Section 11: Introduction to Bootstrap

Yotam Shem-Tov

Fall 2014

The bootstrap

- The bootstrap yields consistent variance estimates under very mild conditions
- Let $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ be the realized sample observations (our data)
- Let $F_{\mathbf{z}}(\cdot)$ be the distribution of $\mathbf{z}$.
- Each observation $\mathbf{z}_i$ is a vector of length p, where p is the number of covariates
- Let $T_n \equiv T_n(\mathbf{z}_1, \ldots, \mathbf{z}_n, F_{\mathbf{z}})$ be a test statistic that is some function of the realized sample observations and the population distribution function of the data $F_{\mathbf{z}}$

The bootstrap

- Let $G_n(t, F_{\mathbf{z}})$ be the distribution of the test statistic (T_n),

$$G_n(t, F_{\mathbf{z}}) = Pr(T_n \leq t | F_{\mathbf{z}})$$

- How can we find the distribution of T_n?
  1. Assume a distribution of the data ($\mathbf{z} \sim F_{\mathbf{z}}$) and derive the distribution of T_n.
  2. Asymptotic distribution of T_n,

$$G_n(t, F) \xrightarrow{d} G(t, F) \Leftrightarrow \lim_{n \to \infty} G_n(t, F) = G(t, F)$$

  3. Bootstrap

  Let $G_n^*(t, F_{\mathbf{z}})$ be the bootstrap approximation of $G_n(t, F_{\mathbf{z}})$

Example: Lalonde 1986

- We consider as a data example the NSW randomized control trail of job training, analysed by Lalonde (1986).
- We restrict the sample to observation with information on prior earnings (*re74, re75*).
- Link to the data is here

Example: Lalonde (1986)

Experimental balance table

	Ave. Treat	Ave. control	T-test	Wilcoxon	KS
age	25.816	25.054	0.266	0.215	0.748
ed	10.346	10.088	0.150	0.056	0.063
black	0.843	0.827	0.647	0.649	1.000
hisp	0.059	0.108	0.064	0.077	0.963
married	0.189	0.154	0.334	0.327	0.999
nodeg	0.708	0.835	0.002	0.001	0.063
re74	2095.574	2107.027	0.982	0.361	0.970
re75	1532.055	1266.909	0.385	0.061	0.164

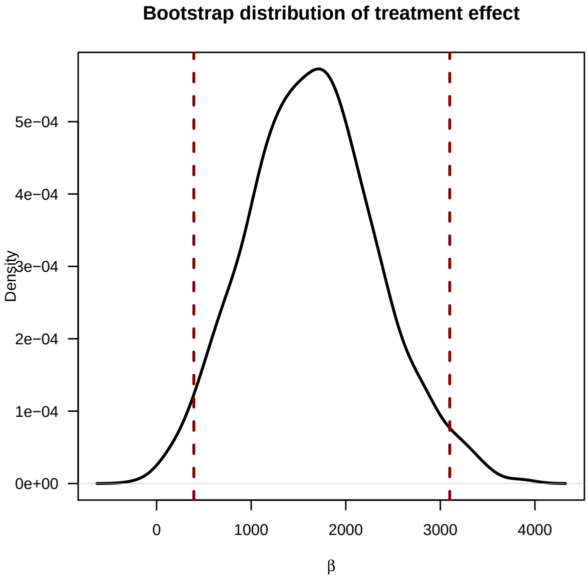
Example: Lalonde (1986)

```
Call:
lm(formula = re78 ~ (.), data = d)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.070e+02  4.808e+03  -0.064  0.94911
treat        1.676e+03  6.393e+02   2.621  0.00907 **
age          1.417e+02  2.744e+02   0.516  0.60578
ed           3.850e+02  2.301e+02   1.673  0.09501 .
black       -2.156e+03  1.170e+03  -1.842  0.06617 .
hisp         1.873e+02  1.553e+03   0.121  0.90406
married     -1.849e+02  8.924e+02  -0.207  0.83597
nodeg       -5.554e+01  1.007e+03  -0.055  0.95602
re74         8.148e-02  7.753e-02   1.051  0.29389
re75         5.082e-02  1.358e-01   0.374  0.70835
age2        -1.435e+00  4.495e+00  -0.319  0.74966
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6520 on 434 degrees of freedom
Multiple R-squared:  0.05505, Adjusted R-squared:  0.03328
F-statistic: 2.528 on 10 and 434 DF,  p-value: 0.005737
```

Lalonde (1986): *Non-parametric* bootstrap code

```
B=1000
beta.boot = rep(NA,B)
for (b in c(1:B)){
  if(b %% 50==0){cat("Iteration: ",b,"\n")}
  index = sample(rownames(d),length(rownames(d)),
   replace=TRUE)
  d0 = d[index,]
  beta.boot[b] = coef(lm(re78~(.),data=d0))[2]
}
```

Lalonde (1986): Bootstrap distribution of the ATE



Lalonde (1986): *Parametric* bootstrap code

```
B=1000
n = dim(d)[1]
beta.boot.parm = rep(NA,B)
for (b in c(1:B)){
  if(b %% 50==0){cat("Iteration: ",b,"\n")}
  epsilon.b = sample(lm1$res,n,replace=TRUE)
  y.b = as.matrix(cbind(rep(1,n),treat,x,x[,"age"]^2))
  %*% matrix(coef(lm1),ncol=1) +epsilon.b
  beta.boot.parm[b] = coef(lm(y.b~(.),data=d))[2]
}
```

Lalonde (1986): Comparison of CI estimations

	2.5%	97.5%
Non-parametric bootstrap	393.45	3099.37
Parametric bootstrap	536.45	2926.94
Analytical	419.27	2932.46

Implementation in *R*

Always know how to write your own code!

- The "boot" package in CRAN implements bootstrap
- The "boot" function allows for parallel computing, bias adjustment and other options.

```
f.lm = function(data,index){
  return(coef(lm(re78~(.),data=data[index,]))[2])
}

boot0 = boot(data=d,statistic=f.lm,R=1000)
```

Implementation in *R*

```
Call:
boot(data = d, statistic = f.lm, R = 1000)


Bootstrap Statistics :
    original    bias    std. error
t1* 1675.862 8.535902    684.6722
```

Implementation in *R*