Practical: Mendelian randomization

Qingyuan Zhao (Statistical Laboratory)[*]

July 21, 2023

Mendelian randomisation is an instance of the instrumental variable method where genetic variation is used as instrumental variables. This exercise demonstrates the so called two-sample summary data Mendelian randomization, where two different GWAS are used to estimate $\mathrm{Cov}(Z, Y)/\mathrm{Cov}(Z, Z)$ and $\mathrm{Cov}(Z, A)/\mathrm{Cov}(Z, Z)$.

1. Load the `TwoSampleMR` package that is hosted on GitHub. Execute `ao <- available_outcomes()` to obtain a data frame of the available GWAS summary datasets in the database. The returned variable `ao` is a `tibble`, a modern implementation of R's basic `data.frame` in the tidyverse framework. If you are not familiar with `tibble`, you can convert it to a `data.frame` using the function `as.data.frame`.
2. Find the traits for the GWAS datasets "ieu-a-2" and "ieu-a-7". *Hint*: `?subset`
3. Use the function `extract_instruments` with the "ieu-a-2" dataset to obtain genetic instruments for the exposure trait. This function uses LD-clumping to greedily find (nearly) independent SNPs that are associated with the exposure trait. The argument `p1` in `extract_instruments` controls the significance threshold. Set `p1` to `1e-3`. This should give you about 480 SNPs (genetic instruments).
4. Obtain the associations of these SNPs with the outcome trait using the function `extract_outcome_data` with the "ieu-a-7" dataset.
5. Harmonise the alleles and effects between the exposure and outcome using the function `harmonise_data`. This should return a data frame that contains the associations of the SNPs with the exposure in column `beta.exposure` and with the outcome in column `beta.outcome`. They estimate $\mathrm{Cov}(Z, Y)/\mathrm{Cov}(Z, Z)$ and $\mathrm{Cov}(Z, A)/\mathrm{Cov}(Z, Z)$ for each instrument Z, and the standard errors of these estimates are given in the columns `se.exposure` and `se.outcome`.
6. Obtain an estimate of the causal effect by taking the sample median of the ratio of the columns `beta.exposure` and `beta.outcome`.
7. Perform the default Mendelian randomisation tests using the function `mr`. This gives you the estimated causal effect (in column `b`) and its standard error (in column `se`) using several simple methods. Compare your median estimate above with the results here.
8. Visualise the results above using the function `mr_scatter_plot`.
9. Perform two additional Mendelian randomisation analyses, one using highly significant SNPs with `pval.exposure` less than 1e-8 and one using less significant SNPs with `pval.exposure` larger than 1e-8 (but smaller than 1e-3, why?). You should find that the results are quite different. Can you offer an intuitive explanation? *Hint*: If we generate `x <- rnorm(500)` but only keep those entries in `x` that are larger than 2, what would be the mean of the remaining entries?

[*] qyzhao@statslab.cam.ac.uk