

Linear Regression and Resampling Methods for Uncertainty

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

- 1 Review
- 2 Uncertainty: Motivation
- 3 What are resampling methods?
- 4 The Bootstrap
 - Example: Elasticity of Demand for Gasoline

Agenda

- 1 Review
- 2 Uncertainty: Motivation
- 3 What are resampling methods?
- 4 The Bootstrap
 - Example: Elasticity of Demand for Gasoline

Predicting Well

$$y = f(X) + u \quad (1)$$

- ▶ Interest on predicting y
- ▶ Model? We treat $f()$ as a black box, and any approximation $\hat{f}()$ that yields a good prediction is good enough (*"Whatever works, works..."*).
- ▶ How do we measure "what works"?

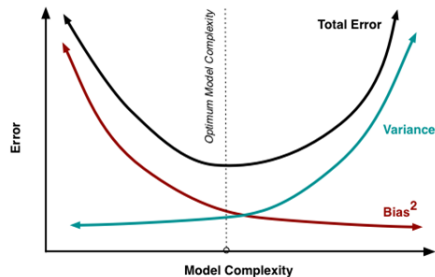
$$E(y - \hat{y})^2 = E(f(X) + u - \hat{f}(X))^2 \quad (2)$$

Predicting Well

$$E(y - \hat{y})^2 = \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(u)}_{\text{Irreducible}} \quad (3)$$

$$\text{MSE} = \text{Bias}^2(\hat{f}(X)) + V(\hat{f}(X)) + \underbrace{\text{Var}(u)}_{\text{Irreducible}} \quad (4)$$

The Bias-Variance Trade-Off



Source: <https://tinyurl.com/y4lvjxpc>

- The best kept secret: tolerating some bias is possible to reduce $V(\hat{f}(X))$ and lower MSE

Linear Regression

$$y = f(X) + u \quad (5)$$

$$= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + u \quad (6)$$

$$= X\beta + u \quad (7)$$

- ▶ If $f(X) = X\beta$, obtaining $f(\cdot)$ boils down to obtaining β
- ▶ where we can obtain β minimizing RSS (SSR)

$$\hat{\beta} = (X'X)^{-1}X'y \quad (8)$$

- ▶ Involves inverting a $k \times k$ matrix $X'X$
- ▶ requires allocating $O(nk + k^2)$ if n is "big" we cannot store this in memory

Linear Regression

► Gauss-Markov Theorem

- If it is assumed that $E(u|X) = 0$ and $E(uu'|X) = \sigma^2 I$ in the linear regression model $y = X\beta + u$, then the OLS estimator $\hat{\beta}$ is more efficient than any other linear unbiased estimator $\tilde{\beta}$, in the sense that $Var(\tilde{\beta}) - Var(\hat{\beta})$ is a positive semidefinite matrix.
- An informal way of stating this theorem is to say that $\hat{\beta}$ is the best linear unbiased estimator, or BLUE for short.
- In other words, the OLS estimator is more efficient than any other linear unbiased estimator.

Linear Regression

- ▶ Let's consider the simple case with two regressors:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u \quad (9)$$

- ▶ with $E(u) = 0$, $cov(x_1, u) = 0$, $cov(x_2, u) = 0$ and $E(u^2|x_1, x_2) = \sigma^2$
- ▶ OLS says we should choose the estimators $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ of $\beta_0, \beta_1, \beta_2$ such that we minimize the Sum of Square Residual (SSR) or the Residual Sum of Squares (RSS)

$$\mathcal{L} = \sum (y_i - \hat{y}_i)^2 \quad (10)$$

$$= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2 \quad (11)$$

Linear Regression

- The solution then comes from solving the FOC (and checking the SOC)

$$\frac{\partial \mathcal{L}}{\partial \hat{\beta}_0} = \sum 2 (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) (-1) = 0 \quad (12)$$

$$\frac{\partial \mathcal{L}}{\partial \hat{\beta}_1} = \sum 2 (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) (-x_{1i}) = 0 \quad (13)$$

$$\frac{\partial \mathcal{L}}{\partial \hat{\beta}_2} = \sum 2 (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) (-x_{2i}) = 0 \quad (14)$$

- Solving, for example, for $\hat{\beta}_2$ we have

$$\hat{\beta}_2 = \frac{\sum (x_{1i} - \bar{x}_1)^2 \sum (x_{2i} - \bar{x}_2) y_i - (\sum x_{1i} x_{2i} - n \bar{x}_1 \bar{x}_2) \sum (x_{1i} - \bar{x}_1) y_i}{\sum (x_{1i} - \bar{x}_1)^2 \sum (x_{2i} - \bar{x}_2)^2 - (\sum x_{1i} x_{2i} - n \bar{x}_1 \bar{x}_2)^2} \quad (15)$$

Goodness-of-fit. In sample performance

- ▶ The mechanics of OLS lead to a very simple measure of *goodness of fit*, the R^2 , or *coefficient of determination*, one of the most used and abused in the practice of econometrics and statistics.
- ▶ The starting point is the following *sum of squares decomposition*:

$$\sum \tilde{y}_i^2 = \sum \hat{y}_i^2 + \sum e_i^2,$$

- ▶ where $\tilde{y}_i \equiv Y_i - \bar{Y}$, $\hat{y}_i \equiv \hat{Y}_i - \bar{Y}$ and e_i are OLS residuals. The decomposition holds for any number of explanatory variables.
- ▶ where $\tilde{y}_i \equiv y_i - \bar{y}$, $\hat{y}_i \equiv \hat{y}_i - \bar{y}$ and e_i are OLS residuals.
- ▶ The decomposition holds for any number of explanatory variables, the derivation uses the FOC above

Goodness-of-fit. In sample performance

- ▶ To get some intuition, divide by n in both sides of the decomposition \rightarrow Each term resembles sort of a variance.
- ▶ The decomposition suggests that the total variability of Y can be 'explained' by the variability in the fitted model (ESS) plus that of the error term (RSS).

$$TSS = ESS + RSS$$

- ▶ The *coefficient of determination*, or R^2 for a given regression model is defined as

$$R^2 \equiv \frac{ESS}{TSS}$$

- ▶ which we can also rewrite as

$$R^2 \equiv 1 - \frac{RSS}{TSS}$$

Goodness-of-fit. In sample performance

- ▶ R_2 is non decreasing in the number of explanatory variables

High Leverage Points

- Suppose the simple model

$$y = \beta_0 + \beta_1 x + u$$

- the solution for $\hat{\beta}_1$ is

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \sum c_i y_i \quad (16)$$

Handwritten notes: $\text{Cov}(x, y)$ above numerator, $V(x)$ below denominator.

- the solution for $\hat{\beta}_0$ is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1}{n} \sum y_i - \bar{x} \sum c_i y_i \quad (17)$$

- With a bit of algebra we can write

$$\hat{y}_i = \sum \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) y_i \quad (18)$$

Handwritten notes: $\hat{y}_{40} = \sum_{i=1}^n h_i y_i$ and $\hat{y}_{40} = \left(\frac{1}{n} + \frac{(x_{40} - \bar{x})^2}{V(x)} \right) y_{40}$ are circled in yellow.

$$\hat{y}_i = \sum h_i y_i \quad (19)$$

Agenda

- 1 Review
- 2 Uncertainty: Motivation**
- 3 What are resampling methods?
- 4 The Bootstrap
 - Example: Elasticity of Demand for Gasoline

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 and hence “overfit.”
- ▶ The ability to deal with this mess and noise is the most important skill you need.

Uncertainty in Linear Regression

- To get a measure of the uncertainty, precision or variability of our estimates we need a measure

$$y = f(x) + u$$

- We can estimate the Variance of our estimators

- For example for

$$V(u) = \sigma^2$$

$$y_i = \beta_0 + \beta_1 x_{1i} + u \quad x \sim \mathcal{N} \quad (20)$$

$$\hat{\beta}_1 = \frac{\sum (x_{1i} - \bar{x}_1) y_i}{\sum (x_{1i} - \bar{x}_1)^2} \quad V(ax) = a^2 V(x) \quad (21)$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_{1i} - \bar{x}_1)^2} = \frac{\sigma^2}{n \text{Var}(x_{1i})} \quad (22)$$

Uncertainty in Linear Regression

- Let's go back to our two variable model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u \quad (23)$$

- the solution for β_2 was:

$$\hat{\beta}_2 = \frac{\sum (x_{1i} - \bar{x}_1)^2 \sum (x_{2i} - \bar{x}_2) y_i - (\sum x_{1i} x_{2i} - n \bar{x}_1 \bar{x}_2) \sum (x_{1i} - \bar{x}_1) y_i}{\sum (x_{1i} - \bar{x}_1)^2 \sum (x_{2i} - \bar{x}_2)^2 - (\sum x_{1i} x_{2i} - n \bar{x}_1 \bar{x}_2)^2} \quad (24)$$

Uncertainty in Linear Regression

- The variance?

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\underbrace{n \text{Var}(x_2)}_1 (1 - \underbrace{R_2^2}_{\uparrow})} \rightarrow \infty \quad (25)$$

- this is a special case of the very general formula

$$\text{Var}(\hat{\beta}_k) = \frac{\sigma^2}{\underbrace{n \text{Var}(x_k)}_1 (1 - R_k^2)} \quad (26)$$

Uncertainty and Resampling

- ▶ Sometimes the analytical expression of the variance can be quite complicated.
- ▶ In these cases we can use the bootstrap
- ▶ The bootstrap provides a way to perform statistical inference by resampling from the sample.

Agenda

- 1 Review
- 2 Uncertainty: Motivation
- 3 What are resampling methods?
- 4 The Bootstrap
 - Example: Elasticity of Demand for Gasoline

What are resampling methods?

- ▶ Tools that involves repeatedly drawing samples from a training set 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: estimate test error rates
 - ▶ They are computationally expensive! But these days we have powerful computers

Agenda

- 1 Review
- 2 Uncertainty: Motivation
- 3 What are resampling methods?
- 4 The Bootstrap
 - Example: Elasticity of Demand for Gasoline

The Bootstrap

Introduction

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

independientes
identicamente
distribuidos

varianza
finita

$X \sim \text{i.i.d.}$

$$V(aX) = a^2 V(X)$$

Desviación

$$\text{Var}(\bar{Y}) \quad (27)$$
$$\text{Var}\left(\frac{\sum y_i}{n}\right) = \frac{\sum \text{Var}(y_i)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$
$$\hat{\sigma}^2 = \frac{1}{n} \sum (y_i - \bar{y})^2 \rightarrow \text{desgasta } (n-1) \text{ para consistente.}$$

↓
conozco

The Bootstrap

Introduction

$$\text{Var}(\bar{y})$$

► Alternative way (no formula!)

1 From the n original data points y_1, y_2, \dots, y_n take a sample *with replacement* of size n

2 Calculate the sample average of this “pseudo-sample” $\bar{y}_{B=1}, \dots, \bar{y}_{B=1000}$

3 Repeat this B times.

4 Compute the variance of the B means

→ Esto se conoce como Bootstrap

$$\text{Var}(\bar{y}) = \frac{1}{B} \sum_{B=1}^{1000} (\bar{y}_B - \bar{\bar{y}})^2$$

The Bootstrap

Introduction

- ▶ The bootstrap is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- ▶ 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.”



The Bootstrap

Introduction

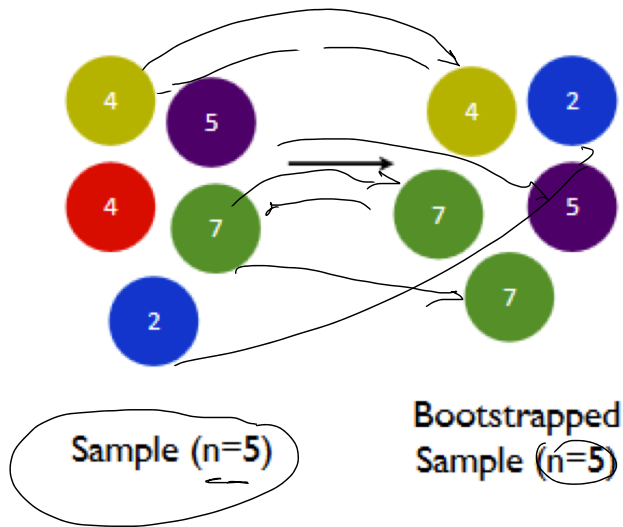
- ▶ **Key Innovation:** The sample itself is used to assess the precision of the estimate.
- ▶ Why would this work?
- ▶ Remember that uncertainty arises from the randomness inherent to our data-generating process
- ▶ So if we can approximately simulate this randomness, then we can approximately quantify our uncertainty.

The Bootstrap

Introduction

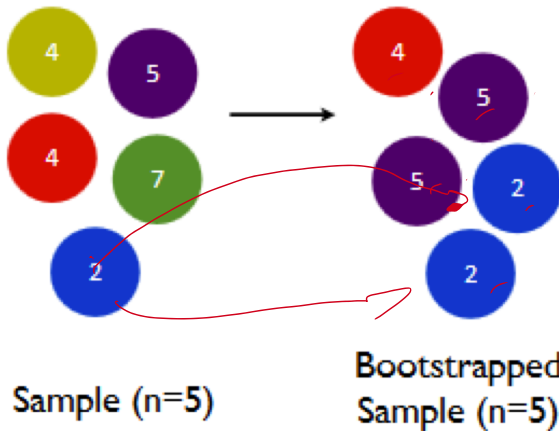
- ▶ There are two key properties of bootstrapping that make this seemingly crazy idea actually work.
 - 1 Each bootstrap sample must be of the same size (N) as the original sample
 - 2 Each bootstrap sample must be taken with replacement from the original sample

Sampling with replacement



Sampling with replacement

Resampling creates synthetic variability

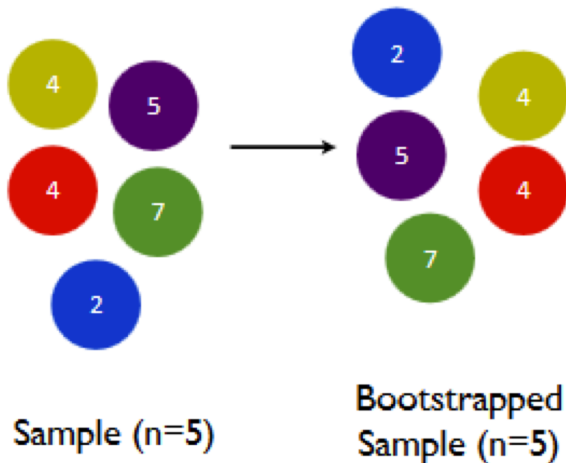


$B = 1000$

(2) (2)
(2) (2)
(2)

Sampling with replacement

Resampling creates synthetic variability



The Bootstrap

$$\text{Var}(\bar{y}) = ??$$

► In general terms:

► $Y_i, i = 1, \dots, n$

► θ is the magnitude of interest $\bar{y}, \beta_0, \beta_1 \rightarrow \text{I.C.}$

► To calculate it's variance

1 Sample of size n with replacement (*bootstrap sample*)

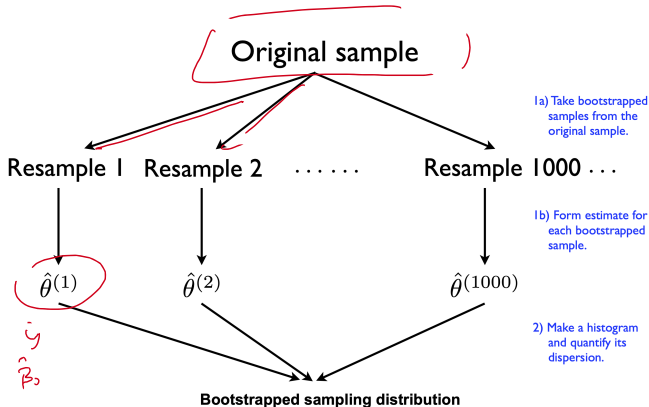
2 Compute $\hat{\theta}_j, j = 1, \dots, B \rightarrow \bar{y}$

3 Repeat B times $\bar{y}_{B=1} \dots \bar{y}_{B=1000}$

4 Calculate

$$\hat{V}(\hat{\theta})_B = \frac{1}{B} \sum_{j=1}^B (\hat{\theta}_j - \bar{\hat{\theta}})^2 \quad (28)$$

The Bootstrap



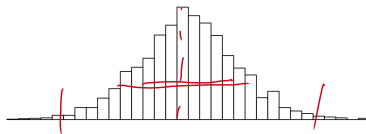
N

$$\hat{\beta} \rightarrow N(\mu, \sigma)$$

$$\bar{y} \rightarrow N(\mu, \sigma^2)$$

$Y_{\text{new}} \rightarrow ?!$
no

\rightarrow sampling distribution of $\hat{\theta}$



Example: Elasticity of Demand for Gasoline

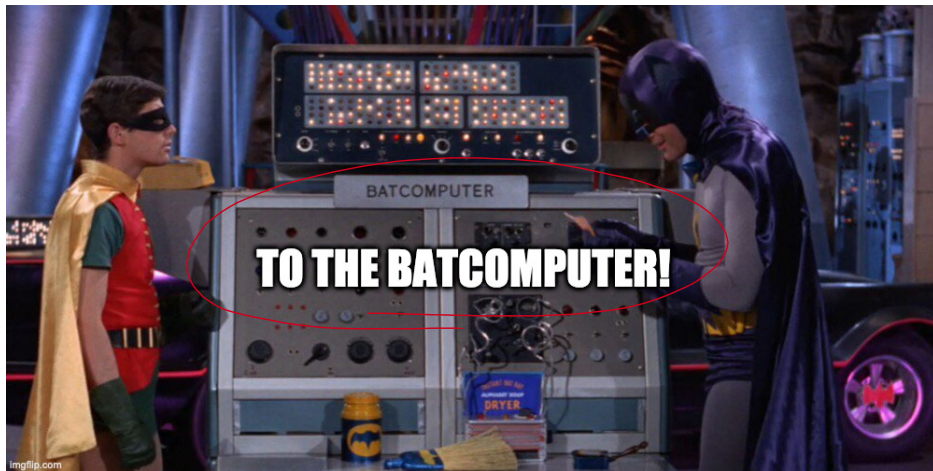


photo from <https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/>

Review and Caveats

- ▶ The bootstrap is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- ▶ The power of the bootstrap, and resampling in general, lies in the fact that it can be easily applied to a wide range of statistical learning methods.
- ▶ In particular, it does not assume that the regression errors are iid so it can accommodate heteroscedasticity.
- ▶ Of course it does still assume that the observations are independent.
- ▶ Resampling dependent observations is an inherently more difficult task which has generated its own rather large literature.