B.5 Modern Survival Problem sheet 5: Relative risks and diagnostics

To be turned in by 2pm on 27 November, 2015

(1) Let $N(t)$ be a counting process with additive hazards $\lambda_i(t) = \lambda_0(t) + \sum_{k=1}^p x_{ik}(t)\beta_k(t)$, with $B_k(t) = \int_0^t \beta_k(s)ds$. As in Lecture 13 we define $\mathbf{N}(t)$ to be the vector of the individual counting processes (so it is a binary vector), and similarly $\mathbf{X}(t)$ the matrix of covariates, and $\hat{B}(t)$ the vector of regression coefficient estimators. Define the martingale residual

$$\mathbf{M}_{res}(t) = \int_0^t J(s)d\mathbf{N}(s) - \int_0^t J(s)\mathbf{X}(s)d\hat{\mathbf{B}}(s),$$

where $J(s)$ is the indicator of $\mathbf{X}(s)^T\mathbf{X}(s)$ having full rank, hence of $\mathbf{X}^-(s)$ being nonzero.

(a) Using the fact that

$$J(s)\big(\mathbf{I} - \mathbf{X}(s)\mathbf{X}^-(s)\big)\mathbf{X}(s) \equiv \mathbf{0},$$

show that $\mathbf{M}_{res}$ is a martingale. (That is, every component is a martingale.)

(b) Suppose now that all covariates are fixed and the data are right-censored, and let τ be the final time under consideration (such that $J(\tau) = 1$). Show that

$$\mathbf{X}(0)^T\mathbf{M}_{res}(\tau) = 0.$$

(For time-fixed covariates we define $\mathbf{X}(t) := \mathbf{Y}(t)\mathbf{X}$, where $\mathbf{Y}(t)$ is the matrix with the at-risk indicators $Y_i(t)$ on the diagonal.)

(c) How might this fact be used as a model-diagnostic for the additive-hazards assumption?

(2) Let

$$\begin{aligned}
\mathbf{X}_i(t) &= \text{ vector of observed covariates for individual } i \text{ at time } t;\\
N_i(t) &= \text{ counting process for individual } i \text{ at time } t;\\
\hat{\beta} &= \text{ estimate of Cox regression coefficients;}\\
\hat{A}_0(t) &= \text{ estimate of baseline hazard in Cox model;}\\
\hat{M}_i(t) &= N_i(t) - \int_0^t Y_i(s)e^{\hat{\beta}^T\mathbf{X}_i(s)}d\hat{A}_0(s) \text{ the martingale residuals;}\\
\bar{X}_k(t) &= \frac{\sum_{i=1}^n Y_i(t)X_{ik}(t)e^{\hat{\beta}^T\mathbf{X}_i(t)}}{\sum_{i=1}^n Y_i(t)e^{\hat{\beta}^T\mathbf{X}_i(t)}};\\
U_k(t) &= \sum_{i=1}^n \int_0^t \big[X_{ik}(s) - \bar{X}_k(s)\big]d\hat{M}_i(s).
\end{aligned}$$

U_k is called the *score process*.

(a) Show that

$$U_k(t) = \sum_{t_j \le t} \left(X_{i_j k} - \bar{X}_k(t_j) \right).$$

(The summands here are called *Schoenfeld residuals*.)

(b) Show that the score process is the conditional expectation of the partial derivative of the log likelihood with respect to the coefficient β_k, conditioned on $\mathcal{F}_t$.

(c) Conclude that $U_k(0) = U_k(\infty) = 0$.

(d) Explain why a plot of $U_k(t)$, suitably scaled, would be expected to look like a random walk conditioned to start and end at 0 (a *discrete bridge*) if the proportional hazards assumption holds.

(3) *(Based on Exercise 11.1 of [KM03].)* The dataset `larynx` in the package `KMsurv` includes times of death (or censoring by the end of the study) of 90 males diagnosed with cancer of the larynx between 1970 and 1978 at a single hospital. One important covariate is the stage of the cancer, coded as 1,2,3,4.

(a) Why would it probably not be a good idea to fit the Cox model with relative risk $e^{\beta \cdot \text{stage}}$?

(b) Use a martingale residual plot to show that `stage` does not enter as a linear covariate.

(c) An alternative is to define three new binary covariates, coding for the patient being in stage 2, 3, or 4 respectively (leaving stage 1, where all three covariates are 0, as the baseline group). Fit this model. Are all of these covariates statistically significant?

(d) An equivalent approach is to replace `stage` in the model definition by `factor(stage)`. Show that this produces the same result.

(e) Try adding year of diagnosis or age at diagnosis as a linear covariate (in the exponent of the relative risk). Is either statistically significant?

(f) Use a residual plot to test whether one or the other of these covariates might more appropriately enter the model in a different functional form — for example, as a step function.

(g) Use a Cox-Snell residual plot to test whether the Cox model is appropriate to these data.