XVI Reunión ASEPELT
Madrid, 20 y 21 de Junio de 2002
Universidad San Pablo CEU

Notas sobre Estadística Robusta

Juan Fco. Ortega Dato

Área de Matemáticas. Facultad de CC. Económicas y Empresariales de Albacete.
Universidad de Castilla-La Mancha.

Socio de Número: 2303

Dirección Postal: Plaza de la Universidad, 1. 02071. Albacete.

Correo Electrónico: JuanFco.Ortega@uclm.es

Artículo Presentado como: Ponencia

Área de conocimiento: Métodos Cuantitativos

Notas sobre Estadística Robusta

Juan Fco. Ortega Dato*

Resumen

La Estadística Robusta, de antecedentes más bien jóvenes, estudia el comportamiento de diferentes procedimientos utilizados en estadística cuando existe una pequeña variación en los supuestos iniciales, o cuando el modelo está contaminado por ciertas observaciones conocidas por el nombre de "Outliers" (Observaciones Atípicas para nosotros), que producen malas influencias en los resultados, proporcionando modelos erróneos. En los últimos años, esta disciplina se ha consolidado, comprobándose su utilidad en diversos ámbitos de la estadística y aconsejándose su uso en toda modelización. Así, con el propósito de divulgar los conceptos definidos en esta rama de la estadística, que son, a mi parecer, imprescindibles en todo estudio estadístico, y el de iniciar al lector en el tema, en este trabajo se pretende mostrar de una manera sencilla, aunque rigurosa, la base sobre los principales elementos que conforman la Estadística Robusta, proporcionando al tiempo, una buena bibliografía de consulta, para aquéllos que deseen conocer o ampliar algunos aspectos concretos sobre el tema.

Palabras Clave: Robustez, "Outliers", Observaciones Atípicas, Características en Robustez, Estimadores Robusto, Detección de "Outliers".

*Área de Matemáticas. Facultad de CC. Económicas y Empresariales de Albacete. Universidad de Castilla-La Mancha. Correo Electrónico: JuanFco.Ortega@uclm.es

1 Introducción

Es frecuente que los modelo estadístico en el que intervienen variables, de las cuales se quiere estudiar su comportamiento o interrelación, se construyan a partir de la información suministrada por una muestra de éstas, de manera que el modelo refleje de una manera fiel la estructura de la muestra de dichas variables, y nos permita, en el caso de interrelaciones, predecir comportamientos generales cuando se ha elegido una variable como variable explicada.

Construir un modelo partiendo de la información suministrada por muestras parece ser una idea bastante acertada, siempre que éstas estén "bien determinadas", es decir, que se hayan tomado en condiciones similares y sin errores. Así, ya que nunca podemos estar seguros de si una determinada observación es un reflejo de la realidad que queremos modelizar, un primer problema que se nos puede plantear en algunos estudios realizados en base a muestras de variables, es que podemos encontrar observaciones no deseables en la muestra, que enturbian y nos proporcionan información errónea sobre la estructura del grueso de las observaciones de ésta, siendo la única manera de tener la certeza de que un dato es genuino, el estudio de su concordancia con el resto de elementos de la muestra.

Dada una muestra de observaciones, los análisis intentan responder a una serie de preguntas tales como; ¿Todos los datos aportan información de la misma realidad, o existe algún conjunto que sigue una dinámica diferente? En el caso en que existan diferentes realidades, ¿Qué cuenta la mayoría de los datos? ¿Qué minoría se distingue en su comportamiento y qué cuenta? El intento de responder a estas preguntas, ha motivado la aparición de técnicas o métodos encuadrados dentro de una rama de la Estadística conocida por *Estadística Robusta*.

En definitiva, la Estadística Robusta estudia el comportamiento de los procedimientos estadísticos, ante pequeños cambios en los supuestos iniciales de un determinado modelo, o ante la presencia de ciertas observaciones no deseables conocidas por *Observaciones "Outliers"*, a las cuales nos referiremos con el término castellano de *Observaciones Atípicas*, cuya definición más extendida es la de aquellas observaciones, que parecen ser sorprendentemente discordantes o inconsistentes con respecto a la mayoría de los datos de la muestra.

Las principales líneas de investigación dentro de la Estadística Robusta, para el estudio de la influencia de las Observaciones Atípicas, se pueden encuadrar en tres grupos, que son: el estudio de las denominadas *Características en Robustez*, conjunto compuesto por las propiedades o características aconsejables para los procedimientos utilizados en presencia de Observaciones Atípicas; la propuesta de *Estimadores Robustos*, los cuales pretenden que la elección del modelo definitivo no sea afectada por estas observaciones indeseables, y así asegurar una respuesta proporcionada por la mayoría de los elementos de la muestra; y por último, la definición de *Métodos de Detección de "Outliers"*, propuestos con la intención de detectar las posibles Observaciones Atípicas, y eliminarlas si fuera necesario, para proteger la información genuina de la muestra dada.

Una vez introducido a grandes rasgos el tema a tratar, el objeto de este papel, desarrollado como consecuencia de un trabajo más amplio recogido en Ortega (2000), es el de presentar y ahondar de una manera simple, con comentarios y ejemplos sencillos, en las definiciones y propiedades de los elementos más destacados de la Estadística Robusta, al tiempo que se pretende proporcionar una amplia bibliografía sobre el tema en cuestión, con el propósito de divulgar y animar a su utilización.

Así, los contenidos de este papel están distribuidos en las siguientes forma: En la siguiente sección, realizaremos un repaso somero por la historia de los estudios estadísticos sobre el tema que nos ocupa, destacando los acontecimientos más significativos que tuvieron como fin la Estadística Robusta. Seguidamente, en la tercera sección, se presentarán los primeros conceptos necesarios en el tema, junto con la notación utilizada en el trabajo. Posteriormente, en la cuarta sección, y utilizando la notación introducida, se estudiarán los elementos más destacados en las tres principales líneas de investigación dentro de la Estadística Robusta que son; las denominadas Características en Robustez; la propuesta de Estimadores Robustos; y la definición de Métodos de Detección de "Outliers". Para finalizar el trabajo, en la última sección se realizará unos comentarios sobre los aspectos más importantes del estudio realizado.

2 Antecedentes

En la historia de la Estadística Robusta se pueden diferenciar dos etapas divididas por las publicaciones, en los años sesenta, de los trabajos de Huber (1964). En estas publicaciones el autor presenta una recopilación de estudios realizados sobre el tema que nos ocupa, y propuso conceptos y teorías nuevas que formaron los cimientos y abrieron un gran campo de investigación de la que hoy se considera una materia imprescindible en toda modelización.

En la primera etapa, los principales estudios en estadística estaban orientados a la predicción de valores futuros de una determinada variable, donde, de una forma matemáticamente formalizados, ya en el siglo XVI estos trabajos estaban relacionados con quizás la primera ciencia, la Astronomía. El estudio y determinación de la estructura física de la Tierra y de las órbitas de los planetas visibles, fue el centro de atención de algunos científicos del momento como J.Kepler (1571-1630) e I.Newton (1642-1727), siendo los modelos lineales y la introducción del método de Mínimos Cuadrados, por su simplicidad, los primeros elementos utilizados para la predicción. Fue poco el tiempo en el que los estudiosos del momento se percataron de lo sensible que eran los procedimientos utilizados para la determinación de un modelo fiable bajo presencia de observaciones extrañas, proponiéndose diferentes procedimientos para proteger y salvar la información de la muestra, de manera que fueran resistentes a estas observaciones indeseables, o que pudieran detectarlas con el fin de estudiar su consideración o no en el modelo.

La mayoría de los primeros procedimientos utilizados para el tratamiento de estas observaciones, fueron de los que podemos llamar *de andar por casa*, sin ninguna base científica, y donde el mayor peso en la decisión del modelo elegido recaía en la predisposición y experiencia del analista. Tenemos que hacer notar, en favor de estos científicos, que las herramientas de las cuales disponían o podían hacer uso, no eran las más apropiadas, de manera que, la mayoría de los procedimientos que hoy se conocen, no se habrían podido utilizar en los comienzos por falta de las teorías y conocimientos en las áreas de matemáticas y estadística, y sin los medios técnicos como el ordenador para el tratamiento de problemas con muchas variables y/u observaciones.

En los trabajos pioneros en estadística que estudian este problema correspondientes a los siglos XIX y principios del XX (Stigler 1973), se utilizaban métodos que en la

literatura se han llamado *Métodos o Test de Rechazo*, propuestos por autores como R.J.Boscovich, D.Bernoulli, F.W.Bessel y J.J.Baeuer, o F.Galton (Allen 1961 y Bustos 1988). Estos métodos no consideraban aquellas observaciones de la muestra que parecían demasiado desviadas del resto, siendo procedimientos muy drásticos y no respaldados por una base científica, siendo su único sustento la experiencia y el buen juicio del analista. Unos procedimientos de este tipo son los propuestos en estas fechas por W.Chauvenet, con el nombre de *Test de Chauvenet*, y por E.J.Stone, basado en un nuevo concepto de nombre *Módulo de Cuidado*. Otros método, también de la misma época, son los propuestos por autores como J.W.L.Glaisher, F.I.Edgeworth, este último mediante el uso de *ponderaciones* en las observaciones con el fin de predeterminar su influencia en el modelo. En estos años, T.W.Wright propuso otro procedimiento, que quedaría encuadrado en los Métodos de Rechazo, basado en la relación entre la diferencia de las observaciones con la media muestral y la desviación típica muestral. En él, se rechazaban las observaciones cuya diferencia con la media era más de tres veces la desviación típica (Bustos 1988).

Ya en el siglo XX, sobre todo en los años cuarenta y cincuenta, se presentan una gran cantidad de estudios referidos a la búsqueda de tests para detectar Observaciones Atípicas, como los realizados por H.N.Goodwin, J.O.Irwin, W.R.Thompson, M.Greenwood, K.R.Nair y F.Mosteller (Barnett y Lewis 1987). En resumen, desde los inicios hasta los años 50 (1950), se presentan una gran variedad de procedimientos que no llegarán a utilizarse de manera general por la comunidad científica (es decir, cada método tenía un propulsor que en definitiva era el que lo utilizaba), por lo que han sido olvidados o han tenido una pobre divulgación y por lo tanto utilización.

A partir de los años sesenta, se realizan estudios en los que se dan los primeros pasos firmes para la construcción de lo que hoy se conoce por Estadística Robusta. Así, esta rama de la Estadística se consolida como consecuencia de una serie de trabajos que culminan con el presentado por P.J.Huber en 1964 con el nombre de *"Robust estimation of location parameter"*. En este trabajo Huber propone conceptos y teorías que motivaron a otros científicos a investigar y utilizar estas herramientas estadísticas.

Los acontecimientos a partir del trabajo de Huber se han desarrollado de manera muy rápida y fulgurante. Al completarse la base de estudio de la problemática por la presencia de Observaciones Atípicas, la aparición de métodos y teorías del tema que

nos ocupa ha inundado la literatura de artículos en los que se pretende avanzar hacia la solución del problema. Estos trabajos, que trataremos en las siguientes secciones, están desarrollados por una gran cantidad de autores que, alentados por lo atractivo del tema y/o por lo útil que éste es para todo estudio estadístico, han inundado de artículos las revistas más importantes que tratan temas de estadística a partir de los años sesenta, como *Journal of the American Statistical Association, The Annals of Statistics, Journal of the Royal Statistical Society, Journal of Applied Statistics*, etc. Como muestra de los autores que han estudiado este tema podemos nombrar, siendo los más importantes y por aproximado orden cronológico, al ya citado P.J.Huber, F.R.Hampel, D.R.Cook, J.W.Tukey, V.Barnett, V.J.Yohai, R.A.Maronna, D.L.Donoho, R.H.Zamar, P.J.Rousseeuw, E.M.Ronchetti, S.Chatterjee, D.Peña, X.He y W.A.Stahel. Naturalmente, la lista dada no incluye a todos los investigadores en este campo, aunque las aportaciones de los aquí nombrados sí que puede representar una gran parte de lo publicado, o al menos, el punto de partida de la mayoría de los estudios realizados por otros estadísticos.

3 Primeros Conceptos y Notación

Para mostrar los elementos más significativos en Estadística Robusta, debemos comenzar por dejar claro cual es el problema a tratar, su primeros elementos básicos, y también, la notación mínima que nos es necesaria para definir y comprender los conceptos que aquí propondremos. Así, tenemos que empezar dejando claro qué entendemos por Observación Atípica, si existen diferentes tipos de estas observaciones y el efecto que producen al estar contenidas en una determinada muestra que es utilizada para estimar un modelo, y posteriormente, presentaremos otros conceptos básicos y la notación necesaria para el estudio del problema por la presencia de estas observaciones.

Asumiendo como definición de Observación Atípica, aquella observación que parece ser sorprendentemente discordante o inconsistente con respecto a la mayoría de las observaciones de la muestra, es posible distinguir entre varios tipos de observaciones que responden a esta definición, dependiendo de su procedencia, su impacto en el cálculo de los parámetros del modelo, o, también, su posición en relación con la muestra en estudio. De esta manera, proponemos una clasificación de diferentes Observaciones Atípicas, la

cual intenta englobar y recoger las encontradas en la literatura. En primer lugar, llamaremos *Observación Contaminante* a aquella que no es una realización del experimento que estamos estudiando, quizá por proceder de una distribución diferente de la que genera los elementos de la muestra, o por errores cometidos a la hora de la construcción de la muestra. Un segundo conjunto, esta formado por las llamadas *Observaciones Influyentes*, concepto el cual se puede definir como, aquellas observaciones que tienen un gran impacto en la estimación de los parámetros, y por lo tanto, su presencia determina la construcción final del modelo. Y el tercer y último tipo de Observación Atípica es el llamado *Observación Extremo*, donde se incluye en este concepto aquella o aquellas observaciones con el menor y mayor valor en la muestra.

Entre los conceptos aquí citados no existe una relación de inclusión estricta. Así, es posible que una Observación Contaminante no produzca una influencia negativa a la hora de estimar un determinado parámetro (por quedar, por ejemplo, dentro de la nube de observaciones genuinas), de manera que no sería considerada como Observación Atípica. Por otra parte, existen Observaciones Influyentes cuyo impacto en la construcción del parámetro a estimar es positivo, reforzando así su determinación. Por último, en el caso de las Observaciones Extremos, éstas pueden quedar, como en el caso de las Observaciones Contaminantes, dentro de la nube de observaciones genuinas, no suponiendo un problema en la estimación. En definitiva, ya que la definición de Observación Atípica pretende recoger aquellas sorpresa por la inconsistencia con el resto de elementos muestrales, de manera que proporcionan una influencia indeseable en los resultados, éstas cumplen ciertas condiciones incluidas en las definiciones de los tres tipos de observaciones citados.

Ahora, al tiempo que introducimos los primeros elementos necesarios para el estudio de los métodos en Estadística Robusta, también iremos presentado la notación más extendida en los trabajos sobre el tema, y que posteriormente será de utilidad.

Unos elementos utilizados en Estadística son los llamados *Estimadores*, denotados genéricamente por T. Las tres familias de estimadores más habituales en el tema que estamos tratando, y para los cuales se definen la mayoría de las propiedades o características en robustez, son las conocidas por los nombres de; *Estimadores de Posición* (T_p), *Estimadores de Escala* (T_e), y *Estimadores de Regresión* (T_r).

Así, dada una muestra de una variable, que denotaremos genéricamente por Z,

podemos definir un Estimador de Posición sobre la muestra, que denotaremos por $T_p(Z)$, como aquél que nos proporciona una aproximación de un cierto lugar predefinido de la distribución de la variable de la que procede. Cuando ese cierto lugar predefinido es una posición central de la distribución, los estimadores reciben el nombre de Estimadores de Posición Central, si bien a lo largo del trabajo, y siempre que no haya lugar a dudas, nos referiremos a ellos como Estimadores de Posición, en pro de simplificar el lenguaje, siendo los más utilizados la *Media* y la *Mediana*.

Por otra parte, un Estimador de Escala definido sobre la muestra considerada, que denotaremos por $T_e(Z)$, nos proporciona una aproximación de la *dispersión* o de la variabilidad de la distribución de la variable de la que procede las observaciones de la muestra. El más conocido y utilizado es la *Varianza*, o su raíz cuadrada positiva la *Desviación Típica*.

Por último, para el caso de Estimadores de Regresión necesitamos la definición de un *Modelo de Regresión Lineal*, el cual supone que existe una relación lineal entre un conjunto de p variables, llamadas *Variables Explicativas* y que denotaremos por $\{x_i\}_{i=1,2,\ldots,p}$, y otra variable a la cual se le llama *Variable Explicada*, y que denotaremos por y. Así, el modelo quedaría definido de la forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + u \tag{1}$$

donde a los $\{\beta_i\}_{i=0,1,\ldots,p}$ se les llama *Parámetros de Regresión*, y donde u es una variable que recoge la influencia en y de otras variables no incluidas en el modelo.

Para el cálculo de los Parámetros de Regresión del modelo (**??**), supondremos conocida una *Muestra* de n elementos de las variables del modelo $z = (y, x_1, x_2, \ldots, x_p)$, denotada por Z. Por otra parte, con el fin de construir el modelo (**??**) en forma matricial, cuando sea necesario, se considerará la matriz compuesta por las observaciones de las variables del modelo $Z = (Y|X)$, es decir, la *Matriz Ampliada* de Y y X, donde Y e X son las matrices de las observaciones de la Variable Explicada y Explicativa, respectivamente. Un caso particular, por su gran difusión, es el conocido por *Estimador de Regresión de Mínimos Cuadrados*, que denotaremos por *LS-Est.*.

Por otra parte, para el estimador T en general, los *Errores de Estimación* serán denotados por $\{e_i\}_{i=1,2,\ldots,n}$.

Otros concepto útiles en el trabajo, son:

Para poder realizar un estudio en muestras con Observaciones Atípicas, debemos tener una manera de introducir este tipo de observaciones en los modelos para comprobar su impacto en los resultados. Naturalmente no pretendemos introducir observaciones indeseables en muestras que han sido bien determinadas, sino estudiar, mediante la construcción de estas muestras, cómo afectan estas observaciones en el comportamiento de diferentes procedimientos. Así, dada una muestra Z de tamaño n, y siendo I un conjunto de subíndices, es decir, $I \subseteq \{1, 2, ..., n\}$, definimos una nueva muestra llamada *Contaminación* de Z en las m Observaciones con Subíndices en I, y denotándola por Zc_m^I, a la muestra construida al reemplazar las observaciones de Z con subíndices en I, con $Card(I)=m$ (donde *Card* denota al *Cardinal*), por otras tantas observaciones arbitrarias. Las simplificaciones con notación Zc_m y Zc^I, corresponden respectivamente a; Contaminación de Z en m observaciones, sin especificar cuales; y a Contaminación de Z en las observaciones con subíndices en I. En general llamaremos Contaminación de una muestra a la sustitución de una o varias observaciones de dicha muestra por otras arbitrarias.

Algunos de los conceptos que estudiaremos posteriormente, utilizan, la ordenación de los elementos en un determinado conjunto, o cálculos sobre conjuntos en los que no se consideran un determinado valor o un subconjunto de ellos. Así, serán necesarias las siguientes notaciones: Siendo $\{t_i\}_{i=1,2,...,n}$ una colección de valores de números reales ($t_i \in \mathbb{R}\ \forall i$), denotamos por $t_{[i]}$ al valor situado en la posición i-ésima en una ordenación de menor a mayor de $\{t_i\}_{i=1,2,...,n}$.

Dada la muestra Z de tamaño n de las variables de un modelo, y siendo I un subconjunto de subíndices, denotaremos por $T_{(I)}(Z)$ al estimador sin considerar las observaciones con índices en I. En general, elementos en los que su notación incluye un subíndice entre paréntesis, formado por índices de observaciones, se debe entender que dicho elemento no considera esas observaciones.

4 Principales Líneas de Trabajo en Estadística Robusta

Las tres principales líneas de acción sobre los que tratan los más recientes e importantes trabajos en Estadística Robusta, y que serán estudiadas en este papel, son; primeramen-

te la compuesta por aquellas propiedades o características que se consideran adecuadas para poder afirmar que un determinado método es un método eficiente para el estudio de Observaciones Atípicas, que llamamos Características en Robustez; la segunda, referida a la propuesta y estudio de los denominados Estimadores Robustos, es decir, de técnicas de estimación que no sean influidas por las Observaciones Atípicas, o lo sean poco; y por último, la tercera, que trata sobre los llamados Métodos de Detección de "Outliers", los cuales intentan determinar qué observaciones tienen una mayor influencia en el proceso de construcción del modelo (entre las cuales se encontrarán las Observaciones Atípicas), con el objeto de analizar su posible exclusión previamente al proceso de estimación.

Veamos los principales elementos dentro de estas tres líneas de trabajo.

4.1 Características en Robustez

Los estudios en Estadística Robusta han generado un conjunto de propiedades y conceptos, en general *características*, que nos informan de lo óptimo que sería la utilización de un determinado procedimiento en presencia de Observaciones Atípicas. Así, conocido el conjunto de propiedades y conceptos definidos en la teoría clásica, y que son pedidas por lo general a cualquier estimador paramétrico (como Insesgadez, Eficiencia, Equivarianzas, Suficiencia, Consistencia u otras), en Estadística Robusta podemos considerar un grupo de conceptos y propiedades más específicas para el tratamiento de contaminaciones, como son la propiedad conocida por el nombre de *Propiedad de Ajuste Exacto*, los conceptos relacionados con las denominadas *Función de Influencia* y *Curva de Sensibilidad*, el concepto llamado *Punto de Ruptura* y por último los conceptos conocidos por *Efectos de Enmascaramiento* (Efectos *"Masking"* y *"Swamping"*).

Así, entendiendo por *Método Robusto*, o en general *Robustez*, la resistencia de éste ante pequeñas contaminaciones en la muestra dada, de manera que el resultado de aplicar ese método sea un reflejo del comportamiento de la mayor parte de las observaciones de la muestra, y que el resultado no esté determinado por unos pocos datos, podemos considerar, en el estudio de la robustez de un determinado método, el conjunto que llamaremos *Características en Robustez* formado por las propiedades y conceptos antes citados, de manera que nos proporcionan información sobre si es o no un método Robusto, o qué grado de Robustez tiene. Los conceptos de la teoría clásica pueden ser consultados en diversos manuales de Estadística (ver por ejemplo Peña 1992), mientras

que aquí nos centraremos en el desarrollo de los conceptos que tienen una interpretación o interés particular en robustez, como la Eficiencia y la Consistencia, o son más específicos de las teorías robustas, como la Propiedad de Ajuste Exacto, los conceptos relacionados con las denominadas Función de Influencia y Curva de Sensibilidad, el Punto de Ruptura y por último los Efectos de Enmascaramiento.

En muchos trabajos se estudian estas Características en Robustez. Así, algunos de los más representativos son los contenidos en Cook y Weisberg (1980), Lopuhaa y Rousseeuw (1991), Yohai y Zamar (1993), y por supuesto en grandes trabajos como Hampel y otros (1987) y Barnett y Lewis (1994).

4.1.1 Eficiencia y Consistencia

En general, en Estadística Robusta el concepto de *Eficiencia* es considerado como sinónimo de Eficiencia Relativa, de manera que se estudia esta característica comparando la varianza del estimador en estudio y el mejor en cada caso. Como es conocido y se comenta en diversos trabajos (por ejemplo Martin y Zamar 1993), la Eficiencia es un concepto que en general se contrapone al de Robustez, de manera que estimadores poco Robustos son muy Eficientes, y en general estimadores Robustos son muy poco Eficientes. Por lo tanto, es necesario tomar una decisión de compromiso de manera que no sea dejada de lado totalmente ninguno de los dos conceptos. Así, por ejemplo, la media tiene la máxima Eficiencia en poblaciones simétricas, siendo, como veremos en esta sección, un Estimador de Posición muy poco Robusto, mientras que la Eficiencia de la mediana, con respecto a la media, es del 64%, con un comportamiento ante la Robustez mucho mas satisfactorio.

Por otra parte, el concepto de *Consistencia* tiene una importancia particular en los estudios de robustez, ya que, por la complejidad en la definición de ciertos estimadores, no es posible más que estudiar su comportamiento asintótico.

4.1.2 Equivarianzas

Se propone en la literatura sobre el tema que tratamos el concepto de *Equivarianzas* (Rousseeuw y Leroy 1987), el cual conduce a tres propiedades que pueden ser consideradas como propiedades analíticas. Así, veamos estas propiedades para los Estimadores de Posición, Escala y Regresión, familias de estimadores más utilizados en la literatura.

Para el caso de Estimadores de Posición o Escala, supongamos dada una muestra de la variable x, $Z_x = \{x_i\}_{i=1,2,...,n}$, y consideremos los siguientes conjuntos $Z_{x+s} = \{x_i + s\}_{i=1,2,...,n}$ y $Z_{rx} = \{rx_i\}_{i=1,2,...,n}$ con $r, s \in \mathbb{R}$ cualesquiera, entonces:

Si T_p es un Estimador de Posición, se dice que:

T_p es Equivariante de Posición $\Leftrightarrow T_p(Z_{x+s}) = T_p(Z_x) + s$

T_p es Equivariante de Escala $\Leftrightarrow T_p(Z_{rx}) = r\ T_p(Z_x)$

T_p es Equivariante Afín $\Leftrightarrow$ es Equivariante de Posición y Escala.

Si T_e es un Estimador de Escala, se dice que:

T_e es Equivariante de Posición $\Leftrightarrow T_e(Z_{x+s}) = T_e(Z_x)$

T_e es Equivariante de Escala $\Leftrightarrow T_e(Z_{rx}) = |r|\ T_e(Z_x)$

T_e es Equivariante Afín $\Leftrightarrow$ es Equivariante de Posición y Escala.

Para el caso de Estimadores de Regresión tenemos que, si T_r es un Estimador de Regresión y Z una muestra de las variables del modelo (**??**), donde en forma matricial $Z = (Y|X)$ con Y la matriz de la variable explicada y X la matriz de variables explicativas, y considerando los siguientes conjuntos: $ZY_a = (aY|X)$, $ZX_A = (Y|XA)$ y $ZYX_V = (Y + XV|X)$, donde $a \in \mathbb{R}$, y A y V son matrices de números reales con $|A| \neq 0$, entonces:

T_r es Equivariante de Escala $\Leftrightarrow T_r(ZY_a) = a\ T_r(Z)$

T_r es Equivariante Afín $\Leftrightarrow T_r(ZX_A) = A\ T_r(Z)$

T_r es Equivariante en Regresión $\Leftrightarrow T_r(ZYX_V) = T_r(Z) + V$

En definitiva, las propiedades de Equivarianzas tratan del comportamiento de los estimadores cuando se realizan cambios de posición y/o escala en los datos muestrales. Así, es fácil comprobar como los estimadores media, mediana y varianza son Equivariante Afín, y que *LS-Est.* es Equivariante de Escala, Afín y de Regresión, de manera que, el cumplimiento de estás propiedades nos asegura su comportamiento ante cambios de posición y escala en las variables, al tiempo que nos permiten operar analíticamente de una manera cómoda con estos estimadores.

4.1.3 Ajuste Exacto

Una propiedad propuesta por el autor Donoho (1982), que pretende recoger de una manera estricta la idea de Estimador Robusto, es la conocida por el nombre de *Propiedad de Ajuste Exacto*. Así, se dice que un estimador tiene la propiedad de Ajuste Exacto si, dada una muestra sobre las variables en estudio, de manera que para la "mayoría" de las observaciones de ésta el estimador proporciona un valor exacto, entonces la respuesta mediante el estimador en estudio nos debe proporcionar este valor. Esta propiedad es demasiado estricta, por lo que en general se utilizan diferentes niveles en su cumplimiento, los cuales están determinados por la interpretación de la palabra "mayoría". En el caso más extremo, podemos sustituir en la definición la palabra mayoría por "la mitad más una". Para los diferentes niveles en los que se puede cumplir esta propiedad, se utiliza el concepto de *Punto de Ajuste Exacto*, definido, para cada familia de estimadores, de la siguiente manera:

Para estimadores de Posición y Escala, consideremos que x es una variable de la cual conocemos una muestra de n elementos dada por $Z = \{x_i\}_{i=1,2,...,n}$, donde $x_i = x_0$ para $i = 1, 2, ..., n$, entonces tenemos que, siendo Zc_m una contaminación de Z en m observaciones:

Si T_p es un Estimador de Posición sobre la muestra Z, se define el Punto de Ajuste Exacto de T_p en Z, denotado por $\delta_n^*(T_p, Z)$, de la forma:

$$\delta_n^*(T_p, Z) = Min_m\{\frac{m}{n} \text{ tal que } \exists Zc_m \text{ con } T_p(Zc_m) \neq x_0\}$$

Si T_e es un Estimador de Escala sobre la muestra Z, se define el concepto de Punto de Ajuste Exacto de T_e en Z de la forma:

$$\delta_n^*(T_e, Z) = Min_m\{\frac{m}{n} \text{ tal que } \exists Zc_m \text{ con } T_e(Zc_m) \neq 0\}$$

Para Estimadores de Regresión, si Z una muestra de n elementos de las variables del modelo (**??**) de manera que las observaciones de Z siguen una relación lineal exacta de parámetro β (es decir, mediante las matrices de observaciones $Y = \beta X$), se define el Punto de Ajuste Exacto del Estimador de Regresión T_r en Z, siendo Zc_m una contaminación de Z en m observaciones, de la forma:

$$\delta_n^*(T_r, Z) = Min_m\{\frac{m}{n} \text{ tal que } \exists Zc_m \text{ con } T_r(Zc_m) \neq \beta\}$$

Notar que, conocida la estimación esperada (x_0, 0 o β, respectivamente) para el estimador T utilizado (T_p, T_e, T_r, respectivamente) sobre la muestra Z de tamaño n, entonces $\delta_n^*(T, Z)$ sería la proporción de observaciones que necesitamos contaminar en Z para que T nos proporcione un valor diferente del esperado. Por lo que, el Punto de Ajuste Exacto mide el porcentaje máximo de contaminación que permite el estimador, antes de dar como respuesta cualquier valor distinto del esperado.

En el caso de los estimadores que estamos utilizando como ejemplos, decir que, sobre una muestra de tamaño n, la media (*LS-Est.*, respectivamente) tiene Punto de Ajuste Exacto de valor $1/n$, el mínimo posible, es decir, suponiendo todas las observaciones iguales (un ajuste exacto entre las variables, respectivamente), tomando una contaminación en una única observación el estimador es diferente del esperado. En cambio, en el caso de la mediana, es necesario contaminar casi la mitad, exactamente $Int(\frac{n+1}{2})$, para conseguir que el estimador obtenido sea diferente del esperado, por lo que su Punto de Ajuste Exacto es de $Int(\frac{n+1}{2})/n$.

4.1.4 Función de Influencia y Curva de Sensibilidad

En Hampel (1968) se propone el concepto conocido por *Función de Influencia*, la cual pretende medir el efecto producido por una contaminación infinitesimal en el modelo. Veamos la definición de esta función.

Siendo z la variable del modelo en estudio, con Función de Distribución F, con Z una muestra de z y T un estimador de un parámetro del modelo, denotando por $T(F) = T(Z)$ al estimador obtenido con la muestra Z generada según F. Entonces, definimos la Función de Influencia de T para F en la observación z' (variable), denotada por $IF(z'; T, F)$, de la forma:

$$IF(z'; T, F) = \lim_{\epsilon \longrightarrow 0} \frac{T((1-\epsilon)F + \epsilon \triangle_{z'}) - T(F)}{\epsilon}$$

donde $\triangle_{z'}$ es una Medida de Probabilidad con peso 1 en z'. Notar que la estructura de esta última definición, se asemeja a la definición de derivada para la función T, de manera que, si la Función de Influencia es no acotada, se podría afirmar que una contaminación infinitesimal produce variaciones no controladas en el estimador, por lo que la estimación sería poco fiable.

Tukey (1970) relaja el concepto de Función de Influencia hasta definir una nueva

función la cual pretende medir la variación, en la respuesta del estimador, por el efecto de una contaminación puntual en la muestra sobre la que se aplica, llamándola *Curva de Sensibilidad*. Así, con las condiciones anteriores, se define el concepto de Curva de Sensibilidad en la observación z', denotándolo por $SC_n(z'; T, Z)$, de la forma:

$$SC_n(z'; T, Z) = (n+1)[T(Z \cup z') - T(Z)]$$

Igual que para la Función de Influencia, el hecho de que la Curva de Sensibilidad esté acotada, significa que la contaminación puntual no es capaz de producir un efecto de manera que la respuesta del estimador sea tan diferente como se quiera, con respecto del valor obtenido con la muestra sin Contaminar. Así, aparece el concepto de estimador *B-Robusto*, que se asocia a aquellos estimadores donde dichas funciones son acotadas, es decir, al producirse una contaminación la respuesta del estimador es previsible.

Como ejemplo, la Función de Influencia de la media es de la forma: $IF(x; T, F) = x$, y la Curva de Sensibilidad $SC_n(x; T, Z) = x - \bar{x}$. Ambas funciones de x son no acotadas, por lo que este estimador es no B-Robusto. Para la varianza y el *LS-Est.*, también se cumple que sus Funciones de Influencia son no acotadas, por lo que estos estimadores son no *B-Robustos*, y por lo tanto poco apropiados en presencia de Observaciones Atípicas, mientras que la mediana si es un Estimador de Posición *B-Robusto*.

Otros conceptos asociados a estas definiciones (recogidos en Hampel y otros (1986)), son los conocidos por el nombre de *Sensibilidad de Error-Grueso* y el de *Cambio-Local* (en inglés *"Gross-Error"* y *"Local-Shift"*), los cuales nos proporciona el valor máximo y la variación máxima de dichas funciones, respectivamente.

4.1.5 Punto de Ruptura

Otro concepto introducido por Hampel (1968), como generalización de una idea de Hodge (1967), es el que lleva el nombre de *Punto de Ruptura* de un estimador. La forma más utilizada de este concepto en los estudios en Robustez, es la propuesta por Donoho y Huber (1983) en la que se define como la fracción de contaminación en una muestra que hace que la respuesta del estimador produzca un valor arbitrario y no controlado. Para comprobar que la respuesta del estimador es no controlada, se utiliza un concepto, parecido al de *Sesgo* pero en muestras contaminadas, llamado *Bias*, el cual mide la máxima desviación causada en el estimador por la contaminación de la muestra.

Veamos la definición de estos conceptos.

Siendo T un Estimador, y Z una muestra de tamaño n de la variable o variables del modelo, y considerando Zc_m como una contaminación de Z en m observaciones, se define el Punto de Ruptura de T en Z, denotado por $\varepsilon_n(T, Z)$ por:

$$\varepsilon_n(T, Z) = Min_m\{\frac{m}{n}/Bias(m; T, Z) \text{ es infinito}\}$$

donde $Bias(m; T, Z)$ (que se debe leer como: *Bias* del Estimador T en Z para una contaminación de m observaciones) se define de la forma:

$$Bias(m; T, Z) = Sup_{Zc_m}\{\mid T(Z) - T(Zc_m) \mid\}$$

donde $\mid . \mid$ es la Norma Euclídea Usual.

Para el estudio del comportamiento asintótico de un estimador (es decir, para muestras muy grandes), se define el concepto de *Punto de Ruptura Asintótico* de T en Z, denotado por $\varepsilon^*(T, Z)$, de la forma:

$$\varepsilon^*(T, Z) = \lim_{n\to\infty} \varepsilon_n^*(T, Z)$$

En definitiva, el Punto de Ruptura nos informa del porcentaje de contaminación que permite un estimador antes de proporcionar como respuesta un valor no controlado. Así, si el Punto de Ruptura Asintótico del estimador es cercano a 1/2, el máximo, entonces obtendremos una estimación como respuesta de la información dada por el grueso de observaciones de la muestra.

Como ejemplo, el Punto de Ruptura de la media, la varianza y el estimador *LS-Est.* es 0, siendo el de la mediana de $Int(\frac{n+1}{2})/n$.

4.1.6 Efectos de Enmascaramiento

Las características conocidas por el nombre de *Efectos de Enmascaramiento* (en inglés *Efecto "Masking"* y *Efecto "Swamping"*), fueron propuestas por Tietjen y Moore (1972). Así, se conoce por Efecto *"Masking"* al hecho de que una observación pueda estar *"enmascarada"* por la presencia de un conjunto de elementos de la muestra, de manera que aunque esta observación fuera Atípica no sería detectada como tal. El otro posible efecto, el llamado Efecto *"Swamping"*, se produce cuando una observación perfectamente legítima en la muestra, es detectada como Observación Atípica por culpa, quizás, de

algún conjunto de observaciones de la muestra. Notar entonces que, estimadores que no estuvieran preparados para estos efectos producirían respuestas sesgadas y por lo tanto erróneas.

Los Efectos de Enmascaramiento son estudiados mediante ciertos test, presentados en trabajos como el de Bendre y Kale (1987), para Estimadores y Métodos de Detección de "Outliers", siendo en general relacionados con otras propiedades como la de Ajuste Exacto, y los conceptos de Curva de Sensibilidad y Punto de Ruptura.

4.2 Estimadores Robustos

La propuesta de nuevos estimadores definidos en pro de conseguir un buen comportamiento ante Observaciones Atípicas, es muy extensa. De estos, algunos han mejorado de manera satisfactoria el comportamiento ante las Características en Robustez de los primeros, aunque no se puede decir que exista uno que sea el mejor en todas las circunstancias. Por otra parte, su cálculo es casi siempre de gran dificultad, planteándose como soluciones de ecuaciones implícitas las cuales sólo se puede resolver mediante algoritmos de aproximación que necesitan la ayuda de sofisticados equipos informáticos.

Dada la gran cantidad de estos nuevos estimadores, realizar un estudio exhaustivo de todos y cada uno de ellos es una tarea que sobrepasa las posibilidades de este trabajo, por lo que nos centramos en la presentación de las familias más importantes, junto con algunas de sus principales propiedades y, naturalmente, unas buenas referencias donde poder completar esta información. Así, los estimadores más representativos y por otra parte más utilizadas en estudios de investigación, son los encuadrados dentro de las familias conocidas por los nombres de *M-Estimadores*, *R-Estimadores* y *L-Estimadores*, de los cuales vamos a presentar sus definiciones y citar sus elementos y características más relevantes, mientras que de otros, menos importantes, daremos cumplidas referencias.

4.2.1 M-Estimadores

La familia de Estimadores Robustos propuestos por Huber (1964), con el nombre de *M-Estimadores*, se construye como una generalización de los Estimadores de Mínimos Cuadrados. Así, los M-Estimadores, que denotaremos por *M-Est.*, se definen utilizando una función de los errores de estimación que dependen de un estimador (es decir, cada

e_i, para i=1,2,...,n, depende del estimador T utilizado), de la forma:

$$\textit{M-Est.} \equiv Min_T \sum_{i=1}^{n} \rho(e_i)$$

donde ρ es una función perteneciente a un conjunto de funciones, que denotaremos por *Rho-Fun.*, definido por:

$$\textit{Rho-Fun.} = \{\rho : \mathbb{R} \to \mathbb{R} / \rho \text{ es par, con un único mínimo en 0, y } \rho' \in \textit{Psi-Fun.}\}$$

siendo $\psi \in$ *Psi-Fun.*, donde el conjunto de funciones *Psi-Fun.* se definen como funciones de $\mathbb{R}$ en $\mathbb{R}$, de manera que, para todo $C \subseteq \mathbb{R}$ con Card(C) finito, se cumple:

- ψ es continua y derivable en $\mathbb{R} \backslash C$.

- Existen los límites a derecha e izquierda de ψ en todo C, ψ es impar para todo elemento de $\mathbb{R} \backslash C$, y $\psi(x) \geq 0$ para todo $x \in \mathbb{R} \backslash C$ y $x \geq 0$.

- Se cumple que:

  Card $\{x \in \mathbb{R} /\ \psi$ es continua, con ψ' no continua o no diferenciable$\}$ es finito.

- Siendo x una variable aleatoria con Función de Distribución F, entonces:

$$\int_{-\infty}^{\infty} \psi^2(x) dF < \infty \quad \text{y} \quad 0 < \int_{-\infty}^{\infty} \psi'(x) dF < \infty$$

Dadas estas definiciones, donde $\psi \in$ *Psi-Función* derivada de $\rho \in$ *Rho-Fun.* (es decir, $\psi = \rho'$), es posible definir el *M-Est.* asociado a ψ como solución de la ecuación implícita:

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

Existe una gran variedad de *M-Est.*, dependiendo de las funciones *Rho-Fun.* o *Psi-Fun.* utilizadas. Un ejemplo es el definido por Huber (1964), que denotaremos por *HU-Est.*, cuya *Psi-Fun.* está dada por:

$$\psi_{HU}(t; a) = Min\{a, Max\{t, -a\}\}$$

donde $t \in \mathbb{R}$ y $a \in \mathbb{R}_+$ es una constante.

Una subfamilia de *Psi-Fun.* son las llamadas funciones *Redescending*, que denotaremos por *Reg-Fun.*, definidas de la forma:

$$\textit{Reg-Fun.} = \{\psi \in \textit{Psi-Fun.} / \psi(x) = 0 \ \forall |x| \geq r\}$$

A esta familia pertenece la función mediante la cual Huber (1964) propone otro *M-Est.*, de la forma:

$$\bar{\psi}_{HU}(t;c) = \begin{cases} t & \text{si } \mid t \mid < c \\ 0 & \text{si } \mid t \mid \geq c \end{cases}$$

donde $t \in \mathbb{R}$ y se cumple que $\psi \in$ *Psi-Fun.*.

Otro *M-Est.* propuesto mediante una *Reg-Fun.* es el definido por Hampel (1974), que denotaremos por *HA-Est.*, donde su *Psi-Fun.* es:

$$\psi_{HA}(t;a,b,c) = \begin{cases} t & \text{si } 0 \leq \mid t \mid < a \\ aSign(t) & \text{si a} \leq \mid t \mid \leq b \\ a\frac{c-|t|}{c-b}Sign(t) & \text{si b} \leq \mid t \mid \leq c \\ 0 & \text{si } \mid t \mid \geq c \end{cases}$$

donde *Sign(t)* es el signo de t, y la función se define para todo $t \in \mathbb{R}$ y valores $a, b, c \in \mathbb{R}$, con $a \leq b \leq c$, donde es fácil comprobar que $\psi_{HA} \in$ *Psi-Fun.*.

El principal avance que se produce con los M-Estimadores (en el sentido de tener las mejores Características en Robustez posible, y por supuesto, siempre dependiendo de la definición o forma de las Rho-Funciones utilizadas) es el controlar o eliminar el efecto de los valores extremos en los residuos, de manera que el ajuste se realiza sobre los valores centrales de éstos (ver Rousseeuw y Leroy 1987).

Una generalización de los M-Estimadores fue dada por Mallow (1975) con el nombre de *GM-Estimadores*, denotados por *GM-Est.*. Estos estimadores utilizan, además de una función dependiente de los errores de estimación, una *Función de Ponderaciones* sobre las observaciones de la muestra, es decir, una función que dé más pesos a los errores de ciertas observaciones.

Así, siendo w una Función de Ponderaciones, se definen el *GM-Est.* con ρ su *Rho-Función* de la forma:

$$GM\text{-}Est. \equiv Min_T \sum_{i=1}^{n} w(e_i)\rho(e_i)$$

donde, siguiendo a Maddala y Rao (1997), la definición de estos estimadores, incluyendo en su construcción una Función de Ponderaciones, mejora en general el comportamiento de éstos respecto de los M-Estimadores.

4.2.2 R-Estimadores

Otra gran familia de estimadores con buenas propiedades en Robustez son los llamados *R-Estimadores*, los cuales se definen como aquellos estimadores que minimizan una expresión en la que interviene una *Función de Rango* de los errores de estimación, que se denota habitualmente por $Rang(.)$, la cual, para un conjunto de valores $\{t_i\}_{i=1,2,...n}$ en $\mathbb{R}$, si t_k es el valor del elemento de este conjunto en la posición r-ésima al ordenarlo de menor a mayor, entonces $Rang(t_k) = r$.

Como ejemplo, uno de los R-Estimadores más conocidos es el propuestos por Adichie (1967), que denotaremos por *A-Est.*, el cual se puede definir de la forma:

$$A\text{-}Est. \equiv Min_T \sum_{i=1}^{n} w(Rang(e_i))e_i$$

donde $\sum_{i=1}^{n} w(Rang(e_i)) = 0$, con w una Función de Ponderaciones monótona.

Se puede encontrar más información sobre los *R-Estimadores* en Maddala y Rao (1997).

4.2.3 L-Estimadores

Los elementos incluidos en la familia de estimadores llamada *L-Estimadores*, utilizan en su definición una función lineal en la que interviene una Función de Rangos de los errores de estimación (Bickel 1973).

Dentro de esta familia, encontramos el primer estimador con Punto de Ruptura máximo, introducido por A.F.Siegel, con el nombre *Medianas Repetidas*, el definido por H.Oja y A.Miinimaa, y el dado por los autores W.A.Stahel y D.L.Donoho con alto Punto de Ruptura (Maronna y Yohai 1995).

Otros estimadores de esta familia de los más difundidos, gracias quizás a su implementación en un paquete informático llamado *PROGRESS*, son los definidos y estudiados por Rousseeuw y Leroy (1987). Así, el primero que vamos a presentar es el de nombre *"Least Median Squares"* (*Mínima Mediana al Cuadrado*) y denotado habitualmente por *LMS-Est.*. Este nace de la idea de sustituir en Mínimos Cuadrados el operador media por uno más Robusto, como es la mediana. Así, el *LMS-Est.* se define de la forma:

$$LMS\text{-}Est \equiv Min_T\{Med_i(\{e_i^2\})\}$$

Este estimador es un caso particular de otro, también de Rousseeuw, con nombre *"Least Quantility Squares"*, (Mínimos Cuantiles al Cuadrado), denotado por *LQS-Est.* y que podemos definir de la forma:

$$LQS\text{-}Est. \equiv Min_T\{(e^2)_{[s]}\}$$

siendo s=[Int((1-α)n)+Int($\alpha(p+1)$)], con $\alpha \in [0, 1/2]$, demostrándose que si $\alpha = 1/2$ entonces *LQS-Est.* es asintóticamente equivalente a *LMS-Est.*.

El último estimador presentado por Rousseeuw es el conocido por el nombre de *"Least Trimmed Squares"* (Mínimos Truncados al Cuadrado), denotado por *LTS-Est.*. Este aparece como mezcla entre *LS-Est.* y *LMS-Est.* con el objeto de perfeccionar a éste último. Así, con la notación anterior, el estimador *LTS-Est.* minimiza la suma, no de todos los errores de estimación al cuadrado, sino sólo de los h más pequeños, para la elección de un $h \in \{1, 2, \ldots, n\}$, es decir:

$$LTS\text{-}Est. \equiv Min_T\{\sum_{i=1}^{h}(e^2)_{[i]}\}$$

donde, si $h = n$, entonces coincide con el *LS-Est.*.

Notar que estos tres últimos estimadores citados, se definen eligiendo un estimador T de manera que se minimice una cierta función de los errores de estimación, dependientes de ese estimador.

En la misma publicación de Rousseeuw y Leroy, se demuestran las buenas propiedades de los Estimadores *LMS-Est.*, *LQS-Est.* y *LTS-Est.*. Así, se comprueba que estos estimadores cumplen la propiedad de Ajuste Exacto en su máximo nivel, que son Equivariantes Afín, y con Punto de Ruptura Asintótico de *LMS-Est.* y *LTS-Est.* el máximo, 1/2.

4.2.4 Otros Estimadores

La lista de estimadores propuestos en la literatura es, como ya hemos comentado, muy grande. En esta subsección citaremos los más destacados de los no encuadrados en las familias anteriores.

Un método diferente a los comentados hasta ahora, lo encontramos en la utilización de procedimientos mediante los cuales se minimice el volumen generado por una cierta figura geométrica que contenga a la mayoría, o a un número dado como mínimo, de

observaciones de la muestra considerada. Un Estimador de Regresión que utiliza este procedimiento es el introducido por Rousseeuw y Leroy (1987), con el nombre *Mínimo Volumen Elipsoidal*. Este estimador, tiene buenas propiedades, siendo la más destacada su Punto de Ruptura, la cual, sobre el modelo (**??**), es de $(\text{Int}(n/2) - p + 1)/n$, y por lo tanto su Punto de Ruptura Asintótico es de 1/2.

Otra práctica habitual recientemente, a la hora de proponer nuevos estimadores Robustos, consiste en utilizar combinaciones de los ya definidos, con el fin de explotar las buenas propiedades que tienen éstos intentando mejorarlas en lo posible. Como muestra de los avances en esta línea podemos citar las familias de los *S-Estimadores* (Rousseeuw y Yohai 1984) y los *τ-Estimadores* (Yohai y Zamar 1986). Los S-Estimadores se definen mediante un M-Estimador, de manera que sus propiedades dependen de éste último. Así, estos autores consiguen, para Estimadores de Regresión sobre el modelo (**??**) y determinadas Rho-Funciones, que el Punto de Ajuste Exacto sea (n-Int(n/2)+p-1)/n, sea Equivariantes Afín, con Punto de Ruptura (Int(n/2)-p+2)/n, por lo que tienen el mejor Punto de Ruptura Asintótico posible, 1/2. Para el caso de los τ-Estimadores, estos tienen Características en Robustez semejantes a la de los S-Estimadores, aunque tanto el Punto de Ruptura como en Eficiencia Asintótica salen mejor parados que éstos.

Otros estimadores que podemos encontrar en la literatura son, por ejemplo; el introducido por J.L.Hodges y E.L.Lehmann, denotado por *HL-Est.*; algunos Estimadores propuestos a partir de los M-Estimadores, como los conocidos con las notaciones de *On-Step-M-Est.*, *MM-Est.*, *M1S-Est.* y *S1S-Est.* estudiados en Rousseeuw y Leroy (1987); el comentado por Bustos (1988), denotado por *D-Est.*, etc.. En Zamar (1989) encontramos referencias y estudios de ampliación interesantes sobre estos estimadores nombrados. Más recientemente, podemos citar; los estudiados en Peña (1992) con el nombre de *Estimadores Contraidos* y *Estimadores Ortogonales*; el nombrado en Lopuhaa (1992); los Estimadores basados en Proyecciones, desarrollados en Zamar (1992) y también en Maronna y otros (1992), y Maronna y Yohai (1993); etc.. Como referencias y estudios de ampliación podemos nombrar el trabajo de Velilla (1995).

4.3 Métodos de Detección de "Outliers"

Como se comentó en la introducción del trabajo, una de las vías de estudio de la presencia de Observaciones Atípicas, junto con la utilización de Estimadores Robustos, es el uso

de unos procedimientos conocidos por el nombre de *Métodos de Detección de "Outliers"*. Estos se basan en ciertas medidas que nos permiten averiguar si, en una muestra dada para la estimación de unos determinados parámetros de un modelo, una observación es Atípica, o si por el contrario su presencia no produce ninguna sorpresa, siendo admisible en la muestra del fenómeno que queremos analizar.

Podemos considerar diferentes procedimientos genéricos para la detección de Observaciones Atípicas, según qué elementos utilicen para su definición. Así, podemos considerar dos familias de medidas diferentes, que llamaremos *Medidas de Relación* y *Medidas de Influencia*, donde; una Medida de Relación la podemos definir intuitivamente como un índice, o un valor numérico, asociado a una o a un conjunto de observaciones de la muestra que deja al descubierto la similitud entre la mayoría de observaciones de la muestra y ésta; mientras que, por otra parte, una Medida de Influencia la podemos definir, también intuitivamente, como un índice asociado a una, o a un conjunto de observaciones de la muestra, que nos proporciona información sobre el impacto producido por ésta en la estimación de los parámetros del modelo en estudio mediante un determinado estimador. Por lo tanto, en la definición de esta última medida en el ámbito de nuestro estudio, debe aparecer el estimador utilizado o algún estadístico relacionado con él.

Desde otro punto de vista, es conveniente estudiar el impacto producido por un conjunto de más de una observación en la estimación de los parámetros del modelo, por ser posible que una observación por sí sola no tenga una influencia destacada en la estimación mientras que si la consideramos como parte de un grupo de observaciones, la influencia de ésta sí sea importante. Así, al conjunto de medidas que nos proporcionan información sobre el impacto producido por un conjunto de más de una observación, se conocen por *Medidas para Conjuntos de Observaciones*. Dentro de este conjunto, también podemos distinguir Medidas de Relación y Medidas de Influencia, dependiendo de su definición. En general, estas medidas son construidas redefiniendo una medida de las anteriores de manera que no considere la presencia de más de una observación, es decir, elimine un conjunto de elementos de la muestra (Besley y otros 1980).

Para el caso particular de Modelos de Regresión, en general se supone que todos los coeficientes de regresión tienen la misma importancia en la determinación del modelo, pero es posible que una observación, o conjunto de observaciones, pueda influir mucho

en una determinada variable, y en cambio no sea nada influyente en otra, por lo que su presencia puede pasar desapercibida. Para resolver este problema se define una familia de medidas que determinan la influencia de cada observación, o conjunto de observaciones, en todos y cada uno de los coeficientes de regresión, conocidas con el nombre de *Medidas de Influencia Parcial*. Estudios sobre estas medidas se pueden encontrar en trabajos como los de , Cook y Wang (1983), Polasek (1984), Rousseeuw (1984), Rousseeuw y Van-Zomeren (1990), Hadi (1992), Davies y Gather (1993), Dawkins (1995), y Rocke y Woodruff (1996).

Revisemos a continuación las llamas Medidas de Relación y Medidas de Influencia, por ser estas las más utilizadas en la práctica.

4.3.1 Medidas de Relación

Un primer tipo de Medida de Relación que se encuentra en la literatura, son aquéllas que utilizan en su definición alguna distancia de las observaciones de la muestra a un determinado representante o centroide de ella, corregida por un determinado elemento. Así, siguiendo a Cuadras (1989), en general es posible definir medidas de la forma $d(a,b) = (a-b)M(a-b)$, donde $a \in \mathbb{R}^n$ y $M \in \mathrm{Mat}_n(\mathbb{R})$ es una matriz Simétrica y Definida Positiva, las cuales sirven para estas definiciones.

Un caso particular de éstas, es la que utiliza para su definición la distancia conocida por *Distancia de Mahalanobis*, mediante la cual es posible determinar la relación entre los elementos de una muestra, con la ayuda de un representante de ésta, calculado mediante un Estimador de Posición, y de la dispersión de los elementos calculado mediante un Estimador de Escala (Rousseew y Leroy 1987).

Así, sobre un modelo como (**??**) con Z una muestra, la Distancia de Mahalanobis (DM^2) queda definida (Peña 1997) por:

$$DM^2(z_i) = (z_i - \mu)^t \Sigma^{-1} (z_i - \mu)$$

donde $z_i \in Z$, y siendo μ y Σ la matriz columna de medias y la matriz de covarianzas, respectivamente. Así, valores altos de $MD^2(z_i)$ nos indica poca relación entre la observación z_i y el resto de elementos de la muestra Z, donde, suponiendo la Normalidad en las variables del modelo, se demuestra que la variable $MD^2(z)$ sigue una Distribución Chi-Cuadrado Centrada con p grados de libertad (χ^2_{p+1}), de manera que se consideran

valores altos para esta medida aquéllos que son mayores a $\chi^2_{p+1}(0.975)$ (Penny 1996).

Otra Medida de Relación relevante, es la dada por los autores Cook y Weisberg (1980) con el nombre de *Ratio de Volumen Elipsoidal*, en cuya definición se utiliza el volumen del elipsoide que contiene a $(1-\alpha)100\%$ de las observaciones de la muestra donde $\alpha \in [0,1]$, una vez eliminada la observación i-ésima, representando por α la proporción de observaciones supuestamente anómalas en la muestra (Chatterjee y Hadi 1986).

Por último, podemos citar la medida conocida por el nombre de *Diagnosis de Resistencia*, dada por Rousseew y Leroy (1987), donde se consideran los residuos de estimación para submuestras de tamaño p de la muestra inicial. Así, sobre la muestra Z de las variables del modelo (**??**), se considera cualquier submuestras $J \subseteq Z$ de tamaño muestral p, denotamos por $\hat{\beta}_J$ el único parámetro de regresión para la submuestra J, y $(e_i)_{\hat{\beta}_J}$ al residuo de la observación i-ésima de la muestra para dicho parámetro, definiendo los valores:

$$u_i = Max_J\{\frac{\mid (e_i)_{\hat{\beta}_J} \mid}{Med_j(\{(e_j)_{\hat{\beta}_J}\})}\}$$

se define el valor Diagnosis de Resistencia para la observación i-ésima, denotándolo por RD_i, por:

$$RD_i = \frac{u_i}{Med_j(\{u_j\})}$$

Esta medida es estudiada a fondo en el citado trabajo de Rousseeuw y Leroy, en el que se concluye que para el 50% de las observaciones estas medidas son menores que 1, y que las observaciones con valores mayores de 2.5 o 3 pueden ser consideradas como Observaciones Atípicas en la muestra y modelo en estudio.

4.3.2 Medidas de Influencia

En el caso de las Medidas de Influencia, la más difundida es la asociada a los estimadores de Mínimos Cuadráticos, mediante la conocida *Matriz Hat*, donde, cada elemento de la diagonal principal de dicha matriz nos proporciona información sobre la influencia de la correspondiente observación en la construcción del estimador Mínimo Cuadrático (Rousseeuw y Leroy 1987). Así, dada la muestra Z, y bajos las condiciones de un modelo de regresión lineal (**??**), con *LS-Est.* el estimador Mínimo Cuadrático de los parámetros del modelo, entonces se define la Matriz Hat, denotada por $\hat{H}$, como una

matriz Cuadrada de orden n, Simétrica e Idempotente ($\hat{H}\hat{H} = \hat{H}$), de la forma:

$$\hat{H} = X(X^tX)^{-1}X^t$$

donde, siendo $\hat{H} = (h_{ij})_{i,j=1,2,\ldots,n}$, cada elemento de la diagonal principal de la Matriz Hat (h_{ii} para i=1,2,...,n) nos proporciona información sobre la influencia de la correspondiente observación en la construcción del estimador *LS-Est.*, ya que, como $\hat{Y} = \hat{H}Y$, si h_{ii} toma un valor cercano a 1, la estimación de la observación i-ésima de la variable y ($\hat{y}$), está muy cercana de la observación i-ésima de y, por lo que entonces la superficie de regresión debe pasar cerca de esta observación, o lo que es lo mismo, la observación z_i de Z influye mucho en la determinación de *LS-Est.*. Para determinar la cercanía de h_{ii} a 1, en el trabajo Rousseeuw y Leroy se compara este valor con $\frac{p+1}{n}$, de manera que, si h_{ii} toma un valor mayor o igual a éste entonces se considera a la observación z_i como Observación Influyente, y candidata a Observación Atípica.

Otras formas comunes de analizar si la estimación de la variable explicada es correcta o no, es; estudiando las diferencias entre los valores reales de las observaciones dados por la muestra y los valores dados por la estimación del Modelo; comparando el estimador resultante de utilizar todas las observaciones de la muestra para su cálculo y el obtenido mediante la eliminación de la observación en estudio de la muestra; o, mediante el uso de distancias entre las Funciones de Verosimilitud de los estimadores con todas, y todas menos una de las observaciones de la muestra. Como muestra de los dos primeros grupos de medidas, en Catterjee y Hadi (1986) se propone una primera, denotada por t_i para la observación i-ésima de la muestra en consideración, en la que se utiliza los errores de estimación (e_i), un corrector, dependiente de la varianza residual corregida ($\hat{S}_R$) y los elementos de la diagonal principal de la Matriz Hat (h_{ii}), de la forma:

$$t_i = \frac{e_i}{\hat{S}_R\sqrt{1-h_{ii}}}$$

y otra, incluida en el segundo conjunto citado, denotada por t_i^* para la observación i-ésima de la muestra en consideración, en la que, siendo $(\hat{S}_R)_{(i)}$ la varianza residual corregida eliminando la observación i-ésima, se define de la forma:

$$t_i^* = \frac{e_i}{(\hat{S}_R)_{(i)}\sqrt{1-h_{ii}}}$$

El comportamiento tanto para t_i como para t_i^* es, bajo el supuesto de Normalidad, el de una Normal Estandarizada, por lo que para valores de estas medidas superiores a 2.5

o 3 son consideradas Observaciones Influyentes, y por lo tanto posibles Observaciones Atípicas.

Otras referencias sobre el tema se pueden encontrar en Welsch y Kuh (1977), Cook y Weisberg (1982), Belsley y otros (1980), y más recientemente en Cook (1996).

5 Comentarios Finales

La problemática que se presenta por la aparición de Observaciones Atípicas, en muestras que son utilizadas para la estimación de parámetros de un modelo, es un problema de enorme dificultad, para el cual se han propuesto diversos procedimientos, muchos de ellos ingeniosos a la vez que rigurosos matemática y estadísticamente, que lo han resuelto en algunos casos, y en otros han ayudado a su comprensión. Así, debemos decir que, aunque todas los procedimientos expuestos cumplen un papel importante dentro del estudio del impacto de las definidas Observaciones Atípicas, lo cierto es que la respuesta de éstos no es siempre la misma, de manera que no se puede decir que uno de ellos en particular pueda ser elegido como el ideal. Así, siguiendo a Andrews y otros (1972), la mejor opción es utilizar un conjunto de ellas de manera que nos proporcionen diferentes puntos de vista sobre el problema a tratar.

La dificultad antes comentada reside en que, si bien en problemas de baja dimensión, la detección de estas posibles observaciones anómalas mediante los procedimientos robustos, suponiendo conocido el modelo, se ven reforzados mediante la construcción de un gráfico de las observaciones de la muestra, en problemas de alta dimensión, donde no es posible realizar ningún tipo de gráfico de las observaciones que nos permitan detectar las indeseables, los procedimientos robustos proporcionan respuestas que debemos considerar como válidas, según la teoría, pero que no podemos tener la certeza de que lo sean.

Los procedimientos estadísticos propuestos en Estadística Robusta todavía no son de gran difusión, por lo que en algunos casos donde su uso sería determinante, obteniéndose unos resultados muy satisfactorios, no son aplicados. Por ello, en este papel se ha realizado un repaso a los principales conceptos y métodos en el ámbito de la Estadística Robusta, centrado en aquellos procedimientos generales más extendidos, con algunos particulares a modo de ejemplo, con objeto de mostrar el estado actual del tema y

proporcionar referencias sobre sus elementos más importantes, con el objeto de que su conocimiento repercuta en su utilización en cualquier estudio estadístico.

Bibliografía

ADICHIE, J.N. (1967): "Estimates of regression coefficients based on rank tests". Annals of Mathematica Statistics, 38.

ALLEN, C.G. (1961): ver BERNOULLI, D. (1977).

ANDREWS, D.F. y otros (1972): *Robust Estimates of Location: Survey and Advances*. Princeton University Press, Princeton, N.J.

BARNETT, V. y LEWIS, T. (1994): *Outliers in Statistical Data*. 3st. Ed. Wiley and Sons.

BELSLEY, D.A., KUH, E. y WELSCH, R.E. (1980): *Regression Diagnostics: Identifying influential data and sources of collinearity*. Ed. Wiley and Sons. N.Y.

BENDRE, S.M. y KALE, B.K. (1987): "Masking effect on tests for outliers in normal samples". Biometrika, 74.

BERNOULLI, D. (1977): "Dijudicatio maxime probabilis plurium observationum discrepantium atque verisimillima inductio inde formanda". Acta Academiae Scientiorum Petropolitannae,1. Traducción de C.G. Allen (1961), Biometrika, 48.

BICKEL, P.J. (1973): "On some analogues to linear combination of order statistics in the linear model". Annals of Statistics. Vol.1, No.4.

BUSTOS, O. (1988): "Outliers y robustez". Documento de trabajo del Conselho Nacional de Desenvolvimiento Cientifico e Tecnologico. BRASIL.

CHATTERJEE, S. y HADI, A.L. (1986): "Influential observation, high leverage points and outliers in linear regression. Comment". Statistical Science, Vol.1, No.3.

COOK, R.D. y WANG, P.C. (1983): "Transformations and influencial cases in regression". Technometrics, No.25.

COOK, R.D. y WEISBERG, S. (1980): "Charecterization of an empirical Influence Function for detecting influential cases in regression". Technometrics, No.22.

COOK, R.D. y WEISBERG, S. (1982): *Residuals and Influence in Regression*. Ed. Chapman and Hall, N.Y.

CUADRAS, C.M. (1989): "Distancias estadísticas (con discusión)". Estadística Española, Vol.30, No.119.

DAVIES, P.L. y GATHER, U. (1993): "The identification of multiple outliers". Technometrics, 35.

DAWKINS, B.P. (1995): "Investigating the geometry of a p-Dimensional data set". J. American Statistical Association, No.429.

DONOHO, D.L. (1982): "Breakdown properties of multivariate location estimators". Cualifying Paper, Harvart Univ., Boston, M.A.

DONOHO, D.L. y HUBER, P.J. (1983). "The notion of breakdown point" en Ed. Bickel, P., Doksum, K. y Hodges, J.L.Jr.: *A Festschrift for Erich Lehmann*.

HADI, A.S. (1992): "Identifying multiple outliers in multivariate data". J. Royal Statistical Society (Serie-B), Vol.54, No.3.

HAMPEL, F.R. (1974): "The Influence Curve and its role in robust estimation". J. of the American Statistical Association, 69.

HAMPEL, F.R. y otros (1986): *Robust Statistics: The approach based on influence functions*. Ed. Wiley and Sons. N.Y.

HOAGLIN, D.C., MOSTELLER, F. y TUKEY, J.W. (1983): *Understanding Robust and Exploratoty Data Analysis*. Ed. Wiley and Sons.

HODGE, J.L. (1967): "Efficiency in normal samples and tolerance of extreme values for some estimates of location". Univ. California, Berkeley.

HUBER, P.J. (1964): "Robust estimation of location parameter". Annals of Mathematical Statistics, 35.

LOPUHAA, H.P. (1992): "Highly efficient estimators of multivariate location with high breakdown point". Annals of Statistics, Vol.20, No.1.

LOPUHAA, H.P. y ROUSSEEUW, P.J. (1991): "Breakdown points of affine equivariant estimators of multivariate location and covariance matrices". Annals of Statistics, Vol.19, No.1.

MADDALA, G.S. y RAO, C.R. (1997): *Robust Inference*. Handbook of Statistics 15. North-Holland. (Elsevier Science B.V.).

MALLOW, C.L. (1975): *On Some Topics in Robustness*. Murray Hill, N.J.

MARONNA, R.A. y YOHAI, V.J. (1993): "Bias-Robust estimates of regression based on projections". Annals of Statistics, Vol.21, No.2.

MARONNA, R.A., STAHEL, W.A. y YOHAI, V.J. (1992): "Bias robust estimators of multivariate scatter based on projections". J. Multivariate Annals, No.42.

MARTIN, R.D. y ZAMAR, R. (1993): "Efficiency-Constrained Bias-Robust estimation of location". Annals of Statistics, Vol.21, No.1.

ORTEGA, J.Fco. (2000): *Nuevas Familias de Estimadores Robustos y Detección de Observaciones Atípicas en Modelos Lineales*. Tesis Doctoral. Univ. de Castilla-La Mancha.

PEÑA, D. (1992): *Estadística: Modelos y Métodos*. Ed. Alianza Universidad Textos.

PENNY, K.I. (1996): "Appropriate critical values when testing for a single multivariate outliers by using the Mahalanobis distance". J. Royal Statistical Society (Serie-C), Vol.45.

POLASEK, W. (1984): "Regression diagnostics for general linear regression models". J. American Statistical Association, No.386.

ROUSSEEUW, P.J. (1984): "Least median of squares regression". J. American Statistical Association, No.388.

ROUSSEEUW, P.J. y LEROY, A.M. (1987): *Robust Regression and Outlier Detection*. Ed. Wiley and Sons.

ROUSSEEUW, P.J. y VAN-ZOMEREN, B. (1990): "Unmasking multivariate outliers and leverange points". J. American Statistical Association, No.411.

ROUSSEEUW, P.J. y YOHAI, V. (1984): "Robust regression by means of S-Estimators". Lectures Notes Statistics, 26.

STIGLER, S.M. (1973): "Simon Newcomb, Percy Daniell and the history of robust estimation 1885-1920". J. American Statistical Association, No.344.

TIETJEN, G.L. y MOORE, R.H. (1972): "Some Grubbs-type statistics for the detection of several outliers". Technometrics, 14.

TUKEY, J.W. (1977): *Exploratory Data Analysis*. Ed. Addison-Welsley P.C.

VELILLA, S. (1995): "Diagnostics and robust estimation in multivariate data transformations". J. American Statistical Association, No.431.

YOHAI, V. y ZAMAR, R. (1986): "High breakdown estimates of regression by means of the minimization of an efficient scale". J. American Statistical Association, No.83.

YOHAI, V. y ZAMAR, R. (1993): "A MiniMax-Bias property of the least α-Quantile estimates". Annals of Statistics, Vol.21, No.4.

ZAMAR, R. (1992): "Bias robust estimation in orthogonal regression". Annals of Statistics, Vol.20, No.4.