Survival Data Analysis (BIOS 7210)
Breheny

Final Project
Due: Monday, December 16

For the final project in this class, you will analyze data from the Infant Feeding Practices Study II, investigating factors associated with cessation of breastfeeding in first-time mothers, and write a paper on your conclusions. The data is available on the course website along with a detailed description of the variables and what they mean. The scientific aims of this study can be summarized as:

- Describe the distribution of breastfeeding duration. In particular, what fraction of mothers continue to breastfeed for the recommended duration?
- Investigate demographic factors and pre-pregnancy beliefs that affect breastfeeding duration.
- Investigate the impact of neonatal breastfeeding experiences on breastfeeding duration, and whether these experiences affect certain types of mothers more than others.

Analysis: The goal is to accurately model the time-to-event distribution, with the explanatory variables explaining as much variability as possible. In complex data sets, there are always a variety of interesting phenomena to explore: effects may be nonlinear, there may be interactions, proportional hazards might not hold, there may be strange outliers . . . I encourage you to be creative in your approach to modeling. Do not just blindly follow automatic stepwise procedures for adding and removing variables.

At the same time, be careful about overfitting. If an effect is fairly linear, don't feel compelled to include a spline just because you can. Furthermore, if your model is so complex that you no longer understand it, you should probably simplify it.

Thinking about whether a model makes biological/scientific sense is also important. Unless you know a lot about breastfeeding already, you will probably need to read about some of the existing literature on the subject to know whether the way in which you are modeling them is reasonable.

For the purposes of this project, your analysis **must** include at least one interaction. Even if you feel that there are no important interactions, choose the one that is closest to being important and describe it (feel free to add something like, "the evidence for this interaction is weak, however").

Finally, don't feel compelled to only include significant terms in your model. If a predictor *should* affect survival, but for whatever reason doesn't, that may be interesting to report as well.

Paper: The paper should have the same format and adhere to the same quality standards as a regular scientific article. In particular, that means: do not include R code or output; format results into nice-looking tables; figures need clear, correct labels, etc.

- Introduction (10): Relevant biological/medical background to the problem being studied.

- Methods (30): For our purposes, this should mainly describe the model-building process. How did you arrive at the final model you did, and why did you rule out other models? In particular, if you checked for things like interactions, nonlinear effects, but didn't include them in your model, or if you decided not to include certain predictors, explain this and justify your decisions.

  Think carefully about the organization of this section. I realize that your actual model-building process may have been a long, complicated journey, with many lessons learned along the way – you're not writing your autobiography. You're attempting to demonstrate that a logical, objective, and rational decision-making process would arrive at the model you are presenting.

- Results (50): This should provide summary and descriptive statistics (10) as well as your final model or models (40). Plots and tables are typically very helpful here, although you must also describe in words what those plots and tables illustrate. Your results section should be the largest section, and should probably have subsections. For example, if you have a complicated term like an interaction your model, you might want to devote a subsection to explaining it.

  You are being graded here on the technical correctness of what you say, as well as being able to explain it clearly and simply – note that these two objectives are often at odds with one another. As a word of general advice, it is often a good idea to explicitly separate specific numeric results from verbal explanations. For example, include a table or figure with specific results (hazard ratios, confidence intervals, etc.), but then say in the text, "As shown in Table 2, men are at considerably higher risk of coronary heart disease than women."

- Discussion (10): A brief discussion of issues/limitations with the data and/or your model, along with your primary conclusions.

Altogether, the paper should be somewhere in the neighborhood of 10 pages (roughly, 6-7 pages of text and 3-4 pages of figures/tables), although your report may be somewhat longer or shorter depending on how you format things and how many plots you include.