Estadística [continuación]

Santander, 2019-2020

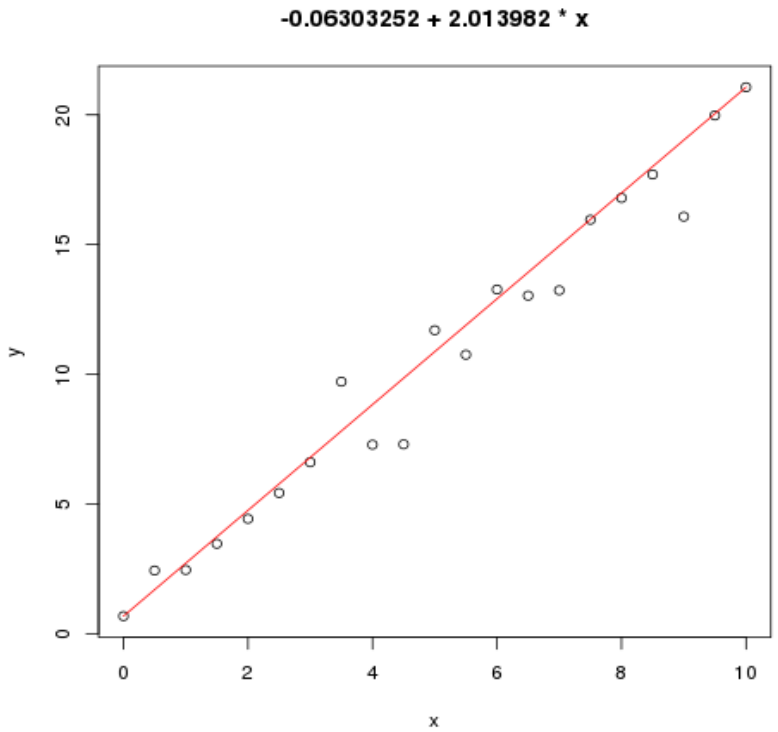
El concepto de máxima verosimilitud (I)

- En las clases previas hemos basado la estimación de parámetros en el concepto de función de loss.
- La función de loss típicamente crece cuando el modelo no se ajusta bien a los datos.
- Consideremos ahora una vez más el caso paradigmático de la regresion lineal:

Modelo

$$p(y|x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2}\frac{(y-2x)^2}{\sigma^2}}$$

Término estocástico



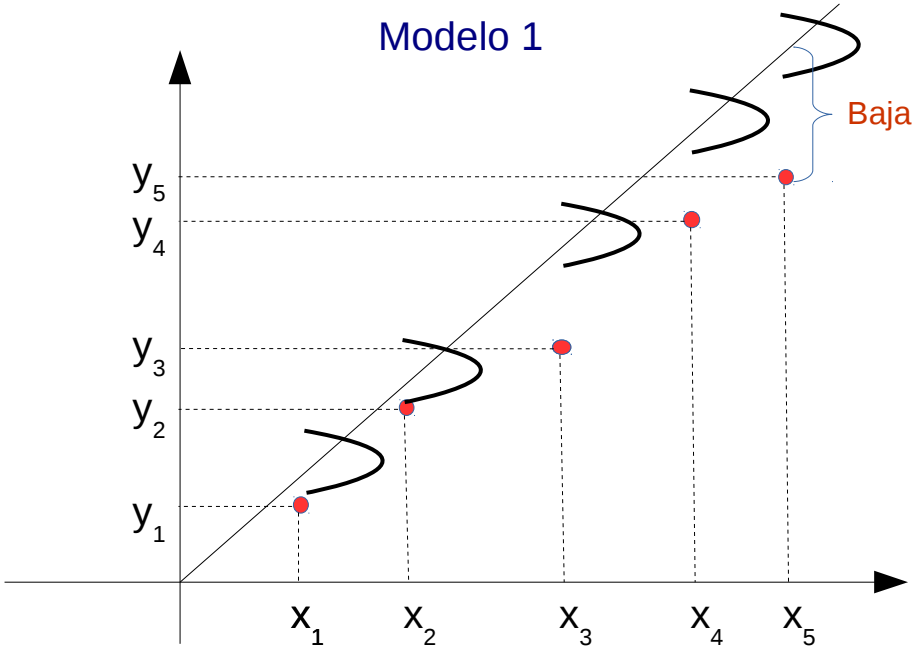
- Es decir, disponemos de una serie de puntos que cumplen y = 2x con un término estocástico.

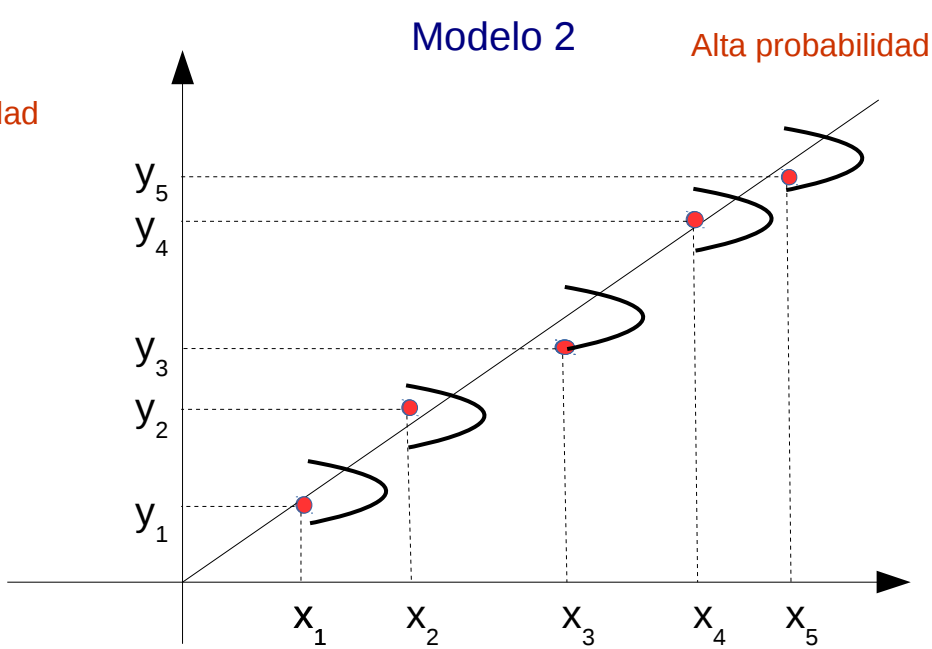
El concepto de máxima verosimilitud (II)

- El término estocástico viene dado por una pdf centrada en el valor dado por el modelo.
- ¿Cuál es la probabilidad de que dado un modelo concreto (parámetros) obtenga la medida y_i?
- Veamos lo que nos dice la intuición:

$$p(y|x)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-1}{2}\frac{(y_5-3x_5)^2}{\sigma^2}}$$

$$p(y|x)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-1}{2}\frac{(y_5-2x_5)^2}{\sigma^2}}$$





La probabilidad de que el punto tenga el valor dado es muy baja dadas estas pdf.

La probabilidad de que el punto tenga el valor dado es relativamente alta dadas estas pdf.

El concepto de máxima verosimilitud (III)

- ¿Cuál es la probabilidad de que dado un model concreto obtenta todas las medidas y_i?
- Asumiendo medidas independientes, dicha probabilidad será el producto de probabilidades.
- Si en general tenemos un modelo y = f(x; theta) entonces la probabilidad de encontrar los datos y_i, para unos x_i y un theta concreto es:

$$p((y_i, x_i); \theta) = \prod_j p(y_i|x_i) = \prod_j \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2}\frac{(y_i - f(x_i;\theta))^2}{\sigma^2}}$$

- Esta probabilidad es conocida como el Likelihood de la muestra de medidas para ese modelo.
- Resulta interesante darse cuenta de que en realidad funciona para cualquier pdf que sigan los datos.
- Por lo tanto permite sacar conclusiones en casos que no son gaussianos.

$$L(\theta; (y_i, x_i)) = \prod_j pdf(y_i|x_i)$$

El concepto de máxima verosimilitud (IV)

- Esta nueva cantidad se comporta de manera parecida a como lo hace una función de coste ya que:
  - Si las medidas no se ajustan al modelo las probabilidades serán bajas y el likelihood también.
  - Si las medidas sí se ajustan al modelo las probabilidades serán altas y el likelihood también.
- Podemos encontrar los parámetros que nos proporcionan el modelo maximizando el likelihood.

$$\nabla_\theta L(\theta;(y_i,x_i))=0(máximo)$$

- Usualmente en lugar de maximizar esa cantidad lo que se hace es minimizar la cantidad "q" o "l".

$$q=-2\log(L(\theta;(y_i,x_i)))=-2\sum_j \log(pdf(y_i|x_i);\theta)$$
$$l=-\log(L(\theta;(y_i,x_i)))=-\sum_j \log(pdf(y_i|x_i);\theta)$$

- Puesto que el logaritmo es una función creciente, el máximo de L corresponderá al mínimo de "q".

$$\nabla_\theta q(\theta;(y_i,x_i))=0(mínimo)$$

Aplicación a una regresión lineal (I)

- Apliquemos esta técnica a nuestro con datos gaussianos en donde:

$$p((y_i, x_i); \theta) = \prod_j pdf(y_i | x_i) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2}\frac{(y_i - f(x_i;\theta))^2}{\sigma^2}}$$

- En este caso recuperamos el valor de "q" es:

$$q = -2\log(L(\theta; (y_i, x_i))) = 2N\log(\sqrt{2\pi}\sigma) + \sum_i \left(\frac{(y_i - x_i^T\theta)}{\sigma}\right)^2$$

Término constante

Función de Loss de la regresion.

- Cuando las pdf de los datos son gaussianas recuperamos casi la función de Loss de la regresion.
- Sin embargo, ahora nos aparece el término de error sigma en la fórmula.
- Ahora la gaussiana no tiene por qué ser igual en todos los casos (diferente sigma):

$$\sum_i \left(\frac{(y_i - x_i^T\theta)}{\sigma_i}\right)^2$$

Aplicación a una regresión lineal (II)

- Dicha expresión puede también expresarse matricialmente de la siguiente forma:

$$\sum_i \left(\frac{(y_i - x_i^T \theta)}{\sigma_i} \right)^2 = (Y - X\theta)^T Cov^{-1} (Y - X\theta)$$

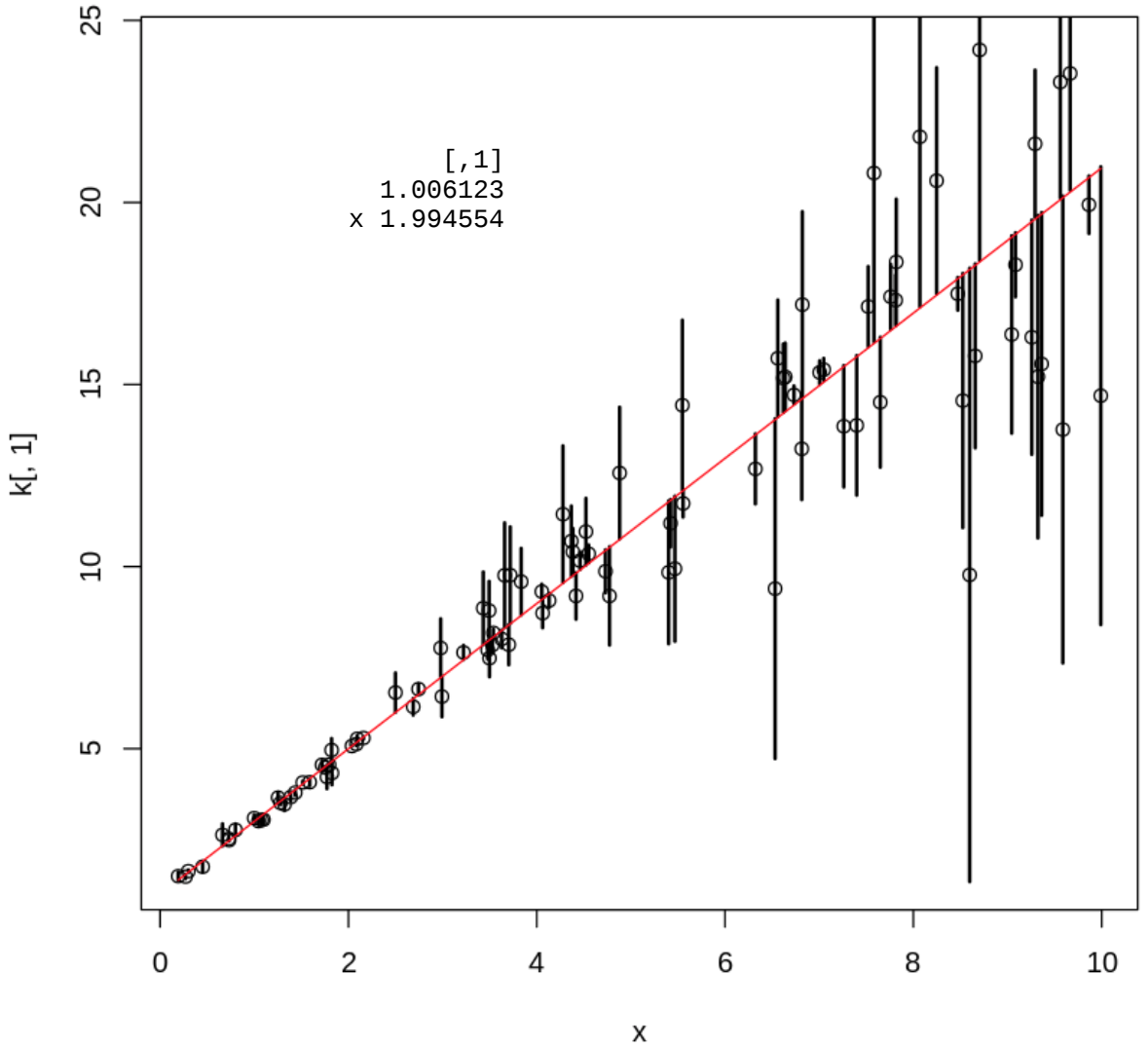
- En donde Y es el vector con las medidas independientes y X la matriz de características.
- C es la matriz de covarianza que tendrá la forma:

$$Cov(y) = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_N^2 \end{bmatrix}$$

- La estructura es esencialmente la misma que para la regresión linea.
- Minimizar q puede ser hecho analíticamente dando el resultado:

$$\theta = (X^T Cov^{-1} X)^{-1} (X^T Cov^{-1}) Y$$

Aplicación a una regresión lineal (III)



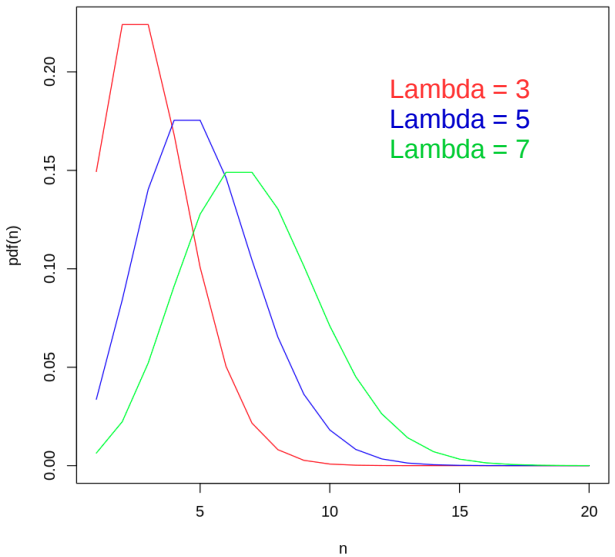
Aplicación al ajuste de un histograma (I)

- La técnica del maximum likelihood nos permite considerar procesos no gaussianos.
- Un buen ejemplo de este tipo de proceso es precisamente el ajuste a un "histograma".
- Un histograma es un conjunto de "bins" con el número de sucesos encontrados en cada segmento.
- El contaje de sucesos en un segmento de una magnitud viene dado por la distribución de Poisson.
  - El número de gotas que caen de un grifo mal cerrado en 10 segundos.
  - El número de desintegraciones de un núcleo observadas por minuto.
  - El número de errores de ortografía que uno comete en una página.
- Lo que tienen en común es que la probabilidad de que el suceso ocurra es constante.

$$pdf(n,\lambda) = \frac{e^{-\lambda}\lambda^{N}}{n!}$$

$$\langle n \rangle = \sum n\, pdf(n,\lambda) = \lambda$$

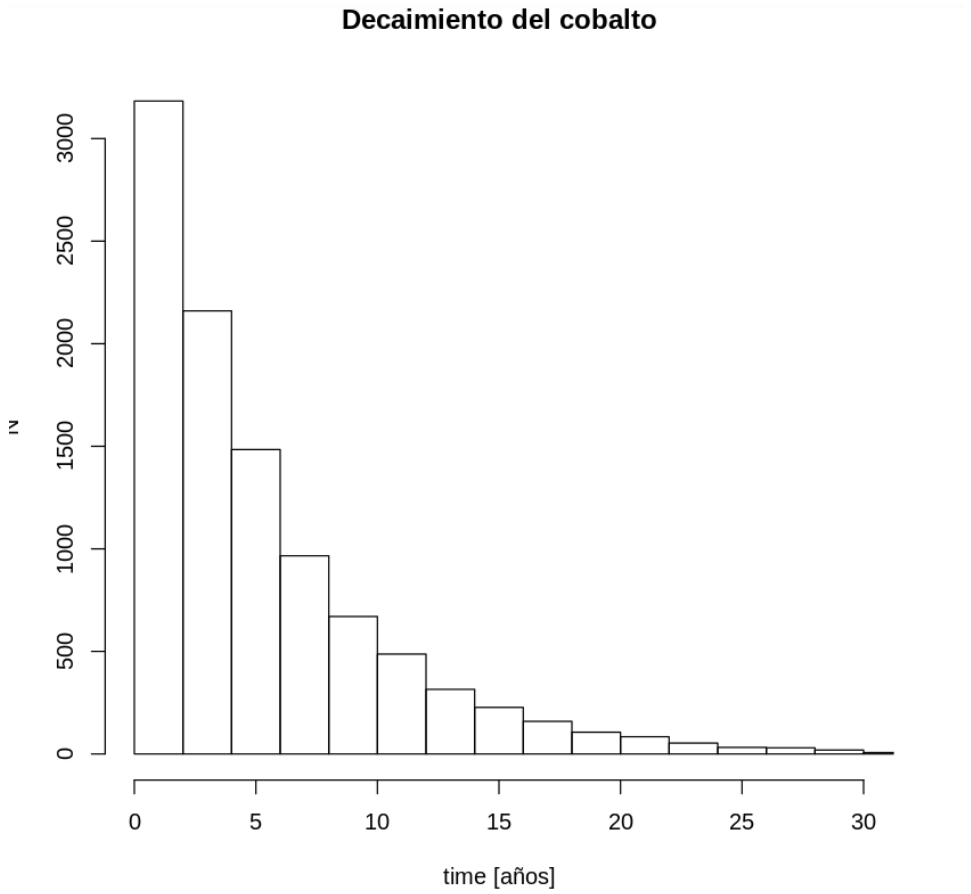
$$var(n) = \sum (n-\lambda)^2\, pdf(n,\lambda) = \lambda$$



Aplicación al ajuste de un histograma (II)

- En un histograma disponemos de muchos bines siendo posible que exista una ley subyacente.
- Ejemplo: Número de partículas que quedan en una muestra radioactiva en función del tiempo.
- La pdf en cada bin viene dada por:

$$pdf\left(n_i\right) = poisson\left(n_i, \lambda = Ne^{-t/\tau}\right)$$



Aplicación al ajuste de un histograma (III)

- De esta forma podemos calcular el likelihood asociado a esta distribución como:

$$L(N,\tau;(y_i,x_i))=\prod_j poisson(n_j, Ne^{-t/\tau})$$

- E igualmente podemos calcular el valor de "q" o de "l"

$$q(N,\tau;(y_i,x_i))=-2\sum_j \log(poisson(n_j, Ne^{-t/\tau}))$$

$$l(N,\tau;(y_i,x_i))=-\sum_j \log(poisson(n_j, Ne^{-t/\tau}))$$

- Minimizando estas funciones podríamos encontrar valores de N y τ que mejor ajustan el histograma.
- Estaríamos teniendo en cuenta de forma intrínseca a la propia pdf generadora del proceso.
- Sin embargo este cálculo no puede hacerse de manera análitica.
- Una posible opción sería utilizar la función "optim" para miniminzar esta cantidad.

La función mlm

mle {stats4}

Maximum Likelihood Estimation

Description

Estimate parameters by the method of maximum likelihood.

Usage

```
mle(minuslogl, start = formals(minuslogl), method = "BFGS",
    fixed = list(), nobs, ...)
```

Arguments

`minuslogl`
: Function to calculate negative log-likelihood.

`start`
: Named list. Initial values for optimizer.

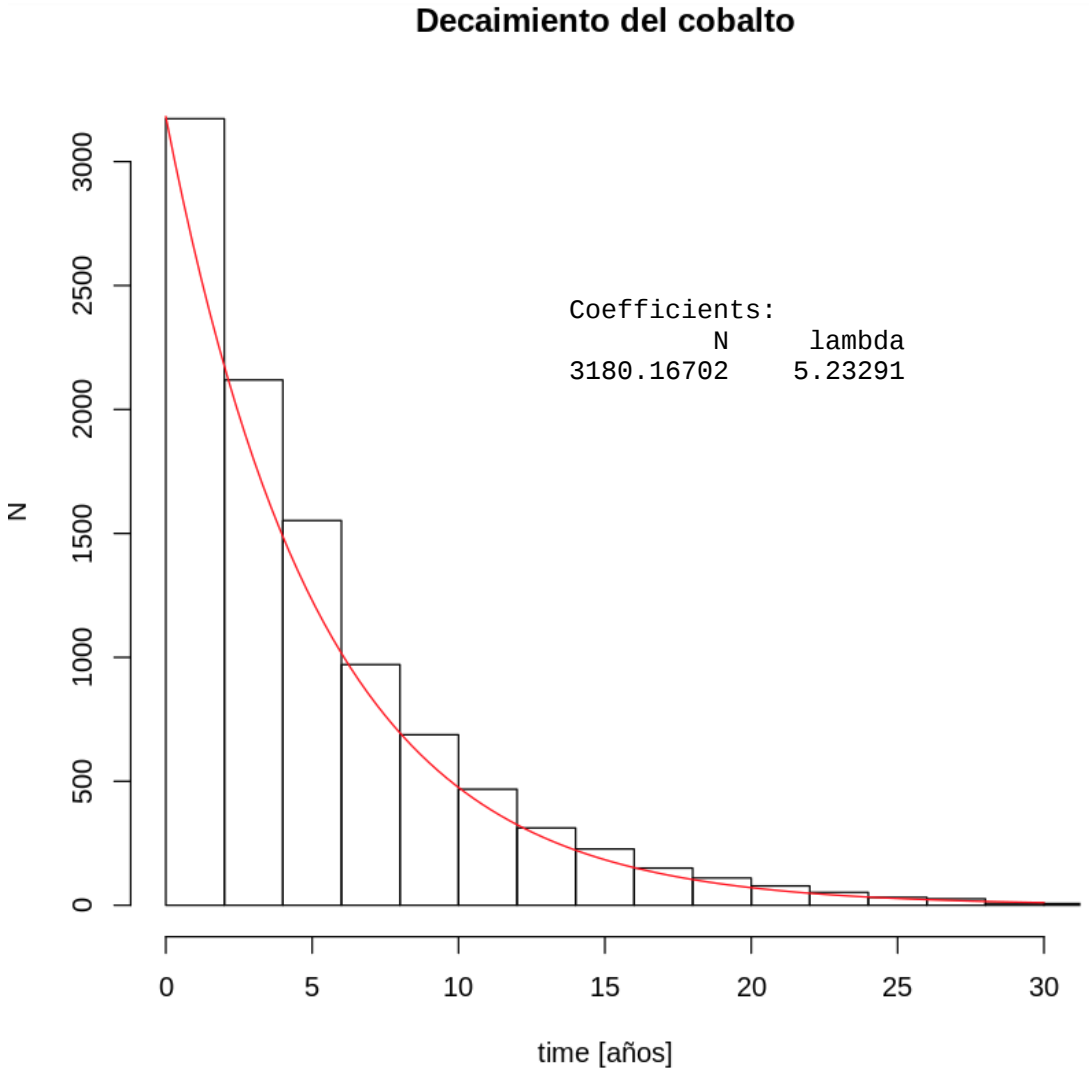
`method`
: Optimization method to use. See `optim`.

`fixed`
: Named list. Parameter values to keep fixed during optimization.

`nobs`
: optional integer: the number of observations, to be used for e.g. computing `BIC`.

`...`
: Further arguments to pass to `optim`.

La función mlm



Ejercicio 6

1) Crea una función a la que se pase como input: un vector “x” de features distribuidos uniformemente, unos valores “a” y “b” parámetros de un modelo lineal “y = a + b * x”, y parámetros “m” y “n” que nos den la sigma de una distribución gaussiana: “sigma = m + n * x^2”. La función debe devolver una matriz que contenga en la primera columna un vector con el término independiente “y = a + b * x + gauss(0, sigma = m + n * x^2)”, y en la segunda columna la “sigma = m + n * x^2”. Nota: Este ejercicio es igual al que ya realizamos, salvo porque ahora la sigma del término estocástico depende de cada punto.
2) Crea una función que encuentre el valor de los parámetros que hace máximo el ML usando la fórmula analítica.
3) Crea un vector x distribuido uniformemente en [0, 10], y encuentra el mínimo para a = 1, b = 2, m = 0.1 y n = 0.04. Pinta los datos, sus errores y la recta de ajuste.
4)
5) Crea una función a la que se le pase como input: la media en el eje x, la media en el eje y, la varianza en el eje x, la varianza en el eje y y la covarianza de x e y, junto con un número de puntos N, y devuelva una matriz con N filas y 2 columnas con los números que salen de la distribución gaussiana de dos dimensiones definidas por los valores de input (usar la función MASS::mvrnorm)
6) Genera una matrix x1 usando la función anterior y tomando: N = 1000, mu_x = 2, mu_y = 4, var_x = var_y = 1, y Cov(x,y) = 0.3. Crea una matriz “y1” con tantas filas como la matriz x y asígnale el valor 0.
7) Repite 2) para otra muestra con N = 1000, mu_x = 6, mu_y = 3, var_x = var_y = 1, y Cov(x,y) = 0.3.