Final Project

STAT 131A, Fall 2021

Due: Wednesday, December 15, 2021 at 11:59 PM

1 Dataset

The dataset (*cholangitis.csv*) comes from a randomized, double-blinded, placebo-controlled clinical trial of the immunosuppressive drug D-penicillamine at the Mayo Clinic. The study consisted of patients living with primary biliary cholangitis, a fatal chronic autoimmune disease of unknown cause affecting the liver. The details of the dataset are provided in the document *cholangitis.pdf*. There are 418 observations of 20 variables, both numeric and categorical. The study lasted about 12 years. The goals of the project are to:

1. Fit a linear regression equation to the **number of days** a patient survives from the time of registration.
2. Fit a logistic regression to model the **status** of a patient at the end of the study (you only need to consider alive vs dead, and may ignore the 25 patients who received a liver transplant).

2 Visualization

1. **Importing the data:** Read the data into R. Make sure your categorical variables are factors. (**5 points**).
2. **Basic exploratory data analysis:** Perform exploratory data analysis of the data, using any appropriate tools we have learned. Note any interesting features of the data. (**20 points**).

3 Multivariate Regression

1. **Multivariate regression analysis:** Perform a regression analysis of the response (number of days) on the explanatory variables. Describe here whether you transformed your data or covariates, or excluded any observations, and why. Here you might include diagnostic plots (i.e. for transformations you considered but did not use), but only show those that are necessary for explaining your choices. (**20 points**).
2. **Variable selection:** Perform variable selection to select a suitable model involving a subset of your explanatory variables. You can use either stepwise methods or regression subsets in conjunction with cross validation. (**10 points**).
3. **Regression diagnostics:** Look at diagnostic plots of this final model and comment on whether any of the regression assumptions are obviously violated for this dataset and the final model. (**10 points**).

4 Logistic Regression

Fit a logistic regression model for the survival status of a patient at the end of the study, given all the explanatory variables (remember, you are considering status as binary, ignoring the patients who receive transplants). You may also perform variable selection. Comment on your model, with visualizations, as in the , text. (**15 points**)

5 Format for Submission

You are expected to create a *Rmd* file for this project from scratch. The text from this instructions pdf should not be part of your *Rmd* file. You will turn in only a compiled *pdf* to gradescope. An actual analysis would blend these components together into a single narrative. However, for grading purposes, we have divided the project into the specific tasks described above, and each of the tasks should be addressed in a separate section and appropriately labeled so that you can tell Gradescope which pages correspond to which task.

This project is intentionally more open-ended than the homework, so as to be more reflective of an actual analysis of the data. For each of the specific tasks, provide commentary on what you are doing, provide R code and output, and also provide commentary on what you deduce from the output. The commentary should be just regular text typed in your Rmd file (it does not need a > in front of it like the homework). DO NOT put any commentary in the comments of your R code.

To make your code in the compiled pdf wrap nicely, add the following into a code chunk at the beginning of your Rmd file:

```
knitr::opts_chunk$set(echo = TRUE, tidy.opts=list(width.cutoff=60), tidy=TRUE)
```