sqrt{\mathrm{Var}(\widehat{\theta}_j)}\right]$$

$$\Rightarrow \left[\widehat{\theta}_j - z_{\alpha/2}\sqrt{\left\{-\ddot{\ell}_n(\widehat{\theta})\right\}^{-1}_{jj}},\ \widehat{\theta}_j + z_{\alpha/2}\sqrt{\left\{-\ddot{\ell}_n(\widehat{\theta})\right\}^{-1}_{jj}}\right], \quad \text{for } j = 1, \ldots, r,$$

where $z_{\alpha/2}$ is the $(\alpha/2)$th quantile of $N(0,1)$, and $\{-\ddot{\ell}_n(\widehat{\theta})\}^{-1}_{jj}$ denotes the (j,j)th element of the inverse information matrix $\{-\ddot{\ell}_n(\widehat{\theta})\}^{-1}$, which is the variance estimator for $\widehat{\theta}_j$.

Hypothesis Testing

We can test $H_0 : \theta = \theta^*$ using one of the following tests.

- Wald test:

$$W_n = (\widehat{\theta} - \theta^*)^{\mathsf{T}} \{n\mathcal{I}(\theta^*)\}(\widehat{\theta} - \theta^*) \xrightarrow{d} \chi^2_r \quad \text{under } H_0$$

- Score test:

$$SC_n = \dot{\ell}_n(\theta^*)^{\mathsf{T}} \{n\mathcal{I}(\theta^*)\}^{-1} \dot{\ell}_n(\theta^*) \xrightarrow{d} \chi^2_r \quad \text{under } H_0$$

- Likelihood ratio test:

$$LRC_n = 2\{\ell_n(\widehat{\theta}) - \ell_n(\theta^*)\} \xrightarrow{d} \chi^2_r \quad \text{under } H_0$$

Given the significance level α, we reject H_0 if the test statistic is greater than $\chi^2_r(\alpha)$, which is the αth upper quantile of χ^2_r.

Hypothesis Testing (Cont.)

- These three tests are asymptotically equivalent.
- In practice, we may replace $n\mathcal{I}(\theta^*)$ by $-\ddot{\ell}_n(\widehat{\theta})$ or $-\ddot{\ell}_n(\theta^*)$.
- An advantage of the score test is that it does not require the computation of $\widehat{\theta}$ and hence no iterative calculation is necessary.

Table of Contents

Exponential Distribution

Suppose that $T_1, T_2, \ldots, T_n \overset{iid}{\sim} \text{Exp}(\lambda)$. The hazard and survival functions are

$$\lambda(t;\lambda) = \lambda, \qquad S(t;\lambda) = e^{-\lambda t}.$$

We consider the general random censoring scenario. The likelihood function arising from the observed data $(Y_i, \delta_i)_{i=1}^n$ is

$$L_n(\lambda) \propto \prod_{i=1}^n \lambda^{\delta_i} e^{-\lambda Y_i}$$

$$\Rightarrow \ell_n(\lambda) = \log(\lambda) \sum_{i=1}^n \delta_i - \lambda \sum_{i=1}^n Y_i.$$

We differentiate $\ell_n(\lambda)$ with respect to λ to obtain the score and information:

$$\dot{\ell}_n(\lambda) = \frac{1}{\lambda} \sum_{i=1}^n \delta_i - \sum_{i=1}^n Y_i, \qquad -\ddot{\ell}_n(\lambda) = \frac{1}{\lambda^2} \sum_{i=1}^n \delta_i.$$

Solving the score equation yields

$$\widehat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} Y_i}.$$

By the MLE theorem, the asymptotic variance estimator of $\widehat{\lambda}$ is

$$\widehat{\mathrm{Var}}(\widehat{\lambda}) = \{-\ddot{\ell}_n(\lambda)\}^{-1} = \frac{\lambda^2}{\sum_{i=1}^{n} \delta_i},$$

and the asymptotic normal approximation is

$$\widehat{\lambda} \dot{\sim} N(\lambda, \frac{\lambda^2}{\sum_{i=1}^{n} \delta_i}).$$

We replace the unknown λ by $\widehat{\lambda}$ to construct its 95% CI:

$$\left[\frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} Y_i} - 1.96\sqrt{\frac{\sum_{i=1}^{n} \delta_i}{\left(\sum_{i=1}^{n} Y_i\right)^2}}, \ \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} Y_i} + 1.96\sqrt{\frac{\sum_{i=1}^{n} \delta_i}{\left(\sum_{i=1}^{n} Y_i\right)^2}} \right].$$

Alternatively, we can consider a log transformation for λ. Applying the Delta method yields

$$\log \widehat{\lambda} \dot{\sim} N(\log \lambda, \frac{1}{\sum_{i=1}^{n} \delta_i}),$$

where the variance no longer depends on the unknown parameter λ.

Thus, the 95% CI of $\log \lambda$ is given by

$$\left[\log \widehat{\lambda} - \frac{1.96}{\sqrt{\sum_{i=1}^{n} \delta_i}}, \log \widehat{\lambda} + \frac{1.96}{\sqrt{\sum_{i=1}^{n} \delta_i}} \right].$$

Exponentiation on both limits yields an alternative 95% CI for λ:

$$\left[\widehat{\lambda} \exp\left(-\frac{1.96}{\sqrt{\sum_{i=1}^{n} \delta_i}} \right), \widehat{\lambda} \exp\left(\frac{1.96}{\sqrt{\sum_{i=1}^{n} \delta_i}} \right) \right].$$

Empirically, this CI has better coverage as it is guaranteed to be positive.

To test the null hypothesis $H_0 : \lambda = \lambda^*$, we compute the Wald, score, and likelihood ratio test statistics:

- Wald test:

$$W_n = (\widehat{\lambda} - \lambda^*)^{\mathsf{T}} \{-\ddot{\ell}_n(\lambda^*)\}(\widehat{\lambda} - \lambda^*) \xrightarrow{d} \chi^2_1 \quad \text{under } H_0$$

- Score test:

$$SC_n = \dot{\ell}_n(\lambda^*)^{\mathsf{T}} \{-\ddot{\ell}_n(\lambda^*)\}^{-1} \dot{\ell}_n(\lambda^*) \xrightarrow{d} \chi^2_1 \quad \text{under } H_0$$

- Likelihood ratio test:

$$LRC_n = 2\{\ell_n(\widehat{\lambda}) - \ell_n(\lambda^*)\} \xrightarrow{d} \chi^2_1 \quad \text{under } H_0$$

Exact Inference

In some special cases, the non-asymptotic distribution of $\widehat{\lambda}$ is easily available, thus exact inference on λ can be made without applying the MLE theorem.

- Without censoring, $\widehat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} Y_i}$ reduces to $\frac{n}{\sum_{i=1}^{n} T_i}$. The denominator follows a Gamma distribution.
- When $C_1, C_2, \ldots, C_n \overset{iid}{\sim} \text{Exp}(\theta)$, $\sum_{i=1}^{n} \delta_i$ and $\sum_{i=1}^{n} Y_i$ follow the Binomial distribution and Gamma distribution, respectively, and they are statistically independent.

Weibull Regression

Now we relate the covariates X to the failure time T through the Weibull regression model

$$\lambda(t; X) = \lambda p(\lambda t)^{p-1} e^{\beta^\top X},$$
$$S(t; X) = \exp\left\{-(\lambda t)^p e^{\beta^\top X}\right\}.$$

It would be easier to reparameterize $\gamma = \lambda^p$ so that the hazard and survival functions become

$$\lambda(t; X) = \gamma p t^{p-1} e^{\beta^\top X},$$
$$S(t; X) = \exp\left\{-\gamma t^p e^{\beta^\top X}\right\}.$$

The likelihood function arising from the observed data $(Y_i, \delta_i, X_i)_{i=1}^n$ is

$$L_n(\gamma, p, \beta) \propto \prod_{i=1}^n \left\{\gamma p Y_i^{p-1} e^{\beta^\top X_i}\right\}^{\delta_i} \exp\left\{-\gamma Y_i^p e^{\beta^\top X_i}\right\}$$

$$\Rightarrow \ell_n(\gamma, p, \beta) = \sum_{i=1}^n \left[\delta_i \left\{\log(\gamma p) + (p-1)\log(Y_i) + \beta^\top X_i\right\} - \gamma Y_i^p e^{\beta^\top X_i}\right].$$

We then obtain the score function

$$\dot{\ell}_n(\gamma, p, \beta) = \begin{pmatrix} \sum_{i=1}^n \left(\delta_i/\gamma - Y_i^p e^{\beta^\top X_i}\right) \\ \sum_{i=1}^n \left\{\delta_i \left(1/p + \log Y_i\right) - \gamma Y_i^p e^{\beta^\top X_i} \log Y_i\right\} \\ \sum_{i=1}^n \left(\delta_i - \gamma Y_i^p e^{\beta^\top X_i}\right) X_i \end{pmatrix}$$

The score equation has no explicit solutions. We can use the Newton Raphson method to compute the MLE $(\widehat{\gamma}, \widehat{p}, \widehat{\beta})$ instead.

The survival function $S(t; X)$ can be estimated by

$$\widehat{S}(t; X) = \exp\left\{-\widehat{\gamma} t^{\widehat{p}} e^{\widehat{\beta}^\top X}\right\},$$

but the construction of its 95% CI is not that straightforward.

A log-log transformation of $S(t; X)$ can help.

Define the log-log transformation of $S(t;X)$ as

$$B(t;X) = \log\{-\log S(t;X)\} = \log\gamma + p\log t + \beta^{\mathsf{T}}X,$$

which can be estimated by

$$\widehat{B}(t;X) = \log\widehat{\gamma} + \widehat{p}\log t + \widehat{\beta}^{\mathsf{T}}X.$$

By the Delta method, the asymptotic variance estimator of $\widehat{B}(t;X)$ is

$$\widehat{\mathrm{Var}}\{B(t;X)\} = (1/\widehat{\gamma},\ \log t,\ X^{\mathsf{T}})\mathrm{Cov}(\widehat{\gamma},\widehat{p},\widehat{\beta})(1/\widehat{\gamma},\ \log t,\ X^{\mathsf{T}})^{\mathsf{T}}.$$

We first construct the 95% CI for $B(t;X)$, then take exponentiation twice to obtain the 95% CI for $S(t;X)$.

The log-log transformation not only restricts the resulting CI within the meaningful range, but also improves greatly the small-sample performance.

Concluding Remarks

- In parametric models, inference of the failure time distribution is simply based on maximum likelihood estimation and its large-sample theory.
- The computation is relatively easy because the number of unknown parameters is usually small.
- However, parametric models are restrictive as they involve strong distributional assumptions that may not be suitable for a particular dataset. Therefore, the inference procedures may not be robust against model misspecification.
- To overcome this limitation, non- and semi-parametric methods can be considered, although they in general require more complex computation.