1 Veteran Lung Cancer

- The data set *VeteranLungCancer.CSV* contains data from the Veteran's Administration Lung Cancer Trial (Kalbfleisch and Prentice). The data are described below.
  - *Treatment* denotes the type of lung cancer treatment; 1 (standard) and 2 (test drug)
  - *CellType* denotes the type of cell involved; 1 (squamous), 2 (small cell), 3 (adeno), 4 (large)
  - *Survival* is the survival time in days since the treatment
  - *Status* denotes the status of the patient as dead or alive; 1 (dead), 0 (alive)
  - *Karnofsky* is the Karnofsky score
  - *Diag* is the time since diagnosis in months
  - *Age* is the age in years
  - *Therapy* denotes any prior therapy; 0 (none), 10 (yes)
- Learning Objective: be able to manipulate and query data, employ a wide range of classical techniques to answer questions of interest, summarize and present results
- Target Audience: undergraduates, first year graduate students
- Techniques: a wide range of classical inferential methods

(a) For the patients which are still alive, determine if the population mean age is less than 45 years old. Use a hypothesis test, p-value, and $\alpha = .05$.

(b) Use a hypothesis test, p-value, and $\alpha = .05$ to determine if the population mean survival times are different for the different types of cancerous cells. If you determine there are population mean survival times which are different, identify pairs of cell types which have similar population mean survival times and provide a graphical summary.

(c) Suppose that we are interested in predicting patient status (either dead or alive). Use a *glm* and the available predictors to model patient status. Use model selection to find the best subset of predictors for predicting patient status. For your best model, report the cross validated prediction error and a confusion matrix.

(d) Repeat (c) by employing a *gam* and smoothing the continuous predictors to predict patient status. For your best model, report the cross validated prediction error and a confusion matrix.

(e) Repeat (c) and (d) by employing a regression tree to predict patient status. For your best model, report the cross validated prediction error and confusion matrix.

(f) Compare the models you found in (c), (d), and (e). Of these 3 techniques, what is the best model for predicting patient status? Use this model to predict the status of a 30 year old patient diagnosed 15 months ago with a squamous cancerous cell type, a Karnofsky score of 50, prior therapy, and who had the test treatment one year ago.