Quantile Regression in the Post GWAS Secondary Phenotype Analysis

Ying Wei [1]

[1] Department of Biostatistics
Columbia University

Table of Contents

Genome Wide Association Study(GWAS)

- Genome Wide Association Study has been widely used to study how individual genetic variants associate with diseases of interest.

- A typical GWAS employs a case-control design.

  Cases — random sample from disease population
  Controls — random sample from disease-free population

0 1 1 1 1 1 0 1 0 2 1 2 2 0 1 0 0 0 1 1	Control
2 0 1 1 1 2 0 0 0 1 0 1 1 0 1 1 0 1 0 0	Control
2 0 1 2 2 0 1 2 1 0 0 1 1 0 1 0 0 1 1 1	Control
1 2 1 1 2 1 1 1 1 0 1 1 1 0 0 2 2 2 0 2	Control
1 1 2 1 0 1 2 1 1 1 1 2 1 2 1 2 1 2 1 1	Case
2 2 1 2 0 1 0 0 0 1 2 2 1 2 1 2 1 0 2 1	Case
0 1 1 0 0 2 1 0 0 2 1 1 1 2 1 1 2 0 1 0	Case
0 1 1 0 0 1 0 2 2 1 1 1 1 2 0 1 2 1 1 2	Case

GWAS

- In past decade, Genomewide Association Study(GWAS) made remarkable progress in our understanding of the role of genetic variation in complex human diseases.
- According to GWAS catalog, https://www.ebi.ac.uk/gwas/home, over 3000 loci were discovered from 2854 publications. (33674 unique SNP-trait associations)

Secondary Phenotype Analysis in GWAS

- Besides the primary disease status (D) and genetic variants (X), GWAS data often include rich information on additional phenotypes (Y), e.g. BMI, blood pressure, cholesterol level. They often are characteristics of the diseases.

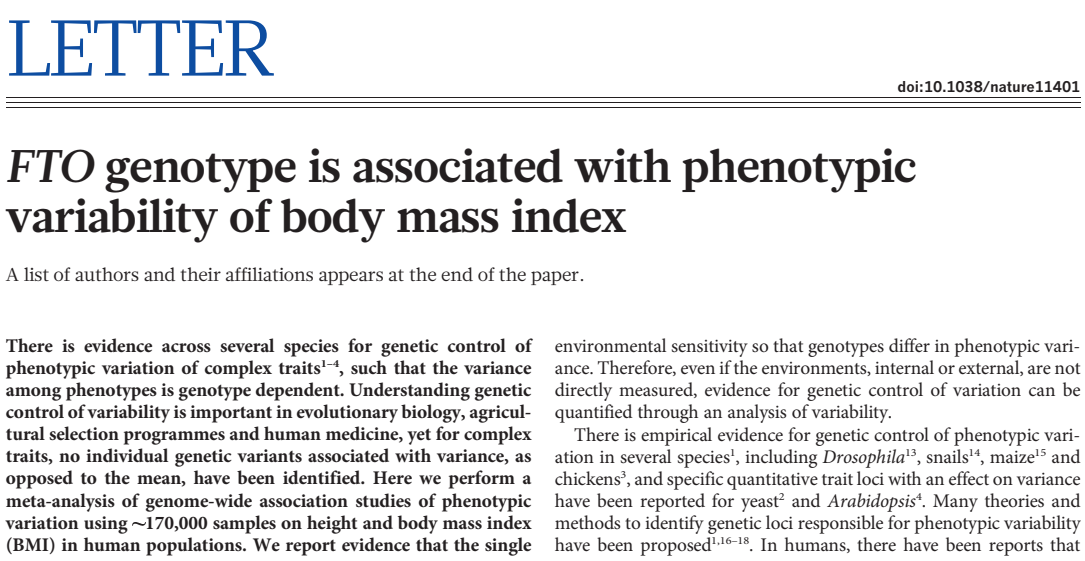
- Analyzubg the genetic association with these secondary phenotypes is an important way to understand underlying biological mechanisms.

  Primary analysis: $X \Rightarrow D$ Secondary analysis $X \Rightarrow Y$

Table of Contents

Why Quantile Analysis?

- Most existing strategies are to identify genetic variants that influence the mean of the phenotypes.
- Genetic effects are more complex than mean-level associations.



LETTER

doi:10.1038/nature11401

FTO genotype is associated with phenotypic variability of body mass index

A list of authors and their affiliations appears at the end of the paper.

There is evidence across several species for genetic control of phenotypic variation of complex traits[1–4], such that the variance among phenotypes is genotype dependent. Understanding genetic control of variability is important in evolutionary biology, agricultural selection programmes and human medicine, yet for complex traits, no individual genetic variants associated with variance, as opposed to the mean, have been identified. Here we perform a meta-analysis of genome-wide association studies of phenotypic variation using ~170,000 samples on height and body mass index (BMI) in human populations. We report evidence that the single

environmental sensitivity so that genotypes differ in phenotypic variance. Therefore, even if the environments, internal or external, are not directly measured, evidence for genetic control of variation can be quantified through an analysis of variability.

There is empirical evidence for genetic control of phenotypic variation in several species[1], including *Drosophila*[13], snails[14], maize[15] and chickens[3], and specific quantitative trait loci with an effect on variance have been reported for yeast[2] and *Arabidopsis*[4]. Many theories and methods to identify genetic loci responsible for phenotypic variability have been proposed[1,16–18]. In humans, there have been reports that

- We would like to consider higher-level associations by using Quantile Regression (Koenker and Bassett, 1978).

Quantile Regression

- Quantile effect

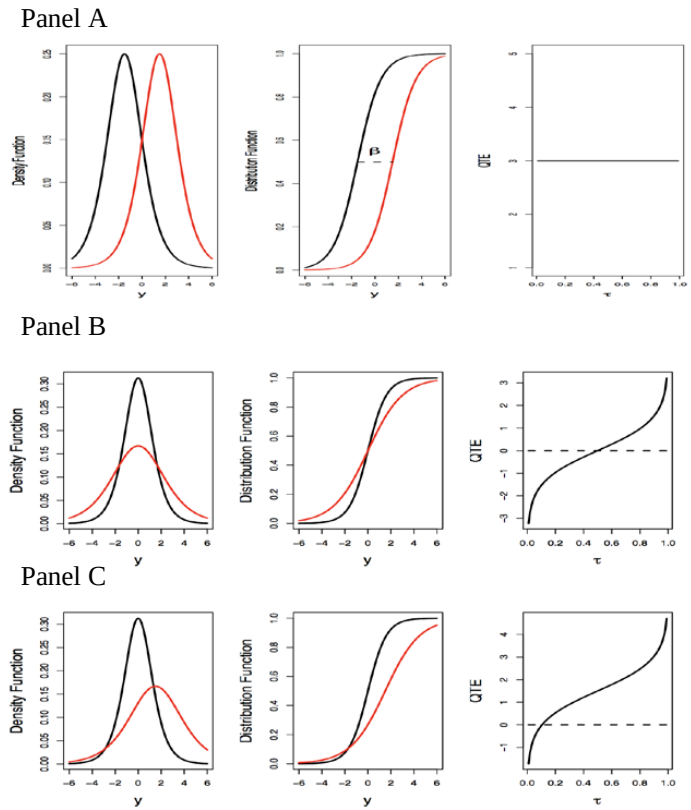
$$Q_Y(\tau \mid X=1,Z) - Q_Y(\tau \mid X=0,Z)$$

- Linear quantile model

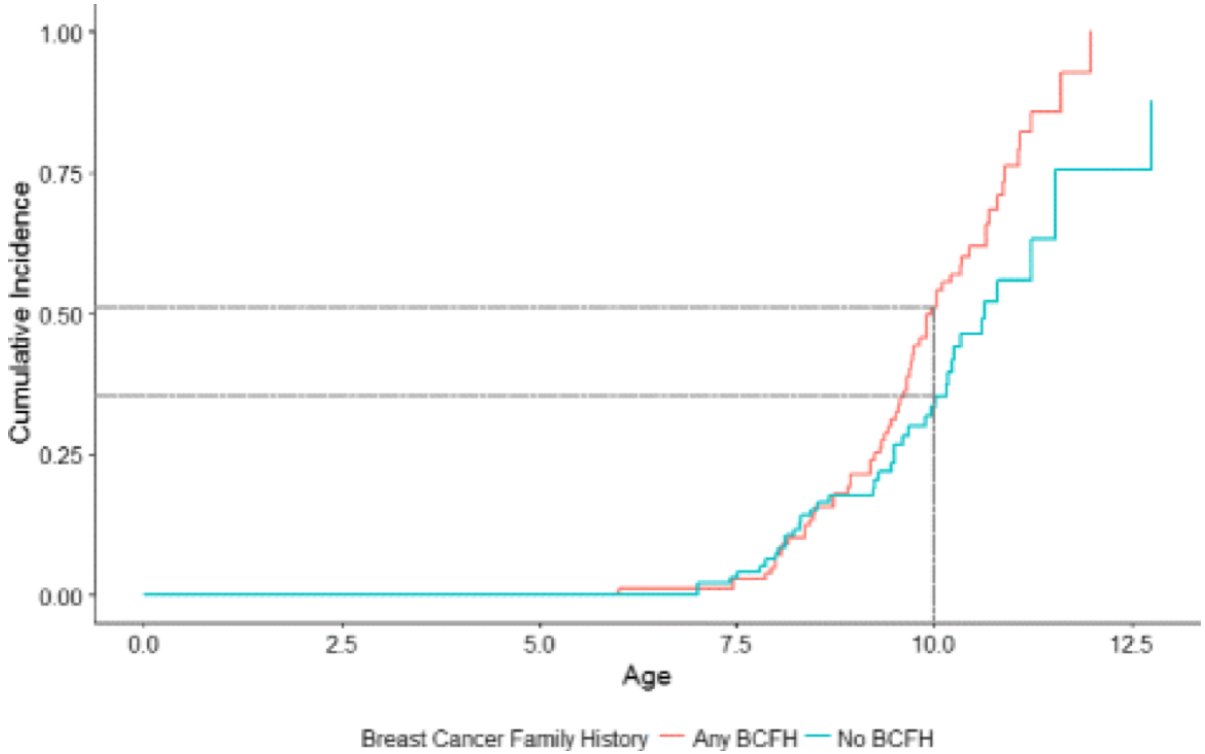
$$Q_Y(\tau \mid X,Z) = X\beta(\tau) + Z^\top \gamma(\tau)$$

- Capture the genetic effects across the entire distribution
- $\beta(\tau)$ can be consistently estimated through quantile regression

Figure 1. The quantile effect under location, scale and location-scale models.



Quantile analysis



- Terry et al. *Breast Cancer Research* 2017

Quantile model for secondary phenotypes

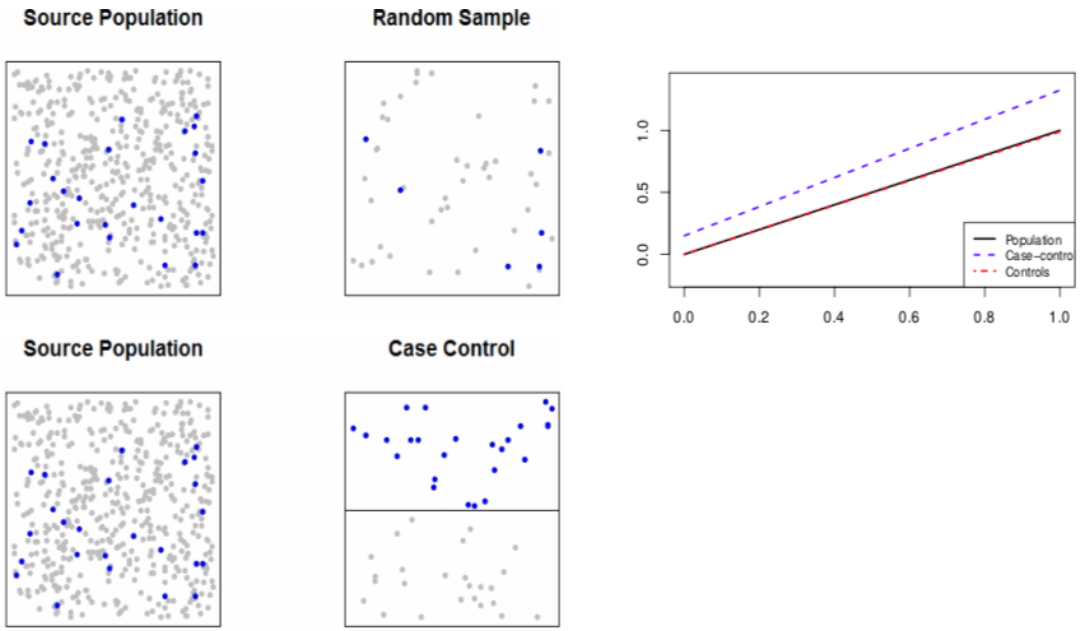
- A case-control GWAS sample: n_1 cases denoted by $\{\mathbf{x}_i, y_i, d_i = 1, \boldsymbol{z}_i\}_{i=1,2,...,n_1}$, and n_2 controls denoted by $\{\mathbf{x}_i, y_i, d_i = 0, \boldsymbol{z}_i\}_{i=n_1+1,...,n}$
  - d_i is the binary indicator for cases;
  - y_i is a continuous phenotype of interest
  - $\mathbf{x}_i$ is a p-dimensional vector of SNP
  - $\boldsymbol{z}_i$ is a q-dimensional controlling variables
- Quantile model in general population

$$Q_\tau(y_i|\mathbf{x}_i) = \mathbf{x}_i^\top \boldsymbol{\beta}_{0,\tau} + \boldsymbol{z}_i^\top \boldsymbol{\gamma}_{0,\tau}.$$

- Goal: to estimate $\boldsymbol{\beta}_{0,\tau}$ consistently using a case-control sample.

Major statistical issues in Secondary (Quantile) Analysis

- Due to the case-control sampling, $(\mathbf{x}_i, y_i, \boldsymbol{z}_i)$ are no longer representative of the general population.
- Ignoring this data structure and directly regressing the secondary trait Y against X and Z could lead to substantive bias.



Major statistical issues in Secondary (Quantile) Analysis

- **Controls only** when disease is rare, but it may lose efficiency.
- **Inverse Probability Weighing (IPW)** If the $P(\text{select} \mid X, D)$ is known, we could use an IPW approach.
- A few **likelihood based** proposals have been made to utilize the entire case control sample, including Roeder, Carroll and Lindsay (1996), Lee, McMurchy and Scott (1997), Jiang, Scott and Wild (2006), and Lin and Zeng (2009).
- Those methods cannot be applied to quantile regression directly, since the latter does not assume a parametric likelihood.
- Review two recently developed methods for secondary quantile analysis
  - A weighted estimating equation approach combining observed and "counterfactual" outcomes (Wei, Song, Liu and Ionita-Laza, JASA, 2016)
  - A superpopulation treatment (Liang, Ma, Wei and Carroll, JRSS-B, 2018)

Estimation of quantile regression.

- Mean regression $E(Y|\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}$,

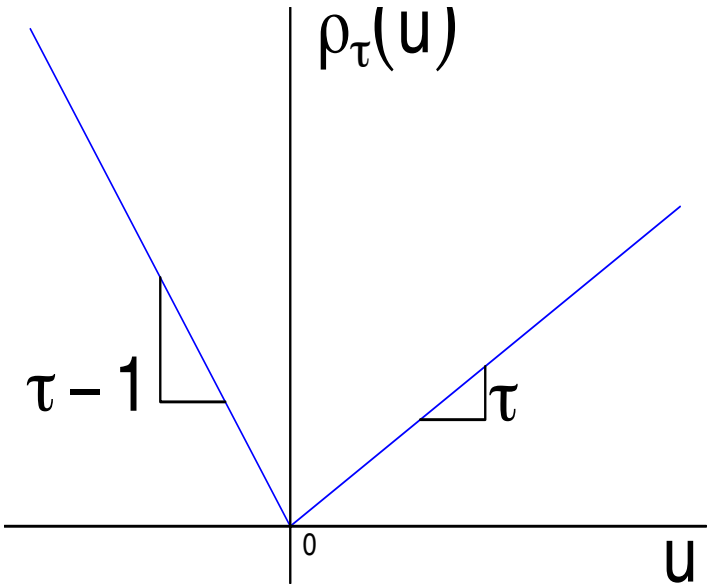
$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{\beta}} E_Y(Y - \mathbf{X}^T\boldsymbol{\beta})^2$$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2.$$

- Assume $Q_\tau(Y|\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}(\tau)$, then

$$\boldsymbol{\beta}(\tau) = \arg\min_{\boldsymbol{\beta}} E_Y(\rho_\tau\{Y - \mathbf{X}^T\boldsymbol{\beta}\})$$

$$\hat{\boldsymbol{\beta}}(\tau) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}\rho_\tau\{y_i - \mathbf{x}_i^T\boldsymbol{\beta}\}.$$



Quantile regression estimating functions

- Estimating function: $\Psi_\tau(\mathbf{X}, Y, \boldsymbol{\beta}) = \left[\tau - I\{Y \leq \mathbf{X}^\top \boldsymbol{\beta}\}\right]\mathbf{X}$
- At the true $\boldsymbol{\beta}_{0,\tau}$,

$$E_Y\left[\Psi_\tau(\mathbf{X}, Y, \boldsymbol{\beta}_{0,\tau}) \mid \mathbf{X}\right] = 0.$$

- With a representative sample, one can obtain a consistent estimate by solving

$$\sum_{i=1}^{n} \Psi_\tau(\mathbf{x}_i, y_i, \boldsymbol{\beta}) = 0$$

New estimating functions with counter-factual observations

- Expand the original estimating functions

$$\begin{aligned} 0 &= E\left[\Psi_\tau(\mathbf{X}, Y, \boldsymbol{\beta}_{0,\tau}) \mid \mathbf{X}\right] \\ &= E_Y\left[\Psi_\tau(\mathbf{X}, Y, \boldsymbol{\beta}_{0,\tau}) \mid X, D=0\right] P(D=0|\mathbf{X}) \\ &\quad + E_Y\left[\Psi_\tau(\mathbf{X}, Y, \boldsymbol{\beta}_{0,\tau}) \mid X, D=1\right] P(D=1|\mathbf{X}) \end{aligned}$$

- Suppose we are able to observe a counter-factual outcomes $\tilde{y}_i$ for each i. Then one could construct unbiased estimation equations following (1) by

$$\sum_{i=1}^{n} \left\{\Psi_\tau(\mathbf{x}_i, y_i, \boldsymbol{\beta}) p(d_i \mid \mathbf{x}_i) + \Psi_\tau(\mathbf{x}_i, \tilde{y}_i, \boldsymbol{\beta}) p(1-d_i \mid \mathbf{x}_i)\right\} = 0,$$

New estimating functions with counter-factual observations

- In reality, those pseudo counter-factual outcomes are unobserved. To get around this, we propose to simulate counter-factual outcomes $\tilde{y}_i$'s from the conditional quantile process

$$\boldsymbol{\beta}_0(\tau \mid d) = \arg\min_{\boldsymbol{\beta}} E_Y[\|\Psi_\tau(Y, \mathbf{X}, \boldsymbol{\beta})\| \mid \mathbf{X}, D = d], \forall \tau \tag{1}$$

  — $\mathbf{x}^\top \boldsymbol{\beta}_0(\tau \mid 0)$ is the conditional quantile function of y given $\mathbf{x}$ among controls,
  – $\mathbf{x}^\top \boldsymbol{\beta}_0(\tau \mid 1)$ defines that among disease population.

Simulating the counter-factual outcomes

- Estimate $\boldsymbol{\beta}_0(\tau \mid 0)$ and $\mathbf{x}^\top \boldsymbol{\beta}_0(\tau \mid 1)$ from cases and controls separately on a fine grid of τ's.
- For the ith subject, $i = 1, \ldots, n$, we simulate its pseudo outcome $\widetilde{y}_i$ by

$$\widehat{\widetilde{y}}_i = \mathbf{x}_i^\top \widehat{\beta}(u_i \mid 1 - d_i), \quad u_i \sim \mathrm{U}(0,1)$$

  where u_i is a random draw from Uniform $(0,1)$ distribution.
- The sampling estimating equations are then

$$\sum_{i=1}^{n} \Psi_\tau(\mathbf{x}_i, y_i, \boldsymbol{\beta}) p(d_i \mid \mathbf{x}_i) + \Psi_\tau(\mathbf{x}_i, \widehat{\widetilde{y}}_i, \boldsymbol{\beta}) p(1 - d_i \mid \mathbf{x}_i) = 0. \tag{2}$$

Simulating the counter-factual outcomes

- Simulating pseudo outcomes is subject to sampling uncertainty, and brings extra variability into parameter estimation.
- To further stabilize the variance, we suggest to repeat the above simulation procedures m time, and use their average as final estimation.

$$\widehat{\beta}_{n,\tau} = m^{-1} \sum_{\ell=1}^{m} \widehat{\beta}_{n,\tau}^{(\ell)}.$$

Estimating $P(D|\mathbf{X})$

- To estimate the conditional disease probability $P(D|\mathbf{X})$, we assume a logistic model
$$P(D=1|\mathbf{X}) = \exp(\gamma_0 + \mathbf{X}^\top \boldsymbol{\gamma}_1)/\{1+\exp(\gamma_0 + \mathbf{X}^\top \boldsymbol{\gamma}_1)\}$$
- Or, use the derived model from the primary analysis.
- The slope $\boldsymbol{\gamma}_1$ can be consistently estimated by regressing d_i over $\mathbf{x}_i$ using the case control sample (Prentice and Pyke, 1979).
- The intercept γ_0 can be calculated by solving the equation
$$P_0 = \int_X \exp(\gamma_0 + \mathbf{X}^\top \widehat{\boldsymbol{\gamma}}_1)/\{1+\exp(\gamma_0 + \mathbf{X}^\top \widehat{\boldsymbol{\gamma}}_1) dF_X, \quad (3)$$
where P_0 is the overall disease prevalence, F_X is the distribution of $\mathbf{X}$.

Estimating $p(D|\mathbf{x})$

- In some cases, F_X can be obtained from external resources. Such as single SNP comparison, F_X can be derived from the minor allele frequency (MAF)
- When the joint distribution of $\mathbf{X}$ is hard to obtain, we proposed to approximate γ_0 by solving its sample version,

$$\widehat{\gamma_0} = \arg\min_{\gamma_0} \left(P_0 - \frac{1}{n} \sum_{i=1}^{n} \exp(\gamma_0 + \mathbf{x}_i^\top \widehat{\boldsymbol{\gamma}}_1) / \{1 + \exp(\gamma_0 + \mathbf{x}_i^\top \widehat{\boldsymbol{\gamma}}_1)\} \right)^2 . \quad (4)$$

- P_0 is estimated sample prevalence.

Large sample properties of the proposed estimator

Theorem

Under Assumptions 1-5, for $k_n \to \infty$ *and* $k_n/n \to 0$, *we have*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{n,\tau} - \boldsymbol{\beta}_{n,\tau}) \to N(0, G_0^{-1}\Sigma_0 G_0^{-1}),$$

where $\Sigma_0 = V_1 + m^{-1}V_2 + 2U_1 + \{(m-1)/m\}U_2$

Applications: Part of New York University Bellevue Asthma Registry

- D: 387 asthmatics and 212 healthy controls
- X: one of the 10 SNPs on the Thymic stromal lymphopoietin (TSLP) gene
- Z: a continuous variable derived as the first principal component scores from 213 ancestry informative markers (AIMs) to adjust for population stratification.
- Y: Forced expiatory volume in one second (FEV_1), an important quantitative measure of lung functions. Values of between 80 and 120 are considered normal.
- Model

$$Q_\tau(\text{FEV}_1) = \beta_{0,\tau} + \beta_{1,\tau} X + \beta_{2,\tau} Z, \tag{5}$$

Estimated allelic effects in the mean regression model and quantile regression models at quantile levels of 0.15, 0.25, 0.5, 0.75 and 0.85

	mean*			$\tau = 0.15$		$\tau = 0.25$		$\tau = 0.5$		$\tau = 0.75$	
SNPs	Est.	P-value	Method	Est.	P-value	Est.	P-value	Est.	P-value	Est.	P-value
rs11466743	-8.1	0.009	SICO(10)	-17.2	0.003	-6.7	0.015	-1.9	0.307	-7.5	0.000
			IPW	-18.4	0.103	-3.9	0.733	-2.5	0.638	-6.8	0.216
rs2289278	3.5	0.041	SICO(10)	0.5	0.691	-1.9	0.140	5.8	0.000	1.4	0.125
			IPW	-1.2	0.641	-1.6	0.633	6.0	0.028	1.0	0.631
rs11241090	-4.8	0.042	SICO(10)	-3.7	0.189	-0.3	0.900	-0.7	0.671	-5.9	0.000
			IPW	-3.5	0.678	2.0	0.696	-2.5	0.486	-7.0	0.099

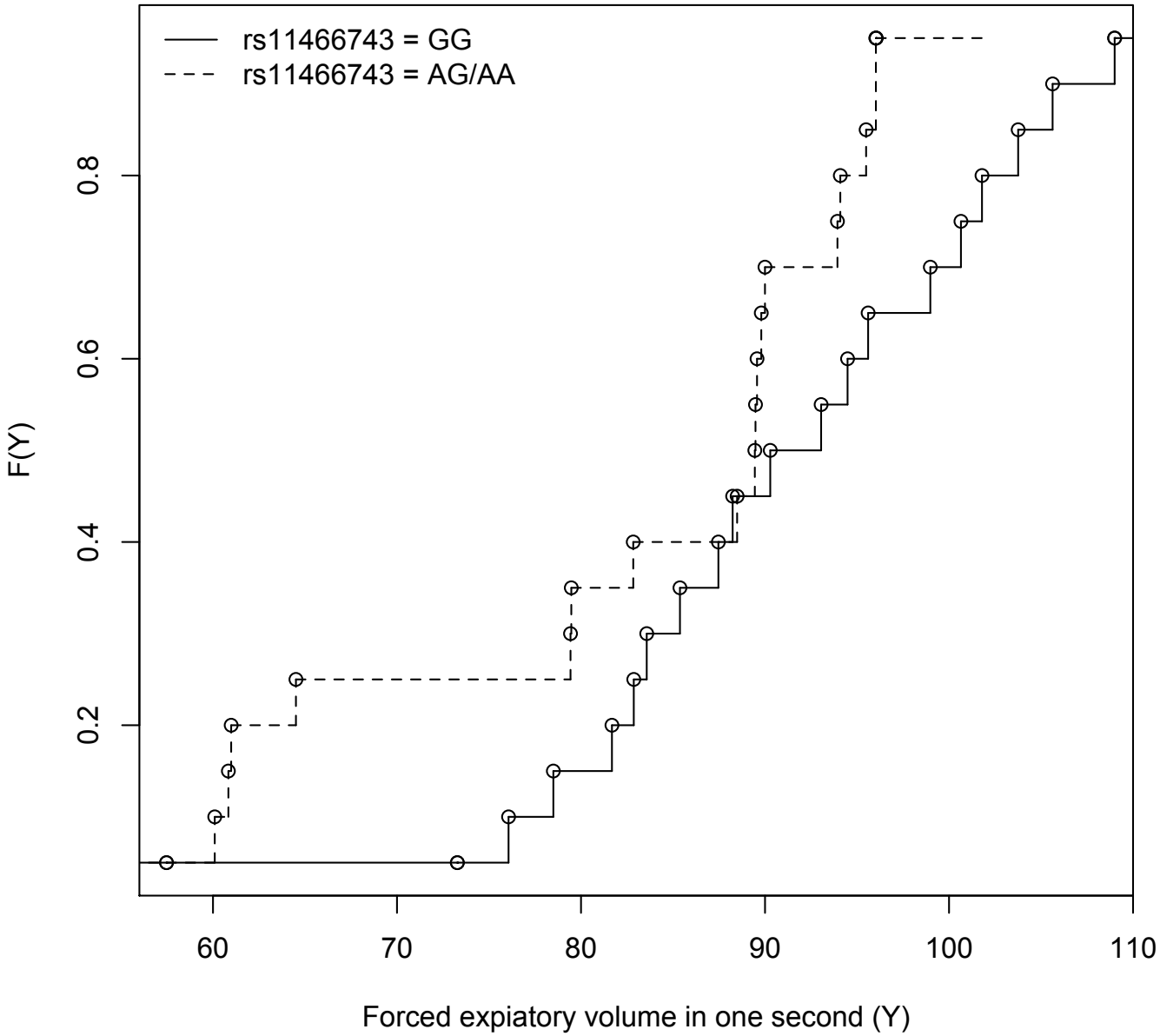
* Mean coefficients were estimated using the profile likelihood methods in Zeng and Lin(2009).

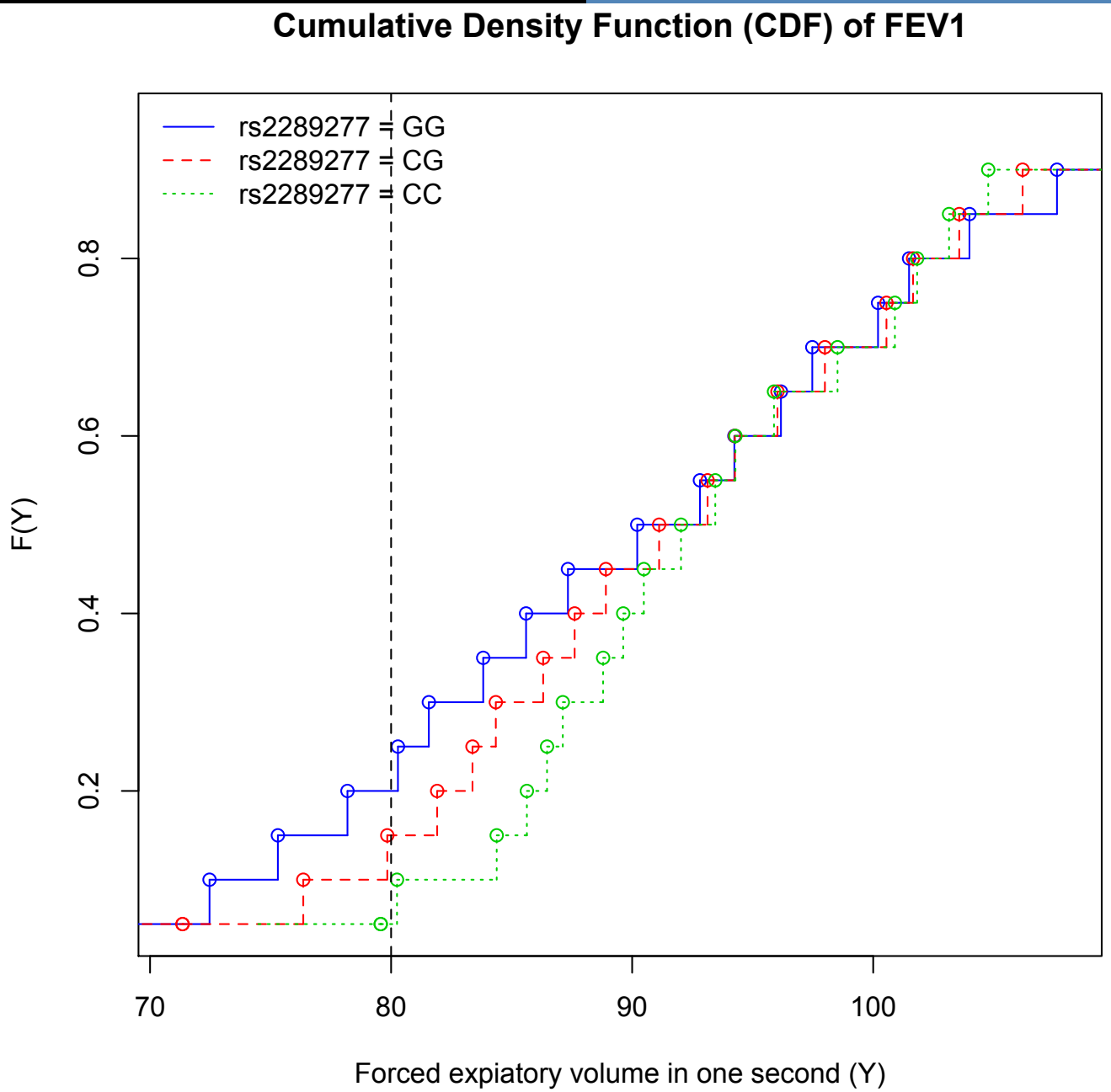
Estimated allelic effects in the mean regression model and quantile regression models at quantile levels of 0.15, 0.25, 0.5, 0.75 and 0.85

	mean*			$\tau = 0.15$		$\tau = 0.25$		$\tau = 0.5$		$\tau = 0.75$	
SNPs	Est.	P-value	Method	Est.	P-value	Est.	P-value	Est.	P-value	Est.	P-value
rs1898671	0.2	0.920	SICO(m=10)	-2.7	0.000	-2.2	0.003	-1.8	0.010	-0.4	0.550
			IPW	-3.1	0.176	-2.1	0.307	-0.7	0.776	0.6	0.800
rs2289277	0.6	0.544	SICO(m=10)	3.0	0.000	2.6	0.000	0.2	0.797	0.5	0.309
			IPW	3.4	0.105	2.1	0.075	-0.5	0.800	0.2	0.872
rs10035870	-1.1	0.659	SICO(m=10)	0.9	0.472	1.5	0.234	2.8	0.063	5.6	0.000
			IPW	-2.2	0.606	0.2	0.966	2.2	0.688	5.7	0.319

* Mean coefficients were estimated using the profile likelihood methods in Zeng and Lin(2009).

Conditional FEV1 distribution Prediction




Cumulative Density Function (CDF) of FEV1
rs2289277 = GG
rs2289277 = CG
rs2289277 = CC
F(Y)
0.2
0.4
0.6
0.8
70
80
90
100
Forced expiatory volume in one second (Y)

More recent development in secondary quantile analysis

- Semi-parametric Efficient Estimation in Quantile Regression of Secondary Analysis by Liang, Ma, Wei and Carroll (2018)
- Based on the concept of "Hypothetical Population", where (1) the disease to non-disease ratio is n_1/n_0 and (2) $F(Y, \mathbf{X}|D) = F_{\text{true}}(Y, \mathbf{X}|D)$.
- A case-control sample from the true population can be treated as a random sample from the super-population.
- Existing semi-parametric efficient estimation for iid sample can be applied to enhance the efficiency.

More recent development in secondary quantile analysis

- The join likelihood of $(Y, \mathbf{X}, D)$ in the super-population can be written and decompose as

$$
\begin{aligned}
& f^s_{\mathbf{X},Y,D}(\mathbf{x}, y, d) \\
&= f^s_D(d) f^t_{\mathbf{X},Y|D}(\mathbf{x}, y, d) = \frac{n_d}{n} f^t_{\mathbf{X},Y|D}(\mathbf{x}, y, d) \\
&= \frac{n_d}{n} \frac{\eta_1(\mathbf{x})\eta_2(\epsilon_\tau, \mathbf{x}) f^t_{D|\mathbf{X},Y}(d, \mathbf{x}, y, \boldsymbol{\alpha})}{\int \eta_1(\mathbf{x})\eta_2(\epsilon_\tau, \mathbf{x}) f^t_{D|\mathbf{X},Y}(d, \mathbf{x}, y, \boldsymbol{\alpha}) d\mu(\mathbf{x})\mu(y)} \\
&= \frac{n_d \eta_1(\mathbf{x})\eta_2(y - \beta_{\tau,c} - \mathbf{x}^\mathsf{T}\boldsymbol{\beta}_\tau, \mathbf{x}) f(d, \mathbf{x}, y, \boldsymbol{\alpha})}{n \int \eta_1(\mathbf{x})\eta_2(y - \beta_{\tau,c} - \mathbf{x}^\mathsf{T}\boldsymbol{\beta}_\tau, \mathbf{x}) f(d, \mathbf{x}, y, \boldsymbol{\alpha}) d\mu(\mathbf{x})\mu(y)},
\end{aligned}
$$

- Semi-parametric efficient estimation equation is consequently constructed.

More recent development in secondary quantile analysis

Efficient score in quantile regression

$$\mathbf{S}_{\text{eff}}(\mathbf{X}, Y, D; \theta, \eta) = \mathbf{S} - \mathbf{g}(Y - r(\mathbf{X}, \alpha), \mathbf{X}) - (1 - D)\mathbf{v}_0 - D\mathbf{v}_1,$$

where

$$\mathbf{S} = \mathbf{S}(\mathbf{x}, y, d, \theta, \eta_2) = \begin{bmatrix} \partial \log\{f(d, \mathbf{x}, y, \beta)\}/\partial\beta \\ \partial \log[\eta_2\{y - r(\mathbf{x}, \alpha), \mathbf{x}\}]/\partial\alpha \end{bmatrix},$$

$b_d \equiv E\{f_{D|\mathbf{X},Y}(1 - d, \mathbf{X}, Y) \mid D = d\}$; $u_\tau = I(\epsilon < 0) - \tau$;
$\mathbf{c}_d \equiv E(\mathbf{S} \mid D = d) - E\{E(\mathbf{S} \mid \epsilon, \mathbf{X}) \mid D = d\}$;
$\kappa(\mathbf{x}, y) \equiv [\sum_{d=0}^{1}\{n_d f(d, \mathbf{x}, y)\}/(n\pi_d)]^{-1}$;
$t_1(\mathbf{x}) \equiv [E^t\{u_\tau^2 \kappa(\mathbf{X}, Y) \mid \mathbf{x}\}]^{-1}$;
$\mathbf{t}_2(\mathbf{x}) \equiv E^t\{u_\tau E(\mathbf{S} \mid \epsilon_\tau, \mathbf{x}) \mid \mathbf{x}\} - (\mathbf{c}_0/b_0)E^t\{u_\tau f_{D|\mathbf{X},Y}(0, \mathbf{x}, Y) \mid \mathbf{x}\}$;
$t_3(\mathbf{x}) \equiv -b_0^{-1}E^t\{u_\tau f_{D|\mathbf{X},Y}(0, \mathbf{x}, Y) \mid \mathbf{x}\}$; $\mathbf{a}(\mathbf{x}) \equiv t_1(\mathbf{x})\{\mathbf{t}_2(\mathbf{x}) + t_3(\mathbf{x})\mathbf{u}_0\}$;
$\mathbf{u}_0 \equiv (1 - E[u_\tau t_1(\mathbf{X})t_3(\mathbf{X})\kappa(\mathbf{X}, Y) \mid D = 0])^{-1}E[u_\tau t_1(\mathbf{X})\mathbf{t}_2(\mathbf{X})\kappa(\mathbf{X}, Y) \mid$
$D = 0]$; $\mathbf{u}_1 \equiv -(n_0/n_1)\mathbf{u}_0$; $\mathbf{v}_0 \equiv (\pi_1/b_0)(\mathbf{u}_0 + \mathbf{c}_0)$;
$\mathbf{v}_1 \equiv -(\pi_0/b_0)(\mathbf{u}_0 + \mathbf{c}_0)$; $\mathbf{g}(\epsilon, \mathbf{x}) \equiv E(\mathbf{S} \mid$
$\epsilon_\tau, \mathbf{x}) - u_\tau \mathbf{a}(\mathbf{x})\kappa(\mathbf{x}, y) - \mathbf{v}_0 f_{D|\mathbf{X},Y}(0, \mathbf{x}, y) - \mathbf{v}_1 f_{D|\mathbf{X},Y}(1, \mathbf{x}, y)$.

More recent development in secondary quantile analysis

- The efficient estimating function can be estimated from the sample
- The resulting quantile estimates are consistent and more efficient
- The population prevalence could be unknown

Numerical example

Table: $Y = 0.5 + X + (1 + 0.3X)\epsilon$, ϵNormal

	τ	0.1	0.25	0.5	0.75	0.9
Truth	β	0.744	0.865	1	1.135	1.256
SICO	mean	0.740	0.862	0.995	1.125	1.244
Semi	mean	0.747	0.867	0.996	1.126	1.246
SICO	sd	0.196	0.159	0.142	0.156	0.192
Semi	sd	0.187	0.144	0.119	0.126	0.155

Numerical example

Table: $Y = 0.5 + X + (1 + 0.3X)\epsilon$, ϵ Gamma

	τ	0.1	0.25	0.5	0.75	0.9
Truth	β	1.097	1.163	1.268	1.412	1.578
SICO	mean	1.096	1.158	1.260	1.404	1.583
Semi	mean	1.100	1.163	1.265	1.408	1.561
SICO	sd	0.089	0.103	0.131	0.178	0.277
Semi	sd	0.084	0.096	0.115	0.150	0.231

Acknowledgement

Collaborators:

Iuliana Ionita-Laza (Columbia)
Xiaoyu Song (Mount Sinai)
Joan Reibman (NYU)
Ray Carroll (Texas A&M)
Mary Beth Terry (Columbia)
Mengling Liu (NYU)
Yanyuan Ma (PSU)
Liang Liang (Texas A&M, Harvard)

Grants:

This research is supported by NIH (R03 HG007443-01, R01HG008980).