Universidad de Valparaíso
CHILE

Facultad de Ciencias
Instituto de Estadística
Ingeniería en Estadística

Modelo con coeficiente variando parcialmente Birnbaum-Saunders reparametrizado

Trabajo de titulación para optar al:

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

Michelle Karina Osorio Órdenes

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

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

Valparaíso, Chile, 28 de diciembre de 2021

Índice general

Índice de figuras

Lista de Tablas

Algunas palabras

A punto de finalizar esta gran etapa, quisiera reconocer a todos aquellos quienes me han acompañado. Primero a mi familia y amigos, quienes siempre me dieron su apoyo incondicional en los momentos más complejos y me permitieron completar este proceso. En particular, a mi mamá y a mi hermano, pilares fundamentales en mi vida. Quisiera agradecer también a mi profesor guía, el Dr. Germán Ibacache–Pulgar y a mi profesora co–guía, la Dra. Carolina Marchant, quienes con su conocimiento, habilidad y buena disposición me permitieron concluir exitosamente el tema desarrollado en este proyecto. Ambos aportaron saberes invaluables en mi formación académica. Además, me gustaría dar las gracias a cada profesor y compañero que fue parte de este camino recorrido, en especial a mi amiga y compañera Marcela, de quienes me llevo muchas experiencias y herramientas que podré aplicar en el futuro. 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 con coeficiente variando parcialmente surgen como una alternativa para modelar el efecto de interacción no lineal entre una variable de respuesta y un conjunto de covariables provenientes de diversas áreas de investigación. En este trabajo se propone un modelo estadístico basado en la distribución Birnbaum–Saunders reparametrizada, cuya componente sistemática permite que los coeficientes de regresión varíen suavemente debido al efecto de alguna(s) covariable(s). Para obtener las estimaciones de máxima verosimilitud penalizada de los parámetros asociados al modelo, se propone el algoritmo Scoring de Fisher y Back–fitting ponderado basados en suavizamiento spline cúbico. Sumado a esto, se desarrollará un análisis de residuos e influencia local, con el objeto de evaluar la potencial influencia que pueden ejercer algunas observaciones en el ajuste del modelo. Finalmente, se presentará una aplicación del modelo propuesto a un conjunto de datos reales de contaminación atmosférica.

Palabras clave: modelo aditivo semiparamétrico, distribución Birnbaum–Saunders, técnica de influencia local, estimadores de máxima verosimilitud penalizada, Scoring de Fisher, algoritmo de Back–fitting ponderado.

Capítulo 1

Introducción

1.1. Planteamiento del problema

El interés por estudiar diversos fenómenos ambientales y el comportamiento de los datos asociados a dichas áreas, ha impulsado la constante investigación y propuesta de nuevos modelos, en adición a los avances tecnológicos en simulación y recursos computacionales. Dada la situación medioambiental crítica, debido a la contaminación atmosférica, que atraviesa todo el mundo y sus repercusiones en la calidad de vida (UNEP, 2019) es que se hace imperativa la inclusión de la Estadística como una ciencia que aporte en el análisis de las variables asociadas a dicha problemática. Una alternativa correspondería al modelamiento semiparamétrico, específicamente, los modelos con coeficiente variando parcialmente (MCVP), los cuales entregan flexibilidad a la hora de incorporar variables explicativas y cuya relación de interacción entre las covariables no es necesariamente lineal. En adición, los procedimientos de estimación y diagnóstico en modelos semiparamétricos ha permitido el estudio de dichas variables, surgiendo de esta manera como una potente herramienta para modelar sus efectos, ver Ibacache–Pulgar *et al.* (2012, 2013, 2021). Por esta razón, el estudio de los modelos de regresión lineal clásicos ha presentado un nuevo enfoque, con la finalidad de sustituir la estructura de regresión paramétrica por funciones suavizadas no paramétricas, las cuales ofrecen nuevas alternativas al proceso del modelamiento estadístico.

Dentro de la literatura clásica en el estudio de los modelos de regresión lineales, destacan Hastie y Tibshirani (1990), quienes describen una generalización de los modelos de regresión lineal; el modelo aditivo, cuya idea central es reemplazar la función lineal habitual de una covariable por una función suave no especificada y sumar dichas funciones. Este modelo no es paramétrico en el sentido de imponer una forma paramétrica a las funciones, sino por el hecho de estimarlas por un proceso iterativo mediante el uso de de suavizadores en gráficos de dispersión. El modelo estimado consta de una función para cada una de las covariables, lo que es útil para un modelo predictivo sin perder la interpretabilidad de un modelo de regresión clásico. En esta línea de autores se encuentran Green y Silverman (1994), los que proponen una visión personalizada del método de penalización por rugosidad, el cual entrega un enfoque unificador para la amplia gama de problemas de suavizado existentes, ya que actúa como el punto de unión entre la estadística clásica y paramétrica, donde el suavizado de funciones ya no es solamente incorporada en la regresión, sino que también aborda problemáticas en el modelamiento lineal generalizado. Birnbaum y Saunders (1969) propusieron un modelo estadístico para los fenómenos de fatiga de estructuras o sistemas, basados en la Ley de

Miner o de daño acumulativo, los cuales al ser sometidos a esfuerzos cíclicos eventualmente alcanzan un umbral que desencadenaría un colapso, considerando de esta manera una sucesión finita de variables aleatorias independientes e idénticamente distribuidas, debido a que la suma de éstas superará el valor límite que generaría aquella falla. No obstante, hoy en día, trabajos recientes han aplicado este modelo a diversas áreas, considerándolo prácticamente como una distribución de probabilidad, más que encasillarlo en un modelamiento de datos de tiempo de vida; Leiva *et al.* (2011). Asimismo, ha sido fuente de estudio para la resolución de problemáticas de contaminación atmosférica; véase Leiva *et al.* (2008, 2015), Vilca *et al.* (2010), Marchant *et al.* (2018), Cavieres *et al.* (2020) y Puentes *et al.* (2021).

Dado que en la literatura no existen muchos estudios que involucren a los modelos semiparamétricos basados en la distribución Birnbaum–Saunders reparametrizada (BSR), la implementación de un modelo con coeficiente variando parcialmente bajo dicha distribución es un aporte a la teoría del modelamiento estadístico, en particular, para llevar a cabo el análisis de datos de contaminación atmosférica.

Sumado a lo anterior, este trabajo de título se justifica principalmente por su propuesta teórica, dado que la forma de abordar el proceso de estimación de parámetros, inferencia y diagnóstico del MCVP-BSR son una propuesta novedosa y desafiante, así como también la importancia de su aplicación en problemas relacionados a las ciencias ambientales, un tópico de alta relevancia en Chile y Latinoamérica, sobretodo considerando los altos índices de contaminación a los que se encuentran expuestos los habitantes en numerosas regiones de nuestro país.

1.2. Preguntas de investigación

1. ¿Es posible incorporar en la componente sistemática de la distribución Birnbaum–Saunders reparametrizada una estructura semiparamétrica?

2. ¿Es posible estimar los parámetros del modelo utilizando suavizamiento spline cúbico y P–spline, y utilizar técnicas de diagnóstico para evaluar la sensibilidad de los estimadores?

3. ¿Es posible aplicar el MVCP–BSR a variables del área de la contaminación atmosférica?

1.3. Hipótesis

1. Es factible la estimación de parámetros y el estudio de ciertos aspectos de la inferencia estadística bajo el modelo propuesto.

2. Es viable la aplicación de algunas técnicas de diagnóstico para evaluar la influencia de los datos en el ajuste del modelo.

1.4. Objetivo general

Proponer un proceso de estimación para los parámetros y desarrollar algunas técnicas de diagnóstico en el MCVP–BSR.

1.5. Objetivos específicos

1. Estudiar aspectos teóricos del MCVP–BSR.
2. Derivar un proceso iterativo para obtener las estimaciones de los parámetros del modelo propuesto.
3. Realizar un análisis de diagnóstico del modelo basado en los residuos y la técnica de influencia local.
4. Implementar computacionalmente los resultados teóricos y aplicar el modelo propuesto a datos medioambientales.

1.6. Sobre contaminación atmosférica

Con relación a los estudios de contaminación atmosférica y calidad del aire, esta problemática aqueja a millones de habitantes en todo el mundo. Según la OPS (2016), cerca de 7 millones de muertes prematuras fueron atribuibles a la contaminación del aire ambiental, además, alrededor del 88 % de estas muertes ocurrieron en países de ingresos bajos y medios en el año 2016. Otra cifra alarmante es que más de 150 millones de personas en América Latina viven en ciudades cuyas concentraciones de contaminantes exceden a las establecidas por las Guías de Calidad del Aire de la OMS.

El Sistema Nacional de información Medioambiental (SINIA) indica que la contaminación del aire es uno de los principales problemas urbanos, ya que su constante exposición deteriora la salud de las personas, perjudica la vegetación, altera la vida silvestre, modifica los suelos, daña y debilita la estructura de ciertos materiales y favorece significativamente el calentamiento global. Según la naturaleza de la fuente emisora, el origen de los contaminantes del aire puede ser clasificado en biogénico o antropogénico, siendo este último la principal causa en nuestro país, donde se ha determinado la existencia de 3 fuentes de contaminación: las actividades industriales, los medios de transporte y el uso de leña para la calefacción de los hogares. Igualmente, algunos sectores dedicados a la producción también han contribuido a generar problemas de contaminación en distintas localidades del país (MMA, 2016).

Cabe señalar que en la atmósfera terrestre existen distintos tipos de contaminantes, dentro de los cuales se pueden considerar el polvo, la emanación de gases y el humo. Uno de estos elementos nocivos corresponde al material particulado (MP) que se define como aquellas partículas líquidas o sólidas que se encuentran en suspensión, siendo posible clasificarlas según su diámetro, característica de la cual depende la intensidad de sus impactos, ya que tiene directa relación con el lugar de las vías respiratorias que alcancen. En Chile se utilizan dos métricas de clasificación: partículas de diámetros menores a 10 micrones conocidas como MP10 (gruesas) y de diámetros menores a 2,5 micrones conocidas como MP2.5 (finas), siendo este último el contaminante más dañino para la

salud de la población. Es importante considerar que las partículas más gruesas no comprometen los órganos respiratorios, debido a que se depositan en el tracto respiratorio superior y son expulsadas por la acción de los cilios. Por otra parte, el MP2.5 está conformado por partículas lo suficientemente pequeñas como para penetrar en las vías respiratorias, llegar a los alveolos pulmonares e ingresar directamente al torrente sanguíneo, lo que genera mayores niveles de mortalidad prematura en la población, padecimiento de enfermedades respiratorias y cardiovasculares. Al mismo tiempo, está altamente relacionada con el aumento de las admisiones hospitalarias en invierno, afectando principalmente a adultos mayores, a menores de 8 años y a pacientes que presentan problemas de salud de carácter crónico (MMA, 2011).

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 (`https://bit.ly/2u40gDq`). Su capital Santiago, es una de las ciudades más contaminadas del mundo en términos de MP2.5 y MP10, debido a una combinación de factores meteorológicos, antropogénicos y topográficos (Marchant *et al.*, 2013). No obstante, existen otras ciudades de nuestro país que presentan graves problemas de contaminación atmosférica, principalmente entre el 1 de abril y el 31 de agosto de cada año. Estas ciudades son Coyhaique, Linares, Osorno, Padre las Casas, Puerto Montt, Rancagua, Temuco y Valdivia, que se encuentran entre las 10 ciudades sudamericanas más contaminadas, según un informe de Greenpeace y AirVisual que mide el índice de calidad del aire en base a los niveles de MP2.5 (`https://bit.ly/2TAwI0P`).

En algunas ocasiones, pueden producirse episodios periódicos de contaminación extrema con determinados contaminantes, los cuales fluctúan respecto de las condiciones meteorológicas y geográficas. Como resultado de esta variación, los niveles de contaminantes atmosféricos se tratan como variables aleatorias con soporte positivo, que pueden modelarse a través de una distribución de probabilidad sesgada a la derecha, véase Cavieres *et al.* (2020). En adición, existen diversos estudios previos a nivel mundial, que estudian la relación entre las variables meteorológicas y las partículas contaminantes. En la actualidad existen más de 140 estaciones de monitoreo a lo largo de todo Chile, las cuales se clasifican según el tipo de origen; las hay privadas, que corresponden a las instaladas como exigencia de Planes de Descontaminación de fundiciones Mineras y las hay públicas, que fueron las instaladas por MINSAL y CONAMA.

Según el Sistema de Información Nacional de Calidad del Aire (SINCA), el monitoreo en línea de las concentraciones ambientales de MP10 y MP2.5 se presenta como promedios móviles de 24 horas, basados en el monitoreo continuo de este contaminante. Además, la información mostrada permite identificar los eventos asociados a situaciones de emergencia ambiental para las últimas 24 horas a partir del momento de consulta. Los datos son actualizados cada 1 hora y los registros pueden variar una vez validados operacionalmente por el operador de la estación. En el presente trabajo se estudiarán datos de la comuna de Pudahuel, localizada en la Región Metropolitana de Chile. Sumado a esto, se considerarán como covariables predictivas a aquellas variables que más resaltan en la literatura consultada.

Capítulo 2

Modelo con coeficiente variando parcialmente Birnbaum–Saunders reparametrizado

En este capítulo se introduce el modelo con coeficiente variando parcialmente Birnbaum–Saunders reparametrizado (MCVP–BSR). En la Sección 2.1 se hace una breve revisión bibliográfica de los MCVP y la distribución BSR. En las secciones 2.2 y 2.3 se realiza una descripción del MCVP y de la distribución BSR, respectivamente. En la Sección 2.3 se presenta el modelo propuesto, la representación matricial del componente sistemático y también cómo el MCVP surge como una extensión de otros modelos que se han propuesto anteriormente en la literatura. Finalmente, en la Sección 2.4, se presenta la función de verosimilitud penalizada del modelo propuesto, asumiendo que las funciones suaves pertenecen al espacio de funciones de Sobolev.

2.1. Introducción

Los autores Hastie y Tibshirani (1993) estudiaron una clase de modelos de regresión en que los coeficientes varían como funciones suaves de otras variables, es decir, modelos de generalizaciones aparentemente diferentes que son lineales en los regresores, pero sus coeficientes pueden cambiar con el valor de otras variables. Esta clase de modelos une los modelos aditivos generalizados y los modelos lineales dinámicos en un marco común y entrega una extensión potencialmente útil de los modelos de regresión. Algunas posibles extensiones hacen referencia a la inclusión de modelos de regresión no lineal o a la expansión de la dirección en el espacio del modificador de efecto, lo que resultaría en grandes cambios en los coeficientes.

Años más tarde, Jiang *et al.* (2013) proponen un nuevo modelo de coeficientes variables denominado modelo de coeficientes variables principales, esto con el objetivo de mejorar las estimaciones de un MCV, especialmente cuando p es grande. En esta propuesta se caracterizan los coeficientes variables a través de combinaciones lineales de algunas funciones principales, con lo que se reduce el número real de funciones no paramétricas, independiente del método de suavizado, motivo por el cual tendría una mejor eficiencia en la estimación. Se aplica la estimación con la penalización L_1 y se comprueba que presenta buenos resultados incluso si el número de covariables es muy grande.

Tiempo después, Ibacache–Pulgar y Reyes (2018) extendieron los modelos de coeficiente variable con errores normales a errores elípticos para permitir distribuciones con colas más pesadas y ligeras que las normales, trabajando particularmente con una Student-.t, que es una distribución continua simétrica de contorno elíptico. En adición, desarrollaron los enfoques de influencia local para el modelo propuesto bajo perturbaciones de ponderación de caso, parámetro de escala y de la variable explicativa. Finalmente, aportaron evidencias sobre los aspectos robustos que poseen los estimadores de máxima verosimilitud penalizados de MCVP en una Student–t con grados de libertad pequeños frente a observaciones atípicas, tal y como señalan Ibacache–Pulgar *et al.* (2013).

La distribución Birnbaum–Saunders (BS) ha atraído una atención considerable en la literatura estadística durante el último período, debido a sus argumentos teóricos físicos, atractivas propiedades y su relación con el modelo normal, ver Santos–Neto *et al.* (2014). Algunos de los investigadores que se vieron involucrados, fueron por ejemplo; Rieck y Nedelman (1991), Galea *et al.* (2004), Paula *et al.* (2012), Marchant *et al.* (2016 a,b) y Dasilva *et al.* (2020). Respecto a la distribución BS de dos parámetros, Santos–Neto *et al.* (2012) estudiaron 11 diferentes parametrizaciones, analizando las propiedades estructurales de éstas, utilizaron el método de máxima verosimilitud para la estimación de los parámetros correspondientes y realizaron un estudio de simulación de Monte Carlo para detectar su rendimiento. La motivación de esta investigación radicó en la búsqueda de estimadores de parámetros insesgados y consistentes, hallar probabilidades de cobertura en los intervalos de confianza que estén cerca del nivel nominal, ortogonalidad de los parámetros y la posibilidad de describir directamente la media de una variable aleatoria de la BS a través de un modelo de regresión sin la necesidad de transformar la variable de respuesta. Dentro de los principales resultados se confirmó que los estimadores de máxima verosimilitud son asintóticamente insesgados y que las probabilidades de cobertura de los respectivos intervalos de confianza están cerca del nivel esperado. Un aspecto que se sugiere para estudios posteriores es la estimación de momentos para estas propuestas de parametrizaciones. Leiva *et al.* (2014) introducen un nuevo enfoque para los modelos de regresión de Birnbaum–Saunders, que permitió analizar los datos en su escala original y modelizar la varianza no constante. Además, propusieron cuatro tipos de residuos para dichos modelos y establecieron, a través de simulación, cuál de ellos tuvo un mejor rendimiento empírico. Sumado a esto, se desarrollaron métodos de influencia local calculando las curvaturas normales bajo diferentes esquemas de perturbación. Es importante mencionar que los modelos de regresión BS propuestos no requieren de una transformación logarítmica en los datos, lo que entregaría una ventaja debido a que no hay reducción en la potencia del estudio, así como tampoco dificultades en la interpretación de los resultados, por otra parte, son una alternativa mucho más flexible, ya que permiten el uso de diferentes funciones de enlace no negativas que relacionan la media con los regresores, en otras palabras, incorporaron una función de enlace para la media en el mismo sentido que los modelos lineales generalizados (GLM), basándose en la distribución BSR que no pertenece a la familia exponencial, logrando de esta manera relacionar la respuesta media con un predictor lineal mediante una de varias funciones de enlace posibles, las que contienen los parámetros a estimar.

Finalmente, dentro de las propuestas más actualizadas, Cárcamo *et al.* (2021) propone un modelo aditivo semiparamétrico basado en la distribución BSR (MAS–BSR), cuya metodología involucró un análisis de diagnóstico, la derivación del algoritmo Back–fitting para la obtención de estimaciones de máxima verosimilitud penalizada y el desarrollo de métodos de influencia local, calculando las curvaturas normales bajo diferentes esquemas de perturbación.

2.2. Modelos con coeficiente variando parcialmente

Según Ibacache–Pulgar y Reyes (2018) los MCVP se caracterizan por poseer una componente paramétrica y otra no paramétrica. En particular, el componente no paramétrico radica en una suma de variables regresoras cuyos coeficientes son funciones suaves arbitrarias unidimensionales, que cuantifican el efecto no paramétrico (no lineal) que poseen otras variables explicativas sobre la variable de respuesta. La parte paramétrica se compone de un vector de parámetros y un vector de variables explicativas.

Este modelo asume que la relación entre la variable de respuesta y las variables explicativas se puede representar como:

$$\mathrm{y}_i = \mathbf{z}_i^\top \boldsymbol{\alpha} + \mathrm{x}_{1_i}\, \beta_1(\mathrm{t}_{1_i}) + \ldots + \mathrm{x}_{s_i}\, \beta_s(\mathrm{t}_{s_i}) + \epsilon_i, \quad (i = 1, 2, \ldots, n), \tag{2.1}$$

donde y_i denota el valor de respuesta asociado con la i-ésima unidad experimental, $\mathbf{z}_i = (\mathbf{z}_{i_1}, ..., \mathbf{z}_{i_p})^\top$ es un vector de variables explicativas, $\boldsymbol{\alpha}$ es un vector $(p \times 1)$ de parámetros desconocidos, β_k $(k = 1, ..., s)$ son funciones arbitrarias suaves desconocidas de la variable explicativa t_k, asociadas con la covariable x_k, y ϵ_i es el error aleatorio.
Es importante señalar que el MCVP es una extensión de otros modelos que ya se han propuesto en la literatura, por ejemplo:
(i) cuando β_k son todas constantes, es decir, $\beta_k(t_k) = \beta_k$ para todo $k = 1, ..., s$, entonces el MCVP se reduce al modelo de regresión múltiple;
(ii) cuando $\boldsymbol{\alpha}$ toma el valor $\boldsymbol{\alpha} = 0$ y la covariable $\mathrm{x}_k \equiv 1$ para todo $k = 1, ..., s$, el MCVP se reduce al modelo aditivo propuesto, en particular, por Hastie y Tibshirani (1993);
(iii) cuando $\mathrm{x}_i \equiv 1$ y $\beta_k \equiv 0$ para todo $k = 3, \ldots s$, entonces el MCVP se reduce al modelo lineal parcial discutido, entre otros, por Green y Silverman (1994); y
(iv) cuando $\boldsymbol{\alpha} = 0$, $\mathrm{x}_1 \equiv 1$ y $\mathrm{x}_k \equiv 0$ para todo $k = 2, ..., s$, el MCVP se reduce al modelo de regresión no paramétrica discutido, por ejemplo, por Silverman (1985).

Para finalizar, mencionar que los MCVP se han convertido en una poderosa herramienta de modelado debido a la maleabilidad que ofrece para explorar las características dinámicas que pueden existir en los datos y su fácil interpretación. Además, desde el punto de vista de la modelización estadística, el MCVP permite que los coeficientes varíen suavemente sobre el grupo estratificado por t_k, permitiendo interacciones no lineales entre t_k y x_k.

2.3. Distribución Birnbaum–Saunders reparametrizada

Santos–Neto *et al.* (2012) desarrolló una reformulación para la distribución BS basada en los parámetros μ y δ, donde $\mu > 0$ es un parámetro de escala y la media de la distribución, mientras que $\delta > 0$ es un parámetro de forma y precisión; ver Santos–Neto *et al.* (2014) y Leiva *et al.* (2014). Con base en esta reformulación para la distribución BS, la función de densidad de probabilidad viene dada por:

$$f(y; \mu, \delta) = \frac{\exp(\delta/2)\sqrt{\delta+1}}{4\, y^{3/2}\sqrt{\pi\mu}} \left[y + \frac{\delta\mu}{\delta+1}\right] \exp\left(-\frac{\delta}{4}\left[\frac{y\{\delta+1\}}{\delta\mu} + \frac{\delta\mu}{y\{\delta+1\}}\right]\right), \quad y > 0. \tag{2.2}$$

En este caso, usamos la notación $Y \sim \text{BSR}(\mu, \delta)$. La media y la varianza de Y están dadas por $E[Y] = \mu$ y $V[Y] = \mu^2/\phi$, respectivamente, donde $\phi = [\delta + 1]^2/[2\delta + 5]$ tal que, como se ha mencionado δ puede interpretarse 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 $\delta \to 0$, entonces $V[Y] \to 5\mu^2$. Se puede ver que $V[Y] = \mu^2/\phi$ 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 $b\,Y \sim \text{BSR}(b\,\mu, \delta)$, con $b > 0$, y $1/Y \sim \text{BSR}(\mu^*, \delta)$, donde $\mu^* = [\delta + 1]/[\delta\mu]$.

2.4. Modelo propuesto

Sea $Y_1, \ldots, Y_n$ variables aleatorias independientes, donde $Y_i \sim \text{BSR}(\mu_i, \delta)$, para $i = 1, \ldots, n$ e $y = (y_1, ..., y_n)^\top$ las observaciones correspondientes. Entonces, el MCVP–BSR se basa en (2.2) mediante el componente sistemático:

$$h(\mu_i) = \eta_i = \mathbf{z}_i^\top \boldsymbol{\alpha} + \mathrm{x}_{0_i}\, \beta_0(\mathrm{t}_{0_i}) + \mathrm{x}_{1_i}\, \beta_1(\mathrm{t}_{1_i}) + \ldots + \mathrm{x}_{s_i}\, \beta_s(\mathrm{t}_{s_i}), \quad i = 1, 2, \ldots, n, \tag{2.3}$$

o, equivalentemente,

$$h(\mu_i) \;\; = \;\; \mathbf{z}_i^\top \boldsymbol{\alpha} + \tilde{\mathbf{n}}_{1_i}^\top \boldsymbol{\beta_1} + \ldots + \tilde{\mathbf{n}}_{s_i}^\top \boldsymbol{\beta_s}$$

donde $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_p)^\top$ para $p < n$, es un vector de parámetros desconocidos a estimar, $\mathbf{z}_i{}^\top = (1, \mathrm{z}_{i_2}, \ldots, \mathrm{z}_{i_p})$ representa los valores de p regresores, $\mu_i = h^{-1}(\mathbf{z}_i^\top \boldsymbol{\alpha} + \mathrm{x}_{1_i}\, \beta_1(\mathrm{t}_{1_i}) + \ldots + \mathrm{x}_{s_i}\, \beta_s(\mathrm{t}_{s_i}))$ con h^{-1} siendo la función inversa de h, $\beta_k(\cdot)$ son funciones arbitrarias suaves desconocidas de la variable explicativa t_{k_i}, asociada con las covariables x_{k_i}, para $k = 1, \ldots, s$, $\tilde{\mathbf{N}}_k = \mathbf{X}^{(k)}\mathbf{N}_k, \mathbf{X}^{(k)} = \text{diag}_{1 \le j \le m_i}(\mathrm{x}_i^{(k)})$, siendo $\mathbf{N}_k$ una matriz de incidencia $(n \times r_k)$ con el (j, l)–ésimo elemento igual a la función del indicador $I(\mathrm{t}_{k_i} = \mathrm{t}_{k_l}^0)$, acá $\tilde{\mathbf{n}}_{k_i}^\top$ denota la i-ésima fila de la matriz de incidencia $\tilde{\mathbf{N}}_k$, $\boldsymbol{\beta}_k = (\xi_{k_1}, \ldots, \xi_{k_{r_k}})^\top$ es un vector de parámetros $(r_k \times 1)$ tal que $\xi_{k_j} = \beta_k(\mathrm{t}_{k_l}^0)$, con $\mathrm{t}_{k_l}^0$ $(l = 1, \ldots, r_k)$ denotando los valores distintos y ordenados de la variable explicativa t_k. En el modelo dado en (2.3), la función de enlace h: $\mathbb{R} \rightsquigarrow \mathbb{R}^+$ es estrictamente monótona, positiva y al menos dos veces diferenciable; por ejemplo, $h(\mu) = \log(\mu)$ o $h(\mu) = \sqrt{\mu}$. Formalmente, tenemos que $V[Y_i]$ es una función de μ_i y, en consecuencia, de los regresores $\mathbf{z}_i$. Entonces, debido a que estamos modelando la media en base a una estructura particular, también estamos modelando la varianza debida a $\mathrm{V}[Y_i] = \mu_i^2/\phi$. Por lo tanto, los problemas en los que existe una varianza no constante podrían analizarse utilizando este modelo.

2.5. Función penalizada

La función de log–verosimilitud del MCVP–BSR dada en (2.3) para $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}_1^\top, ..., \boldsymbol{\beta}_s^\top, \delta)$ es:

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} L_i(\mu_i, \delta; y_i), \tag{2.4}$$

donde

$$L_i(\mu_i, \delta) = \frac{\delta}{2} - \frac{\log(16\pi)}{2} - \frac{1}{2}\log\left(\frac{[\delta + 1]y_i^3\mu_i}{[\delta\, y_i + y_i + \delta\mu_i]^2}\right) - \frac{y_i[\delta + 1]}{4\mu_i} - \frac{\delta^2\mu_i}{4[\delta + 1]y_i}. \tag{2.5}$$

En general, un problema de la maximización directa de $L(\boldsymbol{\theta})$, sin imponer restricciones sobre las funciones β_k's, es que conduce a un sobreajuste y a la no identificación de $\boldsymbol{\alpha}$. Un procedimiento que puede resolver este inconveniente y una alternativa para determinar los estimadores de β_k, consiste en incorporar un término de penalización sobre cada función suave β_k's denotado por $J(\beta_k)$. Si suponemos que β_k pertenece al espacio de funciones de Sobolev, es decir, β_k pertenece al conjunto de todas las funciones continuamente diferenciables sobre $[a_k; b_k]$ con segundas derivadas cuadradas integrables,

$$\mathcal{W}_2^{(2)} \quad = \quad \{\beta_k : \beta_k, f_k^{(1)}\text{abs. cont.}, \beta_k^{(2)} \in \mathcal{L}^2[a_k, b_k]\} \ ,$$

donde $\beta_k^{(2)}(\mathrm{t}_k) = \frac{\mathrm{d}^2}{\mathrm{dt}_k^2}\beta_k(\mathrm{t}_k)$, entonces el estimador de β_k maximiza la función de log–verosimilitud penalizada dada por

$$\ell_{\mathrm{p}}(\boldsymbol{\theta}, \lambda_1, \ldots, \lambda_s) \quad = \quad L(\boldsymbol{\theta}) + \sum_{k=1}^{s} \lambda_k^* J(\beta_k) \ , \tag{2.6}$$

sobre todas las funciones β_k en este conjunto, con $J(\beta_k)$ denotando la función de penalización sobre β_k. En este caso $\lambda_k^* = \lambda^*(\lambda_k)$ es una constante que depende del parámetro de suavización $\lambda_k \geq 0$, el cual depende de alguna aplicación específica. En la literatura podemos encontrar diferentes tipos de penalizaciones en función del método propuesto a las curvas no paramétricas. Se considera como medida de la curvatura de las funciones la norma al cuadrado definida por

$$J(\beta_k) = ||\beta_k||^2 = \int_{a_k}^{b_k} \beta_k^{(2)}(\mathrm{t}_k)^2 \mathrm{dt}_k. \tag{2.7}$$

El primer término del lado derecho de la ecuación (2.6) mide la bondad del ajuste, mientras que el segundo término, definido por (2.7), penaliza la rugosidad de cada β_k con un parámetro λ_k. En este caso, la estimación de β_k conduce a un spline cúbico natural con nodos en los puntos $\mathrm{t}_{k_l}^0$, es decir, a un polinomio de grado 3 a trozos en cada intervalo $[\mathrm{t}_{k_l}, \mathrm{t}_{k_{l+1}}]$, para $l = 1, 2, ..., r_k - 1$. De acuerdo con Green y Silverman (1994), $J(\beta_k)$ puede escribirse como

$$J(f_k) = \int_{a_k}^{b_k} [\beta_k^{(2)}(\mathrm{t}_k)]^2 \mathrm{dt}_k = \boldsymbol{\beta}_k^\top \mathbf{K}_k \boldsymbol{\beta}_k \ ,$$

donde $\mathbf{K}_k$ es una matriz $(r_k \times r_k)$ definida no negativa que depende sólo de los nodos t_k^0. Entonces, si consideramos $\lambda_k^* = -\lambda_k/2$, la función de log–verosimilitud penalizada puede expresarse como

$$\ell_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = L(\boldsymbol{\theta}) - \sum_{k=1}^{s} \frac{\lambda_k}{2} \ \boldsymbol{\beta}_k^\top \mathbf{K}_k \boldsymbol{\beta}_k \ , \tag{2.8}$$

donde $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_s)^\top$ denota un vector $(s \times 1)$ de parámetros de suavizado. Es importante señalar que la selección de estos parámetros de suavizado es un aspecto esencial en el proceso de modelado semiparamétrico, en particular durante la estimación, ya que controlan el equilibrio entre la bondad del ajuste y la suavidad (rugosidad) de la función estimada. 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 en inglés) y el error cuadrático medio (ECM).

Capítulo 3

Estimación de parámetros

En este capítulo se considerará el problema de estimación de parámetros asociado al MCVP-BSR. Inicialmente, en la Sección 3.1 se presenta una discusión sobre algunos trabajos que tratan el problema de estimación en el MCVP, en el modelamiento semiparamétrico y en la distribución BSR. En la Sección 3.2 se obtiene la función Score penalizada para $\boldsymbol{\theta}$. En las secciones 3.3 y 3.4, se calculan las matrices Hessiana y de información de Fisher penalizadas, respectivamente. En la Sección 3.5 se propone un procedimiento iterativo basado en los algoritmos de Scoring de Fisher y Back–fitting para resolver las ecuaciones de estimación asociadas con el coeficiente de regresión $\boldsymbol{\alpha}$ y las funciones suaves $\beta_k's$.

3.1. Introducción

El problema de la estimación en el contexto del MCVP–BSR no ha sido discutido en la literatura, sin embargo, algunos autores han abordado sus componentes por separado en modelos relacionados.Por ejemplo, respecto al modelo con coeficiente variando, Hastie y Tibshirani (1993) estudiaron los modelos aditivos generalizados en 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. Cai *et al.* (2000) estimaron las funciones del coeficiente basándose en la técnica de regresión polinómica local y propusieron un método que implica la resolución de cientos de ecuaciones de verosimilitud local mediante la aplicación de un algoritmo Newton-Raphson de un solo paso. Chiang *et al.* (2001) derivaron un procedimiento para la estimación de splines de suavización para modelos de coeficientes variables considerando variables dependientes medidas repetidamente; véase también Eubank (2004). Por otra parte, los autores Liu y Li (2015) estimaron las curvas de coeficiente en un modelo de coeficiente variable para datos longitudinales utilizando el método de suavización polinómica local y mostraron que el estimador resultante es asintóticamente más eficiente que los que ignoran la estructura de correlación intra–sujeto. Ibacache–Pulgar y Reyes (2018), estimaron las curvas de coeficiente de un MCVP bajo una distribución simétrica de contorno elíptico Student-t, basándose en el criterio de la verosimilitud penalizada y en el smoothing spline. Con relación al modelamiento semiparamétrico, Ibacache–Pulgar *et al.* (2013) estudiaron un modelo de este tipo bajo distribuciones simétricas, estimando el coeficiente de regresión y las funciones suaves a través de un algoritmo de Back–fitting ponderado, lo que condujo a un spline cúbico como solución para

las funciones no paramétricas. En este mismo contexto y más recientemente, Ibacache–Pulgar *et al.* (2021) estudiaron un modelo de regresión beta aditivo y obtuvieron las estimaciones de máxima verosimilitud penalizada, calcularon su correspondiente función de puntuación y desarrollaron un proceso iterativo para estimar sus parámetros. Para el caso del modelo de regresión BSR, los estimadores de máxima verosimilitud de los parámetros no pueden obtenerse de forma explícita y deben obtenerse resolviendo una ecuación no lineal. Leiva *et al.* (2014) utilizan el método de Scoring de Fisher para estimar los parámetros del modelo. Cárcamo *et al.* (2021) propusieron ajustar el MAS–BSR estimando conjuntamente el coeficiente de regresión, las funciones suaves y el parámetro de precisión basado en el criterio de la verosimilitud penalizada.

Se propone ajustar el MCVP–BSR basándose en los trabajos propuestos por Leiva *et al.* (2014) e Ibacache–Pulgar *et al.* (2013), con el fin de estimar conjuntamente el coeficiente de regresión, las funciones suaves y el parámetro de precisión basado en la verosimilitud penalizada y Smoothing spline.

3.2. Funciones score penalizadas

Asumiendo que la función (2.8) es regular con respecto a $\boldsymbol{\alpha}$, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_s$ y δ, se tiene que el vector de la función score penalizada de $\boldsymbol{\theta}$ viene dado por

$$\mathbf{U}_{\mathrm{p}}(\boldsymbol{\theta}) = \frac{\partial L_{\mathrm{p}}(\boldsymbol{\theta}, \lambda)}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \mathbf{U}_{\mathrm{p}}^{\alpha}(\boldsymbol{\theta}) \\ \mathbf{U}_{\mathrm{p}}^{\beta_1}(\boldsymbol{\theta}) \\ \vdots \\ \mathbf{U}_{\mathrm{p}}^{\beta_s}(\boldsymbol{\theta}) \\ \mathbf{U}_{\mathrm{p}}^{\delta}(\boldsymbol{\theta}) \end{pmatrix}, \tag{3.1}$$

donde cada elemento del vector score $\mathbf{U}_{\mathrm{p}}(\boldsymbol{\theta})$ puede escribirse de la forma

$$\begin{aligned} \mathbf{U}_{\mathrm{p}}^{\alpha}(\boldsymbol{\theta}) &= \mathbf{Z}^{\top}\mathbf{D}_a\mathbf{r}\,, \\ \mathbf{U}_{\mathrm{p}}^{\beta_k}(\boldsymbol{\theta}) &= \widetilde{\mathbf{N}}_k^{\top}\mathbf{D}_a\mathbf{r} - \lambda_k\mathbf{K}_k\boldsymbol{\beta}_k \quad (k = 1, \ldots, s) \qquad \text{y} \\ \mathbf{U}_{\mathrm{p}}^{\delta}(\boldsymbol{\theta}) &= \mathrm{tr}(\mathbf{D}_{\mathrm{b}})\,, \end{aligned}$$

en que $\mathbf{Z}$ es una matriz $(n \times p)$, con fila $\mathbf{z}_i^{\top}$, $\widetilde{\mathbf{N}}_k$ es una matriz $(n \times r_k)$, $\mathbf{r} = (\mathrm{r}_1, \ldots, \mathrm{r}_n)^{\top}$, $\mathbf{D}(\mathbf{a}) = \mathrm{diag}(\mathbf{a})$ y $\mathbf{D}(\mathbf{b}) = \mathrm{diag}(\mathbf{b})$ son matrices $(n \times n)$, $\mathbf{a} = (a_1, \ldots, a_n)^{\top}$ y $\mathbf{b} = (b_1, \ldots, b_n)^{\top}$, con

$$\begin{aligned} \mathrm{r}_i &= -\frac{1}{2\mu_i} + \frac{\delta}{[\delta\, y_i + y_i + \delta\mu_i]} + \frac{y_i\,[\delta+1]}{4{\mu_i}^2} - \frac{\delta^2}{4y_i\,[\delta+1]}, \\ b_i &= \left\{ \frac{1}{2} - \frac{1}{2\,[\delta+1]} + \frac{[y_i+\mu_i]}{[\delta\, y_i + y_i + \delta\mu_i]} - \frac{y_i}{4\mu_i} - \frac{\delta[\delta+2]\mu_i}{4[\delta+1]^2 y_i} \right\} \end{aligned}$$

y

$$a_i = \frac{1}{h'(\mu_i)}.$$

3.3. Matriz Hessiana

Corresponde a una matriz cuadrada de las segundas derivadas parciales del vector de parámetros $\boldsymbol{\theta}$ y tiene varias aplicaciones en estadística. Por ejemplo, es utilizada para verificar la existencia del EMVP (estudiando la concavidad de la función de log–verosimilitud penalizada) y construir un proceso iterativo basado en el algortimo de Newton–Raphson.

Sea $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_s^\top, \delta)^\top$ y $\ddot{\mathbf{L}}_{\mathrm{p}}(\boldsymbol{\theta})$ la matriz hessiana $(p^* \times p^*)$ con (j^*, ℓ^*)–elementos dados por $\partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda})/\partial\theta_{j^*}\theta_{\ell^*}$, para $j^*, \ell^* = 1, \ldots, p^*$ y $p^* = 1 + p + \sum_{k=1}^{s} r_k$. Tras algunas manipulaciones algebraicas encontramos que la matriz hessiana penalizada tiene la forma

$$\ddot{\mathbf{L}}_{\mathrm{p}}(\boldsymbol{\theta}) = \begin{pmatrix} \ddot{\mathbf{L}}_{\mathrm{p}}^{\alpha\alpha} & \ddot{\mathbf{L}}_{\mathrm{p}}^{\alpha\beta_1} & \cdots & \ddot{\mathbf{L}}_{\mathrm{p}}^{\alpha\beta_s} & \ddot{\mathbf{L}}_{\mathrm{p}}^{\alpha\delta} \\ \ddot{\mathbf{L}}_{\mathrm{p}}^{\alpha\beta_1^\top} & \ddot{\mathbf{L}}_{\mathrm{p}}^{\beta_1\beta_1} & \cdots & \ddot{\mathbf{L}}_{\mathrm{p}}^{\beta_1\beta_s} & \ddot{\mathbf{L}}_{\mathrm{p}}^{\beta_1\delta} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \ddot{\mathbf{L}}_{\mathrm{p}}^{\alpha\beta_s^\top} & \ddot{\mathbf{L}}_{\mathrm{p}}^{\beta_1\beta_s^\top} & \cdots & \ddot{\mathbf{L}}_{\mathrm{p}}^{\beta_s\beta_s} & \ddot{\mathbf{L}}_{\mathrm{p}}^{\beta_s\delta} \\ \ddot{\mathbf{L}}_{\mathrm{p}}^{\alpha\delta^\top} & \ddot{\mathbf{L}}_{\mathrm{p}}^{\beta_1\delta^\top} & \cdots & \ddot{\mathbf{L}}_{\mathrm{p}}^{\beta_s\delta^\top} & \ddot{\mathrm{L}}_{\mathrm{p}}^{\delta\delta} \end{pmatrix}, \tag{3.2}$$

donde cada elemento de la matriz $\ddot{\mathbf{L}}_{\mathrm{p}}(\boldsymbol{\theta})$ puede escribirse como

$$\begin{aligned}
\ddot{\mathbf{L}}_{\mathrm{p}}^{\alpha\alpha}(\boldsymbol{\theta}) &= \mathbf{Z}^\top \mathbf{D}_c \mathbf{Z}, \\
\ddot{\mathbf{L}}_{\mathrm{p}}^{\beta\beta}(\boldsymbol{\theta}) &= \begin{cases} -\widetilde{\mathbf{N}}_k^\top \mathbf{D}_c \widetilde{\mathbf{N}}_k - \lambda_k \mathbf{K}_k & k = k' \\ \widetilde{\mathbf{N}}_k^\top \mathbf{D}_c \widetilde{\mathbf{N}}_{k'} & k \neq k' \end{cases} \\
\ddot{\mathbf{L}}_{\mathrm{p}}^{\alpha\beta}(\boldsymbol{\theta}) &= \mathbf{Z}^\top \mathbf{D}_c \widetilde{\mathbf{N}}_k \qquad (k = 1, \ldots, s), \\
\ddot{\mathbf{L}}_{\mathrm{p}}^{\alpha\delta}(\boldsymbol{\theta}) &= \mathbf{Z}^\top \mathbf{D}_a \mathbf{m}, \\
\ddot{\mathbf{L}}_{\mathrm{p}}^{\beta\delta}(\boldsymbol{\theta}) &= \widetilde{\mathbf{N}}_k^\top \mathbf{D}_a \mathbf{m} \qquad (k = 1, \ldots, s), \\
\ddot{\mathbf{L}}_{\mathrm{p}}^{\delta\delta}(\boldsymbol{\theta}) &= \mathrm{tr}(\mathbf{D}_d),
\end{aligned}$$

donde $\mathbf{D}_c = \mathrm{diag}\{c_1, \cdots, c_n\}, \mathbf{D}_a = \mathrm{diag}\{a_1, \cdots, a_n\}$ y $\mathbf{m} = (m_1, \cdots, m_n)^\top$, con

$$\begin{aligned}
c_i &= \frac{\partial^2 \ell_i(\mu_i, \delta)}{\partial \mu_i^2} \left[\frac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i}\right]^2 + \frac{\partial \ell_i(\mu_i, \delta)}{\partial \mu_i} \left[\frac{\partial}{\partial \mu_i} \frac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i}\right] + \frac{\mathrm{d}\mu_i}{\mathrm{d}\eta_i}, \\
m_i &= \frac{y_i}{[\delta y_i + y_i + \delta\mu_i]^2} + \frac{y_i}{4\mu_i^2} - \frac{\delta[\delta + 2]}{4[\delta + 1]^2 y_i} \quad \text{y} \\
d_i &= \frac{1}{2[\delta + 1]^2} - \frac{[y_i + \mu_i]^2}{[\delta y_i + y_i + \delta\mu_i]^2} - \frac{\mu_i}{2[\delta + 1]^3 y_i}.
\end{aligned}$$

3.4. Matriz 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 esperada penalizada de dimensión $(p^* \times p^*)$ que está dada por

$$\mathfrak{I}_{\mathrm{p}}(\boldsymbol{\theta}) = \begin{pmatrix} \mathfrak{I}_{\mathrm{p}}^{\alpha\alpha}(\boldsymbol{\theta}) & \mathfrak{I}_{\mathrm{p}}^{\alpha\beta_1}(\boldsymbol{\theta}) & \cdots & \mathfrak{I}_{\mathrm{p}}^{\alpha\beta_s}(\boldsymbol{\theta}) & \mathfrak{I}_{\mathrm{p}}^{\alpha\delta}(\boldsymbol{\theta}) \\ \mathfrak{I}_{\mathrm{p}}^{\alpha\beta_1^\top}(\boldsymbol{\theta}) & \mathfrak{I}_{\mathrm{p}}^{\beta_1\beta_1}(\boldsymbol{\theta}) & \cdots & \mathfrak{I}_{\mathrm{p}}^{\beta_1\beta_s}(\boldsymbol{\theta}) & \mathfrak{I}_{\mathrm{p}}^{\beta_1\delta}(\boldsymbol{\theta}) \\ \vdots & \vdots & \ddots & \vdots & \ddots \\ \mathfrak{I}_{\mathrm{p}}^{\alpha\beta_s^\top}(\boldsymbol{\theta}) & \mathfrak{I}_{\mathrm{p}}^{\beta_1\beta_s^\top}(\boldsymbol{\theta}) & \cdots & \mathfrak{I}_{\mathrm{p}}^{\beta_s\beta_s}(\boldsymbol{\theta}) & \mathfrak{I}_{\mathrm{p}}^{\beta_s\delta}(\boldsymbol{\theta}) \\ \mathfrak{I}_{\mathrm{p}}^{\alpha\delta^\top}(\boldsymbol{\theta}) & \mathfrak{I}_{\mathrm{p}}^{\beta_1\delta^\top}(\boldsymbol{\theta}) & \cdots & \mathfrak{I}_{\mathrm{p}}^{\beta_s\delta^\top}(\boldsymbol{\theta}) & \mathfrak{I}_{\mathrm{p}}^{\delta\delta}(\boldsymbol{\theta}) \end{pmatrix}, \tag{3.3}$$

donde cada elemento de la matriz puede escribirse como

$$\begin{aligned} \mathfrak{I}_{\mathrm{p}}^{\alpha\alpha} &= \mathbf{Z}^\top \mathbf{D}_v \mathbf{Z}, \\ \mathfrak{I}_{\mathrm{p}}^{\alpha\beta_k} &= \mathbf{Z}^\top \mathbf{D}_v \widetilde{\mathbf{N}}_k\,, \qquad (k = 1, \ldots, s) \\ \mathfrak{I}_{\mathrm{p}}^{\alpha\delta} &= \mathbf{Z}^\top \mathbf{D}_a\, \mathbf{s} \\ \mathfrak{I}_{\mathrm{p}}^{\beta_k\beta_k} &= \begin{cases} \widetilde{\mathbf{N}}_k^\top \mathbf{D}_v \widetilde{\mathbf{N}}_k + \lambda_k \mathbf{K}_k & k = k' \\ \widetilde{\mathbf{N}}_k^\top \mathbf{D}_v \widetilde{\mathbf{N}}_{k'} & k \neq k' \end{cases} \\ \mathfrak{I}_{\mathrm{p}}^{\beta_k\delta} &= \widetilde{\mathbf{N}}_k^\top \mathbf{D}_a\, \mathbf{s}, \qquad (k = 1, \ldots, s) \\ \mathfrak{I}_{\mathrm{p}}^{\delta\delta} &= \mathrm{tr}(\mathbf{D}_u)\,, \end{aligned}$$

donde $\mathbf{D}_v = \mathrm{diag}\{v_1, \cdots, v_n\}, \mathbf{D}_u = \mathrm{diag}\{u_1, \cdots, u_n\}$ y $\mathbf{s} = (s_1, \cdots, s_n)^\top$, son matrices diagonales $(n \times n)$ con

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

con

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

Es importante señalar que en esta clase de modelos, la propiedad de ortogonalidad entre los vectores de parámetros $(\boldsymbol{\alpha}, \boldsymbol{\beta}_k)$ y δ no se verifica, a diferencia de lo que se observa en otras clases de modelos tales como los MLG. En general, la propiedad de ortogonalidad es deseable puesto que simplifica el proceso de estimación, en el sentido de que permite estimar los parámetros por separado. Más detalles en el contexto semiparamétrico pueden encontrarse en el trabajo de Ibacache–Pulgar *et al.* (2013).

3.5. Encontrando la solución en la práctica: proceso iterativo

El procedimiento natural para determinar el estimador de $\boldsymbol{\theta}$ basado en la maximización de la función de verosimilitud penalizada equivale a resolver la ecuación $\mathbf{U}_{\mathrm{p}}(\boldsymbol{\theta}) = \mathbf{0}$. Sin embargo, las ecuaciones de estimación son no lineales y requieren un método iterativo de optimización, es decir, no se dispone de expresiones de forma cerrada para el estimador de máxima verosimilitud penalizada (EMVP) de $\boldsymbol{\theta}$. No obstante, considerando el hecho de que la matriz $-\ddot{\mathbf{L}}_{\mathrm{p}}(\boldsymbol{\theta})$ puede ser definida no positiva, se sugiere sustituirla por la matriz $-\boldsymbol{\mathcal{I}}_{\mathrm{p}}(\boldsymbol{\theta})$ y utilizar por ejemplo, el algoritmo Scoring de Fisher, Newton o cuasi–Newton (BFGS). Entonces, el algoritmo iterativo de estimación viene dado por:

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + (\boldsymbol{\mathcal{I}}_{\mathrm{p}}(\boldsymbol{\theta})^{-1})^{(m)}\mathbf{U}_{\mathrm{p}}(\boldsymbol{\theta})^{(m)}, \quad m = 0, 1, 2, \cdots,$$

siendo m el orden de la iteración. Esto equivale a resolver el siguiente sistema de ecuaciones matriciales

$$\begin{pmatrix} \mathbf{Z}^\top\mathbf{D}_v\mathbf{Z} & \mathbf{Z}^\top\mathbf{D}_v\widetilde{\mathbf{N}}_1 & \cdots & \mathbf{Z}^\top\mathbf{D}_v\widetilde{\mathbf{N}}_s & \mathbf{Z}^\top\mathbf{D}_a\,\mathbf{s} \\ \widetilde{\mathbf{N}}_1^\top\mathbf{D}_v\mathbf{Z} & \widetilde{\mathbf{N}}_1^\top\mathbf{D}_v\widetilde{\mathbf{N}}_1 + \lambda_1\mathbf{K}_1 & \cdots & \widetilde{\mathbf{N}}_1^\top\mathbf{D}_v\mathbf{N}_s & \widetilde{\mathbf{N}}_1^\top\mathbf{D}_a\,\mathbf{s} \\ \vdots & \vdots & \ddots & \vdots & \ddots \\ \widetilde{\mathbf{N}}_s^\top\mathbf{D}_v\mathbf{Z} & \widetilde{\mathbf{N}}_s^\top\mathbf{D}_v\widetilde{\mathbf{N}}_1 & \cdots & \widetilde{\mathbf{N}}_s^\top\mathbf{D}_v\widetilde{\mathbf{N}}_s + \lambda_s\mathbf{K}_s & \widetilde{\mathbf{N}}_s^\top\mathbf{D}_a\,\mathbf{s} \\ \mathbf{s}^\top\mathbf{D}_a\mathbf{Z} & \mathbf{s}^\top\mathbf{D}_a\widetilde{\mathbf{N}}_1 & \cdots & \mathbf{s}^\top\mathbf{D}_a\widetilde{\mathbf{N}}_s & \mathrm{tr}(\mathbf{D}_u) \end{pmatrix}^{(m)} \begin{pmatrix} \boldsymbol{\Delta}_\alpha^{(m+1,m)} \\ \boldsymbol{\Delta}_{\beta_1}^{(m+1,m)} \\ \vdots \\ \boldsymbol{\Delta}_{\beta_s}^{(m+1,m)} \\ \boldsymbol{\Delta}_\delta^{(m+1,m)} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{Z}^\top\mathbf{D}_a\mathbf{r} \\ \widetilde{\mathbf{N}}_1^\top\mathbf{D}_a\mathbf{r} - \lambda_1\mathbf{K}_1\boldsymbol{\beta}_1 \\ \vdots \\ \widetilde{\mathbf{N}}_s^\top\mathbf{D}_a\mathbf{r} - \lambda_s\mathbf{K}_s\boldsymbol{\beta}_s \\ \mathrm{tr}(\mathbf{D}_b) \end{pmatrix}^{(m)} \tag{3.4}$$

donde $\boldsymbol{\Delta}_\alpha^{(m+1,m)} = \boldsymbol{\alpha}^{(m+1)} - \boldsymbol{\alpha}^{(m)}$, $\boldsymbol{\Delta}_{\beta_k}^{(m+1,m)} = \boldsymbol{\beta}_k^{(m+1)} - \boldsymbol{\beta}_k^{(m)}$ y $\boldsymbol{\Delta}_\delta^{(m+1,m)} = \delta^{(m+1)} - \delta^{(m)}$.

Entonces, tras algunas manipulaciones algebraicas, se obtienen las siguientes expresiones para las soluciones iterativas:

$$\begin{aligned}
\boldsymbol{\alpha}^{(m+1)} &= (\mathbf{Z}^\top\mathbf{D}_v^{(m)}\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{D}_v^{(m)}\Big[\boldsymbol{\psi}_\alpha^{(m)} - \mathbf{D}_{v,a}^{(m)}\mathbf{s}\boldsymbol{\Delta}_\delta^{(m+1,m)} - \sum_{k=1}^{s}\widetilde{\mathbf{N}}_k\boldsymbol{\Delta}_{\beta_k}^{(m+1,m)}\Big] \\
\boldsymbol{\beta}_\ell^{(m+1)} &= (\widetilde{\mathbf{N}}^\top\mathbf{D}_v^{(m)}\widetilde{\mathbf{N}} + \lambda\mathbf{K})^{-1}\widetilde{\mathbf{N}}^\top\mathbf{D}_v^{(m)}\Big[\boldsymbol{\psi}_{\beta_\ell}^{(m)} - \mathbf{D}_{v,a}^{(m)}\mathbf{s}\boldsymbol{\Delta}_\delta^{(m+1,m)} - \mathbf{Z}\boldsymbol{\Delta}_\alpha^{(m+1,m)} \\
&\qquad - \sum_{k=1,k\neq\ell}^{s}\widetilde{\mathbf{N}}_k\boldsymbol{\Delta}_{\beta_k}^{(m+1,m)}\Big] \qquad (\ell = 1, \ldots, s) \qquad \text{y} \\
\delta^{(m+1)} &= \mathrm{tr}^{-1}(\mathbf{D}_u^{(m)})\Big[\mathrm{tr}(\mathbf{D}_b^{(m)}) + \mathrm{tr}(\mathbf{D}_u^{(m)})\delta^{(m)} - \mathbf{s}^\top\mathbf{D}_a^{(m)}\mathbf{Z}\boldsymbol{\Delta}_\alpha^{(m+1,m)} - \\
&\qquad \mathbf{s}^\top\mathbf{D}_a^{(m)}\sum_{k=1}^{s}\widetilde{\mathbf{N}}_k\boldsymbol{\Delta}_{\beta_k}^{(m+1,m)}\Big],
\end{aligned}$$

donde $\boldsymbol{\psi}_\alpha^{(m)} = \mathbf{D}_{v,a}^{(m)}\mathbf{r}^m + \mathbf{Z}\boldsymbol{\alpha}^{(m)}$ y $\boldsymbol{\psi}_{\beta_\ell} = \mathbf{D}_{v,a}^{(m)}\mathbf{r}^m\widetilde{\mathbf{N}}_\ell\boldsymbol{\beta}_\ell^m$, con $\mathbf{D}_{v,a}^{(m)} = {\mathbf{D}_v^{(m)}}^{-1}\mathbf{D}_a^{(m)}$.

Cuando δ es conocido, es posible obtener expresiones simplificadas para las soluciones iterativas de $\boldsymbol{\alpha}^{(m+1)}$ y $\boldsymbol{\beta}_\ell^{(m+1)}$. En efecto, luego de realizar ciertos cálculos, se obtiene que

$$\begin{aligned}
\boldsymbol{\alpha}^{(m+1)} &= (\mathbf{Z}^\top \mathbf{D}_v^{(m)} \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{D}_v^{(m)} \Big[\boldsymbol{\psi}_\alpha^{(m)} - \sum_{k=1}^{s} \widetilde{\mathbf{N}}_k \boldsymbol{\Delta}_{\beta_k}^{(m+1,m)} \Big] \qquad \text{y} \\
\boldsymbol{\beta}_\ell^{(m+1)} &= (\widetilde{\mathbf{N}}^\top \mathbf{D}_v^{(m)} \widetilde{\mathbf{N}} + \lambda \mathbf{K})^{-1} \widetilde{\mathbf{N}}^\top \mathbf{D}_v^{(m)} \Big[\boldsymbol{\psi}_{\beta_\ell}^{(m)} - \mathbf{Z} \boldsymbol{\Delta}_\alpha^{(m+1,m)} - \\
&\qquad \sum_{k=1, k\neq\ell}^{s} \widetilde{\mathbf{N}}_k \boldsymbol{\Delta}_{\beta_k}^{(m+1,m)} \Big] \qquad (\ell = 1, \ldots, s),
\end{aligned}$$

o, equivalentemente,

$$\begin{aligned}
\boldsymbol{\alpha}^{(m+1)} &= (\mathbf{Z}^\top \mathbf{D}_v^{(m)} \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{D}_v^{(m)} \Big[\mathbf{r}_{v,a}^{(m)} - \sum_{k=1}^{s} \widetilde{\mathbf{N}}_k \boldsymbol{\beta}_k^{(m+1)} \Big] \qquad \text{y} \\
\boldsymbol{\beta}_\ell^{(m+1)} &= (\widetilde{\mathbf{N}}^\top \mathbf{D}_v^{(m)} \widetilde{\mathbf{N}} + \lambda \mathbf{K})^{-1} \widetilde{\mathbf{N}}^\top \mathbf{D}_v^{(m)} \Big[\mathbf{r}_{v,a}^{(m)} - \mathbf{Z} \boldsymbol{\alpha}^{(m+1)} - \\
&\qquad \sum_{k=1, k\neq\ell}^{s} \widetilde{\mathbf{N}}_k \boldsymbol{\beta}_k^{(m+1)} \Big] \qquad (\ell = 1, \ldots, s),
\end{aligned}$$

donde $\mathbf{r}_{v,a} = \mathbf{D}_{v,a}^{(m)} \mathbf{r}^{(m)} + \boldsymbol{\eta}^{(m)}$ con $= \boldsymbol{\eta}^{(m)} \mathbf{Z} \boldsymbol{\alpha}^{(m)} \sum_{k=1, k\neq\ell}^{s} \widetilde{\mathbf{N}}_k \boldsymbol{\beta_k}^{(m+1)}$.

Es posible demostrar que estas expresiones corresponden a las iteraciones de Back–fitting ponderado (Gauss–Seidel) considerando $\mathbf{r}_{v,a}$ como variable dependiente modificada y $\mathbf{D}_v$ como una matriz de pesos que cambia con cada iteración del proceso. Una expresión general para estas iteraciones es la siguiente:

$$\boldsymbol{\beta}_\ell^{(m+1)} = \mathbf{S}_\ell^{(m)} \Big[\mathbf{r}_{v,a}^{(m)} - \sum_{k=0, k\neq\ell}^{s} \widetilde{\mathbf{N}}_k \boldsymbol{\beta}_k^{(m+1)} \Big] \qquad (\ell = 0, 1, \ldots, s), \tag{3.5}$$

donde $\mathbf{r}_{a,b}^{(m)} = \mathbf{D}_{v,n}^{(m)} \mathbf{r}^{(m)} + \boldsymbol{\eta}^{(m)}$, con $\boldsymbol{\eta}^{(m)} = \sum_{k=0}^{s} \widetilde{\mathbf{N}}_k \boldsymbol{\beta}_k^{(m)}$, $\widetilde{\mathbf{N}}_0 = \mathbf{Z}$, $\boldsymbol{\beta}_0 = \boldsymbol{\alpha}$ y

$$\begin{aligned}
\mathbf{S}_0^{(m)} &= (\widetilde{\mathbf{N}}_0^\top \mathbf{D}_v^{(m)} \widetilde{\mathbf{N}}_0)^{-1} \widetilde{\mathbf{N}}_0^\top \mathbf{D}_v^{(m)} \qquad \text{y} \\
\mathbf{S}_k^{(m)} &= (\widetilde{\mathbf{N}}_k^\top \mathbf{D}_v^{(m)} \widetilde{\mathbf{N}}_k + \lambda_k \mathbf{K}_k)^{-1} \widetilde{\mathbf{N}}_k^\top \mathbf{D}_v^{(m)} \qquad (k = 1, \ldots, s).
\end{aligned}$$

El procedimiento de estimación para obtener el estimador de máxima verosimilitud penalizada de $\boldsymbol{\theta}$ itera entre un algoritmo de ajuste posterior ponderado con matriz de pesos $\mathbf{D}_v$ y una estimación de máxima verosimilitud del parámetro de escala.

En general, el sistema de ecuaciones (3.4) es consistente y el algoritmo de ajuste a posteriori (3.5) converge a una solución para cualquier valor inicial si la matriz de pesos $\mathbf{D}_v$ es simétrica y definida positivamente. Además, esta solución es única bajo *"no concurvity"* en los datos (equivalente al concepto de multicolinealidad en regresión múltiple paramétrica), para más información véase Berhane y Tibshirani (1998) e Ibacache–Pulgar *et al.* (2012).

Con relación a los valores iniciales del algoritmo, se consideraron en el caso de $\boldsymbol{\alpha}$ y $\boldsymbol{\beta}_k$ los estimadores de mínimos cuadrados ordinarios, con la diferencia que para $\boldsymbol{\beta}_k$ fueron bajo el modelo de regresión no paramétrico. Por otra parte, para el parámetro de precisión δ se usó el estimador de momentos presentado en el trabajo de Santos–Neto *et al.* (2014). En resumen, la soluciones iterativas aquí presentadas serán encontradas bajo cierto criterio de convergencia, el cual es dado por las correspondientes iteraciones, en particular, cuando la diferencia entre la estimación del paso anterior y el actual sea menor a un valor ϵ pre–establecido (Hastie y Tibshirani, 1990).

Capítulo 4

Algunos aspectos inferenciales

En este capítulo se considerarán algunos aspectos inferenciales asociados al MCVP–BSR. En la Sección 4.1 se realizará una revisión de algunos trabajos que abordan el problema de la inferencia en el modelo de regresión BSR y el contexto semiparamétrico. En la Sección 4.2, se deriva la matriz de varianza–covarianza de los EMVPs a partir de la inversa de la matriz de información de Fisher. En la Sección 4.3, se presenta una breve discusión sobre cómo seleccionar el grado de libertad efectivo (edf en inglés) asociado al componente no paramétrico del modelo. Por último, en la Sección 4.4, se presentan los principales métodos para seleccionar los parámetros de suavizado.

4.1. Introducción

El MCVP–BSR surge como una alternativa al modelo de regresión BSR ya que permite modelar tanto tendencias lineales como no lineales en su componente sistemática, siendo éstas producto de una interacción entre las covariables. Aunque los MCVP se han utilizado ampliamente en la modelización estadística, el estudio de su inferencia bajo distribuciones no normales ha sido bastante limitado. Sin embargo, algunos autores han abordado la estimación e inferencia para algunos casos particulares. Por ejemplo, Wahba (1983) estudió la construcción de intervalos de confidencia para la función suave del modelo de regresión no paramétrico basado en la función de covarianza del estimador de Bayes. Segal *et al.* (1994) derivaron la varianza del EMVP utilizando el algoritmo de maximización de esperanza (EM en inglés); véase también Green (1990). Durban *et al.* (1999) propusieron una aproximación para el error estándar del estimador del coeficiente de regresión en un modelo aditivo semiparamétrico utilizando smoothing spline y loess. Berhane y Tibshirani (1998) propusieron bandas de error estándar puntuales (SEB en inglés) y algunas pruebas aproximadas, como la prueba de razón de verosimilitud y la prueba de puntuación para modelos aditivos generalizados en un conjunto de datos longitudinales. Fan y Jiang (2005) extendieron la prueba de razón de verosimilitud generalizada para modelos aditivos, con el fin de verificar si admite una forma paramétrica. Esta extensión consideró los estimadores de back–fitting para las funciones suaves. Lombardía y Sperlich (2008) desarrollaron una prueba para la hipótesis de un modelo de efectos mixtos paramétrico frente a un modelo de efectos mixtos semiparamétrico. Ibacache–Pulgar *et al.* (2012) aproximaron la matriz de varianza-covarianza del EMVP en el modelo lineal parcial de Student–t. Ibacache–Pulgar y Reyes (2018) propusieron aproximar la matriz de varianza–covarianza del EMVP en un modelo aditivo semiparamétrico bajo distribuciones simétricas, utilizando la matriz de información

de Fisher obtenida a partir de la función de verosimilitud penalizada. En cuanto a la selección de los parámetros de suavización, cuando se utiliza un spline es habitual considerar el método de VC o el método de VCG, estudiada por Craven y Wahba (1979). Como alternativa, estos parámetros pueden seleccionarse aplicando el criterio de información de AIC (Akaike, 1973) o el criterio de información bayesiano (BIC en inglés) (Schwarz, 1978). En cuanto a la selección de los parámetros de suavizado, Hurvich *et al.* (1998) propusieron una versión corregida del AIC, el AICm, para evitar la sobrecarga en modelos no paramétricos; véase también Hurvich *et al.* (1998). De acuerdo a Wahba y Wold (1975), cuando se intenta elegir el modelo óptimo para el conjunto de datos, se pueden utilizar algunos criterios de rendimiento como herramienta de comprobación, como por ejemplo seleccionar aquel ajuste e implícitamente, aquel parámetro λ_k que minimice el error cuadrático promedio de predicción. Además, Simonoff y Tsai (1999) derivaron el AICm para la selección de parámetros y variables de suavización en el contexto de modelos aditivos y semiparamétricos.

4.2. Errores estándar aproximados

En esta sección se considerará el problema de estimar la matriz de varianza–covarianza del EMVP de $\boldsymbol{\theta}$. Teniendo en cuenta que hemos obtenido el EMVP de $\boldsymbol{\theta}$ mediante el algoritmo Fisher's scoring, es razonable derivar la matriz de varianza–covarianza utilizando la inversa de la matriz de información de Fisher penalizada; que a su vez se calcula considerando la función de verosimilitud penalizada dada en (2.8) como una función de verosimilitud habitual; véase, por ejemplo, Segal *et al.* (1994), Wahba (1983) e Ibacache–Pulgar *et al.* (2013). Para calcular la matriz inversa de $\mathfrak{I}_{\mathrm{p}}(\boldsymbol{\theta})$ dada en (3.1), se debe considerar

$$\mathfrak{I}_{\mathrm{p}}^{11} = \begin{pmatrix} \mathfrak{I}_{\mathrm{p}}^{\alpha\alpha} & \mathfrak{I}_{\mathrm{p}}^{\alpha\beta_1} & \cdots & \mathfrak{I}_{\mathrm{p}}^{\alpha\beta_s} \\ \mathfrak{I}_{\mathrm{p}}^{\alpha\beta_1^\top} & \mathfrak{I}_{\mathrm{p}}^{\beta_1\beta_1} & \cdots & \mathfrak{I}_{\mathrm{p}}^{\beta_1\beta_s} \\ \vdots & \vdots & \ddots & \vdots \\ \mathfrak{I}_{\mathrm{p}}^{\alpha\beta_s^\top} & \mathfrak{I}_{\mathrm{p}}^{\beta_1\beta_s^\top} & \cdots & \mathfrak{I}_{\mathrm{p}}^{\beta_s\beta_s} \end{pmatrix}, \qquad \mathfrak{I}_{\mathrm{p}}^{12} = \begin{pmatrix} \mathfrak{I}_{\mathrm{p}}^{\alpha\delta} \\ \mathfrak{I}_{\mathrm{p}}^{\beta_1\delta} \\ \vdots \\ \mathfrak{I}_{\mathrm{p}}^{\beta_s\delta} \end{pmatrix}, \qquad \mathfrak{I}_{\mathrm{p}}^{22} = \mathfrak{I}_{\mathrm{p}}^{\delta\delta}.$$

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

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

Suponiendo que existen todas las inversas necesarias, tras 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:

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

donde $\mathfrak{I}_{\mathrm{p}}^{11,1} = (-\mathfrak{I}_{\mathrm{p}}^{11} - \mathfrak{I}_{\mathrm{p}}^{12}\mathfrak{I}_{\mathrm{p}}^{22^{-1}}\mathfrak{I}_{\mathrm{p}}^{12\top})^{-1}$ y $\mathfrak{I}_{\mathrm{p}}^{22,1} = \mathfrak{I}_{\mathrm{p}}^{22^{-1}} + \mathfrak{I}_{\mathrm{p}}^{22^{-1}}\mathfrak{I}_{\mathrm{p}}^{12\top}\mathfrak{I}_{\mathrm{p}}^{11,1}\mathfrak{I}_{\mathrm{p}}^{12}\mathfrak{I}_{\mathrm{p}}^{22^{-1}}$.

Por lo tanto, la matriz de varianza–covarianza asintótica de $\widehat{\boldsymbol{\theta}}$ viene dada por

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

En particular, se tiene

$$\widehat{\mathrm{Cov}}_{\mathrm{aprox}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}_1, \cdots, \widehat{\boldsymbol{\beta}}_s) = \mathfrak{I}_{\mathrm{p}}^{22,1}\big|_{\widehat{\boldsymbol{\theta}}}.$$

Los autores Hastie y Tibshirani (1993) propusieron un SEB puntual aproximado para funciones no paramétricas β_k's, con el fin de evaluar la precisión de los estimadores $\widehat{\beta_k}$'s para diferentes localizaciones dentro del rango de interés. Para este caso, las bandas se construyen utilizando los elementos diagonales correspondientes de la matriz $\widehat{\mathrm{Cov}}_{\mathrm{aprox}}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}_1, \cdots, \widehat{\boldsymbol{\beta}}_s)$ como estimadores de los errores estándar (SE) de β_k's. En efecto, podemos considerar el siguiente SEB puntual aproximado:

$$\mathrm{SEB}_{\mathrm{approx}}(\beta_k(\mathrm{t}_l^0)) \;=\; \widehat{\beta}_k(\mathrm{t}_l^0) \pm 2\sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_k(\mathrm{t}_l^0))},$$

donde $\mathrm{Var}(\widehat{\beta}_k(\mathrm{t}_l))$ es el l–ésimo elemento de la diagonal principal de la matriz dada en (4.1), para $l = 1, \cdots, r$. Note que los t_l^0 corresponden a los nodos asociados a cada variable cuya contribución al modelo es no paramétrica.

4.3. Sobre los grados de libertad

A continuación, se presenta una definición de los grados de libertad asociada a los componentes paramétricos y no paramétricos del modelo, el cual está basado en la convergencia del proceso iterativo dado en la ecuación (3.5) para las estimaciones $\widehat{\boldsymbol{\beta}}_\jmath (\jmath = 0, 1, \cdots, s)$. Ciertamente, considerando δ y λ_k's fijos, se obtiene

$$\widehat{\boldsymbol{\beta}}_\ell \;=\; \widehat{\mathbf{S}}_\ell\, \widehat{\mathbf{r}}_{v,a}^{*} \quad (\ell = 0, \cdots, s)$$

donde $\widehat{\mathbf{r}}_{v,a}^{*} = \widehat{\mathbf{r}}_{v,a} - \sum_{k=0, k\neq\ell}^{s} \widetilde{\mathbf{N}}_k \widehat{\boldsymbol{\beta}}_k$, con $\widehat{\mathbf{r}}_{v,a} = \widehat{\mathbf{D}}_{v,n}\widehat{\mathbf{z}} + \widehat{\boldsymbol{\eta}}$, $\widehat{\boldsymbol{\eta}} = \sum_{k=0}^{s} \widetilde{\mathbf{N}}_k \widehat{\boldsymbol{\beta}}_k$ y

$$\begin{aligned} \widehat{\mathbf{S}}_0 &= (\widetilde{\mathbf{N}}_0^\top \widehat{\mathbf{D}}_v \widetilde{\mathbf{N}}_0)^{-1} \widetilde{\mathbf{N}}_0^\top \widehat{\mathbf{D}}_v \qquad \text{y} \\ \widehat{\mathbf{S}}_k &= (\widetilde{\mathbf{N}}_k^\top \widehat{\mathbf{D}}_v \widetilde{\mathbf{N}}_k + \lambda_k \mathbf{K}_k)^{-1} \widetilde{\mathbf{N}}_k^\top \widehat{\mathbf{D}}_v \quad (k = 0, \cdots, s). \end{aligned}$$

En la literatura relativa a los modelos de regresión existen diferentes denominaciones para los grados de libertad (g.l), dependiendo del contexto en el que se utilicen; véase, por ejemplo, Buja *et al.* (1989). Aquí, los g.l asociados al componente paramétrico, $\mathbf{Z}\widehat{\boldsymbol{\alpha}}$, se denotan como

$$\mathrm{df}_z = \mathrm{tr}\{\mathbf{X}\big(\mathbf{Z}^\top \widehat{\mathbf{D}}_v \mathbf{Z}\big)^{-1} \mathbf{Z}^\top \widehat{\mathbf{D}}_v\} = p,$$

donde p es el rango de $\mathbf{Z}$. Por otro lado, los g.l para el componente no paramétrico, $\widetilde{\mathbf{N}}_k \widehat{\boldsymbol{\beta}}_k$, se definen como

$$\mathrm{df}(\lambda_k) = \mathrm{tr}\{\widetilde{\mathbf{N}}_k \big(\widetilde{\mathbf{N}}_k^\top \widehat{\mathbf{D}}_v \widetilde{\mathbf{N}}_k\big)^{-1} \widetilde{\mathbf{N}}_k^\top \widehat{\mathbf{D}}_v\}, \tag{4.3}$$

que miden la contribución del efecto individual del k–ésimo componente no paramétrico.

4.4. Sobre los parámetros de suavizado

En las secciones anteriores, los parámetros de suavizado λ_k se han supuesto fijos. Sin embargo, en situaciones prácticas, dichos parámetros deben seleccionarse a partir de los datos. A continuación, se describe un criterio para seleccionar los parámetros de suavizado basado en el AIC.

4.4.1. Criterio AIC

Antes de proponer un criterio para seleccionar los parámetros de suavizado, hay que recordar que bajo el MCVP–BSR se tiene un total de $1+p+\mathrm{df}(\boldsymbol{\lambda})$ parámetros a estimar, con $\mathrm{df}(\boldsymbol{\lambda}) = \sum_{k=1}^{s} \mathrm{df}(\lambda_k)$ denotando aproximadamente el número de parámetros efectivos involucrados en el modelado de las funciones de suavizado. En este caso, se puede utilizar tanto el AIC como el BIC para seleccionar los parámetros de suavizado λ_k's. La idea es minimizar la función objetivo, con respecto a $\boldsymbol{\lambda}$

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

donde $L_p(\widehat{\boldsymbol{\theta}}, \boldsymbol{\lambda})$ denota la función de log–verosimilitud penalizada evaluada en $\widehat{\boldsymbol{\theta}}$ para un valor fijo de $\boldsymbol{\lambda}$. Una cuadrícula (superficie) para diferentes valores de $\boldsymbol{\lambda}$ y su correspondiente $\mathrm{AIC}(\boldsymbol{\lambda})$ es útil para elegir los parámetros óptimos de suavizado.

4.4.2. Fijación de los grados de libertad

Otra forma de seleccionar los parámetros de suavizado es cuando los g.l dados en la ecuación (4.3) dependen sólo de λ_k y, por tanto, se puede especificar el parámetro de suavizado correspondiente. En otras palabras, se especifican previamente $\mathrm{df}(\lambda_k)$ como una función de λ_k. Este enfoque fue utilizado para el modelo aditivo generalizado y el MCV por Hastie y Tibshirani (1990, 1993), respectivamente. Para finalizar, se pueden encontrar más detalles sobre estos métodos en el trabajo de varios autores, tales como, Buja *et al.* (1989), Rigby y Stasinopoulos (2005) e Ibacache–Pulgar y Reyes (2018).

Capítulo 5

Análisis de diagnóstico

En este capítulo se considerará la extensión y aplicación de la técnica de influencia local al MCVP–BSR. Inicialmente, en la Sección 5.1, se hace una breve revisión bibliográfica de los principales trabajos relacionados con el análisis de diagnóstico. En la Sección 5.2, se proponen algunos tipos de residuos basados en el trabajo desarrollado por Leiva *et al.* (2014). En la Sección 5.3, se realizará una descripción general del método de influencia local. Por último, en la Sección 5.4 se derivará la curvatura normal para tres esquemas de perturbación, concretamente, la perturbación de la ponderación de casos, la variable de respuesta y la perturbación del parámetro de precisión.

5.1. Introducción

Es sabido que el análisis de diagnóstico es un proceso fundamental en la modelización estadística de cualquier conjunto de datos. Entre las técnicas más utilizadas en regresión paramétrica se encuentran la influencia global (o eliminación de casos) y la influencia local, esta última introducida por Cook (1986), para evaluar la sensibilidad de los estimadores de los parámetros cuando se introducen pequeñas perturbaciones en las hipótesis del modelo (en este caso los supuestos del MCVP–BSR) o en el conjunto de datos, para ello, se miden los cambios en la función de log–verosimilitud asignando diferentes pesos a las unidades y es muy útil para indagar en las fuentes de aquellas desviaciones. En este trabajo se considerarán la extensión y aplicación de este último método en el modelo propuesto, cuya implementación es nueva. Algunos de los principales trabajos sobre esta metodología son los siguientes: Thomas (1991) construyó diagnósticos de influencia local para evaluar la sensibilidad de la estimación del parámetro de suavizado obtenido por el criterio de validación cruzada, Zhu *et al.* (2003) e Ibacache–Pulgar *et al.* (2012) proporcionaron medidas de influencia local que permitieron evaluar la sensibilidad del EMVP en modelos normales y Student–t parcialmente lineales, respectivamente. Chen *et al.* (2010) propusieron un procedimiento para seleccionar el esquema de perturbación apropiado cuando el método de influencia local se aplica a los modelos lineales mixtos generalizados. Ibacache–Pulgar *et al.* (2013) derivaron la curvatura de influencia local para modelos aditivos semiparamétricos simétricos, y obtuvieron evidencia empírica de la robustez (en el sentido de la distancia de Mahalanobis) del EMVP para distribuciones con colas pesadas. De Bastiani *et al.* (2014) extendieron el método de influencia local a modelos lineales espaciales elípticos y reportaron que la presencia de datos atípicos en la muestra tiene una influencia significativa cuando se modifica la estructura de dependencia espacial. Ferreira y Paula (2016) extendieron la técnica

de influencia local para diferentes esquemas de perturbación considerando un modelo parcialmente lineal sesgado–normal. Respecto al MCVP, Zhang *et al.* (2015) desarrollaron medidas de influencia local para este modelo bajo normalidad, mientras que Ibacache–Pulgar y Reyes (2018) extendieron dichos resultados para el caso elíptico. Por otra parte, en el modelo de regresión BSR, Leiva *et al.* (2014) desarrollaron métodos de influencia local calculando las curvaturas normales bajo diferentes esquemas de perturbación para evaluar la influencia potencial de algunas observaciones en dicho modelo. Estos esquemas de perturbación son la ponderación de casos, la respuesta, el regresor y una perturbación en el parámetro de precisión. Además, se propusieron cuatro tipos de residuos, a saber: residuo estandarizado, residuo de Jørgensen, residuo estandarizado basado en la solución de una regresión lineal ponderada mediante el cálculo de las estimaciones de mínimos cuadrados ordinarios y el residuo del componente de desvío (CD). En los últimos años, 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 MAS–BSR.

5.2. Análisis de los residuos

Es un método estadístico para comprobar si un modelo de regresión presenta un buen ajuste, estudiando la componente de los datos que no es explicada por dicho modelo, el cual es resultado de un procesamiento de múltiples pasos.

Con el fin de detectar las especificaciones erróneas en la distribución, así como la presencia de observaciones periféricas, se proponen dos tipos de residuos basados en los resultados presentados en Leiva *et al.* (2014). Estos residuos son descritos a continuación.

Residuo normalizado
Estos residuos estandarizan los datos en el análisis de regresión y en las pruebas de hipótesis de chi cuadrado.
Además, este tipo de residuo se basa en $y_i - \mu_i$ y puede definirse de la siguiente manera

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

donde $\phi = [\delta + 1]^2/[2\delta + 5]$ y $\mu_i = h^{-1}(\mathbf{z}_i^\top \boldsymbol{\alpha} + \mathrm{x}_{1_i}\, \beta_1(\mathrm{t}_{1_i}) + \ldots + \mathrm{x}_{s_i}\, \beta_s(\mathrm{t}_{s_i}))$.

El residuo 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 MCVP–BSR, se propone

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

donde

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

y

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

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

5.3. Método de influencia local

Sea $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_n)^\top$ un vector n–dimensional de perturbaciones restringido a algún subconjunto abierto $\Omega \in \mathbb{R}^n$ y la función de log–verosimilitud penalizada perturbada denotada por $L_{\mathrm{p}}(\boldsymbol{\theta},\boldsymbol{\lambda}|\boldsymbol{\omega})$. Suponiendo que existe $\boldsymbol{\omega}_0 \in \Omega$, un vector de no perturbación, tal que $L_{\mathrm{p}}(\boldsymbol{\theta},\boldsymbol{\lambda}|\boldsymbol{\omega}_0) = L_{\mathrm{p}}(\boldsymbol{\theta},\boldsymbol{\lambda})$. Para evaluar la influencia de las perturbaciones menores en el EMVP $\widehat{\boldsymbol{\theta}}$, podemos considerar el desplazamiento de verosimilitud

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

donde $\widehat{\boldsymbol{\theta}_\omega}$ es el EMVP bajo $L_{\mathrm{p}}(\boldsymbol{\theta},\boldsymbol{\lambda}|\boldsymbol{\omega})$. La medida $LD(\boldsymbol{\omega})$ es útil para evaluar la distancia entre $\widehat{\boldsymbol{\theta}}$ y $\widehat{\boldsymbol{\theta}_\omega}$. Cook (1986), propone estudiar el comportamiento local de $LD(\boldsymbol{\omega})$ alrededor de $\boldsymbol{\omega}_0$. El procedimiento consiste en seleccionar una dirección de la unidad $\boldsymbol{\ell} \in \Omega$ ($\|\boldsymbol{\ell}\| = 1$) y luego considerar la gráfica de $LD = (\boldsymbol{\omega}_0 + a\boldsymbol{\ell})$ versus a, siendo $a \in \mathbb{R}$. Esta gráfica se denomina 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 $LD(\boldsymbol{\omega})$. Según Cook (1986), la curvatura normal en la dirección unitaria viene dada por

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

donde

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

Nótese que $-\ddot{\mathbf{L}}_{\mathrm{p}}$ es la matriz de información observada penalizada evaluada en $\widehat{\boldsymbol{\theta}}$ (véase la sección 2.4) y $\boldsymbol{\Delta}_{\mathrm{p}}$ 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 y 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 $\mathbf{C} = \boldsymbol{\Delta}_{\mathrm{p}}^\top\ddot{\mathbf{L}}_{\mathrm{p}}^{-1}\boldsymbol{\Delta}_{\mathrm{p}}$. Para tener una curvatura invariante bajo un cambio de escala uniforme, Poon y Poon (1999) propusieron la curvatura normal conforme denotada como

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

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 $\boldsymbol{\Delta}_{\mathrm{p}}$ para los esquemas de perturbación de la ponderación de casos, la variable de respuesta y el parámetro de precisión.

5.4.1. Perturbación de ponderación de casos

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

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

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 de no perturbación. Diferenciando $L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega})$ con respecto a los elementos de $\boldsymbol{\theta}$ y $\boldsymbol{\omega}$ se obtiene

$$\begin{aligned}
\frac{\partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\omega}}\Big|_{\theta=\hat{\theta},\ \omega=\omega_0} &= \mathbf{Z}^{\top} \widehat{\mathbf{D}}_a \widehat{\mathbf{D}}_{\mathbf{Z}}, \\
\frac{\partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial \boldsymbol{\beta}_k \partial \boldsymbol{\omega}}\Big|_{\theta=\hat{\theta},\ \omega=\omega_0} &= \widetilde{\mathbf{N}}_k^{\top} \widehat{\mathbf{D}}_a \widehat{\mathbf{D}}_{\mathbf{Z}} \qquad (k = 1, \ldots, s) \qquad \text{y} \\
\frac{\partial^2 L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial \delta \partial \boldsymbol{\omega}}\Big|_{\theta=\hat{\theta},\ \omega=\omega_0} &= \widehat{\mathbf{b}},
\end{aligned}$$

para $i = 1, \cdots, n$, con $\mathbf{D}_a$ y $\mathbf{b}$ denotados 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 (2.8) con y_i sustituida por $\mathrm{y}_{i\omega}$, es decir

$$L_{\mathrm{p}}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega}) = L(\boldsymbol{\theta}|\boldsymbol{\omega}) - \sum_{k=1}^{s} \frac{\lambda_k}{2}\, \boldsymbol{\beta}_k^{\top} \mathbf{K}_k \boldsymbol{\beta}_k \,, \tag{5.2}$$

donde $L(\cdot)$ se muestra en la expresión (2.4), pero esta vez con $\mathrm{y}_{i\omega_i}$ en lugar de y_i. Aquí, el vector de no perturbación viene dado por $\boldsymbol{\omega}_0 = (0, \cdots, 0)^{\top}$.

Diferenciando $L_{\rm p}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega})$ con respecto a los elementos de $\boldsymbol{\theta}$ y $\boldsymbol{\omega}$, tras realizar ciertos cálculos, se tiene que

$$\begin{aligned}
\frac{\partial^2 L_{\rm p}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\omega}}\Big|_{\theta=\hat{\theta},\, \omega=\omega_0} &= \mathbf{Z}^\top \widehat{\mathbf{D}}_a \widehat{\mathbf{D}}_\psi \widehat{\mathbf{D}}_\vartheta, \\
\frac{\partial^2 L_{\rm p}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial \boldsymbol{\beta}_k \partial \boldsymbol{\omega}}\Big|_{\theta=\hat{\theta},\, \omega=\omega_0} &= \widetilde{\mathbf{N}}_k^\top \widehat{\mathbf{D}}_a \widehat{\mathbf{D}}_\psi \widehat{\mathbf{D}}_\vartheta \qquad (k = 1, \ldots, s) \qquad \text{y} \\
\frac{\partial^2 L_{\rm p}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial \delta \partial \boldsymbol{\omega}}\Big|_{\theta=\hat{\theta},\, \omega=\omega_0} &= \widehat{\boldsymbol{\tau}}^\top \widehat{\mathbf{D}}_\vartheta,
\end{aligned}$$

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

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

y

$$\widehat{\tau}_i = -\frac{\widehat{\mu}_i}{[\widehat{\delta} y_i + y_i + \widehat{\delta}\widehat{\mu}_i]^2} - \frac{1}{4\widehat{\mu}_i} + \frac{\widehat{\delta}[\widehat{\delta}+2]\widehat{\mu}_i}{4[\widehat{\delta}+1] y_i^2} \quad (i = 1, \ldots, n).$$

5.4.3. Perturbación en el parámetro de precisión

La perturbación del parámetro de precisión se utiliza para evaluar la sensibilidad de los EMVPs a pequeñas modificaciones de δ. Inicialmente, el MCVP–BSR asume que el parámetro de precisión es constante en todas las observaciones. Sin embargo, en el modelo perturbado, este parámetro no es constante entre las observaciones, es decir,

$$\text{y}_i \sim \text{BSR}(\mu_i, \delta_i),$$

donde $\delta_i = \delta/\omega_i$, con $\omega_i > 0$, para $i = 1, \cdots, n$. Bajo este esquema de perturbación, el vector de no perturbación viene dado por $\boldsymbol{\omega}_0 = (1, \cdots, 1)^\top$. A continuación, la función de log–verosimilitud penalizada perturbada se construye a partir de la ecuación (2.8) con δ sustituido por δ_i. Tomando diferenciales de $L_{\rm p}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega})$ con respecto a los elementos de $\boldsymbol{\theta}$ y $\boldsymbol{\omega}$, se obtiene después de algunas manipulaciones algebraicas lo siguiente

$$\begin{aligned}
\frac{\partial^2 L_{\rm p}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\omega}}\Big|_{\theta=\hat{\theta},\, \omega=\omega_0} &= \mathbf{Z}^\top \widehat{\mathbf{D}}_a \widehat{\mathbf{D}}_\varpi, \\
\frac{\partial^2 L_{\rm p}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial \boldsymbol{\beta}_k \partial \boldsymbol{\omega}}\Big|_{\theta=\hat{\theta},\, \omega=\omega_0} &= \widetilde{\mathbf{N}}_k^\top \widehat{\mathbf{D}}_a \widehat{\mathbf{D}}_\varpi \qquad (k = 1, \ldots, s) \qquad \text{y} \\
\frac{\partial^2 L_{\rm p}(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\omega})}{\partial \delta \partial \boldsymbol{\omega}}\Big|_{\theta=\hat{\theta},\, \omega=\omega_0} &= \widehat{\boldsymbol{\varphi}}^\top,
\end{aligned}$$

para $i = 1, \cdots, n$, donde $\widehat{\mathbf{D}}_\varpi = \text{diag}\{\widehat{\varpi}_1, \cdots, \widehat{\varpi}_n\}$ y $\widehat{\boldsymbol{\varphi}} = \text{diag}(\widehat{\varphi}_1, \cdots, \widehat{\varphi}_n)^\top$, con

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

y

$$\widehat{\varphi}_i = -\frac{1}{2} + \frac{1}{2[\widehat{\delta}+1]^2} - \frac{y_i[y_i+\widehat{\mu}_i]}{[\widehat{\delta} y_i + y_i + \widehat{\delta}\widehat{\mu}_i]^2} - \frac{1}{4\widehat{\mu}_i} + \frac{y_i}{4\widehat{\mu}_i} + \frac{\widehat{\delta}^2\widehat{\mu}_i[\widehat{\delta}+3]}{4y_i[\widehat{\delta}+1]^3} + \frac{\widehat{\delta}\widehat{\mu}_i}{y_i[\widehat{\delta}+1]^3} \quad (i = 1, \ldots, n).$$

Capítulo 6

Aplicación a datos de contaminación atmosférica

La contaminación atmosférica consiste en la presencia de materia o partículas flotantes en el aire, lo cual se relaciona directamente con el desarrollo de ciertas enfermedades en los seres humanos y daño en la biodiversidad. Aunque la fuente más importante en nuestro país se ha identificado como antropogénica, es considerado como un fenómeno multifactorial, ya que por ejemplo, en su capital Santiago, es debido a sus rasgos meteorólogicos y geomorfológicos, así como su latitud y la frecuente ubicación del anticiclón del Pacífico, que hacen que en su cuenca atmosférica se den circunstancias desfavorables para los procesos de dispersión de contaminantes, los cuales junto al aumento de la población, la constante expansión urbana, la alta concentración de vehículos y las crecientes actividades industriales que existe una acumulación de MP y gases durante el invierno, además de un aumento en la radiación solar durante el verano que favorece las reacciones fotoquímicas; véase a Marchant *et al.* (2013), Cavieres *et al.* (2020). Sumado a esto, los altos niveles de ciertos contaminantes varían según los cambios climáticos y topográficos que dependen de las modificaciones en la fuente y el tipo de emisión. Motivo por el cual, las condiciones meteorológicas son un factor clave e incontrolable en la determinación de la variabilidad de la contaminación atmosférica. En algunos casos, se puede superar la influencia de algunos efectos antropogénicos, como los que son originados por el rastreo de vehículos; Yáñez *et al.* (2017). Además, la relación entre variables meteorológicas y el MP se han analizado en todo el mundo Clements *et al.* (2016), considerándose en este estudio algunas de estas variables. El efecto de los parámetros meteorológicos sobre el MP se ha estudiado utilizando diferentes técnicas estadísticas, incluyendo la regresión lineal múltiple, los modelos aditivos generalizados, splines de regresión adaptativa multivariante y redes neuronales; Puentes *et al.* (2021). Para este trabajo se considerará el conjunto de datos medioambientales relacionados con la contaminación atmosférica. En particular, los datos proporcionados por el SINCA (`www://sinca.mma.gob.cl`) correspondientes a la contaminación del aire en la comuna de Pudahuel en la Región Metropolitana, durante el período GEC (1 de abril al 31 de agosto) del año 2019. El objetivo de este estudio es evaluar la asociación de las concentraciones contaminantes con variables meteorológicas mediante el uso del MCVP–BSR. Para ello, con el propósito de motivar los modelos semiparamétricos, se consideró el MP2.5 como variable de respuesta y como covariables, las concentraciones de MP10 en μg/Nm^3, velocidad del viento en m/s y temperatura en °C. En el periodo GEC se trabaja con un total de 153 observaciones, valor que surge del cálculo de los promedios diarios.

Nivel MP2.5	Nivel MP10	Indicación
[0, 50)	[0, 150)	Bueno
[50, 80)	[150, 195)	Regular
[80, 110)	[195, 240)	Alerta
[110, 170)	[240, 330)	Pre-Emergencia
≥ 170	≥ 330	Emergencia

Tabla 6.1: Norma nacional índice de calidad del aire para los niveles de MP2.5 y MP10.

6.1. Análisis exploratorio

Es importante aclarar que durante el periodo GEC (correspondiente a los meses de otoño e invierno) hay un aumento en los niveles de MP2.5 (Figura 6.1), alcanzando concentraciones que son consideradas peligrosas para la salud humana.

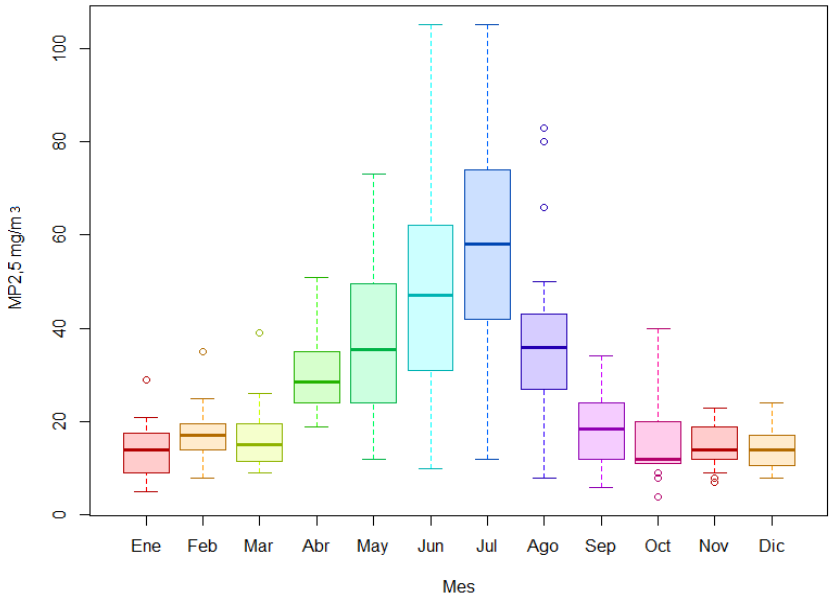

Figura 6.1: Boxplot ajustado para MP2.5 por mes registrado por la estación de monitoreo Pudahuel. Región Metropolitana, Chile 2019.

Además, se analizaron los niveles de MP2.5 para los meses del periodo GEC durante un lapso de 24 hrs. (Figura 6.2). Inicialmente, se puede notar que el mes de Abril es el que presenta los niveles más bajos de material particulado. Asimismo, se observaron dos *peaks* significativos; el primero correspondiente a un alza en las concentraciones del compuesto alrededor de las 7:00 hrs., y el segundo, de mayor magnitud, registrado aproximadamente entre las 23:00 y 01:00 hrs.

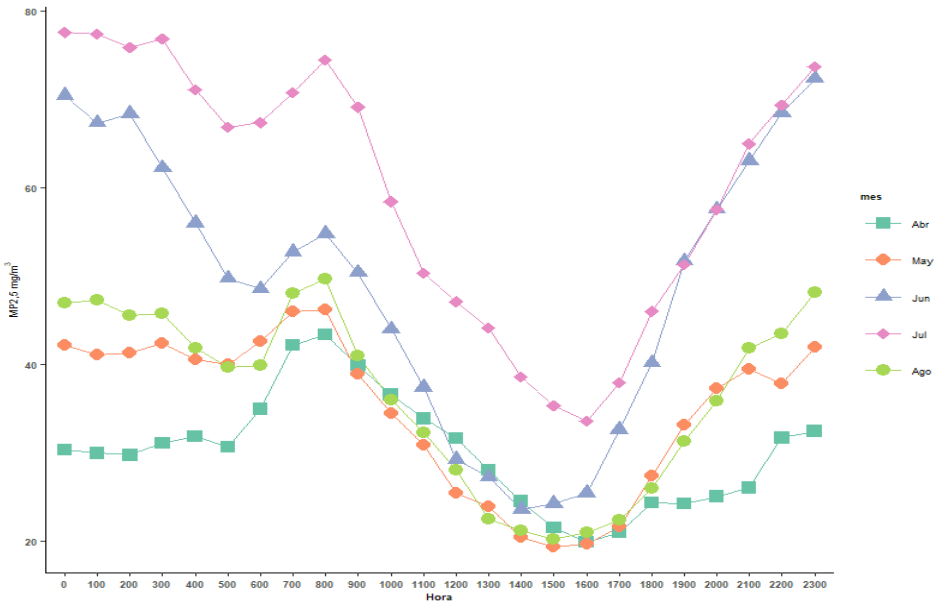

Figura 6.2: Concentraciones promedio de MP2.5 por mes y hora durante el periodo GEC, registrado por la estación de monitoreo Pudahuel. Región Metropolitana, Chile 2019.

Material particulado		
Estadístico	MP2.5	MP10
n	153	153
$\bar{y}$	42.27	100.28
DE	21.52	45.9
MD	37	99
$y_{(1)}$	8	16
$y_{(n)}$	105	239
Rango	97	223
CS	0.88	0.5
CK	3.22	2.84

Variables meteorológicas		
Estadístico	Vel. viento	Temp.
n	153	153
$\bar{y}$	0.74	10.83
DE	0.30	3.49
MD	0.71	10.19
$y_{(1)}$	0.24	3.57
$y_{(n)}$	1.69	20.56
Rango	1.45	16.99
CS	0.72	0.53
CK	3.34	2.62

Tabla 6.2: Estadística descriptiva para variables de contaminación atmosférica, registradas diariamente por la estación de monitoreo de Pudahuel durante el periodo GEC. Región Metropolitana, Chile 2019.

La Tabla 6.2 proporciona un resumen descriptivo de las variables consideradas en este estudio, el cual incluye la media ($\bar{y}$), la desviación estándar (DE), la mediana (MD), el mínimo ($y_{(1)}$), el máximo ($y_{(n)}$), el rango, el coeficiente de sesgo (CS), el coeficiente de curtosis (CK) y el total de observaciones (n). La Figura 6.3 contiene el histograma y el boxplot de las concentraciones (en micrómetros) de la variable MP2.5.

La normativa principal de calidad del aire para material particulado fino MP2.5 es de 20 μg/m^3 como concentración anual y 50μg/m^3 como nivel de 24 horas. En cambio, para el contaminante material particulado respirable MP10, es de 50 μg/m^{3}N por año y 150 μg/m^{3}N por 24 horas (Tabla 6.1). No obstante, según la Tabla 6.2, esta normativa es superada por ambas concentraciones de material particulado.

Con relación a la variable respuesta, se observa que presenta un CS = 0.88, lo que indicaría una ligera asimetría en los datos, además posee un CK = 3.26, valor que señala una distribución de colas más pesadas con respecto a la Normal. Asimismo, a partir del histograma mostrado en la Figura 6.3 (a), se puede constatar que los valores de MP2.5 tienen una distribución empírica positivamente sesgada y en la Figura 6.3 (b), se detecta un dato atípico en el boxplot, correspondiente a la observación 63 (2 de junio del 2019) y una concentración de los datos entre los 30 y 55 μg/m^3. Por consiguiente, tras los resultados presentados en la Tabla 6.2 y las tendencias mostradas en las Figuras 6.3 y 6.5, se sugiere que el MCVP–BSR dado en la subsección 2.3 puede ser adecuado para describir la media, la varianza y la asimetría presentada por este conjunto de datos.

En cuanto a las variables metereológicas, la velocidad del viento presentó valores clasificados como moderados y la variable temperatura una media considerada como templada. De igual forma ambas presentaron asimetría positiva, es decir, sus distribuciones se encuentran sesgadas hacia la derecha. Un dato importante a considerar es que para la comuna de Pudahuel el índice de aridez fue de 2.11, valor clasificado como desértico (GeoAdaptative, 2020).

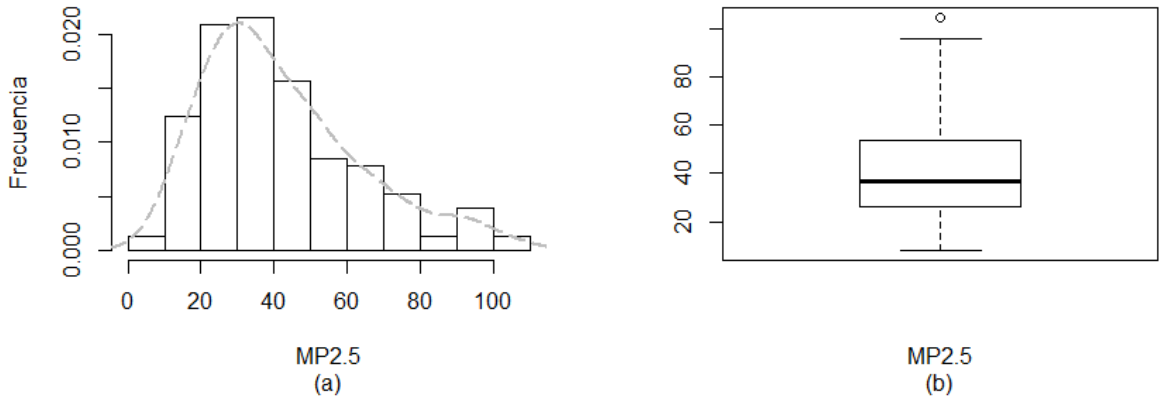

Figura 6.3: Histograma (a) y Boxplot (b) de la variable respuesta MP2.5 en el periodo GEC. Pudahuel, Región Metropolitana, Chile 2019.

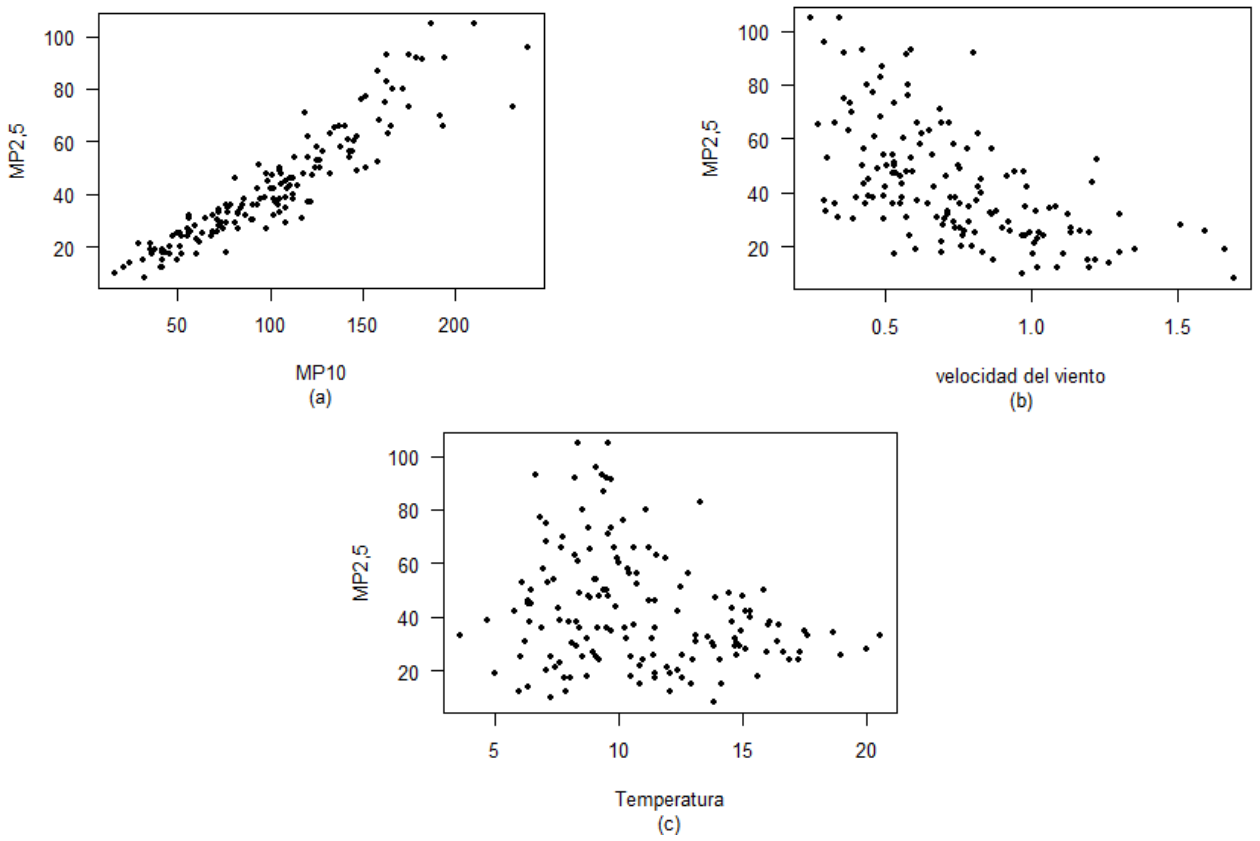

Figura 6.4: Gráfico de dispersión entre variable respuesta MP2.5 y variables MP10 (a), vel. del viento (b) y temperatura (c). Estación de monitoreo Pudahuel, Chile 2019.

La Figura 6.4 contiene los gráficos de dispersión entre la variable de respuesta y cada una de las covariables. En la Figura 6.4 (a) se puede apreciar que la relación entre el MP2.5 y la covariable MP10 parece ser positivamente lineal, lo cual es apoyado por el coeficiente de correlación entre ambas variables que es igual a 0.93, además, se observa que la variabilidad de la variable respuesta tiende a aumentar a medida que los valores de MP10 aumentan.

La observación anterior también podría ser indicio de una varianza no constante en los datos. Por su parte, la Figura 6.4 (a) muestra evidencias de que la recta no pasa por el origen, por ende, podría ser de utilidad considerar el intercepto en la componente paramétrica del modelo propuesto. Mientras que la relación entre MP2.5 y el resto de covariables parece no ser lineal como se muestra en la Figura 6.4 (b) y (c), esta tendencia se presenta más claramente en la asociación de la respuesta y la covariable viento.

Sin embargo, como la asociación no lineal entre las variables MP2.5 y temperatura no es tan evidente, se sugiere la incorporación del efecto de interacción entre las covariables viento y temperatura (Figura 6.5), aquí es posible comprobar la contribución no lineal por parte de la variable viento.

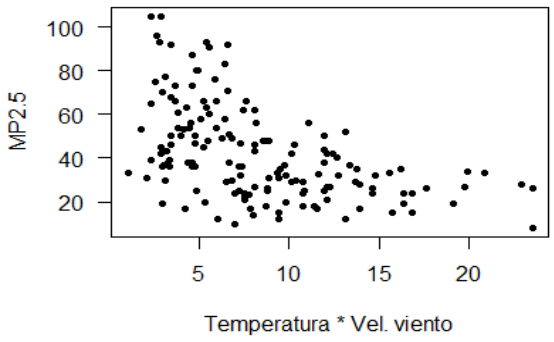

Figura 6.5: Gráfico de interacción temperatura* vel. viento. Estación de monitoreo Pudahuel, Chile 2019.

Tras realizar el análisis exploratorio al conjunto de datos, es posible observar que la variable de respuesta presenta características de la distribución BSR, tales como la asimetría positiva y el soporte no negativo, de igual forma en los gráficos de dispersión se puede constatar el aporte tanto lineal como no lineal de las covariables en la respuesta, razón por la cual se propone un modelo semiparamétrico que modele las concentraciones y tendencias del MP2.5, tanto en función del MP10 como en la interacción del viento y temperatura, las cuales, como ha quedado en evidencia, contribuyen conjuntamente de manera tanto paramétrica como no paramétrica al modelo.

6.2. Estimación y verificación de los supuestos

Las tendencias descritas anteriormente en la Sección 6.1 sugieren un modelo semiparamétrico BSR entre la MP2.5 y las covariables, asumiendo la siguiente estructura para la función de enlace:

$$h(\mu_i) = \mu_i = \alpha_0 + \alpha_1 * \mathrm{z}_{1_i} + \mathrm{x}_{1_i} * \beta_1(\mathrm{t}_{1_i}), \quad i = 1, 2, \ldots, 153.$$

donde y_i denota el i-ésimo valor de la variable MP2.5, z_i corresponde a la i-ésimo valor de la variable MP10, x_{1_i} denota el i-ésimo valor de la variable temperatura y t_{1_i} corresponde a la i-ésima unidad experimental de los valores distintos y ordenados de la variable viento. Por otra parte, $h(\cdot)$ corresponde a la función de enlace del modelo dado en la subsección 2.3, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^\top$ es un vector de parámetros desconocidos y $\beta_1(\cdot)$ es la función suave.

Se aplicó el procedimiento descrito en la subsección 4.4 para estimar el parámetro de suavizado para el componente no paramétrico del modelo, por lo tanto, el estimador para λ_1 obtenido fue $\hat{\lambda}_1 = 0.032$, con el cual se obtuvo un valor para los g.l cercano a 5.5. En cuanto a la estimación de los parámetros del componente paramétrico, se maximizó la función de log–verosimilitud penalizada como se describe en la subsección 2.4 y se obtuvieron los EMVP para $(\alpha_0, \alpha_1)^\top$ y δ son $\hat{\alpha}_0 = 3.55$, $\hat{\alpha}_1 = 0.42$ y $\hat{\delta} = 42.39$, cuyos errores estándar (SE) son 1.93, 0.021 y 4.8 respectivamente.

Para comprobar que el modelo es adecuado para describir la media de la variable de respuesta, se verifican los supuestos establecidos para el modelo. En primer lugar se observa que α_0 y α_1 son altamente significativos al 5 %, ya que ambos p-valor empíricos son cercanos a cero, lo que se esperaba tras el análisis exploratorio de datos realizado previamente. Así pues, el componente paramétrico del modelo seleccionado parece ser apropiado en cuanto a los datos en estudio.

La Figura 6.6 (a) muestra el gráfico de la función suave estimada, cuyos coeficientes se calcularon utilizando el valor del parámetro $\hat{\lambda}_1$. Sus bandas de confianza, construidas a partir del error estándar aproximado, se muestran en la Figura 6.6 (b). Obsérvese que la función estimada prácticamente no oscila, lo que es un indicio de que la estimación del parámetro de suavizado parece ser adecuada, puesto que, como se ha mencionado con anterioridad, λ_1 es el encargado de controlar simultáneamente la bondad de ajuste y la suavidad de la función. Además, se sugiere claramente que las curvas de las funciones estimadas varían conjuntamente con la variable explicativa t_1.

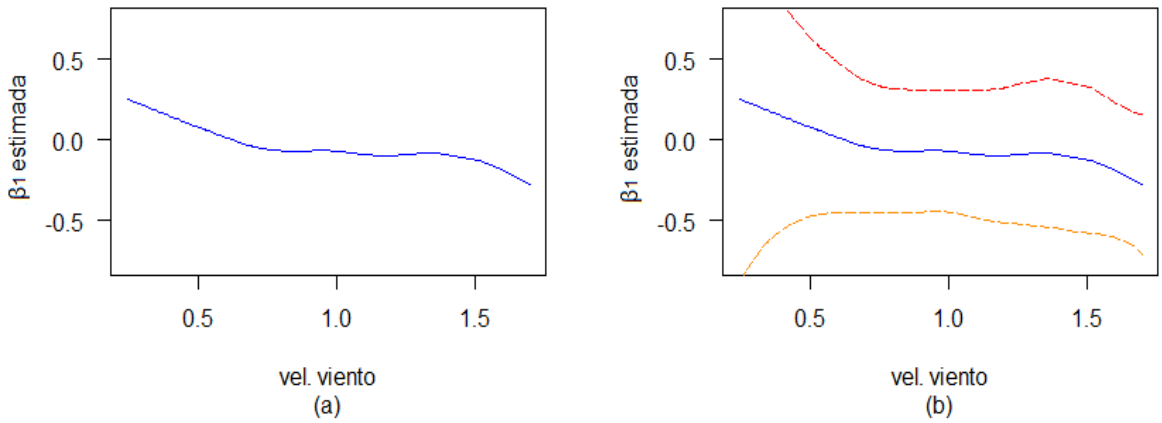

Figura 6.6: Gráfico de los valores estimados de la función (a) y de las bandas de confianza (b).

La Figura 6.7 muestra el gráfico entre los residuos parciales, $r^{(p)}$, definidos como

$$r^{(p)} = h(\hat{\mu}_i) - \hat{\alpha}_0 - \hat{\alpha}_1 \, \mathrm{z}_i$$

y la variable viento en la que se superpone la curva de la función suave estimada. Nótese que la curva parece adaptarse bien, ya que cubre correctamente aquellos valores muy distantes, además de seguir la tendencia de los puntos, esto verifica lo mencionado anteriormente. Por lo tanto, la contribución no lineal de la covariable viento sobre la variable respuesta MP2.5 estaría correctamente cuantificada a través de la función suave estimada. A continuación, se realizará un análisis residual basado en los datos de contaminación atmosférica para verificar el cumplimiento de los supuestos.

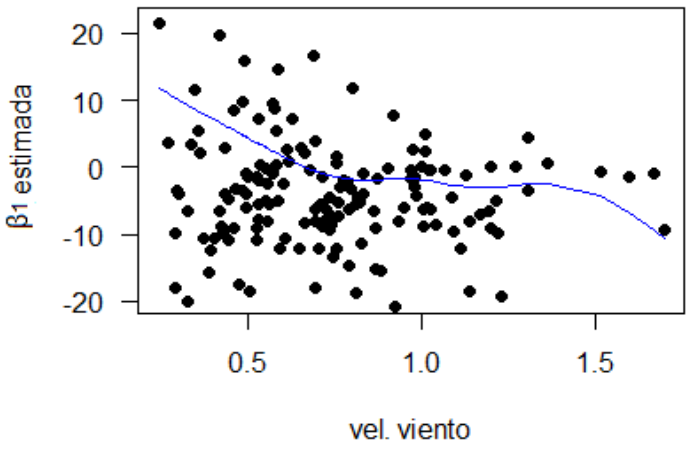

Figura 6.7: Gráfico de los residuos parciales v/s viento con la función suave estimada superpuesta.

La Figura 6.8 contiene los gráficos de valores índice frente a los dos tipos de residuos propuestos en este estudio: los residuos estandarizados, $r^{(1)}$ y los residuos de Jørgensen, $r^{(2)}$. En este caso, la mayoría de los residuos estandarizados están comprendidos en el intervalo [-2, 2], salvo ciertos valores que se ubican fuera o sobre los puntos de corte; denotadas como dos veces la desviación estándar de cada residuo.

Para el caso de los residuos estandarizados, se tienen como posibles candidatos a *outliers* las observaciones $\{30\}$, correspondiente al 30 de abril, $\{32, 33, 48\}$ con fecha 2, 3 y 18 de mayo y los casos $\{123, 151\}$ equivalentes al 1 y 29 de agosto. Mientras que en los residuos de Jørgensen, además de encontrarse los valores mencionados anteriormente, se presentan los casos $\{58, 59\}$ correspondientes a las fechas 28 y 29 de mayo, la observación $\{67\}$ con fecha 6 de junio y los valores $\{149, 150, 152\}$ correspondientes a los días 27, 28 y 30 de agosto. Dado que la distribución BSR está altamente relacionada con la distribución Normal, para las etapas posteriores se considerarán los residuos estandarizados o normalizados.

Para verificar el supuesto de distribución establecido en el modelo, se realiza un gráfico de cuantiles (QQ) para los residuos estandarizados $r^{(1)}$, que se muestran en la Figura 6.9 (a). Esta figura no muestra características inusuales, por lo que el supuesto de distribución de la variable de respuesta no parece ser inadecuado. Además, la hipótesis de normalidad residual también parece verificarse en el QQ *plot*, en el gráfico de residuos presentado en la Figura 6.8 (a) y en el histograma que se muestra en la Figura 6.9 (b), que evidencian una simetría considerable. Por lo tanto, la función de enlace seleccionada para este modelo (identidad) parece ser adecuada dado el comportamiento residual presentado.

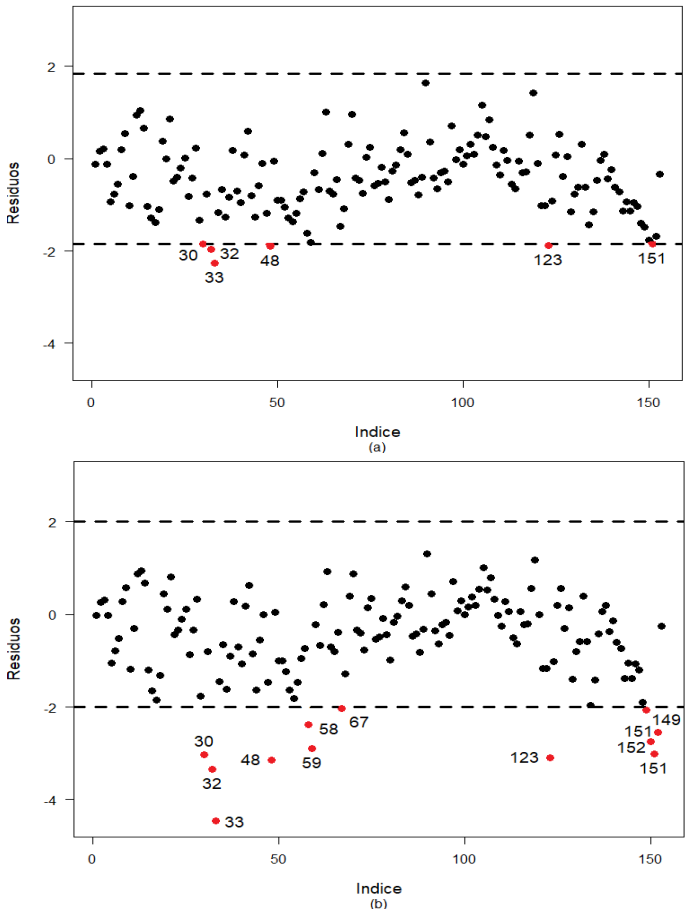

Figura 6.8: Gráfico residual: $r^{(1)}$ (a) y $r^{(2)}$ (b).

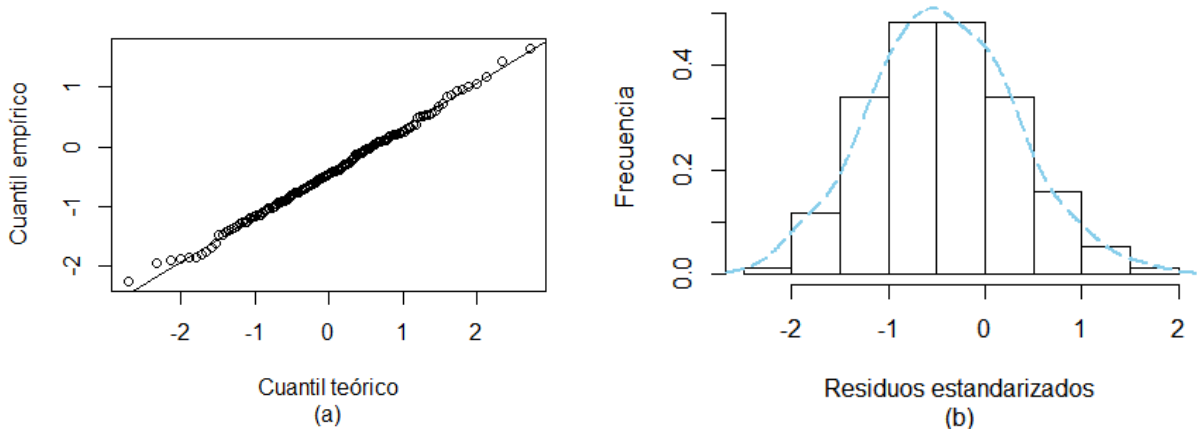

Figura 6.9: QQ *plot* (a) e histograma (b) de los residuos estandarizados.

6.3. Análisis de influencia local

La influencia local permite detectar el efecto de las perturbaciones en la estimación de los parámetros, trabajando en todo momento con el conjunto de datos sin eliminar observaciones. Para identificar posibles casos de influencia o discrepancia en el modelo ajustado, a continuación se presentan las técnicas propuestas en la subsección 2.3 para el MCVP–BSR, las cuales quedan representadas en los gráficos 6.10 al 6.13.

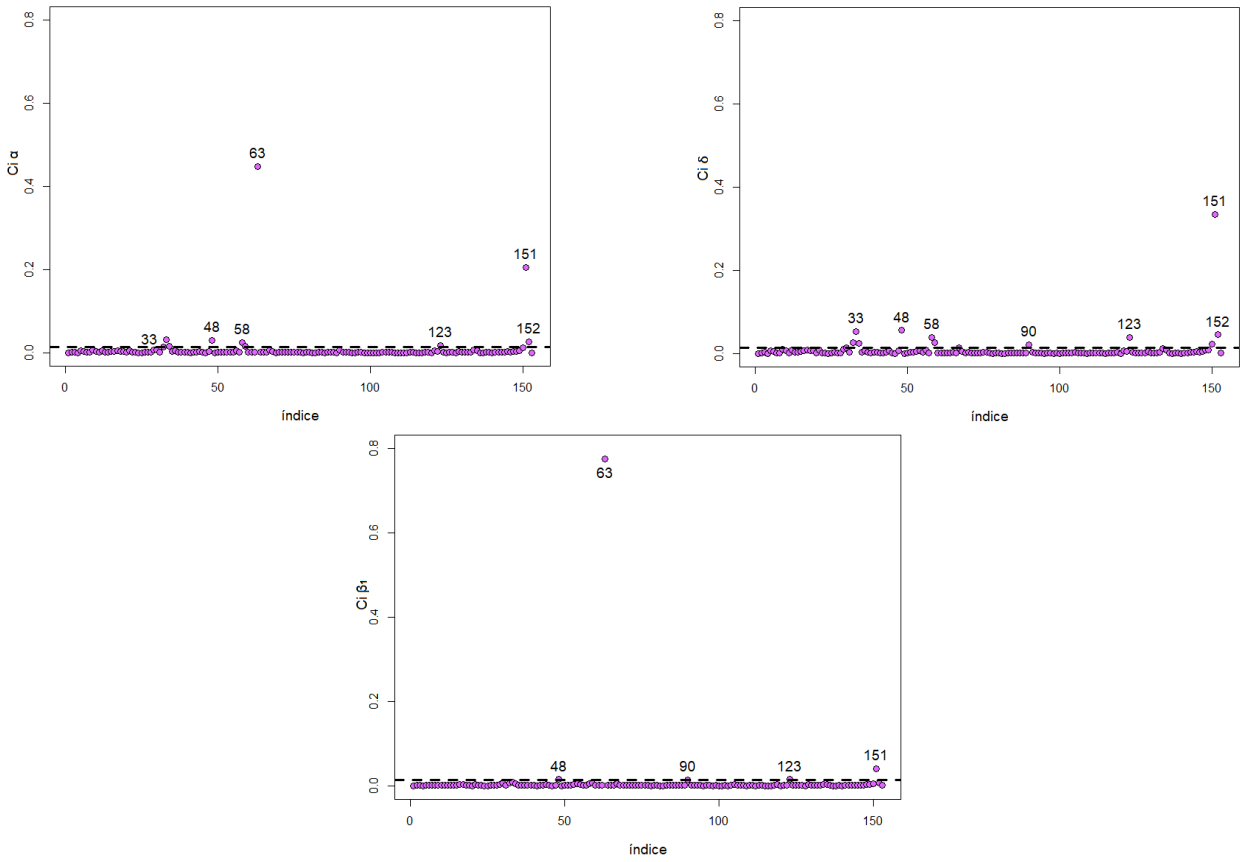

Figura 6.10: Gráficos de índices de C_i para $\boldsymbol{\alpha}$, δ y β_1 bajo la perturbación de la ponderación de casos.

En las Figuras 6.10 a la 6.12 se muestran los gráficos de índices de C_i. En la Figura 6.10 se pueden apreciar las observaciones destacadas que influyen en $\boldsymbol{\alpha}$, δ y β_1 bajo la perturbación de la ponderación (pesos) de casos, la Figura 6.11 presenta las observaciones influyentes en $\boldsymbol{\alpha}$, δ y β_1 bajo el esquema de perturbación de la respuesta y la Figura 6.12 revela las observaciones influyentes en los mismos parámetros mencionados con anterioridad, pero esta vez bajo la perturbación en la precisión.

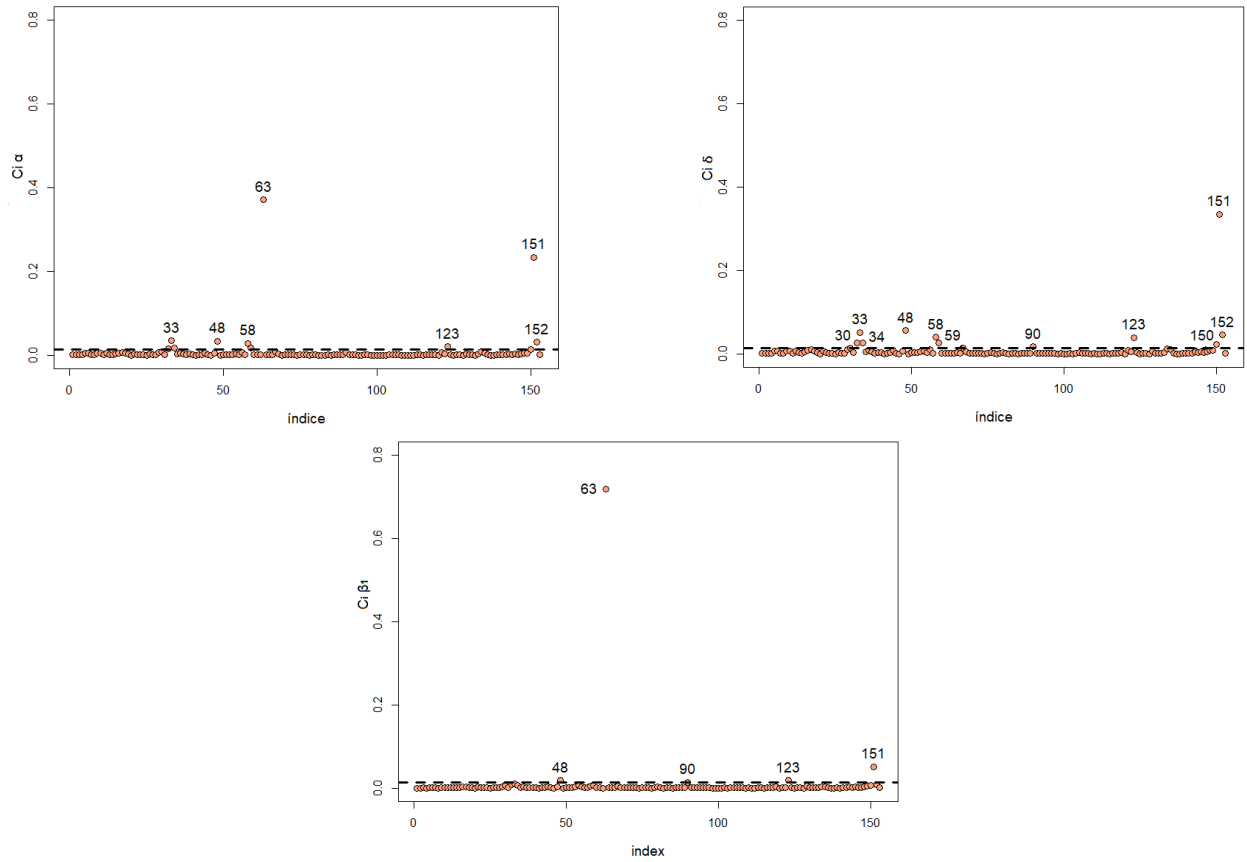

Figura 6.11: Gráficos de índices de C_i para $\boldsymbol{\alpha}$, δ y β_1 bajo la perturbación de la respuesta.

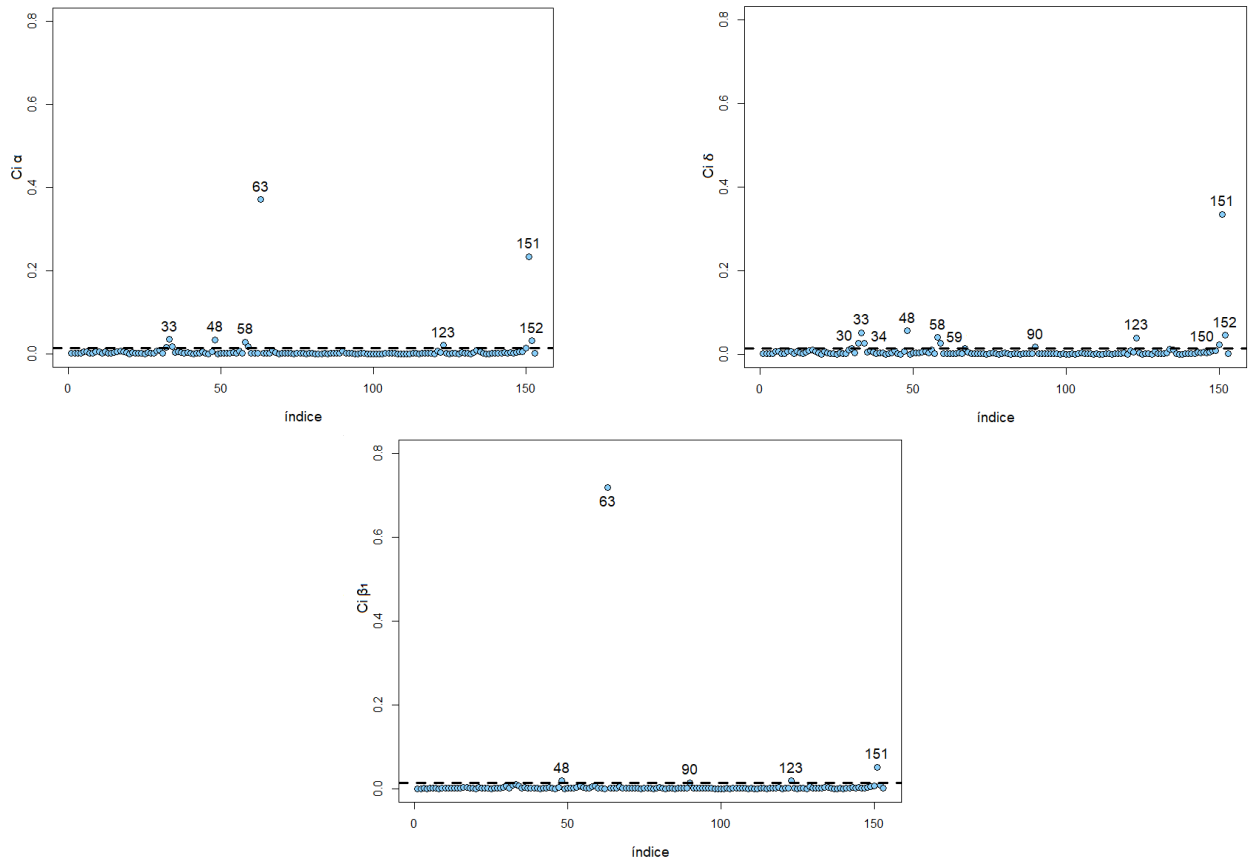

Figura 6.12: Gráficos de índices de C_i para $\boldsymbol{\alpha}$, δ y β_1 bajo la perturbación de la precisión.

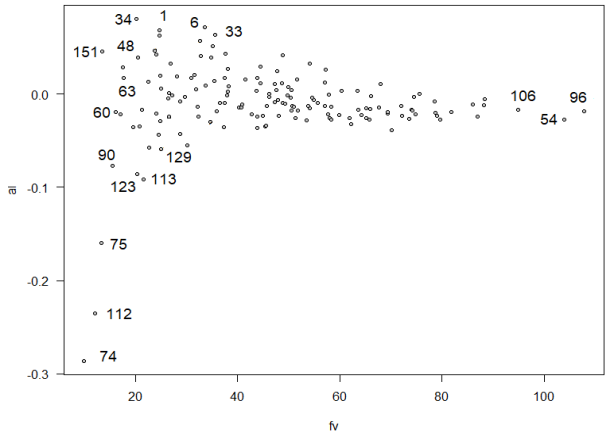

Figura 6.13: Gráfico valores de apalancamiento generalizados v/s la respuesta media estimada.

Considerando los resultados obtenidos de los gráficos de influencia local, nótese que los casos {63, 123, 151} se detectan como influyentes en los componentes paramétrico, no paramétrico y de precisión, respectivamente, debido a que son frecuentes en cada esquema de perturbación. No obstante, la observación 63 está ausente en los tres esquemas de perturbación referentes a delta, razón por la cual, no se consideraría influyente en la estimación de dicho parámetro.

La Figura 6.13 muestra un gráfico de apalancamiento o *leverage* para los valores influyentes (GL), en el que se puede notar estos mismos casos {63, 123, 151} correspondientes a las fechas 2 de junio, 1 de agosto y 29 de agosto del periodo GEC, las cuales ya podrían ser consideradas como valores potencialmente influyentes, debido a que se encuentran presentes en todos (o en la gran mayoría) de los gráficos.

En cuanto a estas observaciones influyentes, cabe señalar que el 2 de junio, se registró la concentración de MP2.5 más alta del periodo GEC 2019, correspondiente a 105 μg/m^3. Como consecuencia, se registraron alertas ambientales ese día y las fechas 4, 5, 8, 9, 10 y 11 de junio y preemergencia el día 3 del mismo mes. Por el contrario, el 29 de agosto, se informaron los niveles de MP2.5 más bajos. Un dato interesante a considerar es que la última alerta ambiental dentro del periodo GEC fue el 14 de agosto (SMA, 2020). Referente a las temperaturas máximas, para el día 2 de junio se registraron 19.3°C, para el 1 de agosto 15.9°C y para el 29 de agosto 27.7°C, temperaturas consideradas altas para invierno. En base a esta información, es posible suponer una disminución en el uso de calefactores para el hogar durante el mes de agosto, esto sumado a la fuerte campaña de fiscalización y seguimiento de los Planes de Prevención y/o Descontaminación ambiental establecidos por el Seremi del Medio Ambiente para el 2019. En cuanto a la presencia de lluvias o tormentas, en junio cayeron 32.26 mm. de precipitaciones, en agosto, por otra parte, los pluviómetros marcaron cero (GeoAdaptative, 2020).

6.4. Análisis confirmatorio

A partir de la sección anterior, se tiene conocimiento sobre qué observaciones podrían ser consideradas como potencialmente influyentes, para ello, se analiza el cambio que hay en las EMVP cuando se excluyen del conjunto de datos las observaciones señaladas en los gráficos de influencia bajo el MCVP–BSR. Para realizar este análisis, la observación o el conjunto de observaciones {63}, {123}, {151}, {63,123}, {63, 151}, {123, 151} y {63, 123, 151} se eliminarán y la estimación de los parámetros del modelo se realizará de nuevo.

La Tabla 6.3 proporciona los cambios relativos (CR) en % de las EMVP de los parámetros, la estimación de sus correspondientes SE y el nuevo p–valor asociado. Estos cambios se calculan a partir de la siguiente expresión

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

donde $\hat{\alpha}_{j(i)}$ y $\text{SE}(\hat{\alpha}_j)_{(i)}$ denotan las estimaciones de máxima verosimilitud de α_j y sus errores estándar, obtenidos después de extraer la i-ésima observación, para $j = 1, 2$ e $i = 1, 2, ..., 153$. En la Tabla 6.3 se observa que los CR más importantes se detectan para las estimaciones de α_0, donde los valores más altos se asocian con dicho parámetro, en particular, para las observaciones {123, 151}, esto sin considerar la remoción de todos los puntos catalogados como influyentes, escenario en que los cambios relativos se disparan. No obstante, analizando cada caso particular, se tiene que la observación {151} es la más influyente en las estimaciones de máxima verosimilitud. También, es posible notar un alto CR en las desviaciones estándar del parámetro de precisión delta. Obsérvese que la significación de los parámetros, al 5 %, no cambia, ya que los p–valor se mantienen por debajo de 0.01. En resumen, los resultados presentados en esta tabla muestran que las medidas de diagnóstico derivadas en este estudio identifican puntos potencialmente influyentes, siendo las observaciones {123, 151} las que afectan principalmente a la inferencia estadística del modelo, pero no de forma significativa.

Es importante recordar que estas observaciones corresponden al 1 y al 29 de agosto del periodo GEC del 2019, fechas que pesentaron concentraciones de material particulado por debajo del promedio y condiciones metereológicas irregulares. Para finalizar, estos análisis de diagnóstico basados en los enfoques de la influencia local y los residuos, confirman que el MCVP–BSR presentado en la subsección 2.3 es estable a los puntos atípicos detectados y a las observaciones potencialmente influyentes, así como también bastante adecuado para modelar datos ambientales.

Casos eliminados	**CR en la estimación**	α_0	α_1	δ
	θ	–	–	–
Ninguno	SE	–	–	–
	p-valor	<0.01	<0.01	–
	θ	0.40	0.80	2.57
{63}	SE	1.39	1.62	2.87
	p-valor	<0.01	<0.01	–
	θ	4.62	0.42	2.85
{123}	SE	0.07	1.25	3.19
	p-valor	<0.01	<0.01	–
	θ	7.91	0.82	5.08
{151}	SE	2.17	2.45	5.43
	p-valor	<0.01	<0.01	–
	θ	5.05	1.22	5.51
{63, 123}	SE	1.34	2.87	6.21
	p-valor	<0.01	<0.01	–
	θ	8.34	1.63	7.90
{63, 151}	SE	3.59	4.10	8.61
	p-valor	<0.01	<0.01	–
	θ	12.94	1.27	8.17
{123, 151}	SE	2.09	3.70	8.88
	p-valor	<0.01	<0.01	–
	θ	13.19	2.08	11.13
{63, 123, 151}	SE	3.54	5.36	12.23
	p-valor	<0.01	<0.01	–

Tabla 6.3: CRs (%) en las estimaciones de máxima verosimilitud y en los correspondientes SE para el(los) caso(s) eliminado(s) y los respectivos p–valor utilizando datos de contaminación atmosférica y el MCVP–BSR.

Capítulo 7

Conclusiones y trabajos futuros

Este trabajo es una aproximación a la teoría del modelamiento semiparamétrico, el cual contempla la relación entre variables tanto de manera lineal, como no lineal. Por una parte, se estudió el modelo con coeficientes variando parcialmente, el cual aporta flexibilidad a la hora de incorporar el efecto de interacción entre un conjunto de covariables, enfocándose principalmente en la contribución no lineal de éstas y por otra parte, se consideró la distribución Birnbaum–Saunders reparametrizada, la cual posee excelentes características al momento de modelar fenómenos que experimenten daño acumulativo o cíclico, tales como la fatiga de material o la contaminación ambiental, motivo por el cual es un muy buen candidato al momento de considerar este tipo de datos. El modelo propuesto combina ambos elementos resultando en una propuesta novedosa con múltiples contribuciones, tales como la incorporación de una componente no paramétrica, un proceso iterativo, su correspondiente análisis de diagnóstico y su aplicación en datos reales. Sumado a lo anterior, este estudio permitió describir la media de una variable aleatoria a través de una componente paramétrica y otra no paramétrica. Además, fue posible mantener la escala original de los datos, ya que al realizar transformaciones en la variable modelada se puede reducir la interpretabilidad de los resultados.

Dentro de las etapas realizadas, se analizó la estimación de parámetros a través del criterio de máxima verosimilitud penalizada bajo el modelo con coeficiente variando parcialmente. También se desarrollaron métodos para realizar un análisis de influencia local bajo diferentes esquemas, con el fin de encontrar observaciones potencialmente influyentes. Un aspecto a considerar tiene relación con la flexibilidad del modelo estudiado, ya que permitió modelar una variable aleatoria cuyo supuesto distribucional se extiendió más allá de la clásica distribución Normal y se hizo de forma similar a los modelos lineales generalizados, por medio de una función de enlace, pero sin la necesidad de pertenecer a la familia exponencial. Finalmente, se aplicó la metodología desarrollada a un conjunto de datos reales de contaminación atmosférica de Santiago, Chile, verificando de esta manera, que el modelo propuesto es adecuado para trabajar con esta clase de variables.

Como trabajo futuro, los modelos con coeficiente variando parcialmente bajo una Birnbaum–Saunders reparametrizada pueden extenderse para casos de parámetro de precisión variable o componentes georreferenciados. Además, se pueden considerar otros tipos de penalizaciones, como por ejemplo utilizar bases P–spline e incluir aplicaciones en otras áreas.

Bibliografía

Akaike, H. (1973). Information theory and the maximum likelihood principle in 2nd International Symposium on Information Theory (B.N. Petrov and F. Csaki, eds.). Akademiai Kiado, Budapest, Hungary, 267–281.

Berhane, K., & Tibshirani , J. (1998). Generalized additive models for longitudinal data. *The Canadian Journal of Statistics*, 26: 517–535.

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.

Cai, Z., Fan, J. & Li, R. (2000) Eficient estimation and inferences for varying- coeficient models. *J. Amer. Statist. Assoc*, 95:888–902.

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

Cavieres, M. F., Leiva, V, Marchant., C., & Rojas, F. (2020). A methodology for data-driven decision making in the monitoring of particulate matter environmental contamination in Santiago of Chile. *Reviews of Environmental Contamination and Toxicology*, 250:5–67.

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.

Chiang, C., Rice, J., & Wu, C. (2001). Smoothing Spline Estimation for Varying Coeficient Models With Repeatedly Measured Dependent Variables. *Journal of the Royal Statistical Society*, 605–619.

Clements, N., Hannigan, M., Miller, S., Peel, J. , & Milford, J. (2016). Comparisons of urban and rural $PM_{10-2,5}$ and $PM_{2,5}$ mass levels and semi-volatile fractions in northeastern Colorado. *Atmospheric Chemistry and Physics*, 16, 7469–7484.

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

Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerical Mathematical*, 377–403.

Dasilva, A., Dias, R., Leiva, V., Marchant, C., & Saulo, H. (2020). Birnbaum-Saunders regression models: A comparative evaluation of three approaches. *Journal of Statistical Computation and Simulation*, 90(14):2552–2570.

De Bastiani, F., Mariz de Aquino Cysneiros, A. H., Uribe–Opazo, M. A., & Galea, M. (2014). Influence diagnostics in elliptical spatial linear models. TEST, 24(2), 322–340.

Durban, M., Hackett, C., & Currie, I. (1999). Approximate standard errors in semi-parametric additive models. *Biometrics*, 55, 699–703.

Escobar, A., & Meeker Q. (1992) Assessing local in uence in regression analysis with censored data. *Biometrics*, 48:507–528.

Eubank, L. (2004). Smoothing spline estimation in varying–coeficient models. *J. R. Statist. Soc*, 653–667.

Fan, J., & Jiang, J. (2005). Nonparametric Inferences forAdditive Models. *Journal of the American Statistical Association*, 100:890–907.

Ferreira, C. S., & Paula, G. A. (2016). Estimation and diagnostic for skew-normal partially linear models. *Journal of Applied Statistics*, 44(16), 3033–3053.

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

GeoAdaptative. (2020). *Informe de riesgos climáticos para la Región Metropolitana*. Paiscircular.cl. Recuperado el 3 de noviembre del 2021, desde: `https://www.paiscircular.cl/wp-content/uploads/2020/02/Informe_Riesgos_Climaticos_RM.pdf.`

Green, P. J. (1990). On Use of the EM Algorithm for Penalized Likelihood Estimation. *Journal of the Royal Statistical Society B*, 52: 443–452.

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., & 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.

Hurvich, C. M., Simonoff, J. S., & Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society B*, 60: 271–293.

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. A., & Cysneiros, F. J. A. (2013). Semiparametric additive models under symmetric distributions. Test, 22, 103–121.

Ibacache–Pulgar, G., & Reyes, S. (2018). Local influence for elliptical partially varying coeficient model. *Statistical Modelling*, 149–174.

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

Jiang, Q., Wang, H., Xia, Y., & Jiang, G. (2013). On a principal varying coefficient model. *Journal of the American Statistical Association*, 108, 228–236.

Jørgensen, B. (1984). The delta algorithm and GLIM. *International Statistical Review*, 52, 283–300.

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., Athayde, E., Azevedo, C., & Marchant, C. (2011). Modeling wind energy flux by a Birnbaum–Saunders distribution with unknown shift parameter. *Journal of Applied Statistics*, 38:2819–2838.

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

Leiva, V., Marchant, C., Ruggeri, F., & Saulo, H. (2015). A criterion for environmental assessment using Birnbaum-Saunders attribute control. *Environmetrics*, 26:463–476.

Liu, S., & Li, G. (2015). Varying-coeficient mean-covariance regression for longitudinal. *Journal of statistical planning and inference*.

Lombardía, M. J., & Sperlich, S. (2008). Semiparametric inference in generalized mixed effects models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,70(5), 913–930.

Marchant, C., Leiva, V., Cavieres, M., & Sanhueza, A. (2013). Air contaminant statistical distributions with application to PM10 in Santiago, Chile. *Reviews of Environmental Contamination and Toxicology*, 223:1–31.

Marchant, C., Leiva, V., & Cysneiros, F. J. A. (2016). A multivariate log-linear model for Birnbaum-Saunders distributions. *IEEE Transactions on Reliability*, 65:816–827.

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.

Marchant, C., Leiva, V., Cysneiros, F.J.A., & Liu, S. (2018). Robust multivariate control charts based on Birnbaum-Saunders distributions. *Journal of Statistical Computation and Simulation*, 88(1):182–202.

MMA. (2011). Informe del estado del medio ambiente. Santiago, Chile: Ministerio del Medio Ambiente.

MMA. (2016). *Guia-para-Docentes-Sobre-Calidad-del-Aire*. Mma.gob.cl. Recuperado el 6 de Mayo del 2021, desde: `https://mma.gob.cl/wp-content/uploads/2018/08/Guia-para-Docentes-Sobre-Calidad-del-Aire-003.pdf`.

OMS. (2018). *Calidad del aire ambiente (exterior) y salud*. Who.int. Recuperado el 15 de julio del 2021, desde: `https://www.who.int/es/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health`.

OPS. (2016). *Programa de calidad del aire de la OPS/OMS. Aire limpio, futuro saludable*. Breathelife2030.org. Recuperado el 4 de Mayo del 2021, desde: `http://breathelife2030.org/wp-content/uploads/2018/11/Road-map-for-a-PAHO-air-quality-program-PWR-SPAfinal-1.pdf`.

OPS. (2017). *Calidad del aire - OPS/OMS — Organización Panamericana de la Salud*. Paho.org. Recuperado el 8 de Mayo del 2021, desde: `https://www.paho.org/es/temas/calidad-aire?page=4`.

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.

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

Puentes, R., Marchant, C., Leiva, V., Figueroa–Zúñiga, J. I., & Ruggeri, F. (2021). Predicting PM2.5 and PM10 Levels during Critical Episodes Management in Santiago, Chile, with a Bivariate Birnbaum-Saunders Log–Linear Model. *Mathematics*, 9, 645.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. (`www.R-project.org`)

Rieck, J. R., & Nedelman, J. R. (1991). A log–linear model for the Birnbaum–Saunders distribution. *Technometrics*, 33, 51–60.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3), 507–554.

Santos–Neto, M., Cysneiros, F. J. A., Leiva, V., & Ahmed, S. E. (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.

Schwarz, C. (1978). Estimating the dimension of a model. *Annals of Statistics*, 461–464.

Segal, M., Bacchetti, P., & Jewell, P. (1994). Variances for Maximum Penalized Likelihood Estimates Obtained via the EM Algorithm. *Journal of the Royal Statistical Society B*, 56:345–352.

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.

Simonoff, J. S., & Tsai, C. L. (1999). Semiparametric and additive model selection using an improved Akaike information criterion. *Journal of Computational and Graphical Statistics*, 8, 2–40

SINCA. (2015). *Sistema de Información Nacional de Calidad del Aire*. Chile. Sinca.mma.gob.cl. Recuperado desde: `www://sinca.mma.gob.cl/`

SINIA. (s. f.). *Calidad del aire*. Sinia.mma.gob.cl. Recuperado el 18 de Abril del 2021, desde: https://sinia.mma.gob.cl/temas-ambientales/calidad-del-aire/.

SMA. (2020).*Programa Integrado de Fiscalización año 2019 período Gestión de Episodios Críticos de contaminación, Plan de Prevención y Descontaminación atmosférica de la Región Metropolitana, DS N° 31/2016 del MMA* (pp. 5–21). Santiago: División de Fiscalización Superintendencia del Medio Ambiente Gobierno de Chile.

Thomas, W. (1991). Influence diagnostics for the cross-validated smoothing parameter in spline smoothing. *Journal of the American Statistical Association*, 693–698.

UNEP. (2019). *United Nations Environment Annual Report*.unep.org. Recuperado el 9 de mayo del 2021, desde: `www.unep.org/annualreport/2019/index.php.`

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.

Wahba, G. & Wold, S. (1975). A completely automatic French curve: Fitting spline functions by cross-validation.*Commun. Statist 4*, 1—17.

Wahba, G. (1983). Bayesian confidence intervals for the cross-validated smoothing spline. *Journal of the Royal Statistical Society*, 133–150.

Yáñez, M., Baettig, R., Cornejo, J., Zamudio, F., Guajardo, J., & Fica, R. (2017). Urban airborne matter in central and southern Chile: Efects of meteorological conditions on fine and coarse particulate matter. *Atmospheric Environment*, 161:221–234.

Zhang, J., Zhang, X., Ma, H., & Zhiya, C. (2015). Local influence analysis of varying coefficient linear model. *Journal of Interdisciplinary Mathematics*, 3, 293–306.

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.

Apéndice

```
#Limpieza entorno de trabajo#
rm(list=ls())
graphics.off()

#Librerias#
library(MASS)
library(sBF)
library(pracma)
library(corpcor)
library(modes)
library(moments)
library(readr)
library(VGAM)
library(gbs)
library(maxLik)
library(betareg)
library(nortest)
library(car)
library(carData)

#file.choose()
library(readxl)
datos <- read_excel("C:\\Users\\ASUS\\Downloads\\BD_final_1.xlsx")
viento <- read_excel("C:\\Users\\ASUS\\Downloads\\Viento.xlsx")
temperatura<- read_excel("C:\\Users\\ASUS\\Downloads\\Temperatura.xlsx")

#Analisis descriptivo variable respuesta#
Y <- datos$‘PM2,5‘

hist(Y,
col = "white",
main = "",sub="(a)",
xlab = "MP2.5",
ylab = "Frecuencia",
freq = FALSE)

lines(density(Y), xlim=c(-2,120), lwd=2, lty=5, col="gray")

boxplot(Y,
col="white", sub="(b)",
horizontal=F,
xlab="MP2.5")

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#
library(naniar)
n_miss(temperatura$XD)

library(imputeTS)
Temperatura<-na_ma(temperatura$XD, k = 4, weighting = "exponential")

#Verificando datos perdidos
n_miss(Temperatura)

X1 <- datos$Temperatura
temperaturaxd <- Temperatura

for (i in 1:152) {
X1[i+1] <- mean(temperaturaxd[((24*i)+1):(((24*i)+1) +23)])
}
X1[1] <- mean(temperaturaxd[1:24])

#Verificando datos perdidos
n_miss(X1)

X2 <- datos$Viento
vientoxd <- viento$XD
for (i in 1:152) {
X2[i+1] <- mean(vientoxd[((24*i)+1):(((24*i)+1) +23)])
```

```
}
X2[1] <- mean(vientoxd[1:24])

#Graficos de dispersion#

layout(matrix(c(1:4), nrow=2, byrow=FALSE))
#layout.show(4)

plot(L1, Y, type="p", col="black",bg="black",pch=20,lwd=1, cex.main=1,cex.sub=1,
main = "", sub="(a)", xlab="MP10", ylab="MP2,5", cex.lab=1, axes=TRUE,
las=1, bty="o")
plot(X2, Y, type="p", col="black",bg="black",pch=20,lwd=1, cex.main=1,cex.sub=1,
main = "", sub="(b)", xlab="velocidad del viento", ylab="MP2,5", cex.lab=1, axes=TRUE,
las=1, bty="o")
plot(X1, Y, type="p", col="black",bg="black",pch=20,lwd=1, cex.main=1,cex.sub=1,
main = "", sub="(c)", xlab="Temperatura", ylab="MP2,5", cex.lab=1, axes=TRUE,
las=1, bty="o")

#Interacciones#
layout(matrix(c(1:1), nrow=2, byrow=FALSE))

plot(X2*X1, Y,type="p",col="black",bg="black",pch=20,lwd=1,cex.main=1,cex.sub=1,
main = "",sub="",xlab="Temperatura * Vel. viento",ylab="MP2.5",cex.lab=1,
axes=TRUE,las=1,bty="o")

#Estadistica descriptiva#

library(psych)
describe(Y)
describe(L1)
describe(X1)
describe(X2)


#Analisis confirmatorio#
#Y <-Y[-c(63,123,151)]
#X1 <- X1[-c(63,123,151)]
#X2 <- X2[-c(63,123,151)]
#L1 <- L1[-c(63,123,151)]

#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#

#X2=vel viento  X1=Temperatura

T1<- X2
T1_0 <-sort(T1[!duplicated(T1)])

#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 N1#
N1 <- matrix(1,nrow=n, ncol=r1, byrow = TRUE)

for (i in 1:n) {
for(j in 1:r1){
if (T1[i] == T1_0[j]) {N1[i,j] <- 1
}else N1[i,j] <- 0
}
}
N1T <- t(N1)

##smoothing Matrix

NN1=(diag(X1))%*%N1   #diagonal variable temperatura por matriz de incidencia 1
NN1T<- t(NN1)

#Parametros smooth#
fi<- var(Y)
m<-11

#Lambda1#
u_k1<-runif(m, min = 0, max = 1)
a_k1 <-seq(0.1,0.2,by= 0.01)
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<-solve(N1T%*%Dv0%*%N1 + a_k1[i]*fi*K1)%*%N1T%*%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 <- 0.032

#Estimacion de los beta y F1#

F1_0  = (Jr1 - (V1%*%V1T)/r1)%*%solve(NN1T%*%NN1 + lambda1*K1)%*%NN1T%*%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)

epsilon_theta <- 0.00001
epsilon_beta <- 0.00001
epsilon_delta <- 0.00001
epsilon_f1 <- 0.00001
norma_theta <- 1000
norma_beta <- 1000
```

```
norma_delta <- 1000
norma_f1 <- 1000

beta_i <- betas
delta_i <- delta
f1_i <- F1_0
L_i <- sum (

( delta_i/2 ) - ( (log(16*pi))/2 ) -
( 0.5*log( ( (delta_i + 1)*(Y^3)*(((W%*%beta_i) + NN1%*%f1_i )^2))/( delta_i*Y + Y + delta_i*(((W%*%beta_i) + NN1%*%f1_i)^2) )^2 ) ) -
( ( Y*(delta_i + 1) )/( 4*(((W%*%beta_i) + NN1%*%f1_i)^2) ) ) -
( ( (delta_i^2)*( ((W%*%beta_i) + NN1%*%f1_i)^2 ) )/( 4*(delta_i + 1)*Y ) ) ) -

( (lambda1/2)*t(f1_i)%*%K1%*%f1_i )

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

I0 <- numeric(1)

while (norma_theta > epsilon_theta) {
while (norma_delta > epsilon_delta) {

while (norma_beta > epsilon_beta & norma_f1 > epsilon_f1) { #Algoritmo Backfitting

f.bs0 <- function(x){   #matriz penalizacion

((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) + (NN1%*%f1_i)
mu0 <- ((W%*%beta_i) + (NN1%*%f1_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
Da0 <- diag(as.vector(a0))
Dv0 <- diag(as.vector(v0))
Dva0 <- solve(Dv0)%*%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_a = eta + D_v,n*z
S0 <- (1/(2*(mu0*(delta_i+1)))) + ((delta_i*(mu0))/((delta_i+1)^3))*I0
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 <- (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))
r.abu0 <- sum(b0) + sum(u0)*delta_i + t(S0)%*%Da0%*%W%*%beta_i + t(S0)%*%Da0%*%NN1%*%f1_i
Betas <- solve(WT%*%Dv0%*%W)%*%WT%*%Dv0%*%(r.va0 - (NN1%*%f1_i))
F1 <-  (Jr1 - (V1%*%V1T)/r1)%*%solve(NN1T%*%Dv0%*%NN1 + (lambda1*K1))%*%NN1T%*%Dv0%*%(r.va0 - (W%*%Betas))

norma_beta <- sqrt((t(Betas-beta_i)%*%(Betas-beta_i))/((t(beta_i)%*%beta_i)))
norma_f1 <- sqrt((t(F1-f1_i)%*%(F1-f1_i))/((t(f1_i)%*%f1_i)))

beta_i <- Betas
f1_i <- F1

conteo1 <- conteo1 + 1
}

eta0   <- (W%*%beta_i) + (NN1%*%f1_i)
mu0    <- (W%*%beta_i) + (NN1%*%f1_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
Da0    <- diag(as.vector(a0))
Dv0    <- diag(as.vector(v0))
Dva0   <- solve(Dv0)%*%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)))  )
Dz0    <- diag(as.vector(z0))
r0     <- eta0 + Dva0%*%z0
S0     <- (1/(2*(mu0*(delta_i+1)))) + ((delta_i*(mu0))/((delta_i+1)^3))*I0
r.va0  <- r0 + Dva0%*%S0%*%delta_i
u0     <- (((delta_i^2)+(3*delta_i)+1)/(2*(delta_i^2)*(delta_i+1)^2)) + (((mu0^2)/(delta_i+1)^4))*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))
r.abu0 <- sum(b0) + sum(u0)*delta_i + t(S0)%*%Da0%*%W%*%beta_i + t(S0)%*%Da0%*%NN1%*%f1_i
Delta <- as.matrix((1/(sum(u0)))%*%(r.abu0 - (t(S0)%*%Da0%*%W%*%beta_i) - (t(S0)%*%Da0%*%NN1%*%f1_i)))

norma_delta <- sqrt(((delta_i - Delta[1])^2)/(delta_i^2))

delta_i <- Delta[1]

conteo2 <- conteo2 + 1

}

L <- sum (
```

```
( delta_i/2 ) - ( (log(16*pi))/2 ) -
( 0.5*log( ( (delta_i + 1)*(Y^3)*(((W%*%beta_i) + NN1%*%f1_i )^2))/( delta_i*Y + Y + delta_i*(((W%*%beta_i) + NN1%*%f1_i)^2) )^2 ) ) -
( ( Y*(delta_i + 1) )/( 4*(((W%*%beta_i) + NN1%*%f1_i)^2) ) ) -
( ( (delta_i^2)*( ((W%*%beta_i) + NN1%*%f1_i)^2 ) )/( 4*(delta_i + 1)*Y ) ) ) -

( (lambda1/2)*t(f1_i)%*%K1%*%f1_i )

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

}

#Estimaciones#
beta_i
delta_i
f1_i

#Grados de libertad#
sum(diag(NN1%*%solve(NN1T%*%Dv0%*%NN1 + (lambda1*K1))%*%NN1T%*%Dv0))

#Residuos parciales (estimaciones de F1  ponderadas por NN1)#
Yn <- (Y - W%*%beta_i)

#Grafico de ajuste de la funcion estimada#
plot(T1, Yn,
type="p",
col="black",
bg="black",
pch=21,
lwd=1,
main="",
cex.main=1,
cex.sub=1,
xlab="vel. viento",
ylab="F1 estimada",
ylim=c(-20,22),
#xlim=c(min(datos_finales$X2),max(datos_finales$X2)),
cex.lab=1,
axes=TRUE,
las=1,
sub="",
bty="o")
par(new=TRUE)
plot(T1_0,f1_i,type="l",lwd=1,col="blue", ylab="",xlab="",yaxt = "n",
xaxt = "n", ylim=c(-0.5,0.5), xlim=c(min(X2),max(X2)))

#Variable respuesta estimada#
Y_e <- W%*%beta_i + NN1%*%f1_i

#Estimados v/s observados
plot(Y,Y_e,type="p",cex.main=1,
cex.sub=1,
xlab="Y",
ylab="Y estimado",
col="black",
bg="gray20",
pch=21,
lwd=1,
bty="o")

#Matriz de informacion esperada de Fisher y covarianza#

Kbb <- WT%*%Dv0%*%W
Kbd <- WT%*%Da0%*%S0
Kbf1 <- WT%*%Dv0%*%NN1

Kdb <- t(Kbd)
Kdd <- sum(u0)
Kf1d <- NN1T%*%Da0%*%S0
Kdf1 <- t(Kf1d)

Kf1b <- t(Kbf1)
Kf1f1 <- NN1T%*%Dv0%*%NN1 + lambda1*K1

A1<-matrix(c(Kbb,Kbd,Kbf1),2,r1+p+1)
A2<-matrix(c(Kdb,Kdd,Kdf1),1,r1+p+1)
A3<-matrix(c(Kf1b,Kf1d,Kf1f1),r1,r1+p+1)

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

#Matriz hessiana#

dmu01 <- (delta_i/(delta_i*Y + Y + delta_i*mu0))+((Y*(delta_i+1))/(4*(mu0^2)))-((delta_i^2)/(4*Y*(delta_i+1)))-(1/(2*mu0))
dmu02 <- (1/(2*(mu0^2)))-((delta_i^2)/((delta_i*Y + Y + delta_i*mu0)^2))-((Y*(delta_i+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]

}
}

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

lbb <- WT%*%Dc%*%W
lbd <- WT%*%Da0%*%m
lbf1 <- WT%*%Dc%*%NN1

ldb <- t(lbd)
ldd <- sum(d)
lf1d <- NN1T%*%Da0%*%m
ldf1 <- t(lf1d)

lf1b <- t(lbf1)
lf1f1 <- NN1T%*%Dc%*%NN1 - lambda1*K1

l1<-matrix(c(Kbb,Kbd,Kbf1),2,r1+p+1)
l2<-matrix(c(Kdb,Kdd,Kdf1),1,r1+p+1)
l3<-matrix(c(Kf1b,Kf1d,Kf1f1),r1,r1+p+1)

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

#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_F1 <- se.f[(p+2):(r1+p+1)]
#SD_F2 <- desv.e[(r1+p+1):(r2+r1+p)] #

#SD_fi <- desv.e[r2+r1+p+1]
SD_fi <- se.f[p+1]
se.f[3]

Band1_F1 <- f1_i + 2*SD_F1
Band2_F1 <- f1_i - 2*SD_F1

mi<- min(f1_i) -0.5
ma <- max(f1_i) +0.5

#Bandas de confianza para F1#

plot(T1_0,f1_i,type='l',lwd=1, col="blue", ylim=c(mi,ma),sub="", xlab = "vel. viento", ylab="F1 estimada", las=1, main="")
par(new=TRUE)
plot(T1_0,Band1_F1,type='l', lty=5, col="red",ylim=c(mi,ma), xlab = "", ylab="",axes = F)
par(new=TRUE)
plot(T1_0,Band2_F1,type='l',lty=5, col="darkorange",ylim=c(mi,ma), xlab = "", ylab="",axes = F)

#Analisis residual#

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

#Residuos estanzarizados#

dif  = Y - mu0
phi  = sqrt((2 * ((delta_i + 1) ^ 2)) / ((2 * delta_i) + 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="black",
pch=21,
lwd=1,
main="",
sub="(a)",

cex.main=1,
xlab="Indice",
ylim=c(-4.5,3),
ylab="Residuos",
axes=TRUE,
las=1,
bty="o")

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=1, n= )

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

#Residuos propuestos por Jorgensen#

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

plot(rbJ,
type="p",
col="black",
bg="black",
pch=21,
lwd=1,
main="",
sub="(b)",
cex.main=1,

ylim=c(-4.5,3),
xlab="Indice",
ylab="Residuos",
axes=TRUE,
las=1,
bty="o")

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

hist(rbJ,
col = "white",
main = "",
sub="(b)",
xlab = "Residuos estandarizados",
ylab = "Frecuencia")

#QQ-plots#

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

qqnorm(rbJ,col="black",xlab="Cuantil teórico" ,sub="(b)", ylab="Cuantil empírico", main="")#cex.lab=1.5, #cex.sub=1.4
qqline(rbJ,col="black")
lillie.test(rbJ)



#Influencia local#

#Matrices delta#

#Esquema de perturbacion ponderacion de casos#

vt     = Y
vu     = mu0
```

```
vd     = delta_i
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))
De     = diag(as.vector(ve))
Deltab = WT%*%Da0%*%De
Deltaf = NN1T%*%Da0%*%De
Deltad = t(as.matrix(vb))

Deltacw    = rbind(Deltab,Deltaf,Deltad)


#Perturbacion en 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))


Deltab = WT%*%Da0%*%Dk%*%Dpsi
Deltaf = NN1T%*%Da0%*%Dk%*%Dpsi
Deltad = t(as.matrix(vro))%*%Dk
Deltares  = rbind(Deltab, Deltaf, Deltad)



#Perturabacion en el parametro de precision#

qsi  = - (vd*vt)/(((vd*vt) + vt +(vd*vu))^2) -
(vd*vt)/(4*(vu^2)) + ((vd^2)*(vd+2))/(4*vt*((vd+1)^2))
vpsi = - (1/2) + 1/(2*((vd+1)^2)) - (vt*(vt+vu))/(((vd*vt)+vt+(vd*vu))^2) +
vt/(4*vu)+ ((vd^2)*vu*(vd+3))/(4*vt*((vd+1)^3)) +
(vd*vu)/(vt*((vd+1)^3))
Dqsi = diag(as.vector(qsi))
Deltab = WT%*%Da0%*%Dqsi
Deltaf = NN1T%*%Da0%*%Dqsi
Deltad = t(as.matrix(vpsi))
Deltapre  = rbind(Deltab, Deltaf, Deltad)

caseweightspert = Deltacw #perturabacion ponderacion de casos
responsepert    = Deltares #perturbacion en la respuesta
precisionpert   = Deltapre #perturbacion en la precision

#leverage Generalizado#

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 = NN1T%*%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)
}

Bpre = function(I, M)
{
B =(t(Deltapre)%*%(I-M)%*%Deltapre)
return(B)
}

#Matrices Auxiliares#

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

b11    = cbind(matrix(0, p, p), matrix(0, p, 1),matrix(0,p,r1))
```

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

b211   = cbind(-solve(Lbeta), matrix(0, p, 1),matrix(0,p,r1))
b212   = cbind(matrix(0, 1, p), matrix(0, 1, 1),matrix(0,1,r1))
b213   = cbind(matrix(0, r1, p), matrix(0,r1,1),-solve(Lf))
B2     = rbind(b211, b212, b213)  # parametro delta

b311   = cbind(-solve(Lbeta), matrix(0, p, 1),matrix(0,p,r1))
b312   = cbind(matrix(0, 1, p), -Ldelta^(-1), matrix(0,1,r1))
b313   = cbind(matrix(0, r1, p), matrix(0,r1,1),matrix(0,r1,r1))
B3     = rbind(b311, b312, b313)  # funcion suavizada

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

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

FPC3            = Bcw(L1, B3)#F1
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))

#Respuesta#

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

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

FPR3 = Bres(L1, B3)#f1
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))

#Precision#

FPP1 = Bpre(L1, B1)#Beta
autovmaxbPP = eigen(FPP1, symmetric = TRUE)$val[1]
vetorpcbPP  = eigen(FPP1, symmetric = TRUE)$vec[,1]

FPP2 = Bpre(L1, B2)#delta
autovmaxdPP = eigen(FPP2, symmetric = TRUE)$val[1]
vetorpcdPP  = eigen(FPP2, symmetric = TRUE)$vec[,1]

FPP3 = Bpre(L1, B3)#f1
autovmaxfPP = eigen(FPP3, symmetric = TRUE)$val[1]
vetorpcfPP  = eigen(FPP3, symmetric = TRUE)$vec[,1]

vCiPP      = 2 * abs(diag(FPP1))
vCidPP     = 2 * abs(diag(FPP2))
vCifPP     = 2 * abs(diag(FPP3))

yest        = Y_e #fitted.values
Lby         = Deltalev
betas       = beta_i
v0          = rep(0, n)
eta         = as.vector(W%*%beta_i+ NN1%*%f1_i)
Da          = Da0
Do          = (Da0%*%W)
GL          = diag(Do%*%(L1[1:p,1:(p+1+r1)])%*%Lby )#leverage generalizado


#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",

axes=TRUE,
las=1,
bty="o")

identify(yest, GL, cex=1, n = )


#Perturbation ponderacion de casos#

#dmax

infl1  = vetorpcbPC
dmaxG1 = abs(infl1)

infl2  = vetorpcdPC
dmaxG2 = abs(infl2)

infl3  = vetorpcfPC
dmaxG3 = abs(infl3)

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 = )

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

plot(1:length(dmaxG3), dmaxG3, xlab = "id", ylim = c(0, 1), ylab = "dmaxpcf1", sub = "", cex = 0.7,bg="yellow",pch=21)
identify(1:length(dmaxG3), dmaxG3, n = )

#Ci

Cb1 = vCiPC
Cb1 = Cb1 / sum(Cb1)

Cb2 = vCidPC
Cb2 = Cb2/sum(Cb2)

Cb3 = vCifPC
Cb3 = Cb3/sum(Cb3)

limi = 2 * mean(Cb1)
plot(1:length(Cb1), Cb1,xlab = "índice", ylab = "Ci alfa",cex.lab = 1.2,  ylim = c(0, 0.5), cex = 1.2,bg="mediumorchid1",pch=21)
abline(h = limi, lty = 2, lwd = 3)
identify(1:length(Cb1),cex = 1.2, Cb1, n = )

limi = 2 * mean(Cb2)
plot(1:length(Cb2), Cb2, xlab = "índice", ylab = "Ci delta", cex.lab = 1.2, ylim = c(0, 0.5), cex = 1.2,bg="mediumorchid1",pch=21)
abline(h = limi, lty = 2, lwd = 3)
identify(1:length(Cb2),cex=1.2, Cb2, n = )

limi = 2 * mean(Cb3)
plot(1:length(Cb3), Cb3, xlab = "índice", ylab = "Ci Beta 1",cex.lab = 1.2, ylim = c(0, 0.5), cex = 1.2,bg="mediumorchid1",pch=21)
abline(h = limi, lty = 2, lwd = 3)
identify(1:length(Cb3),cex=1.2, Cb3, n = )


#Perturbacion en 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 = )

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

plot(1:length(dmaxG3), dmaxG3,  xlab = "id", ylim = c(0, 0.5), ylab = "dmaxpcf1", sub = "", cex = 0.7,bg="yellow",pch=21)
identify(1:length(dmaxG3), dmaxG3, n = )

#Ci
```

```
Cb1 = vCiPR
Cb1 = Cb1 / sum(Cb1)

Cb2 = vCidPR
Cb2 = Cb2 / sum(Cb2)

Cb3 = vCifPR
Cb3 = Cb3 / sum(Cb3)

limi = 2 * mean(Cb1)
plot(1:length(Cb1), Cb1,  xlab = "índice", ylab = "Ci alfa", cex.lab = 1.2,  ylim = c(0, 0.5), cex = 1.2,bg="lightsalmon",pch=21)
abline(h = limi, lty = 2, lwd = 3)
identify(1:length(Cb1),cex=1.2, Cb1, n = )

limi = 2 * mean(Cb2)
plot(1:length(Cb2), Cb2,  xlab = "índice", ylab = "Ci delta", cex.lab = 1.2,  ylim = c(0, 0.5), cex = 1.2,bg="lightsalmon",pch=21)
abline(h = limi, lty = 2, lwd = 3)
identify(1:length(Cb2),cex=1.2, Cb2, n = )

limi = 2 * mean(Cb3)
plot(1:length(Cb3), Cb3,  xlab = "index", ylab = "Ci Beta 1", cex.lab = 1.2,  ylim = c(0, 0.5), cex = 1.2,bg="lightsalmon",pch=21)
abline(h = limi, lty = 2, lwd = 3)
identify(1:length(Cb3),cex=1.2, Cb3, n = )

#Perturbacion en la precision#

#dmax

infl1  = vetorpcbPP
dmaxG1 = abs(infl1)

infl2  = vetorpcdPP
dmaxG2 = abs(infl2)

infl3  = vetorpcfPP
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 = )

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

plot(1:length(dmaxG3), dmaxG3, xlab = "id",  ylim = c(0, 0.5), ylab = "dmaxpcf1", sub = "", cex = 0.7,bg="yellow",pch=21)
identify(1:length(dmaxG3), dmaxG3, n = )

#Ci

Cb1 = vCiPP
Cb1 = Cb1/sum(Cb1)

Cb2 = vCidPP
Cb2 = Cb2 / sum(Cb2)

Cb3 = vCifPP
Cb3 = Cb3 / sum(Cb3)

limi = 2 * mean(Cb1)
plot(1:length(Cb1), Cb1,  xlab = "índice", ylab = "Ci alfa", cex.lab = 1.2,  ylim = c(0, 0.5), cex = 1.2, bg="lightskyblue",pch=21)
abline(h = limi, lty = 2, lwd = 3)
identify(1:length(Cb1),cex=1.2, Cb1, n = )

limi = 2 * mean(Cb2)
plot(1:length(Cb2), Cb2,  xlab = "índice", ylab = "Ci delta", cex.lab = 1.2,  ylim = c(0, 0.5), cex = 1.2, bg="lightskyblue",pch=21)
abline(h = limi, lty = 2, lwd = 3)
identify(1:length(Cb2),cex=1.2, Cb2, n = )

limi = 2 * mean(Cb3)
plot(1:length(Cb3), Cb3,  xlab = "índice", ylab = "Ci Beta 1",cex.lab = 1.2,  ylim = c(0, 0.5), cex = 1.2,bg="lightskyblue",pch=21)
abline(h = limi, lty = 2, lwd = 3)
identify(1:length(Cb3),cex=1.2, Cb3, n = )


#CR Analisis confirmatorio#

100*(abs((42.39048 - delta_i)/42.39048))
100*(abs((4.846624- se.f[3])/4.846624))

100*(abs((3.555832 - beta_i[1])/3.555832))
100*(abs((1.93175150 - se.f[1])/1.93175150))

100*(abs((0.427818 - beta_i[2])/0.427818))
100*(abs((0.02122493 - se.f[2])/0.02122493))
```