Virgilio L. Foglia

April 6, 2008

Contents

1 Funcionales

1.1 Modelo paramétrico

Sea una muestra de n variables aleatorias independientes $X_1, X_2, \cdots, X_n$, todas definidas en el mismo espacio muestral $\Omega \subset \mathbb{R}$, y con la misma distribución F_θ , donde $\theta \in \Theta$, siendo Θ el espacio de parámetros. Se define modelo paramétrico a la familia de todas estas posibles funciones de distribución

$$P_\theta = \{F_\theta : \theta \in \Theta\}$$

En estadística clásica se asume que $X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F$, donde $F \in P_\theta$; o sea será $F = F_\theta$ totalmente identificada para cierto $\theta \in \Theta$. Y el problema de estimación consiste en obtener una función de la muestra $g_n^P(X_1, X_2, \cdots, X_n)$ para estimar θ.

En estadística robusta se acepta que el modelo paramétrico $P_\theta = \{F_\theta : \theta \in \Theta\}$ es solo una aproximación de la realidad, y que en la práctica puede darse que la muestra $X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F$ donde $F \notin P_\theta$, sinó a una familia un poco mayor $\mathcal{F}$, donde $P_\theta \subset \mathcal{F}$(más adelante definiremos esta familia). Es decir, no solo interesarán las distribuciones de los estimadores bajo el modelo P_θ, sinó también suponiendo desviaciones respecto este modelo.

Como se asume que F perteneciente a una familia $\mathcal{F}$ mas amplia, se buscará una función estimadora $g_n(X_1, X_2, \cdots, X_n)$, tal que:

1) si $F \in P_\theta$, el estimador g_n sea casi tan bueno como g_n^P (1)
2) si $F \in \mathcal{F} - P_\theta$, el estimador g_n sea, no obstante, también bastante bueno

1.2 Distribución empírica de una muestra

Sea la muestra $X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F$, se define la distribución empírica

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \quad \text{o también con la función impulso: } F_n = \frac{1}{n}\sum_{i=1}^{n} \Delta_{X_i}$$

que en realidad, para cada x, es una variable aleatoria, obtenida como promedio de las indicadoras $I(X_i \leq x)$. Como

$$\begin{aligned} E\left[I(X_i \leq x)\right] &= F(x) \\ Var\left[I(X_i \leq x)\right] &= F(x)(1 - F(x)) \end{aligned}$$

resulta $\sum_{i=1}^{\infty} \sigma_i^2 / i^2 = \sum_{i=1}^{\infty} \frac{F(x)(1-F(x))}{i^2} < \infty$, y entonces vale la ley fuerte, o sea

$$\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) - F(x) \overset{\textbf{c.t.p.}}{\longrightarrow} 0$$

es decir

$$F_n(x) \overset{\textbf{c.t.p.}}{\longrightarrow} F(x)$$

Es más, según el teorema de Glivenko-Cantelli se prueba que vale una convergencia todavía más fuerte, que es la convergencia uniforme en x

$$\sup_x |F_n(x) - F(x)| \overset{\textbf{c.t.p.}}{\longrightarrow} 0$$

1.3 Funcionales y estimadores

1.3.1 De funcionales a estimadores

Sea una variable aleatoria $X \sim F$, con $F \in P_\theta = \{F_\theta : \theta \in \Theta\}$ un modelo parámetrico. En muchos casos θ puede ser pensado como un funcional, o sea expresarlo así: $\theta = T(F) \ \ \forall F \in P_\theta$.

Por ejemplo si P_θ esta compuesto por todas las distribuciones exponenciales de parámetro θ

$$P_\theta = \left\{ F_\theta = 1 - e^{-x/\theta} : \theta \in \mathbb{R}^+ \right\}$$

como $E_F(X) = \int_0^\infty x dF = \theta$, resulta $\theta = E_F(X) = \int_0^\infty x dF = T(F)$.

Ahora bién, si se tiene una muestra $X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F$ donde $F \in P_\theta$, y su correspondiente distribución empírica F_n, parece natural proponer como estimador de $\theta = T(F)$, a $\widehat{\theta}_n = T(F_n)$. Este es el llamado principio de sustitución(plug-in en ingles). De esta manera queda expresado el estimador $\widehat{\theta}_n$ como un funcional de la distribución empírica.

1.3.2 De estimadores a funcionales

Pero en estadística robusta se estudiarán muestras $X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F$, en donde F se admite que se desvíe un poco de P_θ, (o sea $F \notin P_\theta$, sinó que $F \in \mathcal{F} \supset P_\theta$). En este caso quizás no se pueda expresar $\theta = T(F) \ \forall F \in \mathcal{F}$.Entonces se buscará un estimador que cumpla los dos requisitos de (1) obteniendo

$$\widehat{\theta}_n = g_n(X_1, X_2, \cdots, X_n)$$

siendo deseable que este estimador pueda ser expresado como funcional de la distribución empírica, o sea que $\forall n$ y F_n, exista un funcional $T : \mathcal{F} \longrightarrow \mathbb{R}$ tal que

$$\widehat{\theta}_n = g_n(X_1, X_2, \cdots, X_n) = T(F_n)$$

o al menos que asintóticamente pueda ser reemplazado por un funcional

$$\widehat{\theta}_n = g_n(X_1, X_2, \cdots, X_n) \xrightarrow{\textbf{p}} T(F)$$

donde $T(F)$ es el valor asintótico de la secuencia $\widehat{\theta}_n$ en F. En definitiva, se tratará de expresar los estimadores como funcionales, siendo para esto útil el siguiente resultado.

Proposition 1 *Sea la muestra $X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F$, y F_n la distribución empírica. Se calculará la esperanza de una función cualquiera $g(X)$, pero suponiendo que $X \sim F_n$*

$$E_{F_n}(g(X)) = \frac{1}{n} g(x_1) + \cdots + \frac{1}{n} g(x_n) = \frac{1}{n} \sum_{i=1}^{n} g(x_i) \tag{2}$$

Example 2 *Si se tiene el estimador* $\widehat{\theta}_n = \sqrt[n]{X_1 X_2 \cdots X_n}$, *como* $\ln \widehat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \ln(X_i) = E_{F_n}(\ln(X))$ *de aquí resulta*

$$\widehat{\theta}_n = e^{E_{F_n}(\ln(X))} = e^{\int_{\mathbb{R}} \ln(x) dF_n} = T(F_n)$$

y entonces el funcional T *está definido por*

$$T(F) = e^{E_F(\ln(X))} = e^{\int_{\mathbb{R}} \ln(x) dF}$$

Example 3 *Suponga ahora el estimador* $\widehat{\theta}_n = medp(X_1, X_2, \cdots, X_n)$, *donde la mediana principal es el punto medio del intervalo de medianas. Se calculará, igual que en el ejemplo anterior, la mediana principal de una variable aleatoria* X, *pero con* $X \sim F_n$ *(definiendo mediana principal en este caso el punto medio del intervalo en que* $F_n(u) = 0.5$*). Se demuestra que vale*

$$medp_{F_n}(X) = medp(X_1, X_2, \cdots, X_n) = \widehat{\theta}_n$$

luego queda el estimador expresado como $\widehat{\theta}_n = medp_{F_n}(X) = T(F_n)$. *Y entonces el funcional* T *está definido por*

$$T(F) = medp_F(X)$$

1.4 Consistencia de funcionales

1.4.1 Consistencia

Sea la muestra $X_1, X_2, \cdots, X_n \overset{\mathbf{iid}}{\sim} F$ donde $F \in \mathcal{F} \supset P_\theta$. Si F_n es la distribución empírica, se dice que el funcional T es consistente, si vale la convergencia en probabilidad

$$T(F_n) \xrightarrow{\mathbf{p}} T(F)$$

y en este caso $T(F)$ se interpretará como el valor límite al cual tiende el estimador $\widehat{\theta}_n = T(F_n)$, cuando la muestra proviene de la distribución F.

Como según Glivenko-Cantelli $F_n \xrightarrow{\mathbf{c.t.p.}} F$, y suponiendo que T sea continuo respecto de cierta métrica, no es difícil lograr la consistencia de los funcionales que se usan en problemas de robustez.

1.4.2 Consistencia de Fisher

Con este planteo, como $F \in \mathcal{F} \supset P_\theta$, no necesariamente $T(F) = \theta \quad \forall F \in \mathcal{F}$(como sería deseable si lo que queremos es estimar θ). Entonces, al menos, parece natural pedir que si $F \in P_\theta$, (o sea $F = F_\theta$) resulte

$$T(F_\theta) = \theta \quad \forall \theta \in \Theta$$

Esta es la llamada consistencia de Fisher. Con esta condición, si el modelo paramétrico es el correcto, entonces asintóticamente, los estimadores $\widehat{\theta}_n$ estimaran θ.

Example 4 *Considérese el modelo paramétrico* $P_\theta = \left\{F_{\sigma^2} \sim N(0;\sigma^2) : \sigma^2 \in \mathbb{R}^+\right\} \subset \mathcal{F}$*, y el estimador muestral de la varianza*

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

Primero se lo expresará como funcional de la distribución empírica observando que

$$\widehat{\theta}_n = S_n^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}_n^2 = E_{F_n}(X^2) - E_{F_n}^2(X) = T(F_n)$$

Notar que este funcional T*, es consistente de Fisher, ya que si la muestra tiene distribución* $F = F_{\sigma^2} \in P_\theta$ *resulta*

$$T(F) = T(F_{\sigma^2}) = E_{F_{\sigma^2}}(X^2) - E_{F_{\sigma^2}}^2(X) = \sigma^2$$

y sin embargo el estimador $\widehat{\theta}_n$ *es sesgado ya que* $E(\widehat{\theta}_n) = \frac{n-1}{n}\sigma^2$*.*

El concepto de consistencia de Fisher es más apropiado al trabajar con funcionales, que el de consistencia o el de estimadores asintoticamente insesgados.

1.5 Funcionales y Continuidad

Sea la muestra $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} F$, con $F \in \mathcal{F} \supset P_\theta = \{F_\theta : \theta \in \Theta\}$. Y considere T un funcional consistente de Fisher, o sea si $F \in P_\theta$, resulta $T(F) = \theta$. Este requisito es importante. Pero si la muestra proviene de otra distribución G, con $G \notin P_\theta$, como se vió, no está garantizado que $T(G) = \theta$.

Sin embargo, si G esta "cerca" de F en algún sentido, desde un punto de vista robusto sería deseable que $T(G) \approx T(F) = \theta$. Habrá entonces que definir cierta forma de continuidad aplicable a estos funcionales. Y esto requiere la definición de una conveniente distancia $d(F;G)$, que sea útil en problemas de robustez.

Para comprender como debería ser esta distancia, se supondrá que si con la muestra $X_1, X_2, \cdots, X_n$ de distribución empírica F_n, se construye después la muestra $Y_1, Y_2, \cdots, Y_n$ de distribución empírica G_n, mediante:

- modificación **arbitraria** de una **pequeña** proporción de las X_i, y/o
- modificación **leve** de **todas** las X_i

entonces la distancia $d(F_n;G_n)$ debería ser "pequeña". Algunas distancias que satisfacen estas ideas intuitivas sobre cercanía en robustez son la de Levy, la de Lipschitz acotada y la de Prokhorov.

1.5.1 Robustez cualitativa

Hampel(1971) definió a un **funcional** T como cualitativamente robusto(**CR**) en F, si es continuo en F según una de estas distancias, o sea si

$$\forall \varepsilon > 0, \quad \exists \delta > 0 : d(F, G) < \delta \Longrightarrow |T(F) - T(G)| < \varepsilon \tag{3}$$

Usando el teorema de Glivenko-Cantelli se prueba que los estimadores construidos con funcionales **CR** son consistentes, o sea

$$\widehat{\theta}_n = T(F_n) \xrightarrow{\mathbf{p}} T(F)$$

Por otro lado dada la muestra $X_1, X_2, \cdots, X_n \overset{\mathbf{iid}}{\sim} F$, Hampel también definió a la **secuencia de estimadores** $\widehat{\theta}_n(X_1, X_2, \cdots, X_n)$ como cualitativamente robusta asintóticamente(**CRA**) en F, si

$$\forall \varepsilon > 0, \quad \exists \delta > 0 : d(F, G) < \delta \Longrightarrow d(\mathcal{L}_F(\widehat{\theta}_n); \mathcal{L}_G(\widehat{\theta}_n)) < \varepsilon \tag{4}$$

donde $\mathcal{L}_F(\widehat{\theta}_n)$ y $\mathcal{L}_G(\widehat{\theta}_n)$ son las distribuciones de $\widehat{\theta}_n$ cuando la muestra proviene de F y G respectivamente.

Hampel también demuestra que en el caso que los estimadores puedan ser expresados como un funcionales de la distribución empírica, o sea $\widehat{\theta}_n = T(F_n)$, entonces que el funcional T sea **CR** es equivalente a que la secuencia $\widehat{\theta}_n = T(F_n)$ sea **CRA**.

En definitiva estos conceptos expresan la idea intuitiva de **robustez** de un estimador, en el sentido de que si la distribución de la muestra no coincide con la supuesta (ya sea porque hay pequeños cambios en todas las observaciones, o grandes cambios en unas pocas), el comportamiento del estimador no variará mucho.

Robustez cualitativa es una propiedad muy básica y, por lo tanto, estimadores que no poseen esta propiedad pueden ser descartados desde el punto de vista de la robustez. Por otro lado, esta teoría no es completamente satisfactoria por cuanto no permite comparaciones dentro de la clase de estimadores que son cualitativamente robustos.

1.6 Funcionales y diferenciación

Como en teoría robusta interviene el valor que toma el funcional $T(F)$, y los estimadores son también expresados como funcionales $\widehat{\theta}_n = T(F_n)$, interesará estudiar la variación de estos funcionales para pequeños cambios en la F o F_n respectivamente. Se necesita entonces introducir el concepto de diferenciación de funcionales.

1.6.1 Derivada de Gâteau

Sean $F, G \in \mathcal{F}$, y un funcional $T : \mathcal{F} \longrightarrow \mathbb{R}$. Se define la contaminación de F por G a la distribución $F_t \in \mathcal{F}$ donde

$$F_t = (1 - t)F + tG \quad \text{para } t \in [0; 1]$$

evidentemente $F_0 = F$ y $F_1 = G$. Pero ahora interesará el valor que toma $T(F_t)$ cuando $t \longrightarrow 0^+$, es decir para pequeños valores de contaminación.

Se dice que el funcional T es diferenciable en F en la dirección de G, en el sentido de **Gâteau**, si existe el límite

$$T'_G(F) = \lim_{t \longrightarrow 0^+} \frac{T\left[(1-t)F + tG\right] - T(F)}{t} = \left. \frac{\partial}{\partial t} T(F_t) \right|_{t=0^+}$$

donde $T'_G(F)$ es la derivada de **Gâteau** en F en la dirección de G. Se define también la derivada de **Gâteau** de orden k así

$$T^{(k)}_G(F) = \left. \frac{\partial^k}{\partial t^k} T(F_t) \right|_{t=0^+}$$

En el caso que G corresponda a una variable aleatoria que toma solo el valor x con probabilidad 1, será

$$G = \Delta_x = \begin{cases} 0 & \text{para } u < x \\ 1 & \text{para } u \geq x \end{cases}$$

y la derivada se escribirá $T'_x(F)$. En este caso $T'_x(F)$ mide la variación de $T(F)$ debida a una contaminación infinitesimal de F en el punto x, relativa a la cantidad de contaminación.

1.6.2 Derivada de Fréchet

Sean $F \in \mathcal{F}$, y un funcional $T : \mathcal{F} \longrightarrow \mathbb{R}$. Se dice que T es diferenciable en F en el sentido de **Fréchet**, si existe una función $g : \mathbb{R} \longrightarrow \mathbb{R}$ tal que

$$\lim_{t \longrightarrow 0} \frac{T\left[(1-t)F + tG\right] - T(F)}{t} = E_{G-F}(g(U))$$

uniformemente en $\{G \in \mathcal{F} : d(G; F) < \epsilon\}$. Notar que entonces $E_{G-F}(g(U))$ es un funcional lineal, y es la derivada de Fréchet de T en la dirección de G. Observar además que por su definición, la derivada de Frechet debe existir para toda G en un entorno de F, y no en una sola dirección, como es el caso de la derivada de Gâteau.

Como diferenciable Fréchet $\implies$ diferenciable Gâteau, se puede expresar

$$T'_G(F) = E_{G-F}(g(U)) \quad \forall G \in \mathcal{F} \tag{5}$$

Si se toma $G = \Delta_x$ y usando la notación $T'_{\Delta_x}(F) = T'_x(F)$ se obtiene

$$T'_x(F) = E_{\Delta_x - F}(g(U)) = g(x) - E_F(g(U)) \tag{6}$$

esta función de x da la derivada de Gâteau cuando F está contaminada solo por observaciones en el punto x. No solo es más fácil de calcular, sinó que de ella se derivan dos resultados importantes.

Si se piensa a la contaminación x como un valor aleatorio X, y se calcula la esperanza de esta función respecto de X se tiene que

- si $X \sim F$, $E_F(T'_X(F)) = E_F\left[g(X) - E_F(g(U))\right] = E_F(g(X)) - E_F(g(U)) = 0$, o sea
$$\text{Si } X \sim F \longrightarrow E_F(T'_X(F)) = 0 \tag{7}$$
en palabras: si la contaminación se da solo en un punto X que se distribuye según F, el valor promedio de esta derivada es nulo.

- si $X \sim G$, $E_G(T'_X(F)) = E_G\left[g(X) - E_F(g(U))\right] = E_G\left[g(X)\right) - E_F(g(U)] = E_{G-F}(g(U)) = T'_G(F)$, o sea
$$\text{Si } X \sim G \longrightarrow E_G(T'_X(F)) = T'_G(F) \tag{8}$$
en palabras: si la contaminación se da solo en un punto X que se distribuye según G, el valor promedio de esta derivada coincide con $T'_G(F)$.

El siguiente corolario que se deduce de (8) es muy útil, ya que simplifica mucho los cálculos para obtener $T'_G(F)$

Corollary 5 *Para hallar $T'_G(F)$, bastará con calcular primero $T'_x(F)$, y luego*
$$T'_G(F) = E_G(T'_X(F)) = \int_{-\infty}^{+\infty} T'_x(F)dG \tag{9}$$

Y el siguiente se utilizará al estudiar la distribución asintótica de $T(F_n)$

Corollary 6 *Si se toma $G = F_n = \frac{1}{n}\sum_{i=1}^{n}\Delta_{X_i}$, como $F_n - F = \frac{1}{n}\sum_{i=1}^{n}(\Delta_{X_i} - F)$ y utilizando la linealidad del funcional*
$$\begin{array}{l} T'_{F_n}(F) = E_{F_n - F}(g(U)) = \frac{1}{n}\sum_{i=1}^{n} E_{\Delta_{X_i} - F}(g(U)) \\ = \frac{1}{n}\sum_{i=1}^{n}\left[g(X_i) - E_F(g(U))\right] = \frac{1}{n}\sum_{i=1}^{n} T'_{X_i}(F) \end{array} \tag{10}$$

es decir $T'_{F_n}(F)$, la derivada de **Fréchet** en F en la dirección de la distribución empírica, es el promedio muestral de las derivadas de **Fréchet** en F en la dirección de cada observación $(T'_{X_i}(F))$.

La derivada de Frechet impone condiciones más restrictivas a un funcional, que no siempre las satisfacen los funcionales robustos. Por eso, en la definición de la función de influencia, de usa la derivada de Gâteau. Sin embargo cuando un funcional es diferenciable Frechet, fácilmente se deriva la distribución asintótica de $T(F_n)$, ya que en este caso interesa que la derivada exista en un entorno, y no solamente en una recta.

Example 7 *Considérese el modelo paramétrico de todas las distribuciones exponenciales*
$$P_\theta = \left\{F_\theta = 1 - e^{-x/\theta} : \theta \in \mathbb{R}^+\right\} \subset \mathcal{F}$$
Sea el funcional $T(F) = E_F(X) = \int_0^\infty x dF$. Si $F \in P_\theta$, como $\int_0^\infty x dF_\theta = \int_0^\infty x\frac{1}{\theta}e^{-x/\theta}dx = \theta$, resulta $T(F_\theta) = \theta$, luego es Fisher consistente. Suponiendo

que este funcional es diferenciable Frechet, se calculará su derivada en F_θ, en la dirección de $G = \Delta_x$, o sea suponiendo una contaminación en x fijo. Según la definición

$$T'_x(F_\theta) = \lim_{t\to 0} \frac{T((1-t)F_\theta + t\Delta_x) - T(F_\theta)}{t}$$

pero

$$\begin{aligned} T((1-t)F_\theta + t\Delta_x) &= \int_0^\infty x d((1-t)F_\theta + t\Delta_x) = (1-t)\int_0^\infty x dF_\theta + t\int_0^\infty x d\Delta_x \\ &= (1-t)\theta + tx \end{aligned}$$

luego

$$T'_x(F_\theta) = \lim_{t\to 0} \frac{(1-t)\theta + tx - \theta}{t} = x - \theta \tag{11}$$

Ahora se mostrará la utilidad de esta derivada. Si $F = F_\theta$, el funcional $T(F_\theta) = \theta$, el valor correcto. Sin embargo si F no pertenece al modelo paramétrico P_θ, sinó que admite una pequeña proporción de contaminación, por ejemplo 0.03 en el punto x, será $F_{0.03} = (1-0.03)F_\theta + 0.03\Delta_x$. Entonces, si se usa el mismo funcional T para averiguar θ, resultará $T(F_{0.03}) \neq \theta$, pero cuanto variará? Aumentará o disminuirá? Según el resultado obtenido

$$delta = T(F_{0.03}) - T(F_\theta) \simeq T'_x(F_\theta) * 0.03 = (x-\theta)0.03$$

luego, si el valor que contamina, coincide con la media de la exponencial, o sea $x = \theta$, entonces delta $\simeq 0$, y no habrá variación. Por otro lado, si $x > \theta$, será delta $\gtrsim 0$ lo que indica que en presencia de contaminación por arriba de la media de la exponencial, el funcional dará estimaciones por exceso. Y por defecto cuando $x < \theta$.

1.7 Desarrollo en serie

Sean $F, G \in \mathcal{F}$, un funcional $T : \mathcal{F} \longrightarrow \mathbb{R}$, y la distribución de la contaminación de F por G, $F_t = (1-t)F + tG$ para $t \in [0;1]$. Ahora interesa calcular $T(F_t)$ mediante un desarrollo en serie de Taylor en $t_0 = 0^+$

$$T(F_t) = T(F) + \sum_{k=1}^{n-1} \frac{T_G^{(k)}(F)}{k!} t^k + \frac{1}{n!} \frac{\partial^n}{\partial t^n} T((1-t)F + tG)\Big|_{t=v} t^n \quad \text{para } v \in [0:t]$$

y si enteresa $T(F_1) = T(G)$ resultará

$$T(G) = T(F) + \sum_{k=1}^{n-1} \frac{T_G^{(k)}(F)}{k!} t^k + \frac{1}{n!} \frac{\partial^n}{\partial t^n} T((1-t)F + tG)\Big|_{t=v} t^n \quad \text{para } v \in [0:1]$$

Si G difiere mucho de F, se necesitarán varios términos en este desarrollo.

Sin embargo si T es diferenciable de Frechet, $\forall G \in \{G \in \mathcal{F} : d(G;F) < \epsilon\}$, se podrá aproximar solo con la primera derivada, o sea

$$T(G) - T(F) = T'_G(F) + o(d(G,F)) \tag{12}$$

Notar que la diferenciacion de Frechet es necesaria pues se requiere que la aproximación sea válida para todo G en un entorno de F, y no solo en una dirección.

1.8 Distribución asintótica de $T(F_n)$

Suponiendo una muestra $X_1, X_2, \cdots, X_n \overset{\mathbf{iid}}{\sim} F$, con distribución empírica muestral F_n, y utilizando resultados del estadístico de bondad de ajuste de Kolmogorov-Smirnov, se demuestra que para $n \longrightarrow \infty$

$$d(F_n, F) = O_p(n^{-\frac{1}{2}}) \tag{13}$$

en principio este resultado es válido para la distancia de Kolmogorov: $d_K(F,G) = \sup_{x\in\mathbb{R}} |F(x) - G(x)|$, pero también para otras distancias. Luego, suponiendo que T es diferenciable de Frechet, y tomando en (12) $G = F_n$, se tiene

$$T(F_n) - T(F) = T'_{F_n}(F) + o(d(F_n, F))$$

y reemplazando(13) y la derivada de Frechet en la dirección de F_n de (10)

$$T(F_n) - T(F) = \frac{1}{n}\sum_{i=1}^{n} T'_{X_i}(F) + o_p(n^{-\frac{1}{2}})$$

o sea

$$\sqrt{n}\,[T(F_n) - T(F)] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} T'_{X_i}(F) + o_p(1)$$

además de (7) $E_F(T'_{X_i}(F)) = 0$, y suponiendo que $Var_F(T'_{X_i}(F)) = E_F(T'^2_{X_i}(F)) < \infty$, mediante el teorema central del límite se llega a

$$\sqrt{n}\,[T(F_n) - T(F)] \xrightarrow{\mathbf{d}} N(0; E_F(T'^2_X(F))) \tag{14}$$

2 Caracterización de la Robustez

2.1 Función de influencia

Sea $F \in \mathcal{F} \supset P_\theta$, y un funcional $T : \mathcal{F} \longrightarrow \mathbb{R}$. Considérese la distribución $F_t = (1-t)F + t\Delta_x$ correspondiente a la contaminación de F por una observacion solo en el punto x. Se define la función de influencia a la derivada de **Gâteau** de T en F, en la dirección de Δ_x

$$IF(x, T, F) = T'_x(F) = \lim_{t\to 0^+} \frac{T\,[(1-t)F + t\Delta_x] - T(F)}{t} = \frac{\partial}{\partial t} T(F_t)\bigg|_{t=0^+} \tag{15}$$

En este caso $IF(x, T, F)$ mide la variación de $T(F)$ debida a una contaminación infinitesimal de F en el punto x, relativa a la cantidad de contaminación. En realidad constituye una colección de derivadas direccionales en la direccion de Δ_x, para distintos x, que usualmente son evaluadas para una $F \in P_\theta$. Y se la usa para estudiar la estabilidad infinitesimal del valor asintótico de estimador, o sea de $T(F)$.

Propiedades

- si se piensa a X como aleatorio, con $X \sim F$ de (7)

$$E_F(IF(X,T,F)) = 0$$

luego si la contaminación es un valor aleatorio con la misma distribución F, en promedio, su influencia sobre el estimador es nula.

- si lo que interesa es la variación de $T(F)$ debida a una contaminación infinitesimal de F por G, relativa a la cantidad de contaminación, se requiere la derivada de **Gâteau** $T'_G(F)$, que según (9) se obtiene a partir de la función de influencia mediante

$$T'_G(F) = E_G(IF(X,T,F)) = \int_{-\infty}^{+\infty} IF(x,T,F)dG \tag{16}$$

2.1.1 Varianza y distribución asintótica

Si se tiene $X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F$, con distribución empírica muestral F_n, e interesa el estimador $\widehat{\theta}_n = T(F_n)$, y suponiendo que T es diferenciable de Frechet, de (14) resulta

$$\sqrt{n}\,[T(F_n) - T(F)] \xrightarrow{\textbf{d}} N(0; E_F(IF^2(X,T,F)) \tag{17}$$

luego la varianza asintótica es

$$V(T,F) = E_F(IF^2(X,T,F) \tag{18}$$

2.1.2 Eficiencia asintótica relativa

Dado el estimadore $\widehat{\theta}_n = T(F_n)$, y otro $\widehat{\theta}^0_n = T^0(F_n)$, este último considerado el estándar contra el cual se quiere comparar,(que verifican ambos la (17)), y llamando precisión de un estimador a la inversa de su varianza, se define la eficiencia asintótica relativa de $\widehat{\theta}_n$ respecto de $\widehat{\theta}^0_n$ a

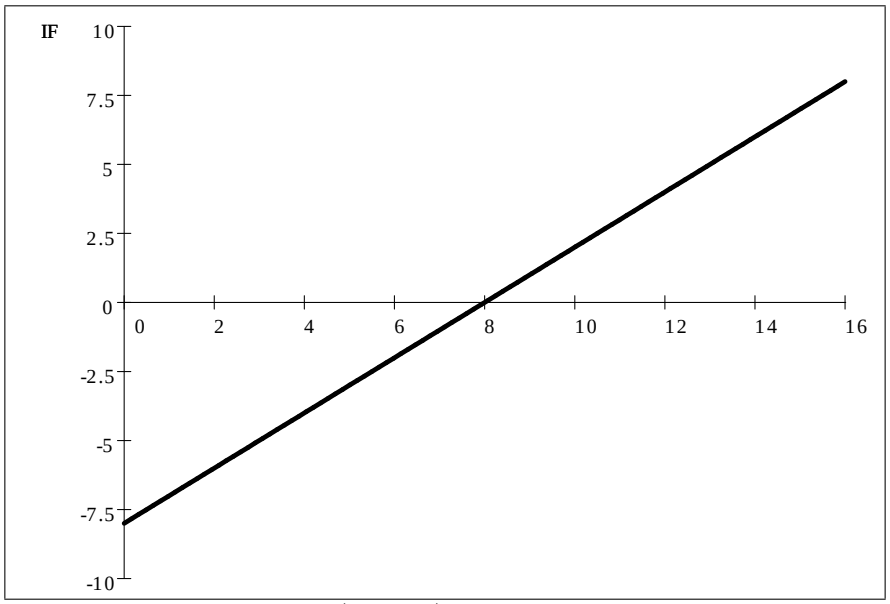
$$ARE_{\widehat{\theta}_n,\widehat{\theta}^0_n} = \frac{V(T,F)^{-1}}{V(T^0,F)^{-1}} = \frac{V(T^0,F)}{V(T,F)} = \frac{E_F(IF^2(X,T^0,F)}{E_F(IF^2(X,T,F)}$$

Por ejemplo si $ARE_{\widehat{\theta}_n,\widehat{\theta}^0_n} = a = 0.8$, quiere decir que la precisión del primer estimador es un 80% de la del estándar. Otra forma de interpretar esto es preguntarse: ¿cuánto mayor debería ser la muestra usada con el primero para que el estimador tenga igual varianza que el segundo? Como $Var(\widehat{\theta}_n) = \frac{V(T,F)}{n}$ y $Var(\widehat{\theta}^0_{n_0}) = \frac{V(T^0,F)}{n_0}$, para que sean iguales deberá ser $n = \frac{V(T,F)}{V(T^0,F)} n_0 = a^{-1} n_0 = 1.25 n_0$. Luego con el primer estimador se requiere una muestra 25% mayor.

Example 8 *Sea la muestra* $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} F$, *con* $F \in \mathcal{F} \supset P_\mu = \left\{N(\mu;\sigma^2) : \mu \in \mathbb{R}\right\}$ *donde en el modelo paramétrico se supondrá que* σ *es conocido. Notar que si* $F \in P_\mu$ *será* $F(x) = \Phi(\frac{x-\mu}{\sigma})$. *Interesa como estimador* $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = E_{F_n}(X)$, *luego el funcional a analizar es* $T(F) = E_F(X)$.

En primer lugar notar que es consistente de Fisher ya que si $F \in P_\mu$, $T(F) = \mu$. Se calculará la función de influencia pero en una $F = F_\mu \in P_\mu$, ya que se quieren estudiar las variaciones en $T(F)$ para pequeñas desviaciones del modelo paramétrico

$$\begin{aligned} IF(x,T,F_\mu) &= T'_x(F_\mu) = \lim_{t\to 0^+} \frac{T\left[(1-t)F_\mu + t\Delta_x\right] - T(F_\mu)}{t} \\ &= \lim_{t\to 0^+} \frac{E_{(1-t)F_\mu + t\Delta_x}(X) - E_{F_\mu}(X)}{t} \\ &= \lim_{t\to 0^+} \frac{(1-t)E_{F_\mu}(X) + tE_{\Delta_x}(X) - E_{F_\mu}(X)}{t} \\ &= \lim_{t\to 0^+} \frac{-t\mu + tx}{t} = x - \mu \end{aligned}$$



$IF(x,T,F_\mu)$ para $\mu = 8$

Además como $V(T,F_\mu) = E_{F_\mu}(X-\mu)^2 = \sigma^2$, resulta $Var(\widehat{\mu}_n) = Var(\overline{X}) = \frac{\sigma^2}{n}$.

Example 9 *Sea la muestra* $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} F$, *con* $F \in \mathcal{F} \supset P_{\sigma^2} = \left\{N(\mu;\sigma^2) : \sigma^2 \in \mathbb{R}^+\right\}$ *donde en el modelo paramétrico se supondrá que* μ *es desconocido pero fijo. Notar que si* $F \in P_{\sigma^2}$ *será* $F(x) = \Phi(\frac{x-\mu}{\sigma})$. *Interesa*

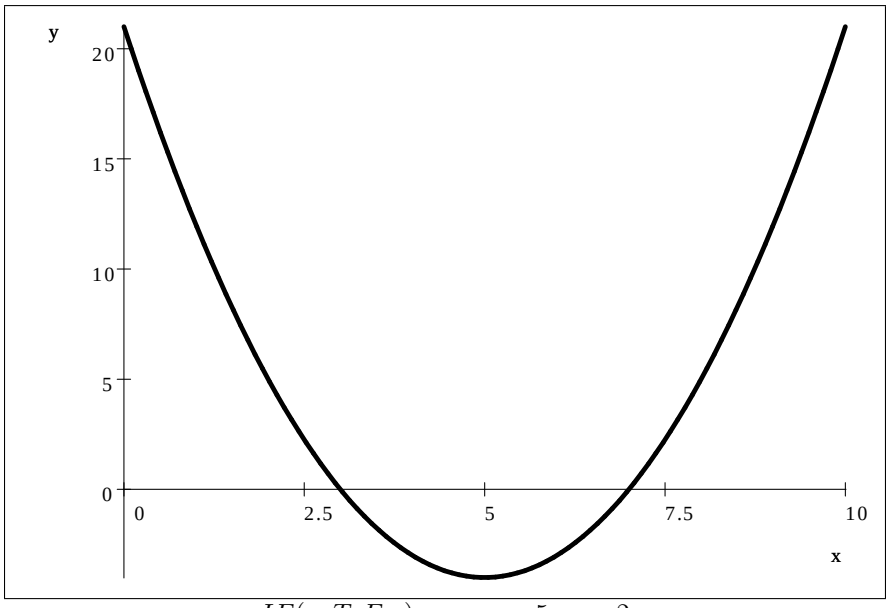
como estimador $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = E_{F_n}(X^2) - E_{F_n}^2(X)$*, luego el funcional a analizar es* $T(F) = E_F(X^2) - E_F^2(X)$.

Este funcional es consistente de Fisher ya que si $F \in P_{\sigma^2}$, $T(F) = \sigma^2$. Se calculará la función de influencia pero en una $F = F_{\sigma^2} \in P_{\sigma^2}$, ya que se quieren estudiar las variaciones en $T(F)$ para pequeñas desviaciones del modelo paramétrico

$$\begin{aligned} IF(x, T, F_{\sigma^2}) &= T'_x(F_{\sigma^2}) = \lim_{t\to 0^+} \frac{T\left[(1-t)F_{\sigma^2} + t\Delta_x\right] - T(F_{\sigma^2})}{t} \\ &= \lim_{t\to 0^+} \frac{\left[E_{(1-t)F_{\sigma^2}+t\Delta_x}(X^2) - E^2_{(1-t)F_{\sigma^2}+t\Delta_x}(X)\right] - \sigma^2}{t} \\ &= \lim_{t\to 0^+} \frac{\left[(1-t)(\mu^2+\sigma^2) + tx^2 - [(1-t)\mu + tx]^2\right] - \sigma^2}{t} \\ &= (x-\mu)^2 - \sigma^2 \end{aligned}$$

O sea, si la proporción de contaminación es ε pequeña se tendrá:

1. si el valor que contamina está cercano a la media ($x \simeq \mu$), entonces asintoticamente el estimador $\widehat{\sigma}_n^2$ tendrá por sesgo $IF(\mu, T, F_{\sigma^2})\varepsilon = -\sigma^2\varepsilon$
2. en cambio si la contaminación esta a 1 desvío de la media ($x \simeq \mu \pm \sigma$), el sesgo asintotico será nulo
3. y por supuesto para $x \ll \mu - \sigma$ o $x \gg \mu + \sigma$, el sesgo asintótico será positivo.



$IF(\mu, T, F_{\sigma^2})$ para $\mu = 5$, $\sigma = 2$

Además como $V(T, F_{\sigma^2}) = E_{F_{\sigma^2}}\left[(X-\mu)^2 - \sigma^2\right]^2 = E_{F_{\sigma^2}}(X-\mu)^4 - \sigma^4 = 2\sigma^4$, resulta $Var(\widehat{\sigma}_n^2) = \frac{2\sigma^4}{n}$.

Por último si el valor que contamina es un valor aleatorio con distribución G, entonces usando(16) el sesgo asintótico será $\varepsilon \int_{-\infty}^{+\infty} IF(x, T, F)dG = \varepsilon \int_{-\infty}^{+\infty} \left[(x-\mu)^2 - \sigma^2\right] dG$.

2.1.3 Curva de sensibilidad

Dado un estimador expresado como funcional de la distribución empírica, $\widehat{\theta}_n = T(F_n)$, y asumiendo que es consistente

$$\widehat{\theta}_n = T(F_n) \xrightarrow{\mathbf{p}} T(F)$$

se desarrolló el concepto de función de influencia analizando pequeñas desviaciones de F en el lado derecho de esta expresión. Ahora se definirá un concepto similar pero trabajando del lado izquierdo.

Dada la muestra $X_1, X_2, \cdots, X_n$, considérese la variación en el estimador al agregar una nueva observación x

$$\widehat{\theta}_{n+1}(X_1, X_2, \cdots, X_n, x) - \widehat{\theta}_n(X_1, X_2, \cdots, X_n)$$

si se divide por la proporción que representa x en la muestra ampliada $(1/(n+1))$, se define la curva de sensibilidad a

$$SC_n(x) = \frac{\widehat{\theta}_{n+1}(X_1, X_2, \cdots, X_n, x) - \widehat{\theta}_n(X_1, X_2, \cdots, X_n)}{1/(n+1)} \tag{19}$$

pero se expresará esto con funcionales. Por de pronto $\widehat{\theta}_n(X_1, X_2, \cdots, X_n) = T(F_n)$, y como

$$F_n = \frac{1}{n}\sum_{i=1}^{n} \Delta_{X_i}$$

para la muestra ampliada se tendrá

$$\begin{aligned} F_{n+1} &= \frac{1}{n+1}(\sum_{i=1}^{n} \Delta_{X_i} + \Delta_x) = \frac{n}{n+1}F_n + \frac{1}{n+1}\Delta_x \\ &= (1 - \frac{1}{n+1})F_n + \frac{1}{n+1}\Delta_x \end{aligned}$$

pudiéndose pensar que F_{n+1} proviene de la contaminación de F_n con Δ_x, en la proporción $t = \frac{1}{n+1}$. Luego (19) se puede expresar como

$$SC_n(x) = \frac{T\left[(1 - \frac{1}{n+1})F_n + \frac{1}{n+1}\Delta_x\right] - T(F_n)}{1/(n+1)}$$

entonces por el aspecto de la parte derecha de esta igualdad se puede pensar que $\lim_{n\to\infty} SC_n(x) = IF(x, T, F)$, y entonces la curva de sensibilidad sería la versión

muestral de la función de influencia. Pero si bién para un x fijo esta convergencia suele cumplirse, la convergencia que importa es la uniforme, ya que interesa la curva para diferentes valores de x. Y la convergencia uniforme no siempre se verifica.

2.2 Sensibilidad a errores groseros(GES)

La función de influencia $IF(x, T, F)$ mide sesgo del valor asintótico del estimador $(T(F))$, debido a una contaminación infinitesimal en el punto x. Obviamente varía con x. Se define la **sensibilidad a errores groseros** del funcional T en F a

$$\gamma^* = \gamma^*(T, F) = \sup_{x \in \Omega} |IF(x, T, F)| \tag{20}$$

donde el supremo se toma en los x en que la IF esté definida. Entonces γ^*mide la peor influencia que una pequeña contaminación puede tener sobre el valor asintótico del estimador. Naturalmente es deseable que γ^* sea finito, en cuyo caso se dirá que T es **B-robusto** en F, donde "B" viene de "bias".

Como se verá muchas veces, los diferentes indicadores para caracterizar la robustez suelen entrar en conflicto. Se darán dos ejemplos:

1. Es deseable que un funcional T sea cualitativamente robusto(**CR**), sin embargo el funcional del R-estimador de posición, si bién es **CR**, su función de influencia no es acotada.

2. Es deseable que un funcional T tenga $IF(x, T, F)$ acotada, y que su varianza asintótica $V(T, F)$ (ver 18) sea mínima. Sin embargo estas dos exigencias no se pueden lograr simultaneamente.

2.3 Sensibilidad a cambios locales

Ahora se quiere medir el efecto de pequeñas fluctuaciones de las observaciones, sobre el valor asintótico del estimador. Como al desplazar una observación desde x a un valor cercano y, el efecto sobre el estimador es proporcional a $IF(y, T, F) - IF(x, T, F)$, si se lo estandariza por $|y - x|$, se obtiene aproximadamente la pendiente de la función de influencia en el punto. Se define la **sensibilidad a cambios locales**, a

$$\lambda^* = \sup_{y \neq x} \frac{|IF(y, T, F) - IF(x, T, F)|}{|y - x|} \tag{21}$$

que es una medida del peor efecto que producen pequeños desplazamientos.

Example 10 *Cuando interesa como estimador la media muestral* $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = E_{F_n}(X)$*, cuyo funcional es* $T(F) = E_F(X)$*, se obtuvo en el ejemplo-8 que* $IF(x, T, F_\mu) = x - \mu$*. Luego* $\gamma^* = \infty$ *(es decir la media muestral es muy sensible a errores groseros), y* $\lambda^* = 1$ *(sin embargo es poco sensible a cambios locales).*

Example 11 *Cuando interesa la varianza muestral* $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = E_{F_n}(X^2) - E_{F_n}^2(X)$*, cuyo funcional es* $T(F) = E_F(X^2) - E_F^2(X)$*, se obtuvo en el ejemplo-9 que* $IF(x, T, F_{\sigma^2}) = (x-\mu)^2 - \sigma^2$*. Luego* $\gamma^* = \infty$*, y además*

$$\lambda^* = \sup_{y \neq x} \frac{\left|(y-\mu)^2 - (x-\mu)^2\right|}{|y-x|} = \sup_{y \neq x} |y + x - 2\mu| = \infty$$

luego la varianza muestral no es robusta tanto para grandes como para pequeños cambios.

2.4 Punto de rechazo

Una idea siempre presente en robustez es la de descartar outliers muy grandes, de manera que no tengan ningún efecto sobre el estimador. Pensando en términos de la función de influencia, si esta se vale cero fuera de cierta area D, cualquier observación fuera de D no tendrá ninguna influencia sobre el estimador. Se define entonces el punto de rechazo

$$\rho^* = \inf\{r > 0 : IF(x, T, F) = 0 \ \text{ para } \ |x| < r\}$$

y si no existe tal r, se define $\rho^* = \infty$. Entonces toda observación x, que verifique $|x| > \rho^*$ será totalmente rechazada.

Lo deseable es que un estimador tenga punto de ruptura ρ^* finito.

2.5 Entorno de contaminación

Sea la muestra $X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F$, con $F \in \mathcal{F} \supset P_\theta$, y un estimador $\widehat{\theta}_n = \widehat{\theta}_n(X_1, X_2, \cdots, X_n)$. En los casos de interés práctico, para $n \to \infty$ se tendrá que

$$\widehat{\theta}_n \xrightarrow{\mathbf{P}} \widehat{\theta}_\infty(F) \tag{22}$$

donde $\widehat{\theta}_\infty(F)$ es el valor asintótico del estimador, que en general dependerá de F. También se supondrá consistencia de Fisher, ya que esto garantiza que si $F \in P_\theta$, entonces asíntoticamente el estimador estima $\widehat{\theta}_\infty(F_\theta) = \theta$.

¿Pero que pasa si la muestra proviene de una $F \notin P_\theta$? Asintóticamente según (22) el estimador converge a $\widehat{\theta}_\infty(F)$, pero no se puede desear ni exigir que $\widehat{\theta}_\infty(F) \cong \theta$, ya que esta F, al no pertenecer a P_θ, y pertenecer en cambio a una familia muy amplia $\mathcal{F}$, no tiene porqué tener "algo que ver" con θ.

Por ello se definirá una tercera familia de distribuciones, más cercana a F_θ, que se llamará entorno de contaminación de F_θ definida por

$$\mathcal{F}_{\theta\varepsilon}^d = \{F \in \mathcal{F} : d(F; F_\theta) < \varepsilon\} \tag{23}$$

notar que $F_\theta \subset \mathcal{F}_{\theta\varepsilon}^d \subset \mathcal{F}$, y si la contaminación ε es pequeña, en $\mathcal{F}_{\theta\varepsilon}^d$ sí podremos desear o exigir que $\widehat{\theta}_\infty(F) \cong \theta$.

Remark 12 *La idea es incluir en esta familia todas las distribuciones que desde el punto de vista robusto queremos considerar "parecidas" a F_θ. Por eso es importante la métrica que se utilice para medir la distancia entre dos distribuciones. Por ejemplo, idealmente desearíamos que la muestra provenga de una $F_\theta \in P_\theta$, pero si las observaciones han sido* ***redondeadas o agrupadas levemente****, la distribución real será F, y no F_θ. Sin embargo en este caso querremos que $d(F; F_\theta)$ sea pequeña. También debemos estar preparados para la situación en que una pequeña proporción de la muestra presente* ***errores groseros u observaciones aberrantes****, y en este caso la distribución real también será F, y no F_θ, pero querremos también que $d(F; F_\theta)$ sea pequeña. Distancias que cumplen estos dos requisitos son las de Prokhorov(Nobel de física 1964) y la de Levy, que se definen a continuación.*

- Distancia de Prokhorov

$$d_p(F, G) = \inf \{\varepsilon > 0 : P_F(A) \leq P_G(A^\varepsilon) + \varepsilon, \ \forall \ \ suceso \ \ A \neq \emptyset\}$$

  donde A^ε es el conjunto de puntos cuya distancia a A es menor que ε. Notar que el primero de los requisitos en negrita, se refleja en A^ε, y el segundo en el "$\cdot\cdot + \varepsilon$".

- Distancia de Levy

$$d_L(F, G) = \inf \{\varepsilon > 0 : F(x - \varepsilon) - \varepsilon \leq G(x) \leq F(x + \varepsilon) + \varepsilon, \forall x \in \mathbb{R}\}$$

2.5.1 Otro entorno de contaminación

En problemas de robustez es más común definir el entorno de contaminación como

$$\mathcal{F}_{\theta\varepsilon} = \{F \in \mathcal{F} : F = (1 - t)F_\theta + tG \text{ para } t \leq \varepsilon, \ \text{ con } G \in \mathcal{F} \text{ arbitraria}\}$$

pero, con una notación menos rigurosa, lo escribiremos así

$$\mathcal{F}_{\theta\varepsilon} = \{F \in \mathcal{F} : F = (1 - \varepsilon)F_\theta + \varepsilon G \ \ \text{con } G \in \mathcal{F} \text{ arbitraria}\} \tag{24}$$

donde a veces se exige que G sea simétrica o una Δ_x(este último caso se usó al definir la función de influencia).

Si bién esta familia no esta definida en base a una distancia, es más fácil de trabajar matemáticamente.

2.6 Punto de ruptura asintótico

La idea es la siguiente: si ε es pequeño, las $F \in \mathcal{F}^d_{\theta\varepsilon}$ serán muy "parecidas" a F_θ, y entonces el valor asintótico del estimador $\widehat{\theta}_n$ se parecerá mucho a $\widehat{\theta}_\infty(F) \simeq \theta$. Sin embargo si se aumenta ε, $\mathcal{F}^d_{\theta\varepsilon}$ será más amplio también, y llegará un momento en que contendrá distribuciones F muy diferentes a F_θ, para las cuales $\widehat{\theta}_n$ no tiene sentido como estimador de θ. El punto de ruptura es la proporción

de contaminación ε^* en que sobreviene esta situación. Pero se formalizará mas esta idea.

Si Θ es el espacio de parámetros, la definición exige que debe existir un compacto $K_\varepsilon \subset \Theta$, que será un conjunto acotado. Y que si $\partial\Theta$ es la frontera de Θ, se supondrá también que $K_\varepsilon \cap \partial\Theta = \emptyset$.

Sea $\widehat{\theta}_n$ una sucesión de estimadores y $\widehat{\theta}_\infty(F)$ su valor asintótico(que depende en general de F), se define el punto de ruptura en F_θ a

$$\varepsilon^* = \varepsilon^*(\widehat{\theta}_n, F_\theta) = \sup_{\varepsilon\in(0,1)} \left\{\varepsilon : \exists K_\varepsilon \text{ con } \forall F \in \mathcal{F}^d_{\theta\varepsilon} \Longrightarrow \widehat{\theta}_\infty(F) \in K_\varepsilon\right\} \qquad (25)$$

o sea, para $\varepsilon < \varepsilon^*$, y para todas las $F \in \mathcal{F}^d_{\theta\varepsilon}$, el valor del estimador se encontrará en el conjunto acotado K_ε. En el caso de usar la versión-24 del entorno de contaminación $\mathcal{F}_{\theta\varepsilon}$, se tendrá la definición

$$\varepsilon^* = \varepsilon^*(\widehat{\theta}_n, F_\theta) = \sup_{\varepsilon\in(0,1)} \left\{\varepsilon : \exists K_\varepsilon \text{ con} \forall G \Longrightarrow \widehat{\theta}_\infty((1-\varepsilon)F_\theta + \varepsilon G) \in K_\varepsilon\right\} \quad (26)$$

¿Pero quién es K_ε? Para la media, se suele pedir que el estimador analizado se encuentre por ejemplo en $K_\varepsilon = [a_\varepsilon, b_\varepsilon]$ con $a_\varepsilon > -\infty$ y $b_\varepsilon < \infty$. Sin embargo para la varianza, como $\Theta = [0, \infty]$, el estimador deberá estar por ejemplo en $K_\varepsilon = [a_\varepsilon, b_\varepsilon]$ donde $a_\varepsilon > 0$ y $b_\varepsilon < \infty$, ya que $K_\varepsilon \cap \partial\Theta = \emptyset$. De esta manera se quiere evitar que el estimador de varianza, asintóticamente de ∞ o cero.

Aunque en principio $\varepsilon^* = \varepsilon^*(\widehat{\theta}_n, F_\theta)$, usualmente no depende de F_θ. Y también si el entorno de contaminación $\mathcal{F}^d_{\theta\varepsilon}$ se define usando otras distancias (Prokhorov, Levy, etc), ε^* suele ser el mismo.

Además, en el caso de un estimador definido por una ecuación que tiene varias soluciones, se pedirá que todas ellas pertenezcan a K_ε.

En el caso que el estimador se exprese como $T(F_n)$ con T consistente, las definiciónes anterióres serán

$$\varepsilon^* = \varepsilon^*(T, F_\theta) = \sup_{\varepsilon\in(0,1)} \left\{\varepsilon : \exists K_\varepsilon \text{ con } \forall F \in \mathcal{F}^d_{\theta\varepsilon} \Longrightarrow T(F) \in K_\varepsilon\right\} \qquad (27)$$

y

$$\varepsilon^* = \varepsilon^*(T, F_\theta) = \sup_{\varepsilon\in(0,1)} \left\{\varepsilon : \exists K_\varepsilon \text{ con} \forall G \Longrightarrow T((1-\varepsilon)F_\theta + \varepsilon G) \in K_\varepsilon\right\} \quad (28)$$

Example 13 *En el caso de la media, el funcional es $T(F) = E_F(X)$. Entonces si se usa la (28), se tiene que si $\varepsilon \neq 0$, no importa cual sea el compacto K_ε, que siempre existirá una G en que*

$$T((1-\varepsilon)F_\mu + \varepsilon G) = (1-\varepsilon)\mu + \varepsilon E(G) \notin K_\varepsilon$$

Luego, el punto de ruptura asintótico es $\varepsilon^* = 0$.

2.7 Punto de ruptura para muestras finitas

Además de la versión asintótica, suele ser útil la definición de punto de ruptura para muestras finitas, ya que proporciona un concepto más simple y no contiene distribuciones de probabilidad.

Si $\mathbf{x} = \{x_1, x_2, \cdot\cdot, x_n\}$ es una muestra fija, y $\widehat{\theta}_n = \widehat{\theta}_n(\mathbf{x})$ un estimador,.se define

$$\varepsilon_n^* = \varepsilon_n^*(\widehat{\theta}_n, \mathbf{x}) = \frac{1}{n} \underset{m \geq 0}{máx} \left\{ m : \exists K_m \text{ tq. cambiando } x_{i_1}, \cdot\cdot, x_{i_m} \text{ arbitrariamente, } \widehat{\theta}_n(\mathbf{x}) \in K_m \right\} \tag{29}$$

y se interpreta como la mayor proporción de observaciones de la muestra que pueden ser arbitrariamente modificados, manteniendo el estimador en K_m.

En la mayoría de los casos ε_n^* no depende de $\mathbf{x}$, solo de n, y tiende al punto de ruptura asintótico cuando $n \to \infty$.

Example 14 *Dada la muestra* $\mathbf{x} = \{x_1, x_2, \cdot\cdot, x_n\}$*, para la media muestral* $\overline{x}_n$*, el punto de ruptura* $\varepsilon_n^*(\overline{x}_n, \mathbf{x}) = 0$*, ya que modificando arbitrariamente solo 1 observación,* $\overline{x}_n$ *puede llevarse fuera de todo límite. Además notar que* $\lim_{n\to\infty} \varepsilon_n^*(\overline{x}_n, \mathbf{x}) = 0$ *que coincide con el punto de ruptura asintótico.*

Example 15 *Dada la muestra* $\mathbf{x}_n = \{x_1, x_2, \cdot\cdot, x_n\}$*, para la mediana principal* $\widetilde{\mathbf{x}}_n$*, si* n *es impar resulta* $\varepsilon_n^*(\widetilde{\mathbf{x}}_n, \mathbf{x}) = \frac{n-1}{2n}$*, y si* n *es par* $\varepsilon_n^*(\widetilde{\mathbf{x}}_n, \mathbf{x}) = \frac{n-2}{2n}$*. Unificando queda* $\forall n \;\; \varepsilon_n^*(\widetilde{\mathbf{x}}_n, \mathbf{x}) = \frac{1}{n}\left[\frac{n-1}{2}\right]$*. Y también resulta que* $\lim_{n\to\infty} \varepsilon_n^*(\widetilde{\mathbf{x}}_n, \mathbf{x}) = \frac{1}{2}$ *coincidente con el punto de ruptura asintótico de la mediana.*

2.8 Sesgo asintótico máximo

Notar que la función de influencia IF y el punto de ruptura **BP**=ε^* consideran dos situaciones extremas en cuanto a la cantidad de contaminación ε. La primera se define para una contaminación infinitesimal($\varepsilon \to 0^+$), y entonces analiza el comportamiento asintótico del estimador para distribuciones que están en un entorno $\mathcal{F}_{\theta\varepsilon}$, que es solo infinitesimalmente mas amplio que F_θ. En cambio el **BP** considera la mayor contaminación que un estimador puede tolerar, resultando en general entornos $\mathcal{F}_{\theta\varepsilon}$ muy grandes. La intención ahora es estudiar el comportamiento del estimador(el peor), para diferentes contaminaciones $\varepsilon < \varepsilon^*$.

Nuevamente se considerará el entorno

$$\mathcal{F}_{\theta\varepsilon} = \{F \in \mathcal{F} : F = (1-\varepsilon)F_\theta + \varepsilon G \;\; \text{con } G \in \mathcal{F} \text{ arbitraria}\}$$

Si se fija un ε queda definido un $\mathcal{F}_{\theta\varepsilon}$, y entonces para cada $F \in \mathcal{F}_{\theta\varepsilon}$ el sesgo asintótico del estimador será $b_{\widehat{\theta}}(F, \theta) = \widehat{\theta}_\infty(F) - \theta$. Luego se define el sesgo asintótico máximo a

$$\mathbf{MB}_{\widehat{\theta}}(\varepsilon, \theta) = máx\left\{\left|b_{\widehat{\theta}}(F, \theta)\right| \;\text{ con } F \in \mathcal{F}_{\theta\varepsilon}\right\} \tag{30}$$

que es función de la cantidad de contaminación ε, para $\varepsilon < \varepsilon^*$. Notar de paso que cuando $\varepsilon = \varepsilon^*$, resultará $\mathbf{MB}_{\widehat{\theta}}(\varepsilon^*, \theta) = \infty$.

En el caso que $\Theta = \mathbb{R}$, existe una relación entre **BP** y $\mathbf{MB}(\varepsilon)$

$$\varepsilon^*(\widehat{\theta}_n, F_\theta) = máx\left\{\varepsilon \geq 0 : MB_{\widehat{\theta}}(\varepsilon, \theta) < \infty\right\}$$

notar que aunque dos estimadores tengan el mismo **BP**, usualmente serán diferentes sus curvas de $\mathbf{MB}(\varepsilon)$.

2.8.1 Sensibilidad a la contaminación

Se define como

$$\gamma_c(\theta) = \frac{d}{d\varepsilon}\mathbf{MB}_{\widehat{\theta}}(\varepsilon, \theta)\Big|_{\varepsilon=0} \tag{31}$$

y mide la pendiente del **MB** para una pequeñisima contaminación. Además, si se desarrolla en serie

$$\mathbf{MB}_{\widehat{\theta}}(\varepsilon, \theta) \approx \mathbf{MB}_{\widehat{\theta}}(\varepsilon, \theta)\big|_{\varepsilon=0} + \frac{d}{d\varepsilon}\mathbf{MB}_{\widehat{\theta}}(\varepsilon, \theta)\Big|_{\varepsilon=0} \varepsilon$$

y como para un ε muy pequeño resulta $\mathcal{F}_{\theta\varepsilon} \simeq F_\theta$, y en el caso de un estimador consistente, $\mathbf{MB}_{\widehat{\theta}}(0, \theta) \simeq 0$, entonces γ_c da una estimación de **MB** para un ε pequeño

$$\mathbf{MB}_{\widehat{\theta}}(\varepsilon, \theta) \approx \varepsilon\gamma_c(\theta) \tag{32}$$

relación que se usa aveces, cuando $\mathbf{MB}(\varepsilon)$ es difícil de obtener, y la contaminación es pequeña.

2.8.2 Relación con la sensibilidad a errores groseros

Según la definición de sensiblidad a errores groseros(20)

$$\gamma^* = \gamma^*(T, F_\theta) = \sup_{x\in\Omega} |IF(x, T, F_\theta)|$$

si se toma el entorno $\mathcal{F}_{\theta\varepsilon} = \{F = (1-\varepsilon)F_\theta + \varepsilon\Delta_x \ \ \forall x\}$, y suponiendo $\widehat{\theta}_n$ consistente, entonces $\forall x$ el sesgo asintótico verificará

$$\left|\widehat{\theta}_\infty((1-\varepsilon)F_\theta + \varepsilon\Delta_x) - \widehat{\theta}_\infty(F_\theta)\right| \leq \mathbf{MB}_{\widehat{\theta}}(\varepsilon, \theta)$$

dividiendo por ε, tomando el límite y usando (32) resulta

$$\gamma^* \leq \gamma_c(\theta) \tag{33}$$

la igualdad vale para M-estimadores con ψ acotada, pero no en general.

3 Estimadores Robustos Clásicos

3.1 Modelo de posición

Se supondrá una muestra donde cada observación X_i depende del "verdadero valor" del parámetro desconocido θ, más un error aleatorio U_i , o sea

$$X_i = \theta + U_i \quad (i = 1, \cdots, n)$$

usualmente se asume también que:

1. $U_1, U_2, \cdots, U_n \overset{\textbf{iid}}{\sim} F_0$ (conocida)
2. F_0 es una distribución simétrica

De 1 surge que cada X_i se distribuye según una $F(x) = F_0(x - \theta)$; y el punto 2 se impone para evitar el error sistemático. Con la notación del capítulo anterior se tiene entonces el modelo paramétrico

$$X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F \quad \text{con} \quad F \in P_\theta = \{F_\theta : F_\theta(x) = F_0(x - \theta)\} \tag{34}$$

y con el se busca un estimador $\widehat{\theta}_n$ del parámetro de **posición** θ. Pero tratandose de un parámetro de posición sería deseable que este estimador satisfaga, los siguientes dos requisitos:

- Equivarianza de traslación: $\forall X_i$, $\forall c$

$$\widehat{\theta}_n(X_1 + c, X_2 + c, \cdots, X_n + c) = \widehat{\theta}_n(X_1, X_2, \cdots, X_n) + c \tag{35}$$

- Equivarianza de escala: $\forall X_i$,$\forall k$,

$$\widehat{\theta}_n(kX_1, kX_2, \cdots, kX_n) = k\widehat{\theta}_n(X_1, X_2, \cdots, X_n) \tag{36}$$

Con el primero nos aseguramos que si los datos de la muestra están todos desplazados c unidades, el valor que proporcione el estimador estará también igualmente desplazado; y con el segundo el estimador se ajustará automaticamente cuando cambien las unidades de la muestra, por ejemplo cuando las X_i se midan en libras en lugar de kilogramos. En el caso de la media muestral $\overline{X}_n$, se cumplen estas dos equivarianzas, sin embargo en estimadores más sofisticados, podrían no cumplirse.

Comparación $\overline{X}$ con $\widetilde{X}$ en P_θ y en $\mathcal{F}_{\theta\varepsilon}$ Ahora bién, si en el modelo paramétrico de posición(34) se supone que $F_0 = N(0; \sigma)$, se tendrá que

$$X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F \quad \text{con} \quad F \in P_\theta = \{F_\theta : F_\theta = N(\theta; \sigma)\} \tag{37}$$

y si se usa el método clásico de máxima verosimilitud, se obtendrá como estimador $\widehat{\theta}_n = \overline{X}_n$, la media muestral. Y este es un estimador óptimo, en el sentido que es **IMVU**(insesgado de mínima varianza uniformemente) con

$$\text{Si } F \in P_\theta : E(\overline{X}_n) = \theta \ \text{ y } \ Var(\overline{X}_n) = \frac{\sigma^2}{n} \tag{38}$$

y además cumple con las dos equivarianzas. Es muy bueno. Pero por supuesto es optimo si tenemos la certeza que $F_0 = N(0;\sigma)$, o sea que $F \in P_\theta$.

Pero, que tal si admitimos que casi siempre(con probabilidad $1-\varepsilon$), F_0 es una $N(0;\sigma)$ como supusimos, pero a veces(con probabilidad ε), F_0 es una $N(0;\tau)$. Esto equivale a que la distribución de la muestra es $F = (1-\varepsilon)N(\theta;\sigma) + \varepsilon N(\theta;\tau)$, luego el modelo de posición(37) quedará ampliado a un entorno de contaminación

$$X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F \ \text{ con } \ F \in \mathcal{F}_{\theta\varepsilon} = \{F : F = (1-\varepsilon)N(\theta;\sigma) + \varepsilon N(\theta;\tau)\} \tag{39}$$

¿Pero, como se comportará nuestro estimador $\widehat{\theta}_n = \overline{X}_n$, que se lo obtuvo suponiendo que $F \in P_\theta$, cuando en realidad $F \in \mathcal{F}_{\theta\varepsilon} \supset P_\theta$? Usando la expresión de la media y varianza de una mezcla en(39) se obtiene

$$\text{Si } F \in \mathcal{F}_{\theta\varepsilon} : E(\overline{X}_n) = \theta \ \text{ y } \ Var(\overline{X}_n) = \frac{(1-\varepsilon)\sigma^2 + \varepsilon\tau^2}{n} = \left[1 + \varepsilon(\frac{\tau^2}{\sigma^2} - 1)\right]\frac{\sigma^2}{n} \tag{40}$$

Para concretar supondremos que la contaminación es baja, $\varepsilon = 0.05$, y que la normal que contamina $N(0;\tau)$ tiene desvío $\tau = 10\sigma$. Comparando (38) y (40) resulta que $\overline{X}_n$ sigue siendo insesgada, pero al admitir una contaminación de solo $\varepsilon = 0.05$, la varianza del estimador pasa de valer $\frac{\sigma^2}{n}$ a $5.95\frac{\sigma^2}{n}$. Se incrementó 5.95 veces! Luego la media muestral se comporta muy bién bajo el modelo paramétrico P_θ, pero cuando hay contaminación, aún baja, su comportamiento en el entorno $\mathcal{F}_{\theta\varepsilon}$ se deteriora mucho. Por eso se dice que $\widehat{\theta}_n = \overline{X}_n$ no es un estimador robusto.

Se analizará ahora como estimador, la mediana muestral $\widehat{\theta}_n = \widetilde{X}_n$, y se estudiará su comportamiento en las dos situaciones anteriores($F \in P_\theta$ y $F \in \mathcal{F}_{\theta\varepsilon}$). Como las expresiones para muestras finitas son complicadas, se utilizará el siguiente resultado asintótico:

Theorem 16 *Si $X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F$, y F tiene única mediana $\widetilde{\mu}$, y densidad f continua en $\widetilde{\mu}$, con $f(\widetilde{\mu}) > 0$, entonces si $\widetilde{X}_n$ es la mediana de la muestra*

$$\sqrt{n}(\widetilde{X}_n - \widetilde{\mu}) \xrightarrow{\textbf{d}} N(0; \frac{1}{2f(\widetilde{\mu})})$$

Corollary 17 *Si $F = N(\mu;\sigma)$ entonces $\sqrt{n}(\widetilde{X}_n - \mu) \xrightarrow{\textbf{d}} N(0; \sqrt{\frac{\pi}{2}}\sigma)$.*

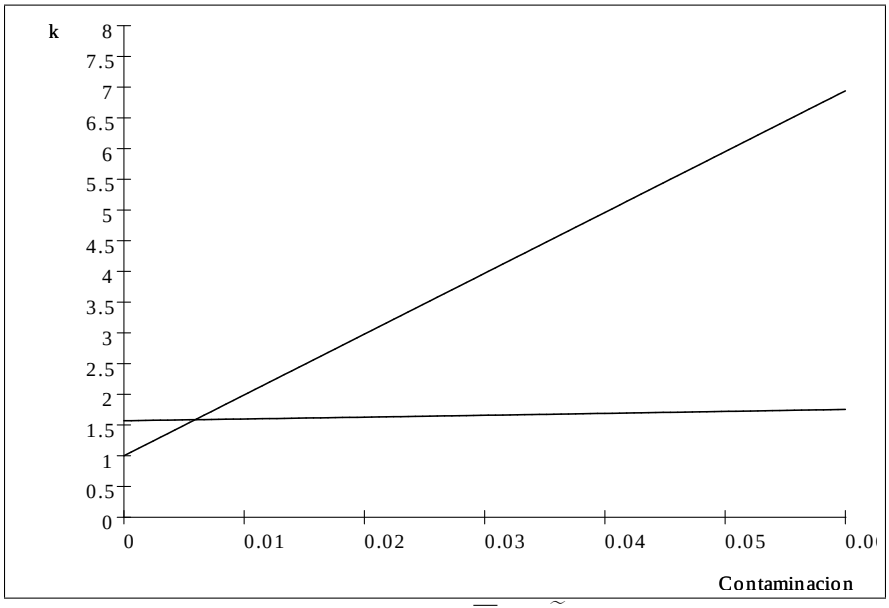
Suponiendo muestras grandes para poder usar el teorema, resulta

$$\text{Si } F \in P_\theta : E(\widetilde{X}_n) = \theta \ \text{ y } \ Var(\widetilde{X}_n) \approx \frac{\pi}{2}\frac{\sigma^2}{n} \tag{41}$$

y si $F \in \mathcal{F}_{\theta\varepsilon}$, $f(x) = (1-\varepsilon)\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\theta)^2}{2\sigma^2}} + \varepsilon\frac{1}{\sqrt{2\pi}\tau}e^{-\frac{(x-\theta)^2}{2\tau^2}}$ y $f(\theta) = \frac{1}{\sqrt{2\pi}\sigma}\left[1+\varepsilon(\frac{\sigma}{\tau}-1)\right]$ luego

$$\text{Si } F \in \mathcal{F}_{\theta\varepsilon} : E(\widetilde{X}_n) = \theta \;\text{ y }\; Var(\widetilde{X}_n) \approx \frac{\pi}{2\left[1+\varepsilon(\frac{\sigma}{\tau}-1)\right]^2}\frac{\sigma^2}{n} \qquad (42)$$

Para comparar, igual que antes supondremos que la contaminación es baja, $\varepsilon = 0.05$, y que la normal que contamina $N(0;\tau)$ tiene desvío $\tau = 10\sigma$. De (41) y (42) resulta que $\widetilde{X}_n$ sigue siendo insesgada, incluso equivariante, y al admitir una contaminación de solo $\varepsilon = 0.05$, la varianza del estimador pasa de valer $1.57\frac{\sigma^2}{n}$ a $1.72\frac{\sigma^2}{n}$. Se incrementó solo 1.095 veces! Luego la mediana muestral bajo el modelo paramétrico P_θ, se comporta un poco peor que la media(su varianza es 1.57 veces mayor), pero cuando hay contaminación, aún baja, su comportamiento en el entorno $\mathcal{F}_{\theta\varepsilon}$ es muy superior al de la media(la varianza de la media es $5.95\frac{\sigma^2}{n}$ v.s.$1.72\frac{\sigma^2}{n}$ para la mediana). Por eso se dice que $\widehat{\theta}_n = \widetilde{X}_n$ es un estimador robusto. A continuación, para $\frac{\tau}{\sigma} = 10$, se grafica la varianza de $\overline{X}_n$ y $\widetilde{X}_n$, en función de la contaminación ε(en realidad se grafica $k(\varepsilon)$, donde $Var() = k(\varepsilon)\frac{\sigma^2}{n}$)



Comparación de varianzas de $\overline{X}_n$ y $\widetilde{X}_n$ en función de ε

La idea subyacente al buscar un estimador robusto es aceptar que si el modelo paramétrico es válido($F \in P_\theta$), "**perderemos algo**" en relación a un estimador óptimo para ese modelo; sin embargo, si hay contaminación, "**ganaremos mucho**" también en relación a dicho estimador.

En este ejemplo se prestó atención al sesgo, la equivarianza y principalmente la varianza, para evaluar el comportamiento de un estimador. Sin embargo en

el capítulo anterior se vieron varios indicadores para caracterizar la robustez, que deben ser tenidos en cuenta al buscar un buen estimador robusto. A continuación se analizarán de acuerdo a estos indicadores varios estimadores clásicos de posición.

3.1.1 Mediana muestral

Suponiendo $P_\theta = \{F_\theta = N(\theta;\sigma)\}$ donde θ es la media(y la mediana también), y funcional $T(F) = F^{-1}(\frac{1}{2})$, resulta $T(F_\theta) = F_\theta^{-1}(\frac{1}{2}) = \theta$, o sea es *consistente de Fisher*. Y también la mediana satisface las dos *equivarianzas de traslación y de escala* vistas en (35) y (36).

- **Función de influencia**: $IF(x,T,F_\theta) = \lim\limits_{\varepsilon\to 0^+} \frac{T((1-\varepsilon)F_\theta+\varepsilon\Delta_x)-\theta}{\varepsilon}$, donde en el numerador habrá que resolver en $\widetilde{u}$: $(1-\varepsilon)F_\theta(\widetilde{u})+\varepsilon\Delta_x(\widetilde{u}) = \frac{1}{2}$. Primero se considerará el caso $x > \theta$, y como entonces debe ser $\widetilde{u} < x$, esto hace $\Delta_x(\widetilde{u}) = 0$, quedando a resolver solo $(1-\varepsilon)F_\theta(\widetilde{u}) = \frac{1}{2}$. Si se desarrolla en serie de Taylor en θ queda

$$(1-\varepsilon)\left[\frac{1}{2} + f_\theta(\theta)(\widetilde{u}-\theta)\right] = \frac{1}{2}$$

despejando sale $\widetilde{u} = \theta + \frac{\varepsilon}{(1-\varepsilon)2f_\theta(\theta)}$. Reemplazando en la IF y tomando límite

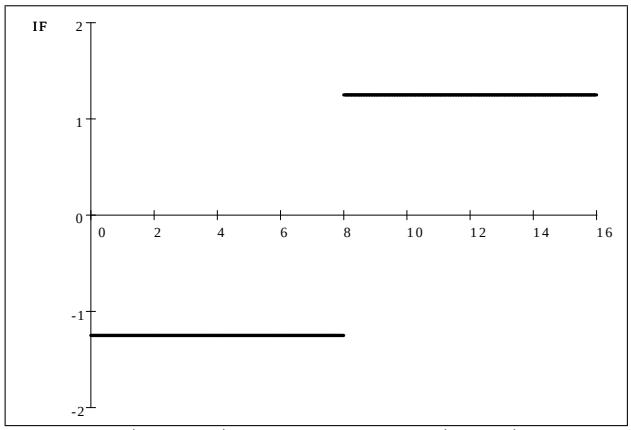
$$\lim_{\varepsilon\to 0^+} \frac{\theta + \frac{\varepsilon}{(1-\varepsilon)2f_\theta(\theta)} - \theta}{\varepsilon} = \frac{1}{2f_\theta(\theta)} = \frac{1}{2\frac{1}{\sqrt{2\pi}\sigma}} = \sqrt{\frac{\pi}{2}}\sigma$$

haciendo un análisis similar para $x < \theta$, resulta el límite $-\sqrt{\frac{\pi}{2}}\sigma$. En definitiva la función de influencia es

$$IF(x,T,F_\theta) = sgn(x-\theta)\sqrt{\frac{\pi}{2}}\sigma = \begin{cases} -\sqrt{\frac{\pi}{2}}\sigma & \text{sí } x < \theta \\ 0 & \text{sí } x = \theta \\ \sqrt{\frac{\pi}{2}}\sigma & \text{sí } x > \theta \end{cases} \tag{43}$$

A continuación se representa IF suponiendo que la mediana $\theta = 8$, y que

$\sigma = 1$



$IF(x, T, F_\theta)$ para la mediana ($\sigma = 1$)

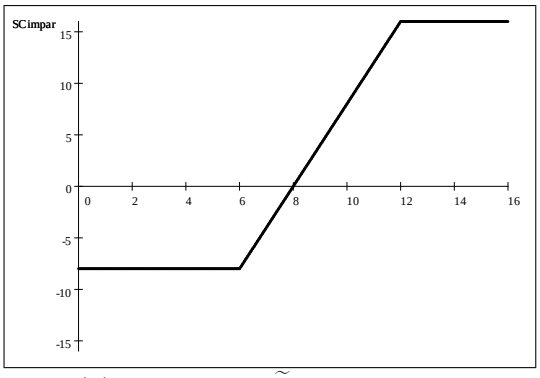
- **Varianza asintótica**: Como $V(T, F_\theta) = E_{F_\theta}(IF^2(X, T, F_\theta) = E_{F_\theta}(\frac{\pi}{2}\sigma^2)$, resulta
$$V(T, F_\theta) = \frac{\pi}{2}\sigma^2$$
que coincide con la del teorema-16.

- **Curva de sensibilidad**: Según la definición(19)
$$SC_n(x) = \frac{medp(X_1, X_2, \cdots, X_n, x) - medp(X_1, X_2, \cdots, X_n)}{1/(n+1)}$$
Si se considera la muestra ordenada, y llamando $k = \left[\frac{n+1}{2}\right]$, se analizará primero el caso n impar, ya que entonces $medp(X_1, X_2, \cdots, X_n) = X_{(k)}$. Para analizar el primer término del numerador se representa la muestra
$$X_{(1)} \quad \cdot \quad \cdot \quad \cdot \quad X_{(k-1)} \quad X_{(k)} \quad X_{(k+1)} \quad \cdot \quad \cdot \quad \cdot \quad X_{(n)}$$
y entonces
$$\begin{array}{ll} \text{sí } x \leq X_{(k-1)} & medp(X_1, X_2, \cdots, X_n, x) = \frac{X_{(k-1)} + X_{(k)}}{2} \\ \text{sí } X_{(k-1)} < x < X_{(k+1)} & medp(X_1, X_2, \cdots, X_n, x) = \frac{x + X_{(k)}}{2} \\ \text{sí } x \geq X_{(k+1)} & medp(X_1, X_2, \cdots, X_n, x) = \frac{X_{(k)} + X_{(k+1)}}{2} \end{array}$$
Finalmente la curva de sensiblidad para n impar es
$$SC_n^{imp}(x) = \begin{cases} \frac{X_{(k-1)} - X_{(k)}}{2}(n+1) & \text{para } x \leq X_{(k-1)} \\ \frac{x - X_{(k)}}{2}(n+1) & \text{para } X_{(k-1)} < x < X_{(k+1)} \\ \frac{X_{(k+1)} - X_{(k)}}{2}(n+1) & \text{para } x \geq X_{(k+1)} \end{cases}$$
y con un análisis similar, para n par es

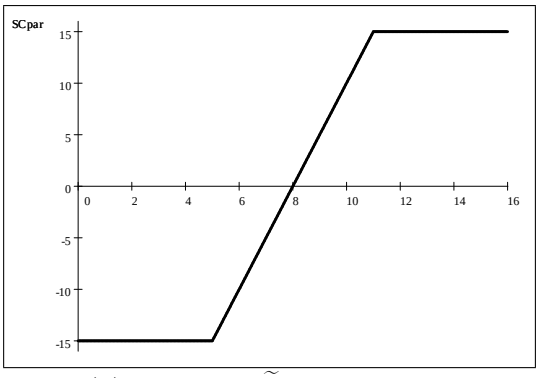
$$SC_n^{par}(x) = \begin{cases} -\frac{X_{(k+1)}-X_{(k)}}{2}(n+1) & \text{para } x \leq X_{(k)} \\ (x - \frac{X_{(k)}+X_{(k+1)}}{2})(n+1) & \text{para } X_{(k)} < x < X_{(k+1)} \\ \frac{X_{(k+1)}-X_{(k)}}{2}(n+1) & \text{para } x \geq X_{(k+1)} \end{cases}$$

A continuación se presentan las gráficas para dos muestras con

$$\begin{aligned} n &= 7: \quad \{1, 2, 6, \mathbf{8}, 12, 14, 15\} \\ n &= 4: \quad \{2, 5, 11, 16\} \end{aligned}$$



$SC_7(x)$:$X_{(k-1)} = 6, \widetilde{X} = 8, X_{(k+1)} = 12$



$SC_4(x) : X_{(k)} = 5, \widetilde{X} = 8, X_{(k+1)} = 11$

donde se observa que siempre $SC_n(\widetilde{X}) = 0$, y para muestra impar, en general los valores máximos positivos y negativos no son opuestos.

Observación: Según Croux(1998) la SC_n de la mediana no converge en probabilidad a la correspondiente IF para ningún x.

- **Sensibilidad a errores groseros**: Según (20) y la(43)

$$\gamma^* = \sup_{x \in \Omega} |IF(x, T, F_\theta)| = \sqrt{\frac{\pi}{2}}\sigma$$

y por lo tanto la mediana es **B-robusto,** a diferencia de la media muestral que tiene una función de influencia no acotada. En otras palabras, si la contaminación es pequeña, el efecto que un outlier muy grande puede tener sobre el sesgo asintótico de la mediana, está acotado. Es más, es fijo, ya que $\sqrt{\frac{\pi}{2}}\sigma$ también es el ínfimo.

- **Sensibilidad a cambios locales**: Según (21) y la(43)

$$\lambda^* = \sup_{y \neq x} \frac{|IF(y, T, F_\theta) - IF(x, T, F_\theta)|}{|y - x|} = \infty$$

y da ∞ ya que la IF tiene un salto en $x = \theta$. Pero notar que λ^* es un indicador global(debido al supremo) de cambios locales(x próximo a y); y

el motivo del ∞ es que si se modifican levemente observaciones cercanas a θ(siempre suponiendo pequeña la contaminación), la variación relativa del sesgo asintótico del estimador será muy importante. Pero si se modifican observaciones alejadas de θ, en realidad el efecto es nulo.

- **Punto de ruptura asintótico**: De (28)

$$\varepsilon^* = \sup_{\varepsilon\in(0,1)} \{\varepsilon : \exists K_\varepsilon \text{ con} \forall G \Longrightarrow medp((1-\varepsilon)F_\theta + \varepsilon G) \in K_\varepsilon\}$$

Se tomará $\varepsilon < 0.5$, y veremos que siempre $\exists K_\varepsilon$. Como $\lim_{u\to\infty} F_\theta(u) = 1$, tomando un δ muy chico, concretamente $\delta < \frac{0.5-\varepsilon}{1-\varepsilon}$, existirá un b_ε tal que $F_\theta(b_\varepsilon) > 1-\delta > \frac{0.5}{1-\varepsilon}$. Ahora notar que $\forall G$

$$(1-\varepsilon)F_\theta(b_\varepsilon) + \varepsilon G(b_\varepsilon) \geq (1-\varepsilon)F_\theta(b_\varepsilon) > (1-\varepsilon)\frac{0.5}{1-\varepsilon} = 0.5$$

luego la mediana deberá cumplir $\widetilde{\theta} \leq b_\varepsilon$. Demostrando en forma similar para el otro lado, se llega a que $\widetilde{\theta} \geq a_\varepsilon$. Luego $K_\varepsilon = [a_\varepsilon, b_\varepsilon]$. Ahora veamos que para $\varepsilon \geq 0.5$, $\nexists K_\varepsilon$. Tómese en particular $G = \Delta_t$ y entonces notar que para $u < t$

$$(1-\varepsilon)F_\theta(u) + \varepsilon\Delta_t(u) < 0.5$$

con solo tomar t suficientemente grande. Luego la mediana no estará en un conjunto acotado. En definitiva, el supremo será $\varepsilon = \varepsilon^* = 0.5$, que es el punto de ruptura asintótico de la mediana.

- **Punto de ruptura para muestras finitas**: Según (29)

$$\varepsilon_n^* = \frac{1}{n}\underset{m\geq 0}{máx}\{m : \exists K_m \text{ tq. cambiando } x_{i_1}, \cdot\cdot, x_{i_m} \text{ arbitrariamente, } medp_n(\mathbf{x}) \in K_m\}$$

y en el ejemplo-15 se llegó a que para la mediana

$$\varepsilon_n^* = \frac{1}{n}\left[\frac{n-1}{2}\right] \quad \text{resultando también } \varepsilon_n^* \underset{n\to\infty}{\longrightarrow} \varepsilon^* = \frac{1}{2}$$

Pero ahora se analizará el sentido de la definición de ε_n^* para, por ejemplo una muestra de tamaño $n = 7$. Dada la muestra, la mediana está en $X_{(4)}$. Si se toma $m = 2$, y se cambian arbitrariamente 2 observaciones, tratando que la mediana **aumente**; (por ejemplo pasando $X_{(1)}$ y $X_{(2)}$ a la derecha de toda la muestra), la nueva mediana como máximo valdrá $X_{(6)}$. Y si ahora se cambian arbitrariamente 2 observaciones, tratando que la mediana **disminuya**; (por ejemplo pasando $X_{(6)}$ y $X_{(7)}$ a la izquierda de toda la muestra), la nueva mediana como mínimo valdrá $X_{(2)}$. Luego con $m = 2$, cambiando arbitrariamente 2 observaciones siempre la mediana estará en $[X_{(2)}; X_{(6)}] = K_2$ conjunto acotado. Si se toma $m = 3$, se obtendrá otro $[X_{(1)}; X_{(7)}] = K_3$, también acotado. Pero para $m = 4$ no se obtiene un K_4 acotado, ya que K_4 depende de las observaciones que se modifican. Luego en este caso $\varepsilon_7^* = \frac{3}{7}$.

- **Sesgo asintótico máximo**: Según (30)

$$\mathbf{MB}_{\widehat{\theta}}(\varepsilon,\theta) = máx\left\{\left|b_{\widehat{\theta}}(F,\theta)\right| \text{ con } F \in \mathcal{F}_{\theta\varepsilon}\right\}$$

se supondrá ε fijo, con $\varepsilon < \varepsilon^* = 0.5$, y entonces para todas las $F \in \{(1-\varepsilon)F_\theta + \varepsilon G \text{ con } G \in \mathcal{F} \text{ arbitraria}\}$ hay que estudiar el máximo $|sesgo|$. Pero notar que el sesgo positivo más grande se dará cuando la $medp((1-\varepsilon)F_\theta + \varepsilon G)$ sea lo mayor posible, y esto se logrará tomando a G bien corrida a la derecha de F_θ. Concretamente se tomará $G = \Delta_t$ para t suficientemente grande. La mediana surge entonces de

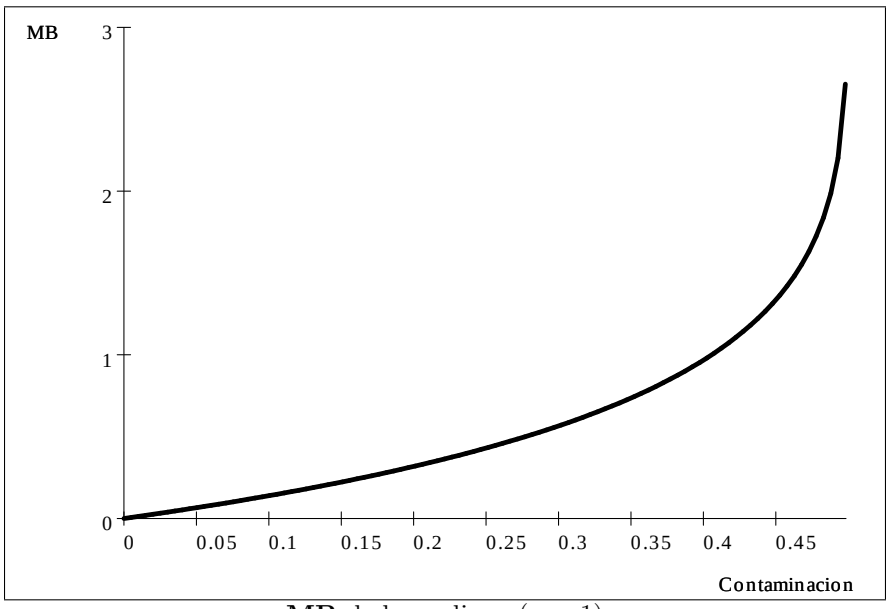
$$(1-\varepsilon)F_\theta(\widetilde{u}) + \varepsilon\Delta_t(\widetilde{u}) = 0.5$$

y como para u grande el primer término es mayor que 0.5, tomando un t más grande aún, $\Delta_t(u) = 0$, y entonces la ecuacion para hallar la mediana máxima es

$$(1-\varepsilon)F_\theta(\widetilde{u}) = 0.5$$

o sea $\theta_{\max} = \theta + \sigma\Phi^{-1}(\frac{1}{2(1-\varepsilon)})$ (con el desarrollo en serie.usado para obtener la función de influencia de la mediana en 3.1.1). Luego el sesgo máximo positivo es $\sigma\Phi^{-1}(\frac{1}{2(1-\varepsilon)})$. Haciendo similar deducción pero tomando $G = \Delta_{-t}$, se llega a que el sesgo máximo negativo es igual pero de diferente signo. En definitiva

$$\mathbf{MB}_{\widetilde{\theta}}(\varepsilon,\theta) = \sigma\Phi^{-1}(\frac{1}{2(1-\varepsilon)}) \tag{44}$$



MB de la mediana ($\sigma = 1$)

- **Sensibilidad a la contaminación**: Según (31) es la pendiente de la $\mathbf{MB}_{\widetilde{\theta}}(\varepsilon, \theta)$ en $\varepsilon = 0$, luego resulta

$$\gamma_c(\theta) = \sqrt{\frac{\pi}{2}}\sigma$$

en particular en este caso $\gamma^* = \gamma_c(\theta)$, o sea vale la acotación($\gamma^* \leq \gamma_c(\theta)$) con la sensibilidad a errores groseros.

3.1.2 Media truncada

Es un estimador del parámetro de posición que consiste en eliminar una proporción de las menores y mayores observaciones de la muestra. Sea $\alpha \in [0; \frac{1}{2})$ y $m = [(n-1)\alpha]$, luego la $\alpha-$media truncada se define mediante

$$\overline{X}_\alpha = \frac{1}{n-2m}\sum_{i=m+1}^{n-m} X_{(i)}$$

es decir se descartan de la muestra los primeros m, y los últimos m estadísticos de orden. Además si una variable aleatoria tiene distribución F, la $\alpha-$media truncada se calcula

$$\theta_\alpha = \frac{1}{1-2\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x dF \tag{45}$$

y luego de un cambio de variable se puede expresar como funcional así

$$\theta_\alpha = T(F) = \frac{1}{1-2\alpha}\int_\alpha^{1-\alpha} F^{-1}(t)dt$$

Suponiendo $P_\theta = \{F_\theta = N(\theta; \sigma)\}$ donde θ es la media(y la $\alpha-$media truncada también), y como $T(F_\theta) = \frac{1}{1-2\alpha}\int_\alpha^{1-\alpha} F_\theta^{-1}(t)dt = \theta_\alpha$, resulta que es *consistente de Fisher*. Y también la $\alpha-$media truncada satisface las dos *equivarianzas de traslación y de escala* vistas en (35) y (36).

Remark 18 *Existe también otro estimador, la "adaptive" media truncada, que es similar, salvo que la proporción α de observaciones que se descartan no es fija, sinó que depende de los datos.*

- **Función de influencia**:Será $IF(x, T, F_\theta) = \lim_{\varepsilon \to 0^+} \frac{\theta_\alpha((1-\varepsilon)F_\theta + \varepsilon\Delta_x) - \theta}{\varepsilon}$, donde para calcular el primer término del numerador en (45) se distinguirán dos casos:

  1. si $F_\theta^{-1}(\alpha) \leq x \leq F_\theta^{-1}(1-\alpha)$, se tendrá

$$\begin{aligned} & \frac{1}{1-2\alpha}\left[\int_{F_\theta^{-1}(\frac{\alpha}{1-\varepsilon})}^{F_\theta^{-1}(\frac{1-\alpha}{1-\varepsilon})} u(1-\varepsilon)f_\theta(u)du + x\varepsilon\right] \\ = & \frac{1-2\alpha-\varepsilon}{1-2\alpha}\int_{F_\theta^{-1}(\frac{\alpha}{1-\varepsilon})}^{F_\theta^{-1}(\frac{1-\alpha}{1-\varepsilon})} \frac{u(1-\varepsilon)f_\theta(u)}{1-2\alpha-\varepsilon}du + \frac{x\varepsilon}{1-2\alpha} \end{aligned}$$

pero la integral vale θ(por ser los límites simétricos respecto de θ), luego queda

$$\frac{1-2\alpha-\varepsilon}{1-2\alpha}\theta+\frac{x\varepsilon}{1-2\alpha}=\theta+\frac{x-\theta}{1-2\alpha}\varepsilon$$

y entonces $IF(x,T,F_\theta)=\lim_{\varepsilon\to 0^+}\frac{\theta+\frac{x-\theta}{1-2\alpha}\varepsilon-\theta}{\varepsilon}=\frac{x-\theta}{1-2\alpha}$

2. si $x>F_\theta^{-1}(1-\alpha)$

$$\frac{1}{1-2\alpha}\int_{F_\theta^{-1}(\frac{\alpha}{1-\varepsilon})}^{F_\theta^{-1}(\frac{1-\alpha+\varepsilon}{1-\varepsilon})}u(1-\varepsilon)f_\theta(u)du$$

y como la integral tiene límites infinitesimalmente asimétricos respecto de θ, se la descompone en dos, una con límites simétricos y normalizandola por $1-2\alpha-\varepsilon$

$$\frac{1-2\alpha-\varepsilon}{1-2\alpha}\int_{F_\theta^{-1}(\frac{\alpha}{1-\varepsilon})}^{F_\theta^{-1}(\frac{1-\alpha}{1-\varepsilon})}\frac{u(1-\varepsilon)f_\theta(u)}{1-2\alpha-\varepsilon}du+\frac{1}{1-2\alpha}\int_{F_\theta^{-1}(\frac{1-\alpha}{1-\varepsilon})}^{F_\theta^{-1}(\frac{1-\alpha+\varepsilon}{1-\varepsilon})}u(1-\varepsilon)f_\theta(u)du$$

nuevamente la primera integral vale θ, y si se llama $F_\theta^{-1}(\frac{1-\alpha}{1-\varepsilon})=a$, aproximando resulta para la diferencia de límites de la segunda integral

$$F_\theta^{-1}(\frac{1-\alpha+\varepsilon}{1-\varepsilon})-F_\theta^{-1}(\frac{1-\alpha}{1-\varepsilon})\approx\frac{\varepsilon}{f_\theta(a)(1-\varepsilon)}$$

luego reemplazando

$$\frac{1-2\alpha-\varepsilon}{1-2\alpha}\theta+\frac{1}{1-2\alpha}\frac{\varepsilon}{f_\theta(a)(1-\varepsilon)}a(1-\varepsilon)f_\theta(a)=\theta+\frac{a-\theta}{1-2\alpha}\varepsilon$$

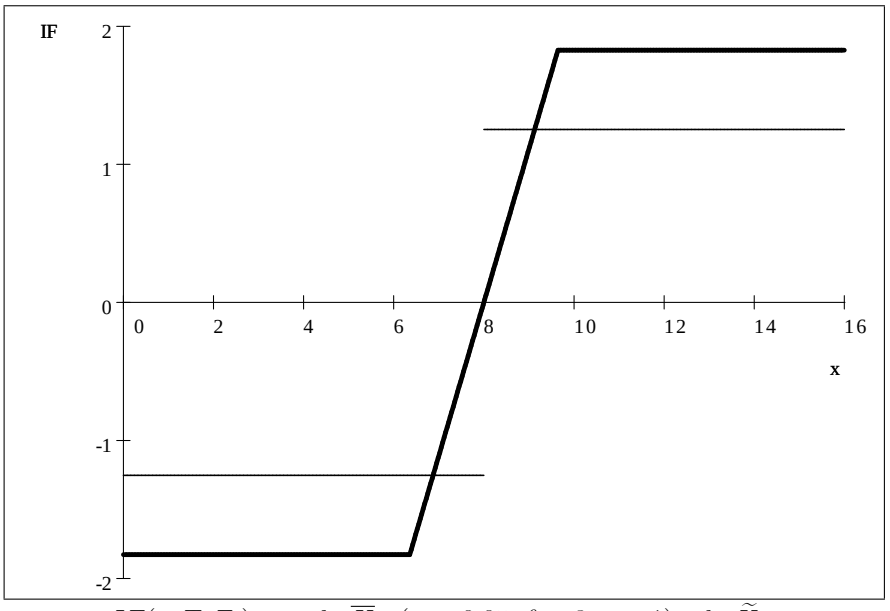
y entonces $IF(x,T,F_\theta)=\lim_{\varepsilon\to 0^+}\frac{\theta+\frac{a-\theta}{1-2\alpha}\varepsilon-\theta}{\varepsilon}=\lim_{\varepsilon\to 0^+}\frac{F_\theta^{-1}(\frac{1-\alpha}{1-\varepsilon})-\theta}{1-2\alpha}=\frac{F_\theta^{-1}(1-\alpha)-\theta}{1-2\alpha}$. Finalmente, luego de considerar el último caso $x<F_\theta^{-1}(\alpha)$, se llega a la expresión de la función de influencia, que es válida también para F continua, unimodal, y con densidad simétrica respecto de θ

$$IF(x,T,F_\theta)=\begin{cases}\frac{F_\theta^{-1}(\alpha)-\theta}{1-2\alpha} & \text{para } x<F_\theta^{-1}(\alpha)\\ \frac{x-\theta}{1-2\alpha} & \text{para } F_\theta^{-1}(\alpha)\le x\le F_\theta^{-1}(1-\alpha)\\ \frac{F_\theta^{-1}(1-\alpha)-\theta}{1-2\alpha} & \text{para } x>F_\theta^{-1}(1-\alpha)\end{cases}$$

En el caso de una normal, teniendo en cuenta que $F_\theta^{-1}(\alpha)=\theta-\sigma\Phi^{-1}(1-\alpha)$ y que $F_\theta^{-1}(1-\alpha)=\theta+\sigma\Phi^{-1}(1-\alpha)$

$$IF(x,T,F_\theta)=\begin{cases}\frac{-\sigma\Phi^{-1}(1-\alpha)}{1-2\alpha} & \text{para } x<\theta-\sigma\Phi^{-1}(1-\alpha)\\ \frac{x-\theta}{1-2\alpha} & \text{para } \theta-\sigma\Phi^{-1}(1-\alpha)\le x\le\theta+\sigma\Phi^{-1}(1-\alpha)\\ \frac{\sigma\Phi^{-1}(1-\alpha)}{1-2\alpha} & \text{para } x>\theta+\sigma\Phi^{-1}(1-\alpha)\end{cases}\tag{46}$$

A continuación se representa IF para la media truncada con $\alpha = 0.05$, $\theta = 8$, y $\sigma = 1$



$IF(x, T, F_\theta)$ para la $\overline{X}_\alpha$ ($\alpha = 0.05$, $\theta = 8, \sigma = 1$) y la $\widetilde{X}$

Ahora que se han obtenido las funciones de influencia de la media, mediana y media truncada, es apropiado hacer algunas comparaciones:

1. Notar que si $\alpha \to 0$, la IF de la media truncada tiende a la IF de la media, y que cuando $\alpha \to 0.5$, tiende a la IF de la mediana.

2. Si se compara en la figura los gráficos de la IF de la $\alpha-$media truncada(para $\sigma = 1$) y de la mediana, surge que el efecto sobre el sesgo asintótico de una contaminación pequeña, en un x cercano a θ, es mayor en la mediana(comparar $\pm\sqrt{\frac{\pi}{2}} = 1.25$ vs $\frac{x-\theta}{0.9}$); en cambio si x está alejado de θ, el efecto es mayor en la $\alpha-$media truncada(1.25 vs $\pm\frac{1.645}{0.9} = \pm 1.82$).

3. Finalmente la IF de la media no depende de σ, o sea, el efecto de una observación x, es $x - \theta$, ya sea que la normal tenga mucha o poca σ. En cambio para las otras dos sí depende de σ.

- **Varianza asintótica**: Como $V(T, F_\theta) = E_{F_\theta}(IF^2(X, T, F_\theta)$ y si se llama $h = \Phi^{-1}(1 - \alpha)$ queda

$$V(T, F_\theta) = \frac{\sigma^2}{(1 - 2\alpha)^2}\left[2\alpha h^2 + (1 - 2\alpha) - 2h\varphi(h)\right] \textbf{(OJO:verificar)}$$

- **Curva de sensibilidad**: Según la definición(19)

$$SC_n(x) = \frac{\overline{X}_\alpha(X_1, X_2, \cdots, X_n, x) - \overline{X}_\alpha(X_1, X_2, \cdots, X_n)}{1/(n+1)}$$

Por simplicidad en lo que sigue se considerará solo el caso en que $[(n-1)\alpha] = [(n)\alpha] = m$, es decir, al ampliar con x la muestra, no cambia m. La muestra ordenada tendrá el aspecto

$$X_{(1)} \quad \cdot \quad X_{(m)} \quad | \quad X_{(m+1)} \quad \cdot \quad \cdot \quad \cdot \quad X_{(n-m)} \quad | \quad X_{(n-m+1)} \quad \cdot \quad X_{(n)}$$

y entonces, llamando $H = \sum_{i=m+1}^{n-m} X_{(i)}$ en el numerador de la SC_n se tendrá:

$$\begin{array}{ll} \text{sí } x < X_{(m)} & \frac{X_{(m)}+H}{(n+1)-2m} - \frac{H}{n-2m} \\ \text{sí } X_{(m)} \leq x \leq X_{(n-m+1)} & \frac{x+H}{(n+1)-2m} - \frac{H}{n-2m} \\ \text{sí } x > X_{(n-m+1)} & \frac{X_{(n-m+1)}+H}{(n+1)-2m} - \frac{H}{n-2m} \end{array}$$

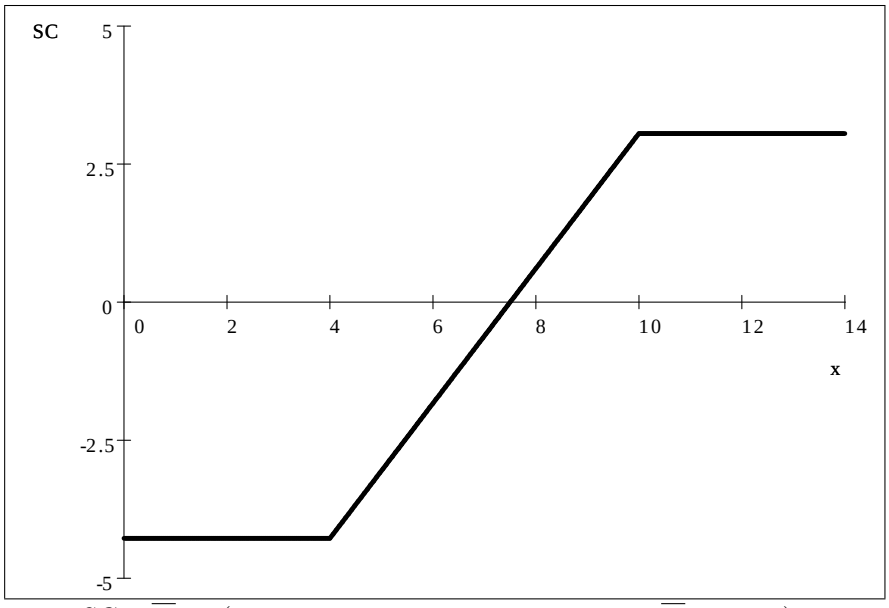
operando y dividiendo por $\frac{1}{n+1}$, queda

$$SC_n(x) = \begin{cases} \frac{X_{(m)}-\overline{X}_{\alpha n}}{n+1-2m}(n+1) & \text{sí } x < X_{(m)} \\ \frac{x-\overline{X}_{\alpha n}}{n+1-2m}(n+1) & \text{sí } X_{(m)} \leq x \leq X_{(n-m+1)} \\ \frac{X_{(n-m+1)}-\overline{X}_{\alpha n}}{n+1-2m}(n+1) & \text{sí } x > X_{(n-m+1)} \end{cases}$$

por ejempo para $\alpha = 0.1$, $n = 10$ y la muestra

$$\{4, 5, 6, 7, 7, 8, 9, 9, 9, 10\}$$

donde $m = 1$, y $X_{(m)} = 4, X_{(n-m+1)} = 10$, y $\overline{X}_{\alpha n} = 7.5$, queda



$SC_n : \overline{X}_{0.1}$, ($n = 10$, $X_{(m)} = 4, X_{(n-m+1)} = 10$, y $\overline{X}_{\alpha n} = 7.5$)

- **Sensibilidad a errores groseros**: Según (20) y la(46)

$$\gamma^* = \sup_{x \in \Omega} |IF(x, T, F_\theta)| = \frac{\sigma \Phi^{-1}(1-\alpha)}{1-2\alpha}$$

  luego la media truncada es **B-robusta.** O sea, si la contaminación es pequeña, el efecto que un outlier muy grande puede tener sobre el sesgo asintótico del estimador está acotado.

- **Sensibilidad a cambios locales**: Según (21) y la(46)

$$\lambda^* = \sup_{y \neq x} \frac{|IF(y, T, F_\theta) - IF(x, T, F_\theta)|}{|y - x|} = \frac{1}{1-2\alpha}$$

  Como λ^* es un indicador global(debido al supremo) de cambios locales(x próximo a y); el motivo del $\frac{1}{1-2\alpha}$ es que si se modifican levemente observaciones cercanas a θ(siempre suponiendo pequeña la contaminación), la variación relatica del sesgo asintótico del estimador será $\frac{1}{1-2\alpha}$. Pero si se modifican observaciones alejadas de θ, en realidad el efecto es nulo.

- **Punto de ruptura asintótico**: De (28)

$$\varepsilon^* = \sup_{\varepsilon \in (0,1)} \{\varepsilon : \exists K_\varepsilon \text{ con} \forall G \Longrightarrow \theta_\alpha((1-\varepsilon)F_\theta + \varepsilon G) \in K_\varepsilon\}$$

Tomando un $\varepsilon < \alpha$, hay que ver si $\exists K_\varepsilon$ acotado, tal que para $F = (1-\varepsilon)F_\theta + \varepsilon G$ con G arbitraria, resulta

$$\theta_\alpha(F) = \frac{1}{1-2\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x dF \in K_\varepsilon$$

pero se analizarán los límites a y b de esta integral. Para el primero hay que resolver

$$(1-\varepsilon)F_\theta(a) + \varepsilon G(a) = \alpha$$

y como $\forall G$, $G(a) \leq 1$, resulta $F_\theta(a) \geq \frac{\alpha-\varepsilon}{1-\varepsilon} > 0$, y de aquí $a \geq F_\theta^{-1}(\frac{\alpha-\varepsilon}{1-\varepsilon}) = a_\varepsilon$(finito). Para el otro se resuelve

$$(1-\varepsilon)F_\theta(b) + \varepsilon G(b = 1-\alpha$$

y como $\forall G$, $G(b) \geq 0$, resulta $F_\theta(b) \leq \frac{1-\alpha}{1-\varepsilon} < 1$, y luego $b \leq F_\theta^{-1}(\frac{1-\alpha}{1-\varepsilon}) = b_\varepsilon$(finito). Como hemos encontrado que los límites de la integral están acotados, entonces resultará $\theta_\alpha(F) \in [a_\varepsilon; b_\varepsilon] = K_\varepsilon$.conjunto acotado, como se quería demostrar. Finalmente si se toma $\varepsilon \geq \alpha$, y $G = \Delta_t$ con t grande, se puede probar que no existe un K_ε acotado donde caiga el sesultado de la integral. En definitiva resulta

$$\varepsilon^* = \alpha$$

- **Punto de ruptura para muestras finitas**: Según (29)

$$\varepsilon_n^* = \frac{1}{n}\underset{m\geq 0}{máx}\left\{m : \exists K_m \text{ tq. cambiando } x_{i_1}, \cdot\cdot, x_{i_m} \text{ arbitrariamente, } \overline{X}_{\alpha n}(\mathbf{x}) \in K_m\right\}$$

Se ilustrará con una muestra de $n = 100$ y $\alpha = 0.05$

$$X_{(1)} \cdots X_{(5)} \mid X_{(6)}X_{(7)}X_{(8)} \cdot\cdot\cdot\cdot\cdot X_{(93)}X_{(94)}X_{(95)} \mid X_{(96)} \cdots X_{(100)}$$

la media truncada se calcula promediando los 90 estadísticos de orden centrales. Si por ejemplo se toma $m = 2$, y se alteran arbitrariamente 2 observaciones, tratando que la nueva media truncada **aumente**, convendrá cambiar los 2 primeros estadísticos de orden de la parte central($X_{(6)}$ y $X_{(7)}$) y pasarlos a la derecha de toda la muestra. En este caso, la nueva media truncada valdrá$\frac{1}{90}\sum_{i=8}^{97} X_{(i)}$. Por otro lado, si lo que se busca es que **disminuya**, habrá que cambiar $X_{(94)}$ y $X_{(95)}$, y pasarlos a la izquierda de toda la muestra, resultando una nueva media truncada de $\frac{1}{90}\sum_{i=6}^{93} X_{(i)}$. En definitiva si $m = 2$ observaciones se cambian arbitrariamente, se tendrá que $\overline{X}_{\alpha n}(\mathbf{x}) \in [\frac{1}{90}\sum_{i=6}^{93} X_{(i)}; \frac{1}{90}\sum_{i=8}^{97} X_{(i)}] = K_u$ acotado. Lo mismo ocurrirá si $m = 3, 4, 6$ pero para $m \geq 6$, siempre quedarán en la parte central algunas

de estas observaciones, y entonces no existirá K_m. En definitiva el punto de ruptura en este caso será $\varepsilon_n^* = \frac{5}{100} = 0.05$. Y en el caso general

$$\varepsilon_n^* = \frac{m}{n} = \frac{[(n-1)\alpha]}{n} \text{ resultando también } \varepsilon_n^* \underset{n\to\infty}{\longrightarrow} \varepsilon^* = \alpha$$

- **Sesgo asintótico máximo**: Según (30)

$$\mathbf{MB}_{\theta_\alpha}(\varepsilon, \theta) = máx\left\{\left|b_{\widehat{\theta}_\alpha}(F, \theta)\right| \text{ con } F \in \mathcal{F}_{\theta\varepsilon}\right\}$$

se supondrá ε fijo, con $\varepsilon < \varepsilon^* = \alpha$, y entonces para todas las $F \in \{(1-\varepsilon)F_\theta + \varepsilon G \text{ con } G \in \mathcal{F} \text{ arbitraria}\}$ hay que estudiar el máximo $|sesgo|$. Pero notar que el sesgo positivo más grande se dará cuando la $\theta_\alpha((1-\varepsilon)F_\theta + \varepsilon G)$ sea lo mayor posible, y esto se logrará(pensarlo!), tomando a $G = \Delta_t$, con $t = b$, el extremo derecho del intervalo de integración. Entonces hay que calcular

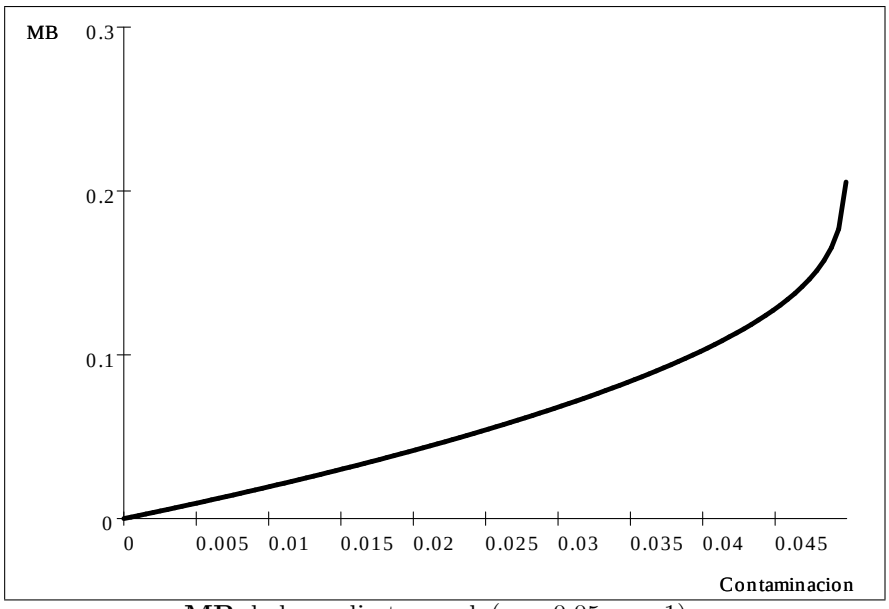
$$\theta_\alpha = \frac{1}{1-2\alpha}\int_a^b ud\left[(1-\varepsilon)F_\theta + \varepsilon\Delta_b\right] = \frac{1}{1-2\alpha}\int_a^b u(1-\varepsilon)f_\theta(u)du + \frac{\varepsilon b}{1-2\alpha}$$

donde $(1-\varepsilon)F_\theta(a) = \alpha$, y $(1-\varepsilon)F_\theta(b) = 1-\alpha$, resultando a y b simétricos respecto de θ, y en particular $b = F_\theta^{-1}(\frac{1-\alpha}{1-\varepsilon})$. Como $(1-\varepsilon)f_\theta(u)$ tiene entre a y b probabilidad $1-2\alpha-\varepsilon$, se puede escribir

$$\frac{1-2\alpha-\varepsilon}{1-2\alpha}\int_a^b u\frac{(1-\varepsilon)f_\theta(u)}{1-2\alpha-\varepsilon}du + \frac{\varepsilon b}{1-2\alpha} = \frac{1-2\alpha-\varepsilon}{1-2\alpha}\theta + \frac{\varepsilon b}{1-2\alpha} = \theta + \frac{(b-\theta)\varepsilon}{1-2\alpha}$$

entonces el sesgo máximo positivo es $\frac{(b-\theta)\varepsilon}{1-2\alpha}$. Y como razonando en forma similar, el sesgo máximo negativo es igual pero de distinto signo, se tendrá que $\mathbf{MB}_{\theta_\alpha}(\varepsilon, \theta) = \frac{(b-\theta)\varepsilon}{1-2\alpha}$. Y como $b = F_\theta^{-1}(\frac{1-\alpha}{1-\varepsilon}) = \theta + \sigma\Phi^{-1}(\frac{1-\alpha}{1-\varepsilon})$, se tendrá en definitiva

$$\mathbf{MB}_{\theta_\alpha}(\varepsilon, \theta) = \frac{\varepsilon\sigma\Phi^{-1}(\frac{1-\alpha}{1-\varepsilon})}{1-2\alpha}$$



MB de la media truncada($\alpha = 0.05, \sigma = 1$)

- **Sensibilidad a la contaminación**: Según (31) es la pendiente de la $\mathbf{MB}_{\theta_\alpha}(\varepsilon, \theta)$ en $\varepsilon = 0$, luego resulta

  $$\gamma_c(\theta) = aaaaa(\textbf{OJO:verificar})$$

  en particular en este caso $\gamma^* = \gamma_c(\theta)$, o sea vale la acotación($\gamma^* \leq \gamma_c(\theta)$) con la sensibilidad a errores groseros.

3.2 Modelo de escala

Se supondrá una muestra donde cada observación X_i satisface el modelo multiplicativo

$$X_i = \sigma U_i \quad (i = 1, \cdots, n)$$

donde $\sigma > 0$ es el parámetro de escala desconocido(se usa la notación σ, pero no necesariamente es el desvío estándar), y usualmente se asume también que:

$$U_1, U_2, \cdots, U_n \overset{\textbf{iid}}{\sim} F_0 \quad \text{(conocida)}$$

Entonces cada X_i se distribuye según una $F(x) = F_0(x/\sigma)$. Con la notación del capítulo anterior se tiene entonces el modelo paramétrico

$$X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F \text{ con } F \in P_\sigma = \{F_\sigma : F_\sigma(x) = F_0(x/\sigma)\}$$

y con el se busca un estimador $\widehat{\sigma}_n$ del parámetro de **escala** σ. Pero tratandose de un estimador de un parámetro de escala, debe cumplir

- Equivarianza de escala: $\forall c > 0$

$$\widehat{\sigma}_n(cX_1, cX_2, \cdots, cX_n) = c\widehat{\sigma}_n(X_1, X_2, \cdots, X_n)$$

3.3 Estimadores de dispersión

Cualquier estimador que satisfaga las siguientes dos equivarianzas, se llamará un **estimador de dispersión**.

- Equivarianza de traslación: $\forall X_i$, $\forall c$

$$\widehat{\sigma}_n(X_1 + c, X_2 + c, \cdots, X_n + c) = \widehat{\sigma}_n(X_1, X_2, \cdots, X_n) \tag{47}$$

- Equivarianza de escala: $\forall X_i$, $\forall k$

$$\widehat{\sigma}_n(kX_1, kX_2, \cdots, kX_n) = |k|\,\widehat{\sigma}_n(X_1, X_2, \cdots, X_n) \tag{48}$$

Con el primero nos aseguramos que si los datos de la muestra están todos desplazados c unidades, el valor que proporcione el estimador no cambiará; y con el segundo el estimador se ajustará automaticamente cuando cambien las unidades de la muestra. Los estimadores que se definirán a continuación cumplen estas dos equivarianzas, sin embargo en estimadores más sofisticados, hay que verificar su cumplimiento.

3.3.1 Desvío estándar

Se lo define mediante

$$SD(\mathbf{X}) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Como la función de influencia es no acotada, $\gamma^* = \infty$, tiene además punto de ruptura $\varepsilon^* = 0$, resultando un estimador decididamente no robusto.

3.3.2 Desviación absoluta media(MD)

Se define mediante

$$MD(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - \overline{X}\right|$$

que tampoco es robusto ($\gamma^* = \infty$ y $\varepsilon^* = 0$), ya que intervienen dos promedios no robustos: $\overline{X}$ y también $\frac{1}{n}\sum|\cdot\cdot|$.

3.3.3 Desviación absoluta mediana(MAD)

Buscando resolver los problemas de robustez del anterior se define

$$MAD(\mathbf{X}) = \underset{1\leq i\leq n}{Med}\left\{|X_i - Med(\mathbf{X})|\right\}$$

3.3.4 Rango intercuartil

3.4 Modelo de posición y escala

Como lo usual es desconocer tanto el parámetro de posición como el de escala, se plantea el modelo

$$X_i = \theta + \sigma U_i \quad (i = 1, \cdots, n)$$

donde comunmente se asume que:

1. $U_1, U_2, \cdots, U_n \overset{\textbf{iid}}{\sim} F_0$ (conocida)
2. F_0 es una distribución simétrica

De 1 surge que cada X_i se distribuye según una $F(x) = F_0(\frac{x-\theta}{\sigma})$; y con la notación del capítulo anterior se tiene el modelo paramétrico

$$X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F \quad \text{con} \quad F \in P_{\theta,\sigma} = \left\{ F_{\theta,\sigma} : F_{\theta,\sigma}(x) = F_0(\frac{x-\theta}{\sigma}) \right\}$$

y con el se buscarán estimadores de los dos parámetros.

4 M-estimadores

4.1 M-estimadores en general

Sea $X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F$, donde $F \in P_\theta = \{F_\theta : \theta \in \Theta\}$. Suponiendo que existe la densidad $f(x;\theta)$, la función de verosimilitud es

$$L(X_1, X_2, \cdot\cdot, X_n; \theta) = \prod_{i=1}^{n} f(X_i; \theta) \tag{49}$$

y el estimador $\widehat{\theta}_n = \widehat{\theta}(X_1, X_2, \cdot\cdot, X_n)$ de máxima verosimilitud de θ se obtiene mediante

$$\widehat{\theta}_n = \arg\max_\theta \prod_{i=1}^{n} f(X_i; \theta)$$

pero lo usual es calcular $-\ln L$, en (49), y luego obtener el mínimo o sea

$$\widehat{\theta}_n = \arg\min_\theta \sum_{i=1}^{n} -\ln f(X_i; \theta) \tag{50}$$

El estimador que se obtiene es muy bueno, con todas las propiedades de optimalidad que tiene el estimador de máxima verosimilitud. Sin embargo notar que hemos supuesto $F = F_\theta \in P_\theta$, es decir será válido solo para F dentro del modelo paramétrico P_θ. Se quiere ahora ampliar esta definición.

Sea ahora $X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F$, donde $F \in \mathcal{F} \supset P_\theta = \{F_\theta : \theta \in \Theta\}$. En la (50) se minimiza la suma de $-\ln f(x;\theta)$ en cada x_i. Pero con el fin de obtener

un estimador mas general, Huber(1964) propuso cambiar esta función por otra, $\rho(x;\theta)$ (que se elegirá de acuerdo a las propiedades que se quiere que tenga $\widehat{\theta}$), quedando

$$\widehat{\theta}_n = \arg\min_{\theta} \sum_{i=1}^{n} \rho(X_i;\theta) \tag{51}$$

y el estimador obtenido se llamará un M-estimador de θ. Por supuesto, si se toma

$$\rho(x,\theta) = -\ln f(x;\theta) \tag{52}$$

el M-estimador será el de MV, que sabemos que funciona muy bién en P_θ; pero la idea ahora es, con una conveniente elección de la función $\rho(x,\theta)$, obtener un estimador que funcione bastante bién dentro de P_θ, pero que también tenga un buén desempeño en $\mathcal{F}$.

Si $\rho(x,\theta)$ es diferenciable en θ, con derivada $\psi(\cdot,\theta) = \frac{\partial}{\partial\theta}\rho(\cdot,\theta)$, entonces también se puede expresar

$$\widehat{\theta}_n \text{ es una de las raíces de: } \sum_{i=1}^{n} \psi(X_i,\theta) = 0 \tag{53}$$

Estas ecuaciones (51 y 53) no siempre son equivalentes ya que la segunda a veces tiene varias raíces, sin embargo esta suele considerarse la que define el M-estimador.

Por último, aunque para definir el M-estimador se partió del estimador de MV para una densidad f, las funciones ρ y ψ no tienen porqué estar relacionadas a ninguna función de densidad.

4.1.1 Existencia y unicidad

Para resolver la (53) habrá que estudiar la función $g(\theta) = \sum_{i=1}^{n} \psi(x_i,\theta)$, prestando atención como posibles raíces a los $\widehat{\theta}$ en que $g(\theta)$ cambia de signo, o sea

$$S = \left\{ \widehat{\theta} : \text{si } \theta < \widehat{\theta},\ g(\theta) \geq 0 \quad \text{y} \quad \text{si } \theta > \widehat{\theta},\quad g(\theta) \leq 0 \right\}$$

Suponiendo que $\theta \in (\theta_a, \theta_b)$, un intervalo, que puede ser finito o infinito(por ejemplo si el parámetro es una media, sería $(-\infty;\infty)$, y si es una varianza $(0;\infty)$), se tiene el siguiente

Theorem 19 *Si $\psi(x,\theta)$ es **no-creciente** en θ y*

$$\lim_{\theta\to\theta_a} \psi(x,\theta) > 0 > \lim_{\theta\to\theta_b} \psi(x,\theta)$$

entonces:

1. $S \neq \emptyset$ y es un intervalo
2. Si ψ es **continua** en $\theta \implies S$ contiene las raíces de $g(\theta)$ (existencia)
3. Si ψ es **continua y estrictamente decreciente** $\implies S$ contiene 1 sola raíz (unicidad)

4.1.2 Expresión como funcional

A continuación se expresara con funcionales la ecuacione que definen el M-estimador(53). Recordando de (2), que si $g(X)$ es una función de la variable aleatoria $X \sim F_n$(la distribución empírica), entonces $E_{F_n}(g(X)) = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$, la (53) se puede expresar

$$\widehat{\theta}_n \text{ es una de las raíces de } \; E_{F_n}\left[\psi(X;\theta)\right] = 0 \tag{54}$$

llamando $\widehat{\theta}_n = T(F_n)$ al estimador, y $T(F)$ a su valor asintótico, la siguiente ecuación implícita define el funcional T de un M-estimador

$$T(F) \text{ es una de las raíces de } \; E_F\left[\psi(X;T(F))\right] = 0 \tag{55}$$

Notar que este funcional es una función $T : \mathcal{F} \longrightarrow \Theta$, pero si es consistente de Fisher($T(F_\theta) = \theta \quad \forall\theta \in \Theta$), también cumplirá

$$E_{F_\theta}\left[\psi(X;\theta)\right] = 0 \;\; \forall\theta \in \Theta \tag{56}$$

4.1.3 Consistencia

Resolviendo la (54) se obtiene $\widehat{\theta}_n = T(F_n)$, y de la (55) surge $T(F)$. ¿Pero se cumplirá que $\widehat{\theta}_n = T(F_n) \xrightarrow{\mathbf{p}} T(F)$? Este es el problema de consistencia del M-estimador.

Para resolver la (55) ahora habrá que estudiar la función $\lambda_F(\theta) = E_F\left[\psi(x;\theta)\right]$, prestando atención como posibles raíces a los $\theta_F = T(F)$ en que $\lambda_F(\theta)$ cambia de signo, o sea

$$S = \{\theta_F : \text{si } \theta < \theta_F, \;\; \lambda_F(\theta) \geq 0 \quad \text{y} \quad \text{si } \theta > \theta_F, \;\; \lambda_F(\theta) \leq 0 \;\}$$

Suponiendo que $\theta \in (\theta_a, \theta_b)$ un intervalo, que puede ser finito o infinito se tiene el siguiente

Theorem 20 *Si $\psi(x,\theta)$ es **no-creciente** en θ y*

$$\begin{array}{rcl} \lim\limits_{\theta\to\theta_a} \psi(x,\theta) & > & 0 > \lim\limits_{\theta\to\theta_b} \psi(x,\theta) \\ E_F\left|\psi(X,\theta)\right| & < & \infty \;\; \forall\theta \end{array}$$

entonces:

1. $S \neq \emptyset$
2. Si $\lambda_F(\theta)$ es **continua** en $\theta \implies S$ contiene las raíces de $\lambda_F(\theta)$ (existencia)
3. Si S contiene 1 sola raíz ($T(F)$) $\implies \widehat{\theta}_n = T(F_n) \xrightarrow{\mathbf{p}} T(F)$ (consistencia)

4.1.4 Función de influencia del M-estimador

Según la definición (15) de función de influencia, y llamando $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_x \in \mathcal{F}$

$$IF(x,T,F) = \lim_{\varepsilon\to 0^+} \frac{T\left[(1-t)F + \varepsilon\Delta_x\right] - T(F)}{\varepsilon} = \left.\frac{\partial T(F_\varepsilon)}{\partial\varepsilon}\right|_{\varepsilon=0}$$

Suponiendo que la (55) tiene solución única, en el primer término del numerador hay que resolver $E_{(1-\varepsilon)F+\varepsilon\Delta_x}\left[\psi(Y;T(F_\varepsilon))\right] = 0$. Por la linealidad de la esperanza(usando para no confundir Y en lugar de X)

$$(1-\varepsilon)E_F\left[\psi(Y;T(F_\varepsilon))\right] + \varepsilon E_{\Delta_x}\left[\psi(Y;T(F_\varepsilon))\right] = 0 \tag{57}$$

ahora de desarrollará en serie $\psi(y;T(F_\varepsilon))$ como función de ε

$$\psi(y;T(F_\varepsilon)) \approx \left.\psi(y;T(F_\varepsilon))\right|_{\varepsilon=0} + \left.\frac{\partial\psi(y;T(F_\varepsilon))}{\partial T(F_\varepsilon)}\frac{\partial T(F_\varepsilon)}{\partial\varepsilon}\right|_{\varepsilon=0}\varepsilon$$

pero

$$\left.\psi(y;T(F_\varepsilon))\right|_{\varepsilon=0} = \psi(y;T(F))$$

y usando la definición de función de influencia(15)

$$\left.\frac{\partial\psi(y;T(F_\varepsilon))}{\partial T(F_\varepsilon)}\frac{\partial T(F_\varepsilon)}{\partial\varepsilon}\right|_{\varepsilon=0} = \left.\frac{\partial\psi(y;\theta)}{\partial\theta}\right|_{\theta=T(F)} IF(x,T,F) = \psi'(y;T(F))IF(x,T,F) \tag{58}$$

luego el desarrollo queda

$$\psi(y;T(F_\varepsilon)) \approx \psi(y;T(F)) + \psi'(y;T(F))IF(x,T,F)\varepsilon$$

y reemplazando en la (57), (usando de (55) que $E_F(\psi(Y;T(F))) = 0$)

$$(1-\varepsilon)\left[E_F(\psi'(Y;T(F)))IF(x,T,F)\varepsilon\right] + \varepsilon\left[\psi(x;T(F)) + \psi'(x;T(F))IF(x,T,F)\varepsilon\right] = 0$$

simplificando y reordenando

$$(1-\varepsilon)\left[E_F(\psi'(Y;T(F)))IF(x,T,F)\right] + \psi(x;T(F)) + \psi'(x;T(F))IF(x,T,F)\varepsilon = 0$$

y tendiendo $\varepsilon \to 0$

$$E_F(\psi'(Y;T(F)))IF(x,T,F) + \psi(x;T(F)) = 0$$

y en definitiva(designando nuevamente X en lugar de Y), y suponiendo que el denominador no es nulo

$$IF(x,T,F) = \frac{\psi(x;T(F))}{-E_F(\psi'(X;T(F)))} \tag{59}$$

Donde, como se definió en (58)

$$\psi'(X;T(F)) = \left.\frac{\partial\psi(y;\theta)}{\partial\theta}\right|_{\theta=T(F)}$$

Aunque esta justificación no es muy rigurosa, si lo hicieron Huber(1981), y Clarke(1983,1984), este último probando la diferenciabilidad de Frechet de los M-estimadores en condiciones muy generales, aún para funciones $\psi(x;\theta)$ poco regulares.

Lo importante de la (59) es que la función de influencia del M-estimador es proporcional a la función ψ. Y en el caso que IF se evalue en F_θ, sindo T consistente de Fisher se tendrá

$$IF(x,T,F_\theta) = \frac{\psi(x;\theta)}{-E_{F_\theta}\left[\psi'(X;\theta)\right]} \tag{60}$$

Expresión alternativa En el caso de consistencia de Fisher, y cuando interesa la IF en F_θ(como en la 60) se dará otra expresión. Esta será útil cuando al derivar para obtener $\psi'(X;\theta)$ aparecen funciones impulsivas. Como por consistencia de Fisher

$$E_{F_\theta}\left[\psi(x;\theta)\right] = \int \psi(x;\theta) f_\theta(x)dx = 0$$

derivando respecto de θ

$$\begin{aligned} \int \psi'(x;\theta) f_\theta(x)dx + \int \psi(x;\theta)\frac{\partial}{\partial\theta} f_\theta(x)dx &= 0 \\ E_{F_\theta}\left[\psi'(X;\theta)\right] + \int \psi(x;\theta)\frac{\partial}{\partial\theta} f_\theta(x)dx &= 0 \end{aligned}$$

y reemplazando en la (60) queda

$$IF(x,T,F_\theta) = \frac{\psi(x;\theta)}{\int \psi(x;\theta)\frac{\partial}{\partial\theta} f_\theta(x)dx} \tag{61}$$

4.1.5 Normalidad asintótica

Suponiendo la consistencia del M-estimador, usando la distribución asintótica de $T(F_n)$ vista en (17), y la expresión de la función de influencia recién analizada(59), (omitiendo algunas condiciones de regularidad), se tiene que

$$\sqrt{n}\left[T(F_n) - T(F)\right] \xrightarrow{\mathbf{d}} N(0; \frac{E_F\left[\psi^2(X;T(F))\right]}{E_F^2\left[\psi'(X;T(F))\right]}) \tag{62}$$

luego la varianza asintótica es

$$V(T,F) = \frac{E_F\left[\psi^2(X;T(F))\right]}{E_F^2\left[\psi'(X;T(F))\right]} \tag{63}$$

Notar que vale la relación

$$V(T,F) = E_F\left[IF^2(X,T,F)\right]$$

4.2 M-estimador de posición

Nuevamente se analizará el modelo de posición, donde cada observación $X_i = \theta + U_i$, y usualmente se asume:

1. $U_1, U_2, \cdots, U_n \overset{\textbf{iid}}{\sim} F_0$ (conocida)
2. F_0 es una distribución simétrica

Según esto, el modelo paramétrico es

$$X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F \quad \text{con} \quad F \in P_\theta = \{F_\theta : F_\theta(x) = F_0(x-\theta)\}$$

pero para el análisis robusto se considerará también el modelo más general

$$X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F \quad \text{con} \quad F \in \mathcal{F} \supset P_\theta$$

La intención es estimar θ mediante un M-estimador como el de (51)

$$\widehat{\theta}_n = \arg\min_\theta \sum_{i=1}^n \rho(X_i;\theta)$$

pero se necesita definir la función $\rho(x;\theta)$. Si conociesemos F_0, lo ideal sería utilizar máxima verosimilitud, que según (52), equivale a tomar

$$\rho(x;\theta) = -\ln f_0(x-\theta) \tag{64}$$

y esta función tiene dos propiedades: 1) depende de $x-\theta$, o sea $\rho(x;\theta) = \rho(x-\theta)$. 2) si F_0 es simétrica, también lo será f_0, y entonces ρ será una función par de $x-\theta$, o sea $\rho(x-\theta) = \rho(-(x-\theta))$. En realidad no se tiene la certeza de la validez del modelo paramétrico, sin embargo como se quiere que el M-estimador funcione bién en el, se exigirá que siempre

$$\rho(x;\theta) = \rho(x-\theta) \;\; \text{y esto arrastra que } \psi(x;\theta) = \psi(x-\theta)$$

Como tanto ρ como ψ son funciones de $u = x-\theta$, para describirlas se estudiará $\rho(u)$ y $\psi(u)$ respectivamente. Y si F_0 es una distribución simétrica, se suele pedir también

$$\rho(x-\theta) = \rho\left[-(x-\theta)\right] \;\; \text{y esto arrastra} \quad \psi\left[-(x-\theta)\right] = -\psi(x-\theta)$$

es decir, $\rho(u)$ es par, y $\psi(u)$ impar. En definitiva el M-estimados de posición lo obtendremos mediante

$$\widehat{\theta}_n = \arg\min_\theta \sum_{i=1}^n \rho(X_i-\theta) \qquad \textbf{o} \qquad \widehat{\theta}_n \text{ es una de las raíces de:} \sum_{i=1}^n \psi(X_i-\theta) = 0 \tag{65}$$

Remark 21 *Notar que si en lugar de ρ se toma $\rho^* = a\rho + b$, con $a > 0$, la solución es la misma. Lo mismo si en lugar de ψ se toma $\psi^* = a\psi$ con $a > 0$.*

4.2.1 F. de influencia, existencia, unicidad, consistencia etc

Veamos algunas propiedades del M-estimador de posición:

- Cumplen la equivarianza de traslación:

$$\widehat{\theta}_n(X_1 + c, X_2 + c, \cdots, X_n + c) = \widehat{\theta}_n(X_1, X_2, \cdots, X_n) + c$$

  pero en general **no la de escala**($\widehat{\theta}_n(kX_1, kX_2, \cdots, kX_n) = k\widehat{\theta}_n(X_1, X_2, \cdot\cdot\cdot, X_n)$). Sin embargo este incumplimiento será resuelto más adelante.

- Se quiere ver si el funcional T de un M-estimador de posición es consistente de Fisher. Debería cumplirse $T(F_\theta) = \theta$, $\forall\theta$. Como el funcional esta definido por la (55)

$$T(F) \text{ es una de las raíces de } \; E_F\left[\psi(X - T(F))\right] = 0$$

  reemplazando F por F_θ, y $T(F_\theta)$ por θ, la consistencia de Fisher se expresa

$$E_{F_\theta}\left[\psi(X-\theta)\right] = \int \psi(x-\theta)dF_0(x-\theta) = \int \psi(u)f_0(u)du = E_{F_0}(\psi(U)) = 0 \tag{66}$$

  Notar que si f_0 es simétrica, y $\psi(u)$ es impar, esta integral es nula. O sea la consistencia de Fisher queda asegurada si F_0 tiene distribución simétrica, y se eligió una ψ impar.

- La función de influencia es, ajustando la(59)

$$IF(x, T, F) = \frac{\psi(x - T(F))}{E_F(\psi'(X - T(F)))}$$

  y si vale la consistencia de Fisher, en F_θ valdrá

$$IF(x, T, F_\theta) = \frac{\psi(x-\theta)}{E_{F_\theta}(\psi'(X-\theta))} = \frac{\psi(x-\theta)}{E_{F_0}(\psi'(U))}$$

- Para estudiar la existencia y unicidad de la solución $\widehat{\theta}_n$ es de aplicación el teorema-19 para M-estimadores en general, prestando atención a que ahora, al ser $\psi(x;\theta) = \psi(x-\theta) = \psi(u)$, "no-creciente en θ", debe leerse "$\psi(u)$ no-decreciente". Sin embargo en el caso de ser $\psi(u)$ impar (útil cuando F_0 es de distribución simétrica) se enunciará el siguiente

Theorem 22 *Sea $\psi(u)$ impar, continua y no-decreciente, con $\psi(a) > 0$ para algún a, entonces*

1. $\sum_{i=1}^{n}\psi(X_i - \theta) = 0$ tiene alguna raíz, y las raíces forman un intervalo.

2. Toda raíz es un mínimo de $\sum_{i=1}^{n}\rho(X_i - \theta)$

3. Si $\psi(u)$ es **estrictamente creciente** $\implies$ la raíz es única

- Para estudiar la consistencia, se podría usar el teorema-20 de consistencia en general, sin embargo en el caso de el M-estimador de posición se dará el siguiente resultado particular, para $F \in P_\theta$

  Theorem 23 *Si* $X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F$, *donde* $F \in P_\theta = \{F_\theta : \theta \in \Theta\}$, *y* $\widehat{\theta}_n(F_\theta)$ *es el M-estimador de posición solución de la (65), y se cumple:*

1. $E_{F_\theta}[\psi(X-\theta)] = 0, \ \ \forall\theta$ (consistencia de Fisher)
2. $\psi(u)$ estrictamente creciente

luego: $\widehat{\theta}_n(F_\theta) \xrightarrow{\textbf{ctp}} \theta$ (consistencia en casi todo punto).

- Suponiendo que $\widehat{\theta}_n = T(F_n)$ sea consistente con valor asintótico $T(F)$, de (62) aplicado al M-estimador de posición

$$\sqrt{n}\,[T(F_n) - T(F)] \xrightarrow{\textbf{d}} N(0; \frac{E_F\left[\psi^2(X-T(F))\right]}{E_F^2\left[\psi'(X-T(F))\right]})$$

$$\text{con} \ \ V(T,F) \ = \ \frac{E_F\left[\psi^2(X-T(F))\right]}{E_F^2\left[\psi'(X-T(F))\right]} = E_F\left[IF^2(X,T,F)\right]$$

y en el caso de $F = F_\theta$ con consistencia de Fisher, resultan las expresiones

$$\sqrt{n}\left[\widehat{\theta}_n - \theta\right] \xrightarrow{\textbf{d}} N(0; \frac{E_{F_0}\left[\psi^2(U)\right]}{E_{F_0}^2\left[\psi'(U)\right]}) \ \ \text{con} \ \ V(T,F_\theta) = \frac{E_{F_0}\left[\psi^2(U)\right]}{E_{F_0}^2\left[\psi'(U)\right]}$$

Example 24 *Suponiendo que* $F_0 = N(0;\sigma)$ *con* σ *conocida, entonces* $P_\theta = \{F_\theta : F_\theta = N(\theta;\sigma)\}$. *Si se quiere el M-estimador que corresponde a máxima verosimilitud, de (64) habrá que tomar* $\rho(u) = -\ln f_0(u) = \ln\sqrt{2\pi}\sigma + \frac{u^2}{2\sigma^2}$, *o sea descartando constantes:* $\rho(u) = \frac{u^2}{2}$, *y* $\psi(u) = u$. *Luego, según la definición*

$$\widehat{\theta}_n \text{ es una de las raíces de: } \sum_{i=1}^{n}(X_i - \theta) = 0 \qquad \text{o sea } \ \widehat{\theta}_n = \overline{X}_n$$

como se esperaba. Y sabemos que este estimador funciona muy bién dentro de P_θ, *pero mal fuera de el. Como* $\rho(u) = \frac{u^2}{2}$ *da mucha importancia a observaciones con* $u = x - \theta$ *grande, se probará ahora con otra función* ρ, *(que debe ser par pues la* $N(\theta;\sigma)$ *es simétrica), por ejemplo* $\rho(u) = |u|$, *y entonces* $\psi(u) = sgn(u)$. *El M-estimador que corresponde a esta* ψ *surge de*

$$\widehat{\theta}_n \text{ es una de las raíces de: } \sum_{i=1}^{n} sgn(X_i - \theta) = 0$$

si bién ψ es impar, no es ni continua, ni estrictamente creciente, así que con los teoremas dados no necesariamente se debe esperar una solución única. Si se considera la muestra ordenada, llamando $k = \left[\frac{n+1}{2}\right]$, el llamado conjunto S de cambios de signo (ver teorema-19) es:

$$\begin{array}{l} \textit{si } n \textit{ es impar} \;\; S = X_{(k)} \\ \textit{si } n \textit{ es par} \;\; S = \left[X_{(k)}, X_{(k+1)}\right] \end{array}$$

luego en definitiva $\widehat{\theta}_n = \widetilde{X}_n$ es una mediana muestral(por ejemplo la mediana principal).

Ahora bién, este M-estimador, la mediana muestral, que tiene función $\psi(u) = sgn(u)$, ¿será un estimador de máxima verosimilitud para alguna F_0?

Como $\rho(u) = -\ln f_0(u)$, y $\rho'(u) = sgn(u)$, resolviendo se obtiene $f_0(u) = \frac{1}{2}e^{-|u|}$ para $-\infty < u < \infty$, o sea una densidad doble exponencial. Resumiendo los dos casos

F_0	ρ_{MV}	ψ_{MV}	$\widehat{\theta}_{MV}$
normal	$\frac{u^2}{2}$	u	$\overline{X}_n$
doble exponencial	$\lvert u\rvert$	$sgn(u)$	$\widetilde{X}_n$

En el siguiente ejemplo se verificarán algunos resultados vistos para la mediana, usando la teoría de M-estimadores de posición.

Example 25 *Como la mediana muestral corresponde al M-estimadores.con $\psi(u) = sgn(u)$, se tiene entonces:*

1. Funcional: hay que resolver $E_F\left[sgn(X - T(F))\right] = 0$ o sea

$$0 = \int sgn(x - T(F))dF = -\int\limits_{x<T(F)} dF + \int\limits_{x>T(F)} dF$$

   o sea $P(X < T(F)) = P(X > T(F))$, y esto define a $T(F)$ como una mediana de F. Por ejemplo se puede tomar la mediana principal $T(F) = Medp(F)$.

2. Consistencia de Fisher: $E_{F_0}(\psi(U)) = \int sgn(u)d\Phi(\frac{u}{\sigma}) = 0$

3. Función de influencia: $IF(x, T, F_\theta) = \frac{\psi(x-\theta)}{E_{F_0}(\psi'(U))} = \frac{sgn(x-\theta)}{E_{F_0}(\psi'(U))}$ donde

$$E_{F_0}(\psi'(U)) = \int \psi'(u)d\Phi(\frac{u}{\sigma})$$

   pero debido a la discontinuidad, $\psi'(u)$ tiene una función delta en cero, siendo: $\psi'(u) = \left[\psi(0^+) - \psi(0^-)\right]\delta(u)$, luego

$$\begin{aligned} \int \psi'(u)d\Phi(\frac{u}{\sigma}) &= \int \left[\psi(0^+) - \psi(0^-)\right]\delta(u)d\Phi(\frac{u}{\sigma}) = \left[\psi(0^+) - \psi(0^-)\right]\varphi(\frac{0}{\sigma})\frac{1}{\sigma} \\ &= \left[1 - (-1)\right]\frac{1}{\sqrt{2\pi}\sigma} = \sqrt{\frac{2}{\pi}}\frac{1}{\sigma} \end{aligned}$$

   y reemplazando queda $IF(x, T, F_\theta) = sgn(x - \theta)\sqrt{\frac{\pi}{2}}\sigma$.

Example 26 *En el ejemplo anterior, para $\psi(u) = sgn(u)$ el funcional del M-estimador resultó ser $T(F) = Medp(F)$. Ahora se quiere estudiar la distribución asintótica de $\widehat{\theta}_n = T(F_n)$ en las dos situaciones siguientes(suponiendo la validez de las correspondientes condiciones de regularidad):*

1. $F \in P_\theta = \{F_\theta = N(\theta; \sigma)\}$: en este caso

$$\sqrt{n}\left[\widehat{\theta}_n - \theta\right] \xrightarrow{\mathbf{d}} N(0; \frac{E_{F_0}\left[\psi^2(U)\right]}{E^2_{F_0}\left[\psi'(U)\right]})$$

donde $E_{F_0}\left[\psi^2(U)\right] = E_{F_0}\left[sgn^2(U)\right] = E_{F_0}\left[1\right] = 1$, y según el ejemplo anterior $E^2_{F_0}\left[\psi'(U)\right] = \frac{2}{\pi}\frac{1}{\sigma^2}$ luego queda

$$\sqrt{n}\left[\widehat{\theta}_n - \theta\right] \xrightarrow{\mathbf{d}} N(0; \frac{\pi}{2}\sigma^2)$$

que es el comportamiento asintótico de la mediana muestral que ya hemos analizado anteriormente, suponiendo que la muestra viene de una población normal.

2. $F \notin P_\theta$, concretamente $f(x) = \frac{1}{\sqrt{2}\sigma}e^{-\left|\frac{\sqrt{2}(x-\theta)}{\sigma}\right|}$, que en lugar de ser una $N(\theta; \sigma)$ como en el caso anterior, es una doble exponencial desplazada, de igual media θ y desvío σ. En este caso se usará la expresión general

$$\sqrt{n}\left[T(F_n) - T(F)\right] \xrightarrow{\mathbf{d}} N(0; \frac{E_F\left[\psi^2(X - T(F))\right]}{E^2_F\left[\psi'(X - T(F))\right]})$$

donde $T(F) = Medp(F) = \theta$, $E_F\left[sgn^2(X - \theta)\right] = 1$, y recordando la delta en el denominador

$$\begin{aligned} E_F\left[\psi'(X-\theta)\right] &= \int \psi'(u)dF(u) = \int \left[\psi(0^+) - \psi(0^-)\right]\delta(u)dF(u) = \left[\psi(0^+) - \psi(0^-)\right]f(0) \\ &= \left[1 - (-1)\right]\frac{1}{\sqrt{2}\sigma} = \frac{\sqrt{2}}{\sigma} \end{aligned}$$

luego en definitiva

$$\sqrt{n}\left[\widehat{\theta}_n - \theta\right] \xrightarrow{\mathbf{d}} N(0; \frac{\sigma^2}{2})$$

notar que la varianza asintótica es menor, ya que la mediana es justamente el estimador de máxima verosimilitud, si F_0 es doble exponencial.

4.2.2 Indicadores de robustez

Aunque obviamente las propiedades del M-estimador dependerán de la función ψ que lo caracteriza(tema que se tratará en la siguiente sección), a continuación se analizarán en general los diferentes indicadores de robustez.

Sensibilidad a errores groseros Según la definición, suponiendo F_0 de distribución simétrica

$$\gamma^* = \sup_{x\in\Omega} |IF(x,T,F_\theta)| = \frac{\sup_{x\in\Omega} |\psi(x-\theta)|}{E_{F_\theta}(\psi'(X-\theta))} = \frac{\sup_{x\in\Omega} |\psi(x-\theta)|}{E_{F_0}(\psi'(U))}$$

Como la IF es proporcional a ψ, entonces si ψ es acotada, la sensibilidad a errores groseros (γ^*) será finita, y por tanto el estimador será B-robusto.

Si además de acotada, ψ es no-decreciente, se tendrá (llamando $\psi(\infty) = k$)

$$\gamma^* = \frac{k}{E_{F_0}(\psi'(U))}$$

(notar que en este caso γ^* no depende de θ).

Y si ψ es acotada y estrictamente creciente, entonces el M-estimador será también cualitativamente robusto.

Punto de ruptura asintótico Según la definición(26)

$$\varepsilon^* = \sup_{\varepsilon\in(0,1)} \left\{ \varepsilon : \exists K_\varepsilon \text{ con} \forall G \Longrightarrow \widehat{\theta}_\infty((1-\varepsilon)F_\theta + \varepsilon G) \in K_\varepsilon \right\}$$

para evitar problemas de existencia y consistencia supondremos que ψ es no-decreciente, y además llamaremos(k^- y k^+ ambos positivos)

$$\lim_{u\to-\infty} -\psi(u) = k^- < 0 < \lim_{u\to\infty} \psi(u) = k^+$$

Como se necesita calcular el valor asintótico $\widehat{\theta}_\infty$ para cada ε dado, habrá que resolver

$$E_{(1-\varepsilon)F_\theta+\varepsilon G}\left[\psi(X-\widehat{\theta}_\infty)\right] = (1-\varepsilon)E_{F_\theta}\left[\psi(X-\widehat{\theta}_\infty)\right] + \varepsilon E_G\left[\psi(X-\widehat{\theta}_\infty)\right] = 0$$

tomemos primero un $\varepsilon < \varepsilon^*$, entonces existirá K_ε(acotado), tal que cualquiera sea G, el valor asintótico $\widehat{\theta}_\infty$ siempre estará en K_ε(acotado). Ahora se tomará $G = \Delta_t$, entonces la expresión anterior queda

$$0 = (1-\varepsilon)E_{F_\theta}\left[\psi(X-\widehat{\theta}_\infty)\right] + \varepsilon\psi(t-\widehat{\theta}_\infty)$$

y si se tiende $t \to \infty$ resultará $\psi(t-\widehat{\theta}_\infty) \to k^+$(ya que $\widehat{\theta}_\infty$ está en un conjunto acotado), y como siempre $\psi \geq -k^-$ resultará

$$0 \geq -(1-\varepsilon)k^- + \varepsilon k^+ \longrightarrow \varepsilon \leq \varepsilon^+ = \frac{k^-}{k^- + k^+}$$

y si se tiende $t \to -\infty$ resultará $\psi(t-\widehat{\theta}_\infty) \to -k^-$(ya que $\widehat{\theta}_\infty$ está en un conjunto acotado), y como siempre $\psi \leq k^+$ resultará

$$0 \leq (1-\varepsilon)k^+ - \varepsilon k^- \longrightarrow \varepsilon \leq \varepsilon^- = \frac{k^+}{k^- + k^+}$$

o sea, de ambas resulta que $\varepsilon \leq \min\{\varepsilon^+, \varepsilon^-\} = \frac{\min\{k^-,k^+\}}{k^-+k^+}$. La demostración continúa suponiendo que $\varepsilon > \varepsilon^*$, y como ahora no existirá K_ε(acotado), $\widehat{\theta}_\infty$ se escapará de todo límite, y procediendo en forma similar resulta que en este caso debe ser $\varepsilon \geq \frac{\min\{k^-,k^+\}}{k^-+k^+}$. En definitiva el punto de ruptura asintótico será

$$\varepsilon^* = \frac{\min\{k^-, k^+\}}{k^- + k^+}$$

y en el caso que ψ sea impar, como $k^- = k^+$, resultará $\varepsilon^* = 0.5$.

Punto de ruptura para muestras finitas Recordando la definición dada en (29), si $\mathbf{x} = \{x_1, x_2, \cdot\cdot, x_n\}$ es una muestra fija, y $\widehat{\theta}_n = \widehat{\theta}_n(\mathbf{x})$ un estimador,.se define

$$\varepsilon_n^* = \varepsilon_n^*(\widehat{\theta}_n, \mathbf{x}) = \frac{1}{n} \underset{m\geq 0}{máx} \left\{ m : \exists K_m \text{ tq. cambiando } x_{i_1}, \cdot\cdot, x_{i_m} \text{ arbitrariamente, } \widehat{\theta}_n(\mathbf{x}) \in K_m \right\}$$

y se interpreta como la mayor proporción de observaciones de la muestra que pueden ser arbitrariamente modificados, manteniendo el estimador en K_u.

En el caso de los estimadores de posición equivariantes (este tema se verá más adelante), se demuestra que

$$\varepsilon_n^* \leq \frac{1}{n}\left[\frac{n-1}{2}\right]$$

y la cota se alcanza para los M-estimadores, si se cumple

$$\text{M-est. equivariante } \ \psi \text{ impar y acotada } \ \Longrightarrow \varepsilon_n^* = \frac{1}{n}\left[\frac{n-1}{2}\right]$$

Sesgo asintótico máximo En general para estimadores de posición equivariantes el sesgo asintótico máximo $\mathbf{MB}_{\widehat{\theta}}(\varepsilon, \theta)$ no depende de θ. Para obtenerlo se puede usar la definición $\mathbf{MB}_{\widehat{\theta}}(\varepsilon, \theta) = máx\left\{\left|b_{\widehat{\theta}}(F, \theta)\right| \text{ con } F \in \mathcal{F}_{\theta\varepsilon}\right\}$, sin embargo se enunciará un teorema, que bajo ciertas condiciones, facilita su cálculo.

Theorem 27 *Sea $F_\theta(x) = F_0(x-\theta)$, con F_0 simétrica, ψ no-decreciente y acotada, y $\psi(\infty) = k$. Entonces* $\mathbf{MB}(\boldsymbol{\varepsilon}) = b_\varepsilon$ *es la solución de*

$$E_{F_0}\left[\psi(U - b_\varepsilon)\right] = \frac{-\varepsilon}{1-\varepsilon}k \tag{67}$$

La idea de la demostración es la siguiente: para obtener un valor del sesgo b_ε, se fija un $\varepsilon < \varepsilon^*$, y luego una G, y con ella se obtiene $F = (1-\varepsilon)F_\theta + \varepsilon G$ y se resuelve en t

$$E_F\left[\psi(X - t)\right] = 0$$

el sesgo será $b_\varepsilon = t - \theta$. Notar que si G es simétrica centrada en θ, F también lo será, y como ψ es impar, resultará $t = \theta$, y por tanto $b_\varepsilon = 0$. Luego, si se

quiere que haya sesgo positivo habrá que tomar por ejemplo $G = \Delta_{\theta+y}$ *con* y *positivo (*G *a la derecha de* θ*). Cuanto mayor sea* y*,* t *necesitará ser mayor,* y *mayor será el sesgo. Pero calculemoslo. Como* $\psi(X-t) = \psi(X-\theta+\theta-t) = \psi(X-\theta-b_\varepsilon)$ *y reemplazando*

$$\begin{aligned}(1-\varepsilon)E_{F_\theta}\left[\psi(X-\theta-b_\varepsilon)\right] + \varepsilon E_{\Delta_{\theta+y}}\left[\psi(X-\theta-b_\varepsilon)\right] &= 0\\ (1-\varepsilon)E_{F_0}\left[\psi(U-b_\varepsilon)\right] + \varepsilon\psi(y-b_\varepsilon) &= 0\end{aligned}$$

notar que si y *es muy grande, variando* b_ε *el primer término valdrá*

$$\begin{array}{ll}\text{para } b_\varepsilon = 0 & \varepsilon k > 0\\ \text{para } b_\varepsilon = y/2 & -(1-\varepsilon)k + \varepsilon k < 0\end{array}$$

luego suponiendo continuidad,la solucion estará en un b_ε *tal que*

$$(1-\varepsilon)E_{F_0}\left[\psi(U-b_\varepsilon)\right] + \varepsilon k = 0$$

y de aquí la tesis.

Sensibilidad a la contaminación: Derivando la (67) respecto de ε

$$E_{F_0}\left[\psi'(U-b_\varepsilon)\right]\frac{db_\varepsilon}{d\varepsilon} = \frac{k}{(1-\varepsilon)^2}$$

y para $\varepsilon = 0$, $\left.\frac{db_\varepsilon}{d\varepsilon}\right|_{\varepsilon=0} = \gamma_c$ quedando

$$\gamma_c = \frac{k}{E_{F_0}\left[\psi'(U)\right]} = \gamma^*$$

o sea en este caso(F_0 simétrica, ψ no-decreciente y acotada), la sensibilidad a la contaminación es igual a la sensibilidad a errores groseros(recordar de (33) que siempre vale la acotación $\gamma_c(\theta) \geq \gamma^*$).

4.3 M-estimador de escala

Aunque lo usual es deconocer tanto el parámetro de posición como el de escala, se analizará primero el caso en que solo la escala es desconocida.

Recordando el modelo de escala donde cada observación X_i satisface el modelo multiplicativo

$$X_i = \sigma U_i \quad (i = 1, \cdots, n)$$

siendo $\sigma > 0$ el parámetro de escala desconocido, y como es usual asumiendo

$$U_1, U_2, \cdots, U_n \overset{\textbf{iid}}{\sim} F_0 \text{ conocida}$$

Entonces cada X_i se distribuye según una $F(x) = F_0(x/\sigma)$. Con la notación del capítulo anterior se tiene entonces el modelo paramétrico

$$X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F \text{ con } F \in P_\sigma = \{F_\sigma : F_\sigma(x) = F_0(x/\sigma)\}$$

pero para un análisis robusto se considerará también el modelo más general

$$X_1, X_2, \cdots, X_n \overset{\textbf{iid}}{\sim} F \quad \text{con} \quad F \in \mathcal{F} \supset P_\sigma$$

y se busca un estimador $\widehat{\sigma}_n$ del parámetro de **escala** σ.

Suponiendo primero que estamos dentro del modelo paramétrico y que existe $f_\sigma(x) = \frac{1}{\sigma} f_0(\frac{x}{\sigma})$, la función de verosimilitud es

$$L(X_1, X_2, \cdot\cdot, X_n; \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma^n} f_0(\frac{X_i}{\sigma})$$

el estimador $\widehat{\sigma}_n = \widehat{\sigma}(X_1, X_2, \cdot\cdot, X_n)$ de máxima verosimilitud de σ se obtiene mediante

$$\widehat{\sigma}_n = \arg\max_{\sigma} \prod_{i=1}^{n} \frac{1}{\sigma^n} f_0(\frac{X_i}{\sigma})$$

pero lo usual es calcular $-\ln L$, y luego obtener el mínimo o sea

$$\widehat{\sigma}_n = \arg\min_{\sigma} \sum_{i=1}^{n} \left[\ln \sigma - \ln f_0(\frac{X_i}{\sigma}) \right]$$

si se deriva respecto de σ (y multiplicando por -1) queda

$$\widehat{\sigma}_n \text{ es una de las raíces de: } \sum_{i=1}^{n} \left[-\frac{f_0'(\frac{X_i}{\sigma})}{f_0(\frac{X_i}{\sigma})} \frac{X_i}{\sigma} - 1 \right] = 0 \tag{68}$$

y ahora sí el M-estimador de escala queda definido mediante

$$\widehat{\sigma}_n \text{ es una de las raíces de: } \sum_{i=1}^{n} \chi(\frac{X_i}{\sigma}) = 0 \tag{69}$$

donde la función $\chi(u)$ juega el mismo papel que $\psi(u)$ en el estimador de posición. Notar que aquí, si se toma $\chi_0(u) = -\frac{f_0'(u)}{f_0(u)} u - 1$, resulta $\widehat{\sigma}_n$ el estimador de máxima verosimilitud, (lo mismo ocurría tomando $\psi_0(u) = -\frac{f_0'(u)}{f_0(u)}$ en el estimador de posición).

Notación de Yohai La anterior es la notación de Huber en su libro. Sin embargo la empleada por Yohai, que es útil en estimadores de escala en regresión, se basa en expresar la (68) así

$$\widehat{\sigma}_n \text{ es una de las raíces de: } \frac{1}{n} \sum_{i=1}^{n} - \frac{f_0'(\frac{X_i}{\sigma})}{f_0(\frac{X_i}{\sigma})} \frac{X_i}{\sigma} = 1$$

luego el M-estimador de escala se define mediante

$$\widehat{\sigma}_n \text{ es una de las raíces de: } \frac{1}{n} \sum_{i=1}^{n} \rho(\frac{X_i}{\sigma}) = \delta \tag{70}$$

donde la función $\rho(u)$ junto con la constante δ caracterizan al estimador de escala. Notar sin embargo que esta ρ no es la misma que la utilizada en el estimador de posición, (aunque se usa la misma notación), pero tiene **en general** las mismas características, a saber:

1. $\rho(u)$ es función no decreciente de $|u|$
2. $\rho(0) = 0$
3. $\rho(u)$ es creciente para $u > 0$, con $\rho(u) < \rho(\infty)$
4. si $\rho(u)$ es acotada, se asume que $\rho(\infty) = 1$

Notar también que para que (70) tenga solución, deberá ser $0 < \delta < \rho(\infty)$. Entonces en el caso que ρ sea acotada, se asumirá sin pérdida de generalidad que $\rho(\infty) = 1$ y que $\delta \in (0, 1)$.

Además si F_0 es de distribución simétrica, $\rho(u)$ será par. Y cuando se elige $\rho_0(u) = -\frac{f_0'(u)}{f_0(u)}u$ y $\delta_0 = 1$ se obtiene el estimador de máxima verosimilitud.

Por último vale la equivalencia entre las dos notaciones

$$\chi(u) = \rho(u) - \delta$$

4.3.1 Características y propiedades

De acuerdo a (69) o (70) el funcional $T(F)$ de un M-estimador de escala está definido por

$$E_F\left[\chi(\frac{X}{T(F)})\right] = 0 \quad \text{o también} \quad E_F\left[\rho(\frac{X}{T(F)})\right] = \delta \tag{71}$$

y si es consistente de Fisher, valdrá $\forall\sigma$

$$E_{F_\sigma}\left[\chi(\frac{X}{\sigma})\right] = E_{F_0}\left[\chi(u)\right] = 0 \quad \text{o también} \quad E_{F_\sigma}\left[\rho(\frac{X}{\sigma})\right] = E_{F_0}\left[\rho(u)\right] = \delta \tag{72}$$

pero notar que a diferencia del estimador de posición, si F_0 es simétrica, $\chi(u)$ y $\rho(u)$ serán pares, pero esto no implica la validez de estas expresiones.

Remark 28 *Notar que $T(F)$ es una medida de la escala de F, pero no necesariamente es el desvío estándar (podría ser el MD, MAD, IRQ, etc). Entonces si $F = F_\sigma$ con σ desvío estándar, las consistencias de Fisher de (72) no se cumplirán, ya que el valor asíntotico es $T(F) \neq \sigma$. Pero siempre se podrá encontrar un $c > 0$, tal que $\frac{T(F)}{c} = \sigma$ coincidente con el desvío estándar. Entonces para hallar c, y dado que $T(F) = c\sigma$ habrá que exigir*

$$E_{F_\sigma}\left[\rho(\frac{X}{c\sigma})\right] = E_{F_0}\left[\rho(\frac{u}{c})\right] = \delta \tag{73}$$

y de aquí se despeja c. Luego en la expresión del M-estimador se utilizará $\rho(\frac{u}{c})$ en lugar de $\rho(u)$. De esta manera queda garantizado que asintóticamente el estimador estima el desvío estándar σ.

Para obtener la función de influencia se usará la definición general(59) para M-estimadores, cuidando de derivar en el denominador $\chi(\frac{X}{T(F)})$ respecto de $T(F)$ quedando

$$IF(x,T,F) = \frac{\chi(\frac{X}{T(F)})T(F)}{E_F\left[\chi'(\frac{X}{T(F)})\frac{X}{T(F)}\right]} = \frac{\left[\rho(\frac{X}{T(F)}) - \delta\right]T(F)}{E_F\left[\rho'(\frac{X}{T(F)})\frac{X}{T(F)}\right]}$$

y si vale la consistencia de Fisher, en F_σ se tendrá

$$IF(x,T,F_\sigma) = \frac{\chi(\frac{X}{\sigma})\sigma}{E_{F_0}\left[\chi'(u)u\right]} = \frac{\left[\rho(\frac{X}{\sigma}) - \delta\right]\sigma}{E_{F_0}\left[\rho'(u)u\right]}$$

Finalmente estos estimadores cumplen la equivarianza de escala, o sea:

$$\forall c > 0, \quad \widehat{\sigma}_n(cX_1, cX_2, \cdots, cX_n) = c\widehat{\sigma}_n(X_1, X_2, \cdots, X_n) \tag{74}$$

y en el caso que $\chi(u), \rho(u)$ sean pares, también la más general que se exige a los estimadores de dispersión, o sea:

$$\forall c, \quad \widehat{\sigma}_n(cX_1, cX_2, \cdots, cX_n) = |c|\,\widehat{\sigma}_n(X_1, X_2, \cdots, X_n)$$

Punto de ruptura Huber demuestra que si

$$\chi(u) \text{ es par, y creciente para } u > 0 \Longrightarrow \varepsilon^* = \frac{-\chi(0)}{\chi(\infty) - \chi(0)} \leq \frac{1}{2} \tag{75}$$

sin embargo hay un aspecto que descarta en su prueba.

Cuando se estudió el punto de ruptura del M-estimador de posición(4.2.2), si había una proporción $\varepsilon > \varepsilon^+$ de valores atípicos muy grandes a la derecha, el estimador se disparaba fuera de toda cota a ∞; y si había una proporción $\varepsilon > \varepsilon^-$ de valores atípicos muy grandes a la izquierda, ocurría lo mismo pero hacia $-\infty$. Y se definía el punto de ruptura a $\varepsilon^* = \min\{\varepsilon^+, \varepsilon^-\}$.

Para el estimador de escala también se presentan las dos alternativas: si hay una proporción $\varepsilon > \varepsilon^+$ de valores atípicos muy grandes en valor absoluto, el estimador de escala se disparará fuera de todo límite a ∞. Estos son los llamados "outliers" Pero si hay una proporción $\varepsilon > \varepsilon^-$ de obervaciones muy cercanas a cero, el estimador se acercará fuera de todo límite a cero. Estos son los "inliers", que son los que descarta Huber en su demostración, y que analizaremos ahora. Por supuesto, el punto de ruptura combina a ambos, mediante $\varepsilon^* = \min\{\varepsilon^+, \varepsilon^-\}$.

Lo que haremos es convertir un estimador de escala en uno de posición, y utilizaremos el punto de ruptura de este.

En el estimador de escala se resuelve $E_F\left[\rho(\frac{X}{\sigma})\right] = \delta$, o sea

$$E_F\left[\rho(\frac{X}{\sigma}) - \delta\right] = 0$$

si en lugar de trabajar con X, lo hacemos con $Y = \ln X$, resultará $X = e^Y$; y si llamamos $\mu = \ln \sigma$, será $\sigma = e^\mu$, reemplazando se tiene

$$E_F\left[\rho(\frac{e^Y}{e^\mu}) - \delta\right] = E_F\left[\rho(e^{Y-\mu}) - \delta\right] = 0$$

y si ahora se define la función $\psi(u) = \rho(e^u) - \delta$ queda

$$E_F\left[\psi(Y - \mu)\right] = 0$$

que es la ecuación que define a un estimador de posición. Notar que como $X = e^Y$, una proporción $\varepsilon > \varepsilon^+$ de observaciones $Y \to \infty$, implica $X \to \infty$, los "outliers"; mientras que una proporción $\varepsilon > \varepsilon^-$ de observaciones $Y \to -\infty$, implica $X \to 0$, los "inliers". Según la notación usada en estimadores de posición

$$\begin{aligned} k^+ &= \lim_{u\to\infty} \psi(u) = \lim_{u\to\infty} \rho(e^u) - \delta = 1 - \delta \\ k^- &= \lim_{u\to-\infty} -\psi(u) = \lim_{u\to-\infty} -\rho(e^u) + \delta = \delta \end{aligned}$$

donde se ha supuesto la validez de las propiedades generales de una función ρ vistas en (4.3), como es que ρ es par con $\rho(0) = 0$, y además acotada con $\rho(\infty) = 1$. De aquí surge que

$$\varepsilon^* = \min\left\{\delta, 1 - \delta\right\} \tag{76}$$

y δ y $1-\delta$ son los puntos de ruptura por "inliers", y "outliers" respectivamente. Obviamente en MV $\delta = 1$, lo que equivale a punto re ruptura $\varepsilon^* = 0$. Sin los supuestos respecto de ρ se tendrá

$$\varepsilon^* = \frac{\min\left\{-\rho(0) + \delta, \rho(\infty) - \delta\right\}}{\rho(\infty) - \rho(0)}$$

Finalmente, los estimadores de dispersión SD, MAD e IQR tienen puntos de ruptura $\varepsilon^* = 0$, 0.5 y 0.25 respectivamente. Y en general los puntos de ruptura de los estimadores de dispersión equivariantes son siempre ≤ 0.5.

Y cuando se elige $\rho_0(u) = -\frac{f_0'(u)}{f_0(u)}u$ y $\delta = 1$ se obtiene el estimador de máxima verosimilitud.

Example 29 *Si $F_0 = N(0,1)$, como $\frac{f_0'(u)}{f_0(u)} = -u$, tomando $\rho_0(u) = -\frac{f_0'(u)}{f_0(u)}u = u^2$, y $\delta_0 = 1$ debe obtenerse el estimador de MV. Aplicando la (70) resulta*

$$\frac{1}{n}\sum_{i=1}^n \rho_0(\frac{X_i}{\sigma}) = \frac{1}{n}\sum_{i=1}^n \frac{X_i^2}{\sigma^2} = 1 = \delta_0 \Longrightarrow \widehat{\sigma}_n = \sqrt{\frac{1}{n}\sum_{i=1}^n X_i^2}$$

que es la varianza muestral, cuando la media poblacional vale cero. Para hallar el funcional de este estimador, usando(71), hay que resolver

$$E_F\left[\rho_0(\frac{X}{T(F)})\right] = E_F\left[\frac{X^2}{T^2(F)}\right] = \frac{E_F(X^2)}{T^2(F)} = 1 = \delta_0$$

resultando $T(F) = \sqrt{E_F(X^2)}$*, que verifica* $T(F_\sigma) = \sqrt{E_{F_\sigma}(X^2)} = \sqrt{0^2 + \sigma^2} = \sigma$*, consistente de Fisher; pero sin embargo de (75) sale que* $\varepsilon^* = 0$*, decididamente un estimador no robusto.*

Example 30 *Se analizará ahora el M-estimador de escala con función* $\rho(u) = I(|u| > 1)$ *y* $\delta = 0.5$*, o sea*

$$\rho(u) = \begin{cases} 0 & para\ |u| \leq 1 \\ 1 & para\ |u| > 1 \end{cases}$$

Si cuando F_σ *es la* $N(0;\sigma)$ *se quiere que el estimador asintóticamente estime* σ*, de (73) deberá buscarse un* $c > 0$ *tal que*

$$E_{F_\sigma}\left[\rho(\frac{X}{c\sigma})\right] = E_\Phi\left[\rho(\frac{U}{c})\right] = E_\Phi\left[I(\left|\frac{U}{c}\right| > 1)\right] = P_\Phi(\left|\frac{U}{c}\right| > 1) = \delta = 0.5$$

o sea

$$1P_\Phi(U < -c) + 1P_\Phi(U > c) = 0.5 \quad o\ sea\ P_\Phi(U < c) = \Phi(c) = \frac{3}{4}$$

luego deberá ser $c = \Phi^{-1}(\frac{3}{4}) = 0.675$*, y se usará* $\rho(\frac{u}{c})$*. Ahora se buscará el funcional* $T(F)$*. Usando (71) y poniendo* t *en lugar de* $T(F)$ *para simplificar la notación, se tiene*

$$E_F\left[\rho(\frac{X}{ct})\right] = E_F\left[I(\left|\frac{X}{ct}\right| > 1)\right] = P_F(\left|\frac{X}{ct}\right| > 1) = 0.5$$

luego

$$P_F(|X| > ct) = 0.5$$

entonces ct *será una mediana de* $|X|$*, o sea* $ct = Med_F(|X|)$*, y en definitiva*

$$T(F) = \frac{Med_F(|X|)}{c} = \frac{Med_F(|X|)}{0.675}$$

y si se utiliza (70) resulta el estimador

$$\widehat{\sigma}_n = \frac{Med(|X_i|)}{c} = \frac{Med(|X_i|)}{0.675}$$

O sea, el estimador es la mediana de los valores absolutos, corregido por $c = 0.675$ *para lograr consistencia a* σ*, cuando la distribución es normal. Finalmente el punto de ruptura es* $\varepsilon^* = \min\{\delta, 1-\delta\} = 0.5$*.*

Remark 31 *Si se usa la misma función* $\rho(u) = I(|u| > 1)$ *pero con otro valor de* δ*, el estimador de escala que se obtiene es el* h*-ésimo estadístico de orden de los* $\{|X_1|, |X_2|, \cdot\cdot, |X_n|\}$*, donde* $h = n - [n\delta]$*, o sea*

$$\widehat{\sigma}_n = \frac{|X_i|_{(h)}}{c}$$

donde c *habrá que calcularlo para consistencia al desvío estándar de la normal. Pero ahora el punto de ruptura* $\varepsilon^* = \min\{\delta, 1-\delta\}$ *será menor que* 0.5*.*

4.4 M-estimadores de posición equivariantes

4.4.1 Problema con la equivarianza de escala

Considere un modelo de posición, con modelo paramétrico

$$X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} F \quad \text{con} \quad F \in P_\theta = \{F_\theta : F_\theta(x) = F_0(x-\theta)\}$$

donde F_0 es totalmente conocida. Se supondrá que teniendo en cuenta apropiados indicadores de robustez, se ha elegido la función $\psi(u)$ que define al M-estimador de θ.

Se cumplirá entonces que

$$\begin{aligned} \text{(consistencia de Fisher)} \qquad E_{F_0}\left[\psi(u)\right] &= 0 \\ \widehat{\theta}_n \ \text{ surge de resolver } \sum_{i=1}^{n}\psi(X_i - \widehat{\theta}_n) &= 0 \end{aligned} \tag{77}$$

Si ahora se considera la muestra desplazada $X_i^* = X_i + c$, es natural esperar que el valor que proporcione el estimador este igualmente desplazado, o sea que $\widehat{\theta}_n^* = \widehat{\theta}_n + c$. Verifiquemoslo:

$$\sum_{i=1}^{n}\psi(X_i^* - \widehat{\theta}_n^*) = \sum_{i=1}^{n}\psi(X_i + c - \widehat{\theta}_n - c) = \sum_{i=1}^{n}\psi(X_i - \widehat{\theta}_n) = 0$$

es decir el M-estimador cumple la equivarianza de traslación. Considerese ahora un cambio de escala en la muestra, $X_i^{**} = kX_i$, también es natural que el valor que proporcione el estimador este proporcionalmente modificado, o sea que $\widehat{\theta}_n^{**} = k\widehat{\theta}_n$. Verifiquemoslo:

$$\sum_{i=1}^{n}\psi(X_i^{**} - \widehat{\theta}_n^{**}) = \sum_{i=1}^{n}\psi(kX_i - k\widehat{\theta}_n) = \sum_{i=1}^{n}\psi\left[k(X_i - \widehat{\theta}_n)\right]$$

pero esto no necesariamente será cero. Como ejemplo, si la ψ elegida verifica $\psi(ku) = g(k)\psi(u)$, que es el caso de $\overline{X}_n$ donde $\psi(u) = u$, o de la mediana $\widetilde{X}_n$ donde $\psi(u) = sgn(u)$, la expresión anterior valdrá cero, y se cumplirá la equivarianza de escala. Pero la idea es no agregar una nueva exigencia a ψ.

Para enfatizar la importancia de la equivarianza de escala, suponga que en este problema las X_i se miden en metros, y el valor del estimador resultó $\widehat{\theta}_n = 3m$. Si después, usando la muestra en centímetros, se resuelve la ecuación del estimador y da $\widehat{\theta}_n^{**} = 250cm$, la pregunta es ¿Cual es el valor correcto de la estimación?

La respuesta es: $\widehat{\theta}_n = 3m$. Lo que pasa es que con la primer muestra, para la cual se eligió ψ, se verifica $E_{F_0}\left[\psi(u)\right] = 0$ ya que esta es una exigencia del M-estimador; pero con la segunda en cm, F_0 cambia y se convierte en F_0^{**}, y entonces $E_{F_0^{**}}\left[\psi(u)\right] \neq 0$. Si se quisiera llegar al valor correcto con la segunda muestra, habría que modificar ψ por otra ψ^{**}, que cumpla la consistencia de Fisher.

4.4.2 Ampliación a un modelo de posición-escala(conocida)

Para soluciona el problema anterior se ampliará el modelo a uno de posición y escala, ya que queremos que el estimador absorba los cambios de escala. Se asume que

$$X_i = \theta + \sigma U_i \text{ donde } U_i \sim F_0 \text{ conocida, } \sigma \text{ conocido}$$

donde σ, es un indicador de escala de σU_i. Primero supondremos que $\sigma = 1$, y entonces $X_i = \theta + U_i$ es un modelo de posición simple. Si se elige una $\psi(u)$, y esta deberá cumplir $E_{F_0}(\psi(u)) = 0$.

A continuación hagamos un cambio de escala en la muestra, por ejemplo $X_i^{**} = \sigma X_i$, entonces el nuevo modelo será $X_i^{**} = \sigma\theta + \sigma U_i = \sigma\theta + V_i$, o sea de posición pero con $V_i \sim F_0^{**}$ y $f_0^{**}(v) = \frac{1}{\sigma} f_0(\frac{v}{\sigma})$. Notar que haciendo un cambio de variable

$$\begin{aligned} 0 &= E_{F_0}\left[\psi(U)\right] = \int \psi(u) f_0(u) du = \int \psi(\frac{v}{\sigma}) f_0(\frac{v}{\sigma}) \frac{dv}{\sigma} \\ &= \int \psi(\frac{v}{\sigma}) f_0^{**}(v) dv = E_{F_0^{**}}\left[\psi(\frac{U}{\sigma})\right] \end{aligned}$$

entonces si cambia la escala en la muestra, lo único que habrá que hacer para que siga valiendo la consistencia de Fisher es usar $\psi(\frac{u}{\sigma})$ en lugar de $\psi(u)$, donde σ es un indicador de escala de la muestra. Y si la escala no cambia, también vale la última expresión ya que $\psi(\frac{u}{1}) = \psi(u)$.

En definitiva entonces generalizamos el M-estimador de posición así:

$$\begin{aligned} \text{Modelo de posición-escala} &: X_i = \theta + \sigma U_i \text{ donde } U_i \sim F_0 \text{ conocida, y } F_\theta(x) = F_0\left[\frac{x-\theta}{\sigma}\right] \\ \text{Consistencia de Fisher} &: \psi(u) \text{ verifica } E_{F_0}\left[\psi(U)\right] = E_{F_\theta}\left[\psi(\frac{X-\theta}{\sigma})\right] = 0 \\ \text{Estimador} &: \widehat{\theta}_n \text{ es una de las raíces de:} \sum_{i=1}^{n} \psi(\frac{X_i - \theta}{\sigma}) = 0 \end{aligned} \quad (78)$$

y ahora sí, con esta definición el M-estimador de posición cumple las dos equivarianzas. Seguidamente se enunciarán algunas relaciones, recordando siempre que σ es un indicador de escala de σU_i:

- La función de influencia es

$$IF(x, T, F) = \frac{\sigma\psi(\frac{x-T(F)}{\sigma})}{E_F(\psi'(\frac{X-T(F)}{\sigma}))}$$

  y en F_θ valdrá

$$IF(x, T, F_\theta) = \frac{\sigma\psi(\frac{x-\theta}{\sigma})}{E_{F_\theta}(\psi'(\frac{X-\theta}{\sigma}))} = \frac{\sigma\psi(\frac{x-\theta}{\sigma})}{E_{F_0}(\psi'(U))}$$

- Suponiendo que $\widehat{\theta}_n = T(F_n)$ sea consistente con valor asintótico $T(F)$

$$\sqrt{n}\,[T(F_n) - T(F)] \xrightarrow{\mathbf{d}} N(0; \frac{\sigma^2 E_F\left[\psi^2(\frac{X-T(F)}{\sigma})\right]}{E_F^2\left[\psi'(\frac{X-T(F)}{\sigma})\right]}) \;\text{ con }\; V(T,F) = \frac{\sigma^2 E_F\left[\psi^2(\frac{X-T(F)}{\sigma})\right]}{E_F^2\left[\psi'(\frac{X-T(F)}{\sigma})\right]}$$

y en el caso de $F = F_\theta$

$$\sqrt{n}\left[\widehat{\theta}_n - \theta\right] \xrightarrow{\mathbf{d}} N(0; \frac{\sigma^2 E_{F_0}\left[\psi^2(U)\right]}{E_{F_0}^2\left[\psi'(U)\right]}) \;\text{ con }\; V(T,F_\theta) = \frac{\sigma^2 E_{F_0}\left[\psi^2(U)\right]}{E_{F_0}^2\left[\psi'(U)\right]}$$

4.5 M-estimadores de posición con escala desconocida

Se considerará nuevamente el modelo de posición-escala, pero ahora suponiendo que la escala es también desconocida

$$X_i = \theta + \sigma U_i \text{ donde } U_i \sim F_0 \text{ conocida, } \sigma \textbf{ desconocido}$$

hay dos alternativas para tratar este modelo.

4.5.1 Estimación previa de la escala

La solución más natural es (ver 78) usar un indicador de escala muestral $\widehat{\sigma}_n$ y luego resolver

$$\widehat{\theta}_n \text{ es una de las raíces de:} \sum_{i=1}^{n} \psi(\frac{X_i - \theta}{\widehat{\sigma}_n}) = 0 \tag{79}$$

pero para que el M-estimador $\widehat{\theta}_n$ obtenido cumpla las dos equivarianzas de posición, el estimador de escala utilizado $\widehat{\sigma}_n$ debe cumplir las dos equivarianzas de escala, que es lo mismo que exigir que debe ser un **estimador de dispersión**.

Obviamente es importante que $\widehat{\sigma}_n$ sea un estimador robusto. Por ejemplo si se usa $\widehat{\sigma}_n = SD(\mathbf{X})$, y hay outliers, $\widehat{\sigma}_n$ estará inflada, y entonces muy probablemente en la (79) los outliers no sean detectados como tales en $\frac{X_i-\theta}{\widehat{\sigma}_n}$.

En el caso general, la IF correspondiente a $\widehat{\theta}_n$ depende también, con una expresión muy complicada, de la IF que corresponde a $\widehat{\sigma}_n$. Pero si F_0 es simétrica, ψ impar y χ par, todo se simplifica, y la IF de $\widehat{\theta}_n$ depende de $\widehat{\sigma}_n$, solo a través de su valor asintótico. Luego, si $\widehat{\sigma}_n \xrightarrow{\mathbf{p}} \sigma_\infty$ (es consistente)

$$IF(x,T,F) = \frac{\sigma_\infty \psi(\frac{x-T(F)}{\sigma_\infty})}{E_F(\psi'(\frac{X-T(F)}{\sigma_\infty}))} \tag{80}$$

y la distribución de $\widehat{\theta}_n$ será aproximadamente normal, o sea

$$\sqrt{n}\left[\widehat{\theta}_n - \theta_\infty\right] \xrightarrow{\mathbf{d}} N(0; \frac{\sigma_\infty^2 E_F\left[\psi^2(\frac{X-\theta_\infty}{\sigma_\infty})\right]}{E_F^2\left[\psi'(\frac{X-\theta_\infty)}{\sigma_\infty})\right]}) \;\text{ con }\; V(\widehat{\theta}_n,F) = \frac{\sigma_\infty^2 E_F\left[\psi^2(\frac{X-\theta_\infty}{\sigma_\infty})\right]}{E_F^2\left[\psi'(\frac{X-\theta_\infty)}{\sigma_\infty})\right]}$$

donde vale también

$$V(\widehat{\theta}_n, F) = E_F\left[IF^2(X, T, F)\right]$$

en otras palabras la eficiencia de $\widehat{\theta}_n$, no depende de la de $\widehat{\sigma}_n$. Y esto significa que a la hora de elegir $\widehat{\sigma}_n$, debe prestarse atención a otros aspectos de su robustez, pero no es importante su varianza asintótica. Sin embargo se verá más adelante que la robustez de $\widehat{\theta}_n$ sí depende de la de $\widehat{\sigma}_n$.

Por último una observación: en principio el estimador de dispersión puede ser cualquiera robusto (no conviene el IQR ya que tiene $\varepsilon^* = 0.25$), pero sí en cambio el MAD. Sin embargo, como en general el parámetro k de la función ψ se ajusta para lograr cierta eficiencia cuando $F_0 = N(0, 1)$, si queremos que $\widehat{\theta}_n$ tenga la misma eficiencia para cualquier normal $N(\theta; \sigma)$, se buscará un $\widehat{\sigma}_n$ que sea consistente a σ, o sea $\widehat{\sigma}_n \xrightarrow{\mathbf{p}} \sigma$. Este es el motivo por el cual se usa el $MADN$ (normalizado).

Punto de ruptura

- Cuando ψ es monotona, acotada e impar, se demuestra que $\varepsilon^*(\widehat{\theta}) = \varepsilon^*(\widehat{\sigma})$. Entonces si $\widehat{\sigma}_n = MADN(\mathbf{X})$, será $\varepsilon^*(\widehat{\theta}) = 0.5$, y si $\widehat{\sigma}_n = SD(\mathbf{X})$ resulta $\varepsilon^*(\widehat{\theta}) = 0$.

Pero, si se observa la expresión de la función de influencia(80), resultan idénticas en ambos casos (para el $MADN(\mathbf{X})$ y $SD(\mathbf{X})$), ya que en ella solo interviene el valor asintótico σ_∞. Sin embargo los puntos de ruptura son muy diferentes. El último es un ejemplo de un estimador con IF acotada, pero $\varepsilon^* = 0$.

- Si para el estimador de θ se usa una ρ acotada, y por lo tanto ψ será re-descendente, el problema se complica ya que $\varepsilon^*(\widehat{\theta})$ depende no solo de $\varepsilon^*(\widehat{\sigma})$ sinó también de la magnitud de $\widehat{\sigma}$. Sin embargo si como estimador de σ, se estima previamente θ por ejemplo con la mediana $\widetilde{\theta}$ (que tiene $\varepsilon^*(\widetilde{\theta}) = 0.5$), y se usa un M-estimador de escala con ρ_0 también acotado, mediante

  $$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{x_i - \widetilde{\theta}}{\sigma}\right) = 0.5$$

  (que por ser $\delta = 0.5$ tiene $\varepsilon^*(\widehat{\sigma}) = 0.5$), entonces si $\rho \leq \rho_0$, resultará $\varepsilon^*(\widehat{\theta}) = 0.5$. Estas ideas se utilizarán más adelante cuando se analice el MM-estimador en regresión.

4.5.2 Estimación simultanea de la posición y escala

La otra alternativa es proceder a la estimación simultanea de $\widehat{\theta}_n$ y $\widehat{\sigma}_n$. Consideremos nuevamente el modelo de posición-escala

$$X_i = \theta + \sigma U_i \text{ donde } U_i \sim F_0 \text{ conocida, } \sigma \textbf{ desconocido}$$

y el modelo paramétrico

$$X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} F \quad \text{con} \quad F \in P_{\theta,\sigma} = \left\{ F_{\theta,\sigma} : F_{\theta,\sigma}(x) = F_0(\frac{x-\theta}{\sigma}) \right\}$$

Suponiendo que existe la densidad $f_{\theta,\sigma}(x) = \frac{1}{\sigma} f_0(\frac{x-\theta}{\sigma})$, la función de verosimilitud es

$$\begin{aligned} L(X_1, X_2, \cdot\cdot, X_n; \theta, \sigma) &= \frac{1}{\sigma^n} \prod_{i=1}^{n} f_0(\frac{X_i - \theta}{\sigma}) \\ \text{y también} \quad -\ln L_{\theta,\sigma} &= n \ln \sigma - \sum_{i=1}^{n} \ln f_0(\frac{X_i - \theta}{\sigma}) \end{aligned}$$

derivando respecto de θ y de σ resulta

$$\begin{aligned} \sum_{i=1}^{n} \frac{f_0'(\frac{X_i-\theta}{\sigma})}{f_0(\frac{X_i-\theta}{\sigma})} &= 0 \\ \frac{1}{n} \sum_{i=1}^{n} - \frac{f_0'(\frac{X_i-\theta}{\sigma})}{f_0(\frac{X_i-\theta}{\sigma})} \frac{X_i - \theta}{\sigma} &= 1 \end{aligned}$$

entonces el M-estimador simultaneo de θ y σ se define mediante

$$(\widehat{\theta}_n, \widehat{\sigma}_n) \text{ es una de las raíces de: } \begin{cases} \sum_{i=1}^{n} \psi(\frac{X_i - \widehat{\theta}_n}{\widehat{\sigma}_n}) = 0 \\ \frac{1}{n} \sum_{i=1}^{n} \rho_s(\frac{X_i - \widehat{\theta}_n}{\widehat{\sigma}_n}) = \delta \end{cases}$$

que para el caso de MV $\psi(u) = -\frac{f_0'(u)}{f_0(u)}$, $\rho_s(u) = \psi(u)u$, $\delta = 1$. Y como siempre los funcionales $T(F)$ y $S(F)$ correspondientes a estos estimadores quedan definidos mediante

$$E_F \left[\psi(\frac{X - T(F)}{S(F)}) \right] = 0 \quad \text{y} \quad E_F \left[\rho_s(\frac{X - T(F)}{S(F)}) \right] = \delta$$

También si F es simétrica, asintoticamente la distribución de $\widehat{\theta}_n$ será aproximadamente normal, o sea

$$\sqrt{n} \left[\widehat{\theta}_n - \theta \right] \xrightarrow{\mathbf{d}} N(0; \frac{\sigma^2 E_F \left[\psi^2(\frac{X-\theta}{\sigma}) \right]}{E_F^2 \left[\psi'(\frac{X-\theta)}{\sigma}) \right]}) \quad \text{con} \quad V(\widehat{\theta}_n, F) = \frac{\sigma^2 E_F \left[\psi^2(\frac{X-\theta}{\sigma}) \right]}{E_F^2 \left[\psi'(\frac{X-\theta)}{\sigma}) \right]}$$

donde θ y σ son soluciones del sistema

$$\begin{cases} E_F \left[\psi(\frac{X-\theta}{\sigma}) \right] = 0 \\ E_F \left[\rho_s(\frac{X-\theta}{\sigma}) \right] = \delta \end{cases}$$

En general el estimador con estimación previa de la dispersión es más robusto que el simultaneo, y con menores problemas numéricos.

5 Optimalidad de M-estimadores

Un M-estimador de posición queda caracterizado por su función $\rho(u)$, o $\psi(u)$. Ahora se abordará el problema de la elección de estas funciones. De lo tratado hasta aquí surgen algunas pautas, por ejemplo:

1. que si F_0 es de distribución simétrica, elegir $\rho(u)$ par, y $\psi(u)$ impar
2. como la IF del M-estimador es proporcional a $\psi(x - T(F))$, si se quiere que la sensibilidad a errores groseros(γ^*) sea finita, habrá que elegir una $\psi(u)$ acotada

Pero como un estimador robusto debe comportarse bien tanto en P_θ como en un entorno $\mathcal{F}_{\theta\varepsilon} = \{F = (1-\varepsilon)F_\theta + \varepsilon G \text{ con } G \text{ arbitraria}\}$, la elección de ψ debe tener en cuenta los diferentes indicadores de robustez que tendrá el estimador propuesto(máximo sesgo, varianza asintótica, eficiencia, punto de ruptura, etc.).

Como usualmente se elegirá un estimador consistente de Fisher, para $F \in P_\theta$ no habrá problema de sesgo, así que aquí trataremos de tener una baja varianza del estimador, o equivalentemente, alta eficiencia(ARE) respecto del estimador de MV correspondiente. Por otro lado para $F \in \mathcal{F}_{\theta\varepsilon}$ en general aparecerá un sesgo además de la varianza. Pero como para n grande la varianza del estimador aparece dividida por n, y el sesgo no, en $\mathcal{F}_{\theta\varepsilon}$ trataremos de que el sesgo del estimador sea bajo. Resumiendo, las recomendaciones son

- En P_θ es importante una baja varianza (o sea, alta eficiencia (ARE))
- En $\mathcal{F}_{\theta\varepsilon}$ es importante bajo sesgo

Pero como estos dos requisitos no se pueden lograr simultaneamente(mínima varianza y mínimo sesgo), habrá que buscar alguna solución intermedia.

Por último, pero no imprescindible, por motivos de unicidad, consistencia o facilidad de cálculo, sería conveniente pedir que ψ sea no-decreciente, o creciente, o continua. Sin embargo muchas veces se dejarán de lado estos últimos requisitos.

5.1 Optimalidad en el sesgo (MB(ε))

Si prestamos atención solo al sesgo, es de utilidad la siguiente proposición de Huber(1964), llamada la "propiedad minimax de la mediana"

Proposition 32 *Si F_0 tiene distribución simétrica y unimodal, $\widetilde{X}$ es la mediana muestral, y $\widehat{\theta}$ es cualquier otro estimador equivariante de posición, entonces resultará* $\mathbf{MB}_{\widetilde{X}}(\varepsilon) \leq \mathbf{MB}_{\widehat{\theta}}(\varepsilon)$.

Según esto, la mediana muestral tiene menor sesgo máximo($\forall \varepsilon < \varepsilon^*$) que cualquier otro estimador equivariante de posición. Así que si esto nos interesa, habrá que tomar $\psi(u) = sgn(u)$, la función ψ de la mediana.

5.2 Optimalidad en la varianza ($V(\psi, F)$)

Suponiendo F_0 de distribución simétrica, entonces dentro de $P_\theta = \left\{F_\theta(x) = F_0(\frac{x-\theta}{\sigma})\right\}$, donde se supone σ conocida, un muy buen estimador es el de máxima verosimilitud, que según (52) corresponde a

$$\rho_0(u) = -\ln f_0(u) \quad \text{y a} \quad \psi_0(u) = -\frac{f_0'(u)}{f_0(u)}$$

Sin embargo Huber(1964) buscó un M-estimador que fuese óptimo (en un sentido que se precisará enseguida), en el entorno (para $0 < \varepsilon < 1$, fijo)

$$\mathcal{F}_{\theta\varepsilon} = \{F = (1-\varepsilon)F_\theta + \varepsilon G \quad \text{con } G \textbf{ simétrica}\} \supset P_\theta$$

notar que en este entorno, a diferencia del usual, la contaminación se exige que sea simétrica. El motivo es que de esta manera, cualquier $F \in \mathcal{F}_{\theta\varepsilon}$ será simétrica, y no habrá problema de sesgo.

Entonces, fijado ε, y elegida una $\psi(u)$, para cada $F \in \mathcal{F}_{\theta\varepsilon}$ la varianza asintótica del estimador será (modificando apropiadamente la 63)

$$V(\psi, F) = \sigma^2 \frac{E_F\left[\psi^2(U)\right]}{E_F^2\left[\psi'(U)\right]}$$

La idea de Huber fué obtener para cada ψ propuesta, la máxima varianza asintótica(variando $F \in \mathcal{F}_{\theta\varepsilon}$), y luego tomar como solución la ψ_H, que da la menor de estas varianzas, o sea

$$\psi_H = \arg \min_{\psi} \sup_{F \in \mathcal{F}_{\theta\varepsilon}} V(\psi, F)$$

la solución obtenida por Huber fué la ψ_H definida por

$$\psi_H(u) = \begin{cases} -k & \psi_0(u) < -k \\ \psi_0(u) & -k \le \psi_0(u) \le k \\ k & \psi_0(u) < k \end{cases} \tag{81}$$

donde k depende de F_0 y de la cantidad de contaminación ε. Notar que esta ψ_H tiene en la parte central, la ψ_0 que corresponde al estimador de MV, y fuera de la parte central es constante y acotada.

Entonces como para obtener el estimador habrá que resolver $\sum_{i=1}^{n} \psi_H(\frac{X_i - \widehat{\theta}}{\sigma}) = 0$, notar que las X_i cercanas a $\widehat{\theta}$, intervienen en esta expresión como $\psi_0(\frac{X_i-\widehat{\theta}}{\sigma})$, es decir usando la ψ_0 que corresponde a máxima verosimilitud; mientras que para observaciones más alejadas intervienen como la constante k.

En el caso de la normal, $F_0 = \Phi$, y como $\psi_0(u) = u$, resulta

$$\psi_H(u) = \begin{cases} -k & u < -k \\ u & |u| \le k \\ k & u < k \end{cases} \qquad \text{donde } 2\Phi(k) - 1 + \frac{2\varphi(k)}{k} = \frac{1}{1-\varepsilon} \tag{82}$$

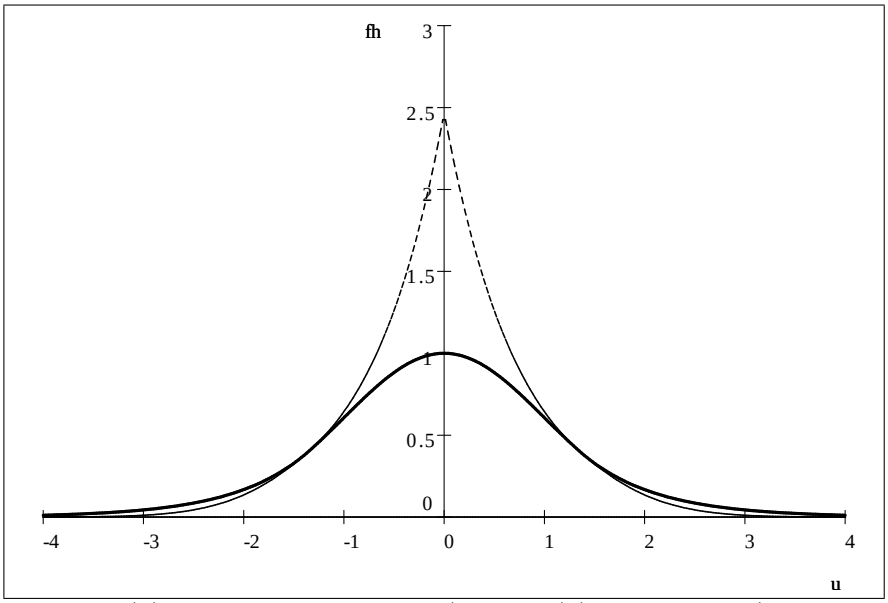
(donde por ejemplo para $\varepsilon = 0.05$ resulta $k = 1.398$).

Observar que si $\varepsilon \to 0(\mathcal{F}_{\theta\varepsilon} = P_\theta)$, resulta $k \to \infty$, esto es $\psi_H(u) \to u$ que corresponde a la media muestral $\widehat{\theta}_H = \overline{X}_n$; por otro lado, si $\varepsilon \to 1(\mathcal{F}_{\theta\varepsilon} =$ todas las F simétricas), resulta $k \to 0$, esto es $\psi_H(u) \to sgn(u)$ que corresponde a la mediana muestral $\widehat{\theta}_H = \widetilde{X}_n$. O sea cuanto mayor es la contaminación, más se parece el estimador de Huber a la mediana muestral.

Si para una $\psi_H(u)$ de un k dado, que corresponde al estimador $\widehat{\theta}_H$, se averigua cual es la densidad f_H de la cual $\widehat{\theta}_H$ sería el estimador de máxima verosimilitud, resulta

$$f_H(u) = \begin{cases} (1-\varepsilon)\varphi(u) & |u| \le k \\ \frac{(1-\varepsilon)}{\sqrt{2\pi}} e^{\frac{k^2}{2} - k|u|} & |u| > k \end{cases} \tag{83}$$

Esta densidad es la que Huber llama "menos favorable", ya que entre todas las de $\mathcal{F}_{\theta\varepsilon}$, es la que da mayor varianza($V(\psi_H, F_H)$). Visto con ojos más optimistas, usando el estimador basado en ψ_H, cualquiera sea $F \in \mathcal{F}_{\theta\varepsilon}$, la varianza del estimador será siempre $\le V(\psi_H, F_H)$(ya que es el $\sup remo$) . Pero si usamos un estimador basado en otra $\psi \ne \psi_H$, su varianza será $> V(\psi_H, F_H)$(puesto que es la *mínima*). A continuación se representa esta densidad



$f_0(x)$ correspondiente a MV, (para $\psi_H(u)$ con $k = 1.345$)

notar que, salvo una constante multiplicativa, para $|u| \le 1.345$ es una normal, y para $|u| > 1.345$, es una doble exponencial, que al tener colas más pesadas, es menos sensible a las observaciones muy desviadas.

Respecto de las propiedades del M-estimador basado en la ψ_H de Huber, se mencionan

1. como ψ_H es acotada, es B-robusto
2. es también cualitativamente robusto
3. para $F_0 = \Phi$, la sensibilidad a errores groseros es

$$\gamma^* = \frac{\sup_{x\in\Omega} |\psi_H(x-\theta)|}{E_{F_0}(\psi'_H(U))} = \frac{k}{\int_{-k}^{k} \varphi(u)du} = \frac{k}{2\Phi(k)-1}$$

4. el punto de ruptura asintótico es $\varepsilon^* = \frac{1}{2}$

Por último, este mismo problema pero con dispersión desconocida es analizado por Li y Zamar(1991).

5.3 Optimalidad en GES (γ^*)

Se buscará ahora el M-estimador que minimiza la sensibilidad a errores groseros γ^*. Se considerará el caso general, donde el parámetro de interés es cualquiera, no necesariamente uno de posición. Por lo tanto se trabajará con la función $\psi(x,\theta)$. Suponiendo un modelo paramétrico $P_\theta = \{F_\theta : \theta \in \Theta\}$, el estimador de máxima verosimilitud es el M-estimador con

$$\psi_0(x,\theta) = -\frac{f'_\theta(x)}{f_\theta(x)}$$

y para cualquier M-estimador la IF y la γ^* son

$$IF(x,T,F) = \frac{\psi(x;T(F))}{-E_F(\psi'(X;T(F)))} \quad \text{y} \quad \gamma^* = \gamma^*(T,F) = \sup_{x\in\Omega} |IF(x,T,F)| \tag{84}$$

la intención es hallar ψ que minimiza γ^*, pero lo haremos en la familia de todos los M-estimadores consistentes de Fisher, de esta manera nos aseguramos que todos estimen el parámetro correcto. Reemplazando entonces $T(F)$ por $T(F_\theta) = \theta$, y la notación T por ψ, la γ^* queda

$$\gamma^* = \gamma^*(\psi, F_\theta) = \frac{\sup_x |\psi(x;\theta)|}{\left|E_{F_\theta}(\psi'(X;\theta))\right|}$$

y el problema de optimización es encontrar $\widetilde{\psi}$ que minimize

$$\widetilde{\psi}(x;\theta) = \min_{\psi} \gamma^*(\psi, F_\theta)\Big|_{E_{F_\theta}(\psi(X;\theta))=0} \tag{85}$$

y se demuestra que la solución, es el M-estimador con función $\widetilde{\psi}$ donde

$$\widetilde{\psi}(x;\theta) = sgn\left[\psi_0(x,\theta) - Med_{F_\theta}(\psi_0(x,\theta))\right] \tag{86}$$

que es consistente de Fisher, y tiene el menor γ^*, en la familia de este tipo de estimadores.

Para obtener el M-estimador, habrá que resolver

$$\sum_{i=1}^{n}\widetilde{\psi}(X_i,\theta)=\sum_{i=1}^{n}sgn\left[\psi_0(X_i,\theta)-Med_{F_\theta}(\psi_0(X,\theta))\right]=0 \tag{87}$$

pero como $Med_{F_\theta}(\psi_0(X,\theta))$ es constante, que la suma de los $sgn()$ sea cero es equivalente a

$$Med\left\{\psi_0(X_1,\theta),\psi_0(X_2,\theta),\cdots,\psi_0(X_n,\theta)\right\}=Med_{F_\theta}\left[\psi_0(X,\theta)\right] \tag{88}$$

Con esta expresión es interesante señalar una analogía. Notar que el estimador de MV pensado como M-estimador surge de

$$\sum_{i=1}^{n}\psi_0(X_i,\theta)=0 \quad \text{o sea} \quad \frac{1}{n}\sum_{i=1}^{n}\psi_0(X_i,\theta)=E_{F_\theta}\left[\psi_0(X,\theta)\right]=0$$

ya que vale la consistencia de Fisher. Y esto también lo podemos expresar

$$Media\left\{\psi_0(X_1,\theta),\psi_0(X_2,\theta),\cdots,\psi_0(X_n,\theta)\right\}=E_{F_\theta}\left[\psi_0(X,\theta)\right] \tag{89}$$

si se compara (88) con (89), resulta que el estimador óptimo en γ^*, tiene la misma expresión que el de MV, salvo que se reemplaza la media (muestral y poblacional), por la mediana(usando la misma $\psi_0(X,\theta)$).

Example 33 *En el caso de $F_\theta=N(\theta,\sigma)$ con σ conocido, $\psi_0(x,\theta)=x-\theta$, y como $Med_{F_\theta}(x-\theta)=0$, resulta*

$$\widetilde{\psi}(x;\theta)=sgn(x-\theta)$$

que es la $\widetilde{\psi}$ que corresponde a la mediana muestral $\widetilde{X}_n$. Esto quiere decir que para la media, en el modelo paramétrico normal, la mediana muestral es el M-estimador con menor γ^*, entre todos los consistentes de Fisher. Y esto habla bién de la mediana, ya que también, por la "propiedad minimax", tiene menor sesgo máximo, entre todos los estimadores equivariantes de posición.

Example 34 *Considerese el caso en que $F_\sigma=N(0,\sigma)$. Ahora interesa como parámetro σ. Como $f_\sigma(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}$ el estimador de MV tendrá(descartando el factor $\frac{1}{\sigma}$) corresponde a*

$$\psi_0(x,\sigma)=-\frac{f'_\sigma(x)}{f_\sigma(x)}=\frac{1}{\sigma}(1-\frac{x^2}{\sigma^2})\approx(1-\frac{x^2}{\sigma^2})$$

Pero aquí corresponde una aclaración: esta $\psi_0(x,\sigma)$ es la que corresponde a MV en la teoría de M-estimadores en general, o sea, no es la ψ de un estimador de posición. Luego

$$\widetilde{\psi}(x;\sigma)=sgn\left[(1-\frac{x^2}{\sigma^2})-Med_{F_\sigma}(1-\frac{X^2}{\sigma^2})\right]$$

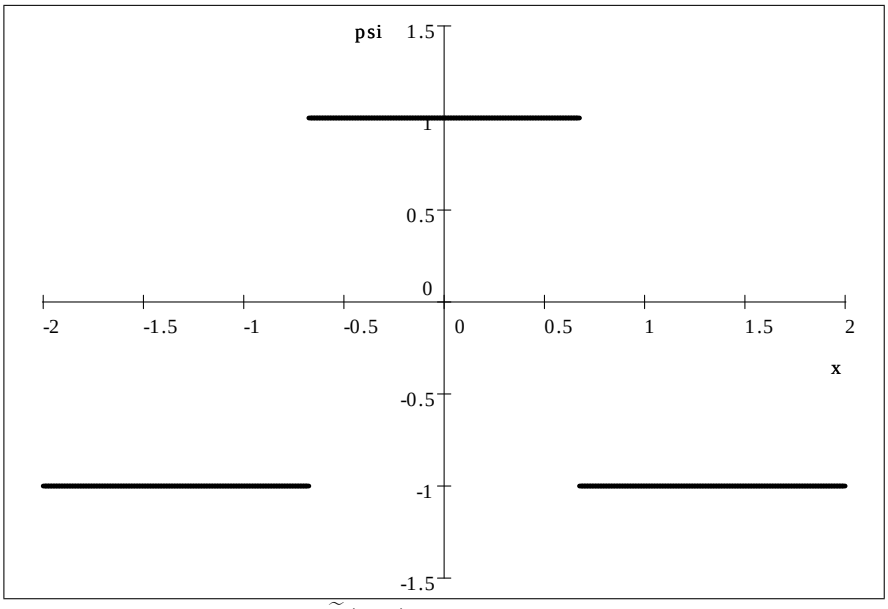
pero $Med_{F_\sigma}(1-\frac{X^2}{\sigma^2}) = 1-z^2_{0.75}$ *donde* $z_{0.75} = \Phi^{-1}(0.75) = 0.675$. *Reemplazando*

$$\widetilde{\psi}(x;\sigma) = sgn(z^2_{0.75} - \frac{x^2}{\sigma^2})$$

como el signo vale 1 *para* $|x| < z_{0.75}\sigma = 0.675\sigma$, *y,* -1 *en el resto, queda en definitiva*

$$\widetilde{\psi}(x;\sigma) = \begin{cases} 1 & para\ |x| < 0.675\sigma \\ -1 & para\ |x| \geq 0.675\sigma \end{cases}$$

cuyo gráfico es



$\widetilde{\psi}(x;\sigma)$ *para* $\sigma = 1$

y el estimador surge de resolver en σ

$$\sum_{i=1}^{n} \widetilde{\psi}(X_i,\sigma) = 0$$

y como para pas observaciones que estén en $|X_i| < 0.675\sigma$, $\widetilde{\psi}(X_i,\sigma) = 1$, *y para* $|X_i| \geq 0.675\sigma$, $\widetilde{\psi}(X_i,\sigma) = -1$, *el cambio de signo del miembro izquierdo se dará cuando* $2 * 0.675\sigma = ICR(X_1, X_2, \cdots X_n)$, *luego el estimador buscado es*

$$\widehat{\sigma} = \frac{ICR(X_1, X_2, \cdots X_n)}{2 * 0.675}$$

que es el rango intercuartil normalizado para la normal(la solución suele no ser única, sinó un intervalo). Y este estmador tendrá γ^* *mínima, en la familia de los consistentes de Fisher.*

Example 35 *Se analizarán para el ejemplo anterior algunos indicadores de robustez. La función de influencia de (84) es*

$$IF(x,\widetilde{\psi},F_\sigma)=\frac{\widetilde{\psi}(x;\sigma)}{-E_{F_\sigma}(\widetilde{\psi}'(X;\sigma))}$$

y como al derivar $\widetilde{\psi}(x;\sigma)$ *respecto de* σ *aparecen impulsos*

$$\widetilde{\psi}'(x;\sigma)=\lim_{\varepsilon\to 0}\frac{\widetilde{\psi}(x;\sigma+\varepsilon)-\widetilde{\psi}(x;\sigma)}{\varepsilon}=2\delta(x+0.675\sigma)+2\delta(x-0.675\sigma)$$

resultará

$$\begin{aligned}E_{F_\sigma}(\widetilde{\psi}'(X;\sigma)) &= \int\widetilde{\psi}'(x;\sigma)dF_\sigma\\ &= \int 2\delta(x+0.675\sigma)d\Phi(\frac{u}{\sigma})+\int 2\delta(x-0.675\sigma)d\Phi(\frac{u}{\sigma})\\ &= 2\varphi(\frac{-0.675\sigma}{\sigma})\frac{1}{\sigma}+2\varphi(\frac{0.675\sigma}{\sigma})\frac{1}{\sigma}=\frac{4}{\sigma}\varphi(0.675)\end{aligned}$$

de acuerdo a esto la sensibilidad a errores groseros es

$$\gamma^*=\frac{1}{\frac{4}{\sigma}\varphi(0.675)}=\frac{\sigma}{4\varphi(0.675)}=0.499\sigma$$

y la varianza asintótica, según (63)

$$V(\widetilde{\psi},F_\sigma)=\frac{E_{F_\sigma}\left[\widetilde{\psi}^2(X;\sigma)\right]}{E^2_{F_\sigma}(\widetilde{\psi}'(X;\sigma))}=\frac{1}{\left[\frac{4}{\sigma}\varphi(0.675)\right]^2}=0.249\sigma^2$$

5.4 Optimalidad en varianza y GES ($V(\psi,F)$ y γ^*)

Al principio de este capítulo se señalaron dos recomendaciones para el comportamiento de un estimador robusto en P_θ y $\mathcal{F}_{\theta\varepsilon}$:

- En P_θ es importante una baja varianza (o sea, alta eficiencia (ARE))
- En $\mathcal{F}_{\theta\varepsilon}$ es importante bajo sesgo

Lo ideal sería analizar este problema para un tamaño de entorno finito, por ejemplo $\varepsilon=0.10$, o 0.15. Sin embargo, para simplificar se presentará la solución de Hampel(1974) para un ε infinitesimal(es decir muy chico).

Cuando se definió la sensibilidad a la contaminación(2.8.1) $\gamma_c(\theta)=\frac{d}{d\varepsilon}\mathbf{MB}_{\widehat{\theta}}(\varepsilon,\theta)\big|_{\varepsilon=0}$, se demostró que para ε muy pequeño vale

$$\mathbf{MB}_{\widehat{\theta}}(\varepsilon,\theta)\approx\varepsilon\gamma_c(\theta)$$

y que aunque en general $\gamma^* \leq \gamma_c(\theta)$, para M-estimadores con ψ acotada, se cumple la igualdad $\gamma^* = \gamma_c(\theta)$, resultando

$$\mathbf{MB}_{\widehat{\theta}}(\varepsilon, \theta) \approx \varepsilon\gamma^*$$

Entonces según esto, minimizar o mantener acotado el sesgo equivale a minimizar o mantener acotado γ^*.

Se desarrollará la propuesta de Hampel para M-estimadores en general. Se supondrá un modelo paramétrico $P_\theta = \{F_\theta : \theta \in \Theta\}$, y como siempre la función ψ que corresponde al M-estimador de MV es

$$\psi_0(x, \theta) = -\frac{f'_\theta(x)}{f_\theta(x)}$$

Como interesará la varianza asintótica dentro de P_θ, se usará la notación $V(\psi, F_\theta)$, y para la sensibilidad a errores groseros usaremos $\gamma^*(\psi, F_\theta)$.

La propuesta de Huber es buscar un M-estimador con función $\psi^*(x, \theta)$ que verifique cualquiera de los siguientes dos criterios, ya que son equivalentes:

Propuesta I : **minimizar varianza**

$$\psi^*(x, \theta) = \arg \min_{\psi} V(\psi, F_\theta) \Big|_{\substack{E_{F_\theta}[\psi(X,\theta)]=0 \\ \gamma^*(\psi,F_\theta) \leq \gamma^*_{\max}(\theta)}}$$

es decir con $\psi = \psi^*$ se logra mínima varianza, entre todas las ψ que cumplen:

a $E_{F_\theta}[\psi(X, \theta)] = 0$ (consistencia de Fisher)

b $\gamma^*(\psi, F_\theta) \leq \gamma^*_{\max}(\theta)$ donde $\gamma^*_{\max}(\theta)$ es una cota que se impone a γ^*.

Propuesta II : **minimizar GES**

$$\psi^*(x, \theta) = \arg \min_{\psi} \gamma^*(\psi, F_\theta) \Big|_{\substack{E_{F_\theta}[\psi(X,\theta)]=0 \\ V(\psi,F_\theta)) \leq V_{\max}(\theta)}}$$

es decir con $\psi = \psi^*$ se logra mínimo γ^* todas las ψ que cumplen:

a $E_{F_\theta}[\psi(X, \theta)] = 0$ (consistencia de Fisher)

b $V(\psi, F_\theta)) \leq V_{\max}(\theta)$ donde $V_{\max}(\theta)$ es una cota que se impone a la varianza.

Son equivalentes en el siguiente sentido:

Equivalencia $I \rightarrow II$ Si se fija una cota $\gamma^*_{\max}(\theta)$ a la sensibilidad a errores groseros y se emplea la propuesta 1, se obtendrá una ψ^*, y una correspondiente varianza mínima $V(\psi^*, F_\theta)$; si a continuación se emplea la propuesta 2 usando como cota de varianza $V_{\max}(\theta) = V(\psi^*, F_\theta)$, se obtendrá igual ψ^*. De igual manera vale la

Equivalencia $II \rightarrow I$ Si se fija una cota $V_{\max}(\theta)$ a la varianza y se emplea la propuesta 2, se obtendrá una ψ^*, y una correspondiente γ^* mínima, $\gamma^*(\psi^*, F_\theta)$; si a continuación se emplea la propuesta 1 usando como cota de γ^*, $\gamma^*_{\max}(\theta) = \gamma^*(\psi^*, F_\theta)$, se obtendrá igual ψ^*.

La solución dada por Hampel a ambos problemas es

$$\psi^*(x, \theta) = \psi_{k(\theta)} \left[\psi_0(x, \theta) - r(\theta)\right] \tag{90}$$

donde $\psi_{k(\theta)}$ es la ψ de Huber, y tanto $k(\theta)$ como $r(\theta)$ deben ajustarse para que la solución $\psi^*(x, \theta)$ verifique la consistencia de Fisher $E_{F_\theta}\left[\psi^*(x, \theta)\right] = 0$.

Si se presta atención a la (90), esencialmente la solución es la $\psi_0(x, \theta)$ que corresponde a máxima verosimilitud, centrandola en $r(\theta)$, y acotandola a través de $\psi_{k(\theta)}$.

Remark 36 *Si en la propuesta 1, se toma $\gamma^*_{\max}(\theta) = \infty$, que equivale a no imponer restricción sobre γ^*, la solución es $\psi^*(x, \theta) = \psi_0(x, \theta)$ que corresponde al estimador de MV, que es no robusto. Por otro lado, si en la propuesta 2 se toma $V_{\max}(\theta) = \infty$, solo se minimimizará γ^* bajo consistencia de Fisher, obteniendo $\psi^*(x, \theta) = \widetilde{\psi}(x; \theta)$ que es la solución vista en Optimalidad en GES(5.3).*

Combinando este Remark y las Equivalencias $I \rightarrow II$ y $II \rightarrow I$, resultan las siguientes restricciones en la elección de las cotas $\gamma^*_{\max}(\theta)$ y $V_{\max}(\theta)$ para que el problema tenga solución

$$\gamma^*_{\max}(\theta) \geq \gamma^*(\widetilde{\psi}, F_\theta) \quad \text{y} \quad V_{\max}(\theta) \geq V(\psi_0, F_\theta)$$

es decir $\gamma^*_{\max}(\theta)$ tiene que ser $\geq$ que el GES mínimo obtenido bajo Optimalidad en GES, y $V_{\max}(\theta)$ debe ser $\geq$ que la varianza que corresponde al estimador de MV.

Pero, no obstante la equivalencia, desde un punto de vista práctico es más razonable fijar una cota a la varianza ($V_{\max}(\theta)$) en la propuesta 2, que imponer un $\gamma^*_{\max}(\theta)$ en la propuesta 1. Con este fin, si ARE es la eficiencia asintótica del estimador obtenido, respecto del de MV, o sea

$$ARE = \frac{V(\psi_0, F_\theta)}{V(\psi^*, F_\theta)}$$

entonces si la eficiencia propuesta es ARE_{prop}, deberá ser $\frac{V(\psi_0, F_\theta)}{V(\psi^*, F_\theta)} \geq ARE_{prop}$, o sea $V(\psi^*, F_\theta) \leq \frac{V(\psi_0, F_\theta)}{ARE_{prop}}$ entonces la sugerencia es tomar como cota

$$V_{\max}(\theta) = \frac{V(\psi_0, F_\theta)}{ARE_{prop}} \tag{91}$$

Hallar $k(\theta)$ y $r(\theta)$ en el caso general suele ser complicado, así que se analizará un caso particular.

Solución para el modelo de posición En este caso $P_\theta = \left\{F_\theta(x) = F_0(\frac{x-\theta}{\sigma})\right\}$ donde σ es conocida, y para el estimador de MV

$$\psi_0(x,\theta) = -\frac{f'_\theta(x)}{f_\theta(x)} = -\frac{f'_0(\frac{x-\theta}{\sigma})\frac{-1}{\sigma^2}}{f_0(\frac{x-\theta}{\sigma})\frac{1}{\sigma}} = \frac{1}{\sigma}\frac{f'_0(\frac{x-\theta}{\sigma})}{f_0(\frac{x-\theta}{\sigma})} = \frac{1}{\sigma}\xi_0(\frac{x-\theta}{\sigma}) \quad \text{con } \xi_0(u) = \frac{f'_0(u)}{f_0(u)}$$

la función de influencia para una ψ cualquiera pero consistente de Fisher, en F_θ es

$$IF(x,\psi,F_\theta) = \frac{\psi(x;\theta)}{-E_{F_\theta}(\psi'(X;\theta))}$$

y para ψ_0 de MV

$$IF(x,\psi_0,F_\theta) = \frac{\psi_0(x;\theta)}{-E_{F_\theta}(\psi'_0(X;\theta))} = \frac{\sigma\xi_0(\frac{x-\theta}{\sigma})}{E_{F_0}\left[\xi'_0(U)\right]}$$

Como necesitamos la cota $V_{\max}(\theta)$, se obtendrá la varianza del estimador de MV

$$V(\psi_0,F_\theta) = E_{F_\theta}\left[IF^2(x,\psi_0,F_\theta)\right] = \sigma^2\frac{E_{F_0}\left[\xi_0^2(U)\right]}{E^2_{F_0}\left[\xi'_0(U)\right]}$$

que no depende de θ, se puede calcular, y entonces con la eficiencia deseada, de (91) resulta la cota $V_{\max}(\theta) = V_{\max}$.

La solución es

$$\psi^*(x,\theta) = \psi_{k(\theta)}\left[\frac{1}{\sigma}\xi_0(\frac{x-\theta}{\sigma}) - r(\theta)\right]$$

donde hay que determinar $k(\theta)$ y $r(\theta)$. Pero como $\psi^*(x,\theta)$ debe verificar la consistencia de Fisher queda la primera relación

$$E_{F_\theta}\left[\psi^*(x,\theta)\right] = E_{F_0}\left[\psi_{k(\theta)}\left[\frac{1}{\sigma}\xi_0(U) - r(\theta)\right]\right] = 0 \tag{92}$$

De la equivalencia $I \to II$, cuando en la propuesta 1 obtenemos la varianza mínima $V(\psi^*,F_\theta)$, si la igualamos a $V_{\max}(\theta)$, obtendremos la misma solución. Luego la otra relación es

$$V_{\max}(\theta) = V(\psi^*,F_\theta) \tag{93}$$

la función de influencia es

$$IF(x,\psi^*,F_\theta) = \frac{\psi^*(x,\theta)}{-E_{F_\theta}(\psi^{*\prime}(x,\theta))} = \frac{\psi^*(x;\theta)}{\int\psi^*(x;\theta)\frac{\partial}{\partial\theta}f_\theta(x)dx}$$

pero para evitar derivar la $\psi^*(x,\theta)$ respecto de θ, usaremos la segunda de estas expresiones(vista en 61) y queda

$$IF(x,\psi^*,F_\theta) = \frac{\sigma\psi_{k(\theta)}\left[\frac{1}{\sigma}\xi_0(\frac{x-\theta}{\sigma}) - r(\theta)\right]}{-\int\psi_{k(\theta)}\left[\frac{1}{\sigma}\xi_0(u) - r(\theta)\right]f'_0(u)du}$$

y la varianza buscada

$$V(\psi^*, F_\theta) = E_{F_\theta}\left[IF^2(X, \psi^*, F_\theta)\right] = \sigma^2 \frac{E_{F_0}\left[\psi^2_{k(\theta)}\left[\frac{1}{\sigma}\xi_0(U) - r(\theta)\right]\right]}{\left[\int \psi_{k(\theta)}\left[\frac{1}{\sigma}\xi_0(u) - r(\theta)\right] f_0'(u)du\right]^2}$$

finalmente de (92) y (93) el sistema de ecuaciones a resolver es(donde se puso k y r, en lugar de $k(\theta)$ y $r(\theta)$, ya que el sistema no dependen de θ)

$$\begin{cases} E_{F_0}\left[\psi_k\left[\frac{1}{\sigma}\xi_0(U) - r\right]\right] = 0 \\ \sigma^2 \frac{E_{F_0}\left[\psi^2_k\left[\frac{1}{\sigma}\xi_0(U) - r\right]\right]}{\left[\int \psi_k\left[\frac{1}{\sigma}\xi_0(u) - r\right] f_0'(u)du\right]^2} = V_{\max}(\theta) = V_{\max} \end{cases}$$

(y en el caso que f_0 sea simétrica, de la primera resulta que $r = 0$).

Example 37 *En el caso de una $F_\theta = N(\theta, \sigma)$ con σ conocido, $F_0 = N(0, 1)$ y entonces $\xi_0(u) = \frac{f_0'(u)}{f_0(u)} = -u$. Además por ser simétrica $r = 0$ y entonces queda solo la segunda ecuación.*

$$\sigma^2 \frac{E_{F_0}\left[\psi^2_k\left[\frac{1}{\sigma}\xi_0(U)\right]\right]}{\left[\int \psi_k\left[\frac{1}{\sigma}\xi_0(u)\right] f_0'(u)du\right]^2} = V_{\max}$$

o sea

$$\sigma^2 \frac{E_{F_0}\left[\psi^2_k[U]\right]}{\left[\int \psi_k(u) f_0'(u)du\right]^2} = V_{\max}$$

pero el primer término es la varianza asintótica del M-estimador con la ψ_k de Huber, así que el k surge de la varianza asintótica deseada. En el próximo capítulo se analizará con más detalle esta varianza.

6 Algunos ejemplos

6.1 Estudio del MAD

Se quiere estudiar el comportamiento del $MAD(\mathbf{X}) = \underset{1\le i\le n}{Med}\{|X_i - Med(\mathbf{X})|\}$. Pero como este es un estimador que incluye a la mediana, se planteará un modelo de posición-escala simultaneo. Se supondrá $X_i = \theta + \sigma U_i$ donde $U_i \sim F_0$ conocida, σ desconocido, y el modelo parametrico

$$X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} F \quad \text{con} \quad F \in P_{\theta,\sigma} = \left\{F_{\theta,\sigma} : F_{\theta,\sigma}(x) = F_0(\frac{x-\theta}{\sigma})\right\}$$

Como queremos que el estimador de posición sea la mediana se elegirá $\psi(u) = sgn(u)$, y para la escala tomaremos $\chi(u) = sgn(\left|\frac{u}{c}\right| - 1)$ (ver ejemplo-30), teniendo la libertad de ajustar c para, por ejemplo, que el estimador de escala estime justo el desvío estándar en el caso de una normal(MADN).

Si se llama $T(F)$ y $S(F)$ a los correspondientes funcionales, estos quedan definidos por

$$E_F\left[sgn(\frac{X-T(F)}{S(F)})\right]=0 \quad \text{y} \quad E_F\left[sgn(\left|\frac{X-T(F)}{cS(F)}\right|-1)\right]=0 \qquad (94)$$

y resolviendo surge que $T(F)=Med(F)=F^{-1}(\frac{1}{2})$, y $S(F)=\frac{MAD(\mathbf{X})}{c}$. Además según Huber(pág 137), en el caso que F sea simétrica, las funciones de influencia son

$$IF(x,T,F)=\frac{sgn(x-T(F))}{2f(T(F))} \quad \text{y} \quad IF(x,S,F)=\frac{sgn\left[\left|\frac{x-T(F)}{cS(F)}\right|-1\right]}{4cf(T(F)+cS(F))}$$

Con la primera, dentro del modelo paramétrico con $F=F_{\theta,\sigma}$, como $T(F_{\theta,\sigma})=\theta$, y $f_{\theta,\sigma}(T(F_{\theta,\sigma}))=\varphi(\theta)=\frac{1}{\sqrt{2\pi}\sigma}$, resulta $IF(x,T,F_{\theta,\sigma})=sgn(x-\theta)\sqrt{\frac{\pi}{2}}\sigma$ que es la función de influencia de la mediana, en una población normal.

Antes de abordar la segunda, también para $F=F_{\theta,\sigma}$, busquemos el valor de c para que el estimador sea el $MADN$. Como en la segunda de las (94), $T(F_{\theta,\sigma})=\theta$ y queremos que $S(F_{\theta,\sigma})=\sigma$, deberá ser

$$E_{F_{\theta,\sigma}}\left[sgn(\left|\frac{X-\theta}{c\sigma}\right|-1)\right]=E_{F_0}\left[sgn(\left|\frac{U}{c}\right|-1)\right]=0$$

y vimos del ejemplo-30, que esto se cumple para $c=\Phi^{-1}(\frac{3}{4})=0.675$. Entonces para la normal, como $f_{\theta,\sigma}(T(F)+cS(F))=f_{\theta,\sigma}(\theta+c\sigma)$ será

$$IF(x,S,F_{\theta,\sigma})=\frac{sgn\left[\left|\frac{x-\theta}{c\sigma}\right|-1\right]}{4cf_{\theta,\sigma}(\theta+c\sigma)}=\frac{\sigma}{4c\varphi(c)}sgn\left[\left|\frac{x-\theta}{c\sigma}\right|-1\right]$$

De aquí sale que $\gamma^*=(4c\varphi(c))^{-1}\sigma=1.167\sigma$. Y el punto de ruptura es

$$\varepsilon^*=\frac{-\chi(0)}{\chi(\infty)-\chi(0)}=\frac{1}{1+1}=\frac{1}{2}$$

Para la varianza asintótica hay que calcular

$$V(T,F_{\theta,\sigma})=E_{F_{\theta,\sigma}}\left[IF^2(X,S,F_{\theta,\sigma})\right]=\frac{\sigma^2}{\left[4c\varphi(c)\right]^2}=1.361\sigma^2$$

Si interesa la eficiencia asintótica, habrá que calcular la varianza asintótica del estimador de MV. Como $\frac{(n-1)S^2}{\sigma^2}\sim\chi^2_{n-1}$, la $Var(S^2)=\frac{2\sigma^4}{n-1}$ o sea $V(S^2,F_{\theta,\sigma})=2\sigma^4$; pero como $S=\left[S^2\right]^{1/2}$ aproximando por la primera derivada, se obtiene $V(S,F_{\theta,\sigma})=\frac{\sigma^2}{2}$ y luego

$$ARE_{MADN,S}=\frac{V(S,F_{\theta,\sigma})}{V(MADN,F_{\theta,\sigma})}=\frac{\sigma^2/2}{\sigma^2/\left[4c\varphi(c)\right]^2}=\frac{\left[4c\varphi(c)\right]^2}{2}=0.367$$

que en realidad es muy baja, sin embargo cuando se usa como estimador previo de dispersión en el M-estimador de posición, no interesa su eficiencia, sinó sus características de robustez.

6.2 Eficiencia asintótica con la ψ de Huber

Suponiendo un modelo parametrico normal, usando estimador de posición basado en la

$$\psi_{H,k}(u) = \psi_k(u) = \begin{cases} -k & u < -k \\ u & |u| \leq k \\ k & u < k \end{cases}$$

y estimación previa de dispersión $MADN$, se tiene de (80)

$$IF(x, T, F_{\theta,\sigma}) = \frac{\sigma\psi_k(\frac{x-\theta}{\sigma})}{E_{F_{\theta,\sigma}}(\psi_k'(\frac{X-\theta}{\sigma}))} = \frac{\sigma\psi_k(\frac{x-\theta}{\sigma})}{E_{F_0}(\psi_k'(U))}$$

siendo la varianza asintótica

$$V(\widehat{\theta}_{Hk}, F_{\theta,\sigma}) = E_{F_{\theta,\sigma}}\left[IF^2(X, T, F_{\theta,\sigma})\right] = \sigma^2 \frac{E_{F_0}(\psi_k^2(U))}{E_{F_0}^2(\psi_k'(U))}$$

donde en el denominador:

$$E_{F_0}(\psi_k'(U)) = E_{F_0}(I(|U| \leq k)) = \Phi(k) - \Phi(-k) = 2\Phi(k) - 1$$

y en el numerador:

$$E_{F_0}(\psi_k^2(U)) = k^2\Phi(-k) + \int_{-k}^{+k} u^2\varphi(u)du + k^2(1 - \Phi(k))$$

pero, por partes, $\int_{-k}^{+k} u^2\varphi(u)du = -2k\varphi(k) + 2\Phi(k) - 1$, y reemplazando, y operando

$$E_{F_0}(\psi_k^2(U)) = 2k^2(1 - \Phi(k)) - 2k\varphi(k) + 2\Phi(k) - 1$$

luego la varianza asintótica es

$$V(\widehat{\theta}_{Hk}, F_{\theta,\sigma}) = \sigma^2 \frac{2k^2(1 - \Phi(k)) - 2k\varphi(k) + 2\Phi(k) - 1}{[2\Phi(k) - 1]^2}$$

como con el estimador de MV $(\overline{X})$, es $V(\overline{X}, F_{\theta,\sigma}) = \sigma^2$, resulta

$$ARE_{\widehat{\theta}_{Hk},\overline{X}} = \frac{V(\overline{X}, F_{\theta,\sigma})}{V(\widehat{\theta}_{Hk}, F_{\theta,\sigma})} = \frac{[2\Phi(k) - 1]^2}{2k^2(1 - \Phi(k)) - 2k\varphi(k) + 2\Phi(k) - 1}$$

A continuación se presenta una tabla, para diferentes k

k	$V(\psi_k\ Huber)/\sigma^2$	ARE
0	1.570	0.637
0.5	1.263	0.792
0.985	1.110	**0.90**
1	1.107	0.903
1.345	1.053	**0.95**
2.01	1.010	**0.99**
2.5	1.002	0.997
∞	1	1

donde para $k = 0$, el estimador es la mediana, con eficiencia 63.7%; y para $k \to \infty$, el estimador es la media muestral, con eficiencia 100%.

Remark 38 *La varianza asintótica obtenida, $V(\widehat{\theta}_{Hk}, F_{\theta,\sigma})$, es para $F = F_{\theta,\sigma} \in P_\theta$, ($\sigma$ es una estimación previa), ya que en P_θ al no haber problema de sesgo, nos va a interesar que el estimador tenga baja varianza o alta eficiencia. También comentamos que para $F \in \mathcal{F}_{\theta\varepsilon}$, además de la varianza del estimador aparece un sesgo. Y como para n grande, la varianza se divide por n, y el sesgo no, en este entorno $\mathcal{F}_{\theta\varepsilon}$ debemos preocuparnos fundamentalmente por el sesgo. Esto lo trataremos en (6.4).*

Sin embargo, con fines ilustrativos, analizaremos la varianza del estimador para un entorno particular $\mathcal{F}_{\theta\varepsilon} \supset P_\theta$. Se recordará parcialmente la propiedad de optimalidad de la $\psi_{H,k}$ de Huber, suponiendo aquí que σ es conocida, ya que es un requisito en el desarrollo de Huber.

Dado k, y definido el entorno

$$\mathcal{F}_{\theta\varepsilon} = \{F = (1-\varepsilon)F_\theta + \varepsilon G \ \text{ con } G \textbf{ simétrica}\} \supset P_\theta$$

donde el tamaño ε del entorno surge de $2\Phi(k) - 1 + \frac{2\varphi(k)}{k} = \frac{1}{1-\varepsilon}$, entonces, para F en este entorno, la varianza del estimador $\widehat{\theta}_{Hk}$ será $\leq$ que la que corresponde a la de la densidad f_H "menos favorable" de Huber(83). Se calculará entonces $V(\widehat{\theta}_{Hk}, F_{H,\theta,\sigma})$, que representa una cota superior de la varianza, pero por supuesto en este entorno, (con la restricción de que su tamaño depende de k, y que incluye solo densidades simétricas). Como

$$IF(x, T, F_{H,\theta,\sigma}) = \frac{\sigma\psi_k(\frac{x-\theta}{\sigma})}{E_{F_H}(\psi_k'(U))} \ \text{ resulta } V(\widehat{\theta}_{Hk}, F_{H,\theta,\sigma} = \sigma^2 \frac{E_{F_H}\left[\psi_k^2(U)\right]}{E_{F_H}^2\left[\psi_k'(U)\right]}$$

y entonces(después de algunas cuentas)

$$V(\widehat{\theta}_{Hk}, F_{H,\theta,\sigma} = \sigma^2 \frac{(2\Phi(k) - 1) + \frac{2\varphi(k)}{k}}{2\Phi(k) - 1}$$

y se tiene la tabla

k	ARE	$Var\widehat{\theta}_{Hk}/\sigma^2$ en P_θ	ε	$Var\widehat{\theta}_{Hk}Max/\sigma^2$ en $\mathcal{F}_{\theta\varepsilon}$
0	0.637	1.57	1	∞
0.985	0.90	1.11	0.148	1.74
1.345	0.95	1.053	0.058	1.29
2.01	0.99	1.01	0.008	1.055
∞	1	1	0	1

donde por ejemplo si $k = 1.345$ que corresponde a eficiencia 0.95 en P_θ, en el entorno $\mathcal{F}_{\theta\varepsilon}$ de tamaño $\varepsilon = 0.058$, la varianza máxima es un poco mayor, 1.29. El problema de este análisis es que el tamaño del entorno varía con k. Por ejemplo para $k = 2.01$, la varianza máxima es 1.055, pero en un entorno muy chico($\varepsilon = 0.008$). De todas maneras muestra que la varianza del estimador de Huber no se incrementa mucho(por supuesto, suponiendo entorno de distribuciones simétricas, y σ conocida).

6.3 Eficiencia asintótica con la ψ bicuadrada de Tukey

A continuación se repetirá el análisis anterior pero usando para el estimador de posición la

$$\psi_{B,k}(u) = \psi_k(u) = \begin{cases} u\left[1-(\frac{u}{k})^2\right]^2 & |u| \le k \\ 0 & |u| > k \end{cases}$$

y estimación previa de dispersión con $MADN$. También aquí resulta

$$IF(x, T, F_{\theta,\sigma}) = \frac{\sigma\psi_k(\frac{x-\theta}{\sigma})}{E_{F_0}(\psi_k'(U))} \quad \text{y} \quad V(\widehat{\theta}_{Hk}, F_{\theta,\sigma}) = \sigma^2 \frac{E_{F_0}(\psi_k^2(U))}{E_{F_0}^2(\psi_k'(U))}$$

donde

$$V(\widehat{\theta}_{Bk}, F_{\theta,\sigma}) = \sigma^2 \frac{\int_{-k}^{+k} u^2\left[1-(\frac{u}{k})^2\right]^4 \varphi(u)du}{\left[\int_{-k}^{+k}(1-6\frac{u^2}{k^2}+5\frac{u^4}{k^4})\varphi(u)du\right]^2}$$

Pero notar que tanto en el numerador como en el numerador hay que calcular integrales del tipo $I(m) = \int_{-k}^{+k} u^{2m}\varphi(u)du$ para $m = 0, 1, 2, .5$. La idea es definir $H(a) = \int_{-k}^{+k} \frac{1}{\sqrt{2\pi}} e^{-a\frac{u^2}{2}} du = \frac{2\Phi(k\sqrt{a})-1}{\sqrt{a}}$, y darse cuenta que $I(m) = (-2)^m \left.\frac{d^m H(a)}{da^m}\right|_{a=1}$. Pero como las expresiones son igualmente complicadas, se calculará la varianza numericamente con el S-PLUS. Como con el estimador de MV ($\overline{X}$), es $V(\overline{X}, F_{\theta,\sigma}) = \sigma^2$, resulta también

$$ARE_{\widehat{\theta}_{Bk},\overline{X}} = \frac{V(\overline{X}, F_{\theta,\sigma})}{V(\widehat{\theta}_{Bk}, F_{\theta,\sigma})} = \frac{\left[\int_{-k}^{+k}(1-6\frac{u^2}{k^2}+5\frac{u^4}{k^4})\varphi(u)du\right]^2}{\int_{-k}^{+k} u^2\left[1-(\frac{u}{k})^2\right]^4\varphi(u)du}$$

A continuación se presenta una tabla, para diferentes k

k	$V(\psi_k\ Tukey)/\sigma^2$	ARE
$\to 0$	$\to \infty$	$\to 0$
1	9.848	0.101
2	2.14	0.466
3.14	1.25	**0.80**
3.44	1.177	**0.85**
3.88	1.111	**0.90**
4.68	1.053	**0.95**
7.04	1.010	**0.99**
∞	1	1

y los comandos de SPLUS correspondientes:

```
# Var asintotica con ψ de Tukey
k<-3.88
num<-function(u)  {((u*(1-(u/k)**2)**2)**2)*dnorm(u)}
den<-function(u)  {(1-6*(u/k)**2+5*(u/k)**4)*dnorm(u)}
a<-integrate(num,lower=-k,upper=k)$integral
b<-integrate(den,lower=-k,upper=k)$integral
v<-a/(b**2)
v;1/v
```

6.4 MB con la ψ de Huber

Suponiendo un modelo parametrico normal, usando estimador de posición basado en la

$$\psi_{H,k}(u) = \psi_k(u) = \begin{cases} -k & u < -k \\ u & |u| \leq k \\ k & u < k \end{cases}$$

y estimación previa de dispersión, por ejemplo $MADN$, se quiere estudiar el sesgo asintótico máximo. Como $F_0 = N(0,1)$ es simétrica, ψ_k no-decreciente y acotada, y $\psi_k(\infty) = k$, entonces según la (67, poniendo b/σ en lugar de b, para tener en cuenta el desvío), $\mathbf{MB}(\boldsymbol{\varepsilon})$ $=b$ es la solución de

$$E_{F_0}\left[\psi_k(U - \frac{b}{\sigma})\right] = \frac{-\varepsilon}{1-\varepsilon}k$$

sin embargo, por simplicidad, resolveremos $E_{F_0}\left[\psi(U-b)\right] = \frac{-\varepsilon}{1-\varepsilon}k$, acordándonos de expresar finalmente $\mathbf{MB}(\boldsymbol{\varepsilon})$ $=b\sigma$

$$(-k)\Phi(b-k) + \int_{b-k\sigma}^{b+k\sigma} \psi_k(u-b)\varphi(u)du + k(1-\Phi(b+k)) = \frac{-\varepsilon}{1-\varepsilon}k$$

operando resulta

$$[(b+k)\Phi(b+k) + \varphi(b+k)] - [(b-k)\Phi(b-k) + \varphi(b-k)] - \frac{k}{1-\varepsilon} = 0$$

donde se observa que el sesgo máximo, b, depende de ε y de k. Se presenta una función de SPLUS para calcularlo

```
# MB con la ψ_k de Huber y la Normal
mbhuber<-function(e,k)
{h<-function(b,e,k)
     {(b+k)*pnorm(b+k)+dnorm(b+k)-(b-k)*pnorm(b-k)-dnorm(b-k)-k/(1-e)}
uniroot(h,lower=0,upper=10,keep.xy=T,e=e,k=k)$root}
```

A continuación se presenta una tabla donde figura MB/σ, para diferentes k y contaminaciones ε, agregando también una columna con las eficiencias asintóticas

MB/σ^2 con la ψ_k de Huber

ARE	$k \downarrow \varepsilon \rightarrow$	**0.05**	**0.10**	**0.20**
0.792	**0.5**	0.069	0.146	0.332
0.90	**0.985**	0.077	0.163	0.371
0.95	**1.345**	0.086	0.182	0.416
0.99	**2.01**	0.111	0.234	0.532
0.997	**2.5**	0.133	0.282	0.637

Se nota el compromiso entre la ARE y el MB/σ, ya que al aumentar k, aumenta la eficiencia asintótica dentro de $P_{\theta,\sigma}$, pero aumenta también el sesgo máximo en el entorno de contaminación $\mathcal{F}_\varepsilon \supset P_{\theta,\sigma}$.

6.5 MB con la ψ Bicuadrada de Tukey

A diferencia del caso anterior como la ψ_k de Tuckey no es, no-decreciente, no se podrá usar una fórmula para hallar el MB. Procederemos usando la definición, suponiendo para simplificar que $\sigma = 1$, ya que si el sesgo obtenido es b, cuando $\sigma = 1$, para $\sigma \neq 1$ el sesgo será $MB(\varepsilon) = b\sigma$. Considerando el entorno

$$\mathcal{F}_{\theta\varepsilon} = \{F \in \mathcal{F} : F = (1-\varepsilon)F_\theta + \varepsilon G \ \text{ con } G \in \mathcal{F} \text{ arbitraria}\}$$

El método a seguir es el siguiente: se fija un $\varepsilon < \varepsilon^*$ quedando definido un $\mathcal{F}_{\theta\varepsilon}$; y entonces para cada G queda determinada una $F \in \mathcal{F}_{\theta\varepsilon}$, y con ella se calcula el valor asintótico $t = \widehat{\theta}_\infty(F)$ mediante

$$E_F\left[\psi_k(X-t)\right] = 0$$

luego el sesgo en valor absoluto será $|t-\theta|$. La idea es variando G, lograr el el máximo de este indicador.

Desarrollando la expresión anterior y poniendo $\psi_k(X-t) = \psi_k(X-\theta+\theta-t) = \psi_k(X-\theta-b)$ se tiene

$$\begin{aligned}(1-\varepsilon)E_{F_\theta}\left[\psi_k(X-\theta-b)\right] + \varepsilon E_G\left[\psi_k(X-\theta-b)\right] &= 0\\ E_{F_0}\left[\psi_k(U-b)\right] + \frac{\varepsilon}{(1-\varepsilon)}E_G\left[\psi_k(X-\theta-b)\right] &= 0\end{aligned}$$

Ahora se tomará $G = \Delta_{y+\theta}$ y entonces queda(y mide la ubicación de la contaminación, medida respecto de θ)

$$E_{F_0}\left[\psi_k(U-b)\right] + \frac{\varepsilon}{(1-\varepsilon)}\psi_k(y-b) = 0$$

Notar que para cada y se obtiene un sesgo $b(y)$ que cumple:

1. $b(0) = 0$
2. si $y \geq k$, resulta $b(y) = 0$ ya que la ψ_k es nula arriba de k

3. $b(-y) = -b(y)$

Esto quiere decir que el sesgo máximo se dará para $0 \leq y \leq k$, y es en este intervalo que tendremos que maximizar. Se presenta un programa en SPLUS.

```
# MB con la psi de Tukey y la Normal
k<-4.68
e<-0.05
f<-function(u,b){psibsq(u-b,k)*dnorm(u)}
int<-function(b){integrate(f,lower=b-k,upper=b+k,b=b)$integral}
g<-function(b,y){int(b)+(e/(1-e))*psibsq(y-b,k)}
sesgo<-function(y){uniroot(g,c(0,y),y=y)$root}
optimize(sesgo,interval=c(0,k),max=T)$objective
```
(95)

A continuación se presenta una tabla donde figura MB/σ, para diferentes k y contaminaciones ε, agregando también una columna con las eficiencias asintóticas

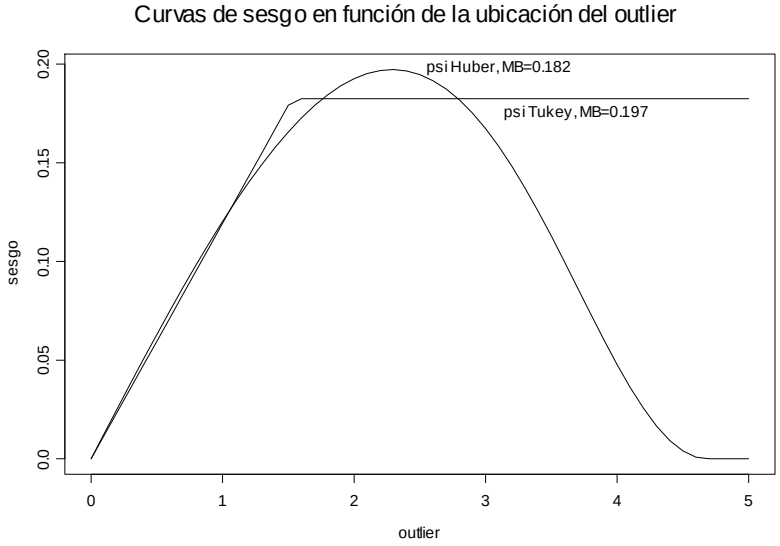
MB/σ^2 con la ψ_k de Tukey

ARE	$k \downarrow \varepsilon \rightarrow$	**0.05**	**0.10**	**0.20**
0.80	**3.14**	0.087	0.185	0.428
0.85	**3.44**	0.087	0.183	0.423
0.90	**3.88**	0.088	0.186	0.426
0.95	**4.68**	0.093	0.197	0.451
0.99	**7.04**	0.120	0.254	0.577

Si se compara esta tabla con la de la ψ_k de Huber, se nota que también al aumentar la eficiencia, aumenta el sesgo máximo, pero ahora no tanto. Y en ambas el MB aumenta casi proporcionalmente a la contaminación. Comparando las dos para eficiencia 0.95, la de Huber resulta ligeramente superior ya que los sesgos son 0.086, 0.182, 0.416, versus, 0.093, 0.197, 0.451 para la de Tukey.

6.6 Balance entre robustez y eficiencia

Sin embargo analizaremos con más cuidado esta última afirmación. Para eficiencia 0.95 y contaminación 0.10, los sesgos máximos son $MB_{Huber} = 0.182$ y $MB_{Tukey} = 0.197$. Pero ahora se graficarán las curvas de sesgo en función de la ubicación y de la contaminación respecto de θ. La función sesgo(y) fué definida en los comandos de SPLUS para la ψ de Tukey(95), pudiéndose definir una similar para la de Huber. Graficando resultan las curvas



donde se observa de que a pesar que con la ψ_{Tukey} es mayor el sesgo máximo(0.197 v.s 0.182 en la de Huber), el sesgo es menor en casi toda la banda de outliers(incluso es cero para outliers que se alejen de θ más de 4.68σ).

Según lo analizado hasta ahora, con un modelo paramétrico normal P_θ, y si se consideran dos estimadores, uno basado en la ψ_{Huber} con $k = 1.345$(para eficiencia 0.95), y el otro con la ψ_{Tukey} con $k = 4.68$ (también para eficiencia 0.95) resultan en cuanto a:

- **Eficiencia** (en P_θ): similares, ya que ambos tienen eficiencia 0.95(por diseño)
- **Robustez** (en $\mathcal{F}_\varepsilon$):

1. Punto de ruptura: similares, ya para ambos vale $\frac{1}{2}$(usando el $MADN$ como estimador de dispersión.
2. Sesgo: para $\varepsilon = 0.10$ el sesgo máximo es levemente menor con la ψ_{Huber}, sin embargo con la ψ_{Tukey} el sesgo es menor para una amplia banda de outliers, siendo cero para grandes outliers.

Con el fin de "desempatar", se tomará una $F \in \mathcal{F}_\varepsilon$, concretamente una distribución de Cauchy, y se analizará la varianza asintótica de estos dos estimadores. Como la Cauchy es simétrica, no habrá problema de sesgo.

Sea entonces $f_{\theta,\gamma}(x) = \frac{1}{\pi\gamma(1+(\frac{x-\theta}{\gamma})^2)}$, donde γ es el parámetro de escala de la Cauchy. Como usaremos el $MADN$ como estimador previo de escala, se necesitará conocer su valor asintótico para la Cauchy. Según el ejemplo-30, y

suponiendo una Cauchy de $\theta = 0$, el estimador de escala lo obtenemos con $\chi(\frac{u}{c}) = sgn(|\frac{u}{c}| - 1)$, buscando c que cumpla

$$\begin{aligned} E_{F_\gamma}\left[sgn(\left|\frac{X}{\gamma c}\right| - 1)\right] &= E_{F_0}\left[sgn(\left|\frac{U}{c}\right| - 1)\right] = 0 \text{ con } f_0(u) = \frac{1}{\pi(1+u^2)} \\ 1P(|U| &> c) + (-1)P(|U| < c) = 0 \\ F_0(c) &= \frac{3}{4} \text{ o sea } c = 1 \end{aligned}$$

Luego, si la población es Cauchy y se usa el $MAD = \frac{MAD}{1}$ asintóticamente estimará γ. Pero como el estimador de dispersión que usaremos lo pensamos para el modelo paramétrico normal, será el $MADN = \frac{MAD}{0.675}$; entonces si F es de Cauchy, asintoticamente $MADN = \frac{MAD}{0.675}$ estimará $\frac{\gamma}{0.675}$.

Con la $\psi_{H,k}$ de Huber, la función de influencia es

$$IF(x, T, F_{\theta,\gamma}) = \frac{\frac{\gamma}{0.675}\psi_{H,k}(\frac{x-\theta}{\frac{\gamma}{0.675}})}{E_{F_{\theta,\gamma}}(\psi'_{H,k}(\frac{X-\theta}{\frac{\gamma}{0.675}}))} = \frac{\frac{\gamma}{0.675}\psi_{H,k}(0.675\frac{x-\theta}{\gamma})}{E_{F_0}(\psi'_{H,k}(0.675U))}$$

y la varianza asintótica evaluada numericamente para $k = 1.345$

$$V(\widehat{\theta}_{H,k}, F_{\theta,\gamma}) = \frac{\gamma^2}{0.675^2}\frac{E_{F_0}\left[\psi^2_{H,k}(0.675U)\right]}{E^2_{F_0}\left[\psi'_{H,k}(0.675U)\right]}$$

De lo anterior surge que si interesa un un estimador de posición, se recomienda la ψ_k de Tukey con el $MADN$ como estimador previo de dispersión, y utilizando la mediana como punto de partida en el algoritmo de aproximación.

6.7 Estudio de la familia exponencial

Sea $P_\theta = \left\{F_\theta : f_\theta(x) = \frac{1}{\theta}e^{-\frac{x}{\theta}}I(x \geq 0)\right\}$ se analizarán:

6.7.1 Estimador óptimo en GES

Según (5.3) $\widetilde{\psi}(x;\theta) = sgn\left[\psi_0(x,\theta) - Med_{F_\theta}(\psi_0(x,\theta))\right]$ pero necesitamos

$$\psi_0(x,\theta) = -\frac{f'_\theta(x)}{f_\theta(x)} = \frac{1}{\theta}(1 - \frac{x}{\theta})$$

y la $Med_{F_\theta}(\psi_0(x,\theta)) = M$

$$\begin{aligned} P\left[\psi_0(X,\theta) \leq M\right] &= 0.5 \\ P\left[\frac{1}{\theta}(1 - \frac{X}{\theta}) \leq M\right] &= 0.5 \\ P\left[X \geq \theta(1 - \theta M)\right] &= 0.5 \\ e^{\theta M - 1} &= 0.5 \\ M &= \frac{1 - \ln 2}{\theta} \end{aligned}$$

luego queda

$$\begin{aligned}\widetilde{\psi}(x;\theta) &= sgn\left[\frac{1}{\theta}(1-\frac{x}{\theta})-\frac{1-\ln 2}{\theta}\right] \\ &\equiv sgn\left[(1-\frac{x}{\theta})-(1-\ln 2)\right] \\ &\equiv sgn(\ln 2-\frac{x}{\theta}) \\ &\equiv sgn(\ln 2-\frac{x}{\theta}) \equiv sgn(\theta\ln 2-x)\end{aligned}$$

y el funcional $E_F\left[\widetilde{\psi}(x;T(F))\right]=0$,

$$\begin{aligned}E_F\left[sgn(T(F)\ln 2-X)\right] &= 0 \\ (-1)P\left[T(F)\ln 2-X<0\right]+(1)P\left[T(F)\ln 2-X>0\right] &= 0\end{aligned}$$

o sea $P\left[X<T(F)\ln 2\right]=P\left[X<T(F)\ln 2\right]$ y esto define $Med(X)=T(F)\ln 2$, luego

$$T(F)=\frac{Med(X)}{\ln 2}$$

que es consistente de Fisher. Para hallar el estimador habría que resolver

$$\begin{aligned}\sum_{i=1}^{n} sgn(\widehat{\theta}_n\ln 2-X_i) &= 0 \\ \#(X_i &< \widehat{\theta}_n\ln 2)=\#(X_i>\widehat{\theta}_n\ln 2)\end{aligned}$$

lo que define una mediana muestral

$$\widehat{\theta}_n=\frac{Med\left\{X_1,X_2,\cdots,X_n\right\}}{\ln 2}$$

Este es el estimador con menor GES en la familia exponencial.

6.7.2 Distribución asintótica

De la (62) para M-estimadores en general, en F_θ

$$\sqrt{n}\left[\widehat{\theta}_n-\theta\right]\xrightarrow{\mathbf{d}}N(0;\frac{E_{F_\theta}\left[\widetilde{\psi}^2(X;\theta)\right]}{E^2_{F_\theta}\left[\widetilde{\psi}'(X;\theta)\right]})$$

donde

$$E_{F_\theta}\left[\widetilde{\psi}^2(X;\theta)\right]=E_{F_\theta}\left[sgn^2(\theta\ln 2-X)\right]=1$$

y en el denominador, para evitar los impulsos, usamos el mismo recurso para obtener la expresión alternativa de la IF en la (61)

$$\begin{aligned} E_{F_\theta}\left[\widetilde{\psi}'(X;\theta)\right] &= -\int_0^\infty \widetilde{\psi}(x;\theta)\frac{\partial}{\partial\theta}f_\theta(x)dx \\ &= -\int_0^\infty sgn(\theta\ln 2 - x)\frac{\partial}{\partial\theta}f_\theta(x)dx \\ &= -\int_0^\infty sgn(\theta\ln 2 - x)\frac{x-\theta}{\theta^2}\frac{1}{\theta}e^{-\frac{x}{\theta}}dx \\ &= -\int_0^{\theta\ln 2}\frac{x-\theta}{\theta^2}\frac{1}{\theta}e^{-\frac{x}{\theta}}dx + \int_{\theta\ln 2}^\infty \frac{x-\theta}{\theta^2}\frac{1}{\theta}e^{-\frac{x}{\theta}}dx \\ &= \frac{\ln 2}{\theta} \end{aligned}$$

luego en definitiva

$$\sqrt{n}\left[\widehat{\theta}_n - \theta\right] \xrightarrow{\mathbf{d}} N(0; \frac{\theta^2}{(\ln 2)^2})$$

6.7.3 Eficiencia asintótica

Se necesita la varianza asintótica del estimador de MV, que tiene $\psi_0(x,\theta) = \frac{1}{\theta}(1-\frac{x}{\theta})$, o sea

$$V(\psi_0, F_\theta) = \frac{E_{F_\theta}\left[\psi_0^2(X;\theta)\right]}{E^2_{F_\theta}\left[\psi_0'(X;\theta)\right]}$$

donde

$$E_{F_\theta}\left[\psi_0^2(X;\theta)\right] = \int_0^\infty \frac{1}{\theta^2}(1-\frac{x}{\theta})^2\frac{1}{\theta}e^{-\frac{x}{\theta}}dx = \frac{1}{\theta^2}$$

y además

$$E_{F_\theta}\left[\psi_0'(X;\theta)\right] = E_{F_\theta}\left[-\frac{1}{\theta^2} + \frac{2X}{\theta^3}\right] = \frac{1}{\theta^2}$$

luego

$$V(\psi_0, F_\theta) = \theta^2$$

y entonces

$$ARE = \frac{V(\psi_0, F_\theta)}{V(\widetilde{\psi}, F_\theta)} = \frac{\theta^2}{\frac{\theta^2}{(\ln 2)^2}} = (\ln 2)^2$$

6.7.4 Estimador óptimo de Hampel

Este estimador tiene la forma $\psi^*(x,\theta) = \psi_{k(\theta)}\left[\psi_0(x,\theta) - r(\theta)\right]$ y reemplazando $\psi_0(x,\theta)$ queda

$$\psi^*(x,\theta) = \psi_k\left[\frac{1}{\theta}(1-\frac{x}{\theta}) - r\right]$$

(solo para facilitar la notación se ha puesto k y r como si no dependiesen de θ), y como debe verificar la consistencia de Fisher debe ser(donde como siempre $k \geq 0$).

$$E_{F_\theta}\left[\psi_k\left[\frac{1}{\theta}(1-\frac{X}{\theta})-r\right]\right]=0$$

pero expresando $\psi^*(x,\theta)$ con mas detalle(notar que es no-creciente ya que estamos trabajando con la función ψ de un M-estimador en general)

$$\psi^*(x,\theta)=\begin{cases} k & \text{para } x<\theta\left[1-\theta(r+k)\right] \\ \frac{1}{\theta}(1-\frac{x}{\theta})-r & \text{para } \theta\left[1-\theta(r+k)\right]\leq x\leq\theta\left[1-\theta(r-k)\right] \\ -k & \text{para } x>\theta\left[1-\theta(r-k)\right] \end{cases} \tag{96}$$

y entonces la consistencia de Fisher se expresa(Suponiendo que $\theta\left[1-\theta(r+k)\right]<0$ y $\theta\left[1-\theta(r-k)\right]>0$, para que el primer punto de corte de la ψ_k esté debajo de cero, y el otro arriba. Si se toman los dos debajo de cero, no se puede cumplir nunca la consistencia de Fisher, y si los dos están arriba, sí se puede cumplir, pero las eficiencias asintóticas que resultan son muy bajas, menores a 0.72, así que no se considerará este caso)

$$\begin{aligned} \int_0^{\theta[1-\theta(r-k)]}\left[\frac{1}{\theta}(1-\frac{x}{\theta})-r\right]\frac{1}{\theta}e^{-\frac{x}{\theta}}dx+\int_{\theta[1-\theta(r-k)]}^{\infty}(-k)\frac{1}{\theta}e^{-\frac{x}{\theta}}dx &= 0 \qquad (97)\\ \frac{1}{\theta}e^{\theta(r-k)-1}-r &= 0 \end{aligned}$$

o sea queda la expresión

$$\theta k=\theta r-\ln(\theta r)-1 \tag{98}$$

Fijada una eficiencia asintótica desada, la cota de varianza será

$$V_{\max}(\theta)=\frac{V(\psi_0,F_\theta)}{ARE_{prop}}=\frac{\theta^2}{ARE_{prop}} \tag{99}$$

Pero se necesita ahora la varianza asintótica $V(\psi^*,F_\theta)$, pero antes

$$\begin{aligned} IF(x,\psi^*,F_\theta) &= \frac{\psi^*(x,\theta)}{-E_{F_\theta}(\psi^{*\prime}(x,\theta))}=\frac{\psi^*(x;\theta)}{\int\psi^*(x;\theta)\frac{\partial}{\partial\theta}f_\theta(x)dx} \\ &= \frac{\psi_k\left[\frac{1}{\theta}(1-\frac{x}{\theta})-r\right]}{\int\psi_k\left[\frac{1}{\theta}(1-\frac{x}{\theta})-r\right]\frac{x-\theta}{\theta^2}\frac{1}{\theta}e^{-\frac{x}{\theta}}dx} \end{aligned}$$

y luego

$$V(\psi^*,F_\theta)=\frac{E_{F_\theta}\left[\psi_k^2\left[\frac{1}{\theta}(1-\frac{X}{\theta})-r\right]\right]}{\left[\int\psi_k\left[\frac{1}{\theta}(1-\frac{x}{\theta})-r\right]\frac{x-\theta}{\theta^2}\frac{1}{\theta}e^{-\frac{x}{\theta}}dx\right]^2}$$

en el numerador luego de simplificar con la igualdad de (97) queda

$$E_{F_\theta}\left[\psi_k^2\left[\frac{1}{\theta}(1-\frac{X}{\theta})-r\right]\right]=\frac{1}{\theta^2}\left[\theta^2r^2+1-2(\theta k+1)\theta r\right]$$

y similarmente en el denominador, luego de simplificar con (97)

$$\int \psi_k \left[\frac{1}{\theta}(1-\frac{x}{\theta})-r\right] \frac{x-\theta}{\theta^2}\frac{1}{\theta}e^{-\frac{x}{\theta}}dx = \frac{\theta r\left[\theta(k-r)+2\right]-1}{\theta^2}$$

luego

$$V(\psi^*, F_\theta) = \theta^2 \frac{\theta^2 r^2 + 1 - 2(\theta k+1)\theta r}{\left[\theta r\left[\theta(k-r)+2\right]-1\right]^2}$$

igualando a la $V_{\max}(\theta)$, utilizando la (98) y operando queda una ecuación función de $K = k\theta$ y de $R = r\theta$ solamente

$$\frac{R^2+1-2R(R-\ln R)}{\left[R(1-\ln R)-1\right]^2} = \frac{1}{ARE_{prop}} \quad \text{con } K = R - \ln R - 1$$

Debido a las restricciones que se impusieron a los límites de la integral (97) debe cumplirse $1-(R+K) < 0$ y $1-(R-K) > 0$, y además $K > 0$ y $K = R - \ln R - 1$. De esto surge que hay que probar con $R \in (0, 0.203185)$.

A continuación se presenta una tabla con los valores de R y K para distintas eficiencias

ARE_{prop}	R	K
0.90	0.062	1.844
0.95	0.029	2.573
0.99	0.005	4.245

y en la $\psi^*(x,\theta)$ habrá que tomar $r(\theta) = \frac{R}{\theta}$ y $k(\theta) = \frac{K}{\theta}$. Por ejemplo para eficiencia 0.95 resultará

$$\psi^*(x,\theta) = \begin{cases} 2.573/\theta & \text{para } x < -1.602\theta \\ \frac{1}{\theta}(1-\frac{x}{\theta}) - \frac{0.029}{\theta} & \text{para } -1.602\theta \le x \le 3.544\theta \\ -2.573/\theta & \text{para } x > 3.544\theta \end{cases}$$

7 Regresión lineal con matriz de diseño fija

7.1 Método de cuadrados mínimos

7.1.1 Cuando se cumple $E(u_i) = 0$

Si

- $\mathbf{Y} \in \mathbb{R}^n$, es un vector de variables aleatorias observadas, y
- $\mathbf{X} \in \mathbb{R}^{n\times p}$ es fija, y usualmente $rg(\mathbf{X}) = p$

se define un modelo lineal cuando vale

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad \text{con} \begin{cases} u_i \text{ \textbf{i.i.d.}} \\ F(u_i) \text{ no depende de } \mathbf{x}_i \\ E(u_i) = 0,\ Var(u_i) = \sigma^2 \end{cases} \tag{100}$$

según el criterio de cuadrados mínimos, $\widehat{\boldsymbol{\beta}}$ se obtiene minimizando la norma de los residuos

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta})$$

derivando resultan las ecuaciones normales

$$(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{X} = \mathbf{0}$$

o también llamando $\mathbf{r}(\boldsymbol{\beta}) = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$ al vector de residuos

$$\mathbf{r}(\boldsymbol{\beta})'\mathbf{X} = \mathbf{0}$$

que expresa la ortogonalidad de $\mathbf{r}(\boldsymbol{\beta})$ con todas las columnas de $\mathbf{X}$. Pero se le dará otra forma a esta ecuación. Escribiendo $\mathbf{X}'\mathbf{r}(\boldsymbol{\beta}) = \mathbf{0}$, y llamando $\mathbf{x}_i$ a los vectores fila de la matriz $\mathbf{X}$ queda

$$\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i = \sum_{i=1}^{n} r_i(\boldsymbol{\beta})\mathbf{x}_i = \mathbf{0}$$

o sea la combinación lineal de los vectores fila de $\mathbf{X}$, usando como coeficientes las componentes del vector de residuos $\mathbf{r}(\boldsymbol{\beta})$ da cero. Esta propiedad se usará mas adelante. Además si $rg(\mathbf{X}) = p$, resulta

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad \text{con} \quad E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \quad \text{y} \quad \boldsymbol{\Sigma}_{\widehat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$$

En el caso $rg(\mathbf{X}) = p^* < p$, hay un problema de identificación en β que habrá que resolver; sin embargo el estimador de σ^2, sea $\mathbf{X}$ de rango completo o no, es siempre

$$\widehat{\sigma}^2 = S^2 = \frac{1}{n - p^*}\sum_{i=1}^{n} r_i^2(\widehat{\boldsymbol{\beta}}) \tag{101}$$

Además si $u_i \sim N(0; \sigma^2)$ vale, en el caso de rango completo

$$\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}; (\mathbf{X}'\mathbf{X})^{-1}\sigma^2) \tag{102}$$

Propiedades de equivarianza de $\widehat{\boldsymbol{\beta}}$ A continuación se estudiarán ciertas propiedades de equivarianza, muy básicas, que cumple el estimador de cuadrados mínimos de $\boldsymbol{\beta}$. Posteriormente se analizarán estimadores robustos de $\boldsymbol{\beta}$, y el deseo será que también las cumplan (aunque no siempre se logrará). Como el estimador $\widehat{\boldsymbol{\beta}}$ depende de $\mathbf{X}$ y de $\mathbf{Y}$, se usará la notación $\widehat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y})$.

- Equivarianza de regresión

$$\widehat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y} + \mathbf{X}\boldsymbol{\gamma}) = \widehat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}) + \boldsymbol{\gamma} \quad \forall \boldsymbol{\gamma} \in \mathbb{R}^p$$

- Equivarianza de escala

$$\widehat{\boldsymbol{\beta}}(\mathbf{X}, \lambda\mathbf{Y}) = \lambda\widehat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}) \quad \forall \lambda \in \mathbb{R}$$

  es decir, si cambian las unidades en que se mide $\mathbf{Y}$, los coeficientes se ajustan automáticamente

- Equivarianza afín

$$\widehat{\boldsymbol{\beta}}(\mathbf{XA},\mathbf{Y}) = \mathbf{A}^{-1}\widehat{\boldsymbol{\beta}}(\mathbf{X},\mathbf{Y}) \quad \forall \mathbf{A} \in \mathbb{R}^{p\times p} \text{ no singular}$$

  que como caso particular, en el caso que $\mathbf{A}$ sea diagonal, significa que los coeficientes se ajustan automáticamente cuando se cambian las unidades de las $\mathbf{X}$.

Las tres se demuestran fácilmente usando la (100), o en el caso de rango completo la (102).

Propiedades de equivarianza e invarianza de $\widehat{\boldsymbol{\sigma}}$ Igual que antes, como el estimador $\widehat{\boldsymbol{\sigma}}$ depende de $\mathbf{X}$ y de $\mathbf{Y}$, se usará la notación $\widehat{\boldsymbol{\sigma}}(\mathbf{X},\mathbf{Y})$. Resulta en este caso

- Invarianza de regresión

$$\widehat{\boldsymbol{\sigma}}(\mathbf{X},\mathbf{Y}+\mathbf{X}\boldsymbol{\gamma}) = \widehat{\boldsymbol{\sigma}}(\mathbf{X},\mathbf{Y}) + \boldsymbol{\gamma} \quad \forall \boldsymbol{\gamma} \in \mathbb{R}^p$$

  geometricamente: si al vector $\mathbf{Y}$ se le suma un vector paralelo al subespacio generado por las columnas de $\mathbf{X}$, no cambia el vector de residuos.

- Equivarianza de escala

$$\widehat{\boldsymbol{\sigma}}(\mathbf{X},\lambda\mathbf{Y}) = |\lambda|\,\widehat{\boldsymbol{\sigma}}(\mathbf{X},\mathbf{Y}) \quad \forall \lambda \in \mathbb{R}$$

  geometricamente: si se multiplica por una constante λ el vector $\mathbf{Y}$, su norma quedará multiplicada por $|\lambda|$, y también multiplicada por $|\lambda|$ la norma de su proyección ortogonal, que es el vector de residuos.

- Invarianza afín

$$\widehat{\boldsymbol{\sigma}}(\mathbf{XA},\mathbf{Y}) = \widehat{\boldsymbol{\sigma}}(\mathbf{X},\mathbf{Y}) \quad \forall \mathbf{A} \in \mathbb{R}^{p\times p} \text{ no singular}$$

  geometricamente: si se cambia la base del subespacio generado por las columnas de $\mathbf{X}$, el subespacio generado será el mismo, y por lo tanto también el vector de residuos.

7.1.2 Cuando $E(u_i) \neq 0$

Si $E(\mathbf{u}) \neq \mathbf{0}$, el estimador de β será sesgado ya que

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}) &= E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left[\mathbf{X}\boldsymbol{\beta}+E(\mathbf{u})\right] \\ &= \boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}) \neq \boldsymbol{\beta} \end{aligned}$$

Sin embargo si $E(\mathbf{u}) = \mathbf{j}\alpha$ y la matriz $\mathbf{X}$ contiene el vector $\mathbf{j} = (1,1,\cdots,1)'$ (ambos requisitos), el estimador de los coeficientes que no corresponden a $\mathbf{j}$, sí sera insesgado. Llamemos $\mathbf{X} = [\mathbf{j},\underline{\mathrm{X}}]$, luego el modelo queda

$$\mathbf{Y} = \mathbf{j}\beta_0 + \underline{\mathrm{X}}\boldsymbol{\beta}_1 + \mathbf{u}$$

sumando y restando $\mathbf{j}\alpha$, y llamando $\mathbf{u}^* = \mathbf{u} - \mathbf{j}\alpha$, queda

$$\mathbf{Y} = \mathbf{j}(\beta_0 + \alpha) + \underset{\sim}{\mathrm{X}}\boldsymbol{\beta}_1 + \mathbf{u}^*$$

y como ahora $E(\mathbf{u}^*) = \mathbf{0}$, será un modelo lineal insesgado, y entonces

$$E(\widehat{\beta}_0) = \beta_0 + \alpha \text{ (sesgado)} \quad \text{pero} \quad E(\widehat{\boldsymbol{\beta}}_1) = \boldsymbol{\beta}_1 \text{(insesgado)}$$

y la matriz de covarianza de $\widehat{\boldsymbol{\beta}}_1$, será

$$\boldsymbol{\Sigma}_{\widehat{\beta}_1} = \mathbf{C}^{-1}\sigma^2$$

donde $\mathbf{C}^{-1}$ surge de quitarle la primera fila y columna a $(\mathbf{X}'\mathbf{X})^{-1}$. Y respecto del estimador de varianza sigue valiendo la (101), ya que $\boldsymbol{\Sigma}_{\mathbf{u}^*} = \boldsymbol{\Sigma}_{\mathbf{u}} = \mathbf{I}_n\sigma^2$.

7.2 M-estimador simultaneo

Aquí ampliaremos la definición (100) del modelo lineal mediante

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad \text{con } u_i \textbf{ i.i.d.} \quad \text{y} \quad F(u_i) = F_0(\frac{u_i}{\sigma})$$

donde σ es un parámetro de escala. Como la densidad de Y_i es $\frac{1}{\sigma}f_0(\frac{Y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma})$, la verosimilitud y el $-\ln L(\boldsymbol{\beta}, \sigma)$ son

$$L(\boldsymbol{\beta}, \sigma) = \frac{1}{\sigma^n}\prod_{i=1}^{n} f_0(\frac{Y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}) \qquad \text{y} \qquad -\ln L(\boldsymbol{\beta}, \sigma) = \sum_{i=1}^{n} -\ln f_0(\frac{Y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}) + n\ln\sigma$$

y entonces

$$\widehat{\boldsymbol{\beta}},\, \widehat{\sigma} = \underset{\beta,\sigma}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} -\ln f_0(\frac{r_i(\boldsymbol{\beta})}{\sigma}) + \ln\sigma \tag{103}$$

El M-estimador simultaneo se define mediante

$$\widehat{\boldsymbol{\beta}},\, \widehat{\sigma} = \underset{\beta,\sigma}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \rho(\frac{r_i(\boldsymbol{\beta})}{\sigma}) + \ln\sigma \tag{104}$$

donde ρ es una función rho, y como siempre si $\rho(u) = -\ln f_0(u)$ se obtiene el estimador de MV. Si ahora usando que $r_i(\boldsymbol{\beta}) = Y_i - \mathbf{x}_i'\boldsymbol{\beta}$ se deriva la (104) respecto de $\boldsymbol{\beta}$ y σ queda el sistema

$$\begin{cases} \frac{1}{n}\sum_{i=1}^{n} \psi(\frac{r_i(\boldsymbol{\beta})}{\sigma})\mathbf{x}_i = \mathbf{0} \\ \frac{1}{n}\sum_{i=1}^{n} \psi(\frac{r_i(\boldsymbol{\beta})}{\sigma})\frac{r_i(\boldsymbol{\beta})}{\sigma} = 1 \end{cases}$$

Se define entonces el M-estimador simultaneo de $\boldsymbol{\beta}$ y σ mediante

$$\begin{cases} \sum_{i=1}^{n} \psi(\frac{r_i(\boldsymbol{\beta})}{\sigma})\mathbf{x}_i = \mathbf{0} \\ \frac{1}{n}\sum_{i=1}^{n} \rho_s(\frac{r_i(\boldsymbol{\beta})}{\sigma}) = \delta \end{cases}$$

donde ψ, ρ_s son funciones psi y rho que cumplen con los requisitos de (4.3). Además en el caso en que $\psi(u) = -\frac{f_0'(u)}{f_0(u)}$, $\rho_s(u) = \psi(u)u$, y $\delta = 1$, los estimadores $\widehat{\boldsymbol{\beta}}$ y $\widehat{\sigma}$ obtenidos coinciden con los de máxima verosimilitud. Notar también que la segunda de estas ecuaciones es escalar, y la primera vectorial, interpretando los $\psi(\frac{r_i(\boldsymbol{\beta})}{\sigma})$ como coeficientes que hacen nula la combinación lineal de los vectores fila de X.

Como en el caso en que ψ sea monótona los estimadores simultaneos son menos robustos que los obtenidos estimando previamente σ, se analizar el siguiente M-estimador.

7.3 M-estimador con estimador previo de escala

Si en la (104) se estima previamente σ, el M-estimador de $\boldsymbol{\beta}$ con estimación previa de escala quedaría

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\beta} \sum_{i=1}^{n} \rho(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}}) \tag{105}$$

donde $\widehat{\sigma}$ es un estimador de escala de los residuos que tenga un alto punto de ruptura, y ρ una función rho.

Derivando respecto de $\boldsymbol{\beta}$ se obtiene

$$\sum_{i=1}^{n} \psi(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}})\mathbf{x}_i = \mathbf{0} \tag{106}$$

donde ψ es una función psi.

Notar que si en (105) se toma $\rho(u) = u^2$, resulta $\widehat{\boldsymbol{\beta}}_{LS} = \arg\min_{\beta} \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta})$ que coincide con el estimador de cuadrados mínimos (también con el de MV para u_i distribuídos normalmente). Y si se toma $\rho(u) = |u|$, se obtiene $\widehat{\boldsymbol{\beta}}_{L^1} = \arg\min_{\beta} \sum_{i=1}^{n} |r_i(\boldsymbol{\beta})|$, el estimador L^1, que estima β a través de la minimización de la suma de valores absolutos de los residuos (y coincide con el de MV si los u_i se distribuyen según una doble exponencial). Lo interesante en estos dos casos es que se puede obtener el estimador de $\boldsymbol{\beta}$, sin necesidad de la estimación previa $\widehat{\sigma}$, ya que $\widehat{\sigma}$ es un factor constante que sale fuera de la sumatoria, y no interviene en la minimización.

Justamente esta última propiedad se utilizará más adelante para obtener $\widehat{\sigma}$: primero se hace una regresión L^1(despreocupandonos de la escala) y se obtiene $\widehat{\boldsymbol{\beta}}_{L^1}$, después se calculan los residuos $r_{L^1 i} = Y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{L^1}$, y con ellos se calcula $\widehat{\sigma}$.

7.3.1 Existencia y unicidad

Llamando $Sol(\rho)$ a las soluciones de la (105), y $Sol(\psi)$ a las soluciones de (106), se tienen los resultados

- **Existencia**: si $\rho(u)$ es continua, no-acotada, y no-decreciente como función de $|u|$, entonces

$$Sol(\rho) \neq \emptyset \quad \text{(es decir, existen soluciones)}$$

- **Unicidad**: si $\psi(u)$ es no-decreciente, entonces

$$Sol(\psi) \subset Sol(\rho) \quad \text{(o sea, toda solución en } Sol(\psi), \text{ lo es de } Sol(\rho)).$$

y si $\psi(u)$ es creciente, la solución es única.

7.3.2 Consistencia y distribución asintótica

Aquí se designará $\widehat{\sigma}_n$ al estimador previo de escala, y $\widehat{\boldsymbol{\beta}}_n$ al M-estimador de $\boldsymbol{\beta}$. En forma similar a lo tratado en cuadrados mínimos se considerarán dos casos.

- Cuando se cumple $E_F\left[\psi(\frac{u}{\sigma})\right] = 0$,(como ψ es impar, esto se cumplirá si la distribución de los u_i es simétrica), y además

1. $\widehat{\sigma}_n \xrightarrow{\mathbf{p}} \sigma_\infty$
2. $\max\limits_{1\leq i\leq n} \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \longrightarrow 0$

entonces resultará $\widehat{\boldsymbol{\beta}}_n \xrightarrow{\mathbf{ctp}} \boldsymbol{\beta}$. Y con algunas hipótesis adicionales también

$$\widehat{\boldsymbol{\beta}}_n \sim N_p(\boldsymbol{\beta}; v(\mathbf{XX})^{-1}) \tag{107}$$

donde

$$v = \sigma_\infty^2 \frac{E_F\left[\psi^2(\frac{u}{\sigma_\infty})\right]}{E_F^2\left[\psi'(\frac{u}{\sigma_\infty})\right]}$$

- Cuando no se cumple $E_F\left[\psi(\frac{u}{\sigma})\right] = 0$, pero $\mathbf{X} = [\mathbf{j}, \underset{\sim}{\mathrm{X}}]$, y además

1. $\widehat{\sigma}_n \xrightarrow{\mathbf{p}} \sigma_\infty$
2. $\max\limits_{1\leq i\leq n} \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \longrightarrow 0$

entonces resultará solamente que $\widehat{\boldsymbol{\beta}}_{1n} \xrightarrow{\mathbf{ctp}} \boldsymbol{\beta}_1$ (ya que $\widehat{\boldsymbol{\beta}}_{0n}$ será asintóticamente sesgado). Y con algunas hipótesis adicionales se cumplirá también

$$\widehat{\boldsymbol{\beta}}_{1n} \sim N_{p-1}(\boldsymbol{\beta}_1; v\mathbf{C}^{-1}) \tag{108}$$

donde $v = \sigma_\infty^2 \frac{E_F\left[\psi^2(\frac{u}{\sigma_\infty})\right]}{E_F^2\left[\psi'(\frac{u}{\sigma_\infty})\right]}$ y $\mathbf{C}^{-1}$ surge de quitarle a $(\mathbf{XX})^{-1}$ la primera fila y columna.

7.3.3 Eficiencia asintótica

Primero de definirá la eficiencia asintótica cuando el parámetro tiene varias dimensiones. Sea $\boldsymbol{\theta} \in \mathbb{R}^p$, y $\widehat{\boldsymbol{\theta}}_n$ un estimador cuya distribución asintótica sea normal, o sea $\widehat{\boldsymbol{\theta}}_n \sim N_p(\boldsymbol{\theta}; \mathbf{V})$. Se supondrá también que el estimador de MV, $\widehat{\boldsymbol{\theta}}_{Mn}$ también asintóticamente es normal, o sea $\widehat{\boldsymbol{\theta}}_{Mn} \sim N_p(\boldsymbol{\theta}; \mathbf{V}_M)$. Como las varianzas ahora son matrices, se compararán las varianzas de una misma combinación lineal de ambos estimadores, $\mathbf{c}'\widehat{\boldsymbol{\theta}}_n$ y $\mathbf{c}'\widehat{\boldsymbol{\theta}}_{Mn}$, cuyas varianzas son respectivamente $\mathbf{c}'\mathbf{V}\mathbf{c}$ y $\mathbf{c}'\mathbf{V}_M\mathbf{c}$. Luego la eficiencia sería $\mathbf{c}'\mathbf{V}_M\mathbf{c}/\mathbf{c}'\mathbf{V}\mathbf{c}$, pero como en general depende de $\mathbf{c}$, se define la eficiencia para el $\mathbf{c}$ más desfavorable, o sea

$$eff(\widehat{\boldsymbol{\theta}}_n) = \min_{\mathbf{c}\neq\mathbf{0}} \frac{\mathbf{c}'\mathbf{V}_M\mathbf{c}}{\mathbf{c}'\mathbf{V}\mathbf{c}} = \lambda_1(\mathbf{V}^{-1}\mathbf{V}_M)$$

donde $\lambda_1(\mathbf{V}^{-1}\mathbf{V}_M)$ es el mayor autovalor de $\mathbf{V}^{-1}\mathbf{V}_M$.

Volviendo al estimador de $\boldsymbol{\beta}$, como el estimador de MV tiene matriz de covarianza $\sigma^2(\mathbf{X}\mathbf{X})^{-1}$ en el numerador de la expresión de la eficiencia pondremos $\sigma^2\mathbf{c}'(\mathbf{X}\mathbf{X})^{-1}\mathbf{c}$; y en el denominador, usando la matriz de covarianza del M-estimador de (107) $\sigma_\infty^2 \frac{E_F\left[\psi^2(\frac{u}{\sigma_\infty})\right]}{E_F^2\left[\psi'(\frac{u}{\sigma_\infty})\right]}\mathbf{c}'(\mathbf{X}\mathbf{X})^{-1}\mathbf{c}$, luego la eficiencia será

$$eff(\widehat{\boldsymbol{\beta}}_n) = \frac{\sigma^2}{\sigma_\infty^2}\left(\frac{E_F\left[\psi^2(\frac{u}{\sigma})\right]}{E_F^2\left[\psi'(\frac{u}{\sigma})\right]}\right)^{-1} = \frac{\sigma^2}{v}$$

siendo la misma para (108). Notar que la eficiencia no depende de la matriz $\mathbf{X}$.

Sin embargo como lo usual es evaluar la eficiencia respecto de residuales $N(0;\sigma)$, si se ajusta $\widehat{\sigma}_n$ de manera que $\widehat{\sigma}_n \xrightarrow{\mathbf{p}} \sigma_\infty = \sigma$ la eficiencia queda

$$eff(\widehat{\boldsymbol{\beta}}_n) = \left(\frac{E_F\left[\psi^2(\frac{u}{\sigma})\right]}{E_F^2\left[\psi'(\frac{u}{\sigma})\right]}\right)^{-1}$$

7.3.4 Estimación previa de la escala

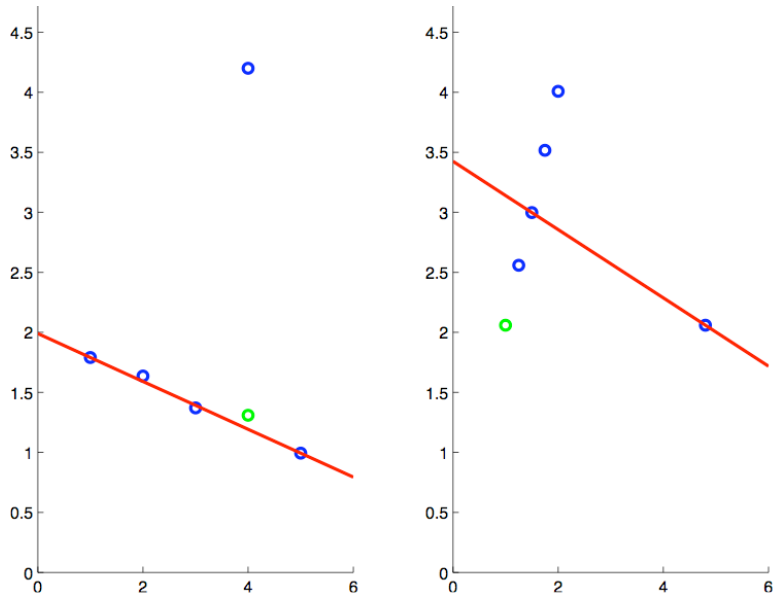
Cuando se definió el M-estimador de posición con estimación previa de escala, se utilizó para este fin el MAD o el $MADN$. Ahora en regresión haremos algo similar. Primero se ajusta una regresión L^1, obteniendo $\widehat{\boldsymbol{\beta}}_{L^1} = \arg\min_{\beta} \sum_{i=1}^{n} |r_i(\boldsymbol{\beta})|$ (que no requiere conocimiento previo de escala), luego se calculan los residuos $r_{L^1 i} = Y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{L^1}$, y finalmente se define

$$\widehat{\sigma} = \frac{1}{0.675} \underset{r_i^{L^1} \neq 0}{Med} \{|r_{L^1 i}|\}$$

Notar que al calcular la mediana se excluyen los residuos nulos. El motivo es que en la regresión L^1 por lo menos p residuales son nulos, y entonces si se los incluye, la σ puede ser subestimada. Por último en el caso de residuales $N(0;\sigma)$ resulta

$$\widehat{\sigma}_n \xrightarrow{\mathbf{p}} \sigma$$

Example 39 *En la figura (sacada de Henri Pesonen-Robust Regression), se muestra que la regresión L^1 es robusta frente a outliers en* $\mathbf{Y}$ *(gráfico de la izquierda), pero poco robusta frente a puntos de alto leverage en* $\mathbf{X}$ *(gráfico de la derecha).*



7.3.5 Equivarianza del M-estimador de regresión

Se presentarán sin demostración las siguientes afirmaciones muy fáciles de probar.

- $\widehat{\boldsymbol{\beta}}_{L^1}(\mathbf{X}, \mathbf{Y})$ es regresión, escala y afín equivariante

utilizando esto se prueba que

- $\widehat{\sigma}(\mathbf{X}, \mathbf{Y})$ es regresión y afín invariante, y escala equivariante

utilizando esta $\widehat{\sigma}$, y resolviendo (105) o (106) para obtener el M-estimador $\widehat{\boldsymbol{\beta}}$ se prueba finalmente que

- $\widehat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y})$ es regresión, escala y afín equivariante

es decir el M-estimador de $\boldsymbol{\beta}$ cumple con las mismas tres equivarianzas que el de cuadrados mínimos.

7.3.6 Estimación final de la varianza

Por último, con $\widehat{\boldsymbol{\beta}}$ se calculan los nuevos residuos $r_i = Y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$, y la estimación muestral de la varianza que figura en (107) la haremos mediante

$$\widehat{v} = \widehat{\sigma}^2 \frac{ave_i \left\{\psi^2(\frac{r_i}{\widehat{\sigma}})\right\}}{ave_i^2 \left\{\psi'(\frac{r_i}{\widehat{\sigma}})\right\}} \frac{n}{n-p}$$

donde se pone $n-p$ para ajustar los grados de libertad(igual que en cuadrados mínimos). Luego resultará

$$\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}; \widehat{v}(\mathbf{XX})^{-1})$$

7.3.7 Consideraciones respecto de la función ψ

Recordando lo analizado para el M-estimador de posición en cuanto al balance entre robustez y eficiencia(6.6), la recomendación fué utilizar una función ψ re-descendente como la ψ_k de Tukey, ya que al dar peso nulo a las grandes desviaciones, se logra alta eficiencia tanto para la normal, como para otras distribuciones de colas mas pesadas. Además se mantiene controlado el sesgo máximo, e incluso nulo para altas desviaciones. También como estimador previo de escala se sugirió el $MADN$ que tiene punto de ruptura $\frac{1}{2}$, que lo hereda también el M-estimador. Y como en el proceso iterativo de cálculo se requiere un punto de partida como estimador inicial, en este caso se usa la mediana.

Por los mismos motivos en el caso de regresión también conviene usar una ψ re-descendente. Pero se necesita un estimador inicial de $\boldsymbol{\beta}$.

Justamente en el caso que X sea fija y cumpliendo ciertas condiciones (que en particular satisfacen los modelos de análisis de varianza), se puede usar una ψ monótona para obtener el estimador inicial. La ventaja del estimador con ψ monótona es que es más fácil de calcular, es robusto, y como se analizó al tratar la unicidad(7.3.1) toda solución en $Sol(\psi)$, lo es de $Sol(\rho)$, y en el caso que ψ sea creciente la solución es única. Entonces la recomendación para el estimador incial es usar una ψ monotona, con estimación previa de escala mediante una regresión L^1.

Finalmente cuando los predictores no sean fijos, o fijos pero de alguna manera "desbalanceados", el estimador previo con ψ monótona deja de ser confiable, y hay que recurrir a otros estimadores que se verán en el próximo capítulo.

7.4 Punto de ruptura cuando X es fija

Se analizará el punto de ruptura para una muestra finita de un M-estimador de regresión con $\mathbf{X}$ fija. Suponiendo que $\mathbf{X}$ es de rango completo, y al ser fija, solo se puede modificar $\mathbf{Y}$. Si $\widehat{\boldsymbol{\beta}}_n(\mathbf{X}, \mathbf{Y})$ es el estimador, se define

$$\varepsilon_n^* = \frac{1}{n} \underset{m \geq 0}{máx} \left\{ m : \exists K_m \text{ tq. cambiando } Y_{i_1}, \cdot\cdot, Y_{i_m} \text{ arbitrariamente, } \widehat{\boldsymbol{\beta}}_n(\mathbf{X}, \mathbf{Y}) \in K_m \right\} = \frac{m^*}{n}$$

Para el estimador de cuadrados mínimos, como $\widehat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$, basta que un solo $Y_i \longrightarrow \infty$ para que el estimador caiga fuera de todo conjunto acotado, luego $\varepsilon_n^* = \varepsilon^* = 0$.

7.4.1 Estimadores equivariantes de regresion

En el caso de un estimador equivariante de regresión se puede encontrar una cota para el punto de ruptura. Se supondrá en la demostración una matriz $\mathbf{X}$ de rango completo, aunque el resultado vale en general.

Sea k^* el máximo número de vectores fila de $\mathbf{X}$ que pertenecen a un mismo subespacio de dimensión $< p$. Debido a esto, existirá un vector $\gamma \in \mathbb{R}^p$, con $\gamma \neq \mathbf{0}$, que es ortogonal a todos ellos. Luego se puede definir

$$k^* = \max\{\#(\mathbf{x}_i'\boldsymbol{\gamma} = 0 : \gamma \in \mathbb{R}^p, \boldsymbol{\gamma} \neq \mathbf{0}\}$$

entonces la combinación lineal $\mathbf{X}\boldsymbol{\gamma}t$ (donde $t \in \mathbb{R}$ es un escalar) será un vector de $\mathbb{R}^n$ que tiene k^* componentes nulas, que por comodidad de notación las supondremos las primeras, y las llamaremos el grupo "a". Y las restantes $n-k^*$, se las separa en el grupo "b" de $r = \left[\frac{n-k^*}{2}\right]$ componentes, y el grupo "c" al resto.

A continuación a partir del vector $\mathbf{Y}$ se obtienen el vector $\mathbf{Y}^*$(sumandole las componentes del grupo "b") y el $\mathbf{Y}^{**}$(restandole las componentes del grupo "c"), quedando

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \cdot \\ Y_{k^*} \\ Y_{k^*+1} \\ \cdot \\ Y_{k^*+a} \\ Y_{k^*+a+1} \\ \cdot \\ Y_n \end{bmatrix} \qquad \mathbf{Y}^* = \begin{bmatrix} Y_1 \\ \cdot \\ Y_{k^*} \\ Y_{k^*+1} + \mathbf{x}'_{k^*+1}\boldsymbol{\gamma}t \\ \cdot \\ Y_{k^*+r} + \mathbf{x}'_{k^*+r}\boldsymbol{\gamma}t \\ Y_{k^*+r+1} \\ \cdot \\ Y_n \end{bmatrix} \qquad \mathbf{Y}^{**} = \begin{bmatrix} Y_1 \\ \cdot \\ Y_{k^*} \\ Y_{k^*+1} \\ \cdot \\ Y_{k^*+r} \\ Y_{k^*+r+1} - \mathbf{x}'_{k^*+r+1}\boldsymbol{\gamma}t \\ \cdot \\ Y_n - \mathbf{x}'_n\boldsymbol{\gamma}t \end{bmatrix}$$

notar que $\mathbf{Y}^* - \mathbf{Y}^{**} = \mathbf{X}\boldsymbol{\gamma}t$. La intención ahora es tender $t \to \infty$, (en cuyo caso r componentes de Y^* tenderán a ∞, y $(n-k^*)-r$ de Y^{**} también), y ver que pasa con los respectivos $\widehat{\boldsymbol{\beta}}$. Como

$$\begin{aligned} \widehat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}^*) - \widehat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}^{**}) &= \widehat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}^*) - \widehat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}^* - \mathbf{X}\boldsymbol{\gamma}t) \\ &= \widehat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}^*) - \left[\widehat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y}^*) - \boldsymbol{\gamma}t)\right] \\ &= \boldsymbol{\gamma}t \end{aligned}$$

donde el el segundo paso se utilizó la equivarianza de regresión. Entonces si $t \to \infty$, alguno de los betas tenderá a ∞, lo que implica que $r > m^*$ o $(n - k^*) - r > m^*$, que equivale a

$$m^* < \max\{r, (n-k^*) - r\} = r = \left[\frac{n-k^*+1}{2}\right]$$

como esto es lo mismo que $m^* \leq \left[\frac{n-k^*-1}{2}\right]$ en definitiva queda

$$\boldsymbol{\varepsilon}^* \leq \frac{1}{n}\left[\frac{n-k^*-1}{2}\right]$$

Por último, siempre $k^* \geq p-1$, y cuando $k^* = p-1$, se dice que $\mathbf{X}$ está en "posición general". Entonces si $\mathbf{X}$ está en "posición general", la cota máxima para estimadores equivariantes es $\frac{1}{n}\left[\frac{n-p}{2}\right]$.

Example 40 *Se analizará el punto de ruptura para un análisis de varianza de un factor, tres tratamientos, usando la ψ_k de Huber, y con estimación previa de escala. Los datos son*

$\mathbf{T}_1$	$\mathbf{T}_1$	$\mathbf{T}_1$
Y_1	Y_5	Y_8
Y_2	Y_6	Y_9
Y_3	Y_7	Y_{10}
Y_4		

el modelo expresado matricialmente, y el vector $\psi_k(\frac{r(\boldsymbol{\alpha})}{\widehat{\sigma}})$ son

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \\ Y_{10} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \\ u_8 \\ u_9 \\ u_{10} \end{bmatrix} \qquad \psi_k(\frac{r(\boldsymbol{\alpha})}{\widehat{\sigma}}) = \begin{bmatrix} \psi_k(\frac{Y_1-\alpha_1}{\widehat{\sigma}}) \\ \psi_k(\frac{Y_2-\alpha_1}{\widehat{\sigma}}) \\ \psi_k(\frac{Y_3-\alpha_1}{\widehat{\sigma}}) \\ \psi_k(\frac{Y_4-\alpha_1}{\widehat{\sigma}}) \\ \psi_k(\frac{Y_5-\alpha_2}{\widehat{\sigma}}) \\ \psi_k(\frac{Y_6-\alpha_2}{\widehat{\sigma}}) \\ \psi_k(\frac{Y_7-\alpha_2}{\widehat{\sigma}}) \\ \psi_k(\frac{Y_8-\alpha_3}{\widehat{\sigma}}) \\ \psi_k(\frac{Y_9-\alpha_3}{\widehat{\sigma}}) \\ \psi_k(\frac{Y_{10}-\alpha_3}{\widehat{\sigma}}) \end{bmatrix}$$

según (106) y llamando a $\mathbf{e}_1$,$\mathbf{e}_2$, y $\mathbf{e}_3$ a los vectores fila de $\mathbf{X}$ queda

$$\left[\sum_{i=1}^{4}\psi_k(\frac{Y_i-\alpha_1}{\widehat{\sigma}})\right]\mathbf{e}_1 + \left[\sum_{i=5}^{7}\psi_k(\frac{Y_i-\alpha_2}{\widehat{\sigma}})\right]\mathbf{e}_2 + \left[\sum_{i=8}^{10}\psi_k(\frac{Y_i-\alpha_3}{\widehat{\sigma}})\right]\mathbf{e}_3 = \mathbf{0}$$

debido a la ortogonalidad entre los $\mathbf{e}_i$, se tiene entonces el sistema de ecuaciones

$$\begin{cases} \psi_k(\frac{Y_1-\alpha_1}{\widehat{\sigma}}) + \psi_k(\frac{Y_2-\alpha_1}{\widehat{\sigma}}) + \psi_k(\frac{Y_3-\alpha_1}{\widehat{\sigma}}) + \psi_k(\frac{Y_4-\alpha_1}{\widehat{\sigma}}) = 0 \\ \psi_k(\frac{Y_5-\alpha_2}{\widehat{\sigma}}) + \psi_k(\frac{Y_6-\alpha_2}{\widehat{\sigma}}) + \psi_k(\frac{Y_7-\alpha_2}{\widehat{\sigma}}) = 0 \\ \psi_k(\frac{Y_8-\alpha_3}{\widehat{\sigma}}) + \psi_k(\frac{Y_9-\alpha_3}{\widehat{\sigma}}) + \psi_k(\frac{Y_{10}-\alpha_3}{\widehat{\sigma}}) = 0 \end{cases}$$

Si por ejemplo en la última ecuación tendemos solo $Y_8 \to \infty$, notar que no puede $\alpha_3 \to \infty$, pues quedaría

$$\psi_k(\frac{Y_8-\alpha_3}{\widehat{\sigma}}) - k - k = 0$$

y no existiría α_3 solución. Con similar argumento tampoco puede haber solución si suponemos que $\alpha_3 \to -\infty$. En realidad la solución $\widehat{\alpha}_3$ es finita, concretamente $\widehat{\alpha}_3 = \frac{Y_9+Y_{10}}{2} + \frac{k}{2}\widehat{\sigma}$ (verificarlo!). Por otro lado si tendemos $Y_8 \to \infty$, $Y_9 \to \infty$, la solución α_3 no puede ser finita, pues quedaría

$$k + k + \psi_k(\frac{Y_{10}-\alpha_3}{\widehat{\sigma}}) = 0$$

que no tiene solución. Luego debe ser $\widehat{\alpha}_3 \to \infty$. Con argumentos similares si en la segunda se tiende $Y_5 \to \infty$ la solución $\widehat{\alpha}_2$ será finita. Idem si en la primera

se tiende por ejemplo $Y_1 \to \infty$. En definitiva tendiendo un solo $Y_i \to \pm\infty$, la solución es siempre finita, pero tendiendo dos, la solución puede ser infinita. Luego el punto de ruptura es $\varepsilon^* = \frac{1}{10}$. (En este análisis no se tuvo en cuenta la ruptura de la regresión L^1 para estimar $\widehat{\sigma}$. Sin embargo como se mostró en un ejemplo anterior, el ajuste L^1 tiene ruptura cero solo si el outlier es un punto de alto leverage en $\mathbf{X}$, que no es el caso en un analisi de varianza como este).

Como se trata de un estimador equivariante de regresión, analizaremos la cota. Observando las filas de la matriz $\mathbf{X}$, resulta $k^* = 7$, y entonces $\left[\frac{n-k^*-1}{2}\right] = \left[\frac{10-7-1}{2}\right] = 1$, y luego $\boldsymbol{\varepsilon}^* \leq \frac{1}{10}$. Se observa que en este caso particular, (un análisi de varianza de un factor), el punto de ruptura coincide con la cota para estimadores equivariantes de regresión en general.

7.4.2 El problema del leverage

Considérese un M-estimador de regresión con una ψ monótona y acotada(no re-descendente), cuya matriz $\mathbf{X}$ tiene uno de sus vectores fila que es linealmente independiente del resto. En este caso el leverage de esta fila es 1. Si se designa $\mathbf{x}_a$ a esta fila, según (106) la ecuación a resolver es

$$\psi(\frac{Y_1 - \mathbf{x}_1'\boldsymbol{\beta}}{\widehat{\sigma}})\mathbf{x}_1 + \cdots + \psi(\frac{Y_a - \mathbf{x}_a'\boldsymbol{\beta}}{\widehat{\sigma}})\mathbf{x}_a + \cdots + \psi(\frac{Y_n - \mathbf{x}_n'\boldsymbol{\beta}}{\widehat{\sigma}})\mathbf{x}_n = \mathbf{0}$$

Ahora tendamos $Y_a \to \infty$, y supongamos que existe solución $\widehat{\boldsymbol{\beta}}$ finita. Notar que deberá ser

$$\psi(\frac{Y_a - \mathbf{x}_a'\widehat{\boldsymbol{\beta}}}{\widehat{\sigma}}) = 0 \tag{109}$$

pues sinó, $\mathbf{x}_a$ quedaría expresado como combinación lineal de los restantes. Pero para que se satisfaga esta igualdad con $Y_a \to \infty$, deberá ser $\widehat{\boldsymbol{\beta}} \to \infty$ en contra de lo supuesto. En definitiva el punto de ruptura será $\boldsymbol{\varepsilon}^* = 0$.

Por motivos similares, aunque el leverage no sea exactamente 1, pero alto, cuando $Y_a \to \infty$, puede no resultar un $\widehat{\boldsymbol{\beta}}$ acotado, provocando también la ruptura y $\boldsymbol{\varepsilon}^* = 0$.

Entonces las matrices $\mathbf{X}$ con altos leverage pueden disminuir seriamente el punto de ruptura de un M-estimador monótono, incluso llevarlo a cero. Sin embargo si ψ es redescendente, como se hace nula fuera de cierto intervalo, este problema no se presenta ya que la (109) sí puede satisfacerse(pues aunque $Y_a \to \infty$, puede existir un $\widehat{\boldsymbol{\beta}}$ finito que la cumpla).

7.4.3 Punto de ruptura para ψ monótona

Los M-estimadores de regresión con ψ monótona tienen en general un punto de ruptura menor que la cota dada para los equivariantes. Sin embargo se prueba que en el caso de análisis de varianza de 1 y 2 factores se alcanza esta cota (ver ejemplo 40). Y se conjetura que este resultado también es válido para matrices $\mathbf{X}$ cuyos elementos sean solo 0 y 1.

Además como se analizó en la sección anterior, cuando ψ es monótona, las filas de $\mathbf{X}$ con alto leverage deterioran mucho el punto de ruptura.

En (7.3.7), y respecto del M-estimador de regresión se sugirió usar una ψ re-descendente, y para el estimador previo de $\boldsymbol{\beta}$ una ψ monótona. Sin embargo de acuerdo a lo recién analizado esto será válido solo para diseños $0-1$, pero no para matrices con puntos de leverage.

Cuando se presentan observaciones con leverage, el estimador inicial se obtendrá con otros métodos que se verán en el próximo capítulo, aplicables cuando $\mathbf{X}$ es aleatoria (y también para $\mathbf{X}$ fija).

8 Regresión lineal con matriz de diseño aleatoria

8.1 M-estimador con parámetros multidimensionales

Sea una muestra de n vectores aleatorios independientes $Z_1, Z_2, \cdots, Z_n$, todos definidos en el mismo espacio muestral $\Omega \subset \mathbb{R}^m$, y con la misma distribución $F_{\boldsymbol{\theta}}$, donde $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$, siendo Θ el espacio de parámetros. Se define modelo paramétrico a la familia de todas estas posibles funciones de distribución

$$P_\theta = \{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p\}$$

Considérese la función $\boldsymbol{\Psi}(Z, \boldsymbol{\theta}) \in \mathbb{R}^p$

$$\boldsymbol{\Psi}(Z, \boldsymbol{\theta}) = \begin{bmatrix} \Psi_1(Z, \boldsymbol{\theta}) \\ \cdot\cdot \\ \Psi_p(Z, \boldsymbol{\theta}) \end{bmatrix} \tag{110}$$

un M-estimador en general, se define mediante

$$\sum_{i=1}^{n} \boldsymbol{\Psi}(Z_i, \boldsymbol{\theta}) = \mathbf{0}$$

y de aquí surge el estimador $\widehat{\boldsymbol{\theta}}_n$. Además como es habitual en estadística robusta si la muestra $Z_1, Z_2, \cdots, Z_n \overset{\text{iid}}{\sim} F$ donde $F \notin P_{\boldsymbol{\theta}}$, sinó a una familia un poco mayor $\mathcal{F}$, donde $P_{\boldsymbol{\theta}} \subset \mathcal{F}$, el valor asintótico de este estimador se designará $\widehat{\boldsymbol{\theta}}_\infty(F)$ ya que en general dependerá de F. Y como siempre si este estimador es consistente de Fisher, resultará $\widehat{\boldsymbol{\theta}}_\infty(F_{\boldsymbol{\theta}}) = \boldsymbol{\theta}$.

Si ahora se considera una contaminación infinitesimal en el punto $Z_0 \in \mathbb{R}^m$, y la muestra $Z_1, Z_2, \cdots, Z_n \overset{\text{iid}}{\sim} F$, se puede definir la función de influencia, cuya expresión se demuestra que es

$$IF_{\widehat{\boldsymbol{\theta}}}(Z_0, F) = -\mathbf{B}^{-1} \boldsymbol{\Psi}(Z_0, \widehat{\boldsymbol{\theta}}_\infty(F)) \tag{111}$$

donde

$$B_{i,j} = E\left[\left. \frac{\partial \Psi_i(Z, \boldsymbol{\theta})}{\partial \theta_j} \right|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_\infty(F)} \right]$$

es decir usando de (110) las componentes de $\boldsymbol{\Psi}(Z,\boldsymbol{\theta})$

$$\mathbf{B} = E\left(\begin{bmatrix} \frac{\partial \Psi_1(Z,\boldsymbol{\theta})}{\partial \theta_1} & .. & \frac{\partial \Psi_1(Z,\boldsymbol{\theta})}{\partial \theta_p} \\ .. & .. & .. \\ \frac{\partial \Psi_p(Z,\boldsymbol{\theta})}{\partial \theta_1} & .. & \frac{\partial \Psi_p(Z,\boldsymbol{\theta})}{\partial \theta_p} \end{bmatrix}\right) \quad \text{con} \quad \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_\infty(F) \tag{112}$$

También se demuestra que el estimador $\widehat{\boldsymbol{\theta}}_n$ es asintóticamente normal con matriz de covarianza asintótica

$$\mathbf{V} = \mathbf{B}^{-1} E\left[\boldsymbol{\Psi}(Z,\boldsymbol{\theta})\boldsymbol{\Psi}'(Z,\boldsymbol{\theta})\right]\mathbf{B}^{-1} \quad \text{con } \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_\infty(F) \tag{113}$$

verificando también

$$\mathbf{V} = E\left[IF(Z,F)IF(Z,F)'\right] \tag{114}$$

8.2 Modelo lineal con X aleatoria

En este caso las observaciones son $(\mathbf{x}_i, Y_i)$ para $1 \leq i \leq n$, y el modelo es

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad \text{con} \begin{cases} u_i \text{ } \mathbf{i.i.d.} \\ u_i \text{ independiente de } \mathbf{x}_i \\ E(u_i) = 0,\ Var(u_i) = \sigma^2 \end{cases}$$

donde ahora se pide que "u_i sea independiente de $\mathbf{x}_i$", ya que los $\mathbf{x}_i$ son aleatorios.

Además el concepto de rango completo cuando $\mathbf{X}$ es aleatoria consiste en exigir que la distribución de $\mathbf{x}$ no este concentrada en un subespacio de dimensión $< p$, o sea

$$\nexists \mathbf{a} \neq \mathbf{0} \quad \text{con} \quad P(\mathbf{a}'\mathbf{x} = 0) = 1$$

cumpliendose esto el estimador de cuadrados mínimos está bién definido, verificándose $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ y condicionalmente

$$E(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}) = \boldsymbol{\beta} \quad \text{y} \quad \boldsymbol{\Sigma}_{\widehat{\beta}|\mathbf{X}} = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$$

Además si $u_i \sim N(0;\sigma^2)$ vale, también condicionalmente

$$\widehat{\boldsymbol{\beta}} \mid \mathbf{X} \sim N_p(\boldsymbol{\beta}; (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$$

Pero si las u_i no se distribuyen normalmente, y asumiendo que existe $\mathbf{V}_x = E(\mathbf{x}\mathbf{x}')$, vale asintóticamente

$$\widehat{\boldsymbol{\beta}} \mid \mathbf{X} \approx N_p(\boldsymbol{\beta}; \frac{\mathbf{V}_x\sigma^2}{n})$$

8.3 M-estimador con una función ρ acotada

En el capítulo anterior, cuando $\mathbf{X}$ es fija, se sugirió usar un M-estimador con ψ re-descendente, y como estimador previo de $\boldsymbol{\beta}$, otro M-estimador pero con ψ monótona(que es más fácil de calcular), y estimador previo de escala mediante una regresión L^1. Sin embargo, aún para $\mathbf{X}$ fija, si había puntos de alto leverage, el estimador inicial con ψ monótona dejaba de ser útil ya que el punto de ruptura disminuía mucho, pudiendo llegar a cero. El mismo problema lo tiene la regresión L^1.

Cuando $\mathbf{X}$ es aleatoria, como no se tiene control sobre las filas de la matriz de diseño, y tanto $\mathbf{Y}$ como $\mathbf{X}$ pueden tener outliers, se deberá utilizar otro método para obtener el estimador inicial de $\boldsymbol{\beta}$, y además este método no debe requerir una estimación previa de σ, (ya que no conviene usar una regresión L^1 por los motivos señalados antes).

Por eso, antes de proponer un buen estimador inicial de $\boldsymbol{\beta}$ se estudiarán ahora las propiedades del M-estimador con función ρ acotada, (y entonces, necesariamente, ψ re-descendente), que será el utilizado como estimador principal de $\boldsymbol{\beta}$..

El M-estimador se define mediante

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\beta} \sum_{i=1}^{n} \rho(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}}) \tag{115}$$

donde ρ es una función rho acotada, y $\widehat{\sigma}$ un estimador preliminar de escala de alto punto de ruptura. Si existe la derivada $\rho' = \psi$ también resulta

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

donde como ψ es re-descendente, esta ecuación puede tener varias soluciones que corresponden a multiples mínimos locales de (115), y generalmente solo uno de ellos corresponde al mínimo global $\widehat{\boldsymbol{\beta}}$.

También, y similarmente a lo que ocurre con el estimador con ψ monótona

$$\begin{array}{ll} & \text{Sí } \widehat{\sigma} \text{ es regresión y afín invariante, y escala equivariante} \\ \Longrightarrow & \widehat{\boldsymbol{\beta}} \text{ es regresión, escala y afín equivariante} \end{array}$$

8.3.1 Punto de ruptura cuando X es aleatoria

Se analizará el punto de ruptura para una muestra finita de un M-estimador de regresión con $\mathbf{X}$ aleatoria. Como ahora también se puede modificar $\mathbf{X}$, entonces llamando $Z_i = (\mathbf{x}_i, Y_i)$, si $\widehat{\boldsymbol{\beta}}_n(\mathbf{X}, \mathbf{Y})$ es el estimador, se define

$$\varepsilon_n^* = \frac{1}{n} \underset{m \geq 0}{máx} \left\{ m : \exists K_m \text{ tq. cambiando } Z_{i_1}, \cdot\cdot, Z_{i_m} \text{ arbitrariamente, } \widehat{\boldsymbol{\beta}}_n(\mathbf{X}, \mathbf{Y}) \in K_m \right\} = \frac{m^*}{n}$$

Notar que este punto de ruptura va a ser $\leq$ que el definido en el capítulo anterior, ya que ahora no solo $\mathbf{Y}$ sinó también $\mathbf{X}$ es aleatoria.

Caso en que $\widehat{\boldsymbol{\beta}}$ es un estimador equivariante de regresión Aquí vale la misma cota dada en el capítulo anterior para $\mathbf{X}$ fija, o sea

$$\varepsilon^* \leq \frac{1}{n}\left[\frac{n-k^*-1}{2}\right]$$

donde también

$$k^* = \max\{\#(\mathbf{x}_i'\boldsymbol{\gamma} = 0 : \gamma \in \mathbb{R}^p, \boldsymbol{\gamma} \neq \mathbf{0}\}$$

Caso en que ψ es monótona Asumiendo que σ es conocida y vale 1, entonces la ecuación que define al M-estimador es

$$\psi(Y_1 - \mathbf{x}_1'\boldsymbol{\beta})\mathbf{x}_1 + \cdots + \psi(Y_a - \mathbf{x}_a'\boldsymbol{\beta})\mathbf{x}_a + \cdots + \psi(Y_n - \mathbf{x}_n'\boldsymbol{\beta})\mathbf{x}_n = \mathbf{0} \qquad (116)$$

Ahora se supondrá que en $Z_a = (\mathbf{x}_a, Y_a)$, se tiende $\mathbf{x}_a \to \infty$ y $Y_a \to \infty$, pero de modo que $\frac{Y_a}{\|\mathbf{x}_a\|} \to \infty$. Notar que si la solución $\widehat{\boldsymbol{\beta}}$ se mantuviese acotada, se tendría

$$Y_a - \mathbf{x}_a'\boldsymbol{\beta} \geq Y_a - \|\mathbf{x}_a\|\left\|\widehat{\boldsymbol{\beta}}\right\| = \|\mathbf{x}_a\|\left(\frac{Y_a}{\|\mathbf{x}_a\|} - \left\|\widehat{\boldsymbol{\beta}}\right\|\right) \to \infty$$

y entonces ya que ψ es no-decreciente, $\psi(Y_a - \mathbf{x}_a'\boldsymbol{\beta}) \to Sup\ \psi > 0$. Luego el término $\psi(Y_a - \mathbf{x}_a'\boldsymbol{\beta})\mathbf{x}_a \to \infty$ pero conservando la suma en (116) nula, lo que es imposible. De aquí surge que $\widehat{\boldsymbol{\beta}}$ no puede pertenecer a un K_1 acotado, y entonces resulta $\varepsilon_n^* = 0$.

Caso en que ψ es re-descendente A continuación se verá que la función de influencia es no-acotada, sin embargo cuando ρ es acotada, es posible lograr un alto punto de ruptura, y por lo tanto el sesgo se mantendrá acotado aún para grandes valores de ε.

8.3.2 Función de influencia

Si se supone σ conocida, la ecuación del estimador es

$$\sum_{i=1}^{n}\psi\left(\frac{Y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\mathbf{x}_i = \mathbf{0}$$

pero esto se puede escribir

$$\sum_{i=1}^{n}\begin{bmatrix}\psi\left(\frac{Y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)x_{i,1}\\ \cdot\cdot\\ \psi\left(\frac{Y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)x_{i,p}\end{bmatrix} = \mathbf{0}$$

Si las observaciones se designan $Z_i = (\mathbf{x}_i, Y_i)$, con $Z_1, Z_2, \cdots, Z_n \overset{\text{iid}}{\sim} F$, esta es la ecuación de un M-estimador general para parametros multidimensionales(8.1),

cuya función $\boldsymbol{\Psi}(Z, \boldsymbol{\beta})$ es

$$\boldsymbol{\Psi}(Z, \boldsymbol{\beta}) = \begin{bmatrix} \Psi_1(Z, \boldsymbol{\beta}) \\ .. \\ \Psi_p(Z, \boldsymbol{\beta}) \end{bmatrix} = \begin{bmatrix} \psi(\frac{Y - \mathbf{x}'\boldsymbol{\beta}}{\sigma}) x_1 \\ .. \\ \psi(\frac{Y - \mathbf{x}'\boldsymbol{\beta}}{\sigma}) x_p \end{bmatrix} = \psi(\frac{Y - \mathbf{x}'\boldsymbol{\beta}}{\sigma})\mathbf{x} \qquad (117)$$

Suponiendo que el valor asintótico del estimador es $\boldsymbol{\beta}$, la matriz $\mathbf{B}$ de (112) será

$$\begin{aligned} \mathbf{B} &= E\left(\begin{bmatrix} \frac{\partial \Psi_1(Z,\boldsymbol{\beta})}{\partial \beta_1} & .. & \frac{\partial \Psi_1(Z,\boldsymbol{\beta})}{\partial \beta_p} \\ .. & .. & .. \\ \frac{\partial \Psi_p(Z,\boldsymbol{\beta})}{\partial \beta_1} & .. & \frac{\partial \Psi_p(Z,\boldsymbol{\beta})}{\partial \beta_p} \end{bmatrix} \right) = E\left(\begin{bmatrix} -\psi'(\frac{Y - \mathbf{x}'\boldsymbol{\beta}}{\sigma})\frac{x_1 x_1}{\sigma} & .. & -\psi'(\frac{Y - \mathbf{x}'\boldsymbol{\beta}}{\sigma})\frac{x_1 x_p}{\sigma} \\ .. & .. & .. \\ -\psi'(\frac{Y - \mathbf{x}'\boldsymbol{\beta}}{\sigma})\frac{x_p x_1}{\sigma} & .. & -\psi'(\frac{Y - \mathbf{x}'\boldsymbol{\beta}}{\sigma})\frac{x_p x_p}{\sigma} \end{bmatrix} \right) \\ &= E\left[-\frac{1}{\sigma}\psi'(\frac{Y - \mathbf{x}'\boldsymbol{\beta}}{\sigma})\mathbf{x}\mathbf{x}' \right] = -\frac{1}{\sigma} E\left[\psi'(\frac{Y - \mathbf{x}'\boldsymbol{\beta}}{\sigma}) \right] E(\mathbf{x}\mathbf{x}') \end{aligned}$$

(el último paso debido a la independencia), luego

$$\mathbf{B} = -\frac{1}{\sigma} E\left[\psi'(\frac{Y - \mathbf{x}'\boldsymbol{\beta}}{\sigma}) \right] \mathbf{V}_x$$

Finalmente llamando $Z_0 = (\mathbf{x}_0, Y_0)$, y usando (111) y (117) la función de influencia es

$$\begin{aligned} IF(\mathbf{x}_0, Y_0, F) &= \frac{\sigma}{E\left[\psi'(\frac{Y - \mathbf{x}'\boldsymbol{\beta}}{\sigma}) \right]} \mathbf{V}_x^{-1} \psi(\frac{Y_0 - \mathbf{x}_0'\boldsymbol{\beta}}{\sigma})\mathbf{x}_0 \\ &= \frac{\sigma}{E\left[\psi'(\frac{U}{\sigma}) \right]} \psi(\frac{Y_0 - \mathbf{x}_0'\boldsymbol{\beta}}{\sigma}) \mathbf{V}_{\mathbf{x}}^{-\mathbf{1}} \mathbf{x}_0 \end{aligned}$$

Notar que no es acotada, y que con ψ monótona, si $\mathbf{x}_0$ es fijo, basta que $Y_0 \to \infty$, para que $IF \to \infty$. Sin embargo si ψ es re-descendente con $\psi(x) = 0$ para $|x| \geq k$, para que $IF \to \infty$ se necesita que $\mathbf{x}_0 \to \infty$ y $\left| \frac{Y_0 - \mathbf{x}_0'\boldsymbol{\beta}}{\sigma} \right| < k$. O sea, grandes outliers no tienen influencia en la estimación.

En el caso que σ sea desconocido, y se lo estime mediante un estimador previo $\widehat{\sigma}_n$, la IF correspondiente a $\widehat{\boldsymbol{\beta}}_n$ dependerá también, con una expresión muy complicada, de la IF que corresponde a $\widehat{\sigma}_n$. Pero si F es simétrica, y ψ impar, todo se simplifica, y la IF de $\widehat{\boldsymbol{\beta}}_n$ depende de $\widehat{\sigma}_n$, solo a través de su valor asintótico. Luego, si $\widehat{\sigma}_n \xrightarrow{\mathbf{p}} \sigma$ (es consistente)

$$IF(\mathbf{x}_0, Y_0, F) = \frac{\sigma}{E\left[\psi'(\frac{U}{\sigma}) \right]} \psi(\frac{Y_0 - \mathbf{x}_0'\boldsymbol{\beta}}{\sigma}) \mathbf{V}_{\mathbf{x}}^{-\mathbf{1}} \mathbf{x}_0$$

8.3.3 Normalidad asintótica

Se demuestra que si

- $\mathbf{V}_x = E(\mathbf{x}\mathbf{x}')$ existe y finita

- $\widehat{\sigma}_n \xrightarrow{\mathbf{P}} \sigma$

entonces bajo condiciones bastante generales la solución $\widehat{\boldsymbol{\beta}}_n$ de (115) es consistente y asintóticamente normal. La varianza asintótica surge de la (114) mediante

$$\begin{aligned}
\mathbf{V} &= E\left[IF(\mathbf{x},Y,F)IF(\mathbf{x},Y,F)'\right] \\
&= E\left[\frac{\sigma^2}{E^2\left[\psi'(\frac{U}{\sigma})\right]}\psi^2(\frac{Y-\mathbf{x}'\boldsymbol{\beta}}{\sigma})\mathbf{V}_{\mathbf{x}}^{-\mathbf{1}}\mathbf{x}\mathbf{x}'\mathbf{V}_{\mathbf{x}}^{-\mathbf{1}}\right] \\
&= \sigma^2\frac{E\left[\psi^2(\frac{Y-\mathbf{x}'\boldsymbol{\beta}}{\sigma})\right]}{E^2\left[\psi'(\frac{U}{\sigma})\right]}\mathbf{V}_{\mathbf{x}}^{-\mathbf{1}} \\
&= \sigma^2\frac{E\left[\psi^2(\frac{U}{\sigma})\right]}{E^2\left[\psi'(\frac{U}{\sigma})\right]}\mathbf{V}_{\mathbf{x}}^{-\mathbf{1}}
\end{aligned}$$

luego asintóticamente se tendrá que

$$\widehat{\boldsymbol{\beta}} \approx N_p(\boldsymbol{\beta}; v\mathbf{V}_{\mathbf{x}}^{-\mathbf{1}})$$

donde

$$v = \sigma^2\frac{E\left[\psi^2(\frac{U}{\sigma})\right]}{E^2\left[\psi'(\frac{U}{\sigma})\right]}$$

y también como con $\mathbf{X}$ fija, la eficiencia no depende de la distribución de $\mathbf{x}$

$$eff(\widehat{\boldsymbol{\beta}}) = \frac{\sigma^2}{v}$$

8.4 MM-estimador

Se analizará ahora un estimador de $\boldsymbol{\beta}$ que tiene un alto punto de ruptura y también alta eficiencia respecto de la distribución normal. Para lograr esto, como **estimador principal** $\widehat{\boldsymbol{\beta}}_{MM}$ se usará un M-estimador con función ρ acotada, y estimación previa de σ. Pero se necesitará un **estimador inicial** $\widehat{\boldsymbol{\beta}}_0$, que servirá también como **estimador previo** $\widehat{\sigma}$. Se describirán a continuación los requisitos que deben cumplir para lograr entre todos, los objetivos deseados (alto punto de ruptura y eficiencia).

8.4.1 Estimador inicial ($\widehat{\boldsymbol{\beta}}_0$)

Este estimador se estudiará en la próxima sección, pero se requiere que

- $\widehat{\boldsymbol{\beta}}_0$ sea regresión, escala y afín equivariante.

El motivo es similar al analizado en el capítulo anterior(7.3.5) al tratar la equivarianza del M-estimador de regresión con una regresión L^1. Si $\widehat{\boldsymbol{\beta}}_0$ es regresión, escala y afín equivariante, esto arrastra que $\widehat{\sigma}$ sea regresión y afín invariante, y escala equivariante. Y finalmente esto implica que $\widehat{\boldsymbol{\beta}}$ sea también regresión, escala y afín equivariante (como $\widehat{\boldsymbol{\beta}}_0$).

- El punto de ruptura $\varepsilon^*(\widehat{\boldsymbol{\beta}}_0)$ sea alto.

Ya que el punto de ruptura del estimador principal $\varepsilon^*(\widehat{\boldsymbol{\beta}}_{MM})$ es siempre $\geq \varepsilon^*(\widehat{\boldsymbol{\beta}}_0)$.

- $\widehat{\boldsymbol{\beta}}_0$ no requiera estimación previa de escala

Pues justamente se lo quiere usar para obtener $\widehat{\sigma}$
Sin embargo no será importante la eficiencia de $\widehat{\boldsymbol{\beta}}_0$.

8.4.2 Estimador previo de escala ($\widehat{\sigma}$)

Se utilizará un M-estimador de escala definido por

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0(\frac{r_i(\widehat{\boldsymbol{\beta}}_0)}{c_0\sigma}) = 0.5$$

donde:

- ρ_0 es una función rho acotada
- $E_F\left[\rho_0(\frac{U}{c_0\sigma})\right] = 0.5$

De esta manera nos aseguramos que $\widehat{\sigma}_n \xrightarrow{\mathbf{p}} \sigma$, el desvío estándar σ, cuando los u_i tienen distribución normal. Por ejemplo si ρ_T es la bicuadrada de Tukey con $k = 1$, dada por

$$\rho_T(x) = \begin{cases} 1-(1-x^2)^3 & \text{si } |x| \leq 1 \\ 1 & \text{si } |x| > 1 \end{cases}$$

entonces $c_0 = 1.56$.

- $\varepsilon^*(\widehat{\sigma}) = 0.5$

Pues δ en la expresión del M-estimador es 0.5.

8.4.3 Estimador principal ($\widehat{\boldsymbol{\beta}}_{MM}$)

- Se define

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\rho(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}}) \tag{118}$$

donde ρ es una función rho acotada, y siendo $\rho' = \psi$, también se tiene

$$\sum_{i=1}^{n}\psi(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}})\mathbf{x}_i = \mathbf{0} \tag{119}$$

En principio el estimador se obtendría buscando el mínimo absoluto de la (118). Sin embargo en general las dificultades de cálculo son muy grandes.

Otra alternativa sería resolver la (119), pero como ψ es re-descendente, existirán varias soluciones que corresponden a diversos mínimos locales de (118). Por suerte, como demostró Yohai(1987), no es necesario encontrar el mínimo absoluto de (118) para obtener un estimador con alto punto de ruptura y alta eficiencia. Pero se requiere como condición que

- $\rho \leq \rho_0$
- $\widehat{\boldsymbol{\beta}}_{MM}$ es una solución cualquiera de la (119) que cumple $L(\widehat{\boldsymbol{\beta}}_{MM}) \leq L(\widehat{\boldsymbol{\beta}}_0)$.

Entonces $\widehat{\boldsymbol{\beta}}_{MM}$ se llamará un MM-estimador, y cumplirá:

1. $\widehat{\boldsymbol{\beta}}_{MM} \xrightarrow{\mathbf{p}} \boldsymbol{\beta}$, o sea, consistente
2. $\varepsilon^*(\widehat{\boldsymbol{\beta}}_{MM}) \geq \varepsilon^*(\widehat{\boldsymbol{\beta}}_0)$, por eso se exigió que $\widehat{\boldsymbol{\beta}}_0$ tuviese alto punto de ruptura.
3. $eff(\widehat{\boldsymbol{\beta}}_{MM})$ coincide con la eficiencia del mínimo global de (118). Entonces no se necesitará buscar el mínimo absoluto.
4. En el proceso numérico se parte de $\widehat{\boldsymbol{\beta}}_0$ y se demuestra que en cada iteración $L(\boldsymbol{\beta})$ decrece, lo que asegura el cumplimiento de $L(\widehat{\boldsymbol{\beta}}_{MM}) \leq L(\widehat{\boldsymbol{\beta}}_0)$.

Remark 41 *Restan elegir las funciones ρ_0 y ρ. Si la bicuadrada de Tukey con $k = 1$ es*

$$\rho_T(x) = \begin{cases} 1 - (1 - x^2)^3 & si\ |x| \leq 1 \\ 1 & si\ |x| > 1 \end{cases}$$

se tomará

$$\rho_0(r) = \rho_T(\frac{r}{c_0}) \qquad y \qquad \rho(r) = \rho_T(\frac{r}{c_1})$$

donde $c_0 = 1.56$ para que $\widehat{\sigma}_n \xrightarrow{\mathbf{p}} \sigma$ en el caso de residuales $N(0; \sigma)$. Y para que $\rho \leq \rho_0$ se pedirá que $c_1 \geq c_0$. Como c_1 es el k de la ρ_T o ψ_T de Tukey, vale la tabla de eficiencias analizada en (6.3) para el M-estimador de posición con la ψ_T de Tukey (se transcribe una parte)

$c_1 = k$	eff
3.14	**0.80**
3.44	**0.85**
3.88	**0.90**
4.68	**0.95**
7.04	**0.99**
∞	1

Sin embargo hay un compromiso entre eficiencia y sesgo. Por eso si bién podría interesar una $eff = 0.95$, lo que supone tomar $c_1 = 4.68$, el sesgo es demasiado grande. La sugerencia es tomar $c_1 = 3.44$, que da un sesgo menor, con una eficiencia razonable de 0.85.

8.5 Estimadores basados en escala robusta

Se estudiarán una familia de estimadores de regresión que no dependen de un estimador previo de escala. Justamente uno de ellos servirá como estimador inicial $\widehat{\boldsymbol{\beta}}_0$ en el MM-estimador, y con sus residuos se obtendrá también el estimador previo de escala $\widehat{\sigma}$.

Si $\mathbf{r}(\boldsymbol{\beta}) = (r_1(\boldsymbol{\beta}), r_2(\boldsymbol{\beta}), \cdot\cdot, r_n(\boldsymbol{\beta}))$ es un vector de residuos, el estimador de cuadrados mínimos se puede expresar

$$\widehat{\boldsymbol{\beta}}_{LS} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} SD(\mathbf{r}(\boldsymbol{\beta})) \text{ donde } SD(\mathbf{r}(\boldsymbol{\beta})) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} r_i^2(\boldsymbol{\beta})}$$

y en la regresión L^1 como

$$\widehat{\boldsymbol{\beta}}_{L^1} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} |r_i(\boldsymbol{\beta})| = \arg\min_{\boldsymbol{\beta}} MD(\mathbf{r}(\boldsymbol{\beta})) \text{ donde } MD(\mathbf{r}(\boldsymbol{\beta})) = \frac{1}{n}\sum_{i=1}^{n} |r_i(\boldsymbol{\beta})|$$

notar que en ambos casos el estimador surge de minimizar una medida del tamaño de los residuos: SD en cuadrados mínimos, y MD con la regresión L^1, que no son robustas. Además, ambos estimadores no requieren un estimador previo de escala.

Con el fin de conservar esta última característica, pero sin embargo obtener un estimador $\widehat{\boldsymbol{\beta}}$ robusto, se propone utilizar como medida del tamaño de los residuos un estimador de escala

$$\widehat{\sigma}(\mathbf{r}(\boldsymbol{\beta})) \text{ robusto y equivariante de escala}$$

y definir

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \widehat{\sigma}(\mathbf{r}(\boldsymbol{\beta})) \tag{120}$$

Estos estimadores son regresión, escala y afín equivariantes.

8.5.1 S estimadores

Se obtienen cuando como estimador $\widehat{\sigma}(\mathbf{r}(\boldsymbol{\beta}))$ en la (120), se usa un M-estimador de escala, definido para cada $\mathbf{r}(\boldsymbol{\beta})$ por

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\sigma}\right) = \delta \tag{121}$$

donde ρ es una función rho acotada. El estimador de $\boldsymbol{\beta}$ que resulta de la (120) fué llamado S-estimador por Yohai y Rousseeuw en 1984, y es justamente el recomendado como estimador inicial $\widehat{\boldsymbol{\beta}}_0$ en el MM-estimador(usando la bicuadrada de Tukey como ρ en la 121).

Punto de ruptura El estimador $\widehat{\sigma}(\mathbf{r}(\boldsymbol{\beta}))$ tiene como punto de ruptura el usual para los M-estimadores de escala $\varepsilon^* = \min\{\delta, 1-\delta\}$. Y para el S-estimador de $\boldsymbol{\beta}$, el punto de ruptura tiene la misma cota que los estimadores equivariantes de regresión, o sea

$$\boldsymbol{\varepsilon}^* \leq \frac{1}{n}\left[\frac{n-k^*-1}{2}\right] = \frac{m^*_{\max}}{n}$$

donde también

$$k^* = \max\{\#(\mathbf{x}_i'\boldsymbol{\gamma} = 0 : \gamma \in \mathbb{R}^p, \boldsymbol{\gamma} \neq \mathbf{0}\}$$

y esta cota se alcanza tomando

$$\delta = \frac{m^*_{\max} + \gamma}{n} \quad \text{con } \gamma \in (0,1)$$

Y como siempre, si $\mathbf{X}$ está en "posición general", $k^* = p-1$, y entonces la cota máxima es $\frac{1}{n}\left[\frac{n-p}{2}\right]$, que vale aproximadamente 0.5 para n grande.

Eficiencia Con la solución $\widehat{\boldsymbol{\beta}}$ del S-estimador se obtiene el $\widehat{\sigma}_{\min}$. Y se usó una función ρ(pensada como función rho de escala) para el M-estimador de escala de(121).

Que tal si ahora se busca el "M-estimador de $\boldsymbol{\beta}$ con estimación previa de escala", pero con la misma función ρ(pensada ahora como función rho de un M-estimador de regresión), y se aprovecha $\widehat{\sigma}_{\min}$ como estimador previo de escala. La ecuación sería

$$\widehat{\boldsymbol{\beta}}^* = \arg\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}_{\min}}\right) \tag{122}$$

Notar que $\widehat{\boldsymbol{\beta}}^*$ debe ser igual a $\widehat{\boldsymbol{\beta}}$. En efecto, si $\widehat{\boldsymbol{\beta}}^*$ fuese otro, como con el se logra el mínimo de (122), deberá ser (usando 121 con $\widehat{\boldsymbol{\beta}}$ y $\widehat{\sigma}_{\min}$)

$$\sum_{i=1}^{n} \rho\left(\frac{r_i(\widehat{\boldsymbol{\beta}}^*)}{\widehat{\sigma}_{\min}}\right) < \sum_{i=1}^{n} \rho\left(\frac{r_i(\widehat{\boldsymbol{\beta}})}{\widehat{\sigma}_{\min}}\right) = n\delta$$

pero como ρ es monótona, debe existir un $\widehat{\sigma}_{\min\,\min} < \widehat{\sigma}_{\min}$ en que se verifique

$$\sum_{i=1}^{n} \rho\left(\frac{r_i(\widehat{\boldsymbol{\beta}}^*)}{\widehat{\sigma}_{\min\,\min}}\right) = n\delta$$

pero entonces $\widehat{\sigma}_{\min}$ no sería el mínimo.

De acuerdo a esto el S-estimador $\widehat{\boldsymbol{\beta}}$ es también un M-estimador con la escala estimada simultaneamente, y función psi $\psi = \rho'$, o sea

$$\begin{cases} \sum_{i=1}^{n} \rho'\left(\frac{r_i(\boldsymbol{\beta})}{\sigma}\right)\mathbf{x}_i = \mathbf{0} \\ \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\sigma}\right) = \delta \end{cases}$$

y entonces si $\widehat{\sigma_{n\min}} \xrightarrow{\mathbf{p}} \sigma$ y ρ es dos veces derivable, serán válidas también

$$\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}; v(\mathbf{XX})^{-1}) \quad \text{con} \quad v = \sigma^2 \frac{E_F\left[\psi^2(\frac{u}{\sigma})\right]}{E_F^2\left[\psi'(\frac{u}{\sigma})\right]}$$

Además la velocidad de convergencia de $\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}$ es la usual de $n^{-\frac{1}{2}}$.

Pero lamentablemente los S-estimadores no pueden tener simultaneamente alto punto de ruptura y alta eficiencia. Hossjer(1992) demostró que un S-estimador con $\varepsilon^* = 0.5$ tiene eficiencia asintótica bajo errores distribuídos normalmente ≤ 0.33. Y cuando se usa una ρ bicuadrada de Tukey, la eficiencia es de aproximadamente 0.29, casi la máxima. Sin embargo debido a su buen punto de ruptura es muy útil como estimador inicial $\widehat{\boldsymbol{\beta}}_0$ para el MM-estimador.

Finalmente como en el S-estimador se usa una ρ re-descendente, observaciones con Y_i muy grande no tienen efecto sobre el estimador.

LMS estimador Otro ejemplo de S-estimador consiste en minimizar la mediana del valor absoluto de los residuos. Y es el llamado LMS-estimador ("least median of squares"), que fué el primer estimador de este tipo, introducido por Hampel(1975) y Rousseeuw(1984). Se obtienen cuando como estimador $\widehat{\sigma}(\mathbf{r}(\boldsymbol{\beta}))$ en la (120), se usa un M-estimador de escala, definido para cada $\mathbf{r}(\boldsymbol{\beta})$ por

$$\frac{1}{n}\sum_{i=1}^{n}\rho(\frac{r_i(\boldsymbol{\beta})}{\sigma}) = \delta \quad \text{donde } \rho(u) = I(|u| > 1) \quad \text{y} \quad \delta = 0.5$$

y en realidad es el M-estimador de escala analizado en el ejemplo-30. Concretamente el LMS-estimador se define mediante

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Med(|r_i(\boldsymbol{\beta})|)$$

El estimador de escala $Med(|r_i(\boldsymbol{\beta})|)$ tiene $\varepsilon^* = 0.5$, y el correspondiente LMS-estimador de $\boldsymbol{\beta}$, tiene la misma cota para el punto de ruptura que cualquier S-estimador

$$\boldsymbol{\varepsilon}^* \leq \frac{1}{n}\left[\frac{n - k^* - 1}{2}\right] = \frac{m^*_{\max}}{n}$$

Pero como ρ es discontinua, la velocidad de convergencia del LMS-estimador es menor, concretamente según Davies(1990), $\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}$ converge según $n^{-\frac{1}{3}}$. Luego, es muy ineficiente para n grande.