Two-Sample Instrumental Variable Analysis: Challenges and Some Progress

Qingyuan Zhao

Department of Statistics, The Wharton School, University of Pennsylvania

November 28, 2017

Outline

Some interesting history

Bristol → Admiral William Penn → William Penn → Pennsylvania (Penn's woods).

This talk is based on joint work with

- Jingshu Wang, Dylan Small (Penn).
- Jack Bowden (Bristol).
- Manuscript and slides are available on my webpage `http://www-stat.wharton.upenn.edu/~qyzhao/`.

Part 0 Primer of instrumental variable (IV) and Mendelian randomization (MR).

Part 1 Two-sample IV using heterogeneous samples.

Part 2 New methods for two-sample MR using GWAS summary statistics.

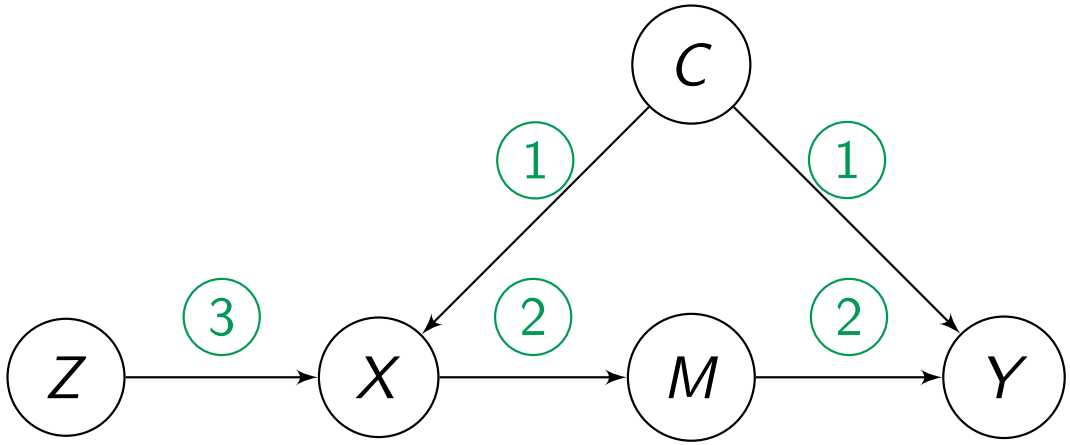
Causal inference

The general problem of causal inference

Without randomized controlled experiments, can we still estimate the **causal effect** of variable X on variable Y?

Three general identification strategies

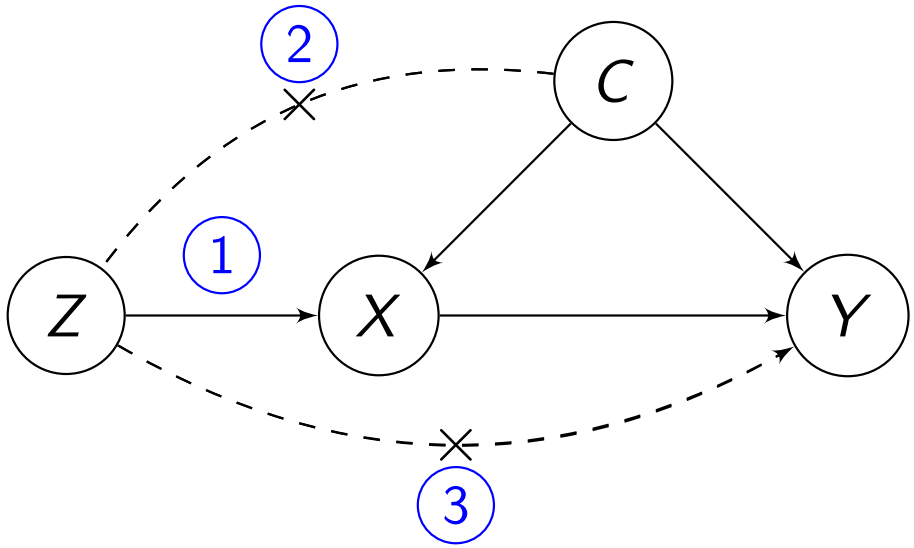
1. Condition on all common causes of X and Y.
2. Study all causal mechanisms by which X influences Y.
3. Use instrumental variables (IV) or natural experiments.



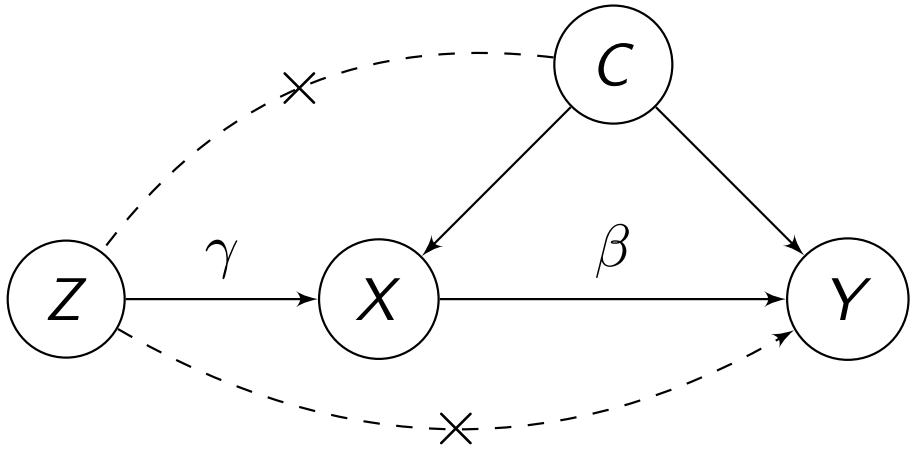
Instrumental variables

Core IV assumptions

1. IV causes the exposure (X).
2. IV is independent of the unmeasured confounder (C).
3. IV cannot have any direct effect on the outcome (Y).



Why does IV work?



Heuristic: Effect of Z on Y entirely goes through X.

Wald ratio estimator

$$\beta = \frac{\mathrm{lm}(Y \sim Z)}{\mathrm{lm}(X \sim Z)}.$$

Two-stage least squares (LS)

$$\beta = \mathrm{lm}(Y \sim \hat{X}), \text{ where } \hat{X} = \mathbb{E}[X|Z] = \mathrm{predict}(\mathrm{lm}(X \sim Z)).$$

Can we trust an IV analysis?

Success of an IV analysis depends on

1. Using *good* instrument(s).
   - Can we reasonably justify the core IV assumptions?
   - Is the IV-exposure association strong enough?
2. Statistical inference.
   - Can we establish consistency and asymptotic normality?
3. Robustness.
   - Can we check if the data satisfies the modeling assumptions?
   - How sensitive is the conclusion to violations of the identification and modeling assumptions?

Mendelian randomization (MR)

A brilliant idea [Katan, 1986, Davey Smith and Ebrahim, 2003]

Use genetic variants as IV.

Recall the three core IV assumptions:

1. Need to find SNPs that are associated with the exposure.
2. Independence of unmeasured confounder is self-evident.
   - The only minor concern is population stratification.
3. Direct effect on the outcome is possible (pleiotropy).

Next

Two great ideas

1. Two-sample IV: don't need the full data (Z, X, Y) for all individuals.
   - Use $(Z, X, \texttt{NA})$ to estimate $\mathrm{lm}(X \sim Z)$.
   - Use $(Z, \texttt{NA}, Y)$ to estimate $\mathrm{lm}(Y \sim Z)$.
   - Dates back at least to Klevmarken [1982] (thanks to David Pacini). The most well known references are Angrist and Krueger [1992], Inoue and Solon [2010].
2. MR with GWAS summary statistics: don't need individual level data.

Next:

Part 1 What if the two samples are from different populations?

Part 2 New statistical methods for two-sample MR.

An example

An easy way to confirm heterogeneity of the two samples: check allele frequency.

SNP	Gene	Allele	Frequency	
			Sample *a*	Sample *b*
rs12916	HMGCR	C	0.40	0.43
rs1564348	LPA	C	0.18	0.16
rs2072183	NPC1L1	C	0.29	0.25
rs2479409	PCSK9	G	0.32	0.35

Table : The instrumental variables usually have different distributions in two-sample Mendelian randomization. In this Table we included four single nucleotide polymorphisms (SNPs) used in Hemani et al. [2016, Figure 2] to estimate the effect of low-density lipoprotein (LDL) cholesterol lowering on the risk of coronary heart disease.

Summary of results

Question

Is this a big problem (for identification and estimation)?

Surprisingly, little is known even though two-sample IV is widely used in econometrics.

Main messages

- Additional untestable assumptions are needed for identification.
- The IV analysis is no longer robust to misspecified instrument-exposure model.
- The two stage LS is not asymptotically efficient.

Some notations

Data: $(\mathbf{z}_i^s, x_i^s, y_i^s)$, $i = 1, 2, \ldots, n^s$ and $s \in \{a, b\}$ is the sample index.

The two-sample instrumental variable problem

Suppose only $\mathbf{Z}^a$, $\mathbf{x}^a$, $\mathbf{Z}^b$, and $\mathbf{y}^b$ are observed (in other words $\mathbf{y}^a$ and $\mathbf{x}^b$ are not observed).
If x is endogenous, what can we learn about the exposure-outcome relationship by using the IVs $\mathbf{z}$?

Message 1: Identification

Assumption	Detail	1	2	3	4
(1) Structural model	$Y \sim X$: $y_i^s = g^s(x_i^s, u_i^s)$ $X \sim Z$: $x_i^s = f^s(\mathbf{z}_i^s, v_i^s)$	✓	✓	✓	✓
(2) Validity of IV	$\mathbf{z}_i^s \perp\!\!\!\perp (u_i^s, v_i^s)$	✓	✓	✓	✓
(3.1) Linearity of $Y \sim X$	$g^b(x_i, u_i) = \beta^b x_i + u_i$	✓	✓		
(3.2) Linearity of $X \sim Z$	$f^s(\mathbf{z}_i, v_i) = (\boldsymbol{\gamma}^s)^T \mathbf{z}_i + v_i$	✓			
(4) Structural invariance	$f^a = f^b$	✓	✓	✓	✓
(5) Sampling homogeneity of noise	$v_i^a \stackrel{d}{=} v_i^b$			✓	
(6) Additivity of $X \sim Z$	$f^s(\mathbf{z}, v) = f_z^s(\mathbf{z}) + f_v^s(v)$		✓		
(7) Monotonicity	$f^s(z, v)$ is monotone in z			✓	✓
Identifiable estimand		β^b	β^b	β^b_{LATE}	β^{ab}_{LATE}

Table : Summary of some identification results and assumptions. Highlighted assumptions (4 and 5) are new due to heterogeneity and untestable. Case 3 and 4 consider binary IV and binary exposure. β^b_{LATE} is the local average treatment effect (LATE) in population b [Angrist, Imbens, and Rubin, 1996]. $\beta^{ab}_{\text{LATE}} = \beta^b_{\text{LATE}} \times \mathbb{P}_b(\text{complier})/\mathbb{P}_a(\text{complier})$.

A robustness property of one-sample IV

A well known fact

In one-sample IV analysis, two stage LS is robust against misspecified IV-exposure model.

Why? β can be identified by the estimating equation

$$\mathbb{E}[h(\mathbf{z})(y - x\beta)] = 0$$

for *any* function h of $\mathbf{z}$.

- IV estimate: $\hat{\beta}_h = \Big[\sum_{i=1}^n y_i h(\mathbf{z}_i)\Big] \Big/ \Big[\sum_{i=1}^n x_i h(\mathbf{z}_i)\Big]$.
- Consistent and asymptotically normal if $\mathrm{Cov}(x, h(\mathbf{z})) \neq 0$.
- The most *efficient* choice is $h^*(\mathbf{z}) = \mathbb{E}[x|\mathbf{z}]$.
- Two-stage LS: $h(\mathbf{z}) = \mathbf{z}^T\gamma$ is the best linear approximation to $h^*(\mathbf{z})$.

Message 2

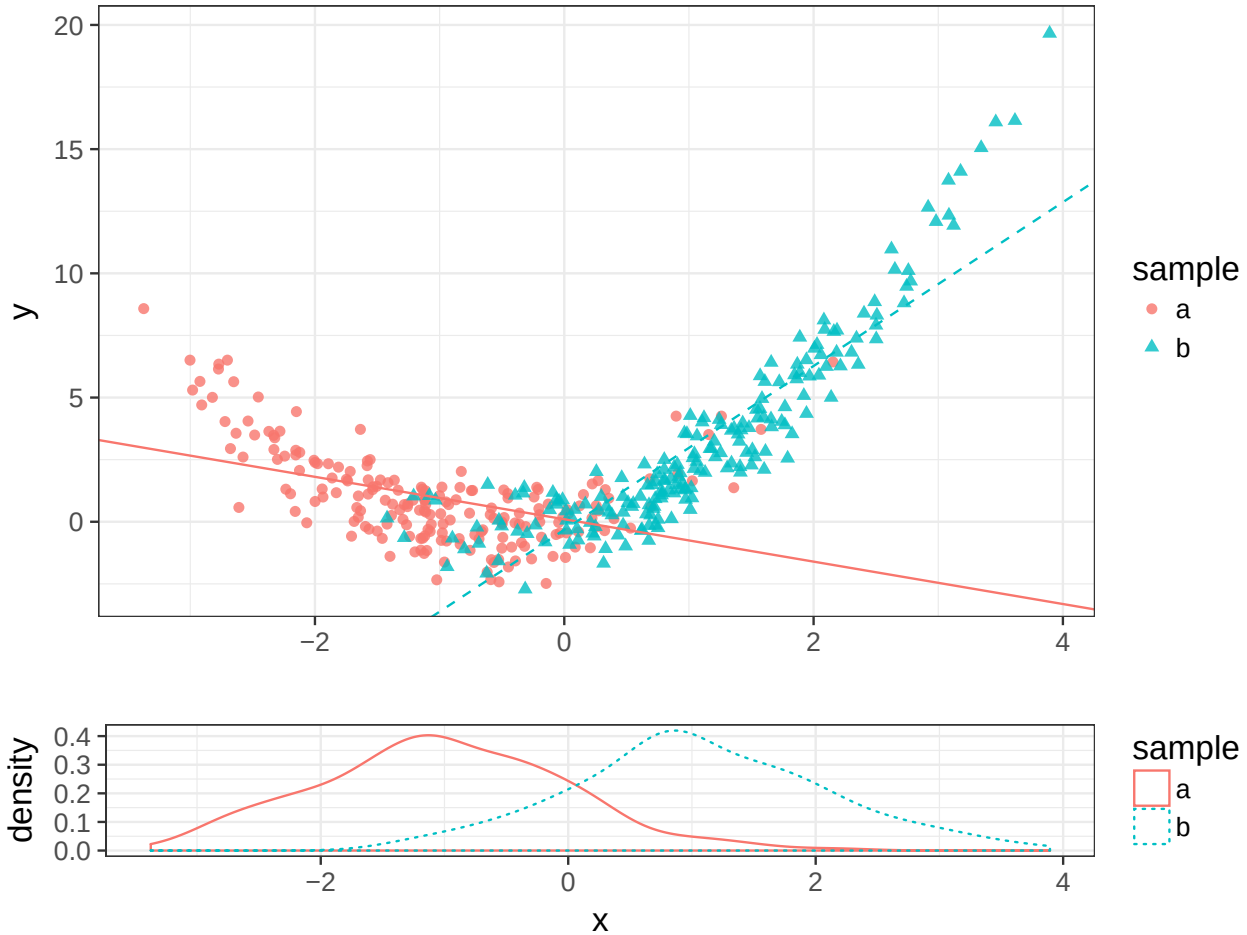
Message 2

This robustness property does not carry to two-sample IV with heterogeneous samples.

Why?

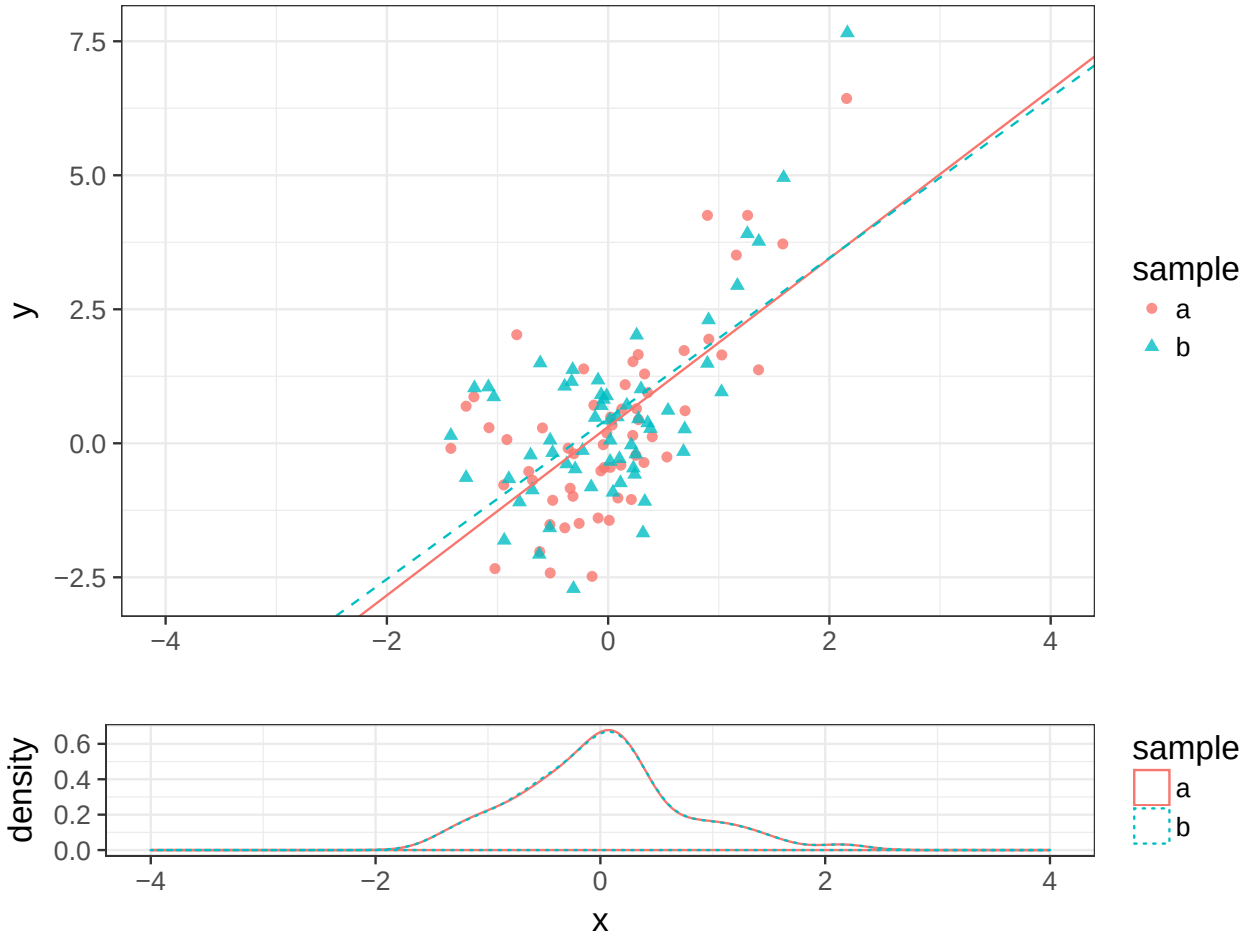
- The best parametric approximation depends on the population!
- Buja et al. [2014] described this "conspiracy" of model misspecification and random design.

An example of the conspiracy



Matching

An intuitive solution: make sure the IVs has the same distribution in both samples, for example by matching.



Message 3

When the linear IV-exposure model is correctly specified, the two-stage LS estimator is asymptotically efficient in the class of limited information estimators

1. In the one-sample setting [Wooldridge, 2010], and
2. In the homogeneous two-sample setting [Inoue and Solon, 2010].

Message 3

The asymptotic efficiency does not carry to two-sample IV with heterogeneous samples.

Generalized method of moments (GMM)

- Assume all the variables are centered. Let $\mathbf{S}$ be the sample covariance matrix. For example, $\mathbf{S}_{zy}^s = (\mathbf{Z}^s)^T \mathbf{y}^s / n^s$.
- Over-identified estimating equations:

$$\mathbf{m}_n(\beta) = (\mathbf{S}_{zz}^b)^{-1} \mathbf{S}_{zy}^b - (\mathbf{S}_{zz}^a)^{-1} \mathbf{S}_{zx}^a \beta.$$

- The class of GMM estimators:

$$\hat{\beta}_{n,\mathbf{W}} = \arg\min_{\beta} \mathbf{m}_n(\beta)^T \mathbf{W} \mathbf{m}_n(\beta).$$

- Two stage LS: $\mathbf{W} = \mathbf{S}_{zz}^b$.
- Optimal choice: $\mathbf{W} \propto \mathrm{Cov}(\mathbf{m}_n(\beta))^{-1} = \frac{1}{n_b} (\mathbf{S}_{zz}^b)^{-1} \mathrm{Var}(y_i^b | \mathbf{z}_i^b) + \frac{1}{n_a} (\mathbf{S}_{zz}^a)^{-1} \beta^2 \mathrm{Var}(x_i^a | \mathbf{z}_i^a)$.

Recap

Three messages of Part I

In two-sample IV with heterogeneous samples,

- Additional untestable assumptions are needed for identification.
- The IV analysis is no longer robust to misspecified instrument-exposure model.
- The two stage LS is not asymptotically efficient.

Next:

Part 2 New statistical methods for two-sample MR using just summary statistics.

Setup

- Suppose we are in an ideal scenario: linearity, homogeneity.

Setup

Suppose we have p SNPs, $Z_1, \ldots, Z_p$.

- IV-exposure sample $\mathrm{lm}(X^a \sim Z_j^a)$.
  - Population parameter: γ_j.
  - Estimator: $\hat{\gamma}_j \sim \mathrm{N}(\gamma_j, \sigma_{j1}^2)$, available from GWAS.
- IV-outcome sample $\mathrm{lm}(Y^b \sim Z_j^b)$.
  - Population parameter: Γ_j.
  - Estimator: $\hat{\Gamma}_j \sim \mathrm{N}(\Gamma_j, \sigma_{j2}^2)$, available from GWAS.

Statistical problem

Suppose $\Gamma_j = \beta\gamma_j$ for all $j = 1, \ldots, p$. Can we provide consistent point estimate and valid confidence interval for β?

Challenges

1. Measurement error: $\hat{\gamma}_j$ is measured with error, so classical linear regression cannot be directly applied.
2. Linkage disequilibrium: $\hat{\Gamma}_j$ and $\hat{\Gamma}_k$ $(j \neq k)$ may be dependent.
   - Can use uncorrelated SNPs (clumping).
3. How many SNPs should we use?
   - Selection bias/winner's curse: typically we only use SNPs such that $|\hat{\gamma}_j|/\sigma_{j1}$ is larger than some threshold.
   - May want toselect SNPs liberally (e.g. p-value $\leq 10^{-4}$) to improve power. However the WR $\hat{\Gamma}_j/\hat{\gamma}_j$ is biased towards 0 due to weak instrument.
4. Pleiotropy: the equation $\Gamma_j = \beta\gamma_j$ might not always be true.
5. ...

A profile likelihood (PL) approach

- A simple setting: $\hat{\gamma}_j \sim \mathrm{N}(\gamma_j, \sigma_{j1}^2)$, $\hat{\Gamma}_j \sim \mathrm{N}(\Gamma_j, \sigma_{j2}^2)$, all independent and variances are known. $\Gamma_j \equiv \beta\gamma_j$.
- Log-likelihood:

$$l(\beta, \boldsymbol{\gamma}) = -\frac{1}{2}\left[\sum_{j=1}^{p} \frac{(\hat{\gamma}_j - \gamma_j)^2}{\sigma_{j1}^2} + \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \gamma_j\beta)^2}{\sigma_{j2}^2}\right].$$

- Challenge: a lot of nuisance parameters $\gamma_1, \ldots, \gamma_p$.
- Profile log-likelihood:

$$l(\beta) = -\frac{1}{2}\sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta\hat{\gamma}_j)^2}{\sigma_{j2}^2 + \sigma_{j1}^2\beta^2}.$$

- Profile likelihood estimator: $\hat{\beta} = \arg\max l(\beta)$.
- Turns out to be the same as the 2nd order weighted estimator [Bowden et al., 2017].

Theoretical results I

Assumption (Variance is $O(1/n)$)

Let $n = \min(n^a, n^b)$ be the sample size. There exists $C \geq 1$ such that $C^{-1}/n \leq \sigma_{j1}^2, \sigma_{j2}^2 \leq C/n$ for all j.

Assumption (Collective strength of IV)

$C^{-1} \leq \|\gamma\|_2^2 \leq C.$

Theorem (Consistency)

If $p/n^2 \to 0$ and the above assumption holds, then $\hat{\beta} \xrightarrow{p} \beta$.

Theoretical results II

Assumption

Suppose $p/n \to \kappa < \infty$. If $\kappa > 0$, there exists $\delta > 0$ such that

$$\frac{1}{p^{1+\delta}} \sum_{j=1}^{p} (n\gamma_j^2 + 1)^{1+\delta} \to 0.$$

Theorem (Asymptotic normality)

Under the preceding assumptions,

$$\frac{V_2}{\sqrt{V_1}}(\hat{\beta} - \beta) \xrightarrow{d} \mathrm{N}(0,1) \text{ as } n \to \infty, \text{ where}$$

$$V_1 = \sum_{j=1}^{p} \frac{\gamma_j^2\sigma_{j2}^2 + \Gamma_j^2\sigma_{j1}^2 + \sigma_{j1}^2\sigma_{j2}^2}{(\sigma_{j2}^2 + \sigma_{j1}^2\beta^2)^2} = O(n+p), \; V_2 = \sum_{j=1}^{p} \frac{\gamma_j^2\sigma_{j2}^2 + \Gamma_j^2\sigma_{j1}^2}{(\sigma_{j2}^2 + \sigma_{j1}^2\beta^2)^2} = O(n).$$

Should we include very weak instruments?

Theorem (Asymptotic normality)

$$\mathrm{Var}(\hat{\beta}) \approx V_1/V_2^2, \text{ where}$$

$$V_1 = \sum_{j=1}^{p} \frac{\gamma_j^2\sigma_{j2}^2 + \Gamma_j^2\sigma_{j1}^2 + \sigma_{j1}^2\sigma_{j2}^2}{(\sigma_{j2}^2 + \sigma_{j1}^2\beta^2)^2}, \quad V_2 = \sum_{j=1}^{p} \frac{\gamma_j^2\sigma_{j2}^2 + \Gamma_j^2\sigma_{j1}^2}{(\sigma_{j2}^2 + \sigma_{j1}^2\beta^2)^2}.$$

An important observation

Including extremely weak instruments ($|\gamma_j|/\sigma_{j1} \ll 1$) may increase the variance of $\hat{\beta}$.

Selection bias/Winner's curse

If we select large $|\hat{\gamma}_j|/\sigma_{j1}$, then $|\hat{\gamma}_j|$ is generally larger than $|\gamma_j|$ (especially if $|\gamma_j|$ is small). The Wald ratio $\hat{\Gamma}_j/\hat{\gamma}_j$ is biased towards 0.

Systematic pleiotropy

- A big concern of MR is $\Gamma_j \equiv \beta\gamma_j$ may not hold.

A random direct effects model (overdispersion)

Suppose $\Gamma_j = \beta\gamma_j + \alpha_j$ and the direct effect $\alpha_j \overset{i.i.d.}{\sim} \mathrm{N}(0, \tau^2)$.

- Profile log-likelihood:

$$l(\beta, \tau^2) = -\frac{1}{2}\left[\sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta\hat{\gamma}_j)^2}{\tau^2 + \sigma_{j2}^2 + \sigma_{j1}^2\beta^2} + \log(\tau^2 + \sigma_{j2}^2)\right].$$

Failure of the profile likelihood

$$\frac{\partial}{\partial\tau^2} l(\beta, \tau^2) = \frac{1}{2}\left[\sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta\hat{\gamma}_j)^2}{(\tau^2 + \sigma_{j2}^2 + \sigma_{j1}^2\beta^2)^2} - \frac{1}{\tau^2 + \sigma_{j2}^2}\right].$$

However, expectation of this score is not 0 at the true (β, τ^2).

Modified score equations

- Estimate β and τ^2 by solving

$$0 = \frac{\partial}{\partial \beta} l(\beta, \tau^2),$$

$$0 = \sum_{j=1}^{p} \sigma_{j1}^2 \left[\frac{(\hat{\Gamma}_j - \beta \hat{\gamma}_j)^2}{(\tau^2 + \sigma_{j2}^2 + \sigma_{j1}^2 \beta^2)^2} - \frac{1}{\tau^2 + \sigma_{j2}^2 + \sigma_{j1}^2 \beta^2} \right].$$

- Can prove consistency and asymptotic normality under similar assumptions as before.

Idiosyncratic pleiotropy

- The random effects model $\alpha_j \sim \mathrm{N}(0, \tau^2)$ may fail to explain some extraordinarily large "outlier".
- Recall the profile log-likelihood

$$l(\beta) = -\frac{1}{2} \sum_{j=1}^{p} \frac{(\hat{\Gamma}_j - \beta \hat{\gamma}_j)^2}{\sigma_{j2}^2 + \sigma_{j1}^2 \beta^2}.$$

Problem: A single SNP can have unbounded influence.

Our solution

Robustify the likelihood/estimating equations, in the same spirit as robust regression (e.g. Huber's loss, Tukey's biweight).

- Consistency is difficult to prove but seems to be true in simulations.
- Asymptotic normality is still true given consistency.

Recap

Three estimators proposed

1. No pleiotropy: PL estimator (compare to IVW).
2. Systematic pleiotropy: modified PL score equation (compare to MR-Egger).
3. Systematic and idiosyncratic pleiotropy: robustified score equation (compare to ???).

Diagnostic tools

1. Residual Quantile-Quantile plot. Standardized residual is

$$\hat{\epsilon}_j = \frac{\hat{\Gamma}_j - \hat{\beta}\hat{\gamma}_j}{\hat{\tau}^2 + \sigma_{j2}^2 + \sigma_{j1}^2\hat{\beta}^2}.$$

2. Leave-one-out plot: investigate the influence of a single SNP.

Next: Three real data examples.

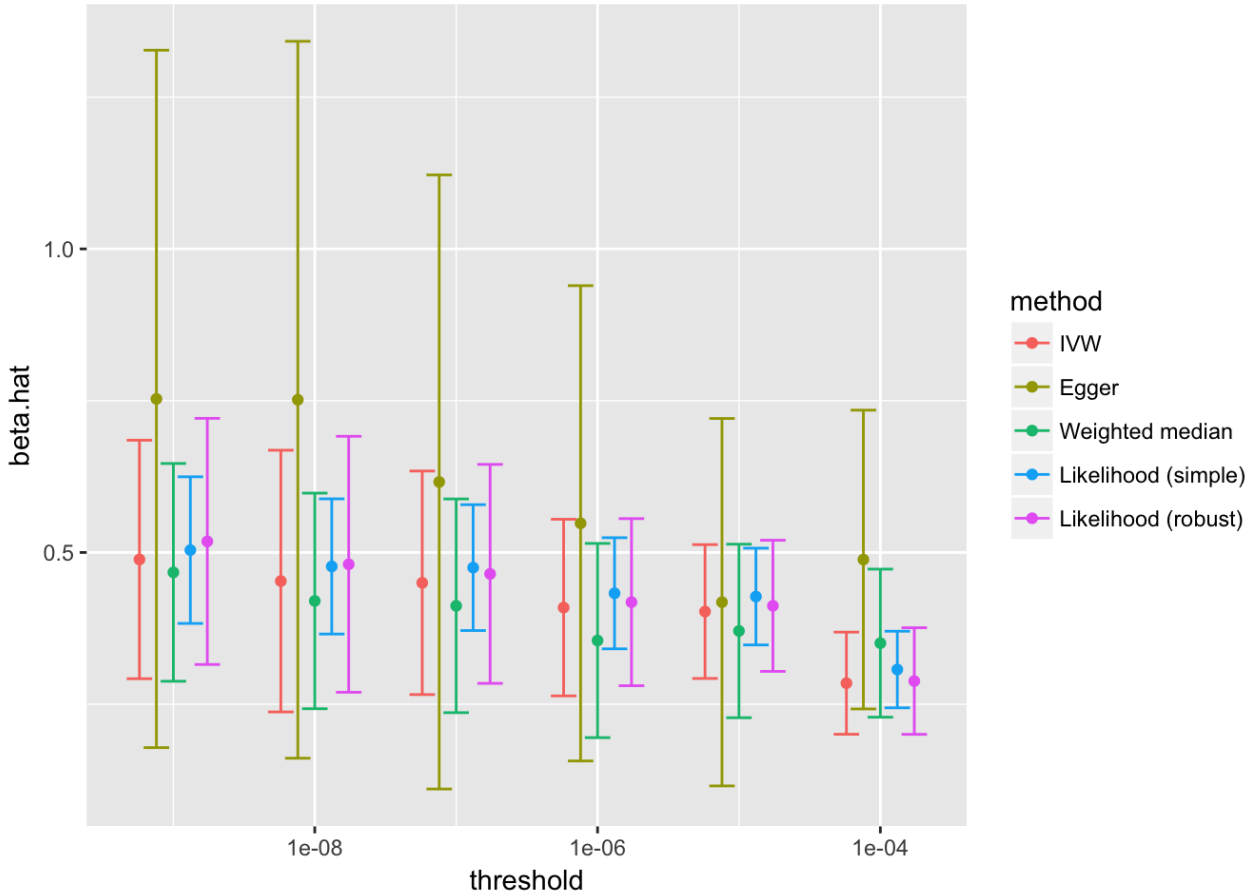
Example 1: BMI and coronary heart disease

Goal of this example

- Theory requires us to select independent and relatively strong instruments.
- In the documentation of `TwoSampleMR`, the same dataset is used for selection and inference. How large is the selection bias?

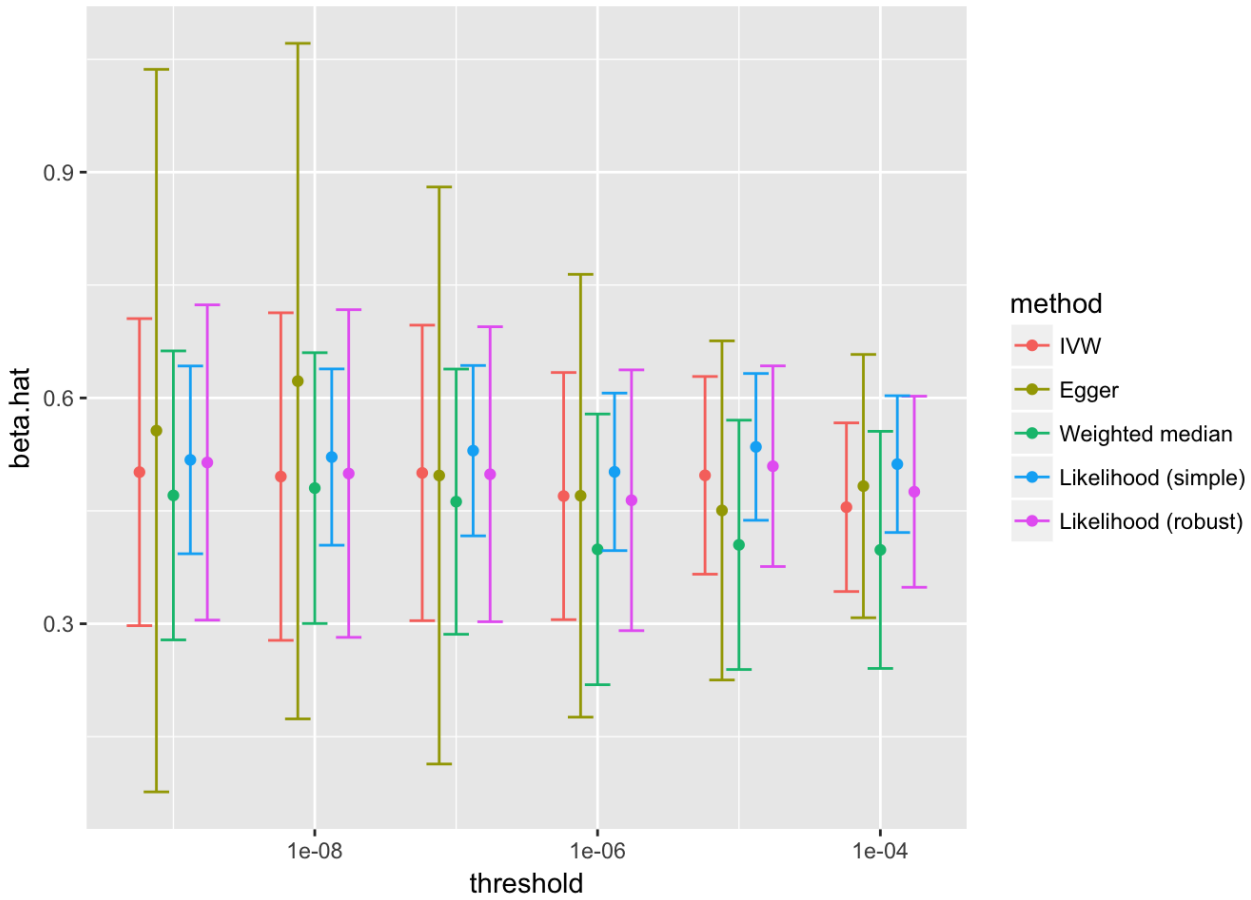
- Locke et al. [2015] reported two independent GWAS of BMI, one for male and one for female.
- Design 1: use the female dataset for both selection (based on $|\hat{\gamma}_j|/\sigma_{j1}$) and statistical inference.
- Design 2: use the female dataset for selection; use the male dataset for inference.

Design 1



- Biased towards 0 due to selection bias/winner's curse.

Design 2



- When there is no selection bias, adding weak instruments (p-value $\approx 10^{-4}$) can still reduce the standard error.

Example 2: LDL-c and coronary heart disease

Goal of this example

Demonstrate the necessity and effectiveness of modifying the profile likelihood score equation.

- Design 2: Two (seemingly) disjoint GWAS are used.
  1. Screening: Kettunen et al. [2016] ($n = 21555$).
  2. Inference: GLGC [2013] ($n = 173082$).
- There are 70 SNPs left after selection.

Example 2: LDL-c and coronary heart disease

- Results of `mr` in `TwoSampleMR`:

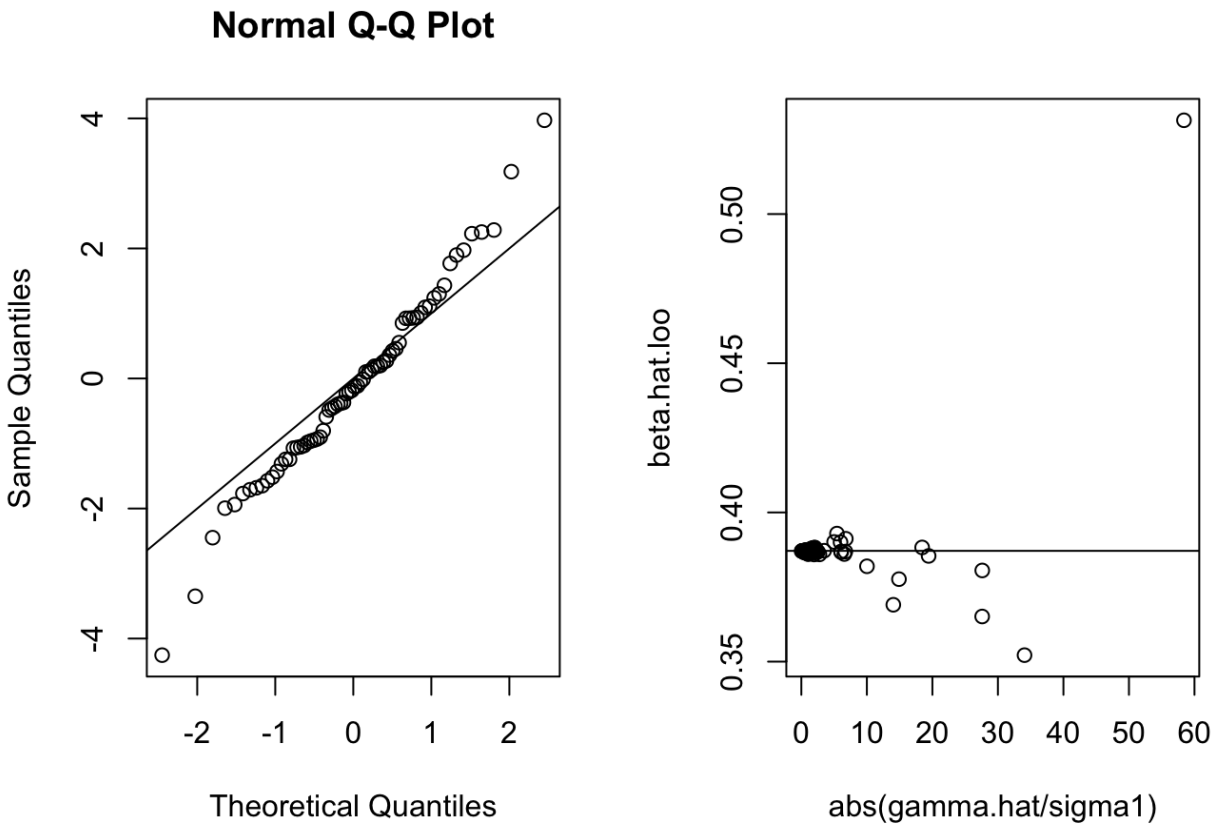
Method	$\hat{\beta}$	se($\hat{\beta}$)
MR-Egger	0.391	0.040
Weighted median	0.233	0.047
Inverse variance weighted	0.377	0.036
Simple mode	0.319	0.513
Weighted mode	0.432	0.435

- Results of our estimators:

Method	$\hat{\beta}$	se($\hat{\beta}$)
PL (Basic)	0.387	0.025
PL (Overdispersed)	0.369	0.031
PL (Overdispersed, Huber)	0.453	0.031
PL (Overdispersed, Tukey)	0.535	0.032

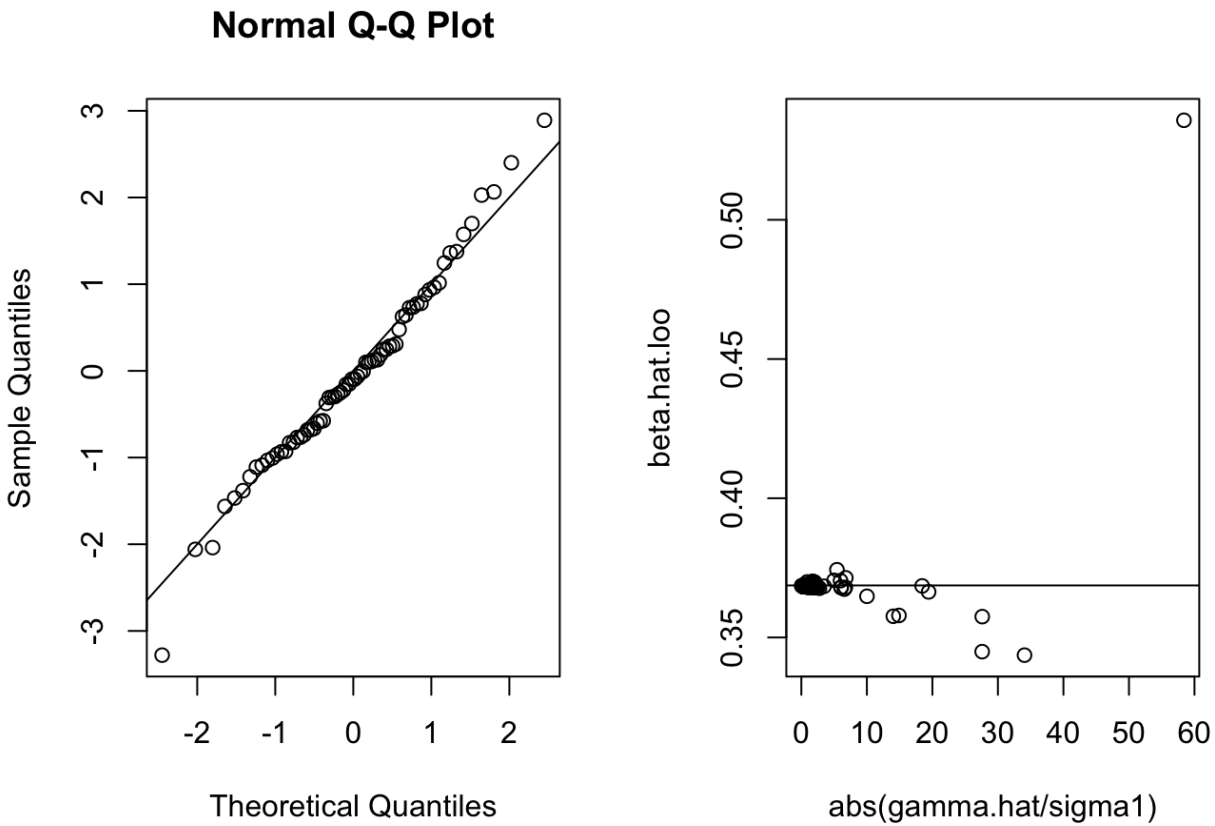
Necessity of considering overdispersion

Diagnostic plots for the PL (basic) estimator:



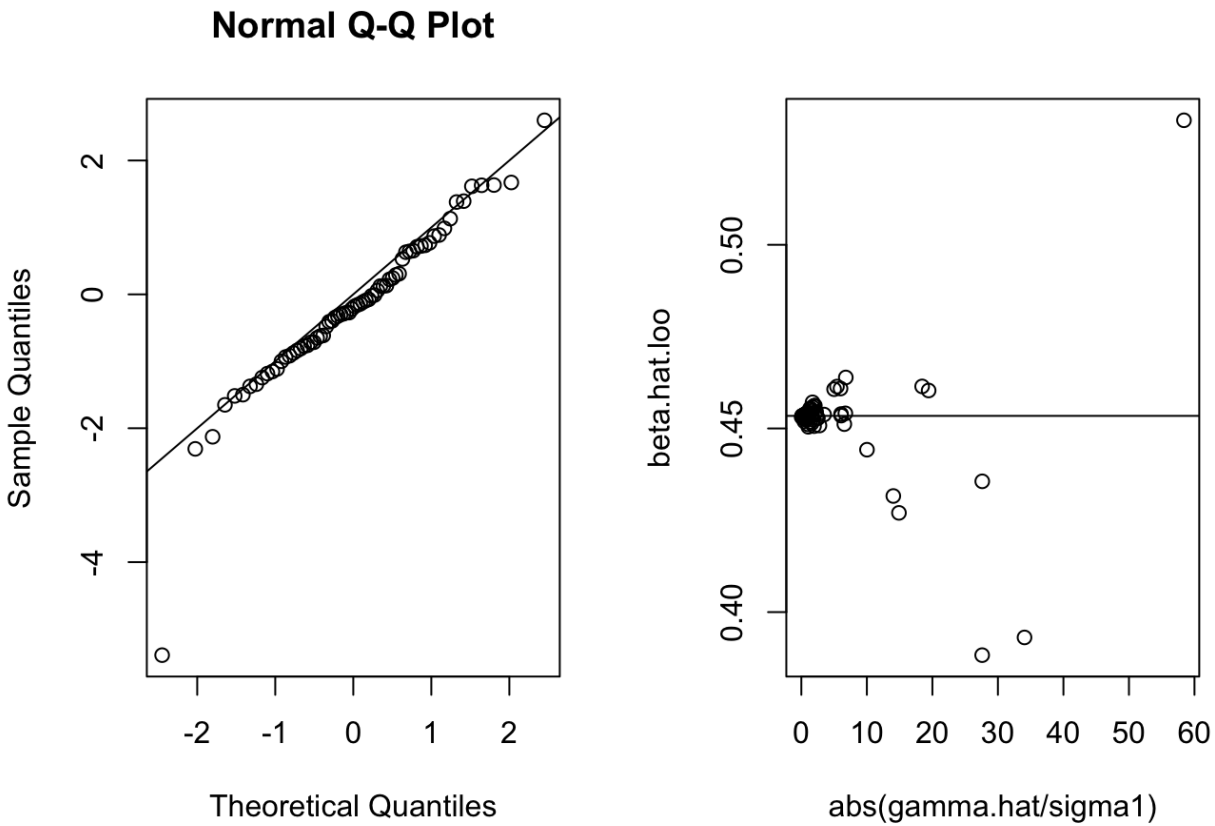
Outlier???

Diagnostic plots for the PL (overdispersed) estimator:



Outlier!!!

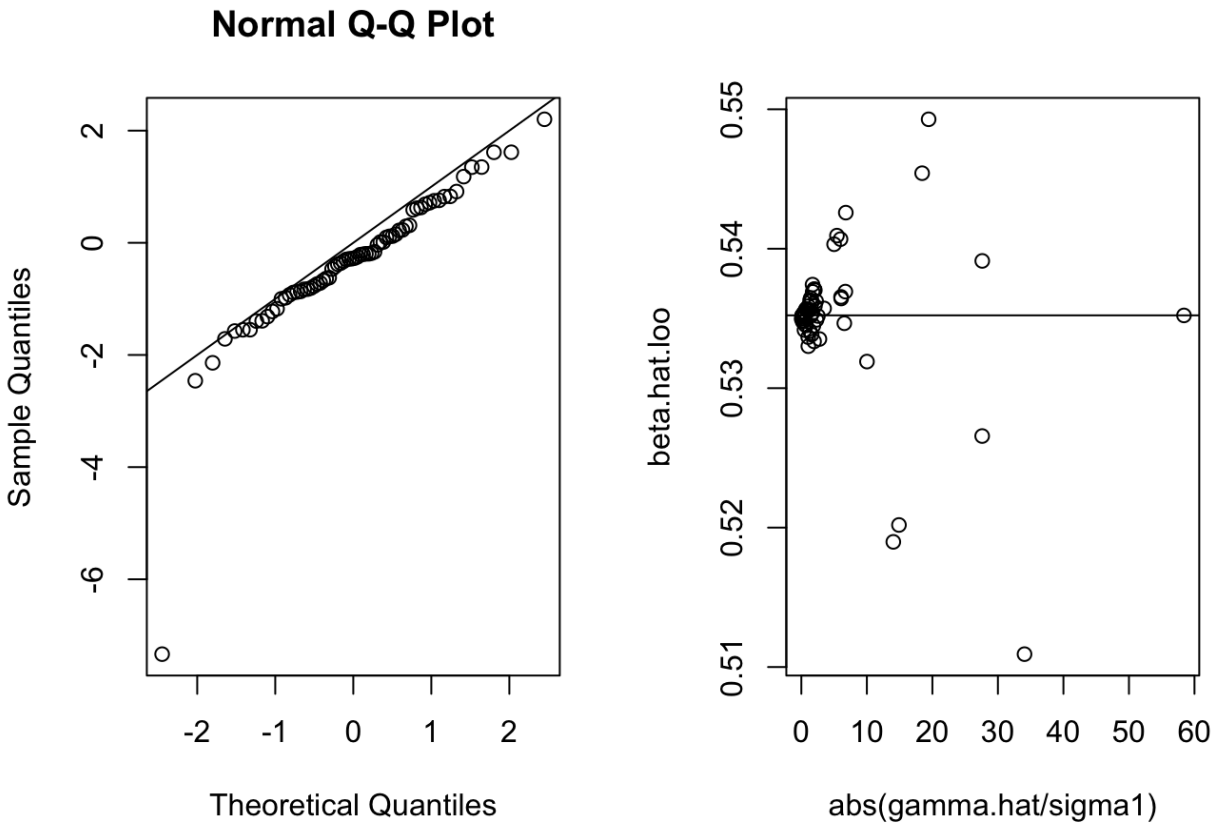
Diagnostic plots for the PL (overdispersed, Huber) estimator:



- The outlier is `rs7412`. I'd appreciate any biological story.

Outlier!!!!!!

Diagnostic plots for the PL (overdispersed, Tukey) estimator:



- To detect outlier, must use robust initial estimator.

Example 3: HDL-c and coronary heart disease

- Design 2: 59 SNPs after selection.
- Results of `mr` in `TwoSampleMR`:

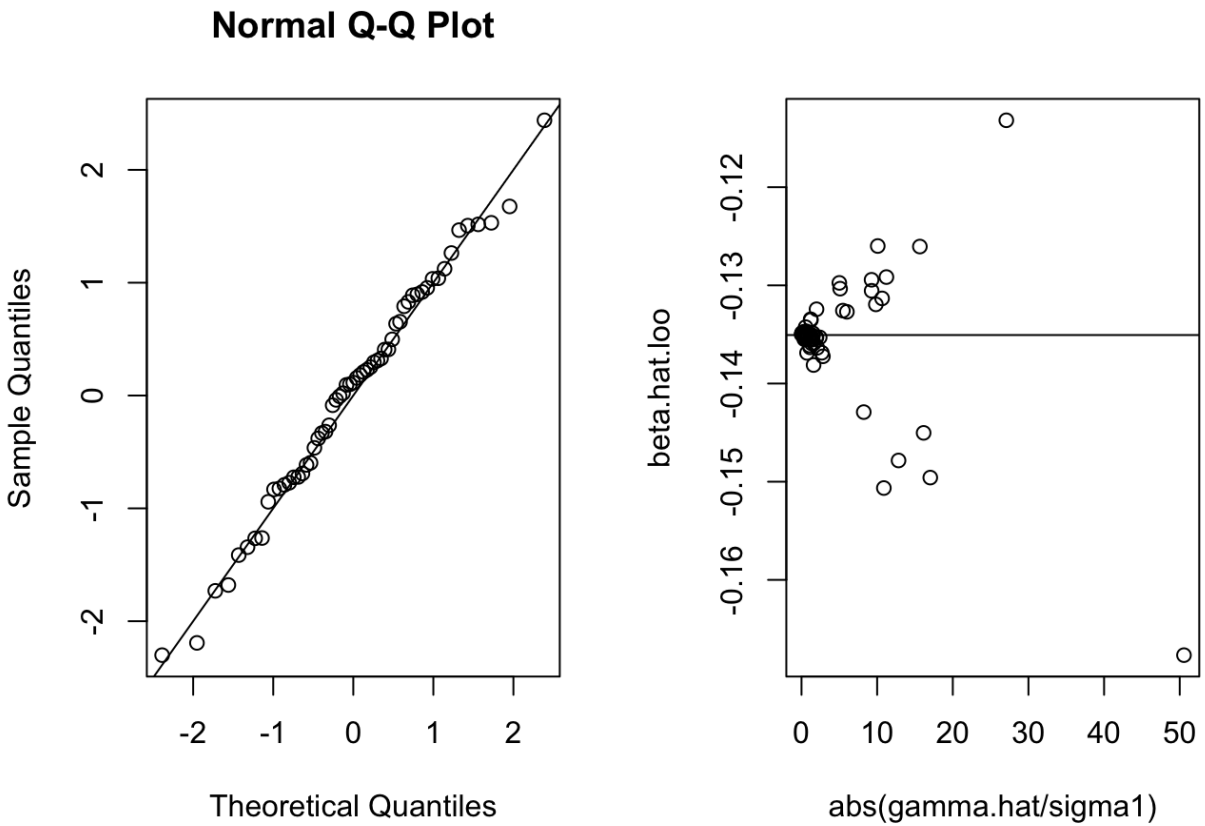
Method	$\hat{\beta}$	se($\hat{\beta}$)
MR-Egger	-0.137	0.047
Weighted median	-0.126	0.040
Inverse variance weighted	-0.138	0.040
Simple mode	0.064	1.438
Weighted mode	-0.103	1.475

- Results of our estimators:

Method	$\hat{\beta}$	se($\hat{\beta}$)
PL (Basic)	-0.142	0.031
PL (Overdispersed)	-0.135	0.041
PL (Overdispersed, Huber)	-0.134	0.043
PL (Overdispersed, Tukey)	-0.135	0.043

Diagnosis

Diagnostic plots for the PL (overdispersed, Tukey) estimator:



- Looks fine (especially the Q-Q plot).

Recap

Three messages of Part 2

1. Sample splitting is very important to obtain unbiased estimator.
2. Pleiotropy (systematic and idiosyncratic) can be handled by modifying the PL score equation.
3. Theoretical guarantees: statistical consistency and asymptotic normality.

Discussion

- Our results for HDL-c are different from previous studies. A possible reason is the sample splitting design.
- Future work: Goodness-of-fit test of the statistical model.
- Good statistical fit $\Rightarrow$ more confidence in the results??

References I

J. D. Angrist and A. B. Krueger. The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *Journal of the American Statistical Association*, 87(418):328–336, 1992.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

J. Bowden, M. Fabiola Del Greco, C. Minelli, D. Lawlor, N. Sheehan, J. Thompson, and G. D. Smith. Improving the accuracy of two-sample summary data mendelian randomization: moving beyond the nome assumption. *bioRxiv*, page 159442, 2017.

A. Buja, R. Berk, L. Brown, E. George, E. Pitkin, M. Traskin, L. Zhao, and K. Zhang. Models as approximations, part i: A conspiracy of nonlinearity and random regressors in linear regression. *arXiv preprint arXiv:1404.1578*, 2014.

G. Davey Smith and S. Ebrahim. "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32(1):1–22, 2003.

GLGC. Discovery and refinement of loci associated with lipid levels. *Nature genetics*, 45(11):1274–1283, 2013.

References II

G. Hemani, J. Zheng, K. H. Wade, C. Laurin, B. Elsworth, S. Burgess, J. Bowden, R. Langdon, V. Tan, J. Yarmolinsky, et al. MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. *bioRxiv*, 2016. doi: 10.1101/078972.

A. Inoue and G. Solon. Two-sample instrumental variables estimators. *The Review of Economics and Statistics*, 92(3):557–561, 2010.

M. Katan. Apoupoprotein e isoforms, serum cholesterol, and cancer. *The Lancet*, 327(8479):507–508, 1986.

J. Kettunen, A. Demirkan, P. Würtz, H. H. Draisma, T. Haller, R. Rawal, A. Vaarhorst, A. J. Kangas, L.-P. Lyytikäinen, M. Pirinen, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of lpa. *Nature Communications*, 7, 2016.

A. Klevmarken. Missing variables and two-stage least-squares estimation from more than one data set. Technical report, IUI Working Paper, 1982.

A. E. Locke, B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers, F. R. Day, C. Powell, S. Vedantam, M. L. Buchkovich, J. Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538): 197–206, 2015.

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.