

Instructions & Introductions

Room .

2022 International Statistical Genetics Workshop

Assumptions and biases in biometrical models and extended twin family designs

FIRST QUESTION (to be answered by ONE person, the Scribe, in your room): Which zoom breakout room are you in (what number)?

Q1.

Today you'll be working together as a group to better understand CTD and NTFD models, their instantiation in openMx scripts, and biases in those models if assumptions are violated.

INSTRUCTIONS:

1) Everyone can open the Qualtrics survey, but **only the group Scribe (more on that person below) should answer questions in the survey. I.e., one set of answers per group.** This

Qualtrics survey will guide you through the practical and help the tutors understand where groups are at and how they're doing.

2) Throughout the practical, you will be asked to work out solutions to problems and to share your solutions using this survey. Some of these problems have exact numerical answers. Others are open-ended thought questions. Please discuss and debate your answers among yourselves before asking the Scribe to put down your group answer in Qualtrics. Everyone should try to participate. After you have collectively come up with your best answer, put that answer in the answer box.

3) You will also be working through an R script (CTD.NTFD.R) alongside the Qualtrics survey. Everyone should have the R script open on their computer and everyone should run it together at the same time. When prompted in the Qualtrics survey, run the relevant Section of the R script line by line (e.g., hitting Ctr + Enter, or Apple + Enter) or in small chunks of code. The R scripts are broken up into Sections. Please don't skip ahead - keep pace with the Qualtrics survey.

IMPORTANT: Please turn on "soft wrapping" in your RStudio Server session so that you can easily read the comments. To do this, in RStudio Server, go to "Tools" -> "Global Options" -> "Code" and then check "Soft-wrap R source files".

4) **Groups should ask one person (Group Leader; see below) to share their screen.** This can be the same person as the Scribe but it might be better to have them be different people because the Scribe will be busy writing. The Leader should share their screen with the group, alternating between the Qualtrics survey window and the RStudio window. Everyone should work through the survey and script together. The Leader should read the Qualtrics questions and also read the parts of the R script denoted "NOTE:"

5) After everyone in the group has finished a particular Section of the R script, the group should pause to discuss it and any questions they have about it.

6) No cheating: Don't skip ahead in the survey or run the whole R script to get answers. The point of all this isn't to get a good "score" - it's to learn!

7) Try to get through all sections, but Section I is a bonus section for those who are up for a challenge. You can skip it and come back to complete it after you've done the other sections.

Q1.2.

Please spend a couple of minutes introducing yourselves. You will need to then choose two people who will serve two different roles in your group:

Group Scribe: This person will submit the groups' answers in this Qualtrics survey.

Group Leader: The Leader doesn't need to know more about the topic; they're just the reader/screen sharer, but that doesn't quite roll off the tongue like "Leader" does. The Leader will share their screen, alternating between the Qualtrics and RStudio windows. They will read aloud the Qualtrics questions as well as specific comments in the R script that begin with the word "NOTE". The comments following the word "NOTE" are ones that I think are especially important for you to understand. Groups should discuss these comments among themselves to ensure that they're understood.

People other than the group Leader are free to share screen if there are particular points to be made. Remember that anyone can also share code or instructions via chat.

As you're going through this tutorial, we'll call you back into the main room once or twice to check in and take a break.

There are hints built into some of the questionnaire. Also, answers to some questions come later on in the Qualtrics survey or the R script (remember, no cheating). If you need more help please click the 'Ask for help' button in Zoom and someone will be by.

To get started ssh into the workshop server. Use `mkdir` to make a working directory for today's tutorial and copy the files from `/faculty/matt/2022/CTD.NTFD/` to your working directory. So navigate to where your working directory is, and then type the following command (the period at the end is part of the code):

```
cp -r /faculty/matt/2022/CTD.NTFD/ .
```

You will find the R script (CTD.NTFD.R) we'll be using in that folder. You can open it in RStudio Server, but don't start working on it quite yet. I'll tell you when to do so.

[Page 1/25]

Q2. Random Ice-Breaker: Each person, tell us about your worst haircut (or worst fashion choice) ever.

Q4. What are the first names of your group members?

Q5. What is the name of your group's Scribe? (The person who will write answers in this survey)

Q6. What is the name of your group's Leader? (The person who will share screen and read the qualtrics questions and "NOTE" comments in R)

Q7. How many people in your group have written and run a classical twin design model in openMx before this workshop?

- 0
- 1
- 2
- 3+

Q8. How many people in your group have written and run an extended twin-family design model in openMx before this workshop?

- 0
- 1
- 2
- 3+

[Page 2/25]. .

R Script Section A: CTD ACE

A1.

CTD Model

Today you'll be working along with an R script that has CTD and NTFD models in it and works on simulated data. Before getting to that, however, let's do a little work by hand.

The trait we will be analyzing today is simulated. Simulation is a crucially important tool in behavioral genetics, especially for models (like ETFDs) that are complicated and provide estimates that are difficult/impossible to verify mathematically. But for sake of fun, let's say the first trait we're analyzing is reading ability among young adults.

As you will see in a bit, this trait has variance ~ 1 . In our sample, there are 2901 MZ twin pairs whose covariance is .594, and there are 3683 DZ twin pairs whose covariance is .303.

Given this, what type of model would you choose to run?

- ACE model

- ADE model
- ADCE model
- AE model

A2. Why did you choose to run the model you selected above?

[Page 3/25]. .

A2.1. ANSWER: If you chose to run an ACE model, good job! $2CV(DZ) > CV(MZ)$, so an ACE model is appropriate. Some groups may have chosen an AE model because, by looking at the covariances, it's obvious that your estimate of VC will be very small. That's not a bad answer, but we should let the analysis inform us if we should drop a parameter. Also, there's some debate on whether we should report estimates from full or reduced models. Reduced models are saying that "this estimate is 0", and thus tend to be more biased than full models. On the other hand, reduced models estimate fewer parameters and tend to have higher precision. There's no one right answer on that: perhaps we should always report both sets of results.

[Page 4/25]. .

A5. OK, it's time to start running the R script, which is named "CTD.NTFD.R". It should be in your working directory.

Run the start of the script. Then please work through to the end of Section A, line by line or perhaps in small chunks of code. Stop at the end of Section A. The group Leader should share their screen, and everyone should work through it at about the same pace, pausing if there are any questions or comments and reading any comments beginning with "NOTE".

Once everyone has reached the end of Section A, spend a few minutes discussing the questions at the end of that section, and then answer them below.

A3.

In any event, as noted, $CV(MZ) = .594$ and $CV(DZ) = .303$. What is the algebraic solution (also called a "Method of Moments" solution) for your estimate of VA in an ACE model? I'm looking for a number.

A4 . What is the algebraic solution for your estimate of VC in an ACE model?

[Page 5/25]. .

A6 . Everyone should have had different start values for their parameters. Did these start values influence your final estimates? What does this tell you?

A7. How did your Maximum Likelihood estimates from openMx compare to the estimates you derived algebraically?

A7.2. ANSWER: Your MoM and ML estimates should be very similar, but not exactly the same - they are different estimators after all. And the fact that you get the same estimates despite different start values tells you that your model is (probably) identified.

[Page 6/25]. .

R Script Sections B & C - Sensitivity Analysis

BC1. Don't run Section B quite yet.

In the usual CTD ACE model above, we assumed $VD=0$. What if $VD \neq 0$? Well, we know that would bias our estimate of VA upwards by $3/2 VD$ and bias our estimate of VC downwards by $1/2 VD$. On the other hand, if we fit a model assuming that $VD > 0$, that would mean our new estimate of VA would be smaller by $3/2$ of whatever we assume VD to be, and similarly for VC , which we know is biased downwards by $1/2$ of whatever VD is.

Use this logic to derive what our estimate of VA will be if we assume that $VD = .05$ (by fixing it to be $.05$ in our model) rather than assuming it is 0 . Write your answer to what this new estimate of VA would be (a numeric answer) in the box below. (Recall that $CV(MZ) = .594$ and $CV(DZ) = .303$).

BC2. What will our estimate of VC be if we assume that $VD = .05$ (by fixing it to be .05 in our model) rather than assuming it is 0?

BC2.1. You will be able to check your answers above by comparing them to the output from your R script.

[Page 7/25]. .

BC3. OK, now run both Sections B and C in R, line by line.

Let's go ahead and run two "sensitivity analyses", where we fix $VD=.05$ instead of 0 (Section B) and then we fix $VD=.12$ instead of 0 (Section C). Stop at the end of each Section to discuss. You should be able to check your answers above by comparing it to your R output. Once you've finished Section C, stop and answer the questions below.

BC4. Why did your estimates of VA and VC change when you fixed VD to be different than its usual assumed value of 0 (in an ACE model)?

[Page 8/25]. .

BC4.2 . ANSWER: These estimates changes because there is a trade-off between estimates of VA and VC in the CTD. When we assume $VD=0$, we're saying that ALL of the MZ-DZ covariance difference is due to VA. If we assume that $VD > 0$, then we're allowing some of the MZ-DZ covariance difference to be due to VD, which lowers estimates of VA and thereby raises estimates of VC.

[Page 9/25]. .

BC5. How biased would your estimates of VA and VC in the usual CTD have been if VD were actually .12?

- Give me a hint
- Answer here

BC5.2.

HINT: You already know what the estimates of VA and VC are in the usual CTD (from Section A). Just compare those to the estimates you got in Section C, which gives you the ML estimates if VD were actually .12. Alternatively, figure it out algebraically.

BC6 . There are two approaches to estimating parameters in biometrical models: the path coefficient approach (i.e., estimating the path coefficients and setting latent variances to 1) and the direct symmetric approach (estimating the latent factor variances and setting the

path coefficients to 1). We've used the path coefficient approach because that's the approach we need for ETFDs and it therefore keeps our modeling approach consistent.

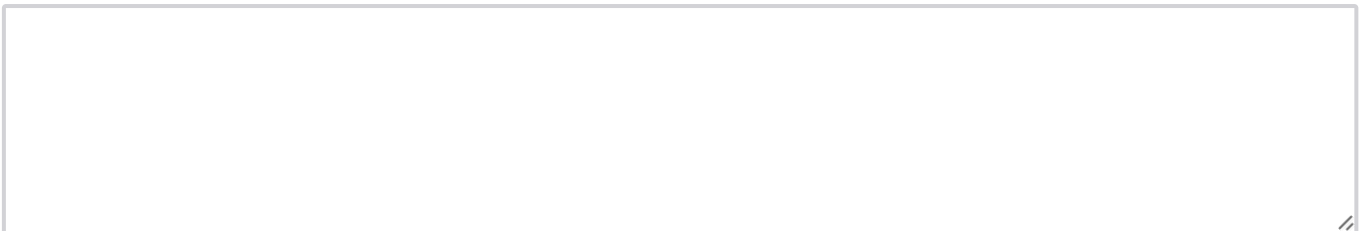
Thought question: had we tried to fit an ADE model to these data using the path coefficient approach, what would our estimate of VD have been?



BC6.2. ANSWER: When $2CV(DZ) > CV(MZ)$ (where we would get a positive estimate of VC), your model would "like" to estimate VD negatively, but it's unable to do so using the path coefficient because variance estimates are inherently constrained to be greater than or equal to 0. So this estimate will hit the 0 boundary (VD estimate will be = 0).

[Page 9.5/25]. .

BC7. Had we tried to fit an ADE model to these data using the "direct symmetric" approach, would our estimate of VD have been some positive number, 0, or some negative number? Why?



[Page 10/25]. .

R Section D: ADCE Twin Models

D1. Don't Run Section D yet.

Briefly explain why we cannot fit an ADCE model using twins only.

D2. If we attempted to fit an ADCE model using twins only, what would alert you to the fact that the model is not identified?

D2.1. ANSWER: We can't fit an ADCE model using twins only because there is insufficient information (2 covariances) to estimate 3 parameters (A, D, and C).

There are two ways to know this:

1. use `mxCheckIdentification()`
2. note that (a) the estimates you get depend on their start values and (b) the -2LL model fits when you get different estimates (from different start values) are all identical.

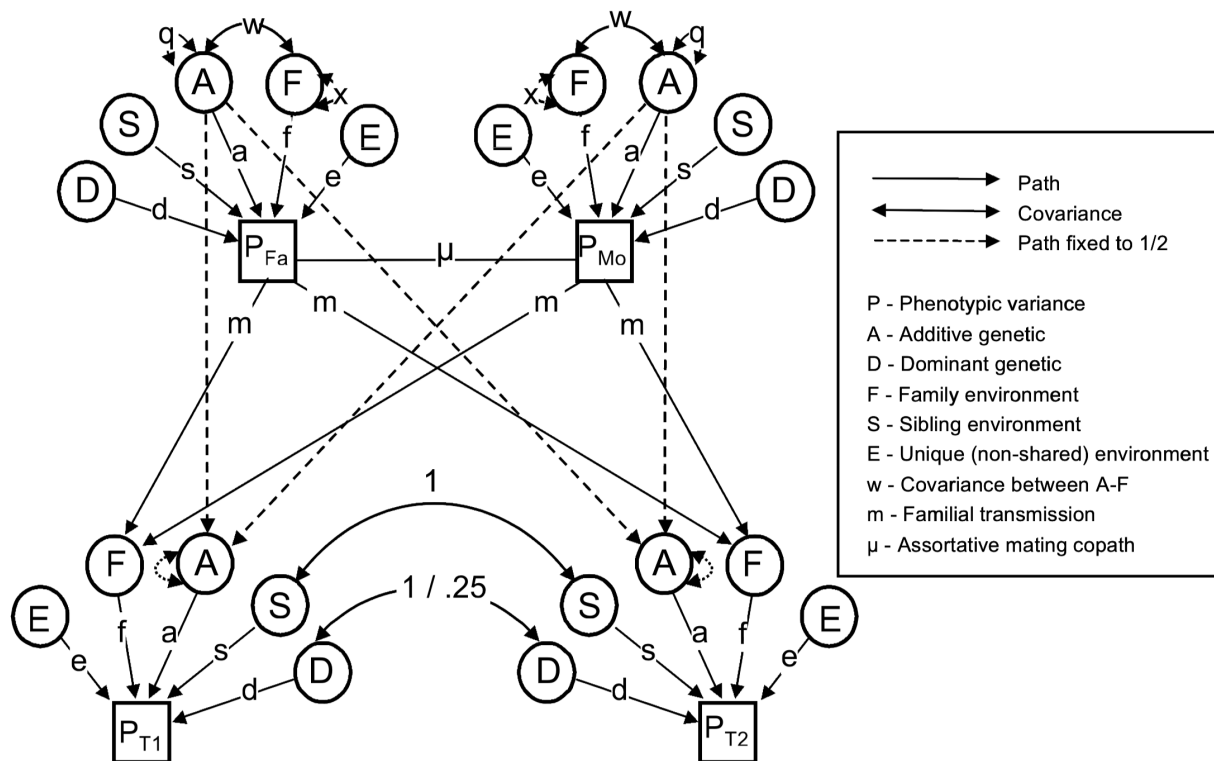
Note: this is not to say that ANY parameters are possible, but there are a set of three

estimates that all have the same -2LL and are mathematically indistinguishable. That's not to say all are equally BIOLOGICALLY plausible. After all, negative variances seem unlikely ;). Also, parts of the space where $VD > VA$ are unlikely.

[Page 11/25]. .

R Script Section E: NTFD on data 1

E1 . Below is a path diagram of the full NTFD model. It includes VA, VD, VS, VF , 3 of the 4 of which are estimable in this model. In our first NTFD model, we will assume $VF = 0$ (thereby setting $w, x, f,$ and m to be 0) and estimate $VA, VD,$ and VS .



Go ahead and run Section E to fit an ADSE NTFD model. This will pull in the parents of the

twins as well (thus 4 relatives per family and 6 bits of information to help estimate factors that lead to within-family similarity per family rather than 1).

After you've finished running Section E, answer the questions below.

E2. How do you know your NTFD model is identified?

E3. What does the parameter "q" quantify in this NTFD?

E3.2. ANSWER: q quantifies the increase in the additive genetic variance arising from the influence of assortative mating (AM). AM causes all causal variants genome-wide to become positive correlated, such that increasing alleles are slightly more likely to be in a genome with other increasing alleles and vice-versa. This increases the additive genetic variance in the population.

[Page 12/25]. .

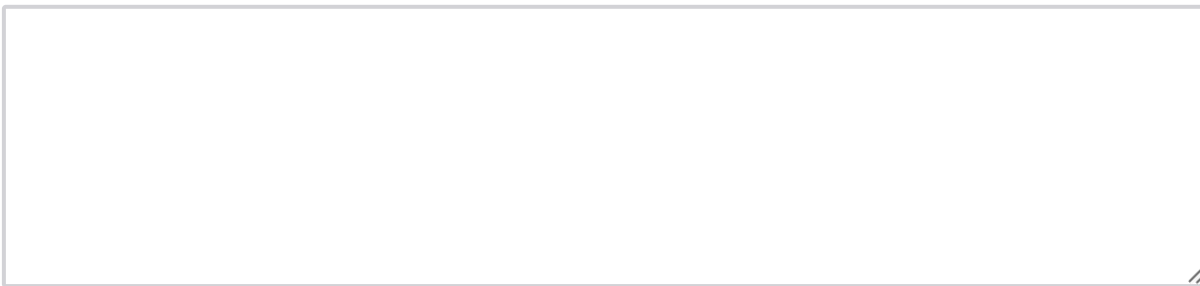
E4. Why does AM tend to bias estimates of VC upwards in the CTD?



E4.2. ANSWER: AM increases the correlation between additive genetic effects of siblings and DZ twins above it's expectation of .50. Such an increase cannot occur for MZ twins because the correlation is already at its max. Thus, AM increases the CV(DZ) relative to the CV(MZ), thereby mimicking VC.

[Page 13/25]. .

E5. The estimate of CV(Spouse) in dataset 1 is about 0. What does this suggest about the estimates of VC that we obtained in the ACE CTD in Section A?



E5.2. ANSWER: That CV(Spouse) is ~ 0 in this data suggests that our estimate of VC is not going to be biased upwards by this factor. This makes it more likely that the estimated VC from the CTD is downwardly biased.

[Page 14/25]. .

R Script Section F: CTD ACE model on dataset 2

F1.

Don't run R script Section F quite yet.

Let's pull in some new data, simulated using different parameters. For sake of fun, let's say the second trait we're analyzing is liberalism/conservatism.

As you will see in a bit, this trait has variance of 1.12. In our sample, there are 2726 MZ twin pairs whose covariance is .661, and there are 3403 DZ twin pairs whose covariance is .421.

Given this, what type of model would you choose to run?

- ACE model
- ADE model
- ADCE model
- AE model

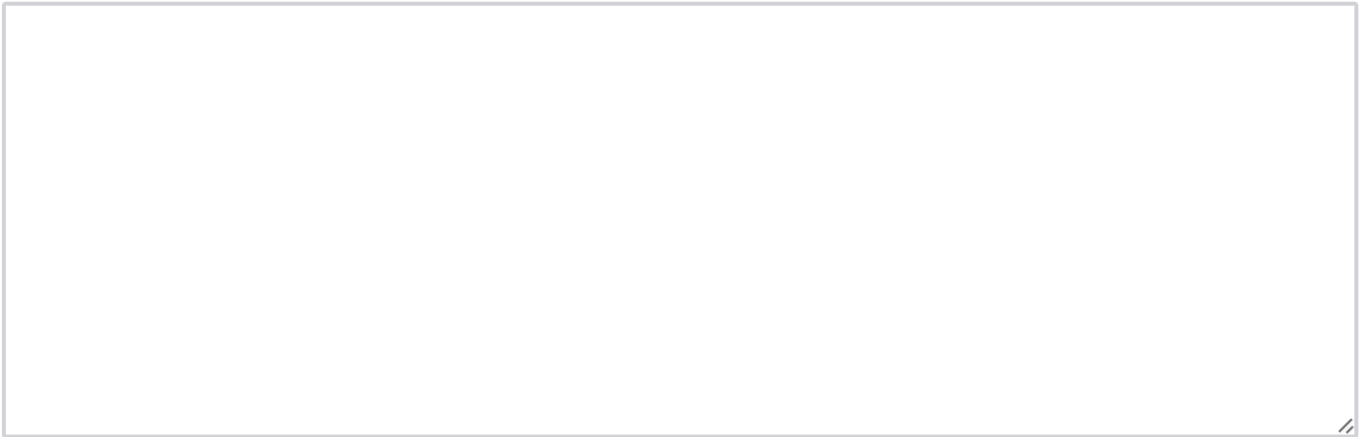
F2. What is your numerical estimate of VA using the MoM (algebraic) approach?

F3. What is your numerical estimate of VC using the MoM (algebraic) approach?

F4. OK, now go ahead and run the script to the end of section F. After you have done that,

answer the following question.

Given that this is a CTD, how confident are you that the estimate of VC we obtained is an underestimate (due to its being partially cancelled out due to any potential VD in the population)? Would it change your opinion if you knew this trait had a pretty high correlation between spouses?



F4.2. ANSWER: There is a headwind against VC in all ACE CTD models due to VD (or non-additive genetic variance) working against it. However, that is but one factor that can bias VC. AM, for example, can inflate it. Thus, you can have multiple different biases, some going in the same directions and some going in opposite directions, simultaneously influencing a given estimate.

Here, knowing that there is VC would decrease our confidence that VC is underestimated: any VD may deflate its estimate but AM may inflate it. Where the actual estimate of VC comes out in the end depends on the strength of these (and potentially other) factors.

[Page 15/25]. .

R Script Section G: NTFD on dataset 2

G1 . Go ahead and run script section G. Then answer the questions below.

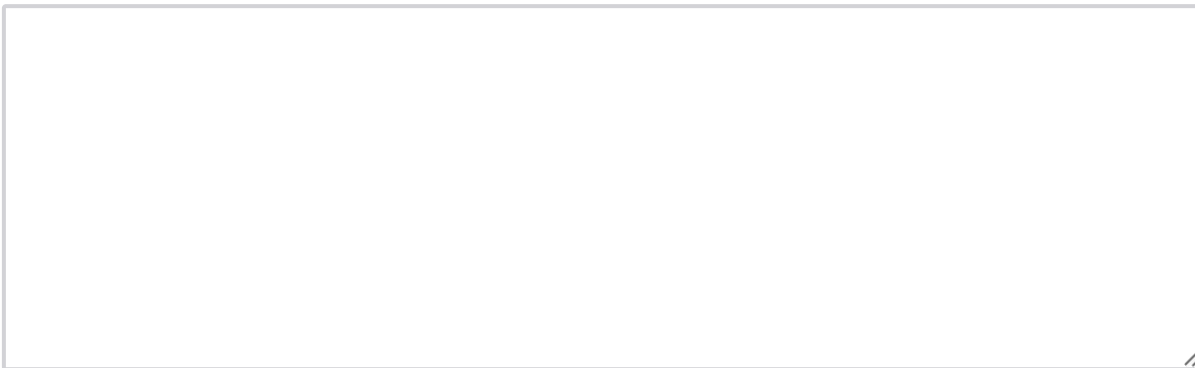
G2. How do your estimates from the NTFD compare to those from the CTD in Section F?
Which estimates do you trust more?



G2.2. ANSWER: I would certainly trust the estimates from the NTFD more. That's true in general, but especially the case here because we know there is a violation of the CTD (there is strong AM).

[Page 16/25]. .

G3. Why was your VC estimate in the CTD higher than the VS estimate in the NTFD, even though they should be estimating the same quantity?



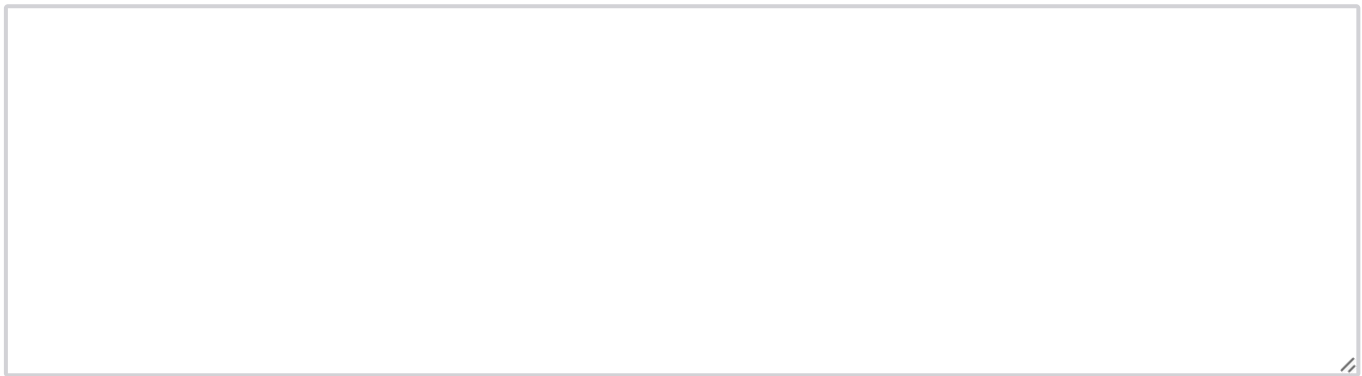
G3.3. ANSWER: The estimate of VC in CTD is higher than the one here because the CTD does not account for AM, which is one of the factors that biases VC upwards.

[Page 17/25]. .

R Script Section H: Overall Results Comparison

H1. Go ahead and run Section H in the script. This will create matrices that compare all your results in one place.

Why were the estimates from the ACE.D12.CTD model about as unbiased as those from the ADSE.NTFD model?



H1.1. ANSWER: We just so happened to choose an assumed value of VD (.12) that was ~ the same as the true value of VD. Whenever we get our assumptions right in the CTD (or in any good model), the estimates from that model should be unbiased.

[Page 18/25]. .

H2. On average, were NTFD estimates more or less biased than CTD estimates? Why?

H3. Which of the following are advantages of the NTFD?

- we can estimate VA, VD, and VS (aka VC) simultaneously, thereby reducing bias in other estimates
- we can estimate and account for the influences of AM on estimates
- we can estimate and account for the influences of passive G-E covariance
- we can account for qualitative G-by-age interactions
- we can estimate VS and VF simultaneously

H4. Reveal answer to the above question

- Reveal

H5. ANSWER: The first 3 answers were correct

[Page 19/25]. .

H4. What are some of the potential problems of NTFD models?

- estimates can be biased if genetic non-additivity is due to epistasis rather than dominance
- qualitative gene-by-age interactions can reduce parent-offspring similarity, reducing

evidence for VF and increasing that for VD

- the biases of the estimates depend on getting the model of AM correct
- the biases of the estimates depend on getting the model of vertical transmission correct
- the data is difficult to collect, and not much of it exists out there

H5. Reveal answer to the above question

- Reveal

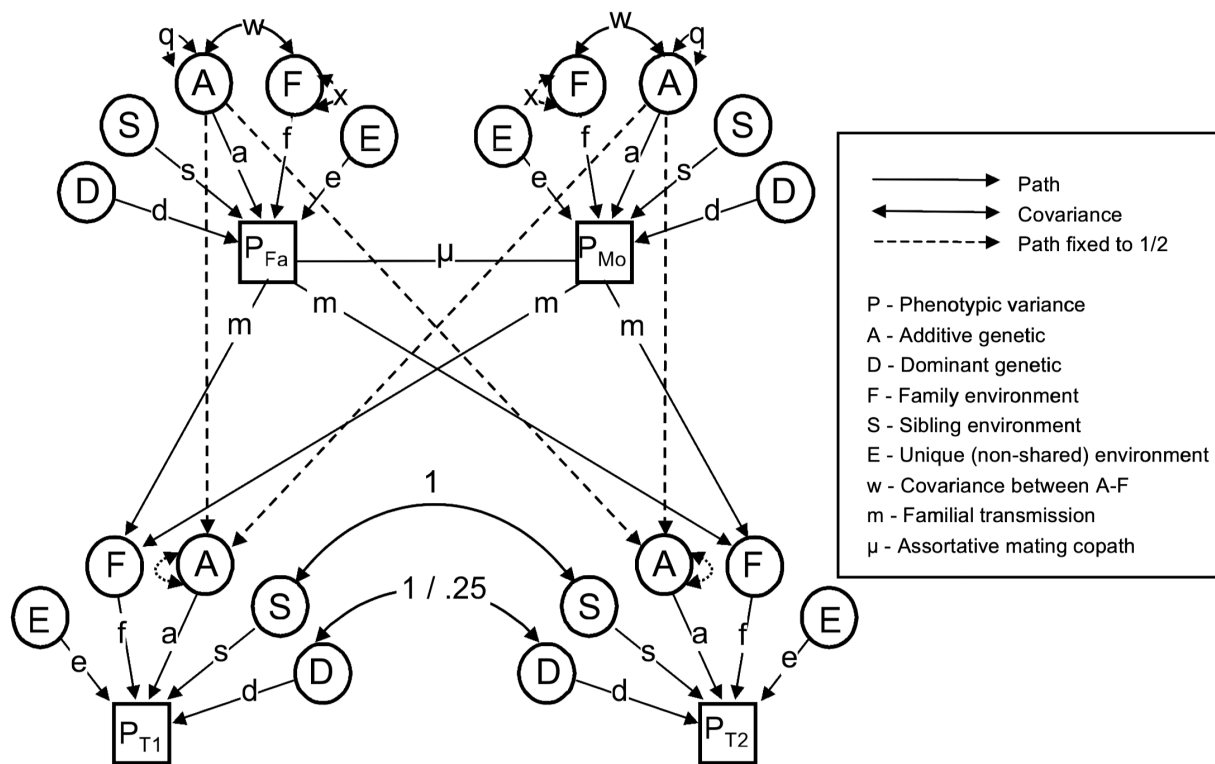
H6. ANSWER: All of the answers were correct

[Page 20/25]. .

R Script Section I: Writing an NTFD ADFE model

11. THIS IS A BONUS SECTION - here if you have sufficient time and you're up for a challenge. If you're running out of time, go ahead and skip forward to the next two sections, which will be very short.

Below is a path diagram of the full NTFD model again. It includes VA, VD, VS, VF, 3 of the 4 of which are estimable in this model. In our second NTFD model, we will fit an ADFE model, assuming $VS = 0$, and estimate VA, VD, and VF. Vertical transmission is parameterized using the "m" path. You cannot estimate both f and m, so we fix $f=1$ (we could alternatively have fixed $m=1$ and estimated f; it makes no difference). The variance of F is parameterized using x, and covariance between A and F is w. Neither of these are independent estimates; they are both non-linear constraints defined from all the other estimates.



Here, your job is to derive the expectations of x and w using path tracing rules as described in the videos. You then need to fill in these constraints in the script in Section I (where the ???'s are). After you've done that, run the model and see if you're getting reasonable answers. Once (or if) you think you have it running well, go to the next question and put in your estimates of the model.

12. Please provide the following estimates from your NTFD ADFE model below

VA estimate

VD estimate

VF estimate (aka, x)

COV(A,F) estimate (aka, w)

13. ANSWER: If your NTFD is working well, the estimates you should have provided in the last question are:

VA est = .448

VD est = .121

VF est (x) = .219

COV(A,F) est (w) = .319

A HUGE congratulations if you got those right!!!

[Page 21/25]. .

RCR

RCR1.

RCR Discussion

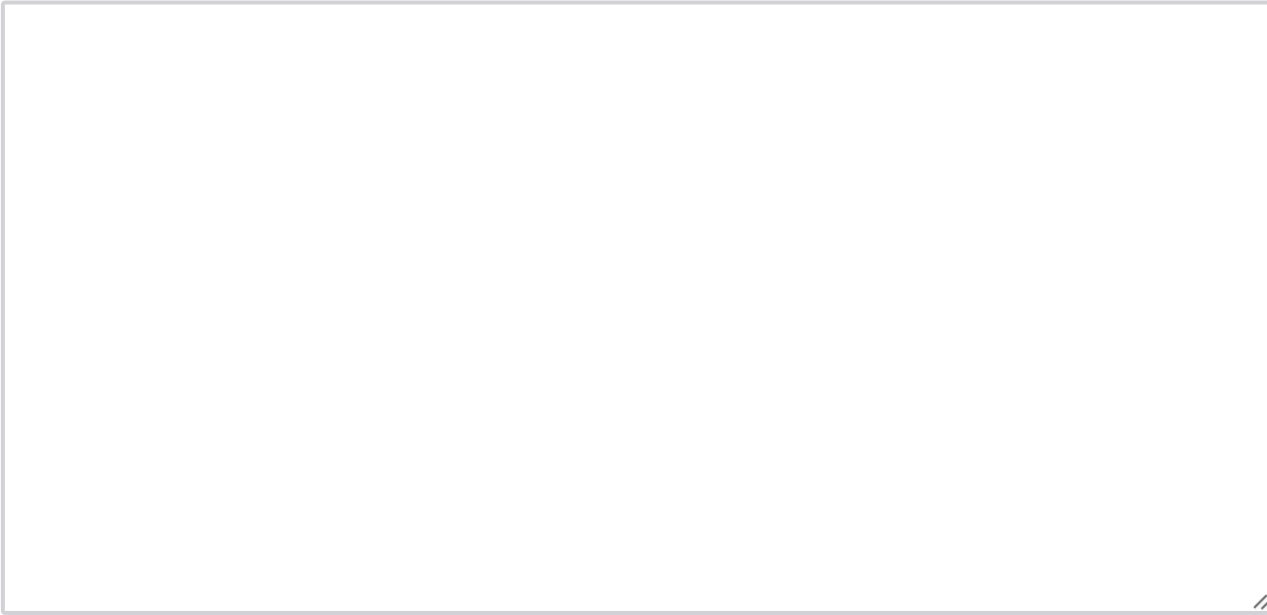
You've been working on a project for a long time, looking at the relationship between a genetic polymorphism known in the literature to be associated with Alzheimer's disease. While you didn't exactly replicate that association, you have found an interesting interaction between the polymorphism and an individual's education predicting Alzheimer's disease. You think this might make sense, because the effect seems to increase with educational attainment, and the studies showing this link have been oversampled for high education. However, you're concerned that you have run a lot of analyses by this point. Your advisor is on your case about publishing this paper, and you feel like you need a publication to result from this project or it could negatively impact your career.

What do you think you should do in this situation? Should you go ahead and submit the paper now? If so, how do you discuss the results in your paper? If not, what do you think the next

steps should be (recognizing that collecting more data will set you back even further)?

- Discuss these questions within your group.

RCR2. Summarise the main points of your discussion in the box below.



[Page 22/25]. .

End

Feedback1. Was the content of this tutorial...

- way too hard
- a bit too hard
- about right
- a bit too easy
-

way too easy

Feedback2. This tutorial...

- would have taken a lot more time than the time allotted
- would have taken a bit more time than the time allotted
- took us about the allotted time to get through
- took a bit less time than the time allotted
- took a lot less time than the time allotted

Feedback3. Please provide any other feedback about this tutorial or the videos related to it that would help me improve it in the future.

[Page 23/25]. .

End1.

Thank you for working through this tutorial. Hopefully you've learned something about biometrical modeling.

You can find a completed NTFD script example that fits VF in /faculty/matt/2022/Answers.

There is also a paper in there that explains the logic underlying the NTFD.

You can make a copy of those files with the following command

```
cp -r /faculty/matt/2022/Answers .
```

If you are SURE that you are done, and have no more answers to provide in the survey, click the right arrow button below to complete the survey. (If you click to the end of the survey, your answers will be final and you can't go back to modify them).

On the other hand, if you would like to go back and revisit some questions, you can click back now - the answers you've provided will still be there when you click back/forward so long as you don't end the survey).

[Page 24/25]. .

Q127. OK, I wasn't 100% truthful. The arrow on the last page didn't end the survey. But the right arrow on this page DOES end it. I promise.

Are you really done? Really? No more love for this survey? I did tell you that if you click to the end, you can't go back to continue working on your answers, right?

Well, ok, if that's how it is, I get it. Don't feel guilty or anything. Just go ahead. Click away. See if I care. Just leave me alone, here in the netherworld of the cloud. It's dark in here all alone.

[Page 25/25]

Powered by Qualtrics