Modul Statistische Aspekte der Analyse molekularbiologischer und genetischer Daten

R-Blatt 6: Mendelsche Randomisierung

Janne Pott

WS 2021/22

Ziel der Übung ist eine Mendelsche Randomisierung (MR) durchzuführen um kausale Beziehungen zu identifizieren.

Ratio-Methode

Beispiele

Das Ziel einer MR ist die Beschreibung von einem kausalen Effekt von einem Risikofaktor X auf ein Outcome bzw. Krankheit Y.

Klassischerweise würde man dazu eine randomisierte kontrollierte Studie (RCT) durchführen, in der ein Studienarm eine Intervention (Medikament, Behandlung o.ä.) erfährt, die einen Effekt auf den Risikofaktor hat. Der andere Arm wird nur mittels Placebo behandelt oder ohne Therapie weiter beobachtet. Wenn die beiden Arme balanciert waren bezüglich aller relevanten Confounder U ist jede Änderung in Y verursacht durch die Änderung in X und ein kausaler Effekt kann geschätzt werden.

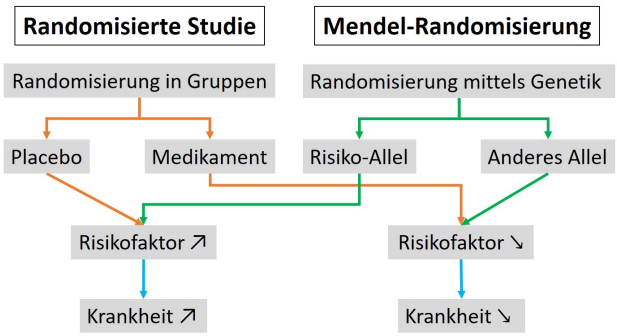


Figure 1: Vergleich RCT vs. MR

Um Zeit und Kosten zu sparen, kann man statt Medikament vs Placebo auch Risiko-Allel vs anderes Allel vergleichen. 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).

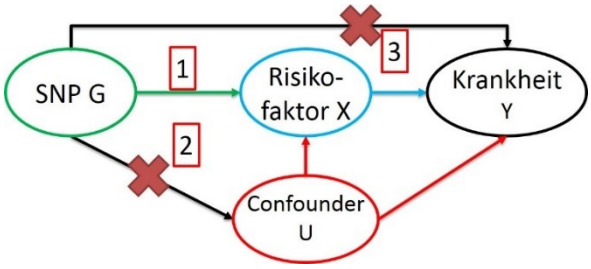


Figure 2: Gerichteter azyklischer Graph der Mendelschen Randomisierung

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):

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

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

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

In diesem Beispiel nutzen wir 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.

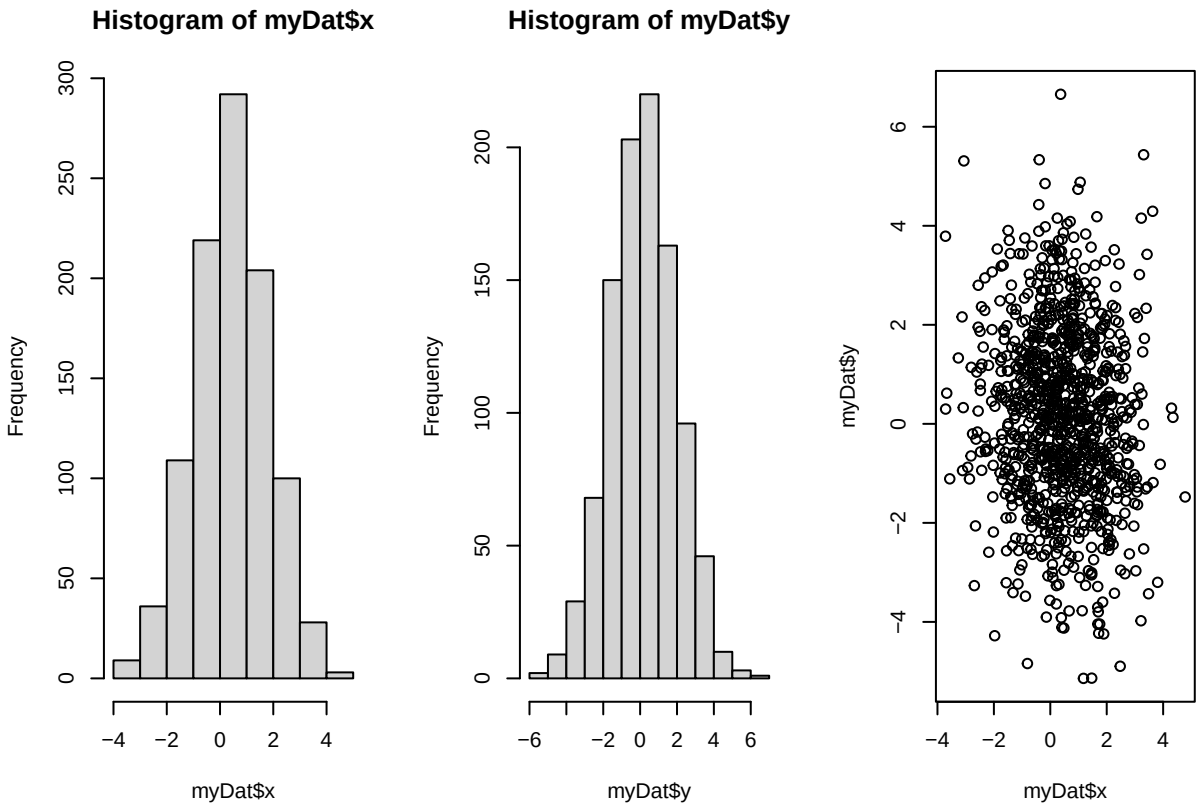
```
load("data2/MR.RData")
dim(myDat)
```

```
## [1] 1000    8
```

```
colnames(myDat)
```

```
## [1] "ID"    "g1"    "g2"    "g3"    "g4"    "x"     "y"     "y.bin"
```

```
par(mfrow=c(1,3))
hist(myDat$x)
hist(myDat$y)
plot(myDat$x,myDat$y)
```




```
cor.test(myDat$y,myDat$x)
```

```
##
##  Pearson's product-moment correlation
##
## data:  myDat$y and myDat$x
## t = -3.0816, df = 998, p-value = 0.002115
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.15812807 -0.03530613
## sample estimates:
##         cor
## -0.09708672
```

```
summary(lm(y~x,data=myDat))
```

```
##
## Call:
## lm(formula = y ~ x, data = myDat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2100 -1.1850  0.0143  1.1599  6.4855
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.21671    0.05900   3.673 0.000252 ***
## x           -0.12547    0.04071  -3.082 0.002115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.788 on 998 degrees of freedom
## Multiple R-squared:  0.009426,   Adjusted R-squared:  0.008433
## F-statistic: 9.496 on 1 and 998 DF,  p-value: 0.002115
```

```
attach(myDat)
mody<-lm(y ~ g1)
modx<-lm(x ~ g1)
by<-summary(mody)$coef[2,1]
byse<-summary(mody)$coef[2,2]
bx<-summary(modx)$coef[2,1]
bxse<-summary(modx)$coef[2,2]
beta.ratio<-by/bx
se.ratio.st<- byse/sqrt(bx^2)
p1<-2*pnorm(-abs(beta.ratio/se.ratio.st))
beta.ratio; se.ratio.st; p1
```

```
## [1] -0.06302634
```

```
## [1] 0.6451987
```

```
## [1] 0.9221823
```

Aufgaben

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)

Beispiele

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

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

Der Vorteil dieser Methode ist, dass man auch gleichzeitig mehrere SNPs verwenden kann:

```
stage1<-lm(x~g1)
fit<-stage1$fitted.values
stage2<-lm(y~fit)
summary(stage2)
```

```
##
## Call:
## lm(formula = y ~ fit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3108 -1.2105 -0.0018  1.2244  6.4947
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.19091    0.27257   0.700    0.484
## fit         -0.06303    0.64520  -0.098    0.922
##
## Residual standard error: 1.797 on 998 degrees of freedom
## Multiple R-squared:  9.561e-06,  Adjusted R-squared:  -0.0009924
## F-statistic: 0.009542 on 1 and 998 DF,  p-value: 0.9222
```

```
mod1<-ivreg(y~x|g1,x=T)
summary(mod1)
```

```
##
## Call:
## ivreg(formula = y ~ x | g1, x = T)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -5.257578 -1.213809  0.008249  1.206072  6.488404
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.19091    0.27160   0.703    0.482
## x           -0.06303    0.64291  -0.098    0.922
##
## Residual standard error: 1.79 on 998 degrees of freedom
## Multiple R-Squared: 0.007091,    Adjusted R-squared: 0.006096
## Wald test: 0.00961 on 1 and 998 DF,  p-value: 0.9219
```

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 $r^2 < 0.1$).

Aufgaben

a) Führen Sie ein TSLS per Hand 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?

Per Hand bezieht sich hier darauf, dass man beide Stufen per Hand rechnet, und nicht alles in einem Schritt wie mit der *ivreg* Funktion von **ivpack**.

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

Beispiele

Oft hat man nur Summarized Data, d.h. nur die βs und SEs der einzelnen SNPs. Damit lässt sich keine TSLS durchführen. Stattdessen kann man den kausalen Schätzer mittels IVW bestimmen:

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

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

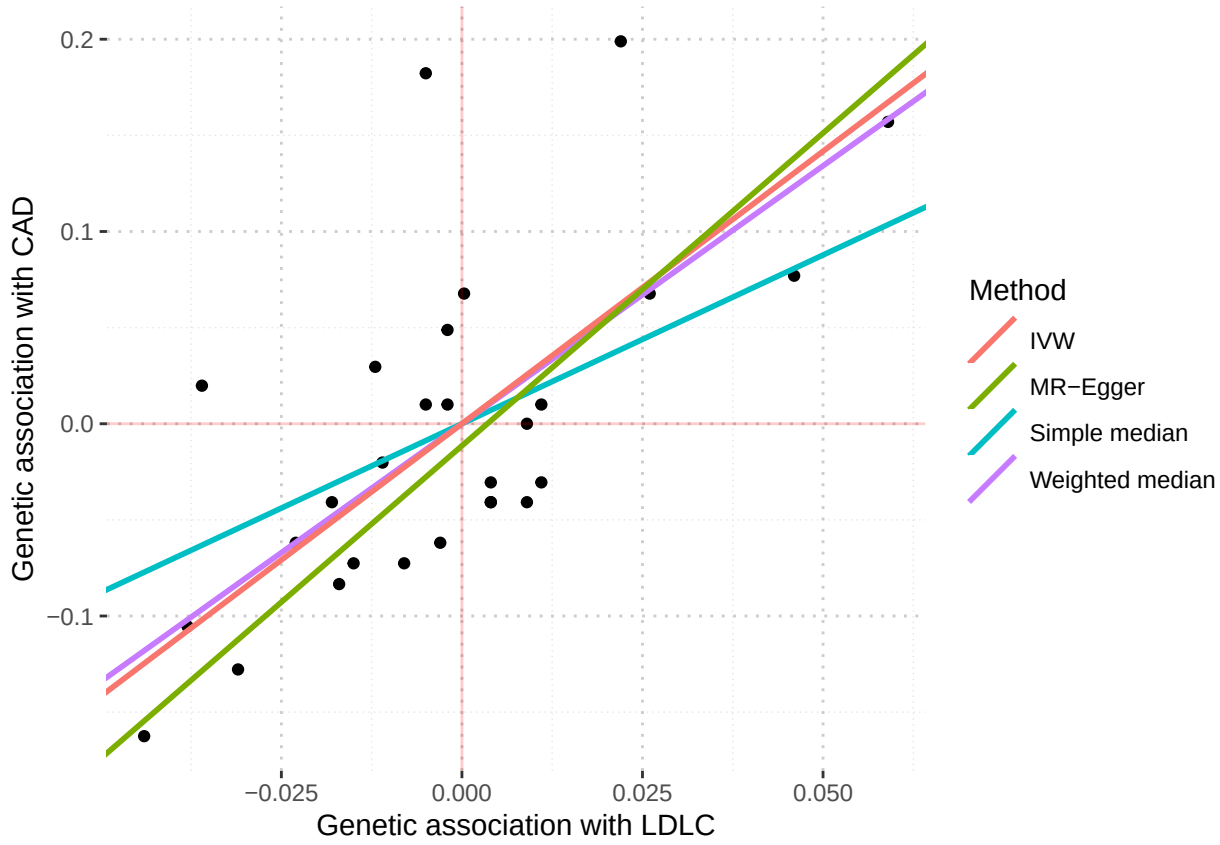
Dies entspricht einer Meta-Analyse der kausalen Schätzer (FEM). Auch hier sollte man vorher sicherstellen, dass die SNPs nicht korreliert sind.

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). Hier im Beispiel wird der Effekt von LDL-Cholesterol auf koronare Herzkrankheit betrachtet:

```
MRObject<- mr_input(bx = ldlc, bxse = ldlcse, exposure = "LDLC",
                    by = chdlodds, byse = chdloddsse, outcome="CAD")
mr_allmethods(MRObject)
```

```
##                      Method Estimate Std Error 95% CI        P-value
##              Simple median    1.755     0.740   0.305 3.205   0.018
##            Weighted median    2.683     0.419   1.862 3.504   0.000
##  Penalized weighted median    2.681     0.420   1.857 3.505   0.000
##
##                         IVW    2.834     0.530   1.796 3.873   0.000
##              Penalized IVW    2.561     0.413   1.752 3.370   0.000
##                 Robust IVW    2.797     0.307   2.195 3.399   0.000
##       Penalized robust IVW    2.576     0.251   2.083 3.069   0.000
##
##                   MR-Egger    3.253     0.770   1.743 4.762   0.000
##                 (intercept)  -0.011     0.015  -0.041 0.018   0.451
##         Penalized MR-Egger    3.421     0.531   2.380 4.461   0.000
##                 (intercept)  -0.022     0.011  -0.043 0.000   0.051
##             Robust MR-Egger    3.256     0.624   2.033 4.479   0.000
##                 (intercept)  -0.015     0.021  -0.055 0.026   0.474
##  Penalized robust MR-Egger    3.502     0.478   2.566 4.438   0.000
##                 (intercept)  -0.026     0.014  -0.054 0.003   0.075
```

```
mr_plot(mr_allmethods(MRObject,method = "main"))
```



Aufgaben

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

Beispiele

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.

```
attach(myDat)
```

```
## Die folgenden Objekte sind maskiert von myDat (pos = 3):
##
##     g1, g2, g3, g4, ID, x, y, y.bin
```

```
g1.con<-g1[y.bin==0]
x.con<-x[y.bin==0]
predict.con.g1<-predict(lm(x.con~g1.con),newdata=list(g1.con=g1))
tsls2.con<-glm(y.bin ~ predict.con.g1, family="binomial")
summary(tsls2.con)
```

```
##
## Call:
## glm(formula = y.bin ~ predict.con.g1, family = "binomial")
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.359  -1.320   1.024   1.041   1.041
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.4920     0.3173   1.551    0.121
## predict.con.g1   0.2600     0.5922   0.439    0.661
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1355.2  on 999  degrees of freedom
## Residual deviance: 1355.0  on 998  degrees of freedom
## AIC: 1359
##
## Number of Fisher Scoring iterations: 4
```

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:

```
beta.tsls.con<-summary(tsls2.con)$coef[2,1]
se.tsls.con<-summary(tsls2.con)$coef[2,2]

lower.bound<-beta.tsls.con - 1.96*se.tsls.con
upper.bound<-beta.tsls.con + 1.96*se.tsls.con

OR<-exp(beta.tsls.con)
exp(lower.bound); OR; exp(upper.bound)
```

```
## [1] 0.4063013
```

```
## [1] 1.296883
```

```
## [1] 4.139553
```

Aufgaben

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