PLSC 503: "Multivariate Analysis for Political Research"

Bootstrapping
February 2, 2017

Bootstrapping Variance Estimates

I want to focus on bootstrap variance estimates because they are a general and tremendously useful practical tool for getting variance estimates for estimates of quantities we care about. One could and should teach an entire course on resampling methods, including not only the bootstrap but related approaches like the jackknife, permutation-based tests, and so forth. We won't do all that today.

Instead, we'll just quickly introduce the bootstrap. It is owed originally to Brad Efron, and was developed as a way of getting around an unfortunate fact: *With one sample, you have one estimate of any parameter* $\hat{\theta}$. This fact makes knowing the variability of $\hat{\theta}$ impossible. The germ of Efron's idea was to sample *from the sample itself*, with replacement; this "bootstrap sample" could then be used to get an estimate of the parameter $\hat{\theta}$. Do this enough times, and you could just "see" what the sampling distribution of θ was.

The intiution is thus that:

The population is to the sample as the sample is to the bootstrap sample.[1]

The process for getting bootstrapped quantities is pretty straightforward. Starting with N observations in your original data:

1. Draw a sample of size N **with replacement** from the original data; this is one *bootstrap sample*.

2. Estimate the parameter(s) of interest; denote this *bootstrap estimate* $\underset{k\times 1}{\tilde{\theta}}$.

3. Repeat steps 1 and 2 R times, where R is a large number (at least in the neighborhood of 1000; $R = 999$ is often used, but see below). This will give you R vectors of estimates $\tilde{\theta}_r$, $r \in \{1, 2, ...R\}$, comprised of elements $\tilde{\theta}_{rk}$ (the rth bootstrap estimate of the kth element of θ).

4. Examine the empirical characteristics of the resulting distribution(s) of the $\tilde{\theta}_{rk}$s. In particular:

   - the quantity $\frac{1}{R}\sum_{r=1}^{R}\tilde{\theta}_{irk}$ is a bootstrap estimate of the (expected value of) θ_k,
   - the "empirical" standard deviation of the $\tilde{\theta}_{rk}$s is an estimate of the standard error of the bootstrap parameter estimate, and

[1] I didn't make this line up. Fox and Weisberg did.

- the $(R+1)/40$th- and $R-(R+1)/40$th-largest values of $\tilde{\theta}_{rk}$ (so, for example, for $R = 999$, the 25th- and 975th-largest values of $\tilde{\theta}_{rk}$) are an empirical two-tailed 95% confidence interval for the bootstrap estimates of θ_k

What's So Great About The Bootstrap?

Several things, actually.

1. **It's intuitive.** It's often easier to explain where a confidence interval comes from using a bootstrap example than a conventional analytical one.

2. **It's simple**, even when the statistics are complicated. Imagine that, for some bizarre reason, you were interested in something like $\frac{\beta_3/\beta_7}{2\ln\beta_2}$. Finding the analytical standard error for this would be hard, but the bootstrap is easy (since it's just the empirical standard deviation of that thing in the R bootstrap samples).

3. **It's robust.** It works in lots of different circumstances, and gives "good" answers (i.e., standard errors that are neither too small or too large, confidence intervals with accurate coverage probabilities, etc.) in almost any context.

Other Practicalities

1. It is not uncommon to see researchers substitute medians for means when calculating bootstrap estimates. In practice, this rarely makes much difference in the estimates one gets.

2. Bootstrapping can be slow, even in 2016. It's common practice to use $R = 1000$ for initial runs, to cut down on processing time. But "in real life," you probably want at least $R = 10000$, if not $R = 50000$ or more.

3. One other important caution: Bootstrap confidence intervals that rely on the normal approximation need to be "smooth" and symmetrical. This is one reason why it's always a good idea to examine graphically the distribution of your bootstrapped estimates $\hat{\theta}_r$.

4. Bootstrapping is most often used (in the social sciences) to get measures of variability for things (like medians) where such measures are otherwise hard to get. But they can also be valuable for *bias correction.* For estimators that are not unbiased but consistent, the bootstrap estimate of θ is often "better" than the typical estimate based on the original data; this is especially true when the bias is due to a lack of asymptotics (e.g., when fitting models via maximum likelihood on small samples).

5. This is just scratching the surface of what we can do with resampling methods (and I'll likely expand this section of the course the next time I teach it). Weisberg lists some good references for additional reading.

6. There are a bunch of bootstrap packages in R. A good one is the `Boot` function in the `car` package by John Fox; Fox and Weisberg discuss it here.

An Example

There's a simulated-data example in the slides, which we'll review. You'll also be asked to do some bootstrapping in homework Exercise Two (I think).