Statistics 255
Derivation of the KM Estimator as a Non-Parametric MLE

Let \( t_1 < t_2 < \cdots < t_D \) represent the observed failure times in a sample of size \( n \) from a population with survival function \( S \). Suppose that \( d_j \) observations fail at time \( t_j \) and \( m_j \) observations are censored in the interval \( [t_j, t_{j+1}) \) at times \( t_{j1}, \ldots, t_{jm_j}, j = 0, 1, \ldots, D \), where \( t_0 = 0 \) and \( t_{D+1} = \infty \). Let

\[
n_j = (m_j + d_j) + \cdots + (m_D + d_D)
\]
denote the number of items at risk just prior to time \( t_j \). The probability of failure at time \( t_j \) is

\[
\Pr[T = t_j] = S(t_j^-) - S(t_j).
\]

Under the assumption of independence between censoring and failure time, the contribution to the likelihood for a censored observation at time \( t_{jl} \) is

\[
\Pr[T > t_{jl}] = S(t_{jl}).
\]

Then the probability of the data is given by

\[
L = \prod_{j=0}^{D} \left\{ [S(t_j^{-}) - S(t_j)]^{d_j} \prod_{l=1}^{m_j} S(t_{jl}) \right\},
\]

which represents a likelihood function on the space of all survivor functions \( S \) (or CDFs \( F = 1 - S \)). Then the non-parametric MLE is the survival function \( \hat{S} \) that maximizes \( L \). In this case, the MLE will be discontinuous at \( t_1 < t_2 < \cdots < t_D \). Further, since \( t_{jl} \geq t_j \), \( S(t_{jl}) \) is maximized by taking \( S(t_{jl}) = S(t_j), j = 1, \ldots, D; l = 1, \ldots, m_j \). From this, we have that the MLE \( \hat{S} \) will be a discrete survival function with hazard components \( \lambda_1, \ldots, \lambda_D \) at \( t_1 < t_2 < \cdots < t_D \). Thus,

\[
\hat{S}(t_j) = \prod_{l=0}^{j-1} (1 - \lambda_l) \quad \text{and} \quad \hat{S}(t_j^{-}) = \prod_{l=0}^{j-1} (1 - \lambda_l),
\]

where \( \lambda_1, \ldots, \lambda_D \) maximize

\[
L(\lambda) = \prod_{j=0}^{D} \left\{ [S(t_j^{-}) - S(t_j)]^{d_j} \prod_{l=1}^{m_j} S(t_{jl}) \right\} = \prod_{j=1}^{D} \left[ \lambda_j^{d_j} \prod_{l=1}^{j-1} (1 - \lambda_l)^{d_l} \prod_{l=1}^{j} (1 - \lambda_l)^{m_j} \right] = \prod_{j=1}^{D} \lambda_j^{d_j} (1 - \lambda_j)^{n_j - d_j}.
\]

To this end,

\[
\ell(\lambda) = \sum_{j=1}^{D} d_j \log \lambda_j + (n_j - d_j) \log(1 - \lambda_j)
\]

\[
\ell(\lambda) = \frac{\partial \ell(\lambda)}{\partial \lambda_j} = \frac{d_j}{\lambda_j} - \frac{n_j - d_j}{1 - \lambda_j},
\]

and solving \( \ell(\lambda) \equiv 0 \) we have \( \hat{\lambda}_j = d_j/n_j \). From this,

\[
\hat{S}(t) = \prod_{j: t_j \leq t} (1 - \hat{\lambda}_j) = \prod_{l=0}^{t} \left( 1 - \frac{d_j}{n_j} \right) \quad \text{(the K-M estimator)}.
\]
Now,

\[ I_{jj}(\lambda) = -E \left[ \frac{\partial U_j(\lambda)}{\partial \lambda_j} \right] = E \left[ \frac{d_j}{\lambda_j^2} + \frac{(n_j - d_j)}{(1 - \lambda_j)^2} \right] \]

\[ I_{jk}(\lambda) = -E \left[ \frac{\partial U_j(\lambda)}{\partial \lambda_k} \right] = 0, \]

and

\[ \hat{I}_{jj}(\lambda) = \left[ \frac{d_j}{\lambda_j^2} + \frac{(n_j - d_j)}{(1 - \lambda_j)^2} \right] = \cdots = \frac{n_j^3}{d_j(n_j - d_j)}. \]

So, from asymptotic likelihood theory we have

\[ \hat{\lambda} \sim N \left( \bar{\lambda}, I^{-1}(\bar{\lambda}) \right), \]

and by the \( \delta \)-method

\[ \log(1 - \hat{\lambda}_j) \sim N \left( \log(1 - \lambda_j), \frac{1}{(1 - \lambda_j)^2} I^{-1}(\bar{\lambda}) \right), \]

where \( \text{Var}[\log(1 - \hat{\lambda}_j)] = \frac{n_j^2}{(1 - \lambda_j)^2 I_{jj}(\bar{\lambda})} = \frac{d_j(n_j - d_j)}{n_j(n_j - d_j)}. \)

From this, \( \log \hat{S}(t) = \sum_{j: t_j \leq t} \log(1 - \hat{\lambda}) \) and

\[ \text{Var}[\log \hat{S}(t)] = \sum_{j: t_j \leq t} \text{Var}[\log(1 - \hat{\lambda})] = \sum_{j: t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}, \]

and again from the \( \delta \)-method

\[ \text{Var}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{j: t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}, \]

yielding Greenwood’s formula.