CONFIDENCE INTERVALS FOR THE SURVIVOR FUNCTION IN THE COX REGRESSION MODEL

In the Cox model, the survival function is given by $S(t|\mathbf{z}) = \exp(-e^{\boldsymbol{\beta}^T\mathbf{z}}\Lambda_0(t))$. It is estimated by $\hat{S}(t|\mathbf{z}) = \exp(-e^{\hat{\boldsymbol{\beta}}^T\mathbf{z}}\hat{\Lambda}_0(t)))$, where $\hat{\boldsymbol{\beta}}$ is the Cox partial likelihood estimate of $\boldsymbol{\beta}$ and $\hat{\Lambda}_0(t)$ is the corresponding Breslow estimate of the cumulative hazard function:

$$\hat{\Lambda}_0(t) = \sum_{i:X_i\le t} \frac{\delta_i}{\sum_{j=1}^n Y_j(X_i)e^{\hat{\boldsymbol{\beta}}^T\mathbf{z}_j}}. \tag{1}$$

The purpose of these notes is to develop a confidence interval for $S(t|\mathbf{z})$. We do this by developing a confidence interval for $\Lambda(t|\mathbf{z})$ and then transforming this interval into a confidence interval for $S(t|\mathbf{z})$.

In the development below, we will use the symbol $\doteq$ to denote approximate equality. This means that the difference between the two sides of the $\doteq$ symbol is negligible for large n.

We have

$$\Lambda(t|\mathbf{z}) = \Lambda_0(t)e^{\boldsymbol{\beta}^T\mathbf{z}}, \quad \hat{\Lambda}(t|\mathbf{z}) = \hat{\Lambda}_0(t)e^{\hat{\boldsymbol{\beta}}^T\mathbf{z}}.$$

By Taylor expansion we have

$$e^{\hat{\boldsymbol{\beta}}^T\mathbf{z}} \doteq e^{\boldsymbol{\beta}^T\mathbf{z}} + e^{\boldsymbol{\beta}^T\mathbf{z}}\mathbf{z}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

Hence

$$\begin{aligned}
\hat{\Lambda}(t|\mathbf{z}) &= \hat{\Lambda}_0(t)e^{\hat{\boldsymbol{\beta}}^T\mathbf{z}} \\
&\doteq \hat{\Lambda}_0(t)[e^{\boldsymbol{\beta}^T\mathbf{z}} + e^{\boldsymbol{\beta}^T\mathbf{z}}\mathbf{z}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \\
&= \Lambda_0(t)[e^{\boldsymbol{\beta}^T\mathbf{z}} + e^{\boldsymbol{\beta}^T\mathbf{z}}\mathbf{z}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] + [\hat{\Lambda}_0(t) - \Lambda_0(t)][e^{\boldsymbol{\beta}^T\mathbf{z}} + e^{\boldsymbol{\beta}^T\mathbf{z}}\mathbf{z}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \\
&= \Lambda_0(t)e^{\boldsymbol{\beta}^T\mathbf{z}} + \Lambda_0(t)e^{\boldsymbol{\beta}^T\mathbf{z}}\mathbf{z}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + e^{\boldsymbol{\beta}^T\mathbf{z}}[\hat{\Lambda}_0(t) - \Lambda_0(t)] \\
&\quad + e^{\boldsymbol{\beta}^T\mathbf{z}}[\hat{\Lambda}_0(t) - \Lambda_0(t)][\mathbf{z}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \\
&= \Lambda(t|\mathbf{z}) + \Lambda_0(t)e^{\boldsymbol{\beta}^T\mathbf{z}}\mathbf{z}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + e^{\boldsymbol{\beta}^T\mathbf{z}}[\hat{\Lambda}_0(t) - \Lambda_0(t)] \\
&\quad + e^{\boldsymbol{\beta}^T\mathbf{z}}[\hat{\Lambda}_0(t) - \Lambda_0(t)][\mathbf{z}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]. \qquad (2)
\end{aligned}$$

Now, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ and $\hat{\Lambda}_0(t) - \Lambda_0(t)$ are both $O_p(n^{-\frac{1}{2}})$, and thus $[\hat{\Lambda}_0(t) - \Lambda_0(t)][\mathbf{z}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]$ is $O_p(n^{-1})$. It follows that the last term on the right hand side of (2) is negligible in comparison with the preceding two terms.

We thus get

$$\hat{\Lambda}(t|\mathbf{z}) - \Lambda(t|\mathbf{z}) \doteq \Lambda_0(t)e^{\boldsymbol{\beta}^T\mathbf{z}}\mathbf{z}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + e^{\boldsymbol{\beta}^T\mathbf{z}}[\hat{\Lambda}_0(t) - \Lambda_0(t)].$$

Applying a Taylor approximation to (1), we get

$$\hat{\Lambda}_0(t) \doteq \tilde{\Lambda}_0(t) - \mathbf{C}(t)^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

where

$$\tilde{\Lambda}_0(t) = \sum_{i:X_i \le t} \frac{\delta_i}{\sum_{j=1}^n Y_j(X_i)e^{\boldsymbol{\beta}^T\mathbf{Z}_j}}, \quad \mathbf{C}(t,\boldsymbol{\beta}) = \sum_{i:X_i \le t} \frac{\sum_{j=1}^n Y_j(X_i)e^{\boldsymbol{\beta}^T\mathbf{Z}_j}\mathbf{Z}_j}{[\sum_{j=1}^n Y_j(X_i)e^{\boldsymbol{\beta}^T\mathbf{Z}_j}]^2}.$$

Thus,

$$\hat{\Lambda}(t|\mathbf{z}) - \Lambda(t|\mathbf{z}) \doteq e^{\boldsymbol{\beta}^T\mathbf{z}}\left[\sum_{i:X_i \le t} \frac{\delta_i}{\sum_{j=1}^n Y_i(X_i)e^{\boldsymbol{\beta}^T\mathbf{Z}_j}} - \Lambda_0(t)\right] + e^{\boldsymbol{\beta}^T\mathbf{z}}\mathbf{Q}(t,\mathbf{z},\boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

where

$$\mathbf{Q}(t,\mathbf{z},\boldsymbol{\beta}) = \Lambda_0(t)\mathbf{z} + \mathbf{C}(t,\boldsymbol{\beta}).$$

Now, it can be shown that the variance of $\tilde{\Lambda}_0(t)$ can be estimated by

$$\widehat{\mathrm{Var}}(\tilde{\Lambda}_0(t)) = \sum_{i:X_i \le t} \frac{\delta_i}{[\sum_{j=1}^n Y_j(X_i)e^{\hat{\boldsymbol{\beta}}^T\mathbf{Z}_j}]^2}.$$

This variance expression is analogous to the variance expression for the Nelson-Aalen cumulative hazard function estimator for the case of univariate data. In addition, it can be shown that $\tilde{\Lambda}_0(t) - \Lambda_0(t)$ is asymptotically independent of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$. (The proof of this requires advanced methods; see Andersen and Gill (1982, *Ann. Stat.*), page 1104.)

We thus get

$$\widehat{\mathrm{Var}}(\hat{\Lambda}_0(t)) = (e^{\hat{\boldsymbol{\beta}}^T\mathbf{z}})^2\left[\sum_{i:X_i \le t} \frac{\delta_i}{[\sum_{j=1}^n Y_j(X_i)e^{\hat{\boldsymbol{\beta}}^T\mathbf{Z}_j}]^2}\right] + \mathbf{Q}(t,\mathbf{z},\hat{\boldsymbol{\beta}})^T\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})\mathbf{Q}(t,\mathbf{z},\hat{\boldsymbol{\beta}}),$$

with $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})$ obtained in the standard manner.