Inferencia utilizando técnicas de remuestreo: el Bootstrap

Contents

1 Introducción

El bootstrap es un método para estimar varianzas de estadísticos e intervalos de confianza mediante el remuestreo de los datos disponibles.

1.1 Estimación de la varianza con bootstrap

Sea $\{X_1, X_2, \cdots, X_n\}$ una muestra aleatoria simple (luego los datos son independientes y con igual distribución). Y sea $T = f(X_1, X_2, \cdots, X_n)$ un estadístico, es decir, T es cualquier función de los datos. Para calcular la varianza del estimador, $Var(T)$, el método bootstrap consiste en:

- Paso 1: generar, mediante simulación, una muestra con reemplazamiento a partir de $\{X_1, X_2, \cdots, X_n\}$ que llamaremos $\{X_1^*, X_2^*, \cdots, X_n^*\}$.
- Paso 2: Calcular la estimación de T a partir de la muestra bootstrap: $T^* = f(X_1^*, X_2^*, \cdots, X_n^*)$.
- Paso 3: Repetir los pasos 1 y 2 un total de B veces, obteniendo $T_1^*, T_2^*, \ldots, T_B^*$.
- Paso 4: estimar la varianza de T mediante la varianza de $T_1^*, T_2^*, \ldots, T_B^*$.

2 Estimación de intervalos de confianza utilizando bootstrap

Hay varios métodos para calcular el intervalo de confianza de un parámetro θ con bootstrap. Nosotros vamos a utilizar el método de los percentiles:

- Paso 1: generar, mediante simulación, una muestra con reemplazamiento a partir de $\{X_1, X_2, \cdots, X_n\}$ que llamaremos $\{X_1^*, X_2^*, \cdots, X_n^*\}$.
- Paso 2: Calcular la estimación de θ a partir de la muestra bootstrap: $\theta^* = T(X_1^*, X_2^*, \cdots, X_n^*)$.
- Paso 3: Repetir los pasos 1 y 2 un total de B veces, obteniendo $\theta_1^*, \theta_2^*, \ldots, \theta_B^*$.
- Paso 4: estimar el intervalo de θ mediante $(\theta^*_{\alpha/2}, \theta^*_{1-\alpha/2})$.

3 Bootstrap en el modelo de regresión

En un problema de regresión la muestra consiste en:

$$(y_1, x_{11}, x_{21}, \cdots, x_{k1})$$

$$(y_2, x_{12}, x_{22}, \cdots, x_{k2})$$

...

$$(y_n, x_{1n}, x_{2n}, \cdots, x_{kn})$$

Para el modelo de regresión el método bootstrap consiste en:

- Paso 1: generar una muestra con reemplazamiento de los pares de datos que llamaremos $\{(y_1^*, x_{11}^*, x_{21}^*, \ldots, x_{k1}^*), (y_2^*, x_{12}^*, x_{22}^*, \ldots, x_{k2}^*), \cdots, (y_n^*, x_{1n}^*, x_{2n}^*, \ldots, x_{kn}^*)\}$.
- Paso 2: estimar los parámetros del modelo a partir de la muestra bootstrap $y^* = X^* \hat{\beta}^* + e^*$.
- Paso 3: Repetir los pasos 1 y 2 un total de B veces, obteniendo $\beta_1^*, \beta_2^*, \ldots, \beta_B^*$.
- Paso 4: calcular la varianza de los estimadores o los intervalos de confianza de los parámetros a partir de los valores calculados en el paso 3.

Por ejemplo, vamos a calcular la varianza de los estimadores y los intervalos de confianza para el modelo:

```
load("datos/kidiq.Rdata")
str(d)
```

```
## 'data.frame':    434 obs. of  5 variables:
##  $ kid_score: int  65 98 85 83 115 98 69 106 102 95 ...
##  $ mom_hs   : Factor w/ 2 levels "no","si": 2 2 2 2 2 1 2 2 2 2 ...
##  $ mom_iq   : num  121.1 89.4 115.4 99.4 92.7 ...
##  $ mom_work : Factor w/ 4 levels "notrabaja","trabaja23",..: 4 4 4 3 4 1 4 3 1 1 ...
##  $ mom_age  : int  27 25 27 25 27 18 20 23 24 19 ...
```

```
# estimamos los parametros del modelo
m1 = lm(kid_score ~ mom_iq + mom_hs, data = d)
beta_e = coef(m1)
```

```
# BOOTSTRAP
# muestreamos los datos CON REPOSICION
n = nrow(d)
B = 1000
beta_e_b = matrix(0, nrow = B, ncol = 3)
for (i in 1:B){
  pos = sample(1:n, rep = T)
  db = d[pos,]
  mb = lm(kid_score ~ mom_iq + mom_hs, data = db)
  beta_e_b[i,] = coef(mb)
}
```

- Varianza de los parémetros estimados

```
# aplicando la teoría
diag(vcov(m1))
```

```
##  (Intercept)       mom_iq     mom_hssi
## 34.518069224  0.003669219  4.892113136
```

```
# aplicando bootstrap
apply(beta_e_b, 2, var)
```

```
## [1] 35.031793928  0.003595535  5.928901043
```

- intervalos de confianza

```
# aplicando la teoría
confint(m1)
```

```
##                  2.5 %     97.5 %
## (Intercept) 14.1839148 37.2791615
## mom_iq       0.4448487  0.6829634
## mom_hssi     1.6028370 10.2973969
```

```
# aplicando bootstrap
t(apply(beta_e_b, 2, quantile, probs = c(0.025,0.975)))
```

```
##            2.5%      97.5%
## [1,] 15.0016393 37.4631291
## [2,]  0.4474774  0.6773451
## [3,]  1.5469493 10.7665515
```