The Bootstrap (EH, Chapters 10 and 11)

Rob McCulloch

December 1, 2019

Background: Standard Errors

A basic idea in frequentist statistics is the *standard error*.

Give a "sample of data" x, we seek to estimate some unknown quantity θ.

Let $\hat{\theta} = s(x)$ denote our estimate from the sample x.

We understand that our sample as given imperfect information information so we seek a standard error $\hat{se}$ (which is also a function of x) such that

$$P(\theta \in \hat{\theta} \pm k_\alpha \hat{se}) = 1 - \alpha$$

The interval,

$$(\hat{\theta} - k_\alpha \, \hat{se}, \hat{\theta} - k_\alpha \, \hat{se})$$

is called a *confidence interval*, which coverage probability $(1 - \alpha)$.

The classic example is estimation of a mean.

If $s = \{X_1, X_2, \ldots, X_n\}$ is our sample where the X_i are iid from some distribution and $\theta = E(X)$.

Our estimator is $\hat{\theta} = \bar{X}$.

We let,

$$s^2 = \frac{1}{n-1} \sum (X_i - \bar{X}), \;\; \hat{se} = \frac{s}{\sqrt{n}}.$$

Then, for large enough n,

$$P(\theta \in \bar{X} \pm 1.96\, \hat{se}) \approx .95$$

About 95% of the time, the true value will be in the interval!

Let $Var(X) = E((X - \mu)^2) = \sigma^2$.

This result relies on some key assumptions

- The X_i are iid.
- $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$
- $Var(\bar{X})$ has the simple form σ^2/n.
- In large samples we can *plug-in* s^2 in place of σ^2.

How can we obtain standard errors and confidence intervals for estimators more complex than $\bar{X}$?

EH:

“Direct standard error formulas exist for various forms of averaging such as linear regression, and for hardly anything else.” (page 155)

The goal of the *Jacknife* and the *bootstrap* is to compute standard errors, or, more generally, confidence intervals for complex estimators (e.g. not averages) without making many assumptions.

And, to do it in a computationally feasible way.

Example

Supose you have the simple linear regression model and you want an interval for

$$E(Y \,|\, x) = \beta_0 + \beta_1 \, x$$

Easy!!

Example

Suppose you have a simple logistic regression model with one x and you want an interval for

$$P(Y = 1 \,|\, x) = F(\beta_0 + \beta_1 x); \;\; F(\eta) = \frac{e^\eta}{1 + e^\eta}$$

Not so easy.
Delta method??

The Jacknife Estimate of Standard Error

Suppose we have

$$x_i \sim F, iid, \quad i = 1, 2, \ldots n.$$

The x can belong to an set.

Let $x = (x_1, x_2, \ldots, x_n)$ and,

$$\hat{\theta} = s(x).$$

Note that s could be a complex algorithm, rather than a simple function.

We want to compute the standard error, that is, we want to estimate the standard deviation of $\hat{\theta} = s(x)$.

Let,

$$x_{(i)} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

and,

$$\hat{\theta}_{(i)} = s(x_{(i)}).$$

Then the jacknife estimate of the standard error for $\hat{\theta}$ is

$$\hat{se}_{\text{jack}} = \left[\frac{n-1}{n} \sum_{i=1}^{n} (\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^2 \right]^{1/2}, \text{ with } \hat{\theta}_{(.)} = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_{(i)}$$

The "fudge factor" $\frac{n-1}{n}$ is chosen to make $\hat{se}_{\text{jack}}$ the same as the classic formula for $\hat{\theta} = \bar{X}$.

Note

- intuitive that $(\hat{\theta}_{(i)} - \hat{\theta}_{(.)})$ captures sample variation in the estimator.
- fudge factor gets the scaling right.
- It is nonparametric, no special form for F need by assumed.
- It is automatic. Just need code for $s(x)$, then the same simple code works for everything.
- $\hat{se}_{\text{jack}}$ is upwardly biased.

Example:

Standard error of a correlation.

The Nonparametric Bootstrap

The standard error is the a measure of the variation we would observe if we repeately sampled x from F and computed $s(x)$ for each draw of x.

This is impossible since F is uknown.

Instead the bootstrap substitutes an estimate $\hat{F}$ for F, and then estimates the frequentist standard error by direct simulation.

That is:

- draw x repeately from $\hat{F}$.
- for each x draw, compute $s(x)$.
- compute the sample standard deviation of the draws.

For formalize this, we need the notion of a *bootstrap sample*.

Given observed $(x_1, x_2, \ldots, x_n)$ let a bootstrap sample

$$x^* = (x_1^*, x_2^*, \ldots, x_n^*)$$

where each x_i^* is drawn with equal probability *and replacement* from $\{x_1, x_2, \ldots, x_n\}$.

From each bootstrap sample we compute

$$\hat{\theta}^* = s(x^*).$$

We then draw B bootstrap samples $x^{*b}, b = 1, 2, \ldots, B$.

At each bootstrap sample we compute $\hat{\theta}$:

$$\hat{\theta}^{*b} = s(x^{*b}), \quad b = 1, 2, \ldots, B.$$

We then have:

$$\hat{se}_{\text{boot}} = \left[\frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}^{*b} - \hat{\theta}^{*\cdot})^2 \right]^{1/2}, \text{ with } \hat{\theta}^{*\cdot} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{*b}$$

We can few the bootstrap as *plugging in* the empirical distribution!!

Our model is

$$F \overset{iid}{\to} x \overset{s}{\to} \hat{\theta}.$$

In principle we would draw x repeatedly and observe the variation in $\hat{\theta}$.

Since we can't do this (don't know F) we *plug-in* an estimate

$$\hat{F} = \sum_{i=1}^{n} \frac{1}{n} \delta_{x_i},$$

where δ_x puts probability 1 on x.

$\hat{F}$ is simply the empirical distribution.

Plugging-in means we replace

$$F \overset{iid}{\rightarrow} x \overset{s}{\rightarrow} \hat{\theta}.$$

with,

$$\hat{F} \overset{iid}{\rightarrow} x^* \overset{s}{\rightarrow} \hat{\theta}^*.$$

We only get one $\hat{\theta}$, but we get $\hat{\theta}^{*b}$, $b = 1, 2, \ldots, B$, and we choose B.

Note, Jackknife and Bootstrap

- completely automatic. Input x and s, get out $\hat{se}_{\text{boot}}$.
- Bootstraping *shakes* the original data more violently than the jackknife.
- There is nothing special about standard errors, we could bootstrap to estimate $E(|\hat{\theta} - \theta|)$.
- The jackknife method is more conservative than the bootstrap method, that is, its estimated standard error tends to be slightly larger.
- Jackknife performs poorly when the the estimator is not sufficiently smooth, i.e., a non-smooth statistic for which the jackknife performs poorly is the median.
- bootstrap can be more computationally demanding.

Bootstrap Confidence Intervals

Why did we want to estimate the se?

We want to have some way of gauging the uncertainty associated with our estimation of θ given the amount of information in the sample x.

Can we use use the bootstrap to construct confidence intervals?

The obvious thing to try is the *standard interval*

$$\hat{\theta} \pm 1.96\,\hat{se}.$$

This interval is useful but may be inaccurate if the sampling distribution of $\hat{\theta}$ is not normal.

Typically we use Central Limit Theorem ideas to argue that $\hat{\theta}$ will be normal in "large samples" but the sample may not be large enough.

In particular the interval $\hat{\theta} \pm 1.96\,\hat{se}$ is always symmetric around $\hat{\theta}$ and that may not be appropriate if the sampling distribution of $\hat{\theta}$ is skewed.

There are a variety of ways to get confidence intervals from the bootstrap that perform better than the standard interval and we will just look at one simple approach, *the percentile method*.

The Percentile Method

The goal is to automate the computation of confidence intervals using the bootstrap distribution of the estimateor $\hat{\theta}$.

The percentile method uses the shape of the bootstrap empirical distribution of the

$$\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}$$

Let, $\hat{G}$ be the empirical CDF of the $\hat{\theta}^{*b}$, so that $\hat{G}(t)$ is the proportion of $\hat{\theta}^{*b}$ less than t

$$\hat{G}(t) = \#\{\hat{\theta}^{*b} \leq t\}/B.$$

Then the αth percentage point $\hat{\theta}^{*(\alpha)}$ given by the inverse function of $\hat{G}$,

$$\hat{\theta}^{*(\alpha)} = \hat{G}^{-1}(\alpha).$$

So, $\hat{\theta}^{*(\alpha)}$ is the value putting proportion α of the bootstrap sample $\hat{\theta}^{*b}$ to its left.

$$\hat{\theta}^{*(\alpha)} = \hat{G}^{-1}(\alpha).$$

Then, for example, the 95% central percentile interval is

$$(\hat{\theta}^{*(.025)}, \hat{\theta}^{*(.975)})$$

Notes:

- the method requires bootstrap samples on the order of $B = 2000$.
- the argument for the method centers around the fact that it is invariant to monotonic transformations of θ.
- two further improvements are "BC" and "BCa", where BC stands for *bias corrected* are covered in EH 11.3.

The Parametric Bootstrap

The nonparametric bootstrap can be described as:

$$\hat{F} \;\stackrel{iid}{\rightarrow}\; x^* \;\stackrel{s}{\rightarrow}\; \hat{\theta}^*.$$

where $\hat{F}$ is the empirical distribution.

The empirical distribution is appealing because it is nonparametric.

But, if we have a parametric family that we belief in or simply want to explore, we can get $\hat{F}$ from our parametric estimation.

Suppose $f(x \mid \mu)$ is a paramtric family.

Now suppose we have an estimate $\hat{\mu}$ (e.g the mle), then we can simply replace the empirical distribution with $f(x \mid \hat{\mu})$:

$$f(x \mid \hat{\mu}) \;\rightarrow\; x^* \;\rightarrow\; \hat{\theta}^*.$$

and get a bootstrap distribution estimate $\hat{se}_{\text{boot}}$ as before.

As before, we could bootstrap to get any quantitly of interest (not just the an se).

Basic Example

Suppose $x = (x_1, x_2, \ldots, x_n)$ are a sample assumed to be iid $N(\mu, 1)$.

Then $\hat{\mu} = \bar{x}$ and a parametric bootstrap sample is

$$x^* = (x_1^*, x_2^*, \ldots, x_n^*), \quad x_i^* \overset{iid}{\sim} N(\bar{x}, 1)$$

Not So Basic Example

Suppose we have

$$x_i = \alpha + \beta x_{i-1} + \epsilon_i, \;\; \epsilon_i \sim N(0, \sigma^2).$$

Given an esimtate $(\hat{\alpha}, \hat{\beta}, \hat{\sigma})$, we can draw bootstrap samples

$$x_i^* = \hat{\alpha} + \hat{\beta} x_{i-1}^* + \epsilon_i, \;\; \epsilon_i \sim N(0, \hat{\sigma}^2), \;\; i = 2, 3, \ldots, n.$$

Then we could, for example, get estimates of (α, β, σ) from each bootstrap sample.

Note:

For time series data there is a *Moving Blocks Bootstrap* (EH 10.3) but it seems tricky.

For more complex non iid models, the parametric bootstrap seems like just a great idea.

Perhaps more generally, we often want to test a complex modeling approach (model + computation).

Often we try it on simulated data and real data.

But, we never are sure the simulate data represent a good “use case” and we never know the truth with the real data.

Simulating data from a model fit to data seems like an approach worth thinking about in general.