Revista Colombiana de Estadística
Nº 17-18, 1988

OBSERVACIONES INFLUENCIALES

Elkin Castaño Velez

Centro de Investigaciones Economicas
Universidad de Antioquia

Resumen. El análisis de regresión lineal múltiple por mínimos cuadrados ordinarios es quizás la técnica estadística más aplicada por muchas ciencias para modelar la relación entre varias variables. En las últimas décadas ha habido un gran desarrollo en el estudio de los factores que determinan el ajuste de la ecuación de regresión, como son las observaciones, las variables y las hipótesis del modelo. En este artículo presentaremos algunos de los procedimientos empleados para medir el efecto que tiene sobre los resultados del ajuste y sobre las conclusiones obtenidas, la eliminación (o la inclusión) de una o de un grupo de observaciones. También se discutira el papel que juegan algunas observaciones en la generación o en el enmascaramiento de relaciones de colinealidad en la matriz de datos del modelo de regresión lineal.

Introducción. Varios elementos determinan el ajuste de una ecuación de regresión: las variables, las observaciones y las hipótesis del modelo. En este artículo nos concentraremos fundamentalmente en el papel que juegan las observaciones, individualmente o en grupo, en la estimación y en las conclusiones obtenidas de la ecuación de regresión ajustada, y en algunas de las metodologías más comúnmente empleadas para describir su

comportamiento.

Con el fin de comprender mejor el papel realizado por los datos en el ajuste de la ecuación de regresión veamos algunos ejemplos sencillos que nos permitan diferenciar las diferentes formas como una observación puede actuar sobre el ajuste de la ecuación. Con el fin de simplificar la exposición consideremos el modelo lineal simple. (Los siguientes ejemplos se encuentran en Chatterjee y Hadi (1988)).

Ejemplo 1.

Consideremos los datos del gráfico 1. Suponga que a los datos marcados con un + queremos agregar, uno a la vez, los datos marcados con las letras A, B y C.

La inclusión del punto A en la regresión generaría un pequeño residual debido a que A está en la dirección de la recta que pasa por todos los demás puntos. Esto implica que la observación A no tiene una gran influencia sobre la ecuación ajustada, es decir no cambia los coeficientes estimados. Por tanto, a pesar de que A es un punto extremo en X (por esto es denominado punto de *alto poder*) y en Y, A no es *influencial* para la estimación de los coeficientes de la ecuación, aunque puede serlo para el error estándar de los coeficientes de regresión.

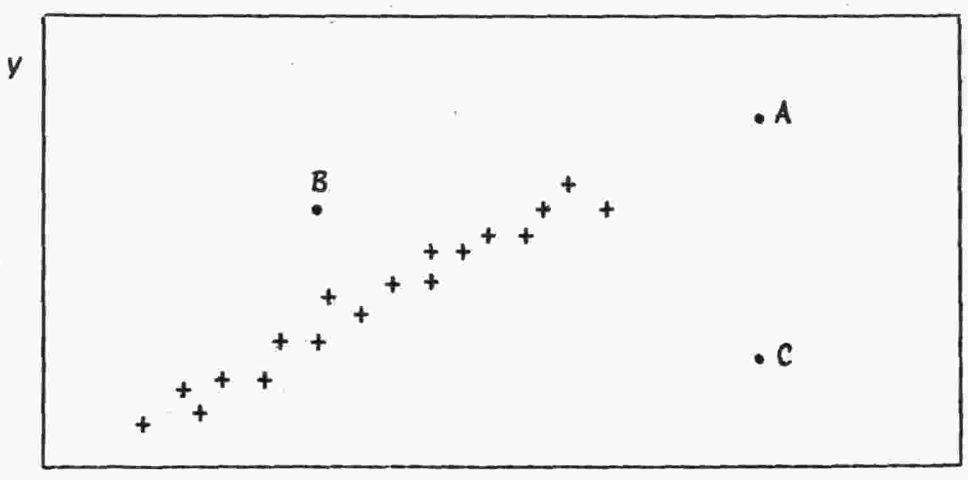
Por otro lado, si B es incluído generará un gran residual (por esto es llamado un punto *outlier*) y aunque puede no cambiar el estimador de la pendiente si alterará el intercepto de la recta. Su inclusión también puede cambiar la varianza estimada del error y, en consecuencia, las varianzas estimadas de los coeficientes. Por tanto una observación como B es un *outlier* (no es de *alto poder* puesto que no es un extremo de X) y

es un punto *influencial*.

Por último, si agregamos el punto C (el cual es un punto de alto poder puesto que es extremo de X) a los datos marcados con +, se generá un gran residual (C será entonces un 'outlier') y se alterarán sustancialmente las características de la regresión ajustada (C es un punto influencial). Por tanto C es un punto *outtlier, influencial y de alto poder*. Este ejemplo ilustra diferentes formas como las observaciones actúan sobre el ajuste de la ecuación: en forma de outliers, puntos de alto poder y puntos influenciales; se observa también que estos conceptos no son excluyentes y se revelan algunas relaciones entre ellos.

El siguiente es un ejemplo para ilustrar que los 'outliers' no necesariamente son observaciones influenciales y que las observaciones influenciales no necesariamente son 'outliers'.

Gráfico 1.



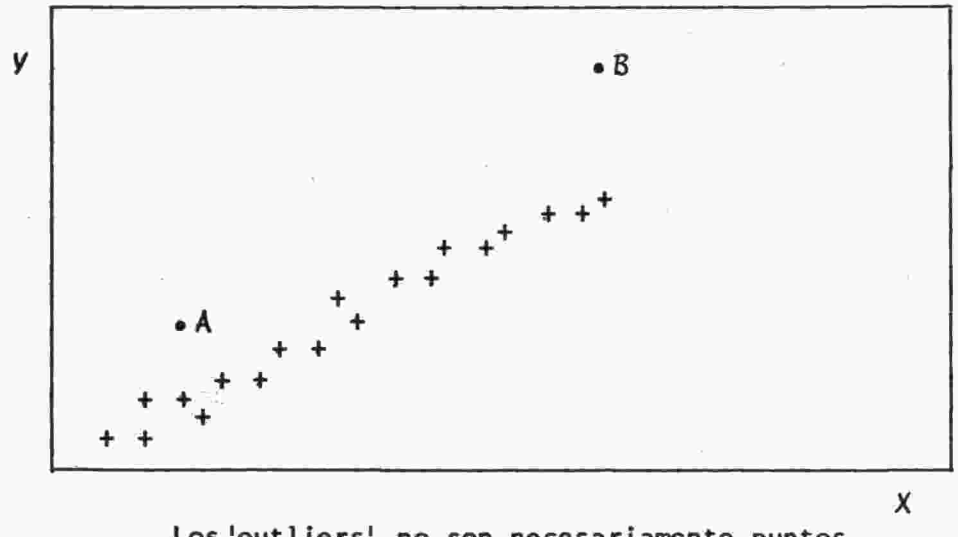
Distinción entre 'outliers', puntos de alto poder y observaciones influenciales.

Ejemplo 2.

Considere el gráfico 2. Si ajustamos una línea recta a los datos del gráfico, claramente A será un 'outlier' (dejará un residual grande). Sin embargo su exclusión dificilmente cambiará los resultados del ajuste. Este es el caso de un 'outlier' que tiene poca influencia sobre $\hat{\beta}$. Ahora bien, la observación B tendrá un residual pequeño pero sin embargo, cuando la excluímos, los coeficientes de regresión estimados varían sustancialmente. Este es el punto influencial que no es un 'outlier'.

Estos ejemplos muestran que el solo exámen de los residuales puede no servir para detectar observaciones aberrantes o inusuales y que los métodos gráficos basados en residuales solamente, pueden fracasar en reconocer dichos puntos. Por tanto se necesitan medidas estadísticas que permitan detectar el *poder* y la *influencia* de las observaciones.

Gráfico 2.



Los 'outliers' no son necesariamente puntos influenciales ni los puntos influenciales son necesariamente 'outliers' .

Existe una vasta literatura estadística sobre construcción de medidas para detectar observaciones influenciales y medir sus efectos sobre varios aspectos del análisis: Besley et al (1980), Cook y Weisberg (1982), Atkinson (1985), Chatterjee y Hadi (1986, 1988), etc. En este artículo solamente desarrollaremos algunas medidas basadas en la eliminación de una observación o de un grupo de observaciones.

El esquema de este artículo es el siguiente: en la sección 1 se recordarán algunos de los resultados básicos del ajuste del modelo lineal general por mínimos cuadrados; la sección 2 presenta algunas propiedades e interpretaciones de los elementos de la matriz proyección $P = X(X'X)^{-1}X'$. Además se presentan definiciones de 'outlier', punto de alto poder y punto influencial así como las relaciones existentes entre estos conceptos. En la sección 3 se discuten algunas medidas para detectar la influencia de una sóla observación y de un grupo de observaciones sobre la ecuación de regresión; una breve discusión de la influencia de múltiples observaciones se encuentra en la sección 4 y la sección 5 introduce el problema de la colinealidad y su relación con los puntos influenciales.

1. Supuestos y Resultados Basicos de la Estimación por Mínimos Cuadrados.

1.1. El Modelo.

Suponemos que la relación existente entre una variable Y y las variables $X_1, X_2, \ldots, X_k$, es de la forma

$$Y = X\beta + \varepsilon$$

donde

Y es el vector *respuesta* de $n \times 1$ o variable *dependiente*,

X es una matriz de $n \times k$ de *predictores*, o *regresores*, o *variables explicativas*,

β es un vector de $k \times 1$ de coeficientes desconocidos (parámetros) que van a ser estimados y,

ε es un vector de $n \times 1$ de perturbaciones o errores aleatorios.

1.2. Los Supuestos.

El procedimiento de estimación por Mínimos Cuadrados Ordinarios está basado en los siguientes supuestos:

(a) Hipótesis de Linealidad:

Este supuesto indica que la i-ésima respuesta observada puede ser escrita como una función lineal de la i-ésima fila x_i' de la matriz X, es decir

$$y_i = x_i'\beta + \varepsilon_i, \quad \text{para } i = 1,2,\ldots,n.$$

(b) Hipótesis Computacional:

Para que exista un único estimador de β es necesario que $(X'X)^{-1}$ exista o equivalentemente,

$$\text{rango } (X) = k$$

(c) Hipótesis Distribucional:

i) X está medida sin error,

ii) ε_i no depende de x_i', $i = 1,\ldots,n$.

iii) $\varepsilon \sim N_n(0,\sigma^2 I)$, es decir, la distribución conjunta de las n perturbaciones aleatorias es multinormal con vector de medias 0 y matriz de covarianzas $\sigma^2 I$.

(d) Hipótesis Implícita:

Todas las observaciones son igualmente confiables y deben jugar un igual papel en la determinación de los resultados

y las conclusiones.

1.3. Resultados Básicos de la Estimación por Mínimos Cuadrados.

El estimador de β por el procedimiento de los Mínimos Cuadrados (E.M.C.) se obtiene minimizando la expresión $(Y-X\beta)'(Y-X\beta)$ con respecto a β. Tomando la derivada con respecto al vector β obtenemos las ecuaciones normales o de estimación

$$(X'X)\beta = X'Y$$

Este sistema tiene solución única si y sólo si $(X'X)^{-1}$ existe y en este caso la solución es

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Si los supuestos de la sección 1.2 se cumplen la teoría de los E.M.C. proporciona los siguientes resultados:
(Graybill (1976), Draper y Smith (1981)).

(a) El vector β tiene las siguientes propiedades:

(i) $E(\hat{\beta}) = \beta$, es decir $\hat{\beta}$ es un estimador insesgado para β.

(ii) $\hat{\beta}$ es el mejor estimador lineal insesgado para β, y su matriz de covarianzas es $\text{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}$.

(iii) $\hat{\beta} \sim N_k(\beta,\sigma^2(X'X)^{-1})$.

(b) $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = PY$ es el vector de los valores ajustados de Y. Entonces

(i) $E(\hat{Y}) = X\beta$

(ii) $\text{Cov}(\hat{Y}) = \sigma^2 P$

(iii) $\hat{Y} \sim N_n(X\beta,\sigma^2 P)$, donde $P = X(X'X)^{-1}X'$.

(c) El vector de $n\times 1$ de residuales (ordinarios)

$$e = Y - \hat{Y} = Y - PY = =(I-P)Y$$

tiene las siguientes propiedades:

(i) $E(e) = 0$

(ii) $\text{Cov}(e) = \sigma^2(I-P)$

(iii) $e \sim N_n(0, \sigma^2(I-P))$

(iv) $e'e/\sigma^2 \sim \chi^2 \ (n-k)$

(v) $e = (I-P)\varepsilon$

(vi) Un estimador insesgado de σ^2 está dado por

$$\hat{\sigma}^2 = e'e/(n-k) = Y'(I-P)Y/(n-k)$$

Observación.

Las matrices P e $I-P$ son singulares y por tanto las distribuciones multinormales de $\hat{Y}$ y de e son singulares (para una definición veáse, por ejemplo, Mardia et al., (1978). En la próxima sección discutiremos las relaciones de los elementos de las matrices P e $I-P$ con los resultados de la estimación, y sus propiedades e interpretación.

2. Propiedades e Interpretación de los Elementos de la Matriz P.

En la sección 1.3 vimos que la matriz P determina muchos de los resultados de la estimación por M.C. Esta matriz recibe el nombre de matriz *predicción* debido a que al ser aplicada a Y produce los valores predichos; también recibe el nombre de matriz *hat* (sombrero, en inglés) puesto que al aplicarla a Y le coloca el sombrero; otro nombre dado a P es el de matriz *proyección*, puesto que proyecta ortogonalmente a y sobre el

espacio k-dimensional generado por las columna de X. Analoga - mente $I-P$ es llamada matriz *residual* puesto que al ser aplicada a Y produce los residuales (ordinarios).

2.1. Propiedades de los Elementos de P.

Denotemos por p_{ij} el ij-ésimo elemento de P, $i,j = 1,2,\ldots,n$; entonces $p_{ij} = x_i'(X'X)^{-1}x_j$ y las siguientes propiedades son ciertas

(a) $\Sigma\Sigma p_{ij}^2 = k = \Sigma p_{ii}$

(b) $0 \leq p_{ij} \leq 1$ para todo $i = 1,2,\ldots,n$.

(c) $-.5 \leq p_{ij} \leq .5$ si $i \neq j$.

(d) Si X contiene una columna constante, $p_{ii} \geq 1/n$, $i = 1,2,\ldots,n$.

(e) Si $p_{ii} = 1$ ó $p_{ii} = 0$ entonces $p_{ij} = 0$.

(f) $(1-p_{ii})(1-p_{jj})-p_{ij}^2 \geq 0$

(g) $p_{ii}p_{jj}-p_{ij}^2 \geq 0$

(h) $p_{ii}+e_i^2/e'e \leq 1$

(i) Si X matriz $n\times k$ es de rango k, entonces para k fijo p_{ii} es no decreciente en n, para $i = 1,\ldots,n$.

2.2. Relaciones Entre los Resultados del Ajuste y la Matriz P.

De la sección 1.3 vimos que $\text{Cov}(\hat{y}) = \sigma^2 P$ y que $\text{Cov}(e) = \sigma^2(I-P)$. De estos resultados obtenemos que los elementos de la diagonal de la matriz P determinan las varianzas de los valores ajustados y de los residuales:

$$\text{var}(\hat{y}_i) = \sigma^2 p_{ii} \quad i=1,2,\ldots,n, \quad \text{y}$$

$$\text{var}(e_i) = \sigma^2(1-p_{ii}) \quad i=1,2,\ldots,n.$$

Además, puesto que la $\text{cov}(e_i,e_j) = -\sigma^2 p_{ij}$, el coeficiente de correlación entre e_i y e_j está completamente determinado por los elementos de P,

$$\text{corr}(e_i,e_j) = -p_{ij}/[(1-p_{ii})(1-p_{jj})]^{1/2}.$$

De 1.3 $e = (I-P)\varepsilon$ lo que indica que la relación entre e y ε sólo depende de P.

Además, también de 1.3, $\hat{y} = PY$, de donde el i-ésimo valor ajustado puede escribirse como

$$\hat{y}_i = \sum_{i=1}^{n} p_{ij}y_j = p_{ii}y_i + \sum_{j\neq i}^{n} p_{ij}y_j, \quad i = 1,2,\ldots,n.$$

2.3. Interpretación de los Elementos de P.

En la sección anterior obtuvimos

$$\hat{y}_i = p_{ii}y_i + \sum_{j\neq i}^{n} p_{ij}y_j, \quad i = 1,2,\ldots,n.$$

Derivando esta expresión con respecto a y_i obtenemos,

$$\partial\hat{y}_i/\partial y_i = p_{ii}, \quad i = 1,2,\ldots,n.$$

y los p_{ii} pueden ser interpretados como la cantidad en que cada valor y_i determina a $\hat{y}_i$ (Hoaglin y Welsch (1978)). Similarmente, p_{ij} puede ser considerado como la cantidad en que cada valor y_j determina a $\hat{y}_i$.

El recíproco de p_{ii} puede interpretarse como el número efectivo (o equivalente) de observaciones que determinan a $\hat{y}_i$ (Huber (1981)). Es decir, si $p_{ii} = 1$, entonces $\hat{y}_i$ está deter-

minado por y_i solamente (una observación). Por otro lado, si $p_{ii} = 0$, y_i no tiene influencia sobre $\hat{y}_i$, mientras que si $p_{ii} = .5$, $\hat{y}_i$ está determinado por el equivalente de dos observaciones.

Los elementos de P también tienen interpretación geométrica. Primero, cuando X contiene una columna constante o cuando las columnas de X están centradas, la forma cuadrática $v'(X'X)^{-1}v = c$, donde v es un vector $k\times 1$ y c es una constante, define los contornos elípticos k-dimensionales centrados en $\bar{x}$, el vector que contiene los promedios de las columnas de X. El más pequeño conjunto convexo que contiene la dispersión de los n puntos de X está contenido en las elipsoides que satisfacen que $c \leqslant \max(p_{ii})$. Por tanto p_{ii} está determinado por la localización de x_i en el espacio X; un gran (pequeño) valor de p_{ii} indica que x_i cae lejos (cerca) de la masa de los otros puntos.

En segundo lugar, el volumen de la elipsoide de $(1-\alpha)\%$ de confianza para β incrementa monotonamente con los p_{ii}. A mayor p_{ii} mayor es el incremento del volumen de la elipsoide de confianza para β cuando la i-ésima observación es omitida (lo veremos más adelante).

Para otra interpretación véase, por ejemplo, a Chatterjee y Hadi (1988).

2.4. La Distribución de los p_{ii}.

La matriz P depende de la matriz de datos X y por los supuestos dados, ella es fija. Sin embargo, en algunos casos, X está medida con error y en ocasiones parece razonable asumir que las filas de X tienen una distribución multinormal con vector de medias μ y matriz de covarianza Σ. En este caso Besley,

Kuh y Welsch (1980) prueban que si las filas de $X = (I-n)^{-1}11')X$, donde 1 es un vector de n unos, son i.i.d., de una población Multinormal $(k-1)$ dimensional entonces,

(a) $$(n-k)(p_{ii}-n^{-1})/\{(k-1)(1-p_{ii})\} \sim F(k-1,n-k)$$

Es importante observar que si $i \neq j$, no podemos asumir las filas $\tilde{x}_i$, $\tilde{x}_j$ sean independientes, pues

$$\text{cov}(\tilde{x}_i,\tilde{x}_j) = E(\tilde{x}_i.\tilde{x}_j)-\{E(\tilde{x}_i)\}\{E(\tilde{x}_j)\}'$$

$$= \mu\mu'-2(\mu\mu'+n^{-1}\Sigma)+\mu\mu'+n^{-1}\Sigma = n^{-1}\Sigma,$$

la cual converge a cero cuando $n \to \infty$.

Por tanto la dependencia desaparece a medida que n crece, y el resultado anterior es cierto sólo aproximadamente.

(b) Si el modelo no contiene constante y las x_i, $i = 1,2,\ldots,n$, son i.i.d., de una $N_k(\mu,\Sigma)$, entonces

$$(n-k-1)\tilde{p}_{ii}/\{(k-1)(1-p_{ii})\} \sim F(k-1,n-k)$$

donde $\tilde{p}_{ii} = \tilde{x}_i'(\tilde{X}'\tilde{X})^{-1}\tilde{x}_i$ y $p_{ii} = n^{-1}+\tilde{p}_{ii}$ (Chaterjee y Hadi, 1988).

2.5. Observaciones 'outliers', de Alto Poder y Observaciones Influenciales.

Antes de presentar algunos de los procedimientos frecuentemente empleados para estudiar el comportamiento de los datos sobre el ajuste de la ecuación de regresión, es conveniente

afianzar un poco más los conceptos de 'outliers', puntos de alto poder y observaciones influenciales. A continuación daremos definiciones para ellos y veremos como interactúan entre sí.

Outliers. En el contexto de la regresión lineal, definimos un 'outliers' como una observación para la cual su residual estandarizado es grande en magnitud (la apropiada desviación estándar para los residuales será definida en la sección 3) con respecto a las otras observaciones en el conjunto de datos. Las observaciones son señaladas como 'outliers' sobre la base de que la regresión ajustada no es capaz de acomodarlas.

Observaciones de alto poder. Son aquellas para las cuales el vector X_i correspondiente cae lejos, en algún sentido, del resto de datos. Equivalentemente, un punto con alto poder es una observación con un P_{ii} grande en comparación con las otras observaciones del conjunto. Las observaciones en el espacio X que se encuentran aisladas tienen un alto poder. Los puntos con alto poder pueden ser considerados como 'outliers' en el espacio X.

Observaciones influenciales. Son aquellas que individualmente o colectivamente tienen una influencia excesiva sobre la regresión ajustada comparadas con las otras observaciones del conjunto. Esta definición es subjetiva pero implica que podemos ordenar las observaciones de manera razonable de acuerdo a alguna medida de su influencia.

Sin embargo, una observación puede no tener la misma influencia sobre todos los resultados de la regresión. Por ejemplo, una observación puede tener influencia sobre $\hat{\beta}$, los valores ajustados, y/o los estadísticos de ajuste. El primer objetivo es entonces determinar cuál influencia debemos considerar.

Por ejemplo, si la estimación de β es el interés primordial, entonces la medición de la influencia de las observaciones sobre $\hat{\beta}$ es la indicada, mientras que si el objetivo es la predicción, entonces una medida de la influencia sobre los valores predichos es más apropiada que la medición de la influencia sobre β.

También debemos tener en cuenta que:

(a) Los puntos 'outliers' no son necesariamente observaciones influenciales.

(b) Las observaciones influenciales no son necesariamente 'outliers'.

(c) Mientras que las observaciones con grandes residuales no son deseables, un residual pequeño no implica que la observación correspondiente sea típica. Esto se debe a que los mínimos cuadrados evita grandes residuales y puede acomodar un punto que no es típico a expensas de otros puntos en el conjunto de datos. En efecto, la tendencia general de puntos con alto poder es la de tener residuales pequeños e influenciar el ajuste en forma desproporcionada.

(d) En forma análoga a los residuales, los puntos de alto poder no son necesariamente influenciales y las observaciones influenciales no son necesariamente puntos de alto poder. Sin embargo, los puntos de alto poder son probablemente influenciales.

Los ejemplos 1 y 2 ilustran claramente las situaciones anteriores.

3. Efectos de una Observación Sobre la Ecuación de Regresión.

La hipótesis implícita en el ajuste de M.C. dice que todos los datos deberían jugar un papel igual en la determinación de los resultados de la estimación y en las conclusiones que se desprenden del análisis del modelo

$$Y = X\beta + \varepsilon$$

Generalmente, sin embargo, no todas las observaciones tienen la misma influencia, y puede ocurrir que una o varias observaciones tengan una influencia excesiva sobre el ajuste y las conclusiones derivadas del análisis. En estos casos es importante poder identificarlas y confirmar sus efectos sobre los diferentes aspectos de la regresión. En la literatura estadística existe una gran cantidad de medidas propuestas para diagnosticar observaciones influenciales. Dos formas diferentes de aproximarse a la medición de la influencia son:

(a) El método de la *eliminación*, y

(b) El método de la *diferenciación*.

En (a) se examina cómo los resultados y las conclusiones del análisis de regresión cambian cuando una (o varias) observación(es) se omite(n).

En (b) se examinan las tasas de cambio (las derivadas) de varios resultados de la regresión con respecto a ciertos parámetros del modelo.

En lo que sigue estudiaremos el método de la eliminación para el caso de *una observación influencial*. La siguiente sección desarrollará el caso más general de un grupo de observa-

ciones influenciales.

Para el método de la diferenciación véase por ejemplo Besley et al (1980), o Chatterjee y Hadi (1988).

El Método de la Eliminación.

Existe un grupo de métodos interrelacionados para detectar una observación influencial y medir sus efectos sobre varios aspectos del análisis. Estos métodos se pueden dividir en siete grupos dependiendo del aspecto del ajuste en que nos interesemos; de acuerdo a esto los dividiremos en métodos basados en

(a) residuales

(b) la lejanía de los puntos en el espacio X-Y,

(c) la curva de influencia (centro de las elipsoides de confianza),

(d) el volumen de las elipsoides de confianza,

(e) la función de verosimilitud,

(f) subconjuntos de coeficientes de regresión, y,

(g) la estructura de valores y vectores propios de $X'X$ o en la descomposición en valores singulares de X.

En lo que sigue daremos algunas de las medidas más empleadas surgidas de cada uno de estos métodos.

(a) Medidas Basadas en Residuales.

De la Sección 1.3(e) vimos que $e = (I-P)\varepsilon$. Esta identidad indica que para que e sea un substituto razonable de ε, los elementos fuera de la diagonal de P deben ser suficientemente pequeños. Además, si los elementos de ε son independientes y tienen la misma varianza, dicho resultado muestra que los re-

siduales no son independientes (a menos que P sea diagonal) y que no tienen la misma varianza (a menos que los elementos de la diagonal de P sean iguales). En consecuencia, los residuales e_i pueden ser considerados como un sustituto razonable de los ε_i si las filas de X son homogéneas (y por tanto los elementos de la diagonal de P son aproximadamente iguales) y los elementos fuera de la diagonal de P son suficientemente pequeños.

Por estas razones, es preferible usar una versión transformada de los resisduales ordinarios para propósitos de diagnósticos.

En lugar de los e_i podemos usar

$$g(e_i,\sigma_i) = e_i/\sigma_i$$

donde σ_i es la desviación estandar del i-ésimo residual.

Hay cuatro posibilidades para la transformación anterior:
$a_i = e_i/\{e'e\}^{1/2}$ el residual normalizado,
$b_i = e_i/\hat{\sigma}$, donde $\hat{\sigma} = \{e'e/(n-k)\}^{1/2}$ el residual estandarizado,
$r_i = e_i/\{\hat{\sigma}(1-p_{ii})\}$ el residual internamente studentizado y,
$r_i^* = e_i/\{\hat{\sigma}_{(i)}(1-p_{ii})\}$ el residual externamente studentizado, donde
$\hat{\sigma}^2(i) = \{Y'_{(i)}(I-P)_{(i)})Y_{(i)}\}/(n-k-1)$, $i = 1,2,\ldots,n$, es el estimador del error cuadrático medio cuando se omite la i-ésima observación y

$$P_{(i)} = X'_{(i)}(X'_{(i)}X_{(i)})^{-1}X'_{(i)}, \quad i = 1,2,\ldots,n,$$

es la matriz de predicción para $X_{(i)}$.

Cuál de las formas anteriores es la más adecuada para diagnosticar? Las cuatro versiones están muy relacionadas entre sí; las dos primeras son muy simples y no reflejan la varianza de

e_i; para diagnosticar son basicamente equivalentes.

Behnken y Draper (1972) sugieren, que si los elementos de la diagonal de P (y por lo tanto las varianzas de e_i, $i = 1,2, \ldots,n$), varían sustancialmente debería preferirse usar r_i.

Muchos autores (por ejemplo Besley et al. (1980); Atkinson, (1981, 1982,1985); y Velleman y Welsch, (1981), prefieren r_i^* sobre r_i por varias razones:

(i) r_i^* puede interpretarse como el estadístico para contrastar la significancia del i-ésimo vector unidad u_i en el modelo de '*outlier*' de cambio de media definido como

$$E(Y) = X\beta + u_i\theta$$

donde θ es el coeficiente de regresión de u_i, (Chatterjee y Hadi (1988)).

(ii) $r_i^* \sim t(n-k-1)$ para la cual existen tablas facilmente disponibles.

(iii) r_i^* es una transformación monótona de r_i pues

$$r_i^* = r_i\{(n-k-1/(n-k-r_i^2)\}^{1/2}$$

y puesto que $r_i^* \to \infty$ cuando $r_i \to n-k$, r_i^* refleja más dramáticamente las grandes desviaciones de lo que lo hace r_i.

(iv) El estimador $\hat{\sigma}_{(i)}$ es robusto a errores burdos en la i-ésima observasión.

Ahora bien, el patrón que siguen los residuales es más informativo que sus magnitudes y por tanto los métodos gráficos son más útiles que los procedimientos formales de contrastes.

Behnken y Draper (1972), sugieren el empleo de los r_i en la construcción de los gráficos, mientras Atkinson (1981) pre-

fiere los r_i^*. Puesto que r_i^* es una transformación monótona de los r_i las conclusiones obtenidas de los dos gráficos son generalmente las mismas; sin embargo es más fácil la identificación de 'outliers' en los gráficos basados en los r_i^*.

Algunos de los gráficos de residuales más comunes son:

(i) Distribución de frecuencia de los residuales: histogramas, gráficos de puntos, gráficos 'stem and leaf' y de cajas esquemáticas.

(ii) Gráfico de los residuales en el tiempo.

(iii) Gráfico de probabilidad normal o seminormal.

(iv) Gráfico de residuales contra los valores ajustados.

(v) Gráfico de residuales contra X_j, $j = 1,2,\ldots,k$.

(vi) Gráficos aditivos.

(vii) Gráficos de componentes-más-residuales.

(viii) Gráficos de residuales parciales aumentados.

Para una descripción de estos métodos gráficos véase, por ejemplo, Seber (1977), Daniel y Wood (1980), Atkinson (1985), Chatterjee y Hadi (1988).

(b) Medidas Basadas en la Lejanía de los Puntos en el Espacio X-Y.

En esta sección discutiremos algunas cantidades para medir el poder de un punto y también combinaremos los valores de poder y los residuales en un gráfico llamado L-R. Este gráfico permite distinguir entre puntos de alto poder y 'outliers'.

(i) Elementos de la diagonal de P.

Como vimos anteriormente los elementos de la matriz P y en particular los elementos de su diagonal, los p_{ii}, juegan un pa-

pel importante en la determinación de los valores ajustados, los residuales y su estructura de varanza y covarianza. Por esta razón Hoaglin y Welsch (1978) sugieren el exámen de r_i^* y p_{ii} y señalan que estos dos aspectos de la búsqueda en los datos son complementarios y que ninguno es suficiente por sí mismo.

Una pregunta natural es, qué tan grande debe ser un p_{ii} grande? A continuación se sugieren dos cotas usadas comunmente: **(1)** Huber (1981) sugiere que puntos con $p_{ii} > 0.2$ sean clasificados como observaciones de alto poder. Esta regla recomienda que se preste atención especial a las observaciones cuyos valores ajustados estén determinados por un equivalente de cinco o menos observaciones.

(2) Hoaglin y Welsch (1978) sugieren que puntos con $p_{ii} > 2/n$ sean clasificados como puntos de alto poder.

Las cotas anteriores no deberían ser usadas mecanicamente; ellas deberían servir como medidas aproximadas de guía general para detectar problemas en los datos. Se recomienda una comparación de los elementos de la diagonal de P através de gráficos tales como gráficos de los p_{ii} contra el número de la observación, diagramas 'stem-leaf' y/o cajas esquemáticas.

(ii) La distancia de Mahalanobis.

El poder de una observasión puede ser medido por medio de la distancia de Mahalanobis,

$$M_i = n(n-2)(p_{ii}-n^{-1})/[(n-1)(1-p_{ii})]$$

Esta distancia es equivalente a p_{ii}.

(iii) La distancia cuadrática estandarizada ponderada

Suponga que el modelo tiene un término constante definamos

$$c_{ij} = \hat{\beta}_j(x_{ij}-\bar{x}_j), \quad i = 1,2,\ldots,n, \quad j = 1,2,\ldots,k,$$

donde $\bar{x}_j$ es el promedio de la j-ésima columna de X. La cantidad c_{ij} puede ser considerada como el efecto de la j-ésima variable sobre el i-ésimo valor ajustado $\hat{y}_i$. Se puede mostrar que $\hat{y}_i-\bar{y} = \sum_j c_{ij}$, donde $\bar{y}$ es el promedio de y. Daniel y Wood (1971) sugieren usar la distancia cuadrática estandarizada ponderada

$$WSSD_i = \sum_j c_{ij}^2 / s_y^2$$

donde $s_y^2 = \sum_i (y_i-\bar{y})^2/(n-1)$, para medir la influencia de la i-ésima observación sobre $\hat{y}_i-\bar{y}$. $WSSD_i$ es una medida de la suma de cuadrados de las distancias de x_{ij} con respecto a la media de la j-ésima variable $\bar{x}_j$, ponderada por la importancia relativa de la j-ésima variable (la magnitud del coeficiente de regresión estimado). Por tanto, $WSSd_i$ será grande si la i-ésima observación es extrema en al menos una variable cuyo coeficiente de regresión estimado es grande en magnitud.

Un gráfico de dispersión que es efectivo para distinguir entre un punto de alto poder y un 'outler' es llamado el gráfico L-R. Este se define como el gráfico de dispersión de los valores p_{ii} contra los residuales normalizados al cuadrado a_i^2 definidos antes. El diagrama de dispersión debe caer dentro del triángulo definido por las siguientes tres condiciones:

(i) $0 \leqslant p_{ii} \leqslant 1$

(ii) $0 \leqslant a_i^2 \leqslant 1$

(iii) $p_{ii} + a_i^2 \leqslant 1$

Por tanto los puntos que caen en la esquina inferior derecha del diagrama son outliers y los que caen en la esquina superior izquierda son puntos de alto poder.

(c) La Curva de Influencia.

Una clase importante de medidas de la influencia de la i-ésima observación sobre los resultados de la regresión está basada en la idea de la *curva de influencia o función de influencia* introducida por Hampel (1974). Para una discusión sobre este concepto véase Chatterjee y Hadi (1988). Algunas de las medidas de influencia derivadas de este concepto son:

(i) La distancia de Cook.

Bajo normalidad, la región de $(1-\alpha)\%$ de confianza conjunta para β se obtiene de

$$(\beta-\hat{\beta})(X'X)(\beta-\hat{\beta})/k\sigma^2 \leq F(\alpha;k,n-k)$$

donde $F(\alpha;k,n-k)$ es el percentil α-superior de una distribución F con k y $n-k$ grados de libertad. Esta desigualdad define una región elipsoidal centrada en $\hat{\beta}$. La influencia de una observación puede ser medida por el cambio en el centro de esta región cuando la i-ésima observación es omitida. Cook (1977) definió la medida

$$C_i = (\hat{\beta}-\hat{\beta}_{(i)})(X'X)(\hat{\beta}-\hat{\beta}_{(i)})/k\sigma^2 \qquad i = 1,2,\ldots,n,$$

$$= p_{ii}r_i^2/(k(1-p_{ii}))$$

para medir la influencia de la i-ésima observación sobre el

centro de la elipsoide 0, equivalente, sobre los coeficientes estimados. Esta medida combina información sobre el alto poder de la observación y r_i que da información sobre 'outliers'. Grandes valores de C_i indican que la observación es influencial. Cook (1977) sugiere que cada C_i sea comparada con el percentil de una F con K y $n-k$ grados de libertad. Para una discusión sobre la distribución de los C_i, (Cook, (1977)).

(ii) La distancia de Welsch-Kuh.

La influencia de la i-ésima observación sobre el valor predicho $\hat{y}_i$ puede medirse como el cambio (con relación al error estandar de $\hat{y}_i$) en la predicción en x_i cuando la i-ésima observasión es omitida, es decir:

$$WK_i = |\hat{y}_i - \hat{y}_{i(i)}| / \sigma(P_{ii})^{1/2} = |x_i'(\hat{\beta} - \hat{\beta}_{(i)})| / \sigma(P_{ii})^{1/2}$$

Besley et al. (1980) sugieren usar $\hat{\sigma}_{(i)}$ en lugar de σ. Haciendo algunos reemplazos obtenemos,

$$WK_i = |r_i^*| (p_{ii}/(1-p_{ii}))^{1/2}.$$

Besley et al. la llamaron $DFFITS_i$ debido a que es la diferencia escalada entre $\hat{y}_i$ y $\hat{y}_{i(i)}$. Grandes valores de WK_i indican que la observación es influencial. La distribución de WK_i aunque no es exactamente una t es similar a ella. Debido a esto Velleman y Welsch (1981) sugieren que valores mayores que 1 o 2 parecen ser razonables para indicar puntos que merezcan atención especial. Para otros posibles puntos de calibración para WK_i, (Chatterjee y Hadi (1988)).

sión de una observación con un gran residual producirá una gran reducción en la suma de cuadrados residual SSE. La influencia de una observación puede ser medida combinando estas dos ideas y calculando entonces el cambio tanto en $e'e$ como en $\det(X'X)$. Andrews y Pregibon (1978) sugieren en cociente:

$$AP_i = SSE_{(i)}\det(X'_{(i)}X_{(i)})/SSE\det(X'X), \quad i = 1,2,\ldots,n.$$

Puede mostrarse que la expresión anterior se reduce a (Chatterjee y Hadi (1988)):

$$AP_i = p_{ii} + e_i^2/e'e, \quad i = 1,2,\ldots,n,$$

y entonces AP_i no distingue entre puntos de alto poder y puntos outliers en el espacio Z formado por $X:Y$. Valores pequeños de AP merecen atención especial. Para otros comentarios sobre AP, (Draper y John (1981)).

(ii) El cociente de varianzas.

Una medida alternativa a AP se basa en medir la influencia de i-ésima observación comparando $\text{cov}(\hat{\beta})$ y $\text{cov}(\hat{\beta}_{(i)})$. Si el $\text{rango}(X_{(i)}) = k$, estas matrices son definidas positivas y existen varias formas de compararlas; el cociente de sus trazas o el cociente de sus determinantes. Besley et al. (1980) sugieren usar el cociente de sus determinantes,

$$VR_i = \det(\sigma^2_{(i)}(X'_{(i)}X_{(i)})^{-1}/\det(\sigma^2(X'X)^{-1}), \quad i = 1,\ldots,n.$$

Besley et al. (1980), la denominan COVRATIO y muestran que:

$$VR_i = \{(n-k-r_i^2)/(n-k-1)\}^k(1/(1-p_{ii})).$$

(iii) La distancia de Cook modificada.

Atkinson (1981) ha sugerido usar una versión modificada de la distancia de Cook para detectar observaciones influenciales. Esta modificación cambia σ^2 por $\hat{\sigma}^2_{(i)}$, toma la raíz cuadrada de C_i y ajusta C_i por el tamaño muestral. Con estas modificaciones obtenemos

$$C_i^* = |r_i^*|(p_{ii}(n-k)/((1-p_{ii})k))^{1/2} = WK_i((n-k)/k)^{1/2}.$$

Atkinson (1981) resalta que esta modificación mejora a C_i en tres formas:

(a) C_i^* da más énfasis a puntos extremos.

(b) C_i^* es más adecuado para procedimientos gráficos tales como gráficos de probabilidad normal, y

(c) para el caso perfectamente balanceado donde $p_{ii} = k/n$, para todo i, el gráfico de los C_i^* es idéntico al de $|r_i^*|$.

(d) Medidas Basadas en el Volumen de Elipsoides.

Una clase alternativa de medidas de la influencia de la i-ésima obsevación está basado en el cambio del volumen de la elipsoide de confianza cuando la i-ésima observación es omitida. Veamos algunas de ellas.

(i) El estadístico de Andrews-Pregibon.

El volumen de la elipsoide de confianza para β es inversamente proporcional a la raíz cuadrada del $\det(X'X)$. Una medida de la influencia de la i-ésima observación sobre el volumen de la elipsoide de confianza para β puede ser obtenida de la comparación del $\det(X'X)$ y $\det(X'_{(i)}X_{(i)})$. Por otro lado, la omi-

VR_i será mayor que uno cuando r_i^2 es pequeño y p_{ii} es grande, y será menor que uno cuando r_i^2 es grande y p_{ii} es pequeño. Pero cuando r_i^2 y p_{ii} son ambos grandes (o ambos pequeños), VR_i tiende a uno. Estos factores reducen la habilidad de VR_i para detectar observaciones influenciales. Sin embargo, del análisis de varios conjuntos de datos, se ha observado que VR_i señala correctamente las observaciones influenciales; esto quizás se deba al hecho de que las observaciones con p_{ii} grande tienden a jalar la ecuación ajustada hacia ellas y en consecuencia tienen un pequeño residual. Idealmente, cuando todas las observaciones tienen la misma influencia sobre la matriz de covarianzas, VR_i es aproximadamente uno. La desviación con respecto a la unidad indica que la i-ésima observación es potencialmente influencial. Besley et al. (1980) proporcionan puntos de calibración para VR_i. Encuentran que la i-ésima observación es posiblemente influencial si $|VR_i - 1| \geq 3k/n$.

(e) Medidas Basadas en la Función de Verosimilitud.

Asumiendo que $Y \sim N_n(X\beta, \sigma^2 I)$, el logaritmo de la función de verosimilitud de β y σ^2 es,

$$l(\beta,\sigma^2) = -n\ln(2\pi)/2 - n\ln(\sigma^2)/2 - (Y-X\beta)\cdot(Y-X\beta)/2\sigma^2.$$

Una región de $(1-\alpha)\%$ de confianza para β y σ^2 está dada por (Cox y Hinkley (1974)):

$$\{(\beta,\sigma^2) : 2[l(\tilde{\beta},\tilde{\sigma}^2) - l(\beta,\sigma^2)] \leq \chi^2_{(\alpha;k+1)}\},$$

donde $\tilde{\beta}$ y $\tilde{\sigma}^2$ son los estimadores máximo verosímiles para β y σ^2 dados por $\tilde{\beta} = \hat{\beta}$ y $\tilde{\sigma}^2 = \hat{\sigma}^2(n-k)/n$.

La influencia de la i-ésima observación puede ser medida por la distancia entre $I(\tilde{\beta},\tilde{\sigma}^2)$ y $I(\tilde{\beta}_{(i)},\tilde{\sigma}^2_{(i)})$. Cook y Weisberg (1982) definen la distancia de verosimilitud como:

$$LD_i(\beta,\sigma^2) = 2[I(\tilde{\beta},\tilde{\sigma}^2)-I(\tilde{\beta}_{(i)},\tilde{\sigma}^2_{(i)})], \quad i = 1,2,\ldots,n.$$

En Chatterjee y Hadi (1988) se prueba que:

$$LD_i(\beta,\sigma^2) =$$

$$= n \ln\{n(n-k-r_i^2)/(n-1)(n-k)\} + \{(n-1)r_i^2/(1-p_{ii})(n-k-r_i^2)\}-1$$

y los autores sugieren que sea comparada con una distribución $\chi^2_{(k+1)}$. Es importante observar que $LD_i(\beta,\sigma^2)$ se basa sobre el modelo de probabilidad usado mientras que las otras medidas de influencia discutidas son estrictamente numéricas. Una ventaja de la distancia de verosimilitud es que puede ser extendida a otros modelos fuera del modelo lineal normal.

(f) Medidas Basadas en Subconjuntos de Coeficientes de Regresión.

Besley et al. (1980) sugieren medir la influencia de la i-ésima observación sobre el j-ésimo coeficiente de regresión estimado como:

$$DFBETAS_{ij} = (\hat{\beta}_j-\hat{\beta}_{j(i)})/\text{var}(\hat{\beta}_j)$$

$$= \{r_i^* w_{ij}/(W_j'W_j)^{1/2}\}(1/(1-p_{ii})^{1/2})$$

donde $W_j = (I-P_{[j]})X_j$ siendo $P_{[j]}$ la matriz predicción de la matriz X sin la j-ésima columna X_j, y donde w_{ij} es el i-ésimo

elemento de W_j. Valores de [DFBETAS] que exceden $2/n$ merecen especial atención.

Chatterjee y Hadi (1988) describen una medida para la influencia de la i-ésima observación sobre q combinaciones lineales independientes de los coeficientes de regresión.

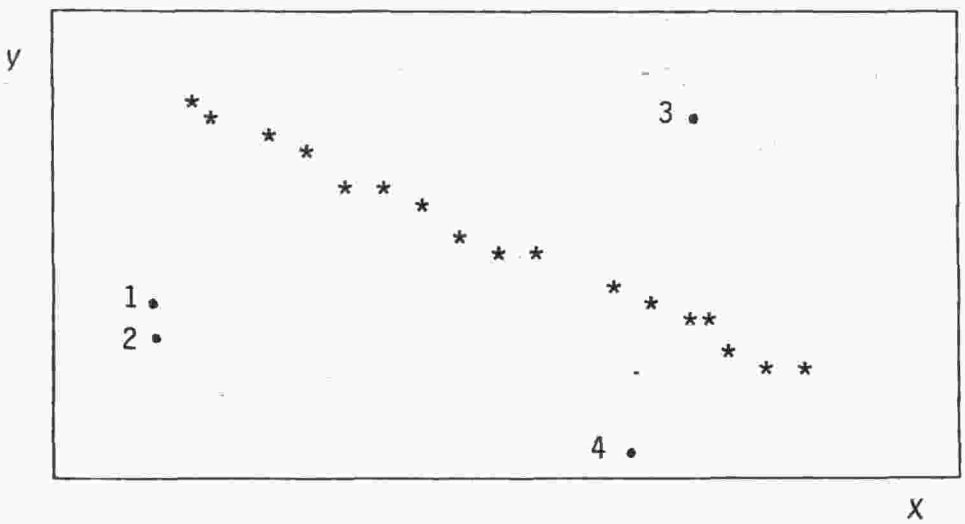
4. Efecto de Multiples Observaciones Sobre la Ecuación de Regresión.

Hasta aquí hemos presentado algunos métodos para la detección de observaciones que *individualmente* pueden ser consideradas 'outliers', de alto poder o influenciales. En esta sección discutiremos la necesidad de extender esas medidas para el caso donde varias observaciones actúan conjuntamente sobre los resultados y las conclusiones de la ecuación ajustada. El problema de múltiples observaciones es importante tanto desde el punto de vista teórico como practico. Desde el punto de vista teórico, pueden existir situaciones en las que las observaciones pueden ser conjuntamente, pero *no individualmente*, influenciales, el gráfico 3 ilustra esta situación. Los puntos 1 y 2 no son individualmente influenciales, pero conjuntamente tienen una gran influencia sobre el ajuste. Esta situación es conocida como *enmascaramiento*, debido a que la influencia de una observación es enmascarada por la presencia de otra (s). Atkinson (1986) y Leroy y Rousseeuw (1984) sugieren procedimientos para tratar estas situaciones. Por otro lado, las observaciones 3 y 4 son individualmente influenciales pero no conjuntamente, es decir, cuando las dos son simultaneamente omitidas.

Desde el punto de vista práctico, cuando la influencia múltiple existe es mucho más severa que en el caso individual y

frecuentemente es pasada por alto en el proceso de ajuste de la ecuación debido a que es más difícil de detectar. En muchas situaciones las observaciones con influencia conjunta pueden detectarse empleando los diagnósticos para el caso de la influencia individual.

Gráfico 3.



Influencia conjunta pero no individual (puntos 1 y 2) e influencia individual pero no conjunta (puntos 3 y 4).

Hay tres problemas inherentes al caso de observaciones múltiples. El primero es cómo determinar el tamaño del subconjuto de observaciones conjuntamente influenciales. El segundo problema es computacional. Suponga que conocemos el tamaño adecuado m para las observaciones conjuntamente influenciales. Habrán entonces $n!/m!(n-m)!$ posibles subconjuntos para los cuales debemos calcular las medidas de influencia. Aún hoy día, esto puede ser prohibitivo si m y n son grandes. El tercer problema es la dificultad de representar graficamente las observaciones. Las observaciones con influencia múltiple frecuente-

mente no permiten ser examinadas por los métodos antes mencionados, como cajas esquemáticas, gráficos de las observaciones contra su índice, gráficos 'stem-and leaf', etc.

Los procedimientos dados anteriormente para el caso de una observación pueden ser generalizados en forma directa, en su mayoría, al caso de múltiples observaciones influenciales. El lector interesado puede referirse a Besley et al. (1980) o, Chatterjee y Hadi (1988). Estos procedimientos tienen el inconveniente de la determinación del subconjunto m. Un procedimiento introducido por Atkinson (1986) proporciona un método, basado en la regresión de la mínima mediana al cuadrado (Rousseuw (1984)) que permite determinar el conjunto de observaciones influenciales a las cuales se les aplicarán las medidas de influencia generalizadas.

Otro procedimiento que no necesita la especificación del tamaño m del número de observaciones conjuntamente influenciales se basa en la técnica del análisis 'Cluster' (Gray y Ling (1984)).

5. Puntos Influenciales y Colinealidad.

Cuando existen relaciones lineales cercanas entre las columnas de la matriz X, se presenta un problema llamado *colinealidad*. Esta puede causar graves daños sobre la estimación por mínimos cuadrados de la ecuación de regresión. Por ejemplo, la colinealidad puede inflar las varianzas de los coeficientes de regresión estimados, alterar los signos esperados de ellos y producir resultados inestables numéricamente.

Para medir el grado de colinealidad, Besley et al. (1980)

emplean el número de condición de la matriz X, el cual se define como:

$$K = d_1/d_k = (\lambda_1/\lambda_k)^{1/2}$$

donde $d_1 \geq d_2 \geq \ldots \geq d_k$ son los valores singulares de X, es decir son los valores de la diagonal de la matriz D, de dimensión $k \times k$, obtenida de la descomposición $X = UDV'$, donde $U'U = V'V = I$ y donde $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$ son los valores propios de $X'X$. Un Valor grande para K es indicativo de la existencia de al menos una relación de dependencia cercana entre al menos dos columnas de la matriz X. El mínimo de K es uno.

Ahora bien, existen dos problemas con K:

(i) K no es invariante bajo cambio de escala de las columnas de X.

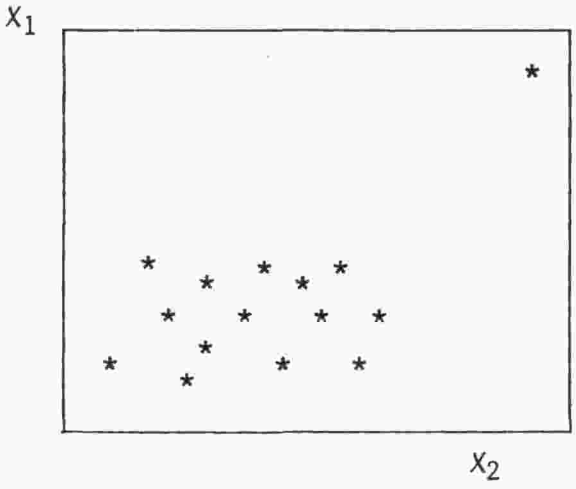
(ii) K puede ser fuertemente influenciado por uno o algunos pocos puntos extemos en el espacio X.

El primer problema se resuelve estandarizando cada columna de X de forma que tenga media cero y varianza uno (véase Besley (1984) para una discusión sobre las ventajas y desventajas de centrar y normalizar).

El segundo problema es más complejo. Los puntos de alto poder tienden a influenciar la estructura de valores y vectores propios y por tanto el número de condición de X. Por ejemplo, en el gráfico 4 vemos que un punto puede crear una colinealidad o encubrir una como lo muestra el gráfico 5.

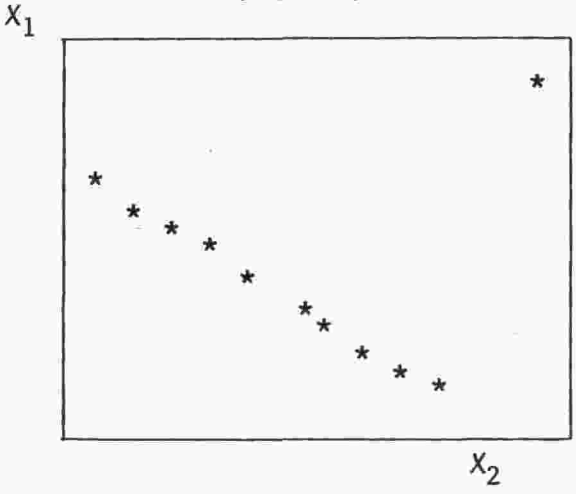
Por tanto un número de condición pequeño no necesariamente significa que X no esté mal condicionada. Además, uno o dos puntos pueden ser los culpables de que exista un número de condición grande. Los puntos que esconden o crean colinealidad los

Gráfico 4.



Colinealidad creada por un punto.

Gráfico 5.



Colinealidad enmascarada por un punto.

llamaremos puntos de *colinealidad influencial*. Estos puntos son generalmente, aunque no necesariamente, observaciones con alto poder. Sin embargo, no todas las observaciones de alto poder son puntos de colinealidad influencial y no todos los puntos de colinealidad influencial son observaciones de alto poder.

El siguiente ejemplo aparece en Chatterjee y Hadi (1988).

Ejemplo 3.

La tabla 1 muestra los datos para tres variables explicativas X_1, X_2, X_3, p_{ii} y el cambio relativo en número de condición definido como $|K_{(i)}-K|/K$, donde $K_{(i)}$ es el número de condición de $X_{(i)}$.

TABLA 1.

Observación	X_1	X_2	X_3	p_{ii}	$\|K_{(i)}-K\|/K$
1	15	-3	-3	.45	.31
2	15	-2	-1	.21	9.24
3	-5	0	1	.07	.35
4	7	-1	2	.17	.70
5	-8	1	-1	.11	.84
6	8	-2	0	.09	1.19
7	-14	2	-1	.22	6.19
8	2	-1	-1	.08	3.36
9	-6	0	0	.11	2.90
10	2	-2	2	.22	5.62
11	-4	1	2	.14	2.70
12	4	0	2	.20	6.68
13	-8	2	-2	.17	2.68
14	3	1	-2	.19	3.93
15	-3	2	-1	.16	5.88
16	-4	-1	2	.22	5.54
17	-14	3	0	.19	7.04

En este ejemplo la observación 1 es de alto poder pero no es de colinealidad influencial y la observación 2 es de colinealidad influencial pero no es un punto de alto poder.

Diagnósticos de Colinealidad Influencial.

Influencia sobre el número de condición. Una de las medidas propuestas para cuantificar la influencia sobre el número de condición de la matriz X es análoga al cambio relativo $|K_{(i)}-K|/K$ y se define como

$$|H_i| = |\tilde{K}_{(i)} - K|/K$$

donde $\tilde{K}_{(i)}$ es una aproximación al número de condición de $X_{(i)}$ dada en Chatterjee y Hadi (1988), página 166. H_i puede interpretarse como el cambio relativo en el número de condición de X que resulta al omitirse la i-ésima observación. Si H_i es grande y positiva, entonces la eliminación de x_i incrementa el número de condición, y si H_i es grande y negativa, la omisión de x_i disminuirá el número de condición

Otras medidas para determinar la influencia sobre el número de condición y los indices de condición, así como métodos gráficos para detectar puntos colineales influenciales se encuentran en Chatterjee y Hadi (1988); estos autores también presentan un procedimiento para el caso de múltiples puntos influenciales debido a Kempthorne (1986), el cual no requiere de la especificación del tamaño del subconjunto.

Conclusiones.

Hemos discutido varias de las medidas más empleadas para el estudio de 'outliers', puntos de alto poder y observaciones influenciales, tanto en el caso de una sola observación como en el de múltiples observaciones.

Las medidas presentadas detectan la influencia sobre dife-

rentes aspectos de la ecuación ajustada: coeficientes de regresión estimados, valores ajustados, matriz de covarianzas de los coeficientes estimados, elipsoides de confianza para β, elipsoides de confianza para β y σ^2.

En el caso de la colinealidad se ilustró la relación entre puntos de alto poder y el número de condición de la matriz X; pueden existir observaciones que crean colinealidad y otras que la enmascaran y de ahí la importancia de detectar la existencia de ellas.

Por último, para la aplicación de las medidas de influencia en el procedimiento de ajuste de la ecuación de regresión, existen paquetes estadísticos que traen programadas algunas de ellas. Hadi (1988) ha construído un paquete interactivo y dirigido por menú, llamado SAILR, en el cual ha implementado todos los procedimientos estadísticos y gráficos presentados en su libro *Sensitivity Analysis in Linear Regression*, escrito con Chatterjee. Este paquete de fácil aplicación, permite chequear eficientemente y de forma muy completa el comportamiento de las observaciones en cada aspecto del análisis.

* *

BIBLIOGRAFIA

Andrews, D.F., y Pregibon, D., (1978). Finding Outliers That Matter. Journal of the Royal Statistical Society, (B), 40, 85-93.

Atkinson, A.C., (1982). Two Graphical Displays for Outlying and Influential Observations in Regression. Biometrika, 68, 13-20.

Atkinson, A.C., (1982). Regression Diagnostics, Transformations, and Constructed Variables (With discussion). Journal of the Royal Statistical Society (B), 44, 1-36.

Atkinson, A.C., (1985). Plots, Transformations and Regression. Oxford University Press.

Atkinson, A.C., (1986). Masking unmasked. Biometika, 73, 3, pp. 533-41.

Behnken, D.W. y Draper, N.R., (1972). Residuals and their Variance. Technometrics, 11, Nº 1, 101-111.

Besley, D.A., (1984). Demeaning Conditioning Diagnostics through Centering (With comments). The American Statistician, 38, Nº 2, 73-93.

Besley, D.A., Kuh, E., y Welsch, R.E., (1980). Regression Diagnostics. New York: Wiley.

Chatterjee, S., y Hadi, A.S., (1986). Influential Observations, High Leverage, and Outliers in Linear Regression. Statistical Science, Vol. 1, Nº 3, 379-416.

Chatterjee, S., y Hadi, A.S., (1988). Sensitivity Analysis in Linear Regression. New York: John Wiley & Sons.

Cook, R.D., (1977). Detection of Influential Observations in Linear Regression. Technometrics, 19, 15-18.

Cook, R.D., y Weisberg, S., (1982). Residuals and Influence in Regression. New York and London: Chapman and Hall.

Cox, D.R., y Hinkley, D.V., (1974). Theoretical Statisticis. London: Chapman and Hall.

Daniel, C., y Wood, F.S., (1980). Fitting Ecuations to Data: Computer Analysis of Multifactor Data. 2a Ed., New York: John Wiley & Sons.

Draper, N.R., y John, J.A. (1981). Influential Observations and Outliers in Regression. Technometrics, 23, 21-26.

Draper, R.R. y Smith, H., (1981) Applied Regression Analysis. 2a. Ed., New York: John Wiley & Sons.

Gray, J.B., y Ling, R.F., (1984). K-clustering as a Detection Tool for Influential Subsets in Regression (With Discussion). Technometrics, 26, 305-330.

Graybill, F.A., (1976). Theory and Application of the Linear Model. MA: Duxbury Press, North Scituate.

Hadi, A.S., (1988). SAILR: User's Guide. New York: Cornell University.

Hampel, F.R., (1974). The Influence Curve and Its Role in Robust Estimation. JASA, 62, 1179-1186.

Hoaglin, D.C. y Welsch, R.E., (1978). The Hat Matrix in Regression and ANOVA. The American Statistician, 32, 17-22.

Huber, P., (1981). Robust Statistics. New York: Wiley.

Kempthorne, P.J., (1986). Identifying Rank-Influencial Groups of Observations in Linear Regression Modelling. Memorandum NS-539, Department of Statistics, Harvard University.

Lerdy, A., y Rousseeuw, P.J., (1984). A multiple regression technique for detecting outliers. Report 84-33, Delft University of Technology.

Mardia, K.V., Kent, J.T. y Bibby, J.M., (1979). Multivariate Analysis. Academic Press, London.

Mosteller, F. y Tukey, J.W., (1977). Data Analysis and Regression. Addison-Wesley. Reading, Mass.

Rousseeuw, P.J., (1984). Least median of squares regression. JASA, 79, 871-880.

Seber, G.A.F., (1977). Linear regression Analysis. New York: John Wiley y Sons.

Velleman, P.F., y Welsch, R.E., (1981). Efficient Computing of Regression Diagnostics. The American Statistician, 35, 234-242.

* *