Quantile Regression for Longitudinal Data

Maria Laura Battagliola*
Department of Mathematical Sciences, University of Copenhagen

Project Outline

Scalar-on-function observations collected in longitudinal studies have become of major interest in recent years. They represent a powerful modeling tool in many fields, from medicine to life science.
The final aim of the research is to endow this framework with *quantile regression models*, since they would allow for a much more flexible and robust analysis than mean regression.

Future Application

As long as intensive animal farming is concerned, one of the major interests is to assure that the environmental conditions do not negatively affect the breeding capabilities of the livestock.

Consider the daily food intake as a measure of health of lactating sows and the stable temperature and humidity profiles as representatives of their living conditions [3].

Our objective is to focus on the feed intake during the very first days after giving birth (i.e. on *low quantiles* of the response conditional distribution), since insufficient nutrition in this period may affect sows' reproductive system.

Quantile Regression

In several contexts, when we are interested in finding a relationship between response $y \in \mathbb{R}^n$ and predictor $X \in \mathbb{R}^{n\times p}$, linear regression might not be the most suited tool, as it only estimates the conditional mean. To have a deeper insight in the phenomenon considered, it can be useful to examine several points of the conditional distribution instead. *Quantile regression allows to estimate the conditional quantiles of y given X.*

Consider the conditional quantile function

$$Q_{y|X}(\tau) = X'\beta_\tau$$

where $\tau \in (0,1)$ is the quantile level and the coefficient $\beta_\tau \in \mathbb{R}^p$ is unknown. Then, we can get the estimate

$$\hat{\beta}_\tau = \arg\min_{\beta\in\mathbb{R}} \sum_{i=1}^{N} \rho_\tau(y_i - X_i'\beta_\tau)$$

with $\rho_\tau(v) = v(\tau - \mathbb{I}_{\{v<0\}})$ loss function called check function.

*Contact

mlbattagliola@math.ku.dk

Quantile Regression for Longitudinal Data

At this stage of the project, the aim is to apply quantile regression to *scalar longitudinal data* $\{y_{ij}, Z_{ij}, X_{ij}\}_{i=1:N}^{j=1:n_i}$, with N subjects and n_i repeated measurements for each subject i over a time domain Ω. The conditional quantile function is then

$$Q_{y_{ij}|u_i}(\tau) = X_{ij}'\beta_\tau + Z_{ij}u_i$$

where β_τ are the fixed effects and u_i the subject specific random effects.

We compare two estimation approaches:

1. From [1], minimization of criterion

$$P(\beta_\tau, u) = \sum_{i=1}^{N}\sum_{j=1}^{n_i} \rho_\tau(y_{ij} - X_{ij}'\beta_\tau - Z_{ij}u_i) + \lambda \sum_{i=1}^{N} |u_i|$$

2. From [2], maximization of the marginal distribution of y_{ij} assuming that

$$y_{ij}|u_i \overset{ind}{\sim} ALD(X_{ij}'\beta_\tau + Z_i u_i, \sigma, \tau) \qquad i \in \{1,...,N\}\ j \in \{1,...,n_i\}$$
$$u_i \overset{ind}{\sim} N(0, \varphi^2) \qquad i \in \{1,...,N\}$$

Simulations

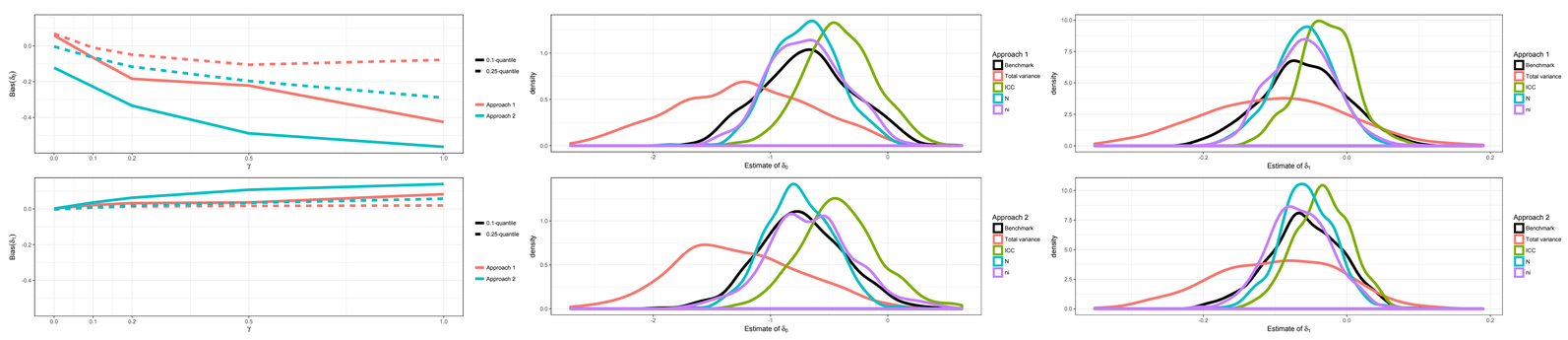
We consider this *shift-scale effect model* from [1]:

$$X_{ij} = v_i + w_{ij}, \quad v_i \overset{iid}{\sim} \chi^2_3 \quad w_{ij} \overset{iid}{\sim} \chi^2_3$$
$$y_{ij} = u_i + \beta X_{ij} + (1 + \gamma X_{ij})e_{ij}$$
$$u_i \overset{iid}{\sim} N(0, \sigma_u^2) \quad e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$$

with consequent conditional quantile function

$$Q_{y_{ij}|u_i}(\tau) = u_i + \Phi^{-1}(\tau) + (\beta + \gamma\Phi^{-1}(\tau))X_{ij}$$

Take $\beta = 0$ and analyse the estimates of $\delta_0 = \Phi^{-1}(\tau)$ and $\delta_1 = \beta + \gamma\Phi^{-1}(\tau)$.
We show how the results are affected increasing one single parameter at the time from the benchmark setting ($\tau = 0.25$, $\gamma = 0.1$, $\sigma_u^2 + \sigma_e^2 = 2$, $\frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2} = 0.5$, $N = 50$, $n_i = 5$).



Approach 1 is implemented in R package `hqreg` and Approach 2 in `lqmm`. The latter is computationally faster of a 10^2 order of magnitude than the former.

Conclusions and Further Developments

- Bias $\uparrow$ as τ tends to extreme values;
- Bias $\downarrow$ as $\frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2} \uparrow$;
- Bias does not vary as either $N \uparrow$ or $n_i \uparrow$.

- Exploit maximization of marginal likelihood with more general distributional hypothesis than Approach 2;
- Tackle model misspecification when $X_{ij} \in (-\infty, \infty)$;
- $X_{ij} \to X_{ij}(s)$, s functional coordinate.

References

[1] R. Koenker Quantile regression for longitudinal data. *Journal of Multivariate Statistics*, 91:74–89, 2004
[2] M. Geraci, M. Bottai Linear quantile mixed models. *Stat Comput*, 24:461–479, 2014
[3] S.Y. Park, C. Li, Z. Han, S.M. Mendoza, E. van Heugten, A.M. Staicu Conditional analysis for mixed covariates, with application to feed intake of lactating sows. *Biometrics*, 59:1–25, 2016