Goal of this lab

- Foundations of Mendelian randomization (this week)
- Confounder correction for molecular QTL analysis (next lab)

Contact: Yongjin Park, `ypp@mit.edu` for any comments and questions.

Mendelian randomization

Sobel's test

In strict sense, causal inference in observational studies is intractable unless we have a way to carry out additional interventional experiments. Given a set of reasonable assumptions, we will make causal effects identifiable and thus estimable from observational (retrospective) data.

Classical two-stage regression approach (a single SNP, a single mediator method)

Before the geneticists realize the value of mediation analysis, researchers from other fields, especially social science, mediation analysis has been conducted. Please read this gentle introduction written by MacKinnon, Fairchild, and Fritz (2007):

The idea is quite simple. Consider the following model:

```
G -γ-> Y
G -α-> M -β-> Y
```

or we could define this more precisely

$$M = G\alpha + \epsilon_M, \quad Y = M\beta + G\gamma + \epsilon_Y$$

- Baron and Kenny (1986) method
  i. Step 1. regression `Y ~ G γ`
  ii. Step 2. regression `M ~ G α`
  iii. Step 3. regression `Y ~ M β + G γ'` or `Y ~ M β` (without the direct effect)
  iv. Step 4. check if `M` explained away the effect of `G` on `Y`

What is a suitable test statistic? There are two options: (1) Non-zeroness of `αβ` (2) or non-zeroness of `γ - γ'`. We will use parameters α, β, γ and corresponding standard errors $\sigma_\alpha, \sigma_\beta, \sigma_\gamma$ estimated from the regression models (typically `lm` or `glm` in `R`).

Here we use $T = \alpha\beta$ as test statistic. Employing asymptotic normality we can estimate the mean and variance of T as follows using Sobel's method (1982).

$$\mathbb{E}[T] \leftarrow \hat{\alpha}\hat{\beta}$$

and

$$\mathbb{V}[T] \leftarrow \sqrt{\hat{\sigma}_\alpha^2 \hat{\beta}^2 + \hat{\sigma}_\beta^2 \hat{\alpha}^2}$$

If you're mathematically oriented and curious (totally optional), see this classical paper.

Lab questions

- How do you calculate p-value in the Sobel's test? Hint: Wald test.
- Using a subset of 1000 genomes genotype matrix (which we used in the previous lab; `../lab2/chr22.Rd`), we can simulate mediation data both `M` and `Y`.
  i. Generate 1 true mediating gene `M` with heritability $h_\alpha^2 > 0$, i.e., $\alpha \sim \mathcal{N}(0, h_\alpha^2)$ and $\epsilon \sim \mathcal{N}(0, 1 - h_\alpha^2)$.
  ii. Generate 99 other "false" / null genes `M0` with heritability $h_{\alpha_0}^2 = 0$.
  iii. Generate downstream phenotype `Y` as a function of true `M` to have mediated effect size $h_\beta^2 > 0$.
- Perform the Sobel's test gene by gene and report p-values.
- Repeat the same experiments with different configuration of gene and phenotype heritability parameters. And summarize your results with qq-plots.

Mendelian randomization

Suppose we have resulting parameters from the following regression models:

$$Y \sim G\Gamma + \epsilon, \quad M \sim G\alpha + \epsilon.$$

Let us assume the following conditions hold:

- IV1: The genetic variant `G` is independent of a potential confounder `U` influencing both `M` and `Y`;
- IV2: The genetic variant `G` is associated with the mediator `M`, i.e., $\alpha \neq 0$;
- IV3: The genetic variant `G` is independent of the outcome `Y` conditional on the mediator (exposure) `M` and confounders U, i.e., in the regression $Y \sim M\alpha\beta + G\gamma$ we have $\gamma = 0$.

We term a genetic variant satisfying IV1-IV3, "a instrumental variable", or genetic instrumental variable.

Then we can decompose the composite effect size Γ as $\Gamma = \alpha\beta$. Since we only have estimates $\hat{\Gamma}$ and $\hat{\alpha}$, we need to test mediation effect size asking non-zeroness of

$$H_0 : \beta \approx \Gamma/\alpha = 0, \quad H_1 : \beta \neq 0$$

As Sobel's test, we estimate the mean of the test by conventional model estimation: $\hat{\Gamma}$ and $\hat{\alpha}$. Or these estimates are provided as input along with standard errors σ_Γ and σ_α.

We can calculate the variance of mediation effect β by the delta method.

$$\mathbb{V}[\beta] \approx \frac{\sigma_\Gamma^2}{\hat{\alpha}^2} + \frac{\hat{\Gamma}^2 \sigma_\alpha^2}{\hat{\alpha}^4}$$

A similar approach was used by a different group, termed SMR method with a binary package, SMR, which you can use.

Lab questions

- Simulate the data with a single true instrumental variable and a single true mediator and 99 null mediators as before.
- Estimate summary statistics $\hat{\Gamma}$ with standard error, and $\hat{\alpha}$ with standard error on 100 genes.
- Perform summary-based Mendelian randomization (either in-house or the publicly available code).
- Repeat the experiments with a genetic variant violating IV conditions (IV1 - IV3). You could make it violate IV assumptions.
- An open question: What will be a remedy even in the cases where none of genetic variants perfectly satisfy IV conditions?

(Optional) Multiple instrumental variable

So far, we only considered a single SNP as a genetic IV, but we can aggregate information across multiple genetic IVs. This is currently actively developing research topic. There is a convenient MR package for multiple SNP analysis. If you were interested, you can simulate data with a multivariate genetic model and estimate mediation effect sizes using this package.