The bootstrap

Patrick Breheny

September 11

Introduction

- Thus far, we have encountered influence functions and the jackknife as two nonparametric methods for assessing the uncertainty surrounding an estimate by observing how the estimate changes upon small changes to the data
- In our next few lectures, we will explore another method based on this same idea: the *bootstrap*
- The bootstrap is an extremely important idea in modern nonparametric statistics; indeed, Casella & Berger (2002) call it "perhaps the single most important development in statistical methodology in recent times"

Derivation of bootstrap

- Suppose we are interested in assessing the variance of an estimate $\hat{\theta} = \theta(\mathbf{x})$
- It's actual variance is given by

$$\mathbb{V}(\hat{\theta}) = \int \cdots \int \{\theta(x_1, \ldots, x_n) - \mathbb{E}(\hat{\theta})\}^2 dF(x_1) \cdots dF(x_n)$$

  where $\mathbb{E}(\hat{\theta}) = \int \cdots \int \theta(x_1, \ldots, x_n) dF(x_1) \cdots dF(x_n)$
- There are two problems with evaluating this expression directly

The ideal bootstrap

- The first is that we do not know F
- A natural solution would therefore be to use the plug-in principle:

$$\hat{\mathbb{V}}(\hat{\theta}) = \int \cdots \int \{\theta(x_1, \ldots, x_n) - \hat{\mathbb{E}}(\hat{\theta})\}^2 d\hat{F}(x_1) \cdots d\hat{F}(x_n)$$

- For reasons that will become clear, we will call this the *ideal bootstrap* estimate

The ideal bootstrap (cont'd)

- The second problem, however, is that this integral is difficult to evaluate
- Because $\hat{F}$ is discrete,

$$\hat{\mathbb{V}}(\hat{\theta}) = \sum_j \frac{1}{n^n}\{\theta(\mathbf{x}_j) - \hat{\mathbb{E}}(\hat{\theta})\}^2$$

  where $\mathbf{x}_j$ ranges over all n^n possible combinations of the observed data points $\{x_i\}$ (note, however, that only $\binom{2n-1}{n}$ are distinct)
- Unless n is very small, this may take a long time to evaluate

Monte Carlo approach

- However, we can approximate this answer instead using *Monte Carlo integration*
- Instead of actually evaluating the integral, we approximate it numerically by drawing random samples of size n from $\hat{F}$ and finding the sample average of the integrand
- This approach gives us the bootstrap – an approximation to the ideal bootstrap
- By the law of large numbers, this approximation will converge to the ideal bootstrap as the number of random samples that we draw goes to infinity

Bootstrap estimate of variance

The procedure for finding the bootstrap estimate of the variance (or "bootstrapping the variance") is as follows:

(1) Draw $\mathbf{x}_1^*, \ldots, \mathbf{x}_B^*$ from $\hat{F}$, where each *bootstrap sample* $\mathbf{x}_b^*$ is a random sample of n data points drawn iid from $\hat{F}$

(2) Calculate $\hat{\theta}_b^*$, where $\hat{\theta}_b^* = \theta(\mathbf{x}_b^*)$; these are called the *bootstrap replications*

(3) Let

$$v_{boot} = \frac{1}{B} \sum_{b=1}^{B} \left\{ \hat{\theta}_b^* - \bar{\theta}^* \right\}^2,$$

where $\bar{\theta}^* = B^{-1} \sum_b \hat{\theta}_b^*$

Resampling

- What does a random sample drawn from $\hat{F}$ look like?
- Because $\hat{F}$ places equal mass at every observed value x_i, drawing a random sample from $\hat{F}$ is equivalent to drawing n values, with replacement, from $\{x_i\}$
- In practice, this is how the $\mathbf{x}_i^*$'s from step 1 on the previous page are generated
- This somewhat curious phenomenon in which we draw new samples by sampling our original sample is called *resampling*

Bootstrap estimation of the CDF of $\hat{\theta}$

- The bootstrap is not limited to the variance
- We can use it to estimate the bias:

$$b_{boot} = \bar{\theta}^* - \hat{\theta}$$

- We can use it to estimate any aspect of the sampling distribution of $\hat{\theta}$, including its entire CDF
- Let G denote the CDF of $\hat{\theta}$; for any t,

$$\hat{G}(t) = \frac{1}{B}\sum_{b=1}^{B} I(\hat{\theta}_b^* \leq t)$$

- If $\theta = T(F)$ is Hadamard differentiable, then $\hat{G}$ is a consistent estimator of G (see our textbook for details)

The boot package

- I will not be making you write your own `bootstrap` function, as a nice R package already exists for doing this, called `boot`
- By default, it is installed but not loaded with R (*i.e.*, you will have to type `require(boot)` to use it)
- The package is fairly intuitive, with the exception of the fact that it requires the $\theta(\cdot)$ be written as a function of two arguments: the first is the original data and the second is a vector of indices specific to the bootstrap sample
- Thus, in order to, say, bootstrap the mean, you will need to define a function like the following:

```
mean.boot <- function(x, ind) {mean(x[ind])}
```

boot example

Once you have defined such a function, its usage is straightforward:

```
> boot(x, mean.boot, 1000)

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = x, statistic = mean.boot, R = 1000)

Bootstrap Statistics :
    original       bias    std. error
t1* 2.002748 -0.02296749   0.3185166
```

However, there is no point in actually bootstrapping the mean, as the ideal bootstrap estimates have closed form solution equal to 0 bias and the usual SE of the mean

How big should B be?

- What is a good value for B?
- On the one hand, computing time increases linearly with B, so we would like to get by with a small B
- This desire is particularly acute if θ is complicated and time-consuming to calculate
- On the other hand, the lower the value of B, the less accurate and more variable our estimated standard error is

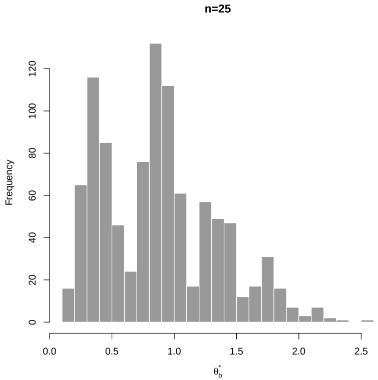
How big should B be? (cont'd)

- How much accuracy do we lose by stopping at B bootstrap samples instead of going to ∞?
- This can be assessed by standard statistical methods: $\{\hat{\theta}^*_b\}$ are iid, our bias estimate is derived from a mean, and our SE is a standard deviation
- Generally speaking, published articles in recent years tend to use 1000 bootstrap replications; however, for highly computer intensive statistics, 100 or even 50 may be acceptable
- However, each application is different – bootstrap data, just like real data, often deserves a closer look: in the words of Brad Efron, "it is almost never a waste of time to display a histogram of the bootstrap replications"

Histogram of bootstrap replications

```
out <- boot(x,var.boot,1000)
hist(out$t)
```

95% CI for bootstrap SE:
(.44, .48)



Homework

- In our last lecture, I introduced a problem asking your to calculate the standard error of the Pearson correlation for the LSAT data using three different methods (normal theory, functional delta method, and jackknife)
- **Homework:**
  (d) Use the bootstrap to estimate the the standard error of $\hat{\rho}$.
  (e) Plot a histogram of your bootstrap replications $\{\hat{\rho}_b^*\}$. Does the sampling distribution appear to be normally distributed?
  (f) Compare the four estimates (a)-(d).

Homework

- One great advantage of the bootstrap is that it can be used to calculate the standard error of arbitrarily complicated statistics
- For example, consider a study of student's test scores in various subjects
- 88 students were given 5 tests, on Mechanics, Vectors, Algebra, Analysis, and Statistics; their scores are on the course website
- The correlation matrix of these data provides a wealth of information about the relationships between these tests; for example, the correlation between statistics test scores and algebra test scores is .66

Homework (cont'd)

- One natural question about this data is the extent to which these tests measure separate skills vs. general tests of quantitative ability
- Educational testing theory often turns to *eigenvalues* to answer this question
- Eigenvalues measure the extent to which orthogonal linear combinations of elements can explain the patterns seen in multivariate data
- They can be used to address the above question by looking at the following statistic:

$$\hat{\theta} = \hat{\lambda}_1 / \sum_{i=1}^{5} \hat{\lambda}_i,$$

where $\{\hat{\lambda}_i\}$ are the eigenvalues, sorted from largest to smallest

Homework (cont'd)

- For the test score data, $\hat{\theta} = .636$, indicating that "latent quantitative ability" is able to explain about 64% of the variation in students' test scores, the rest of the variability coming from subject-specific abilities or just random noise
- **Homework:**
  (a) Use the bootstrap to estimate the standard error of $\hat{\theta}$.
  (b) Plot a histogram of your bootstrap replications $\{\hat{\theta}_b^*\}$. Does the sampling distribution appear to be normally distributed?
- Hints: `cor` is the function in R for calculating correlation matrices; `eigen` calculates eigenvalues

The Parametric Bootstrap

- This feature of the bootstrap — the way in which it can be used to easily calculate standard errors for complicated statistics — has led to its use in parametric problems as well
- The only difference from what we have described is in the first step of the bootstrap, where the bootstrap samples $\mathbf{x}_b^*$ are generated
- In the parametric bootstrap, rather than generate $\mathbf{x}_b^*$ from the empirical CDF $\hat{F}$, we generate $\mathbf{x}_b^*$ from the MLE of F, $F_{\hat{\theta}}$