Bayesian Analysis of Censored and Binary Quantile Regression with Endogeneity

Genya Kobayashi

Second Bayesian Young Statisticians Meeting (BAYSM 2014)
Vienna, September 18–19, 2014

Chiba University, Faculty of Law, Plitics, & Economics, Chiba, Japan
gkobayashi@chiba-u.jp

Abstract

This paper considers Bayesian inference for the p-th quantile regression models for censored and binary response variables with an endogenous variable. We introduce a regression model of the endogenous regressor on the exogenous variables such that thte α-th quantile of the error term is zero. Then the residual of this regression model is included in the p-th quantile regression model as control function such that the corrected error term has zero p-th quantile. The resulting hierarchical model is estimated using the Gibbs sampler. Since the choice of α has an impact on the estimates, we also estimate it along with other parameters. The proposed models are demonstrated using the simulated and real datasets.

Keywords: asymmetric Laplace distribution; control function; Gibbs sampler; hierarchical model;

1 Introduction

Suppose that the response variables $y_i,\ i = 1, \dots, n$ are generated according to either $y_i = \{0, y_i^*\}$ when the observations are censored or $y_i = I(y_i^* > 0)$ when the binary responses are observed. Then consider the quantile regression model given by

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta}_p + \delta_p d_i + \epsilon_i,$$

where $\mathbf{x}_i$ is the vector of exogenous variables, d_i is the endogenous variable, $\boldsymbol{\beta}_p$ and δ_p are the coefficient parameters, and ϵ_i is the error term. When the covariate is endogenous, the standard censored and binary quantile regression models ([4], [1]) would produce biased estimates. Inference for quantile regression models for limited response variables with endogenous variables has been considered to be challenging (*e.g.*, [2]).

2 Approach

Following [3], we introduce the control function that corrects the error term such that its p-th quantile is equal to zero:

$$\begin{aligned} y_i^* &= \mathbf{x}_i'\boldsymbol{\beta}_p + \delta_p d_i + \eta_p(d_i - \mathbf{z}_i'\boldsymbol{\gamma}_\alpha) + e_i, & (1) \\ d_i &= \mathbf{z}_i'\boldsymbol{\gamma}_\alpha + u_i, & (2) \end{aligned}$$

where e_i and u_i are the error terms, $\mathbf{z}_i = (w_i, \mathbf{x}_i')'$, w_i is an exogenous variable, and $\boldsymbol{\gamma}_\alpha$ is the coefficient parameter. It is assumed that $u = d - \mathbf{z}'\boldsymbol{\gamma}_\alpha$ is satisfied and quantile independence of e on $\mathbf{z}$ conditional on u:

$$Q_p(\epsilon|d, \mathbf{z}) = Q_p(\epsilon|u, \mathbf{z}) = Q_p(\epsilon|u) = \eta_p(d - \mathbf{z}'\boldsymbol{\gamma}_\alpha) \quad (3)$$

where $Q_\tau(\cdot|\cdot)$ denotes the τ-th conditional quantile. Also, the α-th quantile of u is assumed to be zero for some $\alpha \in (0, 1)$.

$$Q_\alpha(u|\mathbf{z}) = 0. \quad (4)$$

The value of α is not known and its choice is arbitrary. In [3], it is mentioned that the zero mean restriction $E[u|\mathbf{z}] = 0$ can be used instead. However, since the choice of α is a part of model specification and can have an impact on the conclusion on endogeneity of d, we treat it as a parameter. Assuming that e and u independently follow asymmetric Laplace distributions respectively with shape parameters p and α, the parameters are estimated using the Gibbs sampler.

3 Simulation

We generated $n = 500$ observations from

$$\begin{aligned} y_i^* &= 0 + 2x_{1i} + 2x_{2i} + 2x_{3i} + d_i + v_i, \\ d_i &= 1 + w_i + x_{1i} + x_{2i} + x_{3i} + u_i, \end{aligned}$$

where

$$\begin{aligned} x_{1i}, x_{2i}, x_{3i}, w_i &\sim N(0, 1), \\ v_i &\sim N(u_i, 1), \\ u_i &\sim \exp(0.5w_i)\chi_2^2. \end{aligned}$$

We replicated the data 100 times. For the censored data, we consider the standard Bayesian censored quantile regression model (CQR), the proposed model with restriction (4) (IVCQR_A), the model with (4) but $\alpha = p$ (IVCQR_P), and the model with the zero mean restriction (IVCQR_M). We denote the models for the binary data accordingly by BQR, IVBQR_A, IVBQR_P, and IVBQR_M. For the binary data, the scale of the coefficient vector $\tilde{\boldsymbol{\beta}}_p = (\boldsymbol{\beta}_p', \delta_p)$ is normalised such that the coefficient for x_{1i} equal to 2.

Figure 1 show the bias and RMSE for the posterior means of $\tilde{\boldsymbol{\beta}}_p$ for $p = 0.25$, 0.5, 0.75 for the case of the censored data. As expected, CQR produced the largest bias and RMSE. In contrast, the bias and RMSE for the proposed IVCQR is the smallest. Since the distribution of u is highly skewed to the right (the average of the posterior mean of α is 0.065), (4) is not satisfied for IVCQR_P and IVCQR_M and the resulting performance for them was not as good as that for IVCQR_A. Fig 2 shows a similar result for the case of the binary response data.

4 Conclusion

The proposed models are useful approach to quantile regression analysis with endogeonous variables. It would be interesting to extend the present model to incorporate endogenous binary variables as in the context of endogenous switching.

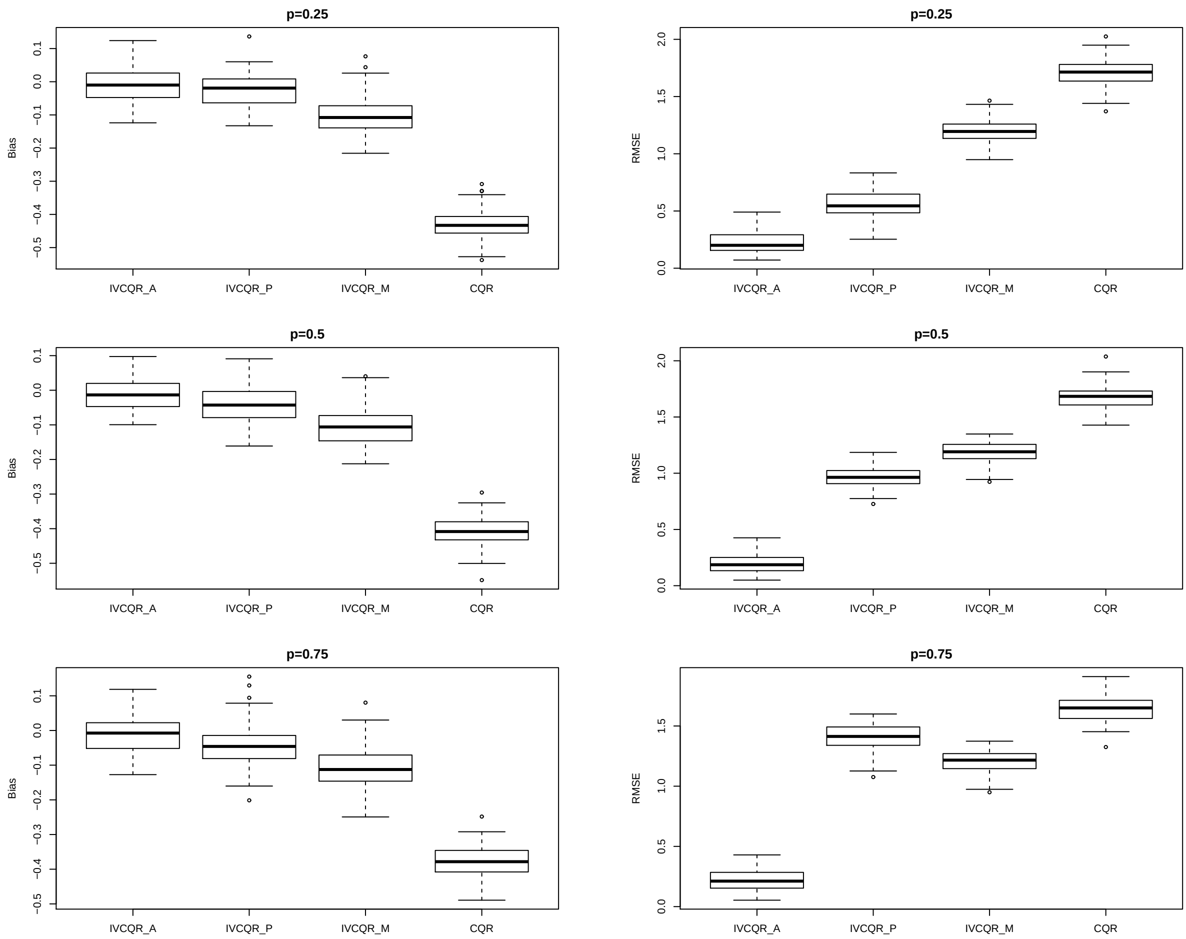

Figure 1: Bias and RMSE for censored data

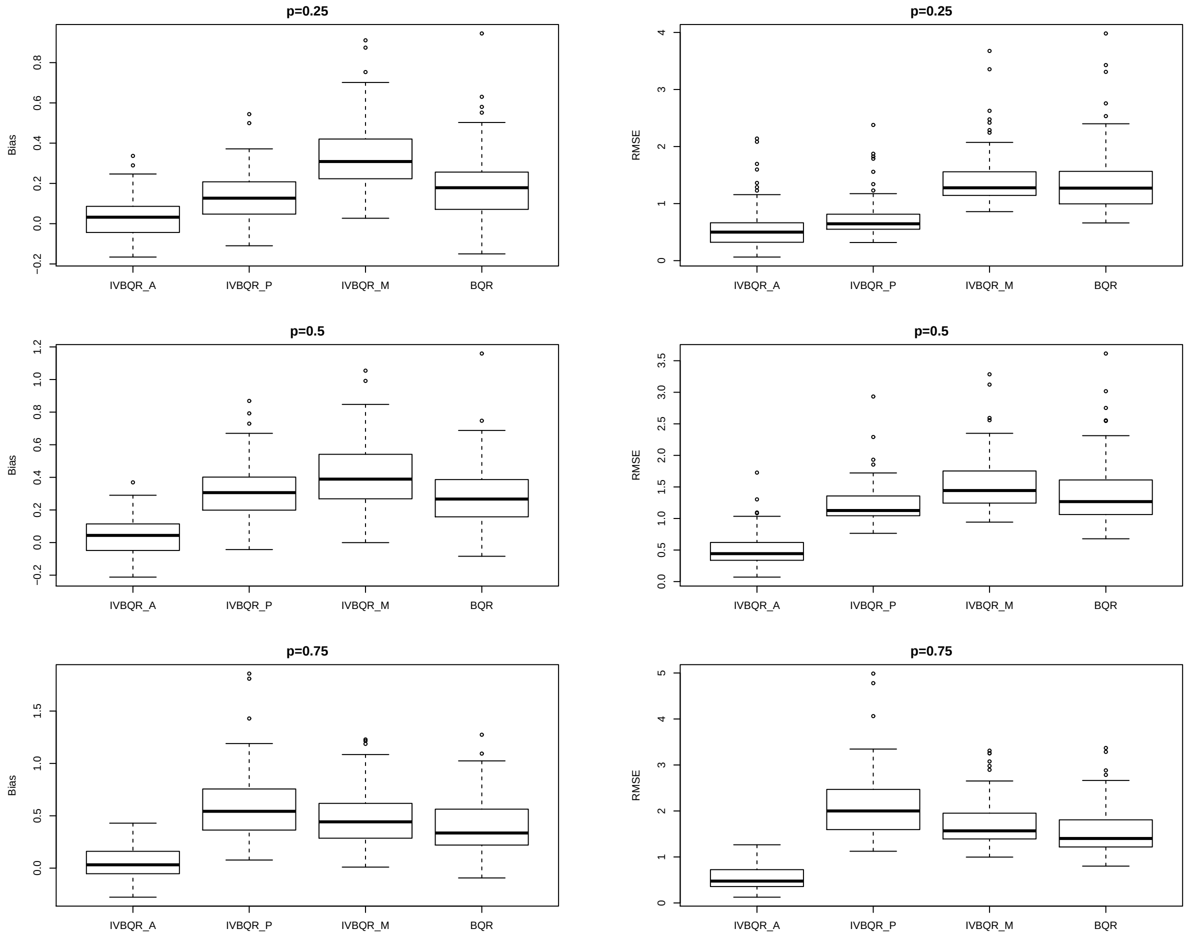

Figure 2: Bias and RMSE for binary data

References

[1] Benoit, D. F. and Van den Poel, D. (2012). "Binary quantile regression: a Bayesian approach based on the asymmetric Laplace distribution," *Journal of Applied Econometrics*, **27**, 1174-1188.

[2] Blundell, R. and Powell, J. L. (2007). "Censored regression quantiles with endogenous regressors," *Journal of Econometrics*, **141**, 65-83.

[3] Lee, S. (2007). "Endogeneity in quantile regression models: a control function approach," *Journal of Econometrics*, **141**, 1131-1158.

[4] Yu, K. and Stander, J. (2007). "Bayesian analysis of a tobit quantile regression model," *Journal of Econometrics*, **137**, 260-276.