Capítulo 6. Estimación por máxima verosimilitud

Índice

El método de máxima verosimilitud es actualmente el más extendido con modelos relativamente sofisticados. Se diferencia del método de los momentos y de mínimos cuadrados en que hace supuestos fuertes acerca de los datos. En concreto asume que se conoce la función de probabilidad o densidad de las variables observadas. Esto tiene ventajas e inconvenientes. La ventaja es que hacer supuestos fuertes se traduce en que el método proporciona más información acerca de los datos, en concreto permite estimar los errores típicos de los parámetros. Además, los estimadores máximo-verosímiles tienen buenas propiedades estadísticas. El inconveniente es que dichos supuestos pueden no cumplirse o puede que la distribución de las variables observadas sea desconocida, por lo que máxima verosimilitud es aplicable en un rango menor de situaciones que los otros métodos.

1. Función de verosimilitud

El método de máxima-verosimilitud requiere que la función de probabilidad o densidad de los datos observados, $f(x)$, sea conocida; puede ser la normal, Poisson, binomial, etc. pero su forma matemática debe estar especificada. Además por sencillez es habitual, aunque no imprescindible, asumir muestreo aleatorio simple. Bajo el supuesto de m.a.s. la función de probabilidad o densidad de probabilidad de una muestra de n observaciones es

$$\begin{aligned} f(\boldsymbol{x}) &= f(x_1, \ldots, x_n) \\ &= f(x_1) \cdots f(x_n) \\ &= \prod_{i=1}^{n} f(x_i), \end{aligned}$$

donde el símbolo $\prod_{i=1}^{n}$ representa el producto de n términos, es similar al símbolo del sumatorio ($\sum_{i=1}^{n}$) pero multiplicando los elementos en lugar de sumarlos. El método de máxima verosimilitud toma $f(\boldsymbol{x})$ como base para realizar la estimación.

Supongamos que se ha tomado una m.a.s., $\boldsymbol{x}$, y el propósito es estimar el parámetro θ. La *función de verosimilitud* es la función $f(\boldsymbol{x})$, entendida como una función de θ y manteniendo $\boldsymbol{x}$ fijo al valor encontrado en la muestra. Esto suele indicarse del modo

$$L(\theta) = f(\boldsymbol{x})$$

donde L procede de *likelihood* (verosimilitud en inglés).

Ejemplo 1. Sea $X > 0$ una variable aleatoria con distribución Weibull, cuya función de densidad es

$$f(x) = \frac{1}{\lambda} \exp\left(-\frac{x}{\lambda}\right).$$

Si tomamos una muestra aleatoria simple de n observaciones, la función de densidad de la muestra es:

$$\begin{aligned} f(\boldsymbol{x}) &= \prod_{i=1}^{n} \frac{1}{\lambda} \exp\left(-\frac{x_i}{\lambda}\right) \\ &= \frac{1}{\lambda} \exp\left(-\frac{x_1}{\lambda}\right) \times \frac{1}{\lambda} \exp\left(-\frac{x_2}{\lambda}\right) \times \cdots \times \frac{1}{\lambda} \exp\left(-\frac{x_n}{\lambda}\right) \\ &= \frac{1}{\lambda^n} \exp\left(-\frac{\sum_{i=1}^{n} x_i}{\lambda}\right) \\ &= \frac{1}{\lambda^n} \exp\left(-\frac{n\overline{X}}{\lambda}\right) \end{aligned}$$

Por tanto, la función $f(\boldsymbol{x})$ depende únicamente de la media muestral $\overline{X}$, y no de ningún otro dato o cantidad observada en la muestra. Cuando esto sucede así se dice que $\overline{X}$ es un *estadístico suficiente* para λ. Es decir, toda la información observada en la muestra se resume en $\overline{X}$, que es el único dato necesario para estimar λ. Dicho de otra manera, dos muestras que tuviesen la misma media producirían el mismo valor estimado de λ con independencia de que variaran en otros aspectos como la varianza, etc.

Supongamos que hemos tomado una m.a.s. de tamaño tres y se encuentra el resultado $\boldsymbol{x} = (2, 7, 3)'$. El estadístico suficiente es $\overline{X} = 4$, a partir del cual la función de verosimilitud es

$$L(\lambda) = \frac{1}{\lambda^3} \exp\left(-\frac{3 \times 4}{\lambda}\right).$$

La figura 1 representa la función $L(\lambda)$ para valores de λ entre 0 y 20. Es importante advertir que $L(\lambda)$ no es la función de densidad de λ, es la función de densidad de $f(\overline{X} = 4)$ para distintos valores de λ

2. Estimación por máxima verosimilitud

La estimación por máxima verosimilitud consiste en asumir que $L(\theta)$ mide la compatibilidad de un valor de θ con los datos observados, por lo que se busca aquel valor del parámetro que es máximamente compatible con la muestra.

> El estimador máximo-verosímil de un parámetro θ es el valor de θ que hace máxima la función de probabildiad o densidad de probabilidad de los datos observados.

Al variar $L(\theta)$ en función de θ no estamos obteniendo la la probabilidad de que cada valor de θ sea el correcto sino como de verosímil es cada valor, entendiendo que un valor es inverosímil cuando, si ese valor fuese el correcto, sería improbable encontrar unos datos como los observados.

La forma práctica de obtener el estimador máximo-verosímil consiste utilizar los conceptos del cálculo diferencial para encontrar el máximo de una función. En primer lugar, en la mayoría de las ocasiones no se trabaja directamente con $L(\theta)$ sino con su logaritmo, denominado $l(\theta) = \log L(\theta)$. Entre otros motivos, esto se debe a que ambas funciones alcanzan su valor máximo en el mismo punto de θ, sin embargo $l(\theta)$ suele ser más sencilla que $L(\theta)$ por lo que es más cómodo trabajar con ella.

Para saber cual es el máximo de $l(\theta)$, se utiliza la propiedad de que en el máximo de una función su derivada toma el valor cero. Por ello, se calcula la derivada de $l(\theta)$ con respecto a θ, y se busca el valor de θ que hace que dicha derivada sea cero. Los siguientes ejemplos ilustran este proceso.

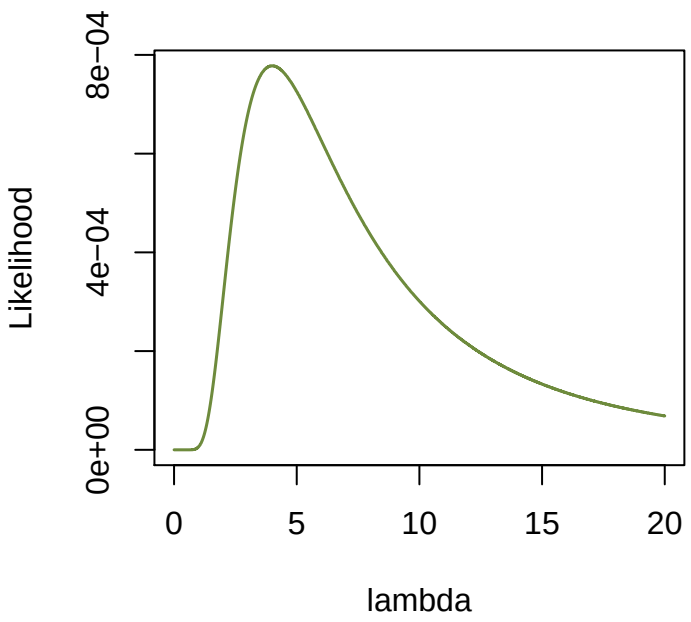

Figura 1: Función de verosimilitud Weibull

Ejemplo 2. La función de verosimilitud Weibull y su logaritmo son

$$L(\lambda) = \frac{1}{\lambda^n} \exp\left(-\frac{n\overline{X}}{\lambda}\right)$$
$$l(\lambda) = -n \log \lambda - \frac{n\overline{X}}{\lambda}$$

En el ejemplo 1 vimos el caso en que $n = 3$ y $\overline{X} = 4$. Entonces tenemos

$$L(\lambda) = \frac{1}{\lambda^3} \exp\left(-\frac{12}{\lambda}\right)$$
$$l(\lambda) = -3 \log \lambda - \frac{12}{\lambda}$$

La figura 2 muestra la forma de $L(\lambda)$ y $l(\lambda)$. Puede apreciarse que ambas alcanzan su valor máximo cuando en el valor del parámetro es 4.

En un análisis real el estimador no se busca mirando la gráfica, aunque esta puede ser útil para obtener una idea cualitativa de la forma de la verosimilitud, sino utilizando el cálculo matemático. En concreto, debemos buscar el valor de λ que anula la primera derivada de $l(\lambda)$. La derivada de $l(\lambda)$ con respecto a λ es

$$l'(\lambda) = -\frac{3}{\lambda} + \frac{12}{\lambda^2}$$

El estimador máximo verosímil es el valor de λ que resuelve la ecuación de estimación

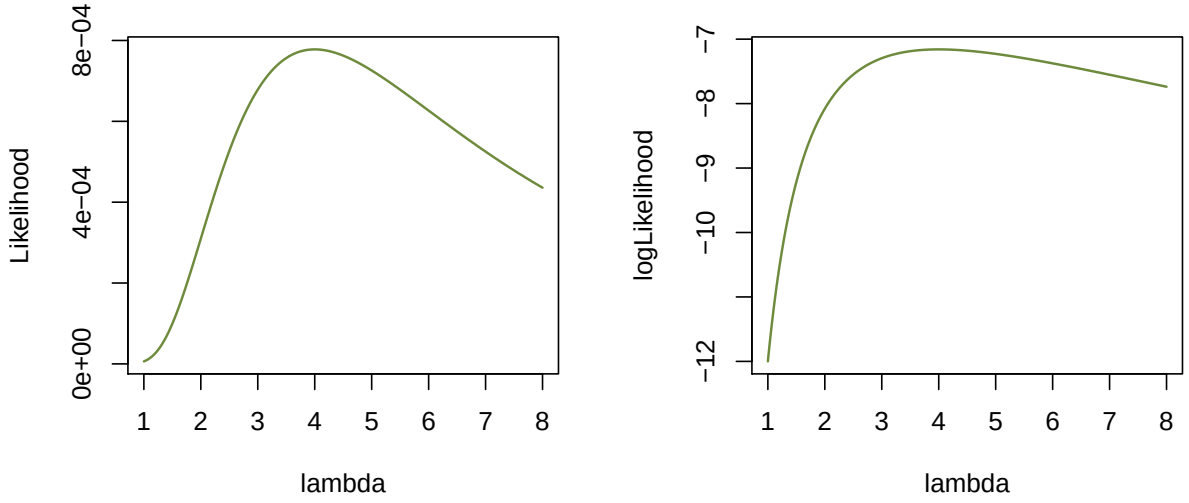

Figura 2: Función de verosimilitud y su logaritmo, distribución Weibull

$$-\frac{3}{\lambda}+\frac{12}{\lambda^2}=0$$
$$\hat{\lambda}=\frac{12}{3}=4.$$

Ejemplo 3. Vamos a obtener la fórmula genera del estimador máximo-verosímil de la distribución Weibull, sin concretar con los datos de una muestra particular. La función de log-verosimilitud es

$$l(\lambda)=-n\log\lambda-\frac{n\overline{X}}{\lambda}$$

Para encontrar el máximo de $l(\lambda)$ se toma su primera derivada con respecto a λ:

$$l'(\lambda)=-\frac{n}{\lambda}+\frac{n\overline{X}}{\lambda^2}.$$

Se denomina *ecuación de estimación* a la primera derivada de $l(\lambda)$ igualada a cero ($l'(\lambda)=0$), es decir

$$-\frac{n}{\lambda}+\frac{n\overline{X}}{\lambda^2}=0.$$

Despejando λ se obtiene el estimador máximo verosímil:

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

La derivada de una función toma el valor cero en los máximos pero también en otros puntos. Para comprobar que en un punto concreto hay un máximo puede utilizarse el criterio de la segunda derivada. Para que en $\hat{\lambda}$ exista un máximo debe cumplirse que $l''(\lambda)<0$. La segunda derivada es

$$l''(\lambda) = \frac{n}{\lambda^2} - \frac{2n\overline{X}}{\lambda^3}.$$

Sustituyendo el valor del estimado ($\hat{\lambda} = \overline{X}$) en la segunda derivada se obtiene:

$$l''(\overline{X}) = \frac{n}{\overline{X}^2} - \frac{2n\overline{X}}{\overline{X}^3} = \frac{n}{\overline{X}^2} - \frac{2n}{\overline{X}^2} = -\frac{n}{\overline{X}^2} < 0.$$

Cómo la variable X sólo toma valores positivos, $\overline{X} > 0$. Por tanto la segunda derivada es negativa cuando $\hat{\lambda} = \overline{X}$, lo que indica que $l(\lambda)$ tiene un máximo en este punto.

Ejemplo 4. Cómo parte de un entrenamiento, un deportista intenta realizar 25 veces una determinada prueba. El resultado de cada intento se clasifica como *éxito* o *fracaso*, y se considera que la probabilidad de éxito π permanece constante a lo largo del experimento. ¿Cuál es la probabilidad estimada de éxito asumiendo independencia entre las distintas realizaciones?

La variable X_i describe el resultado de la prueba i, y sigue la distribución de Bernoulli:

$$f(x_i; \pi) = \pi^{x_i}(1-\pi)^{(1-x_i)} \ .$$

La función de probabilidad del vector de resultados de las 25 ejecuciones es

$$\begin{aligned} f(\boldsymbol{x}; \pi) &= \prod_{i=1}^{25} \pi^{x_i}(1-\pi)^{(1-x_i)} \\ &= \pi^{\sum_{i=1}^{25} x_i}(1-\pi)^{(25-\sum_{i=1}^{25} x_i)} \\ &= \pi^{x}(1-\pi)^{(25-x)} \end{aligned}$$

donde x es el estadístico suficiente que representa el número de éxitos, $x = \sum_{i=1}^{25} x_i$. Entonces, la función de verosimilitud y su logaritmo son

$$\begin{aligned} L(\pi) &= \pi^{x}(1-\pi)^{(25-x)} \\ l(\pi) &= x \log \pi + (25-x)\log(1-\pi) \end{aligned}$$

El estimador máximo-verosímil es el valor de π que maximiza $l(\pi)$. Como hemos visto, en el punto máximo de $l(\pi)$, su derivada es cero:

$$l'(\pi) = \frac{x}{\pi} - \frac{25-x}{1-\pi} = 0 \ .$$

Despejando π en la ecuación de estimación encontramos

$$\begin{aligned} \frac{x}{\pi} &= \frac{25-x}{1-\pi} \\ x - x\pi &= 25\pi - x\pi \\ \hat{\pi} &= \frac{x}{25} \end{aligned}$$

Además necesitamos comprobar que la segunda derivada es negativa en el punto encontrado para asegurar que es un máximo de $l(\pi)$. Derivando $l'(\pi)$ se obtiene

$$l''(\pi) = -\frac{x}{\pi^2} - \frac{25-x}{(1-\pi)^2} \ .$$

La cual es necesariamente negativa, por lo que $l(\pi)$ alcanza un máximo en $\hat{\pi}$.

3. Máxima verosimilitud en lenguaje R

La estimación máximo-verosímil puede programarse en R utilizando las funciones *optimize* y *optim*. La función *optimize* sirve para buscar el máximo de una función univariante, es decir que depende de un solo parámetro. La función *optim* permite realizar la optimización univariante y multivariante, por lo que permite estimar modelos que dependen de uno o más parámetros.

3.1. Estimación de un parámetro

Vamos a ver este método en relación con el ejemplo sobre la distribución Weibull. Teníamos la muestra $\boldsymbol{x} = (2, 7, 3)'$, y la función log verosimilitud era

$$l(\lambda) = -3\log\lambda - \frac{12}{\lambda} \, .$$

En el ejemplo demostramos matemáticamente que el estimador es $\hat{\lambda} = 4$. A continuación utilizaremos R para obtener dicho estimador. Cómo la función de verosimilitud depende de un solo parámetro puede utilizarse la función *optimize*, a la cual le pasamos la función a maximizar y el rango de valores de λ en el que debe trabajar:

```
l <- function(lambda) -3*log(lambda) - 12/lambda
lMax <- optimize(l, c(0, 10), maximum=TRUE)
print(lMax)
```

```
## $maximum
## [1] 4.000008
##
## $objective
## [1] -7.158883
```

Podemos ver que R ha encontrado el estimador y que el máximo de la función de log verosimilitud es $l(\lambda = 4) \approx -7,16$.

También podemos realizar la estimación con *optim*. A esta función le pasamos varios argumentos, el primero es el valor valor inicial de λ a partir del cual comienza a buscar el estimador utilizando un procedimiento iterativo (en el ejemplo este valor es 1). A continuación, el argumento *f=l* le dice a *optim* que optimice la función *l*. El siguiente argumento, *Brent*, es el método de de cálculo que debe emplear. *lower* y *upper* son los valores mínimo y máximo de λ entre los que tiene que buscar el estimador. El argumento *control=list(fnscale=-1)* le dice a *opim* que debe maximizar la función, de no incluir este comando lo que haría *optim* sería buscar el mínimo. El resultado de *optim* se guarda en el objeto *fit*.

```
l <- function(lambda) -3*log(lambda) - 12/lambda
fit <- optim(1, f=l, method="Brent", lower=0, upper=10, control=list(fnscale=-1))
print(fit)
```

```
## $par
## [1] 4
##
```

```
## $value
## [1] -7.158883
##
## $counts
## function gradient
##       NA       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

El resultado de la estimación se recoge en dos elementos del objeto *fit*. En *fit$par* tenemos el valor del estimador y en *fit$value* el valor máximo de $l(\lambda)$, estos valores coinciden con los proporcionados por *optimize*.

3.2. Estimación de dos o más parámetros

Cuando una distribución depende de dos o más parámetros podemos realizar la estimación con *optimize*. Por ejemplo, hemos visto que la función de densidad normal de una muestra aleatoria simple es

$$f(\boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{nS_Y^2 + n(\overline{Y}-\mu)^2}{2\sigma^2}\right)$$

Los estimadores máximo-verosímiles de μ y σ^2 son los valores que maximizan la función de verosimilitud $L(\mu, \sigma^2) = f(\boldsymbol{y})$, cuyo logaritmo es

$$l(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{nS_Y^2 + n(\overline{Y}-\mu)^2}{2\sigma^2}$$

Si en una muestra de tamaño $n = 30$ tenemos $\overline{X} = 85$ y $s^2 = 16$ los estimadores $\hat{\mu}$ y $\hat{\sigma}^2$ los podemos obtener con el siguiente código, en el cual se han eliminado todas las constantes innecesarias de la función de log verosimilitud. Además se ha utilizado el comando *lower* para indicar cual es el valor más pequeño que puede tomar cada parámetro y se especifica que el método de cálculo numérico es "L-BFGS-B", que es el método que debe utilizarse cuando el rango de los parámetros está acotado.

```
n <- 30
media <- 85
varianza <- 16
l <- function(x){ - n*0.5*log(x[2]) - 0.5*(n*varianza + n*(media - x[1])^2)/x[2] }
fit <- optim(par=c(100, 20), fn=l, lower=c(-Inf, 0), method="L-BFGS-B", control=list(fnscale=-1))
print(fit)
```

```
## $par
## [1] 84.99987 15.99981
##
## $value
## [1] -56.58883
##
## $counts
## function gradient
##       15       15
##
## $convergence
```

```
## [1] 0
##
## $message
## [1] "CONVERGENCE: REL_REDUCTION_OF_F <= FACTR*EPSMCH"
```

3.3. Utilización de las funciones de probabilidad y densidad del lenguaje R

La estimación máximo verosímil puede realizarse aprovechando las funciones de probabilidad y densidad de probabilidad incluidas en R. De modo general, si una distribución depende de un vector de parámetros θ, la función de log-verosimilitud es

$$l(\theta) = \sum_{i=1}^{n} \log f(x_i)$$

donde $f(x_i)$ es la función de probabilidad o densidad de la observación X_i. Podemos programar la estimación utilizando las funciones estándar de R para definir $f(x_i)$ y sin tener que preocuparnos por los detalles internos de su formulación matemática.

Por ejemplo, la distribución gamma depende de dos parámetros y se utiliza en el estudio de tiempos de reacción. Supongamos que tenemos la siguiente muestra aleatoria y queremos ajustar la distribución gamma $\boldsymbol{t} = (5, 3, 7, 9, 1)$. El siguiente código realiza la estimación máximo-verosímil utilizando la función *dgamma* para obtener la función de densidad de cada una de las observaciones de la muestra. Como valores iniciales para el procedimiento iterativo de estimación de la función *optim*, utilizamos los parámetros estimados por el método de los momentos.

```
tiempo = c(5, 3, 7, 9, 1)

n <- length(tiempo)

l <- function(p){
    logLk <- 0
    for(i in 1:n){
        logLk <- logLk + log(dgamma(tiempo[i], p[1], p[2]))
    }
    return(logLk)
}

media <- mean(tiempo)
varianza <- var(tiempo)
alpha_Momentos <- media^2 / varianza
beta_Momentos <- media / varianza

fit <- optim(par=c(alpha_Momentos, beta_Momentos), fn=l, control=list(fnscale=-1))
print(fit)
```

```
## $par
## [1] 2.243039 0.448578
##
## $value
## [1] -12.29357
##
## $counts
## function gradient
##       53       NA
```

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

3.4. Estimación de modelos de regresión

Un modelo de regresión distingue entre una variable dependiente, Y, e independiente, X. Además, para utilizar máxima-verosimilitud es necesario asumir que la distribución de Y para un valor de X es conocida. Vamos a estimar una regresión de Poisson. Supongamos que disponemos de los siguientes datos

$$\boldsymbol{x} = (5, 2, 3, 7, 9, 8)$$
$$\boldsymbol{y} = (5, 4, 5, 22, 15, 18)$$

La regresión de Poisson asume que la distribución de Y_i es Poisson(λ_i), donde $\lambda_i = \exp(\alpha + \beta x_i)$. Es decir, la función de log-verosimilitud es

$$l(\alpha, \beta) = \sum_{i=1}^{n} \log P(Y_i)$$

donde $\log P(Y_i)$ es el logaritmo de la función de probabilidad de Poisson. Podemos implementar la estimación mediante el siguiente código R.

```r
x <- c(5, 2, 3, 7, 9, 8)
y <- c(5, 4, 5, 22, 15, 18)

n <- length(x)

l <- function(p){
    logLk <- 0
    for(i in 1:n){
        logLk <- logLk + log(dpois(y[i], exp(p[1] + p[2]*x[i])))
    }
    return(logLk)
}

fit <- optim(par=c(0, 0), fn=l, control=list(fnscale=-1))
print(fit)
```

```
## $par
## [1] 1.0164949 0.2244463
##
## $value
## [1] -16.43331
##
## $counts
## function gradient
##       67       NA
##
## $convergence
## [1] 0
```

```
##
## $message
## NULL
```

4. Ejercicios

Ejercicio 1. Se ha tomado una muestra de la distribución de Poisson: $\boldsymbol{y} = \{4, 2, 6, 4\}$.

1. Calcule el estimador máximo-verosímil matemáticamente.
2. Represente gráficamente las funciones de verosimilitud y log-verosimilitud utilizando R.
3. Obtenga el estimador máximo-verosímil utilizando R.

Ejercicio 2. Se ha tomado la siguiente una muestra de la distribución exponencial: $\boldsymbol{x} = \{2,5,\ 1,5,\ 1,25,\ 0,75\}$. Calcule el estimador máximo-verosímil matemáticamente y utilizando R.

Ejercicio 3. Sea $X = 0,\ 1, \ldots$ una variable aleatoria que sigue la distribución geométrica

$$f(x) = \pi\,(1-\pi)^x\ .$$

obtenga matemáticamente el estimador máximo-verosímil de π si la muestra es $X = 5$.

Ejercicio 4. Hemos tomado la siguiente muestra procedente de una distribución beta

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

Responda a los siguientes apartados utilizando R:

1. Obtenga el estimador máximo verosímil de α y β.
2. Calcule el valor esperado y la varianza de la distribución beta estimada en el punto 1.
3. Represente gráficamente la distribución beta estimada.

Ejercicio 5. Un psicólogo está analizando la curva de aprendizaje de una determinada tarea. Para ello ha recogido las variables $X \equiv$ horas de entrenamiento e $Y \equiv$ número de realizaciones correctas. Su hipótesis es que la relación entre ambas variables es curvilínea, de modo que con pocas horas de entrenamiento se aprende muy rápido pero después el aprendizaje se estabiliza. Por ello, quiere aplicar una regresión de Poisson de Y sobre X en la que el parámetro λ toma la forma

$$\lambda_i = \alpha x_i^{\beta}$$

Por tanto, el espacio paramétrico es $\alpha > 0$ y $\beta \in \mathbb{R}$. Los datos de que dispone el psicólogo son

$$\boldsymbol{x} = (1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10)$$
$$\boldsymbol{y} = (1,\ 2,\ 3,\ 4,\ 4,\ 5,\ 4,\ 6,\ 4,\ 5)$$

Responda a los siguientes apartados utilizando R.

1. Realice la estimación máximo-verosímil de α y β.
2. Obtenga el diagrama de dispersión de X e Y y superponga la curva que representa los pronósticos.