## Intro

*Q1.1.* As you get started, introduce yourself and let us know your zoom breakout room number?

*Q1.2.* What are the first names of your group members?

*Q1.3.* Has anyone in your group run a sex limitation or GxE analysis before?

□ Group members have run a sex limitation analysis before

□ Group members have run a GxE analysis before

□ Group members have not run either of these types of analyses before

*Q1.4.* Welcome to the heterogeneity practical! Please use the following commands to copy the example files into your own directories.

```
# Create a directory to hold your day's work
mkdir day3
# change into that directory, and then copy over the exercises.
cd day3
cp -R /faculty/hmaes/2022/day3/* ./
```

Make sure you have a space and a period after the star to copy the contents of my `day3` directory into yours. You will be running a set of analyses in RStudio. Remember to set your working directory using `setwd('day3').` Use examples of the code in `practicalDay3.R` as instructed by the questions below. It might be helpful to organize your code file by adding your own comments, so you can easily run/or re-run sections throughout the practical. It might also be useful to copy summary tables of goodness-of-fit statistics and estimates into separate files as you go along.

### Sex limitation Intro

*Q2.1.*

## Sex limitation Practical

By now, you are familiar with fitting the basic twin model, to estimate the contributions of additive genetic **VA**, shared environmental **VC** (or dominance **VD**) and unique environmental **VE** factors to your phenotype of interest. We have also shown how confounders/covariates can impact the results of these twin analyses and that we can correct for their main effect by including them in the analyses, estimating their effect on the means and partition the remaining variance into the **ACE**/**ADE** variance components.

The data used in previous examples were simulated to have specific properties. In today's practical, we're analyzing data on body mass index, BMI, a measure of obesity, calculated as weight (in kg) divided by height squared (in meters). For most of the behavioral traits we're interested in, we know that their distributions might vary by sex and age, and potentially other confounders. In addition to mean differences in the trait by sex (or any other covariate), there might also be differences in the total variance or any of the variance components.

Today, we're focusing on testing for heterogeneity in sources of variance and the magnitude of their contributions by sex, age or other binary or continuous covariates. Let's consider BMI. We know from epidemiological studies that there are differences in mean and variance of BMI by sex. We have access to reasonably large samples of BMI data collected in MZ and DZ twin pairs, a dataset that is freely available when installing OpenMx, called `twinData`, which one row of data per twin pair, twin 1's measure of BMI denoted with `bmi1` and twin 2's with `bmi2`.

### Twin Correlations

*Q3.1.*

Open the the `practicalDay3.R` script and run the code in lines 17-44 to inspect the data and estimate the twin correlations in R.

Note that it’s important to know which zygosity code represents which zygosity and sex. To get some extra practice with OpenMx, let’s run a basic twin model, separately for young males and females. We’ll ignore the opposite sex twins for the time being. Before we do that, we'll inspect the correlations to decide which genetic model to run.

Based on the twin correlations, which model is likely going to fit the data best? Is this the same model for males and females?

### 2-group Saturated Models

*Q4.1.*

You can edit code used in previous practicals and change the data, or you can run lines 45-110 of the `practicalDay3.R` script for the 2-group **saturated** model (without any comments), which tests model assumptions, and then run either the **ACE** (lines 111-154) or **ADE** (lines 154-193) models. Note that we've included age as a covariate, and regressed out its effect on the means. Later in the practical, we'll test whether there are differences in variance (components) by age.

Use the proper zygosity codes and run the code -separately for young females and young males. Make sure you write down the -2 log-likelihood and degrees of freedom for each of the analyses as we'll need them later.

Are assumptions about equality of means and variances across twin order and zygosity met for young males? for young females? Put a check mark in the boxes if the assumption is met.

| | males | females |
|---|---|---|
| equal means by twin order | □ | □ |
| equal variances by twin order | □ | □ |
| equal means & variances by zygosity | □ | □ |

## 2-group ACE/ADE Models

*Q5.1.*

What are the estimates of the additive genetic, shared environmental (or dominance) and unique environmental variances the same for males and females? Please list the estimates below.

| | females | males |
|---|---|---|
| additive genetic variance | | |
| shared environmental variance | | |
| unique environmental variance | | |
| dominance variance | | |

*Q5.2.* Do you think they are different from one another? How would you test this?

## 4-group Saturated model

*Q6.1.* We need to be able to constrain the parameters for males to those for females, to test whether they are significantly different from one another, thus we need to extend our multigroup **2-group** analysis to a **4-group** analysis (same-sex female MZ & DZ and same-sex male MZ & DZ groups), thus doubling up data statements, means matrices statements, covariance matrices statements, expectation statements, model statements, etc. and combine them all in one model, as in lines 200-317 of the `practicalDay3.R` script.

We can use a number of functions that summarize the key goodness-of-fit statistics (`fitGofs()`) and the estimated parameters (`fitEsts()`) in just a few lines. These functions are sourced from a file called `miFunctions.R` which can be dowloaded here.

Report the likelihoods of the 2-group **saturated** models you ran separately for males and females and the 4-group **saturated** model ?

| | females | males | both sexes |
|---|---|---|---|
| -2 log-likelihood | | | |

*Q32.* How does the likelihood of the 4-group saturated model compare to the likelihoods of the 2-group saturated models you ran separately for males and females?

## 4-group Saturated Models 2

*Q7.1.* Next, we'll test the significance of age on BMI, and repeat the assumption testing in lines 318-358. By now, you are familiar with testing whether means and/or variances can be equated across twin order and zygosity.

Please edit lines 360-364 of the code which is reproduced below to test whether means and variances can also be equated by sex.

```
# Constrain expected Means and Variances to be equal across twin order and zygosity and sex

modelEMVS <- mxModel( fitEMVZ, name="oneEMVSca" )

modelEMVS <- omxSetParameters( modelEMVS, label=c(______,______ free=TRUE, values=svMe, newlabels='mZ' )

modelEMVS <- omxSetParameters( modelEMVS, label=c("vZf","vZm"), free=______, values=______, newlabels='vZ'
)

fitEMVS   <- mxRun( modelEMVS, intervals=F )
```

Paste a copy of your completed lines of code into the box below.

**4-group Genetic Models**

*Q8.1.* As we established that the means cannot be equated by sex without loss of fit and that the twin correlations are consistent with the **ADE** model in lines 374-469 of the `practicalDay3.R` script, we will estimate one mean for males and one for females. To test whether the variance components estimates vary by sex, we will first fit a model with separate estimates for males and females, which implies that we have to double up statements again - see lines 402-407 . Note that we give different labels for the parameters for males and females - **different labels = different parameters**.

Please complete the expressions for the expectations of the variances and covariances for the four zygosity by sex groups in lines 410-415 of the `practicalDay3.R` script and reproduced below, and paste a copy of your code into the box below:

```
# Create Algebra for expected Variance/Covariance Matrices in MZ & DZ twins

covPf     <- mxAlgebra( expression= ___________, name="Vf" )

covPm     <- mxAlgebra( expression= ___________, name="Vm" )

covMZf    <- mxAlgebra( expression= ___________, name="cMZf" )

covDZf    <- mxAlgebra( expression= ___________, name="cDZf" )

covMZm    <- mxAlgebra( expression= ___________, name="cMZm" )

covDZm    <- mxAlgebra( expression= ___________, name="cDZm" )
```

## 4-group Genetic Models 2

*Q9.1.* How does the -2 log-likelihood of this 4-group **ADE** model compare with those of the sex-specific **ADE** models you ran earlier? What about the degrees of freedom?

| | females | males | both sexes |
|---|---|---|---|
| -2 log-likelihood | | | |
| degrees of freedom | | | |

*Q9.2.* This statement below prints the unstandardized and the standardized variance components, which were combined in one matrix using the `cbind` function. Comment on which of these sets of estimates is more relevant to compare across sex.

```
round(fitADEq$US$result,2)
```

## Test Quantitative Sex Differences

*Q10.1.* Complete the following code statements of lines 474-482 of the `practicalDay3.R` script, paste it in the box below and evaluate whether the magnitude of the contributions of genetic and environmental factors on BMI differs by sex. Remember, **same label = same parameter**.

```
# Run ADE model - Test for Quantitative Sex Differences of ADE model

modelADE  <- mxModel( fitADEq, name="oneADE4vca" )

modelADE  <- omxSetParameters( modelADE, labels=c(___________), free=___________, values=svPa,

newlabels='___________' )

modelADE  <- omxSetParameters( modelADE, labels=c("VDf11","VDm11"), free=TRUE, values=svPa,

newlabels='VD11' )

modelADE  <- omxSetParameters( modelADE, labels=c("VEf11","VEm11"), free=TRUE, values=svPa,

newlabels='VE11' )

fitADE    <- mxRun( ___________, intervals=T ) fitGofs(fitADE); fitEsts(fitADE)

mxCompare( ___________, ___________)

round(rbind(fitADEq$US$result,fitADE$US$result),4)
```

Note that fully executable scripts are available in the `hmaes/2022/day3/scripts` directory. These scripts (and many more) are also downloadable from the OpenMx scripts library: `hermine-maes@squarespace.com`.

### from 4-group to 5-group Models

*Q11.1.* Let's move on and explore what extra information we can obtain when we include data of opposite-sex twins (DZO), so look for lines 488-629 of the `practicalDay3.R` script.  Here we highlight lines pertaining to the DZO twins. It is critical that they are organized such that twin 1 is one sex (i.e. female) and twin 2 the other sex (i.e. male), or alternatively you can create two groups, one group where twin 1 is female and a group where twin 1 is male. We have re-ordered DZO pairs such that twin 1 is female and twin 2 is male.

```
meanGo     <- mxMatrix( type="Full", nrow=1, ncol=ntv, free=TRUE, values=svMe, labels=c("mZf","mZm"),
name="meanGo" )
```

Note that we equated the mean for twin 1 in DZO pairs to that of the same-sex female pairs, and correspondingly equated the mean for twin 2 in DZO pairs to that of the same-sex male pairs.

With the additional observed statistic (the DZO correlation), we can estimate one additional parameter; either the correlation between additive genetic factors across sex (`rg`, see next page) - or a sex-specific source of additive genetic variance (`VAms`) from which we calculate the genetic correlation across sex. Alternatively, one can estimate the correlation between shared environmental or dominance effects across sex.

```
covAms    <- mxMatrix( type="Symm", nrow=nv, ncol=nv, free=TRUE, values=0, label="VAms11", lbound=.0001,
name="VAms" )
```

Given we're using the direct variance estimation approach, we'll need some extra algebra (copied below) that will allow the variance components to go negative. Remember that if you obtain a negative variance component, your model may not provide a good representation of the data.

```
signA     <- mxAlgebra( ((-1)^omxLessThan(VAf,0))*((-1)^omxLessThan(VAm,0)), name="signA")
covAos    <- mxAlgebra( signA*(sqrt(abs(VAf))*t(sqrt(abs(VAm)))), name="VAos")
pathRg    <- mxAlgebra( signA*(sqrt(abs(VAf))*t(sqrt(abs(VAm))))/sqrt(VAf*(VAm+VAms)), name="rg")
```

Please complete the lines 570-571 of code reproduced below for the expected DZm and DZo covariance.

```
covMZm    <- mxAlgebra( expression= VAm+VDm+VAms+VDms, name="cMZm" )
covDZm    <- mxAlgebra( expression= 0.5%x%VAm+0.25%x%VDm+________+________, name="cDZm" )
covDZo    <- mxAlgebra( expression= 0.5%x%_____+0.25%x%VDos, name="cDZo" )
```

## Non-scalar Sex-limitation Model

*Q12.1.* The general non-scalar sex-limitation model fitted to 5 groups of data estimates one additional parameter which allows us to test whether different sets of genes contribute to the variability of BMI in males versus females - which we also call qualitative sex differences. We can test the significance of these differences by dropping the sex-specific variance component or by fixing the genetic correlation across sex to 1. If this test is significant, in other words if there are qualitative sex differences, then it becomes unnecessary to evaluate further whether there are quantitative sex differences in the magnitude of the variance components. If  different genes are operating in males and in females, it seems unlikely that they would explains the exact same amount of variance in both sexes. If you prefer to estimate `rg` (or `rd`) directly, the lines defining `VAms` & `VDms` need to be deleted and the following lines of code have to be changed.

```
pathRg    <- mxMatrix( type="Full", nrow=1, ncol=1, free=TRUE, values=1, label="rg11", lbound=0, ubound=1,
name="rg" )
pathRd    <- mxMatrix( type="Full", nrow=1, ncol=1, free=FALSE, values=1, label="rd11", lbound=0,
ubound=1, name="rd" )
covDZo    <- mxAlgebra( expression= 0.5%*%rg%x%VAos+ 0.25%*%rd%x%VDos, name="cDZo" )
```

To test the significance of qualitative sex differences - please edit lines 636 and 644 of the `practicalDay3.R` script,

we fix the sex-specific parameter (VAms11) to

or alternatively, we fix the genetic correlation (rg) across sex to

## Testing Qualitative & Quantitative Sex Differences

*Q13.1.* Based on the results obtained after fitting these three submodels, please summarize what this means for your hypotheses about heterogeneity.

*Q13.2.* Congratulations, you have finished the sex limitation practical!

For those of you with more experience with OpenMx, you might try to re-specify the 2-group or 4-group models with definition variables for zygosity (and sex), thus reducing the number of groups.  You might find the scripts from day 2's session that include zygosity as a definition variable helpful to get started.

How do you rate this practical?

## GxE Intro

*Q14.1.*

# G x E Interaction Practical

Now we'll move on to a GxE example which tests another version of heterogeneity. Even though we have so far in this workshop strongly advocated for using the direct variance estimation approach as it provides more accurate tests of significance, the approach does not generalize easily to all scenario's, and it's not always possible to generate equivalent code using path versus variance estimation. One such case is GxE, and in particular when we're dealing with a covariate that is shared across twin pairs, which only allows for scalar GxE or testing for quantitative differences in the variance components by a covariate/moderator.

ps. We have not found a direct variance estimation of this model that is completely equivalent to the path coefficients version. If you can come up with one, please share with us!

Note, we're using BMI data in `twinData` on the adult female twins (both young and older)who range in age from 17 to 88 years, as indicated in lines 667-819 of the `practicalDay3.R` script.

Please check the number of variables in the dataset and their means/variance.

| | |
|---|---|
| Nmz | □ |
| Ndz | □ |
| mean BMI | □ |
| variance BMI | □ |

### GxE ADE Model expected means

*Q15.1.* By now, you're familial with using definition variables to adjust the means of the phenotype for the effects of covariates. In previous examples, we have corrected for the linear effects of age on BMI. Here we extend this by also estimating a quadratic effect of age. To do so, we have recoded age (`ageL`, divided by 100 to make optimization a little easier), and pre-calculated age squared (`ageQ`), and included both covariates (`covVars`) in the data objects. Then we created two matrices to hold the definition variables and use matrix algebra to generate the expected means.

Write out the expectation for the means by working out the matrix algebra.

## GxE ADE Model expected covariances

*Q16.1.* Instead of estimating the variance components directly, as you've seen in previous scripts (copied from earlier script) using the direct symmetric approach:

```
# covA      <- mxMatrix( type="Symm", nrow=nv, ncol=nv, free=TRUE, values=svPa, label="VA11", name="VA" )
```

now, we're estimating path coefficients and using `mxAlgebra` statement to generate variance components. Note that this forces the estimated variance components to be positive. Remember that OpenMx/R are case sensitive - we've used lower case **a**, **d**, and **e** for the names of the matrices containing the path coefficients and and upper case **A**, **D** and **E** for the names of the matrices containing the variance components (or the squared path coefficients), shown below for the additive genetic component. Given that we know that the twin correlations are consistent with an **ADE** model, that's what we are specifying. As an exercise, you may attempt to change this into an **ACE** model.

```
pathA     <- mxMatrix( type="Lower", nrow=nv, ncol=nv, free=TRUE, values=svPa, label="a11", name="a" )

covA      <- mxAlgebra( expression= a %*% t(a), name="A" )
```

Can you work out the predicted means and variances/covariances at the starting values for MZ twins? Are those in the right order of magnitude given the observed means and variances? Remember that reasonable start values will help optimization.

## GxE ADE Model moderation paths

*Q17.1.* Now we create additional matrices for the moderated path coefficients (i.e. `aI`), those we are going to multiply with the definition variables containing the age of the twin pairs (`ageL`). We start these parameters close to zero, assuming no moderation. However, if the variance components change significantly as a function of age, then any or all of the moderated paths would be estimated to be non-zero.

```
pathAI    <- mxMatrix( type="Lower", nrow=nv, ncol=nv, free=TRUE, values=svPaI, label="aI11", name="aI" )

covAI     <- mxAlgebra( expression= (a+ ageL%*%aI) %*% t(a+ ageL%*%aI), name="AI" )
```

How many free parameters are being estimated in model `modelADElqi`?

## GxE ADE Model expected variances 2

*Q18.1.* Remember how we drew the path diagram for the moderation model? Both unmoderated and moderated paths start from the same latent variable, i.e. **A**, suggesting that some of the contributions of genes to the variance of BMI are not moderated by age while other contributions of the same set of genes might be moderated by age, and we will estimate how much is accounted for by each of these effects. When you apply the rules of path analysis to this path diagram for the contribution to the variance of BMI due to **A**, it includes not just the variance components `a^2` and `(ageL*aI)^2` but also twice their covariance `2a*ageL*aI`, which is the result of the matrix algebra `(a+ ageL%*%aI) %*% t(a+ ageL%*%aI).`

(Optional). Work out the matrix algebra to match the expectations from applying the path tracing rules.

## GxE ADE Model expected combined variances

*Q19.1.* Next we use the combined variance components, including the unmoderated and moderated effects of genes (and environment), to generate the predicted variances and covariances by zygosity.

```
covPI     <- mxAlgebra( expression= AI+DI+EI, name="VI" )

covMZ     <- mxAlgebra( expression= _____, name="cMZ" )

covDZ     <- mxAlgebra( expression= ______________, name="cDZ" )
```

Objects that include definition variables or calculations with definition variables are only included in the `mxModel` statements that include the actual data objects, thus we create two lists of objects, one called '`pars`' and one called '`defs`' for objects contain definition variables. Note that `defs` is included in `modelMZ` and `modelDZ` but not in the overall model (further down in the script) `modelACElqi`.

```
pars      <- list( pathB, meanG, pathA, pathD, pathE, covA, covD, covE, covP, ______, ______, ______)

defs      <- list( defAgeL, defAgeQ, ______, ______, ______, covPI, meanAge)

modelMZ   <- mxModel( pars, defs, expMean, covMZ, expCovMZ, dataMZ, expMZ, funML, name="MZ" )

modelDZ   <- mxModel( pars, defs, expMean, covDZ, expCovDZ, dataDZ, expDZ, funML, name="DZ" )
```

Please complete lines 751-752 & 766-767 of the code and paste it in the box below:

## GxE ADE Model Calculate Unmoderated Estimates

*Q20.1.* The next blocks of code (lines 772-801) are not necessary to fit the model and can also be generated with different statements after the model has been fitted (and are thus not evaluated for every iteration). However, if you want to estimate confidence intervals (CIs) around any additional calculated quantities, it has to be done as part of the model, as the CIs are likelihood based.

The first block generates the unstandardized and standardized variance components of the unmoderated components. In case moderation of the variance components is not statistically significant, there would be just one estimate for the relative contributions of **A**, **C** and **E**. The next block is used to indicate which confidence intervals we want to estimate. It requires the ‘`intervals=T`’ argument to the `mxRun` statement to actually tell OpenMx to estimate them.

```
estUS     <- mxAlgebra( expression=cbind(A,D,E,A/V,D/V,E/V), name="US", dimnames=list(rowUS,colUS))
```

```
ciADE     <- mxCI( c("US[1,1:3]" ))#,"AI","DI","EI") )
```

Adjust the code to generate confidence intervals around the standardized variance components. Does the interpretation of the results change?

Yes ○ No ○

## GxE ADE Model Calculate Moderated Estimates

*Q21.1.* However, if moderation of the variance components is significant, then the predicted variance components change as a function of the moderator, in this case, age. We can make use of the power of matrix algebra to generate the predicted (means and) variance components for a range of values of the moderator that reflects the range of the moderator in the observed data. In this example, we included adult twins, with most pairs in the 15-75 age range. We can generate a table with the predicted values across this age range, and we do this in the ‘`UxAge`’ matrix for unstandardized components and in the ‘`SxAge`’ for standardized components.

Play with the values of the `vals` object (line 782) and re-generate the graph below and discuss the changes?

### GxE ADE Model Plot Moderated Estimates

*Q22.1.* Of course, we can use the power of R to generate nice plots with the matplot function from the values generated in the algebras. In lines 827-835 we plot the standardized and unstandardized predicted estimates of the contributions of **A**, **D** and **E** by age, which we pre-calculated in R in the `UxAge` and `SxAge` matrices.

It would be nice to have confidence intervals around the estimates of **ADE** by age. This requires that you estimate those CIs first. To generate similar graphs as above with CIs, we have to reformat the output that contains the CIs (lines 837-846), so that we can add extra dotted lines for the lower (**lci**) and upper (**uci**) CI’s around **A**, **D**, and **E** (lines 848-857**)**. If you come up with more efficient ways to do this, please let us know!

What do you conclude about the causes of variation for BMI and do they change as a function of age? Discuss the advantages and disadvantages of plotting the standardized or the unstandardized estimates.

## GxE ADE SubModels

*Q23.1.* This full GxE model allows adjusting the means for the linear and quadratic effects of the moderator, and furthermore moderating the **ADE** variance components by the moderator, which in this example is age.

Formulate three different hypotheses about how a person's age influences their BMI that you can test as submodes of this first model, generate the R code to test them and discuss the results.

If you get stuck, lines 859-888 provide some examples. We typically test the significance of the moderation of the variance components prior to testing the main effects.

## Block 25

*Q24.1.* **RCR topic**. Models discussed in this session test heterogeneity of means and variance components as a function of a moderator. Discuss the considerations and potential consequences of fitting these models when the moderator is race/ethnicity or ancestry.

## Thank you

*Q25.1.* Congratulations, you have successfully completed this practical!

We hope you expanded your model fitting expertise in OpenMx and leaned how to test genetic epidemiological hypotheses about heterogeneity.

How do you rate this practical?

Powered by Qualtrics