Uncertainty and Resampling Methods

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

1. Uncertainty
2. Resampling methods
3. The Non Parametric Bootstrap
4. The Wild Bootstrap
5. The "XY" Bootstrap
   - Example: Elasticity of Demand for Gasoline
6. The Bag of Little Bootstraps

Agenda

1 Uncertainty

2 Resampling methods

3 The Non Parametric Bootstrap

4 The Wild Bootstrap

5 The "XY" Bootstrap
- Example: Elasticity of Demand for Gasoline

6 The Bag of Little Bootstraps

Motivation

- The real world is messy.
- Recognizing this mess will differentiate a sophisticated and useful analysis from one that is hopelessly naive.
- This is especially true for highly complicated models, where it becomes tempting to confuse signal with noise.
- The ability to deal with this mess and noise is the most important skill you need.

Parameter Precision

Variance Sample Mean

- Suppose we have $y_1, y_2, \ldots, y_n$ iid $Y \sim F(\mu, \sigma^2)$ (both finite)
- We want to estimate

$$Var(\bar{Y}) \tag{1}$$

Parameter Precision

Variance in Linear Regression

- Suppose we have $y_i = \beta X_i + u_i \; i = \{1, \ldots, n\} \; E(u_i|X_i) = 0 \; V(u_i|X_i) = \sigma^2$
- We want to estimate $Var(\hat{\beta})$

Parameter Precision

Variance in Nonlinear Inference

- Suppose we have estimated a cost function of the quadratic form,

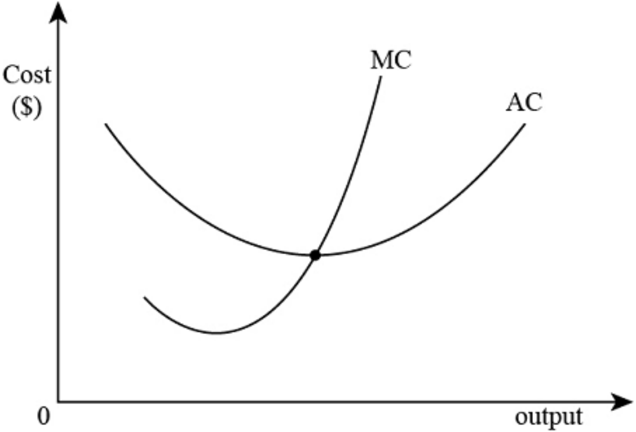
$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \mathbf{z}_i^\top \beta + u_i$$

- where
    - y_i is log cost of firm i,
    - $x_i = \log(q_i)$ is log output, and
    - $\mathbf{z}_i$ is a vector of other characteristics of the ith firm.
- I want to minimize average cost.

Parameter Precision

Variance in Nonlinear Inference

I want to minimize average cost.



Agenda

1. Uncertainty
2. Resampling methods
3. The Non Parametric Bootstrap
4. The Wild Bootstrap
5. The "XY" Bootstrap
   - Example: Elasticity of Demand for Gasoline
6. The Bag of Little Bootstraps

What are resampling methods?

- Tools that involves repeatedly drawing samples and refitting a model of interest on each sample in order to obtain more information about the fitted model
  - Parameter Assessment: estimate standard errors
  - Model Assessment: finding the best model

The Bootstrap

- Sometimes the analytical expression of the variance can be quite complicated.
- In these cases bootstrap can be useful
- In German the expression *an den eigenen Haaren aus dem Sumpf zu ziehen* nicely captures the idea of the bootstrap – *"to pull yourself out of the swamp by your own hair."*

Agenda

1 Uncertainty

2 Resampling methods

3 The Non Parametric Bootstrap

4 The Wild Bootstrap

5 The "XY" Bootstrap
- Example: Elasticity of Demand for Gasoline

6 The Bag of Little Bootstraps

Non Parametric Bootstrap

- Suppose we have $y_1, y_2, \ldots, y_n$ iid $Y \sim F(\mu, \sigma^2)$ (both finite)
- We want to estimate

$$Var(\bar{Y}) \tag{2}$$

- Alternative way (no formula!)
  1. From the n original data points $y_1, y_2, \ldots, y_n$ take a sample *with replacement* of size n
  2. Calculate the sample average of this *"pseudo-sample"* (Bootstrap sample)
  3. Repeat this B times.
  4. Compute the variance

The Non Parametric Bootstrap

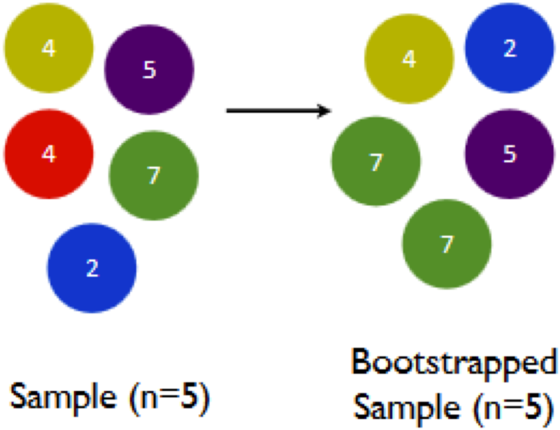
Two key properties

- Two key properties of bootstrapping that make this seemingly crazy idea actually work.

  1. Each bootstrap sample must be of the same size (n) as the original sample
  2. Each bootstrap sample must be taken with replacement from the original sample.

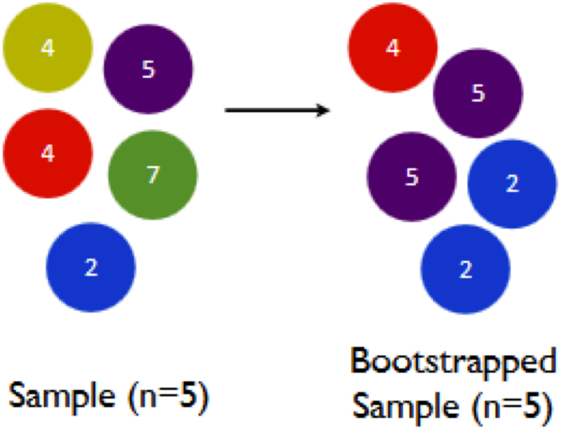
The Non Parametric Bootstrap

Sampling with replacement



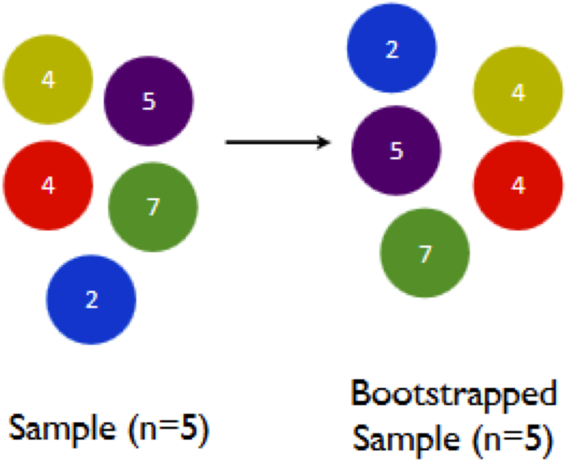
The Non Parametric Bootstrap

Sampling with replacement



The Non Parametric Bootstrap

Sampling with replacement



The Non Parametric Bootstrap

Weaker Assumptions: IID Errors

- The non-parametric bootstrap, relies on empirical data to resample and estimate the sampling distribution.
- This method allows us to understand the variability of a statistic without relying on theoretical distribution assumptions.

Why bootstrap works?

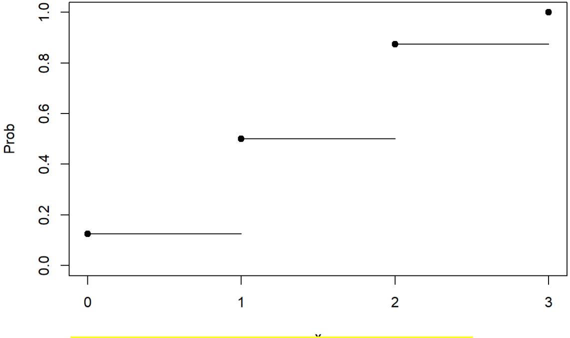
- The key is that the distribution of any estimator or statistic is determined by the distribution of the data.
- While the latter is unknown it can be estimated by the empirical distribution of the data.

Why bootstrap works?

Empirical Distribution Function

- **Empirical distribution function** Given a sample of data $(x_1, \ldots, x_n)$, each an iid realization of some random variable X, we define the empirical cumulative distribution function (ecdf) as:

$$\hat{F}(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i \leq x\} \quad \forall\, x \in \mathbb{R}$$



Why bootstrap works?

Glivenko-Cantelli Theorem

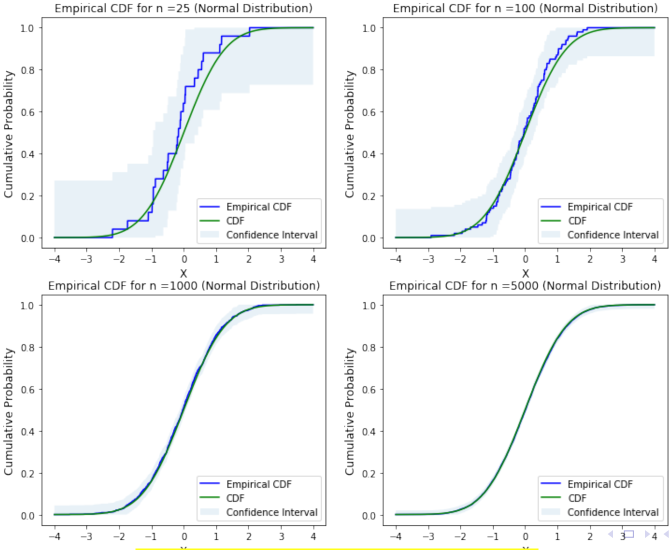
- **Glivenko-Cantelli**: Let $X_1, \ldots, X_n$ be a random sample from a distribution with cdf $F(x)$. Then:

$$\sup_{x \in \mathbb{R}} |\hat{F}(x) - F(x)| \to 0 \quad \text{as } n \to \infty$$

  where convergence is in probability.

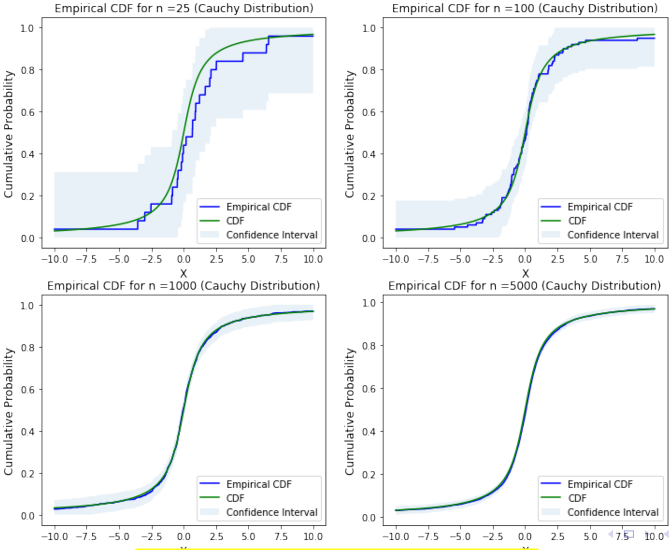
Why bootstrap works?

Glivenko-Cantelli Theorem



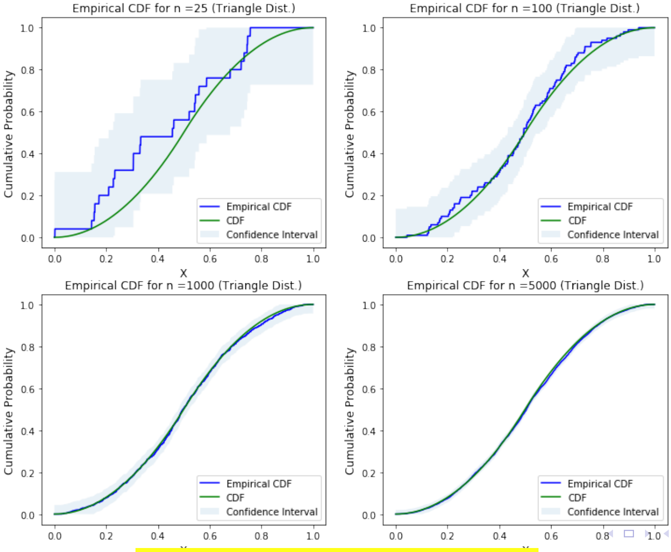
Why bootstrap works?

Glivenko-Cantelli Theorem



Why bootstrap works?

Glivenko-Cantelli Theorem



The Non Parametric Bootstrap

Variance in Linear Regression: IID Errors

- Suppose we have $y_i = \beta X_i + u_i \; i = \{1, \ldots, n\} \; u_i \sim_{iid} F(0, \sigma^2)$
- We want to estimate $Var(\hat{\beta})$

Agenda

1 Uncertainty

2 Resampling methods

3 The Non Parametric Bootstrap

4 The Wild Bootstrap

5 The "XY" Bootstrap
- Example: Elasticity of Demand for Gasoline

6 The Bag of Little Bootstraps

The Wild Bootstrap

Variance in Linear Regression: Heteroskedastic Errors

- For models with heteroskedastic errors, the standard procedure is to use White's heteroskedastic-consistent covariance matrix estimator (HCCME), which can perform poorly in small samples.
- The residual bootstrap can lead to invalid inference by assuming $u_i \mid x_i$ is iid, which may not be the case.
- The wild bootstrap, introduced by Wu (1986) and Liu (1988), refines the bootstrap for heteroskedastic errors.

Agenda

1 Uncertainty

2 Resampling methods

3 The Non Parametric Bootstrap

4 The Wild Bootstrap

5 The "XY" Bootstrap
- Example: Elasticity of Demand for Gasoline

6 The Bag of Little Bootstraps

The "XY" Bootstrap

- In regression, we need not use the residual bootstrap alone.
- A more direct implementation of the bootstrap would be to resample (x, y) pairs
- This approach is less sensitive to assumptions than the residual-based bootstrap introduced earlier.
  - In particular, it does not assume that the regression errors are iid, so it can accommodate heteroscedasticity, for example.

Agenda

1. Uncertainty
2. Resampling methods
3. The Non Parametric Bootstrap
4. The Wild Bootstrap
5. The "XY" Bootstrap
   - Example: Elasticity of Demand for Gasoline
6. The Bag of Little Bootstraps

Example: Elasticity of Demand for Gasoline

photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Agenda

1. Uncertainty
2. Resampling methods
3. The Non Parametric Bootstrap
4. The Wild Bootstrap
5. The "XY" Bootstrap
   - Example: Elasticity of Demand for Gasoline
6. The Bag of Little Bootstraps

The Bag of Little Bootstraps

- Standard bootstrap resampling can be tedious in large samples, often exceeding storage limits.
- Bickel and Sakov (2002) proposed the m out of n bootstrap method, which involves drawing a smaller sample $m < n$ for each replication.
- This method reduces computation by estimating variability from a smaller sample and rescaling for the reduced size.
- However, choosing m effectively is challenging, and without m being much smaller than n, computational gains are minimal.

Bag of Little Bootstraps: Kleiner et al. (2014)

- Kleiner et al. (2014) proposed an alternative scheme for bootstrap resampling that is easily parallelized, improving computational efficiency.
- When n is very large, bootstrap samples can also become very large, even with weighting to reduce the effective sample size.
- Kleiner et al. suggest splitting the sample into G groups, each of size S, with $GS \approx n$.
- Each group undergoes the usual bootstrap on S observations, and the variability measures are averaged across groups to estimate variability for the entire sample.
- Simulations indicate that setting $S = n^{\gamma}$ with $\gamma = 0.7$ works well, significantly speeding up the process.