Survival Data Analysis (BIOS 7210)
Breheny

Assignment 6
Due: Thursday, October 15

1. In this problem, we will write code to solve for score and likelihood ratio confidence intervals. In particular, we will obtain score and likelihood ratio confidence intervals for the `stage` coefficient in our model from the 10-8 notes. Following the design matrix there, we will refer to this coefficient below as β_3. Please refer to the code at http://myweb.uiowa.edu/pbreheny/7210/f15/notes/10-8.R, which fit the model and worked out the Wald CIs, as a starting point for this problem.

   (a) Write a function called `prof` that takes a value, `b3`, for β_3 and returns $\widehat{\boldsymbol{\beta}}_{-3}(\beta_3)$, the profile likelihood MLEs for the other coefficients. To make parts (b) and (c) easier, I recommend having `prof` return $\boldsymbol{\mu}$, $\mathbf{W}$, and $\ell(\beta_3, \widehat{\boldsymbol{\beta}}_{-3}(\beta_3))$ as well.

   (b) Recall that the multivariate score test is based on

   $$\mathbf{I}^{-1/2}\mathbf{u} \dot{\sim} \mathrm{N}(\mathbf{0}, \mathbf{1}).$$

   In this case, only the third element (the one corresponding to `stage`) is of interest. Note that the third element of the score vector is

   $$\mathbf{x}_3^T(\mathbf{d} - \boldsymbol{\mu})$$

   and $\mathbf{I}$ depends on $\mathbf{W}$; if you took my advice in part (a), `prof` will return both $\boldsymbol{\mu}$ and $\mathbf{W}$, so the score test should be easy to calculate. Construct a 95% confidence interval by finding the range of β_3 values for which the score test is not rejected. For the purposes of this problem, calculate the interval out to three decimal places.

   (c) Construct a 95% confidence interval by finding the range of β_3 values for which the likelihood ratio test is not rejected. Again, if `prof` returns $\ell(\beta_3, \widehat{\boldsymbol{\beta}}_{-3}(\beta_3))$, this will be easier to calculate. As in (b), calculate the interval out to three decimal places.

2. In the 9-24 notes, we stratified the GVHD analysis on age, and noted that there seemed to be an interaction between age and treatment. Let us re-analyze this data using exponential regression.

   (a) Obviously, the failure times do not even remotely resemble an exponential distribution, since they all occur within the first 60 days and the survival function is completely flat past that point. To make the distribution more exponential-like, consider all times past 60 days to be censored (to be clear: do not throw any observations out, and do not touch any subjects with times on study less than 60 days). Make a linear diagnostic plot and comment on whether the modified data look reasonably exponential in distribution.

   (b) Fit an exponential regression model with three covariates: Group, Age (as a continuous numeric quantity; do not categorize Age as we did in the 9-24 example), and the interaction between Group and Age. Carry out a Wald test of the interaction term. Report your result and comment on what it means.

   (c) For the model in (b), estimate the hazard ratio for the `Group` term and provide a 95% confidence interval. Comment on the meaning of these results. *Think very carefully about the meaning here.*

(d) Re-fit the model, only this time, subtract 21 (the median age) from `Age`. Repeat part (c) for this new model. Comment on the meaning of the hazard ratios you obtain, and how and why they differ from (c).

(e) For the model from part (d), what is the λ parameter (i.e., the "baseline" hazard) of the exponential regression model? What are the characteristics of this "baseline" individual?