Ejercicios de Bono

Debbie Echanique, Geoconda Molina, Fabián Encarnación

2024-02-29

Ejercicio 1

Sea $\{X_1, \ldots, X_n\}$ una **m.a.s.** de X con distribución F. verificando que $\mathbb{E}[X]^6 < \infty$. Considere la estimación de $\eta = \mu^3$ con $\mu = \mathbb{E}[X]$:

1. Crear el U-estadístico para estimar ηU_n

Mediante $h(x_1, x_2, x_3) = x_1 x_2 x_3$ podemos obtener el U-estadístico para $\eta = \mu^3$

$$U_n = \binom{n}{3}^{-1} \sum_{1<=i<j<k<=n} h(X_i, X_j, X_k)$$

$$= \frac{6}{n(n-1)(n-2)} \sum_{1\leq i<j<k\leq n} X_i X_j X_k$$

donde $\sum_{1\leq i<j<k\leq n}$ es la suma de $\binom{n}{3}$ combinaciones de $\{i, j, k\}$ en $\{1, \ldots, n\}$

2. Calcular el U-estadístico proyectado

Dado que $\eta = \mathbb{E}[h(X_1, X_2, X_3)]$ la proyección de U_n es:

$$\hat{U}_n = \frac{3}{n} \sum_{i=1}^{3} h_1(X_i) + \eta$$

donde $h_1(x) = \mathbb{E}[h(x, X_2, X_3)] - \eta$.

Si tomamos $X_1 = x$, entonces:

$$\begin{aligned}
\mathbb{E}[h(X_1, X_2, X_3)|X_1 = x] &= \mathbb{E}[h(x, X_2, X_3)] \\
&= \mathbb{E}[x X_2 X_3] \\
&= x\mathbb{E}[X_2 X_3] \\
&= x\mu^2
\end{aligned}$$

Luego,

$$\hat{U}_n = \frac{3}{n}[\mu^2 \sum_{i=1}^{3} X_i + (\frac{n}{3} - 3)\eta]$$

$$= \frac{3}{n}\sum_{i=1}^{3}\{\mathbb{E}[h(X_1,X_2,X_3)|X_i] - \eta\} + \eta$$

$$= \frac{3}{n}\sum_{i=1}^{3}\{\mathbb{X}_i\mu^2 - \eta\} + \eta$$

3. Calcular la varianza de U_n

Tenemos que:

$$\mathrm{Var}(U_n) = \binom{n}{3}^{-1}\sum_{k=1}^{3}\binom{3}{k}\binom{n-3}{3-k}\xi_k$$

con $\xi_k = \mathrm{Var}(h_k(X_1,\ldots,X_n))$. Y por tanto:

$$\begin{aligned}
\mathrm{Var}(U_n) &= \binom{n}{3}^{-1}\sum_{k=1}^{3}\binom{3}{k}\binom{n-3}{3-k}(h_k(X_1,\ldots,X_n))\\
&= \binom{n}{3}^{-1}\frac{3}{2}(n-4)(n-5)\mathrm{Var}(h_1(X_1)) + 3(n-3)\mathrm{Var}(h_2(X_1,X_2)) + \mathrm{Var}(h_3(X_1,X_2,X_3))\\
&= \binom{n}{3}^{-1}\frac{3}{2}(n-4)(n-5)\mathrm{Var}(X_1) + 3(n-3)\mathrm{Var}(X_1X_2) + \mathrm{Var}(X_1X_2X_3)
\end{aligned}$$

4. Calcular $\frac{\mathbb{E}[\overline{\mathrm{X}}^3]}{\mathbb{E}[U_n]} = C_n$ y qué sucede cuando $\lim_{n\to\infty}$

Tenemos que:

$$\begin{aligned}
\frac{\mathbb{E}\left[\overline{\mathrm{X}}^3\right]}{\mathbb{E}\left[U_n\right]} &= \frac{\frac{1}{n^3}\sum_i \mathbb{E}X_i^3 + 3\sum_{i\neq j}\mathbb{E}X_iX_j^2 + \sum_{i\neq j\neq k}\mathbb{E}X_iX_jX_k}{\frac{6}{n(n-1)(n-2)}\sum_{1\leq i<j<k\leq n}X_iX_jX_k}\\
&= \frac{n\mathbb{E}X^3 + 3\left(\begin{array}{c} n\\ 2\end{array}\right)\mu\mathbb{E}X^2 + \left(\begin{array}{c} n\\ 3\end{array}\right)\mu^3}{n^3\mu^3}\\
&= C_n
\end{aligned}$$

Para $n\to\infty$,

$$\begin{aligned}
C_n &= \frac{\mathbb{E}X^3}{n^2\mu^3} + \frac{3(n-1)\mathbb{E}X^2}{2n^2\mu^3} + \frac{(n-1)(n-2)}{6n^2}\\
C_n &\to \frac{1}{6}
\end{aligned}$$

Ejercicio 2

A continuación deduciremos la probabilidad de que una observación determinada forme parte de una muestra bootstrap. Supongamos que obtenemos una muestra bootstrap a partir de un conjunto de n observaciones.

a. ¿Cuál es la probabilidad de que la primera observación bootstrap no sea la $j$ésima observación de la muestra original?. Justifique su respuesta.

Esto es 1-probabilidad de que sea la $j$ésima. Es decir:

$$P = 1 - \frac{1}{n}$$

b. ¿Cuál es la probabilidad de que la segunda observación bootstrap no sea la $j$ésima observación de la muestra original?

Como cada observación bootstrap es una muestra aleatoria, la probabilidad es la misma:

$$P = 1 - \frac{1}{n}$$

c. Argumentar que la probabilidad de que la obsersación $j$ésima no esté en la muestra bootstrap es $(1 - 1/n)^n$

Para que la observación $j$ésima no esté en la muestra, tendría que no estar en cada posición n, en otras palabras, no ser elegida en $1, 2, ..., n$. Por lo tanto la probabilidad es:

$$P = \left(1 - \frac{1}{n}\right)^n$$

d. Cuando $n = 5$, ¿cuál es la probabilidad de que la $j$ésima observación esté en la muestra bootstrap?

Utilizando la respuesta del literal anterior, se tiene:

```
n<-5
p<-1-(1-1/n)^n
cat("La probabilidad es:", p)
```

```
## La probabilidad es: 0.67232
```

e. Cuando $n = 100$, ¿cuál es la probabilidad de que $j$ésima observación esté en la muestra bootstrap?

Análogo al literal anterior:

```
n<-100
p<-1-(1-1/n)^n
cat("La probabilidad es:", p)
```

```
## La probabilidad es: 0.6339677
```

e. Cuando $n = 10000$, ¿cuál es la probabilidad de que $j$ésima observación esté en la muestra bootstrap?

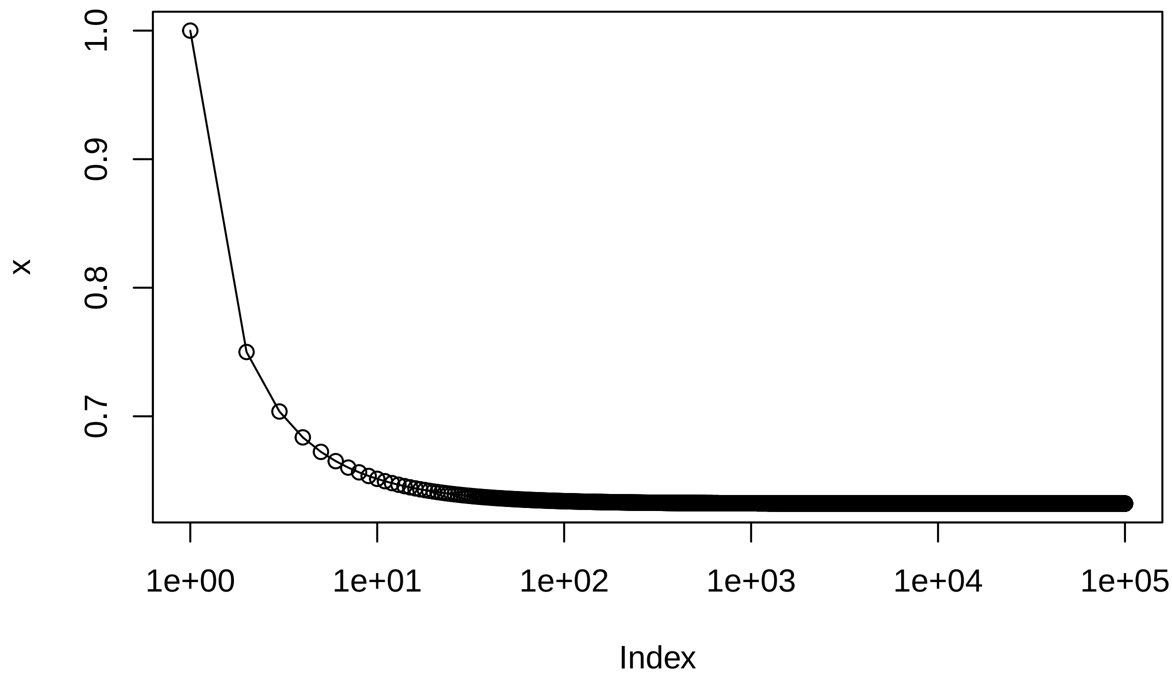
Análogo a los literales anteriores:

```
n<-100000
p<-1-(1-1/n)^n
cat("La probabilidad es:", p)
```

```
## La probabilidad es: 0.6321224
```

g. Crear un gráfico que muestre, para cada valor de n desde 1 a 100000, la probabilidad de que la $j$ésima observación esté en la muestra bootstrap. Comentar lo que se observa

```r
x<-numeric(100000)
for (i in 1:100000) {
  x[i]<-1-(1-1/i)^i
}
plot(x,log="x",type="o")
```



La probabilidad se acerca a 0.63 cuando n va creciendo.

Dado que:

$$e^x = \lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n,$$

Tomando $x=-1$, se tiene que la probabilidad cuando n tiende a infinito es:

$$P = 1 - \frac{1}{e}$$

h. Ahora investigaremos numéricamente la probabilidad de que una muestra bootstrap de tamaño $n = 100$ contenga la $j$ésima observación. Aquí $j = 4$. Creamos repetidamente muestras bootstrap, y cada vez registramos si la cuarta observación está contenida o no en la muestra bootstrap.

```r
store <- rep (NA, 10000)
 for (i in 1:10000) {
store[i] <- sum(sample(1:100, rep = TRUE) == 4) > 0
}
mean(store)
```

```
## [1] 0.6327
```

Vemos que la probabilidad cuando se remuestrea es cercana al límite encontrado en el literal anterior

Ejercicio 9

Seguimos considerando el uso de un modelo de regresión logística para predecir la probabilidad de `default` utilizando `income` y `balance` en el conjunto de datos `Default`. En concreto, ahora calcularemos las estimaciones de los errores estándar de los coeficientes de regresión logística de `income` y `balance` de dos formas diferentes: (1) utilizando el bootstrap, y (2) utilizando la fórmula estándar para calcular los errores estándar en la función `glm()`. No olvide establecer una semilla aleatoria antes de comenzar el análisis.

1. Utilizando las funciones `summary()` y `glm()`, determine los errores estándar estimados para los coeficientes asociados con `income` y `balance` en un modelo de regresión logística múltiple que utiliza ambos predictores.

```r
data(Default)
fit <- glm(default ~ income + balance, data = Default, family = "binomial")
summary(fit)
```

```
##
## Call:
## glm(formula = default ~ income + balance, family = "binomial",
##     data = Default)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.154e+01  4.348e-01 -26.545  < 2e-16 ***
## income       2.081e-05  4.985e-06   4.174 2.99e-05 ***
## balance      5.647e-03  2.274e-04  24.836  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1579.0  on 9997  degrees of freedom
## AIC: 1585
##
## Number of Fisher Scoring iterations: 8
```

Los errores estándar obtenidos por el método bootstrap son: $\beta_1 = 5e^{-6}$ and $\beta_2 = 2.3e^{-4}$.

2. Escriba una función, `boot.fn()`, que tome como entrada el conjunto de datos `Default` así como un índice de las observaciones, y que produzca las estimaciones de los coeficientes para `income` y `balance` en el modelo de regresión logística múltiple.

```r
boot.fn <- function(x, i) {
  fit <- glm(default ~ income + balance, data = x[i, ], family = "binomial")
  coef(fit)[-1]
}
```

3. Utilice la función `boot()` junto con su función `boot.fn()` para estimar los errores estándar de los coeficientes de regresión logística para los ingresos y el saldo.

```
library(boot)
set.seed(42)
boot(Default, boot.fn, R = 1000)
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = Default, statistic = boot.fn, R = 1000)
##
##
## Bootstrap Statistics :
##         original       bias     std. error
## t1* 2.080898e-05 2.737444e-08 5.073444e-06
## t2* 5.647103e-03 1.176249e-05 2.299133e-04
```

4. Comente los errores estándar estimados obtenidos utilizando la función `glm()` y utilizando su función bootstrap.

Podemos notar que los errores estandar ontenidos por el método bootstrap son iguales a los obtenidos por `glm()`

\end{document}