Debiased Inference on Heterogeneous Quantile Treatment Effects with Regression Rank-Scores

Alexander Giessing[1]

Department of Statistics
University of Washington

August, 2023

[1]joint work with Jingshen Wang, Division of Biostatistics, UC Berkeley

Problem Statement

The same treatment may affect different individuals differently – how can we conduct efficient inference on the heterogeneous TE?

- **Personalized medicine**

  Which radiation therapy is most appropriate for a cancer patient?

- **Targeted advertisement**

  What is the best ad to play on Youtube given my subscriptions?

- **Fairness in machine learning/ subgroup analysis**

  Does early screening in college applications discriminate against certain minorities?

Modelling Framework

- **Potential Outcome Framework**
  - Treatment indicator: $D \in \{0, 1\}$.
  - Unobserved potential outcomes: $Y(0), Y(1) \in \mathbb{R}$.
  - Observed outcome: $Y = DY(1) + (1 - D)Y(0)$.
  - High-dim covariates: $X \in \mathbb{R}^p$ with $p \gg n$.

- **Heterogeneous Quantile Treatment Effect (HQTE)**

$$\delta(\tau; z) := Q_{Y(1)}(\tau; z) - Q_{Y(0)}(\tau; z),$$

with $Q_{Y(d)}(\tau|z)$ τth conditional quantile of $Y \mid X = z$ (Doksum, 1986).

- **Identifiability of the HQTE**
  - Unconfoundedness assumption
  - Sparse linear quantile regression function: $Q_{Y(d)}(\tau; z) = z'\theta_d(\tau)$ and $\sup_{\tau \in \mathcal{T}} \|\theta_d(\tau)\|_0 \ll p \wedge n$ for all $\tau \in \mathcal{T} \subset (0, 1)$.

Why estimate a high-dim linear HQTE curve?

$$\delta(\tau; z) := z'\theta_1(\tau) - z'\theta_0(\tau), \quad \tau \in \mathcal{T} \subset (0,1)$$

- **dense $z \in \mathbb{R}^p$, uniform in $\tau \in \mathcal{T}$**
  - heterogeneity across different quantiles τ
  - uniform confidence bands for HQTE curve
  - maximal TE $\sup_{\tau \in \mathcal{T}} \delta(\tau; z)$ (subgroup analysis)
  - integrated TE $\int_{\mathcal{T}} \delta(\tau; z) d\tau$ (robust HQTE)
- **sparse $z \in \mathbb{R}^p$**

  differential TE between sub-populations characterized by a few pre-treatment covariates (e.g. age, race, gender, etc.)
- **unconfoundedness assumption** is more plausible when X is a rich set of covariates (aka "high-dimensional") (Rubin, 2009)

Preliminary thoughts about estimating the HQTE curve

$$\delta(\tau; z) := z'\theta_1(\tau) - z'\theta_0(\tau), \quad \tau \in \mathcal{T} \subset (0, 1)$$

- $\theta_d(\tau) \in \mathbb{R}^p$ **is high-dimensional**

  $\Longrightarrow$ we have to use some regularized estimator which is biased

- $z \in \mathbb{R}^p$ **may be dense**

  $\Longrightarrow$ if $z \notin \mathrm{span}(X_1, \dots, X_n)$ there is an out-of-sample prediction bias

Before we can discuss efficient estimation of the HQTE, we need to think about debiasing procedures!

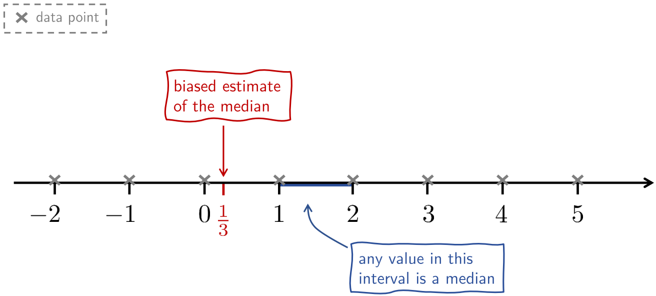
Outline

1. **Heuristics: Efficient Debiasing of Conditional Quantiles**

2. **Theory: Properties of the Rank-Score Debiasing Algorithm**

3. **Illustration: Differential Effect of Statin Usage in Alzheimer's Patients**

How to correct biased quantile estimates? (1)

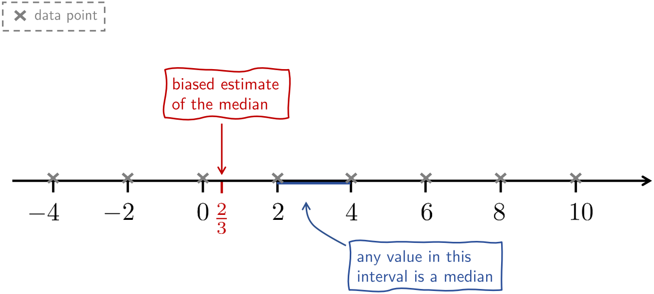


$$\widehat{Q}_{1/2}^{\text{debiased}} = \widehat{Q}_{1/2} + \frac{\#\{\text{data points} > \widehat{Q}_{1/2}\}}{2} - \frac{\#\{\text{data points} \leq \widehat{Q}_{1/2}\}}{2}$$

$$= \frac{1}{3} + \frac{5}{2} - \frac{3}{2} = \frac{4}{3}$$

$$\stackrel{?}{\Longrightarrow} \quad \widehat{Q}_{\tau}^{\text{debiased}} = \widehat{Q}_{\tau} + \sum_{i=1}^{n} \left(\tau - \mathbf{1}\{Y_i \leq \widehat{Q}_{\tau}\}\right), \quad \tau \in (0,1)$$

How to correct biased quantile estimates? (2)

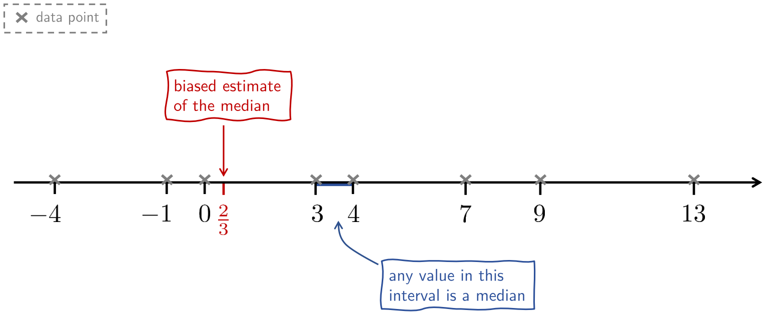


$$\widehat{Q}_{1/2}^{\text{debiased}} = \widehat{Q}_{1/2} + \frac{{\color{red}2} \times \#\{\text{data points} > \widehat{Q}_{1/2}\}}{2} - \frac{{\color{red}2} \times \#\{\text{data points} \leq \widehat{Q}_{1/2}\}}{2}$$

$$= \frac{1}{3} + \frac{{\color{red}2} \times 5}{2} - \frac{{\color{red}2} \times 3}{2} = \frac{8}{3}$$

$$\overset{?}{\Longrightarrow} \quad \widehat{Q}_{\tau}^{\text{debiased}} = \widehat{Q}_{\tau} + {\color{red}\text{scale} \times} \sum_{i=1}^{n} \left(\tau - \mathbf{1}\{Y_i \leq \widehat{Q}_{\tau}\}\right), \quad \tau \in (0,1)$$

How to correct biased quantile estimates? (3)



$$\Longrightarrow \widehat{Q}_Y^{\text{debiased}}(\tau) = \widehat{Q}_Y(\tau) + \text{scale} \times \sum_{i=1}^{n} \text{weight}_i \times \left(\tau - \mathbf{1}\{Y_i \leq \widehat{Q}_Y(\tau)\}\right)$$

How to adapt this idea to the conditional quantile estimate $\widehat{Q}_Y(\tau|z)$?

Debiasing conditional quantile estimates

"Rank-Score Balancing Weights"

Only the signs (rank-scores) of residuals not their magnitude are informative in quantile regression.

$$\widehat{Q}_Y^{\text{debias}}(\tau|z) := z'\hat{\theta}(\tau) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i \frac{\tau - \mathbf{1}\{Y_i \le X_i'\hat{\theta}(\tau)\}}{\hat{f}_i(\tau)}$$

Rank-scores are dimensionless; to compare them to the leading term, put them on roughly the same scale.

$\hat{\theta}(\tau)$ – solution to ℓ_1-penalized QR program

$\hat{f}_i(\tau)$ – an estimate of $f_{Y|X}(X_i'\theta(\tau)|X_i)$

Balancing bias and variance to find the optimal w

Sum of independent and centered random variables, asymp. normal with variance $\tau(1-\tau)\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n w_i^2 f_{Y|X}^{-2}(X_i'\theta(\tau)|X_i)\right]$

$$\widehat{Q}_Y^{\text{debias}}(\tau|z) - Q_Y(\tau|z)$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^n w_i \frac{\tau - \mathbf{1}\{Y_i \leq X_i'\theta(\tau)\}}{f_{Y|X}(X_i'\theta(\tau)|X_i)}$$

Minimize variance

$$+ \left(\mathbb{E}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n w_i X_i\right] - z\right)'(\theta(\tau) - \hat{\theta}(\tau))$$

Bias term, bounded by $\left\|\mathbb{E}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n w_i X_i\right] - z\right\|_\infty \|\theta(\tau) - \hat{\theta}(\tau)\|_1$

$$+ r_n(z, w)$$

Control bias

Rank-score debiasing algorithm

1. Compute ℓ_1-penalized quantile regression vectors:

$$\hat{\theta}_d(\tau) \in \arg\min_{\theta \in \mathbb{R}^p} \left\{ \sum_{i:D_i=d} \rho_\tau(Y_i - X_i'\theta) + \lambda_d \sum_{j=1}^{p} |\theta_j| \right\}.$$

2. Compute rank-score debiasing weights:

$$\widehat{w}(\tau) \in \arg\min_{w \in \mathbb{R}^n} \left\{ \sum_{i=1}^{n} w_i^2 \hat{f}_i^{-2}(\tau) : \left\| z - \frac{1}{\sqrt{n}} \sum_{i:D_i=d} w_i X_i \right\|_\infty \leq \frac{\gamma_d}{n}, d \in \{0,1\} \right\},$$

where $\hat{f}_i(\tau)$ is an estimate of $f_{Y(d)|X}(X_i'\theta_d(\tau)|X_i)$.

3. Construct rank-score debiased estimates:

$$\widehat{Q}^{\text{rank}}_{Y(d)}(\tau; z) := z'\hat{\theta}_d(\tau) + \frac{1}{\sqrt{n}} \sum_{i:D_i=d} \widehat{w}_i(\tau) \frac{\tau - \mathbf{1}\{Y_i \leq X_i'\hat{\theta}_d(\tau)\}}{\hat{f}_i(\tau)},$$

$$\widehat{\delta}^{\text{rank}}(\tau; z) := \widehat{Q}^{\text{rank}}_{Y(1)}(\tau; z) - \widehat{Q}^{\text{rank}}_{Y(0)}(\tau; z).$$

Outline

1. Heuristics: Efficient Debiasing of Conditional Quantiles

2. Theory: Properties of the Rank-Score Debiasing Algorithm

3. Illustration: Differential Effect of Statin Usage in Alzheimer's Patients

(Un)expected statistical properties

- **Consistent and asymptotically unbiased**

  intuition: "box-constraint" in step 2 of the algorithm balances covariates $z, X_1, \ldots, X_n$ and controls out-of-sample prediction bias

- **Asymptotically normal/ weakly convergent to a Gaussian process**

  intuition: leading term of the "Taylor-like expansion" with fixed weights w is a sum of centered i.i.d. random variables

- **Semi-parametric efficient**

  step 2 of the algorithm minimizes the empirical sample version of the asymptotic variance of the leading term of the "Taylor-like expansion"

- **Simple consistent estimate of asymptotic covariance function**

  optimal value of the objective function in step 2 of the algorithm is a consistent estimate of the asymptotic covariance function

The two main challenges in the theoretical analysis

$$\widehat{Q}^{\mathsf{rank}}_{Y(d)}(\tau; z) = z'\hat{\theta}_d(\tau) + \frac{1}{\sqrt{n}} \sum_{i:D_i=d} \widehat{w}_{d,i}(\tau) \frac{\tau - \mathbf{1}\{Y_i \leq X_i'\hat{\theta}_d(\tau)\}}{\hat{f}_i(\tau)}$$

The two main challenges in the theoretical analysis

$$\widehat{Q}^{\mathsf{rank}}_{Y(d)}(\tau; z) = z'\hat{\theta}_d(\tau) + \frac{1}{\sqrt{n}} \sum_{i:D_i=d} \widehat{w}_{d,i}(\tau) \frac{\tau - \mathbf{1}\{Y_i \leq X_i'\hat{\theta}_d(\tau)\}}{\hat{f}_i(\tau)}$$

- consistent estimates $\hat{f}_i(\tau)$ of the conditional densities $f_{Y(d)|X}(X_i'\theta_d(\tau)|X_i)$

The two main challenges in the theoretical analysis

$$\widehat{Q}^{\mathsf{rank}}_{Y(d)}(\tau; z) = z'\hat{\theta}_d(\tau) + \frac{1}{\sqrt{n}} \sum_{i:D_i=d} \widehat{w}_{d,i}(\tau) \frac{\tau - \mathbf{1}\{Y_i \leq X_i'\hat{\theta}_d(\tau)\}}{\hat{f}_i(\tau)}$$

- consistent estimates $\hat{f}_i(\tau)$ of the conditional densities $f_{Y(d)|X}(X_i'\theta_d(\tau)|X_i)$

$\Longrightarrow$ Koenker's nonparametric density estimator

The two main challenges in the theoretical analysis

$$\widehat{Q}^{\text{rank}}_{Y(d)}(\tau; z) = z'\hat{\theta}_d(\tau) + \frac{1}{\sqrt{n}} \sum_{i:D_i=d} \widehat{w}_{d,i}(\tau) \frac{\tau - \mathbf{1}\{Y_i \leq X_i'\hat{\theta}_d(\tau)\}}{\hat{f}_i(\tau)}$$

- consistent estimates $\hat{f}_i(\tau)$ of the conditional densities $f_{Y(d)|X}(X_i'\theta_d(\tau)|X_i)$

$\Longrightarrow$ Koenker's nonparametric density estimator

$\Longrightarrow$ other density estimators?

The two main challenges in the theoretical analysis

$$\widehat{Q}^{\text{rank}}_{Y(d)}(\tau; z) = z'\hat{\theta}_d(\tau) + \frac{1}{\sqrt{n}} \sum_{i:D_i=d} \widehat{w}_{d,i}(\tau) \frac{\tau - \mathbf{1}\{Y_i \leq X_i'\hat{\theta}_d(\tau)\}}{\hat{f}_i(\tau)}$$

- consistent estimates $\hat{f}_i(\tau)$ of the conditional densities $f_{Y(d)|X}(X_i'\theta_d(\tau)|X_i)$

$\Longrightarrow$ Koenker's nonparametric density estimator

$\Longrightarrow$ other density estimators?

- rank-score balanced estimator with optimal weights $\widehat{w}_{d,i}(\tau)$ does not satisfy a "Taylor-like expansion"

The two main challenges in the theoretical analysis

$$\widehat{Q}^{\textsf{rank}}_{Y(d)}(\tau; z) = z'\hat{\theta}_d(\tau) + \frac{1}{\sqrt{n}} \sum_{i:D_i=d} \widehat{w}_{d,i}(\tau) \frac{\tau - \mathbf{1}\{Y_i \leq X_i'\hat{\theta}_d(\tau)\}}{\hat{f}_i(\tau)}$$

- consistent estimates $\hat{f}_i(\tau)$ of the conditional densities $f_{Y(d)|X}(X_i'\theta_d(\tau)|X_i)$

  $\Longrightarrow$ Koenker's nonparametric density estimator

  $\Longrightarrow$ other density estimators?

- rank-score balanced estimator with optimal weights $\widehat{w}_{d,i}(\tau)$ does not satisfy a "Taylor-like expansion"

  $\Longrightarrow$ consider the dual of the rank-score debiasing program

The two main challenges in the theoretical analysis

$$\widehat{Q}^{\mathsf{rank}}_{Y(d)}(\tau; z) = z'\hat{\theta}_d(\tau) - \frac{1}{\sqrt{n}} \sum_{i:D_i=d} \frac{\hat{f}_i^2(\tau)}{2\sqrt{n}} X_i'\hat{v}_d(\tau) \frac{\tau - \mathbf{1}\{Y_i \leq X_i'\hat{\theta}_d(\tau)\}}{\hat{f}_i(\tau)}$$

- consistent estimates $\hat{f}_i(\tau)$ of the conditional densities $f_{Y(d)|X}(X_i'\theta_d(\tau)|X_i)$

  $\Longrightarrow$ Koenker's nonparametric density estimator

  $\Longrightarrow$ other density estimators?

- rank-score balanced estimator with optimal weights $\widehat{w}_{d,i}(\tau)$ does not satisfy a "Taylor-like expansion"

  $\Longrightarrow$ consider the dual of the rank-score debiasing program

  $\Longrightarrow$ $\widehat{Q}^{\mathsf{rank}}_{Y(d)}(\tau; z)$ is an affine function of the dual solution $\hat{v}_d(\tau)$

The two main challenges in the theoretical analysis

$$\widehat{Q}^{\text{rank}}_{Y(d)}(\tau; z) = z'\hat{\theta}_d(\tau) - \left(\frac{1}{2n} \sum_{i: D_i = d} \hat{f}_i(\tau) \big(\tau - \mathbf{1}\{Y_i \leq X_i'\hat{\theta}_d(\tau)\}\big) X_i \right)' \hat{v}_d(\tau)$$

- consistent estimates $\hat{f}_i(\tau)$ of the conditional densities $f_{Y(d)|X}(X_i'\theta_d(\tau)|X_i)$

$\Longrightarrow$ Koenker's nonparametric density estimator

$\Longrightarrow$ other density estimators?

- rank-score balanced estimator with optimal weights $\widehat{w}_{d,i}(\tau)$ does not satisfy a "Taylor-like expansion"

$\Longrightarrow$ consider the dual of the rank-score debiasing program

$\Longrightarrow$ $\widehat{Q}^{\text{rank}}_{Y(d)}(\tau; z)$ is an affine function of the dual solution $\hat{v}_d(\tau)$

$\Longrightarrow$ $\widehat{Q}^{\text{rank}}_{Y(d)}(\tau; z)$ is amenable to high-dim empirical process theory

Rank-score debiased estimate is semi-parametric efficient

Theorem

Under regularity conditions,

$$\sqrt{n}\left(\widehat{Q}^{\text{rank}}_{Y(d)}(\tau|z) - Q_{Y(d)}(\tau|z)\right) \rightsquigarrow \mathcal{N}\left(0, \tau(1-\tau)z' D^{-1}_{2,d}(\tau) z\right).$$

$D_{k,d}(\tau)$ – *denotes* $\mathrm{E}[f_d^k(\tau) X X' \mathbf{1}\{D = d\}]$, $k = 0, 1, 2$
$f_d(\tau)$ – *shorthand for* $f_{Y(d)|X}(X'\beta_d(\tau)|X)$

- Same variance as the weighted QR program (Koenker and Zhao, 1994)

$$\widetilde{\theta}_d(\tau) \in \arg\min_{\theta \in \mathbb{R}^p} \sum_{i: D_i = d} \hat{f}_i^{-1}(\tau) \rho_\tau(Y_i - X_i'\theta).$$

- More efficient than the standard QR estimator in the sense that

$$z' D^{-1}_{2,d}(\tau) z \leq z' D^{-1}_{1,d}(\tau) D_{0,d}(\tau) D^{-1}_{1,d}(\tau) z.$$

- Attains semi-parametric efficiency bound (Newey and Powell, 1990).

Rank-score debiased estimate is semi-parametric efficient

Theorem

Under regularity conditions,

$$\sqrt{n}\left(\widehat{Q}^{\text{rank}}_{Y(d)}(\tau|z) - Q_{Y(d)}(\tau|z)\right) \rightsquigarrow \mathcal{N}\left(0, \tau(1-\tau)z' D^{-1}_{2,d}(\tau) z\right).$$

$D_{k,d}(\tau)$ – *denotes* $\mathrm{E}[f_d^k(\tau)XX'\mathbf{1}\{D=d\}]$, $k = 0, 1, 2$
$f_d(\tau)$ – *shorthand for* $f_{Y(d)|X}(X'\beta_d(\tau)|X)$

Theorem

Under regularity conditions,

$$\sqrt{n}\left(\widehat{\delta}^{\text{rank}}(\tau; z) - \delta(\tau; z)\right) \rightsquigarrow N\left(0, \sigma^2(\tau; z)\right),$$

where

$$\sigma^2(\tau; z) = \tau(1-\tau)z'\left[D^{-1}_{2,1}(\tau) + D^{-1}_{2,0}(\tau)\right] z.$$

Asymptotic variance can be estimated easily

For $\tau \in \mathcal{T}$ define

$$\widehat{\sigma}_n^2(\tau; z) := \tau(1-\tau) \sum_{i=1}^{n} \widehat{w}_i^2(\tau) \hat{f}_i^{-2}(\tau).$$

Uniformly consistent estimate of covariance

Under regularity conditions,

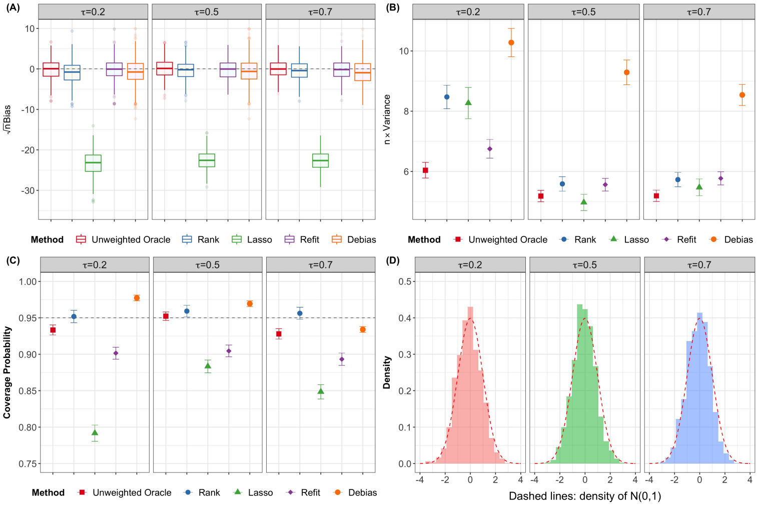
$$\sup_{\tau \in \mathcal{T}} \left| \widehat{\sigma}_n^2(\tau; z) - \sigma^2(\tau; z) \right| = o_p(1).$$

- By-product of estimating the rank-score balancing weights
- We don't have to estimate the inverse of a high-dim. matrix

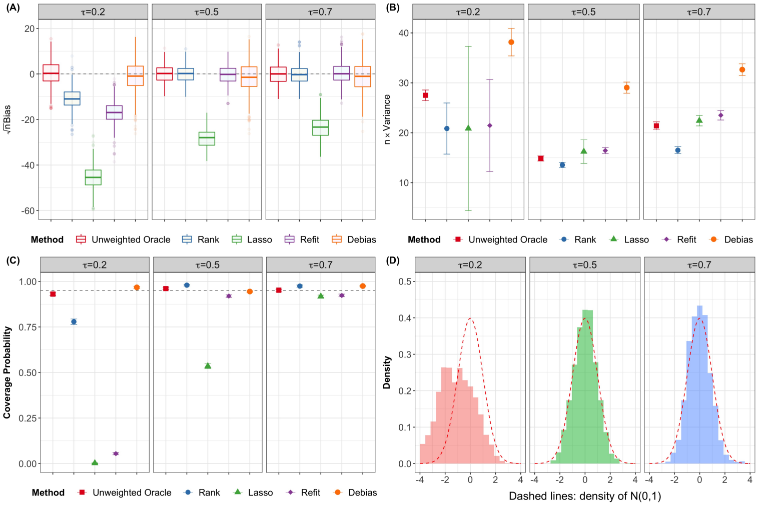
Supporting Monte Carlo Experiments

- **We compare the following estimators:**
  - Unweighted Oracle: Estimator based on covariates in support of θ_d only
  - Rank: Our rank-score debiased estimator
  - Lasso: ℓ_1-penalized quantile regression estimator
  - Refit: Refit based on support of ℓ_1-penalized quantile regression estimator
  - Debias: Estimator using debiased ℓ_1-penalized quantile regression estimate by Zhao et al. (2019)
- **We report (based on 2,000 MC samples):**
  - $\sqrt{n} \times \mathrm{Bias}$
  - $n \times \mathrm{Variance}$
  - 95% Coverage Probability
  - histogram of standardized HQTE

Homoscedastic Design



Heteroscedastic Design



Outline

1. Heuristics: Efficient Debiasing of Conditional Quantiles

2. Theory: Properties of the Rank-Score Debiasing Algorithm

3. Illustration: Differential Effect of Statin Usage in Alzheimer's Patients

Basic Scientific Background

Goal:

Estimate the heterogeneous effect of statin usage on lowering the Low-Density-Lipoprotein Cholesterol (LDL-c) concentration levels in Alzheimer's disease (AD) patients.

Relevance:

- Elevated concentration of LDL-c is considered a risk factor for AD.
- Treating AD patients with statin to reduce their LDL-c concentration appears to slow down progression of AD.

Heterogeneity:

Lifestyle patterns (i.e. diets, levels of physical activity, alcohol consumption, and smoking status) affect LDL-c concentration levels.

Study Design

Subset of UK Biobank data set

- 3713 patients with Alzheimer's disease (and AD proxies), older than 65yrs, no missing covariates, and no cholesterol medication history
- To account for genetic pleiotropy and linkage disequilibrium we include 637 SNPs and lifestyle factors associated with LDL cholesterol.
- To eliminate (some) confounders we do not consider statin usage but the functionally equivalent genetic variant rs12916-T; 3150 subjects carry, 563 subjects don't carry this variant.

Does the effect of a "healthy lifestyle" (defined as a healthy diet, physical activities, and reduced smoking) on lowering the LDL-c concentration differ in control and treatment group?

Does the effect of statin usage on lowering the LDL-c concentration differ between Alzheimer's patients with different lifestyles?

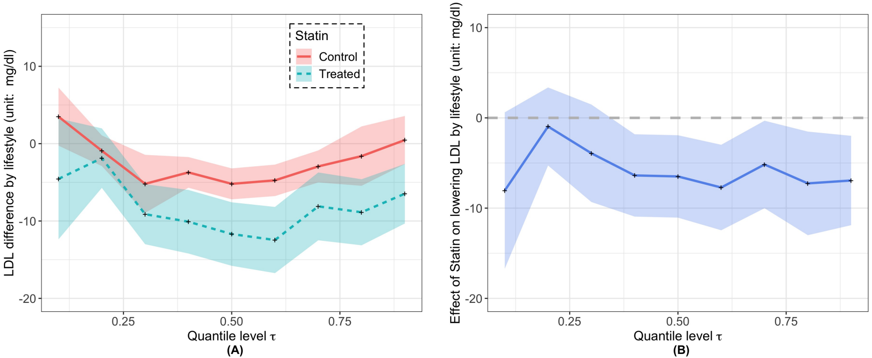
HQTE Regression Model

- Y – LDL-c concentration in mg/dL
- X_1 – intercept
- $X_2, \ldots, X_{18}$ – lifestyle patterns
- X_{19} – gender
- $X_{20}, \ldots, X_{637}$ – SNPs associated with the LDL-c concentration
- Differential effect of statin usage on LDL-c concentration

$$\hat{\delta}^{\text{rank}}(\tau; z) := \widehat{Q}_1^{\text{rank}}(\tau; z) - \widehat{Q}_0^{\text{rank}}(\tau; z),$$

where $z = (0, 0, \underbrace{1, \ldots, 1}_{8}, \underbrace{-1, \ldots, -1}_{6}, 0, \ldots, 0)' \in \mathbb{R}^{637}$.

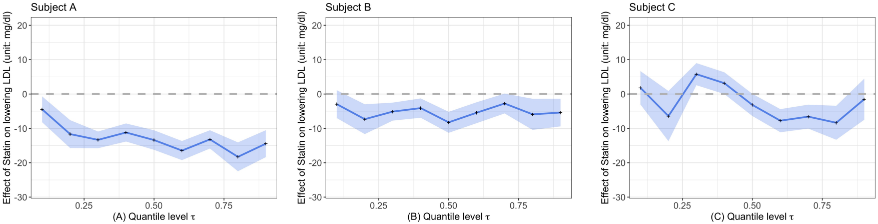
Differential HQTE of statin usage on LDL-c concentration



(A) LDL-c plasma concentration for the treated and control group. (B) Differential HQTE of statin usage on LDL-c concentration by healthy lifestyle. Shaded areas are uniform 95% confidence bands.

Illustrative Individual HQTEs

(subjects characterized by individual z's)



Heterogeneous quantile treatment effects of statin usage for three subjects. Shaded areas are uniform 95% confidence bands.

Summary

- Conditional quantile regression is a flexible semi-parametric framework to model heterogeneous treatment effects.
- Rank-score debiasing removes shrinkage bias and yields a semi-parametric efficient estimator.
- Our methodology can be motivated as either bias-variance trade-off or Neyman orthogonalization.
- The general principle is applicable beyond conditional quantile regression.

Acknowledgment

Paper was published yesterday:

Giessing, A. and Wang, J. (2023). Debiased Inference on Heterogeneous Quantile Treatment Effects with Regression Rank-Scores. *JRSSB.*

This talk is based upon work supported by the National Science Foundation under Grant NSF DMS-2310578.