BST 764: Applied Statistical Modeling
Breheny

Assignment 1: Penalized Regression
Due: Tuesday, September 20

Mathematical concepts and derivations

1. Consider the application of ridge regression in simple linear regression (*i.e.*, $p = 1$), where $y_i \sim N(x_i\beta, \sigma^2)$. For the purposes of simplification, you may assume that $\mathbf{y}$ has been centered and that $\mathbf{x}$ has been centered and scaled (*i.e.*, $\bar{y} = \bar{x} = 0$ and $\mathbf{x}^T\mathbf{x} = n$). Denote the ridge regression estimate $\tilde{\beta}$ and the ordinary least squares estimator $\hat{\beta}$.

   (a) What is the bias of $\tilde{\beta}$?

   (b) What is the variance of $\tilde{\beta}$?

   (c) What is the derivative of the squared bias with respect to λ at $\lambda = 0$?

   (d) What is the derivative of the variance with respect to λ at $\lambda = 0$?

   (e) Given that, at $\lambda = 0$, ridge regression is equivalent to ordinary least squares, what do (c) and (d) imply about the comparative MSEs of ordinary least squares and ridge regression?

   (f) Construct a plot which overlays the following four lines, with λ on the x axis: (i) the squared bias of $\tilde{\beta}$ (ii) the variance of $\tilde{\beta}$ (iii) the MSE of $\tilde{\beta}$ (iv) the MSE of $\hat{\beta}$. Note that in order to do so, you will have to assume values for β, n, and σ^2. Choose whatever values seem reasonable to you, or make a few different plots to see how the plot changes with different values of the parameters. Note also that $\hat{\beta}$ has nothing to do with λ, so (iv) will be a flat line.

   (g) Briefly, comment on what you see in the plot.

2. Consider the application of the lasso in the same setting as above.

   (a) In terms of β and σ, what is the probability that lasso will select $\mathbf{x}$ (*i.e.*, that $\hat{\beta} \neq 0$)?

   (b) Assume that $\sigma^2/n = 1$ and that $\lambda = 1$. Plot the probability in (a) versus β.

   (c) In terms of σ^2/n, what value must λ have in order to limit the false selection probability to 5%?

Simulation

3. Suppose that $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$, where $\mathbf{X}$ is a 50×15 matrix. Consider three settings:

   - Setting 1: $\beta_1, \ldots, \beta_{15} = 0.5$
   - Setting 2: $\beta_1, \ldots, \beta_5 = 1$, $\beta_6, \ldots, \beta_{15} = 0$
   - Setting 3: $\beta_1 = 5$, $\beta_2, \ldots, \beta_{15} = 0$

For each setting, conduct a simulation with at least 100 replications in which you compare subset selection, ridge regression, and the lasso in terms of the mean squared error in their estimates of $\boldsymbol{\beta}$. You may assume whatever distribution on $\mathbf{X}$ you wish and select λ (or p) in any way that you deem reasonable. Comment on the results of your simulation.

Application

4. The presentation of an acutely ill young infant presents health workers, especially those in developing countries, with a very difficult problem. Serious infections are the main cause of morbidity and mortality in infants under 3 months of age in these countries, and diagnosing the severing of the illness is rather difficult.

   To study this problem, the World Health Organization (WHO) collected data on a number of readily accessible variables such as vital signs, family history, and clinical observations resulting from physical examination. The patients' disease status was later determined based on the course of the disease and various laboratory tests. The goal of the study was to develop a early prediction rule for grading the severity of the disease so that timely treatment could be delivered (and costly but unnecessary treatments avoided).

   The WHO study looked at several acute respiratory illnesses, but we will focus our attention here on pneumonia, which was abbreviated `pnsc` and measured on the following scale:

   1: No disease
   2: Cold/cough
   3: Pneumonia
   4: Severe pneumonia
   5: Life-threatening illness

   The data, as well as descriptions of the variables, are available on the course website. For the purposes of this assignment, there are two objectives: (1) Create a linear model to predict the expected value of the disease severity based on available information, and (2) Provide some insight into the nature of this model – which variables appear to be most important? Do they make sense? Why or why not?