A.4 Model diagnostics, repeated events

Questions 1,3,5 are to be done for discussion in class. Questions 2,4,6 are to be handed in for marking.

1. In a relative-risk regression model the intensity for individual i is

$$\lambda_i(t) = Y_i(t)\alpha_0(t)r(\beta, \mathbf{x}_i),$$

where β is the vector of parameters, and $\mathbf{x}_i$ is the vector of covariates associated with individual i, and

$$r(\beta, \mathbf{x}_i) = e^{\beta^T \mathbf{x}_i}.$$

Let $L(\beta)$ be the partial likelihood.

   (a) Compute a formula for the k-th component of the score function;

   (b) Compute a formula for the (k, m) component of the observed information matrix.

2. *(Based on Exercise 11.1 of [17].)* 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 $e^{\beta \cdot \text{stage}}$? What should be done instead?

   (b) Explain how you would use a martingale residual plot to show that `stage` does not enter as a linear covariate.

   (c) Which residual plot would you use to test whether the proportional-hazards assumption holds for `age` or `stage`, or whether the proportional effect of one of these covariates changes over time.

   (d) Explain how you would use a Cox–Snell residual plot to test whether the Cox model is appropriate to these data. Describe the calculations you would perform, the plot that you would create, and describe the visual features you would be looking for to evaluate the goodness of fit.

3. Carry out these computations in `R` for the data set described in the previous question:

   (a) One way of making `R` treat the `stage` variable appropriately is to replace it in the model definition by `factor(stage)`. Show that this produces the same result as defining separate binary variables for three different outcomes.

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

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

   (d) 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.

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

4. We observe survival times T_i satisfying an additive hazards model, so the hazard for individual i is $h_i(t) = \beta_0(t) + \sum_{k=1}^{p} x_{ik}(t)\beta_k(t)$, with $B_k(t) = \int_0^t \beta_k(s)ds$. We define $Y_i(t)$ to be the at-risk indicator for individual i at time t, and $\mathbf{X}(t)$ the matrix of covariates at time t multiplied by at-risk, defined as in section 5.12.2. We also define $\mathbf{N}(t)$ to be the binary vector giving in position i the number of events that individual i has had up to time t. Let $\hat{B}(t)$ be the vector of cumulative regression coefficient estimators.

   We assume the process has been observed up to a final time τ where there is a sufficient range of subjects remaining that $\mathbf{X}(t)$ has full rank. Define the martingale residual vector

$$\mathbf{M}_{res}(t) = N(t) - \sum_{t_j \le t} \mathbf{X}(t_j) d\hat{\mathbf{B}}(t_j),$$

   (a) Show that all components of $\mathbf{M}_{res}(t)$ have expectation 0, for all times $0 \le t \le \tau$.

   (b) Suppose now that all covariates are fixed and the data are right-censored. Show that

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

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

5. Suppose we have a right-censored survival data set where we have accidentally copied every line of data twice. We fit a Cox proportional hazards regression model.

   (a) Show that the point estimate for β and for the baseline hazard will be the same as for the correct (undoubled) data, but that the variance estimate will be wrong. (Use the Breslow method for dealing with the tied observations.) What will the variance estimate be, relative to the correct estimate?

   (b) Show that the sandwich estimate described in section 9.3 will agree asymptotically with the variance estimate for the correct data set.

   (c) Carry out the calculations in `R` for the `bmt` (bone marrow transplant) data set in the `KMsurv` package, referred to in section 7.4. That is, fit a Cox proportional hazards model as described in 7.4, and then fit the same model to the data set where every line of the data object has been duplicated. Compare the conclusions. Then add an `id` variable (so that the two lines corresponding to the same patient have the same `id`) and redo the analysis using a `+cluster(id)` term in the formula, and see if the problem is resolved.

6. Suppose n individuals experience events at constant rate λ_i, $i = 1, \ldots, n$, over the same period of time $[0, T]$. The rate λ_i for individual i is unknown. Suppose the unknown rates λ_i have a gamma distribution with parameters (r, λ), and let N_i be the observed number of events for individual i.

   (a) Show that (N_i) have a negative binomial distribution, and compute the parameters.

   (b) Suppose we fit these data to a Poisson model, to obtain an estimate $\hat{\lambda}$. How will $\hat{\lambda}$ behave as $n \to \infty$?

   (c) How would you test the hypothesis that the Poisson model is correct, against the alternative that it is negative binomial?

   (d) **[optional]** Test these conclusions with simulated data in `R`.