An introduction to quantile regression

Alessio Farcomeni
University of Rome "Tor Vergata"

alessio.farcomeni@uniroma2.it

Linear quantile regression

- Linear regression model: $E[Y|X] = \beta^T X$
- Quantile regression model:

$$Q_\tau(Y|X) = \beta_\tau^T X$$

Linear quantile regression

- No assumptions on a parametric distribution (distribution-free method), no homoscedasticity
- Assumptions on Y: continuity (no ties); it can be relaxed to model count data
- Assumptions on Y: independence and identical distribution; it can be relaxed to model panel data
- Median ($\tau = 0.5$) regression is an alternative to linear regression $E[Y|X] = \beta^T X$
- Other quantiles also of interest: a covariate (e.g., treatment) might have different effects in the tails (even contradictory, sometimes)

Optimisation for mean and quantile estimation

- It is well known that

$$\bar{y} = \arg\min_{\mu} \sum_i (y_i - \mu)^2$$

- You probably also know that

$$\hat{Q}_{0.5}(y) = \arg\min_{m} \sum_i |y_i - m|$$

- More in general, define

$$\rho_\tau(u) = u(\tau - I(u < 0)),$$

you have that

$$\hat{Q}_\tau(y) = \arg\min_{q} \sum_i \rho_\tau(y_i - q)$$

Estimation of quantile regression coefficients

- For least squares estimation you solve

$$\arg\min_b \sum_i (y_i - b^T x_i)^2$$

- Linear quantile regression simply defined as

$$\hat{\beta}_\tau = \arg\min_b \sum_i \rho_\tau(y_i - b^T x_i),$$

- Predictions

$$\hat{Q}_\tau(y|x) = \hat{\beta}_\tau^T x$$

Interpretation

- Intercept: the estimated quantile when the remaining covariates are zero
- Slopes: the variation in τ-quantile of Y for one unit increase of the covariate, holding the others fixed (-ish)

Interval estimation and testing

- Simplest way to obtain standard errors and confidence intervals: bootstrap
- For testing one can use Wald tests

$$\hat{\beta}_\tau / \sqrt{\hat{\mathrm{Var}}(\hat{\beta}_\tau)} | H_0 \sim N(0,1)$$

- There are several other methods (see `help(summary.rq)`)

Example on reaction times

- Case control study of $n = 60$ subjects using a cell phone or not
- Age: 16-24 vs > 24, gender
- Outcome: reaction time to an external event

Example on reaction times

	$E[\log(Y)\|X]$	$\hat{Q}_{.25}(y\|x)$	$\hat{Q}_{.5}(y\|x)$	$\hat{Q}_{.75}(y\|x)$
Intercept	**0.29**	**1.22**	**1.31**	**1.40**
age: 25+	**0.06**	**0.08**	**0.09**	**0.07**
Male	-0.01	-0.00	-0.01	-0.01
Cell use	**0.05**	**0.12**	**0.07**	0.05

Panel data

- Panel data arise when repeated measures are taken over time
- Instead of y_i, the outcome is y_{it}, $i = 1, \ldots, n$, $t = 1, \ldots, T$.

Why bother?

- The assumption that units are independent is not credible anymore
- If you have dependent units and ignore the issue, you get biased results

Basic solution

- The basic solution involves the modeling stage
- A conditional independence assumption is used: there exist subject-specific parameters α_i such that

$$y_{it} \perp\!\!\!\perp y_{ih} \mid \alpha_i(\tau)$$

- The linear model becomes

$$E[Y_{it}|x_{it}] = \alpha_i + \beta' x_{it}$$

- Consequently, the QR is

$$Q_\tau(y|x_{it}) = \alpha_i(\tau) + \beta'_\tau x_{it}$$

nothing else changes.
- If X_{it} includes a leading 1, for identifiability it is assumed that $\sum_i \alpha_i = 0$ or $\alpha_1 = 0$

FE (and almost that)

- If you wish to fit a fixed-effects panel QR model, that is straightforward.
- You can simply include `factor(ID)` (or the equivalent) in the `qr` statement
- There also exist an approach that is in-between FE and RE by Koenker (2004) Quantile regression for longitudinal data. *Journal of Multivariate Analysis*. **91**, 74-89; which we just mention.

Quantile Regression with Random Effects

- If you wish to fit a random-effects model, you need to specify a parametric distribution for the quantiles in order to obtain a likelihood.
- This is also useful in other contexts (e.g., Bayesian inference)
- We have used distribution-free objective functions so far.
- Common assumption:

$$f(y_{it}|\alpha_{it}, \beta, x_{it}, \eta) \sim \frac{\tau(1-\tau)}{\eta} \exp\left\{-\rho_\tau\left(\frac{y_{it}-\alpha_{it}-x_{it}'\beta}{\eta}\right)\right\}$$

- The density above is an asymmetric Laplace distribution.

- Observed likelihood

$$\prod_i \int_{\alpha_i} \prod_t f(y_{it}|\alpha_i, \beta, x_{it}, \eta) f(\alpha_i)\ d\alpha_i$$

- The integrals are usually approximated numerically (through Gaussian quadrature, for instance)
- Parameter estimates are obtained by numerical maximization of the likelihood
- In `R` you can use function `lqmm` in `library(lqmm)`

Marginal Mid-quantiles

What if there are ties?

- Mid-CDF: $G(y) = \Pr(Y \leq y) - 0.5 \Pr(Y = y)$
- Mid-quantiles: $H_\tau(\cdot)$ as generalised inverse of $G(y)$
- If Y is continuous, it reduces to commonly used marginal quantiles.
- Empirical versions are well behaved (sample mid-quantiles are Gaussian asymptotically).

Mid-quantile regression

- Approach described in Geraci and Farcomeni (2022)
- Conditional mid-CDF: $G(y|X) = \Pr(Y \leq y|X) - 0.5\Pr(Y = y|X)$
- We induce continuity through a linear interpolator operator, obtaining $G^C(y|X)$. This substitutes for jittering.
- mid-QR model:

$$H_\tau(h(Y)|X) = X^T\beta_\tau$$

The jittering approach

- Machado & Santos Silva (2005); Hong & He (2010): jittering, standard QR, averaging the noise out
- Let $U \in [0,1)$ and $Y \in \mathcal{N}$. Standard QR for continuous outcomes is applied to $Y + U$. Noise is added M times (default, $M = 50$), and $\hat{\beta}(p)$ is the average across replications.
- Implemented in `rq.counts` in `library(Qtools)` (restricted to $h = \log(\cdot)$)

The jittering approach

It is straightforward, but has limitations:

- Adjiacent values should be equally spaced
- For theoretical results to hold, you need at least one continuous predictor.
- Jittering, even after averaging, might increase standard error.
- Quantile crossing is not uncommon anymore

Censoring

- In some cases you have censored data
- You observe a time-to-event T_i and an indicator Δ_i.
- When $\Delta_i = 0$, you only know that the true time is $> T_i$ for the i-th subject

Censored data

- We focus on random censoring, where censoring is an event independent of the outcome.
- We observe $Y_i = \min(\tilde{Y}_i, C_i)$, with $\Delta_i = I(\tilde{Y}_i < C_i)$
- If $\Delta_i = 0$ we only know that $\tilde{Y}_i > C_i$

Kaplan-Meier

- You clearly can not simply estimate quantiles (or mean and variance) of T_i if you have at least one $\Delta_i = 0$.
- You can not ignore subjects for which $\Delta_i = 0$
- Kaplan-Meier estimator is the solution

$$\hat{S}(t) = 1 - \hat{F}(t) = \prod_{i:T_i \leq \tau} \left(1 - \frac{\sum_j I(T_j \leq T_i) I(\Delta_j = 1)}{\sum_j I(T_j \geq T_i)} \right)$$

- $\hat{Q}_\tau((T, \Delta))$ is the generalised inverse as usual

Quantile function under random censoring

- The main tool is the Kaplan-Meier estimator.
- In the absence of censoring, quantiles solve

$$\arg\min_q \sum_i \rho_\tau(y_i - q)$$

- We rely on Portnoy (2003) to show how to work under random censoring, that is, if $\sum_i \Delta_i < n$.

Quantiles under random censoring

Università di Roma Tor Vergata

- Reorder the sample so that $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$.
- Let $w_i(\tau) = \frac{\tau - i/n}{1 - i/n}$.
- Quantiles solve

$$
\begin{aligned}
\arg\min_q \quad & \sum_{i:\Delta_{(i)}==1} \rho_\tau(y_{(i)} - q) \\
& + \sum_{i:\Delta_{(i)}==0 \& i \leq n\tau} w_{(i)}(\tau)\rho_\tau(y_{(i)} - q) \\
& + \sum_{i:\Delta_{(i)}==0 \& i \leq n\tau} (1 - w_i(\tau))\rho_\tau(\infty - q)
\end{aligned}
$$

- It is now straightforward to work with a linear quantile regression model, by replacing q with $\underline{X}^T\beta$ in the expressions above and minimizing over β.
- Interpretation is as usual, provided you can estimate the quantile of interest in the first place (remember that if you have many censored observations, you can only estimate low quantiles)