Assumptions of the Classical Twin Design and Biases when Violated

Matthew Keller

International Statistical Genetics Workshop 2022

Models in science

- Scientific models represent natural phenomena in a logical and simplified way, allowing for better understanding and/or prediction of the phenomena
- Models must make multiple simplifying assumptions.
- To the degree that these assumptions are unmet (do not reflect the true complexity in the real world), biases result

”All models are wrong, some are useful.” - George Box

The point of this lecture

- We must learn to interpret model estimates in context of their biases. That biases exist DOES NOT invalidate the utility of the model.
- Because the Classical Twin Design (CTD) is the most basic and common design in behavioral genetics, it is crucial that you understand biases in CTD estimates. This enables you to properly interpret CTD estimates.
- As a generalization, the use of the CTD leads to upwardly biased estimates of V_A and downwardly biased estimates of V_D & V_C.
  - CTD provides decent broad-sense h^2, but are poor at differentiating V_A from V_D (or V_{NA}), or at estimating V_C
  - The extent of bias depends on a quantity that is often unknown: *how* violated its assumptions are. Even if unknown, we can guess at this or use alternative designs with different assumptions to triangulate.
- Ala Box, the CTD is undoubtedly useful, but that doesn't imply that its estimates should be taken too literally

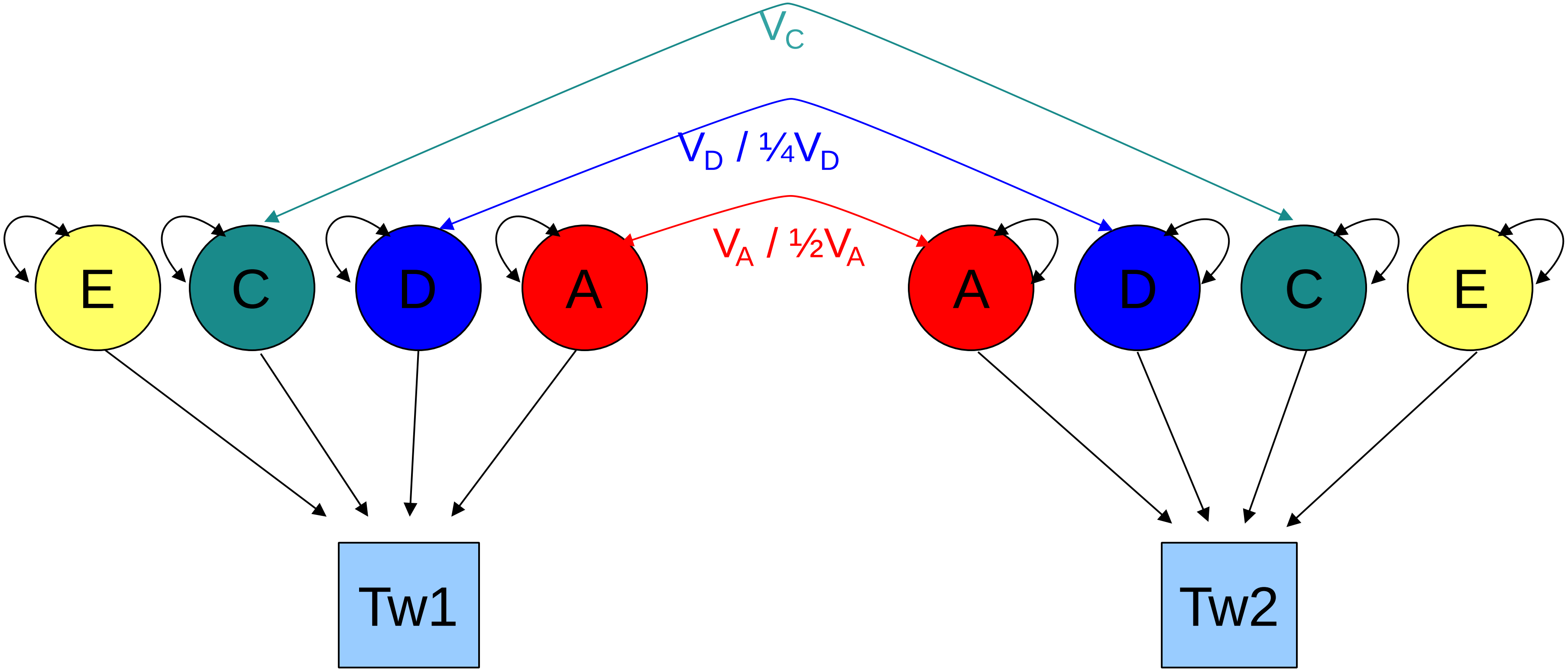
True vs. Estimated parameters

- V_A, V_C, V_D, V_E: population parameters. The **true values** (typically unknowable) in the population
- $\hat{V}_A$, $\hat{V}_C$, $\hat{V}_D$, $\hat{V}_E$: **estimated values** of V_A, V_C, V_D, and V_E
- $\hat{\theta}$ differs from θ due to:
  1) sampling variability
  2) bias (= E[$\hat{\theta}$] - θ)
- In this session, we will discuss assumptions of the CTD, derive some of the biases in CTD estimates when assumptions are violated, and learn how to interpret them in light of these biases.

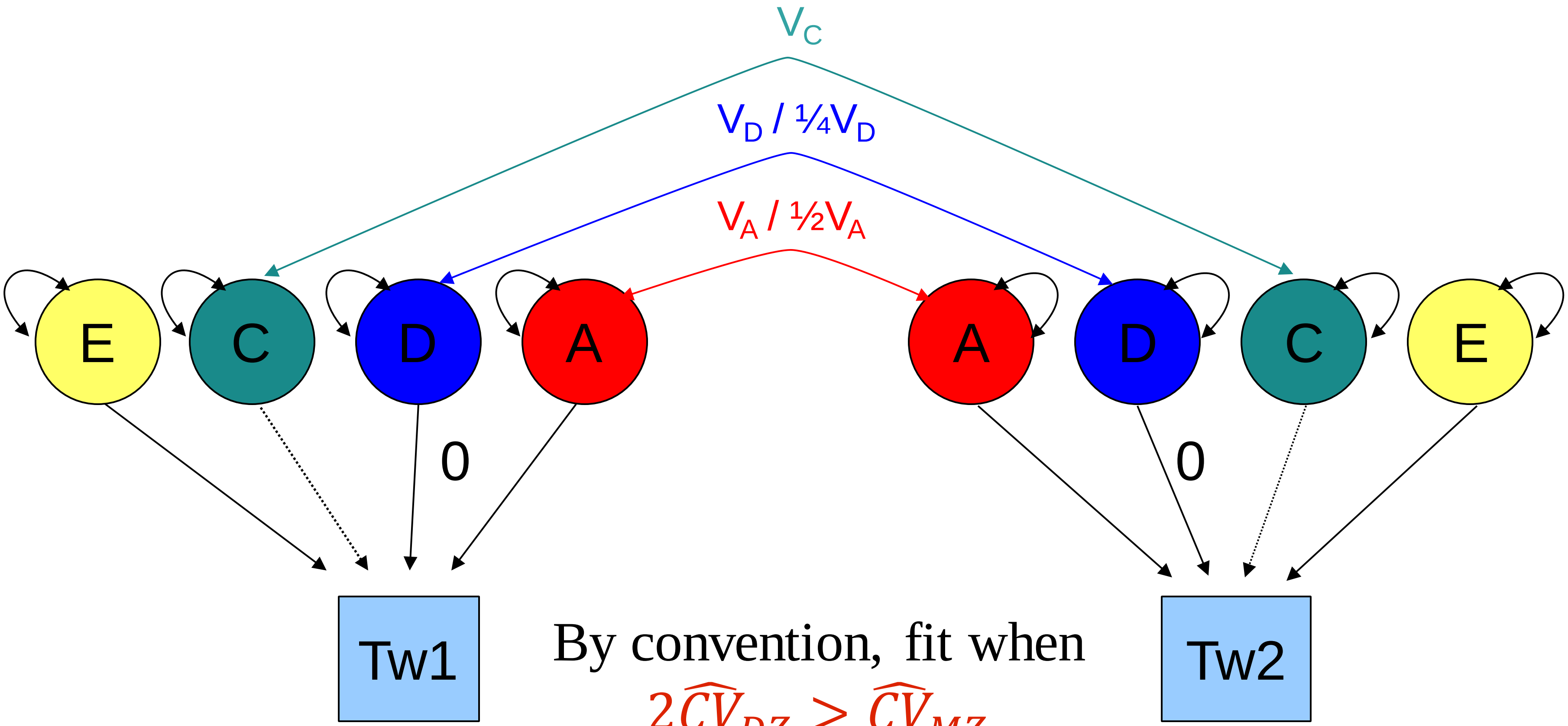
ACE and ADE models in the CTD

- ”ACE” CTD models estimate V_A, V_C, and V_E.
  - It is important to recognize that this implicitly assumes V_D=0.
- “ADE” CTD models estimate V_A, V_D, and V_E.
  - This implicitly assumes V_C=0.
- There are several other assumptions (i.e., simplifications) all CTDs make. We’ll get to these later and focus now just on these two.

The Classical Twin Design



ACE Model



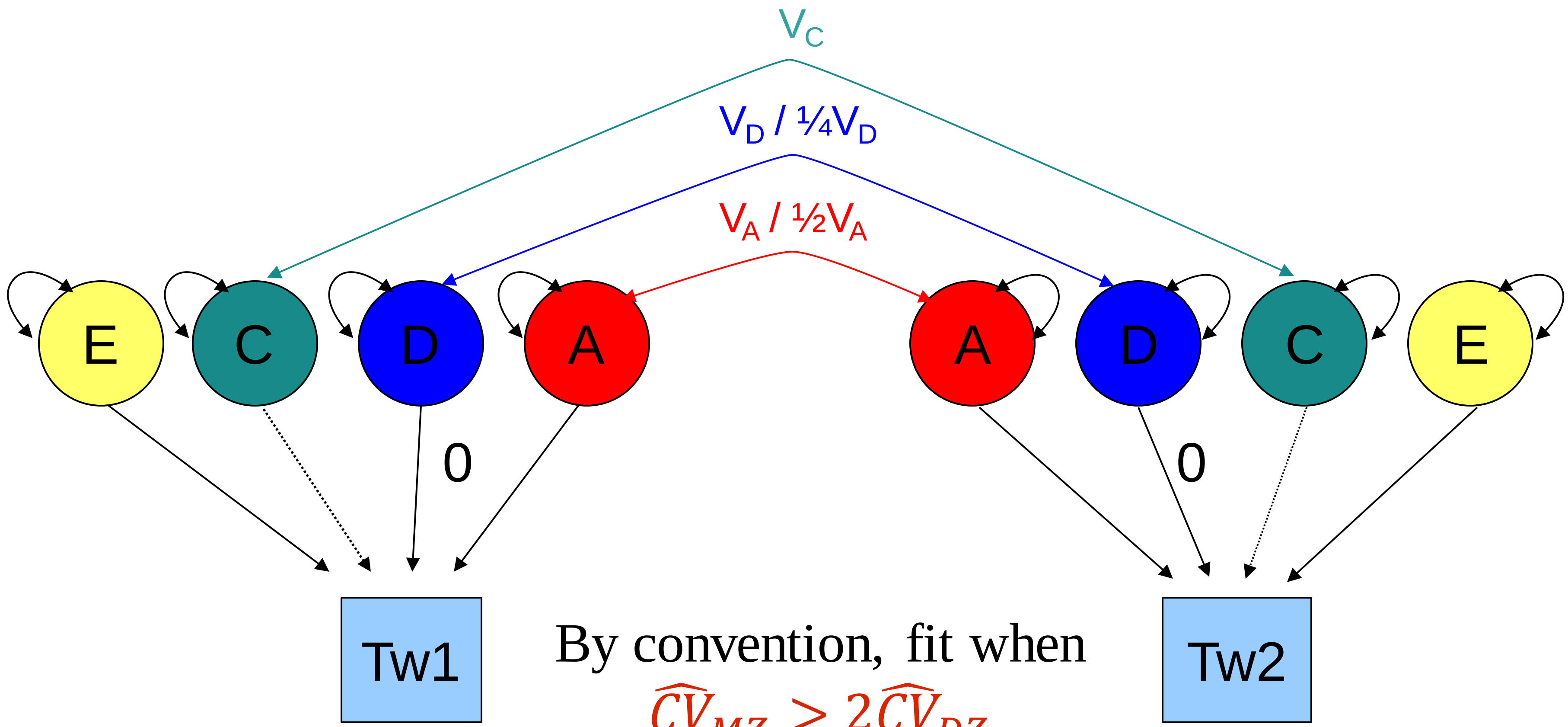
By convention, fit when

$$2\widehat{CV}_{DZ} > \widehat{CV}_{MZ}$$

$\hat{V}_P$	
$\widehat{CV}_{MZ}$	$\hat{V}_P$

$\hat{V}_P$	
$\widehat{CV}_{DZ}$	$\hat{V}_P$

ADE Model



By convention, fit when $\widehat{CV}_{MZ} > 2\widehat{CV}_{DZ}$

$$\begin{array}{|c|c|}\hline \hat{V}_P & \\ \hline \widehat{CV}_{MZ} & \hat{V}_P \\ \hline \end{array} \qquad \begin{array}{|c|c|}\hline \hat{V}_P & \\ \hline \widehat{CV}_{DZ} & \hat{V}_P \\ \hline \end{array}$$

Deriving algebraic expectations of variance component estimates

1) In an ACE model, we assume V_D=0. To get algebraic expectations of $\hat{V}_A$ and $\hat{V}_C$ in an ACE model, write down what CV_{MZ} and CV_{DZ} are assumed to be composed of:

$$CV_{MZ} = V_A + V_C$$

$$CV_{DZ} = \tfrac{1}{2}V_A + V_C$$

2) To get an estimate of one term (e.g., V_A), find a contrast of linear transformations of these two equations that cancel out one parameter (e.g., V_C) and isolate the other (e.g., V_A). E.g.:

$$CV_{MZ} - CV_{DZ} = \tfrac{1}{2}V_A\text{. Thus } 2(CV_{MZ} - CV_{DZ}) = V_A.$$

Thus, an estimator of V_A:

$$\hat{V}_A = 2(\widehat{CV}_{MZ} - \widehat{CV}_{DZ})$$

3) Similarly, to cancel out V_A and isolate V_C:

$$\hat{V}_C = 2\widehat{CV}_{DZ} - \widehat{CV}_{MZ}$$

Pen & Paper Practice 1: Algebraic expectations of ADE model

Use what we just learned to derive algebraic expectations of the estimates of V_A and V_D in an ADE model (where we assume V_C=0). As a hint, in this situation, we assume V_C=0 and therefore:

$$CV_{MZ} = V_A + V_D$$

$$CV_{DZ} = ½V_A + ¼V_D$$

To get $\hat{V}_A$, think of possible contrasts of linear transformations of these equations that cancel out V_D and isolate V_A.(and vice-versa for $\hat{V}_D$)

QUESTION1.1: What is your estimator of V_A ($\hat{V}_A$) in an ADE model?

QUESTION1.2: What is your estimator of V_D ($\hat{V}_D$) in an ADE model?

Pen & Paper Practice 1:
Algebraic expectations of ADE model

QUESTION1.1: What is your estimator of V_A ($\hat{V}_A$) in an ADE model?

QUESTION1.2: What is your estimator of V_D ($\hat{V}_D$) in an ADE model?

How to derive algebraic expectations of bias in estimates due to misspecification

1) We want to know what happens when we "misspecify" the model (here, when a parameter assumed to be 0 in the model is not 0). To do this, first write out one of your estimators. E.g., in an ACE model:

$$\hat{V}_A = 2(\widehat{CV}_{MZ} - \widehat{CV}_{DZ})$$

2) Consider the true compositions of parameters used in the estimators (i.e., if you got an assumption wrong). If V_D is actually non-zero, then:

$$CV_{MZ} = V_A + V_C + V_D$$

$$CV_{DZ} = ½V_A + V_C + ¼V_D$$

3) Finally, just substitute the true compositions of CV_{MZ} and CV_{DZ} into $\widehat{CV}_{MZ}$ and $\widehat{CV}_{DZ}$ used in the estimator. Thus, for $\hat{V}_A$ in an ACE:

$$\hat{V}_A = 2*(V_A + V_D + V_C - ½V_A - ¼V_D - V_C) = V_A + {}^{3}/_{2}V_D$$

In word: when $V_D \neq 0$ but one fits an ACE model, **$\hat{V}_A$ is biased upwards by 1.5 of whatever V_D truly is.**

4) Similarly, $\hat{V}_C = V_C - ½V_D$: **$\hat{V}_C$ is biased down by ½ of what V_D is.**

Pen & Paper Practice 2: Deriving biases of ADE

1) Use what we just learned to derive the bias in $\hat{V}_A$ and $\hat{V}_D$ in an ADE model (where we assume V_C=0). Recall:

$$\hat{V}_A = 4\widehat{CV}_{DZ} - \widehat{CV}_{MZ}$$

$$\hat{V}_D = 2\widehat{CV}_{MZ} - 4\,\widehat{CV}_{DZ}$$

$$CV_{MZ} = V_A + V_D + V_C$$

$$CV_{DZ} = \tfrac{1}{2}V_A + \tfrac{1}{4}V_D + V_C$$

2) Now just substitute the true compositions of CV_{MZ} and CV_{DZ} into $\widehat{CV}_{MZ}$ and $\widehat{CV}_{DZ}$ used in the estimator to see how our estimates are biased.

QUESTION2.1: How is $\hat{V}_A$ is biased in an ADE model when V_C (contrary to our assumption) is actually non-zero?

QUESTION2.2: How is $\hat{V}_D$ biased in an ADE model when V_C (contrary to our assumption) is actually non-zero?

Pen & Paper Practice:
Deriving biases of ADE

QUESTION2.1: How is $\hat{V}_A$ biased in an ADE model when $V_C \neq 0$?

QUESTION2.2: How is $\hat{V}_D$ biased in an ADE model when $V_C \neq 0$?

Quiz Question 1

1) We must fix to zero (and not estimate) either $\hat{V}_C$ or $\hat{V}_D$ in an identified CTD model because: [choose all that are correct]

a) these estimates are too highly correlated (multicolinearity problems)

b) you **can** estimate $\hat{V}_C$ and $\hat{V}_D$ simultaneously - you just have to fix $\hat{V}_A$ to some specific value (e.g., to 0)

c) you **can** estimate $\hat{V}_C$ and $\hat{V}_D$ simultaneously - you just have to allow them to go negative (not use path coefficient approach)

d) there are fewer informative statistics regarding within-family similarity (2) than parameters to be estimated (3), thus the "ADCE" model is unidentified.

Why can’t we estimate $\hat{V}_C$ and $\hat{V}_D$ at same time using twins only?

- Solve the following two equations for $\hat{V}_A$, $\hat{V}_C$ and $\hat{V}_D$

$$CV_{MZ} = V_A + V_D + V_C$$
$$CV_{DZ} = ½V_A + ¼V_D + V_C$$

- 3 unknowns, 2 informative equations. It can't be done. There are no <u>unique</u> solutions. The model is “unidentified”.

- Here, it’s obvious, but sometimes non-identification is challenging to see. You can empirically detect non-identification by noting that (a) model estimates depend on starting values AND (b) all final models have identical likelihoods

- Alternatively, in OpenMx, use mxCheckIdentification(model)

Just because we cannot fit $\hat{V}_C$ and $\hat{V}_D$ simultaneously in CTD doesn't mean one or the other's true value is 0!

- However, when we *try* to fit an ADCE model with just twins, there are an infinite number of combinations of $\hat{V}_A$, $\hat{V}_C$ and $\hat{V}_D$ that fit the data equally well. This is called "parameter indeterminacy" and is a necessary consequence of model non-identification.

- Thus, we just have to fit either an ADE (assuming $V_C = 0$) or ACE model (assuming $V_D = 0$) and live with potentially biased estimates.

- But it's good to quantify this bias to help in interpreting those estimates.

Quiz Question 2

2) If the assumptions of the CTD model that either V_C or V_D is zero is violated (i.e., A, C, and D simultaneously influence phenotypic variation)... [choose all that apply]

a) the interpretation of the estimated parameters should be altered; e.g., $\hat{V}_A$ should be considered an amalgam of V_A and V_D (in ACE model) or of V_A and V_C (in ADE model)

b) there is no point in doing the analysis

c) the estimated parameter values will be biased

Quiz Question 3

3) An ADE model finds that $\hat{V}_A$ = .30 and $\hat{V}_D$ = .10. This implies that shared environmental factors do not influence the trait in question.

a) TRUE

b) FALSE

Quiz Question 4

4) We run an ADE model and find that $\hat{V}_A$ = .69 and that $\hat{V}_D$ = .05. If in truth, V_C = .10, what will the effect on the estimated parameters be? [choose all that apply]

a) $\hat{V}_A$ will be biased (too low)

b) $\hat{V}_A$ will be biased (too high)

c) $\hat{V}_D$ will be biased (too low)

d) $\hat{V}_D$ will be biased (too high)

e) there is no effect on the estimated parameters; however, by not estimating V_C (aka, fixing it to zero), we underestimated V_C

Bias in parameter estimates for violation of assumption that either V_D or V_C is 0

- In ACE Models (bias induced in setting $\hat{V}_D = 0$):

$$\hat{V}_A = V_A + {}^{3}/_{2}\, V_D$$

$$\hat{V}_C = V_C - {}^{1}/_{2}\, V_D$$

- In ADE Models (bias induced in setting $\hat{V}_C = 0$):

$$\hat{V}_A = V_A + 3V_C$$

$$\hat{V}_D = V_D - 2V_C$$

- Thus, V_A is typically over-estimated and V_C and V_D under-estimated.
- However, things are more complicated when one considers the possibility of epistasis, assortative mating, etc.

Effects of epistasis on these biases

- Epistasis (across loci interactions) can increase the degree of the biases because it can reduce the $\widehat{CV}_{DZ} : \widehat{CV}_{MZ}$ ratio even further than the expected 1:4 ratio under dominance.
- However, the degree of bias rests on how strong higher-level epistatic influences are. This is an active area of debate.
- Epistatic effects will generally come out in $\hat{V}_D$. Thus, interpret $\hat{V}_D$ broadly, as a rough estimate of V_{NA} (a weighted amalgam of all the epistatic effects: $A \times A$, $A \times D$, $D \times D$, $A \times A \times A$, etc.)
- My take: V_A is almost certainly greater than V_{NA}, and evidence for much V_D per se is scant. But some traits may show high enough V_{NA} to bias $\hat{V}_C$ and $\hat{V}_D$ ($\sim$ estimate of V_{NA}) down and $\hat{V}_A$ up considerably from twin studies.

Quiz Question 5

5) What are the *typical* assumptions of a classical twin model? [choose all that apply]

a) only genetic factors cause MZ twins to be more correlated than DZ twins

b) either V_D or V_C is zero

c) no epistasis

d) no assortative mating

e) no gene-environment interactions or correlations

What are the typical effects of violations of assumptions in the CTD?

a) Only genetic factors cause MZ twins to be more correlated than DZ twins:

$\hat{V}_A$ & $\hat{V}_D$ overestimated and $\hat{V}_C$ underestimated

b) Either V_D or V_C is zero:

$\hat{V}_A$ overestimated and $\hat{V}_D$ & $\hat{V}_C$ underestimated

c) No epistasis:

$\hat{V}_D$ or $\hat{V}_A$ overestimated and $\hat{V}_C$ underestimated

d) No assortative mating:

$\hat{V}_A$ and/or $\hat{V}_D$ underestimated and $\hat{V}_C$ overestimated

e) No gene-environment interactions or correlations:

AxC: $\hat{V}_A$ overestimated

AxE: $\hat{V}_E$ overestimated

passive Cov(A,C): $\hat{V}_C$ overestimated

Conclusions

- All models require assumptions. Generally, the more these assumptions are violated, the more estimates are biased
- Understanding biases allows you to understand how to interpret estimates with the proper nuance
- In all models, including the CTD, be cautious of reifying parameter estimates!
  - $\hat{V}_A$ is amalgam of mostly V_A but also V_D & V_C.
  - $\hat{V}_C$ & $\hat{V}_D$ may often be underestimates
  - Interpret $\hat{V}_D$ as a (potentially downwardly biased) estimate of V_{NA}
  - $\hat{V}_A/\hat{V}_P$ (in ACE) or $(\hat{V}_A+\hat{V}_D)/\hat{V}_P$ (in ADE) are decent estimates of <u>broad sense *h*²</u>.

Readings related to this lecture

- Eaves LJ, Last KA, Young PA, Martin NG (1978) Model-fitting approaches to the analysis of human behaviour. *Heredity* 41:249-320
- Fulker DW (1982) Extensions of the classical twin method. Human Genetics. Part A: The Unfolding Genome (Progress in Clinical and Biological Research Vol 103A). p. 395-406
- Keller MC & Coventry WL (2005). Quantifying and addressing parameter indeterminacy in the classical twin design. *Twin Research and Human Genetics*, 8, 201-213
- Keller MC, Medland SE, & Duncan LE (2010). Are extended twin family designs worth the trouble? A comparison of the bias, precision, and accuracy of parameters estimated in four twin family models. *Behavior Genetics*.