

INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO



**UN MODELO BAYESIANO Y NO PARAMETRICO DE
REGRESION SOBRE CUANTILES**

TESIS

QUE PARA OBTENER EL TITULO DE
LICENCIADO EN MATEMATICAS APLICADAS

PRESENTA

CARLOS OMAR PARDO GOMEZ

CIUDAD DE MEXICO

2018

INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO



**UN MODELO BAYESIANO Y NO PARAMETRICO DE
REGRESION SOBRE CUANTILES**

TESIS

QUE PARA OBTENER EL TITULO DE
LICENCIADO EN MATEMATICAS APLICADAS

PRESENTA

CARLOS OMAR PARDO GOMEZ

ASESOR: DR. JUAN CARLOS MARTINEZ OVANDO

CIUDAD DE MEXICO

2018

Con fundamento en los artículos 21 y 27 de la Ley Federal del Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada **"UN MODELO BAYESIANO Y NO PARAMÉTRICO DE REGRESIÓN SOBRE CUANTILES"**, otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la Biblioteca Raúl Baillères Jr., la autorización para que fijen la obra en cualquier medio, incluido el electrónico, y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por tal divulgación una contraprestación.

CARLOS OMAR PARDO GÓMEZ

FECHA

FIRMA

A Mago.

Agradecimientos

¡Muchas gracias a todos!

Prefacio

El tema de esta tesis es describir un modelo de regresión sobre cuantiles, debido a diversas bondades que presenta sobre el tradicional análisis de regresión a la media. El modelo tiene una construcción probabilística realizada a través del paradigma Bayesiano y presenta una gran flexibilidad, al contar con componentes no paramétricos. En este trabajo también se describen los modelos tradicionales de regresión, con la intención de entender las áreas de oportunidad que son atacadas por el nuevo modelo.

El capítulo 1 describe la importancia de las aproximaciones distintas al modelo tradicional de regresión (a la media, lineal y con error normal), así como la evolución histórica de los modelos alternativos. El capítulo 2 introduce al paradigma Bayesiano y sus fundamentos generales. El capítulo 3 se centra en los modelos Bayesianos tradicionales de regresión, tanto a la media, como sobre cuantiles. El capítulo 4 plantea como alternativa especificaciones no paramétricas para la tendencia y el error. El capítulo 5 define el modelo propuesto por esta tesis y explica el algoritmo necesario para realizar inferencia y predicción. El capítulo 6 muestra aplicaciones del modelo sobre conjuntos de datos simulados. Finalmente, el capítulo 7 hace referencia a las conclusiones de esta tesis, además de describir el trabajo futuro que se podría desarrollar.

Índice general

1. Introducción	6
2. Paradigma Bayesiano	10
2.1. Inferencia de variables aleatorias	10
2.2. Propiedad conjugada	15
2.3. Inferencia con variables explicativas	16
3. Modelos de regresión	18
3.1. Concepto de regresión	18
3.2. Regresión a la media	19
3.2.1. Planteamiento general	19
3.2.2. Modelo tradicional	20
3.3. Regresión sobre cuantiles	23
3.3.1. Planteamiento general	23
3.3.2. Modelo tradicional	24
4. Especificación no paramétrica	29
4.1. Motivación	29
4.2. Distribución de f_p , mediante procesos Gaussianos	32
4.2.1. Introducción a los procesos Gaussianos	32
4.2.2. Definiciones y notación	34

4.2.3.	Funciones de covarianza	36
4.2.4.	Predicción	38
4.3.	Distribución de ε_p , mediante procesos de Dirichlet	39
4.3.1.	Definición de los procesos de Dirichlet	39
4.3.2.	Distribución posterior	41
4.3.3.	Distribución predictiva	42
4.3.4.	Proceso estocástico de rompimiento de un palo . . .	43
4.3.5.	Modelo general de mezclas infinitas de Dirichlet . . .	44
4.3.6.	Modelo de mezclas infinitas de Dirichlet para la distribución asimétrica de Laplace	45
5.	Modelo GPDP para regresión sobre cuantiles	48
5.1.	Definición	48
5.2.	Inferencia con el simulador de Gibbs	50
5.2.1.	Actualización del error	51
5.2.2.	Actualización de la tendencia	53
5.3.	Predicción	55
5.4.	Hiper-parámetros iniciales del modelo	57
5.4.1.	Función de medias m	57
5.4.2.	<i>Gamma-Inversas</i> de λ y el Proceso de Dirichlet . . .	58
5.4.3.	Parámetro de concentración α	59
5.5.	Consideraciones sobre la bondad de ajuste	60
5.6.	Paquete <i>GPDPQuantReg</i> en R	61
6.	Aplicaciones	63
6.1.	Metodología de simulación de datos y ajuste de los modelos	63
6.2.	Conjuntos de datos simulados	67
6.2.1.	Supuestos tradicionales de regresión a la media . . .	67
6.2.2.	Error de colas pesadas	72
6.2.3.	Heterocedasticidad	76
6.2.4.	Error asimétrico	80

6.2.5. Discontinuidades	84
6.3. Comparación general entre modelos	88
7. Conclusiones y trabajo futuro	92
Bibliografía	95
A. Distribuciones de probabilidad	98
A.1. Distribución Normal condicional	98
A.2. Distribución de Dirichlet	99
B. Algoritmos MCMC	101
B.1. Introducción	101
B.2. Simulador de Gibbs	102
B.2.1. Simulador de Gibbs de dos pasos	102
B.2.2. Simulador de Gibbs de múltiples pasos	102
B.3. Monitoreo de convergencia y adaptación de los algoritmos	
MCMC	104
B.3.1. Monitoreo de convergencia a la <i>estacionariedad</i> . . .	104
B.3.2. Monitoreo de convergencia a los promedios	104
B.3.3. Monitoreo de convergencia a una muestra <i>iid</i>	105
C. Predicción de cuantiles, utilizando el modelo tradicional de	
regresión a la media	106

Capítulo 1

Introducción

Detrás de cualquier modelo de regresión la intención es entender alguna característica asociada con una variable aleatoria, en función de un conjunto de variables potencialmente explicativas o predictivas para tal característica. Ha sido común resumir esta dependencia mediante alguna medida de tendencia central, condicionada a los valores de las covariables.

La medida de tendencia central tradicionalmente usada ha sido la media, dando lugar a los modelos de *regresión a la media*, con sus variantes lineal y no lineal, simple y múltiple, con error normal y no normal. Este tipo de modelos tiene un buen número de ventajas, entre las que destacan el bajo costo de estimación y la facilidad de interpretación de sus parámetros (principalmente cuando es lineal). Sin embargo, como mencionan Hao & Naiman (2007), tienen tres grandes limitaciones, que a continuación se mencionan.

La primera es que, generalmente, la estimación del modelo busca minimizar la diferencia entre el valor esperado teórico y la media observada. Por lo

tanto, la inferencia sobre los valores lejanos a la media, que suelen ser de interés en ciertos contextos, como los seguros o las finanzas, puede ser inexacta.

La segunda es que algunos fenómenos de estudio tienen distribuciones de colas pesadas, principalmente en las ciencias sociales. Esto da lugar a valores atípicos, mismos que pueden sesgar estimaciones de la media, mientras prácticamente no afectan a otros estadísticos, como la mediana.

La tercera es que al focalizar la intervención de las variables explicativas únicamente en la media, y darle una distribución predefinida al error, se suelen dejar de lado características propias del fenómeno que se está estudiando. Por ejemplo, cada cuantil, además de diferir en el nivel, podría tener una forma funcional distinta.

Debido a esto, desde mitades del siglo XVIII han surgido alternativas a este tipo de modelos. Según Hao & Naiman (2007), la primera conocida data de 1760, cuando el jesuita croata Rudjer Josip Boscovich visitó Londres en búsqueda de consejo computacional para su novedoso modelo de *regresión a la mediana*. De nueva cuenta se buscó una medida de tendencia central, pero con otras bondades. Por ejemplo, ser una mejor medida informativa para distribuciones asimétricas y menos susceptible a valores atípicos.

Así como los modelos de regresión a la media son comúnmente relacionados con la minimización de los errores cuadráticos, los modelos de regresión a la mediana lo son con la minimización de los errores absolutos. Debido a la no diferenciabilidad, tuvieron que pasar muchos años para que lograran ser viables, hasta que el poder computacional y los algoritmos de programación lineal lo permitieron.

Cabe recordar que, a grandes rasgos y sin dar aún una definición formal, el

cuantil p -ésimo es aquel valor tal que el $p \times 100\%$ de los valores están por debajo de él, y el $(1 - p) \times 100\%$, por encima. Así, la mediana es un caso particular de un cuantil, específicamente el 0.5-ésimo. Esto abre la idea de que otros cuantiles también podrían ser modelados en función de variables explicativas y no necesariamente tienen que ser una medida de tendencia central.

Los *modelos de regresión sobre cuantiles* fueron introducidos por Koenker & Bassett (1978), y han permitido concentrarse en la estimación de valores de interés para los modeladores, sin importar que estén alejados de la media. Además, el cálculo de diversos cuantiles para un mismo fenómeno ha permitido entender mejor la forma y propiedades de las distribuciones condicionales de la variable de respuesta.

El paradigma Bayesiano ha desarrollado este tipo de modelos de forma posterior a otros paradigmas. Walker & Mallick (1999), Kottas & Gelfand (2001) y Hanson & Johnson (2002) desarrollaron modelos para la mediana, suponiendo una distribución no paramétrica del error. Yu & Moyeed (2001) y Tsionas (2003) desarrollaron inferencia paramétrica, basados en la distribución asimétrica de Laplace para los errores. Por otro lado, Lavine (1995) y Dunson & Taylor (2005) usaron una perspectiva distinta y propusieron una aproximación de la verosimilitud para cuantiles.

Por la misma naturaleza compleja del nuevo desafío, los modelos de regresión sobre cuantiles han retomado un concepto que también ha tomado auge en los de regresión a la media: la distribución no paramétrica de los errores, misma que generaliza la idea tradicional del error normal.

Desafortunadamente, a diferencia de los modelos de regresión a la media, que ya han propuesto alternativas para este tema, los modelos Bayesianos de regresión sobre cuantiles han hecho muy poco por romper con la relación

lineal en los parámetros entre la variable de respuesta y las covariables. Y casi todos aquellos que lo han logrado, han tenido que recurrir a estimaciones no probabilísticas o no Bayesianas, para resolver alguna parte del problema.

Esta tesis tiene la finalidad de sustituir simultáneamente las ideas tradicionales de regresión a la media, linearidad y distribución normal de los errores, por un enfoque más flexible, y que se mantenga totalmente probabilístico. Para ello, rescata las ideas de Kottas *et al.* (2007) y Kottas & Krnjajic (2005), proponiendo un modelo Bayesiano y no paramétrico, útil en el contexto de regresión sobre cuantiles.

Capítulo 2

Paradigma bayesiano^{1,2}

2.1. Inferencia de variables aleatorias

Un problema clásico de la estadística es el de hacer predicción, utilizando la información de los datos que ya han sido observados. Por ejemplo, es posible pensar que ya se tiene el conjunto de n datos observados $\{y_1, \dots, y_n\}$ y se desea hacer predicción acerca del valor del dato y_{n+1} , que aún no ha sido observado. Para esto, se podría usar la probabilidad condicional

$$\mathbb{P}(y_{n+1}|y_1, \dots, y_n) = \frac{\mathbb{P}(y_{n+1} \cap \{y_1, \dots, y_n\})}{\mathbb{P}(y_1, \dots, y_n)} = \frac{\mathbb{P}(y_1, \dots, y_n, y_{n+1})}{\mathbb{P}(y_1, \dots, y_n)},$$

¹Las ideas de este capítulo son retomadas de Denison *et al.* (2002).

²Esta tesis da como aceptados los axiomas de coherencia de la Teoría de la Decisión, mismos que pueden ser encontrados, por ejemplo, en Fishburn (1986). Por lo tanto, entiende al paradigma Bayesiano como el coherente para hacer estadística, cuando una toma de decisión con incertidumbre es el objetivo final del estudio.

pero esto requeriría conocer la función conjunta, misma que puede ser compleja por la estructura de dependencia de los datos.³

Este problema puede ser abordado mediante el uso del Teorema de representación general de de Finetti. Para ello, antes se dará una definición.

Definición. Sea (y_1, y_2, \dots) , una sucesión de variables aleatorias, cuya distribución de probabilidad conjunta está dada por $\mathbb{P}(y_1, y_2, \dots)$. Sea ψ una función biyectiva que crea una permutación finita del conjunto $\{1, 2, \dots\}$, es decir, permuta un número finito de elementos y al resto los deja fijos. Se dice entonces que (y_1, y_2, \dots) es una **sucesión aleatoria infinitamente intercambiable** si se cumple que

$$\mathbb{P}(y_1, y_2, \dots) = \mathbb{P}(y_{\psi(1)}, y_{\psi(2)}, \dots),$$

para cualquier permutación ψ .

En pocas palabras, una sucesión (y_1, y_2, \dots) se considerará infinitamente intercambiable si el orden en que se etiquetan las variables no afecta su distribución conjunta. Es importante hacer notar que la comúnmente usada independencia implica intercambiabilidad, pero lo contrario no se cumple. Es decir, la intercambiabilidad es un supuesto menos rígido que la independencia.

Dicho esto, es momento de plantear el **Teorema de representación general de de Finetti**.⁴

³En este trabajo se usará la notación \mathbb{P} como una forma general de definir una medida de probabilidad, independientemente de los detalles teóricos sobre análisis y medibilidad.

⁴Una demostración de este teorema puede ser encontrada en Schervish (1996).

Teorema. *Sea (y_1, y_2, \dots) una sucesión aleatoria infinitamente intercambiable de valores reales. Entonces existe una distribución de probabilidad F sobre \mathcal{F} , el espacio de todas las distribuciones, de forma que la probabilidad conjunta de (y_1, y_2, \dots) se puede expresar como*

$$\mathbb{P}(y_1, y_2, \dots) = \int_{\mathcal{F}} \left[\prod_{k=1}^{\infty} \mathbb{P}(y_k | G) \right] dF(G),$$

con

$$F(G) = \lim_{n \rightarrow \infty} F(G_n),$$

donde $F(G_n)$ es una función de distribución evaluada en la función de distribución empírica definida por

$$G_n(y) = \frac{1}{n} \sum_{i=1}^n I(y_i \leq y).$$

En otras palabras, el Teorema de de Finetti dice que $\{y_1, y_2, \dots\}$ es un conjunto de variables aleatorias condicionalmente independientes, dado que se supone provienen de la distribución G . Al mismo tiempo, G es una realización de otra variable aleatoria cuya distribución está dada por $F(G)$.

Cabe hacer notar que dicho teorema plantea la distribución conjunta de (y_1, y_2, \dots) como una mezcla de verosimilitudes condicionalmente independientes en G , donde el peso asociada a cada una depende de $F(G)$. Por lo tanto, $F(G)$ expresa la incertidumbre acerca de cuán probable es que G sea idónea para explicar el fenómeno, aún sin observar los datos.

Un subconjunto del espacio de todas las distribuciones \mathcal{F} es el espacio de las distribuciones paramétricas, es decir, aquellas que pueden ser descritas en su totalidad únicamente señalando el valor de un vector de parámetros

θ de tamaño finito, mismo que puede tomar valores en todo un soporte Θ ⁵. Por lo tanto, si se hace el supuesto adicional que la distribución marginal de y_i es paramétrica, con vector de parámetros desconocido θ , se obtiene como corolario del Teorema de de Finetti que

$$\mathbb{P}(y_1, y_2, \dots) = \int_{\Theta} \left[\prod_{k=1}^{\infty} \mathbb{P}(y_k | \theta) \right] \mathbb{P}(\theta) d\theta.$$

Siguiendo el razonamiento anterior, $\mathbb{P}(\theta)$ refleja la incertidumbre del modelador acerca de cuál es el vector θ que originó los datos, antes de observarlos.

La intuición detrás de este resultado es que, al igual que en otros paradigmas, se supone a un sólo vector θ como aquel del que provinieron los datos, pero es desconocido, y la tarea es estimarlo. Una particularidad del paradigma Bayesiano es expresar la incertidumbre que tiene el modelador acerca del valor verdadero mediante la asignación de una distribución de probabilidad para θ , sujeta a la información inicial o conocimiento previo a observar los datos (CP) que se tenga del fenómeno. Es decir, $\mathbb{P}(\theta|CP)$. Como una simplificación de la notación, en la literatura normalmente se escribe $\mathbb{P}(\theta) = \mathbb{P}(\theta|CP)$ y se conoce como la *probabilidad inicial* del parámetro.

Así, el Teorema de representación general garantiza que para variables aleatorias infinitamente intercambiables existe una distribución del vector de parámetros, tal que la probabilidad conjunta se puede expresar como la verosimilitud de variables condicionalmente independientes, dado un vector de parámetros, multiplicada por la probabilidad de que dicho vector de

⁵En lo que resta de esta sección las proposiciones y resultados harán referencia al conjunto de distribuciones paramétricas, apelando a que así los nuevos conceptos tengan mayor claridad para el lector, considerando que este tipo de distribuciones son con las que se suele estar más familiarizado. Todos serán generalizables al conjunto de distribuciones \mathcal{F} .

parámetros sea del que efectivamente provinieron las observaciones. Con el teorema se prueba la existencia de dicha representación, mas no su unicidad. En los siguientes párrafos se hará uso de ella para realizar inferencia en variables aleatorias.

Regresando al problema inicial, y bajo los supuestos recién mencionados, es posible escribir

$$\mathbb{P}(y_{n+1}|y_1, \dots, y_n) = \int_{\Theta} \mathbb{P}(y_{n+1}|\theta) \mathbb{P}(\theta|y_1, \dots, y_n) d\theta,$$

donde a su vez, usando el **Teorema de Bayes**, se obtiene que

$$\mathbb{P}(\theta|y_1, \dots, y_n) = \frac{\mathbb{P}(y_1, \dots, y_n|\theta) \times \mathbb{P}(\theta)}{\mathbb{P}(y_1, \dots, y_n)},$$

que en el paradigma Bayesiano se conoce como la *probabilidad posterior* del parámetro.

Se puede observar que el denominador es constante respecto a θ , por lo que usualmente la probabilidad condicional del vector de parámetros no se expresa como una igualdad, sino con la expresión

$$\mathbb{P}(\theta|y_1, \dots, y_n) \propto \mathbb{P}(y_1, \dots, y_n|\theta) \times \mathbb{P}(\theta),$$

y sólo difiere de la igualdad por una constante que permita que, al integrar sobre todo el soporte de θ , el resultado sea igual a 1.

Cabe resaltar que debido a la independencia condicional dada por el Teorema de de Finetti, el factor de verosimilitud puede ser reescrito como

$$\mathbb{P}(y_1, \dots, y_n|\theta) = \prod_{i=1}^n \mathbb{P}(y_i|\theta).$$

Por lo tanto, se confirma que el aprendizaje en el paradigma Bayesiano se

obtiene como

$$Posterior \propto Verosimilitud \times Inicial.$$

Es decir, el conocimiento final surge de conjuntar el conocimiento inicial con la información contenida en los datos.

Es importante notar que bajo este enfoque se obtiene una distribución de probabilidad para un valor futuro de y_{n+1} . Ésta se puede utilizar para el cálculo de predicciones puntuales o intervalos (que en el caso del paradigma Bayesiano son de *probabilidad*), mediante el uso de la Teoría de la Decisión y funciones de utilidad o pérdida.

2.2. Propiedad conjugada

En los casos en los que la distribución posterior de los parámetros tiene la misma forma funcional que la distribución inicial, se dice que la distribución de los parámetros pertenece a una **familia paramétrica conjugada**, para la verosimilitud respectiva.

Esta propiedad es conveniente, porque permite a la distribución posterior tener forma analítica cerrada, evitando tener que usar métodos numéricos para aproximarla. Además permite ver de forma más clara cómo afectan los datos a la actualización, respecto a la distribución inicial.

Sin embargo, el rango de posibles modelos conjugados puede resultar limitado en algunos contextos prácticos debido a que el fenómeno en estudio puede ser mejor representado con otras distribuciones, que usualmente no pertenecen a familias conjugadas.

2.3. Inferencia con variables explicativas⁶

Como se verá en el siguiente capítulo de esta tesis, un problema común es estimar la distribución de cierta sucesión de variables aleatorias (y_1, y_2, \dots) , condicionada a las observaciones (x_1, x_2, \dots) de otras variables usualmente llamadas explicativas o predictivas (cada x_i es un vector de dimensión finita). En este caso, la sucesión (y_1, y_2, \dots) ya no es intercambiable, porque cada valor y_i depende del valor de su respectiva x_i , por lo que no es posible aplicar de manera directa el Teorema de de Finetti. Para hacerlo de manera indirecta, se introducirá el concepto de intercambiabilidad parcial.

Definición. Sea (y_1, y_2, \dots) dada (x_1, x_2, \dots) , una sucesión numerable de variables aleatorias, asociada con los correspondientes valores de sus variables explicativas, y cuya distribución de probabilidad conjunta está dada por $\mathbb{P}(y_1, y_2, \dots | x_1, x_2, \dots)$. Sea ψ una función biyectiva que crea una permutación finita de un conjunto, es decir, permuta un número finito de elementos y al resto los deja fijos. Sean $\tilde{x}_1, \tilde{x}_2, \dots$ los distintos valores únicos que toman las x 's, y y_{k_i} una y , tal que el valor de su correspondiente variable explicativa es \tilde{x}_k . Entonces, se dice que (y_1, y_2, \dots) es una **sucesión aleatoria infinita y parcialmente intercambiable** si se cumple que

$$\mathbb{P}(y_{k_1}, y_{k_2}, \dots | \tilde{x}_k) = \mathbb{P}(y_{\psi(k_1)}, y_{\psi(k_2)}, \dots | \tilde{x}_k),$$

para cualquier permutación ψ y para todos los diferentes valores k .

Es decir, todas aquellas y 's cuyas x 's tienen el mismo valor, son infinitamente intercambiables entre sí.

⁶Esta sección carece de una formalidad completa, pero busca darle la intuición al lector para generalizar el resultado del Teorema de de Finetti en el contexto en el que se desarrollará este trabajo. Para una explicación más formal, consultar Dawid (2016).

Si además se cumpliera que el orden de los valores únicos \tilde{x} 's es intercambiable en el sentido de de Finetti, entonces, intuitivamente se podría tomar la G del Teorema de de Finetti como dependiente de las x_i 's, $G(x_i)$, y se obtendría que

$$\mathbb{P}(y_1, y_2, \dots | x_1, x_2, \dots) = \int_{\mathcal{F}} \left[\prod_{i=1}^{\infty} \mathbb{P}(y_i | G(x_i)) \right] dF(G(x_1, x_2, \dots)),$$

donde las y 's resultarían ser independientes entre sí, condicionadas a una distribución que depende de la x_i asociada.

El tema del siguiente capítulo será la discusión de métodos de inferencia sobre las variables y , dentro de este contexto específico de dependencia de una variable explicativa o predictiva x , normalmente conocidos como **modelos de regresión**.

Capítulo 3

Modelos de regresión

3.1. Concepto de regresión

Los modelos de regresión tienen como objetivo describir la distribución de una variable aleatoria $y \in \mathbb{R}$, generalmente llamada *variable de respuesta*, condicional a los valores de las variables $x \in \mathbb{R}^n$, conocidas como *covariables*, *variables explicativas* o *variables predictivas*. Visto en términos matemáticos, se puede expresar como

$$y|x \sim \mathbb{P}(y|x).$$

Si bien esta relación se da por hecha y se supone un vínculo concreto del que provinieron los datos, normalmente dicho vínculo es desconocido. Por lo tanto, la intención de estos modelos es realizar alguna estimación de él. Dado que es complicado aproximar con exactitud toda la distribución, comúnmente se utilizan un número finito de parámetros para describirla.

Además, la interpretación de dichos parámetros suele tener relevancia para el modelador, como es el caso de la media o la mediana.

También es importante resaltar que este trabajo tiene interés en modelar a y , dado que ya se observaron los valores de x . Sin embargo, se podría pensar en modelos donde tenga sentido la distribución conjunta de y y x , misma que se podría obtener como

$$\mathbb{P}(y, x) = \mathbb{P}(y|x) \times \mathbb{P}(x).$$

3.2. Regresión a la media

3.2.1. Planteamiento general

La *regresión a la media* es el caso más usado dentro de los modelos de regresión, tanto en el paradigma Bayesiano, como en otros. Esto sucede debido al bajo uso de recursos que requiere su estimación. En el caso particular Bayesiano, diversas familias conjugadas tienen sentido dentro de este contexto; y en otros paradigmas, también suelen ser utilizados debido a que se asocian con la minimización de términos cuadrados, con las bondades de diferenciabilidad que eso implica.

En notación probabilística, retomando el hecho de que $y|x \sim \mathbb{P}(y|x)$, el modelo de regresión a la media busca aproximar a la función f , tal que

$$f(x) = \mathbb{E}[y|x].$$

Para hacer esto, normalmente se vale del supuesto que

$$y = f(x) + \varepsilon,$$

con $\varepsilon \in \mathbb{R}$ (denominado comúnmente como el *error aleatorio*), una variable aleatoria independiente de x , tal que $\mathbb{E}[\varepsilon] = 0$. Además, $f : \mathbb{R}^n \rightarrow \mathbb{R}$ es desconocida, pero fija. Así, debido a que la esperanza de la suma es la suma de las esperanzas, se cumple que $\mathbb{E}[y|x] = f(x)$.

Además, se supone independencia entre los errores. Por lo tanto, sean \tilde{x} las covariables asociadas a la variable de respuesta \tilde{y} , y \hat{x} las asociadas a \hat{y} , se tiene que $\tilde{y}|\tilde{x}, f$ es condicionalmente independiente a $\hat{y}|\hat{x}, f$.

3.2.2. Modelo tradicional ⁷

La *regresión lineal a la media* es el caso particular más usado en el contexto de regresión a la media. Consiste en definir

$$f(x) = x^T \beta,$$

donde $\beta \in \mathbb{R}^n$ se piensa con valores constantes, pero desconocidos, y la tarea es estimarlos. Además, se supone error normal, es decir, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, tal que el parámetro de varianza fijo σ^2 también debe ser estimado.

Cabe señalar que una gran ventaja de plantear de esta manera el modelo, y por la que ha sido ampliamente usado, es la interpretación que se le puede dar al vector de parámetros β , como cuánto aumenta o disminuye el valor de la variable de respuesta, cuando se aumenta o disminuye una unidad de la variable explicativa.

Sea $\{(x_i, y_i) | x_i \in \mathbb{R}^n, y_i \in \mathbb{R}, i \in \{1, \dots, m\}\}$ el conjunto de datos observados de las variables de respuesta y de las covariables. Es posible representar este

⁷Algunas ideas de esta subsección son retomadas de Denison *et al.* (2002) y Bannerjee (2008).

mismo conjunto con la notación matricial $\{X, Y | X \in \mathbb{R}^{m \times n}, Y \in \mathbb{R}^m\}$.⁸ Sea $\mathcal{E} \in \mathbb{R}^m$ el vector de errores aleatorios, tal que $\mathcal{E} \sim \mathcal{N}(0, \sigma^2 I)$. El modelo se puede reescribir como

$$\begin{aligned} Y &= X\beta + \mathcal{E} \\ \implies Y | X, \beta, \sigma^2 &\sim \mathcal{N}(X\beta, \sigma^2 I). \end{aligned}$$

De esta manera, se tiene una distribución para el vector aleatorio Y , dada la matriz X , que para poder ser expresada, requiere de la estimación de β y σ^2 . Para hacer esto, el enfoque Bayesiano le asigna una distribución inicial de probabilidad a ambos parámetros, reflejando la incertidumbre que tiene el modelador acerca de su valor real. Es decir, se tiene que

$$\beta, \sigma^2 \sim \mathbb{P}(\beta, \sigma^2).$$

Por el Teorema de Bayes,

$$\begin{aligned} \mathbb{P}(\beta, \sigma^2 | Y, X) &= \frac{\mathbb{P}(Y | X, \beta, \sigma^2) \times \mathbb{P}(\beta, \sigma^2 | X)}{P(Y | X)} \\ &= \frac{\mathbb{P}(Y | X, \beta, \sigma^2) \times \mathbb{P}(\beta, \sigma^2)}{\mathbb{P}(Y | X)} \\ &\propto \mathbb{P}(Y | X, \beta, \sigma^2) \times \mathbb{P}(\beta, \sigma^2), \end{aligned}$$

donde $\mathbb{P}(Y | X, \beta, \sigma^2)$ es la verosimilitud de los datos observados, es decir, $\mathcal{N}(X\beta, \sigma^2 I) = \prod_{i=1}^m \mathcal{N}(x_i^T \beta, \sigma^2)$.

Por conveniencia analítica, hay una distribución inicial comúnmente usada

⁸Por simplificación y limpieza de notación en este trabajo se escribirán de igual manera variables aleatorias y los datos en efecto observados, considerando que en cada caso el contexto será suficiente para saber de cuál se está hablando. Por otro lado, se asociarán las letras minúsculas a una única observación y las mayúsculas a una matriz de observaciones.

para los parámetros β y σ^2 debido a que es conjugada respecto a la distribución normal de los datos. Su nombre es *Normal-Gamma Inversa (NGI)* y se dice que $\beta, \sigma^2 \sim \mathcal{NGI}(M, V, a, b)$, si

$$\begin{aligned}\mathbb{P}(\beta, \sigma^2) &= \mathbb{P}(\beta|\sigma^2) \times \mathbb{P}(\sigma^2) \\ &= \mathcal{N}(\beta|M, \sigma^2 V) \times \mathcal{GI}(\sigma^2|a, b) \\ &\propto (\sigma^2)^{-(a+(n/2)+1)} \exp\left(-\frac{(\beta - M)^T V^{-1}(\beta - M) + 2b}{2\sigma^2}\right),\end{aligned}$$

donde M es la media inicial de los coeficientes, $\sigma^2 V$ la matriz de varianzas y covarianzas, y a, b son los parámetros iniciales de forma y escala, respectivamente, de σ^2 .

Aprovechando la propiedad conjugada, es posible escribir la densidad posterior de los parámetros como

$$\begin{aligned}\mathbb{P}(\beta, \sigma^2|Y, X) &\propto \mathbb{P}(Y|X, \beta, \sigma^2) \times \mathbb{P}(\beta, \sigma^2), \\ &\propto (\sigma^2)^{-(\bar{a}+(n/2)+1)} \exp\left(-\frac{(\beta - \bar{M})^T \bar{V}^{-1}(\beta - \bar{M}) + 2\bar{b}}{2\sigma^2}\right),\end{aligned}$$

donde

$$\begin{aligned}\bar{M} &= (V^{-1} + X^T X)^{-1}(V^{-1} M + X^T Y), \\ \bar{V} &= (V^{-1} + X^T X)^{-1}, \\ \bar{a} &= a + m/2, \\ \bar{b} &= b + \frac{\bar{M}^T V^{-1} M + Y^T Y - \bar{M}^T \bar{V}^{-1} \bar{M}}{2}.\end{aligned}$$

Es decir, la distribución posterior de (β, σ^2) es *Normal - Gamma Inversa*, con parámetros $\mathcal{NGI}(\bar{M}, \bar{V}, \bar{a}, \bar{b})$.

3.3. Regresión sobre cuantiles

3.3.1. Planteamiento general

Cuando surgió entre la comunidad estadística el problema de regresión sobre cuantiles, inicialmente fue modelado bajo un enfoque no Bayesiano, como se describe en Yu & Moyeed (2001). Posteriormente, Koenker & Bassett (1978) retomaron esas ideas, y las aplicaron en el paradigma Bayesiano.

La *regresión sobre cuantiles* es una alternativa que se ha desarrollado recientemente y que permite enfocarse en aspectos alternativos de la distribución, por ejemplo, centrar la estimación en estadísticos que son valores extremos, o en algún decil de interés.

Definición 1. Sea F_y la función de distribución acumulada de la variable aleatoria y , entonces la función que regresa su cuantil p -ésimo se escribe

$$q_p(y) = \inf \{x \in \mathbb{R} : p \leq F_y(x)\};$$

que se puede simplificar a

$$q_p(y) = F_y^{-1}(p),$$

cuando F_y es continua y estrictamente creciente en el soporte de y .

Dicho de manera más intuitiva, si se tiene un conjunto grande de realizaciones de una variable aleatoria y , se esperará que el $p \times 100\%$ esté por debajo de $q_p(y)$ y el $(1-p) \times 100\%$ esté por arriba. Por ejemplo, la mediana es un caso particular de un cuantil, específicamente el 0.5-ésimo.

Sea el p -ésimo cuantil aquel predefinido como el de interés para el modelador. En notación probabilística, se buscará aproximar a la función f_p , tal que

$$f_p(x) = q_p(y|x),$$

para $p \in (0, 1)$ arbitrario y fijo.

Para hacer esto, normalmente se valdrá del supuesto que

$$y = f_p(x) + \varepsilon_p,$$

con $\varepsilon_p \in \mathbb{R}$, una variable aleatoria tal que $q_p(\varepsilon_p) = 0$. Asimismo, $f_p : \mathbb{R}^n \rightarrow \mathbb{R}$ es desconocida, pero fija. Por ello, debido a que el cuantil de una suma es la suma de los cuantiles, se cumple que $q_p(y|x) = f_p(x)$.

Al igual que con la regresión a la media, se supone independencia entre los errores, y, por lo tanto, hay independencia condicional entre las observaciones.

3.3.2. Modelo tradicional

Al igual que en la regresión a la media, el primer y más popular modelo ha sido el lineal. Es decir, para $p \in (0, 1)$ fijo y elegido por el modelador, se define

$$f_p(x) = x^T \beta_p,$$

donde β_p es el vector de coeficientes, dependiente de p .

Definición 2. Se define a la función

$$\rho_p(u) = u \times [pI_{(u>0)} - (1-p)I_{(u<0)}] = \begin{cases} up & \text{si } u \geq 0 \\ -u(1-p) & \text{si } u < 0 \end{cases}.$$

Se dice que una variable aleatoria u sigue una distribución asimétrica de Laplace ($u|p, \sigma \sim AL_p(\sigma)$) si su función de densidad se escribe como

$$w_p^{AL}(u|\sigma) = \frac{p(1-p)}{\sigma} \exp \left[-\rho_p \left(\frac{u}{\sigma} \right) \right],$$

con p parámetro de asimetría y σ , de escala.

Para entender mejor las propiedades de dicha distribución, se presentan las siguientes gráficas.

Figura 3.1: Función de densidad de la distribución asimétrica de Laplace, con $\sigma = 1$ y p variable.

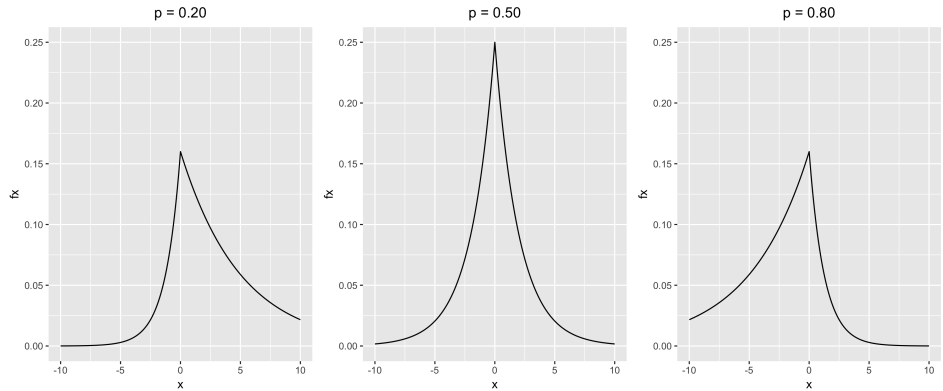
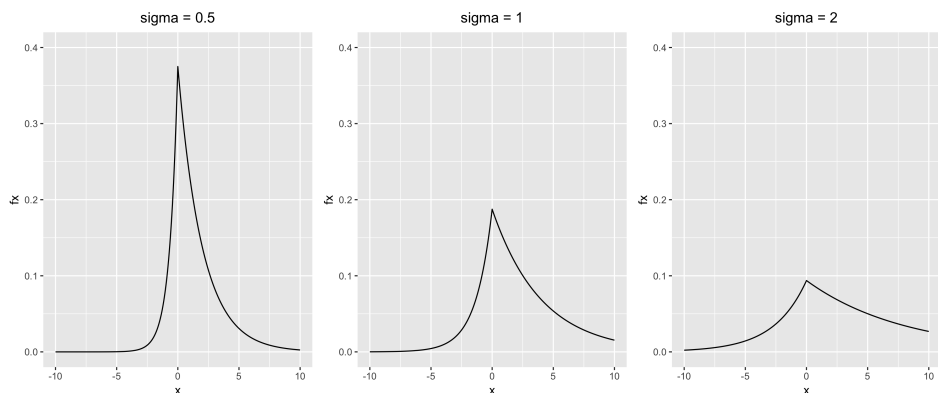


Figura 3.2: Función de densidad de la distribución asimétrica de Laplace, con $p = 0.25$ y σ variable.



Como se puede observar, el parámetro p representa la asimetría de la distribución. Para valores por debajo de 0.5, la distribución está sesgada a la derecha, mientras que para valores superiores a 0.5, presenta sesgo a la izquierda. El único caso en el que es simétrica, es cuando $p = 0.5$.

Por otro lado, el parámetro σ representa la dispersión de la distribución. A menor σ , los datos, aunque sesgados, estarán más concentrados; en cambio, conforme crezca σ , la cola de la distribución será más pesada.

Dicho esto, la propiedad más relevante para este trabajo es que si $\varepsilon_p|\sigma \sim AL_p(\sigma)$, entonces $q_p(\varepsilon_p|\sigma) = 0$, independientemente del valor de σ . Recordando que esa es la única característica que pide el modelo general, el modelo tradicional utiliza esta distribución para explicar la dispersión del error.

Es posible, entonces, reescribir el modelo como

$$y|x, \beta_p, \sigma \sim AL_p(y - x^T \beta_p | \sigma).$$

Sea $\{X, Y | X \in \mathbb{R}^{m \times n}, Y \in \mathbb{R}^m\}$ el conjunto de datos observados. Por el Teorema de Bayes,

$$\mathbb{P}(\beta_p, \sigma | Y, X) \propto \mathbb{P}(Y | X, \beta_p, \sigma) \times \mathbb{P}(\beta_p, \sigma),$$

donde $\mathbb{P}(Y | X, \beta_p, \sigma)$ es la verosimilitud, y debido a la independencia condicional, se puede calcular como

$$\mathbb{P}(Y | X, \beta_p, \sigma) = \prod_{i=1}^m AL_p(y_i - x_i^T \beta_p | \sigma).$$

Por otro lado, $\mathbb{P}(\beta_p, \sigma^2)$ es la distribución inicial de los parámetros, para los que normalmente se usa

$$\beta_p, \sigma \sim \mathcal{NGI}(M, V, a, b).$$

A diferencia del modelo tradicional de regresión a la media, este modelo no es conjugado. Por lo tanto se requieren métodos computacionales (como los que serán descritos en el capítulo 5) para aproximar la distribución posterior.

Además, es importante resaltar que, al igual que el modelo de regresión sobre la media, la salida de este modelo es una distribución completa para y , dado el valor de las covariables x . Ambos modelos difieren únicamente en cuáles son los parámetros a aproximar y que son suficientes para expresar en su totalidad a la distribución respectiva.

Si bien este modelo de regresión sobre cuantiles representa una buena alternativa, aún queda la posibilidad de retomar estas ideas y crear modelos más flexibles, que capturen con mayor precisión las particularidades de cada fenómeno y la interacción entre las variables de salida y las explicativas. En el siguiente capítulo se discutirá la importancia de capturar mayor complejidad en la distribuciones mediante el de uso de métodos no paramétricos.

Capítulo 4

Especificación no paramétrica

4.1. Motivación

En el capítulo anterior se analizaron los modelos tradicionales de regresión, tanto a la media, como sobre cuantiles, para una variable de respuesta y , dado un cierto conjunto de covariables x . Si bien tienen muchas ventajas, es relevante no olvidar que cuentan con un supuesto fuerte: la relación entre la variable dependiente y y las variables independientes x únicamente se da de forma lineal en los parámetros. Pero las funciones lineales sólo son un subconjunto de las funciones existentes. Por ello, valdría la pena analizar si es posible relajar este supuesto y tener un modelo más general.

Una idea inicial para darle la vuelta es redefinir variables, de tal manera que se pueda obtener un polinomio. Por ejemplo, suponga que \hat{x} es un buen

predictor de la media de y , pero como polinomio de orden 3, es decir,

$$y = \beta_0 + \beta_1 \dot{x} + \beta_2 \dot{x}^2 + \beta_3 \dot{x}^3 + \varepsilon.$$

Entonces, se puede definir el vector x de covariables como $x = (1, \dot{x}, \dot{x}^2, \dot{x}^3)$ y aplicar las técnicas de regresión lineal antes mencionadas.

Otra limitación con la que podría contar el modelo tradicional es que no tome en cuenta la interacción entre variables como información relevante. Para ejemplificar esto, se podría pensar en que los datos provengan de

$$y = \beta_0 + \beta_1 \dot{x}_1 + \beta_2 \dot{x}_2 + \beta_3 \dot{x}_1 \dot{x}_2 + \varepsilon.$$

Es posible llevar dicha representación al modelo tradicional, declarando al vector x de variables de entrada como $x = (1, \dot{x}_1, \dot{x}_2, \dot{x}_1 \dot{x}_2)$, y el procedimiento sería análogo.

Y aún es posible dar un siguiente paso, saliendo del terreno de los polinomios y entrando en el de las funciones biyectivas. Se podría pensar en un caso como el siguiente, donde $\dot{y}, \dot{x}_1, \dot{x}_2 > 0$.

$$\begin{aligned} \dot{y} &= \dot{\beta}_0 \dot{x}_1^{\beta_1} \dot{x}_2^{\beta_2} e^{\varepsilon} \\ \iff \ln(\dot{y}) &= \ln(\dot{\beta}_0) + \beta_1 \ln(\dot{x}_1) + \beta_2 \ln(\dot{x}_2) + \varepsilon \\ \iff y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \end{aligned}$$

con

$$\begin{aligned} y &= \ln(\dot{y}), \\ \beta_0 &= \ln(\dot{\beta}_0), \\ x_1 &= \ln(\dot{x}_1), \\ x_2 &= \ln(\dot{x}_2), \end{aligned}$$

y el procedimiento se convierte, de nuevo, en el del modelo tradicional.

La intención de presentar estos ejemplos es que el lector se de cuenta del gran conjunto de funciones que es posible cubrir usando el modelo tradicional de regresión lineal. Sin embargo, también le permitirán observar cómo se puede complicar la relación de dependencia entre y y las covariables x , de tal manera que el comportamiento de la función de la que realmente se originaron los datos podría no ser capturado por el modelo planteado hasta ahora. Así surge la necesidad de buscar un método que permita aproximar cualquier tipo de relación entre y y x .

Por otro lado, en cuanto el error, la distribución asimétrica de Laplace cumple el cometido de que el cuantil p -ésimo sea igual a 0. Es decir, implícitamente provoca la asimetría necesaria para que el porcentaje de valores por debajo de $f_p(x)$ se aproxime al $p \times 100 \%$, y por arriba, al $(1 - p) \times 100 \%$.

Si bien esta es una característica necesaria, puede no ser suficiente debido a que contempla que la distribución de los errores es la misma para distintos valores de las covariables. Dicha problemática podría ser mitigada mediante el uso de una mezcla de distribuciones, con pesos variables para cada observación. Particularmente es posible usar distribuciones asimétricas de Laplace con diferentes valores para σ y probabilidad asociada a cada una, de acuerdo a su factibilidad.

Entonces surgen algunas preguntas como ¿cuántos valores de σ debería de contener el modelo y cuáles deberían ser esos valores? Normalmente no existe una respuesta definitiva a ambas preguntas y se deja la decisión arbitraria al modelador. Pero, ¿qué pasaría si se planteara un modelo de mezclas infinitas de distribuciones? Así, se podría encontrar la mezcla óptima, ya que cualquier mezcla con número fijo de parámetros sería un caso particular.

En resumen, tanto la estimación de la distribución de f_p , como la de ε_p , podrían mejorarse usando modelos de infinitos parámetros, que generalizan a los modelos con un número de parámetros predefinido. Con los métodos estadísticos tradicionales es imposible hacerlo, pero esto abre la puerta a una visión menos explorada para hacer estadística: los **métodos no paramétricos**.

Como menciona Wasserman (2006): *La idea básica de la inferencia no paramétrica es usar los datos para inferir una medida desconocida, haciendo los menos supuestos posibles. Normalmente esto significa usar modelos estadísticos de dimensión infinita. De hecho, un mejor nombre para la inferencia no paramétrica podría ser inferencia de dimensión infinita.*

Y si bien esto puede sonar irreal, la idea intuitiva que está detrás de este tipo de modelos es que el modelador no debería fijar el número de parámetros antes de analizar la información, sino que los datos deben ser los que indiquen cuántos y cuáles son los parámetros que vale la pena usar.

4.2. Distribución de f_p , mediante procesos Gaussianos ⁹

4.2.1. Introducción a los procesos Gaussianos

Retomando las ideas del capítulo anterior, los modelos de regresión tienen como objetivo describir la distribución de una variable aleatoria y , condicional a los valores de las covariables x , es decir $y|x \sim \mathbb{P}(y|x)$. Dado que es complicado aproximar con exactitud toda la distribución, comúnmente se enfocan en un estadístico particular, que en el caso de la regresión sobre

⁹Las ideas de esta sección son inspiradas por Rasmussen & Williams (2006).

cuantiles se define como $f_p(x) = q_p(y|x)$, y luego suponer una distribución para el error aleatorio alrededor de ese valor.

Con el objetivo de ajustar un modelo, se utiliza el supuesto que

$$y = f_p(x) + \varepsilon_p,$$

tal que $q_p(\varepsilon_p) = 0$.

En el modelo tradicional se utiliza el supuesto de relación lineal $f_p(x) = x^T \beta_p$, mismo que se buscará relajar en esta sección, para obtener un modelo más general.

Es importante recordar que la función f_p es pensada fija, pero desconocida. De nueva cuenta, como reflejo de la incertidumbre respecto al valor real con la que cuenta modelador, es posible asignarle una medida de probabilidad. Pero a diferencia del modelo lineal, ya no existirá el parámetro β_p al cual canalizarle esta incertidumbre, por lo que ahora tendrá que ser sobre toda la función.

Para ello, se puede pensar en una medida de probabilidad para la función f_p . Pero como f_p está definida para múltiples valores de x (un número infinito no numerable, cuando tiene dominio en algún intervalo real), ya no se podrá pensar como una variable aleatoria, sino como un conjunto de variables aleatorias que depende de variables de entrada, es decir, un proceso estocástico. Para el caso particular de esta tesis, se pensará que sigue la ley de probabilidad de un *proceso Gaussiano*, concepto que se introduce a continuación.

Definición 3. *Un **proceso Gaussiano** es un conjunto de variables aleatorias, tal que todo subconjunto finito de ellas tendrá una distribución Gaus-*

siana (Normal) conjunta.

Así, sean $x \in \mathbb{R}$ un vector de covariables, cada $f_p(x)$ tiene una distribución Normal univariada, con media $m(x)$ y varianza $k(x)$, las cuales reflejan el conocimiento previo que se tiene del fenómeno de estudio.

Para continuar con la notación matricial del capítulo anterior, sean $Y \in \mathbb{R}^m$ y $X \in \mathbb{R}^{m \times n}$, y $\mathcal{E}_p \in \mathbb{R}^m$ el vector de errores aleatorios, es posible describir al modelo como

$$Y = f_p(X) + \mathcal{E}_p$$

donde

$$f_p(X) = \begin{bmatrix} f_p(x_1) \\ \dots \\ f_p(x_m) \end{bmatrix},$$

$$x_i \in \mathbb{R}^n, \forall i \in \{1, \dots, m\}.$$

Y debido a la definición de proceso Gaussiano, $f_p(X) \in \mathbb{R}^n$ es un vector aleatorio que se distribuye de forma Normal multivariada, con vector de medias $M_{f_p}(X)$ y matriz de covarianzas $K_{f_p}(X)$.

4.2.2. Definiciones y notación

Para las siguientes definiciones se supondrá que $f_p(x)$ es una variable aleatoria y $f_p(X)$ un vector aleatorio, con medias y covarianzas conocidas y finitas.

Definición. Sean $x, x' \in \mathbb{R}^n$.

La función de medias de f_p (m_{f_p}) se define como

$$m_{f_p} : \mathbb{R}^n \rightarrow \mathbb{R}, \text{ tal que}$$

$$m_{f_p}(x) = \mathbb{E}[f_p(x)].$$

La función de covarianzas de f_p (k_{f_p}) se define como

$$k_{f_p} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}, \text{ tal que}$$

$$k_{f_p}(x, x') = \text{Cov}(f_p(x), f_p(x')).$$

Definición. Sea $X \in \mathbb{R}^m \times \mathbb{R}^n$ y $X' \in \mathbb{R}^r \times \mathbb{R}^n$, es decir,

$$X = \begin{bmatrix} x_1 \\ \dots \\ x_m \end{bmatrix},$$

$$X' = \begin{bmatrix} x'_1 \\ \dots \\ x'_r \end{bmatrix}.$$

La función vector de medias de f_p (M_{f_p}) se define como

$$M_{f_p} : \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}^m, \text{ tal que}$$

$$M_{f_p}(X) = \begin{bmatrix} m_{f_p}(x_1) \\ \dots \\ m_{f_p}(x_m) \end{bmatrix}.$$

La función matriz de covarianzas de f_p (K_{f_p}) se define como

$$K_{f_p} : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^r \rightarrow \mathbb{R}^m \times \mathbb{R}^m, \text{ tal que}$$

$$K_{f_p}(X, X') = \begin{bmatrix} k_{f_p}(x_1, x'_1) & \dots & k_{f_p}(x_1, x'_r) \\ \dots & \dots & \dots \\ k_{f_p}(x_m, x'_1) & \dots & k_{f_p}(x_m, x'_r) \end{bmatrix}.$$

Dadas estas definiciones, se puede observar que el proceso Gaussiano está completamente caracterizado por su función de medias m_{f_p} y su función de covarianzas k_{f_p} . Por lo tanto, la manera en que se definan estas dos funciones representará el conocimiento inicial que se tiene del objeto de estudio. A partir de este punto, y cuando el contexto lo permita, por simplicidad de notación se omitirá el uso del subíndice f_p en las funciones recién definidas. Además, cuando se desee referirse al proceso estocástico f_p que sigue la ley de probabilidad de un proceso Gaussiano, se hará con la notación

$$f_p \sim \mathcal{GP}(m, k).$$

4.2.3. Funciones de covarianza

Recordando que k describe la covarianza entre dos variables aleatorias que pertenecen al mismo proceso estocástico f_p , las propiedades que $k(x, x')$ tiene que cumplir son

$$k(x, x') = k(x', x) \text{ (simetría),}$$

$$k(x, x) = \text{Var}(f_p(x)) > 0.$$

Si bien es cierto que dadas esas restricciones hay una variedad muy grande de funciones con las que se puede describir k , por practicidad, y tomando

en cuenta que es un supuesto sensato para la mayoría de los casos, es común describirla en relación a la distancia entre x y x' , escrita usualmente como $\|x, x'\|$. Es decir, $k(x, x') = k(\|x, x'\|)$. A este tipo de funciones de covarianza se les denomina **estacionarias**.

Esta relación entre covarianza y distancia suele ser inversa, es decir, entre menor sea la distancia, mayor será la covarianza, y viceversa. De esta manera, para valores $x \approx x'$, se obtendrá que $f_p(x) \approx f_p(x')$, por lo que se hace de forma implícita el supuesto de que f_p es una función continua.

Un ejemplo de este tipo de funciones son las **γ -exponencial**, mismas que se definen de la siguiente manera:

$$k(x, x') = k(\|x - x'\|_\gamma) = \lambda \exp(-\tau \|x - x'\|_\gamma),$$

donde λ es un parámetro de escala, τ de rango y γ especifica el tipo de norma euclidiana a usar.

Las de uso más común suelen ser la 1 y 2-*exponencial*. Ambas tienen la ventaja de ser continuas, pero la 2-*exponencial* tiene además la peculiaridad de ser infinitamente diferenciable y, por lo tanto, es suave.

Otra posible función de covarianza es la **racional cudrática**, caracterizada como

$$k(x, x') = k(\|x - x'\|_2) = \lambda \left(1 + \tau \frac{\|x - x'\|_2^2}{2\alpha} \right)^{-\alpha},$$

con $\alpha, \lambda, \tau > 0$.

4.2.4. Predicción

Para esta subsección se supondrá que se cuenta con datos de $f_p(X)$, mismos que en la práctica son imposibles de observar directamente y únicamente se pueden aproximar con el modelo descrito anteriormente. La intención de este supuesto es sentar las bases teóricas para realizar predicción con el modelo central de esta tesis (GPDP), tema que será explorado con más detalle en el siguiente capítulo.

Sea un conjunto de observaciones $\{X, f_p(X) | X \in \mathbb{R}^{m \times n}, f_p(X) \in \mathbb{R}^m\}$. Por otro lado, se tiene un nuevo conjunto de covariables $X_* \in \mathbb{R}^{r \times n}$, y se desea predecir $f_p(X_*) \in \mathbb{R}^r$, subconjunto del proceso Gaussiano f_p .

La distribución inicial conjunta de los datos de entrenamiento $f_p(X)$ y los datos a predecir $f_p(X_*)$ es:

$$\begin{bmatrix} f_p(X) \\ f_p(X_*) \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} M(X) \\ M(X_*) \end{bmatrix}, \begin{bmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix} \right)$$

Bajo el supuesto que ya se conocen los valores de $f_p(X)$, es posible condicionar la distribución conjunta, dadas esas observaciones. Utilizando las propiedades de la distribución Normal condicional¹⁰, se obtiene que:

$$f_p(X_*) | f_p(X) \sim \mathcal{N}(\bar{M}(X, X_*), \bar{K}(X, X_*)),$$

con

$$\begin{aligned} \bar{M}(X, X_*) &= M(X_*) + K(X_*, X)K(X, X)^{-1}(f_p(X) - M(X)), \\ \bar{K}(X, X_*) &= K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*). \end{aligned}$$

¹⁰La especificación de la distribución Normal condicional se puede encontrar en el Apéndice A.

De esta manera quedan sentadas las bases de la distribución no paramétrica de f_p . A continuación se analizarán las de la distribución de ε_p , y en el próximo capítulo se estudiará cómo hacer inferencia conjuntando ambas, mediante el uso del modelo central de esta tesis.

4.3. Distribución de ε_p , mediante procesos de Dirichlet ¹¹

Un proceso de Dirichlet, visto de manera intuitiva, es una medida de probabilidad sobre funciones distribución. Es decir, cada realización de él es en sí misma una función distribución. En el caso particular de este trabajo, cuya misión es encontrar un modelo Bayesiano y no paramétrico para la regresión sobre cuantiles, la ley de probabilidad de los procesos de Dirichlet será utilizada para reflejar la incertidumbre que tiene el modelador acerca de la distribución verdadera del error aleatorio ε_p .

4.3.1. Definición de los procesos de Dirichlet

En términos generales, para que una distribución de probabilidad G siga la ley de probabilidad de un proceso de Dirichlet, toda partición finita de su soporte tiene que seguir una distribución Dirichlet¹². A continuación se enuncia una definición más detallada.

Definición 4. Sean G y H dos distribuciones cuyo soporte es el conjunto Θ y sea $\alpha \in \mathbb{R}_+$. Si se toma una partición finita cualquiera (A_1, \dots, A_r)

¹¹Las ideas de esta sección son retomadas de Teh (2010).

¹²Antes de revisar la definición formal de los Procesos de Dirichlet, es conveniente recordar la definición de la distribución de Dirichlet, misma que se ubica en el Apéndice A.

del conjunto Θ , se entenderá que $G(A_i)$ es la probabilidad de que una realización de G pertenezca al conjunto A_i . A su vez, G será realización de otra distribución de probabilidad, por lo que el vector $(G(A_1), \dots, G(A_r))$ también será aleatorio.

Se dice que G sigue la distribución de un **proceso de Dirichlet** ($G \sim DP(\alpha, H)$), con distribución media H y parámetro de concentración α , si

$$(G(A_1), \dots, G(A_r)) \sim Dir(\alpha H(A_1), \dots, \alpha H(A_r)),$$

para cualquier partición finita (A_1, \dots, A_r) del conjunto Θ .

Es momento de analizar el papel que juegan los parámetros. Sea $A_i \subset \Theta$, uno de los elementos de la partición anterior, y recordando las propiedades de la distribución de Dirichlet, entonces

$$\begin{aligned} E[G(A_i)] &= \frac{\alpha H(A_i)}{\sum_{k=1}^p \alpha H(A_k)} \\ &= H(A_i) \end{aligned}$$

$$\begin{aligned} Var(G(A_i)) &= \frac{\alpha H(A_i) (\sum_{k=1}^p (\alpha H(A_k)) - \alpha H(A_i))}{(\sum_{k=1}^p \alpha H(A_k))^2 (\sum_{k=1}^p (\alpha H(A_k)) + 1)} \\ &= \frac{\alpha^2 [H(A_i)(1 - H(A_i))]}{\alpha^2 (1)^2 (\alpha + 1)} \\ &= \frac{H(A_i)(1 - H(A_i))}{\alpha + 1}. \end{aligned}$$

En este orden de ideas, es posible darse cuenta que la distribución H representa la *distribución media* del proceso de Dirichlet. Por otro lado, el parámetro α tiene una relación inversa con la varianza, es decir, es un pa-

rámetro de precisión. Así, a una mayor α , corresponde una menor varianza del proceso de Dirichlet, y, por lo tanto, una mayor concentración respecto a la distribución media H .

4.3.2. Distribución posterior

Sea (ϕ_1, \dots, ϕ_n) una sucesión de realizaciones independientes provenientes de la función distribución G , cuyo soporte es Θ . Pero G es desconocida, y para reflejar la incertidumbre acerca de su ley de probabilidad real, a su vez se asigna una distribución sobre las posibles distribuciones, particularmente la de un proceso de Dirichlet.

Sea de nuevo (A_1, \dots, A_r) una partición finita cualquiera del conjunto Θ , y $n_k = |\{i : \phi_i \in A_k\}|$ el número de valores ϕ observados dentro del conjunto A_k . Por la propiedad conjugada entre la distribución de Dirichlet y la distribución Multinomial, se obtiene que

$$(G(A_1), \dots, G(A_r)) | \phi_1, \dots, \phi_n \sim \text{Dir}(\alpha H(A_1) + n_1, \dots, \alpha H(A_r) + n_r).$$

Es posible reescribir $n_k = \sum_{i=1}^n \delta_i(A_k)$, donde $\delta_i(A_k) = 1$ si $\phi_i \in A_k$, y 0 en cualquier otro caso. Así,

$$\begin{aligned} \alpha H(A_k) + n_k &= \alpha H(A_k) + \sum_{i=1}^n \delta_i(A_k) \\ &= (\alpha + n) \left[\frac{\alpha \times H(A_k) + n \times \frac{\sum_{i=1}^n \delta_i(A_k)}{n}}{\alpha + n} \right] \\ &= \bar{\alpha} \bar{H}(A_k), \end{aligned}$$

con

$$\bar{\alpha} = \alpha + n$$

$$\bar{H}(A_k) = \left(\frac{\alpha}{\alpha + n} \right) H(A_k) + \left(\frac{n}{\alpha + n} \right) \frac{\sum_{i=1}^n \delta_i(A_k)}{n}.$$

Por lo tanto, $G|\phi_1, \dots, \phi_n \sim DP(\bar{\alpha}, \bar{H})$. Es decir, la distribución posterior de G sigue, de nuevo, la ley de probabilidad de un proceso de Dirichlet, con parámetros actualizados. Asimismo, se puede interpretar a la distribución media posterior \bar{H} como una mezcla entre la distribución media inicial H , con peso proporcional al parámetro de concentración inicial α , y la distribución empírica de los datos, con peso proporcional al número de observaciones n .

4.3.3. Distribución predictiva

Continuando con la idea de la sección anterior de que ya se conoce el valor de ϕ_i, \dots, ϕ_n realizaciones provenientes de la función distribución G , se desea hacer predicción de la observación ϕ_{n+1} , condicionada a los valores observados. Así,

$$\begin{aligned} P(\phi_{n+1} \in A_k | \phi_1, \dots, \phi_n) &= \int P(\phi_{n+1} \in A_k | G) P(G | \phi_1, \dots, \phi_n) dG \\ &= \int G(A_k) P(G | \phi_1, \dots, \phi_n) dG \\ &= \mathbb{E}[G(A_k) | \phi_1, \dots, \phi_n] \\ &= \bar{H}(A_k), \end{aligned}$$

es decir,

$$\phi_{n+1} | \phi_1, \dots, \phi_n \sim \left(\frac{\alpha}{\alpha + n} \right) H(\phi_{n+1}) + \left(\frac{n}{\alpha + n} \right) \frac{\sum_{i=1}^n \delta_i(\phi_{n+1})}{n}.$$

Cabe resaltar que dicha distribución predictiva tiene puntos de masa localizados en ϕ_1, \dots, ϕ_n . Esto significa que la probabilidad de que ϕ_{n+1} tome un valor que ya ha sido observado es mayor a 0, independientemente de la forma de H . Yendo aún más allá, es posible darse cuenta que si se obtienen realizaciones infinitas de G , cualquier valor obtenido será repetido eventualmente, casi seguramente. Por lo tanto, G es una distribución discreta también casi seguramente.

4.3.4. Proceso estocástico de rompimiento de un palo

Dado que $G \sim DP(\alpha, H)$ es una distribución discreta casi seguramente, se puede expresar como una suma de centros de masa de la siguiente manera:

$$G(\phi) = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k^*}(\phi),$$

$$\phi_k^* \sim H,$$

siendo π_k la probabilidad de ocurrencia de ϕ_k^* .

Dicha probabilidad de ocurrencia será generada con la siguiente metáfora.¹³ Se piensa un palo de longitud 1. Se genera un número aleatorio $\beta_1 \sim Beta(1, \alpha)$, mismo que estará en el intervalo $(0, 1)$. Esa será la magnitud del pedazo que será separado del palo de longitud 1, y le será asignado a $\pi_1 = \beta_1$. Así, quedará un palo de magnitud $(1 - \beta_1)$ a repartir. Posteriormente se vuelve a generar un número aleatorio $\beta_2 \sim Beta(1, \alpha)$, que representará la proporción del palo restante que le será asignada a π_2 . Es decir, $\pi_2 =$

¹³Una demostración de la equivalencia puede ser encontrada en Paisley (2010).

$\beta_2(1 - \beta_1)$. En general, para $k \geq 2$,

$$\begin{aligned}\beta_k &\sim \text{Beta}(1, \alpha), \\ \pi_k &= \beta_k \prod_{i=1}^{k-1} (1 - \beta_i).\end{aligned}$$

Dada su construcción, es inmediato darse cuenta que $\sum_{k=1}^{\infty} \pi_k = 1$. En algunas ocasiones se nombra a esta distribución $\pi \sim GEM(\alpha)$, en honor a Griffiths, Engen y McCloskey.

Cabe notar que, bajo esta construcción alternativa de los procesos de Dirichlet, el valor de cada centro de masa y su probabilidad de ocurrencia asociada, son independientes entre sí. En el siguiente capítulo esta característica será explotada por el algoritmo que realiza el ajuste del modelo, para calcular de manera separada ambos conjuntos de parámetros.

4.3.5. Modelo general de mezclas infinitas de Dirichlet

Sean $\{y_1, \dots, y_n\}$ un conjunto de observaciones provenientes de la distribución F , condicionalmente independientes, y que se suponen vienen del *Modelo de mezclas de Dirichlet*:

$$\begin{aligned}y_i | \phi_i &\sim F(y_i | \phi_i), \\ \phi_i | G &\sim G(\phi_i), \\ G | \alpha, H &\sim DP(\alpha, H).\end{aligned}$$

En este modelo de mezclas es posible que existan y 's que comparten un mismo valor para ϕ_i (por la propiedad discreta de G), por lo que pueden ser consideradas pertenecientes a una misma subpoblación.

Es posible reescribir este modelo usando la equivalencia entre los procesos

de Dirichlet y el proceso estocástico de rompimiento de un palo, visto anteriormente. Sea z_i la subpoblación a la que pertenece y_i entre las $\Phi_1^*, \Phi_2^*, \dots$ posibles, se tiene entonces que $P(z_i = \Phi_k^*) = \pi_k$. Y si ϕ_k^* es el valor que comparten los miembros de Φ_k^* , se usará la notación $\phi_{z_i} = \phi_k^*$, cuando $z_i = \Phi_k^*$. Por lo tanto, el modelo se puede ahora escribir como

$$\begin{aligned} y_i | z_i, \phi_k^* &\sim F(y_i | \phi_{z_i}), \\ z_i | \pi &\sim Mult_\infty(\pi)^{14}, \\ \pi | \alpha &\sim GEM(\alpha), \\ \phi_k^* | H &\sim H. \end{aligned}$$

De esta manera, el modelo de mezclas de Dirichlet es un modelo de mezclas infinitas, debido a que tiene un número infinito de posibles subpoblaciones. Intuitivamente, la importancia realmente recae sólo en aquellas subpoblaciones que tienen un peso π posterior mayor a cierto umbral. Sin embargo, dichos pesos son detectados hasta después de observar los datos, a diferencia de los modelos de mezclas finitas, que ya tienen un número de subpoblaciones definidas previamente.

4.3.6. Modelo de mezclas infinitas de Dirichlet para la distribución asimétrica de Laplace

Aterrizando las ideas anteriores al caso particular de los modelos de regresión sobre cuantiles, se busca describir la distribución de ε_p como producto de una mezcla infinita de distribuciones asimétricas de Laplace, de la manera siguiente. Sea $w_p^{AL} | \sigma$ la función de densidad de la distribución asimétrica

¹⁴Se usará la notación $Mult_\infty$ para denotar al límite de la distribución Multinomial, cuando el número de posibles categorías tiende a infinito.

de Laplace, condicional en el valor del parámetro σ . Sea $h_p|G$ la función de densidad de ε_p condicional en una distribución $G(\sigma)$, realización de un proceso de Dirichlet con parámetro de concentración α y distribución media H . Se tiene entonces que

$$h_p(\varepsilon|G) = \int w_p^{AL}(\varepsilon|\sigma) dG(\sigma),$$

$$G \sim DP(\alpha, H).$$

Cabe resaltar que a pesar de la mezcla, se sigue cumpliendo la condición de que $q_p(\varepsilon_p|G) = 0$, para toda G .

Además, por construcción, esta formulación es equivalente al modelo de mezclas infinitas de Dirichlet (visto en la subsección anterior), por lo que se puede reescribir como

$$\begin{aligned} \varepsilon_{p_i}|z_i, \sigma_k^* &\sim AL_p(\varepsilon_{p_i}|\sigma_{z_i}), \\ z_i|\pi &\sim Mult_\infty(\pi), \\ \pi|\alpha &\sim GEM(\alpha), \\ \sigma_k^*|H &\sim H. \end{aligned}$$

En este orden de ideas, la tarea del modelador únicamente consistirá en definir el valor p del cuantil que desea modelar, el del parámetro de concentración α , así como a la distribución de H y sus respectivos hiper-parámetros, con la restricción de que su soporte deberá ser un subconjunto de \mathbb{R}_+ . Por lo tanto, la distribución *Gamma* o la *Gamma-Inversa* se postulan como opciones convenientes.

En el siguiente capítulo se retomará este modelo para especificar el error aleatorio de la regresión sobre cuantiles, y conjuntándolo con los procesos Gaussianos (vistos antes en este capítulo), se obtendrá el modelo GPDP,

centro de esta tesis.

Capítulo 5

Modelo GPDP para regresión sobre cuantiles

5.1. Definición

Después de analizar la introducción de componentes no paramétricos en las distribuciones, tanto de f_p , como de ε_p , a continuación se enunciará el modelo central de esta tesis, al cual se le denominará **Modelo GPDP** (por las siglas en inglés de procesos Gaussianos y procesos de Dirichlet).

Sea el p -ésimo cuantil aquel de interés para el modelador, el cual predefine con anterioridad. Sea $\{(y_i, x_i) | i = 1, \dots, m\}$ el conjunto de observaciones de la variable de respuesta y sus respectivas covariables, cuya relación se supone como

$$y_i = f_p(x_i) + \varepsilon_{p_i},$$

donde $f_p : \mathbb{R}^n \times \mathbb{R}$ es la función cuantil y $\varepsilon_{p_i} \in \mathbb{R}$ es el error aleatorio,

ambos desconocidos.

Para reflejar la incertidumbre del modelador acerca del valor real de f_p , se supone a $f_p \sim \mathcal{GP}(m, k)$, con función de medias m dada por el modelador y función de covarianza k del tipo 2-*exponencial*, con parámetro de rango fijo $\tau = 1$. Es decir,

$$k(x_i, x_j | \lambda) = \lambda \exp\{-\|x_i - x_j\|_2\},$$

con $\lambda \sim GI(c_\lambda, d_\lambda)$, siendo c_λ y d_λ los parámetros de forma y escala, respectivamente, de una *Gamma-Inversa*, mismos que deberán ser elegidos por el modelador.

La razón para fijar $\tau = 1$ es para simplificar el proceso computacional de inferencia que se verá en la siguiente sección, pero bien podría también tener una distribución inicial que refleje la incertidumbre acerca de su valor.

En cuanto a la distribución inicial de ε_p , se supondrá un modelo de mezclas infinitas de Dirichlet, cuya distribución media H del proceso de Dirichlet será una *Gamma-Inversa*, con parámetros de forma c_{DP} y escala d_{DP} , elegidos por el investigador.

En resumen, el Modelo GPDP queda descrito de la siguiente forma:

$$\begin{aligned} y_i | f_p(x_i), z_i, \sigma_k^* &\sim AL_p(\varepsilon_{p_i} = y_i - f_p(x_i) | \sigma_{z_i}), \\ f_p | m, k, \lambda &\sim \mathcal{GP}(m, k(\lambda) | \lambda), \\ \lambda &\sim GI(c_\lambda, d_\lambda), \\ z_i | \pi &\sim Mult_\infty(\pi), \\ \pi | \alpha &\sim GEM(\alpha), \\ \sigma_k^* | c_{DP}, d_{DP} &\sim GI(\sigma_k | c_{DP}, d_{DP}), \\ k(x_i, x_j | \lambda) &= \lambda \exp\{-\|x_i - x_j\|_2\}. \end{aligned}$$

Es posible notar que dicha representación del modelo también podría incorporar a p como un parámetro a estimar, dándole su respectiva distribución inicial. Eso sería particularmente útil para mejorar la estimación de la distribución condicional $y|x$. Sin embargo, a pesar de que el modelo ajusta teóricamente tal distribución, el parámetro de mayor interés para este trabajo es $f_p(x) = q_p(y|x)$, una vez que ya se predefinió que el cuantil p -ésimo es el de interés para el modelador.

5.2. Inferencia con el simulador de Gibbs

Dado que el modelo descrito no es conjugado, las distribuciones posteriores tienen que ser aproximadas mediante métodos computacionales. Para hacer esto, es posible hacer uso de algoritmos MCMC (Markov chain Monte Carlo), y particularmente del simulador de Gibbs.¹⁵

En este orden de ideas, a continuación se detallan las distribuciones condicionales posteriores de los parámetros del modelo, así como la inclusión de algunas variables latentes para permitir el funcionamiento del algoritmo. Es oportuno recordar que dichas distribuciones posteriores resultan de multiplicar la verosimilitud por la probabilidad inicial, como se revisó en el capítulo 2 de este trabajo.

Antes de correr los algoritmos, usualmente resulta conveniente estandarizar los datos. En primer lugar, para que la estructura de covarianza tenga más sentido, ya que la escala de las covariables afectaría la correlación que existe entre los datos, al depender esta de la distancia entre ellas. Además, estandarizar los datos suele mejorar el rendimiento computacional de este

¹⁵En caso de que el lector no esté familiarizado con este tipo de algoritmos, puede consultar una breve descripción de ellos en el Apéndice B.

tipo de algoritmos. Asimismo, vuelve más sencillo definir el valor inicial de los hiper-parámetros, como se detallará más adelante.

5.2.1. Actualización del error

Recordando que los centros de masa y los pesos de un proceso de Dirichlet son independientes, pueden ser actualizados por separado, con el inconveniente de que hay un número infinito de parámetros que actualizar. Para resolverlo, se utilizará el algoritmo de truncamiento del *slice sampling*, propuesto por Kalli *et al.* (2009), y adaptado para el modelo propuesto en esta tesis. A grandes rasgos consiste en truncar las posibles subpoblaciones a un número finito, el cual se actualizará de forma dinámica, de acuerdo a lo que vaya aprendiendo de los datos.

Sea $\xi_1, \xi_2, \xi_3, \dots$ una secuencia positiva, generalmente elegida de forma determinista y decreciente. Sea N una variable aleatoria auxiliar con soporte en los números naturales, la cual representa el número de truncamiento de posibles distintas subpoblaciones, y se actualiza en cada iteración.

Actualización de los centros de masa

Para cada $k \in \{1, 2, \dots, N\}$, se obtiene que

$$\begin{aligned} \sigma_k | \{\varepsilon_{p_i} | z_i = k\}, c, d &\sim GI(\bar{c}_{DP}, \bar{d}_{DP}), \\ \bar{c}_{DP} &= c_{DP} + |\{i | z_i = k\}|, \\ \bar{d}_{DP} &= d_{DP} + p \left[\sum_{\{i | z_i = k, \varepsilon_{p_i} \geq 0\}} \varepsilon_{p_i} \right] + (1 - p) \left[\sum_{\{i | z_i = k, \varepsilon_{p_i} < 0\}} -\varepsilon_{p_i} \right]. \end{aligned}$$

Actualización de los pesos

Sea $\bar{\pi}_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_j)$, de modo que para cada $k \in \{1, 2, \dots, N\}$, la distribución condicional posterior de β_k es

$$\begin{aligned} \beta_k | \{z_i\}, a, b &\sim \text{Beta}(\bar{a}, \bar{b}), \\ \bar{a} &= 1 + |\{i | z_i = k\}|, \\ \bar{b} &= \alpha + |\{i | z_i > k\}|. \end{aligned}$$

Dado que existe un número finito de posibles subpoblaciones, ya no se sigue propiamente la distribución *GEM*, sino un truncamiento de ella hasta la N -ésima subpoblación. Posteriormente se realiza un reescalamiento de las probabilidades para que sumen 1. Es decir, se calcula

$$\pi_k = \frac{\bar{\pi}_k}{\sum_{j=1}^N \bar{\pi}_j}$$

Actualización de las clases y variables de truncamiento

Siguiendo el algoritmo de Kalli *et al.* (2009), para cada observación $i \in \{1, \dots, m\}$, se obtiene

$$u_i \sim U(0, \xi_{z_i}),$$

valor que se utiliza para actualizar la probabilidad de pertenencia a cada clase de la siguiente forma. Para cada $k \in \{1, 2, \dots, N\}$,

$$P(z_i = k | \varepsilon_{p_i}, \pi_k, \sigma_k) \propto \mathbb{1}(u_i < \xi_k) \cdot \frac{\pi_k}{\xi_k} \cdot AL_p(\varepsilon_{p_i} | \sigma_k).$$

Posteriormente se actualiza

$$N = \max\{N_i | N_i = \max\{j | \xi_j > u_i\}, i \in \{1, \dots, m\}\}.$$

5.2.2. Actualización de la tendencia

Se define la variable aleatoria auxiliar b , con la finalidad de anticipar si $\varepsilon_p = y - f_p(x)$ será positiva o negativa, y así simplificar el cálculo de la actualización de f_p .

$$b_i | p, \sigma_i \sim \begin{cases} \frac{p}{\sigma_i} & \text{prob} = P(\varepsilon_{p_i} \geq 0) = 1 - p \\ -\frac{1-p}{\sigma_i} & \text{prob} = P(\varepsilon_{p_i} < 0) = p \end{cases},$$

de forma que $b = [b_1, \dots, b_m]^T$.

Actualización de $f_p(X)$

Es pertinente recordar que la función de densidad de una observación y_i , debido a que sigue la distribución asimétrica de Laplace, se escribe

$$P(y_i | f_p(x_i), \sigma_i) = \frac{p(1-p)}{\sigma_i} \exp \left\{ -\rho_p \left(\frac{y_i - f_p(x_i)}{\sigma_i} \right) \right\} \mathbb{1}_{(-\infty, \infty)}.$$

Una vez calculada la variable auxiliar b que recién se acaba de definir, dicha densidad se puede expresar de forma condicional como

$$P(y_i | f_p(x_i), \sigma_i, b_i) \propto \begin{cases} \exp \left\{ \frac{-p(y_i - f_p(x_i))}{\sigma_i} \right\} \mathbb{1}_{\{y_i - f(x_i) \geq 0\}} & \text{si } b_i > 0 \\ \exp \left\{ \frac{(1-p)(y_i - f_p(x_i))}{\sigma_i} \right\} \mathbb{1}_{\{y_i - f(x_i) < 0\}} & \text{si } b_i < 0 \end{cases}.$$

Por lo tanto, la verosimilitud de las observaciones se puede calcular como

$$\begin{aligned}
 & P(Y|f_p(X), \sigma, b) \\
 & \propto \exp \left\{ - \sum_{\{i|b_i>0\}} \frac{p}{\sigma_i} (y_i - f_p(x_i)) - \sum_{\{i|b_i<0\}} -\frac{1-p}{\sigma_i} (y_i - f_p(x_i)) \right\} \\
 & \quad \prod_{\{i|b_i>0\}} \mathbb{1}_{\{y_i \geq f(x_i)\}} \prod_{\{i|b_i<0\}} \mathbb{1}_{\{y_i < f(x_i)\}} \\
 & = \exp \left\{ - \sum_{\{i|b_i>0\}} b_i (y_i - f_p(x_i)) - \sum_{\{i|b_i<0\}} b_i (y_i - f_p(x_i)) \right\} \\
 & \quad \prod_{\{i|b_i>0\}} \mathbb{1}_{\{y_i \geq f(x_i)\}} \prod_{\{i|b_i<0\}} \mathbb{1}_{\{y_i < f(x_i)\}} \\
 & = \exp \left\{ -b^T (y - f_p(X)) \right\} \prod_{\{i|b_i>0\}} \mathbb{1}_{\{y_i \geq f(x_i)\}} \prod_{\{i|b_i<0\}} \mathbb{1}_{\{y_i < f(x_i)\}}.
 \end{aligned}$$

En esta peculiar verosimilitud, cada $f_p(x_i)$ estará condicionada de manera excluyente a estar por arriba o por abajo de y_i . Por ese motivo, al multiplicar la verosimilitud por la distribución inicial Gaussiana de $f_p(X)$, se obtendrá como distribución posterior una Normal Truncada, misma que se detalla a continuación.

$$f_p(X)|Y, X, M, b, \lambda \sim TruncNormal(\bar{M}(X, b), K(X, X|\lambda), \gamma, \eta),$$

$$\bar{M}(X, b) = M(X) + K(X, X|\lambda)b,$$

$$\gamma_i = \begin{cases} -\infty & \text{si } b_i > 0 \\ y_i & \text{si } b_i < 0 \end{cases},$$

$$\eta_i = \begin{cases} y_i & \text{si } b_i > 0 \\ \infty & \text{si } b_i < 0 \end{cases},$$

donde γ es el vector de límites inferiores y η es el vector de límites superiores

de la distribución Normal truncada.

Debido a que $f_p(X)$ se encuentra en la verosimilitud únicamente como un elemento de primer orden, al hacer la multiplicación con la distribución inicial Normal, la varianza queda exactamente igual. La actualización se da únicamente en la media, como una perturbación de la media inicial, dada por las covarianzas que tiene cada observación respecto a las demás, así como el signo de b_i para cada una.

Actualización del parámetro de escala

Condicional a los demás valores obtenidos, se obtiene la distribución posterior de λ como

$$P(\lambda|X, M(X), f_p(X), b, c_\lambda, d_\lambda) \propto \lambda^{-\bar{c}_\lambda - 1} \cdot \exp\left\{-\frac{\bar{d}_\lambda}{\lambda}\right\} \cdot \exp\{-\bar{B}\lambda\},$$

$$\bar{c}_\lambda = c_\lambda + \frac{p}{2},$$

$$\bar{d}_\lambda = d_\lambda + \bar{F},$$

$$\bar{F} = \frac{1}{2}(f_p(X) - M(X))^T[K(X, X|\lambda = 1)^{-1}](f_p(X) - M(X)),$$

$$\bar{B} = \frac{1}{2}b^T[K(X, X|\lambda = 1)]b.$$

5.3. Predicción

Una de las desventajas de los modelos no paramétricos es que, a diferencia de los modelos paramétricos, es complicado interpretar los parámetros de ajuste del modelo. Por ello, resulta particularmente importante la faceta de la predicción, que es donde más se puede explotar la flexibilidad de los

modelos no paramétricos, debido a que su objetivo es tener precisión en la estimación. Específicamente, esta sección se enfocará en la predicción de f_p , que es el parámetro de mayor interés del modelo, para efectos de este trabajo.

Debido al uso del simulador de Gibbs, después de realizar el ajuste, se cuenta con un conjunto grande de realizaciones aproximadas de $f_p(X)$, provenientes de las cadenas de Markov.

Recordando lo visto en la sección 4.2.4, cuando se tienen valores de $f_p(X)$, es posible usar la propiedad de la *Normal condicional* para realizar predicción. Sea $X \in \mathbb{R}^m \times \mathbb{R}^n$ la matriz de datos originales, $X_* \in \mathbb{R}^r \times \mathbb{R}^n$ la matriz de datos a predecir, $f_p(X)$ una realización de la distribución posterior correspondiente a X , y $f_p(X_*)$ el vector aleatorio de los datos a predecir. Se tiene entonces que

$$f_p(X_*)|f_p(X) \sim \mathcal{N}(\bar{M}(X, X_*), \bar{K}(X, X_*|\lambda)),$$

con

$$\begin{aligned}\bar{M}(X, X_*) &= M(X_*) + K(X_*, X)K(X, X)^{-1}(f_p(X) - M(X)), \\ \bar{K}(X, X_*|\lambda) &= \lambda \times \left[K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*) \right].\end{aligned}$$

donde $K(X_1, X_2) = K(X_1, X_2|\lambda = 1)$, y X_1 y X_2 pueden ser X o X_* .

Por lo antes descrito, es posible obtener una realización de $f_p(X_*)$ simulando de dicha distribución Normal. De esta manera, por cada valor de $f_p(X)$ y λ en la cadena de Markov, se simula una realización de $f_p(X_*)$, y entonces es posible aproximar la distribución posterior de $q_p(y|x) = f_p(x)$, para los datos X_* .

5.4. Hiper-parámetros iniciales del modelo

Una complicación que puede tener un modelo jerárquico, como el GPDP, es que no es sencillo darle una interpretación intuitiva a los hiper-parámetros de las diversas capas, de forma que el conocimiento previo del modelador pueda reflejarse en valores asignados a ellos.

Para mitigar este problema, a continuación se proponen una serie de heurísticas para definirlos, mismas que se derivan de algunas ideas que me parecen sensatas, pero no se originan de ningún cuerpo axiomático y bien podrían ser mejoradas. También es importante aclarar que por lo comentado al inicio de la sección 5.2, para todas ellas se pensará que los datos están estandarizados.

5.4.1. Función de medias m

Este es el hiper-parámetro al que se le puede dar una mayor interpretación, debido a que representa el nivel donde el modelador estima que se encontrará el cuantil p -ésimo de y , para cada valor de las covariables x .

Para simplificar el proceso de definición de la función de medias m , se puede partir de la hipótesis que es constante, y, por lo tanto, las variaciones son únicamente producto de la varianza de ε_p . Dada la estructura de probabilidad posterior, la media de f_p podrá actualizarse si los datos brindan información suficiente para suponer lo contrario.

Una vez aceptada esta estructura, resta asignar el valor constante que tomará. Si el modelador tiene una idea del nivel donde espera los datos, puede asignar la constante c . En caso de no tenerla, un valor que se suele usar en

diversos contextos es el de $c = 0$, de forma que

$$m : \mathbb{R}^n \rightarrow \mathbb{R}, \text{ tal que}$$

$$m(x) = c.$$

5.4.2. *Gamma-Inversas* de λ y el Proceso de Dirichlet

Tanto c_λ y d_λ , como c_{DP} y d_{DP} son parámetros de distribuciones *Gamma-Inversa*. Es oportuno recordar que si $U \sim \mathcal{GI}(c, d)$, entonces

$$\mathbb{E}[U] = \frac{d}{c-1}, \quad c > 1,$$

$$Var(U) = \frac{d^2}{(c-1)^2(c-2)}, \quad c > 2.$$

En este orden de ideas, si se elige $c = 2$, $Var(U)$ será infinita y $\mathbb{E}[U] = d$. Asignar a c_λ y c_{DP} de esta manera permitirá darle a d_λ y d_{DP} el valor que se piense como el mejor estimador puntual *a priori* de λ y σ , pero con una varianza holgada, que permitirá a los datos tener el peso principal en la actualización del modelo.

Debido a la estandarización de los datos, la varianza muestral de y es igual a 1. Es posible pensarla como el resultado de sumar la varianza de $f_p(x)$ y la de ε_p , que además se suponen independientes. Entonces, se puede definir una heurística tal que $Var(f_p(x)) = \frac{1}{2}$ y $Var(\varepsilon_p) = \frac{1}{2}$, a falta de mayor información.

La varianza de $f_p(x)$ es igual a λ , por lo que lo coherente con lo dicho en los párrafos anteriores será asignar $d_\lambda = \frac{1}{2}$. Por el otro lado, si únicamente para este ejercicio, y con el afán de volver analítico el cálculo, se piensa a

$\varepsilon_p \sim AL_p(\sigma = d_{DP})$. Entonces, su varianza estaría dada por

$$Var(\varepsilon_p) = \left[\frac{d_{DP}}{p(1-p)} \right]^2 (1 - 2p(1-p)).$$

Dado que se fijará $Var(\varepsilon_p) = \frac{1}{2}$, por la heurística antes mencionada, despejando es posible obtener que

$$d_{DP} = \frac{p(1-p)}{\sqrt{2(1-2p(1-p))}}.$$

5.4.3. Parámetro de concentración α

Este es el parámetro más difícil de definir, por su complejidad de interpretación. Pero cabe recordar que el valor de α tiene una relación positiva con el número de subpoblaciones.

De hecho, sea \bar{m} el número de subpoblaciones y m el número de datos de entrenamiento, Teh (2010) expone que

$$\mathbb{E}[\bar{m}|\alpha, m] \simeq \alpha \log \left(1 + \frac{m}{\alpha} \right), \text{ para } m, \alpha \gg 0.$$

Si se define $\alpha = \frac{\sqrt{m}}{2}$, se tiene que

$$\begin{aligned} \mathbb{E}[\bar{m}|m] &\simeq \frac{\sqrt{m}}{2} \times \log(1 + 2\sqrt{m}) \\ &\simeq \frac{m}{7}, \text{ para } m \approx 100. \end{aligned}$$

Es decir, si se tienen alrededor de 100 observaciones, el número esperado de subpoblaciones será alrededor de la séptima parte de las observaciones.

Valor que a falta de mayor exploración en este tema, no suena descabellado. De nuevo, vale la pena tener presente que a mayor cantidad de datos, se tendrá una menor dependencia de esta arbitraria decisión inicial.

5.5. Consideraciones sobre la bondad de ajuste

El modelo GPDP descrito en este trabajo pretende ser una opción más de modelo de regresión, particularmente útil cuando es interés del modelador aproximar algún cuantil en específico de la variable de respuesta, dados los valores de las variables explicativas.


Sin embargo, no está de más recordar que existen muchos otros modelos de regresión, a la media y sobre cuantiles, lineales y no lineales, paramétricos y no paramétricos. Todos ellos cumplen el mismo cometido de estimar la distribución condicional de $y|x$.

En la siguiente sección se hará un análisis de los resultados obtenidos por el modelo GPDP, que hará uso de herramientas básicas, como la exploración visual, la correlación o el error cuadrático medio. En este sentido, hay una evidente área de mejora, ya que se podría encontrar alguna medida robusta, estadísticamente hablando, de bondad de ajuste. Ésta permitiría saber qué tan bueno resulta el modelo para describir un conjunto dado de datos, y permitiría compararlo con otros, para seleccionar cuál es el que presenta mejores resultados.

Desafortunadamente, realizar esto con un modelo con la estructura del GPDP resulta complejo y la literatura para medir la bondad de ajuste de modelos de regresión bayesianos, sobre cuantiles y no paramétricos, está apenas en sus pininos. Por lo tanto, una estimación robusta de la bondad

de ajuste del modelo GPDP y la selección del mejor modelo quedan fuera del alcance de este trabajo.

5.6. Paquete *GPDPQuantReg* en R

Todas las ideas expuestas en este capítulo han sido implementadas en el paquete *GPDPQuantReg* del lenguaje de programación R, mismo que puede ser encontrado en el repositorio de Github  titulado: **opardo/GPDP-QuantReg**.

Al momento de escribir este trabajo, cuenta con tres funciones públicas: *GPDPQuantReg*, para ajustar el modelo con el simulador de Gibbs; *predict*, para realizar predicción en un nuevo conjunto de datos del modelo ajustado; y *diagnose*, para realizar el diagnóstico de la ergodicidad, la autocorrelación, la correlación cruzada y la traza de las cadenas de Markov, para los distintos parámetros.

A continuación se expone un ejemplo de uso, el cual es similar a lo que se realizó para obtener los resultados del capítulo siguiente.

```

1 # Instalación del paquete
2 install.packages("devtools")
3 library(devtools)
4 install_github("opardo/GPDPQuantReg")
5 library(GPDPQuantReg)
6
7 # Simulación de datos
8 set.seed(201707)
9 f_x <- function(x) return(0.5 * x * cos(x) - exp(0.1 * x))
10 error <- function(m) rgamma(m, 2, 1)
11 m <- 20
12 x <- sort(sample(seq(-15, 15, 0.005), m))
13 sample_data <- data.frame(x = x, y = f_x(x) + error(m))

```

```
14
15 # Ajuste del modelo
16 GPDP_MCMC <- GPDPQuantReg(y ~ x, sample_data, p = 0.250)
17
18 # Predicción, usando el modelo ajustado
19 pred_data <- data.frame(x = seq(-15, 15, 0.25))
20 credibility <- 0.90
21 prediction <- predict(GPDP_MCMC, pred_data, credibility)
22
23 # Diagnóstico de las cadenas de markov
24 diagnose(GPDP_MCMC)
```

Capítulo 6

Aplicaciones

6.1. Metodología de simulación de datos y ajuste de los modelos

Para la elaboración de este capítulo se ajustaron modelos de diversos cuantiles, a distintos conjuntos de datos. Los datos se obtuvieron de la siguiente manera. Sea $y \in \mathbb{R}$ el valor de la variable de respuesta, $x \in \mathbb{R}$ su respectiva covariable, $g : \mathbb{R} \rightarrow \mathbb{R}$ una función denominada *tendencia* y $\omega(x) \in \mathbb{R}$ un error aleatorio, se simuló

$$y = g(x) + \omega(x).$$

Cabe resaltar que se utiliza la notación g , y ω , para diferenciarlas de f_p y ε_p , que son las usadas en el planteamiento de los modelos teóricos, vistos en los capítulos anteriores. La razón es que no necesariamente se cumple que $q_p(\omega|x) = 0$. De hecho, como se verá en las siguientes líneas, la función que se querrá estimar no será g , sino las correspondientes f_p para diferentes

cuantiles de interés.

La función real del cuantil p -ésimo de $y|x$ se puede obtener como

$$q_p(y|x) = g(x) + q_p(\omega|x),$$

función que en la construcción teórica se denominó como f_p , y la cual se desea estimar para diversos valores de p .

En todos los casos presentados a continuación se ajustó el modelo para el cuantil 0.5-ésimo, por ser la mediana y una medida de tendencia central alternativa a la media. También se modeló el cuantil 0.95-ésimo, ya que es un valor extremo. Finalmente, se busco describir al cuantil 0.25-ésimo, debido a que el primer cuartil es de interés dentro de algunos contextos, y no es medida de tendencia central, pero tampoco un valor extremo.

Para todos los casos se simularon 60 datos sin reemplazo dentro del intervalo $(-15, 15)$, con un refinamiento de 2 decimales. Las razones de elegir esa cantidad de datos, que podría parecer reducida, fueron varias. Primero, para validar el poder predictivo del algoritmo y su capacidad de ajuste a los valores originales, aún cuando se tienen pocos datos. Segundo, para obtener viabilidad en tiempos de ejecución, ya que el algoritmo simula en cada iteración, entre otras variables aleatorias, normales multivariadas y truncadas, con dimensión igual al número de datos. Dicha simulación utiliza el método de aceptación y rechazo, por lo que entre mayor sea la dimensión, usualmente requiere de mayor esfuerzo computacional, y, por ende, el tiempo total del ajuste del modelo aumenta de manera significativa. Finalmente, la tercer razón es que esa cantidad de datos permitirá observar qué influencia puede tener un dato particular en el ajuste general del modelo.

Una vez simulados los datos, se ajustaron el modelo tradicional de regresión

a la media (descrito en la sección 3.2.2 de este trabajo) y el GPDP. Es importante recordar que el primero, además de ser sobre la media, es lineal en los parámetros de las covariables y con error normal. Por otro lado, el modelo GPDP propone ideas alternativas, como centrar su estimación en el cuantil de interés, y utilizar componentes no paramétricos, tanto para modelar la tendencia, como el error. En medio queda una amplia gama de modelos, como el modelo tradicional de regresión sobre cuantiles, o algún modelo de regresión a la media que contemple errores no paramétricos, por ejemplo.

Sin embargo, decidí realizar esa particular comparación para descubrir qué tan bueno podría ser un modelo que contempla ideas alternativas en todas las facetas de la regresión (estadístico a modelar, tendencia y error aleatorio), cuando se le comparara con el que utiliza únicamente supuestos tradicionales. Eso no significa que si se mezclan algunas ideas alternativas, con algunas tradicionales, no se pueda obtener un tercer modelo, que resulte aún mejor. Pero el alcance de esta tesis se limita a comparar únicamente esos dos modelos.

Para el ajuste del modelo tradicional se utilizó un polinomio de orden 5. Es decir, se utilizaron como covariables los valores de x, x^2, \dots, x^5 , además del término independiente. Aún así, dicho ajuste fue simple, computacionalmente hablando, ya que requirió únicamente de multiplicaciones matriciales para actualizar sus parámetros posteriores. Como parámetros iniciales de la distribución Normal de β , se utilizó el vector de medias con valores iguales a 0, y como matriz de varianzas y covarianzas, la identidad. En el caso de la *Gamma-Inversa* de σ^2 , el valor de forma usado fue 2, y el de escala, 1.

Por otra parte, el ajuste del GPDP involucró el cálculo de una cadena de Markov con ciertas particularidades. En todos los casos se desearon las

primeras 5,000 simulaciones, para así evitar dependencia de los valores iniciales. Posteriormente, se simularon 15,000 valores para cada parámetro, pero de ellos únicamente se guardó uno de cada cinco, con la intención de buscar replicar el comportamiento independiente, que se supone entre los valores. Por lo tanto, se utilizaron 3,000 valores simulados para aproximar la distribución posterior de cada parámetro. Dicho ajuste se realizó utilizando el paquete *GPDPQuantReg* en R, mismo que, como se detalló en el capítulo anterior, implementa el modelo GPDP para la regresión sobre cuantiles.

Posteriormente, se realizó la predicción para cada $x \in [-15, 15]$, tal que x tuviera valor cero para todos los decimales a partir de las centésimas. Es decir, $-15.0, -14.9, \dots, 14.9, 15.0$. Se utilizaron ambos modelos para simular 3,000 predicciones de cada $f_p(x)$ ¹⁶. Posteriormente, se calcularon los cuantiles muestrales 0.025 y 0.975-ésimos, para crear el intervalo de probabilidad al 95 % de la estimación posterior de la predicción. Además, se calculó el cuantil 0.5-ésimo (la mediana), para tener una medida de tendencia central de dicha estimación.

Finalmente, se obtuvieron métricas que permitieran comparar a grandes rasgos la calidad de estimación de ambos modelos. Se contrastó la estimación mediana de la predicción con el valor real del cuantil, y se obtuvo el error cuadrático medio, así como la correlación al cuadrado. Además se calculó qué porcentaje de los valores reales cayeron dentro del intervalo de probabilidad al 95 %, estimado para la predicción.

¹⁶El algoritmo utilizado para simular cuantiles del modelo tradicional de regresión a la media se detalla en el Apéndice C.

6.2. Conjuntos de datos simulados

A continuación se presentan los diversos conjuntos de datos simulados. Primero, se expone uno que cumple con los supuestos de la regresión tradicional a la media. Posteriormente, se presentan otros que desafían alguno de sus supuestos, con el fin de comparar cómo lidian dicho modelo, y el modelo GPDP, con tal inconveniente. Finalmente, se comparan los resultados.

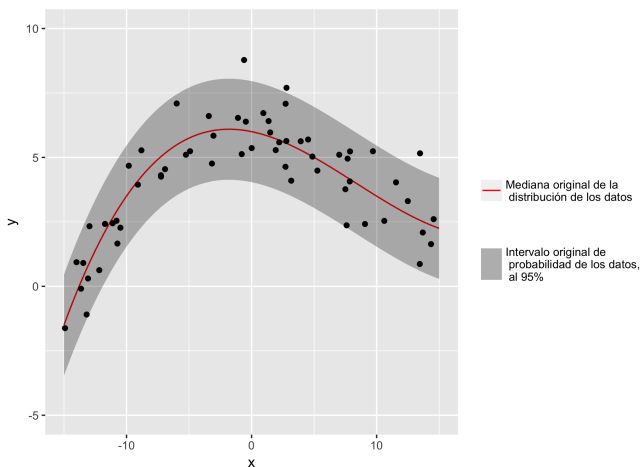
6.2.1. Supuestos tradicionales de regresión a la media

El primer caso que se presenta es el de un conjunto de datos que se simuló con las condiciones comúnmente supuestas dentro de los modelos de regresión a la media: función de tendencia polinomial y error normal alrededor de ella, independiente del valor de x . Es decir, los errores presentaron homocedasticidad. Expresado en términos matemáticos,

$$g(x) = \frac{1}{1000}x^3 - \frac{1}{40}x^2 - \frac{1}{10}x + 6,$$

mientras el error se distribuyó $\omega(x) \sim \mathcal{N}(0, 1)$. Los datos simulados se pueden observar en la figura 6.1.

Figura 6.1: Datos simulados con los supuestos tradicionales de regresión.



Posteriormente se ajustaron los modelos y se realizó predicción de la forma en la que se explicó anteriormente, obteniendo los siguientes resultados.

Cuadro 6.1: Error cuadrático medio de datos que cumplen supuestos tradicionales de regresión.

Cuantil	Modelo Tradicional	Modelo GPDP
0.95	0.84	0.83
0.50	0.02	0.19
0.25	0.23	0.16

Cuadro 6.2: Correlación al cuadrado de datos que cumplen supuestos tradicionales de regresión.

Cuantil	Modelo Tradicional	Modelo GPDP
0.95	0.99	0.91
0.50	0.99	0.94
0.25	0.99	0.96

Cuadro 6.3: Porcentaje de valores reales dentro del intervalo de confianza de datos que cumplen supuestos tradicionales de regresión.

Cuantil	Modelo Tradicional	Modelo GPDP
0.95	68 %	96 %
0.50	100 %	100 %
0.25	100 %	100 %

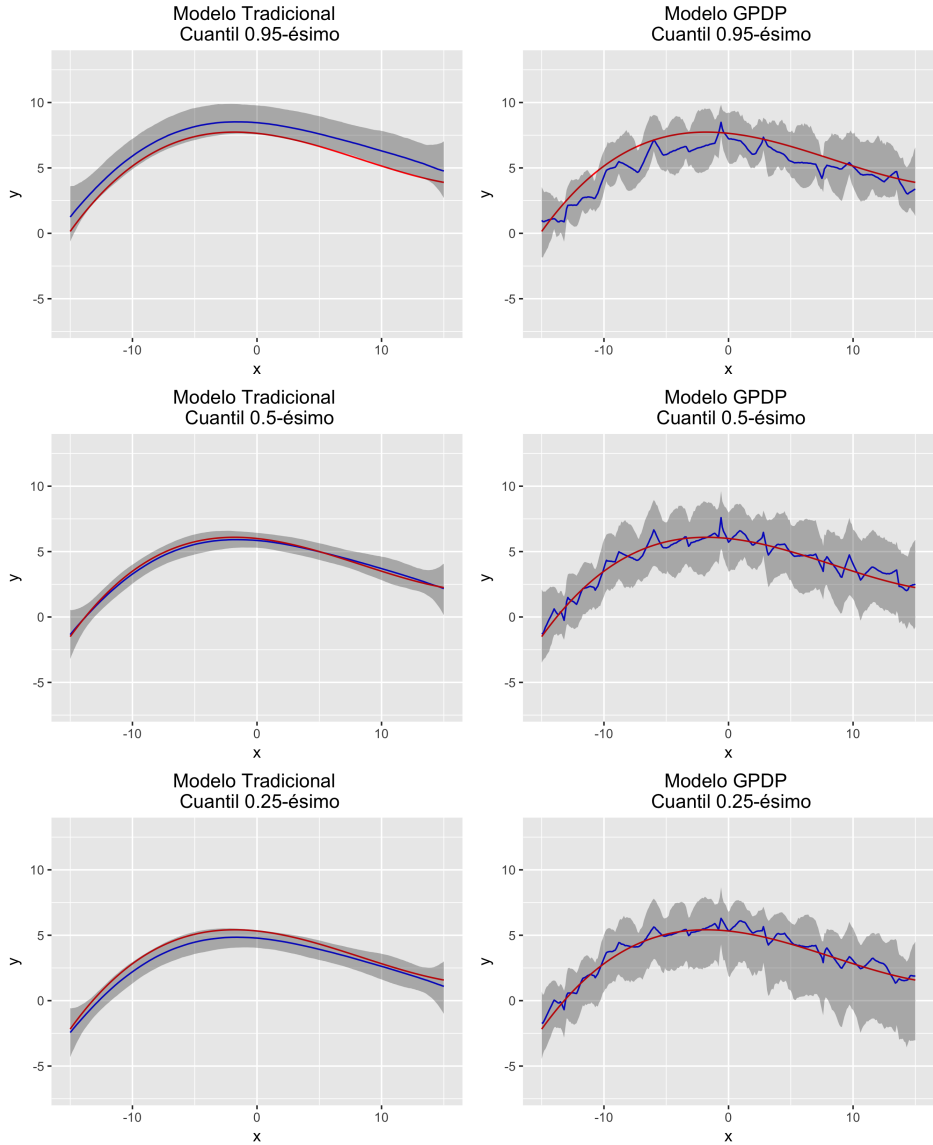
Como era de esperarse, el modelo tradicional replicó de forma muy atinada el comportamiento polinomial de los datos. La métrica que mejor refleja esta situación es la correlación al cuadrado, en donde dicho modelo obtuvo prácticamente un valor perfecto para los tres cuantiles ajustados. Sin embargo, si bien no alcanzó los mismos niveles, el modelo GPDP también obtuvo valores altos.

De hecho, si observa el cuadro 6.1, el lector podrá notar que para el cuantil 0.95-ésimo, el error cuadrático medio es prácticamente el mismo, y para el cuantil 0.25-ésimo es menor. Es decir, mientras que el modelo tradicional comprendió de mejor manera las subidas y bajadas puntuales de la función, el modelo GPDP fue más certero para estimar el nivel, conforme los valores se alejaron de la mediana.

Lo anterior es confirmado por la tabla 6.3, ya que para el caso del ajuste tradicional del cuantil 0.95-ésimo únicamente el 68 % de los valores reales del cuantil quedaron dentro del angosto intervalo de probabilidad. Mi hipótesis es que dicho error de estimación del nivel proviene de un error minúsculo de ajuste de la distribución de σ^2 , el cual queda totalmente desapercibido en la estimación casi perfecta de la mediana, pero en un valor extremo, como es el caso del cuantil 0.95-ésimo, sí se ve reflejado.

Hay comentarios del ajuste de este conjunto de datos que coincidirán con otros, por lo que en la sección 6.3 se harán apuntes generales acerca de

Figura 6.2: Ajuste de los modelos Tradicional y *GPDP*, para un conjunto de datos que cumplen los supuestos tradicionales de regresión.

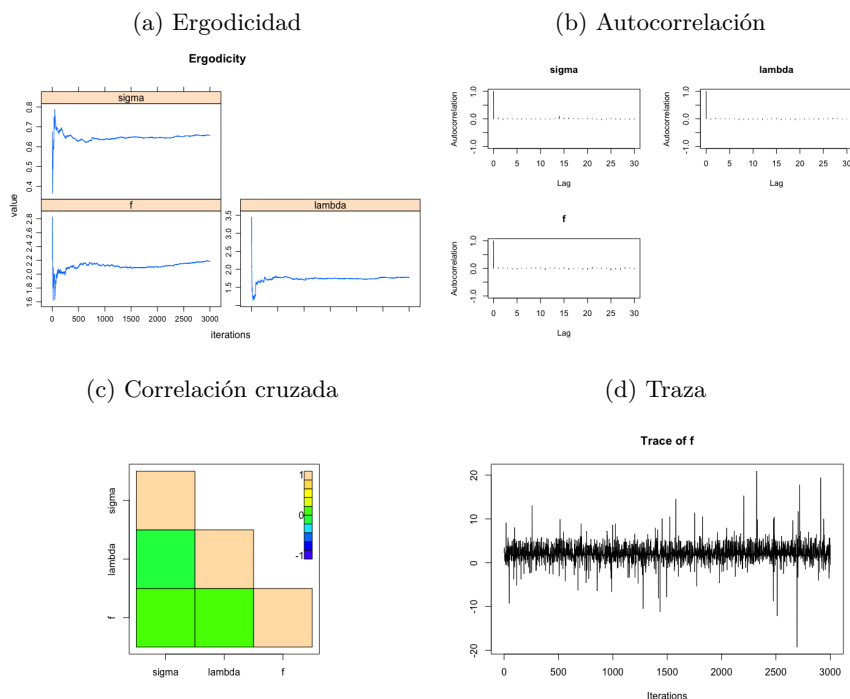


Nota: La línea roja representa el valor real de cada cuantil, la línea azul representa la mediana de la distribución posterior predictiva y el área gris su intervalo de probabilidad al 95 %.

ellos.

Por otro lado, como se detalló en el capítulo anterior, con el uso del paquete *GPDQuantReg* también es posible obtener diagnósticos de las cadenas de Markov del modelo GPD, los cuales se explican de manera más detallada en el Apéndice B. Como ejemplo, se presentan los del ajuste del cuántil 0.5-ésimo para los datos de esta subsección, en la figura 6.3. Los parámetros que se analizan son la λ , parámetro de escala de la función covarianza del proceso Gaussiano; y, además, se eligió una observación de forma aleatoria, para la que se dio seguimiento al valor de σ y $f_p(x)$.

Figura 6.3: Diagnósticos de las cadenas de Markov del cuántil 0.5-ésimo, de datos que cumplen los supuestos tradicionales de regresión.



La ergodicidad parece adecuada, debido a que después de las primeras iteraciones, el promedio del valor de los parámetros se estabiliza en un nivel relativamente fijo para todas las iteraciones siguientes. La autocorrelación es excelente, ya que para los tres parámetros validados se muestra el valor de 1 en la correlación consigo mismo (por construcción, siempre es así), pero a partir del siguiente valor en la cadena de Markov, la correlación es prácticamente 0. Es decir, los valores simulados sí parecen haber sido simulados de la distribución posterior, de forma independiente entre sí.

La correlación cruzada entre parámetros también es cercana a 0, por lo que la simulación de un parámetro de la cadena no muestra una gran dependencia del valor de otro parámetro, en la misma iteración. Por ello, hay motivos para pensar que la cadena de Markov no cayó en ciclos distintos a su distribución estacionaria. De hecho, esto se puede confirmar al observar la traza de f , ya que no muestra comportamientos cíclicos y aunque se centra en cierto intervalo de valores, también hay simulaciones que salen de él de forma esporádica.

6.2.2. Error de colas pesadas

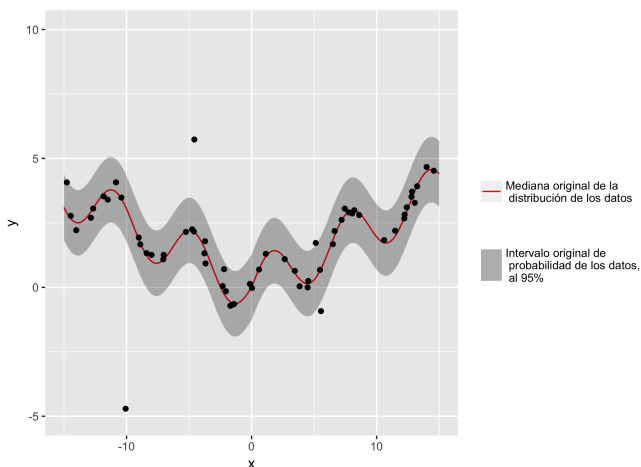
En el siguiente conjunto de datos se utilizó la distribución Cauchy para simular el error. A pesar de ser simétrica y centrada en 0, dicha distribución es peculiar debido a que su esperanza y varianza no están definidas. El motivo de ello son sus colas pesadas, mismas que originan con cierta frecuencia valores muy alejados del centro. Además, se introdujo un componente sinoidal en la función tendencia, mismo que puede ser aproximado con un polinomio tanto como se quiera, utilizando términos de su serie de Taylor. Sin embargo, también tiene un valor absoluto, que requiere un número grande de términos para poder ser aproximado certeramente por la familia de los polinomios.

La expresión de la función tendencia utilizada es

$$g(x) = \frac{1}{4}|x| + \text{sen}(x).$$

Por otro lado, el error se distribuyó $\omega(x) \sim \text{Cauchy}(0, 0.1)$. Los datos obtenidos se pueden observar en la figura 6.4.

Figura 6.4: Datos simulados con error de colas pesadas.



Como se puede observar, la mayoría de los datos están concentrados alrededor de la función tendencia. Sin embargo, ocasionalmente los valores se disparan y se alejan ampliamente de dicha concentración. Claros ejemplos son el dato que está muy por debajo de los demás, cerca de $x = -10$, y uno que está muy por arriba, alrededor de $x = -5$. Lo interesante de este ejemplo fue ver cómo se comportaban ambos modelos ante dichos datos atípicos. Los resultados se muestran a continuación.

Cuadro 6.4: Error cuadrático medio de datos con error de colas pesadas.

Cuantil	Modelo Tradicional	Modelo GPDP
0.95	5.42	0.18
0.50	0.61	0.75
0.25	1.89	1.12

Cuadro 6.5: Correlación al cuadrado de datos con error de colas pesadas.

Cuantil	Modelo Tradicional	Modelo GPDP
0.95	0.64	0.91
0.50	0.64	0.64
0.25	0.64	0.57

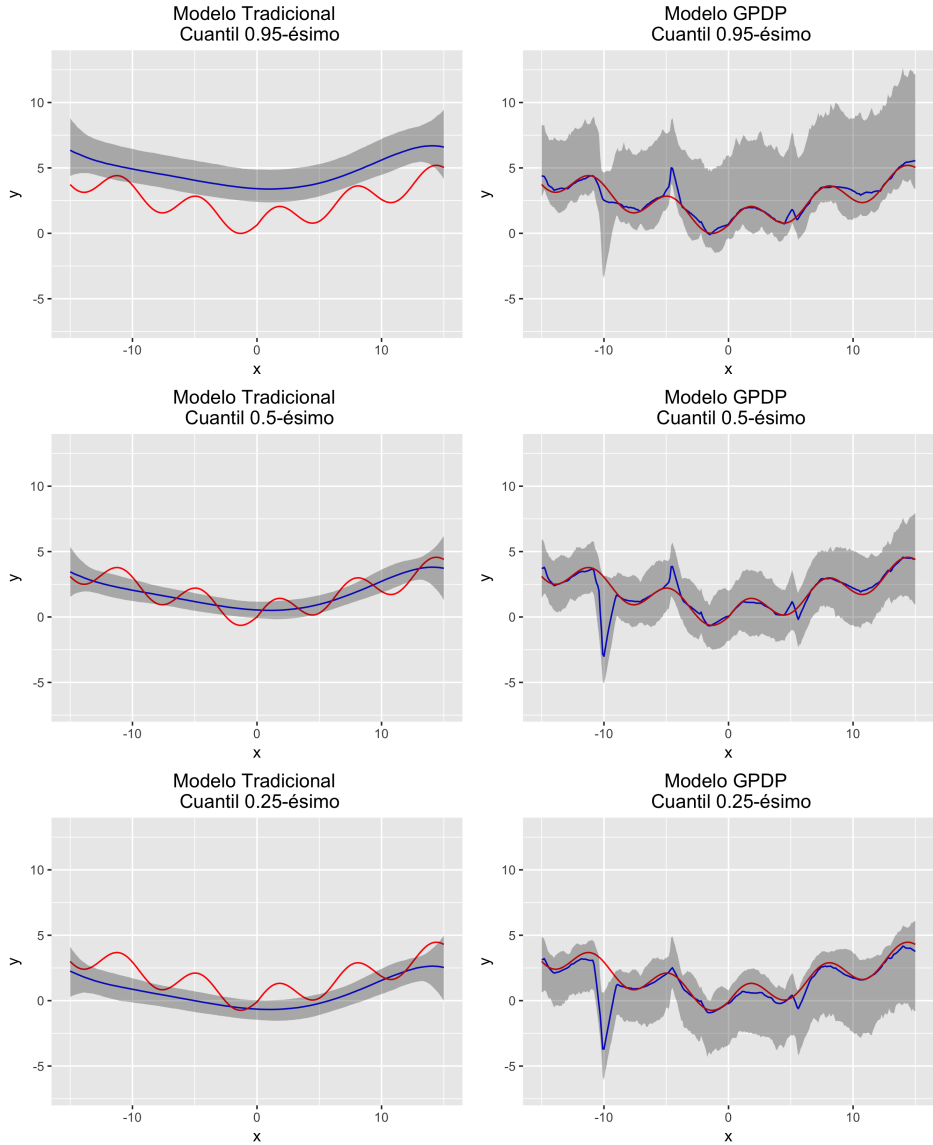
Cuadro 6.6: Porcentaje de valores reales dentro del intervalo de confianza de datos con error de colas pesadas.

Cuantil	Modelo Tradicional	Modelo GPDP
0.95	9 %	98 %
0.50	57 %	100 %
0.25	41 %	96 %

Una vez obtenidos los resultados de ambos modelos, resulta interesante observar varios aspectos de la figura 6.5. Tocando el tema de la influencia de valores atípicos, parece claro que el modelo tradicional los manejó estimando valores muy altos para σ^2 , del error que supone normal. Esto ocasionó que para los cuantiles distintos a la mediana, estimara que estaban más lejos del centro, de lo que realmente estaban. Dicha situación se puede confirmar utilizando el cuadro 6.6, ya que sólo el 41 % de los datos reales cayó dentro del intervalo de probabilidad estimado para el cuantil 0.25-ésimo, y peor aún para el 0.95-ésimo, donde sólo el 9 % lo hizo.

En cuanto a la estimación de la tendencia, el modelo tradicional interpre-

Figura 6.5: Ajuste de los modelos Tradicional y *GPDP*, para un conjunto de datos con error de colas pesadas.



Nota: La línea roja representa el valor real de cada cuantil, la línea azul representa la mediana de la distribución posterior predictiva y el área gris su intervalo de probabilidad al 95 %.

tó que los datos venían de lo que parece ser una parábola, con mínimo y centro en 0, acercándose relativamente al comportamiento del valor absoluto. Desafortunadamente no logró atrapar el comportamiento sinoidal y atribuyó las fluctuaciones al error aleatorio. Eso se puede ver claramente en su estimación del cuantil 0.5-ésimo, donde capturó el nivel de manera adecuada, pero como muestra la figura 6.5 y el cuadro 6.5, muchos valores salieron tanto por arriba, como por abajo, del intervalo.

Por otro lado, el modelo GPDP hizo una estimación muy certera en la mayoría de los puntos, como se refleja en los cuadros 6.4 y 6.6. La vecindad alrededor de $x = 10$, que coincide con el valor atípico más dramático, es donde el ajuste pudo haber sido mucho mejor. Sin embargo, creo que es importante recalcar que el problema se limitó a un error de estimación local, que no afectó a los puntos más alla de cierta pequeña vecindad. De hecho, el siguiente valor atípico en magnitud, alrededor de $x = 5$, fue manejado aceptablemente por el modelo para los tres cuantiles.

6.2.3. Heterocedasticidad

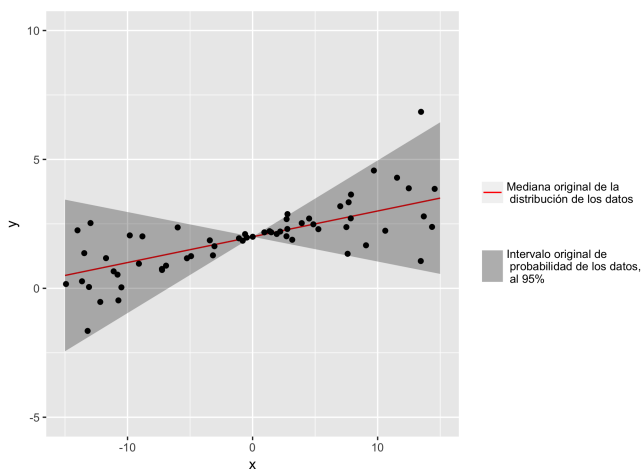
Otro supuesto común en muchos modelos de regresión es la homocedasticidad, es decir, que los errores no dependen de x . Para este ejemplo particular decidí observar cómo afectaría a la estimación el hecho de que la distribución de los errores aleatorios dependiera del valor de la variable explicativa, es decir, datos que presentaran heterocedasticidad.

Para ello se usó como función de tendencia al polinomio de primer orden

$$g(x) = \frac{1}{10}x + 2,$$

y la complejidad de estimación recayó en el error $\omega(x) \sim \mathcal{N}\left(0, \frac{|x|}{10}\right)$.

Figura 6.6: Datos simulados con heterocedasticidad.



Como se puede observar en la figura 6.6, únicamente la mediana de los datos es una línea totalmente recta. Los demás cuantiles se expresan como dos rectas con distinta pendiente, que se unen en el punto $x = 0$. Este comportamiento, similar al valor absoluto, tampoco se puede expresar fielmente con polinomios, sino que sólo se puede aproximar. A continuación se presentan los resultados del ajuste de ambos modelos.

Cuadro 6.7: Error cuadrático medio de datos que presentan heterocedasticidad.

Cuantil	Modelo Tradicional	Modelo GPDP
0.95	0.51	0.58
0.50	0.02	0.17
0.25	0.15	0.14

Cuadro 6.8: Correlación al cuadrado de datos que presentan heterocedasticidad.

Cuantil	Modelo Tradicional	Modelo GPDP
0.95	0.63	0.78
0.50	0.98	0.86
0.25	0.86	0.84

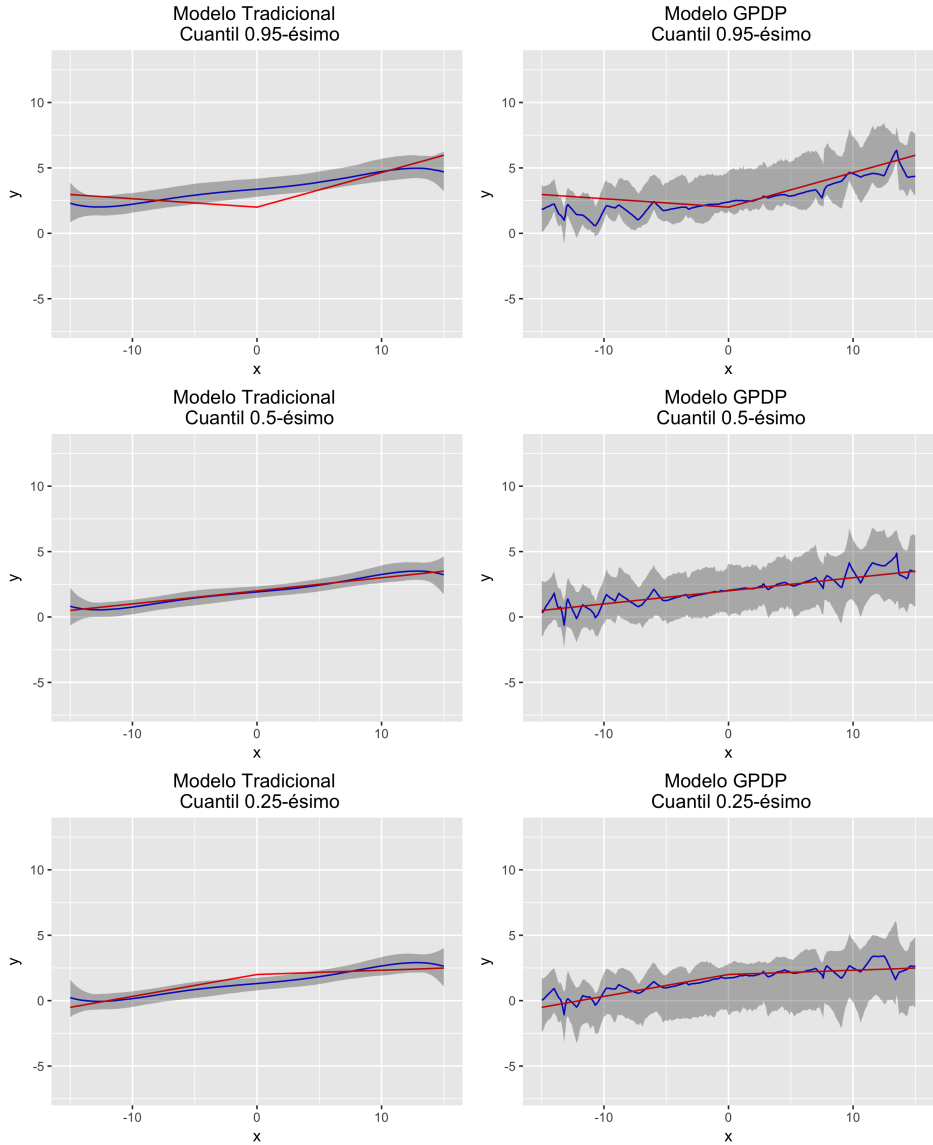
Cuadro 6.9: Porcentaje de valores reales dentro del intervalo de confianza de datos que presentan heterocedasticidad.

Cuantil	Modelo Tradicional	Modelo GPDP
0.95	68 %	95 %
0.50	100 %	100 %
0.25	81 %	100 %

Como se puede observar en la figura 6.7, el modelo tradicional interpretó el valor de todos los cuantiles como una simple línea recta, únicamente variando la ordenada al origen. De hecho, el modelo de la mediana logró una estimación excelente, como lo reafirman los cuadros 6.7, 6.8 y 6.9. Sin embargo, conforme la estimación se fue alejando de la mediana y los valores reales de los cuantiles fueron distinguiéndose de la línea recta, la calidad de la estimación predictiva se redujo drásticamente.

En cuanto al modelo GPDP, los cuadros 6.7, 6.8 y 6.9 también muestran que conforme se tomó un cuantil más lejano a la mediana, la calidad de la estimación empeoró. Sin embargo, la caída fue menos fuerte, particularmente para las medidas de correlación y porcentaje de valores dentro del intervalo. Esto se debe principalmente a que el modelo tradicional modela la media, y a partir de ahí cada cuantil sólo representa un cambio de nivel, pero manteniendo la forma de la función. En cambio, el modelo GPDP contempla que cada cuantil puede tener una forma propia, distinta de los

Figura 6.7: Ajuste de los modelos Tradicional y *GPDP*, para un conjunto de datos con heterocedasticidad.



Nota: La línea roja representa el valor real de cada cuantil, la línea azul representa la mediana de la distribución posterior predictiva y el área gris su intervalo de probabilidad al 95 %.

otros cuantiles.

6.2.4. Error asimétrico

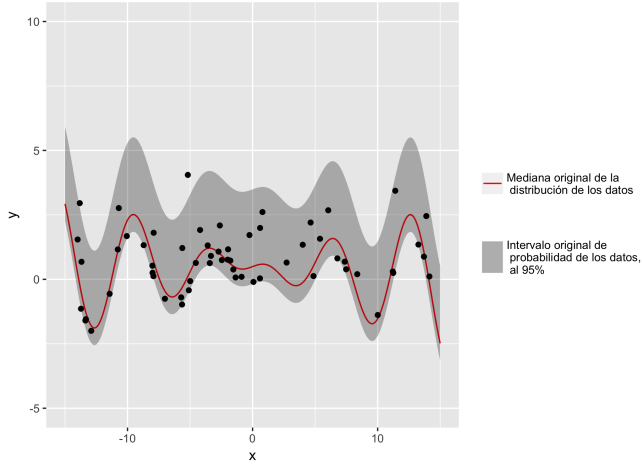
Hasta este punto se han realizado perturbaciones al modelo tradicional de regresión en la forma de la distribución de los datos (colas pesadas y heterocedasticidad). Sin embargo, dichas perturbaciones mantuvieron una característica del modelo tradicional: la simetría del error. En el siguiente conjunto de datos se desafió dicha condición, usando un error aleatorio asimétrico. Particularmente se utilizó $\omega(x) \sim \text{Gamma}(1, 1)$.

La función tendencia involucró a las funciones coseno y exponencial, ambas pudiéndose aproximar de buena forma con polinomios, utilizando sus series de Taylor. Sin embargo, cabe recordar que únicamente se están utilizando polinomios de grado 5 en el modelo tradicional, por lo que podrían resultar insuficientes. Su expresión matemática fue

$$g(x) = \frac{1}{5}x\cos(x) - \frac{1}{5}\exp\left(\frac{x}{10}\right).$$

Los datos obtenidos después de simular se pueden observar en la figura 6.8.

Figura 6.8: Datos simulados con error asimétrico.



Como se podrá dar cuenta el lector, la asimetría del error provoca cambios en la distancia de los cuantiles respecto a la mediana. Por dar un ejemplo, el cuantil 0.025-ésimo (el límite inferior del intervalo) está mucho más cerca de la mediana, que el cuantil 0.975-ésimo (el límite superior del intervalo). Esta situación es atípica, ya que estamos acostumbrados a que ambos estén a la misma distancia, propiciado por la simetría. Por ello, es pronosticable que el modelo tradicional sufrirá para reflejar dicho comportamiento, debido a que supone error normal, el cual presenta la simetría a la que estamos acostumbrados. A continuación se presentan los resultados obtenidos del ajuste de ambos modelos.

Cuadro 6.10: Error cuadrático medio de datos con error asimétrico.

Cuantil	Modelo Tradicional	Modelo GPDP
0.95	1.69	1.23
0.50	1.65	0.64
0.25	1.53	0.44

Cuadro 6.11: Correlación al cuadrado de datos con error asimétrico.

Cuantil	Modelo Tradicional	Modelo GPDP
0.95	0.00	0.50
0.50	0.00	0.61
0.25	0.00	0.69

Cuadro 6.12: Porcentaje de valores reales dentro del intervalo de confianza de datos con error asimétrico.

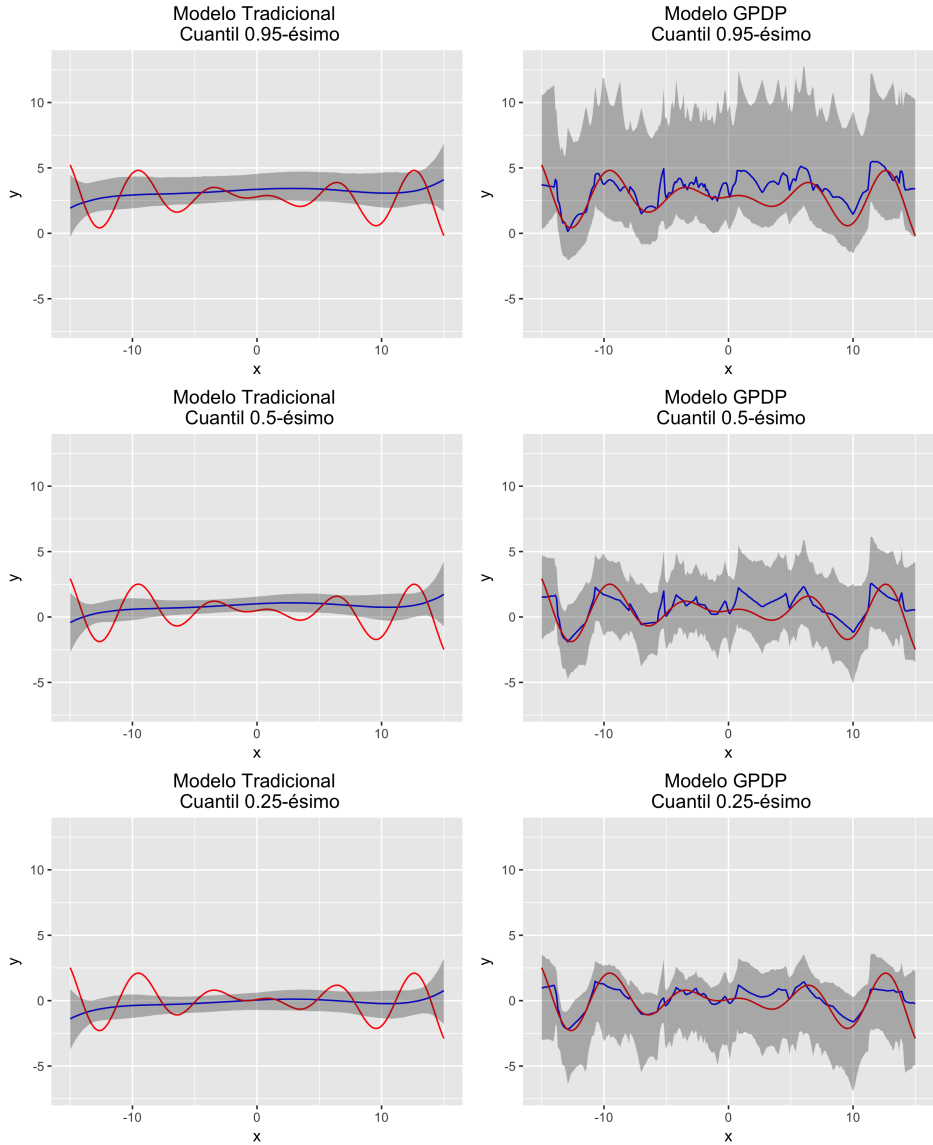
Cuantil	Modelo Tradicional	Modelo GPDP
0.95	58 %	100 %
0.50	46 %	100 %
0.25	52 %	100 %

La figura 6.9 permite observar que el modelo tradicional infiere una mediana, en su mayoría, mayor que la real. Esto se debe a que, por el supuesto de simetría, implícitamente modela a la mediana como si fuera la media, siendo la segunda más grande en su valor real, que la primera.

La misma figura sugiere que el modelo tradicional interpretó a la función de los diversos cuantiles prácticamente como una constante, y atribuyó las variaciones únicamente al error aleatorio. La sugerencia de la constante también se sustenta en el cuadro 6.11, donde es posible ver que la correlación es prácticamente 0 (la primer cifra significativa es el tercer decimal), respecto a los valores originales. Además, la hipótesis de que las variaciones se le atribuyeron al error aleatorio se apoya en los bajos porcentajes de valores reales dentro de los intervalos de predicción, presentados en el cuadro 6.12.

Por otro lado, se puede verificar en las tablas 6.10, 6.11 y 6.12 que el modelo GPDP logró un mejor desempeño en todas las métricas, para todos los

Figura 6.9: Ajuste de los modelos Tradicional y *GPDP*, para un conjunto de datos con error asimétrico.



Nota: La línea roja representa el valor real de cada cuantil, la línea azul representa la mediana de la distribución posterior predictiva y el área gris su intervalo de probabilidad al 95 %.

cuantiles. Sin embargo, no hay que perder de vista que de todos los casos vistos hasta este punto, es el que presentó intervalos de probabilidad más amplios. Es decir, es el caso en el que hay mayor incertidumbre respecto a la predicción. Además, presentó niveles de correlación bajos, comparativamente hablando. No necesariamente la causa fue el error asimétrico, sino pudo provenir de que la forma de la función tendencia es atípica y compleja de inferir para muchos modelos.

6.2.5. Discontinuidades

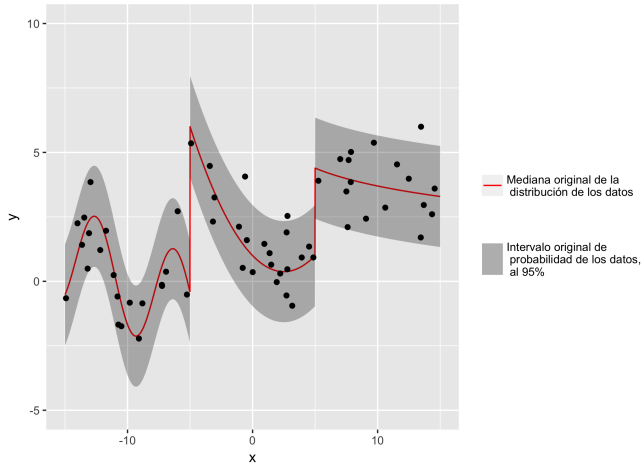
Dejando atrás las modificaciones en el error $\omega(x)$, ahora se presenta una situación atípica en la función tendencia. Tanto el modelo tradicional, como el modelo GPDP, suponen continuidad en las funciones de los cuantiles. Por ello, me pareció interesante averiguar cómo se comportaría su ajuste, en caso de modelar datos que provinieran de una función tendencia discontinua. Particularmente se simuló utilizando la función

$$g(x) = \begin{cases} -\frac{1}{5}x + 2\cos(x) - 2 & \text{si } x \in (-\infty, -5) \\ \frac{1}{10}x^2 - \frac{1}{2}x + 1 & \text{si } x \in [-5, 5) \\ 6 - \ln(x) & \text{si } x \in [-5, \infty) \end{cases},$$

con el error típico de los modelos de regresión, $\omega(x) \sim \mathcal{N}(0, 1)$.

El comportamiento discontinuo se puede entender de mejor manera gráficamente, como lo muestra la figura 6.10. Además, en dicha figura también se pueden ver los puntos obtenidos de la simulación.

Figura 6.10: Datos simulados con discontinuidades.



Como se puede observar en la figura 6.10, y en la expresión matemática antes mencionada, la función tendencia tiene tres secciones: la primera con un comportamiento oscilatorio, la segunda, un trozo de parábola, y la última, un logaritmo, que por el nivel de la covariable, asemeja a una recta. Todas, por separado, son continuas. Por lo tanto, son aproximables con polinomios en un intervalo cerrado. Sin embargo, entre cada sección hay un salto de discontinuidad, por lo que la función global no puede ser expresada fielmente usando dicha familia. Ante dicha dificultad se ajustaron los modelos tradicional y GPDP, obteniendo los resultados que se presentan a continuación.

Cuadro 6.13: Error cuadrático medio de datos simulados con discontinuidades.

Cuantil	Modelo Tradicional	Modelo GPDP
0.95	7.08	1.10
0.50	2.23	0.35
0.25	3.19	0.47

Cuadro 6.14: Correlación al cuadrado de datos simulados con discontinuidades.

Cuantil	Modelo Tradicional	Modelo GPDP
0.95	0.38	0.83
0.50	0.38	0.90
0.25	0.38	0.88

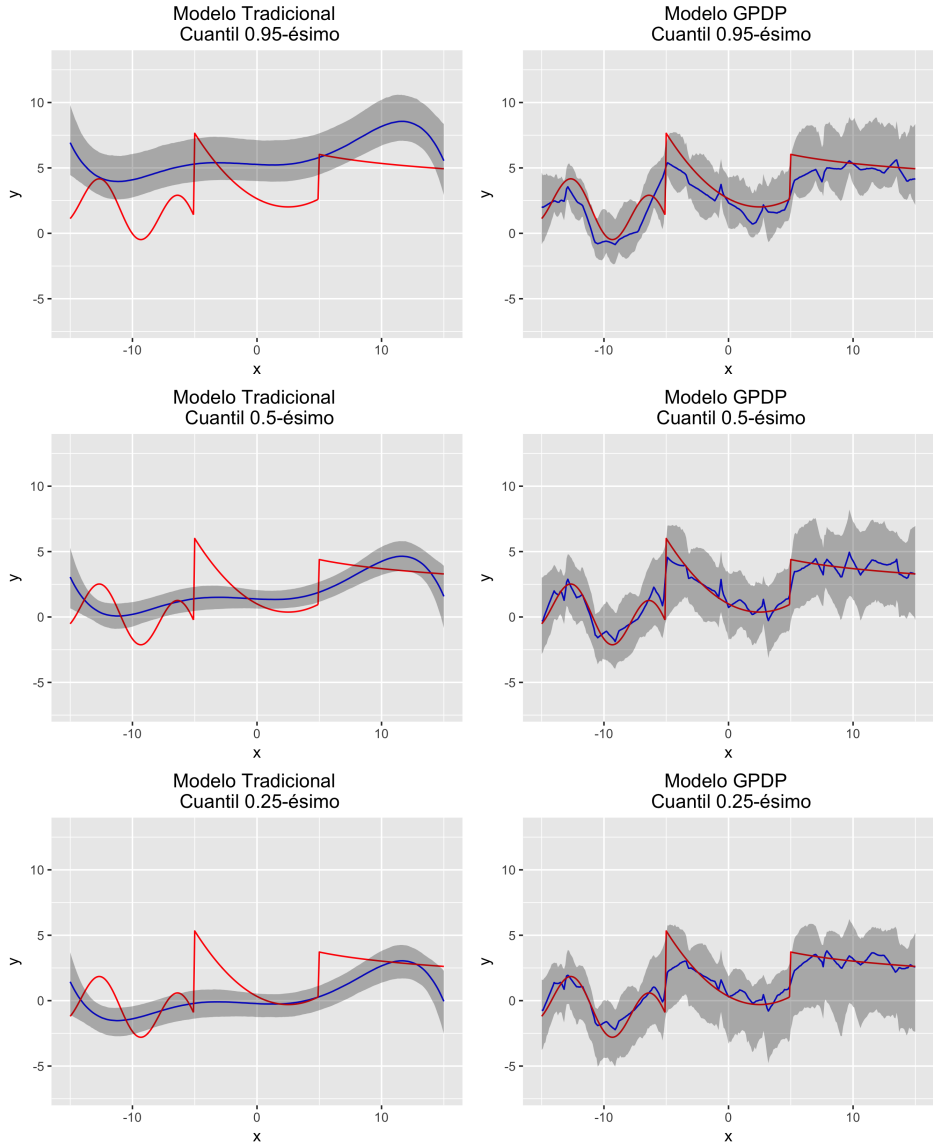
Cuadro 6.15: Porcentaje de valores reales dentro del intervalo de confianza de datos simulados con discontinuidades.

Cuantil	Modelo Tradicional	Modelo GPDP
0.95	28 %	98 %
0.50	46 %	100 %
0.25	53 %	99 %

Como se puede observar en la figura 6.11, el modelo tradicional estuvo muy lejos de inferir correctamente la forma real de la función de los distintos cuantiles. Esto también lo confirma el cuadro 6.15, ya que en la mejor de las estimaciones del modelo tradicional, únicamente el 53 % de los datos reales se pronosticó dentro del intervalo de probabilidad. Al parecer, atribuyó el primer salto únicamente al error aleatorio, el cual estimó de gran magnitud, razón por lo que su estimación de los cuantiles 0.25 y 0.95-ésimos fue más alejada de la mediana de lo que era realmente. El segundo salto parece que sí lo detectó, pero la elevación la realizó paulatinamente, llegando al punto máximo después de 5 unidades de donde realmente ocurrió.

Por otro lado, como lo confirma la exploración visual de la figura 6.11, así como los cuadros 6.13, 6.14 y 6.15, la estimación del modelo GPDP fue muy superior en calidad, debido a que tuvo errores cuadráticos medios consistentemente bajos, correlación cuadrática de al menos 83 % y al menos 98 % de los valores reales adentro del intervalo de probabilidad de la predicción.

Figura 6.11: Ajuste de los modelos Tradicional y *GPDP*, para un conjunto de datos simulados con discontinuidades.



Nota: La línea roja representa el valor real de cada cuantil, la línea azul representa la mediana de la distribución posterior predictiva y el área gris su intervalo de probabilidad al 95 %.

Debido a la construcción del modelo, tuvo una estimación continua de la función de los cuantiles. Sin embargo, lo compensó elevando rápidamente su nivel (dentro de vecindades más pequeñas que la unidad) cuando hubo saltos de discontinuidad en la función real. Por ello, los intervalos donde la estimación fue mala fueron en realidad muy angostos.

6.3. Comparación general entre modelos

El análisis de los datos anteriores brinda información para pensar que el ajuste del modelo GPDP fue, la mayoría de las veces, igual o mejor que el del modelo tradicional de regresión a la media, cuando el objetivo fue estimar el valor de los cuantiles. Esto se dio principalmente cuando se modelaron valores extremos, o los datos provinieron de situaciones que violan los supuestos de la regresión tradicional a la media.

Pasando a un análisis más minucioso, observé detalles que considero vale la pena resaltar. Para empezar, la magnitud de los intervalos de probabilidad de predicción, ya que los que presentó el modelo GPDP fueron comúnmente más anchos que los del tradicional. Esto no significa que su estimación haya sido peor. De hecho, quizás fue lo contrario. Mientras que los intervalos de probabilidad al 95 % del modelo GPDP siempre contuvieron, al menos, al 95 % de los valores reales de los cuantiles (como era esperado), hubo intervalos del modelo tradicional al 95 % de probabilidad que contuvieron únicamente al 9 % de los valores reales.

Hablando de los intervalos de probabilidad, es interesante observar que aquellos provenientes del GPDP se hicieron más angostos, conforme hubo más datos observados cercanos a ese valor, y más anchos, ante la ausencia de datos en la vecindad próxima. Este comportamiento provino de la

estructura de correlación dada por el proceso Gaussiano, recordando que depende de la distancia entre los datos. Por su parte, en el modelo tradicional de regresión a la media, el intervalo de predicción tuvo menor anchura en la parte media de los datos de x , y creció muy ligeramente conforme se acercó a los valores mínimo y máximo observados. En lo personal, prefiero el comportamiento del modelo GPDP, ya que considero que entre mayor sea la cantidad de datos cercanos, se debería tener menor incertidumbre, y viceversa.

También es interesante observar que la mediana de la predicción en el modelo tradicional se encontró usualmente en medio del intervalo. Por otro lado, debido a que supone que el error es una mezcla infinita de procesos de Dirichlet, el modelo GPDP presentó, en algunos casos, asimetría en la distribución posterior predictiva.

Las personas que hemos trabajado con el modelo tradicional de regresión a la media estamos acostumbrados a que cualquier estimación puntual que tomemos es suave, ya que se puede expresar como polinomio, y, por lo tanto, es infinitamente diferenciable. En cambio, debido a los supuestos que introduce el modelo GPDP para el proceso Gaussiano, sus estimaciones puntuales mantuvieron un comportamiento mucho menos suave, pero que fue útil para capturar cambios bruscos en la función. Sin embargo, es importante recordar que la dificultad de interpretación que tienen los modelos no paramétricos representa una de sus mayores desventajas.

Finalmente, es interesante comparar la diferencia en tiempos de ajuste y predicción para ambos modelos, los cuales se muestran en las figuras 6.16 y 6.17.¹⁷ Cabe señalar que los tiempos presentados son el acumulado de ajustar o predecir los modelos de los tres cuantiles.

¹⁷Estos tiempos se obtuvieron corriendo los procesos en una MacBook, con procesador 1.1 GHz Intel Core M, y memoria RAM DDR3, de 8 GB y 1600 MHz. Los procesos de predicción se corrieron en paralelo, usando 2 de los 4 núcleos.

Cuadro 6.16: Tiempo de ajuste por conjunto de datos, para cada modelo.

Datos	Tradicional (seg)	GPDP (seg)
Supuestos tradicionales	menos de 1	2,498
Colas pesadas	menos de 1	4,006
Heterocedasticidad	menos de 1	3,502
Error asimétrico	menos de 1	6,707
Discontinuidades	menos de 1	3,062

Cuadro 6.17: Tiempo de predicción por conjunto de datos, para cada modelo.

Datos	Tradicional (seg)	GPDP (seg)
Supuestos tradicionales	6.4	563.8
Colas pesadas	4.7	529.0
Heterocedasticidad	5.3	534.5
Error asimétrico	5.7	537.4
Discontinuidades	5.0	533.2

Analizando los tiempos de ajuste, y recordando que se usaron 60 datos en todos los casos, es clara la bondad dada por la propiedad conjugada del modelo tradicional. Ésta permite obtener una expresión analítica cerrada para la distribución posterior de los parámetros, y por ende, el ajuste consiste únicamente en multiplicaciones matriciales que toman menos de un segundo. Además, por las características de este modelo, únicamente se requirió ajustar una vez por conjunto de datos, independientemente de que se usara para varios cuantiles.

Por otro lado, el modelo GPDP requiere un ajuste distinto por cuantil, y el tiempo que tarda en correr está siempre en función de la cantidad de iteraciones que se eligen realizar. En el caso particular de este trabajo, se corrieron 20,000 iteraciones para todos los casos, lo cual representa un número poco sobresaliente, comparado con las que normalmente se utilizan

cuando se busca tener una precisión alta. El modelo que menos tardó requirió más de 40 minutos para realizar el ajuste, por lo que la diferencia es abismal entre ambos modelos.

De hecho, es interesante ver cómo hay grandes diferencias dentro de los propios ajustes del modelo GPDP. La razón principal es la simulación de una distribución normal truncada en cada iteración del algoritmo, usando un método de aceptación y rechazo. Cuando los datos cumplieron los supuestos tradicionales, fue cuando le costó menos trabajo realizar las simulaciones. En cambio, cuando se usó una función con muchas subidas y bajadas, y error asimétrico, el algoritmo requirió más del doble de tiempo para poder ajustar el modelo.

Para el caso de las predicciones, el número de puntos aumentó a 300 para todos los casos. Sin embargo, debido a que estas simulaciones ya no pertenecen a una cadena de Markov, se pueden correr en paralelo para los dos modelos. Ambos requieren simular una distribución Normal, pero el tradicional simula una marginal por cada dato, mientras que el GPDP requiere simular una conjunta, de dimensión igual al número de datos. Esa es la razón por la que tarda 100 veces más.

Así, mientras que el modelo GPDP mostró, en general, mejores resultados en la precisión, el modelo tradicional de regresión a la media fue mucho más efectivo en tiempo. Por ello, el modelador debería considerar cuál de ambas prefiere en su estudio, para poder tomar una mejor decisión de qué modelo elegir.

Capítulo 7

Conclusiones y trabajo futuro

Si bien los modelos de regresión a la media han sido de mucha utilidad en las últimas décadas, principalmente cuando el poder computacional era menor, es importante darse cuenta que actualmente existen contextos en los que resultan insuficientes. Ya sea porque sea deseada la mejor aproximación posible de algún estadístico distinto a la media, o debido a que hay ocasiones en las que no se cumplen algunos de sus supuestos.

De manera similar, la relación lineal en los parámetros y la distribución Normal del error han sido fundamentales para que los modelos de regresión hayan proliferado en una gran cantidad de industrias, tanto por su interpretabilidad, como por su bajo costo de estimación. Sin embargo, es imposible ignorar que únicamente representan un subconjunto del universo de funciones y errores aleatorios posibles. Crear modelos que permitan una mayor flexibilidad, como aquellos que utilizan métodos no paramétricos,

logrará una representación más certera de la realidad de la que provienen los datos.

Al momento de hacer aproximaciones estadísticas, se dice que utilizar el paradigma Bayesiano muchas veces presenta la ventaja de poder introducir información de las personas expertas en el fenómeno a estudiar. Desafortunadamente, en un modelo jerárquico, como lo es el expuesto en esta tesis, es complicado transmitir dicho conocimiento, debido a que los hiperparámetros se encuentran varias capas abajo de aquellos parámetros que tienen una interpretación natural. A pesar de ello, vale la pena resaltar que dicha construcción jerárquica, con la flexibilidad que brinda, es posible y es congruente gracias al paradigma Bayesiano, debido a que todas las expresiones son completamente probabilísticas y fundadas en un cuerpo axiomático.

Un reto importante que presentó este trabajo fue el desarrollo del paquete en R para implementar el modelo GPDP. Primero, porque se requirió plantear teóricamente la distribuciones condicionales necesarias para que corriera el simulador de Gibbs. Y segundo, porque tuve que buscar programar de forma general y eficiente, para que el paquete funcionara siempre que recibiera los parámetros predefinidos, y corriera lo más rápido posible, ante la desventaja que representa sólo poder calcular una iteración de la cadena de Markov, a la vez. De hecho, dejé el número de iteraciones como un parámetro a elección del modelador, para que pueda decidir si prefiere precisión o velocidad.

Si bien estos avances son significativos, aún existe mucho que explorar respecto a lo expuesto en esta tesis. Por ejemplo, el modelo planteado en este trabajo no es capaz de darle un peso distinto a cada variable explicativa, sino que las toma a todas por igual al momento de calcular la distancia entre observaciones. Para mejorar esta situación se podría plantear una

descomposición de la función del cuantil en la suma de varios procesos Gaussianos, uno por covariable, lo que brindaría un mayor peso a aquellas que en efecto sean más significativas para explicar el fenómeno en cuestión.

Además, sería conveniente la inclusión de un parámetro de rango que regule dinámicamente la relación entre la distancia y la covarianza entre observaciones. Por ejemplo, aún cuando estén estandarizados los datos, una misma distancia podría significar una covarianza grande entre observaciones para alguna covariable o fenómeno, pero covarianza casi nula para otro. Lograr implementar este parámetro dinámico seguramente mejorará el ajuste.

Finalmente, tendría una gran utilidad el desarrollar una medida robusta de bondad de ajuste para este tipo de modelos. Esto brindaría cualidades importantes al modelo GPDP, como el poder hacer selección de variables, y también permitiría saber qué tan bueno o malo es el modelo, en comparación con lo demás disponibles.

Bibliografía

- Bannerjee, S. 2008. *Bayesian Linear Models: The Gory Details*. Notas del autor. Descargado de <http://www.biostat.umn.edu/ph7440/>.
- Dawid, A. P. 2016. Exchangeability and Its Ramifications. *Chap. 2, pages 19–29 of*: Damien, P., Dellaportas, P., Polson, N., & Stephens, D. A. (eds), *Bayesian Theory and Applications*. Oxford University Press.
- Denison, D. G. T., Holmes, C. C., Mallick, B. K., & Smith, A. F. M. 2002. *Bayesian Methods for Nonlinear Classification and Regression*. Wiley Series in Probability. Wiley.
- Dunson, D.B., & Taylor, J.A. 2005. Approximate Bayesian Inference for Quantiles. *Journal of Nonparametric Statistics*, **17**, 385–400.
- Fishburn, P. C. 1986. The Axioms of Subjective Probability. *Statistical Science*, **1**(3), 335–345.
- Hanson, T., & Johnson, W.O. 2002. Modeling Regression Error With a Mixture of Polya Trees. *Journal of the American Statistical Association*, **97**, 1020–1033.
- Hao, L., & Naiman, D.Q. 2007. *Quantile Regression*. Quantile Regression series, no. 149. Thousand Oaks, California: SAGE Publications.

- Kalli, M., Griffin, J.E., & Walker, S.G. 2009. Slice Sampling Mixture Models. *Statistics and Computing*, **21**(1), 93–105.
- Koenker, R., & Bassett, G. 1978. Regression Quantiles. *Econometrica*, **46**(1), 33–50.
- Kottas, A., & Gelfland, A.E. 2001. Bayesian Semiparametric Median Regression Modeling. *Journal of the American Statistical Association*, **96**, 1458–1468.
- Kottas, A., & Krnjajic, M. 2005. *Bayesian Nonparametric Modeling in Quantile Regression*. Technical Report AMS 2005-06. University of California, Santa Cruz.
- Kottas, A., Krnjajic, M., & Taddy, M. 2007. Model-Based Approaches to Nonparametric Bayesian Quantile Regression. *Pages 1137–1148 of: Proceedings of the 2007 Joint Statistical Meetings*.
- Lavine, M. 1995. On an Approximate Likelihood for Quantiles. *Biometrika*, **82**, 220–222.
- Paisley, J. 2010. *A Simple Proof of the Stick-Breaking Construction of the Dirichlet Process*. Tech. rept. MIT.
- Rasmussen, C.E., & Williams, C.K.I. 2006. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning series. Cambridge, Massachusetts: the MIT Press.
- Robert, C. P., & Casella, G. 2009. *Introducing Monte Carlo Methods with R*. Berlin, Heidelberg: Springer-Verlag.
- Schervish, M.J. 1996. *Theory of Statistics*. Springer Series in Statistics. New York: Springer.
- Teh, Y. W. 2010. Dirichlet Process. *Pages 280–287 of: Sammut, C, & Webb, GI (eds), Encyclopedia of Machine Learning*. New York: Springer.

- Tsionas, E.G. 2003. Bayesian Quantile Inference. *Journal of Statistical Computation and Simulation*, **73**, 659–674.
- Walker, S.G., & Mallick, B.K. 1999. A Bayesian Semiparametric Accelerated Failure Time Model. *Biometrics*, **55**(2), 477–483.
- Wasserman, L. 2006. *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer-Verlag.
- Yu, K., & Moyeed, R. A. 2001. Bayesian Quantile Regression. *Statistics & Probability Letters*, **54**(4), 437–447.

Apéndice A

Distribuciones de probabilidad

A.1. Distribución Normal condicional

Propiedad. Sea $X \in \mathbb{R}^m$ un vector aleatorio que tiene distribución Normal conjunta y está particionado de la siguiente manera:

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix},$$

con dimensiones $\begin{bmatrix} (m - q) \\ q \end{bmatrix}$.

Entonces, la media $\mu \in \mathbb{R}^m$ y varianza $\Sigma \in \mathbb{R}^{m \times m}$ de X se pueden escribir

$$\begin{aligned} \mu &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \\ \text{con dimensiones} &\begin{bmatrix} (m-q) \\ q \end{bmatrix}, y \\ \Sigma &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \\ \text{con dimensiones} &\begin{bmatrix} (m-q) \times (m-q) & (m-q) \times q \\ q \times (m-q) & q \times q \end{bmatrix}. \end{aligned}$$

La distribución condicional de X_2 , sujeta a que $X_1 = a$ es Normal con $X_2|X_1 = a \sim \mathcal{N}(X_2|\bar{\mu}, \bar{\Sigma})$, donde

$$\begin{aligned} \bar{\mu} &= \mu_2 + \Sigma_{2,1}\Sigma_{11}^{-1}(a - \mu_1) \\ \bar{\Sigma} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}. \end{aligned}$$

A.2. Distribución de Dirichlet

Definición 5. Se dice que un vector aleatorio $x \in \mathbb{R}^n$ se distribuye de acuerdo a la **distribución de Dirichlet** ($\mathbf{x} \sim \text{Dir}(\alpha)$) con vector de parámetros α , específicamente,

$$x = \begin{pmatrix} x_1 \\ \dots \\ x_n \end{pmatrix}, \quad \alpha = \begin{pmatrix} \alpha_1 \\ \dots \\ \alpha_n \end{pmatrix},$$

para los cuales se cumplen las restricciones

$$x_i > 0, \forall i \in \{1, \dots, n\}$$

$$\sum_{i=1}^n x_i = 1$$

$$\alpha_i > 0, \forall i \in \{1, \dots, n\},$$

si su función de densidad es

$$f(x|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^n x_i^{\alpha_i-1},$$

donde B es la función Beta multivariada, y puede ser expresada en términos de la función Γ como

$$B(\alpha) = \frac{\prod_{i=1}^n \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^n \alpha_i)}, \quad \alpha = (\alpha_1, \dots, \alpha_n).$$

La esperanza y varianza de cada x_i son los siguientes:

$$\mathbb{E}[x_i] = \frac{\alpha_i}{\sum_{k=1}^n \alpha_k}$$

$$Var(x_i) = \frac{\alpha_i (\sum_{k=1}^n \alpha_k - \alpha_i)}{(\sum_{k=1}^n \alpha_k)^2 (\sum_{k=1}^n \alpha_k + 1)}$$

Es común que esta distribución sea usada como la inicial conjugada de la distribución multinomial, debido a que el vector x tiene las mismas propiedades de una distribución de probabilidad discreta (elementos positivos y que en conjunto suman 1).

Apéndice B

Algoritmos MCMC¹⁸

B.1. Introducción

Los algoritmos MCMC son utilizados para aproximar distribuciones de probabilidad, normalmente complejas. La idea es lograr simular una muestra de la distribución, para poder aproximar sus características. Entre más grande sea la muestra, mejor será la estimación.

Para hacer esto simula cadenas de Markov de los distintos elementos de la distribución compleja, y, bajo el supuesto de que se alcanza la distribución estacionaria, toma al conjunto de dichas esas simulaciones como una muestra de la distribución original. De hecho, el nombre MCMC viene del inglés *Markov chain Monte Carlo*, haciendo también referencia a la simulación de Monte Carlo para cada iteración.

¹⁸Las ideas de este apéndice son retomadas de Robert & Casella (2009)

B.2. Simulador de Gibbs

Se trata de un caso particular de los algoritmos *MCMC*, y a continuación se analizan dos tipos, siendo el segundo una generalización del primero.

B.2.1. Simulador de Gibbs de dos pasos

Funciona de la siguiente manera: si dos variables aleatorias X y Y tienen una densidad conjunta $f(x, y)$, con sus correspondientes densidades condicionales $f_{Y|X}$ y $f_{X|Y}$, se genera una cadena de Markov (X_t, Y_t) de acuerdo al siguiente algoritmo:

Algoritmo 1: Simulador de Gibbs de dos pasos

```

Tomar  $X_0 = x_0$  arbitraria ;
para  $t = 1, 2, \dots, n$  hacer
    | 1.  $Y_t \sim f_{Y|X}(y|x_{t-1})$ 
    | 2.  $X_t \sim f_{X|Y}(x|y_t)$ 
fin

```

La convergencia de la cadena de Markov está asegurada, a menos que los soportes de las condicionales no estén conectados.

B.2.2. Simulador de Gibbs de múltiples pasos

Sea $\mathbb{X} \in \mathcal{X}$ una variable aleatoria que puede ser escrita como $\mathbb{X} = (X_1, \dots, X_p)$, con $p \in \mathbb{Z}^+$, y donde las X_i 's bien pueden ser unidimensionales o multidimensionales. Además, es posible encontrar las distribuciones condicionales,

de forma que

$$X_i | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p \sim f_i(x_i | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p),$$

$$i \in \{1, \dots, p\}.$$

El correspondiente algoritmo de Gibbs está dado por:

Algoritmo 2: Simulador de Gibbs de múltiples pasos

Tomar $\mathbf{x}^{(0)} = (x_1^{(0)}, \dots, x_p^{(0)})$ arbitraria;

para $t = 1, 2, \dots, n$ **hacer**

1. $X_1^{(t)} \sim f_1(x_1 | x_2^{(t-1)}, \dots, x_p^{(t-1)})$

2. $X_2^{(t)} \sim f_2(x_2 | x_1^{(t)}, x_3^{(t-1)}, \dots, x_p^{(t-1)})$

...

k . $X_k^{(t)} \sim f_k(x_k | x_1^{(t)}, \dots, x_{k-1}^{(t)}, x_{k+1}^{(t-1)}, \dots, x_p^{(t-1)})$

...

p . $X_p^{(t)} \sim f_p(x_p | x_1^{(t)}, \dots, x_{p-1}^{(t)})$

fin

Cabe resaltar que el desempeño puede estar fuertemente afectado por la parametrización del modelo. Por ello puede resultar una buena idea reparametrizar el modelo, buscando que las componentes sean lo más independientes posible.

B.3. Monitoreo de convergencia y adaptación de los algoritmos MCMC

B.3.1. Monitoreo de convergencia a la *estacionariedad*

El primer requisito de convergencia de un algoritmo MCMC es que la distribución de la cadena $(x^{(t)})$ sea la distribución estacionaria f . Una meta menos ambiciosa sería que sea independiente del punto inicial $x^{(0)}$, después de muchas realizaciones de la cadena. La principal herramienta para verificar *estacionariedad* es correr varias cadenas en paralelo, para poder comparar sus rendimientos.

Un primer acercamiento empírico al control de convergencia es el dibujar gráficas de las cadenas simuladas (componente a componente o juntas), para detectar valores muy desviados y comportamientos no estacionarios.

Otro diagnóstico gráfico que se puede utilizar es la *traza*, es decir, la gráfica de cada uno de los valores de la cadena en el eje y , contra su respectivo número de iteración en el eje x . Así será posible observar cuando la cadena tiene un comportamiento repetitivo en ciertos valores y a partir de qué momento se distribuye sobre todo el soporte, es decir, a partir de qué iteración alcanza la distribución estacionaria.

B.3.2. Monitoreo de convergencia a los promedios

Una vez cubierta la distribución estacionaria, se verifica la convergencia del promedio aritmético

$$\frac{1}{T} \sum_{t=1}^T h(x^{(t)})$$

a la esperanza $\mathbb{E}_f[h(x)]$, para una función h arbitraria. Esta propiedad se denomina comúnmente *ergodicidad*.

La herramienta inicial y más natural suele ser el graficar la evolución del estimador del promedio, conforme crece T . Si dicha curva no se ha estabilizado después de T iteraciones, habría que incrementar la longitud de la cadena de Markov.

B.3.3. Monitoreo de convergencia a una muestra *iid*

Para finalizar, idealmente, la aproximación de f obtenida de los algoritmos MCMC se debería extender a la producción (aproximada) de muestras *iid* de f . La técnica más usada para lograr esto es el *submuestreo o refinamiento*, donde se consideran sólo los valores $y^{(t)} = x^{(kt)}$, para cierta k .

Como medidas diagnósticas normalmente se usan las siguientes: la autocorrelación dentro de cada variable aleatoria que es parte del simulador de Gibbs; y la correlación cruzada entre las distintas variables aleatorias, dado que se busca independencia entre ellas.

Apéndice C

Predicción de cuantiles, utilizando el modelo tradicional de regresión a la media

Una vez ajustado el modelo tradicional de regresión a la media (descrito en la sección 3.2.2), un problema de interés es predecir algún cuantil en específico de la variable de respuesta, dado un nuevo valor en las covariables.

Recordando los supuestos del modelo, se tiene que

$$y = x^T \beta + \varepsilon,$$

con $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, tal que β y σ^2 se piensan como constantes, pero desconocidas.

Dicha incertidumbre se traduce en una distribución de probabilidad para los parámetros, conocida como *Normal-Gamma Inversa (NGI)*. Una de las mayores ventajas de usar tal distribución es que es conjugada respecto a la distribución Normal de los datos. Por lo tanto, una vez observados los datos, la distribución posterior de los parámetros (β, σ^2) continuará siendo *NGI*.

Suponiendo que se tiene un nuevo vector de covariables $x_* \in \mathbb{R}^n$ para el cual se desea hacer una predicción y_* de la variable de respuesta, su distribución de probabilidad es

$$y_*|x_*, \beta, \sigma^2 \sim \mathcal{N}(x_*^T \beta, \sigma^2).$$

Si el interés del modelador es estimar el cuantil p -ésimo de y_* , se tiene entonces que

$$q_p(y_*|x_*, \beta, \sigma^2) = x_*^T \beta + q_p(\varepsilon|\sigma^2),$$

con $q_p(\varepsilon|\sigma^2)$ el cuantil p -ésimo de $\mathcal{N}(0, \sigma^2)$.

Como es posible notar en el resultado anterior, una vez que se tiene una realización de la distribución posterior de los parámetros (β, σ^2) , obtener una realización de la predicción del cuantil únicamente requiere de la aplicación de una función, dejando a un lado toda aleatoriedad. En ese orden de ideas, será posible estimar la distribución posterior predictiva del cuantil p -ésimo mediante la simulación de un número grande de sus propias realizaciones, a partir de simular valores de la distribución posterior de los parámetros.

Sea k el número de simulaciones a obtener de la distribución predictiva, y recordando que $(\beta, \sigma^2) \sim \mathcal{NGI}(\bar{M}, \bar{V}, \bar{a}, \bar{b})$ (la distribución posterior de los

parámetros), se puede reescribir como

$$\begin{aligned}\sigma^2 &\sim \mathcal{GI}(\bar{a}, \bar{b}), \\ \beta &\sim \mathcal{N}(\bar{M}, \sigma^2 \bar{V}).\end{aligned}$$

Se tiene, entonces, que el algoritmo para obtener realizaciones de la distribución posterior predictiva del cuantil p -ésimo se puede representar de la siguiente forma.

Algoritmo 3: Simulador predictivo del cuantil p -ésimo, usando el modelo tradicional de regresión a la media.

```

para  $t = 1, 2, \dots, k$  hacer
    1. Simular  $\sigma_t^2 \sim \mathcal{GI}(\bar{a}, \bar{b})$ 
    2. Simular  $\beta_t \sim \mathcal{N}(\bar{M}, \sigma_t^2 \bar{V})$ 
    3. Obtener  $q_p(y_*)_t = x_*^T \beta_t + q_p(\varepsilon | \sigma_t^2)$ 
fin
```
