Resampling

Professor: Hammou El Barmi
Columbia University

Introduction

- Finding a single set of estimates for the parameters in a statistical model is not enough. An assessment of the uncertainty in these estimates is also needed.
- Standard errors and confidence intervals are common methods for expressing uncertainty
- Before modern computers, doing statistical analysis involved using mathematics and probability theory to derive statistical formulas for standard errors and confidence intervals. Often these formulas are approximations that rely on large samples
- And it was sometimes difficult, if not impossible, to assess uncertainty, especially for complex models.

Introduction

- Fortunately, the speed of modern computers, and the innovations in statistical methodology inspired by this speed, have largely overcome this problem.
- With modern computers and statistical software, resampling methods (e.g. bootstrapping) can be used to produce standard errors and confidence intervals without the use of formulas that are often more reliable than statistical formulas

Introduction

- In this chapter we apply a computer simulation technique called the bootstrap to find standard errors and confidence intervals.
- The bootstrap method is very widely applicable and is one way that modern computing has revolutionized statistics.
- Advantages of Bootstrapping
    1. Fewer Assumptions (Do not need the data to be normally distributed)
    2. Greater Accuracy (Do not rely on very large sample sizes in contract to Central Limit Theorem)
    3. Generality (Same technique works for a wide variety of statistics

Introduction

- Idea: a sample is a good representative of the population, and we can simulate sampling from the population by sampling from the sample. The process is called resampling.
- Each resample has the same sample size n as the original sample. The reason for this is that we are trying to simulate the original sampling, so we want the resampling to be as similar as possible to the original sampling.
- By bootstrap approximation, we mean the approximation of the sampling process by resampling.

Bootstrap Estimates of Bias, Standard Deviation, and MSE

- Let θ be a one-dimensional parameter, let $\hat{\theta}$ be its estimate based on a sample, $\{R_1, R_2, \dots, R_n\}$ be our sample
- Create B bootstrap samples by sampling with replacement from the original data. Each sample has n observations (same a the original sample)

$$\begin{aligned}
\{R_{11}^*, R_{12}^*, \dots, R_{1n}^*\} &= \text{1st bootstrap sample} \\
\{R_{11}^*, R_{12}^*, \dots, R_{1n}^*\} &= \text{2nd bootstrap sample} \\
\vdots &= \vdots \\
\{R_{11}^*, R_{12}^*, \dots, R_{1n}^*\} &= \text{Bth bootstrap sample}
\end{aligned}$$

(An asterisk indicates a statistic calculated from a resample).

- Let $\hat{\theta}_1^*, \hat{\theta}_1^*, \dots, \hat{\theta}_B^*$ be estimates from B resamples. Also, define $\overline{\hat{\theta}}^*$ to be the mean of $\hat{\theta}_1^*, \hat{\theta}_1^*, \dots, \hat{\theta}_B^*$

Bootstrap Estimates of Bias, Standard Deviation, and MSE

- The bias of $\hat{\theta}$ is defined as $\text{BIAS}(\theta) = E(\hat{\theta}) - \theta$
- Since expectations, which are population averages, are estimated by averaging over resamples, the bootstrap estimate of bias is

$$\text{BIAS}_{\text{boot}}(\hat{\theta}) = \overline{\hat{\theta}}^{*} - \hat{\theta}$$

- In the bootstrap estimate of bias, the unknown population parameter θ is replaced by the estimate $\hat{\theta}$ from the sample.
- The bootstrap standard error for $\hat{\theta}$ is the sample standard deviation of $\hat{\theta}_1^*, \hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$

$$s_{\text{boot}} = \sqrt{\frac{1}{B-1} \sum_{i=1}^{B} (\hat{\theta}_i - \overline{\hat{\theta}}^{*})^2}$$

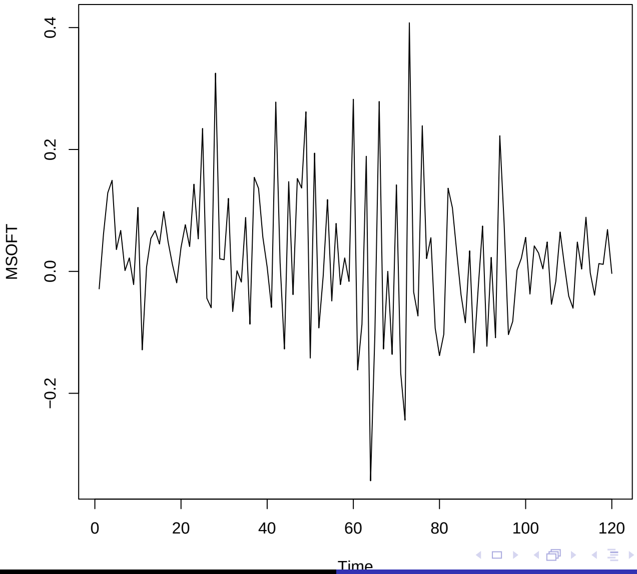
- s_{boot} estimates the standard deviation of $\hat{\theta}$

Bootstrap Estimates of Bias, Standard Deviation, and MSE

- The mean square error (MSE) of $\hat{\theta}$ is $E(\hat{\theta} - \theta)^2$ and is estimated by

$$\text{MSE}_{\text{boot}}(\hat{\theta}) = \frac{1}{B}\sum_{i=1}^{B}(\hat{\theta}_i^* - \hat{\theta})^2$$

Bootstrap Estimates of Bias, Standard Deviation, and MSE



Bootstrap Estimates of Bias, Standard Deviation, and MSE

```
The data consistis of 120 MSOFT returns.
> n=120
> sample(MSOFT,n, replace=TRUE)
```

Suppose $\theta = \mu$.

```
 > B=500
 > muha.boot=rep(0,B)
> for (i in1:B) { data.boot= sample(MSOFT, n, replace=TRUE);
 muhat.boot[i]=mean(booot.data)}
```

Bootstrap bias:

```
> mean(muhat.boot)$-$ mean(MSOFT)
[1] -0.0006678712$

 bootstrap SE:
sd(muhat.boot)
0.01004047
analytics SE:
  > sd(MSOFT)/sqrt(n)
 [1] 0.01060189$
```

Confidence Intervals

- Besides its use in estimating bias and finding standard errors, the bootstrap is widely used to construct confidence intervals.
- When a confidence interval uses an approximation, there are two coverage probabilities
  1. the nominal one that is stated and
  2. the actual one that is unknown.
- Only for exact confidence intervals making no use of approximations will the two probabilities be equal.
- By the accuracy of a confidence interval, we mean the degree of agreement between the nominal and actual coverage probabilities.

Confidence Intervals

- Let $\hat{\theta}$ be an estimate of θ and let $s_{\text{boot}}(\hat{\theta})$ be the estimate of standard error. Then the normal theory confidence interval for θ is

$$\hat{\theta} \pm z_{\alpha/2} s_{\text{boot}}(\hat{\theta})$$

Confidence Intervals for the mean

- Suppose we wish to construct a confidence interval for the population mean.
- Start with the so-called t-statistic:

$$t = \frac{\mu - \overline{X}}{\frac{s}{\sqrt{n}}}$$

- The denominator of t $s/\sqrt{n}$, is just the standard error of the mean.

Confidence Intervals for the mean

- When sampling from a normal population, $t \sim t_{n-1}$.
- Denote by $t_{\alpha/2}$ is the $1 - \alpha/2$ quantile of this distribution, then

$$P(-t_{\alpha/2} \leq t \leq t_{\alpha/2}) = 1 - \alpha$$

-
- This implies that the probability is $1 - \alpha$ that

$$\bar{X} - t_{\alpha/2} s\sqrt{n} \leq \mu \leq \bar{X} + t_{\alpha/2} s/\sqrt{n}$$

- which shows that

$$\bar{X} \pm t_{\alpha/2} s/\sqrt{n}$$

  is a $1 - \alpha$ confidence interval for μ, assuming normally distributed data.

Confidence Intervals for the mean

- What happens if we are not sampling from a normal distribution?
- In that case, the distribution of t will no longer be the t distribution, but rather some other distribution that is not known to us.
- We have now two problems
  1. we don't know the population distribution.
  2. Even we were to know it, we still need to get the distribution of the t-statistic from the population distribution.

Confidence Intervals for the mean

- Considering these difficulties, can we get a confidence interval? The answer is Yes, by resampling.
- Take a large number, say B, resamples from the original sample.

Confidence Intervals for the mean

- Let $\overline{X}_{boot,b}$ and $s_{boot,b}$ be the sample mean and standard deviation of the bth resample. ?
- Define

$$t_{boot,b} = \frac{\overline{X} - \overline{X}_{boot,b}}{\frac{s_{boot,b}}{\sqrt{n}}}$$

- Notice that $t_{boot,b}$ is defined in the same way as t except for two changes.
  1. First, $\overline{X}$ and s in t are replaced by $\overline{X}_{boot,b}$ and $s_{boot,b}$ in $t_{boot,b}$.
  2. Second, μ in t is replaced by $\overline{X}$ in $t_{boot,b}$. (A resample is taken using the original sample as the population. Thus, for the resample, the population mean is $\overline{X}$)
- Resamples are independent. Therefore, $t_{boot,1}, t_{boot,2}, \ldots$ is a random sample from the t-statistic distribution.
- After B values of $t_{boot,b}$ have been calculated, we find the 2.5% and 97.5% percentiles (t_L and t_U) of this collection of $t_{boot,b}$ values.

Confidence Intervals for the mean

- Specifically, we find t_L and t_U as follows:
  1. The B values of $t_{boot,b}$ are sorted from the smallest to the largest
  2. We then calculate $B\alpha/2$ and round to the nearest integer. For example, if $\alpha = 0.05$ and $B = 1000$, then $B\alpha/2 = 25$. The 25th value of the sorted values of $t_{boot,b}$ is t_L. Similarly, calculate $B(1-\alpha/2)$ rounded to the nearest integer say k. Then t_U is the kth sorted value of $t_{boot,b}$.
- The bootstrap confidence interval is then give by

$$(\overline{X} + t_L \frac{s}{\sqrt{n}}, \overline{X} + t_U \frac{s}{\sqrt{n}})$$

Basic Bootstrap Interval

- Let q_L and q_U be the $\alpha/2$-lower and -upper sample quantiles of $\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$. The fraction of bootstrap estimates that satisfy

  $$q_L \leq \hat{\theta}_b^* \leq q_U$$

  is $1 - \alpha$.
- This equation is equivalent to

  $$\hat{\theta} - q_U \leq \hat{\theta} - \hat{\theta}_b^* \leq \hat{\theta} - q_L$$

  so that $\hat{\theta} - q_U$ and $\hat{\theta} - q_L$ are lower and upper quantiles for the distribution of $\hat{\theta} - \theta_b^*$.
- The basic bootstrap interval uses them as lower and upper quantiles of the distribution of $\theta - \hat{\theta}$
- Using the bootstrap approximation, it is assumed

  $$\hat{\theta} - q_U \leq \theta - \hat{\theta} \leq \hat{\theta} - q_L$$

  will occur in a fraction $1 - \alpha$ of samples. Adding $\hat{\theta}$ to all sides give

  $$2\hat{\theta} - q_U \leq \theta \leq 2\hat{\theta} - q_L$$

- The confidence interval is then

  $$(2\hat{\theta} - q_U, 2\hat{\theta} - q_L).$$

Basic Bootstrap Interval

```
>  2*mean(MSOFT)-quantile(muhat.boot, 0.975)

0.005016204

> 2*mean(MSOFT)-quantile(muhat.boot, 0.025)
0.04370343
```

The basic confidence interval for μ is $(0.005016204, 0.04370343)$.

Value at Risk (VaR): Normal case

Recall

$$\text{VaR}(\alpha) = -S_0(\mu + z_\alpha \sigma)$$

```
> b=500; n=120; S0=1000000; var.boot=rep(0,b)

> for (i in 1:b){data.boot=sample(MSOFT,n, replace=TRUE)
var.boot[i]= -S0*(mean(data.boot)+qnorm(0.05)*sd(data.boot))}

> var.MSOF= -S0*(mean(MSOFT)+qnorm(0.05)*sd(MSOFT))

> 167163.8

The basic 95% confidence interval for VaR(0.05) is

(2*mean(MSOFT)-quantile(var.boot, 0.975), 2*mean(MSOFT)-quantile(var.boot,
0.025))
This gives (136542.7,201317.7)
```