Tutorial 1
Economics and Genetics (FEB13089)

Erasmus School of Economics – Bachelor (Block 2)

Learning goals

- Becoming familiar with basic functionality of *R*.
- Being able to estimate heritability using the classical twin study.

Data

On Canvas, you can find the files Tutorial1.RData and Income.RData you need for Tutorial 1 and the Individual Assignment. Please download these files from Canvas and store them in an easily accessible folder on your PC. In this Tutorial, I'll assume you have stored these files in the folder C:\Users\Niels\Documents\EUR\E&G\Tutorials. Whenever this path to the folder with data is mentioned in this Tutorial, please use your own path to it.

Introducing *R*

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. *R* is the most often used (general) statistical software for genetic analysis, because it is free and has enormous possibilities through the availability of many add-on packages. *RStudio* is a powerful and productive user interface for *R*.

R and *RStudio* are installed on the computers in the PC labs. If you don't have *R* and *RStudio* on your own computer yet, please download and install *R* from https://www.r-project.org/ and *RStudio* from https://www.rstudio.com/products/RStudio/ (The tutorial is based on R version 4.0.2 and RStudio Desktop 1.3.1073). We use *R* and *RStudio* in Tutorial 1 and Tutorial 4.

After installing *R* and *RStudio*, open *RStudio* (by doing so, you automatically launch *R*).[1]

Getting familiar with the *RStudio* environment

Make yourself familiar with the *RStudio* environment. A nice and short introduction to the different "panels" in *RStudio* can be found on YouTube: https://www.youtube.com/watch?v=5YmcEYTSN7k. As also explained in the video, I recommend to store all your commands in a script and to run the commands from there (with the "Run" button). Also because your need to supply your analysis code in the Appendix of your Individual Assignment. *You can open a new script by clicking File → New File → R Script.* Store the new script as script_Tutorial1.R via *File → Save As…* in your data folder C:\Users\Niels\Documents\EUR\E&G\Tutorials.

Setting your working directory

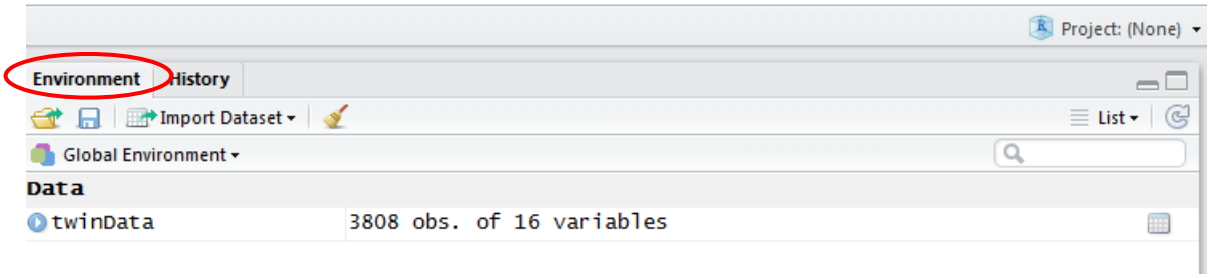
As a final preparation step, we change the working directory of *R* to our folder with data. You can do this with the "setwd()" ("set working directory") command (note that *R* uses *forward* slashes):

```
setwd("C:/Users/Niels/Documents/EUR/E&G/Tutorials")
```

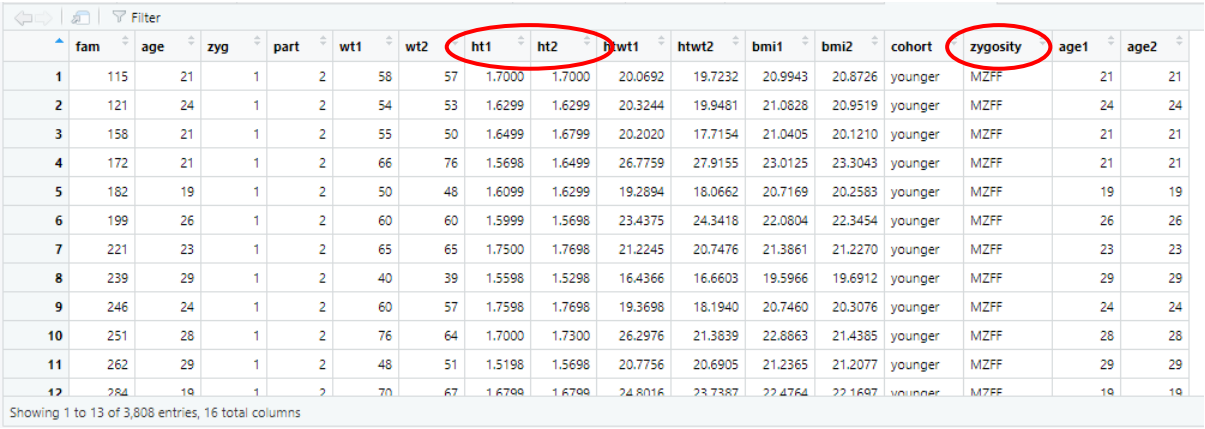
[1] Alternatively, you can also run RStudio via https://rstudio.cloud/. We have checked that you can complete the exercises in the tutorial via this online platform.

Importing data and performing the first calculations

Open the datafile "Tutorial1.RData" via *File* → *Open file*. The data object "twinData" will appear in your "Environment".



Click on `twinData` in your Environment to inspect the contents of this data object. You should see:



Filter

	fam	age	zyg	part	wt1	wt2	ht1	ht2	htwt1	htwt2	bmi1	bmi2	cohort	zygosity	age1	age2
1	115	21	1	2	58	57	1.7000	1.7000	20.0692	19.7232	20.9943	20.8726	younger	MZFF	21	21
2	121	24	1	2	54	53	1.6299	1.6299	20.3244	19.9481	21.0828	20.9519	younger	MZFF	24	24
3	158	21	1	2	55	50	1.6499	1.6799	20.2020	17.7154	21.0405	20.1210	younger	MZFF	21	21
4	172	21	1	2	66	76	1.5698	1.6499	26.7759	27.9155	23.0125	23.3043	younger	MZFF	21	21
5	182	19	1	2	50	48	1.6099	1.6299	19.2894	18.0662	20.7169	20.2583	younger	MZFF	19	19
6	199	26	1	2	60	60	1.5999	1.5698	23.4375	24.3418	22.0804	22.3454	younger	MZFF	26	26
7	221	23	1	2	65	65	1.7500	1.7698	21.2245	20.7476	21.3861	21.2270	younger	MZFF	23	23
8	239	29	1	2	40	39	1.5598	1.5298	16.4366	16.6603	19.5966	19.6912	younger	MZFF	29	29
9	246	24	1	2	60	57	1.7598	1.7698	19.3698	18.1940	20.7460	20.3076	younger	MZFF	24	24
10	251	28	1	2	76	64	1.7000	1.7300	26.2976	21.3839	22.8863	21.4385	younger	MZFF	28	28
11	262	29	1	2	48	51	1.5198	1.5698	20.7756	20.6905	21.2365	21.2077	younger	MZFF	29	29
12	284	19	1	2	70	67	1.6799	1.6799	24.8016	23.7387	22.4764	22.1697	younger	MZFF	19	19

Showing 1 to 13 of 3,808 entries, 16 total columns

The object "twinData" contains data from Australian twins. There are 3,808 observations (twin pairs) in the data. For this tutorial, the most important information in the data object is the zygosity of the twin (zygosity: MZFF = Monozygotic females, MZMM = Monozygotic males, DZFF = dizygotic females, DZMM = dizogytic males, DZOS = dizygotic opposite sex), and the height of each twin (height twin 1 = ht1, height twin 2 = ht2).

Falconer's formula

The easiest way to estimate heritability is to use Falconer's formula, as discussed in Lecture 1. Heritability can be calculated as: $h^2 = 2(r_{mz} - r_{dz})$, where r_{mz} is the phenotypic (trait) correlation among monozygotic twin pairs, and r_{dz} the phenotypic correlation among dizygotic twin pairs. Let's use Falconer's formula to estimate the heritability of height.

We estimate the heritability of height for females and males seperately, which is generally adviced for phenoypes distributed differently (e.g., different means) between sexes. Moreover, dizygotic twins of opposite sex should also be left out of the analysis, because for these brothers and sisters the assumption of sharing half of their genes is not correct because of the sex chromosome difference. Let's start by creating two new dataframes which contain the monozygotic and dizygotic Female twin pairs only, respectively.

```
mz_F = twinData[twinData$zygosity == "MZFF", ]
dz_F = twinData[twinData$zygosity == "DZFF", ]
```

Check in your "Environment" that "mz_F" contains 1,232 observations and "dz_F" 751 observations.

We want to calculate the correlation between the height of twin 1 and the height of twin 2 in our data, and we can access the ht1 and ht2 columns in the object "twinData" with the dollar sign $. The default function to calculate correlations in *R* is "cor". Try running:

```
cor(mz_F$ht1,mz_F$ht2)
```

Unfortunately, we'll get "NA" ("Not applicable") as result. The reason *R* returns "NA" in that there are missing values in the data. You can find them using the "summary" command.

```
summary(mz_F$ht1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  1.420   1.570   1.630   1.623   1.680   1.900      25
```

We see that there are 25 missing values in the column for the height of twin 1. Only by telling *R* explicitly to ignore the observations with missing values, we get the results we want:

```
cor(mz_F$ht1,mz_F$ht2,use="complete.obs")
cor(dz_F$ht1,dz_F$ht2,use="complete.obs")
```

The results show that among the monozygotic twins, the correlation is 0.869. Among the dizygotic twins, the correlation is 0.454. Let's store these estimates as labelled "values":

```
rMZ_F = cor(mz_F$ht1,mz_F$ht2,use="complete.obs")
rDZ_F = cor(dz_F$ht1,dz_F$ht2,use="complete.obs")
```

With these values, we can calculate the heritability of height among females now:

```
2*(rMZ_F-rDZ_F)
```

As answer, we'll get 0.8303215. Hence, h^2 = 0.83 (83%).

- **Exercise:** By repeating the above steps, estimate the heritability of height for males. Start by creating the dataframes mz_M and dz_M containing the data for monozygotic males and dizygotic males, respectively. Make sure you get h^2 = 0.97.

Classical twin study

Let's use a classical twin study now to estimate the heritability of height. For this purpose, we make use of the *R* package "umx". This add-on package needs to be installed first. Go to *Tools* → *Install packages…* Type *umx*, select the package, and click "install". *Alternatively*, install the *umx* package from the command line using the following command:

```
install.packages("umx")
```

The installation of the "umx" package and the associated packages may take some time. To be able to work with add-on *R* packages, you need to load them after their installation using the "library" command:[2]

```
library(umx)
```

The classical twin study can be done with the raw height data, or with the twin pair correlation (similar as for Falconer's formula). Let's practice with the raw data first. The procedure to estimate an ACE model is "umxACE". This command has several inputs: "selDVs" contains the selected

[2] Due to a recent update, some *R* versions (for example, the one running on the university PCs) have some trouble running the "umx" package. This problem can usually be solved by installing the "Rpcc" package using install.packages("Rpcc"). In rare cases, also the "systemfonts" package needs to be installed, using install.packages("systemfonts").

dependent variables from the data frame, in this case we use "ht" (height). It's not needed to specify "ht**1**" and "ht**2**" because "umxACE" expects this structure by default (but "sep" specifies the separator between the variable names as given in "selDVs" and "1" and "2"; in our case there is no spacing, the variables are "ht[no space]1" and "ht2"). "dzData" contains the data for dizygotic twins pairs, and "mzData" the data for monozygotic twin pair. Let's estimate the heritability of height for females as follows:

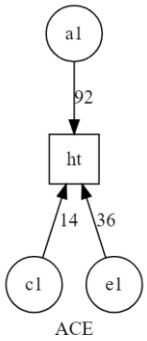
```
umxACE(selDVs = c("ht"), dzData = dz_F, mzData = mz_F, sep = "")
```

You'll get some (polite) warnings when running this command, but the following is the most important output:

```
Running ACE with 4 parameters
ACE -2 × log(Likelihood) = -11985.567
Standardized solution
```

	a1	c1	e1
ht	0.923	0.144	0.357

A visual representation of the results is also given:



Importantly, "umxACE" gives (standardized) path coefficients as output which need to be squared to get the heritability estimates. Hence, h^2 = 0.923 × 0.923 = 0.852. Common environment contributes for 0.144 × 0.144 = 0.021 to the phenotype variance. Unique environment contributes for 0.357 × 0.357 = 0.127 to the outcome variance. Note that 0.852 + 0.021 + 0.127 = 1.000 (100%). The heritability estimate is very close to what Falconer's formula provides.

An advantage of using a classical twin study over using Falconer's formula, it the possibility to compare the ACE model with and AE and CE model. The idea is that you "drop" a path in the model, and that you compare the fit of this reduced model with the fit of the ACE model. If the fit (assessed by the change in likelihood) does not change significantly, then the more parsimonious model is preferred ("Occam's razor"). Let's drop the common environment component first. For this, we use the "umxModify" command. Input is the complete ACE model, which we will save as "ACE_F" first. The other input parameters make sure that the pathway coefficient for the common environment (this is "c_r1c1" in the model) is set equal to 0:

```
ACE_F = umxACE(selDVs = c("ht"), dzData = dz_F, mzData = mz_F, sep = "")
AE_F = umxModify(ACE_F, update = "c_r1c1", name = "AE_F", value=0)
```

Let's use the "umxSummary" function to assess the fit of the reduced model:

```
umxSummary(AE_F, std = TRUE, comparison = ACE_F)
```

The most important output is:

```
Comparison of model with parent model:
```

Model	EP	Δ -2LL	Δ df	p	AIC	Δ AIC	Compare with Model
ACE	4				-19779.57	0.000000	
AE_F	3	0.1391257	1	0.71	-19781.43	-1.860874	ACE

```
Standardized solution
```

	a1	c1	e1
ht	0.93	0	0.36

We see that the fit of the model with the data (as assessed with the change in likelihood of the model, Δ -2LL) has not changed significantly (p =0.71 > 0.05). Hence, this model is preferred over the ACE model (same fit, but more parsimonious). Heritability is estimated to be 0.93 × 0.93 = 0.86 in this model, slightly higher than in the ACE model.

Let's also run a CE model, using similar commands (that is, we now set the path from the additive genetic component "a" to the phenotype, "a_r1c1", to 0 in the model):

```
CE_F = umxModify(ACE_F, update = "a_r1c1", name = "CE_F", value=0)

umxSummary(CE_F, std = TRUE, comparison = ACE_F)
```

Your output should be:

```
Comparison of model with parent model:
```

Model	EP	Δ -2LL	Δ df	p	AIC	Δ AIC	Compare with Model
ACE	4				-19779.57	0.0000	
CE_F	3	519.0178081	1	< 0.001	-19262.55	517.0178	ACE

```
Standardized solution
```

	a1	c1	e1
ht	0	0.84	0.54

We see that in this case, the fit of the model significantly decreased (p < 0.05). Hence, the ACE model is preferred over the CE model.

- By repeating the above steps, estimate the heritability of height among males using an ACE, CE, and AE model. Make sure that in this case also the AE model is the best fitting model, and that in this model you get h^2 = 0.95 × 0.95 = 0.90.

Classical twin study based on correlations

In some cases, you do not have access to raw data but only to twin pair trait *correlations* (for example, the correlations reported in the paper by Nicolaou et al. (2008) discussed in class). In this case, the "umxACE" command can also be used. The covariance (correlations scaled by standard deviations of the phenotype distribution) and the number of twin pairs are then input parameters. Let's calculate these two inputs for the monozygotic and dizygotic (female) twin pairs:

```
covMZ_F = cov(mz_F[,c("ht1", "ht2")],use="complete.obs")

sum(complete.cases(mz_F[,c("ht1", "ht2")]))
[1] 1193
```

```
covDZ_F = cov(dz_F[,c("ht1", "ht2")],use="complete.obs")

sum(complete.cases(dz_F[,c("ht1", "ht2")]))
[1] 730
```

With these inputs, we can run the "umxACE" command again *(Note: this is one line of code)*:

```
ACE_F_cov = umxACE(selDVs = c("ht1", "ht2"), dzData = covDZ_F, mzDat
a = covMZ_F, numObsDZ=730, numObsMZ=1193)

Running ACE with 3 parameters
ACE -2 × log(Likelihood) = -18922.404
Standardized solution

|   |    a1|    c1|    e1|
|:--|-----:|-----:|-----:|
|ht | 0.921| 0.151| 0.358|
```

You can see that the results are very similar to the results obtained with raw data. Let's also run the CE and AE model:

```
AE_F_cov = umxModify(ACE_F_cov, update = "c_r1c1", name = "AE_F_cov", value=0)

umxSummary(AE_F_cov, std = TRUE, comparison = ACE_F_cov)

Comparison of model with parent model:

|Model    | EP|Δ -2LL    |Δ df |p    |AIC |Δ AIC |Compare with Model |
|:--------|--:|:---------|:----|:----|:---|:-----|:------------------|
|ACE      |  3|          |     |     |    |      |                   |
|AE_F_cov |  2|0.1678464 |1    |0.68 |    |      |ACE                |

Standardized solution

|   |   a1| c1|   e1|
|:--|----:|--:|----:|
|ht | 0.93|  0| 0.36|

CE_F_cov = umxModify(ACE_F_cov, update = "a_r1c1", name = "CE_F_cov", value=0)

umxSummary(CE_F_cov, std = TRUE, comparison = ACE_F_cov)

Comparison of model with parent model:

|Model    | EP|Δ -2LL      |Δ df |p       |AIC |Δ AIC |Compare with Model |
|:--------|--:|:-----------|:----|:-------|:---|:-----|:------------------|
|ACE      |  3|            |     |        |    |      |                   |
|CE_F_cov |  2|515.9226548 |1    |< 0.001 |    |      |ACE                |

Standardized solution

|   | a1|   c1|   e1|
|:--|--:|----:|----:|
|ht |  0| 0.84| 0.54|
```

As in the models using the raw data, you'll see that the AE model is the best fitting model.

- By repeating the above steps, estimate the heritability of height among males using covariances as input for "umxACE". Make sure that in this case also the AE model is the best fitting model, and that in this model you get $h^2 = 0.95 \times 0.95 = 0.90$.

Classical twin study: Assortative mating

An important assumption in the classical twin model is that there is no assortative mating. This means that parents have "randomly" chosen their partner (there is no mating on the basis of *phenotypic* similarity; mating based on biological relatedness is termed inbreeding). This assumption is often violated, because *heritable* factors like socio-economic status, cognitive ability, and height often (consciously or unconsciously) play a role in partner choice. When individuals select partners like themselves based on the phenotype, they are also (indirectly) choosing a partner who resembles themselves genetically and culturally. As a result, (positive) phenotypic assortative mating increases the genetic and environmental correlations between relatives. Hence, in a twin study, assortative mating increases the similarity of DZ twins relative to MZ twins.

Silventoinen et al. (2003, https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajhb.10183), show that the genetic correlation among dizygotic twin pairs equals $0.5 \times (1 + \delta_m \times h^2)$. In this formula, δ_m is the selective association between the phenotypes of parents and h^2 the estimate of heritability. Hence, the effect of selection is stronger for phenotypes for which the heritable is higher. You need multi-generational data to estimate δ_m, so we don't do this is our tutorial. However, Silventoinen et al. (2003) estimate that for height the genetic correlation among DZ should be set to 0.62 for females ($0.5 \times (1 + 0.28 \times 0.91 \times 0.91)$) and to 0.61 ($0.5 \times (1 + 0.27 \times 0.92 \times 0.92)$ for males to effectively account for assortative mating.

Let's use this information to re-estimate the heritability of height in our data. This is straightforward, because the only thing that needs to be changed is giving dzAr as additional input in the "umxACE" command. dzAr denotes the (additive) genetic correlation among DZ twins, and equals 0.5 by default. However, we need to set it to 0.62 to account for assortative mating. For females we get:

```
ACE_F = umxACE(selDVs = c("ht"), dzData = dz_F, mzData = mz_F, sep = "", dzAr=0.62)
```

```
Running ACE with 4 parameters
ACE -2 × log(Likelihood) = -11971.628
Standardized solution

|   |    a1| c1|    e1|
|:--|-----:|--:|-----:|
|ht | 0.934|  0| 0.357|
```

So, h^2 is a little higher compared to when we didn't control for assortative mating.[3] Moreover, common environment is estimated to play no role at all.

- Estimate the heritability of height among males, taking into account assortative mating. Make sure you get $h^2 = 0.95 \times 0.95 = 0.90$.

[3] This is somewhat surprising, because based on Falconer's formula we would expect that not controlling for assortative mating leads to an overestimation of h^2 (Lecture 1). When using ACE modelling, we however sometimes see a small upward change in the h^2 estimate when controlling for assortative mating in case the contribution of the common environment to the phenotypic variance is relatively small. This is because the ACE model uses an optimization algorithm to find the best fit between the model and the data.

Individual assignment

For the individual assignment, you need to work with the datafile "Income.RData" (from Canvas). This file contains personal income data from twins. Provide new evidence for the heritability of personal income (among females *and* among males; a separate model for each sex, similar as in this tutorial) using a classical twin study (input = raw data). No adjustment for assortative mating is needed (because we don't know δ_m), but you may want to discuss in your report the influence of assortative mating on your results. In your report, describe the main intuition behind the classical twin study and discuss your findings. Use at least 1 scientifically formatted table to report the estimation results in this section.[4] [+/- 500 words]

[4] For example, have a look (Lecture 1) at the first four columns of Table 4 of Nicolaou et al. (2008) for how to format such a table. Compared to Nicolaou et al. (2008)'s table, you don't need to report confidence intervals (however, note that you can compute these using the "umxCI" command). The umxACE package does not automatically produce all the summary statistics reported in the last five columns of Nicolaou et al. (2008)'s Table 4, but in your table you should at least report the *p*-values for the comparisons between i) the ACE and the AE model, and ii) the ACE and the CE model.