A major issue with antiretroviral drugs is the mutation of the virus’ genes. Because of its high rate of replication (10⁹ to 10¹⁰ virus per person per day) and error-prone polymerase¹, HIV can easily develop mutations that alter susceptibility to antiretroviral drugs. The emergence of resistance to one or more antiretroviral drugs is one of the more common reasons for therapeutic failure in the treatment of HIV.

In the following paper², a sample of in vitro³ HIV viruses were grown and exposed to a particular antiretroviral therapy. The susceptibility of the virus to treatment and the number of genetic mutations of each virus were recorded.

1. Load the data set hiv.rda and create

   \[
   X = \text{hiv.train$x} \\
   Y = \text{hiv.train$y}
   \]

   What would be \( n \) and \( p \) in this problem? What are the covariates in this problem? What are the observations? What is the response? **Note:** Attempt to answer this question before moving on to the rest of the questions.

2. Consider the design matrix \( X \). It is composed of 0’s and 1’s, with a 1 indicating a mutation in a particular gene. Run

   \[
   \text{table}(X)
   \]

   What results do you get? What does this indicate?

3. The response the log transformed susceptibility of a virus to the considered treatment, with large values indicating the virus is resistant (that is, not susceptible). Run

---

¹ An enzyme that ‘stitches’ back together DNA or RNA after replication
² The entire paper is on the website. Try to see what you can get out of it.
³ Latin for in glass, sometimes known colloquially as a test tube
hist(Y)

What plot did you just create? What does this indicate?

4. We may have (at least) two goals with a data set such as this: inference or prediction. An inferential question would be: can we find some genes whose mutation seems to be most related to viral susceptibility? A prediction question would be: can we make a model that would predict whether this therapy would be efficacious, given a virus with a set of genetic mutations.

(a) Let’s do model selection, which can address either of these goals.
   i. Try to find either the best subset solution for this problem. Discuss any problems or findings you discover. In particular, how many possible models are there?
   ii. Now do forward selection with \texttt{regsubsets}. Report the selected covariates using \texttt{bic} as the criterion.

(b) Now, let’s do ridge regression, which only addresses prediction. Using the package \texttt{glmnet}, find the minimum CV ridge solution and report its CV estimate of the prediction risk.

\textbf{Note:} There is no need to report the \( p \) coefficient estimates from the ridge solution.