Linear Regression III: Quantile Estimation

Paul Goldsmith-Pinkham

February 14, 2022

A brief refresher on OLS (and GMM)

- Recall that OLS is the “least-squares” method – it can be defined as the method that minimizes the sum of squared “errors”
  - These errors are the residuals from say, our linear model:

$$E(y_i|x_i) = x_i\beta, \qquad \hat{\beta}_{ls} = \arg\min_{\beta} \sum_i (y_i - x_i\beta)^2 = \arg\min_{\beta} (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$$

  - No surprise – the least squares method is finding the “least” of the squares. In particular, we can use calculus to get our analytic solution, since we’re trying to minimize an objective function:

$$-\mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\beta}) = 0 \qquad -\mathbf{X}'\mathbf{Y} + \mathbf{X}'\mathbf{X}\hat{\beta} = 0 \qquad \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

- The least squares does a lot of work for us by creating a nice objective function
  - Beyond that, what does a quadratic obj. function do?

A brief refresher on OLS (and GMM)

- Key features of OLS:
  - Squared loss function leads to heavily penalization from big outliers
  - Local approximation to the conditional expectation function – OLS finds the closest linear fit to the CEF
  - In context of treatment effects, gives us approximation to the ATE

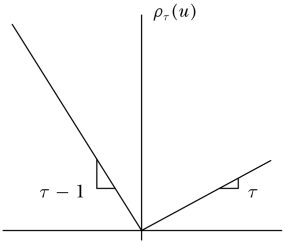
- Most important feature of OLS for today: it characterizes features of the mean of our outcome variable, conditional on covariates (e.g. treatments)
  - What if we care about other things?
  - What are some properties of means that are problematic?
    - Very sensitive to outliers!

Quantiles - some definitions

- First, recall that for any r.v. X we can define its CDF and inverse CDF:

$$F(x) = Pr(X \leq x), \qquad F^{-1}(\tau) = \inf\{x : F(x) \geq \tau\}$$

  - The infimum deals with ties
  - $\tau = 0.5$ is the median!



- Consider now the following loss function:

$$\rho_\tau(u) = u\tau 1(u > 0) + u(\tau - 1)1(u < 0) = u(\tau - 1(u < 0))$$

  - $\tau = 0.5 \longrightarrow \rho_\tau(u) = 0.5|u|$

- We can talk about expected loss (a la OLS):

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

Quantiles as solutions

$$E(\rho_\tau(X - \hat{\mu})) = \tau \int_{\hat{\mu}}^{\infty} (x - \hat{\mu}) dF(x) + (1 - \tau) \int_{-\infty}^{\hat{\mu}} (x - \hat{\mu}) dF(x)$$
$$\rightarrow \hat{\mu} = F^{-1}(\tau)$$

- This problem naturally lends itself to generalization. Let $Q_\tau(Y|X) \equiv \inf\{y : F_Y(y|X) \geq \tau\}$ be the *conditional quantile function*, analogous to the conditional expectation function

- This function minimizes the ρ_τ distance between some function of X and Y:

$$Q_\tau(Y|X) = \arg\min_{q(X)} E(\rho_\tau(Y - q(X)))$$

- Just as we denoted approximated the conditional expectation function with a linear model, we can approximate the $Q_\tau(Y|X)$ with a linear model!

Quantiles as solutions

- Consider now our linear model minimizer:

$$\beta(\tau) \equiv \arg\min_{\beta} E(\rho_\tau(Y - X'\beta))$$

- This is the best linear predictor under the ρ loss function
    - But how does it map to the true $Q_\tau(Y|X)$?

- Key result from Angrist et al. (2006): this linear model is the weighted least squares approximation to the unknown CQF

$$\beta(\tau) = \arg\min_{\beta} E\left[w_\tau(X,\beta)\Delta_\tau^2(X,\beta)\right], \qquad \Delta_\tau(X,\beta) = X'\beta - Q_\tau(Y|X),$$

where the w_τ are *importance* weights, and average over the difference between the true CQF and the linear approximation.

How is it solved?

- Unlike OLS, there is no direct analytic solution for $\beta(\tau)$
  - This implies that the problem needs to be solved numerically
- Key insight: you can redefine the minimization problem of

$$\hat{\beta}(\tau) = \arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau(Y_i - X\beta)$$

  as a linear programming problem.

- We're not going to get into the details of this – others have suffered for us
  - See Chapter 6 of Koenker (2005) or appendix of Koenker and Bassett (1978)

Variance properties

- Let's walk through thinking about the variance of a quantile. Let $\xi_\tau = F^{-1}(\tau)$, with density $f(\xi)$
  - E.g. this is a quantile estimate
  - How can we talk about its limiting properties?

- Key trick: as we move around our estimate of ξ_τ, we can think about the contribution that this has to our objective function (e.g. the gradient):

$$g_n(\xi) = n^{-1} \sum_i 1(Y_i < \xi) - \tau)$$

- As a result, you can think about the variability in our estimate coming from a series of coinflips on whether the data point is above or below the quantile estimate
  - Convergence of the estimate is implied by the convergence of the empirical CDF to the true CDF
  - Normality is a side benefit, and under iid data:

$$\sqrt{n}(\hat{\xi}_\tau - \xi_\tau) \rightarrow \mathcal{N}(0, \tau(1-\tau)f^{-2}(\xi_\tau))$$

Variance properties

- The non-i.i.d. error form of the limiting distribution for $\hat{\beta}(\tau)$ is familiar:

$$\sqrt{n}(\hat{\beta}(\tau) - \beta(\tau)) \to \mathcal{N}(0, \tau(1-\tau)H_n^{-1}J_nH_n^{-1}$$
$$J_n(\tau) = n^{-1}\sum_i x_i' x_i$$
$$H_n(\tau) = n^{-1}\sum_i x_i' x_i f_i(\xi_i(\tau))$$

- The asymptotic variance of the estimator relies on knowledge of the density function
- That makes it harder (and slower!) to compute
- $\tau(1-\tau)$ is smaller in the tails, but f_i is poorly estimated there, which tends to dominate.

Properties of Quantile Regressions (and sometimes OLS)

Equivariance (Koenker and Basset (1978) Consider a linear model $y = x\beta + epsilon$

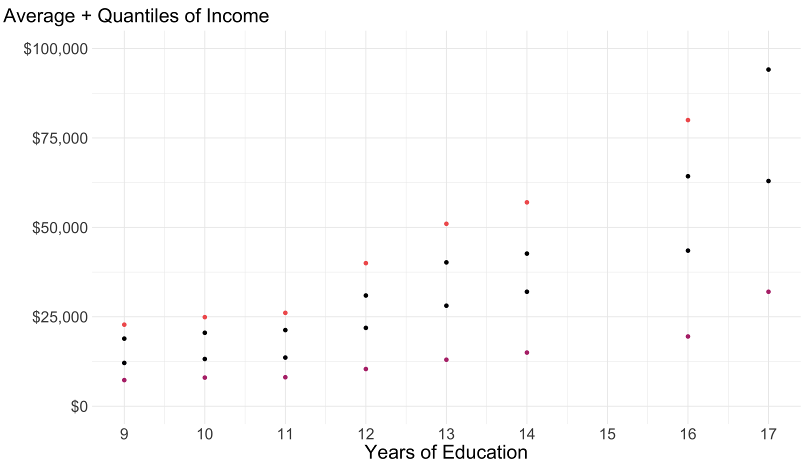
1. Scale equivariance:
   - scaling y by some constant a implies that $\hat{\beta} \to a\hat{\beta}$
2. Shift equivariance
   - adding to y some amount $X\gamma$ implies that $\hat{\beta} \to \hat{\beta} + X\gamma$
3. equivariance to reparametrization of design
   - Linear combinations of regressors leads to linear combinations of coefficients
4. *equivariance to monotone transformations*
   - Let $h(\cdot)$ be monotone function
   - $Q_{h(Y)}(\tau) = h(Q_Y(\tau))$
   - E.g. the median of log(Y) is the log of the median of Y!
   - Something OLS does *not* have
5. The influence function of quantile regression is *bounded* with respect to y
   - This is not the case for OLS (outliers can have unlimited influence)

Practically, why are these properties useful?

- Skewed variables– no more worrying about logs or outliers in the outcome variable
- Censoring – in many datasets, our outcome variables are top-coded or bottom-coded
    - Note that given the influence function results, this is not a problem – we can still identify (some) of the quantile functions
- Let's look at an example

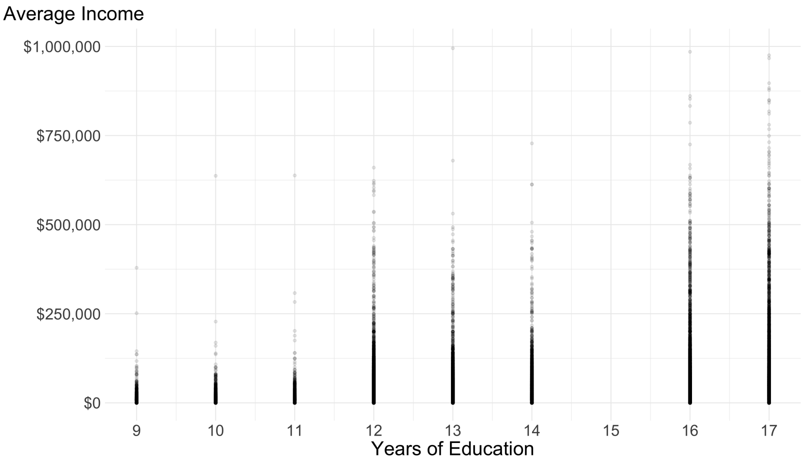
Quantile regression

- Education + Income gradient
- Clear heteroskedasticity



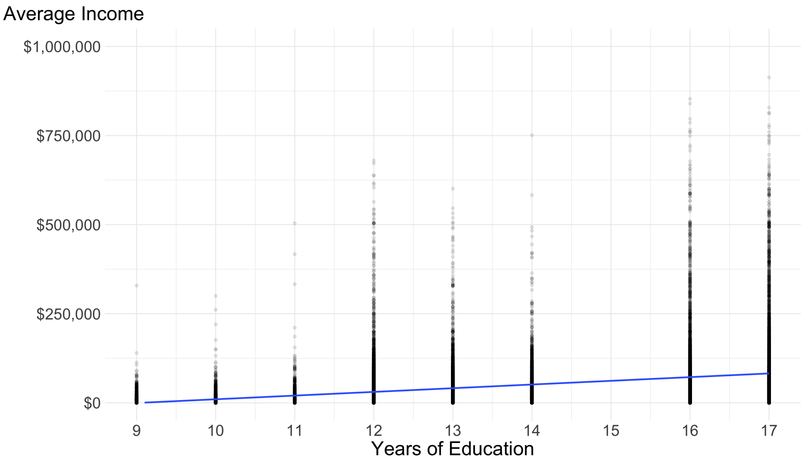
Quantile regression

- Education + Income gradient
- Clear heteroskedasticity
- Very wide variance, especially at high education



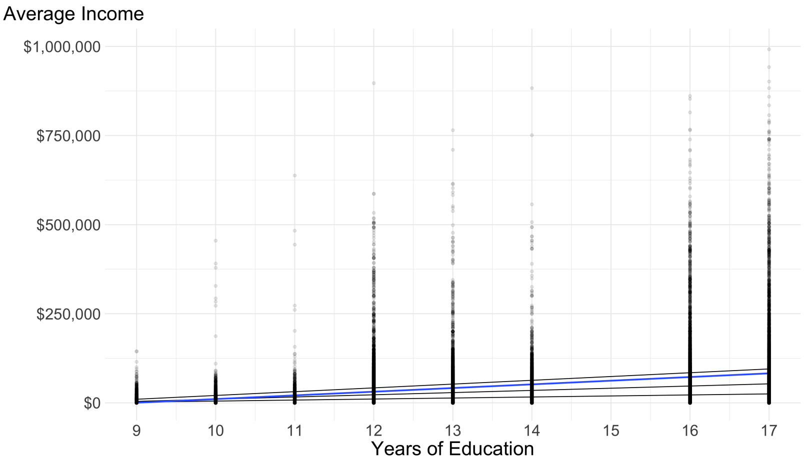
Quantile regression

- Education + Income gradient
- Clear heteroskedasticity
- Very wide variance, especially at high education



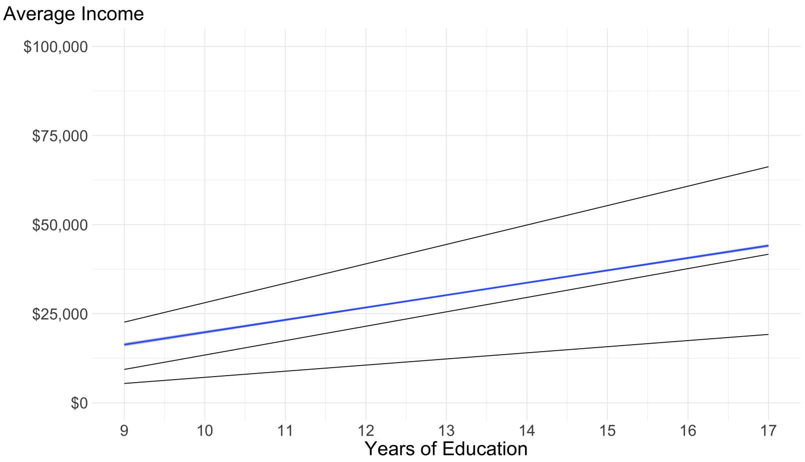
Quantile regression

- Education + Income gradient
- Clear heteroskedasticity
- Very wide variance, especially at high education
- OLS is heaviliy influenced by the tails of income



Quantile regression

- Education + Income gradient
- Clear heteroskedasticity
- Very wide variance, especially at high education
- OLS is heaviliy influenced by the tails of income



Interpreting Quantile Coefficients

- There are some very nice features of this setup.
  - Very robust
- However, interpreting these coefficients from a structural model standpoint is challenging
  - Even Koenker's book punts on this issue – instead pointing out that the OLS interpretions are probably wrong!
- Why is it so hard? Let's dig into this.

Interpreting Quantile Regressions

- Consider a binary treatment variable D_i – in fact, let's use the NSW program from Lalonde
- Consider the very simple OLS verison testing this model using the experimental data:

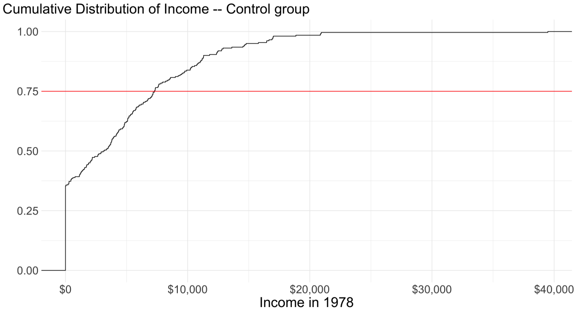
$$y_i = \alpha + D_i\beta + \epsilon_i$$

- Recall that this will estimate our ATE for the treatment
- What is the interpretation of this affect?
  - $E(Y_i(1)) - E(Y_i(0))$ – in other words, the expected change in the outcome for a person moving from untreated to treated
  - That's a useful metric!

Estimate	Point Est.	SE
β_{OLS}	1794.3	(632.9)

Interpreting Quantile Regressions

- Now consider if I did quantile regression instead? What is that doing?
- Previously, we were comparing means of the two distribiutions – e.g. $Y(1)$ and $Y(0)$. We did not need to specify anything about the joint distribution of $Y(1)$, $Y(0)$
- Why does this matter?
  - Consider a person sitting in the control group at the 75 percentile e.g. $Y_{0.75}(0)$
  - What is their relevant treatment effect?



Interpreting Quantile Regressions

- Types of treatment effects can focus on verisons:
    1. Just comparing parts of the *distribution*: $q_{1,\tau} - q_{0,\tau}$ (e.g. Firpo (2005))
    2. Assume rank invariance – e.g. that individuals' rank in the distribiution does not change in moving from control to treatment (e.g. Chernozhukov and Hansen (2005))

- The second approach is very strong, and gets you a lot of mileage (e.g. extremely useful for IVQR)

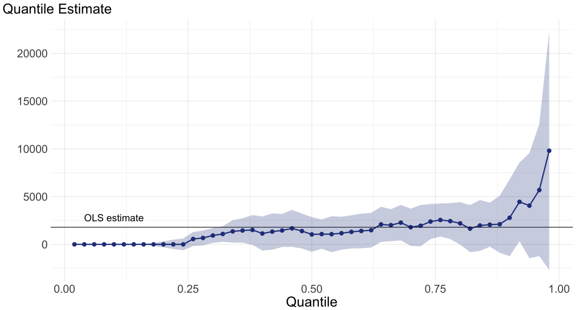
- The first approach requires weaker assumptions, but then we cannot say anything about what the effect of a policy is on a person in a given part of the distribiution.
    - Instead, our policy takeaways are integrated over changes in the full shape

Interpreting Quantile Regressions

- Now we can look at the effect of NSW across the distributions

- Remarkably homogeneous

- 20% of distributions had zero income, so degenerate effects. However, can trace out distributional effects for large groups



Interpreting Quantile Regressions

- How does this compare efficiency-wise?
- Much noisier – compare median, 75th percentile and 95th
- Important to be holistic about estimates in this setting; b/c of joint estimation problem of density and quantiles, different quantiles can be better estimated

Estimate	Point Est.	SE
β_{OLS}	1794.3	(632.9)
$\beta_{0.5}$	1038.3	(872.3)
$\beta_{0.75}$	2342.5	(893.4)
$\beta_{0.95}$	2992.2	(2973.0)

A result from Firpo (2005)

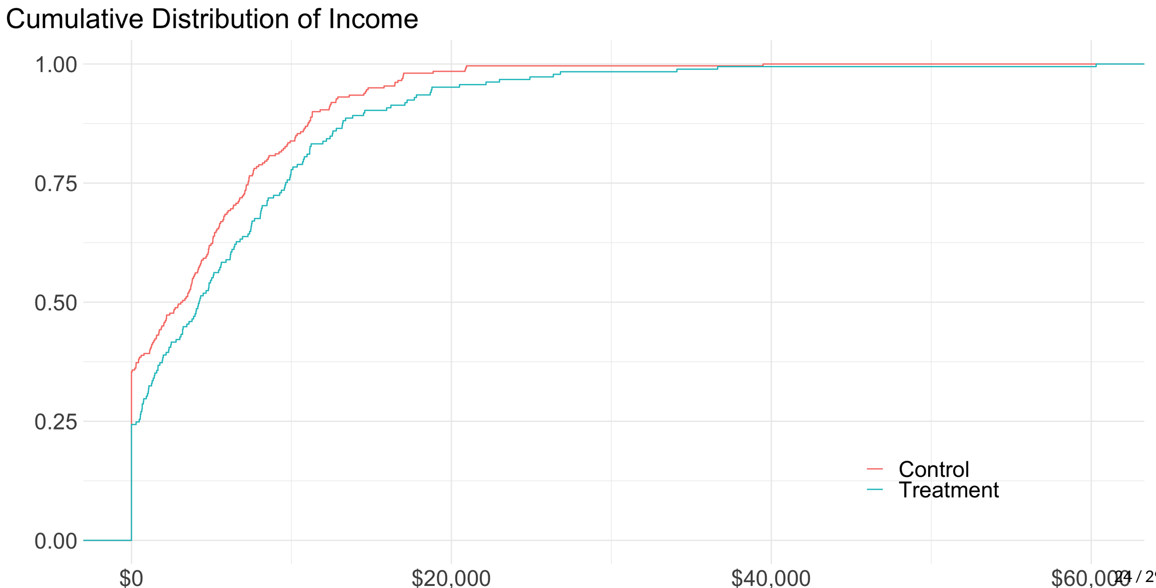
- An analagous IPW estimator which we used for efficient estimation of ATE can be used for estimating QTE: $\beta_\tau = \hat{q}_{1,\tau} - \hat{q}_{0,\tau}$

$$\hat{q}_{j,\tau} = \arg\min_q \sum_{i=1}^{n} \hat{\omega}_{j,i} \rho_\tau (Y_i - q), \qquad \hat{\omega}_{1,i} = \frac{T_i}{n\hat{p}(X_i)} \qquad \hat{\omega}_{0,i} = \frac{1 - T_i}{n(1 - \hat{p}(X_i))}$$

- Indeed, this estimator is the best semiparametric estimator (Firpo (2005))

- Note that this follows the same procedure as with the ATE – using IPW to identify the quantiles of each underyling distribution

Comparing distributions



Last example

- Ok so what? While estimating the range of effects is interesting, it is
  - noisier
  - challenging to interpret in an intuitive way

- However, if you have underyling theory that has implications for distribiution, quantile regression is the empirical approach for you

- A nice paper highlighting this point: Bitler, Gelbach and Hoynes (2006)

Bitler, Gelbach and Hoynes (2006)

- Comparing the “Jobs First” and AFDC programs in CT

- Key difference between programs was significantly more generous tax treatment in Jobs First (shifting budget line out)

- How does implementation of policy affect income?

- Implications:
  1. Very bottom earners will have no effect
  2. Very top is zero or negative
  3. In between, JF should have positive effect

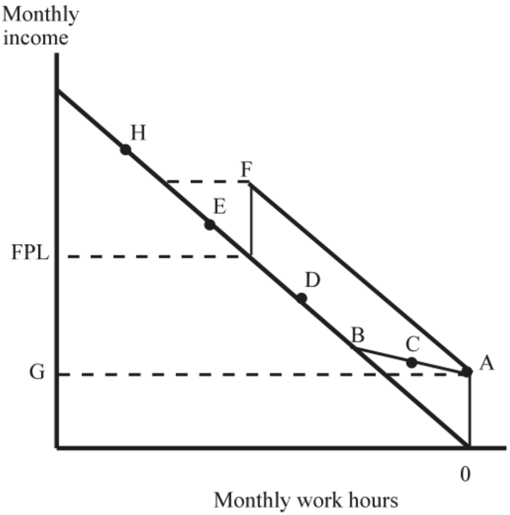
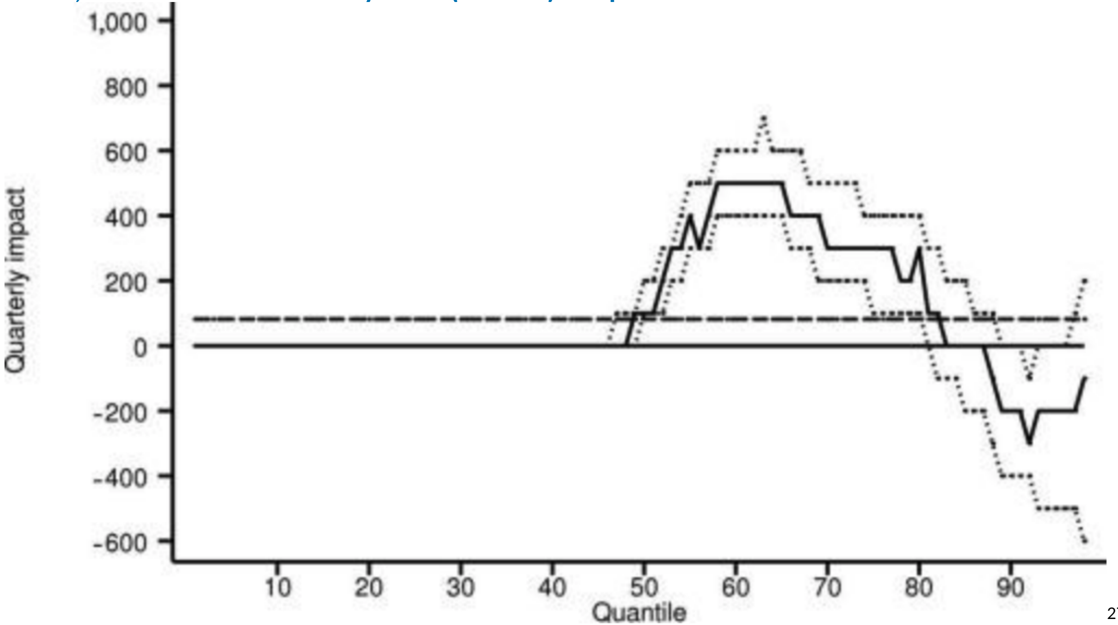


FIGURE 1. STYLIZED CONNECTICUT BUDGET CONSTRAINT UNDER AFDC AND JOBS FIRST

Bitler, Gelbach and Hoynes (2006) impact on income



The Upsides of Quantile Regression

- Allows you to characterize the distribution
  - When considering welfare, can be very useful
  - This can be important for more complicated models
  - We will revisit when considering hierarchical models

- Robust to:
  - issues of functional form (e.g. log)
  - censoring/truncation
  - outliers

- Worth using in your toolkit along with OLS in many applications
  - Easy to plug in
  - `qreg` in Stata and `quantreg` in `R`

Issues with Quantile Regression

- Not that fast– linear programming problem and standard errors
- Not additively combinable. E.g., if $Y = Y_1 + Y_2$, not possible to decompose and have the effects be comparable.
  - This can create issues with fixed effects
- Can be challenging to interpet as structural parameters
  - Shift focus from parameters to understading how the shape of the distribution changes with changes in covariates
  - Change your estimand!
- Standard errors can be wonky – asymptotic theory is less developed, although clustering finally exists! (See Hagemann (2017), also Parente and Santos Silva (2016))