R-Blatt 6: Mendelische Randomisierung - Aufgaben

Statistical Aspects (09-202-2413)

Janne Pott

Last compiled on 06 Oktober, 2022

Session Setup

```
rm(list = ls())
time0<-Sys.time()

source("../sourceFile.R")
setwd(pathToExercise)

knitr::opts_chunk$set(echo = TRUE)
```

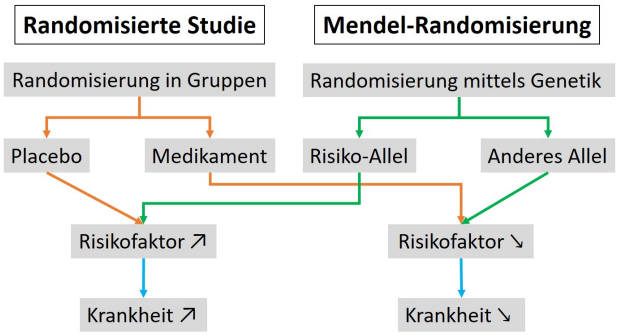

Abbildung 1: Vergleich RCT vs. MR

Das Ziel einer MR ist die Beschreibung von einem kausalen Effekt von einem Risikofaktor X auf ein Outcome bzw. Krankheit Y. Man kann statt Medikament vs Placebo auch Risiko-Allel vs anderes Allel vergleichen ("Randomisierung während Meiose"). Dazu müssen drei Bedingungen gelten:

1) Die Assoziation der genetischen Variante G auf X ist **stark**, z.B. genomweit signifikant.
2) Der SNP G ist unabhängig von **allen** Confoundern U, d.h. G ist nicht auch mit diesen assoziiert.
3) Der SNP G ist unabhängig von Y, bis auf den Effekt der durch X vermittelt wird, d.h. es gibt keinen direkten Effekt von G auf Y.

Während man die erste Bedingung gut nachweisen kann, kann man die anderen beiden nur plausibilisieren (man kennt nicht alle Confounder; man kann nur für die testen, zu denen man Daten hat). Wenn die Bedingungen (plausibel) erfüllt sind, kann man mittels folgenden Model einen Ratio-Schätzer ableiten (β_{IV}, durch die Genetik erklärte Effekt von X auf Y, IV= instrumenal variable):

$$Y \sim \beta_{IV} \cdot X = \beta_{IV} \cdot (\beta_X \cdot G) = \beta_Y \cdot G$$
$$\Longrightarrow \hat{\beta}_{IV} = \frac{\hat{\beta}_Y}{\hat{\beta}_X}$$

Den Standardfehler kann mittels der Delta-Methode bestimmen. Üblicherweise schneidet man nach dem ersten oder zweiten Term (SE_1 bzw. SE_2) ab:

$$SE_1(\hat{\beta}_{IV}) = se(\hat{\beta}_Y)/\hat{\beta}_X$$
$$SE_2(\hat{\beta}_{IV}) = \sqrt{\frac{se(\hat{\beta}_Y)^2}{\hat{\beta}_X^2} + \frac{\hat{\beta}_Y^2 se(\hat{\beta}_X)^2}{\hat{\beta}_X^4}}$$

Ratio-Methode

Bitte laden Sie den Datensatz *Blatt6_MR.RData* mittels *load()*. Er enthält Daten von 1000 Personen und 4 SNPs *g1 - g4*, einem Risikofaktor x und zwei Outcomes y (kontinuierlich) und *y.bin* (binär). Die beiden kontiunierlichen Größen sind annähernd normalverteilt. Die SNPs sind klassisch codiert, d.h. Genotyp AA entspricht 0, AB 1 und BB 2.

a) Bestimmen sie folgende Parameter für alle SNPs *g1 - g4*:
   - Die Schätzer aus den jeweiligen linearen Regressionen: $\hat{\beta}_Y$, $se(\hat{\beta}_Y)$, $\hat{\beta}_X$, $se(\hat{\beta}_X)$
   - Der kausale Schätzer β_{IV} und beide Standardfehler SE_1 und SE_2 sowie die dazugehörigen P-Werte
   - Die F-Statistik der Regression des Risikofaktors.
   - Die MAF

b) Bezogen auf den Standardfehler erster Ordnung, welche genetische Variante liefert das präziseste Ergebnis? Wodurch wird die Präzision beeinflusst? Wann und wo unterscheiden sich die Fehler erster und zweiter Ordnung am meisten?

c) Unterscheidet sich der kausale Schätzer von der beobachteten Assoziation? Welche kausalen Schätzer sind signifikant?

Two-Stage least squares Methode (2SLS oder TSLS)

Bei dieser Methode wird der kausale Schätzer mittels zweifacher Regression bestimmt:

1. Stufe: Regression des Risikofaktors auf den SNP
2. Stufe: Regression des Outcome auf die gefitteten Werte des Risikofaktors aus der 1. Stufe

$$Y \sim \beta_{IV} \cdot X = \beta_{IV} \cdot (\beta_X \cdot G) = \beta_Y \cdot G$$
$$X \sim \beta_0 + \beta_X \cdot G + \epsilon \rightarrow X_{fit} = \beta_x \cdot G$$
$$Y \sim \beta_0 + \beta_{IV} \cdot X_{fit} + \epsilon$$

Ein Vorteil dieser Methode ist, dass man auch gleichzeitig mehrere SNPs verwenden kann, indem man in der ersten Stufe ein multivariates Modell verwendet. Allerdings sollten dazu diese SNPs unkorreliert sein (z.B. LD $r^2 < 0.1$).

a) Führen Sie ein **TSLS** stufenweise für alle SNPs einzeln und gemeinsam durch und notieren Sie sich den Schätzer und dessen Standardfehler!

b) Nutzen Sie nun die *ivreg* Funktion des R-Pakets *ivpack* und führen Sie ebenfalls eine **TSLS** für alle SNPs einzeln und gemeinsam durch!
c) Wie unterscheiden sich die Ergebnisse:
   - von der Ratio- und der TSLS-Methode pro SNP?
   - von der per Hand und der *ivreg* Variante für die gemeinsame Analyse?

Inverse Varianz gewichtete Methode (inverse-variance weighted, IVW)

Auch mit der Ratio-Methode kann man mehrere SNPs kombinieren. Dazu wird eine Meta-Analyse der kausalen Schätzer durchgeführt. Dies funktioniert auch, wenn die Statistiken von unterschiedlichen GWASs stammen und entspricht einem fixed Effekt Modell. Auch hier sollte man vorher sicherstellen, dass die SNPs nicht korreliert sind.

$$\hat{\beta}_{IV,IVW} = \frac{\sum \hat{\beta}_Y \hat{\beta}_X se(\hat{\beta}_Y)^{-2}}{\sum \hat{\beta}_X^2 se(\hat{\beta}_Y)^{-2}}$$

$$SE_3(\hat{\beta}_{IV,IVW}) = \sqrt{\frac{1}{\sum \hat{\beta}_X^2 se(\hat{\beta}_Y)^{-2}}}$$

In dem Paket **MendelianRandomization** von Stephen Burgess sind inzwischen viele Varianten der MR mittels Summary Statistics implementiert. Die IVW-Methode ist nur eine davon. Andere berücksichtigen etwaige Pleiotropie (z.B. *MR_egger*).

Bestimmen Sie den kausalen Meta-Effekt mittels jeweils mit und ohne SNP *g4*:

a) der oben aufgeführten Funktionen
b) der Funktion *metagen* aus dem Paket **meta** (s. letzte R-Übung)
c) der Funktion *mr_allmethods* aus dem Paket **MendelianRandomization**
d) Gibt es Unterschiede bei den Kausalschätzungen?
e) Erstellen Sie einen Scatterplot der vier Instrumente inkl. der Fehlerbalken (Hinweis: *mr_plot* aus **MendelianRandomization**)!

MR mit binären Outcome

Natürlich kann man das Prinzip der MR auch auf binäre Phänotypen anwenden. Hierbei ist zu beachten, dass empfohlen wird, die G-X Assoziation nur auf den Kontrollen bzgl. Y zu rechnen. Die berechneten Schätzer repräsentieren die log-kausalen Odds Ratios für *y.bin* pro Einheitssteigerung von *x*. Zurückrechnen zu normalen OR erfolgt über die Exponential-Funktion, die Konfidenzintervalle kann man mittels Normal-Approximation bestimmen:

Führen Sie die Ratio-Methode bzw. TSLS für *g2* bzw. *g1* - *g3* durch!

Session Information

```
sessionInfo()
message("\nTOTAL TIME : " ,round(difftime(Sys.time(),time0,units = "mins"),3)," minutes")
```