ACTA UNIVERSITARIA
Guanajuato, Gto., México

Una Modificación a la Distancia de Cook.

José Antonio Díaz García* y Oscar Alejandro Martínez Jaime**.

RESUMEN / ABSTRACT

En este trabajo 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 multivariada.

A modification of the classical Cook´s distance is proposed in this paper, being based on the distance of widespread Mahalanobis in the context of multivariate normal linear regression model. Furthermore, the exact distribution of a pivotal type statistics based on this generalized Mahalanobis distance is established, providing critical points for the identification of outliers in data points. The procedure is illustrated with an example, in the case of multiple and multivariate linear regression.

Recibido: 1 de Marzo de 2004

Aceptado: 12 de Agosto de 2004

* Departamento de Estadística y Cálculo. Universidad Autónoma Agraria "Antonio Narro". Buenavista, Saltillo, Coahuila, México. Correo electrónico: jadiaz@uaaan.mx

** Unidad de Estudios Superiores de Salvatierra. Universidad de Guanajuato. Salvatierra, Guanajuato, México. Correo electrónico: oscarja@dulcinea.ugto.mx

INTRODUCCIÓN

El problema de la identificación de "outliers" o puntos influyentes (puntos atípicos), 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; 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 F centrada, tal como lo propuso Cook, 1977. En este trabajo, el propósito es realizar una modificación a esta distancia y derivar su distribución exacta, a partir de la cual se establece una regla de decisión exacta, ver Martínez Jaime (2001).

Suponiendo que $\mathbf{Y} \in \Re^{n\times p}$ tiene una distribución normal con media $\boldsymbol{\mu} \in \Re^{n\times p}$ y matriz de covarianzas $\Sigma \otimes \Theta \in \Re^{np\times np}$ con $\Sigma \in \Re^{p\times p}$, $\Sigma > 0$ y $\Theta \in \Re^{n\times n}$, $\Theta > 0$, entonces la función de densidad está dada por $f_{\mathbf{Y}}(\mathbf{Y}) = (2\pi)^{-\frac{pn}{2}} |\Sigma|^{-\frac{p}{2}} |\Theta|^{-\frac{n}{2}} etr\left(\Sigma^{-1}(\mathbf{Y}-\boldsymbol{\mu})^T \Theta^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right)$. Este hecho, se denota también como $\mathbf{Y} \sim N_{n\times p}(\boldsymbol{\mu}, \Sigma \otimes \Theta)$.

Considerando el modelo de regresión lineal multivariado:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (1)$$

donde $\mathbf{Y} \in \Re^{n\times p}$ es la matriz respuesta, $\mathbf{X} \in \Re^{n\times q}$ con rango $r(\mathbf{X}) = q$, $\boldsymbol{\beta} \in \Re^{q\times p}$ es la matriz de parámetros desconocidos y $\boldsymbol{\varepsilon} \in \Re^{n\times p}$ es una matriz de errores, tal que $\boldsymbol{\varepsilon} \sim N_{n\times p}(\mathbf{0}, \Sigma \otimes \mathbf{I}_n)$.

PALABRAS CLAVE: Medidas de diagnóstico; Distancia de Mahalanobis generalizada; Puntos influyentes.

KEYWORDS: Diagnostic measures; Generalized Mahalanobis distance; Influential points.

Éste es conocido como modelo de regresión lineal normal multivariado. Los estimadores de máxima verosimilitud para $\boldsymbol{\beta}$ y Σ están dados por

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

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

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".

MÉTODOS

DISTANCIA MODIFICADA: UNA OBSERVACIÓN

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

$$\mathbf{Y}_{(i)} = \mathbf{X}_{(i)}\boldsymbol{\beta}_{(i)} + \boldsymbol{\varepsilon}_{(i)}, \ \boldsymbol{\varepsilon}_{(i)} \sim N_{(n-1)\times p}\left(\mathbf{0}, \Sigma_{(i)} \otimes \mathbf{I}_n\right), \tag{2}$$

el cual se obtiene del modelo dado en (1), eliminando la i-ésima fila de $\mathbf{Y}$, $\mathbf{X}$ y $\boldsymbol{\varepsilon}$, esto es, eliminando la i-ésima observación.

Para el modelo modificado, los estimadores son:

$$\hat{\boldsymbol{\beta}}_{(i)} = \left(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)} = \mathbf{X}_{(i)}^{-}\mathbf{Y}_{(i)} \quad \text{y} \quad \hat{\Sigma}_{(i)} = \frac{1}{(n-1)}\left(\mathbf{Y}_{(i)} - \mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)}\right)^T\left(\mathbf{Y}_{(i)} - \mathbf{X}_{(i)}\hat{\boldsymbol{\beta}}_{(i)}\right).$$

En el primer paso, se requiere una representación simple para $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$. Para lo cual, se considera la siguiente partición en las matrices:

$$\mathbf{Y} = \begin{pmatrix} Y_1^T \\ Y_2^T \\ \cdot \\ \cdot \\ \cdot \\ Y_n^T \end{pmatrix}, \ Y_i \in \Re^p; \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1^T \\ \varepsilon_2^T \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n^T \end{pmatrix}, \ \varepsilon_i \in \Re^p; \qquad \mathbf{X} = \begin{pmatrix} X_1^T \\ X_2^T \\ \cdot \\ \cdot \\ \cdot \\ X_n^T \end{pmatrix}, \ X_i \in \Re^q$$

por consiguiente

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} X_1 & X_2 & \cdot & \cdot & \cdot & X_n \end{pmatrix}\begin{pmatrix} X_1^T \\ X_2^T \\ \cdot \\ \cdot \\ \cdot \\ X_n^T \end{pmatrix} = \sum_{k=1}^{n} X_k X_k^T = \sum_{k\neq i}^{n} X_k X_k^T + X_i X_i^T = X_{(i)}^T X_{(i)} + X_i X_i^T \quad , \ \text{y}$$

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

Notando que $\mathbf{e}_i^n$ es el i-ésimo vector de la base canónica en $\Re^n$, esto es, el vector unitario dado por $\mathbf{e}_i^n = (0 \ \ldots \ 0 \ 1 \ 0 \ \ldots \ 0)^T$, entonces $\mathbf{e}_i^{nT}\mathbf{Y} = \mathbf{Y}_i^T$, $\mathbf{e}_i^{nT}\mathbf{X} = \mathbf{X}_i^T$ y $\mathbf{e}_i^{nT}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_i^T$. Rao, 1973, señala que si **A** es no singular, **v** y **u** son dos vectores arbitrarios, entonces

$$(\mathbf{A} - \mathbf{u}\mathbf{v}^T)^{-1} = \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}}{\mathbf{1} - \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}}$$

por lo tanto, si se define $\mathbf{A} = \mathbf{X}^T\mathbf{X}$ y $\mathbf{u} = \mathbf{v} = \mathbf{X}_i$, se obtiene

$$(\mathbf{X}^T\mathbf{X} - \mathbf{X}_i\mathbf{X}_i^T)^{-1} = (\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1} = (\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_i\mathbf{X}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{(\mathbf{1} - \mathbf{p}_{ii})} \tag{3}$$

con $\mathbf{p}_{ii} = \mathbf{X}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_i$.

De (3), se obtiene que

$$\begin{aligned} \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} - (\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)} \\ &= \left((\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1} - \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_i\mathbf{X}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{(\mathbf{1} - \mathbf{p}_{ii})} \right) \mathbf{X}^T\mathbf{Y} - (\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}^T\mathbf{Y}_{(i)} \\ &= (\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}(\mathbf{X}^T\mathbf{Y} - \mathbf{X}_{(i)}^T\mathbf{Y}_{(i)}) - \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_i\mathbf{X}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}}{(\mathbf{1} - \mathbf{p}_{ii})} \\ &= (\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}\mathbf{X}_i\mathbf{Y}_i^T - \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_i\mathbf{X}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}}{(\mathbf{1} - \mathbf{p}_{ii})} \end{aligned} \tag{4}$$

Usando (3) en la primera parte de (4), se obtiene:

$$\begin{aligned} (\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}\mathbf{X}_i\mathbf{Y}_i^T &= \left((\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_i\mathbf{X}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{(\mathbf{1} - \mathbf{p}_{ii})} \right) \mathbf{X}_i\mathbf{Y}_i^T \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_i\mathbf{Y}_i^T + \frac{\mathbf{p}_{ii}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_i\mathbf{Y}_i^T}{(\mathbf{1} - \mathbf{p}_{ii})} \\ &= \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_i\mathbf{Y}_i^T}{(\mathbf{1} - \mathbf{p}_{ii})} \end{aligned} \tag{5}$$

Sustituyendo (5) en (4), resulta:

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \frac{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\mathbf{Y}_i^T - \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\mathbf{X}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}}{\left(\mathbf{1}-\mathbf{p}_{ii}\right)}$$

$$= \frac{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i}{\left(\mathbf{1}-\mathbf{p}_{ii}\right)}\left(\mathbf{Y}_i^T - \mathbf{X}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\right) \tag{6}$$

Ahora, puesto que $\hat{\boldsymbol{\varepsilon}} = \left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{I} - \mathbf{X}\mathbf{X}^-\right)\mathbf{Y} = \left(\mathbf{I}-\mathbf{P}\right)\mathbf{Y}$, donde $\mathbf{P}$ es el proyector ortogonal sobre la imagen de $\mathbf{X}$. Entonces $\mathbf{e}_i^{nT}\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\varepsilon}}_i^T = \mathbf{e}_i^{nT}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{Y}_i^T - \mathbf{X}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$, así, se obtiene:

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \frac{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\hat{\boldsymbol{\varepsilon}}_i^T}{\left(\mathbf{1}-\mathbf{p}_{ii}\right)} \tag{7}$$

Bajo el supuesto de la matriz de distribución normal, se propone la siguiente modificación a la distancia de Cook, denotada como D_m, es decir

$$D_m = vec\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)^T \hat{C}ov\left(vec\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)\right)^- vec\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right) \tag{8}$$

La expresión dada en (8), es una extensión del cuadrado de la distancia de Mahalanobis generalizada, tal como lo afirma Rao y Mitra, 1971.

El segundo paso, es encontrar una expresión simple para la matriz de varianzas y covarianzas $Cov\left(vec\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)\right)$. Para lo cual, se sabe que $\mathbf{Y} \sim N_{n\times p}\left(\boldsymbol{\mu}, \Sigma\otimes\Theta\right)$, entonces $E\left(\mathbf{Y}\right) = \boldsymbol{\mu}$, $Cov\left(vec\mathbf{Y}\right) = \left(\Sigma\otimes\Theta\right)$ (Muirhead, 1982).

Puesto que $\hat{\boldsymbol{\varepsilon}}_i^T = \mathbf{e}_i^{nT}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{e}_i^{nT}\left(\mathbf{I}-\mathbf{P}\right)\mathbf{Y}$, es claro que $vec\left(\hat{\boldsymbol{\varepsilon}}_i^T\right) = \left(\mathbf{I}_p \otimes \mathbf{e}_i^{nT}\left(\mathbf{I}-\mathbf{P}\right)\right)vec\mathbf{Y}$, por consiguiente

$$vec\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right) = \frac{\left(\mathbf{I}_p \otimes \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\right)}{\left(\mathbf{1}-\mathbf{p}_{ii}\right)}\left(\mathbf{I}_p \otimes \mathbf{e}_i^{nT}\left(\mathbf{I}-\mathbf{P}\right)\right)vec\mathbf{Y}$$

$$= \frac{\left(\mathbf{I}_p \otimes \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\mathbf{e}_i^{nT}\left(\mathbf{I}-\mathbf{P}\right)\right)}{\left(\mathbf{1}-\mathbf{p}_{ii}\right)}vec\mathbf{Y}$$

$$= \frac{\left(\mathbf{I}_p \otimes \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\mathbf{H}_i^T\right)}{\left(\mathbf{1}-\mathbf{p}_{ii}\right)}vec\mathbf{Y} \tag{9}$$

donde $\mathbf{H}_i^T = \mathbf{e}_i^{nT}(\mathbf{I}-\mathbf{P})$ es la i-ésima fila de la matriz $(\mathbf{I}-\mathbf{P})$. Entonces

$$Cov\left(vec\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}\right)\right) = \frac{\left(\mathbf{I}_p \otimes \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\mathbf{H}_i^T\right)}{\left(\mathbf{1}-\mathbf{p}_{ii}\right)} Cov(vec\mathbf{Y}) \frac{\left(\mathbf{I}_p \otimes \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\mathbf{H}_i^T\right)^T}{\left(\mathbf{1}-\mathbf{p}_{ii}\right)}$$

$$= \frac{\left(\mathbf{I}_p \otimes \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\mathbf{H}_i^T\right)}{\left(\mathbf{1}-\mathbf{p}_{ii}\right)^2}(\Sigma\otimes\mathbf{I})\left(\mathbf{I}_p \otimes \mathbf{H}_i\mathbf{X}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)$$

$$= \frac{\|\mathbf{H}_i\|^2\left(\Sigma\otimes\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\mathbf{X}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)}{\left(\mathbf{1}-\mathbf{p}_{ii}\right)^2} \qquad (10)$$

Nótese que

$$\begin{aligned}\|\mathbf{H}_i\|^2 &= \mathbf{e}_i^{nT}(\mathbf{I}-\mathbf{P})(\mathbf{I}-\mathbf{P})\mathbf{e}_i^n \\ &= \mathbf{e}_i^{nT}(\mathbf{I}-\mathbf{P})\mathbf{e}_i^n \\ &= \mathbf{e}_i^{nT}\mathbf{e}_i^n - \mathbf{e}_i^{nT}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}_i^n \\ &= \mathbf{1}-\mathbf{X}_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_i^T \\ &= \mathbf{1}-\mathbf{p}_{ii}\end{aligned} \qquad (11)$$

Sustituyendo (11) en (10), se obtiene

$$Cov\left(vec\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}\right)\right) = \frac{\left(\Sigma\otimes\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\mathbf{X}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)}{\left(\mathbf{1}-\mathbf{p}_{ii}\right)^2} \qquad (12)$$

Sea $S_1 = \frac{n\hat{\Sigma}}{n-q}$ y se sabe que $E(S_1)=\Sigma$, (Muirhead, 1982), entonces

$$\hat{Cov}\left(vec\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}\right)\right) = \frac{\left(S_1\otimes\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\mathbf{X}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)}{\left(\mathbf{1}-\mathbf{p}_{ii}\right)^2} \qquad (13)$$

Sea $r_i = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i$, y basándose en los siguientes resultados:

1. Para $a\in\Re^n$, $a^- = \frac{a^T}{\|a\|^2}$,

2. Dada $\mathbf{A}\in\Re^{p\times q}$, entonces $\left(\mathbf{A}\mathbf{A}^T\right)^- = \left(\mathbf{A}^T\right)^-\mathbf{A}^-$ con $\mathbf{A}^{-1}=\mathbf{A}^-$ si $\mathbf{A}$ es no singular,

3. Dadas las matrices $\mathbf{A}$ y $\mathbf{B}$, $(\mathbf{A}\otimes\mathbf{B})^- = \mathbf{A}^-\otimes\mathbf{B}^-$, se obtiene

$$\left(\hat{C}ov\left(vec\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}\right)\right)\right)^{-} = \hat{C}ov\left(vec\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}\right)\right)^{-}$$

$$= \left(\frac{\left(S_1 \otimes r_i r_i^T\right)}{\left(\mathbf{1}-\mathbf{p}_{ii}\right)}\right)^{-}$$

$$= \frac{\left(\mathbf{1}-\mathbf{p}_{ii}\right)}{\left\|r_i\right\|^4}\left(S_1^{-1} \otimes r_i r_i^T\right)$$

Por lo tanto, la distancia de Cook modificada puede ser re-escrita como:

$$D_m = vec\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}\right)^T \hat{C}ov\left(vec\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}\right)\right)^{-} vec\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}\right)$$

$$= \left(\frac{\left(\mathbf{I}_p \otimes \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\mathbf{H}_i^T\right)vec\mathbf{Y}}{\left(\mathbf{1}-\mathbf{p}_{ii}\right)}\right)^T \frac{\left(\mathbf{1}-\mathbf{p}_{ii}\right)\left(S_1^{-1} \otimes r_i r_i^T\right)}{\left\|r_i\right\|^4}\left(\frac{\left(\mathbf{I}_p \otimes \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\mathbf{H}_i^T\right)vec\mathbf{Y}}{\left(\mathbf{1}-\mathbf{p}_{ii}\right)}\right)$$

$$= \frac{\left(\mathbf{1}-\mathbf{p}_{ii}\right)^{-1}}{\left\|r_i\right\|^4} vec^T\mathbf{Y}\left(S_1^{-1} \otimes \mathbf{H}_i r_i^T r_i r_i^T r_i \mathbf{H}_i^T\right)vec\mathbf{Y}$$

$$= \left(\mathbf{1}-\mathbf{p}_{ii}\right)^{-1} vec^T\mathbf{Y}\left(S_1^{-1} \otimes \mathbf{H}_i\mathbf{H}_i^T\right)vec\mathbf{Y} \qquad (14)$$

Alternativamente, ya que $tr\left(\mathbf{B}\mathbf{X}^T\mathbf{C}\mathbf{X}\mathbf{D}\right) = vec^T\mathbf{X}\left(\mathbf{B}^T\mathbf{D}^T \otimes \mathbf{C}\right)vec\mathbf{X} = vec^T\mathbf{X}\left(\mathbf{D}\mathbf{B} \otimes \mathbf{C}^T\right)vec\mathbf{X}$, para matrices de órdenes adecuados, se puede escribir D_m como $D_m = \left(\mathbf{1}-\mathbf{p}_{ii}\right)^{-1} tr\left(S_1^{-1}\mathbf{Y}^T\mathbf{H}_i\mathbf{H}_i^T\mathbf{Y}\right)$.

Por otro lado, puesto que $\hat{\boldsymbol{\varepsilon}}_i^T = \mathbf{e}_i^{nT}\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right) = \mathbf{H}_i\mathbf{Y}$, entonces

$$D_m = \left(\mathbf{1}-\mathbf{p}_{ii}\right)^{-1} tr\left(S_1^{-1}\hat{\boldsymbol{\varepsilon}}_i\hat{\boldsymbol{\varepsilon}}_i^T\right)$$

$$= \left(\mathbf{1}-\mathbf{p}_{ii}\right)^{-1} tr\left(\hat{\boldsymbol{\varepsilon}}_i^T S_1^{-1}\hat{\boldsymbol{\varepsilon}}_i\right)$$

$$= \left(\mathbf{1}-\mathbf{p}_{ii}\right)^{-1}\hat{\boldsymbol{\varepsilon}}_i^T S_1^{-1}\hat{\boldsymbol{\varepsilon}}_i$$

Por lo que se tienen las siguientes expresiones como alternativa para el cuadrado de la distancia de Cook modificada:

$$D_m = \begin{cases} vec\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}\right)^T \hat{C}ov\left(vec\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}\right)\right)^{-} vec\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}\right) \\ \left(\mathbf{1}-\mathbf{p}_{ii}\right)^{-1} vec^T\mathbf{Y}\left(S_1^{-1} \otimes \mathbf{H}_i\mathbf{H}_i^T\right)vec\mathbf{Y} \\ \left(\mathbf{1}-\mathbf{p}_{ii}\right)^{-1} tr\left(S_1^{-1}\mathbf{Y}^T\mathbf{H}_i\mathbf{P}_i^T\mathbf{Y}\right) \\ \left(\mathbf{1}-\mathbf{p}_{ii}\right)^{-1}\hat{\boldsymbol{\varepsilon}}_i^T S_1^{-1}\hat{\boldsymbol{\varepsilon}}_i \end{cases} \qquad (15)$$

De acuerdo con Chatterjee y Hadi, 1988, se puede reemplazar la matriz S_1 por otra obtenida usando la muestra reducida $(n-1)$, denotada por $S_1(i)$.

Cook, 1977, Chatterjee y Hadi, 1988, Díaz-García *et al.*, 2001, y algunos otros autores, utilizan la matriz de varianzas y covarianzas del $vec\left(\hat{\boldsymbol{\beta}}\right)$ para construir las medidas de distancia. La reformulación que se propone, está basada en el reemplazo de esta matriz, por la matriz de varianzas y covarianzas del $vec\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}\right)$. Se puede encontrar esta idea en Chatterjee y Hadi, 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 $vec\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}\right)$.

Teorema 1. *Considerando el modelo de regresión normal multivariado dado en* (1)*, entonces, el cuadrado de la distancia de Cook modificada para detectar un "outlier", puede ser escrita como:*

$$D_m = \begin{cases} vec\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}\right)^T \left(\dfrac{\left(S_1 \otimes \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_i\mathbf{X}_i^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)}{\left(\mathbf{1}-\mathbf{p}_{ii}\right)}\right)^{-} vec\left(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)}\right) \\ \left(\mathbf{1}-\mathbf{p}_{ii}\right)^{-1} vec^T\mathbf{Y}\left(S_1^{-1}\otimes\mathbf{H}_i\mathbf{H}_i^T\right)vec\mathbf{Y} \\ \left(\mathbf{1}-\mathbf{p}_{ii}\right)^{-1} tr\left(S_1^{-1}\mathbf{Y}^T\mathbf{H}_i\mathbf{P}_i^T\mathbf{Y}\right) \\ \left(\mathbf{1}-\mathbf{p}_{ii}\right)^{-1}\hat{\boldsymbol{\varepsilon}}_i^T S_1^{-1}\hat{\boldsymbol{\varepsilon}}_i \end{cases} \tag{16}$$

En (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 D_m, en el caso normal univariado, coincide con el análisis de residuales estudentizados, tal como se observa en Besley *et al.*, 1980 y Chatterjee y Hadi, 1988.

DISTANCIA MODIFICADA: MÚLTIPLES OBSERVACIONES

Sea $\mathbf{I}=\{i_1,i_2,...,i_k\}$ un subconjunto de tamaño k de $\{1,2,...,n\}$, de forma tal que $(n-k)\geq q$. Ahora, bajo el modelo (1), se denotan por $\mathbf{Y}_{(I)}$, $\mathbf{X}_{(I)}$ y $\hat{\boldsymbol{\varepsilon}}_{(I)}$, las matrices de datos, de regresión y de errores, respectivamente, después de eliminar las correspondientes observaciones de acuerdo con los subíndices en I. Sean $\hat{\boldsymbol{\beta}}_{(I)}$ y $\hat{\Sigma}_{(I)}$, los correspondientes estimadores de máxima verosimilitud en el modelo:

$$\mathbf{Y}_{(I)} = \mathbf{X}_{(I)}\boldsymbol{\beta}_{(I)} + \boldsymbol{\varepsilon}_{(I)},\ \boldsymbol{\varepsilon}_{(I)} \sim N_{(n-k)\times p}\left(\mathbf{0}, \Sigma_{(I)}\otimes\mathbf{I}_n\right).$$

Lema 1. 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

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

Para la demostración de este lema, sólo debe verse que

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

Basándose en el Lema 1 y usando procedimientos similares a los de la distancia modificada para una sola observación, es fácil verificar que

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(I)} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_I\left(\mathbf{I}-\mathbf{P}_I\right)^{-1}\hat{\boldsymbol{\varepsilon}}_i,$$

con $\left(\mathbf{I}-\mathbf{P}_I\right)=\left(\mathbf{I}_k - \mathbf{X}_I^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_I\right)$ y $\mathbf{X}_I$ es la matriz con las correspondientes filas de $\mathbf{X}$ de acuerdo con I. Observando que $\hat{\boldsymbol{\varepsilon}}_{(I)} = \mathbf{U}_I^T\hat{\boldsymbol{\varepsilon}} = \mathbf{U}_I^T(\mathbf{I}-\mathbf{P})\mathbf{Y}$, donde

$$\mathbf{U}_I^T = \begin{pmatrix} e_{i\,1}^{n\,T} \\ e_{i\,2}^{n\,T} \\ \cdot \\ \cdot \\ \cdot \\ e_{i\,k}^{n\,T} \end{pmatrix}$$

Se obtiene $vec\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(I)}\right)=\left(\mathbf{I}_p \otimes \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_I\left(\mathbf{I}-\mathbf{P}_I\right)^{-1}\mathbf{H}_I\right)vec\mathbf{Y}$, con $\mathbf{H}_I = \mathbf{U}_I^T(\mathbf{I}-\mathbf{P})$.

De donde $Cov\left(vec\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(I)}\right)\right)=\left(\sum \otimes\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_I\left(\mathbf{I}-\mathbf{P}_I\right)^{-1}\mathbf{X}_I^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)$ y

$\hat{C}ov\left(vec\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(I)}\right)\right)=\left(S_1 \otimes \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_I\left(\mathbf{I}-\mathbf{P}_I\right)^{-1}\mathbf{X}_I^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right).$

Entonces se tiene,

Teorema 2. *Considerando el modelo de regresión normal multivariado dado en* (1)*, entonces, el cuadrado de la distancia de Cook modificada para detectar* k *observaciones influyentes, puede ser escrita como:*

$$D_{mI} = \begin{cases} vec\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(I)}\right)^T\left(S_1 \otimes \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}_I\left(\mathbf{I}-\mathbf{P}_I\right)^{-1}\mathbf{X}_I^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)^{-} vec\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(I)}\right) \\ vec^T\mathbf{Y}\left(S_1^{-1} \otimes \mathbf{H}_I^T\left(\mathbf{I}-\mathbf{P}_I\right)^{-1}\mathbf{H}_I\right)vec\mathbf{Y} \\ tr\left(S_1^{-1}\mathbf{Y}^T\mathbf{H}_I^T\left(\mathbf{I}-\mathbf{P}_I\right)^{-1}\mathbf{H}_I\mathbf{Y}\right) \\ tr\left(S_1^{-1}\hat{\boldsymbol{\varepsilon}}_I^T\left(\mathbf{I}-\mathbf{P}_I\right)^{-1}\hat{\boldsymbol{\varepsilon}}_I\right) \end{cases} \tag{17}$$

RESULTADOS Y DISCUSIÓN

FUNCIONES DE DISTRIBUCIÓN ASOCIADAS CON LAS DISTANCIAS MODIFICADAS

La razón principal para estudiar las modificaciones al cuadrado de la distancia de Cook para una y múltiples observaciones, es que, en vez de utilizar una aproximación a la distribución **F**, se derivará la distribución exacta para D_m. Análogamente se encuentra la distribución exacta para D_{mI}, 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 de múltiples observaciones.

Teorema 3. *Bajo los supuestos del Teorema 1, se tiene que*

$$\frac{D_m}{n-q} \sim \beta_{(p/2,(n-q-p)/2)} \tag{18}$$

donde $\beta_{(p/2,(n-q-p)/2)}$ *denota una distribución beta centrada, con parámetros* $p/2$ *y* $(n-q-p)/2$.

La demostración de este teorema se sigue directamente de Caroni, 1987.

Del Teorema 3, dando un nivel de significancia α, se puede escribir la siguiente regla de decisión: Y_i, $i=1,2,...,n$, es un "outlier", si

$$\frac{D_m}{n-q} \geq \beta_{\alpha:(p/2,(n-q-p)/2)} \tag{19}$$

donde $\beta_{\alpha:(p/2,(n-q-p)/2)}$ es el correspondiente α-*percentil* superior de una distribución beta de parámetros $p/2$ y $(n-q-p)/2$.

Similarmente, para el caso de múltiples observaciones:

Teorema 4. *Bajo los supuestos del Teorema 2, se tiene que*

$$\frac{D_{mI}}{n-q} \sim \rho_{s,m,h} \tag{20}$$

donde $\rho_{s,m,h}$ *denota la distribución centrada para el estadístico de Pillai con parámetros*

$s=\min(p,k)$, $m=(|p-k|-1)/2$ *y* $h=(n-q-p+1)/2$.

La demostración de este teorema es una consecuencia directa del Teorema 10.6.2, Corolario 10.6.3 en Muirhead, 1982.

UNA APLICACIÓN

A continuación, en la tabla 1, se presentan los datos correspondientes a un ejemplo tomado de Srivastava y Carter, 1983. En el cual, 25 depósitos de truchas se sometieron a diferentes dosis de cobre en miligramos por litro y se registró el peso promedio de los peces, es decir, las proporciones $\mathbf{p}_{ij}$, $(i=1,2,...,25;\ j=1,2,3,4,5)$ de peces muertos después de 8, 14, 24, 36 y 48 horas.

Los datos son ajustados por el modelo multivariado de regresión lineal múltiple. Dado el carácter de las variables dependientes, se hace la transformación *arcsen* de las variables respuesta para estabilizar la varianza y se considera el logaritmo de la variable dependiente $\mathbf{X}_1$.

Dadas las asunciones anteriores, tres diferentes métricas fueron aplicadas a los datos para ver los candidatos a ser considerados "outliers", cuyos resultados se presentan en seguida, en la figura 1, donde se muestran las tres gráficas correspondientes.

Tabla 1. Proporción de peces muertos después de 8, 14, 24, 36 y 48 horas de que se les suministraron diferentes dosis de cobre.

Y_1 (8 horas)	Y_2 (14 horas)	Y_3 (24 horas)	Y_4 (36 horas)	Y_5 (48 horas)	X_1 *log* (dosis)	X_2 Pesos Promedio
0,00	0,00	30,00	30,00	30,00	5,60	0,67
0,00	18,44	33,21	33,21	33,21	6,02	0,64
0,00	45,00	60,00	71,57	71,57	6,41	0,73
22,79	53,73	90,00	90,00	90,00	6,85	0,77
42,13	90,00	90,00	90,00	90,00	7,28	0,57
0,00	12,92	26,57	26,57	26,57	5,60	0,78
12,92	18,44	33,21	33,21	33,21	6,02	0,81
12,92	42,13	77,08	90,00	90,00	6,41	0,82
18,44	56,79	90,00	90,00	90,00	6,85	0,87
26,57	67,21	90,00	90,00	90,00	7,28	0,84
0,00	0,00	0,00	0,00	12,92	5,60	0,86
0,00	12,92	22,79	30,00	33,21	6,02	0,91
0,00	22,79	77,08	77,08	77,08	6,41	1,03
0,00	46,79	77,08	90,00	90,00	6,85	1,05
18,44	67,21	90,00	90,00	90,00	7,28	1,05
0,00	0,00	0,00	12,92	18,44	5,60	0,62
0,00	12,92	22,79	26,57	30,00	6,02	0,53
18,44	42,13	77,08	77,08	77,08	6,41	0,60
18,44	56,79	90,00	90,00	90,00	6,85	0,64
36,27	77,08	90,00	90,00	90,00	7,28	0,67
0,00	12,92	26,57	26,57	26,57	5,60	0,57
0,00	0,00	22,79	30,00	30,00	6,02	0,60
0,00	39,23	71,57	90,00	90,00	6,41	0,63
12,92	53,73	90,00	90,00	90,00	6,85	0,69
33,21	67,21	90,00	90,00	90,00	7,28	0,72

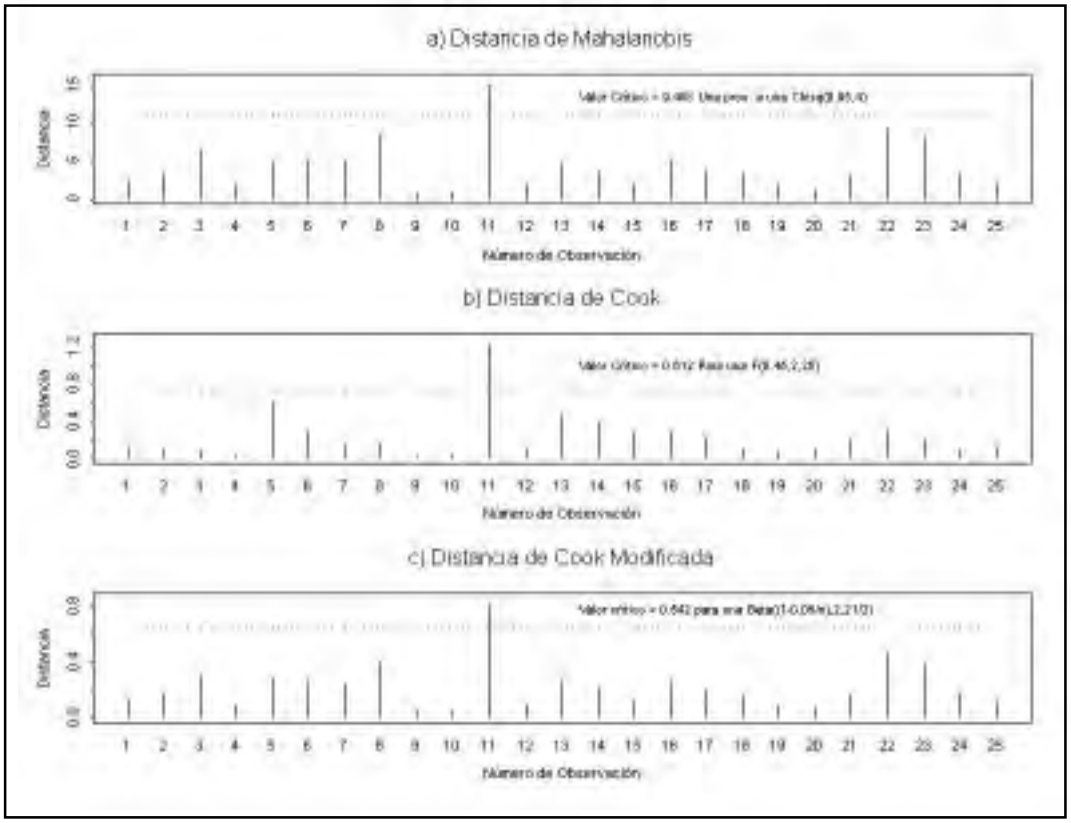

Figura 1. Aplicación de tres diferentes métricas a las observaciones para identificar "outliers".
(**a**) Distancia al cuadrado de Mahalanobis.
(**b**) Distancia de Cook clásica.
(**c**) Distancia de Cook modificada.

En la figura 1 (a) aparece la gráfica de la distancia al cuadrado de Mahalanobis tal como es presentada por Seber (1984). En la figura 1 (b) se muestra la gráfica obtenida al aplicar la distancia de Cook clásica, como lo señalan Chatterjee y Hadi (1988). Finalmente, la gráfica de los resultados aplicando la distancia de Cook modificada propuesta en este trabajo, se observa en la figura 1 (c).

En la figura 1 (a), (b) y (c), se observa que con la aplicación de las tres distancias, la observación número 11 se identifica como un "outlier". Es obvio que sólo hay evidencia de que el punto número 11 es una observación atípica, pero a manera de ejemplo se considera como posible punto atípico a la siguiente observación con la distancia de Cook modificada mayor, que en este caso corresponde al punto número 22. Así, considérense simultáneamente como posibles observaciones atípicas a los datos número 11 y número 22. Luego, aplicando el estadístico de Pillai para identificar si estas observaciones, son o no, puntos influyentes en forma simultánea, se obtiene que $D_{mI} = 1{,}25871$. Si se aplica ahora la aproximación del F-estadístico propuesto en Rencher (1995, p.185), se obtiene un valor de 6,452547, comparándolo con el valor crítico correspondiente de 2,09085, entonces se puede concluir que las observaciones número 11 y número 22 son en conjunto influyentes. En este caso, se llega a tal conclusión, dada la fuerte influencia de la observación número 11, sin embargo en una situación real, hay que tener evidencia suficiente (o a través de métodos gráficos) de la posible influencia conjunta de dos o más observaciones sobre los parámetros del modelo.

REFERENCIAS

Besley, D.A., Kuh, E. y Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* New York. John Wiley & Sons.

Caroni, C. (1987). Residuals and Influence in the Multivariate Linear Model. *The Statistician.* 36: 365-370.

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

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

Cook, R.D. y Weisberg, S. (1982). *Residual and Influence in Regression.* London. Chapman and Hall.

Díaz-García, J.A., Galea, M. y Leiva Sánchez, V. (2001). Influence Diagnostics for Elliptical Regression Linear Models. *Communication in Statistics.* 32: 625-641.

Galea, M., Paula, G. y 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.

Martínez Jaime, O.A. (2001). *Análisis de Sensibilidad en Regresión.* Tesis de Maestría en Ciencias en Estadística Experimental. Universidad Autónoma Agraria "Antonio Narro". Saltillo, Coahuila, México.

Muirhead, R.J. (1982). *Aspects of Multivariated Statistical Theory.* New York. John Wiley & Sons.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications* 2nd. ed. New York. John Wiley & Sons.

Rao, C.R. y Mitra, S.K. (1971). *Generalized Inverse of Matrices and its Applications* 2nd. ed. New York. John Wiley & Sons.

Rencher, A. C. (1995). *Methods of Multivariate Analysis.* New York. John Wiley & Sons.

Seber, G. A. F. (1984). *Multivariate Observations.* New York. John Wiley & Sons.

Srivastava, M.S. y Carter. (1983). *An Introduction to Applied Multivariate Statistics*. New York. North-Holland Publ.