Multivariate Modeling

Analysis of Twin Data

Jacob B. Hjelmborg

Dept. of Epidemiology and Biostatistics, SDU

Spring 2018

University of Southern Denmark

Overview

1. Aims of multivariate twin analyses
2. Examples
3. Classic multivariate models
4. Appendix: SEM- Structural Equation Models
   - Linear regression
   - Factor models
   - Clustered data
   - Structural equation model

Prologue

Effect?

Exposure→Outcome

- Outcome: There are multiple outcomes! (eg. Telomere length, HDL, and BMI).
- What is the contribution of genetic and environmental factors to the variation in outcome?

$$\begin{cases} Y = \text{Genes} + \text{Environment} \\ \Sigma_Y = \Sigma_{\text{Genes}} + \Sigma_{\text{Environment}} \end{cases}$$

- What kind of genetic and environmental influences to expect?
- Are the same or different genes influencing the traits?

Aims of multivariate twin analyses

Scope of study

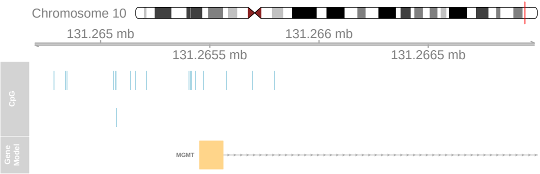
- Co-occurrence or co-morbidity of different diseases.
- Inter-relations, interactions, confounding and moderation effects.
- Genetic or environmental overlap between traits, that is, origin of comorbidity
  - pleiotropic genetic effects
  - environmental overlap: prevention strategies impacting on multiple diseases.
- Developmental changes (longitudinal data).

Overview

1. Aims of multivariate twin analyses
2. Examples
3. Classic multivariate models
4. Appendix: SEM- Structural Equation Models
   - Linear regression
   - Factor models
   - Clustered data
   - Structural equation model

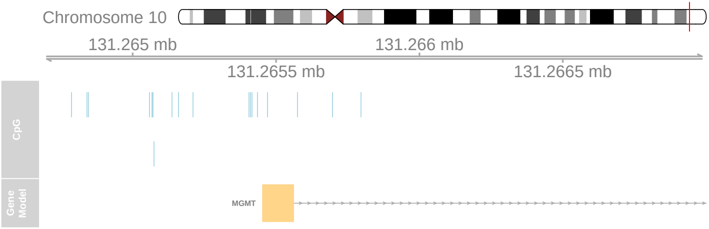
Example: MGMT methylation

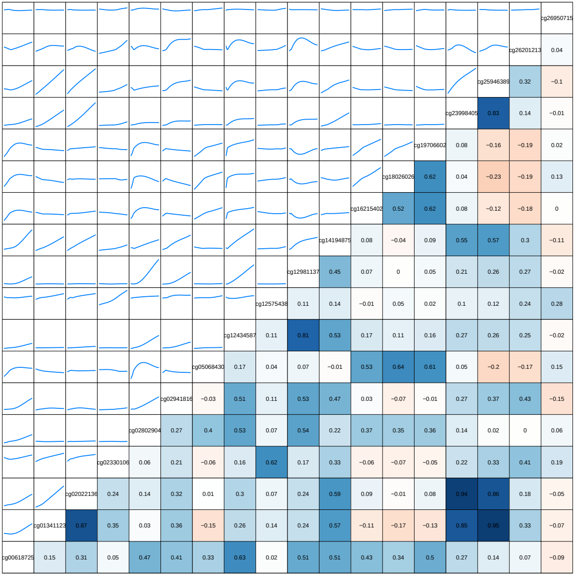
- Can we regulate the DNA repair gene; MGMT?
- -analogy in 5aza-cytidine treatment for acute leukemia
- Epigenetics: Key in evolutionary dynamics of cancer.



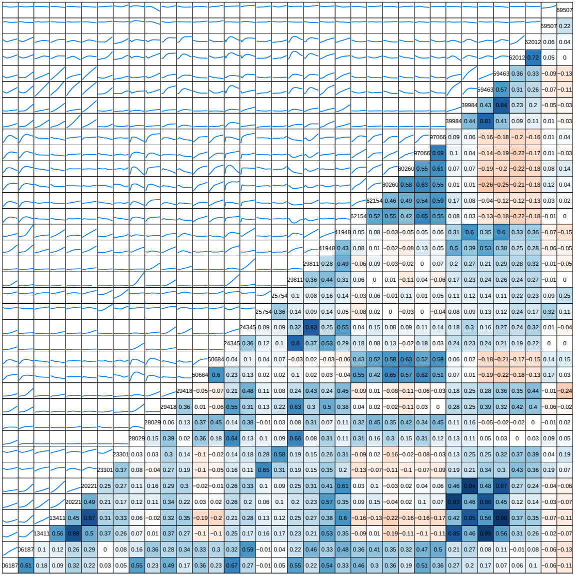
Example: MGMT methylation

- We consider 18 CpG sites at chromosome 10
- -controls the DNA repair gene; MGMT
- How are the 18 sites correlated?
- How are they correlated within MZ pairs? - tells us maximum genetic influence.





cg26950715
cg26201213 0.04
cg25946389 0.32 −0.1
cg23998405 0.83 0.14 −0.01
cg19706602 0.08 −0.16 −0.19 0.02
cg18026026 0.62 0.04 −0.23 −0.19 0.13
cg16215402 0.52 0.62 0.08 −0.12 −0.18 0
cg14194875 0.08 −0.04 0.09 0.55 0.57 0.3 −0.11
cg12981137 0.45 0.07 0 0.05 0.21 0.26 0.27 −0.02
cg12575438 0.11 0.14 −0.01 0.05 0.02 0.1 0.12 0.24 0.28
cg12434587 0.11 0.81 0.53 0.17 0.11 0.16 0.27 0.26 0.25 −0.02
cg05068430 0.17 0.04 0.07 −0.01 0.53 0.64 0.61 0.05 −0.2 −0.17 0.15
cg02941816 −0.03 0.51 0.11 0.53 0.47 0.03 −0.07 −0.01 0.27 0.37 0.43 −0.15
cg02802904 0.27 0.4 0.53 0.07 0.54 0.22 0.37 0.35 0.36 0.14 0.02 0 0.06
cg02330106 0.06 0.21 −0.06 0.16 0.62 0.17 0.33 −0.06 −0.07 −0.05 0.22 0.33 0.41 0.19
cg02022136 0.24 0.14 0.32 0.01 0.3 0.07 0.24 0.59 0.09 −0.01 0.08 0.94 0.86 0.18 −0.05
cg01341123 0.87 0.35 0.03 0.36 −0.15 0.26 0.14 0.24 0.57 −0.11 −0.17 −0.13 0.85 0.95 0.33 −0.07
cg00618725 0.15 0.31 0.05 0.47 0.41 0.33 0.63 0.02 0.51 0.51 0.43 0.34 0.5 0.27 0.14 0.07 −0.09


69507
69507 0.22
62012 0.06 0.04
62012 0.72 0.05 0
59463 0.36 0.33 −0.09 −0.13
59463 0.57 0.31 0.26 −0.07 −0.11
39984 0.43 0.84 0.23 0.2 −0.05 −0.03
39984 0.44 0.81 0.41 0.09 0.11 0.01 −0.03
97066 0.09 0.06 −0.16 −0.18 −0.2 −0.16 0.01 0.04
97066 0.69 0.1 0.04 −0.14 −0.19 −0.22 −0.17 0.01 −0.03
80260 0.55 0.61 0.07 0.07 −0.19 −0.2 −0.22 −0.18 0.08 0.14
80260 0.58 0.63 0.55 0.01 0.01 −0.26 −0.25 −0.21 −0.18 0.12 0.04
62154 0.46 0.49 0.54 0.59 0.17 0.08 −0.04 −0.12 −0.12 −0.13 0.03 0.02
62154 0.52 0.55 0.42 0.65 0.55 0.08 0.03 −0.13 −0.18 −0.22 −0.18 −0.01 0
41948 0.05 0.08 −0.03 −0.05 0.05 0.06 0.31 0.6 0.35 0.6 0.33 0.36 −0.07 −0.15
41948 0.43 0.08 0.01 −0.02 −0.08 0.13 0.05 0.5 0.39 0.53 0.38 0.25 0.28 −0.06 −0.05
29811 0.28 0.49 −0.06 0.09 −0.03 −0.02 0 0.07 0.2 0.27 0.21 0.29 0.28 0.32 −0.01 −0.05
29811 0.36 0.44 0.31 0.06 0 0.01 −0.11 0.04 −0.06 0.17 0.23 0.24 0.26 0.24 0.27 −0.01 0
25754 0.1 0.08 0.16 0.14 −0.03 0.06 −0.01 0.11 0.01 0.05 0.11 0.12 0.14 0.11 0.22 0.23 0.09 0.25
25754 0.36 0.14 0.09 0.14 0.05 −0.08 0.02 0 −0.03 0 −0.04 0.08 0.09 0.13 0.12 0.24 0.17 0.32 0.11
24345 0.09 0.09 0.32 0.83 0.25 0.55 0.04 0.15 0.08 0.09 0.11 0.14 0.18 0.3 0.16 0.27 0.24 0.32 0.01 −0.04
24345 0.36 0.12 0.1 0.8 0.37 0.53 0.29 0.18 0.08 0.13 −0.02 0.18 0.03 0.24 0.23 0.24 0.21 0.19 0.22 0 0
50684 0.04 0.1 0.04 0.07 −0.03 0.02 −0.03 −0.06 0.43 0.52 0.58 0.63 0.52 0.59 0.06 0.02 −0.18 −0.21 −0.17 −0.15 0.14 0.15
50684 0.6 0.23 0.13 0.02 0.02 0.1 0.02 0.03 −0.04 0.55 0.42 0.65 0.57 0.62 0.51 0.07 0.01 −0.19 −0.22 −0.18 −0.13 0.17 0.03
29418 −0.05 −0.07 0.21 0.48 0.11 0.08 0.24 0.43 0.24 0.45 −0.09 0.01 −0.08 −0.11 −0.06 −0.03 0.18 0.25 0.28 0.36 0.35 0.44 −0.01 −0.24
29418 0.36 0.01 −0.06 0.55 0.31 0.13 0.22 0.63 0.3 0.5 0.38 0.04 0.02 −0.02 −0.11 0.03 0 0.28 0.25 0.39 0.32 0.42 0.4 −0.06 −0.02
28029 0.06 0.13 0.37 0.45 0.14 0.38 −0.01 0.03 0.08 0.31 0.07 0.11 0.32 0.45 0.35 0.42 0.34 0.45 0.11 0.16 −0.05 −0.02 −0.02 0 −0.01 0.02
28029 0.15 0.39 0.02 0.36 0.18 0.64 0.13 0.1 0.09 0.66 0.08 0.31 0.11 0.31 0.16 0.3 0.15 0.31 0.12 0.13 0.11 0.05 0.03 0 0.03 0.09 0.05
23301 0.03 0.03 0.3 0.14 −0.1 −0.02 0.14 0.18 0.28 0.58 0.19 0.15 0.26 0.31 −0.09 0.02 −0.16 −0.02 −0.08 −0.03 0.13 0.25 0.25 0.32 0.37 0.39 0.04 0.19
23301 0.37 0.08 −0.04 0.27 0.19 −0.1 −0.05 0.16 0.11 0.65 0.31 0.19 0.15 0.35 0.2 −0.13 −0.07 −0.11 −0.1 −0.07 −0.09 0.19 0.21 0.34 0.3 0.43 0.36 0.19 0.07
20221 0.25 0.27 0.11 0.16 0.29 0.3 −0.02 −0.01 0.26 0.33 0.1 0.09 0.25 0.31 0.41 0.61 0.03 0.1 −0.03 0.02 0.04 0.06 0.46 0.94 0.48 0.87 0.27 0.24 −0.04 −0.06
20221 0.49 0.21 0.17 0.12 0.11 0.34 0.22 0.03 0.02 0.26 0.2 0.06 0.1 0.2 0.23 0.57 0.35 0.09 0.15 −0.04 0.02 0.1 0.07 0.93 0.46 0.86 0.45 0.12 0.14 −0.03 −0.07
13411 0.45 0.87 0.31 0.33 0.06 −0.02 0.32 0.35 −0.19 −0.2 0.21 0.28 0.13 0.12 0.25 0.27 0.38 0.6 −0.16 −0.13 −0.22 −0.16 −0.16 −0.17 0.42 0.85 0.56 0.96 0.37 0.35 −0.07 −0.11
13411 0.56 0.88 0.5 0.37 0.26 0.07 0.01 0.37 0.27 −0.1 −0.1 0.25 0.17 0.16 0.17 0.23 0.21 0.53 0.35 −0.09 0.01 −0.19 −0.11 −0.1 −0.11 0.85 0.46 0.95 0.56 0.31 0.26 −0.02 −0.07
06187 0.1 0.12 0.26 0.29 0 0.08 0.16 0.36 0.28 0.34 0.33 0.3 0.32 0.59 −0.01 0.04 0.22 0.46 0.33 0.48 0.36 0.41 0.35 0.32 0.47 0.5 0.21 0.27 0.08 0.11 −0.01 0.08 −0.06 −0.13
06187 0.61 0.18 0.09 0.32 0.22 0.03 0.05 0.55 0.23 0.49 0.17 0.36 0.23 0.67 0.27 −0.01 0.05 0.55 0.22 0.54 0.33 0.46 0.3 0.36 0.19 0.51 0.36 0.27 0.2 0.17 0.07 0.06 0.1 −0.06 −0.11

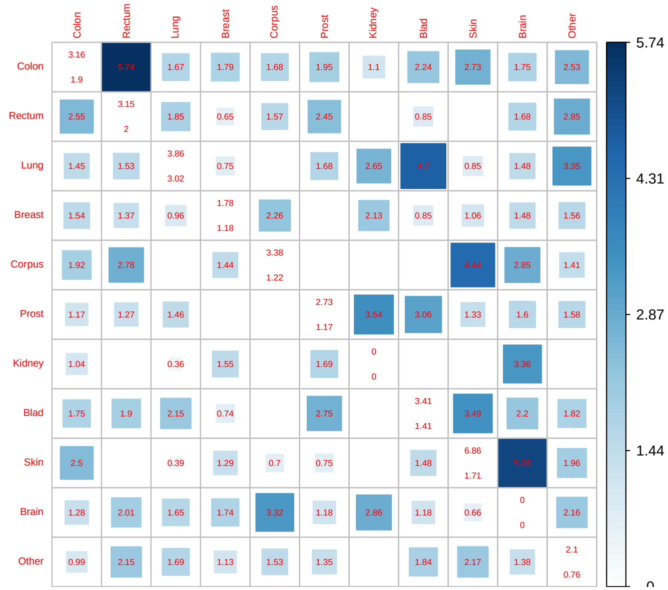
Example: Genetic relatedness of cancer sites

Brain and CNS cancer

- Brain and CNS cancer per se: Less indication of genetic causes like somatic mutations.
- Brain and CNS cancers show genetic relatedness with certain cancer loci.

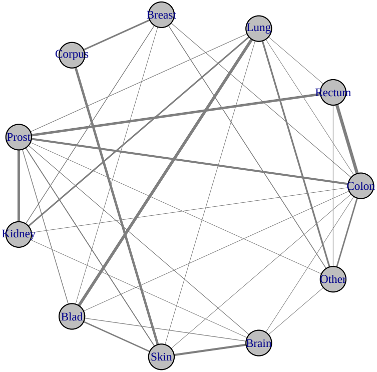
- Beh Genet 2015 Estimating Twin Pair Concordance for Age of Onset. Scheike, Hjelmborg and Holst K.
- JAMA 2016 Familial Risk and Heritability of Cancer. Mucci, Hjelmborg, Harris, Kaprio et al.

NorTwinCan
Nordic Twin Study of Cancer

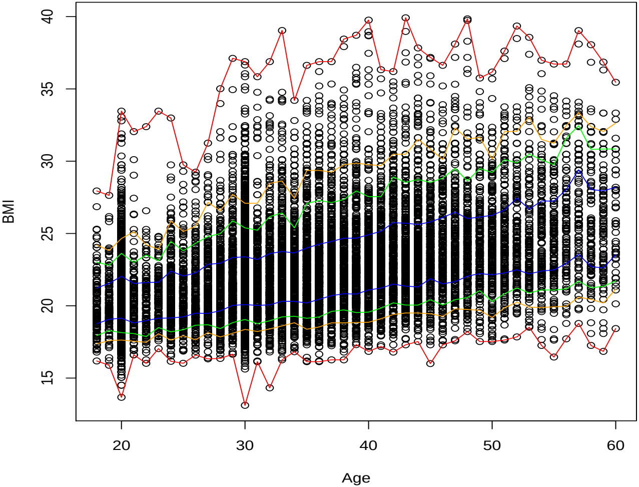


	Colon	Rectum	Lung	Breast	Corpus	Prost	Kidney	Blad	Skin	Brain	Other
Colon	3.16 1.9	5.74	1.67	1.79	1.68	1.95	1.1	2.24	2.73	1.75	2.53
Rectum	2.55	3.15 2	1.85	0.65	1.57	2.45		0.85		1.68	2.85
Lung	1.45	1.53	3.86 3.02	0.75		1.68	2.65	4.7	0.85	1.48	3.35
Breast	1.54	1.37	0.96	1.78 1.18	2.26		2.13	0.85	1.06	1.48	1.56
Corpus	1.92	2.78		1.44	3.38 1.22				4.44	2.85	1.41
Prost	1.17	1.27	1.46			2.73 1.17	3.54	3.06	1.33	1.6	1.58
Kidney	1.04		0.36	1.55		1.69	0 0			3.36	
Blad	1.75	1.9	2.15	0.74		2.75		3.41 1.41	3.49	2.2	1.82
Skin	2.5		0.39	1.29	0.7	0.75		1.48	6.86 1.71	5.28	1.96
Brain	1.28	2.01	1.65	1.74	3.32	1.18	2.86	1.18	0.66	0 0	2.16
Other	0.99	2.15	1.69	1.13	1.53	1.35		1.84	2.17	1.38	2.1 0.76

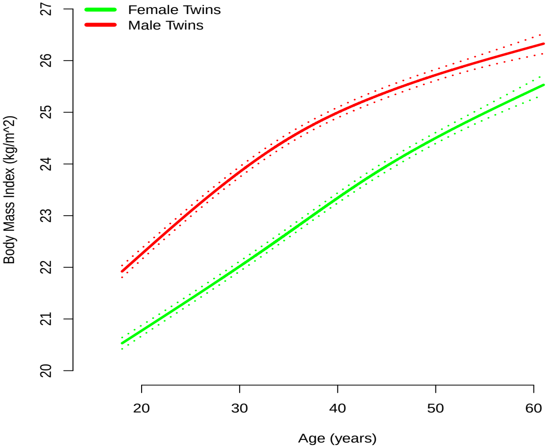
Heritability



Example: Seven waves of BMI measurements



BMI by age (fitted by a 'gamm model')



```r
plot(bmi_1_1 ~ bmi_1_2, data=mzData)
plot(bmi_1_1 ~ bmi_1_2, data=dzData)
```

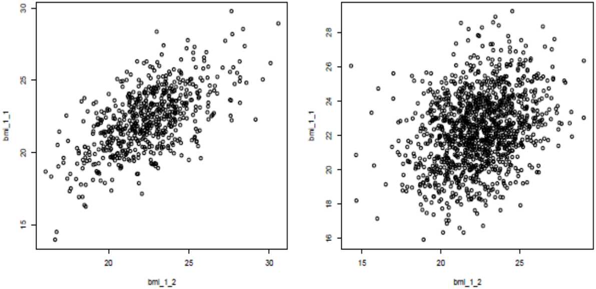


Figure: Male BMI versus co twin BMI for MZ and DZ pairs at first wave

Overview

1. Aims of multivariate twin analyses
2. Examples
3. Classic multivariate models
4. Appendix: SEM- Structural Equation Models
   - Linear regression
   - Factor models
   - Clustered data
   - Structural equation model

UNIVERSITY OF SOUTHERN DENMARK

Outline - Multiple phenotypes measured in twins

Aims of Multivariate Analysis

- Structural Equation Modeling
  - the full multivariate ACE model
  - the independent pathway model
  - the common pathway model
  - the growth curve model
  - the direction of causation model
- Example: BMI in Finnish adult twins

Modelling

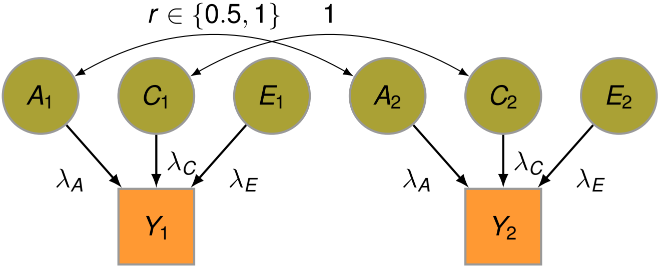
Univariate → multivariate

- What is the contribution of genetic and environmental factors to the variation in several outcomes?

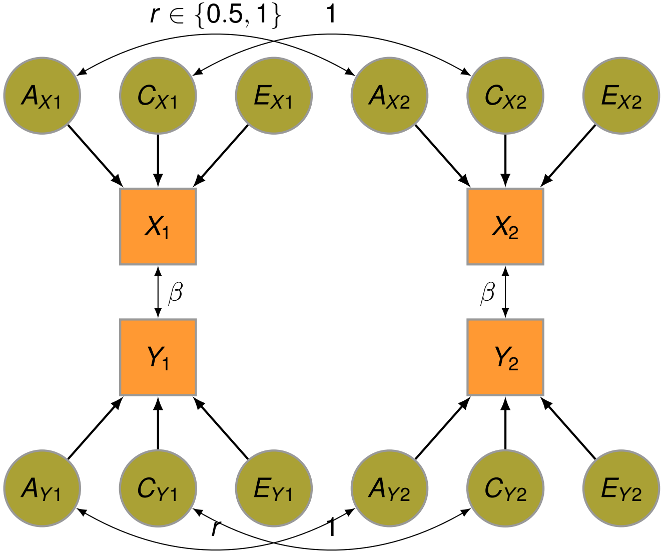
$$\begin{cases} Y = \text{Genes} + \text{Environment} \\ \Sigma_Y = \Sigma_{\text{Genes}} + \Sigma_{\text{Environment}} \end{cases}$$

- What kind of genetic and environmental influences to expect?
- Are the same or different genes influencing the traits?
- The univariate models seen so far are generalized.
- Structural equation models, SEM's, are briefly introduced in the Appendix.

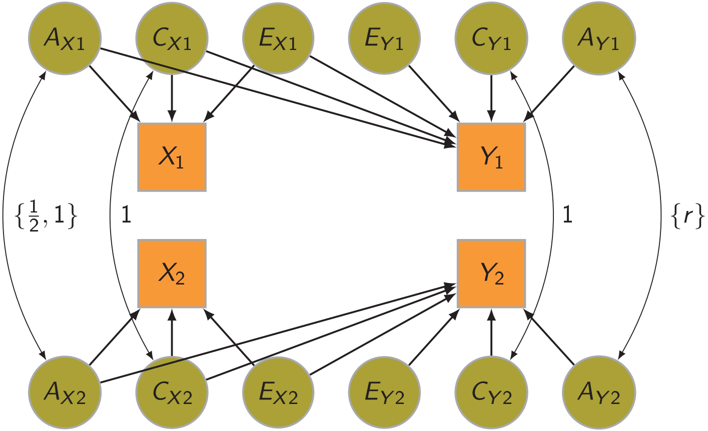
SEM - Univariate ACE Path Diagram representation



SEM - Path Diagram representation



SEM - the full multivariate ACE model



Biometric analyses - polygenic model

- Contributing factors to the variation in outcome:

$$\Sigma_Y = \begin{pmatrix} \Sigma_A & r\Sigma_A \\ r\Sigma_A & \Sigma_A \end{pmatrix} + \begin{pmatrix} \Sigma_C & \Sigma_C \\ \Sigma_C & \Sigma_C \end{pmatrix} + \begin{pmatrix} \Sigma_E & 0 \\ 0 & \Sigma_E \end{pmatrix}$$

where $r = 1$ for MZ pairs and $z = \frac{1}{2}$ for DZ pairs.

In particular, we obtain

- Heritability:

$$h_Y^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_C^2 + \sigma_E^2}$$

- Shared environmental effect:

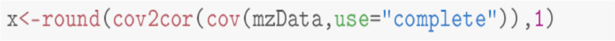
$$c_Y^2 = \frac{\sigma_C^2}{\sigma_A^2 + \sigma_C^2 + \sigma_E^2}$$

Biometric analyses - polygenic model

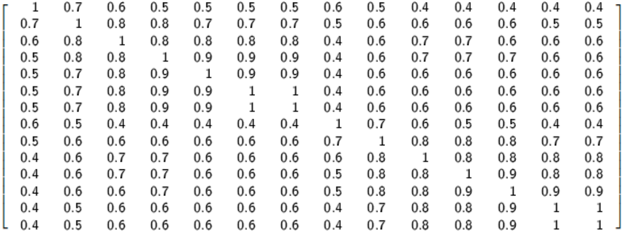
Main assumptions

- Equal environments assumption for MZ and DZ twins.
- No gene-environment interaction and correlation.
- No gene-gene interaction (link: epistasis).
- Equal mean and variance of twin 1 and twin 2, MZ and DZ.
- Estimation and inference by maximum likelihood principle assuming bivariate normality of paired observations (as before).

Correlation matrix of seven waves in MZ pairs

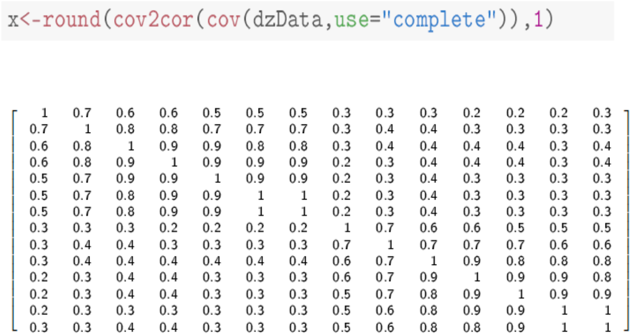


```
x<-round(cov2cor(cov(mzData,use="complete")),1)
```



$$\begin{bmatrix}
1 & 0.7 & 0.6 & 0.5 & 0.5 & 0.5 & 0.5 & 0.6 & 0.5 & 0.4 & 0.4 & 0.4 & 0.4 & 0.4 \\
0.7 & 1 & 0.8 & 0.8 & 0.7 & 0.7 & 0.7 & 0.5 & 0.6 & 0.6 & 0.6 & 0.6 & 0.5 & 0.5 \\
0.6 & 0.8 & 1 & 0.8 & 0.8 & 0.8 & 0.8 & 0.4 & 0.6 & 0.7 & 0.7 & 0.6 & 0.6 & 0.6 \\
0.5 & 0.8 & 0.8 & 1 & 0.9 & 0.9 & 0.9 & 0.4 & 0.6 & 0.7 & 0.7 & 0.7 & 0.6 & 0.6 \\
0.5 & 0.7 & 0.8 & 0.9 & 1 & 0.9 & 0.9 & 0.4 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 \\
0.5 & 0.7 & 0.8 & 0.9 & 0.9 & 1 & 1 & 0.4 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 \\
0.5 & 0.7 & 0.8 & 0.9 & 0.9 & 1 & 1 & 0.4 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 \\
0.6 & 0.5 & 0.4 & 0.4 & 0.4 & 0.4 & 0.4 & 1 & 0.7 & 0.6 & 0.5 & 0.5 & 0.4 & 0.4 \\
0.5 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 0.7 & 1 & 0.8 & 0.8 & 0.8 & 0.7 & 0.7 \\
0.4 & 0.6 & 0.7 & 0.7 & 0.6 & 0.6 & 0.6 & 0.6 & 0.8 & 1 & 0.8 & 0.8 & 0.8 & 0.8 \\
0.4 & 0.6 & 0.7 & 0.7 & 0.6 & 0.6 & 0.6 & 0.5 & 0.8 & 0.8 & 1 & 0.9 & 0.8 & 0.8 \\
0.4 & 0.6 & 0.6 & 0.7 & 0.6 & 0.6 & 0.6 & 0.5 & 0.8 & 0.8 & 0.9 & 1 & 0.9 & 0.9 \\
0.4 & 0.5 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 0.4 & 0.7 & 0.8 & 0.8 & 0.9 & 1 & 1 \\
0.4 & 0.5 & 0.6 & 0.6 & 0.6 & 0.6 & 0.6 & 0.4 & 0.7 & 0.8 & 0.8 & 0.9 & 1 & 1
\end{bmatrix}$$

Correlation matrix of seven waves in DZ pairs



```r
x<-round(cov2cor(cov(dzData,use="complete")),1)
```

$$\begin{bmatrix}
1 & 0.7 & 0.6 & 0.6 & 0.5 & 0.5 & 0.5 & 0.3 & 0.3 & 0.3 & 0.2 & 0.2 & 0.2 & 0.3 \\
0.7 & 1 & 0.8 & 0.8 & 0.7 & 0.7 & 0.7 & 0.3 & 0.4 & 0.4 & 0.3 & 0.3 & 0.3 & 0.3 \\
0.6 & 0.8 & 1 & 0.9 & 0.9 & 0.8 & 0.8 & 0.3 & 0.4 & 0.4 & 0.4 & 0.4 & 0.3 & 0.4 \\
0.6 & 0.8 & 0.9 & 1 & 0.9 & 0.9 & 0.9 & 0.2 & 0.3 & 0.4 & 0.4 & 0.4 & 0.3 & 0.4 \\
0.5 & 0.7 & 0.9 & 0.9 & 1 & 0.9 & 0.9 & 0.2 & 0.3 & 0.4 & 0.3 & 0.3 & 0.3 & 0.3 \\
0.5 & 0.7 & 0.8 & 0.9 & 0.9 & 1 & 1 & 0.2 & 0.3 & 0.4 & 0.3 & 0.3 & 0.3 & 0.3 \\
0.5 & 0.7 & 0.8 & 0.9 & 0.9 & 1 & 1 & 0.2 & 0.3 & 0.4 & 0.3 & 0.3 & 0.3 & 0.3 \\
0.3 & 0.3 & 0.3 & 0.2 & 0.2 & 0.2 & 0.2 & 1 & 0.7 & 0.6 & 0.6 & 0.5 & 0.5 & 0.5 \\
0.3 & 0.4 & 0.4 & 0.3 & 0.3 & 0.3 & 0.3 & 0.7 & 1 & 0.7 & 0.7 & 0.7 & 0.6 & 0.6 \\
0.3 & 0.4 & 0.4 & 0.4 & 0.4 & 0.4 & 0.4 & 0.6 & 0.7 & 1 & 0.9 & 0.8 & 0.8 & 0.8 \\
0.2 & 0.3 & 0.4 & 0.4 & 0.3 & 0.3 & 0.3 & 0.6 & 0.7 & 0.9 & 1 & 0.9 & 0.9 & 0.8 \\
0.2 & 0.3 & 0.4 & 0.4 & 0.3 & 0.3 & 0.3 & 0.5 & 0.7 & 0.8 & 0.9 & 1 & 0.9 & 0.9 \\
0.2 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 & 0.5 & 0.6 & 0.8 & 0.9 & 0.9 & 1 & 1 \\
0.3 & 0.3 & 0.4 & 0.4 & 0.3 & 0.3 & 0.3 & 0.5 & 0.6 & 0.8 & 0.8 & 0.9 & 1 & 1
\end{bmatrix}$$

Multivariate ACE model fitted to seven waves

```
mxCompare(CholAceFit,CholAeFit)  # C seems important.

##       base comparison  ep minus2LL    df      AIC   diffLL diffdf           p
## 1 CholACE       <NA> 119 91335.02 25641 40053.02       NA     NA          NA
## 2 CholACE     CholAE  91 91389.54 25669 40051.54 54.51227     28 0.001945057
```

- Common environmental effects seems important.
- Further model selection and check of assumptions in Practicals to follow.

Multivariate ACE model fitted to seven waves

```
x<-round(CholAceFit$H2$result,2)
```

$$\Sigma_{H^2} = \begin{bmatrix} 0.48 & 0.54 & 0.51 & 0.64 & 0.6 & 0.66 & 0.65 \\ 0.54 & 0.44 & 0.54 & 0.63 & 0.58 & 0.63 & 0.64 \\ 0.51 & 0.54 & 0.48 & 0.62 & 0.6 & 0.62 & 0.63 \\ 0.64 & 0.63 & 0.62 & 0.6 & 0.61 & 0.63 & 0.63 \\ 0.6 & 0.58 & 0.6 & 0.61 & 0.52 & 0.58 & 0.59 \\ 0.66 & 0.63 & 0.62 & 0.63 & 0.58 & 0.52 & 0.55 \\ 0.65 & 0.64 & 0.63 & 0.63 & 0.59 & 0.55 & 0.53 \end{bmatrix}$$

- Heritabilities of seven waves along diagonal.
- Bivariate heritabilities off the diagonal. It is 0.65 between wave 1 and 7, hence 65 percent of phenotypic correlation is mediated by shared genetic influence.

University of Southern Denmark

Multivariate ACE model fitted to seven waves

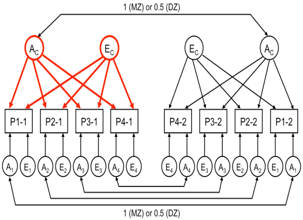
```
x<-round(cov2cor(CholAceFit$A$result),2)
```

$$\Sigma_{\mathrm{Corr}_A} = \begin{bmatrix} 1 & 0.81 & 0.62 & 0.65 & 0.63 & 0.62 & 0.62 \\ 0.81 & 1 & 0.88 & 0.92 & 0.89 & 0.86 & 0.86 \\ 0.62 & 0.88 & 1 & 1 & 1 & 0.97 & 0.97 \\ 0.65 & 0.92 & 1 & 1 & 0.99 & 0.96 & 0.96 \\ 0.63 & 0.89 & 1 & 0.99 & 1 & 0.98 & 0.98 \\ 0.62 & 0.86 & 0.97 & 0.96 & 0.98 & 1 & 1 \\ 0.62 & 0.86 & 0.97 & 0.96 & 0.98 & 1 & 1 \end{bmatrix}$$

- Genetic correlation of seven waves of BMI, that is, correlation of genetic effects regardless of heritability.
- -the likelihood that a gene found to be associated with one trait is also associated with the other trait.

Independent pathway ACE model

- The covariation between traits is caused by genetic and environmental factors common to traits, each having its own paths to each trait

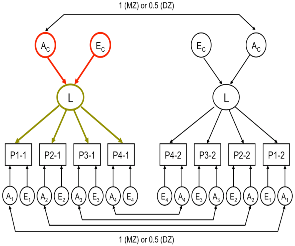


Matrix model representation (one twin)

$$\underset{\text{observed traits}}{\begin{pmatrix}P1\\P2\\P3\\P4\end{pmatrix}} = \overset{\text{common factors coefficients}}{\begin{pmatrix}x_1 & z_1\\x_2 & z_2\\x_3 & z_3\\x_4 & z_4\end{pmatrix}} * \underset{\text{common factors}}{\begin{pmatrix}A_C\\E_C\end{pmatrix}} + \overset{\text{specific genetic coefficients}}{\begin{pmatrix}t_1 & & & \\ & t_2 & & \\ & & t_3 & \\ & & & t_4\end{pmatrix}} * \underset{\text{specific genetic factors}}{\begin{pmatrix}A_1\\A_2\\A_3\\A_4\end{pmatrix}} + \overset{\text{specific unshared environmental coefficients}}{\begin{pmatrix}v_1 & & & \\ & v_2 & & \\ & & v_3 & \\ & & & v_4\end{pmatrix}} * \underset{\text{specific unshared environmental factors}}{\begin{pmatrix}E_1\\E_2\\E_3\\E_4\end{pmatrix}}$$

Common pathway ACE model

- The covariation between traits is caused by a single underlying latent phenotype, that is in turn influenced by genetic and environmental factors



Matrix model representation (one twin)

latent phenotype coefficients

common factors coefficients

specific genetic coefficients

specific unshared environmental coefficients

$$\begin{pmatrix} P1 \\ P2 \\ P3 \\ P4 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix} \cdot \left[\begin{pmatrix} x & z \\ x & z \\ x & z \\ x & z \end{pmatrix} * \begin{pmatrix} A_C \\ E_C \end{pmatrix} \right] + \begin{pmatrix} t_1 & & & \\ & t_2 & & \\ & & t_3 & \\ & & & t_4 \end{pmatrix} * \begin{pmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \end{pmatrix} + \begin{pmatrix} v_1 & & & \\ & v_2 & & \\ & & v_3 & \\ & & & v_4 \end{pmatrix} * \begin{pmatrix} E_1 \\ E_2 \\ E_3 \\ E_4 \end{pmatrix}$$

observed traits

common factors

specific genetic factors

specific unshared environmental factors

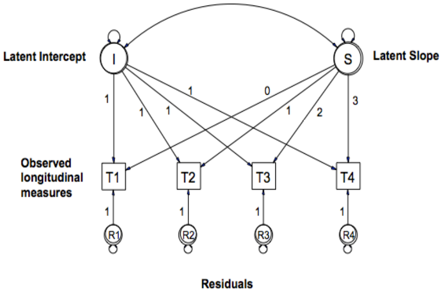
Which model to report?

```
#'  \begin{itemize}
#'\item The multivariate model which has the lowest AIC.
#'\item Choosing among non-nested models can be delicate.
#'\item The CP model can be tested as a sub-model of the

#'$\textrm{nparIP}-\textrm{nparCP}=(\textrm{nfac}-1)*(\te

#'
#'\item If $\chi^2$-test is not significant, the CP shoul



#'\end{itemize}
```

The Growth Curve Model for Longitudinal Data

- The multivariate models above allows for
  - Magnitude of genetic and environmental influences at each time point (wave).
  - Extent to which genetic and environmental influences overlap across time points.
- Focus now on growth variables, eg. initial level (intercept) and rate of change (slope) - to predict level at a series of time points.

The Growth Curve model



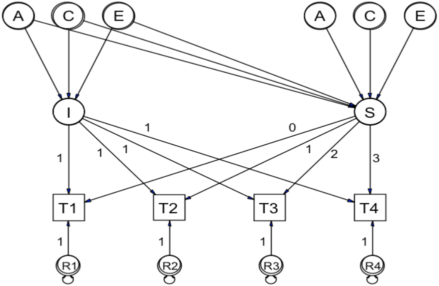
- The growth curve model allows for
  - modeling any number of time points (waves).
  - modeling the individual trajectory.

The Growth Curve model

Aims for the linear growth curve model

- Are there inter-individual differences in initial level and rate of change? - Variance of intercept and slope
- Are initial level and rate of change associated within an individual? - Within-twin correlation between intercept and slope
- Do genetic or environmental factors explain inter-individual differences in initial level and rate of change? - Cross-twin within-trait correlation of intercept and of slope in MZ and DZ twins
- Do genetic or environmental factors explain the within-individual association between initial level and rate of change? - Cross-twin/cross-trait correlation between intercept and slope in MZ and DZ twins
- To what extent are inter-individual differences in each of the longitudinal measures accounted for by initial level and rate of change? - Variance of residuals

The Growth Curve ACE model



- The growth curve ACE model allows for
  - Genetic and environmental influences on initial level and rate of change, and on their mutual interplay modelling any number of time points.
  - Very efficient: number of parameters does not increase with number of measurements.

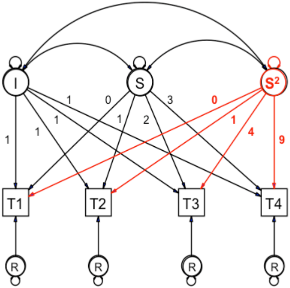
The Growth Curve ACE model

Aims for the linear growth curve ACE model

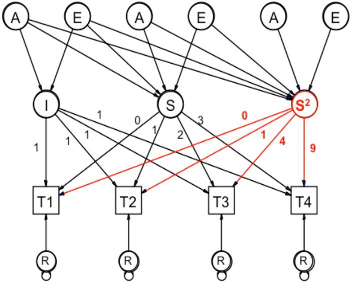
- What is the contribution of genetic factors to inter-individual variation in initial level and rate of change? - Heritability of intercept and slope.
- What is the contribution of environmental factors to inter-individual variation in initial level and rate of change? - Shared and unique environmental proportions of variance of intercept and slope.
- Same or different genes influencing initial level and rate of change? - Genetic correlation between intercept and slope. For BMI: 0.50
- Same or different environments influencing initial level and rate of change? - Shared and unique environmental correlations between intercept and slope

The Growth Curve ACE model - extensions

General



Biometric (AE)



- Easy extension of linear to quadratic model:
  - - loadings of the quadratic factor equal the respective squared loadings of the linear factor.
  - - the quadratic factor covaries with both initial level and linear factor.

Overview

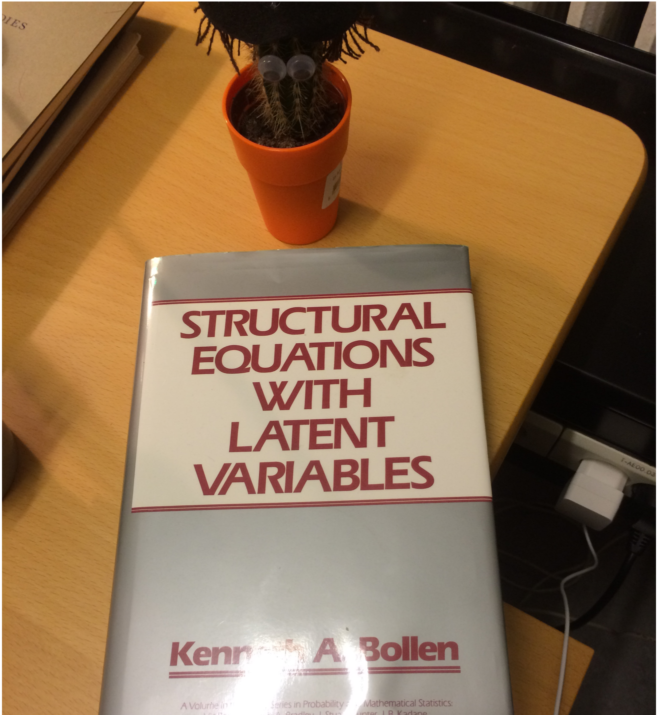
1 Aims of multivariate twin analyses

2 Examples

3 Classic multivariate models

4 Appendix: SEM- Structural Equation Models

- Linear regression
- Factor models
- Clustered data
- Structural equation model


STRUCTURAL EQUATIONS WITH LATENT VARIABLES
Kenneth A. Bollen
A Volume in Wiley Series in Probability and Mathematical Statistics:
A. Bradley, J. Stuart Hunter, J. B. Kadane

CHAPTER ONE

Introduction

Most researchers applying statistics think in terms of modeling the *individual observations*. In multiple regression or ANOVA (analysis of variance), for instance, we learn that the regression coefficients or the error variance estimates derive from the minimization of the sum of squared differences of the predicted and observed dependent variable for each case. Residual analyses display discrepancies between fitted and observed values for every member of the sample.

The methods of this book demand a reorientation. The procedures emphasize *covariances* rather than cases.[1] Instead of minimizing functions of observed and predicted individual values, we minimize the difference between the sample covariances and the covariances predicted by the model. The observed covariances minus the predicted covariances form the residuals. The fundamental hypothesis for these structural equation proce[illegible]s is that the covariance matrix of the observed variables is a function of [illegible] of parameters. If the model were correct and if we knew the [illegible]meters, the population covariance matrix would be exactly reproduced. [illegible]ch of this book is about the equation that formalizes this fundamental hypothesis:

$$\Sigma = \Sigma(\theta) \tag{1.1}$$

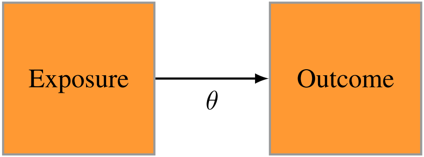
In (1.1), Σ (sigma) is the population covariance matrix of observed variables, θ (theta) is a vector that contains the model parameters, and $\Sigma(\theta)$ is

[1] As is clear from several places in the book, individual cases that are outliers can severely affect covariances and estimates of parameters. Thus, with these techniques, researchers still need to check for outliers. In addition, in many cases (e.g., regression models) the minimizations based on individuals and minimizations based on the predicted and observed covariance matrices lead to the same parameter estimates.

SEM models

- The focus in SEM is the basically the covariance matrix, $\Sigma = \Sigma(\theta)$.
- Great many statistical methods can be formulated via SEM.

A linear regression model



Measuring the effect, θ

- $E(Y|X_1, \ldots, X_p) = \alpha + \beta_1 X_1 + \ldots + \beta_p X_p$ is the regression model, say.
- $E(Y|X_1 = 1) - E(Y|X_1 = 0) = \beta_1$ is the effect of treatment X_1 given covariates $X_2, \ldots, X_p$.

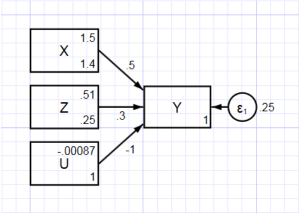
-simulating data

- $E(Y|X,Z,U) = \alpha + \beta_X X + \beta_U U + \beta_Z Z$ is the true model, say.

```r
set.seed(727)
n=10000
U=rnorm(n)
G=rbinom(n,1,0.5)
Z=rbinom(n,1,0.5)
alp0=1;alp1=0.75;alp2=-0.5;alp3=0.3
mu=alp0+alp1*G+alp2*U+alp3*Z
X=rnorm(n,mu,1)

bet0=1;bet1=0.5;bet2=-1;bet3=0.3
thet=bet0+bet1*X+bet2*U+bet3*Z
Y=rnorm(n,thet,0.5)
```

- $E(Y|X,Z,U) = 1 + 0.5X - U + 0.3Z$ is the true model.
- Can we estimate the effect of X? - that is $\beta_X = 0.5$?

-as path-diagram



-in Stata

```
use datLinModel.dta, clear
sum
regress Y X Z U
sem (X -> Y, ) (Z -> Y, ) (U -> Y, ), nocapslatent
```

```
summary(lm(Y~X+Z+U))

##
## Call:
## lm(formula = Y ~ X + Z + U)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77234 -0.33678 -0.00383  0.33905  1.87276
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.003710   0.009549  105.11   <2e-16 ***
## X            0.496471   0.004659  106.56   <2e-16 ***
## Z            0.303819   0.010042   30.25   <2e-16 ***
## U           -1.001372   0.005441 -184.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 0.4966 on 9996 degrees of freed
## Multiple R-squared:  0.8875,Adjusted R-squared:  0.8875
## F-statistic: 2.629e+04 on 3 and 9996 DF,  p-value: < 2.2
```

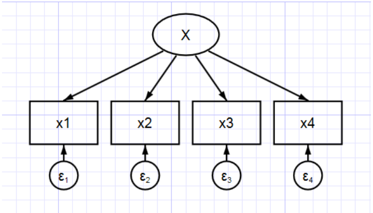
SEM models

- We can estimate the effect of X - that is $\beta_X = 0.5$
- The focus in SEM is basically the covariance matrix, $\Sigma = \Sigma(\theta)$.
- Simple linear regression, $Y = \beta X + \epsilon$, corresponds in SEM to

$$\Sigma_Y = \begin{pmatrix} \beta^2\sigma_X^2 + \sigma_\epsilon^2 & \beta\sigma_X^2 \\ \beta\sigma_X^2 & \sigma_X^2 \end{pmatrix}$$

- -now, find parameters $\beta \in \theta$ minimizing the difference between the sample covariance matrix and the one predicted by the model, the right-side.
- Statistical regression models can be formulated via SEM.

A single factor model



Aims

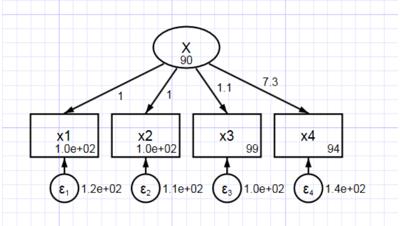
- -is there a factor common to the observables?
- -an underlying latent disease?
- -a latent feature explaining eg. the questionaire outcome?

-simulating data

- $x_j = \alpha_j + \beta_j X + \epsilon_j$, $j = 1, 2, 3, 4$, is the true model (with contraints for identifiability).

```
set seed 83216
set obs 500
gen X = round(rnormal(0,10))
gen x1 = round(100 + X + rnormal(0, 10))
gen x2 = round(100 + X + rnormal(0, 10))
gen x3 = round(100 + X + rnormal(0, 10))
gen x4 = round(100 + 7*X + rnormal(0, 10))
drop X
```

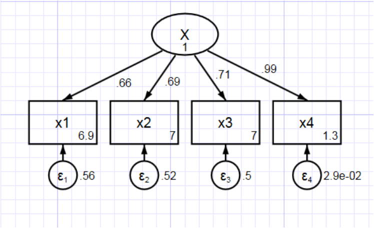
-estimated path-diagram



-in Stata

```
webgetsem sem_1fmm
use http://www.stata-press.com/data/r15/sem_1fmm
summarize
sem (x1 x2 x3 x4 <- X)
```

-standardized estimates



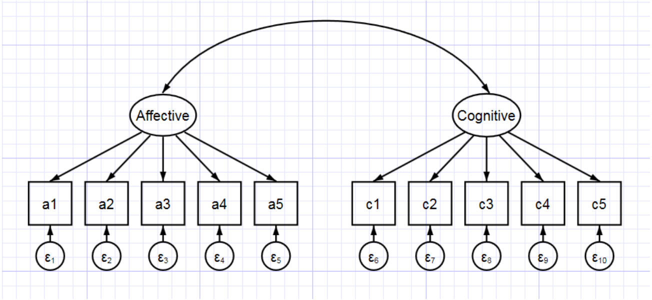
Results

- -is there a factor common to the observables?
- -amount of variation explained by X.
- Goodness of fit: No significant deviation from saturated model, $\chi^2_2 = 1.48$, $p = 0.48$

-submodels: constraining effects for x_2 and x_3

```
sem (X -> x1, ) (X -> x2, ) (X -> x3, ) (X -> x4, )
estimates store modelA
sem (X -> x1, ) (X -> x2@myb, ) (X -> x3@myb, ) (X -> x4, )
estimates store modelB
lrtest modelA modelB
```

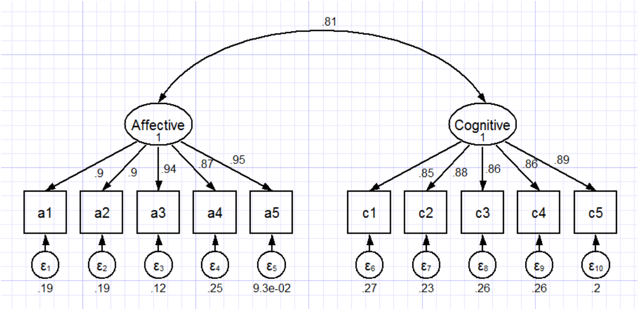
A two factor model



Aims

- cognitive and affective arousal: Test in children (Visual Similes Test II).
- -are there intrinsic factors common to the observables?
- -underlying latent features?

-standardized estimates



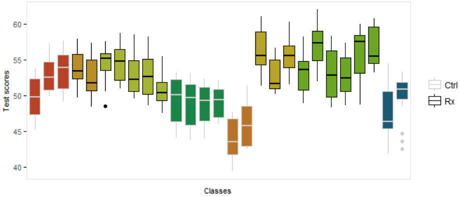
Results

- -two factors, affective and cognition are common to the observables.
- -amount of variation explained by these factors.
- -correlation of latent factors, $r = 0.81$.
- Goodness of fit: See equation-level fit in exercises.

-in Stata

```
webgetsem sem_2fmm
use http://www.stata-press.com/data/r15/sem_2fmm
sem (Affective -> a1 a2 a3 a4 a5) (Cognitive -> c1 c2 c3 c4
sem, standardized
estat eqgof
estat gof, stats(all)
sem (Affective -> a1 a2 a3 a4 a5) (Cognitive -> c1 c2 c3 c4
cov(Affective*Cognitive@0)
```

A hierarchical model



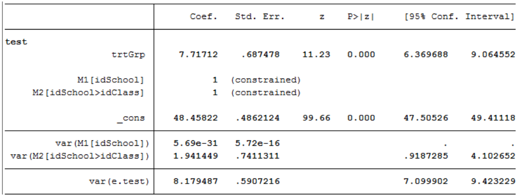
Aims

- -effect of intervention for childrens learning?
- -an individual belongs to a class in a school.
- Data: 403 children in 20 classes within eight schools.
- Test scores for interveened and controls in boxplot by classes by schools.

-in Stata

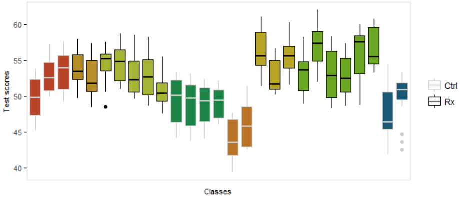
```
use datKiDMmodel.dta, clear
sum
gsem (test <- trtGrp M1[idSchool] M2[idSchool>idClass])
mixed test trtGrp || idSchool: || idClass:, var mle
```

Stata outcome



```
------------------------------------------------------------------------------------------
                             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------------------+------------------------------------------------------------
test                         |
                      trtGrp |    7.71712    .687478    11.23   0.000     6.369688    9.064552
                             |
                M1[idSchool] |          1  (constrained)
       M2[idSchool>idClass]  |          1  (constrained)
                             |
                       _cons |   48.45822   .4862124    99.66   0.000     47.50526    49.41118
-----------------------------+------------------------------------------------------------
           var(M1[idSchool]) |   5.69e-31   5.72e-16                              .           .
  var(M2[idSchool>idClass])  |   1.941449   .7411311                       .9187285    4.102652
-----------------------------+------------------------------------------------------------
                 var(e.test) |   8.179487   .5907216                       7.099902    9.423229
------------------------------------------------------------------------------------------
```

A hierarchical model



Results

- -effect of intervention for childrens learning: $\beta = 7.71$ points higher if intervention ($p < 0.001$)
- -considerable variation between classes, around 20% $(= 1.94/(1.94 + 8.18))$.

SEM - General form

SEM

- A general SEM is of form, $\boxed{\eta = B\eta + \Gamma\xi + \zeta}$, where η and ξ denotes endogenous and exogenous variables respectively, B and Γ are matrices of coefficients and ζ denotes errors.
- -this induces the covariance matrix, $\Sigma(\theta)$, to be compared with the observed covariance matrix from the sample.
- -indeed the purpose of programs LISREL, Mx, OpenMx, M-Plus, lava, Stata SEM, ...