El Estimador de Regresión en R.

Muestreo II

12 de septiembre de 2017

Se presenta un ejemplo para la implementación del estimador de regresión en R. Se trabaja con la población MU281 del Apéndice B del "Libro Amarillo". Se ensayan algunos de los resultados presentados en el apartado 7.9.1. de dicho libro.

1. Introducción

Para empezar, se inicia la sesión en R, luego se carga la libreria `survey` y se leen los datos, que deben estar grabados en en un archivo MU281.txt, en el directorio desde donde se abre el R:

```
> library(survey)
> MU281 <- read.table("MU281.txt", header = TRUE)
```

Los datos:

```
> MU281[1:5, ]
```

```
  LABEL P85 P75 RMT85 CS82 SS82 S82 ME84 REV84 REG CL
1     1  33  27   288   13   24  49 2135  2836   1  1
2     2  19  15   139   14   12  41  957  2035   1  1
3     3  26  20   196   12   14  41 1530  6030   1  1
4     4  19  15   159   12   19  41 1059  4704   1  1
5     5  56  52   536   20   27  61 3951  5183   1  1
```

de dimensión

```
> dim(MU281)
```

```
[1] 281  11
```

En este ejemplo, solamente se usan las siguientes variables:

LABEL: las etiquetas,
RMT85: la recaudación por impuestos municipales en 1985 (en millones de coronas),
CS82: el número de bancas del partido conservador en el legislativo municipal y
SS82: el número de bancas del partido social-demócrata en el legislativo municipal.

Construcción de datos poblacionales:

```
> N <- nrow(MU281)
> k <- seq(1, N)
> fpc <- rep(N, N)
> U <- data.frame(k, MU281$RMT85, MU281$CS82, MU281$SS82, fpc)
> nombres <- c("k", "y", "x.1", "x.2", "N")
> colnames(U) <- nombres
> U[1:5, ]
```

```
  k   y x.1 x.2   N
1 1 288  13  24 281
2 2 139  14  12 281
3 3 196  12  14 281
4 4 159  12  19 281
5 5 536  20  27 281
```

Cálculo de los totales poblacionales:

```
> t.y.U <- sum(U$y)
> t.x1.U <- sum(U$x.1)
> t.x2.U <- sum(U$x.2)
> t.x.U <- t(t(c(N, t.x1.U, t.x2.U)))
> t.y.U
```

```
[1] 53151
```

```
> t.x.U
```

```
     [,1]
[1,]  281
[2,] 2508
[3,] 6193
```

2. Resultados para una muestra *SI* de tamaño $n = 100$.

Se construye una muestra SI de tamaño $n = 100$:

```
> n <- 100
> pw <- rep(n/N, N)
> set.seed(48182)
> s <- sample(seq(1, N), n, replace = FALSE, prob = pw)
```

Con la función `svydesign` se espesifica el diseño muestral a usar:

```
> datos <- U[s, ]
> ps <- svydesign(id = ~1, data = datos, fpc = ~N)
> summary(ps)

Independent Sampling design
svydesign(id = ~1, data = datos, fpc = ~N)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3559  0.3559  0.3559  0.3559  0.3559  0.3559
Population size (PSUs): 281
Data variables:
[1] "k"   "y"   "x.1" "x.2" "N"
```

Se está en condiciones de estimar el total y el desvío del estimador:

```
> svytotal(~y, ps)

  total     SE
y 55110 4538.3
```

Donde:

$\hat{t}_1 = \hat{t}_{y\pi} = 55110 = N \sum_s \frac{y_k}{n} = N\bar{y}_s$ (Ver (7.9.2))

y

$\hat{V}\left(\hat{t}_{y\pi}\right) = 4538{,}3^2 = N^2 (1-f) \frac{S^2_{y_s}}{n}$.

Si se quiere utilizar el estimador de razón con la variable auxiliar `x.1` se debe especificar:

```
> ra1 <- svyratio(~y, ~x.1, ps)
> pop <- data.frame(x.1 = t.x1.U)
> predict(ra1, pop$x.1)

$total
       x.1
y 55578.41
```

```
$se
      x.1
y 3534.81
```

y para la variable `x.2`

```
> ra2 <- svyratio(~y, ~x.2, ps)
> pop <- data.frame(x.2 = t.x2.U)
> predict(ra2, pop$x.2)

$total
       x.2
y 54270.38

$se
       x.2
y 3686.643
```

Se puede verificar, por ejemplo, para `x.2`:

$\hat{t}_3 = \hat{t}_{yra}\,(x,2) = \sum_U x_{2k} \frac{\sum_s \check{y}_k}{\sum_s \check{x}_{2k}}$ (Ver (7.9.4))

y

$\hat{V}\left(\hat{t}_2\right) = \hat{V}\left(\hat{t}_{yra}\,(x,2)\right) = N^2 \frac{1-f}{n} \frac{\sum_s g_{ks}^2 e_{ks}^2}{n-1}$ (Ver (7.9.8))

con $g_{ks} = \frac{\sum_U x_k}{\sum_s \check{x}_{2k}}$, $e_{ks} = y_k - \hat{B} x_{2k}$ y $\hat{B} = \frac{\sum_s \check{y}_k}{\sum_s \check{x}_{2k}}$.

El estimador de regresión tiene la misma lógica pero con la función `svglm`, con los siguientes comandos se realizan las estimaciones puntales y las de los desvíos para los estimadores $\hat{t}_4 = \hat{t}_{yreg}\,(x1)$, $\hat{t}_5 = \hat{t}_{yreg}\,(x2)$ y $\hat{t}_6 = \hat{t}_{yreg}\,(x1, x2)$ de la sección 7.9.1.

```
> reg1 <- svyglm(y ~ x.1, ps)
> summary(reg1)

Call:
svyglm(formula = y ~ x.1, ps)

Survey design:
svydesign(id = ~1, data = datos, fpc = ~N)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  -54.838     27.212  -2.015   0.0466 *
x.1           28.357      3.863   7.341 6.28e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 23334.76)

Number of Fisher Scoring iterations: 2

> pop <- data.frame(x.1 = t.x1.U)
> predict(reg1, newdata = pop, total = N)

   link     SE
1 55710 3501.9

> reg2 <- svyglm(y ~ x.2, ps)
> summary(reg2)

Call:
svyglm(formula = y ~ x.2, ps)

Survey design:
svydesign(id = ~1, data = datos, fpc = ~N)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -197.659     37.443  -5.279 7.82e-07 ***
x.2           17.595      1.994   8.824 4.31e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 23197.1)

Number of Fisher Scoring iterations: 2

> pop <- data.frame(x.2 = t.x2.U)
> predict(reg2, newdata = pop, total = N)

   link     SE
1 53425 3308.5

> reg12 <- svyglm(y ~ x.1 + x.2, ps)
> summary(reg12)

Call:
svyglm(formula = y ~ x.1 + x.2, ps)
```

```
Survey design:
svydesign(id = ~1, data = datos, fpc = ~N)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -334.438     35.676  -9.374 3.01e-15 ***
x.1           23.304      2.621   8.891 3.32e-14 ***
x.2           14.491      1.521   9.528 1.41e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 12145.86)

Number of Fisher Scoring iterations: 2
```

```
> pop <- data.frame(x.1 = t.x1.U, x.2 = t.x2.U)
> predict(reg12, newdata = pop, total = N)

   link     SE
1 54215 2440.3
```

3. El cálculo del estimador de regresión

En esta sección se verifican algunos de los cálculos anteriores.

Los resultados para el estimador $\hat{t}_1 = \hat{t}_{y\pi}$:

```
> y.s <- as.matrix(datos$y)
> tpi <- (N/n) * sum(y.s)
> tpi

[1] 55109.72

> vpi <- N^2 * (1 - n/N) * (1/n) * var(y.s)
> sqrt(vpi)

         [,1]
[1,] 4538.307
```

Para el estimador de regresión, $\hat{t}_6 = \hat{t}_{yreg}\,(x1, x2)$:

```
> x.s <- as.matrix(rbind(1, datos$x.1, datos$x.2))
> T <- (N/n) * x.s %*% t(x.s)
> t <- (N/n) * x.s %*% y.s
```

```
> b <- solve(T) %*% t
> treg <- t(t.x.U) %*% b
> treg

        [,1]
[1,] 54214.6

> e.k.s <- y.s - t(x.s) %*% b
> vreg1 <- N^2 * (1 - n/N) * (1/n) * var(e.k.s)
> sqrt(vreg1)

         [,1]
[1,] 2485.458
```

Donde

$\texttt{vreg1} = 2485{,}46^2 = N^2 \frac{1-f}{n} \frac{\sum_s e_{ks}^2}{n-1}$ (Ver (7.9.9)).

```
> t.x.s <- as.matrix(apply((N/n) * x.s, 1, sum))
> g.k.s <- t(1 + t((t.x.U - t.x.s)) %*% solve(T) %*% x.s)
> (N/n) * sum(y.s * g.k.s)

[1] 54214.6

> (N/n) * apply(x.s %*% g.k.s, 1, sum)

[1]  281 2508 6193

> vreg2 <- N^2 * (1 - n/N) * (1/n) * (n - 1)^(-1) * sum(c((e.k.s)^2) *
+     (g.k.s^2))
> sqrt(vreg2)

[1] 2440.281
```

Donde

$\texttt{vreg2} = 2440{,}28^2 = N^2 \frac{1-f}{n} \frac{\sum_s g_{ks}^2 e_{ks}^2}{n-1}$ (Ver (7.9.8)).

4. Resultados simulados

Se simulan 5000 muestras *SI* y se analizan los resultados para los distintos estimadores utilizados:

```
> R <- 5000
> tpi <- rep(0, R)
> vtpi <- rep(0, R)
> treg <- rep(0, R)
> vreg1 <- rep(0, R)
> vreg2 <- rep(0, R)
> set.seed(987654321)
> for (i in 1:R) {
+     s <- sample(seq(1, N), n, replace = FALSE, prob = pw)
+     y.s <- as.matrix(U[s, "y"])
+     tpi[i] <- (N/n) * sum(y.s)
+     vtpi[i] <- N^2 * (1 - n/N) * (1/n) * var(y.s)
+     x.s <- as.matrix(rbind(1, U[s, "x.1"], U[s, "x.2"]))
+     T <- (N/n) * x.s %*% t(x.s)
+     t <- (N/n) * x.s %*% y.s
+     b <- solve(T) %*% t
+     treg[i] <- t(t.x.U) %*% b
+     e.k.s <- y.s - t(x.s) %*% b
+     vreg1[i] <- N^2 * (1 - n/N) * (1/n) * var(e.k.s)
+     t.x.s <- as.matrix(apply((N/n) * x.s, 1, sum))
+     g.k.s <- t(1 + t((t.x.U - t.x.s)) %*% solve(T) %*% x.s)
+     vreg2[i] <- N^2 * (1 - n/N) * (1/n) * (n - 1)^(-1) * sum(c((e.k.s)^2) *
+         (g.k.s^2))
+ }
```

Algunos de los resultados para las 5000 simulaciones se presentan en el siguiente cuadro que es similar al cuadro 7.2. de la pág 280 del "Libro Amarillo".

Estimador	$\bar{\hat{t}}$	$S^2_{\hat{t}}$	$\bar{\hat{V}}_g$	$\bar{\hat{V}}_{sim}$	AV
$\hat{t}_1$	5,32	0,205		0,204	
$\hat{t}_6$	5,31	0,056	0,052	0,05	

Por último, se presentan los histogramas de las 5000 réplicas de $\hat{t}_1$ y $\hat{t}_6$.

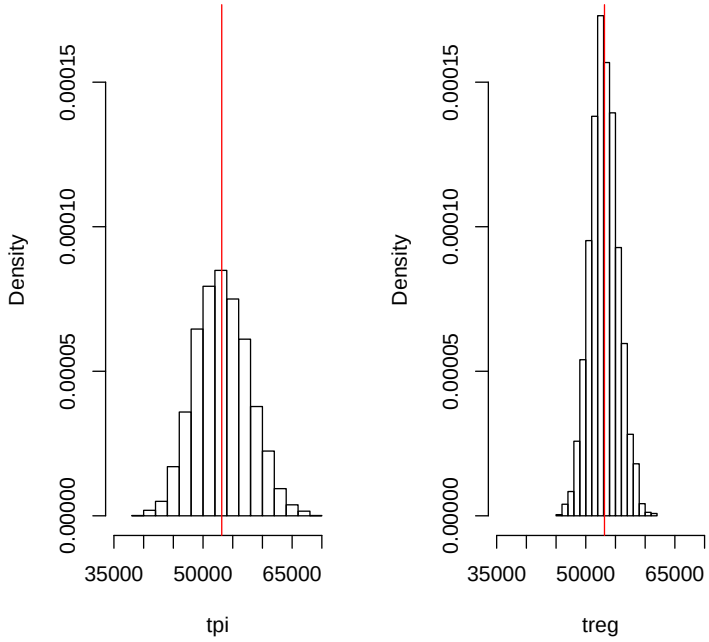

Figura 1: 5000 muestras SI de tamaño $n = 100$