Stat 305A: Linear Models **Autumn 2019/20**

Lecture 18: Quantile regression

Lecturer: Art B. Owen *December 3*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. For instance, citations are mostly omitted or imprecisely made. The notes are meant as a memory aid for students who took stat 305A at Stanford University. They appear in* `http://statweb.stanford.edu/~owen/courses/305a/` *where you may find the most recent version. Stanford University holds the copyright.*

18.1 Alternative loss functions

Regression modeling emphasizes squared error loss over all others. It has a simple theory and simple algorithms and we can use a lot of geometric intuition to understand it.

Recall that the regression function is $\mu(x) = \mathbb{E}(Y \mid X = x)$ and it minimizes squared error $\mathbb{E}((Y - \mu(x))^2 \mid X = x)$. The conditional median minimizes absolute error $\mathbb{E}(|Y - \mu(x)| \mid X = x)$.

Suppose that we want instead to model a conditional quantile $Q^\tau(Y \mid X = x)$ for some $0 < \tau < 1$. The Q^τ quantile satisfies $\Pr(Y \leqslant Q^\tau) = \tau$ and we want a function $Q^\tau(x)$. Then we could use a loss function that is minimized at the desired quantile.

Before presenting such a loss function, let's consider why one might want conditional quantiles. One important example arises with the breaking strength of lumber. There it is customary to model $Q^{0.10}$, the tenth percentile. It is not enough to know that the average strength of beams is high enough because the lowest strength ones cause the problem. It is also not practical to model something like the 10^{-9}th quantile (10^{-7}th percentile) because there won't be enough data to do that and besides, multiple pieces (beams, studs, joists) are used in the typical structure, not just one. If one in ten are too weak things may be ok. Also there are customarily some sizeable safety factors baked into engineering. As another example, if your firm is exposed to financial risk you might be more interested in something like $Q^{0.99}$ of your potential losses. (Better yet is to use something like $\mathbb{E}(Y \mid Y \geqslant Q^{0.99})$.)

Now let's define the ***check function***

$$\rho_\tau(u) = u \times \big(\tau - 1\{u < 0\}\big) = \begin{cases} \tau|u|, & u \geqslant 0 \\ (1-\tau)|u|, & u \leqslant 0. \end{cases}$$

We have $\rho_\tau(0) = 0$, it increases linearly with slope τ as u moves away from zero to the right and it increases linearly with slope $1 - \tau$ as u moves away from zero to the left.

Now suppose that Y has density function f and CDF F. In class we saw that

$$\mathbb{E}(\rho_\tau(Y - \theta)) = \int_{-\infty}^{\infty} \rho_\tau(y - \theta) f(y)\, \mathrm{d}y$$

is minimized at $\theta = F^{-1}(\tau) = Q^\tau(Y)$.

In quantile regression we seek β that minimizes

$$\mathbb{E}(\rho_\tau(Y_i - Z_i^{\mathsf{T}}\beta)) \quad \text{for IID data}$$

© Stanford University 2019

and we estimate it via

$$\hat{\beta} = \hat{\beta}(\tau) = \arg\min_{\beta} R_\tau(\beta), \quad \text{where}$$
$$R_\tau(\beta) = \sum_{i=1}^{n} \rho_\tau(Y_i - Z_i^\mathsf{T}\beta).$$

This function R_τ is more difficult to minimize than a sum of squared errors. It can be minimized by linear programming.

We proved in class that if we were to replace Y_i by $\tilde{Y}_i = Y_i + Z_i^\mathsf{T}\gamma$ then the result would be to change $\hat{\beta}(\tau)$ to $\hat{\beta}(\tau) + \gamma$. This of course is what we would think should be true and it is reassuring that it comes out that way. It would also happen for ordinary least squares.

18.2 Low birthweight data example

We looked in class at data from 198,377 births in Chapter 1.5.3 of Koenker's book. There is special interest in babies born with a low birthweight, taken to be below 2.5kg. A customary analysis would be based on logistic regression predicting $1\{Y = 1\}$ as a function of covariates. See Stat 305B. That seems to be throwing away information when we have the actual birth weights. We can distinguish 2.501 from 3.5 and 2.49 from 1.8. Also in a regression model, 2.499 and 2.501 impact the predictions very similarly and while those may be very different for reporting purposes, we might think that the condition of babies born at these two weights is not very different.

The example fit quantile regressions at every $\tau \in \{0.05, 0.10, 0.15, \ldots, 0.95\}$. The fitted β_j were plotted vs τ with confidence intervals. The 0.05 quantile overall was close to the low birthweight threshold 2.5kg. For each predictor we could compare its impact on the 5'th percentile with others such as the 50'th. Some variables were more important in terms of predicing extreme quantiles than moderate ones. We would not have gotten such insights from ordinary least squares.

18.3 Model and inference

The usual model for quantile regression is

$$Y_i = Z_i^\mathsf{T}\beta + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} F. \tag{18.1}$$

Then

$$Q^\tau(Y \mid X = x) = Z(x)^\mathsf{T}\beta + Q^\tau(F)$$

Let's assume that the first component of $Z(x)$ is an intercept and write

$$Q^\tau(Y \mid X = x) = Z(x)^\mathsf{T}\beta + \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}^\mathsf{T} \begin{pmatrix} Q^\tau(F) \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{18.2}$$

Now we see that changing F only affects the intercept. If F is Gaussian or exponential or even Cauchy, we get the same β apart from the intercept.

Also, if we change the quantile τ then that also only affects the intercept, shifting it by $Q^\tau(F)$. That tells us right away that the birthweight data do not follow (18.1) because they had such interesting patterns in how $\hat\beta(\tau)$ changed with τ. More than just the intercept was changing, so that model misses something. It must of course be missing some predictors but not necessarily important predictors. It is very likely that the variance of ε_i is not constant. We would expect larger fluctuations in a baby predicted to be 4kg than in one predicted to be 2kg. Perhaps $\log(Y_i)$ would fit the model better, though of course it is exceedingly unlikely that $\log(Y)$ would have constant variance.

Koenker considers bootstrapping residuals for quantile regression. The bootstrap is based on the approximate pivots

$$\mathcal{L}(\hat\beta^*(\tau) - \hat\beta(\tau)) \doteq \mathcal{L}(\hat\beta(\tau) - \beta(\tau))$$

with $*$ denoting the resampled quantity as usual.

Recall that from (18.2) we know that

$$\beta(\tau) = \beta + Q^\tau(u)\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}$$

under our model, when there is an intercept.

There is a central limit theorem for quantile regression. If (18.1) holds with $(1/n)\sum_{i=1}^n Z_i Z_i^\mathsf{T} = (Z^\mathsf{T}Z)/n \to D$ for a positive definite matrix $D \in \mathbb{R}^{p\times p}$, then

$$\sqrt{n}(\hat\beta(\tau) - \beta(\tau)) \xrightarrow{\mathrm{d}} \mathcal{N}\Big(0, \frac{\tau(1-\tau)}{f(Q^\tau)^2} D^{-1}\Big).$$

The factor $\tau(1-\tau)/f(Q^\tau)^2$ is the limit of n times the variance of the τ'th quantile of an IID sample from F.

18.4 Quantile crossing

Suppose that $\hat\beta(\tau_1)_j \neq \hat\beta(\tau_2)_j$ where this j'th coefficient is not the intercept and $\tau_1 < \tau_2$. Then as the corresponding $Z_j(x)$ varies from $-\infty$ to ∞ for fixed values of the other Z_k, the predictions $Z(x)^\mathsf{T}\hat\beta(\tau_1)_j$ and $Z(x)^\mathsf{T}\hat\beta(\tau_2)_j$ will cross somewhere. There will then be values of Z for which $Z(x)^\mathsf{T}\hat\beta(\tau_1) > Z(x)^\mathsf{T}\hat\beta(\tau_2)$. The quantile regressions will cross as non-parallel lines are assured to do (in Euclidean geometry). Empirically and not surprisingly, these crossings are more likely at the edge of the data range or outside of it. They do however pose a challenge for interpretable predictions.

18.5 More flexible models

We looked at an example from Koenker's Figure 2.6 with daily maximum temperatures in Melbourne plotted versus the prior days' maximum. Instead of plain linear quantiles there were curvy and not necessarily monotone ones. They were fit by B-spline regressions. They did not cross over the range of plotted values (though on close inspection it appears that some crossings were averted by plotting only over a portion of the temperature range).

18.6 Calibration

Given a quantile regression or other model that estimates the τ quantile we can shift the predictions by an amount δ chosen to make a proportion τ of the $\hat{Y}_i(\tau)$ values below their corresponding Y_i observations. This calibration is fixed for all x and $Z(x)$. More complicate calibrations can make δ a function of x and it is also possible to get combined upper and lower limits $\tau_{up} = 1 - \alpha/2$ and $\tau_{lo} = \alpha/2$. There is interesting ongoing work in that area in an area called ***conformal prediction***. The conformal/calibration idea originated for least squares but applies also to quantile regression.