Topic 10: Brief Introduction to the Bootstrap

History[1]

- "Bootstrap" coined by Efron (1979)
- becomes popular about 10 years later

Notation

In most Econometrics problems, we calculate a statistic, $\hat{\theta}$, from the sample (size n). For example, $\hat{\theta}$ could be an estimator (OLS slope coefficient, MLE), a test statistic, R^2, etc.

[1] Davidson & MacKinnon, 2006

We will denote $\hat{\theta}_j^*$ as the bootstrap statistic calculated from the j^{th} *bootstrap sample*.

Bootstrap Idea

The idea behind the (simple) bootstrap is to use the original sample, n, in place of the *population.* We can then mimic "repeated sampling from the population" in order to emulate the sampling distribution of $\hat{\theta}$. $\hat{\theta}$ is used in place of the true population parameter, θ.

Simple Bootstrap

Bootstrapping allows us to determine properties of $\hat{\theta}$, using only (i) the sample at hand; (ii) an iterative procedure (typically using a computer).

The procedure:

1. Draw n observations with replacement from the original data, in order to create a *bootstrap sample* or *resample*.

2. Calculate the statistic of interest, $\hat{\theta}_j^*$, from the bootstrap sample.

3. Repeat k times (k = 10000, say). Save $\hat{\theta}_j^*$ each time.

4. We now have $\{\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_k^*\}$.

Note: If we make the following substitutions:

original data	=	population
bootstrap sample	=	sample
k	=	infinite

then we have the definition of the sampling distribution.

The 10000 bootstrap statistics ($\hat{\theta}_j^*s$) comprise the *bootstrap distribution* (the bootstrap distribution emulates the sampling distribution of $\hat{\theta}$).

Use of the Bootstrap

In a sense, the 10000 bootstrap statistics comprise the *empirical sampling distribution*. Up to this point, we have only worked with the theoretical sampling distribution (Ryan – recap), and sometimes this theoretical distribution is only asymptotically valid (Ryan – discuss some examples).

So, all of the things we used the theoretical sampling distributions for, we can also use the bootstrap distribution for.

For example:

- calculate the standard error of the $\hat{\theta}_j^* s$: $s_{\hat{\theta}^*} = \sqrt{\frac{\sum_{j=1}^{k}\left(\hat{\theta}_j^* - \bar{\hat{\theta}}^*\right)^2}{k-1}}$
- construct (95%) confidence intervals for $\hat{\theta}$: sort the $\hat{\theta}_j^* s$ in ascending order, get $[\hat{\theta}^*_{k\times.05} \;\;,\;\; \hat{\theta}^*_{k\times.95}]$[2]
- obtain a p-value from a bootstrap test: observe the portion of $\hat{\theta}_j^* s$ which are more extreme than H_0.
- estimate the bias of $\hat{\theta}$ by $\bar{\hat{\theta}}^* - \hat{\theta}$. A bias-corrected estimator for $\hat{\theta}$ is then: $2\hat{\theta} - \bar{\hat{\theta}}^*$

---

[2] Not quite. This is wrong.

Why Use the Bootstrap?

In some situations the sampling distribution will be difficult or impossible to obtain. In addition, it has been shown that the bootstrap can have better finite sample properties than asymptotic approximations (see below).

Properties of the Bootstrap

- Under certain conditions, the bootstrap yields a consistent estimator of the distribution of a statistic. That is, the bootstrap distribution gets the asymptotic distribution of $\hat{\theta}$ right, if the sample is sufficiently large.
- "…the bootstrap often does much more than get the asymptotic distribution right. In a large number of situations that are important in applied econometrics, it provides a higher-order asymptotic approximation to the distribution of a statistic."[3]

---

[3] Horowitz, 2001, pg. 3172

When the Bootstrap Doesn't Work[4]

It doesn't always work!

- if $\hat{\theta}$ is not asymptotically pivotal

There are many versions of the bootstrap. Refinements need to be used in the case of:

- dependent data
- semi-parametric or non-parametric estimators
- non-smooth estimators

---

[4] Horowitz, 2001

Illustrations

These illustrations use the Cobb-Douglas data from Greene, used in class already (n = 25):

```
cobbdata=read.csv("http://home.cc.umanitoba.ca/~godwinrt/7010/cob
      b.csv")
attach(cobbdata)
```

OLS:

```
Call:
lm(formula = log(y) ~ log(k) + log(l))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8444     0.2336   7.896 7.33e-08 ***
log(k)        0.2454     0.1069   2.297   0.0315 *
log(l)        0.8052     0.1263   6.373 2.06e-06 ***
```

Two estimators for σ^2:

```
> s_hat = (sum(res$residuals^2))/22
> s_hat
[1] 0.05555727
```

$$s^2 = \frac{e'e}{n-k} = 0.056$$

```
> sigma_hat = (sum(res$residuals^2))/25
> sigma_hat
[1] 0.0488904
```

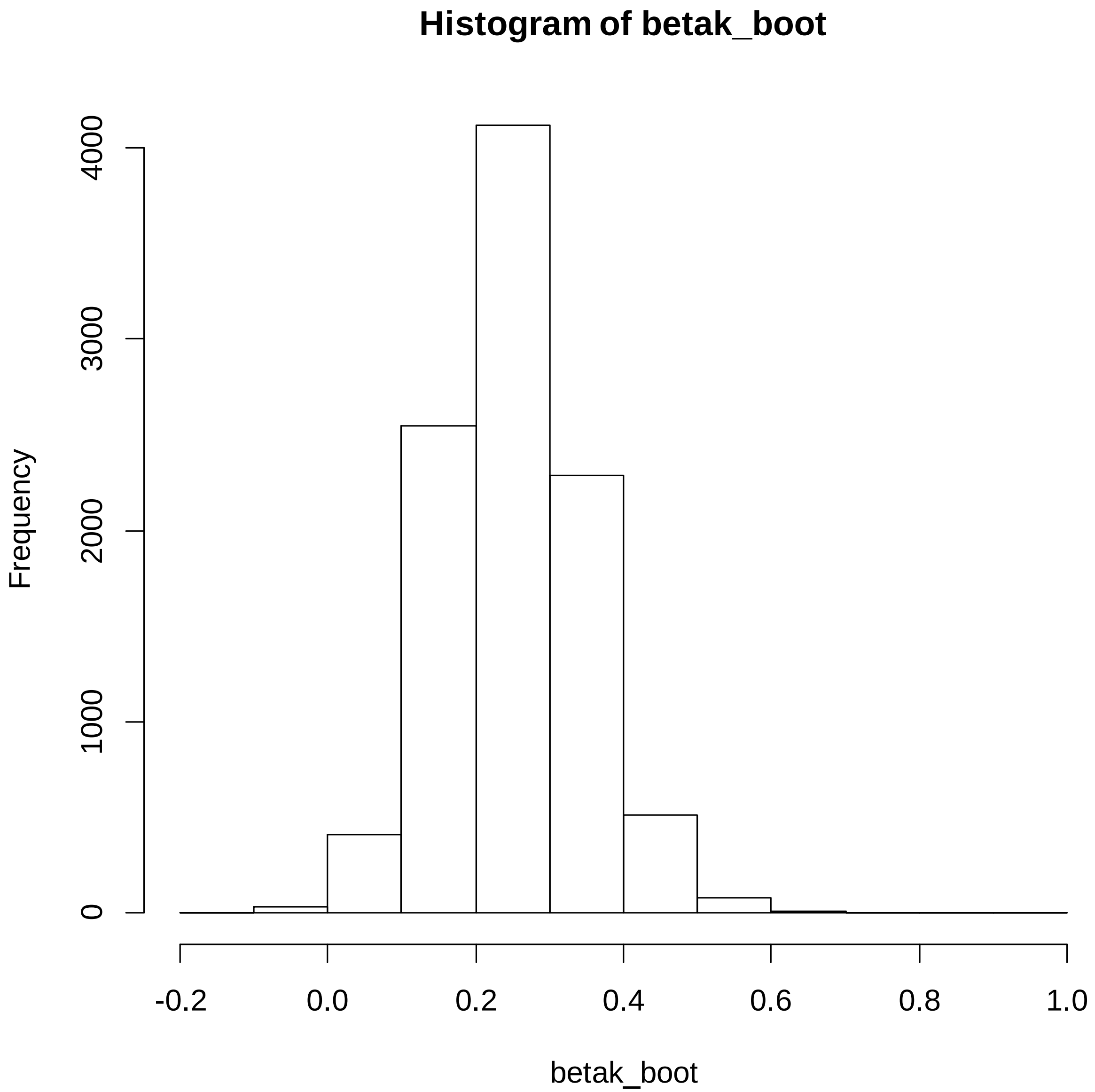
$$\hat{\sigma}^2 = \frac{e'e}{n} = 0.049$$

The bootstrap code:

```
sigma_boot = s_boot = betak_boot = rep(0,10000)

for(j in 1:10000){
resample = round(runif(25, min = 0.5, max = 25.5))
booty = y[resample]
bootk = k[resample]
bootl = l[resample]

resboot = lm(log(booty) ~ log(bootk) + log(bootl))

sigma_boot[j] = (sum(resboot$residuals^2))/25
s_boot[j] = (sum(resboot$residuals^2))/22
betak_boot[j] = resboot$coeff[2]
}
```

Let’s “see” the bootstrap distribution of b_k:

```
> hist(betak_boot)
```



Histogram of betak_boot
Frequency
0
1000
2000
3000
4000
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
betak_boot

Calculating a Bootstrap p-value Directly

$$H_0: \beta_k = 0$$

```
> sum(betak_boot < 0)/10000
[1] 0.0033
```

95% Confidence Interval

```
> betak_boot_sorted = sort(betak_boot)
> betak_boot_sorted[501]
[1] 0.1047377
> betak_boot_sorted[9500]
[1] 0.4101228
```

For this, *k* should equal 9999.[5]

---

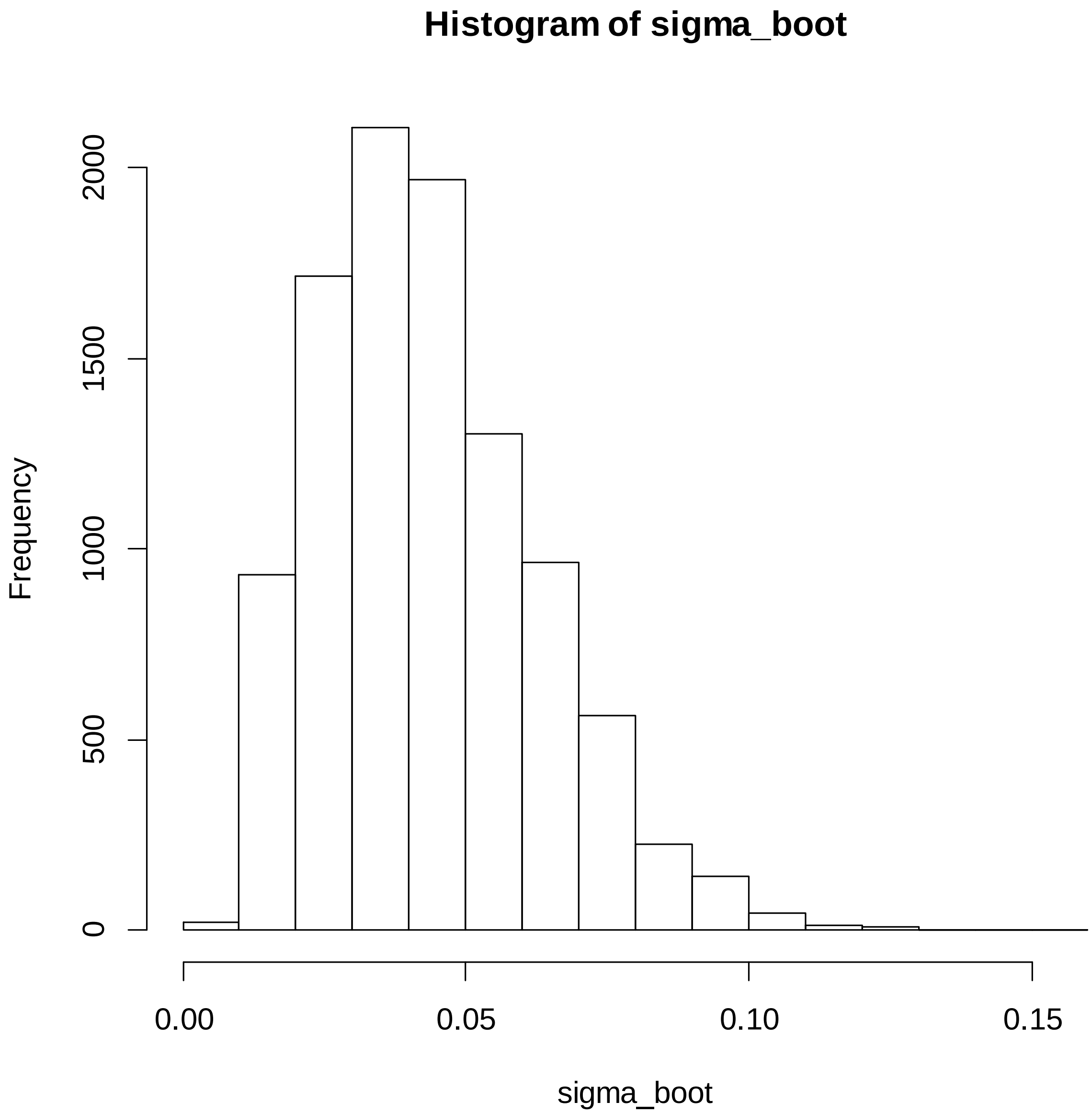
[5] How many bootstraps? See Davidosn & MacKinnon, 2000.

Bias Correction

$$\hat{\sigma}^2 = \frac{e'e}{n} = 0.049$$

If $\frac{e'e}{n}$ is an unbiased estimator for σ^2, and if 0.049 is the true population value, then $E\left[\frac{e'e}{n}\right] = 0.049$.

```
> hist(sigma_boot)
```


Histogram of sigma_boot
Frequency
0
500
1000
1500
2000
0.00
0.05
0.10
0.15
sigma_boot

```
> mean(sigma_boot)
[1] 0.04360541
> 2*sigma_hat - mean(sigma_boot)
[1] 0.05417539
```

References

Davidson, Russell, and James G. MacKinnon. "Bootstrap tests: How many bootstraps?." *Econometric Reviews* 19.1 (2000): 55-68.

Davidson, Russell, and James G. MacKinnon. "Bootstrap methods in econometrics." (2006): 812-838.

Hesterberg, Tim. "What Teachers Should Know about the Bootstrap: Resampling in the Undergraduate Statistics Curriculum." *arXiv preprint arXiv:1411.5279* (2014).

Horowitz, Joel L. "The bootstrap." *Handbook of econometrics* 5 (2001): 3159-3228.