Survival Data Analysis (BIOS 7210)
Breheny

Assignment 8
Due Thursday, November 2

1. This problem consists of deriving the information matrix for the Weibull AFT model and using it to construct a confidence interval for σ. For parts (e) and (f), you will construct a confidence interval using data from a real AFT model and compare your results to what you get from the `survival` package. For these parts, use the Age by Treatment interaction model that we fit in the last homework assignment (i.e., for the GVHD data censored at 60 days, the model with Age, Group, and the Group by Age interaction).

   (a) Using the book's notation (sort of; actually, the book defines $\mathbf{a}$ to be the negative of how we're defining it here), $\mathbf{a}$ is an $n \times 1$ vector with elements

   $$a_i = \frac{\partial g}{\partial w_i},$$

   and $\mathbf{A}$ is an $n \times n$ matrix with elements

   $$A_{ij} = -\frac{\partial^2 g}{\partial w_i \partial w_j}.$$

   Recall that g was defined in the notes as follows:

   $$g(\mathbf{w}) = \sum_i \{d_i \log \lambda(w_i) + \log S(w_i)\},$$

   where $\lambda(\cdot)$ and $S(\cdot)$ are the hazard and survival function for W. Derive $\mathbf{a}$ and $\mathbf{A}$ for the Weibull AFT model.

   (b) Derive $\partial^2 \ell / \partial \boldsymbol{\beta}^2 |_{\hat{\boldsymbol{\theta}}}$; express your answer in terms of σ, $\mathbf{X}$, and $\mathbf{A}$. Note that for (b)-(d), some terms are equal to zero when evaluated at the MLE.

   (c) Derive $\partial^2 \ell / \partial \boldsymbol{\beta} \partial \sigma |_{\hat{\boldsymbol{\theta}}}$; express your answer in terms of σ, $\mathbf{X}$, $\mathbf{A}$, and $\mathbf{w}$.

   (d) Derive $\partial^2 \ell / \partial \sigma^2 |_{\hat{\boldsymbol{\theta}}}$; express your answer in terms of σ, $\mathbf{A}$, $\mathbf{a}$, and $\mathbf{w}$.

   (e) Using your answers from (b)-(d), calculate a confidence interval for σ in the model described above. Note that you can obtain $\hat{\boldsymbol{\eta}}$ from `fit$linear.predictors` and $\mathbf{X}$ from `model.matrix(fit)`.

   (f) Use your confidence interval from (e) to obtain a confidence interval for the Weibull shape parameter γ. How does it compare to the answer from the `survival` pacakge (i.e., your answer part b of the problem from the last assignment)?

2. Parkinson's disease is a neurodegenerative disorder primarily affecting motor function. The disease is typically first diagnosed after age 50, although it can also occur in younger individuals (the actor Michael J. Fox being a well-known example). In an effort to identify genetic variants that affect the age at onset for Parkinson's disease, researchers at the Wadsworth Center genotyped 431 individuals with Parkinson's disease at a number of genetic location.

The course website contains the results of this genotyping (`parkinsons.txt`) for three single nucleotide polymorphisms (SNPs). A single nucleotide polymorphism is a genetic location at which diversity exists in the human population: for example, most people have a "T" at a given location while others have a "C". The "C" in this example would be referred to as a *minor allele*. The SNP columns in the data set record the number of minor alleles in each person's genome at the given position (humans have two copies of their genome, so this number can be 0, 1, or 2).

Fit a Weibull AFT model to this data; note that due to the sampling design of investigating only individuals with Parkinson's disease, there is no censoring in this example.

(a) Provide a one sentence summary of the effect of the SNP that you consider to be the most important. Think of this as a sentence that would appear in the abstract of an article publishing this finding – i.e., it should be short, clearly understandable to non-statisticians, yet still scientific and accurate.

(b) Now consider the model as a proportional hazards model. Again, write a one sentence description of the "most important" SNP's effect on hazard as it might appear in the abstract of a scientific article.

(c) In your opinion, which parameterization of the model (AFT or PH) provides the most clear interpretation in light of the particular scientific goals of this study?

(d) A more sophisticated analysis of this data might try to account for the fact that the individuals are not equally likely to be sampled. In particular, subjects with earlier onset have greater availability and may be over-represented in this sample. What concept that we have discussed in this course most accurately describes this phenomenon?

3. Take the crude `R` code in http://myweb.uiowa.edu/pbreheny/7210/f17/notes/10-24.R (this is the same as the code we discussed in class) and improve it in two ways: (1) The code should not assume a fixed number of iterations; rather, it should check for convergence after every update. (2) The code should check whether the Newton-Raphson update does, in fact, increase the Cox partial log-likelihood, and if it does not (and only then!), apply step-halving.

   This problem does not require any written component. Rather, e-mail your code to me; your score will be determined by whether your code executes correctly on a data set of my choosing. As in the `R` code referenced above, assume that `X` is the design matrix and `d` is the vector of failure indicators, with `X` and `d` sorted by failure time with no ties present. Your code should not modify `X` or `d` in any way. Also, please do not send me 400 lines of code and require me to find the 20 lines of code that actually implement your algorithm.