Quantile Regression in the Post GWAS Secondary Phenotype Analysis

Ying Wei

1 Department of Biostatistics
Columbia University
Overview

1. Overview

2. Quantile regression for secondary phenotypes
   - A weighted estimating equation approach
   - Semiparametric efficient secondary quantile analysis
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
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.

- Analyzing 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

1 Overview

2 Quantile regression for secondary phenotypes
   • A weighted estimating equation approach
   • Semiparametric efficient secondary quantile analysis
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

do: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 \(\sim170,000\) 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\(^6\), including *Drosophila*\(^6\), snails\(^7\), maize\(^8\) and chickens\(^8\), and specific quantitative trait loci with an effect on variance have been reported for yeast\(^9\) and *Arabidopsis*\(^10\). Many theories and methods to identify genetic loci responsible for phenotypic variability have been proposed\(^3\)\(^-\)\(^4\). 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 \( \{x_i, y_i, d_i = 1, z_i\}_{i=1,2,...,n_1} \), and \( n_2 \) controls denoted by \( \{x_i, y_i, d_i = 0, z_i\}_{i=n_1+1,...,n} \)
  - \( d_i \) is the binary indicator for cases;
  - \( y_i \) is a continuous phenotype of interest
  - \( x_i \) is a \( p \)-dimensional vector of SNP
  - \( z_i \) is a \( q \)-dimensional controlling variables

- Quantile model in general population
  \[
  Q_\tau(y_i|x_i) = x_i^T \beta_{0,\tau} + z_i^T \gamma_{0,\tau}.
  \]

- Goal: to estimate \( \beta_{0,\tau} \) consistently using a case-control sample.
Major statistical issues in Secondary (Quantile) Analysis

- Due to the case-control sampling, \((x_i, y_i, z_i)\) 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 $P(\text{select} | X, D)$ 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 \( E(Y|X) = X^T \beta \),
  \[
  \beta = \arg \min_{\beta} E_Y (Y - X^T \beta)^2 \\
  \hat{\beta} = \arg \min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2.
  \]

- Assume \( Q_\tau(Y|X) = X^T \beta(\tau) \), then
  \[
  \beta(\tau) = \arg \min_{\beta} E_Y (\rho_\tau \{ Y - X^T \beta \}) \\
  \hat{\beta}(\tau) = \arg \min_{\beta} \sum_{i=1}^{n} \rho_\tau \{ y_i - x_i^T \beta \}.
  \]
Quantile regression estimating functions

- Estimating function: \( \Psi_{\tau}(X, Y, \beta) = \left[ \tau - I\{Y \leq X^T \beta \} \right] X \)

- At the true \( \beta_{0, \tau} \),
  \[ E_Y \left[ \Psi_{\tau}(X, Y, \beta_{0, \tau}) \mid X \right] = 0. \]

- With a representative sample, one can obtain a consistent estimate by solving
  \[ \sum_{i=1}^{n} \Psi_{\tau}(x_i, y_i, \beta) = 0 \]
New estimating functions with counter-factual observations

- Expand the original estimating functions

\[ 0 = E \left[ \Psi_{\tau}(X, Y, \beta_{0,\tau}) \mid X \right] = E_Y \left[ \Psi_{\tau}(X, Y, \beta_{0,\tau}) \mid X, D = 0 \right] P(D = 0 \mid X) + E_Y \left[ \Psi_{\tau}(X, Y, \beta_{0,\tau}) \mid X, D = 1 \right] P(D = 1 \mid X) \]

- 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

\[ \sum_{i=1}^{n} \left\{ \Psi_{\tau}(x_i, y_i, \beta) p(d_i \mid x_i) + \Psi_{\tau}(x_i, \tilde{y}_i, \beta) p(1 - d_i \mid x_i) \right\} = 0, \]
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

$$\beta_0(\tau | d) = \arg\min_{\beta} E_Y[||\Psi_\tau(Y, X, \beta)|| | X, D = d], \forall \tau$$

(1)

$-$ $x^T \beta_0(\tau | 0)$ is the conditional quantile function of $y$ given $x$ among controls,
$-$ $x^T \beta_0(\tau | 1)$ defines that among disease population.
Simulating the counter-factual outcomes

- Estimate $\beta_0(\tau | 0)$ and $x^\top \beta_0(\tau | 1)$ from cases and controls separately on a fine grid of $\tau$’s.

- For the $i$th subject, $i = 1, \ldots, n$, we simulate its pseudo outcome $\tilde{y}_i$ by

$$\tilde{y}_i = x_i^\top \hat{\beta}(u_i | 1 - d_i), \quad u_i \sim U(0, 1)$$

where $u_i$ is a random draw from Uniform $(0, 1)$ distribution.

- The sampling estimating equations are then

$$\sum_{i=1}^{n} \Psi_\tau(x_i, y_i, \beta)p(d_i | x_i) + \Psi_\tau(x_i, \tilde{y}_i, \beta)p(1 - d_i | x_i) = 0. \quad (2)$$
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.

$$\hat{\beta}_{n,\tau} = m^{-1} \sum_{\ell=1}^{m} \hat{\beta}_{n,\tau}^{(\ell)}.$$
To estimate the conditional disease probability $P(D|X)$, we assume a logistic model

$$P(D = 1|X) = \frac{\exp(\gamma_0 + X^T \gamma_1)}{1 + \exp(\gamma_0 + X^T \gamma_1)}$$

Or, use the derived model from the primary analysis.

The slope $\gamma_1$ can be consistently estimated by regressing $d_i$ over $x_i$ using the case control sample (Prentice and Pyke, 1979).

The intercept $\gamma_0$ can be calculated by solving the equation

$$P_0 = \int_X \frac{\exp(\gamma_0 + X^T \hat{\gamma}_1)}{1 + \exp(\gamma_0 + X^T \hat{\gamma}_1)}dF_X,$$

where $P_0$ is the overall disease prevalence, $F_X$ is the distribution of $X$. 
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 $X$ is hard to obtain, we proposed to approximate $\gamma_0$ by solving its sample version,

$$\hat{\gamma}_0 = \arg \min_{\gamma_0} \left( P_0 - \frac{1}{n} \sum_{i=1}^{n} \frac{\exp(\gamma_0 + x_i^T \hat{\gamma}_1)}{1 + \exp(\gamma_0 + x_i^T \hat{\gamma}_1)} \right)^2. \quad (4)$$

$P_0$ 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*

\[
\sqrt{n} \left( \hat{\beta}_{n,\tau} - \beta_{n,\tau} \right) \to N \left( 0, G_0^{-1} \Sigma_0 G_0^{-1} \right),
\]

*where* \( \Sigma_0 = V_1 + m^{-1} V_2 + 2 U_1 + \{(m - 1)/m\} U_2 \).
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 lymphopoi etin (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
  \[
  Q_{\tau}(FEV_1) = \beta_{0,\tau} + \beta_{1,\tau}X + \beta_{2,\tau}Z, \tag{5}
  \]
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

<table>
<thead>
<tr>
<th>SNPs</th>
<th>mean*</th>
<th>τ = 0.15</th>
<th>τ = 0.25</th>
<th>τ = 0.5</th>
<th>τ = 0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est.</td>
<td>P-value</td>
<td>Est.</td>
<td>P-value</td>
<td>Est.</td>
</tr>
<tr>
<td>rs11466743</td>
<td>-8.1</td>
<td>0.009</td>
<td>-17.2</td>
<td>0.003</td>
<td>-6.7</td>
</tr>
<tr>
<td></td>
<td>SICO(10)</td>
<td></td>
<td>IPW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2289278</td>
<td>3.5</td>
<td>0.041</td>
<td>0.5</td>
<td>0.691</td>
<td>-1.9</td>
</tr>
<tr>
<td></td>
<td>SICO(10)</td>
<td></td>
<td>IPW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs11241090</td>
<td>-4.8</td>
<td>0.042</td>
<td>-3.7</td>
<td>0.189</td>
<td>-0.3</td>
</tr>
<tr>
<td></td>
<td>SICO(10)</td>
<td></td>
<td>IPW</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* 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

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

* Mean coefficients were estimated using the profile likelihood methods in Zeng and Lin(2009).
Conditional FEV1 distribution Prediction

- \( rs11466743 = GG \)
- \( rs11466743 = AG/AA \)
A weighted estimating equation approach
Semiparametric efficient secondary quantile analysis

Overview
Quantile regression for secondary phenotypes

Cumulative Density Function (CDF) of FEV1

Forced expiatory volume in one second (Y)

rs2289277 = GG
rs2289277 = CG
rs2289277 = CC
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) $F(Y, X|D) = F_{\text{true}}(Y, X|D)$.

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

\[
f^s_{X, Y, D}(x, y, d) = f^s_D(d) f^t_{X, Y|D}(x, y, d) = \frac{n_d}{n} f^t_{X, Y|D}(x, y, d)
\]

\[
= \frac{n_d}{n} \int \eta_1(x) \eta_2(\epsilon_\tau, x) f^t_{D|x, Y}(d, x, y, \alpha) d\mu(x) \mu(y)
\]

\[
= \frac{n_d \eta_1(x) \eta_2(y - \beta_{\tau, c} - x^T \beta_\tau, x)}{n \int \eta_1(x) \eta_2(y - \beta_{\tau, c} - x^T \beta_\tau, x) f(d, x, y, \alpha) d\mu(x) \mu(y)}
\]

Semi-parametric efficient estimation equation is consequently constructed.
More recent development in secondary quantile analysis

**Efficient score in quantile regression**

\[
S_{\text{eff}}(X, Y, D; \theta, \eta) = S - g(Y - r(X, \alpha), X) - (1 - D)v_0 - Dv_1,
\]

where

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

\[
b_d = E\{f_D|x, y(1 - d, X, Y) | D = d\}; u_\tau = I(\epsilon < 0) - \tau;
\]

\[
c_d = E(S | D = d) - E\{E(S | \epsilon, X) | D = d\};
\]

\[
\kappa(x, y) = [\sum_{d=0}^{1} n_d f(d, x, y)]/(n_{n_d})^{-1};
\]

\[
t_1(x) = [E^t\{u_\tau \kappa(X, Y) | x\}]^{-1};
\]

\[
t_2(x) = E^t\{u_\tau E(S | \epsilon, x) | x\} - (c_0/b_0)E^t\{u_\tau f_D|x, y(0, x, Y) | x\};
\]

\[
t_3(x) = -b_0^{-1}E^t\{u_\tau f_D|x, y(0, x, Y) | x\}; a(x) = t_1(x)\{t_2(x) + t_3(x)u_0\};
\]

\[
u_0 = (1 - E[u_\tau t_1(X)t_3(X)\kappa(X, Y) | D = 0])^{-1}E[u_\tau t_1(X)t_2(X)\kappa(X, Y) | D = 0];
\]

\[
u_1 = -(\pi_0/b_0)(u_0 + c_0); v_0 = (\pi_1/b_0)(u_0 + c_0);
\]

\[
v_1 = -(\pi_0/b_0)(u_0 + c_0); g(\epsilon, x) = E(S | \epsilon, x) - u_\tau a(x)\kappa(x, y) - v_0 f_D|x, y(0, x, Y) - v_1 f_D|x, y(1, x, y).
\]
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: \( Y = 0.5 + X + (1 + 0.3X)\epsilon, \epsilon \text{Normal} \)

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>0.1</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth ( \beta )</td>
<td>0.744</td>
<td>0.865</td>
<td>1</td>
<td>1.135</td>
<td>1.256</td>
</tr>
<tr>
<td>SICO mean</td>
<td>0.740</td>
<td>0.862</td>
<td>0.995</td>
<td>1.125</td>
<td>1.244</td>
</tr>
<tr>
<td>Semi mean</td>
<td>0.747</td>
<td>0.867</td>
<td>0.996</td>
<td>1.126</td>
<td>1.246</td>
</tr>
<tr>
<td>SICO sd</td>
<td>0.196</td>
<td>0.159</td>
<td>0.142</td>
<td>0.156</td>
<td>0.192</td>
</tr>
<tr>
<td>Semi sd</td>
<td>0.187</td>
<td>0.144</td>
<td>0.119</td>
<td>0.126</td>
<td>0.155</td>
</tr>
</tbody>
</table>
Numerical example

Table: \( Y = 0.5 + X + (1 + 0.3X)\epsilon, \epsilon \text{ Gamma} \)

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>0.1</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth ( \beta )</td>
<td>1.097</td>
<td>1.163</td>
<td>1.268</td>
<td>1.412</td>
<td>1.578</td>
</tr>
<tr>
<td>SICO mean</td>
<td>1.096</td>
<td>1.158</td>
<td>1.260</td>
<td>1.404</td>
<td>1.583</td>
</tr>
<tr>
<td>Semi mean</td>
<td>1.100</td>
<td>1.163</td>
<td>1.265</td>
<td>1.408</td>
<td>1.561</td>
</tr>
<tr>
<td>SICO sd</td>
<td>0.089</td>
<td>0.103</td>
<td>0.131</td>
<td>0.178</td>
<td>0.277</td>
</tr>
<tr>
<td>Semi sd</td>
<td>0.084</td>
<td>0.096</td>
<td>0.115</td>
<td>0.150</td>
<td>0.231</td>
</tr>
</tbody>
</table>
Acknowledgement

Collaborators:

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

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