Twin and Family Strutural Equation Models in R & Lavaan

Michel G Nivard

6/21/2021

Contents

1 Introduction

This document describes various twin structural equation models (SEM) to estimate the nature of the relationship between family members in order to learn about the contribution of genes and the social or rearing environment to complex (behavioral) outcomes. The goal is to provide a basic understands of these

models with the means to fit the models in lavaan (Rosseel 2012), lavaan is an R package that allow the use to define a structural equation model in terms of regression, variances and covariances and will be familiar to users of M-Plus. The package is (IMO) more accessible to beginners then another excellent SEM R package: OpenMx (Neale et al. 2015), but the accessibility comes at the cost of less flexibility, In terms of flexibility OpenMX is truely unrivaled. Its worth pointing out that the developers behind OpenMx have an academic interest in twin models, which means scripts and support for users in those models is often especially excellent.

2 Lavaan

Lavaan requires the user to specify

3 Twin Models & Family models

There is a long history of questioning the nature of the resemblance between family members, specifically sibling and even more specifically twins. Reasons that are commonly put forward for these similarities are the obvious genetic similarity, similarity in socio-economic position and similarity in upbringing shared between siblings. Twin and family models leverage variation in genetic and environmental relatedness between family members to estimate the relative contributions of genetics, the environment, their interacting, their correlation and other process to the similarities between relatives. Twins offer an excellent "natural experiment" where some twins are genetically identical (identical, or monozygotic twins or MZ twins or MZs) and some twins share halve their segregating DNA (fraternal, or dizygotic twins, DZ twins or DZs). Like any natural experiment it comes with various assumptions some of which are specific to twin models, and I'll make sure to catalog the assumptions we are making along the way.

4 Quickly set up my R environment

```
library(lavaan)
```

```
## This is lavaan 0.6-7
```

```
## lavaan is BETA software! Please report any bugs.
```

```
library(MASS)
library(tidySEM)
```

```
## Registered S3 methods overwritten by 'tidySEM':
##   method              from
##   print.mplus.model   MplusAutomation
##   print.mplusObject   MplusAutomation
##   summary.mplus.model MplusAutomation
```

```
library(ggplot2)
```

5 Univariate Twin model

5.0.1 A Common environment

Lets simulate same data where the resemblance between twins is a function of equal parts (33.3%/33.3%/33.4%) additive genetic variance (A), common environmental influences(C) (can be rearing, can be societal influences can be governmental policies), and environment unique to each of the individual twins (E) (can beprivate friends, being in separate classrooms, but also measurement error).

```
A <- matrix(1,2,2) # genetic correlation for MZ's = 1
C <- matrix(1,2,2)
E <- diag(2)
Adz <- matrix(c(1,.5,.5,1),2,2) # genetic correlation for DZ's = 0.5

# make 1000 pairs of MZ twins
MZ <- mvrnorm(1000,mu=c(0,0),Sigma = A+C+E)

# Add a column to label as MZ:
MZ<- cbind.data.frame("MZ",MZ)
colnames(MZ) <- c("zyg","P1", "P2")


# make 1500 DZ twin pairs
DZ <- mvrnorm(1500,mu=c(0,0),Sigma = Adz+C+E)

# add variable too label as DZ:
DZ <- cbind.data.frame("DZ",DZ)
colnames(DZ) <- c("zyg","P1", "P2")

# Combine MZ and DZ twins
dataset <- rbind(MZ,DZ)
```

We then define the lavaan model that can express the variance in the trait P explained by latent variables A, C and E:

```
ace.model<-"
A1=~ NA*P1 + c(a,a)*P1
A2=~ NA*P2 + c(a,a)*P2
C1 =~ NA*P1 + c(c,c)*P1
C2 =~ NA*P2 + c(c,c)*P2
# variances
A1 ~~ 1*A1
A2 ~~ 1*A2
C1 ~~ 1*C1
C2 ~~ 1*C2
P1~~c(e2,e2)*P1
P2~~c(e2,e2)*P2
# covariances
A1 ~~ c(1,.5)*A2
A1 ~~ 0*C1 + 0*C2
A2 ~~ 0*C1 + 0*C2
C1 ~~ c(1,1)*C2"
```

Lets look at some of the critical lines of code in the model:

`A1=~ NA*P1 + c(a,a)*P1` Here we create the latent variable A1, the phenotype P for twin 1 (P1) loads ion this variable, and in both groups (groups being MZ and DZ twins) the influence of this latent variable on the outcome is the same (contained using `c(a,a)`). Similar code is used to defien the latent variables C1 and C2. Now the effect of genes on an outcome is assumed the same for everyone regardless of whether they are twins, or not, the resemblance between twin 1 and twin 2 difference for MZ and DZ twins. We define/fix the resemblance later in the model here: `A1 ~~ c(1,.5)*A2`, because A1 and A2 are variance 1: `A1 ~~ 1*A1` and `A2 ~~ 1*A2` the constrained implies a correlation of 1 for the MZ twins and a correlation of 0.5 for the DZ twins. The common environment is correlated 1 regardless of twin status: `C1 ~~ c(1,1)*C2`, while the unshared environment E is conceptualized as a residual variance of the trait P (P1 or P2 respectively): `P1~~c(e2,e2)*P1`

We assume the latent variables A(1/2) and C(1/2) are uncorrelated, and fix their covariance to 0:

```
A1 ~~ 0*C1 + 0*C2
```

```
## A1 ~ ~0 * C1 + 0 * C2
```

```
A2 ~~ 0*C1 + 0*C2
```

```
## A2 ~ ~0 * C1 + 0 * C2
```

We also assume the residual variance (E) is uncorrelated to A and C, but fortunately for us this is a lavaan default. We proceed to fit the model to the simulated data:

```
# Standard ace model:
ace.fit<-cfa(ace.model, data = dataset,group = "zyg")
summary(ace.fit)
```

```
## lavaan 0.6-7 ended normally after 21 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of free parameters                         16
##   Number of equality constraints                     9
##
##   Number of observations per group:
##     MZ                                            1000
##     DZ                                            1500
##
## Model Test User Model:
##
##   Test statistic                                 3.453
##   Degrees of freedom                                 3
##   P-value (Chi-square)                           0.327
##   Test statistic for each group:
##     MZ                                           1.635
##     DZ                                           1.819
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
```

```
##
##
## Group 1 [MZ]:
##
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   A1 =~
##     P1         (a)    0.930    0.072   12.945    0.000
##   A2 =~
##     P2         (a)    0.930    0.072   12.945    0.000
##   C1 =~
##     P1         (c)    1.080    0.056   19.374    0.000
##   C2 =~
##     P2         (c)    1.080    0.056   19.374    0.000
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   A1 ~~
##     A2                1.000
##     C1                0.000
##     C2                0.000
##   A2 ~~
##     C1                0.000
##     C2                0.000
##   C1 ~~
##     C2                1.000
##
## Intercepts:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .P1                0.050    0.055    0.916    0.360
##    .P2                0.027    0.055    0.496    0.620
##     A1                0.000
##     A2                0.000
##     C1                0.000
##     C2                0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##     A1                1.000
##     A2                1.000
##     C1                1.000
##     C2                1.000
##    .P1        (e2)    1.000    0.044   22.630    0.000
##    .P2        (e2)    1.000    0.044   22.630    0.000
##
##
## Group 2 [DZ]:
##
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   A1 =~
##     P1         (a)    0.930    0.072   12.945    0.000
##   A2 =~
##     P2         (a)    0.930    0.072   12.945    0.000
```

```
##   C1 =~
##     P1         (c)    1.080    0.056   19.374    0.000
##   C2 =~
##     P2         (c)    1.080    0.056   19.374    0.000
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   A1 ~~
##     A2                0.500
##     C1                0.000
##     C2                0.000
##   A2 ~~
##     C1                0.000
##     C2                0.000
##   C1 ~~
##     C2                1.000
##
## Intercepts:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .P1                0.061    0.045    1.355    0.175
##    .P2               -0.006    0.045   -0.133    0.894
##     A1                0.000
##     A2                0.000
##     C1                0.000
##     C2                0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##     A1                1.000
##     A2                1.000
##     C1                1.000
##     C2                1.000
##    .P1         (e2)    1.000    0.044   22.630    0.000
##    .P2         (e2)    1.000    0.044   22.630    0.000
```

5.0.2 Non-additive genetic variance

Lets simulate same data where the resemblance between twins is a function of equal parts (33.3%/33.3%/33.4%) additive genetic variance (A), non additive genetic effects (D) (can be geneXgene interactino, can be dominant inheritance where a single allele is enough to express the trait regardless of the state of the other allele), and environment unique to each of the individual twins (E) (can private friends, being in separate classrooms, but also measurement error).

```
A <- matrix(1,2,2) # genetic correlation for MZ's = 1
D <- matrix(1,2,2)
E <- diag(2)
Adz <- matrix(c(1,.5,.5,1),2,2) # additive genetic correlation for DZ's = 0.5
Ddz <- matrix(c(1,.25,.25,1),2,2)  # non-additive genetic correlation for DZ's = 0.5

# make 1000 pairs of MZ twins
MZ <- mvrnorm(1000,mu=c(0,0),Sigma = A+D+E)

# Add a column to label as MZ:
```

```
MZ<- cbind.data.frame("MZ",MZ)
colnames(MZ) <- c("zyg","P1", "P2")


# make 1500 DZ twin pairs
DZ <- mvrnorm(1500,mu=c(0,0),Sigma = Adz+Ddz+E)

# add variable too label as DZ:
DZ <- cbind.data.frame("DZ",DZ)
colnames(DZ) <- c("zyg","P1", "P2")

# Combine MZ and DZ twins
dataset <- rbind(MZ,DZ)
```

We then define the lavaan model that can express the variance in the trait P explained by latent variables A, C and E:

```
ade.model<-"
A1=~ NA*P1 + c(a,a)*P1
A2=~ NA*P2 + c(a,a)*P2
D1 =~ NA*P1 + c(d,d)*P1
D2 =~ NA*P2 + c(d,d)*P2
# variances
A1 ~~ 1*A1
A2 ~~ 1*A2
D1 ~~ 1*D1
D2 ~~ 1*D2
P1~~c(e2,e2)*P1
P2~~c(e2,e2)*P2
# covariances
A1 ~~ c(1,.5)*A2
A1 ~~ 0*D1 + 0*D2
A2 ~~ 0*D1 + 0*D2
D1 ~~ c(1,.25)*D2"
```

Lets look at some of the critical lines of code in the model:

`A1=~ NA*P1 + c(a,a)*P1` Here we create the latent variable A1, the phenotype P for twin 1 (P1) loads on this variable, and in both groups (groups being MZ and DZ twins) the influence of this latent variable on the outcome is the same (contained using `c(a,a)`). Similar code is used to define the latent variables D1 and D2. Now the effect of genes on an outcome is assumed the same for everyone regardless of whether they are twins, or not, the resemblance between twin 1 and twin 2 difference for MZ and DZ twins. We define/fix the resemblance later in the model here: `A1 ~~ c(1,.5)*A2`, because A1 and A2 are variance 1: `A1 ~~ 1*A1` and `A2 ~~ 1*A2` the constrained implies a correlation of 1 for the MZ twins and a correlation of 0.5 for the DZ twins. The non-additive genetic effects are correlated 1 for MZ twins and .25 for DZ twins `C1 ~~ c(1,.25)*C2`, while the unshared environment E is conceptualized as a residual variance of the trait P (P1 or P2 respectively): `P1~~c(e2,e2)*P1`

We assume the latent variables A(1/2) and D(1/2) are uncorrelated, and fix their covariance to 0:

```
A1 ~~ 0*D1 + 0*D2
```

```
## A1 ~ ~0 * D1 + 0 * D2
```

```
A2 ~~ 0*D1 + 0*D2
```

```
## A2 ~ ~0 * D1 + 0 * D2
```

We also assume the residual variance (E) is uncorrelated to A and D, but fortunately for us this is a lavaan default.We proceed to fit the model to the simulated data:

```
# Standard ace model:
ade.fit<-cfa(ade.model, data = dataset,group = "zyg")
summary(ade.fit)
```

```
## lavaan 0.6-7 ended normally after 22 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of free parameters                         16
##   Number of equality constraints                     9
##
##   Number of observations per group:
##     MZ                                            1000
##     DZ                                            1500
##
## Model Test User Model:
##
##   Test statistic                                 2.133
##   Degrees of freedom                                 3
##   P-value (Chi-square)                           0.545
##   Test statistic for each group:
##     MZ                                           1.264
##     DZ                                           0.869
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
##
## Group 1 [MZ]:
##
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   A1 =~
##     P1         (a)    0.759    0.190    3.987    0.000
##   A2 =~
##     P2         (a)    0.759    0.190    3.987    0.000
##   D1 =~
##     P1         (d)    1.163    0.127    9.157    0.000
##   D2 =~
##     P2         (d)    1.163    0.127    9.157    0.000
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)
```

```
##   A1 ~~
##     A2                1.000
##     D1                0.000
##     D2                0.000
##   A2 ~~
##     D1                0.000
##     D2                0.000
##   D1 ~~
##     D2                1.000
##
## Intercepts:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .P1               -0.054    0.054   -0.997    0.319
##    .P2               -0.038    0.054   -0.709    0.478
##     A1                0.000
##     A2                0.000
##     D1                0.000
##     D2                0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##     A1                1.000
##     A2                1.000
##     D1                1.000
##     D2                1.000
##    .P1        (e2)    0.997    0.044   22.672    0.000
##    .P2        (e2)    0.997    0.044   22.672    0.000
##
##
## Group 2 [DZ]:
##
## Latent Variables:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   A1 =~
##     P1         (a)    0.759    0.190    3.987    0.000
##   A2 =~
##     P2         (a)    0.759    0.190    3.987    0.000
##   D1 =~
##     P1         (d)    1.163    0.127    9.157    0.000
##   D2 =~
##     P2         (d)    1.163    0.127    9.157    0.000
##
## Covariances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   A1 ~~
##     A2                0.500
##     D1                0.000
##     D2                0.000
##   A2 ~~
##     D1                0.000
##     D2                0.000
##   D1 ~~
##     D2                0.250
##
```

```
## Intercepts:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .P1                0.026    0.044    0.588    0.556
##    .P2               -0.036    0.044   -0.820    0.412
##     A1                0.000
##     A2                0.000
##     D1                0.000
##     D2                0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##     A1                1.000
##     A2                1.000
##     D1                1.000
##     D2                1.000
##    .P1        (e2)    0.997    0.044   22.672    0.000
##    .P2        (e2)    0.997    0.044   22.672    0.000
```

5.0.3 Sex specific effects

5.0.4 Binary/Ordinal data

5.1 Gene environment interaction

5.2 Gene-enviornment correlation (model requires a PRS)

5.3 Sibling interactions & Rater (contrast) effects

6 Multivariate models

6.1 A bivariate twin model

6.2 Direction of Causation models

6.2.1 Variatiomns on Direction of Causation models

6.3 Rater bias models

References

Neale, Michael C., Michael D. Hunter, Joshua N. Pritikin, Mahsa Zahery, Timothy R. Brick, Robert M. Kirkpatrick, Ryne Estabrook, Timothy C. Bates, Hermine H. Maes, and Steven M. Boker. 2015. "OpenMx 2.0: Extended Structural Equation and Statistical Modeling." *Psychometrika* 81 (2): 535–49. https://doi.org/10.1007/s11336-014-9435-8.

Rosseel, Yves. 2012. "Lavaan: AnRPackage for Structural Equation Modeling." *Journal of Statistical Software* 48 (2). https://doi.org/10.18637/jss.v048.i02.