Lecture 22: Bootstrap intro

March 9, 2005

We'll start with an introduction to bootstrapping in a simple setting, and then discuss using it for regression.

Your friend in R through all of this is the package called "boot". Your first step is to load this package by typing "library(boot)".

First some review:

- **sample** is randomly selected from a larger population
- **statistic** is a function of the observations in the sample, for example a mean, median, correlation, standard deviation, slope
- **sampling distribution** is the probability distribution for a statistic based on a random sample. The sampling distribution depends on the "parent" distribution, that is, the distribution the data were sampled from, and the Statistic itself.
- **Central Limit Theorem** says that if the sample is selected independently, and the statistic is a linear combination of the observations, then the sampling distribution is approximately normal, and this approximation improves as the sample size increases.
- **bias** is the amount that a statistic varies, on average, from it's true value.
- **standard error** is the standard deviation of the sampling distribution. It tells us how much variability we will see in our statistic in repeated samples.

The above is most famously applied when calculating confidence intervals for the mean. Assume that there is some variable, say the heights of adult males in the US, that is normally distributed. The mean of this population is unknown and represented by the symbol μ. X represents a random height drawn from this population. X_i represents the ith such sample, and we assume that each sample is independent of the others. (This means that either sampling is done with replacement or we assume the population is so large relative to our sample size that it is as if the sampling were done with replacement.)

Suppose we sample n individuals. We will have observations X_1, $X_2 \ldots X_n$, each an independent, random variable taken from the same normally distributed population. We will use $\bar{X}$ to estimate μ. If the variance of this population is represented by the symbol σ^2, then the following can be shown to be true:

- $\mathrm{E}(\bar{X}) = \mu$, which means tht $\bar{X}$ is an unbiased estimate of μ.
- $\mathrm{Var}(\bar{X}) = \sigma^2/n$ This is the standard error of $\bar{X}$
- The sampling distribution is normal, with the above mean and standard deviation

These are put together to help us find confidence intervals. Since we know that the distribution is normal , we know tht a 95% confidence interval will be the estimate plus or minus 1.96 standard errors.

But what do we do if the parent distribution is not normal? The CLT says that if the statistic is a linear combination, we can still assume (with some loss of accuracy) that the sampling distribution is normal. But what if the statistic is not a linear combination of observations? Then we either have some tough mathematical calculations ahead of us (which will certainly have strong asumptions about the parent distribution) or we will be stuck. And wht if we simply don't know what the parent distribution is?

Historically, Statistics is based on a limited number of statistics because we were limited to sampling distributions that could be derived from mathematical theory. But bootstrapping gives a way of going beyond the math. Let's take the sample of a median. Suppose we have collected a random, independent sample from the population and choose to estimate the population median using the sample median. To find a confidence interval, we need to understand how this sample median will vary from sample to sample.

Here's how we do it.

1. Take a random sample of size n **with** replacement from our original sample.
2. Calculate the median of this random sample. Save it. We call this a bootstrap estimate.
3. Repeat the above steps about 1000 times.
4. The standard deviation of the 1000 sampled medians will estimate the standard error.

To calculate a 95% CI, we simply find the bootstrap estimate that has 2.5% of our bootstraps below it and the one that has 2.5% of our bootstrap estimates above it.

See the notes for an example.

This is called the percentile method for confidence intervals. It is simple, and in many contexts works better than using normal theory for small sample sizes. It is not perfect, however, and there have been many improvements on this method. But it is good enough for our purposes.