UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Tesis de Licenciatura

M-estimadores en Modelos de Regresión no Lineales con Respuestas Faltantes

Mariela A. Fiorenzo

Directora: Dra. Ana M. Bianco

19 de Diciembre de 2012

Dedico este trabajo a mi sobrina Lucía,
quien hace más de cuatro años
se ha convertido en mi Universo.

M-estimadores en Modelos de Regresión no Lineales con Respuestas Faltantes

En los modelos no lineales observamos una muestra aleatoria de n observaciones $(y_i, \mathbf{x}_i) \in \mathbb{R}^{p+1}$ independientes e idénticamente distribuidas ($i.i.d.$), con $y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^p$, siendo

$$y_i = g(\mathbf{x}_i, \boldsymbol{\theta}) + \varepsilon_i, \qquad 1 \leq i \leq n,$$

donde los errores ε_i son variables $i.i.d.$ e independientes de $\mathbf{x}_i$ con $E(\varepsilon_i) = 0$ y $Var(\varepsilon_i) = \sigma^2$, y g es una función conocida salvo por un vector de parámetros desconocido $\boldsymbol{\theta}$.

En esta tesis, estamos interesados en estimar dicho parámetro cuando existen respuestas faltantes en nuestro conjunto de datos. De esta forma, asumimos que observamos una muestra $(y_i, \mathbf{x}_i, \delta_i)$, $1 \leq i \leq n$, en la que $\delta_i = 1$ si y_i es observada y $\delta_i = 0$ si no lo es. Asumiremos que la variable de respuesta y presenta observaciones faltantes de forma aleatoria (missing at random, MAR), es decir, dado $\mathbf{x}$, δ e y son condicionalmente independientes $P(\delta = 1|(y, \mathbf{x})) = P(\delta = 1|\mathbf{x}) = p(\mathbf{x})$.

Dado que las estimaciones mediante los métodos clásicos, como el de mínimos cuadrados, son sensibles ante la presencia de datos atípicos, el objetivo de esta tesis es estudiar algunas propiedades de una familia de M–estimadores del parámetro $\boldsymbol{\theta}$ para el caso en que existen valores faltantes en la variable de respuesta, utilizando como estimador incial el LMS-estimador computado mediante el algoritmo propuesto por Stromberg (1993). Se prueba la Fisher-consistencia de dicho estimador y se deduce su función de influencia. Mediante un estudio de simulación comparamos su performance con la de los estimadores clásicos. Ilustramos su comportamiento a través del análisis de un conjunto de datos reales.

Palabras Claves: Modelos no lineales; Respuestas faltantes; Robustez; LMS; Fisher–consistencia.

Agradecimientos

Durante estos largos años de carrera muchas personas me han acompañado y hoy, en la recta final, quiero agradecerles el haber estado a mi lado.

Muy especialmente, agradezco a mis papás, a mi hermana y a toda mi familia. Cada uno de ellos, a su modo, no me dejaron nunca bajar los brazos y siempre confiaron en que lo lograría, a pesar de los tropiezos.

Quiero dedicar un párrafo aparte a mi mamá quien fue mi gran sostén. Gracias por no hacerme ver el mundo tan de color rosa pero tampoco tan negro como muchas veces me sucedió a lo largo de estos diez años.

Agradezco a mis amigos, por escucharme hablar de matemáticas aunque no entendieran nada y por saber comprender mis ausencias debido al estudio.

A mis amigos y compañeros de la facu, por tantas horas compartidas, por haberme regalado su tiempo para explicarme las cosas que yo no entendía y por confiar en mí al dejarme explicarles las que entendía yo.

A mis compañeros de trabajo, por su apoyo diario en esta última etapa, por tantas risas y por los momentos divertidos que pasamos, logrando alejar de mí los nervios y la ansiedad por la llegada de este tramo final.

A mi directora, Ana, a quien le debo el ser matemática ya que con sus excelentes clases de Probabilidades y Estadística para computadores despertó mi verdadera vocación. Gracias por el esfuerzo y las horas dedicadas para hacer de esta tesis un verdadero trabajo de investigación.

Finalmente, agradezco a Dios por no haberme hecho este camino tan fácil, ya que así comprendí que uno de los valores más importantes de la vida es la Perseverancia.

"Caminante no hay camino... se hace camino al andar."

Índice general

Capítulo 1

Introducción

Uno de los objetivos de la Estadística es encontrar las relaciones, si existen, dentro de un conjunto de variables cuando al menos una de ellas es aleatoria, siendo estas sujetas a posibles errores de medición. En los problemas de regresión una de las variables, usualmente llamada de *respuesta* o variable dependiente es de particular interés y se la nota y. Las otras variables $x_1, x_2, \ldots, x_p$ llamadas *regresoras* o variables independientes son aquellas que explican el comportamiento de y. A menudo, los investigadores se encuentran con expresiones matemáticas que relacionan la variable de respuesta y las variables regresoras mediante un modelo no lineal en los parámetros que podría expresarse de la siguiente manera:

$$y_i = g(\mathbf{x}_i, \boldsymbol{\theta}) + \varepsilon_i \qquad 1 \leq i \leq n\,, \tag{1.1}$$

donde $(y_i, \mathbf{x}_i)$ son vectores independientes e idénticamente distribuidos ($i.i.d.$), con $y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^p$, ε_i variables $i.i.d.$ e independientes de $\mathbf{x}_i$ y g conocida, excepto por el vector de parámetros $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^s$ que debe ser estimado.

A continuación ilustraremos con dos modelos no lineales que serán utilizados más adelante en este trabajo. En ambos casos, notamos $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$.

Modelo Exponencial

$$y_i = \beta e^{\alpha \mathbf{x}_i} + \varepsilon_i \qquad 1 \leq i \leq n\,, \quad \mathbf{x}_i \in \mathbb{R} \tag{1.2}$$

Este modelo se utiliza con frecuencia para describir crecimientos poblacionales, difusiones de epidemias, etc.

Modelo de Michaelis-Menten

Este modelo describe la cinética de las enzimas, relaciona la velocidad inicial de una reacción enzimática con la concentración del sustrato $\mathbf{x}$ a través de la siguiente ecuación

$$y_i = \frac{\alpha \mathbf{x}_i}{e^{\beta} + \mathbf{x}_i} + \varepsilon_i \qquad 1 \leq i \leq n\,, \quad \mathbf{x}_i \in \mathbb{R} \tag{1.3}$$

El método clásico para estimar el parámetro $\boldsymbol{\theta}$ es el de *mínimos cuadrados*, propuesto por Legendre (1805), cuya solución viene dada por

$$\hat{\boldsymbol{\theta}} = \arg \min_{\mathbf{t} \in \Theta} \frac{1}{n} \sum_{i=1}^{n} [y_i - g(\mathbf{x}_i, \mathbf{t})]^2 \,.$$

Bajo ciertas condiciones de regularidad, este estimador tiene características óptimas, incluso coincide con el estimador de máxima verosimilitud si suponemos que $\varepsilon_i \sim N(0, \sigma^2)$. Sin embargo, si estas condiciones no se satisfacen, el estimador de mínimos cuadrados es muy sensible a datos atípicos. Por esta razón, es necesario considerar estimadores robustos, es decir, estimadores poco sensibles a outliers, que a su vez sean altamente eficientes cuando los datos sean normales.

Para los modelos de regresión lineal existen varios métodos para obtener estimadores robustos, entre otros, podemos mencionar los *M-estimadores* propuestos por Huber (1973), los *LMS-estimadores* (Least Median of Squares) y *LTS-estimadores* (Least Trimmed Squares) propuestos por Rousseeuw (1984) y (1985), respectivamente, *MM-estimadores* propuestos por Yohai (1987), los *S-estimadores* propuestos por Rousseeuw y Yohai (1984) y τ*-estimadores* propuestos por Yohai y Zamar (1988).

En el caso de regresión no lineal, también se propusieron algunos estimadores robustos. Tiede y Pagano (1979) proponen un algoritmo para el cálculo de M-estimadores aplicados al análisis de radioinmunoensayos, Fraiman (1983) considera estimadores de influencia acotada, Carroll y Ruppert (1987) trabajan con métodos robustos para transformaciones no lineales de los datos. Stromberg (1993) introduce algoritmos para el cálculo de MM-estimadores para regresión no lineal y Tabatabai y Argyros (1993) extienden el τ-estimador al caso no lineal proponiendo además un algoritmo para su cálculo. Markatou y Manos (1996) consideran test de hipótesis en regresión no lineal basados en M-estimadores y Mukherjee (1996) propone estimadores de mínima distancia. Más recientemente, Stromberg, Hossjer y Hawkins (2000) introducen el *LTD-estimador* (Least Trimmed Difference) con la particularidad de que la distribución del modelo puede ser asimétrica y Sakata y White (2001) trabajan con S-estimadores para modelos de regresión no lineal con observaciones dependientes. Por último, Fasano, Maronna, Sued y Yohai (2012) tratan el problema de la continuidad débil, la Fisher–consistencia y diferenciabilidad de los funcionales asociados a los estimadores de alto punto de ruptura tanto en el caso lineal como no lineal, incluyendo S- y MM-estimadores.

En el contexto de regresión no lineal con respuestas faltantes, Müller (2009) estudia el problema de estimar, mediante un estimador completamente imputado, la esperanza marginal de una función de la variable de respuesta bajo el supuesto de que las respuestas son

faltantes al azar, MAR. Con el interés de estimar la distribución marginal de la respuesta en este mismo contexto, Sued y Yohai (2012) proponen un procedimiento que permite estimar en forma consistente cualquier funcional débilmente continuo de la distribución de las respuestas, que incluye la mediana o M-estimadores y que se basan en la utilización de un estimador robusto inicial del parámetro de regresión no lineal.

En cuanto a métodos numéricos para el cálculo de estimadores robustos en modelos de regresión no lineal, Huber (1981) trabaja con M-estimadores y Stromberg (1993) desarrolla un algoritmo para estimadores de alto punto de ruptura como el LMS, que es el que se ha implementado en nuestro estudio de Monte Carlo para el cómputo del estimador inicial.

Estos métodos fueron diseñados para conjuntos de datos completos, sin embargo, en la práctica podemos encontrarnos con datos faltantes. La ausencia de variables de respuesta puede deberse a que, en ciertas ocasiones, medir la variable y es muy costoso o puede que la información se pierda por alguna falla en la recolección de los datos.

En este trabajo, nos enfocaremos en la estimación robusta del parámetro de regresión de un modelo dado por (1.1) cuando faltan observaciones en la variable de respuesta y, pero las covariables $\mathbf{x}$ son completamente observadas. Luego, nuestro conjunto de datos queda definido a partir de $(y_i, \mathbf{x}_i, \delta_i)$, $1 \leq i \leq n$ donde $\delta_i = 1$ si y_i es observada y $\delta_i = 0$ si no lo es. Para esto asumiremos que y presenta observaciones faltantes de forma aleatoria (missing at random, MAR), es decir, dado $\mathbf{x}$, δ e y son condicionalmente independientes $P(\delta = 1|(y, \mathbf{x})) = P(\delta = 1|\mathbf{x}) = p(\mathbf{x})$.

Ya que las estimaciones mediante el método de mínimos cuadrados son sensibles a la presencia de datos atípicos, el objetivo de esta tesis es estudiar algunas propiedades de una familia de M–estimadores del parámetro $\boldsymbol{\theta}$ para el caso en que existen valores faltantes en la variable de respuesta, utilizando como estimador incial el LMS-estimador computado mediante el algoritmo propuesto por Stromberg (1993).

La tesis está organizada como se describe a continuación. En el Capítulo 2, realizamos un revisión sobre los modelos de regresión no lineal y los métodos de estimación que se pueden aplicar en los mismos. En el Capítulo 3, introducimos tanto nociones básicas de la teoría de robustez como su aplicación a los modelos de regresión no lineales. A su vez, describimos el algoritmo mediante el cual se obtiene el LMS-estimador, tal como fue propuesto por Stromberg (1993). En el Capítulo 4, presentamos el problema de la estimación robusta en modelos de regresión no lineales con valores faltantes en la variable de respuesta, proponemos un estimador robusto y analizamos su Fisher-consistencia. En el Capítulo 5, realizamos un estudio de Monte Carlo con el fin de comparar la performance de los estimadores robustos obtenidos con la de los estimadores clásicos de mínimos cuadrados para muestras con diferentes escenarios de pérdida de datos, contaminadas y sin contaminar. Finalmente, en el Capítulo 6, analizamos un ejemplo con datos reales.

Capítulo 2

Modelo de regresión no lineal y su estimación

La regresión lineal es un método poderoso para analizar datos descriptos por un modelo que sea lineal en los parámetros. A menudo, sin embargo, los investigadores se encuentran con expresiones matemáticas que relacionan la variable de respuesta y con las variables regresoras $\mathbf{x}$, mediante modelos que resultan no lineales en los parámetros. En estos casos, las técnicas de regresión lineal deben ser extendidas, lo cual introduce una complejidad considerable.

Un modelo de regresión no lineal puede ser escrito de la siguiente manera

$$y_i = g(\mathbf{x}_i, \boldsymbol{\theta}) + \varepsilon_i, \quad 1 \leq i \leq n\,, \tag{2.1}$$

donde ε mide las fluctuaciones o errores de medición, n es el tamaño de la muestra y g es una función completamente conocida excepto por $\boldsymbol{\theta}$ que es un vector de parámetros desconocidos que necesita ser estimado. Utilizando la función g podemos predecir y a partir de $\mathbf{x}$, las cuales pueden ser aleatorias o fijas. Como los parámetros suelen tener interpretaciones físicas, uno de los principales objetivos de la investigación es estimar dichos parámetros lo más precisamente posible.

Los modelos no lineales tienden a ser utilizados, o bien cuando son sugeridos mediante consideraciones teóricas, o bien para construir dentro de un modelo ciertos comportamientos no lineales ya conocidos. En ocasiones, aún cuando una aproximación lineal pudiera funcionar bien, el investigador podría preferir un ajuste basado en un modelo no lineal a fin de mantener una clara interpretación de los parámetros. Los mismos han sido aplicados a un gran rango de situaciones, incluso en poblaciones finitas. Seber y Wild (1989) dan una completa revisión sobre estos temas, que abordamos a continuación.

2.1. Modelo con Variables Regresoras Fijas

2.1.1. Variables Regresoras Fijas

Supongamos que μ es el tamaño esperado de un organismo en el tiempo x. Sin embargo, debido a fluctuaciones y posibles errores de medición, el tamaño efectivo es y, de modo que, $E(y) = \mu$. Luego el modelo sería

$$y = g(x, \boldsymbol{\theta}) + \varepsilon\,, \tag{2.2}$$

donde $E(\varepsilon) = E(y - \mu) = 0$. Si el tiempo lo medimos con exactitud de modo que su varianza es esencialmente cero, o despreciable comparado con $var(y)$, luego uno podría tratar a x como fija en vez de aleatoria.

2.1.2. Variables Regresoras Condicionales

Existen dos variaciones al enfoque anterior en las cuales x es, de hecho, aleatoria pero puede ser tratada como si fuera fija. La primera ocurre cuando g es una relación teórica y x es aleatoria, pero medida con exactitud, i.e. $x = x_0$. Un modelo adecuado podría ser

$$E(y|x = x_0) = g(x_0, \boldsymbol{\theta})\,,$$

y (2.2) podría ser interpretado como un modelo de regresión condicional, condicional sobre el valor observado de x.

La segunda variación ocurre cuando g es elegida empíricamente para modelar la relación entre y y el valor medido de x (en vez del verdadero), aún cuando x es medida con error. En este caso (2.2) es usado para modelar la distribución condicional de y, dado el valor medido de x.

Hay una tercera situación posible, a saber, g es un modelo teórico conectando a y con el verdadero valor de x, cuando x es medido con error. En este caso el verdadero valor de x es desconocido y se necesita de una nueva propuesta, como describimos a continuación.

2.2. Modelo con Variables Regresoras Aleatorias con Errores

2.2.1. Relaciones Funcionales

Supongamos que existe una relación funcional exacta

$$\mu = g(\xi, \boldsymbol{\theta}) \tag{2.3}$$

entre las realizaciones ξ y μ de dos variables. Sin embargo, ambas variables son medidas con

error, de modo que lo que observamos es

$$y = \mu + \varepsilon \quad \text{y} \quad x = \xi + \delta\,,$$

donde $E(\varepsilon) = E(\delta) = 0$. Luego

$$y = g(\xi, \boldsymbol{\theta}) + \varepsilon = g(x - \delta, \boldsymbol{\theta}) + \varepsilon. \tag{2.4}$$

Por el Teorema del Valor Medio,

$$g(x - \delta, \boldsymbol{\theta}) = g(x, \boldsymbol{\theta}) - \delta \dot{g}(\tilde{x}, \boldsymbol{\theta}), \tag{2.5}$$

donde $\dot{g}$ es la derivada de g, y $\tilde{x}$ se encuentra entre x y $x - \delta$. Sustituyendo (2.5) en (2.4) tenemos que

$$y = g(x, \boldsymbol{\theta}) - \delta \dot{g}(\tilde{x}, \boldsymbol{\theta}) + \varepsilon = g(x, \boldsymbol{\theta}) + \varepsilon^*\,, \tag{2.6}$$

Este modelo no es el mismo que el (2.2), dado que x ahora es considerada aleatoria y, en general, $E(\varepsilon^*) \neq 0$ y, además, no es independiente de x. Si lo analizáramos de la misma manera que (2.2) utilizando mínimos cuadrados, obtendríamos sesgos.

2.2.2. Relaciones Estructurales

Un tipo de modelo diferente es obtenido si la relación (2.3) es una relación entre variables aleatorias (digamos, u y v) en lugar de sus realizaciones. Luego tenemos lo que llamamos una relación *estructural*

$$v = g(u, \boldsymbol{\theta})\,,$$

con $y = v + \varepsilon$ y $x = u + \delta$. Argumentando igual que en (2.5) obtenemos un modelo de la misma forma que (2.6), pero con una estructura diferente para ε^*. Para el caso lineal $v = \alpha + \beta u$ se ha comprobado, que a pesar de su simpleza, existen problemas de identificabilidad al querer estimar parámetros desconocidos.

2.3. Variables Regresoras Controladas con Errores

Ahora estudiaremos un tercer tipo de modelo, comúnmente utilizado en experimentos de laboratorio. Comenzamos con la relación estructural $v = g(u, \boldsymbol{\theta})$ y tratamos de establecer u en el valor puntual x_0. Sin embargo, x_0 no es alcanzado exactamente y en su lugar tenemos $u = x_0 + \delta$, donde $E(\delta) = 0$ y u es desconocida. Para un modelo general,

$$y = v + \varepsilon = g(u, \boldsymbol{\theta}) + \varepsilon = g(\mathbf{x}_0 + \delta, \boldsymbol{\theta}) + \varepsilon\,,$$

el cual puede ser expandido una vez más mediante el Teorema del Valor Medio

$$y = g(x_0, \boldsymbol{\theta}) + \delta \dot{g}(\tilde{x}, \boldsymbol{\theta}) + \varepsilon = g(x_0, \boldsymbol{\theta}) + \tilde{\varepsilon},$$

con $\tilde{x}$ entre x_0 y $x_0 + \delta$. En general $E(\tilde{\varepsilon}) \neq 0$, pero si δ es tan chico que $\tilde{x} \approx x_0$, entonces $E(\tilde{\varepsilon}) \approx 0$. En este caso podemos tratar al modelo como un modelo con variables regresoras fijas.

2.4. Modelos con Errores Autocorrelacionados

En muchos casos en que los modelos de regresión no lineal han sido ajustados para un conjunto de datos recolectados secuencialmente en el tiempo, gráficos de estos datos revelan largas rachas de residuos positivos y largas rachas de residuos negativos. Esto puede ser debido a lo inadecuado del modelo postulado para $E(y|x)$, o puede ser causado por un alto grado de correlación entre sucesivos términos del error ε_i. Una simple estructura de autocorrelación que a veces es aplicada a los datos recolectados en intervalos de tiempo equiespaciados viene dada por un proceso autoregresivo de orden 1 [AR(1)], digamos

$$\varepsilon_i = \rho \varepsilon_{i-1} + a_i,$$

donde los a_i son no correlacionados, $E(a_i) = 0$, $var(a_i) = \sigma_a^2$, y $|\rho| < 1$. Bajo tal estructura

$$corr[\varepsilon_i, \varepsilon_j] = \rho^{|i-j|},$$

de modo que la correlación entre ε_i y ε_j decrece exponencialmente a medida que crece la distancia entre los tiempos en que y_i e y_j fueron medidas.

2.5. Estimadores de Mínimos Cuadrados

2.5.1. Mínimos Cuadrados no Lineales

Supongamos que tenemos n observaciones $(\mathbf{x}_i, y_i)$, $1 \leq i \leq n$ que satisfacen una relación no lineal tal que

$$y_i = g(\mathbf{x}_i, \boldsymbol{\theta}^*) + \varepsilon_i \quad 1 \leq i \leq n\,, \tag{2.7}$$

donde $E[\varepsilon_i] = 0$, $\mathbf{x}_i \in \mathbb{R}^p$ y el valor verdadero $\boldsymbol{\theta}^*$ de $\boldsymbol{\theta}$ se sabe que pertenece a Θ, un subconjunto de $\mathbb{R}^s$. El estimador de mínimos cuadrados de $\boldsymbol{\theta}^*$, $\hat{\boldsymbol{\theta}}$, minimiza la "suma de cuadrados residual"

$$\mathbf{S}(\boldsymbol{\theta}) = \sum_{i=1}^{n} [y_i - g(\mathbf{x}_i, \boldsymbol{\theta})]^2 \tag{2.8}$$

respecto de $\boldsymbol{\theta} \in \Theta$.

Asumiendo que ε_i son independientes e idénticamente distribuidos con varianza σ^2, bajo ciertos supuestos de regularidad, $\hat{\boldsymbol{\theta}}$ y $s^2 = \mathbf{S}(\hat{\boldsymbol{\theta}})/(n-p)$ son estimadores consistentes de $\boldsymbol{\theta}^*$ y

σ^2 respectivamente. Con algunas condiciones de regularidad adicionales, la distribución de $\hat{\boldsymbol{\theta}}$ también es asintóticamente normal para $n \to \infty$.

Si además suponemos que ε_i son normalmente distribuidos, entonces $\hat{\boldsymbol{\theta}}$ coincide con el estimador de máxima verosimilitud.

Cuando $g(\mathbf{x}_i, \boldsymbol{\theta})$ es diferenciable respecto a $\boldsymbol{\theta}$, y $\hat{\boldsymbol{\theta}}$ pertenece al interior de Θ, $\hat{\boldsymbol{\theta}}$ cumple

$$\left.\frac{\partial \mathbf{S}(\boldsymbol{\theta})}{\partial \theta_r}\right|_{\hat{\boldsymbol{\theta}}} = 0, \quad 1 \leq r \leq p\,. \tag{2.9}$$

Usaremos la notación $g_i(\boldsymbol{\theta}) = g(\mathbf{x}_i, \boldsymbol{\theta})$,

$$\mathbf{g}(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), g_2(\boldsymbol{\theta}), \ldots, g_n(\boldsymbol{\theta}))'\,,$$

y

$$\mathbf{G}_\cdot(\boldsymbol{\theta}) = \frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \left[\left(\frac{\partial g_i(\boldsymbol{\theta})}{\partial \theta_j}\right)\right].$$

Llamaremos

$$\mathbf{G}_\cdot = \mathbf{G}_\cdot(\boldsymbol{\theta}^*)\,\text{y}\,\hat{\mathbf{G}}_\cdot = \mathbf{G}_\cdot(\hat{\boldsymbol{\theta}})\,,$$

donde un sólo punto representa la derivada primera y dos puntos la derivada segunda. Usando la notación anterior podemos escribir

$$\mathbf{S}(\boldsymbol{\theta}) = [\mathbf{y} - \mathbf{g}(\boldsymbol{\theta})]'[\mathbf{y} - \mathbf{g}(\boldsymbol{\theta})] = ||\mathbf{y} - \mathbf{g}(\boldsymbol{\theta})||^2.$$

Luego la ecuación (2.9) se puede escribir

$$\sum_{i=1}^{n} \{y_i - g_i(\boldsymbol{\theta})\} \left.\frac{\partial g_i(\boldsymbol{\theta})}{\partial \theta_r}\right|_{\hat{\boldsymbol{\theta}}} = 0, \quad 1 \leq r \leq p,$$

o

$$\mathbf{0} = \hat{\mathbf{G}}_\cdot'\{\mathbf{y} - \mathbf{g}(\hat{\boldsymbol{\theta}})\} = \hat{\mathbf{G}}_\cdot'\hat{\boldsymbol{\varepsilon}}\,. \tag{2.10}$$

Si $\hat{\mathbf{P}}_{\mathbf{G}} = \hat{\mathbf{G}}_\cdot(\hat{\mathbf{G}}_\cdot'\hat{\mathbf{G}}_\cdot)^{-\mathbf{1}}\hat{\mathbf{G}}_\cdot'$, la matriz idempotente proyecta ortogonalmente $\mathbb{R}^p$ sobre $\mathfrak{R}[\hat{\mathbf{G}}_\cdot]$ entonces (2.10) puede ser escrito como

$$\hat{\mathbf{P}}_{\mathbf{G}}\hat{\boldsymbol{\varepsilon}} = \mathbf{0}\,.$$

Las ecuaciones (2.10) se llaman *ecuaciones normales* para el modelo no lineal. Para la mayoría de los modelos no lineales estas no pueden ser resueltas analíticamente y es necesario recurrir a métodos numéricos iterativos, como veremos a continuación.

2.5.2. Aproximación Lineal

Ahora introduciremos un conjunto de resultados en forma heurística. Comenzaremos notando que en un entorno pequeño de $\boldsymbol{\theta}^*$, el valor verdadero del parámetro $\boldsymbol{\theta}$, tenemos la expansión lineal de Taylor dada por

$$g_i(\boldsymbol{\theta}) \approx g_i(\boldsymbol{\theta}^*) + \sum_{r=1}^{p} \left.\frac{\partial g_i}{\partial \theta_r}\right|_{\boldsymbol{\theta}^*} (\theta_r - \theta_r^*),$$

o

$$\mathbf{g}(\boldsymbol{\theta}) \approx \mathbf{g}(\boldsymbol{\theta}^*) + \mathbf{G}.(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \tag{2.11}$$

donde $\mathbf{G}. = \mathbf{G}.(\boldsymbol{\theta}^*)$. Luego,

$$\mathbf{S}(\boldsymbol{\theta}) = ||\mathbf{y} - \mathbf{g}(\boldsymbol{\theta})||^2 \approx ||\mathbf{y} - \mathbf{g}(\boldsymbol{\theta}^*) - \mathbf{G}.(\boldsymbol{\theta} - \boldsymbol{\theta}^*)||^2 = ||\mathbf{z} - \mathbf{G}.\boldsymbol{\beta}||^2, \tag{2.12}$$

donde, $\mathbf{z} = \mathbf{y} - \mathbf{g}(\boldsymbol{\theta}^*) = \boldsymbol{\varepsilon}$ y $\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$. Por las propiedades bien conocidas en el modelo lineal, $\mathbf{S}(\boldsymbol{\theta})$ es minimizada cuando $\boldsymbol{\beta}$ está dado por

$$\hat{\boldsymbol{\beta}} = (\mathbf{G}.'\mathbf{G}.)^{-\mathbf{1}}\mathbf{G}.'\mathbf{z}$$

Para n grande, bajo ciertas condiciones de regularidad, es casi seguro que $\hat{\boldsymbol{\theta}}$ pertenecerá a un pequeño entorno de $\boldsymbol{\theta}^*$, por lo tanto

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \approx (\mathbf{G}.'\mathbf{G}.)^{-1}\mathbf{G}.'\boldsymbol{\varepsilon}\,. \tag{2.13}$$

Más aún, de (2.11) con $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$,

$$\mathbf{g}(\hat{\boldsymbol{\theta}}) - \mathbf{g}(\boldsymbol{\theta}^*) \approx \mathbf{G}.(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \approx \mathbf{G}.(\mathbf{G}.'\mathbf{G}.)^{-1}\mathbf{G}.'\boldsymbol{\varepsilon} = \mathbf{P}_{\mathbf{G}}\boldsymbol{\varepsilon} \tag{2.14}$$

y los residuos

$$\mathbf{y} - \mathbf{g}(\hat{\boldsymbol{\theta}}) \approx \mathbf{y} - \mathbf{g}(\boldsymbol{\theta}^*) - \mathbf{G}.(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \approx \boldsymbol{\varepsilon} - \mathbf{P}_{\mathbf{G}}\boldsymbol{\varepsilon} = (\mathbf{I}_{\mathbf{n}} - \mathbf{P}_{\mathbf{G}})\boldsymbol{\varepsilon}, \tag{2.15}$$

donde $\mathbf{P}_{\mathbf{G}} = \mathbf{G}.(\mathbf{G}.'\mathbf{G}.)^{-1}\mathbf{G}.'$ y $(\mathbf{I}_{\mathbf{n}} - \mathbf{P}_{\mathbf{G}})$ son simétricas e idempotentes.

Si como es habitual, definimos $s^2 = \mathbf{S}(\hat{\boldsymbol{\theta}})/(n-p)$, de (2.15) y (2.14) tenemos

$$(n-p)s^2 = \mathbf{S}(\hat{\boldsymbol{\theta}}) = ||\mathbf{y} - \mathbf{g}(\hat{\boldsymbol{\theta}})||^2 \approx ||(\mathbf{I}_{\mathbf{n}} - \mathbf{P}_{\mathbf{G}})\boldsymbol{\varepsilon}||^2 = \boldsymbol{\varepsilon}'(\mathbf{I}_{\mathbf{n}} - \mathbf{P}_{\mathbf{G}})\boldsymbol{\varepsilon}, \tag{2.16}$$

y

$$||\mathbf{g}(\hat{\boldsymbol{\theta}}) - \mathbf{g}(\boldsymbol{\theta}^*)||^2 \approx ||\mathbf{G}.(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)||^2 = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)'\mathbf{G}.'\mathbf{G}.(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \approx ||\mathbf{P}_{\mathbf{G}}\boldsymbol{\varepsilon}||^2 = \boldsymbol{\varepsilon}'\mathbf{P}_{\mathbf{G}}\boldsymbol{\varepsilon}. \tag{2.17}$$

Luego, usando (2.16) y (2.17), obtenemos

$$\mathbf{S}(\boldsymbol{\theta}^*) - \mathbf{S}(\hat{\boldsymbol{\theta}}) \approx \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}'(\mathbf{I}_{\mathbf{n}} - \mathbf{P}_{\mathbf{G}})\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'\mathbf{P}_{\mathbf{G}}\boldsymbol{\varepsilon} \approx (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)'\mathbf{G}.'\mathbf{G}.(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*). \tag{2.18}$$

Cuando $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I}_{\mathbf{n}})$, bajo condiciones apropiadas de regularidad, Seber y Wild (1989) establecen que para un n suficientemente grande, se cumplen las siguientes aproximaciones

(i) $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \approx Z \sim N_p(0, \sigma^2\mathbf{C}^{-1})$, donde $\mathbf{C} = \mathbf{G}.'\mathbf{G}. = \mathbf{G}.'(\boldsymbol{\theta}^*)\mathbf{G}.(\boldsymbol{\theta}^*)$

(ii) $(n-p)s^2/\sigma^2 \approx \boldsymbol{\varepsilon}'(\mathbf{I}_{\mathbf{n}} - \mathbf{P}_{\mathbf{G}})\boldsymbol{\varepsilon}/\sigma^2 \sim \chi^2_{n-p}$

(iii) $\hat{\boldsymbol{\theta}}$ es independiente de s^2

(iv) $$\frac{[\mathbf{S}(\boldsymbol{\theta}^*) - \mathbf{S}(\hat{\boldsymbol{\theta}})]/p}{\mathbf{S}(\hat{\boldsymbol{\theta}})/(n-p)} \approx \frac{\boldsymbol{\varepsilon}'\mathbf{P}_{\mathbf{G}}\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}'(\mathbf{I}_{\mathbf{n}} - \mathbf{P}_{\mathbf{G}})\boldsymbol{\varepsilon}}\frac{n-p}{p} \sim F_{p,n-p}$$

No se requiere de la normalidad de $\boldsymbol{\varepsilon}$ para demostrar (i) ya que (2.13) implica que $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ es asintóticamente una combinación lineal de los errores $\boldsymbol{\varepsilon}_i$, que son independientes e idénticamente distribuidos. Luego, una versión apropiada del teorema central de límite nos da (i).

Finalmente, usando (iv) y (2.18) tenemos, aproximadamente

$$\frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)' \mathbf{G}.' \mathbf{G}.(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)}{ps^2} \sim F_{p,n-p}.$$

Luego, $\mathbf{G}.$ juega el mismo rol que la matriz de diseño $\mathbf{X}$ en regresión lineal.

2.5.3. Métodos Numéricos

Supongamos que $\boldsymbol{\theta}^{(a)}$ es una aproximación al estimador de mínimos cuadrados $\hat{\boldsymbol{\theta}}$ de un modelo no lineal. Para $\boldsymbol{\theta}$ cerca de $\boldsymbol{\theta}^{(a)}$, utilizando la expansión lineal de Taylor

$$\mathbf{g}(\boldsymbol{\theta}) \approx \mathbf{g}(\boldsymbol{\theta}^{(a)}) + \mathbf{G}.^{(a)}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}), \tag{2.19}$$

donde $\mathbf{G}.^{(a)} = \mathbf{G}.(\boldsymbol{\theta}^{(a)})$. Aplicando esto al vector de residuos $\mathbf{r}(\boldsymbol{\theta})$, tenemos

$$\mathbf{r}(\boldsymbol{\theta}) = \mathbf{y} - \mathbf{g}(\boldsymbol{\theta}) \approx \mathbf{r}(\boldsymbol{\theta}^{(a)}) - \mathbf{G}.^{(a)}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}).$$

Sustituyendo en $\mathbf{S}(\boldsymbol{\theta}) = \mathbf{r}'(\boldsymbol{\theta})\mathbf{r}(\boldsymbol{\theta})$ obtenemos

$$\mathbf{S}(\boldsymbol{\theta}) \approx \mathbf{r}'(\boldsymbol{\theta}^{(a)})\mathbf{r}(\boldsymbol{\theta}^{(a)}) - 2\mathbf{r}'(\boldsymbol{\theta}^{(a)})\mathbf{G}.^{(a)}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}) + (\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)})'\mathbf{G}.^{(a)\prime}\mathbf{G}.^{(a)}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}). \tag{2.20}$$

El lado derecho es minimizado respecto a $\boldsymbol{\theta}$ cuando

$$\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)} = (\mathbf{G}.^{(a)\prime}\mathbf{G}.^{(a)})^{-1}\mathbf{G}.^{(a)\prime}\mathbf{r}(\boldsymbol{\theta}^{(a)}) = \boldsymbol{\delta}^{(a)}. \tag{2.21}$$

Esto sugiere un esquema iterativo para obtener $\hat{\boldsymbol{\theta}}$: si en el paso a obtenemos una aproximación $\boldsymbol{\theta}^{(a)}$, la aproximación del siguiente paso debería ser

$$\boldsymbol{\theta}^{(a+1)} = \boldsymbol{\theta}^{(a)} + \boldsymbol{\delta}^{(a)}. \tag{2.22}$$

La aproximación de $\mathbf{S}(\boldsymbol{\theta})$ por la cuadrática (2.20), y las resultantes fórmulas (2.21) y (2.22) son llamadas usualmente *método de Gauss-Newton*. El mismo es convergente, o sea, $\boldsymbol{\theta}^{(a)} \to \hat{\boldsymbol{\theta}}$ cuando $a \to \infty$ siempre que el punto inicial $\boldsymbol{\theta}^{(1)}$ esté suficientemente cerca de $\boldsymbol{\theta}^*$ y n sea suficientemente grande.

Una propuesta más general, que se puede aplicar a cualquier función suficientemente suave, es el *método de Newton*, en el cual $\mathbf{S}(\boldsymbol{\theta})$ es expandido directamente utilizando una

expansión de Taylor cuadrática. Sean

$$\mathbf{h}(\boldsymbol{\theta}) = \frac{\partial \mathbf{S}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

y

$$\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial^2 \mathbf{S}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}$$

el vector gradiente y la matriz Hessiana de $\mathbf{S}(\boldsymbol{\theta})$, respectivamente. Luego, tenemos la aproximación cuadrática

$$\mathbf{S}(\boldsymbol{\theta}) \approx \mathbf{S}(\boldsymbol{\theta}^{(a)}) + \mathbf{h}'(\boldsymbol{\theta}^{(a)})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)})'\mathbf{H}(\boldsymbol{\theta}^{(a)})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}), \tag{2.23}$$

que difiere de (2.20) sólo en que $\mathbf{H}(\boldsymbol{\theta}^{(a)})$ es aproximada por $2\mathbf{G}.'(\boldsymbol{\theta}^{(a)})\mathbf{G}.(\boldsymbol{\theta}^{(a)})$. Sin embargo, como

$$\frac{\partial^2 \mathbf{S}(\boldsymbol{\theta})}{\partial \theta_r\, \partial \theta_s} = 2\sum_{i=1}^{n}\left\{\frac{\partial g_i(\boldsymbol{\theta})}{\partial \theta_r}\frac{\partial g_i(\boldsymbol{\theta})}{\partial \theta_s} - [y_i - g_i(\boldsymbol{\theta})]\frac{\partial^2 g_i(\boldsymbol{\theta})}{\partial \theta_r\, \partial \theta_s}\right\}, \tag{2.24}$$

luego

$$E\left[\frac{\partial^2 \mathbf{S}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right] = 2\mathbf{G}.'(\boldsymbol{\theta})\mathbf{G}.(\boldsymbol{\theta}) \tag{2.25}$$

y $\mathbf{H}(\boldsymbol{\theta}^{(a)})$ es aproximado por su valor esperado en $\boldsymbol{\theta}^{(a)}$ en (2.20).

El mínimo de la función cuadrática (2.23) con respecto a $\boldsymbol{\theta}$ se obtiene cuando

$$\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)} = -[\mathbf{H}(\boldsymbol{\theta}^{(a)})]^{-1}\mathbf{h}(\boldsymbol{\theta}^{(a)}) = -[\mathbf{H}^{-1}\mathbf{h}]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(a)}}. \tag{2.26}$$

Este es el llamado *método de Newton* y el término de corrección $\boldsymbol{\delta}^{(a)}$ en (2.22) ahora dado por (2.26) es el llamado *Newton step*.

2.5.4. Mínimos Cuadrados Generalizados

Mencionaremos una generalización del método de mínimos cuadrados llamado método de *mínimos cuadrados generalizados* o *pesados* (GLS). La función a minimizar ahora es

$$\mathbf{S}(\boldsymbol{\theta}) = [\mathbf{y} - \mathbf{g}(\boldsymbol{\theta})]'\mathbf{V}^{-1}[\mathbf{y} - \mathbf{g}(\boldsymbol{\theta})],$$

donde $\mathbf{V}$ es una matriz conocida y definida positiva. El método de mínimos cuadrados ordinarios (OLS), que mencionamos previamente, es un caso particular de GLS tomando $\mathbf{V} = \mathbf{I_n}$. Denotemos por $\hat{\boldsymbol{\theta}}_{\mathrm{G}}$ al estimador de mínimos cuadrados generalizados que minimiza a $\mathbf{S}(\boldsymbol{\theta})$.

Sea $\mathbf{V} = \mathbf{U}'\mathbf{U}$ la descomposición de Cholesky de $\mathbf{V}$, donde $\mathbf{U}$ es una matriz triangular superior. Multiplicando el modelo no lineal por $\mathbf{R} = (\mathbf{U}')^{-1}$, obtenemos

$$\mathbf{z} = \mathbf{k}(\boldsymbol{\theta}) + \boldsymbol{\eta},$$

donde $\mathbf{z} = \mathbf{R}\mathbf{y}, \mathbf{k}(\boldsymbol{\theta}) = \mathbf{R}\mathbf{g}(\boldsymbol{\theta})$ y $\boldsymbol{\eta} = \mathbf{R}\boldsymbol{\varepsilon}$. Entonces $E[\boldsymbol{\eta}] = 0$ y $\mathcal{D}[\boldsymbol{\eta}] = \sigma^2\mathbf{R}\mathbf{V}\mathbf{R}' = \sigma^2\mathbf{I_n}$, la matriz de varianza-covarianza. Así es como nuestro modelo GLS original ha sido transformado en un modelo (OLS). Más aún,

$$\mathbf{S}(\boldsymbol{\theta}) = [\mathbf{y} - \mathbf{g}(\boldsymbol{\theta})]'\mathbf{V^{-1}}[\mathbf{y} - \mathbf{g}(\boldsymbol{\theta})] = [\mathbf{y} - \mathbf{g}(\boldsymbol{\theta})]'\mathbf{R}'\mathbf{R}[\mathbf{y} - \mathbf{g}(\boldsymbol{\theta})] = [\mathbf{z} - \mathbf{k}(\boldsymbol{\theta})]'[\mathbf{z} - \mathbf{k}(\boldsymbol{\theta})].$$

Por lo tanto, la suma de cuadrados GLS es la misma que la suma de cuadrados OLS para el modelo transformado, y $\hat{\boldsymbol{\theta}}_{\mathrm{G}}$ es el estimador OLS del mismo.

Sea $\mathbf{K}.(\boldsymbol{\theta}) = \partial\mathbf{k}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$ y $\hat{\mathbf{K}}. = \mathbf{K}.(\hat{\boldsymbol{\theta}})$. Entonces

$$\mathbf{K}.(\boldsymbol{\theta}) = \mathbf{R}\frac{\partial\mathbf{g}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'} = \mathbf{R}\mathbf{G}.(\boldsymbol{\theta}).$$

Como $\hat{\boldsymbol{\theta}}_{\mathrm{G}}$ es el estimador OLS del modelo transformado, para n suficientemente grande, tiene matriz de varianza-covarianza dada por

$$\mathcal{D}[\hat{\boldsymbol{\theta}}_{\mathrm{G}}] \approx \sigma^2[\mathbf{K}.'(\boldsymbol{\theta}^*)\mathbf{K}.(\boldsymbol{\theta}^*)]^{-1} = \sigma^2[\mathbf{G}.'(\boldsymbol{\theta}^*)\mathbf{R}'\mathbf{R}\mathbf{G}.(\boldsymbol{\theta}^*)]^{-1} = \sigma^2[\mathbf{G}.'(\boldsymbol{\theta}^*)\mathbf{V^{-1}}\mathbf{G}.(\boldsymbol{\theta}^*)]^{-1}.$$

Esta matriz es estimada por

$$\hat{\mathcal{D}}[\hat{\boldsymbol{\theta}}_{\mathrm{G}}] = \hat{\sigma}^2(\hat{\mathbf{K}}.'\hat{\mathbf{K}}.)^{-1} = \hat{\sigma}^2(\hat{\mathbf{G}}.'\mathbf{V^{-1}}\hat{\mathbf{G}}.)^{-1},$$

donde

$$\hat{\sigma}^2 = \frac{1}{n-p}[\mathbf{z} - \mathbf{k}(\hat{\boldsymbol{\theta}}_{\mathrm{G}})]'[\mathbf{z} - \mathbf{k}(\hat{\boldsymbol{\theta}}_{\mathrm{G}})] = \frac{1}{n-p}[\mathbf{y} - \mathbf{g}(\hat{\boldsymbol{\theta}}_{\mathrm{G}})]'\mathbf{V^{-1}}[\mathbf{y} - \mathbf{g}(\hat{\boldsymbol{\theta}}_{\mathrm{G}})].$$

El punto importante de este análisis es que, tratando con el problema transformado como si fuera un problema ordinario de mínimos cuadrados, el mismo produce los resultados correctos para el problema de mínimos cuadrados generalizados. Muchas aplicaciones de este método surgen cuando los errores no son homoscedásticos, pero sí son independientes, en cuyo caso $\mathbf{V}$ es diagonal y el problema es computacionalmente más sencillo.

2.6. Estimadores de Máxima Verosimilitud

Si asumimos conocida la distribución conjunta de $\boldsymbol{\varepsilon}_i$ en el modelo (2.1), entonces el estimador de máxima verosimilitud $\boldsymbol{\theta}$ se obtiene maximizando la función de verosimilitud. Supongamos que $\boldsymbol{\varepsilon}_i$ son *i.i.d.* con función de densidad $\sigma^{-1}h(\boldsymbol{\varepsilon}/\sigma)$, de modo que h es la distribución del error para errores estandarizados para tener varianza uno. Luego, la función de verosimilitud es

$$p(\mathbf{y}|\boldsymbol{\theta}, \sigma^2) = \prod_{i=1}^{n}\left[\sigma^{-1}h\left(\frac{y_i - g(\mathbf{x}_i, \boldsymbol{\theta})}{\sigma}\right)\right]. \tag{2.27}$$

A continuación, estudiaremos errores distribuidos normalmente y no-normalmente. Encontraremos que, bajo normalidad, el estimador de máxima verosimilitud $\boldsymbol{\theta}$ coincide con el estimador de mínimos cuadrados.

Si los $\boldsymbol{\varepsilon}_i$ son *i.i.d.* $N(0, \sigma^2)$, entonces (2.27) cumple

$$p(\mathbf{y}|\boldsymbol{\theta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} \frac{[y_i - g(\mathbf{x}_i, \boldsymbol{\theta})]^2}{\sigma^2}\right). \tag{2.28}$$

Despreciando las constantes, denotamos el logaritmo de la función de verosimilitud por $L(\boldsymbol{\theta}, \sigma^2)$ y obtenemos

$$L(\boldsymbol{\theta}, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - g(\mathbf{x}_i, \boldsymbol{\theta})]^2 = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\mathbf{S}(\boldsymbol{\theta}). \tag{2.29}$$

Dado σ^2, (2.29) es maximizado respecto a $\boldsymbol{\theta}$ cuando $\mathbf{S}(\boldsymbol{\theta})$ es minimizado, es decir, cuando $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ (el estimador de mínimos cuadrados). Es más, $\partial L/\partial\sigma^2 = 0$ tiene solución $\sigma^2 = \mathbf{S}(\boldsymbol{\theta})/n$, que da un máximo (para un $\boldsymbol{\theta}$ dado) al dar negativa la derivada segunda. Esto sugiere que $\hat{\boldsymbol{\theta}}$ y $\hat{\sigma}^2 = \mathbf{S}(\hat{\boldsymbol{\theta}})/n$ son los estimadores de máxima verosimilitud, y podemos verificarlo directamente. Como $\mathbf{S}(\boldsymbol{\theta}) \geq \mathbf{S}(\hat{\boldsymbol{\theta}})$,

$$\begin{aligned} L(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2) - L(\boldsymbol{\theta}, \sigma^2) &= -\frac{n}{2}\log(\hat{\sigma}^2) - \frac{n}{2} - L(\boldsymbol{\theta}, \sigma^2) \\ &\geq -\frac{n}{2}\log\left(\frac{\hat{\sigma}^2}{\sigma^2}\right) - \frac{n}{2} + \frac{1}{2}\frac{\mathbf{S}(\hat{\boldsymbol{\theta}})}{\sigma^2} \\ &= -\frac{n}{2}\left(\log\left(\frac{\hat{\sigma}^2}{\sigma^2}\right) + 1 - \frac{\hat{\sigma}^2}{\sigma^2}\right) \\ &\geq 0, \end{aligned}$$

pues $\log(x) \leq x - 1$ para $x \geq 0$. Por lo tanto, $\hat{\boldsymbol{\theta}}$ y $\hat{\sigma}^2$ maximizan $L(\boldsymbol{\theta}, \sigma^2)$. El máximo valor de (2.28) es

$$p(\mathbf{y}|\hat{\boldsymbol{\theta}}, \hat{\sigma}^2) = (2\pi\hat{\sigma}^2)^{-n/2} \exp(-n/2). \tag{2.30}$$

Jennrich (1969) notó que el estimador de mínimos cuadrados es ahora no sólo el estimador de máxima verosimilitud sino que, bajo condiciones de regularidad, es también asintóticamente eficiente. La teoría asintótica de máxima verosimilitud usual no aplica directamente, sino que necesita modificaciones, ya que las y_i no son *i.i.d.* teniendo distintas medias. Así si $\boldsymbol{\delta} = (\boldsymbol{\theta}', v)'$, donde $v = \sigma^2$, entonces de (2.29) y (2.25) la matriz de información (esperada) viene dada por

$$-E\left[\frac{\partial^2 L}{\partial\boldsymbol{\delta}\,\partial\boldsymbol{\delta}'}\right] = \begin{bmatrix} \frac{1}{2\sigma^2}E\left[\frac{\partial^2\mathbf{S}}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right] & -E\left[\frac{\partial^2 L}{\partial\boldsymbol{\theta}\,\partial v}\right] \\ -E\left[\frac{\partial^2 L}{\partial v\,\partial\boldsymbol{\theta}'}\right] & -E\left[\frac{\partial^2 L}{\partial v^2}\right] \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2}\mathbf{G}_\cdot'(\boldsymbol{\theta}^*)\mathbf{G}_\cdot(\boldsymbol{\theta}^*) & \mathbf{0} \\ \mathbf{0}' & \frac{n}{2\sigma^4} \end{bmatrix} \tag{2.31}$$

Luego, bajo condiciones de regularidad, tenemos que

$$\lim_{n\to\infty} \mathcal{D}[\sqrt{n}\hat{\boldsymbol{\theta}}] = \sigma^2\boldsymbol{\Omega}^{-1} = \sigma^2 \lim_{n\to\infty} n[\mathbf{G}_\cdot'(\boldsymbol{\theta}^*)\mathbf{G}_\cdot(\boldsymbol{\theta}^*)]^{-1}.$$

Al igual que en el caso *i.i.d.*, podemos ver que la matriz de varianza–covarianza del estimador de máxima verosimilitud $\hat{\boldsymbol{\theta}}$ viene dada asintóticamente por la inversa de la matriz de información dada en (2.31).

Sea $\hat{\boldsymbol{\delta}} = (\hat{\boldsymbol{\theta}}', \hat{\sigma}^2)'$. Entonces, como $\partial \mathbf{S}(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \mathbf{0}$ en $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, también tenemos

$$\left[-\frac{\partial^2 L}{\partial \boldsymbol{\delta}\, \partial \boldsymbol{\delta}'}\right]^{-1}_{\hat{\boldsymbol{\delta}}} = \begin{bmatrix} \frac{1}{2\hat{\sigma}^2}\left[\frac{\partial^2 \mathbf{S}}{\partial \boldsymbol{\delta}\, \partial \boldsymbol{\delta}'}\right]_{\hat{\boldsymbol{\theta}}} & \mathbf{0} \\ \mathbf{0}' & \frac{n}{2\hat{\sigma}^4} \end{bmatrix}^{-1} = \begin{bmatrix} 2\hat{\sigma}^2\left[\frac{\partial^2 \mathbf{S}}{\partial \boldsymbol{\delta}\, \partial \boldsymbol{\delta}'}\right]^{-1}_{\hat{\boldsymbol{\theta}}} & \mathbf{0} \\ \mathbf{0}' & \frac{2\hat{\sigma}^4}{n} \end{bmatrix}.$$

A modo de completitud ahora vamos a alejarnos de nuestro tema principal de modelos de regresión no lineales y consideraremos el caso en que la función de máxima verosimilitud es una función general de $E[\mathbf{y}] = \mathbf{g}(\boldsymbol{\theta})$, digamos $L([\mathbf{g}(\boldsymbol{\theta})])$. Asumiremos que el modelo es suficientemente regular tal que el estimador de máxima verosimilitud $\hat{\boldsymbol{\theta}}$ sea solución de las ecuaciones de verosimilitud

$$\mathbf{0} = \frac{\partial L}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}'}\right)' \frac{\partial L}{\partial \mathbf{g}}.$$

Sea $\boldsymbol{\theta}^{(a)}$ la *a-ésima* aproximación para $\hat{\boldsymbol{\theta}}$ y consideremos la expansión de Taylor

$$\mathbf{0} = \frac{\partial L}{\partial \boldsymbol{\theta}}\Big|_{\hat{\boldsymbol{\theta}}} \approx \frac{\partial L}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^{(a)}} + \frac{\partial^2 L}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\Big|_{\boldsymbol{\theta}^{(a)}} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(a)}).$$

Luego

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(a)} \approx \left[\left(-\frac{\partial^2 L}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right)^{-1} \frac{\partial L}{\partial \boldsymbol{\theta}}\right]\Bigg|_{\boldsymbol{\theta}^{(a)}} = \boldsymbol{\delta}^{(a)}, \tag{2.32}$$

de modo que un estimador actualizado es $\boldsymbol{\theta}^{(a+1)} = \boldsymbol{\theta}^{(a)} + \boldsymbol{\delta}^{(a)}$. Este es el método de Newton para resolver las ecuaciones, en este contexto los estadísticos se refieren al mismo como el *método de Newton-Raphson*. En general, la matriz negativa de la segunda derivada contiene variables aleatorias y es recomendable que sean reemplazadas por su valor esperado, la llamada matriz de información (esperada). Esta técnica es también conocida como *algoritmo de "scoring" de Fisher*. Ahora encontraremos la matriz esperada, tenemos que

$$\frac{\partial^2 L}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} = \left(\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}'}\right)' \frac{\partial^2 L}{\partial \mathbf{g}\, \partial \mathbf{g}'} \left(\frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}'}\right) + \sum_i \frac{\partial L}{\partial g_i} \frac{\partial^2 g_i}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}. \tag{2.33}$$

Asumiendo que el orden de diferenciación respecto a $\boldsymbol{\theta}$ y la integral (o suma) con respecto a $\mathbf{y}$ puede ser intercambiada, tenemos (con $p(\cdot)$ la función de densidad de $\mathbf{y}$, y $L = \log(p)$)

$$E\left[\frac{\partial L}{\partial \mathbf{g}}\right] = \int \frac{1}{p} \frac{\partial p}{\partial \mathbf{g}} p \, d\mathbf{y} = \frac{\partial}{\partial \mathbf{g}} \int p \, d\mathbf{y} = \mathbf{0}, \tag{2.34}$$

y, usando un argumento similar

$$E\left[\frac{1}{p}\frac{\partial^2 p}{\partial \mathbf{g}\,\partial \mathbf{g}'}\right] = \mathbf{0}.$$

Por lo tanto,

$$E\left[\frac{\partial^2 L}{\partial \mathbf{g}\,\partial \mathbf{g}'}\right] = E\left[\frac{\partial}{\partial \mathbf{g}}\left(\frac{1}{p}\frac{\partial p}{\partial \mathbf{g}'}\right)\right] = E\left[-\frac{1}{p^2}\frac{\partial p}{\partial \mathbf{g}}\frac{\partial p}{\partial \mathbf{g}'} + \frac{1}{p}\frac{\partial^2 p}{\partial \mathbf{g}\,\partial \mathbf{g}'}\right] = E\left[\frac{\partial L}{\partial \mathbf{g}}\frac{\partial L}{\partial \mathbf{g}'}\right] = -\mathbf{J}. \tag{2.35}$$

Notemos que $\mathbf{J}$ es definida positiva

$$\mathbf{a}'\mathbf{J}\mathbf{a} = E\left[\left(\mathbf{a}'\frac{\partial L}{\partial \mathbf{g}}\right)^2\right] \geq 0,$$

la igualdad sólo vale si $\mathbf{a} = \mathbf{0}$ (bajo condiciones generales sobre L). Aplicando (2.35) y (2.34) a (2.33) obtenemos

$$E\left[-\frac{\partial^2 L}{\partial \boldsymbol{\theta}\,\partial \boldsymbol{\theta}'}\right] = \mathbf{G}.'\mathbf{J}\mathbf{G}.,$$

que es definida positiva pues $\mathbf{G}.$ es de rango completo y además no singular. Luego, el algoritmo de scoring de Fisher resulta

$$\boldsymbol{\theta}^{(a+1)} = \boldsymbol{\theta}^{(a)} + \left((\mathbf{G}.'\mathbf{J}\mathbf{G}.)^{-1}\frac{\partial L}{\partial \boldsymbol{\theta}}\right)_{\boldsymbol{\theta}^{(a)}}. \tag{2.36}$$

La matriz negativa de la segunda derivada podría no ser definida positiva en todo $\boldsymbol{\theta}^{(a)}$, y esto puede causar que el método de Newton de (2.32) falle. El método de scoring de Fisher, por lo tanto, utiliza una aproximación de la matriz negativa de la segunda derivada que es siempre definida positiva. Una ventaja es que sólo requiere de las derivadas primeras de L, de modo que la aproximación puede, a menudo, ser calculada más rápidamente que la matriz de la segunda derivada. El precio que se paga por esta ventaja es que el algoritmo converge más lentamente que el método de Newton.

Si tomamos $\mathbf{j} = \partial L/\partial \mathbf{g}$, obtenemos de (2.32) y (2.36)

$$\boldsymbol{\delta}^{(a)} = [(\mathbf{G}.'\mathbf{J}\mathbf{G}.)^{-1}\mathbf{G}.'\mathbf{j}]_{\boldsymbol{\theta}^{(a)}} = [(\mathbf{G}.'\mathbf{V}^{-1}\mathbf{G}.)^{-1}\mathbf{G}.'\mathbf{V}^{-1}\mathbf{v}]_{\boldsymbol{\theta}^{(a)}}, \tag{2.37}$$

donde $\mathbf{V} = \mathbf{J}^{-1}$ y $\mathbf{v} = \mathbf{V}\mathbf{j}$. Notemos que la ecuación (2.37) representa al estimador de mínimos cuadrados generalizados $\boldsymbol{\delta}^{(a)}$ de $\boldsymbol{\delta}$ para el modelo (evaluado en $\boldsymbol{\theta}^{(a)}$)

$$\mathbf{v} = \mathbf{G}_{\cdot}^{(a)}\boldsymbol{\delta} + \mathbf{v}, \tag{2.38}$$

siendo $\mathcal{D}[\mathbf{v}] = \mathbf{V}^{(a)}$. Usando la misma idea que en la Sección 2.5.4, sea $\mathbf{V}^{(a)} = \mathbf{U}'\mathbf{U}$ la descomposición de Cholesky de $\mathbf{V}^{(a)}$ y sea $\mathbf{R} = (\mathbf{U}')^{-1}$. Entonces multiplicando (2.38) por $\mathbf{R}$, obtenemos

$$\mathbf{z} = \mathbf{X}\boldsymbol{\delta} + \boldsymbol{\eta}, \tag{2.39}$$

donde $\mathbf{z} = \mathbf{R}\mathbf{v}$, $\mathbf{X} = \mathbf{R}\mathbf{G}_{\cdot}^{(a)}$, y $\mathcal{D}[\boldsymbol{\eta}] = \sigma^2\mathrm{I_n}$. Por lo tanto, $\boldsymbol{\delta}^{(a)}$ puede computarse usando un programa de OLS y regresión de $\mathbf{z}$ sobre $\mathbf{X}$. Un resultado estándar de un programa de regresión lineal es $(\mathbf{X}'\mathbf{X})^{-1}$, en (2.39)

$$(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{G}_{\cdot}'\mathbf{V}^{-1}\mathbf{G}_{\cdot})^{-1}_{\boldsymbol{\theta}^{(a)}}.$$

Luego, el resultado al final de la iteración es

$$\left(E\left[-\frac{\partial^2 L}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right]_{\hat{\boldsymbol{\theta}}}\right)^{-1}$$

que, bajo ciertas condiciones de regularidad, es la matriz asintótica de varianza-covarianza de $\hat{\boldsymbol{\theta}}$.

Este método es muy general y comunmente se refiere al mismo como *mínimos cuadrados iterativamente pesados* (IRLS).

Capítulo 3

Robustez

Todos los métodos estadísticos clásicos de estimación se basan en fuertes supuestos tales como datos que se distribuyen normalmente. A menudo este supuesto vale, pero sólo aproximadamente, ya que describe el comportamiento de la mayoría de las observaciones, siendo unas pocas las que siguen un patrón diferente. A estos datos atípicos suele llamárselos *outliers*. Pueden deberse a realizaciones del experimento en circunstancias anormales, errores de medición o una equivocación en la transcripción del dato, entre otros factores. Tienen la particularidad de que la presencia de sólo unos pocos de ellos pueden afectar severamente los resultados obtenidos mediante métodos clásicos como, por ejemplo, el de mínimos cuadrados que es óptimo bajo condiciones de normalidad.

En estos casos, lo ideal sería tener estimadores robustos ya que estos ajustan bien a la mayoría de los datos. Si la muestra no contiene datos atípicos tienen poca pérdida de eficiencia respecto de los estimadores clásicos y dan resultados estables aún si la muestra contiene una cantidad moderada de outliers. Es decir, son estimadores poco sensibles a datos atípicos y simultáneamente son altamente eficientes cuando los datos son normales.

En principio, estudiaremos los estimadores de posición y dispersión para luego enfocarnos en estimadores robustos aplicados a los modelos de regresión lineal y no lineal. Maronna, Martin y Yohai (2006) dan una revisión muy completa de los temas que desarrollamos en este capítulo.

3.1. Modelo de posición

Sea el modelo de posición

$$x_i = \mu + u_i, \qquad 1 \leq i \leq n \tag{3.1}$$

donde μ es el parámetro de posición y $u_1, \ldots, u_n$ son variables aleatorias.

Si las observaciones son repeticiones independientes del mismo experimento bajo las mismas condiciones, se asume que

- $u_1, \ldots, u_n$ tienen la misma función de distribución F_0.
- $u_1, \ldots, u_n$ son independientes.

En este caso, resulta que $x_1, \ldots, x_n$ son independientes con función de distribución

$$F(x) = F_0(x - \mu) \tag{3.2}$$

y decimos que las x_i´s son variables aleatorias independientes e idénticamente distribuidas (*i.i.d.*). Notaremos $\mathbf{x}$ la muestra aleatoria constituida por $x_1, \ldots, x_n$.

Un estimador $\hat{\mu}$ es una función de las observaciones: $\hat{\mu} = \hat{\mu}(x_1, \ldots, x_n) = \hat{\mu}(\mathbf{x})$. Lo que buscamos son estimadores $\hat{\mu}$ tal que en algún sentido estén próximos a μ con alta probabilidad. Una forma de medir esta aproximación es mediante el *Error Cuadrático Medio (MSE)*:

$$\mathrm{MSE}(\hat{\mu}) = E(\hat{\mu} - \mu)^2 .$$

Si los datos fueran exactamente normales, la media muestral sería el estimador óptimo: el Estimador de Máxima Verosimilitud (EMV) y minimizaría al MSE entre todos los estimadores insesgados y los equivariantes. Pero los datos raramente tienen tan buen comportamiento.

En la mayoría de las aplicaciones prácticas a lo sumo se puede asegurar que los errores de medición tienen distribución *aproximadamente normal*. Una forma de determinar distribuciones aproximadamente normales es considerar que una proporción $1 - \varepsilon$ de las observaciones son generadas por el modelo normal, mientras que una proporción ε es generada por un mecanismo desconocido. Llamamos *distribución normal contaminada* a

$$F = (1 - \varepsilon)G + \varepsilon H \tag{3.3}$$

donde $G = \mathrm{N}(\mu, \sigma^2)$ y H una distribución arbitraria.

Supongamos que tenemos el modelo de posición dado por (3.1) donde la distribución F de los u_i es simétrica respecto de 0. Como en este caso μ coincide con la mediana, un estimador alternativo sería $\tilde{\mu} = \mathrm{mediana}(x_1, \ldots, x_n)$. Ordenamos los datos $x_1, \ldots, x_n$ de menor a mayor obteniendo los valores $x_{(1)} \leq \ldots \leq x_{(n)}$, luego la mediana se define como

$$\tilde{\mu} = \begin{cases} x_{(m+1)} & n = 2m + 1 \\ \frac{x_{(m)} + x_{(m+1)}}{2} & n = 2m \end{cases}$$

Si bien este estimador es mucho más resistente a outliers que la media muestral, como contrapartida, es menos eficiente en cuanto a varianza asintótica se refiere. En las siguientes secciones trataremos de encontrar estimadores que satisfagan en simultáneo las "buenas" propiedades de estos dos estimadores.

3.1.1. M-estimadores de posición

A continuación, desarrollaremos una familia de estimadores que contienen a la media y a la mediana como casos especiales.

Consideremos nuevamente al modelo de posición (3.1) y asumamos que F_0, la función de distribución de u_i, tiene densidad $f_0 = F_0'$. Luego, la densidad conjunta de las observaciones, es decir, la *función de verosimilitud* es

$$L(x_1, \ldots, x_n; \mu) = \prod_{i=1}^{n} f_0(x_i - \mu) .$$

El *estimador de máxima verosimilitud* (EMV) de μ es el valor $\hat{\mu}(x_1, \ldots, x_n)$ que maximiza $L(x_1, \ldots, x_n; \mu)$:

$$\hat{\mu} = \hat{\mu}(x_1, \ldots, x_n) = \arg \max_{\mu} L(x_1, \ldots, x_n; \mu)\,. \tag{3.4}$$

Si f_0 es positiva en todos lados, como el logaritmo es una función creciente, podemos escribir (3.4) como

$$\hat{\mu} = \arg \min_{\mu} \sum_{i=1}^{n} \rho(x_i - \mu)\,, \tag{3.5}$$

donde

$$\rho = -\log f_0\,.$$

Si ρ es diferenciable, diferenciando (3.5) respecto a μ obtenemos

$$\sum_{i=1}^{n} \psi(x_i - \hat{\mu}) = 0 \tag{3.6}$$

con $\psi = \rho'$. Notemos que si f_0 es simétrica, entonces ρ es par y por lo tanto ψ es impar.

Si $\rho(x) = x^2/2$, entonces $\psi(x) = x$, y (3.6) sería

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

cuya solución viene dada por $\hat{\mu} = \bar{x}$. Y para $\rho(x) = |x|$, se puede demostrar que la mediana de las observaciones es solución de (3.5).

Por lo tanto, dada una función ρ, un *M-estimador de posición* es una solución de (3.5). La función ρ deberá ser elegida de manera tal que el estimador sea

(A) "cercanamente óptimo" cuando F_0 es exactamente normal, y

(B) "cercanamente óptimo" cuando F_0 es aproximadamente normal.

Asumiendo que ψ es monótona no decreciente, con $\psi(-\infty) < 0 < \psi(\infty)$ resulta que (3.6), y por lo tanto (3.5), siempre tiene solución y si ψ es continua y estrictamente creciente, la solución es única.

Supongamos que ψ es estrictamente creciente. Dada una distribución F, definimos $\mu_0 = \mu_0(F)$ como la solución de

$$E_F[\psi(x - \mu_0)] = 0\,.$$

Puede demostrarse que cuando $n \to \infty$, $\hat{\mu} \to_p \mu_0$ (decimos que $\hat{\mu}$ es *consistente para* μ_0), y la distribución de $\hat{\mu}$ es aproximadamente

$$\mathrm{N}\left(\mu_0, \frac{v}{n}\right) \quad \text{con} \quad v = \frac{E_F[\psi(x - \mu_0)^2]}{(E_F[\psi'(x - \mu_0)])^2}\,. \tag{3.7}$$

Notemos que bajo el modelo (3.2) v no depende de μ_0, es decir

$$v = \frac{E_{F_0}[\psi(x)^2]}{(E_{F_0}[\psi'(x)])^2}\,.$$

Si la distribución de un estimador $\hat{\mu}$ es aproximadamente $\mathrm{N}(\mu_0, v/n)$ para un n grande, decimos que $\hat{\mu}$ es *asintóticamente normal*, con valor asintótico μ_0 y varianza asintótica v. La *eficiencia asintótica* de $\hat{\mu}$ es la proporción

$$\mathrm{Eff}(\hat{\mu}) = \frac{v_0}{v},$$

donde v_0 es la varianza asintótica del EMV y mide cuán cerca está $\hat{\mu}$ del óptimo. La expresión de v en (3.7) se llama la *varianza asintótica* de $\hat{\mu}$.

Huber (1964) propuso una familia de funciones-ρ con importantes propiedades

$$\rho(x) = \begin{cases} x^2 & \text{si } |x| \leq k \\ 2k|x| - k^2 & \text{si } |x| > k \end{cases} \tag{3.8}$$

con derivada $2\psi(x)$, donde

$$\psi(x) = \begin{cases} x & \text{si } |x| \leq k \\ sgn(x)k & \text{si } |x| > k \end{cases},$$

siendo k una constante de calibración que es elegida de manera de obtener una eficiencia determinada. Los M-estimadores correspondientes a los casos límites $k \to \infty$ y $k \to 0$ son la media y la mediana respectivamente, y se define $\psi(x)$ como $sgn(x)$.

M-estimadores redescendientes

Una elección popular de funciones-ρ y ψ es la familia de funciones *bicuadrada* dada por

$$\rho(x) = \begin{cases} 1 - [1 - (x/k)^2]^3 & \text{si } |x| \leq k \\ 1 & \text{si } |x| > k \end{cases}, \tag{3.9}$$

con derivada $\rho'(x) = 6\psi(x)/k^2$ donde

$$\psi(x) = x[1 - (x/k)^2]^2 \ \mathrm{I}(|x| \leq k)\,. \tag{3.10}$$

Llamaremos "M-estimadores monótonos" a aquellos estimadores definidos como solución de (3.6) con ψ monótona y "M-estimadores redescendiente" a los definidos mediante (3.5) cuando ψ no es monótona. Los estimadores redescendientes son más robustos frente a una gran cantidad de outliers.

Definición 3.1. *Llamaremos función-ρ a una función ρ que satisfaga*

R1. *$\rho(x)$ es una función no decreciente de $|x|$*

R2. $\rho(0) = 0$

R3. *$\rho(x)$ es estrictamente creciente para $x > 0$ tal que $\rho(x) < ||\rho||_\infty$*

R4. *Si ρ es acotada, también se asume que $||\rho||_\infty = 1$*

Definición 3.2. *Una función-ψ denotará una función ψ que es la derivada de un función-ρ, que implica en particular que*

Ψ1. *ψ es impar y $\psi(x) \geq 0$ para $x \geq 0$.*

3.2. Estimadores de dispersión

La forma clásica de medir la variabilidad de un conjunto de datos $\mathbf{x}$ es con la *desviación standard (SD)*

$$\mathrm{SD}(\mathbf{x}) = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}.$$

Para cualquier constante c el SD es invariante por traslaciones y cambios de escala, es decir,

$$\mathrm{SD}(\mathbf{x}+c) = \mathrm{SD}(\mathbf{x}), \qquad \mathrm{SD}(c\mathbf{x}) = |c|\mathrm{SD}(\mathbf{x})\,. \tag{3.11}$$

Cualquier estadístico que satisfaga (3.11) será un *estimador de dispersión.*

Una alternativa al SD es la *desviación media absoluta (MD)*

$$\mathrm{MD}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|\,, \tag{3.12}$$

que también es sensible a la presencia de datos atípicos en tanto está basada en promedios muestrales.

Una alternativa robusta es reemplazar los promedios muestrales por las medianas, definiendo de esta forma la *desviación mediana absoluta respecto de la mediana (MAD)*

$$\mathrm{MAD}(\mathbf{x}) = \mathrm{Med}(|\mathbf{x} - \mathrm{Med}(\mathbf{x})|)\,, \tag{3.13}$$

que claramente satisface (3.11).

De la misma manera que en (3.12) y (3.13) definimos las desviaciones medias y medianas de una variable aleatoria x como

$$\mathrm{MD}(x) = E(|x - E(x)|)$$

y

$$\mathrm{MAD}(x) = \mathrm{Med}(|x - \mathrm{Med}(x)|)\,,$$

respectivamente.

Notemos que si $x \sim \mathrm{N}(\mu, \sigma^2)$ entonces $\mathrm{SD}(x) = \sigma$ por definición, mientras que $\mathrm{MD}(x)$ y $\mathrm{MAD}(x)$ son múltiplos de σ

$$\mathrm{MD}(x) = 2\varphi(0)\sigma \quad \text{y} \quad \mathrm{MAD}(x) = \Phi^{-1}(0.75)\sigma\,.$$

Si quisiéramos un estimador de dispersión que midiera lo mismo que el SD bajo normalidad deberíamos normalizar la MAD dividiéndola por $c \approx 0.675$. La "MAD normalizada" *(MADN)* se define entonces como

$$\text{MADN}(x) = \frac{\text{MAD}(x)}{0.675}$$

3.3. M-estimadores de escala

Consideremos observaciones x_i que satisfagan el modelo *multiplicativo*

$$x_i = \sigma u_i \,, \tag{3.14}$$

donde las u_i's son *i.i.d.* con función de densidad f_0 y $\sigma > 0$ es el parámetro desconocido. Las distribuciones de las x_i's constituyen una *familia de escala* con densidad

$$\frac{1}{\sigma} f_0 \left(\frac{x}{\sigma}\right) .$$

El EMV de σ en (3.14) es

$$\hat{\sigma} = \arg \max_{\sigma} \frac{1}{\sigma^n} \prod_{i=1}^{n} f_0 \left(\frac{x_i}{\sigma}\right).$$

Tomando logaritmo y diferenciando respecto a σ se tiene que

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{x_i}{\hat{\sigma}}\right) = 1 \,,$$

donde $\rho(t) = t\psi(t)$, con $\psi = -f_0'/f_0$. En general, a cualquier estimador que satisfaga una ecuación de la forma

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{x_i}{\hat{\sigma}}\right) = \delta, \tag{3.15}$$

donde ρ es una función-ρ y δ es una constante positiva, lo llamaremos un *M-estimador de escala*. Notemos que para tener solución en (3.15) debemos tener $0 < \delta < ||\rho||_\infty$. Luego, si ρ es acotada asumiremos, sin pérdida de generalidad, que

$$||\rho||_\infty = 1, \quad \delta \in (0, 1).$$

Se puede verificar fácilmente que los M-estimadores de escala son equivariantes en el sentido de que $\hat{\sigma}(c\mathbf{x}) = c\hat{\sigma}(\mathbf{x})$ para cualquier $c > 0$, y si ρ es par entonces

$$\hat{\sigma}(c\mathbf{x}) = |c|\, \hat{\sigma}(\mathbf{x})$$

para cualquier c.

Para un n grande, la sucesión de estimadores de (3.15) converge a la solución de

$$E\left[\rho\left(\frac{x}{\sigma}\right)\right] = \delta$$

si es única.

Notemos que usando $\rho(x/c)$ en lugar de $\rho(x)$ en (3.15) se obtiene $\hat{\sigma}/c$. Esto se usa para normalizar $\hat{\sigma}$ y obtener un valor asintótico dado.

3.4. M-estimadores de posición con escala desconocida

Los estimadores definidos en (3.5) no son en general equivariantes por cambio de escala, lo cual implica que nuestros resultados dependen fuertemente de nuestras unidades de medida.

Para fijar ideas, supongamos que queremos estimar μ en el modelo (3.1) donde F viene dada por (3.3) con $G = \mathrm{N}(\mu, \sigma^2)$. Si σ fuese conocida, lo natural sería dividir (3.1) por σ y así reducir el problema al caso $\sigma = 1$, lo cual implica estimar a μ mediante

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{\sigma}\right).$$

Para obtener M-estimadores de posición que sean equivariantes por cambios de escala, una propuesta intuitiva sería usar

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{\hat{\sigma}}\right), \tag{3.16}$$

donde $\hat{\sigma}$ es un estimador de dispersión computado previamente, que deberá ser robusto. Es fácil de verificar que $\hat{\mu}$ es realmente invariante por cambios de escala. Como $\hat{\sigma}$ no depende de μ, (3.16) implica que $\hat{\mu}$ es una solución de

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) = 0.$$

Otra posibilidad es considerar un modelo de posición-dispersión, con dos parámetros desconocidos y estimarlos simultáneamente (ver Maronna, Martin y Yohai, 2006), pero en general, la estimación con cómputo previo del parámetro de dispersión, suele ser más robusto que la estimación simultánea.

3.5. Punto de Ruptura

Entre las medidas de robustez más usadas se encuentra el *punto de ruptura.* Hampel (1971) formaliza esta noción de *punto de ruptura* de un estimador $\hat{\boldsymbol{\theta}}$ del parámetro $\boldsymbol{\theta}$ como la mayor cantidad de contaminación que pueden contener los datos de manera tal que $\hat{\boldsymbol{\theta}}$ siga dando información sobre $\boldsymbol{\theta}$.

Sea $\boldsymbol{\theta} \in \Theta$, Θ el espacio de parámetros. Para que el estimador $\hat{\boldsymbol{\theta}}$ dé información sobre $\boldsymbol{\theta}$, la contaminación no deberá conducir a $\hat{\boldsymbol{\theta}}$ a infinito o a la frontera de Θ.

Luego, se define el *punto de ruptura de contaminación asintótico* de $\hat{\boldsymbol{\theta}}$ en F, denotado por $\varepsilon^*(\hat{\boldsymbol{\theta}}, F)$ como el mayor valor $\varepsilon^* \in (0,1)$ tal que $\forall \varepsilon < \varepsilon^*$, $\hat{\boldsymbol{\theta}}_\infty((1-\varepsilon)F + \varepsilon G)$ se mantiene acotado en función de G y también lejos de la frontera de Θ.

Para que un estimador sea razonable es claramente intuitivo que debe haber mayor cantidad de datos "típicos" que "atípicos", por esto $\varepsilon^* \leq 1/2$.

A modo de ejemplo, los M-estimadores de posición y escala conocida con ψ monótona no necesariamente impar y

$$k_1 = -\psi(-\infty), \qquad k_2 = \psi(\infty)$$

finitos. Se puede ver que

$$\varepsilon^* = \frac{\text{mín}(k_1, k_2)}{k_1 + k_2}$$

Si ψ fuese impar, entonces $k_1 = k_2$ y por lo tanto $\varepsilon^* = 1/2$, cota que también alcanzan los estimadores redescendientes. En el caso de estimadores de dispersión tenemos que SD y MAD tienen punto de ruptura igual a 0 y $1/2$, respectivamente.

3.6. Modelo de regresión lineal con predictores fijos

Supongamos que se tienen n observaciones $(x_{i1}, ..., x_{ip}, y_i)$, donde $\mathbf{x}_i = (x_{i1}, ..., x_{ip})'$ son variables predictoras (o variables independientes) e y_i es una variable de respuesta (o variable dependiente) que cumplen el siguiente *modelo lineal*

$$y_i = \sum_{i=1}^{p} x_{ij}\beta_j + u_i, \qquad 1 \leq i \leq n, \tag{3.17}$$

donde $\beta_1, ..., \beta_p$ son los parámetros desconocidos a ser estimados y los errores u_i son variables aleatorias *i.i.d.*, que no dependen de $\mathbf{x}_i$. En esta sección, consideramos a $\mathbf{x}_i$ fijos, es decir, determinados antes de realizar el experimento. Llamando $\boldsymbol{\beta}$ al vector columna $(\beta_1, ..., \beta_p)'$, el modelo puede ser escrito de la siguiente forma

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i\,. \tag{3.18}$$

Si llamamos $\mathbf{X}$ a la matriz de $n \times p$ con elementos x_{ij} y por otro lado, $\mathbf{y}$ y $\mathbf{u}$ a los vectores con elementos y_i y u_i, respectivamente, el modelo lineal puede ser escrito en forma matricial como

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \tag{3.19}$$

3.6.1. M-estimadores

Asumamos que se cumple el modelo (3.19), donde u_i tiene densidad

$$\frac{1}{\sigma} f_0\left(\frac{u}{\sigma}\right)$$

y σ es el parámetro de escala. En el modelo lineal (3.19), las y_i son independientes pero no idénticamente distribuidas, de hecho cada y_i tiene densidad

$$\frac{1}{\sigma} f_0\left(\frac{y - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)$$

y por lo tanto, la función de verosimilitud para $\boldsymbol{\beta}$, asumiendo fijo el valor de σ es

$$L(\boldsymbol{\beta}) = \frac{1}{\sigma^n} \prod_{i=1}^{n} f_0 \left(\frac{y - \mathbf{x}_i' \boldsymbol{\beta}}{\sigma} \right) .$$

Calcular el EMV maximizando $L(\mathbf{b})$ es equivalente a encontrar $\hat{\boldsymbol{\beta}}_n$ tal que minimice

$$\frac{1}{n} \sum_{i=1}^{n} \rho_0 \left(\frac{r_i(\mathbf{b})}{\sigma} \right) + \log(\sigma) ,$$

donde $\rho_0 = -\log f_0$. Derivando respecto a $\mathbf{b}$ obtenemos el equivalente a las ecuaciones normales

$$\sum_{i=1}^{n} \psi_0 \left(\frac{r_i(\mathbf{b})}{\sigma} \right) \mathbf{x}_i = 0 ,$$

donde $\psi_0 = \rho_o' = -f_0'/f_0$.

Dada una función de pérdida ρ que satisface la Definición 3.1, definimos un *M-estimador de regresión* como

$$\hat{\boldsymbol{\beta}}_n = \arg \min_{\mathbf{b}} \sum_{i=1}^{n} \rho \left(\frac{r_i(\mathbf{b})}{\hat{\sigma}} \right) , \tag{3.20}$$

donde $\hat{\sigma}$ es un estimador de escala.

Derivando respecto a $\mathbf{b}$, obtenemos

$$\sum_{i=1}^{n} \psi \left(\frac{r_i(\mathbf{b})}{\hat{\sigma}} \right) \mathbf{x}_i = 0 , \tag{3.21}$$

donde $\psi = \rho'$. Como antes, esta última ecuación no tiene que ser necesariamente la ecuación de un estimador de máxima verosimilitud. Asumimos que la matriz de diseño $\mathbf{X}$ tiene rango completo. En el caso particular en que σ es conocido, se puede verificar que el M-estimador es de regresión, invariante por traslaciones y equivariante por cambio de escala.

Las soluciones de (3.21) con ψ monótona (respectivamente redescendiente) son llamadas *M-estimadores monótonos de regresión* (respectivamente *redescendientes*).

Las soluciones de (3.20) son soluciones de (3.21) y si ψ es estrictamente creciente, la solución es única. En el caso de regresión lineal como en el modelo de posición, los M-estimadores redescendientes tienen un mejor balance entre eficiencia y robustez que los M-estimadores monótonos. Por este motivo, en general se utiliza un M-estimador monótono como punto inicial necesario para computar un M-estimador redescendiente.

Asumiendo que se cumple el modelo (3.18), con u tal que

$$E \left[\psi \left(\frac{u}{\sigma} \right) \right] = 0$$

y bajo condiciones de regularidad, $\hat{\boldsymbol{\beta}}_n$ es consistente a $\boldsymbol{\beta}$ y además

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{D} \mathrm{N}_p(0, v(\mathbf{X}\mathbf{X}')^{-1}) , \tag{3.22}$$

donde v esta dado por

$$v = \sigma^2 \frac{E[\psi(u/\sigma)^2]}{(E[\psi'(u/\sigma)])^2} \,. \tag{3.23}$$

Una demostración general puede encontarse en Maronna y Yohai (1979).

3.7. Modelo de regresión lineal con predictores aleatorios

Asumimos que observamos n vectores aleatorios *i.i.d.* $(\mathbf{x}_i', y_i)$ de dimensión p, que satisfacen el modelo

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i \,,$$

donde los errores u_i son *i.i.d.* e independientes de las covariables $\mathbf{x}_i$. Sea $\mathbf{x}$ un vector aleatorio con la misma distribución que las covariables $\mathbf{x}_i$. En este contexto, el análogo de asumir que $\mathbf{X}$ tiene rango completo, es asumir que la distribución de $\mathbf{x}$ no se concentra en ningún subespacio, es decir, $P(\mathbf{a}'\mathbf{x} = 0) < 1 \ \forall \mathbf{a} \neq 0$.

Bajo estas condiciones, suponiendo que las $\mathbf{x}_i$ tienen varianza finita y que $\hat{\sigma}$ es consistente a σ, se puede probar que el estimador $\hat{\boldsymbol{\beta}}_n$ definido en (3.20) es consistente y asintóticamente normal, más aún

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{D} \mathrm{N}_p(0, v\mathrm{V}_{\mathbf{x}}^{-1}) \,,$$

donde $\mathrm{V}_{\mathbf{x}} = E(\mathbf{x}\mathbf{x}')$ y v fue definido en (3.23).

3.8. Modelo de regresión no lineal

Consideremos un modelo de regresión no lineal dado por

$$y_i = g(\mathbf{x}_i, \boldsymbol{\theta}) + \varepsilon_i, \quad 1 \leq i \leq n \,, \tag{3.24}$$

donde $\mathbf{x}_i$ son los vectores de variables regresoras, $\boldsymbol{\theta}$ es el vector de parámetros desconocidos a ser estimados, ε_i representan los errores y n el tamaño de la muestra. Dado $\mathbf{t} \in \Theta$, siendo Θ el espacio paramétrico, definimos los residuos correspondientes a $\mathbf{t}$ como

$$r_i(\mathbf{t}) = y_i - g(\mathbf{x}_i, \mathbf{t}) \,.$$

A continuación presentamos dos familias de estimadores bajo el marco de la regresión no lineal: los M-estimadores y los LMS-estimadores.

3.8.1. M-estimadores

Supongamos que se cumple el modelo (3.24) donde ε_i tiene densidad

$$\frac{1}{\sigma} f_0\left(\frac{\varepsilon}{\sigma}\right) \,,$$

y σ es un parámetro de escala. En este modelo, las y_i son independientes pero no idénticamente distribuidas, de hecho cada y_i tiene densidad

$$\frac{1}{\sigma} f_0 \left(\frac{y_i - g(\mathbf{x}_i, \boldsymbol{\theta})}{\sigma} \right)$$

y la función de verosimilitud para $\boldsymbol{\theta}$, fijando σ, es

$$L(\boldsymbol{\theta}) = \frac{1}{\sigma^n} \prod_{i=1}^{n} f_0 \left(\frac{y_i - g(\mathbf{x}_i, \boldsymbol{\theta})}{\sigma} \right).$$

Calcular el EMV maximizando $L(\mathbf{t})$ es equivalente a encontrar $(\hat{\boldsymbol{\theta}}, \hat{\sigma})$ tal que minimice

$$\frac{1}{n} \sum_{i=1}^{n} \rho_0 \left(\frac{r_i(\mathbf{t})}{\sigma} \right) + \log(\sigma),$$

donde $\rho_0 = -\log f_0$. Luego, derivando respecto de $\mathbf{t}$ y dejando fijo el valor de σ, obtenemos

$$\sum_{i=1}^{n} \psi_0 \left(\frac{r_i(\mathbf{t})}{\sigma} \right) \frac{\partial g_i(\mathbf{t})}{\partial t_r} = 0, \quad 1 \leq r \leq p,$$

donde $\psi_0 = \rho_0' = -f_0'/f_0$. Definimos a los *M-estimadores de regresión* como

$$\hat{\boldsymbol{\theta}}_n = \arg \min_{\mathbf{t} \in \Theta} \sum_{i=1}^{n} \rho \left(\frac{y_i - g(\mathbf{x}_i, \mathbf{t})}{\hat{\sigma}} \right), \tag{3.25}$$

donde ρ es una función-ρ como la de la Definición 3.1 y $\hat{\sigma}$ es un estimador de la escala σ de los residuos, es decir,

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{r_i(\mathbf{t})}{\hat{\sigma}} \right) = \delta,$$

siendo δ una constante entre 0 y 1. Diferenciando (3.25) obtenemos las ecuaciones

$$\sum_{i=1}^{n} \psi \left(\frac{r_i(\mathbf{t})}{\hat{\sigma}} \right) \frac{\partial g_i(\mathbf{t})}{\partial t_r} = 0, \quad 1 \leq r \leq p, \tag{3.26}$$

donde $\psi = \rho'$ y es una función-ψ como la de la Definición 3.2.

Fasano (2009) estudia el comportamiento asintótico de estos estimadores bajo condiciones de regularidad, en particular, deriva su distribución asintótica bajo los supuestos que detallamos a continuación.

A fin de que los parámetros sean identificables, asume que la función de regresión g satisface la siguiente condición

$$\mathbf{t} \neq \boldsymbol{\theta} \Rightarrow P\{g(\mathbf{x}, \mathbf{t}) = g(\mathbf{x}, \boldsymbol{\theta})\} < 1.$$

Por otro lado, supongamos que $\mathbf{x}$, y y ε tienen la misma distribución que $\mathbf{x}_i$, y_i y ε_i, respectivamente. Sea $G_0(\mathbf{x})$ la distribución de las $\mathbf{x}$ y $F_0(\varepsilon)$ la de ε, entonces la distribución de $\mathbf{z} = (\mathbf{x}, y)$ está dada por

$$H_0(\mathbf{z}) = G_0(\mathbf{x}) F_0(y - g(\mathbf{x}, \boldsymbol{\theta})).$$

Consideremos los siguientes supuestos

A. $E(|\rho(y - g(\boldsymbol{\theta}))|) < \infty, \ \forall \boldsymbol{\theta} \in \Theta$

B. La distribución F_0 tiene densidad f_0 con las siguientes propiedades:

i) f_0 es par.
ii) $f_0(|\varepsilon|)$ monótona creciente.
iii) $f_0(|\varepsilon|)$ estrictamente decreciente en un entorno de 0.

Sea $\dot{\mathbf{g}}$ el gradiente de g, $\psi = \rho'$ y tomemos, sin pérdida de generalidad, $\sigma = 1$, luego se tiene que $\hat{\boldsymbol{\theta}}_n$ es solución de

$$\sum_{i=1}^{n} \psi(y_i - g(\mathbf{x}_i, \boldsymbol{\theta}))\dot{\mathbf{g}}(\mathbf{x}_i, \boldsymbol{\theta}) = 0\,.$$

Bajo las condiciones **A** y **B** y la suposición de que, o bien, Θ es compacto o se cumple que

$$P\left\{\sup_{\boldsymbol{\theta}} |g(\mathbf{x}, \boldsymbol{\theta})| < \infty\right\} = 1\,, \tag{3.27}$$

Fasano (2009) prueba que el M-estimador $\hat{\boldsymbol{\theta}}_n$ definido en (3.25) es consistente a $\boldsymbol{\theta}$. Si además, las dos primeras derivadas de g son F_0-integrables, la matriz $\mathrm{V} = E[\dot{\mathbf{g}}(\mathbf{x}, \boldsymbol{\theta})\dot{\mathbf{g}}(\mathbf{x}, \boldsymbol{\theta})']$ es definida positiva con probabilidad uno y la función ψ es continua, acotada y tiene derivada acotada, entonces

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} \mathrm{N}\left(0, \frac{E[\psi^2(\varepsilon)]}{(E[\psi'(\varepsilon)])^2}\mathrm{V}^{-1}\right)\,.$$

3.8.2. LMS-estimadores

Rousseeuw (1984), basándose en algunas ideas de Hampel (1975), definió el primer estimador de regresión con el mayor punto de ruptura posible, $1/2$. Es decir, el mismo resiste los efectos de aproximadamente un 50 % de datos contaminados, que es lo mejor que se puede esperar de un estimador. Este es el llamado *LMS-estimador* y denotado $\hat{\boldsymbol{\theta}}_{\mathrm{LMS}}$.

Consideremos nuevamente el modelo de regresión lineal dado por

$$y_i = \mathbf{x}_i{}'\boldsymbol{\theta} + \varepsilon_i, \quad 1 \leq i \leq n\,,$$

donde los $\mathbf{x}_i$ son vectores p-dimensionales de variables explicativas y los ε_i son independientes con distribución F, simétrica y fuertemente unimodal. Se define $\hat{\boldsymbol{\theta}}_{\mathrm{LMS}}$ como

$$\hat{\boldsymbol{\theta}}_{\mathrm{LMS}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \operatorname*{med}_{1\leq i\leq n} (y_i - \mathbf{x}_i{}'\boldsymbol{\theta})^2,$$

donde la mediana es definida como el $[\![n/2]\!] + [\![(p+1)/2]\!]$ estadístico de orden y $[\![\cdot]\!]$ es la parte entera.

Rousseeuw probó que $\hat{\boldsymbol{\theta}}_{\mathrm{LMS}}$ siempre existe y que cualquier conjunto de p observaciones determina un único valor de $\hat{\boldsymbol{\theta}}_{\mathrm{LMS}}$. Además, para una muestra finita, el punto de ruptura del método LMS en el caso de regresión lineal es $([\![(n-p)/2]\!] + 1)/n$.

Kim y Pollard (1990) probaron que el orden de convergencia para el estimador LMS en regresión lineal es $O_p(n^{-1/3})$, lo cual implica que la eficiencia asintótica de este estimador es 0. Es por esto que, generalmente, este estimador se utiliza solamente para detectar un posible problema de outliers o de enmascaramiento que otras técnicas de diagnóstico pasan por alto (Rousseeuw y van Zomeren, 1990) y no para inferencia. O, alternativamente, como primer paso a la hora de aplicar un estimador robusto más eficiente como, por ejemplo, un M-estimador (Jureckova y Portnoy, 1987 y Simpson, Ruppert, y Carroll, 1991).

La definición del LMS-estimador puede ser fácilmente generalizada para su uso en modelos de regresión no lineal de la siguiente manera

$$\hat{\boldsymbol{\theta}}_{\text{LMS}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \operatorname*{med}_{1\leq i\leq n} (y_i - g(\mathbf{x}_i, \boldsymbol{\theta}))^2. \tag{3.28}$$

Sin embargo, deben tenerse en cuenta varios temas importantes que justifiquen su uso en la práctica. Stromberg y Ruppert (1992) analizaron las propiedades del punto de ruptura en el caso de regresión no lineal y Stromberg (1995) da una demostración de la consistencia débil del $\hat{\boldsymbol{\theta}}_{\text{LMS}}$ en modelos de regresión no lineal.

Aspectos computacionales

Rousseeuw y Leroy (1987) utilizaron el algoritmo PROGRESS para aproximar al $\hat{\boldsymbol{\theta}}_{\text{LMS}}$. Este algoritmo puede resumirse de la siguiente manera: en un primer paso se calcula el ajuste exacto a p puntos, denotamos esto $\hat{\boldsymbol{\theta}}_{\text{ex}}$ y luego se calcula la mediana residual en $\hat{\boldsymbol{\theta}}_{\text{ex}}$. Lo ideal es repetir este procedimiento para los $\binom{n}{p}$ posibles subconjuntos de p-elementos y el valor de $\hat{\boldsymbol{\theta}}_{\text{ex}}$ que produzca la menor mediana residual utilizarlo para encontrar el estimador LMS. Si repetir $\binom{n}{p}$ veces es computacionalmente difícil, Rousseeuw y Leroy sugieren un método diferente para elegir el número de submuestras: si la proporción de outliers es ξ, entonces el número de submuestras puede ser elegido para asegurar, con alta probabilidad, que al menos una de las submuestras no contiene ninguno de los outliers. Ellos notan que para n/p, esta probabilidad es aproximada por

$$1 - (1 - (1 - \xi)^p)^k, \tag{3.29}$$

donde k es el número de submuestras.

Sugieren que k debería ser elegido para asegurar que (3.29) sea al menos 0.95, pero en su algoritmo eligen k dependiendo de p de la siguiente manera

	p :	1	2	3	4	5	≥ 6
max	k :	500	1000	1500	2000	2500	3000

Elegir un gran número de submuestras podría requerir mucho tiempo de computación en regresión no lineal, por esto, Stromberg (1993) modificó el algoritmo PROGRESS para encontrar un estimador del estimador LMS en el marco de regresión no lineal.

Este algoritmo es un procedimiento en varias etapas. En cada etapa se procura mejorar el mejor estimador presente, denotado por $\hat{\boldsymbol{\theta}}$, de $\hat{\boldsymbol{\theta}}_{\text{LMS}}$.

El algoritmo introducido por Stromberg (1993) se resume de la siguiente manera:

Paso 0: el $\hat{\boldsymbol{\theta}}$ inicial es el estimador de mínimos cuadrados para la muestra completa.

Paso 1: Se computa el estimador de mínimos cuadrados, $\hat{\boldsymbol{\theta}}_{\text{LS}}$, para p puntos elegidos aleatoriamente. Si el cuadrado de la mediana residual en $\hat{\boldsymbol{\theta}}_{\text{LS}}$ es menor que el cuadrado de la mediana residual en $\hat{\boldsymbol{\theta}}$, el $\hat{\boldsymbol{\theta}}_{\text{LS}}$ reemplaza $\hat{\boldsymbol{\theta}}$ como el estimador actual de $\hat{\boldsymbol{\theta}}_{\text{LMS}}$. Este procedimiento se repite k veces, donde k es especificado por el usuario. El método por defecto para computar $\hat{\boldsymbol{\theta}}_{\text{LS}}$ es el método de Newton-Raphson con punto inicial $\hat{\boldsymbol{\theta}}$, pero si $\hat{\boldsymbol{\theta}}_{\text{LS}}$ puede ser hallado algebraicamente es preferible hacerlo ya que se gana tiempo de cómputos. La cantidad de ajustes de mínimos cuadrados se elige tal que (3.29) sea al menos 0.999 cuando $\xi = 50\,\%$.

Paso 2: Este paso toma ventaja del hecho de que $\hat{\boldsymbol{\theta}}_{\text{LMS}}$ está básicamente tratando de encontrar un buen ajuste a la mitad de la muestra. $\hat{\boldsymbol{\theta}}$ se usa como valor inicial para calcular el ajuste de mínimos cuadrados, denotado $\hat{\boldsymbol{\theta}}^*_{\text{LS}}$, para los puntos de la muestra tal que $r_i^2(\hat{\boldsymbol{\theta}}) \leq \underset{1\leq i\leq n}{\text{med}}\, r_i^2(\hat{\boldsymbol{\theta}})$. Si $\underset{1\leq i\leq n}{\text{med}}\, r_i^2(\hat{\boldsymbol{\theta}}^*_{\text{LS}}) < \underset{1\leq i\leq n}{\text{med}}\, r_i^2(\hat{\boldsymbol{\theta}})$, entonces $\hat{\boldsymbol{\theta}}^*_{\text{LS}}$ reemplaza a $\hat{\boldsymbol{\theta}}$ como el estimador actual de $\hat{\boldsymbol{\theta}}_{\text{LMS}}$.

Paso 3: Para encontrar aún un mejor estimador de $\hat{\boldsymbol{\theta}}_{\text{LMS}}$, el algoritmo simplex de Nelder-Mead (Nelder y Mead, 1965), se utiliza para minimizar $\underset{1\leq i\leq n}{\text{med}}\, r_i^2(\boldsymbol{\theta})$, usando $\hat{\boldsymbol{\theta}}$ como valor inicial. El algoritmo se implementa como en Press, Flannery, Teukolsky, and Vetterling (1986), con tolerancia fraccionaria 10^{-4}.

Este algoritmo se utilizó para computar los estimadores iniciales en nuestro estudio de Monte Carlo que desarrollaremos a continuación en el Capítulo 5.

Capítulo 4

Estimación robusta en el modelo de regresión no lineal con respuestas faltantes

4.1. Introducción

Como fue mencionado en los capítulos anteriores, en los modelos no lineales asumimos que observamos n vectores aleatorios $(y_i, \mathbf{x}_i) \in \mathbb{R}^{p+1}$ independientes e idénticamente distribuidos (*i.i.d.*), con $y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^p$, siendo

$$y_i = g(\mathbf{x}_i, \boldsymbol{\theta}) + \varepsilon_i, \qquad 1 \leq i \leq n\,,$$

donde los errores ε_i son variables *i.i.d.* e independientes de x_i y g una función conocida salvo por el parámetro $\boldsymbol{\theta}$ a estimar. En la teoría clásica, se supone que con $E(\varepsilon_i) = 0$ y $Var(\varepsilon_i) = \sigma^2$. En nuestro contexto, asumiremos que los errores tienen distribución simétrica alrededor del 0 con densidad

$$\frac{1}{\sigma} f_0\left(\frac{\varepsilon}{\sigma}\right),$$

siendo σ un parámetro de escala.

Nuestro objetivo es estimar el parámetro $\boldsymbol{\theta}$ cuando existen respuestas faltantes en el conjunto de datos a tratar, que queda definido a partir de $(y_i, \mathbf{x}_i, \delta_i)$, $1 \leq i \leq n$ donde $\delta_i = 1$ si y_i es observada y $\delta_i = 0$ si no lo es. Sea $(Y, \mathbf{X}, \delta)$ un vector aleatorio con la misma distribución que $(y_i, \mathbf{x}_i, \delta_i)$, asumiremos que la variable de respuesta presenta observaciones faltantes al azar (missing at random, MAR), es decir, dado $\mathbf{X}$, δ e Y son condicionalmente independientes

$$P(\delta = 1|(Y, \mathbf{X})) = P(\delta = 1|\mathbf{X}) = p(\mathbf{X})\,. \tag{4.1}$$

Como las estimaciones mediante métodos clásicos, como el de mínimos cuadrados, son sensibles a la presencia de datos atípicos, resulta necesaria la utilización de métodos robustos. En este contexto, Sued y Yohai (2012) estiman la distribución marginal de la respuesta mediante

un procedimiento que permite estimar en forma consistente cualquier funcional débilmente continuo de la distribución de la respuesta Y. Para este propósito utilizan como estimador robusto inicial del parámetro de regresión no lineal $\boldsymbol{\theta}$, MM-estimadores, de los que estudian sus propiedades asintóticas.

En el presente trabajo, proponemos estimar el parámetro de regresión no lineal cuando existen datos faltantes en la variable de respuesta mediante una familia de M-estimadores y estudiar algunas de sus propiedades.

4.2. M-estimadores

Consideremos la función de pérdida $\rho_c(t)$, donde ρ_c es una función-ρ como las definidas en el capítulo anterior, siendo c el parámetro de calibración. Definimos

$$S_n(\mathbf{t}) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \, \rho_c\left(\frac{y_i - g(\mathbf{x}_i, \mathbf{t})}{\hat{\sigma}}\right), \tag{4.2}$$

donde $\hat{\sigma}$ es un estimador preliminar de la escala σ.

Luego, definimos el M-estimador simplificado $\hat{\boldsymbol{\theta}}$ del parámetro de regresión como

$$\hat{\boldsymbol{\theta}} = \arg\min_{\mathbf{t}} S_n(\mathbf{t})\,. \tag{4.3}$$

Cuando ρ es continuamente diferenciable, con $\psi(y,u,t) = \partial\rho(y,u,t)/\partial u$, $\hat{\boldsymbol{\theta}}$ satisface las ecuaciones diferenciales

$$S_n^{(1)}(\mathbf{t}) = \frac{1}{n}\sum_{i=1}^{n} \delta_i \, \psi_c\left(\frac{y_i - g(\mathbf{x}_i, \mathbf{t})}{\hat{\sigma}}\right) \frac{\partial g(\mathbf{x}_i, \mathbf{t})}{\partial \mathbf{t}} = 0\,. \tag{4.4}$$

Un aspecto crítico en todo problema no lineal de estimación es la elección de los puntos iniciales del algoritmo de cómputo. Esto ocurre tanto si se intenta buscar el mínimo de (4.2) como si se intenta resolver (4.4). Por esta razón, proponemos utilizar como punto inicial en la búsqueda del M-estimador, el estimador LMS computado mediante una adaptación del algoritmo basado en la propuesta de Stromberg (1993), que a la vez permite calcular un estimador preliminar robusto de la escala σ.

Proponemos el siguiente procedimiento para la estimación del parámetro de regresión:

- **Paso 1:** Calcular el estimador de mínimos cuadrados tomando un valor inicial θ_0

$$\hat{\boldsymbol{\theta}}_{\text{LS}} = \arg\min_{\mathbf{t}\in\Theta} \sum_{i=1}^{n} \delta_i (y_i - g(\mathbf{x}_i, \mathbf{t}))^2.$$

- **Paso 2:** Calcular el LMS-estimador mediante el algoritmo de Stromberg (1993) tomando como valor inicial $\hat{\boldsymbol{\theta}}_{\text{LS}}$

$$\hat{\boldsymbol{\theta}}_{\text{LMS}} = \arg\min_{\mathbf{t}\in\Theta} \underset{\substack{\delta_i=1\\ 1\le i\le n}}{\text{med}} (y_i - g(\mathbf{x}_i, \mathbf{t}))^2.$$

- **Paso 3:** Para las observaciones completas, sean los residuos

$$r_{\text{LMS}i} = y_i - g(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_{\text{LMS}})\,.$$

Calcular

$$\hat{\sigma}_{\text{LMS}} = \kappa \cdot \underset{\substack{\delta_i=1\\1\le i\le n}}{\text{med}}\,(|r_{\text{LMS}i} - \underset{\substack{\delta_i=1\\1\le i\le n}}{\text{med}}\,(r_{\text{LMS}i})|),\ \kappa = 1.4826,$$

para utilizarla como escala preliminar en (4.4), siendo κ una constante de calibración.

- **Paso 4:** Estimar $\boldsymbol{\theta}$ mediante un M-estimador, tomando como punto inicial $\hat{\boldsymbol{\theta}}_{\text{LMS}}$, minimizando

$$\hat{\boldsymbol{\theta}}_{\text{M}} = \arg\,\underset{\boldsymbol{\theta}\in\Theta}{\text{mín}} \sum_{i=1}^{n} \delta_i\,\rho_c\left(\frac{y_i - g(\mathbf{x}_i, \boldsymbol{\theta})}{\hat{\sigma}_{\text{LMS}}}\right).$$

Observación 4.2. En el estudio de simulación tomamos como función-ρ la función bicuadrada de Tukey definida por

$$\rho(t) = \text{mín}\left\{1, 1 - (1 - t^2)^3\right\}\ .$$

Cabe mencionar que después del Paso 3 podría computarse un M-estimador de escala como los definidos en la Sección 3.3. de manera de utilizar en el Paso 4 un estimador de σ, $\hat{\sigma}$, que sea más eficiente.

4.2.1. Fisher-consistencia del parámetro $\boldsymbol{\theta}$

Consideremos el funcional asociado al M-estimador propuesto definido por

$$\boldsymbol{\theta}(F) = \arg\,\underset{\mathbf{t}\in\Theta}{\text{mín}}\ E\left[\delta\,\rho\left(\frac{y - g(\mathbf{x}, \mathbf{t})}{\sigma}\right)\right].$$

Veremos que bajo ciertas condiciones de regularidad el funcional es Fisher-consistente, es decir, si

$$S(\mathbf{t}) = E\left[\delta\,\rho\left(\frac{y - g(\mathbf{x}, \mathbf{t})}{\sigma}\right)\right],$$

entonces $\boldsymbol{\theta}(F) = \boldsymbol{\theta}$, en otras palabras, el verdadero valor del parámetro minimiza la función objetivo planteada en términos de la esperanza.

El siguiente lema establece condiciones suficientes para asegurar la Fisher-consistencia del funcional asociado al M-estimador.

Asumiremos la siguiente condición sobre la función de pérdida ρ.

A0. Dado $\sigma > 0$, tenemos que $E(\rho((\varepsilon - a)/\sigma)) > E(\rho(\varepsilon/\sigma))$.

Observación 4.1. El uso de funciones de pérdida acotadas para controlar el crecimiento de los residuos requiere condiciones más exigentes para obtener la unicidad, tales como simetría y unimodalidad de la distribución de los residuos. De hecho, sea ρ una función-ρ como las descriptas anteriormente y acotada. Si denotamos $\lambda(a,\tau) = E(\rho((y-a)/\tau))$, entonces ν se define como $\nu = \arg\min_a \lambda(a,\tau_0)$ con τ_0 la escala marginal. Por el Teorema 10.2 en Maronna, Martin y Yohai (2006), si y tiene una densidad f que es una función decreciente en $|y-\nu|$ y ρ es cualquier función-ρ, entonces, $\lambda(a,\tau)$ tiene un único mínimo en $a=\nu$ para cualquier $\tau > 0$.

Lema 4.1. *Supongamos que $(Y, \mathbf{X}, \delta)$ es tal que $Y = g(\mathbf{X},\boldsymbol{\theta}) + \varepsilon$, donde ε es independiente de $\mathbf{X}$ y satisface la condición* MAR *dada en* (4.1). *Si la función de pérdida ρ es una función-ρ que cumple la condición* **A0** *entonces, el funcional $\boldsymbol{\theta}(F) = \arg\min_{\mathbf{t}} E_F\left[\delta\rho\left(\frac{Y-g(\mathbf{X},\mathbf{t})}{\sigma}\right)\right]$ es Fisher–consistente.*

DEM. Dado que se cumple (4.1), tenemos que Y y δ son condicionalmente independientes dado $\mathbf{X}$, por lo tanto

$$\begin{aligned} E\left[\delta\rho\left(\frac{Y-g(\mathbf{X},\mathbf{t})}{\sigma}\right)\right] &= E\left[p(\mathbf{X})\rho\left(\frac{Y-g(\mathbf{X},\mathbf{t})}{\sigma}\right)\right] \\ &= E\left[p(\mathbf{X})\rho\left(\frac{Y-g(\mathbf{X},\boldsymbol{\theta})+g(\mathbf{X},\boldsymbol{\theta})-g(\mathbf{X},\mathbf{t})}{\sigma}\right)\right] \\ &= E\left[p(\mathbf{X})\rho\left(\frac{\varepsilon-(g(\mathbf{X},\mathbf{t})-g(\mathbf{X},\boldsymbol{\theta}))}{\sigma}\right)\right] \\ &= E\left[p(\mathbf{X})E\left\{\rho\left(\frac{\varepsilon-(g(\mathbf{X},\mathbf{t})-g(\mathbf{X},\boldsymbol{\theta}))}{\sigma}\right)\middle|\mathbf{X}\right\}\right] \end{aligned}$$

De la independencia entre los errores y las covariables y la condición **A0**, se concluye que

$$\begin{aligned} E\left\{\rho\left(\frac{\varepsilon-(g(\mathbf{X},\mathbf{t})-g(\mathbf{X},\boldsymbol{\theta}))}{\sigma}\right)\middle|\mathbf{X}\right\} &= E\left\{\rho\left(\frac{\varepsilon-(g(\mathbf{X},\mathbf{t})-g(\mathbf{X},\boldsymbol{\theta}))}{\sigma}\right)\middle|\mathbf{X}=\mathbf{x}\right\} \\ &= E\left\{\rho\left(\frac{\varepsilon-(g(\mathbf{x},\mathbf{t})-g(\mathbf{x},\boldsymbol{\theta}))}{\sigma}\right)\right\} \\ &> E\left\{\rho\left(\frac{\varepsilon}{\sigma}\right)\right\} \end{aligned}$$

Luego, $\boldsymbol{\theta}(F)$ es un estimador Fisher-consistente de $\boldsymbol{\theta}$. □

4.2.2. Función de Influencia

La *función de influencia* mide la robustez con respecto a un outlier y permite estudiar la robustez local. Puede pensarse como la derivada del estimador y bajo condiciones de regularidad, permite derivar la matriz de covarianza asintótica del mismo, de manera que provee un criterio racional para elegir la constante de calibración.

Sea $T(F)$ un funcional. La función de influencia de $T(F)$ se define como

$$\mathrm{IF}(\mathbf{w}_0, T, F) = \lim_{\epsilon \to 0} \frac{T(F_{\mathbf{w}_0,\epsilon}) - T(F)}{\epsilon},$$

donde $F_{\mathbf{w}_0,\epsilon} = (1-\epsilon)F + \epsilon\Delta_{\mathbf{w}_0}$ representa el modelo contaminado, siendo $\Delta_{\mathbf{w}_0}$ la medida de probabilidad que pone masa 1 en el punto $\mathbf{w}_0 = (y_0, \mathbf{x}_0', \delta_0)$.

Bajo condiciones de regularidad (Fernholz, 1983), tenemos la siguiente expansión

$$\sqrt{n}\left\{T(F_n) - T(F)\right\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{IF}\left(\mathbf{w}_i, T, F\right) + o_p(1),$$

siendo F_n la distribución empírica de $\mathbf{w}_i$, $1 \leq i \leq n$. Luego, la varianza asintótica del estimador puede expresarse como

$$\mathrm{AV}\left(T, F\right) = E_F\left\{\mathrm{IF}\left(\mathbf{w}_1, T, F\right)\mathrm{IF}\left(\mathbf{w}_1, T, F\right)'\right\}. \tag{4.5}$$

Sea F_1 una distribución sobre $\mathbb{R}^{k+1}\times\{0,1\}$ y denotemos por $\boldsymbol{\theta}(F_1)$ y $\sigma(F_1)$ los funcionales asociados a $\hat{\boldsymbol{\theta}}$ y $\hat{\sigma}$, respectivamente.

Asumimos que $\boldsymbol{\theta}(F_1)$ es solución de $S^{(1)}(\mathbf{b}, \tau(F_1), F_1) = \mathbf{0}$ donde

$$S^{(1)}(\mathbf{t}, u, F_1) = E_{F_1}\left(\delta\,\psi\left(\frac{Y - g(\mathbf{X}, \mathbf{t})}{u}\right)\frac{\partial g(\mathbf{X}, \mathbf{t})}{\partial \mathbf{t}}\right).$$

Supongamos, además, que $\boldsymbol{\theta}(F_1)$ es un funcional Fisher-consistente en F, es decir, $\boldsymbol{\theta}(F) = \boldsymbol{\theta}$. En la siguiente proposición damos la función de influencia del funcional simplificado $\boldsymbol{\theta}(F_1)$ en $F_1 = F$.

Proposición 4.1. *Supongamos que IF$(\mathbf{w}_0, \sigma, F)$ existe y que son válidas las siguientes condiciones*

A1. *$\psi(s)$ es continuamente diferenciable, $\dfrac{\partial\psi\left(y, x, \mathbf{t}, u\right)}{\partial \mathbf{t}}$ es continua y ε tiene distribución simétrica alrededor del 0.*

A2. *La matriz*

$$\mathbf{A} = E\left(\psi'\left(\frac{\varepsilon}{\sigma}\right)\right)E\left[\frac{\partial g(\mathbf{X}, \boldsymbol{\theta})}{\partial \mathbf{t}}\frac{\partial g(\mathbf{X}, \boldsymbol{\theta})'}{\partial \mathbf{t}}p(\mathbf{X})\right]$$

es no singular.

Luego, IF$\left(\mathbf{w}_0, \boldsymbol{\theta}, F\right)$ existe y cuando $\sigma(F) = \sigma$, tenemos que

$$\mathit{IF}(\mathbf{w}_0, \boldsymbol{\theta}, F) = -\,\sigma\delta_0\psi\left(\frac{y_0 - g(\mathbf{x}_0, \boldsymbol{\theta}(F))}{\sigma}\right)\mathbf{A}^{-1}\frac{\partial g(\mathbf{x}_0, \boldsymbol{\theta}(F))}{\partial \mathbf{t}}.$$

DEM. Notemos que

$$E_{F_{\mathbf{w}_0,\epsilon}}\left(\delta\,\psi\left(\frac{Y-g(\mathbf{X},\boldsymbol{\theta}(F_{\mathbf{w}_0,\epsilon}))}{\sigma(F_{\mathbf{w}_0,\epsilon})}\right)\frac{\partial g(\mathbf{X},\boldsymbol{\theta}(F_{\mathbf{w}_0,\epsilon}))}{\partial \mathbf{t}}\right)=\mathbf{0}_k\,,$$

luego

$$\begin{aligned}\mathbf{0}_k &= (1-\epsilon)E_F\left(\delta\,\psi\left(\frac{Y-g(\mathbf{X},\boldsymbol{\theta}(F_{\mathbf{w}_0,\epsilon}))}{\sigma(F_{\mathbf{w}_0,\epsilon})}\right)\frac{\partial g(\mathbf{X},\boldsymbol{\theta}(F_{\mathbf{w}_0,\epsilon}))}{\partial \mathbf{t}}\right)\\ &+ \epsilon\,\delta_0\left(\psi\left(\frac{y_0-g(\mathbf{x}_0,\boldsymbol{\theta}(F_{\mathbf{w}_0,\epsilon}))}{\sigma(F_{\mathbf{w}_0,\epsilon})}\right)\frac{\partial g(\mathbf{x}_0,\boldsymbol{\theta}(F_{\mathbf{w}_0,\epsilon}))}{\partial \mathbf{t}}\right)\,. \qquad (4.6)\end{aligned}$$

Entonces, diferenciando (4.6) con respecto a ϵ y evaluando en $\epsilon=0$, resulta

$$\begin{aligned}\mathbf{0}_k &= -E_F\left(\delta\,\psi\left(\frac{Y-g(\mathbf{X},\boldsymbol{\theta}(F))}{\sigma(F)}\right)\frac{\partial g(\mathbf{X},\boldsymbol{\theta}(F))}{\partial \mathbf{t}}\right)\\ &+ E_F\left(\delta\,\psi\left(\frac{Y-g(\mathbf{X},\boldsymbol{\theta}(F))}{\sigma(F)}\right)\frac{\partial^2 g(\mathbf{X},\boldsymbol{\theta}(F))}{\partial \mathbf{t}^2}\right)\mathrm{IF}(\mathbf{w}_0,\boldsymbol{\theta},F)\\ &- E_F\left(\delta\,\psi'\left(\frac{Y-g(\mathbf{X},\boldsymbol{\theta}(F))}{\sigma(F)}\right)\frac{\partial g(\mathbf{X},\boldsymbol{\theta}(F))}{\partial \mathbf{t}}\frac{\partial g(\mathbf{X},\boldsymbol{\theta}(F))}{\partial \mathbf{t}}'\right)\frac{1}{\sigma(F)}\mathrm{IF}(\mathbf{w}_0,\boldsymbol{\theta},F)\\ &- E_F\left(\delta\,\psi'\left(\frac{Y-g(\mathbf{X},\boldsymbol{\theta}(F))}{\sigma(F)}\right)\frac{\partial g(\mathbf{X},\boldsymbol{\theta}(F))}{\partial \mathbf{t}}\frac{(Y-g(\mathbf{X},\boldsymbol{\theta}(F))}{\sigma(F)}\right)\frac{1}{\sigma(F)}\mathrm{IF}(\mathbf{w}_0,\sigma,F)\\ &+ \delta_0\left(\psi\left(\frac{y_0-g(\mathbf{x}_0,\boldsymbol{\theta}(F))}{\sigma(F)}\right)\frac{\partial g(\mathbf{x}_0,\boldsymbol{\theta}(F))}{\partial \mathbf{t}}\right)\,.\end{aligned}$$

Teniendo en cuenta que se cumple la condición de MAR dada en (4.1), que ε tiene distribución simétrica alrededor del 0, que ψ y $\psi'(s)$ son funciones impares, obtenemos que

$$\begin{aligned}\mathbf{0}_k &= -E_F\left(\delta\,\psi'\left(\frac{Y-g(\mathbf{X},\boldsymbol{\theta}(F))}{\sigma(F)}\right)\frac{\partial g(\mathbf{X},\boldsymbol{\theta}(F))}{\partial \mathbf{t}}\frac{\partial g(\mathbf{X},\boldsymbol{\theta}(F))}{\partial \mathbf{t}}'\right)\frac{1}{\sigma(F)}\mathrm{IF}(\mathbf{w}_0,\boldsymbol{\theta},F)\\ &+ \delta_0\left(\psi\left(\frac{y_0-g(\mathbf{x}_0,\boldsymbol{\theta}(F))}{\sigma(F)}\right)\frac{\partial g(\mathbf{x}_0,\boldsymbol{\theta}(F))}{\partial \mathbf{t}}\right)\,,\end{aligned}$$

de donde resulta la proposición. □

Vale la pena observar que la función de influencia obtenida depende de una variable dicotómica que toma el valor 0 cuando la variable de respuesta está ausente. Por esta razón, como en Bianco, Boente y Rodrigues (2012) consideramos la esperanza de la función de influencia que denotamos $\mathrm{EIF}(\mathbf{w}_0,T,F)$, de manera tal que

$$\mathrm{EIF}(\mathbf{w}_0^*,T,F)=\mathrm{E}(\mathrm{IF}(\mathbf{w}_0^*,T,F)|(y_0,\mathbf{x}_0))$$

Para el caso del funcional en estudio, obtenemos que

$$\mathrm{EIF}(\mathbf{w}_0^*,\boldsymbol{\theta},F)=-\,\sigma p(\mathbf{x}_0)\psi\left(\frac{y_0-g(\mathbf{x}_0,\boldsymbol{\theta}(F))}{\sigma}\right)\mathbf{A}^{-1}\frac{\partial g(\mathbf{x}_0,\boldsymbol{\theta}(F))}{\partial \mathbf{t}}\,.$$

Para ilustrar el comportamiento de la medida EIF, consideramos en primer lugar el modelo de crecimiento exponencial, como en Fasano (2009), dado por

$$y = \beta \exp(\alpha x) + \varepsilon,$$

donde $(\alpha, \beta) = (2, 5)$ y tomamos 3 modelos diferentes para la probabilidad de ausencia dados por

$p \equiv 1$: sin respuestas faltantes

$p(x) = \dfrac{1}{1 + \exp(-2x - 2)}$: modelo logístico

$p(x) = 0.7 + 0.2(\cos(2x + 0.4))^2$

En las Figuras 4.1, 4.2 y 4.3, graficamos las superficies obtenidas para el M-estimador cuando se utiliza la función de pérdida bicuadrada con constante de calibración $c = 4$ y la del estimador de mínimos cuadrados que resulta de tomar $\rho(s) = s^2$. En estas figuras, se puede apreciar el efecto de aplicar una función de pérdida ρ acotada sobre los residuos, pues mantiene dominada la influencia controlando el crecimiento de los residuos en una gran región donde la norma de la función de influencia de mínimos cuadrados toma valores muy elevados.

$\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$ $\hat{\boldsymbol{\theta}}_{\mathrm{M}}$

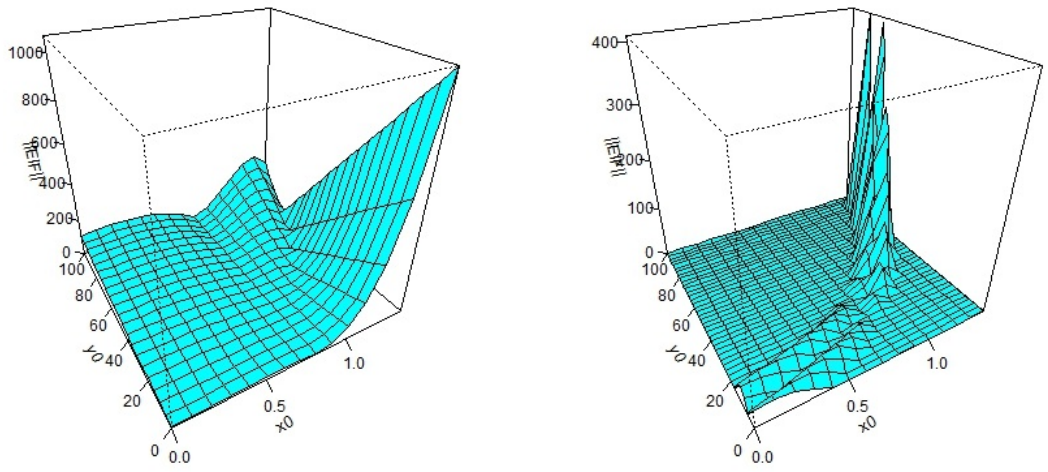

Figura 4.1: $\|EIF\|$ cuando $p \equiv 1$

$\hat{\boldsymbol{\theta}}_{\text{LS}}$ $\hat{\boldsymbol{\theta}}_{\text{M}}$

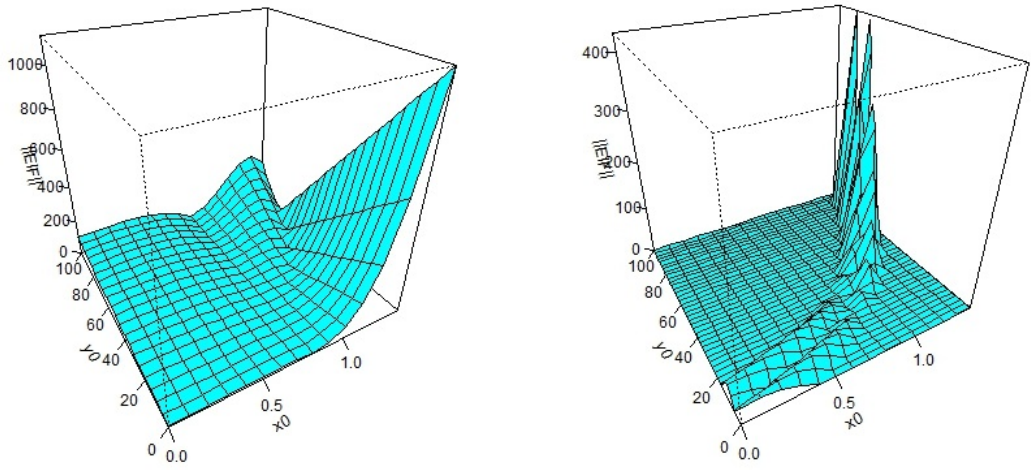

Figura 4.2: $\|EIF\|$ cuando $p(x) = \dfrac{1}{1+\exp(-2x-2)}$

$\hat{\boldsymbol{\theta}}_{\text{LS}}$ $\hat{\boldsymbol{\theta}}_{\text{M}}$

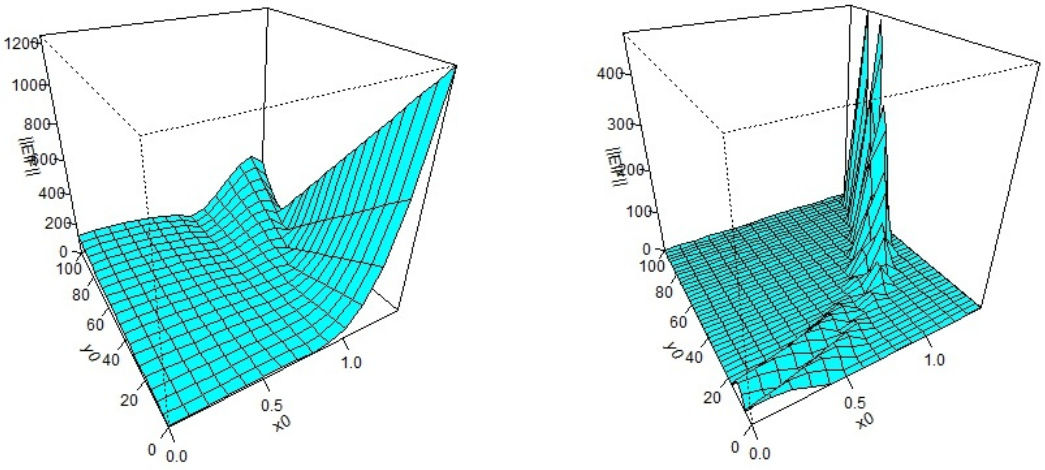

Figura 4.3: $\|EIF\|$ cuando $p(x) = 0.7 + 0.2(\cos(2x+0.4))^2$

En segundo término, consideramos el modelo de Michaelis–Menten parametrizado de la siguiente forma

$$y = \frac{\alpha\, x}{\beta + x} + \varepsilon,$$

con $(\alpha, \beta) = (10, 1)$, $x \sim U(0, 1)$ y $\varepsilon \sim N(0, 1)$ y las mismas probabilidades de ausencia que en el ejemplo anterior. En las Figuras 4.4, 4.5 y 4.6, graficamos las superficies obtenidas para el M-estimador cuando se utiliza la función de pérdida bicuadrada con constante de calibración $c = 4$ y la del estimador de mínimos cuadrados. En estas figuras, además de apreciar el efecto de aplicar una función de pérdida ρ acotada sobre los residuos, se observa claramente el efecto de la función de ausencia p sobre la esperanza de la función de influencia. En particular, en la Figura 4.6 se evidencia el efecto del coseno que introduce una fluctuación en la curva de influencia esperada.

$\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$ $\hat{\boldsymbol{\theta}}_{\mathrm{M}}$

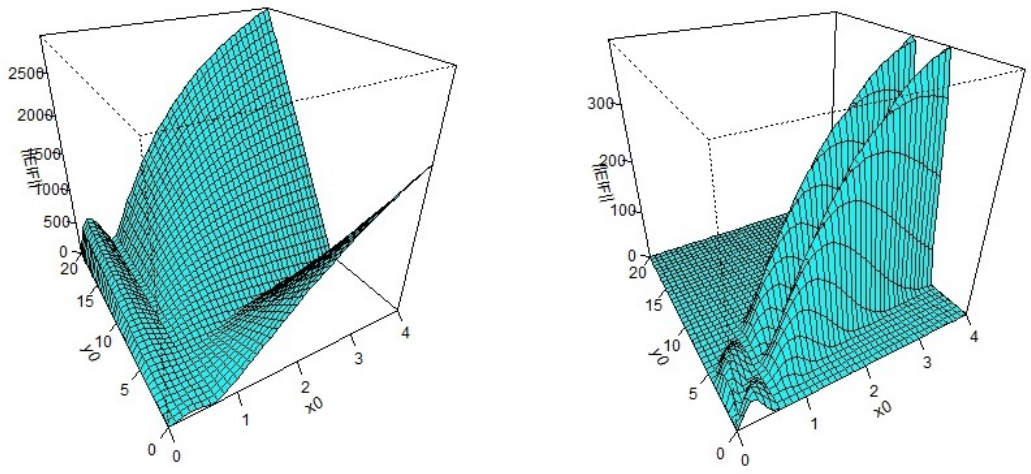

Figura 4.4: $\|EIF\|$ cuando $p \equiv 1$

$\hat{\boldsymbol{\theta}}_{\text{LS}}$ $\hat{\boldsymbol{\theta}}_{\text{M}}$

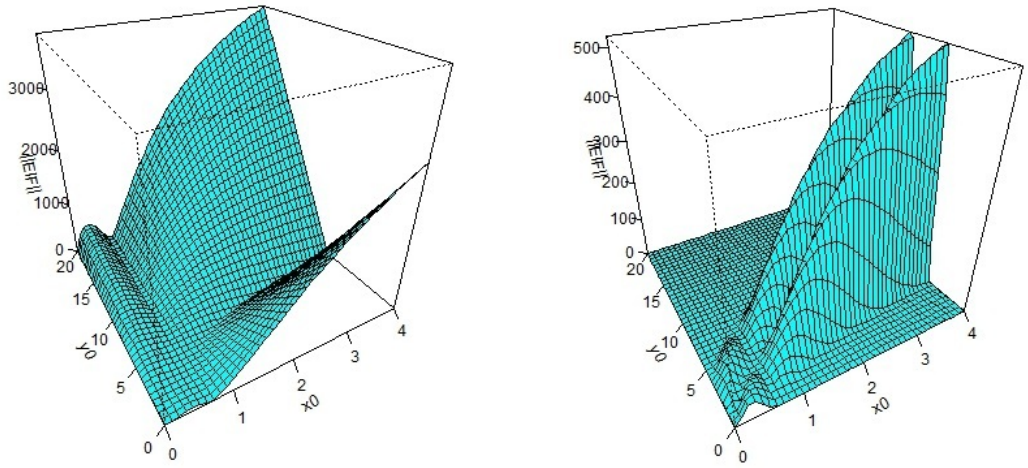

Figura 4.5: $\|EIF\|$ cuando $p(x) = \dfrac{1}{1+\exp(-2x-2)}$

$\hat{\boldsymbol{\theta}}_{\text{LS}}$ $\hat{\boldsymbol{\theta}}_{\text{M}}$

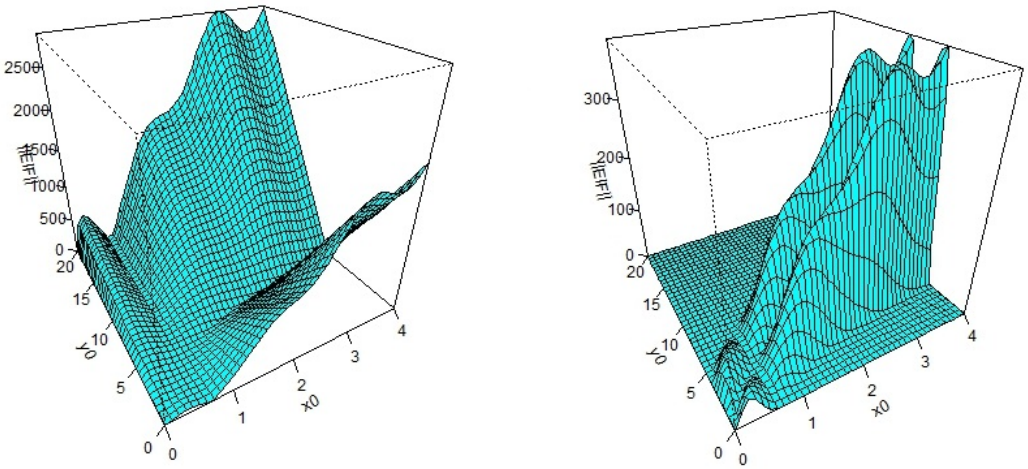

Figura 4.6: $\|EIF\|$ cuando $p(x) = 0.7 + 0.2(\cos(2x+0.4))^2$

Capítulo 5

Estudio de Monte Carlo

Con el fin de investigar el comportamiento del método bajo estudio en muestras finitas se realizó un estudio de simulación. Asimismo, en este análisis se compara el comportamiento del estimador clásico de mínimos cuadrados con la alternativa robusta propuesta.

Se consideran los modelos **Exponencial** y de **Michaelis-Menten**. Para ambos modelos se plantearon diferentes escenarios de análisis, considerando muestras con y sin datos atípicos y diferentes patrones de ausencia o pérdida de respuestas.

Las cuatro funciones de ausencia consideradas son:

- $p \equiv 1$: sin respuestas faltantes
- $p \equiv 0.8$: faltante de respuestas completamente al azar
- $p(x) = \dfrac{1}{1+\exp(-2x-2)}$: modelo logístico
- $p(x) = 0.7 + 0.2(\cos(2x+0.4))^2$

Para el cálculo de los M-estimadores se tomó como función-ρ la función bicuadrada cuya expresión se encuentra en (3.9). En cuanto a la constante c, elegimos $c = 3.25$ y $c = 4$, que corresponden a un valor de eficiencia de 0.82 y 0.90 respectivamente, bajo el modelo con datos completos. Sin embargo, mostraremos los resultados correspondientes a $c = 4$, ya que no hallamos diferencias importantes.

5.1. Modelo Exponencial

Se consideró el modelo de crecimiento exponencial analizado en el estudio de simulación de Fasano (2009), dado por

$$y_i = \beta \exp(\alpha x_i) + \varepsilon_i, \quad 1 \leq i \leq n\,, \tag{5.1}$$

con $\theta_0 = (\alpha_0, \beta_0) = (2, 5)$. La simulación se llevó a cabo generando 1000 muestras independientes de tamaño $n = 100$, siendo $x_i \sim U(0,1)$ y $\varepsilon_i \sim N(0,1)$.

En función de comparar el comportamiento de los estimadores bajo pérdida de datos, se consideraron las cuatro funciones de ausencia p antes mencionadas, así como muestras sin outliers y muestras contaminadas con un 10 % de datos atípicos. Si bien se consideraron diversas contaminaciones, mostraremos a modo de ejemplo una de ellas, que ejemplifica las situaciones consideradas. Se mostrarán los resultados correspondientes a datos contaminados de la siguiente forma

$$y_j = (\beta \exp(\alpha x_j)) \cdot 1.5 \qquad 90 \leq j \leq 100$$

donde $x_j = 1.09 + 0.0001 \cdot Z$, siendo Z un número aleatorio entre -1 y 2. Este modelo de contaminación corresponde al que introdujo un mayor crecimiento en el error cuadrático medio en el estudio numérico de Fasano (2009).

Una vez obtenidos todos los estimadores de α y β se midió su performance mediante las siguientes medidas: Media, Mediana, Varianza, MAD y MSE (error cuadrático medio), que presentamos en las siguientes tablas. Notaremos por LS.cg y LS.nlm al estimador clásico de mínimos cuadrados calculado mediante dos algoritmos diferentes. Por otra parte, indicaremos por LMS y M.cg al LMS-estimador y al M-estimador, respectivamente. Cabe destacar que si bien se consideraron diferentes algoritmos para la minimización del estimador de mínimos cuadrados y del M-estimador, sólo mostraremos los resultados obtenidos para dos de los estimadores calculados en el caso del LS y uno en el caso del M-estimador, ya que son representativos del resto.

En las Tablas 5.1 y 5.2 se presentan los valores obtenidos para α y β respectivamente, cuando no se consideran muestras sin contaminar y en las Tablas 5.3 y 5.4 cuando se introducen datos atípicos en la muestra. En cada tabla se tomó $c = 4$ y se tuvo en cuenta cada una de las funciones de pérdida, nombradas anteriormente.

En la Figura 5.1 se muestran los boxplots de los estimadores de α y β cuando $p \equiv 1$, tanto para muestras sin datos atípicos como para muestras contaminadas. Análogamente, en las Figuras 5.2, 5.3 y 5.4 se presentan los mismos boxplots para las funciones $p \equiv 0.8$, $p(x) = [1 + \exp(-2x - 2)]^{-1}$ y $p(x) = 0.7 + 0.2(\cos(2x + 0.4))^2$, respectivamente.

5.2. Modelo de Michaelis-Menten

En segunda instancia, consideramos el modelo de Michaelis-Menten, según la parametrización de Ratkowsky (1983) dado por

$$y_i = \frac{\alpha x_i}{\exp(\beta) + x_i} + \varepsilon_i \,. \tag{5.2}$$

siguiendo los parámetros de simulación tomados en Stromberg (1993). La simulación se llevó a cabo generando 1000 muestras independientes de tamaño $n = 100$, siendo $\theta_0 = (\alpha_0, \beta_0) = (10, 0)$, $x_i \sim U(0, 10)$ y $\varepsilon_i \sim N(0, 1)$.

Para comparar el comportamiento de los estimadores bajo pérdida de datos, se consideraron las cuatro funciones de p, muestras sin outliers y muestras contaminadas con un 20 %

de datos atípicos que fueron generados a la manera de Stromberg (1993), primero ordenando los datos de menor a mayor según las variables regresoras x_i y luego sumando 20 a los correspondientes valores de respuesta y_i.

Se consideraron los mismos estimadores que para el modelo de crecimiento exponencial. Al igual que en ese caso, una vez obtenidos todos los estimadores de α y β se midió su performance mediante la Media, Mediana, Varianza, MAD y MSE (error cuadrático medio), que presentamos a continuación en las siguientes tablas. Notaremos por LS.cg y LS.nlm al estimador clásico de mínimos cuadrados calculado mediante dos algoritmos diferentes. Por otra parte, indicaremos por LMS y M.cg al LMS-estimador y al M-estimador, respectivamente. Nuevamente, cabe destacar que sólo mostraremos los resultados obtenidos para dos de los estimadores calculados en el caso del LS y uno en el caso del M-estimador, ya que los valores que se obtuvieron con el resto de los algoritmos eran prácticamente idénticos.

En las Tablas 5.5 y 5.6 se presentan los valores obtenidos para α y β respectivamente, cuando no se consideran datos atípicos dentro de la muestra y en las Tablas 5.7 y 5.8 cuando se contaminan las muestras.

En la Figura 5.5 se muestran los boxplots de los estimadores de α y β cuando $p \equiv 1$, tanto para muestras sin datos atípicos como para muestras con datos atípicos. Análogamente, en las Figuras 5.6, 5.7 y 5.8 se presentan los boxplots correspondientes a las funciones $p \equiv 0.8$, $p(x) = [1 + \exp(-2x - 2)]^{-1}$ y $p(x) = 0.7 + 0.2(\cos(2x + 0.4))^2$, respectivamente.

5.3. Conclusiones

Tal como es de esperar, para el modelo exponencial como para el modelo de Michaelis-Menten, al considerar muestras sin datos atípicos observamos que la performance de los estimadores clásicos es mejor que la de los robustos en cuanto a la dispersión de las estimaciones. Sin embargo, no hay una pérdida importante al comparar los errores cuadráticos medios y al comparar media y mediana, son muy similares. Además, este comportamiento es parecido a través de las cuatro funciones de p que se han planteado.

En este estudio de Monte Carlo hemos podido comprobar en los dos modelos que el comportamiento del estimador clásico de mínimos cuadrados es muy inestable ante la presencia de datos atípicos en los cuatro escenarios planteados para la función de p, mientras que los M-estimadores se muestran muy estables y por lo tanto, superan en perfomance al estimador clásico, aún cuando se introducen 20 % de observaciones atípicas como en el caso del modelo de Michaelis-Menten. Más aún, vemos que hay diferencias entre el comportamiento de los estimadores de mínimos cuadrados según el método que se utilice para su cómputo cuando las muestras son contaminadas.

En el modelo de Michaelis-Menten se refleja más claramente el pobre comportamiento del estimador clásico, el cual se ve sumamente afectado por la presencia de datos atípicos. Los estimadores obtenidos distan considerablemente de los valores reales de los parámetros, en especial del parámetro α, los cuales presentan estimaciones de gran magnitud. En este caso, uno de los algoritmos empleados en la estimación por el método de mínimos cuadrados (LS.nlm) mejora la performance del estimador respecto del otro (LS.cg) pero, aún así, estima

valores muy alejados del valor real del parámetro. Por el contrario, los estimadores robustos no se ven afectados, sea cual fuere la función p considerada.

Respecto al LMS-estimador, podemos concluir, que es resistente a los datos atípicos debido a su alto punto de ruptura. Sin embargo, dado que no son altamente eficientes, su cálculo no resulta satisfactorio como estimador final del parámetro $\boldsymbol{\theta}$, pero sí como estimador inicial en el primer paso del algoritmo para obtener el M-estimador, el cual, heredando los beneficios del alto punto de ruptura del LMS resulta robusto y, debido a la constante $c = 4$ considerada, también eficiente.

5.4. Tablas

	LS.cg	LS.nlm	LMS	M.cg	Función de pérdida
Media	1.9995	1.9995	2.0019	1.9993	
Mediana	1.9991	1.999	2.0003	1.9989	
Varianza	0.0007	0.0007	0.0044	0.0007	$p \equiv 1$
MAD	0.0245	0.0245	0.0657	0.0275	
MSE	0.0007	0.0007	0.0044	0.0007	
Media	2.0011	2.0011	1.9992	2.0011	
Mediana	2.0009	2.0009	1.9959	1.9995	
Varianza	0.0008	0.0008	0.0051	0.0009	$p \equiv 0.8$
MAD	0.0265	0.0265	0.0714	0.0297	
MSE	0.0008	0.0008	0.0051	0.0009	
Media	2.0009	2.0009	2.0007	2.0009	
Mediana	1.9997	1.9997	1.9991	1.9994	
Varianza	0.0007	0.0007	0.0049	0.0008	$p(x) = \frac{1}{1+\exp(-2x-2)}$
MAD	0.0256	0.0256	0.0689	0.0279	
MSE	0.0007	0.0007	0.0048	0.0008	
Media	2.0008	2.0008	1.9992	2.0007	
Mediana	2	2	1.9972	1.9998	
Varianza	0.0008	0.0008	0.0054	0.001	$p(x) = 0.7 + 0.2(\cos(2x+0.4))^2$
MAD	0.0278	0.0278	0.0756	0.0304	
MSE	0.0008	0.0008	0.0054	0.001	

Tabla 5.1: Estimación de α con $\alpha_0 = 2$ y $c = 4$, para el Modelo Exponencial, sin outliers.

	LS.cg	LS.nlm	LMS	M.cg	Función de pérdida
Media	5.0026	5.0027	4.9982	5.0031	
Mediana	5.0043	5.0043	4.994	5.0046	
Varianza	0.0105	0.0105	0.0692	0.0116	$p \equiv 1$
MAD	0.103	0.103	0.2744	0.1116	
MSE	0.0105	0.0105	0.0691	0.0116	
Media	4.9979	4.9979	5.0111	4.9979	
Mediana	4.9971	4.9971	5.0153	4.9999	
Varianza	0.0128	0.0128	0.0811	0.0145	$p \equiv 0.8$
MAD	0.1074	0.1074	0.2919	0.1199	
MSE	0.0128	0.0128	0.0812	0.0144	
Media	4.9983	4.9983	5.0064	4.9983	
Mediana	5.0009	5.0009	5.0025	5.0025	
Varianza	0.0108	0.0108	0.0791	0.0122	$p(x) = \frac{1}{1+\exp(-2x-2)}$
MAD	0.1018	0.1018	0.2789	0.1075	
MSE	0.0107	0.0107	0.0791	0.0122	
Media	4.9995	4.9995	5.012	4.9998	
Mediana	5	5	5.0205	4.9997	
Varianza	0.0133	0.0133	0.0879	0.0151	$p(x) = 0.7 + 0.2(\cos(2x+0.4))^2$
MAD	0.1134	0.1134	0.3056	0.1233	
MSE	0.0133	0.0133	0.0879	0.0151	

Tabla 5.2: Estimación de β con $\beta_0 = 5$ y $c = 4$, para el Modelo Exponencial, sin outliers.

	LS.cg	LS.nlm	LMS	M.cg	Función de pérdida
Media	2.8376	2.8352	2.0021	1.9998	
Mediana	2.837	2.8356	2.0015	1.9981	
Varianza	0.0023	0.0028	0.0044	0.0011	$p \equiv 1$
MAD	0.0471	0.0467	0.0638	0.0263	
MSE	0.7039	0.7004	0.0044	0.0011	
Media	2.8329	2.832	1.9999	2.0007	
Mediana	2.83	2.8298	1.9992	2.0005	
Varianza	0.0045	0.0046	0.0046	0.001	$p \equiv 0.8$
MAD	0.0626	0.0627	0.0688	0.0293	
MSE	0.6983	0.6967	0.0046	0.001	
Media	2.8415	2.8395	2.0006	2.0006	
Mediana	2.8397	2.8393	2.0005	1.9997	
Varianza	0.0027	0.0033	0.0043	0.0008	$p(x) = \frac{1}{1+\exp(-2x-2)}$
MAD	0.0495	0.0492	0.0641	0.0281	
MSE	0.7108	0.7081	0.0043	0.0008	
Media	2.8311	2.8293	2.0025	2.0015	
Mediana	2.8291	2.8288	2.0031	2.0003	
Varianza	0.0053	0.0056	0.0053	0.0017	$p(x) = 0.7 + 0.2(\cos(2x+0.4))^2$
MAD	0.0689	0.0684	0.067	0.0312	
MSE	0.6961	0.6933	0.0053	0.0017	

Tabla 5.3: Estimación de α con $\alpha_0 = 2$ y $c = 4$, para el Modelo Exponencial, con outliers.

	LS.cg	LS.nlm	LMS	M.cg	Función de pérdida
Media	2.7802	2.7882	4.9975	5.003	
Mediana	2.782	2.7847	4.9841	5.0036	
Varianza	0.0226	0.0287	0.0651	0.0131	$p \equiv 1$
MAD	0.1527	0.1529	0.262	0.1116	
MSE	4.9501	4.9208	0.065	0.0131	
Media	2.7945	2.7977	5.007	5.0004	
Mediana	2.7987	2.7996	5.0071	4.9991	
Varianza	0.0352	0.0371	0.0744	0.0149	$p \equiv 0.8$
MAD	0.1751	0.1751	0.2831	0.1204	
MSE	4.8992	4.8874	0.0743	0.0149	
Media	2.7651	2.7717	5.0045	5	
Mediana	2.765	2.7662	4.9952	5.0033	
Varianza	0.0242	0.0322	0.0699	0.0124	$p(x) = \frac{1}{1+\exp(-2x-2)}$
MAD	0.1537	0.154	0.2664	0.1095	
MSE	5.0189	4.9973	0.0699	0.0124	
Media	2.7945	2.8006	5.0004	5	
Mediana	2.7945	2.7957	4.9904	5.0026	
Varianza	0.0408	0.0459	0.079	0.0182	$p(x) = 0.7 + 0.2(\cos(2x+0.4))^2$
MAD	0.194	0.1925	0.2797	0.1293	
MSE	4.9051	4.8832	0.0789	0.0182	

Tabla 5.4: Estimación de β con $\beta_0 = 5$ y $c = 4$, para el Modelo Exponencial, con outliers.

	LS.cg	LS.nlm	LMS	M.cg	Función de pérdida
Media	10.0077	10.0077	10.0758	10.0051	
Mediana	10.01	10.01	10.021	10.0004	
Varianza	0.0623	0.0623	0.4853	0.0725	$p \equiv 1$
MAD	0.2487	0.2487	0.6444	0.2659	
MSE	0.0623	0.0623	0.4906	0.0725	
Media	10.0205	10.0205	10.1052	10.0191	
Mediana	10.0199	10.0199	10.068	10.0096	
Varianza	0.083	0.083	0.5668	0.0957	$p \equiv 0.8$
MAD	0.2847	0.2847	0.7538	0.3109	
MSE	0.0833	0.0833	0.5773	0.096	
Media	10.0199	10.0199	10.086	10.0175	
Mediana	10.0206	10.0206	10.033	10.0093	
Varianza	0.0677	0.0677	0.5058	0.0767	$p(x) = \frac{1}{1+\exp(-2x-2)}$
MAD	0.2457	0.2457	0.6664	0.2606	
MSE	0.0681	0.0681	0.5126	0.0769	
Media	10.026	10.026	10.1071	10.0239	
Mediana	10.0162	10.0162	10.0678	10.0081	
Varianza	0.0873	0.0873	0.6162	0.1011	$p(x) = 0.7 + 0.2(\cos(2x+0.4))^2$
MAD	0.288	0.288	0.7294	0.3161	
MSE	0.0879	0.0879	0.6271	0.1015	

Tabla 5.5: Estimación de α con $\alpha_0 = 10$ y $c = 4$, para el Modelo de Michaelis-Menten, sin outliers.

	LS.cg	LS.nlm	LMS	M.cg	Función de pérdida
Media	-0.0034	-0.0035	-0.0061	-0.0054	
Mediana	0.0006	0.0006	-0.0118	-0.0007	
Varianza	0.0158	0.0158	0.1031	0.018	$p \equiv 1$
MAD	0.1203	0.1204	0.2982	0.1323	
MSE	0.0158	0.0158	0.1031	0.018	
Media	0.0013	0.0013	-0.0037	-0.0005	
Mediana	0.001	0.001	0.0104	0.0002	
Varianza	0.0198	0.0198	0.1293	0.0227	$p \equiv 0.8$
MAD	0.1351	0.1351	0.3284	0.1482	
MSE	0.0198	0.0198	0.1292	0.0227	
Media	0.0027	0.0027	0.0009	0.0007	
Mediana	0.0047	0.0047	-0.0036	-0.0029	
Varianza	0.0168	0.0168	0.1094	0.019	$p(x) = \frac{1}{1+\exp(-2x-2)}$
MAD	0.123	0.123	0.3128	0.1339	
MSE	0.0168	0.0168	0.1092	0.019	
Media	0.0026	0.0026	0.006	0.0005	
Mediana	0.0002	0.0002	0.0137	-0.0005	
Varianza	0.0199	0.0199	0.1218	0.0229	$p(x) = 0.7 + 0.2(\cos(2x+0.4))^2$
MAD	0.134	0.134	0.3232	0.1468	
MSE	0.0199	0.0199	0.1217	0.0229	

Tabla 5.6: Estimación de β con $\beta_0 = 0$ y $c = 4$, para el Modelo de Michaelis-Menten, sin outliers.

	LS.cg	LS.nlm	LMS	M.cg	Función de pérdida
Media	28559.8791	2428.4303	10.0675	10.0134	
Mediana	28338.222	1998.2602	10.0208	10.0164	
Varianza	18572833.03	4637701.681	0.5281	0.1131	$p \equiv 1$
MAD	4108.7444	2496.2846	0.6766	0.3372	
MSE	833649855.6	10481868.98	0.5322	0.1132	
Media	28653.1961	2390.9281	10.0817	10.0181	
Mediana	28517.4606	1865.745	10.0284	10.0133	
Varianza	21152741.13	5010523.279	0.6099	0.1412	$p \equiv 0.8$
MAD	4551.7156	2595.2335	0.7632	0.356	
MSE	841564269.7	10674331.35	0.6159	0.1413	
Media	28682.3095	2016.8827	10.0828	10.0231	
Mediana	28542.3652	1346.5702	10.0126	10.0131	
Varianza	18394459.03	4468029.051	0.5432	0.1183	$p(x) = \frac{1}{1+\exp(-2x-2)}$
MAD	4071.7761	1977.8179	0.6913	0.323	
MSE	840477398.2	8491139.12	0.5495	0.1187	
Media	28361.9394	2953.2454	10.0867	10.0263	
Mediana	28297.6178	2818.1634	10.0272	10.0176	
Varianza	20541014.45	5641706.759	0.6117	0.1515	$p(x) = 0.7 + 0.2(\cos(2x+0.4))^2$
MAD	4611.8096	2932.2468	0.7313	0.3909	
MSE	824352942.6	14298758.28	0.6187	0.1521	

Tabla 5.7: Estimación de α con $\alpha_0 = 10$ y $c = 4$, para el Modelo de Michaelis-Menten, con outliers.

	LS.cg	LS.nlm	LMS	M.cg	Función de pérdida
Media	9.3725	5.0544	-0.0023	-0.0044	
Mediana	9.3791	6.717	-0.0041	0.0077	
Varianza	0.0224	18.0858	0.0893	0.0219	$p \equiv 1$
MAD	0.1378	1.2066	0.2735	0.1413	
MSE	87.8652	43.6151	0.0892	0.0219	
Media	9.3743	4.3787	-0.0003	-0.0028	
Mediana	9.3792	6.6634	0.003	0.002	
Varianza	0.0237	26.3912	0.1091	0.0264	$p \equiv 0.8$
MAD	0.1452	1.3846	0.3211	0.1522	
MSE	87.902	45.5378	0.109	0.0263	
Media	9.3752	3.833	0.0079	0.0007	
Mediana	9.381	6.3119	0.0112	0.0004	
Varianza	0.0213	28.592	0.0932	0.0225	$p(x) = \frac{1}{1+\exp(-2x-2)}$
MAD	0.1425	1.7559	0.2885	0.1405	
MSE	87.9148	43.2553	0.0932	0.0225	
Media	9.3516	5.1931	-0.0036	-0.0007	
Mediana	9.3551	7.0439	-0.0001	-0.0008	
Varianza	0.0247	20.3028	0.1037	0.0269	$p(x) = 0.7 + 0.2(\cos(2x+0.4))^2$
MAD	0.1501	1.0365	0.3049	0.16	
MSE	87.4777	47.2512	0.1036	0.0268	

Tabla 5.8: Estimación de β con $\beta_0 = 0$ y $c = 4$, para el Modelo de Michaelis-Menten, con outliers.

5.5. Boxplots

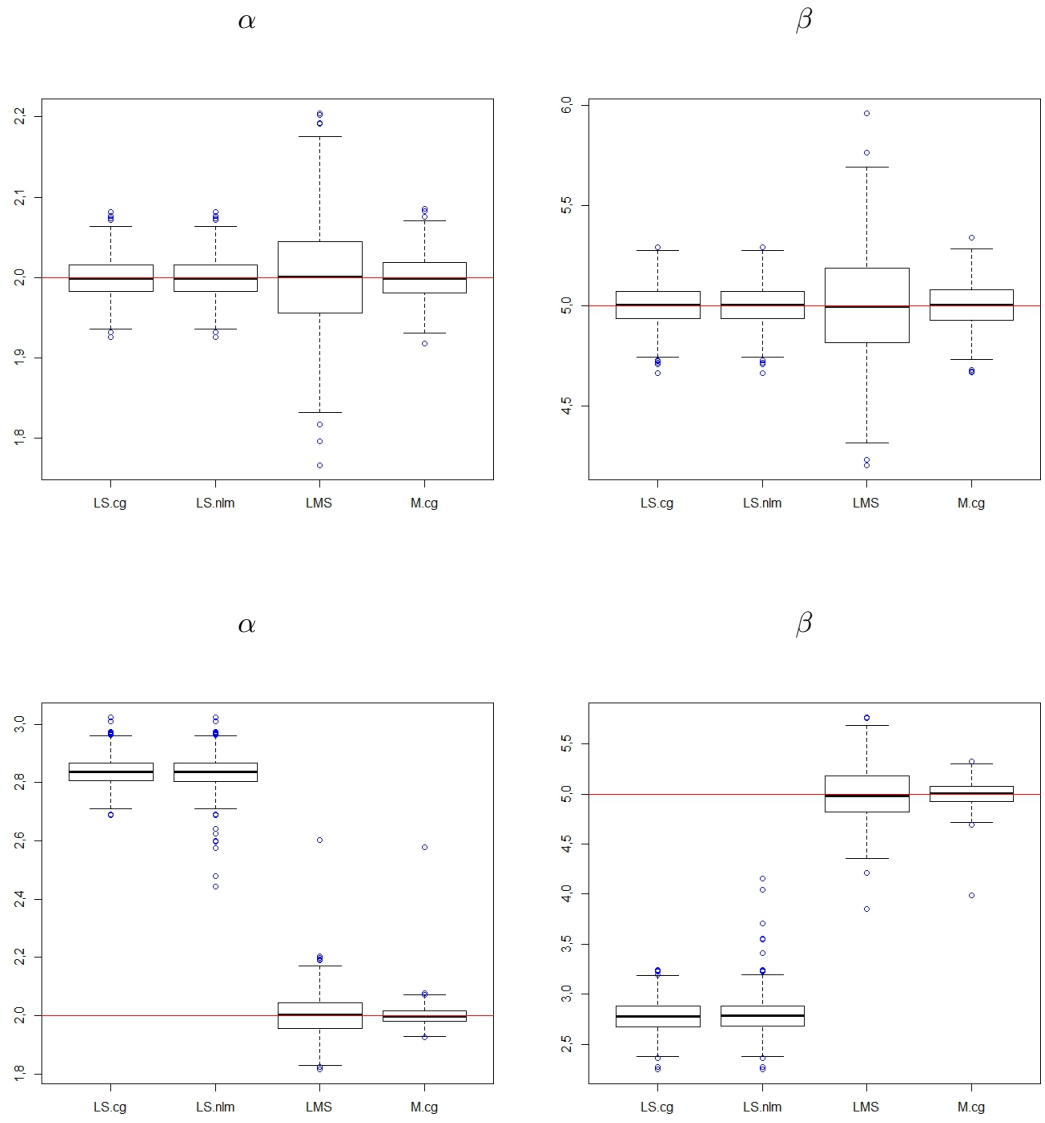

Figura 5.1: Boxplots de los estimadores de α y β para datos sin outliers en la primera línea y con outliers en la segunda, con $c = 4$ y $p \equiv 1$, para el Modelo Exponencial.

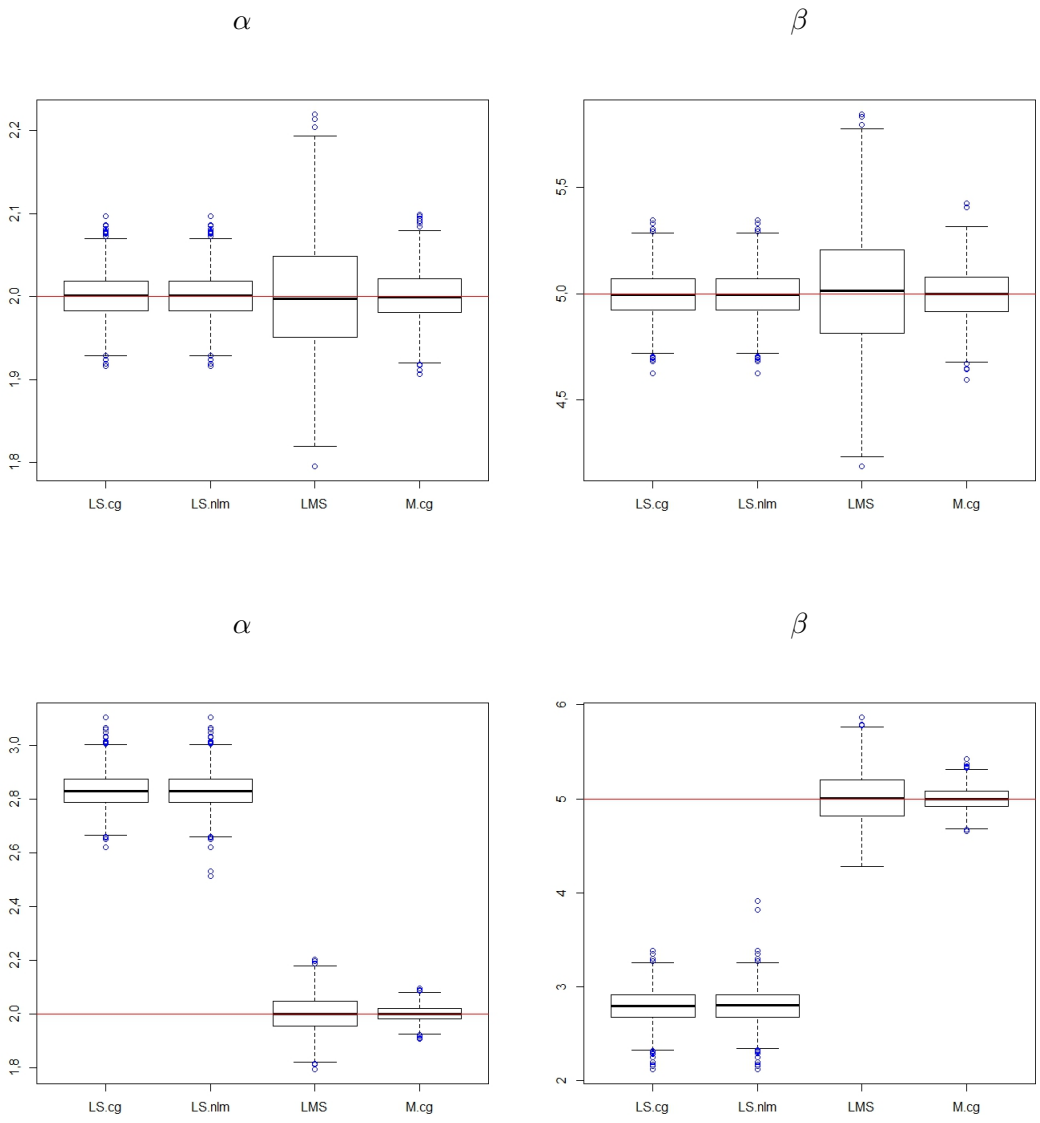

Figura 5.2: Boxplots de los estimadores de α y β para datos sin outliers en la primera línea y con outliers en la segunda, con $c = 4$ y $p \equiv 0.8$, para el Modelo Exponencial.

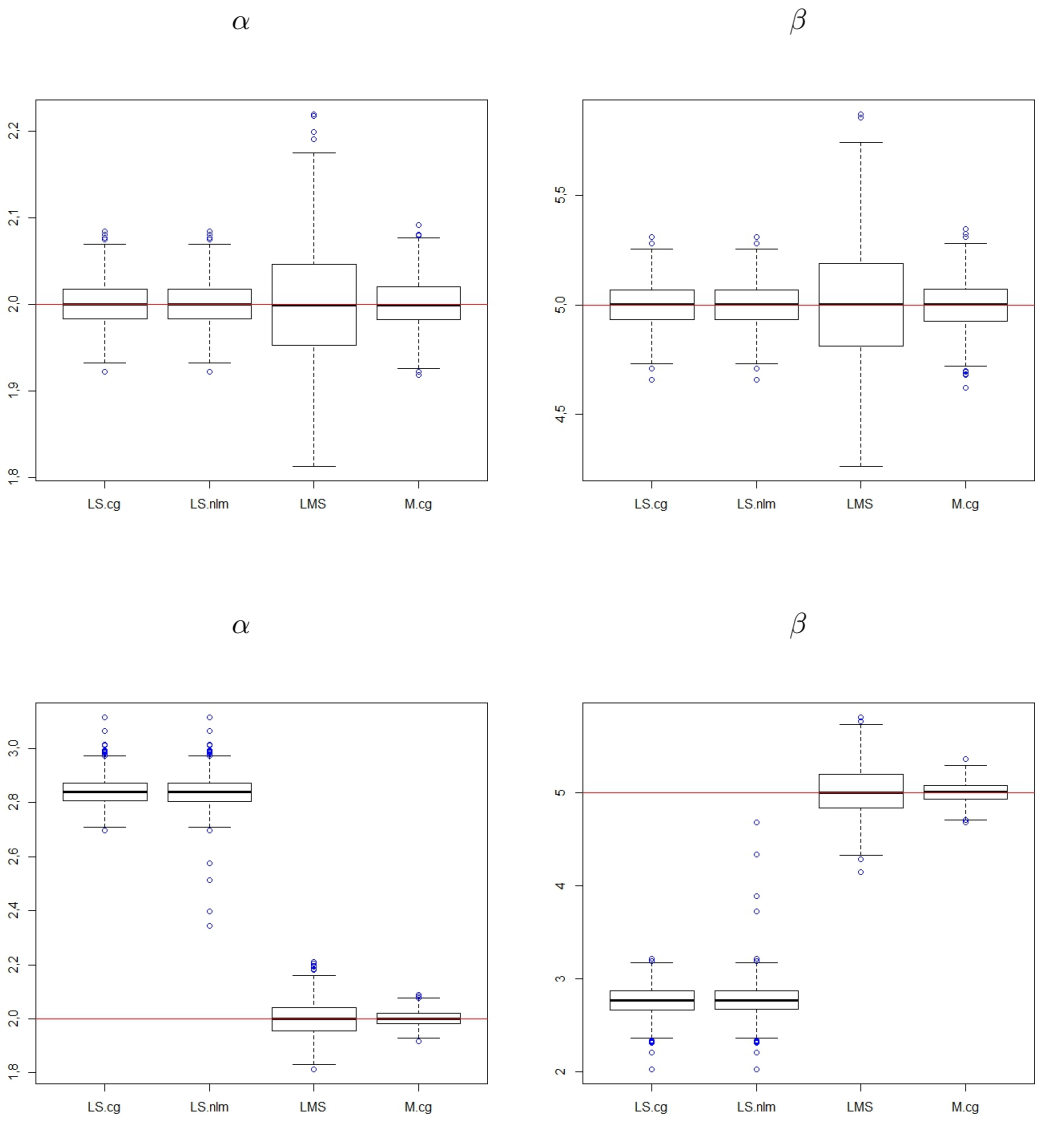

Figura 5.3: Boxplots de los estimadores de α y β para datos sin outliers en la primera línea y con outliers en la segunda, con $c = 4$ y $p(x) = \frac{1}{1+\exp(-2x-2)}$, para el Modelo Exponencial.

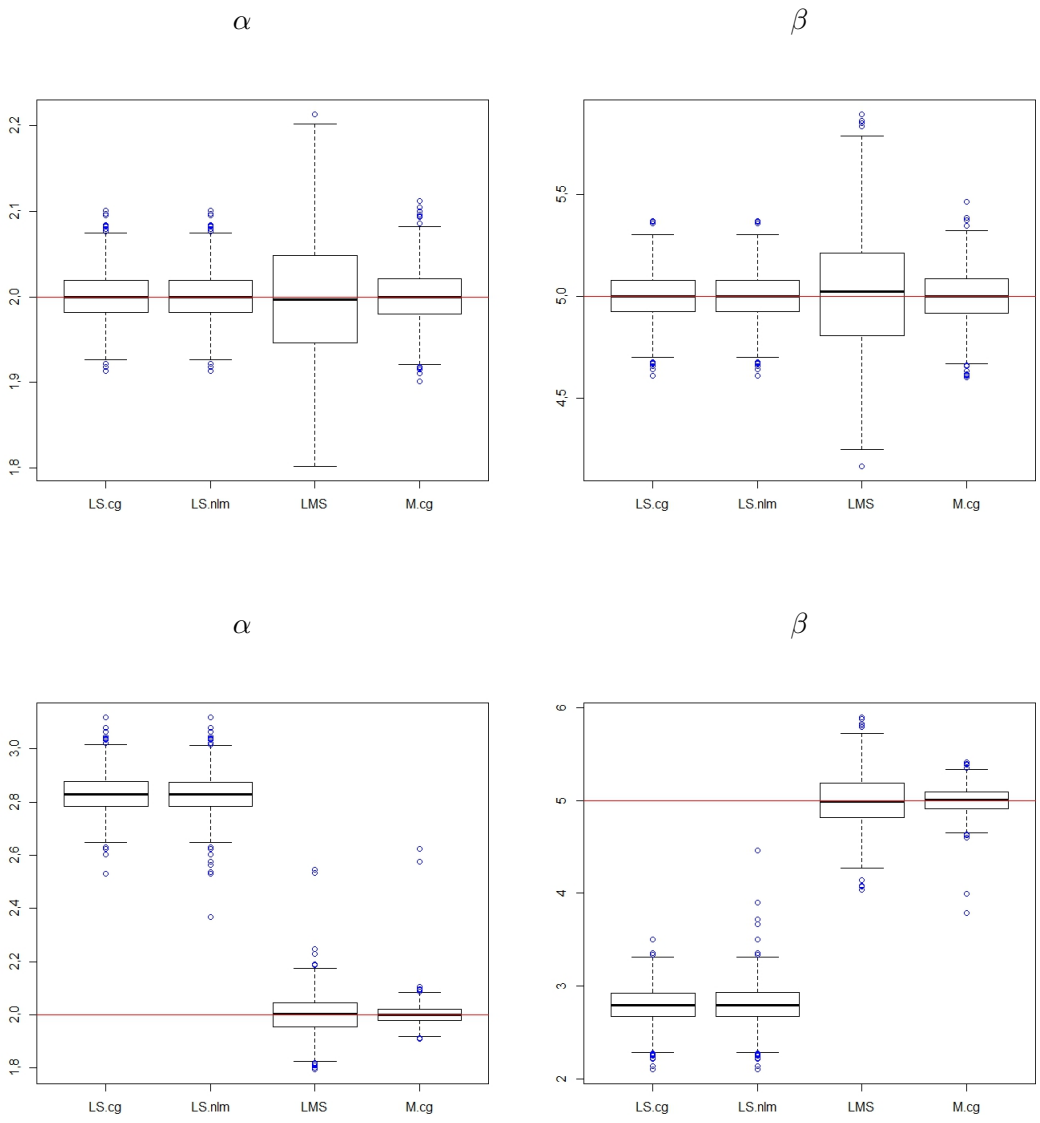

Figura 5.4: Boxplots de los estimadores de α y β para datos sin outliers en la primera línea y con outliers en la segunda, con $c = 4$ y $p(x) = 0.7 + 0.2(\cos(2x + 0.4))^2$, para el Modelo Exponencial.

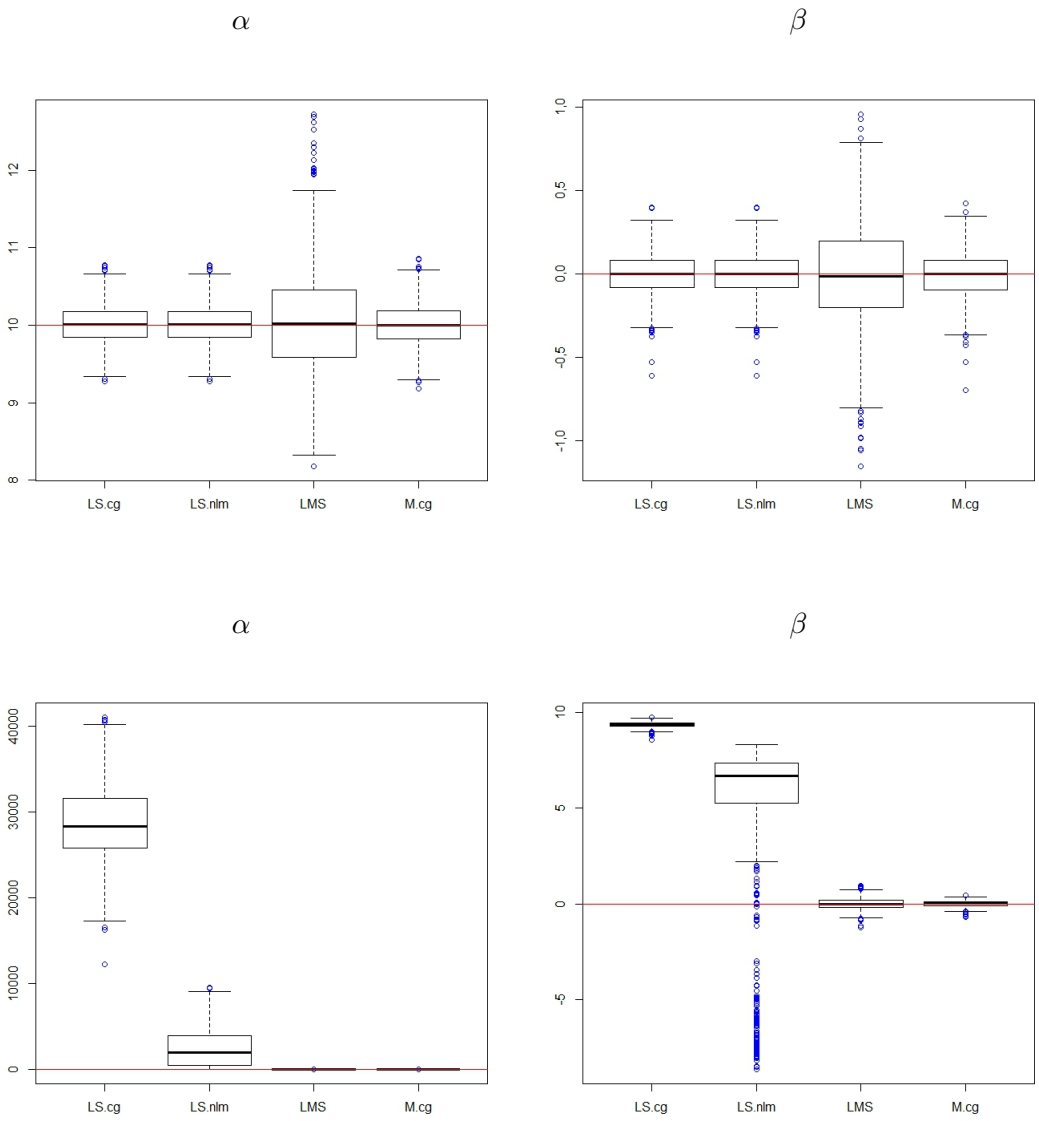

Figura 5.5: Boxplots de los estimadores de α y β para datos sin outliers en la primera línea y con outliers en la segunda, con $c = 4$ y $p \equiv 1$, para el Modelo de Michaelis-Menten.

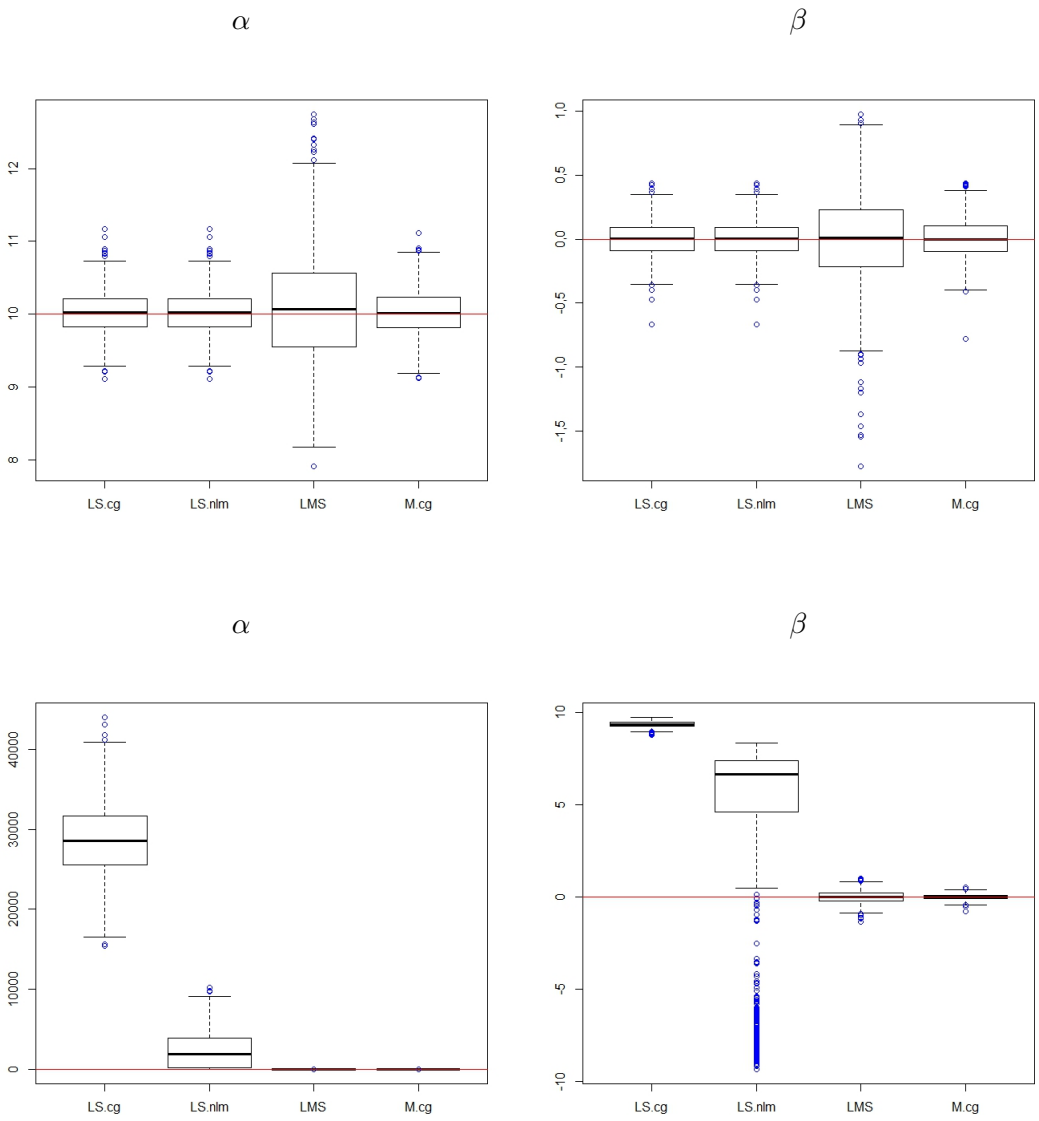

Figura 5.6: Boxplots de los estimadores de α y β para datos sin outliers en la primera línea y con outliers en la segunda, con $c = 4$ y $p \equiv 0.8$, para el Modelo de Michaelis-Menten.

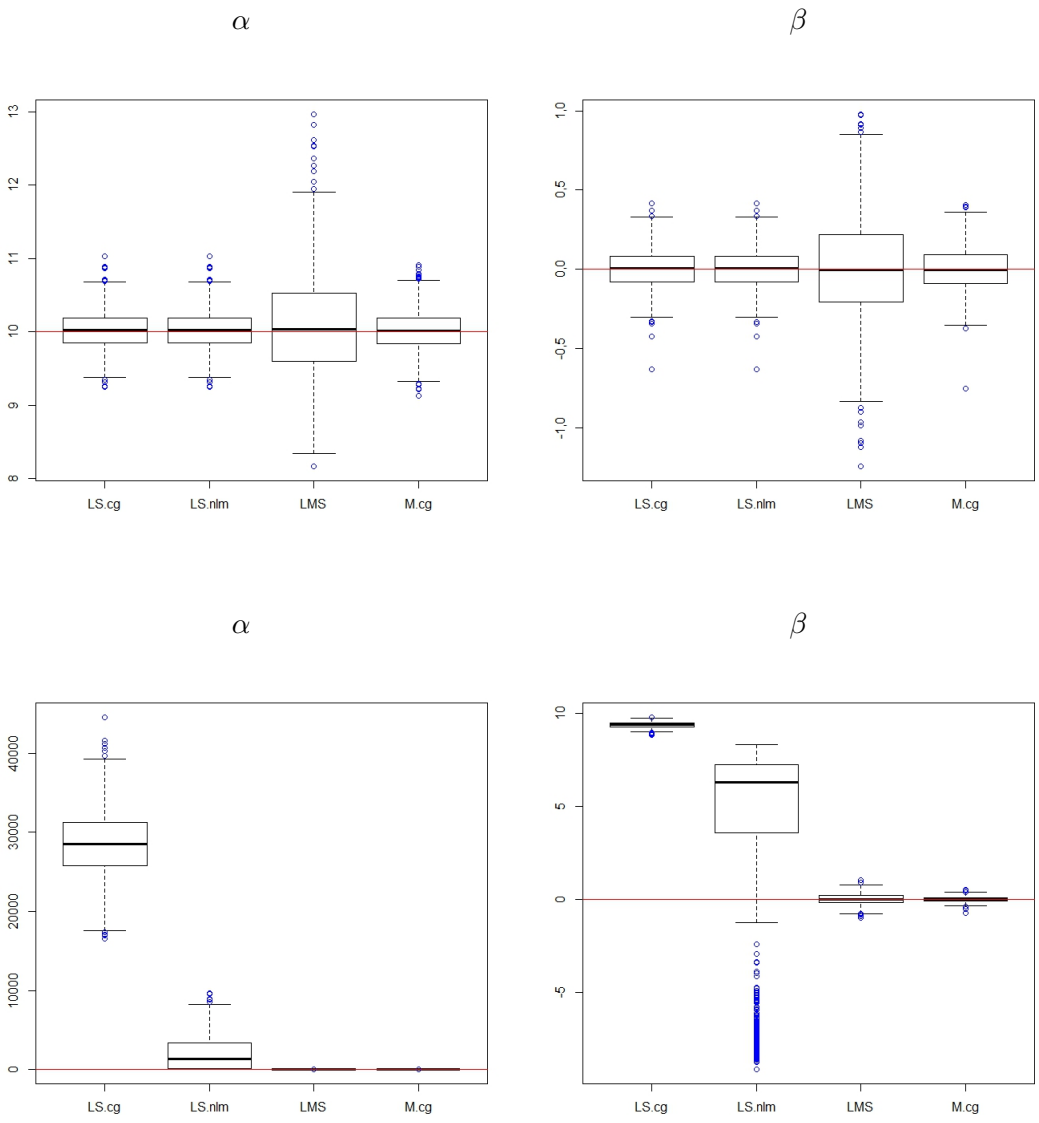

Figura 5.7: Boxplots de los estimadores de α y β para datos sin outliers en la primera línea y con outliers en la segunda, con $c = 4$ y $p(x) = \frac{1}{1+\exp(-2x-2)}$, para el Modelo de Michaelis-Menten.

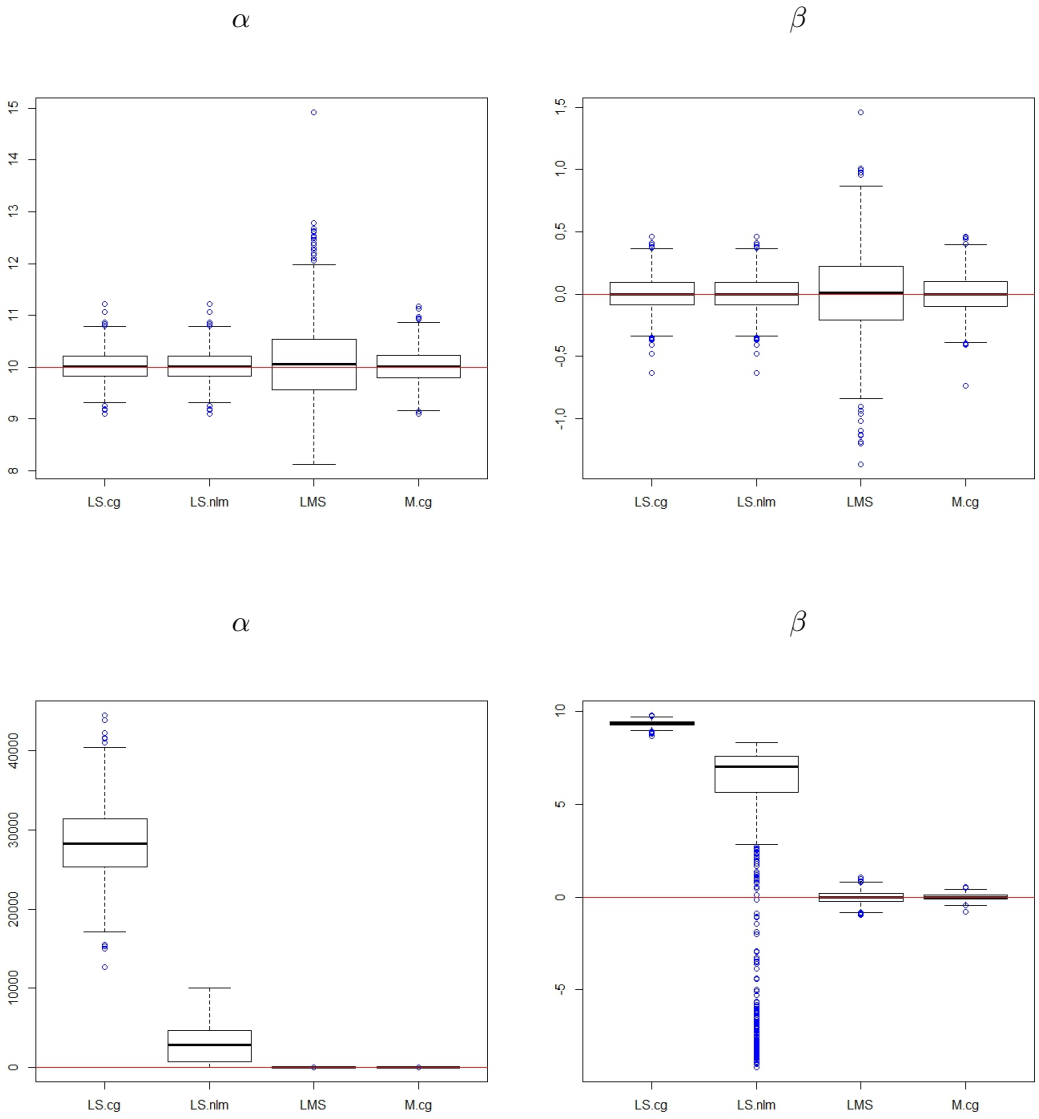

Figura 5.8: Boxplots de los estimadores de α y β para datos sin outliers en la primera línea y con outliers en la segunda, con $c = 4$ y $p(x) = 0.7 + 0.2(\cos(2x + 0.4))^2$, para el Modelo de Michaelis-Menten.

Capítulo 6

Ejemplo

Edad de los Conejos Medida "Por el Ojo"

El conejo europeo *Oryctolagus cuniculus* es una de las principales plagas en Australia. Un método fiable de determinación de la edad para los conejos capturados en el medio silvestre sería de importancia en los estudios ecológicos. En un estudio realizado por Dudzinski y Mykytowycz (1961), se midió el peso seco de la lente del ojo para 71 conejos salvajes, con edades conocidas, viviendo a la intemperie. Como el peso de la lente del ojo tiende a variar mucho menos con las condiciones ambientales que el peso total del cuerpo es, por lo tanto, un mejor indicador de la edad.

Los conejos nacieron y vivieron libres en un recinto de 1.7 acre [1] en Gungahlin, ACT. La información sobre el nacimiento y la historia de cada individuo fue conocida con exactitud. Los conejos que vivían en dicho recinto dependían de la fuente de alimento natural. En este experimento 18 de las lentes oculares se obtuvieron de conejos que murieron en el transcurso de este estudio de varias causas como coccidiosis, depredación de aves o inanición. Los restantes 35 conejos fueron sacrificados deliberadamente, inmediatamente luego de haber sido capturados en el recinto o de haber sido mantenidos durante algún tiempo en jaulas. Las lentes se conservaron y se determinó su peso seco.

Sean

- $Lens$ = peso de la lente del ojo en mg.
- Age = edad en días.

Dudzinski y Mykytowycz (1961) sugieren la relación determinística

$$E(Lens) = \boldsymbol{\theta}_1 \exp\left(\frac{-\boldsymbol{\theta}_2}{\boldsymbol{\theta}_3 + Age}\right),$$

pero este modelo no resulta homocedástico.

[1] 1 acre $\approx 4.047\ m^2$

Un modelo apropiado para este caso sería

$$E(\log(Lens)) = \boldsymbol{\theta}_1 - \frac{\boldsymbol{\theta}_2}{\boldsymbol{\theta}_3 + Age} \ ,$$

el cual sí resulta homocedástico.

Con el fin de comparar, nuevamente, la performance del estimador robusto propuesto frente al estimador clásico se procedió a realizar una estimación del parámetro $\boldsymbol{\theta}$ a partir de los datos *Lens* y *Age*. Para lo cual, se consideraron los mismos de forma completa como así también con ciertas pérdidas en la variable *Lens* mediante las siguientes funciones:

- $p \equiv 1$: sin respuestas faltantes
- $p \equiv 0.8$: faltante de respuestas completamente al azar
- $p(x) = \dfrac{1}{1 + \exp(-0.5x - 1)}$: modelo logístico (se introducen aproximadamente un 25 % de respuestas faltantes)

A su vez, se contaminaron las últimas cinco observaciones de la variable $\log(Lens)$, en los tres escenarios planteados, de la siguiente manera:

- $\log(Lens)_{67} = 7$
- $\log(Lens)_{68} = 7.01$
- $\log(Lens)_{69} = 7.02$
- $\log(Lens)_{70} = 7.03$
- $\log(Lens)_{71} = 7.05$

En la Tabla 6.1 presentamos los estimadores obtenidos mediante el método de mínimos cuadrados y el método robusto propuesto para la tres funciones de pérdida planteadas bajo

- C_0: datos sin contaminar
- C_1: datos contaminados

Podemos observar que, si bien, con el estimador de mínimos cuadrados se obtienen estimaciones muy parecidas entre sí, al cambiar la probabilidad p, su comportamiento es errático al introducir sólo 5 outliers, mientras que el M-estimador se mantiene muy estable. En particular, se puede observar que bajo la contaminación C_1, la performance del estimador de mínimos cuadrados empeora notablemente para el segundo parámetro $\hat{\boldsymbol{\theta}}_{\text{LS},2}$ mientras que la del M-estimador permanece prácticamente invariante.

En las Figuras 6.1 y 6.2, presentamos de manera gráfica el ajuste de los datos completos con ambos métodos para C_0 y C_1, donde se puede apreciar lo mencionado anteriormente. En la Figura 6.1 podemos ver cómo ambos estimadores ajustan apropiadamente los datos y que

	$\hat{\boldsymbol{\theta}}_{\text{LS},1}$	$\hat{\boldsymbol{\theta}}_{\text{LS},2}$	$\hat{\boldsymbol{\theta}}_{\text{LS},3}$	$\hat{\boldsymbol{\theta}}_{\text{M},1}$	$\hat{\boldsymbol{\theta}}_{\text{M},2}$	$\hat{\boldsymbol{\theta}}_{\text{M},3}$	Función de Pérdida
C_0	5.6399	130.5836	37.6028	5.6317	126.9005	35.8215	$p \equiv 1$
C_1	6.9746	800.5597	221.2196	5.6349	128.5574	36.6739	
C_0	5.6435	130.2084	37.4053	5.6372	126.9601	35.5521	$p \equiv 0.8$
C_1	7.2931	1043.2669	265.7357	5.6411	128.536	36.4145	
C_0	5.641	130.0265	37.0525	5.6319	125.3534	34.6558	$p(x) = \frac{1}{1+\exp(-0.5x-1)}$
C_1	7.2765	1016.3011	258.0013	5.6383	128.6059	36.3316	

Tabla 6.1: Estimación del parámetro $\boldsymbol{\theta}$ mediante el LS-estimador y el M-estimador

tanto las estimaciones proporcionadas por el método clásico como por el método robusto dan ajustes prácticamente coincidentes. En la Figura 6.2, notamos cómo las observaciones atípicas influyen en el ajuste mediante el LS-estimador, pero no mediante el M-estimador. Sólo mostramos las gráficas para el caso con datos completos ya que para las restantes funciones de pérdida se observaron resultados análogos.

Por último, se analizó el comportamiento de ambos estimadores al contaminar únicamente la observación número 6 de la variable $\log(Lens)$ cuyo valor original es $\log(Lens)_6 = 3.7$. Los valores que se le asignaron fueron:

- $\log(Lens) = 2$
- $\log(Lens) = 3$
- $\log(Lens) = 3.7$
- $\log(Lens) = 6$
- $\log(Lens) = 8$

para las tres funciones de pérdida por igual.

En las Figuras 6.3, 6.4 y 6.5 mostramos cómo los estimadores se ven afectados por esta contaminación. A simple vista, podemos ver que es el estimador de mínimos cuadrados $\hat{\boldsymbol{\theta}}_{\text{LS}}$ el que sufre variaciones significativas a medida que cambiamos el valor de $\log(Lens)_6$, para cualquiera de las funciones de pérdida y en especial para el párametro $\hat{\boldsymbol{\theta}}_2$. Sin embargo, el comportamiento del M-estimador $\hat{\boldsymbol{\theta}}_{\text{M}}$ se mantiene prácticamente invariante, de hecho, los gráficos muestran una función casi constante respecto a los tres parámetros por igual. Simplemente, por una cuestión de escala, se aprecia más para $\hat{\boldsymbol{\theta}}_1$ en los tres escenarios de pérdida planteados.

En conclusión, pudimos observar mediante un ejemplo con datos reales, que al igual que trabajando con datos simulados, la performance del estimador clásico resulta mejor cuando el modelo no se encuentra contaminado pero que ante la mínima presencia de outliers su comportamiento resulta sumamente afectado. Mientras que el estimador robusto propuesto tiene una performance similar al de mínimos cuadrados cuando la muestra no posee datos atípicos y que, gracias a su eficiencia y robustez, al contaminarse la misma no se ven afectados

en lo absoluto. Y esto lo pudimos observar tanto en muestras completas como en muestras con distintos porcentajes de respuestas faltantes.

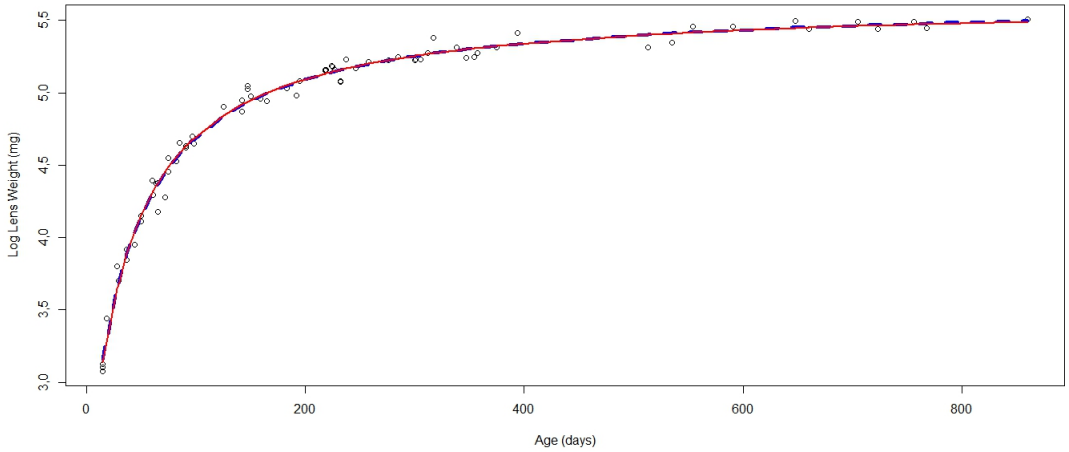

Figura 6.1: Ajuste del modelo mediante método de mínimos cuadrados (azul) y método robusto (rojo) sin outliers.

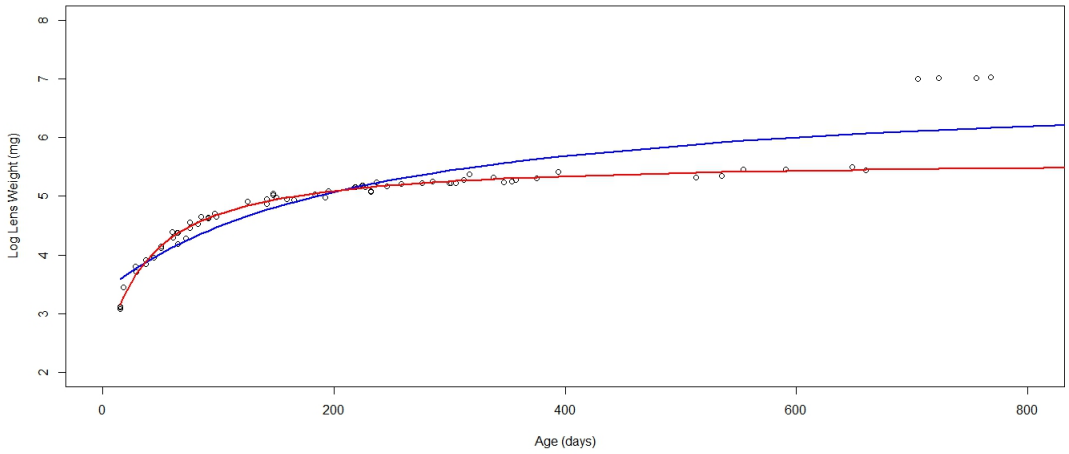

Figura 6.2: Ajuste del modelo mediante método de mínimos cuadrados (azul) y método robusto (rojo) con outliers.

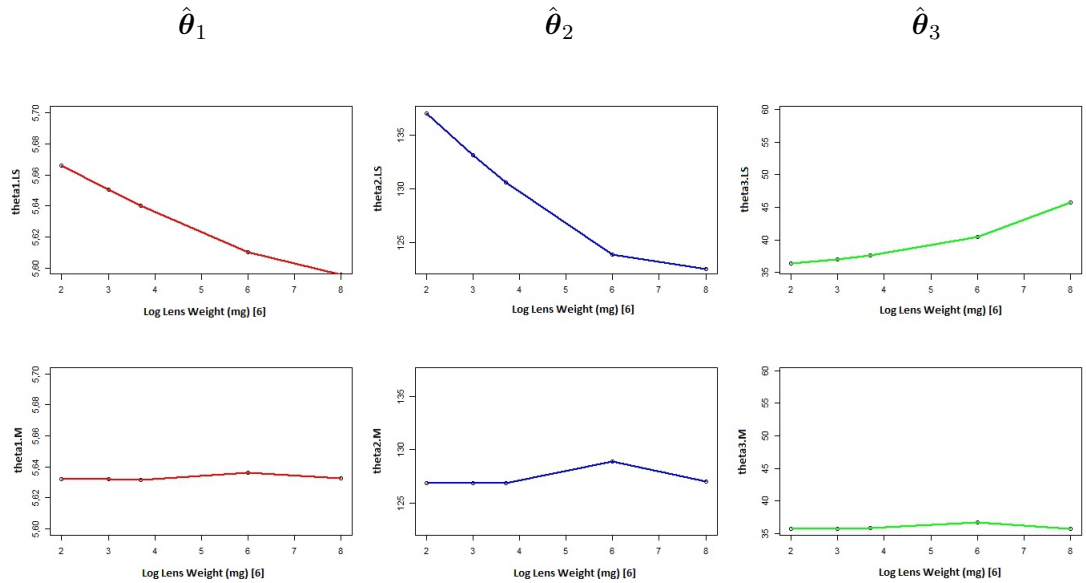

Figura 6.3: Gráfico de $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$ vs. $\log(Lens)_6$, en la primera línea, y de $\hat{\boldsymbol{\theta}}_{\mathrm{M}}$ vs. $\log(Lens)_6$ en la segunda, para $p \equiv 1$.

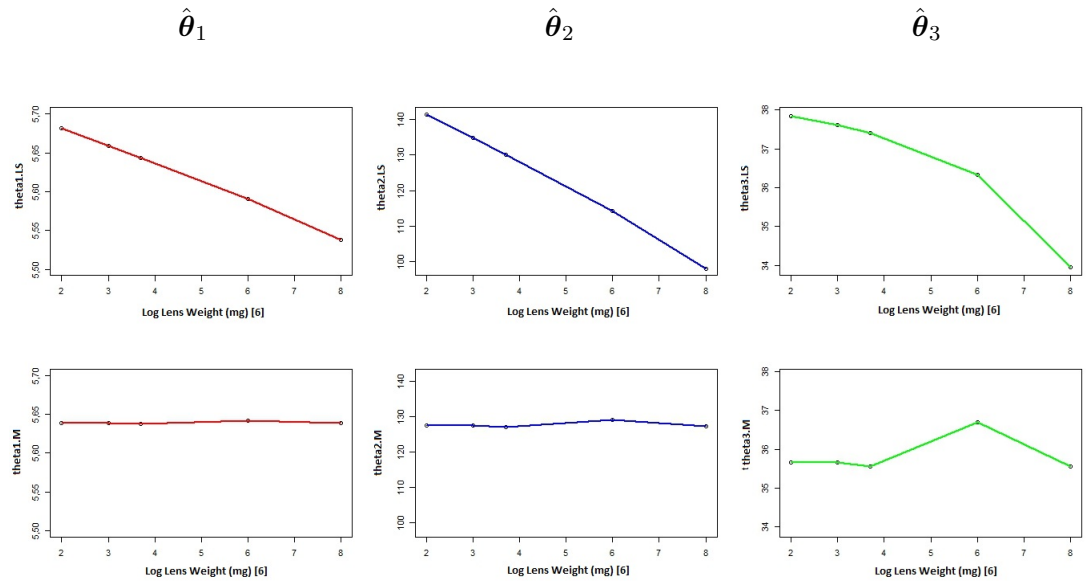

Figura 6.4: Gráfico de $\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$ vs. $\log(Lens)_6$, en la primera línea, y de $\hat{\boldsymbol{\theta}}_{\mathrm{M}}$ vs. $\log(Lens)_6$ en la segunda, para $p \equiv 0.8$.

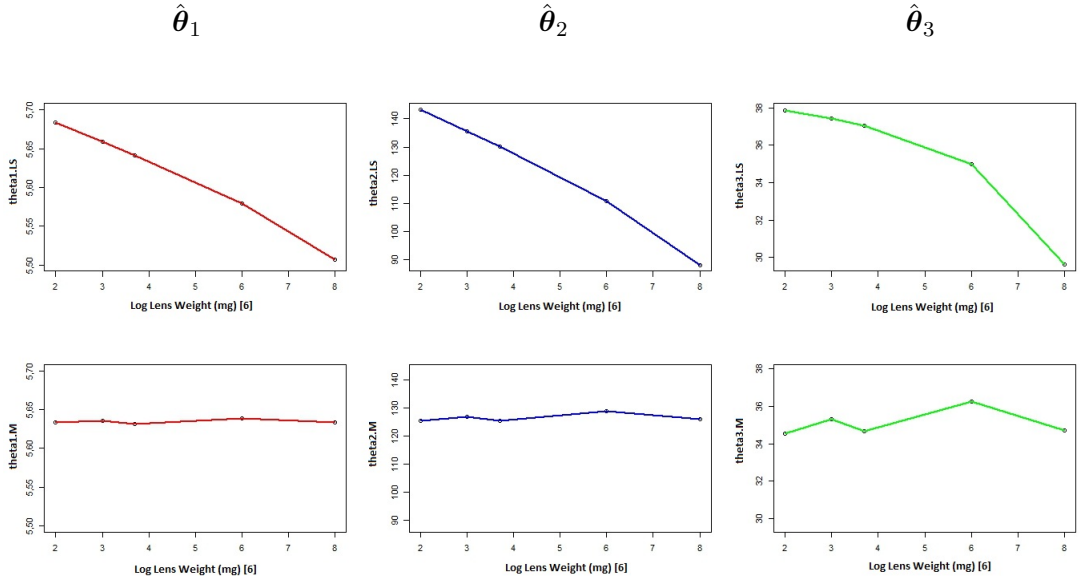

Figura 6.5: Gráfico de $\hat{\boldsymbol{\theta}}_{\text{LS}}$ vs. $\log(Lens)_6$, en la segunda línea, y de $\hat{\boldsymbol{\theta}}_{\text{M}}$ vs. $\log(Lens)_6$ en la segunda, para $p(x) = \frac{1}{1+\exp(-0.5x-1)}$.

Bibliografía

[1] Bates D.M., Watts D.G. (1988). *Nonlinear regression analysis and its applications*. John Wiley and Sons.

[2] Bianco A., Boente G., González-Manteiga W., Pérez-González A. (2010). Estimation of the marginal location under a partially linear model with missing responses. *Elsevier/Computational Statistics and Data Analysis*, **54**, 546-564.

[3] Bianco A., Boente G., Rodrigues I. (2011). Resistant estimators in Poisson and Gamma models with missing responses. *Submitted to Journal of Multivariate Analysis*. Disponible en `http://www.ic.fcen.uba.ar/preprints/biancoboenterodrigues_2011.pdf`.

[4] Bianco A., Boente G., Rodrigues I. (2012). Robust tests in generalized linear models with missing responses. Disponible en `http://www.ic.fcen.uba.ar/preprints/test_GLM_TR.pdf`.

[5] Carroll, R. J. y Ruppert, D. (1987). Diagnostics and robust estimation when transforming the regression model and the response. *Technometrics*, **29**, 287-299.

[6] Conceição E.L.T., Portugal A.A.T.G. (2011). Finite-sample comparison of robust estimators for nonlinear regression using Monte Carlo simulation: Part I. Univariate response models. *Elsevier/Computers and Chemical Engineering*, **35**, 530-544.

[7] Dudzinski, M. L. y Mykytowycz, R. (1961). The Eye Lens as an Indicator of Age in the Wild Rabbit in Australia, *CSIRO Wildl. Res.*, Vol. **6**, pp.156-159.

[8] Fasano M. V. (2009). *Teoría asintótica de estimadores robustos en regresión no lineal*. Tesis Doctoral, Universidad Nacional de la Plata, Argentina.

[9] Fasano, M. V., Maronna, R. A., Sued, M. y Yoahi, V. J. (2011). *Continuity and differentiability of regression M-estimates*. Disponible en `http://arxiv.org/abs/1004.4314`.

[10] Fernholz, L. T. (1983). *Von Mises Calculus for Statistical Functionals*. Springer-Verlag, New York.

[11] Fraiman, R. (1983). General M-estimators and applications to bounded influence estimation for non-linear regression. *Communications in Statistics. Theory and Methods*, Vol. **A12, 22**, 2617-2631.

[12] Hampel, F. (1971). A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics*, **42**, 1887-1896.

[13] Hampel, F. R. (1975). Beyond location parameters: Robust concepts and methods (with discussion). *Bull. Inst. Internat. Statist*, **46 (1)**, 375-391.

[14] Heritier S. y Ronchetti E. (1994). Robust Bounded-Influence Tests in General Parametric Models. *Journal of the American Statistical Association*, Vol. **89**, No. **427**, pp. 897-904.

[15] Huber, P. J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, **1**, 799-821.

[16] Huber, P. (1981). *Robust Statistics*. John Wiley and Sons.

[17] Jennrich, R. I. (1969). Asymptotic Properties of Non-Linear Least Squares Estimators. *The Annals of Mathematical Statistics*, Vol. **40**, Number **2**, 633-643.

[18] Jureckova, J. y Portnoy, S. (1987). Asymptotics for one-step M-estimators in regression with application to combining efficiency and high breakdown point. *Comm. Statis. Theory Methods*, **16**, 2187-2199.

[19] Kim, J. y Pollard D. (1990). Cube root asymptotics. *The Annals of Statistics*, **18(1)**, 191-219.

[20] Markatou, M. y Manos, G. (1996). Robust tests in nonlinear regression models. *Journal of Statistical Planning and Inference*, **55**, 205-217.

[21] Maronna, R. A. y Yohai, V. J. (1979). Asymptotic behavior of M-estimates for the linear model. *The Annals of Statistics*, **7**, 258-268.

[22] Maronna, R., Martin, R. y Yohai V. (2006). *Robust Statistics - Theory and Methods*. John Wiley and Sons.

[23] Mukherjee, K. (1996). Robust estimation in nonlinear regression via minimum distance method. *Mathematical Methods of Statistics*, **5**, 99-112.

[24] Müller, U. U. (2009). Estimating linear functionals in nonlinear regression with responses missing at random. *Annals of Statistics*, **37**, 2245-2277.

[25] Nelder, J. E., and Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, **7**, 308.

[26] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1986). *Numerical Recipes: The Art of Scientific Computing*. New York: Cambridge University Press.

[27] Ratkowsky, D. (1983). *Nonlinear regression modeling*. Marcel Dekker, Inc.

[28] Rousseeuw, P. y Yohai, V. (1984). Robust regression by means of S-estimators, en Robust and nonlinear time series analysis, editor Heidelberg. *Lecture Notes in Statististics*, **26**, 256-272. Springer, New York.

[29] Rousseeuw, P. (1984). Least median of squares regression. *Journal of the American Statistical Association*, **79**, 871-880.

[30] Rousseeuw, P. J. (1985). Multivariate Estimation with High Breakdown Point. Mathematical Statistics and Applications. *Dordrecht: Reidel Publishing Company*, 283-297.

[31] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. John Wiley and Sons.

[32] Rousseeuw, P. J. and van Zomeren B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, **85**, 633-651.

[33] Sakata, S. y White, H. (2001). S-estimation of nonlinear regression models with dependent and heterogeneous observations. *Journal of Econometrics*, **103**, 5-72.

[34] Seber, G. A. F. y Wild, C. J. (1989). *Nonlinear Regression*. John Wiley and Sons.

[35] Simpson, D. G., Ruppert, D. and Carroll, R. J. (1991). One-step GM-estimates and stability of inferences in linear regression. *Journal of the American Statistical Association*.

[36] *STAT 8230 - Applied Nonlinear Regression Course*. Disponible en `http://www.stat.uga.edu/~dhall/8230`

[37] Stromberg, A. J. y Ruppert, D (1992). Breakdown in nonlinear regression. *Journal of the American Statistical Association*, **87**, 991-997.

[38] Stromberg, A. J. (1993). Computation of High Breakdown Nonlinear Regression Parameters. *Journal of the American Statistical Association*, **88**, 237-244.

[39] Stromberg, A. J. (1995). Consistency of the least median of squares estimator in nonlinear regression. *Communications in Statistics: Theory and Methods*, **24**, 1971-1984.

[40] Stromberg, A., Hossjer, O. and Hawkins, D. (2000). The least trimmed difference regression estimator and alternatives. *Journal of the American Statistical Association*, **95**, 853-864.

[41] Sued, M. y Yohai, V. J. (2012). *A robust approach for location estimation in a missing data setting*. Disponible en `http://arxiv.org/abs/1004.5418`.

[42] Tabatabai M. A. y Argyros I. K. (1993). Robust estimation and testing for general nonlinear regression models. *Applied mathematics and computation*, **58**, 85-101.

[43] Tiede, J. J. y Pagano, M. (1979). The application of robust calibration to radioimmunoassay. *Biometrics*, **35**, 567-574.

[44] Yohai, V. J. (1987). High breakdown-point and high efficiency estimates for regression. *The Annals of Statistics*, **15**, 642-656.

[45] Yohai, V. J. y Zamar, R. H. (1988). High breakdown estimates of regression by means of the minimization of an efficient scale. *Journal of the American Association*, **83**, 406-413.