Building a Prediction Model with Bootstrapping

Set Up

- X $n \times p$ matrix of data. n=number of observations (think election sites) and p=number of predictors (county demographics).
- Y $n \times 1$ response, think of Y_i as the percent of voters at site i who voted for Candidate A.

Model Building

Using (X, Y), develop a process for building a model $\hat{f}$. For example, you could use Ridge Regression, KNN, etc.

Prediction

Let X^{real} be a $m \times p$ matrix which represent the "real" data on which you want to make a prediction For example, these could be the demographics at a completely different collection of election sites. Using our model, we get

$$\hat{f}(X^{real}) = \hat{Y}^{real} \quad (m \times 1).$$

For any prediction Y, there is *Summary* function,

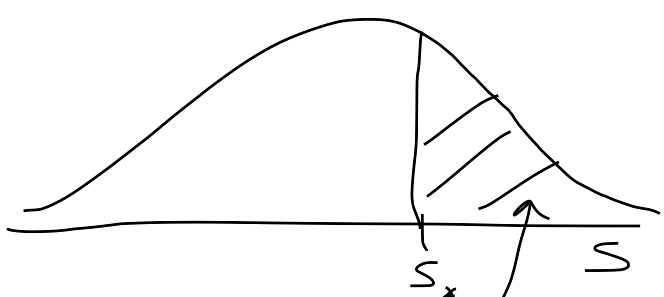
$$\text{SumFunc}(Y).$$

Imagine this is a function that from the predictions Y, computes the overall proportion of candidates that vote for Candidate A. (This involves knowing the total number of voters at each observation site).

Hence, we have

$$\text{SumFunc}(\hat{Y}^{real}) = S.$$

What we really want is the distribution of potential values of S which will represent the variability of the model building process.



Key Question Given a "critical value" of S_*, what is the probability that $S \geq S_*$?, i.e., what is

$$\text{Prob}[S \geq S_*]?$$

Bootstrapping

One way to do this is to reproduce the variability in the data via bootstrapping. Recall that if $z = (z_1, z_2, \ldots, z_n)$ is a sample of size n, a bootstrapped new sample is a set of values $(w_1, w_2, \ldots, w_n)$ where each w_i is a random sample from z. Often we say w is a sample from z of size n taken with replacement.

Bootstrapping in this case involves bootstrapping the observation set $(1, 2, \ldots n)$. Each bootstrap of the observation set results in a "new" data+response set. Bootstrapping B times gives

$$(X_1, Y_1), (X_2, Y_2), \ldots (X_B, Y_B).$$

For each each (X_i, Y_i), we produce a new model $\hat{f}_i$. With each $\hat{f}_i$, we can apply it to the "real" data.

$$\hat{f}_i(X^{real}) = \hat{Y}_i, \quad i = 1, 2 \ldots, B$$

where each $\hat{Y}_i$ is a column of m observation site predictions.

1. Simulating Randomness of the Summary

Put all the columns $\hat{Y}_i$, $i = 1, 2 \ldots, B$ into a $m \times B$ matrix

$$A = [\hat{Y}_1, \hat{Y}_1, \ldots, \hat{Y}_B].$$

Each row of A represents a randomly generated set of values for the i^{th} observation. If these are election sites, then we have an idea of how variable the possible voter proportions are at this site.

We can now use these to generate a truly random prediction of S. To do so, repeat the following process a large number of times.

(1) For each $i = 1, 2, \ldots m$, sample for the i^{th} row of A. Call this number y_i^{sample}

(2) Package all the numbers together in a column vector:

$$\hat{y}^{sample} = \begin{bmatrix} y_i^{sample} \\ y_i^{sample} \\ \vdots \\ y_i^{sample} \end{bmatrix}.$$

(3) Compute

$$\text{SumFunc}(\hat{y}^{sample}) = S^{sample}$$

Repeating steps 1), 2), 3) a large number, K, times produces

$$S_1^{sample}, S_2^{sample}, \ldots, S_K^{sample}.$$

From this, estimate the highly desired Prob$[S \geq S_*]$ with the proportion of these values that are greater than S_*. Formally,

$$\frac{|\{i : S_i^{sample} \geq S_*\}|}{K}.$$