Data generating process underlying causal inference using Mendelian randomization

Gibran Hemani

2020-03-20

Background

Causal inference between two traits, the exposure's (x) effect on the outcome (y) can be made using associations of genetic variants g on x and y. This method is known as Mendelian randomization (MR), a special case of instrumental variable (IV) analysis where the instrument is a genetic variant. Assume the following causal structure:

Our objective is to estimate β_1 the causal effect of the exposure on the outcome.

The exposure x is influenced by a genetic variant g and an unmeasured confounder u:

$$x_i = b_0 + b_1 g_i + b_2 u_i + e_i$$

where $g \sim Binom(2, p)$ and p is the allele frequency; $u \sim N(\mu_u, \sigma_u)$ and the residual $e \sim N(0, \sigma_e)$. The outcome y is causally influenced by the exposure and the confounder:

$$y_u = \beta_0 + \beta_1 x_i + \beta_2 u_i + \epsilon_i$$

where $\epsilon \sim N(0, \sigma_\epsilon)$. A biased estimate of the causal effect of x on y can be obtained from the observational regression of x on y as

$$\beta_{OLS} = \frac{Cov(x, y)}{Var(x)}$$

This is biased because the confounder is not accounted for. In practice it is impossible to prove that all confounders have been accounted for, so an instrumental variable (IV) approach is desirable. An unbiased causal effect estimate can be obtained using using:

$$\hat{\beta}_{IV} = \frac{\hat{\gamma}_1}{\hat{b}_1}$$

where

$$\hat{\gamma}_1 = \frac{Cov(g, y)}{Var(g)}$$

and

$$\hat{b}_1 = \frac{Cov(g, x)}{Var(g)}$$

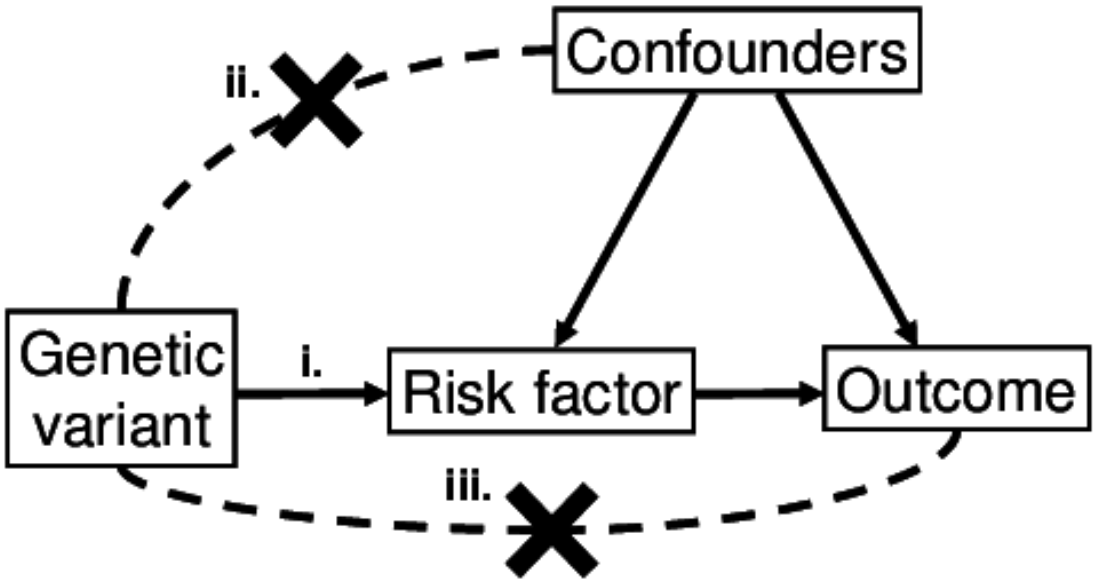

Figure 1: Diagram of instrumental variable assumptions for Mendelian randomization. The three assumptions (i, ii, iii) are illustrated by the presence of an arrow, indicating the effect of one variable on the other (assumption i), or by a dashed line with a cross, indicating that there is no direct effect of one variable on the other (assumptions ii and iii)

Example in R

Set parameters

```
beta_1 = 0.5
b_1 = 0.2
b_2 = 0.5
beta_2 = 0.5
sigma_e = 1
sigma_u = 1
sigma_epsilon = 1
sample_size = 1000000
p = 0.4
set.seed(12345)
```

Generate the data for the confounder, exposure and outcome

```
g <- rbinom(sample_size, 2, p)
u <- rnorm(sample_size, 0, sigma_u)
e <- rnorm(sample_size, 0, sigma_e)
epsilon <- rnorm(sample_size, 0, sigma_epsilon)
x <- g * b_1 + u * b_2 + e
y <- x * beta_1 + u * beta_2 + epsilon
```

Obtain the biased observational estimate

```
beta_obs <- cov(y, x) / var(x)
round(beta_obs, 2)
```

```
## [1] 0.7
```

```
beta_1
```

```
## [1] 0.5
```

Obtain the IV estimate. First get the variant-exposure effect estimate and the variant-outcome effect estimate

```
b_1_hat <- cov(x,g) / var(g)
gamma_1_hat <- cov(y,g) / var(g)
beta_iv <- gamma_1_hat / b_1_hat
round(beta_iv, 2)
```

```
## [1] 0.51
```

```
beta_1
```

```
## [1] 0.5
```