607
Bootstrap, introduction

Jon Faust

http://e105.org/e607

October 13, 2016

▶ Readings

- See the Hansen text chapter on the bootstrap
- See the Wright notes, p. 7–11.
- Efron' 'Bootstrap methods: another look at the jackknife' is a classic

  go

  https://projecteuclid.org/download/pdf%5F1/euclid.aos/1176344552

- Peter Hall's book, 'The bootstrap and Edgeworth expansions' is the best introduction in my view.
- As a somewhat deeper version, I like Davidson and Hinkley, Bootstrap Methods and their Application.
- Many citations are included on the bootstrap section of course website.

▶ Our main problem

- We always start our statistical work with a complete model $Y \sim P_\theta, \theta \in \Theta$, with implied distribution and density F_θ and f_θ, respectively.
- Often times our problem comes down to understanding the distribution of some statistic, $\phi(Y)$, in some relevant sample size T, where that distribution is the one induced by P_{θ^*}—that is, the distribution under the true parameter.
- To be a bit more specific, remember that P, F and f all have T subscripts that we usually omit.

  more completely we write $P_{\theta,T}$, $F_{\theta,T}$ and so forth.

- Call the distribution of ϕ induced by $F_{\theta,T}$, $F^{\phi}_{\theta,T}$.
- The difficulty in macro comes from the fact that under assumptions we find plausible θ is generally very high dimensional

  (that is we often can do no better than to take it to be infinite dimensional).
- And, $F^{\phi}_{\theta,T}$ varies in unknown ways with θ

▶ **The conventional asyptotic solution**

▶ **The conventional asymptotic solution**

- In conventional asymptotics, we pick a ϕ that is asymptotically pivotal

  (that is, the distribution doesn't depend on θ under some hypothesis at least)
- Calling the limiting distribution F^{ϕ}_{∞}, we have that

$$F^{\phi}_{\theta,T}(c) \to F^{\phi}_{\infty}(c)$$

  . (for all c that are points of continuity)
- We then take F^{ϕ}_{∞} to be a good approximation to $F^{\phi}_{\theta,T}$ for all θ and for the T in our application.
- The nutshell version in words is approximate the finite sample distribution of the statistic of interest using the limiting large-sample distribution of the statistic.

  The language '*the* large sample distribution' indicates that there is only one: that is the limiting distribution is the same for all θ—i.e., asymptotically pivotal.

▶ **The bootstrap alternative**

▶ **The bootstrap**

- Instead of approximating the finite sample distribution of the statistic, we approximate the entire finite sample distribution of the data.

  That is we approximate $F_{\theta,T}$ instead of $F^{\phi}_{\theta,T}$
- Then we approximate the distribution of our statistic by the exact finite sample distribution of the statistic under the chosen sample distribution.
- Using our notation, we say we pick a θ_b and take $P_{\theta_b,T}$ to be our approximation to the sample distribution.
- And then we use $F^{\phi}_{\theta_b,T}$ as our approximation to the unknown $F^{\phi}_{\theta^*}$

▶ **A summary in words**

- Conventional Asymptotics: Approximate the distribution of the statistic in a sample of size T by its limiting large sample distribution.
- Bootstrap: Pick some approximation to the distribution of the sample, approximate the distribution of our statistic by the exact finite sample distribution of the statistic under the chosen sample distribution.

    Approximate $F^{\phi}_{\theta^*,T}$ by $F^{\phi}_{\theta_b,T}$ for some cleverly chosen θ_b.

▶ **Key issues:**

- 1. What does it mean to cleverly pick θ_b?

    Once you get the hang of it, this is conceptually trivial. You pick some distribution for the data that mimics important aspects of the sample at hand.

- 2. What are the theoretical foundations for why this may be better than conventional asymptotics?

    This is very deep, and even our best answer is not entirely satisfying. That is, the bootstrap seems to work even better than our justfication would indicate.

- 3. How specifically do we use this to, e.g., evaluate test size, or bias or standard error of estimators, or to form confidence intervals?

    This is not deep, but there are some counterintuitive bits that folks sometimes screw up.

▶ **Remainder of this intro lecture: give a couple of concrete examples**

▶ **Simplest case**

- Suppose our statistic of interest is an estimator of some parameter.

    In the first example it will be an estimator of the sample mean, in the second, an OLS $\hat{\beta}$.

- And what we'd like to know is the first two moments of the estimator.

    In the first example, we wnat to know the bias, in the second the variance of $\hat{\beta}$.

- In both cases, the recipe is trivial:
- 1. Pick a θ_b (that is a distribution for the sample) of a type that we know how to generate new samples according to P_{θ_b}
- 2. Run a Monte Carlo generating a zillion samples of the size available in the actual application.
- 3. Calculate the parameter of interest on each of these samples.
- 4. Use the sample moment (e.g., sample mean or sample variance) across these zillion replicates as our estimate of the unknown moment in the application.

- Above in the definition of the bootstrap principle, I said we use the exact moment under θ_b as our estimate. But here, I am using a Monte Carlo to assess the exact moment.

  True. I perhaps should have said 'the exact moment (up to any numerical approximation error).' But let's be clear, everything we do in applied work has approximation error. Thus, whenever I say 'exact' about something in practice, I mean 'exact up to numerical approximaiton error.' For example, all the work you do in Matlab is done in finite precicsion math. Thus, the answer you get for 1 divided by 3 in matlab is exact (up to numerical approximation error).

- This point is (almost) substantive. Some folks think that the bootstrap is inherently a Monte Carlo technique. This is not correct: we almost always find it convenient to use Monte Carlo as a numerical approximation technique in implementing the bootstrap. The underlying principle says, use the exact distribution under θ_b.

▶ Bias adjustment

▶ Bias adjustment

- $Y \sim P_\theta, \theta \in \Theta$. $Y = (y_1, \ldots, y_N)'$ is $N \times 1$; y_i is iid. Truth is θ^*.
- Call the mean of y under θ μ_θ:

$$E_\theta y_i = \mu_\theta \text{ for all } i$$

- We are estimating μ using some statistic $\phi(Y)$.
- We want to know if ϕ is biased and if we can estimate the bias, we want to form a bias adjusted estimate.
- By definition, the 'true' bias is

$$Bias \equiv E_{\theta^*,T}\phi(Y) - \mu_{\theta^*}$$

- Bootstrap: Pick a θ_b, and use the bias under θ_b as our estimate of the bias under θ^*.
- In our notation:

$$\hat{Bias}_b = E_{\theta_b,T}\phi(Y) - \mu_{\theta_b}$$

  Note we take the expected value under θ_b minus the true or population mean under θ_b.

- If we wanted to bias adjust our original estimator, we might then compute

$$\tilde{\phi}(Y) = \phi(Y) - \hat{Bias}_b$$

- This new estimate is not unbiased because we estimated the bias adjustment instead of using the 'true' (and unknowable) bias.

  But if our estimated bias is good, the bias of the adjusted estimator may be smaller than that of the original.

- We'll formalize 'may be smaller' later.

▶ **How should I think about θ_b?**

- There are many equivalent ways of saying what we do when we pick θ_b:
  - Choose a bootstrap distribution for the data
  - Often this amounts to doing what we call choosing a resampling scheme.

▶ **Aside:: Review what an EDF is**

- Remember for any vector of numbers Z $(N \times 1)$, the empirical distribution functoin (EDF)

$$EDF(c) = \text{share of } z_i \text{s¡c}$$

  or in mathier notation:

$$EDF(c) = \frac{\#(z < c)}{N}$$

  where # means cardinality of the items satisfying the condition.
- Suppose we wanted to draw data such the the underlying population is described by the EDF.
- The probability function that gives rise to data with this distribution function puts mass of $1/N$ on each value in the sample.
- Put another way, if we want to draw from a random variable with distribution function equal to the EDF, simply write each z_i on a piece of paper and throw them in an urn and draw with replacement from the urn.

▶ **Back to the bias example**

- The y_is in our bias example were assumed to be iid.
- Thus, one natural bootstrap distribution to draw from is the one implied by the EDF.
- So our P_{θ_b} is the EDF of the sample at hand.
- We know that as the sample gets large, the EDF will approach F_{θ^*}—the true distribution fuction.

  Thus, this choice intuitively has some appeal, at least if our sample is large.

- Our bias estimate is:

$$\hat{Bias}_b = E_{\theta_b,T}\phi(Y) - \mu_{\theta_b}$$

- Of course, under the EDF, μ_{θ_b} is simply the sample mean of the sample at hand:

$$\mu_{\theta_b} = (1/N)\sum y_i$$

Be sure you understand this. The true mean under θ_b is the sample mean of the orginal sample.

- But how to we compute the expected value of ϕ under θ_b?
- Simplest approach: draw a zillion samples of size T consistent with the EDF, compute $\phi(Y)$ on each and take the sample mean of these.
- Call this $\bar{\phi}_b$ and use it as our numerical approximation to $E_{\theta_b,T}\phi(Y)$
- Now our bias estimate is

$$Bias_b = \bar{\phi}_b - \mu_{\theta_b}$$

where μ_{θ_b} is the mean on the orignal sample.

▶ **Comment 1: This is crazy easy**

- Why didn't we do this in 1930 or 1950?
- Folks have played with this idea for years.

And some folks did some related things under the name, e.g., jackknife.

- But we didn't have a coherent theory that would justify the bootstrap in general until around 1980

And then it took another decade or two to fill out the idea and have it have a big effect on applied work

▶ **Comment 2: A practical problem in these discussions**

- This approach inherently involves multiple versions of very similar items:

True population mean, population mean under the bootstrap distribution, and sample analogs of each. And in our example, the sample mean in the actual data is also the population mean under the bootstrap distribution.

- An in more involved cases, we may have even a greater multiplicity of similar items masquerading as one another.

Notationally and conceptually this can get very confusing.

- Peter Hall uses Russian Matryoshka dolls (those famous nested dolls) to explain the bootstrap.
- I don't find this very helpful except as a cautionary note: All those dolls look a like and it may be easy to confuse one for another if you are not careful.
- Thus, if you find yourself lost occasionally, congratulations: you are just like the rest of us.

▶ **Example 2: White standard errors again**

▶ **White standard errors**

- In this case, our parameter of interest is $\hat{\beta}$ in an OLS regression and instead of the bias, we are interested in estimating the standard error of $\hat{\beta}$.
- In the lecture on 'HAC in relevant sample sizes' we reported an experiment in which t-tests were based on White standard errors.
- And with skewed Xs in the regression, the White standard errors performed poorly and that some modified versions perform well in this case, but others don't.

  perform well or badly in the sense that when used to create t statistics, the the implied tests had exact size that was closer to or further form from the nominal size.
- Now we'll add a bootstrap alternative.

▶ The recipe

- Pick a θ_b
- Generate a zillion samples, compute $\hat{\beta}$ on each, and then take the sample standard error over these zillion replicates as our estimator of the standard eror of $\hat{\beta}$ on the sample at hand.
- Just as in the bias: We take the sample moment from a Monte Carlo under a bootstrap distribution as our estimate if the unknown population moment in the actual sample.g

▶ What θ_b?

- To pick a θ_b we need to be specific about the maintained model.
- Let's take the one from our earlier Monte Carlo.

▶ Review: the White SE Monte Carlo

- DGP: $y_t = \beta^* x_t + \varepsilon_t$.
- Assume $\varepsilon \sim iidN(0,1)$s
- The xs are iid but skewed

$$x_t = N(0, 0.01) + z_t$$

  where z_t is 1 with pr. π, 0 else and indep. of all else.

▶ Picking θ_b

- We can put X and Y in a matrix: $W = [X : Y]$.
- Note that the rows of W are iid: that is, the joint distribution of every (x_t, y_t) pair is iid.
- Thus, we can draw from the bivariate EDF by sampling from the rows of W with replacement.
- In a sufficiently large sample, any heteroskedasticity in θ^* will be reflected in this θ_b.
- This approach is sometimes called the XY bootstrap.

▶ **How does it work**

- Here is one cell from our White SE Monte Carlo extended to include t statistics based on this estimator of the standard error:
- Rej. freq., conventional vs. White t-stat.

T	conv.	White	HC2	HC3	boot
30	5.9	24.5	12.3	5.6	5.5

MC replications: 10,000; share of xs shifted: 0.01

- Note that this almost entirely trivial approach (resample from the EDF and take the standard error of the resulting repicates) works really well in this case.
- Somebody had to prove a justification for this, but it then doesn't require us to think much about the nature of heteroskedasticity, etc.
- This is almost like magic.
- And we will often refer to techniques that sound like magic when first proposed.

  GMM: hey we can easily estimate anything.

- But we soon discover that htere is little magic in life: the techniques perform very badly in standard macro cases.
- The bootstrap turns out to be something close to magic: it is seldom worse than the conventional alternatives and very often better.

▶ **Aside:: Hausman-Palmer**

- In the HAC lecture, I quoted a recent paper by Hausman and Porter about the various versions of White standard errors.

  Heterskedasticity-Robust Inference in Finite Samples go

  http://economics.mit.edu/files/7422

- They also consider the bootstrap described here (and some alternative bootstraps)
- They find that even this simplest bootstrap can be badly behaved and propose a different refinement of the simplest bootstraps.

▶ **Aside::**

- Abstract:

  Since the advent of heteroskedasticity-robust standard errors, several papers have proposed adjustments to the original White formulation. We replicate earlier findings that each of these adjusted estimators performs quite poorly in finite samples. We propose a class of alternative heteroskedasticity-robust tests of linear hypotheses based on an Edgeworth expansions of the test statistic distribution. Our preferred test outperforms existing methods in both size and power for low, moderate, and severe levels of heteroskedasticity.

► **Aside::**

- This paper is a nice entry point into this literature.
- Their preferred alternative shows the kind of advances that we may move toward
- But like much of the bootstrap literature: many of the fanciest refinements rely on the independence assumption and have no natural analog in time series.
- Thus, when you read about 'fancier' techniques (e.g., the ABC bootstrap—accelerated, bias corrected) don't get too interested until you see if there is any plausible way to use the ideas in time series.

  Often there is either no natural way to use the idea or the idea doesn't help much in time series. This doesn't say that you should ignore this stuff, but does say you should be skeptical of its value until you discover otherwise.

► **Bottom line**

- Pick a θ_b (or equivalently, a rule for generating new samples)
- Then use the exact finite sample property of your statistic under θ_b in the sample size at hand as your estimate of the unknown finite sample property under the θ^*.

► **What's left**

- 1. Ways to pick θ_b, especially in the time series case.
- 2. How to use the bootstrap to do other things besides what we've done in this lecture

  e.g., evaluate nominal p-values, or compute confidence intervals with chosen nominal coverage.

- 3. The theoretical foundations of all this.

  What are the foundations of this magic?