PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

# INFERENCIA BAYESIANA EN UN MODELO DE REGRESIÓN CUANTÍLICA SEMIPARAMÉTRICO

Tesis para optar el grado de Magíster en Estadística

AUTOR

Hugo Miguel Agurto Mejía

ASESOR

Dr. Cristian Luis Bayes Rodríguez

JURADO

Dr. Luis Hilmar Valdivieso Serrano

Dr. Cristian Luis Bayes Rodríguez

Mg. José Julio Flores Delgado

LIMA - PERÚ

2013

# Dedicatoria

A Dios por su infinita bondad y misericordia.

A mis padres y hermano, por su amor e incondicional apoyo en todo momento.

A Jovany Elizabeth, por su inagotable amor, paciencia y comprensión en el esfuerzo que implica cumplir con mis objetivos de desarrollo profesional.

# Agradecimientos

Al Dr. Cristian Bayes por su orientación, apoyo, exigencia y tiempo brindado en el proceso de investigación y desarrollo de esta tesis.

A los profesores de la maestría de Estadística de la PUCP, que de una u otra forma colaboraron con sus observaciones y comentarios.

A la Dirección de la Facultad de Ciencias Económicas y Empresariales de la Universidad de Piura y al Jefe del Área de Matemáticas, Enrique Li, por su comprensión y apoyo.

# Resumen

Este trabajo propone un Modelo de Regresión Cuantílica Semiparamétrico. Nosotros empleamos la metodología sugerida por Crainiceanu et al. (2005) para un modelo semiparamétrico en el contexto de un modelo de regresión cuantílica. Un enfoque de inferencia Bayesiana es adoptado usando Algoritmos de Montecarlo vía Cadenas de Markov (MCMC). Se obtuvieron formas cerradas para las distribuciones condicionales completas y así el algoritmo muestrador de Gibbs pudo ser fácilmente implementado. Un Estudio de Simulación es llevado a cabo para ilustrar el enfoque Bayesiano para estimar los parámetros del modelo. El modelo desarrollado es ilustrado usando conjuntos de datos reales.

**Palabras clave:** Regresión Cuantílica, Regresión Semiparamétrica, Inferencia Bayesiana.

# Abstract

This work proposes a Semiparametric Quantile Regression Model. We use the methodology suggested by Crainiceanu et al. (2005) for a semiparametric model in the context of a quantile regression model. A Bayesian inference approach is adopted using Markov Chain Monte Carlo algorithms (MCMC). We obtain closed forms for the full conditional distributions so the Gibbs sampler algorithm can be easily implemented. A simulation study is carried out to illustrate the Bayesian approach to estimate the parameters in the model. The model developed is illustrated using real data sets.

**Keywords:** Quantile Regression, semiparametric Regression, Bayesian Inference.

# Índice general

# Índice de figuras

# Índice de cuadros

# Capítulo 1

# Introducción

## 1.1. Consideraciones Preliminares

En muchas ocasiones podemos estar interesados en estudiar el comportamiento de una variable dependiente dado un conjunto de variables explicativas (o covariables). 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. Aunque la media es una medida importante que representa la tendencia central de la distribución, esta provee poca información acerca del comportamiento de los extremos (colas) de la distribución. En este caso la regresión cuantílica permite estimar diferentes cuantiles de la distribución (incluyendo la mediana,que también representa la tendencia central de los datos) y de esta manera brinda una mayor información sobre la distribución condicional, de la variable en estudio, dada las covariables.

La regresión cuantílica permite determinar la influencia de covariables sobre los cuantiles condicionales de la distribución de una variable dependiente. Por lo tanto una de las principales ventajas sobre la regresión de la media es que la regresión cuantílica permite obtener detallada información acerca de la distribución condicional de la variable en estudio, en lugar de solo la media. Además, los datos o valores extremos y atípicos tienen menor influencia en la regresión cuantílica debido a la inherente robustez de los cuantiles.

Desde el importante trabajo de Koenker y Basset (1978), la regresión cuantílica ha recibido creciente atención. Este es un procedimiento estadístico basado en la minimización de la suma de residuales absolutos ponderados con pesos asimétricos, esto es, se asignan pesos diferentes a los residuales positivos y a los negativos. Este procedimiento puede ser empleado para explorar la relación entre los cuantiles de la distribución de la variable respuesta y las covariables disponibles. Un caso especial de la regresión cuantílica es la regresión de la mediana, en cuyo caso los pesos son simétricos y la regresión tiene por objetivo minimizar la suma de las desviaciones en términos absolutos sin ponderar.

Así como este enfoque inicial de la regresión cuantílica, los tratamientos frecuentistas al respecto son no paramétricos por lo que no requieren de una distribución específica de la variable respuesta. Una formulación bayesiana de la regresión cuantílica propuesta por Yu y Moyeed (2001) asume la distribución asimétrica de Laplace para la distribución de los términos de los errores. El empleo de dicha distribución para los errores provee una forma natural para tratar con el problema de la regresión cuantílica desde la perspectiva Bayesiana.

Por otro lado, la suposición usual de una relación lineal entre un parámetro (por ejemplo, la media o un cuantil de una distribución) con un conjunto de variables explicativas no siempre

es satisfecha. Es decir, muchos de los problemas prácticos de regresión cuantílica requieren, además, formas flexibles semiparamétricas del predictor para modelar la dependencia de la respuesta sobre las covariables. En Hastie y Tibshirani (1986) se propone una clase de modelos denominados Modelos Aditivos Generalizados donde se propone relajar el supuesto de linealidad considerando que esta relación puede ser no lineal utilizando para esto un enfoque semiparamétrico. En Crainiceanu et al. (2005) se presenta cómo realizar la estimación de este tipo de modelos considerando un enfoque Bayesiano.

La presente investigación muestra una variante del trabajo realizado en Crainiceanu et al. (2005), como metodología de regresión semiparamétrica, pero aplicada a un modelo de regresión cuantílica. Este será conocido como un modelo de regresión cuantílica semiparamétrico. En este trabajo estaremos interesados en la estimación de los parámetros de este modelo desde la perspectiva Bayesiana.

## 1.2. Objetivos

El objetivo general de la tesis es estudiar, estimar y aplicar a conjuntos de datos reales el modelo de regresión cuantílica semiparamétrico (mediante splines), desde el punto de vista de la inferencia Bayesiana. De manera específica:

- Revisar la literatura acerca de: modelos de regresión cuantílica, regresión semiparamétrica (mediante splines específicamente).
- Estudiar e implementar la estimación del modelo de regresión cuantílica semiparamétrico (mediante splines) desde la perspectiva Bayesiana.
- Realizar estudios de simulación.
- Aplicar el modelo a conjuntos de datos reales.

## 1.3. Organización del Trabajo

En el Capítulo 2, se presenta una serie de conceptos preliminares involucrados con el modelo que es motivo de este trabajo. Conceptos como el de cuantiles y su importancia que dan lugar a reconocer las ventajas de la regresión cuantílica sobre la común regresión sobre la media. También se explican algunos aspectos importantes en la implementación de la regresión cuantílica relacionados con la función de chequeo resaltando que es la función de pérdida más apropiada para este tipo de regresión. Igualmente se dan detalles sobre la distribución Asimétrica de Laplace y se reconoce que esta función ha permitido el desarrollo del enfoque bayesiano en la regresión cuantílica. Por último, también se tratan aspectos teóricos generales sobre la regresión semiparamétrica y particulares de esta mediante splines.

En el capítulo 3 se dan detalles sobre el modelo que es motivo de este trabajo, el modelo de regresión cuantílica semiparamétrico así como su estimación desde la perspectiva Bayesiana.

En el capítulo 4 se realiza un estudio de simulación con diferentes escenarios para la generación de datos.

En el capítulo 5 se muestran dos aplicaciones del modelo con bases de datos reales: una para datos del estudio Canadian age-income y otra para datos empleados en un estudio sobre la aplicación de la ecuación de Mincer para Lima Metropolitana.

Finalmente, en el Capítulo 6 se discuten algunas conclusiones obtenidas en este trabajo.

En el anexo (Apéndice A) presentamos los programas implementados en la simulación y las aplicaciones a conjuntos de datos reales. En el anexo (Apéndice B) presentamos tablas de resultados y gráficos.

# Capítulo 2

# Conceptos Preliminares

A continuación resaltamos algunos conceptos importantes que emplearemos en nuestra investigación.

## 2.1. Cuantiles y Regresión Cuantílica

Los cuantiles están relacionados a las operaciones de ordenamiento y clasificación de las observaciones de una muestra o población. Así como se puede definir a la media muestral como la solución al problema de minimizar una suma de residuales cuadráticos, se puede definir a la mediana como la solución al problema de minimizar una suma de residuales absolutos (Koenker y Hallock, 2001). La simetría de la función de valor absoluto lineal a trozos implica que la minimización de la suma de residuales absolutos deba igualar el número de residuales positivos y negativos, asegurando que haya por tanto el mismo número de observaciones por debajo y por encima de la mediana.

¿Qué sucede con los otros cuantiles? Ya que la simetría del valor absoluto conduce a la mediana, entonces minimizando una suma asimétricamente ponderada de residuales absolutos (simplemente dando diferentes pesos a los residuales positivos y negativos) nos conduciría a los cuantiles. En efecto ese es el caso. Resolviendo

$$\min_{\xi \in \mathbb{R}} \sum \rho_\tau(y_i - \xi), \tag{2.1}$$

donde

$$\rho_\tau(\mu) = \begin{cases} \mu\tau & \text{si } \mu \geq 0 \\ \mu(\tau - 1) & \text{si } \mu < 0 \end{cases},$$

es la función de valor absoluto inclinada que aparece en la figura 2.1, se obtiene el $\tau$-ésimo cuantil muestral como su solución (ver para más detalles Zevallos, 2012, cap.2).

Al haber logrado definir los cuantiles incondicionales como un problema de optimización, es fácil definir los cuantiles condicionales de forma análoga. La regresión de mínimos cuadrados nos da una idea de cómo proceder. Si, observada una muestra aleatoria $\{y_1, y_2, ..., y_n\}$, resolvemos:

$$\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2,$$

obtenemos la media muestral como un estimado de la media poblacional incondicional, $E(Y)$.

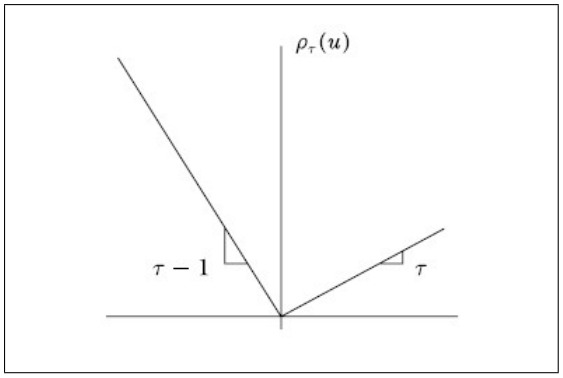


Figura 2.1: Función $\rho$ para la regresión cuantílica, la cual tiene pendiente $\tau$ respecto del eje X sobre la derecha y pendiente $\tau$-1 respecto del eje X sobre la izquierda.

Si reemplazamos el escalar $\mu$ por una función paramétrica $\mu$(x,$\beta$)y resolvemos:

$$\min_{\mu \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \mu(x, \beta))^2,$$

obtenemos un estimado de la función esperanza condicional $E(Y \mid x)$.

En la regresión cuantílica, procedemos exactamente de la misma manera. Para obtener un estimado de la función de la mediana condicional, simplemente reemplazamos el escalar $\xi$ en la ecuación (2.1) por la función paramétrica $\xi(x_i,\beta)$ y establecemos $\tau$ igual a $\frac{1}{2}$. Para obtener estimados de las otras funciones de cuantil condicional, reemplazamos los valores absolutos por $\rho_\tau(.)$ y resolvemos:

$$\min_{\beta \in \mathbb{R}^p} \sum \rho_\tau(y_i - \xi(x_i, \beta))$$

El problema resultante de minimización, cuando $\xi(x, \beta)$ es formulado como una función lineal de parámetros, puede ser resuelto eficientemente por métodos de programación lineal.

Por lo tanto, la regresión cuantilíca, busca extender las ideas del concepto de los cuantiles a la estimación de funciones de cuantil condicional, es decir, modelos en los cuales los cuantiles de la distribución condicional de la variable respuesta (dependiente) sean expresados como funciones de las covariables observadas, para así poder determinar la influencia de las covariables sobre los cuantiles condicionales de la distribución de la variable dependiente.

Algunas de las principales ventajas de la regresión cuantílica sobre la regresión de la media son:

- Permite obtener detallada información acerca de la distribución condicional de la variable en estudio, en lugar de solo la media.
- Los datos o valores extremos y atípicos influyen menos en la regresión cuantílica debido a la inherente robustez de los cuantiles.
- También es más apropiada cuando el modelo subyacente a los datos es no lineal o cuando los términos de los errores siguen una distribución que no es normal o cuando

las colas de la distribución subyacente son de interés para modelar el comportamiento de valores extremos de la población.

Refiriéndonos ahora al modelo de regresión cuantílica podemos decir que, dado un cuantil fijo $\tau \in (0,1)$, el modelo de regresión cuantílica lineal es:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}_\tau + \varepsilon_{\tau i}, \qquad \varepsilon_{\tau i} \sim F_{\tau i}, \text{ sujeto a } F_{\tau i}(0 \mid x_i) = \tau \,, \tag{2.2}$$

donde el error aleatorio $\varepsilon_{\tau i}$ sigue una función distribución acumulada no especificada $F_{\tau i}$ y su $\tau$-ésimo cuantil condicional en $\mathbf{x}_i$ es igual a cero; $y_i$ denota la $i$-ésima observación de la variable dependiente, $\mathbf{x}_i$ es el correspondiente vector de covariables (incluyendo un intercepto). Los efectos lineales de un cuantil específico son dados por $\boldsymbol{\beta}_\tau$. Observamos que el modelo dado en (2.2) asume que:

$$Q_{Y_i}(\tau \mid \mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}_\tau, \tag{2.3}$$

donde $Q_{Y_i}(\tau \mid \mathbf{x}_i) = inf\{y : F_{Y_i}(y \mid \mathbf{x}_i) \geq \tau\}$ es el $\tau$-ésimo cuantil condicional de $y_i$ dado el vector de covariables $\mathbf{x}_i$.

Para la regresión cuantil lineal clásica, como fue presentada por Koenker y Basset (1978), la estimación de los coeficientes $\boldsymbol{\beta}_\tau$ de un cuantil específico, depende de la minimización de la suma de las desviaciones absolutas ponderadas asimétricamente:

$$\min_{\beta_\tau} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\tau) \tag{2.4}$$

donde $(y_i, \mathbf{x}_i)$, i=1,...,n son los valores observados de la variable respuesta y el vector de covariables para $n$ observaciones,

$$\rho_\tau(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\tau) = \begin{cases} \tau(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\tau) & \text{si } y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\tau \geq 0 \\ (1-\tau)(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\tau) & \text{si } y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\tau < 0 \end{cases}$$

es la función de chequeo, la cual también se puede expresar equivalentemente como

$$\rho_\tau(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\tau) = (y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\tau) \left[\tau - I(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_\tau \leq 0)\right], \tag{2.5}$$

donde I(.) es la función indicadora y $\tau \in (0,1)$ es el cuantil de interés. Así, la función de chequeo es la función de pérdida apropiada para los problemas de regresión cuantílica. Este enfoque es completamente no paramétrico, no requiere la suposición de una distribución específica para el término del error o para la respuesta, por lo tanto, tampoco de una función de verosimilitud para la muestra. No existe solución de forma cerrada para el problema de minimización y los estimados de la regresión cuantílica son obtenidos mediante programación lineal (ver Koenker (2005) para más detalles). Ante el escenario anterior, Yu y Moyeed (2001) introdujeron una distribución para el error que permite estimar los parámetros de la regresión cuantílica a través de la verosimilitud de la Distribución Asimétrica de Laplace (ALD). La contribución de Yu y Moyeed (2001) fue identificar que la maximización de la verosimilitud de la ALD se logra en el mismo punto donde la función de chequeo se minimiza.

Ya que la inferencia bayesiana requiere de una verosimilitud, la identificación de la función de verosimilitud de la ALD ha permitido el desarrollo del enfoque Bayesiano de la regresión cuantílica. A continuación, se dará una visión general de la ALD.

## 2.2. Distribución Asimétrica de Laplace

Koenker y Machado (1999) y Yu y Moyeed (2001) fueron los primeros en aplicar esta distribución asimétrica en la regresión cuantílica. En Yu y Zhang (2005) se denota que existen diversas especificaciones para la ALD. En este trabajo emplearemos una de esas especificaciones, que coincide con la propuesta de Yu y Moyeed (2001), la cual tiene la función densidad de probabilidad siguiente:

$$f(y;\mu,\sigma,\tau) = \frac{\tau(1-\tau)}{\sigma} exp\left\{-\frac{1}{\sigma}\rho_\tau(y-\mu)\right\}, \qquad y \in \mathbb{R} \tag{2.6}$$

donde $\tau$ es el parámetro de asimetría, $\sigma$ es el parámetro de escala y $\mu$ es el parámetro de localización, con $0 < \tau < 1$, $\sigma > 0$ y $\mu \in \mathbb{R}$. Utilizaremos la notación: $Y \sim ALD(\mu,\sigma,\tau)$ para representar a una variable aleatoria que sigue una distribución asimétrica de Laplace cuya función de densidad está dada en (2.6). En el trabajo de Yu y Zhang (2005) también se especifican:

Función de distribución acumulada:

$$F_Y(y) = \begin{cases} \tau exp\left(\dfrac{1-\tau}{\sigma}(y-\mu)\right) & \text{si } y < \mu \\ 1-(1-\tau)exp\left(-\dfrac{\tau}{\sigma}(y-\mu)\right) & \text{si } y \geq \mu \end{cases} \tag{2.7}$$

Función generadora de momentos:

$$m_Y(t) = \tau(1-\tau)\frac{exp(\mu t)}{(\tau-\sigma t)(\sigma t+1-\tau)} \tag{2.8}$$

Media y varianza poblacional:

$$E[Y] = \mu + \frac{\sigma(1-2\tau)}{\tau(1-\tau)} \qquad V[Y] = \frac{\sigma^2(1-2\tau+2\tau^2)}{(1-\tau)^2\tau^2} \tag{2.9}$$

La ALD tiene características que la hacen útil en el ámbito de la regresión cuantílica: la maximización de la función de verosimilitud de la ALD ocurre en el punto donde se minimiza la función de pérdida basada en la función de chequeo, facilita la inferencia Bayesiana y el cómputo en los modelos aditivos, es suficientemente flexible para adaptarse a diferentes tipos de distribuciones para los términos de error.

Por otro lado, asumiendo que una variable aleatoria $y \overset{iid}{\sim} ALD(\mu,\sigma,\tau)$, entonces la función de verosimilitud para $n$ observaciones de dicha variable aleatoria es

$$\mathcal{L}(\mu \mid \mathbf{y}) = \left(\frac{\tau(1-\tau)}{\sigma}\right)^n exp\left\{\sum_{i=1}^{n} -\frac{\rho_\tau(y_i-\mu)}{\sigma}\right\}$$

y su log-verosimilitud

$$l = log(L(\mu \mid \mathbf{y})) = n\ln(\tau) + n\ln(1-\tau) - n\ln(\sigma) - \frac{1}{\sigma}\sum_{i=1}^{n}\rho_\tau(y_i - \mu). \tag{2.10}$$

Entonces, si el parámetro $\mu$ puede ser especificado en forma lineal como $\mathbf{x}_i^T\boldsymbol{\beta_\tau}$, la maximización de la función de verosimilitud en torno al argumento $\boldsymbol{\beta_\tau}$ es equivalente a la minimización del esperado de la función de chequeo (ver para más detalles Zevallos, 2012, cap.2). De este modo, la ALD es útil en términos de unificar la verosimilitud y la inferencia para la estimación de la regresión cuantílica. Con la misma suposición sobre el término de error, Yu y Moyeed (2001), entre otros, lograron implementar la inferencia Bayesiana para la regresión cuantílica (paramétrica). Por otra parte, en trabajos como el de Geraci y Bottai (2007) y Yuan y Yin (2010) se estudió para una data longitudinal.

Otra característica de la ALD es que puede ser representada como una mixtura de dos variables aleatorias: una distribución normal estándar y una distribución exponencial estándar independientes. Esta representación fue dada por Kotz et al. (2001) y utilizada por Kozumi y Kobayashi (2011) para implementar la inferencia bayesiana en la regresión cuantílica. Dicha representación mixta se expresa en la proposición siguiente.

**Proposición.** *Si una variable aleatoria $\varepsilon$ sigue una distribución ALD $(0, \sigma, \tau)$, entonces esta variable se puede representar como una mixtura dada por*

$$\varepsilon = \theta_1 v + \theta_2 \sigma^{\frac{1}{2}} v^{\frac{1}{2}} w \ ,$$

donde

$$v \sim exp(\tfrac{1}{\sigma}),$$
$$w \sim N(0,1),$$

$v$ y $w$ son independientes, $\theta_1 = \dfrac{1-2\tau}{\tau(1-\tau)}$ y $\theta_2 = \sqrt{\dfrac{2}{\tau(1-\tau)}}$.
Para la demostración véase Zevallos (2012), cap.3, sección 3.2.2.

En consecuencia, la variable $\varepsilon$ puede expresarse a través de la representación jerárquica siguiente:

$$\varepsilon \mid v \sim N(\theta_1 v, \theta_2^2 \sigma v),$$
$$v \sim exp(\tfrac{1}{\sigma}),$$

El empleo de esta representación jerárquica facilita la implementación de la inferencia bayesiana sobre la regresión cuantílica y permite reescribir el modelo de regresión cuantílica bayesiana como un modelo de regresión normal condicional a variables latentes.

## 2.3. Regresión Semiparamétrica

La regresión semiparamétrica es una fusión entre la regresión paramétrica y la no paramétrica. Este emergente campo de estudio combina la investigación realizada en varias ramas de la estadística como la regresión paramétrica y no paramétrica, el análisis de datos longitudinales y espaciales, los modelos Bayesianos jerárquicos y mixtos y los algoritmos de Montecarlo vía cadenas de Markov.

La regresión semiparamétrica no debe verse como una competencia de los enfoques paramétricos y no paramétricos, sino como un puente entre ellos. La necesidad de modelos estadísticos parsimoniosos es bien conocida y los modelos paramétricos son a menudo un conveniente método para lograr parsimonia. Sin embargo, los modelos no paramétricos son útiles porque hay muchos casos donde los modelos paramétricos no proveen un adecuado ajuste de los datos.

La modelación semiparamétrica permite a un investigador tener lo mejor de ambos enfoques para obtener un modelo de regresión que describa mejor el comportamiento de los datos, es decir, aquellas características de los datos que son adecuadas para la modelación paramétrica son modeladas de esta forma y las componentes no paramétricas son empleadas solo donde sea necesario, Ruppert et al. (2009).

Dos características importantes en gran parte de la regresión semiparamétrica, Ruppert et al. (2009), son:

- Facilitar la parte de regresión no paramétrica utilizando splines[1] penalizados de bajo rango.[2]
- Emplear la representación del modelo mixto de los splines penalizados.

Estas brindan varios beneficios: los efectos longitudinales y espaciales pueden ser fácilmente incorporados en el modelo, el ajuste y la inferencia pueden ser desarrollados dentro de los marcos establecidos de máxima verosimilitud y mejor predicción.

## 2.4. Regresión Semiparamétrica mediante splines

En los modelos de regresión semiparamétricos, los splines penalizados pueden ser usados para describir complejas relaciones no lineales entre la media de la variable respuesta y las covariables. La metodología general de modelado semiparamétrico empleando la equivalencia entre splines penalizados y modelos mixtos es presentada en Ruppert et al. (2003). Consideremos el modelo de regresión:

$$y_i = m(x_i) + \epsilon_i, \tag{2.11}$$

donde los $\epsilon_i$ son i.i.d. N(0,$\sigma_\epsilon^2$) e independientes de $x_i$, y $m(.)$ es una *smooth function*[3] la cual se define como una función suave de los datos porque puede ser modelada fácilmente utilizando splines como sugiere Crainiceanu et al. (2005). Es evidente que un modelo de este tipo es una generalización de un modelo de regresión, que indudablemente, tendrá un coste computacional, pero que nos permitirá estimar la función de una forma más precisa. La estimación de la función $m(x_i)$ se puede realizar mediante distintos métodos divididos en dos grandes grupos: regresión tipo *kernel* y la regresión con *splines*.

Los modelos tipo *kernel* se basan en la idea de que al estimar la función en un punto $x_0$, es deseable dar más peso a las observaciones que están próximas a ese punto. Los pesos son

[1] Un spline es una curva diferenciable definida en tramos mediante polinomios. Estos polinomios a trozos se unen en puntos llamados nodos.

[2] Un spline es de rango bajo, si el tamaño de la base utilizada es mucho menor que la dimensión de los datos.

[3] Es una función que tiene derivadas continuas hasta un orden deseado sobre un dominio determinado.

asignados mediante una *función kernel* e irán disminuyendo conforme nos vayamos alejando del punto $x_0$. En este trabajo no se va a tratar este tipo de métodos.

El segundo grupo de técnicas de suavizado están basadas en splines. Hay dos grandes familias dentro de los modelos de suavizado con splines:

1. **Splines de regresión** (regression splines). En estos modelos es necesario seleccionar el número y la localización de los nodos (para controlar la suavidad de la función ajustada) e imponer restricciones para que los trozos de polinomio se unan de forma suave. Una vez hecha la elección, el modelo se ajusta por mínimos cuadrados.

2. **Splines de suavizado** (smoothing splines). Aparecen como la solución al problema de regresión no-paramétrica: encontrar la función (con derivadas continuas de segundo orden en el intervalo $[a, b]$) que minimice la *suma de residuales cuadrados penalizada*:

$$SCP = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_a^b (f''(x))^2 dx, \tag{2.12}$$

para un valor pre establecido $\lambda > 0$. El primer término en (2.12) denota la suma de los residuales cuadrados y penaliza la falta de ajuste. El segundo término, el cual es ponderado por $\lambda$, es una penalización en la segunda derivada de la curva denotando la penalidad de rugosidad, en otras palabras penaliza la curvatura de la función $f$. El parámetro $\lambda$ es el denominado parámetro de suavizado que controla la suavidad de la función. Como $\lambda$ varía de 0 a $+\infty$, la solución varía desde la interpolación hasta un ajuste lineal. Cuando $\lambda \to \infty$ la penalidad de rugosidad domina en (2.12) y el spline estimado es forzado a ser una constante. Cuando $\lambda \to 0$, la penalidad de rugosidad en (2.12) desaparece y entonces el spline estimado interpola los datos. Así el parámetro de suavizado $\lambda$ juega un rol importante en controlar la compensación entre la bondad de ajuste representada por $\sum_{i=1}^{n}(y_i - f(x_i))^2$ y la suavidad de la estimación medida por $\int_a^b (f''(x))^2 dx$ (ver para más detalles Green y Silverman, 1994).

Sin embargo, ambas técnicas presentan inconvenientes: en los splines de regresión la suavidad de la función ajustada depende de la elección de los nodos y esta elección se hace mediante complicados algoritmos que no son fáciles de extender al caso multidimensional. En el caso de los splines de suavizado los problemas son de tipo computacional, ya que este tipo de splines utilizan tantos nodos (y por lo tanto parámetros) como observaciones.

Los splines con penalizaciones basadas en diferencias entre coeficientes adyacentes (Eilers y Marx, 1996), combinan lo mejor de ambos enfoques: utilizan menos parámetros que los splines de suavizado, pero la selección de los nodos no es tan determinante como en los splines de regresión. Son splines de rango bajo, el número de nodos es mucho menor que la dimensión de los datos, al contrario de lo que ocurre en el caso de los splines de suavizado. El número de nodos, en el caso de los splines penalizados, no supera generalmente los 40, lo que hace que sean computacionalmente eficientes, sobre todo cuando se trabaja con gran cantidad de datos. Además, la introducción de penalizaciones relaja la importancia de la elección del número y la localización de los nodos, cuestión que es de gran importancia en los splines de rango bajo sin penalizaciones. Por último, la correspondencia entre los splines penalizados

y el mejor predictor lineal insesgado (BLUP, por sus siglas en inglés) en un modelo mixto permite, en algunos casos, utilizar la metodología existente en el campo de los modelos mixtos y el uso de software como PROC MIXED en SAS, y lme() en S-PLUS y R.

Regresando al modelo de regresión dado en (2.11) donde los $\epsilon_i$ son i.i.d. N(0,$\sigma_\epsilon^2$), $\epsilon_i$ es independiente de $x_i$, y $m(.)$ es una *smooth function* o función suave de los datos. Esta última podría ser modelada usando splines cúbicos naturales, B-splines, polinomios truncados, splines radiales, etc. En nuestro trabajo, siguiendo a Crainiceanu et al. (2005), nos enfocaremos en los *thin-plate splines*[4] de bajo rango, ya que tienen buenas propiedades matemáticas. En particular, la correlación posterior de los parámetros de los thin plate splines es mucho menor que con otras bases, lo cual los hace más estables. Esto facilita la implementación del algoritmo de Gibbs al estar los parámetros menos correlacionados.

Siguiendo la notación de Crainiceanu et al. (2005), la representación de $m(.)$ como thin-plate spline de bajo rango es

$$m(x) \equiv m(x, \boldsymbol{\theta}) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k \left|x - \kappa_k\right|^3, \tag{2.13}$$

en la nueva notación se enfatiza que la función suave $m$ depende también del vector de coeficientes de regresión, $\boldsymbol{\theta} = (\beta_0, \beta_1, u_1, \ldots, u_K)^\top$ y $\kappa_1 < \kappa_2 < \ldots < \kappa_K$ son nodos fijos. Para la selección y localización de los nodos es suficiente con elegir un número moderadamente grande de nodos equidistantes (de 5 a 20) para asegurar la flexibilidad deseada. De acuerdo con Ruppert et al. (2003) se sugiere elegir los K nodos en los K-cuantiles de $x$; es decir, cada nodo $\kappa_k$ sería el cuantil $k/(K+1)$ de $x$. Ya que la estimación de la curva se logra vía minimización de residuales cuadrados muestra, usualmente, más variación que es justificada por los datos; por lo tanto, para evitar el efecto de sobreajuste se introduce una penalidad de rugosidad y el problema de estimación se reformula como el problema de minimizar

$$\sum_{i=1}^{n} \{y_i - m(x_i, \boldsymbol{\theta})\}^2 + \frac{1}{\lambda} \boldsymbol{\theta}^\top D \boldsymbol{\theta}, \tag{2.14}$$

donde $\lambda$ es el parámetro de suavizado, que controla la compensación entre la bondad de ajuste y el grado de suavizado, y $D$ es una matriz de penalidad semi-definida positiva conocida. La matriz de penalidad thin-plate spline es

$$D = \begin{bmatrix} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times K} \\ \mathbf{0}_{K\times 2} & (\boldsymbol{\Omega}_K^{1/2})^T \boldsymbol{\Omega}_K^{1/2} \end{bmatrix}$$

donde la $(l,k)$ *ésima* entrada de $\boldsymbol{\Omega}_K$ es $|\kappa_l - \kappa_k|^3$ y penaliza solo coeficientes de $|x - \kappa_k|^3$. Luego consideraremos la siguiente notación: sea $\mathbf{Y} = (y_1, y_2, \ldots, y_n)^\top$, $\mathbf{X}$ la matriz con $i - ésima$ fila $\mathbf{X}_i = (1, x_i)$, y $\mathbf{Z}_K$ la matriz con $i - ésima$ fila $\mathbf{Z}_{K_i} = \left\{|x_i - \kappa_1|^3, \ldots, |x_i - \kappa_K|^3\right\}$. Si se divide (2.14) por la varianza del error se obtiene

$$\frac{1}{\sigma_\epsilon^2} \left|\left|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}_K\right|\right|^2 + \frac{1}{\lambda \sigma_\epsilon^2} \boldsymbol{u}^T (\boldsymbol{\Omega}_K^{1/2})^T \boldsymbol{\Omega}_K^{1/2} \boldsymbol{u} \tag{2.15}$$

[4] El thin plate spline es la generalización en dos dimensiones del spline cúbico.

donde $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ y $\boldsymbol{u} = (u_1, \ldots, u_K)^\top$. Luego, se define $\sigma_u^2 = \lambda\sigma_\varepsilon^2$ y se considera el vector $\boldsymbol{\beta}$ como un conjunto de parámetros fijos y el vector $\boldsymbol{u}$ como un conjunto de parámetros aleatorios con $E(u) = 0$ y $cov(u) = \sigma_u^2 \boldsymbol{\Omega}_K^{-1/2}(\boldsymbol{\Omega}_K^{-1/2})^T$. Si $(\boldsymbol{u}^\top, \boldsymbol{\epsilon}^\top)^\top$ es un vector aleatorio normal y además $\boldsymbol{u}$ y $\boldsymbol{\epsilon}$ son independientes, el problema de minimizar la función dada en (2.15) puede ser expresado equivalentemente al de encontrar el estimador de máxima verosimilitud de un modelo lineal mixto de la forma, ver (Ruppert et al., 2003) o (Huaraz, 2012)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_K\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{2.16}$$

con

$$\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 \boldsymbol{\Omega}_K^{-1/2}(\boldsymbol{\Omega}_K^{-1/2})^T & 0 \\ 0 & \sigma_\epsilon^2 \boldsymbol{I}_n \end{pmatrix}\right)$$

Usando la reparametrización $\mathbf{b} = \boldsymbol{\Omega}_K^{1/2} u$ y definiendo $\mathbf{Z} = \mathbf{Z}_K \boldsymbol{\Omega}_K^{-1/2}$ el modelo mixto (2.16) es equivalente a

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \tag{2.17}$$

con

$$\begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_b^2 \boldsymbol{I}_K & 0 \\ 0 & \sigma_\epsilon^2 \boldsymbol{I}_n \end{pmatrix}\right)$$

donde $I_r$ representa una matriz identidad de dimensión $r$.

El modelo mixto (2.17) puede ser ajustado desde un punto de vista frecuentista usando el mejor predictor lineal insesgado (BLUP, por sus siglas en inglés) o estimación de cuasi-verosimilitud penalizada (PQL, por sus siglas en inglés). En nuestro trabajo adoptaremos una perspectiva de inferencia Bayesiana, estableciendo distribuciones a priori sobre los parámetros del modelo y obteniendo simulaciones de la distribución a posteriori.

# Capítulo 3

# Modelo de Regresión Cuantílica Semiparamétrico

## 3.1. Introducción

En muchos de los problemas prácticos de regresión cuantílica se requieren formas flexibles semiparamétricas del predictor para modelar la dependencia de la respuesta sobre las covariables, sobre todo en casos donde la relación entre las covariables y los cuantiles de la variable respuesta es no lineal. Por tanto, la presente investigación muestra una variante del trabajo realizado en Crainiceanu et al. (2005), como metodología de regresión semiparamétrica, pero aplicada a un modelo de regresión cuantílica. Este será conocido como un modelo de regresión cuantílica semiparamétrico. En este trabajo estaremos interesados en la estimación de los parámetros de este modelo bajo una perspectiva Bayesiana. En resumen lo que se busca es aplicar componentes flexibles propios de los modelos de regresión semiparamétrica a la regresión cuantílica y estimar este modelo desde la perspectiva bayesiana.

## 3.2. Modelo

En el modelo propuesto adoptamos la formulación de regresión cuantílica sugerida por Yu y Moyeed (2001) y extendida por Kozumi y Kobayashi (2011) para estimar, desde la perspectiva bayesiana, los parámetros del modelo de regresión cuantílica. Por lo tanto, aplicando la metodología de regresión semiparamétrica, tratada en Crainiceanu et al. (2005), en el modelo de regresión cuantílica, se tiene la siguiente representación para el modelo de regresión cuantílica semiparamétrico

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

donde $(x_i, y_i)$, $i = 1, 2, \ldots, n$ son observaciones independientes de la variable respuesta asociadas a $x_i = (1, x_{i1}, \ldots, x_{ip})$, vector de $p$ covariables conocidas. Los $\epsilon_i$ son los términos de errores asumidos independientes y distribuidos con ALD donde el $\tau - ésimo$ cuantil es igual a cero, es decir $\mathrm{P}(\epsilon_i \leq 0) = \tau$. En cuanto a la función suave $m(.)$, tal como fue revisado en la sección 2.4 (ver ecuación (2.13) para más detalles), será modelada utilizando *thin-plate splines* de bajo rango, siguiendo lo propuesto por Crainiceanu et al. (2005).

Como fue presentado en la sección 2.4 podemos representar el modelo dado en (3.1) como un modelo lineal mixto de la forma

$$y_i = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b} + \epsilon_i, \tag{3.2}$$

donde

- $\epsilon_i \overset{iid}{\sim} ALD(0, \sigma, \tau)$, $\sigma > 0$, $0 < \tau < 1$,
- $\mathbf{b} \sim N(0, \sigma_b^2 \mathbf{I}_K)$, es el vector de parámetros de efectos aleatorios. $\sigma_b > 0$ e $\mathbf{I}_K$ es la matriz identidad,
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$, es el vector de parámetros de efectos fijos,
- $\mathbf{X_i} = (1, x_i)$, es la $i-ésima$ fila de la matriz $\mathbf{X}$, la cual es la matriz de efectos fijos,
- $\mathbf{Z_i}$ es la $i-ésima$ fila de la matriz $\mathbf{Z}$, la cual es la matriz de coeficientes aleatorios dada por $\mathbf{Z} = \mathbf{Z}_K \boldsymbol{\Omega}_K^{-1/2}$, donde

$$\mathbf{Z}_K = \begin{bmatrix} |x_1 - \kappa_1|^3 & |x_1 - \kappa_2|^3 & \cdots & |x_1 - \kappa_K|^3 \\ |x_2 - \kappa_1|^3 & |x_2 - \kappa_2|^3 & \cdots & |x_2 - \kappa_K|^3 \\ \vdots & \vdots & \ddots & \vdots \\ |x_n - \kappa_1|^3 & |x_n - \kappa_2|^3 & \cdots & |x_n - \kappa_K|^3 \end{bmatrix},$$

$$\boldsymbol{\Omega}_K = \begin{bmatrix} 0 & |\kappa_1 - \kappa_2|^3 & \cdots & |\kappa_1 - \kappa_K|^3 \\ |\kappa_2 - \kappa_1|^3 & 0 & \cdots & |\kappa_2 - \kappa_K|^3 \\ \vdots & \vdots & \ddots & \vdots \\ |\kappa_K - \kappa_1|^3 & |\kappa_K - \kappa_2|^3 & \cdots & 0 \end{bmatrix},$$

y, $\kappa_1 < \kappa_2 < \ldots < \kappa_K$ son nodos fijos.

## 3.3. Representación jerárquica del modelo de regresión cuantílica semiparamétrico

Utilizando la proposición de la sección 2.2 del capítulo 2, el modelo dado en (3.2) puede ser reescrito como

$$y_i = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b} + \theta_1 v_i + \theta_2 \sigma^{\frac{1}{2}} v_i^{\frac{1}{2}} w_i \quad , \tag{3.3}$$

donde $\theta_1 = \dfrac{1-2\tau}{\tau(\tau-1)}$, $\theta_2 = \sqrt{\dfrac{2}{\tau(1-\tau)}}$ son dos escalares que dependen de $\tau$. Además los $v_i$ y $w_i$ son independientes entre sí y $v_i \sim exp(\frac{1}{\sigma})$ y $w_i \sim N(0,1)$.

A su vez (3.3) se puede expresar a través de la siguiente representación jerárquica

$$y_i \mid v_i, \mathbf{b} \sim N(\mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b} + \theta_1 v_i,\ \theta_2^2 \sigma v_i) \tag{3.4}$$

Esta representación a través de la mixtura de la normal estándar y exponencial estándar, facilita el planteamiento de algoritmos Gibbs sampling en el contexto de la regresión cuantílica Bayesiana.

## 3.4. Inferencia Bayesiana

Recalcamos que el empleo de la expresión de mixtura de la ALD brinda gran conveniencia para realizar inferencia Bayesiana sobre el modelo de regresión cuantílica, permitiendo

reescribirlo como un modelo de regresión normal condicionalmente con pesos y variables latentes. Como consecuencia, los esquemas de inferencia Bayesiana desarrollados para modelos de regresión normal pueden entonces ser transferidos a la regresión cuantílica.

### 3.4.1. Verosimilitud aumentada

De acuerdo con (3.4), la distribución condicional de $y_i$ dado $v_i$ es normal con media $\mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b} + \theta_1 v_i$ y varianza $\theta_2^2 \sigma v_i$, entonces dadas las observaciones $\mathbf{y} = (y_1, \ldots, y_n)$ la función de verosimilitud aumentada es

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}, \mathbf{b}, \sigma, \sigma_b^2 \mid \mathbf{y}, \mathbf{v}) &= \prod_{i=1}^{n} f(y_i \mid v_i, \mathbf{b}).f(v_i).f(\mathbf{b}) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\theta_2 \sigma^{\frac{1}{2}} v_i^{\frac{1}{2}}} exp\left\{ -\frac{1}{2}\left( \frac{y_i - \mathbf{X_i}\boldsymbol{\beta} - \mathbf{Z_i}\mathbf{b} - \theta_1 v_i}{\theta_2 \sigma^{\frac{1}{2}} v_i^{\frac{1}{2}}} \right)^2 \right\} \frac{1}{\sigma} exp\left\{ -\frac{1}{\sigma} v_i \right\} \\
&\quad \frac{1}{(2\pi)^{\frac{K}{2}} \left|\sigma_b^2 \mathbf{I}_K\right|^{\frac{1}{2}}} exp\left\{ -\frac{1}{2}\mathbf{b}^\top (\sigma_b^2 \mathbf{I}_K)^{-1} \mathbf{b} \right\} \\
&= \left( \frac{\theta_2^{-1}}{\sqrt{2\pi}} \right)^n \sigma^{-\frac{3}{2}n} \left( \prod_{i=1}^{n} v_i^{-\frac{1}{2}} \right) exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \mathbf{X_i}\boldsymbol{\beta} - \mathbf{Z_i}\mathbf{b} - \theta_1 v_i)^2}{\theta_2^2 \sigma v_i} \right. \\
&\quad \left. -\frac{1}{\sigma} \sum_{i=1}^{n} v_i \right\} \frac{1}{(2\pi)^{\frac{K}{2}} \left(\sigma_b^2\right)^{\frac{K}{2}}} exp\left\{ -\frac{1}{2} \frac{\mathbf{b}^\top \mathbf{I}_K \mathbf{b}}{\sigma_b^2} \right\},
\end{aligned} \tag{3.5}$$

donde $\mathbf{v} = (v_1, \ldots, v_n)$ son las variables latentes definidas en (3.3).

### 3.4.2. Distribuciones condicionales completas

Basándonos en la verosimilitud aumentada (3.5), y siguiendo a Kozumi y Kobayashi (2011) y los resultados de Zevallos (2012) asumiremos una distribución Normal Multivariada ($\mathcal{N}$) como distribución a priori independiente para el vector de parámetros $\boldsymbol{\beta}$ y distribuciones Gamma Inversa (GI) como distribuciones a priori independientes para los parámetros $\sigma$ y $\sigma_b^2$:

$$\boldsymbol{\beta} \sim \mathcal{N}(\beta_0, B_0) \tag{3.6}$$

$$\sigma \sim GI\left( \frac{n_0}{2}, \frac{s_0}{2} \right) \tag{3.7}$$

$$\sigma_b^2 \sim GI\left( \frac{n_b}{2}, \frac{s_b}{2} \right) \tag{3.8}$$

En el caso de (3.6), $\beta_0$ (media a priori de $\boldsymbol{\beta}$) es un vector columna de $(p+1)$ filas y $B_0$ (covarianza a priori de $\boldsymbol{\beta}$) es una matriz de varianzas-covarianzas de dimensiones $(p+1) \times (p+1)$. Con respecto a (3.7) y (3.8), los hiperparámetros $n_0$, $s_0$, $n_b$ y $s_b$ son escalares.

#### 3.4.2.1. Condicional Completa de $\boldsymbol{\beta}$

Para hallar la distribución a posteriori de $\boldsymbol{\beta}$ trabajamos con las expresiones de la verosimilitud aumentada dada en (3.5) y la correspondiente distribución a priori definida en (3.6) de manera de expresarlas proporcionalmente en función de los términos relevantes, es decir, donde aparezca $\boldsymbol{\beta}$.

Trabajando con la verosimilitud aumentada obtenemos la expresión (3.9)

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}, \mathbf{b}, \sigma, \sigma_b^2 \mid \mathbf{y}, \mathbf{v}) &\propto exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \mathbf{X_i}\boldsymbol{\beta} - \mathbf{Z_i}\mathbf{b} - \theta_1 v_i)^2}{\theta_2^2 \sigma v_i}\right\} \\
&\propto exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{-2y_i\mathbf{X_i}\boldsymbol{\beta} + 2\theta_1 v_i \mathbf{X_i}\boldsymbol{\beta} + 2\mathbf{Z_i}\mathbf{b}\mathbf{X_i}\boldsymbol{\beta} + \mathbf{X_i}\boldsymbol{\beta}\mathbf{X_i}\boldsymbol{\beta}}{\theta_2^2 \sigma v_i}\right]\right\} \\
&\propto exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[\frac{\mathbf{X_i}\boldsymbol{\beta}\mathbf{X_i}\boldsymbol{\beta}}{\theta_2^2 \sigma v_i} - \frac{2(y_i - \mathbf{Z_i}\mathbf{b} - \theta_1 v_i)\mathbf{X_i}\boldsymbol{\beta}}{\theta_2^2 \sigma v_i}\right]\right\} \\
&\propto exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}\frac{\boldsymbol{\beta}^\top \mathbf{X_i}^\top \mathbf{X_i}\boldsymbol{\beta}}{\theta_2^2 \sigma v_i} - \sum_{i=1}^{n}\frac{2(y_i - \mathbf{Z_i}\mathbf{b} - \theta_1 v_i)\mathbf{X_i}\boldsymbol{\beta}}{\theta_2^2 \sigma v_i}\right]\right\} \\
&\propto exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}^\top \sum_{i=1}^{n}\frac{\mathbf{X_i}^\top \mathbf{X_i}}{\theta_2^2 \sigma v_i}\boldsymbol{\beta} - 2\sum_{i=1}^{n}\frac{(y_i - \mathbf{Z_i}\mathbf{b} - \theta_1 v_i)}{\theta_2^2 \sigma v_i}\mathbf{X_i}\boldsymbol{\beta}\right]\right\}
\end{aligned} \tag{3.9}$$

Por otro lado, trabajando con la distribución a priori Normal Multivariada de (3.6) obtenemos la expresión (3.10)

$$\begin{aligned}
f(\boldsymbol{\beta}) &= \frac{exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \beta_\mathbf{0})^\top B_0^{-1}(\boldsymbol{\beta} - \beta_\mathbf{0})\right\}}{(2\pi)^{\frac{n}{2}} B_0^{\frac{1}{2}}} \\
&\propto exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \beta_\mathbf{0})^\top B_0^{-1}(\boldsymbol{\beta} - \beta_\mathbf{0})\right\} \\
&\propto exp\left\{-\frac{1}{2}(\boldsymbol{\beta}^\top B_0^{-1}\boldsymbol{\beta} - \boldsymbol{\beta}^\top B_0^{-1}\beta_\mathbf{0} - \beta_\mathbf{0}{}^\top B_0^{-1}\boldsymbol{\beta})\right\}
\end{aligned} \tag{3.10}$$

Como $B_0$ es una matriz simétrica y el término $\beta_\mathbf{0}{}^\top B_0^{-1}\boldsymbol{\beta}$ es un escalar, $\boldsymbol{\beta}^\top B_0^{-1}\beta_\mathbf{0} = \beta_\mathbf{0}{}^\top B_0^{-1}\boldsymbol{\beta}$ y por lo tanto la expresión (3.10) puede quedar como

$$f(\boldsymbol{\beta}) \propto exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}^\top B_0^{-1}\boldsymbol{\beta} - 2\beta_\mathbf{0}{}^\top B_0^{-1}\boldsymbol{\beta}\right)\right\} \tag{3.11}$$

Ahora, multiplicando (3.9) y (3.11) se obtiene la distribución a posteriori del vector de parámetros $\boldsymbol{\beta}$

$$\begin{aligned}
f(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{v}, \sigma, \sigma_b^2, \mathbf{b}) &\propto exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}^\top \sum_{i=1}^{n}\frac{\mathbf{X_i}^\top \mathbf{X_i}}{\theta_2^2 \sigma v_i}\boldsymbol{\beta} - 2\sum_{i=1}^{n}\frac{(y_i - \mathbf{Z_i}\mathbf{b} - \theta_1 v_i)}{\theta_2^2 \sigma v_i}\mathbf{X_i}\boldsymbol{\beta} + \right.\right. \\
&\quad \left.\left. \boldsymbol{\beta}^\top B_0^{-1}\boldsymbol{\beta} - 2\beta_\mathbf{0}{}^\top B_0^{-1}\boldsymbol{\beta}\right]\right\} \\
&\propto exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta}^\top\left[\sum_{i=1}^{n}\frac{\mathbf{X_i}^\top \mathbf{X_i}}{\theta_2^2 \sigma v_i} + B_0^{-1}\right]\boldsymbol{\beta} - 2\left[\sum_{i=1}^{n}\frac{(y_i - \mathbf{Z_i}\mathbf{b} - \theta_1 v_i)}{\theta_2^2 \sigma v_i}\mathbf{X_i} + \right.\right.\right. \\
&\quad \left.\left.\left. + \beta_\mathbf{0}{}^\top B_0^{-1}\right]\boldsymbol{\beta}\right]\right\}
\end{aligned} \tag{3.12}$$

Como se aprecia, la distribución a posteriori del vector de parámetros $\boldsymbol{\beta}$ es una normal

multivariada

$$\boldsymbol{\beta} \mid \mathbf{b}, y_i, v_i, \sigma, \sigma_b^2 \sim \mathcal{N}(\mu_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}) \tag{3.13}$$

con parámetros de escala y localización dados, respectivamente en (3.14) y (3.15)

$$\Sigma_{\boldsymbol{\beta}} = \left( \sum_{i=1}^{n} \frac{\mathbf{X_i}^\top \mathbf{X_i}}{\theta_2^2 \sigma v_i} + B_0^{-1} \right)^{-1} \tag{3.14}$$

$$\mu_{\boldsymbol{\beta}} = \left( \sum_{i=1}^{n} \frac{\mathbf{X_i}^\top \mathbf{X_i}}{\theta_2^2 \sigma v_i} + B_0^{-1} \right)^{-1} \left( \sum_{i=1}^{n} \frac{(y_i - \mathbf{Z_i b} - \theta_1 v_i)}{\theta_2^2 \sigma v_i} \mathbf{X_i}^\top + B_0^{-1} \boldsymbol{\beta_0} \right), \tag{3.15}$$

desde que la inversa en (3.14) exista, por ejemplo cuando las matrices simétricas $\sum_{i=1}^{n} \frac{\mathbf{X_i}^\top \mathbf{X_i}}{\theta_2^2 \sigma v_i}$ y $B_0$ sean definidas positivas.

#### 3.4.2.2. Condicional Completa de $v_i$

Para hallar la distribución a posteriori de $v_i$ trabajamos la expresión de la verosimilitud aumentada dada en (3.5) de manera de expresarla proporcionalmente en función de los términos relevantes, es decir, donde aparezca $v_i$. Entonces la parte relevante para cada una de las $v_i$ nos lleva a tener la expresión (3.16)

$$\begin{aligned} \mathcal{L}(v_i \mid \boldsymbol{\beta}, \mathbf{b}, \sigma, \sigma_b^2, y_i) &\propto v_i^{-\frac{1}{2}} exp\left\{ -\frac{1}{2} \frac{(y_i - \mathbf{X_i}\boldsymbol{\beta} - \mathbf{Z_i b} - \theta_1 v_i)^2}{\theta_2^2 \sigma v_i} - \frac{v_i}{\sigma} \right\} \\ &\propto v_i^{-\frac{1}{2}} exp\left\{ -\frac{1}{2} \frac{(y_i - \mathbf{X_i}\boldsymbol{\beta} - \mathbf{Z_i b})^2}{\theta_2^2 \sigma v_i} - \frac{1}{2} \frac{\theta_1^2 v_i}{\theta_2^2 \sigma} - \frac{v_i}{\sigma} \right\} \end{aligned} \tag{3.16}$$

De esta manera, por (3.16) se obtiene que la distribución condicional completa de $v_i$ es

$$\begin{aligned} f(v_i \mid \boldsymbol{\beta}, \mathbf{b}, y_i, \sigma, \sigma_b^2) &\propto v_i^{-\frac{1}{2}} exp\left\{ -\frac{1}{2} \frac{(y_i - \mathbf{X_i}\boldsymbol{\beta} - \mathbf{Z_i b})^2}{\theta_2^2 \sigma v_i} - \frac{1}{2} \frac{\theta_1^2 v_i}{\theta_2^2 \sigma} - \frac{v_i}{\sigma} \right\} \\ &\propto v_i^{-\frac{1}{2}} exp\left\{ -\frac{1}{2} \frac{(y_i - \mathbf{X_i}\boldsymbol{\beta} - \mathbf{Z_i b})^2}{\theta_2^2 \sigma} v_i^{-1} - \frac{1}{2} \left( \frac{\theta_1^2}{\theta_2^2 \sigma} + \frac{2}{\sigma} \right) v_i \right\} \end{aligned} \tag{3.17}$$

Como se aprecia, la expresión (3.17) tiene la forma funcional de una distribución Gaussiana Inversa Generalizada (Barndorff-Nielsen y Shephard, 2001)

$$f(y \mid \nu, a, b) \propto y^{\nu - 1} exp\left\{ -\frac{1}{2}(a^2 y^{-1} + b^2 y) \right\} \tag{3.18}$$

Por tanto, la distribución a posteriori de cada $v_i$ es una Gaussiana Inversa Generalizada

dada por:

$$v_i \mid \boldsymbol{\beta}, \mathbf{b}, y_i, \sigma \sim GIG\left(\frac{1}{2}, \frac{y_i - \mathbf{X_i}\boldsymbol{\beta} - \mathbf{Z_i}\mathbf{b}}{\theta_2\sqrt{\sigma}}, \sqrt{\frac{\theta_1^2}{\theta_2^2\sigma} + \frac{2}{\sigma}}\right) \tag{3.19}$$

#### 3.4.2.3. Condicional Completa de $\sigma$

Para hallar la distribución a posteriori de $\sigma$ trabajamos la expresión de la verosimilitud aumentada dada en (3.5) de manera de expresarla proporcionalmente en función de los términos relevantes, es decir, donde aparezca $\sigma$. Esto nos lleva a obtener la expresión

$$\mathcal{L}(\boldsymbol{\beta}, \mathbf{b}, \sigma, \sigma_b^2 \mid \mathbf{y}, \mathbf{v}) \quad \propto \quad \sigma^{-\frac{3}{2}n} exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \mathbf{X_i}\boldsymbol{\beta} - \mathbf{Z_i}\mathbf{b} - \theta_1 v_i)^2}{\theta_2^2\sigma v_i} - \frac{1}{\sigma}\sum_{i=1}^{n} v_i\right\} \tag{3.20}$$

Por otro lado, al asumir en (3.7) una distribución a priori Gamma Inversa para $\sigma$, los términos relevantes de la distribución son

$$f(\sigma) \quad \propto \quad \sigma^{-\left(\frac{n_0}{2}+1\right)} exp\left\{-\frac{s_0}{2}\sigma^{-1}\right\} \tag{3.21}$$

Ahora multiplicando (3.20) y (3.21), obtenemos la distribución a posteriori de $\sigma$

$$\begin{aligned} f(\sigma \mid \boldsymbol{\beta}, \mathbf{y}, \mathbf{v}, \mathbf{b}, \sigma_b^2) \quad &\propto \quad \sigma^{-\frac{3}{2}n} exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \mathbf{X_i}\boldsymbol{\beta} - \mathbf{Z_i}\mathbf{b} - \theta_1 v_i)^2}{\theta_2^2\sigma v_i} - \frac{1}{\sigma}\sum_{i=1}^{n} v_i\right\}\sigma^{-\left(\frac{n_0}{2}+1\right)} \\ &\quad exp\left\{-\frac{s_0}{2}\sigma^{-1}\right\} \\ &\propto \quad \sigma^{-\left(\frac{3}{2}n+\frac{n_0}{2}+1\right)} exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \mathbf{X_i}\boldsymbol{\beta} - \mathbf{Z_i}\mathbf{b} - \theta_1 v_i)^2}{\theta_2^2\sigma v_i} - \frac{1}{\sigma}\sum_{i=1}^{n} v_i \right. \\ &\quad \left. -\frac{s_0}{2}\sigma^{-1}\right\} \\ &\propto \quad \sigma^{-\left(\frac{3}{2}n+\frac{n_0}{2}+1\right)} exp\left\{-\left(\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \mathbf{X_i}\boldsymbol{\beta} - \mathbf{Z_i}\mathbf{b} - \theta_1 v_i)^2}{\theta_2^2 v_i} + \sum_{i=1}^{n} v_i \right.\right. \\ &\quad \left.\left. + \frac{s_0}{2}\right)\sigma^{-1}\right\} \end{aligned} \tag{3.22}$$

la cual es una distribución Gamma Inversa

$$\sigma \mid \boldsymbol{\beta}, \mathbf{b}, \mathbf{y}, \mathbf{v} \sim GI\left(\frac{3n}{2} + \frac{n_0}{2}, \sum_{i=1}^{n}\frac{(y_i - \mathbf{X_i}\boldsymbol{\beta} - \mathbf{Z_i}\mathbf{b} - \theta_1 v_i)^2}{2\theta_2^2 v_i} + \sum_{i=1}^{n} v_i + \frac{s_0}{2}\right) \tag{3.23}$$

#### 3.4.2.4. Condicional Completa de b

Para hallar la distribución a posteriori del vector de parámetros $\mathbf{b}$ trabajamos la expresión de la verosimilitud aumentada dada en (3.5) de manera de expresarla proporcionalmente en función de los términos relevantes, es decir, donde aparezca $\mathbf{b}$. Entonces la parte relevante

nos lleva a tener la expresión (3.24)

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}, \mathbf{b}, \sigma, \sigma_b^2 \mid \mathbf{y}, \mathbf{v}) &\propto exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}\frac{(y_i - \mathbf{X_i}\boldsymbol{\beta} - \mathbf{Z_i}\mathbf{b} - \theta_1 v_i)^2}{\theta_2^2\sigma v_i} + \frac{\mathbf{b}^\top \mathbf{I}_K \mathbf{b}}{\sigma_b^2}\right]\right\} \\
&\propto exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}\frac{-2y_i\mathbf{Z_i}\mathbf{b} + 2\mathbf{X_i}\boldsymbol{\beta}\mathbf{Z_i}\mathbf{b} + 2\theta_1 v_i\mathbf{Z_i}\mathbf{b} + \mathbf{Z_i}\mathbf{b}\mathbf{Z_i}\mathbf{b}}{\theta_2^2\sigma v_i} + \right.\right. \\
&\quad \left.\left. + \frac{\mathbf{b}^\top \mathbf{I}_K \mathbf{b}}{\sigma_b^2}\right]\right\} \\
&\propto exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}\left[\frac{\mathbf{Z_i}\mathbf{b}\mathbf{Z_i}\mathbf{b}}{\theta_2^2\sigma v_i} - \frac{2(y_i - \mathbf{X_i}\boldsymbol{\beta} - \theta_1 v_i)\mathbf{Z_i}\mathbf{b}}{\theta_2^2\sigma v_i}\right] + \right.\right. \\
&\quad \left.\left. + \frac{\mathbf{b}^\top \mathbf{I}_K \mathbf{b}}{\sigma_b^2}\right]\right\} \\
&\propto exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}\frac{\mathbf{b}^\top \mathbf{Z_i}^\top \mathbf{Z_i}\mathbf{b}}{\theta_2^2\sigma v_i} - \sum_{i=1}^{n}\frac{2(y_i - \mathbf{X_i}\boldsymbol{\beta} - \theta_1 v_i)\mathbf{Z_i}\mathbf{b}}{\theta_2^2\sigma v_i} + \right.\right. \\
&\quad \left.\left. + \frac{\mathbf{b}^\top \mathbf{I}_K \mathbf{b}}{\sigma_b^2}\right]\right\}
\end{aligned} \tag{3.24}$$

De esta manera, por (3.24) se obtiene que la distribución condicional completa del vector de parámetros $\mathbf{b}$ es

$$\begin{aligned}
f(\mathbf{b} \mid \boldsymbol{\beta}, v_i, y_i, \sigma, \sigma_b^2) \propto exp\left\{-\frac{1}{2}\left[\mathbf{b}^\top\left[\sum_{i=1}^{n}\frac{\mathbf{Z_i}^\top \mathbf{Z_i}}{\theta_2^2\sigma v_i} + \frac{1}{\sigma_b^2}\mathbf{I}_K\right]\mathbf{b} + \right.\right. \\
\left.\left. -2\sum_{i=1}^{n}\frac{(y_i - \mathbf{X_i}\boldsymbol{\beta} - \theta_1 v_i)}{\theta_2^2\sigma v_i}\mathbf{Z_i}\mathbf{b}\right]\right\}
\end{aligned} \tag{3.25}$$

Como se aprecia en la expresión (3.25), la distribución a posteriori del vector de parámetros $\mathbf{b}$ es una normal multivariada

$$\mathbf{b} \mid \boldsymbol{\beta}, y_i, v_i, \sigma, \sigma_b^2 \sim \mathcal{N}(\mu_\mathbf{b}, \Sigma_\mathbf{b}) \tag{3.26}$$

con parámetros de escala y localización dados, respectivamente en (3.27) y (3.28)

$$\Sigma_\mathbf{b} = \left(\sum_{i=1}^{n}\frac{\mathbf{Z_i}^\top \mathbf{Z_i}}{\theta_2^2\sigma v_i} + \frac{1}{\sigma_b^2}\mathbf{I}_K\right)^{-1} \tag{3.27}$$

$$\mu_\mathbf{b} = \left(\sum_{i=1}^{n}\frac{\mathbf{Z_i}^\top \mathbf{Z_i}}{\theta_2^2\sigma v_i} + \frac{1}{\sigma_b^2}\mathbf{I}_K\right)^{-1}\left(\sum_{i=1}^{n}\frac{(y_i - \mathbf{X_i}\boldsymbol{\beta} - \theta_1 v_i)}{\theta_2^2\sigma v_i}\mathbf{Z_i}^\top\right) \tag{3.28}$$

desde que la inversa en (3.27) exista, por ejemplo cuando las matrices simétricas $\sum_{i=1}^{n}\frac{\mathbf{Z_i}^\top \mathbf{Z_i}}{\theta_2^2\sigma v_i}$

sean definidas positivas.

#### 3.4.2.5. Condicional Completa de $\sigma_b^2$

Para hallar la distribución a posteriori de $\sigma_b^2$ trabajamos la expresión de la verosimilitud aumentada dada en (3.5) de manera de expresarla proporcionalmente en función de los términos relevantes, es decir, donde aparezca $\sigma_b^2$. Esto nos lleva a obtener la expresión

$$\mathcal{L}(\boldsymbol{\beta}, \mathbf{b}, \sigma, \sigma_b^2 \mid \mathbf{y}, \mathbf{v}) \quad \propto \quad \frac{1}{(\sigma_b^2)^{\frac{K}{2}}} exp\left\{-\frac{1}{2}\frac{\mathbf{b}^\top \mathbf{I}_K \mathbf{b}}{\sigma_b^2}\right\} \tag{3.29}$$

Por otro lado, al asumir en (3.8) una distribución a priori Gamma Inversa para $\sigma_b^2$, los términos relevantes de la distribución son

$$f(\sigma_b^2) \quad \propto \quad (\sigma_b^2)^{-\left(\frac{n_b}{2}+1\right)} exp\left\{-\frac{s_b}{2}(\sigma_b^2)^{-1}\right\} \tag{3.30}$$

Ahora multiplicando (3.29) y (3.30), obtenemos la distribución a posteriori de $\sigma_b^2$

$$\begin{aligned} f(\sigma_b^2 \mid \boldsymbol{\beta}, \mathbf{y}, \mathbf{v}, \mathbf{b}, \sigma) \quad &\propto \quad \frac{(\sigma_b^2)^{-\left(\frac{n_b}{2}+1\right)}}{(\sigma_b^2)^{\frac{K}{2}}} exp\left\{-\frac{1}{2}\frac{\mathbf{b}^\top \mathbf{I}_K \mathbf{b}}{\sigma_b^2}\right\} exp\left\{-\frac{s_b}{2}(\sigma_b^2)^{-1}\right\} \\ &\propto \quad (\sigma_b^2)^{-\left(\frac{K}{2}+\frac{n_b}{2}+1\right)} exp\left\{-\frac{1}{2}\frac{\mathbf{b}^\top \mathbf{b}}{\sigma_b^2} - \frac{s_b}{2}(\sigma_b^2)^{-1}\right\} \\ &\propto \quad (\sigma_b^2)^{-\left(\frac{K}{2}+\frac{n_b}{2}+1\right)} exp\left\{-\left(\frac{1}{2}\mathbf{b}^\top \mathbf{b} + \frac{s_b}{2}\right)(\sigma_b^2)^{-1}\right\} \end{aligned} \tag{3.31}$$

la cual es una distribución Gamma Inversa

$$\sigma_b^2 \mid \boldsymbol{\beta}, \mathbf{b}, \mathbf{y}, \mathbf{v} \sim GI\left(\frac{K}{2} + \frac{n_b}{2}, \frac{1}{2}\mathbf{b}^\top \mathbf{b} + \frac{s_b}{2}\right) \tag{3.32}$$

## 3.5. Criterio de comparación

La medida de bondad de ajuste principal que emplearemos en este contexto de la inferencia Bayesiana es el *Deviance Information Criterion*(DIC), Spiegelhalter et al. (2002). Ella se basa en el *deviance* (D), el cual, para este modelo, se calcula como sigue:

$$D(\boldsymbol{\beta}, \mathbf{b}, \sigma \mid \mathbf{y}) \quad = \quad -2log(\mathcal{L}(\boldsymbol{\beta}, \mathbf{b}, \sigma \mid \mathbf{y})) \tag{3.33}$$

Como se aprecia el *deviance* (D) depende de la función del logaritmo de la verosimilitud del modelo propuesto en (3.2). Entonces asumiendo que $y_i \overset{iid}{\sim} ALD(\mu, \sigma, \tau)$, la verosimilitud

para $n$ observaciones independientes se define como sigue:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, \mathbf{b}, \sigma \mid \mathbf{y}) &= \prod_{i=1}^{n} f(y_i \mid \mathbf{b}).f(\mathbf{b}) \\ &= \frac{\tau^n(1-\tau)^n}{\sigma^n}.exp\left\{-\frac{1}{\sigma}\sum_{i=1}^{n}\rho_\tau\left(y_i - \mathbf{X_i}\boldsymbol{\beta} - \mathbf{Z_i}\mathbf{b}\right)\right\}. \\ &\quad .\frac{1}{(2\pi)^{\frac{K}{2}}\left(\sigma_b^2\right)^{\frac{K}{2}}}.exp\left\{-\frac{1}{2}\frac{\mathbf{b}^\top \mathbf{I}_K \mathbf{b}}{\sigma_b^2}\right\}, \end{aligned} \tag{3.34}$$

por lo tanto, la función de logaritmo de la verosimilitud es:

$$\begin{aligned} l = log(\mathcal{L}(\boldsymbol{\beta}, \mathbf{b}, \sigma \mid \mathbf{y})) &= n\ln(\tau) + n\ln(1-\tau) - n\ln(\sigma) - \frac{1}{\sigma}\sum_{i=1}^{n}\rho_\tau(y_i - \mu_i) - \frac{K}{2}log(2\pi) \\ &\quad -Klog\sigma_b - \frac{1}{2}\frac{\mathbf{b}^\top \mathbf{I}_K \mathbf{b}}{\sigma_b^2} \end{aligned}$$

Entonces el *deviance* (D) para este modelo es:

$$\begin{aligned} D(\beta, \sigma \mid \mathbf{y}) &= -2log(\mathcal{L}(\mu \mid \mathbf{y})) \\ &= -2(n\ln(\tau) + n\ln(1-\tau) - n\ln(\sigma) - \frac{1}{\sigma}\sum_{i=1}^{n}\rho_\tau(y_i - \mu_i) - \frac{K}{2}log(2\pi) \\ &\quad -Klog\sigma_b - \frac{1}{2}\frac{\mathbf{b}^\top \mathbf{I}_K \mathbf{b}}{\sigma_b^2}) \end{aligned} \tag{3.35}$$

donde:

$$\mu_i = m(x_i) = \mathbf{X_i}\widehat{\boldsymbol{\beta}} + \mathbf{Z_i}\widehat{\mathbf{b}}$$

y $\widehat{\beta}$, $\widehat{\mathbf{b}}$ son las estimaciones realizadas para el vector de parámetros de efectos fijos y aleatorios respectivamente.

Luego el *DIC* es igual a la suma del esperado de $D$ al evaluarse en la distribución a posteriori y del número de parámetros efectivo $d_e$. De acuerdo con Congdon (2010):

$$DIC = E_{\beta,b,\sigma|\mathbf{y}}[D] + d_e \tag{3.36}$$

en el cual $E_{\beta,b,\sigma|\mathbf{y}}[D]$, se estima por

$$\overline{D} = \frac{1}{M}\sum_{j=1}^{M} D(\beta^{(j)}, b^{(j)}, \sigma^{(j)}) \tag{3.37}$$

donde $(\beta^{(j)}, b^{(j)}, \sigma^{(j)})$ es la $j$-ésima simulación de la distribución a posteriori y M es el número de simulaciones. Luego $d_e$ se estima por:

$$\widehat{d_e} = \overline{D} - \widehat{D} \tag{3.38}$$

donde $\widehat{D}$ es la estimación del *plug-in deviance*(PID) que se emplea en el enfoque bayesiano y

utiliza la expresión (3.33) evaluada en la media a posteriori ($\overline{\beta}$, $\overline{\mathbf{b}}$ y $\overline{\sigma}$). Siguiendo a Congdon (2010), $\widehat{D}$ es dado por:

$$\widehat{D} = D(\overline{\beta}, \overline{\mathbf{b}}, \overline{\sigma})$$

## 3.6. Implementación

En el presente trabajo de investigación se emplearon como referencia, los programas que se presentan en Crainiceanu et al. (2005), los cuales están escritos en el lenguaje del software WinBUGS (Spiegelhalter et al., 2007) y del software R (R Development Core Team, 2012). A dichos programas se les realizó las modificaciones pertinentes para poder adaptarlos a las bases de datos que nos permitan implementar el respectivo análisis bayesiano y así poder mostrar las aplicaciones del modelo de regresión cuantílica semiparamétrico propuesto en este trabajo tal como aparece en el capítulo 5. La implementación de los programas en WinBUGS (Spiegelhalter et al., 2007) se realizará en el software R (R Development Core Team, 2012) a través del paquete **R2WinBUGS** (Sturtz et al., 2005). Algunos de los programas adaptados y empleados en las aplicaciones que aparecen en capítulo 5, se pueden encontrar por ejemplo en los apéndices: A.3, A.4, A.5 y A.6.

# Capítulo 4

# Estudio de Simulación

En el presente capítulo, se desarrollan tres estudios de simulación. El primero está relacionado a la recuperación de parámetros (a través de la recuperación de una función no lineal $m(x)$ simulada) para comprobar si el modelo de regresión cuantílica semiparamétrico propuesto se ajusta o "recupera" adecuadamente a dicha función $m(x)$. El segundo estudio se relaciona con evaluar la sensibilidad del ajuste del modelo propuesto al considerar diferente número de nodos, empleados en los thin-plate splines de bajo rango para modelar la función $m(x)$. El tercer estudio se relaciona con evaluar la sensibilidad del ajuste del modelo propuesto al considerar diferentes valores de $\sigma^2$ (es decir, diferentes valores de variabilidad en la variable respuesta). Para esos escenarios se distinguen diferentes número de nodos empleados en los thin-plate splines de bajo rango para modelar la función $m(x)$.

## 4.1. Algoritmo para simular los datos

Se simulará un conjunto de datos del modelo:

$$y_i = m(x_i) + \epsilon_i, \tag{4.1}$$

donde $m(x_i) = 0.3x_i + 0.5\sin(2x_i) + \dfrac{1.1}{1+{x_i}^2}$ y $\epsilon_i \overset{iid}{\sim} ALD(0, \sigma, \tau)$. La elección de $m(x_i)$ se debe a que se conoce de antemano que su forma difícilmente pueda ser captada por un modelo de regresión cuantílica lineal. En la figura 4.1 se muestra el modelo no lineal de generación de datos. La resultante función cuantil será: $Q(\tau \mid x_i) = 0.3x_i + 0.5\sin(2x_i) + \dfrac{1.1}{1+{x_i}^2}$.

Para simular un conjunto de datos de este modelo no lineal seguimos el siguiente procedimiento:

- Definir los valores de $\sigma$ y $\tau$, así como el tamaño de muestra $n$.
- Generar $n$ valores de $x_i$ de una distribución uniforme $[-3, 3]$.
- Generar $n$ valores de $\epsilon_i \overset{iid}{\sim} ALD(0, \sigma, \tau)$ utilizando la proposición de la sección 2.2 del capítulo 2.
- Calcular $n$ valores de $y_i$ utilizando (4.1). .

## 4.2. Criterios para la comparación de estimadores

En cada estudio de simulación, se comparan los resultados a través del promedio del error absoluto, MAE (por sus siglas en inglés) y la raíz del error cuadrático medio, RMSE (por sus

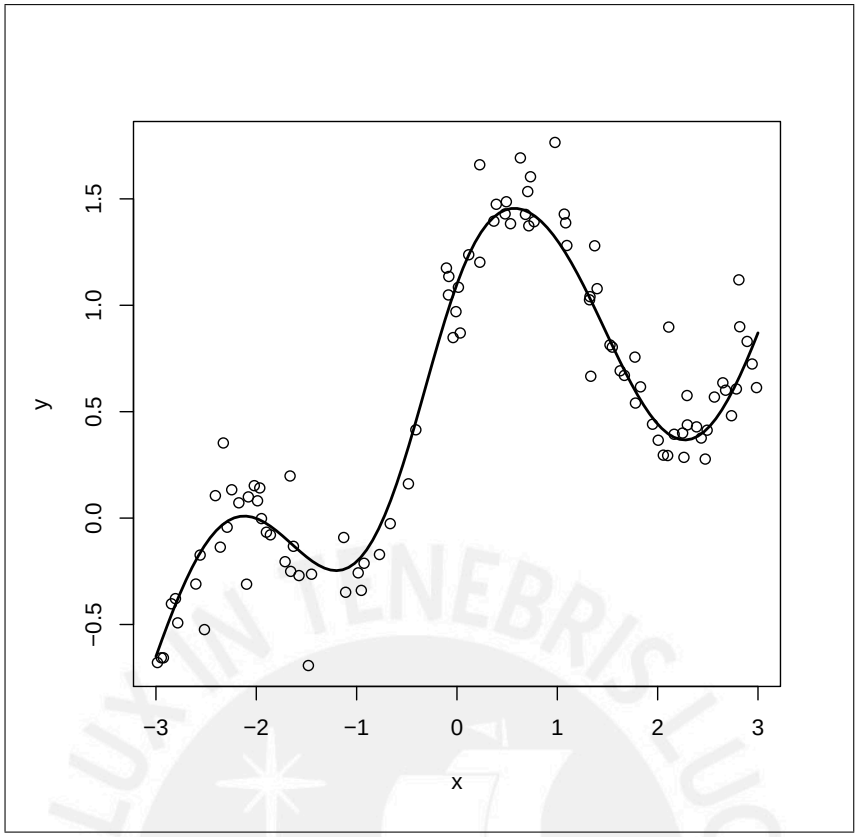


Figura 4.1: Conjunto de datos simulados de la función: $y_i = 0.3x_i + 0.5\sin(2x_i) + \dfrac{1.1}{1+{x_i}^2} + \epsilon_i$

siglas en inglés), ver por ejemplo Willmott (2005).

$$MAE = \sum_{i=1}^{n} \frac{\left|Q(\tau \mid x_i) - \hat{Q}(\tau \mid x_i)\right|}{n} \tag{4.2}$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{\left(Q(\tau \mid x_i) - \hat{Q}(\tau \mid x_i)\right)^2}{n}} \tag{4.3}$$

donde: $Q(\tau \mid x_i)$, es la función cuantil simulada, $\hat{Q}(\tau \mid x_i)$, es la función cuantil estimada y $n$ es el número de simulaciones.

## 4.3. Método de estimación de los parámetros

La estimación de los parámetros se realiza desde la perspectiva bayesiana mediante MCMC implementada en el software WinBUGS. Se calculará la media estimada de la distribución a posteriori de los parámetros a estimar.

## 4.4. Estudio de simulación 1

### 4.4.1. Objetivo

Simular un conjunto de datos del modelo:

$$y_i = m(x_i) + \epsilon_i, \qquad \text{donde:}$$

$m(x_i) = 0.3x_i + 0.5\sin(2x_i) + \frac{1.1}{1+{x_i}^2}$ y $\epsilon_i \overset{iid}{\sim} ALD(0, \sigma, \tau)$. Los datos simulados de este modelo no lineal se emplearán para comparar los estimadores calculados a partir de un modelo de regresión cuantílica lineal y del modelo de regresión cuantílica semiparamétrico propuesto en este trabajo, de manera de observar cuán bien se ajusta (o recupera) este último, al modelo no lineal simulado.

El programa para realizar esta simulación se muestra en el anexo (A.1).

### 4.4.2. Consideraciones para el estudio de simulación

Para la simulación se considera $\sigma = 1.5$, $\tau = 0.50$ y $n = 100$. Para el modelo cuantílico semiparamétrico se consideran 20 nodos. Se realizan 20 réplicas. La estimación de ambos modelos (cuantílico lineal y semiparamétrico) se hace a través de inferencia bayesiana vía MCMC con 500000 iteraciones con un periodo de burn-in de 50000 y saltos de 50. En estas condiciones se observó que la convergencia de la cadena es adecuada.

### 4.4.3. Resultados

En el cuadro 4.1 se muestran los resultados del estudio de simulación 1.

| **Parámetro a recuperar** | Medida | Modelo de Regresión Cuantílica lineal | Modelo de Regresión Cuantílica Semiparamétrico |
|---|---|---|---|
| "función m(x)" | MAE | 0.3587 | 0.0409 |
| | RMSE | 0.4706 | 0.0514 |
| | DIC | 205.9641 | 70.6776 |
| | Tiempo (seg.) | 2389.39 | 13311.47 |

Cuadro 4.1: Resultados del estudio de simulación 1. Se presenta el promedio del MAE, RMSE, DIC y tiempo en segundos que tomó la estimación por MCMC para los modelos de regresión cuantílica lineal y semiparamétrico.

La figura 4.2 muestra el mejor ajuste del modelo de regresión cuantílica semiparamétrico sobre los datos simulados de la función: $y_i = 0.3x_i + 0.5\sin(2x_i) + \frac{1.1}{1+{x_i}^2} + \epsilon_i$.

La conclusión de este estudio de simulación, según los resultados mostrados, es que el modelo de regresión cuantílica semiparamétrico "recupera" o ajusta mejor los datos, considerando los criterios mostrados.

## 4.5. Estudio de simulación 2

### 4.5.1. Objetivo

Simular un conjunto de datos del modelo:

$$y_i = m(x_i) + \epsilon_i, \qquad \text{donde:}$$

$m(x_i) = 0.3x_i + 0.5\sin(2x_i) + \frac{1.1}{1+{x_i}^2}$ y $\epsilon_i \overset{iid}{\sim} ALD(0, \sigma, \tau)$. Los datos simulados de este modelo no lineal se emplearán para comparar los estimadores calculados a partir del modelo de regresión cuantílica semiparamétrico propuesto en este trabajo de manera de observar cuán bien se ajusta (o recupera) al modelo no lineal simulado según el número de nodos empleados en los thin-plate splines de bajo rango para modelar la función $m(x)$.

El programa para realizar esta simulación se muestra en el anexo (A.2).

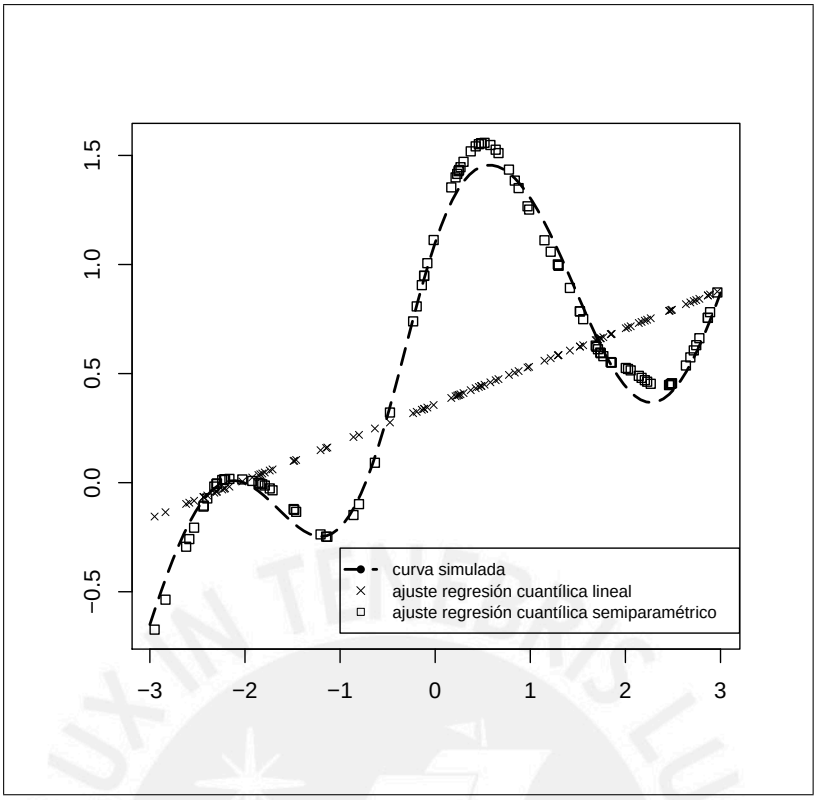


Figura 4.2: Ajuste del modelo de regresión cuantílica lineal y semiparamétrico sobre un conjunto de datos simulados de la función: $y_i = 0.3x_i + 0.5\sin(2x_i) + \dfrac{1.1}{1+{x_i}^2} + \epsilon_i$

### 4.5.2. Consideraciones para el estudio de simulación

Para la simulación se considera $\sigma = 1.5$, $\tau = 0.50$ y distinto tamaños de muestra: $n = 50, 100, 200, 300$. De acuerdo con Crainiceanu et al. (2005), se emplean thin-plate splines de bajo rango para modelar la función $m(x)$. Se prueba diferentes números de nodos (5, 10, 15, 20 y 25) en la modelación. Se realizan 20 réplicas. La estimación del modelo de regresión cuantílica semiparamétrico se realiza mediante inferencia bayesiana vía MCMC con 500000 iteraciones con un periodo de burn-in de 50000 y saltos de 50. En estas condiciones se observó que la convergencia de la cadena es adecuada (anexo B.2).

### 4.5.3. Resultados

En las figuras 4.3 y 4.4 se muestran los resultados resumidos mediante el promedio del MAE, RMSE y DIC para diferentes tamaños de muestra y número de nodos empleado.

La conclusión de este estudio de simulación, según los resultados mostrados, es que el modelo de regresión cuantílica semiparamétrico "recupera" o ajusta mejor los datos, empleando un número de nodos igual a 15 ó 20 para tamaños de muestra pequeño y 20 nodos cuando el tamaño de la muestra se incrementa. En el anexo B.1 aparecen los resultados numéricos en detalle .

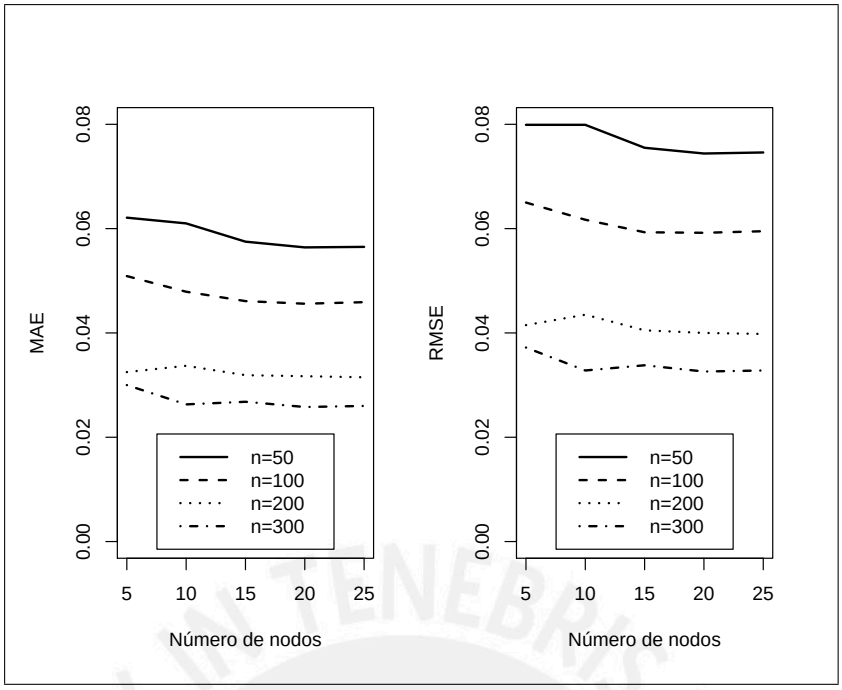


Figura 4.3: Resultados del estudio de simulación 2. Se presenta el promedio del MAE y RMSE para diferentes tamaños de muestra y número de nodos.

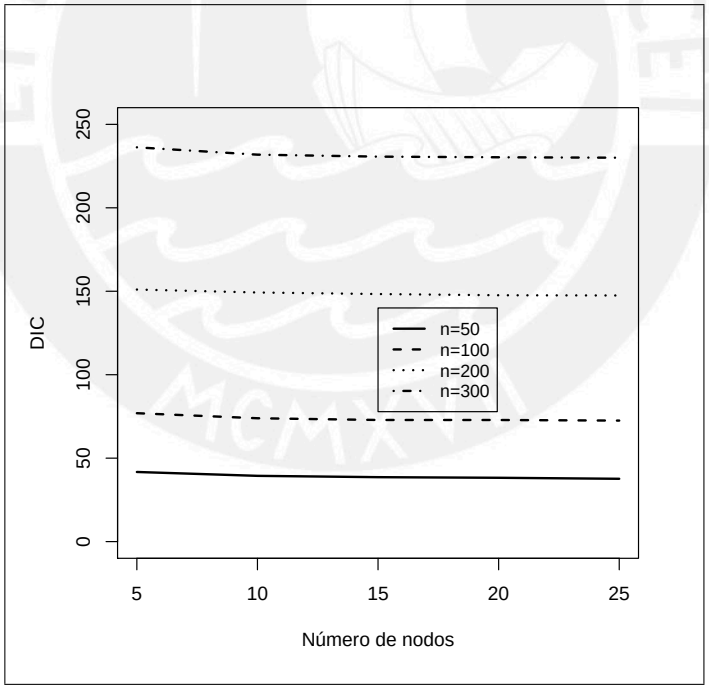


Figura 4.4: Promedio del DIC para el estudio de simulación 2 con diferentes tamaños de muestra y número de nodos.

## 4.6. Estudio de simulación 3

### 4.6.1. Objetivo

Simular un conjunto de datos del modelo:

$$y_i = m(x_i) + \epsilon_i, \qquad \text{donde:}$$

$m(x_i) = 0.3x_i + 0.5\sin(2x_i) + \dfrac{1.1}{1+{x_i}^2}$ y $\epsilon_i \overset{iid}{\sim} ALD(0, \sigma, \tau)$. Los datos simulados de este modelo no lineal se emplearán para comparar los estimadores calculados a partir del modelo de regresión cuantílica semiparamétrico propuesto en este trabajo de manera de observar cuán bien se ajusta (o recupera) al modelo no lineal simulado según el valor asignado a $\sigma^2$ (es decir, diferentes valores de variabilidad en la variable respuesta) y al número de nodos empleados en los thin-plate splines de bajo rango para modelar la función $m(x)$.

El programa para realizar esta simulación es similar al empleado en el estudio de simulación 2 que se muestra en el anexo (A.2), alterando simplemente el valor de $\sigma$ y el número de nodos.

### 4.6.2. Consideraciones para el estudio de simulación

Para la simulación se considera $\tau = 0.50$ y $n = 300$. De acuerdo con Crainiceanu et al. (2005), se emplean thin-plate splines de bajo rango para modelar la función $m(x)$. Se prueban tres diferentes valores de $\sigma$ (0.75, 1.5 y 4) y tres diferentes número de nodos (10, 15 y 20) en la modelación. Se realizan 20 réplicas. La estimación del modelo de regresión cuantílica semiparamétrico se realiza mediante inferencia bayesiana vía MCMC con 500000 iteraciones con un periodo de burn-in de 50000 y saltos de 50. En estas condiciones se observó que la convergencia de la cadena es adecuada (anexo B.4).

### 4.6.3. Resultados

En las figuras 4.5 y 4.6 se muestran los resultados resumidos mediante el promedio del MAE, RMSE y DIC, para tres diferentes valores de $\sigma$ y tres diferentes valores para el número de nodos.

La conclusión de este estudio de simulación, según los resultados mostrados, es que a medida que aumenta la variabilidad de la variable respuesta, la estimación del MAE y de la RMSE es mayor. También se aprecia que el modelo teórico es fácilmente descrito con pocos nodos para los diferentes valores de variabilidad de los datos. En el anexo B.3 aparecen los resultados numéricos en detalle.

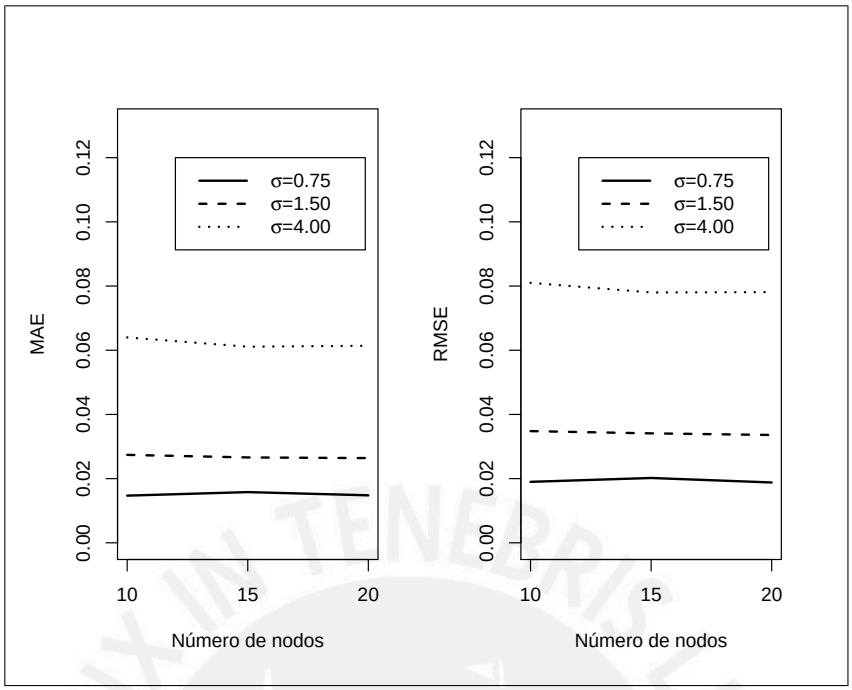


Figura 4.5: Resultados del estudio de simulación 3. Se presenta el promedio del MAE y RMSE para tres diferentes valores de $\sigma$ y tres diferentes valores para el número de nodos.

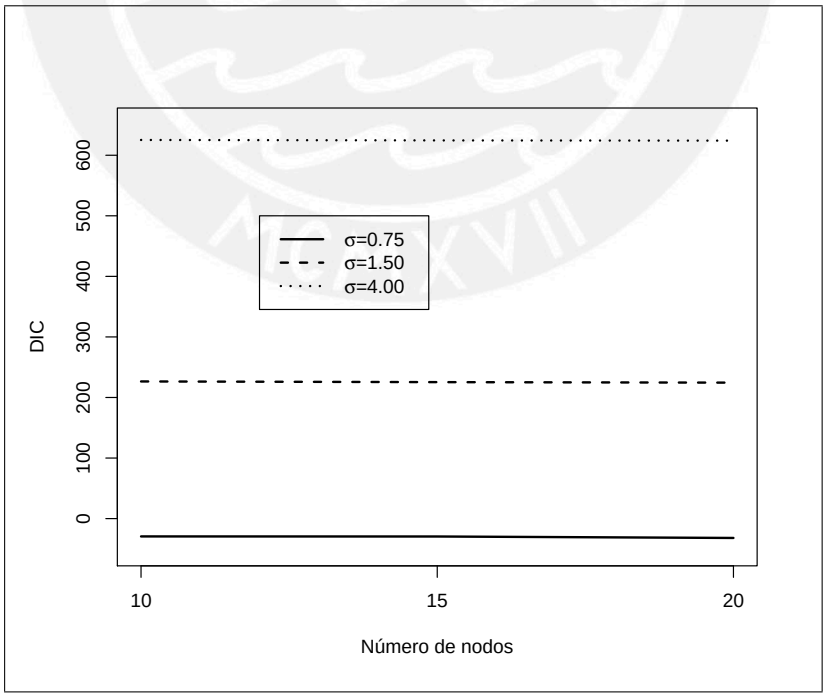


Figura 4.6: Promedio del DIC para el estudio de simulación 3 con tres diferentes valores de $\sigma$ y número de nodos.

# Capítulo 5

# Aplicación del Modelo de Regresión Cuantílica Semiparamétrico

En este capítulo se muestran dos aplicaciones del modelo de regresión cuantílica semiparamétrico propuesto. La primera aplicación es sobre el conjunto de datos Canadian age-income y la segunda sobre los datos empleados en un estudio sobre la aplicación de la ecuación de Mincer para Lima Metropolitana.

Se implementó la inferencia bayesiana del modelo de regresión cuantílica semiparamétrico utilizando el software WinBUGS (Spiegelhalter et al., 2007). También se utilizó el paquete **R2WinBUGS** (Sturtz et al., 2005) del software R (R Development Core Team, 2012) para analizar los resultados.

## 5.1. Aplicación 1: Conjunto de datos Canadian age-income

Este conjunto de datos corresponde a una muestra de $n$=205 trabajadores canadienses. Estos datos fueron empleados en Ullah (1985), y su fuente es un extracto del Censo de Canadá de 1971.

Este conjunto de datos ha sido utilizado en el ámbito de la regresión semiparamétrica en (Ruppert et al., 2003) y en (Crainiceanu et al., 2005). También forma parte del paquete *semipar* del software R.

El conjunto de datos Canadian age-income está conformado por las variables: logaritmo del ingreso (variable dependiente) y la edad (variable explicativa) de cada trabajador. Algunos estadísticos descriptivos de ambas variables se muestran en el cuadro 5.1

| | Edad | Logaritmo del ingreso |
|---|---|---|
| **Estadístico** | | |
| Mínimo | 21.00 | 11.16 |
| Primer cuartil | 27.00 | 13.30 |
| Mediana | 38.00 | 13.61 |
| Media | 38.85 | 13.49 |
| Tercer cuartil | 49.00 | 13.87 |
| Máximo | 65.00 | 15.06 |
| Desviación estándar | 12.23 | 0.64 |

Cuadro 5.1: Estadísticos descriptivos de la base Canadian age-income

Con la finalidad de entender más la relación entre ambas variables, se presenta en la figura 5.1, un diagrama de dispersión de la edad versus el logaritmo del ingreso de los 205

trabajadores. Como se observa la relación entre ambas variables es no-lineal.

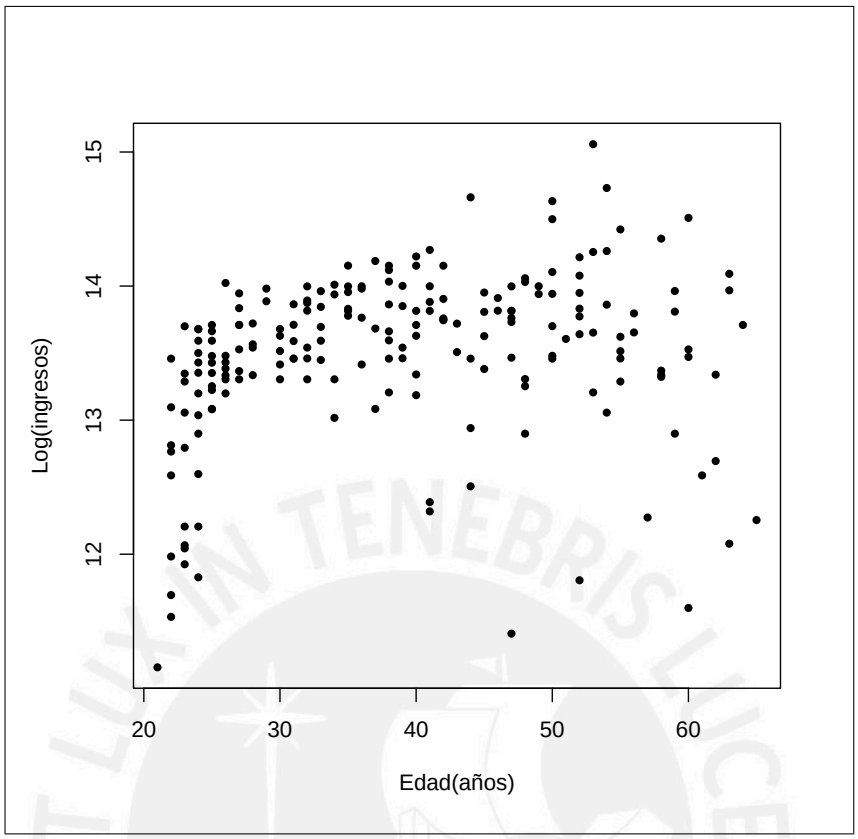


Figura 5.1: Diagrama de dispersión de la Edad vs. Logaritmo de los ingresos de 205 trabajadores canadienses.

### 5.1.1. Modelo y prioris

Siguiendo el modelo dado en (3.1):

$$y_i = m(x_i) + \epsilon_i, \qquad i = 1, 2, \ldots, n \tag{5.1}$$

modelamos un determinado cuantil de la variable respuesta (logaritmo de los ingresos) como una función de la edad del trabajador. Además, como se explicó en la sección 3.2 podemos representar el modelo dado en (5.1) como un modelo lineal mixto de la forma

$$y_i = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b} + \epsilon_i, \qquad i = 1, 2, \ldots, n \tag{5.2}$$

donde

- $y_i$=logaritmo del ingreso del trabajador $i$
- $\mathbf{X_i} = (1, x_i)$, es la $i-ésima$ fila de la matriz $\mathbf{X}$, que en este caso es la matriz de datos de edades, $x_i$=edad del trabajador $i$
- $\epsilon_i \overset{iid}{\sim} ALD(0, \sigma, \tau)$, $\sigma > 0$, $0 < \tau < 1$,

- $\mathbf{b} \sim N(0, \sigma_b^2 \mathbf{I}_K)$, es el vector de parámetros de efectos aleatorios, $\sigma_b > 0$ e $\mathbf{I}_K$ es la matriz identidad,
- $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$, es el vector de parámetros de efectos fijos,
- $\mathbf{Z_i}$ es la $i-ésima$ fila de la matriz $\mathbf{Z}$, la cual es la matriz de coeficientes aleatorios dada por $\mathbf{Z} = \mathbf{Z}_K \boldsymbol{\Omega}_K^{-1/2}$, donde

$$\mathbf{Z}_K = \begin{bmatrix} |x_1 - \kappa_1|^3 & |x_1 - \kappa_2|^3 & \cdots & |x_1 - \kappa_K|^3 \\ |x_2 - \kappa_1|^3 & |x_2 - \kappa_2|^3 & \cdots & |x_2 - \kappa_K|^3 \\ \vdots & \vdots & \ddots & \vdots \\ |x_n - \kappa_1|^3 & |x_n - \kappa_2|^3 & \cdots & |x_n - \kappa_K|^3 \end{bmatrix},$$

$$\boldsymbol{\Omega}_K = \begin{bmatrix} 0 & |\kappa_1 - \kappa_2|^3 & \cdots & |\kappa_1 - \kappa_K|^3 \\ |\kappa_2 - \kappa_1|^3 & 0 & \cdots & |\kappa_2 - \kappa_K|^3 \\ \vdots & \vdots & \ddots & \vdots \\ |\kappa_n - \kappa_1|^3 & |\kappa_n - \kappa_2|^3 & \cdots & 0 \end{bmatrix},$$

y, $\kappa_1 < \kappa_2 < \ldots < \kappa_K$ son nodos fijos. Considerando los resultados obtenidos en el estudio de simulación 2 (ver sección 4.5) y además por lo sugerido en Crainiceanu et al. (2005)) fueron empleados en la modelación $K$=20 nodos.
Además, como se indicó en la sección 3.3, el modelo puede expresarse a través de la representación jerárquica dada en (3.4) y asumiendo que las prioris para los parámetros de interés del modelo, de acuerdo con (3.6), (3.7) y (3.8) son

$$\boldsymbol{\beta} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 10^6 & 0 \\ 0 & 10^6 \end{pmatrix}\right) \tag{5.3}$$

$$\sigma \sim GI\left(10^{-6}, 10^{-6}\right) \tag{5.4}$$

$$\sigma_b^2 \sim GI\left(10^{-6}, 10^{-6}\right). \tag{5.5}$$

Entonces la descripción de dicha representación jerárquica Bayesiana del modelo de regresión cuantílica semiparamétrico aplicada a la base de datos Canadian age-income puede escribirse en código BUGS, el cual puede encontrarse en el anexo (A.3). Como se aprecia, en el modelo participan las matrices $\mathbf{X}$ y $\mathbf{Z} = \mathbf{Z}_K \boldsymbol{\Omega}_K^{-1/2}$. Éstas son obtenidas fuera del WinBUGS y luego son ingresadas como datos. Para esto, se proporciona un programa en R que calcula dichas matrices y emplea el paquete **R2WinBUGS** para invocar a WinBUGS 1.4.3 desde R. Este programa también aparece en el anexo (A.4).

### 5.1.2. Resultados de la Inferencia Bayesiana

Fueron realizadas para la inferencia 500000 iteraciones, definiéndose un periodo de *burning* de 50000 y saltos de 50. En estas condiciones se observó que la convergencia de la cadena es adecuada (anexo B.5). Estas simulaciones tomaron aproximadamente 1125 segundos.

En el cuadro 5.2 se muestra la media, desviación estándar y mediana a posteriori así como el intervalo de credibilidad del 95 % para algunos de los parámetros del modelo, considerando el cuantil 50 ($\tau = 0.50$) de la variable respuesta.

| **Parámetro** | Media | Desviación estándar | 2.5 % | Mediana | 97.5 % |
|---|---|---|---|---|---|
| $\beta_0$ | 14.31000 | 1.88800 | 10.79000 | 14.16000 | 18.84000 |
| $\beta_1$ | -0.01319 | 0.04367 | -0.11670 | -0.00957 | 0.06644 |
| $b_1$ | -0.00680 | 0.00322 | -0.01421 | -0.00640 | -0.00160 |
| $b_2$ | 0.00073 | 0.00517 | -0.00829 | 0.00010 | 0.01315 |
| $b_3$ | 0.00120 | 0.00520 | -0.00892 | 0.00097 | 0.01283 |
| $b_4$ | 0.00012 | 0.00522 | -0.01114 | 0.00037 | 0.00996 |
| $b_5$ | -0.00047 | 0.00577 | -0.01437 | -0.00002 | 0.00967 |
| $b_6$ | 0.00090 | 0.00521 | -0.00970 | 0.00090 | 0.01170 |
| $b_7$ | 0.00155 | 0.00554 | -0.00853 | 0.00119 | 0.01441 |
| $b_8$ | 0.00015 | 0.00534 | -0.01099 | 0.00015 | 0.01111 |
| $b_9$ | -0.00014 | 0.00529 | -0.01094 | -0.00015 | 0.01073 |
| $b_{10}$ | -0.00051 | 0.00524 | -0.01120 | -0.00058 | 0.01064 |
| $b_{11}$ | -0.00127 | 0.00546 | -0.01337 | -0.00094 | 0.00902 |
| $b_{12}$ | -0.00010 | 0.00523 | -0.01131 | -0.00004 | 0.01065 |
| $b_{13}$ | 0.00132 | 0.00548 | -0.00902 | 0.00099 | 0.01366 |
| $b_{14}$ | 0.00241 | 0.00560 | -0.00721 | 0.00184 | 0.01549 |
| $b_{15}$ | 0.00234 | 0.00544 | -0.00737 | 0.00181 | 0.01513 |
| $b_{16}$ | 0.00123 | 0.00550 | -0.00959 | 0.00106 | 0.01309 |
| $b_{17}$ | -0.00015 | 0.00536 | -0.01163 | -0.00003 | 0.01053 |
| $b_{18}$ | -0.00012 | 0.00546 | -0.01104 | -0.00024 | 0.01150 |
| $b_{19}$ | -0.00066 | 0.00500 | -0.01112 | -0.00073 | 0.01023 |
| $b_{20}$ | -0.00193 | 0.00377 | -0.00923 | -0.00200 | 0.00627 |
| $\sigma$ | 0.43500 | 0.01535 | 0.40650 | 0.43450 | 0.46620 |
| $\sigma_b$ | 0.00521 | 0.00271 | 0.00196 | 0.00450 | 0.01230 |

Cuadro 5.2: Media, desviación estándar y mediana a posteriori así como el intervalo de credibilidad del 95 % para algunos de los parámetros del modelo de regresión cuantílica semiparamétrico aplicado al conjunto de datos Canadian age-income, considerando el cuantil 50 ($\tau = 0.50$) de la variable respuesta.

Se muestra además en el cuadro 5.3, una comparación del DIC obtenido cuando se realiza el ajuste del conjunto de datos Canadian age-income al modelo de regresión cuantílica semiparamétrico con el DIC resultante del ajuste del mismo conjunto de datos a un modelo de regresión cuantílica lineal. Ambos ajustes son para el cuantil 50 ($\tau = 0.50$) de la variable respuesta. Se aprecia un mejor desempeño del modelo de regresión cuantílica semiparamétrico.

| Modelo de Regresión Cuantílica | $\overline{D}$ | Número efectivo de parámetros ($d_e$) | DIC |
|---|---|---|---|
| Lineal | 443.02 | 0.89 | 443.91 |
| Semiparamétrico | 246.24 | 15.16 | 261.40 |

Cuadro 5.3: DIC para el ajuste de los datos Canadian age-income al modelo de regresión cuantílica semiparamétrico y del ajuste del mismo conjunto de datos a un modelo de regresión cuantílica lineal, considerando para ambos ajustes el cuantil 50 ($\tau = 0.50$) de la variable respuesta.

Por otro lado, en la figura 5.2 se muestra la mediana y los cuantiles 2.5 % y 97.5 % de las distribuciones a posteriori de la mediana, del cuantil 75 ($\tau = 0.75$) y del cuantil 25 ($\tau = 0.25$), de la variable respuesta (logaritmo del ingreso) para cada valor de la covariable (edad).

En la figura 5.3 se presentan en un solo gráfico las funciones de los cuantiles 25, 50 y

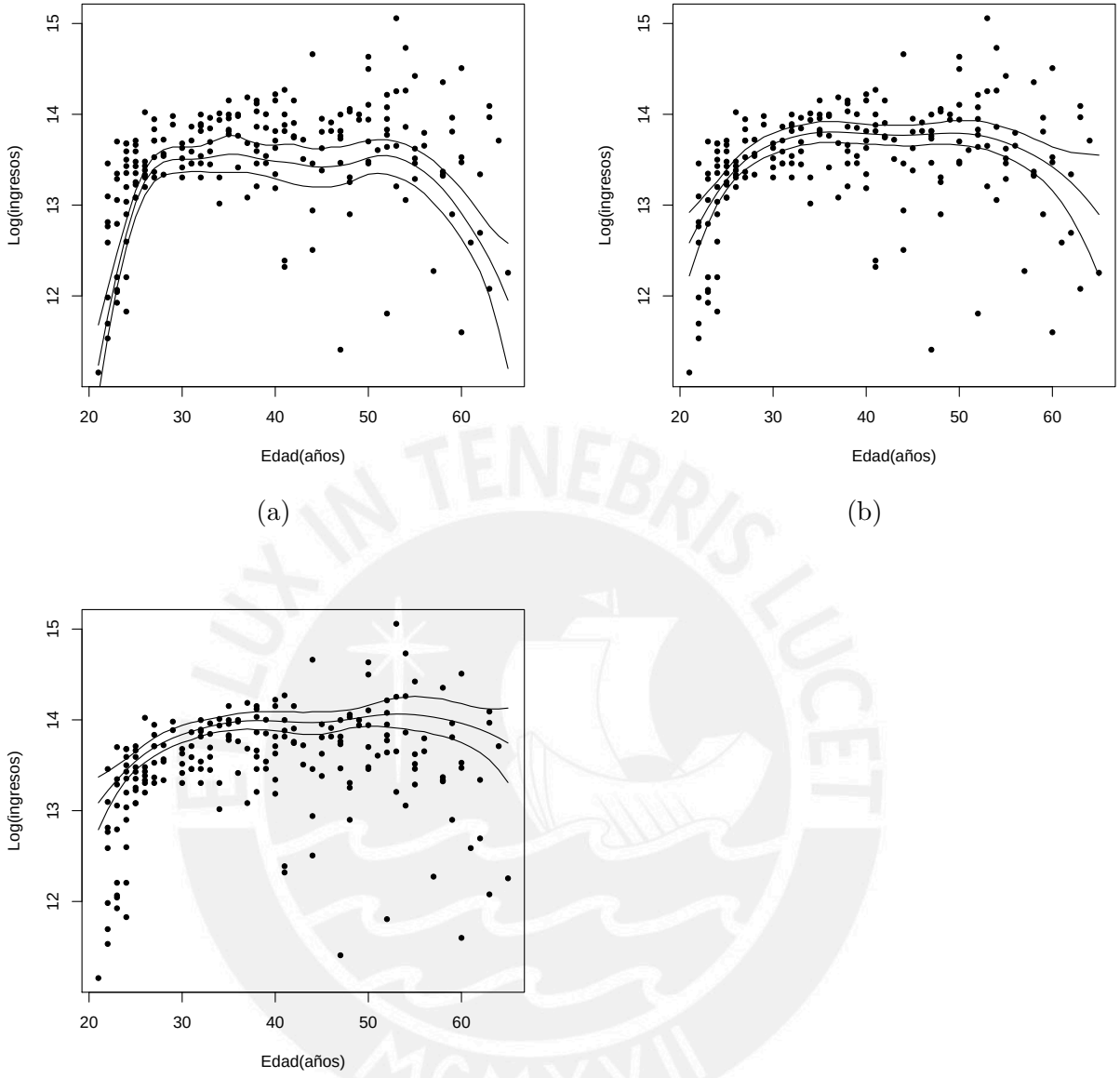


Figura 5.2: Mediana a posteriori e intervalos de credibilidad del 95 % para: (a)el cuantil 25 ($\tau = 0.25$), (b) la mediana ($\tau = 0.5$) y (c) el cuantil 75 ($\tau = 0.75$), de la variable respuesta (logaritmo del ingreso) para cada valor de la covariable (edad).

75 de la variable respuesta (logaritmo del ingreso) para cada valor de la covariable (edad). Como se puede observar, en el cuantil $\tau$=0.25, a partir de los 50 años de edad del trabajador, la reducción de los ingresos es más notoria respecto a lo que sucede en el cuantil $\tau$=0.5, en cambio en el cuantil $\tau$=0.75 se muestra una leve reducción, mostrando así cierta estabilidad en los ingresos.

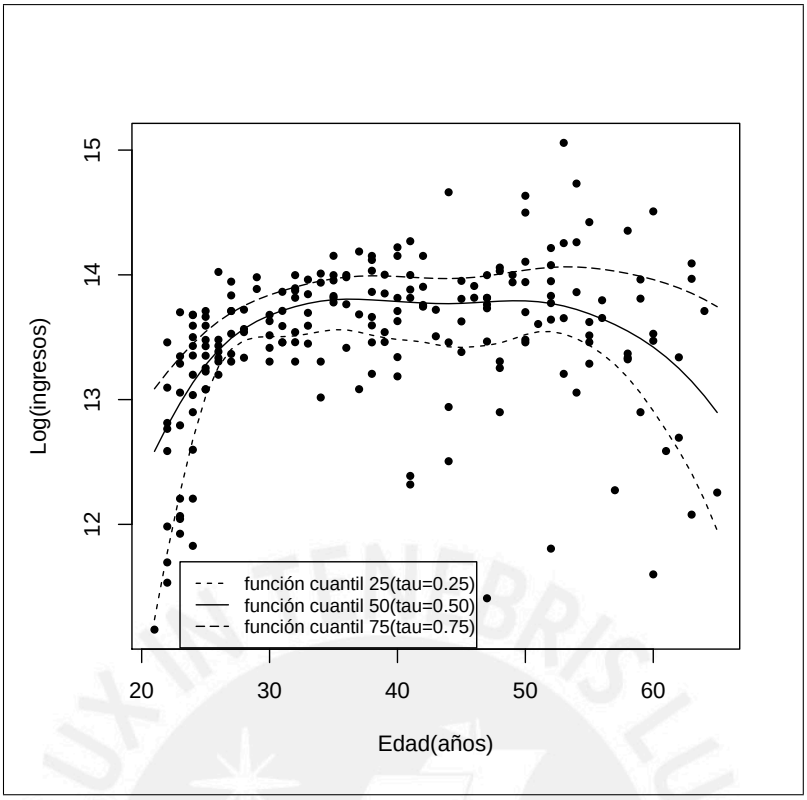


Figura 5.3: Función del cuantil 25, 50 y 75, de la variable respuesta (logaritmo del ingreso) para cada valor de la covariable (edad).

## 5.2. Aplicación 2: Base de datos de Lima Metropolitana ENAHO 2004: Ingreso Laboral versus edad

Lo que se trata de hacer aquí es aplicar el modelo de regresión cuantílica semiparamétrico sobre la información de las variables que se encuentran en la Encuesta Nacional de Hogares (ENAHO) del año 2004, por lo que a través de nuestro modelo expresaremos un determinado cuantil de la variable respuesta (logaritmo del ingreso laboral) como una función de la edad del individuo modelando esta última variable mediante thin-plate splines de bajo rango, tal como se ha definido en nuestro modelo. La inferencia se realizará desde la perspectiva Bayesiana.

### 5.2.1. Base de datos y variables empleadas

Como ya se ha comentado, la base de datos que se emplea es la Encuesta Nacional de Hogares (ENAHO) del año 2004, información recolectada por el Instituto Nacional de Estadística e Informática (INEI). La base fue seleccionada por ser la base utilizada por MTPE (2006) así como por Zevallos (2012). Siguiendo el procedimiento del MTPE (2006), se restringe el análisis a las encuestas de Lima Metropolitana, lo que implicó un tamaño de muestra de 3535 personas[1]. A continuación se presentan las definiciones operacionales

[1]La información laboral de la ENAHO se registra en el capítulo 500. La base de datos original recoge información de 7590 personas para el área de Lima Metropolitana. Sin embargo; se procedió a eliminar observaciones por: no ser parte de la PEA empleada (3008), contener información imputada (698), o no consignar información sobre salario y/o días de trabajo (349)

seguidas en la ENAHO para el cálculo de las variables consideradas:

- **Ingreso laboral por hora:** Variable aleatoria continua. Resultado de la división entre el ingreso anual en soles corrientes de la persona y el número de horas de trabajo al año. Para el cálculo del ingreso anual se consideraron los ingresos y beneficios (monetarios y no monetarios) de las ocupaciones principal y secundaria. Siguiendo la convención de la literatura especializada, el ingreso laboral por hora se presenta en escala logarítmica. Los códigos de las preguntas de la ENAHO empleadas son: d524a1, d529t, d538a1, d540t, d544t, d530a, d536, d541a, d543, p513t p518 y p520.
- **Edad:** Variable aleatoria discreta. Indica la edad de la persona en años. Se emplea la variable p208a de la ENAHO.

Algunos estadísticos descriptivos de estas variables se muestran en el cuadro 5.4.

| | Logaritmo del ingreso laboral | Edad |
|---|---|---|
| **Estadístico** | | |
| Mínimo | -3.55 | 14.00 |
| Primer cuartil | 0.61 | 26.00 |
| Mediana | 1.07 | 35.00 |
| Media | 1.11 | 37.16 |
| Tercer cuartil | 1.61 | 46.00 |
| Máximo | 6.31 | 90.00 |
| Desviación estándar | 0.87 | 13.32 |

Cuadro 5.4: Estadísticas descriptivas de la base Lima metropolitana Ingreso vs. Edad de la ENAHO 2004.

Con la finalidad de entender más la relación entre ambas variables, un diagrama de dispersión de la edad versus el logaritmo del ingreso se presenta en la figura 5.4, observándose que la relación entre ambas variables es no-lineal.

### 5.2.2. Modelo y prioris

Siguiendo el modelo dado en (3.1):

$$y_i = m(x_i) + \epsilon_i, \qquad i = 1, 2, \ldots, n \tag{5.6}$$

modelamos un determinado cuantil de la variable respuesta (logaritmo del ingreso laboral) como una función de la edad del individuo. Además, como se explicó en la sección 3.2, podemos representar el modelo dado en (5.6) como un modelo lineal mixto de la forma

$$y_i = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i}\mathbf{b} + \epsilon_i, \qquad i = 1, 2, \ldots, n \tag{5.7}$$

donde

- $y_i$=logaritmo del ingreso laboral del individuo $i$
- $\mathbf{X_i} = (1, x_i)$, es la $i - ésima$ fila de la matriz $\mathbf{X}$, que en este caso es la matriz de datos de edades, $x_i$=edad del individuo $i$

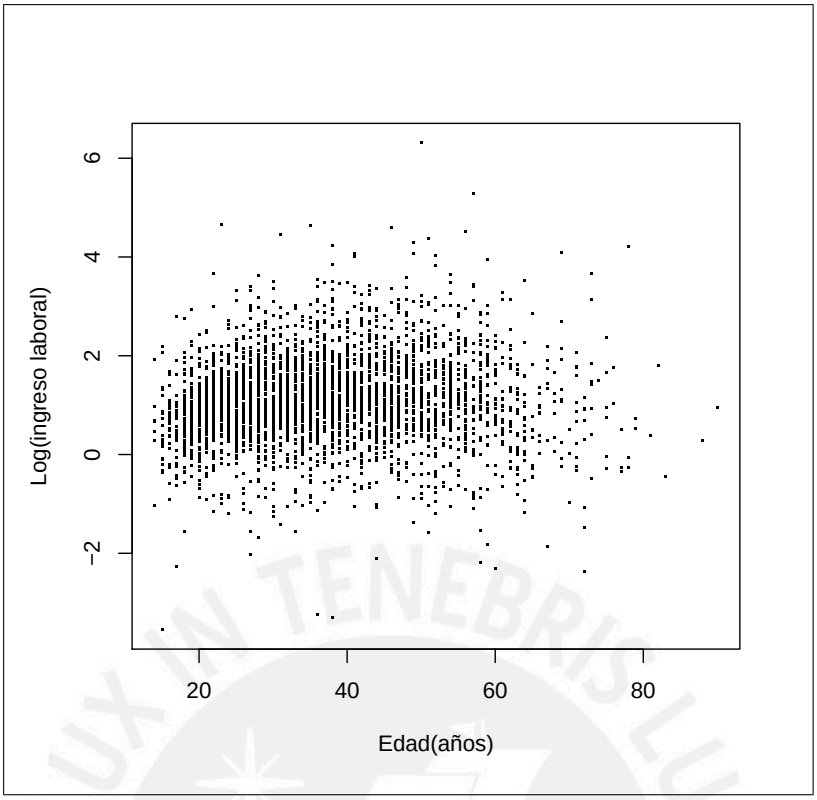


Figura 5.4: Diagrama de dispersión del Logaritmo de los ingresos vs. la Edad en la base Lima metropolitana de la ENAHO 2004.

- $\epsilon_i \overset{iid}{\sim} ALD(0, \sigma, \tau)$, $\sigma > 0$, $0 < \tau < 1$,
- $\mathbf{b} \sim N(0, \sigma_b^2 \mathbf{I}_K)$, es el vector de parámetros de efectos aleatorios, $\sigma_b > 0$ e $\mathbf{I}_K$ es la matriz identidad,
- $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$, es el vector de parámetros de efectos fijos,
- $\mathbf{Z_i}$ es la $i-ésima$ fila de la matriz $\mathbf{Z}$, la cual es la matriz de coeficientes aleatorios dada por $\mathbf{Z} = \mathbf{Z}_K \boldsymbol{\Omega}_K^{-1/2}$, donde

$$\mathbf{Z}_K = \begin{bmatrix} |x_1 - \kappa_1|^3 & |x_1 - \kappa_2|^3 & \cdots & |x_1 - \kappa_K|^3 \\ |x_2 - \kappa_1|^3 & |x_2 - \kappa_2|^3 & \cdots & |x_2 - \kappa_K|^3 \\ \vdots & \vdots & \ddots & \vdots \\ |x_n - \kappa_1|^3 & |x_n - \kappa_2|^3 & \cdots & |x_n - \kappa_K|^3 \end{bmatrix},$$

$$\boldsymbol{\Omega}_K = \begin{bmatrix} 0 & |\kappa_1 - \kappa_2|^3 & \cdots & |\kappa_1 - \kappa_K|^3 \\ |\kappa_2 - \kappa_1|^3 & 0 & \cdots & |\kappa_2 - \kappa_K|^3 \\ \vdots & \vdots & \ddots & \vdots \\ |\kappa_n - \kappa_1|^3 & |\kappa_n - \kappa_2|^3 & \cdots & 0 \end{bmatrix},$$

y, $\kappa_1 < \kappa_2 < \ldots < \kappa_K$ son nodos fijos. Considerando los resultados obtenidos en el estudio de simulación 2 (ver sección 4.5) y además por lo sugerido en Crainiceanu et al. (2005)) fueron empleados en la modelación $K$=20 nodos.

Además, como se indicó en la sección 3.3, el modelo puede expresarse a través de la representación jerárquica dada en (3.4) y asumiendo que las prioris para los parámetros de interés del modelo, de acuerdo con (3.6), (3.7) y (3.8) son

$$\boldsymbol{\beta} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 10^6 & 0 \\ 0 & 10^6 \end{pmatrix}\right) \tag{5.8}$$

$$\sigma \sim GI\left(10^{-6}, 10^{-6}\right) \tag{5.9}$$

$$\sigma_b^2 \sim GI\left(10^{-6}, 10^{-6}\right). \tag{5.10}$$

entonces la descripción de esta representación jerárquica Bayesiana del modelo de regresión cuantílica semiparamétrico aplicada a las variables logaritmo del ingreso laboral y edad del individuo de la base Lima metropolitana de la ENAHO 2004 puede ser escrita en código BUGS, el cual puede encontrarse en el anexo (A.5). Como se aprecia, en el modelo participan las matrices $\mathbf{X}$ y $\mathbf{Z} = \mathbf{Z}_K\boldsymbol{\Omega}_K^{-1/2}$. Éstas son obtenidas fuera del WinBUGS y luego son ingresadas como datos. Para esto, se proporciona un programa en R que calcula dichas matrices y emplea el paquete **R2WinBUGS** para invocar a WinBUGS 1.4.3 desde R. Este programa también aparece en el anexo (A.6).

### 5.2.3. Resultados de la Inferencia Bayesiana

Fueron realizadas para la inferencia 500000 iteraciones, definiéndose un periodo de *burning* de 50000 y saltos de 50. En estas condiciones se observó que la convergencia de la cadena es adecuada (anexo B.6). Estas simulaciones tomaron aproximadamente 21914 segundos.

En el cuadro 5.5 se muestra la media, desviación estándar y mediana a posteriori así como el intervalo de credibilidad del 95 % para algunos de los parámetros del modelo, considerando el cuantil 50 ($\tau = 0.50$) de la variable respuesta.

Se muestra además en el cuadro 5.6, una comparación del DIC obtenido cuando se realiza el ajuste de las variables logaritmo del ingreso laboral y la edad del individuo de la base Lima metropolitana de la ENAHO 2004 al modelo de regresión cuantílica semiparamétrico, con el DIC resultante del ajuste del mismo conjunto de datos a un modelo de regresión cuantílica lineal. Ambos ajustes son para el cuantil 50 ($\tau = 0.50$) de la variable respuesta. Se aprecia un mejor desempeño del modelo de regresión cuantílica semiparamétrico.

Por otro lado, en la figura 5.5 se presentan las funciones de los cuantiles 5, 10, 25, 50, 75, 90 y 95 de la variable respuesta (logaritmo del ingreso laboral) para cada valor de la covariable (edad). Como se puede observar, en el cuantil $\tau$=0.25, a partir de los 50 años de edad del trabajador, la reducción de los ingresos es más notoria respecto a lo que sucede en el cuantil $\tau$=0.5 donde la reducción es más leve, y en cambio en el cuantil $\tau$=0.75 se muestra una mayor estabilidad en los ingresos.

| **Parámetro** | Media | Desviación estándar | 2.5 % | Mediana | 97.5 % |
|---|---|---|---|---|---|
| $\beta_0$ | 2.16200 | 1.42800 | -0.16490 | 1.93900 | 6.46800 |
| $\beta_1$ | -0.01747 | 0.02752 | -0.09341 | -0.01322 | 0.02608 |
| $b_1$ | -0.00162 | 0.00086 | -0.00334 | -0.00159 | -0.00002 |
| $b_2$ | -0.00142 | 0.00230 | -0.00912 | -0.00102 | 0.00171 |
| $b_3$ | 0.00034 | 0.00192 | -0.00305 | 0.00016 | 0.00493 |
| $b_4$ | -0.00034 | 0.00203 | -0.00502 | -0.00033 | 0.00385 |
| $b_5$ | -0.00148 | 0.00240 | -0.00789 | -0.00109 | 0.00240 |
| $b_6$ | -0.00150 | 0.00230 | -0.00765 | -0.00114 | 0.00215 |
| $b_7$ | 0.00019 | 0.00194 | -0.00377 | 0.00014 | 0.00435 |
| $b_8$ | 0.00051 | 0.00214 | -0.00407 | 0.00045 | 0.00506 |
| $b_9$ | -0.00045 | 0.00217 | -0.00610 | -0.00020 | 0.00325 |
| $b_{10}$ | -0.00011 | 0.00213 | -0.00513 | 0.00004 | 0.00372 |
| $b_{11}$ | 0.00047 | 0.00207 | -0.00389 | 0.00048 | 0.00466 |
| $b_{12}$ | 0.00067 | 0.00204 | -0.00333 | 0.00061 | 0.00502 |
| $b_{13}$ | 0.00082 | 0.00217 | -0.00295 | 0.00058 | 0.00604 |
| $b_{14}$ | -0.00007 | 0.00215 | -0.00407 | -0.00021 | 0.00484 |
| $b_{15}$ | -0.00182 | 0.00261 | -0.00867 | -0.00136 | 0.00207 |
| $b_{16}$ | -0.00075 | 0.00233 | -0.00567 | -0.00073 | 0.00409 |
| $b_{17}$ | -0.00015 | 0.00232 | -0.00436 | -0.00029 | 0.00560 |
| $b_{18}$ | 0.00029 | 0.00255 | -0.00374 | -0.00002 | 0.00670 |
| $b_{19}$ | -0.00020 | 0.00216 | -0.00427 | -0.00029 | 0.00456 |
| $b_{20}$ | 0.00006 | 0.00118 | -0.00230 | 0.00002 | 0.00246 |
| $\sigma$ | 0.56110 | 0.00470 | 0.55200 | 0.56100 | 0.57050 |
| $\sigma_b$ | 0.00213 | 0.00118 | 0.00086 | 0.00178 | 0.00551 |

Cuadro 5.5: Media, desviación estándar y mediana a posteriori así como el intervalo de credibilidad del 95 % para algunos de los parámetros del modelo de regresión cuantílica semiparamétrico aplicado a las variables logaritmo del ingreso laboral vs. la edad del individuo de la base Lima metropolitana de la ENAHO 2004, considerando el cuantil 50 ($\tau = 0.50$) de la variable respuesta.

| Modelo de Regresión Cuantílica | $\overline{D}$ | Número efectivo de parámetros ($d_e$) | DIC |
|---|---|---|---|
| Lineal | 9791.43 | 1.20 | 9792.63 |
| Semiparamétrico | 9486.38 | 15.15 | 9501.53 |

Cuadro 5.6: DIC para el ajuste de las variables logaritmo del ingreso laboral vs. la edad del individuo de la base Lima metropolitana de la ENAHO 2004 al modelo de regresión cuantílica semiparamétrico y del ajuste del mismo conjunto de datos a un modelo de regresión cuantílica lineal, considerando para ambos ajustes el cuantil 50 ($\tau = 0.50$) de la variable respuesta.

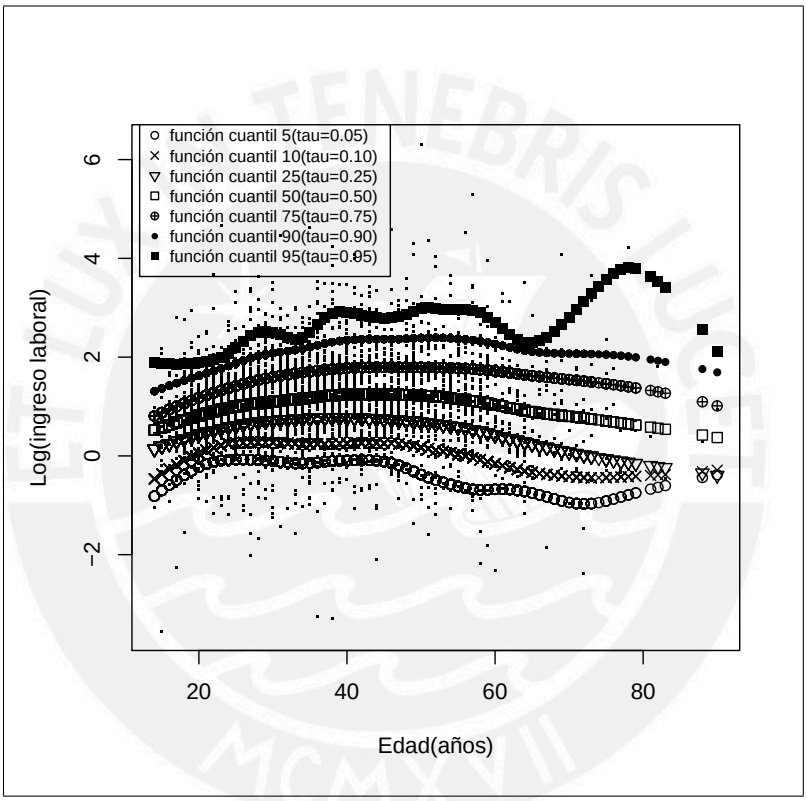


Figura 5.5: Función del cuantil 5, 10, 25, 50, 75, 90 y 95 de la variable respuesta (logaritmo del ingreso laboral) para cada valor de la covariable (edad).

# Capítulo 6

# Conclusiones

## 6.1. Conclusiones

- En este trabajo hemos propuesto un Modelo de Regresión Cuantílica Semiparamétrico el cual surge de emplear la metodología de regresión semiparamétrica sugerida por Crainiceanu, Ruppert y Wand (2005) pero aplicada a la regresión cuantílica. La inferencia del modelo es realizada desde la perspectiva Bayesiana usando MCMC.

- Hemos obtenido las distribuciones condicionales completas para todos los parámetros del modelo propuesto, encontrándose formas cerradas o conocidas para todas ellas, lo cual facilita la implementación del algoritmo Gibbs para generar muestras de las distribuciones a posteriori de los parámetros.

- Basados en los resultados del estudio de simulación 1 (ver sección 4.4) se concluye que el modelo propuesto tiene un mejor ajuste que el modelo de regresión cuantílico lineal, cuando la relación entre la covariable y los cuantiles de la variable respuesta es no-lineal.

- Basados en los resultados del estudio de simulación 2 (ver sección 4.5) se concluye que es suficiente utilizar de 15 a 20 nodos para obtener un ajuste adecuado.

- Basados en los resultados del estudio de simulación 3 (ver sección 4.6) se concluye que a mayor variabilidad presente en la variable respuesta se obtendrán mayores valores del MAE y RMSE. También se aprecia que el modelo teórico es fácilmente descrito con pocos nodos para los diferentes valores de variabilidad de los datos.

- El modelo propuesto ha sido ilustrado aplicándolo al conjunto de datos Canadian age-income, mostrando un adecuado desempeño en el ajuste. Se pudo observar que, en el cuantil $\tau$=0.25, a partir de los 50 años de edad del trabajador, la reducción de los ingresos es más notoria respecto a lo que sucede en el cuantil $\tau$=0.5 donde la reducción es más leve, y en cambio en el cuantil $\tau$=0.75 se muestra una mayor estabilidad en los ingresos.

- El modelo propuesto ha sido ilustrado aplicándolo también a una base de datos de Lima metropolitana de la ENAHO del año 2004, relacionando el logaritmo del ingreso laboral por hora y la edad del individuo. Se pudo observar que, en el cuantil $\tau$=0.25, a partir de los 50 años de edad del trabajador, la reducción de los ingresos es más notoria

respecto a lo que sucede en el cuantil $\tau$=0.5 donde la reducción es más leve, y en cambio en el cuantil $\tau$=0.75 se muestra una mayor estabilidad en los ingresos.

- Al comparar el ajuste del modelo de regresión cuantílica semiparamétrico versus el ajuste a un modelo de regresión cuantílica lineal, de los conjuntos de datos empleados en las aplicaciones, evaluando los valores del DIC, el modelo semiparamétrico resulta ser más adecuado por el menor valor del DIC obtenido.

## 6.2. Sugerencias para investigaciones futuras

- En los estudios de simulación se emplearon thin-plate splines para modelar la función suave $m(x)$ tal como fue sugerido en Crainiceanu, Ruppert y Wand (2005). Es posible investigar y comparar el ajuste obtenido con otros modelos spline.
- La presencia, comportamiento e influencia de outliers en el modelo de regresión cuantílica semiparamétrico podría ser estudiada vía las técnicas de diagnóstico desde la perspectiva Bayesiana.

# Apéndice A

# Programas en WinBUGS y R

## A.1. Programa en R para el Estudio de simulación 1

```
M<-20  # Número de réplicas
n<-100  # Número de datos en la muestra
sigma<-1.5
num.nodos<-20
t<-0.5
theta1<-(1-2*t)/(t*(t-1))
theta2<-sqrt(t/(t*(1-t)))
##############################################################################
musimu<-matrix(0,n,M)
xsim<-matrix(0,n,M)
beta0fitQR<-matrix(0,1,M)
beta1fitQR<-matrix(0,1,M)
sigmafitQR<-matrix(0,1,M)
mufitQR<-matrix(0,n,M)
MAEQR<-matrix(0,1,M)
RMSEQR<-matrix(0,1,M)
tiempoQR<-matrix(0,1,M)
beta0fitQRsemip<-matrix(0,1,M)
beta1fitQRsemip<-matrix(0,1,M)
bfitQRsemip<-matrix(0,num.nodos,M)
mufitQRsemip<-matrix(0,n,M)
MAEQRsemip<-matrix(0,1,M)
RMSEQRsemip<-matrix(0,1,M)
sigmafitQRsemip<-matrix(0,1,M)
sigmabfitQRsemip<-matrix(0,1,M)
tiempoQRsemip<-matrix(0,1,M)
Dtheta<-matrix(0,ME,M)
Despera<-matrix(0,ME,M)
Dtheta1<-matrix(0,ME1,M)
Despera1<-matrix(0,ME1,M)
##############################################################################
```

```
#Estudio de simulación
################################################################################
for (j in 1:M){
x<-runif(n,-3,3)
v<-rexp(n,1/sigma)
w<-rnorm(n,0,0.1)
################################################################################
xsim[,j]<-x
################################################################################
#Simulando valores "y" de una función curva m(x)
################################################################################
y<-0.3*x+0.5*sin(2*x)+1.1/(1+x^2)+theta1*v +theta2*(sigma^(0.5))*(v^(0.5))*w
################################################################################
musim<-  0.3*x+0.5*sin(2*x)+1.1/(1+x^2)
################################################################################
musimu[,j]<-musim
################################################################################
#Estimación bayesiana del modelo de Regresión Cuantílica Lineal con la data
#simulada
################################################################################
library(R2WinBUGS)
inits<-function(){list(beta=c(0,0),tau=0.01)}
parameters<-list("beta","sigma","mu")
n<-length(x)

data<-list(y=c(y),x=x,n=n,t=t)
tiempoBayesQR.fit<-system.time(

BayesQR.fit<- bugs(data, inits, parameters, model.file = "C:/Users/Hugo/
Desktop/codigo modelo QR.txt",
n.chains = 1, n.iter = 500000, n.burnin = 50000,
n.thin = 50,debug = FALSE, DIC = FALSE, digits = 5,
codaPkg = FALSE,bugs.directory = "C:/Users/Hugo/Desktop/winbugs14
/WinBUGS14/"))
attach.all(BayesQR.fit)

beta0fitQR[,j]<-BayesQR.fit$mean$beta[1]
beta1fitQR[,j]<-BayesQR.fit$mean$beta[2]
sigmafitQR[,j]<-BayesQR.fit$mean$sigma
mufitQR[,j]<-BayesQR.fit$mean$mu
MAEQR[,j]<-mean(abs(musimu[,j]-mufitQR[,j]))
RMSEQR[,j]<-sqrt(mean((musimu[,j]-mufitQR[,j])^2))
```

```
tiempoQR[,j]<-tiempoBayesQR.fit[3]

######Cálculo de los D(theta) y el D(theta)medio ######################
ME<-BayesQR.fit$n.sims
n<-n  #tamaño de la muestra
t<-t  #cuantil definido
sigma1<-matrix(BayesQR.fit$sims.matrix[,3],ME,1)
mut<-BayesQR.fit$sims.matrix[,-c(1:3)]
y<-y
#########################################################################
for (i in 1:ME){
mu<-mut[i,]
x1<-y-mu
sigma2<-sigma1[i,]
check <- function(x1,t){
res<-c()
res<-x1*(t-1*(x1<0))
return(res)
}
D<-function(sigma2,x1,t){
sigma2<-sigma2
n<-length(x1)
x1<-x1
t<-t
res<-c()
res<- -2*(n*log(t)+n*log(1-t)-n*log(sigma2) - (1/sigma2)*sum(check(x1,t)))
return(res)
print(res)
}
D(sigma2,x1,t)

Dtheta[i,j]<-D(sigma2,x1,t)

######Cálculo del D(theta) esperado ###################################
sigmamedio<-BayesQR.fit$mean$sigma
mu1<-matrix(BayesQR.fit$mean$mu,n,1)
x2<-y-mu1

check1 <- function(x2,t){
res<-c()
res<-x2*(t-1*(x2<0))
return(res)
```

```
}
Desp<-function(sigmamedio,x2,t){
sigmamedio<-sigmamedio
n<-length(x2)
x2<-x2
t<-t
res<-c()
res<- -2*(n*log(t)+n*log(1-t)-n*log(sigmamedio) -
(1/sigmamedio)*sum(check1(x2,t)))
return(res)
print(res)
}
Desp(sigmamedio,x2,t)

Despera[i,j]<-Desp(sigmamedio,x2,t)
}

##############################################################################
#Estimación bayesiana del modelo Regresión Cuantílica Semiparamétrico con la
#data simulada
##############################################################################
library(R2WinBUGS)
inits.b=rep(0,num.nodos)
inits1<-function(){list(beta=c(0,0),b=inits.b,taub=0.01,tau=0.01)}
parameters1<-list("beta","sigma","b","sigmab","mu")

n<-length(x)
X<-cbind(rep(1,n),x)
num.knots<-num.nodos
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)
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)))

data1<-list(y=c(y),X=X,Z=Z,n=n,num.knots=num.knots,t=t)

tiempoBayesQRsemip.fit<-system.time(
```

```
BayesQRsemip.fit<- bugs(data1, inits1, parameters1, model.file = "C:/Users/
Hugo/Desktop/code semipqr.txt",
n.chains = 1, n.iter = 500000, n.burnin = 50000,
n.thin =50,debug = FALSE, DIC = FALSE, digits = 5,
codaPkg = FALSE,bugs.directory = "C:/Users/Hugo/Desktop/winbugs14/
WinBUGS14/"))
attach.all(BayesQRsemip.fit)

beta0fitQRsemip[,j]<-BayesQRsemip.fit$mean$beta[1]
beta1fitQRsemip[j]<-BayesQRsemip.fit$mean$beta[2]
bfitQRsemip[,j]<-BayesQRsemip.fit$mean$b
mufitQRsemip[,j]<-BayesQRsemip.fit$mean$mu
MAEQRsemip[,j]<-mean(abs(musimu[,j]-mufitQRsemip[,j]))
RMSEQRsemip[,j]<-sqrt(mean((musimu[,j]-mufitQRsemip[,j])^2))
sigmafitQRsemip[,j]<-BayesQRsemip.fit$mean$sigma
sigmabfitQRsemip[,j]<-BayesQRsemip.fit$mean$sigmab
tiempoQRsemip[,j]<-tiempoBayesQRsemip.fit[3]

######Cálculo de los D(theta) y el D(theta)medio ##############################
ME1<-BayesQRsemip.fit$n.sims
n<-n  # tamaño de la muestra
t<-t  #cuantil definido
knot<-20  #número de nodos
sigma3<-matrix(BayesQRsemip.fit$sims.matrix[,3],ME1,1)
sigmab3<-matrix(BayesQRsemip.fit$sims.matrix[,24],ME1,1)
bt<-BayesQRsemip.fit$sims.matrix[,-c(1:3,24:124)]
mut1<-BayesQRsemip.fit$sims.matrix[,-c(1:24)]
y<-y
################################################################################
for (k in 1:ME1){
mu2<-mut1[k,]
x3<-y-mu2
be<-bt[k,]
sigma4<-sigma3[k,]
sigmab4<-sigmab3[k,]
check2 <- function(x3,t){
res<-c()
res<-x3*(t-1*(x3<0))
return(res)
}
D<-function(sigma4,sigmab4,be,x3,t){
sigma4<-sigma4
```

```
sigmab4<-sigmab4
be<-be
n<-length(x3)
x3<-x3
t<-t
res<-c()
res<- -2*(n*log(t)+n*log(1-t)-n*log(sigma4)-(1/sigma4)*sum(check2(x3,t))-
(knot/2)*log(2*pi)-knot*log(sigmab4)-(1/(2*sigmab4*sigmab4))*sum(be^2))
return(res)
print(res)
}
D(sigma4,sigmab4,be,x3,t)

Dtheta1[k,j]<-D(sigma4,sigmab4,be,x3,t)

######Cálculo del D(theta) esperado ##########################################
sigmamedio1<-BayesQRsemip.fit$mean$sigma
sigmabmedio1<-BayesQRsemip.fit$mean$sigmab
bemedio1<-BayesQRsemip.fit$mean$b[]
mu3<-matrix(BayesQRsemip.fit$mean$mu,n,1)
x4<-y-mu3

check3 <- function(x4,t){
res<-c()
res<-x4*(t-1*(x4<0))
return(res)
}
Desp<-function(sigmamedio1,sigmabmedio1,bemedio1,x4,t){
sigmamedio1<-sigmamedio1
sigmabmedio1<-sigmabmedio1
bemedio1<-bemedio1
n<-length(x4)
x4<-x4
t<-t
res<-c()
res<- -2*(n*log(t)+n*log(1-t)-n*log(sigmamedio1) -
(1/sigmamedio1)*sum(check3(x4,t))-(knot/2)*log(2*pi)-knot*log(sigmabmedio1)-
(1/(2*sigmabmedio1*sigmabmedio1))*sum(bemedio1^2))
return(res)
print(res)
}
Desp(sigmamedio1,sigmabmedio1,bemedio1,x4,t)
```

```
Despera1[k,j]<-Desp(sigmamedio1,sigmabmedio1,bemedio1,x4,t)
}
}
colMeans(Dtheta)
colMeans(Despera)
colMeans(Dtheta1)
colMeans(Despera1)
##########################################################################
####Número efectivo de parámetros#########################################
##########################################################################
peQR=colMeans(Dtheta)-colMeans(Despera)
mean(peQR)
peQRsemip=colMeans(Dtheta1)-colMeans(Despera1)
mean(peQRsemip)
##########################################################################
######Cálculo del DIC en el modelo de Regresión Cuantílica Lineal ######
##########################################################################
DICQR=colMeans(Dtheta)+peQR
mean(DICQR)
##########################################################################
#Cálculo del DIC en el modelo de Regresión Cuantílica Semiparamétrico ##
##########################################################################
DICQRsemip=colMeans(Dtheta1)+peQRsemip
mean(DICQRsemip)
##########################################################################
a1=mean(MAEQR)
a2=mean(RMSEQR)
a3=sum(tiempoQR)
a4=mean(MAEQRsemip)
a5=mean(RMSEQRsemip)
a6=sum(tiempoQRsemip)

resultados<-matrix(c(a1,a2,a3,a4,a5,a6),3,2)
rownames(resultados)<-c("MAE","RMSE","Tiempo (seg.)")
colnames(resultados)<-c("Modelo de Regresión Cuantílica Lineal",
"Modelo de Regresión Cuantílica Semiparamétrico")
round(resultados,4)
```

## A.2. Programa en R para el Estudio de simulación 2

```
M<-20  # Número de réplicas
n<-50  # Número de datos en la muestra
sigma<-1.5
```

```
num.nodos<-5
t<-0.5
theta1<-(1-2*t)/(t*(t-1))
theta2<-sqrt(t/(t*(1-t)))
##############################################################################
musimu<-matrix(0,n,M)
xsim<-matrix(0,n,M)
beta0fitQRsemip<-matrix(0,1,M)
beta1fitQRsemip<-matrix(0,1,M)
bfitQRsemip<-matrix(0,num.nodos,M)
mufitQRsemip<-matrix(0,n,M)
MAEQRsemip<-matrix(0,1,M)
RMSEQRsemip<-matrix(0,1,M)
sigmafitQRsemip<-matrix(0,1,M)
sigmabfitQRsemip<-matrix(0,1,M)
tiempoQRsemip<-matrix(0,1,M)
Dtheta1<-matrix(0,ME1,M)
Despera1<-matrix(0,ME1,M)

##############################################################################
#Estudio de simulación
##############################################################################
for (j in 1:M){
x<-runif(n,-3,3)
v<-rexp(n,1/sigma)
w<-rnorm(n,0,0.1)
##############################################################################
xsim[,j]<-x
##############################################################################
#Simulando valores "y" de una función curva m(x)
##############################################################################
y<-0.3*x+0.5*sin(2*x)+1.1/(1+x^2)+theta1*v +theta2*(sigma^(0.5))*(v^(0.5))*w
##############################################################################
musim<-  0.3*x+0.5*sin(2*x)+1.1/(1+x^2)
##############################################################################
musimu[,j]<-musim

##############################################################################
#Estimación bayesiana del modelo Regresión Cuantílica Semiparamétrico con la
#data simulada
##############################################################################
library(R2WinBUGS)
```

```
inits.b=rep(0,num.nodos)
inits1<-function(){list(beta=c(0,0),b=inits.b,taub=0.01,tau=0.01)}
parameters1<-list("beta","sigma","b","sigmab","mu")

n<-length(x)
X<-cbind(rep(1,n),x)
num.knots<-num.nodos
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)
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)))

data1<-list(y=c(y),X=X,Z=Z,n=n,num.knots=num.knots,t=t)

tiempoBayesQRsemip.fit<-system.time(

BayesQRsemip.fit<- bugs(data1, inits1, parameters1, model.file = "C:/Users/
Hugo/Desktop/code semipqr.txt",
n.chains = 1, n.iter = 500000, n.burnin = 50000,
n.thin =50,debug = FALSE, DIC = FALSE, digits = 5,
codaPkg = FALSE,bugs.directory = "C:/Users/Hugo/Desktop/winbugs14/
WinBUGS14/"))
attach.all(BayesQRsemip.fit)

beta0fitQRsemip[,j]<-BayesQRsemip.fit$mean$beta[1]
beta1fitQRsemip[j]<-BayesQRsemip.fit$mean$beta[2]
bfitQRsemip[,j]<-BayesQRsemip.fit$mean$b
mufitQRsemip[,j]<-BayesQRsemip.fit$mean$mu
MAEQRsemip[,j]<-mean(abs(musimu[,j]-mufitQRsemip[,j]))
RMSEQRsemip[,j]<-sqrt(mean((musimu[,j]-mufitQRsemip[,j])^2))
sigmafitQRsemip[,j]<-BayesQRsemip.fit$mean$sigma
sigmabfitQRsemip[,j]<-BayesQRsemip.fit$mean$sigmab
tiempoQRsemip[,j]<-tiempoBayesQRsemip.fit[3]
}

######Cálculo de los D(theta) y el D(theta)medio ##########################
ME1<-BayesQRsemip.fit$n.sims
n<-n  # tamaño de la muestra
```

```
t<-t  #cuantil definido
knot<-5 #número de nodos
sigma3<-matrix(BayesQRsemip.fit$sims.matrix[,3],ME1,1)
sigmab3<-matrix(BayesQRsemip.fit$sims.matrix[,9],ME1,1)
bt<-BayesQRsemip.fit$sims.matrix[,-c(1:3,9:59)]
mut1<-BayesQRsemip.fit$sims.matrix[,-c(1:9)]
y<-y
##############################################################################
for (k in 1:ME1){
mu2<-mut1[k,]
x3<-y-mu2
be<-bt[k,]
sigma4<-sigma3[k,]
sigmab4<-sigmab3[k,]
check2 <- function(x3,t){
res<-c()
res<-x3*(t-1*(x3<0))
return(res)
}
D<-function(sigma4,sigmab4,be,x3,t){
sigma4<-sigma4
sigmab4<-sigmab4
be<-be
n<-length(x3)
x3<-x3
t<-t
res<-c()
res<- -2*(n*log(t)+n*log(1-t)-n*log(sigma4)-(1/sigma4)*sum(check2(x3,t))-
(knot/2)*log(2*pi)-knot*log(sigmab4)-(1/(2*sigmab4*sigmab4))*sum(be^2))
return(res)
print(res)
}
D(sigma4,sigmab4,be,x3,t)

Dtheta1[k,j]<-D(sigma4,sigmab4,be,x3,t)

######Cálculo del D(theta) esperado ###################################
sigmamedio1<-BayesQRsemip.fit$mean$sigma
sigmabmedio1<-BayesQRsemip.fit$mean$sigmab
bemedio1<-BayesQRsemip.fit$mean$b[]
mu3<-matrix(BayesQRsemip.fit$mean$mu,n,1)
x4<-y-mu3
```

```
check3 <- function(x4,t){
res<-c()
res<-x4*(t-1*(x4<0))
return(res)
}
Desp<-function(sigmamedio1,sigmabmedio1,bemedio1,x4,t){
sigmamedio1<-sigmamedio1
sigmabmedio1<-sigmabmedio1
bemedio1<-bemedio1
n<-length(x4)
x4<-x4
t<-t
res<-c()
res<- -2*(n*log(t)+n*log(1-t)-n*log(sigmamedio1) -
(1/sigmamedio1)*sum(check3(x4,t))-(knot/2)*log(2*pi)-knot*log(sigmabmedio1)-
(1/(2*sigmabmedio1*sigmabmedio1))*sum(bemedio1^2))
return(res)
print(res)
}
Desp(sigmamedio1,sigmabmedio1,bemedio1,x4,t)

Despera1[k,j]<-Desp(sigmamedio1,sigmabmedio1,bemedio1,x4,t)
}
}

colMeans(Dtheta1)
colMeans(Despera1)
##########################################################################
####Número efectivo de parámetros#########################################
##########################################################################
peQRsemip=colMeans(Dtheta1)-colMeans(Despera1)
mean(peQRsemip)
##########################################################################
#Cálculo del DIC en el modelo de Regresión Cuantílica Semiparamétrico ##
##########################################################################
DICQRsemip=colMeans(Dtheta1)+peQRsemip
mean(DICQRsemip)
##########################################################################

a1=mean(MAEQRsemip)
a2=mean(RMSEQRsemip)
```

```
a3=sum(tiempoQRsemip)
```

## A.3. Programa en WinBUGS del modelo de regresión cuantílica semiparamétrico aplicado al conjunto de datos Canadian age-income

```
model{
#verosimilitud del modelo

for (i in 1 : n)
{
y[i] ~ dnorm(mu.star[i], pre[i])

mu.star[i] <- mu[i] + theta[1]*v[i]
mu[i] <- mfe[i]+mre110[i]+mre1120[i]

mfe[i]<-beta[1]*X[i,1]+beta[2]*X[i,2]

mre110[i]<-b[1]*Z[i,1]+b[2]*Z[i,2]+b[3]*Z[i,3]+b[4]*Z[i,4]+
b[5]*Z[i,5]+b[6]*Z[i,6]+b[7]*Z[i,7]+b[8]*Z[i,8]+
b[9]*Z[i,9]+b[10]*Z[i,10]

mre1120[i]<-b[11]*Z[i,11]+b[12]*Z[i,12]+b[13]*Z[i,13]+b[14]*Z[i,14]+
b[15]*Z[i,15]+b[16]*Z[i,16]+b[17]*Z[i,17]+b[18]*Z[i,18]+
b[19]*Z[i,19]+b[20]*Z[i,20]

v[i] ~ dexp(tau)
pre[i] <- tau/(theta[2]*theta[2]*v[i])
}

#prioris

for (l in 1 : 2){beta[l]~dnorm(0,1.0E-6)}
for (k in 1 : num.nodos){b[k]~dnorm(0,taub)}
tau~dgamma(1.0E-6,1.0E-6)
taub~dgamma(1.0E-6,1.0E-6)

sigma<-1/sqrt(tau)
sigmab<-1/sqrt(taub)
theta[1] <- (1-2*t)/(t*(1-t))
theta[2] <- sqrt(2/(t*(1-t)))
}
```

*Datos*: Consisten en, la variable respuesta (y[ ]), la matriz de efectos fijos (X[ , ]), la matriz de efectos aleatorios (Z[ , ]), el tamaño de la muestra (n) y el número de nodos (num.nodos).

*Valores iniciales*: Son dados para, los coeficientes de efectos fijos $\beta$ (beta[ ]), los coeficientes de efectos aleatorios $b$ (b[ ]) y las precisiones tau y taub. Los demás valores iniciales se generan aleatoriamente en WinBUGS desde las distribuciones a priori.

Tanto los datos como los valores iniciales son especificados y procesados en R y luego empleados en WinBUGS a través de la función bugs implementada en el paquete **R2WinBUGS**, tal como fue explicado al final de la subsección 5.1.1. El código en R, se muestra a continuación.

## A.4. Programa en R que calcula las matrices X y $\mathbf{Z} = \mathbf{Z}_K\Omega_K^{-1/2}$ , procesa datos y valores iniciales, para el conjunto de datos Canadian age-income

```
library(R2WinBUGS)
data.file.name="C:/Users/Desktop/age income semipqr.txt"
inits.b=rep(0,20)
inits<-function(){list(beta=c(0,0),b=inits.b,taub=0.01,tau=0.01)}
parametros<-list("sigmab","sigma","beta","b")

data<-read.table(file=data.file.name,header=TRUE,sep="\t")
attach(data)
n<-length(edad)
X<-cbind(rep(1,n),edad)
num.nodos<-20
t<-0.5
nodos<-quantile(unique(edad),
seq(0,1,length=(num.nodos+2))[-c(1,(num.nodos+2))])

Z_K<-(abs(outer(edad,nodos,"-")))^3
OMEGA_all<-(abs(outer(nodos,nodos,"-")))^3
svd.OMEGA_all<-svd(OMEGA_all)
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)))

data<-list("y","X","Z","n","num.nodos","t")
Bayes.fit<- bugs(data, inits, parametros,
model.file = "C:/Users/code canadian age income semipqr.txt",
n.chains = 1, n.iter = 500000, n.burnin = 50000,
n.thin =50,debug = TRUE, DIC = FALSE, digits = 5,
codaPkg = FALSE,bugs.directory = "C:/Users/Desktop/winbugs14/WinBUGS14/")
attach.all(Bayes.fit)
```

```
###Calculando los D(theta) y el D(theta)medio##############################
n.iter<-500000
n.burnin<-50000
n.thin<-50
M<-(n.iter-n.burnin)/n.thin # Número de iteraciones efectivas
n<-205  # tamaño de la muestra
t<-0.5  # cuantil definido
k<-20 # número de nodos
####rescatando los M sigmas de las M iteraciones
sigma<-matrix(Bayes.fit$sims.matrix[,2],M,1)
####rescatando los M sigmasb de las M iteraciones
sigmab<-matrix(Bayes.fit$sims.matrix[,1],M,1)
####rescatando los "n" mu de las M iteraciones
mut<-Bayes.fit$sims.matrix[,-c(1:24,230)]
####rescatando los b de las M iteraciones
bt<-Bayes.fit$sims.matrix[,-c(1:4,25:230)]
y<-matrix(data[,2],n,1)
Dtheta<-matrix(0,M,1)

#########################################################################
for (i in 1:M){
mu<-mut[i,]
x<-y-mu
be<-bt[i,]
sigma1<-sigma[i,]
sigmab1<-sigmab[i,]
check <- function(x,t){
res<-c()
res<-x*(t-1*(x<0))
return(res)
}
D<-function(sigma1,sigmab1,be,x,t){
sigma1<-sigma1
sigmab1<-sigmab1
be<-be
n<-length(x)
x<-x
t<-t
res<-c()
res<- -2*(n*log(t)+n*log(1-t)-n*log(sigma1)-(1/sigma1)*sum(check(x,t))-
(k/2)*log(2*pi)-k*log(sigmab1)-(1/(2*sigmab1*sigmab1))*sum(be^2))
```

```
return(res)
print(res)
}
D(sigma1,sigmab1,be,x,t)

Dtheta[i,]<-D(sigma1,sigmab1,be,x,t)
}
Dtheta
Dthetamedio<-mean(Dtheta)
Dthetamedio

###Calculando el D(theta) esperado#######################################
sigmamedio<-Bayes.fit$mean$sigma
sigmabmedio<-Bayes.fit$mean$sigmab
bemedio<-Bayes.fit$mean$b[]
mu1<-matrix(Bayes.fit$mean$mu,n,1)
x1<-y-mu1

check <- function(x1,t){
res<-c()
res<-x1*(t-1*(x1<0))
return(res)
}
Desp<-function(sigmamedio,sigmabmedio,bemedio,x1,t){
sigmamedio<-sigmamedio
sigmabmedio<-sigmabmedio
bemedio<-bemedio
n<-length(x1)
x1<-x1
t<-t
res<-c()
res<- -2*(n*log(t)+n*log(1-t)-n*log(sigmamedio)-
(1/sigmamedio)*sum(check(x1,t))-(k/2)*log(2*pi)-k*log(sigmabmedio)-
(1/(2*sigmabmedio*sigmabmedio))*sum(bemedio^2))
return(res)
print(res)
}
Desp(sigmamedio,sigmabmedio,bemedio,x1,t)

####Número efectivo de parámetros#########################################
p_D=Dthetamedio-Desp(sigmamedio,sigmabmedio,bemedio,x1,t)
p_D
```

```
####Cálculo del DIC#########################################################

DIC=Dthetamedio+p_D
DIC
```

## A.5. Programa en WinBUGS del modelo de regresión cuantílica semiparamétrico aplicado a las variables logaritmo del Ingreso laboral vs. Edad de la base Lima metropolitana de la ENAHO 2004

```
model{

#verosimilitud del modelo

for (i in 1:n)
{
lnw[i] ~ dnorm(mu.star[i], pre[i])

mu.star[i] <- mu[i] + theta[1]*v[i]
mu[i] <- mfe[i]+mre110[i]+mre1120[i]

mfe[i]<-beta[1]*X[i,1]+beta[2]*X[i,2]

pre[i] <- tau/(theta[2]*theta[2]*v[i])

mre110[i]<-b[1]*Z[i,1]+b[2]*Z[i,2]+b[3]*Z[i,3]+b[4]*Z[i,4]+
b[5]*Z[i,5]+b[6]*Z[i,6]+b[7]*Z[i,7]+b[8]*Z[i,8]+
b[9]*Z[i,9]+b[10]*Z[i,10]

mre1120[i]<-b[11]*Z[i,11]+b[12]*Z[i,12]+b[13]*Z[i,13]+b[14]*Z[i,14]+
b[15]*Z[i,15]+b[16]*Z[i,16]+b[17]*Z[i,17]+b[18]*Z[i,18]+
b[19]*Z[i,19]+b[20]*Z[i,20]

v[i] ~ dexp(tau)
}

#prioris

for (l in 1:2){beta[l]~dnorm(0,1.0E-6)}
for (k in 1:num.knots){b[k]~dnorm(0,taub)}
tau~dgamma(1.0E-6,1.0E-6)
taub~dgamma(1.0E-6,1.0E-6)
```

```
sigma<-1/sqrt(tau)
sigmab<-1/sqrt(taub)
theta[1] <- (1-2*t)/(t*(1-t))
theta[2] <- sqrt(2/(t*(1-t)))
}
```

*Datos*: Consisten en, la variable respuesta (y[ ]), la matriz de efectos fijos (X[ , ]), la matriz de efectos aleatorios (Z[ , ]), el tamaño de la muestra (n) y el número de nodos (num.nodos).

*Valores iniciales*: Son dados para, los coeficientes de efectos fijos $\beta$ (beta[ ]), los coeficientes de efectos aleatorios $b$ (b[ ]) y las precisiones tau y taub. Los demás valores iniciales se generan aleatoriamente en WinBUGS desde las distribuciones a priori.

Tanto los datos como los valores iniciales son especificados y procesados en R y luego empleados en WinBUGS a través de la función bugs implementada en el paquete **R2WinBUGS**, tal como fue explicado al final de la subsección 5.2.2. El código en R, se muestra a continuación.

## A.6. Programa en R que calcula las matrices X y $\mathbf{Z} = \mathbf{Z}_K\Omega_K^{-1/2}$ , procesa datos y valores iniciales, para las variables logaritmo del Ingreso laboral vs. Edad de la base Lima metropolitana de la ENAHO 2004

```
library(R2WinBUGS)
data.file.name="C:/Users/Hugo/Desktop/datalima edad ingreso semipqr.txt"
inits.b=rep(0,20)
inits<-function(){list(beta=c(0,0),b=inits.b,taub=0.01,tau=0.01)}
parameters<-list("lambda","sigmab","sigma","beta","b","mu")

data<-read.table(file=data.file.name,header=TRUE,sep="\t")
attach(data)
n<-length(edad)
X<-cbind(rep(1,n),edad)
num.knots<-20
t<-0.50
knots<-quantile(unique(edad),
seq(0,1,length=(num.knots+2))[-c(1,(num.knots+2))])

Z_K<-(abs(outer(edad,knots,"-")))^3
OMEGA_all<-(abs(outer(knots,knots,"-")))^3
svd.OMEGA_all<-svd(OMEGA_all)
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)))
```

```
data<-list("lnw","X","Z","n","num.knots","t")
Bayes.fit<- bugs(data, inits, parameters,
model.file = "C:/Users/Hugo/Desktop/
code limametrop edad ingreso semipqr.txt",
n.chains = 1, n.iter = 500000, n.burnin = 50000,
n.thin =50,debug = TRUE, DIC = FALSE, digits = 5,
codaPkg = FALSE,bugs.directory = "C:/Users/Desktop/winbugs14/WinBUGS14/")
attach.all(Bayes.fit)

###Calculando los D(theta) y el D(theta)medio#############################
n.iter<-500000
n.burnin<-50000
n.thin<-50
M<-(n.iter-n.burnin)/n.thin # Número de iteraciones efectivas
n<-3535  # tamaño de la muestra
t<-0.5  #cuantil definido
k<-20 #número de nodos
####rescatando los M sigmas de las M iteraciones
sigma<-matrix(Bayes.fit$sims.matrix[,3],M,1)
####rescatando los M sigmasb de las M iteraciones
sigmab<-matrix(Bayes.fit$sims.matrix[,2],M,1)
####rescatando los "n" mu de las M iteraciones
mut<-Bayes.fit$sims.matrix[,-c(1:25,3561)]
####rescatando los b de las M iteraciones
bt<-Bayes.fit$sims.matrix[,-c(1:5,26:3561)]
y<-matrix(data[,1],n,1)
Dtheta<-matrix(0,M,1)

#########################################################################
for (i in 1:M){
mu<-mut[i,]
x<-y-mu
be<-bt[i,]
sigma1<-sigma[i,]
sigmab1<-sigmab[i,]
check <- function(x,t){
res<-c()
res<-x*(t-1*(x<0))
return(res)
}
D<-function(sigma1,sigmab1,be,x,t){
sigma1<-sigma1
```

```
sigmab1<-sigmab1
be<-be
n<-length(x)
x<-x
t<-t
res<-c()
res<- -2*(n*log(t)+n*log(1-t)-n*log(sigma1) -(1/sigma1)*sum(check(x,t))-
(k/2)*log(2*pi)-k*log(sigmab1)-(1/(2*sigmab1*sigmab1))*sum(be^2))
return(res)
print(res)
}
D(sigma1,sigmab1,be,x,t)

Dtheta[i,]<-D(sigma1,sigmab1,be,x,t)
}
Dtheta
Dthetamedio<-mean(Dtheta)
Dthetamedio

###Calculando el D(theta) esperado#######################################
sigmamedio<-Bayes.fit$mean$sigma
sigmabmedio<-Bayes.fit$mean$sigmab
bemedio<-Bayes.fit$mean$b[]
mu1<-matrix(Bayes.fit$mean$mu,n,1)
x1<-y-mu1

check <- function(x1,t){
res<-c()
res<-x1*(t-1*(x1<0))
return(res)
}
Desp<-function(sigmamedio,sigmabmedio,bemedio,x1,t){
sigmamedio<-sigmamedio
sigmabmedio<-sigmabmedio
bemedio<-bemedio
n<-length(x1)
x1<-x1
t<-t
res<-c()
res<- -2*(n*log(t)+n*log(1-t)-n*log(sigmamedio)-
(1/sigmamedio)*sum(check(x1,t))-(k/2)*log(2*pi)-k*log(sigmabmedio)-
(1/(2*sigmabmedio*sigmabmedio))*sum(bemedio^2))
```

```
return(res)
print(res)
}
Desp(sigmamedio,sigmabmedio,bemedio,x1,t)

####Número efectivo de parámetros#########################################
p_D=Dthetamedio-Desp(sigmamedio,sigmabmedio,bemedio,x1,t)
p_D

####Cálculo del DIC#######################################################

DIC=Dthetamedio+p_D
DIC
```

# Apéndice B

# Anexos de tablas y gráficos

## B.1. Promedio del error absoluto(MAE), Raíz del error cuadrático medio(RMSE) y DIC para el Estudio de simulación 2

En el cuadro B.1 se muestran los resultados resumidos mediante el promedio del MAE, RMSE y DIC para diferentes tamaños de muestra y número de nodos empleado.

| **Parámetro a recuperar** | Tamaño de muestra | Número de nodos | Medida | | | Tiempo (seg.) |
|---|---|---|---|---|---|---|
| | | | MAE | RMSE | DIC | |
| "función m(x)" | 50 | 5 | 0.0621 | 0.0799 | 41.7267 | 2435.51 |
| | | 10 | 0.0610 | 0.0799 | 39.4036 | 3555.22 |
| | | 15 | 0.0575 | 0.0755 | 38.6020 | 5293.47 |
| | | 20 | 0.0564 | 0.0744 | 38.2435 | 6668.24 |
| | | 25 | 0.0565 | 0.0746 | 37.6409 | 9050.50 |
| | 100 | 5 | 0.0509 | 0.0650 | 76.9414 | 4851.51 |
| | | 10 | 0.0479 | 0.0617 | 73.9115 | 7001.33 |
| | | 15 | 0.0461 | 0.0593 | 72.8722 | 10407.74 |
| | | 20 | 0.0456 | 0.0592 | 72.8633 | 13283.67 |
| | | 25 | 0.0459 | 0.0595 | 72.5120 | 18007.62 |
| | 200 | 5 | 0.0325 | 0.0415 | 151.0323 | 9453.03 |
| | | 10 | 0.0337 | 0.0435 | 149.2668 | 13862.46 |
| | | 15 | 0.0319 | 0.0405 | 148.3335 | 20729.67 |
| | | 20 | 0.0317 | 0.0400 | 147.6084 | 26404.20 |
| | | 25 | 0.0315 | 0.0398 | 147.4688 | 35510.75 |
| | 300 | 5 | 0.0300 | 0.0372 | 236.2391 | 14304.27 |
| | | 10 | 0.0263 | 0.0328 | 231.8376 | 20480.63 |
| | | 15 | 0.0268 | 0.0338 | 230.6589 | 30986.10 |
| | | 20 | 0.0258 | 0.0326 | 230.2817 | 38384.85 |
| | | 25 | 0.0260 | 0.0328 | 230.0233 | 52873.88 |

Cuadro B.1: Resultados de medidas promedio de la simulación para comparar el ajuste del modelo de regresión cuantílica semiparamétrico para diferentes número de nodos, considerando M=20 réplicas.

## B.2. Simulación de 500,000 iteraciones, burn in de 50,000 y thin de 50 para el Estudio de simulación 2

La figura B.1 presenta las iteraciones de los parámetros $\beta_0$; $\beta_1$; $\sigma$; $\sigma_b$ y de los coeficientes $b$ de efectos aleatorios. (n=200 y nodos=20)

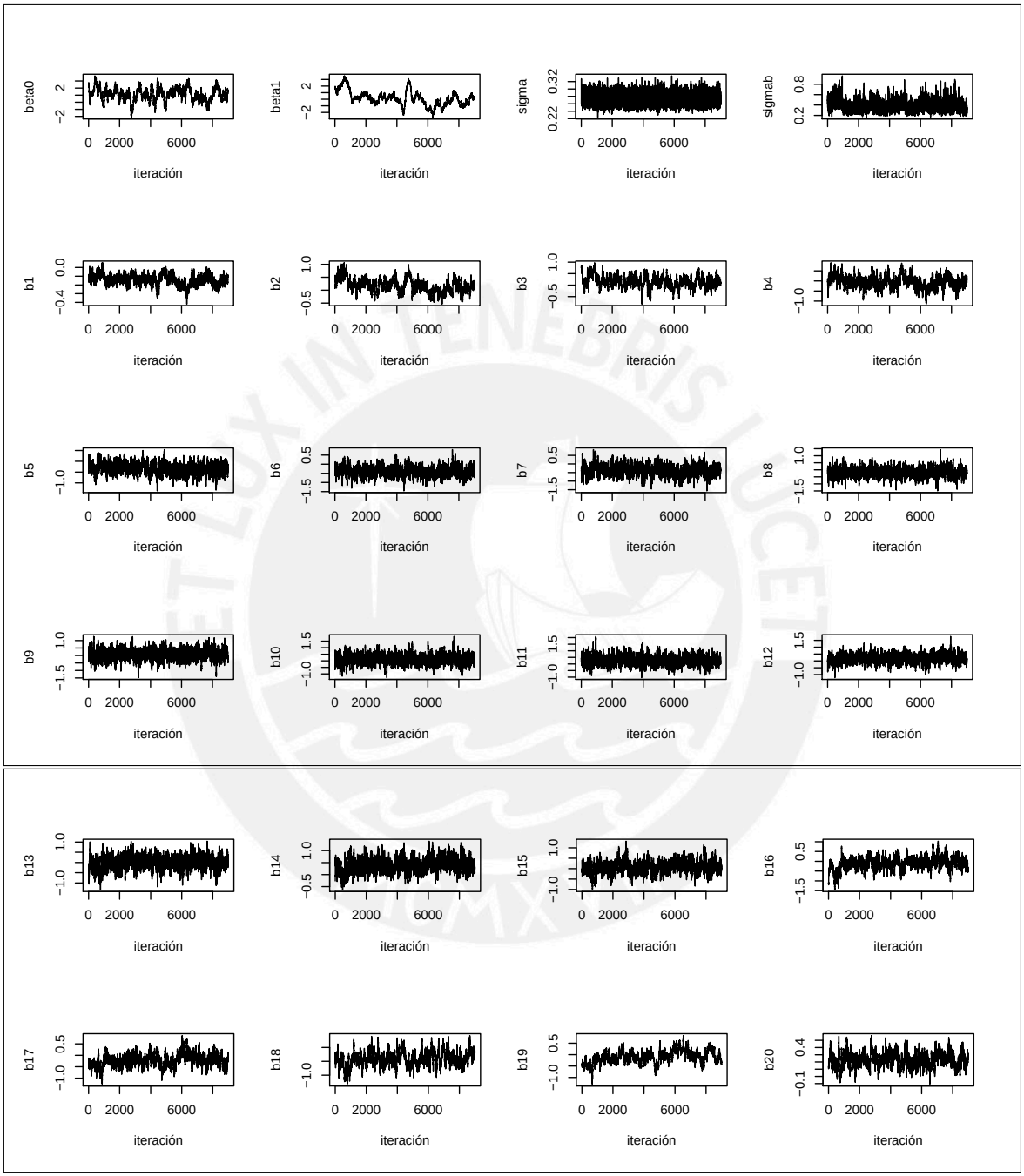


Figura B.1: Valores de los parámetros $\beta_0$; $\beta_1$; $\sigma$; $\sigma_b$ y de los coeficientes $b$ de efectos aleatorios en las iteraciones del estudio de simulación 2.

## B.3. Promedio del error absoluto(MAE), Raíz del error cuadrático medio(RMSE) y DIC para el Estudio de simulación 3

En el cuadro B.2 se muestran los resultados resumidos mediante el promedio del MAE, RMSE y DIC para tres diferentes valores de $\sigma$ y tres diferentes valores para el número de nodos empleado.

| **Parámetro a recuperar** | $\sigma$ | Número de nodos | Medida MAE | Medida RMSE | Medida DIC | Tiempo (seg.) |
|---|---|---|---|---|---|---|
| "función m(x)" | 0.75 | 10 | 0.0147 | 0.0190 | -29.410 | 20729.84 |
| | | 15 | 0.0158 | 0.0202 | -29.558 | 30853.01 |
| | | 20 | 0.0148 | 0.0188 | -31.968 | 38925.13 |
| | 1.5 | 10 | 0.0274 | 0.0348 | 226.485 | 20903.52 |
| | | 15 | 0.0266 | 0.0341 | 225.350 | 31025.68 |
| | | 20 | 0.0264 | 0.0336 | 224.514 | 39206.39 |
| | 4 | 10 | 0.0640 | 0.0810 | 625.232 | 20966.83 |
| | | 15 | 0.0611 | 0.0780 | 624.496 | 30859.44 |
| | | 20 | 0.0614 | 0.0781 | 624.122 | 38599.03 |

Cuadro B.2: Resultados de medidas promedio de la simulación para comparar el ajuste del modelo de regresión cuantílica semiparamétrico para tres diferentes valores de $\sigma$ y número de nodos, considerando M=20 réplicas.

## B.4. Simulación de 500,000 iteraciones, burn in de 50,000 y thin de 50 para el Estudio de simulación 3

La figura B.2 presenta las iteraciones de los parámetros $\beta_0$; $\beta_1$; $\sigma$; $\sigma_b$ y de los coeficientes $b$ de efectos aleatorios.(n=300, nodos=20)

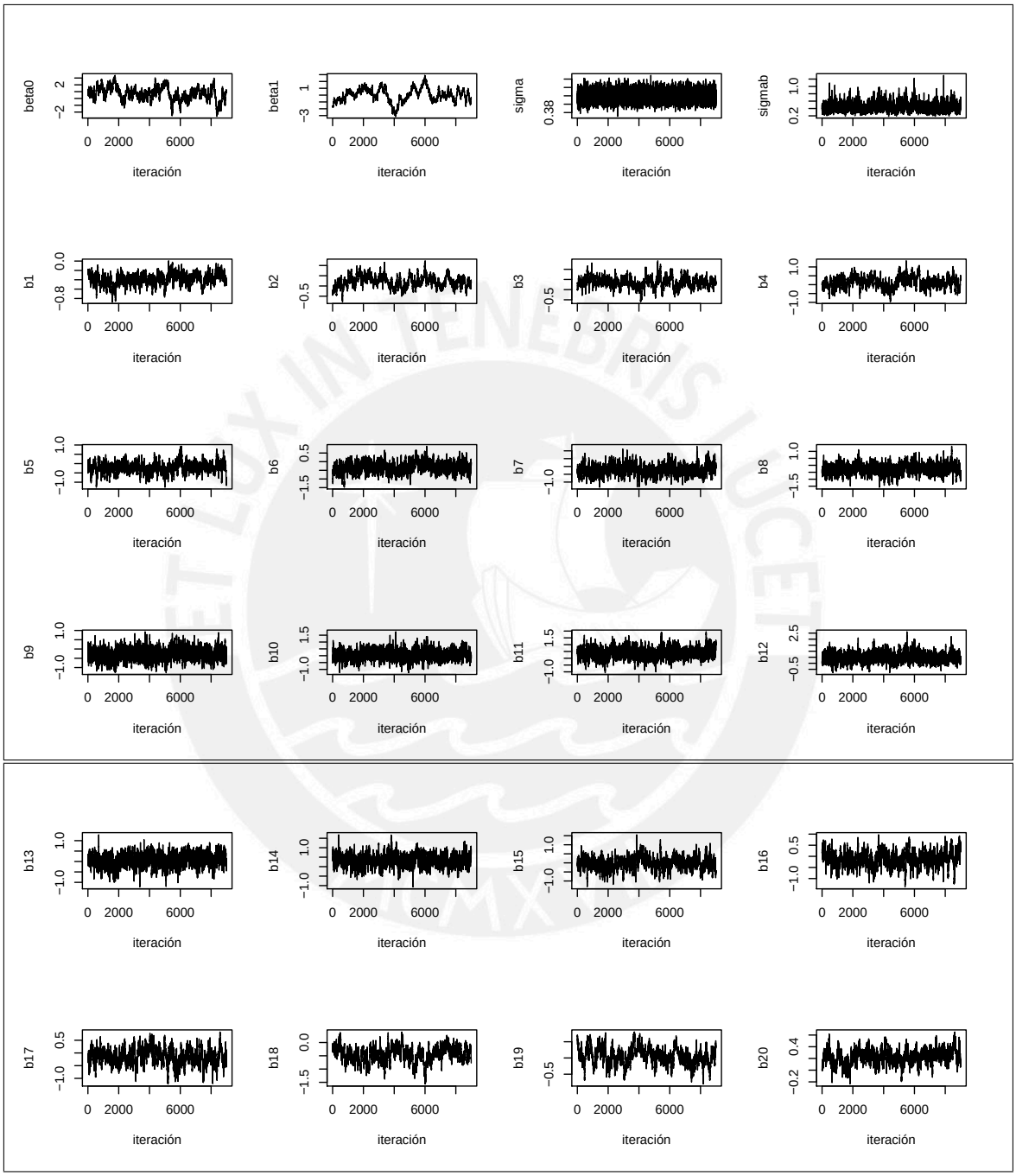


Figura B.2: Valores de los parámetros $\beta_0$; $\beta_1$; $\sigma$; $\sigma_b$ y de los coeficientes $b$ de efectos aleatorios en las iteraciones del estudio de simulación 3.

## B.5. Cadena de valores para los parámetros del modelo de regresión cuantílica semiparamétrico ajustado a los datos Canadian age income.

La figura B.3 presenta la cadena de valores a posteriori de los parámetros del modelo de regresión cuantílica semiparamétrico ajustado a los datos Canadian age income.

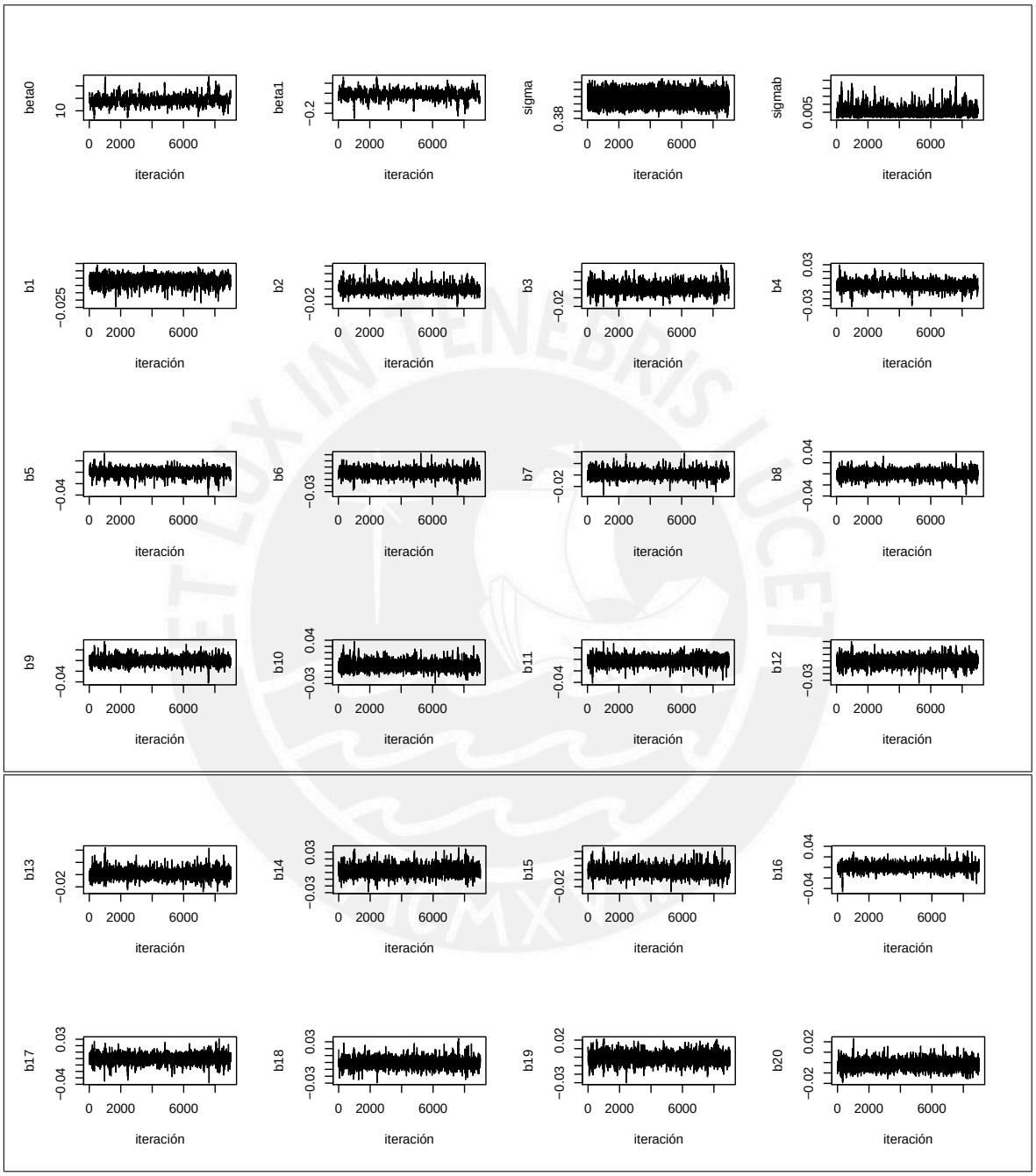


Figura B.3: Valores de los parámetros $\beta_0$; $\beta_1$; $\sigma$; $\sigma_b$ y de los coeficientes $b$ de efectos aleatorios del modelo de regresión cuantílica semiparamétrico ajustado a los datos Canadian age income.

## B.6. Cadena de valores para los parámetros del modelo de regresión cuantílica semiparamétrico ajustado a los datos Lima metropolitana ENAHO 2004.

La figura B.4 presenta la cadena de valores a posteriori de los parámetros del modelo de regresión cuantílica semiparamétrico ajustado a los datos Lima metropolitana ENAHO 2004.

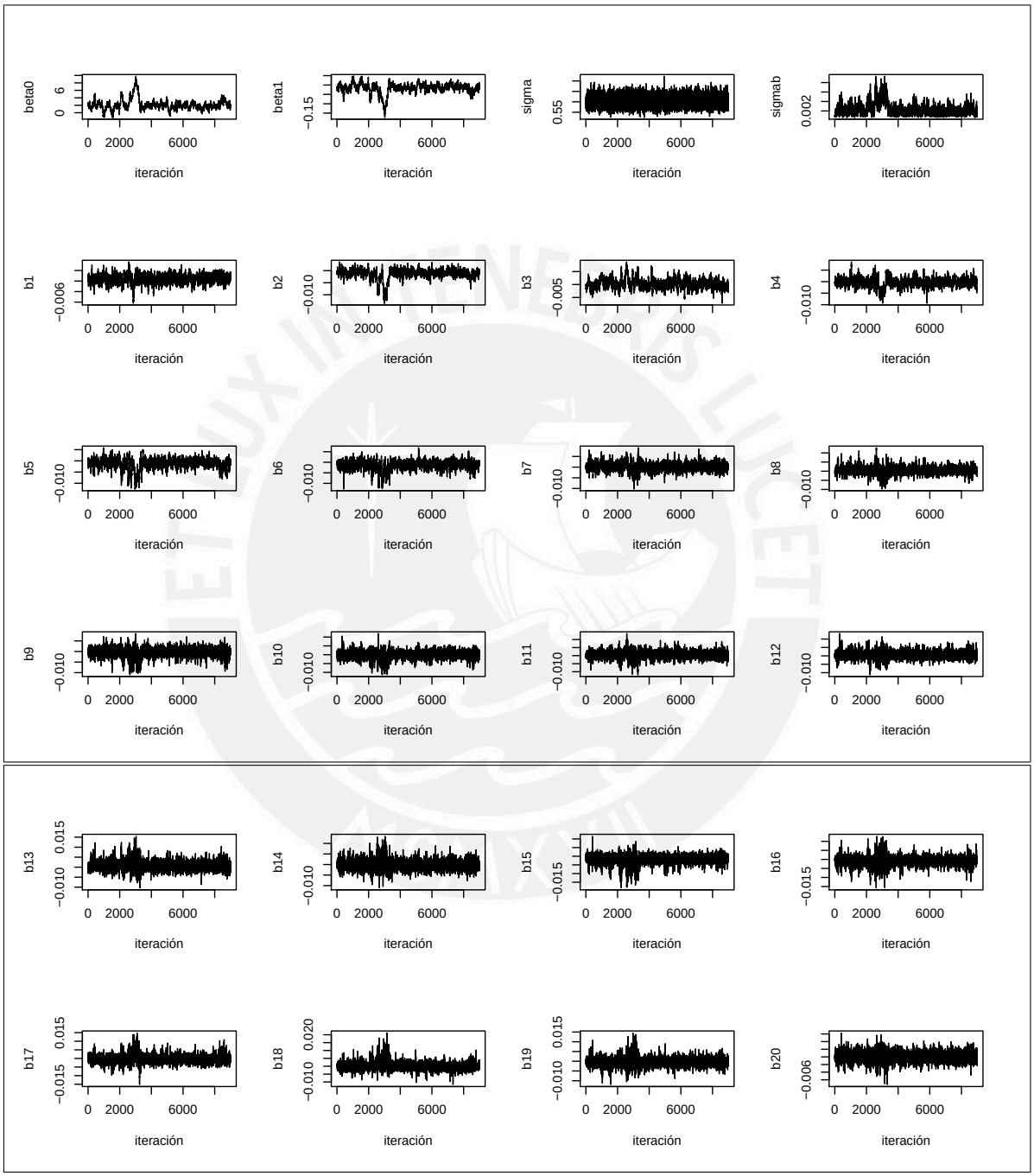


Figura B.4: Valores de los parámetros $\beta_0$; $\beta_1$; $\sigma$; $\sigma_b$ y de los coeficientes $b$ de efectos aleatorios del modelo de regresión cuantílica semiparamétrico ajustado a los datos Lima metropolitana ENAHO 2004.

# Bibliografía


Barndorff-Nielsen, O. y Shephard, N. (2001). Non-Gaussian Ornstein-Uhlenbeck-Based Models and Some of Their Uses in Financial Economics, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **63**(2): 167–241.

Congdon, P. (2010). *Applied Bayesian Hierarchical Methods*, CRC Press.

Crainiceanu, C., Ruppert, D. y Wand, M. P. (2005). Bayesian Analysis for Penalized Spline Regression Using WinBUGS, *Journal of Statistical Software* **14**: 1–24.

Eilers, P. y Marx, B. (1996). Flexible Smoothing with B-splines and Penalties, *Statistical Science* **11**: 89–121.

Geraci, M. y Bottai, M. (2007). Quantile Regression for Longitudinal Data using Asymmetric Laplace Distribution, *Biostatistics* **8**: 140–154.

Green, P. J. y Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*, Chapman and Hall, London.

Hastie, T. y Tibshirani, R. (1986). Generalized Additive Models, *Statistical Science* **1**(3): 297–318.

Huaraz, D. (2012). *Inferencia Bayesiana en el Modelo de Regresión Spline Penalizado con una aplicación a los Tiempos en cola de una agencia bancaria*, Tesis de Maestría, Pontificia Universidad Católica del Perú.

Koenker, R. (2005). *Quantile Regression*, Cambridge University Press,New York.

Koenker, R. y Basset, G. (1978). Regression Quantiles, *Econometrica* **46**(1): 33–50.

Koenker, R. y Hallock, K. (2001). Quantile Regression, *Journal of Economic Perspectives* **15**(4): 143–156.

Koenker, R. y Machado, J. (1999). Goodness of Fit and Related Inference Processes for Quantile Regression, *Journal of the American Statistical Association* **94**(448): 1296–1310.

Kotz, S., Kozubowski, T. y Podgórski, K. (2001). *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*, Birkhäuser.

Kozumi, H. y Kobayashi, G. (2011). Gibbs Sampling Methods for Bayesian Quantile Regression, *Journal of Statistical Computation and Simulation* **81**: 1565–1578.

MTPE (2006). Análisis de la distribución del ingreso laboral en lima metropolitana, 1990-2004, *Boletín de Economía Laboral 33*, PEEL-Ministerio de trabajo y promoción del empleo.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
**URL:** *http://www.R-project.org*

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

Ruppert, D., Wand, M. P. y Carroll, R. J. (2009). Semiparametric regression during 2003-2007., *Electronic Journal of Statistics* **3**: 1193–1256.

Spiegelhalter, D. J., Best, N., Carlin, B. y van der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society,Series B.* **64**(4): 583–639.

Spiegelhalter, D. J., Thomas, A., Best, N. G. y Lunn, D. (2007). *WinBUGS User Manual Version 1.4.3*. http://www.mrc-bsu.cam.ac.uk/bugs.

Sturtz, S., Ligges, U. y Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R., *Journal of Statistical Software* **12**: 1–16.

Ullah, A. (1985). Specification Analysis of Econometric Models, *Journal of Quantitative Economics* **2**: 187–209.

Willmott, C. J. y Matsuura, K. (2005). Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance, *Climate Research* **30**: 79–82.

Yu, K. y Moyeed, R. (2001). Bayesian Quantile Regression, *Statistics Probability Letters* **54**: 437–447.

Yu, K. y Zhang, J. (2005). A Three-Parameter Asymmetric Laplace Distribution and Its Extension, *Communications in Statistics-Theory and Methods* **34**: 1867–1879.

Yuan, Y. y Yin, G. (2010). Bayesian Quantile Regression for Longitudinal Studies with Nonignorable missing data, *Biometrics* **66**: 105–114.

Zevallos, A. (2012). *Inferencia Bayesiana en el Modelo de Regresión Cuantílica*, Tesis de Maestría, Pontificia Universidad Católica del Perú.