Survival Data Analysis (BIOS 7210)
Breheny

Assignment 8
Due: Thursday, October 29

1. The problem re-visits the question of a treatment-age interaction in the GVHD data that we also examined in Assignment 6. As in Assignment 6, censor the times on study at 60 days (for any subject still at risk on day 60). Now, however, fit a Weibull AFT model to the data, again including Age, Group, and the Group by Age interaction.

   (a) Is there significant evidence that the Weibull model provides a superior fit to the data than an exponential model?

   (b) Provide an estimate along with a 95% confidence interval for the Weibull shape parameter γ. Provide some explanation for how you arrived at this interval.

   (c) Consider the question of estimating the treatment effect for an individual of age a. Let $\hat{\boldsymbol{\theta}}$ denote the vector of parameter estimates yielded by the AFT model you fit, with covariance matrix $\mathbb{V}\hat{\boldsymbol{\theta}}$. There is no explicit parameter for the treatment effect at age a. However, the treatment effect can be constructed as a linear combination of the parameters in the model. I.e., there exists a contrast Δ_a such that $\Delta_a^T\boldsymbol{\theta}$ is equal to the treatment effect at age a. What is Δ_a? Also, provide an expression for the Wald 95% confidence interval for $\Delta_a^T\boldsymbol{\theta}$ in terms of Δ_a, $\hat{\boldsymbol{\theta}}$, and $\mathbb{V}\hat{\boldsymbol{\theta}}$.

   (d) Use your result from (c) to construct a plot of the age-specific treatment effect versus age. Include a gray shaded band to represent the confidence intervals. If you have not done this before, a confidence band can be plotted using the `polygon` function using the following code:

   ```
   polygon(c(a, rev(a)), c(lwr, rev(upr)), col='gray85', border=NA)
   ```

   where `lwr` and `upr` are vectors containing the lower and upper bounds of the interval. In one or two sentences, summarize the main implications of the plot as far as what it says about the treatment benefits of MTX+CSP compared to MTX alone.

   (e) Plot the estimated survival functions for four representative patients: (1) a 10 year-old treated with MTX alone (2) a 10 year-old treated with MTX+CSP (3) a 30 year-old treated with MTX alone (4) a 40 year-old treated with MTX+CSP. Provide a single plot with the four survival curves overlaid, along with a legend or some other annotation to indicate which line belongs to which subject.

2. This problem consists of deriving the information matrix for the Weibull AFT model and using it to construct a confidence interval for σ.

   (a) Using the book's notation (sort of),

   $$\frac{\partial \ell}{\partial \mathbf{w}} = \mathbf{a},$$

   where $\mathbf{a}$ is an $n \times 1$ vector, and

   $$-\frac{\partial^2 \ell}{\partial \mathbf{w}^2} = \mathbf{A},$$

   where $\mathbf{A}$ is an $n \times n$ matrix (actually, the book defines $\mathbf{a}$ to be the negative of how we're defining it here). For the Weibull AFT model, give expressions for $\mathbf{a}$ and $\mathbf{A}$.

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

(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 for the regression model in problem 1. 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 your answer from 1(b)?

3. 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?