Bootstrap en el modelo de regresión logística

Contents

1 Bootstrap para regresión

Cuando hablamos de problemas de regresión, los datos adoptan la siguiente forma:

$$\begin{array}{ccccc} y_1 & x_{11} & x_{21} & \cdots & x_{k1} \\ y_2 & x_{12} & x_{22} & \cdots & x_{k2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ y_n & x_{11} & x_{2n} & \cdots & x_{kn} \end{array}$$

que también se pueden representar como $(y_1, x_1), (y_2, x_2), \cdots, (y_n, x_n)$, donde $x_i = \{x_{1i}, x_{1i}, \cdots, x_{1i}\}$. Podemos generar B muestras bootstrap diferentes de dos maneras:

- bootstrap empírico: Tratamos a cada par (y_i, x_i) como un objeto y remuetreamos con reemplazamiento en el conjunto $(y_1, x_1), (y_2, x_2), \cdots, (y_n, x_n)$ para obtener las B réplicas bootstrap.

$$\begin{array}{ccccc} (y_1^{*(1)}, x_1^{*(1)}), & (y_2^{*(1)}, x_2^{*(1)}), & \cdots, & (y_n^{*(1)}, x_n^{*(1)}) \\ (y_1^{*(2)}, x_1^{*(2)}), & (y_2^{*(2)}, x_2^{*(2)}), & \cdots, & (y_n^{*(2)}, x_n^{*(2)}) \\ & \vdots & & \\ (y_1^{*(B)}, x_1^{*(B)}), & (y_2^{*(B)}, x_2^{*(B)}), & \cdots, & (y_n^{*(B)}, x_n^{*(B)}) \end{array}$$

- bootstrap de los residuos: se estima el modelo y se calculan los residuos $e_i = y_i - \hat{y}_i$. A continuación se remuestrean los residuos B veces con reemplazamiento:

$$\begin{array}{cccc} e_1^{*(1)}, & e_2^{*(1)}, & \cdots, & e_n^{*(1)} \\ e_1^{*(2)}, & e_2^{*(2)}, & \cdots, & e_n^{*(2)} \\ & \vdots & & \\ e_1^{*(B)}, & e_2^{*(B)}, & \cdots, & e_n^{*(B)} \end{array}$$

Finalmente se obtienen las nuevas muestras bootstrap mediante $y_i^{*(b)} = \hat{y}_i + e_i^{*(b)}, \quad b = 1, \cdots, B.$

En el modelo de regresión lineal se prefiere utilizar el bootstrap de los residuos, ya que se elimina el efecto de atípicos, puntos influyentes, ... sin embargo, en regresión logística los residuos son del tipo {0,1}, por lo que se tiene que aplicar el bootstrap empírico.

2 Bootstrap en regresión logística

Vamos a calcular, utilizando bootstrap, el standard error y los intervalos de confianza para los parámetros del modelo:

```
d = read.csv("datos/MichelinNY.csv")
str(d)
```

```
## 'data.frame':    164 obs. of  6 variables:
##  $ InMichelin     : int  0 0 0 1 0 0 1 1 1 0 ...
##  $ Restaurant.Name: chr  "14 Wall Street" "212" "26 Seats" "44" ...
##  $ Food           : int  19 17 23 19 23 18 24 23 27 20 ...
##  $ Decor          : int  20 17 17 23 12 17 21 22 27 17 ...
##  $ Service        : int  19 16 21 16 19 17 22 21 27 19 ...
##  $ Price          : int  50 43 35 52 24 36 51 61 179 42 ...
```

Bootstap:

```
set.seed(99)
B = 500
n = nrow(d)
beta_e = matrix(0, nrow = B, ncol = 5)
for (b in 1:B){
  pos_b = sample(1:n, n, replace = T)
  d_b = d[pos_b,]
  m_b = glm(InMichelin ~ Food + Decor + Service + Price, data = d_b, family = binomial)
  beta_e[b,] = coef(m_b)
}
```

- Standard errors calculados con bootstrap:

```
apply(beta_e,2,sd)
```

```
## [1] 3.20370046 0.14221078 0.10042618 0.13469161 0.04590046
```

- Invervalos de confianza calculados con bootstrap:

```
alfa = 0.05
apply(beta_e,2,quantile, probs = c(alfa/2,1-alfa/2))
```

```
##             [,1]      [,2]        [,3]        [,4]       [,5]
## 2.5%  -19.322348 0.1867948 -0.08561463 -0.50060459 0.02105848
## 97.5%  -6.269294 0.7423821  0.30084908  0.04150634 0.20071627
```

Se puede comprobar que los resultados de bootstrap concuerdan con los obtenidos mediante la distribuciones asintóticas.