Statistics 255
Derivation of the KM Estimator and It’s Variance

Derivation of Greenwoods Formula: Let \( t_1 < t_2 < \cdots < t_D \) represent the observed failure times, and let \( F(t_j) \) denote the history, or filtration, of all deaths and censoring occurring up to time \( t_j \) so that
\[
F(t_j) = \{ d_k, w_k; \text{for all } t_k < t_j \}
\]
where \( d_k \) is the total number of deaths and \( w_k \) is the total number of subjects censored at time \( t_k \). Letting \( \hat{\lambda}_j = d_j/n_j \), the K-M estimator is given by
\[
\hat{S}_{KM}(t) = \prod_{j:\ t_j \leq t} \left( 1 - \frac{d_j}{n_j} \right) = \prod_{l=0}^{j-1} \left( 1 - \frac{d_j}{n_j} \right).
\]

Now consider \( E[\hat{\lambda}_j] \), \( \text{Var}[\hat{\lambda}_j] \), and \( \text{Cov}[\hat{\lambda}_j, \hat{\lambda}_k] \). First, note that conditional on \( F(t_j) \), \( d_j \sim \text{Binomial}(\lambda_j, n_j) \). Then,
\[
E[\hat{\lambda}_j] = E\left\{ E[\hat{\lambda}_j | F(t_j)] \right\} = E\{\lambda_j\} = \lambda_j.
\]
In addition,
\[
\text{Var}[\hat{\lambda}_j] = E\left\{ \text{Var}[\hat{\lambda}_j | F(t_j)] \right\} + \text{Var}\left\{ E[\hat{\lambda}_j | F(t_j)] \right\} = E\left\{ \frac{\lambda_j(1 - \lambda_j)}{n_j} \right\} + \text{Var}[\lambda_j] = \lambda_j(1 - \lambda_j)E[n_j^{-1}]
\]
and WLOG assume that \( k < j \) so that
\[
E[\hat{\lambda}_j, \hat{\lambda}_k] = E\left\{ E[\hat{\lambda}_j, \hat{\lambda}_k | F(t_j)] \right\} = E\left\{ \hat{\lambda}_k E[\hat{\lambda}_j | F(t_j)] \right\} = E[\hat{\lambda}_k \lambda_j] = \lambda_j E[\hat{\lambda}_k] = E[\lambda_j] E[\hat{\lambda}_k]
\]
so that
\[
\text{Cov}[\hat{\lambda}_j, \hat{\lambda}_k] = E[\hat{\lambda}_j \hat{\lambda}_k] - E[\hat{\lambda}_j] E[\hat{\lambda}_k] = 0.
\]
Therefore, by the \( \delta \)-method,
\[
\text{Var}[\log(1 - \hat{\lambda}_j)] \approx \frac{1}{(1 - \hat{\lambda}_j)^2} \text{Var}[\hat{\lambda}_j],
\]
which can be estimated by
\[
\hat{\text{Var}}[\log(1 - \hat{\lambda}_j)] = \frac{1}{(1 - \hat{\lambda}_j)^2} \frac{\hat{\lambda}_j(1 - \hat{\lambda}_j)}{n_j} = \frac{n_j^2}{(n_j - d_j)^2} \times \frac{d_j(n_j - d_j)}{n_j} = \frac{d_j}{n_j(n_j - d_j)}.
\]
From this, \( \log \hat{S}_{KM}(t) = \sum_{j: t_j \leq t} \log(1 - \hat{\lambda}) \) and the fact that \( \text{Cov}[\hat{\lambda}_j, \hat{\lambda}_k] = 0 \),

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

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

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

yielding Greenwood’s formula.