Recitation 8: Quantiles and Nonparametric Regressions

Seung-hun Lee

Columbia University
Introduction to Econometrics II Recitation

March 28th, 2022

Quantile Regression

We want to know more than the conditional expected mean

- We kept a linear regression
$$y = X\beta + u$$
with the moment condition $E(Xu) = 0$ or $E[u|X] = 0$
- With this, we get $E[y|X]$, the conditional mean of y given X
- However values of $E[y|X]$ is not preserved in monotonic transformations
- What if we want to know more? (Effect of X at the median, top 10% of X?)
- So now we answer the question, where is y for those with X at τ percentile?
  - Take τ of y are below this point and $1 - \tau$ above
  - and this information is invariant to increasing monotonic transformations

Quantile regression: Capture β at different values of τ

- Capture the parameter of interest at different location of the conditional distribution
- We try to estimate the conditional quantile

$$q_\tau(y|X) = X\beta_\tau$$

where $\tau \in [0, 1]$ satisfies $F_{y|X}(X\beta_\tau|X) = \Pr(y \leq X\beta_\tau|X) = \tau$

- How do we interpret β_τ?
    - $\tau \times 100\%$ of the observations with covariate X has y below $X\beta_\tau$
    - Changes in X by 1 unit raises the τ-quantile of y by β_τ
- $\hat{\beta}_\tau$ is the τ-quantile estimator for β_τ
- Note that we still keep the linearity of our DGP and that X is exogenous
(If not, Chernozhukov-Hansen IVQR: See 2nd year Microeconometrics)

How does our moment condition look like now?

- Since $\Pr(y \leq X\beta_\tau | X) = \tau$, and we have

$$\begin{aligned} \Pr(y \leq X\beta_\tau | X) &= E[\mathbb{I}(y - X\beta_\tau \leq 0)|X] \\ &= E[\mathbb{I}(u \leq 0)|X] (\because y = X\beta + u) \end{aligned}$$

- So the moment condition we get is similar to $E[u|X] = 0$

$$E[\tau - \mathbb{I}(y - X\beta_\tau \leq 0)|X] = E[\tau - \mathbb{I}(u \leq 0)|X] = 0$$

- Or go farther: Use law of iterated expectations to get similar to $E[Xu] = 0$

$$E[(\tau - \mathbb{I}(y - X\beta_\tau \leq 0))X] = E[(\tau - \mathbb{I}(u \leq 0))X] = 0$$

Enter the check function!

- indicator functions are not nice for differentiation, so we need the check function

$$\rho_\tau(u) = u(\tau - \mathbb{I}(u \leq 0))$$

- Median: Let $\tau = 1/2$. Then the check function becomes

$$\rho_{1/2}(u) = \begin{cases} -\frac{1}{2}u & (u \leq 0) \\ \frac{1}{2}u & (u > 0) \end{cases} = \frac{1}{2}|u| = \frac{1}{2}|y - X\beta_{1/2}|$$

  This becomes equivalent to solving the least absolute deviation problem.

- $\tau = 1/3$: Then the check function becomes

$$\rho_{1/3}(u) = \begin{cases} -\frac{2}{3}u & (u \leq 0) \\ \frac{1}{3}u & (u > 0) \end{cases}$$

  which has a kink at $u = 0$ and is asymmetric.

Solving the moment condition

- The quantile regression estimator at τ, which I write as $\hat{\beta}_\tau$ is obtained from the following minimization problem

$$
\begin{aligned}
\hat{\beta}_\tau &= \arg\min_\beta \widehat{E}[\rho_\tau(y - X\beta)] \\
&= \arg\min_\beta \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(y_i - X_i\beta) \\
&= \arg\min_\beta \frac{1}{n}\sum_{i=1}^{n} (y_i - X_i\beta)[\tau - \mathbb{I}[y_i - X_i\beta \leq 0]]
\end{aligned}
$$

- Because of the kink, taking derivatives is not possible

The wayaround: Lipschitz function and subgradient

- Lipschitz function? This is from convex analysis, no harm in accepting this as given
  - *Let f be a convex function and K be a closed, bounded set contained in the relative interior of the domain of f. Then f is Lipschitz continuous on K. That is, $\exists L$ s.t.*

$$|f(x_2) - f(x_1)| \leq L|x_2 - x_1| \forall x_1, x_2 \in K$$

  - This implies that derivatives involving check function are bounded, if not unique
- Subgradient: Range of values for the derivatives
  - *Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function. A* ***subgradient*** *of f at x is any $c \in \mathbb{R}$ such that*

$$f(y) \geq f(x) + c(y - x) \; \forall y \in dom(f) \iff \frac{f(y) - f(x)}{y - x} \geq c$$

A set of all such subgradients of f is called ***subdiffierential*** *and is denoted as $\partial f(x)$.*

Derivative of ρ_τ is bounded and includes 0

- From $\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i - X_i\beta)$, the FOC is

$$0 \in \frac{1}{n}\sum_{i=1}^{n}\partial\rho_\tau(y_i - X_i\beta_\tau)X_i$$

- $\partial\rho_\tau(y_i - X_i\beta) = [\tau - 1, \tau]$: Bounded derivaties with 0 in intervals
- It means that the correct estimate of the β parameter at τ quantile includes 0 as subgradient at FOC
- We then take a limit $n \to \infty$ to get

$$E[\partial\rho_\tau(y - X\beta)X] = X(\tau - \mathbb{I}[y - X\beta \leq 0|X]) = X(\tau - \Pr(y \leq X\beta|X))$$

and this becomes zero iff $\beta = \beta_\tau$ (Also, consistent & asy. normal, CAN)

In practice, we do linear programming to solve this

- Note that

$$
\begin{aligned}
y_i - X_i\beta &= \max(y_i - X_i\beta, 0) + \min(y_i - X_i\beta, 0) \\
&= \max(y_i - X_i\beta, 0) - \max(X_i\beta - y_i, 0)(\because \min(x, y) = -\max(-x, -y)) \\
&\equiv u_i - v_i \;\; \text{(by design, } u_i, v_i \geq 0, u_i v_i = 0)
\end{aligned}
$$

- As such, we can write the minimization equation as

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(y_i - X_i\beta)[\tau - \mathbb{I}[y_i - X_i\beta \leq 0]] &= \frac{1}{n}\tau \sum_{i|u_i>0} u_i + \frac{1}{n}(\tau - 1) \sum_{i|v_i>0} v_i \\
&= \frac{1}{n}\tau \sum_{i=1}^{n} u_i + \frac{1}{n}(\tau - 1)\sum_{i=1}^{n} v_i
\end{aligned}
$$

So the minimization problem can be mapped out on a u_i, v_i plane

Linear programming, but without defining new notations

- Write

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - X_i\beta)[\tau - \mathbb{I}[y_i - X_i\beta \leq 0]] = \frac{\tau}{n}\sum_{i|y_i - X_i\beta > 0}(y_i - X_i\beta) + \frac{1-\tau}{n}\sum_{i|y_i - X_i\beta \leq 0}(y_i - X_i\beta)$$

which again, is a linear programming framework.

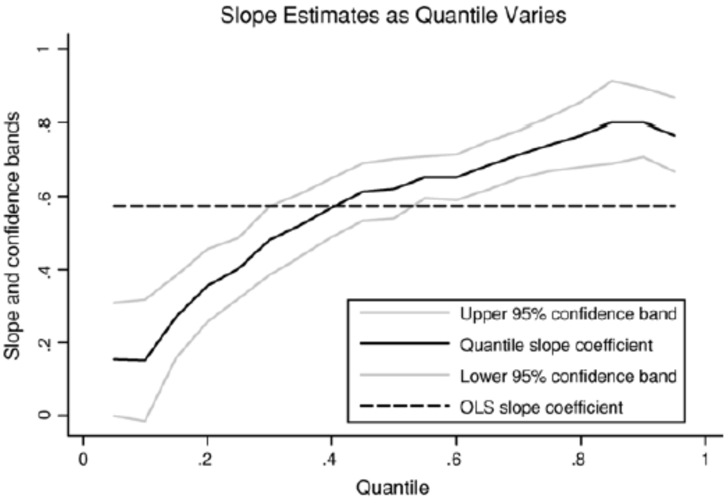
QR in practice: Autor, Houseman, Kerr JOLE 2017

- Using the Detroit's welfare-to-work program, this paper studies the the effect of two government employment programs - direct hire assistance and temporary-help job placements - on distribution of participant's earnings over a 7-quarter period.
- The paper finds that for the **low-tail** of the earnings distribution, neither programs are effective. The direct-hire increases earnings for the high-tail but temporary-help placements negatively affect the distribution for the same group.
- Autor and Houseman (2010) have studied the same program 7 years ago without using the quantile approach and find that on average, the same program lead to earnings gain.
- The key takeaway is that by using quantile regression, you can unmask the effect at a different quantile that you cannot find out through conditional expectations.

QR in practice: From Cameron & Trivedi 2005

- Here, the authors estimate the Engel curve for household annual medical expenditure.
- Data is from 1997 Vietnam Living Standards Survey and $\log$ (medical spending) and $\log$ (total hh expenditure) are dependent and independent variables, so we are estimating elasticity of medical spending w.r.t household expenditure.
- The OLS estimate yield 0.57, indicating inelasticity.
- However, when a quantile regression conditional on expenditure distribution is used elasticity rises with expenditure (and to some extent, income).
    - It ranges from 0.15 for 0.05 quartile, and 0.8 for 0.85 quartile.

OLS vs QR, Cameron & Trivedi 2005



Nonparametric Regression

We now let DGP be anything!

- Assume that an IID data (y_i, x_i) has DGP $P_0(y|X)$
  - DGP being $E[y|X] = X\beta$: we imposed linear (in parameters) model assumption
  - We remove any modeling assumption, other than being a function of X, $E[y|X] = m(X)$
- In essence, we are interested in an estimation problem involving an unknown function
- This approach is called a **nonparametric** approach.
- We normally use nonparametric approach to conduct a diagnostic checking of an estimated parametric model, to conveniently display key features of the dataset in part or in whole, and to conduct an inference under very weak assumptions

Thought experiment: Nonparametric regression in a nutshell

- Discrete Y and X

$$\widehat{P}(y \in A | x \in B) = \frac{n^{-1} \sum_{i=1}^{n} \mathbb{I}(y_i \in A, x_i \in B)}{n^{-1} \sum_{i=1}^{n} \mathbb{I}(x_i \in B)}$$

- Continuous variables: $P_0(Y \leq y|x)$
  - $A \equiv (-\infty, y], B \equiv [x - \epsilon, x + \epsilon]$ and use $\lim_{\epsilon \to 0} \widehat{P}(A|B)$ to back out the DGP
  - As ϵ becomes smaller, there are fewer points to use, leading to highly volatile estimates
  - Even more problematic when we have too many dimensions of X
- Takeaway:
  - We are estimating a distribution f (Kernel density estimation)
  - Interval of X matters in estimation (choice of bandwidth)
  - Much more so with large X dimensions (curse of dimensionality)

Kernel density estimation

- We want an estimation of f that is smooth and non-negative
- Empirical CDF won't do: It is a step function and $f = F'$ has Dirac masses
- Kernel estimation: Idea is that we can estimate $f(y)$ by

$$f(y) \simeq \frac{\int_{y-h}^{y+h} f(u)du}{2h}$$

  over a small interval $[y-h, y+h]$
- Get the numerator estimates with a sample analogue

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{y-h \leq y_i \leq y+h\}$$

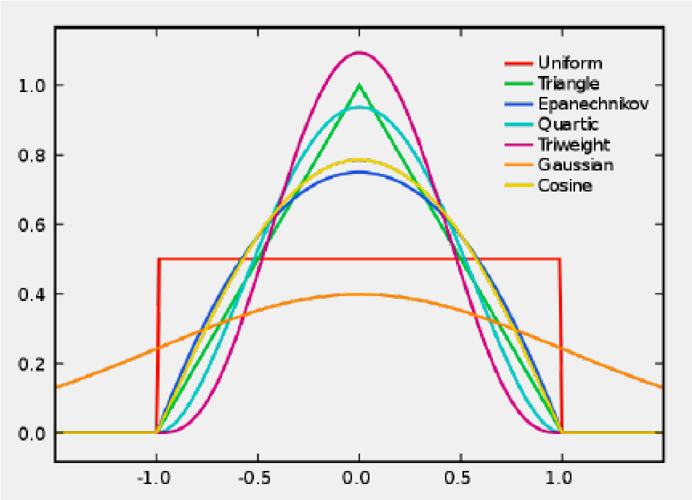
Kernel density estimation: Derivation

- Combining the two expressions, we can approximate $f(x)$ with

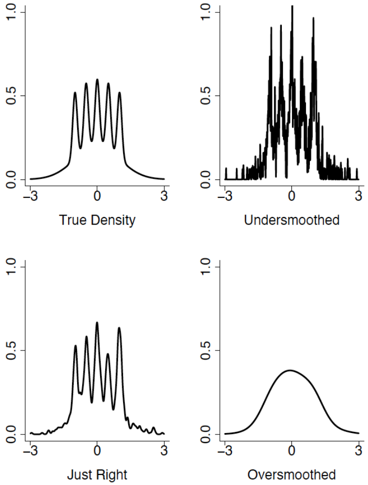
$$\begin{aligned}
\hat{f}_n(y) &= \frac{1}{2nh}\sum_{i=1}^{n}\mathbb{I}[y-h \leq y_i \leq y+h] \\
&= \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{2}\mathbb{I}\left[\left|\frac{y-y_i}{h}\right| \leq 1\right] \\
&= \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{y-y_i}{h}\right)
\end{aligned}$$

- Here, I use $K(u) = \frac{1}{2}1(|u| \leq 1)$. But you can use any other ones that are symmetric, nonnegative over its domain, and integrate to 1.
- $\hat{f}_n(y)$ is the kernel density estimator for $f(y)$

Types of kernels



...but Bandwidth really matters (AKA Bart Simpson Graph)



- Undersmoothed: Too much volatility of $\hat{f}$ in small intervals
- Oversmoothed: Lose accuracy of $\hat{f}$ in large intervals
- What we have: Bias and variance move in opposite directions with respect to the length of the interval (bandwidth) $\rightarrow$ Bias-variance tradeoff!

This is a new problem in nonparametric setups

- In parametric setups, large n took care of everything!

$$
\begin{aligned}
E[(\hat{\theta}-\theta)^2] &= E[((\hat{\theta}-E[\hat{\theta}])+(E[\hat{\theta}]-\theta))^2] \\
&= E[(\hat{\theta}-E[\hat{\theta}])^2+(E[\hat{\theta}]-\theta)^2+2(\hat{\theta}-E[\hat{\theta}])(E[\hat{\theta}]-\theta)] \\
&= E[(\hat{\theta}-E[\hat{\theta}])^2]+E[(E[\hat{\theta}]-\theta)^2]+2E[(\hat{\theta}-E[\hat{\theta}])(E[\hat{\theta}]-\theta)] \\
&= \underbrace{E[(\hat{\theta}-E[\hat{\theta}])^2]}_{V(\hat{\theta})}+\underbrace{(E[\hat{\theta}]-\theta)^2}_{\text{Bias}(\hat{\theta},\theta)^2}
\end{aligned}
$$

- Bias is $O\left(\frac{1}{n^2}\right)$ and that variance is $O\left(\frac{1}{n}\right)$
- Increasing n reduces both!
- Nonparametrics: There is another parameter h....

Smaller h reduces bias...

- We can write the bias as

$$
\begin{aligned}
E[\hat{f}_n(y)] &= E\left[\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y-y_i}{h}\right)\right] = \frac{1}{h}\int_{-\infty}^{\infty} K\left(\frac{y-t}{h}\right) f(t)dt \\
&= \int_{-\infty}^{\infty} K(-u)f(y+uh)du \ \left(\because \frac{t-y}{h} = u \text{ transformation, also } du = \frac{1}{h}dt\right) \\
&= \int_{-\infty}^{\infty} K(-u)\left[f(y) + f'(y)uh + \frac{f''(y)u^2h^2}{2} + o(h^2)\right] dt \ (\because \text{Taylor approximate around } y) \\
&= f(y) + 0 + \frac{1}{2}\int_{-\infty}^{\infty} K(-u)u^2h^2 f''(y)du + o(h^2)
\end{aligned}
$$

  - $\int_{-\infty}^{\infty} K(-u)du = \int_{-\infty}^{\infty} K(u)du = 1$ by symmetry & integrates to 1
  - Symmetry justifies $\int_{-\infty}^{\infty} uK(u)dt = 0$ and $K(u) = K(-u)$
- Bias: $E[\hat{f}(x)] - f(x) = \frac{1}{2}\int_{-\infty}^{\infty} K(u)u^2h^2f''(y)du$

Smaller h increases variance

- We can write variance as

$$
\begin{aligned}
Var[\hat{f}_n(y)] &= E[\hat{f}^2(y)] - (E[\hat{f}(y)])^2 \\
&= E\left[\frac{1}{n^2h^2}\left(\sum_{i=1}^{n} K\left(\frac{y-y_i}{h}\right)\right)^2\right] - (E[\hat{f}(y)])^2 \\
&= E\left[\frac{1}{n^2h^2}\left(\sum_{i=1}^{n} K^2\left(\frac{y-y_i}{h}\right) + 2\sum_{i<j} K\left(\frac{y-y_i}{h}\right)K\left(\frac{y-y_j}{h}\right)\right)\right] - (E[\hat{f}(y)])^2 \\
&= \frac{1}{nh^2}\int_{-\infty}^{\infty} K^2\left(\frac{y-t}{h}\right) f(t)dt + \frac{n(n-1)}{n^2h^2}\left(\int_{-\infty}^{\infty} K\left(\frac{y-t}{h}\right) f(t)dt\right)^2 - \frac{1}{h^2}\left(\int_{-\infty}^{\infty} K\left(\frac{y-t}{h}\right) f(t)dt\right)^2
\end{aligned}
$$

- Focusing on first term, we get (use Taylor expansion and variable transformation)

$$
\frac{1}{nh^2}\int_{-\infty}^{\infty} K^2\left(\frac{y-t}{h}\right) f(t)dt = \frac{1}{nh}\int_{-\infty}^{\infty} K^2(-u)f(y+uh)du \simeq \frac{1}{nh}\int_{-\infty}^{\infty} K^2(u)f(y)du = O\left(\frac{1}{nh}\right)
$$

Choice of h: Minimize asymptotic mean integrated squared error

- AMISE: $\int E[\hat{f}_n(y) - f(y)]^2 dy$
- We can write $E[\hat{f}_n(y) - f(y)]^2$ as

$$\begin{aligned} E[(\hat{f}_n(y) - f(y))^2] &= E[(\hat{f}_n(y) - E[\hat{f}_n(y)] + E[\hat{f}_n(y)) - f(y)]^2] \\ &= E[(\hat{f}_n(y) - E[\hat{f}_n(y)])^2 + (E[\hat{f}_n(y)] - f(y))^2 + 2(\hat{f}_n(y) - E[\hat{f}_n(y)])(E[\hat{f}_n(y)] - f(y))] \\ &= E[(\hat{f}_n(y) - E[\hat{f}_n(y)])^2] + E[(E[\hat{f}_n(y)] - f(y))^2] \\ &= \underbrace{E[(\hat{f}_n(y) - E[\hat{f}_n(y)])^2]}_{V(\hat{f}_n)(y)} + \underbrace{(E[\hat{f}_n(y)] - f(y))^2}_{\text{Bias}(\hat{f}_n(y), f(y))^2} \end{aligned}$$

- Plug in values for bias and variance to get

$$E[(\hat{f}_n(y) - f(y))^2] = \frac{1}{nh} \int_{-\infty}^{\infty} K^2(u) f(y) du + \frac{h^4}{4} (f''(y))^2 \left(\int_{-\infty}^{\infty} K(u) u^2 du \right)^2$$

Optimal 1-dimensional h

- AMISE can be written as

$$\int (\text{Variance} + \text{Bias}^2) dy = \frac{1}{nh} \int_{-\infty}^{\infty} K^2(u) du + \frac{h^4}{4} \int_{-\infty}^{\infty} (f''(y))^2 \left(\int_{-\infty}^{\infty} K(u) u^2 du \right)^2 dy$$

or write $A = \frac{1}{4} \int_{-\infty}^{\infty} (f''(y))^2 \left(\int_{-\infty}^{\infty} K(u) u^2 du \right)^2 dy$, and let $B = \int_{-\infty}^{\infty} K^2(u) du$ to get

$$AMISE = Ah^4 + \frac{B}{nh}$$

- Minimize this w.r.t h to get $h = \left(\frac{B}{4An} \right)^{1/5}$
- Bias and standard errors are both in $n^{-2/5}$ and AMISE will be in $n^{-4/5}$ $\rightarrow$ Estimator not CAN at $n^{-1/2}$ but at slower rate
- We also have $f''(x)$ in A $\rightarrow$ several rules to select h (next class)!