GenomicSEM

Michel Nivard (m.g.nivard@vu.nl)

GenomicSEM

Michel Nivard (m.g.nivard@vu.nl)

Andrew Grotzinger

GenomicSEM

Michel Nivard (m.g.nivard@vu.nl)

Elliot Tucker-Drob

Andrew Grotzinger

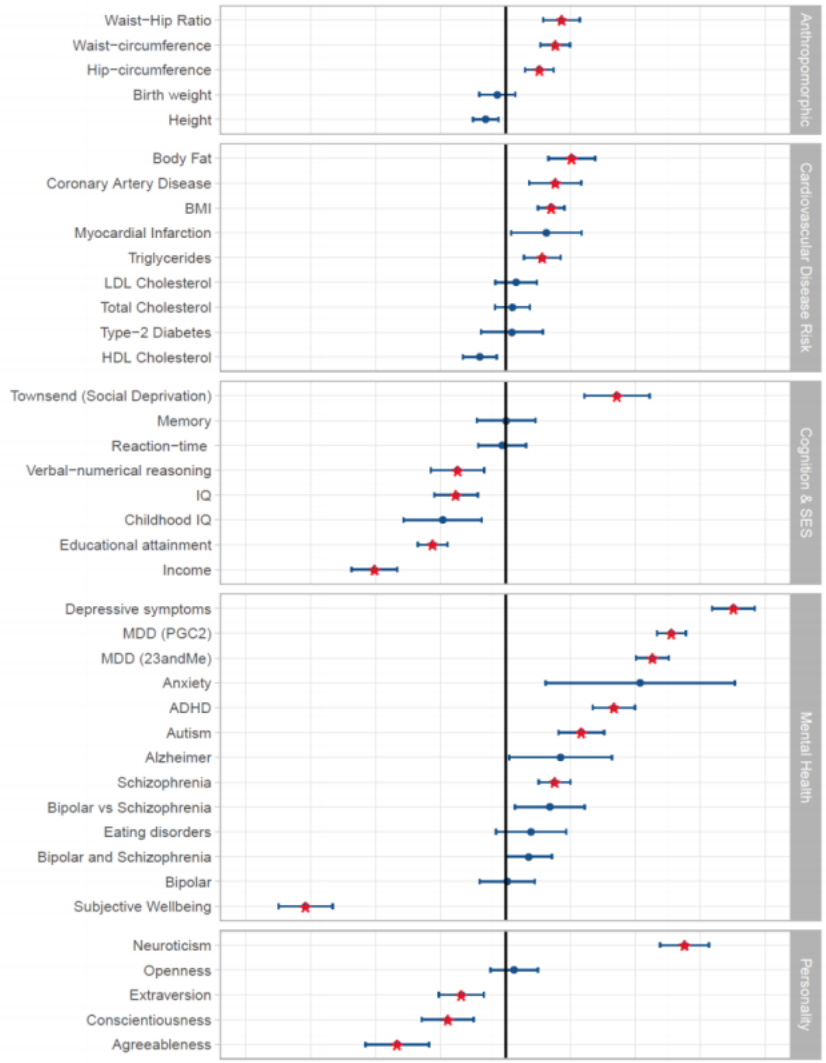
The wish to explore genetic correlations between traits is universal

- Most GWAS correlate their trait of interest with a set of potentially related traits.

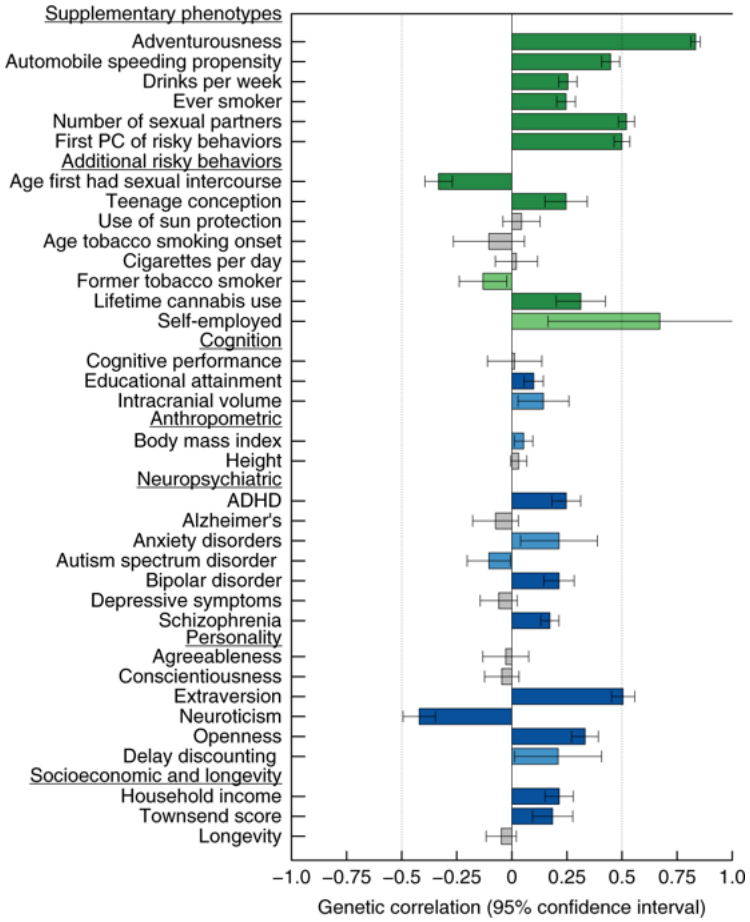
- This has become a standard figure in any large GWAS effort.

AgeSmk	CigDay	SmkInit	SmkCes	DrnkWk	
$h^2 = 0.05$	−0.38**	−0.71**	−0.31**	−0.10*	Age of smoking initiation (AgeSmk)
−0.38**	$h^2 = 0.08$	0.33**	0.42**	0.07*	Cigarettes per day (CigDay)
−0.71**	0.33**	$h^2 = 0.08$	0.40**	0.34**	Smoking initiation (SmkInit)
−0.31**	0.42**	0.40**	$h^2 = 0.05$	0.11**	Smoking cessation (SmkCes)
−0.10*	0.07*	0.34**	0.11**	$h^2 = 0.04$	Drinks per week (DrnkWk)
0.04	−0.01	−0.03	−0.10**	−0.02	Height
0.22**	−0.09*	−0.04	−0.02	0.08*	Age at menarche
0.67**	−0.40**	−0.48**	−0.46**	0.01	Age of first birth
0.55**	−0.26**	−0.40**	−0.51**	0.01	Years of education
−0.21	0.63*	0.16	0.43*	−0.02	Cotinine
−0.32**	0.15**	0.28**	0.12*	0.20**	General risk tolerance
−0.43**	0.09	0.60**	0.06	0.36**	Lifetime cannabis use
−0.31*	0.20*	0.41**	0.17*	0.17	ADHD
0.06	−0.04	0.01	−0.08	−0.03	Autism spectrum disorder
0.02	0.06	0.06	−0.10*	0.04	Bipolar disorder
−0.17*	0.12*	0.19**	0.26**	−0.06	Major depressive disorder
−0.17**	0.13**	0.20**	0.20**	0.02	Neuroticism
−0.05	0.10**	0.14**	0.06*	0.01	Schizophrenia
−0.05	−0.02	−0.06	0.08	0.13	Alzheimer's
−0.03	−0.01	−0.04	0.04	0.03	Multiple sclerosis
−0.02	0.02	0.02	0.01	−0.10*	Parkinson's
−0.16**	0.19**	0.12**	0.12**	−0.08*	Body mass index
−0.20**	0.24**	0.13**	0.17**	−0.11**	Obesity class I
0.03	−0.04	−0.08*	0.02	0.03	Bone density: femoral neck
0.03	0.01	−0.06	0.04	0.02	Lumbar spine
0.16**	−0.17**	−0.09**	−0.15**	0.17**	Cholesterol: HDL
−0.06	0.07*	0.02	0.10*	−0.03	LDL
−0.03	0.07*	0.03	0.08*	−0.01	Total
0.01	0.05	−0.08	0.16*	−0.11	Chronic kidney disease
−0.27**	0.25**	0.19**	0.21**	−0.01	Coronary artery disease
−0.16*	0.15*	0.07*	0.06	−0.08*	Diabetes: type 2
−0.13*	0.17*	0.07*	0.11*	−0.01	Fasting main effect: glucose
−0.24*	0.16*	0.09*	0.10	−0.24**	Insulin
−0.17	0.22*	0.11	0.20	−0.01	Proinsulin
−0.03	0.14*	0.04	0.11*	−0.04	Heart rate
0.04	0.03	0.01	−0.05	−0.06*	Inflammatory bowel disease
0.08	−0.01	−0.03	−0.13*	−0.06	Ulcerative colitis
−0.04	0.08	0.08	0.02	−0.07	Primary biliary cirrhosis
0.06	0.07	0.01	0.06	−0.06	Systemic lupus erythematosus

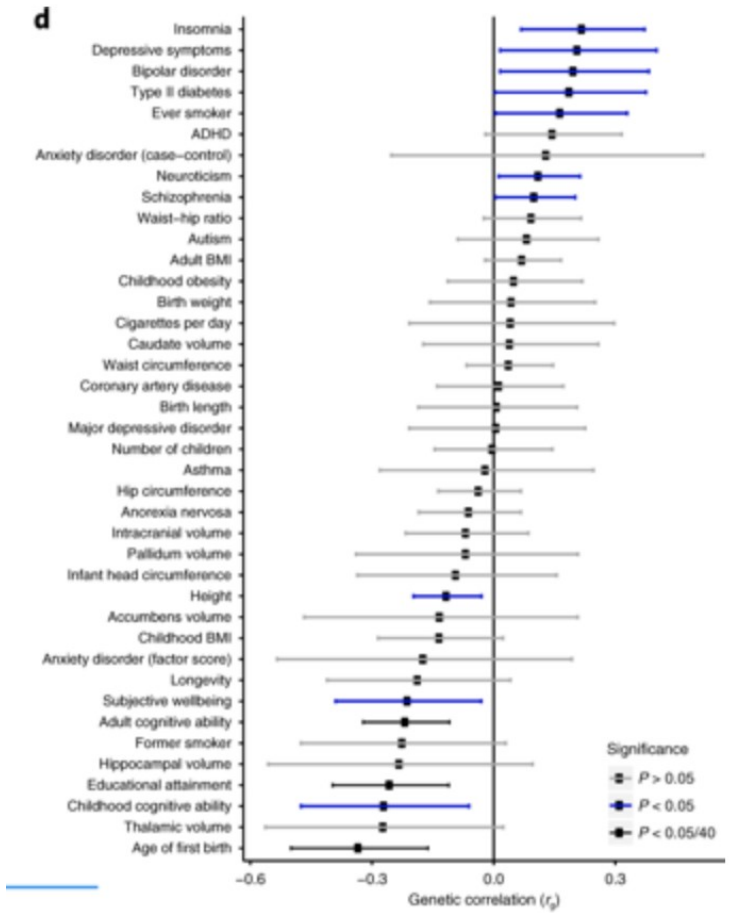
Liu et al Substance use, NG 2019



Abdellaoui loneliness, NG 2019

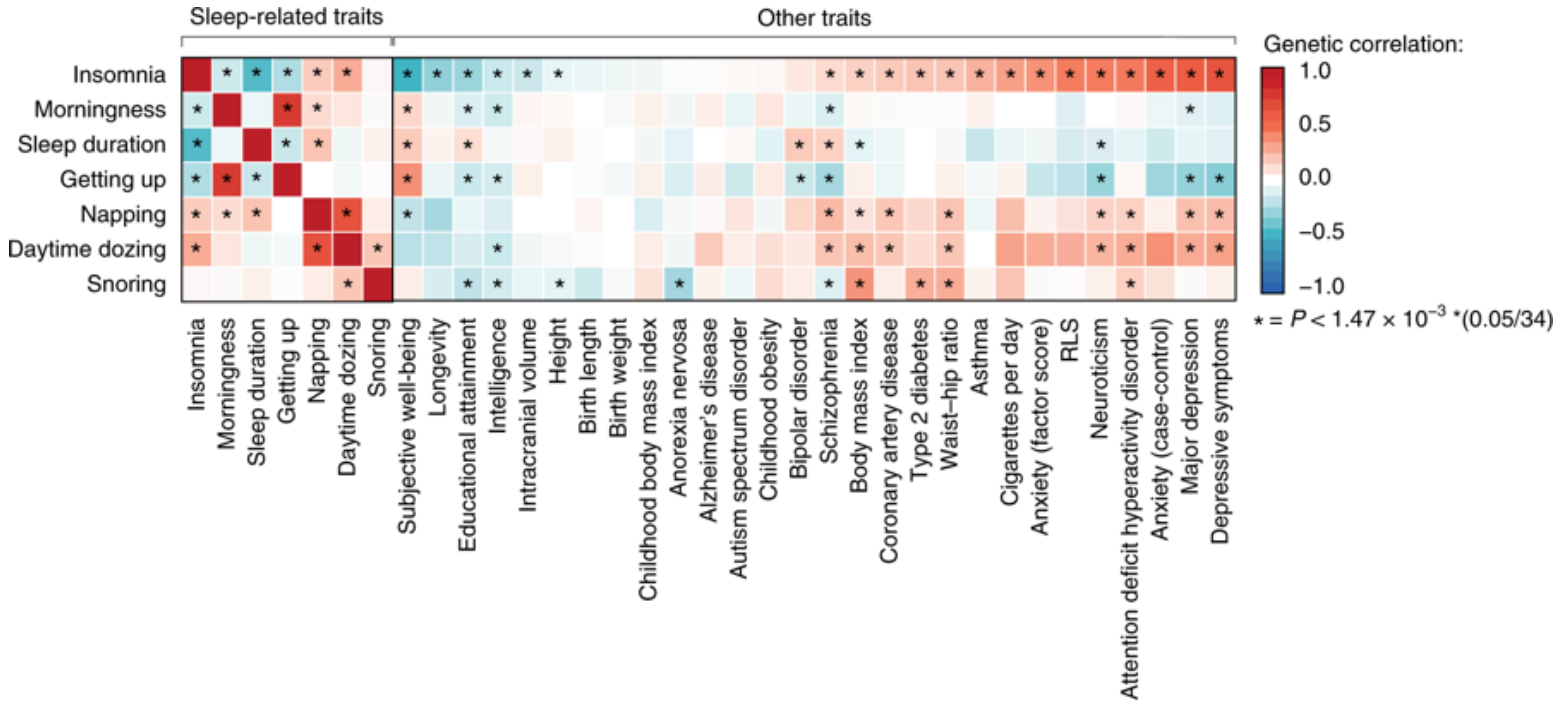


Linner et al. Risk seeking Behaviour, NG 2019



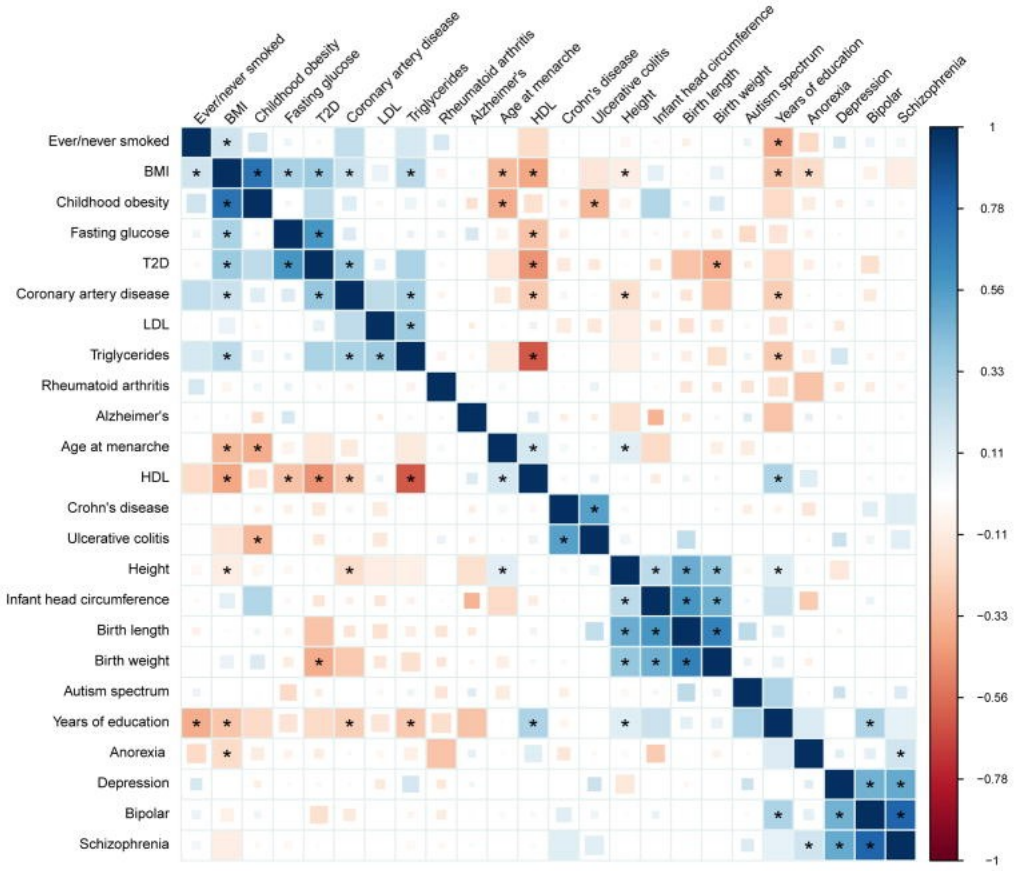
Jansen et al. Alzheimer’s Disease NG, 2019

In GWAS we frequently consider only only a slice of the matrix

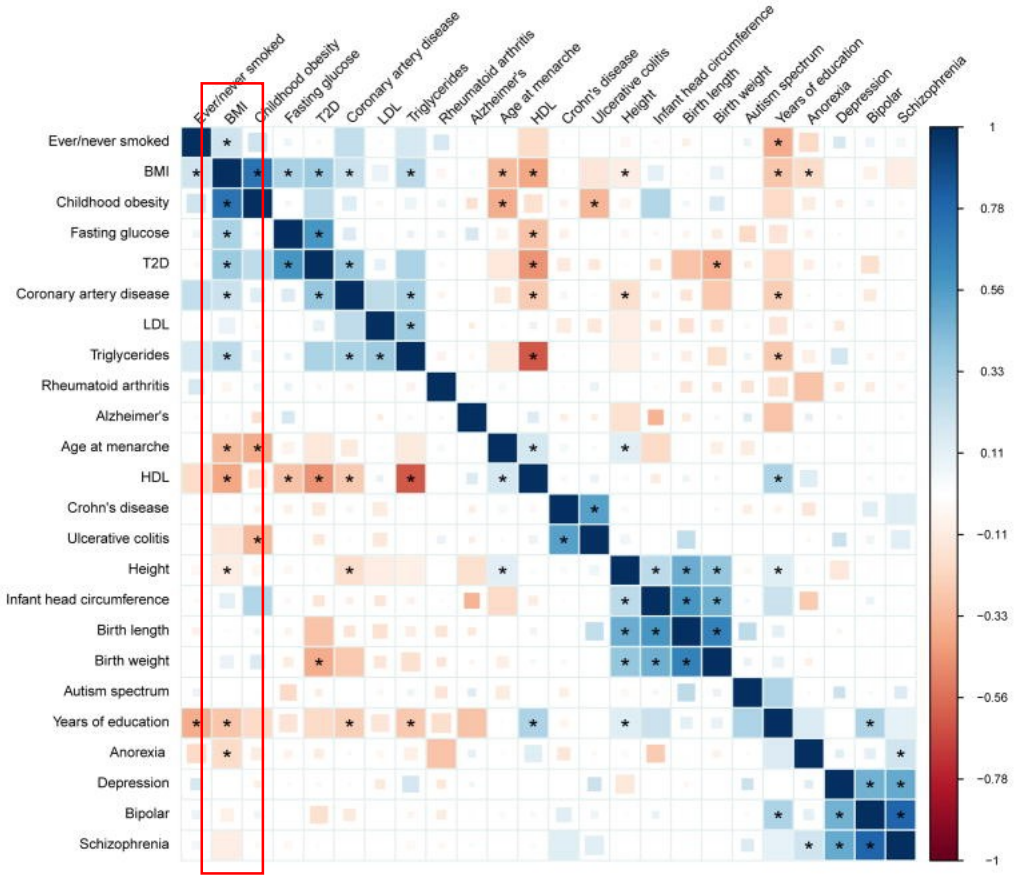


Jansen et al, Insomnia, NG, 2019

So consider the full correlation/covariance matrix:

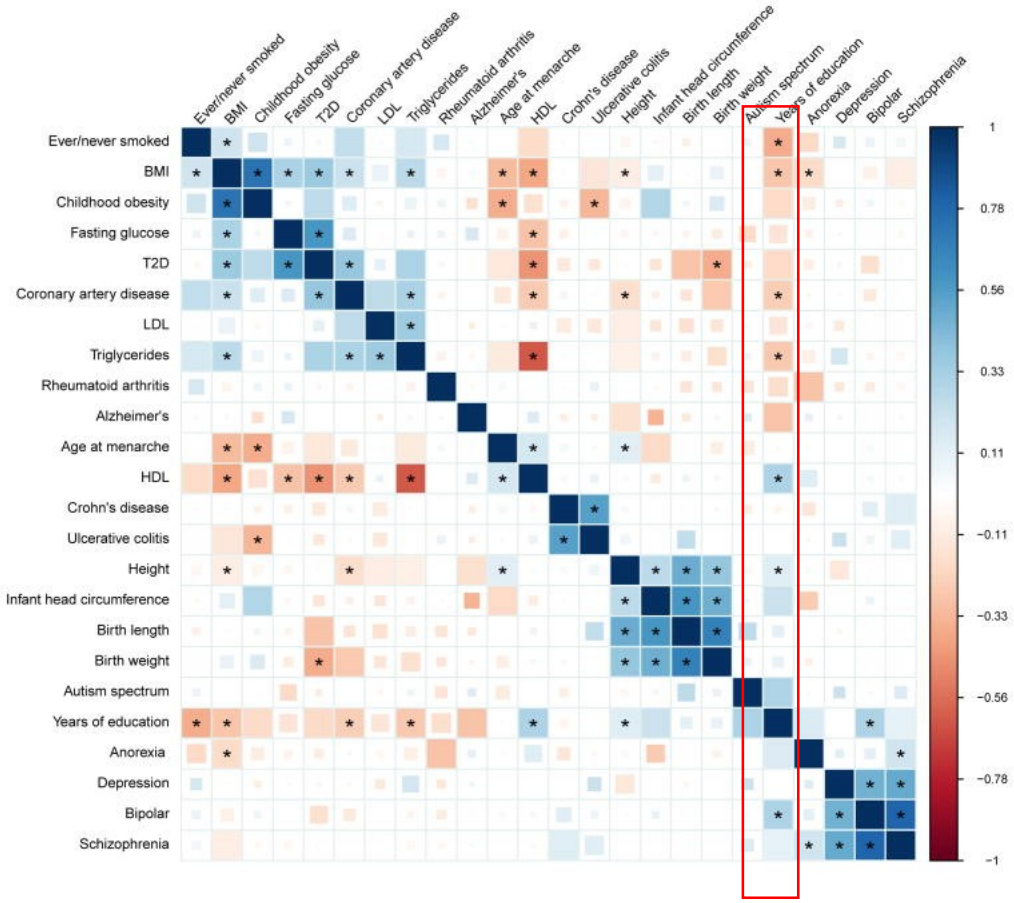


So consider the full correlation/covariance matrix:



Hypothetical GIANT consortium figure

So consider the full correlation/covariance matrix:



Hypothetical SSGAC consortium figure

Our solution: GenomicSEM

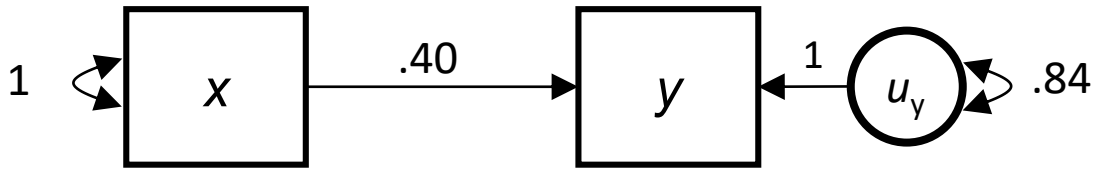
- Apply structural equation model, to estimated genetic covariance matrices

- Make sure we are able to infer things about the parameter (are they equal, are they > 0 are they < 1)

- SEE PREPRINT: https://www.biorxiv.org/content/10.1101/305029v1
- AND GITHUB: https://github.com/MichelNivard/GenomicSEM
- TUTORIALS: https://github.com/MichelNivard/GenomicSEM/wiki

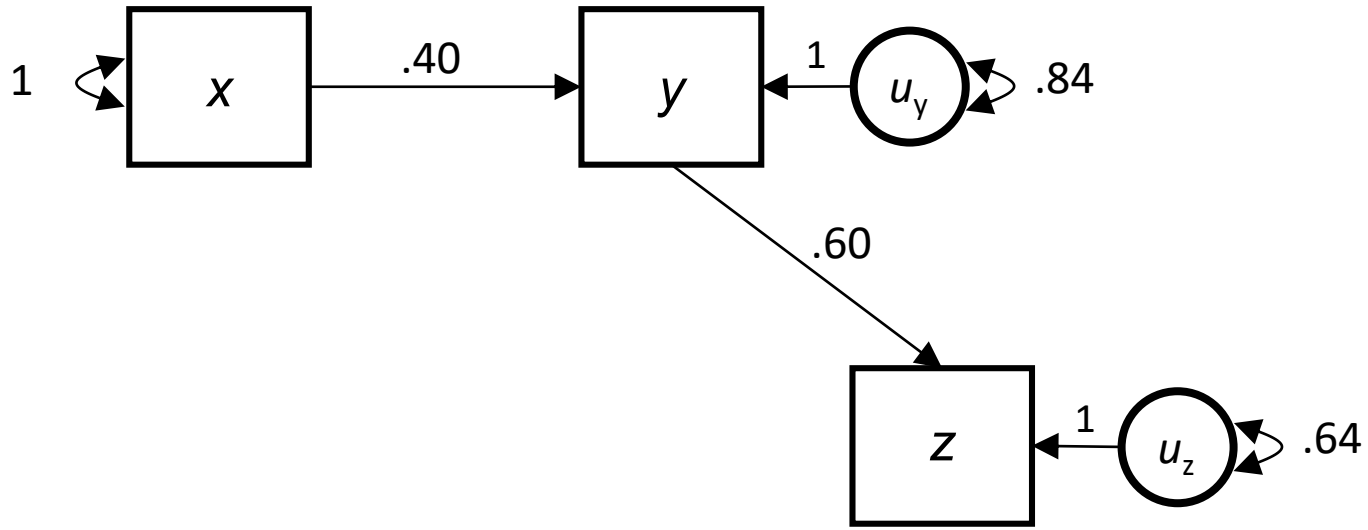
Ultra short primer on SEM

Imagine we knew the generating causal process



$y = .40\,x + u_y$ $\quad\quad$ $x \sim (0,1)$, $u_y \sim (0,.84)$

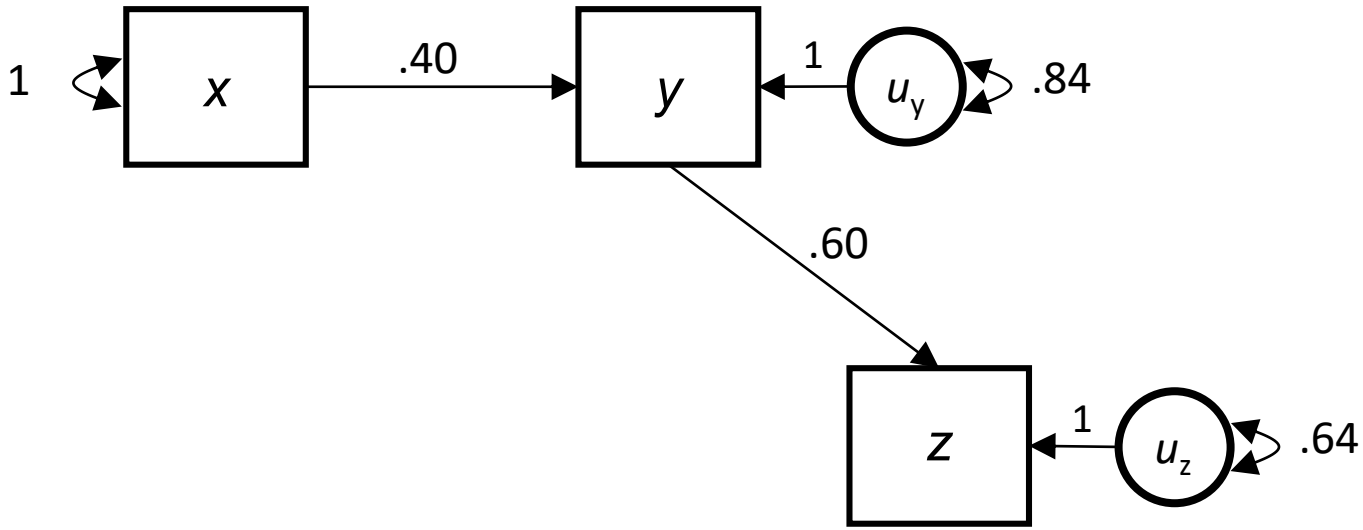
Imagine we knew the generating causal process



$y = .40\,x + u_y$ $\quad\quad$ $x \sim (0,1)$, $u_y \sim (0,.84)$

$z = .60\,y + u_z$ $\quad\quad$ $u_z \sim (0,.64)$

Imagine we knew the generating causal process



Implied covariance matrix
in the population

$\text{cov}(x,y,z)_{pop}$ =

1.00		
.40	1.00	
.24	.60	1.00

$y = .40\,x + u_y$ $\qquad$ $x \sim (0,1)$, $u_y \sim (0,.84)$

$z = .60\,y + u_z$ $\qquad$ $u_z \sim (0,.64)$

In practice, we only observe the sample data, and we propose a model

observed covariance matrix
in a sample

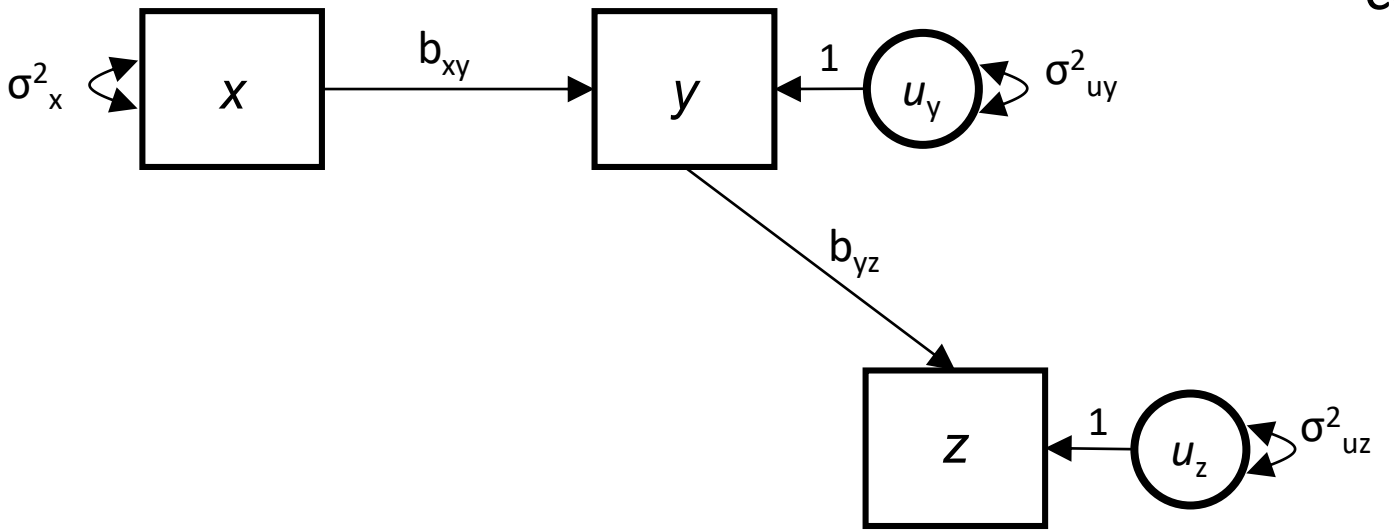
.94		
.33	1.02	
.27	.62	1.02

≈

covariance matrix
in population

1.00		
.40	1.00	
.24	.60	1.00

For the proposed model, estimate parameters from the data, and evaluate model fit to the data



$cov(x,y,z)_{sample}$ =

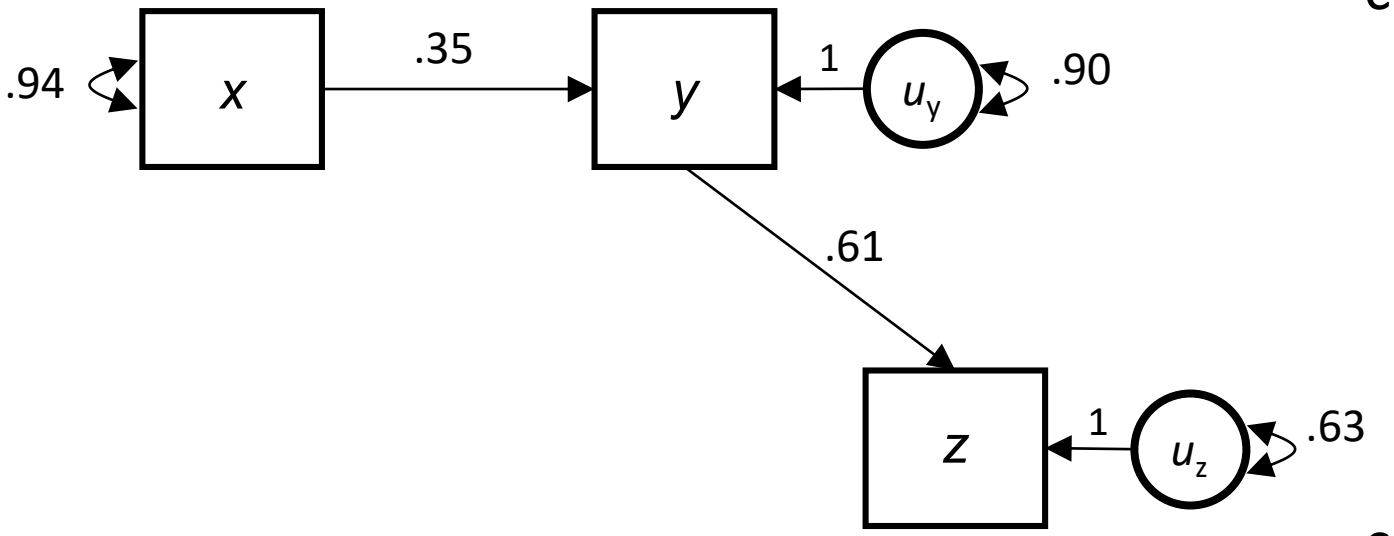
.94		
.33	1.02	
.27	.62	1.02

6 unique elements in the covariance matrix being modeled

5 free model parameters

1 df

For the proposed model, estimate parameters from the data, and evaluate model fit to the data



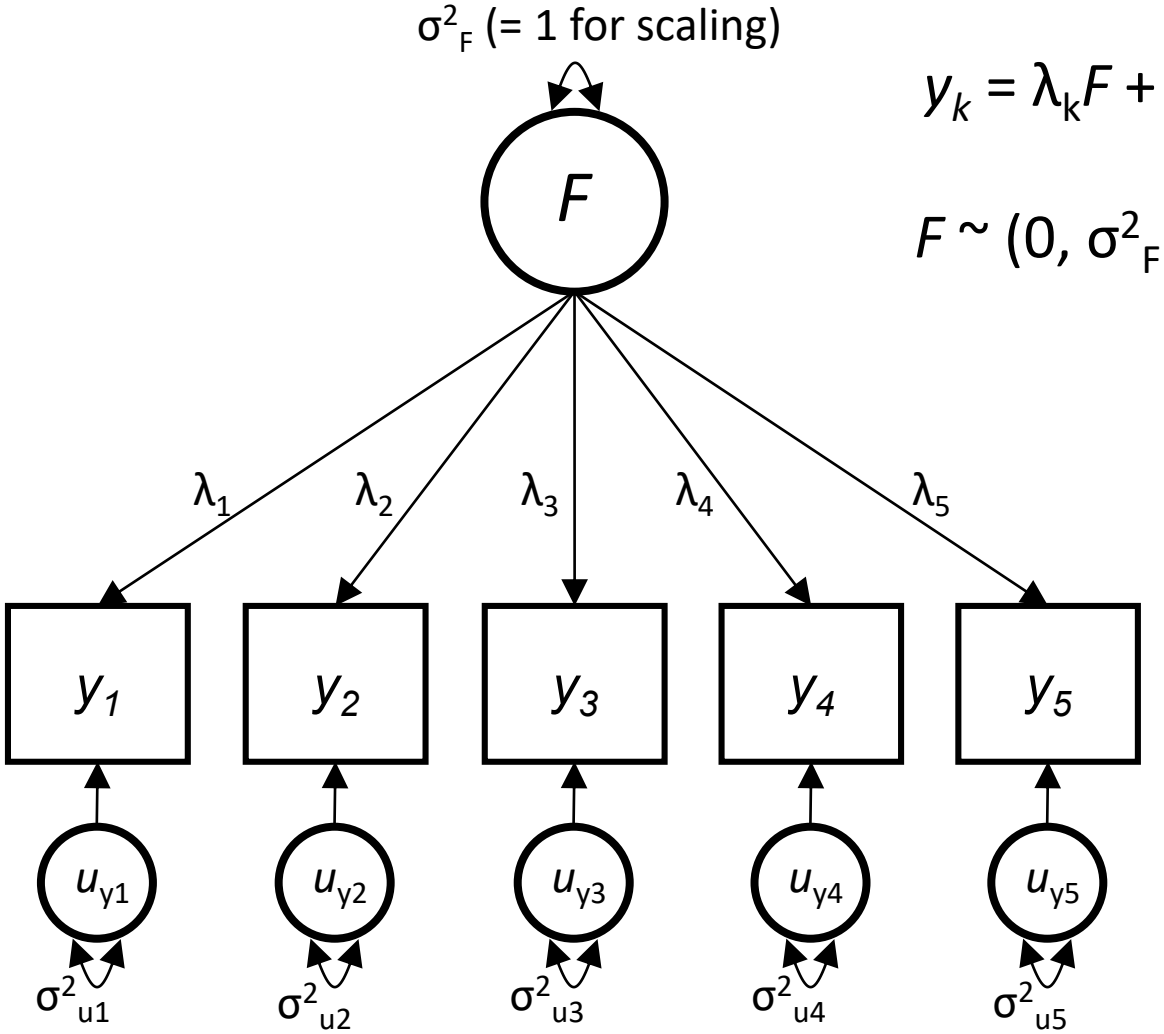
$\text{cov}(x,y,z)_{\text{sample}}$ =

.94		
.33	1.02	
.27	.62	1.02

$\text{cov}(x,y,z)_{\text{implied}}$ =

.94		
.33	1.03	
.20	.63	1.00

The model that we fit may include some variables for which we do not observe data

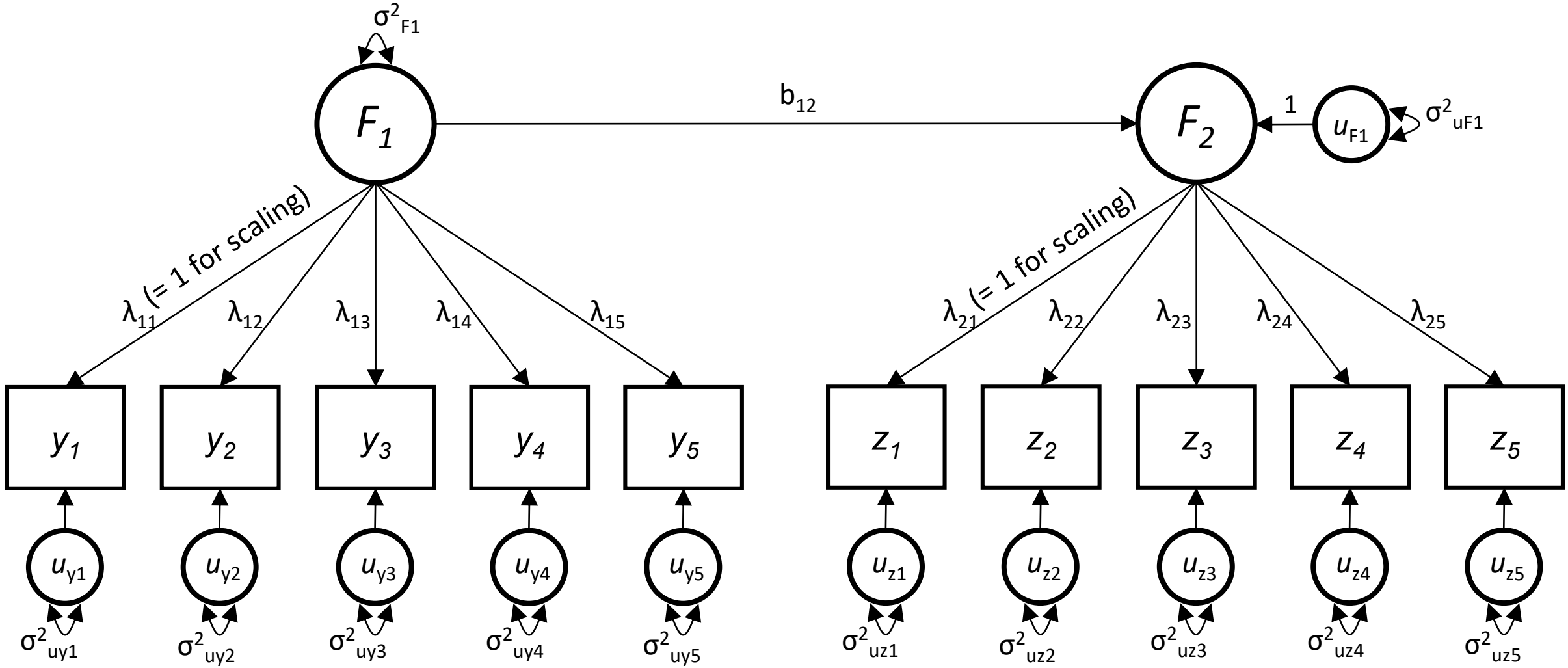


$y_k = \lambda_k F + u_k$

$F \sim (0, \sigma^2_F)$, $u_{yk} \sim (0, \sigma^2_{uk})$

F is unobserved.
Parameters are estimated from,
and fit is evaluated relative to,
the sample covariance matrix for y_1-y_k.

The model that we fit may include some variables for which we do not observe data



Some basic rules for SEM

- Cannot estimate more parameters then UNIQUE ellements in the covariance matrix.

- Each parameter has to have a unique functional relationship to the observed covariances.

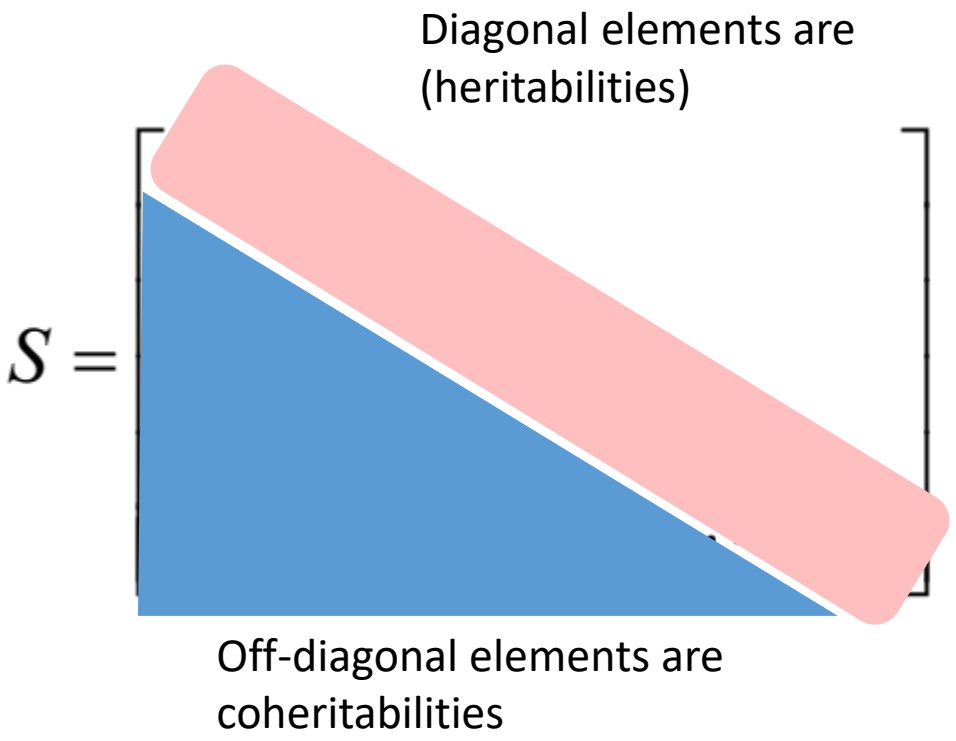
Start with GWAS Summary Statistics for the Phenotypes of Interest

- No need for raw data
- No need to conduct a primary GWAS yourself: Download them online!
  - sumstats for over 3700 phenotypes have been helpfully indexed at http://atlas.ctglab.nl/
  - sumstats for over 4000 UK Biobank phenotypes are downloadable at http://www.nealelab.is/uk-biobank

CHR	SNP	BP	A1	A2	INFO	OR	SE	P	Nca	Nco	MAF
8	rs62513865	101592213	T	C	0.957	1.01461	0.0153	0.3438	59851	113154	0.07330
8	rs79643588	106973048	A	G	0.999	1.02122	0.0136	0.1231	59851	113154	0.09200
8	rs17396518	108690829	T	G	0.980	1.00331	0.0080	0.6821	59851	113154	0.43500
8	rs6994300	102569817	A	G	0.466	0.88126	0.4243	0.7658	16823	25632	0.00556
8	rs138449472	108580746	A	G	0.734	0.97181	0.0598	0.6320	41253	79756	0.00852
8	rs983166	108681675	A	C	0.991	0.99144	0.0080	0.2784	59851	113154	0.43200

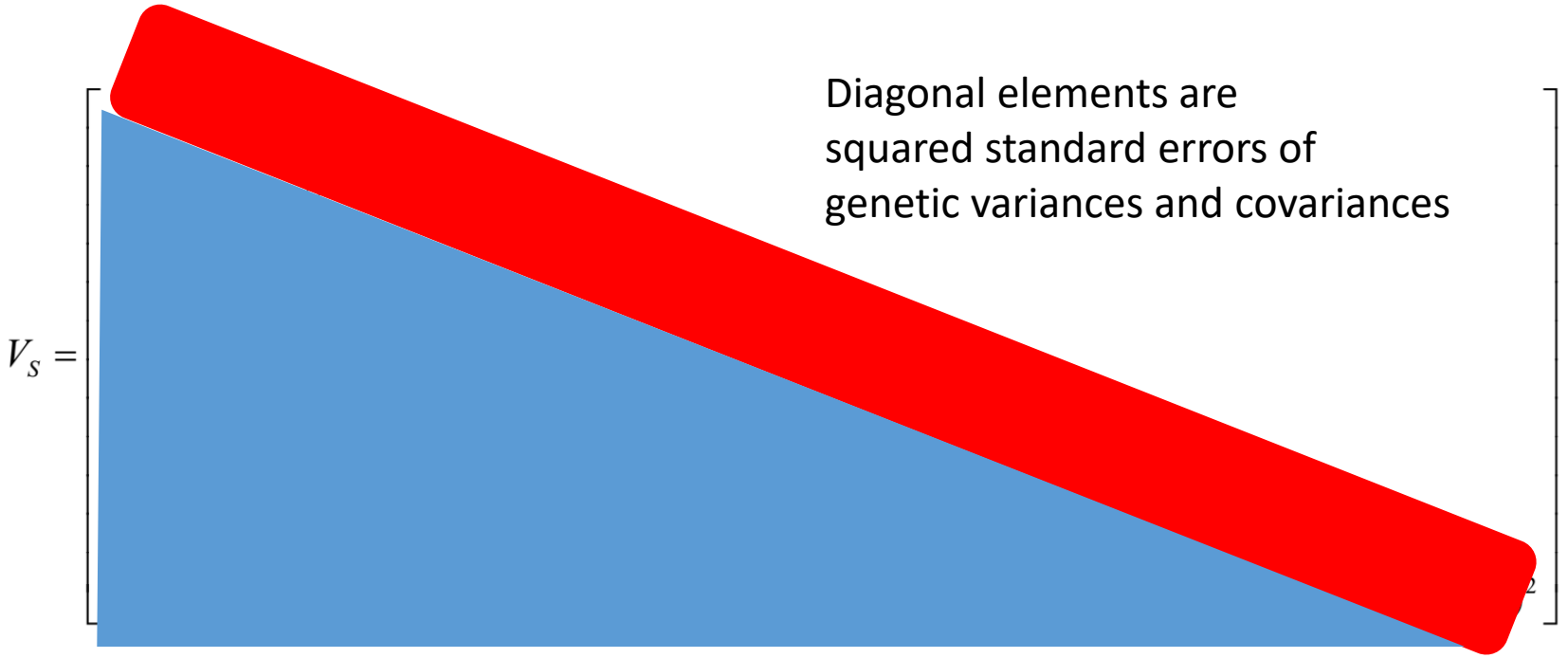
Stage 1 Estimation: Multivariable LDSC

Create a genetic covariance matrix, S: an “atlas of genetic correlations”



Stage 1 Estimation: Multivariable LDSC

Also produced is a second matrix, V, of squared standard errors and the dependencies between estimation errors



Off-diagonal elements are dependencies between estimation errors used to directly model dependencies that occur due to sample overlap from contributing GWASs

Function we minimize:

- s <- vector of unqiue elements in S:

- $F_{WLS} = (s - \Sigma(\theta))' * \text{diag(V)}^{-1} * (s - \Sigma(\theta))$

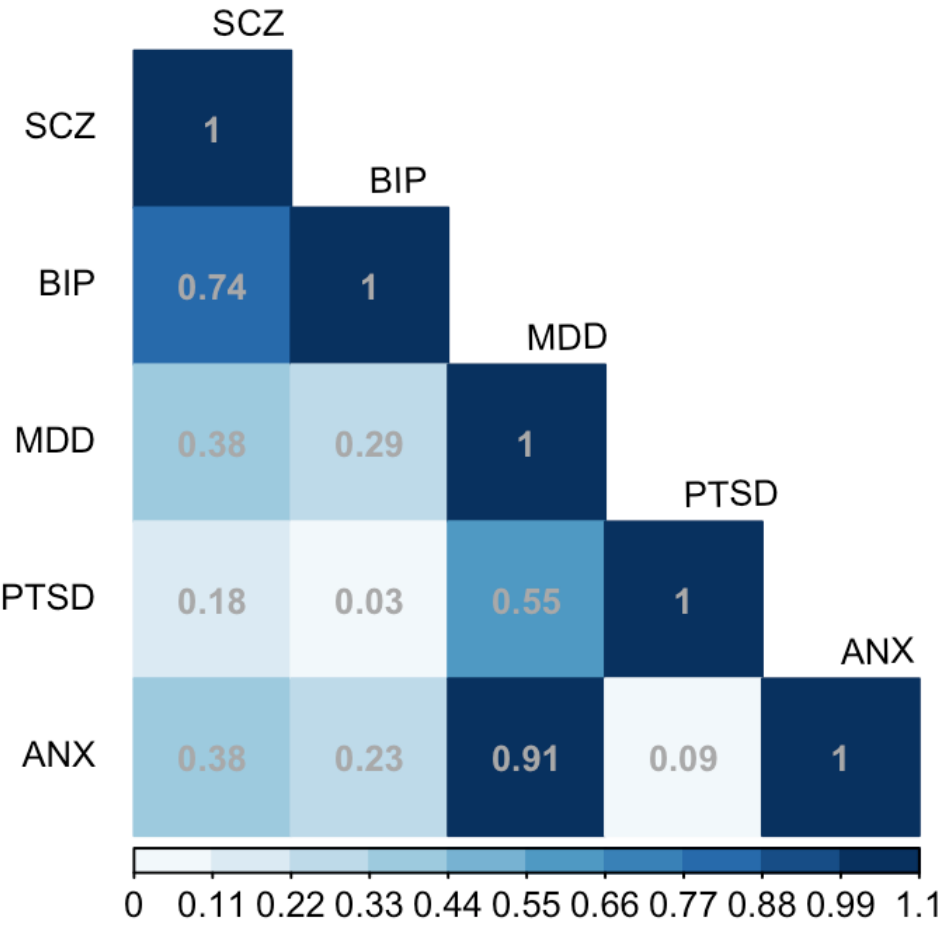
How to use genomicSEM (WITHOUT SNPs)

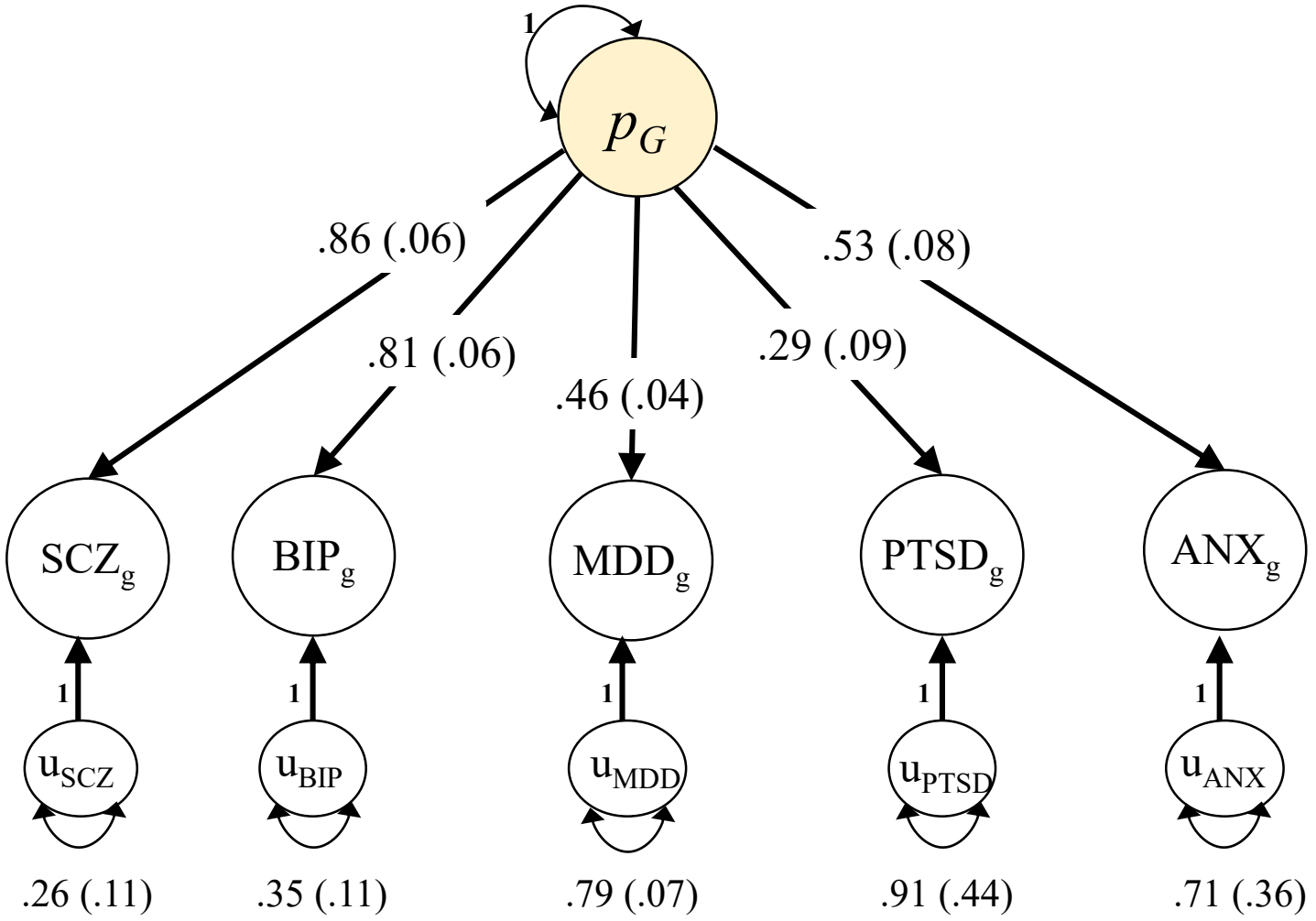
- **Step 1:** download and munge data for traits you want to analyse
  - **R function: munge()**
- Step 1b: Download LD scores
  - (wget https://data.broadinstitute.org/alkesgroup/LDSCORE/eur_w_ld_chr.tar.bz2)
- Step 2: run Ldscore regression
  - **R function: ldsc()**
- Step 3: run a model!
  - R functions: usermodel(),

How to use genomicSEM (WITH SNPs)

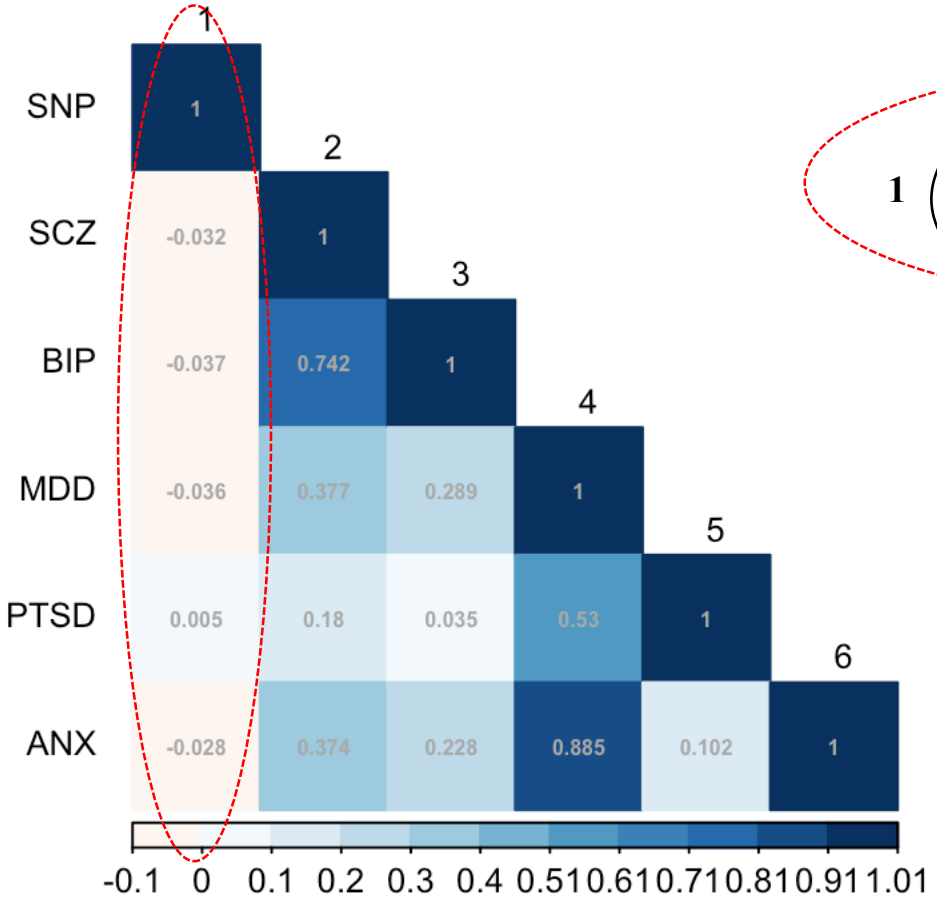
- **Step 1:** download and munge data for traits you want to analyse
  - **R function: munge()**
- Step 1b: Download LD scores
  - (wget https://data.broadinstitute.org/alkesgroup/LDSCORE/eur_w_ld_chr.tar.bz2)
- Step 2: run Ldscore regression
  - **R function: ldsc()**
- Step 3: read the sumstats into R:
  - **R function: sumstats()**
- Step 4: For each SNP, add the SNP to the covariance matrix S and V
  - **R function: addSNP()**
- Step 5: Run the model X million times:
  - **R functions: userGWAS(), userGWASpar() commonfactorGWAS(), commonfactorGWASpar()**

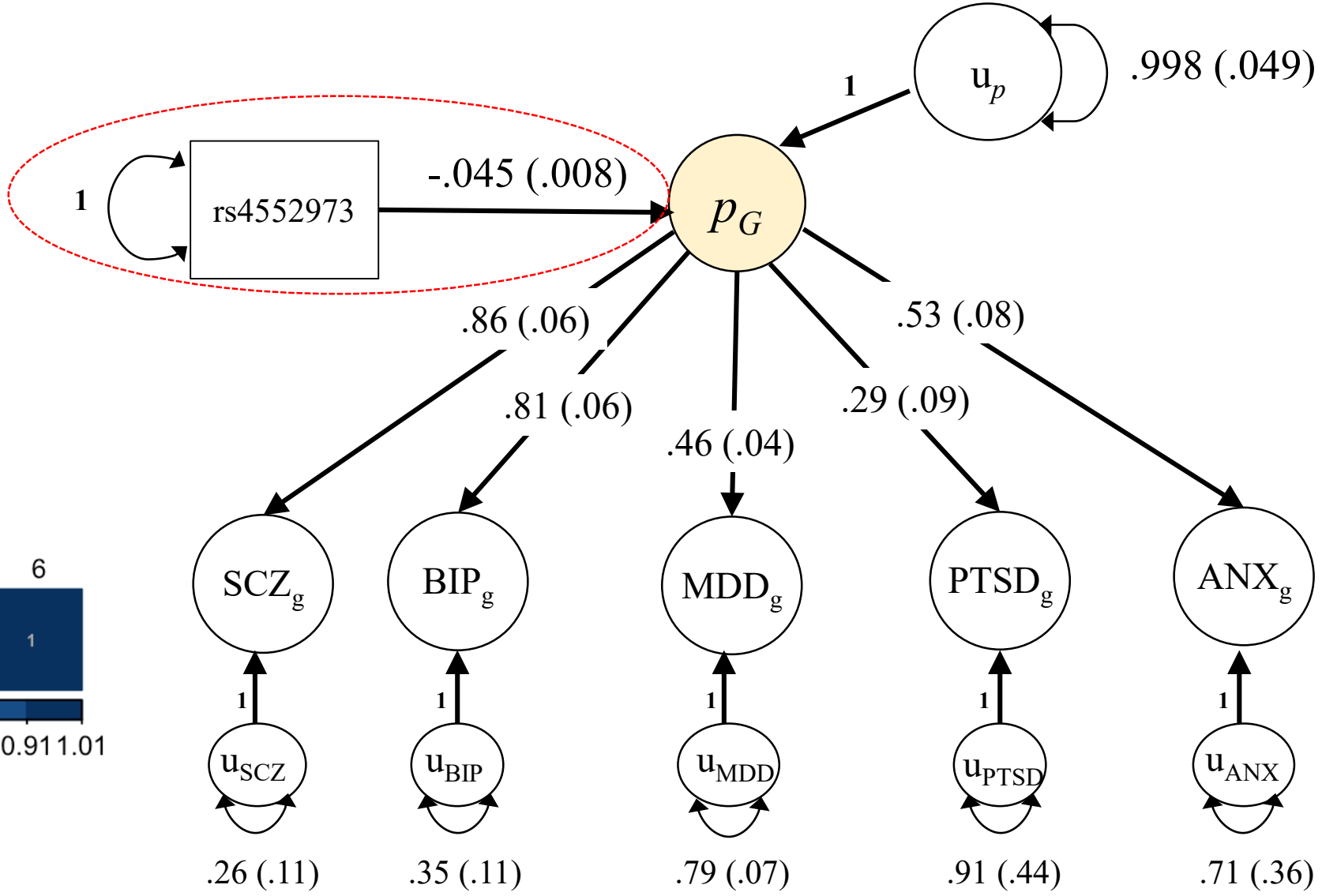
Genetic Correlation Matrix





Genetic Correlation Matrix





Show them the WIKI

- This is just a stage direction for me…..

- https://github.com/MichelNivard/GenomicSEM/wiki

Step 1: munge sumstats : Example code

```
sums <- c("GWAS_EA_excl23andMe (1).txt",
          "adhd_eur_jun2017”)

names <- c("EA", "ADHD")

N <- c(766345,55374)


munge(files = sums,trait.names = names,hm3 =
"w_hm3.noMHC.snplist", N = N)
```

Step 2: run LDSC, example code:

```
# ADHD, sigarets per day, EA, blood prerasure and BMI
traits <-        c("ADHD.sumstats.gz","smoking.sumstats.gz","EA.sumstats.gz",
                   "BP.sumstats.gz","BMI.sumstats.gz")
sample.prev = c(.36,NA,NA,.29,NA)
population.prev = c(.06,NA,NA,.29,NA)
ld ="eur_w_ld_chr/"
wld="eur_w_ld_chr/"

ldsc.object <- ldsc(traits = traits,
                    sample.prev = sample.prev,
                    population.prev = population.prev,
                    ld = "eur_w_ld_chr/",
                    wld = "eur_w_ld_chr/",
                    trait.names = c("ADHD","smoking","EA","BP","BMI"))
```

Step 3a: specify a model

We use the R formula language, slightly extended:

Regression:

A ~ B

(Co)variance:

A ~~ A; A ~~ B

Factor:

F1 =~ A + B + C + D

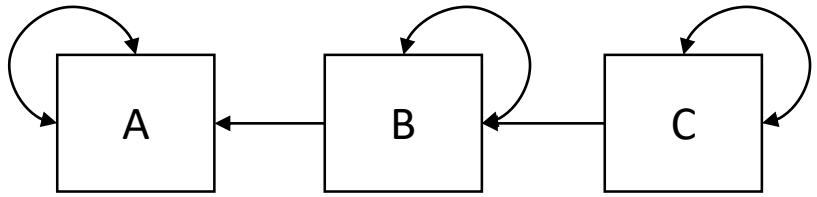
Fix a parameter:

A ~~ 1*B (the covariance between A and B is 1)

Lets make that a bit more specific

Model1 <- “ A ~ B

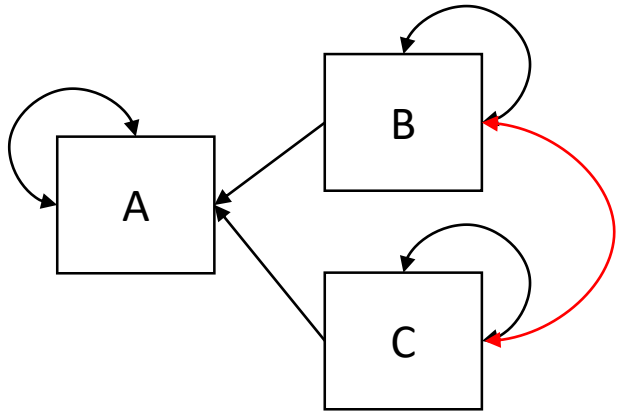
B ~ C”



Model2 <- “ A ~ B

A ~ C

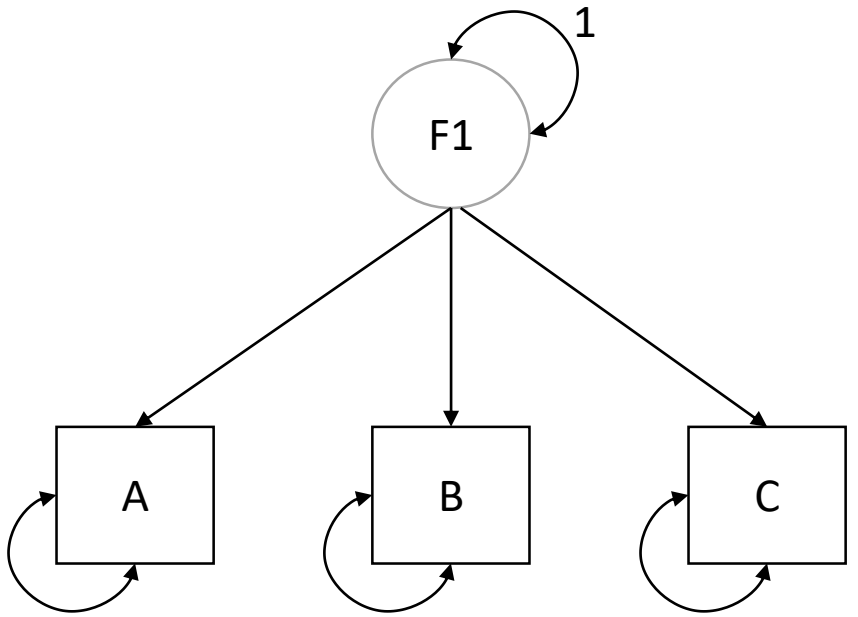
B ~~ C”



Lets make that a bit more specific

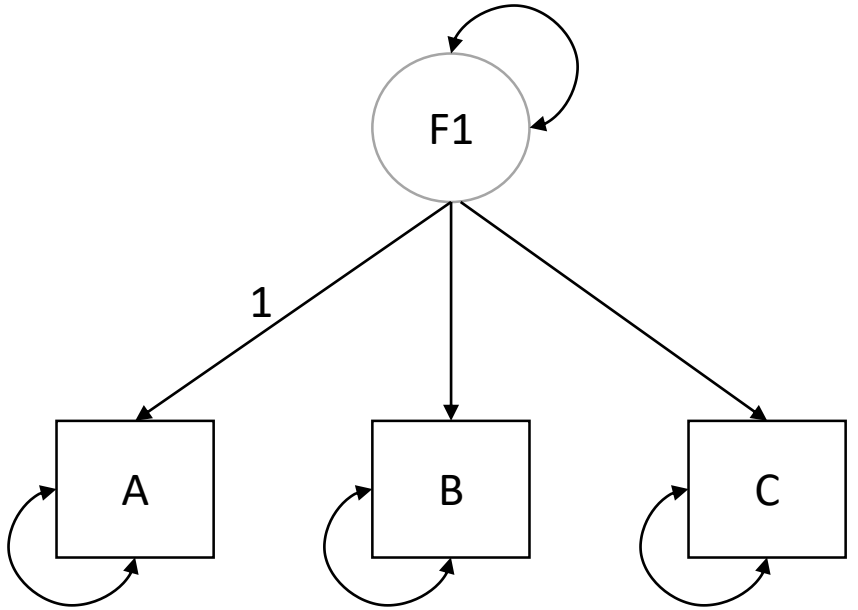
Model3 <- “ F1 =~ NA*A + B + C

F1 ~~ 1*F1”

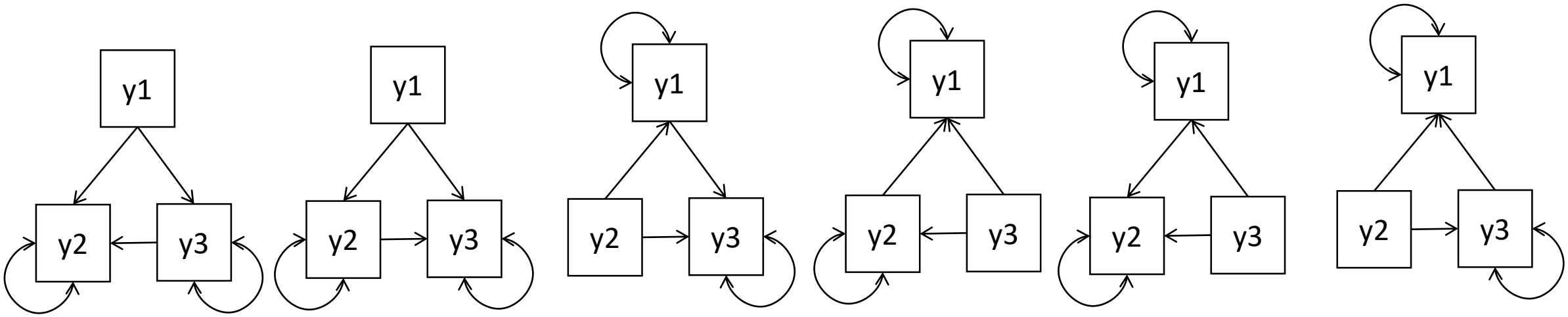


Lets make that a bit more specific

Model3 <- “ F1 =~ 1*A + B + C”



BUT: SEM is not a magical technique!



BUT: SEM is not a magical technique!

1. Their is a risk of overfitting, that is if you fit MANY models, its not likely that the best fitting model will generalize to new data

Solutions: 1) preregister which models you will consider, STICK TO YOUR SELECTION (or explain yourself)

2) develop models on the odd chromosomes, then validate on the even chromosomes.

3) Find other GWASes of the same trait(s) and confirm your chosen model fits those data well.

Some of these best Practices implemented here →

Behavior Genetics
https://doi.org/10.1007/s10519-019-09951-0

ORIGINAL RESEARCH

Check for updates

A Genetic Investigation of the Well-Being Spectrum

B. M. L. Baselmans[1,2] · M. P. van de Weijer[1,2] · A. Abdellaoui[1,5] · J. M. Vink[3] · J. J. Hottenga[1] · G. Willemsen[1] · M. G. Nivard[1] · E. J. C. de Geus[1,2,4] · D. I. Boomsma[1,2,4] · M. Bartels[1,2,4]

Received: 11 July 2018 / Accepted: 29 January 2019

PRACTICAL!

- Copy the folder : faculty/michel/2019/practical_1
- open the Rstudio project by doubleclicking: practical.Rproj
- Run the code up to line 24, this took ~ 3-5 minutes on my laptop (fingers crossed it will finish soonish on yours…)

PRACTICAL!

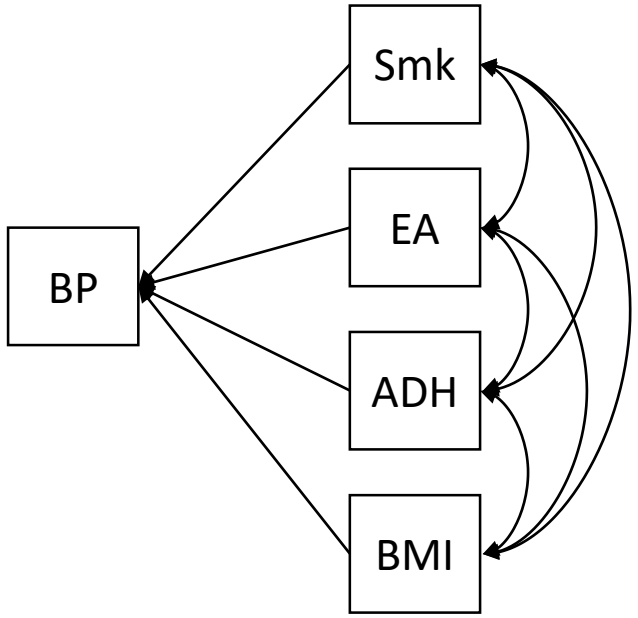
- I have provided munged GWAS sumstats for:
- High blood pressure (BP)
- BMI, smoking (packs per day in smokers), educational attainment (EA), ADHD

- I provide a baseline model for you:
    - BP ~ smoking + EA + ADHD + BMI
    - And all correlations between the variables on the right hand side

PRACTICAL!

I provide a baseline model for you:

- BP ~ smoking + EA + ADHD + BMI
- And all correlations between the variables on the right hand side



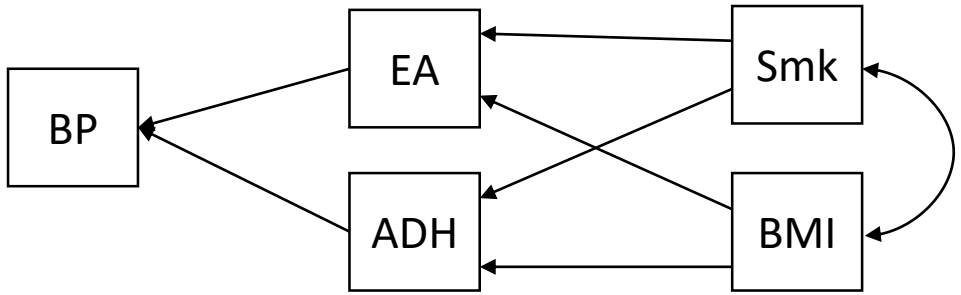
PRACTICAL

- But we are scientists, and we don't want to fool ourselfs, so you must write your model(s) down on paper, as away of preregistering them!

- I preregister the following (silly!!) model:

```
michel.model <- " BP ~ ADHD + EA
                  ADHD ~ BMI + smoking
                  EA ~ BMI + smoking
                  BMI ~~ smoking"
```

Michel's preregistration



Fit statistics I

- CFI values theoretically range from 0 to 1, with higher values indicating good fit. CFI values of .90 and above are typically considered acceptable fit, and values of .95 and above are typically considered good model fit.

- SRMR values below .10 indicate acceptable fit, values less than .05 indicate good fit, and a value of 0 indicates perfect fit. It is positively-biased, with larger bias resulting when the contributing univariate GWAS samples are lower powered.

Bonus practical for on the flight back

- The folder: michel/2019/practical_2 contains munged sumstats from 6 neuroticism items in UK biobank.
- There is a script to read the munged sumstats into R, and run exploratory factor model.
- The script further contains the code to run a 2 factor model,
- You can write a script to fit a single factor model, and compare the fit.

- Which model is ”better”?