PESQUIMAT, Revista de la Fac.CC.MM. de la
UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Vol.IV, $N°2$, pág.17-33, LIMA-PERÚ. Diciembre 2001

# IDENTIFICACIÓN DE MÚLTIPLES OBSERVACIONES DISCREPANTES EN EL ANÁLISIS DE REGRESIÓN

**Ysela Agüero Palacios**[1]

**ABSTRACT.** *Se presentan métodos de diagnóstico de datos discrepantes e influyentes en el análisis de regresión. Estos métodos se basan en el ajuste robusto* LMS *y el elipsoide de volumen mínimo* (MVE), *los cuales no son afectados por el problema de enmascaramiento cuando el conjunto de datos contiene más de una observación discrepante y/o influyente.*

## 1. INTRODUCCIÓN

La confiabilidad de un conjunto de datos obtenidos sobre condiciones similares está basada en la relación entre ellas. Observaciones que en opinión del investigador se encuentran fuera de la masa de los datos son llamados outliers, observaciones discordantes, datos aberrantes, datos contaminantes, observaciones sorprendentes, datos discrepantes, datos groseros, etc. No existe uniformidad en relación al significado exacto de estos términos, a pesar de la gran cantidad de artículos escritos y de que los estudios con relación a este tipo de observaciones se iniciaron hace casi 200 años.

La necesidad de identificar las observaciones discrepantes, en el análisis de datos se debe a que las mismas, pueden distorsionar la información que la masa de datos debe proporcionar sobre el fenómeno en estudio. Por otro lado, las observaciones discrepantes pueden también ser importantes mensajes de que las suposiciones formuladas en relación al mecanismo generador del conjunto de datos no son correctas, consecuentemente, es preciso cambiar la concepción que se tenía obre el fenómeno en estudio.

Supóngase que todos los factores o causas que influyen en una variable respuesta ($Y$), pueden dividirse en dos grupos; el primero contiene $K$ variables, $X_1, X_2, \ldots, X_k$, cuyos valores son conocidos al observar la respuesta y que están relacionadas linealmente con la variable respuesta, $Y$. El segundo grupo incluye un conjunto muy grande de factores que tienen una escasa influencia en $Y$, agrupándose bajo el nombre

[1]Univ. Nac. Mayor de San Marcos. Fac. de Ciencias Matemáticas - Lab.de Series de Tiempo.

de perturbación aleatoria ó error aleatorio, $e$. Supóngase que el modelo que relaciona estas variables es dado por,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K + \varepsilon. \tag{1.1}$$

Las suposiciones asociadas con el modelo (1.1) son:

(i) El error aleatorio tiene esperanza igual a cero,
(ii) La varianza del error es constante y no depende de las variables regresoras (condición de homocedasticidad),
(iii) Los errores son incorrelacionados.

Por la estructura particular de los datos en el análisis de regresión, la discrepancia de una observación puede ser atribuida a que es atípica en la dirección de la variable respuesta, de las variables regresoras ó en ambas direcciones. Las variables regresoras $X_1, \ldots, X_k$ son cantidades observadas sujetas a variabilidad, por lo tanto, pueden ocurrir errores imprevisibles aún tratándose de experimentos planeados. Estas observaciones son conocidas como discrepantes en la dirección de la matriz del modelo ("leverage"), pudiendo tener una fuerte influencia en los parámetros estimados. Esta influencia puede ser buena o perjudicial (mejora el ajuste, o lo desvía) dependiendo del valor de la variable respuesta con la cual se combinan. Es importante resaltar que los "leverage" perjudiciales son discrepantes en la direccin de $X$, pero, las observaciones discrepantes en la dirección de $X$ no necesariamente son puntos "leverage" perjudiciales.

Existen varios métodos de estimación de los parámetros del modelo (1.1), la mayoría de ellos tratan de minimizar una cierta función de los errores. Boscovich en 1757 propuso un estimador que minimiza la suma de los valores absolutos de los errores, este método fue poco difundido por las dificultades computacionales involucradas en su cálculo aún para modelos relativamente simples. Legendre y posteriormente Gauss en 1805 -1809, propusieron un estimador obtenido minimizando la suma de cuadrados de los errores, esto es, *Minimizar* $\sum_{i=1}^{n} \varepsilon_i^2$. Los estimadores así obtenidos son muy populares por la facilidad de cálculo, además, cuando los errores son variables aleatorias independientes e idénticamente distribuidas, con distribución normal, el estimador de Gauss y Legendre denominado "Estimador de mínimos cuadrados" es insesgado, tiene varianza mínima y es equivalente al estimador de máxima verosimilitud.

En la práctica, la suposición de normalidad no se cumple en la mayoría de los datos, por ejemplo, puede existir en el conjunto de datos una cierta proporción de observaciones que no provienen de la distribución normal sino de otra con "colas más

pesadas" como Laplace, Cauchy, etc., ó que existan otros problemas como multicolinealidad, o dependencia de los errores. El no cumplimiento de estas suposiciones se manifiesta en muchos casos a través de la discrepancia de una ó más observaciones con respecto a la masa de datos.

Ha sido ampliamente comprobado que estas observaciones discrepantes pueden distorsionar los resultados del análisis y consecuentemente las conclusiones. Por lo tanto, es necesario emplear métodos de diagnóstico que permitan verificar si las suposiciones establecidas en la formulación del modelo se están verificando en los datos.

En los últimos años se desarrollaron métodos de diagnóstico de observaciones discrepantes e influyentes basados en los residuos del ajuste y en las características de la matriz del modelo. Belsley, Kuh y Welsch (1980), Cook y Weisberg (1982) y otros, propusieron diferentes tipos de diagnósticos basados en el método de mínimos cuadrados. La mayoría de ellos ya forman parte de los programas computacionales que incluyen el análisis de regresión. El problema es que estos métodos sólo identifican observaciones discrepantes individuales y en muchos casos una observación fuertemente discrepante puede enmascarar otras que también están influenciando el ajuste. Este problema es conocido como "efecto de enmascaramiento".

En este artículo se presentan métodos de diagnóstico de múltiples observaciones discrepantes que son robustos en el sentido de no ser afectados en presencia de una ó más observaciones discrepantes en el conjunto de datos.

En la sección 2, se comentan, las características de las observaciones discrepantes y se presentan brevemente los métodos de identificación y evaluación de la influencia de las observaciones discrepantes, basados en el ajuste mínimo cuadrático. En esta misma sección se mencionan algunos conceptos fundamentales de robustez y se estudian los estimadores LMS ("Least Median of Squares") y elipsoide de volumen mínimo ó MVE, los cuales serán utilizados para la identificación de observaciones discrepantes e influyentes, respectivamente. Finalmente, en la sección 3 se presenta un ejemplo de aplicación para ilustrar los métodos de diagnóstico estudiados.

## 2. DIAGNÓSTICOS DE OBSERVACIONES DISCREPANTES E INFLUYENTES.

Uno de los artículos más completos referentes al tema de los datos discrepantes fue presentado por Beckman y Cook (1983), ellos resumen en tres las diferentes denominaciones usadas en los artículos; y dan las definiciones siguientes. Una *observación*

*discrepante (o discordante)* es cualquier observación que parece sorprendente para el investigador, lo cual implica que él supuso un modelo teórico o por lo menos tiene una visión informal de este. Una *observación contaminante* es cualquiera que no corresponde a una realización de la distribución supuesta - La discrepancia de estas observaciones puede no ser tan obvia -. En la práctica, se utiliza la denominación de *observación discrepante* ("outlier"), para referirse a una observación contaminante o discordante.

Nótese que, la definición de observación discrepante es un tanto ambigua, puesto que, lo que realmente caracteriza a la observación discrepante es *su impacto en el observador*, mientras que, una observación puede ser contaminante y no ser percibida.

Se puede entonces concluir que:

(i) Una observación contaminante no necesariamente es discrepante, ella lo será si; además de contaminante fuera también discordante.
(ii) Una observación extrema será discrepante si aparece como discordante bajo el modelo supuesto.

Las causas de la presencia de observaciones discrepantes en un conjunto de datos pueden ser agrupadas en tres categorías. Estas son: debilidad del modelo supuesto, problemas en la obtención de datos y variabilidad natural de los mismos. Se sobrentiende por las dos primeras categorías que las observaciones discrepantes son juzgadas teniendo en mente un modelo determinado.

Las debilidades en el modelo supuesto incluyen causas tales como una variable respuesta en la escala errada, ó el modelo teóricamente supuesto no es adecuado. La primera causa puede llevar a una transformación en la respuesta, entretanto que, la última puede conducir a sustituir el modelo actual. Los problemas en la obtención de los datos; se refieren solamente a las observaciones y no al modelo como un todo. Estas causas pueden indicar que las observaciones discrepantes sean tratadas individualmente. Ejemplos de este tipo de datos son; los errores de medición, de observación o de registro. Finalmente, la variabilidad natural de los datos se refiere a observaciones que pueden aparecer como discrepantes siendo observaciones genuinas de la distribución considerada.

## 2.1 Diagnósticos basados en el método de mínimos cuadrados.

Bajo la suposición de que los errores, del modelo (1.1) son variables aleatorias independientes con esperanza cero y varianza constante, el teorema de Gauss-Markov garantiza que el estimador de mínimos cuadrados

$$\beta = (X^tX)^{-1}X^tY; \tag{2.1}$$

es el único estimador insesgado con varianza mínima en la clase de los estimadores lineales. Además, si los errores del modelo siguen una distribución normal, el vector definido en (2.1) es el estimador de máxima verosimilitud.

A partir de (2.1) se puede expresar el vector de respuestas estimadas como $\hat{y} = X(X^tX)^{-1}X^ty = Hy$ , donde la matriz $H = X(X^tX)^{-1}X^t$ es el operador de proyección ortogonal del vector $y$ sobre el espacio columna generado por $X$. Esta matriz es simétrica e idempotente y de rango $k$ (número de parámetros). Además, el i-ésimo elemento de la diagonal de $H$, denotado por $h_{ii}$ es la medida de la distancia del vector fila $x_i \in X$ al centroide de los datos. Estos valores son de gran importancia tanto en la detección de observaciones discrepantes, como en los diagnósticos de influencia.

Una desventaja al utilizar $h_{ii}$; $i = 1, 2, \ldots, n$, como medida de influencia es que estos sufren el efecto de enmascaramiento, es decir, que si existen múltiples observaciones discrepantes pueden no ser detectadas por este método de diagnóstico (Rousseeuwy Leroy (1987)).

Otro método común de detección de observaciones discrepantes consiste en examinar los residuos estandarizados del ajuste de mínimos cuadrados, denominados también residuos estudentizados internamente ($r_i$) y residuos estudentizados externamente ($t_i$). Este último tipo de residuo es obtenido excluyendo la i-ésima observación en el cálculo de la varianza estimada de los residuos (Belsley, Kuh y Welsch (1980)).

Considerando que los residuos pueden no ser notoriamente grandes, por el hecho de que las observaciones discrepantes dislocan el hiperplano en su dirección, y que $h_{ii}$ no siempre detecta múltiples observaciones discrepantes en la dirección de las variables regresoras, ninguno de los dos tipos de residuos estudentizados son confiables en presencia de múltiples observaciones discrepantes.

Para evaluar la influencia potencial de una observación en determinada fase del análisis de datos se ajusta el modelo con y sin la i-ésima observación y se evalúa la magnitud del cambio en las estimaciones. Las estimaciones de parámetros y respuestas serán, $\beta_{(i)} = (X^t_{(i)}X_{(i)})^{-1}X^t_{(i)}y_{(i)}$ y $\hat{y}_{(i)} = (X^t_{(i)}X_{(i)})^{-1}X^t_{(i)}y_{(i)}$. Donde $X_{(i)}, y_{(i)}$ son la matriz del modelo y el vector de respuestas, respectivamente, sin considerar la i-ésima observación.

Los indicadores de influencia más comunes - que se encuentran incorporados en los programas computacionales - son las estadísticas que se presentan en la tabla siguiente:

| Estadística | Mide el efecto de una observación sobre |
|---|---|
| DFFITS | Una respuestas estimada |
| DFBETAS | Cada uno de los coeficientes de regresión estimados |
| COOK | El vector de coeficientes estimados |

Nuevamente, estos métodos de diagnóstico son útiles para evaluar la influencia de datos individuales pero, no son muy eficientes para identificar outliers múltiples, por lo que se hace necesario buscar otros métodos robustos en presencia de múltiples observaciones discrepantes y/o influyentes.

## 2.2 Diagnósticos basados en Métodos Robustos

El término robustez introducido por George Box (1953), es usado para denominar los procedimientos estadísticos que son resistentes frente a las desviaciones de las suposiciones hechas en la formulación del modelo matemático.

Es importante mencionar que existen diferentes formas de evaluar la robustez de un estimador, Así por ejemplo, el investigador puede estar interesado en evaluar si un estimador es robusto en relación a la presencia de datos discrepantes, ó si es robusto frente a pequeñas desviaciones del modelo supuesto, o tal vez desee averiguar si un estimador es más robusto que otro. Hampel (1971) presenta los conceptos de robustez cualitativa, Función de Influencia y Punto de Ruptura. Cada uno de estos conceptos permite construir estimadores robustos y evaluar la robustez de otros estimadores.

En este artículo el interés está centrado en el punto de ruptura, el cual es una medida global de robustez que da una idea de la tolerancia del estimador a observaciones discrepantes. Donoho y Huber (1983) presentan una definición de punto de ruptura para muestras finitas basada en la definición de sesgo; la cual se define a continuación.

**Definición 2.2.1.-** Sea una muestra de $m$ observaciones $Z = \{z_1, \ldots, z_n\}$ y $Z^*$ un conjunto construido substituyendo $m$ observaciones de $Z$ por valores arbitrarios. Se define $b(m; t, Z^*)$, el sesgo máximo causado por la contaminación del conjunto de datos, como,

$$b(m; t, Z^*) = \sup_{Z^*} \|t(Z^*) - t(Z)\|, \tag{2.2}$$

donde el supremo es obtenido sobre todas las posibles muestras contaminadas. Si $b(m; t, Z^*)$ es infinito, entonces las $m$ observaciones discrepantes tienen un efecto arbitrariamente grande en el estimador.

**Definición 2.2.2.-** Sea $t = t_n$ un estimador aplicado en la muestra $Z$. El punto de ruptura, $\varepsilon_n^*(t, Z)$, de ese estimador es dado por:

$$\varepsilon_n^*(t, Z) = \min_{1 \leq m \leq n} \left\{ \frac{m}{n};\ b(m; t; Z^*) \text{ es infinito} \right\} \tag{2.3}$$

De la expresión (2.3) se tiene que, el punto de ruptura del estimador es igual a la proporción mínima de observaciones discrepantes contenidas en la muestra, que hace que el sesgo del estimador tienda a infinito. En consecuencia, interesarán los estimadores con punto de ruptura próximo de 1/2.

En el caso del estimador de mínimos cuadrados el punto de ruptura del estimador es $\varepsilon_n^*(t, Z) = 1/n$ y tenderá a cero cuando $n \to \infty$, es decir, que una sola observación discrepante puede afectar fuertemente la estimación.

Otro concepto que será de utilidad en la sección siguiente es la definición de punto de ruptura del estimador de la matriz de dispersión de un conjunto de datos multivariado, $X$.

**Definición 2.2.3.-** El punto de ruptura del estimador de la matriz de dispersión, $C_n$ es dado por,

$$\varepsilon^*(C_n, X) = \min_{1 \leq m \leq n} \left\{ \frac{m}{n};\ \sup_{X^*} D(C_n(X), C_n(X^*)) \text{ es infinito} \right\} \tag{2.4}$$

Donde el supremo es calculado sobre todos los conjuntos $X^*$ resultantes de la substitución de $m$ observaciones del conjunto de datos $X = (x_1, \ldots, x_n)$, $x_i \in \Re^k$, $k \geq 1$, por valores arbitrarios. El punto de ruptura de $C_n$ es dado por la menor fracción de datos discrepantes; que pueden hacer que el mayor autovalor, $\lambda_1(C_n)$, sea muy grande y/o que el menor autovalor, $\lambda_1(C_n)$, asuma valores próximos de cero.

La distancias entre las matrices utilizadas en la construcción de la expresión de la derecha de (2.4) es definida por:

$$D(C_n(X), C_n(X^*)) = Max \left\{ \lambda_1(C_n(X)) - \lambda_1(C_n(X^*)), \lambda_k^{-1}(C_n(X)) - \lambda_k^{-1}(C_n(X^*)) \right\}$$

Con $\lambda_1(C_n(\cdot)) \geq \lambda_2(C_n(\cdot)) \geq \ldots \geq \lambda_k(C_n(\cdot))$ autovalores de la matriz $C_n(\cdot)$.

### 2.2.1 Estimadores de regresión con Punto de Ruptura Máximo

A continuación se presentan dos estimadores con alto punto de ruptura, estos son el estimador de regresión LMS ("least Median of Squares") y el estimador Elipsoide de Volumen Mínimo (MVE) los cuales serán utilizados en lugar del estimador de regresión de mínimos cuadrados y los elementos de la diagonal de la matriz $H$, respectivamente.

### a) Estimador LMS

Rousseew (1984) propuso el estimador LMS, el cual es obtenido minimizando la mediana de los cuadrados de los errores.

$$\tilde{\beta} = \min_{\beta} mediana\{\varepsilon_i^2;\ 1 \leq i \leq n\} \qquad \beta \in \Re^k \tag{2.5}$$

Donde la mediana es definida como la estadística de orden $[n/2]+1$ de los errores $\varepsilon_i = (y_i - x_i\tilde{\beta});\ i = 1, \ldots, n$. Para incrementar el punto de ruptura del estimador (2.5) se modifica la función objetivo definiendo,

$$\tilde{\beta} = \min_{\beta} \max_{1 \leq i \leq n} (\varepsilon_{i:n}^2); \qquad \beta \in \Re^k \tag{2.6}$$

Observar que el estimador LMS en (2.6) es un miembro de la familia de estimadores con alto punto de ruptura construidos aplicando la norma del máximo sobre el conjunto de errores "pequeños" (Agüero (1994)). El criterio de división de los errores en "grandes" y "pequeños" es dado por la q-ésima estadística de orden de los cuadrados de los errores del ajuste ($q = [n/2] + [(k+1)/2]$)

La formulación (2.6) como un problema de estimación norma del máximo ó de Chebyshev (Stromberg (1991)) es de utilidad por que facilita la búsqueda de los algoritmos exactos para resolver el problema de optimización (Cheney (1966)).

Geométricamente la solución LMS corresponde a encontrar la banda más estrecha que cubre por lo menos la mitad de las observaciones. La amplitud de la banda es medida en la dirección vertical y se espera que por lo menos $q = [n/2]+[(k+1)/2]+1$ puntos se encuentren contenidos en ella. Este estimador es robusto en el sentido de resistencia a observaciones discrepantes, puesto que, por lo menos 50% de las observaciones están entre los dos hiperplanos $X\tilde{\beta} \pm \rho$, donde $\rho$ es la amplitud de la banda.

El estimador tiene las propiedades de equivarianza sobre transformaciones lineales en la variable respuesta ó sobre transformaciones afines en la matriz del modelo.

### b) Estimador Elipsoide De Volumen Mínimo

El problema de identificación de observaciones discrepantes es resuelto sólo en parte al utilizar un estimador de regresión robusto. Estos estimadores identifican las observaciones discrepantes, pero no dan ninguna información acerca de la dirección de la discrepancia. En está subsección, se presenta un estimador con alto punto de ruptura

denominado elipsoide de volumen mínimo; el cual es útil para la identificación de observaciones discrepantes en la dirección de la matriz del modelo.

Considerando que la matriz $X$ está formada por $n$ vectores fila de dimensión $k$, la búsqueda de métodos robustos para identificar observaciones discrepantes en la dirección de las variables regresoras, será tratada dentro del contexto del análisis multivariado de datos.

La necesidad de caracterizar una observación discrepante en el espacio de dimensión $k$ hace necesario el uso de un método de sub ordenamiento de los puntos a través de una medida de distancia,

$$D(x; x_o, \Gamma) = (x - x_o)\Gamma^{-1}(x - x_o)^t. \tag{2.7}$$

Donde $x_o$ y $\Gamma$ son parámetros de posición y dispersión, respectivamente. Esta medida indica que tan lejos se encuentra un punto de su centroide tomando en consideración la dispersión del conjunto total de puntos en el espacio de las variables regresoras. Notar que, la distancia de Mahalanobis es un caso particular de la distancia definida en (2.7), cuando $x_o$ y $\Gamma$ son estimados por el vector de medias ($\overline{x}$) y la matriz de covarianzas muestrales ($S$), respectivamente.

Rousseuw (1985), propuso el estimador elipsoide de volumen mínimo (MVE), demostrando que es equivariante sobre transformaciones afines y posee un punto de ruptura del 50%.

**Definición 2.2.4.-** Sean $t \in \Re^k$ y $C \in PDS(k)$, donde PDS(K) es la clase de matrices simétricas y definidas positivas. Sea $X = (x_1, x_2, \ldots, x_n), x_i \in \Re^k, n \geq k+1$. Se define el estimador elipsoide de volumen mínimo como el par $(t(X), C(X))$, tal que minimiza el determinante de $C(X)$, sujeto a la restricción:

$$\#\left\{i : (x_i - t)C^{-1}(x_i - t)^t \leq b\right\} \leq q \tag{2.8}$$

$t_n(X)$ y $C_n(X)$ determinan el centroide y la estructura de covarianzas del elipsoide de volumen mínimo cubriendo por lo menos $q$ datos, donde $q = [(n + k + 1)/2]$. El escalar $b$ es fijado y no tiene influencia en el cálculo de $t(X)$, pero es muy importante en la determinación de la magnitud de $C(X)$.

Si $b$ es escogido en el dominio de una distribución de probabilidad de tipo elíptica, obtendremos estimadores consistentes de $x_o$ y $\Gamma$. En particular, bajo la suposición de que el conjunto de datos proviene de una distribución normal con parámetros $\mu \in \Re^k$ y $\Sigma \in PDS(k)$, $b$ será obtenida a partir de,

$$P_{\mu,\Sigma}\left\{(\bar{x}_i - \mu)\Sigma^{-1}(\bar{x}_t - \mu)^t \leq b\right\} = \frac{1}{2}.$$

La forma cuadrática $(\bar{x}_i - \mu)\Sigma^{-1}(\bar{x}_t - \mu)^t$ tiene distribución $\chi^2_k$, de modo que, $b$ corresponde a un percentíl de la distribución chi quadrado con $k$ grados de libertad, cubriendo como máximo 50% del área total de la curva.

### 2.2.3 Análisis exploratorio de Observaciones Discrepantes

Un diagnóstico de observaciones discrepantes, frecuentemente usado en el análisis de regresión, consiste en construir gráficos de residuos del ajuste mínimo cuadrático versus $h_{ii}$ o versus las distancias de Mahalanobis. Estos gráficos pueden también ser construidos relacionando las distancias en la matriz del modelo $(X)$ y los residuos del ajuste, calculados utilizando los estimadores robustos MVE y LMS descritos en la sección anterior. Los gráficos así construidos combinan la información de las observaciones discrepantes de la regresión con los puntos "leverage", posibilitando que sean observadas estructuras no necesariamente reveladas por el método de mínimos cuadrados.

## 3. ESTUDIO DE UN CASO

A continuación se analiza un conjunto de datos empleando los diagnósticos mínimo cuadráticos y robustos presentados en las sección anterior. Note que los puntos de corte generalmente usados en el análisis de residuos son $\pm 2.0$, pero ocurre que, los métodos robustos tienen tendencia a declarar más observaciones discrepantes de las que realmente pueden existir en el conjunto de datos. Por está razón y para facilitar la comparabilidad entre los métodos, en general se considerará una observación como discrepante si su residuo estandarizado cae fuera de la banda $(-2.5, 2.5)$. Asimismo, se identificará una observación como discrepante en la dirección de las variables regresoras sí su distancia excede el percentíl 97.5% de una distribución $\chi^2$ con $p-1$ grados de libertad.

La estimaciones robustas fueron calculadas mediante el paquete computacional PROGRESS (Rousseuw y Leroy (1987)), las distancias robustas se calcularon utilizando el programa computacional MINVOL obtenido de la biblioteca de programas computacionales "The Statlib Collection of Applied Statistics Algorithms" (STATLIBD@STAT.CMU.EDU).

Los datos que se utilizarán para la ilustración fueron reportados por Chambers y colaboradores (1983) y corresponden a las lecturas diarias de valores de la calidad del

aire en la ciudad de Nueva York durante el mes de Mayo de 1973. Las características observadas fueron: Concentración media de Ozono (ozono); Radiación Solar en la banda de frecuencia 4000-7700 Å(RAD-SOL); velocidad del aire (v-aire) y Temperatura del aire (tempera). El conjunto de datos contiene algunas observaciones discordantes genuinamente observados y varios datos contaminantes introducidos intencionalmente (999 y 9999 inicialmente utilizados para codificar datos faltantes). El objetivo es comparar los métodos de diagnósticos basados en el ajuste de mínimos cuadrados y en los métodos robustos.

El modelo formulado es

$$\text{OZONO} = \beta_0 + \beta_1 \text{RAD-SOL} + \beta_2 \text{V-AIRE} + \beta_3 \text{TEMPERA} + \varepsilon$$

La tabla N° 1, presenta los coeficientes de regresión del modelo ajustado por el método de mínimos cuadrados. Notar que con este método sólo el 23.42% de la variación en las lecturas diarias de la concentración media de ozono en el aire ($R^2$) está siendo explicado por el modelo ajustado, además el error estándar estimado ($\hat{\sigma}_\varepsilon$) es 336.923.

**Tabla 1. Estimación de parámetros del modelo de regresión para las lecturas diarias de la concentración de ozono en el aire**

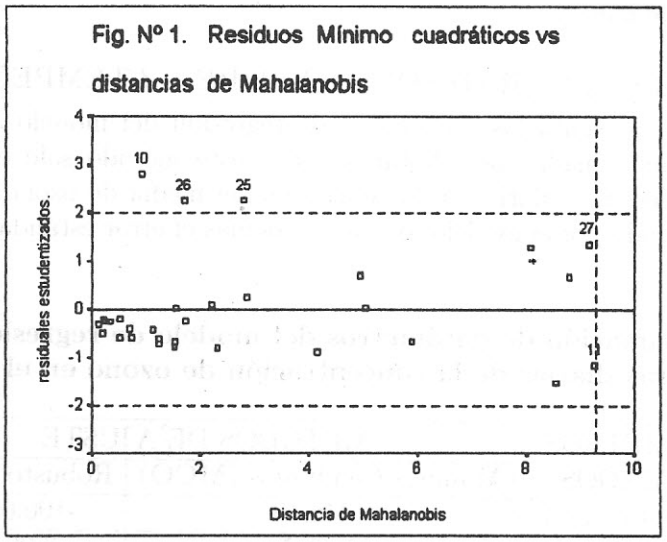
| PARÁMETROS ESTIMADOS | MÉTODOS DE AJUSTE | |
|---|---|---|
| | Mínimos Cuadrados (MCO) | Robusto (LMS) |
| Intercepto ( $\beta_0$ ) | 1340.19 | -100.05050 |
| Radiacin solar ($\beta_1$) | 0.0337 | 0.00104 |
| Veloc. del aire ($\beta_2$) | -1.431 | -0.11204 |
| Temperatura ($\beta_3$) | -18.180 | 1.80876 |

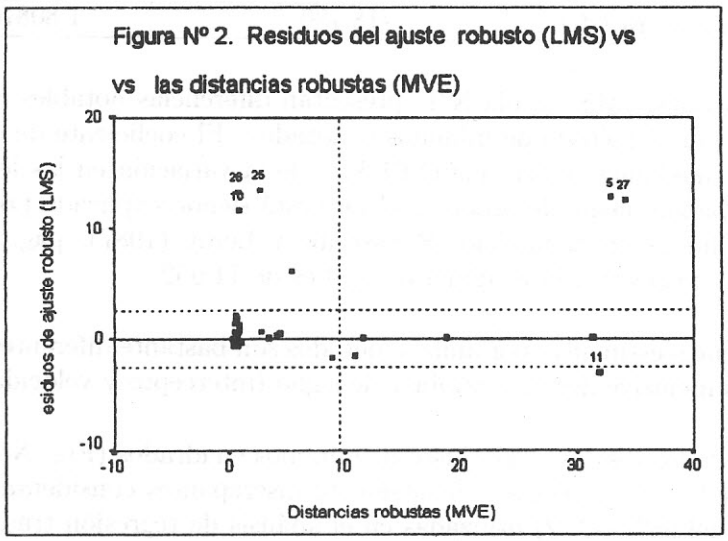
Las estimaciones LMS (Tabla N°1) presentan diferencias notables con respecto a las obtenidas por el método de mínimos cuadrados. El coeficiente de determinación basado en las medianas indica que el 61.55% de la variación en las lecturas diarias de la concentración media de ozono en el aire está siendo explicado por las variables regresoras incluidas en el modelo (Rousseeuw y Leroy (1987), pag. 45), el error cuadrático con respecto a la mediana ($\hat{\sigma}_{LMS}$) es de 11.952.

Los coeficientes estimados por ambos métodos son bastante diferentes en magnitud y dos de ellos inclusive llegan a cambiar de signo (intercepto y velocidad del aire).

El análisis de los residuos del ajuste de mínimos cuadrados (Fig. N°1) revela a las observaciones 10, 25, y 26 como fuertemente discrepantes considerando las bandas de confianza del 95% (-2, 2) utilizadas en el análisis de regresión tradicional. Notar

que, estas observaciones fueron contaminadas en la variable lecturas diarias de concentración media de ozono (999), pero, no son los únicos, las lecturas 5 y 27 también fueron contaminadas y no muestran residuos notables. Por otro lado, si utilizáramos la banda de confianza (-2.5, 2.5), solamente la lectura del décimo día aparecería como discrepante. Notar que ninguna observación excede la línea vertical trazada en el punto 9.34; correspondiente al intervalo de confianza del 95% para la distancia de mahalanobis, aún cuando las lecturas de los días 5, 6, 11 y 27 han sido contaminadas en la dirección de la variable radiación solar (9999).



El diagnóstico basado en el ajuste robusto (Figura N°2) revela claramente las discrepancias de las observaciones 10, 25 y 26 cuyas lecturas de concentración media de ozono en el aire fueron substituidas por valores contaminantes (999) y la lectura del día 30 de Mayo que es una observación verídica. Las observaciones 5 y 27 y 11 son discrepantes en la dirección de la variable lecturas de ozono y de la radiación solar (discrepantes e influyentes). Finalmente, la lectura del día 6 es discrepante solamente en la dirección de las variables regresoras (Tabla A3).

Otros diagnósticos basados en el método de mínimos cuadrados; útiles para medir la influencia de cada una de las observaciones sobre las estimaciones; son las estadística DFFITS, DFBETAS y la distancia de Cook, analicemos los datos utilizando cada uno de estos diagnósticos.

Utilizaremos los puntos de corte sugeridos por Belsley y Welsch (1980), (DFFITS = 0.655, DFBETAS = 0.359, D. Cook = 1). Los datos 5 y 27 son discrepantes en la dirección de las lecturas de ozono y la temperatura, sin embargo, se observa que solamente la lectura del día 27 influye fuertemente sobre el valor de todos los coeficientes estimados, mientras que la observación 5 solamente afecta al coeficiente de la radiación solar. Las observaciones 6, 10, 11, 25 y 26 afectan sólo uno o dos de los coeficientes estimados (Tablas A1, A3).

En conclusión, el método robusto identifica claramente a las observaciones 5, y 27 como fuertemente discrepantes en la dirección de la respuesta y de la matriz del modelo. Notar que estos datos discrepantes han sido enmascarados y no fueron identificados por los diagnósticos basados en el método de mínimos cuadrados. Las observaciones 10, 25 y 26 son discrepantes en la dirección de la variable concentración media del ozono (Fig N°3). Estas han sido detectadas fácilmente por ambos métodos por que muestran residuos grandes pues no son influyentes. Las lecturas 6, 11, 29 son discrepantes en la dirección de la matriz del modelo, pero no muestran residuos notables. Finalmente, la lectura correspondiente al sexto día es la más influyente, pues afecta a casi todos los parámetros estimados.

## 4. CONCLUSIONES Y RECOMENDACIONES

Para poder identificar las observaciones influyentes usando el ajuste mínimo cuadrático ha sido necesario calcular los residuos del ajuste, la distancia de mahalanobis o los elementos de la diagonal de la matriz $H$. Además se ha tenido que analizar las estadísticas de influencia para finalmente identificar los datos gruesamente influyentes y discrepantes que se introdujeron en el conjunto de datos.

Si bien es cierto que los "paquetes" estadísticos incluyen herramientas para calcular todas estas estadísticas, se requiere una cierta práctica para poder interpretarlas.

Los diagnósticos basados en el método de mínimos cuadrados son útiles para identificar observaciones discrepantes cuando el conjunto de datos contiene sólo "outliers" individuales, pero en caso de existir más de un dato discordante, estos no necesariamente son identificados y los efectos que producen en las estimaciones mínimo cuadráticas pueden ser desastrosos. Por lo tanto, es aconsejable utilizar ambos métodos de diagnósticos, los tradicionales y robustos. Posteriormente, si no existen datos discordantes se puede utilizar el método de mínimos cuadrados para analizar los datos.

## 5. BIBLIOGRAFÍA


[1] P. Y. AGÜERO *Estimadores de regressão com alto ponto de ruptura e detecção de multiples observacões discrepantes.* IMECC- UNICAMP (Brasil).(1994).

[2] R. J. BECKMAN & R. D. COOK *OUTLIER S.* Technometrics Vol. 25, N°2.(1983).

[3] D. A. BELSLEY - E. KUH & R. E. WELSCH *Regression Diagnostics Identifying Influential data and sources of Colinearity.* Wiley: New York. (1980).

[4] G. E. P. BOX *Non Normality and Test on Variance.* Biometrica, Vol. 40,(1953), p. 318-335.

[5] J. M. CHAMBERS - W. S. CLEVELAND - B. KIENER - P. A. TUKEY *Graphical Methods for Data Analisys.* Wadsworth International Group Belmont, C.A.(1983).

[6] E. W. CHENEY, *Introduction OF Aproximation Theory.* McGraw-hill.(1966).

[7] R. D. COOK & S. WEISBERG, *Residuals And Influence In Regression.* New York: Chapman And Hall. (1982).

[8] F. R. HAMPEL, *A general qualitative definition of robustness.* The Annals of Mathematical Statistics. Vol. 42, N°6,(1971). p. 1887-96.

[9] P. J. HUBER, *Robust Statistics.* Edit. John Wiley & Sons. (1981).

[10] P. J. ROUSSEEUW & M. A. LEROY, *Robust Regression And Outliers Detection.* John Wiley & Sons.(1987).

[11] A. J. STROMBERG, *Computing The Exact Value Of The Least Median Squares estimate And Stability Diagnostics In Multiple Linear Regression.* Technical Report N 561. Department Of Statistics. University Of Minnesota (1991).

## APÉNDICE

**Tabla A1.** Estudio de la calidad del Aire en cierta ciudad. Mediciones de concentración de Ozono (ozono). Radiación solar (rad-sol), velocidad del aire (v-aire) y temperatura (tempera) y residuos del ajuste del modelo mediante el método de Mínomos cuadrados ordinarios (MCO) y un ajuste de mínimas medianas al cuadrado (LMS).

| Obs. | Observaciones | | | | Residuos del ajuste | |
|---|---|---|---|---|---|---|
| | Rad-sol | V-aire | Temperat. | Ozono | MCO ($t_i$) | LMS ($e_i$) |
| 1 | 190 | 7.4 | 67 | 41 | -0.2386 | 1.71 |
| 2 | 118 | 8.0 | 72 | 36 | 0.0349 | 0.55 |
| 3 | 149 | 12.6 | 74 | 12 | 0.0938 | -1.72 |
| 4 | 313 | 11.5 | 62 | 18 | -0.5698 | 0.57 |
| 5 | 9999 | 14.3 | 56 | 999 | 1.2947 | 82.75 |
| 6 | 9999 | 14.9 | 66 | 28 | -1.5765 | 0.00 |
| 7 | 299 | 8.6 | 65 | 23 | -0.4067 | 0.51 |
| 8 | 99 | 13.8 | 59 | 19 | -0.07093 | 1.15 |
| 9 | 19 | 20.1 | 61 | 8 | -0.6471 | 0.00 |
| 10 | 194 | 8.6 | 69 | 999 | 3.2891 | 81.58 |
| 11 | 9999 | 6.9 | 74 | 7 | -1.1632 | -3.05 |
| 12 | 256 | 9.7 | 69 | 16 | -0.1944 | -0.66 |
| 13 | 290 | 9.2 | 66 | 11 | -0.3807 | -0.64 |
| 14 | 274 | 10.09 | 68 | 14 | -0.2497 | -0.67 |
| 15 | 65 | 13.2 | 58 | 18 | -0.7728 | 1.22 |
| 16 | 334 | 11.5 | 64 | 14 | -0.4709 | -0.06 |
| 17 | 307 | 12.0 | 66 | 34 | -0.2961 | 1.31 |
| 18 | 78 | 18.4 | 57 | 6 | 0.8872 | 0.41 |
| 19 | 322 | 11.5 | 68 | 30 | -0.2038 | 0.67 |
| 20 | 44 | 9.7 | 62 | 11 | -0.5801 | 0.00 |
| 21 | 8 | 9.7 | 59 | 1 | -0.7930 | -0.38 |
| 22 | 320 | 16.6 | 73 | 11 | 0.0393 | -1.63 |
| 23 | 25 | 9.7 | 61 | 4 | -0.6598 | -0.44 |
| 24 | 92 | 12.0 | 61 | 32 | -0.5595 | 1.92 |
| 25 | 66 | 16.6 | 57 | 999 | 2.4837 | 83.48 |
| 26 | 266 | 14.9 | 58 | 999 | 2.4586 | 83.30 |
| 27 | 9999 | 8.0 | 57 | 999 | 1.3632 | 82.54 |
| 28 | 13 | 12.0 | 67 | 23 | -0.2455 | 0.27 |
| 29 | 252 | 14.9 | 81 | 45 | 0.6837 | 0.00 |
| 30 | 223 | 5.7 | 79 | 115 | 0.6913 | 6.07 |
| 31 | 279 | 7.4 | 76 | 37 | 0.2464 | 0.01 |

Fuente: Chambers et al (1983)

**Tabla A2.** Elementos de la diagonal de la matriz de proyección, distancias de mahalanobis y distancias robustas.

| N° de Obs. ($i$) | Diagonal de la Matriz de proyección ($h_{ii}$) | Distancia de Mahalanobis ($MD_i$) | Dist. Robustas ($RD_i$)[1] |
|---|---|---|---|
| 1 | 0.09 | 1.754 | 0.561 |
| 2 | 0.08 | 1.562 | 2.755 |
| 3 | 0.10 | 2.172 | 8.873 |
| 4 | 0.06 | 0.520 | 0.387 |
| 5 | 0.30 | 8.056 | 3297.75 |
| 6 | 0.32 | 8.526 | 3135.79 |
| 7 | 0.07 | 1.149 | 0.719 |
| 8 | 0.07 | 1.229 | 0.837 |
| 9 | 0.23 | 5.872 | 11.59 |
| 10 | 0.06 | 0.942 | 0.807 |
| 11 | 0.34 | 9.278 | 3188.69 |
| 12 | 0.05 | 0.524 | 0.753 |
| 13 | 0.06 | 0.701 | 0.327 |
| 14 | 0.04 | 0.238 | 0.719 |
| 15 | 0.08 | 1.513 | 0.822 |
| 16 | 0.04 | 0.213 | 0.283 |
| 17 | 0.04 | 0.127 | 0.456 |
| 18 | 0.17 | 4.143 | 4.356 |
| 19 | 0.04 | 0.221 | 0.817 |
| 20 | 0.07 | 1.248 | 0.597 |
| 21 | 0.11 | 2.312 | 1.003 |
| 22 | 0.20 | 5.012 | 10.80 |
| 23 | 0.08 | 1.554 | 0.719 |
| 24 | 0.06 | 0.724 | 0.486 |
| 25 | 0.13 | 2.783 | 2.637 |
| 26 | 0.09 | 1.699 | 0.719 |
| 27 | 0.34 | 9.150 | 3423.50 |
| 28 | 0.04 | 0.233 | 4.031 |
| 29 | 0.33 | 8.742 | 21.66 |
| 30 | 0.20 | 4.959 | 5.395 |
| 31 | 0.13 | 2.848 | 3.352 |

[1] El punto de corte para las distancias de Mahalanobis y robustas es $\chi^2_{3.975} = 9.3484$.

**Tabla A3.** Diagnósticos de Influencia del ajuste de mínimos cuadrados: Estadísticas DFFITS, de COOK y DFBETAS.

| Obs. | DFFITS | COOK | DFBETAS (0.359) | | | |
|---|---|---|---|---|---|---|
| ($i$) | (0.686) | (1.0) | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ |
| 1 | -0.08 | 0.001 | -0.03 | 0.02 | 0.06 | 0.02 |
| 2 | 0.01 | 0.000 | -0.00 | -0.00 | -0.01 | 0.00 |
| 3 | 0.03 | 0.000 | -0.02 | -0.00 | 0.01 | 0.03 |
| 4 | -0.13 | 0.004 | -0.08 | 0.05 | 0.03 | 0.07 |
| 5 | 0.85 | 0.176 | 0.14 | 0.70 | 0.16 | -0.22 |
| 6 | -1.07 | 0.273 | 0.40 | -0.95 | -0.48 | -0.30 |
| 7 | -0.11 | 0.003 | -0.06 | 0.04 | 0.08 | 0.04 |
| 8 | -0.20 | 0.010 | -0.10 | 0.07 | -0.03 | 0.11 |
| 9 | -0.35 | 0.032 | 0.10 | 0.03 | -0.31 | -0.03 |
| 10 | 0.86 | 0.135 | 0.13 | -0.48 | -0.48 | 0.08 |
| 11 | -0.84 | 0.173 | 0.22 | 0.69 | 0.16 | -0.31 |
| 12 | -0.04 | 0.001 | 0.00 | 0.01 | 0.02 | -0.01 |
| 13 | -0.09 | 0.002 | -0.04 | 0.03 | 0.05 | 0.02 |
| 14 | -0.05 | 0.001 | 0.00 | 0.02 | 0.01 | -0.01 |
| 15 | -0.23 | 0.014 | -0.15 | 0.08 | 0.00 | 0.16 |
| 16 | -0.10 | 0.002 | -0.04 | 0.03 | 0.02 | 0.03 |
| 17 | -0.06 | 0.001 | -0.00 | 0.02 | -0.01 | -0.00 |
| 18 | -0.40 | 0.041 | -0.03 | 0.06 | -0.27 | 0.11 |
| 19 | -0.04 | 0.000 | 0.01 | 0.01 | -0.00 | -0.01 |
| 20 | -0.16 | 0.007 | -0.12 | 0.06 | 0.10 | 0.09 |
| 21 | -0.28 | 0.020 | -0.24 | 0.10 | 0.16 | 0.21 |
| 22 | 0.02 | 0.000 | -0.02 | 0.00 | 0.02 | 0.01 |
| 23 | -0.20 | 0.010 | -0.15 | 0.08 | 0.12 | 0.13 |
| 24 | -0.14 | 0.005 | -0.08 | 0.05 | 0.02 | 0.08 |
| 25 | 0.98 | 0.185 | 0.25 | -0.21 | 0.46 | -0.41 |
| 26 | 0.77 | 0.124 | 0.32 | -0.20 | 0.24 | -0.41 |
| 27 | 0.97 | 0.229 | 0.47 | 0.68 | -0.40 | -0.42 |
| 28 | -0.05 | 0.001 | 0.01 | 0.02 | -0.01 | -0.01 |
| 29 | 0.47 | 0.057 | -0.43 | 0.02 | 0.29 | 0.42 |
| 30 | 0.34 | 0.030 | -0.11 | -0.04 | -0.15 | 0.19 |
| 31 | 0.09 | 0.002 | -0.03 | -0.01 | -0.03 | 0.05 |