29
7

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

OBSERVACIONES DISCREPANTES EN REGRESION

T E S I S

Que para obtener el Título de

A C T U A R I O

p r e s e n t a

Jesús Bravo Segovia

México, D. F.

1 9 8 6

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

CONTENIDO.

PROLOGO

El problema de observaciones discrepantes es uno de los mas viejos en estadística. Durante los últimos 150 años, el interés en estas observaciones ha tenido sus altas y sus bajas, algunas veces es área activa de investigación para que años mas tarde sufra de un abandono relativo.

En años recientes se han resuelto un gran número de problemas en la teoría de observaciones discrepantes y se han descubierto otros. Sin embargo, la mayoría de los resultados se encuentran dispersos en revistas especializadas.

El objetivo de este trabajo es presentar una pequeña revisión de la filosofía, teoría y métodos inherentes a tres enfoques de la teoría estadística: Clásico o frecuentista, Bayesiano y Robusto.

Esta tésis se presenta en 5 capítulos. En el primer capítulo, se analizan y proponen algunas definiciones del término "observación discrepante", tambien se determinan las causas por las cuales éstas se pueden presentar y las razones de su estudio. Por otro lado y dependiendo del interés sobre dichas observaciones, se clasifican a los métodos para el manejo de éstas. Por último, en este capítulo se presenta una reseña histórica de las observaciones discrepantes.

En el segundo capítulo se efectúa el estudio del enfoque frecuentista. En éste, se asientan muchos elementos necesarios (modelos, elementos de diagnóstico, etc) para la mayoría de los métodos. La parte medular de este capítulo, la sección

2.3 referente a los métodos para el manejo de observaciones discrepantes. La intención de presentar en ese orden los métodos es la de diferenciar cuales de estos sirven para una observación discrepante ($K=1$) y cuales para mas de una ($K>1$) haciendo énfasis en si K es conocida o desconocida. Finalmente, en las últimas secciones se efectúan comentarios generales de los métodos, se aplican algunos de éstos a dos problemas "clásicos" y por último se presentan conclusiones referentes a dichos métodos.

For lo que respecta a los capítulos 3, 4 y 5, podríamos decir que estructuralmente son iguales al capítulo 2, es decir, inicialmente se presenta la filosofía general del enfoque para que en secciones posteriores se presenten los métodos, comentarios, ejemplos y conclusiones. Naturalmente la diferencia radica en el aspecto filosófico con que cada enfoque trata el tema de observaciones discrepantes.

For último, al final de este trabajo se presentan una serie de conclusiones cuya intención es dilucidar que enfoque usar y cuál de los métodos presentados podría ser el adecuado en una situación específica.

NOTACION

En este trabajo el modelo de regresión lineal se manejará regularmente en su forma matricial:

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$$

donde $\underline{Y} = (y_1, y_2, \ldots, y_n)$ es el vector de observaciones de orden $n \times 1$. X es la matriz de dimensiones $n \times p$ y rango $r(X) = p$. cuyos renglones son de la forma $(x_{i1}, x_{i2}, \ldots, x_{ip})$ $i = 1, 2, \ldots, n$ con $x_{i1} = 1$, $i = 1, 2, \ldots, n$ si la recta, plano o hiperplano (en general) de regresión no pasa por el origen. $\underline{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ es un vector de orden $p \times 1$ de parámetros desconocidos. $\underline{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ es un vector aleatorio de $n \times 1$ y además se supone $E(\varepsilon) = 0$; $\text{Var}(\varepsilon) = E[(\varepsilon - E(\varepsilon))(\varepsilon - E(\varepsilon))^T] = \sigma^2 I$. I la matriz identidad de dimensión n y adicionalmente $\varepsilon \sim N(0, \sigma^2 I)$

Se hará referencia a este modelo como el "modelo de regresión lineal con las hipótesis usuales".

Al estimador de $\underline{\beta}$ y $\underline{\varepsilon}$ se les denotará por $\hat{\underline{\beta}}$ y $\underline{e}$ respectivamente, y al de σ^2 por S^2, donde:

$$\hat{\underline{\beta}} = (X^T X)^{-1} X^T \underline{Y},$$
$$\underline{e} = Y - \hat{Y} = (I - X(X^T X)^{-1} X^T), \ \underline{Y} = (I - V)\underline{Y} = M\underline{Y}; \ V \text{ y } M \text{ matrices de proyección.}$$
$$S^2 = \frac{\underline{Y}^T M \underline{Y}}{n - p},$$

K = número de observaciones discrepantes, $i = \{i_1, i_2, \ldots, i_K / 1 \leq i_1 < i_2 < \ldots < i_K \leq n\}$. En este trabajo en ocasiones i, estará como subíndice de alguno de los elementos del modelo de regresión definido anteriormente, ésto puede ser de dos formas: i unicamente como subíndice, en el caso de

X_i indicará a los renglones que están señalados por los elementos de i, es decir X_i representa a la matriz de $K \times n$ cuyos renglones están indicados por los elementos de i, en el caso de vectores, por ejemplo Y_i, representa al vector de $K \times 1$ observaciones cuyas componentes son las K observaciones sospechosas de ser discrepantes, interpretaciones similares tendrán los vectores $\underline{e}_i$, $\underline{\delta}_i$. En el caso que el conjunto indicador este entre paréntesis, significará que se eliminan a los elementos, renglones o componentes según sea el caso, que están indicados por los elementos de i. Por ejemplo $X_{(i)}$ representa a la matriz $(n-K) \times n$ que resulta de eliminar a los renglones indicados por i; $\underline{Y}_{(i)}$ representa al vector de $(n-K) \times 1$ de observaciones resultantes de eliminar a los indicados por los elementos de i, interpretaciones similares tendrán $\underline{e}_{(i)}$, $\underline{\delta}_{(i)}$.

Capítulo 1.

INTRODUCCION.

1.1. ¿Que es una observacion discrepante?

El concepto de observacion discrepante ha interesado a los investigadores desde los primeros intentos de la interpretación de datos. Aun antes del desarrollo del método estadístico había argumentos vehementes sobre si se debian rechazar observaciones de un conjunto de datos, argumentando que no son representativos. Las actitudes variaban de un extremo a otro: Desde el punto de vista que nunca debemos manchar la "santidad" de los datos con sospechas para desechar sus propiedades, hasta un pragmatismo "si dudas, rechazalo".

Hoy contamos con puntos de vista más sofisticados, está reconocida una gran variedad de proposiciones en el manejo de observaciones discrepantes, se han propuesto modelos para la generación de estas observaciones y está disponible una gran variedad de técnicas estadisticas para el procesamiento de datos. A pesar de que existen algunos libros en el tema de observaciones discrepantes, la mayoría de material se encuentra aun disperso en la literatura.

Los primeros trabajos en observaciones discrepantes se caracterizan principalmente por la falta de atención a la modelación del mecanismo de generación de observaciones discrepantes, sin apoyo de un estudio en términos de las propiedades estadísticas de los modelos propuestos.

Ninguna observación puede garantizarse sea totalmente una manifestación del fenómeno bajo estudio. Un evento

con una posibilidad en un millón podría ocurrir con la frecuencia apropiada, no importando que tan sorprendidos estemos de que esto nos ocurra. Intuitivamente la confiabilidad de una observación se refleja por su relación con las otras observaciones que se obtienen bajo condiciones similares. Observaciones que en opinión del investigador se apartan del resto de los datos se llaman: observaciones discrepantes, disidentes, espurias, contaminantes o valores sorprendentes, contaminantes, malos o sucios; solo para mencionar algunos términos que se han usado en épocas pasadas.

El tema de observaciones discrepantes es viejo e indudablemente data de los primeros intentos para obtener conclusiones de un conjunto de datos. Comentarios hechos por Bernoulli (1777) acerca de observaciones astronómicas indican que la practica de descartar observaciones discrepantes era común 200 años atras. Los primeros intentos por desarrollar métodos estadísticos objetivos fueron reportados por 1850. Hoy en día deberíamos esperar métodos extendidamente aceptados para tratar y manejar observaciones discrepantes, ademas de una definición universalmente aceptada. Este no es el caso de acuerdo a nuestra interpretación con la literatura. Aunque mucho se ha escrito, la noción de observaciones discrepantes parece vaga hoy como lo fue 200 años atras. Por ejemplo Edgeworth (1887) escribe:

— Las observaciones discrepantes se pueden definir como aquellas que presentan la apariencia de diferir con respecto a su ley de frecuencias con las otras observaciones con las que están combinadas.—

82 años después Grubbs (1969) establece que:

> — Una observación discrepante, es una que parece desviarse marcadamente de los otros miembros de la muestra en que ocurre. —

en los libros recientes sucede algo similar como a continuación se evidencia. Barnett y Lewis (1978) definen observación discrepante de la siguiente manera:

> — Definimos observación discrepante en un conjunto de datos, como una observación (o subconjunto de observaciones) que parecen ser inconsistentes con el resto del conjunto de datos. —

mientras que Hawkins (1980) escribe:

> — La definición intuitiva de una observación discrepante será: "Una observación que se desvía mucho de las otras observaciones como para despertar sospechas de que fue generada por un mecanismo diferente" —

Estas afirmaciones ilustran que una observación discrepante es un concepto subjetivo después de que los datos se han obtenido. Historicamente, se emplearon métodos "objetivos" para el tratamiento de observaciones discrepantes solamente después de que estos fueron identificados por medio de una inspección visual de los datos.

Para fijar ideas, supongamos que las siguientes 10

observaciones, de Johnson y Hurt (1979) se cree son realizaciones independientes de una población normal:

-1.64, -1.33, -1.10, -0.57, -0.27
1.04, 1.56, 1.34, 2.04, 4.99

¿Cuál de estas observaciones es discrepante? La respuesta será reservada hasta que se construya una prueba o alguna otra medida objetiva en forma natural, pero la decisión para ampliar un criterio objetivo se basa frecuentemente en reacciones subjetivas iniciales de los datos. En este caso, la observación 4.99 parece demasiado grande y también se puede formular algún cuestionamiento para la observación 2.04.

Collet y Lewis (1976) reportan los resultados de un experimento para investigar la naturaleza subjetiva de la decisión para etiquetar una observación como discrepante. Ellos concluyen que la inclinación individual depende del método de presentación (azar, gráfica, orden) experiencia y la escala de los datos; las observaciones extremas tienden a parecer más discrepantes cuando la escala se incrementa.

En conjunto de datos mas estructurados, como, cuando el interés se centra en la forma en que observaciones de una variable de interés principal que varía con respecto a otras variables (análisis de regresión) o con respecto al tiempo (series de tiempo) u otras situaciones, debemos esperar encontrar de vez en cuando, datos no representativos en forma de observaciones discrepantes. Aquí también es importante, como en una muestra simple univariada, poder interpre-

tar y acomodar observaciones discrepantes mediante técnicas estadísticas apropiadamente diseñadas. En este caso las observaciones discrepantes pueden ser de interés intrínseco o pueden ser indicadores de especificaciones inapropiadas de la estructura del error o del modelo básico, con implicaciones de consecuencia para el uso de procedimientos apropiados de inferencia. Con datos mas estructurados surgen dos complicaciones: observaciones sospechosas tienden a ser menos aparentes, mas ocultos en el cuerpo de los datos y además los metodos para su rechazo o su acomodo estan menos desarrollados. En regresión por ejemplo, una inspección visual de los datos dificilmente reflejará cual caso es una observación discrepante. La evidencia para esto se refleja regularmente cuando los parámetros son estimados y los residuales tabulados. Debido a que los residuales no son independientes, se hace difícil juzgar a las observaciones discrepantes, además, también se complica la metodología. Mucho de esta metodología se basa en modelos de observaciones discrepantes y son necesarias algunas decisiones acerca de la naturaleza y frecuencia antes de escoger un criterio. La subjetividad se introduce en la etapa del modelaje.

Debido al énfasis en el modelo en años recientes, "observación discrepante" ahora parece usarse por varios autores para indicar cualquier observación que no provenga de la población seleccionada, aunque en publicaciones recientes falta aun una definición formal de observación discrepante. De las 10 observaciones dadas anteriormente 1.09, 2.09 y 4.99 fueron generados de una

población normal con $\mu=3$ y $\sigma^2=1$, mientras que las 7 restantes fueron generadas por una población normal con $\mu=0$ y $\sigma^2=1$. Algunos autores dirían que esta muestra tiene una observación discrepante si utiliza alguna de las definiciones expuestas anteriormente, mientras que otros dirían que este conjunto tiene 3 observaciones discrepantes por pertenecer éstas a otra distribución.

En un esfuerzo por evitar alguna ambigüedad adoptaremos las siguientes definiciones en el resto de este trabajo:

> Observación anormal :- cualquier observación que parezca sorprendente o discrepante para el investigador.
> Contaminante :- cualquier observación que no sea una realización de la población supuesta.
> Observación discrepante :- para referirse a contaminante o anormal.

1.2 ¿Porqué estudiar observaciones discrepantes?

Una respuesta está indicada en la sección anterior. Parece claro que se necesita un método adecuado para manejar tales observaciones o al menos un fuerte entendimiento de los méritos relativos a los métodos disponibles. Aunque esta es la razón históricamente dominante para el estudio de observaciones discrepantes, existen varias razones adicionales e igualmente importantes que se han enfatizado en años recientes.

Interés especial :- puede suceder que la atención principal sobre la(s) observación (es) discrepante (s) sea aun mayor que, por ejemplo la estimación de un parámetro pobla

cional. En tales situaciones el problema estadístico involucra el hacer inferencias válidas acerca de la(s) observación(es) en cuestión

Barnett (1978) describe un caso legal que es interesante : 349 días después de que el señor Hadlum marchó al extranjero en servicio militar, la señora Hadlum tuvo un parto. El señor Hadlum juzgó a la observación de 349 días discrepante cuando la comparó con el promedio de gestación de 280 días y por lo tanto hace una petición de divorcio. En este caso la ganancia no es cuando descartar la observación discrepante o el efecto en la estimación de algún parámetro, mas bien es como juzgar el peso de la evidencia contra la hipótesis de que la observación discrepante, aunque extrema, es una realización válida de la distribución tiempo de gestación.

Detección de un fenómeno alternativo específico.- Algunos países industrializados para localizar satélites extraños a los suyos han medido los cambios en el nivel de radiación en las regiones en que se sospecha la presencia de estos satélites. Los niveles de radiación que fueran altos relativos al promedio, se toman como indicadores de un fenómeno alternativo. La metodología estadística, básica consiste en la aplicación de esquemas de detección de observaciones discrepantes a gran escala. No hay interés en estimar el nivel promedio de radiación en el ambiente, excepto para proporcionar una guía para juzgar observaciones particulares como discrepantes.

Indicadores de diagnóstico.- En análisis más complica-

das, donde una empresa cree que el modelo es erróneo, la presencia de observaciones discrepantes es frecuente indicación de la inconsistencia en el modelo, en los datos o en ambos. Las observaciones discrepantes, pueden conformarse por ejemplo, cuando las respuestas son transformadas a una escala logaritmica. también las observaciones discrepantes pueden ser indicadoras de observaciones altamente influyentes en el modelo o reflejar errores en los renglones de la matriz (en el caso de regresión), no aditividad o heterocedasticidad.

Los análisis de datos deben incluir la aplicación de técnicas de diagnóstico que puedan suministrar información sobre lo apropiado del análisis y la precisión de las conclusiones. Naturalmente muchas de las técnicas disponibles se basan en la detección de observaciones discrepantes.

Acomodo e influencia:- El estudio de la naturaleza y frecuencia de observaciones discrepantes en un problema particular pueden llevar a modelos y métodos de estimación que acomoden observaciones discrepantes, mejorándose de esta manera la inferencia. Como ejemplo de estos métodos podemos considerar a los métodos robustos.

Las observaciones discrepantes no necesariamente son observaciones influyentes, es decir, los resultados de un análisis pueden permanecer esencialmente sin cambio cuando una observación discrepante se remueve del resto de la muestra. Es útil considerar a las observaciones

influyentes como un tipo especial de observación discrepante. El estudio de las observaciones influyentes hace posible, como veremos más adelante, el obtener un mejor entendimiento de la estructura de un problema, en particular, el problema de regresión. Por ejemplo, tales estudios pueden evidenciar deficiencias inherentes en los datos y llevarnos a experimentaciones adicionales.

1.3 Causas de observaciones discrepantes.

Varios autores (Anscombe 1960; Grubbs 1969) han intentado clasificar las formas diferentes en las que surgen las observaciones discrepantes. En el manejo de observaciones pueden encontrarse diferentes fuentes de variabilidad.

Variabilidad inherente.- Esta es la expresión de la forma en que las observaciones discrepantes varian intrinsecamente en la población; tal variación es una caracteristica incontrolable y natural de la población.

Error de Medición.- Frecuentemente en la toma de mediciones en miembros de una población bajo estudio, las fallas en el instrumental de medición que se usa, introducen un grado adicional de variabilidad, además de redondeos de los valores obtenidos y errores en la recolección entre otros, contribuyen a la variabilidad. En esta situación pueden tomarse algunas precauciones para reducir tal variabilidad.

Error de Ejecución.- también puede surgir una fuente de variabilidad en la colección imperfecta de los datos. Po-

demos escoger inadvertidamente una muestra insesgada o que incluye elementos que no son verdaderamente representativos de la población. Se pueden tomar ciertas precauciones para reducir esta variabilidad.

Beckman - Cook (1983), también clasifican las causas por las que surgen las observaciones discrepantes en 3 categorías: a) modelo global de inconsistencia; b) modelo local de inconsistencia; c) variabilidad natural. Pensando en las dos primeras categorías, las observaciones discrepantes deben juzgarse con algún modelo implícito o explícito en mente. Desde un punto de vista estadístico, es útil pensar que la aparición de una observación discrepante se debe a la incapacidad del modelo para proporcionar una explicación estadística o un ajuste adecuado.

Modelo Global de Inconsistencia:- Son todas aquéllas causas que llevan al reemplazo del modelo en turno por un modelo revisado o nuevo para toda la muestra. Esta categoría incluye causas como variables de respuesta en una escala equivocada, también nos puede llevar a una transformación o al reemplazo del modelo por un modelo mixto.

Modelo Local de Inconsistencia:- Son aquéllas causas que se aplican solamente a las observaciones discrepantes y no al modelo como un total. Tales causas requieren que las observaciones discrepantes sean tratadas con individualidad, ya que regularmente se desconoce algún modelo que incorpore a estas observaciones. Algunos ejemplos incluyen mediciones aisladas, errores en la recolección y observaciones -

altamente influyentes en regresión debido a puntos remotos en el espacio de factores.

Variabilidad Natural.- Finalmente una observación discrepante puede ser el resultado de una variación natural en lugar de una inconsistencia del modelo. Neyman y Scott (1971), Green (1974, 1976) introducen el concepto de familias $\mathcal{F}$ de distribuciones "propensas a discrepantes" y "resistentes a discrepantes" en un esfuerzo por caracterizar la extensión para que se produzcan observaciones discrepantes en forma natural. Sean $Y_{(1)}, Y_{(2)}, \ldots Y_{(n)}$ las estadísticas de orden de una muestra cuya distribución $F \in \mathcal{F}$ entonces se dice que $\mathcal{F}$ es "resistente a discrepante (k,n)" si

$$P(k,n) = \sup \Pr\left[Y_{(n)} - Y_{(n-1)} > k\left(Y_{(n-1)} - Y_{(1)}\right) / F \right] < 1$$

de otra manera se dice que $\mathcal{F}$ es "propensa a discrepante (k,n)". Si $P(k,n) = 1$ para toda $k > 0$ y $n \geq 3$, se dice que $\mathcal{F}$ es completamente propensa a discrepante".

1.4. Identificación y Acomodo

Siguiendo a Barnett (1978) y a Barnett y Lewis (1978), se distinguen dos grandes métodos para el manejo de las posibles observaciones discrepantes. El primer método es simplemente para la identificación de observaciones discrepantes para un estudio adicional. La identificación de una de estas observaciones nos puede llevar a: a) su subsecuente rechazo; b) información nueva importante contenida en variables concomitantes que de otra manera pasaría inadvertida; c) de su incorporación a través de una revisión del modelo (global de inconsistencia) o del método de

estimación ; d) reconocer la inconsistencia inherente en los datos y entonces el reconocimiento también en experimentos posteriores. Posteriormente son utiles los diagnósticos que se basan en la identificación de observaciones discrepantes principalmente, en la construcción iterativa de un modelo donde sea necesaria una evaluación critica en las diferentes etapas de desarrollo. En Box (1979, 1980), Cook y Weisberg (1982), podemos encontrar discusiones mas generales.

El segundo método es para el acomodo de posibles observaciones discrepantes, para modificaciones deseables del modelo y/o método de análisis. Por ejemplo, podemos usar modelos mixtos para acomodar cierto tipo de contaminantes y usar estimadores M para obtener alguna protección contra observaciones discrepantes cuando el modelo mixto sea simétrico.

De estos dos métodos generales, juzgamos que el de identificación es más importante. Los métodos de acomodo de observaciones discrepantes tienden a requerir mucha información, acerca del proceso de generación de observaciones discrepantes o están diseñados para ser inmunes a la presencia de estas observaciones. Los tipos anteriores a menudo obscurecen la información, mientras que los primeros requieren que mucha de la información esencial se haya obtenido ya usualmente por medio del proceso de identificación. Aunque las filosofías fundamentales de identificación y acomodo parecen distintas, estas regularmente son confundidas, ya que un método de acomodo puede producir un método de identificación

como un producto y viceversa.

Un punto que merece un énfasis especial es que los métodos de identificación y acomodo que se basan en modelos alternativos pueden producir resultados que son específicos al modelo en cuestión; entonces la determinación de una observación discrepante puede depender críticamente de nuestro entendimiento del problema. En particular, es útil distinguir entre situaciones donde se suponen observaciones discrepantes que llevan información acerca de los parámetros en el modelo básico y situaciones donde estas observaciones no llevan información relevante. El modelo de la varianza inflada es un ejemplo de la primera clase, mientras que el modelo de la media desfazada es un ejemplo de la segunda.

1.5 Breve historia del estudio de observaciones discrepantes.

La primera discusión acerca de observaciones discrepantes que se ha podido situar es la realizada por Bernoulli (1777). Bernoulli cuestionaba la suposición de que los errores tenían distribución idéntica y criticaba la práctica general de descartar observaciones como si éstas constituyeran la muestra completa. Un ejemplo de esta situación la observamos cuando Boscovich (1755), para determinar la elipticidad de la tierra en base a un promedio de 10 "seconds pendulums" a diferentes latitudes, descarta las dos observaciones más extremas y calcula la media a partir de la muestra reducida. Durante este periodo y hasta la mitad del siglo XIX el punto principal de discusión en la literatura, fué cuando se jus-

tificata el rechazo de observaciones sospechosas.

La primer persona que propuso un criterio objetivo para el rechazo de observaciones discrepantes fue Pierce (1859). El procedimiento de Pierce supone que el modelo que genera los datos es una mezcla de distribuciones con parámetro ρ desconocido. Haciendo la peor suposición acerca de ρ y que la distribución principal es $N(0, \sigma^2)$ con σ conocida, si el valor absoluto de cualquier residual se excede en $c\sigma$, éste debe ser rechazado, donde c se determina de la siguiente manera: Si $\Phi(\cdot)$ denota la distribución de probabilidad acumulada de la normal estandar y s denota la desviación estandar de la muestra completa y s' la de la muestra que resulta de eliminar k observaciones discrepantes. Si $n' = n-k$, entonces c, es tal que satisface la ecuación $(s'/s)^n \frac{1}{\sqrt{2\pi}} \exp\left[\frac{1}{2}c^2 - 1\right] 2\Phi(-c)\}^n = [k^k n'^{n'}/n^n]^{n/k}$. tres años después Gould (1855) proporciona tablas para la implantación del criterio de Pierce.

Sin embargo en este tiempo hay muchos investigadores que plantean objeciones teóricas y prácticas al criterio de Pierce, el astrónomo Airy comenta al respecto:

> — Pienso que toda la teoría es defectuosa en su fundamentación e ilusoria en sus resultados; no se puede obtener una regla para la exclusión de observaciones por un proceso fundamentado en una consideración de la discrepancia de esas observaciones. —

Chauvenet (1866) propone un método diferente: Notando que el valor esperado del número de observaciones en una muestra de tamaño n de una $N(0,\sigma^2)$ excediéndose $\sigma\sigma$ es $n\bar{\Phi}(-c)$, él propuso que cualquier residual que en valor absoluto exceda a $\sigma\sigma$ debe ser rechazado, donde c satisface

$$n\bar{\Phi}(-c) = 0.5$$

entonces el criterio de Chauvenet es tal, que rechaza en promedio media observación de datos buenos por muestra, sin tomar en cuenta el valor de n. Este aspecto de la prueba de Chauvenet es especialmente significante ya que se aboca a la tasa de rechazo del experimento y no a la proporción de datos rechazados, así cuando n crece, c también lo hace y la proporción de observaciones rechazadas decrece.

Un tercer criterio para el rechazo completo de observaciones discrepantes fue creado por Stone (1868). El encuentra al criterio de Pierce fastidioso, y citando a Airy expresa sus dudas acerca de la validez matemática de dicho procedimiento. El afirma que cualquier observador tiene una probabilidad de $\frac{1}{n}$ de cometer un error, y propone que c sea tal que

$$n\bar{\Phi}(-c) = 0.5$$

La aparente similitud entre el criterio de Chauvenet y el de Pierce es engañosa. La regla de Stone rechaza una proporción fija de observaciones buenas y así el número de observaciones discrepantes crece en proporción con

el tamaño de la muestra.

Glaisher (1872-73) fue quizas la primer persona en escribir un procedimiento ponderado. Su método supone que las x_i $(i=1,2,\ldots,n)$ tienen distribución $N(\mu, \sigma_i^2)$, con la media μ común a todas ellas, pero con varianzas σ_i^2 desiguales y desconocidas. Se propone un valor inicial del estimador de la media y de los valores de σ_i^2, con éstos modifica el valor inicial del estimador de la media y de las σ_i^2 y así sucesivamente. Específicamente se calcularon medias ponderadas $m_{(1)}, m_{(2)}, \ldots,$ siendo $m_{(1)}$ la media muestral ordinaria y los pesos w_{ri} para la r-ésima media ponderada $m_{(r)} = \sum_i w_{ri} x_i$ definidos recursivamente por

$$w_{ri} \propto \exp\{-2(w_{r-1,1} + \ldots + w_{r-1,n})(x_i - m_{(r-1)})^2\}$$

Stone (1873), unos meses después critica este método y propone un procedimiento alternativo que se basa en la maximización de la verosimilitud, que es proporcional a

$$\prod_i [(1/\sigma_i) \exp -\{(x_i - \mu)^2 / 2\sigma_i^2\}],$$

con respecto a μ y a todas las σ_i. Esto lleva a una media ponderada $\tilde{\mu}$ dado por la ecuación de grado n-1

$$\sum_{i=1}^{n} (x_i - \tilde{\mu})^{-1} = 0$$

Este mismo método fué publicado independientemente por Edgeworth (1883), aunque en 1887 reconoce que Stone fué el primero en enunciarlo.

Otro método ponderado propuesto en esta época fue el de Newcomb (1886), su procedimiento supone que las n observaciones provienen de una mezcla de r distribuciones normales y desarrolla un estimador final de μ, construido como una media ponderada de r^n medias ponderadas diferentes de las x_i. Tal vez de mayor interés es la referencia de Newcomb en esa publicación al valor "maligno", que resulta ser el error cuadrático medio de un estimador, y es uno de los primeros usos del concepto de función de Pérdida. Una de las cosas notables es que los pesos que obtuvo corresponden, en estimación robusta, con una función $\psi = \max [-c, \min (x, c)]$ una de las funciones favoritas de Huber.

Edgeworth (1887) propone 3 modelos para observaciones discrepantes e ilustra cada uno, usando el método de Monte Carlo. Entonces estos modelos fueron usados como una base para un estudio comparativo de los modelos de Airy, Chauvenet, Glaisher, Newcomb, Pierce y Stone, obteniendo los siguientes resultados: a) El "modulo de descuido" de Stone es legitimo; b) Se refiere a la sugerencia de Airy como el "No-método"; c) Reconoce los modelos fundamentales del trabajo de Stone y Glaisher; d) Indica que el criterio de Chauvenet y Pierce es pesimista; e) Expresa una gran consideración por el trabajo de Newcomb.

Daniell (1920) en su trabajo "observaciones ponderadas de acuerdo a su orden" escribe; además de la media que resulta después de la eliminación de observaciones discrepantes, debemos buscar otros estimadores en los que los pesos son múltiplos de funaones que dependen del orden de las observaciones.

Irwin (1925). sienta las bases de una prueba de varias - observaciones discrepantes en la muestra. Considérese un conjunto de datos x_i $(i = 1, 2, \ldots, n)$ con $x_i \sim iid\ N(\mu, \sigma^2)$, σ co - nocida. Irwin consideró las estadísticas de orden $X_{(1)}, X_{(2)}, \ldots$ $X_{(n)}$ de la muestra y derivó la distribución de la k-ésima "brecha" $(X_{(n-k+1)} - X_{(n-k)})/\sigma$, también derivó expresiones que - computacionalmente son tratables para los momentos de esta k-ésima brecha y demostró que para $k = 1, 2$, la distribución de la brecha se puede aproximar por medio de las dis_ tribuciones normales cuyas medias y varianzas él tabu - ló para varios valores de n.

Goodwin (1913) propuso que en lugar del criterio $(x_i - \bar{x})/s$ usado por los primeros autores para ciertas pruebas de obser_ vaciones discrepantes, uno debería sustituir a $\bar{x}$ y s por la media y desviación estandar que resultarían de eliminar la observación discrepante considerada en la muestra. Al - gunos años tuvieron que pasar antes de que se demos - trara que esta estadística de prueba es una función - monótona de $(x_i - \bar{x})/s$. Este descubrimiento hecho por thomp_ son (1935) lleva a la derivación de la distribución nula - de un residual studentizado arbitrario. De este resultado, él deduce un procedimiento en contra de observaciones dis_ crepantes y rechaza una proporción fija de datos bue - nos, produciendo también una tabla de valores crí - ticos.

En 1936 Pearson y Chandra Sekar, usando los resul_ tados de thompson y algunos lemas matemáticos so - bre residuales studentizados, demuestran que para al - guna c suficientemente grande, el evento $|x_i - \bar{x}|/s > c$ im

plica que para toda $j \neq i$, $|x_j - \bar{x}|/s < c$. En - tonces si c se escoge de tal manera que sea el $\alpha/2n$ percentil de la distribución de $|x_i - \bar{x}|/s$, entonces una prueba de observaciones discrepantes rechazando cualquier x_i que satisfaga $|x_i - \bar{x}|/s > c$ tiene un nivel de significancia experimental de α. Esta publicación también manifiesta la existencia del efecto de enmascaramiento que se analizará más adelante.

Winsor (1941) propone un procedimiento que hoy en día lleva su nombre y que consiste en que, la - observación sospechosa no se rechaza totalmente de la muestra, sino que se sustituya por el valor más - cercano a esta observación que no sea considerada discrepante.

Capítulo 2.

ENFOQUE FRECUENTISTA

Este enfoque recibe también el nombre de - clásico, estándar, ortodoxo o de teoría de muestreo y se distingue por dos características principales. Considera como única forma de información relevante a aquella proporcionada en forma de datos muestrales, y adopta como base para la evaluación y construcción de procedimientos estadísticos, el concepto de probabilidad visto como frecuencia, es decir, para cualquier evento, la probabilidad de éste será una medida de su "regularidad estadística".

2.1 El modelo de regresión lineal.

En innumerables procesos se tiene que el - comportamiento de la variable bajo estudio es influenciada por las variables que intervienen en - el proceso de obtención de las observaciones. Si se cree que la respuesta podría ser aproximada a partir de una relación funcional, en la - cual se consideren únicamente aquéllas variables que se juzgue mas importantes, el modelo que se - propone para esta situación es

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i \quad i = 1,2,\ldots n \qquad (2.1.1)$$

donde Y_i es la i-ésima observación de la variable bajo estudio y es x_{ij} el valor de la j-esima variable cuando se obtiene la i-ésima observación, β_i $i=1,2,\ldots n$ es el i-ésimo parámetro (desconocido) y

ε_i es el error asociado a la i-ésima observación — causado por aquéllos elementos que se consideró no eran determinantes para explicar el comportamiento de la variable bajo estudio. La variable Y es — llamada comúnmente variable dependiente mientras que $x_1, \ldots, x_p$ reciben el nombre de variables respuesta o explicativas.

Con el objeto de complementar la definición del modelo, se hacen algunas suposiciones acerca de la variable aleatoria ε_i

$$E(\varepsilon_i)=0 \; ; \; V(\varepsilon_i)=\sigma^2, \; cov(\varepsilon_i, \varepsilon_j)=0 \quad \forall \; i \neq j \qquad (2.1.2)$$

La notación matricial del modelo (2.1) es

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \qquad (2.1.3)$$

donde Y es un vector n-dimensional cuyas componentes son las y_i, $\underline{\beta}$ es un vector p-dimensional de parámetros desconocidos, X es una matriz de nxp con rango $r(X)=p$ cuyo i-ésimo renglón es de la forma $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ y $\underline{\varepsilon}$ es un vector aleatorio n dimensional cuyas componentes satisfacen (2.1.2).

El problema posterior a la definición del modelo, es encontrar el valor de β, que minimice de alguna manera $\underline{\varepsilon}$. El criterio que mayor difusión ha tenido es el de Mínimos Cuadrados, cuya metodología está fuera del objetivo de es-

te trabajo. La solución que se obtiene para $\underline{\beta}$ por medio de este método es

$$\hat{\underline{\beta}} = (X^t X)^{-1} X^t \underline{Y}. \qquad (2.1.4)$$

Nótese que la matriz X puede pensarse como una transformación lineal, es decir $X : \mathbb{R}^p \longrightarrow \mathbb{R}^n$. La imagen de X, $Im(X)$, es un subespacio de $\mathbb{R}^n$ cuya dimensión es precisamente $r(X) = p$. también es importante hacer notar que la ecuación $\underline{Y} = X\underline{\beta}$ no tiene solución, es decir, $\underline{Y}$ no pertenece a $Im(X)$.

Sea

$$\begin{aligned} \hat{\underline{Y}} &= X\hat{\underline{\beta}} \\ &= X(X^tX)^{-1}X^t\underline{Y}, \end{aligned} \qquad (2.1.5)$$

el vector de valores esperados que resultan de "ajustar" el modelo (2.1.3) por mínimos cuadrados. En la figura (2.1.1) encontramos la interpretación geométrica de la solución al problema de mínimos cuadrados.

Como podemos observar $\hat{\underline{Y}}$ es la proyección ortogonal de $\underline{Y}$ en $Im(X)$. La matriz

$$V = X(X^tX)^{-1}X^t, \qquad (2.1.6)$$

cuyas propiedades se estudian mas adelante, es la matriz que proyecta a $\underline{Y}$ en la imágen de X. también debemos notar que la matriz

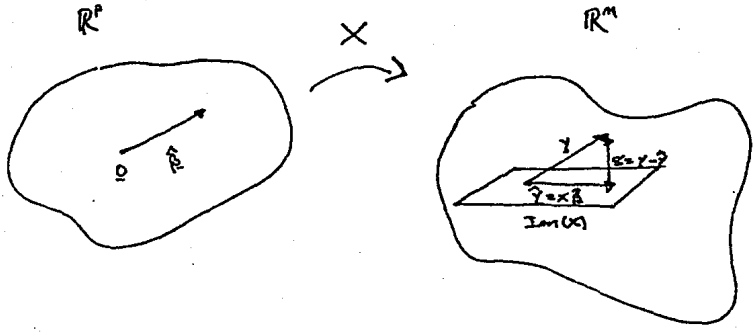


figura 2.1.1 solución al problema de mínimos cuadrados.

$$M = I - V \qquad (2.1.7)$$

es la matriz que proyecta ortogonalmente a $\underline{y}$ en el subespacio de $\mathbb{R}^n$ que es ortogonal a $Im(X)$, esta proyección ortogonal es precisamente el vector de residuales $\hat{\varepsilon}$, cuyas componentes son

$$y_i - \hat{y}_i = y_i - \underline{x}_i^t \hat{\beta} \; ; \; i = 1, 2, \ldots n \qquad (2.1.8)$$

donde $\underline{x}_i^t$ es el i-ésimo renglón de la matriz X.

2.1.1 La Matriz de proyección

(i) Propiedades. Algunas características están presentes en la sección anterior, además por ser una

matriz de proyección; V es simétrica e idempotente y por lo tanto

$$v_{ii} = \sum_{j=1}^{n} v_{ij}^2 = v_{ii}^2 + \sum_{j=1}^{n} v_{ij}^2$$

$$v_{ii} = 0 \Rightarrow v_{ij} = 0$$
$$v_{ii} = 1 \Rightarrow v_{ij} = 0 \; ; \; i \neq j,$$

de lo anterior se desprende que $0 \leq v_{ij} \leq 1$, también se sabe que los valores característicos de una matriz de proyección son cero o uno y que el número de éstos, es igual al rango de la matriz. En este caso $r(V) = r(X) = p$ y por lo tanto la traza de V es p, es decir $\sum_{i=1}^{n} v_{ii} = p$. Entonces el promedio de los elementos de la diagonal es p/n. Una regla empírica señala que un valor v_{ii} es grande si $v_{ii} > 2\, p/n$.

(ii) Importancia de V en el análisis de datos. El vector de residuales ordinarios (2.1.8) está dado por

$$\underline{e} = \underline{Y} - \hat{\underline{Y}}$$
$$= (I - V)\, \underline{Y},$$

la relación entre $\underline{e}$ y $\underline{\varepsilon}$ es

$$\underline{e} = (I - V)(X\underline{\beta} + \underline{\varepsilon})$$
$$= (I - V)\, \underline{\varepsilon}$$

o en forma escalar

$$e_i = \varepsilon_i - \sum_{j=1}^{n} v_{ij}\,\varepsilon_j .$$

Esta igualdad, muestra claramente que la relación entre e y ε depende sólo a través de V. Si los v_{ij}'s son suficientemente pequeñas, e servirá como un sustituto razonable de ε, de otra manera la utilidad de e, estará limitada. Si $\varepsilon \sim N(0, \sigma^2 I)$ entonces $e \sim N(0, \sigma^2(I-V))$. Casos alejados del espacio de factores tendrán valores de v_{ii} relativamente grandes. Ya que $V(\hat{y}_i) = v_{ii}\,\sigma^2$ y $Var(e_i) = (1 - v_{ii})\sigma^2$, los valores ajustados en puntos remotos tendrán varianzas relativamente grandes y los residuales correspondientes tendrán varianzas relativamente pequeñas. Huber llama a $1/v_{ii}$ "el número efectivo de casos que determinan $\hat{y}_i$". Davies y Hutton (1975) señala que si $\max v_{ii}$ no es considerablemente menor que 1, entonces probablemente una observación discrepante pasará desapercibida cuando se examinen los residuales. Box y Draper (1975) sugieren que para diseñar un experimento que sea insensible a observaciones discrepantes, las v_{ii}'s deben de ser pequeñas y aproximadamente iguales.

La importancia de las v_{ii}'s no se limita al análisis de mínimos cuadrados. Huber (1977) menciona que regresión robusta no funciona bien si $\max v_i = 1$.

Hoaglin y Welsch (1978) sugieren el uso directo de las v_{ii} como un diagnóstico para identificar "puntos altamente influyentes". El motivo detrás de esta sugerencia, se basa en la representación

$$\hat{y}_i = v_{ii}\, y_i + \sum_{j \neq i} v_{ij}\, y_i .$$

El valor ajustado $\hat{y}_i$ estará dominado por $v_{ii}\, y_i$ si v_{ii} es grande con respecto a los términos restantes. Ellos interpretan a v_{ii} como la cantidad de influencia ejercida en $\hat{y}_i$ por y_i. Es claro, sin embargo, que para cualquier $v_{ii} > 0$, $\hat{y}_i$ estará dominado por $v_{ii}\, y_i$ si y_i es suficientemente diferente de los otros elementos de $\underline{Y}$ (Esto es, una observación discrepante).

2.1.2 Residuales

Cuando el modelo ajustado es incorrecto, cambiará la distribución de los errores no observables $\underline{\varepsilon}$ y por lo tanto de los residuales $\underline{e}$. El objetivo en el estudio de los residuales $\underline{e}$ es detectar cualquier su posición incorrecta en $\underline{\varepsilon}$. Desafortunadamente, la correspondencia entre $\underline{e}$ y $\underline{\varepsilon}$ es menos que perfecta. En algunos problemas, las fallas en el modelo no se transmiten a $\underline{\varepsilon}$, en otros, los síntomas observados pueden atribuirse a más de una suposición incorrecta.

Para procedimientos de diagnóstico, se han sugerido varias transformaciones de los residuales ordinarios e para superar algunas deficiencias, por ejemplo obtener residuales que sean independientes de los parámetros de escala o para obtener alguna estructura de covarianza específica.

Los residuales ordinarios, tienen una distribución -

que depende de la escala ya que la varianza de cada e_i es una función de σ^2 y v_{ii}. Para muchos procedimientos de diagnóstico, es útil definir la versión "studentizada" de los residuales que no dependen de alguna de estas cantidades.

Cuando se usa mínimos cuadrados en regresión, los residuales studentizados internos se definen por

$$r_i = \frac{e_i}{s(1-v_{ii})^{1/2}} \qquad i = 1, 2, \ldots, n \qquad (2.1.9)$$

donde $s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$ Nos reservaremos el término de residual studentizado al referirnos a (2.1.9).

Los residuales studentizados se usan para reemplazar a los residuales ordinarios en procedimientos gráficos, tal como el diagrama contra valores ajustados (Anscombe y Tukey 1963). También se usan como base para construir la mayoría de las estadísticas que se proponen en este trabajo.

Los residuales studentizados externamente requieren un estimador de σ^2 que sea independiente de e_i. Bajo la suposición de normalidad de los errores, sea $s^2_{(i)}$ el cuadrado medio residual calculado sin el i-ésimo caso, entonces se propone

$$s^2_{(i)} = \frac{(n-p)s^2 - e_i^2/(1-v_{ii})}{n-p-1} = s^2\left(\frac{n-p-v_i^2}{n-p-1}\right). \qquad (2.1.10)$$

Bajo el supuesto de normalidad, $S^2_{(i)}$ y $e_{(i)}$ son independientes y el residual studentizado externamente se define por

$$t_i = \frac{e_i}{s_{(i)}(1-v_{ii})^{1/2}} \qquad i=1,2,\ldots n, \qquad (2.1.11)$$

La distribución de t_i es t-student con $n-p-1$ grados de libertad (Beckman y Trussell 1974). La relación entre t_i y r_i se encuentra sustituyendo (2.1.10) en (2.1.11)

$$t_i = r_i\left(\frac{n-p-1}{n-p-r_i^2}\right)^{1/2}, \qquad (2.1.12)$$

que demuestra que t_i^2 es una transformación monótona de r_i^2.

2.2 Modelos para el manejo de observaciones discrepantes.

En un marco de pruebas de hipótesis, Barnett (1978), Barnett y Lewis (1978) plantean los modelos probables que pueden generar observaciones discrepantes: Para una prueba de observaciones discrepantes, la hipótesis nula H expresará el modelo básico de probabilidad que genera todos los datos sin considerar a las observaciones discrepantes, la hipótesis alternativa $\bar{H}$ expresa una manera en la que puede modificarse el modelo para incorporar o explicar las observaciones discrepantes. La hipótesis nula es solamente una afirmación del modelo básico de probabilidad.

Si en base a una prueba de discrepancia juzgamos que existe una o más observaciones discrepantes, rechazamos implícitamente la hipótesis nula en favor de una hipótesis alternativa que naturalmente debemos conocer. Consideremos algunas formas posibles de hipótesis alternativas para pruebas de discrepancia que nos permita considerar que posibilidades existen en el modelo alternativo (generador de observaciones discrepantes) en el caso de una de estas observaciones.

(i) Alternativa Determinística

Este primer tipo cubre los casos de observaciones discrepantes causados por errores gruesos de medición o recolección. Esta hipótesis esta totalmente especificada por el

conjunto de datos incluyendo a las observaciones discrepantes. Así si los datos $x_1, x_2, \ldots, x_n$ contienen una observación (x_j) que surge de un error de medición o recolección, rechazamos inmediatamente el modelo básico F en favor de un modelo alternativo que diga que todas las x_i $(i \neq j)$ provienen de F mientras que x_j es diferente y debe de rechazarse, corregirse o repetirse. No se necesita prueba alguna.

(ii) Alternativa Inherente.

En este caso debemos tomar en cuenta la posibilidad de que las observaciones discrepantes aparecieron en los datos como resultado de la variabilidad inherente

$$H: x_i \in F, (i = 1, 2, \ldots, n) \text{ VS } \bar{H}: x_i \in G \neq F \ (i=1,2,\ldots n).$$

(iii) Alternativa Mixta

En vez de suponer que las observaciones discrepantes reflejan un grado o forma inesperada de variabilidad inherente, debemos admitir la posibilidad de "errores de ejecución" permitiendo la "contaminación" de la muestra por miembros de una población diferente a la representada por el modelo básico

$$H: x_i \in F \ (i=1,2,\ldots,n) \text{ VS } H: x_i \in (1-\alpha)F + \alpha G; (0 < \alpha < 1) \ (i=1,2,\ldots,n).$$

(iv) Alternativa trasladada.

En esta alternativa, se establece que n-k observaciones independientes surgen de un modelo inicial F con parámetros de posición y escala μ y σ^2 respectivamente, mientras que un número "pequeño" k de observaciones independientes surgen de una versión modificada de F en la que μ y σ^2 están desfazados en su valor (μ en cualquier dirección y σ^2 regularmente un valor mayor al inicial)

$$H : x_i \in F \ (i=1,2,\ldots,n) \quad VS \quad H: \begin{cases} x_i \in F \ (i=1,2,\ldots,j-1,j+1,\ldots,n) \\ x_j \in F' \end{cases}$$

donde F' es la versión modificada de F. Por ejemplo si

$$F \sim (\mu, \sigma^2),$$

F' puede tomar alguna de estas formas

$$a)\ F' \sim (\mu + \lambda, \sigma^2); \qquad b)\ F' \sim (\mu, \delta\sigma^2),$$

formas conocidas como modelo de la media "trasladada" (shift) y modelo de la varianza "inflada" respectivamente.

En la revisión que se ha hecho de los métodos, se nota que la literatura sobre observaciones discrepantes en análisis de regresión y en general en modelos lineales, se encuentra dominada por el modelo de la media trasladada

$$\underline{Y} = X\underline{\beta} + D\underline{\lambda} + \underline{\varepsilon}, \qquad (2.2.1)$$

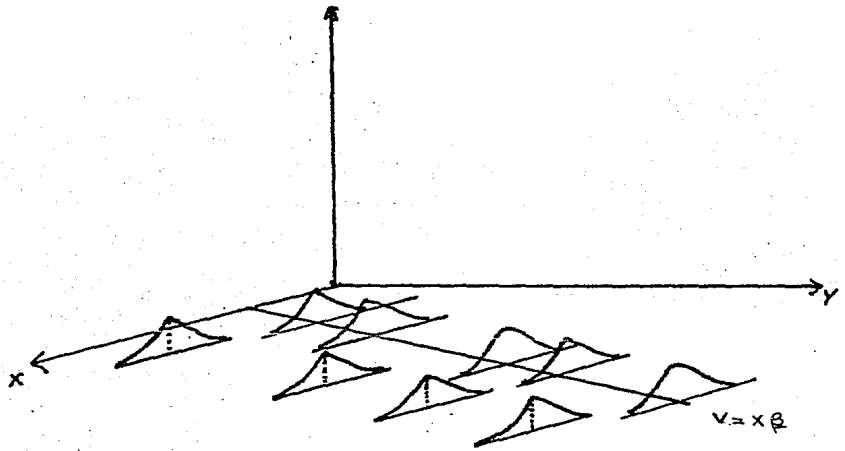
que se detalla a continuación en términos generales. Sea $\mathbf{i} = \{i_1, i_2, \ldots, i_K / 1 \leq i_1 < i_2 < \ldots < i_K\}$ y $\underline{u}_\ell$ un vector de $n \times 1$ cuya ℓ-ésima componente ($\ell \in \mathbf{i}$) vale uno y las restantes valen cero. La única diferencia del modelo (2.2.1) con respecto al modelo

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon}, \qquad (2.2.2)$$

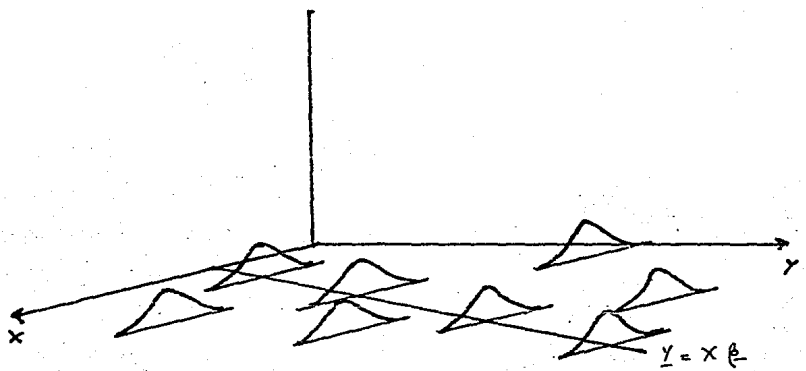
que es el modelo de regresión con las suposiciones de costumbre, es el término $D\underline{\lambda}$ donde D es una matriz de $n \times K$ cuyas columnas son los vectores $u_{i_1}, u_{i_2}, \ldots, u_{i_K}$ y $\underline{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_K)$ es un vector de K parámetros desconocidos. Debe notarse que el planteamiento que se está haciendo con el modelo (2.2.1) es que existen K observaciones que no provienen del modelo (2.2.2) y que provienen de un modelo alternativo, a saber el modelo (2.2.1) que especifica que estas K observaciones surgen de una o varias poblaciones, ésto dependerá según los λ_i ($i = 1, 2, \ldots K$) sean iguales o no, es decir, si $i \in \mathbf{i}$ entonces $E(y_i) = x_i^t \beta + \lambda_i$ con x_i^t el i-ésimo renglón de X. Para una descripción intuitiva, ver figura (2.2.1)

Srikantan (1961) y Ferguson (1961) fueron los primeros en establecer el modelo de la media trasladada como una base para el manejo de observaciones discrepantes, aunque también este enfoque está implícito en el trabajo de Daniel (1960).

El modelo de la media trasladada se considera mas apropiado para la identificación de observaciones



a) Las k observaciones discrepantes provienen de una población alternativa.



b) Las k observaciones discrepantes provienen de poblaciones diferentes

Figura 2.2.1

discrepantes que para su acomodo.

Beckman y Cook (1983) distinguen dos situaciones del modelo (2.2.1): La versión etiquetada en la que se supone conocido el conjunto i es decir, se conocen cuantas observaciones discrepantes hay y sus posiciones; y la versión no etiquetada en la que i es desconocido.

2.3 Métodos para el manejo de observaciones discrepantes.

2.3.1 Versión Etiquetada.

Como esta versión es un modelo lineal estándar, el análisis puede llevarse a cabo usando los métodos usuales y los resultados pueden expresarse en términos del modelo (2.2.2) John (1978), Beckman – Cook (1983).

Así, el estimador mínimos cuadrados de λ es

$$\hat{\lambda} = (I - V_i)\, \underline{e}_i \qquad (2.3.1)$$

y el estimador mínimos cuadrados de $\underline{\beta}$ al suprimir las k observaciones sospechosas resulta

$$\hat{\beta}_{(i)} = (X_{(i)}^{t} X_{(i)})^{-1} X_{(i)}^{t} Y_{(i)}. \qquad (2.3.2)$$

La reducción en la suma de cuadrados debido al ajuste de λ en el modelo (2.2.1) proporciona una medida del efecto de suponer que en realidad las k observaciones son discrepantes. Esta suma de cuadrados es

$$Q_k(i) = \underline{e}^t\underline{e} - \hat{\underline{e}}^t\hat{\underline{e}} \qquad (2.3.3)$$

donde e es el vector de residuales que resulta del modelo (2.2.2) y $\hat{e}$ el vector de "residuales revisados" de Gentleman y Wilk (1975):

$$\hat{\underline{e}} = \underline{Y} - X\beta_{(i)} - \hat{\lambda}$$

donde si $i \in i$, entonces $\hat{\lambda}_i = y_i - x_i \hat{\beta}_{(i)}$, caso contrario $\hat{\lambda}_i = 0$. John (1978), Draper y John (1981) hacen notar que el modelo (2.2.1) es equivalente al usado cuando se estiman k "valores faltantes" en análisis de covarianza con covariables indicadoras (dummy). Entonces es posible considerar a las observaciones discrepantes como observaciones faltantes y reemplazarlas por sus valores faltantes estimados. Los resultados son los mismos.

La estadística que se propone en Cook y Weisberg (1982), Beckman y Cook (1983) suponiendo normalidad para probar la hipótesis $\lambda = 0$ es

$$F_k(i) = \frac{(n-p-k)\,Q_k(i)}{(n-p-r_i^2)}.$$

Si $k=1$ Pekkman y Trussell (1979) demuestran que

$$F_1(i) = \frac{(n-p-1)\,r_i^2}{(n-p-r_i^2)} \qquad (2.3.4)$$

es una función monótona creciente en r_i^2 cuya distribución es t-student con $n-p-1$ grados de libertad y

donde r_i es el i-ésimo residual studentizado.

2.3.2. Versión No Etiquetada.

El análisis de la versión no etiquetada es mucho más complicada ya que tanto i como k son desconocidas. Para una k fija, una de las recomendaciones más comunes es la de examinar todos los $\binom{n}{k}$ subconjuntos y estimar i, usando el conjunto de índices $\hat{i}$ que maximiza $Q_k(i)$. Desgraciadamente el manejo de esta idea implica un alto costo en el aspecto computacional, razón por la cual aun no se tiene una técnica unificada a pesar de que se han sugerido muchas versiones en un esfuerzo por evitar estos costosos cálculos.

Debido a que en los últimos años ha surgido una gran cantidad de técnicas, en este trabajo se tratará de presentar algunas de las mas importantes.

2.3.2.1 Método de Ellenberg

Considere el modelo de regresion lineal con las suposiciones de costumbre y defínase el i-ésimo residual estandarizado por

$$\gamma_i = \frac{r_i}{(n-p)^{1/2}}, \qquad (2.3.5)$$

con r_i el i-ésimo residual studentizado interno definido en (2.1.9).

La regla de decisión es:

Rechazar H_0: "no hay observaciones discrepantes" si $\tau^* \geq c_\alpha$ donde c_α es un valor crítico, α es el nivel de significancia deseado y

$$\tau^* = \max_i |\tau_i|. \tag{2.3.6}$$

Como la distribución de τ^* (aún bajo la suposición de normalidad) resulta intratable, se propone el uso de la primera desigualdad de Bonferroni para encontrar cotas para la distribución de τ^*; así para cualquier constante crítica c_α

$$\sum_{i=1}^{n} (Pr|\tau_i| > c_\alpha) - \sum_{i>j} Pr(|\tau_i| > c_\alpha, |\tau_j| > c_\alpha) \leq P(\tau^* > c_\alpha) \leq \sum_{i=1}^{n} Pr(|\tau_i| > c_\alpha) \tag{2.3.7}$$

o equivalentemente

$$nPr(|\tau_i| > c_\alpha) - \sum_{i>j} Pr(|\tau_i| > c_\alpha, |\tau_j| > c_\alpha) \leq Pr(\tau^* > c_\alpha) \leq nPr(|\tau_i| > c_\alpha) \tag{2.3.8}$$

donde las τ_i se distribuyen idénticamente con

$$f(\tau_i) = \frac{\Gamma(\gamma+1)}{\Gamma(\frac{1}{2})\Gamma(1+\frac{1}{2})} (1 - \tau_i^2)^{\gamma-\frac{1}{2}},$$

para $-1 < \tau_i < 1$, $\quad i = 1, 2, \ldots, n$

y

$$f(\tau_i, \tau_j) = \frac{\gamma}{\pi(1-e_{ij}^2)^{1/2}} \left[1 - \frac{1}{(1-e_{ij}^2)} \left(\tau_i^2 - 2\rho_{ij}\tau_i\tau_j + \tau_j^2\right)\right]^{\gamma-1} \tag{2.3.9}$$

$$\text{para} \quad \tau_i^2 - 2\rho_{ij}\tau_i\tau_j + \tau_j^2 \leq (1-e_{ij}^2),$$

con $\rho_{ij} = -\dfrac{1-v_{ij}}{[(1-v_{ii})(1-v_{jj})]^{1/2}}$ y $\gamma = \dfrac{(n-p-2)}{2}$. la función (2.3.9) es una versión estandarizada de la función student invertida y puede relacionarse implícitamente con la función t-student bivariada.

La solución para la cota inferior de la desigualdad (2.3.8) implica el calculo de todas las $\binom{n}{2}$ funciones del tipo (2.3.9), dicha carga computacional se puede reducir usando el resultado de Srikantan (1961)

$$P(|\tau_i| > c_\alpha, |\tau_j| > c_\alpha) = 0 \quad \text{si } c_\alpha \geq \left(\frac{1+|\rho_{ij}|}{2}\right)^{1/2} \tag{2.3.10}$$

y determinando c_α, con α dado, tal que en la cota superior de (2.3-8)

$$P(|\tau_i| > c_\alpha) = \frac{\alpha}{n},$$

entonces todas las parejas (τ_i, τ_j) donde $c_\alpha \geq \left(\frac{1+|\rho_{ij}|}{2}\right)^{1/2}$ no contribuyen a la cota inferior. Debe de mencionarse que c_α varía inversamente con α, asi que teoricamente puede hacerse decrecer α en cualquier problema hasta que c_α alcance las condiciones de la desigualdad (2.3.10) para todas las $\binom{n}{2}$ ρ_{ij}, de ese modo pueden producirse re-

sultados exactos.

Para aquéllos casos donde la distribución bivariada no sea cero, usar la expresión

$$P(|r_i| > c_\alpha, |r_j| > c_\alpha) = 2P(r_i > c_\alpha, r_j > c_\alpha) + 2P_r(r_i > c_\alpha, r_j < -c_\alpha)$$

y hacer la transformación

$$t_i = (r_i - r_j)/[2(1-e_{ij})]^{1/2},$$

$$t_j = (r_i + r_j)/[2(1-e_{ij})]^{1/2},$$

entonces la distribución conjunta de t_i, t_j es

$$g(t_i, t_j) = \frac{\gamma}{\pi} (1 - t_i^2 - t_j^2)^{\gamma-1} \quad \text{para } t_i^2 + t_j^2 \leq 1.$$

Desgraciadamente este procedimiento sólo funciona bien para valores pequeños de $\gamma = (n-p-2)/2$, ya que para valores grandes pueden surgir problemas de cálculo y de precisión.

Lund (1975) basándose en la cota superior de la desigualdad (2.3.8) propone tablas para una prueba aproximada en modelos lineales para $\alpha = 0.10, 0.05, 0.01$. La estadística básica de prueba es

$$r^* = \max_i |r_i|, \qquad (2.3.11)$$

que es diferente a (2.3.6), pero

$$Pr(r^* > r_0) = P(\tau^* > \tau_0) \leq nP(|\tau_i| > \tau_0) = 2n \int_{\tau_0}^{1} f(\tau)\, d\tau,$$

con $r_0 = \frac{\tau_0}{(n-p)}$. Así, para obtener valores críticos r_0, es necesario resolver para τ_0 $\quad 2n \int_{\tau_0}^{1} f(\tau)\, d\tau = \alpha$

$$\Rightarrow \int_{\tau_0}^{1} f(\tau_i)\, d\tau = \frac{\alpha}{2n}$$

o equivalentemente

$$\int_{0}^{\tau_0} f(\tau) d\tau = \frac{1}{2}\left(1 - \frac{\alpha}{n}\right).$$

Nota: Para encontrar los valores críticos c_α para la prueba de Ellenberg por medio de las tablas de Lund, es necesario dividir τ_0 entre $(n-p)^{1/2}$

2.3.2.2. Método de Cook y Prescott

Considérese el modelo de regresión lineal con las suposiciones de costumbre y recuérdense las definiciones de r_i y τ_i dadas en (2.1.9) (2.3.5) respectivamente. En este método se parte nuevamente de la desigualdad (2.3.7) con $c_\alpha = d$ y $\alpha_i = Pr(|\tau_i| > d)$, $\alpha_{ij} = Pr(|\tau_i| > d, |\tau_j| > d, i \neq j)$, entonces esta desigualdad se transforma a

$$\sum_i \alpha_i - \sum_{i<j} \alpha_{ij} \leq Pr(\tau^* > d) \leq \sum_i \alpha_i. \qquad (2.3.12)$$

Ya que bajo la hipótesis de que no hay observaciones discrepantes, los valores $|\tau_i|$ se distribuyen idéntica-

mente y entonces la cota superior de (2.3.12) puede expresarse convenientemente en términos de una variable aleatoria F cuya distribución es $F(1, n-p-1)$ es decir,

$$\sum_i \alpha_i = nPr(\tau_i^2 > d^2) = nPr[F > d^2(n-p-1)/(1-d^2)].$$

La evaluación de la cota inferior de (2.3.12) es, como se vió anteriormente, más difícil ya que las α_{ij}'s deben determinarse por integración (Ellenberg 1976). En su lugar, se usará una aproximación para las α_{ij}'s como siguen. Ya que la distribución de (τ_i, τ_j) es simétrica.

$$\alpha_{ij} = 2Pr(\tau_i > d, \tau_j > d) + 2Pr(\tau_i > d, -\tau_j > d)$$
$$P(\tau_i > d, \pm\tau_j > d) \leq P(\tau_i \pm \tau_j > 2d),$$

entonces

$$\alpha_{ij} \leq 2Pr(\tau_i + \tau_j > 2d) + 2Pr(\tau_i - \tau_j > 2d)$$
$$= Pr[(\tau_i + \tau_j)^2 > 4d^2] + Pr[(\tau_i - \tau_j)^2 > 4d^2]$$
$$= \beta_{ij}^+ + \beta_{ij}^-; \text{ naturalmente } \beta_{ij}^+ = Pr[(\tau_i + \tau_j)^2 > 4d^2]$$
$$\text{y } \beta_{ij}^- = Pr[(\tau_i - \tau_j)^2 > 4d^2.$$

Este resultado en combinación con (2.3.12) dan como resultado

$$\alpha - \sum_{i<j} (\beta_{ij}^+ + \beta_{ij}^-) \leq Pr(t^* > d) \leq \alpha, \qquad (2.3.13)$$

$$\alpha = \sum_i \alpha_i .$$

Entonces la evaluación de la cota inferior en la desigualdad anterior es sencilla si se reconoce que

$$(\tau_i \pm \tau_j)^2 (n-p-1)/[2(1 \pm \rho_{ij}) - (\tau_i \pm \tau_j)^2] \qquad (2.3.14)$$

tiene distribución $F(1; n-p-1)$. Se puede usar (2.3.14) para demostrar que

$$\beta_{ij}^{\pm} = 0 \quad \text{cuando} \quad 2d^2 > (1 \pm \rho_{ij}), \tag{2.3.15}$$

que es la misma condición dada en (2.3.10). Sea $c(\pm) = \{(i,j) / i<j,\ 2d^2 < (1 \pm \rho_{ij})\}$, usando (2.3.13), (2.3.14) y (2.3.15), la desigualdad de Conferroni resulta ser

$$\alpha - \beta^+ - \beta^- \leq P_r(T^* > d) \leq \alpha \tag{2.3.16}$$

con: $\alpha = nP_r[F > d^2(n-p-1)/(1-d^2)]$,

$$\beta^+ = \sum_{c(+)} P_r[F > d^2(n-p-1)/(\tfrac{1}{2}(1+\rho_{ij}) - d^2)],$$

$$\beta^- = \sum_{c(-)} P_r[F > d^2(n-p-1)/(\tfrac{1}{2}(1-\rho_{ij}) - d^2)].$$

Comentarios finales sobre este procedimiento

i) Se requiere calcular las ρ_{ij}'s ya que la cota inferior en (2.3.16) depende de la distribución conjunta de (T_i, T_j) a través de estas.

ii) La cota superior en (2.3.16) es exacta si $c(+)$, $c(-)$ son vacíos o equivalentemente $1 + \max_{i<j} |\rho_{ij}| < 2d^2$

iii) En algunos casos se puede aproximar la cota inferior reemplazando en β^+ y β^- a ρ_{ij} por $\max_{c(+)} \rho_{ij}$ y $\min_{c(-)} \rho_{ij}$ respectivamente.

iv) Para una prueba de una cola, reemplazar en (2.3.5) la cota superior α por $\alpha/2$ y a la cota inferior por $-(\alpha - \beta^+)/2$.

2.3.2.3 Método de Doornbos.

Este método tiene la misma finalidad que el anterior, es decir, por medio de la desigualdad (2.3.8) encontrar cotas para la distribución de la estadística de prueba (2.3.6) para probar la hipótesis de que el modelo (2.2.2) es correcto contra la alternativa para alguna i desconocida $E(y_i) = x_i^t \beta + \lambda$.

Teniendo en cuenta $\phi(\tau_i, \tau_j)$ definida en (2.3.9) y usando la transformación

$$t_i = \tau_i,$$
$$t_j = (-c_{ij}\tau_i + \tau_j)(1-\tau_i^2)^{-\frac{1}{2}}(1-\rho_{ij}^2)^{-\frac{1}{2}},$$

cuya distribución conjunta (Doornbos y Prints (1958)) es

$$g(t_i, t_j) = \frac{\gamma}{\pi}(1-t_i^2)^{\gamma-\frac{1}{2}}(1-t_j)^{\gamma-\frac{1}{2}}$$

$$= \frac{\Gamma(\gamma+1)}{\Gamma(\frac{1}{2})\Gamma(\gamma+\frac{1}{2})}(1-t_i^2)^{\gamma-\frac{1}{2}} \cdot \frac{\Gamma(\gamma+\frac{1}{2})}{\Gamma(\frac{1}{2})\Gamma(\gamma)}(1-t_j^2)^{\gamma-1}.$$

En otras palabras t_i, t_j son independientes. Por otro lado

$$P(|\tau_i| \geq c_\alpha, |\tau_j| \geq c_\alpha) = 2P(\tau_i \geq c_\alpha, \tau_j \geq c_\alpha) + 2P_r(\tau_i \geq c_\alpha, \tau_j \leq -c_\alpha)$$

$$= 2P_1 + 2P_2.$$

Supongamos por ahora que $\rho_{ij} \geq 0$, entonces

$$P_i = P\{t_i \geq c_\alpha,\ t_j \geq (-\rho_{ij} t_i + c_\alpha)(1-t_i^2)^{-1/2}(1-\rho_{ij}^2)^{-1/2}\}$$

ahora $\frac{\partial}{\partial t_i}(-\rho_{ij} t_i + c_\alpha)(1-t_i^2) = (-\rho_{ij} + t_i c_\alpha)(1-t_i^2)^{-\frac{3}{2}}$

que será no negativa para toda $t_i \geq c_\alpha$ si

$$\rho_{ij} \leq c_\alpha^2. \qquad (2.3.17)$$

Suponiendo que este es el caso y como t_i y t_j son independientes entonces podemos escribir

$$P_i \leq P_r\{t_i \geq c_\alpha,\ t_j \geq (-\rho_{ij} c_\alpha + c_\alpha)(1-c_\alpha^2)^{-\frac{1}{2}}(1-\rho_{ij}^2)^{-\frac{1}{2}}$$

$$= P_r(t_i \geq c_\alpha) \cdot P_r[t_j \geq (-\rho_{ij} c_\alpha + c_\alpha)(1-c_\alpha^2)^{-1/2}(1-\rho_{ij}^2)^{-1/2}]. \qquad (2.3.18)$$

Definamos c_α^* tal que $P_r(|t_i| \geq c_\alpha^*) = \frac{\alpha}{n}$. Debido a que el cociente de las densidades de t_i y t_j decrece con t^2, los percentiles superiores de t_i serán mayores que los correspondientes a t_j. En otras palabras, $c_\alpha^* > c_\alpha$

Ahora calculemos los valores de las ρ_{ij}'s para las que

$$c_\alpha(1-\rho_{ij})(1-c_\alpha^2)^{-\frac{1}{2}}(1-\rho_{ij}^2)^{-\frac{1}{2}} \geq c_\alpha^*$$

o equivalentemente

$$\rho_{ij} \leq \frac{c_\alpha^2 - c_\alpha^{*2} + c_\alpha^* c_\alpha}{c_\alpha^2 + c_\alpha^{*2} - c_\alpha^* c_\alpha} = \rho_{j\alpha}.$$

Además, usando el hecho de que $c_\alpha < c_\alpha^*$ se puede demostrar que

$$g_\alpha < c_\alpha^2,$$

consecuentemente (2.3.19) implica (2.3.17). teniendo en mente que $c_\alpha^*(n,K) = c_\alpha(n, K+1)$, donde $c_\alpha(n, K+1)$ se obtiene de la tabla de Lund al dividir los valores correspondientes por $(n-\rho)^{1/2}$. Entonces c_α^* y g_α se pueden obtener de estas tablas o bien, usando las tablas de Doornbos (1980)

Regresando a (2.3.18) se observa que para ρ_{ij} positiva, si (2.3.19) se cumple, entonces

$$P_1 \leq P(t_i \geq c_\alpha)\, P(t_j \geq c_\alpha^*) = (\tfrac{1}{2}\ \alpha/n)^2. \qquad (2.3.20)$$

Para $\rho_{ij} < 0$, (2.3.20) se verifica fácilmente observando (2.3.17) y (2.3.19). Procediendo de la misma forma.

$$P_2 \leq (\tfrac{1}{2}\ \alpha/n)^2$$

si ρ_{ij} es positiva, o si es negativa, pero $-\rho_{ij} \leq g_\alpha$

Combinando estos resultados obtenemos:

$$Pr(|t_i| \geq c_\alpha, |t_j| \geq c_\alpha) = 2P_1 + 2P_2 \leq 4(\tfrac{1}{2}\ \alpha/n)^2 = (\tfrac{\alpha}{n})^2 \qquad (2.3.21)$$

si

$$|\rho_{ij}| \leq g_\alpha, \qquad (2.3.22)$$

consecuentemente

$$\sum Pr(|t_{i1}| \geq c_\alpha, |t_{j1}| \geq c_\alpha) < \binom{n}{2}\left(\frac{\alpha}{n}\right)^2 < \frac{1}{2}\alpha^2, \qquad (2.3.23)$$

de donde, la desigualdad (2.3.8) puede expresarse como

$$\alpha - \tfrac{1}{2}\alpha^2 \leq Pr(t^* \geq c_\alpha) \leq \alpha \qquad (2.3.24)$$

si

$$\max_{(i,j)} |\rho_{ij}| \leq g_\alpha. \qquad (2.3.25)$$

Note que si (2.3.25) no se cumple, pero (2.3.22) es válida, excepto para pocas parejas (i,j), (2.3.21) - sigue siendo útil para proporcionar una cota inferior aceptable para $Pr(t^* \geq c_\alpha)$.

2.3.2.4. Método de Weisberg

Considerese el modelo de regresión lineal con las hipótesis usuales. Una vez que se cuenta con los elementos necesarios para suponer que la i-ésima observación es discrepante, se sugiere seguir los siguientes pasos para su determinación.

1) Elimine el i-ésimo caso del conjunto de datos.
2) Calcular $\hat{\beta}_{(i)}$ y $s^2_{(i)}$ que son los estimadores de β y σ^2 usando solo $(n-1)$ casos. $s^2_{(i)}$ tiene $(n-1)-p = n-p-1$ grados de libertad.
3) Para el caso eliminado i, calcule $\tilde{y}_i = x_i^t \hat{\beta}_{(i)}$ para predecir y_i, debido a que este caso no se uso en la estimación de β. Como el i-ésimo caso fué excluido antes de la estimación, y_i y $\tilde{y}_i$ (y_i y $s^2_{(i)}$) son independientes. La

varianza de $\tilde{y}_i$ está dada por

$$\text{var}(\tilde{y}_i) = \sigma^2 \underline{x}_i^2 (x_{(i)}^t X_{(i)})^{-1} \underline{x}_i$$

cuyo estimador está dado por

$$\widehat{\text{var}}(\tilde{y}_i) = S_{(i)}^2 \underline{x}_i^t (x_{(i)}^t X_{(i)})^{-1} \underline{x}_i .$$

4) Ahora, si y_i no es una observación discrepante, $E(y_i - \tilde{y}_i) = 0$, pero si y_i es una observación discrepante $E(y_i - \tilde{y}_i) \neq 0$. La diferencia $y_i - \tilde{y}_i$ tiene varianza $\text{var}(y_i - \tilde{y}_i) = \sigma^2 + (1 - \underline{x}_i^t (x_{(i)}^t X_{(i)})^{-1} \underline{x}_i)$ y como se supone que los errores tienen distribución normal, una prueba t-student de la hipótesis "$E(y_i - \tilde{y}) = 0$" está dada por

$$t_i = \frac{y_i - \hat{y}_i}{S_{(i)} \sqrt{1 + \underline{x}_i^t (x_{(i)}^t X_{(i)})^{-1} \underline{x}_i}} .$$

Esta prueba tiene $n-p-1$ grados de libertad.

Una prueba que ayuda a reducir cálculos es

$$t_i = r_i \left(\frac{n-p-1}{n-p-r_i^2} \right)^{1/2} .$$

Naturalmente, está estadística es útil tanto en la versión etiquetada como en la no etiquetada. En el primer caso t_i puede compararse directamente con el valor obtenido en la tabla de distribución t-student usual, mientras que en el segundo caso debe de hacerse una modificación.

Usualmente, el investigador no tiene conocimiento de cual observación es discrepante. Si probamos el caso con el valor máximo de t_i, en realidad estamos efectuando n pruebas de significancia, una para cada caso. Supongamos por ejemplo, que no hay observación discrepante y que $n=65$, $p=5$. La probabilidad que una estadística con 60 grados de libertad exceda a 2.00 es 0.05; sin embargo la probabilidad de que la mayor de las 65 pruebas t independientes se exceda a 2.00 es 0.924. Lo anterior sugiere la necesidad de un valor crítico adecuado. La tecnica que usamos para encontrar valores críticos se basa en la primera desigualdad de Bonferroni, que establece que para n pruebas, cada una de tamaño α, la probabilidad de señalar falsamente una observación como discrepante no es mayor que $n\alpha$. Este procedimiento es conservador, ya que la primera desigualdad de Bonferroni debería de especificar solamente que la probabilidad de que el máximo de 65 pruebas que se exceden en 2.00 no es mayor que $65(0.05)$, que es mayor que 1. Sin embargo, escogiendo el valor crítico como el punto $(\alpha/n)\times 100$ de la distribución t dará un nivel de significancia no mayor que $n(\alpha/n)=\alpha$. Entonces deberíamos escoger un nivel de $0.05/65 = 0.00077$ para cada prueba para dar un nivel global no mayor que $(65)\times(0.00077)=0.05$. La tabla B al final de este trabajo es útil para esta prueba.

Esta prueba puede describirse de la siguiente manera: rechazar la hipótesis de que $E(y_i - \hat{y}_i)=0$ si

$\max_i |t_i| > t(\alpha/n, n-p-1)$ para un nivel nominal α. Cook y Weisberg (1980) sugieren una regla alternativa: rechazar la hipótesis anterior si

$$\max_i |t_i| > t\left(\sqrt{v_{ii}}\,\alpha/p \;;\; n-p-1\right)$$

donde afirman que esta regla mantiene el nivel de significancia total, pero proporciona un incremento en la potencia de la prueba en casos donde v_{ii} es "grande".

2.3.2.5. Sobre la sensibilidad de las pruebas en regresión cuando no hay normalidad en los errores.

Muchas de las pruebas aplicadas para observaciones discrepantes en el modelo de regresión lineal con las suposiciones de costumbre, se basan en los r_i o una función de ellos. Miyashita y Newbold (1983) por medio de simulación y usando como estadística de prueba la expresada en (2.3.6) proporcionan información acerca del comportamiento de las probabilidades en la cola de la distribución de esa estadística cuando la distribución del error no es normal. A continuación se describe este trabajo.

Para evaluar el efecto de no normalidad en el comportamiento de la prueba para observaciones discrepantes, se considerará una clase de distribuciones conocida como "distribuciones potencia exponencial".

Específicamente, si suponemos que los errores ε_i están identicamente distribuidas con funciones de densidad

$$f(\varepsilon_i/\sigma, \theta) = w(\theta)\,\sigma^{-1} \exp\left[-c(\theta)\left|\frac{\varepsilon_i}{\sigma}\right|^{2/(1+\theta)}\right], \qquad (2.3.26)$$

con

$$c(\theta) = \left[\frac{\Gamma\left[\frac{3}{2}(1+\theta)\right]}{\Gamma\left[\frac{1}{2}(1+\theta)\right]}\right]^{1/(1+\theta)}$$

y

$$w(\theta) = \frac{\left[\Gamma\left(\frac{3}{2}(1+\theta)\right)\right]^{1/2}}{(1+\theta)\left[\Gamma\left(\frac{1}{2}(1+\theta)\right)\right]^{3/2}}$$

donde σ es la desviación estándar de la distribución de ε, mientras que el parámetro θ puede tomar valores entre -1 y 1 y puede considerarse como una medida de curtosis que indica el grado de la "no-normalidad" de la población principal. En particular, cuando $\theta = 0$, la distribución es normal. Cuando $\theta = 1$, la distribución es doble exponencial y finalmente se puede demostrar que cuando θ tiende a -1, la distribución es rectangular, ver figura (2-3.1) donde las distribuciones tienen la misma varianza. Podemos observar que para $\theta < 0$ y $\theta > 0$ las distribuciones son platicúrticas y leptocúrticas respectivamente.

El proceso de simulación se realiza con un valor de $\sigma = 1$ y manejando primeramente el modelo de regresión simple

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \qquad (i = 1, 2, \ldots n),$$

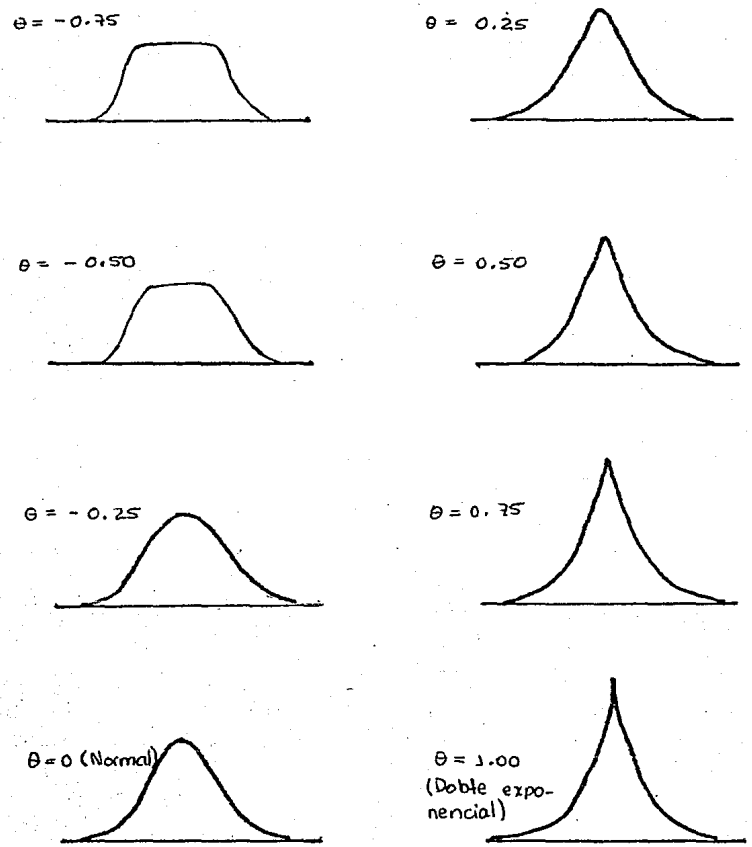


Figura 2.3.1. Distribuciones potencia exponencial con la misma varianza para varios valores de θ

con muestras de tamaño $n = 10, 20, 40$ y 2000 repeticiones para los conjuntos

$A: x_1 = x_2 = 1$, con los demas valores $x_i = 0$; $i = 3, 4, \ldots, n$

$B: x_1 = 1, x_2 = 0$, $x_i = 0.5$ para $i = 3, 4, \ldots, n$

C : $\frac{n}{2}$ de los $x_i = 1$ y las restantes con valor $x_i = 0$

D : x_i con distribución uniforme en el intervalo $0, 1$

La tabla 2.3.1 muestra la proporción de veces que se rechaza la hipótesis nula "no hay observaciones discrepantes" en base a la estadística (2.3.10) con un nivel de significancia $\alpha = 0.05$. De tal forma que las entradas en el cuerpo de la tabla representan niveles de significancia empíricos basados en el máximo residual studentizado en valor absoluto cuando se supone que los errores tienen distribución normal.

Cualitativamente, los resultados en la tabla (2.3.1) son como se esperaban. Para distribuciones platicúrticas, los niveles de significancia empíricos están por abajo del valor supuesto, mientras que para distribuciones leptocúrticas, éstos están por arriba del valor $\alpha = 0.05$.

Mas aún, cuando el tamaño de la muestra crece en distribuciones leptocúrticas, también aumenta este nivel de significancia. Esto se esperaba, ya que para ma-

X	n	θ -0.75	-0.50	-0.25	0	0.25	0.50	0.75	1
A	10	.0205	.0185	.0265	.0470	.0690	.0900	.1060	.1310
B	10	.0160	.0210	.0300	.0510	.0630	.0875	.1100	.1330
C	10	.0215	.0285	.0320	.0515	.0675	.0880	.1050	.1250
D	10	.0200	.0305	.0345	.0505	.0650	.0860	.1145	.1305
A	20	.0020	.0065	.0205	.0390	.0890	.1310	.1810	.2150
B	20	.0010	.0070	.0195	.0500	.0880	.1380	.1755	.2270
C	20	.0045	.0085	.0255	.0495	.0990	.1435	.1790	.2210
D	20	.0035	.0070	.0200	.0520	.0920	.1435	.1850	.2265
A	40	.0000	.0015	.0105	.0545	.1120	.1885	.2420	.3015
B	40	.0000	.0015	.0100	.0565	.1150	.1875	.2505	.2955
C	40	.0000	.0010	.0135	.0565	.1135	.1780	.2310	.3055
D	40	.0000	.0015	.0120	.0560	.1175	.1970	.2370	.3070

tilde (2.3.1)

					θ				
X	n	-0.75	-0.50	-0.25	0	0.25	0.50	0.75	1
A	10	.0205	.0185	.0265	.0470	.0690	.0900	.1060	.1310
B	10	.0160	.0210	.0300	.0510	.0630	.0875	.1100	.1330
C	10	.0215	.0285	.0320	.0515	.0675	.0880	.1050	.1250
D	10	.0200	.0305	.0345	.0505	.0650	.0860	.1145	.1305
A	20	.0020	.0065	.0205	.0390	.0890	.1370	.1810	.2150
B	20	.0010	.0070	.0195	.0500	.0880	.1380	.1755	.2270
C	20	.0045	.0085	.0255	.0475	.0990	.1435	.1790	.2210
D	20	.0035	.0080	.0200	.0520	.0920	.1435	.1850	.2265
A	40	.0000	.0015	.0105	.0545	.1120	.1885	.2420	.3015
B	40	.0000	.0015	.0100	.0565	.1150	.1875	.2505	.2955
C	40	.0000	.0010	.0135	.0565	.1135	.1780	.2310	.3055
D	40	.0000	.0015	.0120	.0560	.1175	.1870	.2370	.3070

table (2.3.1)

yor tamaño de muestra, mayor es la probabilidad — de encontrar errores en las colas de distribuciones leptocúrticas. Así, si la suposición de normalidad no es válida estos errores podrían señalarse como discrepantes.

Cuantitativamente los resultados son motivo de - alguna preocupación, enfatizando la importancia de - la suposición de normalidad, particularmente cuando la muestra es de un tamaño moderadamente gran de. En el caso extremo de una distribución de error doble exponencial, los niveles de significancia empíricos están alrededor del 13%, 22% y 30% para muestras de tamaño 10, 20 y 40 respectivamente. Entonces, aún para muestras de tamaño moderado, es bastante probable que una observación sea señalada (erroneamente) como discrepante si la distribución del error es doble exponencial. Quizas sea de mayor - importancia la influencia substancial sobre los niveles de significancia de las menores desviaciones de normalidad en los errores cuando la muestra contiene alrededor de 40 observaciones. Aquí, para $\theta = -0.25$ los niveles de significancia son del orden de 1.25% — mientras que para $\beta = 0.25$ están alrededor del 11.5%. Esto, indica que la prueba tiene propiedades robustas muy pobres en muestras moderadamente - grandes

Ahora, consideremos el modelo de regresión lineal múltiple

$$Y_i = X_{i1}\beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \quad (i = 1, 2, \ldots, n).$$

con $X_{i1} = 1$ para los conjuntos A, B, C, D definidos anteriormente. En este modelo, como matrices de diseño se escogieron las 4 combinaciones $(X_2 : X_3) = (A : B)$, $(B : C)$, $(C : D)$, $(D : A)$ y nuevamente se realizaron 2000 repeticiones con muestras de tamaño 10, 20, 40 para cada una.

Aunque los autores no muestran los resultados de este procedimiento de simulación, reportan que los niveles de significancia empiricos son muy similares a los contenidos en la tabla (2.3.1). Esto sugiere que el número de variables independientes tiene muy poco efecto en el comportamiento de la prueba estadística para estas distribuciones del error también afirman que los resultados son muy similares cuando se proponen matrices de diseños diferentes

Los métodos expuestos anteriormente, se refieren a pruebas para la identificación de una observación ($k=1$). A continuación se presentan algunas pruebas que dominan la literatura para la identificación de mas de una observación discrepante ($k > 1$).

2.3.2.6 Método de Gentleman y Wilk.

Usando el modelo de la media trasladada (2.2.1) este método se resume en los siguientes pasos:

i) Escoger k, número máximo de observaciones discrepantes.

ii) Calcular $Q_K(i)$ para cada una de las $\binom{n}{k}$ particiones de los datos.

iii) Sea $Q_K^*(i) = \max_{\binom{n}{k}} Q_K(i)$. Se dice que el subconjunto al que le corresponde $Q_K^*(i)$ es "el subconjunto de las k observaciones que probablemente son las mas discrepantes".

iv) Como la distribución de $Q_K^*(i)$ es desconocida, se proponen métodos informales para evaluar la significancia de esta. Una ayuda para juzgar cuando la magnitud de $Q_K^*(i)$ es suficientemente grande con respecto a otros valores $Q_K(i)$ es una forma de "diagrama de probabilidad" en el que, so ó 100 de los valores mas grandes de $Q_K(i)$ se grafican contra "valores típicos", por ejemplo la mediana, de los correspondientes $Q_K(i)$ mas grandes que se obtienen por simulación bajo el supuesto de que no hay observaciones discrepantes

v) Si se concluye que $Q_K^*(i)$ no es estadísticamente significante, entonces el procedimiento anterior se repite para $(K-1)$ y así sucesivamente, caso contrario, el método se da por terminado señalando a las K observaciones discrepantes en los datos.

Como se puede constatar este método es muy laborioso, es mas, difícilmente se podría efectuar sin la ayuda de una computadora. Con la finalidad de reducir esta carga computacional, Gentleman y Wilk (1975) demuestran que

$$Q_k(i) = \underline{e}_i^t (I - V_i)^{-1} \underline{e}_i$$

donde $\underline{e}_i$ y V_i son respectivamente, el subvector de $\underline{e}$ y la submatriz de $k \times k$ de V indicados por i. Entonces el cálculo de $Q_k(i)$, bajo la suposición de normalidad, requiere únicamente de los residuales ordinarios y de la matriz de proyección V

Gentleman (1980) con la misma finalidad, propone explotar la estructura de la matriz V, por ejemplo, usa la descomposición Cholesky de $I - V_i$; Existe una matriz triangular $R_{k \times k}$ tal que $R^T R = I - V_i$. Entonces si resolvemos para $\underline{a}$; $R\underline{a} = \underline{e}_i$ podemos calcular

$$Q_k(i) = \underline{a}^T \underline{a}$$

que es una forma cuyas propiedades computacionales la hace mas deseable.

Alternativamente, Cook y Weisberg (1982) usando la descomposición espectral de $V_i = \Gamma \Lambda \Gamma$, donde Γ es la matriz ortogonal cuyas columnas son los vectores característicos y Λ es la matriz diagonal cuyos elementos son los valores característicos correspondientes, demuestran que

$$Q_k(i) = (\Gamma^t \underline{e}_i)^t (I - \Lambda)^{-1} (\Gamma^t \underline{e}_i).$$

2.3.2.7 Método de Andrews y Pregibon.

Este método alternativo es capaz de encontrar a

las observaciones discrepantes y/o influyentes mas importantes, es decir, identifica varias observaciones discrepantes, con atención directa en aquellas que tienen una influencia substancial en los estimadores resultantes. Para la construcción de esta estadística se combinan 2 elementos de diagnóstico. Inicialmente, considera los efectos de una observación discrepante en $\underline{Y}$ y del renglón correspondiente en forma separada. Primero, al eliminar el caso correspondiente a una observación discrepante en $\underline{Y}$, tiende a reducir marcadamente la suma de cuadrados del residual. La suma de cuadrados del residual es, por lo tanto, un diagnóstico para detectar casos influyentes que surgen debido a una observación discrepante en $\underline{Y}$. Segundo, la influencia de un renglón de X está en parte reflejado por el cambio en $|X^TX|$ cuando se suprime éste. Si $|X^TX|$ cambia substancialmente cuando $\underline{x}_i^2$ se suprime, entonces el caso correspondiente $(y_i, \underline{x}_i^t)$ tendrá una gran influencia en $\hat{\beta}$.

Andrews y Pregibon (1978) sugieren estos diagnósticos separados, que se basan en el cambio en la suma de cuadrados del residual y en $|X^TX|$ que al combinar los nos dan un diagnóstico que se base en el cambio de $(n-p)S^2|X^TX|$ que resulta de la eliminación de uno o mas casos. Específicamente se sugiere

$$\tilde{R}_i = \frac{|X_{(i)}^{*t} X_{(i)}^{*}|}{|X^{*t} X^{*}|} \qquad (2.3.30)$$

donde $X^* = (X/\underline{y})$, es decir, la matriz X aumentada

con Y. Esta cantidad es adimensional. Geometricamente, $R_i^{-1/2} - 1$ corresponde al cambio proporcional en el volúmen del elipsoide generado por $X^{*t} X^*$ cuando se eliminan los casos indicados por i. Así, valores "pequeños" de R_i estarán asociados a observaciones discrepantes y/o influyentes. Sin importar que tipo de observaciones se están considerando, es adecuado aislar subconjuntos de observaciones que produzcan valores pequeños de R_i para un análisis adicional.

Draper y John (1981), Cook y Weisberg (1982) demuestran que

$$\begin{aligned} R_i &= (1 - Q_K(i)/\underline{e}^t\underline{e}) \, |I - V_i| \\ &= \frac{(n-p-K)\, s^2_{(i)} \, |X^t_{(i)} X_{(i)}|}{(n-p)\, s^2 \, |X^t X|} \\ &= \frac{(n-p-K)\, s^2_{(i)}}{(n-p)\, s^2} \, |I - V_i| \\ &= \left(1 - \frac{r_i^2}{n-p}\right) \, |I - V_i| \end{aligned} \qquad (2.3.31)$$

donde $r_i^2 = \frac{\underline{e}_i^2 (I - V_i)^{-1} \underline{e}_i}{s^2} = \frac{Q_K(i)}{s^2}$ es el caso multiple de los residuales studentizados internos definidos en (2.1.9). Estas expresiones, ademas de simplificar un poco los cálculos de R_i, permiten las siguientes interpretaciones de sus componentes, (Draper y John 1981): El primer factor será pequeño si $Q_K(i)$ es grande, en este caso podemos identificar conjuntos de observaciones discrepantes como en Gentleman y Wilk (1975), el segundo factor $(I - V_i)$ proporciona una medida de influencia de las K

observaciones en el ajuste, valores pequeños de $|I - V_i|$ indicarán los "puntos mas remotos" en el espacio de los valores observados. Cuando $K=1$, este factor se reduce a $1 - v_{ii}$, asi, cuando $|1 - v_{ii}|$ sea "pequeño" se debe a que v_{ii} es grande, que como se vió en la sección 2.1.1 corresponde a puntos "altamente influyentes".

Bajo el supuesto de normalidad, $(n-p-K)\, s^2_{(i)} / (n-p)s^2$ tiene distribución Beta con parámetros $(n-p-K)/2$, $K/2$. Asi R_i es proporcional a una variable aleatoria Beta, donde $|I - V_i|$ es la constante de proporcionalidad, de tal manera que, basados en momentos se puedan obtener valores de referencia.

Se pueden encontrar valores exactos de probabilidad para valores extremos de

$$R_i^* = \min_{\binom{n}{k}} R_i . \qquad (2.3.32)$$

El método que se propone es el de Andrews (1971). Esto es, si R_i° es el mínimo valor observado de los R_i, el conjunto de valores mas extremos está dado por $\{ R_i / R_i^* \leq R_i^\circ \}$. El nivel de significancia es la probabilidad de este conjunto, y está dado por

$$\alpha_K = Pr\{ R_i / R_i^* \leq R_i^\circ \} = Pr\left\{ \bigcup_{J \in I \in \binom{n}{p}} Z_I \leq \xi_{(1)} \right\} \qquad (2.3.33)$$

donde Z_I tiene distribución Beta y $\xi_{(1)} = R_i^\circ / R_i(x)$; $R_i(x) = \frac{|X_{(i)}^T X_{(i)}|}{|X^T X|} \neq 0$.

Nótese que los pesos diferenciales $R_i(X)$ incrementan fuertemente la significancia de observaciones discrepantes influyentes. Para conjuntos pequeños de datos ($n \leq 30$) es deseable calibrar la "pequeñez" de los R_i mínimos, vía pruebas de significancia. Sin embargo para conjuntos grandes de datos, solamente tiene un pequeño efecto, mientras que observaciones discrepantes pueden tener un gran efecto y el número de éstos es probablemente mayor. Por consiguiente, en estos casos se hace más énfasis en la detección, que en la evaluación de un nivel de significancia. En la práctica la expresión

$$R_i = |M_i^*| = |I - V_i^*|$$

que es un determinante de $K \times K$, puede incorporarse a un procedimiento secuencial que se basa en $1, 2, \ldots, K_{max}$ posibles observaciones discrepantes, donde K_{max} es el número máximo de observaciones discrepantes, este número lo escoge el analista.

2.3.2.8 Método de Aitkin y Wilson

Este método aplica el algoritmo EM (Dempster, Laird y Rubin, 1977) a modelos de la forma

$$F(y) = \sum_{i=1}^{m} p_i f_i(y) \qquad (2.3.34)$$

donde $f_i(y)$ $i = 1, 2, \ldots, m$, son funciones de densidad normal, donde posiblemente las medias y/o varianzas correspondientes son diferentes. El algoritmo comienza

con estimadores iniciales máximo verosimiles de los parámetros (basados convenientemente en la especificación de una o mas observaciones discrepantes), como el primer paso M, entonces calcular $\pi_j(y_i)$, la probabilidad de que la i-ésima observación provenga de la j-ésima componente; que se basa en los estimadores iniciales, paso E. Usando estas probabilidades, calcular los nuevos estimadores mínimos cuadrados ponderados de los estimadores para un nuevo paso M y asi sucesivamente hasta alcanzar la convergencia. La maximización de la función log verosimilitud puede realizarse y compararse con la maximización log-verosimil para una muestra y dar indicación de la existencia de una mezcla.

La programación del algoritmo es muy facil (Aitkin y Wilson 1980) proporciona los estimadores máximo verosimiles de los parámetros, incluyendo la proporción de cada componente en la mezcla. La matriz de covarianza asintótica y la función log verosimilitud maximizada pueden servir para proporcionar una indicación del número de componentes en la mezcla.

Por ejemplo, considere el caso donde $F(y)$ tiene 2 componentes, es decir

$$F(y_i) = p f_1(y_i) + (1-p) f_2(y_i) \qquad (2.3.35)$$

donde $$f_1(y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[- \frac{(y_i - \sum_{r=0}^{p} \beta_r x_{ri})^2}{2\sigma^2} \right]$$

y $$f_2(y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[- \frac{y_i - \mu}{2\sigma^2} \right]$$

Las ecuaciones máximo verosimil son

$$\hat{p} = \sum_i^n (\hat{P}(1/y_i))/n)$$

$$\hat{\underline{e}} = (x^T W x)^{-1} x^T W \underline{y}$$

$$\hat{\mu} = \sum_i y_i \hat{P}(2/y_i) / \sum_i \hat{P}(2/y_i)$$

$$S^2 = \sum_{j=1}^{2} \sum_i^n \{e_{ji}^2 \hat{P}(j/y_i)/n\}$$

donde $e_{ij} = y_i - \sum_{r=1}^{p} \hat{\beta}_r x_{ri}$ y $e_{2i} = y_i - \hat{\mu}$, x y $\underline{y}$ elementos del modelo (2.2.1) y $W = \text{diag}(\hat{P}(1-y_i))$. Entonces $\hat{\underline{e}}$ es un estimador minimos cuadrados ponderados cuyos pesos son $\hat{P}(1/y_i)$ y

$$-2\log L = n \log S^2 - 2 \sum_i^n \log\{\hat{p} \exp(-e_{1i}^2/2S^2) + (1-\hat{p}) \exp(-e_{2i}^2/2S^2)\}.$$

2.3.2.9 METODO DE MARASINGHE.

El procedimiento que aquí se propone, se basa en el uso de residuales studentizados para construir una estadística F_k similar a la estadística E_k propuesta por Tietjen y Moore (1972) para muestras simples, modificado por Rosner y para la cual Hawkins (1979) proporciona una tabla de valores criticos para la prueba de k observaciones discrepantes. Para generalizar esta estadística y usarla en la detección de k observaciones discrepantes en regresión, se obtiene un subconjunto de k observaciones donde la primer observación que se incluye, es la correspondiente al máximo residual studentizado en valor absoluto que resulta de ajustar el modelo de

regresión con las hipótesis usuales.

Después, se elimina esta observación y se realiza el ajuste con estas $(n-1)$ observaciones usando el mismo modelo. La segunda observación que ingresa a dicho subconjunto se obtiene de la misma forma que la primera. Este procedimiento de eliminar observaciones y calcular residuales studentizados con las observaciones restantes se continúa hasta obtener el subconjunto deseado (naturalmente, κ se determina de antemano). Sea $Q_\kappa^{**}(i)$, la reducción en la suma de cuadrados que resulta de eliminar las observaciones indicadas por i, obtenidas como se indica líneas arriba. Entonces, la estadística de prueba se define por

$$F_\kappa = \frac{S - Q_\kappa^{**}(i)}{S}, \qquad (2.3.36)$$

donde $S=(n-p)s^2$. Se puede demostrar que

$$Q_\kappa^{**}(i) = \sum_{i \in I} t_i^2 \qquad (2.3.37)$$

con $t_i = \frac{e_i}{(1-v_{ii})^{1/2}}$ $\quad i=1,2,\ldots,n$, el i-ésimo residual "ajustado".

Aunque el orden precedente de las observaciones no asegura que el subconjunto de las κ-observaciones escogidas proporcione siempre el valor máximo de $Q_\kappa(i)$ (esto es, que localice las observaciones probablemente mas discrepantes según Gentleman y Wilk (1975)), se puede observar en los resultados de simulación, proporcionado por el

autor, que más del 95% de las veces $Q^{**}_K(i)$ es idéntica a $Q^{*}_K(i) - \max Q_K(i)$. Además, el método de construcción garantiza que cada observación agregada al subconjunto, produciría un incremento máximo en $Q^{**}_K(i)$ para cada $K=1,2,\ldots$

Usando F_K, se rechaza la hipótesis nula: no hay observaciones discrepantes, cuando F_K es menor que un valor crítico específico. En la tabla 2.3.2., se proporcionan valores críticos obtenidos por simulación de F_K para el caso de regresión lineal simple para algunos valores de n y K. Una comparación de estos valores críticos con los obtenidos por Hawkins (ver Hawkins 1980, tabla A7.1) para la estadística E_K de Rosner, indica que éstos pueden usarse como una aproximación razonable para la estadística F_K con la muestra reducida en uno, debido al ajuste extra de un parámetro. También se sugiere usar esta tabla para el caso de regresión múltiple tomando como tamaño de muestra n-p+1.

Aunque la estadística F_K puede usarse de la misma forma que se usa E_K para el caso de muestras simples, ésta tiene los mismos problemas asociados con E_K, ver Hawkins (1980), en particular, es suceptible a empantanamiento (declarar más observaciones discrepantes de las existentes) y enmascaramiento (declarar menos observaciones discrepantes de las existentes) dependiendo de si el valor de K es sobrestimado o subestimado. También se ha observado que las pruebas que se basan en estadísticas de bloque como E_K y F_K, tienen bastante potencia cuando K es sobrestimada. Para sacar ventaja de esta propiedad se propone el siguiente método: Seleccione un conjunto de K observaciones para probarse como dis-

tabla 2.3.2, valores críticos de F para $\rho = 2$.

	$k = 2$		$k = 3$		$k = 4$		$k = 5$	
n	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$
6	.0007	.0041						
7	.0006	.0209						
8	.0196	.0487	.0025	.0084				
9	.0403	.0834	.0086	.0219				
10	.0667	.1195	.0203	.0430	.0053	.0125		
11	.0947	.1551	.0358	.0643	.0111	.0239		
12	.1227	.1880	.0511	.0887	.0217	.0390	.0071	.0156
14	.1752	.2522	.0932	.1404	.0471	.0779	.0218	.0398
16	.2292	.3048	.1326	.1906	.0793	.1173	.0444	.0696
18	.2814	.3543	.1725	.2340	.1105	.1570	.0699	.1019
20	.3147	.3939	.2095	.2708	.1399	.1888	.0947	.1318
24	.3828	.4599	.2782	.3451	.2042	.2595	.1524	.1956
28	.4439	.5134	.3406	.4028	.2622	.3173	.1998	.2516
32	.4890	.5529	.3867	.4472	.3065	.3631	.2458	.2994
36	.5242	.5879	.4291	.4859	.3530	.4086	.2937	.3440
40	.5601	.6165	.4632	.5195	.3935	.4439	.3330	.3815
50	.6273	.6771	.5411	.5899	.4714	.5193	.4135	.4596
60	.6742	.7179	.5951	.6388	.5300	.5735	.4748	.5174
70	.7101	.7489	.6371	.6765	.5763	.6160	.5240	.5633
80	.7384	.7733	.6708	.7065	.6138	.6502	.5644	.6007
90	.7614	.7931	.6984	.7311	.6449	.6788	.5989	.6318
100	.7804	.8095	.7214	.7516	.6710	.7022	.6265	.6580

crepantes. Entonces probar con F_K la hipótesis H_1: no hay observaciones discrepantes. Si H_1 se rechaza, se elimina a la observación que tenga el mayor residual ajustado y se efectúa la hipótesis H_2: no hay observaciones discrepantes usando F_{K-1} calculada en base a las $(n-1)$ observaciones restantes, donde $F_{K-1} = (S_1 - Q^{**}_{K-1}(i'))/S_1$ y $Q^{**}_K(i') = \sum_{i'} t_i^2$, i' no contiene el índice de la observación eliminada. Este procedimiento continúa hasta que la hipótesis nula no se rechaza, entonces se determina el número de observaciones discrepantes. Los valores críticos para F_K, que se usan en cada etapa están calculados en base a que n y K están disminuidos en uno con respecto a la etapa anterior. Esto supone que la observación removida en cada etapa en realidad es una observación discrepante, en cuyo caso, el error tipo I es el nivel especificado para la prueba.

2.4 COMENTARIOS.

Analizando el caso donde $K=1$, el problema para la detección de observaciones discrepantes en la versión etiquetada es muy simple. Para la versión no etiquetada, la estadística básica de prueba es el máximo en el valor absoluto de los residuales studentizados cuya distribución, como se menciona, es intratable. El problema se resuelve, como se puede ver al analizar los métodos propuestos, de diversas formas. Por ejemplo, de los métodos no expuestos. Mickey, Dunn y Clark (1967) sugieren que una observación será discrepante si su eliminación induce una gran

reducción en la suma de cuadrados de los residuales; método que implica el cálculo de n ecuaciones de regresión, una por cada observación eliminada. La estadística de prueba que usan.

$$F_i(1, n-p-1) = \frac{S^2 - S_{(i)}}{S^2_{(i)}/(n-p-1)} \quad ; \quad i = 1, 2, \ldots n, \qquad (2.4.1)$$

donde $S^2_{(i)}$ es la suma de cuadrados del error si la i-ésima observación es eliminada. Siguiendo la misma idea, Snedecor y Cochran (1968) proponen la estadística

$$t_i(n-p-1) = \frac{y_i - \hat{y}_{(i)}}{\hat{V}(y_i - y_{(i)})^{1/2}} \qquad i = 1, 2, \ldots n \qquad (2.4.2)$$

donde $\hat{y}_{(i)}$ es el estimador mínimos cuadrados de y_i, calculado al eliminar la i-esima observación. Por cierto Ellenberg (1976) demuestra que las estadísticas (2.4.1), (2.4.2) y (2.3.5) son equivalentes

Otra forma para obtener la distribución de (2.4.11) o cualquier estadística similar es efectuar simulación a gran escala como lo hicieron Tietjen, Moore y Beckman (1973) y Prescott (1975) para (2.3.11) en el caso de regresión lineal simple.

Tal vez la idea más común para resolver dicho problema es el uso de la desigualdad de Bonferroni propuesta por Ellenberg. Dicha desigualdad acota superior e inferiormente la distribución de (2.3.6) y (2.3.11) y los mé-

todos que aquí se presentan para detectar una observación discrepante se basa en ella.

El método de Ellenberg proporciona las dos cotas de la desigualdad de Bonferroni y una de las grandes dificultades que enfrenta es el cálculo de la cota inferior y en particular de $P(|\tilde{r}_i| > c_\alpha \mid |\tilde{r}_j| > c_\alpha)$ ya que ésta requiere del cálculo de algunas de las $\binom{n}{2}$ integraciones, por que $P(|\tilde{r}_i| > c_\alpha, |\tilde{r}_j| > c_\alpha) = 0$ si $c_\alpha \geq \left(\frac{1+|\rho_{ij}|}{2}\right)^{1/2}$. Así, para la aplicación de este método, se requiere aparte de algunas integraciones para el cálculo de probabilidades bivariadas la correlación entre los residuales razón por la cual este proceso no se puede usar como una técnica rápida de diagnóstico.

El cálculo de la cota superior que realiza Lund y la presentación de ésta en tablas para varios niveles de significancia nos permite usar este procedimiento rapidamente. Con respecto al método de Weisberg, él usa la estadística de prueba max de $|t_i|$, donde t_i esta definido en (2.1.12) y para la cual, usando la desigualdad de Bonferroni, calcula la cota superior. Debe de mencionarse que no existen diferencias importantes entre usar el método de Lund o el método de Weisberg para la detección de una observación discrepante. Al final de este trabajo se presentan las tablas de Lund y Weisberg.

Uno de los principales inconvenientes es que la potencia de la prueba puede estar influenciada por la posición de la observación discrepante en espacio de diseño;

ver Ellenberg (1976), Tietjen, Moore y Beckman (1973) entre otros.

Por otro lado los métodos de Cook y Prescott, y el de Doornbos son métodos que al igual que el de Ellenberg nos permiten evaluar la precisión de la desigualdad de Bonferroni, solo que éstos no requieren de integración numérica. En cuanto al método de Cook y Prescott, la idea básica para calcular la cota superior es expresada en términos de una distribucion F con 1 y n-p-1 grados de libertad a partir de la cuál se puede calcular el valor crítico deseado tal que dicha desigualdad se cumple. Para el método de Doornbos la idea básica radica en el hecho de si $\max |\rho_{ij}| \leq g_\alpha$ entonces se garantiza que el tamaño de prueba esta entre $\alpha - \frac{1}{2}\alpha^2$ y α. Por cierto que g_α puede obtenerse mediante la tabla de Lund.

Para la cota de los dos últimos métodos su precisión dependerá de las correlaciones, así, en situaciones en donde $\rho_{ij} = \pm 1$ para alguna $i \neq j$, las cotas nunca serán exactas.

Debe aclararse que, cualquier método de detección de observaciones solo debe considerarse como un elemento en el proceso de análisis de datos y no el único determinante. Una observación con un residual studentizado grande puede ser el resultado de uno o más problemas serios; fallas en las suposiciones del modelo, no normalidad de los errores, etc., puede llevar a la apariencia de uno o mas datos sospechosos. Un análisis de la estructuctura de la muestra ayudaría al analista a evaluar la validez de las

suposiciones básicas del modelo, procurando que el tamaño de muestra sea grande. Con muestras pequeñas es difícil distinguir entre un error grueso y fallas en las suposiciones del modelo, y por lo tanto se debe tener cuidado en atribuir la presencia de un residual studentizado a un error grueso.

Por lo expuesto anteriormente, sería aconsejable que antes de pensar en cualquier prueba para la detección de observaciones discrepantes, deben de verificarse todas las posibles causas que podrían ocacionar que una observación tenga la apariencia de discrepante. En particular debe de checarse la suposición de normalidad que, como se vió en el proceso de simulación realizado por Miyashita y Newbold, es un aspecto fundamental para el buen funcionamiento de la estadística (2.3.11) para detectar observaciones discrepantes y probar la necesidad de una transformación de los datos.

Igualmente, cuando $K>1$ la versión etiquetada no ofrece dificultad alguna para su prueba. Por otro lado, se mencionó al principio de la sección 2.3.2. que la versión no etiquetada representa un verdadero reto para detectar las observaciones discrepantes ya que hay varios problemas inmersos en los métodos propuestos y principalmente, el que se refiere al aspecto computacional, que hace difícil la aplicación de estas técnicas en forma rutinaria; y los problemas de enmascaramiento y empantanamiento.

Cuando uno se enfrenta con la posibilidad de observacio-

nes discrepantes multiples el problema que surge es como hacer inferencias acerca de i y λ simultaneamente, (teniendo en mente el modelo de la media trasladada 2.2.1). En esta situación uno también se enfrenta a la especificación del número exacto K de observaciones discrepantes o una cota superior de éste. Mickey, Dunn y Clark (1967) evitan la selección de K_a usando un procedimiento secuencial que se basa en la eliminación de ciertos casos tal que su eliminación alcanza el máximo decrecimiento en la suma de cuadrados residual. Las reglas usuales de este tipo de procedimiento indican cuando éste debe detenerse. Gentle (1978) propone también un método secuencial que se basa en la aplicación de estadística (2.3.11) en cada paso. El principal inconveniente para este tipo de pruebas es la pérdida de potencia, ademas de enfrentar problemas de enmascaramiento.

En cuanto a los métodos que se presentan, Gentleman y Wilk (1975) recomiendan que en general las K observaciones mas discrepantes deben determinarse escogiendo un K_a y encontrar el valor máximo de $Q_K(i)$ para $K=K_a$. Si $\max Q_K(i)$ no es estadisticamente discrepante el procedimiento se repite con $K=K_{a-1}$, pero en caso contrario el proceso se detiene con la identificación de $\hat{i}$ que maximiza $Q_K(i)$. Con este proceso se descarta la posibilidad de observaciones discrepantes cuando $Q_K(i)$ no es estadísticamente significante para $K=K_a, K_{a-1}, \ldots, 1$. Naturalmente el primer inconveniente de este método es la explosión computacional, principalmente para valores de n grandes (también para valores moderados), a pesar de los esfuerzos de algunos auto

res para poder calcular $Q_K(i)$ en forma simple ayudandose principalmente de la descomposición de la matriz V; ver Gentleman y Wilk (1975), Gentleman (1980), Cook y Weisberg (1982). Otro inconveniente es el hecho de no existir elementos claros para evaluar estadísticamente la discrepancia de $Q_K(i)$ ya que se desconoce la distribucion de $Q_K(i)$ y se proponen métodos informales para evaluar dicha discrepancia. Estos incluyen diagramas de algún subconjunto grande de las $Q_K(i)$ mas grandes contra "valores típicos" de éstas, obtenidos (aparentemente por simulación) bajo el modelo nulo o también de diagramas de residuales del modelo (2.2.1) que corresponden al conjunto de las observaciones discrepantes detectados por la $Q_K(i)$ mas grande. El primero de estos diagramas se apoya fuertemente en las suposiciones del modelo y se piensa están influenciados por no normalidad, heterosedasticidad y no aditividad de los errores; ver Barnett y Lewis (1978). Una ventaja posible del método se da cuando K se escoge suficientemente grande para que el efecto de enmascaramiento no se presente.

Con respecto al método de Andrews y Pregibon, los comentarios que podemos hacer son de alguna manera similares a los hechos a Gentleman y Wilk. Observando a R_i en (2.3.30), por ejemplo Cook y Weisberg (1982) comentan: (a) R_i es una medida sin unidad; (b) $R_i^{\frac{1}{2}} - 1$ corresponde al cambio proporcional en el volúmen de un elipsóide generado por $X^{*T}X^*$ cuando se eliminan los casos influyentes. Finalmente R_i es invariante bajo permutaciones de las columnas de X^* y entonces al vector de

respuestas $\underline{Y}$ no se le da reconocimiento especial.

A veces es útil un diagrama de $\log R_i/R_K^0$ para valores pequeños de K, donde R_0 denota el valor mínimo observado de R_i en todas las $\binom{n}{k}$ posibles. En estos diagramas se observa que para cierto valor de K para el cual $\log R_i/R_0$ es el más pequeño, este valor se aisla a la izquierda del resto, declarando este valor de K como el número de observaciones discrepantes y/o influyentes.

Aitkin y Wilson (1980) sugieren que un tratamiento adecuado de observaciones discrepantes se puede apoyar en modelos mezcla de la forma

$$F(\underline{Y}) = \sum_{i=1}^{m} P_i N(\mu_i, \sigma_i^2),$$

que consiste de varias componentes normales donde posiblemente las medias y/o las varianzas son diferentes, donde $0 \leq P_i \leq 1$ y $\sum P_i = 1$. Ellos argumentan que el algoritmo EM proporciona una forma relativamente simple para obtener estimadores máxima verosimilitud y que la identificación de observaciones discrepantes se puede basar en la probabilidad a posteriori de que el i-ésimo caso provenga de la j-ésima componente $j=0,1,2,\ldots,m$. Para ilustrar su metodología, ellos analizan uno de los grupos de datos que proponemos aquí, en el cual (como veremos más adelante), se muestra que los resultados dependen críticamente del número m de componentes usados en el modelo. Pruebas de cociente de verosimilitud para el número de com-

ponentes puede ser útil, pero parece que este enfoque, - básicamente reemplaza el problema de escoger K en la versión no etiquetada por el problema de escoger m. La elección de m estará sujeta al mismo tipo de dificultades e inevitablemente nos llevará a procedimientos secuenciales cuando m sea desconocida. Este enfoque será más útil cuando el número de contaminantes surjan consistentemente de un número conocido de componentes.

Por último, el procedimiento propuesto por Marasinghe propone una estadística que es una generalización de una que se usa para observaciones discrepantes múltiples en muestras simples, y por lo tanto tiene asociados los mismos problemas a esa estadística. Para superar dichas desventajas, el autor propone un - método modificado que aparentemente elimina el problema de enmascaramiento y empantamiento.

Mediante simulación, se obtienen las siguientes conclusiones: (a) Parece que el nuevo procedimiento controla el error tipo I cuando se escoge el valor de K adecuadamente. Esto implica que la probabilidad de declarar más observaciones discrepantes de las que hay es α y por lo tanto el problema de empantamiento no se da. (b) La potencia del procedimiento contra una observación discrepante es comparable con R_n (método secuencial, que usa al max $|r_i|$, r_i el residual studentizado) cuando $K=2$ y decrece ligeramente para valores - grandes de K. La potencia para detectar múltiples observaciones discrepantes mayor o igual al número real de dichas observaciones se mantiene alta. Un hecho

importante es que el proceso de simulación reporta que un porcentaje muy alto de las veces este método coincide con el de Gentleman y Wilk pero con la ventaja de la eliminación del proceso tedioso que se requiere para calcular $\max Q_k(i)$.

2.5. Ejemplos.

A continuación se aplicarán algunas de las técnicas presentadas a dos problemas que dominan la literatura referente a observaciones discrepantes. El primero de ellos, es

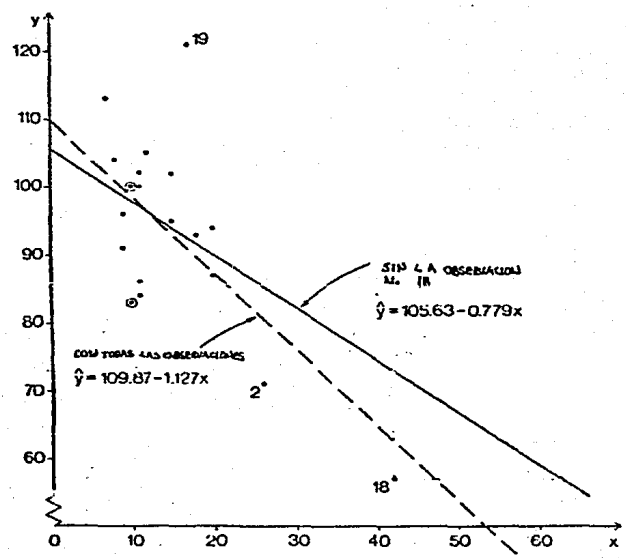
tabla 2.5.1.

Caso.	x	y	Caso	x	y
1	15	95	11	7	113
2	26	71	12	9	96
3	10	83	13	10	83
4	9	91	14	11	84
5	15	102	15	11	102
6	20	87	16	10	100
7	18	93	17	12	105
8	11	100	18	42	57
9	8	104	19	17	121
10	20	94	20	11	86
			21	10	100

analizando inicialmente en Mickey, Dunn y Clark (1967) y se refiere al estudio sobre cierta enfermedad del corazón en niños durante sus primeros meses de vida. Los datos están dados en la tabla 2.5.1. La variable Y es el puntaje adaptable de Gessel y X es la edad del niño (en meses).

La gráfica (2.5.1) representa el diagrama de puntos de las variables de este problema.

Gráfica 2.5.1.



Ahora consideremos el modelo de regresión con las hipótesis usuales para este grupo de datos. A partir de la gráfica, parece posible que dicho modelo se pueda establecer a pesar de que las observaciones 19, 18, 2 dominan nuestra percepción. Si los 3 casos fueran removidos, la linealidad es menos pronunciada.

La tabla (2.6.2), contiene los estimadores de los parámetros y la tabla de análisis de varianza al ajustar estos datos por mínimos cuadrados.

tabla 2.5.2.

ESTIMADORES DE LOS PARAMETROS

	ESTIMADOR	VARIANZAS
	109.8738	25.6826
	-1.1270	0.0962

LA VARIANZA ESTIMADA ES 121.505

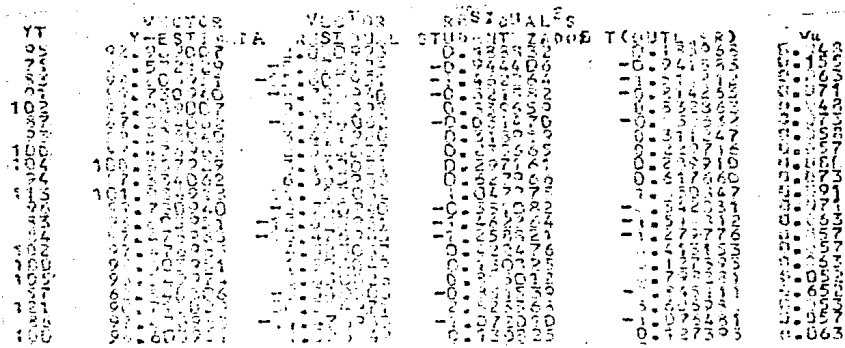
TABLA DE ANALISIS DE VARIANZA

FV	GL	SC	CM	F
HO : B=0	1	1604.081	1604.081	13.202
ERROR	19	2308.586	121.505	
TOTAL	20	5912.667		

EL COEFICIENTE DE CORRELACION **2 ES = 0.410

Por otro lado en la tabla (2.5.3) se exhiben algunos elementos necesarios para la aplicación de los métodos expuestos anteriormente.

TABLA 2.5.3.



METODO DE ELLENBERG. Si lo único que interesa es probar la hipótesis nula: no hay observaciones discrepantes, entonces, para el nivel de significancia α deseado, se tendría que resolver para C_α la expresión $P(|T_i| > C_\alpha) = \frac{\alpha}{n}$. Desgraciadamente, este procedimiento involucra métodos de integración numérica. Ellenberg (1976), calcula dicho valor crítico para los valores de $\alpha = 0.05, 0.01, 0.001$; obteniendo los siguientes resultados $C_{0.05} = 0.690$, $C_{0.01} = 0.708$ y $C_{0.001} = 0.781$. Observando la tabla (2.5.3), tenemos que el valor maximo de $|r_i|$ (residual studentizado en valor absoluto) corresponde a la observación 19: $r^* = 2.823$ y - por lo tanto $\gamma^* = (2.823)/(21-2)^{1/2} = 0.647$. Entonces se puede concluir que con un nivel de confianza de 0.05 la observación

19 es declarada como discrepante, no así, para $\alpha=0.01$ y $\alpha = 0.001$.

Con la finalidad de obtener el valor de c_α, sin necesidad de efectuar integraciones, en algunas ocaciones se podrán usar las tablas de Lund, pero unicamente se podría recomendar su uso cuando se disponga de un método adecuado de interpolación. Para nuestro problema n=21, interpolando linealmente, se consigue $c_\alpha = 0.6359$ y como $\tau^* = 0.647$, la conclusión seria la misma.

Si usamos directamente las tablas de Lund y como estadística de prueba a r^* las conclusiones son las mismas ya que estos métodos son equivalentes, ver Palacios, R. y Castañon V. (1985).

Si además, se requiere la cota inferior de la desigualdad de Bonferroni, entonces, se necesitará la matriz de correlaciones de los residuales y evaluar algunas probabilidades del tipo $P(\tau_i > c_\alpha, |\tau_j| > c_\alpha)$ dependiendo de si $c_\alpha \geq \left(\frac{1+|\rho_{ij}|}{2}\right)^{1/2}$. Usando la tabla (2.5.4), se encuentra que para $\alpha = 0.05$, la última desigualdad no sirve ya que ninguna correlación residual satisface $0.640 \geq \left(\frac{1+|\rho_{ij}|}{2}\right)^{1/2}$. Para $\alpha = 0.01$ solo la satisfacen dos: $\rho_{4,2}$ y $\rho_{12,2}$. Sin embargo para $\alpha=0.01$, solo se requieren calcular 4 probabilidades bivariadas. Los resultados, después de evaluar las probabilidades bivariadas son:

$$0.0499 \leq P(\max_i |\tau_i| > 0.640) \leq 0.05,$$
$$0.0100 \leq P(\max_i |\tau_i| > 0.708) \leq 0.0100,$$
$$0.00100 \leq P(\max_i |\tau_i| > 0.781) \leq 0.00100.$$

tabla 2.5.4.

MATRIZ V DE PROYECCION

1	0.05																			
2	0.05	0.15																		
3	0.05	0.01	0.06																	
4	0.04	−0.00	0.07	0.07																
5	0.05	0.05	0.05	0.04	0.05															
6	0.05	0.10	0.03	0.02	0.05	0.07														
7	0.05	0.08	0.04	0.03	0.05	0.06	0.06													
8	0.05	0.02	0.06	0.06	0.05	0.03	0.04	0.06												
9	0.04	−0.01	0.07	0.07	0.04	0.02	0.03	0.06	0.08											
10	0.05	0.10	0.03	0.02	0.05	0.07	0.06	0.03	0.02	0.07										
11	0.04	−0.02	0.07	0.08	0.04	0.01	0.03	0.07	0.08	0.01	0.09									
12	0.04	−0.00	0.07	0.07	0.04	0.02	0.03	0.06	0.07	0.02	0.08	0.07								
13	0.05	0.01	0.06	0.07	0.05	0.03	0.04	0.06	0.07	0.03	0.07	0.07	0.06							
14	0.05	0.02	0.05	0.05	0.05	0.03	0.04	0.06	0.06	0.03	0.07	0.06	0.06	0.06						
15	0.05	0.02	0.06	0.06	0.05	0.03	0.04	0.06	0.06	0.03	0.07	0.06	0.06	0.06	0.05					
16	0.05	0.01	0.06	0.07	0.05	0.03	0.04	0.06	0.07	0.03	0.07	0.07	0.06	0.06	0.06	0.06				
17	0.05	0.03	0.06	0.06	0.05	0.04	0.04	0.05	0.06	0.04	0.06	0.06	0.06	0.05	0.05	0.06	0.05			
18	0.06	0.30	−0.05	−0.07	0.06	0.17	0.13	−0.03	−0.09	0.17	−0.11	−0.07	−0.05	−0.03	−0.03	−0.05	−0.00	0.65		
19	0.05	0.07	0.04	0.04	0.05	0.06	0.06	0.04	0.03	0.06	0.03	0.04	0.04	0.04	0.04	0.04	0.04	0.10	0.05	
20	0.05	0.02	0.06	0.06	0.05	0.03	0.04	0.06	0.06	0.03	0.07	0.06	0.06	0.06	0.06	0.06	0.05	−0.03	0.04	0.06
21	0.05	0.01	0.06	0.07	0.05	0.03	0.04	0.06	0.07	0.03	0.07	0.07	0.06	0.06	0.06	0.06	0.06	−0.05	0.04	0.06
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

0.06																					

CORRELACIONES RECIPROCAS

1	1.000																				
2	−0.059	1.000																			
3	−0.048	−0.008	1.000																		
4	−0.048	0.002	−0.071	1.000																	
5	−0.050	−0.059	−0.048	−0.048	1.000																
6	−0.054	−0.112	−0.030	−0.026	−0.054	1.000															
7	−0.052	−0.091	−0.037	−0.034	−0.052	−0.068	1.000														
8	−0.048	−0.018	−0.063	−0.066	−0.048	−0.035	−0.040	1.000													
9	−0.048	0.013	−0.075	−0.081	−0.048	−0.021	−0.032	−0.069	1.000												
10	−0.054	−0.112	−0.030	−0.026	−0.054	−0.078	−0.068	−0.035	−0.021	1.000											
11	−0.047	0.023	−0.079	−0.086	−0.047	−0.016	−0.029	−0.073	−0.093	−0.016	1.000										
12	−0.048	0.002	−0.071	−0.076	−0.048	−0.026	−0.034	−0.066	−0.081	−0.026	−0.086	1.000									
13	−0.048	−0.008	−0.067	−0.071	−0.048	−0.030	−0.037	−0.063	−0.075	−0.030	−0.079	−0.071	1.000								
14	−0.048	−0.018	−0.063	−0.066	−0.048	−0.035	−0.040	−0.060	−0.069	−0.035	−0.073	−0.066	−0.063	1.000							
15	−0.048	−0.018	−0.063	−0.066	−0.048	−0.035	−0.040	−0.060	−0.069	−0.035	−0.073	−0.066	−0.063	−0.060	1.000						
16	−0.048	−0.008	−0.067	−0.071	−0.048	−0.030	−0.037	−0.063	−0.075	−0.030	−0.079	−0.071	−0.067	−0.063	−0.063	1.000					
17	−0.049	−0.029	−0.059	−0.062	−0.049	−0.039	−0.043	−0.057	−0.064	−0.039	−0.066	−0.062	−0.059	−0.057	−0.057	−0.059	1.000				
18	−0.106	−0.556	0.084	0.123	−0.106	−0.300	−0.221	0.046	0.162	−0.300	0.202	0.123	0.084	0.046	0.046	0.084	0.008	1.000			
19	−0.052	−0.080	−0.041	−0.039	−0.052	−0.063	−0.058	−0.043	−0.037	−0.063	−0.035	−0.039	−0.041	−0.043	−0.043	−0.043	−0.041	−0.045	−0.183	1.000	
20	−0.048	−0.018	−0.063	−0.066	−0.048	−0.035	−0.040	−0.060	−0.069	−0.035	−0.073	−0.066	−0.063	−0.060	−0.060	−0.060	−0.063	−0.057	0.046	−0.043	1.000
21	−0.048	−0.008	−0.067	−0.071	−0.048	−0.030	−0.037	−0.063	−0.075	−0.030	−0.079	−0.071	−0.067	−0.063	−0.063	−0.067	−0.059	0.084	−0.041	−0.063	1.000
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21

MÉTODO DE COOK Y PRESCOTT. Nuevamente, si nuestro interés solo radica en probar si una observación es discrepante o no lo es, unicamente se requiere encontrar el valor crítico d que satisface $P(|\tau_i| > d) = P[F > d^2(n-p-1)/(1-d^2)]$ o equivalentemente $\alpha = nPr[F > d^2(n-p-1)/(1-d)^2]$. Si $\tau^* > d$ entonces se rechaza la hipótesis nula: no existen observaciones discrepantes, donde τ^* está definido en (2.3.6). En nuestro ejemplo, se sigue una línea un poco diferente. Tomamos el valor $\tau^* = d = 0.6475$ y obtenemos $\alpha = 0.0425$. Entonces no se requiere probar la hipótesis nula con $\alpha = 0.05$ ya que el verdadero valor es menor a 0.05, por lo tanto, podemos declarar a la observación 19 como discrepante.

Sin embargo, cuando se juzga el peso de la evidencia contra la hipótesis nula, sería deseable conocer que tanto $P(\max_i |\tau_i|)$ es menor que el valor α propuesto. Esto último, puede determinarse calculando la cota inferior de (2.3.16), que suele facilitarse al seguir la recomendación de sustituir en β^+ y β^- a ρ_{ij} por el $\max_{c(+)}$ y $\min_{c(-)}$ ρ_{ij} respectivamente. Al observar en la tabla (2.5.4) nos damos cuenta que todas las correlaciones están contenidas en el intervalo $[-0.556, 0.202]$ y aplicar el criterio mencionado, $\alpha - \beta^+ - \beta^-$ es negativo. Para mejorar este resultado, se hace un análisis exhaustivo de la matríz de correlación y obtenemos que $\rho_{18,2} = -0.56$, $\rho_{18,6} = 0.3$, de las restantes 17 caen en el intervalo $[0.002, 0.202]$ y 190 en $[-0.221, -0.016]$. Entonces, para calcular una segunda cota inferior, se usan $\{-0.556, -0.3, -0.016, 0.202\}$ con sus respectivas frecuencias $\{1, 2, 190, 17\}$ para calcular β^+ y $\{-0.556, -0.3, -0.221, 0.002\}$ con las mismas frecuencias para calcular β^-. Como podemos observar este procedi-

miento solo involucra el cálculo de 8 probabilidades para obtener que $\beta^+ + \beta^- < 0.0016$. Entonces, la probabilidad del máximo $|\tilde{r}_i|$, tiene como cotas los límites del intervalo $[0.0409, 0.0425]$ para $d = 0.6475$. Igual procedimiento se aplicaría si se quiere encontrar las cotas para un nivel de significancia específico, por ejemplo, para $\alpha = 0.01$, 0.05 y 0.1, las cotas inferiores resultan 0.00997, 0.0476 y 0.086 respectivamente.

METODO DE DOORNBOS. Lo relevante de este método es obtener una cota superior para la probabilidad bivariada de la desigualdad de Bonferroni, expresada en (2.3.21) siempre y cuando $\max_{i,j} |\rho_{ij}| \leq g_\alpha = \frac{c_\alpha^2 - c_\alpha^4 + c_\alpha^2}{c_\alpha^2 + c_\alpha^4 - c_\alpha^2}$

A continuación, se presentan algunos resultados aproximados para $\alpha = 0.1$, $\alpha = 0.05$, $\alpha = 0.01$

$$\alpha = 0.1 \;;\; g_\alpha = 0.1908,$$
$$\alpha = 0.05 \;;\; g_\alpha = 0.225,$$
$$\alpha = 0.01 \;;\; g_\alpha = 0.2969.$$

Como los resultados lo muestran, para ningún valor de α se consigue la condición (2.3.25) y por lo tanto no podemos garantizar que $\Pr(\tilde{r}^* \geq c_\alpha)$ este en el intervalo $[\alpha - \frac{1}{2}\alpha^2, \alpha]$.

METODO DE WEISBERG. La aplicación de este método es muy sencilla. Observando en la tabla (2.5.3) encontramos que $\max |t_i|$ $(t_i = r_i (n-p-1 / n-p-r_i^2)^{1/2}$ es $t_{19} = 3.6069$. Nuevamente para $\alpha = 0.05$, la observación 19 es declarada como discrepante ya que el valor de t en la tabla de Weisberg $(n = 21, p = 2)$ es 3.53. Pero, para $\alpha = 0.01$ la conclusión es que no hay observa

ciones discrepantes ya que el valor de t es 4.26.

METODO DE ANDREWS Y PREGIBON. En la tabla (2.5.5) se encuentran resumidos los valores de $\hat{R}_i$ y $\hat{a}_k$ para diferentes valores de K, al observarla en K=1, nuestra atención se centra en la observación 18, debido principalmente a que es la más alejada en el espacio de factores. Sin embargo las observaciones 2 y 19 también requieren de un análisis adicional. Por lo que respecta a la observación 19, ésta ha sido declarada como discrepante en el uso de los métodos anteriores, con un nivel de confianza del 6%. Este hecho se puede confirmar al observar la segunda y tercer columna de la tabla (2.5.6) que nos proporciona $Q_i(i)$ y la primer componente de la estadística de Andrews y Pregibon (AP_i) que como se ha mencionado anteriormente, son elementos de diagnóstico para detectar observaciones discrepantes. Los valores correspondientes son .969 y .58, valores que confirman que la observación 19 puede declararse como discrepante. A pesar de todo, la estadística de Andrews y Pregibon señala a la observación 18 como influyente y/o discrepante. Debido a su posición en el espacio de factores, la observación 18 es identificada como influyente. Por otro lado, para la observación 2 no hay evidencia de que esta observación sea influyente y/o discrepante.

Para K=2, la tabla (2.5.7) proporciona sólo aquéllas parejas que potencialmente pudieran importar. Por un lado la estadística A-P selecciona a la pareja (2,18) como más importante con respecto a la pareja (18,19) por pequeño margen. Para esta pareja, (2,18), AP_2 es pequeña indicando que estas observaciones son potencialmente influyentes, mientras que para las

tabla 2.5.5.

k	R_k^0	$\hat{\alpha}_k$	*Observaciones*
1	·3350940	·3838198	18
2	·1645985	·2310607	18, 2
3	·0897915	·2262575	18, 2, 19
4	·0593024	·7475751	18, 2, 19, 11

Gráfica 2.5.2.

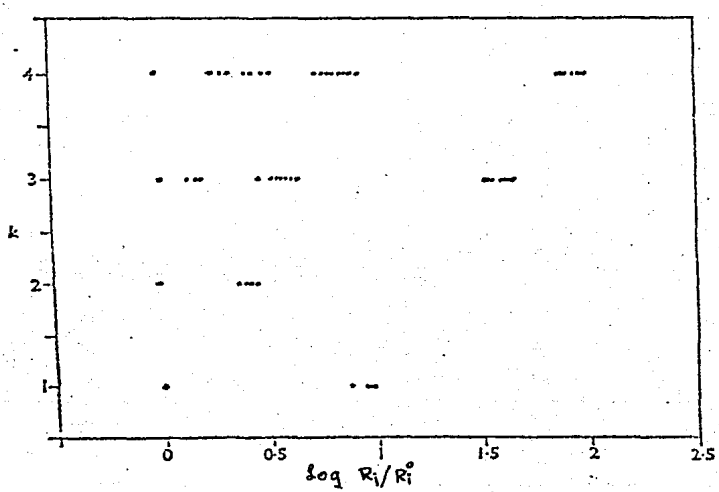


tabla 2.5.6.

Observación Eliminada	Q_1	AP_1 $100(1-Q_1/SCR)$	AP_2 $100(1-V_{ii})$	Estadística AP $100R_i^{(1)}$
1	4	100	95	95
2	108	96	85	81
3	260	89	94	83
4	82	97	93	90
5	86	97	95	92
6	0	100	93	93
7	12	100	94	94
8	7	100	94	94
9	11	100	92	92
10	48	98	93	91
11	133	95	91	86
12	15	99	93	92
13	260	89	94	83
14	193	91	94	86
15	22	99	94	93
16	2	100	94	94
17	79	97	95	92
18	88	95	35	33
19	969	58	95	55
20	140	94	94	89
21	2	100	94	94
Valores usuales	Alto	Bajo	Bajo	Bajo

tabla 2.5.7.

Observación Eliminada	$Q_2(i)$	AP_1 $100(1-Q_2(i)/Q_2)$	AP_2 $100\lvert 1-v_i \rvert$	Estadística AP $100R_{ij}^{(2)}$
18, 2	442	81	20	16
18, 3	324	86	32	28
18, 11	277	88	30	27
18, 19	983	57	32	18
19, 2	1031	55	80	44
19, 3	1189	48	89	43
19, 11	1128	51	86	44
Valores Usuales	Alto	Bajo	Bajo	Bajo

observaciones (18,19), AP, es pequeño y entonces podrían señalarse como discrepantes. Como se ha visto, la observación 19, ha sido anteriormente declarada discrepante y entonces al eliminar dos observaciones en los datos originales que incluya a la observación 19, ésta contribuye fuertemente para que $Q_2(i)$ sea un valor grande, así $Q_2(19,3)$ es el valor mas grande, pero mucho de $Q_2(19,3)$ se debe a la observación 19.

De lo expuesto anteriormente, únicamente podremos señalar una observación como discrepante y es, en este caso, la nú

mero 18. Esta misma conclusión puede obtenerse al observar la gráfica (2.5.2) en la que para K=1, hay mayor separación entre los valores respectivos de $\log\{R_i/R_j\}$.

Uno de los principales inconvenientes de la estadística AP es que no necesariamente pone atención a puntos que son discrepantes y/o influyentes en términos de la estimación de parámetros ya que, como se puede observar, AP_2 da un peso considerable a conjuntos de puntos que pueden considerarse como remotos en el espacio de factores, para una discusión más amplia acerca de observaciones influyentes. N. R. Draper y J. A. John (1981).

El segundo grupo de datos fué tomado de Brownlee, K. A. (1965), están representados en la tabla (2.5.8) y son el resultado de 21 días de operación de una planta para la oxidación de NH_3 (Amonia) a HNO_3 (ácido nítrico) y las variables son:

Y= Diez veces el porcentaje de amonia que se pierde como ácido nítrico no absorbido (esta es una medida indirecta del rendimiento del ácido nítrico).

X_1 = Flujo de aire en la planta.

X_2 = Temperatura del agua al entrar.

X_3 = Concentración de ácido nítrico en el líquido absorbido.

Al investigar las operaciones de la planta, se puede considerar a las corridas (1,2), (4,5,6), (7,8) y (18,19) como re

Tabla 2.5.8.

Observación Número	Y	x_1	x_2	x_3
1	42	80	27	89
2	37	80	27	88
3	37	75	25	90
4	28	62	24	87
5	18	62	22	87
6	18	62	23	87
7	19	62	24	93
8	20	62	24	93
9	15	58	23	87
10	14	58	18	80
11	14	58	18	89
12	13	58	17	88
13	11	58	18	82
14	12	58	19	93
15	8	50	18	89
16	7	50	18	86
17	8	50	19	72
18	8	50	19	79
19	9	50	20	80
20	15	56	20	82
21	15	70	20	91

plicas.

En la tabla (2.5.9) se presentan los estimadores de los parámetros, la tabla de análisis de varianza y algunos elementos de diagnóstico obtenidos de ajustar estos datos por mínimos cuadrados. El modelo resultante es

$$\hat{Y} = -39.9 + 0.72 X_1 + 1.30 X_2 - 0.15 X_3 . \qquad (2.5.1)$$

Daniel y Wood (1971) analizan detalladamente este conjunto de datos y la conclusión a la que llegan es que las observaciones 1, 3, 4 y 21 son discrepantes, y que la variable X_3 no es necesaria. Su modelo final es

$$\hat{Y} = -15.4 - 0.07 X_1 + 0.53 X_2 + 0.0068 X_1^2 . \qquad (2.5.2)$$

Por otro lado si el ajuste (Mínimos cuadrados) se efectúa unicamente eliminando a las observaciones identificadas como discrepantes se obtiene el modelo.

$$\hat{Y} = -37.6 + 0.80 X_1 + 0.58 X_2 - 0.07 X_3 . \qquad (2.5.3)$$

Como los métodos de Ellenberg, Lund y Weisberg son similares, usaremos este último para el caso $k=1$ (una observación discrepante). Buscando en la tabla (2.5.9), encontramos que $\max |t_{i1}|$ corresponde a $t_{21} = -3.33$ cuando usamos al modelo (2.5.1). Pero como el valor $t(\alpha/n, n-p) = 3.60$ de la tabla Weisberg no rechazariamos la hipótesis nula (no existen observaciones discrepantes) para $\alpha = 0.05$, lo mismo ocurre con $\alpha = 0.01$, ya que $t(\alpha/n, n-p) = 4.42$. Como podemos constatar la prueba para detectar una observación discrepante no es útil en esta situación a pesar de que t_1, t_3,

tabla 2.6.9.
ESTIMADORES DE LOS PARAMETROS

	BETAS	VARIANZAS
	-30.9197	141.5147
	0.7156	0.0182
	1.2953	0.1554
	-0.1521	0.0244

LA VARIANZA ESTIMADA ES 10.519

TABLA DE ANALISIS DE VARIANZA

FV	GL	SC	CM	F
MO : REG	3	1320.403	440.136	59.992
ERROR	17	178.300	10.519	
TOTAL	20	2067.303		

EL COEFICIENTE DE CORRELACION AL 2 ES = 0.914

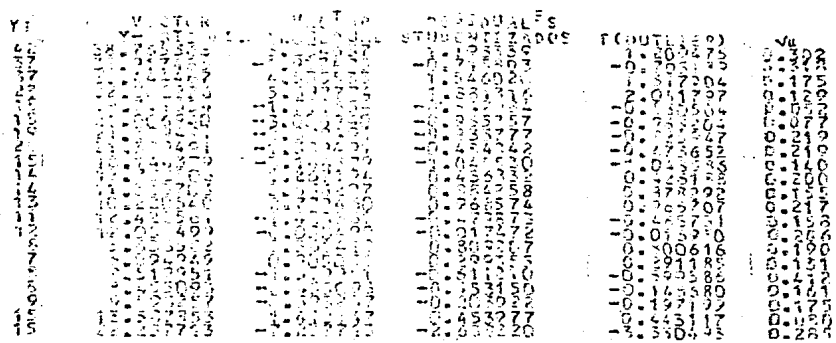
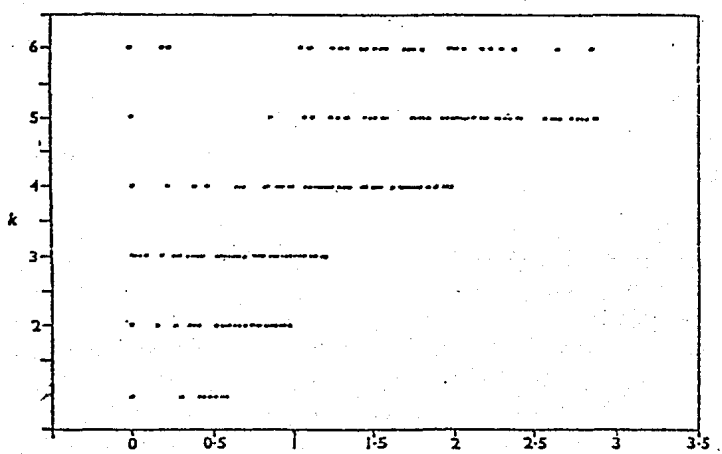


tabla 2.5.10.

k	R_k^0	$\hat{\alpha}_k$	*Observaciones*
1	·4225375	·0596515	21
2	·2072535	·1093887	21, 4
3	·1147183	·3962458	21, 4, 2
4	·0337735	·0257674	21, 4, 3, 1
5	·0080269	·0068339	21, 4, 3, 1, 2
6	·0039551	·0029340	21, 4, 3, 1, 2, 13



cente al comparar $-2 \log L$ con $\chi^2_{1-\alpha}(4)$ para varios valores de α. (esto es consistente con la prueba de Weisberg aplicada anteriormente). Los parámetros estimados están en la tabla (2.5.11) y la probabilidad a posteriori de que la observación $i=1,2,\ldots 21$ sea miembro de la componente j (según el modelo) para los diferentes modelos están en la tabla (2.5.12).

Si inicialmente a la observación 4 (esta tiene el residual positivo más grande en el modelo original) se especifica como discrepante, entonces la convergencia ocurre en 7 iteraciones y los parámetros estimados son idénticos a los obtenidos para el modelo completo de regresión. La observación es "reabsorbida" (reabsorbed) en la primer componente y todas las probabilidades a posteriori de las observaciones de ser miembros de la componente 2 son menores a 10^{-3}.

Si hay varias observaciones discrepantes, el modelo anterior podría no detectarlas si estas se presentan en el lado opuesto del plano de regresión. Para este propósito será necesario un modelo con al menos 3 componentes.

Consideremos un modelo con 3 componentes (modelo 4), en el cuál, la observación 21 y 4 se asignan inicialmente a la segunda y tercera componente respectivamente. La convergencia ocurre en 3 iteraciones y para esta solución $-2 \log L = 58.08$. La reducción comparada con el modelo de 2 componentes es apenas de 3.80. también los parámetros se encuentran en la tabla (2.6.11) y las probabilidades en la (2.5.12). La evidencia para un modelo de 3 componentes no es entonces convincente.

Si las 3 observaciones, 1, 3 y 4 se asignan inicialmente a

tabla 2.5.11.

Parámetros estimados para los modelos del ejemplo no. 2.

Modelo	No. de componentes	Observaciones Discrepantes	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\mu}_5$	$\hat{\sigma}^2$ (sesgado)	-2 log L
1	1	-	-39.9	.716	1.30	-.152	-	-	-	-	8.52	65.98
2	2	(21)	-43.9	.883	.836	-.105	15.0	-	-	-	5.08	61.88
3	2	(4)	-39.9	.716	1.30	-.152	-	.003	-	-	8.51	65.98
4	3	(21) (4)	-42.6	.955	.560	-.108	14.8	28.0	-	-	2.88	58.08
5	4	(21) (4) (1,3)	-38.0	.794	.589	-.006	14.8	28.0	39.5	-	1.54	58.12
6	5	(21) (4) (1) (3)	-38.3	.785	.591	-.005	14.9	28.0	42.0	37.0	0.84	49.96

tabla 2.5.11.

Parametros estimados para los modelos del ejemplo no. 2.

Modelo	No. de componentes	Observaciones Discrepantes	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\mu}_5$	$\hat{\sigma}^2$ (sesgado)	-2 log L
1	1	-	-39.9	.716	1.30	-.152	-	-	-	-	8.52	65.98
2	2	(21)	-43.9	.883	.836	-.105	15.0	-	-	-	5.08	61.88
3	2	(4)	-39.9	.716	1.30	-.152	-	.003	-	-	8.51	65.98
4	3	(21) (4)	-42.6	.955	.560	-.108	14.0	28.0	-	-	2.88	58.08
5	4	(21) (4) (1,3)	-38.0	.794	.589	-.006	14.8	28.0	39.5	-	1.54	58.12
6	5	(21) (4) (1) (3)	-38.3	.785	.591	-.005	14,9	28.0	42.0	37.0	0.84	49.96

tabla 2.5.12.

Modelo	2	4		5			6			
Componente	2	2	3	2	3	4	2	3	4	5
Observación										
1						1			1	
2						.027				.198
3						1				1
4			.998		1			1		
5	.051	.022		.004						
6	.076	.032		.016			.001			
7	.027	.005								
8	.008	.001								
9	.133	.101		.124			.193			
10	.072	.068		.078			.091			
11	.079	.072		.100			.138			
12	.057	.043		.035			.019			
13	.036	.032		.006						
14	.012	.028		.013			.003			
15	.001									
16										
17	.001									
18	.001									
19	.003									
20	.092	.117		.260			.013			
21	1	1		1			1			

. LOS ESPACIOS EN BLANCO INDICAN UNA PROBABILIDAD MENOR A 10^{-3}

la tercera componente, la convergencia ocurre en 8 iteraciones y los estimadores de los parámetros son idénticos a los obtenidos con el modelo 2. Las tres observaciones son "reabsorbidas" en la primer componente y todas las probabilidades de pertenecer a la componente 3 son menores que 10^{-4} también se consideran modelos con 4, 5 y 6 componentes a pesar de que se requieren sin gran número de parámetros para tales modelos que los hace muy complejos para 21 observaciones. Los modelos 5 y 6 están resumidos en la tabla (2.5.11) y las probabilidades correspondientes en (2.5.12).

Para 6 componentes, las 5 observaciones que se encontraron en Andrews y Pregibon como discrepantes fueron asignadas inicialmente a componentes diferentes. La convergencia ocurre lentamente (23 iteraciones) con una solución de $-2 \log L = 42.11$. Las componentes que contenían a las observaciones 2 y 3 se transforman en una sola y las observaciones 10, 11, 12 y 20 fueron separadas de la observación 21.

Existen varios dificultades con este enfoque: (a) con muestras pequeñas la distribución asintótica ($-2 \log L$) no se cumple satisfactoriamente, (b) La más seria, χ^2 con el número de grados de libertad igual al número de parámetros por probarse en la hipótesis nula no es válida, para la mezcla de modelos es no-regular, como se puede ver en el modelo de 2 componentes, donde p no es identificable si $\mu_1 = \mu_2$, o equivalentemente, la hipótesis de "no existen observaciones discrepantes.

METODO DE MARASINGHE. Para la aplicación de este método usaremos el modelo final (2.5.2) obtenido por Daniel y Wood

(1971). En la tabla (2.5.13) se presentan los valores de $\hat{\nu}_{ii}$ y t_i para el modelo completo y sin las observaciones 21, (4, 21), (2,4,21), (12, 4, 21) y (3, 3, 4, 21) respectivamente. Por otro lado, en la tabla (2.5.14) se presentan los residuales ajustados para el modelo completo - sin la observación 21, (21, 4) y (21, 4, 2).

Analizando las últimas 4 columnas de la tabla (2.5.14) observamos que existen 3 residuales ajustados que difieren fuertemente de los demás, 8.45, 7.68, 4.70 que corresponden a las observaciones 21, 4 y 2 respectivamente.

Por estas razones, podríamos sospechar la presencia de 3 observaciones discrepantes. Para efectos de evitar enmascaramiento, supongamos inicialmente que están presentes cuatro observaciones discrepantes (k=4). Las 4 observaciones seleccionadas por el procedimiento son 21, 4, 2, 20 respectivamente, que corresponden a las cuatro observaciones con el residual ajustado (en valor absoluto) más grande. Debe notarse que estas 4 observaciones son las "4 observaciones probablemente más discrepantes" de Gentleman y Wilk seleccionadas por $Q_i^*(k) = 159.34$. También debe mencionarse que la prueba para una sola observación discrepante falla en declarar a la observación 21 como discrepante al 5% de significancia. Entonces el procedimiento secuencial basado en R_n (procedimiento que podría usar el criterio de Ellenberg, Lund y Weisberg indistintamente), no detectaría observaciones discrepantes en este modelo.

El procedimiento secuencial descrito anteriormente se aplica a este conjunto de datos, cuyos resultados se encuentran resumidos en la tabla (2.5.15). Como podemos recordar

tabla 2.5.13

Valores de v_{ii}						residuales studentizados						
observaciones eliminadas							observaciones eliminadas					
ninguna	(21)	(4,21)	(2,4,21)	(1,2,4,21)	(1,3,4,21)		ninguna	(21)	(4,21)	(2,4,21)	(1,2,4,21)	(1,3,4,21)
0.409	0.421	0.421	0.727	•	•	1	0.97	0.77	1.08	−1.75	•	•
0.409	0.421	0.421	•	•	0.993	2	−1.06	−1.81	−2.71	•	•	0.57
0.176	0.199	0.201	0.308	0.983	•	3	1.54	1.40	2.30	1.91	1.21	•
0.191	0.192	•	•	•	•	4	2.27	3.01	•	•	•	•
0.103	0.108	0.125	0.125	0.125	0.131	5	−0.31	−0.63	−0.31	−0.40	−0.40	−0.12
0.134	0.134	0.164	0.164	0.165	0.169	6	−0.73	−0.97	−0.62	−0.80	−0.82	−0.64
0.191	0.192	0.238	0.239	0.240	0.242	7	−0.84	−0.92	−0.30	−0.35	−0.31	−0.18
0.191	0.192	0.238	0.239	0.240	0.242	8	−0.50	−0.48	0.36	−0.54	0.66	0.84
0.163	0.170	0.206	0.208	0.218	0.208	9	−0.94	−0.89	−0.39	−0.36	−0.18	−0.66
0.139	0.175	0.175	0.176	0.179	0.179	10	0.84	0.38	0.49	0.75	0.94	0.96
0.139	0.175	0.175	0.176	0.179	0.179	11	0.84	0.38	0.49	0.75	0.94	0.96
0.212	0.272	0.275	0.276	0.279	0.280	12	0.96	0.31	0.16	0.30	0.45	0.53
0.139	0.175	0.175	0.176	0.179	0.179	13	−0.17	−0.91	−1.42	−1.82	−1.87	−1.98
0.092	0.110	0.111	0.112	0.116	0.113	14	−0.25	−0.79	−1.04	−1.30	−1.29	−1.41
0.188	0.189	0.191	0.191	0.195	0.193	15	0.17	0.34	0.29	0.29	0.19	0 32
0.188	0.189	0.191	0.191	0.195	0.193	16	−0.17	−0.10	−0.35	−0.57	−0.76	−0.67
0.187	0.195	0.195	0.195	0.198	0.197	17	−0.26	−0.01	−0.01	−0.10	−0.23	−0.21
0.187	0.195	0.195	0.195	0.198	0.197	18	−0.26	−0.01	−0.01	−0.10	−0.23	−0.21
0.212	0.232	0.234	0.234	0.236	0.237	19	−0.35	0.09	0.33	0.37	0.31	0.27
0.064	0.064	0.069	0.070	0.076	0.070	20	0.68	0.75	1.42	2.04	2.38	2.16
0.288	•	•	•	•	•	21	−2.63	•	•	•	•	•

Table 2.5 14. Residuales Ajustados

Datos				Observaciones			
Observación	X1	X2	Y	Ninguna	(21)	(21, 4)	(21, 4, 2)
1	80	27	42	3.11	1.96	1.87	−2.25
2	80	27	37	−3.40	−4.61	−4.70	—
3	75	25	37	4.93	3.58	4.00	2.45
4	62	24	28	7.30	7.68	—	—
5	62	22	18	−.98	−1.61	−.54	−.51
6	62	23	13	−2.35	−2.48	−1.07	−1.02
7	62	24	19	−2.71	−2.34	−.52	−.45
8	62	24	20	−1.59	−1.22	.62	.69
9	58	23	15	−3.02	−2.27	−.67	−.46
10	58	18	14	2.69	.98	.64	.96
11	58	18	14	2.69	.98	.84	96
12	58	17	13	3.09	.79	.28	.39
13	58	18	11	−.54	−2.33	−2.46	−2.34
14	58	19	12	−.79	−2.02	−1.81	−1.68
15	50	16	8	.55	.86	.50	.37
16	50	18	7	−.56	−.25	−.61	−.74
17	50	19	8	−.84	−.01	−.02	−.13
18	50	19	8	−.84	−.01	−.02	−.13
19	50	20	9	−1.13	.22	.58	.48
20	56	20	15	2.17	1.91	2.46	2.62
21	70	20	15	−8.45	—	—	—
Suma de Cuadrados Res.				175.642	104.206	45.240	23.131

tabla (2.5.15)

n	K	observaciones probadas	Q_K^{**}	F_k	valores críticos	observación eliminada
21	4	21, 4, 2, 20	159.34	.0928	.1348(1%)	21
20	3	4, 2, 20	87.93	.1562	.1815(1%)	4
19	2	2, 20	20.96	.36	.3366(5%)	

$Q_K^{**}(i)$ puede obtenerse calculando la reducción en la suma de cuadrados resultante de eliminar K observaciones, o usando (2.3.37). Por ejemplo $Q_4^{**}(21,4,2,20) = (-8.46)^2 + (7.68)^2 + (-4.7)^2 + (2.62)^2 = 159.34$, entonces, $F_4 = (175.642 - 159.34)/175.642 = 0.0928$. Ya que F es significante al 1% entonces la observación 21 es eliminada, y se calcula $Q_3^{**}(4,2,20) = (7.68)^2 + (-4.7)^2 + (2.62)^2 = 87.93$; $F_3 = (109.206 - 87.93)/109.206 = 0.1562$ y así sucesivamente. Claramente las observaciones 21 y 4 son declaradas como discrepantes ya que el procedimiento termina en la tercer etapa. Entonces este análisis nos lleva a concluir que existen dos observaciones discrepantes en estos datos con respecto a éste modelo.

Capitulo 3.

ENFOQUE BAYESIANO.

3.1. Introducción.

Como podemos recordar, en el enfoque frecuentista, las inferencias se hacen poniendo atención directa a un conjunto hipotético de datos $x_1, x_2, \ldots, x_n$ de una muestra aleatoria $X_1, X_2, \ldots X_n$ que se supone son generados por el modelo $p(x/\underline{\theta})$ donde $\underline{\theta} \in \Theta$ (espacio parametral) y además $\underline{\theta} = (\theta_1, \theta_2 \ldots \theta_r)$ es fijo y representa supuestamente a los valores de los parámetros desconocidos. Posteriormente, se seleccionan estimadores $\hat{\underline{\theta}}$ que son funciones del vector de datos X y suponiendo que se puede repetir un número "grande" de veces esta situación, se obtienen valores $\hat{\underline{\theta}}(x_1), \hat{\underline{\theta}}(x_2) \ldots$; ahora se pueden efectuar inferencias comparando los valores de $\hat{\underline{\theta}}(x)$ con sus "distribuciones de muestreo" generadas al considerar muchas repeticiones. Las funciones $\hat{\underline{\theta}}(x)$ usualmente se escogen de tal manera que las distribuciones de muestreo de los estimadores $\hat{\underline{\theta}}(x_i)$ estan, en algún sentido, lo más concentradas posible alrededor de los valores verdaderos ?

En estadística Bayesiana, el modelo básico se entiende en otra dirección Sea $X_1, X_2, \ldots X_n$ una muestra aleatoria de tamaño n que proviene de una densidad $f(x/\underline{\theta}_0)$ con $\underline{\theta}_0$ un valor de Θ Entre las primeras diferencias que se presentan en este enfoque es que $\underline{\theta}$ es considerada como una variable aleatoria cuya función de densidad $g(\underline{\theta})$, se supone conocida y que expresa el grado de certeza (degree of belive) o conocimiento inicial acerca de $\underline{\theta}$; a $g(\underline{\theta})$ se le cono-

ce con el nombre de distribución a priori para $\underline{\theta}$ ó simplemente a priori. Mientras que en enfoque clásico, se considera a la función de verosimilitud como la única expresión que contiene toda la información, en estadística Bayesiana la expresión que nos proporciona la información contenida en la verosimilitud y en la función a priori es precisamente el teorema de Bayes.

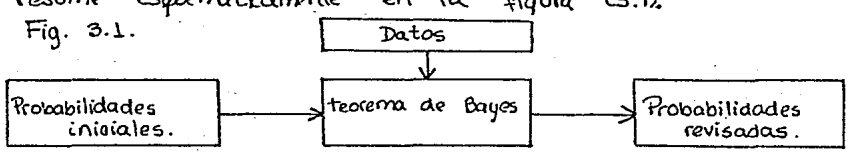
$$p(\underline{\theta}/\underline{x}) = \frac{f(\underline{x}/\underline{\theta}) \cdot g(\underline{\theta})}{f(\underline{x})}$$

$$= \frac{[\prod_{i=1}^{n} f(x_i/\underline{\theta})]\ g(\underline{\theta})}{\int [\prod_{i=1}^{n} f(x_i/\underline{\theta})]\ g(\underline{\theta})\, d\underline{\theta}} \qquad (3.1.1)$$

Así, lo que distingue al esquema Bayesiano de otros enfoques estadísticos es que, antes de la obtención de la muestra, considera su grado de certeza para $\underline{\theta}$ y lo representa en forma de probabilidad. Una vez obtenidos los datos, el teorema de Bayes permite calcular un nuevo conjunto de probabilidades que representan los grados de certeza revisados en los modelos posibles, tomando en cuenta la nueva información proporcionada por los datos.

El proceso básico bajo la metodología bayesiana se resume esquemáticamente en la figura (3.1).

Fig. 3.1.



Como se ha mostrado inicialmente, las componentes que caracterizan a la estadística Bayesiana son: La información a priori, los datos muestrales, el cálculo de la densidad a posteriori para los parámetros y en algunas ocaciones el calculo de la distribución predictiva de futuras observaciones. La información a priori está expresada por la densidad $p(\underline{\theta})$ del parámetro $\underline{\theta}$ correspondiente al modelo de probabilidad $f(\underline{x}/\underline{\theta})$, donde f es la densidad de la variable aleatoria X. La información en los datos $x_1, x_2, \ldots x_n$, está contenida en la función de verosimilitud $(\underline{\theta}/x_1, x_2, \ldots x_n)$ que es la densidad conjunta de los datos muestrales que al combinarse con la densidad a priori de $\underline{\theta}$ por medio del teorema de Bayes, se obtiene la densidad a posteriori de θ

$$p(\underline{\theta}/x_1, x_2, \ldots x_n) = L(\underline{\theta}/x_1, x_2, \ldots x_n) \cdot p(\underline{\theta})$$

a partir de la cual, se puede resolver el problema de inferencia.

Así la investigación se inicia con un modelo de probabilidad que nos dice como se distribuyen las observaciones $\underline{X} = x_1, x_2, \ldots x_n)$; expresamos nuestra información con la densidad $p(\underline{\theta})$ (antes de obtener los datos) y una vez que se observa $\underline{X}$ se calcula la distribución a posteriori de $\underline{\theta}$ vía el teorema de Bayes.

La base de la inferencia Bayesiana es la distribución a posteriori, porque los aspectos restantes dependen de esta distribución. A partir de la distribución a posteriori se pueden hacer inferencias directamente, aunque también se pueden construir estimadores puntuales o de intervalos. Si $\underline{\theta}$ es unidimensional, una gráfica de la distribución a posteriori de $\underline{\theta}$ nos

revelará el comportamiento de este parámetro; si $\underline{\theta}$ es de más de una dimensión, se podrían aislar aquellas componentes de $\underline{\theta}$ que sean de interés.

Información a priori.

De entre todos los aspectos de la teoría bayesiana, éste es el más difícil y controversial; muchos objetan tratar a los parámetros no observables $\underline{\theta}$, como variables aleatorias y darles distribución de probabilidad subjetiva o densidades $p(\underline{\theta})$. Sin embargo no mucha gente objeta darle a las cantidades observables X una distribución de probabilidad de tipo frecuentista; suponer que la variable aleatoria X tiene distribución de frecuencia, es suponer la existencia de una sucesión interminable de repeticiones del experimento, y por supuesto, ya que tales repeticiones en efecto no ocurren, uno debe suponer que ocurren. Por lo tanto suponer que X tiene una distribución de probabilidad tipo frecuentista podría ser tan subjetivo como suponer que $\underline{\theta}$ tiene asociada una ley de probabilidad subjetiva, sin embargo, habría que aceptar que la ley de probabilidad de X es más objetiva en el sentido de que X puede observarse y su ley de probabilidad checarse contra los datos. Apesar de las dificultades para proponer las distribuciones a priori, existen ciertos principios útiles para tal efecto como son: el principio de razón insoficiente, distribuciones a priori [illegible] medición precisa; principios que, a pesar de que se usan, no se describen en este trabajo.

Análisis posterior

Supongamos que se está satisfecho con el modelo de probabilidad $f(\underline{x}/\underline{\theta})$ para las observaciones y también con la densidad a priori para los parámetros. Ahora en la etapa de análisis posterior, lo primero que debe de realizarse es encontrar la distribución a posteriori de $\underline{\theta}$

$$p(\underline{\theta}/\underline{x}) \propto f(\underline{x}/\underline{\theta}) \cdot P(\underline{\theta})$$

por medio del teorema de Bayes y dependiendo de cual sea el objetivo (estimar $\underline{\theta}$ o alguna de sus componentes, probar hipótesis acerca de los parámetros, producir valores futuros de una observación etc) se procede al trabajo de inferencia tomando como base dicha distribución.

Aunque regularmente son de interés todas las componentes de $\underline{\theta}$, en algunas ocaciones se requieren solo algunos parámetros; sea $\underline{\theta}_1$, un subconjunto de las componentes de $\underline{\theta}$, entonces para hacer inferencias acerca de $\underline{\theta}_1$, se requiere calcular la densidad marginal a posteriori de $\underline{\theta}_1$

$$P_1(\underline{\theta}_1/\underline{x}) = \int_{\Omega_2} P(\underline{\theta}_1, \underline{\theta}_2/\underline{x})\, d\underline{\theta}_2, \underline{\theta} = (\underline{\theta}_1, \underline{\theta}_2)$$

naturalmente, se puede usar esta distribución para estimar y probar hipótesis concernientes a $\underline{\theta}_1$, pero estos esfimadores son caracteristicas de resumen de $P(\underline{\theta}_1/\underline{x})$ y solo nos dice parte de la historia acerca de los parámetros y no son un sustituto para la distribución total

3.2 Métodos Bayesianos para el Manejo de Observaciones Discrepantes.

3.2.1. Método de Box-Tiao

George E. P. Box y George C. Tiao (1968) son de las primeras personas en formular un modelo general para la detección de observaciones discrepantes en modelos lineales. Su modelo es el de la varianza inflada y el objetivo fundamental de éste es el de obtener inferencias acerca de $\underline{\beta}$.

Consideremos el modelo de regresión lineal con las hipótesis usuales

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \ ,$$

cuyas dimensiones se han mencionado anteriormente. Supongamos que el error asociado con cada una de las observaciones surge de alguna de estas dos fuentes: Un modelo central $N(0,\sigma^2)$ o del modelo alternativo $N(0,a^2\sigma^2)$ con una probabilidad $1-\alpha$ y α respectivamente, α conocida y también que a es fija, $\underline{\beta}$ y $\log\sigma$ son localmente independientes y uniformes a priori. Usando la misma notación empleada en el enfoque frecuentista, podemos descomponer a los elementos del modelo en $\underline{Y} = \begin{pmatrix} Y_i \\ Y_{(i)} \end{pmatrix}$; $X = \begin{pmatrix} X_i \\ X_{(i)} \end{pmatrix}$ y $\underline{\varepsilon} = \begin{pmatrix} \varepsilon_i \\ \underline{\varepsilon}_{(i)} \end{pmatrix}$. Definamos también E_i como el evento; κ observaciones son discrepantes donde κ se supone conocida; $i = \{i_1, i_2, \ldots i_\kappa\}$. Dadas las observaciones $\underline{Y}$, la verosimilitud del $(\underline{\beta}, \sigma, E_i)$ es

$$\begin{aligned} l(\underline{\beta}, \sigma, E_i) &= f(\underline{\varepsilon}_{(i)}, \sigma) \cdot g(\varepsilon_i | \sigma) \\ &\propto f(\underline{Y}_{(i)} - X_{(i)}\underline{\beta}/\sigma) \cdot g(Y_i - X_i\underline{\beta}/\sigma) \end{aligned} \qquad (3.2.1)$$

donde f, g representan el producto de las funciones de -

densidad de $\underline{\varepsilon}_{(i)}$ y $\underline{\varepsilon}_i$ respectivamente, E_i representa el evento, κ de los errores surgen de $g(\underline{\varepsilon}_i/\sigma)$ y el resto de $f(\underline{\varepsilon}_{(i)}/\sigma)$ con

$$\left.\begin{aligned} f(\underline{\varepsilon}/\sigma) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{\underline{\varepsilon}^2}{2\sigma^2}\right\} \\ g(\underline{\varepsilon}/\sigma) &= \frac{1}{a\sigma\sqrt{2\pi}} \exp\left\{-\frac{\underline{\varepsilon}^2}{2a^2\sigma^2}\right\} \end{aligned}\right\} \qquad (3.2.2.)$$

entonces la verosimilitud completa consiste de 2^n expresiones del tipo (3.2.1) correspondientes a todas las $\binom{n}{\kappa}$ combinaciones posibles para los errores; $\kappa = 0, 1, 2, \ldots n$.

Así, la distribución a posteriori para $(\underline{\beta}, \sigma, E_i)$ es

$$p(\underline{\beta}, \sigma, E_i/\underline{y}) = \frac{P(E_i)\cdot p(\underline{\beta},\sigma)\cdot f(\underline{y}_{(i)} - X_{(i)}\underline{\beta}/\sigma)\cdot g(\underline{y}_i - x_i\underline{\beta}/\sigma)}{\sum_i P(E_i)\, h(\underline{y}_i, \underline{y}_{(i)})} \qquad (3.2.3)$$

donde $P(E_i) = \alpha^\kappa(1-\alpha)^{n-\kappa}$ y tomando $p(\underline{\beta}, \sigma) \propto \frac{1}{\sigma}$ con

$$h(\underline{y}_i, \underline{y}_{(i)}) = \int_{\mathbb{R}} p(\underline{\beta}, \sigma)\cdot f(\underline{y}_{(i)} - X_{(i)}\underline{\beta}/\sigma)\cdot g(\underline{y}_i - x_i\underline{\beta}/\sigma)\, d\sigma\, d\underline{\beta}$$

es la distribución marginal de $\underline{y}$ bajo el supuesto de que los errores surgen de las fuentes citadas. Después de integrar (3.2.3) con respecto a σ, la distribución de $(\underline{\beta}, E_i)$ puede escribirse como

$$p(\underline{\beta}, E_i/\underline{y}) = P(E_i/\underline{y})\cdot P(\underline{\beta}/E_i, \underline{y})$$

donde

$$P(E_i/\underline{y}) = c\left(\frac{\alpha}{a(1-\alpha)}\right)^\kappa \cdot \frac{|X^tX|^{1/2}}{|X^tX - (1-\frac{1}{a^2})X_i^tX_i|^{1/2}} \left\{\frac{d_i^2}{A^2}\right\}^{-\frac{1}{2}(n-p)} \qquad (3.2.4)$$

con $s_i^2 = \frac{1}{n-p} S_i(\hat{\beta}_i)$;

$$\hat{\beta}_i = (X_{(i)}^t X_{(i)} + \tfrac{1}{a^2} x_i^t x_i)^{-1} (X_{(i)}^t \underline{y}_{(i)} + \tfrac{1}{a^2} x_i^t \underline{y}_i),$$

c es una constante que hace que la suma de $P(E_i/y)$ sea la unidad y

$$P(\beta/E_i, \underline{y}) \propto \int_0^\infty \sigma^{-(n+1)} \exp\left\{\frac{1}{2\sigma^2} S_i(\beta)\right\} d\sigma \qquad (3.2.5)$$
$$\propto \left\{S_i(\beta)\right\}^{-\frac{1}{2}n}$$

donde $S_i(\beta) = (\underline{y}_{(i)} - X_{(i)}\beta)^t(\underline{y}_{(i)} - X_{(i)}\beta)$
$+ \frac{1}{a^2}(\underline{y}_i - x_i\beta)^t(\underline{y}_i - x_i\beta)$ (3.2.6)

$= S_i(\hat{\beta}_i) + (\beta - \hat{\beta}_i)^t(X_{(i)}^t X_{(i)} + \frac{1}{a^2} x_i^t x_i)(\beta - \hat{\beta}_i)$ (3.2.7)

además si usamos la identidad

$$(X_{(i)}^t X_{(i)} + \tfrac{1}{a^2} x_i^T x_i)^{-1} = (X^T X - (1 - \tfrac{1}{a^2}) x_i^T x_i)^{-1}$$

$$= (X^T X)^{-1} + (1 - \tfrac{1}{a^2})(X^T X)^{-1} x_i^t$$
$$\{I - (1 - \tfrac{1}{a^2}) x_i (X^T X)^{-1} x_i^T\}^{-1} x_i (X^T X)^{-1}$$

entonces

$$\hat{\beta}_i = \hat{\beta} - (1 - \tfrac{1}{a^2})(X^t X)^{-1} x_i^t \{I - (1 - \tfrac{1}{a^2}) x_i (X^t X)^{-1} x_i^t\}^{-1} (\underline{y}_i - x_i \hat{\beta})$$
y
$$(n-p) s_i^2 = (n-p) s^2 - (1 - \tfrac{1}{a^2})(\underline{y}_i - x_i \hat{\beta})^t \{I - (1 - \tfrac{1}{a^2}) x_i (X^t X)^{-1} x_i^T\}(\underline{y}_i - x_i \hat{\beta}) \qquad (3.2.8)$$

usando de (3.2.4) a (3.2.8) y después de normalizar, la distribución a posteriori resulta

$$P(\underline{\beta}/\underline{y}\ E_i)=\frac{\Gamma(\frac{n}{2})\left|X^tX-(1-\frac{1}{a^2})X_i^tX_i\right|^{1/2}}{\Gamma(\frac{n-p}{2})(\pi(n-p)s_i^2)^{p/2}}\left\{1+\frac{(\underline{\beta}-\hat{\underline{\beta}}_i)^tX^tX-(1-\frac{1}{a^2})X_i^tX_i(\underline{\beta}-\hat{\underline{\beta}}_i)}{(n-p)s_i^2}\right\}^{-\frac{n}{2}} \quad (3.2.9)$$

que es una distribución t multivariada (p-dimensional) con $n-p$ grados de libertad, media $\hat{\underline{\beta}}_i$ y matriz de dispersión $s_i^2(X^tX-(1-\frac{1}{a^2})X_i^tX^t)^{-1}$. En particular, si $k=0$, la expresión (3.2.9) se reduce a la distribución t ordinaria con $n-p$ g.l. centrada en el estimador mínimos cuadrados $\hat{\underline{\beta}}$ y matriz de dispersión $s^2(X^tX)^{-1}$. Entonces, la distribución a posteriori para $\underline{\beta}$ es un promedio ponderado de distribuciones t-student, es decir

$$P(\underline{\beta}/\underline{y})=\sum_i P(E_i/\underline{y})\cdot P(\underline{\beta}/E_i,\underline{y}) \quad (3.2.10)$$

donde los pesos $w_i = P(E_i/\underline{y})$

Distribución marginal de β_j. Como se mencionó, el objetivo fundamental de este modelo es básicamente lograr inferencias acerca de los parámetros, entonces, si se cuenta con una distribución t-student, (3.2.9), (3.2.9) y (3.2.10); la distribución marginal de la j-ésima componente de $\underline{\beta}$ es

$$P(\beta_j/\underline{y})=\sum_i w_i P_i(\beta_j/\underline{y})$$

$$\text{con}\quad P_i(\beta_j/\underline{y})=s_i^{-1}(v_i^{jj})^{-\frac{1}{2}}P\left\{t_{n-p}=\frac{\beta_j-\hat{\beta}_{ij}}{s_i(v_i^{jj})^{1/2}}\right\},$$

β_j el j-ésimo elemento de $\hat{\underline{\beta}}_i$ y v_i^{jj} el j-ésimo elemento de la matriz $(X^tX-(1-\frac{1}{a^2})X_i^tX_i)^{-1}$. Se puede verificar que

$$\bar{\beta}_j=E(\beta_j/\underline{y})=\sum_i w_i\hat{\beta}_{ij} \qquad y$$

$$Var(\beta_j/y) = \sum_i \omega_i \left\{ \frac{n-\rho}{n-\rho-2} \right\} \Delta_i^2 v_i^{jj} + (\hat{\beta}_{ij} - \bar{\beta}_j)^2 \}$$

Mezcla de distribuciones. En el desarrollo del modelo se ha supuesto que el error asociado con cada observación puede surgir de cualquiera de las distribuciones expresadas en (3.2.2) con probabilidad $(1-\alpha)$ y α respectivamente, obteniendose que la distribución a posteriori para $\underline{\beta}$ es un promedio ponderado de 2^n distribuciones t-multivariadas. Dicho resultado, también se puede obtener de la siguiente manera: como

$$p(\xi/\sigma) = (1-\alpha)\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\xi^2}{2\sigma^2}\right) + \alpha \frac{1}{a\sigma\sqrt{2\pi}} \exp\left(-\frac{\xi^2}{2a^2\sigma^2}\right)$$

es una mezcla de dos distribuciones normales que contiene 3 parámetros (α, σ, a), entonces, la función de verosimilitud de $(\underline{\beta}, \sigma, a, \alpha)$ es

$$l(\underline{\beta}, \sigma, a, \alpha) \propto \sigma^{-n} \prod_{i=1}^{n} [(1-\alpha)\exp\{\tfrac{1}{2\sigma^2}(y_i - \underline{x}_i^t\underline{\beta})^2\} + \alpha a^{-1}\exp\{-\tfrac{1}{2a^2\sigma^2}(y_i - \underline{x}_i^t\underline{\beta})^2\}] \qquad (3.2.11)$$

donde $\underline{x}_i$ es el i-ésimo renglón de X. Adoptando $p(\underline{\beta}, \sigma) \propto \frac{1}{\sigma}$, entonces, para a, α fijos, la distribución a posteriori para $\underline{\beta}$ es

$$p(\underline{\beta}/a, \alpha, \underline{y}) = \gamma \int_0^\infty \sigma^{-1} l(\underline{\beta}, \sigma, a, \alpha/\underline{y})\, d\sigma$$

con $\gamma^{-1} = \int_{\mathbb{R}} \sigma^{-1} l(\underline{\beta}, \sigma, a, \alpha/\underline{y})\, d\sigma$,

Esta distribución es matematicamente equivalente a la distribución (3.2.10). En efecto, al expander el integrando en (3.2.11) y efectuando la integración término a término, se

obtiene exactamente la expresión (3.2.10)

En aplicaciones practicas de la distribución a posteriori de $\underline{\beta}$ en (3.2.10), es adecuado, reacomodar la suma de las 2^n posibilidades de la siguiente forma:

$$P(\underline{\beta}/\underline{Y}) = W_0 P_0(\underline{\beta}/\underline{y}) + \sum_{i=1}^{n} W_i P_i(\underline{\beta}/\underline{Y}) + \sum_{i<j}^{n} W_{ij} P_{ij}(\underline{\beta}/\underline{Y}) + \ldots \quad (3.2.12)$$

La distribución $P_0(\underline{\beta}/\underline{Y})$ en la expresión anterior sería apropiada si todas las observaciones surgen del modelo f de la expresión (3.2.2); La distribución $P_i(\underline{\beta}/\underline{Y})$ corresponde a la posibilidad de que toda observación en turno sea discrepante y así sucesivamente. Similarmente, W_0 es la probabilidad a posteriori de que ninguna observación sea mala; W_i representa a la probabilidad de que la i-ésima observación sea discrepante etc..

3.2.2. Método de Abraham - Box

Al igual que el método de Box-Tiao, B. Abraham y G. Box (1978) desarrollan un método cuyo objetivo fundamental también es el de hacer inferencias acerca de los parámetros en modelos lineales en presencia de observaciones discrepantes. Su modelo, considera una fuente distinta de contaminación

$$\underline{Y} = X\underline{\beta} + \lambda \underline{z} + \underline{\varepsilon} \qquad (3.2.13)$$

donde $\underline{z}$ es un vector de $n \times 1$, cuyos elementos tienen una probabilidad α de ser 1 y $1-\alpha$ de ser 0, la cantidad de contaminación λ se supone la misma para todas las observaciones discrepantes y los demás elementos del modelo están contenidos en el modelo de regresión lineal con las hipótesis usuales. Naturalmente $\underline{z}^t = \{z_1, z_2 \cdots z_n\}$ es un vector de k unidades y $n-k$ ceros, tal que

$$z_i = \begin{cases} 1 & \text{si la } i\text{-ésima observación es discrepante} \\ 0 & \text{si no lo es.} \end{cases}$$

Podemos observar que el modelo que se maneja en este método, es el modelo de la media trasladada cuya representación geométrica se encuentra reflejada en la figura 2.1.2. a

$\underline{z}$ conocida. Debe notarse que suponer que $\underline{z}$ es conocida, implica que $i = \{i_1, i_2, \ldots i_k\}$ es conocida y por lo tanto se conoce k, el número de observaciones discre

pantes. La distribución a posteriori de $\underline{\beta}$ dado $\underline{Y}$ y $\underline{Z}$ es

$$P(\underline{\beta}/\underline{Y},\underline{Z}) = \int P(\underline{\beta},\lambda,\sigma/\underline{Z})\cdot P(\underline{Y}/\underline{\beta},\lambda,\sigma,\underline{Z})\,d\lambda\,d\sigma. \qquad (3.2.14)$$

donde $P(\underline{\beta},\lambda,\sigma/\underline{Z})$ es la apriori de $(\underline{\beta},\lambda,\sigma/\underline{Z})$ y $P(\underline{Y}/\underline{\beta},\lambda,\sigma,\underline{Z})$ es la distribución conjunta de $\underline{Y}$ dado $(\underline{\beta},\lambda,\sigma,\underline{Z})$. Esta distribución conjunta puede escribirse como

$$P(\underline{Y}/\underline{\beta},\lambda,\sigma,\underline{Z}) = (2\pi\sigma^2)^{-\frac{1}{2}n}\exp\left[(2\sigma^2)^{-1}\{(\underline{Y}-X\underline{\beta})^t(\underline{Y}-X\underline{\beta}) + k\lambda^2 - 2(\underline{Y}-X\underline{\beta})^t\underline{Z}\lambda\}\right] \qquad (3.2.15)$$

donde $k = \underline{Z}^t\underline{Z}$. Suponiendo que $k \neq 0$, n y que X^tI_iX es no singular, entonces (3.2.15) puede reescribirse como

$$P(\underline{Y}/\underline{\beta},\lambda,\sigma,\underline{Z}) = (2\pi\sigma^2)^{-\frac{1}{2}n}\exp\left[-(2\sigma^2)^{-1}\{(\underline{\beta}-\underline{\beta}^*)^t(X^tI_iX)(\underline{\beta}-\underline{\beta}^*) + k(\lambda-\lambda_{\underline{\beta}})^2 + \nu S_k^2\}\right] \qquad (3.2.16)$$

donde $I_i = I - \frac{1}{k}\underline{Z}\underline{Z}^t$; $\lambda_{\underline{\beta}} = \frac{1}{k}(\underline{Y}-X\underline{\beta})^t\underline{Z}$; $\underline{\beta}_i = (X^tI_iX)^{-1}X^tI_i\underline{Y}$ y

$\nu S_k^2 = \underline{Y}^tI_i(I - X(X^tI_iX)^{-1}X^t)I_i\underline{Y}$.

Supongamos que $\underline{\beta}, \lambda, \log\sigma$ son uniformes a priori y localmente independientes, es decir,

$$P(\underline{\beta},\lambda,\sigma/\underline{Z})\,d\underline{\beta}\,d\lambda\,d\sigma \propto \frac{1}{\sigma}. \qquad (3.2.17)$$

Ahora, multiplicando (3.2.16) por (3.2.17) e integrando con respecto a λ y σ obtenemos

$$P(\underline{\beta}/\underline{Y},\underline{Z})\,d\underline{\beta} \propto \left(\frac{1+(\underline{\beta}-\underline{\beta}_i)^t(X^tI_iX)(\underline{\beta}-\underline{\beta}_i)}{\nu S_k^2}\right)^{-\frac{1}{2}(\nu+p)} \qquad (3.2.18)$$

; donde $\nu = n-p-1$.

Entonces, si tenemos un conjunto de k-observaciones discrepantes, la distribución a posteriori de $\underline{\beta}$, será una distribución t-dimensional con media $\underline{\beta}_i$ y matriz de dispersión

$S_K^2 (X^t V X)^{-1}$ con $\nu = n-p-1$ g. l.

$\underline{z}$ Desconocida. Supongamos que existe una probabilidad α (conocida) de que cualquier elemento de $\underline{z}$ sea la unidad y $1-\alpha$ de que sea cero. Si $\underline{u}_K$ denota un vector de $n \times 1$ con K unidades y $(n-K)$ ceros, entonces, a priori tenemos

$$P(\underline{z} = \underline{u}_K / \alpha) = \alpha^K (1-\alpha)^{n-K} \qquad K = 0, 1, 2, \ldots n.$$

De aquí en adelante se escribirá $\underline{z}$ para indicar $\underline{z} = \underline{u}_K$ y se evitará la condicional sobre α

Dado $\underline{y}$ la distribución a posteriori de $(\beta, \lambda, \sigma, \underline{z})$ está dada por

$$P(\beta, \lambda, \sigma, \underline{z}/\underline{y}) = \frac{P(\underline{z}) \cdot P(\beta, \lambda, \sigma) \cdot P(\underline{y}/\beta, \lambda, \sigma, \underline{z})}{\sum_{\underline{z}} P(\underline{z}) \cdot P(\underline{y}/\underline{z})} \qquad (3.2.19)$$

donde $P(\underline{y}/\beta, \lambda, \sigma, \underline{z})$ está dada en (3.2.16) y

$$P(\underline{y}/\underline{z}) = \int P(\beta, \lambda, \sigma / \underline{z}) P(\underline{y}/\beta, \lambda, \sigma, \underline{z}) d\lambda \, d\beta \, d\sigma$$

La suma en (3.2.19) es sobre los 2^n posibles valores de $\underline{z}$. Integrando (3.2.19) con respecto a λ y σ se obtiene

$$P(\beta, \underline{z}/\underline{y}) = P(\underline{z}) \cdot P(\beta/\underline{y}, \underline{z}),$$

donde $P(\beta/\underline{y}, \underline{z})$ está dada en (3.2.18) y

$$P(\underline{z}/\underline{y}) = P(\underline{z}) \cdot P(\underline{y}/\underline{z}) / \sum_{\underline{z}} P(\underline{z}) \cdot P(\underline{y}/\underline{z})$$

$$= c\{P(\underline{z})/P(\underline{z}=0)\}\{P(\underline{y}/\underline{z}=0)\} \qquad (3.2.20)$$

con $c = P(\underline{z}=0)\ P(\underline{y}/\underline{z}=0)\ /\ \Sigma P(\underline{z})\ P(\underline{y}/\underline{z})$,

Por lo tanto $P(\underline{\beta}/\underline{y}) = \sum_{\underline{z}} P(\underline{z}/\underline{y})\, P(\underline{\beta}/\underline{y}\,\underline{z})$

donde $P(\underline{\beta}/\underline{y},\underline{z})$ esta dada en (3.2.18) y $P(\underline{z}/\underline{y})$ en (3.2.20) que también se puede escribir como

$$P(\underline{z}/\underline{y}) = c\left(\frac{\alpha_i}{\sum \alpha_i}\right)^{\kappa}\left(\frac{\kappa|x^t x|}{\nu_0 s_0^2 |x^t x|}\right)^{-\frac{1}{2}}\left(\frac{\nu r s_R^2}{\nu_0 s^2}\right)^{-\frac{1}{2}\nu} B\left(\tfrac{1}{2}, \tfrac{1}{2}\nu\right).$$

con c una constante, tal que, $\Sigma P(\underline{z}/\underline{y}) = 1$. Entonces la distribución a posteriori de $\underline{\beta}$ dadas las observaciones es un promedio ponderado de distribuciones t.

Inferencias acerca de λ. El modelo que se ha considerado aquí, es una mezcla de dos distribuciones normales, donde solo una de ellas se considera correcta. Este modelo tiene implicaciones en el caso donde se sospecha de observaciones "malas" y se esta interesado en encontrar, cuanto se desvian dichas observaciones de la media de las observaciones buenas.

Como $P(\underline{\beta}, \lambda, \sigma) \propto \frac{1}{\sigma}$ y

$$P(\underline{y}/\underline{\beta}, \lambda, \sigma, \underline{z}) = (2\pi\sigma^2)^{-\frac{1}{2}n} \exp\{(2\sigma^2)^{-1}[\underline{y} - x\underline{\beta}) + \kappa\lambda^2 - 2(\underline{y} - x\underline{\beta})^t \underline{z}\lambda]\}$$

y suponiendo que $\kappa \neq 0$ y $\underline{z}^t M \underline{z} \neq 0$; $M = I - x(x^t x)^{-1} x^t$,

entonces $P(\underline{\beta}, \lambda, \sigma/\underline{y}, \underline{z}) = (2\pi)^{-\frac{1}{2}n} \sigma^{-(n+1)} \exp\{-(2\sigma^2)^{-1}[(\underline{\beta} - \underline{\beta}_A)^t x^t x (\underline{\beta} - \underline{\beta}_A) + (\underline{z}^t M \underline{z})(\lambda - \bar{\lambda}^*)^2 + \nu S^2 d]\}$,

con $\hat{\beta}_\lambda = (X^t X)^{-1}(\underline{Y} - \lambda \underline{z})$, $\lambda^* = (\underline{z}^t M \underline{z})^{-1} \underline{z}^t M \underline{Y}$ y

$$\nu S_d^2 = \underline{y}^t M [I - \underline{z}(\underline{z}^t M \underline{z})^{-1} \underline{z}^t] M \underline{y} .$$

Integrando la expresión (3.2.21) con respecto a β y σ obtenemos

$$P(\lambda / \underline{Y}, \underline{z}) \propto \left(1 + \frac{\underline{z}^t M \underline{z} (\lambda - \lambda^*)^2}{\nu S_d^2}\right)^{-\frac{\nu+1}{2}},$$

que es una distribución "escalada" t-student.

3.2.2 Método de Guttman, Freeman y Dutter

I. Guttman, P. R. Freeman y R. Dutter (1978) plantean un modelo más general que los anteriores para la detección de observaciones discrepantes. Ellos suponen que K de las observaciones son generadas por el modelo

$$Y_{ij} = x_{ij}^t \beta + \lambda_j + \varepsilon_{ij} \quad ; \quad j = 1, 2, \ldots K \qquad (3.2.22)$$

con $i_j \in \underline{i} = \{i_1, i_2, \ldots i_K\}$ y el número restante $(n-K)$ son generadas por el modelo tradicional

$$\underline{Y} = X \beta + \underline{\varepsilon} . \qquad (3.2.23)$$

Podemos observar que el modelo propuesto, es el modelo de la media trasladada donde las observaciones discrepantes surgen de diferentes poblaciones, situación reflejada geometricamente en la figura 3.2.1 b.

Supongamos que cualquier subconjunto de K-observaciones puede ser generado por el modelo (3.2.22) con una probabilidad uniforme a priori de $1/\binom{n}{K}$ y que la información

a priori acerca de $\underline{\beta}, \sigma^2$ y λ es muy vaga y por lo tanto

$$p(\lambda, \underline{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}. \tag{3.2.23}$$

entonces la verosimilitud de λ, $\underline{\beta}$ y σ^2 dado $\underline{Y}$ es

$$l(\lambda, \underline{\beta}, \sigma^2 | \underline{y}) \propto \sum_i^{\prime} \frac{1}{(n)} (\sigma^2)^{-(n-k)/2} \exp\left\{-\frac{1}{2\sigma^2}[(\underline{y}_{(i)} - X_{(i)}\underline{\beta})^t (\underline{y}_{(i)} - X_{(i)}\underline{\beta})]\right\} \times$$
$$\left\{(\sigma^2)^{-\frac{k}{2}} \exp\left\{\frac{1}{2\sigma^2}[(\underline{y}_i - X_i\underline{\beta} - \lambda)(\underline{y}_i - X_i\underline{\beta} - \lambda)]\right\}\right\} \tag{3.2.24}$$

combinando (3.2.23) con (3.2.24) y usando el teorema de Bayes, encontramos que la distribución a posteriori de $(\lambda, \underline{\beta}, \sigma^2)$ dado $\underline{Y}$ es

$$p(\lambda, \underline{\beta}, \sigma^2 | \underline{y}) = \sum_i^{\prime} (\sigma)^{-\frac{n}{2}+1} \exp Q, \tag{3.2.25}$$

donde

$$Q = S_{(i)} + (X_{(i)} X_{(i)}^{-} \underline{y}_{(i)} - X_{(i)}\underline{\beta})^t (X_{(i)} X_{(i)}^{-} \underline{y}_{(i)} - X_{(i)}\underline{\beta}) + [\lambda - (\underline{y}_i - X_i\underline{\beta})]^t [\lambda - (\underline{y}_i - X_i\underline{\beta})] \tag{3.2.26}$$

con $S_{(i)} = (\underline{y}_{(i)} - X_{(i)} X_{(i)}^{-} \underline{y}_{(i)})^t (\underline{y}_{(i)} - X_{(i)} X_{(i)}^{-} \underline{y}_{(i)})$

y $X_{(i)}^{-}$ satisface

$$\left.\begin{aligned} &a) \quad (X_{(i)} X_{(i)}^{-})^t = X_{(i)} X_{(i)}^{-}, \\ &b) \quad X_{(i)} X_{(i)}^{-} X_{(i)} = X_{(i)}, \end{aligned}\right\} \tag{3.2.27}$$

Si $X_{(i)}$ es el rango de p, entonces $X_{(i)}^{-} = (X_{(i)}^t X_{(i)})^{-1} X_{(i)}^t$ y $S_{(i)}$ es la suma de cuadrados residual usual

Para encontrar la distribución marginal a posteriori para $\underline{\beta}$, debemos integrar (3.2.25) con respecto a λ y σ^2.

$$P(\underline{\beta}, \sigma^2/\underline{y}) \propto \sum_i (\sigma^2)^{-\left(\frac{n}{2}+1\right)} \exp \frac{1}{2\sigma^2} Q \cdot \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp\left[\frac{1}{2\sigma^2}(\underline{\beta}-\underline{\varepsilon}_i)^t(\underline{\beta}-\underline{\varepsilon}_i)\right] d\alpha_1, \ldots d\alpha_n$$

$$\propto \sum_i (\sigma^2)^{-\left(\frac{n-\kappa}{2}+1\right)} \exp\left(-\frac{1}{2\sigma^2} Q_{(i)}\right) \qquad (3.2.2)$$

con $Q_{(i)} = S_{(i)} + (\underline{\beta} - \hat{\underline{\beta}}_{(i)})^t X_{(i)}^t X_{(i)} (\underline{\beta} - \hat{\underline{\beta}}_{(i)})$ y $\hat{\underline{\beta}}_{(i)}$ satisface:

$$X_{(i)}^t X_{(i)} \hat{\underline{\beta}}_{(i)} = X_{(i)} \underline{y}_{(i)} \quad \beta_{(i)} = X_{(i)}^- \underline{y}_{(i)},$$

ahora, integrando (3.2.2) con respecto a σ^2, se obtiene

$$P(\underline{\beta}/\underline{y}) \propto \sum_i \left[S_{(i)} + (\underline{\beta} - \hat{\underline{\beta}}_{(i)})^t X_{(i)}^t X_{(i)} (\underline{\beta} - \underline{\beta}_{(i)}) \right]^{-\frac{n-\kappa}{2}}$$

$$\propto \sum_i \left\{ S_i^{-\frac{n-\kappa}{2}} \left[1 + (\underline{\beta} - \hat{\underline{\beta}}_{(i)})^t (S_{(i)}^{-1} (X_{(i)}^t X_{(i)})) (\underline{\beta} - \hat{\underline{\beta}}_{(i)}) \right] \right\} \qquad (3.2.29)$$

Como se puede observar, $P(\underline{\beta}/\underline{y})$ es un promedio ponderado de 2^n distribuciones t de student.

$$t(\underline{w}/\underline{w}_0 ; B, m; \rho) = \frac{\Gamma\left(\frac{\rho+m}{2}\right) |B|^{1/2}}{(\Gamma(\tfrac{1}{2}))^{\rho} (m)^{\rho/4} \Gamma\left(\frac{m}{2}\right)} \left[1 + \frac{(\underline{w} - \underline{w}_0)^t B (\underline{w} - \underline{w}_0)}{m} \right]^{-\frac{\rho+m}{2}} \qquad (3.2.30)$$

es decir, si suponemos que $X_{(i)}^t X_{(i)}$ es positiva definida para todo $i = \{i_1, i_2, \ldots i_\kappa\}$, podemos escribir (3.2.29) como.

$$P(\underline{\beta}/\underline{y}) = \sum_i w_i T\left(\underline{\beta} / \hat{\underline{\beta}}_{(i)} ; \frac{n-\kappa-\rho}{S_{(i)}} X_{(i)}^t X_{(i)} ; n-\kappa-\rho ; \rho\right)$$

donde $w_i = \left[S_{(i)}^{-(n-\kappa)} \left(S_{(i)}^{\rho} / (X_{(i)}^t X_{(i)})^{-1} / \right) \right]^{1/2} \Big/ \sum_i \left[S_{(i)}^{-(n-\kappa)} \left(S_{(i)}^{\rho} \left[(X_{(i)}^t X_{(i)})^{-1} \right] \right) \right]^{1/2}$

son las probabilidades a posteriori de que las observaciones $Y_{i_1}, \ldots Y_{i_\kappa}$ sean discrepantes.

3.2.4. Método de Dutter y Guttman

Este trabajo también utiliza el modelo de la media trasladada

$$\underline{y} = X\underline{\beta} + \lambda \underline{z} + \underline{\varepsilon},$$

que utilizan Abraham y Box (1978), con la diferencia de que $\underline{z}$ toma el valor uno o el valor cero con la misma probabilidad. Entonces, para este modelo con K observaciones trasladadas de la media (K fija y conocida) por una cantidad λ (desconocida) la verosimilitud de $(\underline{\beta}, \sigma^2, \lambda)$ dadas las observaciones es

$$l(\underline{\beta}, \sigma^2, \lambda/\underline{y}) = \sum_i \left\{ \prod_{i \notin i} \left(\tfrac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\tfrac{1}{2\sigma^2}(y_i - \underline{x}_i^T\underline{\beta})^2\right] \right\} \left\{ \prod_{i \in i} \left(\tfrac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\tfrac{1}{2\sigma^2}(y_i - \underline{x}_i^T\underline{\beta} - \lambda)^2\right] \right\}, \tag{3.2.31}$$

donde i es el conjunto de indices que identifican a las K observaciones discrepantes y que ha sido definido anteriormente. Bajo el supuesto de que K observaciones están trasladadas y que $\underline{\beta}$, $\log \sigma^2$ y λ son localmente uniformes e independientemente distribuidas se tiene que la distribución a priori de $(\underline{\beta}, \sigma^2, \lambda)$ es proporcional a $\frac{1}{\sigma^2}$. Con esta información se obtiene la distribución a posteriori de $(\underline{\beta}, \sigma^2, \lambda)$ como:

$$p(\underline{\beta}, \sigma^2, \lambda/\underline{y}) \propto \sum_{i \in i} \frac{1}{\sigma^{n+2}} \exp\left\{-\tfrac{1}{2}\left[\sum_{i}(y_i - \underline{x}_i^T\underline{\beta}) + \sum_{i \in i}(y_i - \underline{x}_i^T\underline{\beta} - \lambda)^2\right]\right\}$$

$$= \sum_i \frac{1}{\sigma^{n+2}} \exp\left\{-\tfrac{1}{2\sigma^2}\left[S_i + (\underline{\beta} - \hat{\underline{\beta}}_i)^T X^T I_i X (\underline{\beta} - \hat{\underline{\beta}}_i) + K(\lambda - \tfrac{1}{K}(\underline{y} - X\underline{\beta})^T\underline{z})^2\right]\right\} \tag{3.2.32}$$

con $\underline{x}_i^T$ al iésimo renglón de la matriz X,

$$S_i = (\underline{y} - X\hat{\underline{\beta}})^T I_i (\underline{y} - X\hat{\underline{\beta}}),$$

$$I_i = I - \frac{1}{k} \underline{z}\underline{z}^T \ ,$$

$$\hat{\underline{\beta}}_i = (X^T I_i X)^{-1} X^T I_i \underline{y}.$$

Se puede demostrar que I_i es simétrica e idempotente, y que $X^T I_i X$ es positiva definida.

A partir de la expresión (3.2.31) se puede encontrar la distribucion marginal de $\underline{\beta}$ dadas las observaciones al integrar con respecto a λ y σ^2 para obtener que:

$$P(\underline{\beta}/\underline{y}) \propto \sum_i [S_i + (\underline{\beta} - \hat{\underline{\beta}}_i)^T X^T I_i X (\underline{\beta} - \hat{\underline{\beta}}_i)]^{-\frac{n-1}{2}}$$

$$\propto \sum_i S_i^{-\frac{n-1}{2}} [1 + \frac{1}{S_i} (\underline{\beta} - \hat{\underline{\beta}}_i)^T X^T I_i X (\underline{\beta} - \hat{\underline{\beta}}_i)]^{-\frac{n-1}{2}}$$

$$= \sum_i W_i \, t_i (\underline{\beta}/\hat{\beta}_i \, ; \frac{S_i}{n-p-1} (X^T I_i X)^{-1}, \, n-p-1) \qquad (3.2.33)$$

donde $t_p(\underline{\beta}, \hat{\underline{\beta}}, \Delta^2 (X^T X)^{-1}, \nu) = \frac{\Gamma(\frac{\nu+p}{2}) / (X^T X)^{1/2}}{[\Gamma(\frac{\nu}{2})]^p \, \Gamma(\frac{\nu}{2}) (\Delta \sqrt{\nu})^p} \left[1 + \frac{(\underline{\beta} - \hat{\underline{\beta}})^T (X^T X) (\underline{\beta} - \hat{\underline{\beta}})}{\Delta^2 \nu}\right]^{-\frac{\nu+p}{2}}$,

$$y \ W_i = \frac{(S_i)^{-\frac{n-p-1}{2}}}{|X^T I_i X|^{1/2}} \Big/ \sum_i \frac{(S_i)^{-\frac{n-p-1}{2}}}{|X^T I_i X|^{1/2}}$$

Como se puede ver $P(\underline{\beta}/\underline{y})$ es una suma ponderada de $\binom{n}{k}$ distribuciones t multivariadas.

Aunque en algunos casos, se requieren algunas modificaciones se encuentran las distribuciones marginales a posteriori de σ^2 y λ con los siguientes resultados:

$$P(\sigma^2/\underline{y}) = \sum_i W_i \, f(\sigma^2/S_i, n-p-1) \qquad (3.2.34)$$

donde f denota la densidad de la χ^2 invertida con $(n-p-1)$ grados de libertad (Box - Tiao 1978).

$$f(\sigma^2/\gamma, \nu) = [\Gamma(\tfrac{\nu}{2})\ 2^{\nu/2}]^{-1}\ (\sigma^2)^{-(\frac{\nu}{2}+1)}\ (\gamma)^{\nu/2}\ \exp\{-\frac{\gamma}{2\sigma^2}\}$$

y

$$p(\lambda/\underline{y}) = \sum_i W_i\, t_i(\lambda \mid \tfrac{1}{m_i};\ \underline{y}^T M \underline{z},\ \tfrac{V_i}{(n-p-1)},\ n-2p) \qquad (3.235)$$

con $m_i = \underline{z}^T M \underline{z}^T$,

$M = I - x(x^T x)^{-1} x^T$,

$V_i = \underline{y}^T M (I - \frac{1}{m_i}\ \underline{z}\ \underline{z}^T) M \underline{y}$

y $W_i = \dfrac{[V_i]^{-\frac{n-p-1}{2}}}{[m_i]^{1/2}} \Big/ \sum_i \dfrac{[V_i]^{-\frac{n-p-1}{2}}}{[m_i]^{1/2}}$.

Es claro que $p(\lambda/\underline{y})$ también es una suma ponderada de distribuciones t- univariadas.

3.3 Comentarios.

Como podemos observar, los métodos Bayesianos que aquí se presentan modelan ligeramente diferente el aspecto de observaciones discrepantes. Box y Tiao (1968), suponen que la distribución del error es del tipo

$$F = (1-\alpha)\, N(0\sigma^2) + \alpha N(0, a^2\sigma^2),$$

$a^2 > 1$; a y a^2 son valores que, dependiendo de sus intereses, el investigador debe especificarlos. Por otro lado, Abraham y Box (1978) aplican estas ideas al modelo de la media trasladada con la restricción que todas las observaciones contaminantes provienen de la misma población. Este modelo es casi idéntico al de Dutter y Guttman con la única diferencia de que en este modelo todas las observaciones tienen la misma probabilidad de ser discrepantes. Esta última característica también es común al modelo de Guttman, Freeman y Dutter (que tambien utiliza el modelo de la media trasladada) pero la diferencia radica en que el trasladado (shift) para cada observación no es el mismo.

Al aplicar el paradigma Bayesiano para cada situación descrita, y dado que el aspecto principal es la estimación de los parámetros, observamos que en todos los casos, la distribución de $\underline{\beta}$ dado $\underline{Y}$ es una suma ponderada de distribuciones t multivariadas. ver ecuaciones (3.2.10), (3.2.18), (3.2.29), (3.2.33).

Cada uno de estos enfoques puede considerarse como un método de acomodo, cuya base es la distribución a posteriori respectiva de $\underline{\beta}$ o un método de identificación basado en la probabilidad a posteriori W_i, donde i identifica a las observaciones contaminantes. Por ejemplo, Box y Tiao sugieren solo incluir términos seleccionados en la expresión (3.2.12) y "renormalizar" los pesos: Si no se sospecha de observaciones discrepantes, entonces sólo el primer término de dicha expresión es relevante y $P(\underline{\beta}/\underline{Y}) = P_0(\underline{\beta}/\underline{Y})$; si solo se sospecha de a lo más una observación discrepante, $P(\underline{\beta}/\underline{Y}) \propto w_0 P_0(\underline{\beta}/\underline{Y}) + \sum_{i=1}^{n} w_i P_i(\underline{\beta}/\underline{Y})$; si se sospecha de dos contaminantes, se agrega el término $\sum_{i<j} w_{ij} P_{ij}(\underline{\beta}/\underline{Y})$ y así sucesivamente. En este enfoque, el problema de especificar el número de observaciones discrepantes se transporta en escoger el

número de términos para $P(\beta/\underline{y})$.

Por otro lado, el uso de cualquiera de los modelos expuestos dependerá de los intereses del investigador.

3.4 Ejemplos.

Con la finalidad de mostrar algunos de los hechos mencionados anteriormente con respecto a los métodos Bayesianos, en esta sección únicamente se desarrollará un ejemplo en el cuál se usará el Método de Ditter y Guttman.

Los datos que manejaremos en este ejemplo, estan contenidos en la tabla (3.4.1) y se obtuvieron por medio de un generador normal pseudo-aleatorio tal que la distribución de las primeras 8 observaciones es $N(0, 1)$, mientras que las dos últimas están trasladadas con $\theta = 2$.

tabla 3.4.1

i	y_i
1	-0.44168
2	0.80814
3	-0.01549
4	0.01265
5	-1.08758
6	-1.38324
7	0.48106
8	-0.59308
9	2.32219
10	2.05387

Sin conocimiento de la forma en que se generan los datos y si suponemos que hay una observación discrepante, entonces, al concentrar nuestra atención en la figura (3.4.1) observamos que precisamente las densidades t correspondientes a estas dos últimas observaciones están separadas claramente del resto de las densidades, mas aún, $P(\underline{\beta}/\underline{Y})$ también se encuentra dominada por estas dos densidades, de tal forma que, podríamos pensar que estas observaciones son discrepantes. Este hecho también puede constatarse si observamos los valores de $\beta_{(i)}$, W_i y S_i que se encuentran en la tabla (3.4.2). Como podemos notar, las observaciones con mayor peso también lo son la 9 y 10 para las cuales se obtiene una mayor reducción en la suma de cuadrados al eliminar dichas observaciones.

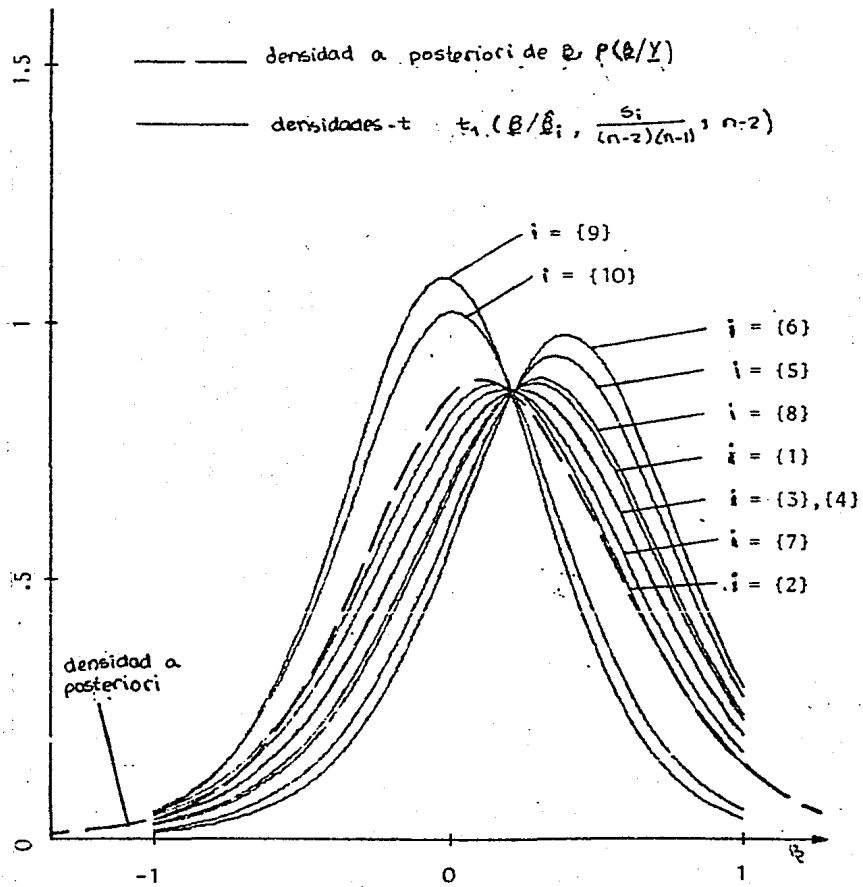
tabla 3.4.2.

$k = 1$

i=i	$\hat{\beta}_{(i)}$	S_i	W_i
1	.2887	13.19	.0562
2	.1499	13.28	.0547
3	.2414	13.61	.0496
4	.2382	13.63	.0494
5	.3605	11.79	.0883
6	.3933	10.83	.1237
7	.1862	13.60	.0499
8	.3056	12.95	.0606
9	-.0184	8.74	.2916*
10	.0114	9.92	.1760*

* pesos dominantes

figura 3.4.1.


densidad a posteriori de β P(β/Y)
densidades-t $t_{n-2}(\beta/\hat{\beta}_i, \frac{s_i}{(n-2)(n-1)}, n-2)$
i = {9}
i = {10}
i = {6}
i = {5}
i = {8}
i = {1}
i = {3},{4}
i = {7}
i = {2}
densidad a posteriori
1.5
1
.5
0
-1
0
1
β

Igualmente, para $k=2$, la tabla (3.4.3) nos proporciona los valores $\hat{\beta}_{(i)}$, $\hat{\beta}_i$, S_i y w_i para las $\binom{10}{2}$ posibles parejas. Los pesos dominantes son para $i=\{9,10\}$; $w_i=0.6489$ y $i=\{5,6\}$; $w_i=0.032$ que es más de 20 veces menor que $w(9,10)$. Por estas razones, a las observaciones 9 y 10 se les declara discrepantes.

tabla 3.4.3.

$k = 2$

	i		$\hat{\beta}_{(i)}$	$\hat{\beta}_i$	S_i	w_i
1	1	2	.2258	0.183	13.67	.0045
2	1	3	.3268	-0.229	13.18	.0052
3	1	4	.3252	-0.215	13.21	.0052
4	1	5	.4608	-0.765	11.27	.0098
5	1	6	.4977	-0.912	10.49	.0130
6	1	7	.2647	0.020	13.58	.0046
7	1	8	.3990	-0.517	12.33	.0068
8	1	9	.0345	0.940	12.36	.0068
9	1	10	.0681	0.806	12.80	.0059
10	2	3	.1705	0.396	13.59	.0046
11	2	4	.1670	0.410	13.58	.0046
12	2	5	.3045	-0.140	13.36	.0050
13	2	6	.3415	-0.288	13.04	.0055
14	2	7	.1085	0.645	13.21	.0052
15	2	8	.2427	0.108	13.64	.0046
16	2	9	-.1217	1.565	9.12	.0228
17	2	10	-.0882	1.431	9.98	.0159
18	3	4	.2700	-0.001	13.56	.0047
19	3	5	.4075	-0.552	12.20	.0071
20	3	6	.4445	-0.699	11.58	.0088
21	3	7	.2114	0.233	13.67	.0045
22	3	8	.3457	-0.304	13.00	.0055
23	3	9	-.0187	1.153	11.48	.0091
24	3	10	.0148	1.019	12.06	.0075
25	4	5	.4040	-0.537	12.26	.0070
26	4	6	.4409	-0.685	11.64	.0086
27	4	7	.2079	0.247	13.67	.0045
28	4	8	.3422	-0.290	13.03	.0055
29	4	9	-.0223	1.167	11.41	.0093
30	4	10	.0113	1.033	12.00	.0076

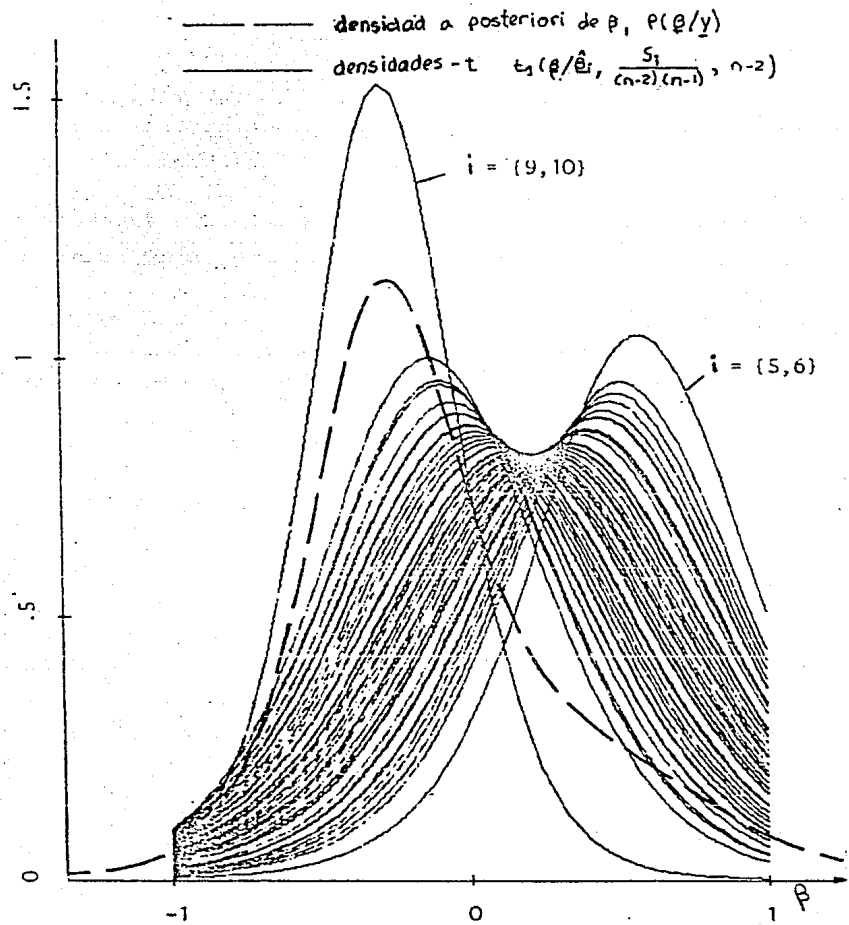
Tabla 3.4.3. (continuación)

31	5	6	.5785	-1.235	8.41	.0315
32	5	7	.3454	-0.303	13.00	.0055
33	5	8	.4797	-0.840	10.89	.0112
34	5	9	.1153	0.617	13.27	.0051
35	5	10	.1488	0.483	13.50	.0048
36	6	7	.3824	-0.451	12.56	.0063
37	6	8	.5167	-0.988	10.05	.0155
38	6	9	.1522	0.469	13.51	.0047
39	6	10	.1858	0.335	13.64	.0046
40	7	8	.2836	-0.056	13.49	.0048
41	7	9	-.0808	1.402	10.16	.0148
42	7	10	-.0473	1.267	10.91	.0111
43	8	9	.0535	0.865	12.62	.0062
44	8	10	.0870	0.730	13.01	.0055
45	9	10	-.2774	2.188	3.95	.6489*)

* pesos dominantes

Al igual que para K=1, esta conclusión también puede obtenerse al observar la figura (3.4.2) en donde claramente la densidad para la pareja (9,10) domina a las restantes incluyendo a $P(\underline{B}/\underline{Y})$ que es la más cercana a esta densidad.

figura 3.4.2.


densidad a posteriori de β, P(β/y)
densidades - t
1.5
1
.5
0
i = {9, 10}
i = {5, 6}
-1
0
1
β

3.5 Conclusiones

Una de las primeras cosas que resultan al exponer los métodos bayesianos es que, al menos teoricamente, se puede modelar cualquier fuente de discrepancia concebible (modelo de la media trasladada, de la varianza inflada, etc), obteniendose para cada situación un modelo. Naturalmente, en la medida en que se agreguen más condiciones a éste, la dificultad y complejidad en la interpretación de los resultados aumentará.

Por otro lado, parece ser que los métodos bayesianos unicamente contemplan a las observaciones discrepantes en el sentido de contaminantes y no en el sentido de sorprendentes o discrepantes, ver L.I. Pettit y A.F.M. Smith (1983).

Existen varios problemas asociados a los metodos bayesianos, el primero en el aspecto teórico: ya se mencionó que en la medida en que se aumenten el número de condiciones a modelarse, el modelo resultante será mucho más complejo. El segundo, en el aspecto práctico: como cada situacion requiere un modelo, entonces, es difícil usar los métodos en forma rutinaria porque cada uno requiere un programa especial. Otro problema práctico es la explosión computacional que surge al tratar de detectar el numero de observaciones discrepantes.

Capítulo 4

ENFOQUE ROBUSTO.

4.1. Introducción

La palabra robusto tiene varias connotaciones (algunas de ellas inconsistentes), en este trabajo usaremos este concepto en un sentido relativamente restringido: para nuestros propósitos, robustes significa insensibilidad a pequeñas desviaciones de las suposiciones inherentes al modelo

Como se sabe, las inferencias estadísticas se basan solo en parte en las observaciones Otro aspecto igualmente importante está formado por las suposiciones a priori acerca de la distribución principal. Aún en los casos mas simples, existen suposiciones implícitas o explícitas acerca de la aleatoridad e independencia de las observaciones, acerca de los modelos de distribución y en algunas ocaciones para las distribuciones a priori para algunos parámetros desconocidos. Naturalmente, dichas suposiciones no estan impuestas por ser exactas, mas bien, son supuestos simplificadores usados en la mayoria de las ramas matemáticas que expresan un conocimiento o certeza del fenómeno estudiado y su uso queda justificado por el principio de estabilidad: "El menor error en el modelo matemático solo debe de causar un error pequeño en las condiciones finales. Desafortunadamente, no siempre se presenta esta situación.

Supongamos que tenemos un modelo paramétrico, con la esperanza de que sea una buena aprozimación a la situación verdadera, pero no se debe ni se puede suponer que sea exactamente el correcto, es por eso que cualquier procedimiento estadístico debe de tener las siguientes carácte-

rísticas deseables.

(a) Debe tener una eficiencia (óptima o cerca del óptimo) razonablemente buena para el modelo propuesto.

(b) Debe de ser robusto en el sentido de que pequeñas desviaciones de las suposiciones del modelo deben de afectar sólo ligeramente su funcionamiento.

(c) Algunas desviaciones grandes del modelo no deben de causar una catastrofe.

¿Que es un procedimiento robusto?. Peter J. Huber (1972), plantea esta pregunta y comenta al respecto lo siguiente: Al principio "robustez" era un concepto vago, por ejemplo, Box y Anderson (1955) introdujeron la noción diciendo que "se requieren procedimientos que sean robustos" (insensibles a cambios en factores extraños no contemplados en la prueba) y poderosos (sencibles a factores especificos bajo la prueba)

Pero, si se desea escoger en forma racional entre diferentes competidores robustos para un procedimiento, se deben precisar los objetivos que se quieren alcanzar. Desafortunadamente, no se ha llegado a un consenso: aunque los objetivos raramente se establecen en forma explícita, se pueden discutir cinco o seis situaciones, de las cuales pienso que no todas deberían de llevar el nombre de robustas.

Para fijar ideas, consideremos el problema de estimar un parámetro de posición para un número grande de observaciones independientes $x_1, x_2, \ldots x_n$, distribuidas de acuerdo con $P(x_i < x) = F(\frac{x-\theta}{\sigma})$ donde la forma de F no se conoce exactamente. Uno juzgará a los estimadores en términos de

(a) varianza asintótica $\sigma_F^2(T)$.
(b) eficiencia absoluta $1/I(F)\sigma_F^2(T)$; con $I(F)$ la información de Fisher.
(c) Eficiencia relativa $\sigma_F^2(T')/\sigma_F^2(T)$.

De acuerdo al primer objetivo, un estimador robusto debe de poseer

(i) Una eficiencia absoluta alta para todas las formas suaves de F que se deseen. Mientras que este objetivo puede alcanzarse asintoticamente para muestras grandes, la convergencia parece ser muy lenta para propósitos prácticos. Entonces, modificamos los requerimientos a uno de los siguientes.

(ii) Una eficiencia relativa alta para la media muestral para toda F.

(iii) Una eficiencia absoluta alta sobre un conjunto $\{F_i\}$ de formas seleccionadas estrategicamente (p.e. normal, doble exponencial, cauchy, etc).
Una variante de (iii) es
(ii') Una eficiencia absoluta alta sobre una familia paramétrica seleccionadas estrategicamente.

(iv) Una varianza asintótica pequeña sobre una vecindad

de una familia, en particular la de la normal.

Ninguno de estos objetivos garantiza la estabilidad cualitativa requerida (la convergencia en (i) y en (iv) no necesariamente es uniforme):

(v) La distribución del estimador cambiará poco bajo variaciones arbitrarias pequeñas de la distribución principal F, y esto uniformemente con el tamaño muestral n.

Huber considera a los objetivos locales (iv) y (v) como los más importantes y comenta que al igual que Anscombe (1960), se inclina a ver robustes como un problema de Seguros: "Estoy dispuesto a pagar una prima (una pérdida de entre el 5% y 10% del modelo ideal) para protegerme contra malos efectos causados por pequeñas desviaciones de él; aunque estoy feliz si el procedimiento funciona bien bajo grandes desviaciones."

Primeramente se estara interesado con robustes distribucional: la forma de la verdadera distribución principal se desvia ligeramente del modelo supuesto (usualmente el modelo normal). Este es el caso más importante y más difundido.

Hay otros aspectos de robustes acerca de los cuales se conoce poco, por ejemplo, insensibilidad respecto a desviaciones de independencia, distribución idéntica etc. mucho menos es conocido de que pasa cuando las otras suposiciones estándar de estadística no se satisfacen y acerca de la protección en estos otros casos.

Debe enfatizarse una vez más que, la ocurrencia de errores gruesos en una fracción pequeña de observaciones es con

siderada como una desviación pequeña, y que, en vista de la sensibilidad extrema de algunos problemas clásicos un objetivo fundamental de los procedimientos robustos es protegerse contra errores gruesos.

4.2. Estimadores robustos en regresión

4.2.1. Estimadores M. Reciben este nombre todos los estimadores que minimizan una función ρ de los residuales

$$\min_{\underline{\beta}} \sum_{i=1}^{n} \rho(e_i) = \min_{\underline{\beta}} \sum_{i=1}^{n} \rho(y_i - \underline{x}_i^t \underline{\beta}) \qquad (4.2.1)$$

donde $\underline{x}_i^t$ denota el i-ésimo renglón de X. Un estimador de este tipo recibe el nombre de estimador M, debido a que puede pensarse como un estimador semejante al obtenido por máxima verosimilitud, esto es, la función ρ, está referida a la función de verosimilitud para una elección apropiada de la distribución del error, por ejemplo, si se usa el método de mínimos cuadrados (se supone que la distribución del error es normal), entonces $\rho(e) = \frac{1}{2} e^2$; $-\infty < e < \infty$.

Uno de los inconvenientes de los estimadores M, es que no siempre son invariantes bajo reescalamiento, es decir, si los residuales $y_i - \underline{x}_i^t \hat{\underline{\beta}}$ se multiplican por una constante, la solución a (4.2.1) no será la misma que se obtuvo anteriormente. Para obtener una versión invariante bajo escala de este estimador, usualmente se resuelve

$$\min_{\underline{\beta}} \sum_{i=1}^{n} \rho\left(\frac{e_i}{d}\right) = \min \sum_{i=1}^{n} \rho[(y_i - \underline{x}_i^t \underline{\beta})/d] \qquad (4.2.2)$$

por los métodos: a) H ; b) Newton c) Mínimos cuadrados ponde

rados ver Hogg (1979 b).

En la expresión (4.2.2) d es un parámetro robusto de escala cuya elección mas común es

$$d = \frac{\text{mediana}\,|e_i - \text{mediana}\,(e_i)|}{0.6745}$$

La constante 0.6745 hace a d un estimador casi insesgado de σ cuando n es grande y la distribución del error es normal.

Para minimizar (4.2.2), encuentre las primeras derivadas parciales de ρ con respecto a β_j $(j = 1, 2, \ldots p)$ (suponiendo que ρ es una función derivable y convexa) e igualando a cero; ésto da un sistema de p ecuaciones

$$\sum_{i=1}^{n} x_{ij}\, \psi[(y_i - \underline{x}_i^t \underline{\beta})/d] = 0 \quad ; \quad j = 1, 2, \ldots p \qquad (4.2.3)$$

donde $\psi = \rho'$ y x_{ij} es la i-ésima observación de la variable j. En general la función ψ no es una función lineal y (4.2.3) debe de resolverse por algún método iterativo mencionado anteriormente. Por ejemplo, si usamos el método de mínimos cuadrados ponderados y además contamos con un estimador inicial de β y σ; $\hat{\beta}_0$, do entonces, (4.2.3) puede reescribirse como

$$\sum_{i=1}^{n} x_{ij}\, \psi[(y_i - \underline{x}_i^t \underline{\beta})/d] = \sum_{i=1}^{n} x_{ij}\, \frac{\psi[(y_i - \underline{x}_i^t \underline{\beta})/d]}{(y_i - \underline{x}_i^t \underline{\beta})/d} \cdot (y_i - x_i^t \beta)/d = 0$$

$$= \sum_{i=1}^{n} x_{ij}\, w_{i_0} (y_i - x_i^t \beta) = 0 \; ; \; j = 1, 2, \ldots p \qquad (4.2.4)$$

donde $w_{io} = \begin{cases} \dfrac{\psi[(y_i - x_i^t\beta_o)/d]}{(y_i - x_i^t\beta_o)/d} & \text{si } y_i \neq x_i^t\beta_o^t \\ 1 & \text{si } y_i = x_i^t\hat{\beta}_o \end{cases}$ (4.2.5)

La expresión (4.2.4) se puede escribir en notación matricial como sigue:

$$X^t W_o X\beta = X^t W_o Y \,, \qquad (4.2.6)$$

donde W_o es la matriz diagonal de pesos, cuyos elementos son w_{1o}, w_{2o}, w_{no}, dados en (4.2.5). La expresión (4.2.6), son las ecuaciones usuales de mínimos cuadrados ponderados Consecuentemente el estimador en la primera iteración es

$$\hat{\beta}_1 = (X^t W_o X)^{-1} X^t W_o Y \qquad (4.2.7)$$

En la siguiente iteración, se vuelven a calcular los pesos en (4.2.5) usando $\hat{\beta}_1$ en lugar de $\hat{\beta}_o$. Para alcanzar la convergencia del método, regularmente se requieren pocas iteraciones, aunque claro, éste depende de que tan buenos son los estimadores iniciales de β y σ. Para obtener estos estimadores, algunos autores sugieren usar los estimadores de norma L_1 (más adelante se hace referencia a ellos), mientras que Hogg (1979.b), sugiere el uso del algoritmo desarrollado por Dutter (1977) que proporciona simultaneamente buenos estimadores de β y σ.

En la tabla (4.2.1) se muestran las funciones ρ mas usuales y sus correspondientes funciones ψ. La función ψ controla los pesos que se asignan a cada residual y es por ello que algu-

tabla 4.2.1.

Estimador	$\rho(e)$	$\psi(e)$	$W(e)$	RANGO
MINIMOS CUADRADOS	$\frac{1}{2}e^2$	e	1.0	$\lvert e\rvert < \infty$
HUBER *t = 2	$\frac{1}{2}e^2$	e	1.0	$\lvert e\rvert \le t$
	$\lvert e\rvert t - \frac{1}{2}t^2$	t signo(e)	$t/\lvert e\rvert$	$\lvert e\rvert > t$
ANDREWS *a = 1.339	$a[1-\cos(e/a)]$	$\operatorname{sen}(e/a)$	$\frac{\operatorname{sen}(e/a)}{e/a}$	$\lvert e\rvert \le a\pi$
	$2a$	0	0	$\lvert e\rvert > a\pi$
RAMSAY *a = 0.3	$a^2[1-\exp(-a\lvert e\rvert)\cdot(1+a\lvert e\rvert)]$	$e\ \exp(-a\lvert e\rvert)$	$\exp(-a\lvert e\rvert)$	$\lvert e\rvert < \infty$
HAMPEL * a = 1.7 b = 3.4 c = 8.6	$\frac{1}{2}e^2$	e	1.0	$\lvert e\rvert \le a$
	$a\lvert e\rvert - \frac{1}{2}a^2$	a signo(e)	$a/\lvert e\rvert$	$a < \lvert e\rvert \le b$
	$\frac{a(c\lvert e\rvert - \frac{1}{2}e^2)}{c-b} - (\frac{7}{6})a^2$	$\frac{a\ \text{signo}(e)(c-\lvert e\rvert)}{c-b}$	$\frac{a(c-\lvert e\rvert)}{\lvert e\rvert(c-b)}$	$b < \lvert e\rvert \le c$
	$a(b+c-a)$	0	0	$\lvert e\rvert > c$
TUKEY	$\frac{e^2}{2}[1-(\frac{e}{k})^2+\frac{1}{3}(\frac{e}{k})^4]$	$e[1-(e/k)^2]^2$	$[1-(e/k)^2]^2$	$\lvert e\rvert \le k$
	$\frac{k^2}{2}$	0	0	$\lvert e\rvert > k$
	donde $e = \frac{y_i - x_i^T \cdot \beta}{s}$ y			
	t, a, b, c pueden tomar otros valores, esto dependerá de la robustez deseada.			

nas veces se les llama funciones de influencia, por ejemplo, la función mínimos cuadrados no está acotada y entonces, el estimador resultante, no es robusto si los datos surgen de una distribución con colas pesadas.

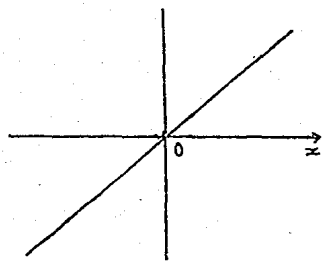
La función de Huber, es idéntica a la de mínimos cuadrados, en el intervalo $(-t, t)$, ver (b) en figura (4.2.1), pero fuera de este la función de influencia toma valores $\psi(x) = t \text{ signo } (x)$, dando así un peso menor a los residuales en la medida en que estos crecen.

La función de Hampel, (e) de (4.2.1), puede pensarse como una versión discreta de la función de Andrews o la de Tukey. También puede observarse que estas últimas funciones son muy similares y por lo tanto una sería sustituto de la otra.

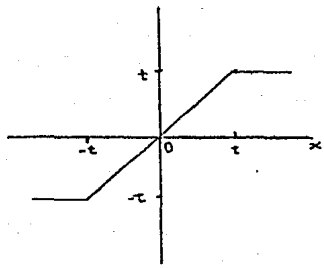
Ahora, si una observación es discrepante, el residual correspondiente es grande. Si se utiliza como estimador el propuesto por Andrews, Tukey o Hampel, entonces estas observaciones tendrán poco peso en el ajuste. Pero si se utiliza el estimador de Huber, el residual tendría que ser muy grande para que el peso $\frac{t}{x}$, fuera pequeño y esta observación no afectará al ajuste.

Debido a que las funciones ρ asociadas con las funciones de influencia no son convexas, pueden surgir ciertos problemas de convergencia en el proceso iterativo, a pesar de esto, se cuenta con buenos procedimientos que pueden usarse con cierto cuidado. En caso de que se utilice el método de mínimos cuadrados, no hay manera de protegerse de las observaciones -

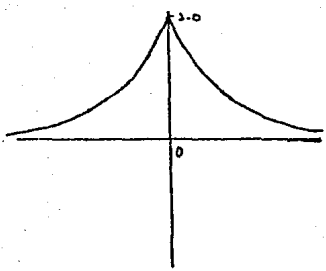
figura 4.2.1.



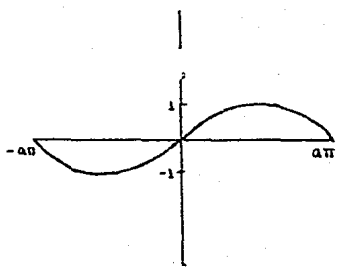
(a) MINIMOS CUADRADOS.



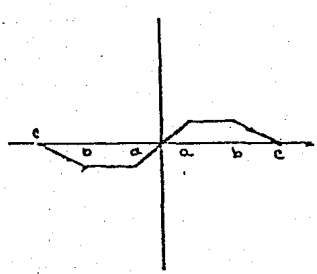
(b) HUBER



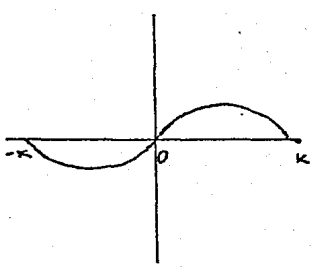
(c) RAMSAY



(d) ANDREWS



(e) HAMPEL



(f) TUKEY

discrepantes.

Entonces, para que el estimador $\hat{\beta}$ del parámetro no se vea influenciado por algunas observaciones lejanas al plano de regresión, la función $\psi(x)$ debe de ser acotada y tender a una constante es decir

$$\lim_{|x| \to \infty} \psi(x) = 0$$

Como resultado de esta condición se pueden tener varios mínimos locales, por lo que la solución dada por el método iterativo de optimización dependerá del punto $\hat{\beta}_0$ que se usa para iniciar el proceso.

4.2.2 Estimadores R. Los métodos conocidos como estimación R, son procedimientos que se basan en los rangos. El procedimiento general reemplaza un factor de cada término en la función objetivo de mínimos cuadrados.

$S_0(\underline{\beta}) = \sum_{i=1}^{n} (y_i - \underline{x}_i^t \underline{\beta})^2$ por su rango correspondiente, así, si R_i es el rango de $y_i - \underline{x}_i^t \underline{\beta}$, se desea minimizar

$$\sum_{i=1}^{n} (y_i - \underline{x}_i^t \underline{\beta}) \cdot R_i$$

Más en general, se reemplazan los rangos por una función de puntaje $a_n(i)$, $i = 1, 2, \ldots n$, entonces la función objetivo que se debe de minimizar es

$$S_1(\underline{y} - X\underline{\beta}) = \sum_{i=1}^{n} (y_i - \underline{x}_i^t \underline{\beta}) \cdot a_n(R_i) \qquad (4.2.8)$$

Como S_1 es una función no negativa de $\underline{\beta}$, continua y convexa (Jaeckel 1972) y por estar definida para toda $\underline{\beta}$, además de ser derivable en casi todas partes, el número puede hallarse por algún método iterativo, como el de "stupest descent" de Luenberger (1973).

Para puntajes $a_n(i)$ tales que

$$\sum a_n(i) = 0 ,$$

S_1 es invariante en posición, ya que: $S_1(\underline{y} - X\underline{\beta} + c) = S_1(\underline{y} - X\underline{\beta})$ y por lo tanto la ordenada al origen, para la ecuación de regresión, no podrá ser estimada por medio de (4.2.8). Para superar este problema, se hace lo siguiente: con los coeficientes que se puedan obtener de (4.2.8), calcular los residuales

$e_i = (y_i - \underset{\sim}{x}_i^{+} \underset{\sim}{\beta})$ y entonces minimizar $\sum_{i=1}^{n} a_n^{+}(R_i^{+})$ signo $(e_i - \alpha)$ con respecto a α; donde R_i^{+} es el rango de $|e_i - \alpha|$, a_n^{+} es una función de puntaje con signo.

Algunas de las funciones de puntaje mayormente utilizadas:

Wilcoxon: $a_n(i) = i \qquad 1 \leq i \leq n$

Mediana: $a_n(i) = \begin{cases} 1 & \text{si } i > (n+1)/2 \\ -1 & \text{si } i \leq (n+1)/2 \end{cases}$;

Normal: $a_n(i) = \Phi^{-1}\left(\frac{i - 1/2}{n}\right)$; $1 \leq i \leq n$.

Vander Warden: $a_n(i) = \Phi^{-1}\left(\frac{i}{n+1}\right) \qquad 1 \leq i \leq n$

con Φ la función de distribución normal $(0, 1)$

4.3 Comentarios acerca de los estimadores Robustos.

Naturalmente, los estimadores robustos discutidos en las secciones anteriores no son los únicos, se cuenta también con los estimadores L (combinaciones lineales de estadísticas de orden), de norma Lp (minimizar la suma de las desviaciones absolutas del hiperplano de regresión elevados a la potencia ρ), método de Andrews, etc; entre otros, ver Gracia - Medrano - (1989).

En este trabajo, se ha hecho énfasis en los estimadores M debido principalmente a que estos pueden hacer frente a casi cualquier tipo de distribución (fuerte contaminación de la distribución normal, distribuciones con colas pesadas, etc.). Además de que

éstos pueden obtenerse facilmente de un programa estandar de M.C. ponderados y los pesos en la iteración final identifica puntos discrepantes. Los estimadores L no se prestan para una generalización de regresión múltiple. Además bajo ciertas condiciones, los estimadores R son asintoticamente equivalentes a los estimadores M.

Uno de los inconvenientes de los estimadores M, es la estimación de la matríz covarianza y en general para hacer inferencias.

Algunos factores para poder elegir algún estimador pueden encontrarse en Gracia - Medrano V (1989) y son

a) Conocimiento acerca de la distribución de la variable respuesta.

b) tamaño de muestra.

c) Recursos con los que se cuenta para el cómputo.

d) La situación, ya sea de observación o experimentación.

e) Número de variables explicativas.

En general, la mayoría de los estimadores robustos requieren de programas más elaborados para su obtención que los de M.C., así como mayor tiempo de cómputo. En el caso de los estimadores M que se obtienen por métodos iterativos, el estimador inicial debe elegirse con mucho cuidado ya que juega un papel importante. Los factores principales para decidir que tipo de estimador inicial usar, son el tamaño

de muestra y recursos con que se cuenta para el cómputo.

Los métodos robustos de regresión se ven fuertemente afectados por el mal condicionamiento de la matriz X (multicolinealidad). Si el modelo lineal no es el adecuado, usar un método robusto no resolverá este problema. En algunos casos será necesario transformar las variables y despues aplicar un método robusto.

El uso de los métodos robustos no se ha generalizado debido a varias razones.

a) Su cómputo es mas complicado que el de mínimos cuadrados.

b) Los criterios que se siguen para la obtención de estimadores robustos en la mayoría de los casos son más difíciles de interpretar en comparación con el criterio de minimizar una suma de residuales al cuadrado.

c) Los estimadores robustos no son invariantes bajo escala.

d) La obtención de resultados teóricos, distribuciones asintóticas, estimación de matrices de covarianza, pruebas asociadas, es mucho más difícil y complicada que para el caso de mínimos cuadrados.

En cuanto al cómputo de los procedimientos robustos, actualmente, gracias a los avances en la computación, se ha desarrollado una gran cantidad de rutinas para obtenerlos. Ge

neralmente éstas son más complicadas y requieren de mayor tiempo que el procedimiento de mínimos cuadrados. Desgraciadamente éstas rutinas, en su mayoría no están todavía implantadas en los paquetes usuales de regresión y esto hace difícil su acceso para la mayoría de investigadores interesados en aplicar métodos robustos.

Los métodos robustos de regresión tienen mucho que ofrecer, éstos son extremadamente útiles en la localización de obvervaciones discrepantes e influyentes. Así, cuando se efectue un análisis mínimos cuadrados sería útil ejecutar también un ajuste robusto. Si los resultados de los dos procedimientos substancialmente concuerdan, entonces usar los obtenidos por mínimos cuadrados debido principalmente a que en la actualidad son los más comprendidos. Sin embargo, si los resultados de los dos análisis difieren, entonces, deben de identificarse la(s) razon(es) de esta(s) diferencia(s). Las observaciones que están ponderadas en el ajuste robusto, deben examinarse cuidadosamente. Algunas de las causas por las cuales los dos procedimientos no concuerdan pueden ser:

a) Errores gruesos, ya sea en la variable respuesta, ó en las explicativas.

b) Un modelo lineal inadecuado.

c) El análisis debió hacerse en otra escala

d) El error tiene una distribución con colas más pesadas que la normal.

4.4. Ejemplos.

Por tener solamente disponibilidad de los estimadores M con la inclusión de la función de Huber, se llevo a cabo el análisis de los dos ejemplos presentados en la sección 2.5. En las tablas 4.4.1 y 4.4.2 se encuentran resumidos los resultados para diferentes valores de t, del primer y segundo ejemplo respectivamente. Para el primero de ellos, observamos que para los cuatro valores de t propuestos, la observación que conserva el menor peso es precisamente la número 19, que por los métodos anteriores había sido declarada como discrepante. Por otro lado, en la medida que t es pequeño ($t < 2$), nos damos cuenta que los estimadores (de los parámetros) robustos difieren marcadamente de los obtenidos por mínimos cuadrados, y para valores grandes de t los valores son mas parecidos. Esto es natural, ya que mientras más grande sea t, la función $\ell(e)$ de Huber se parece más a la de mínimos cuadrados. Por lo que respecta a la suma de cuadrados residual, esta entre 2 263.960 1 para $t=1$ y 2 127.625 9 para $t=4$ con lo que podemos observar que se obtiene una reducción en la suma de cuadrados residual con respecto a mínimos cuadrados a pesar de tener valores extremos para t.

En términos generales podríamos decir que los resultados obtenidos por mínimos cuadrados y métodos robustos no difieren cualitativamente ni cuantitativamente para este ejemplo, y siguiendo nuestras recomendaciones podríamos manejar con confianza los resultados obtenidos por mínimos cuadrados.

Un análisis similar en el grupo de datos nos lleva -

tabla 4.4.1

ESTIMACION ROBUSTA

EL VALOR DE t ES 1.000000

PARAMETROS ESTIMADOS

RETA

110.065
-1.209

EL VALOR DE t ES 1.200000

PARAMETROS ESTIMADOS

RETA

110.022
-1.181

VECTOR Y	VECTOR Y ESTIMADO	RESIDUALES	PESOS	DATO	VECTOR Y ESTIMADO	RESIDUALES	PESOS
95.000000	97.445686	-2.445686	0.408883	1	97.700198	2.700198	0.444412
71.000000	62.741400	8.258600	0.121656	2	62.696301	-8.303699	0.144514
83.000000	67.402179	15.597821	0.064112	3	67.792878	-15.207122	0.078910
91.000000	82.193478	8.806522	0.113552	4	82.611414	-8.388586	0.143052
102.000000	111.445686	9.445686	0.105868	5	111.700198	9.700198	0.123709
87.000000	87.489193	0.489193	1.000000	6	87.607518	0.607518	1.000000
93.000000	97.071790	4.071790	0.245592	7	97.244590	4.244590	0.282713
100.000000	102.610881	2.610881	0.383013	8	102.974342	2.974342	0.403451
104.000000	106.984777	2.984777	0.335033	9	107.429550	3.429550	0.349859
94.000000	101.489193	7.489193	0.133526	10	101.607518	7.607518	0.157739
113.000000	123.776076	10.776076	0.092798	11	124.246486	11.246486	0.106681
96.000000	92.193478	-3.806522	0.262707	12	92.611414	-3.388586	0.354430
83.000000	67.402179	-15.597821	0.064112	13	67.792878	-15.207122	0.078910
84.000000	70.610881	-13.389119	0.074668	14	70.974342	-13.025658	0.092126
102.000000	106.610881	4.610881	0.216878	15	106.974342	4.974342	0.241238
100.000000	101.402179	1.402179	0.713175	16	101.792878	1.792878	0.669315
105.000000	113.819582	8.819582	0.113394	17	114.155806	9.155806	0.131064
57.000000	54.080621	-2.919379	0.342559	18	53.599724	-3.400276	0.552913
121.000000	151.863089	30.863089	0.032401*	19	152.063126	31.063126	*0.038631
86.000000	74.610881	-11.389119	0.087803	20	74.974342	-11.025658	0.108837
100.000000	101.402179	1.402179	0.713175	21	101.792878	1.792878	0.669315

tabla 4.4.1

ESTIMACION PUNTUAL

EL VALOR DE t ES 1.000000

PARAMETROS ESTIMADOS

BETA

110.665
-1.209

VECTOR Y	VECTOR Y ESTIMADO	RESIDUALES	PESOS
95.000000	97.445686	-2.445686	0.408883
77.000000	82.741400	-5.741400	0.174161
53.000000	67.402179	-15.597821	0.064112
91.000000	82.193478	8.806522	0.113552
102.000000	111.445686	-9.445686	0.105868
97.000000	97.488906	-0.488906	1.000000
93.000000	97.071790	-4.071790	0.245592
100.000000	102.610881	-2.610881	0.383013
104.000000	106.987777	-2.987777	0.334553
94.000000	101.489193	-7.489193	0.133529
113.000000	103.772677	10.227323	0.097798
96.000000	92.193478	3.806522	0.262707
63.000000	67.402179	-15.597821	0.064112
84.000000	70.610881	13.389119	0.074688
102.000000	106.610881	-4.610881	0.216978
102.000000	101.402179	1.402179	0.713175
105.000000	113.819582	-8.819582	0.113384
57.000000	67.090221	-10.090221	0.099104
121.000000	91.896308	29.103692	0.034401
86.000000	74.610881	11.389119	0.087803
100.000000	101.402179	-1.402179	0.713175

EL VALOR DE t ES 1.200000

PARAMETROS ESTIMADOS

BETA

110.022
-1.181

DATO	VECTOR Y ESTIMADO	RESIDUALES	PESOS
1	97.700198	-2.700198	0.444612
2	82.665301	-5.665301	0.176510
3	67.792878	-15.207122	0.078910
4	92.611416	-8.388584	0.143052
5	111.700198	-9.700198	0.123709
6	97.607519	-0.607519	1.000000
7	97.244342	-4.244342	0.235591
8	102.974342	-2.974342	0.403451
9	107.429551	-2.429551	0.349159
10	101.607519	-7.607519	0.131479
11	124.248486	11.248486	0.106681
12	92.611416	-3.388584	0.354130
13	67.792878	-15.207122	0.078910
14	70.974342	-13.025658	0.092126
15	106.974342	4.974342	0.241238
16	101.792878	1.792878	0.669315
17	114.155806	9.155806	0.131064
18	93.599724	-3.400276	0.339915
19	152.063126	31.063126	0.038631
20	74.974342	-11.025658	0.101837
21	101.792878	1.792878	0.669315

tabla 4.4.1.(Continuación)

ESTIMADOR HUBER

EL VALOR DE t ES 1.500000	EL VALOR DE t ES 2.500000
PARAMETROS ESTIMADOS	PARAMETROS ESTIMADOS
BETA	BETA
109.598	109.802
-1.159	-1.135

VECTOR Y	VECTOR Y ESTIMADO	RESIDUALES	PESOS	VECTOR Y ESTIMADO	RESIDUALES	PESOS	DATO
95.000000	97.790064	2.790064	0.537622	97.228728	2.228728	1.000000	1
71.000000	62.541019	-8.458981	0.177326	61.717842	-9.282158	0.269334	2
83.000000	67.994175	-15.005825	0.099961	67.551859	-15.448141	0.161832	3
91.000000	82.834998	-8.165002	0.183711	82.416485	-8.583515	0.291256	4
102.000000	111.790064	9.790064	0.153217	111.228728	9.228728	0.270693	5
87.000000	87.585953	0.585953	1.000000	86.905598	-0.094402	1.000000	6
93.000000	97.267597	4.267597	0.351480	96.634850	3.634850	0.687786	7
100.000000	103.153353	3.153353	0.475684	102.687232	2.687232	0.930325	8
104.000000	107.675820	3.675820	0.408072	107.281111	3.281111	0.761937	9
94.000000	101.585953	7.585953	0.197734	100.905598	6.905598	0.362025	10
113.000000	124.516642	11.516642	0.130246	124.145737	11.145737	0.224301	11
96.000000	92.834998	-3.165002	0.473933	92.416485	-3.583515	0.697639	12
83.000000	67.994175	-15.005825	0.099961	67.551859	-15.448141	0.161832	13
84.000000	71.153353	-12.846647	0.116762	70.687232	-13.312768	0.187790	14
102.000000	107.153353	5.153353	0.291073	106.687232	4.687232	0.533364	15
100.000000	101.994175	1.994175	0.752191	101.551859	1.551859	1.000000	16
105.000000	114.312531	9.312531	0.161073	113.822606	8.822606	0.283363	17
57.000000	53.087863	-3.912137	0.383422	51.883826	-5.116174	0.488646	18
121.000000	152.108419	31.108419	*0.048215	151.499476	30.499476	*0.081969	19
86.000000	75.153353	-10.846647	0.138292	74.687232	-11.312768	0.220989	20
100.000000	101.994175	1.994175	0.752191	101.551859	1.551859	1.000000	21

también a obtener las conclusiones de Brownlee, es decir, declarar a las observaciones 1, 3, 4, 21 como disorepantes. Como en este caso si existen ciertas diferencias en los estimadores finales con respecto a los obtenidos por mínimos cuadrados, podríamos concluir que se requiere un análisis adicional, éste se hará en el capítulo 5.

tabla 4.4.2

ESTIMADOR HUBER

EL VALOR DE C ES 1.000000

PARAMETROS ESTIMADOS

BETA

-40.677
0.831
0.875
-0.121

EL VALOR DE C ES 1.500000

PARAMETROS ESTIMADOS

BETA

-41.107
0.[illegible]01
1.041
-0.135

C = 1.000000:

VECTOR Y	VECTOR Y ESTIMADO	RESIDUALES	PESOS	DATO
42.000000	45.381861	3.381861	*0.295695	1
37.000000	35.260865	-1.739135	0.574998	2
37.000000	41.405195	4.405195	*0.227005	3
28.000000	34.714129	6.714129	+0.148940	4
18.000000	16.463733	-1.536267	0.650928	5
18.000000	15.588931	-2.411069	0.414754	6
19.000000	17.440106	-1.559894	0.641069	7
20.000000	19.440106	-0.559894	1.000000	8
15.000000	12.911121	-2.088879	0.478726	9
14.000000	14.438157	0.438157	1.000000	10
14.000000	15.527121	1.527121	0.654827	11
13.000000	14.280927	1.280927	0.780685	12
11.000000	8.680149	-2.319851	0.431062	13
12.000000	11.136304	-0.863696	1.000000	14
8.000000	10.171552	2.171552	0.460511	15
7.000000	7.[illegible]	0.[illegible]	1.000000	16
8.000000	7.239767	-0.760233	1.000000	17
8.000000	8.085740	0.085740	1.000000	18
9.000000	9.332534	0.332534	1.000000	19
15.000000	16.591641	1.591641	0.628284	20
15.000000	6.052980	-8.947021	*0.111767	21

C = 1.500000:

VECTOR Y	VECTOR Y ESTIMADO	RESIDUALES	PESOS	DATO
42.000000	44.905216	2.905216	*0.516312	1
37.000000	34.770492	-2.229508	0.672794	2
37.000000	41.127273	4.127273	*0.363436	3
28.000000	34.177768	6.177768	*0.242806	4
18.000000	16.259865	-1.740135	0.862002	5
18.000000	15.218816	-2.781184	0.539339	6
19.000000	16.986109	-2.013891	0.744827	7
20.000000	18.986109	-1.013891	1.000000	8
15.000000	12.422996	-2.577004	0.582074	9
14.000000	14.[illegible]	0.[illegible]	1.000000	10
14.000000	15.897694	1.897694	0.790435	11
13.000000	14.804018	1.804018	0.831477	12
11.000000	8.954630	-2.045370	0.733364	13
12.000000	11.[illegible]	-0.[illegible]	1.000000	14
8.000000	10.306174	2.306174	0.650456	15
7.000000	7.701903	0.701903	1.000000	16
8.000000	6.[illegible]	-1.[illegible]	1.000000	17
8.000000	7.[illegible]	-0.[illegible]	1.000000	18
9.000000	9.011427	0.011427	1.000000	19
15.000000	16.474829	1.474829	1.000000	20
15.000000	6.472474	-8.527526	*0.175901	21

tabla 4.9.2. (Continuación)

ESTIMADOR HUBER

EL VALOR DE c ES [illegible]

PARAMETROS ESTIMADOS

BETA

-40.313
0.732
1.246
-0.147

EL VALOR DE c ES 3.000000

PARAMETROS ESTIMADOS

BETA

-39.920
0.711
1.295
-0.152

VECTOR Y ESTIMADO	RESIDUALES	PESOS	DATO	VECTOR Y	VECTOR Y ESTIMADO	RESIDUALES	PESOS
43.112765	3.112745	*0.942519	1	42.000000	43.234637	-3.234637	*0.927461
42.265305	-4.734695	*0.633947	2	37.000000	41.055515	-1.517485	1.000000
41.419611	4.419611	*0.678526	3	37.000000	41.055515	4.555515	*0.658541
39.760566	-5.760566	*0.347188	4	28.000000	33.697774	5.697774	*0.526521
16.251610	-1.748390	1.000000	5	18.000000	16.288346	-1.711654	1.000000
15.405561	-2.405561	1.000000	6	18.000000	14.993641	3.006359	1.000000
16.645244	-2.354756	1.000000	7	19.000000	16.610509	2.389491	1.000000
18.644344	-1.355656	1.000000	8	20.000000	18.610509	1.389491	1.000000
11.940723	-1.059277	1.000000	9	15.000000	11.555621	-3.144379	0.954083
15.136277	1.136277	1.000000	10	14.000000	12.267194	1.267194	1.000000
16.463685	2.463685	0.611923	11	14.000000	16.636427	2.636427	1.000000
15.501366	-2.498634	1.000000	12	13.000000	10.779460	2.779460	1.000000
9.431655	-1.568345	1.000000	13	11.000000	9.571439	-1.428561	1.000000
11.807550	-1.192450	1.000000	14	12.000000	11.949501	-0.050499	1.000000
11.552587	-0.447413	0.397561	15	8.000000	10.361413	2.361413	1.000000
7.990247	-0.009753	1.000000	16	7.000000	7.905051	0.905051	1.000000
6.590556	-1.411442	1.000000	17	6.000000	6.480049	-1.519951	1.000000
7.611670	-0.388330	1.000000	18	6.000000	7.544507	-0.455493	1.000000
8.514627	-1.485373	1.000000	19	5.000000	5.401743	-0.598257	1.000000
16.407471	-1.592529	1.000000	20	15.000000	16.412147	1.412147	1.000000
7.463040	-7.536960	*0.269259	21	15.000000	7.762787	-7.237713	*0.414496

Capitulo 5.

TRANSFORMACIONES.

5.1. Introducción

Como hemos observado, las observaciones discrepantes son casos para los cuales, y por diversas razones, el modelo hipotético no funciona bien, es decir, el modelo no ajusta adecuadamente dichas observaciones. Los candidatos para considerarse como observaciones discrepantes, son casos cuyos residuales estudentizados resultan extremos en comparación con los restantes, el valor de v_{ii} o de la distancia D_i de Cook. Si D_i de Cook. Si D_i es pequeña, entonces la observación discrepante (si existe) tiene poco efecto en los estimadores de los parámetros y por lo tanto no es importante excluir el caso del análisis, a pesar de que realmente sea una observación discrepante. Por otro lado, los casos con valores v_{ii} grandes requieren de atención especial, aún cuando los correspondientes r_i's ya que si D_i es grande, el caso puede tener una influencia excesiva en los estimadores de los parámetros.

Un paso generalmente útil, es efectuar nuevamente la regresión sin el (los) caso(s) sospechoso(s) y evaluar los cambios en el análisis. Este es un enfoque razonable, especialmente si la forma del modelo (por ejemplo la escala de todas las x's es conocida. Entonces, casos con valores grandes de v_{ii}, D_i o r_i pueden ser indicadores de la necesidad de una transformación y no de una observación discrepante

Una vez detectadas las observaciones discrepantes, se podrían efectuar cualesquiera de las siguientes cosas, Atkinson (1982):

a) Efectuar nuevamente la regresión sin los casos discrepantes y evaluar los cambios registrados.

b) Considerar transformaciones que hagan a los casos influyentes menos importantes o mejoren el ajuste de casos que se consideran en principio discrepantes.

Los métodos para escoger transformaciones requieren de un conocimiento específico acerca de las relaciones entre las variables y hacen uso de técnicas de diagnóstico para sugerir transformaciones posibles, éstos generalmente escogen una transformación para maximizar alguna función de interés.

5.2. Métodos de transformación de Variables.

5.2.1. Método de máxima verosimilitud. Una clase importante de transformaciones puede obtenerse cuando $\underline{Y}$ toma únicamente valores positivos. Si solo consideramos transformaciones de $\underline{Y}$, esto es equivalente a ajustar el modelo

$$\underline{Y}^{\lambda} = X\underline{\beta} + \underline{\varepsilon}\ , \quad \mathrm{Var}(\underline{\varepsilon}) = \sigma^2 I, \qquad (5.2.1.)$$

donde λ es la potencia de la transformación.

Cuando λ es desconocido, puede verse a (5.2.1) como un modelo con un parámetro desconocido adicional a $\underline{\beta}$ y σ^2. Entonces se puede estimar λ de la misma manera que $\underline{\beta}$ y σ^2.

Box y Cox (1964) sugieren, para estimar λ, encontrar el estimador conjunto $(\hat{\underline{\beta}}, \hat{\sigma}^2, \hat{\lambda})$ máximo verosimil efectuando suposiciones acerca de $\underline{\varepsilon}$; en este trabajo se supone que ε tiene distribución normal.

Por un momento, supongamos que λ es conocida, entonces la estimación de $\hat{\beta}$ y el cálculo de suma de cuadrados del residual es inmediata

$$\hat{\underline{\beta}}_\lambda = (x^t x)^{-1} x^t \underline{y}^\lambda \qquad (5.2.2)$$

$$SRC_\lambda = (\underline{y}^\lambda)^t (I - V)\underline{y}^\lambda \qquad (5.2.3)$$

que podrá obtenerse fácilmente con un programa de regresión lineal simple.

Ya que λ es desconocido, (5.2.3) puede calcularse para un rango de valores de λ. Para cada $\lambda \neq 0$ calcular

$$L(\lambda) = \frac{n}{2} \ln(\lambda^2) - \frac{n}{2} \ln(SCR_\lambda) + (\lambda - 1)\sum_{i=1}^{n} \ln(y_i) \qquad (5.2.4)$$

y si $\lambda = 0$

$$L(\lambda) = -\frac{n}{2} \ln(SCR_\lambda) - \sum_{i=1}^{n} \ln y_i \qquad (5.2.5)$$

Se puede demostrar que la λ que maximiza (5.2.4) y (5.2.5) es el estimador máximo verosímil (suponiendo normalidad de los errores). Debe notarse que no tiene sentido escoger la λ que maximiza a SCR_λ porque para cada λ la SCR_λ es medida en una escala diferente, mientras que las ecuaciones (5.2.4) y (5.2.5) convierte cada SCR_λ

a una escala común.

5.2.2. Método de Carroll

En analogía con los estimadores M y en particular con lo desarrollado por Huber, R.J. Carroll (1980), propone una prueba para obtener una transformación en y. Así, usando

$$\rho(x) = \begin{cases} \frac{1}{2}x^2, & |x| \le K, \\ K(|x| - K/2), & |x| > K, \end{cases}$$

La verosimilitud resulta ser

$$L(\beta, \sigma, \lambda) = \sigma^{-n} \sum_{i=1}^{n} \exp\left\{ -\rho\left(\frac{y_i^{(\lambda)} - x_i^T \beta}{\sigma} \right) + (\lambda - 1)\log y_i \right\} \quad (5.2.6)$$

Para λ fija, tomar un valor inicial de σ y maximizar (5.2.6) con respecto a β tal que

$$\sum_{i=1}^{n} \psi\{(y_i^{\lambda} - x_i^T\beta)/\sigma\} x_i = 0 \quad (5.2.7)$$

entonces se actualiza σ, el procedimiento iterativo

$$(n-p)^{-1} \sum_{i=1}^{n} \psi^2\{(y_i^{(\lambda)} - x_i\beta)/\sigma\} = E_{\Phi}\psi^2(z),$$

donde Φ es la función de distribución normal estándar.

Para una λ dada, denotamos a los estimadores como $\hat{\beta}_R(\lambda)$, $\hat{\sigma}(\lambda)$. Para encontrar λ, entonces maximizamos $L(\hat{\beta}_R(\lambda), \hat{\sigma}(\lambda),$

λ), la solución la denotaremos por λ_R. Note que cuando $k = \infty$, $e(x) = \frac{1}{2}x^2$ y el resultado obtenido se reduce al método de Box-Cox.

Para probar la hipótesis $H_0 : \lambda = \lambda_0$, la estadística que se usará bajo el modelo (5.2.6) es

$$\Lambda_R = -2 \log \left(\frac{L(\hat{\beta}(\lambda_0), \hat{\delta}(\lambda_0), \lambda_0)}{L(\hat{\beta}(\lambda_R), \hat{\delta}(\lambda_R), \lambda_R)} \right),$$

que, bajo condiciones apropiadas, es asintoticamente χ^2 con un grado de libertad.

5.2.3. Método de Box y Tidwell.

Supongamos que tenemos p variables independientes $x_1, x_2, \ldots x_p$, y se quiere transformar cada una por una transformación de potencia para obtener $w_1, w_2, \ldots, w_p$, donde para $j = 1, 2, \ldots, p$

$$w_j = \begin{cases} x_j^{\alpha_j} & \text{si } \alpha_j \neq 0 \\ \ln(x_j) & \text{si } \alpha_j = 0 \end{cases} \qquad (5.2.8)$$

Entonces ajustaremos el modelo lineal

$$\underline{Y} = \beta_0 + \sum \beta_j w_j + \underline{\varepsilon} \qquad (5.2.9)$$

En (5.2.8) y (5.2.9) no se requiere que las α_j sean iguales. Ya que para toda $\underline{\alpha} = (\alpha_1, \alpha_2, \ldots \alpha_p)$, la variable dependiente en (5.2.9) es la misma y podemos encontrar una $\hat{\alpha}$ y una $\hat{\beta}$ que sean estimadores mínimos cuadrados (no lineales).

Esto sugiere que para cada α podemos ajustar el modelo (5.2.9); calcular la suma de cuadrados del residual — SCR_α y encontrar $\hat{\alpha}$ tal que minimice SCR_α.

Un procedimiento iterativo para encontrar $\hat{\alpha}$ los proporcionan Box y Tidwell (1962). Ellos inician con un valor arbitrario de α que se cree es el más adecuado, usualmente — $\hat{\alpha}_1 = \hat{\alpha}_2 = \ldots = \hat{\alpha}_p = 1$. Entonces, ajustando un modelo lineal aumentado, obtiene un estimador mejorado de α, repitiendo este proceso hasta que se alcanza un nivel deseado de convergencia.

Especificando, iniciar con $\hat{\alpha}_1 = \ldots = \hat{\alpha}_p = 1$, así $w_j = x_j^{\hat{\alpha}_j} = x_j$ y ajustar el modelo de regresión (5.2.9) para obtener

$$\underline{Y} = \hat{\beta}_0 + \sum \hat{\beta}_j w_j \qquad (5.2.10)$$

Despues, construir p variables nuevas $z_1, z_2, \ldots, z_p$ definidas por

$$z_j = w_j \ln(w_j) \qquad j = 1, 2, \ldots p \qquad (5.2.11)$$

Las z_j's se definen de tal manera que, en el modelo lineal aumentado

$$\hat{Y} = \hat{\beta}_0^* + \sum \beta_j^* w_j + \sum \hat{\gamma}_j z_j \qquad (5.2.12)$$

entonces, si se requiere una transformación, cada $\hat{\gamma}_j$ sera grande (sin considerar signo) y pequeña en caso contrario. Las β_j^*'s no son iguales a las $\hat{\beta}_j$ del modelo (5.2.10). Box y Tidwell muestran que un estimador apropiado de α_j es

$$\hat{\alpha}_j = \left(\frac{\hat{\sigma}_j}{\hat{\sigma}_0} + 1\right)(\text{valor corriente de } \hat{\alpha}_j). \qquad (5.2.13)$$

El procedimiento anterior, con las $\hat{\alpha}_j$ obtenidas de (5.2.13), puede repetirse hasta hacer decrecer la SCR suficientemente. Sin embargo, frecuentemente un ciclo es adecuado.

5.2.4. Método de Andrews

Este método se basa en la prueba de hipótesis $\lambda = \lambda_0$. La estadística de prueba se construye expandiendo $\underline{Y}^{(\lambda)}$ con λ_0

$$\underline{Y}^{(\lambda_0)} \cong \underline{Y}^{(\lambda)} + (\lambda_0 - \lambda)\underline{G}_Y^{(\lambda_0)} \quad \text{donde } \underline{G}_Y^{(\lambda)} = \frac{\partial \underline{Y}^{(\lambda)}}{\partial \lambda}\Big/_{\lambda=\lambda_0}$$

Ya que $\underline{Y}^{(\lambda)} = X\underline{\beta} + \underline{\varepsilon}$; $\underline{Y}^{(\lambda_0)} \cong X\underline{\beta} + (\lambda_0 - \lambda)\underline{G}_Y^{(\lambda_0)} + \underline{\varepsilon}$.

La estadística para la prueba de Andrews es igual a la estadística t para la hipótesis $\lambda_0 - \lambda = 0$ en el modelo

$$\underline{Y}^{(\lambda_0)} = X\underline{\beta} + (\lambda_0 - \lambda)\,\hat{\underline{G}}_Y^{(\lambda_0)} + \underline{\varepsilon}$$

donde $\hat{\underline{G}}_Y^{(\lambda_0)}$ es igual a $\underline{G}_Y^{(\lambda_0)}$ evaluada en los valores ajustados del modelo nulo $\underline{Y}^{(\lambda_0)} = X\underline{\beta} + \underline{\varepsilon}$, se sigue inmediatamente, a partir del trabajo de Milliken y Graybill (1970) que la estadística t, tiene una distribución t-student con $n-p-1$ grados de libertad, es decir, la prueba de Andrews es exacta.

5.2.5. Método de Atkinson.

El enfoque que Atkinson (1982) usa es similar al usado en Box y Cox (1964). Se supone que existe una transfor-

mación paramétrica de los datos $\underline{z} = \underline{z}^{\lambda}$ tal que, para alguna λ, las observaciones satisfacen las suposiciones de segundo orden para el modelo lineal $E(\underline{z}^{\lambda}) = X\underline{\beta}$ y además que, al menos aproximadamente, las observaciones transformadas pueden tratarse como una muestra de una distribución normal. Para un análisis de verosimilitud $\hat{\lambda}$. Como una forma alternativa para determinar la evidencia para una transformación, Atkinson usa la estadística t_D para probar la hipótesis $\lambda = \lambda_0$.

$$t_D = \frac{-\underline{z}^T M \underline{w}}{s_{\underline{z}} [\underline{w}^T M \underline{w}]^{1/2}} \qquad (5.2.14)$$

donde $\underline{w} = \frac{\partial \underline{z}}{\partial \lambda}$; $M = I - X(X^tX)^{-1}X$, y s_z es un estimador de la varianza de $\underline{z}$ derivada de la suma de cuadrados después de la regresión en X y $\underline{w}$, entonces

$$(n-p-1)s_z^2 = \underline{z}^T M \underline{z} - \frac{(\underline{z}^T M \underline{w})^2}{\underline{w}^T M \underline{w}}, \qquad (5.2.15)$$

además, todas las cantidades se calculan usando $\lambda = \lambda_0$. Asintoticamente, la distribución de $t_D(\lambda_0)$ es normal estándar, pero su distribución en muestras pequeñas es inestable ya que $\underline{z}$ y $\underline{w}$ son variables aleatorias con distribuciones no estandar. La transformación paramétrica estandar que se usa regularmente es

$$\underline{z}^{\lambda} = \begin{cases} \frac{\underline{y}^{\lambda} - 1}{\lambda \dot{y}^{\lambda - 1}}, & \lambda \neq 0 \\ \dot{y} \log \underline{y}, & \lambda = 0 \end{cases} \qquad (5.2.16)$$

donde $\dot{y}$ es la media geométrica de las observaciones

En realidad, $\underline{w}$ también se puede manejar como una variable adicional al modelo de regresión con las suposiciones de costumbre, es decir

$$E(\underline{y}) = X\underline{\beta} + \underline{w}\gamma \qquad (5.2.17)$$

El estimador mínimos cuadrados de γ en (5.2.17) puede escribirse en términos de los residuales de la regresión de $\underline{w}$ y X:

$$\underline{e} = M\underline{Y} = (I-V)\underline{Y}$$

$$\underline{w}^* = M\underline{w} = (I-V)\underline{w}$$

donde M y V están definidos en la sección 2.1 de este trabajo. Las fórmulas estandar de análisis de covarianza muestran que el estimador mínimos cuadrados de γ está dado por

$$\hat{\gamma} = \frac{\underline{w}^t(I-H)\underline{y}}{\underline{w}^t(I-H)\underline{w}} = \frac{\underline{w}^t M\underline{y}}{\underline{w}^t M\underline{w}} = \frac{\underline{w}^{*t}\underline{e}}{\underline{w}^{*t}\underline{w}^*}$$

esta expresión demuestra que $\hat{\gamma}$ es el estimador mínimos cuadrados de la pendiente de la regresión de los residuales $\underline{e}$ en los residuales $\underline{w}$. Los diagramas de estos dos conjuntos de residuales, llamados diagrama de variable agregada, se usan frecuentemente para evaluar la evidencia de términos adicionales en el modelo de regresión y observaciones influyentes ver Cook y Weisberg (1982, fig 2.3.8, 2.3.10), Belsley, Kuh y Welsch (1980 fig. 2.9 - 2.13)

En realidad para el estimador $\hat{\gamma}$, requerimos la prue

ba t para la significancia de la regresión, la varianza de $\hat{\gamma}$; $\sigma^2_{\hat{\gamma}} = \sigma^2 / \underline{w}^t M \underline{w}$, así que, la estadística para la hipótesis de no regresion es

$$t_{\underline{w}} = \frac{\hat{\gamma}}{\sqrt{\hat{\sigma}^2_{\hat{\gamma}}}} = \frac{\underline{w}^t M \underline{Y}}{s_w (\underline{w} M \underline{w})^{1/2}} = \frac{S(\underline{w}, \underline{Y})}{s_w \{S(\underline{w}, \underline{w})\}^{1/2}} \qquad (5.2.18)$$

donde $S(\underline{a}, \underline{b}) = \underline{a}^t (I - V) \underline{b} = \underline{a}^t M \underline{b}$ y con

$$(n-p-1) s^2_{\underline{w}} = \underline{Y}^t M \underline{Y} - (\underline{Y}^t M \underline{w})^2 / \underline{w}^t M \underline{w}$$
$$= S(\underline{Y}, \underline{Y}) - S^2(\underline{Y}, \underline{w}) / S(\underline{w}, \underline{w}) \qquad (5.2.19)$$

Entonces, como quedaría indicado por una tabla de análisis de covarianza, el efecto de agregar una variable $\underline{w}$ depende solamente de la suma de cuadrados y productos de $\underline{w}$ y $\underline{Y}$. Nótese que excepto por un signo sin importancia la expresion (5.2.18) y (5.2.14) son iguales.

Al escribir la transformación (5.2.16) en series de taylor evaluada en λ_0 obtenemos

$$\underline{Z}(\lambda) \simeq \underline{Z}(\lambda_0) + (\lambda - \lambda_0) \underline{w}(\lambda_0) \qquad (5.2.20)$$

$$\text{con } \underline{w}(\lambda_0) = \frac{\partial \underline{Z}(\lambda)}{\partial \lambda} \Big|_{\lambda = \lambda_0}$$

y se encuentra que el modelo lineal aproximado es

$$\underline{Z}(\lambda_0) = X\beta - (\lambda - \lambda_0) \underline{w}(\lambda_0) + \underline{\varepsilon} \qquad (5.2.21)$$

Al comparar esta última expresión con (5.2.17) se muestra que podemos obtener un estimador rápido (aproximado) para

λ como

$$\tilde{\lambda} = \tilde{\lambda}(\lambda_0) = \lambda_0 - \underline{w}^t(\lambda_0) M \underline{z}(\lambda_0) / \underline{w}^t(\lambda_0) M \underline{w}(\lambda_0)$$
$$= \lambda_0 - S(\underline{w}, \underline{z}; \lambda_0) / S(\underline{w}, \underline{w}; \lambda_0) \qquad (5.2.22)$$

similarmente la estadística aproximada de prueba para la transformación es

$$t_{\underline{w}}(\lambda_0) = -\underline{z}^t(\lambda_0) M \underline{w}(\lambda_0) / s_{\underline{z}} \{\underline{w}^t(\lambda_0) M \underline{w}(\lambda_0)\}^{1/2}$$
$$= -S(\underline{w}, \underline{z}; \lambda_0) / \{s_{\underline{z}}^2 S(\underline{w}, \underline{w}; \lambda_0)\}^{1/2} \qquad (5.2.23)$$

En (5.2.23) el estimador de $s_{\underline{z}}^2$, por analogía con (5.2.19) es

$$(n-p-1) s_{\underline{z}}^2 = S(\underline{z}, \underline{z}; \lambda_0) - \frac{S^2(\underline{w}, \underline{z}; \lambda_0)}{S(\underline{w}, \underline{w}; \lambda_0)} \qquad (5.2.24)$$

Un valor significante de $T_{\underline{w}}(\lambda_0)$ indicaría que se requiere una transformación de un valor diferente a λ_0. Debido a que la log-verosimilitud maximizada parcialmente es raramente cuadrática, la estadística $t_{\underline{w}}(\lambda_0)$ no tiene exactamente una distribución t con $n-p-1$ grados de libertad. Similarmente $\tilde{\lambda}$, el estimador rápido de la transformación, puede estar un poco distante al estimador maxima verosimilitud de λ.

Ahora desarrollaremos unas aproximaciones del efecto en $\tilde{\lambda}$ y en $t_{\underline{w}}(\lambda_0)$ al eliminar algunas observaciones. En las muestras se está interesado regularmente en una observación. La teoría sin embargo, se desarrolla facilmente para el caso más general en la eliminación de k observaciones señaladas por i.

Empecemos con la suma de cuadrados del residual del modelo usual de regresión, que es

$$(n-p)s^2 = \underline{Y}^t M \underline{Y} = S(\underline{Y}, \underline{Y})$$

Después de eliminar k observaciones, la suma de cuadrados residual está dado por

$$(n-p-k)s^2_{(i)} = S(\underline{Y}, \underline{Y}) - \underline{e}_i^t (I - V_i)^{-1} \underline{e}_i \qquad (5.2.25)$$

donde (i) y i como subíndice tiene la misma connotación que en el enfoque frecuentista. Así, e_i es el conjunto de k residuales indicados por i y V_i es la submatriz de $k \times k$ de V. Por conveniencia, escribiremos una notación general para la suma de cuadrados como los presentados en (5.2.25). Para vectores $\underline{a}, \underline{b}$ si los residuales, definidos por $\underline{a}^* = (I - V)\underline{a} = M\underline{a}$, $\underline{b}^* = (I-V)\underline{b} = M\underline{b}$. Entonces podemos escribir

$$S_{(i)}(\underline{a}, \underline{b}) = S(\underline{a}, \underline{b}) - \underline{a}_i^{*t} (1 - V_i)^{-1} \underline{b}_i^* \qquad (5.2.26)$$

donde $\underline{a}_i^*$ y $\underline{b}_i^*$ son los k miembros de los vectores de residuales. Para el modelo aumentado (5.2.19), el coeficiente estimado de regresión en $\underline{w}$ después de eliminar k observaciones esta dado por

$$\hat{\gamma} = S_{(i)}(\underline{w}, \underline{Y}) / S_{(i)}(\underline{w}, \underline{w}) \qquad (5.2.27)$$

Usando esta expresión con w reemplazada por $\underline{w}(\lambda_0)$ para obtener un estimador rápido del parámetro de la transformación después de eliminar k observaciones. De (5.2.26) y (5.2.27).

$$\tilde{\lambda}_{(i)}(\lambda_0) = \lambda_0 - S_{(i)}(\underline{w}, \underline{z}; \lambda_0) / S_{(i)}(\underline{w}, \underline{w}; \lambda_0)$$

$$= \lambda_0 - \frac{\underline{w}^t M \underline{z} - \underline{w}_i^t (I - V_i)^{-1} \underline{z}_i^*}{\underline{w}^t M \underline{w} - \underline{w}_i^t (I - V_i)^{-1} \underline{w}_i^*} \qquad (5.2.28)$$

donde los residuales $\underline{w}_i^*$ y $\underline{z}_i^*$ se evalúan en λ_0.

La derivación de un estimador rápido de $\lambda(\lambda_0)$ requiere aproximación de series de taylor para la transformación del modelo. La derivación de $\tilde{\lambda}_{(i)}(\lambda_0)$ vía (5.2.27) incluye una aproximación adicional, ya que las variables $\underline{z}(\lambda_0)$ son funciones de la media geométrica de todas las observaciones. Cuando k observaciones son eliminadas, la media geométrica de las $n-k$ observaciones restantes no tendrá el mismo valor. En general podemos esperar que el cambio en $\dot{y}$ sea despreciable si k es pequeña relativa a n, excepto si una de las observaciones eliminadas fuera muy cercana a cero como lo son las otras. Como el propósito de esta técnica es proporcionar cantidades de diagnóstico que se calculen rapidamente, esta dificultad será ignorada en el desarrollo teórico.

El efecto de la eliminación de observaciones en la estadística (5.2.23) puede derivarse en forma similar. Se sigue de (5.2.25) que el efecto de eliminación para producir una estadística aproximada

$$\tilde{t}_{\underline{w}(i)}(\lambda_0) = -S_{(i)}(\underline{w}, \underline{z}; \lambda_0) / \{ s^2_{\underline{z}(i)} S_{(i)}(\underline{w}, \underline{w}; \lambda_0) \}^{1/2} \qquad (5.2.29)$$

donde, por analogía con (5.2.17)

$$(n-p-k-1)\hat{\sigma}^2_{z(i)} = S_{(i)}(\underline{z}, \underline{z}; \lambda_0)/s_{(i)}(\underline{w}, \underline{w}; \lambda_0) \qquad (5.2.30)$$

En particular, si solo eliminamos la observación

$$\lambda_{(i)}(\lambda_0) = \lambda_0 - \frac{\underline{w}^t M \underline{z} - w_i^* z_i/(1-v_i)}{\underline{w}^T M \underline{w} - \underline{w}_i^t \underline{z}/(1-v_i)} \qquad (5.2.31)$$

tambien se puede obtener una simplificación similar para la estadística $\tilde{T}_{w(i)}(\lambda_0)$ definida en (5.2.29) y (5.2.30). Adicionalmente, se puede obtener una simplificación para la variable $\underline{w}(\lambda)$ definida en (5.2.20) para la transformación de potencia (5.2.16). Para cualquier λ

$$\underline{w}(\lambda) = \frac{\partial \underline{z}(\lambda)}{\partial \lambda}\Big/_{\lambda=\lambda_0} = \frac{\underline{y}^\lambda \log \underline{y}}{\lambda \dot{y}^{\lambda-1}} - \frac{\underline{y}^\lambda - 1}{\lambda \dot{y}^{\lambda-1}}\left(1/\lambda + \log \dot{y}\right) \qquad (5.2.32)$$

Debido a que las cantidades de diagnostico dependen solo de los residuales de $\underline{w}(\lambda)$, pueden encontrarse formas simples para $\underline{w}(\lambda)$ para ciertos valores de λ. En particular

$$\underline{w}(1) = \underline{y}\{\log(\underline{y}/\dot{y}) - 1\}$$

$$\underline{w}(0) = \dot{y} \log \underline{y}(\log \underline{y}/2 - \log \dot{y}) \qquad (5.2.33)$$

5.3 COMENTARIOS

Uno de los principales inconvenientes que presentan los

métodos de máxima verosimilitud para estimar el parámetro de la transformación λ, es que, éstos se ven afectados por observaciones discrepantes, confirmando el hecho de que máxima verosimilitud en el modelo lineal normal tiende a no ser robusto: Andrews (1971). Además, estos métodos requieren de calculos repetidos al realizar varias transformaciones de los datos originales, resultando por esto, un método bastante tedioso cuando se usa. Adicionalmente, las pruebas y los límites de confianza que se basan en ellos solo tienen validez asintotica y el número de parámetros debe ser chico comparado con el numero de observaciones

Por otro lado, el método de Andrews presenta las siguientes ventajas sobre el de máxima verosimilitud.

a) La prueba de significancia es exacta

b) La cantidad de cálculos involucrados se reduce cuando se prueba una o pocas transformaciones.

c) La precisión de la transformación puede evaluarse teoricamente.

Una de sus principales desventajas es que no se pueden obtener conclusiones por medio de graficas como las que se obtienen con diagramas de verosimilitud y cuando la distribución de los errores es normal, la prueba de Andrews es menos potente que la prueba del cociente de verosimilitud: Atkinson (1973). En cuanto al método de Carrol (1980), este demuestra por medio de un experimento

de Montecarlo que su procedimiento es preferible al de verosimilitud en términos de inferencias acerca de λ y β cuando se consideran modelos para el error diferentes al normal. también el método parece preferible al de Andrews debido a la falta de potencia de éste último; esta ventaja se ve disminuida un poco, por un pequeño incremento en el nivel de la prueba de una transformación particular. Además el método de Carroll parece ligeramente menos robusto contra observaciones discrepantes que el método de Andrews.

5.4. Ejemplos.

En esta sección unicamente se aplican algunos métodos de transformaciones al grupo de datos Brownlee, K.A. (1965) para determinar si es necesaria una transformación, se calcula la estadística T_D de Atkinson (1973) cuyo valor es -3.29. Dado que este resultado indica la necesidad de una transformación se calcula $\hat{\lambda}$ (estimador de máxima verosimilitud), donde $\hat{\lambda} = 0.30$ y un intervalo de confianza del 95% $(-0.19, 0.74)$ que da evidencia de una transformación logaritmica. Cuándo se efectúa dicha transformación y calcular T_D, ésta no proporciona elementos para una nueva transformación, ver Atkinson (1981), (1982) para mayores detalles.

Como hemos visto la observación 21 es la más discrepante, si la eliminamos $T_D = -2.53$ y $\hat{\lambda} = 0.48$ con $(0.05, 0.87)$ un intervalo de confianza del 95%. De lo anterior podemos concluir que en la observación más discrepante no esta contenida toda la información necesaria para una transformación.

Por otro lado, al aplicar procedimientos estándar de regresión

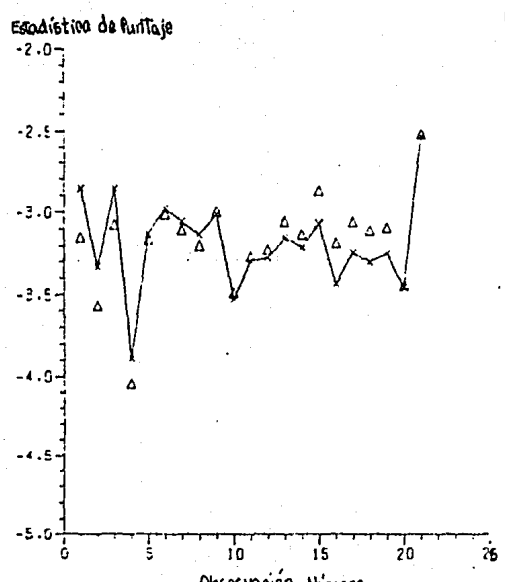
para la adición y eliminación de variables nos conduce (Atkinson (1981)) al modelo $\log Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_1 X_2 + \varepsilon$ para el cuál la suma de cuadrados residual es 64.4 que es todavía mayor que 20.40, valor que se obtiene con el modelo de primer orden (2.5.3); pero a pesar de todo, no existe con este modelo evidencia de observaciones discrepantes.

En la gráfica (5.4.1) se muestra un diagrama de la estadística $\tilde{T}_{W(i)}(\lambda_0)$ y $T_{W(i)}(\lambda_0)$ definido en (5.2.23), $T_{W(i)}(\lambda_0)$ la estadística (5.2.19) cuándo se eliminan del modelo (2.5.1) las K observaciones indicadas por i; para este caso $i=\{i\}$ y $\lambda_0 = 1$. Aunque la iluminación de la observación 21 causa solo una pequeña reducción en la evidencia para una transformación, las dos estadísticas, $T_{W(i)}(\lambda_0)$ y $\tilde{T}_{W(i)}(\lambda_0)$ concuerdan en que los datos deben de ser transformados. Por último, la gráfica (5.4.2) muestra los valores de $\tilde{\lambda}_{(i)}$, $\hat{\lambda}_{(i)}^{(a)}$ y $\hat{\lambda}_{(i)}$ y de la cuál podemos concluir que la eliminación de la observación 21 causa que la transformación estimada que era de aproximadamente de 1/3 con todos los datos a un valor muy cercano a 1/2.

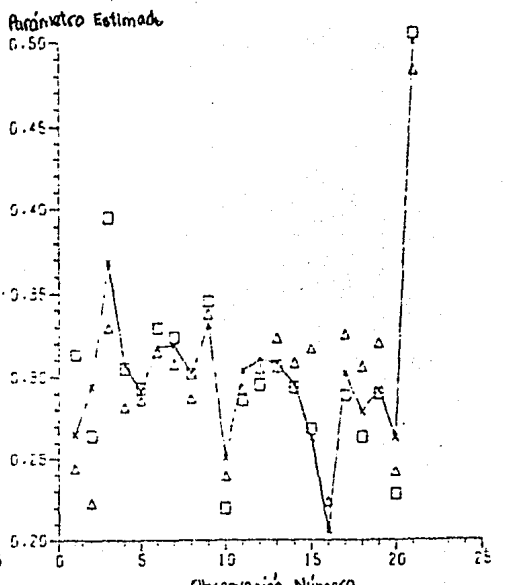
5.5. Conclusiones

Como ya se mencionó anteriormente, antes de aplicar algún método para el manejo de observaciones discrepantes, sería adecuado verificar todas las hipótesis subyacentes a los modelos, si en esta etapa no existe algo anormal, entonces buscar alguna transformación que permita - "unir" los datos. De entre estos métodos, tal vez el que mayor difusión a tenido es el de mínimos cuadrados, pero debido a que es un proceso tedioso se buscan nuevas alternativas. Aunque siempre es conveniente tener un método robusto, de los métodos presentados y que más desarrollo ha tenido es el de Atkinson. Este

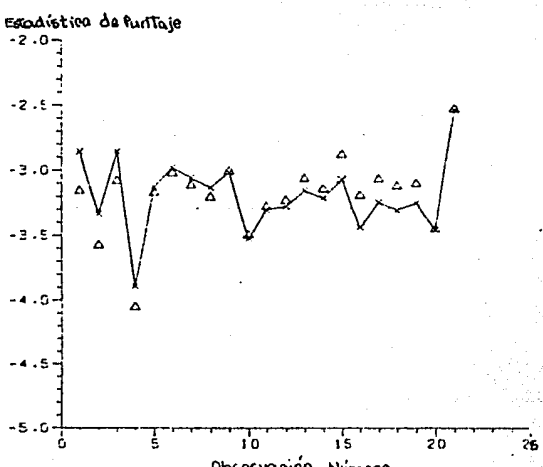
Gráfica 5.4.1.



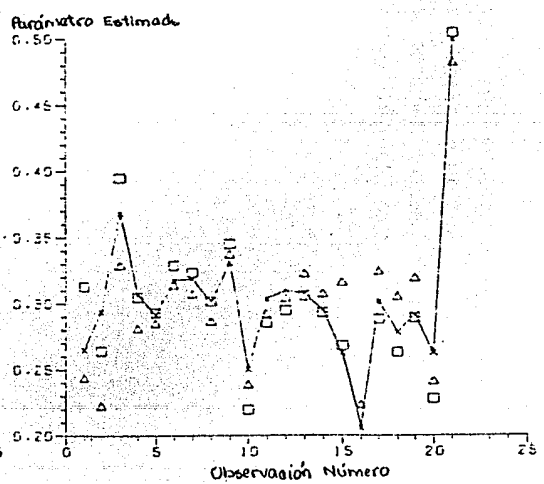
Gráfica 5.4.2.



Gráfica 5.4.1.



Gráfica 5.4.2.



método presenta aparentemente mayores ventajas debido principalmente a que su aplicación puede efectuarse casi directamente a partir de los programas usuales de regresión, además pueden calcularse valores rápidos para el valor λ de la transformación.

Conclusiones Generales.

En el desarrollo de este trabajo, hemos visto una gran variedad de métodos para el manejo de observaciones discrepantes y notado, al menos en modelos lineales, no existe un método óptimo en todos sentidos, y además, en la mayoría de los casos éstos proporcionan esencialmente los mismos resultados. Por otro lado, también se observa que dichos métodos, para el caso de $K=1$, se basan principalmente en residuales y matrices de proyección, y para $K>1$ en cantidades equivalentes como son $G_K(i)$, $V(i)$.

Naturalmente la elección de un método no es fácil y ésto dependerá de la filosofía que cada uno de nosotros considere tiene mayores ventajas, aunque claro está, uno podría mezclar varios enfoques para un análisis más profundo, tambien de la información a priori disponible, y por supuesto de los objetivos específicos del análisis. La verdad es que muchos de los métodos que hemos analizado pueden tener un lugar en un problema específico, dependiendo de los requisitos del investigador: desde métodos cuya aplicación no representa problema alguno (gráficas, resultados tradicionales de mínimos cuadrados, etc.); métodos ligeramente sofisticados (métodos robustos principalmente) hasta métodos para los cuales, casi en toda situación, se tenga que construir un modelo particular (métodos Bayesianos).

Pero la cuestión de cuando una observación es mala o informativa acerca de lo inapropiado del modelo y cuál es la mejor técnica para distinguir tales observaciones, no puede decidirse en base a un conjunto de datos analizados a través de los años, para los cuales cada autor de un método propone su modelo. Lo

que se requiere hacer es aplicar éstos métodos a "datos reales" donde se puedan checar las mediciones sospechosas y si es necesario, repetirlos.

¿Debemos sujetar a rutina todas las muestras a algún tipo de observaciones discrepantes?. La respuesta es sí, en modelos lineales, por ejemplo, una inspección de los residuales studentizados, la diagonal de V y una medida de influencia por lo menos para un caso deberían de incluirse como parte de la rutina en fase de diag. nóstico de cada análisis. Generalmente, dichos análisis son una herramienta importante para guiar el análisis subsecuente y puede ser un indicador de la necesidad de acomodo, transformación o revisión del modelo.

¿Debemos buscar en forma rutinaria rechazar observaciones discrepantes en base a pruebas formales?. La respuesta es no. Durante la fase de construcción del modelo, por ejemplo, las pruebas formales no deben de usarse ya que requieren que el modelo nulo sea exacto en términos de la esperanza y distribución de los errores. Deben de buscarse pruebas formales solo en situaciones en donde se sabe que el modelo nulo es exacto en ausencia de contaminantes, y el interés se centra en el estudio del fenómeno alternativo. Existe poca información sobre el funcionamiento de los estimadores mínimos cuadrados usuales en combinación con rechazo vía pruebas formales.

En modelos lineales normales, quizás el problema más difícil y que no se ha resuelto a satisfacción, es el concerniente a como tratar con la posibilidad de más de una observación discrepante ($k>1$); cuando hay poca información a priori importante sobre el número y tipos de observaciones discrepantes cuando, identificación

es lo más importante. Algunos métodos recomendados de identificación surgen como derivado de métodos de acomodo. Los pesos de regresión robusta y métodos bayesianos, por ejemplo, caen en esta categoría. Sin embargo, los métodos de acomodo se basan en información a priori específica como simetría, por ejemplo y se desarrolla con las propiedades de los estimadores resultantes en mente.

Tabla de Lund ($\alpha = .10$)

n	c									
	1	2	3	4	5	6	9	10	15	25
5	1.97									
6	2.00	1.89	-							
7	2.10	2.02	1.98							
8	2.18	2.12	2.03	1.91						
9	2.24	2.20	2.13	2.05	1.92					
10	2.30	2.26	2.21	2.15	2.06	1.92				
12	2.39	2.37	2.33	2.29	2.24	2.17	1.93			
14	2.47	2.45	2.42	2.39	2.36	2.32	2.19	1.94		
16	2.53	2.51	2.50	2.47	2.45	2.42	2.34	2.20		
18	2.58	2.57	2.56	2.54	2.52	2.50	2.44	2.35		
20	2.63	2.62	2.61	2.59	2.58	2.56	2.52	2.46	2.11	
25	2.72	2.72	2.71	2.70	2.69	2.68	2.66	2.63	2.50	
30	2.80	2.79	2.79	2.78	2.77	2.77	2.75	2.73	2.66	2.13
35	2.86	2.85	2.85	2.85	2.84	2.84	2.82	2.81	2.77	2.55
40	2.91	2.91	2.90	2.90	2.90	2.89	2.88	2.87	2.84	2.72
45	2.95	2.95	2.95	2.95	2.94	2.94	2.93	2.93	2.90	2.82
50	2.99	2.99	2.99	2.99	2.98	2.98	2.98	2.97	2.95	2.89
60	3.06	3.06	3.05	3.05	3.05	3.05	3.05	3.04	3.03	3.03
70	3.11	3.11	3.11	3.11	3.11	3.11	3.10	3.10	3.09	3.07
80	3.16	3.16	3.16	3.15	3.15	3.15	3.15	3.15	3.14	3.12
90	3.20	3.20	3.19	3.19	3.19	3.19	3.19	3.19	3.18	3.17
100	3.23	3.23	3.23	3.23	3.23	3.23	3.23	3.22	3.22	3.21

($\alpha = .05$)

n	c									
	1	2	3	4	5	6	9	10	15	25
5	1.92									
6	2.07	1.93								
7	2.19	2.08	1.94							
8	2.28	2.20	2.10	1.94						
9	2.35	2.29	2.21	2.10	1.95					
10	2.42	2.37	2.31	2.22	2.11	1.95				
12	2.52	2.49	2.45	2.39	2.33	2.24	1.96			
14	2.61	2.59	2.55	2.51	2.47	2.41	2.25	1.96		
16	2.68	2.65	2.63	2.60	2.57	2.53	2.43	2.26		
18	2.73	2.72	2.70	2.68	2.65	2.62	2.55	2.44		
20	2.78	2.77	2.76	2.74	2.72	2.70	2.64	2.57	2.15	
25	2.89	2.88	2.87	2.86	2.84	2.83	2.80	2.76	2.60	
30	2.96	2.96	2.95	2.94	2.93	2.93	2.90	2.88	2.79	2.17
35	3.03	3.02	3.02	3.01	3.00	3.00	2.98	2.97	2.91	2.64
40	3.08	3.08	3.07	3.07	3.06	3.06	3.05	3.03	3.00	2.84
45	3.13	3.12	3.12	3.12	3.11	3.11	3.10	3.09	3.06	2.96
50	3.17	3.16	3.16	3.16	3.15	3.15	3.14	3.14	3.11	3.04
60	3.23	3.23	3.23	3.23	3.22	3.22	3.22	3.21	3.20	3.15
70	3.29	3.29	3.28	3.28	3.28	3.28	3.27	3.27	3.26	3.23
80	3.33	3.33	3.33	3.33	3.33	3.33	3.32	3.32	3.31	3.29
90	3.37	3.37	3.37	3.37	3.37	3.37	3.36	3.36	3.36	3.34
100	3.41	3.41	3.40	3.40	3.40	3.40	3.40	3.40	3.39	3.38

Continuación ($\alpha = .01$)

n	p									
	1	2	3	4	5	6	8	10	15	25
5	1.93									
6	2.17	1.98								
7	2.32	2.17	1.98							
8	2.44	2.32	2.18	1.98						
9	2.54	2.44	2.33	2.18	1.99					
10	2.62	2.55	2.45	2.33	2.18	1.99				
12	2.76	2.70	2.64	2.56	2.46	2.34	1.99			
14	2.86	2.82	2.78	2.72	2.65	2.57	2.35	1.99		
16	2.95	2.92	2.88	2.84	2.79	2.73	2.53	2.35		
18	3.02	3.00	2.97	2.94	2.90	2.85	2.75	2.59		
20	3.08	3.06	3.04	3.01	2.98	2.95	2.87	2.76	2.20	
25	3.21	3.19	3.18	3.16	3.14	3.12	3.07	3.01	2.73	
30	3.30	3.29	3.28	3.26	3.25	3.24	3.21	3.17	3.04	2.21
35	3.37	3.36	3.35	3.34	3.34	3.33	3.30	3.28	3.19	2.91
40	3.43	3.42	3.42	3.41	3.40	3.40	3.38	3.36	3.30	3.08
45	3.48	3.47	3.47	3.46	3.46	3.45	3.44	3.43	3.38	3.23
50	3.52	3.52	3.51	3.51	3.51	3.50	3.49	3.48	3.45	3.34
60	3.60	3.59	3.59	3.59	3.58	3.58	3.57	3.56	3.54	3.48
70	3.65	3.65	3.65	3.65	3.64	3.64	3.64	3.63	3.61	3.57
80	3.70	3.70	3.70	3.70	3.69	3.69	3.69	3.68	3.67	3.64
90	3.74	3.74	3.74	3.74	3.74	3.74	3.73	3.73	3.72	3.70
100	3.78	3.78	3.78	3.77	3.77	3.77	3.77	3.77	3.76	3.74

n = número de observaciones.

p = número de variables independientes.

Tabla de Weisberg para la estadística t_i

$\alpha = .05$

n / p	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
6	4.85	6.23	10.89	76.39											
7	4.38	5.07	6.58	11.77	89.12										
8	4.12	4.63	5.26	6.90	12.59	101.9									
9	3.95	4.22	4.66	5.44	7.18	13.36	114.6								
10	3.83	4.03	4.32	4.77	5.60	7.45	14.09	127.3							
11	3.75	3.90	4.10	4.40	4.88	5.75	7.70	14.78	140.1						
12	3.69	3.81	3.96	4.17	4.49	4.98	5.89	7.94	15.44	152.8					
13	3.65	3.74	3.86	4.02	4.24	4.56	5.08	6.02	8.16	16.08	165.5				
14	3.61	3.69	3.79	3.91	4.07	4.30	4.63	5.16	6.14	8.37	16.69	178.2			
15	3.58	3.65	3.73	3.83	3.95	4.12	4.36	4.70	5.25	6.25	8.58	17.28	191.0		
16	3.56	3.62	3.68	3.77	3.87	4.00	4.17	4.41	4.76	5.33	6.36	8.77	17.85	203.7	
17	3.54	3.59	3.65	3.72	3.80	3.90	4.04	4.21	4.46	4.82	5.40	6.47	8.95	18.40	216.4
18	3.53	3.57	3.62	3.68	3.75	3.83	3.94	4.08	4.26	4.51	4.88	5.47	6.57	9.13	18.93
19	3.52	3.56	3.60	3.65	3.71	3.78	3.86	3.97	4.11	4.30	4.55	4.93	5.54	6.67	9.30
20	3.51	3.54	3.58	3.62	3.67	3.73	3.81	3.89	4.00	4.15	4.33	4.59	4.98	5.60	6.76
21	3.50	3.53	3.57	3.60	3.65	3.70	3.76	3.83	3.92	4.03	4.18	4.37	4.64	5.03	5.67
22	3.50	3.52	3.55	3.59	3.63	3.67	3.72	3.78	3.86	3.95	4.06	4.21	4.40	4.68	5.09
23	3.49	3.52	3.54	3.57	3.61	3.65	3.69	3.75	3.81	3.88	3.98	4.09	4.24	4.44	4.71
24	3.49	3.51	3.53	3.56	3.59	3.63	3.67	3.71	3.77	3.83	3.91	4.00	4.12	4.27	4.47
25	3.49	3.50	3.53	3.55	3.58	3.61	3.65	3.69	3.73	3.79	3.85	3.93	4.02	4.14	4.30
26	3.48	3.50	3.52	3.54	3.57	3.60	3.63	3.66	3.70	3.75	3.81	3.87	3.95	4.05	4.17
27	3.48	3.50	3.52	3.54	3.56	3.58	3.61	3.65	3.68	3.72	3.77	3.83	3.89	3.97	4.07
28	3.48	3.50	3.51	3.53	3.55	3.58	3.60	3.63	3.66	3.70	3.74	3.79	3.84	3.91	3.99
29	3.48	3.49	3.51	3.53	3.55	3.57	3.59	3.62	3.64	3.68	3.71	3.76	3.81	3.86	3.93
30	3.48	3.49	3.51	3.52	3.54	3.56	3.58	3.60	3.63	3.66	3.69	3.73	3.77	3.82	3.88
31	3.48	3.49	3.50	3.52	3.54	3.55	3.57	3.59	3.62	3.64	3.67	3.71	3.74	3.79	3.84
32	3.48	3.49	3.50	3.52	3.53	3.55	3.57	3.59	3.61	3.63	3.66	3.69	3.72	3.76	3.80
33	3.48	3.49	3.50	3.52	3.53	3.54	3.56	3.58	3.60	3.62	3.64	3.67	3.70	3.74	3.77
34	3.48	3.49	3.50	3.51	3.53	3.54	3.56	3.57	3.59	3.61	3.63	3.66	3.68	3.71	3.75
35	3.48	3.49	3.50	3.51	3.52	3.54	3.55	3.57	3.58	3.60	3.62	3.64	3.67	3.70	3.73
36	3.48	3.49	3.50	3.51	3.52	3.54	3.55	3.56	3.58	3.60	3.61	3.63	3.66	3.68	3.71
37	3.48	3.49	3.50	3.51	3.52	3.53	3.55	3.56	3.57	3.59	3.61	3.62	3.65	3.67	3.69
38	3.48	3.49	3.50	3.51	3.52	3.53	3.54	3.56	3.57	3.58	3.60	3.62	3.64	3.66	3.68
39	3.49	3.49	3.50	3.51	3.52	3.53	3.54	3.55	3.57	3.58	3.59	3.61	3.63	3.65	3.67
40	3.49	3.49	3.50	3.51	3.52	3.53	3.54	3.55	3.56	3.58	3.59	3.60	3.62	3.64	3.66
50	3.51	3.51	3.51	3.52	3.53	3.53	3.54	3.54	3.55	3.56	3.57	3.57	3.58	3.59	3.60
60	3.53	3.53	3.53	3.54	3.54	3.54	3.55	3.55	3.56	3.56	3.57	3.57	3.58	3.58	3.59
70	3.55	3.55	3.55	3.55	3.56	3.56	3.56	3.56	3.57	3.57	3.57	3.58	3.58	3.59	3.59
80	3.57	3.57	3.57	3.57	3.57	3.58	3.58	3.58	3.58	3.58	3.59	3.59	3.59	3.60	3.60
90	3.58	3.59	3.59	3.59	3.59	3.59	3.59	3.60	3.60	3.60	3.60	3.60	3.60	3.61	3.61
100	3.60	3.60	3.60	3.60	3.61	3.61	3.61	3.61	3.61	3.61	3.61	3.62	3.62	3.62	3.62
200	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.73	3.74
300	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.81	3.82	3.82	3.82	3.82	3.82
400	3.87	3.87	3.87	3.87	3.87	3.87	3.87	3.88	3.88	3.88	3.88	3.88	3.88	3.88	3.88
500	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92	3.92

tabla de Weisberg - Continuación

$\alpha = .01$

n / p	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
6	7.53	10.87	24.46	382.0											
7	6.35	7.84	11.45	26.43	445.6										
8	5.71	6.54	8.12	11.98	28.26	509.3									
9	5.31	5.84	6.71	8.38	12.47	29.97	573.0								
10	5.04	5.41	5.96	6.87	8.61	12.92	31.60	636.6							
11	4.85	5.12	5.50	6.07	7.01	8.83	13.35	33.14	700.3						
12	4.71	4.91	5.19	5.58	6.17	7.15	9.03	13.75	34.62	763.9					
13	4.60	4.76	4.97	5.25	5.66	6.26	7.27	9.22	14.12	36.03	827.6				
14	4.51	4.64	4.81	5.02	5.32	5.73	6.35	7.39	9.40	14.48	37.40	891.3			
15	4.44	4.55	4.68	4.85	5.05	5.37	5.80	6.43	7.50	9.57	14.82	38.71	954.9		
16	4.38	4.48	4.59	4.72	4.90	5.12	5.43	5.86	6.51	7.60	9.73	15.15	39.98		
17	4.34	4.41	4.51	4.62	4.76	4.94	5.17	5.48	5.92	6.59	7.70	9.83	15.46	41.21	
18	4.30	4.36	4.44	4.54	4.66	4.80	4.98	5.21	5.53	5.98	6.66	7.80	10.03	15.76	42.41
19	4.26	4.32	4.39	4.47	4.57	4.69	4.83	5.01	5.25	5.57	6.03	6.72	7.89	10.17	16.05
20	4.23	4.29	4.35	4.42	4.50	4.60	4.72	4.86	5.05	5.29	5.62	6.08	6.79	7.98	10.31
21	4.21	4.26	4.31	4.37	4.44	4.52	4.62	4.74	4.89	5.08	5.33	5.66	6.13	6.85	8.06
22	4.19	4.23	4.28	4.33	4.39	4.46	4.55	4.65	4.77	4.92	5.11	5.36	5.70	6.18	6.91
23	4.17	4.21	4.25	4.30	4.35	4.41	4.49	4.57	4.67	4.80	4.95	5.14	5.40	5.74	6.22
24	4.15	4.19	4.22	4.27	4.32	4.37	4.43	4.51	4.59	4.70	4.82	4.98	5.17	5.43	5.78
25	4.14	4.17	4.20	4.24	4.28	4.33	4.39	4.45	4.53	4.62	4.72	4.85	5.00	5.20	5.46
26	4.12	4.15	4.18	4.22	4.26	4.30	4.35	4.41	4.47	4.55	4.64	4.74	4.87	5.03	5.23
27	4.11	4.14	4.17	4.20	4.24	4.27	4.32	4.37	4.43	4.49	4.57	4.66	4.76	4.89	5.05
28	4.10	4.13	4.15	4.19	4.21	4.25	4.29	4.33	4.38	4.44	4.51	4.59	4.68	4.78	4.91
29	4.09	4.12	4.14	4.17	4.20	4.23	4.26	4.30	4.35	4.40	4.46	4.53	4.60	4.69	4.80
30	4.09	4.11	4.13	4.15	4.18	4.21	4.24	4.28	4.32	4.36	4.42	4.47	4.54	4.62	4.71
31	4.08	4.10	4.12	4.14	4.17	4.19	4.22	4.26	4.29	4.33	4.38	4.43	4.49	4.56	4.64
32	4.07	4.09	4.11	4.13	4.15	4.18	4.21	4.24	4.27	4.31	4.35	4.39	4.45	4.50	4.57
33	4.07	4.08	4.10	4.12	4.14	4.17	4.19	4.22	4.25	4.28	4.32	4.36	4.41	4.46	4.52
34	4.06	4.08	4.09	4.11	4.13	4.15	4.18	4.20	4.23	4.26	4.29	4.33	4.37	4.42	4.47
35	4.06	4.07	4.09	4.11	4.12	4.14	4.16	4.19	4.21	4.24	4.27	4.31	4.34	4.39	4.43
36	4.05	4.07	4.08	4.10	4.12	4.13	4.15	4.18	4.20	4.22	4.25	4.28	4.32	4.36	4.40
37	4.05	4.06	4.08	4.09	4.11	4.13	4.14	4.16	4.19	4.21	4.24	4.26	4.29	4.33	4.37
38	4.05	4.06	4.07	4.09	4.10	4.12	4.13	4.15	4.17	4.20	4.22	4.25	4.27	4.31	4.34
39	4.04	4.06	4.07	4.08	4.10	4.11	4.13	4.14	4.16	4.18	4.21	4.23	4.26	4.28	4.32
40	4.04	4.05	4.06	4.08	4.09	4.10	4.12	4.14	4.15	4.17	4.19	4.22	4.24	4.27	4.29
50	4.03	4.03	4.04	4.05	4.06	4.07	4.07	4.08	4.09	4.10	4.12	4.13	4.14	4.15	4.17
60	4.03	4.03	4.04	4.04	4.05	4.05	4.06	4.06	4.07	4.08	4.08	4.09	4.10	4.11	4.12
70	4.03	4.03	4.04	4.04	4.05	4.05	4.05	4.06	4.06	4.07	4.07	4.08	4.08	4.09	4.09
80	4.04	4.04	4.04	4.05	4.05	4.05	4.06	4.06	4.06	4.07	4.07	4.07	4.08	4.08	4.09
90	4.05	4.05	4.05	4.05	4.06	4.06	4.06	4.06	4.07	4.07	4.07	4.07	4.08	4.08	4.08
100	4.06	4.06	4.06	4.06	4.06	4.07	4.07	4.07	4.07	4.07	4.08	4.08	4.08	4.08	4.09
200	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15	4.15
300	4.21	4.21	4.21	4.21	4.21	4.21	4.22	4.22	4.22	4.22	4.22	4.22	4.22	4.22	4.22
400	4.26	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27	4.27
500	4.31	4.31	4.31	4.31	4.31	4.31	4.31	4.31	4.31	4.31	4.31	4.31	4.31	4.31	4.31

BIBLIOGRAFÍA.

Abraham, B., and Box, G.E.P (1978) "Linear Models and Spurious Observations", Applied Statistics, 27, 131-138.

Aitkin, M., and Wilson, G.t. (1980) "Mixture Models, Outliers and the E.M. Algorithm", Technometrics 22, 325-332.

Andrews, D.F.(1971), "A note on the Selección of data transformations"; Biometrika, 58, 249-254.

————, Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and tukey, J.W. (1972), "Robust Estimates of Location", Princeton, N.J.: Princeton University Press.

————, and Pregibon, D. (1978) "Finding the Outliers that — Matter", Journal of the Royal Statistical Society Ser. B, 40 87-93.

Anscombe, F.J. (1960) "Rejection of Outliers" technometrics 2, 123-147.

Atkinson, A.C. (1981), "Robustness transformations and two Graphical Displays for Outlying and Influential Observations in Regression", Biometrika, 68, 13-20.

————, (1982), "Robust and Diagnostic Regression Analysis" Commun. Statist. theor. Meth., 11(22), 2559-2571.

————,(1982), "Regression Diagnostics Transformations and Constructed -

Variables" (with discussion), Journal of the Royal Statistics Society, Ser. B, 44, 1-35.

———, (1986), "Diagnostics for transformations", technometrics, 28, 29-34.

Barnett, V. (1973), "Comparative Statistical Inference", John Wiley and Sons, Inc. New York.

———, (1978), "the Study of Outliers: Purpose and Model", Applied Statistics 27, 242-250.

———, and Lewis, T (1978), "Outliers in Statistical Data", New York: John Wiley.

Beckman, R.J. and trussell, H.J., (1974) "the Distribution of an Arbitrary Studentized Residual and the Effects of Updating in Mustiple Regression", Journal of the American Statistical Association 69, 199—201.

———, and Cook, R.D. (1983) "Outlier....s", technometrics 25, 119-149.

Becroolli, D. (1777), "the Most Probable Choice Between Several Discrepant Observations and the Formation there from of the Most Likely Induction" in C.G. Allen (1961), Biometrika 48, 3-13.

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

Bikel, P.J. and Doksum, K.A. (1981), "An Analysis of transformations Revisited", Journal of the Royal Statistical Association, 76, 296-311.

Box, G.E.P (1979), "Robustness in Strategy of Scientific Model - Building" eds. R.L. Launer and G.N. Wilkinson, New York, Academic Press.

———, (1980), "Sampling and Bayes Inference in Scientific Modelling and Robustness (with discution)." J.R. Statist. Soc. A., 143, 383-480.

———, and Anderson S.L. (1955), "Permutation theory in the Derivation of Robust Criterio and the Study of Departures from Assumption," J. Roy Statist. Soc. Ser: B, 17, 1-34.

———, and Cox, D.R. (1964), "An Analysis of transformations (With discussion)", J.R. Statist Soc. A, 143, 383-430.

———, and tiao, G.C. (1968), "A Bayesian Approach to Some Outlier Problems", Biometrika, 55, 119-129.

———, and tiao, G.C. (1973), "Bayesian Inference in Statistical Analysis. Addison-Wesley. Reading, Massachusetts.

———, and tidwell, P.W. (1962), "transformations of the Independent Variables", technometrics 4, 531-550.

Broemeling, L.D. (1985) "Bayesian Analysis of Linear Models", Oklahoma State University Marcel Dekker, Inc: New York and Basel.

Brownlee, K. A. (1965). Statistical theory and Methodology in Science and Engineering, New York: John Wiley.

Carroll, R.J. (1980 a) "A Robust method for testing transformations to Achieve Approximate normality", J.R. Statist. Soc. B, 42, 71-78.

———, (1980b), "Robust Methods for factorial Experiments with Outliers" Applied Statics 29, 246-251.

———, (1982), "two Examples of transformations when there are Possible Outliers". Appl. Statist. 31, 149-152.

Chambers, R.L. and Heatcote, C.E. (1981), "On the estimation of slope and the Identification of Outliers in Linear Regression", Biometrika, 68, 21-33.

Collet, D. and Lewis, T (1976), "the subjetive Nature of Outlier Rejection Proaedures", Applied Statistics, 25, 228-237.

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

———, (1979), "Influential Observations in Linear Regression", Journal of the American Statistical Association 74, 169-174.

———, Holschuh, N., and Weisberg, S. (1982), "A note on an Alternative Outlier Model", Journal of the Royal Statistical Societus, Ser B. 44, 370-376.

———, and Prescott, P. (1981), "On the Accuracy of Bonferroni Significance Levels for Detecting Outliers in Linear Models", Technometrics 29, 59-63.

———, and Wang, P.C. (1983), "transformations and Influential Cases in Regression", Technometrics 25, 337-343.

———, and Weisberg, S. (1980), "Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression", Technometrics, 22, 495-508

———, (1982), "Residuals and Influence in Regression, New York: Chapman Hall.

Daniel, C., and Woods, F.S. (1971), "Fitting Equations to Data", New York: John Wiley.

Danniel, P.J. (1920), "Observations Weighted According to Order", American Journal of Mathematics, 42, 222-236.

Davies, R.B, and Hutton, B. (1975) "the Effect of Errors in the Independent Variables in Linear Regression" Biometrika 62, 383-391.

Demster, A.P. and Gasko-Green, M. (1981), "New tools for Residual Analysis", Annals of Statistics, 9, 945-959.

———, Laird, N.M. and Rubin, D.B. (1977), "Maximun Likelihood from Incomplete Data via the E.M. Algorithm (with discussion) J. Roy. Statist. Soc. B, 39, 1-38.

Doornbos, R. (1981), "testing for a Single Outlier in a Linear Model", Biometrika 37, 705-712.

Draper, N. R., and John J. A. (1980), "testing for three or Fewer Outliers in two-way tables", Technometrics 22, 9-15.

———, and John J. A. (1981), "Influential Observations and Outliers in Regression" Technometrics 23, 21-26.

———, and Smith, H. (1981) "Applied Regression Analysis, Wiley, New York.

Dutter, R. (1977), "Numerical Solution of Robust Regression Problems: Computational Aspects, a Comparacion," Journal of Statistical Computation and Simulation, 5, 207-238.

———, and Guttman (1979), "On Estimation in the linear Model when Spurios Observations are Present a Bayesian Approach" Commun. Statist. theor. Meth., A 8 (7; 61 - 635).

Edgeworth. F. Y. (1887), "On Discordant Observations", Philosophical Magazine, 23 Ser. 5, 364-375

Efron, B. (1965), "the Convex Hull of a Random Set of Points", Biometrika, 52, 331-343

Ellenberg, J. H. (1973) "the Joint Distribution of the Standardized Least Squares Residuals From a General linear Regression", Journal of the American Statistical Association, 68, 941-963.

———(1976) "testing for a Single Outlier From a General linear Regression" Biometrics 32, 637-645.

Freeman, P.R. (1981) "On the number of Outliers in data from a Linear Model", In Bayesian Statistics (eds. Bernardo, J.M.), 349 - 365. University Press. Valencia.

Furnival, G.M. and Wilson, R.W. Jr. (1974) Regressions by Leaps and Bounds. Technometrics. 16, 499-551.

Geisser, S. (1980). In Discussion of Box (1980)

Gentle, J.E. (1978), "Testing for Outliers in Linear Regression" in Contributions to Surey Sampling and Applied Statics in Honor of H.O. Hartley, ed. H.A. David, New York: Academic Press.

Gentleman, J.F. (1980), "Finding the K Most Likely Outliers in two-Way tables" Technometrics 22, 591-600.

———, and Wilk, M.B. (1975a), "Detecting Outliers in a two-Way Table: I Statistical Behavior of Residuals" technometrics, 17, 1-14.

———, "Detecting Outliers. II. Suplementing the Direct Analysis of Residuals". Biometrics 31, 387-410.

Glaisher, J.W.L. (1873), "On the Rejection of Discordant Observations", Montly Notices of the Royal Astronomical Society, 23, 141-160.

Gould, B.A: Jr (1855) "On Peirce's criterion for the Rejection of Doubtful Observations with tables for facilitating its

Application", Astronomical Journal 6, 81-83.

Gracia-Medrano, L.E. (1984), "Aplicación de técnicas de Regresión Robusta"; tesis para obtener el título de Actuario, Facultad de Ciencias : UNAM

Green, P.J. and Silverman, B.W. (1979), "Constructing the Convex Hull of a Set of Points, in the Plane. the Computer Jornal 22, 262, 266.

Grubbs, F.E. (1969), "Procedures for Detecting Outlying Observations in Samples", Technometrics, 15, 385-404.

Guttman, I. (1973 a), "Care and Handling of Univariate or Multivariate Outliers in Detecting Spuriosty. A Bayesian Approach", Technometrics, 15, 723-738.

———, and Dutter, R, and Freeman, P.R. (1978). Care and Handling of Univariate Outliers in the General Linear Model to detect Spuriosity : A Bayesian Approach", Technometrics 20, 187-193.

———, and Dutter, R., (1976), "Procedures for Investigating Outliers when Estimating in the General Univariate Linear Situation Nonfull Rank Case"; Communications in Statistics A5, 819-835.

———, and Kshatri, C.G., (1975), "A Bayesian Approach to Some Problems Involving the detection of Spuriosity"; in Applied Statistics Symposium, ed. R.P. Gupta, Amsterdam : North. Holland, 111-146.

———, and Smith, D.F. (1969), "Investigation of Rules for Dealing with Outliers in Small Samples from the Normal Distribution: I: Estimation of the Mean" technometrics, 11, 527 - 550.

Hawkins, D.M. (1979), "Fractiles of an Extended Multiple Outliers test", Journal of Statical Computation and Simulation, 8, 227 - 236.

———, (1980), "Identification of Outliers", London: Chapman and Hall.

Hewett, J. and Bulgren, W. G. (1971), "Inequalities for Some Multivariate F- Distributions with Applications", technometrics, 13, 397 - 402.

Hoaglin, D.C., and Welsch, R. (1978), "the hat Matrix in Regression and ANDVA", the American Statist. an 32, 17-22.

Hocking, R.R. (1983), "Developments in linear Regression, Methodology: 1859 - 1982", technometrics, 25, 219-249.

Hogg, R.V. (1979), "Statistical Robustness: One View of its - Use in Applications today", the American Statistical 33, 108 - 115.

———, (1979b) "An Introduction to Robust Estimation" Robustness in Statistics, Academic Press.

Holland, P.W. and Welsch, R.E. (1977)", Robust Regression using iteratively reweighted least squares" Communications in Statistics, A6,

813 - 828.

Huber, P.J. (1972), "the 1972 Wald Lecture Robust Statistics : A Review" Annals of Mathematical Statistics, 35, 73-101.

————, (1981), Robust Statistics, New York : John Wiley.

Irwin, J.O. (1925), "On a Criterion for the Rejection of Outlying Observations", Biometrika, 17, 238-250.

Jaeckel, L.A. (1972) "Estimating Regression Coefficients by minimizing the Dispersion of the Residual", Ann. Math. Statistics, 43, 1449 - 1468.

John, J.A. (1978), "Outliers in Factorial Experiments", Applied. - Statistics, 27, 111-119.

————, and Draper, N. R. (1978), "On testing for two Outliers One Outlier in two-way tables", technometrics 20, 69-78.

————, and Prescott (1975), "Critical values of a test Detect Outliers in Factorial Experiments," Applied Statistics 24, 56-59.

Johnson, B.A. and Hurt, H.H. (1979), "Performance characteristics for Certain tests to Detect Outliers" Proceedings of the Statistical Computing Section, American Statistical Association 247-249.

Joshi, P.C. (1972) "Some Slippage tests of Mean for a Simple Outlier in Linear Regression", Biometrika, 59, 109-120.

Kitagawa, G. (1979), "On the use of AIC. for Detection of Outliers", Technometrics 21, 193-199.

Krasker, W.S., and Welch, R.E. (1982), "Efficient Bounded-Influence Regression Estimation", Journal of the American Statistical Association, 595-609.

Lindley, D.V., and Smith, A.J.M. (1972), "A Bayes Estimates for the Linear Model (With discussion). J. R. Statist Soc. B., 34, 1-41.

Luendberger, D.G. (1973) Introduction to Linear and non linear Programming. Addison-Wesley. Menlo Park California. 148-155.

Lund, R.E. (1975) "tables for an Approximate test for Outliers in Linear Regression", technometries, 17, 473-476.

Mickey, M.R., Dunn, O.J., and Clark, V. (1967), "Note on the Use of Stepwise Regression, in Detecting Outliers", Computers and Biomedical Research, 1, 105-111.

Miyashita, H. and Newbold, P. (1983), "On the Sensivity to Non-Normality of a test of Outliers in Linear Models", Comman Statist.-theor. Meth., 12(12), 1413-1419.

Montgomery, D.C. and Peck, E.A. (1982), "Introduction to Linear Regression Analysis", New York: John Wiley.

Neave, H.R. (1978) "Statistics tables", Allen and Unwin. London.

Newcomb, S. (1886) "A Generalized theory of the Combination of

Observations to obtain the Best Result", American Journal of Mathematics, 8, 343-366.

Neyman, J. and Scott, E.L. (1971), "Outlier Proneness of phenomena and of Related Distributions", in Optimizing Methods in Statistics, ed. J.S. Rustagi New York: Academic Press.

Palacios, R.B. y Castañón, V.I., "Observaciones Discordantes en Análisis de Regresión: Detección y tratamiento", tesis para obtener título de Actuario, Facultad de Ciencias, UNAM.

Pearson, E.S., and Chandra Sekar C. (1936), the Efficiency of Statistical tools and Criterion for the Rejection of Outlying Observations", Biometrika 28, 308-320.

Peirce, B. (1852) "Criterion for the Rejection of Douftful Observations", Astronomical Journal. 2, 161-163.

Pettit, L.I., and Smith, A.F.M. (1983), "Outliers and Influential Observations in Linear Models", Second Valencia International Meetting on Bayesian Statistics.

Plackett, R.L. (1952), "Some theorems in Least Saveres", Biometrika 37, 149-157.

Prescott, P. (1975), "An Approximate test for Outliers in Linear Models" technometrics 17, 129-132.

Rosner, B. (1975), "On the Detection of Many Outliers" technometrics 17, 221-227.

Schweder, T. (1976), "Some Optimal Methods to Detect Structural Shift or Outliers in Regression" Journal of the American Statistical Association 68, 872–879.

Snedecor, G.W. and Cochan, W.G. (1967), "Statistical Methods" (6 h ed).,) Iowa State University Press.

Srikantan, K.S. (1961) "Testing for a single Outlier in a Regression Model", Sankya, Ser. A, 251–260.

Stefansky, W. (1971), "Rejecting Outliers by Maximum Normed Residual", the Annals of Mathematical of Statistics 42, 35–45

———, (1972). "Rejection Outliers in Factorial Designs", technometries, 14, 469–479.

Stigler, S.M. (1973), "Simon Newcomb, Percy Danniel and the History of Robust Estimation (1885–1920)", Journal of the American Statistical Association, 68, 872–879.

Stone, E.J. (1868) "On the Rejection of Discordant Observations" Monthly Notices of the Royal Astronomical Society, 34, 9–15.

Tiao, G.C. and Box, G.E.P. (1974) "Some comments on Bayes Estimators", in S.E. Fienberg and A. Zellner (eds.) Studies in Bayesian Econometries and Statistics: North-Holland

Tietjen, G.L., Moore, R.L. and Beckman, R.J. (1973), "testing for a Outlier in Simple Linear Regression": technometries 15, 717–721.

Weisberg, S. (1980), Applied Linear Regression" New York: John Wiley.