Estimación robusta: una introducción

Ricardo A. Maronna

Universidad de La Plata y Universidad de Buenos Aires

1 ¿Para qué necesitamos mètodos robustos?

Un ejemplo: El siguiente dataset contiene 24 mediciones (ordenadas) del contenido de cobre en muestras de harina (en partes por millón).

2.20 2.20 2.40 2.40 2.50 2.70 2.80 2.90 3.03 3.03 3.10 3.37

3.40 3.40 3.40 3.50 3.60 3.70 3.70 3.70 3.70 3.77 5.28 28.95

El último valor 28.95 es llamativo. Podría deberse a un error de transcripción. Para ver su efecto, comparamos los valores de la media, el desvío standard y la mediana muestrales, con y sin él.

	con	sin
media	4.28	3.21
desvío	5.30	0.69
mediana	3.40	3.37

Vemos que el efecto en la media es notorio, en el desvío es espectacular, y en la mediana imperceptible.

1.1 Pensándolo desde los datos

Queremos ajustar un modelo a un conjunto de datos ("dataset") y hacer inferencia usando métodos "clàsicos" (medias y desvíos muestrales, regresión por mínimos cuadrados, etc.).

Puede haber algunas observaciones atípicas ("outliers") que no parecen corresponder al conjunto y que pueden tener una desmedida influencia en los resultados.

El enfoque más antiguo ("diagnósticos"): detectar outliers, eliminarlos, y volver a estimar.

Usar un buen diagnóstico para outliers es claramente mucho mejor que no hacer nada, pero tiene sus inconvenientes:

- ¿Cuándo se justifica la eliminación?. La eliminación requiere una decisión subjetiva. ¿Cuando es una observación "lo bastante atípica" como para ser eliminada?

- El usuario o el autor de los datos puede pensar que "una observación es una observación" (o sea, las observaciones deberían hablar por símismas) y por lo tanto no estar de acuerdo con su eliminación.

- Dado que generalmente existe cierta incertidumbre sobre si una observación es realmente atípica, existe el riesgo de eliminar observaciones "buenas", con lo que se subestima la variabilidad de los datos.

- Como los resultados dependen de las decisiones subjetivas del usuario, es difícil determinar el comportamiento estadistico del procedimiento completo.

- Last but not least: los valores atípicos no son necesriamente "malos". ¡Pueden contener información valiosa!

Métodos robustos: automáticamente ajustan los datos "típicos"

Pueden ser descriptos generalmente como métodos clásicos pesados, donde los pesos dependen del conjunto de datos, y dan menor peso a los outliers.

1.2 Pensándolo desde el modelo

Enfoque clásico: creemos que el modelo vale exactamente.

Métodos "óptimos": generalmente Estimador de Máxima Verosimilitud (EMV) y Likelihood Ratio Tests (LRT)).

Pero enla vida real los modelos valen –en el mejor de los casos– sólo aproximadamente.

Si el modelo vale sólo aptorimadamente, los métodos clásicos *no* son aproximadamente óptimos.

Métodos robustos: en vez de ser óptimos para un modelo, son aproximadamente óptimos en un èntorno" del modelo.

1.3 Un poco de historia

La media muestral, el desvío standard muestral, el estimador de Mínimos Cuadrados (EMC), los tests T, están justificados por ser òptimos para datos normales.

¿Por qué suponer normalidad?

Muchos datasets tienen distribución con forma de campana en el centro, pero con colas más peadas que la normal.

En realidad Gauss (1820) justificó la suposición de normalidad porque ésta justificaba el uso del EMC, que era el único estimador de regresión calculable en esa época (y todavía mucho más tarde).

Puede parecer natural proceder como sigue:

- testear la hipótesis de que los datos son normales
- si no es rechazada, usar la media
- si no, usar la mediana

O, mejor aún, ajustar una distribución a los datos y usar el EMV correspondiente.

Pero esto tienen el incoveniente de que hacen falta muestras *muy grandes* para distingur la distribución correcta, especialmente porque las *colas* –precisamente las regiones con menos datos– son las más influyentes.

2 El modelo de posición

Suponemos que el resultado x_i de cada observación depende del "valor verdadero" μ del parámetro desconocido, y también de algún proceso de error. La suposición más simple es que el erro actúa en forma aditiva, o sea

$$x_i = \mu + u_i \ \ (i = 1, ..., n)$$

donde los errores $u_1, .., u_n$ son variables aleatorias.

Esto es el llamado *modelo de posición.*

Si las observaciones son réplicas independientes del mismo experimento bajo las mismas condiciones, se puede suponer que

- $u_1, .., u_n$ tienen la misma función de distribución F_0

- $u_1, .., u_n$ son independientes.

Esto implica que $x_1, ..., x_n$ son independientes con la misma función de distribución:

$$F(x) = F_0(x - \mu),$$

y diremos que las x_i son variables aleatorias *i.i.d.* (independientes e idénticamente distribuidas).

La suposición de que no hay errores sistemáticos de medición se puede representar con:

- u_i y $-u_i$ tienen la misma distribución, y por lo tanto $F_0(x) = 1 - F_0(-x)$ ("F_0 es una distribución *simétrica* respecto de 0").

Un *estimador* $\widehat{\mu}$ es una función de las observaciones: $\widehat{\mu} = \widehat{\mu}(x_1, ..., x_n) = \widehat{\mu}(\mathbf{x})$

Una manera de medir la aproximación de $\widehat{\mu}$ a μ es con el *error medio cuadrático (EMC)*

$$\mathsf{EMC}(\widehat{\mu}) = \mathsf{E}(\widehat{\mu} - \mu)^2$$

El EMC se puede descomponer como

$$\text{EMC}(\widehat{\mu}) = \text{Var}(\widehat{\mu}) + \text{Sesgo}(\widehat{\mu})^2,$$

con

$$\text{Sesgo}(\widehat{\mu}) = \text{E}\widehat{\mu} - \mu.$$

Nótese que si $\widehat{\mu}$ es la media muestral y c es cualquier constante, entonces

$$\widehat{\mu}(x_1 + c, ..., x_n + c) = \widehat{\mu}(x_1, ..., x_n) + c$$

and

$$\widehat{\mu}(cx_1, ..., cx_n) = c\widehat{\mu}(x_1, ..., x_n)$$

Lo mismo ocurre con la mediana.

Estas dos propiedades se llaman respectivamente *equivariancia por posición y por escala.*

2.1 Alternativas a la normalidad

Una manera tradicional de repreentar datos "buenos" –o sea, sin outliers– es suponer que F_0 es normal con media 0 y varianza desconocida σ^2, lo que implica

$$F = \mathcal{D}(x_i) = \mathsf{N}(\mu, \sigma^2),$$

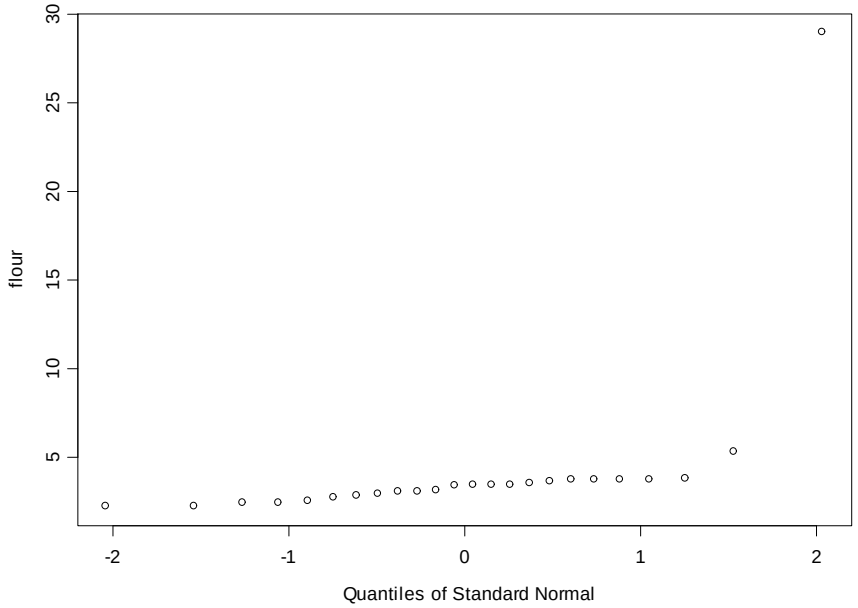
donde $D(x)$ denota la distribución de la variable x, y $\mathsf{N}(\mu, v)$ es la distribución normal con media μ y varianza v.

Los métodos clásicos suponen que F pertenece a una familia de distribuciones *exactamente conocida.*

Si los datos fueran *exactamente* normales, la media muestral sería un estimador "óptimo": es el de máxima verosimilitud, y minimiza el EMC entre los estimadores insesgados, y también entre los equivariantes.

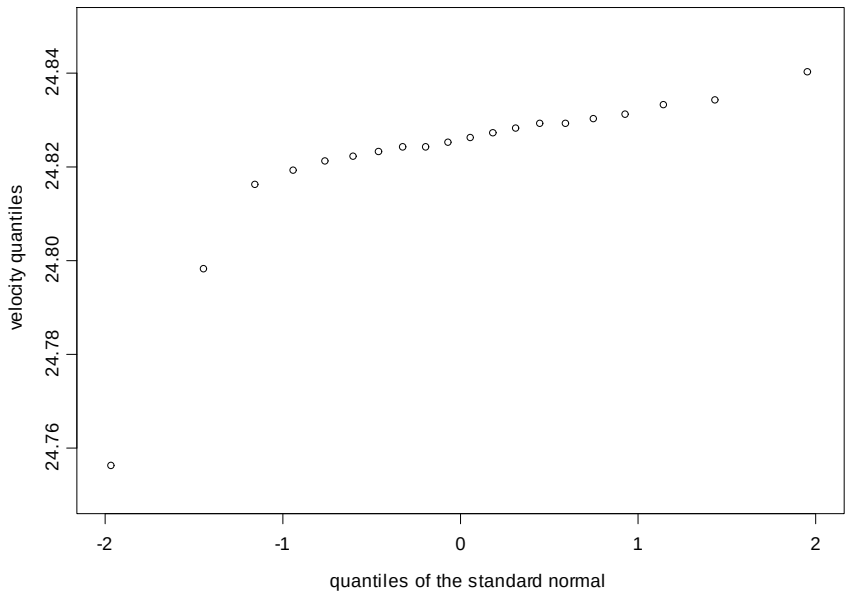
Pero la realidad no suele comportarse siempre tan bien.

La siguiente figura muestra el Q-Q del dataset de contenido de cobre en harina mostrado al principio:



Cobre en harina: QQ-plot

Y el siguiente es de un dataset històrico de mediciones de la velocidad de la luz:



Velocidad de la luz: QQ-plot de los tiempos medidos

Vemos que la *mayoría* de los datos pueden ser descriptos por una distribuión normal, pero no la totalidad de ellos.

En este sentido, podemos considerar a F como *aproximadamente normal*, con la normalidad fallando en las colas.

Podemos entonces plantear nuestro objetivo inicial como: buscar estimadores

que son "casi tan buenos" como la media muestral cuando F es exactamente normal
pero que son también "buenos" en algún sentido cuando F es sólo aproximadamente normal

Para formalizar la idea de "normalidad aproximada" , podemos imaginar que una proporción $1 - \varepsilon$ de los datos es generada por el modelo normal, mientras que una proporción ϵ es generada por un mecanismo desconocido, o sea

$$F = (1 - \epsilon)G + \epsilon H$$

donde $G = \mathsf{N}(\mu, \sigma^2)$ y H puede ser cualquier otra distribución; por ejemplo, otra normal con una varianza mayor y una media posiblemente distinta.

Esto se llama una *distribución normal contaminada.*

En general, F es llamada una *mezcla* de G y H, y se la llama una *mezcla normal* cuando tanto G como H son normales.

Si G y H tienen densidades g y h, respectivamente, F tiene densidad

$$f = (1 - \varepsilon)g + \varepsilon h.$$

Otro modelo para outliers son las llamadas *distribuciones con colas pesadas,* es decir, distribuciones cuya densidad tiende a cero más lentamente que las colas de la normal.

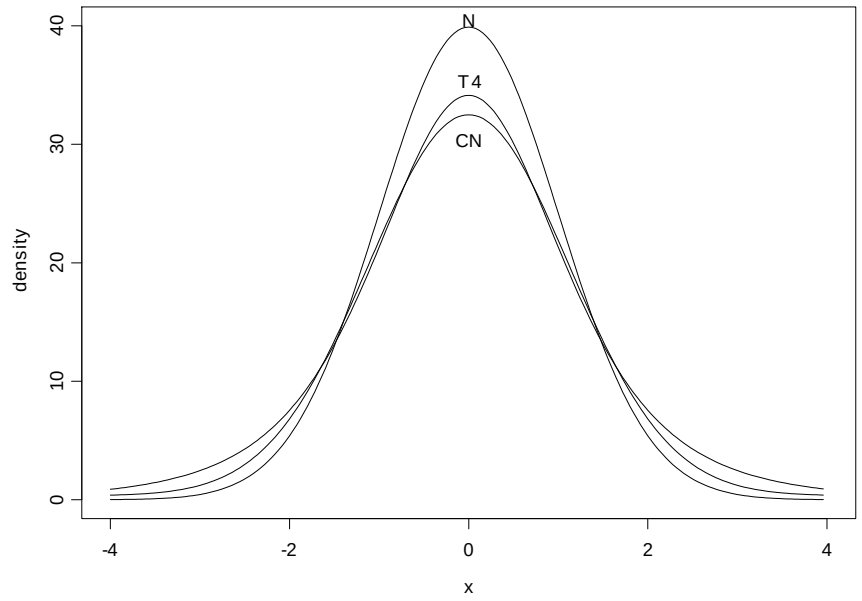
Un caso extremo es la llamada *distribución de Cauchy*, con densidad

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

La siguiente figura muestra las densidades de $\mathsf{N}(0,1)$, la de la distribución de Student con 4 grados de libertad, y la de la distribución normal contaminada

$$0.9\mathsf{N}(0,1) + 0.1\mathsf{N}(0,100)$$

indicadas por N, T4 y CN respectivamente.



Densidades de la Normal Standard (N), Student (T4), y normal contaminada(CN) escaladas para tener el mismo recorrido intercuartiles.

Para facilitar la comparación, las tres distribuciones están normalizadas para tener el mismo recorrido intercuartiles.

2.1.1 Eficiencia

Recordemos que si $x \sim \mathsf{N}(0, \sigma^2)$, entonces la media muestral $\overline{x} \sim \mathsf{N}(\mu, \sigma^2/n)$ (donde "$\sim$" significa "tiene distribución").

Como veremos luego, la mediana muestral es aproximadamente $\mathsf{N}(\mu, 1.57\sigma^2/n)$, de modo que para muestras normales, la mediana tiene un aumento de la varianza de 57% comparada con la media.

Decimos que la mediana tiene *baja eficiencia* respecto de la media, para datos normales.

La siguiente Tabla muestra para n "grande" el valor de $n \times$ **varianza**, de la media y le mediana, para diferentes valores τ en el modelo

$$F = (1-\varepsilon)\mathsf{N}(\mu, 1) + \varepsilon \mathsf{N}(\mu, \tau^2)$$

ε	0.05		0.10	
τ	$n\mathsf{Var}(\overline{x})$	$n\mathsf{Var}(\mathrm{Med})$	$n\mathsf{Var}(\overline{x})$	$n\mathsf{Var}(\mathrm{Med})$
3	1.40	1.68	1.80	1.80
4	1.75	1.70	2.50	1.84
5	2.20	1.70	3.40	1,86
6	2.75	1.71	4.50	1.87
10	5.95	1.72	10.90	1.90
20	20.9	1.73	40.90	1.92

Table 1: Varianzas ($\times n$) ofde media y mediana para n grande

Se ve que los valores para la media aumentan rápidamente con τ, mientras que los de la mediana se estabilizan.

En lo que sigue desarrollaremos estimadores que combinan la baja varianza de la media en la normal, con la robustez de la mediana bajo contaminación.

Po el momento nos limitaremos a distribuciones *simétricas.*

3 M-estimadores de posición

Ahora desarrollaremos una familia general de estimadores que contiene la media y la mediana como casos especiales.

3.1 Generalizando Màxima Verosimilitud

Volvemos al modelo de posición:

$$x_i = \mu + u_i \text{ con } u_i \sim F_0.$$

Si conociéramos F_0 exactamente, el EMV sería "óptimo": tiene la menor varianza asintótica dentro de una clase "razonable" de estimadores.

Pero como la conocemos sólo aproximadamente, nuestro objetivo será encontrar estimadores que sean

(A) "casi òptimos" cuando F_0 es exactamente normal,

y también

(B) "casi òptimos" cuando F_0 es aproximadamente normal (por ejemplo, normal contaminada).

Si la densidad $f_0 = F_0'$ es siempre positiva, dado que el logaritmo es una función creciente, el EMV puede escribirse como

$$\widehat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho\left(x_i - \mu\right)$$

donde

$$\rho = -\ln\ f_0.$$

Si $F_0 = \mathsf{N}(0,1)$ es

$$f_0(x) = \frac{1}{\sqrt{2\pi}}\ \exp^{-x^2/2}$$

y (salvo una constante) es $\rho(x) = x^2/2$.

Por lo tanto el EMV es

$$\widehat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} (x_i - \mu)^2,$$

lo que implica $\widehat{\mu} = \overline{x}$.

Si F_0 es la distribución "doble exponencial"

$$f_0(x) = \frac{1}{2} e^{-|x|}$$

resulta $\rho(x) = |x|$, y el EMV es

$$\widehat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} |x_i - \mu|,$$

lo que implica $\widehat{\mu} =$ mediana muestral.

Si ρ es diferenciable, derivar respecto de μ produce la "ecuación de estimación"

$$\sum_{i=1}^{n} \psi(x_i - \widehat{\mu}) = 0$$

con

$$\psi = \rho'.$$

Si ψ es discontinua, la solución puede no existir, y en tal caso interpretamos la ecuación en el sentido de que el primer miembro cambia de signo en μ.

Si $\rho(x) = x^2/2$, entonces $\psi(x) = x$,
y la ecuación de estimación resulta

$$\sum_{i=1}^{n} (x_i - \widehat{\mu}) = 0.$$

cuya solución es $\widehat{\mu} = \overline{x}$.

Definición: Dada una función ρ, un *M-estimador de posición* es una solución de

$$\widehat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho\left(x_i - \mu\right).$$

Si ρ es diferenciable, derivar respecto de μ produce la "ecuación de estimación"

$$\sum_{i=1}^{n} \psi(x_i - \widehat{\mu}) = 0$$

con

$$\psi = \rho'.$$

Esta familia contiene a los EMV, pero también estimadores que no son EMV para ninguna distribución.

Si ρ es convexa, entonces ψ es no decreciente.

Si ψ es estrictamente creciente, la solución es única, y si no, es todo un intervalo.

Si ρ no es convexa, hay soluciones múltiples.

Es fácil mostrar que son equivariantes por posición.

La media y la ediana son también equivariantes por escala, pero esto no se cumple en general para los M-estimadores. Màs tarde se verá cómo se arregla esto.

Desde ahora se supondrá ρ par, y por lo tanto ψ impar.

3.2 La distribución de los M-estimadores

Para evaluar la performance de los M-estimadores, hay que calcular su distribución.

Salvo para la media y la mediana, no hay expresiones explícitas para la distribución de M-estimadores en muestras finitas, pero se pueden encontrar aproximaciones para muestras "grandes".

Para una distribución F, sea $\mu_0 = \mu_0(F)$

$$\mu_0 = \arg\min_{\mu} \mathsf{E}_F \rho\left(x - \mu\right),$$

que cumple

$$\mathsf{E}_F \psi(x - \mu_0) = 0.$$

Para la media muestral, $\psi(x) = x$, lo que iplica $\mu_0 = \mathsf{E}x$, o sea, la media de F.

Para la mediana muestral es $\mu_0 = \mathsf{Med}(x)$, o sea, la mediana de F.

En general, si F es simétrica, μ_0 coincide con el centro de simetría de F.

Se puede mostrar que cuando $n \to \infty$,

$$\widehat{\mu} \to_p \mu_0$$

donde "$\to_p$" significa "tiende en probabilidad" (decimos que $\widehat{\mu}$ es "*consistente* para μ_0"), y que la distribución de $\widehat{\mu}$ es aproximadamente

$$\mathsf{N}\left(\mu_0, \frac{v}{n}\right) \text{ con } v = \frac{\mathsf{E}_F \psi(x-\mu_0)^2}{\left(\mathsf{E}_F \psi'(x-\mu_0)\right)^2}.$$

Si la distribuciòn de un estimador $\widehat{\mu}$ es aproximadamente $\mathsf{N}(\mu_0, v/n)$ para n grande, decimos que $\widehat{\mu}$ es *asintóticamente normal,*
con *valor asintótico* μ_0 y *varianza asintótica* v.

La *efciencia asintótica* de $\widehat{\mu}$ bajo la distribuciòn F es el cociente

$$\mathsf{Efi}(\widehat{\mu}) = \frac{v_0}{v},$$

donde v_0 es la varianza asintótica del EMV
y mide cuán cerca está $\widehat{\mu}$ del òptimo bajo F.

Para la media muestral es $\psi' \equiv 1$ y por lo tanto $v = \mathsf{Var}(x)$.

Para la mediana muestral, puede probarse que

$$v = \frac{1}{4f(\mu_0)^2},$$

donde $f = F'$; y por lo tanto para $F = \mathsf{N}(0, 1)$ tenemos

$$v = \frac{2\pi}{4} = 1.571,$$

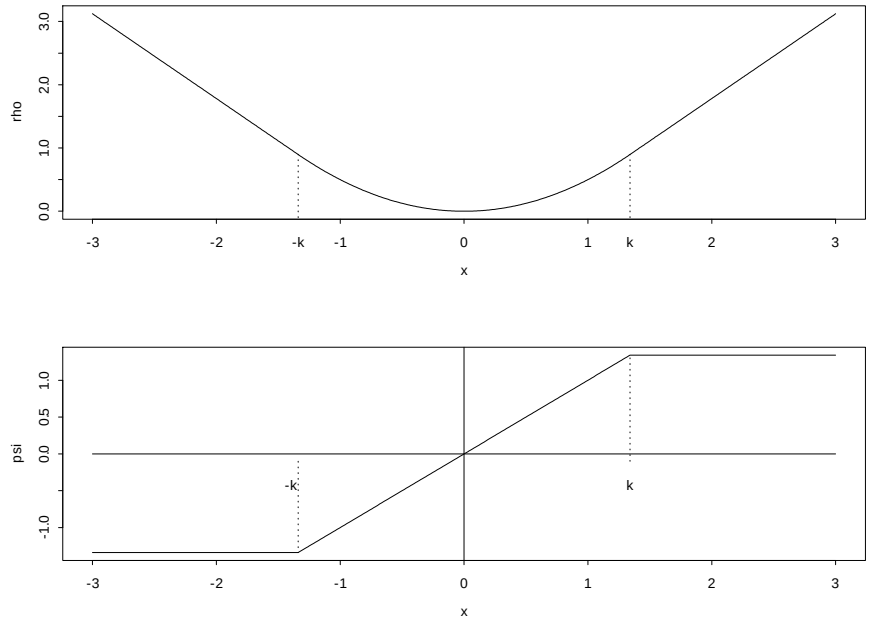
así que la eficiencia de la mediana es $1/1.571$=0.636.

Una familia de funciones ρ y ψ con propiedades importantes es la de *funciones de Huber* graficadas en la prósima figura:

$$\rho_k(x) = \begin{cases} x^2 & \text{si} \quad |x| \leq k \\ 2k\,|x| - k^2 & \text{si} \quad |x| > k \end{cases}$$

con derivada $2\psi_k(x)$ donde

$$\psi_k(x) = \begin{cases} x & \text{si} \quad |x| \leq k \\ \mathsf{sgn}(x)k & \text{si} \quad |x| > k \end{cases}$$



Funciones ρ y ψ de Huber

Se ve que ρ_k es cuadrática en un intervalo central, pero crece sólo linealmente fuera de él

Los M-estimadores correspondientes a los casos límites $k \to \infty$ y $k \to 0$ son respectivamente la media y la mediana (definimos $\psi_0(x)$ como $\mathsf{sign}(x)$).

El valor de k se elige para obtener una varianza (y por lo tanto una eficiencia) deseada en la normal.

La siguiente tabla muestra la varianza asintótica del estimador bajo el modelo normal contaminado

$$F = (1 - \varepsilon)\mathsf{N}(0, 1) + \varepsilon\mathsf{N}(0, 10)$$

para diferentes valores de ε y k.

Aquí se ve el "regateo" entre robustez y eficiencia:

k	$\varepsilon = 0$	$\varepsilon = 0.05$	$\varepsilon = 0.10$
0	1.571	1.722	1.897
1.0	1.107	1.263	1.443
1.4	1.047	1.227	1.439
2.0	1.010	1.259	1.550
∞	1.000	5.950	10.900

Table 2: Varanzas asintóticas del estimador de Huber

- Cuando $k = 1.4$, la varianza del M-estimador bajo la normal es sólo 4.7% mayor que la de $\overline{x}$ (que corresponde a $k = \infty$) y mucho menor que la de la mediana (que corresponde a $k = 0$), mientras

- paa normales contaminadas el claramente menor que ambas.

3.3 Una visión intuitiva de los M-estimadores

Un M-estimador de posición se puede ver como una media pesada.

En la mayoría de los casos que interesan, es $\psi(0) = 0$, y $\psi'(0)$ existe, de modo que ψ es aproximadamente lineal en el origen. Sea

$$W(x) = \begin{cases} \psi(x)/x & \text{si} \quad x \neq 0 \\ \psi'(0) & \text{si} \quad x = 0 \end{cases}$$

(la "función de peso"). Entonces la ecuación de estimación puede escribirse como

$$\sum_{i=1}^{n} W(x_i - \widehat{\mu})(x_i - \widehat{\mu}) = 0,$$

lo que implica

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \text{ con } w_i = W(x_i - \widehat{\mu}),$$

lo que expresa al estimador como una media pesada.

Notar que aunque esto parece una expresión explícita para $\widehat{\mu}$, en realidad los pesos w_i dependen de $\widehat{\mu}$.

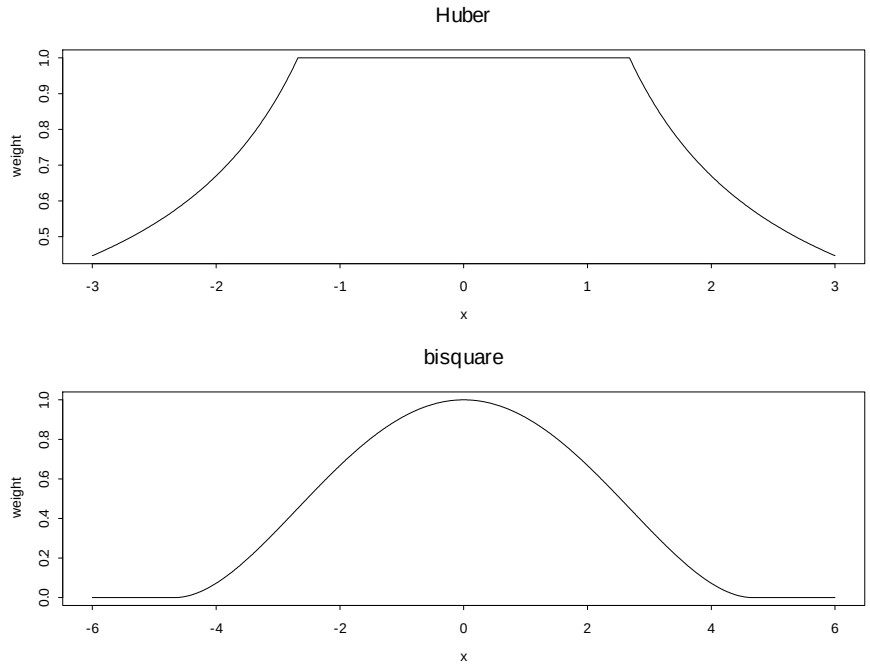
Como en general W es decreciente, las observaciones recibirán pesos menores cuanto más alejadas estén del "centro" de los datos.

Además de su valor intuitivo, esta representación será útil para el cáculo numérico.

La función de peso correspondiente a la ψ de Huber es

$$W_k(x) = \min\left\{1, \frac{k}{|x|}\right\}$$

que se muestra en el panel superior de la próxima figura, junto con otra que trataremos luego.



Funciones de peso W Huber y bisquare

3.4 M-estimadores redescendientes

Es fácil mostrar que el EMV para la densidad de Student con ν grados de libertad es

$$\psi(x) = \frac{x}{x^2 + \nu},$$

que tiende a 0 cuando $x \to \infty$.

Esto sugiere que para distribuciones simétricas con colas pesadas, es mejor usar una ψ "redescendiente" que tienda a 0 en infinito.

Esto implica que para x grande, la ρ correspondiente crece más lentamente que la de Huber, que es lineal para $|x| > k$.

Una elección popular es la familia *bisquare* (o *biweight*):

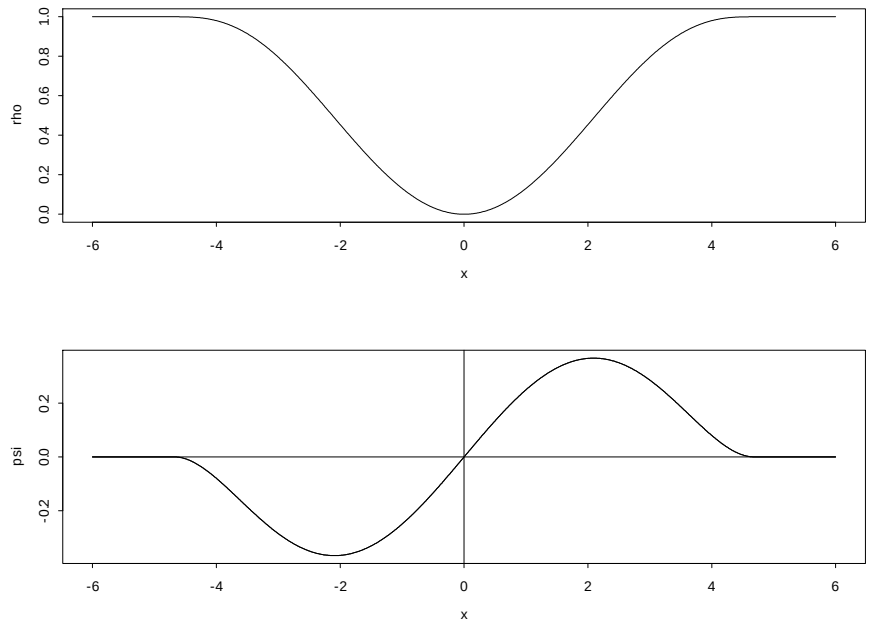
$$\rho(x) = \begin{cases} 1 - \left(1 - (x/k)^2\right)^3 & \text{if} \quad |x| \leq k \\ 1 & \text{if} \quad |x| > k \end{cases}$$

con derivada $\rho'(x) = 6\psi(x)/k^2$ donde

$$\psi(x) = x\left(1 - \left(\frac{x}{k}\right)^2\right)^2 \mathsf{I}(|x| \leq k).$$

donde "I" es el indicador.

Estas funciones se muestran en la próxima figura.



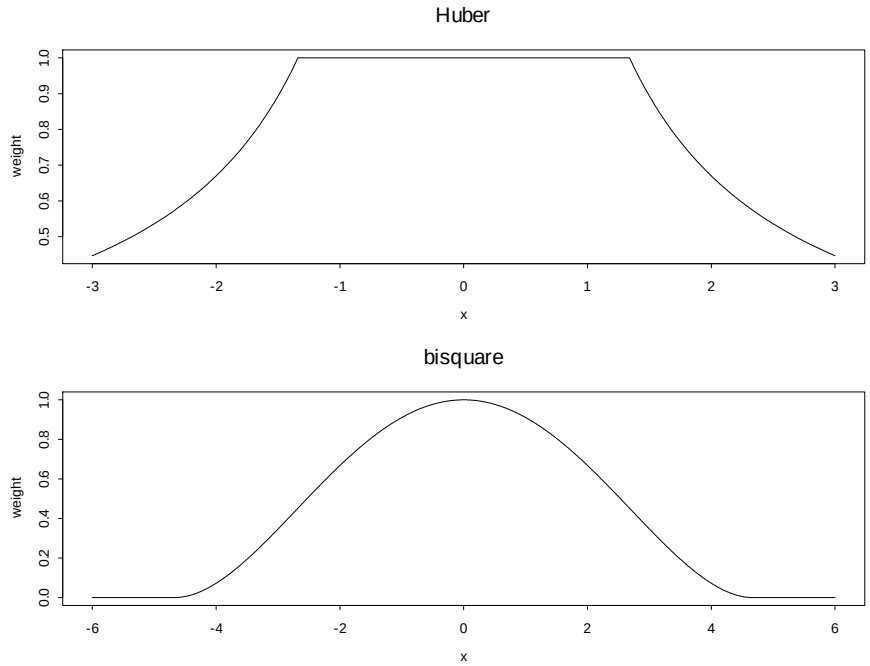
Funciones ρ y ψ para el estimador bisquare

Notar que ψ es diferenciable y se anula fuera de $[-k, k]$.

La correspondiente función de peso es

$$W(x) = \left(1 - \left(\frac{x}{k}\right)^2\right)^2 \mathsf{I}(|x| \leq k)$$

que volvemos a mostrar.



Funciones de peso W Huber y bisquare

Estos estimadores mejoran la resistencia a outlers grandes.

4 M-estimadores de posición con escala desconocida

Los estimadores definidos por

$$\widehat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho\left(x_i - \mu\right)$$

no son equivariantes por escala. Es natural exigir que si multiplico los datos por 100, el estimador se multiplique por 100.

Para arreglar este problema necesitamos un estimador robusto de dispersión.

Para fijar ideas, supongamos que queremos estimar μ en el modelo de posición con

$$F = (1 - \varepsilon)\mathsf{N}(\mu, \sigma^2) + \varepsilon H.$$

Si σ fuera conocida, sería natural estimar μ con

$$\widehat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{\sigma}\right).$$

Una idea intuitiva es usar

$$\widehat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{\widehat{\sigma}}\right),$$

donde $\widehat{\sigma}$ es un estimador de dispersión calculado previamente.

Entonce $\widehat{\mu}$ resulta equivariante por escala.

Como $\widehat{\sigma}$ no depende de μ, resulta que $\widehat{\mu}$ es solución de

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \widehat{\mu}}{\widehat{\sigma}}\right) = 0.$$

No podemos usar como $\widehat{\sigma}$ el desvío standard (DS), porque estaría afectado por los outliers. Se necesita un $\widehat{\sigma}$ *robusto.*

Al mismo tiempo, para datos normales $\widehat{\sigma}$ deberia coincidir con el DS para mantener la eficiencia elegida.

Se define la MAD (median absolute deviation) como

$$\mathsf{MAD}\,(x) = \mathsf{Med}\,(|x - \mathsf{Med}\,(x)\,|)\,.$$

Se prueba que si $x \sim \mathsf{N}\left(\mu, \sigma^2\right)$ es $\mathsf{MAD}\,(x) = 0.675\sigma$.

En consecuencia se define la MAD normalizada como

$$\mathsf{MADN}\,(x) = \frac{\mathsf{MAD}\,(x)}{0.675},$$

y entonces para x normal es $\mathsf{MADN}\,(x) = \sigma$.

Este es un estimador muy robusto de dispersión. Para los datos de harina, la MAD con y sin el outlier es 0.35 y 0.34.

Elegimos entonces tomar $\widehat{\sigma}$ como la MADN de la muestra.

5 Cálculo numérico de los M-estimadores

Describiremos un procedimiento llamado *reponderación iterativa* basado en la presentación intuitiva que ya vimos.

La expresiòn de $\widehat{\mu}$ como media pesada

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \text{ con } w_i = W(x_i - \widehat{\mu})$$

sugiere un procedimiento iterativo.

Sean $\widehat{\sigma}_0$ un estimador robusto de dispersión (por ejemplo MADN) y $\widehat{\mu}_0$ un estimador inicial (por ejemplo la mediana). Ponemos $m = 0$.

Dado $\widehat{\mu}_m$ calcular

$$w_{m,i} = W\left(\frac{x_i - \widehat{\mu}_m}{\widehat{\sigma}}\right) \ (i = 1, .., n)$$

y

$$\widehat{\mu}_{m+1} = \frac{\sum_{i=1}^{n} w_{m,i} x_i}{\sum_{i=1}^{n} w_{m,i}}.$$

Si $W(x)$ es acotada y es decreciente para $x > 0$, la sucesiòn $\widehat{\mu}_m$ converge a una sollución

Si ψ es estrictamente creciente la solución es única, y el punto de partida $\widehat{\mu}_0$ sólo influye en la cantidad de iteraciones.

Si ψ es redescendiente, $\widehat{\mu}_0$ tiene que ser robusto para asegurar la convergencia a una solución "buena".

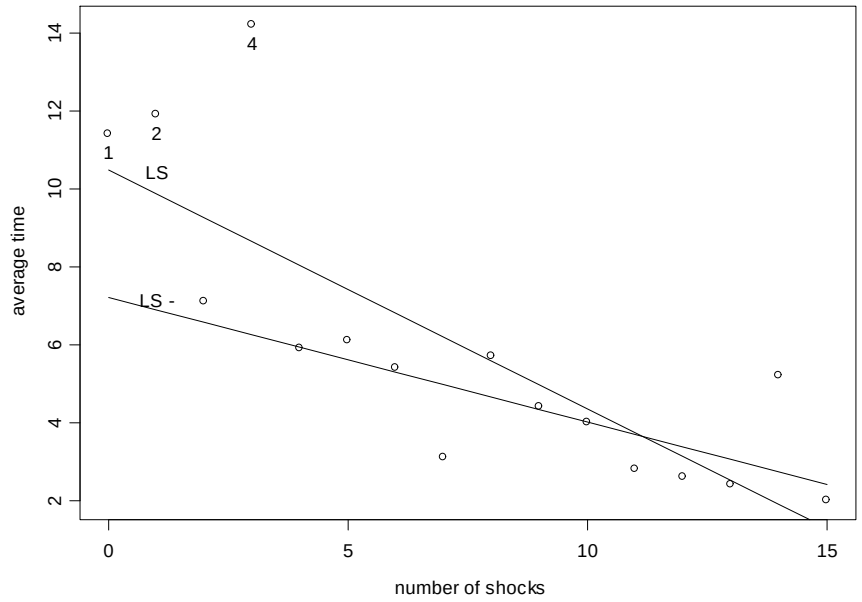
Para los datos de harina, los valores del M-estimador bisquare con eficiencia 0.95, con y sin el outlier, son 3.14 y 3.12, o sea que no es casi afectado por éste. La media sin el outlier era 3.40.

6 Regresión lineal (trailer)

El estimador clásico para regresión lineal es el de *mínimos cuadrados* (EMC).

Pese a sus virtudes, es muy sensible a outliers.

Ejemplo: Los datos corresponden a un experimento sobre la velocidad de aprendizaje de ratas. Para abreviar se omite la descripción detallada. El grafico muestra en la abscisa la cantidad de shocks, y en la ordenada el tiempo promedio de los intentos, para cada rata.



Ajuste del EMC con todos los datos, y omitiendo los puntos 1-2-4.

The Figure muestra los datos y las líneas ajustadas por MC para el m odelo

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

La relación entre las variables se ve como aproximadamente lineal, excepto para los 3 puntos en el ángulo superior izquierdo.

La línea del EMC no ajusta la mayoría de los datos, sindo un compromiso entre esos 3 puntos y el resto.

El ajuste sin los 3 puntos da una mejor representación de la mayoria de los datos, y muestra el carácter excepcional de 1-2-4.

Buscamos procedimientos que den un buen ajuste a la mayoria de los datos, sin ser perturbados por una pequeña proporción de outliers.

Consideremos en general un dataset de n observaciones $\left(x_{i1}, ..., x_{ip}, y_i\right)$, $i = 1, ..., n$, donde

- $x_{i1}, ..., x_{ip}$ son *predictores*, y
- y_i es la *respuesta.*

Se supone que los datos cumplen el *modelo lineal*

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + u_i, \qquad i = 1, ..., n$$

donde $\beta_1, ..., \beta_p$ son parámetros desconocidos que hay que estimar, y las u_i son variables aletorias ("errores").

Poniendo

$$\mathbf{x}_i = (x_{i1}, ..., x_{ip})', \quad \beta = (\beta_1, ..., \beta_p)',$$

el modelo puede ser escrito en forma más compacta como

$$y_i = \mathbf{x}_i'\beta + u_i$$

donde $\mathbf{x}'$ is el traspuesto de $\mathbf{x}$.

Los *valores ajustados* $\widehat{y}_i$ y los *residuos* r_i correspondientes a un vector β se definen como

$$\widehat{y}_i = \widehat{y}_i(\beta) = \mathbf{x}_i'\beta \text{ y } r_i = r_i(\beta) = y_i - \widehat{y}_i(\beta).$$

El EMC cumple las *ecuaciones normales*

$$\sum_{i=1}^{n} r_i(\widehat{\beta})\mathbf{x}_i = \mathbf{0}.$$

Trataremos los M-estimadores de regresión $\widehat{\beta}$ definidos como solución de

$$\sum_{i=1}^{n} \rho\left(\frac{r_i(\widehat{\beta})}{\widehat{\sigma}}\right) = \min .$$

Aquí ρ es una función como las que vimos en el modelo de posición, y $\widehat{\sigma}$ es un estimador auxiliar de escala, calculado previamente, necesario para hacer $\widehat{\beta}$ equivariante por escala.

Esta clase contiene a los EMV.

El EMC corresponde a $\rho(t) = t^2$.

Para $\rho(t) = |t|$ tenemos el *estimador* L_1 (o "LAD": Least Absolute deviations"), que es el equivalente de la mediana para regresión.

Derivando respecto de β resulta el análogo de las ecuaciones normales:

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\widehat{\beta})}{\sigma}\right)\mathbf{x}_i = \mathbf{0},$$

Recordemos que para M-estimadores de posición, estimàbamos σ mediante la MAD.

Aquí el procedimiento equivalente es calcular el estimador L_1, $\widehat{\beta}_{L_1}$ y los correspondientes residuos, y de allí obtener el análogo de la MAD normalizada tomando la mediana de los resiudos no nulos:

$$\widehat{\sigma} = \frac{1}{0.675}\mathsf{Med}_i(|r_i|\left(\widehat{\beta}_{L_1}\right) \;\Big|\; r_i \neq 0).$$

El motivo para usar sólo los residuos no nules es que el L_1 produce al menos p resiuos nulos, y por lo tanto usar todos los residuos puede producir una σ subestimada.

Nótese que L_1 no necesita una escala auxiliar.

7 Interpretaciòn intuitiva y cálculo numérico

Supongamos ψ diferenciable y definamos

$$W(x) = \frac{\psi(x)}{x}.$$

Entonces las ecuaciones de estimaciòn pueden escribirse como

$$\sum_{i=1}^{n} w_i r_i \mathbf{x}_i = \sum_{i=1}^{n} w_i \mathbf{x}_i \left(y_i - \mathbf{x}_i' \widehat{\beta} \right) = \mathbf{0}$$

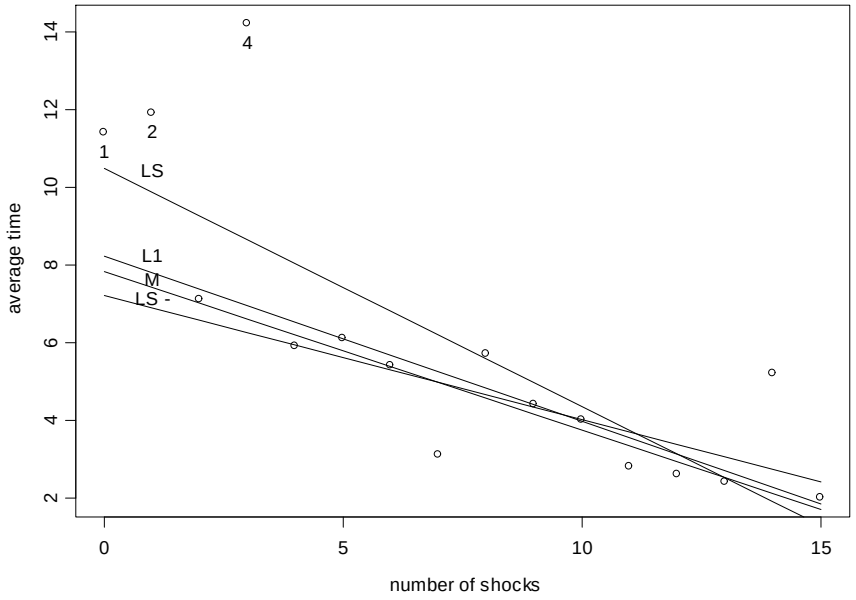
con $w_i = W(r_i/\widehat{\sigma})$.

Esto muestra al estimador como un EMC ponderado.

Al mismo tiempo sugiere un algoritmo iterativo para su cómputo, análogo al que se mostró para posiciòn.

Volvemos al ejemplo inicial. Se muestran los ajustes por EMC, EMC sin los outliers, L_1, y el M-estimador con la función bisquare que se definió para posición.

Se ve que el bisquare es el que más se parece al EMC sin outliers.



Ajustes por EMC (LS), L_1, M-estimador Bisquare (M) y EMC sin los outliers (LS-)

Referencias

Maronna, R.A., Martin, R.D. Yohai, V.J. y Salibián-Berrera, M. (20019). *Robust Statistics: Theory and Methods (with R). Second Edition.* . John Wiley and Sons, New York.