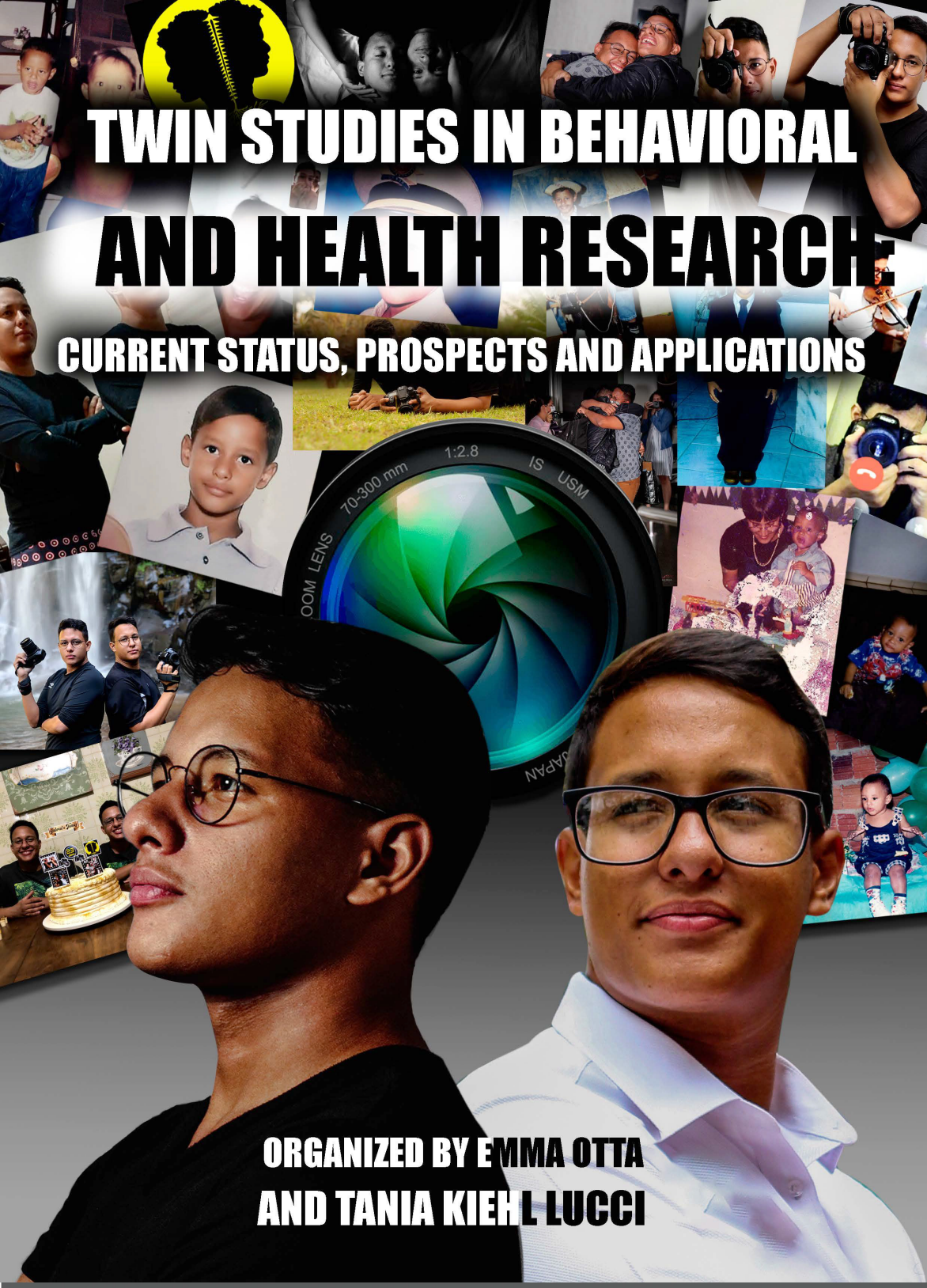

DOI: 10.11606/9786587596167

This book is made available under Creative Commons license to allow others to freely access, copy and use provided the authors are correctly attributed.

Universidade de São Paulo
Reitor – Prof. Dr. Vahan Agopyan
Vice-Reitor – Prof. Dr. Antonio Carlos Hernandes
Pró-Reitor de Pesquisa – Prof. Dr. Sylvio Roberto Accioly Canuto
Pró-Reitor de Pós-Graduação – Prof. Dr. Carlos Gilberto Carlotti Júnior
Pró-Reitor de Pós-Graduação – Prof. Dr. Edmund Chada Baracat
Pró-Reitora de Cultura e Extensão Universitária – Profa. Dra. Maria Aparecida de Andrade Moreira Machado

Instituto de Psicologia
Diretora – Profa. Dra. Ana Maria Loffredo
Vice-Diretor – Prof. Dr. Gustavo Martineli Massola

Departamento de Psicologia Experimental
Chefe – Prof. Dr. Marcelo Fernandes Da Costa
Vice-Chefe – Prof. Dr. Marcelo Frota Lobato Benvenuti

Organizing Committee of the Twin Studies in Behavioral and Health Research: Current Status, Prospects and Applications
Profa. Dra. Emma Otta (IPUSP)
Profa. Dra. Patrícia Ferreira Monticelli (FFCLRP)
Prof. Dr. Claudio Possani (IME-USP)
Dra. Tania Kiehl Lucci (IPUSP)
Dr. Ricardo Prist (IPUSP)

Cover Photo: Tomaz Maranhão
Book formatting: Sofia Barbosa Lima
English Editing Services: Michael Germain and Lisa Burger, MC TRADUÇÕES S/S LTDA

Apoio:

Catalogação na publicação
Serviço de Biblioteca e Documentação
Instituto de Psicologia da Universidade de São Paulo

Twin studies in behavioral and health research: current status, prospects and aplications / Organized by Emma Otta e Tania Kiehl Lucci -- São Paulo, Instituto de Psicologia da Universidade de São Paulo, 2021.

170 p.

ISBN: **978-65-87596-16-7**

DOI: **10.11606/9786587596167**

1. Gêmeos. 2. Pesquisa comportamental. 3. Pesquisa em saúde. I. Otta, E., ed. II. Lucci, T. K., ed. III. Título.

Ficha elaborada por: Cristiane de Almeida Camara - CRB 5384/08

Chapter 3

Mixed models and statistical analysis of twin data

Vinicius Frayze David

Although I am a psychologist, I have been working with statistics for a number of years now. Usually, psychologists do not receive good statistics training in their undergraduate course, and I can say that, in the beginning, it is difficult to understand even the basics, but it is definitely worth the effort. Knowing statistics helps us better understand our data and other researcher`s studies, and think about new approaches to our research questions. My aim here is to address the overall aspects of using linear mixed models in twin designs. I will use almost no mathematics, because the aim is to show what these models can do more than how they work, and I will also show an example of how they can be applied using Stata software. More information on the mathematics involved can be found elsewhere (Wang et al, 2011). I hope that this chapter serves as an introduction for researchers who are not well versed in the issues surrounding twin data mixed models.

When I talk about statistics with other researchers, most of them view it according to its purpose: to teach people how to use a limited sample and make intelligent and accurate conclusions about a large population (Lammers & Badia, 2004). In this sense, statistics is interpreted as a tool and a means to an end. However, it is also a constantly changing field of knowledge, and we have to keep track of new developments that can help us in our studies. We should always be careful about statistics in any field of Experimental Psychology, but when we work with twin designs, even data from the most straightforward experimental design can be challenging to deal with.

What are the issues involving twin designs? One of the most uncomplicated designs is comparing the distribution of two groups with a particular observable trait. In this case, we have a "control group" and an "experimental group" with its participants and their measured trait (Figure 3.1a). This design is simple and allows us to compare different characteristics of intergroup trait distribution using means, standard deviations, medians, and frequencies, among others.

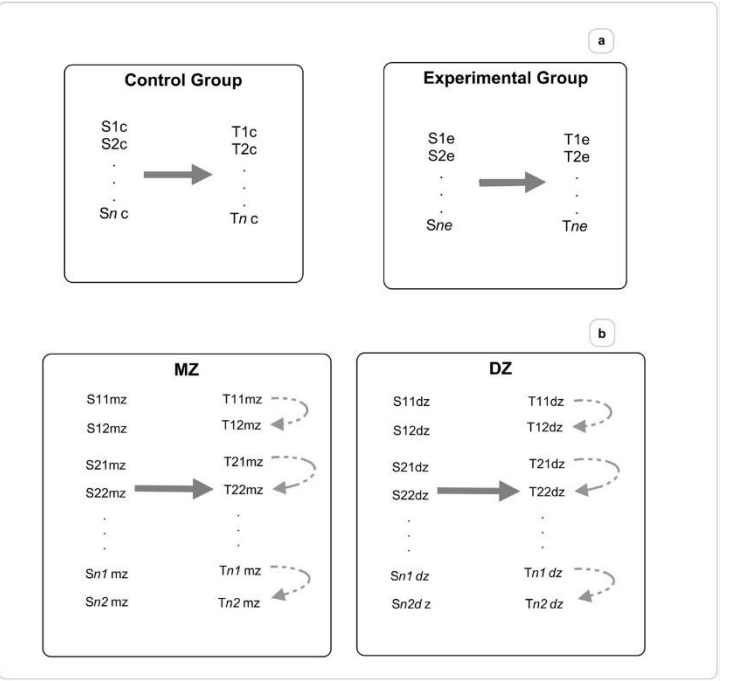

FIGURE 3.1. DATA STRUCTURE IN (A) TYPICAL DESIGNS, AND (B) TWIN DESIGNS

When we use the same design with twins, at first glance, it does not seem very different. We still have two groups of participants and their measured trait. Here, I am separating monozygotic (MZ) and dizygotic (DZ) participants because this is usually what we want to compare (Figure 3.1b). The problem is that these are not independent participants, as in the first case. In a twin design, we have pairs of participants: 11mz and 12mz, 21mz and 22mz, and so on. Since they are pairs, we expect some covariance between them in the measured trait and are interested in the value of this covariance. After all, if we find that MZ twins have a greater covariance than DZ twins, we can surmise that this observed trait involves some genetic influence.

For example, if we measure MZ and DZ pairs data, we may find that MZ pairs are much more similar than DZ, so there is probably a genetic influence on this trait (Figure 3.2).

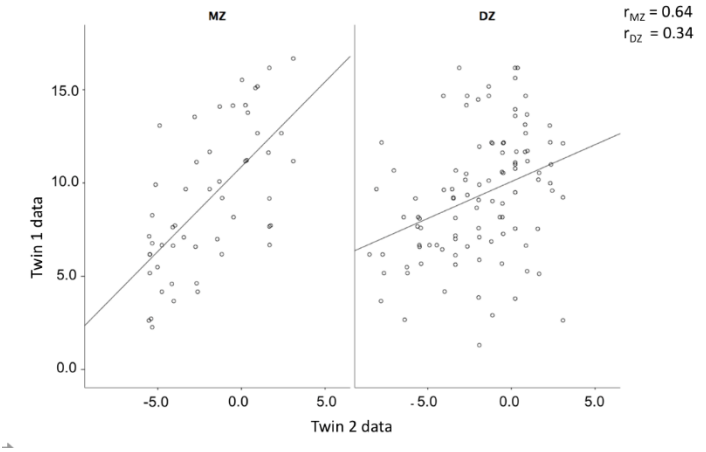

FIGURE 3.2. RELATIONSHIP BETWEEN HYPOTHETICAL DATA OF MZ AND DZ PAIRS

What do I mean when I say that there is a genetic influence? We know that heritability is defined as the portion of phenotype variability attributed to genetic variation. One of the most common approaches to calculate heritability is to use the ADCE model.

The ADCE model assumes that the variation of any individual trait is influenced by genetic and environmental variability, which can be divided into five different effects. The genetic effect is composed of the (1) Additive effect, (2) Dominance effect, and (3) Epistasis effect, while the environmental effect consists of the (4) Common environmental effect, and (5) Unique environmental effect.

Briefly, additive genetic effects (A) are those involving direct action of each allele of homologous chromosomes, so that each adds a direct value to the phenotype; dominant genetic effects (D) result from the joint action of homologous chromosomes; epistasis (I) is an effect resulting from the joint action of alleles on different loci. The common environment effect (C) is the result of the twins' common experiences, usually the family environment, parents, home, and others; and the individual environment effect (E) is the sum of the different experiences of each individual, along with errors of measurement, which are also individual.

Given that heritability is a relationship between the genetic and phenotypic variances, we can formulate heritability according to the ADCE model as:

$$H^2 = \frac{VarG}{VarF} = \frac{VarA + VarD + VarI}{VarA + VarD + VarI + VarC + VarE}$$

We can use variations of the ADCE model if we exclude some of the factors. One of the most common is the ACE model (Maes, 2005), in which we consider all genetic variation to be an additive effect. Mathematically, it is an easier model to work with because it assumes that the increase in observed trait differences is directly related to a difference in the genotype. In this case, we assume that the similarity in an observable trait due to genetic variation in MZ should be twice as large as in DZ.

We are interested in variances and covariances, so how can we calculate them? Two of the most widely used techniques are the intraclass correlation coefficient (ICC), and structural equation modeling (SEM) (Franic et al, 2012). The ICC quantifies the degree to which individuals with a fixed degree of relatedness resemble each other. It can be interpreted in the same way as a Pearson correlation, which varies from -1.0 to +1.0, where the closer to 1.0, the greater the similarity between the siblings (negative values are not typically expected). The most significant difference from a regular Pearson correlation is that the ICC uses the pooled mean of all the data and its standard deviation, whereas in the Pearson correlation, each variable is centered and scaled by its own mean and standard deviation. ICC is more accurate for twin designs because, when using it, the order of the pair is not important and there is usually no good reason to select a twin as number one or number two (which would be the variables in a Pearson correlation). There are different ICC models, but I will not discuss them here, and more information can be found in Koo and Li (2016).

Structural equation modeling (SEM) has been widely used in twin studies. It is a highly complex and versatile model, containing a set of methods that check hypotheses about the structure of the relationships between observed and non-observed (latent) variables (Kaplan, 2008), as defined by the researcher. It is typically represented as a path diagram, in which the paths constitute the set of model parameters. Covariances can be established or calculated for all paths as well as the variances, making it a very interesting model for twin designs. For example, the covariances of additive effects can be set at 1 for MZ, and 0.5 for DZ, and/or dominance effects at 1 for MZ, and 0.25 for DZ, and then calculate the other parameters. Several parameters can be obtained from the models, which also allows researchers to determine model goodness-of-fit and compare different models.

Although ICC and SEM are interesting approaches for studies, they have some limitations. The most notable limitation of ICC is that it compares only two sets of data, such as MZ or DZ siblings. If we are interested in studying other variables such as sex or age, several analyses must be conducted. For sex, we will have to calculate one coefficient for male MZ, another for female MZ, and then for male DZ and female DZ. This increases the likelihood of type I error and creates a need for larger sample sizes.

With respect to SEM, we know that most of the procedures that have been suggested involve non-standard and complex model specifications that are challenging for the average user and therefore susceptible to error, especially because some of the most promising models are not easily available in conventional SEM software (Tomarken & Waller, 2005). Moreover, convergence problems have been observed with some procedures, which may not work properly. Finally, it requires large sample sizes - some rules of thumb suggest at least 25 observations per parameter.

As such, we have mixed models as an alternative. The main difference between a linear mixed model (LMM) and a general linear model such as analysis of variance (ANOVA) is that an LMM includes both fixed and random effects (Baltagi, 2008). Random effects assume that the data come from a hierarchy of different populations and that the differences are related to this hierarchy. In other words, there is an assumption that individual traits are not related only to the independent (fixed) variables because non-random errors are present.

Mixed models are widely used in educational and health studies. One example is the comparison between the performance of male and female children on a test when we have data from more than one school in each group. The children's sex is our fixed factor, but we have to consider the school in our model because we expect to have some covariance in our data due to the school. Some schools may have better facilities and more qualified teachers than others, among several other differences. If we assume that test results can be influenced by the school, although not to the same extent as sex, we can include it in our model as a random factor. The idea of including schools as a random factor can serve both to control for this possible effect and calculate how large this effect can be. Mixed models can be used at several hierarchy levels, such as classrooms, schools and type of school (public or private), and can also include different effects for each level. However, for the purposes of this chapter, we will only discuss the inclusion of covariances between siblings in twin designs.

The logic of having a random effect has been adapted to twin designs. We are usually interested in some fixed factors and covariates

such as sex or age, but we also expect the errors of our participants' individual trait measures to be related to the error of their siblings. Thus, in twin designs, we use the pair of twins as our random factor, making it possible to calculate the covariance between them. Since we know that the total variance of any mixed model is the sum of the variance of the random factors and the residual variance, in our case the total variance will be the variance of the pairs plus the residual variance.

When dealing with twin data, we also have a second problem. The first problem I discussed was how to consider and calculate sibling covariances, and this is similar to many other studies, and not much different from the school example I used before. But when we use the school as a random factor, we can assume some form of regular distribution among schools, and use the school as a unique random factor. With twins, we want to calculate at least two different and very specific covariances to investigate the extent to which our trait can be considered heritable. So, what must we do? We need to separate DZ from MZ covariance in our model. This can be done using mixed models. Here, I used an adaptation of what can be found in Twins Research Australia (https://www.twins.org.au/). Covariance, which is a function of the twins, whether they are MZ or DZ, can be separated from the "extra" covariance because they are an MZ pair. In other words, we can examine the difference between the covariance of MZ and DZ pairs. Thus, our total variance will now be the sum of the variance of the pair, the extra variance of MZ pairs and the residual variance.

How can this be achieved? I will show you an example using Stata software. This analysis can also be carried out in R, SAS, or SPSS using the appropriate commands.

First, we need to organize our data set (Figure 3.3). Each participant must be in a different row and we need a variable to identify each pair. You can use any number, as long as it is the same for each pair and different pairs have different numbers. Then, the next columns can contain your variables of interest, such as zygosity, sex, or any other – the same as in any other analysis. The "trick" is to create three additional variables responsible for separating MZ covariation from DZ covariation. First, you have to identify your pair of participants as MZ or DZ twins, and the easiest way to do that is to create a column in which you assign 0s to DZ and 1s to MZ twins. Remember that you have to assign these values to each participant, even knowing that the sibling will have the same number. Next, you create two new variables that I call dz1 and dz2 in this example. You will have to assign each of the DZ twins from each pair to one of these

new columns, using 0s and 1s again. The first twin will be 1 and 0, and the second 0 and 1. It does not matter which one is which, as long as they are assigned differently. MZ will only have 0s here since they were already defined in the previous column.

	A	B	C	D	E	F	G	H
1	pairid	zygozity	male	age	mz	dz1	dz2	Closeness
2	1	1	0	3	1	0	0	1.32
3	1	1	0	3	1	0	0	0.26
4	2	1	1	2	1	0	0	0.85
5	2	1	1	2	1	0	0	0.16
6	3	0	1	6	0	1	0	1.32
7	3	0	1	6	0	0	1	1.34
8	4	1	0	2	1	0	0	0.81
9	4	1	0	2	1	0	0	3.32
10	5	0	0	6	0	1	0	1.32
11	5	0	0	6	0	0	1	0.81
12	6	0	0	5	0	1	0	3.36
13	6	0	0	5	0	0	0	1.34

FIGURE 3.3. ORGANIZATION OF THE DATASET FOR MIXED MODEL ANALYSES

Then we have our commands. Below are two examples that can be modified to fit different designs. First you declare that you are using a mixed model, then you have your observed trait; subsequently the fixed factors you are interested in – in the first command, I used only "male", the sex variable, and in the second, I declared a more complete model. The most important part is what comes next, when you need to declare your two random effects. The first is the pair random effect, irrespective of whether it is MZ or DZ, and it will calculate the covariance of the pair that is common to MZ and DZ twins. The other effect is only valid for MZ, and it will calculate the difference of covariance between MZ and DZ pairs. Then you can specify the structure for the covariance matrices of the twins. You can usually consider it to be identity.

The main difference between these two commands is that the first uses maximum likelihood estimation with a chi-square distribution, and the second a restrictive maximum likelihood estimation (reml) with a t distribution. As a rule of thumb, if you do not have reliable information to choose between them, and your sample is small, you should use the reml, and if it is large, you can use maximum likelihood, a more powerful model (less chance of type II error).

Commands:

- mixed Closeness male, || pairid: || pairid: mz dz1 dz2, covariance (identity) nocons

- mixed Closeness mz##male age || pairid: || pairid: mz dz1 dz2, reml cov(id) nocons dfmethod(residual)

Now we can look at our outputs. This first one is like any other general linear analysis: there are estimates, errors, statistical values, and p-values for each of the fixed factors and covariates. There was a significant effect of zygosity and age, but not sex (Figure 3.4a). It is important to underscore that these effects take into account the covariation between pairs.

a.

```
Mixed-effects REML regression                   Number of obs     =      1,300
Group variable: pairid                          Number of groups  =        650

                                                Obs per group:
                                                              min =          2
                                                              avg =        2.0
                                                              max =          2
DF method: Residual                             DF:           min =   1,295.00
                                                              avg =   1,295.00
                                                              max =   1,295.00

                                                F(4, 1295.00)     =       8.80
Log restricted-likelihood = -2668.199           Prob > F          =     0.0000
```

Closeness	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mz						
MZ	.5619628	.1813237	3.10	0.002	.2062424	.9176831
male						
Masculino	-.2294931	.1631751	-1.41	0.160	-.5496096	.0906235
mz#male						
MZ#Masculino	.2772218	.19641	1.41	0.158	-.1080949	.6625385
age	.096499	.0281407	3.43	0.001	.0412927	.1517054
_cons	-.6525122	.186734	-3.49	0.000	-1.018847	-.2861779

b.

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
pairid: Identity				
var(_cons)	2.187427	.2297072	1.78052	2.687326
pairid: Identity				
var(mz dz1 dz2)	1.444014	.2060063	1.091785	1.909878
var(Residual)	1.12134	.0873257	.9626069	1.306248

LR test vs. linear model: chi2(2) = 327.14 Prob > chi2 = 0.0000

FIGURE 3.4. OUTPUTS OF MIXED MODEL ANALYSIS

The exciting part is in the other table, which contains the values of our variances (Figure 3.4b). The first, var(_cons), is the portion of the variance due to pair covariances, which shows how the pairs of siblings are related to each other, regardless of their zygosity. The second,

var(mz, dz1, dz2), is the increase of variance explained by the covariance being MZ and is only valid for MZ twins, and the third is residual variance, the portion not explained by the fact that they are siblings. Remembering the previous formula, total variance is the sum of pair variance, the extra variance of MZ and residual variance. Now we have the following:

$$VarTOTAL = VarPAIR + VarMZ + VarRESIDUAL = 2.19 + 1.44 + 1.12 = 4.75$$

What is the common variance of the pair, our first component? If we are using an ACE model, what is common for every pair? We have at least half of the genetic similarity (1/2A) and the common environment (C). The extra MZ variance is half of the genetic covariance that was missing (1/2 A), since MZ twins are expected to double their genetic similarity compared to DZ twins. Residual variance is what is explained by neither the additive effect, nor the common environment. In an ACE model, we can assume it is the portion of variance due to the unique environment effect (E). The sum of these three variances is the total variance in our sample. Putting this in numbers, we can conclude that, in this example, 15.6% of the trait variation is due to common environmental variations, 60.7% to additive effects, and the unique environmental variation is responsible for the remaining 23.6%.

It is important to consider that mixed models also have limitations. First, I showed you how to perform the analysis with an ACE model, and we know that this model can overestimate the genetic effects, so we need to keep this in mind. It is possible to include Dominance and Epistasis effects in the analysis, but it becomes much more complicated, resulting in the loss of the advantage of a simple model. More details can be found in Maes (2014).

Here, I only analyzed one independent variable at a time, but several independent variables could have been considered simultaneously. This is not difficult to do, since we just need to include another random factor: the participant. I also used a linear model, but could have used generalized mixed models, which consider different distributions, such as binary, ordinal, gamma, Poisson, and others. Finally, as far as I know, if you have both observed and latent independent variables, there is as yet no solution for mixed models

References

Baltagi, B. (2008). *Econometric analysis of panel data*. New Jersey, EUA: John Wiley & Sons.

Franić, S., Dolan, C. V., Borsboom, D., & Boomsma, D. I. (2012). Structural equation modeling in genetics. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (p. 617–635). New York, EUA: The Guilford Press.

Kaplan, D. (2008). *Structural equation modeling: Foundations and extensions* (Vol. 10). Los Angeles: Sage Publications.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15(2)*, 155-163.

Lammers, W. J., & Badia, P. (2004). *Fundamentals of behavioral research*. Belmont, Australia: Wadsworth Publishing.

Maes, H. H. (2005). ACE Model. *Encyclopedia of Statistics in Behavioral Science.* Chichester, UK: John Wiley & Sons

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology, 1,* 31-65.

Wang, X., Guo, X., He, M., & Zhang, H. (2011). Statistical inference in mixed models and analysis of twin and family data. *Biometrics, 67(3),* 987-995.