MR Maths

June 5, 2025

1 Introduction

Premise of MR MR stems from the fact that observed correlations in epidemiology do not automatically imply a causal association between an exposure and a disease, as this correlation may be due to unobserved confounders, or reverse causation (the outcome causes the exposure, not vice verse). The main premise of MR are instrumental variables (IV), that is genetic variants that are associated with an exposure, but do not affect the outcome except possibly through the exposure. Grouping individuals according to their allele dosage (i.e. 0, 1, or 2), can be thought of as a proxy for the treatment group in a randomized controlled trial: individuals with differing numbers of alleles of an IV can be thought of as receiving a different average level of exposure throughout their life. If a person's genetic variant is independent of all factors except that of exposure, then the differences in outcome between genetic subgroups can be causally attributed to the exposure.

1.1 Assumptions of instrumental variables (IV)

A genetic variant is a valid instrument if it satisfies:

1. **Relevance (IV1):** the variant is predictive of the exposure.
2. **Independence (IV2):** the variant is independent of any confounders of the exposure-outcome association. By Mendel's laws, germline variants are unlinked to environmental confounders (barring population stratification).
3. **Exclusion restriction (IV3):** the variant is conditionally independent of the outcome given the exposure and the confounding factors.

1.2 General modeling framework

Consider an MR analysis with J genetic variants $G_1, \ldots, G_J$, where $G_j \in \{0, 1, 2\}$ encodes the allele dosage, and where all variants are uncorrelated. Let X be a continuous exposure and Y a continuous outcome, and let U represent all confounding variables. We assume that all associations are linear, which yields the following structural equations model for variant G_j:

$$X = \gamma_0 + \gamma_j G_j + \epsilon_{Xj}, \tag{1}$$

$$Y = \Gamma_0 + \Gamma_j G_j + \epsilon_{Yj}, \tag{2}$$

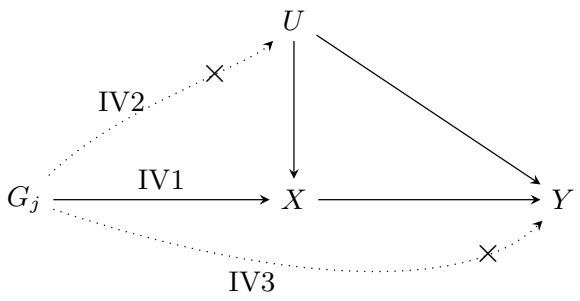

Figure 1: Enter Caption

where γ_j is variant-exposure association, and Γ_j is the variant-outcome association. Exposure and outcome confounders effects are modeled via the error terms ϵ_{Xj} and ϵ_{Yj}, respectively. The standard MR model assumes $\epsilon_{Xj}, \epsilon_{Yj}$ to be normal with zero mean and uncorrelated with the genetic variant.We can decompose the association between genetic variant and outcome into a sum of a direct (pleiotropic) effect and an indirect (causal) effect:

$$\Gamma_j = \gamma_j \beta + \alpha_j \tag{3}$$

where α_j is the effect of the genetic variant on the outcome that is not mediated via the exposure, and β is the causal effect of the exposure on the outcome. For the variant to be a valid IV, assumption IV1 requires that $\gamma_j \neq 0$, while assumptions IV2 and IV3 require no pleiotropy, i.e. $\alpha_j = 0$.

TODO: Mention the most common point-estimate assumptions (homogeneity etc). Also, Under IV1–IV3, the ratio $\gamma^{-1} \operatorname{Cov}(Z, Y)$ consistently estimates β.

TODO: Mention gene-environment equivalence assumption

1.3 Weak instruments bias

When instruments are "weak," the sampling distribution of the MR estimate is pulled toward the ordinary least squares (OLS) estimate of the exposure-outcome association rather than the true causal effect.

TODO: More details

1.4 Overview of approaches

Different MR approaches vary by sample design, instrument choice, and data type. In one-sample MR, SNP–exposure and SNP–outcome associations come from individual-level data; two-sample MR uses summary statistics from separate cohorts. Single-SNP MR relies on a single variant, while allele-score MR combines multiple SNPs into a genetic risk score. Individual-level analyses allow flexible modeling (e.g., covariate adjustment), whereas summary-statistic MR enables large-scale two-sample designs. Single-SNP methods suit strong instruments, whereas allele-scores boost power but require SNP independence.

2 Estimating causal effects with individual-level data

2.1 Allele scores

When using individual-level data in MR estimation, genetic variants can either be used as separate instruments or combined into an allele score. An allele score is generated by adding up the number of risk-increasing alleles for all the variants selected as instruments. This score can be *unweighted*, so that each SNP makes the same contribution, or *weighted*, so that the count of risk-increasing alleles at each SNP is multiplied by the estimated effect of that SNP on the exposure. Weighted scores can provide increased instrument strength and statistical power.

2.2 Two-stage least-squares (2SLS) estimation

A common way of estimating causal effects with individual level data is done via two-stage least-square (2SLS) regression, which uses genetic variants to first predict the value of the exposure, and then estimates the causal effect on the outcome. Using the notation for the variant-exposure association similar to 2, the first regression can be written as

$$X = \gamma_0 + \gamma_j G_j + \epsilon_{Xj} \tag{4}$$

and estimates $\hat{\gamma}_j$ via OLS such that

$$\hat{X} = E[X|G_j] = \hat{\gamma}_0 + \hat{\gamma}_j G_j \tag{5}$$

In the second stage the outcome Y is regressed on the predicted value of the exposure $\widehat{X}$ which estimates $\hat{\beta}$ such that

$$E[Y|\hat{X}] = \hat{\beta}_0 + \hat{X}\hat{\beta} \tag{6}$$

Under IV assumptions, $\hat{\beta}$ then yields the estimated exposure-outcome association.

The 2SLS estimation is most commonly performed on multiple genetic variants simultaneously, which extends 4 to

$$X = \gamma_0 + \sum_{j=1}^{J} \pi_j G_j + \epsilon_{Xj} \tag{7}$$

while the rest of the approach stays the same.

2.3 Strong vs. weak instruments

A genetic variant G_j is *strong* if its association with the exposure X is large enough that the first-stage regression yields an F-statistic

$$F \;=\; \frac{\hat{\gamma_j}^2}{\mathrm{Var}(\hat{\gamma}_j)} \;>\; 10,$$

equivalently, $\hat{\gamma}_j$ is far from zero and explains a nontrivial fraction of $\mathrm{Var}(X)$. A *weak* instrument has $\hat{\gamma}_j \approx 0$, implying the resulting IV estimator will be biased and imprecise.

3 Estimating causal effects with summary-level statistics

3.1 Inverse-variance weighted (IVW) estimate

The IVW estimator is one of the most commonly used causal effect estimation methods on summary-level data.

3.1.1 Derivation via meta-analysis of Wald ratios

Under the standard IV assumptions plus the IVW homogeneity assumption, we posit that there is a single true causal parameter β such that

$$\Gamma_j \;=\; \beta\,\gamma_j \quad \text{for all } j. \tag{8}$$

In the context of meta-analysis this is referred to as "fixed-effect model". Equivalently, each SNP j yields a Wald ratio estimate

$$\widehat{\beta}_j \;=\; \frac{\widehat{\Gamma}_j}{\widehat{\gamma}_j},$$

Hence each $\hat{\beta}_j$ "targets" the same β, and any differences among the $\hat{\beta}_j$ arise solely from sampling variability in $\hat{\Gamma}_j$ and $\hat{\gamma}_j$. The inverse-variance weighting estimation is a meta-analysis of these SNP-specific Wald ratios $\hat{\beta}_j$. The IVW estimator $\widehat{\beta}_{\mathrm{IVW}}$ is computed as

$$\widehat{\beta}_{\mathrm{IVW}} \;=\; \frac{\displaystyle\sum_{j=1}^{L} w_j\,\hat{\beta}_j}{\displaystyle\sum_{j=1}^{L} w_j}, \tag{9}$$

where

$$w_j \;:=\; \frac{1}{\mathrm{Var}(\hat{\beta}_j)} \;\approx\; \frac{\hat{\gamma}_j^2}{\mathrm{Var}\,(\hat{\Gamma}_j)}, \tag{10}$$

with the approximation arising from a first-order Taylor expansion (Delta method). Then $\widehat{\beta}_{IVW}$ is an unbiased estimator for the common causal effect β.

3.1.2 Derivation via weighted least squares (WLS)

The IVW estimate can be equivalently obtained from linear regression of the genetic associations with the outcome ($\hat{\Gamma}_j$) on the genetic assocation with the exposure ($\hat{\gamma}_j$)

$$\widehat{\Gamma_j} = \beta\,\hat{\gamma}_j + \nu_j \tag{11}$$

Because $\nu_j \sim \mathcal{N}(0, \mathrm{Var}\,(\hat{\Gamma}_j))$ is heteroskedastic, a weighted leas-square regression is performed with inverse-variance weights as defined by

$$v_j := \frac{1}{\mathrm{Var}\,(\hat{\Gamma}_j)} \tag{12}$$

The WLS objective is

$$Q(\beta) \;=\; \sum_{j=1}^{J} v_j \left(\hat{\Gamma}_j - \beta\,\hat{\gamma}_j\right)^2.$$

Minimizing with respect to β yields (9).

Heterogeneity testing (Cochran's Q)

A core assumption of IVW estimates is the "homogeneity" of the genetic-variants specific causal estimates, i.e. any differences among the $\hat{\beta}_j$ arise as a result of sampling variability ("fixed-effect model"). In contrast, heterogeneity of the causal estimates makes the variants invalid IVs. Cochran's Q statistic can be used to check for between-variant heterogeneity. Between-variant heterogeneity in individual ratios $\widehat{\beta}_j = \widehat{\Gamma}_j/\widehat{\gamma}_j$ is quantified by

$$Q \;=\; \sum_{j=1}^{J} w_j \left(\widehat{\beta}_j - \widehat{\beta}_{\mathrm{IVW}}\right)^2, \quad w_j = \frac{\widehat{\gamma}_j^2}{\mathrm{Var}\,(\Gamma_j)}.$$

Under homogeneity, $Q \sim \chi^2_{J-1}$; large Q indicates that variant specific causal estimates differ more than by chance. Outliers and influential points, which may be associated with invalid IVs, can be identified using studentized residuals and Cook's distances in the WLS IVW model.

3.2 MR-Egger method

MR-Egger generalizes the IVW estimation by allowing directional pleiotropic effects. Recall that the fixed-effects IVW model assumes a zero intercept. Instead, MR-Egger introduces an additional parameter θ_{0E}, yielding

$$\widehat{\Gamma}_j \;=\; \alpha \;+\; \beta\,\widehat{\gamma}_j \;+\; \epsilon, \tag{13}$$

where β is the causal slope and α captures the average pleiotropic effect across SNPs. For the MR-Egger estimate of the causal effect to be asymptotically consistent under pleiotropy ($\alpha \neq 0$), we require that the pleiotropic effects α_j are independent from the exposure causal effects β_j. This is referred to as the InSIDE assumption (INstrument Strength Independent of Direct Effect). When it holds, the intercept from the MR-Egger analysis can be interpreted as the average pleiotropic effect of a genetic variant, that is the weighted mean of the α_j using the inverse variance weights. The InSIDE assumption is likely to be satisfied if pleiotropic effects on the outcome are direct (i.e., not via a confounder). Testing the intercept from the MR-Egger analysis provides an assessment of the validity of the IV3 assumption, with a nonzero intercept indicating a biased IVW estimate. Note that MR-Egger provides a consistent estimate of the causal effect even if all genetic variants are invalid due to a violation of IV3, as long as InSIDE holds.

3.3 Weighted median MR

Simliar to MR-Egger, the weighted median MR relaxes the assumption that all genetic variants are non-pleiotropic. However, instead of the InSIDE assumption, weighted median MR requires that at least 50% of the instruments are valid (no-pleiotropy) variants. The weighted-median MR estimator is motivated by robust statistics: by ordering the causal estimates $\hat{\beta}_j$ and assigning normalized weights

based on precision, one selects the $\hat{\beta}_k$ at which the cumulative weight first reaches 0.5, i.e. the median value. This construction yields a consistent causal estimate $\hat{\beta}_{WM}$ even when many invalid instruments exist, so long as invalid SNPs do not collectively account for more than half the information. Intuitively, invalid SNPs appear as outliers; the median resists influence of those extremes, ensuring robustness.

Define the *unnormalized weight* for the j-th ratio estimate by

$$\tilde{w}_j \;=\; \frac{1}{\mathrm{Var}(\hat{\beta}_j)}.$$

Normalize these weights so that they sum to unity. Specifically, set

$$w_j \;=\; \frac{\tilde{w}_j}{\sum_{i=1}^{J} \tilde{w}_i} \quad\Longrightarrow\quad \sum_{j=1}^{J} w_j \;=\; 1.$$

Treat $\{w_j\}$ as a discrete probability mass on the ordered ratio estimates

$$\hat{\beta}_{(1)} \;\le\; \hat{\beta}_{(2)} \;\le\; \cdots \;\le\; \hat{\beta}_{(J)}. \tag{14}$$

Define the cumulative sum of weights up to index k by

$$s_k \;=\; \sum_{j=1}^{k} w_{(j)}, \qquad k = 1, \ldots, J.$$

In particular, let $s_J = \sum_{j=1}^{J} w_{(j)} = 1$.

The weighted-median estimate $\hat{\beta}_{\mathrm{WM}}$ is defined to be the statistic $\hat{\beta}_{(k)}$ for which the cumulative weight first exceeds or equals 0.5. Equivalently, let

$$k^* \;=\; \min\bigl\{\, k : s_k \;\ge\; 0.5 \bigr\}.$$

Then

$$\hat{\beta}_{\mathrm{WM}} \;=\; \hat{\beta}_{(\,k^*)}. \tag{15}$$

In other words, $\hat{\beta}_{\mathrm{WM}}$ is the smallest ordered ratio estimate whose cumulative weight s_{k^*} satisfies $s_{k^*-1} < 0.5 \le s_{k^*}$. If $s_{k^*} = 0.5$ exactly, one may take $\hat{\beta}_{(k^*)}$ (or average with $\hat{\beta}_{(k^*+1)}$), but exact ties are rarely encountered in practice.

Under standard MR assumptions (relevance, independence, exclusion restriction) but allowing for horizontal pleiotropy in some instruments, the weighted-median estimator $\hat{\beta}_{\mathrm{WM}}$ is *consistent* for the true causal effect β provided that at least 50 % of the total weight derives from valid instruments.

3.4 Mode-based estimate

The Mode-based estimate (MBE) is another approach designed to provide a consistent causal effect estimate even when some instruments are invalid due to horizontal pleiotropy. The core idea behind MBE is that if the SNP-specific causal estimates ($\hat{\beta}_j$) are grouped based on their true underlying causal effect, the largest group (i.e., the mode of the distribution of these estimates) will correspond to the true causal effect, assuming that the largest number of individual variants estimate the same true causal effect.

This method relies on the "ZEro Modal Pleiotropy Assumption" (ZEMPA), which states that the mode of the pleiotropic effects is zero. The MBE can be derived using the Wald ratios $\hat{\beta}_j = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j}$. The distribution of these estimates is considered, and the mode of this distribution is taken as the causal effect estimate $\hat{\beta}_{MBE}$. In practice, this often involves kernel density estimation to find the mode of the weighted distribution of $\hat{\beta}_j$. The weights can be defined similarly to those in the IVW method.

References