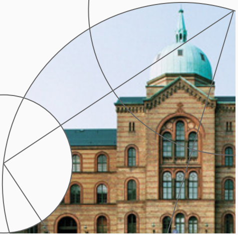
Quantile Regression

Advanced Microeconometrics

Anders Munk-Nielsen

2022



Plan for lectures: Helicopter

Part I: Linear methods. ✓

Part II: High-dimensional methods. ✓

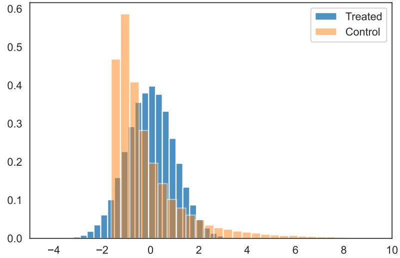
Part III: M-estimation, theory ✓

Part IV: M-estimation, examples ←

Where are we in the course?

Part	Topic	Parameterization non-linear	Estimation non-linear	Dimension dim(x)	Numerical optimization	M-estimation (Part III)	Outcome (y_i)	Panel (c_i)
I	OLS	$\div$	$\div$	low	$\div$	✓	$\mathbb{R}$	✓
II	LASSO	$\div$	✓	high	✓	$\div$	$\mathbb{R}$	$\div$
IV	Probit	✓	✓	low	✓	✓	$\{0,1\}$	$\div$
	Logit	✓	✓	low	✓	✓	$\{1,2,...,J\}$	$\div$
	Tobit	✓	✓	low	✓	✓	$[0;\infty)$	$\div$
	Simulated Likelihood	✓	✓	low	✓	✓	Any	✓
	Quantile Regression	$\div$	✓	(low)	✓	✓	$\mathbb{R}$	$\div$
	~~Non-parametric~~	~~✓~~	~~(✓)~~	~~∞~~	~~$\div$~~	~~$\div$~~	~~$\mathbb{R}$~~	~~$\div$~~

Discussion time!



In the example in the graph, the *average* treatment effect is zero.

Question

What else can we say about the impact of treatment based on this graph?

Topic for today

- **Most** empirical work focuses on the conditional **mean**...
- **Today:** explore conditional **quantiles**.

Outline

Earnings Example

- **Bitler, Gelbach, & Hoynes (2006; AER):** The effect on earnings of Connecticut's Job First waiver program.
- **Finding:** Behind the mean response lies a large response by a small group.

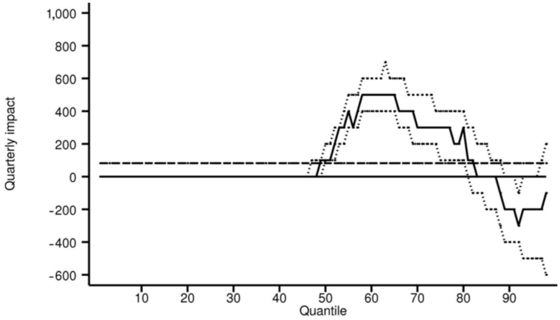
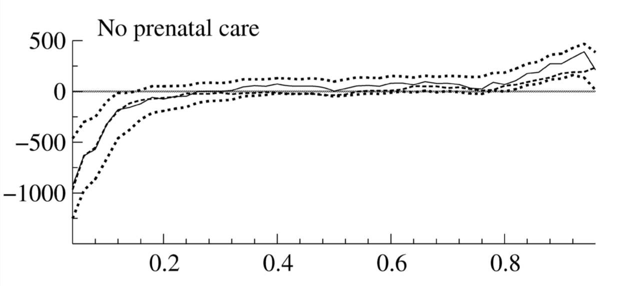


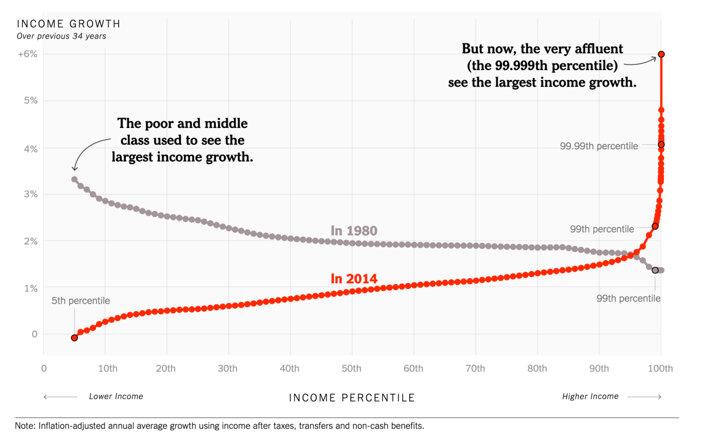
FIGURE 3. QUANTILE TREATMENT EFFECTS ON THE DISTRIBUTION OF EARNINGS, QUARTERS 1–7

Notes: Solid line is QTE; dotted lines provide bootstrapped 90-percent confidence intervals; dashed line is mean impact; all statistics computed using inverse propensity-score weighting. See text for more details.

- **Abrevaya & Dahl (2008):** Birth weight of newborns.
- **Example:** No prenatal visits makes a big difference at bottom quantiles (the “at risk” babies).



Income Inequality



Source: David Leonhardt, NYTimes, `nytimes.com`

Outline

Introduction

- **Normally:** we are interested in the *conditional mean*,

$$\mathbb{E}(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}_m$$

- **Instead:** we might consider

$$\mathrm{Median}(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}_{0.5}$$

- **More generally:** conditional τ-percentiles, $\tau \in (0;1)$

$$\mathrm{Percentile}_\tau(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}_\tau$$

Quantiles

Definition: quantile

The τ'th quantile ($\tau \in (0;1)$) of y is μ_τ such that

$$\tau = \Pr(y \leq \mu_\tau) \equiv F_y(\mu_\tau).$$

Hence

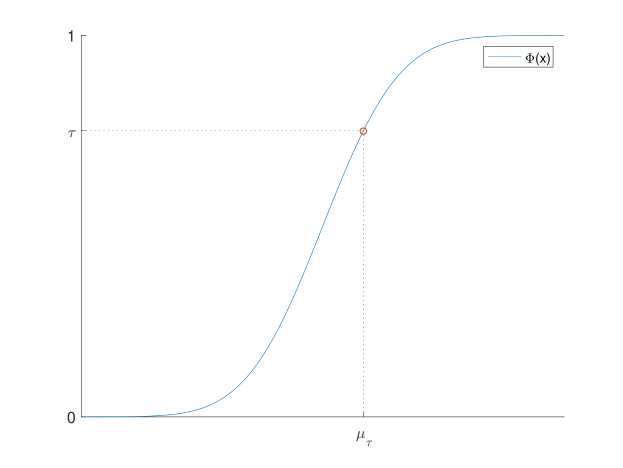
$$\mu_\tau = F_y^{-1}(\tau).$$

Definition: percentile

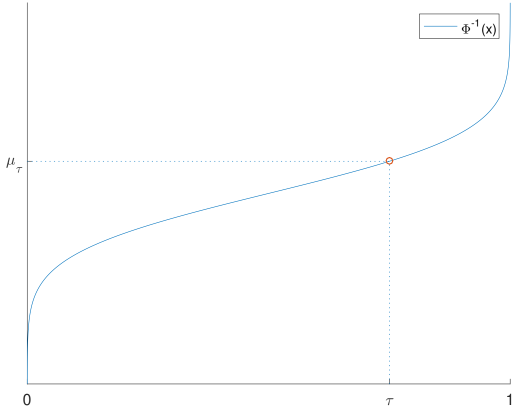
The τ'th quantile is the 100τ'th percentile.

E.g., the 0.23 quantile and the 23rd percentile are the same.

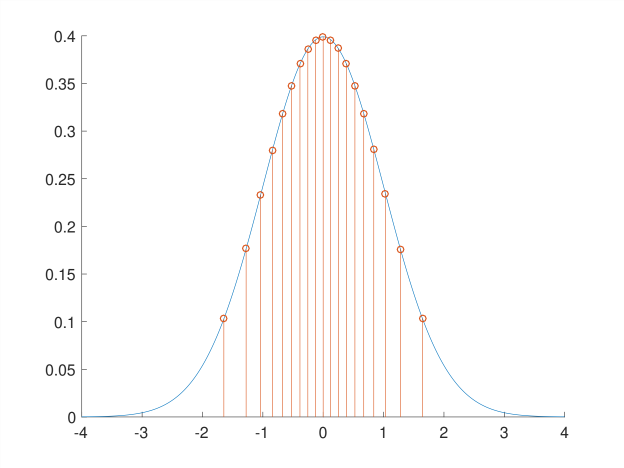
Graphically, $\tau \equiv F_y(\mu_\tau)$: The CDF



Graphically, $\mu_\tau = F^{-1}(\tau)$: The Quantile Function



Quantiles $\{.05, .1, .15, ..., .95\}$



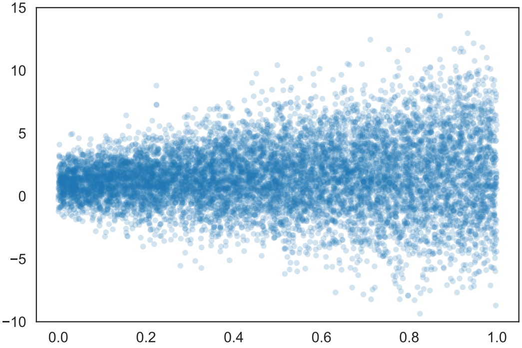
Conditional Quantiles

Definition: conditional quantile

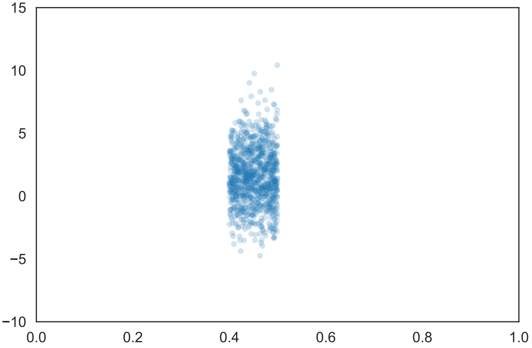
The conditional τ quantile of y is

$$\mu_\tau(\mathbf{x}) = F_{y|\mathbf{x}}^{-1}(\tau|\mathbf{x}).$$

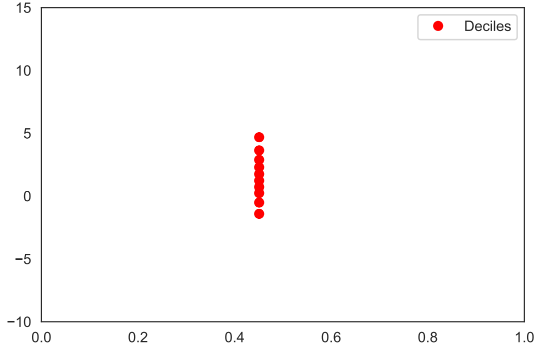
Conditional Quantiles: Dataset



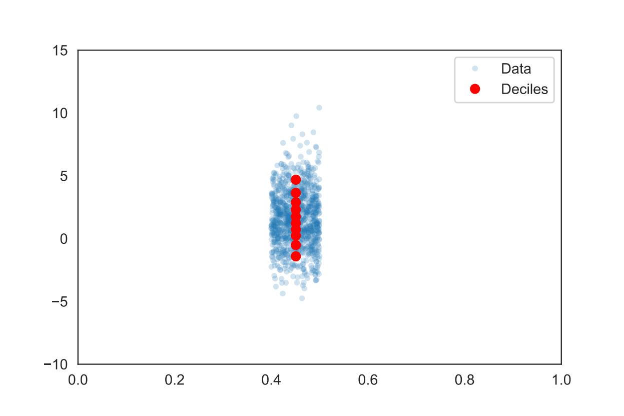
Conditional Quantiles: Focus on x_i



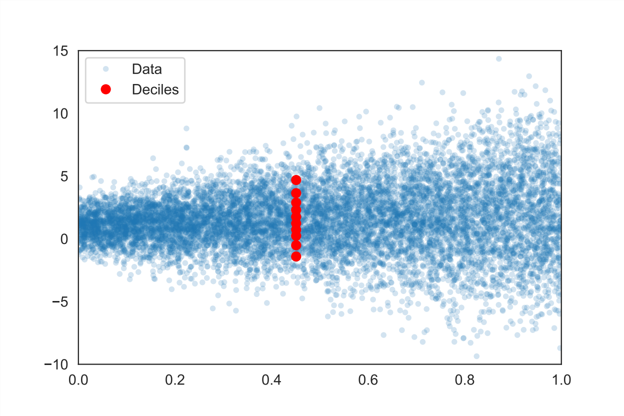
Conditional Quantiles, 0, 5, 10, ..., 100

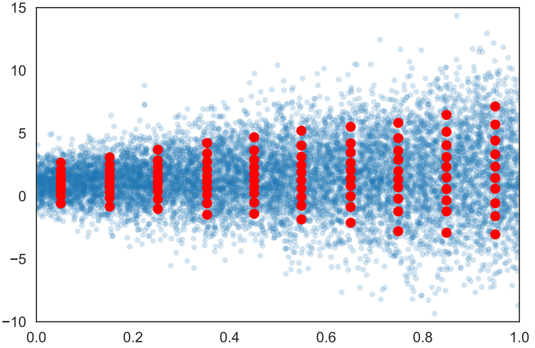


Conditional Quantiles over x



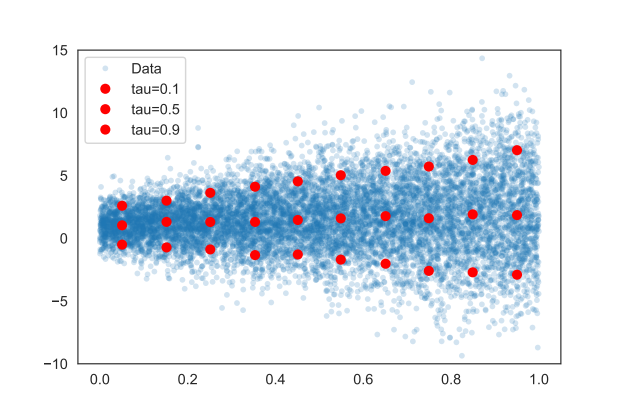
Conditional Quantiles over x



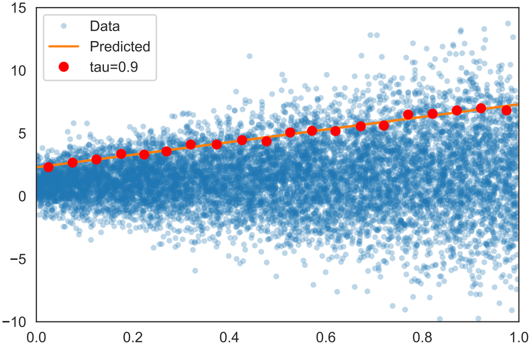


15
10
5
0
−5
−10
0.0
0.2
0.4
0.6
0.8
1.0

Conditional Quantiles over x



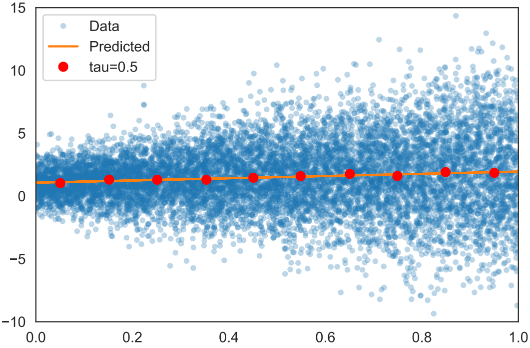
Quantile Regression, $\tau = 0.90$



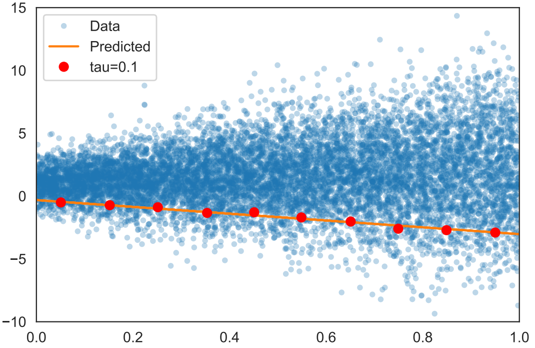

```
why are the dots below the
line

well, finite sample
had we access to an
infinitely large dataset,
we could essentially have
dots perfectly line up
with the line itself
```

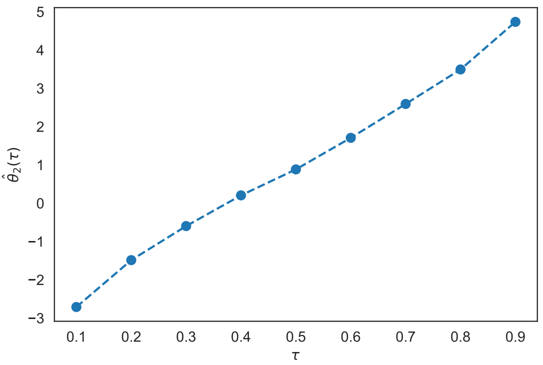
Quantile Regression, $\tau = 0.50$



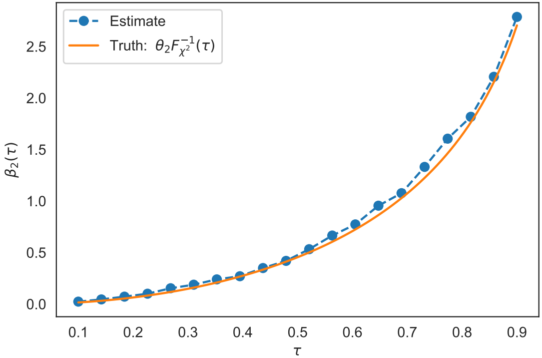
Quantile Regression, $\tau = 0.10$



Quantile Regression, $\tau = 0.10$



Quantile Regression, $\tau = 0.10$



Outline

Model

- Will cover **two specific models** as examples;
  1. Heteroscedasticity model,
     Pro: error-term form,
  2. Heterogenous parameter model.
     Pro: simpler to grasp what we estimate.
- **Finally:** Will cover the more general model.

Model

$$
\begin{aligned}
y &= \mathbf{x}\beta + u \\
u &= (\mathbf{x}\boldsymbol{\alpha})\,\varepsilon, \quad \varepsilon \sim F_\varepsilon,
\end{aligned}
$$

where we assume $\mathbf{x}\boldsymbol{\alpha} > 0$.

- **Goal:** Derive an expression for $\mu_\tau(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha})$.

$$
\begin{aligned}
\tau &= \Pr\left[y \leq \mu_\tau(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha})\right] \\
&= \Pr\left[\mathbf{x}\boldsymbol{\beta} + (\mathbf{x}\boldsymbol{\alpha})\,\varepsilon \leq \mu_\tau(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha})\right] \\
&= \Pr\left\{\varepsilon \leq \frac{1}{\mathbf{x}\boldsymbol{\alpha}}\left[\mu_\tau(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha}) - \mathbf{x}\beta\right]\right\} \\
&\equiv F_\varepsilon\left\{\frac{1}{\mathbf{x}\boldsymbol{\alpha}}\left[\mu_\tau(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha}) - \mathbf{x}\beta\right]\right\}.
\end{aligned}
$$

- **Qs:** Where did we need $\mathbf{x}\boldsymbol{\alpha} > 0$?

Heteroscedasticity model (cf. C&T)

- We have

$$\tau \quad = \quad F_\varepsilon \left\{ \frac{1}{\mathbf{x}\boldsymbol{\alpha}} \left[\mu_\tau(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha}) - \mathbf{x}\boldsymbol{\beta} \right] \right\}.$$

- It follows,

$$\begin{aligned} F_\varepsilon^{-1}(\tau) \quad &= \quad \frac{1}{\mathbf{x}\boldsymbol{\alpha}} \left[\mu_\tau(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha}) - \mathbf{x}\boldsymbol{\beta} \right] \\ \Leftrightarrow \mu_\tau(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha}) \quad &= \quad \mathbf{x}\boldsymbol{\beta} + (\mathbf{x}\boldsymbol{\alpha}) F_\varepsilon^{-1}(\tau). \end{aligned}$$

\mu's are scaled up or down(?)

Discussion

$$\mu_\tau(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathbf{x}\boldsymbol{\beta} + (\mathbf{x}\boldsymbol{\alpha})F_\varepsilon^{-1}(\tau).$$

Suppose $\mathbf{x}_i\alpha = 1$ for all i.

Discuss

What is the difference between two quantile regression lines in this case?

Heterogenous parameter model

Model

```
there is no error term
OR whatever \tau we draw is 'the error term'
```

$$y_i \quad = \quad \mathbf{x}_i\beta(\tau_i), \quad \tau_i \sim \text{Uniform}(0,1),$$

where $\beta(\cdot)$ maps $[0;1]$ into $\mathbb{R}^K$ subject to the *no quantile crossing* restriction [next slide].

- τ_i is interpreted as the conditional **quantile draw** of i.
- **Alternatively,** we can think of the model as

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + e_i, \quad e_i \equiv \mathbf{x}_i\beta(\tau_i) - \mathbf{x}_i\boldsymbol{\beta}.$$

  - we then require $\Pr(e_i \leq \tau_i|\mathbf{x}_i) = 0$.

No quantile crossing

The mapping, $\tau_i \mapsto \mathbf{x}_i\beta(\tau_i)$ must be monotonic for all $\tau_i \in (0;1)$ where $\mathbf{x}_i$ has support (almost surely). `you can't have that the 90th percentile is below 10th percentile`

Discussion

Discuss

Why do we need the no quantile crossing property?

[hint, consider the "constant only," example where $\mathbf{x}_i = 1$, so $\mathbf{x}_i\beta(\tau_i) = \beta_0(\tau_i)$]

Heteroscedasticity DGP in Python

```
def sim_data(theta, N, alpha):
    K = theta.size
    assert alpha.size == K, f'alpha must be same size as beta'

    # 1. regressors
    oo = np.ones((N,1))
    xx = np.random.uniform(size=(N,K-1)) # uniform => ensures posit
    x = np.hstack([oo, xx])

    # 2. error term
    epsilon = np.random.normal(size=(N,))
    sigma_i = x@alpha # individual variance (function of x)
    u = sigma_i * epsilon # heteroskedastic error

    # 3. outcome
    y = x@theta + u
    return y,x
```

- **Note**: we estimate $\hat{\beta}(\tau)$, but the DGP requires θ_o (K-vector) and α_o (K-vector)

DGP in Python

```
def sim_data(theta, N, icdf):
    K = theta.size
    assert K == 2 # for simplicity
    assert callable(icdf)

    oo = np.ones((N,1))
    xx = np.random.uniform(size=(N,1)) + theta[0]
    x = np.hstack([oo, xx]) # (N,K)

    tau = np.random.uniform(size=(N,1))
    beta2 = icdf(tau) * theta[1] INVERSE CDF!!
    beta1 = np.ones((N,1)) # first coefficient constant across quar
    beta = np.hstack([beta1, beta2]) # (N,K)

    y = np.sum(x * beta, axis=1)
    return y,x
```

- **Note**: we estimate $\hat{\beta}(\tau)$, but the DGP requires θ_o (K-vector) and $F^{-1}(\cdot)$ (function)

Outline

Overview

- **Recall:** We minimize $\mathbb{E}[(y-\mu)^2]$ by setting $\mu = \mathbb{E}(y)$. `minimizing sq distances = mean`
- **Similarly:** We minimize $\mathbb{E}|y-\mu|$ by setting $\mu = \mathrm{Median}(y)$. `minimizing abs dist = median`
- **Regression:** We minimize $\mathbb{E}[(y-\mu(x))^2|x]$ by setting $\mu(x) = \mathbb{E}(y|x)$.
  - Typically, we then further assume $\mathbb{E}(y|x) = \mathbf{x}\beta$.
- **Median regression:** minimize $\mathbb{E}\left(|y-\mu(x)|x\right)$ by setting $\mu(x) = \mathrm{Median}(y|x)$.
  - Similarly, we might assume $\mathrm{Median}(y|x) = \mathbf{x}\beta$.

- **In C&T:** They write

$$Q(\beta_\tau) = \sum_{i:y_i \geq \mathbf{x}_i\beta_\tau} \tau\,|y_i - \mathbf{x}_i\beta_\tau| + \sum_{i:y_i < \mathbf{x}_i\beta_\tau} (1-\tau)|y_i - \mathbf{x}_i\beta_\tau|.$$

- **Equivalently**, we can write this as

$$Q(\beta_\tau) = \textstyle\sum_{i=1}^{N} \left[\mathbf{1}_{\{y_i \geq \mathbf{x}_i\beta_\tau\}}\tau|y_i - \mathbf{x}_i\beta_\tau| + \mathbf{1}_{\{y_i < \mathbf{x}_i\beta_\tau\}}(1-\tau)|y_i - \mathbf{x}_i\beta_\tau|\right]$$

- **Getting** rid of the "$|\cdot|$",

$$Q(\beta_\tau) = \textstyle\sum_{i=1}^{N} \left[\mathbf{1}_{\{y_i \geq \mathbf{x}_i\beta_\tau\}}\tau(y_i - \mathbf{x}_i\beta_\tau) + \mathbf{1}_{\{y_i < \mathbf{x}_i\beta_\tau\}}(\tau-1)(y_i - \mathbf{x}_i\beta_\tau)\right]$$

- **In my slides** and everywhere else,

$$\begin{aligned} Q(\beta_\tau) &= \textstyle\sum_{i=1}^{N} \left(\tau - \mathbf{1}_{\{y_i < \mathbf{x}_i\beta_\tau\}}\right)(y_i - \mathbf{x}_i\beta_\tau) \\ &\equiv \textstyle\sum_{i=1}^{N} \rho_\tau(y_i - \mathbf{x}_i\beta_\tau). \end{aligned}$$

"CHECK FUNCTION"

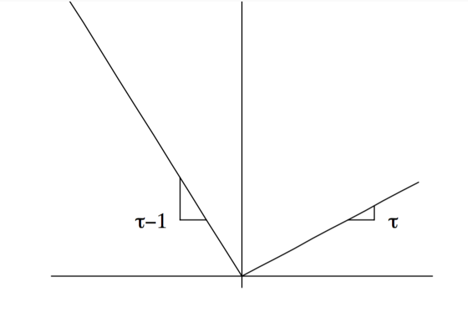
For all obs where y is above predicted, we penalize by \tau
when below, we penalize by 1-\tau

at the 90th percentile (\tau = 0.9)
we punish obs above prediction much more harshly than we do for obs below prediction

More generally

Definition: Check function

$$\rho_\tau(u) \equiv \left(\tau - \mathbf{1}_{\{u<0\}}\right) u.$$



From Koenker (2000).

Criterion function

$$q(y_i, \mathbf{x}_i, \boldsymbol{\beta}_\tau) = \rho_\tau(y_i - \mathbf{x}_i\boldsymbol{\beta}_\tau),$$

where ρ_τ is the *check function*.

- **Claim:** $\boldsymbol{\beta}_\tau = \arg\min_\beta \mathbb{E}[q(y_i, \mathbf{x}_i, \boldsymbol{\beta})|\mathbf{x}_i]$.
- ... we will show this for the intercept-only model.
    - [not the full proof]

Minimizer

- **To show:** $\beta_{0\tau} = \arg\min_{\beta_0} \mathbb{E}[\rho_\tau(y - \beta_0)|x]$.
- **Idea:** Split up $(-\infty; \infty)$ into $(-\infty; \beta_0] \cup [\beta_0; \infty)$ to get rid of $\mathbf{1}\{\cdot\}$.
- Re-write

$$
\begin{aligned}
\mathbb{E}[\rho_\tau(y - \beta_0)|x] &\equiv \int_{-\infty}^{\infty} [\tau - \mathbf{1}\{y - \beta_0 < 0\}] (y - \beta_0) \, \mathrm{d}F(y) \\
&= \int_{-\infty}^{\beta_0} [\tau - \mathbf{1}\{y < \beta_0\}] (y - \beta_0) \, \mathrm{d}F(y) \\
&\quad + \int_{\beta_0}^{\infty} [\tau - \mathbf{1}\{y < \beta_0\}] (y - \beta_0) \, \mathrm{d}F(y) \\
&= (\tau - 1) \int_{-\infty}^{\beta_0} (y - \beta_0) \, \mathrm{d}F(y) \\
&\quad + \tau \int_{\beta_0}^{\infty} (y - \beta_0) \, \mathrm{d}F(y).
\end{aligned}
$$

Step 2: FOC

- **We have**

$$\mathbb{E}[\rho_\tau(y-\beta_0)|x] = (\tau-1)\int_{-\infty}^{\beta_0}(y-\beta_0)\,\mathrm{d}F(y) + \tau\int_{\beta_0}^{\infty}(y-\beta_0)\,\mathrm{d}F(y).$$

- **FOCs** imply

$$\begin{aligned} \frac{\partial\mathbb{E}[\rho_\tau(y-\beta_0)|x]}{\partial\beta_0} &= 0 \\ &[\ldots] \\ \Leftrightarrow F(\beta_0) &= \tau. \end{aligned}$$

- **Intuition:** Choose β_0 so that τ of the mass is to the left.
  - [cdf(x) $\equiv$ % of mass to the left of x]

Identification

- **Intuitively:** similar to OLS.
- **Further** details get much more complicated.

Question

Why can I not identify a model with two intercepts?

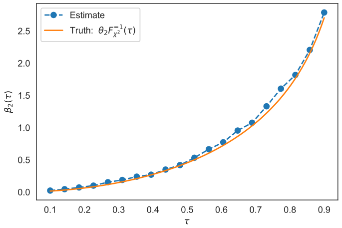
```
suppose I know the true parameters

Not changing "x", yields the same value criterion ->
thus unable to identify a model with two intercept (?)
```

Model

$$y_i = \mathbf{x}_i \beta(\tau_i), \quad \tau_i \sim \text{Uniform}(0,1),$$

where $\beta_0(\tau) = 1 \forall \tau$ and $\beta_1(\tau) = F^{-1}(\tau)$, and F is the χ^2 cdf.



we can evaluate the function at infinitely many points

Discuss

What are the parameter estimates? A vector? A function?

the thing we are going to learn about is a function (see above) -

Outline

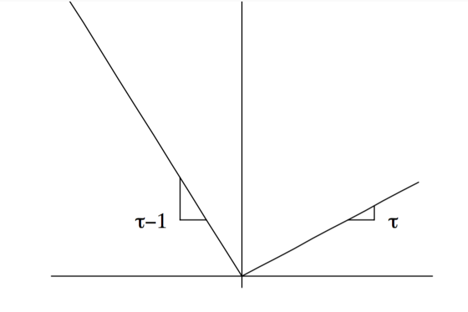
- **If** we are interested in how x_{ik} affects conditional quantiles...
  - ... it is linear!
- **That is,**
$$\frac{\partial \mu_\tau(\mathbf{x}_i, \boldsymbol{\beta}_\tau)}{\partial x_{ik}} = \beta_{\tau k}.$$
- **Often** we will plot all the estimates together, comparing with OLS.
  - [see examples earlier]

Outline

Recall

Definition: Check function

$$\rho_\tau(u) \equiv [\tau - \mathbf{1}\{u < 0\}]\, u.$$



From Koenker (2000).

Criterion function

$$q(y_i, \mathbf{x}_i, \beta_\tau) = \rho_\tau(y_i - \mathbf{x}_i\beta_\tau),$$

where $\rho_\tau(u) \equiv [\tau - \mathbf{1}\{u < 0\}]\, u$.

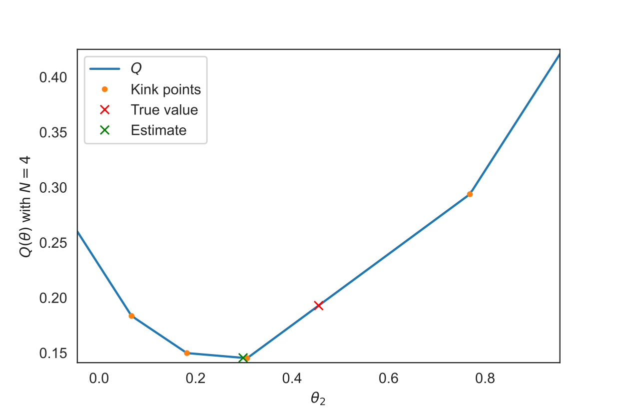
- Two problems;
  1. Non-differentiable at $y_i = \mathbf{x}_i\beta_\tau$,
  2. Hessian is zero everywhere. second derivatives are all zero
- **Implications:**

  Newton-method is bad (step-sizes are gonna be too large, f'/f'' are infinitely large)

  1. Optimization may be hard.
  2. Theorem 12.3 doesn't guarantee asymptotic normality,

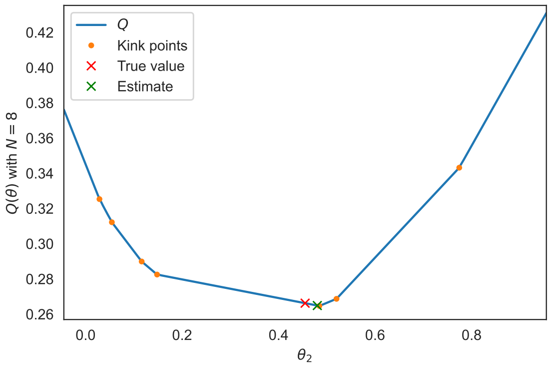
Criterion for various N

Figure 1: $N = 4$



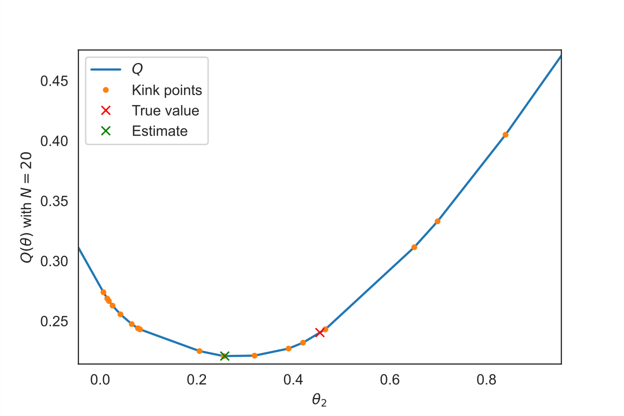
Criterion for various N

Figure 2: $N = 8$



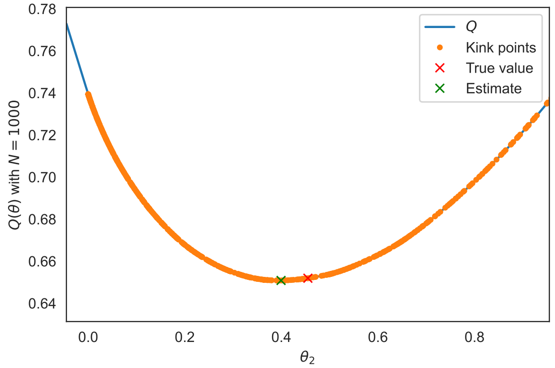
Criterion for various N

Figure 3: $N = 20$



Criterion for various N

Figure 4: $N = 1000$



Discuss

Discuss

Which problems do we face here:

1. The gradients may point the wrong way,
2. We cannot divide by the Hessian, even though we can't see,
3. There are local optima. function will never problem with this

Also: Does the primary problem get better or worse as $N \to \infty$? Why?

USE NON-GRADIENT BASED OPTIMIZER

NELDER MEAD

Standard errors

- **Note:** Criterion is not smooth,
  - Non-differentiable at $y_i = \mathbf{x}_i \beta_\tau$,
  - Hessian is zero everywhere else.
- **Fortunately:** Normality still comes about...

$$\sqrt{N}(\hat{\beta}_\tau - \beta_\tau^o) \xrightarrow{d} \mathcal{N}(0, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}).$$

  - $\mathbf{A}_o = \text{plim} \frac{1}{N} \sum_{i=1}^{N} f_{u_\tau|x}(0|\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'$,
  - $\mathbf{B}_o = \text{plim} \frac{1}{N} \sum_{i=1}^{N} \tau(1-\tau)\mathbf{x}_i\mathbf{x}_i'$,
  - $u_\tau \equiv y - \mathbf{x}\beta_\tau$, $f_{u_\tau|x}$ is the conditional pdf of u_τ given $\mathbf{x}_i$.

- **Bootstrap:** Use when in doubt.
- **Many variants** exist (non-parametric, parametric, block, ...).

Bootstrap

1. Draw a new dataset *with replacement* from the original dataset,
2. Estimate parameters, $\hat{\beta}_\tau^b$,
3. Repeat 1–2 for $b = 1, ..., B$.
4. The distribution of $\{\hat{\beta}_\tau^b\}_{b=1}^B$ is the approximation of the asymptotic distribution of $\hat{\beta}_\tau$.

Intuition

- **Original** dataset was an IID draw from the population.
- **Thus,** IID draws from this sub-population will match the original distribution closely.
- **Panel data:** Do the bootstrap sampling over i (not i, t);
  - i.e. take entire T-paths for individual i when he is drawn.