Mendelian Randomization - Comparison of Different Estimators

Lakshika Ruberu
STAT 6390 - Causal Inference

Table of Contents

1 Introduction

2 Estimators

3 Simulation Study

4 Application

What is Mendelian Randomization (MR) ?

- Ideally, evaluate whether a exposure of interest (X) causes Outcome (Disease) (Y) using randomized control trials - Not always feasible
- Rely on observational studies using standard analytical methods

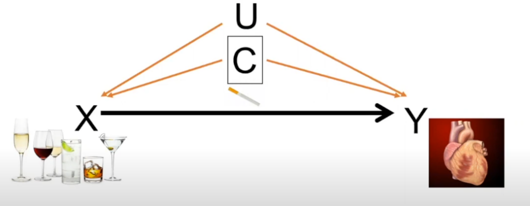


Figure: DAG to show effect of alcohol consumption on CVD
source: https://www.youtube.com/watch?v=PKHiOKgLmWk

What is Mendelian Randomization (MR) ?

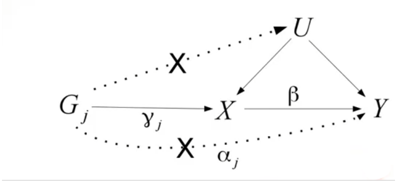
- MR overcomes these issues of confounding and bias in non trial designs using genetic information
- Participants are allocated to different exposure levels due to their genetic liability -Randomized at conception!!
- People with a genetic variant are in exposure group and without are in control group.
- If the genetic variant is associated with the exposure of interest, we can look for outcomes that co-vary with its presence or absence
- We assume that the genetic variants do not associate with the confounders in the environment

What is Mendelian Randomization (MR) ?

- Genetic Variants - Single Nucleotide Polymorphisms (SNP)
- There are SNPs found to be associated with alcohol consumption. Place individuals into drinking somewhat more or somewhat less alcohol.
- These SNPs are not associated with CVD except through exposure
- Then we look at the effect of alcohol on Cardio vascular disease

Instrument Variables

The basic premise of Mendelian randomization relies on genetic variants that explain variation in the exposure, but do not affect the disease outcome except possibly through the exposure. Such genetic variants are known as instrumental variables (IV)..



1. G_j must be associated with the exposure
2. G_j must not be associated with the confounder U
3. G_j is independent of Y apart from through the exposure (No horizontal plietropy $\implies \alpha_j = 0$)

Challenges

- Small sample sizes. Only a handful of genetic variants were used in early analysis.

- Overcome by the proliferation of GWAS whose summary data comprises of beta-coefficients and standard errors from regression of the trait of interest (either exposure or outcome) on each SNP - Increased power.

- Large number of genetic variants would imply that at least some of those would be associated with Y (horizontal plietropy).

- Various methods have been introduced that produce unbiased causal estimates robust to horizontal pleiotropy.

Notation

- Number of Genetic Variants: J
- Genetic Variants: $G1, \ldots, G_J$
- continuous exposure X and continuous outcome Y.
- All confounding variables are subsumed into U

$$X|G_j = \gamma_0 + \gamma_j G_j + \epsilon_{X_j}$$

$$Y|G_j = \Gamma_0 + \Gamma_j G_j + \epsilon_{Y_j}$$

- Causal effect of j^{th} variant is $\beta_j = \dfrac{\Gamma_j}{\gamma_j}$

Inverse-Variance Weighted Method

- There is one underlying true effect
- All deviations from the true effect are due to chance

$$\hat{\beta}_{IVW} = \frac{\sum_j w_j \times \hat{\beta}_j}{\sum_j w_j} \text{ where } \quad w_j = \frac{1}{var(\beta_j)}$$

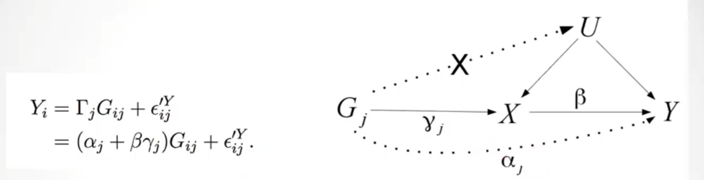
$$se(\hat{\beta}_{IVW}) = \sqrt{\frac{1}{\sum_j w_j}}$$

where $var(\beta_i) = \hat{\gamma}_j^2 \sigma_{Y_j}^2$ with $\sigma_{Y_j}^2$ the standard error of Γ_j

- For this estimator to be consistent, all of the analyzed genetic variants should be valid - No horizontal pleiotropy

MR Egger Regression

- Used in two aspects.
  1. Identify directional pleiotropy (horizontal pleiotropy in data that biases the corresponding causal estimate)
  2. In the presence of pleiotropy it provides a less biased causal estimate.



$$\begin{aligned} Y_i &= \Gamma_j G_{ij} + \epsilon'^{Y}_{ij} \\ &= (\alpha_j + \beta\gamma_j)G_{ij} + \epsilon'^{Y}_{ij}. \end{aligned}$$

- However, lacks power
- Relies on INSIDE (Instrument strength is independent of direct effect) assumption.

MR Egger Regression

- Performs a weighted linear regression of the gene-outcome coefficients Γ_j on the gene-exposure coefficients γ_j

$$\hat{\Gamma}_j = \beta_{0E} + \beta_E \hat{\gamma}_j$$

- Weights in the regression are the inverse variances of the gene-outcome associations $1/\sigma^2_{Y_j}$.
- The value of the intercept term $\hat{\beta}_{0E}$ can be interpreted as an estimate of the average pleiotropic effect across the genetic variants
- Provides an estimate for the true causal effect $\hat{\beta}E$ that is consistent even if all genetic variants are invalid due to violation of 3 given INSIDE assumption holds.

Simple Median Estimator

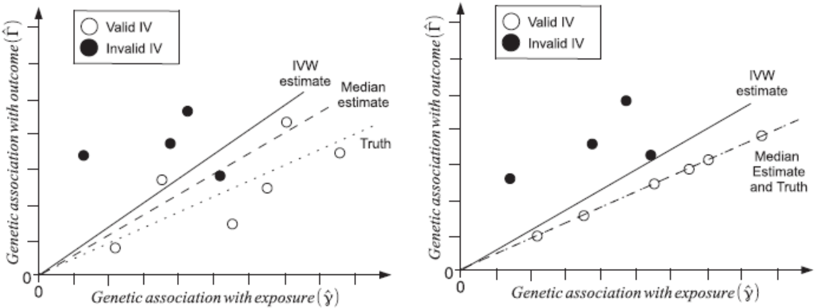


Figure: Finite sample Infinite Sample

Weighted Median Estimator

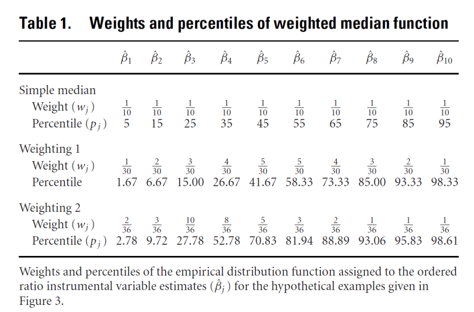


Table 1. Weights and percentiles of weighted median function

	$\hat\beta_1$	$\hat\beta_2$	$\hat\beta_3$	$\hat\beta_4$	$\hat\beta_5$	$\hat\beta_6$	$\hat\beta_7$	$\hat\beta_8$	$\hat\beta_9$	$\hat\beta_{10}$
Simple median										
Weight (w_j)	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$
Percentile (p_j)	5	15	25	35	45	55	65	75	85	95
Weighting 1										
Weight (w_j)	$\frac{1}{30}$	$\frac{2}{30}$	$\frac{3}{30}$	$\frac{4}{30}$	$\frac{5}{30}$	$\frac{5}{30}$	$\frac{4}{30}$	$\frac{3}{30}$	$\frac{2}{30}$	$\frac{1}{30}$
Percentile	1.67	6.67	15.00	26.67	41.67	58.33	73.33	85.00	93.33	98.33
Weighting 2										
Weight (w_j)	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{5}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$
Percentile (p_j)	2.78	9.72	27.78	52.78	70.83	81.94	88.89	93.06	95.83	98.61

Weights and percentiles of the empirical distribution function assigned to the ordered ratio instrumental variable estimates ($\hat\beta_j$) for the hypothetical examples given in Figure 3.

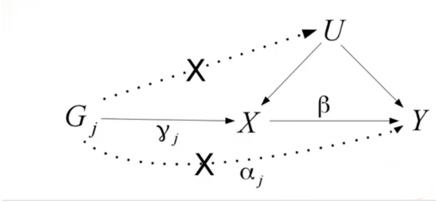
- Use the inverse of the variance of the ratio estimates as weights. $w_j = \frac{\hat\gamma_j}{\sigma^2_{Yj}}$
- $p_j = 100(s_j – w_j/2)^{th}$ percentile where $sj = \sum_{k=1}^{j} w_k$.
- SNPs that provide outlying causal effects can be downweighed $\longrightarrow$ Penalized weighted median

Data Generation

$$U_i = \sum_{j=1}^{J} \phi_j G_{ij} + \epsilon_i^U$$

$$X_i = \sum_{j=1}^{J} \gamma_j G_{ij} + U_i + \epsilon_i^X$$

$$Y_i = \sum_{j=1}^{J} \alpha_j G_{ij} + \beta X_i + U_i + \epsilon_i^Y$$



- G_{ij} is from a trinomial distribution with $p = (0.49, 0.42, 0.09)$
- Error terms were each drawn independently from standard normal distributions
- The genetic effects on the exposure $\gamma_j \sim U(0.03, 0.1)$
- ϕ_j and α_j plietropic effect (set to zero if the genetic variant was a valid instrumental variable.)

Simulation Settings

- Three Scenarios
  1. Balanced pleiotropy: InSIDE assumption satisfied— pleiotropic effects are equally likely to be positive as negative, these effects are uncorrelated with the instrument strength. $\alpha_j \sim U(-0.2, 0.2), \phi_j = 0$
  2. Directional pleiotropy, InSIDE assumption satisfied— only positive pleiotropic effects are simulated, these effects are uncorrelated with the instrument strength. $\alpha_j \sim U(0, 0.2), \phi_j = 0$
  3. Directional pleiotropy, InSIDE assumption not satisfied—pleiotropic effects are via a confounder, these effects on the outcome are therefore positive and are correlated with the instrument strength $\alpha_j \sim U(0, 0.2), \phi_j = U(-0.2, 0.2)$

Simulation Settings

- One/ two sample settings
  1. One sample: SNP exposure and SNP outcome associations are calculated on same sample
  2. Two sample: SNP exposure and SNP outcome associations are calculated on different samples
- Sample sizes – 10000/20000
- probability of being a invalid IV – 0.1/0.2/0.3
- Different causal effects – $\beta = 0$, $\beta = 0.1$
- For each setting results are based on 10000 simulated samples
- Abbreviations: inverse-variance weighted (IVW), weighted median(WM), penalized weighted median (PWM) and MR-Egger regression (EGGER)

Simulation Results

Table: **Results from simulation study in two-sample setting (Mean(SE) and Power=P for each method given different proportions of Invalid IVS–N=20000)**

	IVW		WM		PWM		Egger	
Scenario 1. Balanced pleiotropy, InSIDE assumption satisfied								
$\beta = 0$								
	Mean(SE)	P	Mean(SE)	P	Mean(SE)	P	Mean(SE)	P
0.1	0.001 (0.107)	4.9	0.001 (0.067)	3.2	0.002 (0.067)	3.4	0.000 (0.303)	6.2
0.2	0.001 (0.149)	5.8	0.001 (0.071)	4.8	0.002 (0.071)	5.1	-0.004 (0.426)	6.3
0.3	0.004 (0.183)	6.3	0.001 (0.075)	6.7	0.001 (0.076)	6.5	0.004 (0.523)	6.4
$\beta = 0.1$								
0.1	0.095 (0.108)	19.7	0.094 (0.071)	19.3	0.093 (0.071)	20.2	0.064 (0.309)	6.6
0.2	0.095 (0.150)	12.5	0.091 (0.075)	19.5	0.093 (0.075)	20.1	0.061 (0.425)	6.4
0.3	0.091 (0.183)	10	0.090 (0.079)	19.6	0.091 (0.080)	20	0.061 (0.521)	6.4

Simulation Results

	IVW		WM		PWM		Egger	
Scenario 2. Directional pleiotropy, InSIDE assumption satisfied								
$\beta = 0$								
	Mean(SE)	P	Mean(SE)	P	Mean(SE)	P	Mean(SE)	P
0.1	0.136 (0.105)	15.5	0.027 (0.068)	5.2	0.027 (0.068)	5.4	0.007 (0.296)	6.1
0.2	0.268 (0.141)	43.0	0.062 (0.072)	12.6	0.080 (0.078)	15.7	0.010 (0.395)	6.3
0.3	0.404 (0.166)	71.2	0.115 (0.080)	26.3	0.175 (0.095)	36.2	0.016 (0.464)	6.3
$\beta = 0.1$								
0.1	0.230 (0.105)	61.4	0.120 (0.071)	37.2	0.119 (0.072)	36.0	0.067 (0.298)	7.1
0.2	0.366 (0.143)	80.3	0.157 (0.077)	52.5	0.173 (0.082)	53.9	0.076 (0.401)	6.7
0.3	0.500 (0.168)	92.3	0.213 (0.086)	66.3	0.269 (0.099)	72.0	0.081 (0.469)	6.4

Simulation Results

	IVW		WM		PWM		Egger	
Scenario 3. Directional pleiotropy, InSIDE assumption not satisfied								
$\beta = 0$								
	Mean(SE)	P	Mean(SE)	P	Mean(SE)	P	Mean(SE)	P
0.1	0.189 (0.084)	53.5	0.131 (0.072)	32.4	0.059 (0.070)	13.2	0.412 (0.184)	57.5
0.2	0.327 (0.100)	81.0	0.290 (0.075)	63.8	0.176 (0.077)	40.5	0.607 (0.198)	77.1
0.3	0.427 (0.105)	93.5	0.428 (0.072)	83.9	0.321 (0.077)	68.4	0.697 (0.197)	86.9
$\beta = 0.1$								
0.1	0.285 (0.085)	81.1	0.229 (0.076)	63.8	0.153 (0.074)	47.6	0.491 (0.189)	62.1
0.2	0.423 (0.101)	93.4	0.391 (0.079)	85.0	0.274 (0.081)	71.1	0.694 (0.201)	81.2
0.3	0.525 (0.106)	98.0	0.529 (0.076)	94.7	0.420 (0.082)	87.1	0.788 (0.200)	90.2

Lipid Concentrations and Coronary Artery Disease Risk (CAD)

- Focused on association of low density (LDL-c) and high density (HDL-c) lipid concentrations on CAD
- Observational studies found
  1. LDL-c is positively associated with CAD risk (agrees with MR studies and RCT)
  2. HDL-c is inversely associated with CAD risk (contradict to MR studies and RCT)
- Use summary data from the GLGC on genetic associations with lipid fractions, and from CARDIoGRAM on associations with CAD risk, and perform the analysis methods discussed.
- use data provided by Do et al. [2013] on 185 genetic variants for this analysis

Analysis Details

- Two strategies
  1. Consider all genetic variants associated with the exposure at a genome-wide level of significance (taken as $P < 10^-8$).
  2. Restrict to genetic variants forwhich the P-value for association with the target exposure (say, HDL-c) is less than the P-values for association with the nontarget exposure(LDL-c)
- The genetic associations with the lipid fractions are in standard deviation units
- With the outcome are log odds ratios
- causal effects represent log odds ratios per 1 standard deviation increase in the lipid fraction.

Results

p-values are indicated as: * $= p < 0.05$, ** $= p < 0.01$, *** $= p < 0.0001$

	All genetic variants		Primary association with target exposure	
Analysis method	Estimate (SE)	p-value	Estimate (SE)	p-value
Low-density lipoprotein cholesterol (LDL-c)				
Inverse-variance weighted	0.482 (0.060)	***	0.472 (0.055)	***
Weighted median	0.458 (0.065)	***	0.457 (0.065)	***
Penalized weighted median	0.457 (0.063)	***	0.457 (0.067)	***
MR-Egger regression: slope	0.616 (0.103)	***	0.560 (0.094)	***
intercept	-0.009 (0.005)		-0.006 (0.005)	
High-density lipoprotein cholesterol (HDL-c)				
Inverse-variance weighted	-0.254 (0.070)	***	-0.137 (0.066)	*
Weighted median	-0.069 (0.070)		-0.066 (0.065)	
Penalized weighted median	-0.071 (0.070)		-0.064 (0.066)	
MR-Egger regression: slope	-0.013 (0.115)		0.092 (0.107)	
intercept	-0.014 (0.005)	*	-0.013 (0.005)	*

References

- Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. International journal of epidemiology, 44(2):512–525.
- Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent estima- tion in mendelian randomization with some invalid instruments using a weighted median estimator. Genetic epidemiology, 40(4):304–314.
- Do, R., Willer, C. J., Schmidt, E. M., Sengupta, S., Gao, C., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., et al. (2013). Common variants associated with plasma triglycerides and risk for coronary artery disease. Nature genetics, 45(11):1345–1352.

THANK YOU
FOR LISTENING
ANY
QUESTIONS