Quantile Regression
&
Quantile Regression Averaging

Notes from chapter 1 of the textbook "Quantile Regression", Koenker
&
the article:
"Computing electricity spot price prediction intervals using quantile regression and forecast averaging",
Nowotarski, Weron, 2014

Can Hakan Dagidir
28.11.2021

Motivation

Mosteller and Tukey (1977) remark:

> What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of xs. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set. Ordinarily this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions.

- Quantile regression is intended to offer a comprehensive strategy for completing the regression picture

What is a "quantile"?

- Any real-valued random variable X may be characterized by its **distribution function:** $F(x) = P(X \leq x)$
- And for any $0 < \tau < 1$,
    - $F^{-1}(\tau) = inf\{x : F(x) \geq \tau\}$ is called the τ'th quantile of X

Least Squares Error

Why widely used?

- Computationally nice.
- If noise is normally distributed, performs well.
- Provides a general approach to estimating conditional **mean** functions.

- But as Mosteller and Tukey stated, **mean is rarely satisfactory**.
  - When we might be interested in describing the relationship at different points in the conditional distribution of y, **Quantile Regression** is helpful.

Quantile Regression

- **Classical linear regression** methods are based on: **minimizing sums of squared residuals** to estimate models for **conditional mean functions.**
- **Quantile regression** methods offer a mechanism for estimating **conditional median function,** and/or the full range of other **conditional quantile functions.**

- **What does quantile regression minimize?**

The Minimization Problem

- Least Squares Loss Function:
    - $L = (y - X\beta)^2$, where
      β is the coefficient of the linear model and X is the feature used for prediction
    - i.e. $L = (y - \hat{y})^2$
- Quantile Loss Function:
    - $L = \tau(y - \hat{y})$, if $y \geq \hat{y}$
    - $L = (1 - \tau)(\hat{y} - y)$, if $y < \hat{y}$

The Minimization Problem

- Least Squares Loss Function:
  - $L = (y - X\beta)^2$, where
    β is the coefficient of the linear model and X is the feature used for prediction
  - i.e. $L = (y - \hat{y})^2$
- Quantile Loss Function:
  - $L = \tau(y - \hat{y}),$ if $y \geq \hat{y}$
  - $L = (1 - \tau)(\hat{y} - y)$, if $y < \hat{y}$

<u>i.e.</u>
we want to penalize loss if:
the percentile τ is low, but the prediction $\hat{y}$ is high
the percentile τ is high, but the prediction $\hat{y}$ is low

The Minimization Problem

- A (Least Square) Linear Regression Model tries to minimize:
  - $\sum_{i} (y_i - \hat{y}_i)^2$
- Quantile Regression Model tries to minimize:
  - $\sum_{i:y_i \geq \hat{y}_i} [(\tau(y_i - \hat{y}_i)] + \sum_{i:y_i < \hat{y}_i} [(1 - \tau)\,|\,y_i - \hat{y}_i\,|\,]$

Some notes:

- For example, if an underestimate is marginally **three times more costly** than an overestimate, we will choose $\hat{x}$ so that $P(X \leq \hat{x})$ is three times greater than $P(X > \hat{x})$ to compensate. That is, we will choose $\hat{x}$ to be the **75th percentile of** F.

Median Regression / Least Absolute Deviations

- We can see that choosing $q = 0.5$ gives us Least Absolute Deviation (LAD) (minimizing L1-norm)

Advantages of quantile regression (QR)

- While OLS can be inefficient if the errors are highly non-normal, QR is **more robust to non-normal errors and outliers**.

- QR also provides a richer characterization of the data, allowing us to consider the impact of a covariate on the entire distribution of y, not merely its conditional mean.

- Furthermore, QR is **invariant to monotonic transformations.** So the inverse transformation may be used to translate the results back.

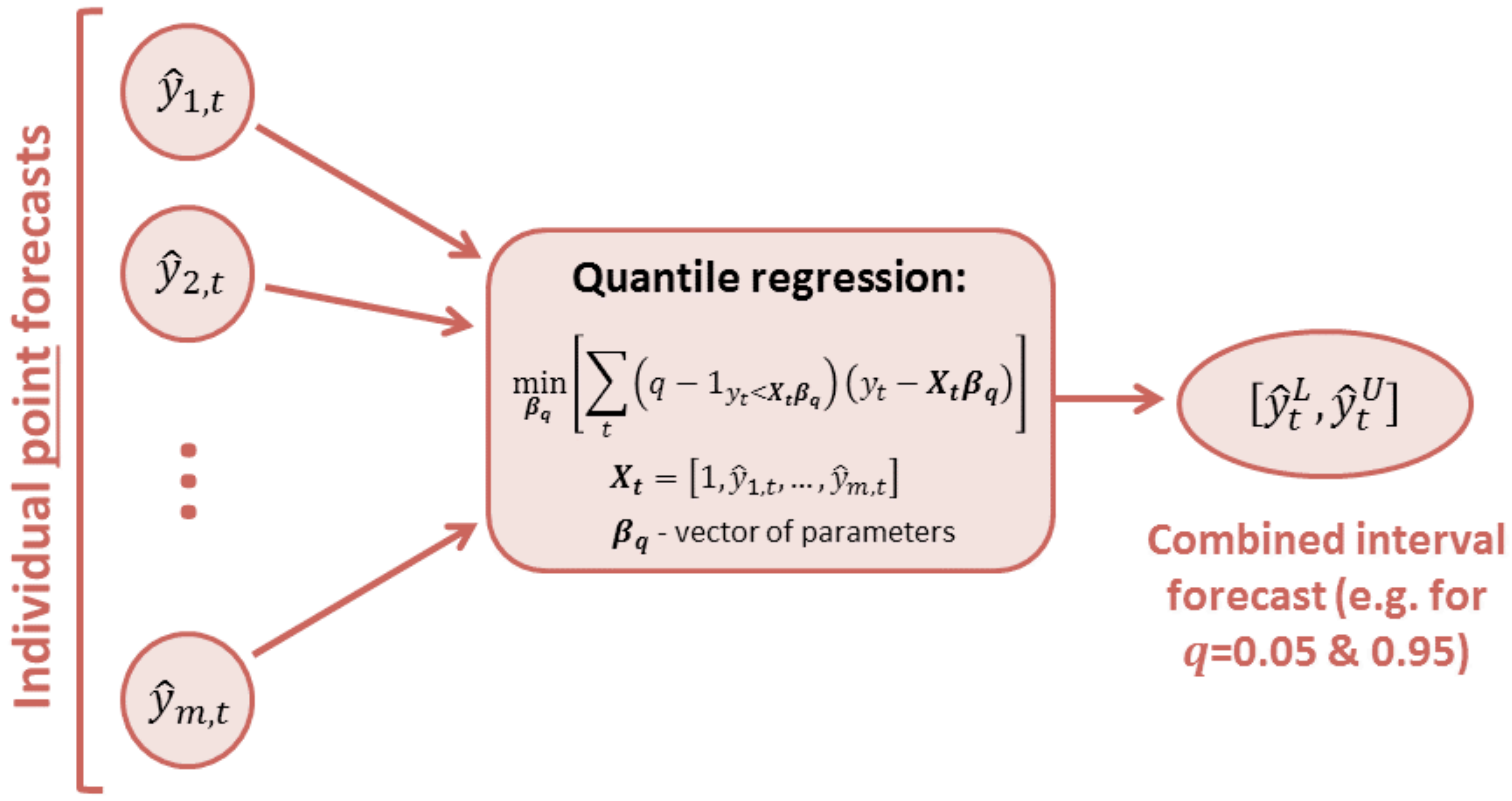
Quantile Regression Averaging

Notes from the paper of Jakub Nowotarski and Rafał Weron

- QRA is first defined/published in Nowotarski and Weron's paper in 2014.
- Published as a new method for constructing Prediction Intervals (PI).

Quantile Regression Averaging

- QRA uses quantile regression with **point forecasts** from other individual models:



QRA

Why Point Forecast?

- "Quantile Regression Averaging (QRA) yields an **interval forecast of the spot price**, but **does not use the PI** (prediction intervals) **of the individual methods**.
- This is an important point, since as Wallis (2005) remarks: **combining intervals directly will not in general give an interval with the correct probability**.
- For instance, Granger et al. (1989) attempt to overcome this difficulty by **estimating combining weights from data on past forecasts** that in effect **recalibrate the forecast quantiles**."
- From wikipedia:
  - One of the reasons for using point forecasts (and not interval forecasts) is their availability. For years, forecasters have focused on obtaining accurate point predictions.

QRA

The Minimization Problem (Almost same as QR)

In our case the averaging problem is given by:

$$Q_p(q|\widehat{\boldsymbol{p}}_t) = \widehat{\boldsymbol{p}}_t \boldsymbol{w}_q, \tag{3}$$

where $Q_p(q|\cdot)$ is the conditional qth quantile of the electricity spot price distribution, $\widehat{\boldsymbol{p}}_t$ are the regressors (explanatory variables) and $\boldsymbol{w}_q$ is a vector of parameters (q in the subscript emphasizes the fact that the parameters are varying for different quantiles). The weights are estimated by minimizing the loss function for a particular qth quantile:

$$\begin{aligned}
&\min_{\boldsymbol{w}_t} \left[\sum_{\{t: p_t \geq \widehat{\boldsymbol{p}}_t \boldsymbol{w}_t\}} q|p_t - \widehat{\boldsymbol{p}}_t \boldsymbol{w}_t| + \sum_{\{t: p_t < \widehat{\boldsymbol{p}}_t \boldsymbol{w}_t\}} (1-q)|p_t - \widehat{\boldsymbol{p}}_t \boldsymbol{w}_t| \right] \\
&= \min_{\boldsymbol{w}_t} \left[\sum_t (q - \mathbb{1}_{p_t < \widehat{\boldsymbol{p}}_t \boldsymbol{w}_t})(p_t - \widehat{\boldsymbol{p}}_t \boldsymbol{w}_t) \right]. \qquad (4)
\end{aligned}$$

QRA

One example for understanding the difference

- Recall that LAD = Least Absolute Deviation (minimizing L1 error)
- QRA-based $50\,\%$ PI $\neq$ LAD-based $50\,\%$ PI:
- QRA: running quantile regression for $q = 0.25$ and $q = 0.75$
- LAD: running quantile regression for $q = 0.5$ then taking 25 and $75\,\%$ quantiles of the distribution of forecast errors (residuals).

To Sum Up

What I need from the Traffic Models for QRA?

- Only the predictions from different **individual(?)** models.

A more detailed look on the book

Consider a simple problem:

If loss is described by the function ,

$\rho(u) = u(\tau - \mathbf{1}(u < 0))$ **,for some $\tau \in (0,1)$.**

Find $\hat{x}$ to minimize expected loss.

We seek to minimize:

$$E\rho_t(X - \hat{x}) = (\tau - 1)\int_{-\infty}^{\hat{x}} (x - \hat{x})dF(x) + \tau\int_{\hat{x}}^{\infty} (x - \hat{x})dF(x) .$$

Differentiating wrt to $\hat{x}$, we have:

$$0 = (1 - \tau)\int_{-\infty}^{\hat{x}} dF(x) - \tau\int_{\hat{x}}^{\infty} dF(X) = F(\hat{x}) - \tau$$

Since F is monotone, any element of $\{x : F(x) = \tau\}$ minimizes expected loss.

When the solution is unique $\hat{x} = F^{-1}(\tau)$, ow, we have an "interval of τth quantiles" from which we may choose the smallest element. (To adhere to the convention that the empirical quantile function to be left-continuous)

A more detailed look on the book

- It is natural that our optimal point estimator for asymmetric linear loss should lead us to the quantiles.
- In the symmetric case of absolute value loss, it yields the median.
- When loss is linear and asymmetric we prefer a point estimate more likely to leave us on the flatter of the two branches of marginal loss.
- Thus, for example: if an underestimate (i.e. $x > \hat{x}$) is *marginally* three times more costly than an overestimate, we will choose $\hat{x}$ so that $P(X \leq \hat{x})$ is three times greater than $P(X > \hat{x})$ to compensate. That is, we will choose $\hat{x}$ to be the 75th percentile of F.

A more detailed look on the book
Empirical case

When F is replaced by the empirical distribution function,

$$F_n(x) = n^{-1} \sum_{i=1}^{n} \mathbf{1}(X_i \leq x)$$

We may still choose $\hat{x}$ to minimize expected loss

$$\int \rho_\tau(x - \hat{x}) dF_n(x) = n^{-1} \sum_{i=1}^{n} \rho_\tau(x_i - \hat{x}) = min!$$

And doing so now yileds the τth *sample* quantile. When τn is an integer there is again some ambiguity in the solution, because we really have an interval of solutions, $\{x : F_n(x) = \tau\}$, but we shall see that this is of little practical consequence.

Much more important is the fact that we have expressed the problem of finding the τth sample quantile, which seems inherently tied to the notion of an ordering, as the solution to a simple **optimization** problem.
In effect we have replaced **sorting** by **optimizing**.

- **Skipping for now. Not sure if really useful**
- **Adding +-error values as 2n dimensional vectors -> linear programming, polyhedral.**

The simple case of ordinary sample quantiles

The problem of finding the τth sample quantile

$$min_{\xi \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \xi),$$

May be reformulated as a linear program by introducing $2n$ artificial, or “slack”, variables $\{u_i, v_i : 1,...,n\}$to represent the positive and negative parts of the vector of residuals. This yields the new problem,

$$min_{\xi,u,v} \in \mathbb{R} \times \mathbb{R}_+^{2n} \{\tau 1_n' u + (1-\tau) 1_n' v \,|\, 1_n' \xi + u - v = y\}$$

where 1_n denotes an n-vector of ones. Above, we are minimizing a linear function on a polyhedral constraint set, consisting of the intersection of the $(2n+1)$ dimensional hyperplane determined by the linear equality constraints and the set $\mathbb{R} \times R_+^{2n}$. Many feature of the solution are immediately apparent from this simple fact. For example, $\min\{u_i, v_i\}$ must be zero for all i, since otherwise, the objective function may be reduced without violating the constraint by shrinking such a pair toward zero.
This is usually called complementary slackness in linear programming. Indeed, for this same reason we can restrict attention to “basic solutions” of the form $\xi = y_i$ for some observation i.