2 E.
15

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

APLICACION DE TECNICAS DE REGRESION ROBUSTA

TRABAJO ESCRITO

Que para obtener el título de:

A C T U A R I O

P r e s e n t a :

Leticia Eugenia Gracia-Medrano Valdelamar

México, D. F. 1984

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

DERECHOS RESERVADOS © PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis está protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

C O N T E N I D O

N O T A C I O N

De aquí en adelante, se considera el modelo de regresión lineal siguiente:

$$Yi = x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots + x_{Ki}\beta_K + \varepsilon i \qquad i = 1,\ldots,n$$

donde:

$\underset{\sim}{\beta} = (\beta_1, \beta_2, \ldots, \beta_K)$ es un vector de parámetros desconocidos que se desea estimar.

n es el tamaño de la muestra,

x_{ji} es la observación i-ésima de la variable explicativa j,

εi es un error aleatorio. (con determinada función de distribución).

En forma matricial puede escribirse como

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} ,$$

donde

$\underset{\sim}{Y}$ es un vector de dimensiones nx1,

X es la matriz de dimensiones nxK,

cuyo renglón i-ésimo es de la forma

$xi = (x_{1i}, x_{2i}, x_{3i}, \ldots, x_{Ki})$,

$\underset{\sim}{\beta}$ es el vector $(\beta_1, \beta_2, \ldots, \beta_K)'$,

$\underset{\sim}{\varepsilon}$ es un vector aleatorio de dimensiones nx1.

El modelo

$$Yi = \beta_0 + x_{i1}\beta_1 + \varepsilon_i \qquad i = 1,\ldots,n$$

en un caso especial del anterior y se le conoce como modelo de regresión lineal simple.

El estimador de $\underset{\sim}{\beta}$ se denota como $\hat{\underset{\sim}{\beta}}$, ó se le da el nombre del autor que los propone.

El valor de la Y estimada es:

$$\hat{Y}i = xi_1 \hat{\beta}_1 + xi_2 \hat{\beta}_2 + \ldots + x_{1K} \hat{\beta}_K, \qquad i = 1,\ldots,n.$$

Se le llama residual i-ésimo a la diferencia

$$ei = Yi - \hat{Y}i.$$

En general se denota a los estimadores de la escala y de la desviación estándar con una S, una S seguida de algún identificador, o con una $\hat{\sigma}$.

A los estimadores de la media μ se les denotará con $\hat{\mu}$.

Cuando se escribe $f(z) = \min$, significa que f es una función objetivo a minimizar.

I N T R O D U C C I O N

Las inferencias estadísticas están basadas sólo parcialmente en las observaciones. Otra parte importante la forman las suposiciones subyacentes. Aún en los casos más simples, existen suposiciones implícitas o explícitas acerca de la aleatoriedad, independencia o distribución. Estas suposiciones no siempre se satisfacen exactamente. Este incumplimiento podría justificarse con el principio de estabilidad: un error menor en el modelo matemático debe causar sólo pequeños errores en las conclusiones finales. Pero esto no siempre sucede.

En las últimas décadas se ha visto que la mayoría de los procedimientos estadísticos (en particular aquellos optimizados para una distribución normal) son excesivamente sensibles a pequeñas desviaciones de las suposiciones. Es por esto que una gran cantidad de procedimientos robustos han sido propuestos.

La palabra robusto se interpreta como: insensibilidad a pequeñas desviaciones de las suposiciones.

Cuando la distribución de la población se desvía ligeramente de la supuesta en el modelo, se necesita de una robustez del tipo distribucional, que es la más importante y la más conocida. Bastante menos se sabe de lo que sucede cuando otras suposiciones no son satisfechas y de las medidas a tomar para hacer frente a estas violaciones.

En el modelo de regresión lineal

$$\underset{\sim}{Y} = X\beta + \underset{\sim}{\varepsilon}$$

donde

$\underset{\sim}{Y}$ es un vector de (nx1) observaciones;

X es una matriz de (nxk) de términos conocidos
$\underset{\sim}{\beta}$ es un vector de (kx1) parámetros
y $\underset{\sim}{\varepsilon}$ es un vector de (nx1) errores;
cuando $\underset{\sim}{\varepsilon}$ se distribuye como normal, el método de mínimos cuadrados (MC) produce un estimador de $\underset{\sim}{\beta}$, $\underset{\sim}{\hat{\beta}}_{MC}$, con buenas propiedades estadísticas ($\underset{\sim}{\hat{\beta}}_{MC}$ es insesgado, lineal de varianza mínima que al coincidir con el estimador de máxima verosimilitud resulta ser además consistente, eficiente, suficiente). Pero cuando los errores provienen de una población cuya distribución es no normal, en particular una con colas más pesadas que la normal, entonces el método de MC puede no ser adecuado, ya que este tipo de distribuciones genera usualmente observaciones discrepantes y éstas a su vez pueden influir fuertemente en el estimador de MC. En efecto, estas observaciones discrepantes tienden a modificar el ajuste de tal manera que a menudo son difíciles de identificar ya que sus residuales se hacen pequeños de manera artificial.

El siguiente ejemplo tomado de Huber (1981 pág. 153) - ilustra muy bien la falta de robustez distribucional del procedimiento de MC.

Si se ajusta una línea recta a los seis puntos, cuyas -- coordenadas aparecen en la tabla I.1, el ajuste de MC (ajuste 1) da la recta de la figura I.1 Los valores de los residuales de este ajuste (Tabla I.1) dan la impresión de que todo está bien, ninguno de los residuales $e_i=y_i - \hat{y}_i$ es excepcionalmente grande comparado con la desviación estándar estimada (tabla I.1). Pero observando con mayor atención la figura I.1 puede sospecharse que algo puede estar mal con el punto 1 (que tiene el mayor residual) ó con el punto 6. Si el punto 6 se elimina del ajuste se obtiene el ajuste 2 (figura I.2). Pero posiblemente un modelo lineal es inapropiado y los puntos debieron ajustarse mediante una parábola (figura I.3).

TABLA I.1

			AJUSTE 1		AJUSTE 2		AJUSTE 3	
PUNTO	x	y	$\hat{y}$	$y-\hat{y}$	y	$y-\hat{y}$	y	$y-\hat{y}$
1	-4	2.48	0.39	2.09	2.04	0.44	2.23	0.25
2	-3	0.73	0.31	0.42	1.06	-0.33	0.99	-0.26
3	-2	-0.04	0.23	-0.27	0.08	-0.12	-0.09	-0.13
4	-1	-1.44	0.15	-1.59	-0.90	-0.54	-1.00	-0.44
5	0	-1.32	0.07	-1.39	-1.87	0.55	-1.74	0.42
6	10	0	-0.75	0.75	-11.64	(11.64)	0.01	-0.01

d.s. $\hat{\sigma} = 1.55$ $\hat{\sigma} = 0.55$ $\hat{\sigma} = 0.41$

$e_{max/\hat{\sigma}} = 1.35$ $e_{max/\hat{\sigma}} = 1.00$ $e_{max/\hat{\sigma}} = 1.08$

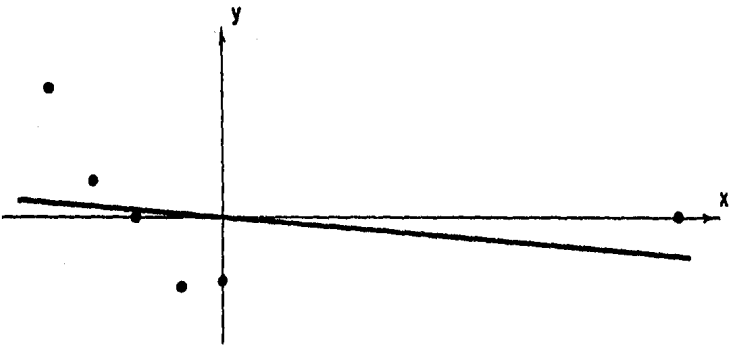


FIGURA I.1 AJUSTE DE MINIMOS CUADRADOS

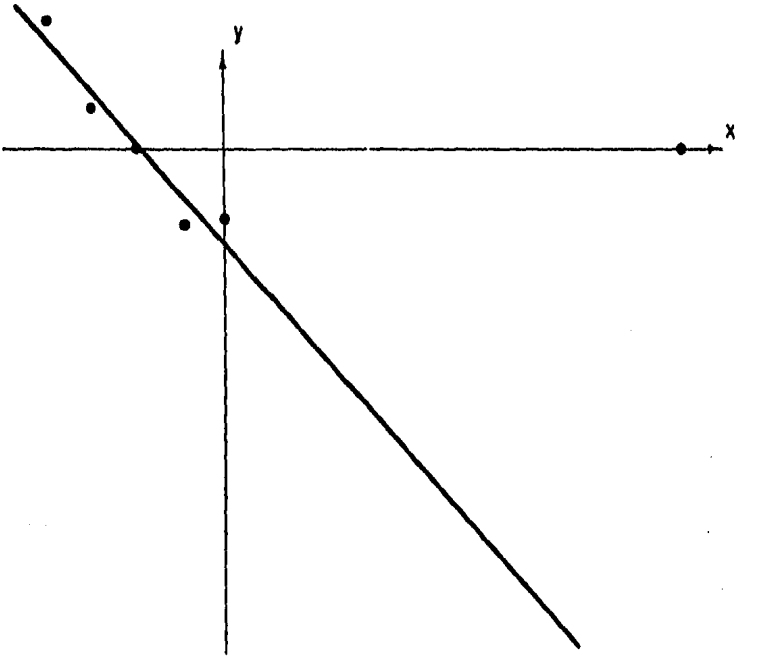


FIGURA I.2 AJUSTE DE MINIMOS CUADRADOS ELIMINANDO EL PUNTO 6.

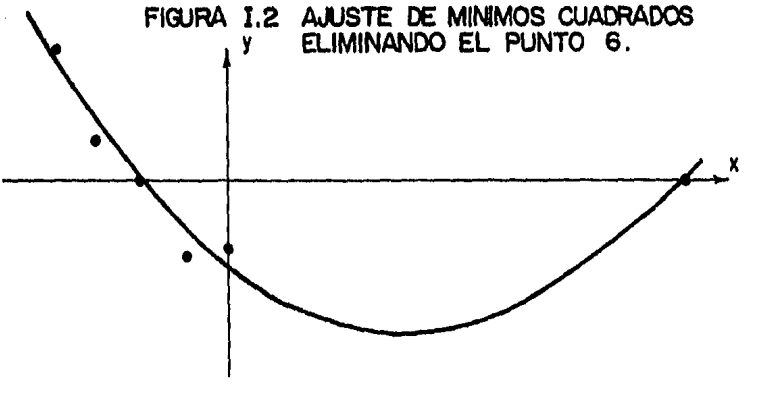


FIGURA I.3 AJUSTE MEDIANTE UNA PARABOLA

Con los datos disponibles es difícil distinguir entre estas tres posibilidades. Guiándose por el tamaño de la desviación estándar estimada ($\hat{\sigma}$) se eligiría la tercera variante.

Este ejemplo es artificial, los puntos fueron generados al tomar la recta $y = -2 - x$, y a los puntos 1,2,3,4 y 5 se les sumaron errores aleatorios normales (con media 0 y desviación estándar 0.6) y al punto 6 se le sumó un error más grueso.

Entonces el ajuste 2 es el apropiado. En este caso se involucran sólo dos parámetros y las gráficas dan evidencia del problema del punto 6. En casos más complicados de regresión múltiple, detectar este tipo de problemas es bastante más difícil.

En este ejemplo al no cumplirse la hipótesis de homoscedasticidad, $\hat{\beta}_{MC}$ resulta ser un mal estimador.

Podría pensarse que en vez de utilizar un método robusto sería igualmente bueno y más sencillo seguir estos dos pasos:

(1) Limpiar los datos, mediante algún criterio para eliminar las observaciones discordantes.

(2) Usar los procedimientos clásicos con los datos restantes.

Pero no es así, Huber (1981) da las siguientes razones:

(a) Es difícil hacer los pasos completamente por separado. Por ejemplo en los casos de regresión múltiple es difícil hallar a las observaciones discrepantes, a menos de que se cuente con estimadores robustos de los parámetros.

(b) Si el conjunto original de observaciones proviene de una población normal, con algunos errores gruesos, el conjunto de datos "limpio" no es normal (habrá falsas eliminaciones y falsas retenciones), y la situación es peor que si los datos proviniesen de una población no normal.

La teoría clásica normal no es aplicable a estas muestras "limpias" y tal vez seguir este procedimiento de los dos pasos sea más difícil que seguir un procedimiento robusto.

(c) Es un hecho empírico que los mejores procedimientos de eliminación de observaciones discrepantes no alcanzan a obtener los resultados de los procedimientos robustos. Estos últimos aparentemente son mejores porque hacen una transición suave entre una completa aceptación y un completo rechazo de la observación.

Si se cuenta con un modelo paramétrico que es una buena aproximación a la situación real subyacente, pero que no se puede y no se supondrá ser exactamente el correcto, cualquier procedimiento estadístico debe satisfacer (Huber 1981).

(1) Tener una eficiencia razonablemente buena en el modelo supuesto.

(2) Debe ser robusto en el sentido de que pequeñas desviaciones de las suposiciones del modelo deben afectar sólo ligeramente su funcionamiento.

(3) Algunas grandes desviaciones del modelo no deben causar una catástrofe.

En este trabajo se presentan: en el Capítulo 1, estimadores robustos de los coeficientes de regresión del tipo L, R y M, en el Capítulo 2, cinco técnicas de regresión robusta, algunas de ellas basadas en estimadores Lp, en el Capítulo 3, dos estudios que comparan a algunos estimadores robustos, con el fin de dar ideas sobre cual de ellos es conveniente utilizar en determinada situación y finalmente en los Capítulos 4 y 5, se tratan las inconveniencias y precauciones que deben tenerse en el uso de los estimadores robustos.

Cabe advertir que en todo este trabajo, al referirse a estimadores robustos, se alude únicamente por ser el tema a desarrollar a los estimadores en regresión, omitiendo tratar a los de algún otro tipo.

HISTORIA DE LA ESTIMACION ROBUSTA

En el siglo XVIII la palabra robusto en la lengua inglesa era utilizada para referirse a alguien que era fuerte, rudo y vulgar. Dentro de nuestro contexto este significado ha evolucionado con el tiempo.

Es en el año de 1953 que Box la utiliza dándole un sentido estadístico por vez primera. Se entenderá por robustez la

"insensibilidad de los procedimientos a desviaciones de las suposiciones".

Los científicos trataron con procedimientos que podrían llamarse "robustos" desde el siglo XVIII. Por ejemplo, casi medio siglo antes de que Legendre anunciara su principio de Mínimos Cuadrados (en 1805), Ruggiero Giussepe Boscovich (1711-1787) formuló y aplicó el principio en el que, dadas más de dos observaciones de x y de y, relacionadas por una función lineal de la forma,

$$y = \beta_0 + \beta_1 x \quad ,$$

los valores de β_0 y β_1 se determinaban de tal manera que la recta obtenida fuese la más acorde con las observaciones, según estas dos condiciones:

$$(1) \quad \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x) = 0$$

$$(2) \quad \sum_{i=1}^{n} | y_i - \beta_0 - \beta_1 x | = \min_{\beta_0 , \beta_1} !$$

Boscovich desarrolló este criterio entre 1755 y 1757 para determinar, junto con su compañero Maire, la figura de la tie

rra a partir de las longitudes de los meridianos y los "seconds pendulums" a diferentes latitudes. En 1760 Boscovich propone un procedimiento geométrico para hallar la solución al problema dado por las condiciones anteriores.

Este procedimiento resulta ser insensible a observaciones extremas.

En 1789 Laplace en su Segunda Memoria de la Figura de la tierra adopta el criterio de Boscovih para determinar la recta de mejor ajuste y da una solución analítica pero no hace mención alguna acerca de Boscovich.

Legendre en el primer trabajo publicado acerca de MC en 1805, considera la existencia de observaciones discrepantes y escribe: "si dentro de los errores hay alguno que parezca demasiado grande como para ser admitido, entonces las ecuaciones asociadas a estos errores deben ser rechazadas, por provenir de experimentos con fallas, las incógnitas deberán ser determinadas por medio de las ecuaciones restantes que dan errores bastante más pequeños".

Otro trabajo matemático relacionado con la estimación robusta es el realizado por Laplace, en 1818, acerca de la distribución de la mediana.

El siguiente problema conectado con la estimación robusta fue el tratamiento para las observaciones discordantes.

El primer criterio para eliminar observaciones discordantes fue publicado por Peirce, un matemático y astrónomo de Harvard, en 1859. En el trabajo de Peirce y en otros relacionados con el mismo tema, los autores no se ocupaban de las propiedades de los estimadores resultantes, ellos implícitamente suponían que después de realizar la prueba para las observaciones discordantes, la estimación podía hacerse sin considerar la información que se podía haber perdido.

En 1856 Airy, un astrónomo, hace la primera crítica al

uso de los criterios para rechazar las observaciones discordantes y escribe: "Ninguna regla para la exclusión de observaciones puede obtenerse por ningún proceso basado únicamente en la consideración de la discordancia de estas observaciones"

Para la segunda mitad del siglo XIX, el principio de Mínimos Cuadrados Ponderados se había convertido en un tema estándar en la literatura de la teoría de los errores y era frecuente la práctica (al menos en investigaciones astronómicas) de la ponderación desigual de las observaciones, de acuerdo a estimaciones científicas (a menudo subjetivas) del error probable de la observación.

Podría decirse que alrededor de 1885 ya se tomaban precauciones en el uso de la media muestral, algunas veces se ponderaba y otras se usaba después de eliminar las observaciones discordantes. Sin embargo, aún se usaba la media muestral únicamente.

A partir de 1885 empieza un período más activo y de más innovaciones en el área de la estadística matemática.

Simon Newcomb (1878-1909) introduce por primera vez una mezcla de densidades normales como un modelo para las distribuciones con colas pesadas, y explora este modelo para obtener un estimador de localización que fuese más robusto que la media muestral.

Se considera a Newcomb como el más grande de los astrónomos del siglo XIX pues determinó constantes que son aceptadas aún actualmente. Newcomb hizo uso frecuente de medias ponderadas en la determinación de constantes astronómicas. También rechazó observaciones discordantes cuando fue necesario, pero basado únicamente en evidencias externas o desviaciones verdaderamente grandes.

En 1886 Newcomb critica el uso excesivo de los criterios de eliminación de observaciones discordantes, presenta su modelo de mezclas y desarrolla un estimador basándose en la teoría de decisión Bayesiana que da "menos peso a las obser

vaciones discordantes".

Las combinaciones lineales de estadísticas de orden fueron consideradas quizá por primera vez en un anâlisis matemático extenso por Laplace. En su segundo suplemento a su Theorie Analytique des Probabilite's (1818), Laplace considera el problema de regresión lineal siguiente:

$$yi = \beta xi + \varepsilon i$$

donde yi y xi son conocidos;

β debe ser estimada;

los errores ε i tienen una distribución continua y simétrica.

Laplace busca el estimador que minimice la suma de los valores absolutos de los residuales (siguiendo el criterio de Boscovich). Considera el caso en el que xi = 1, que se reduce a estimar la mediana de las yi's y encuentra la densidad de este estimador y muestra que esta densidad se aproxima a la normal cuando el tamaño de la muestra crece y da la condición suficiente y necesaria sobre la distribución del error para que la mediana tenga una varianza asintótica menor que la de la media muestral.

Más tarde en el siglo XIX, Galton en 1875 y Edgeworth, entre 1887-1888 sugieren el uso de la mediana en situaciones donde la distribución tuviese colas más pesadas que la normal.

Específicamente Edgeworth en 1888 utiliza los resultados de Laplace para concluir que la mediana podría ser mejor que la media muestral para el caso en el que la distribución sea una mezcla de distribuciones normales del tipo desarrollado por Newcomb.

Edgeworth en 1886 se da cuenta de que la mediana aventaja a la media muestral cuando se trata de datos correlacionados serialmente.

Estimadores lineales más complicados aparecen en 1889 cuando Galton sugiere estimar la media y la desviación están dar de una distribución normal mediante las expresiones siguientes:

$$\hat{\mu} = \frac{\xi_p \; x_{(nq)} - \xi_q \; x_{(np)}}{\xi_p - \xi_q}$$

$$\hat{\sigma} = \frac{x_{(np)} - x_{(nq)}}{\xi_p - \xi_q}$$

donde: ξ_p y ξ_q son los p y q percentiles de una normal estandarizada;

$x_{(np)}$ y $x_{(nq)}$ son los p y q percentiles muestrales

p y q son arbitrarios pero fijos $(o < p < q < 1)$.

El siguiente trabajo que aparece acerca de estadísticas de orden fue el de Karl Pearson en 1902; al dar la distribución conjunta de dos estadísticas de orden consecutivas y encontrar el valor esperado de su diferencia.

En 1920 Dantell en su trabajo "Observaciones ponderadas de acuerdo a su orden" escribe: Además de la media que resulta después de la eliminación de observaciones discordantes, debemos inventar otros (estimadores) en los que los pesos sean asignados de acuerdo al orden. De hecho la media, la mediana, la media recortada, pueden ser vistas como cálculos mediante procesos en los que los pesos son múltiplos de funciones que dependen del orden de las observaciones".

Más adelante, Winsor en 1941 propone un procedimiento que ahora lleva su nombre, en el que una observación de la que se sospecha sea discordante, no debe rechazarse por completo sino que su valor original debe sustituirse por el valor más cercano considerado como no discordante.

Recientemente, alrededor de los años sesentas, a raíz de los adelantos en la computación, la estimación robusta ha tenido un resurgimiento que se vió iniciado cuando Tukey y el Grupo de Investigación Estadística de Princeton propagan los inconvenientes de los estimadores clásicos y establecen propiedades para estimadores alternativos.

El interés de Tukey en este campo se inicia con la discusión de la media "Winzorizada" y la media recortada en un artículo relacionado con el rechazo de las observaciones discordantes. (Tukey, 1962).

Tukey trabaja con procedimientos resistentes, entendiéndose por resistente que el valor del estimador sea insensible a pequeños cambios en la muestra (sin importar la distribución subyacente).

Hampel en 1974 introduce el uso de la función de influencia y sus propiedades para derivar nuevos estimadores con ciertas características robustas.

En cuanto a la robustez en regresión, Andrews y Huber han hecho grandes contribuciones. En particular Huber (1973) ha dado resultados asintóticos de la distribución de los estimadores de regresión robustos y considerado las funciones de influencia descendientes.

Bickel (1976) ha hecho importantes aportaciones en cuanto a métodos y pruebas robustas.

Con este resumen de la historia de la estimación robusta, se da uno cuenta de que este tipo de estimadores se necesitaron ya desde el siglo XVIII, principalmente en la astronomía.

El desarrollo de estos métodos se vió entorpecido por las dificultades existentes en su cálculo, pero actualmente con ayuda de las computadoras se ha podido de nuevo avanzar en ellos.

CAPITULO 1

ESTIMADORES ROBUSTOS

1.1 ESTIMADORES -L

Se conoce como estimadores L a aquellos que involucran combinaciones lineales de estadísticas de orden, o más generalmente alguna función h de ellas.

$$T = \sum_{i=1}^{n} ai \quad h\ (x\ (i)\ ).$$

donde x (i) i = 1,...,n son las estadísticas de orden de la muestra aleatoria $x_1, x_2, \ldots, x_n$, y satisfacen

$$x(1) \leq x(2) \leq \cdots \leq x\ (n).$$

Un ejemplo de estimador de este tipo es la mediana.

Bickel (1973) extiende los estimadores L a los modelos lineales, pero son Koenker y Basset (1978) quienes extienden los cuantiles a los modelos de regresión.

Sea $0<\alpha<1$ y

$$\rho_\alpha(x) = \begin{cases} \alpha x & x \geq 0 \\ (\alpha-1)\, x & x < 0 \end{cases}$$

Entonces el cuantil α de regresión está definido como la solución $\hat{\beta}_\alpha$ del siguiente problema de minimización.

$$\sum_{i=1}^{n} (yi - \underset{\sim}{x}i\ \hat{\beta}_\alpha\ ) = \min$$

Los cuantiles $\hat{\beta}_\alpha$ de regresión se obtienen mediante la resolución del problema de programación lineal siguiente:

Si ei = ui-vi donde ui, vi $\geq$ 0

Minimizar

$$\sum_{i=1}^{n} \rho_\alpha(ui) + \sum_{i=1}^{n} \rho_\alpha(vi)$$

Sujeto a

$$yi = \sum_{j=1}^{K} xij\,\hat{\beta}_{\alpha}j + ui - vi$$

$\hat{\beta}_{\alpha}j$ sin restricción de signo;
$ui, vi \geq 0$

El ejemplo más simple es la mediana de regresión cuando $\alpha = 1/2$, que determina un hiperplano π, la mitad de las observaciones tienen una distancia a π no negativa y la otra mitad tienen distancia a π no positiva.

Dado los cuantiles de regresión, cualquier combinación lineal de éstos puede construirse. Dos ejemplos de estimadores - L son:

La trimedia igual a
$1/4\ \hat{\beta}.25 + 1/2\ \hat{\beta}.50 + 1/4\ \hat{\beta}.75$
y el estimador de Gastwirth es:
$0.3\ \hat{\beta}.33 + 0.4\ \hat{\beta}.50 + 0.3\ \hat{\beta}.66$

1.2 ESTIMADORES - R

Los procedimientos para obtener este tipo de estimadores están basados en los rangos de las observaciones, y es por eso que se les denota con una R.

El procedimiento general reemplaza un factor de cada término en la función objetivo de MC.

$S(\underset{\sim}{\beta}) = \sum_{i=1}^{n} (yi - \underset{\sim}{x}i\underset{\sim}{\beta})^2$ por su rango correspondiente,

así, si Ri es el rango de $yi - \underset{\sim}{x}i\underset{\sim}{\beta}$, se desea minimizar

$$\sum_{i=1}^{n} (yi - \underset{\sim}{x}i\underset{\sim}{\beta})\ Ri.$$

Más en general, se reemplazan los rangos por una función de puntaje ("score function") $a_n(i)$, i=1, ...,n. La función objetivo a minimizar se convierte entonces en

$$D(y - x\underset{\sim}{\beta}) = \sum_{i=1}^{n} (yi - \underset{\sim}{x}i\underset{\sim}{\beta})\ a_n(Ri) \qquad (1.2.1)$$

Jaeckel (1972) prueba que D es una función no negativa, continua y convexa de $\underset{\sim}{\beta}$. Por ser D valuable en todas partes y derivable casi en todas partes el mínimo puede hallarse mediante un método iterativo, como es el método "steepest descent" que utiliza primeras derivadas (Luenberger 1973).

Para puntajes $a_n(i)$ tales que

$$\sum_{i=1}^{n} a_n(i) = 0$$

D es invariante en localización, ya que:

$$D(y-x\underset{\sim}{\beta} + c) = \sum_{i=1}^{n} (yi-\underset{\sim}{x}i\underset{\sim}{\beta}+ c)\ a_n\ (Ri), \text{c cualquier constante.}$$

$$= \sum_{i=1}^{n} (yi-\underset{\sim}{x}i\underset{\sim}{\beta})\ a_n\ (Ri) + c\sum_{i=1}^{n} a_n\ (Ri)$$

$$= \sum_{i=1}^{n} (yi-\underset{\sim}{x}i\underset{\sim}{\beta})\ a_n\ (Ri)$$

$$= D\ (y-x\ \underset{\sim}{\beta})$$

De aquí que la ordenada al origen no puede ser estimada usando (1.2.1). Para calcularla se hace lo siguiente:

Con los coeficientes que se pueden obtener de (1.2.1) se calculan los residuales $ei = (yi - \underset{\sim}{x}i\hat{\underset{\sim}{\beta}})$ y se procede a minimizar con respecto a α la expresión.

$$\sum_{i=1}^{n} a_n^{+}\ (Ri^{+})\ \text{signo}\ (ei - \alpha)$$

donde

Ri^{+} es el rango de $|ei - \alpha|$

an^{+} es una función de puntaje con signo.

Algunas de las funciones de puntaje más utilizadas son:

Wilcoxon: $a_n(i) = i \quad , \quad 1 \leqslant i \leqslant n\ ;$

Mediana: $a_n(i) = \begin{cases} -1, & i \leqslant (n+1)/2; \\ 1, & i > (n+1)/2; \end{cases}$

Normal: $a_n(i) = \Phi^{-1}\left(\frac{i - 1/2}{n}\right)$, $1 \leq i \leq n$;

Vander Warden: $a_n(i) = \Phi^{-1}\left(\frac{i}{n+1}\right)$, $1 \leq i \leq n$.

donde Φ es la función de distribución de la normal $(0,1)$.

1.3 ESTIMADORES - M

A la clase de estimadores que minimizan una función de los residuales ei (i = 1,...n), esto es

$$\min_{\theta} \sum_{i=1}^{n} \varphi(yi - \theta) = \min_{\theta} \sum_{i=1}^{n} \varphi(ei) \ ,$$

Se les conoce como estimadores M, la M se debe a que - éstos pueden pensarse como estimadores semejantes a los de Máxima verosimilitud. (Huber, 1981).

Para el caso de regresión, se puede escribir

$$\min_{\beta} \sum_{i=1}^{n} \varphi(yi - \underset{\sim}{x}i' \underset{\sim}{\beta}) \qquad (1.3.1.)$$

donde xi' es el i-ésimo renglón de x.

La función φ depende del tipo de distribución que se suponga para los errores.

Por ejemplo, para el caso en que los errores se distribuyan normalmente

$$\varphi(Z) = 1/2 \ Z^2, -\infty < Z < \infty$$

donde 1/2 se incluye por conveniencia.
y el método a seguir sería mínimos cuadrados.

Si los errores se distribuyeran como doble exponencial

$$f(\varepsilon i) = \frac{1}{2} e^{-|\varepsilon i|/\sigma}, \ \varphi(Z) = |Z| .$$

donde f es función de densidad de ε .

Es importante hacer notar que los estimadores M no son invariantes bajo reescalamiento, esto es, si los residuales son multiplicados por una constante, la nueva solución a (1.3.1) no será, en general, la misma que la obtenida anterior

mente.

Para evitar este problema, se prefiere resolver

$$\min_{\beta} \sum_{i=1}^{n} \varphi\left[(yi - \underset{\sim}{xi}'\underset{\sim}{\beta}) \;/S\right] = \min_{\beta} \sum_{i=1}^{n} \varphi\,(ei/S) \qquad (1.3.2)$$

donde S es un estimador robusto de la escala.

Una posible opción para S es

$$S = \frac{\text{mediana}\,|\,ei - \text{mediana}\,(ei)\,|}{0.6745}$$

donde ei denota el i-ésimo residual obtenido de una estimación inicial de $\underset{\sim}{\beta}$. En adelante se considera a S fijo.

(la constante 0.6745 se incluye para que aún bajo normalidad S sea un estimador insesgado de la escala, para n suficientemente grande).

Mediante el cálculo de las derivadas parciales de φ (considerando a φ como una función derivable y convexa) con respecto a βj (j= 0,1,...,K) en (1.3.1.), al igualar a cero, se obtiene un sistema de K+1 ecuaciones

$$\sum_{i=1}^{n} xij\left[\psi(yi - xt'\hat{\underset{\sim}{\beta}}) \;/S\right] = 0 \quad j=0,1,\ldots,K \qquad (1.3.3)$$

donde xij es la i-êsima observación de la variable j (j = 1,...K) y xi0 = 1 (i=1,...,n),

la función $\psi = \varphi'$,

además es tal que $\psi = \varphi'$ es una función impar,

esto es $\psi(Z) = -\psi(-Z)$.

Generalmente ψ es una función no lineal y el sistema (1.3.3) debe resolverse por métodos iterativos.

Si se tiene un estimador inicial $\hat{\beta}_0$ y S es un estimador de la escala, el problema puede resolverse reescribiendo (1.3.3) como sigue:

$$\sum_{i=1}^{n} xij \; \frac{\Psi(yi-\underset{\sim}{x}i'\hat{\underset{\sim}{\beta}}_o)\ /S}{(yi-\underset{\sim}{x}i'\hat{\underset{\sim}{\beta}}_o)\ /S} \; (yi-\underset{\sim}{x}i'\hat{\beta})\ /S = 0$$

que a su vez puede escribirse como

$$\sum_{i=1}^{n} xij \quad wio\ (yi-\underset{\sim}{x}i'\hat{\underset{\sim}{\beta}})\ = 0$$

donde

$$wio = \begin{cases} \dfrac{\Psi(yi-\underset{\sim}{x}i'\hat{\underset{\sim}{\beta}}o)\ /S}{(yi-\underset{\sim}{x}i'\hat{\underset{\sim}{\beta}}o)\ /S} & \text{Si } yi \neq \underset{\sim}{x}i'\hat{\underset{\sim}{\beta}}o \\ 1 & \text{Si } yi = \underset{\sim}{x}i'\hat{\underset{\sim}{\beta}}o \end{cases}$$

que en forma matricial es

$$X'\ Wo\ X\underset{\sim}{\beta} = X'\ Wo\ \underset{\sim}{y}$$

con

$$Wo = \begin{bmatrix} W_{10} & & & \\ & W_{20} & & \\ & & \ddots & \\ & & & W_{no} \end{bmatrix}$$

Lo que finalmente se tiene es un caso de mínimos cuadrados ponderados, por lo que puede obtenerse un nuevo estimador mediante la fórmula

$$\hat{\underset{\sim}{\beta}}_1 = (X'\ Wo\ X)^{-1}\ X'\ Wo\ \underset{\sim}{y}$$

Para la siguiente iteración se calculan los pesos pero con $\hat{\underset{\sim}{\beta}}_1$ en vez de $\hat{\underset{\sim}{\beta}}_o$ y así se continua hasta obtener cierto grado de convergencia.

Huber (1973) muestra que $\hat{\underset{\sim}{\beta}}$ (estimador M) tiene una distribución asintóticamente normal con matriz de covarianzas

$$\sigma^2 \frac{E\left[\psi^2 \quad (\varepsilon/\sigma)\right]}{\left[E\{\psi'(\varepsilon/\sigma)\}\right]^2} \qquad (X'X)^{-1}$$

A continuación se presenta una tabla con algunos estimadores-M, y sus funciones asociadas φ, ψ y w.

TABLA 1.3.1 (Montgomery 1982)

Estimador	$\varphi(Z)$	$\Psi(Z)$	$W(Z)$	Rango de Z
Mínimos cuadrados	$1/2\ Z^2$	Z	1	$/z/ < \infty$
Huber	$1/2\ Z^2$	Z	1	$/Z/ \leq t$
	$/Z/\ t - 1/2\ t^2$	t signo (Z)	$\frac{t}{/Z/}$	$/Z/ > t$
Hampel	$1/2\ Z^2$	Z	1	$/Z/ \leq a$
	$a/Z/ - (1/2)\ a^2$	a signo (Z)	$\frac{a}{/Z/}$	$a < /Z/ \leq b$
	$\frac{a(C/Z/-(1/2)\ Z^2}{C - b}$	$\frac{a \text{ signo } (Z)(C-/Z/)}{C - b}$	$\frac{a(C-/Z/)}{/Z/\ (C-b)}$	$b < /Z/ \leq c$
	$a(b+C-a)$	0	0	$/Z/ > C$
Andrews	$\frac{C'(1-\cos(Z/C'))}{2C'}$	$\frac{\text{Sen } (Z/C')}{0}$	$\frac{\text{Sen}(Z/C')}{Z/C'}$	$/Z/ \leq C'$ $/Z/ > C'$
Tukey	$\frac{Z^2}{2}\left[1- (\frac{Z}{K})^2+\frac{1}{3}(\frac{Z}{K})^4\right]$	$Z\ 1\ (Z/K)^2\ ^2$	$1-(Z/K)^2\ ^2$	$/Z/ \leq K$
	$\frac{K^2}{2}$	0	0	$/Z/ > K$

en donde $Z = \frac{yi - xi\hat{\beta}}{S}$

y las constantes t, a, b, c, C' y K para los distintos estimadores, dependerán del grado de - robustez que se quiera obtener.

La función Ψ controla, por así decirlo, el peso que se le da a cada residual y con frecuencia se le llama función de influencia.

La figura 1.3.1 (a), en el caso de mínimos cuadrados nos muestra que si los residuales crecen, la función Ψ también crece, y siendo ésta no acotada, el método tiende a ser no-robusto para datos provenientes de distribuciones con colas pesadas.

La función de Huber, ver fig. 1.3.1 (b), es igual a la de mínimos cuadrados en el intervalo (-t, t), pero fuera de él la función Ψ toma valores $\Psi(z) = t \text{ signo } (z)$ dando así un peso menor a los residuales conforme éstos crecen.

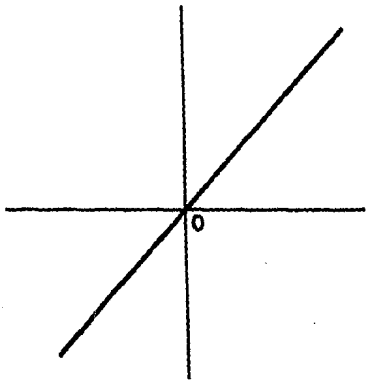
La función de Hampel fig. 1.3.1 (c) puede verse como una discretización de la función de Andrews ó la Tukey.

Fijando la atención en el intervalo (o, ∞), sucede que estas tres funciones crecen en un principio, pero luego decrecen y por esto se les conoce como funciones decrecientes.

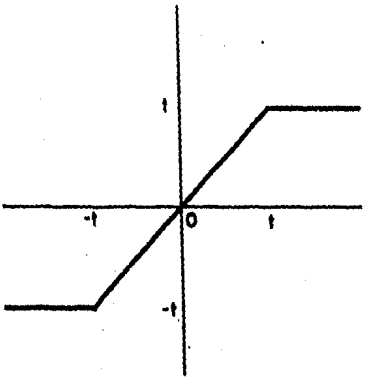
Si el valor absoluto de los residuales es muy pequeño, podría esperarse que las observaciones correspondientes a éstos fueran influyentes como puede verse con el punto 6 del ejemplo de la introducción, página 4 , a estas observaciones habría que darles entonces poco peso en el ajuste, esta situación se ve reflejada en el hecho de que

$\Psi(0) = 0$ y es una función continua e impar para las cinco funciones presentadas en la tabla 1.3.1

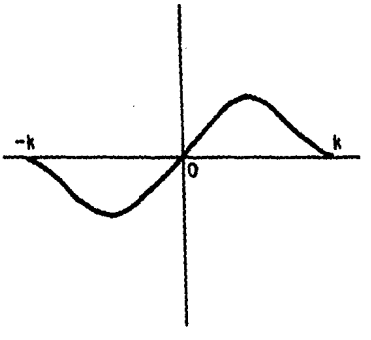
Ahora, si una observación es discrepante (outlier), el valor de su residual es grande. Si se utiliza como estimador el propuesto por Andrews, Tukey ó Hampel, entonces estas observaciones tendrán poco peso en el ajuste. Pero si se utiliza el estimador de Huber, el residual tendría que ser muy grande para que el peso $\frac{t}{z}$, fuera pequeño y esta obser-



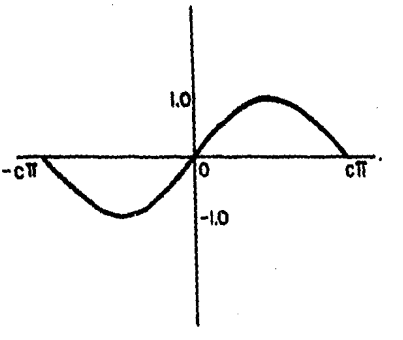
(a) MINIMOS CUADRADOS



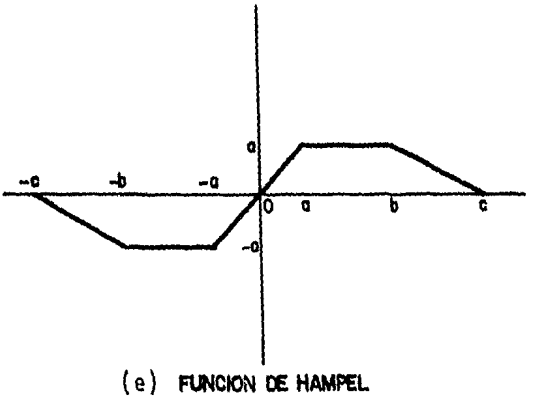
(b) FUNCION DE HUBER



(c) FUNCION DE TUKEY



(d) FUNCION DE ANDREWS



(e) FUNCION DE HAMPEL

FIGURA 1.3.1 FUNCIONES DE INFLUENCIA

vación no afectara al ajuste.

En el caso de que se utilice el método de mínimos cuadrados, no hay manera de protegerse de las observaciones discrepantes.

Entonces, para que el estimador $\tilde{\beta}$ del parámetro no sea influenciado por algunas observaciones lejanas al plano de regresión, la función $\psi(Z)$ debe ser acotada y tender a una constante, es decir para una φ suave

$$\lim_{|Z| \to \infty} \varphi'(Z) = 0$$

Como resultado de esta condición se pueden tener varios mínimos locales, por lo que la solución dada por el método iterativo de optimización dependerá del punto $\hat{\beta}_0$ que se use para iniciar el proceso.

CAPITULO 2

OTROS ESTIMADORES ROBUSTOS

2.1 EL METODO PROPUESTO POR ANDREWS

Considerando el modelo

$$yi = x_{i_1}\beta_1 + x_{i_2}\beta_2 + \ldots + x_{i\ K}\beta_K + \sigma\varepsilon i$$

$$yi = x_i'\underset{\sim}{\beta} + \sigma e_i$$

donde

β es el vector de parámetros desconocidos
x_i' es el vector renglón de variables explicativas,
σ es el parámetro desconocido de escala
y ε_i es el error para la observación.

Dado cualquier vector b se pueden formar los residuales

$$e_i(\underset{\sim}{b}) = yi - xi'\underset{\sim}{b}$$

y $\underset{\sim}{\beta}$ puede estimarse al localizar el mínimo de la función

$\sum_{i=1}^{n} \varphi(ei\ (b) / S\ (b)\ )$ como se discutió en la sección de estimadores M.

El estimador de escala propuesto por Andrews (1974) es

$$c\ s(b) = c\ \text{mediana}\ \{|e_i\ (b)|\} \quad \text{con} \quad c = 2.1$$

Debe recordarse que a partir de la condición

$$\lim_{|Z| \to \infty} \varphi'\ (Z) = 0$$

se pueden tener varios mínimos locales, surgiendo así el problema de que el método iterativo de optimización depende del punto inicial $\hat{\beta}_0$.

Un punto inicial podría ser

$$\hat{\beta}_0 = \hat{\beta}_{MC}$$

pero si los datos están lejos de ser normales podría encontrarse un mínimo local lejano al mínimo global.

Para el caso de localización, cuando se desea hallar la media de un grupo de datos $(x_1, x_2, \ldots, x_m)$, se utiliza la mediana como punto inicial. Andrews (1974) presenta el análogo para el caso de regresión.

Escribiendo el modelo en forma matricial

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \sigma\underset{\sim}{\varepsilon} = x_1 \beta_1 + x_2 \beta_2 + \ldots + x_K \beta_K + \sigma\varepsilon$$

donde xi es un vector columna de X, (i = 1,...,K),
se desea encontrar un estimador de $\underset{\sim}{\beta}$, $\hat{\underset{\sim}{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K)'$

Andrews lo define a través de un operador "R" de barrido (Sweep, ver Goodnight, 1979), que además suprime la dependencia de una variable con otra, el cual se describe a continuación:

$$\text{Sea } M = \left[X \mid Y \right]_{n \times (K+1)}$$

Rij ajusta la columna j-ésima Mj sumándole un múltiplo de la columna i-ésima Mi, es decir

$$Rij: Mj - b\ Mi,$$

donde b es una función de Mi y Mj y se define como sigue:

Por comodidad denotaremos a Mi como Z y a Mj como W.

Se forman dos grupos como sigue:

i) se ordenan las parejas (Zj, Wj) de acuerdo a Zj, j = 1,..., n.

ii) se eliminan dos conjuntos de $p_1 n$ puntos, correspondiendo a las Zj más grandes y más pequeñas.

iii) se eliminan dos conjuntos de $p_2 n$ puntos, cada uno con las Zj inmediatamente arriba y abajo de la mediana de $\{Zj\}$

Los puntos restantes forman dos grupos que se denotan como B y A, correspondiendo a los valores bajos y altos de Zj.

Ejemplo:

Si n = 20, $p_1 = .15$ y $p_2 = .1$, entonces

$$B = \{(Z_{(4)}, W_{(4)}), (Z_{(5)}, W_{(5)}) (Z_{(6)}, W_{(6)}), (Z_{(7)}, W_{(7)}), (Z_{(8)}, W_{(8)})\}$$

$$y\ A = \{(Z_{(13)}, W_{(13)}), (Z_{(14)}, W_{(14)}), (Z_{(15)}, W_{(15)}), (Z_{(16)}, W_{(16)}), (Z_{(17)}, W_{(17)})\}$$

El coeficiente b se define entonces como:

$$b = \frac{\underset{A}{\text{med}} \{W_k\} - \underset{B}{\text{med}} \{W_k\}}{\underset{A}{\text{med}} \{Z_k\} - \underset{B}{\text{med}} \{Z_k\}}$$

R es el resultado de aplicar sucesivamente operadores Rij.

Este operador es no-lineal y no-indempotente.

La primera variable se usa para modificar las k restantes al aplicar

$$R_{1,K+1} \quad (\ldots(R_{1,3} \quad (R_{1,2} \quad (M))) \ldots) = M^*$$

La segunda variable se usa para modificar las siguientes K-1 variables, aplicando R a M*:

$$R_{2,\ K+1} \quad (\ldots(R_{2,4} \ (R_{2,3} \ (M^*)))\ldots)$$

Este proceso se sigue para todas las variables explicativas.

Generalmente se necesita más de una iteración, el número de iteraciones dependerá del número de variables explicativas que se tengan.

Si el número de iteraciones necesarias son m, las operaciones a seguir están representadas por el algoritmo.

```
hacer para ℓ = 1   hasta m
hacer para i = 1   hasta K
hacer para j = i+1 hasta K+1
     aplicar Rij
```

Se usará m = (K/2) + 2

Una vez aplicado R a M, las columnas de la matriz resultante se usan como variables explicativas en la regresión de y, los coeficientes de MC encontrados forman el vector inicial bo.

Si se considera la matriz aumentada

$$M^+ = \left[\frac{M}{I}\right] \quad (n+K+1) \times (K+1)$$

el cálculo de bo se facilita si se aplica R no a M, sino a M^+.

$$bo = - (M^+_{n+1,\ K+1}, \ \ldots, \ M^+_{n+K,\ K+1})$$

y el vector de residuales

$$e(bo) = Y - X \ bo \ es$$

$$e(bo) = (M_{1,k+1} , \ldots, M_{n,k+1})$$

Ver Goodnight (1979).

Este procedimiento es usado únicamente para tener un punto inicial para una optimización posterior, Andrews (1974) señala que el procedimiento tiene al menos un punto fijo.

Teniéndose ya el punto inicial bo, se procederá a mejorarlo minimizando la función

$$\sum_{j=1}^{n} \Psi (ej(bi) / S (bi-1)),$$

donde $S(bi)$ = mediana $\{ | ej (bi)| \}$

usando como valor inicial bo.

Esto puede hacerse mediante el uso de un programa de mínimos cuadrados ponderados, como se mencionó anteriormente.

A continuación se presenta un ejemplo de 21 observaciones y 3 variables explicativas. Los datos aparecen en la tabla 2.1.1 . Estos datos fueron estudiados por Daniel y Wood (1971).

La gráfica en papel normal de los residuales del ajuste por medio de mínimos cuadrados, muestra que la observación 21 tiene un residual anormalmente grande.

El ajuste por mínimos cuadrados a estos datos fué:

$$\hat{Y} = -36.9 + 0.72\, x_1 + 1.30\, x_2 - 0.15\, x_3 \quad (2.1).$$

Los residuales de este modelo aparecen en la tabla 2.1.2 y la gráfica en papel normal en la figura 2.1.1 .

Después de un estudio más profundo, Daniel y Wood separaron las observaciones 1, 3, 4 y 21. Ver la tabla 2.1.3 .

El ajuste de MC para los datos restantes fue

$$\hat{Y} = -37.6 + 0.80\, x_1 + 0.58\, x_2 - 0.07\, x_3 \quad (2.2)$$

Tabla 2.1.1

Datos de operación de una planta de oxidación de Amonia a Acido nítrico.

Número de Observación	Pérdida del Material	Flujo de Aire	Temperatura del agua de enfriamiento al entrar	Concentración del ácido
1	42	80	27	89
2	37	80	27	88
3	37	75	25	90
4	28	62	24	87
5	18	62	22	87
6	18	62	23	87
7	19	62	24	93
8	20	62	24	93
9	15	58	23	87
10	14	58	18	80
11	14	58	18	89
12	13	58	17	88
13	11	58	18	82
14	12	58	19	93
15	8	50	18	89
16	7	50	18	86
17	8	50	19	72
18	8	50	19	79
19	9	50	20	80
20	15	56	20	82
21	15	70	20	91

TABLA 2.1.2 RESIDUALES PARA VARIOS AJUSTES PARA LOS DATOS DE PERDIDA DE MATERIAL

	RESIDUALES			
	MINIMOS CUADRADOS		AJUSTE ROBUSTO	
OBSERVACION	LOS 21 PUNTOS	SIN LAS OBSERVACIONES 1,3,4,21.	LOS 21 PUNTOS	SIN LAS OBSERVACIONES 1,3,4,21
1	3.24	6.08	6.11	6.11
2	-1.92	1.15	1.04	1.04
3	4.56	6.44	6.31	6.31
4	5.70	8.18	8.24	8.24
5	-1.71	-0.67	-1.24	-1.24
6	-3.01	-1.25	-0.71	-0.71
7	-2.39	-0.42	-0.33	-0.33
8	-1.39	0.58	0.67	0.67
9	-3.14	-1.06	-0.97	-0.97
10	1.27	0.35	0.14	0.14
11	2.64	0.96	0.79	0.79
12	2.78	0.47	0.24	0.24
13	-1.43	-2.51	-2.71	-2.71
14	-0.05	-1.34	-1.44	-1.44
15	2.36	1.34	1.33	1.33
16	0.91	0.14	0.11	0.11
17	-1.52	-0.37	-0.42	-0.42
18	-0.46	0.10	0.08	0.08
19	-0.60	0.59	0.63	0.63
20	1.41	1.93	1.87	1.87
21	-7.24	-8.63	-8.91	-8.91

TABLA 2.1.3

OBSERVACION	ESTADISTICAS	ESTADISTICA Ti
i	Di de Cook*	Para observaciones discordantes
1	0.153710	1.209475
2	0.059683	-0.705139
3	0.126414	1.617904
4	0.130542	2.051794
5	0.004048	-0.530504
6	0.019565	-0.963204
7	0.048802	-0.825947
8	0.016502	-0.473652
9	0.044556	-1.048586
10	0.011930	0.426188
11	0.035866	0.878292
12	0.065066	0.966707
13	0.010765	-0.408731
14	0.000020	-0.016950
15	0.038516	0.800616
16	0.003379	0.291185
17	0.065473	-0.599586
18	0.001122	-0.148680
19	0.002179	-0.197199
20	0.004492	0.443117
21	0.692000	-3.330493

Los valores D_1, D_3, D_4 y D_{21} son los más grandes, lo que significa que las observaciones asociadas son influyentes en el ajuste Análogamente los valores T_1, T_3, T_4 y T_{21} son los valores más grades, por lo que para un cierto nivel de significancia estas observaciones podrían ser declaradas como observaciones discrepantes.

* (ver Weisberg, 1980)

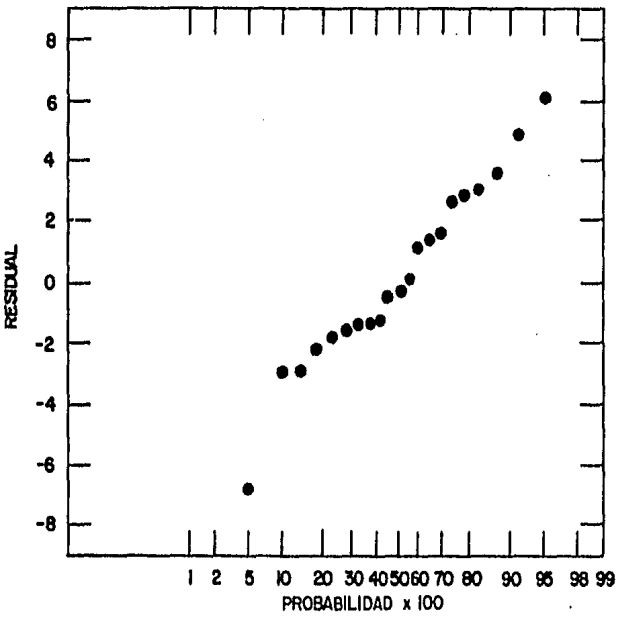


FIGURA 2.1.1 GRAFICACION EN PAPEL NORMAL DE LOS RESIDUALES DEL AJUSTE DE MINIMOS CUADRADOS INCLUYENDO LAS 21 OBSERVACIONES

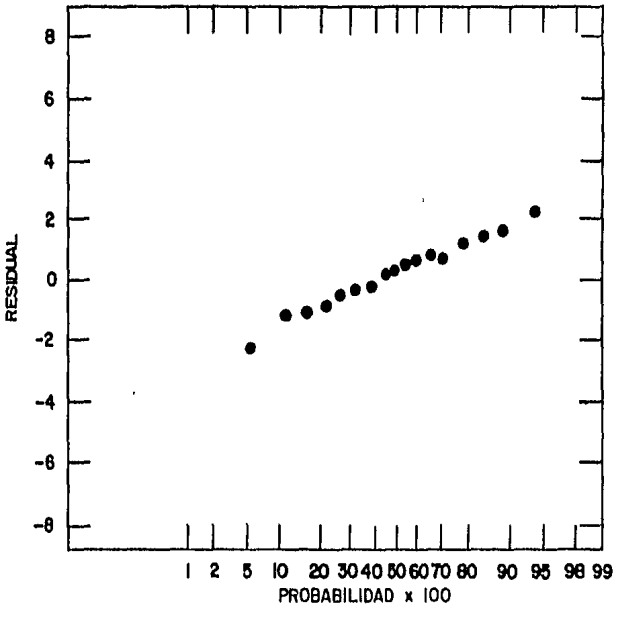


FIGURA 2.1.2 GRAFICACION EN PAPEL NORMAL DE LOS RESIDUALES DEL AJUSTE DE MINIMOS CUADRADOS ELIMINANDO LAS OBSERVACIONES 1, 3, 4 y

Los residuales aparecen en la tabla 2.1.2 y la gráfica correspondiente en papel normal en la figura 2.1.2.

Andrews (1974) hace notar que la mayoría de los investigadores que utilizan modelos de regresión lineal no son tan meticulosos como Daniel y Wood y podrían emplear con precaución métodos robustos produciendo resultados equivalentes.

El ajuste robusto usando C=1.5 ver tabla 1.3.1 produce el modelo

$$\hat{y} = -37.2 + 0.82\, x_1 + 0.52\, x_2 - 0.07\, x_3, \quad (2.1.3)$$

que es casi igual a la ecuación hallada por Daniel y Wood después de un análisis profundo. Los residuales aparecen en la tabla 2.1.2 y su gráfica en papel normal en la figura 2.1.3. Los puntos 1, 3, 4 y 21 son claramente identificados en esta gráfica.

El ajuste robusto quitando a los puntos 1,3, 4 y 21 es exactamente igual al hallado usando todos los puntos, por lo que el ajuste resulta independiente de estos puntos. Los residuales aparecen en la tabla 2.1.2 y la gráfica en papel normal aparece en la figura 2.1.4. Estas gráficas y la del ajuste MC sin las observaciones 1, 3, 4 y 21 son casi las mismas.

Cabe recalcar que este método requiere de un ajuste inicial seguro (bueno), que será refinado para obtener un procedimiento relativamente eficiente para datos "cercanos" a la distribución normal.

Este procedimiento es iterativo y comparado con MC es más caro en cuestión de cómputo, pero es insensible a un número moderado de observaciones extremas, que además pueden ser identificadas fácilmente al examinar los residuales y un nuevo ajuste ya no será necesario.

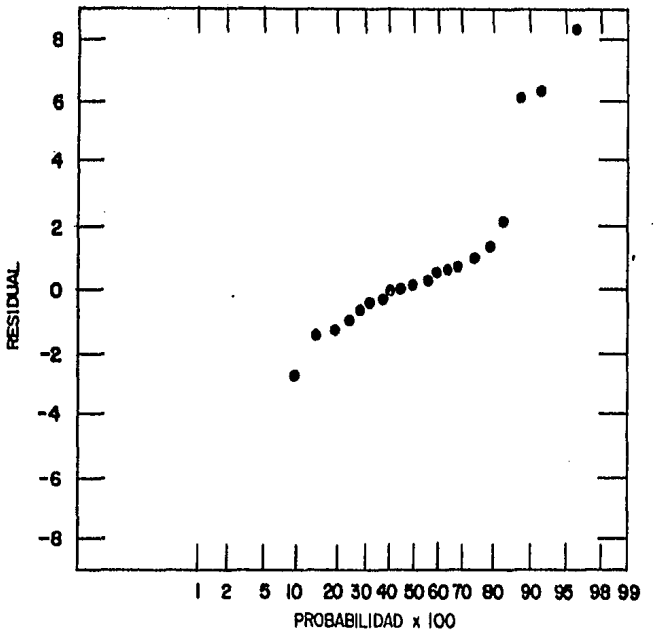


FIGURA 2.1.3 GRAFICACION EN PAPEL NORMAL DE LOS RESIDUALES DEL AJUSTE ROBUSTO INCLUYENDO LAS 21 OBSERVACIONES

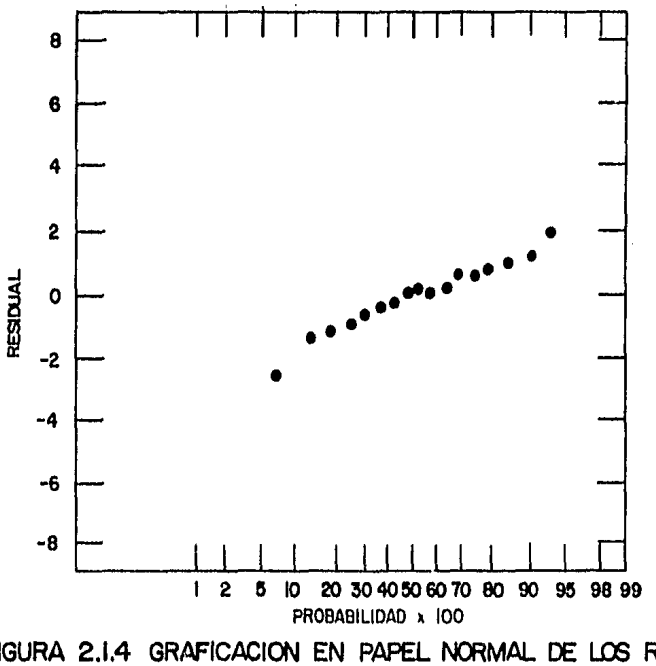


FIGURA 2.1.4 GRAFICACION EN PAPEL NORMAL DE LOS RESIDUALES DEL AJUSTE ROBUSTO ELIMINANDO LAS OBSERVACIONES 1, 3, 4 y 21

2.2 EL METODO PROPUESTO POR HILL Y HOLLAND

Hill y Holland (1979) consideran el modelo lineal

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

donde $\underset{\sim}{Y}$, X y $\underset{\sim}{\beta}$ son como se han considerado antes y $\underset{\sim}{\varepsilon}$ es un vector aleatorio, con coordenadas independientes e idénticamente distribuídas con función de densidad, donde f es normal o una normal contaminada.

Se denotará al estimador de MC de $\underset{\sim}{\beta}$ como $\hat{\underset{\sim}{\beta}}_{MC}$. Las dos alternativas que los autores proponen son:

(1) El estimador de desviaciones mínimas absolutas o estimador de norma L_1, (detalles de este estimador se ven más adelante) $\hat{\beta}_{L_1}$ que minimiza:

$$\sum_{i=1}^{n} \left| Yi - \sum_{j=1}^{n} xij\beta_j \right|$$

y (2) el estimador SENO 1, basado en el estimador propuesto por Andrews visto en la sección anterior, que consiste en utilizar una sola iteración empezando con $\hat{\underset{\sim}{\beta}}_0 = \hat{\underset{\sim}{\beta}}_{L_1}$.

Este estimador SENO 1 ó $\hat{\underset{\sim}{\beta}}_s$ puede verse como un mejoramiento de $\hat{\underset{\sim}{\beta}}_{L_1}$.

Para comparar $\hat{\underset{\sim}{\beta}}_{L_1}$ y $\hat{\underset{\sim}{\beta}}_s$ entre sí y con $\hat{\underset{\sim}{\beta}}_{MC}$ a través de varias X y f, se tomó como medida de ineficiencia relativa de $\hat{\underset{\sim}{\beta}}$ con respecto a $\hat{\underset{\sim}{\beta}}_{MC}$.

$$\text{inef}\left(\hat{\underset{\sim}{\beta}}\right) = \frac{E\left\|\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}\right\|^2}{E\left\|\hat{\underset{\sim}{\beta}}_{MC} - \underset{\sim}{\beta}\right\|^2}$$

donde $\|\cdot\|$ es la norma euclideana. (se comparan errores cuadráticos medios esperados)

para $\hat{\underset{\sim}{\beta}}$ insesgados

$$\text{inef}\left(\hat{\underset{\sim}{\beta}}\right) = \frac{\sum_{j=1}^{k} \text{var}\left(\hat{\beta}_j\right)}{\sum_{j=1}^{k} \text{var}\left(\hat{\beta}_{j\ MC}\right)}$$

Así entre más pequeño sea el valor de inef $\left(\hat{\underset{\sim}{\beta}}\right)$ mejor será $\hat{\underset{\sim}{\beta}}$ respecto a $\hat{\underset{\sim}{\beta}}_{MC}$.

El estudio de estos estimadores se llevó a cabo con datos simulados mediante el método Montecarlo (Ver método de Montecarlo apéndice 3).

Todas las matrices X usadas se derivaron de una matriz básica de 20 x 6 tabla 2.2.1 constituída por tres grupos de dos columnas. Las columnas 1 y 2 fueron elegidas de tal manera que las 20 parejas $(x_1 i, x_2 i)$ $i=1,\ldots,$ 20 caen en un cuadrado centrado en el origen con lados paralelos a los ejes y corresponderían a las variables de un experimento diseñado.

Las columnas 3 y 4 son una muestra tamaño 20 de una distribución bivariada normal que representan una situación bien portada pero no diseñada como en las columnas 1 y 2. Las columnas 5 y 6 fueron elegidas de una muestra tamaño 20 de una distribución con observaciones discrepantes, representando una situación con puntos muy influyentes en el ajuste.

Cada columna fue estandarizada de manera que tiene media cero y la suma de cuadrados es uno.

En la tabla 2.2.2 se ve que las columnas no están altamente correlacionadas.

Para los errores con distribución no normal se utilizó la distribución normal contaminada dada por

$$g(u) = (2\pi)^{-1/2}\left\{(1-\alpha)\, e^{-u^2/2} + (\alpha/q)\, e^{-(u/q)^2/2}\right\}$$

$$\text{con} \qquad 0<\alpha<1 \qquad y \qquad 1<q<\infty$$

La interpretación de esta distribución es que con una probabilidad $1-\alpha$, la observación proviene de una población con distribución normal con varianza 1 y con probabilidad α proviene de una población cuya distribución es q veces una normal con varianza 1.

TABLA 2.2.1 COLUMNAS ESTANDARIZADAS DE DATOS

	COLUMNA					
RENGLON	1	2	3	4	5	6
1	0.2712	0.2712	-0.0453	0.0257	-0.0880	0.0288
2	0.2712	0.1627	0.1092	-0.1268	-0.0809	0.0470
3	0.2712	0.0542	0.4513	0.0963	0.0140	0.0682
4	0.2712	-0.0542	-0.1605	0.2977	-0.1065	0.0225
5	0.2712	-0.1627	0.2242	-0.3618	0.2463	0.3193
6	0.2712	-0.2712	0.0107	0.1246	-0.0814	0.0461
7	0.1627	-0.2712	0.1937	0.1006	-0.0373	0.0583
8	0.0542	-0.2712	-0.2435	0.3205	-0.1373	0.0404
9	-0.0542	-0.2712	-0.0094	-0.4123	-0.0852	0.0228
10	-0.1627	-0.2712	0.1382	0.4631	-0.0630	-0.0112
11	-0.2712	-0.2712	0.0956	0.0984	-0.0489	0.0388
12	-0.2712	-0.1627	0.0597	-0.1136	-0.0732	0.0327
13	-0.2712	-0.0542	-0.0613	-0.1263	-0.0944	0.0303
14	-0.2712	0.0542	0.1282	0.0598	-0.0680	0.0691
15	-0.2712	0.1627	-0.0966	-0.0085	0.1387	-0.0672
16	-0.2712	0.2712	-0.1060	-0.3819	-0.1340	0.0559
17	-0.1627	0.2712	0.2013	0.0145	-0.0290	0.0966
18	-0.0542	0.2712	-0.4324	-0.2083	-0.1520	-0.9198
19	0.0542	0.2712	0.0914	0.0840	-0.0417	-0.0620
20	0.1627	0.2712	-0.5486	0.0544	0.8917	0.0833

TABLA 2.2.2

MATRIZ DE CORRELACION DE LAS COLUMNAS DE DATOS

	COLUMNA					
RENGLON	1	2	3	4	5	6
1	1.	0.	0.0573	0.1462	0.2150	0.1576
2	0.	1.	-0.2786	-0.2456	0.2382	-0.3034
3	0.0573	-0.2786	1.	0.0404	-0.3499	0.4818
4	0.1462	-0.2456	0.0404	1.	-0.0297	0.0744
5	0.2150	0.2382	-0.3499	-0.0297	1.	0.2448
6	0.1576	-0.3034	0.4818	0.0744	0.2448	1

Las distintas distribuciones contaminadas normales se denotarán como $CNq\alpha$.

Para que todos los errores tuviesen varianza 1, se tomó la distribución

$$f(u) = (1/t)\, g(u/t)$$

con g como se definió antes y $t^2 = 1 - \alpha + q^2$

Ambos estimadores $\hat{\beta}_{L_1}$ y $\hat{\beta}_S$ son invariantes, en el sentido de que, si el vector "Y" de valores observados es transformado a $Y + X\beta(o)$, para algún $\beta(o)$, entonces $\hat{\beta}$ se transforma en $\hat{\beta} + \beta(o)$, esta invarianza implica que

$$E_{\beta} \| \hat{\beta} - \beta \|^2 = E_o \| \hat{\beta} \|^2$$

para cualquier distribución de error.

Todos los resultados del método de Montecarlo fueron calculados con el verdadero valor de β igual a 0.

Una gran diferencia entre $\hat{\beta}_{L_1}$ y $\hat{\beta}_S$ es que el segundo depende en la elección del estimador de escala CS.

Es claro que si S es muy grande, los pesos Wi estarán más cerca de 1 y $\hat{\beta}_S$ se parecerá a $\hat{\beta}_{MC}$.

Si S es muy pequeña, los pesos serán todos cero y $\hat{\beta}_S$ no podrá definirse.

$$Wi = \begin{cases} 1 & \text{Si } Zi = 0 \\ [\text{Sen } (zi/c)] / (Zi/C) & \text{Si } |Zi| < \pi C \\ 0 & \text{Si } |Zi| > \pi C \end{cases}$$

donde $Zi = (Yi - \sum xij\, \hat{\beta}j_{L1}) / S$, $\quad C = 2.1$

El estimador robusto propuesto por Andrews como se dijo antes es

$$CS = C \text{ mediana } \{|ei|\} \quad , \quad C = 2.1 \quad (2.2.1)$$

La elección del factor C depende la función de peso y del nivel deseado de ineficiencia para el caso de la normal. (con C=2.1 se alcanza una ineficiencia de 1.07 para el caso normal).

Para obtener los valores de la ineficiencia se necesitan los valores de

$$\sum_j \text{var}\,(\hat{\beta}j_{L_1}) \;,\quad \sum_j \text{var}\,(\hat{\beta}j_{s}) \;,\quad \sum_j \text{var}\,(\hat{\beta}j_{MC})$$

donde

$\sum_j \text{var}\,(\hat{\beta}j_{MC})$ se obtiene fácilmente de $(X'X)^{-1}$

$$\sum_j \text{var}\,(\hat{\beta}j_{L_1}) = \text{traza}\,(\Sigma_{L_1})$$

y

$$\sum_j \text{var}\,(\hat{\beta}j_{s}) = \text{traza}\,(\Sigma_{s})$$

donde

Σ_{L_1} y Σ_s son las matrices de covarianza de $\hat{\beta}_{L_1}$ y $\hat{\beta}_s$. Estas covarianzas fueron estimadas por un método del tipo Montecarlo descrito por Holland (1973) con 500 replicaciones. Para todos los casos, N = 20 y se varió la dimensión de la matriz X, con 1, 2, 3, 4, 5 y 6 columnas.

En la tabla 2.2.3 se ve que conforme aumenta el número de parámetros a estimar $\hat{\beta}_s$ se hace menos eficiente y de seguir este comportamiento $\hat{\beta}_s$ no sería un mejoramiento de $\hat{\beta}_{L_1}$, para casos con más de 6 columnas. De manera contraria $\hat{\beta}_{L_1}$ se vuelve más eficiente conforme aumentan las columnas en X.

La ineficiencia creciente de $\hat{\beta}_s$ puede suprimirse ajustando el estimador de escala, de manera que ésta dependa de k.

La mayoría de los algoritmos para hallar $\hat{\beta}_{L_1}$ hacen que k de los residuales sean idénticamente 0. Esto sugiere eliminar k-1 de estos residuales cero y calcular la escala como sigue

2.1 $S = 2.1$ mediana $\{|r_i|$ tal que r_i pertenece a los

TABLA 2.2.3

INEFICIENCIA DE $\hat{\beta}_{L_1}$ Y $\hat{\beta}_S$ BAJO ERRORES NORMALES USANDO (2.2.1) COMO DEFINICION DE σ , CON n=20

INEFICIENCIA	K				
	1	2	3	4	6
$\hat{\beta}_{L_1}$	1.63	1.63	1.57	1.54	1.47
$\hat{\beta}_S$	1.10	1.12	1.17	1.23	1.32

n-k+1 residuales restantes} (2.2.2)

Con este nuevo estimador de escala se obtuvieron:

$k = 2$ inef $(\hat{\beta}_s) = 1.08$

$k = 4$ inef $(\hat{\beta}_s) = 1.08$

$k = 6$ inef $(\hat{\beta}_s) = 1.09$

Con esta modificación, la ineficiencia de $\hat{\beta}_s$ es independiente de la dimensión de la matriz X. Obviamente esto es conveniente porque ya no depende de la cantidad de variables que se tengan como explicativas.

De aquí en adelante los resultados dados para $\hat{\beta}_s$ utilizan (2.2.2)

Los resultados de Montecarlo, para el caso de la normal contaminada se hallaron para $\alpha = 0.1$, 0.25 y 0.5, $q = 3,5$ y 10 y para $k = 2$, 4 y 6. Cubriendo un amplio rango de las distribuciones contaminadas.

En la tabla 2.2.4 se puede ver que tanto $\hat{\beta}_{L_1}$ como $\hat{\beta}_s$ son mejores que $\hat{\beta}_{MC}$, exceptuando 5 casos.

Y exceptuando otros 4 casos, $\hat{\beta}_s$ es mejor que $\hat{\beta}_{L_1}$.

Se nota también que $\hat{\beta}_s$ y $\hat{\beta}_{L_1}$ son más eficientes según aumenta el grado de contaminación.

Para ambos $\hat{\beta}_s$ y $\hat{\beta}_{L_1}$ la ineficiencia crece según aumenta el número de columnas excepto para $\alpha = .10$ y K=3, pero más bien puede sospecharse que esta tendencia se debe a la naturaleza de las columnas que a la dimensión de X.

También se calcularon las ineficiencias para una muestra mayor, n = 40. Ver tabla 2.2.5 .

Los valores son más pequeños que los de la tabla 2.2.4 indicando esto que la robustez es más fácil de alcanzar con muestras de mayor tamaño.

TABLA 2.2.4

INEFICIENCIA DE $\hat{\beta}_{L_1}$ Y $\hat{\beta}_S$ RESPECTO A $\hat{\beta}_{MC}$

α	K	q = 3		q = 5		q = 10	
		$\hat{\beta}_{L1}$	$\hat{\beta}_S$	$\hat{\beta}_{L1}$	$\hat{\beta}_S$	$\hat{\beta}_{L1}$	$\hat{\beta}_S$
.10	2	1.09	.78	.60	.44	.19	.13
	4	1.03	.80	.59	.47	.20	.15
	6	1.03	.87	.72	.60	.38	.33
.25	2	.87	.71	.42	.39	.16	.15
	4	.91	.78	.49	.47	.26	.25
	6	.96	.83	.69	.63	.45	.43
.50	2	.91	.81	.58	.63	.35	.46
	4	1.00	.87	.74	.75	.55	.62
	6	1.08	.94	.92	.87	.80	.79

TABLA 2.2.5

INEFICIENCIAS DE $\hat{\beta}_{L_1}$ Y $\hat{\beta}_S$ PARA n=40

K	CN3.5		CN10.1	
	$\hat{\beta}_{L_1}$	$\hat{\beta}_S$	$\hat{\beta}_{L_1}$	$\hat{\beta}_S$
2	.79	.72	.18	.12
4	.82	.75	.18	.12
6	1.00	.86	.25	.17

Matriz de Covarianza para estimadores robustos de regresión.

Como se dijo en sección 1.3, $\hat{\beta}$ tiene una distribución asintóticamente normal con matriz de covarianza de la forma

$$h^2 (X' X)^{-1} \qquad (2.2.3)$$

donde h^2 depende potencialmente de:

(1) El estimador (para los estimadores obtenidos por mínimos cuadrados ponderados incluirá la manera en que se dan los pesos).

(2) La distribución del error.

(3) El método de escalar los residuales.

(4) El nivel deseado de ineficiencia frente a una distribución estándar.

(5) La matriz X.

Si (2.2.3) se da

y como $E \left\| \hat{\beta}_{MC} - \beta \right\|^2 = var(f)\ tr\ (X'X)^{-1} = tr\ (X'X)^{-1}$

(pues var(f) = 1)

entonces

$$inef\ (\hat{\beta}) = h^2 \qquad para\ \hat{\beta} = \hat{\beta}_S\ ,\hat{\beta}_{L_1}$$

Los autores consideraron que h^2 dependería únicamente del tamaño de n y k es decir de (5).

Para el caso normal se hizo un intento para suprimir el efecto de X en h^2, pero la tabla 2.2.4 muestra que no fue posible, y es evidente que la dependencia de X no es una función de n y k exclusivamente.

Si (2.2.3) se da, se tiene que la ineficiencia de cada coeficiente es constante, ya que

$$inef\ (\hat{\beta}_j) = \frac{E[\hat{\beta}_j^2}{E[\hat{\beta}_{j_{MC}}^2]} = \frac{h^2 (X'X)^{-1}_{jj}}{(X'X)_{jj}} = h^2$$

Pero al calcular estas ineficiencias para CN3.1 y CN10.1 se obtuvieron los resultados de la tabla 2.2.6 que muestran que para CN10.1 $\hat{\beta}_5$ y $\hat{\beta}_6$ tienen ineficiencias bastante más grandes que los otros coeficientes. Convendría entonces introducir una medida de la kurtosis de las columnas de X para obtener un buen estimador de la matriz de covarianza de $\hat{\beta}$.

TABLA 2.2.6

INEFICIENCIAS DE LOS ESTIMADORES DE LOS COEFICIENTES PARA

n = 20

$\frac{E[\beta_j^2]}{E[\hat{\beta}_j\ MC^2]}$	CN3.1		CN10.1	
	β_{L_1}	β_S	β_{L_1}	β_S
j = 1	1.00	.75	.20	.15
2	1.12	.84	.23	.20
3	1.04	.85	.25	.22
4	1.09	.83	.26	.21
5	1.11	.78	.58	.49
6	1.12	.78	.71	.65

2.3 EL METODO PROPUESTO POR HINICH Y TALWAR

Hinich y Talwar (1975) consideran el modelo

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

donde Y_{nx1}, X_{nxk}, $\underset{\sim}{\beta}_{kx1}$ son como se han venido definiendo y $\underset{\sim}{\varepsilon}_{nx1}$ es un vector de errores independientes e idénticamente distribuidos con función característica.

$$\phi_{\varepsilon}(t) = \exp\left\{- |\sigma t|^{\alpha}\right\}, \quad 1 \leq \alpha \leq 2 \qquad \sigma > 0$$

(si ε se distribuye como normal $N(0,\sigma)$ la función característica es

$$\phi_{\varepsilon}(t) = \exp\left\{- (\sigma t)^2 / 2\right\}$$

El primer paso es dividir la muestra en $m=n/K$ submuestras ajenas, (por conveniencia se supone $n=mk$). Las submuestras se formaron por agrupamiento consecutivo, esto significa que la submuestra ℓ-ésima está formada por los elementos $(\ell-1)\,k+1, \ldots, (\ell-1)\,k+k$.

El estimador preliminar para el coeficiente β_j $(j=1\ldots,K)$ es:

$$\hat{\beta}_j = \text{mediana}\,(\hat{\beta}_{MCj}{}^{1}, \ldots, \hat{\beta}_{MCj}{}^{n}) \qquad m=n/K$$

donde $\hat{\beta}_{MCj}{}^{\ell}$ es el estimador del coeficiente β_j de mínimos cuadrados de la submuestra ℓ-ésima dada por

$$Y_{(\ell)} = (Y_{(\ell-1)K+1}, \ldots, Y_{\ell K})' = X_{(\ell)}\underset{\sim}{\beta} \qquad \ell = 1, 2, \ldots, m$$

$X_{(\ell)}$ es de $K \times K$ cuyo ij-ésimo elemento es

$x_{(\ell-1)K+i,j}$

y $\hat{\beta}(\ell) = X_{(\ell)}^{-1} Y_{(\ell)} = (\hat{\beta}_{1\ell}, \ldots, \hat{\beta}_{K\ell})'$

Suponiendo la no dependencia lineal para que $X_{(\ell)}$ sea no singular.

Al escoger los tamaños de submuestras iguales a K se reduce la varianza de la mediana a expensas de la precisión de los estimadores en cada submuestra. Dado que los estimadores de MC son no robustos cuando los errores tienen una distribución con colas muy pesadas es razonable tratar de maximizar el número de submuestras dada una n.

Dado el estimador preliminar $\hat{\beta}$ se ordenan los residuales

$$ei = yi - \sum_{j=1}^{k} xij\ \hat{\beta}_j$$

Aquí los autores toman como estimador robusto de la desviación estándar σ el sugerido por Fama y Roll (1971).

$$S = \frac{1}{1.654}\ (e_{.72} - e_{.28}) \qquad (2.3.1.)$$

donde $e_{.72}$ y $e_{.28}$ son las estadísticas de orden utilizados para estimar los cuantiles .28 y .72 de la distribución de los errores ε. Este estimador tiene un sesgo asintótico de menos del 0.4 porciento y se distribuye asintóticamente como normal con desviación estándar.

$$\sigma(s) = (0.3/f_{\varepsilon}\ (\alpha,\ .72))\ n^{-1/2}$$

donde $f_{\varepsilon}\ (\alpha, .72)$ es la altura de la densidad de ε en el cuantil .72 (o de su cuantil simétrico .28).

Dado el estimador de σ, la muestra es reducida al quitar las observaciones (yi, xi1..., xik) para las cuales su correspondiente residual sea mayor que 4S. El estimador final de β es el estimador de MC calculado con las observaciones restantes.

Los autores estudiaron el mejor tamaño de las submuestras para el caso K = 1, es decir cuando

$$yi = \beta xi + \varepsilon i, \qquad i = 1 \ldots n$$

Agrupando las n observaciones yi en $m = \frac{n}{R}$ vectores $y(\ell)$ con $y(\ell) = X(\ell)\beta + \varepsilon(\ell)$.

donde

$$Y_{(\ell)} = \begin{bmatrix} Y_{(\ell-1)\ R+1} \\ \\ Y_{(\ell-1)\ R+R} \end{bmatrix}, \quad X_{(\ell)} = \begin{bmatrix} x_{(\ell-1)\ R+1} \\ \\ x_{(\ell-1)\ R+R} \end{bmatrix}, \quad \varepsilon_{(\ell)} = \begin{bmatrix} \varepsilon_{(\ell-1)\ R+1} \\ \\ \varepsilon_{(\ell-1)\ R+R} \end{bmatrix}$$

$$\ell = 1, 2, \ldots, m$$

suponiendo que R divide a n.

Sean

$$Z_\ell = (X'_{(\ell)} X_{(\ell)})^{-1} X'_{(\ell)} Y_{(\ell)} = \beta + (X'_{(\ell)} X_{(\ell)})^{-1} \varepsilon_{(\ell)} \qquad (2.3.1)$$

$$\ell = 1, 2, \ldots, m$$

donde $Z_1, Z_2, \ldots,$ Zm son independientes.

Considérense dos casos:

Caso a: Suponiendo que $x_1, x_2 \ldots, x_n$ son conocidas y fijas, entonces Z tiene función característica

$$\exp\left\{ i\beta t - \left| C_\ell t \right|^\alpha \right\}$$

Hinich (1975), con

$$C_\ell = \sigma \left(X'_{(\ell)} X_{(\ell)} \right)^{-1/\alpha} \left(\left| X_{(\ell-1)\ R+1} \right|^\alpha + \ldots + \left| X_{(\ell-1)\ R+R} \right|^\alpha \right)^{1/\alpha} \qquad (2.3.2)$$

Claramente de (2.3.1) se ve que Z_ℓ tiene la misma localización β, pero diferente escala $C_{\ell,\ell} = 1, 2, \ldots m$.

Caso b: Suponiendo que $x_1, x_2, \ldots, x_n$ son variables aleatorias independientes e idénticamente distribuídas (v.a.i.i.) con varianza finita e independientes de ε_1. Entonces $Z_1, Z_2 \ldots Zm$ son v.a.i.i.d. con localización β

Se estudiará el estimador Mm=mediana $(Z_1, \ldots, Z_m)$.

En cualquier caso el modelo de regresión es reducido a un submodelo de localización. Y es por esto que los autores estudian el estimador

$$Mm = \text{mediana}\ (Z_1, Z_2 \ldots, Z_m),$$

y también medias recortadas de $Z_1, Z_2, \ldots, Zm$, que se exponen más adelante.

Los autores (ver Hinich y Talwar, 1973) argumentan que cuando α tiene valores cercanos a uno, es decir cuando los errores se distribuyen como Cauchy el mejor tamaño para las submuestras es R = 1, en el sentido que minimiza la varianza asintótica de Mm cuando $m \to \infty$

Mientras que para $\alpha = 2$, cuando los errores se distribuyen como normal, R = n, es el mejor.

Los autores muestran en la figura 2.3.1 que R = 1 es la mejor para α entre 1 y 1.5, entonces para errores con distribuciones con colas más pesadas que la normal conviene perder un poco de la precisión del estimador en cada grupo para ganar precisión en la mediana, debido a que hay más grupos (de menor tamaño).

Y si la α crece, esto es si las colas van siendo menos pesadas, el tamaño R de las submuestras también crece.

Para probar los resultados teóricos, los autores consideraron el modelo

$$yi = xi\beta + \varepsilon i \qquad i = 1,\ldots,n. \quad (k = 1).$$

se generaron siete grupos de números pseudoaleatorios correspondientes a

$$\alpha = 1.0,\ 1.1,\ 1.3,\ 1.5,\ 1.7,\ 1.9 \text{ y } 2.0$$

y éstos dieron origen a los valores de los εi. También se generaron 99 números pseudoaleatorios provenientes de una distribución normal, como observaciones xi.

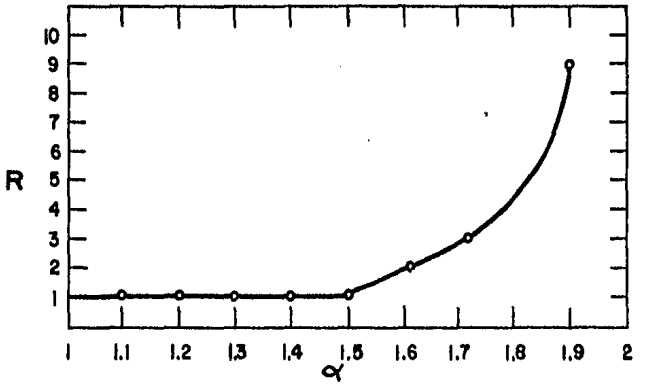


FIGURA 2.3.1 GRAFICA DEL MEJOR TAMANO DE GRUPO R COMO FUNCION DE α

Las muestras con n elementos se dividieron en m subgrupos los tamaños de éstos fueron R = 1,2,3 y n. Además para α = 1.9 para n = 51 y n = 99 se utilizó R = 5 y R = 9 respectivamente.

Los estimadores considerados fueron la mediana y dos medias recortadas correspondientes al 25 y 50 por ciento de las m observaciones ordenadas, denotadas por MR.25 y MR.50.

La media α-recortada se define como:

$$\bar{X} = \left(P \cdot (X_{[\alpha' n+1]} + X_{[n-\alpha' n]}) + \sum_{i=[\alpha' n+2]}^{n-[\alpha' n+1]} x_i \right) \Big/ n\,(1-2\alpha')$$

con

$$P = 1 + [\alpha' n] - \alpha' n$$

$$\alpha' = \alpha/2$$

Para cada valor α y tamaño n, las medias y desviaciones estándar de la distribución Montecarlo de cada estimador β fueron calculados y los resultados concordaron con las predicciones teóricas.

En la figura 2.3.2 se muestran las desviaciones estándar del estimador β = 1 para la mediana y para MC en una muestra n = 51 y con mil replicaciones.

De nuevo sucede que R = 1 es el mejor, en el sentido que tiene la menor desviación estándar para los valores de α entre 1 y 1.5; y R=n es la mejor para α=1.9 y α= 2.

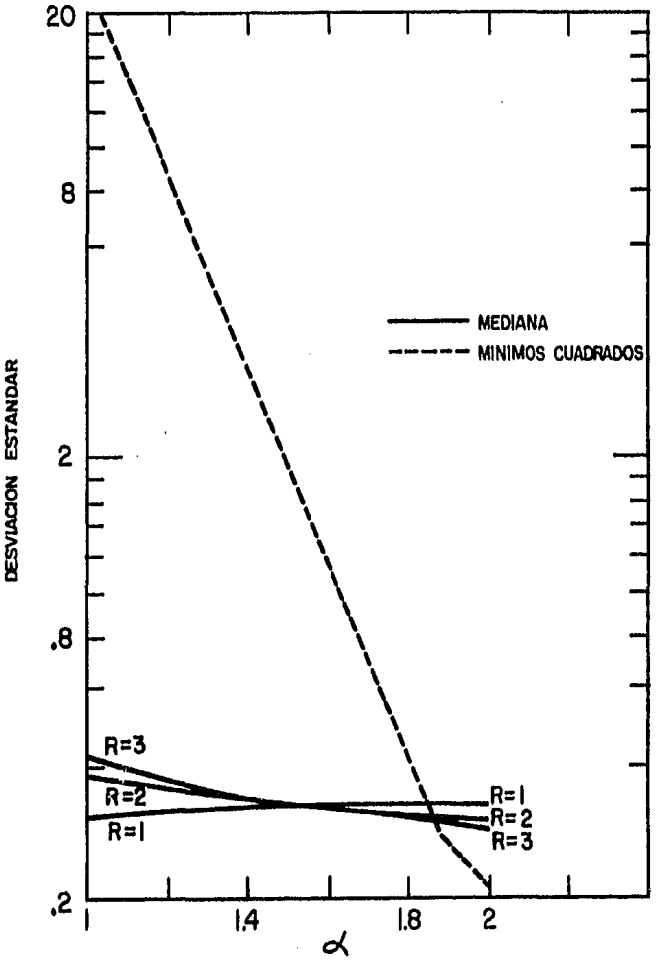


FIGURA 2.3.2 GRAFICA DE LA DESVIACION ESTANDAR DEL ESTIMADOR $\beta = 1$

En la tabla 2.3.1 se presentan los resultados de Monte carlo con 1000 replicaciones para una muestra tamaño n=51 y subgrupos tamaño R = 1,2,3 y n.

Para el caso $\alpha = 1$ la desviación estándar más pequeña se alcanza en R = 1 y para $\alpha = 2$ cuando R = n.

Ningún estimador tiene la menor desviación estándar para todos los casos, pero puede verse que MR.25 tiene un mejor comportamiento que los demás estimadores.

Haciendo uso de los resultados obtenidos, Hinich y Talwar (1975) desarrollan un procedimiento para análisis de regresión en dos etapas. En la primera etapa se utiliza R=k y como estimador preliminar se utiliza MR.25, en la segunda etapa se hace igual rechazando las observaciones con residuales grandes, como se indica adelante, y con las restantes se calcula el estimador de MC.

Para el modelo

$$yi = \beta_0 + \beta_1 \; xi + \varepsilon i, \qquad i = 1,\ldots,n.$$

se hicieron 990 replicaciones para un tamaño de muestra n=60 para cada valor = 2, 1.9, 1.5 y 1, la escala de los errores εi fue $\sigma = 1$.

En la tabla 2.3.2 se muestra el comportamiento del estimador de la primera etapa, de la segunda etapa y el de MC.

Los autores proponen rechazar las observaciones, tales que

$$\left| Yi - \beta_0 - \hat{\beta}_1 \quad xi \right| > 4 \quad S$$

donde S se define en (2.3.1) y $(\hat{\beta}_0, \hat{\beta}_1)$ es el estimador de

TABLA 2.3.1

DESVIACIONES ESTANDAR MUESTRALES DE LOS ESTIMADORES DE $\beta = 1$

		α				
R		2.	1.9	1.5	1.3	1.
	Mediana	.33	.33	.33	.32	.31
1	MR.25	.31	.31	.31	.32	.31
	MR.50	.31	.32	.33	.35	.37
	Mediana	.30	.30	.33	.35	.39
2	MR.25	.28	.29	.31	.34	.39
	MR.50	.28	.28	.32	.35	.44
	Mediana	.29	.30	.33	.36	.43
3	MR.25	.26	.27	.31	.34	.42
	MR.50	.27	.26	.31	.34	.47
n	MC	.21	.27	1.79	5.03	29.96

TABLA 2.3.2.

DESVIACIONES ESTANDAR DE LOS ESTIMADORES DE 1 ETAPA, 2 ETAPAS Y MC.

	$\beta_0 = .5$		$\beta_1 = 1$		β_{MC}	
α	ETAPA 1	ETAPA 2	ETAPA 1	ETAPA 2	β_0	β_1
2.	.29	.18	.35	.19	.18	.19
1.9	.29	.19	.36	.20	.23	.24
1.5	.32	.22	.41	.22	1.80	1.02
1.	.38	.27	.53	.28	27.15	26.16

la primera etapa.

Los estimadores de la segunda etapa tienen menor desviación estándar que los de la primera, que a su vez son mejores que los de MC, excepto para $\alpha = 1.9$ y 2.

Para el modelo $yi = \beta_1 xi_1 + \beta_2 xi_2 + \varepsilon i$. $i=1,\ldots n$, se analizaron dos tipos de rechazo en la segunda etapa.

(1) Recorte variable: rechazando yi si

$$|Yi - \hat{\beta}_1 xi_1 - \hat{\beta}_2 xi_2| > cS$$

donde $(\hat{\beta}_1, \hat{\beta}_2)$ es el estimador de la primera etapa.

Y los valores utilizados de c fueron 3 y 4.

(2) El recorte fijo: rechazando una proporción fija q de observaciones que corresponden a las qn más extremas. Fueron usados dos valores para q, q=0.1 y 0.2

En la tabla 2.3.3 se muestra el comportamiento de los estimadores utilizando los 4 rechazos y del estimador MC.

El estimador de rechazo variable con c=4 tiene las menores desviaciones estándar excepto para $\alpha = 2$ y 1. Para $\alpha = 2$ MC es el mejor y para $\alpha = 1$ el estimador de rechazo variable con c=3 es el mejor.

Este método de dos etapas permite obtener un buen estimador inicial para técnicas robustas más sofisticadas (como las vistas en las dos secciones anteriores) o para recortar aquellas observaciones que tienen residuales grandes.

Este procedimiento requiere sólo un poco más de cálculos que el de mínimos cuadrados y protege al analista de valores grandes en los residuales que son difíciles de detectar en un modelo con muchas variables explicativas.

Sin embargo, los autores no indican como proseguir cuando el tamaño de la muestra no es múltiplo de R. Esto podría resolverse de varias formas.

Si $n = mR + t \qquad o < t < R.$

puede elegirse algún grupo no de tamaño R, sino de tamaño R+t, ó podrían distribuirse las t observaciones restantes entre los m subgrupos.

TABLA 2.3.3. DESVIACIONES ESTANDAR DE LOS ESTIMADORES $\beta_1=1$ Y $\beta_2=1$ USANDO RECORTE VARIABLE, RECORTE FIJO Y MC

	RECORTE VARIABLE				MC		RECORTE FIJO			
	3S		4S				10%		20%	
α	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
2	.22	.25	.19	.21	.18[b]	.20[b]	.24	.28	.29	.35
1.9	.22	.25	.20[b]	.22[b]	.22	.25	.24	.28	.28	.35
1.5	.23	.27	.22[b]	.25[b]	1.43	1.76	.24	.27	.27	.32
1.3	.24	.28	.24[b]	.27[b]	4.06	4.99	.26	.29	.27	.31
1	.27[b]	.30[b]	.28	.31	20.36	22.4	.36	.39	.29	.33

2.4 EL METODO PROPUESTO POR ATKINSON Y COX. POR ANALISIS DISCRIMINANTE LINEAL

Se utilizará el modelo de regresión lineal simple

$$yi=\beta_0 + \beta_1 xi + \varepsilon i, \qquad i=1,\ldots,n.$$

donde los εi representan los errores aleatorios centrados alrededor del cero.

Este método propuesto por Atkinson y Cox (1977) requiere para cada valor xi, $i=1,\ldots,n$ hacer r observaciones de la variable y Se denotarán los valores ordenados de y en $x=xi$ por

$$Yi_1 \leq Yi_2 \leq \cdots \leq Yi_r$$

Ahora, un estimador lineal de localización basado en estos valores es:

$$Yi^{(a)} = \sum_{s=1}^{r} as\ Yis, \quad \sum_{s=1}^{r} as=1 \qquad (2.4.1)$$

$$Yi^{(a)} = a' Yi; \quad \text{donde } a'=(a_1, a_2,\ldots,ar),$$

son las constantes adecuadas, las mismas para toda i.
Y $Yi=(Yi_1, Yi_2,\ldots,Yi_r)'$.

Las a's deben ser escogidas de tal forma que la relación de regresión obtenida sea lo más fuerte posible, esto es, hacer que el coeficiente de correlación, R^2 sea lo más grande posible.

Si existen uno o más observaciones discrepantes, éstas afectarán el ajuste, se espera entonces que las a's den a estos valores poco o ningún peso.

Una manera de hacer esto es escoger las a's de tal forma que se maximice en la suma de cuadrados corregida por la media, la proporción debida a la regresión.

Así en: $\sum (Yi^{(a)} - \bar{Y}^{(a)})^2 = \sum (Yi^{(a)} - \hat{Y}i^{(a)})^2 + \sum (\hat{Y}i^{(a)} - \bar{Y}^{(a)})^2$

el sumando subrayado tratará de maximizarse.

Las a's son escogidas por análisis discriminante lineal. (Mardia, 1979).

El estimador robusto que se requiere es el que resulta de hacer la regresión de $yi^{(a)}$ en xi.

Sea Y la matriz de mxr observaciones y sea x el vector de variables independientes centrados de tal forma que

$$x_1 + \ldots + xm = 0.$$

Y sea X la matriz de mx2 con renglón i-ésimo (1 xi).

El cociente de la suma de cuadrados debida a la regresión entre la suma de residuales al cuadrado es:

$$\frac{(a'Y'x)^2 / (x'x)}{a'Y'(I-X(X'X)^{-1}X')Ya}$$

(Ver nota 1 final sección 2.4)

que es maximizado como función de a por el vector $\hat{\alpha}$ que satisface:

$$Y' \quad I - X\,(X'X)^{-1}X \quad Y\hat{\alpha} = Y'x \qquad (2.4.2)$$

(Ver nota 2 final sección 2.4)

Entonces los coeficientes requeridos son:

$$Cs = \frac{\hat{\alpha}_s}{\sum_{t=1}^{r} \hat{\alpha}_t}$$

y el estimador robusto para β_1 es

$$\hat{\beta}_1^{*} = (c'Y'x) \;/\; (x'x) = \sum_{i=1}^{m} xi \sum_{s=1}^{r} cs\, Yis \;/\; \sum_{s-1}^{r} xi^2$$

En lo sucesivo se considerará $m \geq 3$ y $r \leq m-2$, ya que si $r > m-2$ puede ocurrir que se de una linealidad exacta para una adecuada elección de C_s.

En el procedimiento hasta ahora descrito existen dos dificultades.

Primera, en la medida de localización (2.4.1) podría con

siderarse natural el haber impuesto la condición de no-negatividad.

Si las C's definidas anteriormente, son arbitrarias, existe la posibilidad de que el análisis produzca una combinación lineal que no podría verse con naturalidad como medida de localización y peor aún, podría llevar a un mal estimador de la pendiente. (Los autores señalan que investigaciones numéricas muestran que los estimadores que permiten C's negativas son pobres).

Dos modificaciones al procedimiento pueden hacerse:

(i) Aplicar el procedimiento descrito y desechar las observaciones con pesos (C's) negativos. Repitiendo el procedimiento completo hasta que sólo queden observaciones con pesos no-negativos.

(ii) Ajustar todas las regresiones posibles usando matriz y submatrices de Y y tomar la mejor con pesos no-negativos.

Trabajos numéricos muestran que (ii) da la solución exacta del problema de optimización con las condiciones (2.4.1) y la de no-negatividad, pero la ganancia de (ii) sobre (i) no es grande y por simplicidad los autores utilizaron (i).

Segunda, el estimador de la pendiente es no-invariante bajo transformaciones simples. Si se añade Zxi a cada Yi la pendiente cambia de β a $\beta+Z$. Los estimadores de MC lo hacen de manera similar, pero el estimador descrito por el primer procedimiento ó usando las modificaciones (i) ó (ii) no se transforma de manera simple.

Trabajos numéricos detallados muestran que:

(a) para errores con varianza finita y si se opera en una región fuerte de regresión definida por los autores como aquella que cumple que

$$|\beta| \Big/ \left[rm \left(\sigma \left(\hat{\beta}_{MC} \right) \right] \right. > 0.1$$

entonces los estimadores descritos tienen una invarianza - dentro del 1%.

(b) Si el método es usado dentro de una región de regresión cercana al cero, i.e. $\beta \simeq 0$), entonces los estimadores obtenidos son de calidad inferior.

Una consecuencia de (a) es que para datos que muestran una regresión muy ligera ($\beta \simeq 0$) es mejor añadir Zxi a cada Yi, para una Z adecuada, aplicar el procedimiento (i) y después restar Z del estimador de la pendiente resultante.

Los estudios de simulación para el procedimiento (i) se hicieron con r=5 y m=10, con $\sigma^2 = 1$.

Resultados para r=3 y 8 y para m=5 y 15 se hicieron también, pero no mostraron ningún aspecto inesperado.

Para m=10 los valores de xi se tomaron equiespaciados en (-9,9).

La varianza de $\hat{\beta}_{MC}$ fue de 0.606×10^{-3}. En las simulaciones se tomó $\beta = 1$, así que se está en una región de regresión fuerte, ya que

$$1 = |\beta| > 0.1 \; (\text{rm } \sigma (\hat{\beta}_{MC}) = 5 \; (0,303 \times 10^{-2})^{1/2}$$
$$= 0.2752272$$

La tabla 2.4.1 justifica las siguientes conclusiones:

- Si el error ϵ se distribuye normal, la pendiente estimada es insesgada y tiene una varianza que no excede en más del 25% a la del estimador de MC.
- El efecto de observaciones discrepantes fue estudiado al sumar o restar 10 de una de las observaciones para las cuales xi = 9. En este caso, el estimador de MC dió una varianza de un poco más de 4 veces la del estimador robusto $\hat{\beta}_1^*$.

TABLA 2.4.1

MEDIAS Y VARIANZAS ESTANDARIZADAS PARA ESTIMADORES ROBUSTOS Y MC CUANDO $\beta=1$

MODELO	ESTIMADOR	MEDIA	VARIANZA ESTANDARIZADA
Normal	MC	0.999	1.006
	RANGOS PONDERADOS	1.001	1.241
Normal más observación discrepante	MC	1.001	6.008
	RANGOS PONDERADOS	1.003	1.468
Laplace	MC	1.000	0.992
	RANGOS PONDERADOS	1.001	0.929
	MEJOR ESTIMADOR LINEAL	----	0.792
Cauchy	MC	1.041	14530
	RANGOS PONDERADOS	1.003	5.482
	MEDIANA	----	6.106

LA VARIANZA ESTANDARIZADA ES EL COCIENTE DE LA VARIANZA EMPIRICA ENTRE LA VARIANZA TEORICA PARA MC CON $\sigma^2 = 1$.

- Cuando el error se distribuye como Laplace, el estimador resulta ser mejor que β_{MC}.
- Si los errores son Cauchy, el estimador de MC de la pendiente tiene una varianza demasiado grande (14350) Y el estimador $\hat{\beta}_1^*$ tiene una varianza un poco menor que el estimador basado en la mediana.

En la tabla 2.4.2 se muestran los pesos promedio de 1000 replicaciones. Los pesos variaron mucho de una realización a otra; aún en el caso de la distribución normal. De aquí que los valores de los pesos promedio son útiles como guía de la distribución involucrada sólo cuando se analice un número grande de conjuntos similares de datos. En la tabla, el efecto de las observaciones discrepantes se consiguió restando 10 de las observaciones con $x_i = 9$.

Anteriormente se habló de la región de regresión fuerte que depende del valor de β_1.

Para valorar el efecto del valor de β_1 en la varianza del estimador robusto se hicieron 1000 simulaciones, ver tabla 2.4.3 , para la distribución normal y la Cauchy. Para el caso normal en que $\beta_1=0$ se tiene que la varianza es casi tres veces que la obtenida cuando $\beta_1=5$ ($\sigma(\hat{\beta}_{1_{MC}})$).Para el caso de la Cauchy la varianza es muy grande aún cuando $\beta_1=5$ ($\sigma(\hat{\beta}_{1_{MC}})$), de aquí que la regresión deberá ser más fuerte.

El método descrito requiere de r observaciones para cada x_i $i = 1, \ldots, m$, si esta situación no se tiene, los valores de x pueden agruparse artificialmente.

Hàjtovsky (1973) muestra que la agrupación de los valores de x (aún en forma no muy refinada) tiene un pequeño

TABLA 2.4.2 PROMEDIO DE PESOS DE 1,000 SIMULACIONES DE LAS ESTADISTICAS DE ORDEN EN EL ESTIMADOR ROBUSTO PROPUESTO

MODELO	ESTIMADORES	PESOS				
Normal	RANGOS PONDERADOS	0.195	0.199	0.199	0.210	0.197
	MC	0.2	0.2	0.2	0.2	0.2
Normal más una observación discrepante negativa.	RANGOS PONDERADOS	0.015	0.264	0.246	0.250	0.225
	MC OMITIENDO LA OBSERVACION DISCREPANTE	0	0.25	0.25	0.25	0.25
Laplace	RANGOS PONDERADOS	0.083	0.246	0.340	0.247	0.084
	MEJOR ESTIMADOR LINEAL	0.017	0.221	0.524	0.221	0.017
Cauchy	RANGOS PONDERADOS	0.019	0.206	0.539	0.221	0.015
	MEDIANA	0	0	1	0	0

TABLA 2.4.3 DEPENDENCIA DE LA VARIANZA ESTANDARIZADA DE LOS ESTIMADORES ROBUSTOS CON EL VALOR DE β

		$\beta/\sqrt{\text{var}(\beta_{MC})}$		
MODELO	ESTIMADOR	0	1	5
Normal	Rangos Ponderados	3.760	2.377	1.180
Cauchy	Rangos Ponderados	34.80	41.42	14.95

efecto sobre el análisis de MC, a menos de que haya pocas observaciones extremas y lo mismo sucede en este método propuesto, por Atkinson y Cox. (1977).

Atkinson y Cox (1977) hacen notar que si la varianza de ε varía mucho con x, o sea cuando existe heteroscedasticidad, este método de pesos asignados según los rangos, resulta peor que el método de MC.

NOTA 1

Como

$$a' = (a_1, a_2, \ldots, a_r)$$

$$Y = \begin{bmatrix} Y1_1 & Y1_2 \ldots & Y1_r \\ Y2_1 & Y2_2 \ldots & Y2_r \\ Yn_1 & Yn2 \ldots & Yn_r \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_n \end{bmatrix}$$

$yi^{(a)}$ Puede escribirse como

$yi^{(a)} = \underset{\sim}{a}'Yi'$ Yi el i-ésimo renglón de Y

Entonces a'Y' es el vector que contiene las cantidades sobre las cuales se hará la regresión en x.

La suma de cuadrados debida a la regresión es

$$\hat{\beta}1\left[\sum xiyi^{(a)} - \frac{(\sum xi)\ (\sum yi^{(a)})}{n}\right]$$

como $\sum xi = 0$, entonces puede escribirse como

$$\hat{\beta}_1\ (\sum xi\ yi^{(a)})$$

y $\hat{\beta}_1$ es: $\hat{\beta}_1 = \dfrac{\sum(xi - \bar{x})\ (yi^{(a)} - \bar{y}^{(a)})}{\sum(xi-\bar{x})^2} = \dfrac{\sum xi(yi^{(a)} - \bar{y}^{(a)})}{\sum xi^2}$

$$\hat{\beta}_1 = \frac{\sum xi\ yi^{(a)} - \sum xi\ (\bar{y}^{(a)})}{\sum xi^2} \quad \text{y como } \sum xi = 0$$

$$\hat{\beta}_1 = \frac{\sum xi\ yi^{(a)}}{\sum xi^2}$$

escribiéndolo en forma matricial se tiene la suma de cuadrados debida a la regresión es

$$\frac{(a'Y'x)^2}{x'x}$$

Ahora suma de residuales al cuadrado en forma matricial es:

$$a'Y' \left(I - X \left(X'X \right)^{-1} X' \right) Ya$$

Por tanto, el cociente de la suma de cuadrados debida a la regresión entre la suma de residuales al cuadrado es

$$\frac{(a'Y'x)^2 \ / \ x'x}{a'Y'(I-X(X'X)^{-1}X') \ Ya}$$

NOTA 2

Maximizar $$\frac{(\alpha' Y'x)^2 \;/\; (x'x)}{\alpha'Y'(I-X'(X'X)^{-1} X')\, Y\alpha}$$

es equivalente a maximizar con respecto a

$$\frac{(\alpha'Y'x)^2}{\alpha'A\alpha}$$

donde $A = Y'(I-X(X'X)^{-1} X')\, Y$, A es simétrica.

Derivando, se tiene

$$\frac{\partial}{\partial \alpha}\left(\frac{(\alpha'Y'x)^2}{\alpha'A\alpha}\right) = \frac{(\alpha' A\alpha)\; 2\;(\alpha'Y'x)\;(Y'x) - (\alpha'Y'x)^2\; 2\;(A\alpha)}{(\alpha'A\alpha)^2}$$

Igualando a cero.

$$2\,(\alpha'Y'x)\left[(\alpha' A\alpha)\;(\alpha Y' x) - (\alpha' Y' x)\;A\alpha\right] = 0$$

Si $\alpha'Y'x=0$ significaría que el coeficiente de correlación es cero.

Entonces

$$(\alpha' A\alpha)\,(Y'x) - (\alpha' Y'x)\;A\alpha = 0$$

de aquí que

$$Y'x = \frac{\alpha'Y'x}{\alpha'A\alpha} \qquad (2.4.3)$$

La solución dada en (2.4.2) puede escribirse como

$$A\hat{\alpha} = Y'x \qquad \text{de aquí se tiene}$$

$$\hat{\alpha} = A^{-1}Y'x \qquad (2.4.4) \text{ y por ser A simétrica}$$

$$\hat{\alpha}' = x'Y\,A^{-1} \qquad (2.4.5)$$

(2.4.4) satisface (2.4.3), es decir

$$Y'x = \frac{(x'Y\,A^{-1})\;Y'x \quad A\,A^{-1}\,Y'x}{(x'Y\,A^{-1})A(A^{-1}\,Y'x)}$$

$$Y'x = \frac{x'Y\ A^{-1}\ Y'\ x}{x'Y\ A^{-1}\ Y'x} \qquad Y'x = Y'x$$

2.5 LOS ESTIMADORES DE NORMA Lp EN EL MODELO DE REGRESION LINEAL.

Los estimadores de norma Lp (estimadores Lp) son aquellos que minimizan la suma de las desviaciones absolutas del hiperplano de regresión lineal elevados a la potencia p.

Son una generalización de la técnica de MC.

Así para el modelo lineal

$$Y = X\underset{\sim}{\beta} + \varepsilon$$

El estimador de norma Lp es aquel que minimiza

$$\sum_{i=1}^{n} |ei|^p = \sum_{i=1}^{n} \left| yi - xi\hat{\underset{\sim}{\beta}} \right|^p$$

Para $p \neq 2$ los estimadores que se obtienen son no lineales, de aquí que pueden tener varianza menor que los de MC.

El caso $p = 1$ puede formularse como un problema de programación lineal, considerando a ei como la diferencia de dos variables no-negativas, Wagner (1959). Esto es, si

$$ei = ui - vi \ ,$$

$$\text{minimizar} \quad \sum_{i=1}^{n} |(ui + vi)|$$

sujeto a

$$yi = \sum_{i=1}^{k} xij \ \hat{\beta}_j + ui - vi,$$

βj sin restricción de signo,

$ui, vi \geqslant 0$

El caso $p = \infty$ corresponde a la minimización del máximo error, se le conoce también como la estimación de Chebychev. Wagner (1959) muestra que el problema de estimación L_∞ también puede formularse como uno de programación lineal, como se muestra en seguida:

Sea $D = \text{Máx} \{ |e_t| \}$

Minimizar D

Sujeto a

$$D \geq Y_t - X_t \hat{\beta} = e_t,$$
$$D \geq -Y_t + X_t \hat{\beta} = -e_t,$$
$$D \geq 0$$

$\hat{\beta}$ sin restricción de signo.

No existe ninguna razón teórica por la cual no deban considerarse otros valores de p diferentes a 1, 2 ó ∞

En la sección 3.2 se presenta un estudio comparativo de estimadores Lp, con valores de

$$p = 1, 1.25, 1.5, 1.75, 2.0 \text{ ó } \infty .$$

CAPITULO 3

COMPARACION Y ELECCION DE ESTIMADORES ROBUSTOS

3.1 ESTUDIO COMPARATIVO DE ESTIMADORES

ROBUSTOS

A continuación se presenta el estudio realizado por Heiler (1981) en el que se comparan los siguientes estimadores: dentro de los del tipo M:

	Denotados por:
- Mínimos cuadrados	MC
- Andrews	AN
- Huber	HU
- Hampel	HA
- Tukey	TU

Usando dos propuestas para estimar la escala. La primera s_1, que se determina de manera simultánea con $\underline{\beta}$, de la ecuación

$$\sum_{t=1}^{n} \chi\left(\frac{yt - xt'\hat{\underline{\beta}}}{s_1}\right) = 0 \quad \text{(propuesto por Huber, 1964)}$$

donde

$$\chi = \psi^2(x) - E_\phi(\psi^2)$$

y E_ϕ es la esperanza bajo la distribución normal.

La segunda, s_2, es una propuesta de Hampel, y resultó ser un estimador más robusto que s_1, su expresión es la siguiente:

$$s_2 = \frac{(\text{mediana} \left| e(t) - e.5 \right|}{0.6754}$$

donde $e_{.5}$ es la mediana de los residuales ordenados

$e(t)$, $t = 1, \ldots, n$.

Los cuatro estimadores fueron utilizados con la escala s_1 (AN_1, HU_1, HA_1, TU_1), y con la escala s_2 (AN_2, HU_2, HA_2 TU_2).

Dentro de los del tipo L considera entre otros:

	Denotado por:
- La mediana	MED
- La trimedia	TRI
- El de Gastwirth	GAS

Y dentro de los estimadores R incluye a los estimadores que se obtienen al utilizar las siguientes funciones de puntaje:

Puntajes	Denotado por:
- de Wilcoxon	WIL
- de la mediana	MSC
- de la normal	NOR
- de Vander Warden	VDW

Las propuestas de Atkinson y Cox, Hinich y Talwar, Hill y Holland y los estimadores Lp no están comprendidos en este estudio.

Pero aún así este estudio da una visión general del comportamiento de los estimadores robustos, ya que:

i) el estimador de Atkinson y Cox se utiliza únicamente cuando se tiene una sola variable explicativa y se cuenta - con r observaciones para cada xi.

ii) el de Hinich y Talwar por su fácil obtención puede - ser utilizado como un estimador inicial para técnicas más sofisticadas, tal como la de Andrews, Huber, etc.

iii) El estimador de Hill y Holland es un caso especial de la propuesta de Andrews, ya que utiliza al estimador L_1 como estimador inicial y después utiliza una sola iteración del proceso.

y iv) los estimadores Lp, en muchas ocasiones no son considerados como robustos. *

El estudio de Heiler (1981) simula cuatro situaciones:

El diseño 1:

$$yi = 1 + 0.6\, x_{2i} + ei \quad \text{con } i = 1, \ldots, 20$$

$$x_{2i} = -.95 + .1 \quad (i - 1).$$

Que corresponde a un modelo de regresión lineal simple, con un diseño balanceado en el que la variable x_{2i} se encuentra equiespaciada.

El diseño 2:

$$yi = 1 + 0.5\, x_{2i} + ei \text{ con } i = 1, \ldots, 40$$

$$x_{2i} = -0.342 + 0.00064\ i\ (i-1)$$

Con un diseño desbalanceado en la variable x_{2i}.

El diseño 3:

$$yi = 1 + 0.5\, x_{2i} + 0.25\, x_{3i} + ei \text{ con } i = 1, \ldots, 30$$

$$x_{2i} = (2i - 31)\ /20$$

$$x_{3i} = 0.601481 + (i - 1)\ (i - 30)\ /\ 225$$

En este caso, las variables explicativas son ortogonales.

El diseño 4:

$$yi = 1 + 0.5\, x_{2i} + 0.25\, x_{3i} + ei \text{ con } i = 1, \ldots, 30$$

$$x_{2i} = -0.34435 + 0.001149 \quad i(i - 1)$$

$$x_{3i} = (x_{2i} + 0.34465)^2 - 0.21374$$

Aquí se presenta una alta colinealidad (correlación = 0.96) entre las variables explicativas.

* Un estudio comparativo de estos estimadores se presenta más adelante.

Para generar las distribuciones de los errores (18 distribuciones en total) Heiler utiliza un programa de números pseudoaleatorios de una distribución uniforme (0, 1).- Como distribución de referencia se utilizó la normal (0,0.02) Todas las demás distribuciones fueron "estandarizadas" de manera que su densidad tuviese el mismo valor en cero que la - distribución de referencia.

Las distribuciones asimétricas fueron escogidas de manera que su esperanza fuera cero.

Este es el índice de las distribuciones del error consideradas

Simétricas:	Denotadas por:
- Normal	N
- Doble exponencial (Laplace)	DE
- t_5 de Student (5 grados de libertad)	T
- Cauchy	C
- Logística	L

Asimétricas:	Denotadas por:
- χ_1^2	JI
- Log-normal	LN

Normales contaminadas:	Denotadas por:
$F = (1-C)\ N\ (0,0.02) + CN_j(0,\sigma^2 j)$	N5 N1
con $P\ (C=1) = d\ P\ (C=0) = 1-d$	N10 N1
$d = 0.05,\ 0.10,\ 0.20$	N20 N1
$j = 1,2$	N5 N2
$N_1 = N\ (0, \sigma_1^2),\ \sigma_1^2 = 0.25$	N10 N2
$N_2 = N\ (0, \sigma_2^2),\ \sigma_2^2 = 2.25$	N20 N2
$F = (1-C)\ N\ (0,\ 0.02) + cC$	N5 C
$d = 0.05,\ 0.10,\ 0.20$	N10 C
	N20 C

Normal autocorrelacionada: Denotada por:

$ei = 0.5\, e_{i-1} + \varepsilon i$, $\varepsilon i \sim N(0, 0.02)$ NA

Normal contaminada autocorrelacionada:

$ei = 0.5\, e_{i-1} + i, \delta i \sim N1ON1$ NCA

Propiedades de los Estimadores:

Para cada una de las 18 distribuciones se generaron 500 replicaciones de números pseudoaleatorios y se aplicaron a los cuatro diseños. Con estas 500 simulaciones las propiedades de los estimadores fueron evaluadas por los siguientes cuatro criterios.

(i) La media de las estimaciones

$$M(\hat{\beta}i) = \frac{1}{M} \sum_{j=1}^{M} \beta ij, \quad i = 1,2,3 \quad M = 500$$

donde βij es el estimador de βi, en la j-ésima simulación y $\beta_1 = 1$, $\beta_2 = 0.5$ y $\beta_3 = 0.25$

(ii) La media de los errores de los estimadores al cuadrado

$$MEC(\hat{\beta}i) = \frac{1}{M} \sum_{j=1}^{M} (\hat{\beta}ij - \beta i)^2 \quad i = 1,2,3.$$

(iii) La media de las desviaciones de la recta de regresión (ó del plano en su caso) al cuadrado.

$$MDC = \frac{1}{M} \sum_{j=1}^{M} \left(\sum_{i=1}^{n} (\hat{y}ij - yi^{\circ})^2 \right)$$

con $yi^{\circ} = 1 + 0.5\, x_{2i}\ (+ 0.25\, x_{ei})$ y $\hat{y}ij$ el estimador de yi° en la j-ésima simulación.

(iv) Para cada diseño y para cada uno de los 18 errores los estimadores fueron ordenados según su rango de acuerdo a los valores de MDC y se calcularon las deficiencias DEF.

DEF = 1 - MDC del mejor estimador / MDC

(v) El tiempo de cómputo requerido.

Resultados del Estudio.

La influencia del diseño.-

No se encontró una clara tendencia, de que un procedimiento ó una familia de estimadores tuviera alguna ventaja ó desventaja específica en alguno de los cuatro diseños.

Las posiciones en el ordenamiento de cada uno de los procedimientos varían muy poco de un diseño a otro, y en los casos donde hay mayores diferencias, los valores de MDC difieren muy poco. Un resultado interesante de este estudio es - que las ventajas y desventajas de los procedimientos dependen principalmente en el tipo de distribución del error y poco en el diseño específico de la situación.

En las distribuciones simêtricas.

Para las distribuciones simêtricas, la mayoría de los estimadores alcanzaron casi la misma precisión, tanto para el término constante como para las pendientes.

Los estimadores R WIL y VDW fueron ligeramente inferiores en la estimación del têrmino constante, comparado con la estimación de las pendientes.

Para el caso de la distribución normal N (0,0.02) , MC alcanzó los valores más pequeños en MDC, como era de esperarse, las siguientes posiciones fueron tomadas por los estimadores M (descendientes) AN1, TU1 y HA1 (en este orden), con deficiencias que no excedieron el 3.6% (ver tabla 3.1.1). Los valores de MDC dependen directamente de la longitud de la fase creciente de la función de peso correspondiente a cada estimador. Ver la figura 3.1.1 Siguiendo al anterior grupo M1 están los estimadores R, VDW y NOR, y todavía son considerados como buenos estimadores: WIL, HU1 y AN_2.

Pero para el caso de la normal con 5% de contaminación

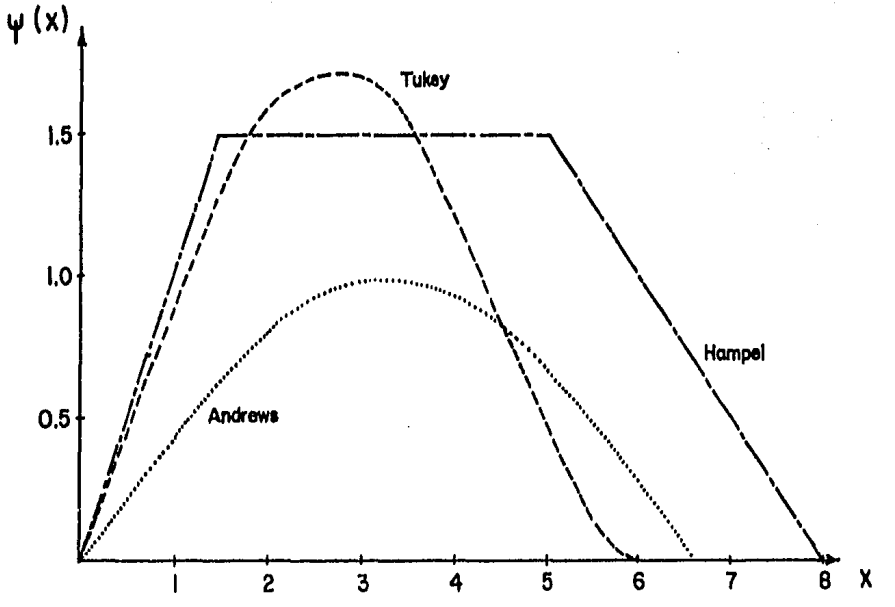


FIGURA 3.1.1 FUNCIONES DE INFLUENCIA DE LOS ESTIMADORES M

TABLA 3.1.1

MEDIA DE LAS DESVIACIONES DEL PLANO DE REGRESION AL CUADRADO

DE 10 ESTIMADORES PARA N (0, 0.02).

	DISEÑO 1			DISEÑO 2			DISEÑO 3			DISEÑO 4		
ESTIMADOR	MDC	DEF	RANGO	MDC	DEF	RANGO	MDC	DEF	RANGO	MDC	DEF	RANGO
CLS	0.0405	0.0	1	0.0395	0.0	1	0.0597	0.0	1	0.0606	0.0	1
HU_1	0.0445	0.090	5	0.0433	0.088	7	0.0653	0.386	7	0.0668	0.093	7
AM_1	0.0408	0.007	2	0.0399	0.010	2	0.0601	0.007	2	0.0610	0.007	2
HA_1	0.0420	0.036	3	0.0407	0.029	3	0.0613	0.026	3	0.0621	0.024	3
HU_2	0.0495	0.182	8	0.0473	0.165	9	0.0740	0.193	8	0.0759	0.202	9
AN_2	0.0453	0.106	7	0.0425	0.071	6	0.0651	0.083	6	0.0661	0.083	6
HA_2	0.0496	0.183	9	0.0468	0.156	8	0.0742	0.195	9	0.0758	0.201	8
WIL	0.0446	0.093	6	0.0424	0.068	5	0.0643	0.072	5	0.0655	0.075	5
VDW	0.0427	0.052	4	0.0414	0.045	4	0.0618	0.034	4	0.0629	0.036	4
GAS	0.0505	0.198	10	0.0491	0.196	10	0.0767	0.222	10	0.0772	0.215	10

Cauchy MC falla completamente, mientras que otros estimadores se enfrentan a esta situación muy bien. Los estimadores M,AN1, TU1 y HA1 tienen los valores menores de MDC, seguidos por AN2 y los estimadores R, VDW y NOR.

Para el caso donde hay un 20% de contaminación Cauchy, el grupo que va a la cabeza es HA1, TU1 y AN2, seguidos por HU1, AN1 y de nuevo éstos seguidos por WIL, VDW y NOR.

De las tablas 3.1.2 y 3.1.3 puede verse que cuando la proporción de contaminación crece ó el tipo de contaminación es más severa (N20 N1, N10 N2, N20 N2), los estimadores M con la escala estimada s_2 (más robusta) mejoran tomando las primeras posiciones, mientras que las posiciones de los estimadores R se deterioran.

En la tabla 3.1.3 puede verse que el estimador GAS es una alternativa muy aceptable.

Para el caso de la Cauchy, los valores de MDC varían considerablemente de un procedimiento a otro. Los estimadores basados en la mediana, MED y MSC son buenas alternativas a HA_2, TU_2 y HU_2.

Para la normal autocorrelacionada se tienen más o menos las mismas posiciones que en la normal, pero los valores de MDC son casi del triple. Pero los valores más altos para MDC se obtuvieron de la normal contaminada autocorrelacionada (NCA) y las posiciones para ésta fueron primero TU_2 HA_2 seguidos por AN_2, HU_1, HU_2, y esto a su vez, seguidos por WIL, TRI y GAS.

En las distribuiones asimétricas.

Para las situaciones con error con distribución asimétrica, el término constante estimado resulta ser sesgado en dirección de la mediana, para la mayoría de los procedimientos.

Los estimadores M que utilizan la escala robusta s_2 son

TABLA 3.1.2

MEDIA DE LAS DESVIACIONES DEL PLANO DE REGRESION AL CUADRADO DE 10 ESTIMADORES

PARA N (0, 0.2) 5N (0,0.25)

SIMETRIA POCO CONTAMINADA)

	DISEÑO 1			DISEÑO 2			DISEÑO 3			DISEÑO 4		
ESTIMADOR	MDC	DEF	RANGO	MDC	DEF	RANGO	MDC	DEF	RANGO	MDC	DEF	RANGO
MC	0.0598	0.204	10	0.0503	0.225	10	0.0967	0.260	10	0.0962	0.257	10
HU_1	0.0487	0.023	3	0.0470	0.038	5	0.0741	0.034	3	0.0738	0.031	3
AN_1	0.0492	0.033	4	0.0468	0.034	4	0.0746	0.040	4	0.0757	0.055	5
HA_1	0.0476	0.0	1	0.0452	0.0	1	0.0716	0.0	1	0.0715	0.0	1
HU_2	0.0537	0.114	7	0.0511	0.115	8	0.0811	0.117	8	0.0810	0.117	7
AN_2	0.0479	0.006	2	0.0455	0.007	2	0.0731	0.021	2	0.0730	0.021	2
HA_2	0.0538	0.115	8	0.0499	0.094	7	0.0794	0.098	7	0.0817	0.125	8
WIL	0.0497	0.042	6	0.0465	0.027	3	0.0753	0.050	5	0.0753	0.050	4
VDW	0.0495	0.039	5	0.0475	0.049	6	0.0767	0.066	6	0.0770	0.072	6
GAS	0.0546	0.128	9	0.0528	0.144	9	0.0849	0.157	9	0.0835	0.144	9

TABLA 3.1.3

MEDIA DE LAS DESVIACIONES AL PLANO DE REGRESION AL CUADRADO DE 10 ESTIMADORES PARA

N (0, 0.25) 20N (0, 2.25).

(CONTAMINADA SEVERAMENTE).

	DISEÑO 1			DISEÑO 2			DISEÑO 3			DISEÑO 4		
ESTIMADOR	MDC	DEF	RANGO	MDC	DEF	RANGO	MDC	DEF	RANGO	MDC	DEF	RANGO
MC	0.9133	0.905	10	0.9367	0.934	10	1.4727	0.934	10	1.4792	0.922	10
HU_1	0.1175	0.262	5	0.0859	0.279	4	0.1773	0.448	5	0.1933	0.401	4
AN_1	0.3544	0.755	9	0.1952	0.633	9	0.4830	0.198	9	0.5268	0.780	9
HA_1	0.1566	0.446	7	0.0886	0.301	6	0.1886	0.481	6	0.2097	0.448	5
HU_2	0.1013	0.144	3	0.0786	0.212	3	0.1398	0.300	3	0.1713	0.324	3
AN_2	0.0970	0.106	2	0.0668	0.073	2	0.1094	0.106	2	0.1195	0.031	2
HA_2	0.0867	0.0	1	0.0619	0.0	1	0.0978	0.0	1	0.1158	0.0	1
WIL	0.1243	0.302	6	0.0969	0.361	7	0.2060	0.525	7	0.2214	0.477	7
VDW	0.1655	0.476	8	0.1312	0.528	8	0.2774	0.647	8	0.2970	0.610	8
GAS	0.1074	0.193	4	0.0885	0.301	5	0.1746	0.440	4	0.2140	0.459	6

muy sensibles a la asimetría, mientras que los que utilizan la escala s_1 apenas se ven influenciados.

Para el caso de la JI, MC tiene los menores valores de MEC para el término constante, pero los valores de MEC para las pendientes están entre los más grandes.

Las mejores posiciones tanto para el término constante como para las pendientes fueron tomadas por los estimadores R NOR y VDW (ver tabla 3.1.4). El estimador WIL tuvo la tercera posición para las pendientes, pero fue ligeramente inferior para la estimación del término constante. El estimador AN_1 funcionó muy bien. En general los estimadores M tuvieron posiciones medianas para la estimación de pendientes pero dieron estimadores pobres del término constante.

De estos resultados, puede decirse que los estimadores M adecuadamente escogidos, pueden hacer frente a casi cualquier tipo de situación concebible. En los casos cuando la contaminación en la distribución normal es grande, ó la distribución tiene colas pesadas y no se espera asimetría, el estimador robusto de la escala s_2 es superior a s_1. Pero esta escala s_2 no debe utilizarse si la suposición de asimetría en la distribución no puede excluirse.

Los estimadores HU_1 y AN_2 estuvieron entre los mejores estimadores (ver tablas 3.1.5 y 3.1.6). HA y TU mostraron resultados similares, y en las situaciones más extremas (C, L, DE) con la escala s_2 puede decirse que tuvieron resultados excelentes.

En varias ocasiones los estimadores R son buenas alternativas, con la ventaja de que son invariantes bajo escala y de que no necesitan de un estimador inicial de $\underset{\sim}{\beta}$, pero por otro lado éstos consumen más tiempo de cómputo.

En las situaciones con asimetría en la distribución del error VDW y NOR logran muy buenas posiciones, pero en las situaciones simétricas sólo alcanzan buenas posiciones en aquellas distribuciones que no se alejan mucho de la normal. En

TABLA 3.1.4

MEDIA DE LAS DESVIACIONES DEL PLANO DE REGRESION AL CUADRADO DE 10 ESTIMADORES PARA LA JI (ASIMETRICA).

	DISEÑO 1			DISEÑO 2			DISEÑO 3			DISEÑO 4		
ESTIMADOR	MDC	DEF	RANGO	MDC	DEF	RANGO	MDC	DEF	RANGO	MDC	DEF	RANGO
MC	0.0315	0.269	5	0.0291	0.127	2	0.0449	0.377	6	0.0440	0.321	5
HU_1	0.0313	0.264	4	0.0499	0.491	6	0.0441	0.366	5	0.0453	0.340	6
AM_1	0.0302	0.238	3	0.0363	0.300	3	0.0402	0.304	3	0.0400	3.253	3
HA_1	0.0319	0.278	6	0.0461	0.449	5	0.0437	0.360	4	0.0435	0.313	4
HU_2	0.0422	0.454	9	0.0732	0.0653	9	0.0611	0.542	9	0.0622	0.519	9
AN_2	0.0383	0.399	7	0.0593	0.572	7	0.0556	0.497	7	0.0580	0.485	7
HA_2	0.0517	0.555	10	0.0947	0.732	10	0.0770	0.637	10	0.0807	0.630	10
WIL	0.0276	0.166	2	0.0396	0.359	4	0.0380	0.264	2	0.0398	0.248	2
VDW	0.0230	0.0	1	0.0254	0.0	1	0.0280	0.0	1	0.0299	0.0	1
GAS	0.0411	0.440	8	0.0723	0.649	8	0.0610	0.542	8	0.0610	0.510	8

TABLA 3.1.5

DEFICIENCIAS Y RANGOS DE HU_1 (ESTIMADOR M)

	DISEÑO 1		DISEÑO 2		DISEÑO 3		DISEÑO 4	
DISTRIBUCION	DEF	RANGO	DEF	RANGO	DEF	RANGO	DEF	RANGO
N	0.0899	5	0.0878	7	0.0858	7	0.0928	7
T	0.0	1	0.0	1	0.0041	2	0.0352	3
LN	0.0729	5	0.0782	7	0.0702	6	0.0775	6
J1	0.2644	4	0.4911	8	0.3658	5	0.3402	6
N_5N1	0.0226	3	0.0383	5	0.0337	3	0.0312	3
$N1OM_1$	0.0138	2	0.0439	3	0.0293	2	0.0243	2
N2ON1	0.0	1	0.0373	3	0.0507	4	0.0386	3
N5 C	0.0730	5	0.0734	6	0.0700	5	0.0611	6
N1OC	0.0600	4	0.0626	6	0.0549	4	0.0556	4
N2OC	0.0381	3	0.0505	4	0.0456	4	0.0590	3
N5N2	0.0817	3	0.1029	4	0.1057	4	0.1122	5
N1ON2	0.1352	4	0.1652	4	0.1942	4	0.1971	4
N2ON2	0.2621	6	0.2794	4	0.4484	5	0.4009	4
L	0.0079	2	0.0	1	0.0140	2	0.0305	5
DE	0.0520	4	0.0736	5	0.0298	2	0.0078	3
C	0.3879	6	0.2829	6	0.3170	5	0.2103	3
AN	0.0390	7	0.0355	7	0.0371	7	0.0346	7
ACN	0.0292	5	0.0510	4	0.0198	3	0.0260	3

TABLA 3.1.6

DEFICIENCIAS Y RANGOS DE AN_2 (ESTIMADOR M)

	DISEÑO 1		DISEÑO 2		DISEÑO 3		DISEÑO 4	
DISTRIBUCION	DEF	RANGO	DEF	RANGO	DEF	RANGO	DEF	RANGO
N	0.1060	7	0.0706	6	0.0829	6	0.0832	6
T	0.0407	4	0.0432	6.5	0.0357	4	0.0352	3
LN	0.0895	7	0.0609	6	0.0731	7	0.0789	7
J1	0.3989	9	0.5718	9	0.4970	9	0.4846	9
N5N1	0.0063	2	0.0066	2	0.0205	2	0.0205	2
N10N1	0.0	1	0.0	1	0.0	1	0.0	1
N20N1	0.0360	4	0.0315	2	0.0465	3	0.0540	4
N5C	0.0668	4	0.0404	5	0.0590	3	0.0387	3
N10C	0.0457	3	0.0210	3	0.0387	3	0.0244	3
N20C	0.0244	2	0.0	1	0.0108	2	0.0	1
N5N2	0.0	1	0.0	1	0.0209	2	0.0	1
N10N2	0.0	1	0.0	1	0.0	1	0.0	1
N20N2	0.1062	2	0.0734	2	0.1060	2	0.0310	2
L	0.0492	6	0.0356	5	0.0346	6	0.0467	6
DE	0.1007	8	0.1435	8	0.0973	7	0.0410	6
C	0.1822	3	0.1863	4	0.1964	3	0.2509	6
AN	0.0252	5	0.0197	6	0.0382	6	0.0332	6
ACN	0.0236	3	0.0295	2	0.0051	2	0.0	1

cambio WIL tiene una amplia aplicabilidad en los casos no extremos (ver tabla 3.1.7)

Los estimadores L alcanzan buenas posiciones sólo en - situaciones muy extremas (C, DE, N29 N2). En particular GAS tuvo algunos buenos resultados para N20 N1, N20 N2, DE y NCA (ver tabla 3.1.8). Pero en general los alcances de los estimadores L van de mediocres a malos.

Este trabajo que Heiler (1981) realiza es un estudio - de gran utilidad, ya que considera bastantes tipos de distribuciones para el error y de estimadores, pero sin llegar a - ser éstos demasiados como para crear una situación confusa al momento de querer elegir alguno como alternativa.

El autor no señala por qué elige como distribución de referencia a la normal con varianza 0.02 y media 0.

Como ya se ha dicho antes, el estimador inicial utilizado para obtener los estimadores M es de gran importancia, sin embargo Heiler no hace mención alguna acerca de los estimadores iniciales que eligió para su estudio.

TABLA 3.1.7. DEFICIENCIAS Y RANGOS DE WIL (ESTIMADOR R)

	DISEÑO 1		DISEÑO 2		DISEÑO 3		DISEÑO 4	
DISTRIBUCION	DEF	RANGO	DEF	RANGO	DEF	RANGO	DEF	RANGO
N	0.0925	6	0.0684	5	0.0719	5	0.0753	5
T	0.0360	3	0.0286	2	0.0208	3	0.0404	5
LN	0.0748	6	0.0519	5	0.0566	5	0.0540	5
J1	0.1664	2	0.3589	4	0.2639	2	0.2483	2
N5N1	0.0421	6	0.0274	3	0.0497	5	0.0499	4
N1ON1	0.0489	4	0.0614	5	0.0558	5	0.0598	5
N2ON1	0.0514	5	0.1045	8	0.1019	6	0.0832	6
N5C	0.0737	6	0.0359	4	0.0708	6	0.0589	5
N1OC	0.0711	6	0.0413	4	0.0640	5	0.0609	5
N2OC	0.0548	4	0.0563	5	0.0761	6	0.0878	6
N5M2	0.1212	5	0.1092	6	0.1316	6	0.1416	6
N1ON2	0.1950	7	0.2115	6	0.2639	7	0.2494	6
N2ON2	0.3025	8	0.3613	7	0.5253	8	0.4769	7
L	0.0344	4	0.0134	3	0.0168	4	0.0252	3
DE	0.0735	6	0.0952	6	0.0452	3	0.0117	4
C	0.3953	7	0.3379	9	0.3689	7	0.2960	8
AN	0.0321	6	0.0185	5	0.0366	5	0.0304	5
ACN	0.0504	7	0.0809	7	0.0447	5	0.0486	5

TABLA 3.1.8

DEFICIENCIAS Y RANGOS DE GAS (ESTIMADOR L)

	DISEÑO 1		DISEÑO 2		DISEÑO 3		DISEÑO 4	
DISTRIBUCION	DEF	RANGO	DEF	RANGO	DEF	RANGO	DEF	RANGO
N	0.1980	12.5	0.1955	14	0.2216	12	0.2150	12
T	0.0730	10	0.0719	9	0.1066	9	0.1115	9
LN	0.1844	13	0.2012	14	0.2116	12	0.2065	12
JI	0.4398	12	0.6488	12	0.5415	11	0.5100	10
N5N1	0.1292	11	0.1439	13	0.1567	11	0.1437	9
N10N1	0.0970	9	0.1407	13	0.1106	9	0.1136	7
N20N1	0.0574	6	0.1029	7	0.0797	5	0.0900	5
N5C	0.1832	11	0.1789	13	0.1944	11	0.1752	10
N10C	0.1654	11	0.1637	13	0.1696	10.5	0.1584	9
N20C	0.1209	9	0.1277	13	0.1417	9	0.1447	8
N5N2	0.1616	9	0.1926	13	0.2209	10	0.2080	9
N10N2	0.1767	6	0.2301	8	0.2543	6	0.2835	7
N20N2	0.1927	4	0.3006	5	0.4399	4	0.4589	6
L	0.0888	12	0.0791	14	0.1259	11	0.1242	9
DE	0.0293	2	0.0320	2	0.0482	4	0.0033	2
C	0.2862	4	0.1595	3	0.3301	6	0.1931	2
AN	0.0931	12	0.1006	14	0.0860	10	0.0612	9
ACN	0.0279	4	0.0535	5	0.0512	6	0.0541	6

3.2 ESTUDIO COMPARATIVO DE LOS ESTIMADORES Lp

Money et al (1982) presentan un estudio de simulación para valores de p=1.00, 1.25, 1.5, 1.75, 2.00 é ∞. Para p=1, 2 é ∞ las soluciones que se obtienen son exactas, mientras que para los otros valores de p, las soluciones son halladas por métodos iterativos hasta alcanzar un cierto grado de convergencia.

El estudio se hizo con el modelo de simulación

$$yi = \beta_0 + x_{1i}\,\beta_1 + x_{2i}\,\beta_2 + \varepsilon i \qquad i=1,\ldots,n$$

El tamaño de la muestra fue n=25 para todas las repeticiones. Se hicieron 500 repeticiones para cada tipo de error estudiado. Se seleccionaron arbitrariamente $\beta_0 = 10$, $\beta_1 = 8$ y $\beta_2 = -6$. Los 25 valores de x_1 y los 25 de x_2 fueron seleccionados aleatoriamente de una distribución uniforme en el rango (0, 40). Se verificó que x_1 y x_2 fuesen no-correlacionadas (es decir $|\rho x_1 x_2| * < 0.01$)

Para cada repetición se calcularon 25 errores, tales que $E(\varepsilon i) = 0$ y $var(\varepsilon i) = 9$, $i = 1, 2 \ldots, 25$, calculando los yi usando el modelo $yi = 10 + 8\,x_{1i} - 6\,x_{2i} + \varepsilon i, i = 1, 2, \ldots, 25$, obteniendo así los distintos estimadores Lp de β_0, β_1, y β_2.

Para los casos L_1, L_∞, los estimadores obtenidos fueron únicos, ya que la matriz de diseño no contenía renglones iguales.

Las distribuciones utilizadas para generar los errores aleatorios fueron todas simétricas y cubren un amplio rango de kurtosis ** y fueron las siguientes:

* $\rho x_1 x_2$ coeficiente de correlación de Pearson de x_1 y x_2

** $E\left[(x-\mu)^4\right]/\sigma^4$ es usado como medida de kurtosis

(i) Distribución uniforme: (kurtosis = 1.8)

$$f_X(x) = \frac{1}{\beta - \alpha} \qquad \alpha < x < \beta$$

para que la media fuera cero y la desviación estándar 3

$$\alpha = -\beta \text{ y } \beta = \sqrt{3\sigma^2} = \sqrt{27}$$

(ii) Distribución normal: (kurtosis = 3)
con $\mu = 0$ y $\sigma^2 = 9$

(iii) Distribución normal contaminada: (kurtosis = 3.5, 4.0, 4.5, 5.0 y 5.5)

$$f_X(x) = \left(\frac{w}{\sigma 1}\right)\phi\left(\frac{x - \mu 1}{\sigma 1}\right) + \left(\frac{1 - w}{\sigma 2}\right)\phi\left(\frac{x - \mu 2}{\sigma^2}\right) \quad 0<w<1$$

donde $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-1/2\, t^2}$, $t \in \mathbb{R}$.

En este caso $w = 1/2$, dando igual peso a ambos componentes y $\mu_1 = \mu_2 = 0$ para asegurar simetría. Los diferentes coeficientes de kurtosis requerida se obtienen al elegir ${\sigma_1}^2$ y ${\sigma_2}^2$ apropiadamente, de tal manera que la varianza total sea igual a 9. Para más detalles de esta distribución ver Johnson and Kotz (1970).

(iv) Distribución Laplace o Doble exponencial (kurtosis=6)

$$f_X(x) = \frac{1}{2\beta} \exp\left(-\left|x - \alpha\right|/\beta\right) \qquad x \in \mathbb{R}$$

Para que x tuviese media cero y varianza 9, $\alpha = 0$ y $\beta = (\sigma^2/2)^{1/2} = (4.5)^{1/2}$

(v) Distribución Cauchy: (kurtosis indefinida)

$$f_X(x) = \frac{1}{\pi\beta\left[1+(x-\alpha)/\beta\right]^2} \qquad x \in \mathbb{R}$$

Esta distribución es simétrica alrededor de la mediana α y ésta se eligió como cero, la β fue determinada de manera que el cuantil .95 de la distribución Cauchy coincida con el cuantil

.95 de la normal (0, 9) usada en (ii).

Para los casos (i), (iv) y (v) se calcularon las funciones inversas de las distribuciones y los errores fueron generados usando números pseudoaleatorios de una distribución uniforme en el (0,1). Los errores para el caso normal y normal contaminada se obtuvieron usando números aleatorios de una uniforme y la transformación de Box - Muller (1958).

Los resultados del estudio de simulación fueron los siguientes:

Las medias de los 500 estimadores de los coeficientes de regresión se calcularon para p = 1.00, 1.25, 1.50, 1.75, 2.00 é ∞ y para cada una de las distribuciones. La tabla 3.2.1 muestra estos resultados.

Usando la aproximación normal de la distribución binomial y para una muestra de 500 estimadores el intervalo de confianza para el número de estimadores que caen por debajo del valor del verdadero parámetro es (228, 272).

Para todas las elecciones de p y para todas las distribuciones ninguna cayó fuera de esos límites.

La tabla 3.2.1 muestra que las medias de los estimadores muestrales son cercanos a los verdaderos valores de los parámetros. Este estudio como los hechos por Forsythe (1972) Kiountouzis (1973) no muestra evidencias de sesgo en los estimadores de norma Lp, $p \geq 1$ para distribuciones simétricas.

También se calcularon las varianzas muestrales de los tres coeficientes de regresión para los diferentes valores de p.

Para el caso p = 2 las varianzas muestrales fueron aproximadamente las mismas para todos los errores, excepto el caso Cauchy, pero eso es debido a que la matriz de covarianza es de la forma $\sigma^2 \ (X' X)^{-1}$, que es independiente de la distribución del error. Para el caso Cauchy la varianza σ^2 es indefinida y por tanto $E (e' e) \neq \sigma^2 I$.

TABLA 3.2.1

COMPARACION DE LAS MEDIAS DE LOS ESTIMADORES MUESTRALES
CON LOS VALORES DE LOS VERDADEROS PARAMETROS

DISTRIBUCION	Kurtosis	Verdaderos valores de $\beta_0, \beta_1, \beta_2$		p = 1.00	1.25	1.50	1.75	2.00	∞
Uniforme	1.8	0	10	9.98	9.99	10.00	10.00	10.00	9.95
		1	8	8.00	8.01	8.01	8.01	8.01	8.00
		2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-5.99
Normal	3.0	0	10	9.98	9.99	9.99	10.00	10.00	9.99
		1	8	7.99	8.00	8.00	8.00	8.00	8.00
		2	-6	-5.99	-5.99	-5.99	-5.99	-6.00	-6.00
Normal Contaminada	3.5	0	10	9.99	10.00	10.01	10.02	10.03	10.01
		1	8	8.00	8.00	8.00	8.00	8.00	8.00
		2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-6.00
Normal Contaminada	4.0	0	10	10.04	10.05	10.06	10.05	10.04	10.06
		1	8	8.00	8.00	8.00	8.00	8.00	8.00
		2	-6	-6.01	-6.00	-6.00	-6.01	-6.01	-5.99
Normal Contaminada	4.5	0	10	9.96	9.96	9.95	9.96	9.96	10.01
		1	8	8.01	8.01	8.01	8.01	8.01	7.99
		2	-6	-5.99	-6.00	-6.00	-6.00	-6.00	-6.01
Normal Contaminada	5.0	0	10	9.99	10.00	9.99	9.97	9.97	10.10
		1	8	8.00	8.00	8.00	8.00	8.00	8.00
		2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-6.01
Normal Contaminada	5.5	0	10	9.98	9.99	9.99	10.01	10.02	10.13
		1	8	8.00	8.00	8.00	8.00	7.99	7.97
		2	-6	-6.00	-6.00	-6.00	-6.00	-6.00	-6.00
Laplace	6.0	0	10	10.04	10.05	10.03	10.02	10.01	9.89
		1	8	8.00	8.00	8.00	8.00	8.00	8.02
		2	-6	-6.00	-6.01	-6.00	-6.00	-6.00	-5.99
Cauchy	-	0	10	10.01	10.03	10.07	10.25	10.39	9.74
		1	8	8.00	8.00	7.97	7.98	7.97	8.04
		2	-6	-6.00	-6.00	-5.98	-6.01	-6.02	-6.04

Para los casos $p \neq 2$ las varianzas cambiaron de acuerdo al tipo de distribución del error, conforme la kurtosis del error crece. Así, para el caso uniforme con kurtosis 1.8 las varianzas muestrales de los coeficientes fueron menores para los estimadores L_∞, mientras que para el caso de la Cauchy los estimadores L_1 fueron los de menor varianza. Entonces, mientras la kurtosis de la distribución crece la elección de $p < 2$ da estimadores con varianza muestral menor que la obtenida al usar MC.

El estudio considera una medida que comprende la varianza de todos los estimadores de los coeficientes, y la llama: varianza empírica generalizada, definiendo a ésta como el determinante de la matriz de covarianzas empíricas.

Como los estimadores Lp no mostraron evidencia de sesgo, en el estudio consideran razonable basar la elección de p en la varianza generalizada. La p que de menor varianza generalizada es el criterio seguido en el estudio.

Las varianzas generalizadas de los 500 estimadores muestrales son presentados en la tabla 3.2.2 , para todos los errores y todos los valores de p. La figura 3.2.1 es la gráfica de la kurtosis de la distribución del error contra el valor de p que da la mínima varianza generalizada, usando los datos de la tabla 3.2.2 .

La figura 3.2.1 muestra claramente que la "mejor" p depende de la kurtosis de la distribución del error.

Si la distribución tiene colas pesadas, es decir con mayor kurtosis que la normal, el valor óptimo de p será pequeño, si la distribución tiene colas cortas, el valor óptimo de p será grande. Esto puede verse en la tabla 3.2.2 para la distribución de Cauchy la mínima varianza generalizada se da cuando $p = 1$, y para el caso de la uniforme la mínima varianza generalizada se da cuando $p = \infty$.

TABLA 3.2.2. VARIANZA GENERALIZADA DE LOS ESTIMADORES DE REGRESION

($n = 25, \sigma^2 = 9$)

		p					
Distribución	Kurtosis	1.00	1.25	1.50	1.75	2.00	∞
Uniforme	1.8	24.55	8.85	4.52	2.77	1.87	0.63
Normal	3.0	8.35	3.93	2.64	2.21	2.09	39.00
Normal contaminada	3.5	6.08	3.24	2.37	2.10	2.20	77.79
Normal contaminada	4.0	2.86	1.56	1.35	1.48	1.82	150.90
Normal contaminada	4.5	2.14	1.31	1.17	1.33	1.73	193.00
Normal contaminada	5.0	1.19	0.93	1.10	1.53	2.30	321.00
Normal contaminada	5.5	0.31	0.30	0.53	1.07	2.12	591.00
Laplace	6.0	1.06	0.63	0.66	0.90	1.36	286.95
Cauchy	-	0.0018	0.0051	10.40	416.30	195000	$5.8 \times 10^{+8}$

Los valores de la tabla deben multiplicarse por 10^{-4}

El estudio propone la siguiente función como relación entre p y la kurtosis, basándose en la figura 3.2.1

$$p = \frac{9}{k^2} + 1 \qquad (3.2.1)$$

donde

k = kurtosis de la distribución del error.

La p que resulte de (3.2.1) es llamada p óptima

Para analizar el comportamiento de los estimadores cuando es utilizado un valor de p distinto a la p óptima, se consideró la eficiencia de la varianza generalizada, esto es, el cociente de la varianza generalizada usando la p óptima entre la obtenida usando cualquier otro valor de p. Esta eficiencia está dada en la tabla 3.2.3 . Por ejemplo, cuando la distribución del error es la contaminada normal con kurtosis 4.5 el estimador de MC tiene una eficiencia del 67% con respecto a estimador obtenido con la p óptima.

Al examinar la tabla 3.2.3 se llegan a las mismas conclusiones que en la tabla 3.2.2 . El uso de la ecuación (3.2.1) da resultados mucho mejores que el uso de MC.

Para examinar la validez de los resultados anteriores, el estudio de simulación fue repetido para los siguientes casos:

(i) $n = 25; \sigma^2 = 1$

(ii) $n = 25; \sigma^2 = 100$

(iii) $n = 10; \sigma^2 = 9$

(iv) $n = 50; \sigma^2 = 9$

Los resultados que se obtuvieron fueron esencialmente los mismos que los ya mencionados. Los autores del artículo concluyen que los resultados obtenidos son válidos para todos los tamaños de muestra y varianzas considerados.

La tabla 3.2.4 muestra la eficiencia de los estimadores de MC relativa a los estimadores de p óptima al variar el

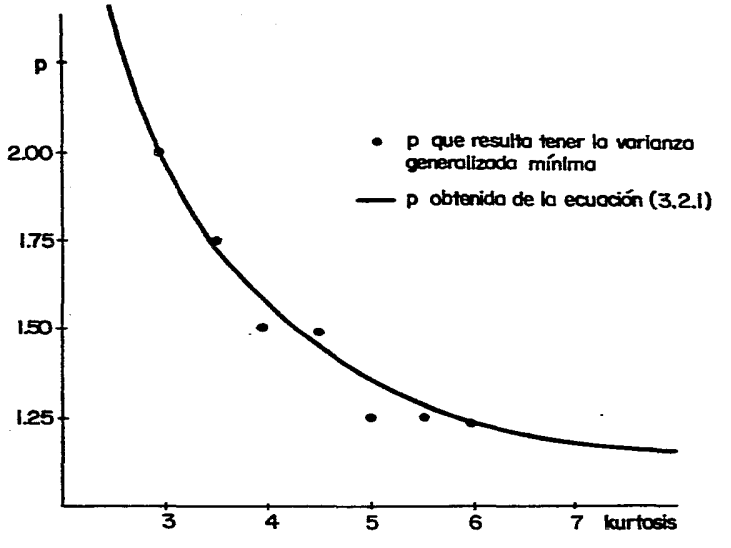

p
2.00
1.75
1.50
1.25
3
4
5
6
7
kurtosis
p que resulta tener la varianza generalizada mínima
p obtenida de la ecuación (3.2.1)

FIGURA 3.2.1

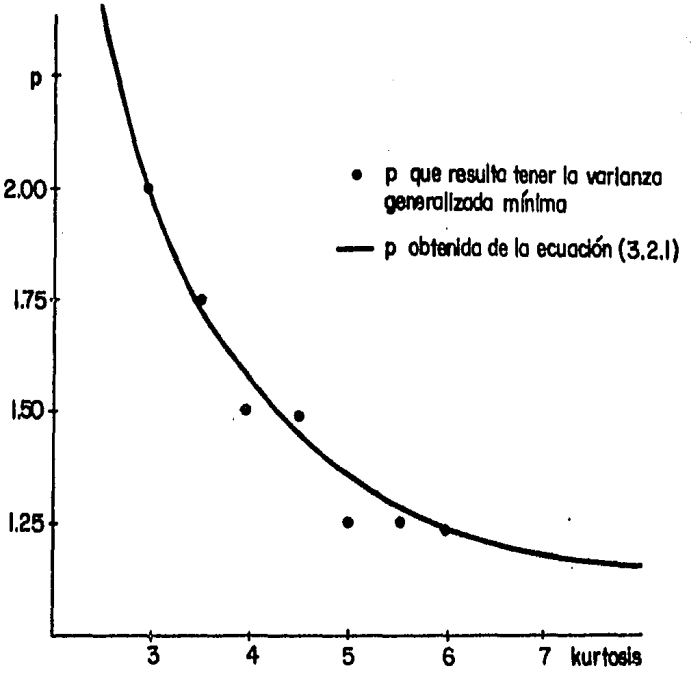

p
2.00
1.75
1.50
1.25
3
4
5
6
7
kurtosis
p que resulta tener la varianza generalizada mínima
p obtenida de la ecuación (3.2.1)

FIGURA 3.2.1

TABLA 3.2.3 EFICIENCIA(BASADA EN LA VARIANZA GENERALIZADA) DE LOS ESTIMADORES DE REGRESION ($n = 25, \sigma^2 = 9$)

Distribución	Kurt.	p 1.00	1.25	1.50	1.75	2.00	∞
Uniforme	1.8	0.026	0.071	0.140	0.227	0.337	1.000
Normal	3.0	0.250	0.532	0.792	0.946	1.000	0.054
Normal contaminada	3.5	0.345	0.648	0.886	1.000	0.995	0.027
Normal contaminada	4.0	0.472	0.865	1.000	0.912	0.742	0.009
Normal contaminada	4.5	0.547	0.893	1.000	0.880	0.676	0.006
Normal contaminada	5.0	0.782	1.000	0.845	0.608	0.404	0.003
Normal contaminada	5.5	0.968	1.000	0.566	0.280	0.142	0.001
Laplace	6.0	0.594	1.000	0.955	0.700	0.463	0.002
Cauchy		1.000	0.353	0.000	0.000	0.000	0.000

TABLA 3.2.4 EFICIENCIA DE LOS ESTIMADORES DE MC RELATIVA AL ESTIMADOR DE p OPTIMA. (BASADA EN LA VARIANZA GENERALIZADA)

Distribución	Kurtosis	n		
		10	25	50
Uniforme	1.8	1.00	0.34	0.09
Normal	3.0	1.00	1.00	1.00
Normal contaminada	3.5	1.00	1.00	0.93
Normal Contaminada	4.0	0.86	0.74	0.77
Normal contaminada	4.5	0.79	0.68	0.59
Normal contaminada	5.0	0.74	0.40	0.31
Normal contaminada	5.5	0.53	0.14	0.09
Laplace	6.0	0.59	0.46	0.30

tamaño de la muestra.

Claramente puede verse que si el tamaño de muestra crece, los estimadores de MC se hacen menos eficientes si el error es no-normal. De aquí que entre más grande sea el tamaño de la muestra más crítica es la elección de p.

En general la kurtosis de la distribución del error es desconocida y debe ser estimada de los datos muestrales.

Money et al. (1982) sugieren que la kurtosis sea reemplazada en la ecuación (3.2.1) por un estimador de la misma, basado en los residuales de MC para obtener el valor de p que será utilizado para determinar los estimadores Lp de los coeficientes de regresión.

Otro estudio similar se realizó para examinar si el uso del estimador de la kurtosis en (3.2.1) resulta con las mismas ventajas sobre MC. El modelo utilizado fue igual al usado con la kurtosis teórica. La matriz de diseño se eligió de manera que no contuviese renglones iguales, y el tamaño de muestra fue de nuevo 25.

La tabla 3.2.5 contiene la eficiencia de los estimadores de MC con respecto a los estimadores Lp, con p óptima.

Para las nueve distribuciones consideradas, los estimadores de MC son superiores en sólo tres casos. En cada uno de estos tres casos la eficiencia de los estimadores Lp, con p óptima tienen una eficiencia mayor al 90%. Para los otros seis casos la eficiencia de los estimadores de MC es mucho menor que la de los estimadores Lp de p óptima.

De aquí que para las distribuciones (del error) consideradas el método de estimación Lp es superior al de MC, al menos en términos de varianza generalizada.

TABLA 3.2.5 EFICIENCIA (BASADA EN LA VARIANZA EMPIRICA GENERALIZADA) DE LOS ESTIMADORES DE MC RELATIVA A LOS ESTIMADORES Lp DE LOS COEFICIENTES DE REGRESION ($n = 25$, $\sigma^2 = 9$)

Distribución	Kurtosis	Varianza Generalizada de los estimadores de MC	Varianza Generalizada de los estimadores Lp	Eficiencia de los estimadores de MC relativa a los estimadores Lp
Uniforme	1.8	4.61	2.50	0.54
Normal	3.0	4.38	4.73	1.08
Normal contaminada	3.5	4.32	4.82	1.12
Normal contaminada	4.0	3.74	3.86	1.03
Normal contaminada	4.5	4.15	3.62	0.87
Normal contaminada	5.0	4.50	2.88	0.64
Normal contaminada	5.5	4.45	1.91	0.43
Laplace	6.0	4.67	2.90	0.62
Cauchy	-	2.2×10^{-1}	7.8×10^{-7}	0

Los valores deben multiplicarse por 10^{-4}, excepto el renglón de la distribución Cauchy

3.3 ACERCA DE LA ELECCION DE LOS ESTIMADORES ROBUSTOS

Hasta ahora se han expuesto varios estimadores robustos y dos estudios que comparan a algunos de éstos. En seguida surge la pregunta ¿cuál de los estimadores robustos es conveniente utilizar?

Esta pregunta no es fácil de contestar.

Los siguientes puntos son algunos de los factores que intervienen en la respuesta:

- conocimiento acerca de la distribución de la variable respuesta;
- tamaño de la muestra;
- recursos con que se cuente para el cómputo;
- la situación, ya sea de observación o de experimentación;
- el número de variables explicativas.

Cuando se conoce la distribución de la variable respuesta, el siguiente cuadro puede ayudar a elegir el tipo de estimador robusto apropiado.

Distribución	Estimadores que obtuvieron mejores resultados según Heiler
Normal	MC, AN_1, TU_1, HA_1
N5C	AN_1, TU_1, HA_1, AN_2, VDW
N5N1	HA_1, TU_1, AN_2, HU_1, AN_1
N2OC	HA_1, TU_1, AN_2, HU_1
N2ON1, N1ON2, N2ON2	HA_2, AN_2, HU_2, GAS
C	MED, MSC, HA_2, TU_2, HU_2
NA	MC, AN_1, TU_1, HA_1
NCA	TU_2, HA_2, AN_2, HU1, HU2
J1	NOR, VDW, WIL, AN_1
T	HU_1, WIL
L	HU_1, WIL
DE	GAS, HU_1

Si además se conoce ó se puede estimar la kurtosis y si se desea emplear un estimador Lp, entonces el siguiente cuadro es de gran ayuda para elegir un estimador de este tipo

Distribución	Kurtosis	Estimador con menor varianza generalizada
Uniforme	1.8	L_∞
Normal	3.0	L_2 (MC)
Normal contaminada	3.5	$L_{1.75}$
Normal contaminada	4.0	$L_{1.5}$
Normal contaminada	4.5	$L_{1.5}$
Normal contaminada	5.0	$L_{1.25}$
Normal contaminada	5.5	$L_{1.25}$
Laplace	6.0	$L_{1.25}$
Cauchy	-	L_1

Es importante recalcar que en el estudio de Money et al. (1982) los estimadores no dan indicios de sesgo, pero estos autores - sólo consideran casos con distribuciones simétricas.

En el caso de elegir un estimador M, el tamaño de la muestra y los recursos con que se cuente para el cómputo son factores importantes para decidir que tipo de estimador inicial se utilizará. Si el tamaño de muestra es suficientemente grande y se cuenta con pocos recursos para el cómputo (software) se puede elegir el propuesto por Hinich y Talwar. Cuando se cuente con mayores recursos puede utilizarse el propuesto por Andrews (1974).

Si la situación se trata de un experimento diseñado con un cierto número común de observaciones para cada x (regresión simple), podría utilizarse la propuesta de Atkinson y Cox. (1977).

Cuando se desconozca la distribución de la variable respuesta será necesario aproximarse a ésta utilizando la distribución de la variable respuesta de estudios iguales o semejantes realizados con anterioridad para poder elegir así algún estimador robusto apropiado.

CAPITULO 4

EL USO DE LA ROBUSTEZ EN REGRESION

4. EL USO DE LA ROBUSTEZ EN REGRESION

Hogg (1979) dice:

> "Un buen estadístico aplicado está siempre alerta de la existencia de observaciones discrepantes, en unas ocasiones las descarta, en otras las investiga más a fondo, según sea lo apropiado en una situación dada. De cualquier modo, cuando se trata con datos complicados, es muy difícil detectar estas observaciones, es aquí donde los procedimientos robustos pueden ayudar"

Además Hogg, recomienda seguir estos pasos:

(a) Realizar el análisis usual (probablemente MC)

(b) Usar también un procedimiento robusto, al menos de una iteración. Idealmente, usar la Ψ de Huber con c=1.5, en la primera iteración y seguir con la Ψ de Andrews con c = 1.5 (posiblemente se puedan modificar estas constantes mediante una investigación preliminar).

(c) Si los resultados de (a) y (b) concuerdan, reportar entonces los resultados estadísticos usuales asociados con (a).

(d) Si los resultados de (a) y (b) difieren, revisar el grupo de datos. Especialmente aquellos puntos con pequeños pesos y preguntar desde ¿este error se debe a algún descuido al recolectar los datos?, hasta ¿esta observación discrepante trata de decir algo?

En el caso de llegar al inciso (d), surge la pregunta ¿A qué se deben estas diferencias?

Atkinson (1982) señala que si los resultados de MC y los de un procedimiento robusto no concuerdan las posibles causas pueden ser:

(i) errores gruesos, ya sea en la variable respuesta, ó en las explicativas.

(ii) un modelo lineal inadecuado.

(iii) que el análisis debió hacerse en otra escala (por ejemplo, después de haber sacado logaritmos).

(iv) el error tiene una distribución con colas más pesadas que la normal.

Atkinson utiliza herramientas de diagnóstico para identificar deficiencias en los datos e inadecuaciones del modelo.

Las herramientas de diagnóstico que utiliza son:

Para probar la concordancia entre y_i y la predicción $\hat{Y}_i = X_i \hat{\underset{\sim}{\beta}}_{(i)}$ utiliza los residuales "jack-knife".

$$e_i^* = e_i / \left(s_{(i)}^2 (1 - h_i) \right)^{1/2}$$

donde

$$e_i = y_i - \hat{y}_i$$

h_i es el elemento i-ésimo de la diagonal de la matriz $H = X (X' X)^{-1} X'$,

y la i entre paréntesis debe leerse como "con la observación i-ésima eliminada"

Como:

$$(n-k-1)\ s_{(i)}^2 = (n-k)\ s^2 - e_i^2 / (1 - h_i)$$

los residuales jack-knife pueden calcularse fácilmente.

Para examinar el efecto de la observación i-ésima en los estimadores de los parámetros utiliza la estadística modificada de Cook.

$$Ti = \left(\frac{n-k}{k} \cdot \frac{hi}{1-hi}\right) \cdot |ei^*|$$

Valores grandes de Ti indican una observación influyente, esto es, que tiene un gran efecto sobre los estimadores de los parámetros.

En algunos casos, observaciones aparentemente discordantes pueden ser unidas al resto de los datos mediante una transformación en la variable de respuesta. Para esto utiliza la familia de transformaciones Box y Cox (1964).

$$Z(\lambda) = Z = \frac{y^{\lambda} - 1}{\dot{y}^{\lambda-1}}$$, donde $\dot{y}$ es la media geométrica.

Ahora, si $W(\lambda) = W = \frac{\partial Z}{\partial \lambda}$, Atkinson (1973) para probar la hipótesis $\lambda = \lambda_0$ utiliza la estadística

$$T_D = \frac{Z^T (I - H) W}{S_Z (W^T (I - H) W)^{T/2}}$$

donde todas las cantidades son calculadas con $\lambda = \lambda_0$ y S_Z^2 es el estimador de la varianza de Z.

$$(n-k-1)S_Z^2 = Z^T (I - H) Z - \frac{(Z^T (I-H) W)^2}{W^T (I-H) W}$$

La estadística T_D se distribuye como t de student con n-k-1 grados de libertad. Dada un α, nivel de significancia, queda determinada la zona de rechazo.

Por ejemplo, si la hipótesis que se desea probar es $\lambda_0 = 1$, esto es que no se necesita transformación ó si se desea probar $\lambda_0 = 0$, la transformación logaritmo.

A continuación se presentan tres ejemplos:

El primer ejemplo trata el grupo de datos de salinidad estudiado por Ruppert y Carrol (1980), quienes encuentran una diferencia apreciable entre los estimadores (de β) de MC y los robustos.

La gráfica de papel normal de la estadística de Cook modificada Ti, muestra que hay algo "raro" con la observación 16. Al inspeccionar los datos se encuentra que una de sus variables explicativas está apartada del rango del resto de las observaciones.

Para el conjunto de datos original, la estadística T_D toma el valor de - 0.0844, cercana a cero, indicando que no hay necesidad de una transformación.

Pero si los datos son corregidos, la estadística T_D crece tomando el valor - 1.61. Otra observación tiene un valor alto de Ti, si esta observación es eliminada también T_D= -2.50, y la λ_0 obtenida por máxima verosimilitud es -0.15, para este caso la transformación logaritmo en la variable respuesta es congruente con los datos.

De aquí se concluye que hay dos observaciones "sospechosas" y hay alguna evidencia de transformar la respuesta.

Este ejemplo sirve para ilustrar el uso de la herramienta de diagnóstico para formularse preguntas acerca de la calidad de los datos y de la concordancia entre los datos y el modelo propuesto.

Otro ejemplo lo constituyen los datos de pérdida de material * debidos a Brownlee (1965). Daniel y Wood (1980) después de una amplia discusión concluyen que 4 observaciones son discordantes y deben ser eliminadas. A una conclusión similar llega Andrews (1974).

En cambio Atkinson (1981) y (1982) (a) usando la herramienta de diagnóstico, encuentra que hay una observación discrepante y que la estadística T_D da evidencia de la necesidad de una transformación, además cuando la observación discrepante es eliminada, la estadística T_D sigue dando evidencia de la necesidad de una transformación; la transformación indicada es el logaritmo.

* estos datos son utilizados en la sección 2.1 (Andrews).

La conclusión de ambos artículos es que todas las observaciones pueden ser modeladas adecuadamente en un modelo de segundo orden en x_1 y x_2, con variable de respuesta transformada mediante logaritmo.

Este ejemplo sugiere que las supuestas observaciones discrepantes del análisis robusto son muestra de un modelo inadecuado.

En el tercer ejemplo, basados en los datos de envenenamiento, Box y Cox (1964) encuentran que $T_D = -13.54$, que implica que una transformación debe hacerse. La $\hat{\lambda}$ de máxima verosimilitud es -0.75, y la transformación inverso ($\hat{\lambda} = -1$) está dentro del intervalo del 95% de confianza. En cambio Andrews (1971) y Carroll (1980) consideran el efecto de cambiar la variable respuesta para una de las observaciones. Este cambio hace decrecer T_D a -10.42, que de todas formas da evidencia de una transformación. Pero en este caso, el valor de λ es - 0.15, de tal manera que la transformación logaritmo debe ser considerada en vez de la transformación "inverso".

Este es un ejemplo en el que la presencia de una observación discordante hace que los procedimientos basados en un modelo lineal normal (MC) y los procedimientos robustos difieran.

En estos tres ejemplos Atkinson (1982) hace ver que las diferencias entre los análisis robustos y los de MC pueden elucidarse mediante el uso de técnicas de diagnóstico.

Los métodos robustos son alternativas al método de MC, cuando algunas suposiciones de éste son violadas. Pero éstos deben utilizarse siempre considerando que:

- Los estimadores robustos de regresión se ven afectados también por el mal condicionamiento de la matriz X (multicolinealidad).

- Si el modelo lineal es inadecuado, usar un método robusto no resolverá este problema. En algunas ocasiones será necesario transformar las variables primero y después aplicar el método robusto.

Además, Andrews (1979) señala:

" Las gráficas de residuales de MC contra otras variables no exhiben claramente dependencias no lineales con respecto a variables omitidas en el modelo. Los residuales robustos no mejorarán estas gráficas".

CAPITULO 5

INCONVENIENTES DE LOS ESTIMADORES ROBUSTOS

5.1 LA MATRIZ DE COVARIANZA DE $\hat{\beta}$

La matriz de covarianzas es un punto importante en la construcción de intervalos de confianza y en general para hacer inferencias. Huber (1973), mostró que $\hat{\beta}$ * tiene una distribución asintóticamente normal con matriz de covarianza.

$$\sigma^2 \frac{E\psi^2(\varepsilon/\sigma)}{[E(\psi'(\varepsilon/\sigma))]^2} (X'X)^{-1} \quad ;$$

Welsch (1975) y Gross (1977) reportan resultados de algunas sugerencias para estimar la matriz de covarianza. En ambos estudios resulta que los estimadores estudiados son muy sensitivos a ciertos diseños de la matriz X; por ejemplo si algunos renglones de la matriz X están alejados del resto de los datos.

Huber (1973), Tukey (1973) y Mallows (1975) han dado varias sugerencias basadas en la teoría asintótica. Hill (1979) muestra cuatro sugerencias fundamentales. Para simplificar la notación, sean

$$P = \sum_{i=1}^{n} \psi^2 \left(\frac{ei}{S}\right), \quad Q = \sum_{i=1}^{n} \psi' \left(\frac{ei}{S}\right)$$

donde

$\psi(t)$, la función de influencia.

$\psi'(t)$, es la derivada de $\psi(t)$

y $ei = yi - xi'\hat{\beta}$

* un estimador del tipo M

Sin tomar en cuenta el factor de corrección n / (n-p), las cuatro sugerencias son:

$$S_1 = n \frac{p}{Q^2} s^2 (X' X)^{-1}$$

$$S_2 = \frac{p}{Q} s^2 (X' \psi' (E/S) X)^{-1}$$

$$S_3 = \frac{p}{n} s^2 (X' \psi' (E/S) X)^{-1} (X'X) (X' \psi' (E/S) X)^{-1}$$

$$S_4 = s^2 (X' \psi (E/S) X)^{-1} (X' \psi^2 (E/S) x) (x' \psi' (E/S) x)^{-1}$$

donde $E = Y - X\beta$ y $\psi' (E/S)$ es una matriz diagonal.

S_1 y S_2 se siguen naturalmente de la teoría asintótica. S_3 y S_4 fueron sugeridas por Tukey.

Welsch (1975) estudia S_1 y S_2 para 3 diseños de matrices con n = 20, k = 3, 5 y 7. El concluye que S_1 da mejor aproximación de la matriz de correlación de $\hat{\beta}$ que S_2, pero ambas fallan cuando la matriz de diseño contiene columnas que son muestras de distribuciones con colas pesadas.

Hill y Holland (1977) (ver la sección 2.2) concluyen que S_1, es muy sensible a observaciones discrepantes en la matriz de diseño. Gross (1977) encuentra resultados semejantes.

Estos estudios dan evidencia de que S_1 y S_2 no son siempre los estimadores adecuados.

Hill (1979) dice no tener conocimiento de investigaciones empíricas sobre S_3 y S_4.

Este autor hace ver que S_1, S_2 y S_3 no son convenientes para el caso en que la regresión es de k puntos, esto es, que la matriz de diseño sólo contiene k renglones distintos (independientes) y el número de parámetros a estimar es k.

De S_4 hace ver que no lleva a la respuesta correcta ($s^2 (X'X)^{-1}$) si $\psi(t) = t$, (esto es, si el estimador M es el de MC)

Hill (1979) sugiere para las regresiones de k puntos el siguiente estimador

$$S_5 = \frac{1}{2}\,(S_5^* + S_5^{*\prime})$$

donde

$$S_5^* = \frac{\sum_{i=1}^{n} e_i^2}{n}\; s^2\,(X'\psi'\,(\tfrac{E}{s})\,X)^{-1}(X'\psi^2(\tfrac{E}{s})X)\;(X'E^2X)^{-1}\;(X'X)$$

$$(X'\psi'(\tfrac{E}{s})\,X)^{-1}$$

Hill (1979) dice "El principal objetivo de este artículo es mostrar que el estimar las matrices de covarianzas de los estimadores M es un problema difícil".

Este problema de la estimación de la matriz de covarianza es uno de los principales inconvenientes de los estimadores M.

5. 2 DIFICULTAD Y TIEMPO DEL COMPUTO DE LOS ESTIMADORES ROBUSTOS.

En general, la mayoría de los estimadores robustos requieren de programas más elaborados para su obtención que los de MC. Así como de mayor tiempo de cómputo.

En el caso de los estimadores M que se obtienen por métodos iterativos (ver sección 1.3) el estimador inicial juega un papel importantísimo en la obtención del estimador final.

Un estimador robusto inicial fácil de obtener es el propuesto por Hinich y Talwar (ver sección 2.3). Cuando se cuente con medios y recursos de cómputo mayores(software) se podrá utilizar como estimador inicial el propuesto por Andrews (ver la sección 2.1), ó el de Hill y Holland (ver la sección 2.2).

La obtención de los estimadores L se hace mediante la resolución de problemas de programación lineal (ver sección 1.1), se requerirá entonces tener disponible algún paquete de programación lineal para poder calcularlos.

En cuanto a los estimadores R son relativamente fáciles de evaluar, ya que se obtienen mediante la minimización de una función no-negativa, convexa y continua del vector como lo prueba Jaeckel (1972) (ver sección 1.2), en este caso se necesitará sólo algún programa que implemente alguno de los métodos iterativos sugeridos por el análisis numérico para hallar el mínimo de esa función.

De los estimadores de norma Lp se han elaborado muchos algoritmos, los casos L_1 y L_∞ pueden reducirse a la solución de problemas de programación lineal.

Para Lp con $1<p<2$ existen varios algoritmos, algu-

nos utilizan el método de Newton ó modificaciones de éste, en el apéndice 1.1 se habla de otro basado en mínimos cuadrados ponderados.

La mayor complejidad y mayor tiempo en el cómputo son parte de los inconvenientes que se tienen a cambio de la robustez.

CONCLUSIONES

En el modelo de regresión lineal se hacen varias suposiciones acerca del error. Como ya se ha dicho, el procedimiento de MC no es el adecuado cuando el error no se distribuye normalmente, en particular si tiene una distribución con las colas más pesadas que la normal, que usualmente generan observaciones discordantes. En estos casos, es conveniente utilizar en forma paralela al procedimiento de MC un procedimiento robusto adecuado. Ya que éstos permiten detectar a las observaciones discordantes más fácilmente y en general arrojan un mejor ajuste que el de MC y para el caso en el que el error se distribuye normalmente los estimadores obtenidos mediante estos procedimientos son sumamente eficientes.

Del estudio comparativo de Heiler puede concluirse que los estimadores M adecuadamente escogidos hacen frente casi a cualquier tipo de situación concebible. En particular los estimadores HU1, AN2 , HA y TU obtuvieron las mejores posiciones dentro de los estimadores considerados.

Para las situaciones donde hay mayor contaminación el estimador S_2 es superior a S_1, siempre y cuando no se espere asimetría en la distribución de la variable respuesta.

Los estimadores R, VDW y NOR en las situaciones con asimetría logran muy buenas posiciones.

En cambio, los estimadores L en general alcanzan resultados sólo mediocres y malos.

En cuanto a los estimadores Lp, puede decirse que éstos no dan muestra de sesgo, tanto en el estudio de Money et al (1982), como en los de Forsythe (1972) y de Kountouzis (1973)

Los estimadores de p óptima dada por la ecuación (3.2.1) resultan ser en la mayoría de los casos más eficientes que el

de MC, en términos de la varianza generalizada.

Para elegir el estimador robusto de β se deberán tomar en cuenta los siguientes puntos.

- distribución de la variable respuesta.
- tamaño de la muestra.
- recursos con que se cuente para el cómputo.
- la situación, ya sea de observación o de experimentación.
- número de variables explicativas.

Si se conoce la distribución de la variable respuesta, las tablas que aparecen en la sección 3.3 son de gran ayuda para elegir un estimador robusto de β.

Es importante señalar que los estimadores robustos de regresión, al igual que los estimadores de MC se ven afectados por el mal condicionamiento de la matriz de diseño. Además si el modelo que se está suponiendo no es el adecuado, es obvio que aunque se trate de un ajuste robusto, no se obtendrá un buen ajuste; en estos casos una transformación de variable o variables sería conveniente antes de aplicar un procedimiento robusto. Ver la sección 4.1.

El uso de los procedimientos robustos no se ha generalizado debido a varias razones:

- Su cómputo es más complicado que el de MC.
- Los criterios que se siguen para obtener a los estimadores robustos en la mayoría de las ocasiones son más difíciles de interpretar, en comparación con el criterio de minimizar una suma de residuales al cuadrado.
- Existe el problema que los estimadores robustos no son invariantes bajo escala. Esta es una dificultad muy seria, ya que a menudo a los grupos de datos se les multiplica por constantes.

- La obtención de resultados teóricos, distribuciones asintóticas, estimación de matrices de covarianza, pruebas asociadas, es bastante más difícil y complicada que para el caso de MC, como se hace ver en la sección 5.1.

En cuanto al cómputo de los procedimientos robustos actualmente, gracias a los avances en la computación, se han podido desarrollar una gran cantidad de rutinas para obtenerlos. Generalmente éstas son más complicadas y requieren de mayor tiempo que el procedimiento de MC.

APENDICE 1

A L G O R I T M O S

APENDICE 1.1

METODO ITERATIVO QUE MINIMIZA LAS DESVIACIONES "ABSOLUTAS" PROPUESTO POR SCHLOSSMACHER.

Karst (1958) es quizá el primero en presentar un procedimiento para estimar parámetros que minimicen las desviaciones absolutas (la norma L_1) pero la técnica iterativa que propone está limitada a modelos con dos parámetros.

La programación lineal puede manejar casos con k parámetros, pero puede resultar que la dimensión del problema sea muy grande, ó también que un paquete de programación lineal no sea accesible desde un paquete estadístico.

Schlossmacher (1973) propone un procedimiento para la minimización de las desviaciones absolutas que utiliza el método de mínimos cuadrados ponderados y que permite manejar k parámetros.

Considérese el modelo

$$yi = \beta_0 + \beta_1 \; x_1 i \; + \ldots + \beta_k \; x_{ik} \; + \; \varepsilon i.$$

El procedimiento de mínimos cuadrados ponderados minimiza el criterio

$$I = \sum_{i=1}^{n} wi \; ei^2, \quad ei \text{ son los residuales donde}$$

los pesos wi son dados por el investigador.

Considerando ahora la (r+1) ésima iteración de un proceso iterativo particular, el criterio puede escribirse como

$$I\ (r+1) = \sum_{i=1}^{n} \frac{1}{\left| e\ (r)_i \right|}\ e\ (r+1)_i^{\,2}$$

donde $e(r)_i$ es el residual i-ésimo en la r-ésima iteración si

$$\left| e(r)_i - e\ (r+1)_i \right| \cong 0 \quad , \quad i=1,\ldots,n \quad ,$$

el criterio se convierte en

$$I\ (r+1) \cong \sum_{i=1}^{n} \left| e\ (r+1)_i \right|$$

que aproxima las desviaciones absolutas deseadas.

El procedimiento propuesto es:

1.- Hacer el ajuste usando MC (wi=1, i=1, ...,n)

2.- Con los parámetros estimados calcular los residuales $e(r)_i$, i=1,...,n.

3.- Resolver el problema de MC ponderados con pesos,

$$wi = \left| \frac{1}{e(r)_i} \right|, \quad \text{si } e(r)_i \cong 0 \text{ entonces } wi=0$$

4.- Repetir los pasos 2 y 3, hasta que

$$\left| e(r+1)_i - e(r)_i \right| \cong 0$$

Durante este proceso iterativo algunos $e(r)_i$, serán cercanos a cero.

Fisher (1961) muestra que el ajuste de mínimas desviaciones absolutas pasa por una o más observaciones, de aquí que haya residuales cero. Al ocurrir esto, en el proceso iterativo los pesos correspondientes a residuales cercanos a cero, es decir

$$wi = \left| 1/e\ (r)_i \right|$$

tomarán valores muy grandes y pueden resultar problemas numéricos.

El procedimiento descrito elimina este problema, si un residual es muy pequeño en comparación al resto de los residuales, entonces a la observación correspondiente se le da peso cero.

Esto puede justificarse, ya que el efecto de un residual casi cero en la suma total de desviaciones absolutas es despreciable.

Pero si en alguna de las siguientes iteraciones el residual para esa observación crece, el proceso vuelve a considerarlo.

Los problemas de regresión por medio de la minimización de desviaciones absolutas pueden clasificarse en dos, aquellos con soluciones únicas y aquellos con soluciones no únicas.

Fisher (1961) muestra que para el caso en que la solución es única el ajuste de mínimas desviaciones absolutas pasa al menos por k puntos, donde k es el número de parámetros a estimar.

Si el ajuste de mínimas desviaciones absolutas pasa por menos de k puntos entonces existen varias soluciones.

Si el ajuste pasa por k puntos exactamente, la solución puede ser o no única.

Considérese el ejemplo que aparece en la tabla A.1.1 , éste fué tratado por Karst (1958) y encuentra que el mejor - ajuste de mínimas desviaciones absolutas es:

$$\hat{y} = 0.659\ x.$$

La técnica propuesta converge y produce este resultado en 8 iteraciones. (Ver tabla A.1.2).

Otro ejemplo con tres parámetros a estimar fue generado usando la ecuación:

$$y = 0.3 + 0.2\ x_1 + 0.1\ x_2 + \varepsilon i$$

usando varios valores de x_1 y x_2 y utilizando errores con distribucion doble exponencial (Laplace). Los valores para las x's fueron todas las posibles combinaciones de

$$x_1 = 0,1,2,3,4,5 \text{ y } x_2 = 0,1,2,3,4,5$$

En este caso el proceso iterativo propuesto por Schlossmacher converge en 7 iteraciones a los valores 0.3, 0.2 y 0.1 y fue comparado con la técnica de programación lineal de Fisher (1961). Se obtuvieron los mismos coeficientes, pero la técnica de Fisher necesitó mayor tiempo de computadora y mayor espacio en memoria.

Schlossmacher (1973) apunta que su técnica fuè 12.7 veces más rápida y requirió de un 90 por ciento menos de memoria que

TABLA A.1.1.

DATOS PARA EL EJEMPLO UNIDIMENSIONAL

x_i	y_i
-12.5	-8.4
- 8.5	-5.4
- 6.5	3.6
- 3.5	-2.4
- 2.5	-4.4
- 1.5	1.6
- 0.5	-0.4
2.5	-0.4
4.5	-2.4
8.5	3.6
8.5	5.6
11.5	9.6

TABLA A.1.2

SECUENCIA DE ESTIMADORES DE β
PARA EL EJEMPLO UNIDIMENSIONAL

j	$\hat{\beta}j$
1	.540
2	.599
3	.633
4	.651
5	.651
6	.654
7	.658
8	.659

la técnica de programación lineal.

Los dos ejemplos anteriores tuvieron solución única.

El siguiente ejemplo ilustra un caso donde la solución no es única.

En la tabla A.1.3 se presentan los datos de un ejemplo trabajado por Karst (1958), quien encuentra que el haz de rectas que pasan por el punto (30, 7.21) y con pendientes entre -0.1078 y 0.1250 es el que minimiza las desviaciones absolutas.

La tabla A.1.4 presenta la secuencia iterativa de las pendientes calculadas usando la tênica de Schlossmacher.

Convergiendo en los valores $\beta_0 = 3.95$ y $\beta_1 = 0.1088$ que definen una recta que pasa a través del punto (30, 7.21) y el valor de la pendiente no exceden los límites encontrados por Karst.

La recta óptima pasa por un punto de los datos, indicando que la solución no es única.

Los rangos donde $\hat{\beta}_1$ puede variar pueden ser hallados al hacer girar la recta en sentido positivo sobre el punto - - (xp, yp) con residual cero, hasta encontrar una observación, la recta que pasa por estas dos observaciones nos determina - uno de los límites del rango y el otro es hallado en forma similar pero rotando en sentido negativo.

Un procedimiento equivalente y más adecuado para explicarlo en una computadora es examinar los siguientes residuales.

TABLA A.1.3

DATOS PARA EL EJEMPLO BIDIMENSIONAL

x_i	y_i
12	5.27
18	5.68
24	6.25
30	7.21
36	8.02
42	8.71
48	8.42

TABLA A.1.4

SECUENCIA DE COEFICIENTES GENERADA PARA EL EJEMPLO BIDIMENSIONAL

j	$\hat{\beta}_{0_j}$	$\hat{\beta}_{1_j}$	I_j
1	3.994	0.1029	1.780
2	3.982	0.1044	1.746
3	3.971	0.1058	1.716
4	3.961	0.1069	1.691
5	3.954	0.1078	1.671
6	3.948	0.1085	0.658
7	3.946	0.1088	1.651
8	3.945	0.1088	1.650

$$R_1 i = \hat{\beta}_1 - (yi - yp) / (x_1 - x_p)$$

Los puntos que tengan el residual $R_1 i$ positivo más pequeño y el residual $R_1 i$ negativo más grande son las observaciones que acotarán el rango de $\hat{\beta}_1$.

Es difícil dar una generalización de estos procedimientos para los casos donde la solución no es única en problemas con más de dos dimensiones.

Estos ejemplos muestran que la técnica de Schlossmacher converge al valor del criterio de las mínimas desviaciones absolutas. La convergencia absoluta del método no ha sido probada, pero la técnica ha convergido para todos los problemas intentados por su autor.

Una gran ventaja de esta técnica es que requiere sólo de un paquete de mínimos cuadrados ponderados.

APENDICE 1.2

ALGORITMO PARA ENCONTRAR LOS ESTIMADORES DE NORMA Lp

($1 \leq p < 2$)

El algoritmo fue dado por Sposito, V.A., Kennedy, W.J., y Gentle, J. E. (1977).

Este algoritmo determina los estimadores de los parámetros del modelo lineal simple.

$$y = \beta_0 + x\beta_1 \quad ,$$

como ya se dijo en la sección 2.5, mediante el criterio de minimizar la norma Lp

$$\sum_{i=1}^{n} \left| yi - \beta_0 - xi\beta_1 \right|^p \qquad i = 1,\ldots,n$$

Considerando los casos con $1 \leq p < 2$.

Siguiendo la idea del algoritmo de Schlossmacher (sección anterior) y extendiéndola al intervalo (1,2), se considera minimizar

$$I = \sum_{i=1}^{n} Wi \; ei^2$$

en donde

e_i son los residuales y

W_i son los pesos

En el proceso iterativo consideran

$$I\,(r+1) = \sum_{i-1}^{n} \frac{1}{|e(r)_i|^{2-p}} (e(r+1)_i)^2$$

si $\left| e(r) - e\,(r+1)_i \right| \cong 0$ para $i = 1,2,\ldots,n$

entonces $$I\,(r+1) \cong \sum_{i=1}^{n} \left| e\ \ (r+1)_i \right|^p$$

Como lo hace notar Schlossmacher en su algoritmo, los residuales con valor absoluto pequeño pueden causar problemas. Schlossmacher sugiere la eliminación temporal de las observaciones asociadas a estos residuales. Las observaciones se reincorporan en las subsiguientes iteraciones en el caso que los residuales crecieran.

La eliminación de observaciones cercanas a la recta de ajuste puede hacer que la subsiguiente recta se aleje de la recta de norma Lp, en algunos casos patológicos, esta rutina incluye un indicador para el caso en que la norma crezca.

La convergencia del algoritmo no ha sido probada, pero la rutina ha convergido para todos los problemas intentados por los autores. Más aún, la convergencia es bastante rápida haciendo a este procedimiento definitivamente conveniente.

Tiempo y precisión de este algoritmo

El tiempo depende en el tamaño de muestra y el número de iteraciones.

La eliminación de observaciones cercanas a la recta de ajuste puede causar que la siguiente recta ajusta se aleje de la recta Lp en algunos casos patológicos, esta rutina verifica si la norma crece.

En las corridas de prueba, la rutina terminó frecuentemente por un incremento en la norma. Pero resultados empíricos sugieren fuertemente que el incremento en la norma ocurre sólo cuando el proceso ha convergido a una solución con tolerancia de errores de redondeo.

La idea de este algoritmo puede extenderse a un modelo de regresión múltiple sin ningún problema.

El programa está escrito en el lenguaje ISO FORTRAN.

APENDICE 1.3

ALGORITMO PARA HALLAR LOS ESTIMADORES DE NORMA L_∞

El algoritmo fue creado por Armstrong, R.D. y Kung, D.S. (1978).

El algoritmo determina los estimadores de norma L_∞ de los parámetros del modelo lineal

$$yi = x_{i1}\beta 1 + x_{i2}\beta 2 + \ldots + xi_k\beta k + \varepsilon_i,\ i=1,2,\ldots,n.$$

Mediante el siguiente criterio

$$\text{minimizar} \left\{ \underset{i=1,2,\ldots n}{\text{máximo}} \left| yi - \sum_{i=1}^{k} x_{ij}\ \beta\ j \right| \right\} \qquad (3.3.1)$$

Con anterioridad se dijo que (3.3.1) es equivalente al problema de programación lineal siguiente:

$$\text{Minimizar } D \qquad (D=\text{máx } |ei| \ )$$

sujeto a

$$yi - D \leqslant \sum_{j=1}^{k} x_{ij}\ \beta_j \leqslant yi + D \qquad i= 1,2,\ldots,n$$

β_1 sin restricción de signo.

Es un algoritmo SIMPLEX revisado que mantiene una base de tamaño kxk en vez de (k+1) x (k+1). Emplea descomposición LU para obtener las soluciones del sistema lineal.

Debido a la estructura especial del problema, el número total de iteraciones requeridas por el algoritmo símplex puede reducirse considerablemente.

El programa está escrito en lenguaje ISO FORTRAN.

APENDICE 2

FUNCIONES DE DISTRIBUCION

APENDICE 2.1

DISTRIBUCIONES SIMETRICAS

La Distribución Normal Apéndice 2.1.1

La función de densidad de una variable aleatoria X que se distribuye como normal es

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (\sigma>0)$$

donde μ es la media
y σ es la desviación estándar
Su función característica es

$$E(e^{i\,tx}) = e^{it\mu - \frac{1}{2}t^2\sigma^2}$$

Los primeros que trabajaron con la distribución normal, consideraban a ésta sólo como una aproximación conveniente a la distribución binomial.

Más adelante fue ampliamente aceptada como base para trabajos estadísticos prácticos, especialmente en astronomía.

La distribución normal tiene una posición única en la teoría de probabilidad, y ésta puede ser utilizada como una aproximación a otras distribuciones.

La mayoría de los argumentos teóricos para el uso de la distribución normal están basados en formas del teorema del límite central.

La Distribución logística Apéndice 2.1.2

Su función de densidad es

$$f_X(x) = e^{-x}(1+e^{-x})^{-2} = \frac{1}{4}\operatorname{sech}^2\frac{1}{2}x$$

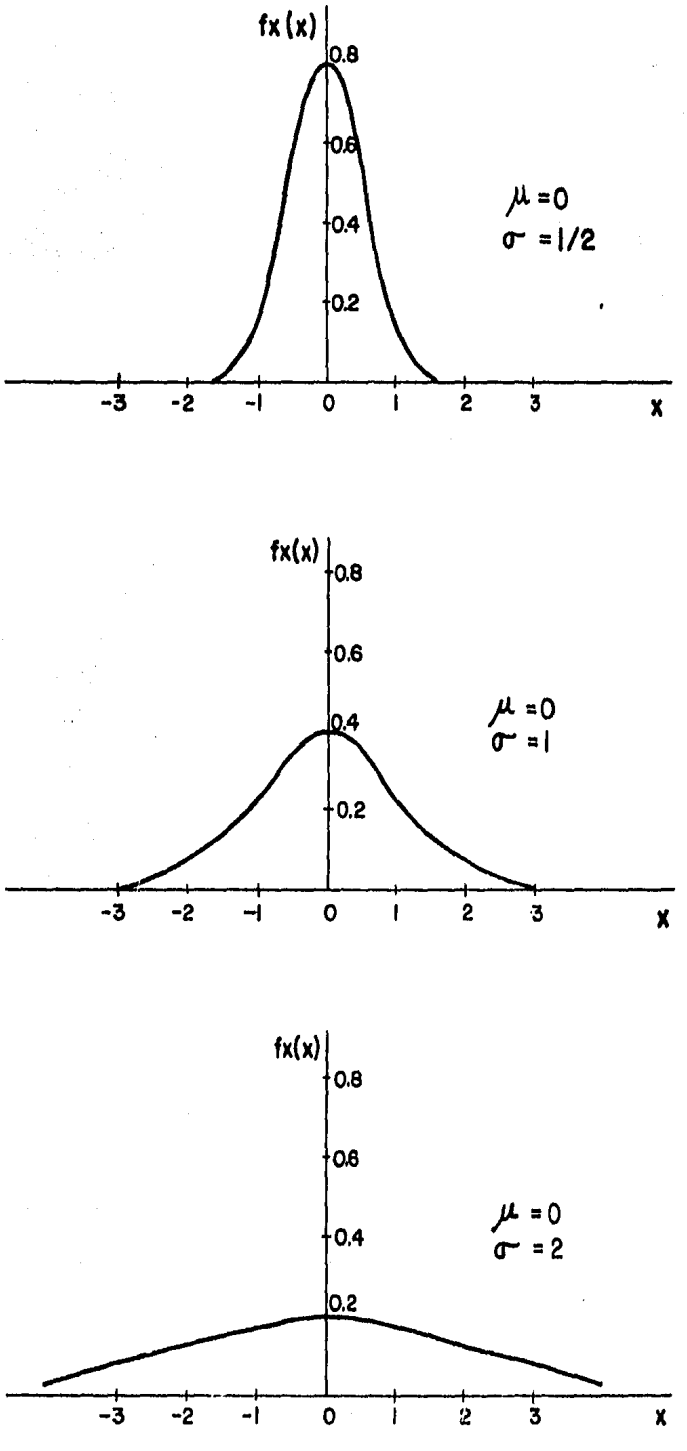


FIGURA A.2.1.1 FUNCIONES DE DENSIDAD NORMAL

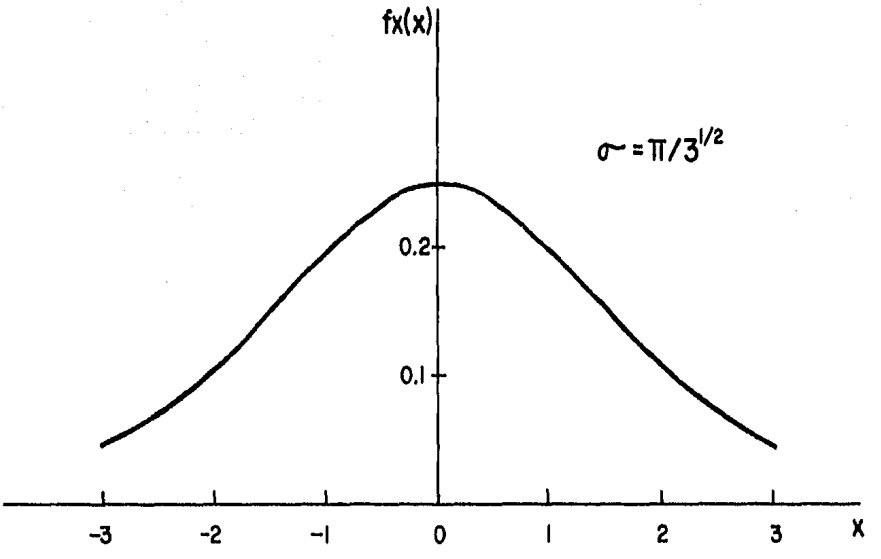


FIGURA A.2.1.2 FUNCION DE DE DENSIDAD LOGISTICA

La función característica es

$$E(e^{itx}) = \pi t \cos \operatorname{sech} \pi t$$

La función logística es utilizada para modelar curvas de crecimiento.

La Distribución Laplace o Doble Exponencial Apéndice 2.1.3

Su función de densidad es

$$f_x(x) = \frac{1}{2} \phi^{-1} \exp. \ (-|x-\theta| / \phi), \phi > 0.$$

La función característica es

$$E \ (e^{itx}) = (1+t^2)^{-1}$$

La distribución de la figura A.2.3 fue propuesta por Laplace en 1774, como una forma de distribución para la cual la función de verosimilitud es maximizada al hacer igual el parámetro de localización con la mediana de los valores observados independientes e idénticamente distribuídos de una variable aleatoria.

La Distribución t Apéndice 2.1.4

La función de densidad de una variable con distribución t con v grados de libertad es:

$$f_{tv} \ (t) = \frac{1}{\sqrt{v}\, B(\frac{1}{2}, \frac{1}{2} v)} \ (1+ \frac{t^2}{v}) - \frac{1}{2} \ (v+1)$$

$$\text{con } B \ (\tfrac{1}{2}, \tfrac{1}{2} v) = \frac{\Gamma(1/2)\Gamma(1/2 \ v)}{\Gamma(1/2 + 1/2 \ v)} = \frac{\sqrt{\pi}\Gamma(1/2 \ v)}{\Gamma(1/2 + 1/2 \ v)}$$

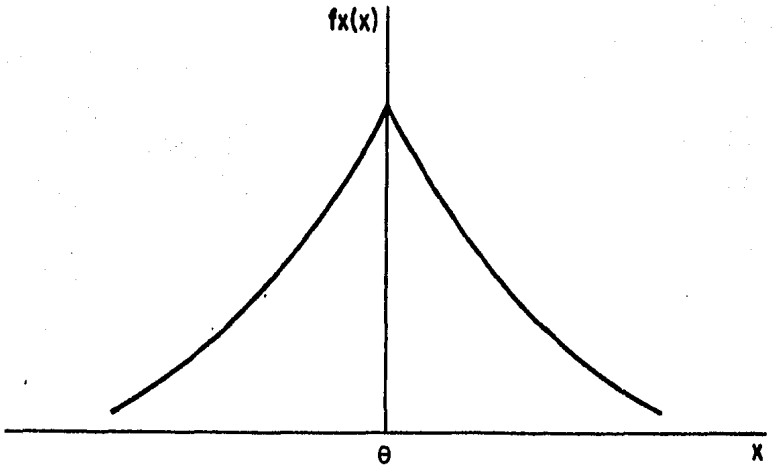


FIGURA A.2.1.3 FUNCION DE DENSIDAD LAPLACE O DOBLE EXPONENCIAL

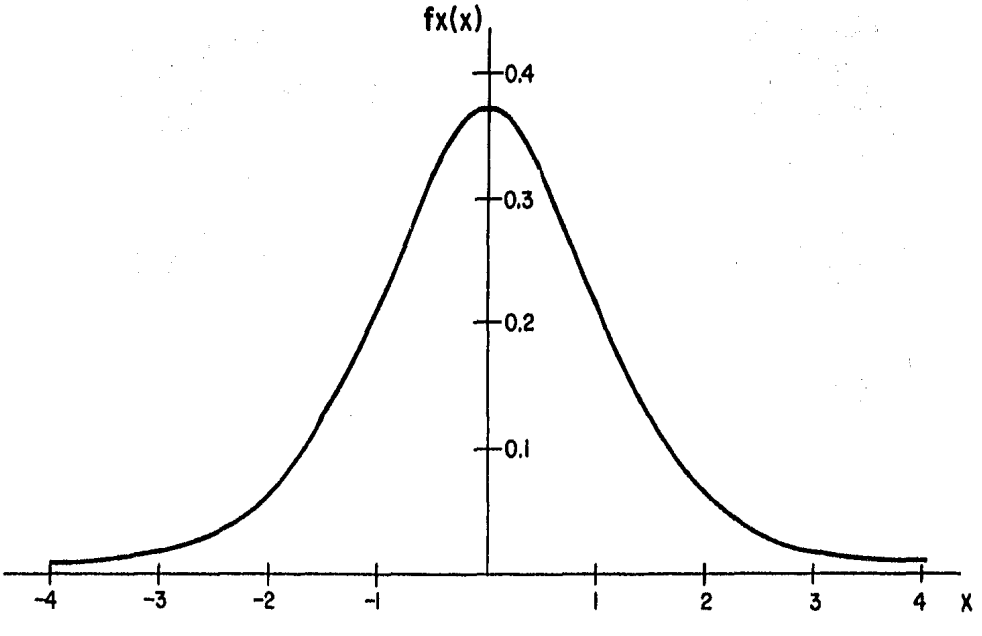


FIGURA A.2.1.4 FUNCION DE DENSIDAD t_4

Cuando se tienen $x_1, x_2, \ldots, x_n$ variables aleatorias independientes todas con la misma distribución normal $N(\mu, \sigma^2)$ entonces $\sqrt{n}(\bar{x} - \mu)/\sigma$ se distribuye como $N(0,1)$. Esta estadística se utiliza mucho en el cálculo de intervalos de confianza y estadísticas de prueba para

Pero para el caso en que σ es desconocida, se adopta

$$n(\bar{x} - \mu)/S \text{ con } S^2 \quad \frac{\sum_{i=1}^{n} (xi - \bar{x})^2}{n - 1}$$

que tiene una distribución t, con n - 1 grados de libertad.

La distribución Cauchy Apéndice 2.1.5

Su función de densidad es:

$$f_x(x) = (\pi\lambda)^{-1} \quad (1 + (x - \theta)/\lambda^2)^{-1} \quad (\lambda > 0)$$

La función característica es

$$E(e^{itx}) = \exp(it\theta - |t|\lambda)$$

La distribución Cauchy nace a raíz de querer escribir la distribución de un punto P de intersección, de una línea recta con otra línea recta variable, orientada aleatoriamente en dos dimensiones a través de un punto A. La distancia OP desde el punto de intersección P al punto O de intersección de la perpendicular trazada desde A a la línea recta fija, tiene una distribución Cauchy. Ver la figura A 2.1.5 (a).

Basados en este modelo, la distribución Cauchy puede ser usada para describir la distribución de los puntos de impacto de partículas de una fuente con una línea recta fija.

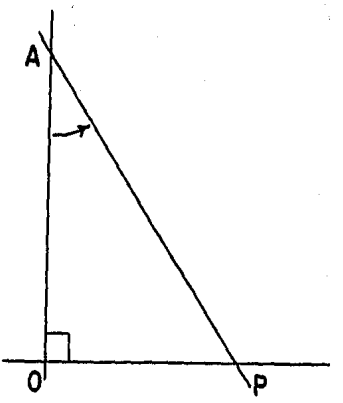


FIGURA A.2.1.5 (a)

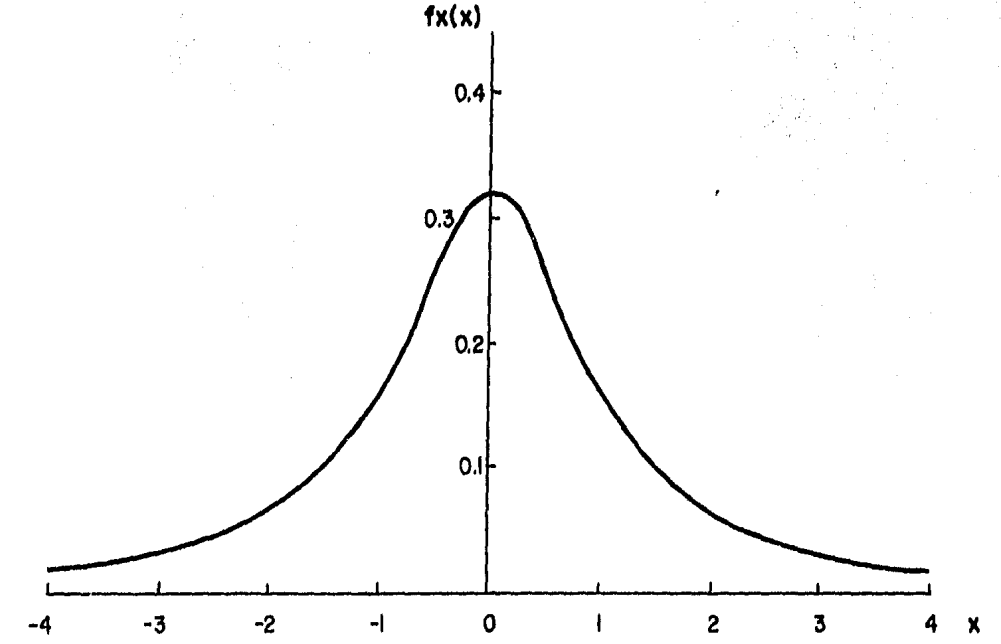


FIGURA A.2.1.5(b) FUNCION DE DENSIDAD CAUCHY

La distribución Uniforme Apéndice 2.1.6

La función de densidad es:

$$f_X(x) = (2h)^{-1} \qquad (a-h \leq x \leq a+h, \quad h>0)$$

y su función característica es:

$$E(e^{itx}) = \frac{e^{i(a+ht)} - e^{i(a-ht)}}{ith}$$

La distribución uniforme, con $a = 0$ y $h = \frac{1}{2} \times 10^{-k}$ es usada frecuentemente para representar la distribución de los errores de redondeo en valores tabulados con k cifras -- decimales.

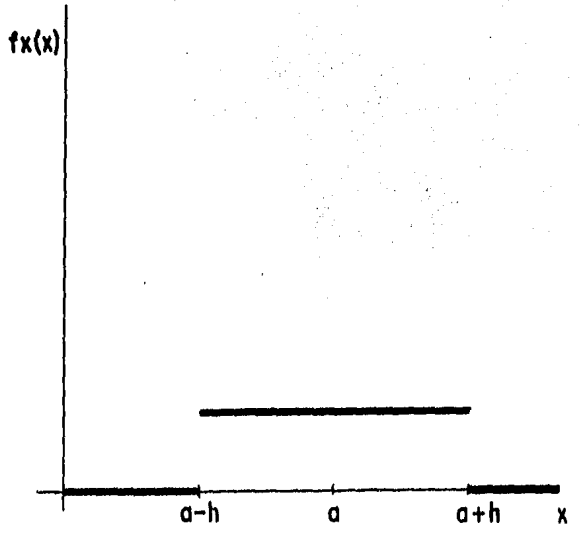


FIGURA A.2.1.6 DISTRIBUCION UNIFORME

APENDICE 2.2

DISTRIBUCIONES ASIMETRICAS

Distribución Ji-Cuadrada (χ^2) Apêndice 2.2.1

La función de densidad es

$$f_X(x) = \frac{1}{\Gamma(r/2)\ 2^{r/2}} \quad x^{r/2-1}\ e^{-x/2} \qquad 0 < x < \infty$$

y la función característica es:

$$E(e^{itx}) = (1 - 2it)^{-r/2}$$

Si $U_1, U_2, \ldots, U.$ son variables aleatorias independientes normales, N (0,1), entonces

$$\sum_{j=1}^{r} U_j^2 \quad \text{tiene una distribución } X^2_r$$

Si $x_1, x_2, \ldots, x_n$ son variables aleatorias independientes normales, y la desviación estándar común es σ, entonces

$$\sum_{j=1}^{n} (x_j - \bar{x})^2 \quad \text{es } \sigma^2 \text{ veces una variable con distribución } \chi^2 \text{ con } n - 1 \text{ grados de libertad.}$$

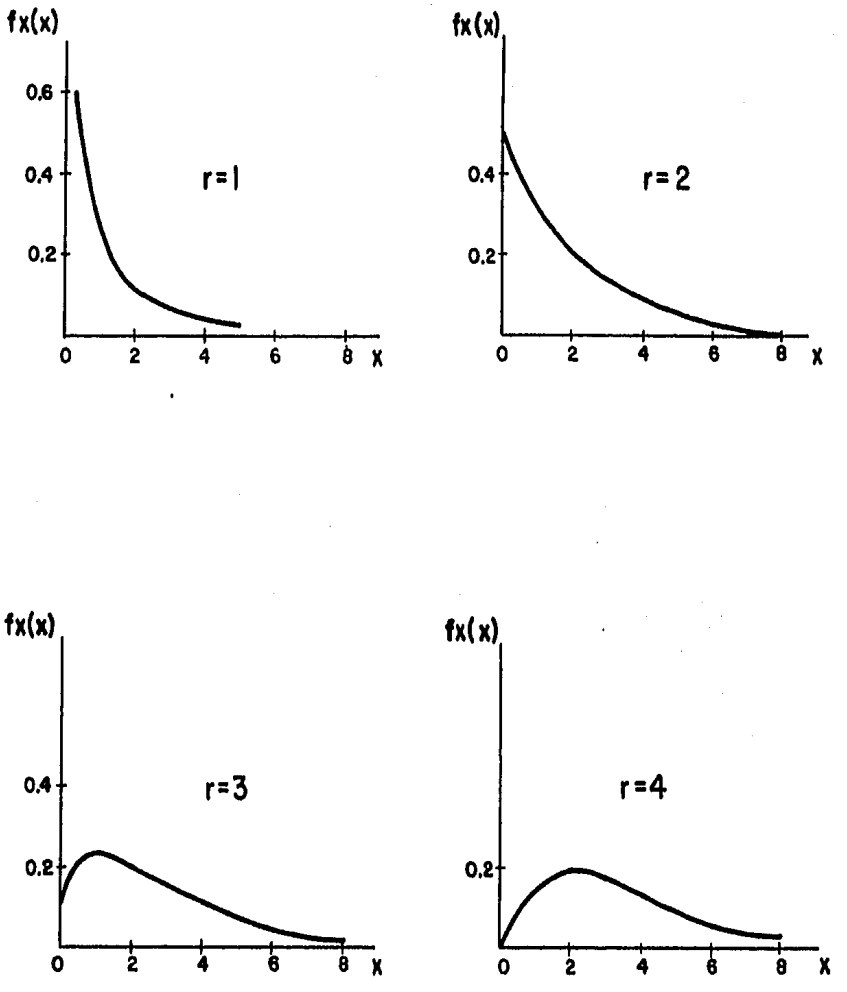


FIGURA A.2.2.1 FUNCIONES DE DENSIDAD JI CUADRADA

Distribución lognormal Apéndice 2.2.2

La función de densidad es:

$$f_X(x) = ((x-\theta)\sqrt{2\pi}\sigma)^{-1} \exp\left[-\frac{1}{2}\left[\log(x-\theta) - \xi\right]^2 / \sigma^2\right]$$

con $x > \theta$

Si $x_1, x_2, \ldots, x_n$ son variables aleatorias positivas independientes

y

$$Tn = \prod_{j=i}^{n} X_j$$

entonces

$$\log Tn = \sum_{j=1}^{n} \log X_j$$

y además las variables aleatorias, $\log X_j$ son tales que si el resultado del límite central se aplica, la distribución estandarizada de $\log Tn$ tiene una distribución que tiende a ser una normal N (0,1), conforme n tiende a infinito.

Entonces la distribución límite de Tn sería lognormal.

La distribución lognormal es aplicable a la distribución del tamaño de partículas en agregados naturales.

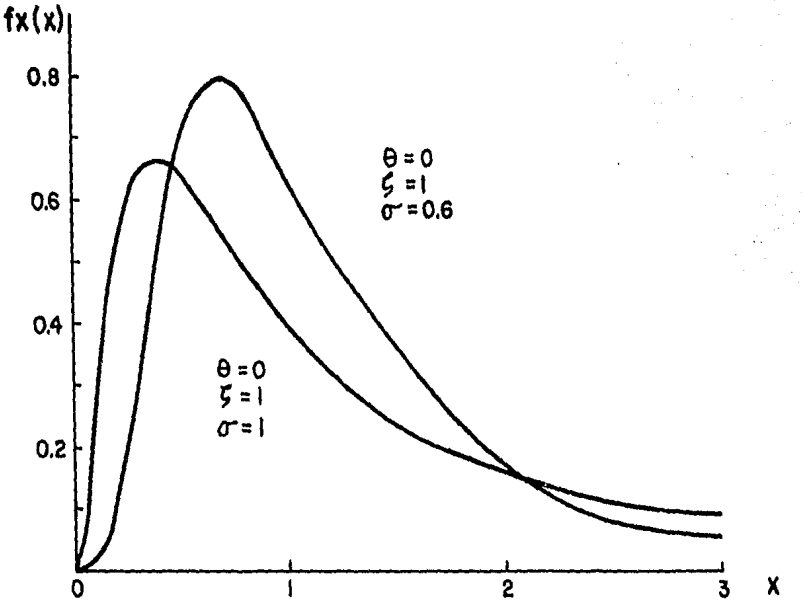


FIGURA A.2.2.2 FUNCION DE DENSIDAD LOGNORMAL

APENDICE 3

EL METODO DE MONTECARLO

APENDICE 3

EL METODO DE MONTECARLO

Durante la segunda guerra mundial se planteó a los científicos del Laboratorio Científico de Los Alamos un problema relativo al comportamiento de los neutrones.

¿Cuál era el poder de penetración de los neutrones en materiales diversos? La realización de un estudio experimental hubiera resultado cara, larga y arriesgada, y por otro lado, el problema parecía fuera del alcance de los cálculos teóricos.

Los físicos conocían la distancia media que podía recorrer un neutrón de una velocidad dada en un material sin colisionar con un núcleo atómico, la probabilidad de que un neutrón fuese rechazado, en vez de ser absorbido por el núcleo, la energía perdida en cada colisión, etc. Sin embargo resultaba imposible sintetizar todo esto en una fórmula manejable que fuese capaz de predecir el resultado de la sucesión de estos acontecimientos.

Los matemáticos John Von Neumann y Stanislaw Ulam resolvieron el problema mediante un procedimiento sencillo.

La solución que sugirieron se reducía a someter el problema a las veleidades de una rueda de ruleta. Las probabilidades de los distintos sucesos iban integrándose paso a paso en un todo que proporcionaba una respuesta aproximada, pero calculable del problema.

La técnica matemática aplicada por Von Neumann y Ulam se conocía desde hacía muchos años; al ser resucitada para el trabajo de Los Alamos, Von Neumann le dió el nombre de "Montecarlo.

Actualmente este método es empleado en diversos campos, principalmente en investigación de operaciones.

Supóngase que se desea conocer el porcentaje de los neutrones de un cierto haz que atraviesa un depósito de agua de un espesor determinado sin ser absorbidos ni perder gran parte de su velocidad. Y además se conoce la distancia media recorrida por un neutrón antes de chocar con un núcleo, la probabilidad relativa de que éste choque sea con un núcleo de oxígeno o de hidrógeno, las probabilidades de que el neutrón sea absorbido o rechazado por el núcleo, así como otra información pertinente.

Se elige un neutrón y se sigue su historia. Si se considerô un neutrón lento y que su primer incidente fue un choque con un átomo de hidrógeno. Sabiendo (empíricamente) que las probabilidades de que el neutrón sea rechazado están en la proporción de 100 a 1. Para decidir si será rechazado o no, en este caso concreto, imagínese el giro de una rueda de ruleta con 100 compartimentos iguales marcados con "rechazado" y uno con "absorbido" Si la ruleta para en "absorbido" la historia del neutrón concluye. Si se para en "rechazado", se hará girar otra rueda convenientemente construída para decidir cuál es la nueva direc-

ción tomada por el neutrón y la pérdida de energía que ha sufrido. De nuevo se hará girar otra rueda para decidir la longitud del camino recorrido hasta la próxima colisión y si ésta será con un núcleo de hidrógeno o de oxígeno. De esta forma se sigue a un neutrón hasta que es absorbido ó pierde tanta energía que se deja de tener interés en él, o se escapa -- del depósito de agua. Al acumular un número suficientemente grande de estas "historias", se obtendrá una descripción más o menos exacta del porcentaje de neutrones que escapan del depósito. El grado de precisión dependerá del número de casos considerados.

Naturalmente, en la práctica no se emplean ruedas de ruleta, sino números aleatorios.

El método de Montecarlo es empleado, en general, para la resolución de problemas que dependen de modo importante de factores probabilísticos, problemas en los que la experimentación física no resulta factible y en los que es imposible llegar a una fórmula exacta. Algunas propiedades de ciertos estimadores estadísticos no pueden ser determinadas usando análisis matemático únicamente. En este caso el método de muestreo de -- Montecarlo es extremadamente útil. Con la simulación de una - estructura económica (que incluya elementos estocásticos) cuyos parámetros son conocidos, se generan muestras (de cierto - tamaño) de "observaciones". Cada muestra es usada para esti-

mar los parâmetros por varios métodos. Para cada método, la distribuciôn de los estimadores es comparada con los valores verdaderos de los parámetros para determinar las propiedades de un estimador, dado el tamaño de la muestra.

B I B L I O G R A F I A .

- Andrews, D.F. (1971)
" A note on the select of data transformations"
Biometrika, 58, 249-254

- Andrews, D.F., (1974).
" A robust method for multiple linear regression".
Technometrics 16, 523-531.

- Andrews, D.F., (1979)
"The robustness of residual displays"
En "Robustness in statistics". Ed. Launer, R.L. y Wilkinson, G.N.
Academic Press. Nueva York.

- Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Roger, W.H. y Tukey (1972)
" Robust estimation of location: Survey and advances"
Princeton Univ. Press.

- Armstrong, R.D., y Kung, D.S., (1979)
AS 135 " Min-Max estimates for a linear multiple regression problem"
Applied Statist. 28 (1), 93-100.

- Atkinson, A.C., (1973).
"Testing transformations to normality"
J.R. Statist. Soc. B 35, 473 - 479.

- Atkinson, A.C. (1981)
"Two graphical displays for outlying and influential observations in regression"
Biometrika 68, 13 - 20.

- Atkinson, A.C. (1982).
"Robust and diagnostic regression analysis"
Comm. statist. theory and methods. 11 (22), 2559 -2571

- Atkinson, A.C., (1982) (a).
"Regression diagnostics, transformations and constructed variables"
J.R. Statist. Soc. B. 44, 1-36

- Atkinson, A.C. y Cox, D.R., (1977)
"Robust regression via discriminant analysis"
Biometrika 64 (1), 15 - 19

- Bickel, P. J., (1973).
"On some analogues to linear combinations of order statistics in the linear model"
Ann. Statist. 1, 597-616

- Bickel, P. J., (1976)
"Another look at robustness: A review of reviews and some developments".
Scandinavian I of Statisticas. 3,145-168

- Box, G.E.P. y Cox, D.R., (1964)
"An analysis of transformations"
J.R. Statist. Soc. B 26, 211-252

- Box, G.E.P. y Muller, M.E. (1958)
"A note on the generation of random normal deviates"
Ann. Math. Statist. 29, 610-611

- Brownlee, K.A., (1965)
"Statistical theory and methodology in science and engineering"
Wiley, Nueva York.

- Carnap. R., Morgenstern, O., Wiener, N., (1974)
"Matemáticas en las ciencias del comportamiento"
Alianza Editorial. Madrid. 78-86

- Carroll, R.J., (1980)
"A robust method for testing transformations to achieve" aproximate normality"
J.R. Statist. Soc. B. 42, 71-78

- Daniel y Wood., (1980)
"Fitting equations to data"
Wiley, Nueva York.

- Fama, E.F. y Roll, R. (1971)
"Parametric estimates for symetric stable distribution"
J. Am. Statist. Assoc. 66, 331-338

- Fisher., (1961)
"A note on curve fitting with minimun deviations by linear programming"
J. Am. Statist. Assoc. 56, 359-362

- Forsythe, A.B., (1972)
"Robust estimation of straight line regression coefficients by minimizing pth power deviations"
Technometrics 14, 159 - 166

- Goodnight, J. H., (1979)
"A tutorial on the sweep operator"
Am. Statist. 33 (3), 149 - 158

- Gross, A.M., (1977).
"Confidence Intervals for bisquare regression estimates"
J. Am. Statist. Assoc. 72, 341-354.

- Haitovsky., (1973)
"Regression Estimation from grouped observations"
Griffin. Londres

- Hardy G. H., Littlewood, J.E. y Polya, G., (1952).
"Inequalities"
Cambridge Univ. Press. Inglaterra.

- Hastings, N.A. J y Peacock, J.B., (1975)
"Statistical distributions"
Butterworths, Londres.

- Heiler, S., (1981)
"Robust estimates in linear regression. A simulation approach"
En Computational Statistics. Ed. Buning, H. y Naeve, P.
Walter de Gruyter, Berlin.

- Hill, R.W. y Holland, P. W., (1977).
"Two robust alternatives to least - squares regression"
J. Am. Statist. Assoc. 72 (360), 828 - 833.

- Hill, R.W., (1979)
"On estimating the covariance matrix of robust regression M-estimates"
Comm. Statist. A 8, 1183 - 1196.

- Hinich, M. J. y Talwar, P. P., (1975).
"A simple method for robust regression"
J. Am. Statist. Assoc. 75 (349), 113 - 119.

- Hogg, R., (1979)
"An introduction to robust estimation"
En "Robustness in Statistics". Ed. Launer, R.L. y Wilkinson, G.N.
Academic Press Nueva York

- Holland, P. W. (1973)
"Monte Carlo for robust regression: The swindle unmasked"
NBER Working Paper No. 10 National Bureau of Economic Research
Cambridge, Mass.

- Huber, P. J., (1964).
"Robust estimation of a location parameter"
Ann. Math Statist. 35, 73-101

- Huber, P. J., (1973)
"Robust regression: Asymptotics, conjetures and Montecarlo"
Ann Statist. 1, 799 - 821

- Huber, P. J., (1981)
"Robust statistics"
J. Wiley & Sons. Nueva York.

- Jaeckel, L. A., (1972)
"Estimating regression coefficients by minimizing the dispersion of the residual"
Ann. Math Statist. 43, 1449-1468

- Johnson, H. L. y Kotz, S.K. (1970)
"Distributions in Statistics, Continuos Univariate distributions"
Vol. 1 y 2 Houghton Mifflin Co. Boston

- Karst, O. J., (1958)
"Linear curve fitting using least derivations"
J. Am. Statist. Assoc. 53, 118-132

- Kiountouzis, E.A. (1973)
"Optimal Lp approximation techniques and data analysis"
Extrait du Bull. de la Soc. Matematique de Grece,
Novelle serie. Tome 12. Fasc. 1, 191-206

- Koenker, R. y Bassett, G., (1978)
"Regression quantiles"
Econométrica, 46, 33-50

- Kruskal, W.H. y Tanur, J. M. (1978).
"International Encyclopedia of Statistics"
The Free Press. Macmillan Publishing Co.

- Launer, R.L. y Wilkinson, G.N., (1979)
"Robustness in Statistics"
Academic Press. Nueva York.

- Luenberger, D. G., (1973)
Introduction to linear and non linear programming.
Addison-Wesley. Menlo Park, California. 148 - 155.

- Mallows, C.L., (1975)
On some topics in robustness. Plática presentada en una reunión regional del IMS y ASA en Nueva York
Mayo 21 - 23

- Mardia, K.V., Kent, J.T., Bibby, J. M. (1979)
"Multivariate Analysis"
Academic Press. Nueva York.

- Money, A. H., Affleck - Graves, J.F., Hart, M.L. Barr, G.D.I. (1982)
"The linear regression model Lp norm estimation and the choice of p."
Comm. Statist. Simulation and Computation II (1) 89-109

- Montgomery D.F. y Peck, E.A., (1982)
"Introduction to linear regression analysis"
Wiley, Nueva York

- Ruppert, D. y Carroll R. J., (1980)
"Trimmed least squares estimation in the linear model"
J. Am Statist. Assoc. 75, 828 - 838.

- Schlossmacher, E. J., (1973)
" An iterative technique for absolute deviations curve fitting"
J. Am. Statist Assoc. 68 (344) 857 - 859

- Seber, G.A.F., (1977)
"Linear regression analysis"
Wiley 351-400

- Sposito, V.A., Kennedy, W.J. y Gentle, J.E.,(1977)
AS110 "Lp Norm fit of a Straight line"
Applied Statist. 26 (1) 114-118

- Talwar, P., (1974)
"Robust estimation of regression parameters"
Ph. D. dissertation, Carnegie - Mellon University, Pittsburgh Penn

- Tukey, J. W. (1962)
"The future of Data Analysis"
Annals of Mathematical Statistics. 32: 1-67

- Tukey, J. W., (1973)
"A way for ward for robust regression"
no publicado. Bell Laboratories Technical report.

- Wagner, H. M., (1959)
"Linear programming techniques for regression analysis"
J. Am. Statist. Assoc. 54, 206-212

- Weisberg, S., (1980)
"Applied Linear Regression"
Wiley, Nueva Yor 108 - 109

- Welsch, R.E., (1975)
"Confidence regions for robust regression estimates"
Statist. Computing section, Proceedings of the Amer. Statist. Assoc. Washington, D.C. p. 36 - 42