Aula 10

Bootstrap

Claudio R. Lucinda

FEA/USP

Agenda

1. Bootstrap

Agenda

1. Bootstrap
2. Bias Reduction

Agenda

1. Bootstrap
2. Bias Reduction
3. Jackknife

Agenda

Bootstrap

Bootstrap. O uso do termo bootstrap deriva da frase para puxar-se para cima por meio de bootstraps, amplamente considerada baseada em uma das "As Surpreendentes Aventuras do Barão de Munchausen" do século XVIII, de Rudolph Erich Raspe: O Barão havia caído no fundo do poço um lago profundo. Justamente quando parecia que tudo estava perdido, ele pensou em se recompor por conta própria.

Bootstrap- Ideia geral

Seja $T(\cdot)$ um funcional de interesse, por exemplo estimador de um parâmetro. Estamos interessados na estimativa de $T(F)$, onde F é a distribuição da população. Seja F_n uma distribuição empírica baseada na amostra $x = (x_1, \ldots, x_n)$. Inicialização:

1. gera uma amostra $x^* = (x_1^*, \ldots, x_n^*)$ com substituição da distribuição empírica F_n para os dados (amostra boostrap);
2. calcula $T(F_n^*)$ a estimativa bootstrap de $T(F)$. Esta é uma substituição da amostra original x por uma amostra bootstrap x^* e a estimativa bootstrap de $T(F)$ no lugar da estimativa amostral de $T(F)$;
3. M vezes repita as etapas 1 e 2 onde M é grande, digamos 100000 .

Bootstrap

Agora, uma coisa muito importante a lembrar é que, com a aproximação de Monte Carlo para o bootstrap, existem duas fontes de erro:

1. a aproximação de Monte Carlo para a distribuição de bootstrap, que pode ser tão pequena quanto você quiser tornando M grande;
2. a aproximação da distribuição bootstrap F_n^* à distribuição populacional F.

Se $T(F_n^*)$ converge para $T(F)$ como $n \to \infty$, então o bootstrap funciona.

Bootstrap- R code

"Uma função nos pacotes *R* básicos que está no centro da reamostragem é a função sample(), cuja sintaxe é

```
sample(x, size, replace= FALSE, prob=NULL)
```

O primeiro argumento x é o vetor de dados, ou seja, a amostra original. size é o tamanho da reamostragem desejada. replace é TRUE se a reamostragem for com substituição e FALSE se não (o padrão). prob é um vetor de pesos de probabilidade se o padrão equalweight não for usado. Quaisquer argumentos omitidos assumirão o padrão. Se o tamanho for omitido, o padrão será o comprimento de x."

Bootstrap- R code

"Para nossos propósitos, geralmente será mais fácil reamostrar os índices dos dados de uma amostra de tamanho *n*, em vez dos próprios dados. Por exemplo, se tivermos cinco dados em nosso conjunto, digamos

```
> x=c(-0.3, 0.5, 2.6, 1.0, -0.9)
> x
[1] -0.3 0.5 2.6 1.0 -0.9
then
> i = sample(1:5, 5, replace=TRUE)
> i
[1] 3 2 3 2 2
> x[i]
[1] 2.6 0.5 2.6 0.5 0.5
```

Bootstrap- Erro-Padrão

A partir da amostragem bootstrap, podemos estimar qualquer aspecto da distribuição de $\hat{\theta} = s(y)$ (que é qualquer quantidade calculada a partir dos dados $y = (y_1, \ldots, y_n)$, por exemplo, seu erro padrão é

$$\text{s.e.b. } (\widehat{\theta}) = \left(\frac{1}{B-1} \sum_{b=1}^{B} \left(\widehat{\theta}^*(b) - \widehat{\theta}^*(\cdot) \right)^2 \right)^{1/2}$$

onde $\hat{\theta}^*(b)$ é a replicação bootstrap de $s(y)$ e

$$\widehat{\theta}^*(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \widehat{\theta}^*(b).$$

Redução de Viés

Seja $\hat{\theta}$ um estimador consistente, mas viesado. Alvo: reduzir o viés do estimador. O viés de $\hat{\theta}$ é o viés de erro sistemático $= \mathbb{E}_F\widehat{\theta} - \theta$. Em geral o bias depende do parâmetro desconhecido θ, por isso não podemos ter $\widehat{\theta}-$ bias. Considere a seguinte correção de viés de bootstrap

$$\widehat{\theta}_B = \widehat{\theta} - \widehat{bias}$$

em que

$$\widehat{bias} = \hat{\mathbb{E}}_F^{\hat{\theta}} - \widehat{\theta} = \widehat{\theta}^*_{(\cdot)} - \widehat{\theta}$$

em que $\hat{\theta}^*_{(\cdot)}$ é a média das estimativas de bootstrap

$$\widehat{\theta}^*_{(\cdot)} = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^*_b.$$

Portanto

$$\widehat{\theta}_B = \widehat{\theta} - \widehat{bias} = 2\widehat{\theta} - \hat{\theta}^*_{(\cdot)}$$

Exemplo

```
theta=6
n=15
set.seed(123)
Data=theta*runif(n)
MLE=max(Data)
B=1000
for (i in 1:B){
    j=sample(1:15,15, replace=TRUE)
    T[i]=max(Data[j])
}
2*MLE-mean(T)
[1] 5.8199
MLE
[1] 5.741
```

Jackknife

Em certo sentido, o método bootstrap é uma generalização do método jackknife, no sentido de que a reamostragem é feita aleatoriamente e não de forma determinística como no jackknife "leave-one-out".

1. Temos uma amostra $y = (y_1, \ldots, y_n)$ e um estimador $\widehat{\theta} = s(y)$.
2. Alvo: estima o viés e o erro padrão do estimador.
3. As amostras de observação "leave-one-out"

$$y_{(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)\,,$$

para $i = 1, \ldots, n$ são chamados de amostras de jackknife.
Os estimadores de Jackknife são $\hat{\theta}_{(i)} = s\left(y_{(i)}\right)$.

Redução do Viés com Jackknife

O viés de $\widehat{\theta} = s(y)$ é definido como

$$\text{bias}_J(\hat{\theta}) = (n-1)\left(\widehat{\theta}_{(\cdot)} - \widehat{\theta}\right),$$

onde $\widehat{\theta}_{(\cdot)}$ é a média dos estimadores Jackknife $\widehat{\theta}_{(i)}$

$$\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}.$$

Isso leva a um estimador jackknife com viés reduzido do parâmetro θ

$$\widehat{\theta}_J = \widehat{\theta} - \text{bias}_J(\widehat{\theta}) = n\widehat{\theta} - (n-1)\hat{\theta}_{(\cdot)}$$

Exemplo

```
> theta=6
> n=15
> set.seed(123)
> Data=theta*runif(n)
> Data
 [1] 1.7254651 4.7298308 2.4538615 5.2981044 5.6428037 0.2733390
3.1686329 5.3545143 3.3086101 2.7396884
 [11] 5.7410001 2.7200049 4.0654238 3.4358004 0.6175481
```

Exemplo – Continuação

O valor máximo é 5,7410001 e o segundo valor máximo é 5,6428037.
A média dos estimadores Jackknife $\hat{\theta}_{(i)}$

$$\widehat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta}_{(i)} = \frac{5.6428037 + 14 \cdot 5.7410001}{15} = 5.734454$$

O estimador jackknife com viés reduzido do parâmetro θ

$$\begin{aligned}\widehat{\theta}_{J} &= n\widehat{\theta} - (n-1)\widehat{\theta}_{(\cdot)} \\ &= 15 \cdot 5.7410001 - 14 \cdot 5.734454 = 5.832645.\end{aligned}$$

O estimador bootstrap com viés reduzido do parâmetro θ foi 5,815999.

Bootstrap e Testes de Hipótese

- Defina as duas hipóteses.
- Escolha uma estatística de teste T que possa discriminar entre as duas hipóteses. Não nos importamos que nossa estatística tenha uma distribuição conhecida sob a hipótese nula.
- Calcula o valor observado t_{obs} da estatística para a amostra.
- Gera B amostras da distribuição implícita pela hipótese nula.
- Para cada amostra calcule o valor $t_{(i)}$ da estatística, $i = 1, \ldots, B$.
- Encontre a proporção de vezes que os valores amostrados são mais extremos do que os observados.
- Aceite ou rejeite de acordo com o nível de significância.

Bootstrap Tests

Suponha duas amostras $x = (x_1, \ldots, x_n)$ e $y = (y_1, \ldots, y_m)$. Queremos testar a hipótese de que as médias de duas populações são iguais, ou seja,

$$H : \mu_x = \mu_y \quad \text{vs} \quad A : \mu_x \neq \mu_y$$

Use como uma estatística de teste $T = \bar{x} - \bar{y}$.
Sob a hipótese nula, uma boa estimativa da distribuição da população é a amostra combinada $z = (x_1, \ldots, x_n, y_1, \ldots, y_m)$
Para cada amostra de bootstrap, calcule $T^*_{(i)}, i = 1, \ldots, B$.
Estime o valor-p do teste como

$$\widehat{p} = \frac{1}{B} \sum_{i=1}^{B} \mathbb{1}\left(T^*_{(i)} \geq t_{obs}\right) \text{ ou } \tilde{p} = \frac{1}{B+1}\left(1 + \sum_{i=1}^{B} \mathbf{1}\left(T^*_{(i)} \geq t_{obs}\right)\right).$$

Outras estatísticas de teste são aplicáveis, como por exemplo t-statistics.