Bayesian Quantile Regression

Tyler Bagwell, Meredith Kruse, Kevin McCoy

Department of Statistics, Rice University

November 21st, 2023

Table of Contents

Section 1

Standard Quantile Regression

What is Quantile Regression?

As opposed least squares regression which estimates conditional mean functions, quantile regression focuses on estimating families of conditional quantile functions.

$$\mathbb{E}[y_i|\boldsymbol{x_i}] = \boldsymbol{x_i^T}\beta \qquad \text{traditional regression}$$

$$q_\tau(y_i|\boldsymbol{x_i}) = \boldsymbol{x_i^T}\beta(\tau) \qquad \text{quantile regression}$$

Notation

- Standard Linear Model: $y_t = \mu(\boldsymbol{x_t}) + \epsilon_t$
  - $\mu(x_t) = \boldsymbol{x_t'\beta}$
- The τth regression quantile is defined as any solution, $\hat{\beta}(\tau)$,to the quantile regression minimisation problem:

$$\min_{\beta} \sum_{t} \rho_\tau(y_t - \mathbf{x_t'}\beta)$$

- Loss Function: $\rho_\tau(u) = u(\tau - I(u < 0))$

Asymmetric Laplace Distribution

- U has an asymmetric Laplace distribution if its probability density is given by:

$$f_\tau(u) = \tau(1-\tau) exp\{-\rho_\tau(u)\}$$

  where $0 < \tau < 1$.
- (Note that location and scale parameters can be incorporated into this density)

Bayesian Quantile Regression

- In the Bayesian framework we are interested in the conditional quantile, $q_\tau(y_i|\mathbf{x_i})$, rather than the conditional expectation.
- Given the observations $\mathbf{y} = (y_1, ..., y_n)$, the posterior distribution of $\boldsymbol{\beta}$ is given by:

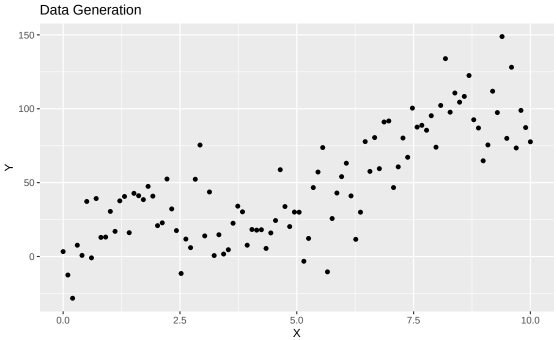
  $$\pi(\boldsymbol{\beta}|\boldsymbol{y}) \propto L(\boldsymbol{y}|\boldsymbol{\beta} p(\boldsymbol{\beta}))$$

  where $p(\boldsymbol{\beta})$ is the prior distribution of $\boldsymbol{\beta}$ and the likelihood is:

  $$L(\boldsymbol{y}|\boldsymbol{\beta}) = p^n(1-p)^n exp\{-\sum_i \rho_\tau(y_i - \mathbf{x_i'}\beta)\}$$

  using the location parameter $\mu_i = \mathbf{x_i'}\beta$
- Note: In the absence of a realistic prior we can use an imporoper uniform prior for all components of β

Data Generation - Normal error

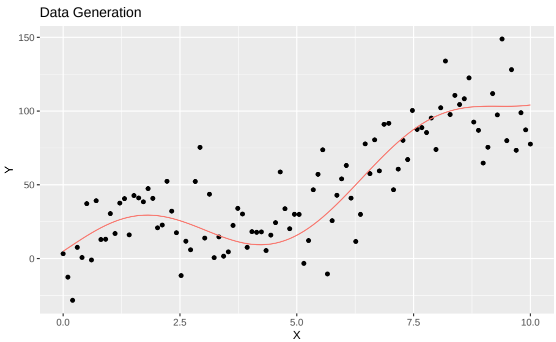


The data was generated via

$$y_i = 5 + x + x^2 + 20\sin(x) + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 20^2) \implies SNR \approx 2.42$$

Data Generation - Normal Error

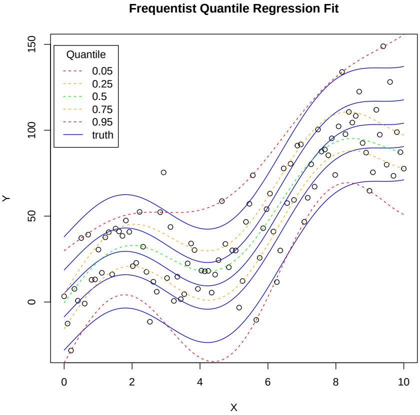


The data was generated via

$$y_i = 5 + x + x^2 + 20\sin(x) + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, 20^2) \implies SNR \approx 3.63$$

Bayesian vs Frequentist - Normal error

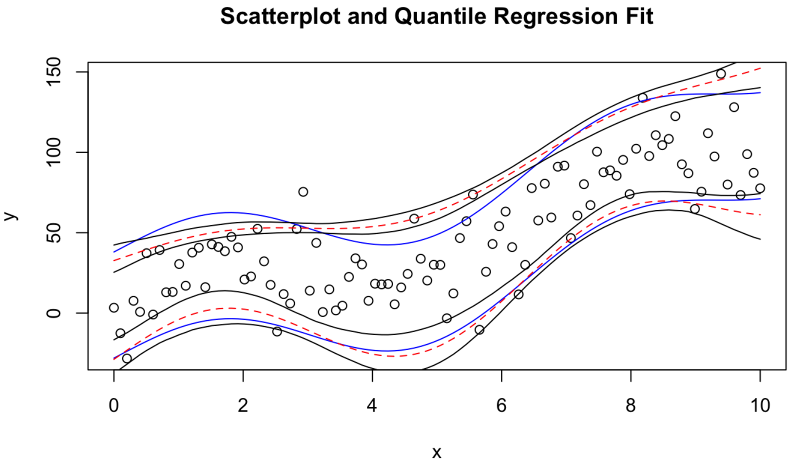


(a) (b)

Figure: (a) Frequentist QR. (b) Bayesian QR.

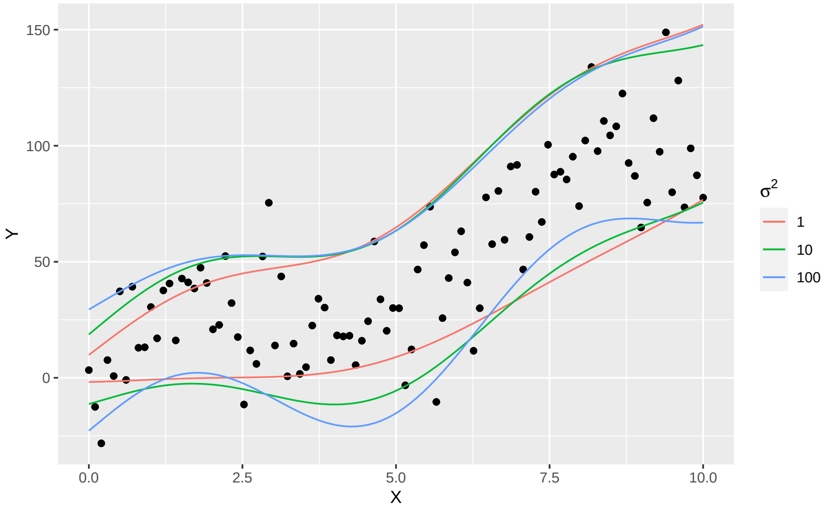
Quantile Credible Intervals - Normal error

Figure: 0.05, 0.95 Quantiles and their 95% Credible Interval



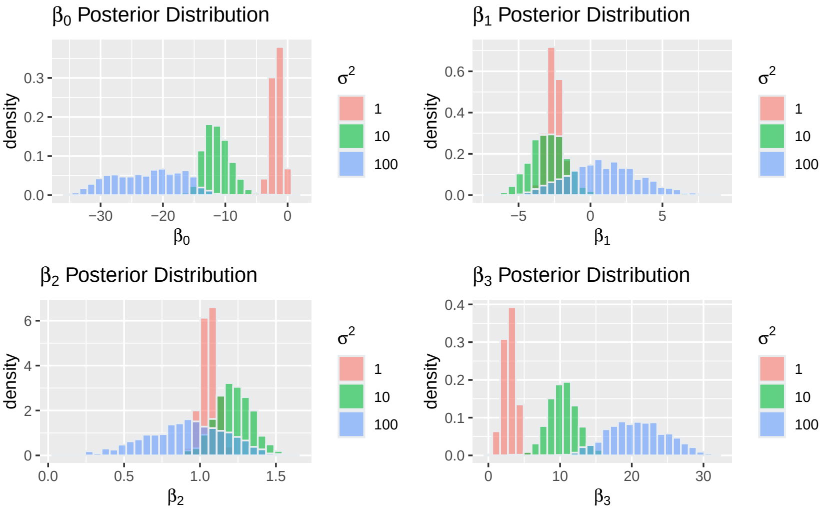
Sensitivity Analysis

Figure: 0.05, 0.95 Quantile Regression Lines



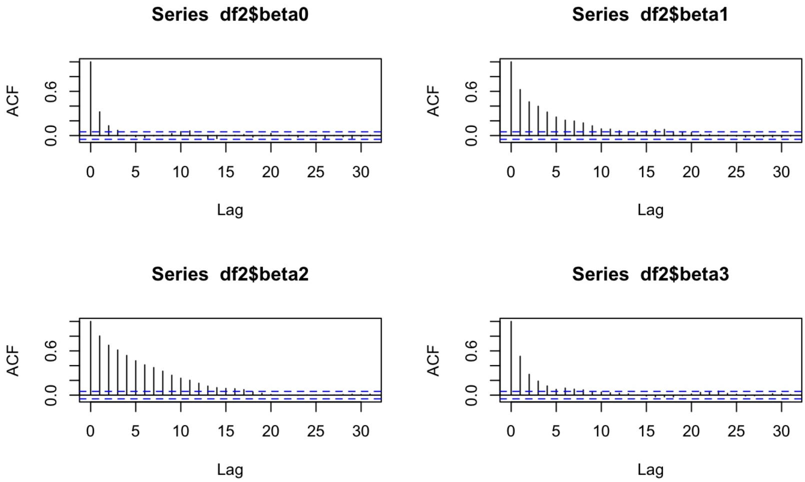
Sensitivity Analysis

Figure: 95% Beta Posteriors

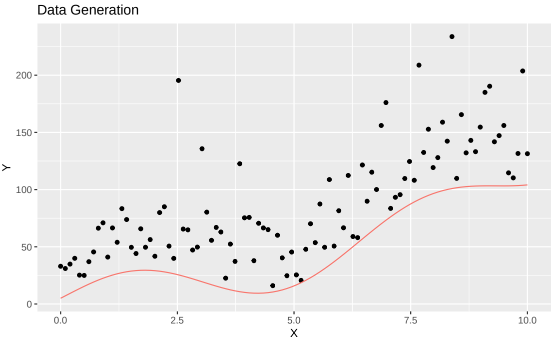


MCMC Diagnostics

Figure: Autocorrelation Plots



Data Generation - χ^2 error

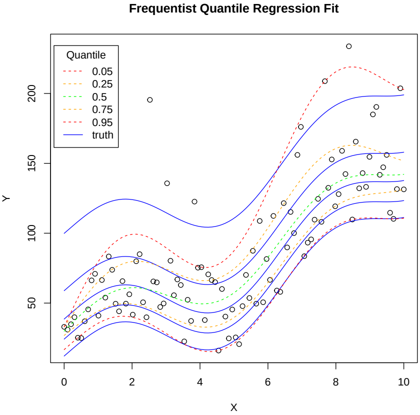


The data was generated via

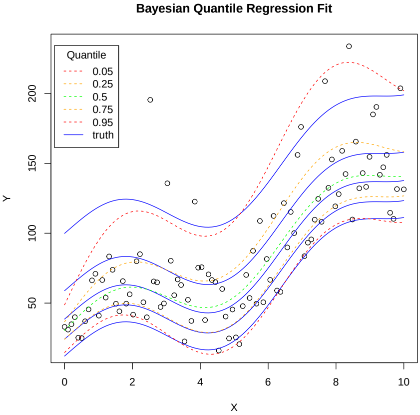
$$y_i = 5 + x + x^2 + 20\sin(x) + 10 * \epsilon$$

$$\epsilon \sim \chi^2_4 \implies SNR \approx 3.06$$

Bayesian vs Frequentist - χ^2 error



(a)

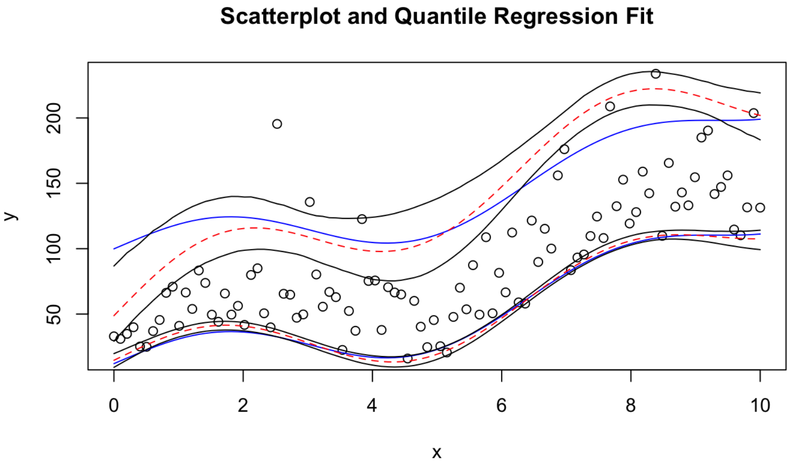


(b)

Figure: (a) Frequentist QR. (b) Bayesian QR.

Quantile Credible Intervals - χ^2 error

Figure: 0.05, 0.95 Quantiles and their 95% Credible Interval



Subsection 3

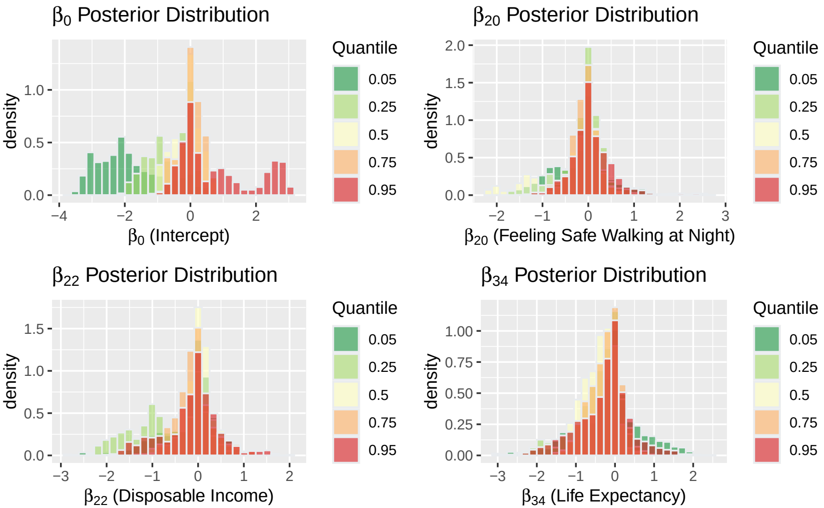
Real Data Example

Figure: Data Composition

	BMI[, 1]	Age group	Labour market insecurity[, 1]	Stakeholder engagement for developing regulations[, 1]	Dwellings without basic facilit
1	-1.147033255	18-19	-0.43110631	1.141138822	
2	-1.904771749	18-19	-0.62165053	-1.508899200	-0
3	-1.509479344	18-19	-0.59783251	-0.183880189	-0
4	-1.471243637	18-19	NA	0.194696671	0
5	-1.276634097	18-19	-0.26438011	1.519715682	-0
6	-0.923220259	18-19	0.49779679	-1.508899200	0
7	-1.560517393	18-19	NA	-1.319610770	1
8	-1.105922375	18-19	NA	-0.562457050	-0
9	-2.146164838	18-19	-0.62165053	-0.941033910	-0
10	-1.786361680	18-19	-0.09765392	-0.183880189	-0
11	-1.816586709	18-19	0.11670834	1.141138822	0
12	-1.959147603	18-19	-0.64546856	0.194696671	-0
13	-2.160005883	18-19	-0.43110631	0.005408241	-0
14	-1.610007698	18-19	-0.83601279	-0.562457050	-0
15	-1.767856834	18-19	3.99904693	-0.562457050	-0
16	-1.808632729	18-19	-0.26438011	-1.698187631	-0
17	-1.519427236	18-19	-0.93128490	0.005408241	-0
18	-1.213053798	18-19	-0.55019645	-1.508899200	-0
19	-1.465434868	18-19	-0.07383589	0.762561962	
20	-2.121626478	18-19	0.87888524	0.762561962	-0
21	-2.478171094	18-19	-0.52637842	-1.319610770	0
22	-2.293674383	18-19	0.33107059	0.194696671	0
23	-2.174829783	18-19	NA	0.573273531	0
24	-1.898703462	18-19	-0.64546856	-0.751745480	-0
25	-1.010575245	18-19	-0.21674406	2.087580973	2

World Development Factors and BMI

Figure: Beta Posteriors Over Various Quantiles



Section 2

Bayesian Quantile Regression for Longitudinal Data

What is Longitudinal Data?

- Data characterized by repeated measurements on the same subject over time $\rightarrow$ introduces within-subject dependence/correlation
  - For example: clinical trials, epidemiological studies, etc.
- **Q: How to handle?**
  - **A**: Adding **random effects** to your model can account for over dispersion caused by unobserved heterogeneity or for correlation in longitudinal data.
- **Next**: Extend the ordinary Bayesian quantile regression model from above by introducing two **linear mixed-effects** Bayesian quantile regression models based on work by Luo et al.

Hierarchical Bayesian Quantile Regression Model

- Linear mixed-effects quantile function of the response:

$$Q_{y_{ij}}(\boldsymbol{\tau}|\boldsymbol{x_{ij}}, \boldsymbol{\alpha_i}) = \boldsymbol{x_{ij}^T}\boldsymbol{\beta} + \boldsymbol{z_{ij}^T}\boldsymbol{\alpha_i},$$

  where $Q_{y_{ij}}(\cdot) \equiv F_{y_i}^{-1}(\cdot)$ is the inverse of the cumulative distribution function of y_{ij} conditional on the covariates.
- y_{ij}: response of ith individual measured at time j.
- General hierarchical model

$$y_{ij} \sim ALD(\boldsymbol{x_{ij}^T}\boldsymbol{\beta} + \boldsymbol{z_{ij}^T}\boldsymbol{\alpha_i}, \sigma, \tau)$$

$$\boldsymbol{\beta} \sim \pi(\boldsymbol{\beta})$$

$$\boldsymbol{\alpha}_i|\boldsymbol{\Sigma} \sim f(\boldsymbol{\alpha}_i|\boldsymbol{\Sigma})$$

$$\boldsymbol{\Sigma} \sim \pi(\boldsymbol{\Sigma})$$

$$\sigma \sim \pi(\sigma)$$

Two Proposed Models

Need to compute: $\pi(\boldsymbol{\beta}, \sigma, \boldsymbol{\Sigma}, \boldsymbol{\alpha}|\boldsymbol{y}) \propto f(\boldsymbol{y}, \boldsymbol{\alpha}|\boldsymbol{\beta}, \sigma, \boldsymbol{\Sigma})\pi(\boldsymbol{\beta})\pi(\sigma)\pi(\boldsymbol{\Sigma})$, but the posterior densities of both the fixed and random effect parameters are very complicated.

Luo et al. propose two models:

1. **BQRMH**: ME Metropolis-Hastings Bayesian Quantile Regression
   - Draw posterior samples of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ via MH, all other parameters can be sampled from well known distributions.
   - **Con**: Must tune parameters of the proposal distributions to attain decent acceptance rates. Different for every value of τ!
2. **BQRGS**: ME Bayesian Quantile Regression Gibbs Sampler.
   - Transforms the likelihood of $\boldsymbol{y}$ from ALD into a MVN with clever manipulations $\rightarrow$ posteriors are easier to derive with conjugate priors.
   - **Plus**: Do not need to deal with any tuning parameters!

Next, we coded up the two algorithms.

Model Comparisons via Simulated Mixed-effects Data

We compare the performance of three Bayesian QR models via a simulated simple linear mixed-effects model.

1. **OBQR**: Ordinary Bayesian Quantile Regression
2. **BQRMH**: ME Metropolis-Hastings Bayesian Quantile Regression
3. **BQRGS**: ME Bayesian Quantile Regression Gibbs Sampler.

Use the following simple linear mixed-effects model to generate the data:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_{i0} + \alpha_{i1} x_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, 20, \; j = 1, \ldots, 5,$$

where $x_{ij} \sim \mathrm{U}(0,1)$, $\boldsymbol{\beta} = (\beta_0, \beta_1)^T = (1,5)^T$, $\boldsymbol{\alpha}_i = (\alpha_0, \alpha_1)^T \sim \mathrm{N}_2(\mathbf{0}, \mathbb{I}_2)$, and $\epsilon_{ij} \sim \mathrm{N}(0,1)$. Assume weak prior information on $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} \sim \mathrm{N}_2(\mathbf{0}, 100\mathbb{I}_2).$$

We investigate five different quantiles $\tau = (0.10, 0.25, 0.50, 0.75, 0.90)$.

A Simulated Dataset

example_dataset.png

Model Diagnostics

For a single simulated dataset with 5000 sample draws for $\tau = 0.50$.

trace_ACF.png

Three Algorithms and Their Parameter Estimation

- Obtained via 100 simulated datasets with 5000 sample draws (500 burn-in) for $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$ for each model.

bias_MSE.png

Longitudinal Studies with Missing Data

A final (third) model: **penalized model** with applications to longitudinal studies with missing data. Want to minimize:

$$\sum_{i=1}^{n} \sum_{j \in J_{obs}} \rho_\tau \left(y_{ij} - \boldsymbol{x}_{ij}^T \boldsymbol{\beta} - \boldsymbol{z}_{ij}^T \boldsymbol{\alpha}_i \right) + \frac{1}{2} \sum_{i=1}^{n} \boldsymbol{\alpha}_i^T \Lambda^{-1} \boldsymbol{\alpha}_i,$$

where Λ is a symmetric matrix $\rightarrow$ can be recast into random-effects model.

- Link missing data process with the longitudinal outcome process assuming they share the same random effects $\boldsymbol{\alpha}_i$.
- Adds a new dimension to the likelihood: need to account for likelihood of observing or no observing an observation.
- Need to compute transition probabilities $\pi_{ij}^{(\mathcal{O})}$ and $\pi_{ij}^{(\mathcal{M})}$.

References

Y. Yuan and G. Yin, "Bayesian Quantile Regression for Longitudinal Studies with Nonignorable Missing Data," *Biometrics*, vol. 66, pp. 105–114, Mar. 2010.

K. Yu and R. A. Moyeed, "Bayesian quantile regression," 2001.

Y. Luo, H. Lian, and M. Tian, "Bayesian quantile regression for longitudinal data models," *Journal of Statistical Computation and Simulation*, 2012.

M. Fleurbaey, "Beyond gdp: The quest for a measure of social welfare," *Journal of Economic literature*, vol. 47, no. 4, pp. 1029–1075, 2009.

J. Bentham, M. Di Cesare, V. BIlano, L. M. Boddy, *et al.*, "Worldwide trends in children's and adolescents' body mass index, underweight and obesity, in comparison with adults, from 1975 to 2016: a pooled analysis of 2,416 population-based measurement studies with 128.9 million participants," *Lancet*, 2017.