Project Proposal: Mendelian Randomization

STAT 6390.001: Introduction to Causal Inference

Thanthirige Lakshika Maduwanthi Ruberu

1 Introduction

Determining whether a exposure to a certain factor causes a disease is a crucial area in causal inference. Ideally, this should be evaluated using evidence from well conducted randomized control trials (RCT). However, conducting RCT 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 which makes the analysis susceptible to various types of biases that arise from 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. In fact, MR is sometimes referred to as nature's RCT. Here participants are allocated to different exposure levels due to their genetic variations(SNPs) which is randomized at conception. If a SNP is associated with the exposure of interest then we can observe outcomes that co-vary with the presence or absence of the genetic variant. Such SNPs are known as instrument variables (IVs)

For a given genetic variant to be considered as a IV, three assumptions should hold.(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 plausibility of the above mentioned assumptions determines the validity of a causal conclusion from a Mendelian randomization analysis. Therefore, identification of IVs that satisfy this assumption 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 regression of the trait of interest (either risk factor or 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 horizontal pleiotropy with the outcome. 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 will be conducted to compare the methods under different scenarios. 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.

2 Existing literature

The inverse variant estimator (IVW) is one of the methods to estimate causal effect of the exposure on the outcome (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.

One of the notable differences among MR Egger and weighted median estimator is that MR-Egger regression can give consistent estimates when 100% of genetic variants are invalid IVs, 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).

3 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).25 genetic variants are assumed to be candidate IVs and three scenarios are considered: (1) Balanced pleiotropy - pleiotropic effects are equally likely to be positive as negative; (2) Directional pleiotropy - only positive pleiotropic effects are simulated; (3) Directional pleiotropy - pleiotropic effects are via a confounder. Here, InSIDE assumption is only valid in scenarios (1) and (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. Further, cases with 10,000 and 20,000 participants are considered. 10,000 simulated datasets are generated for each scenario in a two-sample setting with two values of the causal effect ($\beta = 0$, null causal effect; $\beta = 0.1$, positive causal effect);

4 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. The analysis the analysis will be restricted to genetic variants for which the P-value for association with the target exposure (say, HDL-c) is less than the P-values for association with the nontarget exposures (say, LDL-c and triglycerides).

5 Timeline

- 10/17 – 10/24, read the selected papers;
- 10/25 – 11/05, run the simulations;

- 11/06 – 11/14, analyze real data sets;
- 11/15 – 11/20, prepare the presentation;
- 11/20 – 12/09, write final report.

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.

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.