Estimation simultanée de quantiles de régression via l'approche Bayésienne

Josephine Merhi Bleik [1] & Ghislaine Gayraud [1]

[1] *Sorbonne universités, Université de Technologie de Compiègne, Laboratoire de Mathématiques Appliquées de Compiègne, EA 2222, Département de Génie Informatique, Centre de Recherches de Royallieu, 60 203 Compiègne cedex, France*
email : josephine.merhi-bleik@utc.fr, ggayraud@utc.fr

Résumé. On s'intéresse à estimer simultanément plusieurs quantiles dans un contexte de régression via l'approche Bayésienne. En supposant le terme d'erreur distribué selon la distribution asymétrique de Laplace (ALD) et en utilisant une relation qui lie deux quantiles distincts d'une ALD, on propose une nouvelle approche qui est simple, *full*-Bayésienne et qui satisfait à la propriété de non-croisement des quantiles. Pour évaluer la performance de notre méthode, nous utilisons une méthode de Monte-Carlo par chaîne de Markov (MCMC) pour simuler dans la loi a posteriori qui n'admet pas de forme analytique explicite.

Mots-clés. Régression quantile, Approche Bayésienne, méthodes MCMC.

Abstract. We are interested in estimating several quantiles simultaneously in a regression context via the Bayesian approach. Assuming that the error term has an asymetric Laplace distribution ALD and using a relation that links two distinct quantiles of the ALD distribution, we propose a simple fully Bayesian method that satisfies the property of non-crossing quantiles. To evaluate the perfomance of our method, we use Monte Carlo Markov Chain methods to simulate in the full conditional distributions since they do not admit a closed form representation.

Keywords. Quantile regression, Bayesian approach, MCMC methods.

1 Introduction

On considère le modèle de régression quantile suivant,

$$Y = q_p(X) + \epsilon, \tag{1}$$

en supposant que l'erreur ϵ est telle que $P(\epsilon < 0|X = x) = p$ pour tout $x \in \mathbb{R}$ si bien que q_p correspond à la fonction quantile d'ordre p de la loi conditionnelle $Y|X$. À partir des observations issues du modèle (1), on s'intéresse à estimer simultanément plusieurs quantiles de la loi $Y|X$ via l'approche Bayésienne.

La régression quantile permet une description plus riche que la régression classique, de part l'impact des variables explicatives X sur une variable réponse Y puisqu'elle fournit une description complète de la distribution conditionnelle de Y sachant X. En outre, elle est réputée plus adaptée dans le cas de certains types de données comme les données censurées ou bien en présence de valeurs aberrantes. En fréquentiste, Koenker & Basset (1978) propose d'estimer dans le cas d'un quantile linéaire, $q(x) = \beta_0 + \beta_1 x$, avec β_0 et β_1 réels, au travers du critère de minimisation suivant

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \rho_p(Y_i - (\beta_0 + \beta_1 x_i))$$

puisque

$$(\beta_0, \beta_1) = \arg\min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} \rho_p(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)),$$

avec $\rho_p(u) = u(p - \mathbb{1}(u < 0))$ la fonction de perte, appelée "check function". L'analogue en Bayésien a été proposée par Yu & Moyeed (2001) et consiste à supposer la distribution des erreurs comme étant une distribution asymétrique de Laplace, $ALD(0, p, \sigma)$, de paramètres de localisation nul, d'asymétrie p et d'échelle σ dont la densité est :

$$f(\epsilon) = \frac{p(1-p)}{\sigma} \exp\{-\frac{\rho_p(\epsilon)}{\sigma}\}. \tag{2}$$

Ainsi, supposer $\epsilon \sim ALD(0, p, \sigma)$ dans le modèle (1), fournit une manière naturelle de résoudre le problème d'estimation Bayésien du quantile q_p d'ordre p.

Dans le modèle de régression quantile (1) et en supposant $\epsilon \sim ALD(0, p, \sigma)$, avec $\sigma > 0$ et p fixé dans $(0, 1)$, notre méthode généralise le problème d'inférence Bayésienne d'un seul quantile à celui de l'estimation Bayésienne simultanée de r quantiles distincts $q_{\tau_1}, \ldots, q_{\tau_r}$ d'ordre $\tau_1, \ldots, \tau_r$ respectivement.

Afin d'exprimer la vraisemblance au travers de $q_{\tau_1}, \ldots, q_{\tau_r}$, on a recourt à la relation entre deux quantiles d'ordre différents sous l'hypothèse de la distribution ALD proposée par Yu et Zang (2005), i.e.,

$$q_p(X) = q_{\tau_j}(X) - \sigma g(\tau_j, p), \; \forall j = 1, \ldots, r \tag{3}$$

$$\text{où } g(\tau_j, p) = \begin{cases} \frac{1}{1-p} \log(\frac{\tau_j}{p}) & si \quad 0 < \tau_j < p, \\ -\frac{1}{p} \log(\frac{1-\tau_j}{1-p}) & si \quad p < \tau_j < 1. \end{cases}$$

Il existe plusieurs articles qui traitent, en Bayésien, de l'estimation simultanée de plusieurs quantiles dans un contexte de régression quantile (Fen & al (2015), Tokdar & al (2011), Rodriguez & Fan (2016)) ; néanmoins, nous proposons une approche alternative qui est simple, *full*-Bayésienne et qui satisfait à la propriété de non-croisement des quantiles.

2 Méthodologie : estimation de quantiles linéaires

On se restreint au cas de deux quantiles linéaires d'ordre τ_1 et τ_2 respectivement ($\tau_1 \neq \tau_2$) i.e.,

$$q_{\tau_1}(X) = \beta^0_{\tau_1} + \beta^1_{\tau_1} X \text{ et } q_{\tau_2}(X) = \beta^0_{\tau_2} + \beta^1_{\tau_2} X,$$

et on note $\beta_{\tau_1} = (\beta^0_{\tau_1}, \beta^1_{\tau_1}) \in \mathbb{R}^2$ et $\beta_{\tau_2} = (\beta^0_{\tau_2}, \beta^1_{\tau_2}) \in \mathbb{R}^2$.

On considère le cas de X déterministe et on suppose disposer de n valeurs $x_1, \ldots, x_n$ de X à partir desquelles on observe les réalisations de $Y_1, \ldots, Y_n$ indépendantes et toutes issues du modèle (1) en supposant que $\epsilon \sim ALD(0, p, \sigma)$. On fixe p à $1/2$ tandis que σ est inconnu.

On partitionne l'ensemble des indices en deux sous-ensemble I_1 et I_2,

$$I_1 = \{1, \ldots, \left[\tfrac{n}{2}\right]\},\ I_2 = \{\left[\tfrac{n}{2}\right] + 1, \ldots, n\}$$

et on note $Y_{I_j} = \{Y_i; i \in I_j\}$ et $X_{I_j} = \{X_i; i \in I_j\}$ pour $j = 1, 2$.

En s'appuyant sur l'indépendance des Y_i et sur l'équation (3), on propose de considérer le système des modèles indépendants suivant :

$$\begin{cases} Y_i = q_{\tau_1}(x_i) - \sigma g(\tau_1, p) + \epsilon_i, & i \in I_1, \\ Y_i = q_{\tau_2}(x_i) - \sigma g(\tau_2, p) + \epsilon_i, & i \in I_2. \end{cases} \tag{4}$$

Comme suggéré dans Kozumi et Kobayashi (2009), on passe à la représentation gaussienne de l'ALD puisque la loi a posteriori des quantiles n'est pas tractable analytiquement en raison de la présence de la fonction indicatrice dans la vraisemblance. On a ainsi

$$Y = q_p(X) + \gamma\omega + \delta\sqrt{\sigma\omega}Z, \tag{5}$$

où les variables ω et Z sont indépendantes de loi respective, ω de loi exponentielle de paramètre $1/\sigma$ et Z une loi normale centrée réduite ; les constantes γ et δ dépendent de p comme suit $\gamma = \frac{1-2p}{p(1-p)}$ et $\delta^2 = \frac{2}{p(1-p)}$.

En se référant à l'équation (5), le système (4) admet alors conditionnellement à ω une représentation gaussienne, d'où le système des modèles gaussiens :

$$\begin{cases} Y_i = q_{\tau_1}(x_i) - \sigma g(\tau_1, p) + \gamma\omega_i + \delta\sqrt{\sigma\omega_i}Z,\ i \in I_1 & (6) \\ Y_i = q_{\tau_2}(x_i) - \sigma g(\tau_2, p) + \gamma\omega_i + \delta\sqrt{\sigma\omega_i}Z,\ i \in I_2. & (7) \end{cases}$$

On propose alors de construire la vraisemblance de la manière suivante :

$$L(Y|X, \omega, \beta_{\tau_1}, \beta_{\tau_2}) = L_{I_1}(Y_{I_1}|X_{I_1}, \omega_{I_1}, \beta_{\tau_1}) \times L_{I_2}(Y_{I_2}|X_{I_2}, \omega_{I_2}, \beta_{\tau_2})$$

où $L_{I_1}(Y_{I_1}|X_{I_1}, \omega_{I_1}, \beta_{\tau_1})$ et $L_{I_2}(Y_{I_2}|X_{I_2}, \omega_{I_2}, \beta_{\tau_2})$ sont les vraisemblances associées aux modèles (6) et (7) respectivement.

Finalement, en faisant un choix classique pour les lois a priori sur β_{τ_1}, β_{τ_2} et σ, on propose de les échantilloner selon leur loi a posteriori via un algorithme de Gibbs qui contient une étape de Metropolis-Hastings sur σ.

3 Simulations

On simule 3000 réalisations de (Y, X) issues du modèle de régression linéaire suivant :

$$Y = 1 + 2X + \epsilon \tag{8}$$

avec $X \sim N(0, 1.5)$ et $\epsilon \sim st(df = 6)$. On s'intéresse à estimer les quantiles d'ordre 0.25 et 0.75 i.e., $q_{0.25} = 0.2824 + 2X$ et $q_{0.75} = 1.7175 + 2X$. On applique l'algorithme proposé de Gibbs avec une étape de Metropolis-Hastings avec 1500 itérations et une période de "burn-in" de 1500 itérations. Les lois a priori pour σ et β sont : $\sigma \sim IGamma(0.1, 0.01)$ et $\beta \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$.

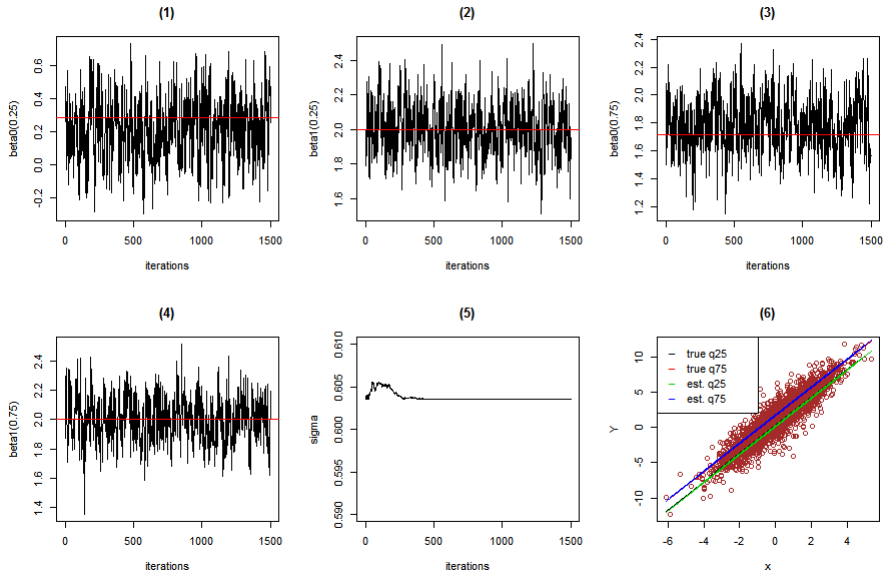

FIGURE 1 – Chaînes de Markov simulées de $\beta^0_{0.25}$ (1), $\beta^0_{0.75}$(2), $\beta^1_{0.25}$(3), $\beta^1_{0.75}$(4) et σ (5) et le tracé de la moyenne a posteriori des $q_{0.25}(X)$ et $q_{0.75}(X)$ (6).

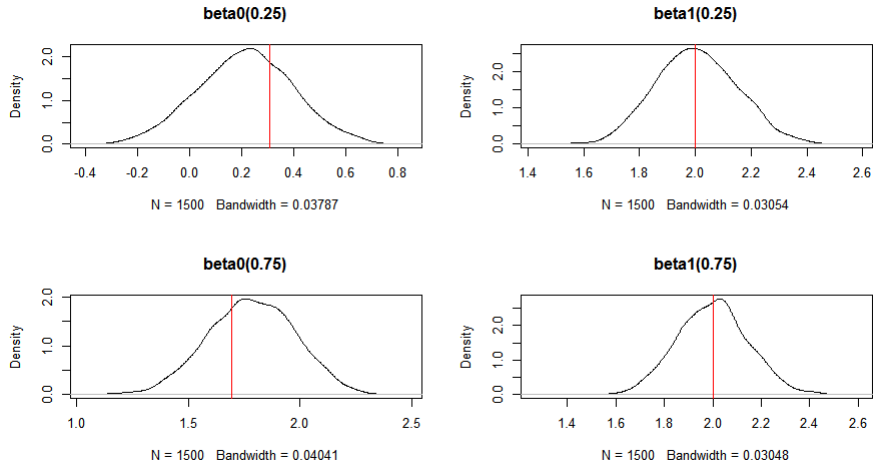

FIGURE 2 – Lois marginales a posteriori de $\beta^0_{0.25}$, $\beta^0_{0.75}$, $\beta^1_{0.25}$ et $\beta^1_{0.75}$ et vraies valeurs du paramètre en (ligne rouge verticale).

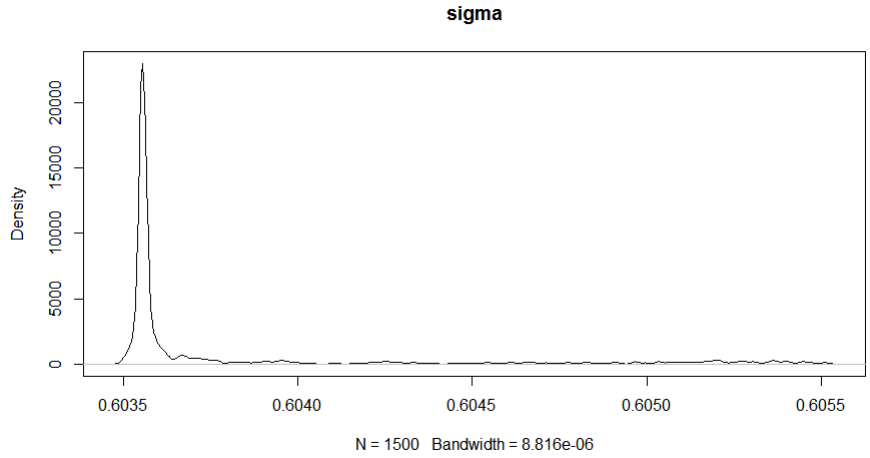

FIGURE 3 – Loi marginale a posteriori de σ.

Dans la Figure 1 ci-dessus, les cinq premiers graphes représentent les chaînes de Markov simulées par notre algorithme de Gibbs qui inclut une étape de Metropolis-Hastings. Le graphe (6) représente le tracé de la moyenne a posteriori des quantiles $q_{0.25}$ et $q_{0.75}$ (en bleu et vert) qui sont quasiment confondus avec les vraies fonctions de quantiles (en noir et rouge). Les Figures 2 et 3 représentent les densités a posteriori marginales des paramètres du modèle. On constate d'après la Figure 1 que notre algorithme converge et d'après la Figure 2 que les lois a posteriori marginales des paramètres convergent vers la vraie valeur des paramètres.

4 Conclusion

On a proposé une méthode d'estimation simultanée de plusieurs quantiles de régression via l'approche Bayésienne. Nos premiers résultats issus des simulations montrent que l'estimation simultanée est efficace quelque soit la distribution des données. En outre, cette méthode satisfait à la propriété de non-croisement de différents quantiles et peut aisément se généraliser au cas d'une estimation simultanée de r ($r > 2$) quantiles.

Bibliographie

[1] Feng, Y., Chen, Y., & He, X. (2015). Bayesian quantile regression with approximate likelihood. Bernoulli, 21(2), 832-850.
[2] Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. Econometrica : journal of the Econometric Society, 33-50.
[3] Kozumi, H., & Kobayashi, G. (2011). Gibbs sampling methods for Bayesian quantile regression. Journal of statistical computation and simulation, 81(11), 1565-1578.
[4] Rodrigues, T., & Fan, Y. (2016). Regression Adjustment for Noncrossing Bayesian Quantile Regression. Journal of Computational and Graphical Statistics, (just-accepted), 00-00.
[5] Tokdar, S. T., & Kadane, J. B. (2011). Simultaneous linear quantile regression : A semiparametric bayesian approach. Bayesian Analysis, 6(4), 1-22.
[6] Yu, K., & Moyeed, R. A. (2001). Bayesian quantile regression. Statistics & Probability Letters, 54(4), 437-447.
[7] Yu, K., & Zhang, J. (2005). A three-parameter asymmetric Laplace distribution and its extension. Communications in Statistics—Theory and Methods, 34(9-10), 1867-1879.