

# The Bootstrap

November 4, 2009

# The Sample Mean and the Sample Median

- Let  $X_i$  be IID for  $i = 1, \dots, n$ , with mean  $\mu$  and variance  $\sigma^2$ . We use the sample mean  $\bar{X}$  to estimate  $\mu$ .
- Is the estimator biased? What is its standard error?
- Of course, we know it's unbiased and the SE is  $\sigma/\sqrt{n}$ , where  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$
- What about the median, particularly the difference in medians? Except for special circumstances, we don't have closed-form formulas for the uncertainty associated with these quantities.

## The Bootstrap Algorithm for estimating standard errors

- 1 Select  $B$  independent bootstrap samples  $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \dots, \mathbf{x}^{*B}$ , each consisting of  $n$  data values drawn **with replacement** from  $x$ .
- 2 Evaluate the bootstrap replication corresponding to each bootstrap sample,

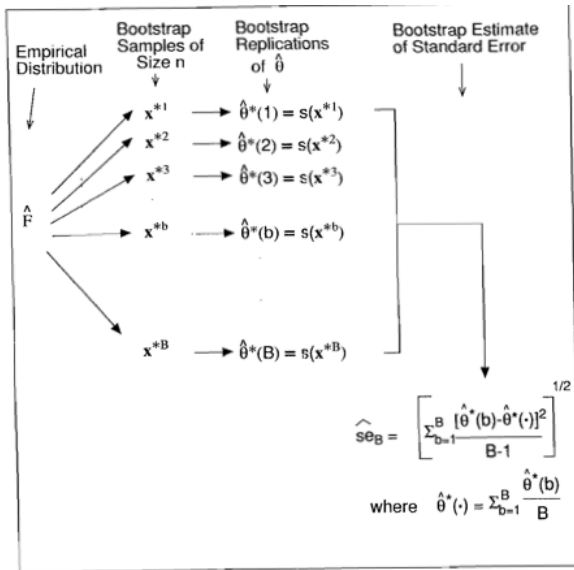
$$\hat{\theta}^*(b) = s(\mathbf{x}^{*b}) \quad b = 1, 2, \dots, B.$$

- 3 Estimate the  $se_F(\hat{\theta})$  by the sample standard deviation of the  $B$  replications

$$\widehat{se}_B = \left\{ \sum_{b=1}^B [\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2 / (B - 1) \right\}^{1/2};$$

where  $\theta^*(\cdot) = \sum_{b=1}^B \hat{\theta}^*(b) / B$

# The Bootstrap Algorithm for SE



# The Plug-in Principle

- We observe a random sample of size  $n$  from a probability distribution  $F$ ,

$$F \rightarrow (x_1, x_2, \dots, x_n)$$

the empirical distribution function  $\hat{F}$  is defined to be the discrete distribution that puts probability  $1/n$  on each value  $x_i, i = 1, \dots, n$ .

- The plug-in estimate of a parameter  $\theta = t(F)$  is defined to be  $\hat{\theta} = t(\hat{F})$ .

# Confidence Intervals

- Standard confidence intervals depend on the large sample or asymptotic result:

$$\frac{\hat{\theta} - \theta}{\hat{se}} \sim N(0, 1)$$

- In finite samples, we may not want to rely on that result and we instead can use bootstrapped confidence intervals.

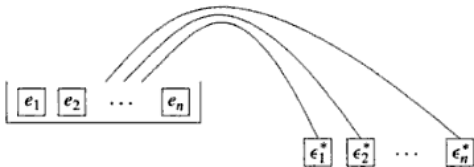
## Percentile Method

- Many methods of bootstrapped confidence intervals, but the **percentile** method is probably the easiest and most intuitive.
- To proceed we generate  $B$  independent bootstrap data sets  $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \dots, \mathbf{x}^{*B}$  and compute the bootstrap replications  $\hat{\theta}^*(b) = s(\mathbf{x}^{*b})$  for  $b = 1, 2, \dots, B$ .
- Let  $\hat{\theta}_B^{*(a)}$  be the  $100 \cdot \alpha$ th empirical percentile of the  $\hat{\theta}^*(b)$  values, that is the  $B \cdot \alpha$ th value in the ordered list of  $B$  replications of  $\hat{\theta}^*$ . Likewise let the  $\hat{\theta}_B^{*(1-a)}$  be the  $100 \cdot (1 - \alpha)$ th empirical percentile.
- The approximate  $1 - 2\alpha$  percentile interval is

$$[\hat{\theta}_{\%,lo}, \hat{\theta}_{\%,up}] = [\hat{\theta}_B^{*(a)}, \hat{\theta}_B^{*(1-a)}]$$

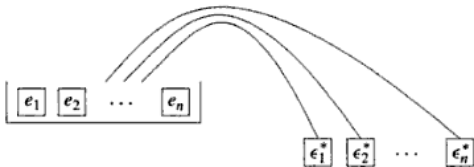
## Regression Models

- Suppose  $Y = X\beta + \epsilon$ , where the design matrix is  $n \times p$ ,  $X$  is fixed and has full rank. The parameter vector  $\beta$  is  $p \times 1$ , unknown, to be estimated by OLS. The errors  $\epsilon_1, \dots, \epsilon_n$  are IID with mean 0 and variance  $\sigma^2$ .
- If we forgot the formulas and wanted to estimate the bias and variance of OLS, we could use the bootstrap. What's random here?
- In the model, the  $Y_i$ 's are random, but not IID. The  $\epsilon_i$  are random and IID but unobserved. What do we do?
- We can re-sample the residuals:  $e = Y - X\hat{\beta}$ .



## Regression Models

- Suppose  $Y = X\beta + \epsilon$ , where the design matrix is  $n \times p$ ,  $X$  is fixed and has full rank. The parameter vector  $\beta$  is  $p \times 1$ , unknown, to be estimated by OLS. The errors  $\epsilon_1, \dots, \epsilon_n$  are IID with mean 0 and variance  $\sigma^2$ .
- If we forgot the formulas and wanted to estimate the bias and variance of OLS, we could use the bootstrap. What's random here?
- In the model, the  $Y_i$ 's are random, but not IID. The  $\epsilon_i$  are random and IID but unobserved. What do we do?
- We can re-sample the residuals:  $e = Y - X\hat{\beta}$ .



# Regression Models

- We draw  $n$  times at random with replacement from this population to get bootstrap errors  $\epsilon_1^*, \dots, \epsilon_n^*$ . These are IID (because you sample them that way).
- Next we generate the  $Y_i^*$ :

$$Y^* = X\hat{\beta} + \epsilon^*$$

- With the  $Y^*$  and  $X$ , we can directly examine the distribution of  $\hat{\beta}^*$ , where  $\hat{\beta}^* = (X'X)^{-1}X'Y^*$ .
- Note that this is known as the “parametric” bootstrap.
- The distribution of  $\hat{\beta}^* - \hat{\beta}$  is a good approximation for the distribution of  $\hat{\beta} - \beta$ . In addition, the empirical covariance matrix of the  $\hat{\beta}^*$  is a good approximation to the theoretical covariance matrix of  $\hat{\beta}$ .