STA 4250 Project Guidelines

Last Updated: April 20, 2020

General Comments

- Deadline: Sunday May 17 2020
- Counts as an extra 10% to your grade
- See **To Be Handed In** below for what to hand in

Objective

- You are given a modified data set of `colon`, which removes some covariates (to make your job easier), and uses only a subset of the observations.
- Your job is to produce your *best* Cox PH model for the given data set, to predict survival times for possible future patients that have the same covariates as in your data set.
- You must explore various possible models, test assumptions, and end up with your *best* model. (Note that there technically is no best model, so you are just doing the best job that you can.)
- After you hand in your project, I will subsequently compute a type of out of test model performance metric on the model that you chose, to see how well it did.
- Exceptional prediction jobs may receive additional extra credit, beyond the 10%.

Things To Pay Attention To

- Predictive model performance - we covered AIC in this class. The model with the *lowest* AIC is the best model (on the basis of AIC).
- AIC is only interpretable if your model satisfies model assumptions.
  - Check the Cox PH assumption for your models.
  - Martingale residuals will check if you correctly identified the functional relationship with numeric variables
- Rinse and repeat the above - that is, model fitting is a process that repeats itself as you add/remove predictors and test assumptions.

Grade Breakdown

- (20%) Thoughtful consideration of all predictors.
- (20%) Tested an adequate number of models, correctly added/removed variables (using either p-values and/or AIC)
- (20%) Tested Cox-PH assumptions for models in consideration, and acted appropriately when the model failed these assumptions
- (20%) Examined martingale residuals for continuous covariates, at appropriate times, and acted appropriately
- (20%) Overall performance. (Is the final model a good choice compared to the other models? Does it pass model assumptions?)

To Be Handed In

Note: please do not spend too much time on the writeup. I intend this to be more of a glorified homework problem rather than a long in-depth project.

- Option 1:
  - `.R` file containing *all* code that you used during your entire model selection process (it doesn't have to be neat or readable, except that you must clearly distinguish your final model, as well as the assumption testing of that model)
  - A short writeup (2-3 pages) explaining the data set and a short summary of your model selection process. You should also include
    - the final model (its output from the `R` summary, and its AIC value).
    - Plots showing accurate choices for numerical predictors (e.g. martingale residual plots).
    - Plots and tables showing that your model satisfies assumptions.
- Option 2:
  - `.R` file containing *all* code that you used during your entire model selection process (it doesn't have to be neat or readable, except that you must clearly distinguish your final model, as well as the assumption testing of that model), **OR**, if this was done all in a `.rmd` file, then the output of that file
  - The output (`.pdf` or `.html`) of a `.rmd` file, containing all of the information requested in the writeup for option 1.