Regresión Cuantílica Bayesiana

Lautaro Lasorsa

2024

Índice

1. Prefacio

En este trabajo explicaré con mis palabras el capítulo 4 del libro *Handbook of Quantile Regression*, titulado *Bayesian Quantile Regression* y escrito por Huixia Judy Wang y Yunwen Yang. Cuando sea oportuno, expandiré con los conceptos.

Las secciones de esta explicación coincidirán con las secciones del capítulo original.

2. Introducción

El problema de la regresión cuantílica consiste en, justamente, modelar no la esperanza sino un cierto cuantil de una distribución condicional.

Es decir, queremos modelar:

$$Y \sim F_{Y|X}(y|x) \quad \text{donde} \quad F_{Y|X}(y|x) = P(Y \leq y|X = x)$$

Y expresar:

$$q_\tau(x) = F^{-1}_{Y|X}(\tau|x) = \phi(x, \beta_\tau)$$

Donde τ es el cuantil que queremos modelar y ϕ es una función que depende de los parámetros β_τ y las variables predictoras x.

Es un problema de interés por ejemplo en economía, permitiendo modelar el impacto de las variables predictoras en distintos estratos sociales.

Es notable que el modelo $q_\tau(x) = \phi(x, \beta_\tau)$ habla de ese cuantil en específico sin ser un modelo de la distribución completa parametrizado por β_τ. Un aspecto interesante será traducir nuestro requerimiento de aproximar un cuantil al problema de estimar la distribución de los parámetros de una distribución, eligiendo la familia de distribuciones adecuada.

Por otro lado, las motivaciones para utilizar modelos bayesianos son:

- Obtienen de forma natural tanto el punto como la incertidumbre de la estimación.

- El cómputo con MCMC del posterior es más sencillo que resolver optimizaciones complejas, como la optimización no-convexa usada en Powell (1986).

Este trabajo enfatiza que un modelo que funciona es una herramienta específica para un caso de uso, y no necesita ni garantiza funcionar para otras situaciones.

3. Verosimilitud Asimétrica de Laplace

Dado un cuantil τ, podemos definir la regresión lineal cuantílica como:

- $Q_Y(\tau|X = x) = x^T\beta_\tau$
- $\rho_\tau(u) = u(\tau - I(u < 0))$
- $\hat{\beta_\tau} = \arg\min_{\beta_\tau} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T\beta_\tau)$

Lo interesante de este modelo es que la función de pérdida $\rho_\tau(\cdot)$ es proporcional al logaritmo de la función de densidad de la distribución asimétrica de Laplace. Por tanto, podemos utilizar esta distribución en un modelo bayesiano:

$$L(\beta_\tau; Data) = \frac{\tau^n(1-\tau)^n}{\sigma^n} \exp\left(-\sum_{i=1}^{n} \frac{\rho_\tau(y_i - x_i^T\beta_\tau)}{\sigma}\right)$$

De esta forma, sea $\pi(\beta_\tau)$ el prior de β_τ, el posterior es:

$$p_n(\beta_\tau|Data) \propto \pi(\beta_\tau) \exp(-\sum_{i=1}^{n} \frac{\rho_\tau(y_i - x_i^T\beta_\tau)}{\sigma})$$

Algo que no enfatiza el libro pero creo que es importante es tener en cuenta que en este caso la likelihood $L(\beta_\tau; Data)$ no modela la distribución de los datos (es decir, no se modela que $y \sim AL(\beta_\tau, \sigma)$), sino que modela la verosimilitud de observar esos datos dado que el cuantil τ condicional es $x^T\beta_\tau$. Por otro lado, el posterior $p_n(\beta_\tau|Data)$ sí modela la distribución de los parámetros dados los datos.

El paper marca como un hecho importante que incluso tomando un prior impropio (que no integra 1 en todo el espacio), el posterior sigue siendo propio (integrable), por lo que puede ser muestreado mediante MCMC.

Además, para facilitar la generación del posterior este capítulo comenta una igualdad interesante (de Kotz et al. (2001)):

Si:

$$f_{AL}(x) = \frac{\tau(1-\tau)}{\sigma} \exp\left(-\frac{\rho_\tau(x)}{\sigma}\right)$$

Entonces:

- $x = \sigma * (\theta * v + \gamma * \sqrt{v} * z)$
- $\theta = \frac{1-2*\tau}{1-\tau}$
- $\gamma^2 = \frac{2}{\tau*(1-\tau)}$
- $z \sim N(0,1)$
- $v \sim exp(1)$

Esta representación, también llamada mezcla escalada de normales, así como otras de la misma categoría, permitieron desarrollar algoritmos de muestreo y aumento de datos para la regresión cuantílica.

El estimador $\hat{\beta}_\tau$ obtenido mediante el modelo posterior es un estimador consistente y asintóticamente normal, sin embargo esto no garantiza que sea útil para el problema que se busca resolver. El primer trabajo en poner esto a prueba es Chernozhukov y Hong (2003). Sobre esta base, Yang et al. (2016) propone una forma específica de corregir el posterior para que sea útil en la práctica.

Lo que indican es que las condiciones que garantizan la normalidad asintótica de $\hat{\beta}_\tau$ implican que $||\beta_\tau - \beta_\tau^0|| = O(n^{-1/2})$, donde β_τ^0 es el verdadero valor de β_τ, y que, si σ es conocido, el posterior es:

$$p_n(\beta_\tau|Data) \propto \pi(\beta_\tau) \exp(-\frac{n * (\beta_\tau - \hat{\beta}_\tau)^T D_1 (\beta_\tau - \hat{\beta}_\tau) + o_p(1)}{2\sigma})$$

Donde:

- Si $\pi(\beta_\tau)$ es uniforme, $p_n(\beta_\tau|Data)$ es aproximadamente una normal multivariada con media $\hat{\beta}_\tau$ y varianza $\hat{\Gamma}$.
- $\hat{\Gamma} = \frac{\sigma}{n} D_1^{-1}$
- $$D_1 = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f_{Y_i}(x_i^T \beta_\tau^0) x_i x_i^T$$

  Notar que permite que los distintos y_i tengan distintas distribuciones condicionales.
- Por otro lado, la covarianza asintótica de $n^{1/2}\hat{\beta}_\tau$ es $\tau(1-\tau) D_1^{-1} D_0 D_1^{-1}$.
- $$D_0 = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T$$

Sin embargo, como la varianza del posterior no es la mejor aproximación a la varianza del sampleo de $\hat{\beta_\tau}$, Yang et al. (2016) propone una corrección a la varianza del posterior para que sea más útil en la práctica. La corrección es:

$$\hat{\Gamma}_{adj} = \frac{\tau(1-\tau)}{\sigma^2}\hat{\Gamma}(\sum_{i=1}^{n} x_i x_i^T)\hat{\Gamma}$$

Donde además $\hat{\Gamma}_{adj}$ es asintóticamente independiente de σ, por lo que la elección de σ afecta menos a este estimador. Con una idea similar se propusieron ajustes de la varianza del posterior para problemas relacionados, como regresión cuantílica censurada.

Un ejemplo que visualiza la utilidad de este estimador, propuesto en el propio capítulo, es:

- $T_i = 2{,}5 + 5x_i + \left(1 + (x_i - 0{,}5)^2\right) e_i$
- $e_i \sim t_3$ y $x_i \sim N(0,1)$, variables independientes.
- Aplicando una censura, la respuesta latente T_i será reemplazada por $y_i = max(T_i, 0)$ en un 30 % de los casos.
- Se estudia el cuantil $\tau = 0{,}5$, donde el cuantil condicional de T dado x es $a(\tau) + b(\tau)x$. $a(0{,}5) = 2{,}5$ y $b(0{,}5) = 5$.
- Para los demás cuantiles es no lineal en x.
- Se probaron 3 estimaciones: la bayesiana con likelihood asimétrica de Laplace, sin corregir (BAL) y corregida (BAL_{adj}), y un bootstrap ($Boot$) de 100 aplicaciones del estimador de Powell. Con cada uno se buscó construir un intervalo del 90 % de confianza para $a(0{,}5)$ y $b(0{,}5)$.

Los resultados anteriores pueden resumirse en la tabla **??**:

	100xECP		EML		100xECP		EML	
Method	$a(\tau)$	$b(\tau)$	$a(\tau)$	$b(\tau)$	$a(\tau)$	$b(\tau)$	$a(\tau)$	$b(\tau)$
	n = 200				n = 500			
BALadj	91	90	0.71	1.16	90	91	0.44	0.72
BAL	86	78	0.56	0.78	84	76	0.35	0.49
Boot	85	87	0.59	1.04	84	85	0.36	0.63

Cuadro 1: Resultados de las estimaciones

Donde ECP es la probabilidad empírica de cobertura y EML la longitud empírica media del intervalo. Se observa que BAL_{adj} es el único que satisface el nivel de confianza requerido.

4. Verosimilitud empírica

Un posible problema de utilizar una regresión cuantílica paramétrica, como la Laplace Asimétrica, es que un modelado incorrecto puede llevar a una estimación sesgada. Por tal motivo, se busca también estudiar un marco de trabajo pseudo-bayesiano utilizando verosimilitud empírica, y que además permite extender a la regresión multi-cuantílica.

Asumiendo el modelo $Q_Y(\tau|X = x) = x^T\beta_\tau$, entonces el estimador $\hat{\beta}_\tau$ satisface:

$$\frac{1}{n}\sum_{i=1}^{n}(\tau - I(y_i - x_i^T\hat{\beta}_\tau))x_i \approx 0$$

Para el modelo multi-cuantílico, si estamos interesados en k cuantiles $0 < \tau_1 < \tau_2 < ... < \tau_k < 1$, definimos $\zeta^0 = (\beta^0_{\tau_1}, \beta^0_{\tau_2}, ..., \beta^0_{\tau_k}) \in R^{k*p}$. Sea $m(x, y, \zeta)$ la función de estimación kp dimensional, sus componentes son:

$$m_{dp+j}(x, y, \zeta) = x_j(\tau_{d+1} - I(y < x^T\beta_{\tau_{d+1}}))$$

Para $d = 0, 1, ..., k-1$ y $j = 0, 2, ..., p-1$.
Entonces, el perfil de verosimilitud empírica es:

$$R(\zeta) = \text{máx}\left(\prod_{i=1}^{n}(n * w_i)|\sum_{i=1}^{n} w_i * m(x_i, y_i, \zeta) = 0, w_i \geq 0, \sum_{i=1}^{n} w_i = 1\right)$$

y además:

- La función de verosimilitud empírica es $R(\zeta) * n^{-n} = \prod_{i=1}^{n} w_i(\zeta)$
- $w_i(\zeta) = (n * (1 + \lambda_n(\zeta)^T * m(x_i, y_i, \zeta)))^{-1}$
- $$\sum_{i=1}^{n}\frac{m(x_i, y_i, \zeta)}{1 + \lambda_n^T(\zeta) * m(x_i, y_i, \zeta)} = 0$$
- $p(\zeta|Data) \propto R(\zeta) * \pi(\zeta)$. Notar que si el prior es impropio el posterior también puede serlo.

La ventaja de este método frente a un enfoque frecuentista es que en vez de buscar el ζ que maximiza la verosimilitud, solo es necesario poder evaluar la verosimilitud de un ζ dado.

Sin profundizar demasiado en las fórmulas, dada la intención de resumir, esto lleva a resultados interesantes:

- Si el prior es informativo esto puede pasar de pseudo-bayesiano a un enfoque bayesiano real. Esto permite agregar información en el modelo sobre la relación entre los β de distintos cuantiles y/o para distintas variables predictoras.

- $||\zeta - \zeta^0|| = O(n^{-1/2})$
- El enfoque bayesiano empírico es asintóticamente válido.

Agrego, como comentario, que esta posibilidad de estimar simultáneamente distintos cuantiles es muy interesante y algo que no vimos en el enfoque frecuentista en la otra materia que cursé este cuatrimestre donde la regresión cuantílica fue uno de los temas.

5. Verosimilitudes no paramétricas y semiparamétricas

5.1. Verosimilitud por mezcla

Planteando el modelo:

$$y_i = x_i^T \beta_\tau + \epsilon_i$$

Con $Q_\epsilon(\tau) = 0$. Este enfoque lo que hace es buscar representar el error como una mezcla de distribuciones, especialmente utilizando un proceso de Dirichlet (que es una forma particular de mezcla de distribuciones).
Entonces:

- f_ϵ es la distribución de ϵ.
- Sea $p(.;\theta)$ una distribución con cuantil τ igual a 0.
- Consideremos el modelo $f_1(\epsilon; G) = \int p(\epsilon;\theta) dG(\theta)$, $G \sim DP(\alpha, G_0)$.
- DP es un proceso de Dirichlet, que controla la distribución de los parámetros de la familia p.
- Notar que f_1 preserva que $Q_{f_1}(\tau) = 0$.
- Por ejemplo, Kottas and Gelgand (2001) plantea: $p(\epsilon;\theta,\psi) = f_N(\epsilon|0,\psi\theta)I(\epsilon < 0) + f_N(\epsilon|0,\theta/\psi)I(\epsilon \geq 0)$. Donde $f_N(\cdot|\mu,\sigma^2)$ es la densidad de una normal.

Incluso es posible utilizar mezcla de densidades para generar modelos jerárquicos, por ejemplo:

- $Y_i|\beta_\tau,\theta_i \sim^{ind.} AL(y_i - x_i^T\beta_\tau;\theta_i)$
- $\theta_i|G \sim^{i.i.d} G$
- $G|\alpha, G_0 \sim DP(\alpha, G_0)$
- Elegimos priors para β_τ, α y G_0.

Esto puede ser combinado con densidades no paramétricas para ganar flexibilidad y construir una amplia variedad de modelos.

5.2. Verosimilitud aproximada vía proceso de cuantiles

En este capítulo se plantea inicialmente, con una notación que encontré particularmente críptica, lo que pude entender como:

- $u_y = \{t \in (0,1) : x^T \beta_\tau = y\}$
- $$f_Y(y|x) = \lim_{\delta \to 0} \frac{\delta}{x^T(\beta_{\tau+\delta} - \beta_\tau)}$$

Luego, podemos generalizar a un modelo de regresión de múltiples cuantiles y obtenemos $\beta_K = (\beta_{\tau_1}, \beta_{\tau_2}, ..., \beta_{\tau_K})$, introduciendo el modelo de densidad interpolada linealmente:

$$\begin{aligned}\hat{f}_Y(y_i|x_i, \beta_K) = I\{y_i \in (-\infty, x_i^T \beta_{\tau_1})\}\tau_1 f_l(y_i) + \\ \sum_{k=1}^{K-1} I\{y_i \in [x_i^T \beta_{\tau_k}, x_i^T \beta_{\tau_{k+1}})\} \frac{\tau_{k+1} - \tau_k}{x_i^T(\beta_{\tau_{k+1}} - \beta_{\tau_k})} + \\ I\{y_i \in [x_i^T \beta_{\tau_K}, \infty)\}(1 - \tau_K) f_r(y_i)\end{aligned} \tag{1}$$

Donde f_l y f_r son densidades de las colas de la distribución.

Posteriormente esta sección llega a algo más extremo y propone modelar no una grilla de cuantiles sino a todos los cuantiles a la vez mediante el siguiente modelo:

$$\beta_\tau^j = \sum_{m=1}^{M} B_m(\tau) * \alpha_{j,m}$$

donde:

- β_τ^j es el coeficiente asociado a la j-ésima variable predictora y el cuantil τ.
- $B_m(\tau) = \binom{M}{m} \tau^m (1-\tau)^{M-m}$
- $\alpha_{j,m}$ son coeficientes desconocidos.
- $\sum_{i=1}^{p} x_{i,j}\alpha_{j,m} \geq \sum_{i=1}^{p} x_{i,j}\alpha_{j,m-1}$ es suficiente para garantizar la monotonía de los cuantiles.

Posteriormente, se discute la posibilidad de introducir una *variable espacial s* y hacer que los coeficientes $\alpha_{j,m}$ dependan de esta variable, permitiendo modelar la dependencia espacial de los cuantiles. Sin embargo, se llega a que los dos enfoques que afrontan este problema son computacionalmente muy demandantes y aún no tienen suficiente validación teórica.

6. Discusión

Los puntos clave de esta parte son:

- La regresión cuantílica bayesiana es aún un campo en desarrollo, falta mucho por explorar y este capítulo es una visión incompleta del mismo.
- El framework AL es computacionalmente más simple y más fácil de integrar en un marco bayesiano. Es necesario ajustar la varianza del posterior para hacer las predicciones útiles.
- Los modelos de mezcla de varianzas son más flexibles, pero siguen limitados a estimar un solo cuantil a la vez.
- Los modelos de cuantiles múltiples trabajan automáticamente con la heterocedasticidad.
- La verosimilitud bayesiana empírica es especialmente prometedora en el caso de cuantiles múltiples, pero hay mucho trabajo pendiente. Una ventaja es que no necesita modelar la distribución del error.