Estimación puntual

1 Introducción

Teoría de probabilidad	Estadística
$\downarrow$	$\downarrow$
Parámetros conocidos	A partir de los datos Inferencia sobre los parámetros desconocidos

Una vez que se estimaron los parámetros, el modelo debe ser comparado con la información dada por los datos para ver si es razonable.

$\mathbf{X} = (X_1, \ldots, X_n) \sim F$,

$F \in \mathcal{F} = \{F(x_1, \ldots, x_n, \boldsymbol{\theta}) \text{ donde } \boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \Theta \subset IR^p\}$

Interesa conocer *aproximadamente* $q(\boldsymbol{\theta})$, donde $q(\cdot) : \Theta \to IR$.

Definición. Un *estimador puntual* de $q(\boldsymbol{\theta})$ será cualquier función $\delta(\mathbf{X})$ de las observaciones, o sea, una función $\delta : IR^n \to IR$ que se usa para conocer el valor de $q(\boldsymbol{\theta})$.

Por ejemplo, en el caso de una $N(\mu, \sigma^2)$ uno puede querer estimar $q_1(\mu, \sigma^2) = \mu$ o $q_2(\mu, \sigma^2) = \sigma^2$.

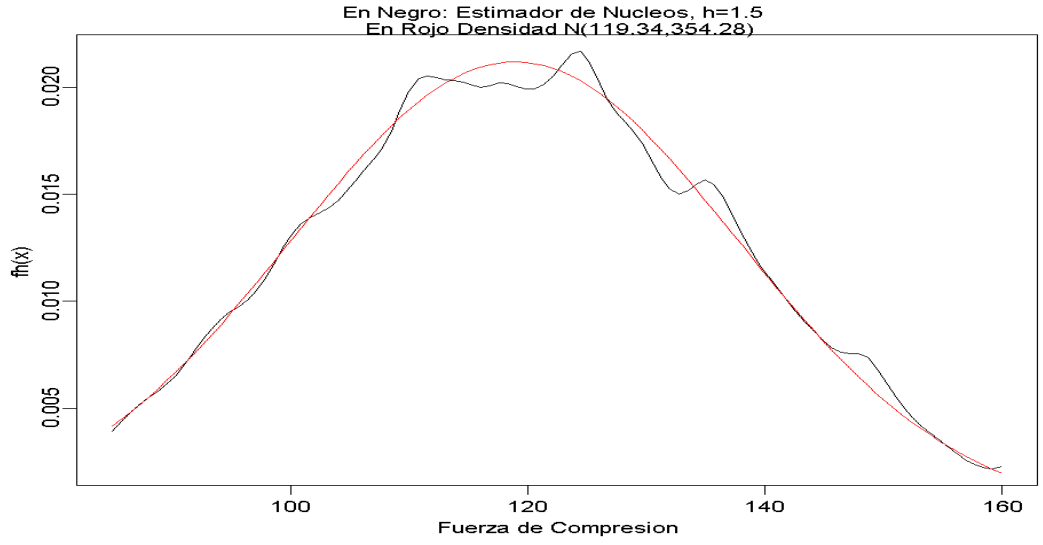
Usualmente X_i son i.i.d. con densidad $f(x, \boldsymbol{\theta})$ y por lo tanto, su densidad conjunta será $\Pi_{i=1}^n f(x_i, \boldsymbol{\theta})$.

Un estimador de $\boldsymbol{\theta}$, o más generalmente de $q(\boldsymbol{\theta})$ será una función de $X_1, \ldots, X_n$ y por lo tanto, será una variable aleatoria con una distribución llamada *distribución muestral*.

Daremos aproximaciones a la distribución muestral para conocer la variabilidad del estimador que se conoce usualmente con el nombre de *error estándar*.

Un buen estimador $\delta(\mathbf{X})$ deberá tener la propiedad de que cualquiera sea el valor de $\boldsymbol{\theta}$, que es desconocido, la diferencia $\delta(\mathbf{X}) - q(\boldsymbol{\theta})$ sea pequeña. En qué sentido esta diferencia es pequeña será especificado más adelante.

Ejemplo 1. Fuerza de compresión de 45 Muestras de Aleaciones de Aluminio-Litio. El conjunto original tenía 4152 observaciones. El siguiente gráfico muestra un estimador de la densidad basado en núcleos con una abertura de ventana igual a 1.5 al que se sobreimpuso en rojo la densidad normal.

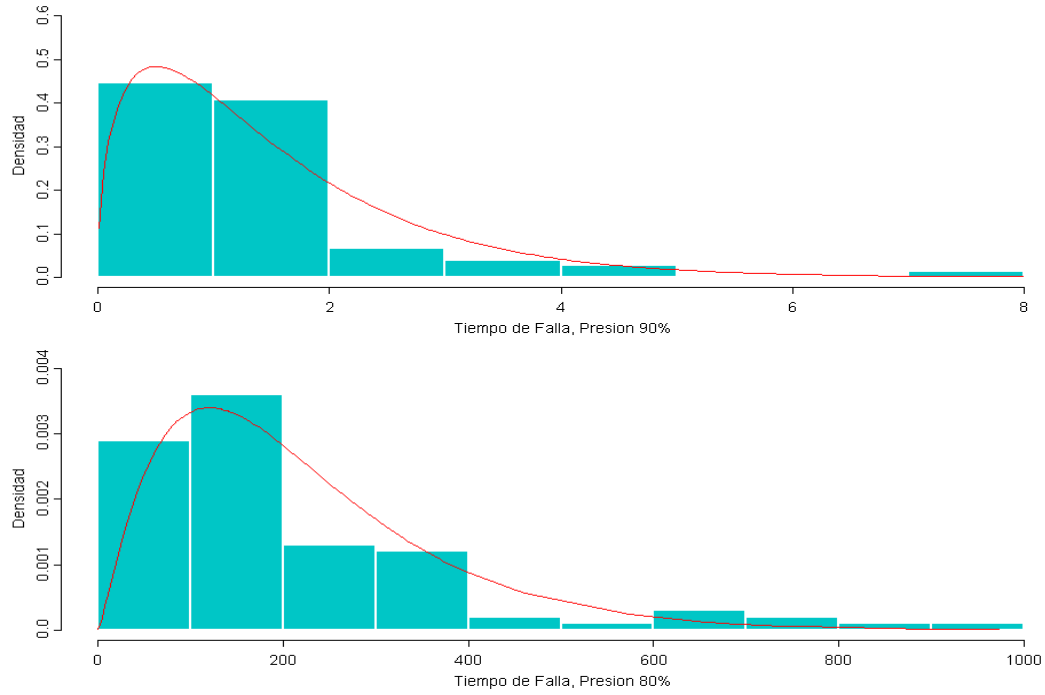


Ejemplo 2. Tiempo de falla de cables Kevlar 49/epoxy sometidos a presión sostenida.

90% de Presión							
0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.04
0.05	0.06	0.07	0.07	0.08	0.09	0.09	0.10
0.10	0.11	0.11	0.12	0.13	0.18	0.19	0.20
0.23	0.80	0.80	0.83	0.85	0.90	0.92	0.95
0.99	1.00	1.01	1.02	1.03	1.05	1.10	1.10
1.11	1.15	1.18	1.20	1.29	1.31	1.33	1.34
1.40	1.43	1.45	1.50	1.51	1.52	1.53	1.54
1.54	1.55	1.58	1.60	1.63	1.64	1.80	1.80
1.81	2.02	2.05	2.14	2.17	2.33	3.03	3.03
3.24	4.20	4.69	7.89				

80% de Presión									
0.18	3.1	4.2	6	7.5	8.2	8.5	10.3	10.6	24.2
29.6	31.7	41.9	44.1	49.5	50.1	59.7	61.7	64.4	69.7
70	77.8	80.5	82.3	83.5	84.2	87.1	87.3	93.2	103.4
104.6	105.5	108.8	112.6	116.8	118	122.3	123.5	124.4	125.4
129.5	130.4	131.6	132.8	133.8	137	140.2	140.9	148.5	149.2
152.2	152.8	157.7	160	163.6	166.9	170.5	174.9	177.7	179.2
183.6	183.8	194.3	195.1	195.3	202.6	220	221.3	227.2	251
266.5	267.9	269.2	270.4	272.5	285.9	292.6	295.1	301.1	304.3
316.8	329.8	334.1	346.2	351.2	353.3	369.3	372.3	381.3	393.5
451.3	461.5	574.2	656.3	663	669.8	739.7	759.6	894.7	974.9

El gráfico siguiente muestra los histogramas de estas observaciones a los que se sobreimpuso un ajuste de la densidad $\Gamma(\alpha, \lambda)$.



Diferencias entre las distribuciones entre los tiempos de falla a presión sostenida del 90% y 80% pueden describirse a través de diferencias en los parámetros.

2 Propiedades asintóticas de un estimador

2.1 Consistencia de estimadores

Una propiedad deseable para un estimador $\widehat{\boldsymbol{\theta}}_n$ de $\boldsymbol{\theta}$, es que cuando n es grande la sucesión $\widehat{\boldsymbol{\theta}}_n$ se aproxime en algún sentido a $\boldsymbol{\theta}$.

Sea $\mathcal{F} = \{F(x, \boldsymbol{\theta}) \text{ con } \boldsymbol{\theta} \in \Theta\}$ una familia de distribuciones y supongamos que para cada n se tiene un estimador $\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_n(X_1, \ldots, X_n)$ de $\boldsymbol{\theta}$ basado en una muestra aleatoria, $X_1, \ldots, X_n$, de tamaño n.

Definición 1. $\widehat{\boldsymbol{\theta}}_n$ es una *sucesión fuertemente consistente de estimadores* de $\boldsymbol{\theta}$ si

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow{c.t.p.} \boldsymbol{\theta}$$

o sea si $P_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}) = 1$ para todo $\boldsymbol{\theta} \in \Theta$.

Definición 2. $\widehat{\boldsymbol{\theta}}_n$ es una *sucesión débilmente consistente de estimadores* de $\boldsymbol{\theta}$ si

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}$$

Es decir, para todo $\varepsilon > 0$ y $\boldsymbol{\theta} \in \Theta$

$$\lim_{n\to\infty} P_{\boldsymbol{\theta}}(|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}| > \varepsilon) = 0\,.$$

Observación 1. Puesto que convergencia en c.t.p. implica convergencia en probabilidad, entonces toda sucesión fuertemente convergente también lo será débilmente.

Ejemplo. Sea $X_1, \ldots, X_n$ una muestra aleatoria de una función de distribución $F(x)$ totalmente desconocida, tal que $E_F(X_1)$ existe. Supongamos querer estimar $\mu = E_F(X_1)$. Si $\widehat{\mu} = \overline{X}_n$, por la ley

fuerte de los grandes números este estimador resulta fuertemente consistente para μ.

Observación 2. Si $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución $N(\mu, \sigma^2)$ se tiene que $\overline{X}_n$ es fuertemente consistente para μ y s_n^2 es fuertemente consistente para σ^2.

El siguiente Teorema nos da una condición suficiente para que una sucesión de estimadores sea débilmente consistente.

Teorema 1. *Sea $\hat{\theta}_n$ un estimador de θ basado en una muestra aleatoria de tamaño n. Si $Var_\theta(\hat{\theta}_n) \to 0$ y $E_\theta(\hat{\theta}_n) \to \theta$, entonces $\hat{\theta}_n$ es débilmente consistente.*

2.2 Estimadores asintóticamente normales

Definición 1. Se dice que $\hat{\theta}_n$ es una sucesión de *estimadores asintóticamente normal* si $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} N(0, \sigma^2)$. σ^2 se llama la varianza asintótica.

Dados dos estimadores, $\hat{\theta}_n^1$ y $\hat{\theta}_n^2$ asintóticamente normales con varianzas asintóticas σ_1^2 y σ_2^2 respectivamente, se define la eficiencia asintótica de $\hat{\theta}_n^2$ respecto de $\hat{\theta}_n^1$ como el cociente $\frac{\sigma_1^2}{\sigma_2^2}$.

3 Criterios para medir la bondad de un estimador

Para poder elegir el estimador $\delta(\mathbf{X})$ de $q(\theta)$ que se utilizará, se deberá dar un criterio para comparar dos estimadores cualesquiera.

La función de pérdida cuadrática fue la primera utilizada en Estadística, y aún hoy la más difundida. La pérdida media, o riesgo, correspondiente está dada por

$$R(\delta, \theta) = E(\delta(\mathbf{X}) - q(\theta))^2$$

y será llamada en adelante error cuadrático medio, e indicada por $\mathrm{ECM}_\theta(\delta)$. Luego

$$\mathrm{ECM}_\theta(\delta) = R(\delta, \theta) = E_\theta(\delta(\mathbf{X}) - q(\theta))^2 \tag{1}$$

Un estimador $\delta_1(\mathbf{X})$ de θ es mejor que $\delta_2(\mathbf{X})$ si

$$\mathrm{ECM}_\theta(\delta_1) \leq \mathrm{ECM}_\theta(\delta_2) \qquad \forall\, \theta \in \Theta$$

Diremos que δ^* es un estimador óptimo si para cualquier otro estimador δ se tiene

$$\mathrm{ECM}_\theta(\delta^*) \leq \mathrm{ECM}_\theta(\delta) \qquad \forall\, \theta \in \Theta \tag{2}$$

Salvo en casos triviales no existirán tales estimadores óptimos. Por lo tanto, una manera de obtener estimadores óptimos consistirá en restringir primero la clase de los estimadores δ considerados, y luego buscar aquél con error cuadrático medio uniformemente menor dentro de esta clase. Otra forma de obtener estimadores óptimos consistirá en minimizar algún criterio general basado en la función de riesgo, como el máximo riesgo o el riesgo bayes.

3.1 Estimadores insesgados

Una propiedad "razonable" que se puede exigir a un estimador está dada por la siguiente definición que hemos mencionado anteriormente.

Definición 1. Se dice que $\delta(\mathbf{X})$ es un *estimador insesgado* para $q(\theta)$ si $E_\theta(\delta(\mathbf{X})) = q(\theta) \quad \forall \theta \in \Theta$.

Definición 2. Si un estimador no es insesgado, se dice *sesgado*, definiéndose el *sesgo* del estimador como $b(\theta) = E_\theta(\delta(\mathbf{X})) - q(\theta)$.

Cuando $\delta(\mathbf{X})$ es un estimador insesgado, su ECM coincide con su varianza ya que

$$\mathrm{ECM}_\theta(\delta) = E_\theta\left[(\delta(\mathbf{X}) - q(\theta))^2\right] = E_\theta\left[(\delta(\mathbf{X}) - E_\theta(\delta(\mathbf{X})))^2\right] = \mathrm{Var}_\theta(\delta(\mathbf{X})).$$

Ejemplo H. Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de F con $E_F(|X|) < \infty$ y supongamos que se quiere estimar $\mu = E_F(X)$. Un posible estimador para μ es $\overline{X} = (1/n)\,\Sigma_{i=1}^n X_i$. El estimador $\overline{X}$ es insesgado ya que

$$E_F(\overline{X}) = \frac{1}{n}\sum_{i=1}^n E_F(X_i) = E_F(X) = \mu$$

Ejemplo J. $X \sim F$ tal que $E_F(X^2) < \infty$. Supongamos que se quiere estimar $\sigma^2 = \mathrm{Var}_F(X)$ a partir de una muestra aleatoria $X_1, X_2, \ldots, X_n$. Ya hemos visto que un estimador adecuado podría ser

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$$

que en el caso de observaciones normales $\widehat{\sigma}^2$ es el estimador de máxima verosimilitud.

Sin embargo, $\widehat{\sigma}^2$ no es un estimador insesgado de σ^2 ya que

$$E_F(\widehat{\sigma}^2) = \frac{n-1}{n}\sigma^2 \ ,$$

aunque el sesgo tiende a 0 cuando n tiende a infinito. El sesgo puede corregirse dividiendo $\widehat{\sigma}^2$ por $(n-1)/n$, obteniendo así el estimador insesgado

$$s^2 = \frac{n}{n-1}\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \ .$$

Si nos restringimos a la clase de los estimadores insesgados, se podrá encontrar frecuentemente, estimadores óptimos.

Definición 2. Se dirá que $\delta(\mathbf{X})$ es un *estimador insesgado de mínima varianza* para $q(\theta)$, uniformemente en $\theta \in \Theta$ (IMVU) si:

(a) $\delta(\mathbf{X})$ es insesgado para $q(\theta)$

(b) dado otro estimador insesgado para $q(\theta)$, $\delta^*(\mathbf{X})$, se cumple $\mathrm{Var}_\theta(\delta(\mathbf{X})) \leq \mathrm{Var}_\theta(\delta^*(\mathbf{X})) \quad \forall\, \theta \in \Theta$.

Sin embargo, en algunos casos, un estimador IMVU puede ser mejorado en su error cuadrático medio por otro estimador no insesgado.

3.2 Consistencia de estimadores IMVU

El siguiente teorema muestra que si $\delta_n(X_1, \ldots, X_n)$ es una sucesión de estimadores IMVU para $q(\theta)$ entonces cumple la hipótesis del Teorema 1.

Teorema 2: *Sea $\delta_n(X_1, \ldots, X_n)$ una sucesión de estimadores IMVU para $q(\theta)$, donde $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución perteneciente a la familia $F(x, \theta)$, $\theta \in \Theta$. Luego $Var_\theta(\delta_n(X_1, \ldots, X_n)) \to 0$ si $n \to \infty$.*

Corolario 1: *Si $\delta_n(X_1, \ldots, X_n)$ es una sucesión de estimadores IMVU para θ donde $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución perteneciente a la familia $\mathcal{F} = \{F(x, \theta)$ con $\theta \in \Theta\}$ entonces $\delta_n(X_1, \ldots, X_n)$ es una sucesión de estimadores débilmente consistentes.*

4 Método de los momentos

Definición 1. Se define el momento de orden k de una distribución F como $\mu_k = E\left(X^k\right)$ donde suponemos que la esperanza existe y que $X \sim F$.

Definición 2. Dadas $X_1, \ldots, X_n$ i.i.d. $X_i \sim F$, se llama momento muestral de orden k a

$$\widehat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k \ .$$

$\widehat{\mu}_k$ es un estimador de μ_k y por la L.G.N. $\widehat{\mu}_k \xrightarrow{c.t.p.} \mu_k$.

El método de los momentos estima parámetros buscando expresiones de ellos en términos de los momentos de menor orden posible y luego sustituye los momentos por los momentos muestrales en esas expresiones.

4.1 Caso de un parámetro

Sea $g : IR \to IR$, luego el método de los momentos estima θ, por el valor $\hat{\theta} = \delta(\mathbf{X})$ que satisface la ecuación

$$\frac{1}{n} \sum_{i=1}^{n} g(X_i) = E_{\hat{\theta}}(g(X_1)), \tag{3}$$

donde $E_\theta(X)$ significa la esperanza de X cuando $X \sim F(x, \theta)$.

$g(x) = x^k$ para algún $k \geq 1$.

Justificación heurística: Por la ley de los grandes números si $E_\theta\,|g(X_1)| < \infty$

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{c.t.p.} E_\theta(g(X_1))\ ,$$

y por lo tanto, si θ puede expresarse como una función continua de $E_\theta(g(X_1))$, se puede esperar que cuando n es grande el valor $\hat{\theta}$ que satisface la ecuación (3) estará cerca de θ.

Ejemplo A. Sean $X_1, X_2, \ldots, X_n$ i.i.d. $X_i \sim N(\mu, 1)$. Con $g(x) = x$

$$\frac{1}{n}\sum_{i=1}^{n} X_i = E_{\widehat{\mu}}(X_1) = \widehat{\mu}\ .$$

$\widehat{\mu} = (1/n)\,\Sigma_{i=1}^{n}\, X_i$ es el estimador de μ resultante y sabemos que la $\widehat{\mu} \sim N(\mu, \frac{1}{n})$

Ejemplo B. Sea $X_1, X_2, \ldots, X_n$ i.i.d. $X_i \sim N(0, \sigma^2)$. Con $g(x) = x^2$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = E_{\widehat{\theta}}(X_1^2) = \widehat{\sigma}^2\ .$$

$\widehat{\sigma}^2 = \delta(X_1, \ldots, X_n) = (1/n)\,\Sigma_{i=1}^{n}\, X_i^2$ es el estimador de σ^2 resultante. Por otra parte, tenemos que $n\,\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi_n^2$.

Ejemplo C. Sea $X_1, X_2, \ldots, X_n$ i.i.d. $X_i \sim \mathcal{P}(\lambda)$, El estimador de los momentos resultantes usando el momento de orden 1 resulta

$$\hat{\lambda}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i\ .$$

El estimador de los momentos resultantes usando el momento de

orden 2 es

$$\hat{\lambda}_2 = \delta_2(X_1, X_2, \ldots, X_n) = -\frac{1}{2} + \sqrt{\frac{1}{4} + \sum_{i=1}^{n} \frac{X_i^2}{n}}$$

Luego observamos que eligiendo distintas funciones g, obtenemos diferentes estimadores. Todavía no estamos en condiciones de comparar uno con otro.

Estudio hecho en el Instituto Nacional de Ciencia y Tecnología de U.S.A. Fibras de asbesto en filtros. Se tomaron muestras de 3 mm de diámetro del filtro y mediante un microscopio electrónico, se contaron el número de fibras en cada una de las 23 grillas obteniéndose

31	29	19	18	31	28	34	17	27	34	30	16
18	26	27	24	27	18	24	22	28	24	21	

$$\hat{\lambda}_1 = 24.913 \qquad \hat{\lambda}_2 = 24.988$$

Si repitieramos la experiencia, los resultados y por lo tanto, el estimador $\hat{\lambda}_1$ de λ no serían exactamente iguales.

Distribución de $\hat{\lambda}_1$

$S = \Sigma_{i=1}^{n} X_i \sim \mathcal{P}(n\lambda) \Rightarrow$

$$P\left(\hat{\lambda}_1 = \nu\right) = P\left(S = n\,\nu\right) = \frac{(n\lambda)^{n\nu} e^{-n\lambda}}{(n\nu)!} \quad \text{si} \quad n\nu \in N_0\ . \qquad (4)$$

$S \sim \mathcal{P}(n\lambda) \Rightarrow E(S) = n\lambda$, $\mathrm{Var}(S) = n\lambda$.

$$E\left(\hat{\lambda}_1\right) = \lambda \qquad \mathrm{Var}\left(\hat{\lambda}_1\right) = \frac{\lambda}{n}$$

$E\left(\hat{\lambda}_1\right) = \lambda$ dice que $\hat{\lambda}_1$ es un estimador insesgado de λ.

Por otra parte, $\text{Var}\left(\hat{\lambda}_1\right) = \frac{\lambda}{n}$ dice que la distribución es más concentrada a medida que n aumenta.

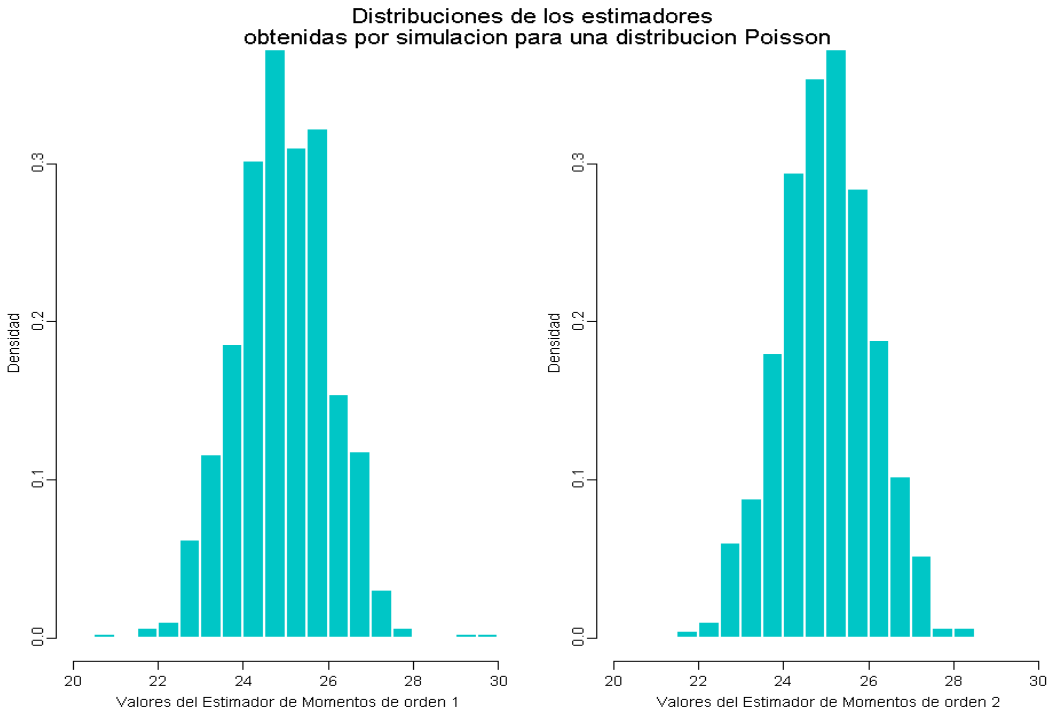
El error estándar de $\hat{\lambda}_1$ es

$$\sigma_{\hat{\lambda}_1} = \sqrt{\frac{\lambda}{n}} \longrightarrow s_{\hat{\lambda}_1} = \sqrt{\frac{\hat{\lambda}_1}{n}}$$

o sea, en este caso $s_{\hat{\lambda}_1} = \sqrt{\frac{24.913}{23}} = 1.04$.

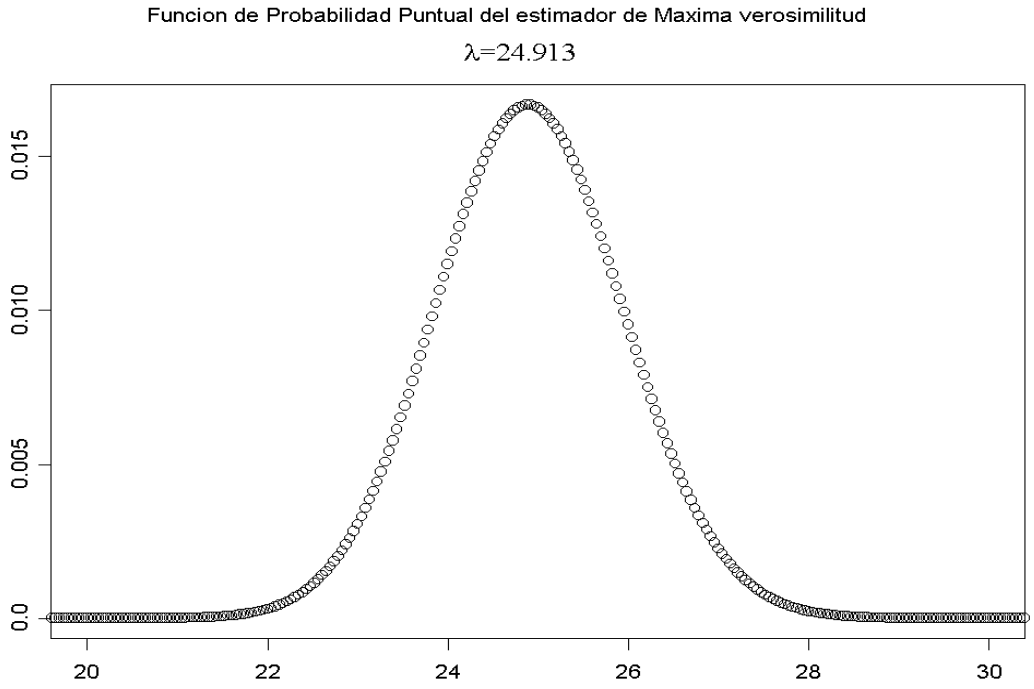
Si usamos el Teorema Central del Límite obtenemos que $\hat{\lambda}_1$ es aproximadamente normal con media λ y con un desvío estándar que puede estimarse por 1.04.

Distribución de $\hat{\lambda}_1$ y $\hat{\lambda}_2$



Como se observa ambos tienen un aspecto semejante.

Para su comparación se grafica a constinuación la función de probabilidad puntual de una variable con probabilidad puntual dada por (4) donde $\lambda = \hat{\lambda}_1$



Si aplicamos el Teorema Central del Límite a $\frac{1}{n}\Sigma_{i=1}^{n} X_i^2$, utilizando que si $X \sim \mathcal{P}(\lambda)$

$$E(X^4) = \lambda^4 + 6\lambda^3 + 7\lambda^2 + \lambda$$

podríamos obtener que $\hat{\lambda}_2$ es aproximadamente normal con media λ y con un desvío estándar

$$\frac{\sigma_2}{\sqrt{n}}$$

donde

$$\sigma_2^2 = \frac{\lambda + 6\lambda^2 + 4\lambda^3}{1 + 4\lambda + 4\lambda^2}$$

que puede estimarse por 1.05. Es decir, el estimador basado en el segundo momento es un poco más disperso asintóticamente que el basado en el primer momento. Veremos que este hecho se debe a que $\hat{\lambda}_1$ es el estimador de máxima verosimilitud.

4.2 Generalización cuando hay varios parámetros

$X_1, X_2, \ldots, X_n$ i.i.d. $X_i \sim F \in \mathcal{F} = \{F(x, \theta_1, \ldots, \theta_p)$ con $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \Theta \subset I\!R^p\}$.

Para estimar $\theta_1, \ldots, \theta_p$ por el método de los momentos se procede como sigue: Se consideran k funciones $g_1, \ldots, g_p$ de $I\!R$ en $I\!R$ y se resuelve el siguiente sistema

$$\frac{1}{n} \sum_{i=1}^{n} g_j(X_i) = E_{\widehat{\boldsymbol{\theta}}}(g_j(X_1)) \qquad j = 1, 2, \ldots, p \, .$$

Podemos tomar, por ejemplo, $g_j(x) = x^j$.

Ejemplo D. Sea $X_1, X_2, \ldots, X_n$ i.i.d. $X_i \sim N(\mu, \sigma^2)$. Consideremos $g_1(x) = x$ y $g_2(x) = x^2$. Luego,

$$\begin{aligned} \widehat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} X_i \\ \widehat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 \end{aligned}$$

Ejemplo E. Sea $X_1, X_2, \ldots, X_n$ i.i.d. $X_i \sim \Gamma(\alpha, \lambda)$. Consideremos $g_1(x) = x$ y $g_2(x) = x^2$. Luego, si $\overline{X} = \frac{1}{n} \Sigma_{i=1}^{n} X_i$ y por $\widehat{\sigma}^2 = \frac{1}{n} \Sigma_{i=1}^{n} (X_i - \overline{X})^2$, los estimadores de los momentos para λ y α resultan ser

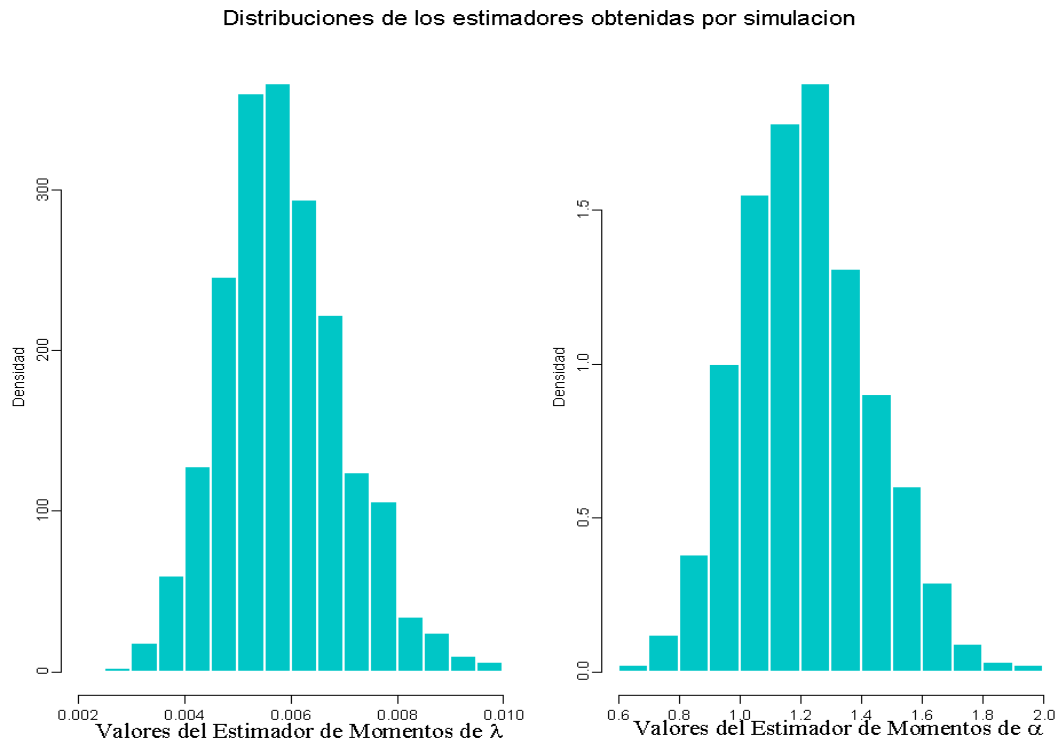
$$\hat{\lambda} = \frac{\overline{X}}{\widehat{\sigma}^2}$$

$$\widehat{\alpha} \;=\; \frac{\overline{X}^2}{\widehat{\sigma}^2}\;.$$

Consideremos los datos del Ejemplo 2 sobre el Tiempo de Falla de cables Kevlar 49/epoxy sometidos a presión sostenida del 80%. Para esos 100 datos obtenemos $\overline{X} = 209.1828$ y $\widehat{\sigma}^2 = 37640$, con lo cual $\widehat{\alpha} = 1.1625$ y $\hat{\lambda} = 0.00556$.

Distribución de $\widehat{\alpha}$ y $\hat{\lambda}$

La podemos obtener por simulación generando muchas muestras de tamaño 100 de una distribución $\Gamma(1.1625, 0.00556)$.



Estos gráficos muestran que los estimadores parecen tener una distribución aproximadamente normal. La variabilidad mostrada en

los histogramas puede resumirse calculando los desvíos de las 1000 estimaciones, lo que provee el error estándar estimado de $\widehat{\alpha}$ y de $\hat{\lambda}$ y que en este caso son 0.2093 y 0.00114 respectivamente.

Este uso de la simulación es un ejemplo de lo que en estadística se denomina *bootstrap.*

En conclusión, de los ejemplos anteriores vemos que los pasos básicos para estimar $\boldsymbol{\theta} \in IR^p$ son:

a) Calcular p momentos de la distribución de las observaciones

b) Expresar $\boldsymbol{\theta}$ en función de ellos

c) Reemplazar los momentos por los momentos muestrales en la expresión obtenida en b) para obtener $\widehat{\boldsymbol{\theta}}$.

Bajo condiciones de regularidad los estimadores de los momentos resultan consistentes.

La consistencia de los estimadores de los momentos sirve de justificación para el procedimiento dado en el Ejemplo C para estimar el error estándar.

5 Método de máxima verosimilitud

$\mathbf{X} = (X_1, X_2, \ldots, X_n)$ discreto o continuo cuya función de densidad discreta o continua pertenece a una familia $f(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$ y que se quiera estimar $\boldsymbol{\theta}$.

Definición 1. Diremos $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{X})$ es un estimador de máxima verosimilitud (E.M.V.) de $\boldsymbol{\theta}$, si se cumple

$$f(\mathbf{X}, \widehat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta} \in \Theta} f(\mathbf{X}, \boldsymbol{\theta})$$

5.1 Cómputo del E.M.V..

Supongamos

- $\Theta \subset I\!R^p$ abierto,
- el soporte de $f(\mathbf{x}, \boldsymbol{\theta})$ no depende de $\boldsymbol{\theta}$
- $f(\mathbf{x}, \boldsymbol{\theta})$ tiene derivadas parciales respecto a θ_i.

Luego, el E.M.V. $\widehat{\boldsymbol{\theta}}$ debe verificar

$$\frac{\partial \ln f(\mathbf{X}, \boldsymbol{\theta})}{\partial \theta_i} = 0 \qquad\qquad i = 1, 2, \ldots, p\ . \tag{5}$$

En particular, si $X_1, \ldots, X_n$ son i.i.d. con función de probabilidad puntual o densidad $f(x, \boldsymbol{\theta})$, se verifica

$$f(\mathbf{x}, \boldsymbol{\theta}) = f(x_1, \ldots, x_n, \boldsymbol{\theta}) = \prod_{j=1}^{n} f(x_j, \boldsymbol{\theta})$$

y bajo las condiciones dadas anteriormente, el sistema de ecuaciones (5) se transforma en

$$\sum_{i=1}^{n} \frac{\partial \ln f(X_i, \widehat{\boldsymbol{\theta}})}{\partial \theta_j} = 0 \qquad\qquad j = 1, 2, \ldots, p \; . \tag{6}$$

Sea

$$\psi_j(x, \boldsymbol{\theta}) = \; - \frac{\partial \ln f(x, \widehat{\boldsymbol{\theta}})}{\partial \theta_j}$$

entonces (6) puede escribirse como

$$\sum_{i=1}^{n} \psi_j(X_i, \boldsymbol{\theta}) = 0 \qquad\qquad j = 1, 2, \ldots, p \; .$$

Esta ecuación corresponde a la forma general de los denominados $M-$estimadores que han visto para el caso de posición y escala.

Ejemplo D. Continuación. Sea $X_1, \ldots, X_n$, $X_i \sim N(\mu, \sigma^2)$. Los E.M.V. de μ y σ^2 son

$$\begin{aligned} \widehat{\mu} &= \sum_{i=1}^{n} X_i/n = \overline{X} \\ \widehat{\sigma}^2 &= \sum_{i=1}^{n} (X_i - \overline{X})^2/n \end{aligned}$$

que son los mismos estimadores que encontramos por el método de los momentos.

Ejemplo E. Continuación. Sean $X_i \sim \Gamma(\alpha, \lambda)$, $1 \leq i \leq n$. Los E.M.V. de α y λ verifican

$$\begin{aligned} n \, \ln \hat{\lambda} + \sum_{i=1}^{n} \ln(X_i) - n \, \frac{\Gamma'(\widehat{\alpha})}{\Gamma(\widehat{\alpha})} &= 0 \\ \frac{n \, \widehat{\alpha}}{\hat{\lambda}} - n \, \overline{X} &= 0 \, , \end{aligned}$$

donde $\Gamma'(\alpha)$ indica la derivada de la función $\Gamma(\alpha)$.

Este sistema no tiene una solución explícita aunque

$$\hat{\lambda} = \frac{\widehat{\alpha}}{\overline{X}} \, .$$

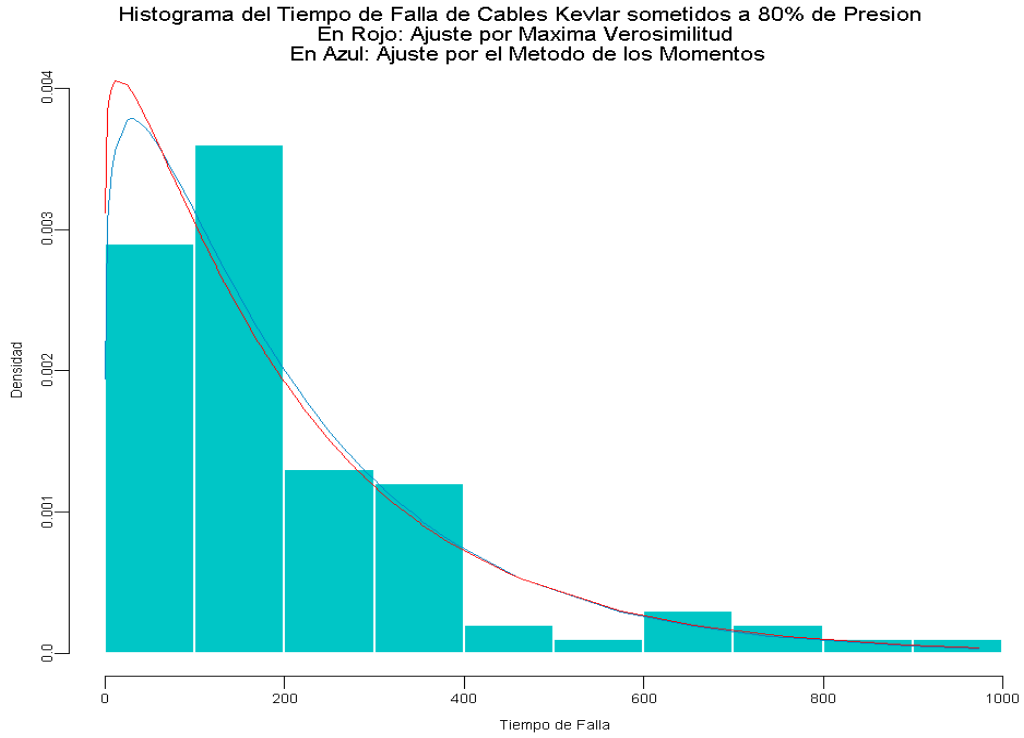
Al reemplazar el valor de $\hat{\lambda}$ obtenemos la ecuación no lineal

$$n \, (\ln \widehat{\alpha} - \ln(\overline{X})) + \sum_{i=1}^{n} \ln(X_i) - n \, \frac{\Gamma'(\widehat{\alpha})}{\Gamma(\widehat{\alpha})} = 0 \, ,$$

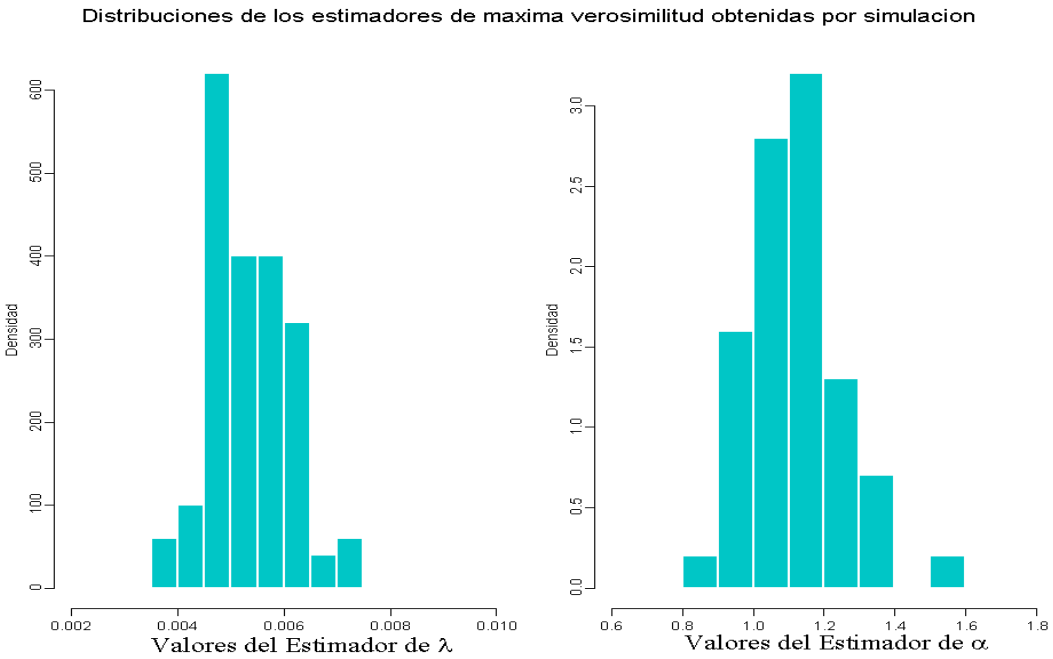
que puede resolverse, por ejemplo mediante, el algoritmo de Newton-Raphson. Para iniciar el proceso, se puede tomar como estimador inicial el estimador de los momentos, por ejemplo.

En este caso, el estimador de máxima verosimilitud no coincide con el estimador de los momentos.

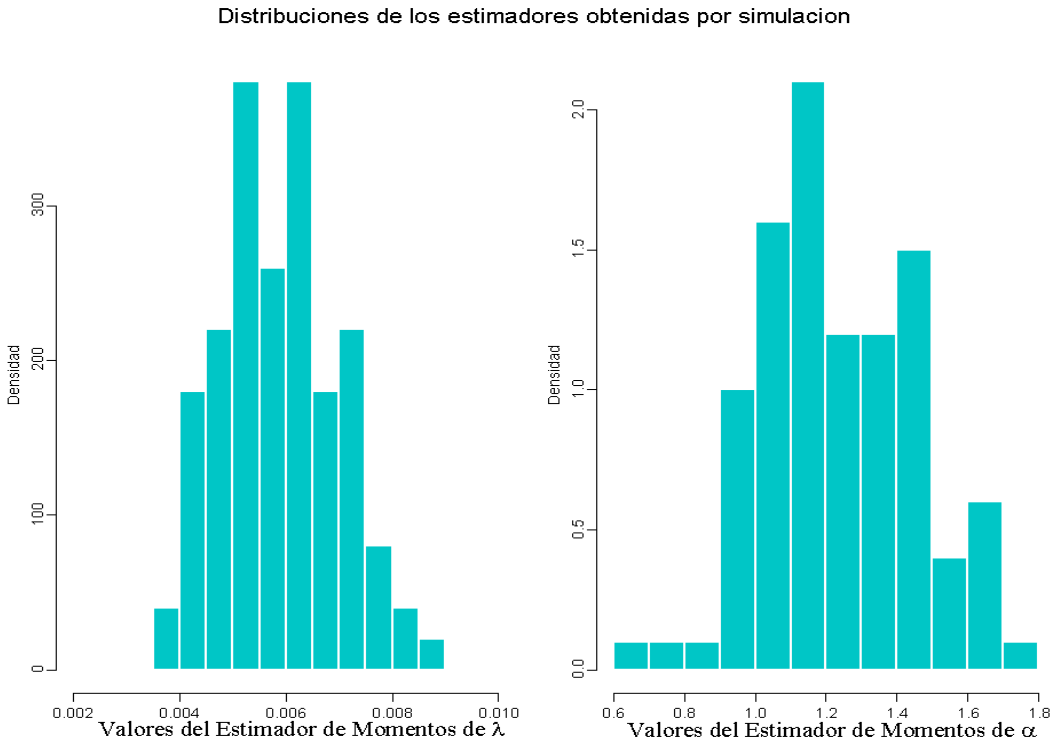
Datos sobre el Tiempo de Falla de cables Kevlar 49/epoxy sometidos a presión sostenida del 80%. Para esos datos los E.M.V. resultan ser $\widehat{\alpha} = 1.0774$ y $\hat{\lambda} = 0.00515$.



Los E.M.V. no tienen una forma explícita en el caso de la distribución $\Gamma(\alpha, \lambda)$.**Bootstrap:** generamos $nr = 100$ muestras cada una de tamaño $n = 100$ de una variable $\Gamma(1.0774, 0.00515)$.



Comparamos los histogramas obtenidos con los estimadores de los momentos utilizando $nr = 100$



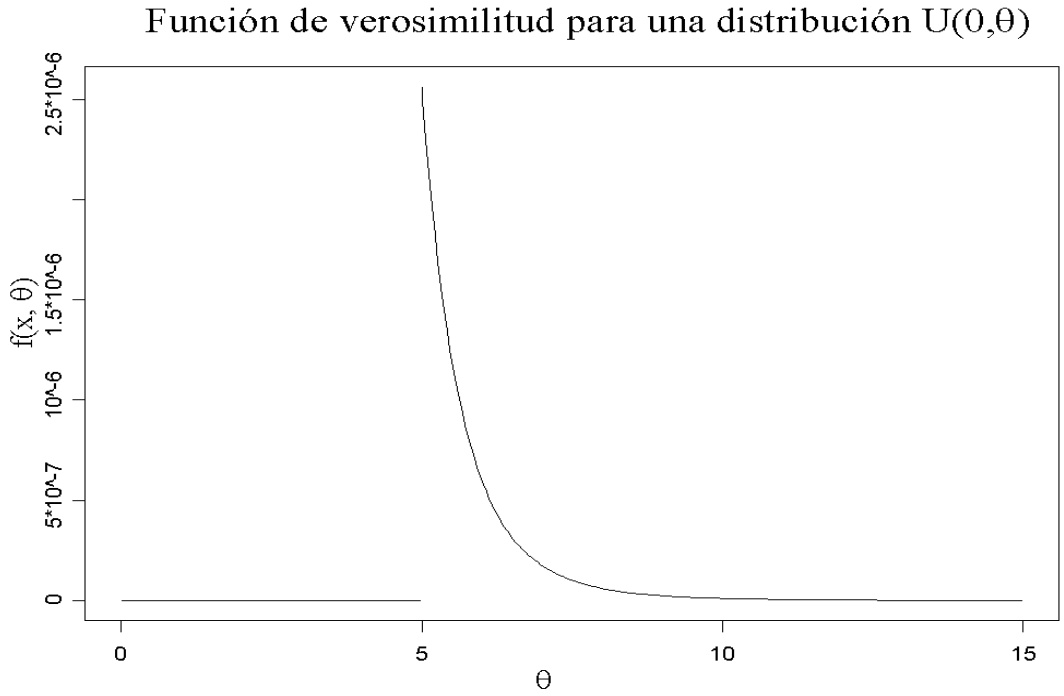
$$s_{\widehat{\alpha}_{\text{Momentos}}} = 0.2172 \qquad s_{\widehat{\lambda}_{\text{Momentos}}} = 0.00109$$

$$s_{\widehat{\alpha}_{\text{MV}}} = 0.1344 \qquad s_{\widehat{\lambda}_{\text{MV}}} = 0.00072$$

Ejemplo F. $X_i \sim \mathcal{U}(0, \theta)$.

$$f(\mathbf{x}, \theta) = \begin{cases} \dfrac{1}{\theta^n} & \text{si} \quad 0 < \text{mín}(x_i) \leq \text{máx}(x_i) < \theta \\ 0 & \text{otro caso} \end{cases}$$

El E.M.V. de θ es $\hat{\theta}_n = \max_{1 \leq i \leq n}(X_i)$. El siguiente es el gráfico de la función de verosimilitud, $f(\mathbf{x}, \theta)$ correspondiente a $n = 8$ observaciones tales que $\max_{1 \leq i \leq n}(X_i) = 5$.



La distribución de $\hat{\theta}_n = \max_{1 \leq i \leq n}(X_i)$ está dada por

$$F_{\hat{\theta}_n}(t, \theta) = \begin{cases} \dfrac{t^n}{\theta^n} & \text{si} \quad 0 < t < \theta \\ 0 & \text{otro caso} \end{cases}$$

con lo cual

$$E\left(\max_{1\leq i\leq n}(X_i)\right) = \frac{n}{n+1}\theta \qquad \text{Var}\left(\max_{1\leq i\leq n}(X_i)\right) = \frac{n}{(n+2)(n+1)^2}\theta^2 \ .$$

Por lo tanto, por el Teorema 1, $\hat{\theta}_n$ es débilmente consistente para θ.

5.2 Consistencia de los estimadores de máxima verosimilitud

Sea $\hat{\theta}_n$ el E.M.V de θ, o sea,

$$\max_{\theta\in\Theta}\prod_{i=1}^{n} f(x_i,\theta) = \prod_{i=1}^{n} f(x_i,\hat{\theta}_n) \tag{7}$$

El siguiente Teorema da condiciones muy generales bajo las cuales $\hat{\theta}_n$ definido por (7) es fuertemente consistente.

Teorema 3. *Sea $X_1,\ldots,X_n$ una muestra aleatoria de una distribución discreta o continua con densidad en la familia $f(x,\theta)$ con $\theta\in\Theta$, donde Θ es un intervalo abierto de IR. Supongamos que $f(x,\theta)$ es derivable respecto de θ y que el conjunto $S=\{x : f(x,\theta)\neq 0\}$ es independiente de θ para todo $\theta\in\Theta$. Sea $\hat{\theta}_n$ el estimador de máxima verosimilitud de θ, que satisface*

$$\sum_{i=1}^{n}\frac{\partial \ln f(X_i,\hat{\theta}_n)}{\partial\theta} = 0 \tag{8}$$

Supongamos finalmente que la ecuación (8) tiene a lo sumo una solución y que $\theta\neq\theta'$ implica que $f(x,\theta)\neq f(x,\theta')$. Entonces $\hat{\theta}_n \xrightarrow{c.t.p.} \theta$, es decir, $\hat{\theta}_n$ es una sucesión de estimadores fuertemente consistente.

Un resultado análogo de consistencia se puede obtener bajo condiciones de regularidad para los M–estimadores.

Sea $\Psi(x, \theta)$ una función de escores, continua en θ. Sea $T(F)$ la solución de la ecuación

$$E_F \Psi(X, \theta) = 0 \tag{9}$$

donde E_F indica la esperanza cuando $X \sim F$.

Supongamos que tenemos observaciones de una distribución $F(x, \theta)$, con $\theta \in \Theta$. Se dice que T es *Fisher–consistente* si $T(F(\cdot, \theta)) = \theta$.

Teorema 4. *Sean $X_1, \ldots, X_n$ i.i.d. con distribución en la familia $F(x, \theta)$ con $\theta \in \Theta$, donde Θ es un intervalo abierto de IR. Supongamos además que el soporte de $F(x, \theta)$ no depende de θ. Sea $\hat{\theta}_n$ la solución de la ecuación*

$$\sum_{i=1}^{n} \Psi(X_i, \hat{\theta}_n) = 0 \tag{10}$$

Supongamos finalmente que la ecuación (9) tiene única solución. Entonces, $\hat{\theta}_n \xrightarrow{c.t.p.} T(F(\cdot, \theta))$.

En particular, si T es Fisher–consistente para θ se obtiene que

$$\hat{\theta}_n \xrightarrow{c.t.p.} \theta$$

6 Método de cuadrados mínimos

Supongamos que $Y_1, \ldots, Y_n$ son variables aleatorias de la forma

$$Y_i = g_i(\theta_1, \ldots, \theta_p) + \varepsilon_i \qquad 1 \leq i \leq n \tag{11}$$

donde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ es un vector de parámetros desconocido, del cual lo único que se conoce es que está en un conjunto $\Theta \subset I\!R^p$ y ε_i son variables aleatorias tales que

(i) $E(\varepsilon_i) = 0$

(ii) $\mathrm{Var}(\varepsilon_i) = \sigma^2$

(iii) $\varepsilon_1, \ldots, \varepsilon_n$ son variables aleatorias independientes.

Ejemplo G. Supongamos tener el modelo de regresión simple, o sea,

$$Y_i = \theta_1 X_i + \theta_2 + \varepsilon_i \qquad 1 \leq i \leq n$$

donde las variables ε_i satisfacen (i), (ii) y (iii). Luego, si llamamos

$$g_i(\theta_1, \theta_2) = \theta_1 X_i + \theta_2 \qquad 1 \leq i \leq n$$

estamos en la situación descripta por la ecuación (11).

Definición 1. Llamaremos *estimador de mínimos cuadrados* (E.M.C.) al valor $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(Y_1, \ldots, Y_n)$ que hace mínima la expresión $\Sigma_{i=1}^n (Y_i - g_i(\theta_1, \ldots, \theta_p))^2$, es decir

$$\sum_{i=1}^n (Y_i - g_i(\widehat{\boldsymbol{\theta}}))^2 = \min_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^n (Y_i - g_i(\boldsymbol{\theta}))^2 \, . \tag{12}$$

Si además de satisfacer (i), (ii) y (iii), los ε_i tienen distribución normal, entonces el E.M.C. coincide con el E.M.V.

6.1 Cálculo de los E.M.C.

Si Θ es abierto y si las funciones $g_i(\theta_1, \ldots, \theta_p)$ son derivables respecto a cada θ_i, $\widehat{\boldsymbol{\theta}}$ deberá satisfacer el sistema de ecuaciones siguiente

$$\frac{\partial \, \Sigma_{i=1}^{n} (Y_i - g_i(\widehat{\boldsymbol{\theta}}))^2}{\partial \theta_j} = 0 \qquad\qquad j = 1, 2, \ldots, p \; ,$$

o sea,

$$\sum_{i=1}^{n} (Y_i - g_i(\widehat{\boldsymbol{\theta}})) \frac{\partial g_i(\widehat{\boldsymbol{\theta}})}{\partial \theta_j} = 0 \qquad\qquad j = 1, 2, \ldots, p \; .$$

Ejemplo G. Continuación. Como $g_i(\theta_1, \theta_2) = \theta_1 X_i + \theta_2$, se tiene

$$\frac{\partial g_i(\boldsymbol{\theta})}{\partial \theta_1} = X_i \qquad \text{y} \qquad \frac{\partial g_i(\boldsymbol{\theta})}{\partial \theta_2} = 1 \; .$$

Luego, (12) se transforma en el sistema

$$\begin{aligned} \sum_{i=1}^{n} (Y_i - \hat{\theta}_1 X_i - \hat{\theta}_2) X_i &= 0 \\ \sum_{i=1}^{n} (Y_i - \hat{\theta}_1 X_i - \hat{\theta}_2) &= 0 \; . \end{aligned}$$

con solución

$$\begin{aligned} \hat{\theta}_1 &= \frac{\Sigma_{i=1}^{n} (Y_i - \overline{Y})(X_i - \overline{X})}{\Sigma_{i=1}^{n} (X_i - \overline{X})^2} \; , \\ \hat{\theta}_2 &= \overline{Y} - \hat{\theta}_1 \overline{X} \; , \end{aligned}$$

Geométricamente la recta $y = \hat{\theta}_1 x + \hat{\theta}_2$ tiene la propiedad siguiente: Minimiza la suma de los cuadrados de las distancias de los puntos (X_i, Y_i) a la recta, si esta distancia se la mide paralelamente al eje de las y. Por otra parte, pasa por el punto $(\overline{X}, \overline{Y})$

Otra posible técnica es la que minimiza la distancia ortogonal de los puntos a la recta, es decir, si $\mathbf{Z}_i = (X_i, Y_i)$, el método de componentes principales busca la recta $\mathcal{L}_o$: $y = \theta_{1o}x + \theta_{2o}$ tal que $\Sigma_{i=1}^n \|\mathbf{Z}_i - P(\mathbf{Z}_i, \mathcal{L})\|^2$ sea mínima, donde $P(\mathbf{Z}, \mathcal{L})$ indica la proyección ortogonal del punto $\mathbf{Z}$ a la recta $\mathcal{L}$, es decir,

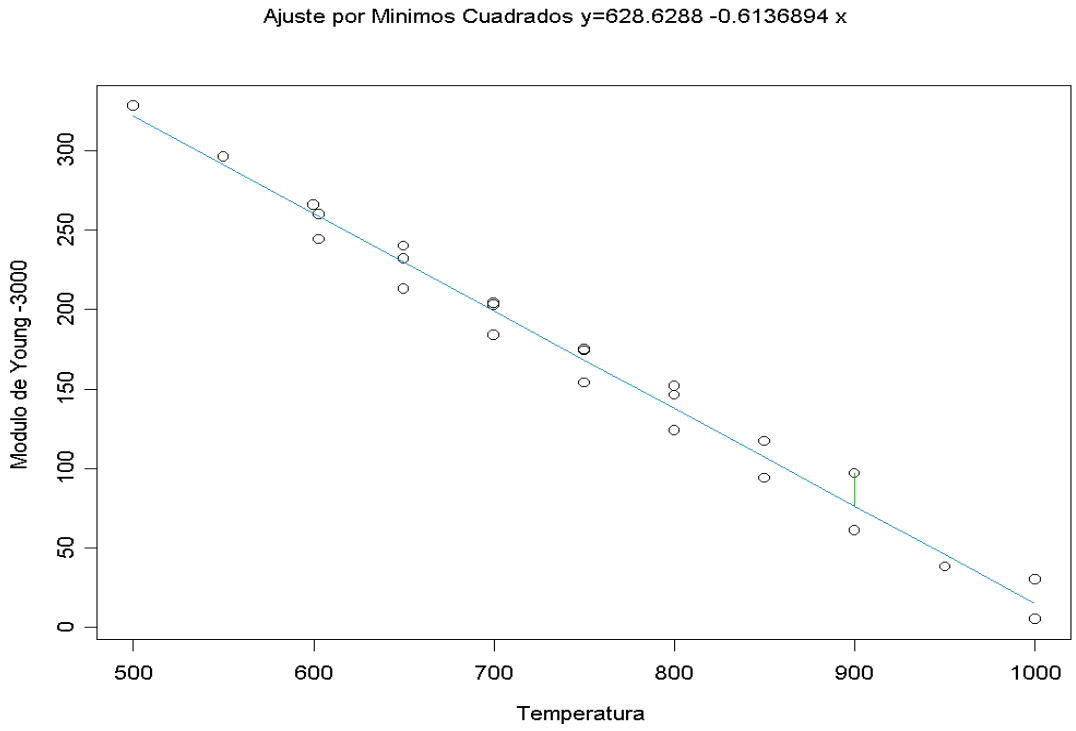
$$\sum_{i=1}^{n} \|\mathbf{Z}_i - P(\mathbf{Z}_i, \mathcal{L}_o)\|^2 = \min_{\mathcal{L}} \sum_{i=1}^{n} \|\mathbf{Z}_i - P(\mathbf{Z}_i, \mathcal{L})\|^2$$

Esta recta tiene la propiedad de que su dirección $\boldsymbol{\beta}_o$ es la dirección que maximiza la variabilidad de los puntos proyectados $\boldsymbol{\beta}_o'\mathbf{Z}$. Ambos procedimientos no dan los mismos resultados.

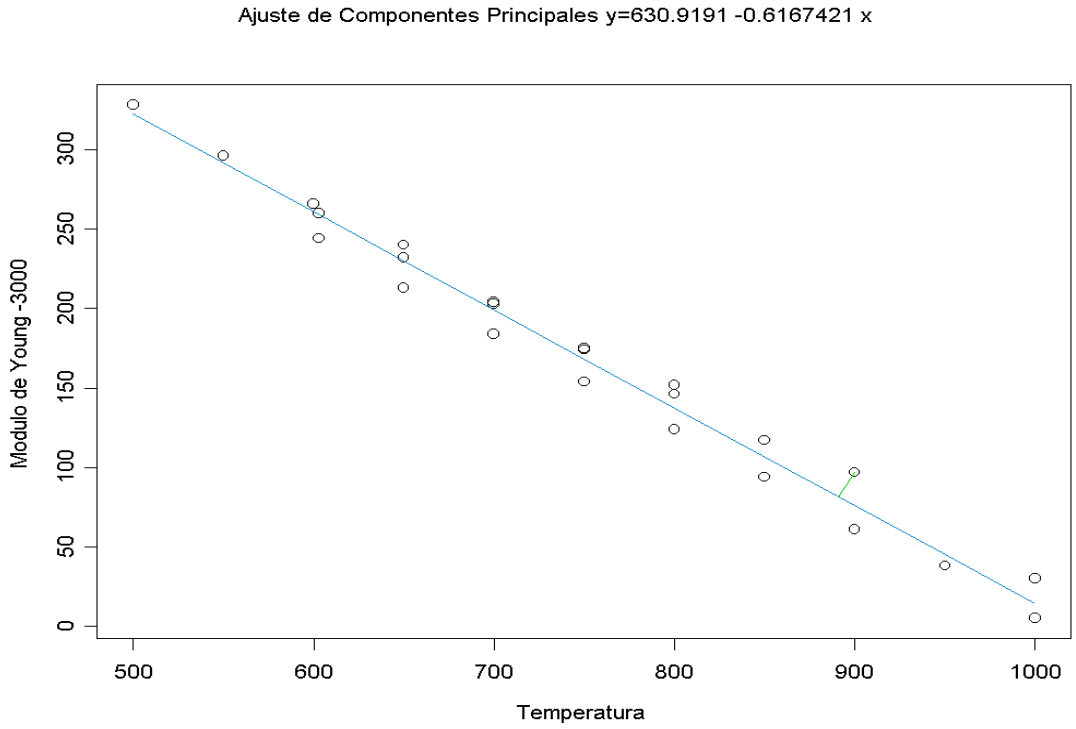
Ejemplo 3. La Tabla siguiente muestra los valores observados del Módulo de Young (G) medido a varias Temperaturas (T) para varillas de zafiro.

T	G	T	G	T	G
500	328	550	296	600	266
603	260	603	244	650	240
650	232	650	213	700	204
700	203	700	184	750	174
750	175	750	154	800	152
800	146	800	124	850	117
850	94	900	97	900	61
950	38	1000	30	1000	5

El gráfico siguiente muestra el ajuste obtenido mediante mímimos cuadrados de la recta $G = \theta_1 T + \theta_2$. Se indica además en verde la distancia vertical que se ha minimizado.


Ajuste por Minimos Cuadrados y=628.6288 -0.6136894 x
Modulo de Young -3000
0
50
100
150
200
250
300
500
600
700
800
900
1000
Temperatura

En el gráfico siguiente se muestra la recta obtenida mediante el procedimiento de componentes principales. Se indica además la distancia que se ha minimizado



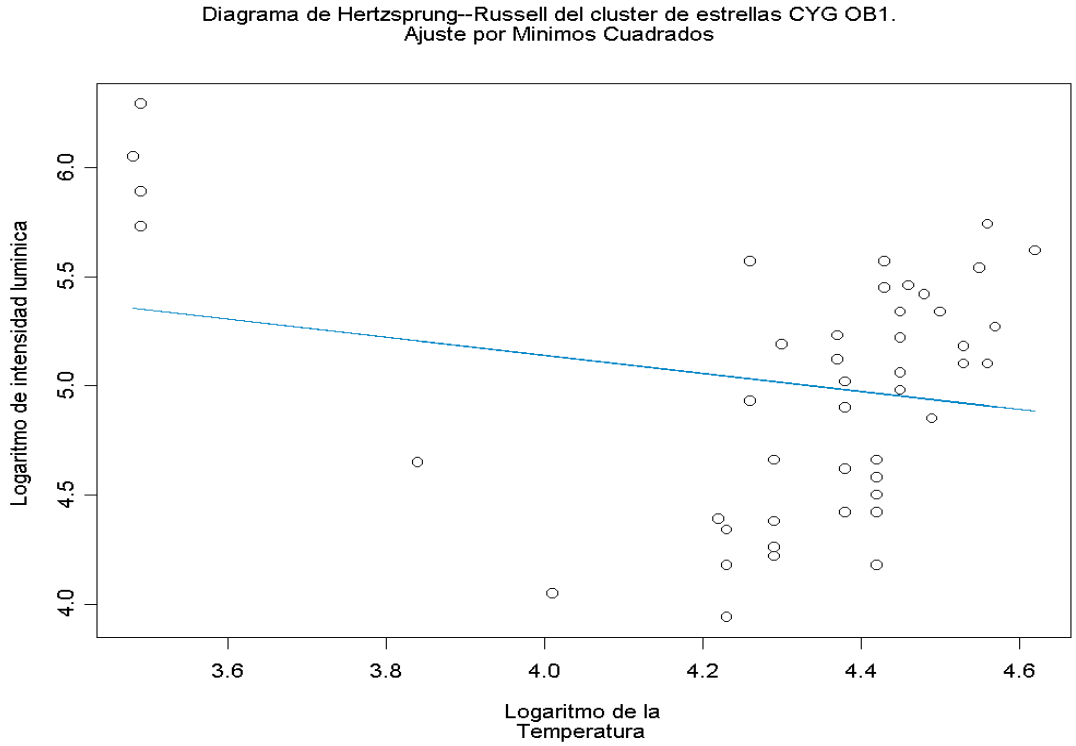
6.2 Sensibilidad a observaciones atípicas.

Al igual que el método de máxima verosimilitud y que el de los momentos, el método de mínimos cuadrados es altamente sensible a observaciones atípicas. ya han visto la sensibilidad de la media y mediana a observaciones atípicas y cómo la mediana y los M–estimadores proveían alternativas más robustas.

Ejemplo 4. Los datos de la Tabla siguiente forman el diagrama de Hertzsprung–Russell del cluster de estrellas CYG OB1, que contiene 47 estrellas en la dirección de Cygnus. X es el logaritmo de la temperatura en la superficie de la estrella e Y es el logaritmo de intensidad lumínica.

obs	X	Y	obs	X	Y
1	4.37	5.23	25	4.38	5.02
2	4.56	5.74	26	4.42	4.66
3	4.26	4.93	27	4.29	4.66
4	4.56	5.74	28	4.38	4.90
5	4.30	5.19	29	4.22	4.39
6	4.46	5.46	30	3.48	6.05
7	3.84	4.65	31	4.38	4.42
8	4.57	5.27	32	4.56	5.10
9	4.26	5.57	33	4.45	5.22
10	4.37	5.12	34	3.49	6.29
11	3.49	5.73	35	4.23	4.34
12	4.43	5.45	36	4.62	5.62
13	4.48	5.42	37	4.53	5.10
14	4.01	4.05	38	4.45	5.22
15	4.29	4.26	39	4.53	5.18
16	4.42	4.58	40	4.43	5.57
17	4.23	3.94	41	4.38	4.62
18	4.42	4.18	42	4.45	5.06
19	4.23	4.18	43	4.50	5.34
20	3.49	5.89	44	4.45	5.34
21	4.29	4.38	45	4.55	5.54
22	4.29	4.22	46	4.45	4.98
23	4.42	4.42	47	4.42	4.50
24	4.49	4.85			

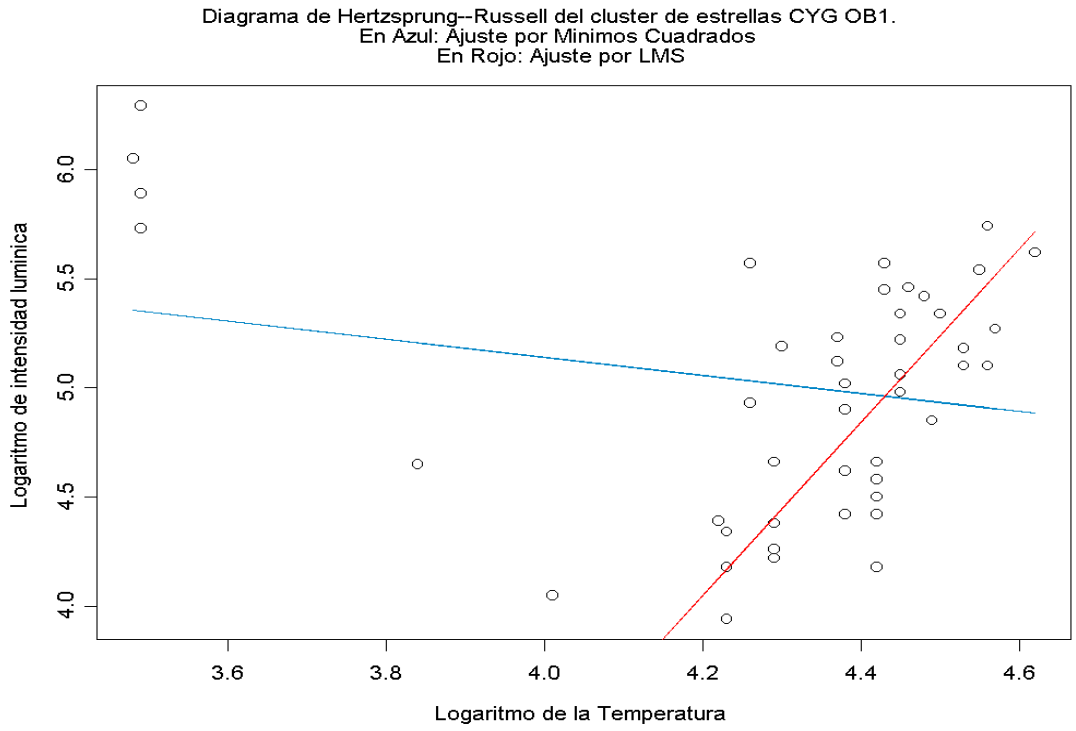
El siguiente es el gráfico del ajuste por mínimos cuadrados de la recta $g(x) = \theta_1\, x + \theta_2$



Los estimadores de mínimos cuadrados $\hat{\theta}_1 = -0.4133039$ y $\hat{\theta}_2 = 6.793467$ no parecen dar un buen ajuste del grueso de los datos. Para resolver esta situación una manera simple consiste en reemplazar la suma de cuadrados por la mediana de los cuadrados, es decir el estimador propuesto por Rousseeuw (1984) es la solución de

$$med\left(Y_i - \left[\tilde{\theta}_2 + \tilde{\theta}_1\, X_i\right]\right)^2 = \min_{\boldsymbol{\theta}\in\Theta} med\left(Y_i - \left[\theta_2 + \theta_1\, X_i\right]\right)^2$$

y se denomina *least median of squares.* El ajuste para este conjunto de datos resulta ser $\tilde{\theta}_1 = 3.970588$ $\tilde{\theta}_2 = -12.62794$ y a continuación se da el gráfico con el ajuste obtenido.



La desventaja de este estimador es su velocidad de convergencia ya que converge a velocidad $n^{\frac{1}{3}}$ en lugar de $n^{\frac{1}{2}}$.

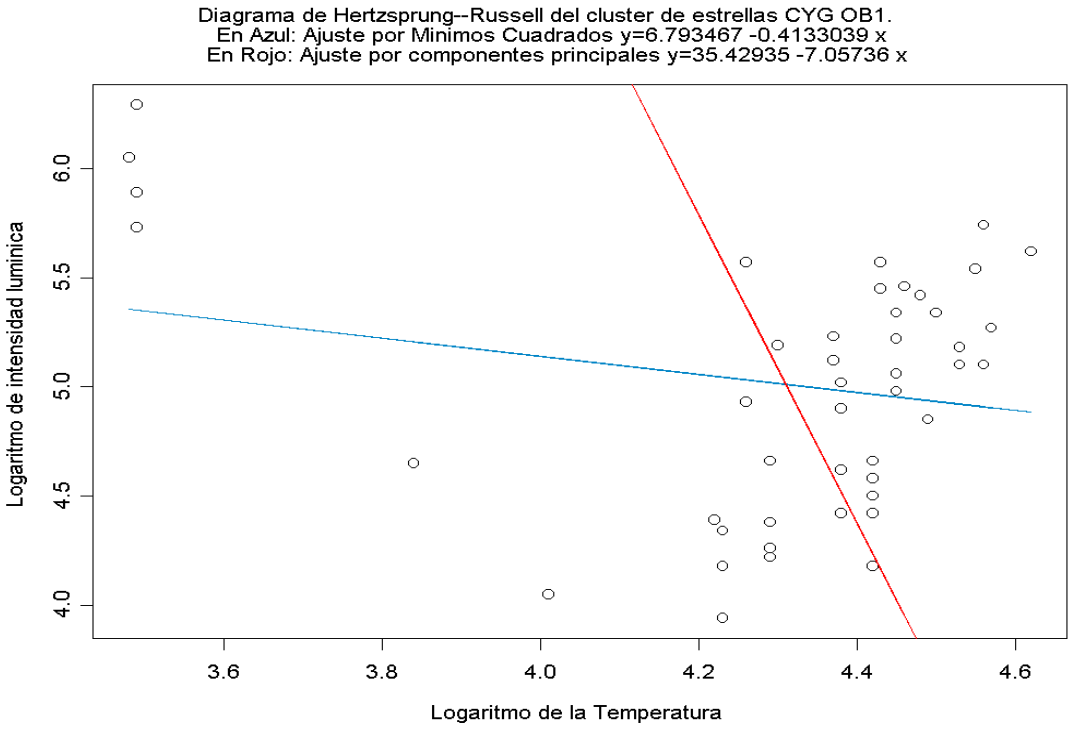
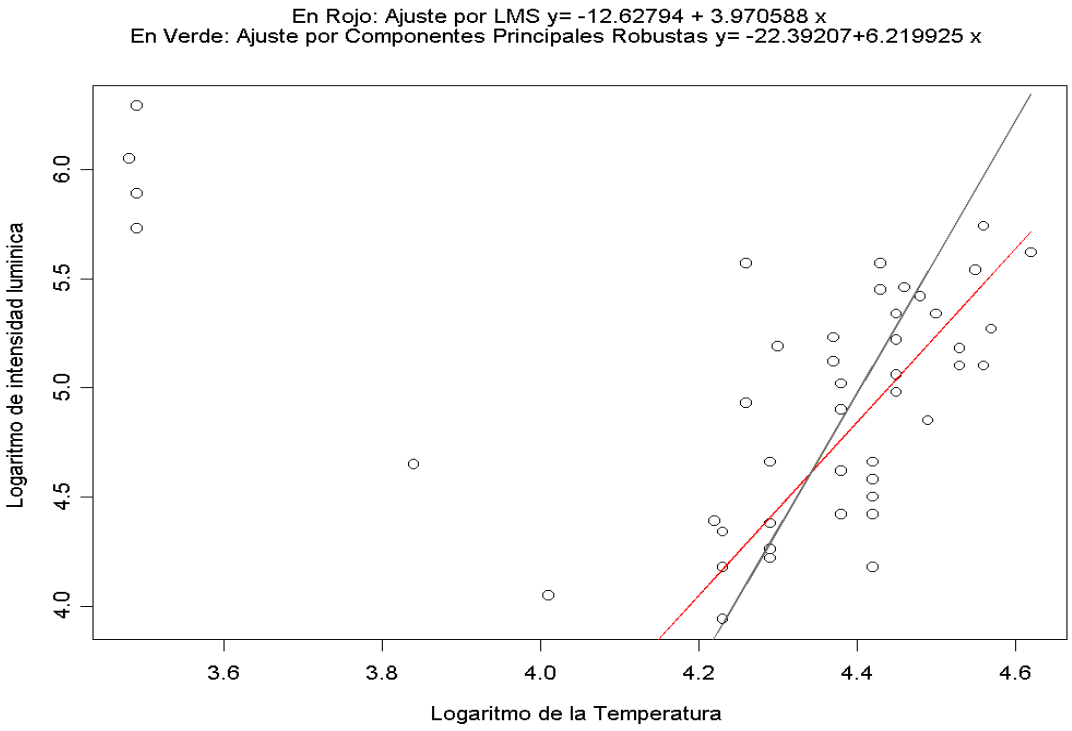
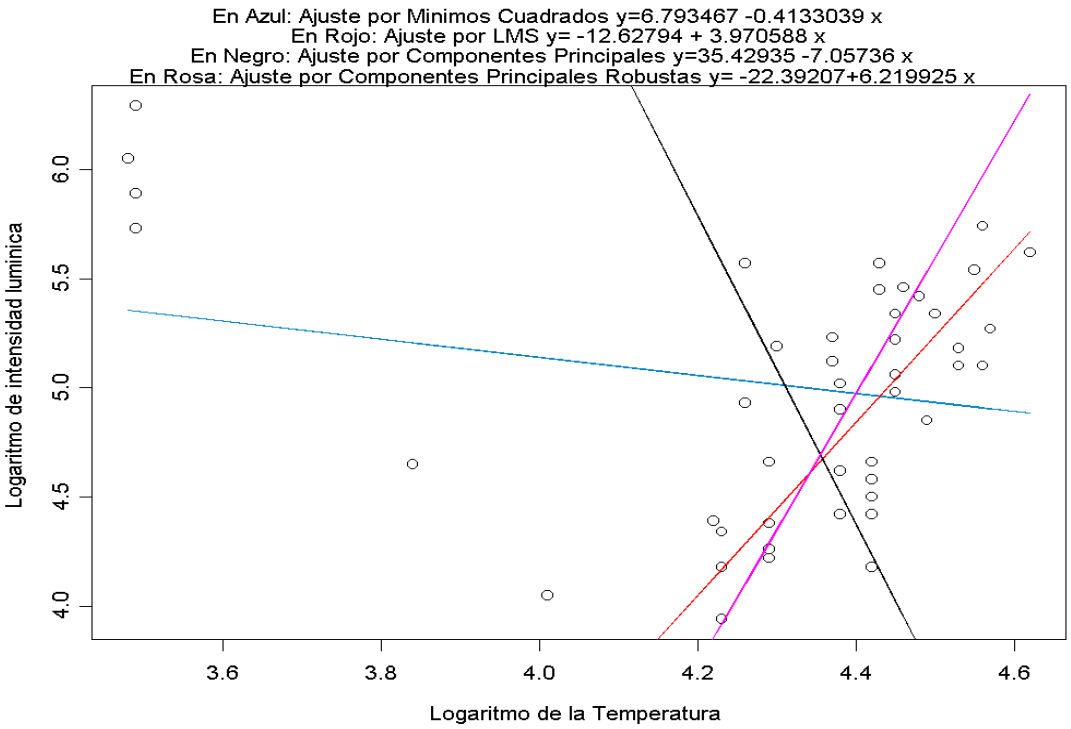


Diagrama de Hertzsprung--Russell del cluster de estrellas CYG OB1.
En Azul: Ajuste por Minimos Cuadrados y=6.793467 -0.4133039 x
En Rojo: Ajuste por componentes principales y=35.42935 -7.05736 x
Logaritmo de intensidad luminica
4.0
4.5
5.0
5.5
6.0
3.6
3.8
4.0
4.2
4.4
4.6
Logaritmo de la Temperatura


En Rojo: Ajuste por LMS y= -12.62794 + 3.970588 x
En Verde: Ajuste por Componentes Principales Robustas y= -22.39207+6.219925 x
Logaritmo de intensidad luminica
4.0
4.5
5.0
5.5
6.0
3.6
3.8
4.0
4.2
4.4
4.6
Logaritmo de la Temperatura


En Azul: Ajuste por Minimos Cuadrados y=6.793467 -0.4133039 x
En Rojo: Ajuste por LMS y= -12.62794 + 3.970588 x
En Negro: Ajuste por Componentes Principales y=35.42935 -7.05736 x
En Rosa: Ajuste por Componentes Principales Robustas y= -22.39207+6.219925 x
Logaritmo de intensidad luminica
4.0
4.5
5.0
5.5
6.0
3.6
3.8
4.0
4.2
4.4
4.6
Logaritmo de la Temperatura