Capítulo 5. Métodos de estimación libres de distribución

Índice

1. Estimación de parámetros

Los parámetros son características de las poblaciones de las que proceden los datos. Uno de los objetivos primordiales de la estadística aplicada es obtener información acerca del valor de los parámetros a partir de los datos contenidos en una muestra, a esto se le denomina *inferencia estadística*. Hay dos formas fundamentales de realizar la inferencia: la estimación de parámetros y el contraste de hipótesis. El problema de la estimación consiste en asignar un valor a los parámetros a partir de la muestra. La estimación se realiza fundamentalmente por tres métodos, el método de los momentos, mínimos cuadrados y máxima verosimilitud.

En ocasiones, las poblaciones vienen descritas por un conjunto de parámetros y por una función de distribución, por ejemplo, si asumimos que la distribución poblacional de una variable sigue la distribución de Poisson, su parámetros es λ. En este caso la estimación de parámetros se realiza por máxima-verosimilitud, que veremos en el capítulo siguiente. El método de máxima-verosimilitud requiere supuestos fuertes, en concreto requiere conocer la función de distribución de la variable aleatoria.

También existen situaciones en las que un problema estadístico está descrito únicamente por los parámetros que representan la población, sin que se asuma una determinada distribución para las variables aleatorias. Por ejemplo, un modelo de regresión lineal asume que la media de una variable Y para un valor fijo de otra variable X es $Y = a + bX$; sin embargo, el modelo de regresión lineal no asume nada acerca de cual es la distribución de X. En estos casos la estimación se realiza por los métodos de los momentos o el de mínimos cuadrados. Estos métodos realizan menos supuestos que máxima-verosimilitud, por tanto es posible aplicarlos en un rango más amplio de situaciones. Por el contrario, los supuestos más fuertes que realiza máxima-verosimilitud redundan en que proporciona más información acerca de los parámetros.

De modo genérico los parámetros se indican por letras griegas, cómo θ y el símbolo $\hat{\theta}$ representa el estimador de θ. Es decir θ es un parámetro y toma un valor fijo mientras que $\hat{\theta}$ es una variable aleatoria cuyo valor depende de la muestra.

2. Estimación por el método de los momentos

Supongamos que se desean estimar k parámetros $(\theta_1, \theta_2, \ldots, \theta_k)$ de una distribución $f(x; \theta_1, \theta_2, \ldots, \theta_k)$. El método de los momentos consiste en buscar los valores de los parámetros que igualan los k primeros momentos con relación al origen en la muestra y en la población. Por tanto, no se utiliza la distribución completa para obtener estimadores sino únicamente los k primeros momentos.

Supongamos que α_r es el momento de orden r, es decir

$$\alpha_r = E(X^r).$$

Supongamos que el correspondiente momento muestral es

$$\hat{\alpha}_r = \frac{\sum_{i=1}^{n} x^r}{n}.$$

Utilizando los momentos poblacionales y muestrales, es posible formar el siguiente sistema de k ecuaciones con k incógnitas:

$$\left.\begin{array}{rcl} \alpha_1 & = & \hat{\alpha}_1 \\ \alpha_2 & = & \hat{\alpha}_2 \\ & \vdots & \\ \alpha_k & = & \hat{\alpha}_k \end{array}\right\}$$

Los valores de los parámetros que satisfacen el sistema de ecuaciones constituyen los estimadores por el método de los momentos, $(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k)$.

Ejemplo 1. Supongamos que queremos estimar el parámetro λ de una distribución de Poisson. Cómo solo hay un parámetro utilizamos nada más el primer momento, que en la distribución de Poisson es $\alpha_1 = E(X) = \lambda$. El primer momento muestral es $\hat{\alpha}_1 = \overline{X}$. Entonces, el estimador es

$$\hat{\lambda} = \overline{X}$$

Si la muestra obtenida fuera $\mathbf{x}^T = \{5, 3, 4\}$, entonces $\hat{\lambda} = 4$.

Ejemplo 2. En la distribución exponencial se cumple que $E(X) = 1/\omega$. Por tanto el estimador se obtiene de la ecuación

$$\frac{1}{\omega} = \overline{X}$$

Es decir $\hat{\omega} = \overline{X}$. Si se obtiene la muestra \$$\boldsymbol{x}$ = (3, 10, 2) \$, entonces $\hat{\omega} = 0{,}2$.

Ejemplo 3. En el ámbito de la psicología cognitiva se utiliza el siguiente modelo para formalizar el tiempo de ejecución de una tarea

$$E(T) = \frac{\alpha}{\beta}\ ,$$
$$Var(T) = \frac{\alpha}{\beta^2}\ ,$$

Para estimar α y β podemos utilizar la media y varianza muestral. Sustituyendo

$$\overline{X} = \frac{\alpha}{\beta}\ ,$$
$$S^2 = \frac{\alpha}{\beta^2}\ ,$$

De donde $S^2 = \frac{\overline{X}}{\beta}$, y los estimadores son

$$\hat{\alpha} = \frac{\overline{X}^2}{S^2}$$
$$\hat{\beta} = \frac{\overline{X}}{S^2}$$

Ejemplo 4. Supongamos que un investigador en psicología cognitiva toma datos de tiempo de reacción en segundos al realizar una determinada tarea y encuentra el resultado

$$\boldsymbol{x} = (5,\ 3,\ 7,\ 9,\ 1)$$

Vamos a realizar la estimación de la distribución gamma. Cuando estudiamos esta distribución vimos que sus parámetros pueden calcularse del siguiente modo a partir del valor esperado y la varianza de la distribución

$$\alpha = \frac{\mu^2}{\sigma^2}$$
$$\beta = \frac{\mu}{\sigma^2}$$

Estas fórmulas nos permiten realizar la estimación de los parámetros α y β por el método de los momentos

```
x = c(5, 3, 7, 9, 1)
media <- mean(x)
varianza <- var(x)
alpha_Momentos <- media^2 / varianza
beta_Momentos <- media / varianza
cat(sprintf("Metodo de los momentos, a = %5.2f, b = %5.2f", alpha_Momentos, beta_Momentos))
```

```
## Metodo de los momentos, a =  2.50, b =  0.50
```

Supongamos que el investigador tiene razones teóricas para suponer que $\mu = 4$. Entonces la relación entre los parámetros y los momentos de la distribución es

$$\alpha = \frac{16}{\sigma^2}$$
$$\beta = \frac{4}{\sigma^2}$$

El siguiente código estima el modelo gamma con esta restricción

```
alpha_Momentos_R <- 16 / varianza
beta_Momentos_R <- 4 / varianza
cat(sprintf("Metodo de los momentos, modelo restringido, a = %5.2f, b = %5.2f",
alpha_Momentos_R, beta_Momentos_R))
```

```
## Metodo de los momentos, modelo restringido, a =  1.60, b =  0.40
```

Por último, podemos comparar la media observada en la muestra con la media reproducida por el modelo restringido. El siguiente código realiza los cálculos. Podemos ver que existe discrepancia entre el valor observado y el reproducido, aunque sería necesario realizar un contraste de hipótesis para valorar si dicha discrepancia es significativa.

```
cat(sprintf("Media observada = %5.2f, reproducida = %5.2f",
            mean(x), alpha_Momentos_R/beta_Momentos_R))
```

```
## Media observada =  5.00, reproducida =  4.00
```

3. Estimación por mínimos cuadrados

El método de mínimos cuadrados es un principio matemático muy general que se aplica para el ajuste de funciones en ámbitos muy diferentes, sea en estadística o en otros. La idea del método es la siguiente, sea

$x_1, \ldots, x_n$ un conjunto de datos y $f_i(\theta)$ el valor pronosticado para el dato x_i. La función $f(\theta)$ depende de un parámetro θ, y el estimador mínimo-cuadrático de θ es el valor que minimiza la distancia cuadrática

$$d(\theta) = \sum_{i=1}^{n} (x_i - f_i(\theta))^2$$

Ejemplo 5. En un estudio de psicología del aprendizaje disponemos de datos referidos a las variables $X \equiv$ número de horas de entrenamiento y $Y \equiv$ número de tareas realizadas correctamente. Vamos a ajustar un modelo de regresión lineal de Y sobre X, además asumimos que el intercepto de la regresión es cero porque los sujetos que entrenen cero horas no deberían completar ninguna tarea. Entonces el pronóstico de la regresión depende sólo de la pendiente

$$y'_i = \beta x_i$$

Según el criterio de mínimos cuadrados, el estimador de β es el valor que minimiza la distancia cuadrática

$$d(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

Según sabemos del cálculo diferencial, en el máximo de una función su derivada es cero. Por tanto derivamos la distancia cuadrática con respecto a β:

$$d'(\beta) = 2 \sum_{i=1}^{n} x_i (y_i - \beta x_i)$$

La ecuación de estimación es $d'(\beta) = 0$. Despejando

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Supongamos que al recoger datos obtenemos

$$\boldsymbol{x} = (3, 1, 5, 4)$$
$$\boldsymbol{y} = (5, 1, 7, 9)$$

Podemos utilizar el siguiente código R para representar los datos y calcular el estimador $\hat{\beta}$. El resultado aparece en la figura 1.

```
x <- c(3, 1, 5, 4)
y <- c(5, 1, 7, 9)
plot(x,y)
beta <- sum(x*y) / sum(x*x)
xp<-seq(1,5,by=0.001)
yp<-xp*beta
lines(xp, yp, col="firebrick", lwd=1.5)
legend(1,8, sprintf("y' = %3.1fx", beta), lty=1, col="firebrick", lwd=1.5)
```

También es posible estimar la ecuación de regresión utilizando la función *optimize* de R para minimizar la discrepancia cuadrática, $d(\beta)$, de forma numérica. En el siguiente código se ha programado $d(\beta)$ y luego se le pasa esta función a *optimize* para que busque el valor de β que la minimiza. Con *print* se muestran los resultados de la estimación, almacenados en el objeto *dMin*, y podemos ver que el parámetro estimado mediante *optimize* coincide con el que aparece en la figura 1.

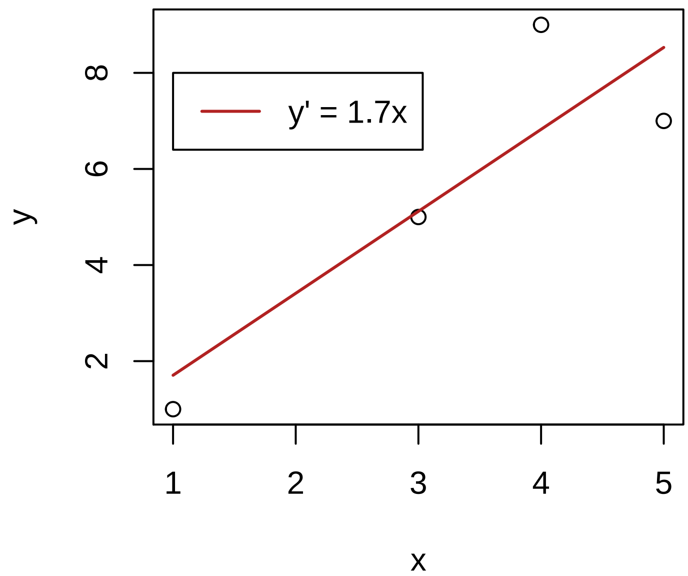

Figura 1: Regresión lineal

```
d <- function(b) sum((y - x*b)^2)
dMin <- optimize(d, c(0, 5))
print(dMin)
```

```
## $minimum
## [1] 1.705882
##
## $objective
## [1] 7.588235
```

Podemos entender mejor lo que significa la estimación mínimo cuadrática mediante la figura 2. En el eje de abcisas tenemos valores de β entre 0 y 5. En el eje de ordenadas aparece el valor de $d(\beta)$ con los datos del ejemplo. Vemos que el valor de β que minimiza $d(\beta)$ es el estimador mínimo-cuadrático.

Ejemplo 6. En un modelo de regresión lineal, el valor pronosticado para la variable dependiente es

$$y_i' = \alpha + \beta x_i$$

La distancia cuadrática entre los datos y los pronósticos es

$$d(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta_i)^2$$

Para estimar α y β no podemos emplear *optimize* porque esta función busca el mínimo con respecto a un parámetro y la discrepancia depende de dos parámetros. Para estos casos disponemos de la función *optim*.

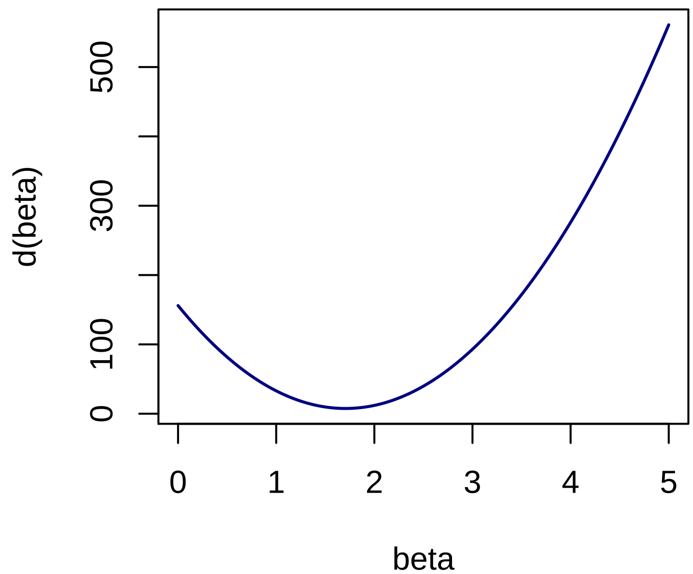

Figura 2: Discrepancia cuadrática

Para utilizar *optim* es necesario que la función a minimizar reciba un único argumento consistente en un vector de parámetros. Por ejemplo, vamos a estimar la regresión a partir del siguiente conjunto de datos

$$\boldsymbol{x} = (2, 4, 3, 5)$$
$$\boldsymbol{y} = (5, 8, 7, 10)$$

El siguiente código de R la función *d(v)* recibe como argumento un vector de dos elementos, que son los parámetros α y β. A continuación le pasamos *d(v)* a *optim* para que busque el mínimo. El primer argumento de *optím* es el valor inicial de los parámetros a partir de los que realizar el proceso de búsqueda de los estimadores, y el segundo argumento es la función a minimizar.

```
x <- c(2, 4, 3, 5)
y <- c(5, 8, 7, 10)
d <- function(v) sum((y - v[1] - x*v[2])^2)
dMin <- optim(c(0,0), d)
print(dMin)
```

```
## $par
## [1] 1.901288 1.599696
##
## $value
## [1] 0.2000007
##
## $counts
## function gradient
##       73       NA
##
```

```
## $convergence
## [1] 0
##
## $message
## NULL
```

La salida de resultados de *optim* tiene varios campos, los principales son *par*, que contiene el valor estimado para los parámetros, *value* es el mínimo de *d(v)*, y *counts*, que indica cuantas pruebas ha realizado en el proceso de búsqueda de los valores estimados.

Ejemplo 7. La ley potencial de Stevens se utiliza en el ámbito de la psicofísica para predecir la sensación S producida por un estímulo de magnitud X. El valor pronosticado de S es

$$y'_i = \alpha x_i^{\beta}$$

Disponemos de los siguientes datos

$$\begin{aligned}\boldsymbol{x} &= (5, 8, 3, 4)\\ \boldsymbol{y} &= (72, 196, 22, 50)\end{aligned}$$

y vamos a realizar la estimación minimizando la distancia cuadrática

$$d(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha x_i^{\beta})^2$$

El siguiente código R realiza la estimación

```r
x <- c(5, 8, 3, 4)
y <- c(72, 196, 22, 50)
d <- function(v) sum((y - v[1]*x^v[2])^2)
dMin <- optim(c(0,0), d)
print(dMin)
```

```
## $par
## [1] 2.535801 2.090769
##
## $value
## [1] 28.10067
##
## $counts
## function gradient
##      133       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

4. Ejercicios

Ejercicio 1. Supongamos que disponemos de los siguientes datos, que representan la proporción de tareas realizadas correctamente por un grupo de cinco sujetos:

$$\boldsymbol{p} = (0,5,\ 0,8,\ 0,6,\ 0,7,\ 0,4)$$

Cuando los datos son proporciones, es común utilizar un modelo estadístico que asume que la media y varianza son

$$\begin{aligned} E(P) &= \mu \\ Var(P) &= \mu(1-\mu)\phi \end{aligned}$$

Realice la estimación de μ y π por el método de los momentos.

Ejercicio 2. En un estudio sobre tiempos de reacción se han encontrado los siguientes datos de cinco sujetos

$$\boldsymbol{t} = (9,\ 2,\ 5,\ 6,\ 8)$$

Vamos a utilizar el siguiente modelo para estos datos

$$\begin{aligned} E(T) &= \kappa\theta \\ Var(T) &= \kappa\theta^2 \end{aligned}$$

Obtenga los estimadores por el método de los momentos.

Ejercicio 3. En un contexto clínico estamos analizando el número de cigarrillos fumados, Y, en función del número de días de tratamiento contra el tabaquismo, X. Los datos son

$$\begin{aligned} \boldsymbol{x} &= (5,3,7,8) \\ \boldsymbol{y} &= (5,5,21,16) \end{aligned}$$

Con este tipo de datos la regresión lineal no es adecuada porque Y no puede ser negativa, mientras que un modelo lineal predeciría que Y es menor de cero para valores bajos de X. Por este motivo vamos a estimar un modelo de regresión en el cual el valor pronosticado de Y es

$$y'_i = \exp(\alpha + \beta x_i)$$

Realice la estimación mínimo-cuadrática utilizando R.

Ejercicio 4. Disponemos de los siguientes datos acerca del número de segundo en se tarda en realizar una tarea, Y, en función de su dificultad, X:

$$\begin{aligned} \boldsymbol{x} &= (4,\ 7,\ 2,\ 4,\ 6) \\ \boldsymbol{y} &= (4,6,\ 0,9,\ 18,6,\ 5,0,\ 1,4) \end{aligned}$$

Ajuste el siguiente modelo de regresión utilizando mínimos-cuadrados en R:

$$y'_i = \frac{1}{\exp(\alpha - \beta x_i)}$$

Ejercicio 5. Un economista ha recogido los siguientes datos acerca del gasto semanal en ocio cultural, Y, en función de los ingresos mensuales en cientos de euros, X:

$$\begin{aligned} \boldsymbol{x} &= (1,\ 2,\ 3,\ 4,\ 5) \\ \boldsymbol{y} &= (10,\ 15,\ 40,\ 55,\ 90) \end{aligned}$$

Ajuste el modelo de regresión polinómica, $y'_i = \alpha + \beta x_i + \delta x_i^2$, mediante mínimos-cuadrados utilizando R.