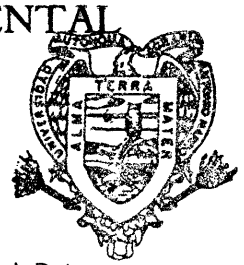
ANALISIS DE SENSIBILIDAD EN REGRESION

OSCAR ALEJANDRO MARTINEZ JAIME

TESIS

PRESENTADA COMO REQUISITO PARCIAL
PARA OBTENER EL GRADO DE
MAESTRO EN CIENCIAS
EN ESTADISTICA EXPERIMENTAL



Universidad Autónoma Agraria
"Antonio Narro"
PROGRAMA DE GRADUADOS
Buenavista, Saltillo, Coah.

NOVIEMBRE DE 2001

Universidad Autónoma Agraria Antonio Narro

Subdirección de Postgrado

ANÁLISIS DE SENSIBILIDAD EN REGRESIÓN

TESIS

Por:

OSCAR ALEJANDRO MARTÍNEZ JAIME

Elaborada bajo la supervisión del comité particular

de asesoría y aprobada como requisito parcial, para optar al grado de:

MAESTRO EN CIENCIAS

EN ESTADÍSTICA EXPERIMENTAL

Comité Particular

Asesor principal: _____

Dr. José Antonio Díaz García

Asesor: _____

Dr. Alejandro Zárate Lupercio

Asesor: _____

M.C. Luis Rodríguez Gutiérrez

_____

Dr. Ramiro López Trujillo

Subdirector de Postgrado

Buenavista, Saltillo, Coahuila. Noviembre de 2001

AGRADECIMIENTOS

A mi asesor principal, el Dr. José Antonio Díaz García, por sus consejos, sugerencias e invaluable ayuda, sin la cual, no hubiera sido posible la culminación de este trabajo de tesis.

A mis asesores, el Dr. Alejandro Zárate Lupercio y al M.C. Luis Rodríguez Gutiérrez, por su participación en la revisión del presente trabajo de tesis.

A mis profesores, el Dr. Mario Cantú Sifuentes, y a los M.C. Emilio Padrón Corral, Roberto Coronado Niño y Félix de Jesús Sánchez Pérez, por contribuir en mi formación académica.

DEDICATORIAS

A mi esposa, **Erika Elizabeth**,

por su gran amor

y apoyo incondicional en todo momento.

y

A mi hija, **Alison Elizabeth**,

por ser el regalo más preciado

que Dios me ha dado.

COMPENDIO

Análisis de Sensibilidad en Regresión

POR

OSCAR ALEJANDRO MARTÍNEZ JAIME

MAESTRÍA

ESTADÍSTICA EXPERIMENTAL

UNIVERSIDAD AUTÓNOMA AGRARIA "ANTONIO NARRO"

BUENAVISTA, SALTILLO, COAHUILA. NOVIEMBRE DE 2001

Dr. José Antonio Díaz García -Asesor-

Palabras clave: Medidas de Diagnóstico, Distancia de Mahalanobis Generalizada, Outliers, Puntos Palanca, Puntos Influyentes.

En el presente trabajo, se describen algunas herramientas básicas de Álgebra Lineal, y varias medidas de diagnóstico que permiten detectar observaciones influyentes, y se propone una modificación de la distancia de Cook, basándose en la distancia de Mahalanobis generalizada, en el contexto del modelo de regresión lineal con distribución normal multivariada. Se establece además, la distribución exacta del estadístico basado en esta distancia de Mahalanobis generalizada, la cual proporciona puntos críticos para identificar "outliers" en un conjunto de datos. Este procedimiento, se ilustra con un ejemplo, en el caso de la regresión lineal múltiple y multivariada.

ABSTRACT

Sensitivity Analysis in Regression

BY

OSCAR ALEJANDRO MARTÍNEZ JAIME

MASTER OF SCIENCE

EXPERIMENTAL STATISTICS

UNIVERSIDAD AUTÓNOMA AGRARIA "ANTONIO NARRO"

BUENAVISTA, SALTILLO, COAHUILA. NOVEMBER 2001

Dr. José Antonio Díaz García -Advisor-

Key Words: Diagnostic Measures, Generalized Mahalanobis Distance, Outliers, High Leverage Points, Influential Points.

In this work, it describe some basic tools of Linear Algebra, and several diagnostics measures that it permit to detect influential observations, and a modification of the classical Cook's distance is proposed, providing us with a generalized Mahalanobis distance in the context of multivariate normal linear regression models. It establish the exact distribution of a pivotal type statistics based on this generalized Mahalanobis distance, which provides critical points for the identification of data points. It illustrate the procedure with an example, in the context of multiple and multivariate linear regression.

ÍNDICE DE CONTENIDO

Página

ÍNDICE DE FIGURAS ix

INTRODUCCIÓN 1

LA MATRIZ DE PREDICCIÓN 6

Notación 6

Resultados 7

Supuestos 9

Las Matrices $\boldsymbol{P}$ y $(\boldsymbol{I}-\boldsymbol{P})$ 10

Propiedades de la Matriz de Predicción 13

Omitiendo o Adicionando una Observación 18

Condiciones para Valores Grandes de $\boldsymbol{p_{ii}}$ 21

Omitiendo Múltiples Filas de $\boldsymbol{X}$ 22

Eigenvalores de $\boldsymbol{P}$ y $(\boldsymbol{I}-\boldsymbol{P})$ 23

Distribución de $\boldsymbol{p_{ii}}$ 27

EFECTOS DE UNA OBSERVACIÓN SOBRE LA ECUACIÓN DE REGRESIÓN 29

Método de la Eliminación de una Observación 30

Medidas Basadas en Residuales 30

Outliers, Puntos Palanca y Puntos Influyentes . 36

Medidas Basadas en la Curva de Influencia..... 40

Medidas Basadas en el Volumen de Elipsoides de Confianza 52
Medidas Basadas en la Función de Verosimilitud 58
Medidas Basadas en un Subconjunto de Coeficientes de Regresión 60
Método Basado en la Diferenciación 64
EFECTOS DE MÚLTIPLES OBSERVACIONES SOBRE LA ECUACIÓN DE REGRESIÓN 67
Método de la Eliminación de Múltiples Observaciones 68
Medidas Basadas en Residuales 68
Medidas Basadas en la Curva de Influencia 71
Medidas Basadas en el Volumen de Elipsoides de Confianza 75
Medidas Basadas en la Función de Verosimilitud 76
Medidas Basadas en un Subconjunto de Coeficientes de Regresión 78
ARTÍCULO
"Una modificación de la distancia de Cook" 79
LITERATURA CITADA 96

ÍNDICE DE FIGURAS

Figura No.		Página
1	Datos originales para el ajuste del *SCORE* y un gráfico Q-Q para los residuales en la regresión simple de *SCORE* sobre *AGE*	91
2	Identificación de la influencia y "outliers", basados en a) Distancia de Cook, b) Distancia de Cook Modificada de acuerdo con (5.20), c) Distancia de Draper y John. $n = 21$, $q = 2$, $p = 1$ y $s^2(i)$. La varianza de los residuales sin la i-ésima observación se usa en todos los casos, de acuerdo con la Nota 2	92
3	Identificación de "outliers" basada en la Distancia de Mahalanobis sobre la matriz de residuales, y detección de "outliers" basada en la Distancia de Cook Modificada dada en el Teorema 3	93

INTRODUCCIÓN

La *regresión* es una herramienta analítica inferencial que establece la relación funcional entre variables, y constituye uno de los métodos estadísticos más ampliamente utilizado. Al conjunto de técnicas que permiten construir y evaluar modelos que describan la relación entre variables, y formular inferencias basadas en los modelos obtenidos, se les conoce como *técnicas de regresión*; y al análisis estadístico que resulta de aplicarlas, se le denomina *análisis de regresión.*

Considerando el modelo de regresión general

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

donde:

$\boldsymbol{Y}$, es un vector de tamaño $n \times 1$ de variables dependientes (variables respuesta),

$\boldsymbol{X}$, es una matriz de tamaño $n \times k$ (con $n > k$) de variables independientes (variables predictoras o explicatorias),

$\boldsymbol{\beta}$, es un vector de tamaño $k \times 1$ de coeficientes o parámetros desconocidos, y

$\boldsymbol{\epsilon}$, es un vector de tamaño $n \times 1$ de errores aleatorios.

Si el modelo dado en la ecuación (1.1) es lineal, entonces se le llama *modelo de regresión lineal*, en el cual, si se involucra sólo una variable independiente, la regresión lineal se dice ser simple, pero si existen dos o más variables independientes en el modelo, entonces se dice ser múltiple. Uno de los propósitos del análisis de regresión lineal, es el de utilizar las variables explicatorias, cuyos valores son

conocidos, para predecir el comportamiento de la variable respuesta, aunque se puede decir que el objetivo general de esta técnica, es hacer mejores inferencias sobre la variable de mayor interés.

Los elementos que determinan una ecuación de regresión lineal, son las observaciones, las variables y los supuestos asociados al modelo, estos últimos se dan en la tercer sección del segundo capítulo. El ajuste de esta ecuación por el método de *cuadrados mínimos*, es el procedimiento de modelación que será utilizado.

En la práctica, es común encontrar algunos problemas relacionados con la información disponible (observaciones), tal como ocurre cuando se tiene la presencia de una o varias observaciones discrepantes (también llamados puntos anómalos), las cuales son inherentes a un conjunto de datos determinado. En este caso, el diagrama de dispersión de los datos, constituye un procedimiento informal que permite detectar con facilidad dichos puntos anómalos, pues aparecen en forma aislada, es decir, se alejan de la nube (o patrón) que siguen los demás puntos.

A las observaciones extremas, que individualmente o en conjunto, afectan la estimación mínimo cuadrática de los parámetros de la regresión, es decir, β y σ^2, al ajustar el modelo de regresión, se les llama *observaciones influyentes*, de las cuales, Chatterjee y Hadi (1988), distinguen tres tipos, a saber: *"outliers"*, *puntos palanca* y *puntos influyentes*, cuyas definiciones se exponen en la primera sección del tercer capítulo.

Se han desarrollado algunos métodos para detectar *observaciones influyentes*, entre los que se encuentran, el *método de la eliminación de observaciones* y el *método basado en la diferenciación*. El primero, estudia cómo el ajuste del modelo de regresión se ve perturbado al omitir una observación en particular, o bien, un conjunto de observaciones; mientras que en el segundo, la perturbación del ajuste es con respecto a ciertos parámetros del modelo.

Dentro del gran número de medidas de diagnóstico correspondientes al primer método, cabe destacar aquellas que miden la influencia de la i-ésima observación (potencialmente influyente), en la estimación de $\boldsymbol{\beta}$ exclusivamente, cuando dicha observación es omitida. Esto es, si se suprime la i-ésima observación, y se compara el vector que no considera esta observación $\widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})}$, con el vector que considera el conjunto de datos completo $\widehat{\boldsymbol{\beta}}$, a través de su diferencia $(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})})$, se establece el impacto que tiene la observación i sobre la estimación.

En el diagnóstico del análisis de regresión lineal, comúnmente, se utilizan medidas basadas en residuales, en las que la versión estudentizada, juega un papel fundamental, tal como lo sugieren Atkinson (1981 y 1982) y Velleman y Welsch (1981); estas medidas se describen en la primera sección del tercer capítulo.

Por su parte Hampel (1968 y 1974), propone las medidas basadas en la curva de influencia para $\widehat{\boldsymbol{\beta}}$, dentro de las cuales se encuentra la *distancia de Cook*, cuya definición requiere considerar, que bajo normalidad, la región de confianza $100(1-\alpha)\%$ para $\boldsymbol{\beta}$, se obtiene a partir de

$$\frac{(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})}{k\widehat{\sigma}^2} \leq \boldsymbol{F}(\alpha; k, n-k),$$

donde $\boldsymbol{F}(\alpha; k, n-k)$ es el punto superior α de la distribución $\boldsymbol{F}$ centrada con k y $(n-k)$ grados de libertad. Esta desigualdad define una región elipsoidal centrada en $\widehat{\boldsymbol{\beta}}$, y permite medir la influencia de la i-ésima observación por el cambio en el centro del elipsoide de confianza, cuando la i-ésima observación es omitida. En forma análoga, Cook (1977), propone la medida

$$\boldsymbol{C_i} = \frac{(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})})^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})})}{k\widehat{\sigma}^2}, \quad i = 1, 2, ..., n,$$

para considerar la influencia de la i-ésima observación sobre el centro del elipsoide de confianza, es decir, sobre $\widehat{\boldsymbol{\beta}}$. Asimismo, Welsch y Kuh (1977), presentan la distancia de Welsch y la de Welsch-Kuh, mientras que Atkinson (1981), realiza una

modificación de la distancia de Cook. Estas medidas se muestran en la primera sección del tercer capítulo.

Una alternativa a las medidas anteriores, la constituyen, medidas de la influencia de la i-ésima observación, que se basan en el cambio en el volumen del elipsoide de confianza para $\boldsymbol{\beta}$, cuando esta observacón es omitida, las cuales se muestran en la primera sección del tercer capítulo, y son: el estadístico Andrews-Pregibon sugerido por Andrews y Pregibon (1978), la relación de varianzas señalada en Belsley *et al.* (1980) y el estadístico Cook-Weisberg propuesto por Cook y Weisberg (1980).

Se concluye esta sección con las definiciones de las medidas basadas en la función de verosimilitud propuestas por Cook y Weisberg (1982).

La mayor parte de las aportaciones son recopiladas en Chatterjee y Hadi (1988) y en Belsley *et al.* (1980), en donde se observa que para la mayoría de las medidas de diagnóstico, el estadístico de prueba no tiene una distribución exacta, que mediante una regla de decisión permita establecer el punto crítico (o punto de corte) que facilite la detección de observaciones influyentes.

En el presente trabajo se plantean las bases y la importancia del análisis de sensibilidad en regresión. Como tema central, se propone una modificación de la distancia de Cook, basándose en la distancia de Mahalanobis generalizada, en el contexto de la regresión lineal con distribución normal multivariada. Se establecen además, las distribuciones exactas de los estadísticos basados en esta distancia modificada, a través de la cual, se obtiene un punto crítico que permita detectar una o varias observaciones influyentes.

En el segundo capítulo, se da la notación, algunos resultados y supuestos asociados con el modelo dado en la ecuación (1.1), que son requeridos en los capítulos posteriores; asimismo, se describen las propiedades de las matrices de predicción $\boldsymbol{P}$

y de residuales $(\boldsymbol{I} - \boldsymbol{P})$, y se mencionan los efectos sobre $\boldsymbol{P}$, que resultan al omitir una o varias observaciones, y por último, se muestra la distribución que siguen los elementos de la diagonal de la matriz de predicción, es decir, los $\boldsymbol{p_{ii}}$.

Las medidas de diagnóstico, correspondientes al método de la eliminación de una observación, y el método basado en la diferenciación son descritas en el tercer capítulo.

Por su parte, en el cuarto capítulo, se generaliza algunas de las medidas descritas en el tercer capítulo, al caso de la influencia de múltiples observaciones sobre el ajuste de la ecuación de regresión, limitándose sólo a aquellas que corresponden al método de la eliminación de múltiples observaciones.

En el quinto capítulo, se presenta el artículo **"Una modificación de la distancia de Cook"**, donde considerando el modelo de regresión lineal multivariado con distribución normal, se propone una modificación de la distancia de Cook para el caso de una observación, basándose en la distancia generalizada de Mahalanobis, vea el Teorema 1. En el Teorema 2 se define cómo calcular el cuadrado de la distancia de Cook modicada para detectar k observaciones influyentes. En los teoremas 3 y 4 se encuentran las distribuciones de las distancias de Cook modificadas propuestas en los teoremas 1 y 2, respectivamente. Se concluye el artículo con una aplicación de las técnicas descritas a un ejemplo de regresión lineal simple y otro de regresión lineal múltiple multivariado.

LA MATRIZ DE PREDICCIÓN

Notación

A continuación se propone la notación que será utilizada durante el desarrollo del presente trabajo.

En primera instancia, se define $\boldsymbol{I}$ como la matriz identidad, $\mathbf{0}$ como la matriz nula y $\mathbf{1}$ como el vector de unos.

Sea $\boldsymbol{u_i}$, el i-ésimo vector unitario, es decir, un vector con uno en la posición i y ceros en el resto de las posiciones.

Asimismo, se utiliza $\boldsymbol{M^T}$, $\boldsymbol{M^{-1}}$, $\boldsymbol{M^{-T}}$, para denotar la transpuesta, la inversa y la inversa de la transpuesta de la matriz $\boldsymbol{M}$, respectivamente. Mientras que el rango, la traza, el determinante y la norma de la matriz $\boldsymbol{M}$ son dados por $\boldsymbol{r(M)}$, $\boldsymbol{tr(M)}$, $\boldsymbol{det(M)}$ y $\|\boldsymbol{M}\|$, respectivamente.

Sean $\boldsymbol{y_i}$, $i = 1, 2, ..., n$, el i-ésimo elemento de $\boldsymbol{Y}$; $\boldsymbol{x_i^T}$, la i-ésima fila de $\boldsymbol{X}$; y sea $\boldsymbol{X_j}$, $j = 1, 2, ..., k$, la j-ésima columna de $\boldsymbol{X}$.

Cuando se hace referencia a la i-ésima observación, se considera el vector fila $(\boldsymbol{x_i} : \boldsymbol{y_i})$, esto es, la i-ésima fila de la matriz aumentada $\boldsymbol{Z} = (\boldsymbol{X} : \boldsymbol{Y})$.

La matriz $\boldsymbol{X}$ con la i-ésima fila eliminada se denota por $\boldsymbol{X_{(i)}}$. Del mismo modo, $\boldsymbol{\widehat{\beta}_{(i)}}$ es conocido como el vector de parámetros estimado cuando la i-ésima observación ha sido eliminada.

Se define a $\widehat{\epsilon}$ o $\widehat{\epsilon}_{Y \cdot X}$, como el vector de residuales, cuando se hace una regresión de la variable $\boldsymbol{Y}$ con la(s) variable(s) $\boldsymbol{X}$.

Las expresiones $\boldsymbol{E}(\cdot)$, $\boldsymbol{Var}(\cdot)$, $\boldsymbol{Cov}(\cdot,\cdot)$ y $\boldsymbol{Corr}(\cdot,\cdot)$, son usadas para definir el valor esperado, la varianza, la covarianza y el coeficiente de correlación de la(s) variable(s) aleatoria(s) que se indican por puntos en los paréntesis, respectivamente.

Resultados

Algunos resultados importantes que serán requeridos, son dados a continuación.

1. El método de estimación más comúnmente utilizado es el de Cuadrados Mínimos, el cual requiere minimizar la expresión

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\boldsymbol{T}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}),$$

para lo cual se deriva con respecto a $\boldsymbol{\beta}$, y resulta el sistema de ecuaciones normales

$$(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y},$$

el cual tiene solución única, si y solo si $(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}$ existe, dando lugar al estimador de $\boldsymbol{\beta}$, es decir,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y},$$

el cual tiene las siguientes propiedades:

(a) $\widehat{\boldsymbol{\beta}}$ es un estimador insesgado para $\boldsymbol{\beta}$, esto es, $\boldsymbol{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$,

(b) $\widehat{\boldsymbol{\beta}}$ es el mejor estimador linealmente insesgado para $\boldsymbol{\beta}$, lo cual indica que la varianza del estimador, es decir, $\boldsymbol{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}$ es la más pequeña, de entre la clase de estimadores linealmente insesgados, y

(c) Asumiendo que los errores siguen una distribución normal, como se verá en la siguiente sección, por (a) y (b), se tiene que

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_k\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\right).$$

2. El vector de valores ajustados (predichos), es decir, $\widehat{\boldsymbol{Y}}$, se obtiene así,

$$\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{Y}, \tag{2.1}$$

donde

$$\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}, \tag{2.2}$$

este vector $\widehat{\boldsymbol{Y}}$, cuenta con las siguientes propiedades:

(a) $\boldsymbol{E}(\widehat{\boldsymbol{Y}}) = \boldsymbol{X}\boldsymbol{\beta}$,

(b) $\boldsymbol{Var}(\widehat{\boldsymbol{Y}}) = \sigma^2\boldsymbol{P}$, y

(c) Bajo normalidad, por (a) y (b), se tiene que

$$\widehat{\boldsymbol{Y}} \sim \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{P}\right),$$

la cual es una distribución singular de rango k, pues $\boldsymbol{P}$ es singular.

3. El vector de residuales $\widehat{\boldsymbol{\epsilon}}$, se calcula así,

$$\widehat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{P}\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y},$$

y presenta las siguientes propiedades:

(a) $\boldsymbol{E}(\widehat{\boldsymbol{\epsilon}}) = \boldsymbol{0}$,

(b) $\boldsymbol{Var}(\widehat{\boldsymbol{\epsilon}}) = \sigma^2(\boldsymbol{I} - \boldsymbol{P})$,

(c) Nuevamente asumiendo normalidad, por (a) y (b), se tiene que

$$\widehat{\epsilon} \sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2(\boldsymbol{I}-\boldsymbol{P})\right),$$

la cual es una distribución singular de rango $n-k$, pues $(\boldsymbol{I}-\boldsymbol{P})$ es singular, y

(d) $\frac{\widehat{\epsilon}^{\boldsymbol{T}}\widehat{\epsilon}}{\sigma^2} \sim \chi^2(n-k)$, donde $\widehat{\epsilon}^{\boldsymbol{T}}\widehat{\epsilon}$ es la suma de cuadrados de residuales.

Supuestos

El análisis de regresión general se basa en suposiciones, en seguida se muestran aquellas que serán de utilidad en el presente estudio.

1. El supuesto de linealidad implícito en el modelo, el cual establece que cada valor respuesta observado $\boldsymbol{y_i}$, puede ser escrito como una función lineal de la i- ésima fila de $\boldsymbol{X}$, o sea,

$$\boldsymbol{y_i} = \boldsymbol{x_i^T}\boldsymbol{\beta} + \boldsymbol{\epsilon_i}, \quad i = 1, 2, ..., n$$

2. El supuesto que permite encontrar un estimador único de $\boldsymbol{\beta}$, para lo cual es necesario que $(\boldsymbol{X^T X})^{-1}$ exista, o equivalentemente, que el $\boldsymbol{r(X)} = k$.

3. El supuesto distribucional, en el que se requiere que

   (a) $\boldsymbol{X}$ sea medida sin errores,

   (b) ϵ_i no dependa de $\boldsymbol{x_i^T}$, para $i = 1, 2, ..., n$, y

   (c) $\epsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\boldsymbol{I})$.

4. El supuesto que indica que todas las observaciones son igualmente confiables y deben influir de igual manera en la determinación de los estimadores de $\boldsymbol{\beta}$ y σ^2.

Las Matrices P y $(I - P)$

Recordando que la matriz $\boldsymbol{P}$ dada en la ecuación (2.2), juega un papel muy importante en el análisis de regresión lineal, ya que es necesaria para obtener los estimadores de los parámetros incluídos en el modelo. Chatterjee y Hadi (1986), le llaman *matriz de predicción*, porque es la transformación matricial que aplicada a $\boldsymbol{Y}$, produce el vector de los valores predichos, es decir, $\widehat{\boldsymbol{Y}}$. Similarmente, a $(\boldsymbol{I} - \boldsymbol{P})$ se le denomina *matriz de residuales*, ya que al aplicarse a $\boldsymbol{Y}$, genera el vector de residuales estimado, esto es, $\widehat{\epsilon}$.

En cuanto a notación se refiere, los elementos de $\boldsymbol{P}$ se definen de la siguiente manera,

1. $\boldsymbol{p_{ij}}$, $i, j = 1, 2, ..., n$ es el ij-ésimo elemento de $\boldsymbol{P}$, y se obtiene así,

$$\boldsymbol{u_i^T P u_j = u_i^T X (X^T X)^{-1} X^T u_j = x_i^T (X^T X)^{-1} x_j = p_{ij}},$$

2. $\boldsymbol{p_{ii}}$, $i = 1, 2, ..., n$ es el i-ésimo elemento de la diagonal de $\boldsymbol{P}$, y similarmente, se encuentra así,

$$\boldsymbol{u_i^T P u_i = u_i^T X (X^T X)^{-1} X^T u_i = x_i^T (X^T X)^{-1} x_i = p_{ii}},$$

Algunos resultados donde están involucrados los elementos de $\boldsymbol{P}$, son dados a continuación.

1. $\boldsymbol{Var}(\widehat{y_i})=\sigma^2 \boldsymbol{p_{ii}}$,

2. $\boldsymbol{Var}(\widehat{\epsilon_i})=\sigma^2(1 - \boldsymbol{p_{ii}})$,

3. $\boldsymbol{Cov}(\widehat{\epsilon_i}, \widehat{\epsilon_j})=-\sigma^2 \boldsymbol{p_{ij}}$, y

4. $\boldsymbol{Corr}(\widehat{\epsilon_i}, \widehat{\epsilon_j}) = \dfrac{-\boldsymbol{p_{ij}}}{\sqrt{(1 - \boldsymbol{p_{ii}})}\sqrt{(1 - \boldsymbol{p_{jj}})}}$,

notando que el coeficiente de correlación es determinado totalmente por los elementos de $\boldsymbol{P}$.

Una característica importante de las matrices de predicción y de residuales, es que son simétricas e idempotentes, es decir, son proyecciones ortogonales. $\boldsymbol{P}$ proyecta ortogonalmente cualquier vector n-dimensional en el espacio k-dimensional de las columnas de $\boldsymbol{X}$, esto es, $\boldsymbol{PX} = \boldsymbol{X}$, por lo que se deduce que $(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X} = \boldsymbol{0}$, y el vector de residuales estimado se puede expresar como

$$\widehat{\epsilon} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{P})(\boldsymbol{X}\boldsymbol{\beta} + \epsilon) = (\boldsymbol{I} - \boldsymbol{P})\epsilon,$$

en donde la relación entre $\widehat{\epsilon}$ y ϵ depende sólo de $\boldsymbol{P}$.

Ahora, si se considera la igualdad dada en la ecuación (2.1), se observa que el i-ésimo valor predicho $\widehat{\boldsymbol{y}}_{\boldsymbol{i}}$, se puede calcular de la siguiente manera,

$$\widehat{\boldsymbol{y}}_{\boldsymbol{i}} = \sum_{j=1}^{n} \boldsymbol{p}_{\boldsymbol{ij}}\boldsymbol{y}_{\boldsymbol{j}} = \boldsymbol{p}_{\boldsymbol{ii}}\boldsymbol{y}_{\boldsymbol{i}} + \sum_{j \neq i} \boldsymbol{p}_{\boldsymbol{ij}}\boldsymbol{y}_{\boldsymbol{j}}, \qquad i = 1, 2, ..., n,$$

de donde se sigue que,

$$\frac{\partial \widehat{\boldsymbol{y}}_{\boldsymbol{i}}}{\partial \boldsymbol{y}_{\boldsymbol{i}}} = \boldsymbol{p}_{\boldsymbol{ii}}, \qquad i = 1, 2, ..., n$$

Por consiguiente, $\boldsymbol{p}_{\boldsymbol{ii}}$ puede interpretarse como el valor *"palanca"* que cada $\boldsymbol{y}_{\boldsymbol{i}}$ tiene en determinado $\widehat{\boldsymbol{y}}_{\boldsymbol{i}}$ (Hoaglin y Welsch, 1978). Similarmente, $\boldsymbol{p}_{\boldsymbol{ij}}$ puede ser interpretado como el valor *"palanca"* que cada $\boldsymbol{y}_{\boldsymbol{j}}$ tiene en determinado $\widehat{\boldsymbol{y}}_{\boldsymbol{i}}$.

Por otro lado, el recíproco de $\boldsymbol{p}_{\boldsymbol{ii}}$ se puede considerar como el número de observaciones que determinan $\widehat{\boldsymbol{y}}_{\boldsymbol{i}}$, Huber (1977 y 1981); esto es, si $\boldsymbol{p}_{\boldsymbol{ii}} = 1$, $\widehat{\boldsymbol{y}}_{\boldsymbol{i}}$ está determinado por una sola observación (solamente por $\boldsymbol{y}_{\boldsymbol{i}}$), por lo tanto se requiere un grado de libertad para ajustar $\boldsymbol{y}_{\boldsymbol{i}}$, Belsley *et al.* (1980); si $\boldsymbol{p}_{\boldsymbol{ii}} = 0$, $\boldsymbol{y}_{\boldsymbol{i}}$ no tiene influencia sobre $\widehat{\boldsymbol{y}}_{\boldsymbol{i}}$; y si $\boldsymbol{p}_{\boldsymbol{ii}} = 0.5$, $\widehat{\boldsymbol{y}}_{\boldsymbol{i}}$ está determinado por dos observaciones. Una amplia variación en los valores que toma $\boldsymbol{p}_{\boldsymbol{ii}}$, indican un espaciamiento no homogéneo en las filas de $\boldsymbol{X}$ (Behnken y Draper, 1972).

Se observa que los elementos de $\boldsymbol{P}$ tienen una interesante interpretación geométrica, esto es,

1. Cuando $\boldsymbol{X}$ contiene una columna constante o cuando las columnas de $\boldsymbol{X}$ son centradas, la forma cuadrática $\boldsymbol{V^T(X^TX)^{-1}V} = c$, donde $\boldsymbol{V}$ es un vector $k \times 1$ y c es una constante, define contornos elípticos k-dimensionales centrados al promedio de las columnas de $\boldsymbol{X}$, es decir, $\bar{\boldsymbol{X}}$. El conjunto convexo más pequeño que contiene a los n puntos en el espacio $\boldsymbol{X}$, está contenido en los elipsoides que satisfacen la desigualdad $c \leq max(\boldsymbol{p_{ii}})$. En consecuencia, $\boldsymbol{p_{ii}}$ está determinado por la colocación de $\boldsymbol{x_i}$ en el espacio $\boldsymbol{X}$, es decir, un valor grande (o pequeño) de $\boldsymbol{p_{ii}}$, indica que $\boldsymbol{x_i}$ se encuentra lejos (o cerca) de la nube de puntos del conjunto de datos.

2. El volumen del elipsoide de confianza $100(1-\alpha)\%$ para $\boldsymbol{\beta}$ es monótono creciente con $\boldsymbol{p_{ii}}$, es decir, un $\boldsymbol{p_{ii}}$ grande, provoca un incremento en el elipsoide de confianza para $\boldsymbol{\beta}$ cuando la i-ésima observación es eliminada.

3. Considerando la regresión de $\boldsymbol{u_i}$ sobre $\boldsymbol{X}$. El vector de residuales para esta regresión es

$$\widehat{\boldsymbol{\epsilon}}_{u_i \cdot X} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{u_i},$$

el cual, simplemente es la i-ésima columna (fila) de $(\boldsymbol{I} - \boldsymbol{P})$. La norma del vector de residuales es

$$\|\widehat{\boldsymbol{\epsilon}}_{u_i \cdot X}\| = \sqrt{\boldsymbol{u_i^T}(\boldsymbol{I} - \boldsymbol{P})^{\boldsymbol{T}}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{u_i}} = \sqrt{\boldsymbol{u_i^T}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{u_i}} = \sqrt{1 - \boldsymbol{p_{ii}}}$$

El ángulo θ_i entre $\widehat{\boldsymbol{\epsilon}}_{Y \cdot X}$ y $\widehat{\boldsymbol{\epsilon}}_{u_i \cdot X}$ se encuentra de la siguiente forma,

$$\begin{aligned} cos\theta_i &= \frac{\widehat{\boldsymbol{\epsilon}}_{Y \cdot X}\widehat{\boldsymbol{\epsilon}}_{u_i \cdot X}}{\|\widehat{\boldsymbol{\epsilon}}_{Y \cdot X}\| \cdot \|\widehat{\boldsymbol{\epsilon}}_{u_i \cdot X}\|} \\ &= \frac{\boldsymbol{Y^T}(\boldsymbol{I} - \boldsymbol{P})^{\boldsymbol{T}}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{u_i}}{\sqrt{\boldsymbol{Y^T}(\boldsymbol{I} - \boldsymbol{P})^{\boldsymbol{T}}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}}\sqrt{1 - \boldsymbol{p_{ii}}}} \\ &= \frac{\boldsymbol{Y^T}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{u_i}}{\sqrt{\widehat{\boldsymbol{\epsilon}}^{\boldsymbol{T}}\widehat{\boldsymbol{\epsilon}}}\sqrt{1 - \boldsymbol{p_{ii}}}} \\ &= \frac{\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}}{\sqrt{\widehat{\boldsymbol{\epsilon}}^{\boldsymbol{T}}\widehat{\boldsymbol{\epsilon}}(1 - \boldsymbol{p_{ii}})}} \end{aligned}$$

De este modo, el i-ésimo residual se puede expresar como

$$\widehat{\epsilon}_{\boldsymbol{i}} = cos\theta_i \sqrt{\widehat{\epsilon}^{\,\boldsymbol{T}} \widehat{\epsilon}(1 - \boldsymbol{p_{ii}})}$$

Ellenberg (1973, 1976), ha demostrado que si $\boldsymbol{p_{ii}} \neq 1$, entonces

$$cos^2\theta_i \sim Beta\left(\frac{1}{2}, \frac{1}{2}(n-k-1)\right)$$

Por su parte, Beckman y Trussel (1974), han mostrado que si $\boldsymbol{p_{ii}} \neq 1$, entonces

$$cot\theta_i \sqrt{n-k-1} \sim \boldsymbol{t}(n-k-1)$$

Propiedades de la Matriz de Predicción

A continuación se muestran algunas propiedades de la matriz $\boldsymbol{P}$, las cuales se dan en los siguientes teoremas.

Teorema 1. *La matriz de predicción $\boldsymbol{P}$ es invariante bajo transformaciones lineales no singulares de la forma $\boldsymbol{X} \rightarrow \boldsymbol{XE}$, donde $\boldsymbol{E}$ es cualquier matriz de tamaño $k \times k$ no singular.*

Demostración. Sea $\boldsymbol{E}$ cualquier matriz de tamaño $k \times k$ no singular y sean las matrices de predicción para $\boldsymbol{X}$ y para $\boldsymbol{XE}$, denotadas por $\boldsymbol{P}_X$ y $\boldsymbol{P}_{XE}$, respectivamente, entonces sólo hay que mostrar que $\boldsymbol{P}_{XE}$=$\boldsymbol{P}_X$, para lo cual, se ve que

$$\boldsymbol{P}_{XE} = (\boldsymbol{XE})\left((\boldsymbol{XE})^{\boldsymbol{T}}(\boldsymbol{XE})\right)^{-1}(\boldsymbol{XE})^{\boldsymbol{T}} = \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}} = \boldsymbol{P}_X \blacksquare$$

Este teorema indica que las matrices de predicción para $\boldsymbol{X}$ y $\boldsymbol{XE}$ son las mismas, por lo tanto, cuando se hace una regresión, el vector de residuales estimado para $\boldsymbol{XE}$ es el mismo que para $\boldsymbol{X}$, es decir,

$$\widehat{\boldsymbol{\epsilon}}_{Y \cdot XE} = (\boldsymbol{I} - \boldsymbol{P}_{XE})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{Y} = \widehat{\boldsymbol{\epsilon}}_{YX}$$

De este modo, si el modelo de la ecuación (1.1) contiene un término constante, los residuales y sus varianzas estimadas son invariantes bajo transformaciones de parámetros de escala y de localización en $\boldsymbol{X}$; mientras que si el modelo no contiene un término constante, solo son invariantes bajo transformaciones de parámetro de escala en $\boldsymbol{X}$.

Este teorema, además, permite cualquier reparametrización no singular del modelo de la ecuación (1.1), es decir, si $\boldsymbol{\alpha}=\boldsymbol{E}^{-1}\boldsymbol{\beta}$, entonces los modelos

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon \qquad \text{y} \qquad \boldsymbol{Y} = \boldsymbol{X}\boldsymbol{E}\boldsymbol{\alpha} + \epsilon,$$

son equivalentes, en el sentido de que producen el mismo $\widehat{\boldsymbol{Y}}$.

Teorema 2. *$\boldsymbol{P}$ y $(\boldsymbol{I}-\boldsymbol{P})$ son matrices simétricas e idempotentes.*

Demostración. Para demostrar simetría, hay que ver que

$$\boldsymbol{P} = \boldsymbol{P}^{\boldsymbol{T}} \quad \text{y} \quad \text{que} \quad (\boldsymbol{I}-\boldsymbol{P}) = (\boldsymbol{I}-\boldsymbol{P})^{\boldsymbol{T}}$$

Mientras que para mostrar idempotencia, se ve que

$$\boldsymbol{P} = \boldsymbol{P}^{\boldsymbol{2}} \quad \text{y} \quad \text{que} \quad (\boldsymbol{I}-\boldsymbol{P}) = (\boldsymbol{I}-\boldsymbol{P})^{\boldsymbol{2}} \blacksquare$$

Teorema 3. *Sea $\boldsymbol{X}$ una matriz de tamaño $n \times k$, entonces*

1. $\boldsymbol{tr}(\boldsymbol{P})=\boldsymbol{r}(\boldsymbol{P})=k$,
2. $\boldsymbol{tr}(\boldsymbol{I}-\boldsymbol{P})=n-k$, y
3. $\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{p}_{\boldsymbol{ij}}^{\boldsymbol{2}}=k$.

Demostración.

1. Para toda matriz de proyección $\boldsymbol{P}$, la $\boldsymbol{tr}(\boldsymbol{P})=\boldsymbol{r}(\boldsymbol{P})=k$.
2. Se ve que la $\boldsymbol{tr}(\boldsymbol{I}-\boldsymbol{P})=\boldsymbol{tr}(\boldsymbol{I})$ - $\boldsymbol{tr}(\boldsymbol{P})=n-k$.

3. Si $\boldsymbol{P} = \boldsymbol{P}^{\boldsymbol{T}} = \boldsymbol{P}^{\boldsymbol{2}}$, entonces, la $\boldsymbol{tr(P)}$=$\boldsymbol{tr(P^2)}$=$\boldsymbol{tr(P^T P)}$=$\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}^2$=$k$. ■

Para la demostración del Teorema 4, se requiere citar el siguiente lema.

Lema 1. *La inversa de una matriz particionada.*

Se define

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A_{11}} & \boldsymbol{A_{12}} \\ \boldsymbol{A_{21}} & \boldsymbol{A_{22}} \end{bmatrix}$$

1. *Si* $\boldsymbol{A}$ y $\boldsymbol{A_{11}}$ *son no singulares, entonces:*

$$\boldsymbol{A^{-1}} = \begin{bmatrix} \boldsymbol{A_{11}^{-1} + A_{11}^{-1} A_{12} M A_{21} A_{11}^{-1}} & \boldsymbol{-A_{11}^{-1} A_{12} M} \\ \boldsymbol{-M A_{21} A_{11}^{-1}} & \boldsymbol{M} \end{bmatrix}$$

donde $\boldsymbol{M = (A_{22} - A_{21} A_{11}^{-1} A_{12})^{-1}}$

2. *Si* $\boldsymbol{A}$ y $\boldsymbol{A_{22}}$ *son no singulares, entonces:*

$$\boldsymbol{A^{-1}} = \begin{bmatrix} \boldsymbol{N} & \boldsymbol{-N A_{12} A_{22}^{-1}} \\ \boldsymbol{-A_{22}^{-1} A_{12} N} & \boldsymbol{A_{22}^{-1} + A_{22}^{-1} A_{21} N A_{12} A_{22}^{-1}} \end{bmatrix}$$

donde $\boldsymbol{N = (A_{11} - A_{12} A_{22}^{-1} A_{21})^{-1}}$

Demostración. Si se considera $\boldsymbol{B}$=$\boldsymbol{A^{-1}}$, entonces sólo se debe mostrar que $\boldsymbol{BA = I}$.

■

Teorema 4. *Sea* $\boldsymbol{X = (X_1 : X_2)}$, *donde* $\boldsymbol{X_1}$ *es una matriz de tamaño* $n \times r$ *de rango* r *y* $\boldsymbol{X_2}$ *es una matriz* $n \times (k-r)$ *de rango* $(k-r)$. *Sea* $\boldsymbol{P_1 = X_1 (X_1^T X_1)^{-1} X_1^T}$ *la matriz de predicción para* $\boldsymbol{X_1}$, *y* $\boldsymbol{W = (I - P_1) X_2}$ *es la proyección ortogonal de* $\boldsymbol{X_2}$ *sobre el complemento de* $\boldsymbol{X_1}$. *Por último, sea* $\boldsymbol{P_2 = W (W^T W)^{-1} W^T}$ *la matriz de predicción para* $\boldsymbol{W}$; *entonces,* $\boldsymbol{P}$ *puede ser expresada como*

$$\boldsymbol{P = P_1 + P_2}$$

$$\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T + (\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{R}_2(\boldsymbol{I}-\boldsymbol{P}_1)$$

donde $\boldsymbol{R}_2 = \boldsymbol{X}_2\left(\boldsymbol{X}_2^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2^T$.

Demostración. Se ve que $\boldsymbol{P}$ se puede calcular así,

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_1^T\boldsymbol{X}_1 & \boldsymbol{X}_1^T\boldsymbol{X}_2 \\ \boldsymbol{X}_2^T\boldsymbol{X}_1 & \boldsymbol{X}_2^T\boldsymbol{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}_1^T \\ \boldsymbol{X}_2^T \end{bmatrix}$$

Al invertir la matriz particionada que se encuentra en la parte central de la expresión anterior, utilizando para ello el Lema 1, se llega a que $\boldsymbol{P} = \boldsymbol{P}_1 + \boldsymbol{P}_2$, quedando probado el teorema. ∎

Este teorema permite además, particionar la matriz $\boldsymbol{P}$ en dos o más matrices de predicción.

Teorema 5. *Para $i = 1, 2, ..., n$ y $j = 1, 2, ..., n$, se tiene que*

1. $0 \leq \boldsymbol{p_{ii}} \leq 1$, *para todo* i,
2. $0.5 \leq \boldsymbol{p_{ij}} \leq 0.5$, *para todo* $j \neq i$,
3. *Si* $\boldsymbol{X}$ *contiene una columna constante, entonces*
   (a) $\boldsymbol{p_{ii}} \geq n^{-1}$, *para todo* i,
   (b) $\boldsymbol{P1} = \boldsymbol{1}$, *y*
4. *Suponga que* $\boldsymbol{x_i}$ *ocurre* a *veces y que* $-\boldsymbol{x_i}$ *ocurre* b *veces, entonces*

$$\boldsymbol{p_{ii}} \leq (a+b)^{-1}.$$

Demostración. Para el inciso 3, se tiene

(a) Si $\boldsymbol{X}$ contiene una columna constante, entonces se puede definir como

$$\boldsymbol{X} = (\boldsymbol{1} : \boldsymbol{X_2}),$$

por el Teorema 4, se tiene que

$$\boldsymbol{P_1} = 1(1^T 1)^{-1} 1^T = 1n^{-1}1^T = n^{-1}11^T$$

$$\boldsymbol{W} = (\boldsymbol{I} - \boldsymbol{P_1})\boldsymbol{X_2} = (\boldsymbol{I} - n^{-1}11^T)\boldsymbol{X_2} \equiv \tilde{\boldsymbol{X}}$$

$$\boldsymbol{P_2} = \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T$$

De este modo, la matriz de predicción $\boldsymbol{P}$ se puede expresar como

$$\boldsymbol{P} = \boldsymbol{P_1} + \boldsymbol{P_2} = n^{-1}11^T + \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T$$

Se ve que cada uno de los elementos de la diagonal de $\boldsymbol{P_1}$ es igual a n^{-1}. Como $\boldsymbol{P_2}$ es una matriz de predicción, por el Teorema 5(1), se deduce que los elementos de la diagonal son no negativos, por lo que $\boldsymbol{p_{ii}} \geq n^{-1}$, para todo i.

(b) Se ve que $\tilde{\boldsymbol{X}}^T\mathbf{1} = \mathbf{0}$, y que $\boldsymbol{P_2}\mathbf{1} = \mathbf{0}$, entonces $\boldsymbol{P}\mathbf{1} = \boldsymbol{P_1}\mathbf{1} = \mathbf{1}$.

La demostración completa aparece en Chatterjee y Hadi (1988). ■

Teorema 6. *Para $i = 1, 2, ..., n$ y $j = 1, 2, ..., n$, se tiene que*

1. *Si* $\boldsymbol{p_{ii}} = 1 \quad o \quad 0$, *entonces* $\boldsymbol{p_{ij}} = 0$,
2. $(1 - \boldsymbol{p_{ii}})(1 - \boldsymbol{p_{jj}}) - \boldsymbol{p_{ij}^2} \geq 0$,
3. $\boldsymbol{p_{ii}p_{jj}} - \boldsymbol{p_{ij}^2} \geq 0$, *y*
4. $\boldsymbol{p_{ii}} + \dfrac{\widehat{\epsilon_i}^2}{\widehat{\epsilon}^T\widehat{\epsilon}} \leq 1$.

Demostración.

1. Se observa que el i-ésimo elemento diagonal de $\boldsymbol{P}$, puede ser escrito como

$$\boldsymbol{p_{ii}} = \sum_{j=1}^{n} \boldsymbol{p_{ij}^2} = \boldsymbol{p_{ii}^2} + \sum_{j \neq i}^{n} \boldsymbol{p_{ij}^2},$$

   por lo que, si $\boldsymbol{p_{ii}} = 0$, entonces $\boldsymbol{p_{ij}} = 0$; y si $\boldsymbol{p_{ii}} = 1$, entonces $\boldsymbol{p_{ij}} = 0$.

2. Se puede ver en Chatterjee y Hadi (1988).

3. Simplificando la expresión del Teorema 6(2), se comprueba que

$$p_{ii}p_{jj} - p_{ij}^2 \geq 0.$$

4. Se define $Z = (X : Y)$, $P_X = X(X^TX)^{-1}X^T$ y $P_Z = Z(Z^TZ)^{-1}Z^T$. Utilizando el Teorema 4, se ve que

$$P_Z = P_X + \frac{(I - P_X)YY^T(I - P_X)}{Y^T(I - P_X)Y} = P_X + \frac{\widehat{\epsilon}\widehat{\epsilon}^T}{\widehat{\epsilon}^T\widehat{\epsilon}}$$

De esta manera, los elementos de la diagonal de P_Z son menores o iguales a uno, tomando el i-ésimo elemento, resulta

$$p_{ii} + \frac{\widehat{\epsilon_i}^2}{\widehat{\epsilon}^T\widehat{\epsilon}} \leq 1 \blacksquare$$

El Teorema 6(1), 6(2) y 6(3), indica que si p_{ii} es grande (cercano a 1), o pequeño (cercano a 0), entonces p_{ij} es pequeño para toda $j \neq i$. El Teorema 6(4) indica que para p_{ii} grande, el i-ésimo residual estimado $\widehat{\epsilon_i}$ es pequeño.

Omitiendo o Adicionando una Observación

Para determinar los efectos que resultan de omitir o adicionar observaciones en P, son necesarios los siguientes resultados.

Lema 2. *Sean A y D matrices no singulares de órdenes $k \times k$ y $m \times m$, respectivamente. Sean además B y C, ambas de orden $k \times m$; entonces, existe la siguiente inversa,*

$$(A + BDC^T)^{-1} = A^{-1} - A^{-1}B(D^{-1} + C^TA^{-1}B)^{-1}C^TA^{-1}$$

Demostración. Sólo debe verse que

$$(A + BDC^T)\left(A^{-1} - A^{-1}B(D^{-1} + C^TA^{-1}B)^{-1}C^TA^{-1}\right) = I \blacksquare$$

Un caso para el que la inversa de esta matriz no existe, se tiene cuando $\boldsymbol{A = I}$, $\boldsymbol{B = X}$, $\boldsymbol{C^T = X^T}$ y $\boldsymbol{D = -(X^T X)^{-1}}$, así, $\boldsymbol{(A + BDC^T) = (I - P)}$, ésta última, como se sabe, es singular.

Este lema, también es conocido como el Teorema de Sherman-Morrison-Woodbury (Henderson y Searle, 1981).

Sustituyendo $\boldsymbol{B = x_i}$, $\boldsymbol{C^T = x_i^T}$, $\boldsymbol{D = 1}$ y $\boldsymbol{A = (X_{(i)}^T X_{(i)})}$, en la expresión del Lema 2, se obtiene

$$\begin{aligned}(X^T X)^{-1} &= (X_{(i)}^T X_{(i)} + x_i x_i^T)^{-1} \\ &= (X_{(i)}^T X_{(i)})^{-1} - \frac{(X_{(i)}^T X_{(i)})^{-1} x_i x_i^T (X_{(i)}^T X_{(i)})^{-1}}{1 + x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i} \end{aligned} \tag{2.3}$$

Sustituyendo ahora, $\boldsymbol{B = x_i}$, $\boldsymbol{C^T = x_i^T}$, $\boldsymbol{D = -1}$ y $\boldsymbol{A = X^T X}$, resulta

$$\begin{aligned}(X_{(i)}^T X_{(i)})^{-1} &= (X^T X - x_i x_i^T)^{-1} \\ &= (X^T X)^{-1} + \frac{(X^T X)^{-1} x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i} \end{aligned} \tag{2.4}$$

Para ver los efectos sobre la matriz de predicción, al adicionar una observación $\boldsymbol{x_i^T}$, se utiliza la ecuación (2.3). A continuación se presentan algunos ejemplos.

1. Si $\boldsymbol{p_{rc} = x_r^T (X^T X)^{-1} x_c}$, entonces

$$\boldsymbol{p_{rc} = p_{rc(i)} - \frac{p_{ri(i)} p_{ic(i)}}{1 + p_{ii(i)}}}, \quad \text{para} \quad r, c \neq i,$$

donde $\boldsymbol{p_{rc(i)} = x_r^T (X_{(i)}^T X_{(i)})^{-1} x_c}$.

2. Si $\boldsymbol{p_{ri} = p_{ir} = x_r^T (X^T X)^{-1} x_i}$, entonces

$$\boldsymbol{p_{ir} = p_{ir(i)} - \frac{p_{ii(i)} p_{ir(i)}}{1 + p_{ii(i)}} = \frac{p_{ir(i)}}{1 + p_{ii(i)}}}, \quad \text{para} \quad r \neq i,$$

donde $\boldsymbol{p_{ir(i)} = x_r^T (X_{(i)}^T X_{(i)})^{-1} x_i}$.

3. Si $p_{ii}=x_i^T(X^TX)^{-1}x_i$, entonces

$$p_{ii} = p_{ii(i)} - \frac{p_{ii(i)}^2}{1+p_{ii(i)}} = \frac{p_{ii(i)}}{1+p_{ii(i)}},$$

donde $p_{ii(i)}=x_i^T(X_{(i)}^TX_{(i)})^{-1}x_i$.

Por otra parte, para ver los efectos sobre la matriz de predicción, al omitir una observación x_i^T, se usa la ecuación (2.4). Observando por ejemplo lo que sucede con $p_{cr(i)}$, esto es,

$$p_{cr(i)} = p_{cr} + \frac{p_{ci}p_{ir}}{1-p_{ii}}, \quad \text{para} \quad r,c \neq i$$

y cuando $r = c$, entonces la expresión anterior se simplifica, resultando

$$p_{rr(i)} = p_{rr} + \frac{p_{ir}^2}{1-p_{ii}}.$$

Cuando la i-ésima observación es omitida, la anterior expresión muestra que para $r \neq i$, $p_{rr(i)}$ es grande, si p_{ii}, p_{rr} y/o $|p_{ir}|$ son grandes. Ahora, cuando la i-ésima y r-ésima filas de X son iguales, entonces $p_{rr(i)}$ se reduce a

$$p_{rr(i)} = \frac{p_{ii}}{1-p_{ii}},$$

observando que si p_{ii} está próximo a 0.5, entonces $p_{rr(i)}$ está próximo a 1, en tal caso la i-ésima (r-ésima) observación puede no ser detectada, cuando ambas observaciones están presentes.

Notando que, implícitamente se supone que existe la inversa de $(X_{(i)}^TX_{(i)})$, o equivalentemente, el $r(X_{(i)})=k$. No existe $(X_{(i)}^TX_{(i)})^{-1}$, si la eliminación de la i-ésima observación resulta en un modelo de rango deficiente.

Condiciones para Valores Grandes de p_{ii}

Para el caso de una regresión simple de la variable $\boldsymbol{Y}$ con $\boldsymbol{X}$, se tiene

$$\boldsymbol{P} = \frac{\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}}{\sum_{r=1}^{n} \boldsymbol{x}_{\boldsymbol{r}}^{\boldsymbol{2}}},$$

de donde se deduce que

$$\boldsymbol{p}_{\boldsymbol{ii}} = \frac{\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{2}}}{\sum_{r=1}^{n} \boldsymbol{x}_{\boldsymbol{r}}^{\boldsymbol{2}}}, \quad i = 1, 2, ..., n$$

Si un término constante es incluído en $\boldsymbol{X}$, y $\bar{\boldsymbol{X}}$ denota el promedio de $\boldsymbol{X}$, entonces utilizando la prueba del Teorema 5(3(a)), $\boldsymbol{p}_{\boldsymbol{ii}}$ se puede escribir como

$$\boldsymbol{p}_{\boldsymbol{ii}} = \frac{1}{n} + \frac{(\boldsymbol{x}_{\boldsymbol{i}} - \bar{\boldsymbol{X}})^{\boldsymbol{2}}}{\sum_{r=1}^{n} (\boldsymbol{x}_{\boldsymbol{r}} - \bar{\boldsymbol{X}})^{\boldsymbol{2}}}, \quad i = 1, 2, ..., n$$

De este modo, $\boldsymbol{p}_{\boldsymbol{ii}}$ será grande, si $\boldsymbol{x}_{\boldsymbol{i}}$ se encuentra alejado de la nube de puntos de un determinado conjunto de datos.

En el caso de una regresión múltiple, considerando λ_r y $\boldsymbol{V}_{\boldsymbol{r}}$, $r = 1, 2, ..., k$, los eigenvalores y sus correspondientes eigenvectores normalizados de $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$, respectivamente. Si θ_{ir} es el ángulo entre $\boldsymbol{x}_{\boldsymbol{i}}$ y $\boldsymbol{V}_{\boldsymbol{r}}$, entonces

$$\boldsymbol{p}_{\boldsymbol{ij}} = \|\boldsymbol{x}_{\boldsymbol{i}}\| \cdot \|\boldsymbol{x}_{\boldsymbol{i}}\| \sum_{r=1}^{k} \lambda_r{}^{-1} cos^2\theta_{ir},$$

y así,

$$\boldsymbol{p}_{\boldsymbol{ii}} = \boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{T}} \boldsymbol{x}_{\boldsymbol{i}} \sum_{r=1}^{k} \lambda_r{}^{-1} cos^2\theta_{ir}.$$

Notando que si $\boldsymbol{x}_{\boldsymbol{i}}$ y $\boldsymbol{V}_{\boldsymbol{r}}$, tienen la misma dirección, entonces el $cos^2\theta_{ir}$=1, y si son perpendiculares, entonces $cos^2\theta_{ir}$=0. De igual forma, $\boldsymbol{p}_{\boldsymbol{ii}}$ será grande si $\boldsymbol{x}_{\boldsymbol{i}}$ se encuentra alejado de la nube de puntos de un determinado conjunto de datos (Cook y Weisberg, 1982).

Omitiendo Múltiples Filas de X

Los efectos sobre P, que se tienen al omitir un subconjunto de filas de X, se muestran en el siguiente teorema.

Teorema 7. *Suponiendo que X es de orden $n \times k$ de rango k y que existen $m < k$ elementos de la diagonal de P iguales a uno, y sea I=$\{i : p_{ii} = 1, i = 1, 2, ..., n\}$ el conjunto de sus índices; entonces, para cualquier $J \supseteq I$, el $r(X_{(J)}) \leq k - m$, con la igualdad sólo si $J = I$.*

Demostración. Ajustando las filas de X, de tal forma que las m filas indexadas por I, sean las últimas m filas de X, entonces X se puede escribir así,

$$X = \begin{bmatrix} X_{(I)} \\ X_I^T \end{bmatrix} \begin{matrix} (n-m) \times k \\ m \times k \end{matrix}$$

Se define a P_I como el menor principal de P correspondiente a X_I^T. Y como p_{ii}=1, $i \in I$, entonces P_I=I, donde I es de dimensión $m \times m$. Utilizando el Teorema 6(1), P puede expresarse como

$$P = \begin{bmatrix} X_{(I)}(X^TX)^{-1}X_{(I)}^T & 0 \\ 0 & I \end{bmatrix},$$

en donde puede observarse que $\left(X_{(I)}(X^TX)^{-1}X_{(I)}^T\right)$ es idempotente con rango $(k - m)$, por lo tanto,

$$(k - m) = r\left(X_{(I)}(X^TX)^{-1}X_{(I)}^T\right) = r(X_{(I)}),$$

de donde se sigue que, para cualquier $J \supseteq I$, el

$$r(X_{(J)}) \leq r(X_{(I)}) = (k - m) \blacksquare$$

Beckman y Trussel (1974), señalan que si se emplean las propiedades de la inversa generalizada para X, bajo las condiciones del modelo de la ecuación (1.1),

entonces si $\boldsymbol{p_{ii}}$=1, el $\boldsymbol{r}(\boldsymbol{X}_{(i)})$=$(k-1)$, y por lo tanto $(\boldsymbol{X}_{(i)}^{\boldsymbol{T}}\boldsymbol{X}_{(i)})^{-1}$ no existe. Esto representa un caso especial de este teorema, donde $m = 1$, de manera que si I={i} y $\boldsymbol{p_{ii}}$=1, entonces se tiene que

$$\boldsymbol{r}(\boldsymbol{X}_{(J)}) \leq \boldsymbol{r}(\boldsymbol{X}_{(i)}) = (k-1)$$

Notando que este teorema establece que si para algún $i \in I$, $\boldsymbol{p_{ii}}$=1, entonces el

$$\boldsymbol{r}(\boldsymbol{X}_{(I)}) \leq k.$$

Eigenvalores de $\boldsymbol{P}$ y $(\boldsymbol{I}-\boldsymbol{P})$

Las propiedades de los eigenvalores de $\boldsymbol{P}$ y $(\boldsymbol{I}-\boldsymbol{P})$, son dadas en los tres teoremas siguientes.

Teorema 8. *Los eigenvalores de $\boldsymbol{P}$ y $(\boldsymbol{I}-\boldsymbol{P})$ son 0's o 1's.*

Demostración. La demostración se sigue directamente del hecho de que los eigenvalores de matrices idempotentes son 0's o 1's.

■

Teorema 9. *Existen $(n-k)$ eigenvalores de $\boldsymbol{P}$ iguales a 0, y los restantes k eigenvalores son iguales a 1. Similarmente, k eigenvalores de $(\boldsymbol{I}-\boldsymbol{P})$ son iguales a 0 y $(n-k)$ eigenvalores son iguales a 1.*

Demostración. Sean λ_i, $i = 1, 2, ..., n$, los eigenvalores de $\boldsymbol{P}$. Usando el Teorema 3(1), se tiene

$$\sum_{i=1}^{n} \lambda_i = \boldsymbol{tr}(\boldsymbol{P}) = k,$$

pero el Teorema 8 afirma que λ_i=0 o 1 para todo i, por lo tanto, k de los eigenvalores de $\boldsymbol{P}$ son iguales a 1 y los restantes $(n-k)$ eigenvalores son iguales a 0. Se aplica un argumento similar para $(\boldsymbol{I}-\boldsymbol{P})$. ■

Para la prueba del Teorema 10 es necesario ver los siguientes dos lemas.

Lema 3. *El determinante de una matriz particionada. Se define la partición de* $\boldsymbol{S}$, *como sigue,*

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$$

1. *Si* $\boldsymbol{A}$ *es no singular, entonces el* $\boldsymbol{det(S)=det(A)det(D-CA^{-1}B)}$, *y*

2. *Si* $\boldsymbol{D}$ *es no singular, entonces el* $\boldsymbol{det(S)=det(D)det(A-BD^{-1}C)}$.

Demostración.

1. Sea $\boldsymbol{S}$ definida como antes, y sea además $\boldsymbol{S^*}$ definida de la siguiente manera,

$$\boldsymbol{S^*} = \begin{bmatrix} \boldsymbol{A^{-1}} & \boldsymbol{0} \\ \boldsymbol{-CA^{-1}} & \boldsymbol{I} \end{bmatrix},$$

entonces se ve que el $\boldsymbol{det(S^*)=det(A^{-1})=\dfrac{1}{det(A)}}$.

Considerando ahora el producto

$$\boldsymbol{S^*S} = \begin{bmatrix} \boldsymbol{A^{-1}} & \boldsymbol{0} \\ \boldsymbol{-CA^{-1}} & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{A^{-1}B} \\ \boldsymbol{0} & \boldsymbol{-CA^{-1}B+D} \end{bmatrix},$$

por lo que el

$$\boldsymbol{det(S^*S) = det(D-CA^{-1}B)},$$

o sea,

$$\boldsymbol{det(S^*)det(S) = det(D-CA^{-1}B)},$$

finalmente se llega a

$$\begin{aligned} \boldsymbol{det(S)} &= \frac{\boldsymbol{det(D-CA^{-1}B)}}{\boldsymbol{det(S^*)}} \\ &= \frac{\boldsymbol{det(D-CA^{-1}B)}}{\dfrac{1}{\boldsymbol{det(A)}}} \\ &= \boldsymbol{det(A)det(D-CA^{-1}B)} \end{aligned}$$

2. Se aplica un argumento similar al utilizado en la prueba del Lema 3(1). ■

Lema 4. *Sean* $\boldsymbol{B}$ *y* $\boldsymbol{C}$ *matrices de orden* $k \times m$. *Si* $\boldsymbol{A}$ *es una matriz de orden* $k \times k$ *no singular, entonces el*

$$\boldsymbol{det(A - BC^T) = det(A)det(I - C^T A^{-1} B).}$$

Demostración. Por el Lema 3(1), se tiene que el

$$\boldsymbol{det} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C^T} & \boldsymbol{I} \end{bmatrix} = \boldsymbol{det(A)det(I - C^T A^{-1} B),}$$

y por el Lema 3(2), se tiene que el

$$\boldsymbol{det} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C^T} & \boldsymbol{I} \end{bmatrix} = \boldsymbol{det(I)det(A - BI^{-1}C^T) = det(A - BC^T),}$$

Por lo tanto se comprueba que el $\boldsymbol{det(A - BC^T) = det(A)det(I - C^T A^{-1} B).}$

■

Teorema 10. *Sea* $\boldsymbol{P_I}$ *de orden* $m \times m$, *un menor de* $\boldsymbol{P}$, *dado por las filas y columnas de* $\boldsymbol{P}$ *indexadas por* I. *Si* $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_m$ *son los eigenvalores de* $\boldsymbol{P_I}$, *entonces,*

1. *Los eigenvalores de* $\boldsymbol{P_I}$ *y de* $(\boldsymbol{I - P_I})$ *están entre 0 y 1, esto es,* $0 \leq \lambda_j \leq 1$, $j = 1, 2, ..., m$,

2. $(\boldsymbol{I - P_I})$ *es una matriz positiva definida si* $\lambda_m < 1$, *si no, es positiva semi-definida, y*

3. *El* $\boldsymbol{r(X_{(I)})} < k$, *si* $\lambda_m = 1$.

Demostración.

1. Sea $\boldsymbol{P_{22}}$ de orden $m \times m$ un menor de $\boldsymbol{P}$, es decir, $\boldsymbol{P_{22}}$ son las últimas m filas y columnas de $\boldsymbol{P}$. Se considera la partición de $\boldsymbol{P}$, como

$$\boldsymbol{P}=\begin{bmatrix} \boldsymbol{P_{11}} & \boldsymbol{P_{12}} \\ \boldsymbol{P_{12}^T} & \boldsymbol{P_{22}} \end{bmatrix}$$

Como $\boldsymbol{P_{22}}$ es simétrica, entonces por el Teorema de la descomposición espectral (Searle, 1982), puede escribirse como $\boldsymbol{P_{22}}=\boldsymbol{V}\Lambda\boldsymbol{V^T}$, donde Λ es una matriz diagonal con los eigenvalores de $\boldsymbol{P_{22}}$ en la diagonal, y $\boldsymbol{V}$ es una matriz que contiene como columnas los correspondientes eigenvalores normalizados. Ahora, se ve que $\boldsymbol{P}=\boldsymbol{PP}$, entonces

$$\boldsymbol{P_{22}} = \boldsymbol{P_{12}^T P_{12}} + \boldsymbol{P_{22}P_{22}},$$

y como $\boldsymbol{P_{12}^T P_{12}} \geq 0$, entonces $\boldsymbol{P_{22}} \geq \boldsymbol{P_{22}P_{22}}$. Sustituyendo en ésta última expresión el valor de $\boldsymbol{P_{22}}$, se tiene

$$\boldsymbol{V\Lambda V^T} \geq \boldsymbol{V\Lambda V^T V\Lambda V^T} = \boldsymbol{V\Lambda^2 V^T},$$

o sea,

$$\boldsymbol{V\Lambda V^T} - \boldsymbol{V\Lambda^2 V^T} \geq 0,$$

luego,

$$\boldsymbol{V}(\Lambda - \Lambda^2)\boldsymbol{V^T} \geq 0,$$

y finalmente se llega a

$$(\Lambda - \Lambda^2) \geq 0,$$

la cual también se puede poner de la forma $(\lambda - \lambda^2) \geq 0$, de donde se deduce que los eigenvalores de $\boldsymbol{P_{22}}$ están entre 0 y 1.

2. Si $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_m$ son los eigenvalores de $\boldsymbol{P_{22}}$, entonces $(1 - \lambda_j)$, $j = 1, 2, ..., m$, son los eigenvalores de $(\boldsymbol{I} - \boldsymbol{P_{22}})$. Ahora, si $\lambda_m = 1$, entonces $(1 - \lambda_m) = 0$ y por lo tanto $(\boldsymbol{I} - \boldsymbol{P_{22}})$ es positiva semidefinida. Pero si $\lambda_m < 1$, entonces $(1 - \lambda_i) > 0$, $i = 1, 2, ..., m$, y por lo tanto $(\boldsymbol{I} - \boldsymbol{P_{22}})$ es positiva definida.

3. Se ve que la expresión $(\boldsymbol{X}_{(I)}^{\boldsymbol{T}}\boldsymbol{X}_{(I)})$ se puede escribir como $(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} - \boldsymbol{X}_I\boldsymbol{X}_I^{\boldsymbol{T}})$, y al aplicar el Lema 4, se obtiene

$$\begin{aligned} \boldsymbol{det}(\boldsymbol{X}_{(I)}^{\boldsymbol{T}}\boldsymbol{X}_{(I)}) &= \boldsymbol{det}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} - \boldsymbol{X}_I\boldsymbol{X}_I^{\boldsymbol{T}}) \\ &= \boldsymbol{det}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}) \quad \boldsymbol{det}\left(\boldsymbol{I} - \boldsymbol{X}_I^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_I\right) \\ &= \boldsymbol{det}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}) \quad \boldsymbol{det}(\boldsymbol{I} - \boldsymbol{P}_I), \end{aligned}$$

de este modo, el

$$\boldsymbol{det}(\boldsymbol{I} - \boldsymbol{P}_I) = \prod_{j=1}^{m}(1 - \lambda_j),$$

concluyéndose que si $\lambda_m = 1$, entonces, el $\boldsymbol{det}(\boldsymbol{X}_{(I)}^{\boldsymbol{T}}\boldsymbol{X}_{(I)}) = 0$, o equivalentemente, el $\boldsymbol{r}(\boldsymbol{X}_{(I)}) < k$, si $\lambda_m = 1$.

■

Distribución de p_{ii}

Recordando que la matriz de predicción $\boldsymbol{P}$, depende solamente de $\boldsymbol{X}$, la cual se supone es medida sin error. En algunos casos, sin embargo, $\boldsymbol{X}$ es medida con error, y entonces es razonable suponer que las filas de $\boldsymbol{X}$ tienen una distribución Normal multivariada con media μ y varianza Σ. En este sentido, la distribución de $\boldsymbol{p_{ii}}$ puede ser encontrada, tal como proponen Belsley, *et al.* (1980), quienes establecen que si las filas de $\tilde{\boldsymbol{X}}=(\boldsymbol{I} - n^{-1}\boldsymbol{1}\boldsymbol{1}^{\boldsymbol{T}})\boldsymbol{X}$ son independientes e idénticamente distribuídas como una distribución Normal $(k-1)$-dimensional, entonces,

$$\frac{(n-k)}{(k-1)}\frac{(\boldsymbol{p_{ii}} - n^{-1})}{(1 - \boldsymbol{p_{ii}})} \quad \sim \quad \boldsymbol{F}(k-1, n-k),$$

donde $\boldsymbol{F}\ (k-1, n-k)$ es la distribución $\boldsymbol{F}$ con $(k-1, n-k)$ grados de libertad. Pero también, recordando que $\boldsymbol{1}^{\boldsymbol{T}}\tilde{\boldsymbol{X}} = \boldsymbol{0}$, por lo que no se puede suponer que las filas de $\tilde{\boldsymbol{X}}$, es decir, $\tilde{\boldsymbol{x}}_{\boldsymbol{i}}$, $i = 1, 2, ..., n$, sean independientes. Afortunadamente, se da

la dependencia de las filas de $\tilde{\boldsymbol{X}}$ en cada disminución o incremento que se tiene en n. Lo anterior puede verse al calcular la covarianza entre $\tilde{\boldsymbol{x_j}}$ y $\tilde{\boldsymbol{x_i}}$, es decir,

$$\begin{aligned} \boldsymbol{Cov}(\tilde{\boldsymbol{x_j}}, \tilde{\boldsymbol{x_i}}) &= \boldsymbol{E}(\tilde{\boldsymbol{x_j}}\tilde{\boldsymbol{x_i}}^{\boldsymbol{T}}) - (\boldsymbol{E}(\tilde{\boldsymbol{x_j}}))\,(\boldsymbol{E}(\tilde{\boldsymbol{x_i}}))^{\boldsymbol{T}}, \quad i \neq j \\ &= \mu\mu^{\boldsymbol{T}} - 2(\mu\mu^{\boldsymbol{T}} + n^{-1}\Sigma) + \mu\mu^{\boldsymbol{T}} + n^{-1}\Sigma \\ &= -n^{-1}\Sigma, \end{aligned}$$

la cual converge a cero cuando n se aproxima a infinito.

EFECTOS DE UNA OBSERVACIÓN SOBRE LA ECUACIÓN DE REGRESIÓN

Cuando un conjunto de datos determinado, se somete a un análisis de regresión, en donde se ajusta el modelo dado en la ecuación (1.1), utilizando el método de Cuadrados Mínimos, entonces, los resultados de este ajuste, pueden ser influenciados por la eliminación o adición de una o varias observaciones. Normalmente, no todas las observaciones presentes en un conjunto de datos, influyen de igual manera en el ajuste de dicho modelo.

Las observaciones, que individual o colectivamente, influyen (en cierta medida) en la ecuación de regresión ajustada, se les llama observaciones influyentes, y éstas pueden ser: "outliers", puntos palanca o puntos influyentes.

Las medidas que se utilizan para detectar observaciones influyentes se han agrupado en dos métodos, que son:

1. **El método de la eliminación de observaciones**. Este procedimiento estudia cómo es que varios resultados del análisis de regresión, cambian cuando alguna o algunas de las observaciones son omitidas.

2. **El método basado en la diferenciación**. Por su parte, este método estudia las modificaciones que sufren varios de los resultados del análisis de regresión con respecto a ciertos parámetros del modelo.

En este capítulo, se analizará la detección de una sola observación influyente.

Método de la Eliminación de una Observación

Existen algunas medidas que permiten detectar una observación influyente, entre las que se pueden mencionar:

1. Medidas basadas en residuales.
2. Outliers, puntos palanca y puntos influyentes.
3. Medidas basadas en la curva de influencia.
4. Medidas basadas en el volumen de elipsoides de confianza.
5. Medidas basadas en la función de verosimilitud.
6. Medidas basadas en un subconjunto de coeficientes de regresión.

Medidas Basadas en Residuales

A pesar de que los errores aleatorios ϵ no pueden ser observados, si pueden ser medidos en cierta forma por los residuales $\hat{\epsilon}$, por lo que estos últimos juegan un papel fundamental en el diagnóstico del análisis de regresión. Recordando que los residuales $\hat{\epsilon}$ pueden ser escritos como $\hat{\epsilon}=(\boldsymbol{I}-\boldsymbol{P})\epsilon$, en donde se observa que $\hat{\epsilon}$ mide razonablemente a ϵ, sólo si los elementos de la diagonal de $\boldsymbol{P}$ son lo suficientemente pequeños. Notando además que, aún cuando los elementos de ϵ sean independientes y con la misma varianza, por la identidad $\hat{\epsilon}=(\boldsymbol{I}-\boldsymbol{P})\epsilon$ se ve que los residuales $\hat{\epsilon}$ no son independientes (a menos que $\boldsymbol{P}$ sea diagonal), y no tienen la misma varianza (a menos que los elementos de la diagonal de $\boldsymbol{P}$ sean iguales). Por lo que puede mencionarse que, $\hat{\epsilon}$ estima aproximadamente a ϵ, si las filas de $\boldsymbol{X}$ son homogéneas, así, los elementos de la diagonal de $\boldsymbol{P}$ son aproximadamente iguales y lo suficientemente pequeños.

Por lo anterior, en vez de usar $\widehat{\epsilon_i}$ en el diagnóstico del análisis de regresión, es preferible utilizar una versión transformada de los residuales, esto es,

$$f(\widehat{\epsilon_i}, \sigma_i) = \frac{\widehat{\epsilon_i}}{\sigma_i}, \tag{3.1}$$

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

De la ecuación (3.1), se obtienen los residuales normalizados, estandarizados y dos casos de residuales estudentizados.

Residual Normalizado

El i-ésimo residual normalizado es obtenido al reemplazar σ_i por $\sqrt{\widehat{\epsilon}^{\boldsymbol{T}}\widehat{\epsilon}}$ en la ecuación (3.1), con lo que se obtiene

$$\boldsymbol{a_i} \equiv f\left(\widehat{\epsilon_i}, \sqrt{\widehat{\epsilon}^{\boldsymbol{T}}\widehat{\epsilon}}\right) = \frac{\widehat{\epsilon_i}}{\sqrt{\widehat{\epsilon}^{\boldsymbol{T}}\widehat{\epsilon}}}, \quad i = 1, 2, ..., n. \tag{3.2}$$

Residual Estandarizado

El i-ésimo residual estandarizado es encontrado al sustituir la expresión

$$\widehat{\sigma} = \sqrt{\frac{\widehat{\epsilon}^{\boldsymbol{T}}\widehat{\epsilon}}{n-k}},$$

en lugar de σ_i en la ecuación (3.1), resultando

$$\boldsymbol{b_i} \equiv f(\widehat{\epsilon_i}, \widehat{\sigma}) = \frac{\widehat{\epsilon_i}}{\widehat{\sigma}}, \quad i = 1, 2, ..., n. \tag{3.3}$$

Residual Interno Estudentizado

Si en la ecuación (3.1), se considera que

$$\sigma_i = \widehat{\sigma}\sqrt{1 - \boldsymbol{p_{ii}}},$$

entonces resulta el i-ésimo residual interno estudentizado, que es

$$\boldsymbol{r_i} \equiv f\left(\widehat{\epsilon_i}, \widehat{\sigma}\sqrt{1-\boldsymbol{p_{ii}}}\right) = \frac{\widehat{\epsilon_i}}{\widehat{\sigma}\sqrt{1-\boldsymbol{p_{ii}}}}, \quad i = 1, 2, ..., n. \tag{3.4}$$

Residual Externo Estudentizado

Tomando ahora

$$\sigma_i = \widehat{\sigma}_{(i)}\sqrt{1 - \boldsymbol{p_{ii}}}$$

en la ecuación (3.1), se obtiene el i-ésimo residual externo estudentizado, dado por

$$\boldsymbol{r_i}^* \equiv f\left(\widehat{\epsilon}_{\boldsymbol{i}}, \widehat{\sigma}_{(i)}\sqrt{1 - \boldsymbol{p_{ii}}}\right) = \frac{\widehat{\epsilon}_{\boldsymbol{i}}}{\widehat{\sigma}_{(i)}\sqrt{1 - \boldsymbol{p_{ii}}}}, \quad i = 1, 2, ..., n, \tag{3.5}$$

donde

$$\widehat{\sigma}^2_{(i)} = \frac{\boldsymbol{Y}^{\boldsymbol{T}}_{(\boldsymbol{i})}(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{i})})\boldsymbol{Y}_{(\boldsymbol{i})}}{n - k - 1}, \quad i = 1, 2, ..., n,$$

es el residual que estima el Cuadrado Medio del Error cuando la i-ésima observación es omitida, y

$$\boldsymbol{P}_{(\boldsymbol{i})} = \boldsymbol{X}_{(\boldsymbol{i})}(\boldsymbol{X}^{\boldsymbol{T}}_{(\boldsymbol{i})}\boldsymbol{X}_{(\boldsymbol{i})})^{-1}\boldsymbol{X}^{\boldsymbol{T}}_{(\boldsymbol{i})}, \quad i = 1, 2, ..., n,$$

es la matriz de predicción para $\boldsymbol{X}_{(\boldsymbol{i})}$.

Para encontrar las relaciones entre los cuatro casos de residuales, primero, es conveniente ver que están en función de cuatro cantidades que son: $(n-k)$, $\widehat{\epsilon}^{\,\boldsymbol{T}}\widehat{\epsilon}$, $\widehat{\epsilon}_{\boldsymbol{i}}$ y $(1 - \boldsymbol{p_{ii}})$. Y segundo, es necesario escribir $\widehat{\sigma}_{(i)}$ en términos de $\widehat{\sigma}$. Esto puede verse al omitir la i-ésima observación, pues es equivalente a ajustar el modelo

$$\boldsymbol{E(Y)} = \boldsymbol{X\beta} + \boldsymbol{u_i}\theta,$$

donde θ es el coeficiente de regresión del i-ésimo vector unitario $\boldsymbol{u_i}$.

Usando la expresión dada en el Teorema 4, se tiene que

$$\begin{aligned}
\boldsymbol{SSE}_{(\boldsymbol{i})} &= \boldsymbol{Y}^{\boldsymbol{T}}_{(\boldsymbol{i})}(\boldsymbol{I} - \boldsymbol{P}_{(\boldsymbol{i})})\boldsymbol{Y}_{(\boldsymbol{i})} \\
&= \boldsymbol{Y}^{\boldsymbol{T}}\left(\boldsymbol{I} - \boldsymbol{P} - \frac{(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{u_i}\boldsymbol{u}^{\boldsymbol{T}}_{\boldsymbol{i}}(\boldsymbol{I} - \boldsymbol{P})}{\boldsymbol{u}^{\boldsymbol{T}}_{\boldsymbol{i}}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{u_i}}\right)\boldsymbol{Y} \\
&= \boldsymbol{SSE} - \frac{\widehat{\epsilon}^2_{\boldsymbol{i}}}{1 - \boldsymbol{p_{ii}}}
\end{aligned} \tag{3.6}$$

donde $\boldsymbol{SSE}$ es la Suma de Cuadrados de Residuales y $\boldsymbol{SSE_{(i)}}$ es la Suma de Cuadrados de Residuales cuando la i-ésima observación es omitida. Observando que al dividir ambos lados de la última expresión por $(n-k-1)$, se obtiene

$$\begin{aligned} \widehat{\sigma}^2_{(i)} &= \frac{(n-k)}{(n-k-1)}\widehat{\sigma}^2 - \frac{\widehat{\epsilon}_i^2}{(n-k-1)(1-\boldsymbol{p_{ii}})} \\ &= \widehat{\sigma}^2\left(\frac{n-k-\boldsymbol{r_i}^2}{n-k-1}\right) \end{aligned} \tag{3.7}$$

Además, $\boldsymbol{b_i} = \boldsymbol{a_i}\sqrt{n-k}$, y por lo tanto

$$\boldsymbol{r_i} = \frac{\boldsymbol{b_i}}{\sqrt{1-\boldsymbol{p_{ii}}}} = \boldsymbol{a_i}\sqrt{\frac{n-k}{1-\boldsymbol{p_{ii}}}}$$

Si en la ecuación (3.5), se reemplaza $\widehat{\sigma}_{(i)}$ por $\sqrt{\dfrac{\boldsymbol{SSE_{(i)}}}{(n-k-1)}}$, y posteriormente se utiliza la ecuación (3.6), entonces $\boldsymbol{r_i}^*$ se puede escribir como

$$\boldsymbol{r_i}^* = \frac{\widehat{\epsilon}_i}{\sqrt{\frac{(1-\boldsymbol{p_{ii}})SSE_{(i)}}{n-k-1}}} = \frac{\widehat{\epsilon}_i}{\sqrt{\left(\frac{1-\boldsymbol{p_{ii}}}{n-k-1}\right)\left(SSE - \frac{\widehat{\epsilon}_i^2}{1-\boldsymbol{p_{ii}}}\right)}} = \frac{\boldsymbol{a_i}\sqrt{n-k-1}}{\sqrt{(1-\boldsymbol{p_{ii}}) - \boldsymbol{a_i}^2}}$$

Asimismo, la sustitución de la ecuación (3.7) en la ecuación (3.5), resulta otra forma de escribir $\boldsymbol{r_i}^*$, esto es,

$$\boldsymbol{r_i}^* = \boldsymbol{r_i}\sqrt{\frac{n-k-1}{n-k-\boldsymbol{r_i}^2}}$$

Las distribuciones de $\boldsymbol{r_i}$ y de $\boldsymbol{r_i}^*$ se presentan en los siguientes teoremas, cuyas pruebas pueden verse en Chatterjee y Hadi (1988).

Teorema 11. *Suponiendo que* $\boldsymbol{X}$ *en el modelo dado en la ecuación (1.1) es de rango* k*, entonces:*

1. *Ellenberg (1973), señala que si el* $\boldsymbol{r}(\boldsymbol{X_{(i)}})=k \quad y \quad \epsilon \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$*, entonces* $\left(\dfrac{\boldsymbol{r_i}^2}{n-k}\right)$*, para* $i = 1, 2, ..., n$*, son idénticamente distribuídos como* $Beta\left(\frac{1}{2}, \frac{1}{2}(n-k-1)\right)$*, y*

2. *Beckman y Trussel (1974), indican que si el* $\boldsymbol{r}(\boldsymbol{X_{(i)}})=(k-1)$*, entonces* $\boldsymbol{r_i}$ *no está definido, esto es,* $\widehat{\epsilon}_i=0$ *y* $\boldsymbol{Var}(\widehat{\epsilon}_i)=0$.

Teorema 12. *Beckman y Trussel (1974), afirman que si se supone que* $\boldsymbol{X}$ *en el modelo dado en la ecuación (1.1) es de rango* k*, entonces:*

1. *Si el* $\boldsymbol{r}(\boldsymbol{X}_{(i)})=k$ *y* $\epsilon \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$*, entonces* $\boldsymbol{r_i}^*$*, para* $i = 1, 2, ..., n$*, son idénticamente distribuídos como* $\boldsymbol{t}(n-k-1)$*, y*

2. *Si el* $\boldsymbol{r}(\boldsymbol{X}_{(i)})=(k-1)$*, entonces* $\boldsymbol{r_i}^*$ *no está definido.*

Para probar la suficiencia de los supuestos del modelo de la ecuación (1.1), no se debe preferir alguna de las formas de residuales sobre otra, ya que ninguna es independientemente distribuída, pero para la mayoría de los problemas prácticos, especialmente cuando el tamaño de muestra es grande, se pueden considerar independientes, tal como lo señalan Behnken y Draper (1972), Chatterjee y Price (1977) y Draper y Smith (1981).

Behnken y Draper (1972), mencionan además, que los residuales normalizados y los estandarizados son básicamente equivalentes, pues son múltiplos constantes de $\widehat{\epsilon_i}$, sin embargo, también sugieren que si varían los elementos de la diagonal de $\boldsymbol{P}$, y por consiguiente, las varianzas de $\widehat{\epsilon_i}$, para $i = 1, 2, ..., n$, sería preferible usar $\boldsymbol{r_i}$.

Algunos otros autores como Atkinson (1981 y 1982) y Velleman y Welsch (1981), prefieren $\boldsymbol{r_i}^*$ sobre $\boldsymbol{r_i}$, ya que como se ve en el Teorema 12(1),

$$\boldsymbol{r_i}^* \sim \boldsymbol{t}(n-k-1),$$

para lo cual las tablas de $\boldsymbol{t}$ estan disponibles.

Existen dos maneras en que los residuales $\boldsymbol{r_i}$ y $\boldsymbol{r_i}^*$, permiten detectar "outliers", es decir, dos procedimientos de prueba que ayudan a descubrir observaciones que se alejan de la nube de puntos en un diagrama de dispersión de un conjunto de datos, estos son la prueba formal e informal para detectar un "outlier".

Prueba Formal para Detectar un "Outlier"

En regresión lineal, existen algunos métodos, los cuales suponen que existe máximo un "outlier" en un conjunto de datos dado y generalmente son de dos formas,

1. Para ciertos estadísticos $\boldsymbol{T(x_i, y_i)}$, se encuentra $\boldsymbol{C\alpha}$ tal que,

$$\boldsymbol{P}\{\boldsymbol{T(x_i, y_i)} > \boldsymbol{C\alpha} : \text{máximo un "outlier" está presente}\} \leq \alpha, \quad \text{y}$$

2. Se declara la i-ésima observación como un "outlier" si $\boldsymbol{T(x_i, y_i)} > \boldsymbol{C\alpha}$.

Tietjen *et al.* (1973), sugieren utilizar $\boldsymbol{T(x_i, y_i)}=max|\boldsymbol{r_i}| \equiv \boldsymbol{r_{max}}$, para el caso de una regresión lineal simple; de manera que los valores críticos aproximados para $\boldsymbol{r_{max}}$, están dados por

$$\boldsymbol{r_{max}} = \sqrt{\frac{(n-k)\boldsymbol{F}}{n-k-1+\boldsymbol{F}}},$$

donde $\boldsymbol{F}$, es el $100(1-\frac{\alpha}{n})$ punto de $\boldsymbol{F}$ $(1, n-k-1)$. Lund (1975), proporciona valores críticos para $\boldsymbol{r_{max}}$ para el caso general de regresión múltiple.

Prueba Informal para Detectar un "outlier" (Métodos Gráficos)

Los métodos gráficos proporcionan información valiosa acerca de la presencia de "outliers", de la suficiencia del modelo y de la validación de sus supuestos asociados. Behnken y Draper (1972), sugieren utilizar $\boldsymbol{r_i}$ en estos gráficos, mientras que, Atkinson (1981) prefiere $\boldsymbol{r_i}^*$, pero como $\boldsymbol{r_i}^*$ es una transformación monótona de $\boldsymbol{r_i}$, las conclusiones hechas de gráficos basados en estos residuales, usualmente son las mismas.

Algunas de las gráficas de residuales más comúnmente usadas son:

1. Distribución de frecuencias de residuales. Las gráficas de una dimensión, tales como

   (a) Histogramas,

(b) Gráficas de puntos,

(c) Gráficas de tallos y hojas, y

(d) Gráficas de cajas,

son una vía fácil para observar si la muestra fue extraída de una población Normal.

2. Gráficas de residuales en series de tiempo. Estos gráficos proveen información para validar el supuesto de que los ϵ_i, $i = 1, 2, ..., n$, son independientes, si no lo son, entonces el problema es conocido como "autocorrelación".

3. Gráficas de probabilidad normal. Este gráfico debe ser aproximadamente una línea recta, donde el intercepto y la pendiente, son estimaciones de μ y σ, respectivamente.

4. Gráficas de residuales contra valores ajustados (predichos). Este gráfico proporciona información acerca de la validación de los supuestos de linealidad y homogeneidad de varianzas.

Outliers, Puntos Palanca y Puntos Influyentes

Los outliers, puntos palanca y puntos influyentes son tres conceptos relacionados entre sí, a continuación se define cada uno de ellos.

"Outliers"

En regresión lineal, se define a un "outlier" como una observación para la cual, el residual estudentizado (ya sea r_i o r_i^*), es grande en magnitud comparado con el que presentan las demás observaciones de un conjunto de datos determinado. El "outlier" se considera como tal, cuando la ecuación de regresión no se ajusta con éxito.

Puntos Palanca

Los puntos palanca, son aquellas observaciones para las cuales el vector $\boldsymbol{x_i}$, se encuentra lejos del resto de las observaciones en un conjunto de datos dado. Equivalentemente, un punto palanca es una observación que presenta un $\boldsymbol{p_{ii}}$ grande en comparación con el que se observa en las otras observaciones del conjunto de datos. Las observaciones que se encuentran aisladas en el espacio $\boldsymbol{X}$ son consideradas como puntos palanca, de manera que en el espacio $\boldsymbol{X}$ los puntos palanca pueden ser también considerados como "outliers", por lo tanto, puede decirse que las variables predictoras en $\boldsymbol{X}$ son las que determinan los puntos palanca, no la variable respuesta en $\boldsymbol{Y}$.

Puntos Influyentes

Por su parte, los puntos influyentes, se definen como observaciones que, individual o colectivamente, influyen excesivamente en la ecuación de regresión ajustada, comparadas con las demás observaciones del conjunto de datos.

Sin embargo, una observación puede no tener la misma influencia sobre los resultados del análisis de regresión, por ejemplo, sobre el vector de parámetros $\hat{\boldsymbol{\beta}}$, o sobre el vector de valores ajustados o predichos $\widehat{\boldsymbol{Y}}$. De este modo, en un análisis de regresión, lo importante estriba en saber qué es lo primordial, si la estimación de $\boldsymbol{\beta}$ o la predicción en $\boldsymbol{Y}$, para posteriormente, dedicarse a medir la influencia de las observaciones sobre $\widehat{\boldsymbol{\beta}}$ o sobre $\widehat{\boldsymbol{Y}}$, respectivamente.

Es interesante notar que:

1. Los "outliers" no necesariamente son puntos influyentes, y viceversa.

2. Las observaciones con residuales grandes no son deseables, pero se debe tener cuidado, pues un residual pequeño, no necesariamente significa que la observación correspondiente es adecuada, esto generalmente es debido a que en el

ajuste por Cuadrados Mínimos se evitan residuales grandes, lo cual puede hacer que un punto indeseable se acomode a los otros puntos del conjunto de datos. De hecho, existe una tendencia general en los puntos palanca que muestran residuales pequeños pero influyen en el ajuste, ocasionando una ecuación de regresión desproporcionada.

3. Los puntos palanca no necesariamente son puntos influyentes, y viceversa.

Recordando la desigualdad del Teorema 6(4), la cual implica que las observaciones con p_{ii} grande, tienden a tener residuales pequeños, por lo que se consideran puntos palanca. A continuación se mencionan tres cantidades relacionadas con la medición de este tipo de observaciones, que son los elementos de la diagonal de P, la distancia de Mahalanobis y los elementos de la diagonal de P_Z.

Elementos de la Diagonal de P

Los p_{ii} juegan un papel importante en la determinación de los valores ajustados $\widehat{Y}$, y la magnitud de los residuales, por lo que Hoaglin y Welsch (1978), sugieren la utilización de r_i^* para detectar "outliers" y p_{ii} para detectar puntos palanca que son potencialmente influyentes. Comúnmente se aplican los siguientes tres puntos de corte para p_{ii}:

1. Se puede considerar a $\frac{1}{p_{ii}}$ como el número de observaciones que determinan $\widehat{y_i}$. Por esto, Huber (1981), señala que puntos con $p_{ii} > 0.2$ se clasifican como puntos palanca. En este sentido, se requiere poner atención especial a aquellas observaciones cuyos valores predichos $\widehat{y_i}$, estén determinados por cinco o menos observaciones.

2. Por el Teorema 3(1), se puede deducir que el promedio de los elementos de la diagonal de P es $\frac{k}{n}$, esto condujo a Hoaglin y Welsch (1978) a indicar que puntos con $p_{ii} > \frac{2k}{n}$, se consideran puntos palanca.

3. Basándose en las condiciones de la distribución de p_{ii}, descritas en el segundo capítulo, se llega entonces a considerar puntos con

$$p_{ii} \leq \frac{nF(k-1)+(n-k)}{nF(k-1)+n(n-k)}$$

como puntos palanca, donde F es el $100(1-\alpha)$ punto de F $(k-1, n-k)$.

Distancia de Mahalanobis

Se supone que X contiene una columna de unos y que $\tilde{X}$ denota la X centrada excluyendo la columna constante, por lo tanto, un estadístico que mide que tan lejos está x_i del centro del conjunto de datos, comúnmente se calcula como

$$(n-1)^{-1}\tilde{x}_i^T(\tilde{X}^T\tilde{X})^{-1}\tilde{x}_i,$$

donde $\tilde{x}_i$ es la i-ésima fila de $\tilde{X}$. Sin embargo, lo que interesa medir es que tan lejos está x_i del resto de las observaciones, por lo que debe excluirse x_i cuando se calcula la media y la matriz de varianzas de X, entonces la distancia de Mahalanobis es definida como

$$M_i = (n-2)(\tilde{x}_i - \bar{\tilde{X}}_{(i)})^T\left(\tilde{X}_{(i)}^T\left(I-(n-1)^{-1}11^T\right)\tilde{X}_{(i)}\right)^{-1}(\tilde{x}_i - \bar{\tilde{X}}_{(i)}),$$

donde $\bar{\tilde{X}}_{(i)}$ es el promedio de $\tilde{X}_{(i)}$. Usando la ecuación (2.4) y notando además, que

$$\bar{\tilde{X}}_{(i)} = (n-1)^{-1}\tilde{X}_{(i)}^T 1 = -(n-1)^{-1}\tilde{x}_i,$$

puesto que $\tilde{X}$ es centrada, entonces M_i se reduce a

$$M_i = \frac{n(n-2)}{(n-1)}\frac{(p_{ii}-\frac{1}{n})}{(1-p_{ii})}, \quad i = 1, 2, ..., n,$$

donde puede verse que M_i es equivalente a p_{ii}.

Elementos de la Diagonal de P_Z

Una desventaja de utilizar solamente a p_{ii} como medida de diagnóstico para identificar puntos palanca, es que no se dispone de la información contenida en Y;

en este caso, se utiliza la matriz aumentada $\boldsymbol{Z} = (\boldsymbol{X} : \boldsymbol{Y})$, donde se define al i-ésimo elemento de la diagonal de la matriz de predicción $\boldsymbol{P_Z}$ como $\boldsymbol{p_{z_{ii}}}$, y recordando el Teorema 6(4), $\boldsymbol{p_{z_{ii}}}$ se puede escribir como

$$\boldsymbol{p_{z_{ii}}} = \boldsymbol{z_i}^{\boldsymbol{T}}(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{Z})^{-1}\boldsymbol{z_i} = \boldsymbol{p_{ii}} + \frac{\widehat{\epsilon}_{\boldsymbol{i}}^{\boldsymbol{2}}}{\widehat{\boldsymbol{\epsilon}}^{\boldsymbol{T}}\widehat{\boldsymbol{\epsilon}}}$$

En consecuencia, $\boldsymbol{p_{z_{ii}}}$ será grande, siempre que $\boldsymbol{p_{ii}}$ sea grande o $\widehat{\epsilon}_{\boldsymbol{i}}^{\boldsymbol{2}}$ sea grande, o ambas lo sean; es por esto, que $\boldsymbol{p_{z_{ii}}}$ no debe usarse como medida, pues no distingue entre un punto palanca en el espacio $\boldsymbol{X}$ y un "outlier" en el espacio $\boldsymbol{Z}$.

Medidas Basadas en la Curva de Influencia

Una clase importante de medidas de la influencia de la i-ésima observación sobre los resultados del análisis de regresión, están basadas en la curva de influencia, cuyo concepto fue introducido por Hampel (1968 y 1974), y el cual es dado a continuación.

Curva de Influencia

Esta definición es general, en el sentido de que puede aplicarse a diferentes modelos y métodos de estimación, pero sólo se menciona lo referente a modelos de regresión y estimación mínimo cuadrática. Para estimar algunos parámetros de interés, suponiendo que se tiene un estadístico $\boldsymbol{T}$, basado en una muestra aleatoria muy grande, $z_1, z_2, ..., z_n$, proveniente de una función de distribución acumulada $\boldsymbol{F}$; al adicionar una observación más z a esta muestra, se observa cómo se modifica al estadístico $\boldsymbol{T}$ y las conclusiones basadas en $\boldsymbol{T}$. Por lo que, la curva de influencia es definida como

$$\psi(\boldsymbol{z}, \boldsymbol{F}, \boldsymbol{T}) = \lim_{\epsilon \to 0} \frac{\boldsymbol{T}\left((1-\epsilon)\boldsymbol{F} + \epsilon\delta_z\right) - \boldsymbol{T}(\boldsymbol{F})}{\epsilon}$$

donde se observa que el límite existe en z, donde $\delta_z = 1$.

Definiendo $\boldsymbol{F_\epsilon}$=$(1-\epsilon)\boldsymbol{F} + \epsilon\delta_z$ y escribiendo nuevamente $\psi(\boldsymbol{z}, \boldsymbol{F}, \boldsymbol{T})$, así,

$$\psi(\boldsymbol{z}, \boldsymbol{F}, \boldsymbol{T}) = \lim_{\epsilon \to 0} \frac{\boldsymbol{T}(\boldsymbol{F_\epsilon}) - \boldsymbol{T}(\boldsymbol{F})}{\epsilon} = \frac{d}{d\epsilon}\boldsymbol{T}(\boldsymbol{F_\epsilon})|_{\epsilon=0}$$

Entonces, la curva de influencia es la derivada de la función $\boldsymbol{T}(\boldsymbol{F}_\epsilon)$ con respecto a ϵ, evaluada en $\epsilon = 0$.

La curva de influencia tiene varias aplicaciones, una de las cuales, se usa para estudiar las propiedades asintóticas de un estimador, para lo cual, considerando a $\boldsymbol{F_n}$ como la función de distribución empírica basada en n observaciones, y suponiendo que $\boldsymbol{F_n}$ converge a $\boldsymbol{F}$ y que $\boldsymbol{T}(\boldsymbol{F_n})$ converge a $\boldsymbol{T}(\boldsymbol{F})$, entonces bajo condiciones de regularidad apropiadas, se tiene que

$$\sqrt{n}\,(\boldsymbol{T}(\boldsymbol{F_n}) - \boldsymbol{T}(\boldsymbol{F})) \cong \sqrt{n} \int \psi(\boldsymbol{z}, \boldsymbol{F}, \boldsymbol{T}) d\boldsymbol{F_n}(\boldsymbol{z}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(\boldsymbol{z}, \boldsymbol{F}, \boldsymbol{T})$$

Por el Teorema Central del Límite, el lado derecho de esta expresión es asintóticamente Normal con media cero y varianza $\Sigma_\psi = \int \psi(\boldsymbol{z}, \boldsymbol{F}, \boldsymbol{T}) \psi^{\boldsymbol{T}}(\boldsymbol{z}, \boldsymbol{F}, \boldsymbol{T}) d\boldsymbol{F}(\boldsymbol{z})$.

Otra aplicación de la curva de influencia, se utiliza para comparar estimadores y sugerir estimadores robustos como otra alternativa de estimación. A continuación se da otro uso de las curvas de influencia, el cual trata sobre la influencia de las observaciones sobre los estimadores de $\boldsymbol{\beta}$ y σ^2.

Curvas de Influencia para $\hat{\boldsymbol{\beta}}$ y $\hat{\sigma}^2$

Si la curva de influencia para $\boldsymbol{T}$ no es acotada, entonces $\boldsymbol{T}$ se dice ser un estimador no robusto, puesto que es sensible a observaciones extremas. Los estimadores de interés para $\boldsymbol{\beta}$ y σ^2, en este caso, son los que resultan de la estimación mínimo cuadrática, los cuales son obtenidos al solucionar el sistema de las $(k+1)$ ecuaciones lineales

$$\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x_i}(\boldsymbol{y_i} - \boldsymbol{x_i^T}\boldsymbol{\beta}) = 0 \qquad (3.8)$$

$$\frac{1}{n-k} \sum_{i=1}^{n} (\boldsymbol{y_i} - \boldsymbol{x_i^T}\boldsymbol{\beta})^2 = \sigma^2 \qquad (3.9)$$

Suponiendo que el vector $(k+1)$-dimensional, dado por $(\boldsymbol{x^T}, \boldsymbol{y})$ tiene una función de distribución acumulada conjunta $\boldsymbol{F}$, entonces las ecuaciones (3.8) y (3.9), pueden

escribirse como

$$\int_{\boldsymbol{x^T},\boldsymbol{y}} \boldsymbol{x}(\boldsymbol{y}-\boldsymbol{x^T\beta})d\boldsymbol{F_n}(\boldsymbol{x^T},\boldsymbol{y}) = 0 \tag{3.10}$$

$$\int_{\boldsymbol{x^T},\boldsymbol{y}} (\boldsymbol{y}-\boldsymbol{x^T\beta})d\boldsymbol{F_n}(\boldsymbol{x^T},\boldsymbol{y}) = \sigma^2 \tag{3.11}$$

La solucion de (3.10) y (3.11) produce funciones para $\boldsymbol{\beta}$ y σ^2.

Se supone ahora que,

$$\boldsymbol{E_F}\left\{\begin{bmatrix} x \\ y \end{bmatrix}\begin{bmatrix} x^T & y \end{bmatrix}\right\} = \begin{bmatrix} \Sigma_{xx}(\boldsymbol{F}) & \Sigma_{xy}(\boldsymbol{F}) \\ \Sigma_{xy}^{\boldsymbol{T}}(\boldsymbol{F}) & \sigma_{yy}(\boldsymbol{F}) \end{bmatrix},$$

entonces resulta que

$$\boldsymbol{\beta}(\boldsymbol{F}) = {\Sigma_{xx}}^{-1}(\boldsymbol{F})\Sigma_{xy}(\boldsymbol{F})$$

y que

$$\sigma^2(\boldsymbol{F}) = \sigma_{yy}(\boldsymbol{F}) - \Sigma_{xy}^{\boldsymbol{T}}(\boldsymbol{F}){\Sigma_{xx}}^{-1}(\boldsymbol{F})\Sigma_{xy}(\boldsymbol{F})$$

Así, la curva de influencia puede ser obtenida sustituyendo z por el vector $(\boldsymbol{x^T},\boldsymbol{y})$ y reemplazando $\boldsymbol{T}$ por $\boldsymbol{\beta}(\boldsymbol{F})$ o $\sigma^2(\boldsymbol{F})$, entonces se tiene que

$$\psi(\boldsymbol{x^T},\boldsymbol{y},\boldsymbol{F},\boldsymbol{T}) = \lim_{\epsilon\to 0}\frac{\boldsymbol{T}\left((1-\epsilon)\boldsymbol{F}+\epsilon\delta_{\boldsymbol{x^T y}}\right)-\boldsymbol{T}(\boldsymbol{F})}{\epsilon} \tag{3.12}$$

Se requiere del siguiente lema, para probar el Teorema 13, pues éste último, proporciona formas explícitas de curvas de influencia para $\widehat{\boldsymbol{\beta}}$ y $\widehat{\sigma}^2$, utilizando la ecuación (3.12).

Lema 5. *Sea* $\boldsymbol{A}$ *cualquier matriz, tal que* $(\boldsymbol{I}+\epsilon\boldsymbol{A})^{-1}$ *exista, entonces*

$$(\boldsymbol{I}+\epsilon\boldsymbol{A})^{-1} = \boldsymbol{I}+\sum_{i=1}^{\infty}(-1)^i\epsilon^i\boldsymbol{A}^i$$

Demostración. Usando repetidamente $(\boldsymbol{I}+\epsilon\boldsymbol{A})^{-1} = \boldsymbol{I}-\epsilon\boldsymbol{A}(\boldsymbol{I}+\epsilon\boldsymbol{A})^{-1}$, se tiene que

$$(\boldsymbol{I}+\epsilon\boldsymbol{A})^{-1} = \boldsymbol{I}-\epsilon\boldsymbol{A}(\boldsymbol{I}+\epsilon\boldsymbol{A})^{-1}$$

$$
\begin{aligned}
&= \boldsymbol{I} - \epsilon\boldsymbol{A} + \epsilon^2\boldsymbol{A}^2(\boldsymbol{I} + \epsilon\boldsymbol{A})^{-1} \\
&= \boldsymbol{I} - \epsilon\boldsymbol{A} + \epsilon^2\boldsymbol{A}^2 - \epsilon^3\boldsymbol{A}^3(\boldsymbol{I} + \epsilon\boldsymbol{A})^{-1} \\
&= \quad . \\
&= \quad . \\
&= \quad . \\
&= \boldsymbol{I} + \sum_{i=1}^{\infty}(-1)^i\epsilon^i\boldsymbol{A}^i \blacksquare
\end{aligned}
$$

Teorema 13.

1. *La curva de influencia para* $\widehat{\boldsymbol{\beta}}$ *es*

$$\psi\left(\boldsymbol{x}^{\boldsymbol{T}}, \boldsymbol{y}, \boldsymbol{F}, \widehat{\boldsymbol{\beta}}(\boldsymbol{F})\right) = {\Sigma_{xx}}^{-1}(\boldsymbol{F})\boldsymbol{x}\left(\boldsymbol{y} - \boldsymbol{x}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}(\boldsymbol{F})\right)$$

2. *La curva de influencia para* $\widehat{\sigma}^2$ *es*

$$\psi\left(\boldsymbol{x}^{\boldsymbol{T}}, \boldsymbol{y}, \boldsymbol{F}, \widehat{\sigma}^2(\boldsymbol{F})\right) = \left(\boldsymbol{y} - \boldsymbol{x}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}(\boldsymbol{F})\right)^2 - \sigma_{yy}(\boldsymbol{F}) + \Sigma_{xy}^{\boldsymbol{T}}(\boldsymbol{F})\widehat{\boldsymbol{\beta}}(\boldsymbol{F})$$

Demostración.

1. La sustitución de $\boldsymbol{\beta}(\boldsymbol{F})$ en la ecuación (3.12), produce la curva de influencia para $\widehat{\boldsymbol{\beta}}$, es decir,

$$\psi\left(\boldsymbol{x}^{\boldsymbol{T}}, \boldsymbol{y}, \boldsymbol{F}, \boldsymbol{\beta}(\boldsymbol{F})\right) = \lim_{\epsilon \to 0} \frac{\widehat{\boldsymbol{\beta}}\left((1-\epsilon)\boldsymbol{F} + \epsilon\delta_{\boldsymbol{x}\boldsymbol{T}_{\boldsymbol{y}}}\right) - \widehat{\boldsymbol{\beta}}(\boldsymbol{F})}{\epsilon}$$

y luego al utilizar el Lema 5, se prueba el Teorema 13(1).

2. La sustitución de $\sigma^2(\boldsymbol{F})$ en la ecuación (3.12), produce la curva de influencia para $\widehat{\sigma}^2$, entonces se tiene

$$
\begin{aligned}
\psi\left(\boldsymbol{x}^{\boldsymbol{T}}, \boldsymbol{y}, \boldsymbol{F}, \sigma^2(\boldsymbol{F})\right) &= \lim_{\epsilon \to 0} \frac{\widehat{\sigma}^2\left((1-\epsilon)\boldsymbol{F} + \epsilon\delta_{\boldsymbol{x}\boldsymbol{T}_{\boldsymbol{y}}}\right) - \widehat{\sigma}^2(\boldsymbol{F})}{\epsilon} \\
&= \frac{d}{d\epsilon}\operatorname{plim}\widehat{\sigma}^2(\boldsymbol{F}_{\boldsymbol{\epsilon}})|_{\epsilon=0},
\end{aligned}
$$

resolviendo la derivada y evaluando en $\epsilon = 0$, se prueba el Teorema 13(2). ■

Notando que las curvas de influencia para $\widehat{\boldsymbol{\beta}}$ y $\widehat{\sigma}^2$, no son acotadas, pues $\left(y - \boldsymbol{x}^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}(\boldsymbol{F})\right)$ tampoco lo es, por lo que no existen estimadores robustos para $\boldsymbol{\beta}$ y σ^2.

Aproximación de la Curva de Influencia

Las curvas de influencia dadas en el Teorema 13, miden la influencia sobre $\widehat{\boldsymbol{\beta}}$ y $\widehat{\sigma}^2$, al adicionarse una observación $(\boldsymbol{x}^{\boldsymbol{T}}, y)$ a una muestra muy grande. En la práctica no siempre se tiene la posibilidad de contar con muestras muy grandes, por lo que es necesario aproximar la curva de influencia para una muestra finita. A continuación se mencionan cuatro aproximaciones de la curva de influencia para $\widehat{\boldsymbol{\beta}}$.

La Curva de Influencia Empírica basada en n Observaciones. Se obtiene a partir de la curva de influencia para $\widehat{\boldsymbol{\beta}}$, aproximando $\boldsymbol{F}$ por la función de distribución acumulada empírica $\widehat{\boldsymbol{F}}$, y sustituyendo $(\boldsymbol{x}_i^{\boldsymbol{T}}, y_i)$, $n^{-1}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})$ y $\widehat{\boldsymbol{\beta}}$ por $(\boldsymbol{x}^{\boldsymbol{T}}, y)$, $\Sigma_{xx}(\boldsymbol{F})$ y $\widehat{\boldsymbol{\beta}}(\boldsymbol{F})$, respectivamente, obteniéndose

$$\boldsymbol{EIC}_i = n(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}) = n(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_i\widehat{\epsilon_i}, \quad i = 1, 2, ..., n$$

La Curva de Influencia Muestral. Se obtiene de la ecuación (3.12), considerando las igualdades $(\boldsymbol{x}^{\boldsymbol{T}}, y) = (\boldsymbol{x}_i^{\boldsymbol{T}}, y_i)$, $\boldsymbol{F} = \widehat{\boldsymbol{F}}$, $\epsilon = -(n-1)^{-1}$ y omitiendo el límite, resulta

$$\begin{aligned} \boldsymbol{SIC}_i &= -(n-1)\left(\boldsymbol{T}\left(\frac{n}{n-1}\widehat{\boldsymbol{F}} + \frac{-1}{n-1}\delta_{\boldsymbol{x}_i^{\boldsymbol{T}}, y_i}\right) - \boldsymbol{T}(\widehat{\boldsymbol{F}})\right) \\ &= (n-1)\left(\boldsymbol{T}(\widehat{\boldsymbol{F}}) - \boldsymbol{T}(\widehat{\boldsymbol{F}}_{(i)})\right), \end{aligned}$$

donde $\widehat{\boldsymbol{F}}_{(i)}$ es la función de distribución acumulada empírica cuando la i-ésima observación es omitida. Tomando ahora las igualdades $\boldsymbol{T}(\widehat{\boldsymbol{F}}) = \widehat{\boldsymbol{\beta}}$ y $\boldsymbol{T}(\widehat{\boldsymbol{F}}_{(i)}) = \widehat{\boldsymbol{\beta}}_{(i)}$, se tiene que $\boldsymbol{SIC}_i$ se puede escribir simplemente como

$$\boldsymbol{SIC}_i = (n-1)(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}), \tag{3.13}$$

donde $\widehat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{X}_{(i)}^{\boldsymbol{T}}\boldsymbol{X}_{(i)})^{-1}\boldsymbol{X}_{(i)}^{\boldsymbol{T}}\boldsymbol{Y}_{(i)}$ es el estimador de $\boldsymbol{\beta}$ cuando la i-ésima observación es omitida. Al usar la ecuación (2.4), se obtiene

$$\widehat{\boldsymbol{\beta}}_{(i)} = \left((\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}}{1 - \boldsymbol{p}_{ii}}\right)\boldsymbol{X}_{(i)}^{\boldsymbol{T}}\boldsymbol{Y}_{(i)} \tag{3.14}$$

También, notando que

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\begin{bmatrix}\boldsymbol{X}_{(i)}^{\boldsymbol{T}} & \boldsymbol{x}_i\end{bmatrix}\begin{bmatrix}\boldsymbol{Y}_{(i)}\\ \boldsymbol{y}_i\end{bmatrix}\\ &= (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_{(i)}^{\boldsymbol{T}}\boldsymbol{Y}_{(i)} + (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_i\,\boldsymbol{y}_i,\end{aligned}$$

despejando, se obtiene

$$(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_{(i)}^{\boldsymbol{T}}\boldsymbol{Y}_{(i)} = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{y}_i, \tag{3.15}$$

sustituyendo la ecuación (3.15) en la ecuación (3.14), y simplificando, se obtiene

$$(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}) = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_i\frac{\widehat{\boldsymbol{\epsilon}}_i}{1-\boldsymbol{p}_{ii}}, \tag{3.16}$$

si se usa esta identidad, entonces el $\boldsymbol{SIC}_i$ dado en la ecuación (3.13), puede ser escrito como

$$\boldsymbol{SIC}_i = (n-1)(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_i\frac{\widehat{\boldsymbol{\epsilon}}_i}{1-\boldsymbol{p}_{ii}}, \quad i = 1, 2, ..., n \tag{3.17}$$

<u>La Curva de Sensibilidad</u>. Considerando $\boldsymbol{F} = \widehat{\boldsymbol{F}}_{(i)}$ y $\epsilon = n^{-1}$ y omitiendo el límite en la ecuación (3.12), resulta

$$\boldsymbol{SC}_i = n\left(\boldsymbol{T}\left(\frac{n-1}{n}\widehat{\boldsymbol{F}}_{(i)} + \frac{1}{n}\delta_{\boldsymbol{x}_i^{\boldsymbol{T}},\boldsymbol{y}_i}\right) - \boldsymbol{T}(\widehat{\boldsymbol{F}}_{(i)})\right) = n\left(\boldsymbol{T}(\widehat{\boldsymbol{F}}) - \boldsymbol{T}(\widehat{\boldsymbol{F}}_{(i)})\right),$$

si nuevamente se consideran las identidades $\boldsymbol{T}(\widehat{\boldsymbol{F}}) = \widehat{\boldsymbol{\beta}}$ y $\boldsymbol{T}(\widehat{\boldsymbol{F}}_{(i)}) = \widehat{\boldsymbol{\beta}}_{(i)}$, se llega a

$$\boldsymbol{SC}_i = n(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}) = n(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_i\frac{\widehat{\boldsymbol{\epsilon}}_i}{1-\boldsymbol{p}_{ii}}, \quad i = 1, 2, ..., n \tag{3.18}$$

<u>La Curva de Influencia Empírica basada en $(n-1)$ Observaciones</u>. De la misma manera que las demás, ésta también se obtiene a partir de la curva de influencia para $\widehat{\boldsymbol{\beta}}$. Sólo que ahora tomando las siguientes igualdades, $\boldsymbol{F} = \widehat{\boldsymbol{F}}_{(i)}$, $(\boldsymbol{x}^{\boldsymbol{T}}, \boldsymbol{y}) = (\boldsymbol{x}_i^{\boldsymbol{T}}, \boldsymbol{y}_i)$, $\Sigma_{xx}(\boldsymbol{F}) = (n-1)^{-1}(\boldsymbol{X}_{(i)}^{\boldsymbol{T}}\boldsymbol{X}_{(i)})$ y $\widehat{\boldsymbol{\beta}}(\boldsymbol{F}) = \widehat{\boldsymbol{\beta}}_{(i)}$, y entonces se produce

$$\boldsymbol{EIC}_{(i)} = (n-1)(\boldsymbol{X}_{(i)}^{\boldsymbol{T}}\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i(\boldsymbol{y}_i - \boldsymbol{x}_i^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}}_{(i)}), \quad i = 1, 2, ..., n, \tag{3.19}$$

a la cantidad $(y_i - x_i^T \widehat{\beta}_{(i)})$, se le llama *residual predicho*, la cual, utilizando las ecuaciones (2.4) y (3.15), puede escribirse como

$$(y_i - x_i^T \widehat{\beta}_{(i)}) = \frac{\widehat{\epsilon}_i}{1 - p_{ii}}$$

Sustituyendo esta última expresión en la ecuación (3.19) y utilizando nuevamente la ecuación (2.4), se obtiene

$$EIC_{(i)} = (n-1)(X^T X)^{-1} x_i \frac{\widehat{\epsilon}_i}{(1-p_{ii})^2}, \quad i = 1, 2, ..., n \tag{3.20}$$

Notando que EIC_i es el menos sensible a puntos palanca, mientras que $EIC_{(i)}$ es el más sensible. También notando que SIC_i y SC_i pueden considerarse equivalentes, pues son proporcionales a la distancia $(\widehat{\beta} - \widehat{\beta}_{(i)})$.

La curva de influencia para $\widehat{\sigma}^2$ dada en el Teorema 13(2), es un escalar. Sin embargo, la curva de influencia para $\widehat{\beta}$ dada en el Teorema 13(1), así como sus aproximaciones, son vectores; por lo tanto, es necesario reducir esta curva de influencia a una cantidad escalar y que las observaciones puedan ser ordenadas de acuerdo a su influencia sobre $\widehat{\beta}$ o una función lineal de $\widehat{\beta}$. Para realizar tal reducción se puede utilizar,

$$\sup_{x^T, y} \| \psi \left(x^T, y, F, \widehat{\beta}(F) \right) \|,$$

esta expresión es considerada por Hampel (1974), como una medida de la influencia máxima posible de cualquier observación sobre los coeficientes estimados, y la llama el *error de sensibilidad*, pues tiene la desventaja de no ser invariante bajo transformaciones de parámetros de localización y escala.

Una alternativa al error de sensibilidad, es del tipo

$$\sup_{x^T, y} \sup_a \frac{| a^T \psi \left(x^T, y, F, \widehat{\beta}(F) \right) |}{\sqrt{a^T \Sigma_\psi a}},$$

la cual, es invariante bajo transformaciones no singulares de X, y mide la influencia máxima de cualquier observación sobre cualquier combinación lineal de los

coeficientes relativos al error estándar de la combinación lineal $\sqrt{a^{\boldsymbol{T}}\Sigma_{\psi}a}$, donde $\Sigma_{\psi}=\int \psi(\boldsymbol{z},\boldsymbol{F},\boldsymbol{T})\psi^{\boldsymbol{T}}(\boldsymbol{z},\boldsymbol{F},\boldsymbol{T})d\boldsymbol{F}(\boldsymbol{z})$. Se ve que si a se restringe a $\boldsymbol{x}$, entonces se convierte en una medida de la influencia de cualquier observación sobre los valores predichos.

Otra cantidad equivalente a la anterior es

$$sup_{\boldsymbol{x}\boldsymbol{T},y}\frac{\psi^{\boldsymbol{T}}\left(\boldsymbol{x}^{\boldsymbol{T}},y,\boldsymbol{F},\widehat{\boldsymbol{\beta}}(\boldsymbol{F})\right)M\psi\left(\boldsymbol{x}^{\boldsymbol{T}},y,\boldsymbol{F},\widehat{\boldsymbol{\beta}}(\boldsymbol{F})\right)}{c}$$

En la práctica, lo que interesa es encontrar la influencia máxima posible de cualquier observación sobre los resultados de la regresión, ordenando las n observaciones de acuerdo a su nivel de influencia, para lo cual es necesario hacer una apropiada elección de M y c, por lo tanto, se puede usar

$$\boldsymbol{D_i}(M,c)=\frac{\psi^{\boldsymbol{T}}\left(\boldsymbol{x_i^T},\boldsymbol{y_i},\boldsymbol{F},\widehat{\boldsymbol{\beta}}(\boldsymbol{F})\right)M\psi\left(\boldsymbol{x_i^T},\boldsymbol{y_i},\boldsymbol{F},\widehat{\boldsymbol{\beta}}(\boldsymbol{F})\right)}{c}, \tag{3.21}$$

para considerar la influencia de la i-ésima observación sobre los coeficientes de regresión relativos a M y c. De manera que si se registra un valor grande en $\boldsymbol{D_i}(M,c)$, este indica que la i-ésima observación tiene una fuerte influencia sobre $\widehat{\boldsymbol{\beta}}$ relativa a M y c.

Se presentan cuatro alternativas de M y c, en $\boldsymbol{D_i}(M,c)$, las cuales son, las distancias de Cook, de Welsch-Kuh, de Welsch y una modificación a la primera.

Distancia de Cook

Para definir esta distancia, es necesario considerar que, bajo normalidad, la región de confianza $100(1-\alpha)\%$ para $\boldsymbol{\beta}$, se puede obtener a partir de

$$\frac{(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})}{k\widehat{\sigma}^2}\leq \boldsymbol{F}(\alpha;k,n-k),$$

donde $\boldsymbol{F}(\alpha;k,n-k)$ es el punto superior α de la distribución $\boldsymbol{F}$ centrada con $(k,n-k)$ grados de libertad. Esta desigualdad, define una región elipsoidal centrada

en $\widehat{\boldsymbol{\beta}}$, y permite medir la influencia de la i-ésima observación por el cambio en el centro del elipsoide de confianza, cuando la i-ésima observación es omitida.

En forma análoga, Cook (1977), propone la medida

$$\boldsymbol{C_i} = \frac{(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})})^{\boldsymbol{T}}(\boldsymbol{X^T X})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})})}{k\widehat{\sigma}^2}, \quad i = 1, 2, ..., n, \tag{3.22}$$

para considerar la influencia de la i-ésima observación sobre el centro del elipsoide de confianza, es decir, sobre $\widehat{\boldsymbol{\beta}}$. Esta medida se conoce como la *distancia de Cook*, y es simplemente, una distancia a escala entre $\widehat{\boldsymbol{\beta}}$ y $\widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})}$.

Bingham (1977), señala que $\boldsymbol{C_i}$, también puede ser escrita como

$$\boldsymbol{C_i} = \frac{(\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{(\boldsymbol{i})})^{\boldsymbol{T}}(\widehat{\boldsymbol{Y}} - \widehat{\boldsymbol{Y}}_{(\boldsymbol{i})})}{k\widehat{\sigma}^2}, \quad i = 1, 2, ..., n$$

donde $\widehat{\boldsymbol{Y}}_{(\boldsymbol{i})}$=$\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})}$ es el vector de valores predichos, cuando se realiza una regresión de $\widehat{\boldsymbol{Y}}_{(\boldsymbol{i})}$ con $\widehat{\boldsymbol{X}}_{(\boldsymbol{i})}$. Por lo anterior $\boldsymbol{C_i}$, se puede interpretar como la distancia euclideana a escala, entre los vectores de valores predichos, incluyendo y excluyendo la i-ésima observación, cuando se hace el ajuste.

A simple vista, pareciera que para encontrar $\boldsymbol{C_i}$, $i = 1, 2, ..., n$, se requiere realizar $(n + 1)$ análisis de regresión, uno utilizando todos los datos, y n usando los datos sin cada observación. Sin embargo, si se sustituye la ecuación (3.16) en la ecuación (3.22), se produce

$$\boldsymbol{C_i} = \frac{\boldsymbol{x_i^T}(\boldsymbol{X^T X})^{-1}(\boldsymbol{X^T X})(\boldsymbol{X^T X})^{-1}\boldsymbol{x_i}}{k(1 - \boldsymbol{p_{ii}})} \frac{\widehat{\boldsymbol{\epsilon_i}}^2}{\widehat{\sigma}^2(1 - \boldsymbol{p_{ii}})} = \frac{1}{k}\frac{\boldsymbol{p_{ii}}}{(1 - \boldsymbol{p_{ii}})}\boldsymbol{r_i}^2 \tag{3.23}$$

En la ecuación (3.23), se observa que se combinan dos medidas, $\boldsymbol{p_{ii}}$ y $\boldsymbol{r_i}$, los cuales, como se sabe, proporcionan información sobre puntos palanca y "outliers", respectivamente.

Observando además que $\frac{\boldsymbol{p_{ii}}}{(1-\boldsymbol{p_{ii}})}$, es la relación dada entre la varianza del i-ésimo valor predicho $\boldsymbol{Var}(\widehat{\boldsymbol{y_i}})$=$\sigma^2\boldsymbol{p_{ii}}$, y la varianza del i-ésimo residual

$$\boldsymbol{Var}(\widehat{\boldsymbol{\epsilon_i}}) = \sigma^2(1 - \boldsymbol{p_{ii}}).$$

Cook y Weisberg (1982), se refieren a esta relación, como el potencial de la i-ésima observación en la determinación de $\widehat{\boldsymbol{\beta}}$, con $M = \boldsymbol{X^T X}$ y $c = (n-1)^2 k\widehat{\sigma}^2$. Otra interpretación de esta relación, es dada por Huber (1981), para la cual se ve que si se utilizan la ecuación (3.16) y el residual predicho antes descrito, $\widehat{y_i}$ se puede escribir como

$$\begin{aligned}
\widehat{y_i} &= \boldsymbol{x_i^T}\widehat{\boldsymbol{\beta}} \\
&= \boldsymbol{x_i^T}\left(\frac{\widehat{\epsilon}_i^2}{(1-\boldsymbol{p_{ii}})}(\boldsymbol{X^T X})^{-1}\boldsymbol{x_i} + \widehat{\boldsymbol{\beta}}_{(i)}\right) \\
&= \boldsymbol{p_{ii}}\frac{\widehat{\epsilon}_i}{(1-\boldsymbol{p_{ii}})} + \boldsymbol{x_i^T}\widehat{\boldsymbol{\beta}}_{(i)} \\
&= \boldsymbol{p_{ii}}\left(\boldsymbol{y_i} - \boldsymbol{x_i^T}\widehat{\boldsymbol{\beta}}_{(i)}\right) + \boldsymbol{x_i^T}\widehat{\boldsymbol{\beta}}_{(i)} \\
&= (1-\boldsymbol{p_{ii}})\boldsymbol{x_i^T}\widehat{\boldsymbol{\beta}}_{(i)} + \boldsymbol{p_{ii}}\boldsymbol{y_i}
\end{aligned}$$

De esta forma, $\frac{\boldsymbol{p_{ii}}}{(1-\boldsymbol{p_{ii}})}$, puede también ser considerada como la relación que existe entre la parte de $\widehat{y_i}$ debida a $\boldsymbol{y_i}$ y la parte debida al valor predicho $\boldsymbol{x_i^T}\widehat{\boldsymbol{\beta}}_{(i)}$.

Es claro ver en la ecuación (3.23), que $\boldsymbol{C_i}$ será grande si $\boldsymbol{p_{ii}}$ es grande y $\boldsymbol{r_i}^2$ también lo es.

Se ve que $\boldsymbol{C_i}$ es una función monótona de $\boldsymbol{r_i}^2$, de manera que si se cumplen las condiciones del Teorema 12(1) y utilizando la expresión

$$\boldsymbol{r_i}^* = \boldsymbol{r_i}\sqrt{\frac{n-k-1}{n-k-\boldsymbol{r_i}^2}},$$

entonces, resulta que

$$\frac{\boldsymbol{r_i}^2(n-k-1)}{n-k-\boldsymbol{r_i}^2} \sim \boldsymbol{F}(1, n-k-1),$$

lo cual indica que $\boldsymbol{C_i}$ no tiene estrictamente una distribución $\boldsymbol{F}$.

La distancia de Cook, puede ser obtenida directamente de la curva de influencia para $\widehat{\boldsymbol{\beta}}$ dada en el Teorema 13(1), si ésta se aproxima por la curva de influencia muestral $\boldsymbol{SIC_i}$ dada en la ecuación (3.13) y utilizando para ello la ecuación

(3.21), entonces C_i puede expresarse como

$$C_i = D_i\left(X^T X, k\hat{\sigma}^2(n-1)^{-2}\right)$$

Distancia de Welsch-Kuh

La influencia de la i-ésima observación sobre los valores predichos $\widehat{y_i}$, puede ser medida por el cambio en la predicción en x_i, cuando la i-ésima observación es omitida, relativa al error estándar de $\widehat{y_i}$, esto es,

$$\frac{|\widehat{y} - \widehat{y}_{i(i)}|}{\sigma\sqrt{p_{ii}}} = \frac{|x_i^T(\widehat{\beta} - \widehat{\beta}_{(i)})|}{\sigma\sqrt{p_{ii}}} \tag{3.24}$$

Welsch y Kuh (1977), Welsch y Peters (1978) y Belsley *et al.* (1980), prefieren utilizar $\widehat{\sigma}_{(i)}$ como un estimador de σ en la ecuación (3.24). Si se usan las ecuaciones (3.16) y (3.5), entonces la ecuación (3.24) puede ser escrita como

$$WK_i = \frac{|\frac{\widehat{\epsilon}_i}{(1-p_{ii})}x_i^T(X^T X)^{-1}x_i|}{\widehat{\sigma}_{(i)}\sqrt{p_{ii}}} = |r_i^*|\sqrt{\frac{p_{ii}}{1-p_{ii}}} \tag{3.25}$$

Valores grandes de WK_i, indican que la i-ésima observación influye sobre el ajuste del modelo.

Del mismo modo que la distancia de Cook, puede mencionarse que aunque la distancia de Welsch-Kuh sea un estadístico parecido a t, no tiene estrictamente una distribución t. Pero si se dan las condiciones del Teorema 12(1), entonces $r_i^* \sim t(n-k-1)$, lo cual hace que pueda sugerirse como un punto de corte a

$$t\sqrt{\frac{k}{n-k}}$$

Belsley *et al.* (1980), recomiendan usar $2\sqrt{\frac{k}{n}}$, como punto de corte para WK_i, pero parece ser más apropiado utilizar $2\sqrt{\frac{k}{n-k}}$.

La distancia de Welsch-Kuh, también puede ser obtenida directamente de la curva de influencia para $\widehat{\beta}$ dada en el Teorema 13(1), si ésta se aproxima por la

curva de influencia muestral $\boldsymbol{SIC_i}$ dada en la ecuación (3.17) y utilizando para ello la ecuación (3.21), entonces $\boldsymbol{WK_i}$ puede expresarse como

$$\boldsymbol{WK_i} = \sqrt{\boldsymbol{D_i}\left(\boldsymbol{X^T X}, (n-1)\widehat{\sigma}^2_{(i)}\right)}$$

Si en vez de aproximar con la curva de influencia muestral, se aproxima con la curva de sensibilidad, dada en la ecuación (3.18), entonces $\boldsymbol{WK_i}$ se escribe así,

$$\boldsymbol{WK_i} = \sqrt{\boldsymbol{D_i}\left(\boldsymbol{X^T X}, n\widehat{\sigma}^2_{(i)}\right)}$$

Distancia de Welsch

Si se utiliza la curva de influencia empírica basada en $(n-1)$ observaciones, la cual está dada en la ecuación (3.20), como una aproximación a la curva de influencia para $\widehat{\boldsymbol{\beta}}$, y considerando además que, $M = \boldsymbol{X^T_{(i)} X_{(i)}} = (\boldsymbol{X^T X} - \boldsymbol{x_i x_i^T})$ y $c = (n-1)\widehat{\sigma}^2_{(i)}$, entonces la ecuación (3.21) se convierte en

$$\begin{aligned}
\boldsymbol{W_i}^2 &= \boldsymbol{D_i}\left(\boldsymbol{X^T_{(i)} X_{(i)}}, (n-1)\widehat{\sigma}^2_{(i)}\right) \\
&= (n-1)\frac{\widehat{\epsilon}^2_{\boldsymbol{i}}}{\widehat{\sigma}^2_{(i)}(1-\boldsymbol{p_{ii}})^4}\boldsymbol{x_i^T}(\boldsymbol{X^T X})^{-1}(\boldsymbol{X^T X} - \boldsymbol{x_i x_i^T})(\boldsymbol{X^T X})^{-1}\boldsymbol{x_i} \\
&= (n-1)\boldsymbol{r_i}^{*2}\frac{\boldsymbol{p_{ii}}}{(1-\boldsymbol{p_{ii}})^2} \qquad (3.26)
\end{aligned}$$

Al comparar las ecuaciones (3.25) y (3.26), resulta

$$\boldsymbol{W_i} = \boldsymbol{WK_i}\sqrt{\frac{n-1}{1-\boldsymbol{p_{ii}}}} \qquad (3.27)$$

Se observa claramente en la ecuación (3.27), que $\boldsymbol{W_i}$ da un mayor énfasis a los puntos palanca que $\boldsymbol{WK_i}$, sin embargo, algunos autores prefieren ésta última, por considerarla más fácil de interpretar. Asimismo, en la ecuación (3.27) se puede ver que un punto de corte para $\boldsymbol{W_i}$, puede obtenerse al multiplicar el punto de corte para $\boldsymbol{WK_i}$ por $\left(\frac{n(n-1)}{(n-k)}\right)^{\frac{1}{2}}$; entonce, si el punto de corte para $\boldsymbol{WK_i}$ es $2\sqrt{\frac{k}{n-k}}$, el correspondiente punto de corte para $\boldsymbol{W_i}$ es $\frac{2}{(n-k)}\sqrt{kn(n-1)}$, sin embargo, si n es grande comparada con k, esta cantidad se convierte en $3\sqrt{k}$.

Distancia de Cook Modificada

Para la detección de observaciones influyentes, Atkinson (1981), propone usar una versión modificada de la distancia de Cook. Esta modificación implica reemplazar $\widehat{\sigma}^2$ por $\widehat{\sigma}^2_{(i)}$, tomar la raíz cuadrada de $\boldsymbol{C_i}$ y ajustarlo por tamaño de muestra, lo cual produce

$$\begin{aligned} \boldsymbol{C_i}^* &= \sqrt{\boldsymbol{D_i}\left(\boldsymbol{X^T X}, \frac{k(n-1)^2}{(n-k)}\widehat{\sigma}^2_{(i)}\right)} \\ &= |\boldsymbol{r_i}^*|\sqrt{\frac{\boldsymbol{p_{ii}}}{(1-\boldsymbol{p_{ii}})}\frac{(n-k)}{k}} \\ &= \boldsymbol{WK_i}\sqrt{\frac{n-k}{k}} \end{aligned} \tag{3.28}$$

Atkinson (1981), afirma que esta modificación, mejora a $\boldsymbol{C_i}$, ya que $\boldsymbol{C_i}^*$ proporciona mayor énfasis a puntos extremos.

Si se utilizan diferentes aproximaciones a la curva de influencia para $\widehat{\boldsymbol{\beta}}$, en combinación con varias elecciones de M y c de la ecuación (3.21), se pueden obtener algunas mediciones de la influencia de la i-ésima observación sobre los coeficientes de regresión; entonces, la diferencia fundamental en $\boldsymbol{C_i}$, $\boldsymbol{WK_i}$, $\boldsymbol{W_i}$ y $\boldsymbol{C_i}^*$, estriba en la elección de M y c. Así, $\boldsymbol{C_i}$ usa $\boldsymbol{X^T X}$ y $\widehat{\sigma}^2$; $\boldsymbol{WK_i}$ y $\boldsymbol{C_i}^*$ usan $\boldsymbol{X^T X}$ y $\widehat{\sigma}^2_{(i)}$, y $\boldsymbol{W_i}$ usa $(\boldsymbol{X^T_{(i)} X_{(i)}})$ y $\widehat{\sigma}^2_{(i)}$, entonces se puede concluir que $\boldsymbol{C_i}$ mide la influencia de la i-ésima observación solamente sobre $\widehat{\boldsymbol{\beta}}$, mientras que $\boldsymbol{WK_i}$, $\boldsymbol{W_i}$ y $\boldsymbol{C_i}^*$ miden la influencia sobre $\widehat{\boldsymbol{\beta}}$ y sobre $\widehat{\sigma}^2$.

Medidas Basadas en el Volumen de Elipsoides de Confianza

Hasta el momento, se ha visto que las medidas de diagnóstico basadas en la curva de influencia, se pueden interpretar como medidas que se basan en el cambio en el centro del elipsoide de confianza, cuando la i-ésima observación es omitida; una alternativa, la constituyen, medidas de la influencia de la i-ésima observación, que

se basan en el cambio en el volumen del elipsoide de confianza, cuando la i-ésima observación es omitida. A continuación se describen tres de estas medidas.

Estadístico Andrews-Pregibon

El volumen de la región de confianza para $\boldsymbol{\beta}$, como se ha visto, se obtiene de

$$\frac{(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})}{k\widehat{\sigma}^2} \leq \boldsymbol{F}(\alpha; k, n-k),$$

la cual, es inversamente proporcional a la raíz cuadrada del $\boldsymbol{det}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})$. Así, la influencia de la i-ésima observación sobre el volumen de los elipsoides de confianza, puede ser medida comparando el $\boldsymbol{det}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})$ con el $\boldsymbol{det}(\boldsymbol{X}_{(i)}^{\boldsymbol{T}}\boldsymbol{X}_{(i)})$. Por otra parte, al omitir una observación con un residual grande, ocasionará una fuerte reducción en la suma de cuadrados de residuales $\boldsymbol{SSE}$. De tal forma que, la influencia de la i-ésima observación puede ser medida combinando éstas dos ideas, y calculando el cambio en $\widehat{\epsilon}^{\,T}\widehat{\epsilon}$ y en el $\boldsymbol{det}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})$ cuando la i-ésima observación es omitida. Por lo que Andrews y Pregibon (1978), sugieren la relación

$$\frac{\boldsymbol{SSE}_{(i)}\boldsymbol{det}(\boldsymbol{X}_{(i)}^{\boldsymbol{T}}\boldsymbol{X}_{(i)})}{\boldsymbol{SSE}\boldsymbol{det}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})}, \quad i = 1, 2, ..., n. \tag{3.29}$$

El cociente dado en la ecuación (3.29), se puede simplicar de la siguiente forma, definiendo la matriz aumentada $\boldsymbol{Z} = (\boldsymbol{X} : \boldsymbol{Y})$, y se ve que

$$\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} & \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y} \\ \boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{X} & \boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{Y} \end{bmatrix}$$

al utilizar el Lema 3(1), se tiene que

$$\begin{aligned} \boldsymbol{det}(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{Z}) &= \boldsymbol{det}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})\boldsymbol{det}\left(\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{Y} - \boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y}\right) \\ &= \boldsymbol{det}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})\boldsymbol{det}\left(\boldsymbol{Y}^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}\right) \\ &= \boldsymbol{det}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})\boldsymbol{SSE} \end{aligned}$$

Equivalentemente, se tiene

$$\boldsymbol{det}(\boldsymbol{Z}_{(i)}^{\boldsymbol{T}}\boldsymbol{Z}_{(i)}) = \boldsymbol{det}(\boldsymbol{X}_{(i)}^{\boldsymbol{T}}\boldsymbol{X}_{(i)})\boldsymbol{SSE}_{(i)}$$

Entonces, la ecuación dada en (3.29), se convierte en

$$\frac{SSE_{(i)}det(X_{(i)}^T X_{(i)})}{SSEdet(X^T X)} = \frac{det(Z_{(i)}^T Z_{(i)})}{det(Z^T Z)},$$

la cual mide el cambio relativo en el $det(Z^T Z)$, debido a la eliminación de la i-ésima observación, es decir, la proporción del volumen generado por Z que no es debida a la i-ésima observación. Si se omite una observación que está lejos del centro de los datos, resultará en una reducción en el determinante y un incremento en el volumen; por lo que, valores pequeños en la ecuación (3.29) requieren una atención especial. Por conveniencia, se ha definido a

$$AP_i = 1 - \frac{det(Z_{(i)}^T Z_{(i)})}{det(Z^T Z)} \tag{3.30}$$

por lo que ahora, los que requieren de atención especial, son los valores grandes de la ecuación (3.30). Por otro lado, si en el Lema 4, se considera que $A = Z^T Z$ y $B = C = z_i$, resulta que

$$\begin{aligned} det(Z_{(i)}^T Z_{(i)}) &= det(Z^T Z - z_i z_i^T) \\ &= det(Z^T Z)\left(1 - z_i^T (Z^T Z)^{-1} z_i\right) \\ &= det(Z^T Z)(1 - p_{z_{ii}}) \end{aligned}$$

Si este resultado se sustituye en la ecuación (3.30), resulta

$$AP_i = p_{z_{ii}},$$

puede concluirse entonces que, AP_i no distingue entre puntos palanca en el espacio X y "outliers" en el espacio Z. Asimismo, se ve que AP_i es el i-ésimo elemento de la diagonal de la matriz de predicción P_Z, se sigue del Teorema 5(1), que $0 \leq AP_i \leq 1$. Recordando que

$$p_{z_{ii}} = z_i^T (Z^T Z)^{-1} z_i = p_{ii} + \frac{\hat{\epsilon}_i^2}{\hat{\epsilon}^T \hat{\epsilon}},$$

entonces la relación entre AP_i, r_i y p_{ii}, está dada por

$$AP_i = p_{ii} + \frac{\hat{\epsilon}_i^2}{\hat{\epsilon}^T \hat{\epsilon}} = p_{ii} + (1 - p_{ii})\frac{r_i^2}{(n-k)},$$

la cual, también puede escribirse como

$$(1 - \boldsymbol{AP_i}) = (1 - \boldsymbol{p_{ii}}) \left(1 - \frac{\boldsymbol{r_i}^2}{(n-k)}\right)$$

Como señalan Draper y John (1981), $(1 - \boldsymbol{AP_i})$ es un producto de dos cantidades, la primera identifica puntos palanca y la segunda detecta "outliers", por lo que resulta más informativa que $\boldsymbol{AP_i}$.

Relación de Varianzas

Constituye una alternativa al estadístico Andrews-Pregibon, para medir la influencia de la i-ésima observación, comparando la varianza estimada de $\widehat{\boldsymbol{\beta}}$ dada por $\widehat{\sigma}^2(\boldsymbol{X^T X})^{-1}$, y la varianza estimada de $\widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})}$ dada por $\widehat{\sigma}^2_{(i)}(\boldsymbol{X^T_{(i)} X_{(i)}})^{-1}$. Si el $\boldsymbol{r(X_{(i)})} = k$, estas matrices son positivas definidas, las cuales, se comparan mediante la relación de sus trazas o sus determinantes. Belsley *et al.* (1980), prefieren utilizar la relación entre sus determinantes, esto es,

$$\begin{aligned} \boldsymbol{VR_i} &= \frac{\boldsymbol{det}\left(\widehat{\sigma}^2_{(i)}(\boldsymbol{X^T_{(i)} X_{(i)}})^{-1}\right)}{\boldsymbol{det}\left(\widehat{\sigma}^2(\boldsymbol{X^T X})^{-1}\right)} \\ &= \left(\frac{\widehat{\sigma}^2_{(i)}}{\widehat{\sigma}^2}\right)^k \frac{\boldsymbol{det}(\boldsymbol{X^T X})}{\boldsymbol{det}(\boldsymbol{X^T_{(i)} X_{(i)}})}, \quad i = 1, 2, ..., n \end{aligned} \tag{3.31}$$

Observando ahora que, esta relación de determinantes tiene que ver con $\boldsymbol{r_i}$ y $\boldsymbol{p_{ii}}$, para lo cual, definiendo $\boldsymbol{A} = \boldsymbol{X^T X}$ y $\boldsymbol{B} = \boldsymbol{C} = \boldsymbol{x_i}$ y utilizando el Lema 4, resulta que

$$\begin{aligned} \boldsymbol{det}(\boldsymbol{X^T_{(i)} X_{(i)}}) &= \boldsymbol{det}(\boldsymbol{X^T X} - \boldsymbol{x_i x_i^T}) \\ &= \boldsymbol{det}(\boldsymbol{X^T X})\left(1 - \boldsymbol{x_i^T}(\boldsymbol{X^T X})^{-1}\boldsymbol{x_i}\right) \\ &= \boldsymbol{det}(\boldsymbol{X^T X})(1 - \boldsymbol{p_{ii}}). \end{aligned} \tag{3.32}$$

Sustituyendo las ecuaciones (3.7) y (3.32) en la ecuación (3.31), se produce

$$\boldsymbol{VR_i} = \left(\frac{n - k - \boldsymbol{r_i}^2}{n - k - 1}\right)^k \frac{1}{(1 - \boldsymbol{p_{ii}})} \tag{3.33}$$

Un análisis de la ecuación (3.33), indica que $\boldsymbol{VR_i} > 1$, cuando $\boldsymbol{r_i}^2$ es pequeño y $\boldsymbol{p_{ii}}$ es grande, mientras que si $\boldsymbol{r_i}^2$ es grande y $\boldsymbol{p_{ii}}$ es pequeño, entonces $\boldsymbol{VR_i} < 1$; pero si $\boldsymbol{r_i}^2$ y $\boldsymbol{p_{ii}}$ son grandes o pequeños, entonces $\boldsymbol{VR_i}$ se aproxima a 1. En la práctica, se ha observado que, $\boldsymbol{VR_i}$ elige exitosamente observaciones influyentes.

Lo ideal ocurre cuando todas las observaciones tienen igual influencia sobre la matriz de covarianzas, en este caso, $\boldsymbol{VR_i}$ es aproximadamente igual a 1, cualquier desviación de la unidad, indica que la i-ésima observación es potencialmente influyente. Belsley *et al.* (1980), proporcionan dos aproximaciones de puntos de corte para $\boldsymbol{VR_i}$, que son:

1. Cuando $|\boldsymbol{r_i}| \geq 2$ con $\boldsymbol{p_{ii}} = \frac{1}{n}$, se produce aproximadamente

$$\boldsymbol{VR_i} \leq 1 - \frac{3k}{n-k},$$

el cual, es útil sólo cuando $n > 4k$.

2. Cuando $\boldsymbol{r_i} = 0$ con $\boldsymbol{p_{ii}} \geq \frac{2k}{n}$, resulta aproximadamente

$$\boldsymbol{VR_i} \geq 1 + \frac{3k}{n-k}.$$

Estos autores, reemplazan $(n-k)$ por n y ponen estos limites de la forma

$$|\boldsymbol{VR_i} - 1| \geq \frac{3k}{n}$$

Estadístico Cook-Weisberg

Bajo normalidad, el $100(1-\alpha)\%$ del elipsoide de confianza conjunta para $\boldsymbol{\beta}$ es dado por

$$E = \left\{ \boldsymbol{\beta} : \frac{(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^{\boldsymbol{T}} (\boldsymbol{X^T X})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})}{k\widehat{\sigma}^2} \leq \boldsymbol{F}(\alpha; k, n-k) \right\},$$

pero, cuando la i-ésima observación es omitida, E se convierte en

$$E_{(i)} = \left\{ \boldsymbol{\beta} : \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})^{\boldsymbol{T}} (\boldsymbol{X}_{(i)}^{\boldsymbol{T}} \boldsymbol{X}_{(i)})(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{k\widehat{\sigma}^2_{(i)}} \leq \boldsymbol{F}(\alpha; k, n-k-1) \right\}$$

Cook y Weisberg (1980), proponen el logaritmo de la relación entre E y $E_{(i)}$ como una medida de la influencia de la i-ésima observación sobre el volumen del elipsoide de confianza para $\boldsymbol{\beta}$, así, se tiene

$$\boldsymbol{CW_i} = log\frac{\text{Volumen}(E)}{\text{Volumen}(E_{(i)})}, \quad i = 1, 2, ..., n, \tag{3.34}$$

y como el volumèn de un elipsoide es proporcional a la inversa de la raíz cuadrada del determinante de la matriz de forma cuadrática asociada, entonces la ecuación (3.34), se convierte en

$$\boldsymbol{CW_i} = log\left(\left(\frac{\boldsymbol{det}(\boldsymbol{X_{(i)}^T X_{(i)}})}{\boldsymbol{det}(\boldsymbol{X^T X})}\right)^{\frac{1}{2}}\left(\frac{\hat{\sigma}}{\hat{\sigma}_{(i)}}\right)^{k}\left(\frac{\boldsymbol{F}(\alpha; k, n-k)}{\boldsymbol{F}(\alpha; k, n-k-1)}\right)^{\frac{k}{2}}\right) \tag{3.35}$$

Sustituyendo las ecuaciones (3.7) y (3.32) en la ecuación (3.35), se obtiene

$$\boldsymbol{CW_i} = \frac{1}{2}log(1-\boldsymbol{p_{ii}}) + \frac{k}{2}log\left(\frac{n-k-1}{n-k-\boldsymbol{r_i}^2}\right) + \frac{k}{2}log\left(\frac{\boldsymbol{F}(\alpha; k, n-k)}{\boldsymbol{F}(\alpha; k, n-k-1)}\right) \tag{3.36}$$

Cook y Weisberg (1980), señalan que si $\boldsymbol{CW_i}$ es grande y positivo, entonces la eliminación de la i-ésima observación ocasionará una reducción en volumen, y si es grande y negativo, resultará en un incremento en volumen. Un análisis de la ecuación (3.36), indica que $\boldsymbol{CW_i}$ será grande y positivo, cuando $\boldsymbol{r_i}^2$ es grande y $\boldsymbol{p_{ii}}$ pequeño, y será grande y negativo, cuando $\boldsymbol{r_i}^2$ es pequeño y $\boldsymbol{p_{ii}}$ grande. Pero si $\boldsymbol{r_i}^2$ y $\boldsymbol{p_{ii}}$ son grandes o pequeños, entonces $\boldsymbol{CW_i}$ se aproxima a cero.

Se observa en la ecuación (3.33), que se puede relacionar a $\boldsymbol{CW_i}$ con $\boldsymbol{VR_i}$ por medio de

$$\boldsymbol{CW_i} = -\frac{1}{2}log(\boldsymbol{VR_i}) + \frac{k}{2}log\left(\frac{\boldsymbol{F}(\alpha; k, n-k)}{\boldsymbol{F}(\alpha; k, n-k-1)}\right), \tag{3.37}$$

en donde se observa que $\boldsymbol{CW_i}$ y $\boldsymbol{VR_i}$ se pueden considerar como cantidades equivalentes.

Medidas Basadas en la Función de Verosimilitud

Sea $log(\boldsymbol{\beta}, \sigma^2)$ el logaritmo de la función de verosimilitud basada en las n observaciones; sean $\tilde{\boldsymbol{\beta}}$ y $\tilde{\sigma}^2$ los estimadores de máxima verosimilitud (MLE) de $\boldsymbol{\beta}$ y σ^2, respectivamente; y sea $log(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2)$ el logaritmo de la función de verosimilitud evaluada en $\tilde{\boldsymbol{\beta}}$ y $\tilde{\sigma}^2$. Se sabe que, el $100(1-\alpha)\%$ de la región de confianza asintótica para $\boldsymbol{\beta}$ y σ^2 es dada por

$$\left\{(\boldsymbol{\beta}, \sigma^2) : 2\left(log(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2) - log(\boldsymbol{\beta}, \sigma^2)\right) \leq \chi^2(\alpha; k+1)\right\}, \tag{3.38}$$

donde $\chi^2(\alpha; k+1)$ es el punto superior α de la distribución ji-cuadrada con $(k+1)$ grados de libertad.

Suponiendo ahora que $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, entonces el logaritmo de la función de verosimilitud para $\boldsymbol{\beta}$ y σ^2 es

$$log(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}log2\pi - \frac{n}{2}log\sigma^2 - \frac{(\boldsymbol{Y}-\boldsymbol{X\beta})^{\boldsymbol{T}}(\boldsymbol{Y}-\boldsymbol{X\beta})}{2\sigma^2} \tag{3.39}$$

de la cual se sigue que los MLE de $\boldsymbol{\beta}$ y σ^2 son

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} \quad y \quad \tilde{\sigma}^2 = \hat{\sigma}^2\left(\frac{n-k}{n}\right)$$

Y cuando la i-ésima observación es omitida, los MLE de $\boldsymbol{\beta}$ y σ^2 son

$$\tilde{\boldsymbol{\beta}}_{(\boldsymbol{i})} = \hat{\boldsymbol{\beta}}_{(\boldsymbol{i})} \quad y \quad \tilde{\sigma}^2_{(i)} = \hat{\sigma}^2_{(i)}\left(\frac{n-k-1}{n-1}\right) = \hat{\sigma}^2\left(\frac{n-k-\boldsymbol{r_i}^2}{n-1}\right)$$

Tomando $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ y $\sigma^2 = \tilde{\sigma}^2$ en la ecuación (3.39) y simplificando, se obtiene

$$log(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2) = -\frac{n}{2}log2\pi - \frac{n}{2}log\hat{\sigma}^2\left(\frac{n-k}{n}\right) - \frac{n}{2} \tag{3.40}$$

Similarmente, tomando $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}_{(\boldsymbol{i})}$ y $\sigma^2 = \tilde{\sigma}^2_{(i)}$ en la ecuación (3.39), se produce

$$\begin{aligned} log(\tilde{\boldsymbol{\beta}}_{(i)}, \tilde{\sigma}^2_{(i)}) &= -\frac{n}{2}log2\pi - \frac{n}{2}log\tilde{\sigma}^2_{(i)} - \frac{1}{2\tilde{\sigma}^2_{(i)}}\sum_{r=1}^{n}\left(\boldsymbol{y_r} - \boldsymbol{x_r^T}\tilde{\boldsymbol{\beta}}_{(\boldsymbol{i})}\right)^2 \\ &= -\frac{n}{2}log2\pi - \frac{n}{2}log\tilde{\sigma}^2_{(i)} - \frac{n-1}{2} - \frac{\left(\boldsymbol{y_i} - \boldsymbol{x_i^T}\tilde{\boldsymbol{\beta}}_{(\boldsymbol{i})}\right)^2}{2\tilde{\sigma}^2_{(i)}} \end{aligned} \tag{3.41}$$

La influencia de la i-ésima observación sobre la función de verosimilitud, puede ser medida por la distancia que existe entre las funciones de verosimilitud dadas en las ecuaciones (3.40) y (3.41). Por analogía a la ecuación (3.38), Cook y Weisberg (1982), definen la distancia de verosimilitud por

$$\boldsymbol{LD_i}(\boldsymbol{\beta}, \sigma^2) = 2\left\{log(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2) - log(\tilde{\boldsymbol{\beta}}_{(\boldsymbol{i})}, \tilde{\sigma}^2_{(i)})\right\}, \quad i = 1, 2, ..., n \tag{3.42}$$

Sustituyendo los MLE $\tilde{\boldsymbol{\beta}}_{(\boldsymbol{i})} = \widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})}$ y $\tilde{\sigma}^2_{(i)} = \widehat{\sigma}^2\left(\dfrac{n-k-\boldsymbol{r_i}^2}{n-1}\right)$, así como las ecuaciones (3.40) y (3.41) en la ecuación (3.42), se tiene

$$\boldsymbol{LD_i}(\boldsymbol{\beta}, \sigma^2) = nlog\left(\frac{n(n-k-\boldsymbol{r_i}^2)}{(n-1)(n-k)}\right) + \frac{(n-1)(\boldsymbol{y_i} - \boldsymbol{x_i^T}\widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})})^2}{\widehat{\sigma}^2(n-k-\boldsymbol{r_i}^2)} - 1$$

Usando el residual predicho, y simplificando, $\boldsymbol{LD_i}(\boldsymbol{\beta}, \sigma^2)$ se convierte en

$$\boldsymbol{LD_i}(\boldsymbol{\beta}, \sigma^2) = nlog\left(\frac{n(n-k-\boldsymbol{r_i}^2)}{(n-1)(n-k)}\right) + \frac{(n-1)\boldsymbol{r_i}^2}{(1-\boldsymbol{p_{ii}})(n-k-\boldsymbol{r_i}^2)} - 1$$

La similaridad existente entre las ecuaciones (3.38) y (3.42), sugiere que $\boldsymbol{LD_i}(\boldsymbol{\beta}, \sigma^2)$ se pueda comparar con la distribución ji-cuadrada con $(k+1)$ grados de libertad.

Notando que $\boldsymbol{LD_i}(\boldsymbol{\beta}, \sigma^2)$ se utiliza para estimar $\boldsymbol{\beta}$ y σ^2, pero si solamente se desea estimar $\boldsymbol{\beta}$, entonces el $100(1-\alpha)\%$ de la región de confianza asintótica para $\boldsymbol{\beta}$ es

$$\left\{\boldsymbol{\beta} : 2\left(log(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2) - max_\sigma \quad log(\boldsymbol{\beta}, \sigma^2)\right) \leq \chi^2(\alpha; k)\right\},$$

y la distancia de verosimilitud se convierte en

$$\boldsymbol{LD_i}(\boldsymbol{\beta}|\sigma^2) = 2\left\{log(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2) - max_\sigma \quad log(\tilde{\boldsymbol{\beta}}_{(\boldsymbol{i})}, \sigma^2)\right\}, \quad i = 1, 2, ..., n, \tag{3.43}$$

donde

$$log(\tilde{\boldsymbol{\beta}}_{(\boldsymbol{i})}, \sigma^2) = -\frac{n}{2}log2\pi - \frac{n}{2}log\sigma^2 - \frac{1}{2\sigma^2}\sum_{r=1}^{n}(\boldsymbol{y_r} - \boldsymbol{x_r^T}\tilde{\boldsymbol{\beta}}_{(\boldsymbol{i})})^2$$

El valor de σ^2 que maximiza $(\tilde{\boldsymbol{\beta}}_{(\boldsymbol{i})}, \sigma^2)$, se encuentra así,

$$\tilde{\sigma}^2(\tilde{\boldsymbol{\beta}}_{(\boldsymbol{i})}) = \frac{1}{n}\sum_{r=1}^{n}(\boldsymbol{y_r} - \boldsymbol{x_r^T}\tilde{\boldsymbol{\beta}}_{(i)})^2 = \tilde{\sigma}^2_{(i)}\left(\frac{n-1}{n}\right) + \frac{\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}^2}{n(1-\boldsymbol{p_{ii}})^2},$$

de manera que

$$max_{\sigma} \quad log(\tilde{\boldsymbol{\beta}}_{(i)}, \sigma^2) = -\frac{n}{2} log 2\pi - \frac{n}{2} log \tilde{\sigma}^2(\tilde{\boldsymbol{\beta}}_{(\boldsymbol{i})}) - \frac{n}{2} \tag{3.44}$$

Por lo que al sustituir las ecuaciones (3.40) y (3.44) en la ecuación (3.43), se produce

$$\begin{aligned}
\boldsymbol{LD_i}(\boldsymbol{\beta}|\sigma^2) &= nlog\left(\frac{\tilde{\sigma}^2(\tilde{\boldsymbol{\beta}}_{(\boldsymbol{i})})}{\tilde{\sigma}^2}\right) \\
&= nlog\left(\frac{(n-1)}{n}\frac{\tilde{\sigma}^2_{(i)}}{\tilde{\sigma}^2} + \frac{\widehat{\boldsymbol{\epsilon}}^2_{\boldsymbol{i}}}{n\tilde{\sigma}^2(1-\boldsymbol{p_{ii}})^2}\right) \\
&= nlog\left(\frac{(n-k-1)}{(n-k)}\frac{\widehat{\sigma}^2_{(i)}}{\widehat{\sigma}^2} + \frac{\widehat{\boldsymbol{\epsilon}}^2_{\boldsymbol{i}}}{(n-k)\widehat{\sigma}^2(1-\boldsymbol{p_{ii}})^2}\right) \\
&= nlog\left(\frac{(n-k-\boldsymbol{r_i}^2)}{(n-k)} + \frac{\boldsymbol{r_i}^2}{(n-k)(1-\boldsymbol{p_{ii}})}\right) \\
&= nlog\left(1 + \frac{\boldsymbol{r_i}^2}{(n-k)}\frac{\boldsymbol{p_{ii}}}{(1-\boldsymbol{p_{ii}})}\right) \\
&= nlog\left(1 + \frac{k}{(n-k)}\boldsymbol{C_i}\right), \quad i = 1, 2, ..., n,
\end{aligned} \tag{3.45}$$

donde $\boldsymbol{C_i}$ está dada en la ecuación (3.22), entonces $\boldsymbol{LD_i}(\boldsymbol{\beta}|\sigma^2)$ es equivalente a la distancia de Cook, pero $\boldsymbol{LD_i}(\boldsymbol{\beta}|\sigma^2)$ puede compararse con la distribución $\chi^2(\alpha; k)$.

Medidas Basadas en un Subconjunto de Coeficientes de Regresión

A continuación se presentan medidas que consideran la influencia que una observación tiene sobre un simple coeficiente de regresión y sobre combinaciones lineales de los coeficientes de regresión.

Influencia sobre un Simple Coeficiente de Regresión

Se supone que la j-ésima variable $\boldsymbol{X_j}$ es la última columna de $\boldsymbol{X}$ y particionando $\boldsymbol{X}$ como $\boldsymbol{X} = (\boldsymbol{X_{(j)}} : \boldsymbol{X_j})$, donde $\boldsymbol{X_{(j)}}$ es la matriz $\boldsymbol{X}$ sin $\boldsymbol{X_j}$. El modelo dado en la ecuación (1.1), puede entonces ser escrito como

$$\boldsymbol{Y} = \boldsymbol{X_{(j)}}\boldsymbol{\beta_{(j)}} + \boldsymbol{X_j}\boldsymbol{\beta_j} + \boldsymbol{\epsilon}$$

Utilizando el Teorema 4, se puede descomponer la matriz de predicción $\boldsymbol{P}$ como

$$\begin{aligned} \boldsymbol{P} &= \boldsymbol{P}_{(j)} + \frac{(\boldsymbol{I} - \boldsymbol{P}_{(j)})\boldsymbol{X}_j \boldsymbol{X}_j^{\boldsymbol{T}} (\boldsymbol{I} - \boldsymbol{P}_{(j)})}{\boldsymbol{X}_j^{\boldsymbol{T}} (\boldsymbol{I} - \boldsymbol{P}_{(j)}) \boldsymbol{X}_j} \\ &= \boldsymbol{P}_{(j)} + \frac{\boldsymbol{W}_j \boldsymbol{W}_j^{\boldsymbol{T}}}{\boldsymbol{W}_j^{\boldsymbol{T}} \boldsymbol{W}_j}, \end{aligned} \tag{3.46}$$

donde

$$\boldsymbol{P}_{(j)} = \boldsymbol{X}_{(j)} (\boldsymbol{X}_{(j)}^{\boldsymbol{T}} \boldsymbol{X}_{(j)})^{-1} \boldsymbol{X}_{(j)}^{\boldsymbol{T}}, \tag{3.47}$$

es la matriz de predicción para $\boldsymbol{X}_{(j)}$, y

$$\boldsymbol{W}_j = (\boldsymbol{I} - \boldsymbol{P}_{(j)}) \boldsymbol{X}_j, \tag{3.48}$$

es el vector de residuales, cuando se hace una regresión de $\boldsymbol{X}_j$ sobre $\boldsymbol{X}_{(j)}$, y $(\boldsymbol{W}_j^{\boldsymbol{T}} \boldsymbol{W}_j)$ es la suma de cuadrados de residuales.

Definiendo a $\widehat{\boldsymbol{\beta}}_j$ y $\widehat{\boldsymbol{\beta}}_{j(i)}$, como los estimadores de $\boldsymbol{\beta}_j$ obtenidos de los datos completos y los datos sin la i-ésima observación, respectivamente. Observando entonces que

$$(\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\beta}}_{j(i)}) = \frac{\widehat{\boldsymbol{\epsilon}}_i}{(1 - \boldsymbol{p}_{ii})} \frac{\boldsymbol{w}_{ij}}{\boldsymbol{W}_j^{\boldsymbol{T}} \boldsymbol{W}_j}, \tag{3.49}$$

donde $\boldsymbol{w}_{ij}$ es la i-ésima componente de $\boldsymbol{W}_j$.

La influencia de la i-ésima observación sobre el j-ésimo coeficiente de regresión estimado, se obtiene al dividir la ecuación (3.49) por el error estándar de $(\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\beta}}_{j(i)})$, entonces se tiene

$$\boldsymbol{Var}(\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\beta}}_{j(i)}) = \frac{\sigma^2}{(1 - \boldsymbol{p}_{ii})} \left(\frac{\boldsymbol{w}_{ij}}{\boldsymbol{W}_j^{\boldsymbol{T}} \boldsymbol{W}_j} \right)^2$$

y así,

$$\frac{(\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\beta}}_{j(i)})}{\sqrt{\boldsymbol{Var}(\widehat{\boldsymbol{\beta}}_j - \widehat{\boldsymbol{\beta}}_{j(i)})}} = \frac{\widehat{\boldsymbol{\epsilon}}_i}{\sigma \sqrt{1 - \boldsymbol{p}_{ii}}},$$

el cual, dependiendo de la elección del estimador de σ, se puede obtener r_i o r_i^*. Alternativamente, la ecuación (3.49), puede ser dividida por el error estándar de $\widehat{\beta}_j$, y así obtener

$$\frac{(\widehat{\beta}_j - \widehat{\beta}_{j(i)})}{\sqrt{Var(\widehat{\beta}_j)}} \tag{3.50}$$

Se ve que $\widehat{\beta}_j$ puede expresarse como

$$\widehat{\beta}_j = \frac{X_j^T(I - P_{(j)})Y}{X_j^T(I - P_{(j)})X_j} = \frac{W_j^T Y}{W_j^T W_j} \tag{3.51}$$

Y de la ecuación (3.51), se sigue que

$$Var(\widehat{\beta}_j) = \frac{\sigma^2}{W_j^T W_j} \tag{3.52}$$

Sustituyendo ahora, las ecuaciones (3.49) y (3.52) en la ecuación (3.50), se produce

$$\frac{(\widehat{\beta}_j - \widehat{\beta}_{j(i)})}{\sqrt{Var(\widehat{\beta}_j)}} = \frac{\widehat{\epsilon}_i}{\sigma\sqrt{1 - p_{ii}}} \frac{w_{ij}}{\sqrt{W_j^T W_j}} \frac{1}{\sqrt{1 - p_{ii}}} \tag{3.53}$$

Usando $\widehat{\sigma}$ como estimador de σ en la ecuación (3.53), se obtiene

$$D_{ij} = r_i \frac{w_{ij}}{\sqrt{W_j^T W_j}} \frac{1}{\sqrt{1 - p_{ii}}} \tag{3.54}$$

Mientras que, usando $\widehat{\sigma}^2_{(i)}$ como estimador de σ en la ecuación (3.53), se llega a

$$D_{ij}{}^* = r_i{}^* \frac{w_{ij}}{\sqrt{W_j^T W_j}} \frac{1}{\sqrt{1 - p_{ii}}} \tag{3.55}$$

Belsley *et al.*, (1980), sugieren poner especial atención, a aquellos puntos con valores de $|D_{ij}{}^*|$, que excedan $\frac{2}{n}$.

Influencia sobre Funciones Lineales de $\widehat{\beta}$

Suponiendo ahora, que la i-ésima observación influye un subconjunto de coeficientes de regresión dado, es decir, q combinaciones linealmente independientes de

$\boldsymbol{\beta}$, o sea, $\boldsymbol{L}^T\boldsymbol{\beta}$, donde $\boldsymbol{L}$ es una matriz $k \times q$ con rango q. Sea $\Theta = \boldsymbol{L}^T\boldsymbol{\beta}$, el máximo MLE de Θ es $\widehat{\Theta} = \boldsymbol{L}^T\widehat{\boldsymbol{\beta}}$. La curva de influencia para $\widehat{\Theta}$ es $\boldsymbol{L}^T\psi\left(\boldsymbol{x}^T, \boldsymbol{y}, \boldsymbol{F}, \widehat{\boldsymbol{\beta}}(\boldsymbol{F})\right)$. Si se utiliza la curva de influencia muestral, $\boldsymbol{SIC_i}$ dada en la ecuación (3.13), para aproximar $\psi(\cdot)$, y se usa además, la ecuación (3.21) como una medida de la influencia de la i-ésima observación sobre $\widehat{\Theta}$, se obtiene

$$\boldsymbol{C_i(L)} = \boldsymbol{D_i}(M, c) = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})})^T M(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})})}{c} \tag{3.56}$$

donde $c = q(n-1)^2\widehat{\sigma}^2$ y $M = \boldsymbol{L}\left(\boldsymbol{L}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}\right)^{-1}\boldsymbol{L}^T$. Usando la ecuación (3.16), entonces la ecuación (3.56) se reduce a

$$\boldsymbol{C_i(L)} = \frac{\boldsymbol{r_i}^2}{q(1-\boldsymbol{p_{ii}})}\boldsymbol{x_i^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}M(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x_i}, \quad i = 1, 2, ..., n \tag{3.57}$$

Se supone que el interés está centrado sólo en q elementos de $\boldsymbol{\beta}$, las cuales, son las últimas q componentes de $\boldsymbol{\beta}$, y así $\boldsymbol{L}^T = (\boldsymbol{0} : \boldsymbol{I})$, donde $\boldsymbol{0}$ es una matriz nula de tamaño $q \times (k-q)$. La partición de $\boldsymbol{X}$ es $\boldsymbol{X} = (\boldsymbol{X_1} : \boldsymbol{X_2})$ y la de $\boldsymbol{x_i}$ es $\boldsymbol{x_i^T} = (\boldsymbol{x_{1_i}^T} : \boldsymbol{x_{2_i}^T})$. Usando el Lema 1(1), se tiene que $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ se puede escribir como

$$\begin{bmatrix} (\boldsymbol{X_1^T X_1})^{-1} + (\boldsymbol{X_1^T X_1})^{-1}\boldsymbol{X_1^T X_2} A \boldsymbol{X_2^T X_1}(\boldsymbol{X_1^T X_1})^{-1} & -(\boldsymbol{X_1^T X_1})^{-1}\boldsymbol{X_1^T X_2} A \\ -A\boldsymbol{X_2^T X_1}(\boldsymbol{X_1^T X_1})^{-1} & A \end{bmatrix}$$

donde $\boldsymbol{A^{-1}} = \boldsymbol{X_2^T}\left(\boldsymbol{I} - \boldsymbol{X_1}(\boldsymbol{X_1^T X_1})^{-1}\boldsymbol{X_1^T}\right)\boldsymbol{X_2}$, y así,

$$\begin{aligned}
&(\boldsymbol{X}^T\boldsymbol{X})^{-1}M(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\
&= \begin{bmatrix} (\boldsymbol{X_1^T X_1})^{-1}\boldsymbol{X_1^T X_2} A \boldsymbol{X_2^T X_1}(\boldsymbol{X_1^T X_1})^{-1} & -(\boldsymbol{X_1^T X_1})^{-1}\boldsymbol{X_1^T X_2} A \\ -A\boldsymbol{X_2^T X_1}(\boldsymbol{X_1^T X_1})^{-1} & A \end{bmatrix} \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1} - \begin{bmatrix} (\boldsymbol{X_1^T X_1})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix},
\end{aligned}$$

de lo cual se sigue que

$$\boldsymbol{x_i^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}M(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x_i} = \boldsymbol{p_{ii}} - \boldsymbol{x_{1_i}^T}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x_{1_i}},$$

por lo tanto, la ecuación (3.57) puede escribirse como

$$C_i(L) = \frac{r_i^2}{q(1-p_{ii})}\left(p_{ii} - x_{1_i}^T(X^TX)^{-1}x_{1_i}\right), \quad i = 1, 2, ..., n \tag{3.58}$$

Así, la ecuación (3.58), mide la influencia de la i-ésima observación sobre q combinaciones linealmente independientes de los coeficientes de regresión especificados para L. Si sólo interesa un coeficiente β_j, entonces la ecuación (3.58) puede simplificarse, esto es, si $q = 1$, $L^T = (0^T : 1)$, $X_1 = X_{(j)}$ y $x_{1_j} = x_{i(j)}$, se produce

$$C_i(L) = \frac{r_i^2}{(1-p_{ii})}(p_{ii} - p_{ii(j)}) \tag{3.59}$$

De la ecuación (3.46), se tiene que

$$p_{ii} = p_{ii(j)} + \frac{w_{ij}^2}{W_j^TW_j},$$

y por consiguiente, la ecuación (3.59) se convierte en

$$D_{ij}^2 = C_i(L) = r_i^2\frac{w_{ij}^2}{W_j^TW_j}\frac{1}{(1-p_{ii})}, \quad i = 1, 2, ..., n, \quad j = 1, 2, ..., k, \tag{3.60}$$

la cual mide la influencia de la i-ésima observación sobre $\widehat{\beta}_j$. Se ve que el D_{ij} en (3.60), es el que está dado en la ecuación (3.54).

Método Basado en la Diferenciación

Hasta el momento, se ha visto el efecto que causa una observación individual sobre la ecuación de regresión ajustada, mediante el método de la eliminación de esa observación. Una alternativa, es el método basado en la diferenciación, el cual, en vez de realizar una modificación en los datos, se induce un pequeño cambio en algunos parámetros del modelo y luego se estudian los resultados del análisis de regresión como una función de estos parámetros modificados, Welsch y Kuh (1977), Pregibon (1979 y 1981) y Belsley *et al.* (1980).

$$\nabla\widehat{\boldsymbol{\beta}}(0) = \frac{\partial\widehat{\boldsymbol{\beta}}(\boldsymbol{w_i})}{\partial \boldsymbol{w_i}}|_{w_i=0} = (\boldsymbol{X^T X})^{-1}\boldsymbol{x_i}\frac{\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}}{(1-\boldsymbol{p_{ii}})^2} \tag{3.67}$$

Existe equivalencia entre las medidas de influencia basadas en $\nabla\widehat{\boldsymbol{\beta}}(\boldsymbol{w_i})$ y las basadas en la curva de influencia, esto es, $\boldsymbol{EIC_i}$ es proporcional a $\nabla\widehat{\boldsymbol{\beta}}(1)$, y $\boldsymbol{EIC_{(i)}}$ es proporcional a $\nabla\widehat{\boldsymbol{\beta}}(0)$. Por lo que $\boldsymbol{EIC_i}$ mide la tasa de cambio en $\widehat{\boldsymbol{\beta}}$ con la i-ésima observación ($\boldsymbol{w_i} = 1$), y $\boldsymbol{EIC_{(i)}}$ mide la tasa de cambio en $\widehat{\boldsymbol{\beta}}$ cuando la i-ésima observación es eliminada ($\boldsymbol{w_i} = 0$).

Puesto que $\widehat{\boldsymbol{\beta}}(\boldsymbol{w})$ es diferenciable sobre todo el rango de $\boldsymbol{w}$, el Teorema del Valor Medio, garantiza la existencia de $\boldsymbol{w_i^*}$, tal que $\nabla\widehat{\boldsymbol{\beta}}(1) \leq \nabla\widehat{\boldsymbol{\beta}}(\boldsymbol{w_i^*}) \leq \nabla\widehat{\boldsymbol{\beta}}(0)$. Entonces

$$\begin{aligned}
\nabla\widehat{\boldsymbol{\beta}}(\boldsymbol{w_i^*}) &= \int_0^1 \nabla\widehat{\boldsymbol{\beta}}(\boldsymbol{w_i})d\boldsymbol{w_i} \\
&= (\boldsymbol{X^T X})^{-1}\boldsymbol{x_i}\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}\int_0^1 \frac{1}{(1-\boldsymbol{p_{ii}}(1-\boldsymbol{w_i}))^2}d\boldsymbol{w_i} \\
&= (\boldsymbol{X^T X})^{-1}\boldsymbol{x_i}\frac{\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}}{(1-\boldsymbol{p_{ii}})}
\end{aligned} \tag{3.68}$$

En las ecuaciones (3.17), (3.18) y (3.68), se ve que

$$\nabla\widehat{\boldsymbol{\beta}}(\boldsymbol{w_i^*}) = \frac{\boldsymbol{SIC_i}}{(n-1)} = \frac{\boldsymbol{SC_i}}{n},$$

por lo que, $\boldsymbol{SIC_i}$ y $\boldsymbol{SC_i}$ son proporcionales a $\nabla\widehat{\boldsymbol{\beta}}(\boldsymbol{w})$, $0 \leq \boldsymbol{w} \leq 1$.

EFECTOS DE MULTIPLES OBSERVACIONES SOBRE LA ECUACIÓN DE REGRESIÓN

Hasta el momento se han descrito algunos métodos para detectar observaciones que individualmente pueden ser consideradas como "outliers", puntos palanca o puntos influyentes; en este capítulo se extienden los procedimientos de una simple observación al caso general, en el que se consideran los efectos conjuntos de múltiples observaciones, sobre varios resultados del análisis de regresión.

Existen tres problemas inherentes al caso general de observaciones múltiples, que son:

1. La determinación del tamaño del subconjunto de observaciones conjuntamente influyentes. Suponiendo que se desean detectar todos los subconjuntos de tamaño $m = 2, 3, 4, ...,$ de observaciones que se consideran potencialmente influyentes, en donde un método secuencial puede usarse para determinar m, iniciando con $m = 2$, $m = 3$, $m = 4$, etc.

2. Se ha visto, en el caso de una simple observación, que para cada medida de diagnóstico, se calculan n cantidades, una para cada observación en el conjunto de datos. En el caso de observaciones múltiples, sin embargo, se tiene la desventaja de que para cada subconjunto de tamaño m, existen $\binom{n}{m}$ posibles subconjuntos para los cuales cada medida de diagnóstico puede ser calculada.

3. Las medidas de influencia, para el caso de una simple observación, pueden ser facilmente graficadas. Pero en el caso de observaciones múltiples, no es posible la representación gráfica, especialmente para n y m grandes.

Método de la Eliminación de Múltiples Observaciones

Como en el caso de la detección de una observación influyente individual, en el caso general, se tienen las medidas correspondientes que permiten detectar múltiples observaciones, y que son:

1. Medidas basadas en residuales.
2. Medidas basadas en la curva de influencia.
3. Medidas basadas en el volumen de elipsoides de confianza.
4. Medidas basadas en la función de verosimilitud.
5. Medidas basadas en un subconjunto de coeficientes de regresión.

Medidas Basadas en Residuales

Recordando que si la i-ésima observación es omitida, equivale a ajustar el modelo

$$\boldsymbol{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u_i}\theta, \tag{4.1}$$

donde θ es el coeficiente de regresión del i-ésimo vector unitario $\boldsymbol{u_i}$.

Esta aproximación en que se adiciona una variable indicadora $\boldsymbol{u_i}$ como una forma para omitir la i-ésima observación, puede ser generalizada al caso de omitir m observaciones y adicionar m variables indicadoras.

Sea $I = \{i_1, i_2, ..., i_m\}$, $m < (n-k)$, el conjunto que contiene los índices de las m observaciones que serán eliminadas, y sea $\boldsymbol{U}_I = \{u_{i_1}, u_{i_2}, ..., u_{i_m}\}$, la matriz que contiene las correspondientes variables indicadoras. Se supone ahora que, las m

observaciones que serán omitidas, son las últimas m observaciones, tales que $\boldsymbol{Y}$, $\boldsymbol{X}$ y $\boldsymbol{U}_I$, pueden escribirse como

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_{(I)} \\ \boldsymbol{Y}_I \end{bmatrix} \begin{matrix} (n-m)\times 1 \\ m \times 1 \end{matrix} \qquad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_{(I)} \\ \boldsymbol{X}_I^{\boldsymbol{T}} \end{bmatrix} \begin{matrix} (n-m)\times k \\ m \times k \end{matrix}$$

$$\boldsymbol{U}_I = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I} \end{bmatrix} \begin{matrix} (n-m)\times m \\ m \times m \end{matrix}$$

Así, omitiendo las m observaciones indexadas por I, equivale a ajustar el modelo

$$\boldsymbol{E(Y)} = \boldsymbol{X\beta} + \boldsymbol{U}_I\theta \tag{4.2}$$

donde θ es un vector $m\times 1$ de los coeficientes de regresión de las variables indicadoras $\boldsymbol{U}_I$.

El estadístico para probar,

$$\boldsymbol{H_0} : \boldsymbol{E(Y)} = \boldsymbol{X\beta} \quad \text{vs} \quad \boldsymbol{H_1} : \boldsymbol{E(Y)} = \boldsymbol{X\beta} + \boldsymbol{U}_I\theta,$$

es

$$\boldsymbol{F}_I = \frac{\dfrac{\boldsymbol{SSE(H_0)} - \boldsymbol{SSE(H_1)}}{m}}{\dfrac{\boldsymbol{SSE(H_1)}}{(n-k-m)}} \tag{4.3}$$

Sean $\boldsymbol{P}$ y $\boldsymbol{P}_{X,U_I}$, las matrices de predicción para $\boldsymbol{X}$ y $(\boldsymbol{X} : \boldsymbol{U}_I)$, respectivamente. Utilizando el Teorema 4, se tiene

$$\boldsymbol{Y^T(I - P_{X,U_I})Y}$$

$$= \boldsymbol{Y^T(I-P)Y} - \boldsymbol{Y^T(I-P)U}_I \left(\boldsymbol{U}_I^{\boldsymbol{T}}\boldsymbol{(I-P)U}_I\right)^{-1}\boldsymbol{U}_I^{\boldsymbol{T}}\boldsymbol{(I-P)Y} \tag{4.4}$$

Notando que

$$\left(\boldsymbol{U}_I^{\boldsymbol{T}}\boldsymbol{(I-P)U}_I\right)^{-1} = \boldsymbol{(I-P}_I)^{-1}, \tag{4.5}$$

donde

$$\boldsymbol{P}_I = \boldsymbol{X}_I^{\boldsymbol{T}}\boldsymbol{(X^TX)^{-1}X}_I, \tag{4.6}$$

es la submatriz de $\boldsymbol{P}$, cuyas filas y columnas están indexadas por I. También notando que

$$\boldsymbol{U}_I^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y} = \boldsymbol{U}_I^{\boldsymbol{T}}\widehat{\epsilon} = \widehat{\epsilon}_I, \tag{4.7}$$

donde

$$\widehat{\epsilon}_I = \boldsymbol{Y}_I - \boldsymbol{X}_I^{\boldsymbol{T}}\widehat{\beta}, \tag{4.8}$$

es el subconjunto de los residuales cuyos índices están contenidos en I. Sustituyendo las ecuaciones (4.5) y (4.7) en la ecuación (4.4), y reordenando, se obtiene

$$\boldsymbol{Y}^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y} - \boldsymbol{Y}^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P}_{X,U_I})\boldsymbol{Y} = \widehat{\epsilon}_I^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\widehat{\epsilon}_I, \tag{4.9}$$

entonces, se deduce que

$$\boldsymbol{SSE(H_0)} - \boldsymbol{SSE(H_1)} = \widehat{\epsilon}_I^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\widehat{\epsilon}_I \tag{4.10}$$

Sustituyendo la ecuación (4.10) en la ecuación (4.3), se obtiene

$$\boldsymbol{F}_I = \frac{(n-k-m)}{m}\frac{\widehat{\epsilon}_I^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\widehat{\epsilon}_I}{\boldsymbol{Y}^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P}_{X,U_I})\boldsymbol{Y}} = \frac{\widehat{\epsilon}_I^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\widehat{\epsilon}_I}{m\widehat{\sigma}_I^2}, \tag{4.11}$$

donde

$$\widehat{\sigma}_{(I)}^2 = \frac{\boldsymbol{SSE}_{Y_{(I)}X_{(I)}}}{(n-k-m)} = \frac{\boldsymbol{Y}^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P}_{X,U_I})\boldsymbol{Y}}{(n-k-m)}, \tag{4.12}$$

es el estimador de σ^2 cuando las m observaciones indexadas por I son omitidas. Si se dividen ambos lados de la ecuación (4.10) por $(n-k-m)$, y reordenando, se obtiene

$$\widehat{\sigma}_{(I)}^2 = \frac{(n-k)}{(n-k-m)}\widehat{\sigma}^2 - \frac{\widehat{\epsilon}_I^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\widehat{\epsilon}_I}{(n-k-m)} = \widehat{\sigma}^2\left(\frac{n-k-\boldsymbol{r}_I^2}{n-k-m}\right), \tag{4.13}$$

donde

$$\boldsymbol{r}_I^2 = \frac{\widehat{\epsilon}_I^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\widehat{\epsilon}_I}{\widehat{\sigma}^2}, \tag{4.14}$$

es el análogo de los $\boldsymbol{r_i}^2$.

Suponiendo que existen máximo m "outliers" en el conjunto de datos y que se conocen las etiquetas de esas observaciones, entonces bajo los supuestos de normalidad, el estadístico $\boldsymbol{F}_I$ dado en la ecuación (4.11) se distribuye como $\boldsymbol{F}(m, n-k-m)$. En la práctica, se calcula $\boldsymbol{F}_I$, para todo $\binom{n}{m}$ posible subconjunto de tamaño m y se busca el $\boldsymbol{F}_I$ más grande. La distribución de $max(\boldsymbol{F}_I)$ es difícil de obtener, como en el caso de $\boldsymbol{r_{max}}$.

Medidas Basadas en la Curva de Influencia

Se han visto varias medidas de la influencia de la i-ésima observación sobre $\widehat{\boldsymbol{\beta}}$, las cuales, están basadas en la ecuación (3.21). Se presenta ahora, una generalización de $\boldsymbol{D_i}(M, c)$, la cual es dada por

$$\boldsymbol{D_i}(M,c) = \frac{\psi^{\boldsymbol{T}}\left(\boldsymbol{X}_I^{\boldsymbol{T}}, \boldsymbol{Y}_I, \boldsymbol{F}, \widehat{\boldsymbol{\beta}}(\boldsymbol{F})\right) \boldsymbol{M} \psi\left(\boldsymbol{X}_I^{\boldsymbol{T}}, \boldsymbol{Y}_I, \boldsymbol{F}, \widehat{\boldsymbol{\beta}}(\boldsymbol{F})\right)}{c}, \tag{4.15}$$

donde $\psi\left(\boldsymbol{X}_I^{\boldsymbol{T}}, \boldsymbol{Y}_I, \boldsymbol{F}, \widehat{\boldsymbol{\beta}}(\boldsymbol{F})\right)$ es una generalización de la versión muestral de la curva de influencia para múltiples observaciones.

Al igual que para el caso de una simple observación, se presentan a continuación dos aproximaciones a la curva de influencia para $\widehat{\boldsymbol{\beta}}$.

Curva de Influencia Muestral

Una generalización de la curva de influencia muestral $\boldsymbol{SIC_i}$ puede obtenerse de la ecuación (3.12), si se omite el límite y se reemplaza $(\boldsymbol{x}^{\boldsymbol{T}}, \boldsymbol{y})$ por $(\boldsymbol{X}_I^{\boldsymbol{T}}, \boldsymbol{Y}_I)$ y $\boldsymbol{F}$ por $\widehat{\boldsymbol{F}}$, y se define a $\epsilon = \frac{m}{m-n}$, entonces se obtiene

$$\begin{aligned} \boldsymbol{SIC}_I &= \frac{(n-m)}{-m}\left(\boldsymbol{T}\left(\frac{n}{(n-m)}\widehat{\boldsymbol{F}} - \frac{m}{(n-m)}\delta_{\boldsymbol{X}_I^{\boldsymbol{T}}, \boldsymbol{Y}_I}\right) - \boldsymbol{T}(\widehat{\boldsymbol{F}})\right) \\ &= \frac{(n-m)}{m}\left(\boldsymbol{T}(\widehat{\boldsymbol{F}}) - \boldsymbol{T}(\widehat{\boldsymbol{F}}_{(I)})\right), \end{aligned} \tag{4.16}$$

donde $\widehat{\boldsymbol{F}}_{(I)}$ es la función de distribución acumulada empírica cuando las m observaciones indexadas por I son omitidas. Sustituyendo $\widehat{\boldsymbol{\beta}}$ por $\boldsymbol{T}(\widehat{\boldsymbol{F}})$ y $\widehat{\boldsymbol{\beta}}_{(I)}$ por $\boldsymbol{T}(\widehat{\boldsymbol{F}}_{(I)})$ en la ecuación (4.16), se obtiene

$$\boldsymbol{SIC}_I = \frac{(n-m)}{m}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(I)}), \tag{4.17}$$

donde

$$\widehat{\boldsymbol{\beta}}_{(I)} = (\boldsymbol{X}_{(I)}^{\boldsymbol{T}}\boldsymbol{X}_{(I)})^{-1}\boldsymbol{X}_{(I)}^{\boldsymbol{T}}\boldsymbol{Y}_{(I)}, \tag{4.18}$$

son los coeficientes de regresión estimados, cuando las m observaciones indexadas por I son eliminadas. Para simplificar la ecuación (4.17), se expresa la diferencia $(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(I)})$ en términos de cantidades obtenidas del ajuste del modelo a los datos completos. Reescribiendo $\widehat{\boldsymbol{\beta}}_{(I)}$ como

$$\widehat{\boldsymbol{\beta}}_{(I)} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} - \boldsymbol{X}_I\boldsymbol{X}_I^{\boldsymbol{T}})^{-1}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y} - \boldsymbol{X}_I\boldsymbol{Y}_I) \tag{4.19}$$

Y usando el Lema 2 para evaluar la inversa de $(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} - \boldsymbol{X}_I\boldsymbol{X}_I^{\boldsymbol{T}})$, se obtiene

$$\begin{aligned}(\boldsymbol{X}_{(I)}^{\boldsymbol{T}}\boldsymbol{X}_{(I)})^{-1} &= (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} - \boldsymbol{X}_I\boldsymbol{X}_I^{\boldsymbol{T}})^{-1} \\ &= (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1} + (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I} - \boldsymbol{P}_I)^{-1}\boldsymbol{X}_I^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\end{aligned} \tag{4.20}$$

Sustituyendo la ecuación (4.20) en la ecuación (4.19), se obtiene

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{(I)} &= \left((\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1} + (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I} - \boldsymbol{P}_I)^{-1}\boldsymbol{X}_I^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\right)\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y} - \boldsymbol{X}_I\boldsymbol{Y}_I\right) \\ &= \widehat{\boldsymbol{\beta}} + (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I} - \boldsymbol{P}_I)^{-1}\boldsymbol{X}_I^{\boldsymbol{T}}\widehat{\boldsymbol{\beta}} - (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_I\boldsymbol{Y}_I \\ &\quad -(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I} - \boldsymbol{P}_I)^{-1}\boldsymbol{P}_I\boldsymbol{Y}_I\end{aligned}$$

Adicionando y sustrayendo $(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I} - \boldsymbol{P}_I)^{-1}\boldsymbol{Y}_I$, y reordenando, se obtiene

$$\begin{aligned}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(I)}) &= (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I} - \boldsymbol{P}_I)^{-1}\widehat{\boldsymbol{\epsilon}}_I \\ &\quad +(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_I\left(\boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{P}_I)^{-1} + (\boldsymbol{I} - \boldsymbol{P}_I)^{-1}\boldsymbol{P}_I\right)\boldsymbol{Y}_I\end{aligned}$$

y como $(\boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{P}_I)^{-1} + (\boldsymbol{I} - \boldsymbol{P}_I)^{-1}\boldsymbol{P}_I) = \boldsymbol{0}$, entonces resulta que

$$(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(I)}) = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I} - \boldsymbol{P}_I)^{-1}\widehat{\boldsymbol{\epsilon}}_I, \tag{4.21}$$

la cual constituye una generalización de la ecuación (3.16). Se puede encontrar una expresión más simple para $\boldsymbol{SIC}_I$, al sustituir la ecuación (4.21) en la ecuación (4.17), lo cual produce

$$\boldsymbol{SIC}_I = \frac{(n-m)}{m}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\widehat{\boldsymbol{\epsilon}}_I, \tag{4.22}$$

la cual se reduce a la ecuación (3.17), cuando m=1.

Curva de Influencia Empírica

La curva de influencia empírica basada en $(n-m)$ observaciones, se encuentra a partir de la curva de influencia para $\widehat{\boldsymbol{\beta}}$, definida en el Teorema 13(1). Aproximando $\boldsymbol{F}$ por $\widehat{\boldsymbol{F}}_{(I)}$, y considerando que $(\boldsymbol{x}^T, \boldsymbol{y}) = (\boldsymbol{X}_I^T, \boldsymbol{Y}_I)$, $\Sigma_{xx}(\boldsymbol{F}) = (n-m)^{-1}(\boldsymbol{X}_{(I)}^T\boldsymbol{X}_{(I)})$ y $\widehat{\boldsymbol{\beta}}(\boldsymbol{F}) = \widehat{\boldsymbol{\beta}}_{(I)}$, se obtiene

$$\boldsymbol{EIC}_{(I)} = (n-m)(\boldsymbol{X}_{(I)}^T\boldsymbol{X}_{(I)})^{-1}\boldsymbol{X}_I(\boldsymbol{Y}_I - \boldsymbol{X}_I^T\widehat{\boldsymbol{\beta}}_{(I)}) \tag{4.23}$$

De la ecuación (4.20), se tiene que

$$\begin{aligned}(\boldsymbol{X}_{(I)}^T\boldsymbol{X}_{(I)})^{-1}\boldsymbol{X}_I &= \left((\boldsymbol{X}^T\boldsymbol{X})^{-1} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\boldsymbol{X}_I^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\right)\boldsymbol{X}_I \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_I\left(\boldsymbol{I} + (\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\boldsymbol{P}_I\right) \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-1} \end{aligned} \tag{4.24}$$

Mientras que de la ecuación (4.21), se tiene

$$\begin{aligned}(\boldsymbol{Y}_I - \boldsymbol{X}_I^T\widehat{\boldsymbol{\beta}}_{(I)}) &= \boldsymbol{Y}_I - \boldsymbol{X}_I^T\left(\widehat{\boldsymbol{\beta}} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\widehat{\boldsymbol{\epsilon}}_I\right) \\ &= \widehat{\boldsymbol{\epsilon}}_I + \boldsymbol{P}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\widehat{\boldsymbol{\epsilon}}_I \\ &= (\boldsymbol{I} + \boldsymbol{P}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-1})\widehat{\boldsymbol{\epsilon}}_I \\ &= (\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\widehat{\boldsymbol{\epsilon}}_I \end{aligned} \tag{4.25}$$

Sustituyendo las ecuaciones (4.24) y (4.25) en la ecuación (4.23), y simplificando, se obtiene

$$\boldsymbol{EIC}_{(I)} = (n-m)(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-2}\widehat{\boldsymbol{\epsilon}}_I, \tag{4.26}$$

la cual se reduce a la ecuación (3.20) cuando m=1.

Las medidas de diagnóstico basadas en la curva de influencia pueden también ser generalizadas al caso de observaciones múltiples. A continuación se muestran dos de ellas, las cuales se obtienen utilizando la ecuación (4.15), una aproximación a la curva de influencia para $\widehat{\boldsymbol{\beta}}$ (tal como $\boldsymbol{SIC}_I$ y $\boldsymbol{EIC}_{(I)}$) y una elección apropiada para M y c.

Distancia de Cook Generalizada

El análogo a la distancia de Cook $\boldsymbol{C_i}$, se obtiene sustituyendo $\boldsymbol{SIC}_I$ en la ecuación (4.15), después de una simplificación, se obtiene

$$\boldsymbol{C}_I = D_I\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}, k\widehat{\sigma}^2\left(\frac{m}{n-m}\right)^2\right) = \frac{\widehat{\boldsymbol{\epsilon}}_I^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\boldsymbol{P}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\widehat{\boldsymbol{\epsilon}}_I}{k\widehat{\sigma}^2} \qquad (4.27)$$

Un valor grande de $\boldsymbol{C}_I$ en la ecuación (4.27), indica que las observaciones indexadas por I, son conjuntamente influyentes sobre $\widehat{\boldsymbol{\beta}}$. Si se considera el caso específico en que $m = 2$ y $I = \{i, j\}$, entonces la ecuación (4.27) se convierte en

$$\begin{aligned} k\widehat{\sigma}^2\boldsymbol{C}_I &= k\widehat{\sigma}^2(\boldsymbol{C_i}+\boldsymbol{C_j})\left(1+\frac{\boldsymbol{p_{ij}}}{\delta_{ij}}\right)^2 + \frac{\boldsymbol{p_{ij}^2}\left(\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}^2(2-\boldsymbol{p_{jj}})+\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{j}}^2(2-\boldsymbol{p_{ii}})\right)}{\delta_{ij}^2} \\ &\quad + \frac{2\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{j}}\boldsymbol{p_{ij}}(1+\boldsymbol{p_{ij}^2}-\boldsymbol{p_{ii}}\boldsymbol{p_{jj}})}{\delta_{ij}^2}, \end{aligned} \qquad (4.28)$$

donde $\delta_{ij} = (1-\boldsymbol{p_{ii}})(1-\boldsymbol{p_{jj}}) - \boldsymbol{p_{ij}^2}$, y por el Teorema 6(2), $\delta_{ij} \geq 0$. Un análisis de la ecuación (4.5), muestra que

$$\text{si} \quad (\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{j}}\boldsymbol{p_{ij}}) > 0, \quad \text{entonces} \quad \boldsymbol{C}_I > (\boldsymbol{C_i}+\boldsymbol{C_j}),$$

se ve además que si δ_{ij} es pequeño, y $(\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{j}}\boldsymbol{p_{ij}})$ es grande, entonces $\boldsymbol{C}_I$ es grande.

Distancia de Welsch Generalizada

Si se utiliza $\boldsymbol{EIC}_{(I)}$ dado en la ecuación (4.26) como una aproximación a la curva de influencia para $\widehat{\boldsymbol{\beta}}$, y se considera además que

$$M = \boldsymbol{X}_{(I)}^{\boldsymbol{T}}\boldsymbol{X}_{(I)} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} - \boldsymbol{X}_I\boldsymbol{X}_I^{\boldsymbol{T}}) \quad \text{y} \quad c = (n-m)\widehat{\sigma}_{(I)}^2,$$

entonces, se obtiene

$$\begin{aligned} \boldsymbol{W_I^2} &= \left(\frac{n-m}{\widehat{\sigma}^2_{(I)}}\right)\widehat{\epsilon}_I^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P}_I)^{-2}\boldsymbol{X}_I^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}M(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-2}\widehat{\epsilon}_I \\ &= \left(\frac{n-m}{\widehat{\sigma}^2_{(I)}}\right)\widehat{\epsilon}_I^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P}_I)^{-2}\boldsymbol{P}_I(\boldsymbol{I}-\boldsymbol{P}_I)(\boldsymbol{I}-\boldsymbol{P}_I)^{-2}\widehat{\epsilon}_I \\ &= \left(\frac{n-m}{\widehat{\sigma}^2_{(I)}}\right)\widehat{\epsilon}_I^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P}_I)^{-2}\boldsymbol{P}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\widehat{\epsilon}_I \end{aligned} \tag{4.29}$$

Medidas Basadas en el Volumen de Elipsoides de Confianza

A continuación se presenta una generalización de las medidas $\boldsymbol{AP_i}$ y $\boldsymbol{VR_i}$. No se generaliza $\boldsymbol{CW_i}$, pues es una función monótona de $\boldsymbol{VR_i}$.

Estadístico Andrews-Pregibon Generalizado

Definiendo la matriz aumentada $\boldsymbol{Z} = (\boldsymbol{X} : \boldsymbol{Y})$, y sean $\boldsymbol{Z_I^T}$ las m filas de $\boldsymbol{Z}$ que serán eliminadas y $\boldsymbol{Z}_{(I)}$ la matriz $\boldsymbol{Z}$ sin $\boldsymbol{Z_I^T}$, entonces, utilizando el Lema 4, se tiene

$$\begin{aligned} \boldsymbol{det}(\boldsymbol{Z}^{\boldsymbol{T}}_{(I)}\boldsymbol{Z}_{(I)}) &= \boldsymbol{det}(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{Z} - \boldsymbol{Z}_I\boldsymbol{Z}_I^{\boldsymbol{T}}) \\ &= \boldsymbol{det}(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{Z})\boldsymbol{det}\left(\boldsymbol{I} - \boldsymbol{Z}_I^{\boldsymbol{T}}(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{Z})^{-1}\boldsymbol{Z}_I\right) \\ &= \boldsymbol{det}(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{Z})\boldsymbol{det}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}_I}), \end{aligned} \tag{4.30}$$

donde $\boldsymbol{P_Z}$ es la matriz de predicción para $\boldsymbol{Z}$, y $\boldsymbol{P}_{\boldsymbol{Z}_I}$ es la submatriz de $\boldsymbol{P_Z}$, cuyas filas y columnas están indexadas por I. Una generalización obvia de la ecuación (3.30), es

$$\boldsymbol{AP}_I = 1 - \frac{\boldsymbol{det}(\boldsymbol{Z}^{\boldsymbol{T}}_{(I)}\boldsymbol{Z}_{(I)})}{\boldsymbol{det}(\boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{Z})}$$

De la ecuación (4.30), se sigue que $\boldsymbol{AP}_I$ puede expresarse como

$$\boldsymbol{AP}_I = 1 - \boldsymbol{det}(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}_I}) \tag{4.31}$$

Del Teorema 6(4), se tiene que

$$\boldsymbol{P_Z} = \boldsymbol{P} + \frac{\widehat{\epsilon}\widehat{\epsilon}^{\boldsymbol{T}}}{\widehat{\epsilon}^{\boldsymbol{T}}\widehat{\epsilon}},$$

y así,

$$\boldsymbol{P_{Z_I}} = \boldsymbol{P_I} + \frac{\widehat{\epsilon}_I\widehat{\epsilon}_I^{\boldsymbol{T}}}{\widehat{\epsilon}^{\boldsymbol{T}}\widehat{\epsilon}}$$

Aplicando el Lema 4, se obtiene

$$\begin{aligned} \boldsymbol{AP_I} &= 1 - \boldsymbol{det}(\boldsymbol{I} - \boldsymbol{P_I})\left(1 - \frac{\widehat{\epsilon}_I^{\boldsymbol{T}}(\boldsymbol{I} - \boldsymbol{P_I})^{-1}\widehat{\epsilon}_I}{\widehat{\epsilon}^{\boldsymbol{T}}\widehat{\epsilon}}\right) \\ &= 1 - \boldsymbol{det}(\boldsymbol{I} - \boldsymbol{P_I})\left(1 - \frac{\boldsymbol{r_I^2}}{(n-k)}\right), \end{aligned} \tag{4.32}$$

donde $\boldsymbol{r_I^2}$ se muestra en la ecuación (4.14).

Relación de Varianzas Generalizada

En forma análoga a la ecuación (3.31), la influencia de las m observaciones indexadas por I sobre la varianza de los coeficientes de regresión estimados es medida por

$$\boldsymbol{VR_I} = \frac{\boldsymbol{det}\left(\widehat{\sigma}^2_{(I)}(\boldsymbol{X}^{\boldsymbol{T}}_{(I)}\boldsymbol{X}_{(I)})^{-1}\right)}{\boldsymbol{det}\left(\widehat{\sigma}^2(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\right)} = \left(\frac{\widehat{\sigma}^2_{(I)}}{\widehat{\sigma}^2}\right)^k \frac{\boldsymbol{det}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})}{\boldsymbol{det}(\boldsymbol{X}^{\boldsymbol{T}}_{(I)}\boldsymbol{X}_{(I)})} \tag{4.33}$$

Usando la demostración del Teorema 10(3) y sustituyendo la ecuación (4.14) en la ecuación (4.33), se obtiene

$$\boldsymbol{VR_I} = \left(\frac{n-k-\boldsymbol{r_I^2}}{n-k-m}\right)^k (\boldsymbol{det}(\boldsymbol{I} - \boldsymbol{P_I}))^{-1} \tag{4.34}$$

donde $\boldsymbol{r_I^2}$ es dada en la ecuación (4.14).

Medidas Basadas en la Función de Verosimilitud

Cuando las m observaciones indexadas por I son omitidas, los MLE de $\boldsymbol{\beta}$ y σ^2 son dados por

$$\tilde{\boldsymbol{\beta}}_{(I)} = \widehat{\boldsymbol{\beta}}_{(I)}, \tag{4.35}$$

y

$$\tilde{\sigma}^2_{(I)} = \hat{\sigma}^2_{(I)}\left(\frac{n-k-m}{n-m}\right) = \hat{\sigma}^2\left(\frac{n-k-\boldsymbol{r}_I^2}{n-m}\right) \tag{4.36}$$

Sustituyendo estos estimadores, por sus respectivos parámetros en la ecuación (3.39), se obtiene

$$log(\tilde{\boldsymbol{\beta}}_{(I)}, \tilde{\sigma}^2_{(I)}) = -\frac{n}{2}log2\pi - \frac{n}{2}log\tilde{\sigma}^2_{(I)} - \frac{(\boldsymbol{Y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}}_{(I)})^{\boldsymbol{T}}(\boldsymbol{Y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}}_{(I)})}{2\tilde{\sigma}^2_{(I)}} \tag{4.37}$$

Usando las ecuaciones (4.21) y (4.35), se tiene

$$\begin{aligned}(\boldsymbol{Y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}}_{(I)}) &= (\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{(I)}) \\ &= \boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\hat{\boldsymbol{\epsilon}}_I \\ &= \hat{\boldsymbol{\epsilon}}+\boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\hat{\boldsymbol{\epsilon}}_I,\end{aligned}$$

y así,

$$\begin{aligned}(\boldsymbol{Y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}}_{(I)})^{\boldsymbol{T}}(\boldsymbol{Y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}}_{(I)}) &= \hat{\boldsymbol{\epsilon}}^{\boldsymbol{T}}\hat{\boldsymbol{\epsilon}}+2\hat{\boldsymbol{\epsilon}}^{\boldsymbol{T}}\boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\hat{\boldsymbol{\epsilon}}_I \\ &\quad +\hat{\boldsymbol{\epsilon}}_I^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\boldsymbol{P}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\hat{\boldsymbol{\epsilon}}_I \qquad (4.38)\end{aligned}$$

Puesto que $\hat{\boldsymbol{\epsilon}}^{\boldsymbol{T}}\boldsymbol{X} = 0$, entonces la ecuación (4.38) se reduce a

$$\begin{aligned}(\boldsymbol{Y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}}_{(I)})^{\boldsymbol{T}}(\boldsymbol{Y}-\boldsymbol{X}\tilde{\boldsymbol{\beta}}_{(I)}) &= \hat{\boldsymbol{\epsilon}}^{\boldsymbol{T}}\hat{\boldsymbol{\epsilon}}+\hat{\boldsymbol{\epsilon}}_I^{\boldsymbol{T}}(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\boldsymbol{P}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\hat{\boldsymbol{\epsilon}}_I \\ &= \hat{\sigma}^2\left((n-k)+k\boldsymbol{C}_I\right), \qquad (4.39)\end{aligned}$$

donde $\boldsymbol{C}_I$ es definido en la ecuación (4.27).

Sustituyendo las ecuaciones (4.36) y (4.39) en la ecuación (4.37), y simplificando se obtiene

$$log(\tilde{\boldsymbol{\beta}}_{(I)}, \tilde{\sigma}^2_{(I)}) = -\frac{n}{2}log2\pi - \frac{n}{2}log\hat{\sigma}^2\left(\frac{n-k-\boldsymbol{r}_I^2}{n-m}\right) - \frac{(n-m)\left((n-k)+k\boldsymbol{C}_I\right)}{2(n-k-\boldsymbol{r}_I^2)} \tag{4.40}$$

Finalmente, la distancia de verosimilitud para $\boldsymbol{\beta}$ y σ^2, cuando m observaciones indexadas por I son eliminadas, es definida por

$$\begin{aligned}\boldsymbol{LD}_I(\boldsymbol{\beta},\sigma^2) &= 2\left(log(\tilde{\boldsymbol{\beta}},\tilde{\sigma}^2)-log(\tilde{\boldsymbol{\beta}}_{(I)},\tilde{\sigma}^2_{(I)})\right) \\ &= nlog\left(\frac{n(n-k-\boldsymbol{r}_I^2)}{(n-m)(n-k)}\right)+\frac{(n-m)(n-k+k\boldsymbol{C}_I)}{(n-k-\boldsymbol{r}_I^2)}-n \qquad (4.41)\end{aligned}$$

Medidas Basadas en un Subconjunto de Coeficientes de Regresión

La influencia de las m observaciones indexadas por I pueden medirse, análogamente a la ecuación (3.53), por

$$\frac{(\widehat{\beta}_j - \widehat{\beta}_{j(I)})}{\sqrt{Var(\widehat{\beta}_j)}}, \tag{4.42}$$

el numerador de la ecuación (4.42), representa el cambio en el j-ésimo coeficiente de regresión debido a la eliminación de las m observaciones indexadas por I, y es dado por

$$(\widehat{\beta}_j - \widehat{\beta}_{j(I)}) = \widehat{\epsilon}_I^T (I - P_I)^{-1} W_{jI} (W_j^T W_j)^{-1}, \tag{4.43}$$

donde W_{jI} es el subconjunto de W_j indexado por I, y W_j es el vector de residuales obtenido de la regresión de X_j sobre los demás componentes de X, para esto, se ve la ecuación (3.48).

De las ecuaciones (3.52) y (4.43), se sigue que

$$\frac{(\widehat{\beta}_j - \widehat{\beta}_{j(I)})}{\sqrt{Var(\widehat{\beta}_j)}} = \sigma \widehat{\epsilon}_I^T (I - P_I)^{-1} W_{jI} (W_j^T W_j)^{-\frac{1}{2}}, \tag{4.44}$$

la cual, es el análogo de la ecuación (3.53). Para $m = 1$, la ecuación (4.44) se reduce a las ecuaciones (3.54) o (3.55), dependiendo de cual cantidad se utilice para estimar σ.

Una modificación de la distancia de Cook

José A. Díaz-García

Departamento de Estadística y Cálculo

Universidad Autónoma Agraria Antonio Narro

25350 Buenavista, Saltillo, Coahuila, MÉXICO.

jadiaz@narro.uaaan.mx

y

Oscar Alejandro Martínez-Jaime

Instituto de Ciencias Agrícolas

Universidad de Guanajuato

Ex-Hda. "El Copal", Irapuato, Guanajuato, MÉXICO.

PALABRAS Y FRASES CLAVES: Medidas de diagnóstico, Distancia de Mahalanobis Generalizada, Puntos influyentes.

ABSTRACT

A modification of the classical Cook's distance is proposed, providing us with a generalized Mahalanobis distance in the context of multivariate normal linear regression models. We establish the exact distribution of a pivotal type statistics based on this generalized Mahalanobis distance, which provides critical points for the identification of data points. We illustrate the procedure with an example, in the context of multiple and multivariate linear regression.

RESUMEN

En el presente artículo se propone una modificación de la distancia de Cook, basándose en la distancia de Mahalanobis generalizada, en el contexto del modelo de regresión lineal multivariado con distribución normal. Se establece además, la

distribución exacta del estadístico basado en esta distancia de Mahalanobis generalizada, la cual proporciona puntos críticos para identificar "outliers" en un conjunto de datos. Este procedimiento, se ilustra con un ejemplo, en el caso de la regresión lineal múltiple y multivariada.

1. INTRODUCCIÓN

El problema de la identificación de "outliers" o puntos influyentes, en el caso de la regresión lineal univariada o multivariada y bajo el supuesto de que los errores se distribuyen normales, ha sido estudiado por varios autores, tales como Cook (1977), Besley et al. (1980), Cook y Weisberg (1982) y Chatterjee y Hadi (1988), sólo por mencionar algunos. Muchos de estos resultados han sido extendidos al caso de las distribuciones de contorno elíptico, ver por ejemplo Galea et al. (1997), Liu (2000) y Díaz-García et al. (2001), entre otros. En todos estos trabajos, la idea es utilizar la distancia de Cook como una medida de diagnóstico, para identificar observaciones influyentes, individualmente o en conjunto. Sin embargo, cuando se usa este criterio, se cuenta solamente con puntos críticos, los cuales son proporcionados por una aproximación a la distribución $\mathcal{F}$ centrada, tal como lo propuso Cook (1977). En este artículo, el propósito es realizar una modificación a esta distancia y derivar su distribución exacta.

Suponiendo que $\boldsymbol{Y} \in I\!R^{n\times p}$ tiene una distribución normal con media $\mu \in I\!R^{n\times p}$ y matriz de covarianzas $\Sigma \otimes \Theta \in I\!R^{np\times np}$ con $\Sigma \in I\!R^{p\times p}$, $\Sigma > 0$ y $\Theta \in I\!R^{n\times n}$, $\Theta > 0$, entonces la función de densidad es dada por

$$\boldsymbol{f}_{\boldsymbol{Y}}(\boldsymbol{Y}) = (2\pi)^{-pn/2}|\Sigma|^{-p/2}|\Theta|^{-n/2}\,\mathrm{etr}(\Sigma^{-1}(\boldsymbol{Y}-\mu)^T\Theta^{-1}(\boldsymbol{Y}-\mu))$$

Este hecho, se denota también, como $\boldsymbol{Y} \sim \mathcal{N}_{n\times p}(\mu, \Sigma \otimes \Theta)$.

Considerando el modelo de regresión lineal multivariado:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{5.1}$$

donde $\boldsymbol{Y} \in I\!R^{n\times p}$ es la matriz respuesta, $\boldsymbol{X} \in I\!R^{n\times q}$, con $\boldsymbol{r}(\boldsymbol{X}) = q$, $\boldsymbol{\beta} \in I\!R^{q\times p}$

es la matriz de parámetros desconocidos y $\epsilon \in I\!R^{n\times p}$ es una matriz de errores, tal que $\epsilon \sim \mathcal{N}_{n\times p}(0, \Sigma \otimes \boldsymbol{I}_n)$. Este modelo es conocido como modelo de regresión lineal normal multivariado. Los estimadores de máxima verosimilitud para $\boldsymbol{\beta}$ y Σ están dados por, ver Muirhead (pp. 83, 1982),

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y} = \boldsymbol{X}^{-}\boldsymbol{Y} \quad \text{y} \quad \widehat{\Sigma} = \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^{\boldsymbol{T}}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}),$$

donde $\boldsymbol{X}^{-}$ es la inversa de Moore-Penrose de $\boldsymbol{X}$.

Se considera entonces la regresión lineal con distribución normal multivariada, y se propone una extensión y modificación a la distancia de Cook. Esto permite derivar la distribución exacta para la nueva distancia, la cual, a su vez, proporciona un punto crítico para decidir si una observación en particular (o un conjunto de observaciones) se comportan como un "outlier".

2. DISTANCIA MODIFICADA : UNA OBSERVACIÓN

Considerando el modelo de regresión lineal normal multivariado con la siguiente modificación,

$$\boldsymbol{Y}_{(\boldsymbol{i})} = \boldsymbol{X}_{(\boldsymbol{i})}\boldsymbol{\beta}_{(\boldsymbol{i})} + \boldsymbol{\epsilon}_{(\boldsymbol{i})}, \qquad \boldsymbol{\epsilon}_{(\boldsymbol{i})} \sim \mathcal{N}_{(n-1)\times p}(0, \Sigma_{(i)} \otimes \boldsymbol{I}_n), \tag{5.2}$$

el cual se obtiene del modelo dado en (5.1), eliminando la i-ésima fila de $\boldsymbol{Y}$, $\boldsymbol{X}$ y ϵ, esto es, eliminando la i-ésima observación.

Para el modelo modificado, se tiene que:

$$\widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})} = (\boldsymbol{X}_{(\boldsymbol{i})}^{\boldsymbol{T}}\boldsymbol{X}_{(\boldsymbol{i})})^{-1}\boldsymbol{X}_{(\boldsymbol{i})}^{\boldsymbol{T}}\boldsymbol{Y}_{(\boldsymbol{i})} = \boldsymbol{X}_{(\boldsymbol{i})}^{-}\boldsymbol{Y}_{(\boldsymbol{i})} \quad \text{y}$$

$$\widehat{\Sigma}_{(\boldsymbol{i})} = \frac{1}{(n-1)}(\boldsymbol{Y}_{(\boldsymbol{i})} - \boldsymbol{X}_{(\boldsymbol{i})}\widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})})^{\boldsymbol{T}}(\boldsymbol{Y}_{(\boldsymbol{i})} - \boldsymbol{X}_{(\boldsymbol{i})}\widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})})$$

Primero, se requiere una representación simple para $\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})}$. Para esto, considerando la siguiente partición en las matrices:

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1^{\boldsymbol{T}} \\ \boldsymbol{Y}_2^{\boldsymbol{T}} \\ \vdots \\ \boldsymbol{Y}_n^{\boldsymbol{T}} \end{pmatrix}, \quad \boldsymbol{Y}_i \in I\!R^p \quad \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1{}^{\boldsymbol{T}} \\ \boldsymbol{\epsilon}_2{}^{\boldsymbol{T}} \\ \vdots \\ \boldsymbol{\epsilon}_n{}^{\boldsymbol{T}} \end{pmatrix}, \quad \boldsymbol{\epsilon}_{\boldsymbol{i}} \in I\!R^p \quad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1^{\boldsymbol{T}} \\ \boldsymbol{X}_2^{\boldsymbol{T}} \\ \vdots \\ \boldsymbol{X}_n^{\boldsymbol{T}} \end{pmatrix}, \quad \boldsymbol{X}_{\boldsymbol{i}} \in I\!R^q.$$

por consiguiente

$$\boldsymbol{X^T X} = (\boldsymbol{X_1 X_2 \cdots X_n}) \begin{pmatrix} \boldsymbol{X_1^T} \\ \boldsymbol{X_2^T} \\ \vdots \\ \boldsymbol{X_n^T} \end{pmatrix} = \sum_{k=1}^{n} \boldsymbol{X_k X_k^T} = \sum_{k \neq i}^{n} \boldsymbol{X_k X_k^T} + \boldsymbol{X_i X_i^T}$$

$$= \boldsymbol{X_{(i)}^T X_{(i)}} + \boldsymbol{X_i X_i^T},$$

y

$$\boldsymbol{X^T Y} = (\boldsymbol{X_1 X_2 \cdots X_n}) \begin{pmatrix} \boldsymbol{Y_1^T} \\ \boldsymbol{Y_2^T} \\ \vdots \\ \boldsymbol{Y_n^T} \end{pmatrix} = \sum_{k=1}^{n} \boldsymbol{X_k Y_k^T} = \sum_{k \neq i}^{n} \boldsymbol{X_k Y_k^T} + \boldsymbol{X_i Y_i^T}$$

$$= \boldsymbol{X_{(i)}^T Y_{(i)}} + \boldsymbol{X_i Y_i^T}.$$

Notando que $\boldsymbol{e_i^n}$ es el i-ésimo vector de la base canónica en $I\!R^n$, esto es, el vector unitario dado por $\boldsymbol{e_i^n} = (0 \cdots 0\ 1\ 0 \cdots 0)^{\boldsymbol{T}}$, entonces $\boldsymbol{e_i^{n^T} Y} = \boldsymbol{Y_i^T}$, $\boldsymbol{e_i^{n^T} X} = \boldsymbol{X_i^T}$ y $\boldsymbol{e_i^{n^T}} \epsilon = \epsilon_i{}^{\boldsymbol{T}}$.

Rao (pp. 33, 1973), señala que si $\boldsymbol{A}$ es no singular, $\boldsymbol{v}$ y $\boldsymbol{u}$ son dos vectores arbitrarios, entonces

$$(\boldsymbol{A - uv^T})^{-1} = \boldsymbol{A^{-1}} + \frac{\boldsymbol{A^{-1} u v^T A^{-1}}}{1 - \boldsymbol{v^T A^{-1} u}}$$

por lo tanto, si se define $\boldsymbol{A = X^T X}$ y $\boldsymbol{u = v = X_i}$, se obtiene

$$(\boldsymbol{X^T X - X_i X_i^T})^{-1} = (\boldsymbol{X_{(i)}^T X_{(i)}})^{-1}$$

$$= (\boldsymbol{X^T X})^{-1} + \frac{(\boldsymbol{X^T X})^{-1} \boldsymbol{X_i X_i^T} (\boldsymbol{X^T X})^{-1}}{(1 - \boldsymbol{p_{ii}})}, \quad (5.3)$$

con $\boldsymbol{p_{ii}} = \boldsymbol{X_i^T (X^T X)^{-1} X_i}$.

De (5.3), se obtiene que

$$\boldsymbol{\hat{\beta} - \hat{\beta}_{(i)}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T Y} - (\boldsymbol{X_{(i)}^T X_{(i)}})^{-1} \boldsymbol{X_{(i)}^T Y_{(i)}}$$

$$
\begin{aligned}
&= \left((X_{(i)}^T X_{(i)})^{-1} - \frac{(X^TX)^{-1}X_iX_i^T(X^TX)^{-1}}{(1-p_{ii})}\right) X^TY \\
&\quad -(X_{(i)}^T X_{(i)})^{-1}X_{(i)}^T Y_{(i)} \\
&= (X_{(i)}^T X_{(i)})^{-1}(X^TY - X_{(i)}^T Y_{(i)}) - \frac{(X^TX)^{-1}X_iX_i^T(X^TX)^{-1}X^TY}{(1-p_{ii})} \\
&= (X_{(i)}^T X_{(i)})^{-1}X_iY_i^T - \frac{(X^TX)^{-1}X_iX_i^T(X^TX)^{-1}X^TY}{(1-p_{ii})}. \qquad (5.4)
\end{aligned}
$$

Usando (5.3) en la primera parte de (5.4), se obtiene

$$
\begin{aligned}
(X_{(i)}^T X_{(i)})^{-1}X_iY_i^T &= \left((X^TX)^{-1} + \frac{(X^TX)^{-1}X_iX_i^T(X^TX)^{-1}}{(1-p_{ii})}\right) X_iY_i^T \\
&= (X^TX)^{-1}X_iY_i^T + \frac{p_{ii}(X^TX)^{-1}X_iY_i^T}{(1-p_{ii})} \\
&= \frac{(X^TX)^{-1}X_iY_i^T}{(1-p_{ii})}. \qquad (5.5)
\end{aligned}
$$

Sustituyendo (5.5) en (5.4), resulta

$$
\begin{aligned}
\widehat{\beta} - \widehat{\beta}_{(i)} &= \frac{(X^TX)^{-1}X_iY_i^T - (X^TX)^{-1}X_iX_i^T(X^TX)^{-1}X^TY}{(1-p_{ii})} \\
&= \frac{(X^TX)^{-1}X_i}{(1-p_{ii})}(Y_i^T - X_i^T(X^TX)^{-1}X^TY) \qquad (5.6)
\end{aligned}
$$

Ahora, puesto que $\widehat{\epsilon} = (Y - X\widehat{\beta}) = (I - XX^-)Y = (I-P)Y$, donde P es el proyector ortogonal sobre la imágen de X. Entonces

$$
e_i^{n^T}\widehat{\epsilon} = \widehat{\epsilon}_i^T = e_i^{n^T}(Y - X\widehat{\beta}) = Y_i^T - X_i^T(X^TX)^{-1}X^TY,
$$

así, se obtiene:

$$
\widehat{\beta} - \widehat{\beta}_{(i)} = \frac{(X^TX)^{-1}X_i\widehat{\epsilon}_i^T}{(1-p_{ii})} \qquad (5.7)
$$

Bajo el supuesto de la matriz de distribución normal, se propone la siguiente modificación a la distancia de Cook, denotada como $\mathcal{D}_m$:

$$
\mathcal{D}_m = \text{vec}(\widehat{\beta} - \widehat{\beta}_{(i)})^T\widehat{\text{Cov}}(\text{vec}(\widehat{\beta} - \widehat{\beta}_{(i)}))^- \text{vec}(\widehat{\beta} - \widehat{\beta}_{(i)}) \qquad (5.8)
$$

Nota 1. La expresión dada en (5.8), es una extensión del cuadrado de la distancia de Mahalanobis generalizada, ver Rao y Mitra (203-206, 1971).

El segundo paso, es encontrar una expresión simple para la matriz de varianzas y covarianzas $\text{Cov}(\text{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})}))$. Para lo cual, recordando que $\boldsymbol{Y} \sim \mathcal{N}_{n\times p}(\mu, \Sigma \otimes \Theta)$, entonces $\boldsymbol{E}(\boldsymbol{Y}) = \mu$, $\text{Cov}(\text{vec}(\boldsymbol{Y})) = (\Sigma \otimes \Theta)$, ver Muirhead (pp. 79, 1982). Puesto que $\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}^{\boldsymbol{T}} = \boldsymbol{e}_{\boldsymbol{i}}^{\boldsymbol{n}^T}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{e}_{\boldsymbol{i}}^{\boldsymbol{n}^T}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$, es claro que

$$\text{vec}\,\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}^{\boldsymbol{T}} = (\boldsymbol{I}_p \otimes \boldsymbol{e}_{\boldsymbol{i}}^{\boldsymbol{n}^T}(\boldsymbol{I} - \boldsymbol{P}))\,\text{vec}\,\boldsymbol{Y},$$

por consiguiente

$$\begin{aligned}
\text{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})}) &= \frac{(\boldsymbol{I}_p \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i)}{(1 - \boldsymbol{p}_{ii})}(\boldsymbol{I}_p \otimes \boldsymbol{e}_{\boldsymbol{i}}^{\boldsymbol{n}^T}(\boldsymbol{I} - \boldsymbol{P}))\,\text{vec}\,\boldsymbol{Y} \\
&= \frac{(\boldsymbol{I}_p \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i\boldsymbol{e}_{\boldsymbol{i}}^{\boldsymbol{n}^T}(\boldsymbol{I} - \boldsymbol{P}))}{(1 - \boldsymbol{p}_{ii})}\,\text{vec}\,\boldsymbol{Y} \\
&= \frac{(\boldsymbol{I}_p \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i\boldsymbol{H}_i^T)}{(1 - \boldsymbol{p}_{ii})}\,\text{vec}\,\boldsymbol{Y} \qquad (5.9)
\end{aligned}$$

donde $\boldsymbol{H}_i^T = \boldsymbol{e}_{\boldsymbol{i}}^{\boldsymbol{n}^T}(\boldsymbol{I} - \boldsymbol{P})$ es la i-ésima fila de la matriz $(\boldsymbol{I} - \boldsymbol{P})$. Entonces

$$\begin{aligned}
\text{Cov}(\text{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})})) &= \frac{(\boldsymbol{I}_p \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i\boldsymbol{H}_i^T)}{(1 - \boldsymbol{p}_{ii})}\,\text{Cov}(\text{vec}\,\boldsymbol{Y})\frac{(\boldsymbol{I}_p \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i\boldsymbol{H}_i^T)^T}{(1 - \boldsymbol{p}_{ii})} \\
&= \frac{(\boldsymbol{I}_p \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i\boldsymbol{H}_i^T)}{(1 - \boldsymbol{p}_{ii})^2}(\Sigma \otimes \boldsymbol{I})(\boldsymbol{I}_p \otimes \boldsymbol{H}_i\boldsymbol{X}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \\
&= \frac{||\boldsymbol{H}_i||^2(\Sigma \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1})}{(1 - \boldsymbol{p}_{ii})^2} \qquad (5.10)
\end{aligned}$$

Notando que

$$\begin{aligned}
||\boldsymbol{H}_i||^2 &= \boldsymbol{e}_{\boldsymbol{i}}^{\boldsymbol{n}^T}(\boldsymbol{I} - \boldsymbol{P})(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{e}_{\boldsymbol{i}}^{\boldsymbol{n}} \\
&= \boldsymbol{e}_{\boldsymbol{i}}^{\boldsymbol{n}^T}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{e}_{\boldsymbol{i}}^{\boldsymbol{n}} \\
&= \boldsymbol{e}_{\boldsymbol{i}}^{\boldsymbol{n}^T}\boldsymbol{e}_{\boldsymbol{i}}^{\boldsymbol{n}} - \boldsymbol{e}_{\boldsymbol{i}}^{\boldsymbol{n}^T}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{e}_{\boldsymbol{i}}^{\boldsymbol{n}} \\
&= 1 - \boldsymbol{X}_i(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i^T \\
&= 1 - \boldsymbol{p}_{ii} \qquad (5.11)
\end{aligned}$$

Sustituyendo (5.11) en (5.10), se obtiene

$$\text{Cov}(\text{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})})) = \frac{(\Sigma \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1})}{(1 - \boldsymbol{p}_{ii})} \qquad (5.12)$$

Sea $S_1 = n\widehat{\Sigma}/(n-q)$ y observando que $E(S_1) = \Sigma$, ver Muirhead (pp. 84, 1982), entonces

$$\widehat{\mathrm{Cov}}(\mathrm{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})) = \frac{(S_1 \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i\boldsymbol{X}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1})}{(1-\boldsymbol{p_{ii}})} \tag{5.13}$$

Sea $\boldsymbol{r_i} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X_i}$. Basándose en los siguientes resultados:

1. Para $a \in I\!R^n$, $a^- = a^T/||a||^2$,
2. Dada $\boldsymbol{A} \in I\!R^{p\times q}$, entonces $(\boldsymbol{AA^T})^- = \boldsymbol{A^{T-}A^-}$ con $\boldsymbol{A^{-1}} = \boldsymbol{A^-}$ si $\boldsymbol{A}$ es no singular,
3. Dadas las matrices $\boldsymbol{A}$ y $\boldsymbol{B}$, $(\boldsymbol{A}\otimes\boldsymbol{B})^- = \boldsymbol{A^-}\otimes\boldsymbol{B^-}$,

se obtiene

$$\begin{aligned}
(\widehat{\mathrm{Cov}}(\mathrm{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})))^- &= \widehat{\mathrm{Cov}}(\mathrm{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}))^- \\
&= \left(\frac{(S_1 \otimes \boldsymbol{r_i}\boldsymbol{r_i^T})}{(1-\boldsymbol{p_{ii}})}\right)^- \\
&= \frac{(1-\boldsymbol{p_{ii}})}{||\boldsymbol{r_i}||^4}({S_1}^{-1} \otimes \boldsymbol{r_i}\boldsymbol{r_i^T})
\end{aligned}$$

Por lo tanto, la distancia de Cook modificada puede ser re-escrita como:

$$\begin{aligned}
\mathcal{D}_m &= \mathrm{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})^T\widehat{\mathrm{Cov}}(\mathrm{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}))^- \mathrm{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}) \\
&= \left(\frac{(\boldsymbol{I}_p \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i\boldsymbol{H}_i^T)\,\mathrm{vec}\,\boldsymbol{Y}}{(1-\boldsymbol{p_{ii}})}\right)^T \frac{(1-\boldsymbol{p_{ii}})({S_1}^{-1} \otimes \boldsymbol{r_i}\boldsymbol{r_i^T})}{||\boldsymbol{r_i}||^4} \\
&\quad \left(\frac{(\boldsymbol{I}_p \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_i\boldsymbol{H}_i^T)\,\mathrm{vec}\,\boldsymbol{Y}}{(1-\boldsymbol{p_{ii}})}\right) \\
&= \frac{(1-\boldsymbol{p_{ii}})^{-1}}{||\boldsymbol{r_i}||^4}\,\mathrm{vec}^T\boldsymbol{Y}({S_1}^{-1} \otimes \boldsymbol{H_i}\boldsymbol{r_i^T}\boldsymbol{r_i}\boldsymbol{r_i^T}\boldsymbol{r_i}\boldsymbol{H_i^T})\,\mathrm{vec}\,\boldsymbol{Y} \\
&= (1-\boldsymbol{p_{ii}})^{-1}\,\mathrm{vec}^T\boldsymbol{Y}({S_1}^{-1} \otimes \boldsymbol{H_i}\boldsymbol{H_i^T})\,\mathrm{vec}\,\boldsymbol{Y}.
\end{aligned} \tag{5.14}$$

Alternativamente, ya que

$$\boldsymbol{tr}(\boldsymbol{BX^TCXD}) = \mathrm{vec}^T\boldsymbol{X}(\boldsymbol{B^TD^T}\otimes\boldsymbol{C})\,\mathrm{vec}\,\boldsymbol{X} = \mathrm{vec}^T\boldsymbol{X}(\boldsymbol{DB}\otimes\boldsymbol{C^T})\,\mathrm{vec}\,\boldsymbol{X},$$

para matrices de órdenes adecuados, se puede escribir $\mathcal{D}_m$ como

$$\mathcal{D}_m = (1 - \boldsymbol{p_{ii}})^{-1} \boldsymbol{tr}(S_1{}^{-1} \boldsymbol{Y^T H_i H_i^T Y}).$$

Por otro lado, puesto que $\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}{}^{\boldsymbol{T}} = \boldsymbol{e_i^{n^T}}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{H_i Y}$, entonces

$$\begin{aligned} \mathcal{D}_m &= (1 - \boldsymbol{p_{ii}})^{-1} \boldsymbol{tr}(S_1{}^{-1} \widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}} \widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}^{\boldsymbol{T}}) \\ &= (1 - \boldsymbol{p_{ii}})^{-1} \boldsymbol{tr}(\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}^{\boldsymbol{T}} S_1{}^{-1} \widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}) \\ &= (1 - \boldsymbol{p_{ii}})^{-1} \widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}^{\boldsymbol{T}} S_1{}^{-1} \widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}} \end{aligned}$$

Por lo que se tienen las siguientes expresiones como alternativa para el cuadrado de la distancia de Cook modificada:

$$\mathcal{D}_m = \begin{cases} \operatorname{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})})^{\boldsymbol{T}} \widehat{\operatorname{Cov}}(\operatorname{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})}))^{-} \operatorname{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})}) \\ (1 - \boldsymbol{p_{ii}})^{-1} \operatorname{vec}^{\boldsymbol{T}} Y (S_1{}^{-1} \otimes \boldsymbol{H_i H_i^T}) \operatorname{vec} \boldsymbol{Y} \\ (1 - \boldsymbol{p_{ii}})^{-1} \operatorname{tr}\left(S_1{}^{-1} \boldsymbol{Y^T H_i P_i^T Y}\right) \\ (1 - \boldsymbol{p_{ii}})^{-1} \widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}^{\boldsymbol{T}} S_1{}^{-1} \widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}} \end{cases} \tag{5.15}$$

Nota 2. De acuerdo con Chatterjee y Hadi (pp. 124, 1988), se puede reemplazar la matriz S_1 por otra obtenida usando la muestra reducida $(n-1)$, denotada por $S_1(i)$.

Nota 3. Cook (1977), Chatterjee y Hadi (pp. 117, 1988), Díaz-García et. al (2001), y muchos otros autores, utilizan la matriz de varianzas y covarianzas del $vec(\widehat{\boldsymbol{\beta}})$ para construir la medida de distancia. La reformulación que se propone, está basada en el reemplazo de ésta matriz, por la matriz de varianzas y covarianzas del $\operatorname{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(\boldsymbol{i})})$. Se puede encontrar esta idea en Chatterjee y Hadi (pp. 150, 1988), para el caso univariado, pero para la evaluación de observaciones influyentes sobre un particular coeficiente de regresión, solamente se utiliza la varianza de un coeficiente, en vez de la varianza de la diferencia. El problema que se presenta, cuando esta idea es extendida al caso multivariado, es que tal matriz es singular,

por lo que se necesita considerar la inversa de Moore-Penrose para la matriz de varianzas y covarianzas del $\text{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})$.

Teorema 1. *Considerando el modelo de regresión normal multivariado dado en (5.1). Entonces, el cuadrado de la distancia de Cook modificada para detectar un "outlier", puede ser escrita como:*

$$\mathcal{D}_m = \begin{cases} \text{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})^{\boldsymbol{T}} \left(\dfrac{(S_1 \otimes (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_i\boldsymbol{X}_i^{\boldsymbol{T}}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1})}{(1-\boldsymbol{p}_{ii})} \right) - \text{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}) \\ (1-\boldsymbol{p}_{ii})^{-1} \,\text{vec}^{\boldsymbol{T}}\boldsymbol{Y}(S_1{}^{-1} \otimes \boldsymbol{H}_i\boldsymbol{H}_i^{\boldsymbol{T}}) \,\text{vec}\,\boldsymbol{Y} \\ (1-\boldsymbol{p}_{ii})^{-1} \,\text{tr}\left(S_1{}^{-1}\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{H}_i\boldsymbol{P}_i^{\boldsymbol{T}}\boldsymbol{Y}\right) \\ (1-\boldsymbol{p}_{ii})^{-1}\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}^{\boldsymbol{T}} S_1{}^{-1}\widehat{\boldsymbol{\epsilon}}_{\boldsymbol{i}} \end{cases} \tag{5.16}$$

Nota 4. En (5.16) es fácil ver que si se quiere implementar esta medida para todo el conjunto de datos, es suficiente ajustar el modelo una sola vez, y de la forma usual se puede construir la distancia modificada para cada punto. Notando que la expresión $\mathcal{D}_m$, en el caso normal univariado, coincide con el análisis de residuales estudentizados, ver Besley et al. (p. 201, 1980) y Chatterjee y Hadi (p. 78, 1988).

3. DISTANCIA MODIFICADA : MULTIPLES OBSERVACIONES

Sea $I = \{i_1, i_2, \ldots, i_k\}$ un subconjunto de tamaño k de $\{1, 2, \ldots, n\}$, de forma tal que $(n-k) \geq q$. Ahora, bajo el modelo (5.1), denote por $\boldsymbol{X}_{(I)}$, $\boldsymbol{Y}_{(I)}$ y $\widehat{\boldsymbol{\epsilon}}_{(I)}$, las matrices de regresión, de datos y de errores, respectivamente, después de eliminar las correspondientes observaciones de acuerdo con los subíndices en I. Sea $\widehat{\boldsymbol{\beta}}_{(I)}$, y $\widehat{\Sigma}_{(I)}$ los correspondientes estimadores de máxima verosimilitud en el modelo

$$\boldsymbol{Y}_{(I)} = \boldsymbol{X}_{(I)}\boldsymbol{\beta}_{(I)} + \boldsymbol{\epsilon}_{(I)}, \qquad \boldsymbol{\epsilon}_{(I)} \sim \mathcal{N}_{(n-k)\times p}(0, \Sigma_{(I)} \otimes \boldsymbol{I}_n).$$

Basándose en el Lema 2, y usando procedimientos similares a los de la sección 2, es fácil verificar que

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(I)} \;=\; (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I} - \boldsymbol{P}_I)^{-1}\widehat{\boldsymbol{\epsilon}}_I,$$

con $(\boldsymbol{I}-\boldsymbol{P}_I) = (\boldsymbol{I}_k - \boldsymbol{X}_I{}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_I)$ y $\boldsymbol{X}_I$ es la matriz con las correspondientes filas de $\boldsymbol{X}$ de acuerdo con I. Observando que $\widehat{\epsilon}_I = \boldsymbol{U}_I{}^T\widehat{\epsilon} = \boldsymbol{U}_I{}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}$, donde

$$\boldsymbol{U}_I{}^T = \begin{pmatrix} e_{i_1}^{n^T} \\ e_{i_2}^{n^T} \\ \vdots \\ e_{i_k}^{n^T} \end{pmatrix}$$

Se obtiene

$$\operatorname{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(I)}) = (\boldsymbol{I}_p \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\boldsymbol{H}_I)\operatorname{vec}\boldsymbol{Y},$$

con $\boldsymbol{H}_I = \boldsymbol{U}_I{}^T(\boldsymbol{I}-\boldsymbol{P})$.

De donde

$$\operatorname{Cov}(\operatorname{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(I)})) = (\Sigma \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\boldsymbol{X}_I{}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

y

$$\widehat{\operatorname{Cov}}(\operatorname{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(I)})) = (S_1 \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\boldsymbol{X}_I{}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

Entonces se tiene,

Teorema 2. *Considerando el modelo de regresión normal multivariado dado en (5.1). Entonces, el cuadrado de la distancia de Cook modificada para detectar k observaciones influyentes, puede ser escrita como:*

$$\mathcal{D}_{m_I} = \begin{cases} \operatorname{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(I)})^T(S_1 \otimes (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_I(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\boldsymbol{X}_I{}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1})^- \operatorname{vec}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(I)}) \\ \operatorname{vec}{}^T\boldsymbol{Y}(S_1{}^{-1} \otimes \boldsymbol{H}_I{}^T(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\boldsymbol{H}_I)\operatorname{vec}\boldsymbol{Y} \\ tr(S_1{}^{-1}\boldsymbol{Y}^T\boldsymbol{H}_I{}^T(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\boldsymbol{H}_I\boldsymbol{Y}) \\ tr\left(S_1{}^{-1}\widehat{\epsilon}_I^T(\boldsymbol{I}-\boldsymbol{P}_I)^{-1}\widehat{\epsilon}_I\right) \end{cases} \tag{5.17}$$

4. FUNCIONES DE DISTRIBUCIÓN ASOCIADAS CON LAS DISTANCIAS MODIFICADAS

La razón principal para estudiar las modificaciones dadas en las secciones 2 y 3 para el cuadrado de la distancia de Cook es que, en vez de utilizar una aproximación a la distribución $\mathcal{F}$, se derivará la distribución exacta para $\mathcal{D}_m$. Análogamente se encuentra la distribución exacta para $\mathcal{D}_{m_I}$, es decir, para el caso de la detección de varias observaciones influyentes, simultáneamente.

Se derivará la distribución del estadístico para el caso de una observación influyente a la vez y para el caso múltiples observaciones.

Teorema 3. *Bajo los supuestos del Teorema 1, se tiene*

$$\frac{(n-q-p+1)}{p(n-q)}\mathcal{D}_m \sim \mathcal{F}_{p,(n-q-p+1)} \tag{5.18}$$

donde $\mathcal{F}_{p,(n-q-p+1)}$ *denota una distribución* $\mathcal{F}$ *centrada, con* p *y* $(n-q-p+1)$ *grados de libertad (gl).*

Prueba: Se sigue directamente del Teorema 5.2.2 en Anderson (p. 163, 1984). ■

Del Teorema 3, dando un nivel de significancia α, se puede escribir la siguiente regla de decisión:

Y_i, $i = 1, 2, \ldots, n$, es un "outlier", si

$$\frac{(n-q-p+1)}{p(n-q)}\mathcal{D}_m \geq \mathcal{F}_{\alpha:p,(n-q-p+1)} \tag{5.19}$$

donde $\mathcal{F}_{\alpha:p,(n-q-p+1)}$ es el correspondiente $\alpha - percentil$ superior de una distribución $\mathcal{F}$ con p y $(n-q-p+1)$ *gl.*

Nota 5. Para el caso univariado, $p = 1$, la regla de decisión se convierte en: Y_i, $i = 1, 2, \ldots, n$, es un "outlier" si

$$\mathcal{D}_m \geq \mathcal{F}_{\alpha:1,(n-q)} \tag{5.20}$$

donde $\mathcal{F}_{\alpha:1,(n-q)}$ es el $\alpha - percentil$ de una distribución $\mathcal{F}$ con 1 y $(n-q)$ gl, o equivalentemente:

$$\mathcal{D}_m^{1/2} \geq \boldsymbol{t}_{\alpha/2:(n-q)} \tag{5.21}$$

donde $\boldsymbol{t}_{\alpha/2:(n-q)}$ es el $\alpha/2 - percentil$ superior de una distribución $\boldsymbol{t}$ con $(n-q)$ gl.

Similarmente, para el caso de múltiples observaciones:

Teorema 4. *Bajo los supuestos del Teorema 2, se tiene que*

$$\mathcal{D}_{m_I} \sim \mathcal{LH}_{s,m,h} \tag{5.22}$$

donde $\mathcal{LH}_{s,m,h}$ denota la distribución centrada para el estadístico Lawley-Hotelling con parámetros $s = \min(p,k)$, $m = (|p-k|-1)/2$ y $h = (n-q-p+1)/2$.

Prueba: Se sigue directamente del Teorema 10.6.2, Corolario 10.6.3, en Muirhead [pp. 468-471 y p. 471, 1982]. ∎

5. UNA APLICACION

Se ilustra la utilización de la prueba exacta dada en la sección 4, bajo una regresión simple y bajo una regresión múltiple multivariada.

El primer conjunto de datos fue presentado por Cook y Weisberg (pp. 204-207, 1994). Este es un conjunto de datos pequeño, con observaciones sobre 21 niños, dando la edad (AGE) en meses, en la que dijeron su primera palabra, y un $SCORE$, el cual es una medida del desarrollo del niño. En la figura *1(a)*, se presenta la gráfica $(AGE, SCORE)$. Es claro que las observaciones 18, 1 y 17, muestran un comportamiento diferente. Si se considera un ajuste lineal usando cuadrados mínimos ordinarios, el $SCORE$ parece decrecer con AGE. El caso 18 parece tener un pobre ajuste hacia una tendencia lineal, comparado con los demás datos. Los casos 1 y 17, tienen relativamente valores grandes de AGE. La figura *1(b)*, muestra el gráfico Q-Q de los residuales provenientes del ajuste lineal, usando cuadrados mínimos or-

dinarios. Es claro que, solamente la observación 18 parece ser un candidato para ser considerado "outlier", tal como es definido en Chatterjee y Hadi (pp.94-95, 1988). La figura *2* muestra la identificación y detección de "outliers" y puntos influyentes, usando diferentes técnicas. Este análisis enfatiza el hecho mencionado antes, es decir, los residuales estudentizados coinciden con la distancia modificada para el caso univariado. Además, los residuales estudentizados son la base para la distancia de Draper y John, y se discuten en Draper y Smith (pp. 169-175, 1981).

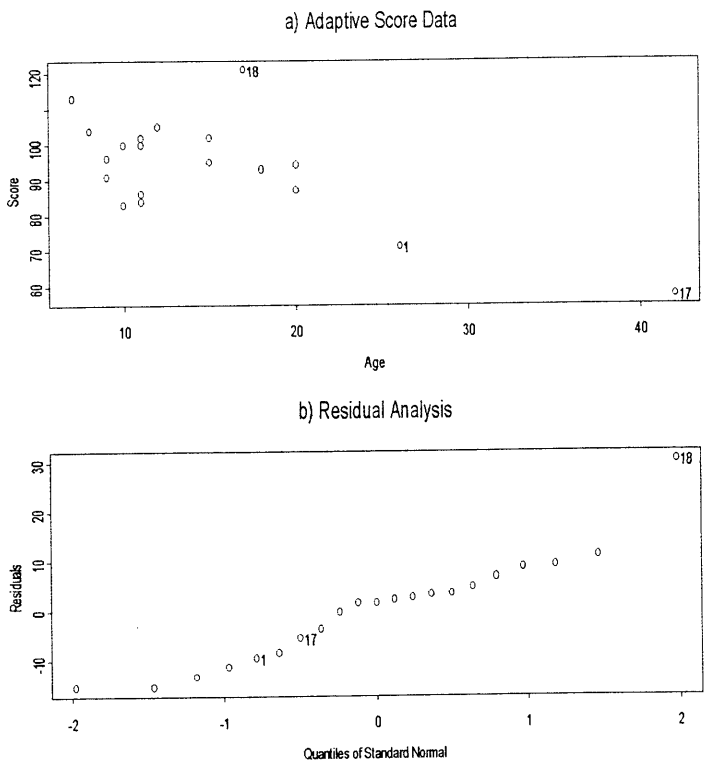


Figura 1. Datos originales para el ajuste del *SCORE* y un gráfico Q-Q para los residuales en la regresión simple de *SCORE* sobre *AGE*.

La figura *2(a)*, muestra cómo la distancia de Cook detecta la observación 17, la cual es un punto palanca, tal como se describe en Cook y Weisberg (1994). Considerando la figura *1(b)* y de acuerdo con las figuras *2(b)* y *2(c)*, la observación 18, es un candidato para ser considerado "outlier", y las pruebas están a favor para declararla

como tal. Notando que el valor crítico en la figura *2(c)*, es el valor aproximado de una distribución $\mathcal{F}$, multiplicado por $s^2(i)$ para $i = 18$, de acuerdo al gráfico de la distancia original de Draper y John.

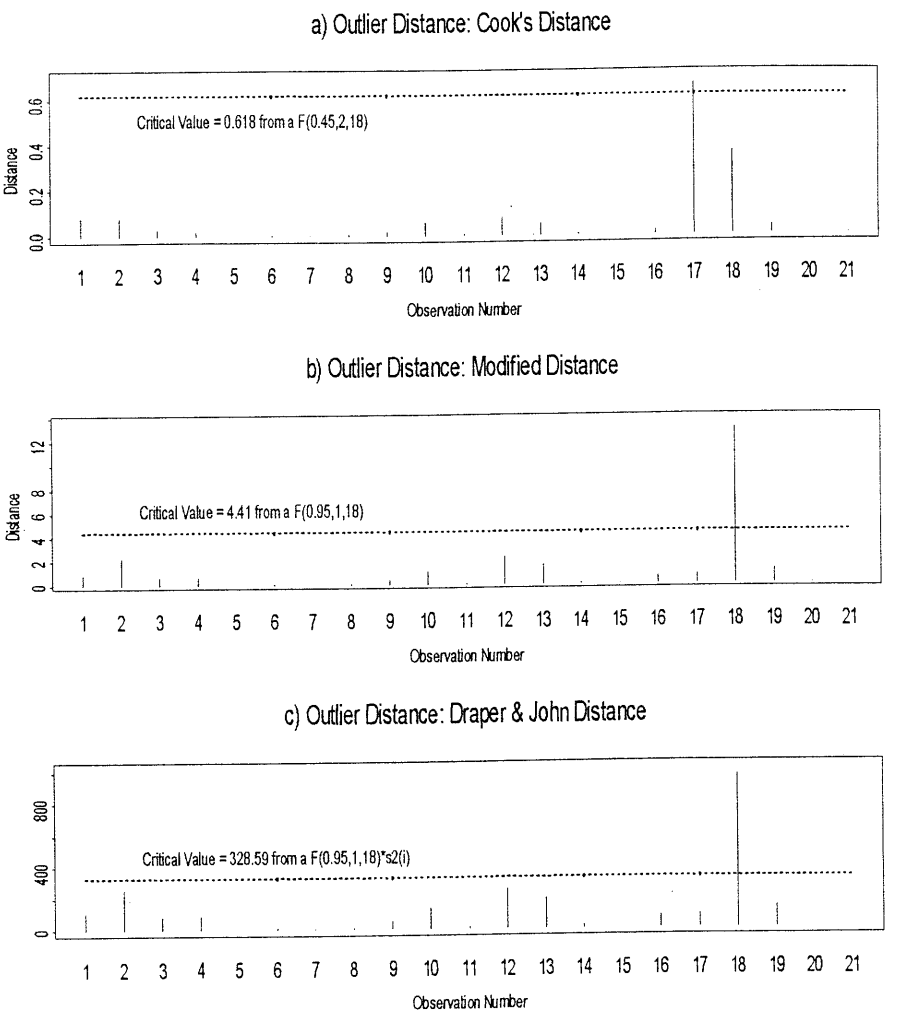


Figure 2. Identificación de la influencia y "outliers", basados en a) Distancia de Cook, b) Distancia de Cook Modificada de acuerdo con (5.20), c) Distancia de Draper y John. $n = 21$, $q = 2$, $p = 1$ y $s^2(i)$. La varianza de los residuales sin la i-ésima observación se usa en todos los casos, de acuerdo con la Nota 2.

Para el ejemplo de la regresión múltiple multivariada, se generan 19 observaciones de un modelo $y = X\beta + \epsilon$, con errores normales, $p = 2$, y $q = 4$.

Se ajusta el modelo, la matriz de la suma de cuadrados de residuales, entonces aplicando el cuadrado de la distancia de Mahalanobis, como se muestra en Seber (pp. 152-153, 1984). La figura *3(a)* presenta el cuadrado de la distancia de Mahalanobis y sugiere a las observaciones 10 y 11 como posibles "outliers". Aplicando (5.22), con $k = 2$ y $Dm = 4.77$, comparado con un valor crítico de 3.015, el percentil es aproximado por una distribución $\mathcal{F}$, tal como sugiere Seber (pp. 38-39, 563-564, 1984). La prueba está a favor de considerar las observaciones 10 y 11 como "outliers".

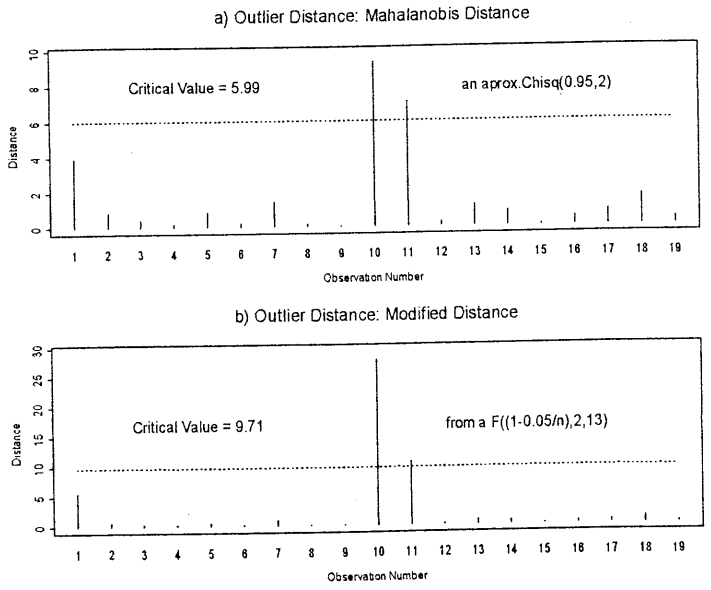


Figure 3. Identificación de "outliers" basada en la Distancia de Mahalanobis sobre la matriz de residuales, y detección de "outliers" basada en la Distancia de Cook Modificada dada en el Teorema 3.

La figura *3(b)*, muestra el mismo análisis tomando una observación a la vez. Se usa una prueba $\mathcal{F}$, basada en $(1 - \alpha/n)$ en vez de $(1 - \alpha)$, para obtener una prueba simultánea con un nivel mínimo α. En este caso, se obtiene la misma conclusión, como con la prueba basada en k observaciones.

Se recomienda el uso de la prueba basada en k observaciones, dada en (5.22), y se

seleccionan los k puntos, usando un método gráfico, como el mostrado en la figura *3(a)*.

REFERENCIAS

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, New York.

Besley, D.A., Kuh, E., and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley & Sons, New York.

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

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

Cook, R. D., and Weisberg, S. (1982). *Residual and Influence in Regression*, Chapman and Hall, London.

Cook, R. D., and Weisberg, S. (1994). *An Introduction to Regression Graphics*, John Wiley & Sons. New York.

Díaz-García, J.A., Galea, M., and Leiva- Sánchez, V. (2001). "Influence Diagnostics for Elliptical Regression Linear Models", Submitted for Publication.

Draper, N., and Smith, H. (1981). *Applied Regression Analysis*, (2nd ed.), John Wiley & Sons, New York.

Galea, M., Paula, G., and Bolfarine, H. (1997). "Local Influence in Elliptical Linear Regression Models", *The Statistician* 46, 71-79.

Liu, S.Z. (2000). "On Local Influence for Elliptical Linear Models", *Statistical Papers*, 41, 211-224.

Muirhead, R. J. (1982). *Aspects of Multivariated Statistical Theory*, John Wiley & Sons, New York.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications* (2nd ed.), John

Wiley & Sons, New York.

Rao, C. R. and Mitra, S. K.(1971). *Generalized Inverse of Matrices and its Applications* (2nd ed.), John Wiley & Sons, New York.

Seber, G.A.F.(1984). *Multivariate Observations* , John Wiley & Sons, New York.

LITERATURA CITADA

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, New York.

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

Atkinson, A.C. (1981). "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.

Beckman, R.J., and Trussel, H.J. (1974). "The Distribution of an Arbitrary Studentized Residual and the Effects of Updating in Multiple Regresion," *Journal of the American Statistical Association*, 69, 199-201.

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

Besley, D.A., Kuh, E., and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley & Sons, New York.

Bingham, C. (1977). "Some Identities Useful in the Analysis of Residuals from Linear Regression," Technical Report 300, School of Statistics, University of Minesota.

Chatterjee, S., and Hadi, A.S. (1986). "Influential Observation, High Leverage Points, and Outliers in Linear Regression," *Statistical Science*, 1, No. 3, 379-416.

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

Chatterjee, S., and Price, B. (1977). *Regression Analysis by Example*, John Wiley & Sons, New York.

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

Cook, R.D., and Weisberg, S. (1980). "Characterization of an Empirical Influence Function for Detecting Influential Cases in Regression," *Technometrics*, 22, 495-508.

Cook, R.D., and Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman and Hall, London.

Cook, R. D., and Weisberg, S. (1994). *An Introduction to Regression Graphics*, John Wiley & Sons. New York.

Díaz-García, J.A., Galea, M., and Leiva- Sánchez, V. (2001). "Influence Diagnostics for Elliptical Regression Linear Models", Submitted for Publication.

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

Draper, N.R., and Smith, H. (1981). *Applied Regression Analysis*, 2nd ed., John Wiley & Sons, New York.

Ellenberg, J.H. (1973). "The Joint Distribution of the Standardized Least Squares Residuals From a General Linear Regression," *Journal of the American Statistical Association*, 68, 941-943.

Ellenberg, J.H. (1976). "Testing for Single Outlier From a General Linear Regression Model," *Biometrics*, 32, 637-645.

Galea, M., Paula, G., and Bolfarine, H. (1997). "Local Influence in Elliptical Linear Regression Models", *The Statistician* 46, 71-79.

Hampel, F.R. (1968). "Contributions to the Theory of Robust Estimation," Ph.D. thesis, University of California, Berkeley.

Hampel, F.R. (1974). "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association*, 62, 1179-1186.

Henderson, H.V., and Searle, S.R. (1981). "On Deriving the Inverse of a Sum of Matrices," *SIAM Review*, 23, 53-60.

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

Huber, P. (1977). *Robust Statistical Procedures*, No. 27, Regional Conference Series in Applied Mathematics, Philadelphia: SIAM.

Huber, P. (1981). *Robust Statistics*, John Wiley & Sons, New York.

Liu, S.Z. (2000). "On Local Influence for Elliptical Linear Models", *Statistical Papers*, 41, 211-224.

Lund, R.E. (1975). "Tables for an Approximate Test for Outliers in Linear Models," *Technometrics*, 17, 473-476.

Muirhead, R. J. (1982). *Aspects of Multivariated Statistical Theory*, John Wiley & Sons, New York.

Pregibon, D. (1979). "Data Analytic Methods for Generalized Linear Models," Unpublished Ph.D. Thesis, University of Toronto.

Pregibon, D. (1981). "Logistic Regression Diagnostics," *The Annals of Statistics*, 9, No. 4, 705-724.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications* (2nd ed.), John Wiley & Sons, New York.

Rao, C. R. and Mitra, S. K.(1971). *Generalized Inverse of Matrices and its Applications* (2nd ed.), John Wiley & Sons, New York.

Searle, S.R. (1982). *Matrix Algebra Useful for Statistics*, John Wiley & Sons, New York.

Seber, G.A.F.(1984). *Multivariate Observations*, John Wiley & Sons, New York.

Tietjen, G.L., Moore, R.H., and Beckman, R.J. (1973). "Testing for a Single Outlier in Simple Linear Regression," *Technometrics*, 15, 717-721.

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

Welsch, R.E., and Kuh, E. (1977). "Linear Regression Diagnostics," *Technical Report 923-77*, Sloan School of Management, Massachussetts Institute of Technology.

Welsch, R.E., and Peters, S.C. (1978). "Finding Influential Subsets of Data in Regression Models," *Proceedings of the Eleventh Interface Symposium on Computer Science and Statistics*, (A.R. Gallant and T.M. Gerig, eds.), Raleigh: Institute of Statistics, North Carolina State University.