Project Progress Pronouncement

Joshua N. Pritikin

Virginia Institute for Psychiatric and Behavioral Genetics
Virginia Commonwealth University

Feb 2021

Progress Plunge

Variance decomposition of ordinal indicators from ABCD

Genome-wide structural equation modeling

Progress Plunge

Variance decomposition of ordinal indicators from ABCD

Genome-wide structural equation modeling

Variance decomposition of ordinal indicators

Mentors and collaborators

- Mike Neale
- Hermine Maes
- Daniel Bustamante

Adolescent Brain Cognitive Development (ABCD)

A unique data resource:

- 21 research sites
- About 12k children recruited at ages 9-10
- Assessments of neurocognition, physical and mental health, social and emotional functions, and culture and environment
- Multimodal structural and functional brain imaging and bioassays

Under way since 2018

Data snapshot from 26 Mar 2019 (Wave 2)

Adolescent Brain Cognitive Development (ABCD)

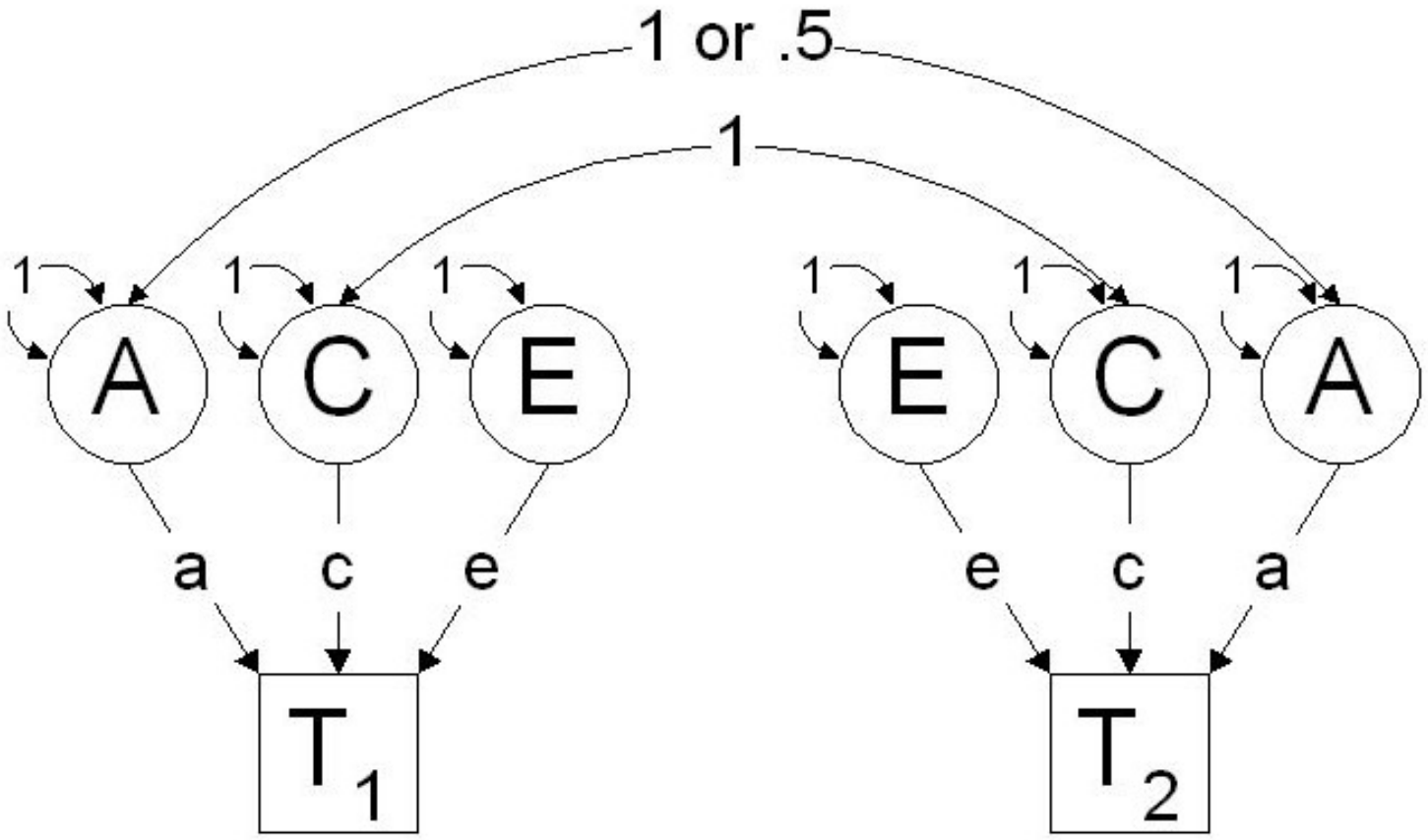
A unique data resource:

- 21 research sites
- About 12k children recruited at ages 9-10
- Assessments of neurocognition, physical and mental health, social and emotional functions, and culture and environment
- Multimodal structural and functional brain imaging and bioassays

Under way since 2018

Data snapshot from 26 Mar 2019 (Wave 2)

ACE model



ABCD Covariates

Adjust each person for

- Age
- Sex
- Race (white, black, hispanic, asian, other)
- Income (< 50k, >= 50k, >= 100k)
- Parents' education (< HS, HS, some, bachelor, post)
- Parents currently married (yes, no)

Site – How much variance due to site?

ABCD Covariates

Adjust each person for

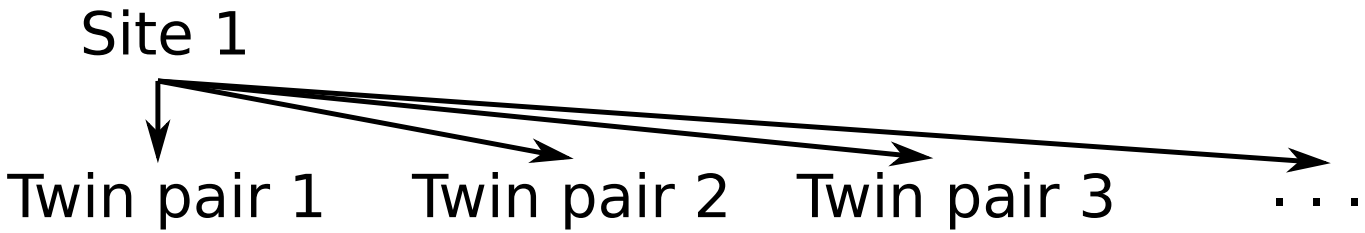
- Age
- Sex
- Race (white, black, hispanic, asian, other)
- Income ($<$ 50k, $>=$ 50k, $>=$ 100k)
- Parents' education ($<$ HS, HS, some, bachelor, post)
- Parents currently married (yes, no)

Site – How much variance due to site?

A multilevel conceptualization

Each site could have a different mean T

When T is continuous



OpenMx's uses the Rampart optimization[1]

[1] Pritikin, Hunter, von Oertzen, Brick, and Boker (2017)

When T is ordinal

Cannot use Rampart optimization

Maximum likelihood

- Slow (quadrature integration over site variance)
- Custom software development
- Point estimates and standard errors

Full Bayes

- Slow (MCMC sampler)
- Custom software development
- Full posterior

When T is ordinal

Cannot use Rampart optimization

Maximum likelihood

- Slow (quadrature integration over site variance)
- Custom software development
- Point estimates and standard errors

Full Bayes

- Slow (MCMC sampler)
- Custom software development
- Full posterior

When T is ordinal

Cannot use Rampart optimization

Maximum likelihood

- Slow (quadrature integration over site variance)
- Custom software development
- Point estimates and standard errors

Full Bayes

- Slow (MCMC sampler)
- Custom software development
- Full posterior

Stan

State-of-the-art Hamiltonian Monte Carlo sampler[2]

Model definition

- Probabilistic programming language
- C/C++-like syntax
- Automatic derivatives

Generally more efficient than BUGS/JAGS[3]

[2] `https://chi-feng.github.io/mcmc-demo/app.html`
[3] Plummer (2013)

Probit ordinal likelihood (1/4)

Let

- $H \geq 2$ be the number of response options
- $y_i \in \{1, \ldots, H\}$ be observed data for person i

Probability is assigned to less-than inequalities and a difference is used to obtain the probability of an observation,[4]

$$\pi(y_i = h) = \begin{cases} \pi(y_i \leq h) - 0 & \text{if } 1 = h \\ \pi(y_i \leq h) - \pi(y_i \leq h-1) & \text{if } 1 < h < H \\ 1 - \pi(y_i \leq h-1) & \text{if } h = H. \end{cases}$$

[4] Samejima (1969)

Probit ordinal likelihood (2/4)

Let

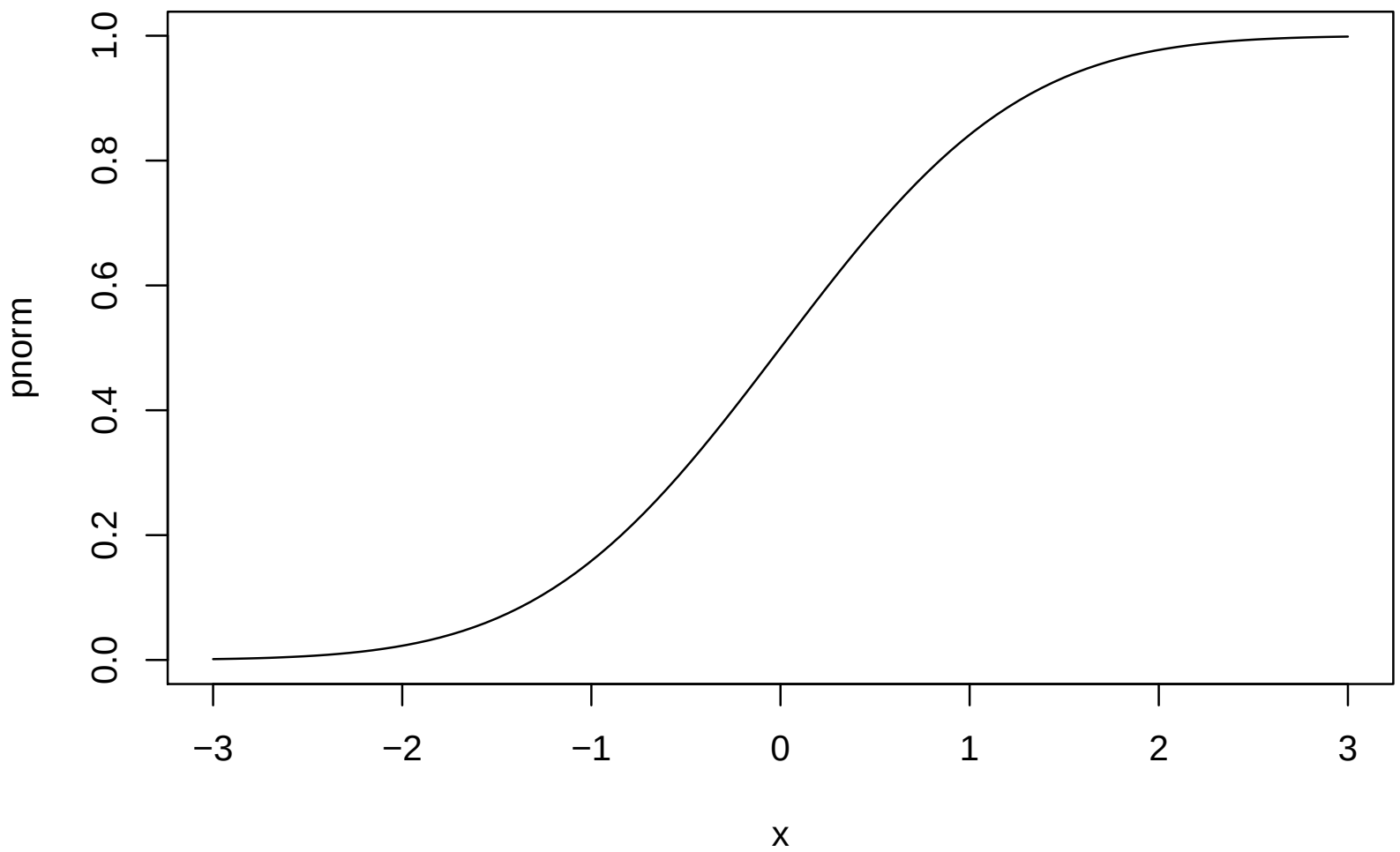
- Δ_h for $h \in \{1, \ldots, H-1\}$ be thresholds
- cumulative sum $\delta_h \equiv \sum_{q=1}^{h} \Delta_q$ for $h \in \{1, \ldots, H-1\}$
- θ_i be the latent continuous trait for person i

We define our response inequality as

$$\pi(y_i \leq h \mid \theta_i, \delta_h) = \Phi(\theta_i - \delta_h) \quad \text{for } h \in \{1, \ldots, H-1\}$$

where Φ is the cumulative standard normal distribution.

Probit ordinal likelihood (3/4)



Probit ordinal likelihood (4/4)

Single item, therefore thresholds Δ are fixed, not estimated

Set Δ_h to the standard normal quantile of the proportion of responses less than or equal to h,

$$\Delta_h = \Phi^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} 1_{y_i \leq h}\right)$$

Model log likelihood is $\sum_{i=1}^{N} \log \pi(y_i = h \mid \theta_i, \Delta)$

Probit ordinal likelihood (4/4)

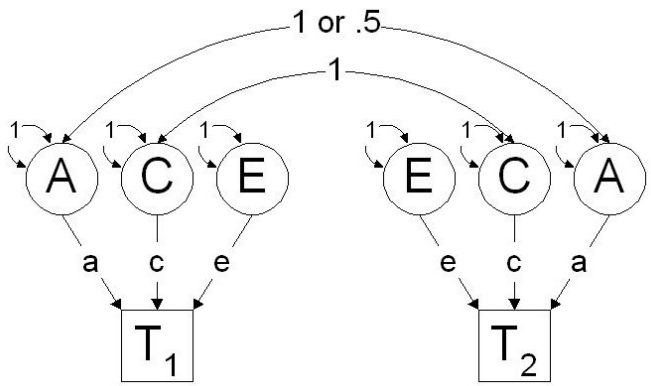
Single item, therefore thresholds Δ are fixed, not estimated

Set Δ_h to the standard normal quantile of the proportion of responses less than or equal to h,

$$\Delta_h = \Phi^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} 1_{y_i \leq h}\right)$$

Model log likelihood is $\sum_{i=1}^{N} \log \pi(y_i = h \mid \theta_i, \Delta)$

Doing variance decomposition with regression

Roughly 

$\longrightarrow$

$$\theta_i = r^{0.5} a_f + c_f + (1 + (1 - r)^{0.5}) e_i$$
$$a \sim \mathcal{N}(0, 1)$$
$$c \sim \mathcal{N}(0, 1)$$
$$e \sim \mathcal{N}(0, 1)$$

where r is the relatedness (1 or .5) and f indexes families.[5]

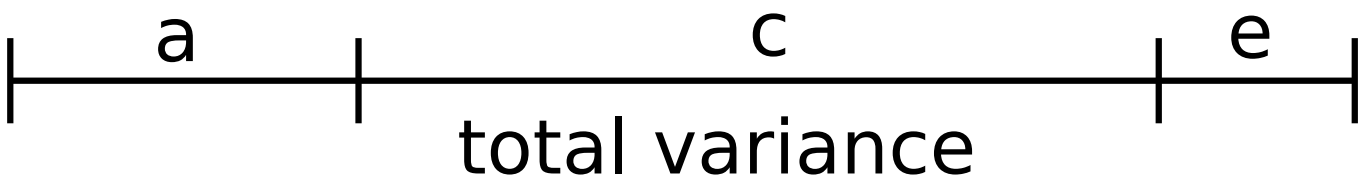
[5] Kuhnert and Do (2003)

Preserving scale

Reconcile

- Total variance is fixed at 1.0
- MCMC sampler can't deal with boundaries

$\longrightarrow$ Only consider proportions



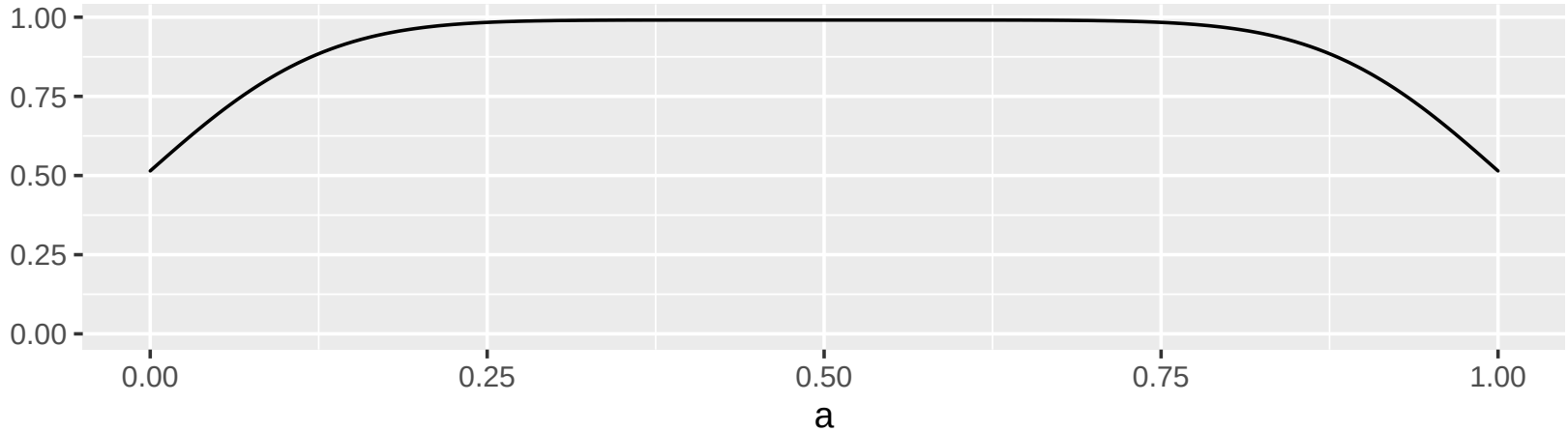
Bounded scalar

Stan offers a built-in log odds transformation

$$a \in (0, 1) \qquad \text{logit}(a) = \frac{a}{1 - a}$$

$$v \in (-\infty, \infty) \qquad \text{logit}^{-1}(v) = \frac{1}{1 + \exp(-v)}$$

Prior $\beta(1.2, 1.2)$



CE Model

Let i index persons, f index families

$$C \sim \beta(1.2, 1.2) \qquad c_f \sim \mathcal{N}(0,1) \quad \text{for } f \in \{1 \dots F\}$$
$$E \sim \beta(1.2, 1.2) \qquad r_i \sim \mathcal{N}(0,1)$$

Person i's family f known

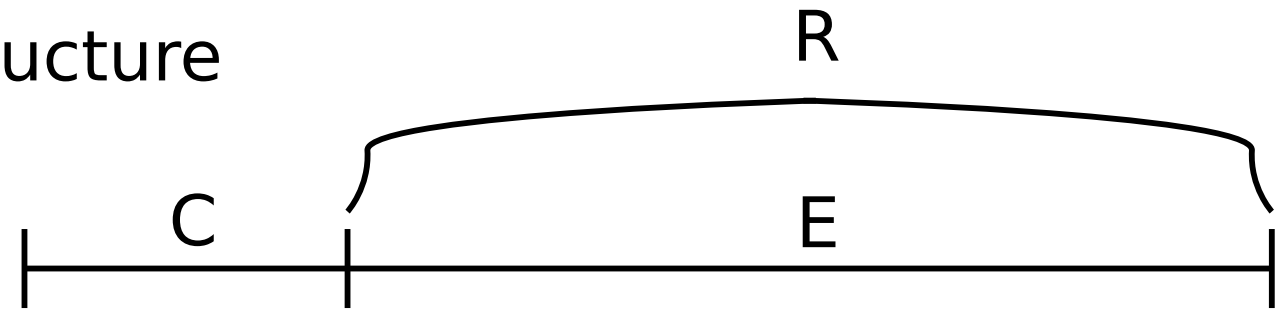
$$R = E$$
$$\theta_i = C^{0.5} c_f + R^{0.5} r_i$$

Person i's family f unknown

$$R = \frac{C}{F-1} \sum_{f=1}^{F} c_f^2 + E$$
$$\theta_i = R^{0.5} r_i$$

CE Model; Family known

variance structure

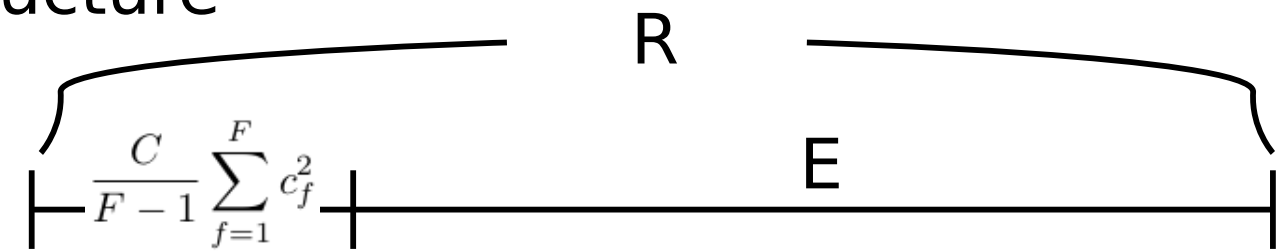


mean structure

$$C^{0.5}\, c_f \quad + \quad R^{0.5}\, r_i$$

CE Model; Family unknown

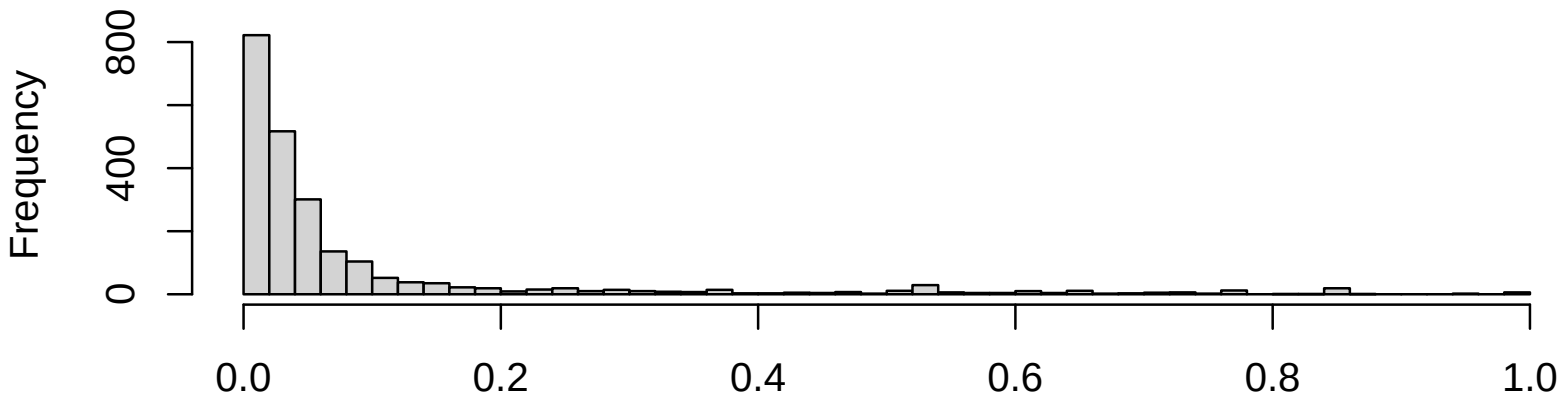
variance structure



mean structure

$R^{0.5}\, r_i$

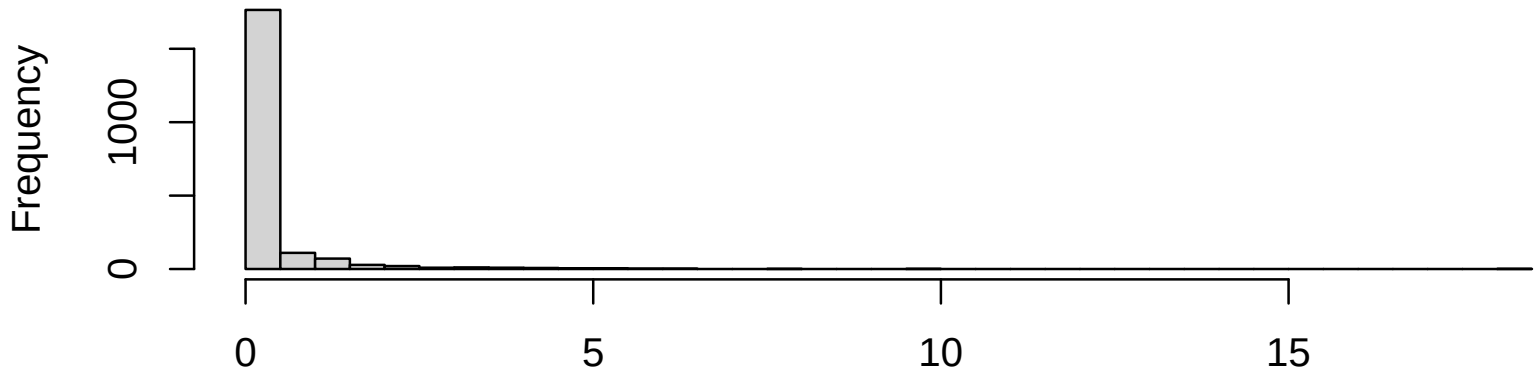
Initial exploration



Ordinal probit regression w/ covariates

- Nominally 5988 ordinal indicators
- 3673 excluded due to more than 50% missing or optimization failure
- Histogram of 2315 indicators by pseudo-R^2
- Roughly: proportion of variance accounted for by covariates

Variance



Indicators of interest

- Exclude 269 indicators that are 20% or more predicted by covariates
- 2046 indicators remain
- Histogram of total variance (treating ordinal as continuous)

Method Matters

Out of 2046 indicators:

- Bayesian sampling succeeded on 1565
- Maximum likelihood (ML) succeeded on 1026
- Both succeeded on 812

Many ML fits are hard to interpret due to

- optimization failure
- negative proportion estimates

Bayesian results generally look sane?

Method Matters

Out of 2046 indicators:

- Bayesian sampling succeeded on 1565
- Maximum likelihood (ML) succeeded on 1026
- Both succeeded on 812

Many ML fits are hard to interpret due to

- optimization failure
- negative proportion estimates

Bayesian results generally look sane?

ABCD Parent Diagnostic Interview for DSM-5 Background Items Full

`ksads_back_conflict_causes_p___2` "Click the things that cause conflict between you and your child"

Response options

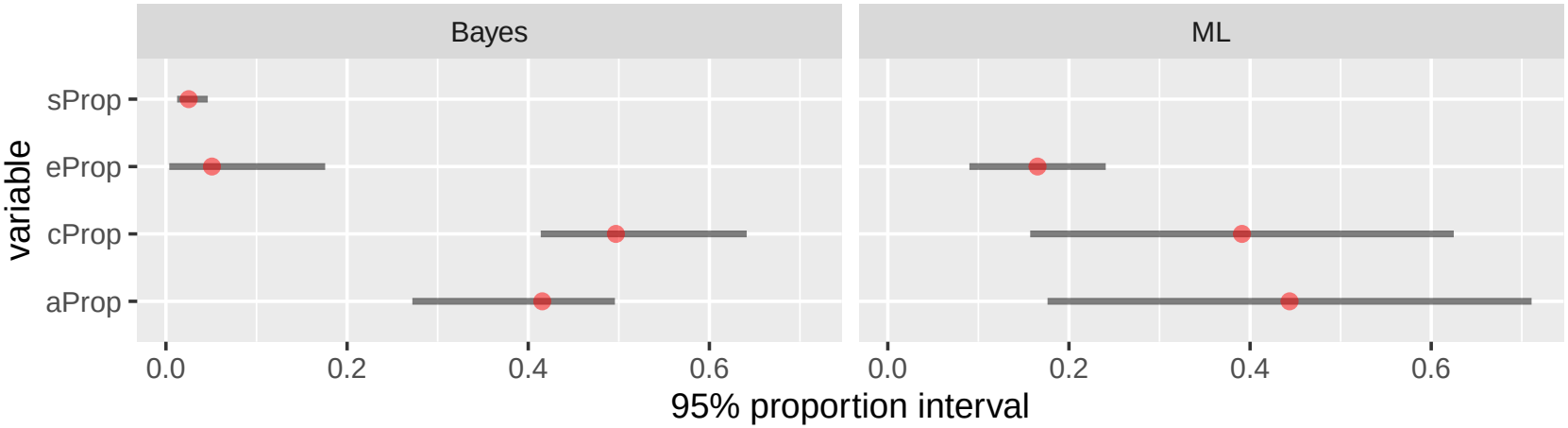
- Messy room
- Not endorsed

Variance of 0.1, in top 6% among KSADS items

Approx 1% of variance accounted for by covariates

ML polychorics: MZ $\begin{bmatrix} 1.00 & 0.83 \\ 0.83 & 1.00 \end{bmatrix}$ DZ $\begin{bmatrix} 1.00 & 0.61 \\ 0.61 & 1.00 \end{bmatrix}$

ksads_back_conflict_causes_p__2



Messy room is about 40% genetic! Why SEs so different?

ABCD Parent Diagnostic Interview for DSM-5

`ksads_14_425_p` "Symptoms interfere with social academic or occupational functioning Past"

Response options

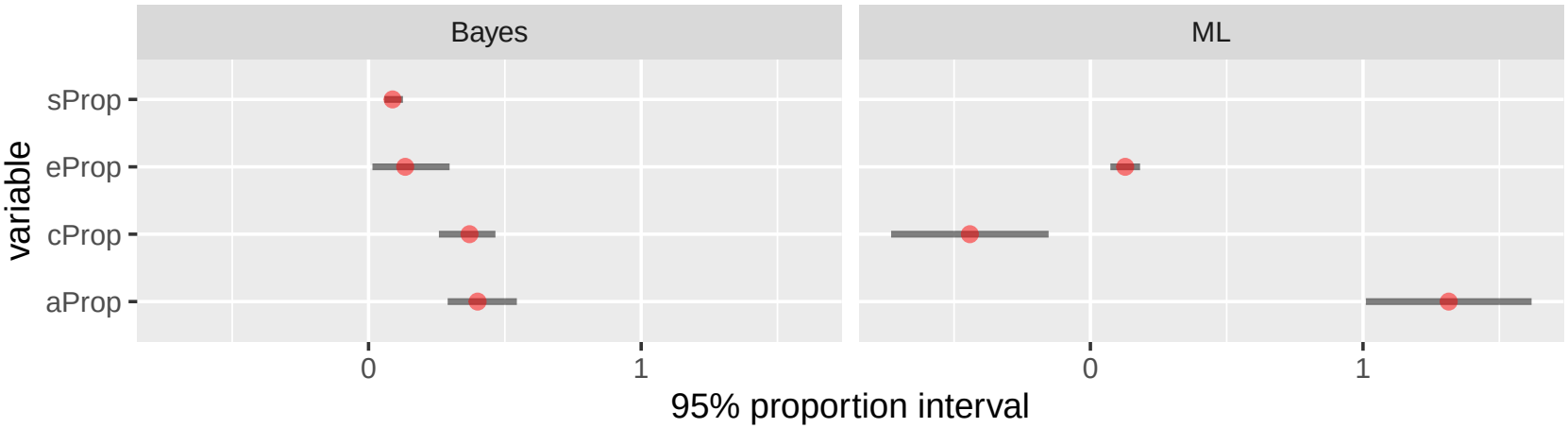
- Yes
- No

Variance of 0.23, in top 5% among KSADS items

Approx 3% of variance accounted for by covariates

ML polychorics: MZ $\begin{bmatrix} 1.00 & 0.87 \\ 0.87 & 1.00 \end{bmatrix}$ DZ $\begin{bmatrix} 1.00 & 0.22 \\ 0.22 & 1.00 \end{bmatrix}$

ksads_14_425_p



???

Next steps

Dissemination stage

- Re-run simulations, double check everything
- Refresh for wave 3 data (Nov 2020 snapshot)
- Write & Submit paper
- Resubmit paper
- Re-resubmit paper
- Re-re-resubmit paper
- Celebrate acceptance
- Write grant to support further work

Next steps

Dissemination stage

- Re-run simulations, double check everything
- Refresh for wave 3 data (Nov 2020 snapshot)
- Write & Submit paper
- Resubmit paper
- Re-resubmit paper
- Re-re-resubmit paper
- Celebrate acceptance
- Write grant to support further work

Next steps

Dissemination stage

- Re-run simulations, double check everything
- Refresh for wave 3 data (Nov 2020 snapshot)
- Write & Submit paper
- Resubmit paper
- Re-resubmit paper
- Re-re-resubmit paper
- Celebrate acceptance
- Write grant to support further work

Next steps

Dissemination stage

- Re-run simulations, double check everything
- Refresh for wave 3 data (Nov 2020 snapshot)
- Write & Submit paper
- Resubmit paper
- Re-resubmit paper
- Re-re-resubmit paper
- Celebrate acceptance
- Write grant to support further work

Next steps

Dissemination stage

- Re-run simulations, double check everything
- Refresh for wave 3 data (Nov 2020 snapshot)
- Write & Submit paper
- Resubmit paper
- Re-resubmit paper
- Re-re-resubmit paper
- Celebrate acceptance
- Write grant to support further work

Progress Plunge

Variance decomposition of ordinal indicators from ABCD

Genome-wide structural equation modeling

GW-SEM Update

History

- Initial prototype[6]
- Rewritten as 2.0, published on CRAN[7]

In preparation

- Gene-age interaction[8]
- Comparison to summary GWAS analyses (e.g., Genomic SEM[9])

[6] Verhulst, Maes, and Neale (2017)
[7] Pritikin, Verhulst, and Neale (in press)
[8] Verhulst, Pritikin, Clifford, and Prom-Wormley (submitted)
[9] Grotzinger et al. (2019)

Single-nucleotide polymorphism

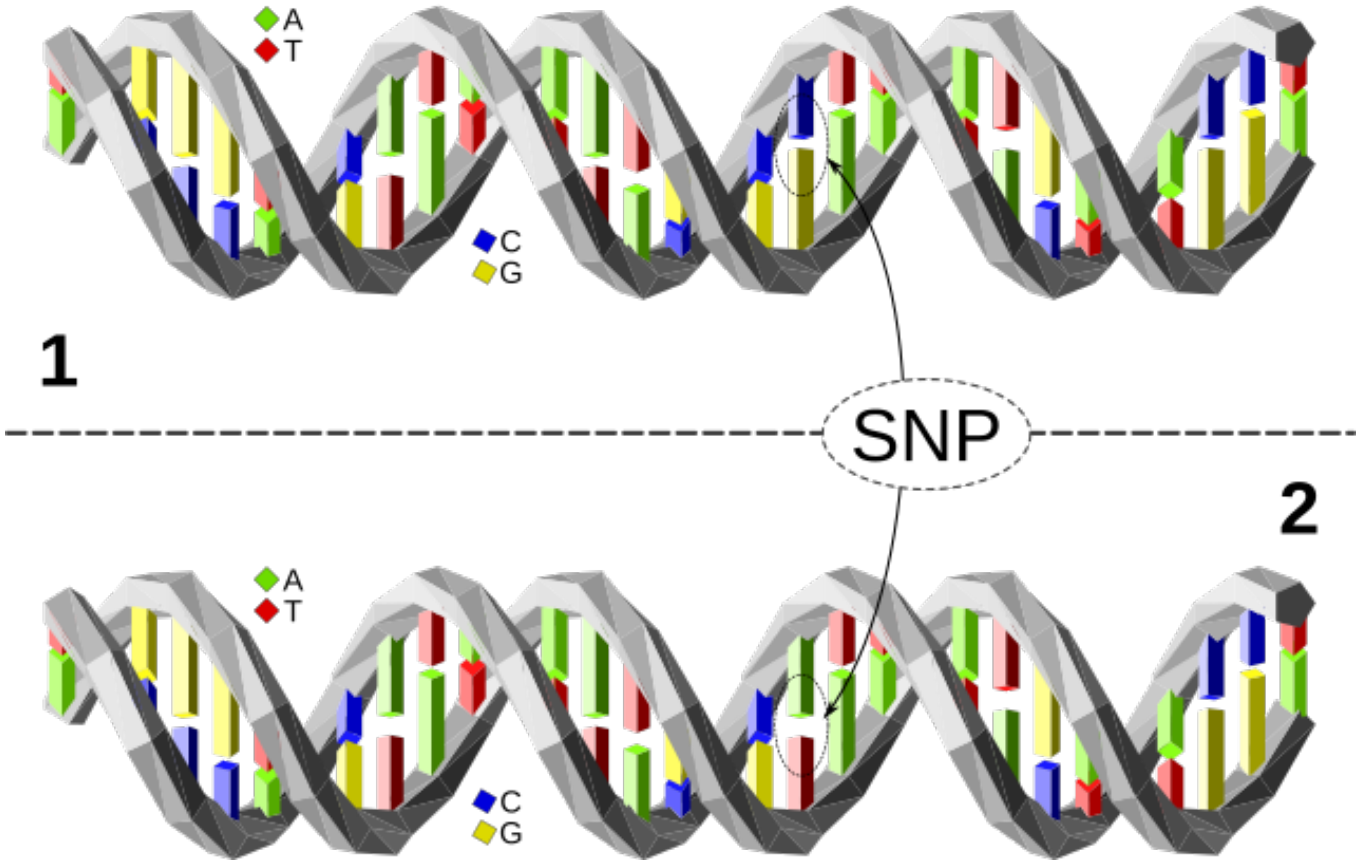
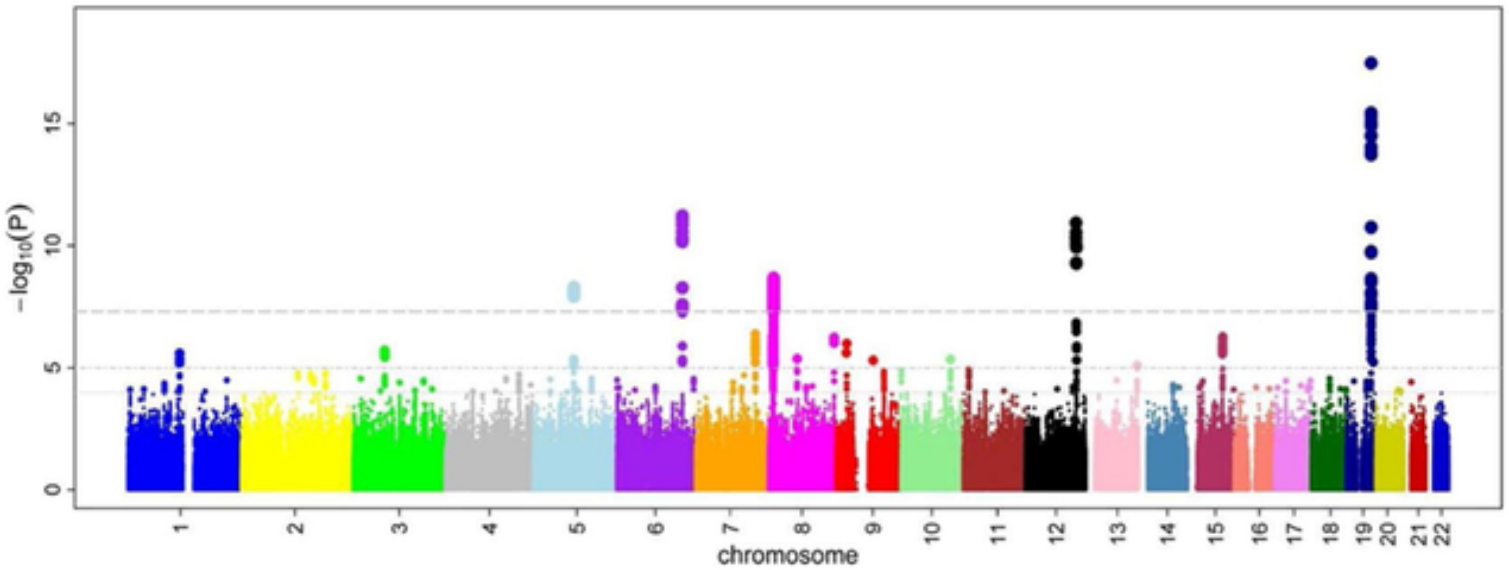


Image by David Eccles (gringer), CC BY 4.0,
`https://commons.wikimedia.org/w/index.php?curid=2355125`

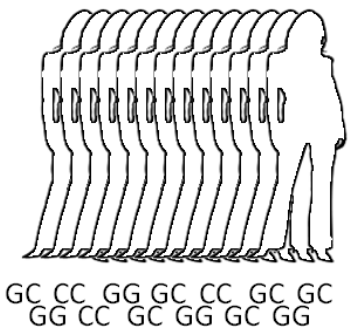
Genome-wide association studies



Ordinary regressions of SNP on microcirculation[10]

[10] By M. Kamran Ikram et al (2010) PLoS Genet. Oct 28;6(10):e1001184. CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=18056138

Case control design



SNP1	*SNP2*	*SNP…*
Cases Count of G: 2104 of 4000 Frequency of G: 52.6%	**Cases** Count of G: 1648 of 4000 Frequency of G: 41.2%	*Repeat for all SNPs*
Controls Count of G: 2676 of 6000 Frequency of G: 44.6%	**Controls** Count of G: 2532 of 6000 Frequency of G: 42.2%	
P-value: $5.0 \cdot 10^{-15}$	**P-value:** 0.33	

Probit or logit regression[11]

[11] Lasse Folkersen CC BY 3.0, https://commons.wikimedia.org/w/index.php?curid=18062562

Predefined models

Model construction

- `buildItem` – regression, but can do multiple items
- `buildOneFac` – single factor model
  similar to GEMMA & plink MANOVA
- `buildOneFacRes` – single factor residuals model
- `buildTwoFac` – two factor model (pleiotropy, comorbidity)

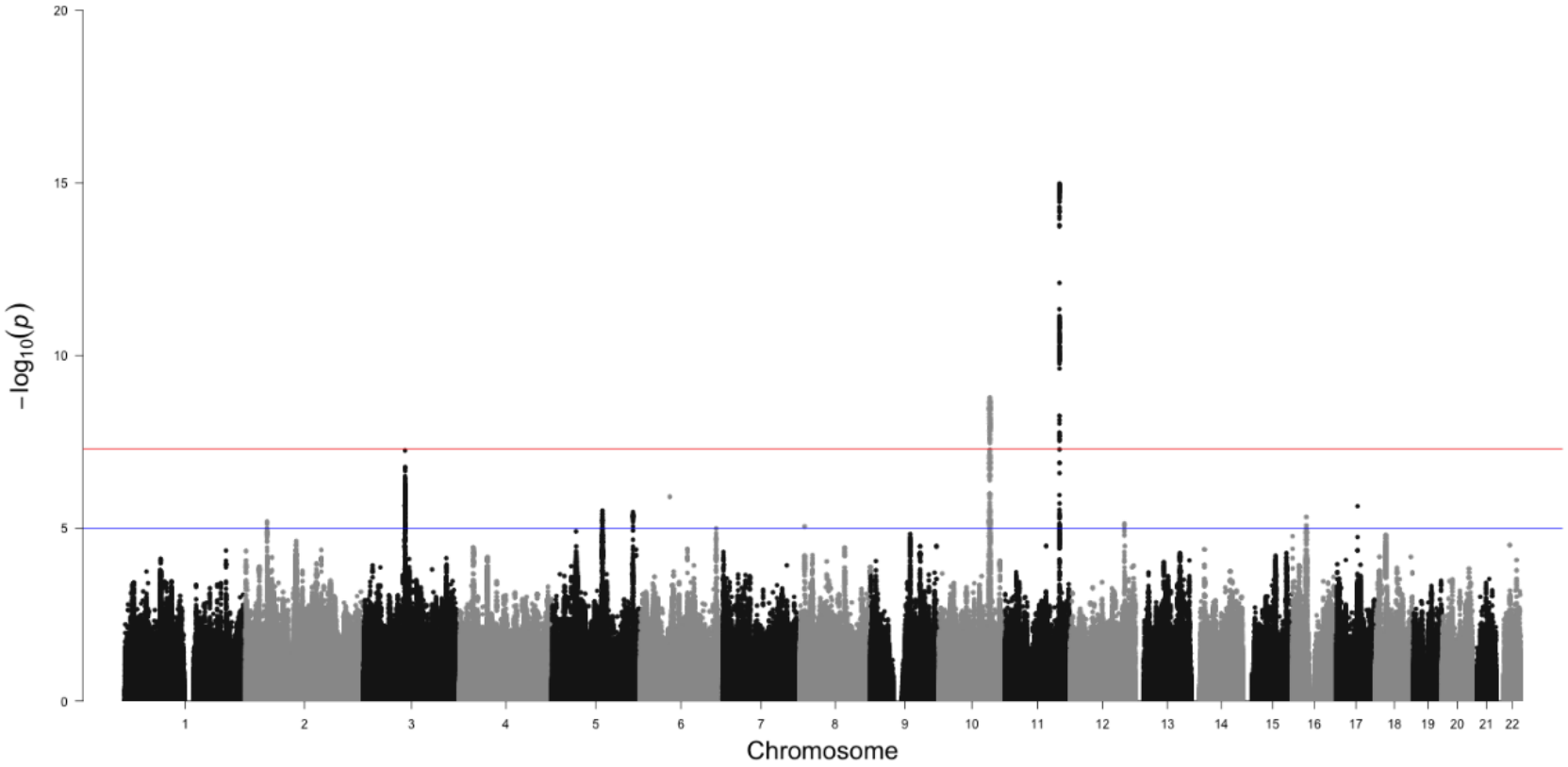
Continuous or ordinal indicators

Recent work

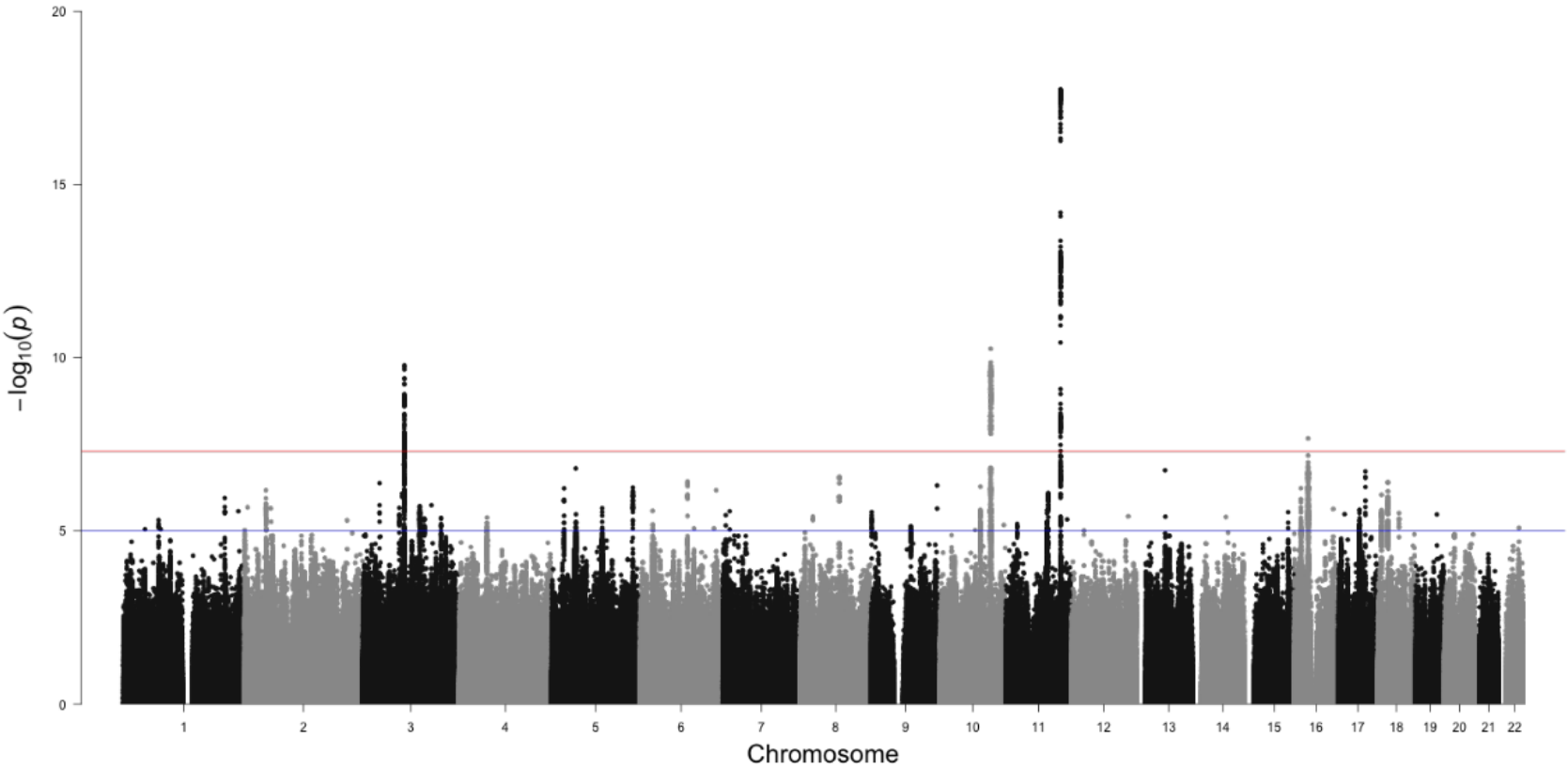
Does it matter whether we treat ordinal data as continuous or ordinal?

Examples of gene-age interactions

Continuous


20
15
10
5
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome

Ordinal

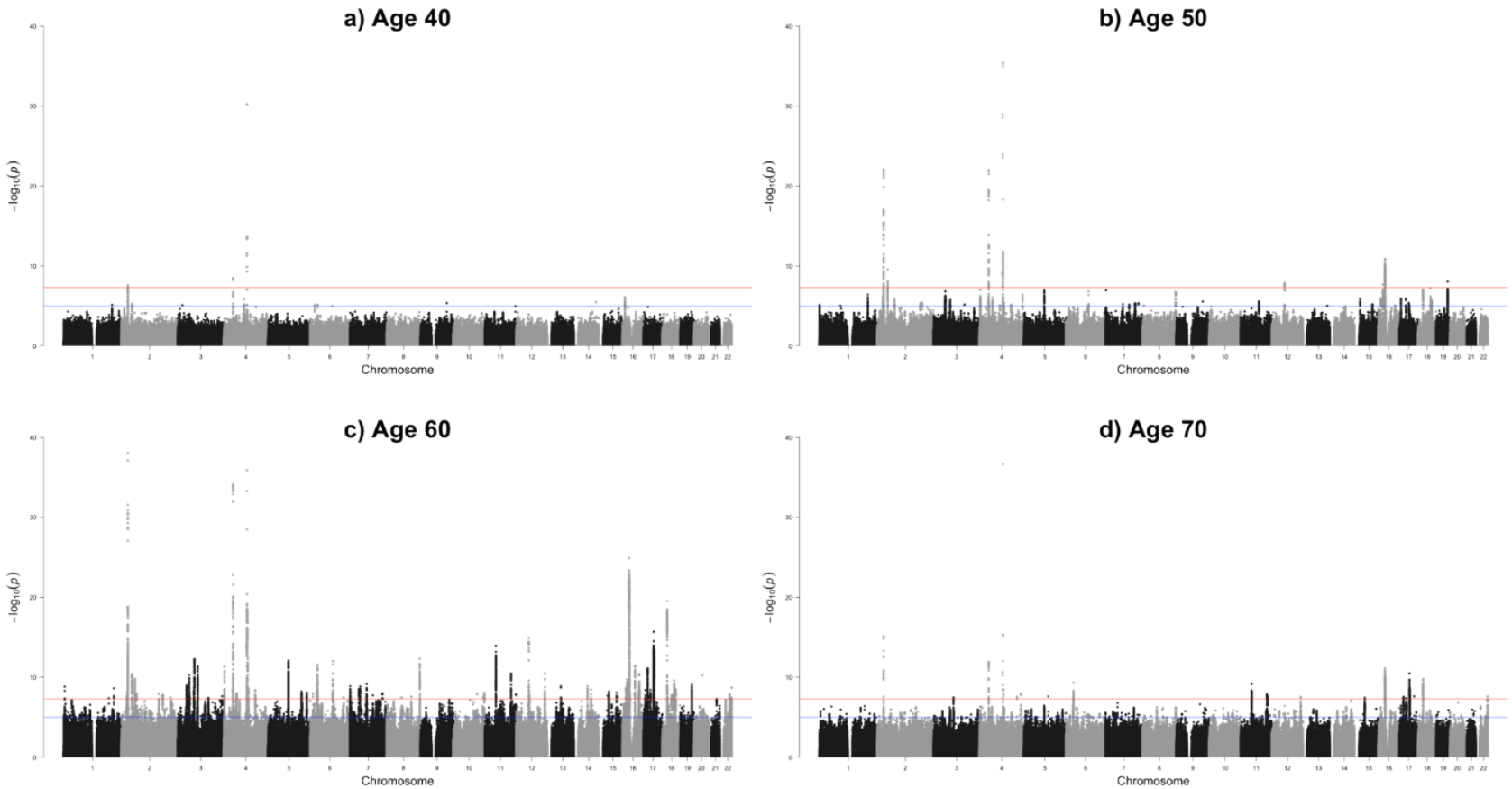

20
15
10
5
0
$-\log_{10}(p)$
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome

Recent work

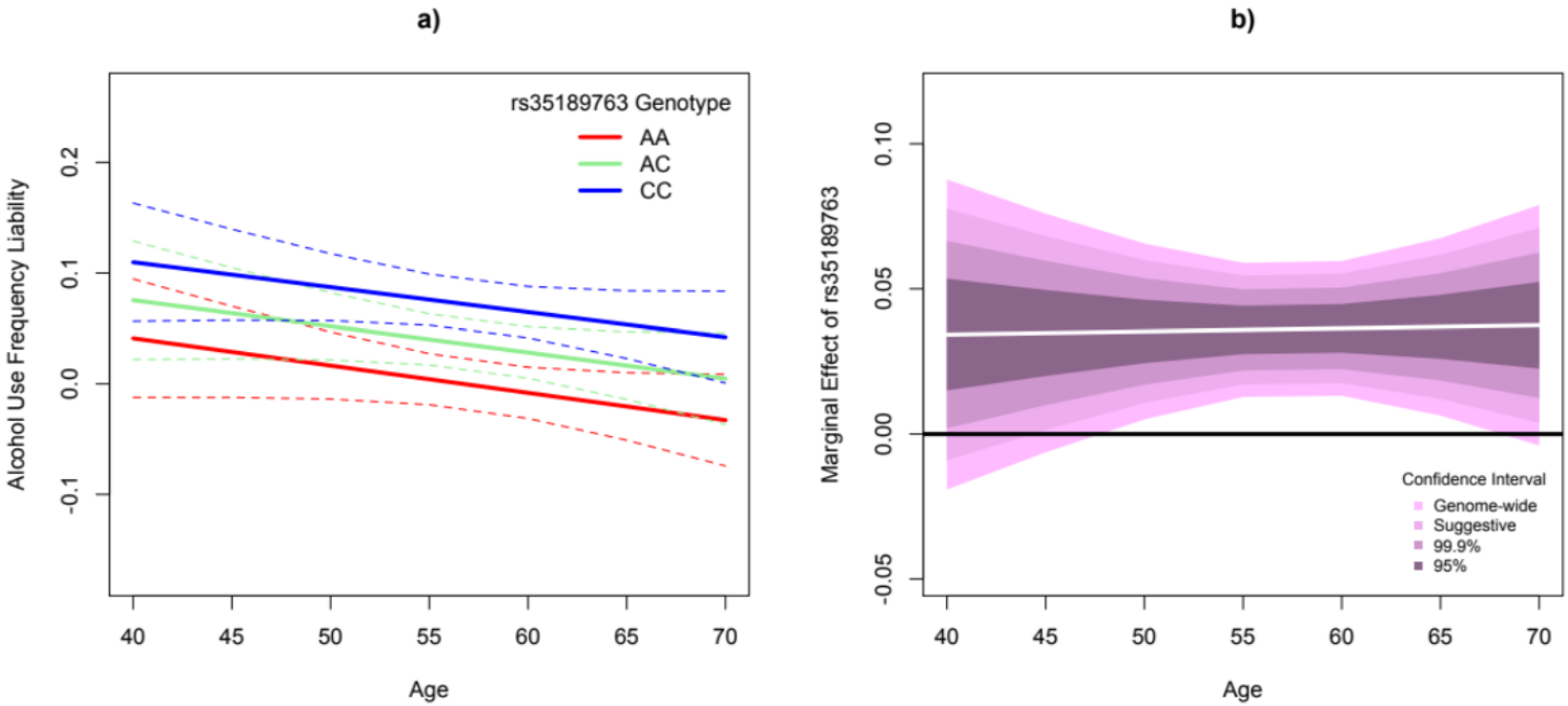
Does it matter whether we treat ordinal data as continuous or ordinal?

Examples of gene-age interactions

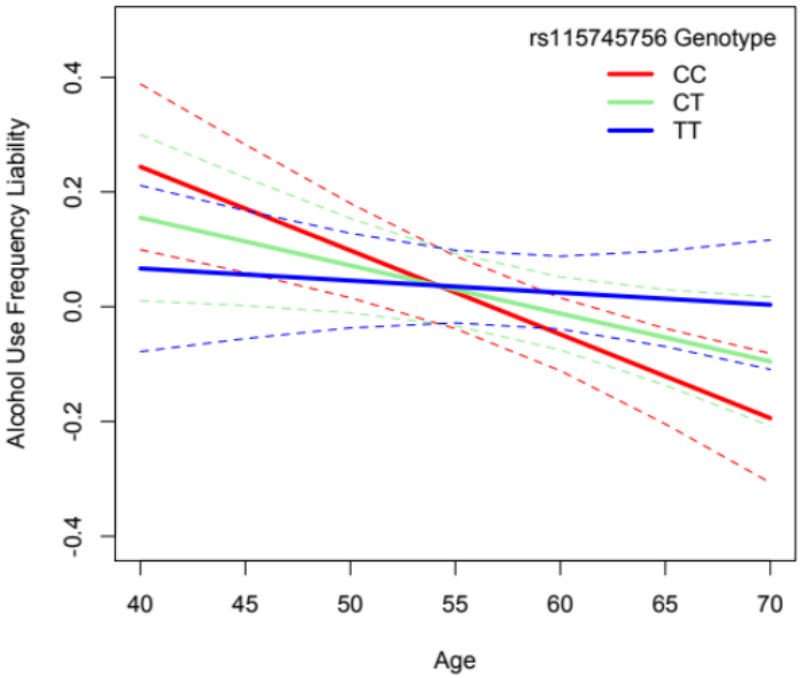
Hits by age

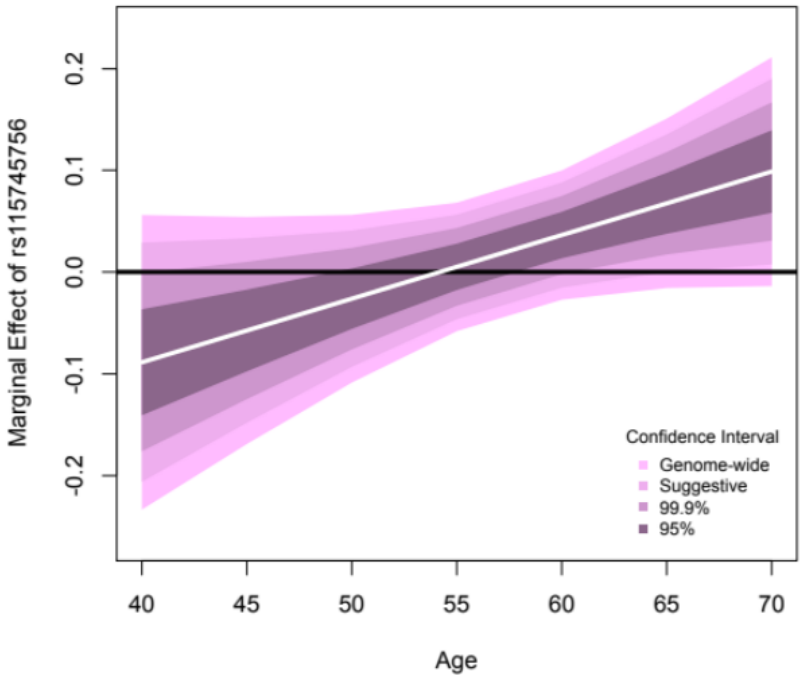


rs35189763 on Alcohol by Age



rs115745756 on Alcohol by Age





Next steps

Compare

- GW-SEM w/ factor model
- GW-SEM w/ sum-score
- GenomicSEM on GWAS summary stats
- TATES principle component analysis

Preliminary results look too good to be true

Next steps

Compare

- GW-SEM w/ factor model
- GW-SEM w/ sum-score
- GenomicSEM on GWAS summary stats
- TATES principle component analysis

Preliminary results look too good to be true

Entrancing beauty of our backyard

Grotzinger, A. D., Rhemtulla, M., de Vlaming, R., Ritchie, S. J., Mallard, T. T., Hill, W. D., ... others (2019). Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature Human Behaviour*, *3*(5), 513–525. doi: 10.1038/s41562-019-0566-x

Kuhnert, P. M., & Do, K.-A. (2003). Fitting genetic models to twin data with binary and ordered categorical responses: a comparison of structural equation modelling and bayesian hierarchical models. *Behavior Genetics*, *33*(4), 441–454.

Plummer, M. (2013). JAGS version 3.4.0 user manual [Computer software manual]. Retrieved from `http://mcmc-jags.sourceforge.net/`

Pritikin, J. N., Hunter, M. D., von Oertzen, T., Brick, T. R., & Boker, S. M. (2017). Many-level multilevel structural equation modeling: An efficient evaluation strategy. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(5), 684-698. doi: 10.1080/10705511.2017.1293542

Pritikin, J. N., Verhulst, B., & Neale, M. C. (in press). GW-SEM 2.0:

Efficient, flexible and accessible multivariate GWAS.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(1), 1–97. doi: 10.1007/BF03372160

Verhulst, B., Maes, H. H., & Neale, M. C. (2017). GW-SEM: A statistical package to conduct genome-wide structural equation modeling. *Behavior Genetics*, *47*(3), 345–359. doi: 10.1007/s10519-017-9842-6

Verhulst, B., Pritikin, J. N., Clifford, J., & Prom-Wormley, E. (submitted). Using genetic marginal effects to study gene-environment interactions with GWAS data.