Survival Data Analysis (BIOS 7210)
Breheny

Final Project
Due: Monday, December 10

For the final project in this class, you will analyze a large, complex, time-to-event data set 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.

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 breast cancer already, you will probably need to read about some of the variables involved in this study to know whether the way in which you are modeling them is reasonable.

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.

Special instructions:

For this particular data set, your final analysis must:

1. Include at least one interaction (or stratification).
2. Pay particular attention to the continuous variables: (a) is their effect linear? (b) Would dichotomizing them make sense, or would it result in a loss of information?

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 – Relevant biological/medical background to the problem being studied
- Methods – 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.

- Results – This should provide summary and descriptive statistics as well as your final model(s). 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.
- Discussion – 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, although your report may be somewhat longer or shorter depending on how you format things and how many plots you include.