Statistical Learning, Fall 2018-19

Class notes: quantile regression

1 Motivation

Understanding the dependence of Y on X and the conditional distribution $P(Y|X)$ can be much more than modeling its mean! Figure 1 demonstrates that if the noise ϵ in $Y = f(X) + \epsilon$ is "nicely" behaved, e.g. Normal(0,1) then modeling the conditional mean (or median) can actually be enough to understand the complete $P(Y|X)$, however if the noise is more erratic we may want to explicitly model the quantiles to understand $P(Y|X)$. In both panes we have

$$\text{median}(Y|X = x) = 2 \cdot \left(\exp(-30 \cdot (x - 0.25)^2) + \sin(\pi \cdot x^2)\right)$$

and we see the median of $Y|X$ (solid) and the $0.25, 0.75$ quantiles (dashed).

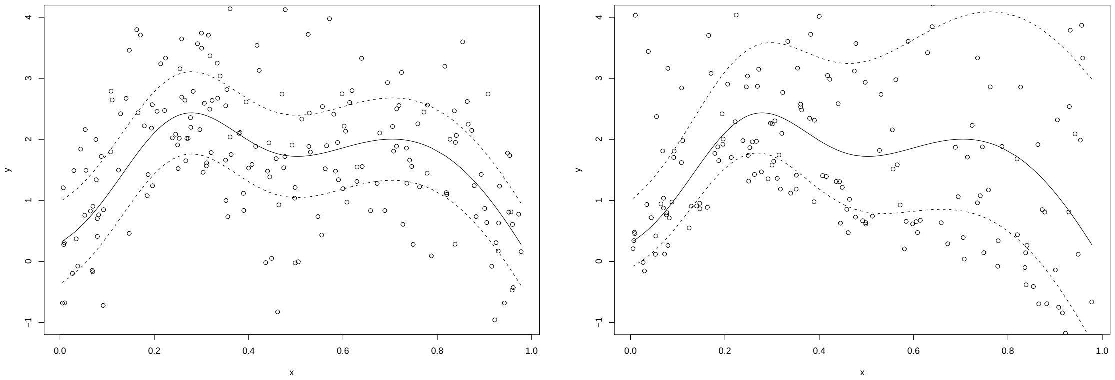

Figure 1: Quantiles for data with Normal(0,1) noise are parallel (left), while those for data with noise that is asymmetric with non-constant variance can look different (right).

2 Modeling

As we showed in class & homework, if we use as EPE criterion the τ-quantile loss:

$$L_\tau(Y, f(X)) = \begin{cases} \tau \times (Y - f(X)) & \text{if } Y - f(X) > 0 \\ -(1-\tau) \times (Y - f(X)) & \text{otherwise} \end{cases}$$

Then the EPE is minimized by setting $f^*(x)$ to be the τth quantile of $P(Y|X = x)$, i.e., $P(Y \leq f^*(x)|X = x) = \tau$. Figure 2 shows the quantile loss for some values of τ. The τ-quantile loss function has angle $\tan^{-1}(\tau)$ with X-axis on the right, and $\tan^{-1}(1-\tau)$ on the left.

If we want to build a model for the τ-quantile of $P(Y|X)$, we have several approaches, comparable to what we did in the case of estimating conditional means:

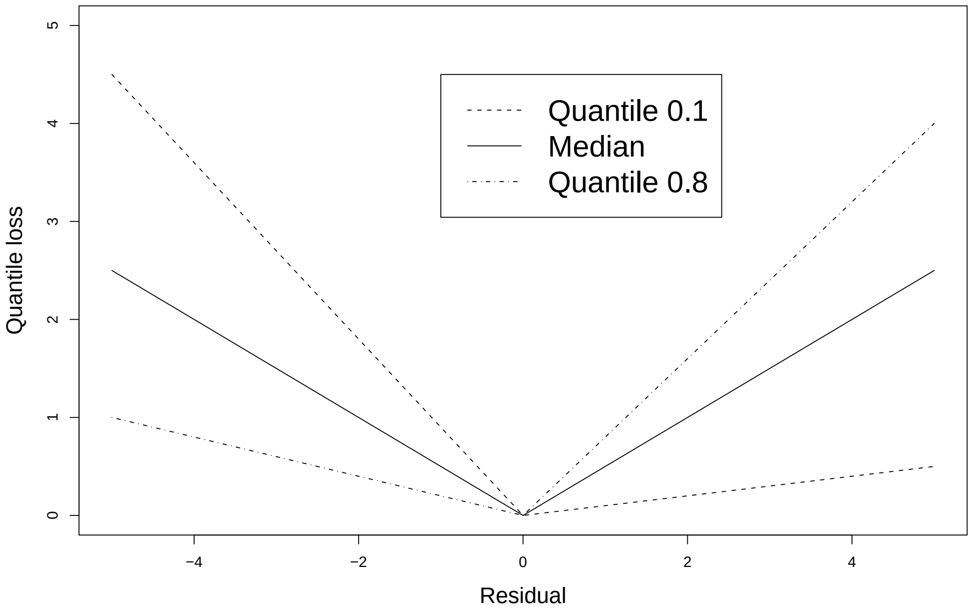

Figure 2: Quantile loss function for some values of τ

- **Empirical risk minimization**, i.e., minimize the EPE loss criterion on the training data, in the spirit of least squares regression for estimating conditional means:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} L_{\tau}(y_i, \mathbf{x}_i^{\mathsf{T}}\beta)$$

- **Local estimation of conditional quantiles**, i.e., approximate the conditional quantile at a point $X = x$ by the quantile of the training observations "in the neighborhood". Given: a point $\mathbf{x}$ for prediction, a neighborhood size k and assuming for simplicity that $(k+1)\cdot\tau$ is integer:

$$\begin{aligned} \hat{Y}_{\tau}(\mathbf{x}) &= \tau\text{th quantile of } Y \text{ in } N_k(\mathbf{x}) \\ &= (y_i \text{ s.t. } \mathbf{x}_i \in N_k(\mathbf{x}) \text{ and } |\{l : \mathbf{x}_l \in N_k(\mathbf{x}), y_l \leq y_i\}| = (k+1)\cdot\tau) \end{aligned}$$

The tradeoffs between bias and variance in this case are essentially similar to those in the conditional mean case.

A few interesting and favorable properties of quantile regression:

1. Since the loss function is *piecewise linear*, solving linear quantile regression is actually a *linear programming* problem, and an easy one at that. The formulation uses the standard *doubling trick* to replace the absolute values by positivity constraints:

$$\begin{aligned} \min_{\beta} \quad & \sum_i \tau\epsilon_i^+ + (1-\tau)\epsilon_i^- \\ \text{s.t.} \quad & -\epsilon_i^- \leq y_i - \mathbf{x}_i^{\mathsf{T}} \leq \epsilon_i^+ \\ & \epsilon^+, \epsilon^- \geq 0 \end{aligned}$$

2. Quantile regression in general and absolute-loss regression (τ=0.5) in particular are robust to outliers and gross errors in the measurement of both Y and X. Detailed and formal discussion of robustness is outside the scope of this overview, but in your HW you will prepare an example of how contamination destroys least squares regression, but leaves quantile regression practically unaffected.

Further reading

Koenker, R. (2005). *Quantile regression.* New York : Cambridge University Press.

Buchinsky, M. (1994). Changes in the u.s. wage structure 1963-1987: Application of quantile regression. *Econometrica*, *62*, 405–458.

Eide, E., & Showalter, M. (1998). The effect of school quality on student performance: A quantile regression approach. *Economics Letters*, *58*, 345–350.

Perlich, C., Rosset, S., Lawrence, R., & Zadrozny, B. (2007). High quantile modeling for customer wallet estimation with other applications. *Proceedings of the Twelfth International Conference on Data Mining, KDD-07.*

Rosset, S. (2008). Bi-Level Path Following for Cross Validated Solution of Kernel Quantile Regression. *Journal of Machine Learning Research*, *10*(Nov), 2473–2505.