



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

Tesis de Licenciatura

Modelo de Regresión no Lineal

Selva Carina Figueroa

Director: Dra. Graciela Boente Boente

26 de Diciembre 2013

Índice general

1. Introducción	2
2. Métodos de estimación	6
2.1. Mínimos cuadrados no lineal	6
2.1.1. Métodos Numéricos	8
2.1.2. Mínimos cuadrados generalizados	10
2.2. Estimadores de máxima verosimilitud	11
2.2.1. Método de Verosimilitud Concentrado	13
2.3. Estimación de cuasi-verosimilitud	13
2.4. Métodos Robustos	14
3. Problemas comunes	17
3.1. Convergencia de los métodos iterativos	17
3.2. Validez de la inferencia asintótica	18
3.3. Problemas de identificación y de mal condicionamiento	19
4. Medidas de curvatura y no linealidad	22
4.1. Interpretación geométrica	23
5. Inferencia Estadística	25
5.1. Intervalos de Confianza asintóticos	25
5.2. Intervalos simultáneos y regiones de confianza	26
6. Software y Ejemplos	28
6.1. Ejemplo de datos de cebolla	28

6.2. Ejemplo de datos de ciervos	31
6.3. Ejemplo de datos de ultrasonido	35
7. Implementación del algoritmo de Stromberg en R	39
Bibliografía	50

Agradecimientos

A Juan Pablo por acompañarme en todo este largo proceso.

A mi mamá por ayudarme y alentarme a terminar.

A mi amiga María José por darme ánimos y no dejarme decaer.

A Graciela, por tenerme paciencia.

Y al resto de mis amigos y compañeros de trabajo que insistieron para que concluyera tantos años de estudio.

A todos, gracias.

Capítulo 1

Introducción

Una de las situaciones más frecuentes en el análisis estadístico es que los datos provengan de observaciones o respuestas conocidas y_i que dependen de un conjunto de valores, fijos o aleatorios, \mathbf{x}_i .

Una manera adecuada de modelar esta relación es a través de un modelo de regresión de la forma

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}) + \varepsilon_i \quad i = 1, \dots, n$$

donde $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})^T$, $f(\mathbf{x}, \boldsymbol{\theta})$ es una función generalmente conocida, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$ es un vector de parámetros p -dimensional desconocido y los ε_i representan los errores experimentales.

En muchas situaciones, el modelo de regresión lineal

$$f(\mathbf{x}, \boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 + \dots + \theta_k x_k$$

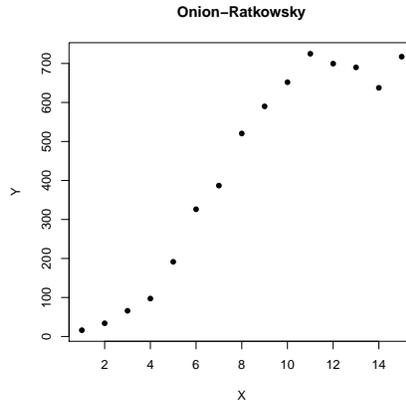
es adecuado para modelar la relación entre la variable respuesta y y las covariables \mathbf{x} . En algunos casos donde no hay relación lineal evidente entre las variables respuesta y regresoras, con una transformación de las variables podemos linealizar la relación.

Pero cuando las consideraciones teóricas lo sugieren o los datos no muestran un comportamiento lineal, hacemos uso de los modelos no lineales. Aún cuando una aproximación lineal funciona bien, un modelo no lineal puede dar una mejor interpretación del problema.

Algunos ejemplos que analizaremos son:

- Ejemplo 1 : Datos de crecimiento de cebollas.

Los datos corresponden al crecimiento de la cebolla estudiado por Ratkowsky , ver [17] y tomado por Lawrence y Arthur , ver [15].



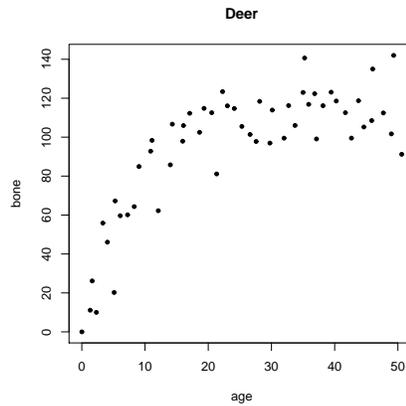
Modelos de crecimiento sigmoidal fueron propuestos para este tipo de datos, entre ellos el modelo de Gompertz, basado en la distribución Weibull:

$$f(x, \boldsymbol{\theta}) = \theta_1 \exp(-\exp(\theta_2 - \theta_3 x)).$$

- Ejemplo 2: Huesos de ciervo.

Este conjunto de datos muestra la longitud del hueso de la mandíbula como una función de la edad en los ciervos. La teoría indica que la relación es una función asintótica exponencial con dos parámetros:

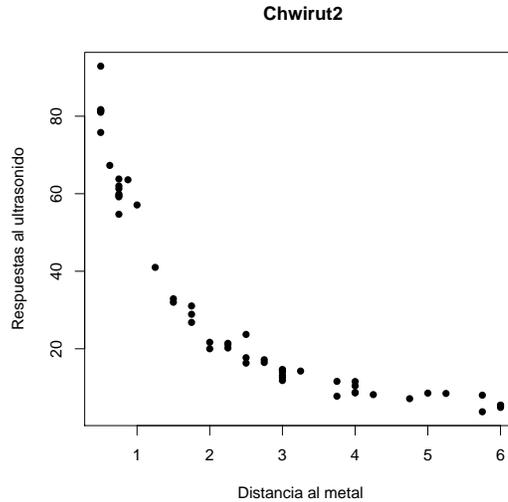
$$f(x, \boldsymbol{\theta}) = \theta_1(1 - e^{-\theta_2 x}).$$



- Ejemplo 3 : Chwirut2.

El conjunto de datos original contiene las mediciones provenientes de un experimento que estudia como la respuesta al ultrasonido depende de la distancia al metal. Los datos son de libre acceso y pertenecen al *NIST (National Institute of Standards and Technology)* (1979). Ultrasonic Reference Study Block (D. Chwirut). Technical report.

<http://www.itl.nist.gov/div898/strd/nls/data/chwirut2.shtml>



El modelo teórico propuesto es

$$f(x, \boldsymbol{\theta}) = \frac{\exp(-\theta_1 x)}{\theta_2 + \theta_3 x}.$$

El estudio de los modelos de regresión no lineal es muy extenso y complejo, existiendo una amplia literatura sobre el tema. Textos de referencia son los de Bates y Watts [3] y el de Seber y Wild [25].

El modelo de regresión no lineal con covariables fijas es de la forma

$$\begin{aligned} y_i &= f(\mathbf{x}_i, \boldsymbol{\theta}^*) + \varepsilon_i & i = 1, \dots, n \\ \mathbb{E}(\varepsilon_i) &= 0, \end{aligned}$$

donde y_i , $1 \leq i \leq n$ son independientes.

Para la i -ésima observación tenemos la variable respuesta y_i , el vector de efectos fijos \mathbf{x}_i , de dimensión k y los errores ε_i que se suponen independientes e idénticamente distribuidos con media 0 y varianza σ^2 desconocida. El vector de parámetros $\boldsymbol{\theta}^* \in \Theta \subset \mathbb{R}^p$, es desconocido.

El estimador clásico $\hat{\boldsymbol{\theta}}$ del vector de parámetros $\boldsymbol{\theta}^*$, se halla como el valor que minimiza la suma de cuadrados $S(\boldsymbol{\theta})$

$$S(\boldsymbol{\theta}) = \sum_{i=1}^n (y_i - f(\mathbf{x}_i, \boldsymbol{\theta}))^2, \quad \boldsymbol{\theta} \in \Theta.$$

A diferencia del modelo de regresión lineal, la minimización resulta en un problema que no tiene solución explícita, debido a la no linealidad de la función $f(\mathbf{x}, \boldsymbol{\theta})$. Por esta razón, necesitamos métodos numéricos de minimización no lineal. Estos son métodos iterativos que, en forma ideal, nos acercan a la solución óptima en un número determinado de pasos. El algoritmo más utilizado es el método de Gauss–Newton.

Las mayores complicaciones que encontramos al utilizar estos métodos son:

- Cómo iniciar el procedimiento y la forma de elegir el parámetro inicial.
- Cómo garantizar que el procedimiento alcanzó el mínimo global y no un mínimo local.

Estos dos temas están relacionados entre sí ya que si los valores de los parámetros iniciales están lo suficientemente cerca de los valores de los parámetros óptimos, entonces el procedimiento converge en unos pocos pasos. Por otro lado, una mala elección del valor inicial puede derivar en la falta de convergencia. Si dicha falta persiste independientemente de la elección de los valores de partida, podría indicar que el modelo en su forma actual no es adecuado para los datos en estudio.

El estimador $\hat{\theta}$ es óptimo siempre y cuando los errores tengan distribución normal. Si la distribución no es exactamente normal, resulta sensible a la presencia de observaciones atípicas, lo que puede provocar que el ajuste no resulte adecuado. Por esta razón se utilizan estimadores robustos que son más estables ante perturbaciones del modelo y que, a su vez, son altamente eficientes bajo normalidad.

En este trabajo comenzaremos con una breve descripción de la teoría sobre modelos no lineales: métodos de estimación, diversos problemas relacionados con la convergencia y validez de la inferencia, medidas de curvatura y no linealidad e inferencia.

Después de una breve descripción sobre las funciones *nls()* y *nlrob()* del programa *R* y de la función desarrollada por Stromberg para *MM*-estimadores, finalizaremos con el análisis de sus comportamientos sobre varios conjuntos de datos.

Capítulo 2

Métodos de estimación

Dados n pares de observaciones (y_i, \mathbf{x}_i) , el modelo de regresión no lineal con regresor fijo es de la forma

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}^*) + \varepsilon_i \quad (2.0.1)$$

donde los errores ε_i son independientes tales que $\mathbb{E}(\varepsilon_i) = 0$ para $1 \leq i \leq n$. El vector de parámetros $\boldsymbol{\theta}^* \in \Theta \subseteq \mathbb{R}^p$ es desconocido.

Existen varios métodos para estimar el valor de $\boldsymbol{\theta}^*$, con distintas propiedades de acuerdo a la información que tengamos sobre las observaciones: Mínimos cuadrados no lineal, Máxima verosimilitud, Cuasi-verosimilitud y Métodos robustos. En este capítulo haremos una revisión de estos procedimientos.

2.1. Mínimos cuadrados no lineal

El estimador de mínimos cuadrados $\hat{\boldsymbol{\theta}}$, se obtiene al minimizar la suma de cuadrados:

$$S(\boldsymbol{\theta}) = \sum_{i=1}^n (y_i - f(\mathbf{x}_i; \boldsymbol{\theta}))^2 = \|\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})\|^2 \quad \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$$

donde

$$\begin{aligned} \mathbf{y} &= (y_1, \dots, y_n)^T \\ f_i(\boldsymbol{\theta}) &= f(\mathbf{x}_i; \boldsymbol{\theta}) \\ \mathbf{f}(\boldsymbol{\theta}) &= (f_1(\boldsymbol{\theta}), f_2(\boldsymbol{\theta}), \dots, f_n(\boldsymbol{\theta}))^T. \end{aligned}$$

Bajo condiciones de regularidad, que incluyen que los errores ε_i , $1 \leq i \leq n$ sean independientes e idénticamente distribuidos con varianza σ^2 , tenemos que $\hat{\boldsymbol{\theta}}$ y $s^2 = S(\hat{\boldsymbol{\theta}})/(n-p)$ son estimadores consistentes de $\boldsymbol{\theta}^*$ y σ^2 , respectivamente.

Si llamamos

$$\mathbf{F}_\bullet(\boldsymbol{\theta}) = \frac{\partial \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \left[\left(\frac{\partial f_i(\boldsymbol{\theta})}{\partial \theta_j} \right) \right]$$

y

$$\mathbf{F}_\bullet = \mathbf{F}_\bullet(\boldsymbol{\theta}^*) \quad \widehat{\mathbf{F}}_\bullet = \mathbf{F}_\bullet(\widehat{\boldsymbol{\theta}}),$$

cuando $f(\mathbf{x}_i; \boldsymbol{\theta})$ es diferenciable con respecto a $\boldsymbol{\theta}$, tenemos las ecuaciones

$$\left. \frac{\partial S(\boldsymbol{\theta})}{\partial \theta_r} \right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \sum_{i=1}^n 2(y_i - f_i(\boldsymbol{\theta}))(-1) \left. \frac{\partial f_i(\boldsymbol{\theta})}{\partial \theta_r} \right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = 0 \quad r = 1, \dots, p,$$

lo que equivale a tener la ecuación

$$\mathbf{0} = \widehat{\mathbf{F}}_\bullet^T (\mathbf{y} - \mathbf{f}(\widehat{\boldsymbol{\theta}})) = \widehat{\mathbf{F}}_\bullet^T \widehat{\boldsymbol{\epsilon}}.$$

Igual que en la teoría lineal, si $\widehat{\mathbf{P}}_{\mathbf{F}} = \widehat{\mathbf{F}}_\bullet (\widehat{\mathbf{F}}_\bullet^T \widehat{\mathbf{F}}_\bullet)^{-1} \widehat{\mathbf{F}}_\bullet^T$ es la matriz idempotente que proyecta ortogonalmente \mathbb{R}^n sobre el espacio columna de $\widehat{\mathbf{F}}_\bullet$, la ecuación anterior deriva en la ecuación normal para el modelo no lineal

$$\widehat{\mathbf{P}}_{\mathbf{F}} \widehat{\boldsymbol{\epsilon}} = \mathbf{0}.$$

Al ser $f(\mathbf{x}; \boldsymbol{\theta})$ una función no lineal en $\boldsymbol{\theta}$, no resulta sencillo hallar una solución explícita. Esto nos obliga al uso de métodos iterativos para la solución del problema, siendo uno de los habituales el de *Gauss-Newton*.

Este se basa en la “linealización” de la función $\mathbf{f}(\boldsymbol{\theta})$ en un entorno de $\boldsymbol{\theta}^*$, verdadero valor de $\boldsymbol{\theta}$.

Para esto, hagamos un desarrollo de Taylor de primer orden de $f_i(\boldsymbol{\theta})$ en $\boldsymbol{\theta}^*$:

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

que en forma matricial resulta en la ecuación

$$\mathbf{f}(\boldsymbol{\theta}) \approx \mathbf{f}(\boldsymbol{\theta}^*) + \mathbf{F}_\bullet (\boldsymbol{\theta} - \boldsymbol{\theta}^*). \quad (2.1.1)$$

Si llamamos a $\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{f}(\boldsymbol{\theta}^*)$ y $\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$ tenemos que

$$\begin{aligned} S(\boldsymbol{\theta}) &= \|\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})\|^2 \\ &\approx \|\boldsymbol{\epsilon} - \mathbf{F}_\bullet \boldsymbol{\beta}\|^2, \end{aligned}$$

y en forma similar que en el modelo lineal, vemos que el mínimo del lado derecho se obtiene en $\widehat{\boldsymbol{\beta}}$ donde

$$\widehat{\boldsymbol{\beta}} = (\mathbf{F}_\bullet^T \mathbf{F}_\bullet)^{-1} \mathbf{F}_\bullet^T \boldsymbol{\epsilon}.$$

Cuando n es grande y bajo ciertas condiciones de regularidad, tenemos que con alta probabilidad $\widehat{\boldsymbol{\theta}}$ está en un entorno pequeño de $\boldsymbol{\theta}^*$. Por lo tanto, como $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \approx \widehat{\boldsymbol{\beta}}$, nos queda que

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \approx (\mathbf{F}_\bullet^T \mathbf{F}_\bullet)^{-1} \mathbf{F}_\bullet^T \boldsymbol{\epsilon}. \quad (2.1.2)$$

Tomando $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, de (2.1.1) deducimos que

$$\begin{aligned} \mathbf{f}(\widehat{\boldsymbol{\theta}}) - \mathbf{f}(\boldsymbol{\theta}^*) &\approx \mathbf{F}_\bullet (\mathbf{F}_\bullet^T \mathbf{F}_\bullet)^{-1} \mathbf{F}_\bullet^T \boldsymbol{\epsilon} = \mathbf{P}_{\mathbf{F}} \boldsymbol{\epsilon} \\ \mathbf{y} - \mathbf{f}(\widehat{\boldsymbol{\theta}}) &\approx \boldsymbol{\epsilon} - \mathbf{P}_{\mathbf{F}} \boldsymbol{\epsilon} = (\mathbf{I}_n - \mathbf{P}_{\mathbf{F}}) \boldsymbol{\epsilon}. \end{aligned}$$

De esta forma, tenemos que

$$(n-p)s^2 \approx \boldsymbol{\epsilon}^T(\mathbf{I}_n - \mathbf{P}_F)\boldsymbol{\epsilon} \quad (2.1.3)$$

$$\left\| \mathbf{f}(\hat{\boldsymbol{\theta}}) - \mathbf{f}(\boldsymbol{\theta}^*) \right\|^2 \approx \boldsymbol{\epsilon}^T \mathbf{P}_F \boldsymbol{\epsilon}. \quad (2.1.4)$$

Usando (2.1.3) y (2.1.4), resulta que

$$\begin{aligned} S(\boldsymbol{\theta}^*) - S(\hat{\boldsymbol{\theta}}) &\approx \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} - \boldsymbol{\epsilon}^T(\mathbf{I}_n - \mathbf{P}_F)\boldsymbol{\epsilon} \\ &= \boldsymbol{\epsilon}^T \mathbf{P}_F \boldsymbol{\epsilon} \\ &\approx (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \mathbf{F}_\bullet^T \mathbf{F}_\bullet (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*). \end{aligned} \quad (2.1.5)$$

Podemos formalizar, los cálculos anteriores en el siguiente resultado:

Teorema 2.1.1. *Dado $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ y con condiciones de regularidad apropiadas, para n grande, tenemos*

$$(i) \quad \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \sim N_p(\mathbf{0}, \sigma^2 \mathbf{C}^{-1}) \text{ donde } \mathbf{C} = \mathbf{F}_\bullet^T \mathbf{F}_\bullet = \mathbf{F}_\bullet^T(\boldsymbol{\theta}^*) \mathbf{F}_\bullet(\boldsymbol{\theta}^*).$$

$$(ii) \quad \frac{(n-p)s^2}{\sigma^2} \approx \frac{\boldsymbol{\epsilon}^T(\mathbf{I}_n - \mathbf{P}_F)\boldsymbol{\epsilon}}{\sigma^2} \sim \chi_{n-p}^2.$$

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

$$(iv) \quad \frac{[S(\boldsymbol{\theta}^*) - S(\hat{\boldsymbol{\theta}})]/p}{S(\hat{\boldsymbol{\theta}})/(n-p)} \approx \frac{\boldsymbol{\epsilon}^T \mathbf{P}_F \boldsymbol{\epsilon}}{\boldsymbol{\epsilon}^T(\mathbf{I}_n - \mathbf{P}_F)\boldsymbol{\epsilon}} \cdot \frac{n-p}{p} \sim \mathcal{F}_{p, n-p}.$$

Por (iv) y (2.1.5) tenemos que, aproximadamente

$$\frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \mathbf{F}_\bullet^T \mathbf{F}_\bullet (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)}{p s^2} \sim \mathcal{F}_{p, n-p}. \quad (2.1.6)$$

Si reemplazamos \mathbf{F}_\bullet por $\hat{\mathbf{F}}_\bullet$, podemos obtener una región de confianza aproximada con un nivel $100(1 - \alpha)\%$ para $\boldsymbol{\theta}$ como

$$\{\boldsymbol{\theta} : (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \hat{\mathbf{F}}_\bullet^T \hat{\mathbf{F}}_\bullet (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leq p s^2 \mathcal{F}_{p, n-p, \alpha}\},$$

donde $\mathbb{P}(\mathcal{F}_{p, n-p} \geq \mathcal{F}_{p, n-p, \alpha}) = \alpha$.

2.1.1. Métodos Numéricos

Los procedimientos más usados son el de *Gauss-Newton* y el de *Newton*.

Método de Gauss–Newton. Este procedimiento se basa en la aproximación lineal de la función mediante un desarrollo de Taylor de primer orden de la función f .

Llamemos $\boldsymbol{\theta}^{(a)}$ a una aproximación del estimador de mínimos cuadrados $\widehat{\boldsymbol{\theta}}$ del modelo no lineal

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}^*) + \varepsilon_i.$$

Para $\boldsymbol{\theta}$ cercano a $\boldsymbol{\theta}^{(a)}$ sabemos que es válida la aproximación lineal de Taylor

$$\mathbf{f}(\boldsymbol{\theta}) \approx \mathbf{f}(\boldsymbol{\theta}^{(a)}) + \mathbf{F}_{\bullet}^{(a)}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)})$$

donde $\mathbf{F}_{\bullet}^{(a)} = \mathbf{F}_{\bullet}(\boldsymbol{\theta}^{(a)})$.

Así, al vector de residuos $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{y} - \mathbf{f}(\boldsymbol{\theta})$ lo podemos aproximar por $\mathbf{r}(\boldsymbol{\theta}^{(a)}) - \mathbf{F}_{\bullet}^{(a)}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)})$, que sugiere la aproximación

$$S(\boldsymbol{\theta}) \approx \mathbf{r}(\boldsymbol{\theta}^{(a)})^T \mathbf{r}(\boldsymbol{\theta}^{(a)}) - 2\mathbf{r}(\boldsymbol{\theta}^{(a)})^T \mathbf{F}_{\bullet}^{(a)}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}) + (\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)})^T \mathbf{F}_{\bullet}^{(a)T} \mathbf{F}_{\bullet}^{(a)}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}). \quad (2.1.7)$$

El lado derecho tiene un mínimo con respecto a $\boldsymbol{\theta}$ cuando

$$\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)} = (\mathbf{F}_{\bullet}^{(a)T} \mathbf{F}_{\bullet}^{(a)})^{-1} \mathbf{F}_{\bullet}^{(a)T} \mathbf{r}(\boldsymbol{\theta}^{(a)}) = \boldsymbol{\delta}^{(a)}.$$

Así, dada una aproximación $\boldsymbol{\theta}^{(a)}$, la siguiente podría ser

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

Este procedimiento es el que se conoce como *el método de Gauss–Newton*.

Para $\boldsymbol{\theta}^{(1)}$ cercano a $\boldsymbol{\theta}^*$ y el tamaño de la muestra suficientemente grande, tenemos que $\boldsymbol{\theta}^{(a)} \rightarrow \widehat{\boldsymbol{\theta}}$ cuando $a \rightarrow \infty$.

Método de Newton. Este es un método más general que el Gauss-Newton y se basa en el desarrollo de Taylor de segundo orden de $S(\boldsymbol{\theta})$.

Llamando

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

al gradiente de $S(\boldsymbol{\theta})$ y

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

al Hessiano de $S(\boldsymbol{\theta})$, tenemos la aproximación cuadrática

$$S(\boldsymbol{\theta}) \approx S(\boldsymbol{\theta}^{(a)}) + \mathbf{g}(\boldsymbol{\theta}^{(a)})^T (\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}) + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)})^T \mathbf{H}(\boldsymbol{\theta}^{(a)}) (\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}). \quad (2.1.8)$$

Como

$$\frac{\partial^2 S(\boldsymbol{\theta})}{\partial \theta_r \partial \theta_s} = 2 \sum_{i=1}^n \left\{ \frac{\partial f_i(\boldsymbol{\theta})}{\partial \theta_r} \frac{\partial f_i(\boldsymbol{\theta})}{\partial \theta_s} - (y_i - f_i(\boldsymbol{\theta})) \frac{\partial^2 f_i(\boldsymbol{\theta})}{\partial \theta_r \partial \theta_s} \right\},$$

usamos como aproximación de $\mathbf{H}(\boldsymbol{\theta}^{(a)})$ su valor esperado en $\boldsymbol{\theta}$

$$\mathbb{E} \left(\frac{\partial^2 S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right) = 2 \mathbf{F}_\bullet^T(\boldsymbol{\theta}) \mathbf{F}_\bullet(\boldsymbol{\theta}).$$

El mínimo con respecto a $\boldsymbol{\theta}$ en la aproximación cuadrática (2.1.8) ocurre cuando

$$\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)} = - \left[\mathbf{H}(\boldsymbol{\theta}^{(a)}) \right]^{-1} \mathbf{g}(\boldsymbol{\theta}^{(a)}) = - \left[\mathbf{H}^{-1} \mathbf{g} \right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(a)}}.$$

Así, el paso es $\boldsymbol{\delta}^{(a)} = - \left[\mathbf{H}^{-1} \mathbf{g} \right]_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(a)}}.$

2.1.2. Mínimos cuadrados generalizados

Este método se utiliza para resolver el problema de regresión no lineal cuando tenemos $\text{VAR}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{V}$, donde \mathbf{V} es una matriz definida positiva conocida.

En esta situación, tenemos que minimizar la suma de cuadrados

$$S_{\text{GLS}}(\boldsymbol{\theta}) = [\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})]^T \mathbf{V}^{-1} [\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})].$$

Usando la descomposición de Cholesky de la matriz \mathbf{V} , podemos transformar el modelo no lineal en un problema de Mínimos Cuadrados Ordinarios (OLS).

Si $\mathbf{V} = \mathbf{U}^T \mathbf{U}$ es la descomposición de Cholesky de \mathbf{V} , con \mathbf{U} una matriz triangular superior, llamando \mathbf{R} a $(\mathbf{U}^T)^{-1}$, tenemos que

$$\mathbf{R}\mathbf{y} = \mathbf{R}\mathbf{f}(\boldsymbol{\theta}) + \mathbf{R}\boldsymbol{\epsilon}.$$

Si $\mathbf{Z} = \mathbf{R}\mathbf{y}$, $\mathcal{K}(\boldsymbol{\theta}) = \mathbf{R}\mathbf{f}(\boldsymbol{\theta})$ y $\boldsymbol{\eta} = \mathbf{R}\boldsymbol{\epsilon}$, tenemos

$$\mathbf{Z} = \mathcal{K}(\boldsymbol{\theta}) + \boldsymbol{\eta},$$

con $\mathbb{E}(\boldsymbol{\eta}) = \mathbf{R}\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ y $\text{VAR}[\boldsymbol{\eta}] = \sigma^2 \mathbf{R}\mathbf{V}\mathbf{R}^T = \sigma^2 \mathbf{I}_n.$

Así, la suma de cuadrados del método de mínimos cuadrados generalizados (GLS), es la misma que la del método de mínimos cuadrados ordinarios (OLS) del problema transformado.

$$\begin{aligned} S_{\text{GLS}}(\boldsymbol{\theta}) &= [\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})]^T \mathbf{V}^{-1} [\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})] \\ &= [\mathbf{Z} - \mathcal{K}(\boldsymbol{\theta})]^T [\mathbf{Z} - \mathcal{K}(\boldsymbol{\theta})] \end{aligned}$$

y el estimador OLS ($\hat{\boldsymbol{\theta}}_{\text{OLS}}$) para el modelo transformado, es también el estimador GLS ($\hat{\boldsymbol{\theta}}_{\text{GLS}}$) del modelo sin transformar.

Si llamamos $\mathcal{K}_\bullet(\boldsymbol{\theta}) = \frac{\partial \mathcal{K}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \mathbf{R} \frac{\partial \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \mathbf{R}\mathbf{F}_\bullet(\boldsymbol{\theta})$ y $\hat{\mathcal{K}} = \mathcal{K}_\bullet(\hat{\boldsymbol{\theta}})$, tenemos que para n grande, la matriz de varianza-covarianza de $\hat{\boldsymbol{\theta}}_{\text{GLS}}$ estará dada por

$$\begin{aligned} \text{VAR}(\hat{\boldsymbol{\theta}}_{\text{GLS}}) &\approx \sigma^2 \left[\mathcal{K}_\bullet^T(\boldsymbol{\theta}^*) \mathcal{K}_\bullet(\boldsymbol{\theta}^*) \right]^{-1} \\ &= \sigma^2 \left[\mathbf{F}_\bullet^T(\boldsymbol{\theta}^*) \mathbf{V}^{-1} \mathbf{F}_\bullet(\boldsymbol{\theta}^*) \right]^{-1} \end{aligned}$$

que puede estimarse por

$$\widehat{\text{VAR}}(\widehat{\boldsymbol{\theta}}_{\text{GLS}}) = \hat{\sigma}^2 [\mathcal{K}_{\bullet}^{\text{T}} \mathcal{K}_{\bullet}]^{-1} = \hat{\sigma}^2 \left(\widehat{\mathbf{F}}_{\bullet}^{\text{T}} \mathbf{V}^{-1} \widehat{\mathbf{F}}_{\bullet} \right)^{-1}.$$

Con la aproximación $\mathbf{f}(\boldsymbol{\theta}) \approx \mathbf{f}(\boldsymbol{\theta}^{(a)}) + \mathbf{F}_{\bullet}^{(a)}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)})$, nos queda que $S_{\text{GLS}}(\boldsymbol{\theta})$ es, aproximadamente,

$$\left(\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}^{(a)}) - \mathbf{F}_{\bullet}^{(a)}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}) \right)^{\text{T}} \mathbf{V}^{-1} \left(\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}^{(a)}) - \mathbf{F}_{\bullet}^{(a)}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)}) \right).$$

Una aproximación para el valor mínimo del problema de Mínimos Cuadrados Generalizados se obtiene entonces mediante

$$\boldsymbol{\theta} - \boldsymbol{\theta}^{(a)} = \left(\mathbf{F}_{\bullet}^{(a)\text{T}} \mathbf{V}^{-1} \mathbf{F}_{\bullet}^{(a)} \right)^{-1} \mathbf{F}_{\bullet}^{(a)\text{T}} \mathbf{V}^{-1} \left(\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}^{(a)}) \right).$$

Esto lleva al método iterativo $\boldsymbol{\theta}^{(a+1)} = \boldsymbol{\theta}^{(a)} + \boldsymbol{\delta}^{(a)}$ donde

$$\begin{aligned} \boldsymbol{\delta}^{(a)} &= \left(\mathbf{F}_{\bullet}^{(a)\text{T}} \mathbf{V}^{-1} \mathbf{F}_{\bullet}^{(a)} \right)^{-1} \mathbf{F}_{\bullet}^{(a)\text{T}} \mathbf{V}^{-1} \left(\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}^{(a)}) \right) \\ &= \left(\mathcal{K}_{\bullet}^{(a)\text{T}} \mathcal{K}_{\bullet}^{(a)} \right)^{-1} \mathcal{K}_{\bullet}^{(a)\text{T}} \left(\mathbf{Z} - \mathcal{K}(\boldsymbol{\theta}^{(a)}) \right) \end{aligned}$$

donde la última ecuación es el método de Gauss-Newton para el modelo transformado.

2.2. Estimadores de máxima verosimilitud

Supongamos que los ε_i son i.i.d con función de densidad $\sigma^{-1}g\left(\frac{\varepsilon}{\sigma}\right)$, donde g es la distribución de errores estandarizados con varianza 1. La función de verosimilitud está dada por

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

El *Estimador de Máxima Verosimilitud*, $\widehat{\boldsymbol{\theta}}_{\text{EMV}}$, de $\boldsymbol{\theta}$ puede obtenerse maximizando la función de verosimilitud.

En particular, cuando los errores tienen distribución $N(0, \sigma^2)$, tenemos la función de verosimilitud

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

y la función de log-verosimilitud está dada por

$$\begin{aligned} L(\boldsymbol{\theta}, \sigma^2) &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - f(\mathbf{x}_i, \boldsymbol{\theta}))^2 \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} S(\boldsymbol{\theta}). \end{aligned}$$

Dado σ^2 , el máximo en $\boldsymbol{\theta}$ se obtiene cuando $S(\boldsymbol{\theta})$ es mínimo. O sea, que el EMV de $\boldsymbol{\theta}$ resulta ser el estimador de mínimos cuadrados $\hat{\boldsymbol{\theta}}_{\text{EMV}} = \hat{\boldsymbol{\theta}}_{\text{OLS}}$.

Dado $\boldsymbol{\theta}$, tenemos que $\partial L(\boldsymbol{\theta}, \sigma^2)/\partial \sigma^2 = 0$ cuando $\sigma^2 = S(\boldsymbol{\theta})/n$. Luego, $\hat{\sigma} = S(\hat{\boldsymbol{\theta}}_{\text{EMV}})/n$ y $\hat{\boldsymbol{\theta}}_{\text{EMV}}$ maximizan la función de log-verosimilitud $L(\boldsymbol{\theta}, \sigma^2)$ y bajo condiciones apropiadas de regularidad de la función f , $\hat{\boldsymbol{\theta}}_{\text{EMV}}$ es asintóticamente eficiente. Cuando los datos no son normales, el EMV no coincide con el de mínimos cuadrados por la dependencia de la función de verosimilitud L de $\mathbf{f}(\boldsymbol{\theta})$.

Debido a esta dependencia escribimos,

$$\begin{aligned} L(\boldsymbol{\theta}, \sigma^2) &= \log(p(y|\boldsymbol{\theta}, \sigma^2)) \\ &= -\frac{n}{2} \log(\sigma^2) + \sum_{i=1}^n \log \left(g \left(\frac{y_i - f(\mathbf{x}_i, \boldsymbol{\theta})}{\sigma} \right) \right) \\ &= L(\mathbf{f}(\boldsymbol{\theta})). \end{aligned}$$

En este caso, suponemos que el modelo es suficientemente regular como para que el EMV de $\hat{\boldsymbol{\theta}}_{\text{EMV}}$ resulte ser la solución de

$$\mathbf{0} = \frac{\partial L}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}^T} \right)^T \frac{\partial L}{\partial \mathbf{f}}.$$

Si $\boldsymbol{\theta}^{(a)}$ es la a -ésima aproximación de $\hat{\boldsymbol{\theta}}_{\text{EMV}}$, dado el desarrollo de Taylor de $\frac{\partial L}{\partial \boldsymbol{\theta}}$ en $\hat{\boldsymbol{\theta}}_{\text{EMV}}$, tenemos que la estimación $\boldsymbol{\theta}^{(a+1)} = \boldsymbol{\theta}^{(a)} + \boldsymbol{\delta}^{(a)}$ esta dada por

$$\left[\left(-\frac{\partial^2 L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right)^{-1} \frac{\partial L}{\partial \boldsymbol{\theta}} \right]_{\boldsymbol{\theta}^{(a)}} = \boldsymbol{\delta}^{(a)}.$$

El método de Newton falla ya que la matriz $-\frac{\partial^2 L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$ no es definida positiva para todo $\boldsymbol{\theta}^{(a)}$.

El *método de Scoring de Fisher* reemplaza dicha matriz por su valor esperado

$$\mathbb{E} \left[-\frac{\partial^2 L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right] = \mathbf{F}^T \mathbf{G} \mathbf{F},$$

donde

$$\mathbf{G} = \mathbb{E} \left[\frac{\partial L}{\partial \mathbf{f}} \frac{\partial L}{\partial \mathbf{f}^T} \right]$$

es una matriz definida positiva.

Esto da origen al procedimiento de estimación

$$\boldsymbol{\theta}^{(a+1)} = \boldsymbol{\theta}^{(a)} + \left[(\mathbf{F}^T \mathbf{G} \mathbf{F})^{-1} \frac{\partial L}{\partial \boldsymbol{\theta}} \right]_{\boldsymbol{\theta}^{(a)}}$$

que por ser $\mathbf{F}^T \mathbf{G} \mathbf{F}$ definida positiva, asegura que el paso $\boldsymbol{\delta}^{(a)} = \boldsymbol{\theta}^{(a+1)} - \boldsymbol{\theta}^{(a)}$ siempre tiene dirección ascendente.

Este método tiene la ventaja de usar derivadas de primer orden a cambio de tener una convergencia más lenta que el método de Newton.

Ratkowsky y Dolby [18] probaron que este algoritmo y el método de Gauss-Newton son equivalentes.

2.2.1. Método de Verosimilitud Concentrado

Para encontrar estimadores de máxima verosimilitud de funciones de verosimilitud en general, a veces es conveniente utilizar este método de maximización paso a paso.

Llamamos $L(\boldsymbol{\theta}, \boldsymbol{\tau}|\mathbf{z})$ a la función de log-verosimilitud a maximizar con respecto a $\boldsymbol{\theta}$ y $\boldsymbol{\tau}$, donde \mathbf{z} son los datos y $\boldsymbol{\tau}$ es un vector q -dimensional de parámetros adicionales o de ruido. Suponemos que para cada conjunto de datos \mathbf{z} , L tiene un único máximo con respecto a $\boldsymbol{\theta}$ y a $\boldsymbol{\tau}$.

El primer paso consiste en encontrar $\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}, \mathbf{z})$, el único valor de $\boldsymbol{\tau}$ que maximiza L considerando a $\boldsymbol{\theta}$ como una constante.

El segundo paso es reemplazar $\boldsymbol{\tau}$ por $\tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}, \mathbf{z})$ en $L(\boldsymbol{\theta}, \boldsymbol{\tau}|\mathbf{z})$. Esto nos da la función de log-verosimilitud concentrada

$$M(\boldsymbol{\theta}) = M(\boldsymbol{\theta}, \mathbf{z}) \equiv L[\boldsymbol{\theta}, \tilde{\boldsymbol{\tau}}(\boldsymbol{\theta}, \mathbf{z})|\mathbf{z}].$$

En el tercer paso utilizamos $M(\boldsymbol{\theta})$ como la verdadera función de log-verosimilitud, que depende solo de $\boldsymbol{\theta}$ y aplicamos el método de máxima verosimilitud habitual para hallar $\hat{\boldsymbol{\theta}}$.

Tenemos varias ventajas al usar $M(\boldsymbol{\theta})$ en lugar de $L(\boldsymbol{\theta}, \boldsymbol{\tau}|\mathbf{z})$. La primera es que obtenemos $\hat{\boldsymbol{\theta}}$ como solución de $\partial M(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = 0$, eliminando el vector de ruido $\boldsymbol{\tau}$. La segunda es que, bajo ciertas condiciones, un estimador asintótico de la matriz de varianza-covarianza de $\hat{\boldsymbol{\delta}} = (\hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\tau}}^T)^T$ es

$$\left\{ -\frac{\partial^2 L(\boldsymbol{\delta})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T} \right\}^{-1} \hat{\boldsymbol{\delta}},$$

una matriz de dimensión $(p+q) \times (p+q)$ que puede ser partida para obtener un estimador de $\text{VAR}[\hat{\boldsymbol{\theta}}]$.

Sin embargo, si solo estamos interesados en $\hat{\boldsymbol{\theta}}$, podemos estimar $\text{VAR}[\hat{\boldsymbol{\theta}}]$ por medio de la matriz $\left\{ -\frac{\partial^2 M(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right\}^{-1} \hat{\boldsymbol{\theta}}$.

2.3. Estimación de cuasi-verosimilitud

El estimador de cuasi-verosimilitud tiene propiedades similares al EMV, con la diferencia que no necesitamos conocer la función de distribución de \mathbf{y} , solo sus dos primeros momentos.

Recordemos que $\mathbb{E}(\mathbf{Y}) = \mathbf{f}(\boldsymbol{\theta}) = \mathbf{f}$ y supongamos más generalmente que $\text{VAR}(\mathbf{Y}) = \sigma^2 \mathbf{V}(\mathbf{f})$.

La función de log cuasi-verosimilitud $l(\mathbf{f})$ esta definida por el sistema de ecuaciones diferenciales

$$\frac{\partial l(\mathbf{f})}{\partial \mathbf{f}} = \mathbf{V}^{-1}(\mathbf{f})(\mathbf{y} - \mathbf{f}). \quad (2.3.1)$$

Por comodidad, usamos $l(\boldsymbol{\theta})$ y $\mathbf{V}(\boldsymbol{\theta})$ en lugar de $l[\mathbf{f}(\boldsymbol{\theta})]$ y $\mathbf{V}[\mathbf{f}(\boldsymbol{\theta})]$. El estimador de cuasi-verosimilitud, $\hat{\boldsymbol{\theta}}$, es solución de la ecuación

$$\mathbf{0} = \frac{\partial l}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}^T} \right)^T \frac{\partial l}{\partial \mathbf{f}} = \mathbf{F}^T(\boldsymbol{\theta}) \mathbf{V}^{-1}(\boldsymbol{\theta}) (\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})). \quad (2.3.2)$$

Para resolver esta ecuación podemos utilizar el algoritmo de Gauss–Newton, usando el desarrollo de Taylor de $\mathbf{f}(\check{\boldsymbol{\theta}})$ alrededor de $\boldsymbol{\theta}^{(a)}$ en (2.3.2). Aproximando $\mathbf{F}_\bullet(\boldsymbol{\theta})$ por $\mathbf{F}_\bullet^{(a)} = \mathbf{F}_\bullet(\boldsymbol{\theta}^{(a)})$ y a $\mathbf{V}(\boldsymbol{\theta})$ por $\mathbf{V}^{(a)} = \mathbf{V}(\boldsymbol{\theta}^{(a)})$, obtenemos

$$\check{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(a)} \approx \left((\mathbf{F}_\bullet^{(a)})^\top (\mathbf{V}^{(a)})^{-1} \mathbf{F}_\bullet^{(a)} \right)^{-1} (\mathbf{F}_\bullet^{(a)})^\top (\mathbf{V}^{(a)})^{-1} [\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}^{(a)})]. \quad (2.3.3)$$

Llamando $\boldsymbol{\delta}^{(a)}$ al miembro derecho de (2.3.3), tenemos $\boldsymbol{\theta}^{(a+1)} = \boldsymbol{\theta}^{(a)} + \boldsymbol{\delta}^{(a)}$.

El paso $\boldsymbol{\delta}^{(a)}$ lo obtenemos utilizando la técnica de mínimos cuadrados generalizados para minimizar la expresión

$$[\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}^{(a)}) - \mathbf{F}_\bullet^{(a)} \boldsymbol{\delta}]^\top (\mathbf{V}^{(a)})^{-1} [\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}^{(a)}) - \mathbf{F}_\bullet^{(a)} \boldsymbol{\delta}]$$

con respecto a $\boldsymbol{\delta}$.

Si $\check{\boldsymbol{\theta}}$ cumple ciertas condiciones (Seber [25] - pag 43; McCullagh [16]) se puede probar que

a)
$$\sqrt{n}(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \approx N_p(\mathbf{0}, \sigma^2 [n^{-1} \mathbf{I}_n(\boldsymbol{\theta}^*)]^{-1}) + O_{\mathbb{P}}(n^{-1/2}) \quad (2.3.4)$$

b) Si el tercer momento es infinito entonces en (2.3.4) tenemos $o_{\mathbb{P}}(1)$ en lugar de $O_{\mathbb{P}}(n^{-1/2})$.

c) $\check{\sigma}^2 = [\mathbf{y} - \mathbf{f}(\check{\boldsymbol{\theta}})]^\top \mathbf{V}^{-1}(\boldsymbol{\theta}) [\mathbf{y} - \mathbf{f}(\check{\boldsymbol{\theta}})] / (n - p)$

d) Entre todos los estimadores $\check{\boldsymbol{\theta}}_L$ que cumplen $\check{\boldsymbol{\theta}}_L - \boldsymbol{\theta}^* = \mathbf{L}^\top (\mathbf{y} - \mathbf{f}(\boldsymbol{\theta}^*)) + o_{\mathbb{P}}(n^{-1/2})$, con $\mathbf{L} \in \mathbb{R}^{n \times p}$ cuyos elementos son funciones de $\boldsymbol{\theta}^*$, $\check{\boldsymbol{\theta}}$ tiene la matriz de dispersión “mínima”.

Observaciones: Una de las ventajas de este procedimiento es que no se requiere una familia de distribuciones para la cual $l(\mathbf{f})$ sea de máxima verosimilitud. Por otra parte, no se requiere la forma explícita de $l(\mathbf{f})$ para el cálculo de $\check{\boldsymbol{\theta}}$. A veces, una solución para $l(\mathbf{f})$ puede encontrarse mediante la construcción de funciones $\boldsymbol{\beta}(\mathbf{f})$ y $c(\boldsymbol{\beta})$ tales que

$$\mathbf{f} = \frac{\partial c}{\partial \boldsymbol{\beta}} \quad V(\mathbf{f}) = \frac{\partial^2 c}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \quad l(\mathbf{f}) = \boldsymbol{\beta}^\top \mathbf{y} - c(\boldsymbol{\beta}) - d(\mathbf{y}, \sigma)$$

con $d(\mathbf{y}, \sigma)$ arbitrario. Para ver que esta función $l(\mathbf{f})$ es solución de (2.3.1), bastará ver que

$$\frac{\partial \boldsymbol{\beta}}{\partial \mathbf{f}^\top} = \left(\frac{\partial \mathbf{f}}{\partial \boldsymbol{\beta}^\top} \right)^{-1} = \left(\frac{\partial^2 c}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \right)^{-1} = \mathbf{V}^{-1}(\mathbf{f}).$$

2.4. Métodos Robustos

En la mayoría de las aplicaciones, muy rara vez podemos asegurar, que los mecanismos aleatorios que producen los datos, provienen de un modelo normal.

Para reflejar esto suponemos que la distribución F pertenece al entorno de contaminación de radio ϵ , ν_ϵ , de la distribución normal definido como

$$\nu_\epsilon = \{F : F = (1 - \epsilon)\Phi + \epsilon F^* \text{ con } F^* \text{ arbitraria}\}.$$

El estimador de mínimos cuadrados no lineal no funciona bien en este tipo de distribuciones: los valores atípicos generados pueden tener una gran influencia sobre los resultados obtenidos por este método.

Para poder amortiguar el efecto de las observaciones atípicas podemos utilizar un procedimiento robusto que tiende a dejar grandes los residuos asociados a dichos valores, facilitando su identificación.

Para el caso lineal, existen varios tipos de estimadores robustos: los estimadores de mínimos cuadrados medianos (LMS) y los de mínimos cuadrados podados (LTS) introducidos por Rousseeuw [22], [23], los estimadores tipo M o M -estimadores (Huber, [13]), los MM -estimadores (Yohai, [27]), los S -estimadores (Rousseeuw y Yohai, [24]) y los τ -estimadores (Yohai y Zamar, [28]). Una extensión de los M -estimadores a modelos de regresión no lineal fue dada por Fraiman [11] mientras que los MM -estimadores fueron considerados en la tesis doctoral de Fasano [10].

Los estimadores robustos tienen dos propiedades importantes a los fines prácticos: el alto **punto de quiebre o ruptura** (Breakdown Point - BP) y la buena **eficiencia** bajo el modelo central.

El **punto de ruptura o BP** de una muestra finita es la mínima fracción de valores atípicos que pueden hacer inútil al estimador. El punto de ruptura del estimador de mínimos cuadrados es $1/n$. Lo deseable para un estimador robusto es de 0,5.

Si aplicamos un estimador robusto a un conjunto de datos que no tiene errores grandes, sin valores atípicos y que provienen de una distribución normal, queremos que los resultados sean idénticos a los obtenidos por el estimador de mínimos cuadrados (LS). La **eficiencia para muestras finitas** de un estimador robusto se define como el cociente entre el cuadrado medio residual de los estimadores LS y el cuadrado medio residual obtenido al calcular el estimador robusto. Lo ideal es que este cociente sea aproximadamente 1.

Dado el modelo de regresión no lineal

$$\mathbf{y}_i = f(\mathbf{x}_i, \boldsymbol{\theta}^*) + \varepsilon_i$$

definimos un M -estimador como el valor

$$\hat{\boldsymbol{\theta}}_n = \operatorname{argmin}_{\boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i=1}^n \rho \left(\frac{y_i - f(\mathbf{x}_i, \boldsymbol{\theta})}{\hat{\sigma}} \right)$$

donde $\hat{\sigma}$ es un estimador robusto de σ y ρ es una función real, llamada función- ρ , que cumple:

- i) $\rho(0) = 0$
- ii) $\rho(-u) = \rho(u)$
- iii) $0 \leq u \leq v \Rightarrow \rho(u) \leq \rho(v)$
- iv) ρ es continua
- v) Si $\rho(u) < \sup_u \rho(u)$ y $0 \leq u < v \Rightarrow \rho(u) < \rho(v)$

Dada una muestra z_1, \dots, z_n , y una función- ρ acotada, definimos el M -estimador de escala $s(\mathbf{z})$ como el valor s , solución de

$$\frac{1}{n} \sum_{i=1}^n \rho\left(\frac{z_i}{s}\right) = b$$

donde b es una constante entre 0 y 1, que en el caso en que $\sup_{t \in \mathbb{R}} \rho(t) = 1$ da el punto de ruptura del estimador de escala.

En un modelo de regresión lineal, el M -estimador puede verse afectado por valores de \mathbf{x} con alto leverage, por lo que su punto de ruptura es $1/n$, al igual que el de mínimos cuadrados. Para solucionar este problema, se desarrollaron diversos estimadores robustos con alto punto de ruptura.

En 1987, Yohai [27] introdujo los MM -estimadores para el caso lineal. Estos combinan la estimación M con la propiedad de alto punto de ruptura. En su tesis, Fasano [10] los extendió al modelo de regresión no lineal.

Para construir un estimador eficiente con alto punto de ruptura tenemos 3 etapas:

1. Tomamos un estimador $\mathbf{T}_{0,n}$ de $\boldsymbol{\theta}_0$ con alto punto de ruptura.
2. Calculamos los residuos $r_i(\mathbf{T}_{0,n}) = y_i - f(\mathbf{x}_i, \mathbf{T}_{0,n})$, $1 \leq i \leq n$, y calculamos el M -estimador de escala s_n usando una ρ_0 acotada y $b = 0,5$.
3. Sea ρ_1 otra función- ρ que cumple que $\rho_1(s) \leq \rho_0(s)$ para todo s . Si $\psi_1 = \rho_1'$ y \dot{f} es el gradiente de f , el MM -estimador $\mathbf{T}_{1,n}$ se define como cualquier solución de

$$\sum_{i=1}^n \psi_1\left(\frac{r_i(\boldsymbol{\theta})}{s_n}\right) \dot{f}(\mathbf{x}_i, \boldsymbol{\theta}) = 0,$$

que cumple $S(\mathbf{T}_{1,n}) \leq S(\mathbf{T}_{0,n})$, donde

$$S(\boldsymbol{\theta}) = \sum_{i=1}^n \rho_1\left(\frac{r_i(\boldsymbol{\theta})}{s_n}\right).$$

El algoritmo antes descrito necesita un estimador inicial de alto punto de ruptura. Un buen candidato es el estimador que minimiza la mediana de los residuos, es decir, el LMS -estimador es el valor tal que

$$\hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \operatorname{mediana}_{1 \leq i \leq n} |y_i - f(\mathbf{x}_i, \boldsymbol{\theta})|.$$

En esta tesis, nos enfocaremos en en M y MM -estimadores.

Cabe mencionar que muchos algoritmos para obtener estimadores con alto punto de ruptura para modelos lineales se basan en hallar estimadores por mínimos cuadrados en muchos conjuntos de p puntos, donde p es la dimensión de \mathbf{X} . Esto no es computacionalmente posible en los problemas no lineales. Por esto, Stromberg [26] diseñó un algoritmo en S-plus donde se requieren pocos conjuntos de p puntos donde realizar ajustes por mínimos cuadrados. En esta tesis se implementó en R dicho algoritmo para modelos lineales generales.

Capítulo 3

Problemas comunes

Al ajustar un modelo no lineal, surgen los siguientes problemas:

1. La convergencia de los métodos iterativos.
2. La limitada aplicación de la teoría de inferencia asintótica.
3. La estimación de parámetros mal condicionada.

Las dificultades que surgen en el último punto son similares al problema de multicolinealidad en la regresión lineal. Estas pueden deberse tanto a la elección de los \mathbf{x}_i en el modelo como a la naturaleza misma del modelo.

3.1. Convergencia de los métodos iterativos

La mayoría de los algoritmos para hallar $\hat{\boldsymbol{\theta}}$ por medios iterativos pueden considerarse adaptaciones del método de Newton: aproximamos

$$h(\boldsymbol{\theta}) = \|\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})\|^2$$

por una función cuadrática $q_h^{(a)}(\boldsymbol{\theta})$, en un entorno de $\boldsymbol{\theta}^{(a)}$ y minimizamos con respecto a $\boldsymbol{\theta}$ para encontrar $\boldsymbol{\theta}^{(a+1)}$.

El método de Newton funciona mejor sobre las funciones más “cercanas” a una cuadrática y converge a $\hat{\boldsymbol{\theta}}$ para un valor inicial $\boldsymbol{\theta}^{(1)}$ suficientemente cercano: ya que cualquier función suave es aproximadamente cuadrática en un pequeño entorno de $\hat{\boldsymbol{\theta}}$, en teoría, el método converge para valores iniciales $\hat{\boldsymbol{\theta}}^{(1)}$ suficientemente cercanos a $\hat{\boldsymbol{\theta}}$. Sin embargo, la función podría tener varios mínimos locales. Por lo que podríamos no llegar a un mínimo global y terminaríamos obteniendo uno local.

3.2. Validez de la inferencia asintótica

Si suponemos válido el modelo no lineal, tenemos que, asintóticamente,

$$\widehat{\boldsymbol{\theta}} \sim N_p(\boldsymbol{\theta}^*, \sigma^2(\mathbf{F}_\bullet^T \mathbf{F}_\bullet)^{-1}).$$

La matriz \mathbf{F}_\bullet juega el mismo rol que \mathbf{X} en la teoría del modelo lineal. En particular, por analogía con las regiones lineales de confianza, tenemos que

$$\{\boldsymbol{\theta} : (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T \widehat{\mathbf{F}}_\bullet^T \widehat{\mathbf{F}}_\bullet (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) \leq p s^2 \mathcal{F}_{p, n-p}(\alpha)\} \quad (3.2.1)$$

es una región de confianza $100(1-\alpha)\%$ aproximada.

Como la aproximación lineal es válida asintóticamente, esta última podría ser la región de confianza asintótica de nivel $1 - \alpha$ correcta. Variando el nivel α las regiones anteriores están encerradas por elipsoides, las cuales también son los contornos de $\widehat{\boldsymbol{\theta}}$ de la función de densidad normal multivariada. Puesto que $S(\boldsymbol{\theta})$ mide la "cercanía" de las observaciones a la ecuación ajustada para cualquier $\boldsymbol{\theta}$, los contornos $S(\boldsymbol{\theta})$ podrían ser apropiadas como base de regiones de confianza para $\boldsymbol{\theta}$. Tales regiones podrían tomar la forma $\{\boldsymbol{\theta} : S(\boldsymbol{\theta}) \leq c S(\widehat{\boldsymbol{\theta}})\}$, para algún $c > 1$.

Las regiones de confianza de este tipo son llamadas Regiones de Confianza "Exactas" ya que no se basan en aproximaciones. Pero el nivel de confianza de tales regiones es generalmente desconocido, por lo cual, niveles aproximados pueden ser obtenidos a través de la teoría asintótica:

Para n suficientemente grande, $\widehat{\boldsymbol{\theta}}$ podría estar suficientemente cercana de $\boldsymbol{\theta}^*$ por lo que vale la aproximación lineal (2.1.5), $S(\boldsymbol{\theta}^*) - S(\widehat{\boldsymbol{\theta}}) \approx (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \mathbf{F}_\bullet^T \mathbf{F}_\bullet (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$, aún reemplazando \mathbf{F}_\bullet por $\widehat{\mathbf{F}}_\bullet$.

$$S(\boldsymbol{\theta}^*) - S(\widehat{\boldsymbol{\theta}}) \approx (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T \widehat{\mathbf{F}}_\bullet^T \widehat{\mathbf{F}}_\bullet (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*).$$

Reemplazando en (3.2.1) tenemos la región de confianza propuesta por Beale [4]

$$\left\{ \boldsymbol{\theta} : S(\boldsymbol{\theta}) \leq S(\widehat{\boldsymbol{\theta}}) \left(1 + \frac{n}{n-p} \mathcal{F}_{p, n-p}(\alpha) \right) \right\} \quad (3.2.2)$$

de nivel de confianza asintótico $1 - \alpha$, ya que asintóticamente es idéntico a (3.2.1).

Una suposición importante para desarrollar inferencia sobre los parámetros a través de intervalos y regiones de confianza es que la aproximación lineal $\mathbf{f}(\widehat{\boldsymbol{\theta}}) + \widehat{\mathbf{F}}_\bullet(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})$ dada a $\mathbf{f}(\boldsymbol{\theta}) = \Omega$ sea adecuada y para esto necesitamos que Ω sea bastante "plano" en un entorno de $\boldsymbol{\theta}$.

La curvatura local de $\mathbb{E}(\mathbf{Y}) = \mathbf{f}(\boldsymbol{\theta})$, puede describirse en términos de 2 componentes ortogonales:

- La componente normal, llamada *Curvatura intrínseca*, que mide cuánto cambia la superficie cerca de $\widehat{\boldsymbol{\theta}}$ cuando cambia $\boldsymbol{\theta}$.
- La componente tangencial, llamada *Curvatura efecto-parámetro*, que refleja cómo cambia una grilla rectangular de valores de $\boldsymbol{\theta}$, centrada en $\widehat{\boldsymbol{\theta}}$, al ser proyectada sobre la superficie esperada $\Omega = \mathbf{f}(\boldsymbol{\theta})$.

Una curvatura intrínseca alta indica (a) un modelo altamente no lineal, (b) que la aproximación lineal por el plano tangente podría no ser adecuada y (c) que inferencias basadas en $\hat{\boldsymbol{\theta}}$ podrían ser engañosas.

La curvatura efecto-parámetro indica que podrían no ser válidos los procedimientos de inferencia y que una transformación paramétrica adecuada podría reducir o eliminar este componente para mejorar el problema.

Los efectos de la curvatura también son necesarios para poder establecer un criterio de diseño del problema para poder minimizar estos efectos.

3.3. Problemas de identificación y de mal condicionamiento

Cuando existen $\boldsymbol{\theta}_1$ y $\boldsymbol{\theta}_2$ tales que $f(\mathbf{x}; \boldsymbol{\theta}_1) = f(\mathbf{x}; \boldsymbol{\theta}_2) \forall \mathbf{x}$, tenemos un problema de identificación de los parámetros para el modelo $\mathbb{E}(Y) = f(\mathbf{x}; \boldsymbol{\theta})$.

Como ejemplo de esto, consideremos el modelo estudiado por Bird Y Milliken [5]:

$$f(\mathbf{x}; \boldsymbol{\theta}) = \exp(-\beta\tau x_1) + \frac{\alpha}{\beta}[1 - \exp(-\beta\tau x_1)]x_2,$$

con $\boldsymbol{\theta} = (\alpha, \beta, \tau)^T$. Vemos que β y τ solo aparecen en la forma $\beta\tau$. Entonces, si consideramos el vector de parámetros $\boldsymbol{\theta}_c = (c\alpha, c\beta, \tau/c)^T$ tenemos que $f(\mathbf{x}; \boldsymbol{\theta}) = f(\mathbf{x}; \boldsymbol{\theta}_c)$ para cualquier valor de c .

Un camino para solucionar este inconveniente es reparametrizar la función usando $\phi_1 = \beta\tau$ y $\phi_2 = \alpha/\beta$.

En un modelo simple, una cuidadosa inspección de los parámetros podría revelar cualquier pérdida de indentificabilidad. Un enfoque más formal, basado en la propiedad de invarianza de las funciones lineales estimables, fue dado por Bird y Milliken [5].

Los problemas asociados a la indentificabilidad se derivan de la estructura del modelo y del método de parametrización, pero no de los puntos x_i del diseño. Algebraicamente, la falta de indentificabilidad esta señalada por $\mathbf{F}^T(\boldsymbol{\theta})\mathbf{F}(\boldsymbol{\theta})$ desde $S(\boldsymbol{\theta})$ cuando es singular, o casi singular.

En el modelo lineal los parámetros no son identificables si la matriz de datos $\mathbf{X} \in \mathbb{R}^{n \times p}$, cuyas filas son las observaciones \mathbf{X}_i^T , tiene rango menor que p , lo que genera que $\mathbf{X}^T\mathbf{X}$ es singular.

Un problema más importante surge cuando las columnas de \mathbf{X} están altamente correlacionadas en lugar de estar relacionadas linealmente. Esto genera que \mathbf{X} tenga rango completo pero que sus columnas esten cerca de ser linealmente dependientes, es decir que, $\mathbf{X}^T\mathbf{X}$ es casi singular. A esto se lo denomina *mal condicionamiento* de la matriz \mathbf{X} .

Los modelos no lineales, heredan estos problemas y se les agrega el problema de los contornos de $S(\boldsymbol{\theta})$, y por lo tanto de los contornos de confianza (3.2.2), que son curvos.

Otro de los problemas es el de los *parámetros estables*. Estos constituyen un vector de parámetros $\boldsymbol{\theta}$ para el cual los elementos de $\hat{\boldsymbol{\theta}}$ están muy cerca de estar correlacionados.

Su nombre se debe a que un cambio en el valor de un parámetro tiene poco efecto sobre la estimación de los otros. Debido a que en los modelos no lineales el mal condicionamiento del prob-

lema puede ser una característica propia del modelo, un buen diseño experimental puede reducir el problema, pero quizás no pueda eliminarlo.

El problema de *parámetros redundantes* es cuando funciones diferentes, dan curvas visualmente indistinguibles.

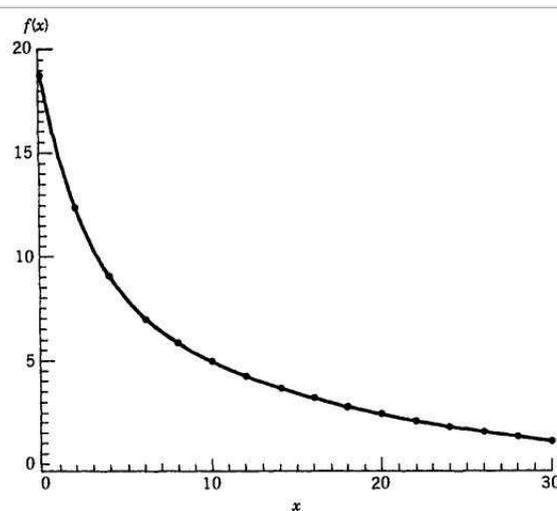
En la figura (Cornish - Bowden [8] - extraído del libro de Seber y Wild [25]), tenemos la superposición de los gráficos de

$$f(x; \boldsymbol{\theta}) = 5e^{-x/1,5} + 5e^{-x/4} + 9e^{-x/14},$$

indicado por los puntos y

$$f(x; \boldsymbol{\theta}) = 8,54e^{-x/2,1} + 10,40e^{-x/12,85},$$

indicado por la línea sólida. A pesar de ser curvas diferentes, visualmente no lo son. Este es un problema habitual cuando tenemos combinaciones lineales de exponenciales.



Reich [19] propuso un método para chequear la redundancia de los parámetros antes de recolectar los datos:

Si el modelo es sensible, queremos que

$$\mathbf{f}(\boldsymbol{\theta} + d\boldsymbol{\theta}) - \mathbf{f}(\boldsymbol{\theta}) \approx \mathbf{F}_{\bullet}(\boldsymbol{\theta})d\boldsymbol{\theta},$$

tenga valor absoluto grande. Por esto, definimos la medida de sensibilidad

$$\mathbf{T}_n = \frac{1}{n} \|\mathbf{f}(\boldsymbol{\theta} + d\boldsymbol{\theta}) - \mathbf{f}(\boldsymbol{\theta})\|^2.$$

Usando la aproximación anterior, definimos $x_{\text{máx}} = \text{máx}_i x_i$ y $\delta_x = x_{\text{máx}}/n$ y tomamos puntos x_i equiespaciados en el intervalo $[0, x_{\text{máx}}]$. Tomando límite cuando $n \rightarrow \infty$, tenemos que la *medida de sensibilidad para un cambio $d\boldsymbol{\theta}$* es

$$T_{\boldsymbol{\theta}} = d\boldsymbol{\theta}^T \mathbf{M}_{\boldsymbol{\theta}} d\boldsymbol{\theta},$$

donde $\mathbf{D} = \text{diag}(\sqrt{m_{11}}, \sqrt{m_{22}}, \dots, \sqrt{m_{pp}})$, y donde

$$\mathbf{M}_\theta = \frac{1}{x_{\text{máx}}} \int_0^{x_{\text{máx}}} \frac{\partial f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial f(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} dx.$$

Reich y Zinke ([20]) y Reich ([19]) propusieron la descomposición de \mathbf{M}_θ en

$$\mathbf{R}_\theta = \mathbf{D}^{-1} \mathbf{M}_\theta \mathbf{D}^{-1},$$

donde $\mathbf{D} = \text{diag}(\sqrt{m_{11}}, \sqrt{m_{22}}, \dots, \sqrt{m_{pp}})$.

Los elementos de la diagonal describen la sensibilidad de los parámetros individuales y \mathbf{R}_θ representa el efecto de la redundancia (Reich [19]).

En resumen, algunos de los problemas descritos son:

- Los diversos tipos de problemas de mal condicionamiento pueden ser inherentes al modelo.
- El mal diseño del experimento puede empeorar los problemas del mal condicionamiento. Aún con un buen diseño y una gran cantidad de datos no tenemos garantía de solución del problema.
- Aún los datos experimentales muy precisos pueden dar lugar a estimaciones imprecisas y altamente correlacionadas, a pesar de tener el modelo correcto.

Aparte de una cuidadosa elección de los puntos del diseño, algunos de los problemas anteriores se pueden superar mediante el uso de transformaciones apropiadas. Ross, [21], enumera cinco tipos de transformaciones, cada una con un propósito diferente. Ellas incluyen transformaciones de los parámetros para lograr correlaciones casi nulas entre las estimaciones de los nuevos parámetros, transformaciones para mejorar los procesos computacionales ya sea inicialmente o aplicados a lo largo de la optimización, transformaciones para “enderezar” valles y curvas y transformaciones para minimizar los efectos de curvatura.

Capítulo 4

Medidas de curvatura y no linealidad

Los métodos de regresión no lineal cuentan con que la función $f(\mathbf{x}, \boldsymbol{\theta})$ puede ser aproximada localmente por una función lineal. Esta aproximación puede ser usada o no por los algoritmos de estimación, pero casi siempre es tomada como válida para poder obtener errores estándar, intervalos de confianza y test t. Por esto necesitamos saber que tan buena es dicha aproximación.

La calidad de la aproximación puede ser resumida por medio de dos medidas referentes al modelo: la *Curvatura Intrínseca* y la *Curvatura efecto-parámetros*. Bates y Watts ([1], [2]), y Goldberg [12], extendieron el trabajo de Beale [4] y, usando conceptos de geometría diferencial, desarrollaron estas medidas de no linealidad basadas en la noción de curvatura.

Estas medidas son independientes de los cambios de escala en los datos y en los parámetros y podemos usarlas para comparar diferentes conjuntos de datos, así como diferentes parametrizaciones del mismo conjunto.

Estas buscan cuantificar qué tan diferentes son en un entorno de $\boldsymbol{\theta}$ la superficie $\Omega = \{\mu : \mu = f(\mathbf{x}, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ y la aproximación lineal.

La validez de dicha aproximación lineal se basa en las suposiciones de linealidad y de coordenadas uniformes:

1. La suposición de linealidad asegura que es posible aproximar la función $f(x, \boldsymbol{\theta})$ en un punto dado usando el plano tangente.
2. La suposición de coordenadas uniformes significa que cualquier descripción paramétrica de $\Omega = \{\mu : \mu = f(\mathbf{x}, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ manda valores equiespaciados de $\boldsymbol{\theta}$ en valores equiespaciados de μ en Ω .

La curvatura intrínseca está relacionada con la suposición de linealidad y depende del conjunto de datos analizado y de la función f pero no de la parametrización.

La curvatura efecto-parámetros está relacionada con la suposición de coordenada uniforme, que depende del modelo y de todos los efectos de la parametrización.

Grandes valores de estas dos medidas indican una pobre aproximación lineal. Éstas buscan

cuantificar cuánto se desvian las suposiciones del modelo no lineal con respecto a las siguientes propiedades del modelo lineal:

- i) $\Omega = \{\boldsymbol{\mu} : \boldsymbol{\mu} = \mathbf{X}^T \boldsymbol{\theta}, \boldsymbol{\theta} \in \mathbb{R}^p\}$ es un subespacio lineal del espacio de la muestra.
- ii) cualquier descripción paramétrica de Ω manda valores equiespaciados de $\boldsymbol{\theta}$ en valores de $\boldsymbol{\mu}$ equiespaciados en Ω .

Al considerar la aproximación de Taylor de segundo orden, tenemos que

$$\mathbf{f}(\boldsymbol{\theta}) - \mathbf{f}(\widehat{\boldsymbol{\theta}}) \approx \widehat{\mathbf{F}}_{\bullet}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^T \widehat{\mathbf{F}}_{\bullet\bullet}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) = \widehat{\mathbf{F}}_{\bullet} \boldsymbol{\delta} + \frac{1}{2} \boldsymbol{\delta}^T \widehat{\mathbf{F}}_{\bullet\bullet} \boldsymbol{\delta}$$

Ignorando el término cuadrático, la aproximación lineal para $\boldsymbol{\theta}$ en un entorno de $\widehat{\boldsymbol{\theta}}$ queda

$$\mathbf{f}(\boldsymbol{\theta}) - \mathbf{f}(\widehat{\boldsymbol{\theta}}) \approx \widehat{\mathbf{F}}_{\bullet}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}).$$

Con esto se ve que las columnas de $\widehat{\mathbf{F}}_{\bullet}$ generan el plano tangente a la superficie esperada, Ω , en $\widehat{\boldsymbol{\theta}}$.

Usando esta aproximación, se puede ver que la región de confianza aproximada con un nivel $100(1 - \alpha)\%$ para $\boldsymbol{\theta}$ dada por $\{\boldsymbol{\theta} : (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \widehat{\mathbf{F}}_{\bullet}^T \widehat{\mathbf{F}}_{\bullet} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leq p s^2 \mathcal{F}_{p, n-p, \alpha}\}$ es el conjunto de los $\boldsymbol{\theta}$ en el plano tangente tales que

$$\|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}\|^2 \approx \left\| \widehat{\mathbf{F}}_{\bullet}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) \right\|^2 \leq p s^2 \mathcal{F}_{p, n-p, \alpha},$$

donde $\boldsymbol{\mu} = \mathbf{f}(\boldsymbol{\theta})$. Por lo tanto, se deduce que $\boldsymbol{\mu}$ vive en una esfera con centro $\widehat{\boldsymbol{\mu}}$ y radio $(\rho^2 \mathcal{F}_{p, n-p, \alpha})^{1/2}$, donde $\rho = s\sqrt{p}$ se define como *radio estándar*.

La validez de dicha aproximación depende de la magnitud del término cuadrático respecto del término lineal.

Para hacer esta comparación, dividimos el término cuadrático (un vector de n -coordenadas) en dos componentes ortogonales: la proyección ortogonal sobre el plano tangente y la normal al plano tangente.

$$\widehat{\mathbf{F}}_{\bullet\bullet} = \widehat{\mathbf{F}}_{\bullet\bullet}^{(T)} + \widehat{\mathbf{F}}_{\bullet\bullet}^{(N)} \quad (4.0.1)$$

Así tenemos dos medidas de curvatura dadas por Bates y Wates ([1]), la *Curvatura efectos-parámetros en la dirección \mathbf{h}* , $K_{\mathbf{h}}^{(T)}$, y la *Curvatura intrínseca en la dirección \mathbf{h}* , $K_{\mathbf{h}}^{(N)}$ dadas por

$$K_{\mathbf{h}}^{(T)} = \frac{\left\| \mathbf{h}^T \widehat{\mathbf{F}}_{\bullet\bullet}^{(T)} \mathbf{h} \right\|}{\left\| \widehat{\mathbf{F}}_{\bullet} \mathbf{h} \right\|^2}, \quad K_{\mathbf{h}}^{(N)} = \frac{\left\| \mathbf{h}^T \widehat{\mathbf{F}}_{\bullet\bullet}^{(N)} \mathbf{h} \right\|}{\left\| \widehat{\mathbf{F}}_{\bullet} \mathbf{h} \right\|^2}. \quad (4.0.2)$$

4.1. Interpretación geométrica

Tomenos una recta arbitraria a través de $\widehat{\boldsymbol{\theta}}$ con dirección \mathbf{h} en Ω , $\boldsymbol{\theta}(b) = \widehat{\boldsymbol{\theta}} + b\mathbf{h}$ con b arbitrario. Esto genera en la superficie esperada Ω la curva $\boldsymbol{\mu}_{\mathbf{h}}(\boldsymbol{\theta}) = \boldsymbol{\mu}(\widehat{\boldsymbol{\theta}} + b\mathbf{h})$.

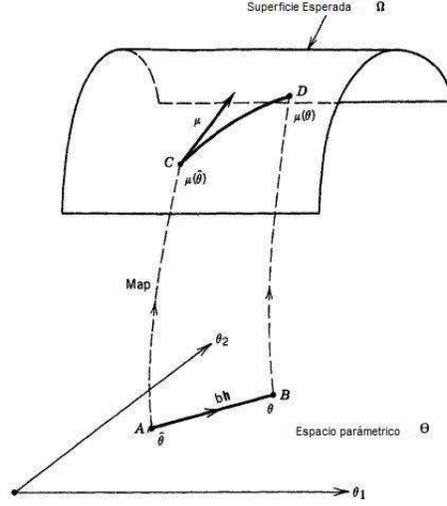


Figura 4.1: Gráfico extraído de Seber y Wild [25] - pag 132

El vector tangente en $\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathbf{f}(\boldsymbol{\theta})$ cuando $b = 0$ es

$$\hat{\boldsymbol{\mu}}_{\mathbf{h}} = \sum_{r=1}^p \hat{f}_r h_r = \hat{\mathbf{F}}_{\bullet} \mathbf{h}.$$

El vector aceleración $\ddot{\boldsymbol{\mu}}_{\mathbf{h}}$ en $b = 0$ está dado por

$$\ddot{\boldsymbol{\mu}}_{\mathbf{h}} = \sum_r \sum_s \hat{f}_{rs} h_r h_s = \mathbf{h}^T \hat{\mathbf{F}}_{\bullet\bullet} \mathbf{h}.$$

Si descomponemos $\hat{\mathbf{F}}_{\bullet\bullet}$ como en (4.0.1) tenemos las medidas de curvatura definidas en (4.0.2).

$$K_{\mathbf{h}}^{(T)} = \frac{\|\ddot{\boldsymbol{\mu}}_{\mathbf{h}}^T\|}{\|\dot{\boldsymbol{\mu}}_{\mathbf{h}}\|^2} = \frac{\|\mathbf{h}^T \hat{\mathbf{F}}_{\bullet\bullet}^{(T)} \mathbf{h}\|}{\|\hat{\mathbf{F}}_{\bullet} \mathbf{h}\|^2}, \quad K_{\mathbf{h}}^{(N)} = \frac{\|\ddot{\boldsymbol{\mu}}_{\mathbf{h}}^N\|}{\|\dot{\boldsymbol{\mu}}_{\mathbf{h}}\|^2} = \frac{\|\mathbf{h}^T \hat{\mathbf{F}}_{\bullet\bullet}^{(N)} \mathbf{h}\|}{\|\hat{\mathbf{F}}_{\bullet} \mathbf{h}\|^2}.$$

Para deshacerse del problema de las escalas, Bates y Watts sugieren dividir por ρ a y , $\boldsymbol{\mu}$, $\hat{\boldsymbol{\mu}}$, $\hat{\mathbf{F}}_{\bullet}$ y a $\hat{\mathbf{F}}_{\bullet\bullet}$. Entonces las curvaturas del problema y la inversa del radio (la curvatura de la esfera) pasan a ser

$$\gamma_{\mathbf{h}}^T = K_{\mathbf{h}}^{(T)} \rho, \quad \gamma_{\mathbf{h}}^N = K_{\mathbf{h}}^{(N)} \rho, \quad y \quad 1/\sqrt{\mathcal{F}_{p,n-p,\alpha}}.$$

Esto nos muestra que la curvatura intrínseca, $\gamma_{\mathbf{h}}^N$, es el recíproco del radio del círculo que mejor aproxima a la superficie esperada Ω en la dirección $\boldsymbol{\mu}_{\mathbf{h}}$. Por lo tanto, si la suposición de linealidad es buena, la curvatura intrínseca será pequeña.

Capítulo 5

Inferencia Estadística

Tomando a $\boldsymbol{\theta}$ como el verdadero valor del parámetro, en este capítulo consideramos el modelo

$$y_i = f(\mathbf{x}_i; \boldsymbol{\theta}) + \epsilon_i, \quad (5.0.1)$$

donde ϵ_i son i.i.d con distribución $N(0, \sigma^2)$, $i = 1, \dots, n$.

5.1. Intervalos de Confianza asintóticos

Bajo condiciones apropiadas de regularidad, tenemos que, asintóticamente

$$\widehat{\boldsymbol{\theta}} \sim N_p(\boldsymbol{\theta}, \sigma^2 \mathbf{C}^{-1}), \quad \mathbf{C} = \mathbf{F}_\bullet^T \mathbf{F}_\bullet. \quad (5.1.1)$$

Entonces, para una combinación lineal $\mathbf{a}^T \boldsymbol{\theta}$, podemos aplicar el teorema (2.1.1) y, en forma asintótica, tenemos que

$$\mathbf{a}^T \widehat{\boldsymbol{\theta}} \sim N_p(\mathbf{a}^T \boldsymbol{\theta}, \sigma^2 \mathbf{a}^T \mathbf{C}^{-1} \mathbf{a}),$$

que es independiente de $S^2 = \left\| \mathbf{y} - \mathbf{f}(\widehat{\boldsymbol{\theta}}) \right\| / (n - p)$, el estimador insesgado de σ^2 . Para n grande tenemos que, aproximadamente,

$$T = \frac{\mathbf{a}^T \widehat{\boldsymbol{\theta}} - \mathbf{a}^T \boldsymbol{\theta}}{s(\mathbf{a}^T \mathbf{C}^{-1} \mathbf{a})^{1/2}} \sim \mathcal{T}_{n-p}. \quad (5.1.2)$$

Un intervalo de confianza $100(1 - \alpha)\%$, aproximado, para $\mathbf{a}^T \boldsymbol{\theta}$ es $\mathbf{a}^T \widehat{\boldsymbol{\theta}} \pm t_{n-p, \alpha} s(\mathbf{a}^T \widehat{\mathbf{C}}^{-1} \mathbf{a})^{1/2}$, donde hemos aproximado a \mathbf{C} por $\widehat{\mathbf{C}} = \widehat{\mathbf{F}}_\bullet^T \widehat{\mathbf{F}}_\bullet$.

Tomando $\mathbf{a}^T = (0, 0, \dots, 0, 1, 0, \dots, 0)$ y definiendo $(\widehat{\mathbf{C}})^{(rs)}$ como el elemento (r, s) de $\widehat{\mathbf{C}}^{-1}$, el intervalo de confianza para θ_r está dado por

$$\widehat{\theta}_r \pm t_{n-p, \alpha} s \sqrt{\widehat{\mathbf{C}}^{(rr)}}. \quad (5.1.3)$$

La experiencia sugiere que la teoría asintótica da resultados que parecen razonables en las aplicaciones. Pero la utilidad de este último depende de las propiedades de curvatura de $\mathbb{E}(\mathbf{Y}) = \mathbf{f}(\boldsymbol{\theta})$. Por esto, necesitamos un chequeo adecuado para que (5.1.3) no resulte engañoso.

Para obtener un conjunto de intervalos (uno por cada θ_r), se puede usar el *método de Bonferroni*. Así, para cada $r = 1, \dots, p$, tenemos $\hat{\theta}_r \pm t_{n-p, \alpha/2p} s \sqrt{\widehat{\mathbf{C}}^{(rr)}}$. Estos intervalos tienen un nivel de confianza menor que $100(1 - \alpha) \%$ y pueden estar afectados por la curvatura.

Usando la linealización asintótica de (5.0.1) se pueden aplicar métodos lineales para encontrar un intervalo de predicción para y en $\mathbf{x} = \mathbf{x}_0$.

Si consideramos

$$\mathbf{f}_0^T = \left(\frac{\partial f(\mathbf{x}_0, \boldsymbol{\theta})}{\partial \theta_1}, \frac{\partial f(\mathbf{x}_0, \boldsymbol{\theta})}{\partial \theta_2}, \dots, \frac{\partial f(\mathbf{x}_0, \boldsymbol{\theta})}{\partial \theta_p} \right),$$

y tomamos el desarrollo usual de Taylor de primer orden $f(\mathbf{x}_0, \hat{\boldsymbol{\theta}}) \approx f(\mathbf{x}_0, \boldsymbol{\theta}) + \mathbf{f}_0^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$, para $\hat{y}_0 = f(\mathbf{x}_0, \hat{\boldsymbol{\theta}})$, tenemos que $y_0 - \hat{y}_0 \approx y_0 - f(\mathbf{x}_0, \boldsymbol{\theta}) - \mathbf{f}_0^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \epsilon_0 - \mathbf{f}_0^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$.

Teniendo en cuenta la distribución (5.1.1) de $\hat{\boldsymbol{\theta}}$ y la independencia de $\hat{\boldsymbol{\theta}}$ y ϵ_0 , tenemos que, en forma asintótica $y_0 - \hat{y}_0 \sim N(0, \sigma^2(1 + v_0))$, donde $v_0 = \mathbf{f}_0^T (\mathbf{F}_\bullet^T \mathbf{F}_\bullet)^{-1} \mathbf{f}_0$.

Ahora s^2 es independiente de y_0 y es asintóticamente independiente de $\hat{\boldsymbol{\theta}}$, de modo que s^2 es asintóticamente independiente de $y_0 - \hat{y}_0$. Por lo tanto, asintóticamente,

$$\frac{y_0 - \hat{y}_0}{s \sqrt{(1 + v_0)}} \sim \mathcal{T}_{n-p}.$$

Esto nos permite obtener un intervalo de predicción aproximado de nivel $100(1 - \alpha) \%$ para y_0 de la forma

$$\hat{y}_0 \pm t_{n-p, \frac{\alpha}{2}} s [1 + v_0]^{1/2},$$

donde v_0 puede ser estimado reemplazando $\boldsymbol{\theta}$ por $\hat{\boldsymbol{\theta}}$.

5.2. Intervalos simultáneos y regiones de confianza

Sea $\mathbf{C} = \mathbf{F}_\bullet^T \mathbf{F}_\bullet = \mathbf{F}_\bullet^T(\boldsymbol{\theta}^*) \mathbf{F}_\bullet(\boldsymbol{\theta}^*)$ la matriz de covarianza asintótica del estimador de mínimos cuadrados. Llamemos

$$\widehat{\mathbf{C}} = \widehat{\mathbf{F}}_\bullet^T \widehat{\mathbf{F}}_\bullet = \mathbf{F}_\bullet^T(\hat{\boldsymbol{\theta}}) \mathbf{F}_\bullet(\hat{\boldsymbol{\theta}}),$$

al estimador de \mathbf{C} . Del teorema (2.1.1) podemos deducir la región de confianza para $\boldsymbol{\theta}$

$$\left\{ \boldsymbol{\theta} : (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \widehat{\mathbf{C}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq p s^2 \mathcal{F}_{p, n-p}(\alpha) \right\}, \quad (5.2.1)$$

algunas veces llamada *Región linealizada*

Siguiendo la teoría lineal usual, para el *método de Scheffé* tenemos

$$\begin{aligned} 1 - \alpha &= \mathbb{P}(\mathcal{F}_{p, n-p} \leq \mathcal{F}_{p, n-p}(\alpha)) \\ &\approx \mathbb{P}\left(\boldsymbol{\theta} : (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \widehat{\mathbf{C}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq p s^2 \mathcal{F}_{p, n-p}(\alpha)\right) \\ &= \mathbb{P}\left(\sup_{\mathbf{b}} \frac{\{\mathbf{b}^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\}^2}{\mathbf{b}^T \widehat{\mathbf{C}}^{-1} \mathbf{b}} \leq c\right) = \mathbb{P}\left(\left|\mathbf{b}^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\right| \leq (c \mathbf{b}^T \widehat{\mathbf{C}} \mathbf{b})^{1/2} \quad \forall \mathbf{b} \neq \mathbf{0}\right), \end{aligned}$$

donde $c = p s^2 \mathcal{F}_{p,n-p}(\alpha)$. Entonces, con probabilidad aproximada $1 - \alpha$, tenemos los intervalos simultáneos

$$\mathbf{b}^T \hat{\boldsymbol{\theta}} \pm (p \mathcal{F}_{p,n-p}(\alpha))^{1/2} s (\mathbf{b}^T \hat{\mathbf{C}}^{-1} \mathbf{b})^{1/2}. \quad (5.2.2)$$

Estos incluyen los intervalos para θ_r , siendo más largos que los de Bonferroni. Los intervalos (5.2.2) son usados para mirar todas las combinaciones sugeridas por los datos, mientras que los generados por el método de Bonferroni son útiles para combinaciones pre elegidas.

Recordando que $S(\boldsymbol{\theta}) = \|\mathbf{r}(\boldsymbol{\theta})\|^2 = \|\mathbf{y} - \mathbf{f}(\boldsymbol{\theta})\|^2$, de (3.2.2) deducimos que

$$\left\{ \boldsymbol{\theta} : \frac{S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}})}{S(\hat{\boldsymbol{\theta}})} \leq \frac{n}{n-p} \mathcal{F}_{p,n-p}(\alpha) \right\}. \quad (5.2.3)$$

Si tomamos $\hat{\mathbf{r}} = \mathbf{r}(\hat{\boldsymbol{\theta}})$, nos queda

$$\left\{ \boldsymbol{\theta} : \|\mathbf{r}(\boldsymbol{\theta})\|^2 \leq (1 + f) \|\hat{\mathbf{r}}\|^2 \right\}, \quad \text{con} \quad f = \frac{p}{n-p} \mathcal{F}_{p,n-p}(\alpha). \quad (5.2.4)$$

Aunque las regiones (5.2.1) y (5.2.3) son asintóticamente la misma para los modelos no lineales y exactamente de la misma para los modelos lineales, ellas pueden ser muy diferentes para muestras pequeñas. En general, (5.2.1) parece ser superior, pues está más cerca del valor nominal y se ve menos afectada por la curvatura (Donaldson y Schnabel, [9]).

La función $S(\boldsymbol{\theta})$ es invariante por reparametrizaciones, por lo cual podemos asumir que encontramos una transformación para eliminar la curvatura efecto-parámetros. Por esto (5.2.4) sólo se ve afectada por la curvatura intrínseca, que a menudo es insignificante. Box y Coutie [6] propusieron una versión aproximada de (5.2.3)

$$\left\{ \boldsymbol{\theta} : (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \hat{\mathbf{H}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq p s^2 \mathcal{F}_{p,n-p}(\alpha) \right\}, \quad (5.2.5)$$

donde $\hat{\mathbf{H}} = \frac{\partial^2 S}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \hat{\mathbf{F}}_•^T \hat{\mathbf{F}}_• - [\mathbf{r}(\hat{\boldsymbol{\theta}})^T][\hat{\mathbf{F}}_•]$.

Se espera que (5.2.5) aproxime mejor a (5.2.3) que (5.2.1) ya que este incluye términos de segundo orden del desarrollo de Taylor. Sin embargo, regiones del tipo (5.2.3) son difíciles de computar y de representar gráficamente. Por este motivo, se está más interesado en intervalos de confianza para cada θ_r , $r = 1, \dots, p$.

La validez de la teoría asintótica depende mucho del grado de no linealidad del modelo y de la elección de los parámetros $\boldsymbol{\theta}$, reflejadas en las curvaturas intrínseca y de efectos-parámetros. Si la aproximación lineal es buena, podemos aplicar la teoría lineal al problema tomando $\mathbf{X} = \hat{\mathbf{F}}_•$. Cuando no podamos ignorar los efectos de la curvatura, podremos realizar algunas modificaciones a la teoría lineal.

Capítulo 6

Software y Ejemplos

Los programas de software estadístico más importantes incluyen funciones para ajustar modelos de regresión no lineal. Para los casos de estudio, utilizamos el programa *R* que cuenta con 2 funciones para realizar estimación no lineal. Una es la función *nls()* que busca estimadores por mínimos cuadrados no lineales y que por defecto utiliza el algoritmo de Gauss-Newton. La otra función es *nlrob()* que realiza un ajuste robusto usando *M*-estimadores por mínimos cuadrados reponderados iterativos (IWLS). También implementamos la rutina de Stromberg [26] *mmnl()* que fue desarrollada originalmente para el software comercial *S-Plus* y que figura en el Capítulo 7. Esta última estima los parámetros de un modelo de regresión no lineal utilizando *MM*-estimadores. El *MM*-estimador se calcula utilizando como función ρ la función de Hampel, ver Huber [14]. En este capítulo, analizamos algunos conjuntos de datos existentes en la literatura a los que se agregaron datos atípicos para evaluar el comportamiento de los estimadores robustos.

6.1. Ejemplo de datos de cebolla

Estos datos son un ejemplo de curvas de crecimiento sigmoideal, tratadas por Ratkowsky [17] y se conocen usualmente en la literatura como **Onion Data**. Son 16 pares de datos donde la variable de respuesta (y) es el peso seco de los bulbos de cebolla y la variable predictora (x) es el tiempo de crecimiento. Los datos se grafican en la Figura 6.1.

Varios modelos de crecimiento sigmoideal fueron propuestos para este tipo de datos, entre ellos el modelo de Gompertz, basado en la distribución Weibull:

$$v = f(x, \boldsymbol{\theta}) = \theta_1 \exp(-\exp(\theta_2 - \theta_3 x))$$

Como estimación inicial de θ_1 , tomamos el último valor de Y , redondeado al entero siguiente (consideramos el valor asintótico de X), es decir, tomamos $\theta_1 = 718$. Para obtener los valores iniciales de θ_2 y θ_3 , linealizamos la ecuación

$$z_0 = \ln(-\ln\left(\frac{v}{\theta_1}\right)) = \theta_2 - \theta_3 x$$

y aplicamos mínimos cuadrados lineales. Con esto obtenemos los valores iniciales

$$\theta_1 = 718, \theta_2 = 2.3794, \theta_3 = 0.4673.$$

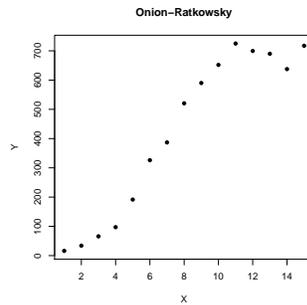


Figura 6.1: datos de Cebolla: Ratkowsky (1983)

Aplicamos mínimos cuadrados no lineales con el metodo de Gauss-Newton usando la función *nls()* del R. El gráfico de residuos se muestra en la Figura 6.1. Calculamos también los *M*-estimadores usamos la función *nlob()*. Los resultados obtenidos con ambos métodos están dados en la Tabla 6.1.

Parámetro	$\hat{\theta}_{OLS}$	$\hat{\theta}_M$
θ_1	723.10870	731.82003
θ_2	2.50018	2.43288
θ_3	0.45010	0.43653

Tabla 6.1: Estimaciones de los parámetros para el conjunto de datos **Onion**.

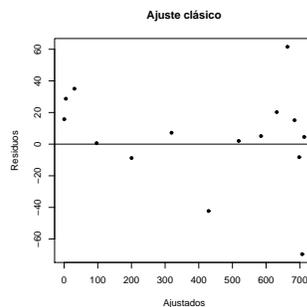


Figura 6.2: Gráfico de residuos

Tenemos que tener en cuenta que los valores del error estándar dados en las Tablas 6.2 y 6.3 para los métodos robustos son asintóticos.

$\hat{\theta}_{OLS}$				
Parámetro	Estimador	Error Estándar	Valor t	Pr(> t)
θ_1	723.10870	22.06049	32.778	4.12e-13
θ_2	2.50018	0.28470	8.782	1.43e-06
θ_3	0.45010	0.05178	8.693	1.59e-06
$\hat{\theta}_M$				
Parámetro	Estimador	Error Estándar	Valor t	Pr(> t)
θ_1	731.82003	21.05600	34.76	2.05e-13
θ_2	2.43288	0.23589	10.31	2.56e-07
θ_3	0.43653	0.04374	9.98	3.66e-07

Tabla 6.2: Información del estimador de mínimos cuadrados y del M -estimador para el conjunto de datos **Onion**.

Estimador	Error Estándar Residual	Cantidad de iteraciones
$\hat{\theta}_{OLS}$	33.67	7
$\hat{\theta}_M$	20.47	11

Tabla 6.3: Errores estándar residuales y número de iteraciones hasta convergencia del estimador de mínimos cuadrados y del M -estimador para el conjunto de datos **Onion**.

Con el método robusto se logró un Error Residual Estándar menor que con el método clásico, pero en una mayor cantidad de iteraciones.

Como puede verse en la Figura 6.3, los valores atípicos no son tan notorios, lo que se traduce en una poca influencia en el método clásico.

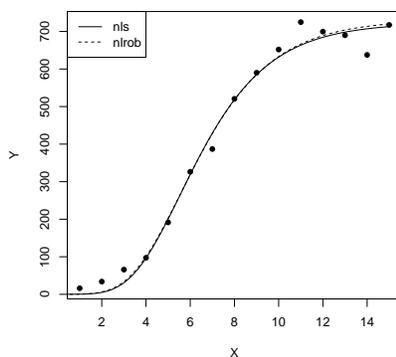


Figura 6.3: Curvas de ajuste

Debido a la poca cantidad de datos no pudimos utilizar la función $mmnl()$.

6.2. Ejemplo de datos de ciervos

Estos pares de datos representan la edad y la longitud de la mandíbula de cierto tipo de ciervos. Fueron objeto de estudio por Crawley [7] y figuran con el nombre **Deer Data** en R. Se muestran en la Figura 6.4.

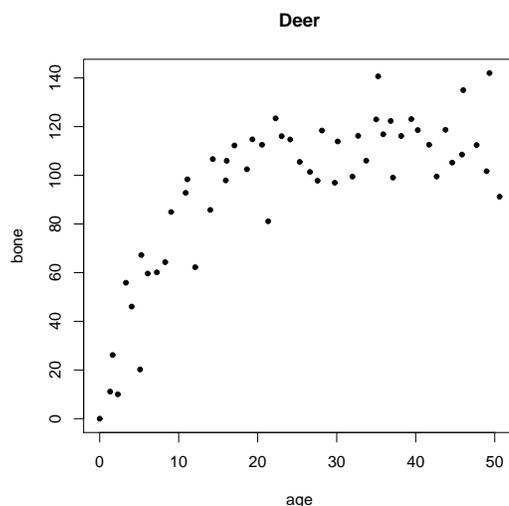


Figura 6.4: Datos de Ciervos : Crawley (2007)

La teoría indica que la relación entre la Edad (x) y la Longitud de la mandíbula (y) es una función exponencial asintótica con 2 parámetros:

$$v = f(x, \boldsymbol{\theta}) = \theta_1(1 - \exp(-\theta_2 x)).$$

Con un análisis previo, Crawley [7] decidió tomar como valores iniciales $\theta_1 = 120$, $\theta_2 = 0.064$. los resultados obtenidos, al realizar el ajuste de los datos por Mínimos Cuadrados No Lineales y por los métodos robustos con el M -estimador y con MM -estimador, se dan en la tabla 6.4.

Parámetro	$\hat{\boldsymbol{\theta}}_{OLS}$	$\hat{\boldsymbol{\theta}}_M$	$\hat{\boldsymbol{\theta}}_{MM}$
θ_1	115.581	114.621	115.491
θ_2	0.11882	0.12647	0.11970

Tabla 6.4: Estimaciones de los parámetros para el conjunto de datos **Deer**.

Los errores estándar de los estimadores OLS y M -estimadores así como los respectivos valores del estadístico t para la hipótesis $H_0 : \theta_j = 0$ se dan en la Tabla 6.5. La función $mmnl()$ no da información con respecto al error estándar ni al t -valor del MM -estimador. Como antes la Tabla 6.6 da los valores del error estándar residual y la cantidad de iteraciones necesarias hasta obtener convergencia. Los resultados para ambos métodos son comparables. La Figura 6.5 muestra los ajustes obtenidos, no observándose diferencias entre los ajustes provistos por el estimador de mínimos cuadrados y el MM -estimador.

$\hat{\theta}_{OLS}$			
Parámetro	Estimador	Error Estándar	Estadístico t
θ_1	115.581	2.84365	40.645
θ_2	0.11882	0.01233	9.635
$\hat{\theta}_M$			
Parámetro	Estimador	Error Estándar	Estadístico t
θ_1	114.621	3.19597	35.864
θ_2	0.12647	0.01552	8.147

Tabla 6.5: Errores estándar de los estimadores de los parámetros para el conjunto de datos contaminados y valores del estadístico t para la hipótesis $H_0 : \theta_j = 0$.

Estimador	Error Estándar Residual	Cantidad de iteraciones
$\hat{\theta}_{OLS}$	13.10	5
$\hat{\theta}_M$	13.19	7

Tabla 6.6: Errores estándar residuales y número de iteraciones hasta convergencia del estimador de mínimos cuadrados y del M -estimador para el conjunto de datos sobre medidas de Ciervo.

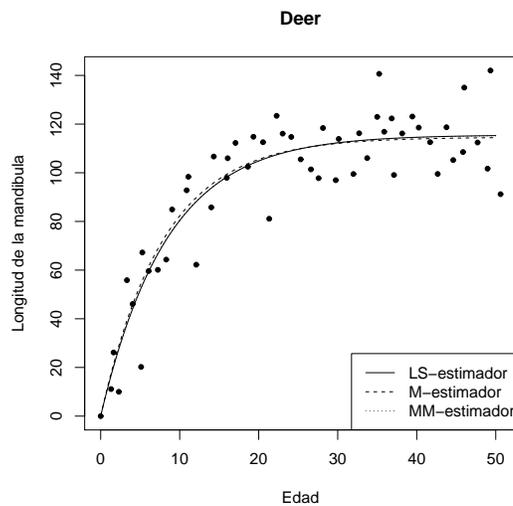


Figura 6.5: Gráfico de valores ajustados $v = \theta_1 (1 - \exp(-\theta_2 x))$ por los 3 métodos de estimación.

Modificamos 4 respuestas en el conjunto de datos para observar el comportamiento de los tres estimadores en presencia de datos atípicos. Los datos modificados se dan en la Tabla 6.7 y se grafican en la Figura 6.6.

Dato	original	modificado
47	112.43	302
48	101.68	310
49	142	320
50	91.20	340

Tabla 6.7: datos modificados en el conjunto de datos de Ciervo.

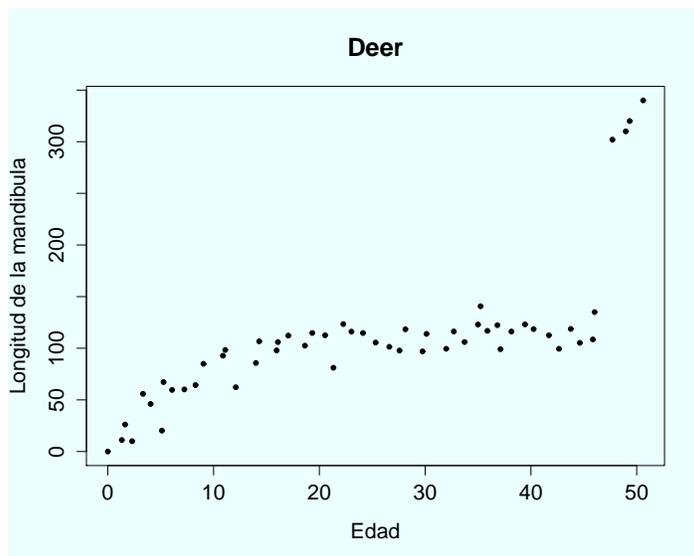


Figura 6.6: Grafico de los datos de Ciervo con datos contaminados

Al aplicar los distintos métodos de estimación al nuevo conjunto de datos obtuvimos los resultados dados en la Tabla 6.8.

Parámetro	$\hat{\theta}_{OLS}$	$\hat{\theta}_M$	$\hat{\theta}_{MM}$
θ_1	676.3	119.8966	116.2906
θ_2	0.00689	0.11169	0.11703

Tabla 6.8: Estimaciones de los parámetros para el conjunto de datos contaminados.

Se puede ver que la contaminación provocó la ruptura del estimador por mínimos cuadrados. El M -estimador no sufrió tanto cambio y el MM -estimador casi no muestra variación. A partir de los resultados reportados en la Tabla 6.9, podemos ver que, con el método por mínimos cuadrados, ninguno de los dos parámetros resultan significativamente distintos de cero, a diferencia de los M -estimadores que conservan la significación del conjunto de datos original.

También se produjeron cambios en el Error Residual Estándar que se reportan en la Tabla 6.10. Para el M -estimador es el Error Residual Estándar Robusto asintótico.

$\hat{\theta}_{OLS}$			
Parámetro	Estimador	Error Estándar	Estadístico t
θ_1	676.3	1.033e+03	0.655
θ_2	0.00689	1.200e-02	0.574
$\hat{\theta}_M$			
Parámetro	Estimador	Error Estándar	Estadístico t
θ_1	119.8966	3.89383	30.795
θ_2	0.11169	0.01417	7.884

Tabla 6.9: Errores estándar de los estimadores de los parámetros para el conjunto de datos contaminados y valores del estadístico t para la hipótesis $H_0 : \theta_j = 0$.

Estimador	Error Estándar Residual	Cantidad de iteraciones
$\hat{\theta}_{OLS}$	48.18	17
$\hat{\theta}_M$	12.93	12

Tabla 6.10: Errores estándar residuales y número de iteraciones hasta convergencia del estimador de mínimos cuadrados y del M -estimador para el conjunto de datos contaminado.

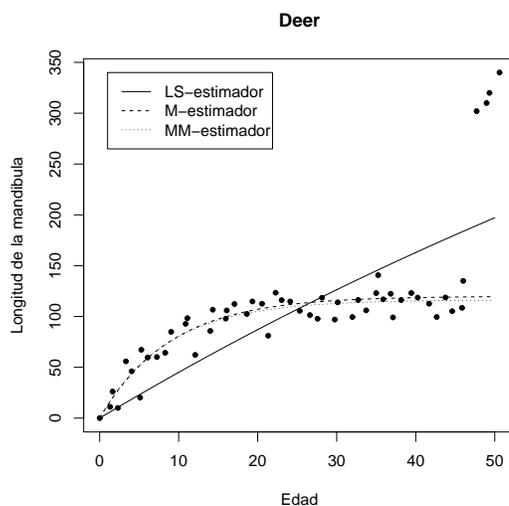


Figura 6.7: Curvas de ajuste para los datos de Ciervo al modificar 4 observaciones.

En la Figura 6.7 se ve como los datos atípicos influyen en el ajuste por mínimos cuadrados y no en los ajustes robustos. La figura 6.8 permite detectar claramente los 4 datos modificados cuando se utilizan procedimientos robustos, mientras que si se utilizase sólo el procedimiento clásico se pensaría que el modelo no es el adecuado.

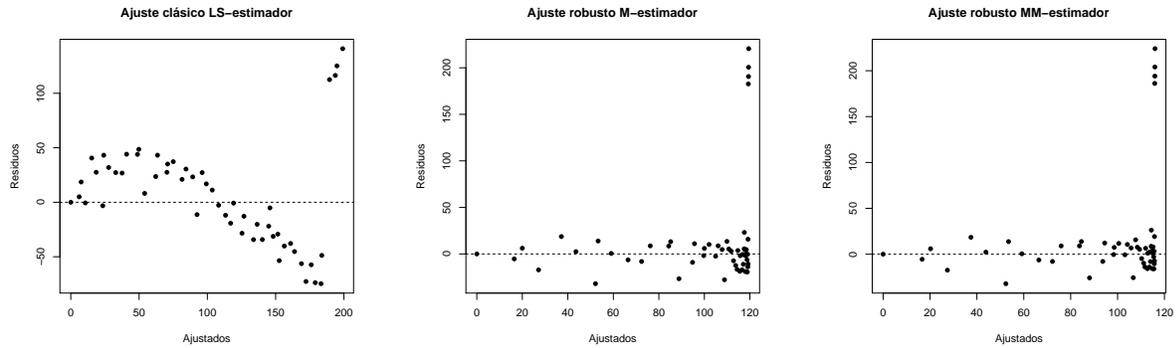


Figura 6.8: Gráficos de valores ajustados versus residuos según el método de ajuste

6.3. Ejemplo de datos de ultrasonido

Este conjunto de datos, usualmente conocido como **Chwirut2**, contiene las mediciones provenientes de un experimento que estudia como la respuesta al ultrasonido depende de la distancia al metal. Los datos son de libre acceso y pertenecen al *National Institute of Standards and Technology* (NIST). Se grafican en la Figura 6.9.

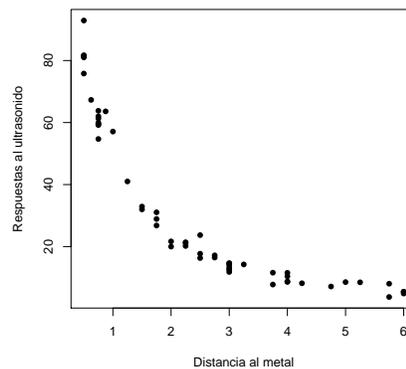


Figura 6.9: datos de ultrasonido (**Chwirut2**): NIST

El modelo teórico propuesto por el NIST es

$$v = f(x, \boldsymbol{\theta}) = \frac{\exp(-\theta_1 x)}{\theta_2 + \theta_3 x}.$$

Los valores de los estimadores clásicos y robustos se dan en la Tabla 6.11. Por otra parte, en la Tabla 6.12, se presentan los errores estándar de los estimadores *OLS* y *M*-estimadores así como los respectivos valores del estadístico *t* para la hipótesis $H_0 : \theta_j = 0$. Como antes la Tabla 6.13 da los valores del error estándar residual y la cantidad de iteraciones necesarias hasta obtener convergencia. Los resultados para ambos métodos son comparables. La Figura 6.10 muestra los ajustes obtenidos, no observándose diferencias entre los ajustes provistos por los distintos métodos.

Parámetro	$\hat{\theta}_{OLS}$	$\hat{\theta}_M$	$\hat{\theta}_{MM}$
θ_1	0.16658	0.15120	0.15632
θ_2	0.005167	0.00495	0.00512
θ_3	0.01215	0.01296	0.01267

Tabla 6.11: Estimadores de los parámetros.

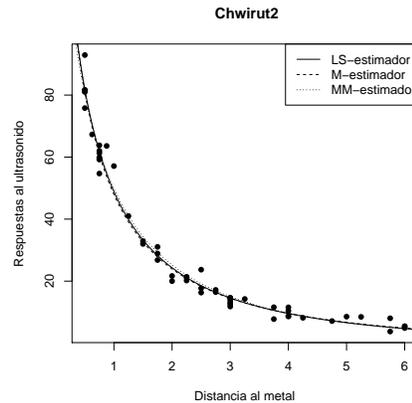


Figura 6.10: Curvas de ajuste

$\hat{\theta}_{OLS}$			
Parámetro	Estimador	Error Estándar	Estadístico t
θ_1	0.16658	0.0383033	4.349
θ_2	0.00517	0.0006662	7.753
θ_3	0.01215	0.0015304	7.939
$\hat{\theta}_M$			
Parámetro	Estimador	Error Estándar	Estadístico t
θ_1	0.15120	0.0251953	6.001
θ_2	0.00495	0.0004854	10.197
θ_3	0.01296	0.0010919	11.872

Tabla 6.12: Errores estándar de los estimadores de los parámetros para el conjunto de ultrasonido y valores del estadístico t para la hipótesis $H_0 : \theta_j = 0$.

Estimador	Error Estándar Residual	Cantidad de iteraciones
$\hat{\theta}_{OLS}$	3.172	4
$\hat{\theta}_M$	1.575	8

Tabla 6.13: Errores estándar residuales y número de iteraciones hasta convergencia del estimador de mínimos cuadrados y del M -estimador para el conjunto de datos sobre ultrasonido.

Modificamos los datos para tener 4 datos atípicos y ver cómo se comportan los procedimientos

de estimación clásicos y robustos del modelo no lineal. Los datos modificados se dan en la Tabla 6.14 y el nuevo conjunto de datos se grafica en la Figura 6.11.

Dato	original	modificado
9	8.4750	60
23	5.4400	62
33	8.5500	61
37	3.7500	60

Tabla 6.14: datos modificados en el conjunto de datos de Ultrasonido.

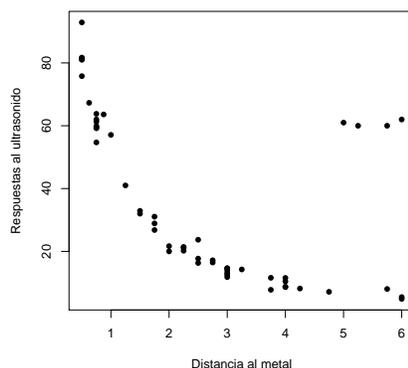


Figura 6.11: Grafico de los datos de ultrasonido con datos contaminados

Para estos datos, obtuvimos los estimadores dados en la Tabla 6.15. Como en el ejemplo anterior, vemos como la modificación de los datos, altera notablemente el estimador $\hat{\theta}_{OLS}$. Los estimadores robustos $\hat{\theta}_M$ y $\hat{\theta}_{MM}$ también sufren cambios, pero no tan drásticos. En la Figura 6.12 se muestran los datos junto con las funciones ajustadas.

Parámetro	$\hat{\theta}_{OLS}$	$\hat{\theta}_M$	$\hat{\theta}_{MM}$
θ_1	-0.25357	0.12834	0.17014
θ_2	-0.00358	0.00464	0.00531
θ_3	0.03380	0.01379	0.01214

Tabla 6.15: Estimaciones de los parámetros para el conjunto de datos contaminados.

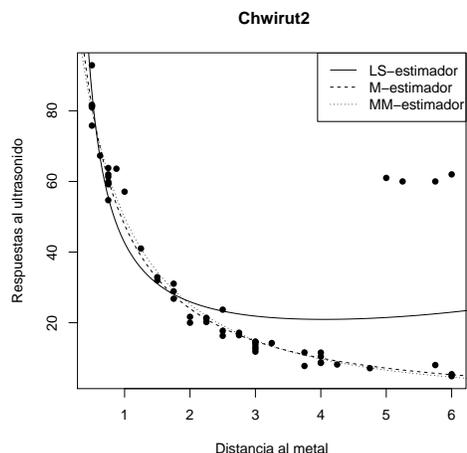


Figura 6.12: Curvas de ajuste para los datos de ultrasonido contaminados

La Tabla 6.16, presenta los errores estándar de los estimadores *OLS* y *M*-estimadores así como los respectivos valores del estadístico *t* para la hipótesis $H_0 : \theta_j = 0$ obtenidos sobre el conjunto de datos contaminados, mientras que la Tabla 6.17 da los valores del error estándar residual y la cantidad de iteraciones necesarias hasta obtener convergencia. El estimador $\hat{\beta}_{OLS}$ de θ_2 no da significativo, mientras que al utilizar los procedimientos robustos, todos los estimadores son significativos.

$\hat{\theta}_{OLS}$			
Parámetro	Estimador	Error Estándar	Estadístico <i>t</i>
θ_1	-0.25357	0.060333	-4.203
θ_2	-0.00358	0.003778	-0.949
θ_3	0.03380	0.007473	4.523
$\hat{\theta}_M$			
Parámetro	Estimador	Error Estándar	Estadístico <i>t</i>
θ_1	0.12834	0.0283401	4.528
θ_2	0.00464	0.0005591	8.302
θ_3	0.01379	0.0012620	10.929

Tabla 6.16: Errores estándar de los estimadores de los parámetros para el conjunto de ultrasonido contaminado y valores del estadístico *t* para la hipótesis $H_0 : \theta_j = 0$.

Estimador	Error Estándar Residual	Cantidad de iteraciones
$\hat{\theta}_{OLS}$	13.68	8
$\hat{\theta}_M$	1.785	8

Tabla 6.17: Errores estándar residuales y número de iteraciones hasta convergencia del estimador de mínimos cuadrados y del *M*-estimador para el conjunto de datos contaminado de ultrasonido.

Capítulo 7

Implementación del algoritmo de Stromberg en R

```
mmnl <- function(X,y,K,Vm,fun,grad,scale=NULL,trace=F)
{
z <- (y-fun(X,K,Vm))
if(all(z==0)) {
if(is.null(scale)) scale <- 0
return(b=b,fitted=y,residuals=rep(0,length(y)),
scale=scale,criterion=0)
}
if(is.null(scale)) scale <- mscale(z)
z <- z/0.9014/scale
iter <- 0
repeat {
iter <- iter+1
G <- grad(X,K,Vm)/0.9014/scale
w <- wt.hampel(z,a=1.5,b=3.5,c=8)
wG <- vecmat(sqrt(w),G)
d1 <- crossprod(G,psi.hampel(z))
d2 <- crossprod(wG)
maxd2 <- max(d2)
if(iter==1) {
if(all(w==0)) stop("Specified scale too small")
rho <- sum(rho.hampel(z))
lambda <- sqrt(mean(d2^2))/1000
I <- diag(dim(G)[2])
}
# Levenberg damping
bold <- b
old <- rho
lev <- 0
```

```

repeat {
lev <- lev+1
db <- solve( d2 + lambda*I, d1 )
b <- bold + db
z <- (y-fun(X,K,Vm))/0.9014/scale
rho <- sum(rho.hampel(z))
if(rho < old - 1e-15 || rho==0) break
if(lambda/maxd2 > 1e15) {
b <- bold
warning("Levenberg tolerance not achievable")
break
}
lambda <- 2*lambda
if(trace) cat("Lambda",lambda,"\n")
}
if(trace)
cat("Iteration",iter,"\nb",b,"\nObjective Function",rho,"\n")
if(lambda/maxd2 > 1e15) break
if(lev==1) lambda <- lambda/10
# Test for convergence
if( crossprod(d1,db) < 1e-8 ) break
if(iter > 40) {
warning("mmnl: Max iterations exceeded")
break
}
}
mu <- fun(X,K,Vm)
list(b=b,fitted=mu,residuals=y-mu,scale=scale,
criterion=rho)
}

mscale <- function(u)
{
if(mean(u==0) >= 0.5) return(0)
U <- abs(u)
s <- median(U)/0.6744898
iter <- 0
repeat {
iter <- iter+1
z <- u/0.212/s
d1 <- mean(rho.hampel(z))-3.75
d2 <- mean(z*psi.hampel(z))
s <- s*(1+d1/d2)
if(iter > 50) {
cat("mscale: Max iterations exceeded")
}
}
}

```

```

break
}
if(abs(d1/d2) < 1e-14) break
}
s
}

rho.hampel <- function(u, a = 1.5, b = 3.5, c = 8)
{
U <- abs(u)
A <- (U <= a) #increasing
B <- (U > a) & (U <= b) #flat
C <- (U > b) & (U <= c) #descending
D <- (U > c) # zero
rho <- U
rho[A] <- (U[A] * U[A])/2
rho[B] <- a * (U[B] - a/2)
rho[C] <- a * (b - a/2) + a * (U[C] - b) *
+(1 - (U[C] - b)/(c - b)/2)
rho[D] <- (a * (b - a + c))/2
rho
}

psi.hampel <- function(u, a = 1.5, b = 3.5, c = 8)
{
U <- abs(u)
B <- (U > a) & (U <= b) #flat
C <- (U > b) & (U <= c) #descending
D <- (U > c) # zero
psi <- u
psi[B] <- sign(u[B]) * a
psi[C] <- sign(u[C]) * a * (c - U[C])/(c - b)
psi[D] <- 0
psi
}

mmfreq <- function(y,x=NULL,freq,coef=NULL,constant=F,
scale=NULL,trace=F){
if(is.null(x)) x <- 0:(length(y)-1)
if(any(is.na(y))) {
x <- x[!is.na(y)]
y <- na.omit(y)
}
nfreq <- length(freq)
if(!is.null(coef)) if(length(coef) != constant+2*nfreq)

```

```

stop("Need 2 coefficients for each frequency")
if(length(y) < constant+3*nfreq)
stop("Need at least 3 observations for each frequency")
fun <- function(x,b) {
nfreq <- length(b) %/% 3
constant <- length(b) %% 3
f <- matrix(b[1:nfreq],1,nfreq)
cosi <- constant + (nfreq+1):(2*nfreq)
sini <- cosi + nfreq
  Xb <- cos(x%%f) %% b[cosi] + sin(x%%f) %% b[sini]
if(constant) Xb <- Xb + b[nfreq+1]
as.vector(Xb)
}
grad <- function(x,b) {
nfreq <- length(b) %/% 3
constant <- length(b) %% 3
f <- matrix(b[1:nfreq],1,nfreq)
cosi <- constant + (nfreq+1):(2*nfreq)
sini <- cosi + nfreq
G <- x*( matvec(cos(x%%f),b[sini])-matvec(sin(x%%f),b[cosi]))
if(constant)
return( cbind( G, 1, cos(x%%f), sin(x%%f) ) )
else
return( cbind( G, cos(x%%f), sin(x%%f) ) )
}
if(is.null(coef)) {
f <- matrix(freq,1,nfreq)
G <- cbind( cos(x%%f), sin(x%%f) )
coef <- ltsreg(G,y,intercept=constant)$coefficients
}
out <- mmnl(x,y,c(freq,coef),fun,grad,scale=scale,trace=trace)
out$freq <- Arg(exp(out$b[1:nfreq]*1i))
out$coef <- out$b[(nfreq+1):(3*nfreq+constant)]
out$b <- NULL
out
}

robfreq <- function(y,nfreq=1,s=NULL,breakdown=0.5,trace=F
,maxregs=6000) {
n <- length(y)
x <- 0:(n-1)
if(breakdown > 0.5) breakdown <- 0.5
if(breakdown < 0) breakdown <- 0
trim <- floor(n*(1-breakdown))+floor(3*nfreq/2)
cosi <- 1:nfreq

```

```

sini <- cosi + nfreq

# Make sure data is not trivial
if(all(y==0)) return(list(freq=0,coef=c(0,0),fitted=y,
residuals=y,scale=0,criterion=0))
if(var(y)==0) {
m <- mean(y)
return(list(freq=0,coef=c(m,0),fitted=rep(m,n),
residuals=rep(0,n),scale=0,criterion=0))
}

# Prony ORA estimator
if(trace) cat(" PronyFreq")
ora <- pronyfreq(y,nfreq,constant=F,maxit=5,warnings=F)
e <- ora$residuals^2
tsmin <- sum( sort(e)[1:trim] )
omegats <- ora$freq
alphats <- ora$coef
ets <- e
if(trace) {
cat(" (")
cat(round(omegats,4),sep="," )
cat(")")
}

# Check for exact fit
if(max(abs(ora$residual))/max(abs(y)) < 1e-11)
return(c(ora,list(scale=0,criterion=0)))

# If low breakdown wanted, return immediately
if(trim >= n) return(ora)

# Number of steps or spacings in elemental set
steps <- 3*nfreq-1

# Stromberg's recommendation for number of element sets
stromberg <- ceiling( log(0.001)/log(1-breakdown^(3*nfreq)) )
if(stromberg < n) stromberg <- n
if(stromberg > maxregs) stromberg <- maxregs

# Find maxs corresponding to Stromberg
nelemsets <- s <- 0
repeat {
s <- s+1
new <- n - steps*s

```

```

if(new < 1) {
  smax <- s-1
  prob <- rep(1,smax)
  break
}
if(s>1) {
  nelemsetsthin <- nelemsets+new*(1+(s-2)*nfreq)/(1+(s-1)*nfreq)
  if(nelemsetsthin >= stromberg) {
    smax <- s
    prob <- rep(1,smax)
    prob[smax] <- (1+(s-2)*nfreq) / (1+(s-1)*nfreq)
    break
  }
}
nelemsets <- nelemsets + new
if(nelemsets >= stromberg) {
  smax <- s
  prob <- rep(1,smax)
  break
}
}

# Make sure widest elemental sets at least at
# least 7% of data range. If smax is increased,
# thin out to required number of elemental sets
if(new > 1) {
  mins <- ceiling(n/14/steps)
  if(smax < mins) {
    scut <- smax
    smax <- mins
    s <- 1:smax
    prob <- pmin( 1, (1+(scut-1)*nfreq) / (1+(s-1)*nfreq) )
  }
  s <- 1:smax
  nsets <- n-steps*s
  prob <- prob*stromberg/sum(nsets*prob)
}

# Make sure no more than maxregs regressions,
# including harmonics
# If necessary, thin out to maximum number of regressions
s <- 1:smax
nregs <- (n-steps*s) * (1+(s-1)*nfreq)
if(sum(nregs*prob) > maxregs) {
  scut <- smax

```

```

repeat {
scut <- scut-1
prob <- pmin( 1, (1+(scut-1)*nfreq) / (1+(s-1)*nfreq) )
if(sum(nregs*prob) <= maxregs) {
prob <- pmin( 1, (1+scut*nfreq) / (1+(s-1)*nfreq) )
prob <- prob/sum(nregs*prob)*maxregs
break
}
if(scut == 1) {
prob <- prob/sum(nregs*prob)*maxregs
break
}
}
}

# Try elemental estimators
Xbase <- matrix(0,3*nfreq,2*nfreq)
X <- Xbase <- matrix(0,n,2*nfreq)
i <- 0:steps

# Step through all spacings
if(trace) cat(" Elemental")
for (spacing in s) {
if(trace) cat(" ",spacing,sep="")

# Sets within each spacing
nsets <- n-(3*nfreq-1)*spacing
x1sample <- sample(nsets,round(prob[spacing]*nsets))
for (x1 in x1sample) {
xe <- x1+spacing*i
ye <- y[xe]
eomega <- elemfreq(ye,nfreq=nfreq)
if ( !is.null(eomega) ) {

# Set omega to base frequency
eomegabase <- eomega/spacing

# Compute criteria at base frequency
eo <- matrix(eomegabase,1,nfreq)
xoe <- (xe-1)%*%eo
Xbase[,cosi] <- cos(xoe)
Xbase[,sini] <- sin(xoe)
xo <- x%*%eo
Xbase[,cosi] <- cos(xo)
Xbase[,sini] <- sin(xo)

```

```

qrXe <- qr(Xebase)
alpha <- qr.coef(qrXe,ye)
mu <- Xbase %*% alpha
e <- (y-mu)^2
tsbase <- sum( sort(e)[1:trim] )
# cat(spacing,x1,"0 0",tsbase,"\n")
eomegats <- eomegabase
ealphats <- alpha
eets <- e
tsminh <- tsbase

if (spacing>1) {
# Try to find a better combination of harmonics for
# this elemental set
Xemin <- Xebase
Xmin <- Xbase
for (f in 1:nfreq) {
Xe <- Xemin
X <- Xmin
# Search harmonics of frequency f
for (h in 1:(spacing-1)) {
eomegah <- acos(cos(eomegabase[f]+2*pi*h/spacing))
xoe <- (xe-1)*eomegah
xo <- x*eomegah
Xe[,f] <- cos(xoe)
Xe[,f+nfreq] <- sin(xoe)
X[,f] <- cos(xo)
X[,f+nfreq] <- sin(xo)
qrXe <- qr(Xe)
alpha <- qr.coef(qrXe,ye)
mu <- X %*% alpha
e <- (y-mu)^2
ts <- sum( sort(e)[1:trim] )
# cat(spacing,x1,f,h,ts,"\n")
if (ts < tsminh) {
tsminh <- ts;
eomegats[f] <- eomegah;
ealphats <- alpha;
eets <- e
Xemin <- Xe;
Xmin <- X
}
}
}
}
}

```

```

# Now have best harmonic. Test again global minimum.
if (tsminh < tsmin) {
  tsmin <- tsminh
  omegats <- eomegats
  alphats <- ealphats
  ets <- eets
  if(trace) {
    cat(" (")
    cat(round(omegats,4),sep=",")
    cat(")")
  }
}
}
}

# Terminate if all frequencies are zero
if(all(omegats==0))
return(list(freq=omegats,coef=alphats,fitted=y-ets,
residuals=ets))

# If some but not all frequencies are zero,
# replace with constant term
constant <- any(omegats==0)
if(constant) omegats <- omegats[omegats != 0]

# Least squares on best half of data
good <- sort(sort.list(e)[1:trim])
ls.ts <- lsfreq(y=y[good],x=x[good],freq=omegats,
constant=constant)
if(trace) {
  cat(" LS (")
  cat(round(ls.ts$freq,4),sep=",")
  cat(")")
}

# Compute scale M estimate
s <- mscale(ls.ts$residuals)
if(trace) cat(" s=",round(s,3),sep="")

# Terminate if scale is zero
if(s/max(ls.ts$coef) < 1e-12) return(c(ls.ts,list(scale=s)))

# Location M estimate

```

```

mm <- mmfreq(y,freq=ls.ts$freq,coef=ls.ts$coef,scale=s,
constant=constant)
mm.ora <- mmfreq(y,freq=ora$freq,coef=ora$coef,scale=s)
if(mm.ora$criterion < mm.ora$criterion) mm <- mm.ora
if(trace) {
cat(" MM (")
cat(round(mm$freq,4),sep=","))
cat(")\n")
}
mm
}

```

```

elemfreq <- function(y,nfreq=NULL) {
# Fit frequencies to elemental set
# Gordon Smyth, U of Queensland, gks@maths.uq.edu.au
# 12 Jul 99. Last revised 15 Oct 99.

```

```

if(is.null(nfreq)) nfreq <- length(y) %/% 3
if (nfreq==1) {
if(y[2]==0) return(NULL)
d2 <- -(y[1]+y[3])/y[2]
if (abs(d2)>2 ) return(NULL) else return(acos(-d2/2))
}

```

```

i <- 0:(nfreq-1)
i1 <- 1
i2 <- 2*nfreq+1
b <- y[i1+i] + y[i2+i]
B <- matrix(0,nfreq,nfreq)
for (j in 1:(nfreq-1)) {
i1 <- i1+1
i2 <- i2-1
B[,j] <- y[i1+i] + y[i2+i]
}
B[,nfreq] <- y[nfreq+1+i]
qrB <- qr(B)
if (qrB$rank < nfreq) return(NULL)
d <- -qr.coef(qrB,b)
f <- log(polyroot( c(1,d,d[(nfreq-1):1],1) ))
if (any(abs(Re(f))>1e-8))
return(NULL)
else
return(sort(Im(f))[(nfreq+1):(2*nfreq)])
}

```

```
wt.hampel<-  
function(u, a = 2, b = 4, c = 8) ¿cambiar los valores?  
{  
  U <- abs(u)  
  A <- (U <= a) #increasing  
  B <- (U > a) & (U <= b) #flat  
  C <- (U > b) & (U <= c) #descending  
  D <- (U > c) # zero  
  w <- u  
  w[A] <- 1  
  w[B] <- a/U[B]  
  w[C] <- (a * ((c - U[C])/(c - b)))/U[C]  
  w[D] <- 0  
  w  
}
```

Bibliografía

- [1] Bates, D. M., and Watts, D. G. (1980). Relative curvature measures of nonlinearity (with Discussion). *J. R. Stat. Soc.*, Series B, **42**, 1-25.
- [2] Bates, D. M., and Watts, D. G. (1981). Parameter transformations for improved approximate confidence regions in nonlinear least squares. *Ann. Stat.*, **9**, 1152-1167.
- [3] Bates, D. M. and Watts, D. G. (1988) *Nonlinear Regression Analysis and Its Applications*. John Wiley and Sons, New York.
- [4] Beale, E.M.L.(1960). Confidence regions in non-linear estimation (with Discussion). *J. R. Stat. Soc.*, Series B **22**, 41-88.
- [5] Bird, H. A., and Milliken, G. A. (1976). Estimable functions in the nonlinear model. *Commun. Statist., Theory and Methods*, **6**, 999-1012.
- [6] Box, G. E. P., and Coutie, G. A. (1956). Application of digital computers in the exploration of functional relationships. *Proc. I.E.E.E.*, **103**, Part B, Suppl. 1, 100-107.
- [7] Crawley, M. J. (2007). *The R Book*. John Wiley and Sons, New York.
- [8] Cornish-Bowden, A. J. (1976). *Principles of Enzyme Kinetics*. Butterworths: London.
- [9] Donaldson, J. R., and Schnabel, R. B. (1987). Computational experience with confidence regions and confidence intervals for nonlinear least squares. *Technometrics*, **29**, 67-82.
- [10] Fasano, M. V. (2009). *Teoría asintótica de estimadores robustos en regresión no lineal*. Tesis Doctoral, Universidad Nacional de La Plata.
- [11] Fraiman, R. (1983). General M-estimators and applications to bounded influence estimation for non-linear regression. *Comm. Statist., Theory and Methods*, **22**, 2617-2631.
- [12] Goldberg, M. L., Bates, D. M., and Watts, D. G. (1983). Simplified methods of assessing nonlinearity. *Amer. Stat. Assoc. Proc. Bus. Econ. Statist.*, 67-74.
- [13] Huber, P. J. (1973) Robust Regression: Asymptotics, Conjectures and Monte Carlo. *Ann. Statist*, **1**, 799-821.
- [14] Huber, P. J. (1981). *Robust Statistics*. John Wiley and Sons, New York.
- [15] Lawrence, K.D. and Arthur, J.L. (1990). *Robust Regression - Analysis and Applications*, .

- [16] McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Stat.*, **11**, 59-67.
- [17] D.A. Ratkowsky. (1983). *Nonlinear Regression Modeling*. Marcel Dekker, New York.
- [18] Ratkowsky, D. A., and Dolby, G. R. (1975). Taylor series linearization and scoring for parameters in nonlinear regression. *Appl. Stat.*, **24**, 109-111.
- [19] Reich, J. G. (1981). On parameter redundancy in curve fitting of kinetic data. In L. Endrenyi (Ed.), *Kinetic Data Analysis: Design and Analysis of Enzyme and Pharmacokinetic Experiments*, pp. 39-50. Plenum Press: New York.
- [20] Reich, J. G., and Zinke, I. (1974). Analysis of kinetic and binding measurements IV. Redundancy of model parameters. *Studio Biophys.*, **43**, 91-107.
- [21] Ross, G. J. S. (1980). Uses of non-linear transformation in non-linear optimisation problems. In M. M. Barritt and D. Wishart (Eds.), COMPSTAT1980, *Proceedings in Computational Statistics*, pp. 382-388. Physica-Verlag: Vienna.
- [22] Rousseeuw, P. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.*, **79**, 871-880.
- [23] Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. En *Mathematical Statistics and its Applications (Vol. B)*, Editores: W. Grossmann, G. Pflug, I. Vincze y W. Wertz, pp. 283-297. Dordrecht: Riedel.
- [24] Rousseeuw, P y Yohai, V. (1984). Robust regression by means of S-estimators En *Robust and nonlinear time series analysis, Lecture Notes in Statistics*, **26**, 256-272. Springer, New York.
- [25] Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear Regression*. John Wiley and Sons, New York.
- [26] Stromberg, A. J. (1993). Computation of high breakdown nonlinear regression parameters. *J. Amer. Statist. Assoc.*, **88**, 237-244.
- [27] Yohai, V. J. (1987). High breakdown-point and high efficiency estimates for regression. *Ann. Statist.*, **15**, 642-656.
- [28] Yohai, V. J. y Zamar, R. H. (1988). High breakdown estimates of regression by means of the minimization of an efficient scale. *J. Amer. Statist. Assoc.*, **83**, 406-413.