Un modelo Bayesiano y no paramétrico de regresión sobre cuantiles

Tesis para obtener el título de Licenciado en Matemáticas Aplicadas

Carlos Omar Pardo Gómez

Asesor: Dr. Juan Carlos Martínez Ovando

Instituto Tecnológico Autónomo de México

20 de abril del 2018

Contenido

Contenido

Introducción

- Propósito **modelos de regresión**: aproximar la **distribución** de una **variable aleatoria**, **condicional** al valor de otras **variables explicativas**.

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

- Comúnmente se supone a ***y*** como la **suma** de un **parámetro** que está en **función** del valor de las **variables explicativas**, y un **error aleatorio**, **independiente** de ellas.

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

Introducción

- La **media** ha sido parámetro tradicionalmente usado, dando lugar a los **modelos de regresión a la media**.

- Ventajas:
  - **Bajo costo** de estimación, en tiempo y recursos.
  - Facilidad de **interpretación** de sus **parámetros** (principalmente del modelo lineal).

Introducción

- Sin embargo, según Hao & Naiman (2007), estos modelos tienen 3 grandes limitaciones:
  - **Inferencia** puede ser acertada para la media, pero **inexacta** para **valores lejanos** a ella.
  - Los **valores atípicos** pueden **sesgar** la estimación de la **media**.
  - Forma funcional de los **cuantiles** depende de la elección del **error aleatorio**.

Introducción

- **Alternativas** a los modelos de regresión a la media:
  - Modelos de regresión a la **mediana** (1760).
  - Modelos de regresión sobre **cuantiles**[1] (1978), siendo la mediana un caso particular. (Ya no se usa necesariamente una medida de tendencia central).

[1] El **cuantil p-ésimo** es aquel valor, tal que el $\mathbf{p} \times \mathbf{100}\,\%$ de los valores están por **debajo** de él, y el $(\mathbf{1} - \mathbf{p}) \times \mathbf{100}\,\%$, por **encima**.

Introducción

- **Objetivo: proponer un modelo**:
  - De regresión sobre **cuantiles**
  - **Bayesiano**[2]
  - Con **error no paramétrico**
  - Relación **no lineal** entre variable de respuesta y covariables,

retomando las ideas de *Kottas et al.(2007)* y *Kottas & Krnjajic (2005)*.

[2] Esta tesis da como aceptados los axiomas de coherencia de la Teoría de la Decisión.

Contenido

Modelos de regresión a la media

- **Modelo general**

$$y = f(x) + \varepsilon, \text{ tal que } \mathbb{E}[\varepsilon] = 0$$
$$\implies \mathbb{E}[\boldsymbol{y}|\boldsymbol{x}] = \boldsymbol{f}(\boldsymbol{x})$$

- **Modelo tradicional**

$$f(x) = x^T\beta \textbf{ (relación lineal)},$$
$$\varepsilon \sim \mathcal{N}(0, \sigma^2) \textbf{ (error paramétrico)}$$
$$\beta, \sigma^2 \sim \mathcal{NGI}(M, V, a, b)$$

Función cuantil

Definición

Sea F_y la función de distribución acumulada de la variable aleatoria $\boldsymbol{y}$, entonces la **función que regresa su cuantil p-ésimo** se escribe

$$q_p(y) = \text{ínf}\left\{x \in \mathbb{R} : p \leq F_y(x)\right\};$$

que se puede simplificar a

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

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

Modelos de regresión sobre cuantiles

- El modelador **elige el parámetro p** de su interés.

- **Modelo general**

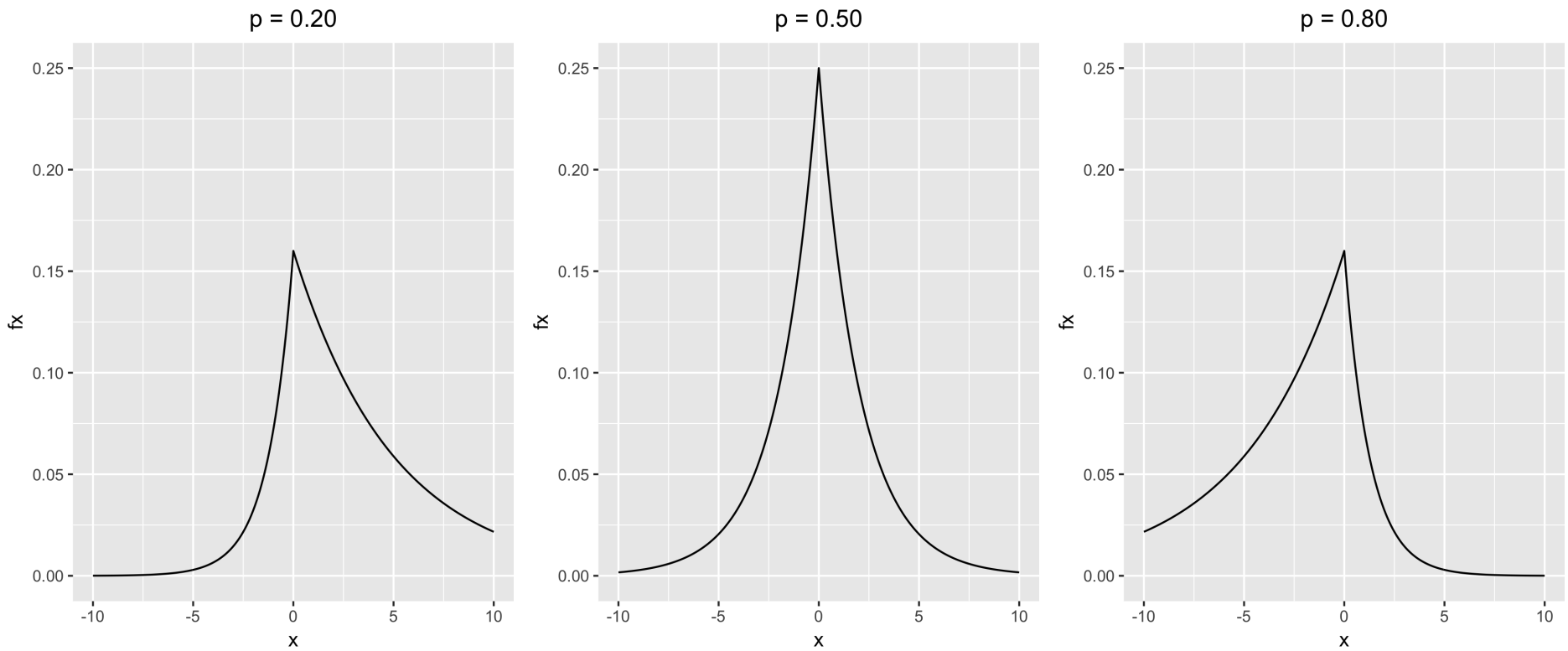
$$y = f_p(x) + \varepsilon_p, \text{ tal que } q_p(\varepsilon_p) = 0$$
$$\Longrightarrow \mathbf{q_p(y|x) = f_p(x)}$$

- **Modelo tradicional**

$$f_p(x) = x^T \beta_p \textbf{ (relación lineal)},$$
$$\varepsilon_p \sim \mathcal{AL}_p(\sigma) \textbf{ (error paramétrico)}$$
$$\beta_p, \sigma \sim \mathcal{NGI}(M, V, a, b)$$

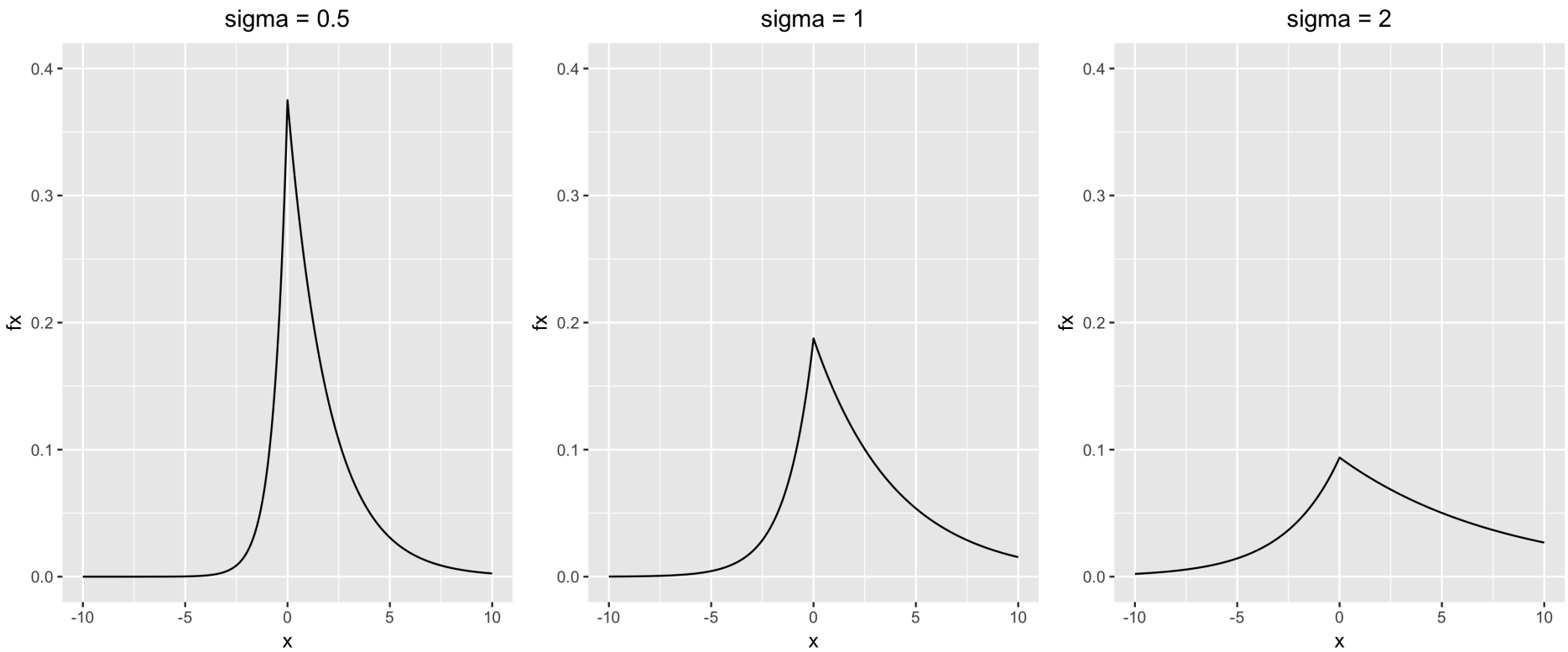
Distribución asimétrica de Laplace

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



Distribución asimétrica de Laplace

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



Contenido

Estadística no paramétrica

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**."

Contenido

Generalización de la relación lineal

- Idea: **generalizar relación lineal** $f_p(x) = x^T \beta_p$, a **cualquier posible función**.
- Medida de probabilidad para $\boldsymbol{\beta_p} \rightarrow$ Medida de probabilidad para $\boldsymbol{f_p}$.
- Como f_p está definida para múltiples valores de x, se trata de un conjunto de variables aleatorias que **depende de variables de entrada**: un **proceso estocástico**.
- Para el caso particular de esta tesis, se pensará que f_p sigue la ley de probabilidad de un **proceso Gaussiano**.

Contenido

Error no paramétrico

Intuición de los procesos de Dirichlet

Las realizaciones $\varepsilon_1, ..., \varepsilon_n$ provienen de una **distribución G**, la cual es **desconocida** para el modelador.

Para reflejar su **incertidumbre**, le asigna una **ley de probabilidad** a los posibles valores de G, particularmente la de un **proceso de Dirichlet**.

Es decir, **la realización de un proceso de Dirichlet es una distribución de probabilidad**.

Contenido

Descripción Modelo GPDP

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

Github : **opardo/GPDPQuantReg**

3 funciones públicas:

- **GPDPQuantReg**: ajusta el modelo con los datos que recibe
- **predict**: realiza predicción para un conjunto de covariables
- **diagnose**: diagnóstico de las cadenas de Markov del simulador de Gibbs [3]

[3] Algoritmo MCMC, para realizar inferencia Bayesiana.

Contenido

Comparación de modelos

	Modelo Tradicional	**Modelo GPDP**
Parámetro de regresión	A la media	Sobre cuantiles
Relación entre *x* y *y*	Lineal	No lineal
Tipo de error	Paramétrico	No paramétrico

Simulación de datos

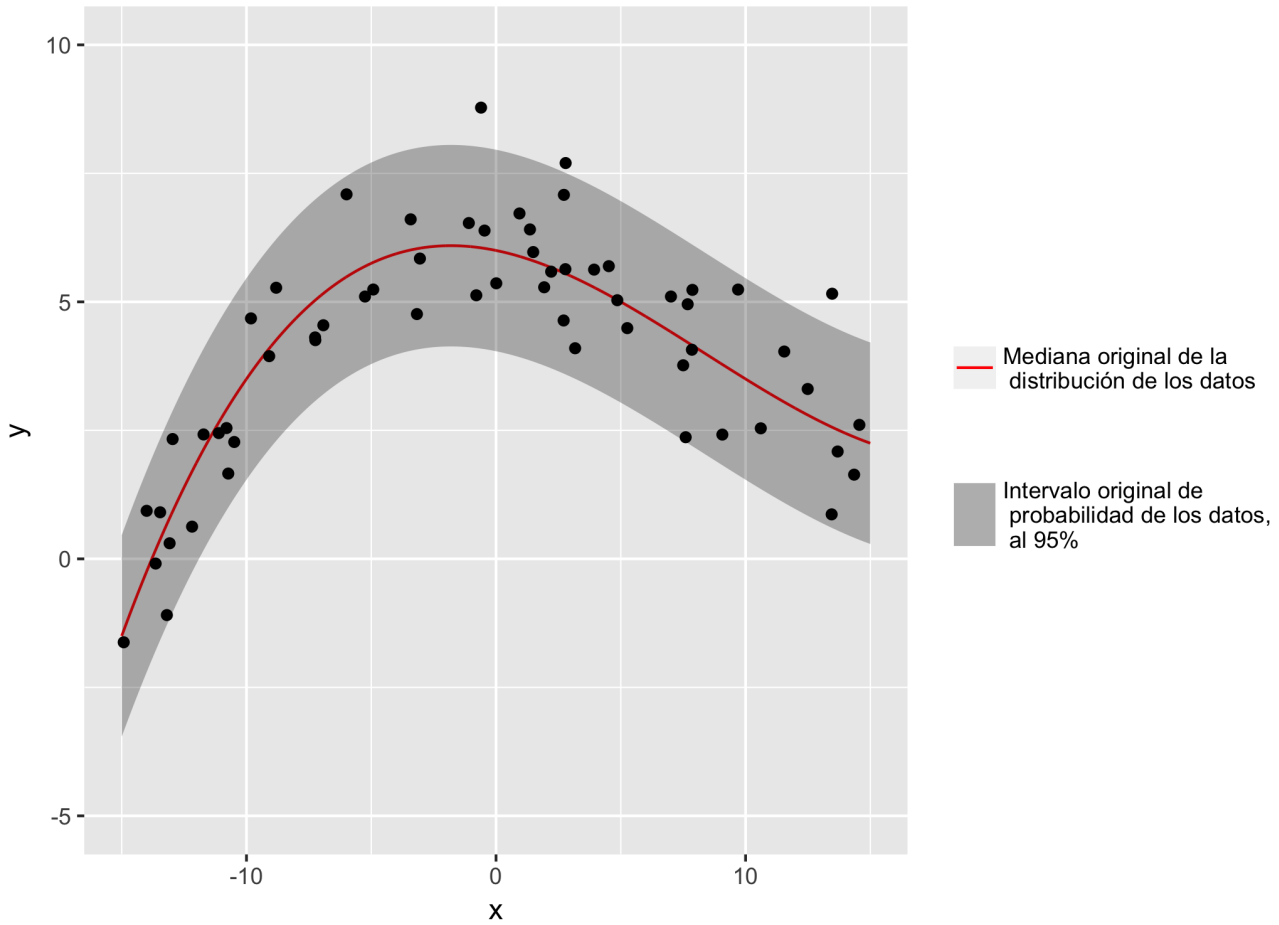
$$y = g(x) + \omega,$$

$g(x)$: función determinista,

ω : variable aleatoria

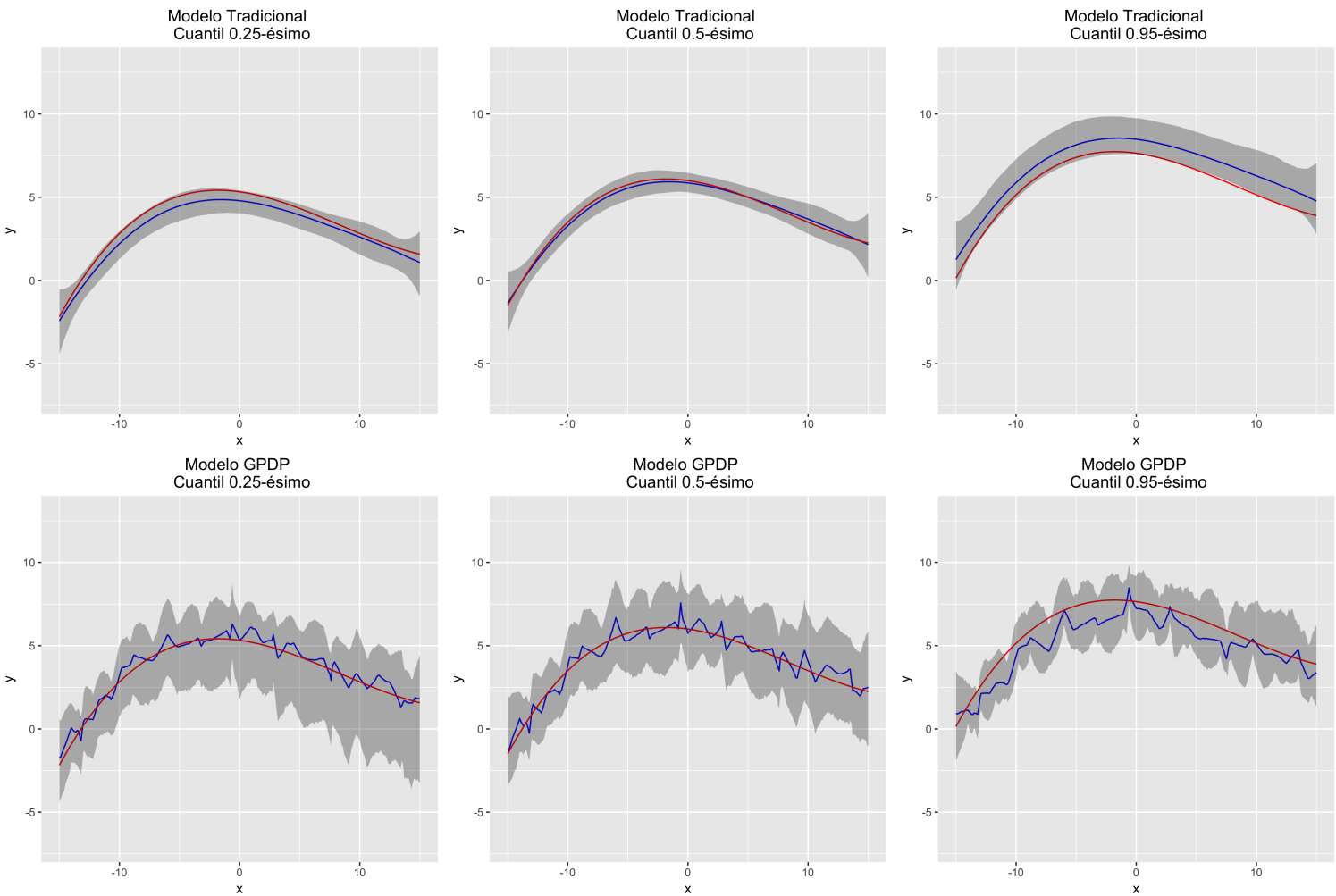
$$\Longrightarrow \boldsymbol{f_p(x) = g(x) + q_p(\omega)}$$

Supuestos tradicionales de regresión



$$g(x) = \frac{1}{1000}x^3 - \frac{1}{40}x^2 - \frac{1}{10}x + 6,$$
$$\omega \sim \mathcal{N}(0, 1)$$

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 %.

Supuestos tradicionales de regresión

Cuadro: Error cuadrático medio entre mediana predictiva y cuantil real

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

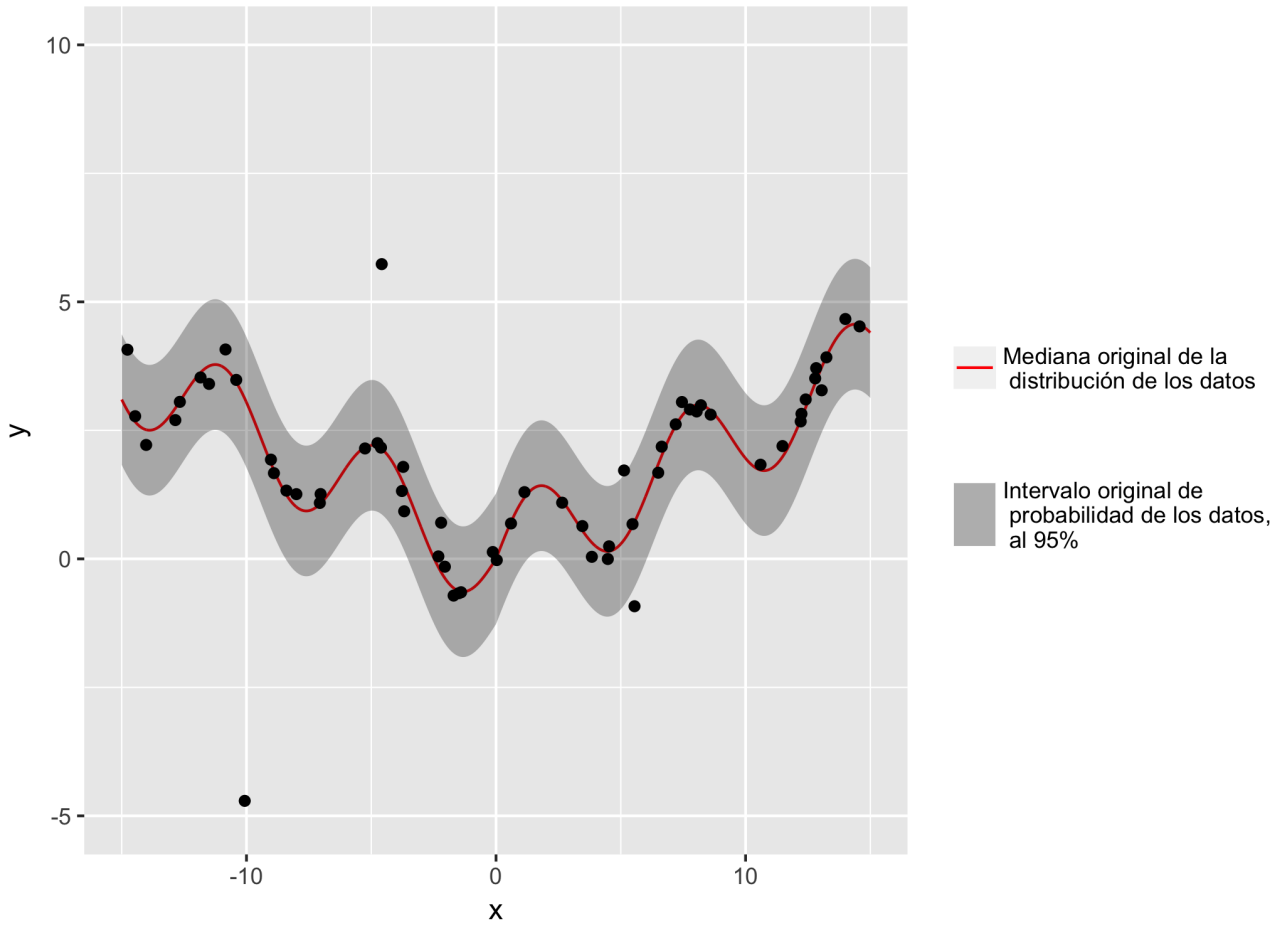
Cuadro: Correlación al cuadrado entre mediana predictiva y cuantil real

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

Cuadro: Porcentaje de valores reales dentro del intervalo de confianza al 95 %

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

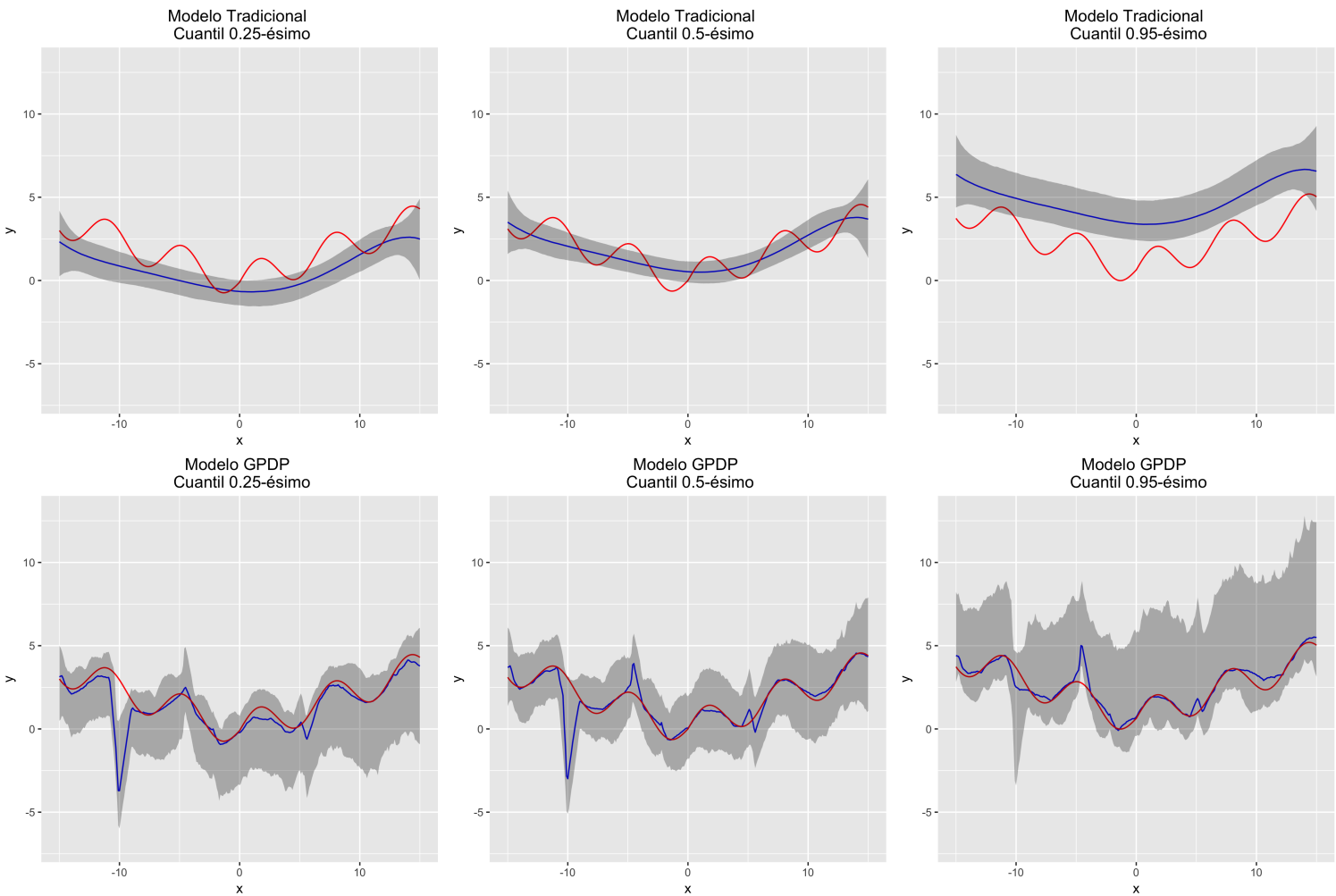
Error aleatorio de colas pesadas



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

$$\omega \sim Cauchy(0, 0.1)$$

Error aleatorio 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 %.

Error aleatorio de colas pesadas

Cuadro: Error cuadrático medio entre mediana predictiva y cuantil real

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

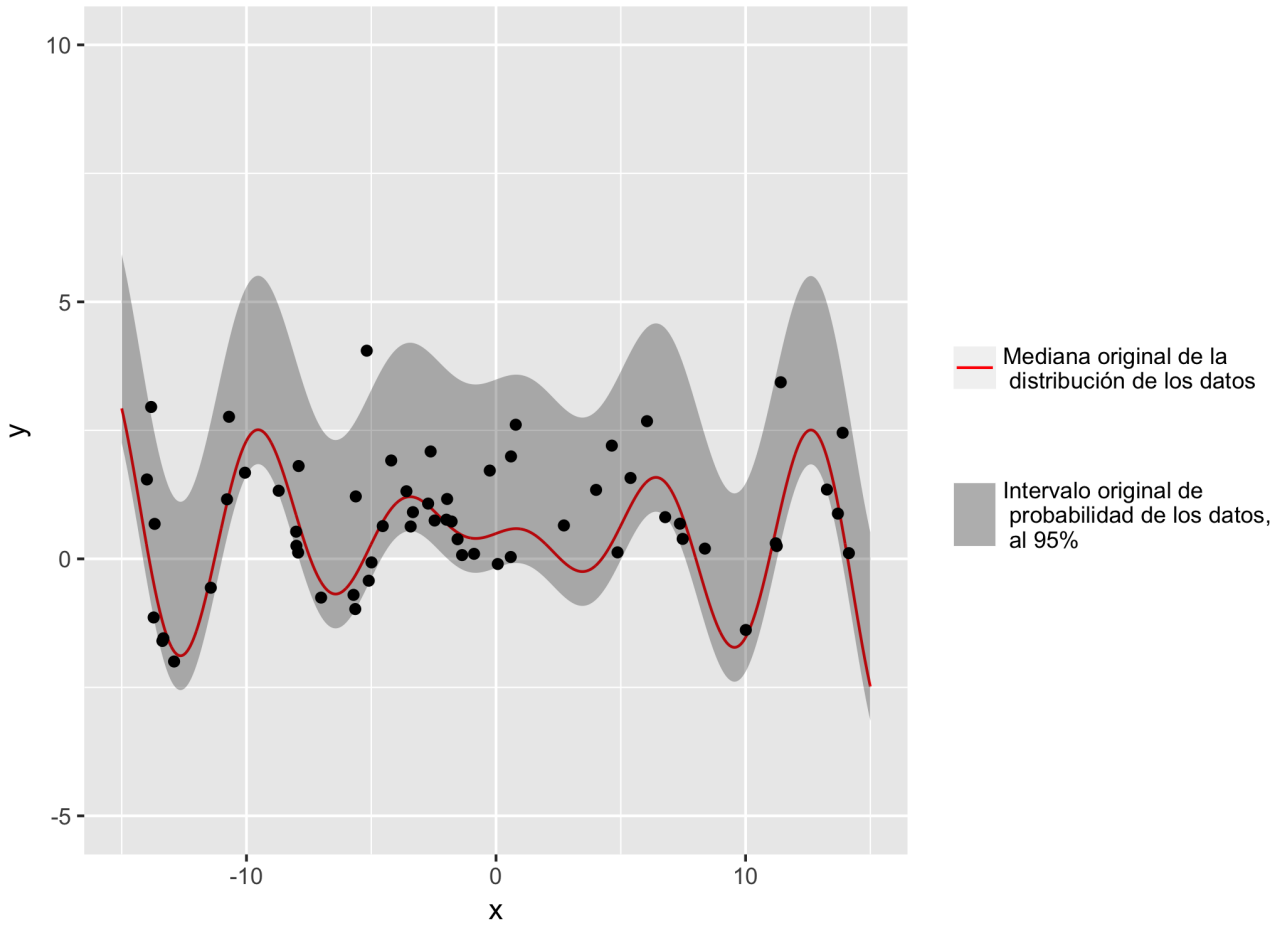
Cuadro: Correlación al cuadrado entre mediana predictiva y cuantil real

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

Cuadro: Porcentaje de valores reales dentro del intervalo de confianza al 95 %

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

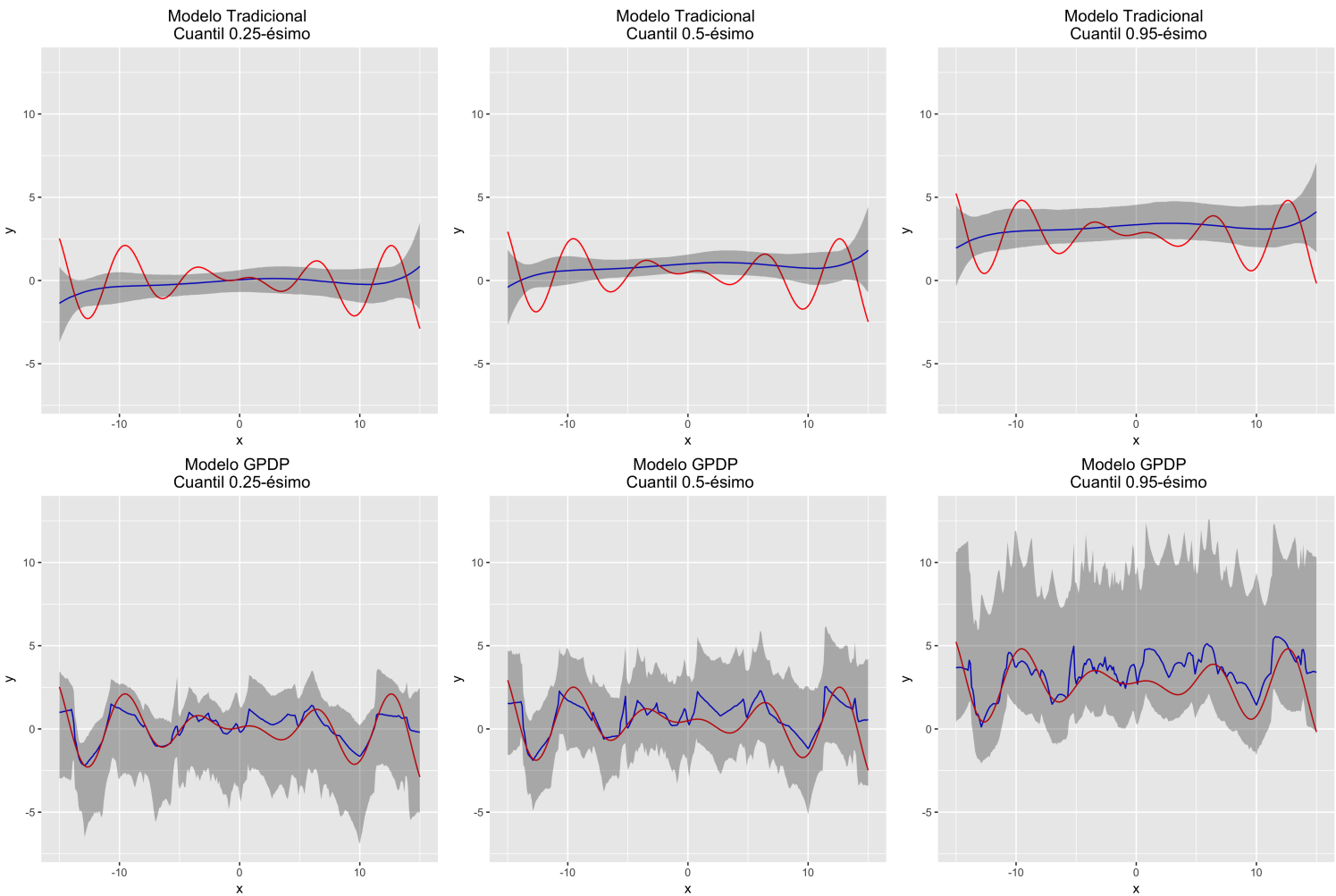
Error aleatorio asimétrico



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

$$\omega \sim Gamma(1,1)$$

Error aleatorio 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 %.

Error aleatorio asimétrico

Cuadro: Error cuadrático medio entre mediana predictiva y cuantil real

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

Cuadro: Correlación al cuadrado entre mediana predictiva y cuantil real

Cuantil	Modelo Tradicional	Modelo GPDP
0.95	0.01	0.50
0.50	0.01	0.61
0.25	0.01	0.69

Cuadro: Porcentaje de valores reales dentro del intervalo de confianza al 95 %

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

Comparación de tiempos

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

Datos	Tradicional (seg)	GPDP (seg)
Supuestos tradicionales	6	564
Colas pesadas	5	529
Heterocedasticidad	5	534
Error asimétrico	6	537
Discontinuidades	5	533

Contenido

Conclusiones

- Si bien los **modelos de regresión a la media** han sido de mucha utilidad en las últimas décadas, existen **contextos** en los que resultan **insuficientes**.

- 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.

- Un reto importante que presentó este trabajo fue el **desarrollo del paquete en R**, tanto por el **planteamiento teórico del simulador de Gibbs**, como por la búsqueda de una **programación general y eficiente**.

Trabajo futuro

- Proponer alguna manera de darle un **peso distinto a cada variable explicativa**, en el proceso Gaussiano. Actualmente toma una única distancia, dando igual peso a cada variable.

- Sería conveniente la inclusión de un **parámetro de rango** que regule dinámicamente la relación entre la **distancia y la covarianza** entre observaciones, en el proceso Gaussiano.

- Desarrollar una medida robusta de **bondad de ajuste**, que permita hacer selección de variables y comparación con otros modelos disponibles).

Un modelo Bayesiano y no paramétrico de regresión sobre cuantiles

Tesis para obtener el título de Licenciado en Matemáticas Aplicadas

Carlos Omar Pardo Gómez

Asesor: Dr. Juan Carlos Martínez Ovando

Instituto Tecnológico Autónomo de México

20 de abril del 2018