

3 Likelihood and Censored (or Truncated) Survival Data

Review of Parametric Likelihood Inference

Suppose we have a random sample (*i.i.d.*) X_1, X_2, \dots, X_n from distribution $f(x; \theta)$ (here $f(x; \theta)$ is either the density function if the random variable X is continuous or probability mass function if X is discrete; θ can be a scalar parameter or a vector of parameters). The distribution $f(x; \theta)$ is totally determined by the parameter θ . For example, if X_i is known from a log-normal distribution, then

$$f(x; \theta) = \frac{1}{\sqrt{2\pi x\sigma}} e^{-(\log x - \mu)^2 / (2\sigma^2)}, \quad (3.1)$$

and $\theta = (\mu, \sigma)$ are the parameters of interest. Any quantity w.r.t. X can be determined by θ . For example, $E(X) = e^{\mu + \sigma^2/2}$. The likelihood function of θ (given data X) is

$$L(\theta; X) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi X_i \sigma}} e^{-(\log X_i - \mu)^2 / (2\sigma^2)} \quad (3.2)$$

$$= (\sqrt{2\pi\sigma})^{-n} \prod_{i=1}^n \frac{e^{-(\log X_i - \mu)^2 / (2\sigma^2)}}{X_i}. \quad (3.3)$$

In general, the likelihood function of θ (given data X) is given by

$$L(\theta; X) = \prod_{i=1}^n f(X_i; \theta) \quad (3.4)$$

and the log-likelihood function θ is

$$\ell(\theta; X) = \log\{L(\theta; X)\} = \sum_{i=1}^n \log\{f(X_i; \theta)\}. \quad (3.5)$$

Note that the (log) likelihood function of θ is viewed more as a function of θ than of data X . We are interested in making inference on θ : estimating θ , constructing confidence interval (region) for θ , and performing hypothesis testing for (part) of θ .

In the likelihood inference for a regression problem, the function $f(\cdot)$ in the above likelihood function is the conditional density of X_i given covariates. For example, suppose X_i is from the following model

$$\log X_i = \beta_0 + z_i \beta_1 + \epsilon_i, \quad (3.6)$$

where β_0 is an intercept and regression coefficient β_1 characterizes the effect of z on X . If we assume $\epsilon_i \sim N(0, \sigma^2)$, then the likelihood of $\theta = (\beta_0, \beta_1, \sigma^2)$ is

$$L(\theta; X) = \prod_{i=1}^n f(X_i | z_i, \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi} X_i \sigma} e^{-(\log X_i - \beta_0 - z_i \beta_1)^2 / (2\sigma^2)}. \quad (3.7)$$

The maximum likelihood estimate (MLE) $\hat{\theta}$ of θ is defined as the maximizer of $\ell(\theta; X)$, which can be usually obtained by solving the following *likelihood equation* (or *score equation*)

$$U(\theta; X) = \frac{\partial \ell(\theta; X)}{\partial \theta} = \sum_{i=1}^n \frac{\partial \log\{f(X_i; \theta)\}}{\partial \theta} = 0,$$

and $U(\theta; X)$ is often referred to as the *score*. Usually $\hat{\theta}$ does not have a closed form, in which case an iterative algorithm such as Newton-Raphson algorithm can be used to find $\hat{\theta}$.

Obviously, the MLE of θ , denoted by $\hat{\theta}$, is a function of data $X = (X_1, X_2, \dots, X_n)$, and hence a statistic that has a sampling distribution. Asymptotically (*i.e.*, for large sample size n), $\hat{\theta}$ will have the following distribution

$$\hat{\theta} \stackrel{a}{\sim} N(\theta, C),$$

where $C = J^{-1}$ or $C = J_0^{-1}$ and

$$J = -E \left[\frac{\partial^2 \ell(\theta; X)}{\partial \theta \partial \theta^T} \right] = - \sum_{i=1}^n E \left[\frac{\partial^2 \log\{f(X_i; \theta)\}}{\partial \theta \partial \theta^T} \right]$$

is often referred to as the *Fisher information matrix* and

$$J_0 = - \left. \frac{\partial^2 \ell(\theta; X)}{\partial \theta \partial \theta^T} \right|_{\theta=\hat{\theta}} = - \sum_{i=1}^n \left. \frac{\partial^2 \log\{f(X_i; \theta)\}}{\partial \theta \partial \theta^T} \right|_{\theta=\hat{\theta}}$$

is often referred to as the *observed information matrix*. Asymptotically J and J_0 are the same. So we usually just use J to mean either information matrix. These results can be used to construct confidence interval (region) for θ .

Suppose $\theta = (\theta_1, \theta_2)$ and we are interested in testing $H_0 : \theta_1 = \theta_{10}$ v.s. $H_A : \theta_1 \neq \theta_{10}$. Under mild conditions, the following test procedures can be used to test H_0 .

Wald test: Suppose the corresponding decompositions of $\hat{\theta}$ and C are

$$\hat{\theta} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}.$$

Then under H_0 ,

$$\chi_{obs}^2 = (\hat{\theta}_1 - \theta_{10})^T C_{11}^{-1} (\hat{\theta}_1 - \theta_{10}) \stackrel{a}{\sim} \chi_k^2,$$

where k is the dimension of θ_1 . Therefore, we reject H_0 if $\chi_{obs}^2 > \chi_{1-\alpha}^2$, where $\chi_{1-\alpha}^2$ is the $(1-\alpha)$ th percentile of χ_k^2 .

Score test: The score test is based on the fact that the score $U(\theta; X)$ has the following asymptotic distribution

$$U(\theta; X) \stackrel{a}{\sim} N(0, J).$$

Decompose $U(\theta; X)$ as $U(\theta; X) = (U_1(\theta; X), U_2(\theta; X))$ and let $\tilde{\theta}_2$ be the MLE of θ_2 under $H_0 : \theta_1 = \theta_{10}$, *i.e.*, $\tilde{\theta}_2$ maximizes $\ell(\theta_{10}, \theta_2; X)$. Then under $H_0 : \theta_1 = \theta_{10}$,

$$\chi_{obs}^2 = U_1^T C_{11} U_1 \sim \chi_k^2,$$

where $U_1(\theta; X)$ and C_{11} are evaluated under H_0 . We reject H_0 if $\chi_{obs}^2 > \chi_{1-\alpha}^2$.

Likelihood ratio test: Under $H_0 : \theta_1 = \theta_{10}$,

$$\chi_{obs}^2 = -2 * (\ell(\theta_{10}, \tilde{\theta}_2; X) - \ell(\hat{\theta}; X)) \sim \chi_k^2.$$

Therefore, we reject H_0 if $\chi_{obs}^2 > \chi_{1-\alpha}^2$.

An example of score tests: Suppose the sample x_1, x_2, \dots, x_n is from a Weibull distribution with survival function $s(x) = e^{-\lambda x^\alpha}$. We want to construct a score test for testing $H_0 : \alpha = 1$, *i.e.*, the data is from an exponential distribution.

The likelihood function of (α, λ) is

$$\begin{aligned} L(\alpha, \lambda; x) &= \prod_{i=1}^n [\lambda \alpha x_i^{\alpha-1} e^{-\lambda x_i^\alpha}] \\ &= \lambda^n \alpha^n e^{-\lambda \sum_{i=1}^n x_i^\alpha + (\alpha-1) \sum_{i=1}^n \log(x_i)}. \end{aligned}$$

Therefore, the log-likelihood function of (α, λ) is

$$\ell(\alpha, \lambda; x) = n \log(\lambda) + n \log(\alpha) - \lambda \sum_{i=1}^n x_i^\alpha + (\alpha - 1) \sum_{i=1}^n \log(x_i).$$

So the components of the score are:

$$\begin{aligned}\frac{\partial \ell(\alpha, \lambda; x)}{\partial \alpha} &= \frac{n}{\alpha} - \lambda \sum_{i=1}^n x_i^\alpha \log(x_i) + \sum_{i=1}^n \log(x_i) \\ \frac{\partial \ell(\alpha, \lambda; x)}{\partial \lambda} &= \frac{n}{\lambda} - \sum_{i=1}^n x_i^\alpha,\end{aligned}$$

and the components of the information matrix is

$$\begin{aligned}\frac{\partial^2 \ell(\alpha, \lambda; x)}{\partial \alpha^2} &= -\frac{n}{\alpha^2} - \lambda \sum_{i=1}^n x_i^\alpha (\log(x_i))^2 \\ \frac{\partial^2 \ell(\alpha, \lambda; x)}{\partial \alpha \partial \lambda} &= -\sum_{i=1}^n x_i^\alpha \log(x_i) \\ \frac{\partial^2 \ell(\alpha, \lambda; x)}{\partial \lambda^2} &= -\frac{n}{\lambda^2}.\end{aligned}$$

For a given data, calculate the above quantities under $H_0 : \alpha = 1$ and construct a score test. For example, for the complete data in home work 2, we have $n = 25$, $\sum x_i = 6940$, $\sum \log(x_i) = 132.24836$, $\sum x_i \log(x_i) = 40870.268$, $\sum x_i (\log(x_i))^2 = 243502.91$, $\hat{\lambda} = 1/\bar{x} = 0.0036023$ so the score $U = 25 - 0.0036023 * 40870.268 + 132.24836 = 10$, the information matrix and its inverse are

$$J = \begin{bmatrix} 902.17186 & 40870.268 \\ 40870.268 & 1926544 \end{bmatrix}, \quad C = J^{-1} = \begin{bmatrix} 0.0284592 & -0.000604 \\ -0.000604 & 0.0000133 \end{bmatrix},$$

so the score statistic is $\chi^2 = 10 * 0.0284592 * 10 = 2.8$ and the p-value is 0.09.

Back to Censored Data

Suppose we have a random sample of individuals of size n from a specific population whose true survival times are T_1, T_2, \dots, T_n . However, due to right censoring such as staggered entry, loss to follow-up, competing risks (death from other causes) or any combination of these, we don't always have the opportunity of observing these survival times. Denote by C the censoring process and by C_1, C_2, \dots, C_n the (potential) censoring times. Thus if a subject is not censored we have observed his/her survival time (in this case, we may not observe the censoring time for this individual), otherwise we have observed his/her censoring time (survival time is larger than

the censoring time). In other words, the observed data are the minimum of the survival time and censoring time for each subject in the sample and the indication whether or not the subject is censored. Statistically, we have observed data (X_i, Δ_i) , $i = 1, 2, \dots, n$, where

$$\begin{aligned} X_i &= \min(T_i, C_i), \\ \Delta_i &= I(T_i \leq C_i) = \begin{cases} 1 & \text{if } T_i \leq C_i \text{ (observed failure)} \\ 0 & \text{if } T_i > C_i \text{ (observed censoring)} \end{cases} \end{aligned}$$

Namely, the potential data are $\{(T_1, C_1), (T_2, C_2), \dots, (T_n, C_n)\}$, but the actual observed data are $\{(X_1, \Delta_1), (X_2, \Delta_2), \dots, (X_n, \Delta_n)\}$.

Of course we are interested in making inference on the random variable T , *i.e.*, any one of following functions

$$f(t) = \text{density function}$$

$$F(t) = \text{distribution function}$$

$$S(t) = \text{survival function}$$

$$\lambda(t) = \text{hazard function}$$

Since we need to work with our data: $\{(X_1, \Delta_1), (X_2, \Delta_2), \dots, (X_n, \Delta_n)\}$, we define the following corresponding functions for the censoring time C :

$$g(t) = \text{density function}$$

$$G(t) = \text{distribution function} = P[C \leq t]$$

$$H(t) = \text{survival function} = P[C \geq t] = 1 - G(t)$$

$$\mu(t) = \text{hazard function} = \frac{g(t)}{H(t)}$$

Usually, the density function $f(t)$ of T may be governed by some parameters θ and $g(t)$ by some other parameters ϕ . In these cases, we are interested in making inference on θ .

In order to derive the density of (X, Δ) , we assume independent censoring, *i.e.*, random

variables T and C are independent. The density function of (X, Δ) is defined as

$$f(x, \delta) = \lim_{h \rightarrow 0} \frac{P[x \leq X < x + h, \Delta = \delta]}{h}, \quad x \geq 0, \delta = \{0, 1\}.$$

Note: Do not mix up the density $f(t)$ of T and $f(x, \delta)$ of (X, Δ) . If we want to be more specific, we will use $f_T(t)$ for T and $f_{X, \Delta}(x, \delta)$ for (X, Δ) . But when there is no ambiguity, we will suppress the subscripts.

1. Case 1: $\delta = 1$, *i.e.*, $T \leq C$, $X = \min(T, C) = T$, we have

$$\begin{aligned} & P[x \leq X < x + h, \Delta = 1] \\ &= P[x \leq T < x + h, C \geq T] \\ &\approx P[x \leq T < x + h, C \geq x] \quad (\text{Note: } x \text{ is a fixed number}) \\ &= P[x \leq T < x + h] * P[C \geq x] \quad (\text{by independence of } T \text{ and } C) \\ &= f(\xi)h * H(x), \quad \xi \in [x, x + h), \quad (\text{Note: } H(x) \text{ is the survival function of } C). \end{aligned}$$

Therefore

$$\begin{aligned} f(x, \delta = 1) &= \lim_{h \rightarrow 0} \frac{P[x \leq X < x + h, \Delta = 1]}{h} \\ &= \lim_{h \rightarrow 0} \frac{f(\xi)h * H(x)}{h} \\ &= f_T(x)H_C(x). \end{aligned}$$

2. Case 2: $\delta = 0$, *i.e.*, $T > C$, $X = \min(T, C) = C$, we have

$$\begin{aligned} & P[x \leq X < x + h, \Delta = 0] \\ &= P[x \leq C < x + h, T > C] \\ &\approx P[x \leq C < x + h, T \geq x] \quad (\text{Note: } x \text{ is a fixed number}) \\ &= P[x \leq C < x + h] * P[T \geq x] \quad (\text{by independence of } T \text{ and } C) \\ &= g_C(\xi)h * S(x), \quad \xi \in [x, x + h). \end{aligned}$$

Therefore

$$\begin{aligned}
 f(x, \delta = 0) &= \lim_{h \rightarrow 0} \frac{P[x \leq X < x + h, \Delta = 0]}{h} \\
 &= \lim_{h \rightarrow 0} \frac{g_C(\xi)h * S(x)}{h} \\
 &= g_C(x)S(x).
 \end{aligned}$$

Combining these two cases, we have the density function of (X, Δ) :

$$\begin{aligned}
 f(x, \delta) &= [f_T(x)H_C(x)]^\delta [g_C(x)S(x)]^{1-\delta} \\
 &= \{[f_T(x)]^\delta [S(x)]^{1-\delta}\} \{[g_C(x)]^{1-\delta} [H_C(x)]^\delta\}.
 \end{aligned}$$

Sometimes it may be useful to use hazard functions. Recalling that the hazard function

$$\lambda_T(x) = \frac{f_T(x)}{S_T(x)}, \quad \text{or} \quad f_T(x) = \lambda_T(x) * S_T(x),$$

we can write $[f_T(x)]^\delta [S(x)]^{1-\delta}$ as

$$[f_T(x)]^\delta * [S(x)]^{1-\delta} = [\lambda_T(x) * S_T(x)]^\delta * [S(x)]^{1-\delta} = [\lambda_T(x)]^\delta * [S(x)].$$

Another useful way of defining the distribution of the random variable (X, Δ) is through the cause-specific hazard function.

Definition: The cause-specific hazard function is defined as

$$\lambda(x, \delta) = \lim_{h \rightarrow 0} \frac{P[x \leq X < x + h, \Delta = \delta | X \geq x]}{h}.$$

For example, $\lambda(x, \delta = 1)$ corresponds to the probability rate of observing a failure at time x given an individual is at risk at time x (*i.e.*, neither failed nor was censored prior to time x).

If T and C are statistically independent, then through the following calculations, we obtain

$$\begin{aligned}
 P[x \leq X < x + h, \Delta = \delta | X \geq x] &= \frac{P[(x \leq X < x + h, \Delta = \delta) \cap (X \geq x)]}{P[X \geq x]} \\
 &= \frac{P[x \leq X < x + h, \Delta = \delta]}{P[X \geq x]}.
 \end{aligned}$$

Hence

$$\begin{aligned}\lambda(x, \delta = 1) &= \frac{\lim_{h \rightarrow 0} \frac{P[x \leq X < x+h, \Delta=1]}{h}}{P[X \geq x]} \\ &= \frac{f(x, \delta = 1)}{P[X \geq x]}.\end{aligned}$$

Since $f(x, \delta = 1) = f_T(x)H_C(x)$ and

$$\begin{aligned}P[X \geq x] &= P[\min(T, C) \geq x] \\ &= P[(T \geq x) \cap (C \geq x)] \\ &= P[T \geq x] * P[C \geq x] \quad (\text{by independence of } T \text{ and } C) \\ &= S_T(x)H_C(x).\end{aligned}$$

Therefore,

$$\begin{aligned}\lambda(x, \delta = 1) &= \frac{f_T(x)H_C(x)}{S_T(x)H_C(x)} \\ &= \frac{f_T(x)}{S_T(x)} = \lambda_T(x).\end{aligned}$$

Remark:

1. This last statement is very important. It says that if T and C are independent then the cause-specific hazard for failing (of the observed data) is the same as the underlying hazard of failing for the variable T we are interested in. This result was used implicitly when constructing the life-table, Kaplan-Meier and Nelson-Aalen estimators.
2. If the cause-specific hazard of failing is equal to the hazard of underlying failure time, the censoring process is said to be *non-informative*. Except for some pathological examples, non-informative censoring is “equivalent to” independent censoring.
3. We assumed independent censoring when we derive the density function for (X, Δ) and the cause-specific hazard. All results depend on this assumption. If this assumption is violated, all the inferential methods will yield biased results.

4. To make matters more complex, we cannot tell whether or not T and C are independent based on the observed data (X_i, Δ_i) , $i = 1, 2, \dots, n$. This is an inherent non-identifiability problem; See Tsiatis (1975) in Proceeding of the National Academy of Science.
5. To complete, if T and C are independent, then

$$\lambda(x, \delta = 0) = \mu_C(x).$$

Now we are in a position to write down the likelihood function for a parametric model given our observed data (x_i, δ_i) (under independence of T and C): $i = 1, 2, \dots, n$.

$$L(\theta, \phi; x, \delta) = \prod_{i=1}^n \{ [f(x_i; \theta)]^{\delta_i} [S(x_i; \theta)]^{1-\delta_i} \} \{ [g(x_i; \phi)]^{1-\delta_i} [H(x_i; \phi)]^{\delta_i} \}.$$

Keep in mind that we are mainly interested in making inference on the parameters θ characterizing the distribution of T . So if θ and ϕ have **no** common parameters, we can use the following likelihood function to make inference on θ :

$$L(\theta; x, \delta) = \prod_{i=1}^n [f(x_i; \theta)]^{\delta_i} [S(x_i; \theta)]^{1-\delta_i}. \quad (3.8)$$

Or equivalently,

$$L(\theta; x, \delta) = \prod_{i=1}^n [\lambda(x_i; \theta)]^{\delta_i} [S(x_i; \theta)]. \quad (3.9)$$

Note: Even if θ and ϕ may have common parameters, we can still use (3.8) or (3.9) to draw **valid** inference on θ . Of course, we may lose some efficiency in this case.

Likelihood for general censoring case

The likelihood function (3.8) has the following form

$$L(\theta; x, \delta) = \prod_{d \in D} f(x_d) \prod_{r \in R} S(x_r), \quad (3.10)$$

where D is the set of death times, R is the set of right censored times. For a death time x_d , $f(x_d)$ is proportional to the probability of observing a death at time x_d . For a right censored

observed x_r , the only thing we know is that the real survival time T_r is greater than x_r . Hence we have $P[T_r > x_r] = S(x_r)$, the probability that the real survival time T_r is greater than x_r , for a right censored observation.

The above likelihood can be generalized to the case where there might be any kind of censoring:

$$L(\theta; x, \delta) = \prod_{d \in D} f(x_d) \prod_{r \in R} S(x_r) \prod_{l \in L} [1 - S(x_l)] \prod_{i \in I} [S(U_i) - S(V_i)], \quad (3.11)$$

where L is the set of left censored observations, I is the set of interval censored observations with the only knowledge that the real survival time T_i is in the interval $[U_i, V_i]$. Note that $S(U_i) - S(V_i) = P[U_i \leq T_i \leq V_i]$ is the probability that the real survival time T_i is in $[U_i, V_i]$.

Likelihood for left truncated observations

Suppose now that the real survival time T_i is left truncated at Y_i . Then we have to consider the conditional distribution of T_i given that $T_i \geq Y_i$:

$$g(t|T_i \geq Y_i) = \frac{f(t)}{P[T_i \geq Y_i]} = \frac{f(t)}{S(Y_i)}. \quad (3.12)$$

Therefore, the probability to observe a death at x_d is proportional to

$$g(x_d|T_d \geq Y_d) = f(x_d)/S(Y_d).$$

The probability that the real survival time T_r is right censored at x_r is

$$P[T_r \geq x_r|T_r \geq Y_r] = S(x_r)/S(Y_r).$$

The probability that the real survival time T_l is left censored at x_l is

$$P[T_l \leq x_l|T_l \geq Y_l] = [S(Y_l) - S(x_l)]/S(Y_l).$$

And the probability that the real survival time T_i is in $[U_i, V_i]$ ($U_i \geq Y_i$) is

$$P(U_i \leq T_i \leq V_i|T_i \geq Y_i) = P(T_i \geq U_i|T_i \geq Y_i) - P(T_i \geq V_i|T_i \geq Y_i) = [S(U_i) - S(V_i)]/S(Y_i).$$

In this case, the likelihood function is given by

$$L(\theta; x, \delta) = \prod_{d \in D} \frac{f(x_d)}{S(Y_d)} \prod_{r \in R} \frac{S(x_r)}{S(Y_r)} \prod_{l \in L} \frac{[S(Y_l) - S(x_l)]}{S(Y_l)} \prod_{i \in I} \frac{[S(U_i) - S(V_i)]}{S(Y_i)} \quad (3.13)$$

$$= \left[\prod_{d \in D} f(x_d) \prod_{r \in R} S(x_r) \prod_{l \in L} (S(Y_l) - S(x_l)) \prod_{i \in I} (S(U_i) - S(V_i)) \right] / \prod_{i=1}^n S(Y_i) \quad (3.14)$$

Likelihood for right truncated observations

We consider the special case of right truncation, that is, only deaths are observed. In this case the probability to observe a death at Y_i conditional on the survival time T_i is less than or equal to Y_i is proportional to

$$f(Y_i)/(1 - S(Y_i)).$$

So the likelihood function is

$$L(\theta; x, \delta) = \prod_{i=1}^n \frac{f(Y_i)}{1 - S(Y_i)}. \quad (3.15)$$

An Example of right censored data: Suppose the underlying survival time T is from an exponential distribution with parameter λ (here the parameter θ is λ) and we have observed data: (x_i, δ_i) , $i = 1, 2, \dots, n$. Since $\lambda(t; \theta) = \lambda$ and $S(t; \theta) = e^{-\lambda t}$, we get the likelihood function of λ :

$$\begin{aligned} L(\lambda; x, \delta) &= \prod_{i=1}^n \lambda^{\delta_i} e^{-\lambda x_i} \\ &= \lambda^{\sum_{i=1}^n \delta_i} * e^{-\lambda \sum_{i=1}^n x_i}. \end{aligned}$$

So the log-likelihood of λ is

$$\ell(\lambda; x, \delta) = \log(\lambda) \sum_{i=1}^n \delta_i - \lambda \sum_{i=1}^n x_i.$$

Obviously, the likelihood equation is

$$U(\lambda; x, \delta) = \frac{d\ell(\lambda; X, \Delta)}{d\lambda} = \frac{\sum_{i=1}^n \delta_i}{\lambda} - \sum_{i=1}^n x_i = 0.$$

So the MLE of λ is given by

$$\begin{aligned}\hat{\lambda} &= \frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n x_i} \\ &= \frac{\# \text{ of failures}}{\text{person time at risk}} \\ &= \frac{D}{PT},\end{aligned}$$

where D is the number of observed deaths and PT is the total patient time. Since

$$\frac{d^2 \ell(\lambda; X, \Delta)}{d\lambda^2} = -\frac{\sum_{i=1}^n \delta_i}{\lambda^2},$$

so the estimated variance for $\hat{\lambda}$ is

$$\begin{aligned}\widehat{\text{Var}}(\hat{\lambda}) &= -\left[\frac{d^2 \ell(\lambda; X, \Delta)}{d\lambda^2} \Big|_{\lambda=\hat{\lambda}} \right]^{-1} \\ &= \frac{\sum_{i=1}^n \delta_i}{[\sum_{i=1}^n x_i]^2} \\ &= \frac{\hat{\lambda}^2}{D},\end{aligned}$$

and asymptotically, we have

$$\hat{\lambda} \stackrel{a}{\sim} N\left(\lambda, \frac{\sum_{i=1}^n \delta_i}{[\sum_{i=1}^n x_i]^2}\right) = N\left(\lambda, \frac{\hat{\lambda}^2}{D}\right).$$

This result can be used to construct confidence interval for λ or perform hypothesis testing on

λ . For example, a $(1 - \alpha)$ confidence interval for λ is given by

$$\hat{\lambda} \pm z_{\alpha/2} * \frac{\hat{\lambda}}{\sqrt{D}}.$$

Note:

1. Sometimes the exponential distribution is parameterized in terms of the mean parameter

$\theta = 1/\lambda$. In this case the MLE of θ is given by

$$\begin{aligned}\hat{\theta} &= \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n \delta_i} \\ &= \frac{\text{total person time at risk}}{\# \text{ of failures}} \\ &= \frac{PT}{D},\end{aligned}$$

and asymptotically,

$$\hat{\theta} \stackrel{a}{\sim} N\left(\theta, \frac{\hat{\theta}^2}{D}\right).$$

(The estimated variance of $\hat{\theta}$ can be obtained by inverting the observed information or using delta-method.)

2. If we ignored censoring and treated the data x_1, x_2, \dots, x_n from the exponential distribution, then the “MLE” of θ would be

$$\tilde{\theta} = \frac{\sum_{i=1}^n x_i}{n},$$

which, depending on the percentage of censoring, would severely underestimate the true mean (note that the sample size n is always larger than D , the number of deaths).

A Data Example: The data below show survival times (in months) of patients with certain disease

$$3, 5, 6^*, 8, 10^*, 11^*, 15, 20^*, 22, 23, 27^*, 29, 32, 35, 40, 26, 28, 33^*, 21, 24^*,$$

where * indicates right censored data. If we fit exponential model to this data set, we have $D = 13$ and $PT = \sum x_i = 418$, so

$$\hat{\lambda} = \frac{D}{PT} = \frac{13}{418} = 0.0311/\text{month},$$

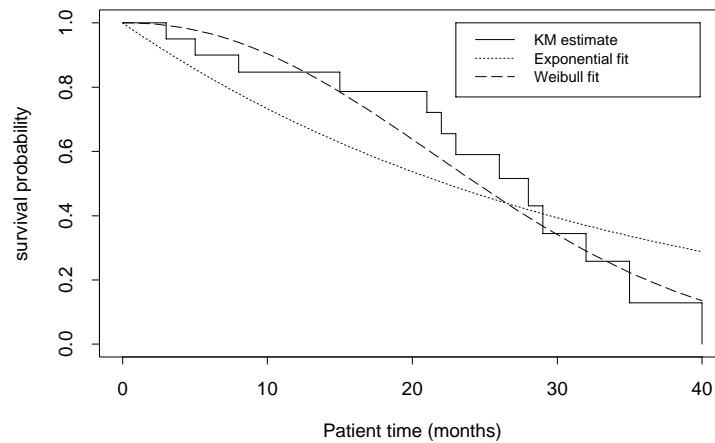
and the estimated standard error of $\hat{\lambda}$ is

$$\text{se}(\hat{\lambda}) = \frac{\hat{\lambda}}{\sqrt{D}} = \frac{0.0311}{\sqrt{13}} = 0.0086,$$

and a 95% confidence interval of λ is

$$\hat{\lambda} \pm z_{0.025} * \text{se}(\hat{\lambda}) = 0.0311 \pm 1.96 * 0.0086 = [0.0142, 0.0480].$$

To see how well the exponential model fits the data, the fitted exponential survival function is superimposed to the Kaplan-Meier estimate as shown in Figure 3.1 using the following R functions:

Figure 3.1: *Three fits to the survival data*

```

> example <- read.table(file="tempsurv.dat", header=T)

> fit <- survfit(Surv(survtime, status), conf.type=c("plain"), example)
> plot(0,0, xlim=c(0,40), ylim=c(0,1),
xlab="Patient time (months)", ylab="survival probability", pch=" ")
> lines(fit, lty=1)
> x <- seq(0,40, by=0.5)
> sx <- exp(-0.0311*x)
> lines(x, sx, lty=2)

```

where the data file `tempsurv.dat` looks like the following

```

survtime status
3 1
5 1
6 0
8 1
10 0
11 0
15 1
20 0
22 1
23 1
27 0
29 1
32 1
35 1
40 1
26 1
28 1
33 0
21 1
24 0

```

Obviously, the exponential distribution is a poor fit. In this case, we can choose one of the following options

1. Choose a more flexible model, such as the Weibull model.
2. Be content with the Kaplan-Meier estimator which makes no assumption regarding the shape of the distribution. In most biomedical applications, the default is to go with the Kaplan-Meier estimator.

To complete, we fit a Weibull model to the data set. Recall that Weibull model has the following survival function

$$S(t) = e^{-\lambda t^\alpha}$$

and the following hazard function

$$\lambda(t) = \alpha \lambda t^{\alpha-1}.$$

So the likelihood function of $\theta = (\lambda, \alpha)$ is given by

$$L(\lambda, \alpha; x, \delta) = \prod_{i=1}^n [\alpha \lambda x_i^{\alpha-1}]^{\delta_i} e^{-\lambda x_i^\alpha}.$$

However, there is no closed form for the MLEs of $\theta = (\lambda, \alpha)$. So we used Proc Lifereg in SAS to fit Weibull model implemented using the following SAS program

```
options ls=80 ps=200;

Data tempurv;
  infile "tempurv.dat" firstobs=2;
  input survtime status;
run;

Proc lifereg data=tempurv;
  model survtime*status(0)= / dist=weibull;
run;
```

The above program produced the following output:

10:41 Friday, January 28, 2005

The LIFEREG Procedure

Model Information

| | |
|--------------------------|---------------|
| Data Set | WORK.TEMPSURV |
| Dependent Variable | Log(survtime) |
| Censoring Variable | status |
| Censoring Value(s) | 0 |
| Number of Observations | 20 |
| Noncensored Values | 13 |
| Right Censored Values | 7 |
| Left Censored Values | 0 |
| Interval Censored Values | 0 |
| Name of Distribution | Weibull |
| Log Likelihood | -16.67769141 |

Algorithm converged.

Analysis of Parameter Estimates

| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|---------------|----|----------|----------------|-----------------------|---------|------------|------------|
| Intercept | 1 | 3.3672 | 0.1291 | 3.1141 | 3.6203 | 679.81 | <.0001 |
| Scale | 1 | 0.4653 | 0.1087 | 0.2943 | 0.7355 | | |
| Weibull Scale | 1 | 28.9964 | 3.7447 | 22.5121 | 37.3483 | | |
| Weibull Shape | 1 | 2.1494 | 0.5023 | 1.3596 | 3.3979 | | |

This SAS program fits a Weibull model with two parameters: intercept β_0 and a scale parameter σ . Two parameters we use λ and α are related to β_0 and σ by (the detail will be discussed in Chapter 5)

$$\lambda = e^{-\beta_0/\sigma}$$

$$\alpha = \frac{1}{\sigma}$$

Since the MLE of β_0 and σ are $\hat{\beta}_0 = 3.36717004$ and $\hat{\sigma} = 0.46525153$, the MLEs $\hat{\lambda}$ and $\hat{\alpha}$ are

$$\hat{\lambda} = e^{-\hat{\beta}_0/\hat{\sigma}} = e^{-3.36717004/0.46525153} = 0.00072,$$

$$\hat{\alpha} = \frac{1}{\hat{\sigma}} = \frac{1}{0.46525153} = 2.149.$$

So $\hat{\alpha}$ is the Weibull Shape parameter in the SAS output. However, SAS uses the parame-

parameterization $S(t) = e^{-(t/\tau)^\alpha}$ for Weibull distribution so that τ is the Weibull scale parameter.

Comparing this to our parameterization, we see that

$$\left(\frac{1}{\tau}\right)^\alpha = \lambda, \implies \tau = \left(\frac{1}{\lambda}\right)^{1/\alpha}.$$

The estimate of this Weibull scale parameter is

$$\hat{\tau} = \left(\frac{1}{0.00072}\right)^{1/2.149} = 28.99.$$

The fitted Weibull survival function was superimposed to the Kaplan-Meier estimator in Figure 3.1 using the the following *R* functions

```
> alpha <- 1/0.46525153
> lambda <- exp(-3.36717004/0.46525153)
> sx <- exp(-lambda * x^alpha)
# the object "x" was created before
> lines(x, sx, lty=4)
> legend(25,1, c("KM estimate", "Exponential fit", "Weibull fit"),
lty=c(1,2,4), cex=0.8)
```

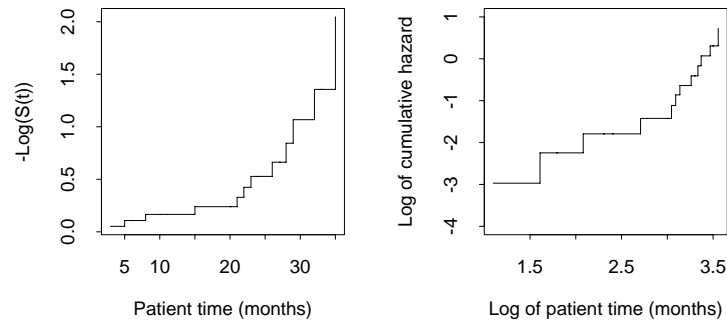
Compared to the exponential fit, the Weibull model fits the data much better (since its estimated survival function tracks the Kaplan-Meier estimator much better than the estimated exponential survival function). In fact, since the exponential model is a special case of the Weibull model (when $\alpha = 1$), we can test $H_0 : \alpha = 1$ using the Weibull fit. Note that $H_0 : \alpha = 1$ is equivalent to $H_0 : \sigma = 1$. Since

$$\left(\frac{\hat{\sigma} - 1}{\text{se}(\hat{\sigma})}\right)^2 = \left(\frac{0.46525153 - 1}{0.108717}\right)^2 = 24.194,$$

and $P[\chi^2 > 24.194] = 0.0000$, we reject $H_0 : \alpha = 1$, *i.e.*, we reject the exponential model. Note also that $\hat{\alpha} = 2.149 > 1$, so the estimated Weibull model has an increasing hazard function.

The inadequacy of the exponential fit is also demonstrated in the first plot of Figure 3.2. If the exponential model were a good fit to the data, we would see a straight line. On the other hand, plot 2 in Figure 3.2 shows the adequacy of the Weibull model, since a straight line of the plot of $\log\{-\log(\hat{s}(t))\}$ vs. $\log(t)$ indicates a Weibull model. Here $\hat{s}(t)$ is the KM estimate.

This graph was plotted using the following *R* codes:

Figure 3.2: *Two empirical plots*

```

postscript(file="fig4.2.ps", horizontal = F,
  height=6, width=8.5, font=3, pointsize=14)
par(mfrow=c(1,2), pty="s")

example <- read.table(file="tempsurv.dat", header=T)

fit <- survfit(Surv(survtime, status), conf.type=c("plain"), example)

plot(fit$time, -log(fit$surv), type="s", xlab=c("Patient time (months)"),
  ylab=c("-Log(S(t))"))

plot(log(fit$time), log(-log(fit$surv)), type="s", ylim=c(-4,1),
  xlab=c("Log of patient time (months)"),
  ylab=c("Log of cumulative hazard"))
dev.off()

```