PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

MODELO DE REGRESIÓN SEMIPARAMÉTRICO ROBUSTO

TESIS PARA OPTAR POR EL GRADO DE MAGÍSTER EN ESTADÍSTICA

Presentado por:

Henry John Esquivel Segura

Asesor: Dr. Cristian Luis Bayes Rodríguez

Miembros del jurado:
Dr. Cristian Luis Bayes Rodríguez
Dr. Luis Hilmar Valdivieso Serrano
Dra. Rocio Paola Maehara Aliaga

Lima, Diciembre 2018

Dedicatoria

Dedico está tesis a mi querida madre, Bertha Segura Atencia, quién fue la más sacrificada desde el inicio de mi vida universitaria, quíen me esperaba despierta hasta altas horas de la noche, 12:30 A.M. aproximadamente, para calentarme la comida; y se despertaba a la misma hora que yo lo hacia 5:00 A.M. aproximadamente, para prepararme el desayuno para irme a estudiar y/o trabajar. Sé que no fue fácil lograr todo lo que le prometí, que cada día tuviste que darme frases de aliento como "hijo todo te va ir bien", "Los problemas van a pasar, tú lo vas a lograr" y otras más, de esa manera poder continuar en la carrera de matemática, sacrifico reuniones familiares, viajes y más, pero quiero decirte que lo logró y que es ella quíen se merece este grado, lo único que yo hice bien fue estudiar.

A mi padre, con su carácter fuerte, disciplinado y su forma de decir las cosas claras y concisas, me volviste una persona fuerte y franca. A mi querido abuelo, Juan Pablo Segura, gracias a él aprendi a tener pasión por el fútbol y ser patriota, aún recuerdo las historias que me contabas y las frases tan suyas " Acaso no puedes patear la pelota, mis hijos son unos buenos para nada", "si nos tiramos un trancazo", estoy feliz porque desde el cielo me guiaste y aconpañaste en esta ardua aventura. A mis hermanos, que teniendo carácteres diferentes llegamos a entendernos y apoyarnos en cada proyecto que emprendimos.

A mi compadre y casi como un hermano, Frank Alvarez, que en la vida universitaria pasamos por bastantes aventuras y compartimos un pan cuando no había comida en el comedor, y no me puedo olvidar de mi amiga "la nera", mi confidente y salvadora de algunas metidas de patas. Por último, a la "estadística", aquella persona que respeto, valoro y estimo mucho, la que me hablaba de temas estadísticos en todo el viaje de la universidad a su casa, gracias que por ti conocí está carrera y por ti ahora presentaré este trabajo de tesis.

Agradecimentos

En la vida nos pasamos el tiempo tratando de hacer cosas que cambie el rumbo de nuestra sociedad, pero cada cambio empieza por manifestaciones pequeñas, y una de esas cosas pequeñas fue la elaboración de está tesis que me ha costado muchas horas de esfuerzo y estres, pero cuando sentía que ya no podía continuar, recibía algunas frases sarcásticas e irónicas que me hacian seguir adelante, por eso quiero agradecer profundamente al quién considero una gran persona, y más que un maestro, un buen amigo, el Dr. Cristian Bayes Rodriguez, mi asesor del presente tema de tesis, por su apoyo continuo durante el desarrollo de la investigación.

Cuando inicié el proyecto de estudiar la carrera de estadística, ese cambio generó un poco de temor en está nueva aventura, pero eso cambió al recibir las primeras clases de probabilidad, por eso agradezco de corazón a la Mg. Rosario Bullón Cuadrado, quién tuvo el tiempo de sentarse conmigo muchas horas y darme consejos de vida que me sirvieron para continuar y no flaquear en mi nuevo amor, la estadística.

La vida en la Pontificia Universidad Católica del Perú fue otro reto más que tuve que enfrentar, y la primera persona que conocí con una personalidad parco, sensible, amable y bondadoso y sus consejos que fueron una herramienta para conocer otro mirada de la estadística, por eso agradezco al Dr. José Julio Flores Delgado, quién me apoyo desde el primer día que me conoció.

Por último me queda agradecer la gran predisposición a algunas inquietudes que tenía y su personalidad concienzuda al Dr. Luis Valdivieso Serrano. Así mismo, agradecer al Dr. Giancarlo Sal Y Rosas, quién tuvo la paciencia y una gran tolerancia de aceptar de buena forma mis intervenciones en sus horas de clase y fuera de ellas.

Resumen

El presente trabajo de tesis presenta un modelo de regresión semiparamétrico con errores t-Student, que permite estudiar el comportamiento de una variable dependiente dado un conjunto de variables explicativas cuando los supuestos de linealidad y normalidad no se cumplen. La estimación de los parámetros se realiza bajo el enfoque bayesiano a través del algoritmo de Gibbs. En el estudio de simulación se observa que el modelo propuesto es más robusto ante la presencia de valores atípicos que el usual modelo regresión semiparamétrico normal. Asimismo se presenta una aplicación con datos reales para ilustrar esta característica.
Palavras-clave: Regresión spline penalizada, modelo lineal mixto, distribución t-student, inferencia bayesiana, algoritmo de Gibbs.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Consideraciones preliminares

En muchos casos un investigador está interesado en estudiar el comportamiento de una variable dependiente dado un conjunto de variables explicativas. Un enfoque común a este problema consiste en especificar un modelo de regresión y estimar la media como una función lineal de las variables explicativas. Sin embargo, los supuestos de normalidad y linealidad no son siempre satisfechos. Hastie y Tibshirani (1986) proponen como alternativa, una clase de modelos semiparamétricos denominados modelos aditivos generalizados donde se relaja el supuesto de linealidad. Crainiceanu et al. (2005) presentan como realizar la estimación de este tipo de modelos bajo un enfoque bayesiano.
Un supuesto usual en los modelos aditivos generalizados es el de normalidad de los datos; sin embargo, se sabe que la estimación de los paramétros del modelo puede verse fuertemente afectadas ante la presencia de valores atípicos. Por este motivo, distribuciones de probabilidad con colas más pesadas que la distribución normal han sido propuestas en la literatura como una alternativa a la usual suposición de normalidad de los errores en modelos de regresión. Estas distribuciones tienen la ventaja de incorporar observaciones que son consideradas atípicas bajo el supuesto de normalidad. Entre las distribuciones que presentan colas pesadas se tiene a la distribución t-Student, la distribución Slash, la distribución normal contaminada, entre otras.

En la presente tesis se propone estudiar un modelo de regresión semiparamétrico, bajo el enfoque bayesiano, considerando que los errores siguen una distribución t-Student.

1.2. Objetivos

El objetivo general de la tesis es estudiar las propiedades del modelo de regresión semiparamétrico con una función splines de base radial cúbica con errores t-Student; así como, estimar y aplicar este modelo a un conjunto de datos reales bajo el enfoque bayesiano.

- Revisar la literatura acerca de las diferentes propuestas de modelos de regresión semiparamétricos (MRS).
- Estudiar las propiedades e implementar la estimación del modelo de regresión semiparamétrico con una función splines de base radial cúbica con errores t-Student bajo el enfoque bayesiano.

- Realizar estudios de simulación para comparar el modelo de regresión semiparamétrico con errores t-Student con el que asume errores normales.
- Aplicar el modelo a un conjunto de datos reales.

1.3. Organización del trabajo

En el Capítulo 2, presentamos conceptos previos al desarrollo del modelo de regresión semiparamétrico robusto (MRSR). Se revisará en primer lugar, el modelo lineal mixto con errores normales, posteriormente se describirá el modelo de regresión semiparamétrica utilizando una base de splines cúbicos y su conexión con el modelo antes mencionado. También, se estudiará la distribución t-Student multivariada presentando algunas propiedades, en particular se estudiará su presentación como una mixtura en la escala de una distribución Normal con una distribución Gamma.

En el Capítulo 3 se estudiará el modelo de regresión semiparamétrico robusto (MRSR) utilizando bases de splines cúbicos bajo un enfoque bayesiano. Se obtendrán las distribuciones a posteriori y las distribuciones condicionales con las que se implementará el algoritmo de Gibbs. Finalmente se presentará el criterio de información de desvio (DIC), el cuál es una medida de comparación de modelos.

En el Capítulo 4 se realizará un estudio de simulación para comparar el modelo de regresión semiparámetrico robusto con splines cúbico, cuando se asume que los errores presentan una distribución t-Student y cuando se asume que tienen una distribución normal bajo tres escenarios distintos: sin la presencia de valores atípicos y con la presencia de dos y cuatro valores atípicos. La estimación desde la perspectiva bayesiana se realizará en el programa R implementando para ello un código propio con el algoritmo de Gibbs descrito en la sección 3.3. En ambos modelos se evaluará como medida de bondad de ajuste al error cuadrático medio y al criterio de información de desvio como medida de comparación del modelo.

En el Capítulo 5 se presentará una aplicación del modelo de regresión semiparamétria robusta utilizando los datos analizados previamente por Staudenmayer, Lake y Wand (2009). Finalmente, en el Capítulo 6 discutimos algunas conclusiones obtenidas en este trabajo. Para lo cual, se analizará las ventajas y desventajas de los modelo de regresión semiparamétrico con una función de splines de base radial cúbica con errores normales y con errores t-Student.

En el Apéndice A se presentará la implementación del algoritmo de Gibbs para el modelo de regresión semiparamétrico robusto en el programa R. En el Apéndice B se presentará todas las gráficas de las cadenas y gráficos de la función de autocorrelación de la aplicación.

Capítulo 2

Conceptos Preliminares

En el presente capítulo se procederá a revisar la formulación del modelo de regresión semiparamétrico utilizando una base de funciones spline radial cúbica y su conexión con el modelo lineal mixto. También se estudiará la distribución t-Student multivariada presentando algunas de sus propiedades, en particular se estudiará su representación como una mixtura en la escala de una distribución normal con una distribución gamma.

2.1. El modelo lineal mixto

El modelo lineal mixto es dado por

$$\boldsymbol{Y} = \boldsymbol{X}\ \boldsymbol{\beta} + \boldsymbol{Z}\ \boldsymbol{b} + \boldsymbol{\varepsilon} \tag{2.1}$$

donde $\boldsymbol{Y}$ es un vector de las observaciones de la variable respuesta de dimensión $n \times 1$, $\boldsymbol{\beta}$ es el vector de parámetros fijos desconocidos de dimensión $p \times 1$, $\boldsymbol{X}$ es una matriz de dimensión $n \times p$ que contiene los valores de las variables explicativas de los efectos fijos, $\boldsymbol{b}$ es el vector de efectos aleatorios de dimensión $q \times 1$, $\boldsymbol{Z}$ es una matrix de dimensión $n \times q$ que contiene los valores de las variables explicativas de los efectos aleatorios, y $\boldsymbol{\varepsilon}$ es un vector de errores no observables $n \times 1$. Se asume que $\boldsymbol{b}$ y $\boldsymbol{\varepsilon}$ sean normalmente distribuidos e independientes, tal que

$$\begin{aligned} \boldsymbol{b} &\sim N_q(\boldsymbol{0}, \boldsymbol{G}) \\ \boldsymbol{\varepsilon} &\sim N_n(\boldsymbol{0}, \boldsymbol{R}) \end{aligned} \tag{2.2}$$

donde $\boldsymbol{G}$ es una matriz de dimensión $q \times q$ y $\boldsymbol{R}$ es una matriz de dimensión $n \times n$, las cuales son definidas positivas y $\boldsymbol{0}$ es un vector de ceros.

De las ecuaciones (2.1) y (2.2) se obtiene la distribución condicional de $\boldsymbol{Y}$ dado $\boldsymbol{b}$,

$$\boldsymbol{Y}|\boldsymbol{b} \sim N_n(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \boldsymbol{R}). \tag{2.3}$$

Luego, la función verosimilitud aumentada de $\boldsymbol{Y}$ y $\boldsymbol{b}$ esta dada por:

$$f(\boldsymbol{y}, \boldsymbol{b}) = f(\boldsymbol{y}|\boldsymbol{b}) f(\boldsymbol{b}). \tag{2.4}$$

Reemplazando las distribuciones de $\boldsymbol{Y}|\boldsymbol{b}$ y $\boldsymbol{b}$ en (2.4), se obtiene:

$$f(\boldsymbol{y},\boldsymbol{b}) = (2\pi)^{-n/2}|\boldsymbol{R}|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})^T\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})\right\}$$

$$\times (2\pi)^{-p/2}|\boldsymbol{G}|^{-1/2}\exp\left\{-\frac{1}{2}\boldsymbol{b}^T\boldsymbol{G}^{-1}\ \boldsymbol{b}\right\}$$

$$= (2\pi)^{-(n+p)/2}\left|\begin{pmatrix}\boldsymbol{R} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{G}\end{pmatrix}\right|^{-1/2}\exp\left\{-\frac{1}{2}[(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})^T\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})+\boldsymbol{b}^T\boldsymbol{G}^{-1}\ \boldsymbol{b}.]\right\} \tag{2.5}$$

donde $|\boldsymbol{A}|$ denota la determinante de la matriz $\boldsymbol{A}$. Para maximizar la ecuación (2.5) para $\boldsymbol{\beta}$ y $\boldsymbol{b}$ se requiere minimizar la siguiente expresión

$$(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})^T\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})+\boldsymbol{b}^T\boldsymbol{G}^{-1}\ \boldsymbol{b}. \tag{2.6}$$

2.2. Regresión semiparamétrica por funciones splines

El modelo de regresión semiparamétrico propuesto en Crainiceanu et al. (2005) es dado por

$$y_i = m(x_i) + \varepsilon_i \tag{2.7}$$

donde y_i es el valor de la variable respuesta, ε_i es un error aleatorio con $E[\varepsilon_i] = 0$ y $Var(\varepsilon_i) = \sigma^2_\varepsilon$, x_i es el i-esimo valor de la covariable y m es una función suave [1] no especificada que necesita ser estimada usando los datos.

Según Crainiceanu et al. (2005) la función m puede ser modelada utilizando splines de base radial cúbica. En particular los autores recomiendan esta elección debido a que reduce la correlación entre los parámetros de las distribuciones a posteriori del modelo, ya que como es conocido, una alta correlación entre los parámetros genera problemas de convergencia en el algoritmo de Gibbs (Gelman et al., 2014).

Bajo esta especificación $m(x_i)$ es dada por

$$m(x_i) = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K}\boldsymbol{u}_k|x_i-\kappa_k|^3. \tag{2.8}$$

donde $\beta_0,\ \beta_1$ son los coeficientes de regresión de los parámetros fijos, $u = (u_1,\ u_2,\ldots,\ u_K)^T$ es el vector de coeficientes de la base de las funciones spline radial cúbica, x_i es el valor de la covariable en la i-ésima observación y $\kappa_1 < \kappa_2 < \ldots < \kappa_K$ son nodos fijos. Según Ruppert et al. (2003) se aconseja considerar un número de nodos que sea suficientemete grande (tipicamente de 5 hasta 20) para asegurar la flexibilidad deseada, y donde κ_K es el cuantil

[1] $m(.)$ es suave si es de clase C^∞; es decir si existen sus derivadas de todos las ordenes y son continuas

$\frac{\kappa}{K+1}$ de los valores de la covariables $x_1, x_2, \ldots, x_n$.

Crainiceanu et al. (2005) y Ruppert et al. (2003) proponen que para estimar $\boldsymbol{u}$ se debe minimizar

$$\sum_{i=1}^{n}\{y_i - m(x_i)\}^2 + \lambda \boldsymbol{u}^T \boldsymbol{\Omega}_k \boldsymbol{u} \tag{2.9}$$

con el fin de evitar un sobreajuste, donde λ es denominado el parámetro de suavizado, y la (l,k)_entrada de $\boldsymbol{\Omega}_k$ es $|\kappa_l - \kappa_k|^3$ y como podemos notar se penalizan únicamente los coeficientes $|x - \kappa_k|^3$.

Sea $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$, $\boldsymbol{X}$ la matriz con i-esima fila $\boldsymbol{X}_i = (1, x_i)$, y $\boldsymbol{Z}_K$ la matriz con la i-esima fila $\boldsymbol{Z}_{Ki} = [|x_i - \kappa_1|^3, \ldots, |x_i - \kappa_K|^3]$. Si se divide la ecuación (2.9) por $\sigma_\varepsilon^2 > 0$ se obtiene

$$\frac{1}{\sigma_\varepsilon^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}_K\boldsymbol{u})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}_K\boldsymbol{u}) + \frac{\lambda}{\sigma_\varepsilon^2}\, \boldsymbol{u}^T\boldsymbol{\Omega}_K\boldsymbol{u},$$

donde $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ y $\boldsymbol{u} = (u_1, \ldots, u_K)^T$ son un vector de parámetros fijos y un vector de efectos aleatorios, respectivamente. Luego si reparametrizamos $\sigma_u^2 = \sigma_\varepsilon^2/\lambda$ obtenemos

$$\frac{1}{\sigma_\varepsilon^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z_K}\boldsymbol{u})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z_K}\boldsymbol{u}) + \frac{1}{\sigma_b^2}\, \boldsymbol{u}^T\boldsymbol{\Omega}_K\boldsymbol{u}. \tag{2.10}$$

Finalmente, si consideramos $\boldsymbol{b} = \boldsymbol{\Omega}_K^{1/2}\boldsymbol{u}$ y a $\boldsymbol{Z} = \boldsymbol{Z}_K\boldsymbol{\Omega}_K^{-1/2}$, la expresión (2.10) puede ser escrita como

$$\frac{1}{\sigma_\varepsilon^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{b}) + \frac{1}{\sigma_b^2}\, \boldsymbol{b}^T\boldsymbol{b}. \tag{2.11}$$

Podemos observar que la expresión en (2.11) es similar a la que obtendriamos con un modelo lineal mixto dada en la expresión (2.6) con $\boldsymbol{G} = \sigma_b^2\, I_K$ y $\sigma_\varepsilon^2\, I_n$,

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon} \\ \boldsymbol{b} &\sim N(\boldsymbol{0}, \sigma_b^2\, I_K) \\ \boldsymbol{\varepsilon} &\sim N(\boldsymbol{0}, \sigma_\varepsilon^2\, I_n) \end{aligned}$$

donde I_r representa la matriz identidad de dimensión $r \times r$. Por lo tanto para estimar el modelo de regresión semiparamétrico dado en (2.7) y (2.8) podemos utilizar las metodologías propuestas para un modelo lineal mixto. En este trabajo consideraremos la estimación de los parámetros del modelo bajo inferencia bayesiana.

2.3. Distribución t-Student multivariada

Un vector aleatorio $\boldsymbol{T}$ de dimensión $p \times 1$ tiene distribución t-Student multivariada con ν grados de libertad, con parámetro de localización $\boldsymbol{\mu}$ de dimensión $p \times 1$, y matriz de varianzas-covarianzas $\boldsymbol{\Sigma}$ de dimensión $p \times p$, simétrica y definida positiva, si su función de densidad de

probabilidad es dada por

$$f(\boldsymbol{t};\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)=\frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}}\left[1+\frac{1}{\nu}(\boldsymbol{t}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})\right]^{-(\nu+p)/2},\quad \boldsymbol{t}\in\mathbb{R}^p$$

Denotaremos en adelante por $\boldsymbol{T}\sim t_\nu(\boldsymbol{\mu},\boldsymbol{\Sigma})$ para indicar que la distribución de probabilidades de $\boldsymbol{T}$ es t-Student multivariada con parámetros $\boldsymbol{\mu},\boldsymbol{\Sigma}$ y ν.

Si $\boldsymbol{T}\sim t_\nu(\boldsymbol{\mu},\boldsymbol{\Sigma})$ no es difícil probar que $E[\boldsymbol{T}]=\boldsymbol{\mu}$, si $\nu>1$ y su matriz de varianzas-covarianzas es $V(\boldsymbol{T})=\frac{\nu}{\nu-2}\boldsymbol{\Sigma}$, si $\nu>2$.

Proposición. 2.1. *Sea $\boldsymbol{T}$ un vector aleatorio de dimensión $p\times 1$ y W una variable aleatoria tal que*

$$\boldsymbol{T}|W=w\sim N_p\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\frac{1}{w}\right) \tag{2.12}$$

y

$$W\ \sim Gamma\left(\frac{\nu}{2},\frac{\nu}{2}\right),$$

entonces $\boldsymbol{T}\sim t_\nu(\boldsymbol{\mu},\boldsymbol{\Sigma})$.

Demostración. *En este caso la distribución conjunta de $(\boldsymbol{T},W)$, se puede escribir como:*

$$\begin{aligned}f_{\boldsymbol{T},W}(\boldsymbol{t},w) &= f_{\boldsymbol{T}|W=w}(\boldsymbol{t})f_W(w)\\ &= \frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}\frac{1}{w}\right|^{1/2}}\exp\left\{\frac{-1}{2}(\boldsymbol{t}-\boldsymbol{\mu})^T\left(\boldsymbol{\Sigma}\frac{1}{w}\right)^{-1}(\boldsymbol{t}-\boldsymbol{\mu})\right\}\\ &\quad\times\frac{\left(\frac{\nu}{2}\right)^{\nu/2}w^{\nu/2-1}}{\Gamma\left(\frac{\nu}{2}\right)}\exp\left\{-\frac{\nu w}{2}\right\}.\end{aligned}$$

Utilizando la propiedad de la determinante de una matriz y realizando las respectivas operaciones, se obtiene

$$\begin{aligned}f_{\boldsymbol{T},W}(\boldsymbol{t},w) &= \frac{w^{p/2}w^{\nu/2-1}\left(\frac{\nu}{2}\right)^{\nu/2}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}\Gamma\left(\frac{\nu}{2}\right)}\exp\left\{\frac{-1}{2}(\boldsymbol{t}-\boldsymbol{\mu})^T\left(\boldsymbol{\Sigma}\frac{1}{w}\right)^{-1}(\boldsymbol{t}-\boldsymbol{\mu})-\frac{\nu w}{2}\right\}.\\ &= \frac{w^{(\nu+p)/2-1}\left(\frac{\nu}{2}\right)^{\nu/2}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}\Gamma\left(\frac{\nu}{2}\right)}\exp\left\{-\frac{\nu w}{2}[1+\frac{1}{\nu}(\boldsymbol{t}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})]\right\}.\end{aligned}$$

Integrando la función de densidad conjunta de $(\boldsymbol{T},W)$ con respecto a W, se obtiene la función

de densidad de $\boldsymbol{T}$:

$$\begin{aligned} f_{\boldsymbol{T}}(\boldsymbol{t}) &= \int_0^\infty f_{\boldsymbol{T},W}(\boldsymbol{t},w)\,dw \\ &= \frac{\left(\frac{\nu}{2}\right)^{\nu/2}}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}\Gamma\left(\frac{\nu}{2}\right)}\int_0^\infty \exp\left\{-\frac{\nu w}{2}[1+\frac{1}{\nu}(\boldsymbol{t}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})]\right\} w^{(\nu+p)/2-1}dw. \end{aligned} \tag{2.13}$$

Si consideramos el cambio de variable

$$x = \frac{\nu w}{2}\left[1+\frac{1}{\nu}(\boldsymbol{t}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})\right] \tag{2.14}$$

se obtiene :

$$w = \frac{2x}{\nu}\left[1+\frac{1}{\nu}(\boldsymbol{t}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})\right]^{-1} \tag{2.15}$$

$$\frac{dw}{dx} = \frac{2}{\nu}\left[1+\frac{1}{\nu}(\boldsymbol{t}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})\right]^{-1} \tag{2.16}$$

donde $x \to 0$ cuando $w \to 0$ y $x \to \infty$ cuando $w \to \infty$.

En la ecuación (2.13) se reemplazan las ecuaciones (2.14), (2.16) y (2.15) del cambio de variable

$$\begin{aligned} f_{\boldsymbol{T}}(\boldsymbol{t}) &= \frac{\left(\frac{\nu}{2}\right)^{\nu/2}}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}\Gamma\left(\frac{\nu}{2}\right)}\int_0^\infty \exp\{-x\}\left(\frac{2x}{\nu}\left[1+\frac{1}{\nu}(\boldsymbol{t}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})\right]^{-1}\right)^{(\nu+p)/2-1} \\ &\quad \times\frac{2}{\nu}\left[1+\frac{1}{\nu}(\boldsymbol{t}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})\right]^{-1}dx \\ &= \frac{1}{(\nu\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}\Gamma\left(\frac{\nu}{2}\right)}\left[1+\frac{1}{\nu}(\boldsymbol{t}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})\right]^{-(\nu+p)/2}\int_0^\infty \exp\{-x\}\,x^{(\nu+p)/2-1}dx \end{aligned}$$

De la definición de la función gamma en la ecuación anterior, se obtiene la función de densidad de la t-Student multivariada:

$$f_{\boldsymbol{T}}(\boldsymbol{t}) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{(\nu\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}\Gamma\left(\frac{\nu}{2}\right)}\left[1+\frac{1}{\nu}(\boldsymbol{t}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})\right]^{-(\nu+p)/2}.$$

Capítulo 3

Modelo de Regresión Semiparamétrico Robusto

El modelo de regresión semiparamétrico robusto es dado por:

$$y_i = m(x_i) + \varepsilon_i,\ \forall i = 1, 2, \ldots, n \tag{3.1}$$

$$\varepsilon_i \sim t_{\nu_\varepsilon}(0, \sigma_\varepsilon^2)$$

donde y_i es el valor de la variable respuesta para la i-ésima observación, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ son errores aleatorios independientes entre si, $\sigma_\varepsilon^2 > 0$, $\nu_\varepsilon > 0$ es el grado de libertad de la distribución t-Student de los errores, y x_i es el valor de la covariable para la i-ésima observación, siendo $m(x_i)$ modelado por una función spline de base radial cúbica

$$m(x_i) = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k |x_i - \kappa_K|^3. \tag{3.2}$$

Como se demostró en la sección 2.2 el modelo dado en (3.1) y (3.2) se puede expresar como un modelo mixto:

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{Z}\,\boldsymbol{b} + \boldsymbol{\varepsilon} \\ \boldsymbol{b} &\sim N_K(0, \sigma_b^2 I_K) \\ \varepsilon_i &\sim t_{\nu_\varepsilon}(0, \sigma_\varepsilon^2),\ \forall i = 1, 2, \cdots, n \end{aligned} \tag{3.3}$$

donde $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)^T$ es el vector de los valores de las observaciones de la variable respuesta de dimensión $n \times 1$, $\boldsymbol{X}$ es una matriz de dimensión $n \times 2$ con la i-ésima fila $\boldsymbol{X}_i = [1, x_i]$ que contiene los valores de la variable explicativa, $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ es el vector de parámetros fijos, $\boldsymbol{Z}$ es una matriz de dimensión $n \times K$, donde K es el número de nodos fijos de la base de las funciones spline radial cúbica, que por la ecuación (2.10) tiene forma $\boldsymbol{Z} = \boldsymbol{Z}_K \boldsymbol{\Omega}_K^{-1/2}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$ es el vector de los errores aleatorios con distribución $t_{\nu_\varepsilon}(0, \sigma_\varepsilon^2)$ de dimensión $n \times 1$ y $\boldsymbol{b} = (b_1, b_2, \ldots, b_K)^T$ es dado por $\boldsymbol{b} = \boldsymbol{\Omega}_K^{1/2} \boldsymbol{u}$ e $\boldsymbol{I}_K$ es la matriz identidad $K \times K$, y $\sigma_b^2 > 0$.

Luego, utilizando la proposición 2.1, el modelo dado en (3.3) se puede escribir como:

$$\begin{aligned} \boldsymbol{Y}|\boldsymbol{b}, w_{\varepsilon_1}, ..., w_{\varepsilon_n} &\sim N_n(\boldsymbol{X\beta} + \boldsymbol{Zb}, \sigma_\varepsilon^2 \boldsymbol{W}_\varepsilon^{-1}) \\ \boldsymbol{b} &\sim N_K(0, \sigma_b^2 \boldsymbol{I}_K) \\ w_{\varepsilon_i} &\sim Gamma\left(\frac{\nu_\varepsilon}{2}, \frac{\nu_\varepsilon}{2}\right) \end{aligned} \tag{3.4}$$

donde $\boldsymbol{W}_\varepsilon = diag(w_1, w_2, ..., w_n)$. Luego, la función de verosimilitud aumentada es dada por:

$$\begin{aligned} L(\boldsymbol{y}|\boldsymbol{\theta}) &= f(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{W}_\varepsilon) f(\boldsymbol{b}|\sigma_b^2) \prod_{i=0}^{n} f(w_{\varepsilon_i}) \\ &\propto \frac{1}{\left|\sigma_\varepsilon^2 \boldsymbol{W}_\varepsilon^{-1}\right|^{1/2}} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zb})^T \boldsymbol{W}_\varepsilon (\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zb})\right\} \\ &\quad \times \left|\sigma_b^2 \boldsymbol{I}\right|^{-1/2} \exp\left\{-\frac{1}{2\sigma_b^2}\boldsymbol{b}^T\boldsymbol{b}\right\} \prod_{i=0}^{n} \left(\frac{\left(\frac{\nu_\varepsilon}{2}\right)^{\frac{\nu_\varepsilon}{2}}}{\Gamma(\frac{\nu_\varepsilon}{2})} w_{\varepsilon_i}^{\frac{\nu_\varepsilon}{2}-1} \exp\left\{-\frac{\nu_\varepsilon}{2} w_i\right\} \right) \end{aligned} \tag{3.5}$$

el vector de parámetros del modelo estaría conformado por $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{b}, \sigma_\varepsilon^2, \sigma_b^2, \nu_\varepsilon)^T$.

3.1. Distribuciones a priori y a posteriori

Según Hoff (2009), la distribución a posteriori se puede escribir de la siguiente manera:

$$f(\boldsymbol{\theta}|\boldsymbol{y}) \propto L(\boldsymbol{y}|\boldsymbol{\theta}) f(\boldsymbol{\beta}) f(\sigma_b^2) f(\sigma_\varepsilon^2) f(\nu_\varepsilon) \tag{3.6}$$

donde se asume independencia entre las distribuciones a priori. De las ecuaciones (3.6) y (3.5) se tiene a la distribución a posteriori como:

$$\begin{aligned} f(\boldsymbol{\theta}|\boldsymbol{y}) &\propto \frac{1}{\left|\sigma_\varepsilon^2 \boldsymbol{W}_\varepsilon^{-1}\right|^{1/2}} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zb})^T \boldsymbol{W}_\varepsilon (\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{Zb})\right\} \\ &\quad \times \left|\sigma_b^2 \boldsymbol{I}_K\right|^{-1/2} \exp\left\{-\frac{1}{2\sigma_b^2}\boldsymbol{b}^T\boldsymbol{b}\right\} \prod_{i=0}^{n} \left(\frac{\left(\frac{\nu_\varepsilon}{2}\right)^{\frac{\nu_\varepsilon}{2}}}{\Gamma(\frac{\nu_\varepsilon}{2})} w_i^{\frac{\nu_\varepsilon}{2}-1} \exp\left\{-\frac{\nu_\varepsilon}{2} w_i\right\} \right) \\ &\quad \times f(\boldsymbol{\beta})\ f(\sigma_b^2)\ f(\sigma_\varepsilon^2)\ f(\nu_\varepsilon) \end{aligned}$$

Siguiendo a Crainiceanu et al. (2005), Geweke (1993) y Rosa et al. (2003) las distribuciones

a priori que consideraremos son las siguientes:

$$\begin{aligned} \boldsymbol{\beta} &\sim N_2(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0) \\ \sigma_\varepsilon^2 &\sim Gamma - inv(a_\varepsilon, c_\varepsilon) \\ \sigma_b^2 &\sim Gamma - inv(a_b, c_b) \\ \nu_\varepsilon &\sim Exp(d_\varepsilon) I_{(2,\infty)}(\nu_\varepsilon) \end{aligned} \tag{3.7}$$

Geweke (1993) considera que ν_ε tiene una distribución exponencial truncada en el intervalo abierto $(2, \infty)$ para que la varianza y media de y_i existan y $d_\varepsilon = 0.1$. Asimismo, $H \sim Gamma - inv(a, c)$ denota que H sigue una distribución gamma inversa con función de densidad

$$f(h) = \frac{c^a}{\Gamma(a)} h^{-(a+1)} \exp\left(-\frac{c}{h}\right). \tag{3.8}$$

Similar a Lunn, Jackson, Best, Thomas y Spiegelhalter (2013) se considerará $a_\varepsilon = a_b = c_\varepsilon = c_b = 0.001$ de tal manera que se obtenga una distribución a priori aproximadamente no informativa. Entonces, teniendo todas las distribuciones a priori definidas, la distribución a posteriori de la ecuación (3.7) quedaría de la siguiente forma:

$$\begin{aligned} f(\boldsymbol{\theta}|\boldsymbol{y}) \propto & \left|\sigma_\varepsilon^2 \boldsymbol{W}_\varepsilon^{-1}\right|^{-1/2} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b})^T \boldsymbol{W}_\varepsilon (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{b})\right\} \\ & \times \prod_{i=1}^{n} \left(\frac{\left(\frac{\nu_\varepsilon}{2}\right)^{\frac{\nu_\varepsilon}{2}}}{\Gamma(\frac{\nu_\varepsilon}{2})} w_i^{\frac{\nu_\varepsilon}{2}-1} \exp\left\{-\frac{\nu_\varepsilon}{2} w_i\right\} \right) \times \left|\sigma_b^2 \boldsymbol{I}_K\right|^{-1/2} \exp\left\{-\frac{1}{2\sigma_b^2}\boldsymbol{b}^T\boldsymbol{b}\right\} \\ & \times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\} \frac{c_\varepsilon^{a_\varepsilon}}{\Gamma(a_\varepsilon)} (\sigma_\varepsilon^2)^{-(a_\varepsilon+1)} \exp\left(-\frac{c_\varepsilon}{\sigma_\varepsilon^2}\right) \\ & \times \frac{c_b^{a_b}}{\Gamma(a_b)} (\sigma_b^2)^{-(a_b+1)} \exp\left(-\frac{c_b}{\sigma_b^2}\right) \exp\left\{-d_\varepsilon \nu_\varepsilon\right\} I_{(2,\infty)(\nu_\varepsilon)} \end{aligned} \tag{3.9}$$

Se debe notar que la ecuación (3.9) es analíticamente intratable, razón por la cual se empleará el algoritmo de Gibbs dado en Gelman et al. (2014) para obtener simulaciones de la distribución a posteriori.

3.2. Distribuciones condicionales completas

En la sección anterior se puede observar que la distribución a posteriori es analíticamente intratable. Por lo tanto en esta sección se obtendrá las distribuciones condicionales completas de las parámetros $\boldsymbol{\beta}, b, \sigma_\varepsilon^2, \sigma_b^2, \nu_\varepsilon$ para implementar el algoritmo de Gibbs.

Para el caso de $\boldsymbol{\beta}$, en la ecuación (3.9) solamente se seleciona las expresiones que dependen del parámetro $\boldsymbol{\beta}$. Por lo tanto, la distribución condicional de este parámetro es proporcional

a:

$$f(\boldsymbol{\beta}|\boldsymbol{W}_\varepsilon, \boldsymbol{b}, \sigma^2_\varepsilon, \boldsymbol{y}) \propto \exp\left\{-\frac{1}{2\sigma^2_\varepsilon}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{b})^T\boldsymbol{W}_\varepsilon(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{b})\right\}$$
$$\times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}$$

operando la ecuación anterior, se obtiene:

$$\begin{aligned} f(\boldsymbol{\beta}|\boldsymbol{W}_\varepsilon, \boldsymbol{b}, \sigma^2_\varepsilon, \boldsymbol{y}) &\propto \exp\left\{-\frac{1}{2\sigma^2_\varepsilon}\left[(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{b})^T\boldsymbol{W}_\varepsilon(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{b})\right.\right. \\ &\quad \left.\left. + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}^T(\boldsymbol{X}^T\frac{\boldsymbol{W}_\varepsilon}{\sigma^2_\varepsilon}\boldsymbol{X}+\boldsymbol{\Sigma}_0^{-1})\boldsymbol{\beta} - 2\boldsymbol{\beta}^T(\boldsymbol{X}^T\frac{\boldsymbol{W}_\varepsilon}{\sigma^2_\varepsilon}(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{b})+\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0)\right)\right\}. \end{aligned} \tag{3.10}$$

Si se definen $\boldsymbol{\Sigma}_\beta$ y $\boldsymbol{\mu}_\beta$ de la siguiente manera:

$$\begin{aligned} \boldsymbol{\Sigma}_\beta &= \left(\frac{1}{\sigma^2_\varepsilon}\boldsymbol{X}^T\boldsymbol{W}_\varepsilon\boldsymbol{X}+\boldsymbol{\Sigma}_0^{-1}\right)^{-1} \\ \boldsymbol{\mu}_\beta &= \boldsymbol{\Sigma}_\beta\left(\frac{1}{\sigma^2_\varepsilon}\boldsymbol{X}^T\boldsymbol{W}_\varepsilon(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{b})+\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\beta}_0\right) \end{aligned} \tag{3.11}$$

la ecuación (3.10) se puede escribir como:

$$f(\boldsymbol{\beta}|\boldsymbol{W}_\varepsilon, \boldsymbol{b}, \sigma^2_\varepsilon, \boldsymbol{y}) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}^T\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta\right)\right\} \tag{3.12}$$

donde realizando operaciones para completar cuadrados, se deduce fácilmente que

$$\boldsymbol{\beta}|\boldsymbol{W}_\varepsilon, \boldsymbol{b}, \sigma^2_\varepsilon, \boldsymbol{y} \sim N_2(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta). \tag{3.13}$$

Asimismo la distribución condicional completa de $\boldsymbol{b}$ es proporcional a:

$$f(\boldsymbol{b}|\boldsymbol{\beta}, \boldsymbol{W}_\varepsilon, \sigma^2_\varepsilon, \sigma^2_b, \boldsymbol{y}) \propto \exp\left\{-\frac{1}{2\sigma^2_\varepsilon}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{b})^T\boldsymbol{W}_\varepsilon(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{b})\right\}$$
$$\times \exp\left\{-\frac{1}{2\sigma^2_b}\boldsymbol{b}^T\boldsymbol{b}\right\}$$

donde operando la ecuación anterior, se obtiene:

$$\begin{aligned} f(\boldsymbol{b}|\boldsymbol{\beta}, \boldsymbol{W}_\varepsilon, \sigma_\varepsilon^2, \sigma_b^2, \boldsymbol{y}) &\propto \exp\left\{-\frac{1}{2}\left[(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})^T\frac{\boldsymbol{W}_\varepsilon}{\sigma_\varepsilon^2}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})\right.\right. \\ &\quad \left.\left. + \boldsymbol{b}^T\frac{\boldsymbol{I}_K}{\sigma_b^2}\boldsymbol{b}\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{b}^T(\boldsymbol{Z}^T\frac{\boldsymbol{W}_\varepsilon}{\sigma_\varepsilon^2}\boldsymbol{Z}+\frac{\boldsymbol{I}_K}{\sigma_b^2})\boldsymbol{b}\right.\right. \\ &\quad \left.\left. -2\boldsymbol{b}^T\boldsymbol{Z}^T\frac{\boldsymbol{W}_\varepsilon}{\sigma_\varepsilon^2}(\boldsymbol{y}-\boldsymbol{X\beta})\right]\right\}. \end{aligned} \tag{3.14}$$

Si se definen $\boldsymbol{\Sigma}_b$ y $\boldsymbol{\mu}_b$ de la siguiente manera:

$$\begin{aligned} \boldsymbol{\Sigma}_b &= \left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{Z}^T\boldsymbol{W}_\varepsilon\boldsymbol{Z}+\frac{1}{\sigma_b^2}\boldsymbol{I}_K\right)^{-1} \\ \boldsymbol{\mu}_b &= \boldsymbol{\Sigma}_b\left(\frac{1}{\sigma_\varepsilon^2}\boldsymbol{Z}^T\boldsymbol{W}_\varepsilon(\boldsymbol{y}-\boldsymbol{X\beta})\right) \end{aligned} \tag{3.15}$$

la ecuación (3.14), se puede reescribir como

$$f(\boldsymbol{b}|\boldsymbol{\beta}, \boldsymbol{W}_\varepsilon, \sigma_\varepsilon^2, \sigma_b^2, \boldsymbol{y}) \propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{b}^T\boldsymbol{\Sigma}_b^{-1}\boldsymbol{b}-2\boldsymbol{b}^T\boldsymbol{\Sigma}_b^{-1}\boldsymbol{\mu}_b\right]\right\} \tag{3.16}$$

y completando cuadrados, se puede ver que:

$$\boldsymbol{b}|\boldsymbol{\beta}, \boldsymbol{W}_\varepsilon, \sigma_\varepsilon^2, \sigma_b^2, \boldsymbol{y} \sim N_K(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b). \tag{3.17}$$

Para w_i, se tiene que

$$f(w_i|\boldsymbol{\beta}, \boldsymbol{b}, \sigma_\varepsilon^2, \nu_\varepsilon, \boldsymbol{y}) \propto w_i^{\frac{\nu_\varepsilon+1}{2}-1}\exp\left\{-w_i\left(\frac{(y_i-\boldsymbol{X_i^T\beta}-\boldsymbol{Z_i^Tb})^2}{2\sigma_\varepsilon^2}+\frac{\nu_\varepsilon}{2}\right)\right\}$$

por lo tanto

$$w_i|\boldsymbol{\beta}, \boldsymbol{b}, \sigma_\varepsilon^2, \nu_\varepsilon, \boldsymbol{y} \sim Gamma\left(\frac{\nu_\varepsilon+1}{2}, \frac{(y_i-\boldsymbol{X_i^T\beta}-\boldsymbol{Z_i^Tb})^2}{2\sigma_\varepsilon^2}+\frac{\nu_\varepsilon}{2}\right). \tag{3.18}$$

Para σ_ε^2, se puede ver en la ecuación (3.9) que su distribución condicional completa es proporcional a:

$$\begin{aligned} f(\sigma_\varepsilon^2|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{W}_\varepsilon, \boldsymbol{y}) &\propto \left|\sigma_\varepsilon^2\boldsymbol{W}_\varepsilon^{-1}\right|^{-1/2} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})^T\boldsymbol{W}_\varepsilon(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})\right\} \\ &\times (\sigma_\varepsilon^2)^{-(a_\varepsilon+1)}\exp\left(-\frac{c_\varepsilon}{\sigma_\varepsilon^2}\right) \end{aligned}$$

luego, operando el determinante $\left|\sigma_\varepsilon^2 \boldsymbol{W}_\varepsilon^{-1}\right|^{-1/2}$ en la ecuación anterior, se obtiene:

$$f(\sigma_\varepsilon^2|\boldsymbol{\beta},\boldsymbol{b},\boldsymbol{W}_\varepsilon,\boldsymbol{y}) \propto (\sigma_\varepsilon^2)^{-(\frac{n}{2})} \exp\left\{-\frac{1}{\sigma_\varepsilon^2}\left(\frac{(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})^T\boldsymbol{W}_\varepsilon(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})}{2}+c_\varepsilon\right)\right\}$$
$$\times (\sigma_\varepsilon^2)^{-(a_\varepsilon+1)}$$

y sumando los exponentes del parámetro σ_ε^2, se obtiene

$$f(\sigma_\varepsilon^2|\boldsymbol{\beta},\boldsymbol{b},\boldsymbol{W}_\varepsilon,\boldsymbol{y}) \propto \left(\sigma_\varepsilon^2\right)^{-(\frac{n}{2}+a_\varepsilon+1)} \exp\left\{-\frac{1}{\sigma_\varepsilon^2}\left(\frac{(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})^T\boldsymbol{W}_\varepsilon(\boldsymbol{Y}-\boldsymbol{X\beta}-\boldsymbol{Zb})}{2}+c_\varepsilon\right)\right\}$$

por lo tanto

$$\sigma_\varepsilon^2|\boldsymbol{\beta},\boldsymbol{b},\boldsymbol{W}_\varepsilon,\boldsymbol{y} \sim Gamma-inv\left(\frac{n}{2}+a_\varepsilon, \frac{(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})^T\boldsymbol{W}_\varepsilon(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zb})}{2}+c_\varepsilon\right). \tag{3.19}$$

Por el mismo razonamiento que en la distribución condicional completa de σ_ε^2, se puede obtener la distribución condicional completa de σ_b^2 de la ecuación (3.9), obteniéndose que:

$$\sigma_b^2|\boldsymbol{b},\boldsymbol{W}_\varepsilon,\boldsymbol{y} \sim Gamma-inv\left(\frac{K}{2}+a_b, \frac{\boldsymbol{b}^T\boldsymbol{b}}{2}+c_b\right), \tag{3.20}$$

donde K es la cantidad de nodos de la base spline radial cúbica. De la ecuación (3.9) la distribución condicional para ν_ε es proporcional a:

$$f(\nu_\varepsilon|\boldsymbol{b},\boldsymbol{W}_\varepsilon,\boldsymbol{y}) \propto \prod_{i=1}^{n}\left(\frac{\left(\frac{\nu_\varepsilon}{2}\right)^{\frac{\nu_\varepsilon}{2}}}{\Gamma(\frac{\nu_\varepsilon}{2})} w_i^{\frac{\nu_\varepsilon}{2}-1} \exp\left\{-\nu_\varepsilon\left(\frac{w_i}{2}+\frac{d_\varepsilon}{n}\right)\right\}\right) I_{(2,\infty)(\nu_\varepsilon)} \tag{3.21}$$

3.3. Algoritmo de Gibbs

El algoritmo de Gibbs es un caso particular del algoritmo Metropolis-Hasting, el cual puede generar vía simulación valores de la distribución a posteriori $f(\boldsymbol{\theta}|y)$, esto se puede ver en Hoff (2009).

De las ecuaciones (3.10),(3.14), (3.19), (3.20) y (3.21) se tienen todas las distribuciones condicionales completas de los parámetros del modelo $\boldsymbol{\theta} = (\boldsymbol{\beta},\boldsymbol{b},\sigma_\varepsilon^2,\sigma_b^2,\nu_\varepsilon)^T$, con lo cual es posible implementar el algoritmo de Gibbs. Este algoritmo se resume a continuación.

Sea $\boldsymbol{\theta}^{(s)} = (\boldsymbol{\beta}^{(s)},\boldsymbol{b}^{(s)},(\sigma_\varepsilon^2)^{(s)},(\sigma_b^2)^{(s)},\nu_\varepsilon^{(s)})^T$ el s-ésimo valor simulado de $f(\boldsymbol{\theta}|\boldsymbol{y})$ por el algoritmo. Para obtener la siguiente simulación $\boldsymbol{\theta}^{(s+1)}$ se procede como a continuación se detalla, observándose que en todo esto se requiere de una estimación inicial
$\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)},\boldsymbol{b}^{(0)},(\sigma_\varepsilon^2)^{(0)},(\sigma_b^2)^{(0)},\nu_\varepsilon^{(0)})^T$

1. Simular $\boldsymbol{\beta}^{(s+1)}$ de $f(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{W}_\varepsilon^{(s)},\boldsymbol{b}^{(s)},(\sigma_\varepsilon^2)^{(s)},(\sigma_b^2)^{(s)})$ mediante (3.10).

2. simular $\boldsymbol{b}^{(s+1)}$ de la distribución $f(\boldsymbol{b}|\boldsymbol{\beta}^{(s+1)}, \boldsymbol{y}, \boldsymbol{W}_\varepsilon^{(s)}, (\sigma_\varepsilon^2)^{(s)}, (\sigma_b^2)^{(s)})$ mediante (3.14).
3. Simular $w_i^{(s+1)}$ de la distribución $f(w_i|\boldsymbol{\beta}^{(s+1)}, \boldsymbol{y}, \boldsymbol{b}^{(s+1)}, (\sigma_\varepsilon^2)^{(s)}, (\sigma_b^2)^{(s)}, \nu_\varepsilon^{(s)})$ mediante (3.18).
4. Simular $(\sigma_\varepsilon^2)^{(s+1)}$ de la distribución $f(\sigma_\varepsilon^2|\boldsymbol{\beta}^{(s+1)}, \boldsymbol{y}, \boldsymbol{b}^{(s+1)}, \boldsymbol{W}_\varepsilon^{(s+1)})$ mediante (3.19).
5. Simular $(\sigma_b^2)^{(s+1)}$ de la distribución $f(\sigma_b^2|\boldsymbol{y}, \boldsymbol{b}^{(s+1)}, \boldsymbol{W}_\varepsilon^{(s+1)})$ mediante (3.20).
6. Simular $\nu_\varepsilon^{(s+1)}$ de la distribución $f(\nu_\varepsilon|\boldsymbol{y}, \boldsymbol{b}, \boldsymbol{W}_\varepsilon)$ mediante (3.21).
7. Actualizar $\boldsymbol{\theta}^{(s+1)}$, y regresar al paso 1 e iterar hasta que los valores simulados se comporten de manera estacionaria, es decir, la cadena alcanza la convergencia.

Todas las distribuciones en los pasos del 1 al 5 son conocidas, el paso 6 involucra sin embargo una distribución desconocida que no es log-concava, y no se podrá utilizar el algoritmo de rechazo adaptativo (ARS) propuesto por Gilks y Wild (1992) para simular valores de esta distribución. Por lo tanto, simularemos los valores de ν_ε utilizando el algoritmo de Metropolis de rechazo adaptativo (ARMS) propuesto por Gilks y Tan (1995), que no requiere esta condición, usando la función *arms* que está implementada por Petris y Tardella (2013) en la librería HI del software R.

Después de M iteraciones del algoritmo de Gibbs se obtiene una cadena de vectores: $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(M)}$. Esta cadena, luego de alcanzar convergencia debe tener como distribución estacionaria $f(\boldsymbol{\theta}|y)$, la distribución a posteriori. Por lo tanto, se deben descartar los primeros vectores de la cadena, aquellos simulados antes de obtener convergencia. En este trabajo evaluaremos la convergencia de la cadena por inspección visual. Adicionalmente, para reducir la posible autocorrelación de los vectores generados, se tomarán saltos de q en q (por ejemplo, en la aplicación se considera $q = 100$), a fin de obtener una cadena de vectores simulados con menor autocorrelación.

3.4. Criterio de información

En el presente trabajo para la comparación entre modelos para un mismo conjunto de datos, consideraremos el criterio de información de desvio (DIC) estimado propuesto por Lunn, Jackson, Best, Thomas y Spiegelhalter (2013). Este criterio se define como

$$DIC = -2\log(f(\boldsymbol{y}|\hat{\boldsymbol{\theta}})) + 2\rho_D \tag{3.22}$$

donde $f(\boldsymbol{y}|\boldsymbol{\theta})$ es la función de verosimilitud aumentada del modelo, $\hat{\boldsymbol{\theta}} = E[\boldsymbol{\theta}|Y]$ es la media a posteriori; y ρ_D es el número efectivo de parámetros, que es definido por

$$\rho_D = \log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) - E\left(\log f(\boldsymbol{y}|\boldsymbol{\theta})\right).$$

Si consideramos que $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \cdots, \boldsymbol{\theta}^{(M)}$ es una muestra de la distribución a posteriori $f(\boldsymbol{\theta}|\boldsymbol{y})$, el DIC puede ser estimado como

$$\widehat{DIC} = -2\log(f(\boldsymbol{y}|\bar{\boldsymbol{\theta}})) + 2\widehat{\rho_D} \tag{3.23}$$

donde

$$\begin{aligned} \bar{\boldsymbol{\theta}} &= \frac{1}{M}\sum_{j=1}^{M}\boldsymbol{\theta}^{(j)} \\ \widehat{\rho_D} &= \log(f(\boldsymbol{y}|\bar{\boldsymbol{\theta}})) - \frac{1}{M}\sum_{j=1}^{M}\log(f(\boldsymbol{y}|\boldsymbol{\theta}^{(j)}) \end{aligned} \tag{3.24}$$

En nuestro trabajo la función de verosimilitud aumentada $f(\boldsymbol{y}|\boldsymbol{\theta})$ es dada en (3.5) denotada por $f(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{W})$ y la distribución a posteriori $f(\boldsymbol{\theta}|\boldsymbol{y})$, es dada en (3.6).

Capítulo 4

Estudio de simulacion

4.1. Introduccion

En el presente capítulo, se realiza un estudio de simulación para comparar el modelo de regresión semiparámetrico con splines cúbico cuando se asume que los errores presentan una distribución t-Student y cuando se asume que tienen distribución normal ante la presencia de valores atípicos. La estimación desde la perspectiva bayesiana se realizará en el programa **R**, implementando un código propio con el algoritmo de Gibss descrito en la sección 3.3. En ambos modelos, se evaluará como medida de bondad de ajuste al error cuadrático medio y el criterio de información de desvio como medida de comparación del modelo.

4.2. Descripción del estudio

Se simularan los valores de una variable respuesta mediante el modelo

$$y_i = g(x_i) + \varepsilon_i,\ i = 1, 2, \ldots, n,$$

donde la covariable x_i se generará de una distribución uniforme entre 0 y 15, y la función g viene dada por

$$g(x) = x^2\ \sin(x) - 5\ \exp(-x^2) + 120,\ \text{ donde } x \in [0, 15] \tag{4.1}$$

y $\varepsilon_i \sim N(0, 16)$. La gráfica de la función g se ilustra en la Figura 4.1

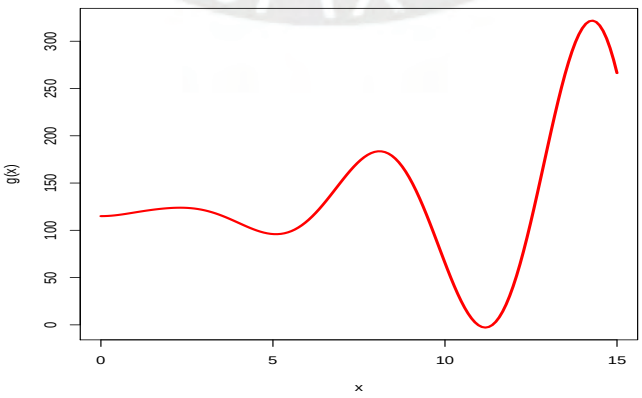


Figura 4.1: Curva que se empleará en el estudio de simulación

Según lo revisado en la ecuación (3.7), para la implementación del modelo bayesiano se considerarán las siguientes distribuciones a priori:

$$\begin{aligned} \beta_0, \beta_1 &\sim N(0, 10^2) \\ \sigma_b^2 &\sim Gamma - inv(0.01, 0.1) \\ \sigma_\varepsilon^2 &\sim Gamma - inv(0.01, 0.1) \\ \nu_\varepsilon &\sim Exp(0.1) I_{(2,\infty)}(\nu_\varepsilon) \end{aligned}$$

Luego, se introducirán valores atípicos reemplazando r valores simulados y_i por

$$y_i^* = y_i + c,$$

donde c se elegirá del siguiente conjunto $\{400, 450, 550, 600\}$ para distintos escenarios. En el presente estudio se considerará la introducción de r valores atípicos con $r = 0, 2, 4$.

Por lo tanto se considerará 3 escenarios (sin valores atípicos, con dos valores atípicos y con cuatro valores atípicos) para el modelo de regresión spline penalizado. En cada escenario se estimará el modelo dado en la ecuación (3.3) para las cuatro cantidades de nodos (K = 5, 10, 15 y 20) bajo errores normales y t-Student.

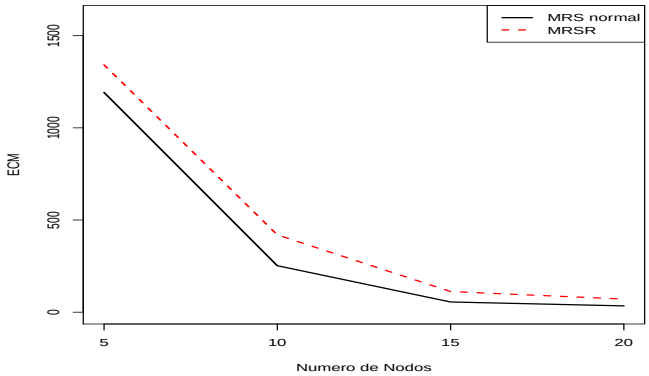
Como medida de comparación de modelos se considerará el criterio de información de desvio (DIC). Como medida de bondad de ajuste se considerará al error cuadrático medio de la curva estimada a la curva original, la expresión es dada por:

$$E\hat{C}M = \frac{1}{n} \sum_{i=1}^{n} (\hat{m}(x_i) - g(x_i))^2$$

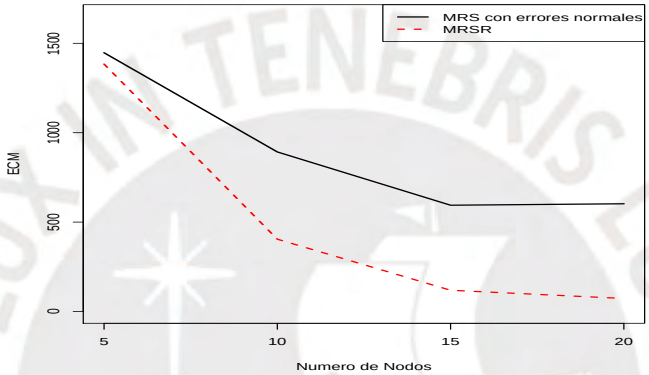
donde $\hat{m}(x_i)$ es la curva estimada por el modelo de regresión semiparamétrico evaluada en las observaciones x_i y n es el tamaño de muestra.

4.3. Resultados

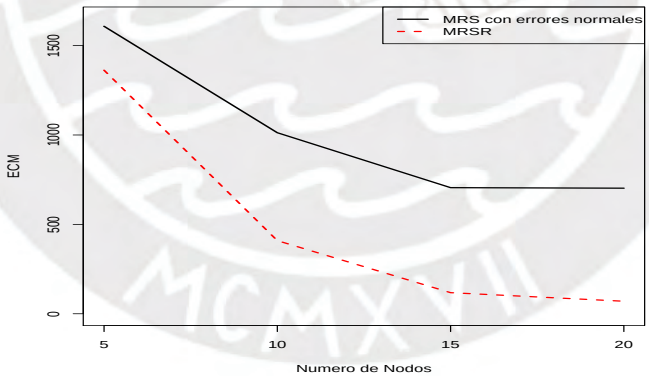
La Figura 4.2 presentan el error cuadratico medio, mientras que la Figura 4.3 presenta el criterio de información de desvio (DIC) mediante un gráfico de lineas para el modelo de regresión semiparamétrico con errores normales y errores t-Student, para los conjuntos de datos sin atípicos, con dos atípicos y con cuatro atípicos. De estas figuras se puede observar lo siguiente:



(a) Primer escenario: sin valores atípicos



(b) Segundo escenario: con 2 valores atípicos

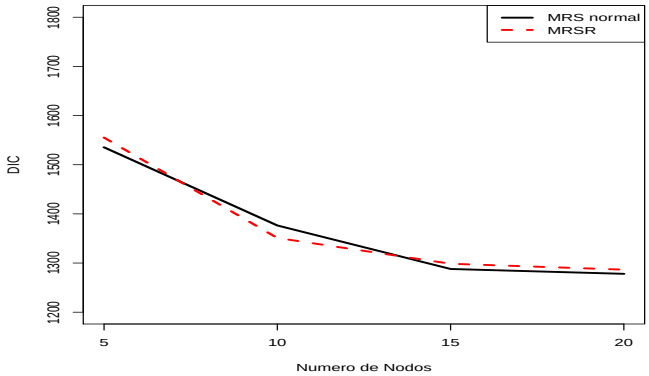


(c) Tercer escenario: con 4 valores atípicos

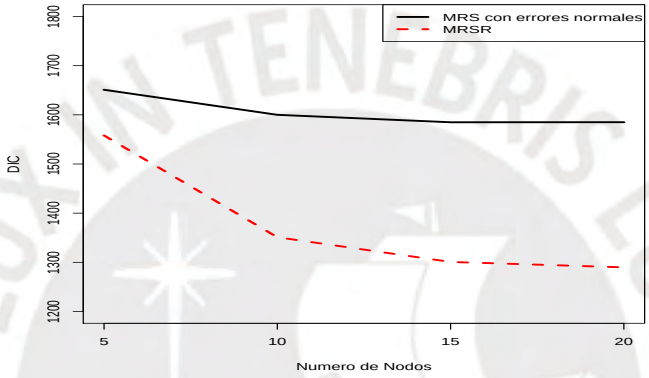
Figura 4.2: Error cuadratico medio (ECM) aplicado a los conjuntos de datos simulados en este capítulo para los tres escenarios, en cada subgráfico se muestra el ECM sobre el número de nodos utilizados en cada modelo de regresión semiparamétrico. La línea sólida de color negro corresponde al modelo de regresión semiparamétrico con errores normales y la línea punteada de color rojo corresponde al modelo de regresión semiparamétrico con errores t-Student

Respecto a la cantidad de nodos, en general, al considerar una mayor cantidad de nodos el error cuadrático medio disminuye. En la mayoría de casos se presenta una mejora significativa en el error cuadrático medio al considerar 20 nodos, pero no disminuye mucho respecto de 15 nodos.

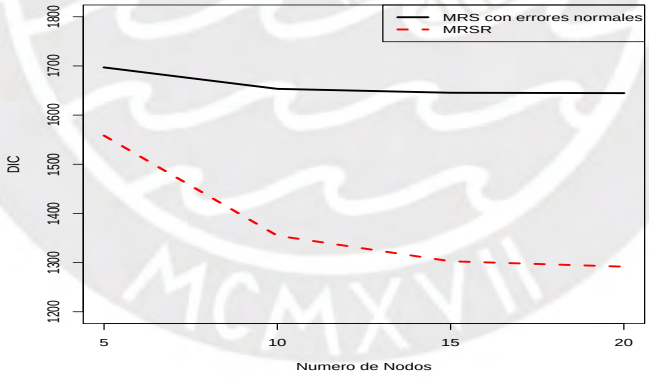
Respecto al modelo, en el caso sin valores atípicos, el modelo con errores normales y el modelo con errores t-Student presentan errores cuadráticos medios similares, mientras, que el modelo con errores t-Student con dos y cuatro atípicos presentan un menor error cuadrático medio que el modelo con errores normales.



(a) Primer escenario: sin valores atípicos



(b) Segundo escenario: con 2 valores atípicos



(c) Tercer escenario: con 4 valores atípicos

Figura 4.3: Criterio de información de desvio (DIC) aplicado a la base de datos simulados en este capítulo para los tres escenarios, en cada subgráfico se muestra el DIC sobre el número de nodos utilizados en cada modelo de regresión semiparamétrico. La línea sólida de color negro corresponde al modelo de regresión semiparamétrico con errores normales y la línea punteada de color rojo corresponde al modelo de regresión semiparamétrico con errores t-Student

		Sin valores atípicos		Con 2 valores atípicos		Con 4 valores atípicos	
	Nodos	ρ_p	DIC	ρ_p	DIC	ρ_p	DIC
MRS con errores normales	5	7.2	1535.4	5.9	1650.9	5.1	1696.9
	10	12.1	1376.4	8.9	1600.0	8.4	1653.3
	15	14.7	1287.9	10.8	1584.8	10.1	1645.4
	20	17.5	1278.1	11.7	1585.0	10.6	1644.7
MRSR	5	7.5	1555.1	7.2	1558.0	7.2	1558.3
	10	11.7	1350.9	11.7	1351.0	12.6	1354.7
	15	15.2	1298.6	14.8	1300.7	14.9	1302.7
	20	18.0	1286.5	17.8	1289.7	17.9	1291.7

Cuadro 4.1: Valores de criterio de información de desvio (DIC) y el número efectivo de parámetros (ρ_D) para los tres escenarios de los MRS con errores normales y MRSR para el conjunto de datos simulados con cuatro valores atípicos para las cantidades $K = 5, 10, 15$ y 20 nodos

Respecto a la cantidad de nodos, en general, al considerar una mayor cantidad de nodos el criterio de información de desvio disminuye, excepto para el caso que no presenta valores atípicos. Para los escenarios con 2 y 4 valores atípicos, se presenta una mejora significativa en el criterio de información de desvio al considerar 20 nodos, pero no disminuye mucho respecto de 15 nodos para ambos modelos.

Respecto al modelo, en general, el modelo con errores t-Student presentan un menor criterio de información de desvio que el modelo con errores normales en los escenarios con valores atípicos esto se muestra en cuadro 4.1. Mientras que en el escenario sin valores atípicos el modelo de regresión semiparamétrico con errores normales presenta menores valores del DIC para las cuatros cantidades de nodos.
Por otro lado, si bien se han revisado los resultados del ECM y el DIC para cada escenario, es necesario también ver de manera gráfica el ajuste del modelo a la curva simulada. Por este motivo, las Figuras 4.5, 4.7 y 4.9 presentan dicho ajuste para el modelo con errores normales y el modelo con errores t-Student. De estos gráficos se puede observar lo siguiente: en el modelo de regresión semiparamétrico sin valores atípicos de la Figura 4.5, el modelo con errores normales y el modelo con errores t-Student presentan curvas similares. En el modelo de regresión semiparamétrico con valores atípicos, en las Figuras 4.7 y 4.9, el modelo con errores normales y el modelo de regresión semiparamétrico con errores t-Student presentan curvas similares para el caso de 5 nodos. Mientras, que el modelo de regresión semiparamétrico con errores normales se ve afectado por los valores atípicos para los otros casos. En la Figura 4.5, 4.7 y 4.9, el modelo de regresión semiparamétrico con errores t-student recupera la curva que se empleó en el estudio de simulación (Figura 4.1) para el caso de 15 y 20 nodos.

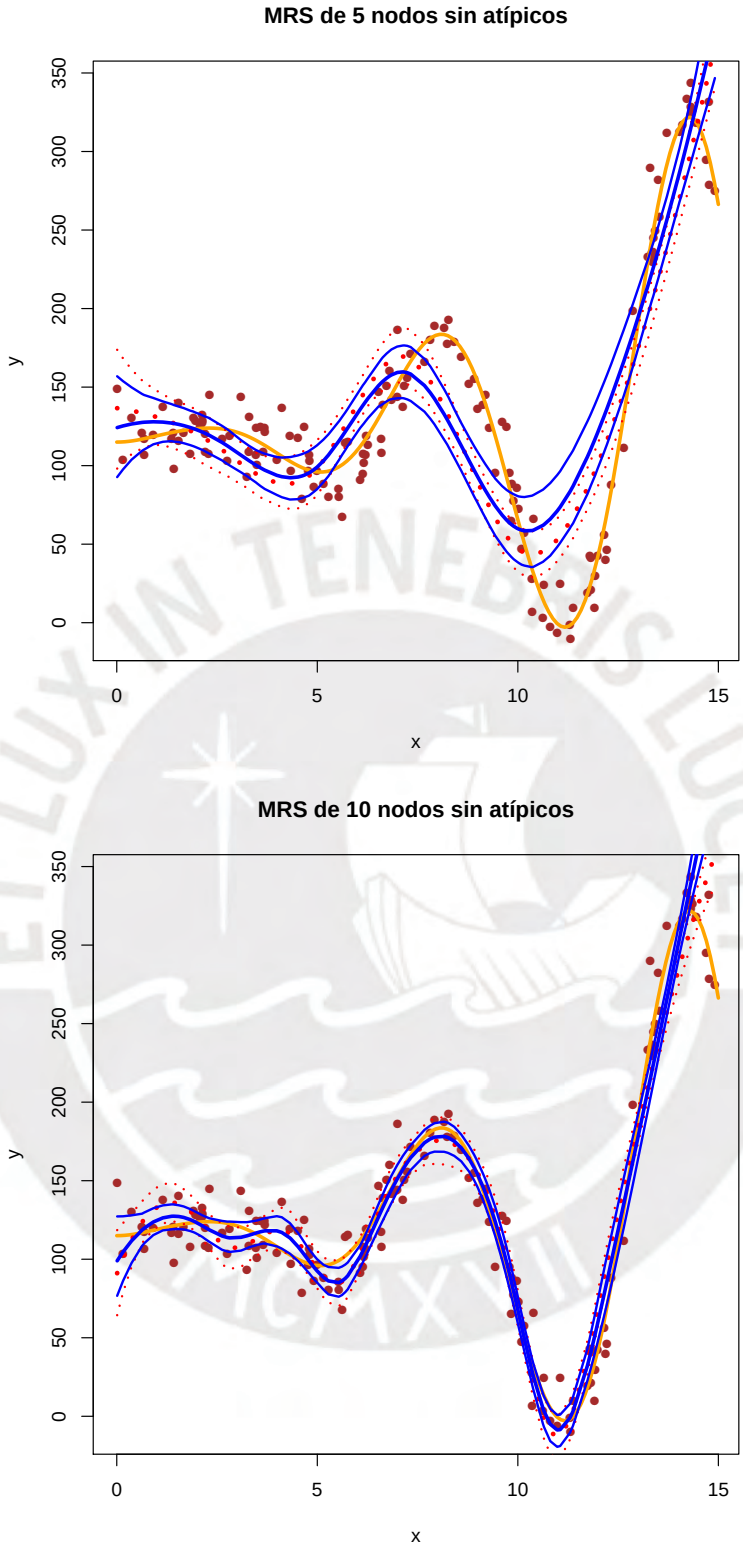


Figura 4.4: Ajuste de los modelos de regresión semiparamétrico aplicado a la base de datos sin valores atípicos simulados en este capítulo. Cada subgráfico ajusta la variable respuesta (Y) sobre la covariable(X) para el número de nodos de $K = 5$ y 10. La línea sólida de color azul corresponde al modelo de regresión semiparamétrico con errores t-student y su intervalo de credibilidad $(\pm p_{97.5}, p_{2.5})$; La línea sólida de color amarillo corresponde a la curva original; y la línea punteada de color roja corresponde al modelo de regresión semiparamétrico con errores normales y su intervalo de credibilidad $(\pm p_{97.5}, p_{2.5})$

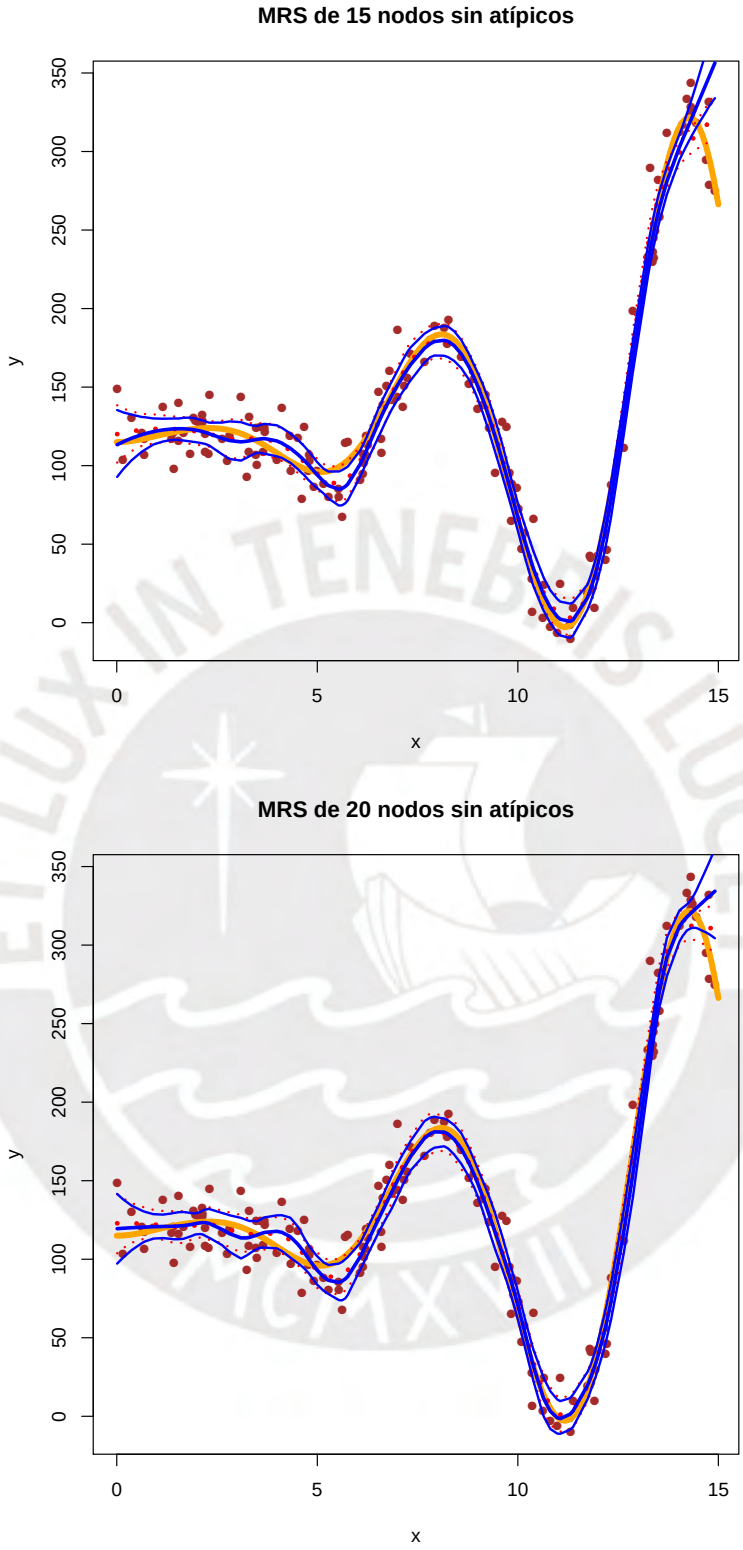


Figura 4.5: Ajuste de los modelos de regresión semiparamétrico aplicado a la base de datos sin valores atípicos simulados en este capítulo. Cada subgráfico ajusta la variable respuesta (Y) sobre la covariable (X) para el número de nodos de $K = 15$ y 20. La línea sólida de color azul corresponde al modelo de regresión semiparamétrico con errores t-Student y su intervalo de credibilidad $(\pm p_{97.5}, p_{2.5})$; La línea sólida de color amarillo corresponde a la curva original; y la línea punteada de color roja corresponde al modelo de regresión semiparamétrico con errores normales y su intervalo de credibilidad $(\pm p_{97.5}, p_{2.5})$

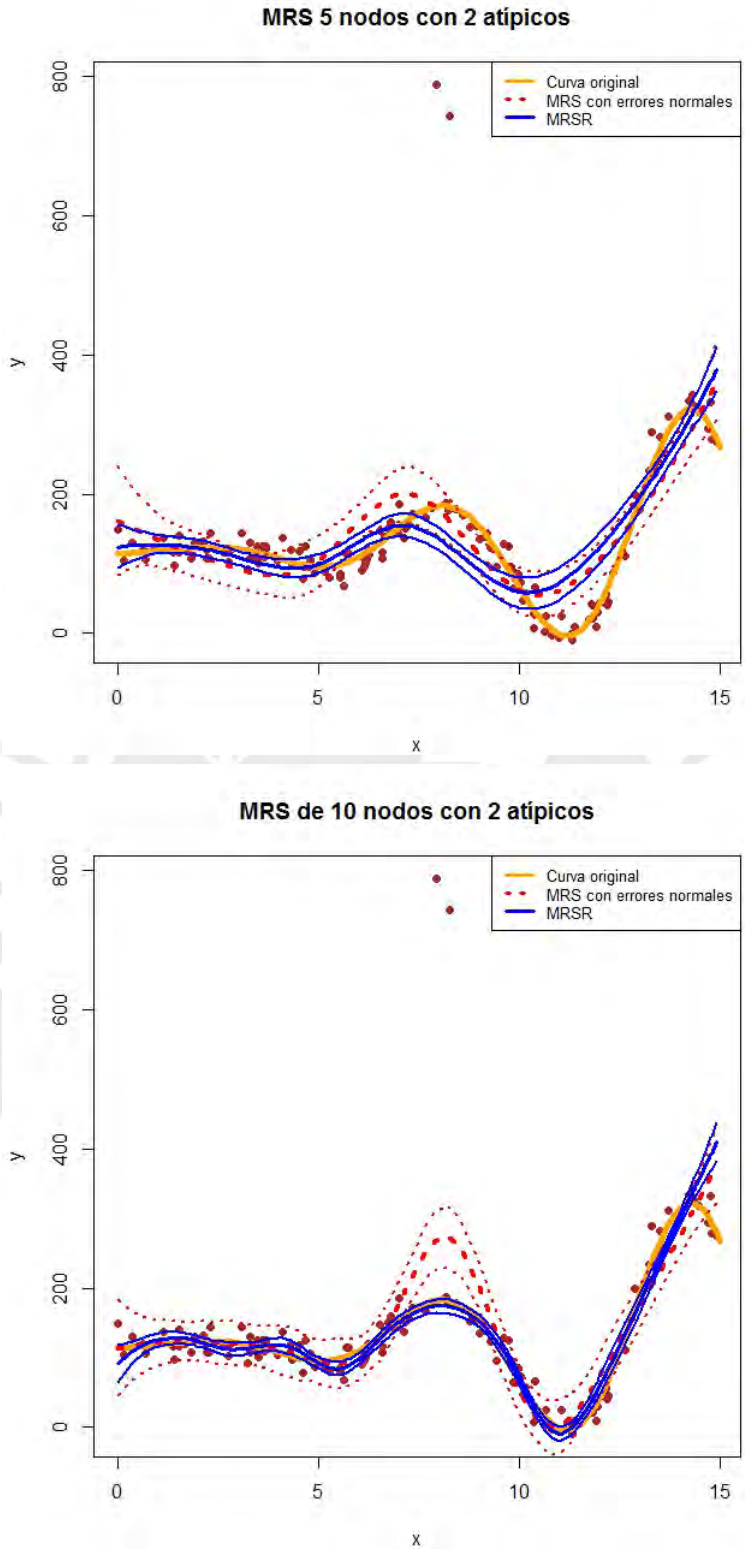


Figura 4.6: Ajuste de los modelos de regresión semiparamétrico aplicado a la base de datos con dos valores atípicos simulados en este capítulo. Cada subgráfico ajusta la variable respuesta (Y) sobre la covariable (X) para el número de nodos de $K = 5$ y 10. La línea sólida de color azul corresponde al modelo de regresión semiparamétrico con errores t-Student y su intervalo de credibilidad $(\pm p_{97.5}, p_{2.5})$; La línea sólida de color amarillo corresponde a la curva original; y la línea punteada de color roja corresponde al modelo de regresión semiparamétrico con errores normales y su intervalo de credibilidad $(\pm p_{97.5}, p_{2.5})$

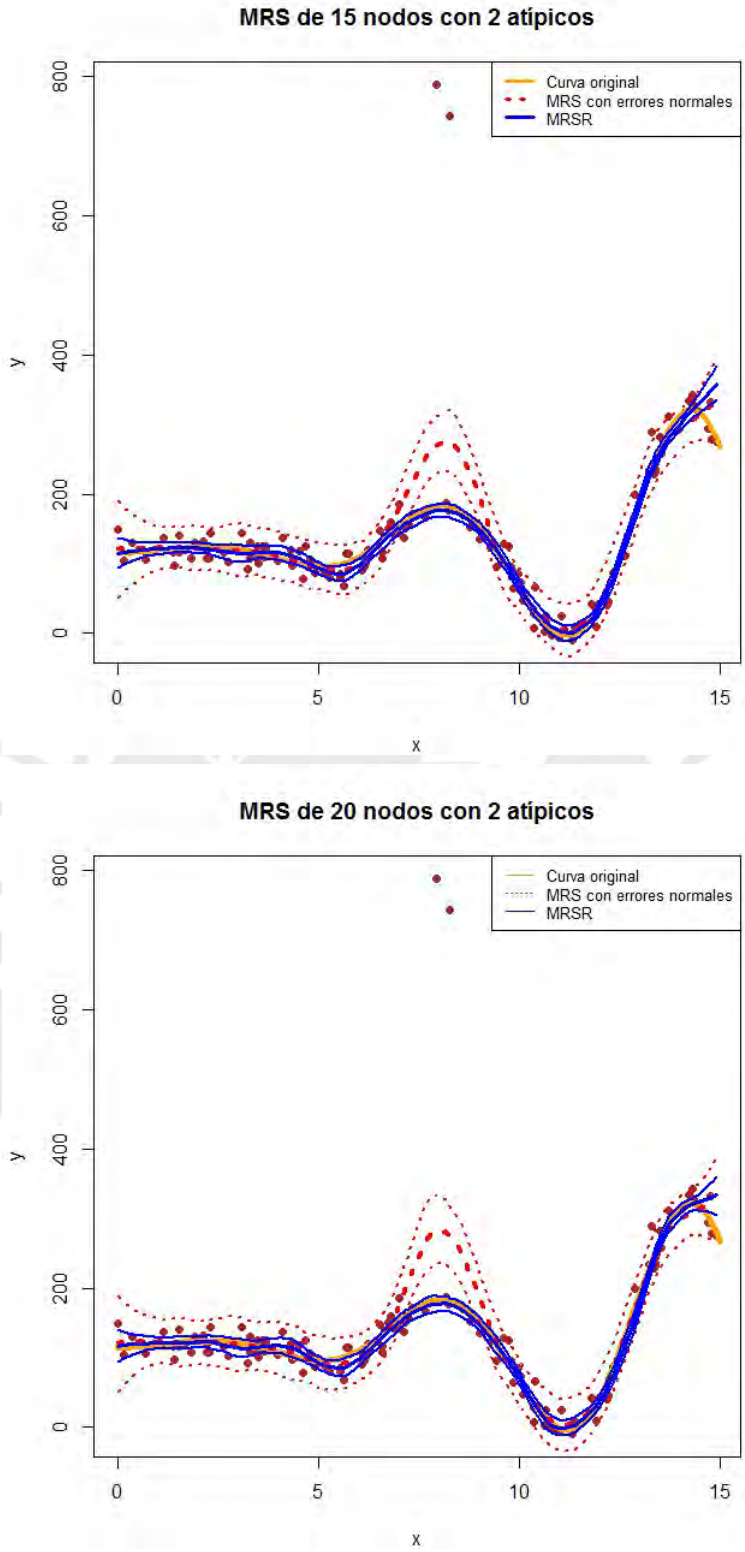


Figura 4.7: Ajuste de los modelos de regresión semiparamétrico aplicado a la base de datos con dos valores atípicos simulados en este capítulo. Cada subgráfico ajusta la variable respuesta (Y) sobre la covariable (X) para el número de nodos de $K = 15$ y 20. La línea sólida de color azul corresponde al modelo de regresión semiparamétrico con errores t-Student y su intervalo de credibilidad $(\pm p_{97.5}, p_{2.5})$; La línea sólida de color amarillo corresponde a la curva original; y la línea punteada de color roja corresponde al modelo de regresión semiparamétrico con errores normales y su intervalo de credibilidad $(\pm p_{97.5}, p_{2.5})$

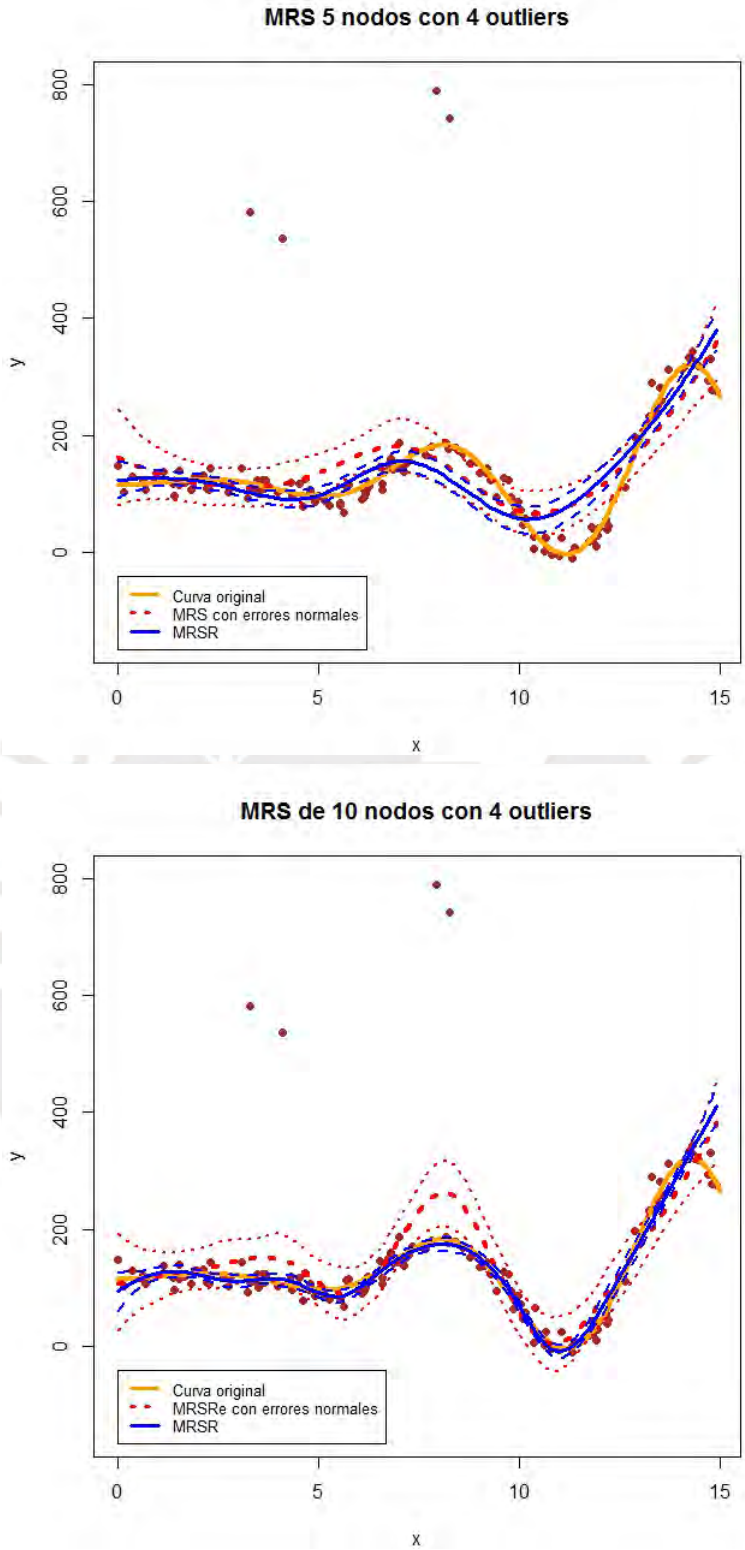


Figura 4.8: Ajuste de los modelos de regresión semiparamétrico aplicado a la base de datos con cuatro valores atípicos simulados en este capítulo. Cada subgráfico ajusta la variable respuesta (Y) sobre la covariable (X) para el número de nodos de $K = 5$ y 10. La línea sólida de color azul corresponde al modelo de regresión semiparamétrico con errores t-Student y su intervalo de credibilidad $(\pm p_{97.5}, p_{2.5})$; La línea sólida de color amarillo corresponde a la curva original; y la línea punteada de color roja corresponde al modelo de regresión semiparamétrico con errores normales y su intervalo de credibilidad $(\pm p_{97.5}, p_{2.5})$

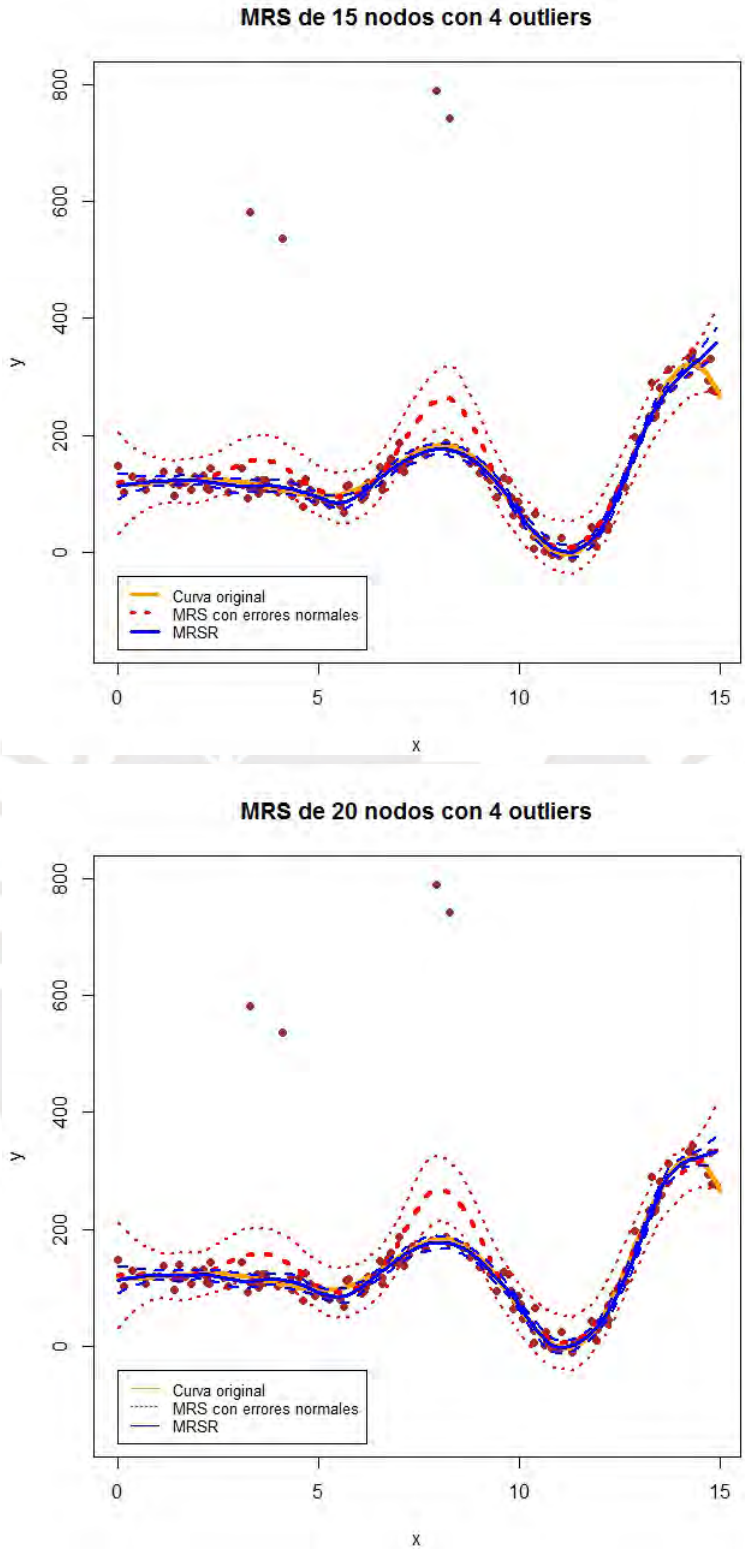


Figura 4.9: Ajuste de los modelos de regresión semiparamétrico aplicado a la base de datos con cuatro valores atípicos simulados en este capítulo. Cada subgráfico ajusta la variable respuesta (Y) sobre la covariable (X) para el número de nodos de $K = 15$ y 20. La línea sólida de color azul corresponde al modelo de regresión semiparamétrico con errores t-Student y su intervalo de credibilidad $(\pm p_{97.5}, p_{2.5})$; La línea sólida de color amarillo corresponde a la curva original; y la línea punteada de color roja corresponde al modelo de regresión semiparamétrico con errores normales y su intervalo de credibilidad $(\pm p_{97.5}, p_{2.5})$

En resumen se ha realizado una comparación entre el modelo de regresión semiparamétrico con errores normales y con errores t-Student cuando los datos presentan valores atípicos. De las Figuras 4.5, 4.7 y 4.9, se concluye que el modelo de regresión semiparamétrico con errores de distribución t-Student con 20 nodos es el menos afectado por la presencia de valores atípicos, asimismo obtuvo los menores DIC y ECM en la mayoría de los escenarios analizados. Con relación a la cantidad de nodos, se obtuvo menores DIC y ECM al considerar 15 y 20 nodos, observándose que no se presenta una mejora significativa al usar 20 en lugar de 15 nodos.

,

Capítulo 5

Aplicación

5.1. Descripción de los datos

Se considera para la aplicación un conjunto de datos previamente analizado por Staudenmayer et al. (2009). Estos datos corresponden a un experimento de laboratorio que describe las características del flujo respiratorio de un participante cuando se lo expuso a aire filtrado. Las variables de estudio en esta aplicación son el logaritmo del tiempo de exhalación ajustado que se explica por el tiempo en segundos. Los autores señalan que debido a un error en la instrumentación, una tos ocasional o una respiración esporádica del participante que se somete al experimento, se registraron algunas observaciones atípicas, como se aprecia en la Figura 5.1.

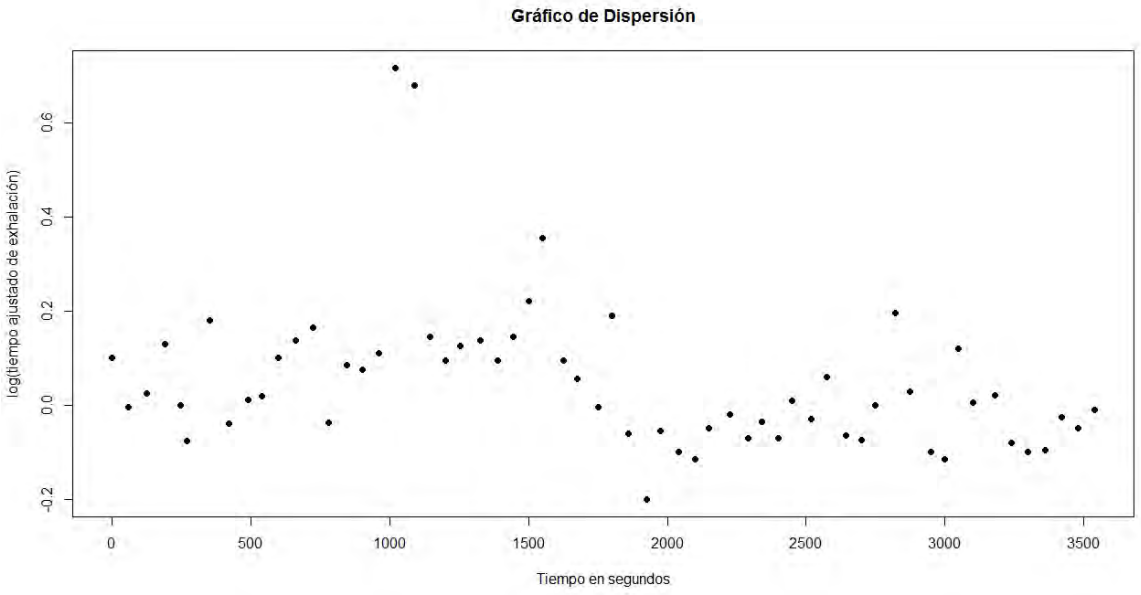


Figura 5.1: Datos recuperados de Staudenmayer et al. (2009), donde se muestra el logaritmo del tiempo de exhalación ajustado sobre el tiempo en segundos x de un participante en el experimento

5.2. Resultados

Se considera el siguiente modelo dado en (3.2)

$$y_i = m(x_i) + \varepsilon_i$$

donde y_i es el logaritmo del tiempo de exhalación ajustado y x_i es el tiempo en segundos. Se consideraran los siguientes dos modelos de regresión semiparamétrica con 20 nodos en cada caso

$$\begin{aligned} M_1: \quad \varepsilon_i &\sim N(0, \sigma_\varepsilon^2) \\ M_2: \quad \varepsilon_i &\sim t_{\nu_\varepsilon}(0, \sigma_\varepsilon^2). \end{aligned}$$

Para ambos modelos se aplicó el algoritmo de Gibbs, descrito en la sección 3.3 con el fin de estimar sus parámetros. Para cada modelo se simuló una cadena de 100000 valores de los cuales se descartaron las 10000 primeros antes de obtener convergencia. Por lo tanto, luego de que la cadena converge se obtiene los valores simulados de la distribución a posteriori. En este trabajo la convergencia fue evaluada por inspección visual de la cadena. El gráfico de las cadenas de los modelos considerados se encuentran en el apéndice B, en estos se observa que hubo convergencia después de las 10000 iteraciones descartadas. Adicionalmente, para reducir la autocorrelación se tomarán saltos de 100 en 100, resultando un tamaño de 900 valores simulados de la distribución a posteriori. El criterio de información de desvio (DIC), mostrado en el Cuadro 5.1, indica que el modelo de regresión semiparamétrico robusto con errores t-Student presenta un mejor ajuste para el conjunto de datos de respiración de Staudenmayer et al. (2009).

	DIC
Modelo con errores nomales	156.2
Modelo con errores t	125.9

Cuadro 5.1: Medidas de comparación de los MRS para el conjunto de datos de respiración de Staudenmayer et al. (2009)

En la Figura 5.2 se presenta los dos modelo estimados asi como sus intervalo de credibilidad al 95 %. Observamos que el modelo normal se ve fuertemente afectado por la presencia de valores atípicos en particular las medidas de exhalación para los tiempos (1020, 0.716) y (1090, 0.680) marcados con $*$ mientras que el modelo de regresión semiparamétrico con errores t-Student no se ve afectado por estas observaciones. Además el intervalo de credibilidad del modelo normal es más amplio que el del modelo t-Student debido a los valores atípicos.

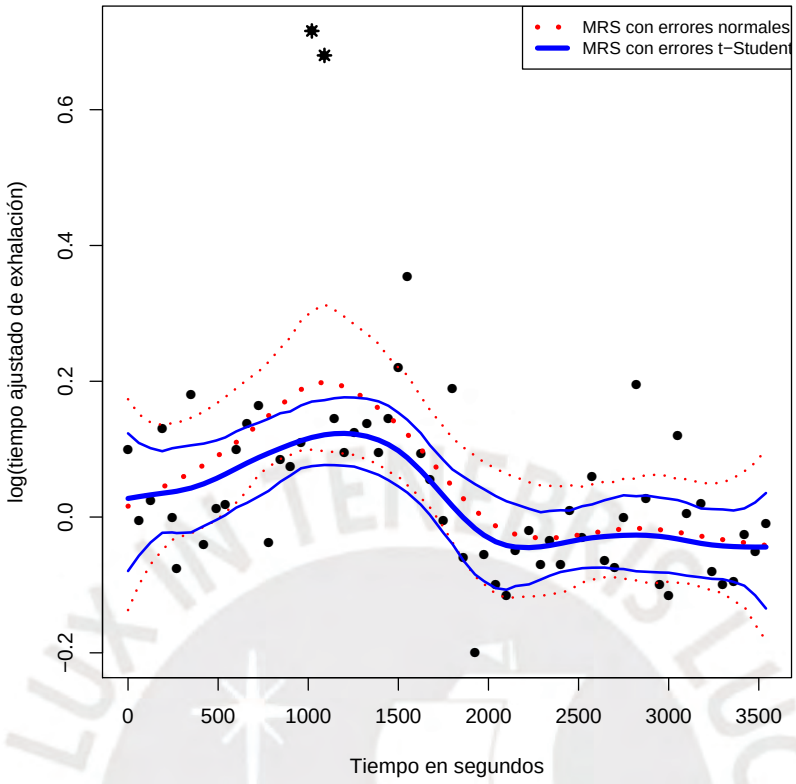


Figura 5.2: Ajuste de los modelos de regresión semiparamétrico aplicados a la base de datos de respiración de un sujeto en el experimento descrito en este capítulo. La línea sólida de color azul corresponde al modelo de regresión semiparamétrico con errores t-Student y su intervalo de credibilidad al 95 %; y la línea punteada de color roja corresponde al modelo de regresión semiparamétrico con errores normales y su intervalo de credibilidad 95 %

Capítulo 6

Conclusiones

6.1. Conclusiones

En el presente trabajo de tesis se ha propuesto un modelo de regresión semiparamétrico robusto, considerando que los errores siguen una distribución t-Student. Asimismo, se ha desarrollado un algoritmo de Gibbs para su estimación desde el punto de vista bayesiano.

A través de un estudio de simulación, se ha comparado el modelo propuesto con el usual modelo de regresión semiparamétrico con errores normales, observándose que el modelo normal puede ser fuertemente afectado por observaciones atípicas; mientras que el modelo propuesto reduce significativamente el efecto que puedan tener estas observaciones en la estimación de la media de la variable respuesta.

Además, se ha realizado una aplicación con datos previamente analizados en la literatura, en donde el modelo de regresión semiparamétrico robusto tiene un mejor ajuste que el modelo normal. En particular, se observa que disminuye el efecto de algunas observaciones atípicas y que tiene una estimación más precisa de la media de la variable respuesta, esto último se evidencia por el hecho que se obtienen intervalos de credibilidad con menor longitud que los obtenidos por el modelo normal.

6.2. Sugerencias para investigaciones futuras

- Proponer nuevos modelos de regresión semiparamétricos robustos, considerando otras distribuciones con colas pesadas para la distribución de los errores.
- Proponer métodos para la detección de valores atípicos en el modelo propuesto, por ejemplo considerando la metodología propuesta por Peng y Dey (1995).

Apéndice A

Implementación del algoritmo de Gibbs

```
library(MASS)
library(Rcpp)
library(coda)
library(MCMCpack)
library(HI)
library(faraway)
library(car)
library(reshape)
library(lubridate)
library(arm)
library(DMwR)
library(zoo)
Exhalation <- read.csv("Exhalacion.csv")
attach(Exhalation)
# gráfico de la base de datos
windows()
plot(time,exhalación,pch=19,col="black",
xlab = "Tiempo en segundos",ylab = "log(tiempo ajustado de exhalación)",
lty=2,lwd=1, main ="Gráfico de Dispersión")
locator(n = 2, type = "p")

#-----------------------------------------
# Contenedores de los valores simulador
#-----------------------------------------
n <- length(time)
Y <- as.numeric(scale(exhalación))
x <- as.numeric(scale(time))
X <- as.matrix(data.frame(intercpt=rep(1,n),x))
windows()
plot(x,Y,pch=19,col="black",
xlab = "Tiempo en segundos",ylab = "log(tiempo ajustado de exhalación)",
lty=2,lwd=1, main ="Datos estandarizados de Staudenmeyer")
```

points(x[25:26], Y[25:26], pch=4,lty=1,lwd=4,col=red")
#————————————————————————————————-
num.knots < − 20
knots¡-quantile(unique(x), seq(0,1,length=(num.knots+2))[-c(1,(num.knots+2))])
$Z_K < -(abs(outer(x, knots, " - ")))^3$
$OMEGA_all < -(abs(outer(knots, knots, " - ")))^3$
svd.OMEGA_all < − svd(OMEGA_all)# descomposición de valores singulares de una matriz
sqrt.OMEGA_all < − t(svd.OMEGA_all$v %* %(t(svd.OMEGA_all$u)*sqrt(svd.OMEGA_all$d)))
Z < − t(solve(sqrt.OMEGA_all,t(Z_K)))

#————————————————————————————————
M=100000
#——- 20 nodos————————————————————-
beta < − matrix(0,M,2)
b < − matrix(0,M,num.knots)
MWe < − matrix(0,M,n)
sigma2e < − numeric(M)
sigma2b < − numeric(M)
Ve < − numeric(M)
valores iniciales
beta[1,] < − coefficients(glm(Y X[,2], family = gaussian, start = NULL,model = TRUE,
method = "glm.fit"))
b[1,] < − rnorm(num.knots,0,1)
Ve[1] < − 5
MWe[1,] < − rgamma(n,2,2)
sigma2e[1] < − 0.012
sigma2b[1] < − 0.4
parametros de la distribucion a priori
beta0 < − c(0,0)
sigmabeta0 < − $(100^2) * diag(2)$
Ae < − 0.01
Ce < − 0.01
Ab < − 0.01
Cb < − 0.01
de < − 0.1 # parametro de la exponencial restringida exp(de)I(2,+inf)(Ve)
#————————————————————————————————
Código del modelo con errores t-Student
#————————————————————————————————
#——- 20 nodos————————————————————-
for(h in 1:M)
We < − diag(MWe[h,])
generando el nuevo Beta

sigma.beta < − solve(solve(sigmabeta0)+(1/sigma2e[h])*(t(X) %* %We %* %X))
media.beta < − sigma.beta %* %(solve(sigmabeta0) %* %beta0 + (1/sigma2e[h])*(t(X) %* %We %* %(Y-Z %* %b[h,])))
beta[h+1,] < − mvrnorm(n = 1, mu=media.beta, Sigma=sigma.beta, tol = 1e-6, empirical = FALSE, EISPACK = FALSE)
generando el nuevo b
sigma.b < − solve(solve(sigma2b[h]*diag(num.knots))+(1/sigma2e[h])*(t(Z) %* %We %* %Z))
media.b < − sigma.b %* %((1/sigma2e[h])*(t(Z) %* %We) %* %(Y-X %* %beta[h+1,]))
b[h+1,] < − mvrnorm(n = 1, mu=media.b, Sigma=sigma.b, tol = 1e-6, empirical = FALSE, EISPACK = FALSE)
# generando el nuevo We
for(i in 1:n) $MWe[h+1,i] \ < - \ rgamma(1,(Ve[h]+1)/2,(1/(2*sigma2e[h]))*(Y[i]-X[i,]\%*\%beta[h+1,]-Z[i,]\%*\%b[h+1,])^2+(Ve[h]/2))$
#————————————
generando el nuevo sigma2e
Prate < − t(Y-X %* %beta[h+1,]-Z %* %b[h+1,]) %* %diag(MWe[h+1,]) %* %(Y-X %* %beta[h+1,]-Z %* %b[h+1,])
ts < − rgamma(1, (n/2)+Ae,Prate/2 + Ce)
sigma2e[h+1] < − 1/ts
#————————————-
generando el nuevo sigma2b
tts < − rgamma(1, (num.knots/2)+Ab,(1/2)*(t(b[h+1,]) %* %b[h+1,]) + Cb)
sigma2b[h+1] < − 1/tts
#————————————————————
De aqui hasta la linea final estoy haciendo el
metroplis-hasting para Ve usando arms del package HI
#————————————————————
# Funcion a utilizar
logp < − function(alpha,We,de){
m < − alpha/(2*(1-alpha))
$sum(m*log(m)-log(gamma(m))+(m-1)*log(We)-(2*m)*(We/2+de/n),na.rm=TRUE)+log((2*m+1)^2)$
}
al < − arms(Ve[h]/(1+Ve[h]), logp, function(alpha,...) (alpha¿2/3)*(alpha¡0.99999), 1, We=MWe[h+1,], de=de)
Ve[h+1] < − al/(1-al)
}
#————————————————————————————————-
Código del modelo con errores normales
#————————————————————————————————-
#—— 20 nodos————————————-
Nbeta < − matrix(0,M,2)# 2 columas por el numero de parámetros fijos

```
Nb < − matrix(0,M,num.knots)
Nsigma2e < − numeric(M)
Nsigma2b < − numeric(M)
# valores iniciales
Nbeta[1,] < − coefficients( glm(Y X[,2], family = gaussian, start = NULL,model = TRUE,
method = "glm.fit"))
Nb[1,] < − rnorm(num.knots,0,1)
Nsigma2e[1] < − 0.012
Nsigma2b[1] < − 0.5
#—— 20 nodos————————————-
for(h in 1:M)
# generando el nuevo Beta
sigma.betaN < − solve(solve(sigmabeta0)+(1/Nsigma2e[h])*(t(X) %* %X))
media.betaN < − sigma.betaN %* %(solve(sigmabeta0) %* %beta0 + (1/Nsigma2e[h])*(t(X) %* %(Y-
Z %* %Nb[h,])))
Nbeta[h+1,] < − mvrnorm(n = 1, mu=media.betaN, Sigma=sigma.betaN, tol = 1e-6, em-
pirical = FALSE, EISPACK = FALSE)
# generando el nuevo b
sigma.bN < − solve(solve(Nsigma2b[h]*diag(num.knots))+(1/Nsigma2e[h])*(t(Z) %* %Z))
media.bN < − sigma.bN %* %((1/Nsigma2e[h])*(t(Z) %* %(Y-X %* %Nbeta[h+1,])))
Nb[h+1,] < − mvrnorm(n = 1, mu=media.bN, Sigma=sigma.bN, tol = 1e-6, empirical =
FALSE, EISPACK = FALSE)
#————————————
# generando el nuevo sigma2e
PrateN < − t(Y-X %* %Nbeta[h+1,]-Z %* %Nb[h+1,]) %* %(Y-X %* %Nbeta[h+1,]-Z %* %Nb[h+1,])
tsN < − rgamma(1, (n/2)+Ae,PrateN/2 + Ce)
Nsigma2e[h+1] < − 1/tsN
#————————————————-
# generando el nuevo sigma2b
ttsN < − rgamma(1, (num.knots/2)+Ab,(1/2)*(t(Nb[h+1,]) %* %Nb[h+1,]) + Cb)
Nsigma2b[h+1] < − 1/ttsN

#————————————————————————————
# Retiro de las primeras 10000 simulaciones de la cadena MCMC
#————————————————————————————
M.inix < − 10001
beta < − beta[M.inix:M,]
b < − b[M.inix:M,]
sigma2b < − sigma2b[M.inix:M]
sigma2e < − sigma2e[M.inix:M]
Ve < − Ve[M.inix:M]
Nbeta < − Nbeta[M.inix:M,]
```

```
Nb < − Nb[M.inix:M,]
Nsigma2b < − Nsigma2b[M.inix:M]
Nsigma2e < − Nsigma2e[M.inix:M]
M.new < − dim(beta)[1]
stin < − c(seq(1,M.new,by=100))
MM < − length(stin)
#————————————————————————————————
# Estimaciones de los Mu (MRSR)
# hallando los valores esperados de Y (E[Y]=Mu(x))
#————————————————————————————————
Mu < − matrix(0,MM,n)
for(i in 1:MM)
for(j in 1:n)
Mu[i,j] < − beta[100*(i-1)+1,1]+ X[j,2]*beta[100*(i-1)+1,2] + t(Z[j,]) %* %b[100*(i-1)+1,]


# estimaciones media, Intervalo de credibilidad
Mu_est20 < − cbind(apply(Mu,2,function(x)quantile(x,0.025)),
colMeans(Mu),apply(Mu,2,function(x)quantile(x,0.975)))
Mu_est20 < − mean(exhalación)+sqrt(var(exhalación))*Mu_est20
colnames(Mu_est20) < − c("Q1","Yi_estimados","Q3")
#————————————————————————————————
# Estimaciones de los Ys (modelo normal)
#————————————————————
Mu_N < − matrix(0,MM,n)
for(i in 1:MM)
for(j in 1:n)
Mu_N[i,j] < − Nbeta[100*(i-1)+1,1]+ X[j,2]*Nbeta[100*(i-1)+1,2] + t(Z[j,]) %* %Nb[100*(i-
1)+1,]


Mu_estN20 < − cbind(apply(Mu_N,2,function(x)quantile(x,0.025)),
colMeans(Mu_N),apply(Mu_N,2,function(x)quantile(x,0.975)))
Mu_estN20 < − mean(exhalación)+sqrt(var(exhalación))*Mu_estN20
colnames(Mu_estN20) < − c("Q1","Yi_estimados","Q3")
#————————————————————————————————-
# Gráficos
#— 20 nodos ——————————————————
windows()
par(mfrow=c(1,1),mar=c(4,4,4,2))
plot(time, exhalación, type="p", pch=16, col="black",
xlab="Tiempo en segundos", ylab="log(tiempo ajustado de
```

exhalación)",main = .ExhalaciónestandarizadosÇ")
points(time[25:26], exhalación[25:26], pch=8,lty=1,lwd=2,col="black")
# Add a line
lines(time[order(time)], Mu_estN20[order(time),1], pch=18, col=red", type="l", lty=3,lwd=2)
lines(time[order(time)], Mu_estN20[order(time),2], pch=18, col=red", type="l", lty=3,lwd=4)
lines(time[order(time)], Mu_estN20[order(time),3], pch=18, col=red", type="l", lty=3,lwd=2)
lines(time[order(time)], Mu_est20[order(time),1], pch=18, col="blue", type="l", lty=1,lwd=2)
lines(time[order(time)], Mu_est20[order(time),2], pch=18, col="blue", type="l", lty=1,lwd=4)
lines(time[order(time)], Mu_est20[order(time),3], pch=18, col="blue", type="l", lty=1,lwd=2)
# Add a legend
legend(x="topright", legend=c("MRSR con errores normales","MRSR con errores t-student"),
col=c(red", "blue"), lty=c(3,1), cex=0.8)
#————————————————————————————
Estimación de los parámetros (beta,b,sigma)
#————————————————————————————
MM es el número de elemtos despues del burn in and thin(100)
c1 <− c("beta1","beta2","sigma2e")
c2 <− c("b1","b2","b3","b4","b5","b6","b7","b8","b9","b10","b11","b12","b13","b14"
,"b15","b16","b17","b18","b19","b20","sigma2b")
resultados20 <− data.frame(cbind(beta,sigma2e,b,sigma2b,Ve))
colnames(resultados20) <− c(c1,c2,"Ve")
resultados20 <− resultados20[stin,]
resultados20_normal <− data.frame(cbind(Nbeta,Nsigma2e,Nb,Nsigma2b))
colnames(resultados20_normal) <− c(c1,c2)
resultados20_normal <− resultados20_normal[stin,]
Est_20_t <− sapply(resultados20,mean)
Est_20_N <− sapply(resultados20_normal,mean)
#————————————————————————————-
Comparación de los modelos (ECM - DIC)
#————————————————————————————-
modelo normal
#——— 20 nodos
logverN20 <− matrix(0,MM,n)
for(i in 1:MM)
for(j in 1:n)
$logverN20[i,j] <- -dnorm(Y[j], X[j,] \%* \%Nbeta[100*(i-1)+1,] + Z[j,] \%* \%Nb[100*(i-1)+1,], sqrt(Nsigma2e[100*(i-1)+1]), log = TRUE)$

SDhatN20 <− 0
for(j in 1:n)
$SDhatN20 <- -dnorm(Y[j], X[j,] \%* \%Est\_20\_N[1:2] + Z[j,] \%* \%Est\_20_N[4:23],$

$sqrt(Est\_20\_N[3]), log = TRUE) + SDhatN20$

Dhat_N20 <- -2*SDhatN20 #Dhat
DbarN20 <- mean(-2*apply(logverN20,1, sum)) #Dbar
PD_N20 <- DbarN20-Dhat_N20 #pD
DIC_20N <- Dhat_N20+2*PD_N20
# modelo t-Student
#——— 20 nodos
logverT20 <- matrix(0,MM,n)
for(i in 1:MM)
for(j in 1:n)
logverT20[i,j] <- dt((Y[j]-X[j,]%*%beta[100*(i-1)+1,]-Z[j,]%*%b[100*(i-1)+1,])/sqrt(sigma2e[100*(i-1)+1]),Ve[100*(i-1)+1],log=TRUE)-
log(sqrt(sigma2e[100*(i-1)+1]))

SDhatT20 <- 0
for(j in 1:n) $SDhatT20 < -dt((Y[j] - X[j,] \% * \% Est_20_t[1:2] - Z[j,] \% * \% Est_20_t[4:23])/sqrt(Est_20_t[3]), Est_20_t[25], log = TRUE) - log(sqrt(Est_20_t[3])) + SDhatT20$
Dhat_T20 <- -2*SDhatT20 #Dhat
DbarT20 <- mean((-2)*apply(logverT20,1, sum))#Dbar
PD_T20 <- DbarT20-Dhat_T20 #pD
DIC_20T <- Dhat_T20+2*PD_T20
#
DICs <- c(DIC_20N,DIC_20T)

Apéndice B

Gráficos de la aplicación

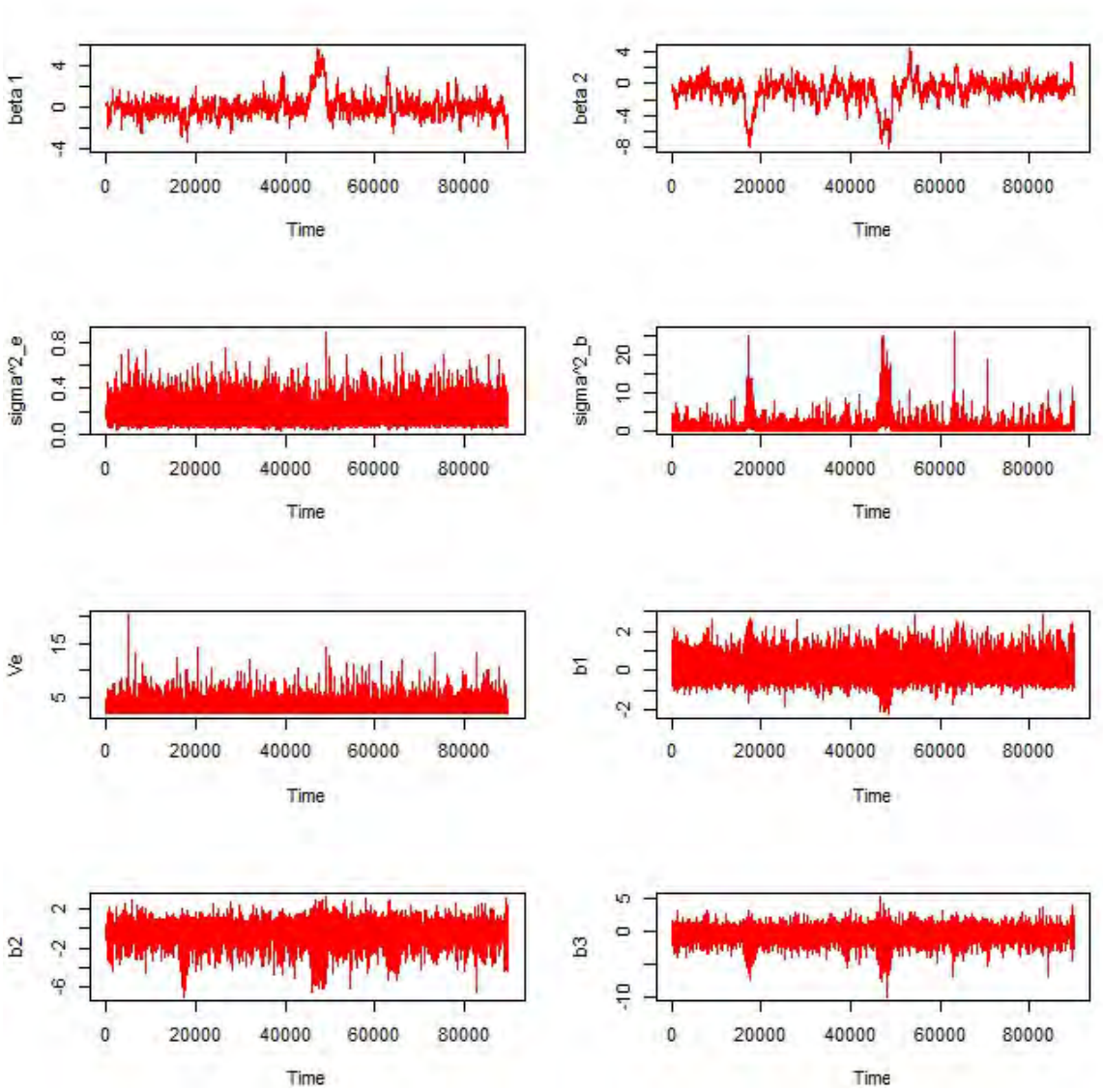


Figura B.1: Cadenas de la iteración del algoritmo de Gibbs para cada parámetro del MRSR para la base de datos de exhalación

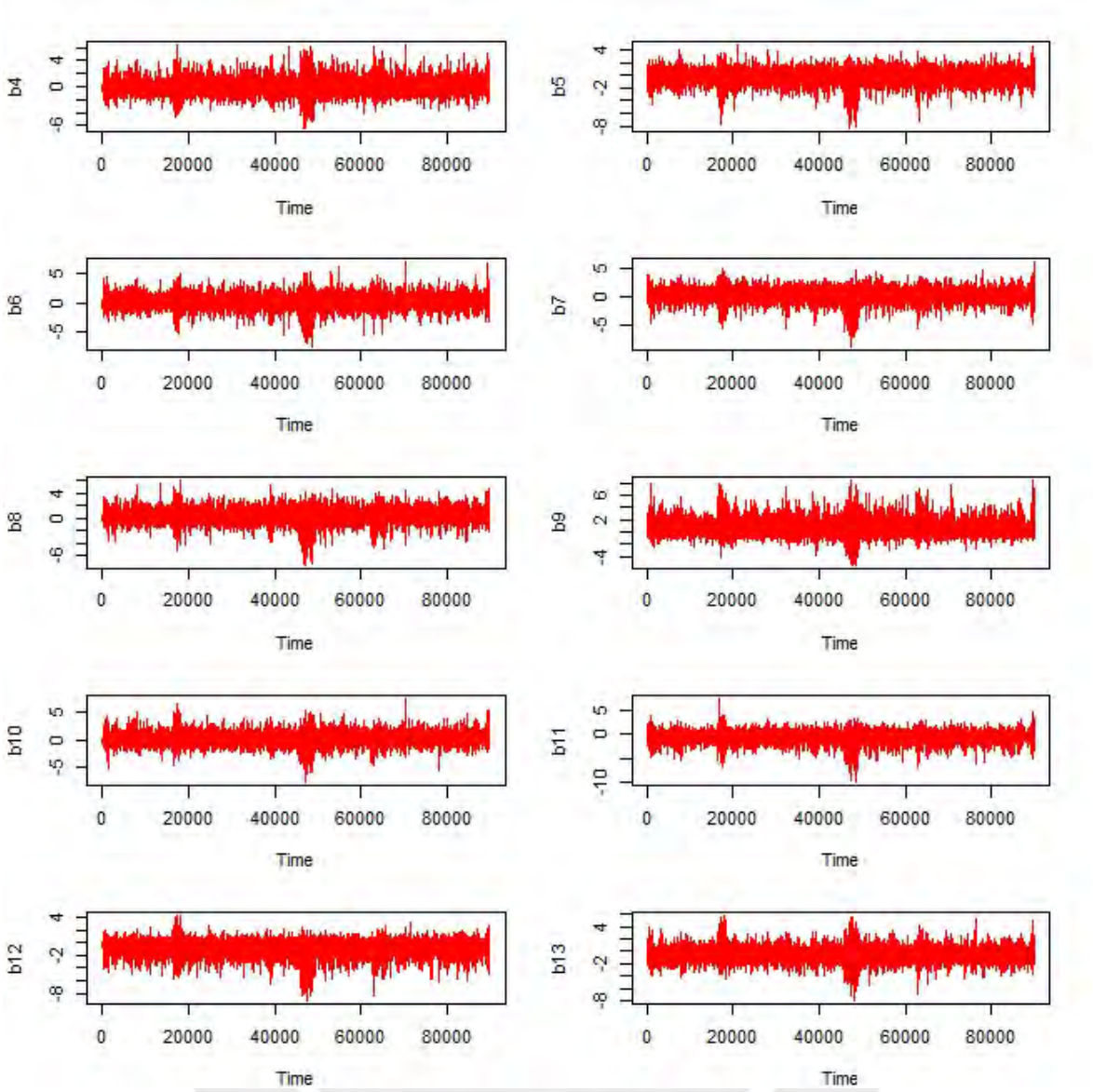


Figura B.2: Cadenas de la iteración del algoritmo de Gibbs para cada parámetro del MRSR para la base de datos de exhalación

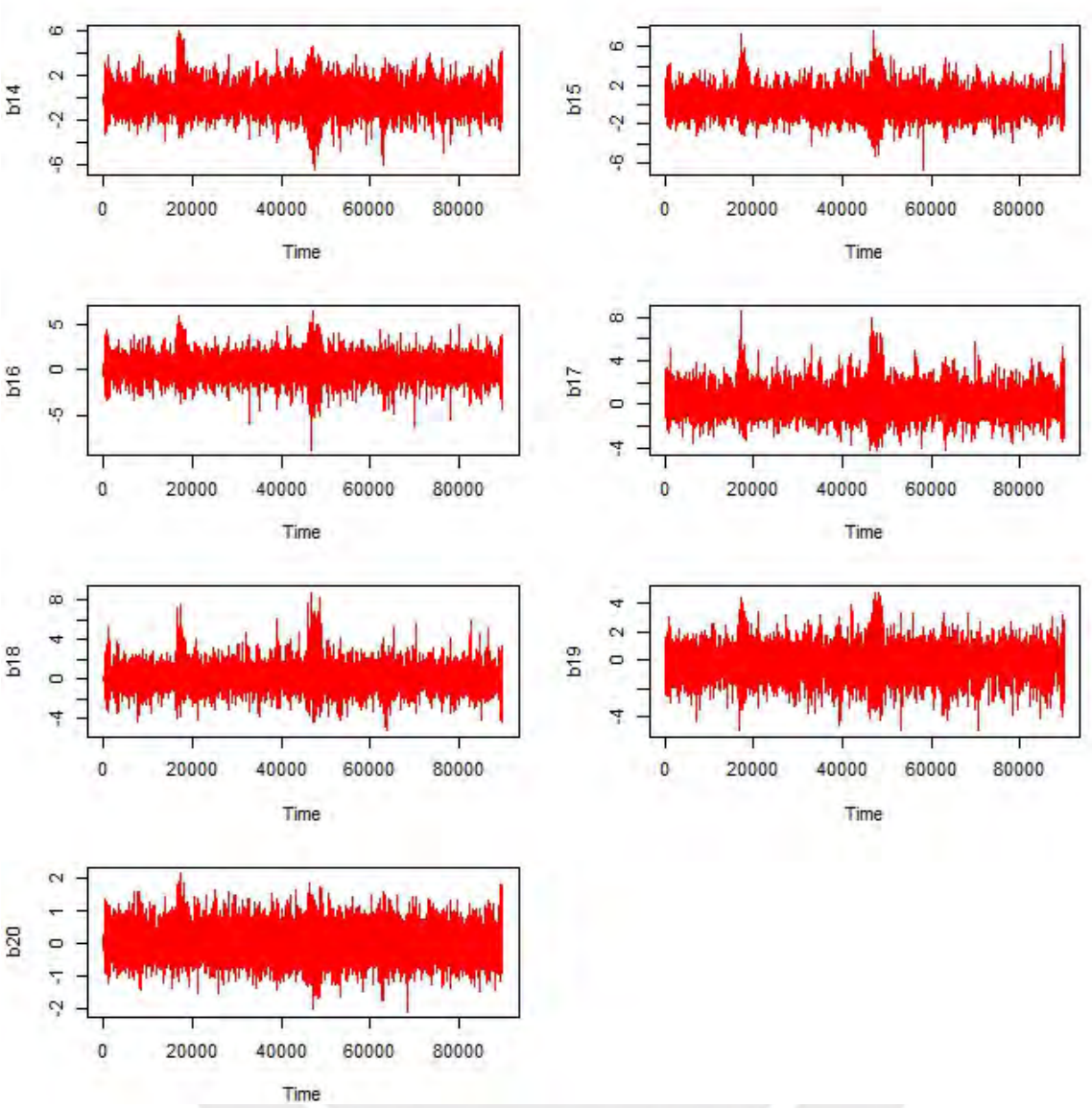


Figura B.3: Cadenas de la iteración del algoritmo de Gibbs para cada parámetro del MRSR para la base de datos de exhalación

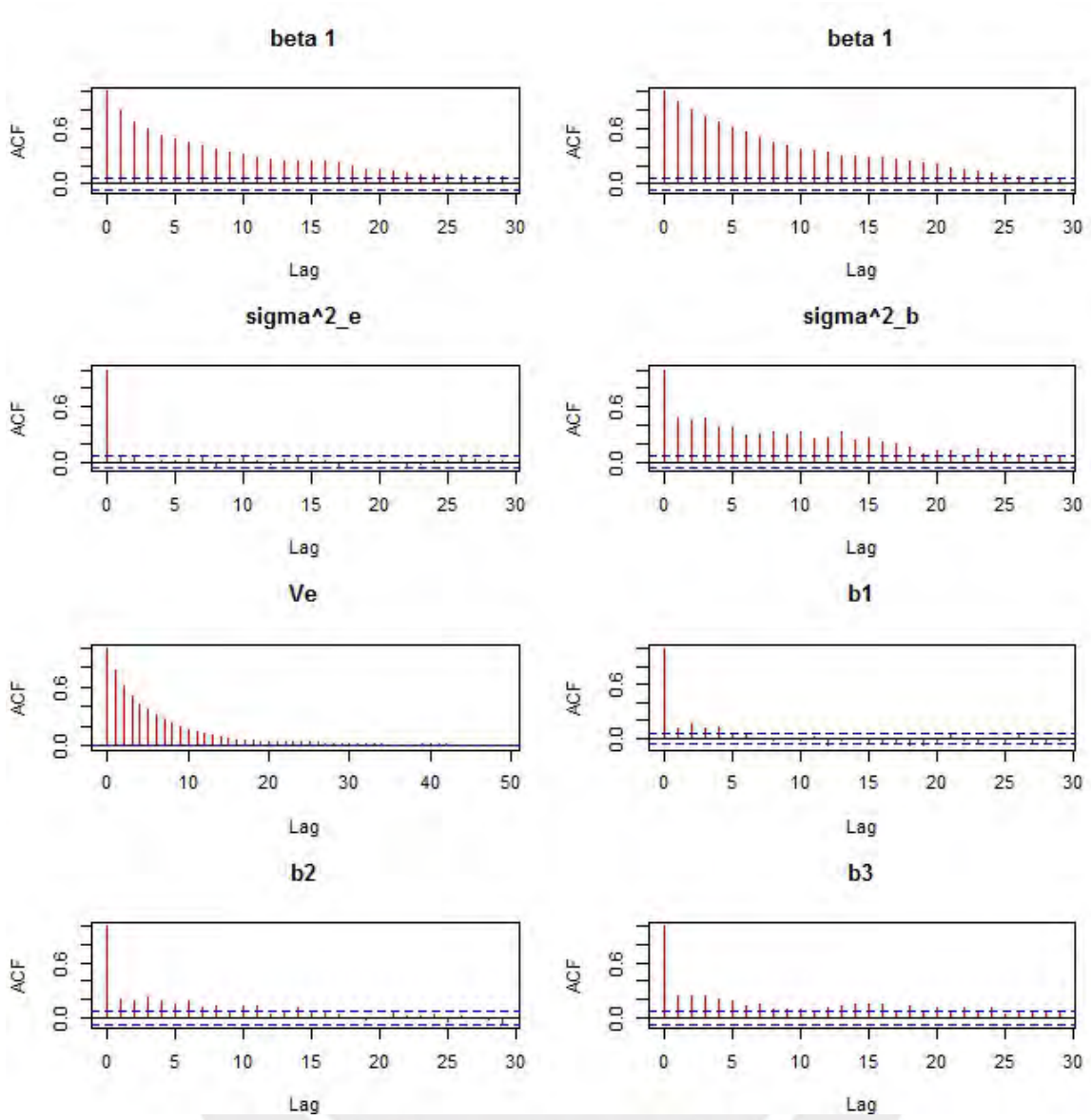


Figura B.4: Función de autocorrelación del MRSR para la base de datos exhalación

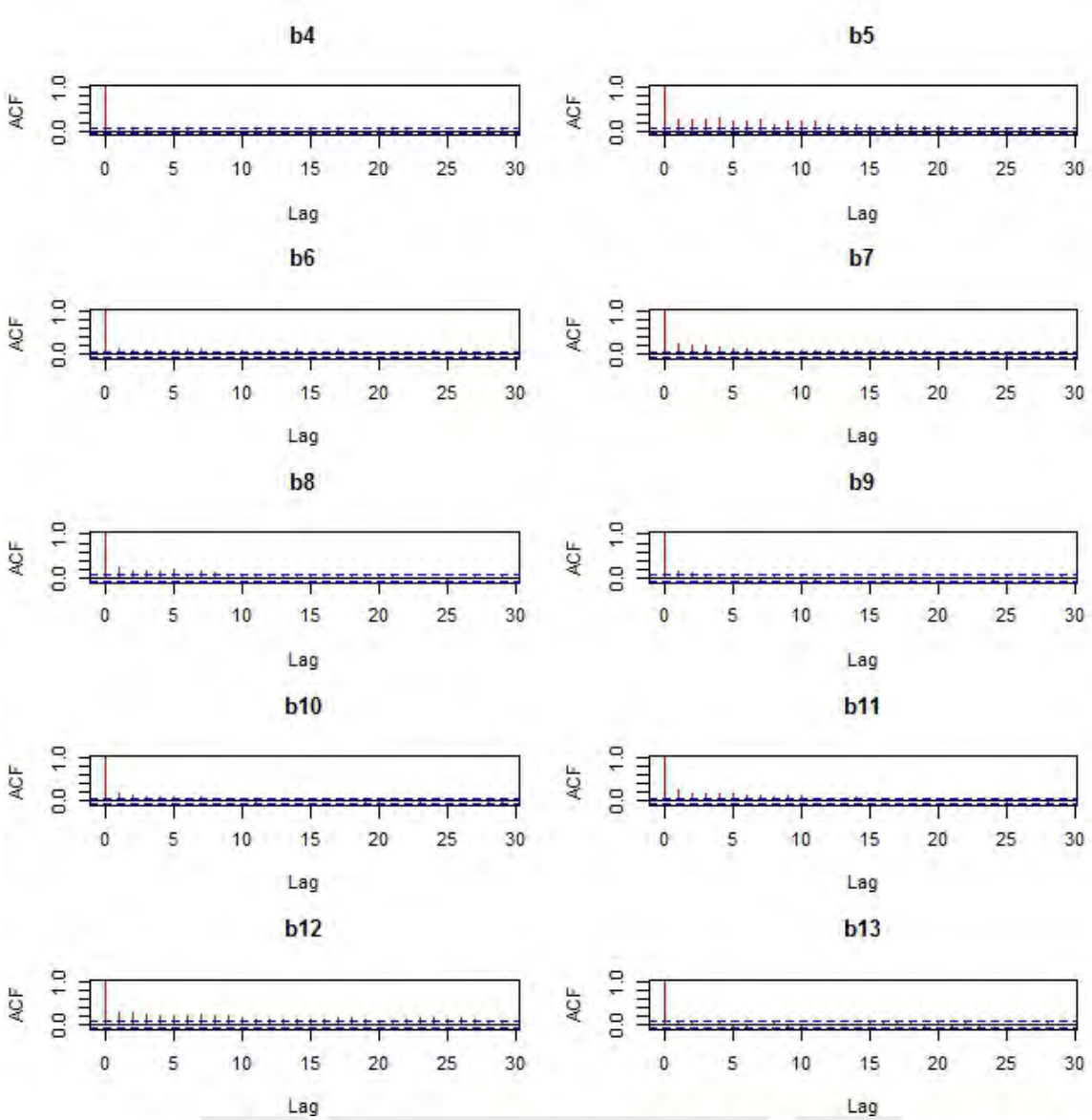


Figura B.5: Función de autocorrelación del MRSR para la base de datos exhalación

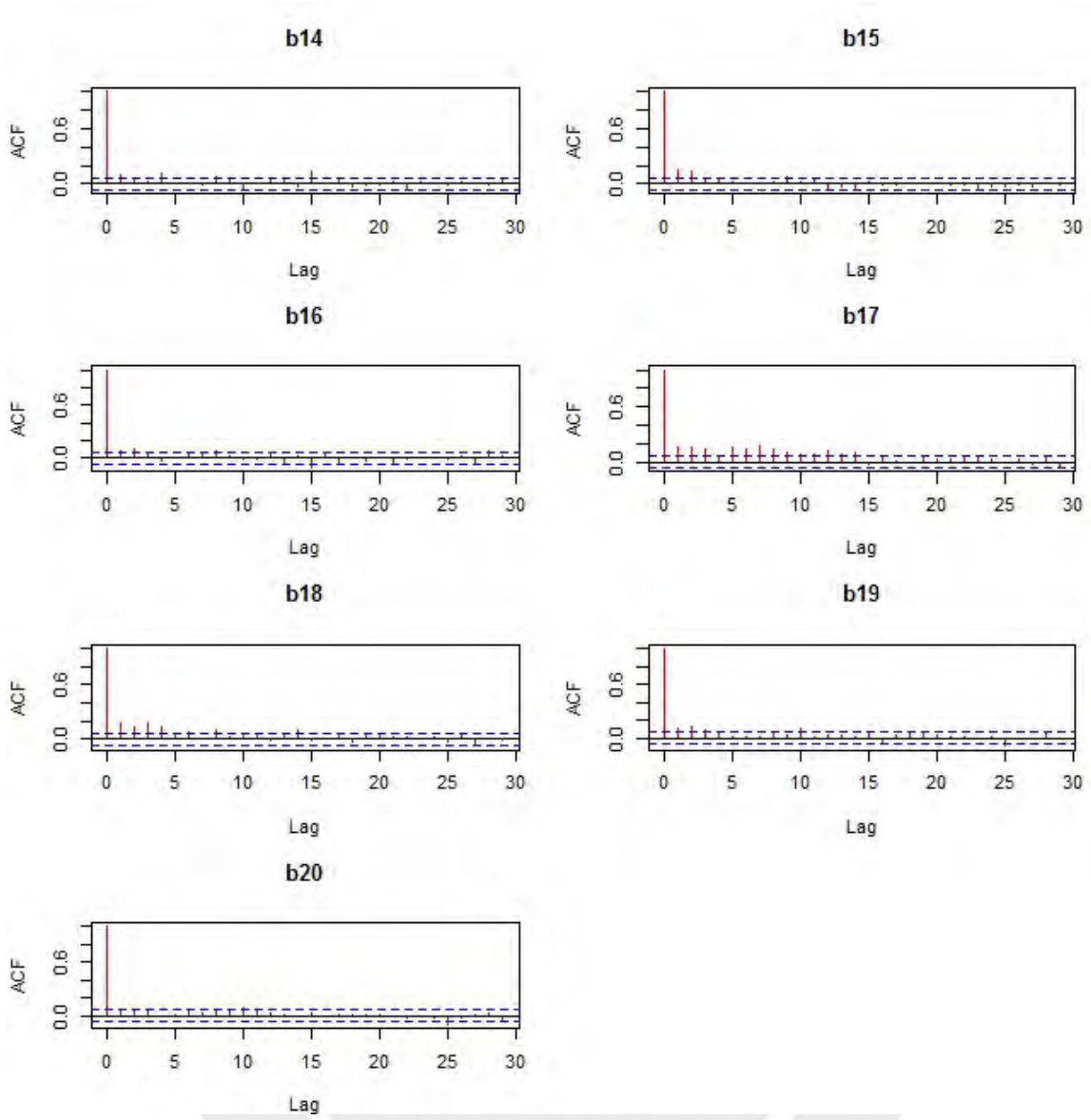


Figura B.6: Función de autocorrelación del MRSR para la base de datos exhalación

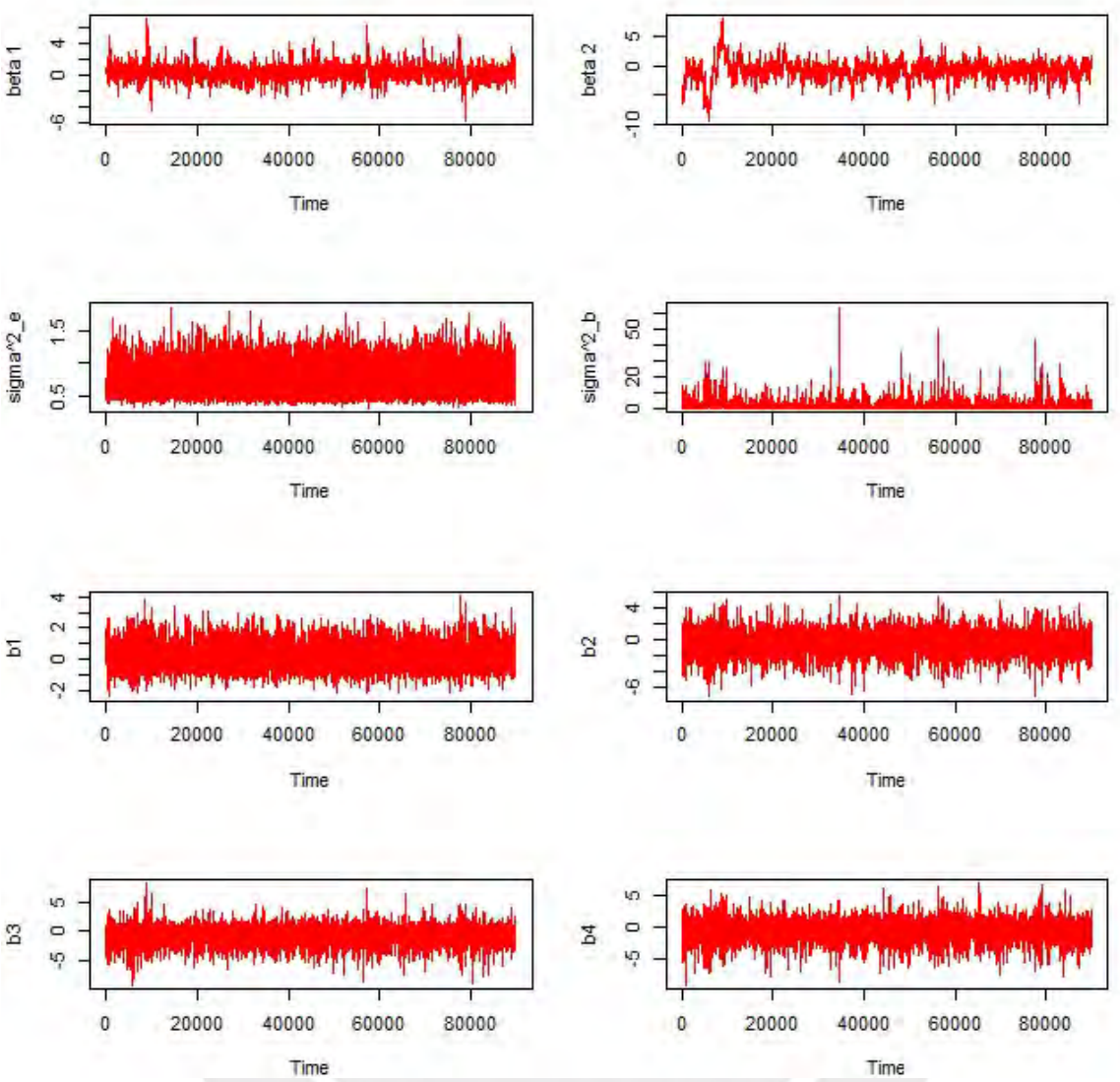


Figura B.7: Cadenas de la iteración del algoritmo de Gibbs para cada parámetro del MRS normal para la base de datos de exhalación

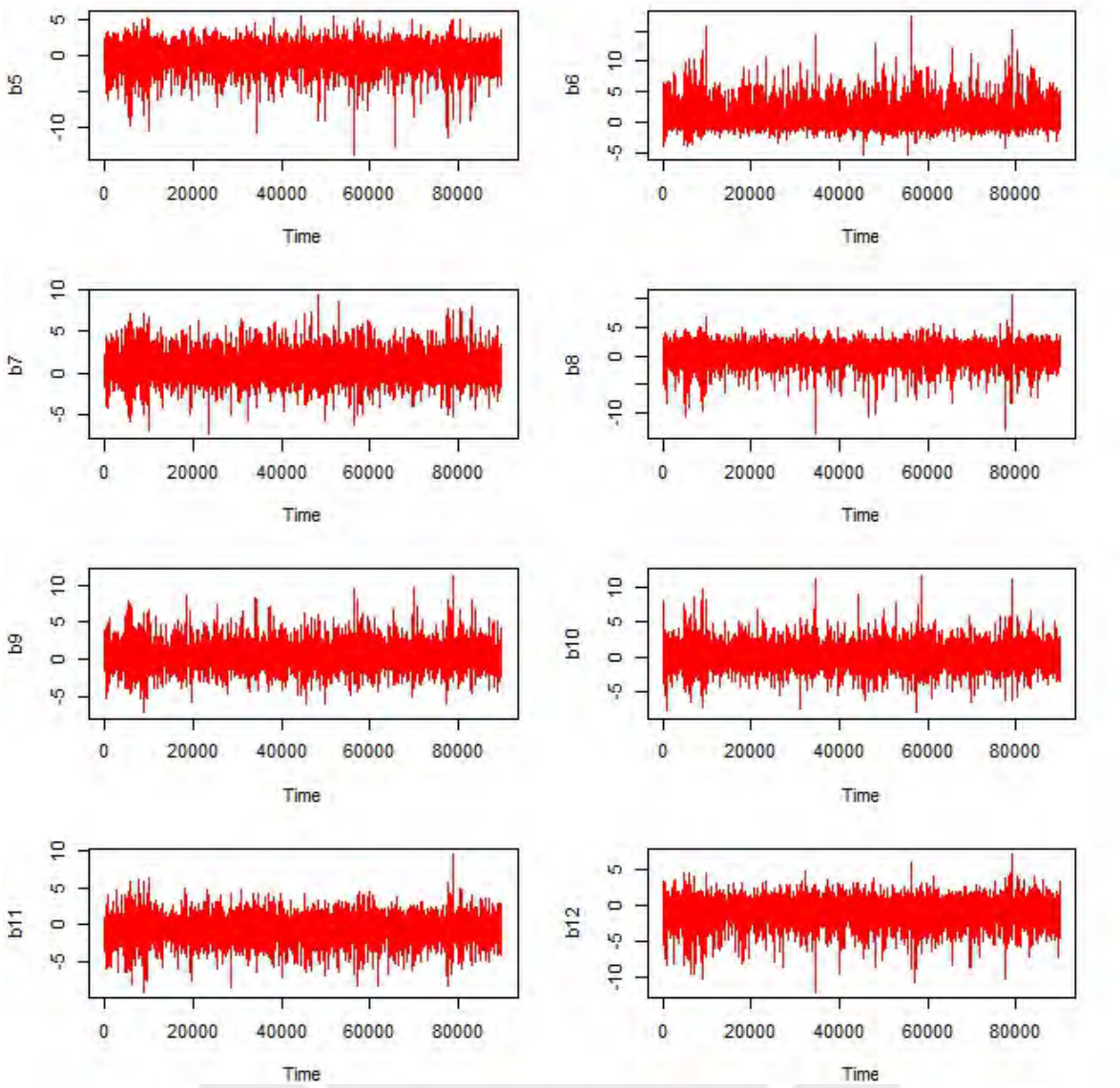


Figura B.8: Cadenas de la iteración del algoritmo de Gibbs para cada parámetro del MRS normal para la base de datos de exhalación

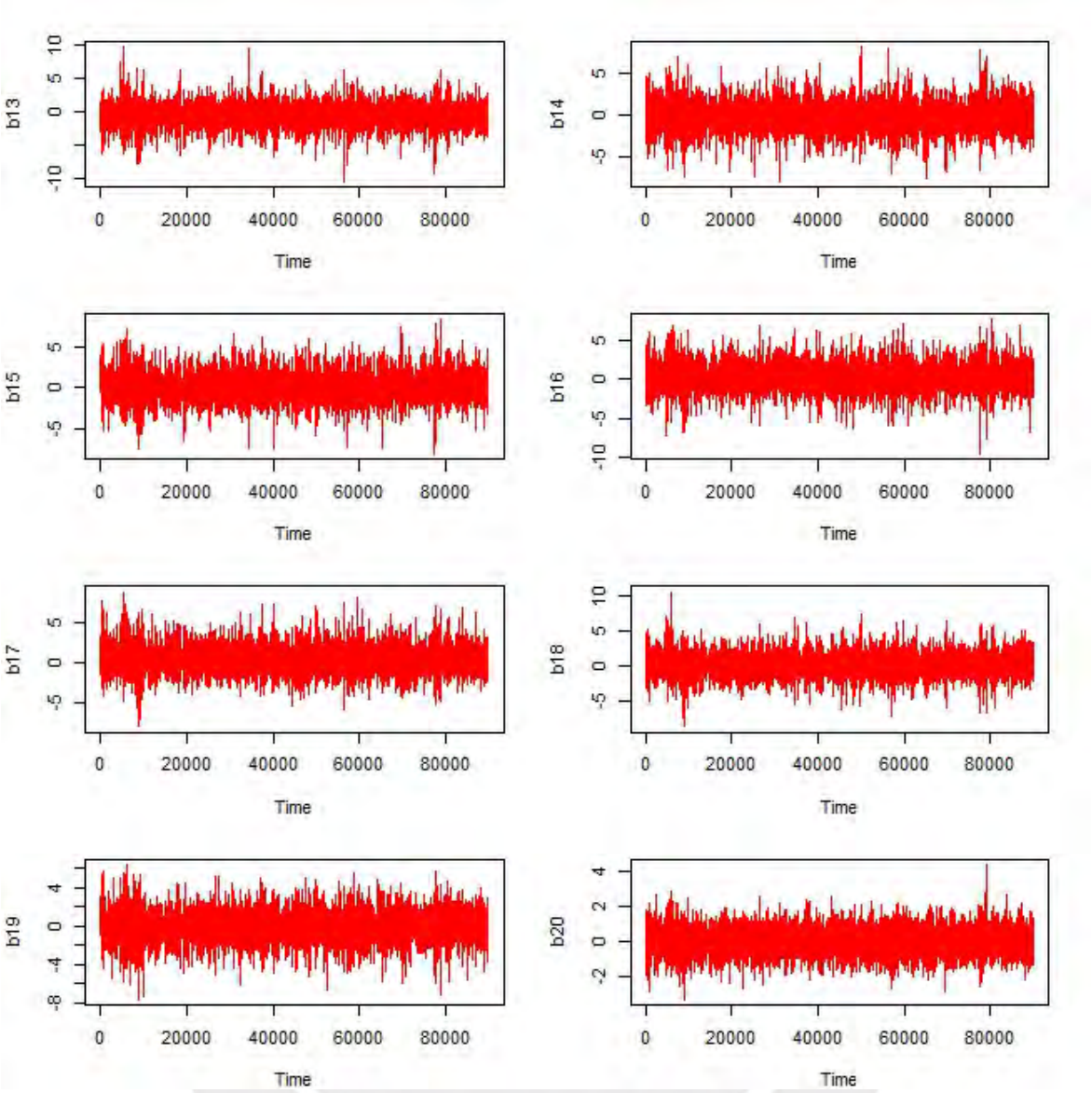


Figura B.9: Cadenas de la iteración del algoritmo de Gibbs para cada parámetro del MRS normal para la base de datos de exhalación

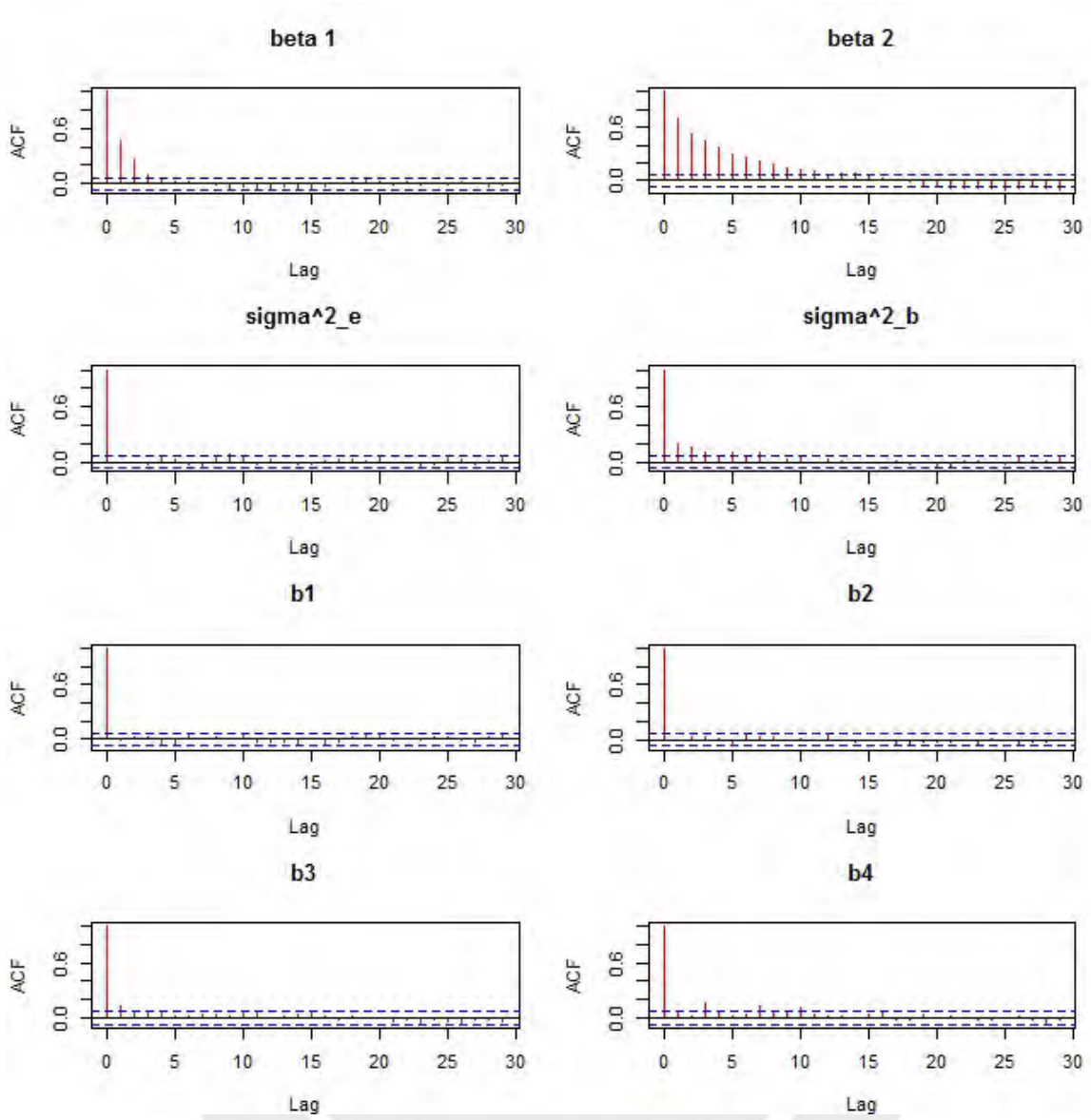


Figura B.10: Función de autocorrelación del MRS normal para la base de datos exhalación

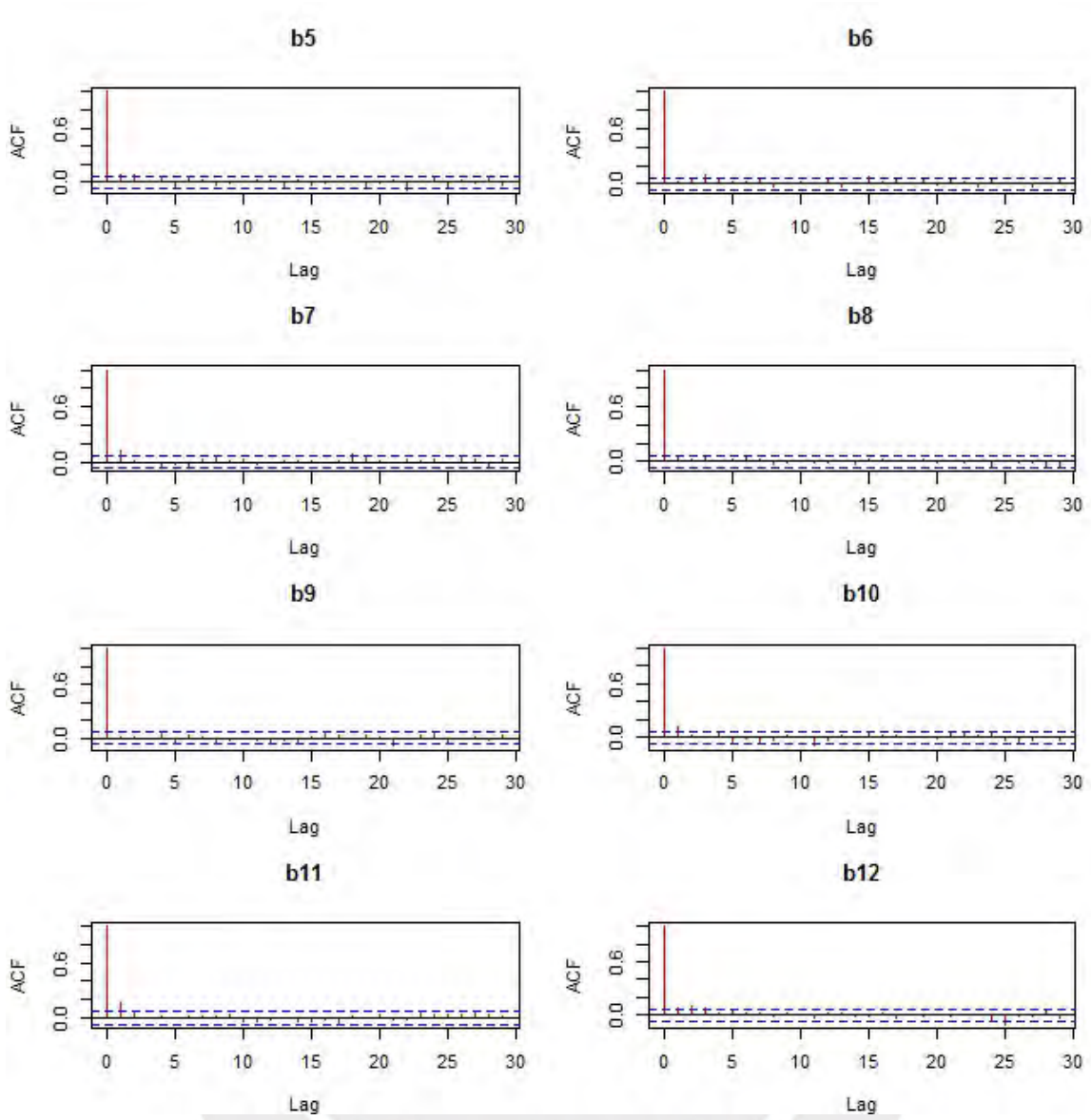


Figura B.11: Función de autocorrelación del MRS normal para la base de datos exhalación

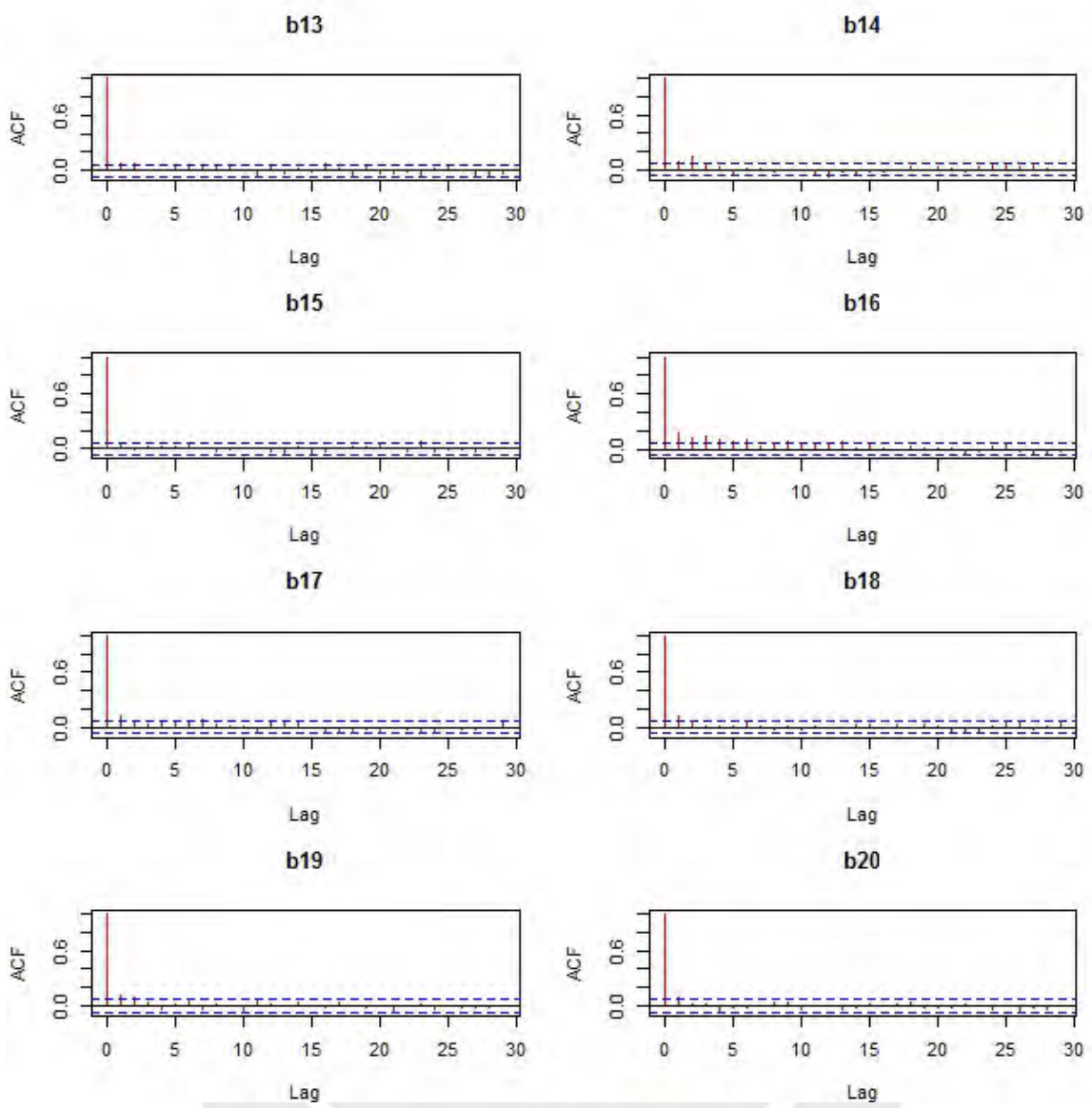


Figura B.12: Función de autocorrelación del MRS normal para la base de datos exhalación

Bibliografía

Crainiceanu, C. M., Ruppert, D. y Wand, M. P. (2005). Bayesian analysis for penalized spline regression using winbugs, *journal of statistical software* **14**: 1–24.

Gelman, A., Carlin, J. y Stern, H. (2014). *Bayesian Data Analysis-Third Edition*, Chapman and Hall.

Geweke, J. (1993). Bayesian treatment of the independent student-t linear model, *journal of applied econometrics* **8**: 19–40.

Gilks, N. G. B. y Tan, K. K. C. (1995). Adaptive rejection metropolis sampling within gibbs sampling, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **4**: 455–472.

Gilks, W. R. y Wild, P. (1992). Adaptive rejection sampling for gibbs sampling, *Applied Statistics* **41**: 337–348.

Hastie, T. y Tibshirani, R. (1986). Generalized additive models, *Statistical Science* **1**: 297–318.

Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*, Springer Texts in Statistics.

Lunn, D., Jackson, C., Best, N., Thomas, A. y Spiegelhalter, D. (2013). *The BUGS Book A Practical Introduction to Bayesian Analysis*, Chapman and Hall.

Peng, F. y Dey, D. K. (1995). Bayesian analysis of outlier problems using divergence measures, *The Canadian Journal of Statistics* **23**: 199–213.

Petris, G. y Tardella, L. (2013). *HI: Simulation from distributions supported by nested hyperplanes*. R package version 0.4.
URL: *https://CRAN.R-project.org/package=HI*

Rosa, G. J. M., Padovani, C. R. y Gianola, D. (2003). Robust linear mixed models with normal/independent distributions and bayesian MCMC implementation, *Biometrical Journal* **45**(5): 573–590.

Ruppert, D., Wand, M. P. y Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge University Press.

Staudenmayer, J., Lake, E. E. y Wand, M. P. (2009). Robustness for the general design mixed models using the t-distribution, *Statistical Modelling* **9(3)**: 235–255.