Asymptotic Likelihood Theory

1 Likelihood function and the maximum likelihood estimator

In general, we have $y_1, y_2, \ldots, y_n$, independent observations, their distribution depending on the parameter $\boldsymbol{\theta}^T = (\theta_1, \theta_2, \ldots, \theta_p)$. Frequently it is the case that $\boldsymbol{\theta}$ is partitioned into two sub-vetors $\boldsymbol{\theta}^T = (\boldsymbol{\theta}_1^T, \theta_2^T)$ where $\boldsymbol{\theta}_1$ is the parameter of interest and $\boldsymbol{\theta}_2$ is a nuisance paramter.

- The likelihood is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta}).$$

  1. Uncensored data:

  $$L_i(\boldsymbol{\theta}) = f(t_i|\boldsymbol{\theta}, x_i)$$

  where $y_i = (t_i, x_i)$

  2. Random censorship data with noninformative censoring:

  $$L_i(\boldsymbol{\theta}) = [\lambda(t_i|\boldsymbol{\theta}, x_i)]^{\delta_i} S(t_i|\boldsymbol{\theta}, x_i)$$

  where $y_i = (t_i, \delta_i, x_i)$

- The score function

$$U(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ln L(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial}{\partial \theta_1} \ln L(\boldsymbol{\theta}) \\ \frac{\partial}{\partial \theta_2} \ln L(\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \ln L(\boldsymbol{\theta}) \end{pmatrix}$$

- Under some regularity conditions

$$E(U(\boldsymbol{\theta})) = \mathbf{0} \qquad \text{and} \qquad \text{Var}(U(\boldsymbol{\theta})) = I(\boldsymbol{\theta})$$

where $I(\boldsymbol{\theta})$ is the information matrix

$$
\begin{aligned}
\mathcal{J}(\boldsymbol{\theta}) = E[U(\boldsymbol{\theta})U^T(\boldsymbol{\theta})] &= -E\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\ln L(\boldsymbol{\theta})\right] \\
&= -E\begin{pmatrix}
\frac{\partial^2}{\partial\theta_1^2}\ln L(\boldsymbol{\theta}) & \frac{\partial^2}{\partial\theta_1\partial\theta_2}\ln L(\boldsymbol{\theta}) & \dots & \frac{\partial^2}{\partial\theta_1\partial\theta_p}\ln L(\boldsymbol{\theta}) \\
\frac{\partial^2}{\partial\theta_2\partial\theta_1}\ln L(\boldsymbol{\theta}) & \frac{\partial^2}{\partial\theta_2^2}\ln L(\boldsymbol{\theta}) & \dots & \frac{\partial^2}{\partial\theta_2\partial\theta_p}\ln L(\boldsymbol{\theta}) \\
\vdots & \vdots & \vdots & \vdots \\
\frac{\partial^2}{\partial\theta_p\partial\theta_1}\ln L(\boldsymbol{\theta}) & \frac{\partial^2}{\partial\theta_2\partial\theta_2}\ln L(\boldsymbol{\theta}) & \dots & \frac{\partial^2}{\partial\theta_p^2}\ln L(\boldsymbol{\theta})
\end{pmatrix}
\end{aligned}
$$

- Likelihood Equations

$$
U(\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}\ln L(\boldsymbol{\theta}) = \begin{pmatrix}
\frac{\partial}{\partial\theta_1}\ln L(\boldsymbol{\theta}) \\
\frac{\partial}{\partial\theta_2}\ln L(\boldsymbol{\theta}) \\
\vdots \\
\frac{\partial}{\partial\theta_p}\ln L(\boldsymbol{\theta})
\end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{0}
$$

- The maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{MLE}$ has the property

$$U(\hat{\boldsymbol{\theta}}_{MLE}) = \mathbf{0}.$$

- Under regularity conditions

$$n^{-1/2}U(\boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \mathcal{J}(\theta))$$

as $n \to \infty$.

2 Hypotheses tests

2.1 The score test

- We can use the result above to test $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $H_a : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

- The test is the score test and is give by

$$RS = U^T(\boldsymbol{\theta}_0)[\mathcal{J}(\theta_0)]^{-1}U(\boldsymbol{\theta}_0)$$

Under H_0,

$$RS \xrightarrow{d} \chi^2_p$$

as $n \to \infty$. We reject H_0 if

$$RS > \chi^2_p(\alpha).$$

- Approximate $100(1-\alpha)$ confidence interval for $\boldsymbol{\theta}$ is

$$\{\boldsymbol{\theta}, U^T(\boldsymbol{\theta})[\mathcal{J}(\boldsymbol{\theta})]^{-1}U(\boldsymbol{\theta}) \leq \chi^2_p(\alpha)\}$$

- To handle the situation with nuisance parameters, partition $\boldsymbol{\theta}^T$ in $(\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)$ where $\boldsymbol{\theta}_1^T = (\theta_1, \theta_2, \ldots, \theta_k)^T$ and $\boldsymbol{\theta}_2 = (\theta_{k+1}, \theta_{k+2}, \ldots, \theta_p)^T$ and similarly
$$[\mathcal{J}(\boldsymbol{\theta})]^{-1} = \begin{pmatrix} \mathcal{J}^{11}(\boldsymbol{\theta}) & \mathcal{J}^{12}(\boldsymbol{\theta}) \\ \mathcal{J}^{21}(\boldsymbol{\theta}) & \mathcal{J}^{22}(\boldsymbol{\theta}) \end{pmatrix}$$
Where $I^{11}(\boldsymbol{\theta})$ is a k by k matrix. Suppose we want to test

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0 \quad (\text{ nothing is hypothesized about } \boldsymbol{\theta}_2)$$

against
$$H_0 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_1^0.$$

- Let $\hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1^0)$ be the mle of $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0$, i.e.

$$\hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1^0) = \arg\max L(\boldsymbol{\theta}_1^0, \boldsymbol{\theta}2)$$

- Define
$$U(\boldsymbol{\theta}_1^0) = \frac{\partial}{\partial \boldsymbol{\theta}} \ln L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)|_{\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2 = \hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1^0)}$$

- The score statistic for testing $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0$ is

$$RS = U^T(\boldsymbol{\theta}_1^0)\mathcal{J}^{11}(\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2 = \hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1^0))U(\boldsymbol{\theta}_1^0)$$

- This test statistic has approximately under H_0 a chi-square distribution with k degrees of freedom. We reject H_0 if

$$RS > \chi^2_k(\alpha)$$

where k is the dimension of $\boldsymbol{\theta}_1$.

2.2 Wald test

- Under regularity conditions

$$\hat{\boldsymbol{\theta}}_{MLE} \overset{Approx}{\sim} N(\boldsymbol{\theta}, \mathcal{J}^{-1}(\boldsymbol{\theta})).$$

- This gives another test called the Wald test for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. The test statistic is

$$W = (\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}_0)^T \mathcal{J}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}_0).$$

- Under $H0$

$$W \xrightarrow{d} \chi^2_p$$

  and we reject H_0 if $W > \chi^2_p(\alpha)$

- Approximate $100(1-\alpha)$ confidence interval for $\boldsymbol{\theta}$ is

$$\{\boldsymbol{\theta}, (\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta})^T \mathcal{J}(\boldsymbol{\theta})((\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}) \leq \chi^2_p(\alpha)\}$$

- To test $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0$, use

$$W = (\hat{\boldsymbol{\theta}}_{1,MLE} - \boldsymbol{\theta}_1^0)^T [\mathcal{J}^{11}(\hat{\boldsymbol{\theta}}_{MLE})]^{-1}(\hat{\boldsymbol{\theta}}_{1,MLE} - \boldsymbol{\theta}_1^0)$$

  and reject H_0 if $W > \chi^2_k(\alpha)$.

- Practical Example ($k = 1$). Suppose we want to test

$$H_0 : \theta_j = \theta_j^0$$

  Here $theta_1 = \theta_j$ and $\boldsymbol{\theta}_2 =$ rest. Let $\imath^{jj}(\boldsymbol{\theta})$ be the (j, j)th element of $[I(\boldsymbol{\theta})]^{-1}$. Then

$$\frac{\hat{\theta}_j - \theta_j^0}{\sqrt{\imath^{jj}(\hat{\boldsymbol{\theta}})}} \overset{Approx}{\sim} N(0,1)$$

- An approximate 95% confidence interval for θ_j is

$$\hat{\theta}_j \pm 1.96\sqrt{\imath^{jj}(\hat{\boldsymbol{\theta}})}$$

2.3 Likelihood Ratio Test (LRT)

- We have always
$$\frac{L(\boldsymbol{\theta}_0)}{L(\hat{\boldsymbol{\theta}})} \leq 1$$

- When $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, the likeliho od ration $\Lambda = \frac{L(\boldsymbol{\theta}_0)}{L(\hat{\boldsymbol{\theta}})}$ is close to 1.

- For this reason, the LRT for $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ rejects H_0 for small value of Λ

- It turns out that under H_0
$$-2 \ln \frac{L(\boldsymbol{\theta}_0)}{L(\hat{\boldsymbol{\theta}})} \xrightarrow{d} \chi_p^2$$

- We reject H_0 if
$$-2 \ln \frac{L(\boldsymbol{\theta}_0)}{L(\hat{\boldsymbol{\theta}})} > \chi_p^2(\alpha).$$

- For testing $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0$, again let $\hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1^0)$ be the mle of $\boldsymbol{\theta}_2$ given that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^0$ the test statistic is
$$-2 \ln \frac{L(\boldsymbol{\theta}_1^0, \hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1^0))}{L(\hat{\boldsymbol{\theta}})}$$
and we reject H_0 if this test statistics is greater than $\chi_k^2(\alpha)$.

2.4 Information Matrix

$I(\theta)$ is called the "Fisher information" or "expected information", But how should one calculate expectation (i.e $-E\left(\frac{\partial^2}{\partial\theta_i\theta_j} \ln L(\boldsymbol{\theta})\right)$.) when there censoring.

In survival analysis, we typically use "observed information"

$$I(\boldsymbol{\theta}) = -\left(\frac{\partial^2}{\partial\theta_1\theta_j} \ln L(\boldsymbol{\theta})\right)_{p\times p}$$

3 Examples

Example 1: Suppose we observe $(t_i, \delta_i), i = 1, 2, \ldots, n$.

$$L(\lambda) = \prod_{i=1}^{n} L_i(\lambda) = \prod_{i=1}^{n} [\lambda e^{-\lambda t_i}]^{\delta_i} [e^{-\lambda t_i}]^{1-\delta_i} = \prod_{i=1}^{n} \lambda^{\delta_i} e^{-\lambda t_i}$$

then

$$\ln L(\lambda) = \sum_{i=1}^{n} \delta_i \ln(\lambda) - \lambda \sum_{i=1}^{n} t_i$$

Then

$$U(\lambda) = \frac{\partial}{\partial \lambda} \ln L(\lambda) = \frac{\sum_{i=1}^{n}}{\lambda} - \sum_{i=1}^{n} t_i = 0 \Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i}$$

The observed Fisher information is

$$I(\lambda) = -\frac{\partial^2}{\partial \lambda^2} \ln L(\lambda) = \frac{\sum_{i=1}^{n} \delta_i}{\lambda^2}$$

and

$$\widehat{\text{Var}}(\hat{\lambda}) = I^{-1}(\lambda)|_{\lambda=\hat{\lambda}} = \frac{\hat{\lambda}}{\sum_{i=1}^{n} \delta_i} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i^2}$$

and

$$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{d} N(0, I^{-1}(\lambda))$$

An approximate 95% confidence interval for λ is

$$\hat{\lambda} \pm 1.96 \sqrt{\widehat{\text{Var}}(\hat{\lambda})}$$

that is

$$\frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i} \pm 1.96 \frac{\sqrt{\sum_{i=1}^{n} \delta_i}}{\sum_{i=1}^{n} t_i}$$

Example 2 : Exponential Regression
Data : $(t_i, \delta_i, x_i), \mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})^T, i = 1, 2, \ldots, n$
Hazard Model : $\lambda(t|x) = \lambda e^{\boldsymbol{\beta}^T \mathbf{x}}$ where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$
Likelihood:

$$
\begin{aligned}
L(\lambda, \boldsymbol{\beta}) &= \prod_{i=1}^{n} [\lambda(t_i|\mathbf{x}_i)]^{\delta_i} S(t_i|\mathbf{x}_i) \\
&= \prod_{i=1}^{n} \lambda^{\delta_i} e^{\delta_i \boldsymbol{\beta}^T \mathbf{x}_i} e^{-\lambda e^{\boldsymbol{\beta}^T \mathbf{x}_i} t_i} \\
&= \lambda^{\sum \delta_i} e^{\boldsymbol{\beta}^T \sum \mathbf{x}_i \delta_i} e^{-\lambda \sum e^{\boldsymbol{\beta}^T \mathbf{x}_i} t_i}
\end{aligned}
$$

log-likelihood

$$
\sum \delta_i \lambda + \boldsymbol{\beta}^T \sum \mathbf{x}_i \delta_i - -\lambda \sum e^{\boldsymbol{\beta}^T \mathbf{x}_i} t_i
$$

Likelihood equations:

$$
\begin{aligned}
\frac{\partial}{\partial \lambda} \ln L &= \frac{\sum \delta_i}{\lambda} - \sum e^{\boldsymbol{\beta}^T \mathbf{x}_i} t_i = 0 \\
\frac{\partial}{\partial \beta_j} \ln L &= \sum \mathbf{x}_{ji} \delta_i - \lambda \sum x_{ji} e^{\boldsymbol{\beta}^T \mathbf{x}_i} t_i = 0, j = 1, 2, \ldots, p.
\end{aligned}
$$

How do we solve the likelihood equations for $\hat{\lambda}$ and $\hat{\boldsymbol{\beta}}$?

Special Case: Two sample problem

$$
p = 1, x_i = \begin{cases} 0, & \text{if i is in group 1} \\ 1, & \text{if i is in group 2} \end{cases}
$$

so the hazard rate for goup 1 is λ and for group 2 the hazard rate is λe^{β}.

Suppose

$$
\begin{aligned}
d_j &= \text{number of individuals who failed in group } j, j = 1, 2 \\
V_j &= \text{total observed time under study in group j, } j, j = 1, 2
\end{aligned}
$$

that is

$$
\begin{aligned}
d_1 &= \sum_{1}^{n} \delta_i (1 - x_i), \quad d_1 = \sum_{1}^{n} \delta_i x_i \quad \text{this implies that} \quad \sum \delta_i = d_1 + d_2 \\
V_1 &= \sum_{1}^{n} t_i (1 - x_i), \quad V_1 = \sum_{1}^{n} t_i x_i
\end{aligned}
$$

Then from above, the likelihood equations are

$$\begin{aligned} \frac{\partial}{\partial\lambda}\ln L &= \frac{d_1+d_2}{\lambda} - V_1 - e^{\beta}V_2 = 0 \quad (1) \\ \frac{\partial}{\partial\beta}\ln L &= d_2 - \lambda e^{\beta}V_2 = 0 \quad (2) \end{aligned}$$

$$(2) \Rightarrow \hat{\lambda} = e^{-\hat{\beta}}\frac{d_2}{V_2}$$

Substitution into (1) gives

$$(d_1+d_2)e^{\hat{\beta}}\frac{d_2}{V_1} - V_1 - e^{\hat{\beta}}V_2 = 0$$

this implies that

$$e^{\hat{\beta}} = \frac{V_1/d_1}{V_2/d_2} \quad \text{and} \quad \hat{\lambda} = d_1/V_1$$

Notice that $e^{\hat{\beta}}$ is the ratio of failure rates.

The information matrix in this case is

$$I(\lambda,\beta) = \begin{pmatrix} \lambda^{-2}\sum_i \delta_i & \sum_i x_i t_i e^{\beta x_i} \\ \sum_i x_i t_i e^{\beta x_i} & \lambda \sum_i x_i^2 t_i e^{\beta x_i} \end{pmatrix}$$

Note that

$$\begin{aligned} \sum_i \delta_i/\hat{\lambda}^2 &= \frac{d_1+d_2}{d_1^2}V_1^2 \\ \sum_i x_i t_i e^{\hat{\beta}x_i} &= e^{\hat{\beta}}V_2 = \frac{V_1 d_2}{d_1 V_2}V_2 \\ \sum_i x_i^2 t_i e^{\beta x_i} &= \sum_i x_i t_i e^{\beta x_i} \end{aligned}$$

This implies that

$$I(\hat{\lambda},\hat{\beta}) = \begin{pmatrix} \frac{d_1+d_2}{d_1^2}V_1^2 & \frac{V_1 d_2}{d_1} \\ \frac{V_1 d_2}{d_1} & d_2 \end{pmatrix}$$

Therefore

$$\widehat{Var}(\hat{\lambda},\hat{\beta}) = I^{-1}(\hat{\lambda},\hat{\beta}) = \begin{pmatrix} \frac{d_1}{V_1^2} & -V_1^{-1} \\ -V_1^{-1} & \frac{d_1+d_2}{d_1 d_2} \end{pmatrix}$$

We will use the following data to illustrate the mle based procedures. the data is time measured in 100 days

$$\begin{array}{lcl} \text{Group 1} & : & 43, 64, 88, 88, 90, 92, 106, 109, 113, 116, 120, 127, 130, 134, 146, 165, 204, 116+, 144+ \\ \text{Group 2} & : & 42, 56, 73, 98, 105, 132, 132, 133, 133, 133, 133, 139, 140, 161, 180, 196, 223, 104+, 244+ \end{array}$$

It is easy to see that $d_1 = 17, V_1 = 2195, d_2 = 19, V_2 = 2923$. This implies that

$$\hat{\lambda} = \frac{17}{2195} = 0.007745, \quad \hat{\beta} = \ln\left(\frac{V_1 d_2}{V_2 d_1}\right) = \ln(0.839) = -0.175.$$

To test $H_0 : \beta = 0$, the Wald test is

$$\frac{\hat{\beta}}{\sqrt{\widehat{Var}(\hat{\beta})}} = \frac{\hat{\beta}}{\sqrt{\frac{d_1+d_2}{d_1 d_2}}} = \ln\left(\frac{V_1 d_2}{V_2 d_1}\right)\sqrt{\frac{d_1 d_2}{d_1 + d_2}}$$

This is equal to -0.524 (not significant)

A 95% confidence interval for β is

$$\hat{\beta} \pm 1.96\sqrt{\frac{d_1 + d_2}{d_1 d_2}} = (-0.830, 0.480)$$