Aula 8

Métodos de Reamostragem

Rafaela Medeiros

Curso de Estatística Básica
WiMLDS

26 de Setembro de 2020

O que são, quais são, como aplicar e/ou analisar e que perguntas os métodos de reamostragem nos respondem?

O que são, quais são, como aplicar e/ou analisar e que perguntas os métodos de reamostragem nos respondem?

O que você vai aprender hoje

1. Definição de reamostragem.
2. Tipos de técnicas de reamostragem.
   - Validação Cruzada.
   - Bootstrapping.
3. Aplicações em R

`https://github.com/aga-rafa/ReamostragemWiMLDS`

helloanamedeiros@gmail.com

Conceito

“Eu vou samplear, eu vou te roubar”

Conceito

- As técnicas de reamostragem são um conjunto de métodos que têm como fundamento dois passos:
  1. seleção aleatória de observações de um banco de dados (ou conjunto de treino);
  2. aplicação de um modelo de interesse em cada conjunto amostral.
- Permite obter informações que não seriam observadas a partir de modelos tradicionais. Exemplo: variação de coeficientes.
- São imprescindíveis na estatística moderna, graças ao desenvolvimento da tecnologia. Formam uma base para modelos de aprendizagem mais avançados.

Indicações

- N $\downarrow$;
- Poder $\downarrow$;
- Violação de pressupostos;
- Calcular erros de teste;
- Comparação entre modelos distintos.
- Métodos não-paramétricos: dispensam pressupostos rígidos sobre a distribuição das variáveis na população real e suas relações entre si.

A reamostragem pode ser utilizada para calcular diversas estimativas, como taxas de erro de teste, viés, coeficientes, intervalos de confiança, desvio-padrão, entre outras.

Precauções

Métodos de reamostragem fazem um ou outra das seguintes suposições:

- A função de densidade empiricamente observada é uma boa aproximação da função populacional para as variáveis de interesse.
- A distribuição dos parâmetros calculados por reamostragem, em comparação à distribuição amostral, é similar à relação que a distribuição amostral mantém com a populacional.

Precauções

Métodos de reamostragem fazem um ou outra das seguintes suposições:

- A função de densidade empiricamente observada é uma boa aproximação da função populacional para as variáveis de interesse.
- A distribuição dos parâmetros calculados por reamostragem, em comparação à distribuição amostral, é similar à relação que a distribuição amostral mantém com a populacional.

Em ambos, o problema de generalização inferencial se reduz a aprender sobre a distribuição dos parâmetros na reamostragem em si.

Precauções

Métodos de reamostragem fazem um ou outra das seguintes suposições:

- A função de densidade empiricamente observada é uma boa aproximação da função populacional para as variáveis de interesse.
- A distribuição dos parâmetros calculados por reamostragem, em comparação à distribuição amostral, é similar à relação que a distribuição amostral mantém com a populacional.

Em ambos, o problema de generalização inferencial se reduz a aprender sobre a distribuição dos parâmetros na reamostragem em si.

ALERTA

No bootstrap, a amostragem deve ser sempre feita COM substituição, garantindo a independência entre as observações.

Tipos de Reamostragem

Validação Cruzada

Conj. Treino

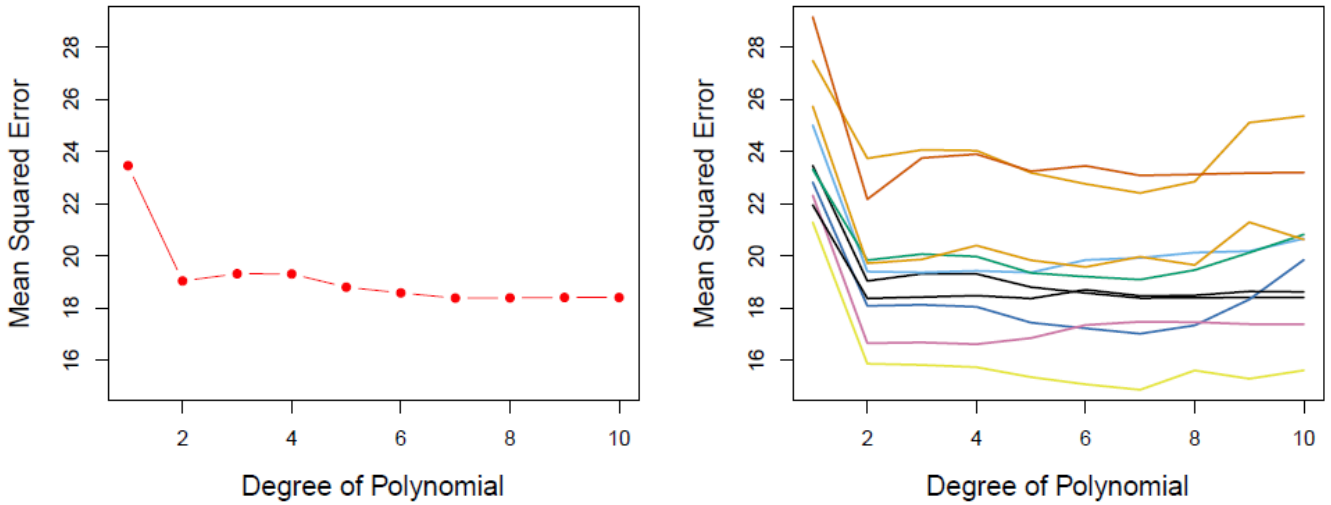
Conj. Teste

Bootstrap

Método de Validação

Uma classe de métodos que divide a amostra aleatoriamente entre conjunto de treino e teste. O primeiro é usado para calcular o modelo; o segundo, para comparar a precisão dos resultados previstos a partir do modelo de treino.

A validação cruzada pode ser usada para selecionar o nível apropriado de flexibilidade de um modelo *(model selection, esq. na figura abaixo)* ou estimar o viés de um modelo e avaliar sua performance *(model assessment, direita abaixo)*.



Validação - Aplicações em R

1. Selecionar uma amostra aleatória como conjunto de treino:
   $train = sample(392, 196)$
2. Aplicar o modelo (p.e.:regresão linear)
   $lm.fit = lm(mpg\ \ horsepower, data = Auto, subset = train)$
3. Estimar o viés (MSE)
   $mean((mpg - predict(lm.fit, Auto))[-train]^2)$
4. Comparar os acima resultados com modelos polinomiais
   $lm.fit2 = lm(mpg\ \ poly(horsepower, 2), data = Auto, subset = train)$
   $mean((mpg - predict(lm.fit2, Auto))[-train]^2)$
   $lm.fit3 = lm(mpg\ \ poly(horsepower, 3), data = Auto, subset = train)$
   $mean((mpg - predict(lm.fit3, Auto))[-train]^2)$

LOOCV

Problemas

1. O viés pode ser muito variável, dependendo das observações contidas no conjunto de treino;
2. A composição do conjunto de teste pode inflar o viés do modelo para o conjunto do banco de dados.

LOOCV: Também chamado de método de Jack-knife, o *Leave-one-out cross validation* é semelhante ao método de validação anterior, mas apenas 01 observação é selecionada para o conjunto de teste. Isso reduz os problemas de enviesamento da amostra e diminui a variabilidade dos paramêtros calculados.

LOOCV

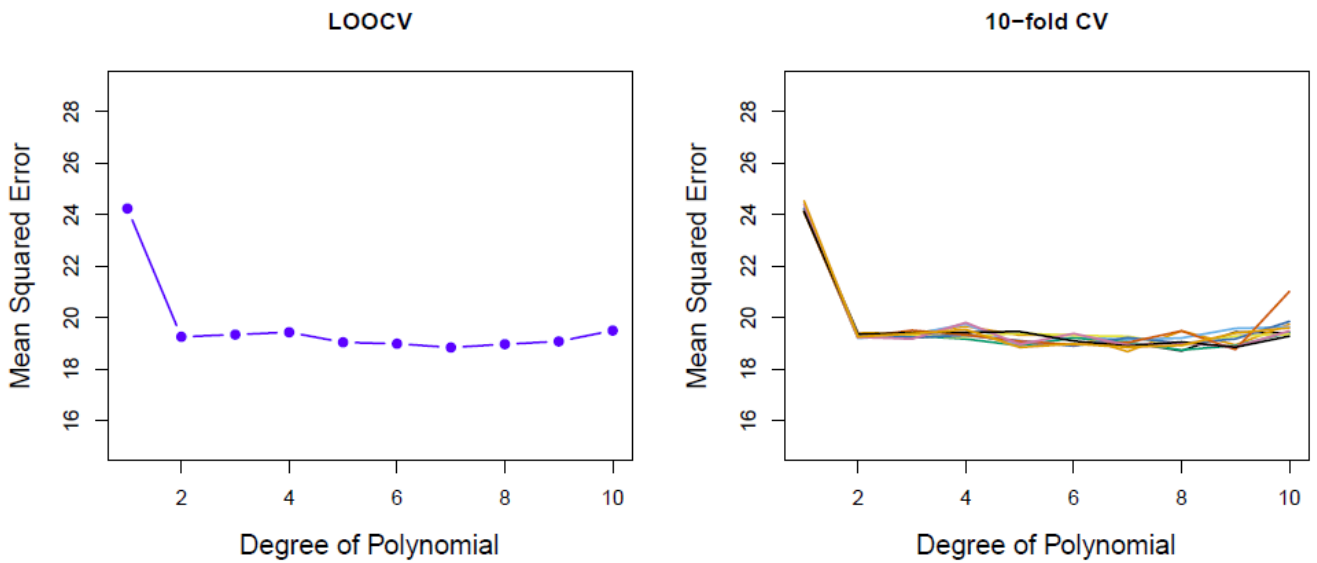
Com o LOOCV, o processo de divisão, modelagem e comparação com a observação de treino é repetido *n* vezes, de modo a garantir uma maior confiabilidade. A estimativa para o MSE então torna-se:

$$\mathrm{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{MSE}_i .$$

LOOCV

Com o LOOCV, o processo de divisão, modelagem e comparação com a observação de treino é repetido *n* vezes, de modo a garantir uma maior confiabilidade. A estimativa para o MSE então torna-se:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i.$$

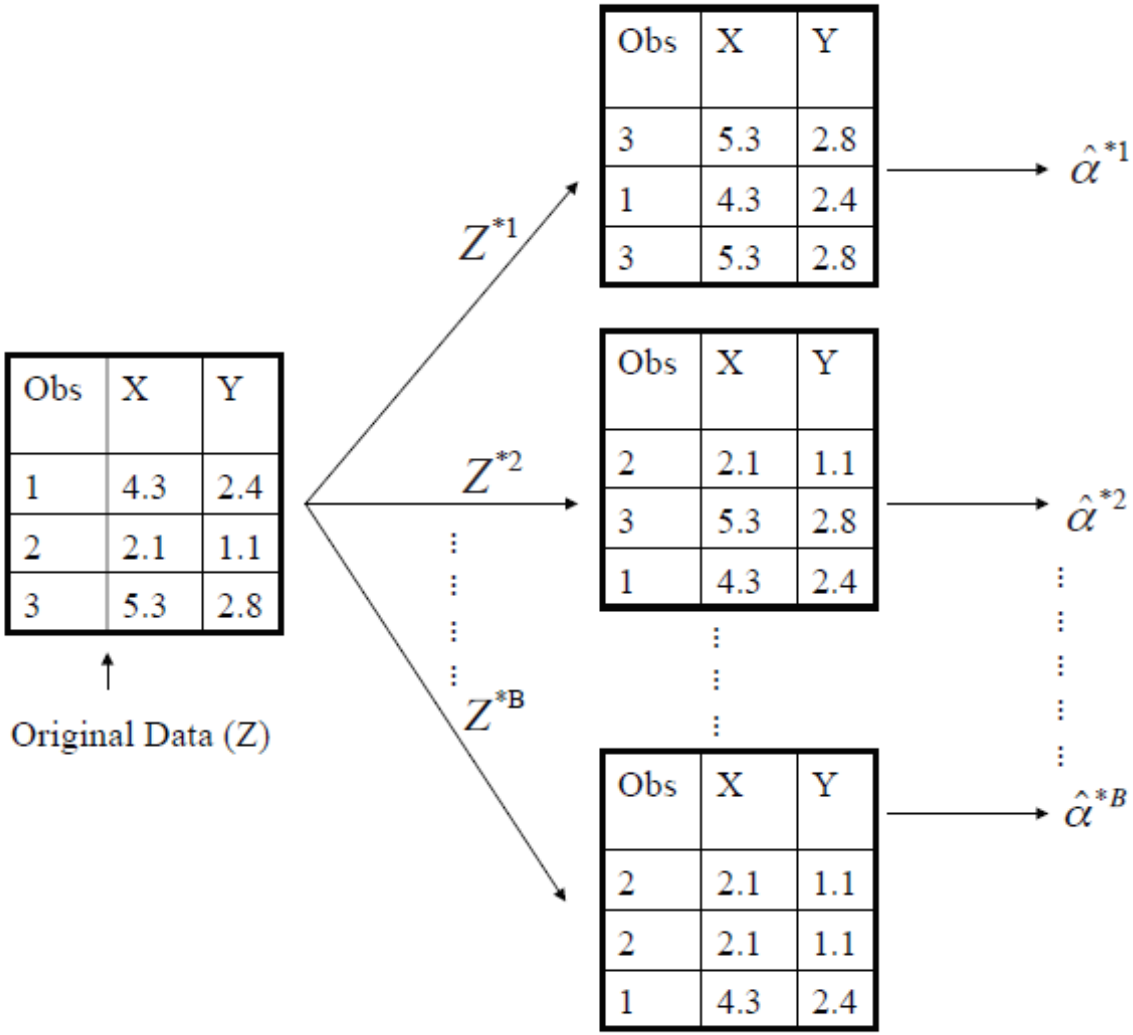


Bootstrap

O bootstrap é uma técnica na qual as observações de um banco de dados são "sampleadas"inúmeras vezes, com igual probabilidade de seleção individual. O modelo é então aplicado a cada amostra individualmente, de modo que a precisão de um parâmetro estimado ou de um modelo específico pode ser acessada.

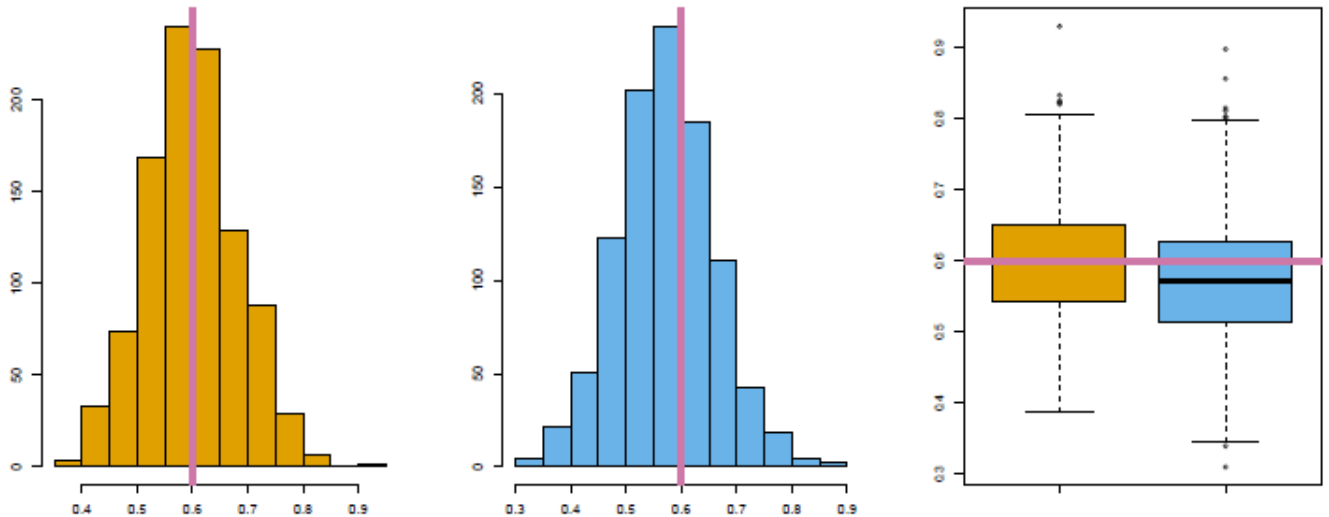
É muito utilizada para estimar parâmetros que não seguem uma distribuição normal, que têm propriedades estatísticas desconhecidas, ou que não possuem uma fórmula especificada (exemplo: intervalo de confiança para R^2)

Bootstrap - Ilustração



Bootstrap

A técnica permite utilizar um algoritmo para emular o processo de obtenção de novas amostras da população real. Ao invés de obter repetidas amostras reais, obtém-se distintos conjuntos de dados ao criar amostras repetidas das observações no próprio banco de dados original. Na figura abaixo, compara-se a variância para diversas amostras de uma população (esquerda) com a variância de amostras geradas por bootstrap apartir de uma única amostra.



Bootstrap - Cont.

A fórmula abaixo calcula o erro padrão para as estimativas bootstrap:

$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2}.$$

Onde: B é o número de amostras bootstrapped e alpha a variância estimada a partir de cada amostra.

Aplicações em R

1. Criar uma função alpha para estimar alpha:
   $alpha.fn = function(data, index)\{$
   $X = data\$X[index]$
   $Y = data\$Y[index]$
   $return((var(Y) - cov(X,Y))/(var(X) + var(Y) - 2 * cov(X,Y)))\}$
2. Selecionar aleatoriamente 100 observações: $set.seed(1)$
   $alpha.fn(Portfolio, sample(100, 100, replace = T))$
3. Repetir esse processo com n = 1000.
   $boot(Portfolio, alpha.fn, R = 1000)$

```
> boot(Portfolio,alpha.fn,R=1000)

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = Portfolio, statistic = alpha.fn, R = 1000)

Bootstrap Statistics :
     original       bias      std. error
t1*  0.5758      -7.315e-05   0.0886
```