The Bootstrap

Ethan Ligon

April 25, 2022

Bootstrap

Issue with Monte Carlo is that we have to construct a model to build estimates. This will often require us to assume more than we wish to about the Real World DGP.

Alternative

Use the RWDGP! We begin by observing a sample of N observations X_j once; say D_N. If these are independent (they're identically distributed by construction) we just need to figure out how to repeat this draw.

Sampling

Since D_N is comprised of N iid observations we can use this sample to construct an empirical distribution function of X, say $\hat{F}$. Then think of simply drawing samples from this empirical distribution.

Non-parametric estimator of empirical distribution function

$$\hat{F}(x) = \frac{1}{N}\sum_j \mathbb{1}(X_j \leq x)$$

Simplification

Since the probability of drawing a particular X from $\hat{F}$ is proportional to the frequency with which X appears in D_N, there's an trivial simplification: instead of constructing $\hat{F}$ just:

1. Draw X_j from D_N.
2. Repeat until you have the sample size you want; often (usually?) this will be N, the size of the original sample. Call the resulting "bootstrap" sample D_N^1.

Basic Bootstrap estimation

Suppose we want to estimate a vector of parameters β. We can construct an estimate of this using the original sample, say b_N. But we may not know much about the distribution of this estimator.

Procedure

1. Choose some positive tolerance ϵ.
2. Having drawn a bootstrap sample D_N^1, use it to estimate b_N^1.
3. Draw a new sample D_N^2, and compute b_N^2
4. ...Repeat 30 times...
5. Calculate the sample covariance matrix of the estimates of β,

$$\hat{V}_N^{30} = \frac{1}{30}\sum_m (b_N^m - \bar{b}_N)(b_N^m - \bar{b}_N)^\top$$

6. Repeat: compute additional bootstrap samples until

$$\|\hat{V}_N^M - \hat{V}_N^{M-1}\| < \epsilon$$

Use

We've just described the construction of a covariance matrix for the estimator b_N via the bootstrap, so this can be used for testing and inference in the usual way. But note that the "usual way" assumes that the distribution of b_N is normal.

Non-normal distributions

In finite samples our distributions may be decidedly *non*-normal. But we have an estimate of the distribution! Just construct the empirical distribution of the M bootstrapped estimates of β.

- Tests of normality available
- Simple construction of confidence intervals

When Sample isn't Simple Random

Or, what's an observation? What *is* selected randomly?

Panel data

We often work with longitudinal *panels* comprising, say, N households observed over T periods.

Stratified samples

Suppose we're interested in the effects of an experimental intervention on both men & women. It may make sense to *stratify* the sample so that we're powered to detect effects for both sexes.

Clustered samples

Surveys of households are often *clustered* geographically, with randomization conducted in two stages: (i) geographical locations (clusters) are randomly selected; then (ii) households who live within a cluster are randomly sampled.

Bootstrapping when a sample isn't simple random

The basic idea is for your bootstrap samples to mimic the randomness used to construct the original sample. So:

Panel data

Resample *households* and their entire histories, not household-periods.

Stratified samples

Think of each strata as it's own random sample, and resample within each strata.

Clustered samples

Resample in two stages: (i) clusters (with replacement); then (ii) households within clusters.

Latent variables

Suppose there are some sets $\{L_i\}$ that an randomly selected observation may belong to (e.g., male and female), and we think membership in these sets is important for determining some outcome.
Then we might have, e.g.,

$$y_j = \sum_i \alpha_i \mathbb{1}(j \in L_i) + \beta^\top X_j + u_j$$

Here α_i is interpreted as something like the mean of y conditional on being in the set L_i.
Suppose the sample is simple random. How should you construct a bootstrap estimator?

Residual Bootstrap

One solution is to hold fixed observables X. Then:

1. Use full dataset to estimate, e.g.,
$$y = X\beta + u,$$
obtaining some estimate $b^{(1)}$ of β.
2. Construct residuals
$$e^{(1)} = y - Xb^{(1)}.$$
3. Now, instead of resampling (y, X) just resample the residuals $e^{(1)}$ obtaining $\tilde{e}^{(1)}$, and construct
$$y^{(1)} = Xb^{(1)} + \tilde{e}^{(1)}$$
4. Re-estimate
$$y^{(1)} = X\beta + \tilde{u},$$
obtaining an estimate $b^{(2)}$.
5. Repeat until convergence.

Wild Bootstrap

The residual bootstrap relies on the disturbances being homoskedastic. But what if $\mathbb{E}(u^2|X)$ is a function of X?

Wild Bootstrap

One idea: generate an auxiliary random variable π_j which takes values $\{-1, 1\}$ with equal probability. Then modify the residual bootstrap algorithm:

1. Use full dataset to estimate, e.g.,
$$y = X\beta + u,$$
obtaining some estimate $b^{(1)}$ of β.
2. Construct residuals
$$e^{(1)} = y - Xb^{(1)}.$$
3. Now, instead of resampling (y, X) or e, hold (X, e) fixed and just draw realizations π_j, $j = 1, \ldots, N$, and construct
$$y_n = X\hat{\beta} + \pi_n e$$
4. Re-estimate
$$y_n = Xb_n + u_n$$
5. Repeat until convergence.