Project Report: Mendelian Randomization

STAT 6390.001: Introduction to Causal Inference

Thanthirige Lakshika Maduwanthi Ruberu

1 Introduction

In causal inference, often we are interested in determining whether a exposure of interest causes an outcome. Typically the outcome considered is a disease. Ideally, this should be evaluated using evidence from well conducted randomized control trials (RCT). However, conducting such trials might not be feasible at times where manipulating the exposure of interest is not practical or ethical. Therefore, we rely upon standard analytical methods using observational data. However, there might be other risk factors besides the exposure of interest that effect the likelihood of a disease, and also varies between different groups of exposure (confounding variables). If such variables are not properly accounted for, we might see spurious associations between the exposure of interest and outcome. There is also bias introduced by unmeasured confounding variables.

Mendelian Randomization (MR) overcomes these issues of confounding and other issues of reverse causation in non-trial designs using genetic information providing stronger ability for causal inference. Here participants are allocated to different exposure levels due to their genetic liability which is randomized at conception. For an example, people with a given genetic variant is allocated to the treatment group and others to the control group. This is equivalent to the random assignment of individuals to treatment and control groups in RCT. In fact, because of this reason MR is sometimes referred to as nature's RCT. If the genetic variant considered is associated with the exposure of interest then we can observe outcomes that co-vary with the presence or absence of the genetic variant. We assume that the genetic variants do not associate with the confounders in the environment

Usually, the genetic variants considered in MR analysis are single nucleotide polymorphisms (SNPs). A genetic variant should satisfy three assumptions to be used effectively in a MR analysis. Such genetic variants are known as instrumental variables (IV). The three assumptions are; (1) The variant is predictive of the exposure; (2) The variant is independent of any confounding factors of the exposure—outcome association; (3) The variant is conditionally independent of the outcome given the exposure and the confounding factors (horizontal pleiotropy). 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. Therefore, identification of IVs that satisfy these assumptions is extremely important in MR.

One of the early challenges in MR was limitation of the power due to small sample sizes. Only a handful of genetic variants (each explaining a small proportion of the variance in the exposure) were used in early analysis. This was overcome by the proliferation of genome wide association studies whose summary data comprises of beta-coefficients and standard errors from separate regression exposure and outcome on each genetic variant. MR use this summary data on large number of genetic variants for analysis leading to increased power. However, a large number of genetic variants would imply that at least some of those would have direct association with the outcome which is not mediated by the exposure of interest. Therefore, various methods have been introduced that produce unbiased causal estimates robust to horizontal pleiotropy.

In this project, I will be focusing on two such methods, known as MR-Egger regression and weighted median estimator. Simulations were conducted to compare the methods under different scenarios of horizontal pleiotropy. Lastly, the two methods will be applied to estimate the causal effect of low-density lipoprotein cholesterol (LDL-c) and high-density lipoprotein cholesterol (HDL-c) on coronary artery disease risk. Weighted median estimates were substantially more precise than those from MR-Egger regression in the simulation study and applied example.

2 Existing literature

One of the earliest estimators used to do MR analysis is the inverse variance weighted estimator (IVW) (Burgess et al., 2013). The estimate can be obtained by taking a weighted mean of the variant-specific causal estimates using inverse variance weights, as in a meta-analysis. However, for this estimator to be consistent, all of the analyzed genetic variants should be valid (assumes horizontal pleiotropy does not exist).

One of the later introduced horizontal pleiotropy robust methods is the MR Egger regression (Bowden et al., 2015) which performs a weighted regression of SNP-outcome associations on SNP-exposure associations. This method is in fact useful in two aspects. Firstly, it can identify directional pleiotropy (horizontal pleiotropy in data that biases the corresponding causal estimate) in data. Secondly, in the presence of pleiotropy it provides a less biased causal estimate. However, the method has less power because it relies on an assumption called InSIDE (instrument strength is independent of direct effect). Here instrument strength is between SNP and exposure and direct effect is between SNP and the outcome. This might not always be a feasible assumption. When InSIDE assumption is satisfied, the causal effect is the slope estimated regression line. Its significance provides evidence of causal effect while the significance of the intercept term provides evidence of directional pleiotropy.

The next method which is robust to horizontal is the weighted median estimator (Bowden et al., 2016). This approach can provide a consistent estimate of the causal effect even when upto 50% of the information contributing to the analysis comes from genetic variants that are invalid IVs. As the name suggests, this estimator is an weighted median of the individual causal effects from SNPs. This method does not have to rely on InSIDE assumption mentioned above. Another median related estimator is penalized weighted median which is equivalent to weighted median when there is no causal effect heterogeneity. This estimator down weights the contribution of heterogeneous variants and may have better finite sample properties than the weighted median estimator particularly if there is directional pleiotropy.

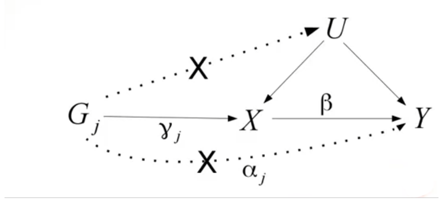

Figure 1: DAG representing the causal effect of X on Y in a MR analysis

3 Theoretical results

Through out the rest of the report we will denote the exposure of interest by X, outcome by Y, all confounding variables subsumed into single variable U and genetic variants (SNPs) used in MR analysis by $G_1, \dots G_J$, where J is the number of genetic variants used.

Figure 1 represents the DAG used to estimate causal effect of X on Y. The crosses in the paths of G_j and U and G_j and Y indicate that assumption of independence between the genetic variant and confounding factors is valid and that there is no horizontal pleiotropy. The regression equations of X and Y on SNPs on are given by the followin equations.

$$X|G_j = \gamma_0 + \gamma_j G_j + \epsilon_{X_j} \tag{1}$$

$$Y|G_j = \Gamma_0 + \Gamma_j G_j + \epsilon_{Y_j} \tag{2}$$

Causal effect of j^{th} variant is given by $\beta_j = \dfrac{\Gamma_j}{\gamma_j}$. This leads to

3.1 Inverse-Variance Weighted (IVW) Method

As mentioned above, the IVW method only gives consistent estimates of the causal effect if all the genetic variants used are valid IV. Therefore, it assumes that each β_j estimates a true underlying causal effect β and the variability between the β_js are sampling variability. The final estimate combining β_js is

$$\hat{\beta}_{IVW} = \frac{\sum_j w_j \times \hat{\beta}_j}{\sum_j w_j} \quad \text{where} \quad w_j = \frac{1}{var(\beta_j)} \tag{3}$$

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

where $var(\beta_i) = \hat{\gamma}_j^2 \sigma_{Y_j}^2$ with $\sigma_{Y_j}^2$ the standard error of Γ_j. As evident by equation 3, β_js with less variability contribute more, to the final estimate. But this method is said to have a 0% breakdown level as all 100% of the genetic variants used in its estimation must be valid.

3.2 MR-Egger Regression Method

MR Egger Regression assumes horizontal pleiotropy. Therefore the regression of Y on SNPs earlier given by equation 2, now changes to

$$Y|G_j = \Gamma_0 + (\gamma_j \beta_j + \alpha_j) G_j + \epsilon_{Y_j} \tag{4}$$

Then the following weighted linear regression of the gene-outcome coefficients Γ_j on the gene-exposure coefficients γ_j is carried out

$$\hat{\Gamma}_j = \beta_{0E} + \beta_E \hat{\gamma}_j \tag{5}$$

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

3.3 Weighted Median Estimator (WM)

I will introduce the simple median estimator as a preface to the WM estimator. It is simply the median of the individual causal effects β_j. Simple median estimator can provide consistent estimates of causal effect even when 50% of IV s used in MR analysis is invalid. However this is an inefficient estimator, especially when the precision of the individual estimates varies considerably. A weighted median is defined as follows to account for this.

1. Let w_j be the weight given to the j^{th} ordered β_j with $w_j = \frac{\hat{\gamma}_j}{\sigma^2_{Yj}}$

2. compute $s_j = \sum_{k=1}^{j} w_k$, the sum of weights up to and including the weight of the j^{th} ordered β_j

3. compute percentiles $p_j = 100(s_j - \frac{w_j}{2})$

4. The weighted median estimator is the β_j when $p_j = 50$ which is obtained through a linear extrapolation between neighboring β_js which contain 50^{th} percentile.

The simple median estimator can be thought of as a weighted median estimator with equal weights. The simple median provides a consistent estimate of causal effect if at least 50% of IVs are valid, whereas the weighted median will provide a consistent estimate if at least 50% of the weight comes from valid IVs. Here, it is assumed that no single IV contributes more than 50% of the weight. Because otherwise, the 50% validity assumption is just equivalent to assuming that this IV is valid (in which case, an analysis should simply be based on this one IV).

3.4 Penalized Median Estimator (PWM)

Even though, the Invalid IVs do not contribute directly to the causal estimate in WM method they do influence it. This could be minimized by down weighting the contribution to the MR analysis of genetic variants with outlying (heterogeneous) ratio estimates. The heterogeneity could be quantified by Cochran's Q statistic.

$$Q = \sum_j Q_j = \sum_j w'_j(\hat{\beta}_j - \hat{\beta})^2$$

Here, $\hat{\beta}$ is the IVW estimate (Greco M et al., 2015).

Under, the null hypothesis that all genetic variants are valid IVs and the same causal effect is identified by all variants, the Q statistic has a chi-squared distribution on J–1 degrees of freedom (df) and Q_j follow a a chi-squared distribution with 1 df. A penalization is proposed with $w^*_j = w'_j \times \min(1, 20q_j)$ where q_j the one-sided upper p-value of a chi-squared distribution with 1 df corresponding to Q_j. If the p-value is greater than 0.05 the weights will not be penalized. However, the weights of outlying variants ($P < 0.05$) will be severely penalized.

As a summary of all the methods, all four methods require that the genetic variants must be associated with the exposure of interest. The association is typically identified by a p value less that 510^{-8}. IVW method has a 0% breakdown assuming no horizontal plietropy in 100% of genetic variants used. MR-Egger regression can give consistent estimates when 100% of genetic variants are invalid IVs given that INSIDE assumption is satisfied, whereas the weighted median method requires 50% of the weight to come from valid IVs. However, the weighted median approach allows the IV assumptions to be violated in a more general way for the invalid IVs, whereas MR-Egger regression replaces one set of untestable assumptions (2 and 3) with a weaker, but still untestable assumption (the InSIDE assumption).

4 Simulation setting

To compare the performance of the three methods (IVW, weighted median estimator, and MR Egger) in realistic settings a simulation study is conducted based on the settings from Bowden et al. (2016).

The data were generated using the following equations.

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

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

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

Here, G_{ij} is from a trinomial distribution with $p = (0.49, 0.42, 0.09)$. This is equivalent to a single nucleotide polymorphism with minor allele frequency 0.3. 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 were set to zero if the genetic variant was a valid instrumental variable.

25 genetic variants are assumed to be candidate IVs and three scenarios are considered:

1. In Scenario 1 (balanced pleiotropy, InSIDE satisfied), the α_j parameter was drawn from a uniform distribution between -0.2 and 0.2.

2. In Scenario 2 (directional pleiotropy, InSIDE satisfied), the α_j parameter was drawn from a uniform distribution between 0 and 0.2.

3. In Scenario 3 (directional pleiotropy, InSIDE not satisfied), the ϕ_j parameter was drawn from a uniform distribution between -0.2 and 0.2.

The validity status of a genetic variant is determined by a random draw for each variant. where probability of being an invalid variant is varied in 3 scenarios with probabilities taken as 0.1, 0.2, and 0.3. I only considered the case with 20000 participants in a two sample setting where causal association between SNPs and Y is measured on one sample and those between SNPs and X are measured on a separate sample. 10,000 simulated datasets are generated for each scenario with two values of the causal effect ($\beta = 0$, null causal effect; $\beta = 0.1$, positive causal effect).

4.1 Simulation Results

Table 1 and 2 provides the results for simulations under null and positive causal effects. For Scenario 1, all he four methods have estimates close to zero in Table 1 and close to 0.1 in Table 2 indicating that the estimates from the four methods are unbiased. Further, the power to detect causal effect is reasonable under null causal effect for the four methods (power under the null is equal to Type 1 error). However, under positive causal effect, the power of the estimates differ substantially. The standard errors under weighted median methods are lower compared to IVW method, and generally have greater power with a positive causal effect, particularly as the proportion of invalid IVs increases. MR EGGER regression mehod has the highest standard error and it lacks power under positive causal effect.

	IVW		WM		PWM		Egger	
	Mean(SE)	P	Mean(SE)	P	Mean(SE)	P	Mean(SE)	P
Scenario 1. Balanced pleiotropy, InSIDE assumption satisfied								
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
Scenario 2. Directional pleiotropy, InSIDE assumption satisfied								
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
Scenario 3. Directional pleiotropy, InSIDE assumption not satisfied								
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

Table 1: Mean estimates, mean standard errors, and power(P) of 95% confidence interval to reject null hypothesis of inverse-variance weighted (IVW), weighted median (WM/ PWM), and MR-Egger(Egger) regression methods in simulation study for two-sample Mendelian randomization with a null ($\beta = 0$) causal effect.

	IVW		WM		PWM		Egger	
	Mean(SE)	P	Mean(SE)	P	Mean(SE)	P	Mean(SE)	P
Scenario 1. Balanced pleiotropy, InSIDE assumption satisfied								
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
Scenario 2. Directional pleiotropy, InSIDE assumption satisfied								
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
Scenario 3. Directional pleiotropy, InSIDE assumption not satisfied								
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

Table 2: Mean estimates, mean standard errors, and power(P) of 95% confidence interval to reject null hypothesis of inverse-variance weighted (IVW), weighted median (WM/ PWM), and MR-Egger(Egger) regression methods in simulation study for two-sample Mendelian randomization with a positive ($\beta = 0.1$) causal effect.

In Scenario 2 with INSIDE assumption satisfied, estimates from the IVW method are biased with higher Type 1 error rates under both null and positive causal effects. The weighted median methods provide less biased estimates under both causal effect. Under null causal effect, nominal Type 1 error rates are observed when only 10% of genetic variants are invalid IVs. However Type 1 error rates are far lower for the median-based methods than those from the IVW method. Under positive causal effect, although power to detect a causal effect is greater in the IVW method, this is achieved at the cost of the Type 1 error rate. Estimates from MR-Egger regression are close to being unbiased under both causal effects. Moreover, under null causal effect Type 1 error rates are at nominal levels.However, they are less precise and under positive causal effect it lacks power.

In Scenario 3 with INSIDE assumption not satisfied all methods suffer from bias and inflated Type 1 error rates. Results from MR EGGER regression are more biased compared to IVW method with simillar type 1 error rates. The weighted median methods give lower Type 1 error rates. MR Egger method lacks power and estimates are less precise as observed in other scenarios.

5 A real example: Lipid Concentrations and Coronary Artery Disease (CAD) Risk

Following Bowden et al. (2016) I use summary data on 185 common variants recently mapped for plasma lipids to examine the role of LDL-c and HDL-c in risk for CAD (Do et al., 2013). The data contains estimates of the effect of each SNP on plasma triglyceride, LDL-C and HDL-c levels in a sample exceeding 180,000 individuals (glo, 2013); and estimates of the effect of each SNP on CAD in a sample exceeding 86,000 individuals (Schunkert et al., 2011). The genetic associations with the lipid fractions are in standard deviation units, and with the outcome are log odds ratios, so the causal effects represent log odds ratios per 1 standard deviation increase in the lipid fraction. I will be exploring two causal effects. First, causal effect of LDL-c on CAD and second, causal effect of LDL-c on CAD. Two strategies will be

			Primary association	
	All genetic variants		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)	*

Table 3: Results; p-values are indicated as: * $= p < 0.05$, ** $= p < 0.01$, *** $= p < 0.0001$

explored.

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)

5.1 Application Results

Results of the application are given on Table 3. The causal effects of LDL-c detected by all the analysis methods. However, results from both MR-Egger regression and the weighted median methods suggest that there is null causal effect for HDL-c. This is contradicted by the results of the IVW method.

MR Egger regression method suggests directional pleiotropy for HDL-c, but not for LDL-c. Simillar to simulation results, The median-based methods were consistently.

6 Conclusion and discussion

Here, I discussed several estimators used to do MR analysis. Inverse variance weighted median estimators, weighted median estimator and MR Egger regression. The latter two methods are robust to horizontal pleiotropy. I have shown how the different methods performs in a simulation study and in an applied example.

Simulations results suggest that weighted median methods could be used as as a sensitivity analysis method when InSIDE is not satisfied, as well as when it is satisfied. This is because MR Egger method provides concerning estimates when InSIDE assumption is not satisfied where is weighted median estimates are consistent in both cases.

However, the methods discussed here do not address selection bias which could arise from different sources. One such source is differential selection of individuals in the datasets from which the genetic associations are obtained. Another is the selection of genetic variants based on their strength in the dataset under analysis, or if genetic variants were discovered as associated with the exposure in the dataset under analysis. Selection bias may affect all genetic variants in a particular analysis and therefore not likely to be addressed by the methods discussed.

References

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

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 estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic epidemiology*, 40(4):304–314.

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic epidemiology*, 37(7):658–665.

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.

Greco M, F. D., Minelli, C., Sheehan, N. A., and Thompson, J. R. (2015). Detecting pleiotropy in mendelian randomisation studies with summary data and a continuous outcome. *Statistics in medicine*, 34(21):2926–2940.

Schunkert, H., König, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., Preuss, M., Stewart, A. F., Barbalic, M., Gieger, C., et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics*, 43(4):333–338.