Lista 5

Técnicas Computacionais em Estatística

Tailine J. S. Nonato

June 3, 2025

Lista 5 - Reamostragem

Exercício 1

Considere o conjunto de dados de ar condicionado (`aircondit`) disponível no pacote `boot` do R. São 12 observações dos tempos (em horas) entre falhas do equipamento de ar condicionado:

3, 5, 7, 18, 43, 85, 91, 98, 100, 130, 230, 487.

Suponha que os tempos entre falhas sigam um modelo exponencial $Exp(\lambda)$. Calcule intervalos de confiança de bootstrap de 95% para o tempo médio entre falhas $1/\lambda$ pelos métodos normal, básico, percentil e BCa. Considere o estimador de máxima verossimilhança $\hat{\lambda}$ para o problema.

Resolução

```r
pacman::p_load(boot, alr4, ggplot2, knitr)
```

```r
data(aircondit)
aircondit <- aircondit$hours
```

Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória de tamanho n de uma distribuição exponencial com parâmetro λ. O estimador de máxima verossimilhança para λ é dado por:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i}$$

A média da distribuição exponencial é dada por $E[X] = \theta = \frac{1}{\lambda}$, logo o estimador de máxima verossimilhança para a média é:

$$\hat{\theta} = \frac{1}{\hat{\lambda}} = \frac{\sum_{i=1}^{n} X_i}{n}$$

```r
theta_hat <- function(data, i) {
  amostra <- data[i]
  return(mean(amostra))
}
```

Assim,

```r
set.seed(451)
boot_aircondit <- boot(data = aircondit, statistic = theta_hat, R = 1000)

# Intervalos de confiança
boot.ci(boot_aircondit, type = "norm")
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot_aircondit, type = "norm")

Intervals :
Level      Normal
95%   ( 31.7, 181.9 )
Calculations and Intervals on Original Scale
```

```r
boot.ci(boot_aircondit, type = "basic")
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot_aircondit, type = "basic")

Intervals :
Level      Basic
95%   ( 23.1, 168.7 )
Calculations and Intervals on Original Scale
```

```r
boot.ci(boot_aircondit, type = "perc")
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot_aircondit, type = "perc")

Intervals :
Level     Percentile
95%   ( 47.4, 193.1 )
Calculations and Intervals on Original Scale
```

```r
boot.ci(boot_aircondit, type = "bca")
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = boot_aircondit, type = "bca")

Intervals :
Level       BCa
95%   ( 57.1, 222.8 )
Calculations and Intervals on Original Scale
Some BCa intervals may be unstable
```

Exercício 2

Considere novamente os dados sobre consumo de combustível nos 50 estados + DC dos Estados Unidos (Federal Highway Administration, 2001). Os dados foram coletados pela Federal Highway Administration dos EUA; ver Weisberg, S. (2014). *Applied Linear Regression*. John Wiley & Sons, Hoboken, New Jersey, fourth edition.

Temos as seguintes variáveis:

- **Drivers**: Número de motoristas com carteira de habilitação no estado.
- **FuelC**: Gasolina vendida para uso rodoviário, milhares de galões.
- **Income**: Renda pessoal por pessoa no ano 2000 em dólares.
- **Miles**: Milhas de rodovias com ajuda federal no estado.
- **Pop**: População de 2001 com 16 anos ou mais.

- **Tax**: Taxa de imposto estadual da gasolina, centavos por galão.
- **MPC**: Milhas estimadas percorridas per capita.

Considere ainda as variáveis derivadas:

- $Fuel = 1000 \times \frac{FuelC}{Pop}$
- $Dlic = 1000 \times \frac{Drivers}{Pop}$
- $lMiles = \log(Miles)$

Podemos tentar entender como o combustível ($Fuel$) está relacionado aos preditores (Tax, $Dlic$, $Income$ e $lMiles$), usando um modelo de regressão linear múltipla.

```r
data(fuel2001)
fuel2001$lMiles <- log(fuel2001$Miles)
fuel2001$Dlic   <- 1000*(fuel2001$Drivers/fuel2001$Pop)
fuel2001$Fuel   <- 1000*(fuel2001$FuelC/fuel2001$Pop)
fuel2001$Income <- fuel2001$Income/1000
```

Considere os seguintes modelos:

$$
\begin{aligned}
\text{Modelo 1:} &\quad \text{Fuel} = \beta_0 + \beta_1\,\text{Tax} + \varepsilon \\
\text{Modelo 2:} &\quad \text{Fuel} = \beta_0 + \beta_1\,\text{Tax} + \beta_2\,\text{Dlic} + \varepsilon \\
\text{Modelo 3:} &\quad \text{Fuel} = \beta_0 + \beta_1\,\text{Tax} + \beta_2\,\text{Dlic} + \beta_3\,\text{Income} + \varepsilon \\
\text{Modelo 4:} &\quad \text{Fuel} = \beta_0 + \beta_1\,\text{Tax} + \beta_2\,\text{Dlic} + \beta_3\,\text{Income} + \beta_4\,l\text{Miles} + \varepsilon
\end{aligned}
$$

Utilize o procedimento leave-one-out para estimar os erros das predições e escolher o melhor dos modelos acima. Como critérios, use o RMSE e MAE, dados por:

$$
\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}, \qquad \text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|
$$

```r
loo <- function(formula, data) {
  n <- nrow(data)
  sq_errors <- abs_errors <- numeric(n)

  for (i in 1:n) {
    train <- data[-i, ]
    test  <- data[i, ]

    model <- lm(formula, data = train)
    pred  <- predict(model, newdata = test)
```

```r
    sq_errors[i]  <- (test$Fuel - pred)^2
    abs_errors[i] <- abs(test$Fuel - pred)
  }

  rmse <- sqrt(mean(sq_errors))
  mae  <- mean(abs_errors)
  return(c(RMSE = rmse, MAE = mae))
}
```

```r
results <- list(
  mod1 = loo(Fuel ~ Tax, data = fuel2001),
  mod2 = loo(Fuel ~ Tax + Dlic, data = fuel2001),
  mod3 = loo(Fuel ~ Tax + Dlic + Income, data = fuel2001),
  mod4 = loo(Fuel ~ Tax + Dlic + Income + lMiles, data = fuel2001))

results_df <- do.call(rbind, results)
results_df <- as.data.frame(results_df)
kable(results_df)
```

	RMSE	MAE
mod1	89.24353	64.21652
mod2	84.73785	62.30178
mod3	75.85411	53.65871
mod4	71.60401	52.60657

Como podemos observar, o modelo 4 apresenta o menor RMSE e MAE, indicando que é o melhor modelo entre os quatro considerados, ou seja, a inclusão das variáveis `Dlic`, `Income` e `lMiles` melhora a capacidade preditiva do modelo.