Universidad de Valparaíso
CHILE

Facultad de Ciencias
Instituto de Estadística
Ingeniería en Estadística

Modelo semiparamétrico Birnbaum-Saunders reparametrizado con parámetro de precisión variando

Marcela Zenteno Zenteno
28 de diciembre del 2021

Profesor Guía
Germán Ibacache-Pulgar, Ph.D.
Instituto de Estadística, Universidad de Valparaíso

Profesora Co-Guía
Carolina Marchant Fuentes, Ph.D.
Facultad de Ciencias Básicas, Universidad Católica del Maule

Proyecto de titulación para optar al:

grado académico de: *Licenciado en Estadística*
título profesional de: *Ingeniero en Estadística*
minor en: *Estadística Financiera*

Agradecimientos

Este trabajo de título constituye la parte final de mis estudios de pregrado, razón por la cual aprovecho para agradecer a las personas que me han ayudado y apoyado en esta etapa de mi vida. Agradezco a Dios por todo lo bueno que me ha dado, sobretodo a mi familia, especialmente a mi madre y a mi hermana Nicole por todo su apoyo incondicional y por estar conmigo en cada proceso de mi vida. A mi profesor guía el Dr. Germán Ibacache-Pulgar y a mi profesora co-guía la Dra. Carolina Marchant por todos sus conocimientos, habilidades y buena disposición en el desarrollo de este proyecto. Le doy las gracias a todos mis profesores que estuvieron en esta etapa entregándome sus conocimientos durante estos cinco años y a mis compañeros por su amistad y compañía. A mi compañera y amiga Michelle, con quien forme un gran equipo durante todo este tiempo y a todas mis amigas y amigos por su cariño. A una de las personas más importante en mi vida, Javier Zenteno, quien no logro estar conmigo durante todo este proceso, pero que sin duda estaría orgulloso y lo recuerdo cada día con mucho amor. Finalmente, destacar la ayuda financiera proporcionada por el proyecto Fondecyt 11190636 y la beca de apoyo brindada por el Centro Interdisciplinario de Estudios Atmosféricos y Astroestadística.

Resumen

Los modelos semiparamétricos tienden a ser más flexibles, dado que permiten una estructura de regresión compuesta entre una componente paramétrica y otra no paramétrica. En este trabajo, se estudia e implementa el modelo semiparamétrico de Birnbaum-Saunders reparametrizado con parámetro de precisión variando. Se utiliza el algoritmo de Backfitting para obtener las estimaciones de máxima verosimilitud penalizada mediante el uso de splines cúbicos suavizados. Se procede a realizar el análisis residual del modelo y se desarrollan métodos de influencia local. Finalmente, se presenta una aplicación computacional mediante el software R-project del modelo propuesto a datos reales de contaminación, cuyo tema posee gran relevancia tanto a nivel nacional como mundial. De este modo, diversos profesionales tendrán a su disposición estas metodologías para ser utilizadas libremente.

Palabras claves: distribución Birnbaum-Saunders, estimador de máxima verosimilitud penalizada, método de influencia local, algoritmo de Backfitting ponderado, modelo semiparamétrico con precisión variando.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

La demanda de datos y las ciencias que giran en torno a su análisis crecen vertiginosamente en paralelo a las soluciones que el medio ambiente requiere. Las fuentes de datos ambientales se han multiplicado, y es por eso que las comunidades de ciencias naturales, oceanografía, ingenierías, entre otras, buscan soluciones flexibles y adaptadas para analizar cuya información; ver González (2020). La necesidad de analizar conjuntos de datos provenientes del área ambiental ha impulsado el desarrollo de nuevas herramientas estadísticas. Por ejemplo, los modelos semiparamétricos han recibido una atención especial los últimos años, debido a la flexibilidad de su estructura de regresión comparada con los modelos de regresión paramétricos clásicos. Por su parte, la distribución Birnbaum-Saunders (BS) ha despertado un interés no menor en el modelamiento de datos ambientales, principalmente por sus argumentos teóricos, sus propiedades y su relación con la distribución normal. Esta distribución posee un parámetro de precisión que, debido a su relación inversamente proporcional con el parámetro de dispersión, es utilizado para cuantificar la variabilidad de las observaciones. En este contexto, Cárcamo *et al*. (2021) desarrolló un modelo semiparamétrico basado en una reparametrización de la distribución Birnbaum-Saunders y lo usó para modelar un conjunto de datos de contaminación atmosférica. Este modelo, flexible en su componente sistemática debido a la presencia de funciones suaves para modelar los efectos no lineales de algunas covariables, asume que la precisión es constante a través de las observaciones, supuesto que en la práctica muchas veces no se verifica. Con todo, la literatura respecto del modelamiento semiparamétrico basado en la distribución Birnbaum-Saunders reparametrizada con parámetro de precisión variando (heterocedástico) es limitada. En virtud de esto, el estudio de tales modelos estadísticos puede contribuir al análisis de datos de diferentes áreas de investigación, en especial de datos ambientales.

1.1. Modelo de regresión semiparamétrico

Los modelos de regresión semiparamétricos permiten modelar conjuntos de datos cuya relación entre la variable de respuesta y el conjunto de covariables no necesariamente son de tipo lineal. Tales modelos, asumen la siguiente estructura:

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + f(\mathrm{t}_i) + \epsilon_i \qquad (i = 1, \ldots, n)\,, \tag{1.1}$$

en que y_i corresponde a la respuesta de la i-ésima unidad experimental, $\boldsymbol{x}_i^T$ es un vector de covariables

$(p \times 1)$, $\boldsymbol{\beta}$ es un vector de parámetros desconocidos $(p \times 1)$, t_i es una covariable escalar, f es una función suave arbitraria, y ϵ_i es un error aleatorio, para $i = 1, \ldots, n$. En esta clase de modelos, la tendencia no lineal es modelada a través de una función suave, f, y no mediante una estructura de regresión paramétrica definida. En la literatura existen diferentes trabajos relacionados a esta clase de modelos. Por ejemplo, Cook (1986) estudió la normalidad asintótica del estimador del coeficiente de regresión para el caso balanceado. Grenn (1987) evaluó el compartamiento asintótico del estimador de máxima verosimilitud penalizada y propuso algunas definiciones para los grados de libertad y los residuos. Speckman (1988) utilizó dos métodos para estimar el coeficiente de regresión y la función no paramétrica basados en suavizamiento spline parcial y análsis residual parcial. Gannaz (2007) desarrolló un método de estimación basado en una expansión wavelet de la componente no paramétrica del modelo semiparamétrico. En el contexto del análisis de diagnóstico, Eubank (1984) derivó algunas medidas de diagnóstico basadas en los leverage y los residuos. Silverman (1985) definió dos tipos de residuos en analogía con regresión paramétrica. Zhu *et al*. (2003) desarrolló la técnica de influencia local para diferentes esquemas de perturbaciones bajo el modelo semiparamétrico normal, mientras que Ibacache-Pulgar & Paula (2011) extendieron tales resultados para el modelo semiparamétrico t-Student. Mayores detalles acerca de los métodos de estimación y extensiones de los modelos de regresión semiparamétricos pueden ser encontrados en Hastie & Tibshirani (1990) y Green & Silverman (1994).

1.2. Datos de contaminación atmosférica

La contaminación por partículas es uno de los principales problemas medioambientales urbanos a nivel mundial, afecta a la salud humana y a la calidad de vida. Según la Organización Mundial de la Salud (OMS), nueve de cada diez personas del planeta respiran aire con altos niveles de contaminantes y siete millones de personas mueren cada año por esta causa (`www.who.int`). El material particulado (MP) es una mezcla compleja de partículas sólidas y líquidas de diferente origen, tamaño, forma y composición química, con impacto sobre la salud humana, los ecosistemas, la visibilidad y la infraestructura, y con consecuencias económicas y sociales Grantz *et al*. (2003). El material particulado se clasifica según su diámetro, debido a que el tamaño de las partículas determina los lugares de deposición en el tracto respiratorio; las más gruesas (las que tienen un diámetro superior a 10 micrómetros -μm-) no penetran en las vías respiratorias, estas se depositan en el tracto respiratorio superior y son eliminadas por la acción de los cilios; estas partículas inhalables que miden menos de 10 μm se denominan MP10 y las menores de 2.5 μm MP2.5. A medida que el tamaño disminuye, hay una mayor posibilidad de que el material particulado penetre más profundamente en los alvéolos y las vías respiratorias más pequeñas. En particular, la frecuente exposición a altos niveles de MP produce diversos efectos, pero la naturaleza de ellos varía según su composición. Existe evidencia de un aumento del riesgo de enfermedades cardiovasculares y de mortalidad por la exposición a MP2.5, que se produce incluso después de períodos de tiempo cortos, como horas o semanas; véase Cavieres *et al*. (2020). En Chile, según el Ministerio de Medio Ambiente, 3.494 personas murieron prematuramente debido a los niveles críticos del aire durante 2017, principalmente por concentraciones extremas de MP2.5, y nueve millones de habitantes están expuestos a niveles de contaminación que superan los estándares de calidad del aire (`https://bit.ly/2u40gDq`). A medida que la calidad del aire urbano disminuye, aumenta el riesgo de accidente cerebrovascular, enfermedades cardíacas, cáncer de pulmón y enfermedades respiratorias crónicas y agudas, para las personas que viven en ciudades con altos niveles de contaminación del aire. En el país existen ciudades que presentan graves problemas de contaminación atmosférica, principalmente en el periodo GEC contemplado entre el 1 de abril y el 31 de agosto de cada año. Estas ciudades son Coyhaique, Linares, Osorno, Padre las Casas, Puerto Montt, Santiago, Rancagua, Temuco y Valdivia, las cuales se encuentran entre las 10 ciudades sudamericanas más contaminadas, según un informe de Greenpeace que mide

el índice de calidad del aire en base a los niveles de MP2.5, la contaminación en estas ciudades puede llegar a ser insalubre en algunos meses del año (`https://bit.ly/2TgF1NY`). En ocasiones se producen episodios periódicos de concentraciones extremas de contaminación atmosférica para determinados contaminantes atmosféricos, las altas concentraciones asociadas a estos varían con las fluctuaciones geográficas y meteorológicas y dependen de los cambios tanto en la fuente como en el tipo de emisiones, debido a estas variaciones, las concentraciones de MP se tratan como variables aleatorias no negativas que pueden modelarse mediante distribuciones estadísticas, con frecuencia, estas distribuciones son asimétricas y presentan una asimetría positiva Marchant *et al*., (2013). La actual metodología oficial utilizada por la autoridad chilena de Santiago para predecir las concentraciones de MP10 se basa en un modelo de regresión múltiple que utiliza como covariables las concentraciones de contaminantes atmosféricos y las variables meteorológicas Morales *et al*., 2012, esto ayuda a predecir el valor máximo de la concentración media de 24 horas de PM10 en g/metros cúbicos normalizados (Nm^3) para el periodo comprendido entre las 00:00 a 24:00 horas del día siguiente. En el año 2015, a través del Decreto Supremo número 15/2015 y la resolución número 9664/2015, se instruyó por parte del Ministerio de Salud declarar alerta sanitaria empleando también concentraciones de MP2.5. En este trabajo se propone desarrollar un modelo semiparamétrico Birnbaum-Saunders reparametrizado con parámetro de precisión variando para modelar concentraciones de contaminantes en la comuna de Pudahuel, Chile. Esta aplicación está motivada dado que la inclusión de funciones no paramétricas ha mejorado profundamente la flexibilidad de modelación para acomodar efectos no lineales de las covariables, en este caso, las concentraciones de contaminantes y las variables meteorológicas como velocidad del viento, temperatura, humedad relativa. Las estructuras semiparamétricas se han utilizado con éxito para modelar componentes no lineales; véase, por ejemplo, Ibacache-Pulgar *et al*. (2021). Además, esta aplicación se apoya en argumentos teóricos que permiten justificar el uso de distribuciones Birnbaum-Saunders para datos ambientales; véase Leiva *et al*. (2015) para una novedosa justificación matemática en ciencias ambientales basada en leyes físicas.

1.3. Hipótesis

1. Es viable estimar los parámetros y estudiar algunos aspectos de inferencia estadística en el modelo semiparamétrico Birnbaum-Saunders reparametrizado con parámetro de precisión variando.
2. Es posible desarrollar un análisis de influencia para el modelo semiparamétrico Birnbaum-Saunders reparametrizado con parámetro de precisión variando.

1.4. Objetivos

Objetivo general

Desarrollar un proceso iterativo para estimar los parámetros asociados a la componente sistemática y a la estructura heterocedástica del modelo, y llevar a cabo un análisis de influencia en el modelo semiparamétrico Birnbaum-Saunders reparametrizado con parámetro de dispersión variando

Objetivos específicos

1. Discutir algunos aspectos teóricos del modelo propuesto.
2. Desarrollar un procedimiento de estimación para los parámetros del modelo basados en la función de log-verosimilitud doblemente penalizada y los algoritmos Scoring de Fisher y Backfitting ponderado.
3. Desarrollar un análisis de influencia e implementarlo computacionalmente en el software estadístico R-project.
4. Aplicar el modelo propuesto a conjuntos de datos provenientes del área medioambiental.

Capítulo 2

Modelo semiparamétrico Birnbaum-Saunders reparametrizado con parámetro de precisión variando

En este capítulo se considera la extensión del modelo de regresión de Birnbaum-Saunders, generando el modelo denominado modelo semiparamétrico Birnbaum-Saunders reparametrizado con parámetro de precisión variando. En la Sección 2.1 se presenta a modo de introducción la revisión bibliográfica de los modelos de regresión Birnbaum-Saunders. En la Sección 2.2 se realizá la descripción de la distribución Birnbaum-Saunders, para luego, en la Sección 2.3 se presenta la distribución Birnbaum-Saunders reparametrizada (BSR). En el apartado la Sección 2.4 se expone el modelo propuesto correspondiente al modelo semiparamétrico BSR con parámetro de precisión variando y la representación matricial de las componentes sistemáticas. Finalmente, en la Sección 2.5 se muestra la función de verosimilitud penalizada asumiendo que la función suave pertenece al espacio de funciones de Sobolev.

2.1. Introducción

Es sabido que, para la cantidad de datos de fatiga que suelen obtenerse, casi cualquier familia de distribuciones paramétricas bidimensionales puede ajustarse razonablemente bien. De hecho distribuciones tales como la lognormal, la Weibull, la Gamma, etc., pueden ajustarse mediante una estimación paramétrica y, debido a los tamaños de muestra relativamente pequeños, casi ninguna puede rechazarse, por ejemplo, mediante una prueba de bondad de ajuste de Chi-cuadrado. Sin embargo, cuando se trata de predecir la ”vida segura”, por ejemplo, el milésimo percentil, existe una gran discrepancia entre estos modelos. Por esta razón, una familia de distribuciones que se obtenga a partir de consideraciones sobre las características básicas del proceso de fatiga debería ser más persuasiva en sus implicaciones que cualquier familia ad hoc elegida por razones extrañas Birnbaum & Saunders (1969). La distribución Birnbaum-Saunders ha recibido atención considerable durante los últimos años, debido a sus argumentos teóricos asociados a los procesos de daño acumulativo, sus propiedades y su relación con la distribución normal. Además, se ha usado con bastante eficacia para modelar tiempos de falla de materiales sujetos a la fátiga. En concreto, se supone que la cantidad de daño acumulado que permite generar la distribución Birnbaum-Saunders sigue una distribución normal. Este modelo corresponde a una distribución unimodal, positivamente sesgada de dos parámetros y con soporte positivo. En las últimas décadas, se han estudiado ampliamente aspectos teóricos, metodológi-

cos y prácticos del modelo Birnbaum-Saunders; ver Leiva *et al*. (2008), Vilca *et al*. (2010), Paula *et al*. (2012), entre otros. Tiempo después, Leiva *et al*. (2014) presentan un nuevo enfoque para los modelos de regresión de Birnbaum-Saunders, que permite analizar los datos en su escala original y modelar la varianza no constante. Importante mencionar que, la modelización de la variabilidad ha sido ampliamente discutida en la literatura relacionada con la heteroscedasticidad; véase Van Keilegom & Wang (2010). Por ejemplo, Cook & Weisberg (1983) estudiaron los modelos normales heteroscedásticos. Taylor & Verbyla (2004) describieron conjuntamente los parámetros de localización y dispersión del modelo Student-t. Lin *et al*. (2009) consideraron las pruebas de heteroscedasticidad en los modelos de regresión de t de Student. Sin embargo, existen pocos trabajos que modelen heteroscedasticidad mediante parámetros de precisión. Ferrari *et al*. (2011) consideraron los modelos de regresión beta para los que el parámetro de precisión no es constante y lo describieron como una función de las variables explicativas (covariables). Simas *et al*. (2010) asumieron una estructura de regresión no lineal para el parámetro de precisión utilizando una distribución beta, mientras que Rocha & Simas (2011) derivaron el método de influencia local en este modelo. Paula, G. (2013) modeló simultáneamente la media y precisión de la distribución gamma, y realizó un análisis de diagnóstico en modelos lineales generalizados dobles. Para las regresiones Birnbaum-Saunders basadas en su parametrización original, Rieck & Nedelman (1991) y Galea *et al*. (2004) asumieron que el parámetro de forma es homogéneo entre los datos. Xie & Wei (2007) propusieron una prueba de homogeneidad de este parámetro de forma. El problema de la estimación en el modelo semiparamétrico Birnbaum-Saunders reparametrizado con parámetro de precisión variando aún no se ha discutido en la literatura. Sin embargo, algunos autores han abordado este problema para algunos modelos relacionados. Por ejemplo, Santos–Neto *et al*. (2016) proponen un modelo de regresión BSR con precisión variando, que permite describir la heteroscedasticidad, ampliando el trabajo de Leiva *et al*. (2014). Estimó parámetros del modelo a través del método de máxima verosimilitud, introdujó pruebas de hipótesis para el parámetro de precisión, así presentó cuatro tipos de residuos para el modelo RBS, además, mediante simulaciones de Monte Carlo, analizó la sensibilidad de los estimadores de máxima verosimilitud mediante el método de influencia local. También, Ibacache-Pulgar *et al*. (2012) estudiaron los modelos aditivos semiparamétricos bajo distribuciones simétricas y estimaron el coeficiente de regresión y las funciones suaves a través de un algoritmo de backfitting ponderado, que conduce a un spline cúbico como solución para las funciones no paramétricas. Recientemente, Ibacache-Pulgar *et al*. (2021) estudiaron el modelo de regresión beta aditivo semiparamétrico y propusieron el algoritmo backfitting para obtener las estimaciones de máxima verosimilitud penalizada. Cárcamo *et al*. (2021) formuló un modelo aditivo semiparamétrico basado en la distribución Birnbaum-Saunders reparametrizada y realiza un análisis de diagnóstico basado en el método de influencia local. Además obtiene un algoritmo de backfitting para obtener las estimaciones de máxima de verosimilitud penalizada mediante el uso de splines cúbicos.

2.2. Distribución Birnbaum-Saunders

Birnbaum & Saunders (1969) propusieron un modelo estadístico para la resistencia a la fatiga de estructuras sometidas a esfuerzos cíclicos. Este modelo, que recibe el nombre de distribución Birnbaum-Saunders (BS), es unimodal y sesgada positivamente, posee soporte positivo y dos parámetros, tiene buenas propiedades y una estrecha relación con la distribución normal. Formalmente, si la variable aleatoria Y sigue una distribución BS con parámetro de forma $\alpha > 0$ y parámetro de escala $\varrho > 0$, denotado por $Y \sim BS(\alpha, \varrho)$, su función de densidad de probabilidad toma la forma:

$$f_Y(y;\alpha,\varrho) = \frac{\exp(\alpha^{-2})}{2\alpha\sqrt{2\pi\varrho}} y^{-3/2}\,[y+\varrho]\exp\left(-\frac{1}{2\alpha^2}\left[\frac{y}{\varrho}+\frac{\varrho}{y}\right]\right), \quad y>0.$$

Es importante señalar que el parámetro ϱ corresponde a la mediana de la distribución de Y. Además, se puede demostrar que $\frac{1}{Y} \sim \text{BS}(\alpha, \frac{1}{\varrho})$ y $bY \sim \text{BS}(\alpha, b\varrho)$ cuando $b > 0$. Con respecto a la media y la varianza de la variable aleatoria Y, se tiene que $\text{E}(Y) = \varrho[1 + \frac{\alpha^2}{2}]$ y $\text{Var}(Y) = [\alpha\varrho]^2[1 + \frac{5\alpha^2}{4}]$, respectivamente. Trabajos recientes han demostrado que el uso de esta distribución no obedece a una simple extensión matemática, sino mas bien a la necesidad de modelar fenómenos aleatorios complejos de diversas áreas de investigación, incluyendo la área ambiental; ver, por ejemplo, Galea *et al*. (2004), Leiva *et al*. (2008), Vilca *et al*. (2010), Marchant *et al*. (2016 b), entre otros. Leiva *et al*. (2008) introdujeron el modelo log-lineal t-Birnbaum-Saunders y desarrollaron algunas medidas de diagnóstico. Además, aplicaron sus resultados a un conjunto de datos reales asociado a pacientes con cáncer de pulmón.

2.3. Distribución Birnbaum-Saunders reparametrizada

Santos–Neto *et al*. (2012) propusieron una nueva parametrización de la distribución Birnbaum-Saunders con el propósito de obtener estimadores insesgados y consistentes, ortogonalidad entre los parámetros, y la posibilidad de describir la media de la distribución a través de una estructura de regresión sin tener que transformar los datos. Especificamente, la variable aleatoria Y sigue una distribución Birnbaum-Saunders reparametrizada si su función de densidad es dada por:

$$f(y;\mu,\delta) = \frac{\exp(\delta/2)\sqrt{\delta+1}}{4\,y^{3/2}\sqrt{\pi\mu}}\left[y + \frac{\delta\mu}{\delta+1}\right]\exp\left(-\frac{\delta}{4}\left[\frac{y\{\delta+1\}}{\delta\mu} + \frac{\delta\mu}{y\{\delta+1\}}\right]\right), \quad y > 0. \tag{2.1}$$

En este caso, se utiliza la notación $Y \sim \text{RBS}(\mu, \delta)$. La media y la varianza de Y están dadas por $\text{E}[Y] = \mu$ y $\text{Var}[Y] = \mu^2/\phi$, respectivamente, donde $\phi = (\delta+1)^2/(2\delta+5)$, tal que, δ se puede interpretar como un parámetro de precisión, es decir, para valores fijos de μ, cuando $\delta \to \infty$, la varianza de Y tiende a cero. Además, para μ fijo, si $\delta \to \infty$, se tiene que $\text{Var}[Y] \to 5\mu^2$. Se puede observar que $\text{Var}[Y] = \mu^2/\phi$ es similar a la función de varianza de la distribución gamma, en cuyo caso, la varianza tiene una relación cuadrática con su media. También, es posible demostrar que $bY \sim \text{RBS}(b\mu, \delta)$, con $b > 0$, y $\frac{1}{Y} \sim \text{RBS}(\mu^*, \delta)$, donde $\mu^* = (\delta+1)/(\delta\mu)$.

2.4. Modelo semiparamétrico Birnbaum-Saunders reparametrizado con parámetro de precisión variando

Sean $y_1, \ldots, y_n$ variables aleatorias independientes, donde $y_i \sim \text{BSR}(\mu_i, \delta_i)$, para $i = 1, \ldots, n$, y $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ las observaciones correspondientes. A continuación, se define un modelo estadístico basado en (2.1) por las siguientes componentes sistemáticas

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta} + f(\mathrm{t}_i) \qquad \text{y} \qquad h(\delta_i) = \tau_i = \mathbf{z}_i^T\boldsymbol{\alpha} \qquad (i = 1, 2, \ldots, n), \tag{2.2}$$

o equivalente

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{n}_i^T\mathbf{f} \qquad \text{y} \qquad h(\delta_i) = \tau_i = \mathbf{z}_i^T\boldsymbol{\alpha} \qquad (i = 1, 2, \ldots, n),$$

donde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, para $p < n$, es un vector de parámetros desconocidos a estimar, $\mathbf{x}_i^T = (1, x_{i_2}, \ldots, x_{i_p})$

representa los valores de p regresores, tal que $\mu_i = g^{-1}(\mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{n}_i^T\mathbf{f})$, con g^{-1} siendo la función inversa de g, $f(\cdot)$ es una función suave, univariada y arbitraria que cuantifica el efecto de la variable explicativa t, $\mathbf{n}_i^T$ denota la *i*-ésima fila de la matriz de incidencia $\mathbf{N}$, cuyo (i, l)ésimo elemento es igual a la función indicadora $I(\mathrm{t}_i = \mathrm{t}_l^0)$. Formalmente, se tiene que $\mathrm{Var}(y_i)$ es una función de μ_i y, en consecuencia, de los regresores $\boldsymbol{x}_i$. Entonces, debido a que se esta modelando la media basada en una estructura en particular, también se esta modelando la varianza debido a que $\mathrm{Var}(y_i) = \mu_i^2/\phi$. Además, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^T$, para $q < n$, es un vector de parámetros desconocidos a estimar, $\mathbf{z}_i^T = (1, z_{i_2}, \ldots, z_{i_p})$ representa los valores de q regresores, tal que $\delta_i = h^{-1}(\mathbf{z}_i^T\boldsymbol{\alpha})$, con h^{-1} siendo la función inversa de h. En el modelo dado en (2.2), las funciones de enlace g: $\mathbb{R} \rightsquigarrow \mathbb{R}^+$ y h: $\mathbb{R} \rightsquigarrow \mathbb{R}^+$ son estrictamente monótonas, positivas y al menos dos veces diferenciables; por ejemplo, $g(\mu) = \log(\mu)$ y $h(\delta) = \log\delta$, entre otros.

Función de enlace	g	h	$\frac{d\mu}{d\eta}$	$\frac{d\delta}{d\tau}$	$\frac{d^2\mu}{d\eta^2}$	$\frac{d^2\delta}{d\tau^2}$
Identidad	$\mu = \eta$	$\delta = \tau$	1	1	0	0
Logarítmica	$\log(\mu) = \eta$	$\log(\delta) = \tau$	μ	δ	μ	δ
Raíz cuadrada	$\sqrt{\mu} = \eta$	$\sqrt{\delta} = \tau$	$2\sqrt{\mu}$	$2\sqrt{\delta}$	2	2

Cuadro 2.1: Derivadas para la función de enlace indicada.

En el Cuadro 2.1 se muestran las funciones de enlace más comunes para g y h, junto con sus primeras y segundas derivadas.

2.5. Función de log-verosimilitud penalizada

La función de log-verosimilitud para el parámetro $\boldsymbol{\theta} = [\boldsymbol{\beta}^\top, \boldsymbol{f}^\top, \boldsymbol{\alpha}^\top]^\top$ obtenida de (2.1) relacionada con $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_n]^\top$ y $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_n]^\top$ para la clase de modelos con funciones de enlace en (2.2), viene dada por:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\mu_i, \delta_i; y_i), \tag{2.3}$$

donde (ignorando el término constante)

$$\ell_i(\mu_i, \delta_i; y_i) = \frac{\delta_i}{2} - \frac{log(\delta_i + 1)}{2} - \frac{log(\mu_i)}{2} - \frac{3log(y_i)}{2} + log(\delta_i y_i + y_i + \delta_i\mu_i) - \frac{y_i[\delta_i + 1]}{4\mu_i} - \frac{\delta_i^2\mu_i}{4y_i[\delta_i + 1]}. \tag{2.4}$$

Para evitar el problema de ajuste y la no identificación del parámetro $\boldsymbol{\beta}$, algunos autores sugieren incorporar en la función de log-verosimilitud del modelo, un término de penalización, denotado aquí por $J(f)$, sobre la función suave f que pertenece al espacio de funciones de Sobolev,

$$\mathcal{W}_2^{(2)} \quad = \quad \{f : f, f^{(1)}\text{abs. cont.}, f^{(2)} \in \mathcal{L}^2[a, b]\} \ .$$

En este caso, la función log-verosimilitud penalizada se expresa como:

$$L_p(\boldsymbol{\theta}, \boldsymbol{\lambda}) = L(\boldsymbol{\theta}) + \lambda^* J(f), \tag{2.5}$$

donde $\lambda^* = \lambda^*(\lambda)$ es una constante que depende del parámetro de suavizamiento $\lambda \geq 0$ que controla el equilibrio entre la bondad de ajuste y suavidad de la función estimada. En la literatura podemos encontrar diferentes tipos de penalizaciones dependiendo del método propuesto para ajustar las curvas no paramétricas. Se considera la norma al cuadrado como medida de curvatura de las funciones; esto es,

$$J(f) = ||f||^2 = \int_a^b f^{(2)}(\mathrm{t})^2 d\mathrm{t}. \tag{2.6}$$

El primer término del lado derecho de la ecuación (2.5) mide la bondad de ajuste, mientras que el segundo término, denotado por (2.6), penaliza la rugosidad de f con un parámetro fijo λ. En este caso, la estimación de f conduce a una spline cúbica natural con nodos en el punto t_l^0, es decir, es por partes un polinomio de grado 3 en cada intervalo $[\mathrm{t}_l, \mathrm{t}_{l+1}]$ para $l = 1, 2, \ldots, r-1$. Según Green & Silverman (1994), $J(f)$ se puede escribir como

$$J(f) = \int_a^b [f^{(2)}(\mathrm{t})]^2 \mathrm{dt} = \mathbf{f}^T \mathbf{K} \mathbf{f} \ ,$$

donde $\boldsymbol{K}$ es una matriz definida no-negativa $(r \times r)$ que depende sólo de los nodos t^0. Entonces, si se considera $\lambda^* = -\lambda/2$, la función de log-verosimilitud penalizada (2.5) se puede expresar como

$$L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda) = L(\boldsymbol{\theta}) - \frac{\lambda}{2} \mathbf{f}^T \mathbf{K} \mathbf{f} \ . \tag{2.7}$$

Es importante tener en cuenta que un aspecto esencial del proceso de modelamiento semiparamétrico está relacionado con la estimación o selección del parámetro de suavizamiento. En la literatura existen varios métodos eficientes de selección, entre los que destacan la Validación Cruzada (VC), la Validación Cruzada Generalizada (VCG), el Criterio de Akaike (AIC) y el Error Cuadrático Medio (ECM).

Capítulo 3

Estimación de Parámetros

En este capítulo se considera el problema de estimación de parámetros asociado al modelo semiparamétrico Birnbaum-Saunders reparametrizado con parámetro de precisión variando. Inicialmente, en la Sección 3.1 se presenta una introducción, donde se muestran algunos trabajos que tratan el problema de estimación en el contexto semiparamétrico. En la Sección 3.2 se obtienen las funciones score penalizadas. En la Sección 3.3 y 3.4, se calcula la matriz hessiana penalizada y la matriz de información de Fisher penalizada, respectivamente.

3.1. Introducción

La estimación de parámetros es una rama de la estadística que implica el uso de datos de muestra para estimar parámetros de una distribución. El problema de la estimación en el modelo semiparamétrico Birnbaum-Saunders reparametrizado con parámetro de precisión variando aún no se ha discutido en la literatura. Sin embargo, algunos autores han abordado este problema para algunos modelos relacionados. Por ejemplo, Santos–Neto *et al*. (2016) proponen una metodología basada en un modelo de regresión Birnbaum-Saunders reparametrizado con precisión variable, que generaliza los trabajos existentes en la literatura sobre el tema. Esta metodología incluye la estimación de los parámetros del modelo, ampliando así el trabajo de Leiva *et al*. (2014) quienes estimaron parámetros del modelo a través del método de máxima verosimilitud. También, Ibacache-Pulgar *et al*. (2012) estudiaron los modelos aditivos semiparamétricos bajo distribuciones simétricas y estimaron el coeficiente de regresión y las funciones suaves a través de un algoritmo de backfitting ponderado, que conduce a un spline cúbico como solución para las funciones no paramétricas. Hastie & Tibshirani (1986) utilizaron el algoritmo de puntuación local para ajustar el componente aditivo de un modelo aditivo generalizado, y utilizaron datos de respuesta binaria y de supervivencia para ilustrar la aplicación del método; véase también Hastie & Tibshirani (1987). Tiempo después, Hastie & Tibshirani (1993) estudiaron los modelos aditivos generalizados en los que los coeficientes de regresión varían suavemente en función de otras covariables, y demostraron, basándose en el criterio de mínimos cuadrados penalizados, que los estimadores de las funciones no paramétricas corresponden a un spline cúbico natural. Recientemente, Ibacache-Pulgar *et al*. (2021) estudiaron el modelo semiparamétrico de regresión beta aditivo y propusieron el algoritmo backfitting para obtener las estimaciones de máxima verosimilitud penalizada mediante el uso de splines cúbicos naturales de suavizamiento. Calculan su correspondiente función score y desarrollan un proceso iterativo para estimar sus parámetros.

3.2. Funciones score penalizadas

Suponiendo que la función (2.7) es regular con respecto a $\boldsymbol{\beta}$, $\mathbf{f}$ y $\boldsymbol{\alpha}$, se tiene que el vector de funciones score penalizadas de $\boldsymbol{\theta}$ está dado por

$$\mathbf{U}_{\mathrm{p}}(\boldsymbol{\theta}) = \frac{\partial L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda)}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \mathbf{U}_{\mathrm{p}}^{\beta}(\boldsymbol{\theta}) \\ \mathbf{U}_{\mathrm{p}}^{\mathrm{f}}(\boldsymbol{\theta}) \\ \mathbf{U}_{\mathrm{p}}^{\alpha}(\boldsymbol{\theta}) \end{pmatrix}, \tag{3.1}$$

cuyos elementos se pueden escribir de la siguiente forma

$$\begin{aligned} \mathbf{U}_{\mathrm{p}}^{\beta}(\boldsymbol{\theta}) &= \mathbf{X}^T\mathbf{A}(\mathbf{y}^* - \boldsymbol{\mu}^*)\,, \\ \mathbf{U}_{\mathrm{p}}^{\mathrm{f}}(\boldsymbol{\theta}) &= \mathbf{N}^T\mathbf{A}(\mathbf{y}^* - \boldsymbol{\mu}^*) - \lambda\mathbf{K}\mathbf{f} \qquad \text{y} \\ \mathbf{U}_{\mathrm{p}}^{\alpha}(\boldsymbol{\theta}) &= \mathbf{Z}^T\mathbf{B}(\mathbf{y}^* - \boldsymbol{\delta}^*)\,, \end{aligned}$$

donde $\mathbf{y}^*$, $\boldsymbol{\mu}^*$ y $\boldsymbol{\delta}^*$ son vectores, tales que

$$\begin{aligned} y_i^* &= \frac{\delta_i}{[\delta_i\, y_i + y_i + \delta_i\mu_i]} + \frac{y_i\,[\delta_i + 1]}{4{\mu_i}^2} - \frac{{\delta_i}^2}{4y_i\,[\delta_i + 1]}, \quad \mu_i^* = \frac{1}{2\mu_i}, \quad a_i = \frac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i} = \frac{1}{\mathrm{d}g(\mu_i)/\mathrm{d}\mu_i} \\ y_i^* &= \frac{[y_i + \mu_i]}{[\delta_i\, y_i + y_i + \delta_i\mu_i]} - \frac{y_i}{4\mu_i} - \frac{\delta_i[\delta_i + 2]\mu_i}{4[\delta_i + 1]^2 y_i}, \quad \delta_i^* = -\frac{\delta_i}{2\,[\delta_i + 1]}, \quad b_i = \frac{\mathrm{d}\delta_i}{\mathrm{d}\tau_i} = \frac{1}{\mathrm{d}h(\delta_i)/\mathrm{d}\delta_i}, \end{aligned}$$

donde $\mathbf{A} = [a_i\delta_{ij}^n]$ y $\mathbf{B} = [b_i\delta_{ij}^n]$ son matrices diagonales $(n \times n)$, con δ_{ij}^n el delta de Kronecker.

3.3. Matriz hessiana penalizada

Sea $\ddot{\mathbf{L}}_{\mathrm{p}}(\boldsymbol{\theta})$ la matriz hessiana $(p^* \times p^*)$ con el (j^*, ℓ^*) elemento dado por $\partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda)/\partial\theta_{j^*}\theta_{\ell^*}$, para $j^*, \ell^* = 1, \ldots, p^*$ y $p^* = p + q + r$. Después de algunas manipulaciones algebraicas, se encuentra que la matriz hessiana penalizada esta dada por

$$\ddot{\mathbf{L}}_{\mathrm{p}}(\boldsymbol{\theta}) = \frac{\partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} = \begin{pmatrix} \ddot{\mathbf{L}}_{\mathrm{p}}^{\beta\beta}(\boldsymbol{\theta}) & \ddot{\mathbf{L}}_{\mathrm{p}}^{\beta\,\mathrm{f}}(\boldsymbol{\theta}) & \ddot{\mathbf{L}}_{\mathrm{p}}^{\beta\delta}(\boldsymbol{\theta}) \\ \ddot{\mathbf{L}}_{\mathrm{p}}^{\beta\,\mathrm{f}^T}(\boldsymbol{\theta}) & \ddot{\mathbf{L}}_{\mathrm{p}}^{\mathrm{f}\,\mathrm{f}}(\boldsymbol{\theta}) & \ddot{\mathbf{L}}_{\mathrm{p}}^{\mathrm{f}\,\delta}(\boldsymbol{\theta}) \\ \ddot{\mathbf{L}}_{\mathrm{p}}^{\beta\delta}(\boldsymbol{\theta}) & \ddot{\mathbf{L}}_{\mathrm{p}}^{\mathrm{f}\,\delta}(\boldsymbol{\theta}) & \ddot{\mathbf{L}}_{\mathrm{p}}^{\delta\delta}(\boldsymbol{\theta}) \end{pmatrix}, \tag{3.2}$$

cuyos elementos de la matriz pueden escribirse como:

$$\begin{aligned} \ddot{\mathbf{L}}_{\mathrm{p}}^{\beta\beta}(\boldsymbol{\theta}) &= \mathbf{X}^T\mathbf{C}\mathbf{X}, \\ \ddot{\mathbf{L}}_{\mathrm{p}}^{\beta\,\mathrm{f}}(\boldsymbol{\theta}) &= \mathbf{X}^T\mathbf{C}\mathbf{N}, \\ \ddot{\mathbf{L}}_{\mathrm{p}}^{\beta\delta}(\boldsymbol{\theta}) &= \mathbf{X}^T\mathbf{M}\mathbf{Z}, \\ \ddot{\mathbf{L}}_{\mathrm{p}}^{\mathrm{f}\,\mathrm{f}}(\boldsymbol{\theta}) &= \mathbf{N}^T\mathbf{C}\mathbf{N} - \lambda\mathbf{K}, \\ \ddot{\mathbf{L}}_{\mathrm{p}}^{\mathrm{f}\,\delta}(\boldsymbol{\theta}) &= \mathbf{N}^T\mathbf{M}\mathbf{Z} \qquad \text{y} \\ \ddot{\mathbf{L}}_{\mathrm{p}}^{\delta\delta}(\boldsymbol{\theta}) &= \mathbf{Z}^T\mathbf{W}\mathbf{Z}, \end{aligned}$$

donde $\mathbf{C} = [c_i \delta_{ij}^n]$, $\mathbf{M} = [m_i \delta_{ij}^n]$ y $\mathbf{W} = [w_i \delta_{ij}^n]$ son matrices diagonales $(n \times n)$, $c_i = d_{\mu^2}^{(i)}(a_i)^2 + d_{\mu}^{(i)} a_i^{'} a_i$, $m_i = d_{\mu\delta}^{(i)} a_i b_i$ y $w_i = d_{\delta^2}^{(i)}(b_i^2) + d_{\delta}^{(i)} b_i^{'} b_i$, con

$$
\begin{aligned}
d_{\mu}^{(i)} &= \frac{\partial L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda)}{\partial \mu_i} = \frac{\delta_i}{[\delta_i\, y_i + y_i + \delta_i \mu_i]} + \frac{y_i\,[\delta_i + 1]}{4{\mu_i}^2} - \frac{{\delta_i}^2}{4 y_i\,[\delta_i + 1]} - \frac{1}{2\mu_i} = y_i^* - \mu_i^*, \\
d_{\mu^2}^{(i)} &= \frac{\partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda)}{\partial \mu_i^2} = \frac{1}{2\mu_i^2} - \frac{\delta_i^2}{[\delta_i\, y_i + y_i + \delta_i \mu_i]^2} - \frac{y_i[\delta_i + 1]}{2\mu_i^3}, \qquad a_i^{'} = -\frac{\mathrm{d}^2 g(\mu_i)/d\mu_i^2}{[\mathrm{d}g(\mu_i)/d\mu_i]^2}, \\
d_{\mu\delta}^{(i)} &= \frac{\partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda)}{\partial \mu_i \partial \delta_i} = \frac{y_i}{[\delta_i\, y_i + y_i + \delta_i \mu_i]^2} + \frac{y_i}{4\mu_i^2} - \frac{\delta_i[\delta_i + 2]}{4[\delta_i + 1]^2 y_i}, \\
d_{\delta_i}^{(i)} &= \frac{L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda)}{\partial \delta_i} = \frac{[y_i + \mu_i]}{[\delta_i\, y_i + y_i + \delta_i \mu_i]} - \frac{y_i}{4\mu_i} - \frac{\delta_i[\delta_i + 2]\mu_i}{4[\delta_i + 1]^2 y_i} + \frac{\delta_i}{2[\delta_i + 1]} = y_i^* - \delta_i^*, \\
d_{\delta^2}^{(i)} &= \frac{\partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda)}{\partial \delta_i^2} = \frac{1}{2[\delta_i + 1]^2} - \frac{[y_i + \mu_i]^2}{[\delta_i y_i + y_i + \delta_i \mu_i]^2} - \frac{\mu_i}{2[\delta_i + 1]^3 y_i}, \qquad b_i^{'} = -\frac{\mathrm{d}^2 h(\delta_i)/\mathrm{d}\delta_i^2}{[\mathrm{d}h(\delta_i)/\mathrm{d}\delta_i]^2}.
\end{aligned}
$$

3.4. Matriz de información de Fisher penalizada

Por otro lado, al calcular la esperanza de la matriz $-\ddot{\mathbf{L}}_{\mathrm{p}}(\boldsymbol{\theta})$, se obtiene la matriz de información penalizada $(p^* \times p^*)$ dada por

$$
\boldsymbol{\mathfrak{I}}_{\mathrm{p}}(\boldsymbol{\theta}) = \begin{pmatrix} \mathfrak{I}_{\mathrm{p}}^{\beta\beta}(\boldsymbol{\theta}) & \mathfrak{I}_{\mathrm{p}}^{\beta\,\mathrm{f}}(\boldsymbol{\theta}) & \mathfrak{I}_{\mathrm{p}}^{\beta\delta}(\boldsymbol{\theta}) \\ \mathfrak{I}_{\mathrm{p}}^{\beta\,\mathrm{f}^T}(\boldsymbol{\theta}) & \mathfrak{I}_{\mathrm{p}}^{\mathrm{f}\,\mathrm{f}}(\boldsymbol{\theta}) & \mathfrak{I}_{\mathrm{p}}^{\mathrm{f}\,\delta}(\boldsymbol{\theta}) \\ \mathfrak{I}_{\mathrm{p}}^{\beta\delta^T}(\boldsymbol{\theta}) & \mathfrak{I}_{\mathrm{p}}^{\mathrm{f}\,\delta^T}(\boldsymbol{\theta}) & \mathfrak{I}_{\mathrm{p}}^{\delta\delta}(\boldsymbol{\theta}) \end{pmatrix}, \tag{3.3}
$$

donde cada elemento de la matriz se puede escribir como

$$
\begin{aligned}
\mathfrak{I}_{\mathrm{p}}^{\beta\beta} &= \mathbf{X}^T \mathbf{V} \mathbf{X}, \\
\mathfrak{I}_{\mathrm{p}}^{\beta\,\mathrm{f}} &= \mathbf{X}^T \mathbf{V} \mathbf{N}, \\
\mathfrak{I}_{\mathrm{p}}^{\beta\delta} &= \mathbf{X}^T \mathbf{S}\, \mathbf{Z}, \\
\mathfrak{I}_{\mathrm{p}}^{\mathrm{f}\,\mathrm{f}} &= \mathbf{N}^T \mathbf{V} \mathbf{N} + \lambda \mathbf{K}, \\
\mathfrak{I}_{\mathrm{p}}^{\mathrm{f}\,\delta} &= \mathbf{N}^T \mathbf{S}\, \mathbf{Z} \quad \text{y} \\
\mathfrak{I}_{\mathrm{p}}^{\delta\delta} &= \mathbf{Z}^T \mathbf{U} \mathbf{Z},
\end{aligned}
$$

donde $\mathbf{V} = [v_i]$, $\mathbf{U} = [u_i]$ y $\mathbf{S} = [s_i]$ son matrices diagonales $(n \times n)$ con

$$
\begin{aligned}
v_i &= \frac{\delta_i a_i^2}{2\mu_i^2} + \frac{\delta_i^2 a_i^2}{[\delta_i + 1]^2} \mathcal{J}(\boldsymbol{\theta}), \\
s_i &= \left[\frac{1}{2\mu_i[\delta_i + 1]} + \frac{\delta_i \mu_i}{[\delta_i + 1]^3} \mathcal{J}(\boldsymbol{\theta}) \right] a_i b_i \qquad \text{y} \\
u_i &= \left[\frac{[\delta_i^2 + 3\delta_i + 1]}{2\delta_i^2 [\delta_i + 1]^2} + \frac{\mu_i^2}{[\delta_i + 1]^4} \mathcal{J}(\boldsymbol{\theta}) \right] b_i^2,
\end{aligned}
$$

donde

$$\begin{aligned}\jmath(\boldsymbol{\theta}) &= \mathrm{E}\left[\left\{Y+\frac{\mu\delta_i}{(\delta_i+1)}\right\}^{-2}\right] \\ &= \int_0^\infty \frac{\sqrt{\delta_i+1}\exp(\delta_i/2)}{4\sqrt{\pi\mu}y^{3/2}}\left[y+\frac{\delta_i\mu}{\delta_i+1}\right]^{-2}\exp\left(-\frac{\delta_i}{4}\left[\frac{(\delta_i+1)y}{\delta_i\mu}+\frac{\delta_i\mu}{(\delta_i+1)y}\right]\right)\mathrm{d}y\,.\end{aligned}$$

3.5. Encontrando la solución en la práctica: proceso iterativo

Para estimar los parámetros del modelo por el método de máxima verosimilitud penalizada, se resuelve la ecuación $\mathbf{U}_{\mathrm{p}}(\boldsymbol{\theta}) = \mathbf{0}$. Sin embargo, no se dispone de expresiones de forma cerrada para los estimadores de máxima verosimilitud penalizada. Luego, se necesita un método iterativo para la optimización no lineal, como los algoritmos de Scoring de Fisher, Newton-Raphson o cuasi-Newton. Considerando que la matriz $-\ddot{\mathbf{L}}_{\mathrm{p}}(\boldsymbol{\theta})$ puede ser definida no-positiva, se sugiere sustituirla por la matriz $-\mathfrak{I}_{\mathrm{p}}(\boldsymbol{\theta})$ y utilizar el método de Scoring de Fisher. Entonces, el algoritmo para estimar $\boldsymbol{\theta}$ viene dado por

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + (\mathfrak{I}_p(\boldsymbol{\theta})^{-1})^{(m)}\mathbf{U}_{\mathrm{p}}(\boldsymbol{\theta})^{(m)}. \tag{3.4}$$

La ecuación anterior es equivalente a

$$\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)} = (\mathfrak{I}_p(\boldsymbol{\theta})^{-1})^{(m)}\mathbf{U}_{\mathrm{p}}(\boldsymbol{\theta})^{(m)},$$

lo que es igual a resolver la ecuación matricial

$$\begin{pmatrix}\mathbf{X}^T\mathbf{V}\mathbf{X} & \mathbf{X}^T\mathbf{V}\mathbf{N} & \mathbf{X}^T\mathbf{S}\mathbf{Z}\\ \mathbf{N}^T\mathbf{V}\mathbf{X} & \mathbf{N}^T\mathbf{V}\mathbf{N}+\lambda\mathbf{K} & \mathbf{N}^T\mathbf{S}\mathbf{Z}\\ \mathbf{Z}^T\mathbf{S}\mathbf{X} & \mathbf{Z}^T\mathbf{N}\mathbf{S} & \mathbf{Z}^T\mathbf{U}\mathbf{Z}\end{pmatrix}^{(m)}\begin{pmatrix}\boldsymbol{\Delta}_\beta^{(m+1,m)}\\ \boldsymbol{\Delta}_f^{(m+1,m)}\\ \boldsymbol{\Delta}_\alpha^{(m+1,m)}\end{pmatrix} = \begin{pmatrix}\mathbf{X}^T\mathbf{A}(\mathbf{y}^*-\boldsymbol{\mu}^*)\\ \mathbf{N}^T\mathbf{A}(\mathbf{y}^*-\boldsymbol{\mu}^*)-\lambda\mathbf{K}\mathbf{f}\\ \mathbf{Z}^T\mathbf{B}(\mathbf{y}^*-\boldsymbol{\alpha}^*)\end{pmatrix} \tag{3.5}$$

donde $\boldsymbol{\Delta}_\beta^{(m+1,m)} = \boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}$, $\boldsymbol{\Delta}_f^{(m+1,m)} = \mathbf{f}^{(m+1)} - \mathbf{f}^{(m)}$ y $\boldsymbol{\Delta}_\alpha^{(m+1,m)} = \boldsymbol{\alpha}^{(m+1)} - \boldsymbol{\alpha}^{(m)}$. Luego de algunas manipulaciones algebraicas, se obtienen las siguientes expresiones para las soluciones iterativas;

$$\begin{aligned}&\boldsymbol{a)}\ \mathbf{X}^T\ \mathbf{V}^{(m)}\ \mathbf{X}\ \boldsymbol{\Delta}_\beta^{(m+1,m)} + \mathbf{X}^T\ \mathbf{V}^{(m)}\mathbf{N}\boldsymbol{\Delta}_{\mathbf{f}}^{(m+1,m)} + \mathbf{X}^T\mathbf{S}\mathbf{Z}\boldsymbol{\Delta}_\alpha^{(m+1,m)} = \mathbf{X}^T\mathbf{A}^{(m)}\ (\mathbf{y}^*-\boldsymbol{\mu}^*)\\ &= \mathbf{X}^T\ \mathbf{V}^{(m)}\mathbf{X}\ [\boldsymbol{\beta}^{(m+1)}-\boldsymbol{\beta}^{(m)}] + \mathbf{X}^T\ \mathbf{V}^{(m)}\ \mathbf{N}\ [\mathbf{f}^{(m+1)}-\mathbf{f}^{(m)}] + \mathbf{X}^T\ \mathbf{S}^{(m)}\ \mathbf{Z}\ [\boldsymbol{\alpha}^{(m+1)} - \boldsymbol{\alpha}^{(m)}]\\ &= \mathbf{X}^T\ \mathbf{V}^{(m)}\ [\ \mathbf{D}_{v,a}^{(m)}\ (\ \mathbf{y}^* - \boldsymbol{\mu}^*)^{(m)} + \boldsymbol{\eta}^{(m)} + \mathbf{D}_{v,s}^{(m)}\ \boldsymbol{\tau}^{(m)} - \mathbf{N}\ \mathbf{f}^{(m+1)} - \mathbf{D}_{v,s}^{(m)}\ \tau^{(m+1)}\ ],\end{aligned}$$

donde

$\mathbf{D}_{v,a}^{(m)} = (\mathbf{V}^{(m)})^{-1}\mathbf{A}^{(m)}, \quad \mathbf{D}_{v,s}^{(m)} = (\mathbf{V}^{(m)})^{-1}\ \mathbf{S}^{(m)}, \quad \boldsymbol{\eta}^{(m)} = \mathbf{X}\ \boldsymbol{\beta}^{(m)} + \mathbf{N}\mathbf{f}^{(m)}$ y $\quad \boldsymbol{\tau}^{(m)} = \ \mathbf{Z}\ \alpha^{(m+1)}$.

Luego, $\boldsymbol{\beta}^{(m+1)}$ esta dado por

$$\boldsymbol{\beta}^{(m+1)} = (\mathbf{X}^T\mathbf{V}^{(m)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{(m)}[\mathbf{D}_{v,a}^{(m)}(\mathbf{y}^* - \boldsymbol{\mu}^*)^{(m)} + \boldsymbol{\eta}^{(m)} + \mathbf{D}_{v,s}^{(m)}\boldsymbol{\tau}^{(m)} - \mathbf{N}\mathbf{f}^{(m+1)} - \mathbf{D}_{v,s}^{(m)}\boldsymbol{\tau}^{(m+1)}].$$

b) $\mathbf{X}^T\mathbf{V}^{(m)}\ \mathbf{X}\boldsymbol{\Delta}_{\beta}^{(m+1,m)} + (\mathbf{N}^T\mathbf{V}^{(m)}\mathbf{N} + \lambda\mathbf{K})\boldsymbol{\Delta}_{\mathbf{f}}^{(m+1,m)} + \ \mathbf{N}^T\mathbf{S}^{(m)}\mathbf{Z}\ \boldsymbol{\Delta}_{\alpha}^{(m+1,m)} = \mathbf{Z}^T\mathbf{B}(\mathbf{y}^* - \boldsymbol{\mu}^*)$

$= \mathbf{N}^T\mathbf{V}^{(m)}\mathbf{X}\ [\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}]\ + (\mathbf{N}^T\ \mathbf{V}^{(m)} + \lambda\mathbf{K})\ [\mathbf{f}^{(m+1)} - \mathbf{f}^{(m)}] + \mathbf{N}^T\mathbf{S}^{(m)}\mathbf{Z}[\boldsymbol{\alpha}^{(m+1)} - \boldsymbol{\alpha}^{(m)}]$

$= \mathbf{N}^T\ \mathbf{V}^{(m)}\ [\ \mathbf{D}_{v,a}^{(m)}\ (\ \mathbf{y}^* - \ \boldsymbol{\mu}^*)^{(m)}\ +\ \boldsymbol{\eta}^{(m)} + \ \mathbf{D}_{v,s}^{(m)}\ \boldsymbol{\tau}^{(m)} - \mathbf{X}\ \boldsymbol{\beta}^{(m+1)}\ -\ \mathbf{D}_{v,s}^{(m)}\ \boldsymbol{\tau}^{(m+1)}\ ].$

.

Así,

$$\mathbf{f}^{(m+1)} = (\mathbf{N}^T\mathbf{V}^{(m)}\mathbf{N}+\lambda\mathbf{K})^{-1}\mathbf{N}^T\mathbf{V}^{(m)}[\mathbf{D}_{v,a}^{(m)}(\mathbf{y}^*-\boldsymbol{\mu}^*)^{(m)}+\boldsymbol{\eta}^{(m)}+\mathbf{D}_{v,s}^{(m)}\boldsymbol{\tau}^{(m)}-\mathbf{X}\boldsymbol{\beta}^{(m+1)}-\mathbf{D}_{v,s}^{(m)}\boldsymbol{\tau}^{(m+1)}].$$

c) $\mathbf{Z}^T\ \mathbf{S}^{(m)}\ \ \mathbf{X}\ \ \boldsymbol{\Delta}_{\beta}^{(m+1,m)}\ +\ \mathbf{Z}^T\ \ \mathbf{S}^{(m)}\mathbf{N}\boldsymbol{\Delta}_{\mathbf{f}}^{(m+1,m)}\ + \mathbf{Z}^T\ \mathbf{U}\ \mathbf{Z}\boldsymbol{\Delta}_{\alpha}^{(m+1,m)} = \mathbf{Z}^T\mathbf{B}^{(m)}(\mathbf{y}^* - \boldsymbol{\mu}^*)$

$= \mathbf{Z}^T\ \mathbf{S}^{(m)}\mathbf{X}\ [\boldsymbol{\beta}^{(m+1)} - \ \boldsymbol{\beta}^{(m)}] + \ \mathbf{Z}^T\mathbf{S}^{(m)}\ \mathbf{N}\ [\ \mathbf{f}^{(m+1)} - \mathbf{f}^{(m)}] + \ \mathbf{Z}^T\ \mathbf{U}^{(m)}\ \mathbf{Z}\ [\ \boldsymbol{\alpha}^{(m+1)} - \boldsymbol{\alpha}^{(m)}]$

$$= \mathbf{Z}^T\mathbf{U}^{(m)}[\mathbf{D}_{v,b}^{(m)}(\mathbf{y}^* - \boldsymbol{\mu}^*) + \mathbf{D}_{v,s}^{(m)}\mathbf{X}\boldsymbol{\beta}^{(m)} + \mathbf{D}_{v,s}^{(m)}\mathbf{N}\mathbf{f}^{(m)} - \mathbf{D}_{v,s}^{(m)}\mathbf{X}\boldsymbol{\beta}^{(m+1)} - \mathbf{D}_{v,s}^{(m)}\mathbf{N}\mathbf{f}^{(m+1)} + \mathbf{Z}\boldsymbol{\alpha}^{(m)}].$$

Así,

$$\boldsymbol{\alpha}^{(m+1)} = (\ \mathbf{Z}^T\mathbf{U}^{(m)}\mathbf{Z})^{-1}\ \mathbf{Z}^T\ \mathbf{U}^{(m)}\ [\ \mathbf{D}_{v,b}^{(m)}\ (\mathbf{y}^* - \boldsymbol{\mu}^*)\ +\ \mathbf{D}_{v,s}^{(m)}\ \boldsymbol{\eta}^{(m)} - \mathbf{D}_{v,s}^{(m)}\ \boldsymbol{\eta}^{(m+1)} + \ \boldsymbol{\tau}^{(m)}],$$

donde

$\mathbf{D}_{v,b}^{(m)} = (\mathbf{U}^{(m)})^{-1}\mathbf{B}^{(m)}, \quad \mathbf{D}_{v,s}^{(m)} = (\mathbf{V}^{(m)})^{-1}\mathbf{S}^{(m)}, \quad \boldsymbol{\eta}^{(m)} = \mathbf{X}\boldsymbol{\beta}^{(m)} + \mathbf{N}\mathbf{f}^{(m)}$ y $\quad \boldsymbol{\eta}^{(m+1)} = \mathbf{X}\boldsymbol{\beta}^{(m+1)} + \mathbf{N}\mathbf{f}^{(m+1)}$.

Capítulo 4

Aspectos Inferenciales

En este capítulo se estudian algunos aspectos inferenciales asociados al modelo semiparamétrico BSR con parámetro de precisión variando. En la Sección 4.1 se realiza una breve revisión de trabajos que abordan aspectos inferenciales en el modelo de regresión Birnbaum-Saunders reparametrizado. En la Sección 4.2 se deriva la matriz de varianza-covarianza de la estimación de máxima verosimilitud penalizada a partir de la matriz inversa de información de Fisher. En la Sección 4.3, se presenta una breve discusión sobre cómo seleccionar los grados de libertad (g.l) asociados con el componente no parámetrico del modelo. Finalmente, en la Sección 4.4 se presentan algunos métodos para seleccionar parámetros de suavizamiento.

4.1. Introducción

Los aspectos inferenciales son un conjunto de métodos y técnicas que permiten inducir, a partir de la información empírica proporcionada por un conjunto de datos. Cabe destacar que el modelo semiparamétrico BSR con parámetro de precisión variando surge como una alternativa al modelo de regresión BSR ya que permite modelar tendencias lineales y no lineales en su componente sistemático. Este modelo tiene una serie de aplicaciones potenciales, su análisis inferencial es bastante limitado (si es que no nulo). Sin embargo, algunos autores han abordado este tema para algunos modelos relacionados. Por ejemplo, Santos–Neto *et al*. (2016) proponen un modelo de regresión BSR con precisión variando, que permite describir la heteroscedasticidad en el cual introducieron pruebas de hipótesis para el parámetro de precisión y evaluaron su rendimiento. Algo similar a lo que realizo Lombardia & Sperlich (2008) quienes desarrollaron una prueba para la hipótesis de un modelo paramétrico de efectos mixtos versus un modelo semiparamétrico de efectos mixtos. Segal *et al*. (1994) derivan la varianza del estimador de máxima verosimilitud penalizado utilizando el algoritmo de *Expectation-Maximization (EM)*; véase también Green & Silverman (1990). Fan & Jiang (2005) ampliaron la prueba de la relación de probabilidad generalizada para modelos aditivos con el fin de verificar si admite una forma paramétrica. Lin & Zhang (1999) estudiaron algunos aspectos de la estimación y la inferencia para modelos generalizados aditivos mixtos basados en el criterio de la función de cuasi-verosimilitud penalizada. Se notificaron comparaciones entre inferencia frecuentista y bayesiana. Berhane & Tibshirani (1998) propusieron bandas de error estándar y algunas pruebas aproximadas, como la prueba de cociente de probabilidad y la prueba de score, para modelos aditivos generalizados para datos longitudinales. Importante mencionar además que Santos-Neto *et al*. (2014) proponen la estimación e inferencia de la distribución Birnbaum-Saunders reparametrizada basada en métodos de máxima verosimilitud, de momentos, de momentos modificados y de momentos generalizados. Ibacache-Pulgar & Paula (2011) aproximaron la matriz de varianza-covarianza en el modelo lineal parcial de t-Student. Ibacache-Pulgar & Reyes (2018)

propusieron aproximar la matriz de varianza-covarianza en un modelo aditivo semiparamétrico bajo distribuciones simétricas utilizando la matriz de información de Fisher obtenida a partir de la función de función de verosimilitud penalizada.

4.2. Errores estándar aproximados

En esta sección se considera el problema de estimar la matriz de varianza-covarianza del estimador de máxima verosimilitud penalizada $\widehat{\boldsymbol{\beta}}$. Teniendo en cuenta el hecho de que se obtuvo el estimador de máxima verosimilitud penalizada de $\boldsymbol{\theta}$ a través del algoritmo Scoring de Fisher, es razonable derivar la matriz de varianza-covarianza utilizando la inversa de la matriz de información de Fisher penalizada; véase, por ejemplo, Segal *et al* (1994), Wahba (1983) e Ibacache-Pulgar *et al*. (2013). Con el propósito de calcular la matriz inversa de $\mathfrak{I}_{\mathrm{p}}(\boldsymbol{\theta})$ dada en (3.3), considere que

$$\mathfrak{I}_{\mathrm{p}}^{11} = \begin{pmatrix} \mathfrak{I}_{\mathrm{p}}^{\beta\beta}(\boldsymbol{\theta}) & \mathfrak{I}_{\mathrm{p}}^{\beta\,\mathrm{f}}(\boldsymbol{\theta}) \\ \mathfrak{I}_{\mathrm{p}}^{\beta\,\mathrm{f}^T}(\boldsymbol{\theta}) & \mathfrak{I}_{\mathrm{p}}^{\mathrm{f}\,\mathrm{f}}(\boldsymbol{\theta}) \end{pmatrix}, \qquad \mathfrak{I}_{\mathrm{p}}^{12} = \begin{pmatrix} \mathfrak{I}_{\mathrm{p}}^{\beta\delta}(\boldsymbol{\theta}) \\ \mathfrak{I}_{\mathrm{p}}^{f\delta}(\boldsymbol{\theta}) \end{pmatrix} \quad y \quad \mathfrak{I}_{\mathrm{p}}^{22} = \mathfrak{I}_{\mathrm{p}}^{\delta\delta}.$$

Así, la matriz $\mathfrak{I}_{\mathrm{p}}(\boldsymbol{\theta})$ puede escribirse como

$$\mathfrak{I}_{\mathrm{p}}(\boldsymbol{\theta}) = \begin{pmatrix} \mathfrak{I}_{\mathrm{p}}^{11} & \mathfrak{I}_{\mathrm{p}}^{12} \\ \mathfrak{I}_{\mathrm{p}}^{12^T} & \mathfrak{I}_{\mathrm{p}}^{22} \end{pmatrix}. \tag{4.1}$$

Suponiendo que existen todas las inversas necesarias y después de haber aplicado algunas manipulaciones algebraicas sobre la expresión (4.1) se muestra que la matriz inversa de $\mathfrak{I}_{\mathrm{p}}(\boldsymbol{\theta})$ asume la siguiente forma de bloque:

$$\mathfrak{I}_{\mathrm{p}}^{-1}(\boldsymbol{\theta}) = \begin{pmatrix} \mathfrak{I}_{\mathrm{p}}^{11,1} & -\mathfrak{I}_{\mathrm{p}}^{11,1}\mathfrak{I}_{\mathrm{p}}^{12}\mathfrak{I}_{\mathrm{p}}^{22^{-1}} \\ -\mathfrak{I}_{\mathrm{p}}^{22^{-1}}\mathfrak{I}_{\mathrm{p}}^{12^T}\mathfrak{I}_{\mathrm{p}}^{11,1} & \mathfrak{I}_{\mathrm{p}}^{22,1} \end{pmatrix}, \tag{4.2}$$

donde $\mathfrak{I}_{\mathrm{p}}^{11,1} = (\mathfrak{I}_{\mathrm{p}}^{11} - \mathfrak{I}_{\mathrm{p}}^{12}\mathfrak{I}_{\mathrm{p}}^{22^{-1}}\mathfrak{I}_{\mathrm{p}}^{12^T})^{-1}$ y $\mathfrak{I}_{\mathrm{p}}^{22,1} = \mathfrak{I}_{\mathrm{p}}^{22^{-1}} + \mathfrak{I}_{\mathrm{p}}^{22^{-1}}\mathfrak{I}_{\mathrm{p}}^{12^T}\mathfrak{I}_{\mathrm{p}}^{11,1}\mathfrak{I}_{\mathrm{p}}^{12}\mathfrak{I}_{\mathrm{p}}^{22^{-1}}$. Por lo tanto, la matriz de varianza-covarianza asintótica de $\widehat{\boldsymbol{\theta}}$ está dado por

$$\widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\theta}})_{\mathrm{aprox}} = \mathfrak{I}_{\mathrm{p}}^{-1}(\boldsymbol{\theta})\Big|_{\widehat{\boldsymbol{\theta}}}. \tag{4.3}$$

En particular, se tiene

$$\widehat{\mathrm{Cov}}_{\mathrm{aprox}}(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{f}}) = \mathfrak{I}_{\mathrm{p}}^{11,1}\Big|_{\widehat{\boldsymbol{\theta}}} \qquad \mathrm{y} \qquad \widehat{\mathrm{Cov}}_{\mathrm{aprox}}(\widehat{\boldsymbol{\alpha}}) = \mathfrak{I}_{\mathrm{p}}^{22,1}\Big|_{\widehat{\boldsymbol{\theta}}}.$$

4.3. Grados de Libertad

En esta sección se presenta una definición de los grados de libertad asociados a las componentes paramétricas y no paramétricas del modelo basada en la convergencia del proceso iterativo dado en la Sección (3.5) para la estimación de **f**. Considerando un valor de λ fijo, se obtiene

$$\widehat{\mathbf{f}} \;=\; \widehat{\mathbf{S}}\widehat{\mathbf{r}^*}_{v,a},$$

donde $\widehat{\mathbf{r}^*}_{v,a} = \widehat{\mathbf{r}}_{a,b} - \mathbf{X}\widehat{\boldsymbol{\beta}}$, con $\widehat{\mathbf{r}}_{a,b} = \widehat{\mathbf{D}}_{v,n}\widehat{\mathbf{z}} + \widehat{\boldsymbol{\eta}}$, $\widehat{\boldsymbol{\eta}} = \mathbf{N}\widehat{\mathbf{f}}$ y

$$\widehat{\mathbf{S}} \;=\; (\mathbf{N}^T\widehat{\mathbf{D}}_v\mathbf{N})^{-1}\mathbf{N}\widehat{\mathbf{D}}_v.$$

En la literatura existen diferentes efectos para los grados de libertad, dependiendo del contexto en el que se utilicen; ver Buja *et al*. (1989). Aquí los grados de libertad asociados con el componente paramétrico $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ se definen como

$$\mathrm{gl}_x = \mathrm{tr}\{\mathbf{X}(\mathbf{X}^T\widehat{\mathbf{D}}_v\mathbf{X})^{-1}\mathbf{X}^T\widehat{\mathbf{D}}_v\} = p,$$

donde p es el rango de $\mathbf{X}$. Por otro lado, los grados de libertad para la componente no paramétrica $\mathbf{N}\widehat{\mathbf{f}}$ se definen como

$$\mathrm{gl}(\lambda) = \mathrm{tr}\{\mathbf{N}(\mathbf{N}^T\widehat{\mathbf{D}}_v\mathbf{N} + \lambda\mathbf{K})^{-1}\mathbf{N}^T\widehat{\mathbf{D}}_v\}, \tag{4.4}$$

que mide la contribución del efecto no paramétrico sobre el modelo.

4.4. Sobre el parámetro de suavizamiento

En la sección anterior se considero el parámetro λ fijo. Sin embargo, en la práctica los parámetros de suavizamiento deben seleccionarse a partir de los datos. A continuación se describe un criterio para seleccionar los parámetros de suavizamiento basado en el Criterio de Información de Akaike (AIC).

Criterio AIC

El AIC o el criterio de información bayesiana (BIC) se pueden utilizar para seleccionar el parámetro de suavizamiento λ. La idea es minimizar la siguiente función con respecto a λ

$$\mathrm{AIC}(\lambda) = -2L_p(\widehat{\boldsymbol{\theta}}, \lambda) + 2[1 + p + \mathrm{df}(\lambda)],$$

donde $L_p(\widehat{\boldsymbol{\theta}}, \lambda)$ denota la función de probabilidad logarítmica penalizada encontrada en $\widehat{\boldsymbol{\theta}}$ para un parámetro fijo λ. Una cuadrícula (superficie) para diferentes valores de λ y su correspondiente $\mathrm{AIC}(\lambda)$ es útil para elegir el parámetro de suavizamiento óptimo.

Fijación de los grados de libertad

Otra forma de seleccionar los parámetros de suavizamiento es cuando los grados de libertad de la ecuación (4.4) dependen solo de λ y por lo tanto, se puede especificar el parámetro de suavizamiento correspondiente. En otras palabras, se especifica $\text{gl}(\lambda)$ para una función y luego se establece el valor de λ que cumple con este objetivo. Este enfoque es utilizado por varios autores, entre ellos Buja *et al*. (1989) y Rigby & Stasinopoulos (2005).

Capítulo 5

Análisis de diagnóstico

En este capítulo se considera la extensión y aplicación de la técnica de influencia local al modelo semiparamétrico BSR con parámetro de precisión variando. En la Sección 5.1, se realizá una breve revisión bibliográfica de los principales trabajos relativos a la técnica de influencia local. En la Sección 5.2, se proponen dos tipos de residuos basados en el trabajo desarrollado por Leiva *et al*. (2014). En la Sección 5.3, se presenta una descripción general del método de influencia local. Finalmente, en la Sección 5.4 se deriva la curvatura normal para dos esquemas de perturbación, en concreto, la perturbación del peso del caso y la variable de respuesta.

5.1. Introducción

Es sabido que el análisis de diagnóstico es un proceso fundamental en la modelización estadística. Entre las técnicas más utilizadas en regresión paramétrica se encuentran la influencia global e influencia local, este último propuesto por Cook (1986) con el propósito de evaluar la sensibilidad de los estimadores de los parámetros cuando se introducen pequeñas perturbaciones en los supuestos del modelo semiparamétrico BSR con parámetro de precisión variando o también en los datos. En esta sección se considera la extensión y aplicación del método de influencia local en el modelo propuesto, existen trabajos en los cuales se aplica esta metodología, entre ellos Zhu *et al*. (2003) proporcionan un medio conveniente para extender el análisis de influencia local de Cook al EMVP en los modelos normales, mientras que Ibacache-Pulgar (2011) lo hace para modelos t de Student. Chen *et al*. (2010) propusieron un procedimiento para seleccionar el esquema de perturbación apropiado cuando el método de influencia local se aplica a modelos lineales mixtos generalizados. Recientemente Ibacache-Pulgar *et al*. (2021) desarrollaron el método de influencia local para modelos semiparamétricos de regresión beta aditiva y Cárcamo *et al*. (2021) para el modelo BSR-SAM. Ibacache-Pulgar *et al*. (2012) lo hace con modelos t-Student parcialmente lineales. Ferreira & Paula (2016) extendieron la técnica de influencia local para diferentes esquemas de perturbación considerando un modelo parcialmente lineal sesgado-normal. De Bastiani *et al*. (2014) utilizan la metodología de influencia local para evaluar la sensibilidad de los estimadores de máxima verosimilitud a pequeñas perturbaciones en un conjunto de datos reales y/o en los supuestos del modelo lineal espacial.

5.2. Análisis residual

El análisis residual es un método matemático para verificar si un modelo de regresión se ajusta bien, estudiando la parte de los datos que no es explicada por el modelo. Con el fin de detectar las especificaciones erróneas de la distribución, así como la presencia de observaciones observaciones periféricas, proponemos dos tipos de residuos basados en los resultados presentados en Leiva *et al*. (2014). Estos residuos se describen a continuación.

Residuos estandarizados

Introducimos los residuos del modelo con la función de enlace definida en (2.2) a partir de en los residuos estandarizados de Pearson, este tipo de residuo se basa en el $y_i - \mu_i$ y se denota como sigue

$$r_i^s = \frac{y_i - \mu_i}{\sqrt{\widehat{\mathrm{Var}}(Y_i)}} = \frac{\widehat{\phi}_i^{1/2}[y_i - \widehat{\mu}_i]}{\sqrt{\widehat{\mu}_{\mathrm{i}}}} \qquad (i = 0, \cdots, n),$$

donde $\widehat{\phi}_i = [\widehat{\delta}_i + 1]^2/[2\widehat{\delta}_i + 5]$, con $\widehat{\mu}_i = g^{-1}(\mathbf{x}_i^T\widehat{\boldsymbol{\beta}} + \widehat{f(\mathrm{t}_i)})$ y $\widehat{\delta}_i = h^{-1}(\mathbf{z}_i^T\widehat{\boldsymbol{\alpha}})$ las estimaciones de máxima verosimilitud de μ_i y δ_i, respectivamente.

Residuos de Jørgensen

Otro tipo de residuo que se puede considerar se basa en el trabajo desarrollado por Jørgensen (1984) y extendido a los modelos de regresión BSR por Leiva *et al*. (2014). En el marco del modelo semiparamétrico BSR con parámetro de precisión variando, se propone

$$r_i^J = \mathrm{J}_i(\widehat{\mu}_i)^{1/2}\kappa_i\widehat{\mu}_i \qquad (i = 0, \ldots, n)$$

donde

$$\kappa_i\widehat{\mu}_i = -\frac{1}{2\widehat{\mu}_i} + \frac{\widehat{\delta}_i}{[\widehat{\delta}_i y_i + y_i + \widehat{\delta}_i\widehat{\mu}_i]} + \frac{y_i[\widehat{\delta}_i + 1]}{4\widehat{\mu}_i^2} - \frac{\widehat{\delta}_i^{\,2}}{4y_i[\widehat{\delta}_i + 1]}$$

y

$$\mathrm{J}_i(\widehat{\mu}) = -\frac{1}{2\widehat{\mu}_i^2} + \frac{\widehat{\delta}_i^{\,2}}{[\widehat{\delta}_i y_i + y_i + \widehat{\delta}_i\widehat{\mu}_i]^2} + \frac{[\widehat{\delta}_i + 1]y_i}{2\mu_i^2},$$

con $\kappa_i\widehat{\mu}_i$ siendo el i-ésimo elemento del vector $\boldsymbol{\kappa}(\boldsymbol{\mu}) = \partial L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda)/\partial(\boldsymbol{\mu})$ y $J_i(\widehat{\mu}_i)$ el i-ésimo elemento diagonal de $\boldsymbol{J}(\boldsymbol{\mu}) = \partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda)/\partial\boldsymbol{\mu}^\top\boldsymbol{\mu}$ evaluado en $\widehat{\boldsymbol{\theta}}$.

5.3. Método de Influencia Local

Para el modelo dado en (2.2) se tiene que $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^\top$ es un vector n-dimensional de perturbaciones restringido a algún subconjunto abierto $\Omega \in \mathbb{R}$ y el logaritmo de la probabilidad penalizada perturbada denotada por $L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda|\boldsymbol{\omega})$. Suponiendo que existe $\boldsymbol{\omega}_0 \in \Omega$, un vector de no perturbación, tal que

$L_{\rm p}(\boldsymbol{\theta}, \lambda|\boldsymbol{\omega}_0) = L_{\rm p}(\boldsymbol{\theta}, \lambda)$. Para evaluar la influencia de las perturbaciones menores en el estimador de máxima verosimilitud penalizada $\widehat{\boldsymbol{\theta}}$, podemos considerar el desplazamiento de verosimilitud

$$\mathrm{LD}(\boldsymbol{\omega}) = 2\Big[L_{\rm p}(\widehat{\boldsymbol{\theta}}, \lambda) - L_{\rm p}(\widehat{\boldsymbol{\theta}_\omega}, \lambda)\Big] \geq 0,$$

donde $\widehat{\boldsymbol{\theta}_\omega}$ es el estimador de máxima verosimilitud penalizada bajo $L_{\rm p}(\boldsymbol{\theta}, \lambda|\boldsymbol{\omega})$. La medida $\mathrm{LD}(\boldsymbol{\omega})$ es útil para evaluar la distancia entre $\widehat{\boldsymbol{\theta}}$ y $\widehat{\boldsymbol{\theta}_\omega}$. Cook (1986) propone estudiar el comportamiento local de $\mathrm{LD}(\boldsymbol{\omega})$ alrededor de $\boldsymbol{\omega}_0$. El procedimiento consiste en seleccionar una dirección de la unidad $\boldsymbol{\ell} \in \Omega$ $(\|\boldsymbol{\ell}\| = 1)$ y luego a considerar la gráfica de $\mathrm{LD} = (\boldsymbol{\omega}_0 + a\boldsymbol{\ell})$ contra a, siendo $a \in \mathbb{R}$. Esta gráfica se llama línea levantada. Cada línea levantada se puede caracterizar considerando la curvatura normal $C_\ell(\boldsymbol{\omega})$ alrededor de $a = 0$. La sugerencia es considerar la dirección $\boldsymbol{\ell} = \boldsymbol{\ell}_{max}$ correspondiente a la mayor curvatura $C_{\ell max}(\boldsymbol{\omega})$. El gráfico del índice de $\boldsymbol{\ell}_{max}$ puede revelar aquellas observaciones que bajo pequeñas perturbaciones ejercen una notable influencia en $\mathrm{LD}(\boldsymbol{\omega})$. Según Cook (1986), la curvatura normal en la dirección unitaria viene dada por

$$C_\ell(\boldsymbol{\theta}) = -2\{\boldsymbol{\ell}^\top \boldsymbol{\Delta}_{\rm p}^\top \ddot{\mathbf{L}}_{\rm p}^{-1} \boldsymbol{\Delta}_{\rm p} \boldsymbol{\ell}\},$$

donde

$$\ddot{\mathbf{L}}_{\rm p} = \frac{\partial^2 L_{\rm p}(\boldsymbol{\theta}, \lambda)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\Big|_{\widehat{\boldsymbol{\theta}}} \qquad \text{y} \qquad \boldsymbol{\Delta}_{\rm p} = \frac{\partial^2 L_{\rm p}(\boldsymbol{\theta}, \lambda|\boldsymbol{\omega})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\omega}^\top}\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}},\, \boldsymbol{\omega}=\boldsymbol{\omega}_0}.$$

Nótese que $-\ddot{\mathbf{L}}_{\rm p}$ es la matriz de información observada penalizada evaluada en $\widehat{\boldsymbol{\theta}}$ (véase la Sección 2.4) y $\boldsymbol{\Delta}_{\rm p}$ es la matriz de perturbación penalizada evaluada en $\widehat{\boldsymbol{\theta}}$ y $\boldsymbol{\omega}_0$. $C_\ell(\boldsymbol{\theta})$ denota la influencia local en la estimación de $\widehat{\boldsymbol{\theta}}$ después de perturbar el modelo o los datos. Escobar & Meeker (1992) propusieron estudiar la curvatura normal en la dirección $\boldsymbol{\ell} = \mathbf{e}_i$, donde $\mathbf{e}_i$ es un vector n-dimensional con ceros en la posición i-ésima y ceros en las posiciones restantes. En este caso, la curvatura normal, llamada influencia local total del i-ésimo individuo, toma la forma $C_{\mathbf{e}_i}(\boldsymbol{\theta}) = 2|\,\mathrm{c}_{ii}\,|(i = 1, \cdots, n)$, donde c_{ii} es el i-ésimo elemento diagonal principal de la matriz $\mathbf{C} = \boldsymbol{\Delta}_{\rm p}^\top \ddot{\mathbf{L}}_{\rm p}^{-1} \boldsymbol{\Delta}_{\rm p}$. Para tener una curvatura invariante bajo un cambio de escala uniforme, Poon & Poon (1999) propusieron la curvatura normal conforme denotada como

$$B_\ell(\boldsymbol{\theta}) = \frac{C_\ell(\boldsymbol{\theta})}{2\sqrt{\mathrm{tr}(\boldsymbol{\Delta}_{\rm p}^\top \mathbf{L}_{\rm p}^{-1} \boldsymbol{\Delta}_{\rm p})^2}} = -\frac{\boldsymbol{\ell}^\top \boldsymbol{\Delta}_{\rm p}^\top \mathbf{L}_{\rm p}^{-1} \boldsymbol{\Delta}_{\rm p} \boldsymbol{\ell}}{\sqrt{\mathrm{tr}(\boldsymbol{\Delta}_{\rm p}^\top \mathbf{L}_{\rm p}^{-1} \boldsymbol{\Delta}_{\rm p})^2}}\,.$$

Esta curvatura se caracteriza por permitir cualquier dirección unitaria $\boldsymbol{\ell}$ tal que $0 \leq B_\ell(\boldsymbol{\theta}) \leq 1$. Una sugerencia es considerar la dirección $\boldsymbol{\ell} = \boldsymbol{\ell}_{max}$ correspondiente a la mayor curvatura $B_{\ell max}(\boldsymbol{\theta})$ o, alternativamente, evaluar la curvatura normal en la dirección $\boldsymbol{\ell} = \mathbf{e}_i$ y observar el gráfico de índices de $B_{\mathbf{e}_i}(\boldsymbol{\theta})$.

5.4. Esquemas de perturbación

En la siguiente sección se presentará la expresión de $\boldsymbol{\Delta}_{\rm p}$ para los esquemas de perturbación de la ponderación de los casos, la variable de respuesta, de las covariables y conjunta de las covariables.

5.4.1. Perturbación de ponderación de casos

La perturbación de la ponderación de casos se considera para detectar observaciones con una gran contribución a la función de verosimilitud y que pueden ejercer una gran influencia en los estimadores de máxima verosimilitud penalizados. Se considerarán los pesos atribuidos a las observaciones en la función de log-verosimilitud penalizada como

$$L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda|\boldsymbol{\omega}) = \sum_{k=1}^{n} \omega_i L_i(\boldsymbol{\theta}) - \frac{\lambda}{2}\, \mathbf{f}^\top \mathbf{K}\mathbf{f}\ , \tag{5.1}$$

donde $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_n)^\top$ es el vector de pesos, con $0 \leq \omega_i \leq 1$ $(i = 1, \cdots, n)$, y $\omega_0 = (1, \cdots, 1)^\top$ denota el vector no perturbado. Diferenciando $L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda|\boldsymbol{\omega})$ con respecto a los elementos de $\boldsymbol{\theta}$ y $\boldsymbol{\omega}$ se obtiene

$$\begin{aligned}
\frac{\partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda|\boldsymbol{\omega})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\omega}}\Big|_{\theta=\hat{\theta},\ \omega=\omega_0} &= \mathbf{X}^\top a_i d_{\mu}^{(i)} \delta_{ij}^{n}, \\
\frac{\partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda|\boldsymbol{\omega})}{\partial \mathbf{f} \partial \boldsymbol{\omega}}\Big|_{\theta=\hat{\theta},\ \omega=\omega_0} &= \mathbf{N}^\top \widehat{\mathbf{D}}_a \widehat{\mathbf{D}}_{\mathbf{Z}} \qquad \text{y} \\
\frac{\partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda|\boldsymbol{\omega})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\omega}}\Big|_{\theta=\hat{\theta},\ \omega=\omega_0} &= \mathbf{Z}^\top b_i d_{\delta}^{(i)} \delta_{ij}^{n},
\end{aligned}$$

para $i = 1, \cdots, n$ con $\mathbf{D}_a$ denotado en secciones anteriores, mientras que $\mathbf{D}_{\mathbf{Z}} = \mathrm{diag}\{z_1, \cdots, z_n\}$.

5.4.2. Perturbación de la variable de respuesta

De acuerdo a Leiva *et al*. (2014), la perturbación aditiva sobre la i-ésima variable de respuesta es dada por $\mathrm{y}_{i\omega_i} = \mathrm{y}_i + \omega_i s(\mathrm{y}_i)$ donde $s(\mathrm{y}_i) = \sqrt{\widehat{\mu}_i^2/\widehat{\phi}}$ y $\omega_i \in \mathbb{R}$ para $i = (1, \cdots, n)$. A continuación, la función de log-verosimilitud penalizada se construye a partir de la ecuación (2.7) con y_i sustituida por $\mathrm{y}_{i\omega}$, es decir

$$L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda|\boldsymbol{\omega}) = L_i(\boldsymbol{\theta}|\boldsymbol{\omega}) - \frac{\lambda}{2}\, \mathbf{f}^\top \mathbf{K}\mathbf{f}\ , \tag{5.2}$$

donde $L(\cdot)$ se muestra en la expresión (2.3) con $\mathrm{y}_{i\omega_i}$ en el lugar de y_i. Aquí, el vector no perturbado viene dado por $\boldsymbol{\omega}_0 = (0, \cdots, 0)^\top$. Diferenciando $L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda|\boldsymbol{\omega})$ con respecto a los elementos de $\boldsymbol{\theta}$ y $\boldsymbol{\omega}$, obtenemos, tras algunas manipulaciones algebraicas, que

$$\begin{aligned}
\frac{\partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda|\boldsymbol{\omega})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\omega}}\Big|_{\theta=\hat{\theta},\ \omega=\omega_0} &= \mathbf{X}^\top a_i d_{y\mu}^{(i)} S_{Y_i} \delta_{ij}^{n}, \\
\frac{\partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda|\boldsymbol{\omega})}{\partial \mathbf{f} \partial \boldsymbol{\omega}}\Big|_{\theta=\hat{\theta},\ \omega=\omega_0} &= \mathbf{N}^\top \widehat{\mathbf{D}}_a \widehat{\mathbf{D}}_{\psi} \widehat{\mathbf{D}}_{\vartheta} \\
\frac{\partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda|\boldsymbol{\omega})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\omega}}\Big|_{\theta=\hat{\theta},\ \omega=\omega_0} &= \mathbf{Z}^\top b_i d_{y\delta}^{(i)} S_{Y_i} \delta_{ij}^{n},
\end{aligned}$$

para $i = 1, \cdots, n$ donde $\widehat{\mathbf{D}}_{\vartheta} = \mathrm{diag}\{\widehat{\vartheta}_1, \cdots, \widehat{\vartheta}_n\}, \widehat{\mathbf{D}}_{\psi} = \mathrm{diag}\{\widehat{\psi}_1, \cdots, \widehat{\psi}_n\},$ con $\widehat{\vartheta}_i = s(\mathrm{y}_i)$, y

$$\widehat{\psi}_i = -\frac{\widehat{\delta}[\widehat{\delta}+1]}{[\widehat{\delta}y_i + y_i + \widehat{\delta}\widehat{\mu}_i]^2} + \frac{[\widehat{\delta}+1]}{4\widehat{\mu}_i^2} + \frac{\widehat{\delta}^2}{4[\widehat{\delta}+1]y_i^2}.$$

Para la parte parámetrica del modelo, se presentan adicionalmente los esquemas de Perturbación de las covariables y Perturbación conjunta de covariables.

5.4.3. Perturbación de las covariables

En este caso, se perturba aditivamente una covariable continua X sustituyendo x_l por $x_l + \omega S_{X_l}$, donde S_{X_l} es la desviación estándar de X_l. Aquí el componente del predictor lineal i es $\eta_i(\omega) = \beta_1 + \ldots + \beta_l[x_{il} + \omega_i S_{X_l}] + \ldots + \beta_{\mathrm{p}} x_{ip}$ y $\boldsymbol{\omega}_0 = \mathbf{0}_{nx1}$ lo que ahora $\ell(\boldsymbol{\theta}|\boldsymbol{\omega}) = \sum_{i=1}^n \ell_i(\mu_i(\omega_i), \delta_i)$. Entonces, se obtiene $\boldsymbol{\Delta}$, cuyos elementos son $\boldsymbol{\Delta}(\boldsymbol{\alpha})_{ri} = \beta_l m_i z_{ir} S_{X_l}$ y

$$\boldsymbol{\Delta}(\boldsymbol{\beta})_{ji} = \begin{cases} \beta_l c_i x_{ij} S_{X_l}, & \text{para } j \neq l; \\ \beta_l c_i x_{il} S_{X_l} + d_\mu^{(i)} a_i S_{X_l}, & \text{para } j = l; \end{cases}$$

para $i = 1, \ldots, n$, $j = 1, \ldots, p$ y $r = 1, \ldots, q$, que debe ser evaluado en $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. En forma de matriz, $\boldsymbol{\Delta}(\boldsymbol{\beta}) = \beta_l S_{X_l} \mathbf{X}^T c_i \delta_{ij}^n + S_{Xn_l} \mathbf{1}_{nx1}^{(k)T} d_\mu^{(i)} a_i \delta_{ij}^n$ y $\boldsymbol{\Delta}(\boldsymbol{\alpha}) = \beta_l S_{X_l} \mathbf{Z}^T m_i \delta_{ij}^n$. Ahora, se perturba aditivamente una covariable continua Z sustituyendo z_k por $z_k + \omega S_{Z_k}$, donde S_{Z_k} es la desviación estándar de $\mathbf{Z}_k$. Aquí, el componente del predictor lineal i es $\tau_i(\omega) = \alpha_1 + \ldots + \alpha_k[z_{ik} + \omega_i S_{Z_k}] + \ldots + \alpha_q z_{iq}$ y $\boldsymbol{\omega}_0 = \mathbf{0}_{nx1}$ lo que ahora $\ell(\boldsymbol{\theta}|\boldsymbol{\omega}) = \sum_{i=1}^n \ell_i(\mu_i, \delta_i(\omega_i))$, $\boldsymbol{\Delta}(\boldsymbol{\beta})_{ji} = \alpha_k m_i x_{ij} S_{Z_k}$ y

$$\boldsymbol{\Delta}(\boldsymbol{\alpha})_{ri} = \begin{cases} \alpha_k w_i z_{ir} S_{Z_k}, & \text{para } r \neq k; \\ \alpha_k w_i z_{ik} S_{Z_k} + d_\delta^{(i)} b_i S_{Z_k} & \text{para } r = k; \end{cases}$$

lo que debe ser evaluado en $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. En forma de matriz, $\boldsymbol{\Delta}(\boldsymbol{\alpha}) = \alpha_k S_{Z_k} \mathbf{Z}^T w_i \delta_{ij}^n + S_{X_l} \mathbf{1}_{nx1}^{(l)T} d_\delta^{(i)} b_i \delta_{ij}^n$ y $\boldsymbol{\Delta}(\boldsymbol{\beta}) = \alpha_k S_{Z_k} \mathbf{X}^T m_i \delta_{ij}^n$

5.4.4. Perturbación conjunta de covariables

Otro esquema de perturbación que se puede considerar es cuando alguna covariable perturbada en X también está presente en Z. Por ejemplo, $x_{il} = z_{ik}$. Entonces, $\tau_i(\omega) = \alpha_1 + \ldots + \alpha_k[x_{il} + \omega_i S_{X_l}] + \ldots + \alpha_q z_{iq}$ y $\boldsymbol{\omega}_0 = \mathbf{0}_{nx1}$. Entonces, ahora $\ell(\boldsymbol{\theta}|\boldsymbol{\omega}) = \sum_{i=1}^n \ell_i(\mu_i(\omega_i), \delta_i(\omega_i))$ y $\boldsymbol{\Delta}$ tiene los elementos

$$\boldsymbol{\Delta}(\boldsymbol{\beta})_{ji} = \begin{cases} \beta_l c_i x_{ij} S_{X_l} + \alpha_k m_i x_{ij} S_{X_l}, & \text{para } j \neq l; \\ \beta_l c_i x_{il} S_{X_l} + \alpha_k m_i x_{ij} S_{X_l} + d_\mu^{(i)} a_i S_{X_l}, & \text{para } j = l; \end{cases}$$

$$\boldsymbol{\Delta}(\boldsymbol{\alpha})_{ri} = \begin{cases} \alpha_k w_i z_{ir} S_{X_l} + \beta_l m_i z_{ij} S_{X_l}, & \text{para } r \neq k; \\ \alpha_k w_i x_{il} S_{X_l} + \beta_l m_i x_{ij} S_{X_l} + d_\delta^{(i)} b_i S_{X_l} & \text{para } r = k; \end{cases}$$

En forma de matriz, $\boldsymbol{\Delta}(\boldsymbol{\beta}) = S_{X_l}[\mathbf{X}^T\{\beta_l c_i \delta_{ij}^n + \alpha_k m_i \delta_{ij}^n\} + \mathbf{1}_{nxp}^{(k)T} a_i d_\mu^{(i)} \delta_{ij}^n]$ y $\boldsymbol{\Delta}(\boldsymbol{\alpha}) = S_{X_l}[\mathbf{Z}^T\{\alpha_k w_i \delta_{ij}^n + \beta_l m_i \delta_{ij}^n\} + \mathbf{1}_{nxp}^{(l)T} b_i d_\delta^{(i)} \delta_{ij}^n]$. Como se ha mencionado, las matrices $\boldsymbol{\Delta}(\boldsymbol{\beta})$ y $\boldsymbol{\Delta}(\boldsymbol{\alpha})$ deben ser evaluadas en $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$

Capítulo 6

Aplicación a datos de contaminación atmosférica

En este capítulo se considera la aplicación del modelo semiparamétrico BSR con parámetro de precisión variando. En la Sección 6.1 se realiza una breve introducción de la importancia e impacto del material particulado en la humanidad, junto con una revisión de trabajos que abordan el tema desde distintos aspectos. En la Sección 6.2 se muestra un análisis exploratorio de los datos proporcionados por el Sistema de Información de la Calidad del Aire correspondientes a la contaminación del aire en la comuna de Pudahuel durante el periodo GEC del año 2019. En la Sección 6.3 se presenta la estimación de parámetros del modelo y la verificación de supuestos, considerando dos tipos de residuos. En la Sección 6.4, el análisis de influencia local, considerando dos esquemas de perturbación, los cuales son perturbación de la ponderación de casos y de respuesta. Finalmente, en la Sección 6.5 se realiza el análisis confirmatorio, en donde se identifican las observaciones que se consideran potencialmente influyentes.

6.1. Introducción

Las partículas que tienen un diámetro inferior a 2.5 micrómetros (PM2.5) están compuestas por partículas suficientemente pequeñas que penetran las vías respiratorias hasta alcanzar los pulmones y los alvéolos causando riesgos en la salud MMA (2011). Estudios epidemiológicos, toxicológicos y de exposición humana controlada relacionados han sido revisados Stanek *et al*. (2011). Concluyendo que varias investigaciones, centradas en fuentes individuales de MP, proporcionan evidencia sobre una fuente específica que afecta a la salud humana. Este es el caso de la contaminación atmosférica derivada de la tracción vehicular que provoca algunos efectos en la salud humana como asma, exacerbación de enfermedades respiratorias crónicas, problemas respiratorios y mortalidad cardiovascular total, entre otros Stanek *et al*. (2011). Otros trastornos causados por los contaminantes atmosféricos son la epilepsia, las cefaleas y la enfermedad tromboembólica venosa Cakmak *et al*. (2010). Desde hace más de tres décadas, la ciudad de Santiago de Chile es uno de los lugares urbanos que ha presentado niveles que superan los límites de contaminación nacionales e internacionales Ostro (2003). Su ubicación, topografía y meteorología provocan condiciones críticas en la salud humana cuando existe interacción con las emisiones antropogénicas, condición que ocurre cuando la contaminación del aire se combina con el calor Kinney (2008). Así, durante los meses de otoño e invierno, los contaminantes quedan atrapados en el valle de Santiago, lo que produce contaminación atmosférica en la ciudad. Debido a factores meteorológicos y topográficos, existe una acumulación de material particulado y

contaminantes gaseosos durante el invierno, y se observa un aumento de la radiación solar durante el verano que favorece las reacciones fotoquímicas Marchant *et al*. (2013); Cavieres *et al*. (2021). Pueden producirse episodios periódicos de contaminación extrema con determinados contaminantes. Dichos contaminantes y sus altos niveles varían en función de las fluctuaciones meteorológicas y geográficas, que dependen de los cambios en la fuente y el tipo de emisión. Debido a esta variación, los niveles de contaminantes atmosféricos se tratan como variables aleatorias, que pueden modelarse mediante una distribución de probabilidad Cavieres *et al*. (2021). Las condiciones meteorológicas son un factor clave e incontrolable en la determinación de la variabilidad de la contaminación atmosférica. En algunos casos, puede superar la influencia de algunos efectos antropogénicos, como los que se originan en el rastreo vehicular Yañez *et al*. (2017). Además, la relación entre las variables meteorológicas y material particulado han sido analizadas en todo el mundo Clements *et al*. (2016), algunas de estas variables han sido consideradas como variable explicativa en el modelo propuesto en este estudio. El efecto de los parámetros meteorológicos sobre el material particulado ha sido estudiado utilizando diferentes técnicas estadísticas, incluyendo regresión lineal múltiple, modelos aditivos generalizados, splines de regresión adaptativa multivariante y redes neuronales; Puentes *et al*. (2021).

6.2. Análisis exploratorio de los datos

Para este estudio se considera el conjunto de datos medioambientales relacionados con la contaminación atmosférica. En particular, los datos proporcionados por el Sistema Nacional de Información de la Calidad del Aire (www://sinca.mma.gob.cl) correspondientes a la contaminación del aire en la comuna de Pudahuel durante el período GEC (1 de abril al 31 de agosto) del año 2019. Cuyo conjunto de datos contiene las variables de concentraciones de MP2.5, concentraciones de MP10 y velocidad del viento (m/s), estos datos fueron obtenidos por un registro horario (cada hora del día), en donde existe un porcentaje pequeño de datos perdidos; para aquello se aplicará el método de imputación por medias móviles para estimar los valores faltantes y luego se procede a promediar los registros diariamente. El objetivo del estudio es evaluar la asociación de las concentraciones contaminantes con variables meteorológicas mediante el uso del modelo semiparamétrico Birnbaum-Saunders reparametrizado con parámetro de precisión variando. Para este estudio, con el propósito de motivar el uso de los modelos semiparamétricos, se consideran las concentraciones promedio diarias de MP2.5 como variable de respuesta, se trabajará así con un total de 153 observaciones.

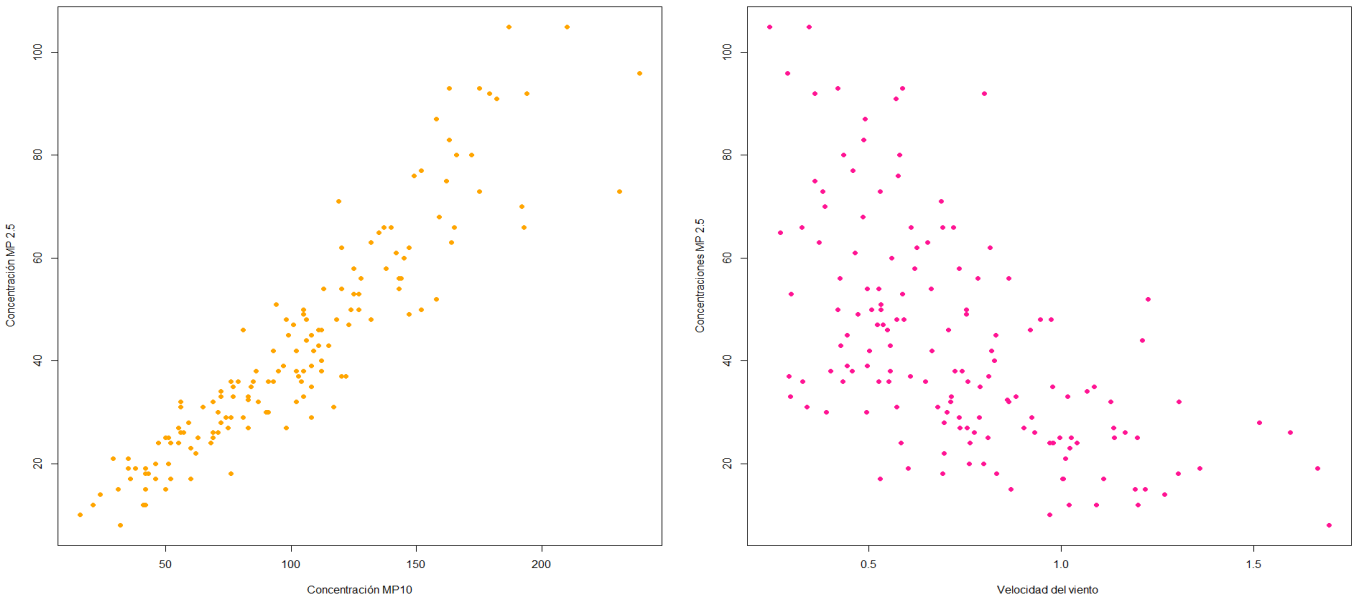

Figura 6.1: Scatter plots de MP2.5 vs MP10 (izquierdo) y MP2.5 vs velocidad del viento (derecho).

La Figura 6.1 contiene los gráficos de dispersión entre la variable de respuesta y la covariable MP10 y velocidad del viento respectivamente. En el gráfico de la izquierda se aprecia la relación entre las concentraciones, la cual es positivamente lineal, esto puede ser apoyado por la correlación entre ambas variables que es igual a 0.93, también se observa que la variabilidad de MP2.5 tiende a aumentar a medida que aumentan los valores de MP10 y muestra evidencias de que la línea no pasa por el origen, por lo que podría ser útil considerar el intercepto en el componente paramétrico del modelo seleccionado. Mientras que, la relación entre concentraciones MP2.5 y velocidad del viento la cual corresponde al gráfico derecho existe una variación clara, pero no es del tipo lineal, más bien los puntos forman una especie de curvatura.

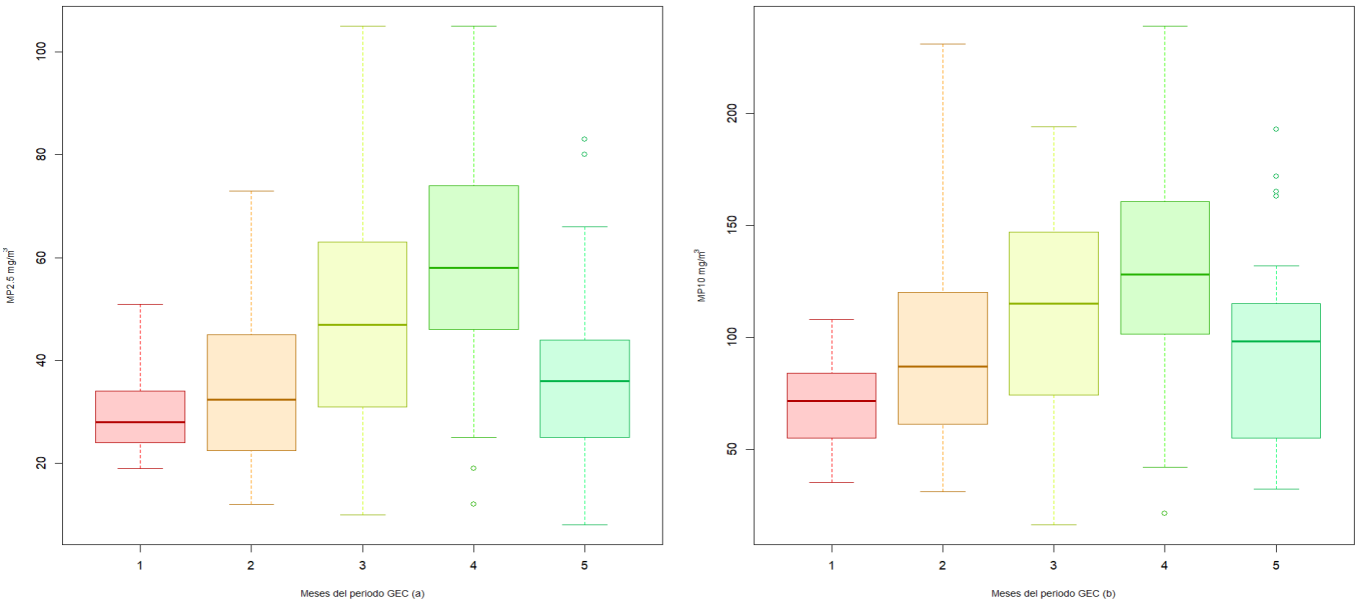

Figura 6.2: Boxplots ajustados para MP2.5 (a) y MP10 (b) por meses del periodo GEC registrado por la estación de monitoreo Pudahuel, Chile 2019.

Se aprecia en la Figura 6.2 que tanto el MP2.5 como el MP10 respectivamente, presentan concentraciones altas en los meses del periodo GEC, los cuales son meses fríos. Esto se puede explicar por dos condiciones; primero, debido a la presencia en otoño e invierno de condiciones meteorológicas que no favorecen la dispersión de los contaminantes, y segundo, la existencia de un aumento de las emisiones de material particulado en el área, debido a métodos de calefacción y/o uso de leña, cuyo uso aumenta con las bajas temperaturas de los meses entre abril y agosto.

6.2.1. Variable de respuesta MP2.5

Como fue señalado anteriormente, resulta de gran interés estudiar los promedios diarios de concentraciones de MP2.5 durante el periodo GEC, dado al incremento de material particulado que se observa durante esos meses en la comuna de Pudahuel.

Material Particulado	Media	Mediana	SD	CV	CS	CK	Mínimo	Máximo	n
MP2.5	42.27	37	21.5	0.51	0.89	3.26	8	105	153

Cuadro 6.1: Estadística descriptiva para niveles de MP2.5 (en mg/m^3) registrado por la estación de monitoreo de Pudahuel, Chile 2019

El cuadro 6.1 proporciona un resumen descriptivo de la variable de respuesta MP2.5 (en micrómetros), incluye la media, la mediana, la desviación estándar (SD), el coeficiente de variación (CV), el coeficiente de asimetría (CS), el coeficiente de curtosis (CK), el mínimo, el máximo y el total de observaciones (n). Se puede observar que CS = 0.89, lo que indica una ligera asimetría en los datos de la variable de respuesta, y su CK = 3.26, que indica una densidad con colas pesadas con respecto a la distribución normal, se logra identificar que los valores están distribuidos de forma asimétrica. Para verificar esto se precede a gráficar la variable de respuesta, como se presenta a continuación

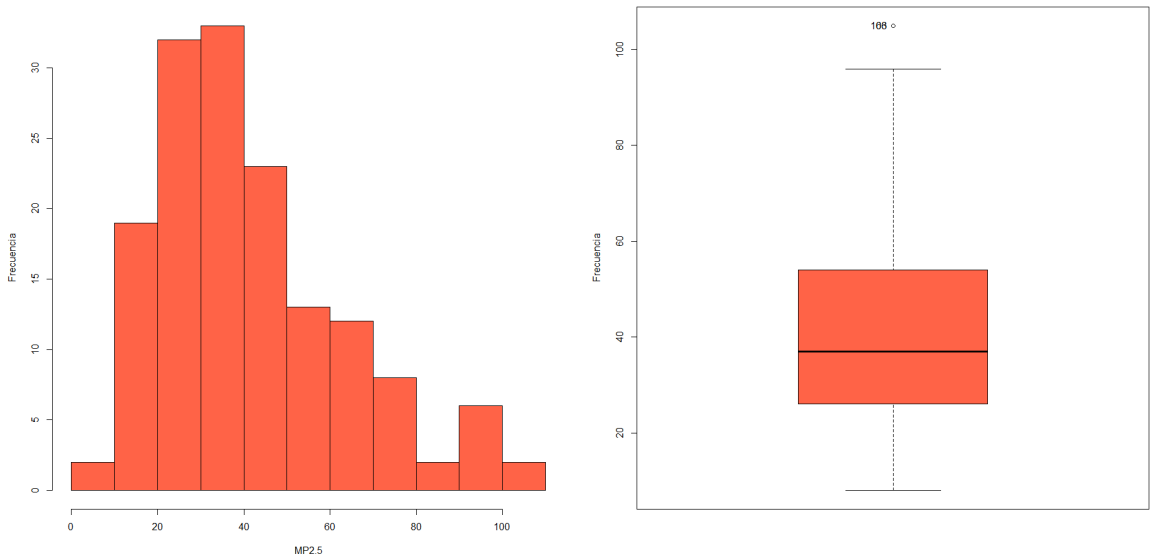

Figura 6.3: Histograma y boxplot de la variable de respuesta MP2.5, durante el periodo GEC del año 2019 en la comuna de Pudahuel.

A partir de la Figura 6.3, en el histograma se aprecia que valores de las concentraciones de PM2.5 tienen una distribución empírica que está sesgada positivamente y en el boxplot se detectan un posible valor atípico. A partir de este análisis exploratorio de los datos, se puede ver que la variable de respuesta presenta características de la distribución BSR, como la asimetría positiva, también en los gráficos de dispersión se puede que las covariables contribuyen de forma lineal y no lineal a la variable de respuesta, por lo que se puede proponer un modelo semiparamétrico para modelar las concentraciones de MP2.5 en función de MP10 y velocidad del viento que contribuyen de forma paramétrica y no paramétrica, respectivamente, al modelo.

6.3. Estimación y verificación de supuestos

Una vez realizado el análisis exploratorio y haber comprobado las tendencias presentadas en los datos de la Sección 6.1, se sugiere usar un modelo semiparamétrico BSR entre MP2.5 y las covariables MP10 y velocidad de viento, asumiendo las siguientes estructuras para las funciones de enlace

$$g(\mu_i) = \beta_1 + \beta_2\, \mathrm{x}_i + f(\mathrm{t}_i) \qquad \text{y} \qquad h(\delta_i) = \alpha_1 + \alpha_2\, \mathrm{z}_i \qquad (i = 1, 2, \ldots, 153), \tag{6.1}$$

donde μ_i denota el valor del MP2.5 asociado a la i-ésima medición, x_i denota el valor de MP10 asociado a la i-ésima medición y t_i denota el valor de la i-ésima medición de la velocidad del viento, $g(\cdot)$ es la función de enlace del modelo dado en la sección 2.3, $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ es un vector de parámetros y $f(\cdot)$ es la función suave y z_i denota la i-ésima unidad experimental de la variable MP10. Se aplica el procedimiento descrito en la Sección 4.4 para estimar el parámetro de suavizamiento para el componente no paramétrico del modelo, entonces el estimador de λ obtenido es $\widehat{\lambda}$ igual a 0.0054. Para estimar los parámetros de los componentes paramétricos del modelo, se maximiza la función de log-verosimilitud penalizada tal y como se describe en la Sección 2.4, entonces, los estimadores de máxima verosimilitud penalizada para $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ y $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^T$ son $\widehat{\beta_1} = 4.11$, $\widehat{\beta_2} = 0.42$ y $\widehat{\alpha_1} =$ -15.4, $\widehat{\alpha_2} = 0.92$, cuyos errores estándar son 0.000236043, 0.01903877, 0.00000076 y 0.0007287524, respectivamente. Para comprobar que el modelo

es adecuado para describir la media de la variable de respuesta MP2.5, se verifican los supuestos establecidos para el modelo, primero se observa que β_1 y β_2 son altamente significativos al 5 %, ya que ambos p-valores empíricos son cercanos a cero, tal y como se esperaba en el análisis exploratorio que se realizó. Así, el componente paramétrico seleccionado para el modelo parece ser adecuado.

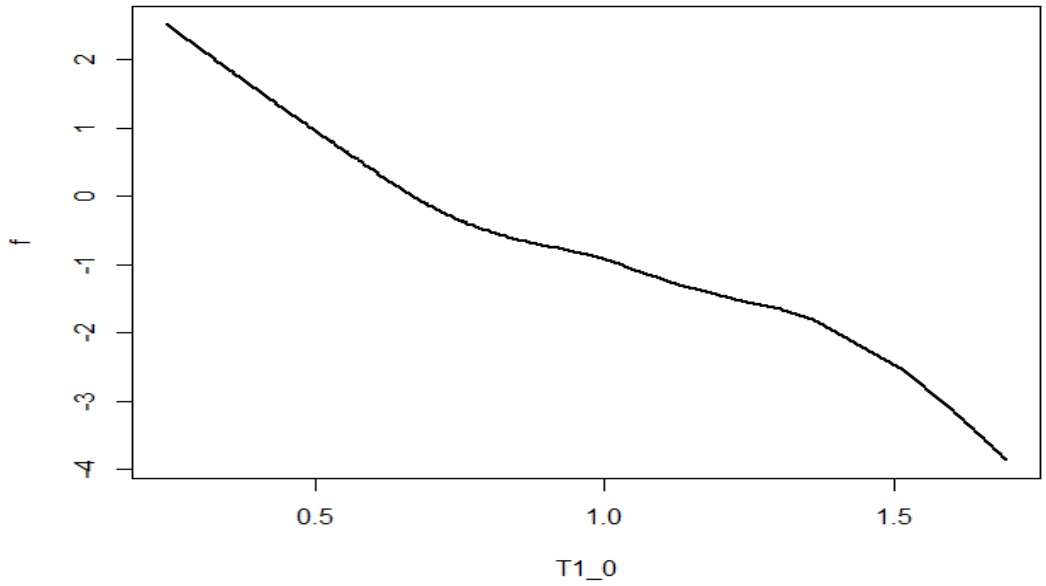

Figura 6.4: Gráfico de la función estimada.

En la Figura 6.4 se presenta el gráfico de la función estimada suavizada teniendo en cuenta el parámetro de suavizamiento estimado definido anteriormente. En donde se aprecia una curva suave, que no presenta saltos ni quiebres pronunciados.

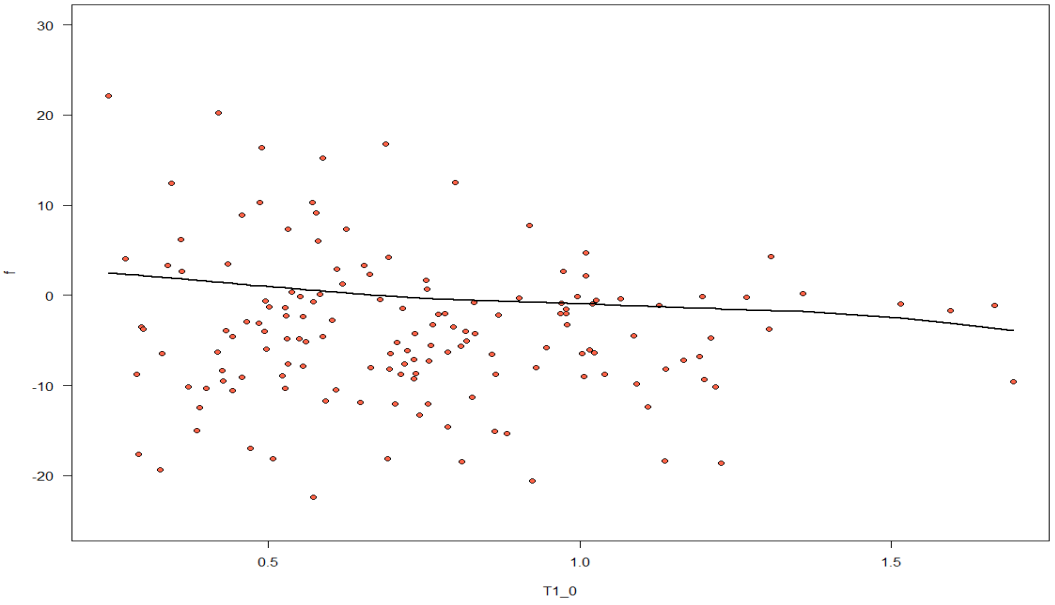

Figura 6.5: Gráfico de residuos parciales vs función de velocidad del viento estimada superpuesta.

En la Figura 6.5 se presenta el gráfico de los residuos parciales, $r^{(p)}$, definidos como

$$r^{(p)} = g(\widehat{\mu}_i) - \widehat{\beta}_1 - \widehat{\beta}_2 x_i$$

y la variable velocidad del viento en la que se solapa la curva de la función suave estimada. Nótese que la curva parece adaptarse bien, ya que cubre correctamente esos valores tan distantes, esto verifica lo mencionado anteriormente. Por lo tanto, la contribución no lineal de la covariable velocidad del viento sobre la variable de respuesta MP2.5 se cuantifica correctamente a través de la función estimada función suave. Ahora se realizá un análisis residual basado en los datos ambientales para verificar los supuestos.

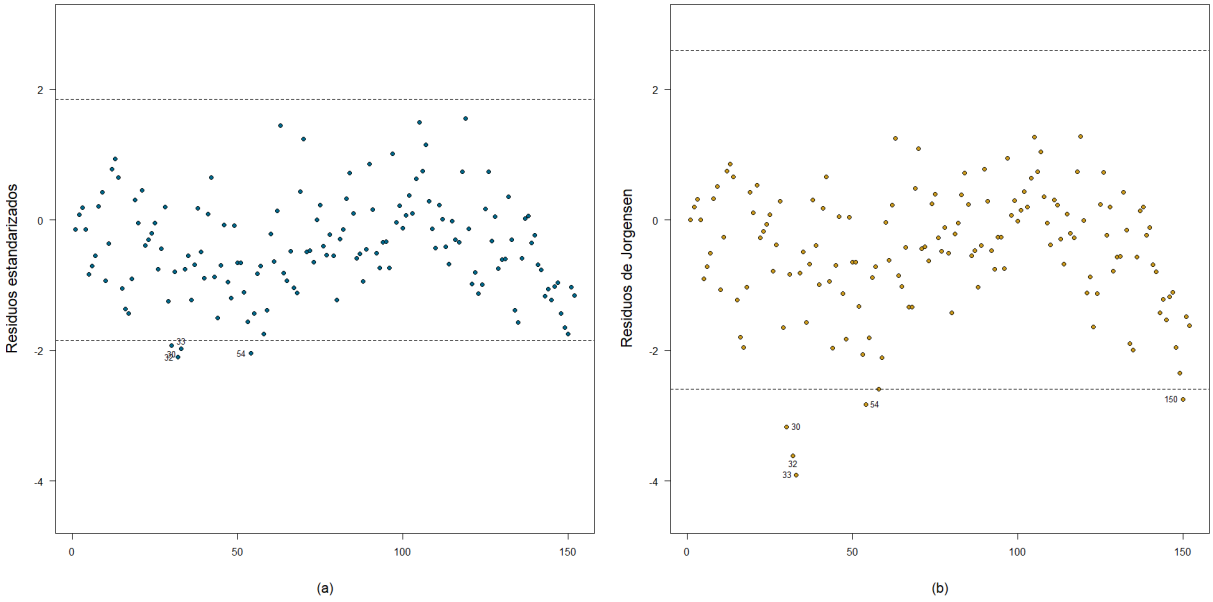

Figura 6.6: Gráfico de residuos estandarizados (a) y residuos de Jørgensen (b).

La Figura 6.6 contiene los gráficos para los índices contra dos tipos de residuos propuestos a lo largo de este trabajo, los cuales son residuos estandarizados $r^{(1)}$ (a) y residuos de Jørgensen $r^{(2)}$ (b). En este caso, los residuos estandarizados toman valores entre el intervalo [-2, 2] aproximadamente, excepto por algunos valores que van fuera de las bandas (descritas como dos veces la desviación estándar de cada residuo), estos valores son las observaciones 30, 32, 33, 54, que corresponden a las fechas 30 de abril, 02, 03 y 24 de mayo del año 2019, respectivamente. Por otra parte, los residuos de Jørgensen toman valores en el intervalo [-2.5, 2.5] y también se presentan un par de valores fuera de la banda de confianza, estos son el 30, 32, 33, 54 y 150 correspondientes a las fechas 30 de abril, 02, 03, 24 de mayo y 28 de agosto del año 2019, respectivamente. Como los dos tipos de residuos son similares, se utilizarán los residuos estandarizados, ya que la distribución BSR está relacionada con la distribución Normal. Para verificar el supuesto distribucional establecido en el modelo, a continuación se realizá una gráfica QQplot y un histograma para los residuos estandarizados $r^{(1)}$.

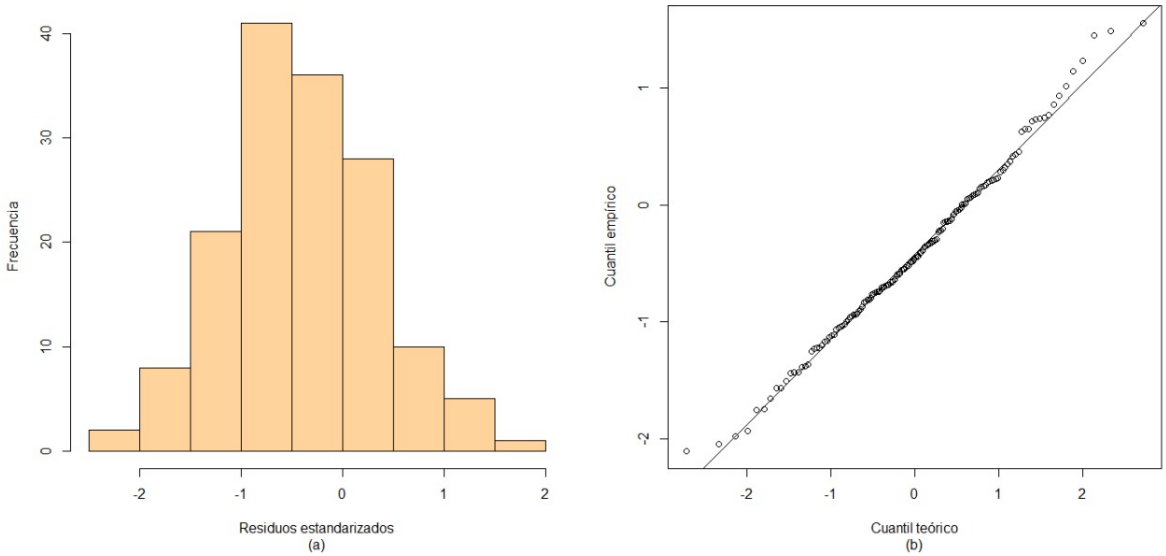

Figura 6.7: Histograma (a) y QQplot (b) de los residuos estandarizados.

La Figura 6.7 no muestra rasgos inusuales, por lo que la hipótesis de distribución de la variable de respuesta parece ser adecuada. También, el supuesto de independencia se verifica por ambas figuras, que muestran una simetría considerable. Así, la función de enlace identidad que fue seleccionada anteriormente para este modelo parece ser apropiada dado el comportamiento residual presentado.

6.4. Análisis de Influencia Local

Las técnicas de influencia local para el modelo semiparamétrico BSR con parámetro de precisión variando propuestas en la Sección 2.3 se presentan gráficamente en las Figuras 6.8 y 6.9, en donde se aprecian las observaciones 32, 54 y 96 como potencialmente influyentes.

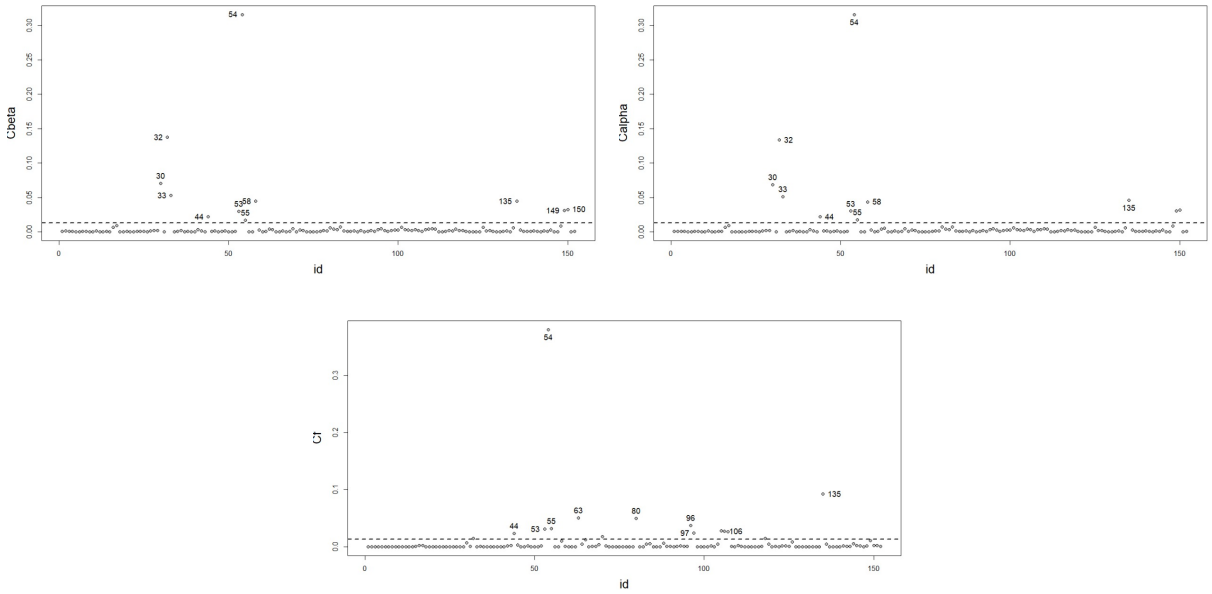

Figura 6.8: Gráficos de índice C_i para $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ y $\mathbf{f}$ bajo el esquema de perturbación de ponderación de casos.

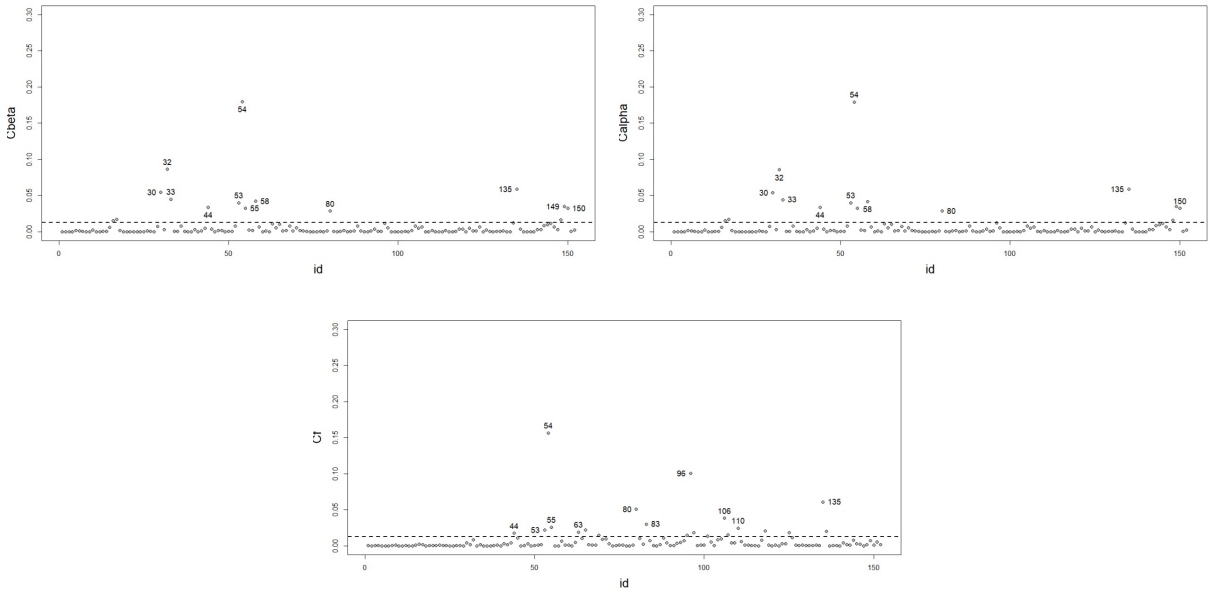

Figura 6.9: Gráficos de índice C_i para $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ y $\mathbf{f}$ bajo el esquema de perturbación de la variable de respuesta.

6.5. Análisis confirmatorio

De la sección anterior, se logró identificar con éxito las observaciones que se consideran como posibles potenciales influyentes, es por ello que se analizá el cambio que presentan las estimaciones realizadas con anterioridad cuando se excluyen estas observaciones influyentes. Para realizar este análisis, el conjunto de observaciones {32}, {54}, {96}, {32, 54}, {32, 96}, {32, 54, 96} se eliminarán y las estimaciones de los parámetros del modelo se realizarán nuevamente. En la tabla presentada a continuación se proporcionan los cambios relativos (CR) en las estimaciones de parámetros, en sus correspondientes errores estándar estimados y en el nuevo valor-p. Estos cambios se calculan a partir de

$$CR_{\hat{\theta}_{j(i)}} = \left| \frac{\hat{\theta}_j - \hat{\theta}_{j(i)}}{\hat{\theta}_j} \right| X100\,\% \qquad \text{y} \qquad CR_{SE(\hat{\theta}_j)_{(i)}} = \left| \frac{SE(\hat{\theta}_j) - SE(\hat{\theta}_j)_{(i)}}{SE(\hat{\theta}_j)} \right| X100\,\%,$$

donde $\hat{\theta}_{j(i)}$ y $SE(\hat{\theta}_j)_{(i)}$ denotan las estimaciones de máxima verosimilitud de θ_j y SE sus correspondientes errores estándar, obtenido después de extraer la observación i-ésima, para $j = 1, 2$ y $i = 1, 2, \ldots, 153$. Del cuadro (6.2) nótese que los cambios relativos más importantes se detectan para las estimaciones del parámetro de precisión, es decir, α_1 y α_2. En particular para la observación $\{96\}$. En las estimaciones de máxima verosimilitud, de manera individual la observación $\{54\}$, tanto en los betas como en los alphas, fue la observación más influyente; de manera conjunta, al eliminar las observaciones $\{32, 54, 96\}$ hubo cambios relativos importantes, los resultados presentados en esta tabla muestran que las medidas de diagnóstico derivadas en este estudio identifican estos puntos potencialmente influyentes que afectan principalmente a la inferencia estadística del modelo presentado, pero no de forma significativa. Cabe destacar que estas observaciones corresponden a los días 02 de mayo, 24 de mayo y 05 de julio del periodo GEC del año 2019 en la comuna de Pudahuel. Se puede apreciar además que la significación de los parámetros, al 5 %, no cambia, ya que los p-valores se mantienen por debajo de 0.01.

Casos eliminados	CR en la estimación de	β_1	β_2	α_1	α_2
Ninguno	θ	-	-	-	-
	SE	-	-	-	-
	p-valor	< 0.01	< 0.01	-	-
{32}	θ	0.7540757	0.4627552	0.2233433	52.8791
	SE	1.106462	0.7199899	37.3746	49.47
	p-valor	< 0.01	< 0.01	-	-
{54}	θ	3.200001	0.548869	6.55443	51.76142
	SE	1.34685	0.6892648	37.7177	49.9
	p-valor	< 0.01	< 0.01	-	-
{96}	θ	0.4467491	0.2255878	2.04896	52.6584
	SE	0.2673927	0.7424298	33.64495	44.793
	p-valor	< 0.01	< 0.01	-	-
{32, 54}	θ	4.019828	1.026363	6.448812	51.6418
	SE	2.488492	0.7424298	38.53939	50.03952
	p-valor	< 0.01	< 0.01	-	-
{32, 96}	θ	1.250974	0.2462318	1.951693	52.53937
	SE	0.8556559	0.014451	34.53927	45.70581
	p-valor	< 0.01	< 0.01	-	-
{32, 54, 96}	θ	4.572811	0.810226	8.778648	51.18353
	SE	2.206816	0.6539416	35.2801	46.1563
	p-valor	< 0.01	< 0.01	-	-

Cuadro 6.2: Cambios relativos (en %) en las estimaciones de máxima verosimilitud, los correspondientes errores estándar estimados para los casos eliminados indicados, y los respectivos p-valor.

En síntesis, este análisis de diagnóstico basado en el enfoque de la influencia local y los residuos confirman que el modelo semiparamétrico Birnbaum-Saunders reparametrizado con parámetro de precisión variando, presentado en la Subsección 2.4, es adecuado para modelar datos ambientales, incluso si existen valores atípicos y observaciones potencialmente influyentes.

7. Conclusión y trabajos futuros

En este trabajo se estudió el modelo semiparamétrico Birnbaum-Saunders reparametrizado con parámetro de precisión variando, generalizando los trabajos existentes en la literatura sobre el tema, en donde se evidencia que la distribución Birnbaum-Saunders tiene excelentes características que no se pueden omitir al querer modelar datos ambientales a través de modelos de regresión desarrollados en la literatura cuando la media y la precisión se modelan simultáneamente. Fue posible mantener la escala original de los datos y no realizar alguna clase de transformación, ya que al hacer esto la variable modelada puede reducir su interpretabilidad. Todo lo anterior evidencia el potencial de utilizar la nueva metodología basada en un modelo semiparamétrico Birnbaum-Saunders reparametrizado con parámetro de precisión variando. Dentro de las etapas desarrolladas en el modelo se realizó un procedimiento de estimación para los parámetros del modelo basados en la función de log-verosimilitud doblemente penalizada y los algoritmos Scoring de Fisher y Backfitting ponderado, se desarrolló un análisis de diagnóstico que incluía análisis de residuos y métodos para realizar un análisis de influencia local con el propósito de encontrar observaciones potencialmente influyentes. Un aspecto importante a tener en consideración tiene que ver con la flexibilidad del modelo, ya que permitió modelar una variable aleatoria cuyo supuesto de distribución se extiende más allá de la distribución Normal, de forma similar a los modelos lineales generalizados. Finalmente, se aplicó la metodología desarrollada a un conjunto de datos de contaminación atmosférica de Santiago, Chile, verificando de esta forma que el modelo propuesto es adecuado para trabajar con esta clase de variables. Como trabajo futuro, el modelo Birnbaum-Saunders reparametrizado con precisión variando se puede extender para los casos de coeficientes parcialmente variando o componentes estacionarios o espaciales, se pueden considerar además otras formas de penalizar. Este modelo resultó ser útil en datos de contaminación y puede servir en otros tópicos de interés.

Anexos

```
library(readxl)
library(plyr)
library(tidyr)
library(dplyr)
library(sBF)
library(ggplot2)
library(psych) #estadistica descriptiva
library(robustbase)
library(imputeTS)
library(pracma)
library(corpcor)
library(modes)
library(moments)
library(readr)
library(VGAM)
library(gbs)
library(MASS)
library(maxLik)
library(betareg)
library(nortest)
library(car)
library(agricolae)

datos <- read_excel("C:/Users/56990/OneDrive - alumnos.uv.cl/Escritorio/final.xlsx")

#Funciones para el promedio diario de las variables viento, humedad y temperatura

X1<-datos$vientonuevo
X2<-datos$humedadnuevo
X3<-datos$temperaturanuevo

viento<-datos$viento
humedad<-datos$humedad
temperatura<-datos$temperatura

#viento
for (i in 1:153) {
X1[i+1] <- mean(viento[((24*i)+1):(((24*i)+1) +23)])
}
X1[3] <- mean(viento[1:24])

#humedad
for (i in 1:153) {
X2[i+1] <- mean(humedad[((24*i)+1):(((24*i)+1) +23)])
}
X2[1] <- mean(humedad[1:24])

#temperatura
for (i in 1:153) {
X3[i+1] <- mean(temperatura[((24*i)+1):(((24*i)+1) +23)])
}
X3[1] <- mean(temperatura[1:24])



#_________________Estimación_____________________________

#Análisis descriptivo variable respuesta#
Y <- datos$`PM2,5`

length(Y)
summary(Y)
modes(Y)
R = max(Y)-min(Y)
IQR(Y)
sd(Y)
var(Y)
cv = sd(Y)/mean(Y)
skewness(Y)
kurtosis(Y)
missing(Y)

#Regresores lineales#
L1 <- datos$PM10

#Regresores no lineales#
X1 <- datos$Viento

cor(Y,L1)

#Constantes#
c <- as.numeric(length(Y))
CONS <- numeric(c)

for (i in 1:c) {

CONS[i] <- 1
```

```
}

#Matriz que contiene los regresores lineales#
W <- cbind(CONS,L1)
WT <- t(W)


#Regresores no lineales distintos y ordenados#
T1 <- X1

T1_0 <- sort(T1[!duplicated(T1)])

#Matriz que contiene los regresores no lineales#
Z <- cbind(CONS,L1)
ZT <- t(Z)


#Dimensiones#
n <- as.numeric(length(Y))
p <- as.numeric(length(W[1,]))
k1 <- as.numeric(length(T1))
r1 <- as.numeric(length(T1_0))

#Vectores de 1's#
V1 <- numeric(r1)
for (i in 1:r1) {

V1[i] <- 1

}

V1T <- t(V1)

#Matrices de unos en la diagonal principal#
Jn <- matrix(0, nrow = n, ncol = n, byrow = TRUE)
for (i in 1:n) {

Jn[i,i] <- 1

}

Jr1 <- matrix(0 , nrow = r1, ncol = r1, byrow = TRUE)
for (i in 1:r1) {

Jr1[i,i] <- 1

}

#Matriz Q1#
h1 <- numeric(r1-1)

for (i in 1:(r1-1)) {

h1[i] <- T1_0[i+1] - T1_0[i]

}

Q1 <- matrix(0, nrow = r1, ncol = (r1-1), byrow = TRUE)

for (i in 1:r1){

for (j in 2:(r1-1)){

if(abs(i-j)<2) {Q1[j-1,j] <- solve(h1[j-1])
Q1[j,j] <- -(solve(h1[j-1])+solve(h1[j]))
Q1[j+1,j] <- solve(h1[j])
}else Q1[i,j] <- 0


}

}

Q1 <- Q1[1:r1,2:(r1-1)]
Q1T <- t(Q1)

#Matriz K1#
R1 <- matrix(0,nrow=r1, ncol=r1, byrow = TRUE)

for (i in 2:(r1-1)){

for (j in 2:(r1-2)){

if(abs(i-j)<2) {R1[i,i] <- (1/3)*(h1[i-1]+h1[i])
R1[i,i+1] <- (1/6)*h1[i]
R1[i+1,i] <- (1/6)*h1[i]
}else  R1[i,j] <- 0


}

}
```

```
R1 <- R1[2:(r1-1),2:(r1-1)]
R1I <- solve(R1)

K1 <- Q1%*%R1I%*%Q1T

#Matriz N#
N <- matrix(1,nrow=n, ncol=r1, byrow = TRUE)

for (i in 1:n) {

for(j in 1:r1){

if (T1[i] == T1_0[j]) {N[i,j] <- 1

}else N[i,j] <- 0

}
}

NT <- t(N)

##Parametros smooth#
fi<- var(Y)
m<-3002

##Lambda1#

u_k1<-runif(m, min = 0, max = 1)
a_k1 <-seq(0.00001,0.00004,by= 0.00000001)

v_if<-numeric(c)
for (i in 1:c) {

v_if[i]<-1

}

s_k1<-matrix(0,nrow = c , ncol=c )

df_ak_1<-numeric(m)
for (i in 1:m) {

s_k1<-ginv(NT%*%Dv0%*%N + a_k1[i]*fi*K1)%*%NT%*%Dv0
xD1<- sum(diag(N1%*%s_k1))

df_ak_1[i]<- xD1


}


plot(a_k1,df_ak_1,xlab ="S.Parameter",ylab ="Grados de libertad", main="Inicial")
cbind(a_k1,df_ak_1)

lambda1 <- 0.00057


#Estimacion de los beta y F#

F_0  = solve(NT%*%N + lambda1*K1)%*%NT%*%as.matrix(Y)

ybar  = mean(Y)
vart  = (n / (n - 1)) * var(Y)
delta =  ((ybar^2) - vart + sqrt((ybar^4) + (3 * (ybar^2) * vart))) / vart
betas = solve(WT%*%W)%*%WT%*%as.matrix(Y)
alfas =  solve(ZT%*%Z)%*%ZT%*%as.matrix(Y)

epsilon_theta <- 0.0001
epsilon_beta <- 0.0001
epsilon_delta <- 0.0001
epsilon_precision <- 0.0001

epsilon_f <- 0.0001
norma_theta <- 1000
norma_beta <- 1000
norma_delta <- 1000
norma_precision <- 1000
norma_f <- 1000


beta_i <- betas
alfa_i<-alfas
delta_i <- delta
f_i <- F_0
L_i <- sum (

( delta_i/2 ) - ( (log(16*pi))/2 ) -
( 0.5*log( ( (delta_i + 1)*(Y^3)*(((W%*%beta_i) + N%*%f_i )^2))/( delta_i*Y + Y + delta_i*(((W%*%beta_i) + N%*%f_i)^2) )^2 ) ) -
( ( Y*(delta_i + 1) )/( 4*(((W%*%beta_i) + N%*%f_i)^2) ) ) -
( ( (delta_i^2)*( ((W%*%beta_i) + N%*%f_i)^2 ) )/( 4*(delta_i + 1)*Y ) ) ) -
```

```
( (lambda1/2)*t(f_i)%*%K1%*%f_i )


conteo1 <- 0
conteo2 <- 0
conteo3 <- 0
conteo4 <- 0

I0 <- numeric(1)

while (norma_theta > epsilon_theta) {

while (norma_precision > epsilon_precision) {

while (norma_beta > epsilon_beta & norma_f > epsilon_f) {


f.bs0 <- function(x){

((sqrt(delta_i+1)*(exp(1)^(delta_i/2)))/(4*sqrt(pi*ybar)*(x^(3/2)))) * ((x + ((delta_i*ybar)/(delta_i+1)))^(-2)) *
(exp(1)^(-(delta_i/4)*((((delta_i+1)*x)/(delta_i*ybar)) + ((delta_i*ybar)/((delta_i+1)*x)))))

}
integral0 <- integrate(f.bs0, lower = 0, upper = Inf)
I0<- integral0$value

eta0 <- (W%*%beta_i) + (N%*%f_i)
mu0 <- ((W%*%beta_i) + (N%*%f_i))
a0 <- matrix(1,nrow=n,ncol=1)
v0 <- (delta_i*(a0^2)/(2*(mu0^2))) + (((delta_i^2)*(a0^2))/((delta_i+1)^2))*I0
b0 <- (1/2) - (1/(2*(delta_i+1))) + ((Y+ mu0)/(delta_i*Y + Y + delta_i*mu0)) - (Y/(4*mu0)) - ((delta_i*(delta_i+2)*mu0)/(4*((delta_i+1)^2)*Y))
S0 <- ((1/(2*(mu0*(delta_i+1)))) + ((delta_i*(mu0))/((delta_i+1)^3))*I0)*a0*b0


Da0 <- diag(as.vector(a0))
Dv0 <- diag(as.vector(v0))
Dva0 <- solve(Dv0)%*%Da0
Ds0 <- diag(as.vector(S0))
Dvs0<-solve(Ds0)%*%Da0
z0 <- ((-1/(2*mu0) + delta_i/((delta_i*Y) + Y + (delta_i*mu0))) + ((Y*(delta_i+1))/(4*(mu0)^2)) - ((delta_i^2)/(4*Y*(delta_i+1))))


r0 <- eta0 + Dva0%*%z0

r.va0 <- r0
u0 <- ((((delta_i^2)+(3*delta_i)+1)/(2*(delta_i^2)*(delta_i+1)^2)) + (((mu0^2)/(delta_i+1)^4))*I0)*b0^{2}

r.abu0 <- sum(b0) + sum(u0)*delta_i + t(S0)%*%Da0%*%W%*%beta_i + t(S0)%*%Da0%*%N%*%f_i
Betas <- solve(WT%*%Dv0%*%W)%*%WT%*%Dv0%*%(Dva0%*%z0+eta0+ Dvs0%*%(Z%*%alfa_i)- (N%*%f_i)-Dvs0%*%(Z%*%alfa_i))


F1 <- (Jr1 - (V1%*%V1T)/r1)%*%solve(NT%*%Dv0%*%N + (lambda1*K1))%*%NT%*%Dv0%*%(r.va0 - (W%*%Betas))
norma_beta <- sqrt((t(Betas-beta_i)%*%(Betas-beta_i))/((t(beta_i)%*%beta_i)))
norma_f <- sqrt((t(F1-f_i)%*%(F1-f_i))/((t(f_i)%*%f_i)))

beta_i <- Betas
f_i <- F1

conteo1 <- conteo1 + 1


}


tau <- (Z%*%alfa_i)
mu0 <- ((W%*%beta_i) + (N%*%f_i))
a0 <- matrix(1,nrow=n,ncol=1)
u0 <- ((((delta_i^2)+(3*delta_i)+1)/(2*(delta_i^2)*(delta_i+1)^2)) + (((mu0^2)/(delta_i+1)^4))*I0)*b0^{2}
Da0 <- diag(as.vector(a0))
Du0 <- diag(as.vector(u0))
Dv0 <- diag(as.vector(v0))
Dva0 <- solve(Dv0)%*%Da0
Dua0 <- solve(Du0)%*%Da0

j0<-(Y+mu0)/((delta_i*Y) + Y + (delta_i*mu0)) - (Y/(4*mu0)) - ((delta_i*(delta_i+2)*mu0)/(4*((delta_i + 1)^{2})*Y)) + delta_i/(2*(delta_i+1))

r0 <- tau + Dva0%*%j0
r.va0 <- r0
r.abu0 <- sum(b0) + sum(u0)*delta_i + t(S0)%*%Da0%*%W%*%beta_i + t(S0)%*%Da0%*%N%*%f_i
norma_precision <- sqrt((t(Alphas-alphas_i)%*%(Alphas-alphas_i))/((t(Alphas)%*%alphas_i)))
Alphas <- solve(ZT%*%Dv0%*%Z)%*%ZT%*%Dv0%*%(Dua0%*%j0 + Dvs0%*%eta0-Dvs0%*%eta0+ (Z%*%alfa_i))

alphas_i<-Alphas

conteo2<- conteo2 + 1

}


L <- sum (

( delta_i/2 ) - ( (log(16*pi))/2 ) -
```

```
( 0.5*log( ( (delta_i + 1)*(Y^3)*(((W%*%beta_i) + N%*%f_i )^2))/( delta_i*Y + Y + delta_i*(((W%*%beta_i) + N%*%f_i)^2) )^2 ) ) -
( ( Y*(delta_i + 1) )/( 4*(((W%*%beta_i) + N%*%f_i)^2) ) ) -
( ( (delta_i^2)*( ((W%*%beta_i) + N%*%f_i)^2 ) )/( 4*(delta_i + 1)*Y ) ) ) -

( (lambda1/2)*t(f_i)%*%K1%*%f_i )

norma_theta <- abs((L_i-L)/(L))
L_i <- L
conteo3 <- conteo3 +1

}

#Estimaciones#
beta_i
alphas_i
f_i


#Grados de libertad#
sum(diag(N%*%solve(NT%*%Dv0%*%N + (lambda1*K1))%*%NT%*%Dv0))

#Residuos parciales (estimaciones de F1 ponderadas por N1)#
Yn <- (Y - W%*%beta_i)


plot(Yn,
type="p",
col="black",
bg="gray",
pch=21,
lwd=1,
main="",
cex.main=1,
xlab="",
ylab="",
axes=TRUE,
las=1,
bty="o",
cex.lab=1.5,
cex.sub=1.4)


#Grafico de ajuste de la funcion estimada#
plot(X1, Yn,
type="p",
col="black",
bg="tomato",
pch=21,
lwd=1,
main="",
cex.main=1,
xlab="",
ylab="",
ylim=c(-25,30),
xlim=c(min(X1),max(X1)),
axes=TRUE,
las=1,
bty="o",
cex.lab=1.5,
cex.sub=1.4)

par(new=TRUE)
plot(T1_0,f_i,type="l",ylab="f",xlab="T1_0",xlim=c(min(X1),max(X1)),ylim=c(-25,30),main="",axes=F,lwd=2,col="black")


#Variable respuesta estimada#
Y_e <- W%*%beta_i + N%*%f_i


plot(Y, Y_e,
type="p",
col="black",
bg="aquamarine",
pch=21,
lwd=1,
main="",
cex=1.4,
cex.main=1,
xlab="MP2.5",
ylab="MP2.5 estimado",
axes=TRUE,
las=1,
bty="o",
cex.lab=1.5,
cex.sub=6.1)


#Matriz de información esperada de Fisher y covarianza#


Kbb <- WT%*%Dv0%*%W
Kbd <- WT%*%Ds0%*%Z
```

```
Kbf<- WT%*%Dv0%*%N


Kdb <- t(Kbd)
Kdd <- ZT%*%Du0%*%Z
Kfd <- NT%*%Ds0%*%Z
Kdf <- t(Kfd)

Kfb <- t(Kbf)
Kff <- NT%*%Dv0%*%N + lambda1*K1

A1<-matrix(c(Kbb,Kbd,Kbf),2,p+q+r1)
A2<-matrix(c(Kdb,Kdd,Kdf),2,p+q+r1)
A3<-matrix(c(Kfb,Kfd,Kff),r1,p+q+r1)

fisher <- rbind(A1,A2,A3)
se.f = sqrt(diag(ginv(fisher)))

#Matriz hessiana#


dmu01 <- (tau/(tau*Y + Y + tau*mu0))+((Y*(tau+1))/(4*(mu0^2)))-((tau^2)/(4*Y*(tau+1)))-(1/(2*mu0))
dmu02 <- (1/(2*(mu0^2)))-((tau^2)/((tau*Y + Y + tau*mu0)^2))-((Y*(tau+1))/(2*(mu0^3)))

ci <- numeric(c)
for (i in 1:c) {

ci[i] <- dmu02[i]*(b0[i]^2) + dmu01[i]*t(b0[i])*b0[i]

}

delta_ikro <- matrix(0,nrow = n,ncol = n, byrow = TRUE)
for (i in 1:c) {

delta_ikro[i,i] <- 1

}

Dc <- matrix(0, nrow = n, ncol = n, byrow = TRUE)
for (i in 1:n) {
for (j in 1:n) {

Dc[j,i] <- ci[i]*delta_ikro[j,i]

}
}

dm <- matrix(1,nrow=n,ncol=1)
dmu00 <- (Y/(tau*Y + Y + tau*mu0)^{2})+((Y/(4*(mu0^2))))-(((tau)*(tau+2))/(4*Y*(tau+1)^{2}))
Dmu00 <- diag(as.vector(dmu00))
ddel01 <- ((tau + mu0)/(tau*Y + Y + tau*mu0))-((Y)/(4*(mu0^2)))-((tau*mu0*(tau+2))/(4*Y*(tau+1)^{2}))+(tau/(2*(tau+1)))
ddel02 <- (1/(2*(tau+ 1)^{2}))-(((Y + mu0)^{2})/((tau*Y + Y + tau*mu0)^2))-((mu0)/(2*Y*(tau +1)^{3}))

bi <- numeric(c)
for (i in 1:c) {

ci[i] <- ddel02[i]*(b0[i]^2) + ddel01[i]*t(b0[i])*b0[i]

}

delta_ikro <- matrix(0,nrow = n,ncol = n, byrow = TRUE)
for (i in 1:c) {

delta_ikro[i,i] <- 1

}

Wc <- matrix(0, nrow = n, ncol = n, byrow = TRUE)
for (i in 1:n) {
for (j in 1:n) {

Wc[j,i] <- bi[i]*delta_ikro[j,i]

}
}

m <- (Y/((tau*Y + Y + tau*mu0)^2)) + (Y/((4*mu0)^2)) - ((tau*(tau+2))/(4*((tau+1)^2)*Y))
d <- (1/(2*((tau+1)^2)))-(((Y+mu0)^2)/((tau*Y + Y + tau*mu0)^2))-((mu0)/(2*(tau+1)^3)*Y)



lbb <- WT%*%Dc%*%W
lbd <- WT%*%Dmu00%*%Z
lbf <- WT%*%Dc%*%N

ldb <- t(lbd)
ldd <- ZT%*%Wc%*%Z



l1<-matrix(c(Kbb,Kbd,Kbf),2,r1+p+q)
```

```
l2<-matrix(c(Kdb,Kdd,Kdf),2,r1+p+q)
l3<-matrix(c(Kfb,Kfd,Kff),r1,r1+p+q)

H <- rbind(l1,l2,l3)
se.h<-sqrt(diag(ginv(H)))
L1 <- -ginv(H)
#Valores p#

zstatbeta  = beta_i / se.h[1:p]
pvalorbeta = 2 * pnorm(abs(zstatbeta), lower.tail = F)

#Bandas de confianza para las funciones suaves F1 y F2#
SD_BETA <- se.f[1:p]
SD_BETA <- se.f[1:q]

SD_F <- se.f[(r1+q+q):(r1+p+q)]

SD_fi <- se.f[p+q]
se.f[3]

Band1_F <- f_i + 2*SD_F
Band2_F <- f_i - 2*SD_F

mi<- min(f_i) -0.5
ma <- max(f_i) +0.5


#Bandas de confianza para F1#
plot(T1_0,f_i,type='l',col="black", ylim=c(mi,ma),sub="", xlab = "Viento", ylab="f(Viento)", las=1, main="",cex.lab=1.5,
cex.sub=1.4)
par(new=TRUE)
plot(T1_0,Band1_F,type='l',col="red",ylim=c(mi,ma), xlab = "", ylab="",axes = F)
par(new=TRUE)
plot(T1_0,Band2_F,type='l',col="red",ylim=c(mi,ma), xlab = "", ylab="",axes = F)

#Análisis residual

alpha   = sqrt(2 / tau)
bet_a   = as.vector((tau*mu0) / (tau + 1))


#Residuos estanzarizados#

dif  = Y - mu0
phi  = sqrt((2 * ((tau + 1) ^ 2)) / ((2 * tau) + 5))
raiz = sqrt(2 * mu0 * mu0)
res.est  = ((dif)*phi) / raiz
mean(res.est)
sd(res.est)
min(res.est)
max(res.est)

par(mfrow = c(1, 2))

plot(res.est,
type="p",
col="black",
bg="deepskyblue4",
pch=21,
lwd=1,
main="",
sub="(a)",
cex.main=1.4,
xlab="",
ylim=c(-4.5,3),
ylab="Residuos estandarizados",
axes=TRUE,
las=1,
bty="o",
cex.lab=1.5,
cex.sub=1.4)

abline(h=2.5*sd(res.est),lwd=1,col="black",lty=2)
abline(h=-2.5*sd(res.est),lwd=1,col="black",lty=2)
identify(res.est,cex=0.8, n= 4 )



dif  = Y - mu0
phi  = sqrt((2 * ((tau + 1) ^ 2)) / ((2 * tau) + 5))
raiz = sqrt(2 * mu0 * mu0)
res.est  = ((dif)*phi) / raiz
mean(res.est)
sd(res.est)
min(res.est)
max(res.est)


plot(res.est,
type="p",
col="black",
```

```
bg="gray",
pch=21,
lwd=1,
main="",
sub="(a)",
cex.main=1.4,
xlab="Index",
ylim=c(-4.5,3),
ylab="Residuals",
axes=TRUE,
las=1,
bty="o",
cex.lab=1.5,
cex.sub=1.4)

abline(h=2.5*sd(res.est),lwd=1,col="black",lty=2)
abline(h=-2.5*sd(res.est),lwd=1,col="black",lty=2)


par(mfrow = c(1, 2))



hist(res.est,
col = "gray",
main = "",
sub="(a)",
xlab = "Residuos estandarizados",
ylab = "Frecuencia")




qqnorm(res.est,col="black", xlab="Cuantil teórico",sub="(b)", ylab="Cuantil empírico", main="")
qqline(res.est,col="black")
lillie.test(res.est)


#Residuos propuestos por Jorgensen#

vJ  = (- 1 / (2 * (mu0 ^ 2))) + (tau ^ 2) / (((Y * tau) + Y  + (tau * mu0)) ^ 2) + ((tau + 1) / 2) * (Y / (mu0 ^ 3))
vmu = 1 / (((tau + 1) / tau) * Y + mu0) + (Y * (tau + 1)) / (4 * (mu0 ^ 2)) -((tau ^ 2) / (4 * tau + 4)) * (1 / Y) - 0.5 / mu0
rbJ = vmu / sqrt(vJ)

plot(rbJ,
type="p",
col="black",
bg="goldenrod3",
pch=21,
lwd=1,
main="",
cex.main=1,
sub="(b)",
ylim=c(-4.5,3),
xlab="",
ylab="Residuos de Jorgensen",
axes=TRUE,
las=1,
bty="o",cex.lab=1.5,
cex.sub=1.4)

abline(h=2.7*sd(rbJ),lwd=1,col="black",lty=2)
abline(h=-2.7*sd(rbJ),lwd=1,col="black",lty=2)
identify(rbJ, cex=0.8, n = 5)

hist(rbJ,
col = "gray",
main = "",
sub="(b)",
xlab = "Residuals",
ylab = "Frequency")

#QQ-plots#

qqnorm(rbJ,col="black",xlab="Theoretical quantile",sub="(b)", ylab="Empirical quantile", main="",cex.lab=1.5, cex.sub=1.4)
qqline(rbJ,col="black")
lillie.test(rbJ)


#Influencia Local
#Perturbación de la ponderación por casos
#Matrices

vt     = Y
vu     = mu0
vd     = tau
ve     = (-1/(2*vu)) + vd /((vt*vd) + vt + (vd*vu)) +
((vd+1)*vt)/(4*(vu^2)) - (vd^2)/(4*vt*(vd+1))

vb     = 1/2 - (1/2)*(vd+1)^(-1) + (vt+vu)/((vt*vd) + vt + (vd*vu)) -
(1/4)*(vt/vu) - (vd*(vd+2)*vu)/(4*vt*((vd+1)^2))
```

```
va=(Y+vu)/((vd*Y) + Y + (vd*vu)) - (Y/(4*vu)) - ((vd*(vd+2)*vu)/(4*((vd + 1)^{2})*Y)) + vd /(2*(vd+1))


Dz     = diag(as.vector(va))

ai <- W%*%beta_i
bi<-Z%*%alphas_i

D1     = diag(as.vector(ai))
D2     = diag(as.vector(ve))

D3     = diag(as.vector(bi))
D4     = diag(as.vector(va))

Deltaf = NT%*%Da0%*%De
Deltab<-WT%*%D1%*%D2
Deltad<-ZT%*%D3%*%D4

Deltacw    = rbind(Deltab,Deltaf,Deltad)

#Perturbación de la respuesta

phi = ((2*vd)+5)/((vd+1)^2)
vk = sqrt((vu^2)*phi)
Dk = diag(as.vector(vk))

vpsi = -(vd*(vd+1))/( ((vd*vt)  + vt  + (vd*vu) )^2) +
(vd+1)/(4*(vu^2)) +  ((vd^2)/(4*(vd+1)*(vt^2)))

Dpsi = diag(as.vector(vpsi))

vro  = (-vu/( ((vd*vt) + vt + (vd*vu))^2)) -
1/(4*vu) + (vd*(vd+2)*vu)/(4*(vt^2)*((vd+1)^2))

Dvro = diag(as.vector(vro))

Deltaf = NT%*%Da0%*%Dk%*%Dpsi
Deltab <- WT%*%D1%*%Dpsi%*%Dk
Deltad <- ZT%*%D3%*%Dvro%*%Dk

Deltares  = rbind(Deltab, Deltaf, Deltad)


caseweightspert = Deltacw #perturbación de la ponderación por casos
responsepert    = Deltares #Perturbación de la respuesta


#Generalized leverage#

phi = (2*((vd+1)^2))/((2*vd) +5)
vpsi = -(vd*(vd+1))/( ((vd*vt)  + vt  + (vd*vu) )^2) +
(vd+1)/(4*(vu^2)) +  ((vd^2)/(4*(vd+1)))*(1/(vt^2))
Dpsi = diag(as.vector(vpsi))
vro  = (-vu/( ((vd*vt) + vt + (vd*vu))^2)) +
1/(4*vu^2) + (vd*(vd+2)*vu)/(4*(vt^2)*((vd+1)^2))
Deltab = WT%*%Da0%*%Dpsi
Deltaf = NT%*%Da0%*%Dpsi
Deltad = t(as.matrix(vro))
Deltalev = rbind(Deltab,Deltaf, Deltad)

Bcw = function(I, M)
{
B =(t(Deltacw)%*%(I-M)%*%Deltacw)
return(B)
}

Bres = function(I, M)
{
B =(t(Deltares)%*%(I-M)%*%Deltares)
return(B)
}

#matrices auxiliares


Lbeta  = H[1:p,1:p]
Lf     = H[(p+2):(p+1+r1),(p+2):(p+1+r1)]
Ldelta  = H[1:q,1:q]

b11    = cbind(matrix(0, p, p), matrix(0, p, q),matrix(0,p,r1))
b12    = cbind(matrix(0, p, p), -Ldelta^(-1), matrix(0,p,r1))
b13    = cbind(matrix(0, r1, p), matrix(0,r1,q),-solve(Lf))
B1     = rbind(b11, b12, b13)    #parametro beta


b211   = cbind(-solve(Lbeta), matrix(0, p, q),matrix(0,p,r1))
b212   = cbind(matrix(0, q, p), matrix(0, q, q),matrix(0,q,r1))
b213   = cbind(matrix(0, r1, p), matrix(0,r1,q),-solve(Lf))
B2     = rbind(b211, b212, b213)  # parametro alpha


b311   = cbind(-solve(Lbeta), matrix(0, p, q),matrix(0,p,r1))
```

```
b312   = cbind(matrix(0, q, p), -Ldelta^(-1), matrix(0,q,r1))
b313   = cbind(matrix(0, r1, p), matrix(0,r1,q),matrix(0,r1,r1))
B3     = rbind(b311, b312, b313)  #función f

##perturbación de la ponderación por casos##

FPC1            = Bcw(L1, B1)#beta
autovmaxbPC     = eigen(FPC1)$val[1]
vetorpcbPC      = eigen(FPC1)$vec[,1]

FPC2            = Bcw(L1, B2)#alpha
autovmaxdPC     = eigen(FPC2)$val[1]
vetorpcdPC      = eigen(FPC2)$vec[,1]

FPC3            = Bcw(L1, B3)#F
autovmaxfPC     = eigen(FPC3)$val[1]
vetorpcfPC      = eigen(FPC3)$vec[,1]


vCiPC           = 2 * abs(diag(FPC1))
vCidPC          = 2 * abs(diag(FPC2))
vCifPC           = 2 * abs(diag(FPC3))

##perturbación de la respuesta##

FPR1 = Bres(L1, B1)#beta
autovmaxbPR = eigen(FPR1, symmetric = TRUE)$val[1]
vetorpcbPR  = eigen(FPR1, symmetric = TRUE)$vec[,1]

FPR2 = Bres(L1, B2)#alpha
autovmaxdPR = eigen(FPR2, symmetric = TRUE)$val[1]
vetorpcdPR  = eigen(FPR2, symmetric = TRUE)$vec[,1]

FPR3 = Bres(L1, B3)#f
autovmaxfPR = eigen(FPR3, symmetric = TRUE)$val[1]
vetorpcfPR  = eigen(FPR3, symmetric = TRUE)$vec[,1]

vCiPR      = 2 * abs(diag(FPR1))
vCidPR     = 2 * abs(diag(FPR2))
vCifPR     = 2 * abs(diag(FPR3))

#Plots leverage, dmax y Cii#

#Leverage#

plot(yest,GL,
type="p",
col="black",
bg="gray",
cex = 0.7,
pch=21,
lwd=1,
main="",
cex.main=1,
xlab="fv",
ylab="al",
cex.lab=1.5,
cex.sub=1.4,
axes=TRUE,
las=1,
bty="o")

identify(yest, GL, cex=1.5, n = )

###perturbación de la ponderación por casos##

#dmax

infl1  = vetorpcbPC
dmaxG1 = abs(infl1)

infl2  = vetorpcdPC
dmaxG2 = abs(infl2)

infl3  = vetorpcfPC
dmaxG3 = abs(infl3)

par(mfrow = c(1, 1))
plot(1:length(dmaxG1), dmaxG1, xlab = "id", ylim = c(0, 1), ylab = "dmaxpcbeta", sub = "", cex = 0.7,bg="green",pch=21)
identify(1:length(dmaxG1), dmaxG1, n = 7 )

par(mfrow = c(1, 1))
plot(1:length(dmaxG2), dmaxG2, xlab = "id", ylim = c(0, 1), ylab = "dmaxpcalpha", sub = "", cex = 0.7,bg="green",pch=21)
identify(1:length(dmaxG2), dmaxG2, n = 6)

plot(1:length(dmaxG3), dmaxG3, xlab = "id", ylim = c(0, 1), ylab = "dmaxpcf", sub = "", cex = 0.7,bg="yellow",pch=21)
identify(1:length(dmaxG3), dmaxG3, n = 2)

#Ci

Cb1 = vCiPC
Cb1 = Cb1 / sum(Cb1)
```

```
Cb2 = vCidPC
Cb2 = Cb2/sum(Cb2)

Cb3 = vCifPC
Cb3 = Cb3/sum(Cb3)

par(mfrow = c(1, 1))

limi = 2 * mean(Cb1)
plot(1:length(Cb1), Cb1,xlab = "id", ylab = "Cbeta", ylim =c(0, 0.3), sub = "", cex = 0.7,bg="gray",pch=21,cex.lab=1.5,
cex.sub=1.4)
abline(h = limi, lty = 2, lwd = 2)
identify(1:length(Cb1),cex=0.8, Cb1, n = 11)

limi = 2 * mean(Cb2)
plot(cex.lab=1.5,
cex.sub=1.4,1:length(Cb2), Cb2, xlab = "id", ylab = "Calpha", sub = "", cex = 0.7,bg="gray",pch=21)
abline(h = limi, lty = 2, lwd = 2)
identify(1:length(Cb2),cex=0.8, Cb2, n = 9)

limi = 2 * mean(Cb3)
plot(cex.lab=1.5,
cex.sub=1.4,1:length(Cb3), Cb3, xlab = "id", ylab = "Cf",sub = "", cex = 0.7,bg="gray",pch=21)
abline(h = limi, lty = 2, lwd = 2)
identify(1:length(Cb3),cex=0.8, Cb3, n =14)

###perturbación de la respuesta##

#dmax

infl1  = vetorpcbPR
dmaxG1 = abs(infl1)

infl2  = vetorpcdPR
dmaxG2 = abs(infl2)

infl3  = vetorpcfPR
dmaxG3 = abs(infl3)



plot(1:length(dmaxG1), dmaxG1,  xlab = "id", ylim = c(0, 0.5), ylab = "dmaxpcbeta", sub = "", cex = 0.7,bg="red",pch=21)
identify(1:length(dmaxG1), dmaxG1, n =12 )

plot(1:length(dmaxG2), dmaxG2,  xlab = "id", ylim = c(0, 0.5), ylab = "dmaxpcalpha", sub = "", cex = 0.7,bg="green",pch=21)
identify(1:length(dmaxG2), dmaxG2, n = 9)

plot(1:length(dmaxG3), dmaxG3,  xlab = "id", ylim = c(0, 0.5), ylab = "dmaxpcf", sub = "", cex = 0.7,bg="yellow",pch=21)
identify(1:length(dmaxG3), dmaxG3, n = 7 )

#Ci

Cb1 = vCiPR
Cb1 = Cb1 / sum(Cb1)

Cb2 = vCidPR
Cb2 = Cb2 / sum(Cb2)

Cb3 = vCifPR
Cb3 = Cb3 / sum(Cb3)



limi = 2 * mean(Cb1)
plot(cex.lab=1.5,
cex.sub=1.4,1:length(Cb1), Cb1,  xlab = "id", ylim =c(0, 0.3),  ylab = "Cbeta",sub = "", cex = 0.7,bg="gray",pch=21)
abline(h = limi, lty = 2, lwd = 2)
identify(1:length(Cb1),cex=0.8, Cb1, n = 15 )



limi = 2 * mean(Cb2)
plot(cex.lab=1.5,
cex.sub=1.4,1:length(Cb2), Cb2,  xlab = "id", ylim =c(0, 0.3), ylab = "Calpha",sub = "", cex = 0.7,bg="gray",pch=21)
abline(h = limi, lty = 2, lwd = 2)
identify(1:length(Cb2),cex=0.8, Cb2, n = 14)




limi = 2 * mean(Cb3)
plot(cex.lab=1.5,
cex.sub=1.4,1:length(Cb3), Cb3,  xlab = "id", ylab = "Cf", ylim =c(0, 0.3), sub = "", cex = 0.7,bg="gray",pch=21)
abline(h = limi, lty = 2, lwd = 2)
identify(1:length(Cb3),cex=0.8, Cb3, n = 19)

###########Para influencia local###########


100*(abs((4.099949 - beta_i[1])/4.0999485))
100*(abs((0.000236043- se.f[1])/0.000236043))
```

```
100*(abs((0.4211408 - beta_i[2])/0.4211408))
100*(abs((0.01903877- se.f[2])/0.01903877))

100*(abs((-15.4112 - alphas_i[1])/-15.4112))
100*(abs((0.00000701- se.f[3])/0.00000701))

100*(abs((0.92341 - alphas_i[2])/0.92341))
100*(abs((0.000729- se.f[4])/0.000729))
```

Bibliografía

Birnbaum, Z. W., & Saunders, S.C. (1969a). A new family of life distributions. *Journal of Applied Probability*. **6**: 319–27.

Buja A, Hastie T, Tibshirani R. (1989). Linear smoothers and additive models. Annals of Statistics. 17:453-555.

Cárcamo, E., Marchant, C., & Ibacache-Pulgar, G. (2021). Birnbaum- Saunders semiparametric additive model.

Cook, R. D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society B*, **48**: 133-169.

Cook, R. & Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression.*Biometrika*, **70**, 1–10.

Chen, F., Zhu, H.-T., Song, X.-Y., & Lee, S.-Y. (2010). Perturbation Selection and Local Influence Analysis for Generalized Linear Mixed Models. *Journal of Computational and Graphical Statistics*, 19(4), 826–842.

De Bastiani, F., Cysneiros, A., Uribe-Opazo, M., Galea, M. (2014). Influence diagnostics in elliptical spatial linear models. *TEST*,

Eubank, R.L. (1984). The hat matrix for smoothing splines. *Statistics and Probability Letters*. **2**, 9–14.

Ferrari, S. Espinheira, P. Cribari-Neto, F. (2011). Diagnostic tools in beta regression with varying dispersion. *Statistica Neerlandica*, **65**, 337–351.

Galea, M., Paula, G.A., & Leiva, V. (2004). Influence diagnostics in log–Birnbaum–Saunders regression models. *Journal of Applied Statistics*. **31**, 1049–1064.

Gannaz, I. (2007). Robust estimation and wavelet thresholding in partially models. *Statistics and Computing*. **17**, 239–310.

González, A. (2020). *La ciencia y los datos para el medio ambiente, al alcance de todos los profesionales*. Comunidadism. Recuperado el 09 de mayo del 2020, desde: http://www.comunidadism.es/blogs/la-ciencia-y-los-datos-para-el-medio-ambiente-al-alcance-de-todos-los-profesionales.

Grantz, DA., Garner, J.H. y Johnson, D.W. Ecological effects of particulate matter, Environment International, 29(2), 213-239 (2003)

Green, P.J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review*. **55**, 245–259.

Green, P. & Silverman, B.(1990) On use of the EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society*, 443-452.

Green, P., & Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models: A roughness penalty*, Chapman and Hall/CRC.

Hastie, T. e Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, **82**: 371-386.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. Chapman and Hall.

Hastie, T., & Tibshirani, R. (1993). Varying-Coefficient Models. *Journal of the Royal Statistical Society*, **55**: 757-796.

Heckman, N. (1986). Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society*. **48**, 244–248.

Ibacache-Pulgar, G., Paula, G. A., & Galea, M. (2012). Influence diagnostics for elliptical semiparametric mixed models. *Statistical Modelling*, **12**:165-193.

Ibacache-Pulgar, G., Paula, G. (2011). Local influence for Student-t partially linear models. *Computational Statistics and Data Analysis*. **55**: 1462–1478.

Ibacache-Pulgar, G., Figueroa-Zúñiga, J., & Marchant, C. (2021). Semiparametric additive beta regression models: inference and local influence diagnostics. *REVSTAT* (accepted).

Leiva ,V., Barros, M., Paula, G.A., & Sanhueza, A. (2008). Generalized Birnbaum–Saunders distributions applied to air pollutant concentration. *Environmetrics*. **19**, 235–249.

Leiva, V., Santos-Neto, M., Cysneiros, FJA, Barros, M. (2014). Birnbaum–Saunders statistical modelling: A new approach. *Statistical Modelling*. **14**: 21–48.

Leiva, V., Santos-Neto, M., Cysneiros, F.J.A., & Barros, M. (2014). Birnbaum–Saunders statistical modelling: A new approach. *Statistical Modelling*, 14: 21–48.

Lin, J. Zhu, L. Xie, F. (2009). Heteroscedasticity diagnostics for t linear regression models. *Metrika*, **70**, 59–77.

Marchant, C., Leiva, V., Cysneiros, F.J.A., & Vivanco, J.F. (2016). Diagnostics in multivariate Birnbaum-Saunders regression models. *Journal of Applied Statistics*. **43(15)**: 2829-2849.

Ministerio del Medio Ambiente. (2021). *Sistema de Información Nacional de Calidad del Aire (SINCA)*. Recuperado el 08 de mayo del 2021, desde: https://sinca.mma.gob.cl/.

Paula, G.A., Leiva, V., Barros, M., & Liu, S. (2012). Robust statistical modeling using the Birnbaum–Saunders-t distribution applied to insurance. *Applied Stochastic Models in Business and Industry*, **28**, 16–34.

Paula, G. (2013). On diagnostics in double generalized linear models. *Computational Statistics and Data Analysis*, **68**, 44–51.

Poon, W. y Poon, Y. S. (1999). Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical Society B*, **61**: 51-61.

Rieck, J. & Nedelman, J. (1991). A log-linear model for the Birnbaum- Saunders distribution. *Technometrics*, **3**, 51–60.

Rigby y Stasinopoulos. (2005). Generalized additive models for location, scale and shape. *Applied Statistics*, **54**:507-554.

Rocha, A. Simas, A. (2011). Influence diagnostic in a general class of beta regression models. *TEST*, **20**, 95–119.

Santos–Neto M, Cysneiros FJA, Leiva V, Ahmed SE (2012) On new parametrizations of the Birnbaum–Saunders distribution. *Pakistan Jounal of Statistics*. **1**, 1–26.

Santos-Neto, M., Cysneiros, F.J.A., Leiva, V., & Barros, M. (2014). On a reparameterized Birnbaum-Saunders distribution and its moments, estimation and applications. *REVSTAT–Statistical Journal*, **12(3)**: 247-272.

Santos-Neto, M. Cysneiros, F. Leiva, F. Barros, M. (2016). Reparameterized Birnbaum-Saunders regression models with varying precision. *Electronic Journal of Statistics*. **10**, 2825-2855.

Silverman, B. (1985). Some aspects of the Spline Smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society*. **47**:1-52.

Simas, A. Barreto-Souza, W. Rocha, A. (2010). Improved estimators for a general class of beta regression models. Computational Statistics and Data Analysis, 54 348–366.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society* . **50**, 413–436.

Taylor, J. & Verbyla, A. (2004). Joint modeling of location and scale parameters of the t distribution. Statistical Modelling, 4 91–112.

Van Keilegom, I. & Wang, L. (2010). Semiparametric modeling and estimation of heteroscedasticity in regression analysis of cross-sectional data. *Electronic Journal of Statistics*, **4**, 133–160.

Vilca, F. Sanhueza, A. Leiva, V. & Christakos, G. (2010). An extended Birnbaum–Saunders model and its application in the study of environmental quality in Santiago, Chile. *Stochastic Environmental Research and Risk Assessment*. **24**, 771–782.

Xie, F. & Wei, B. (2007). Diagnostics analysis for log-Birnbaum- Saunders regression models.*Computational Statistics and Data Analysis*, **51**, 4692–4706.

Zhu, Z.Y., He, X., Fung, W.K. (2003). Local influence analysis for penalized Gaussian likelihood estimators in partially linear models. *Scandinavian Journal of Statistics*. **30**, 767–780.