Analysis of continuous twin data

Univariate analysis

Jacob Hjelmborg

University of Southern Denmark

Spring 2018

Overview

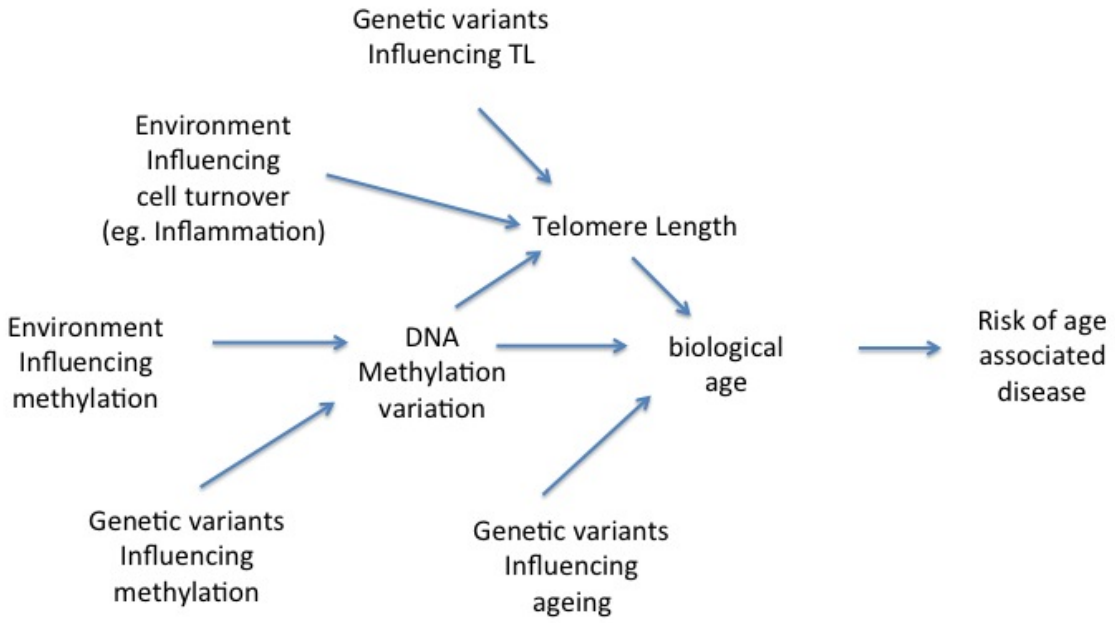
University of Southern Denmark

Prologue

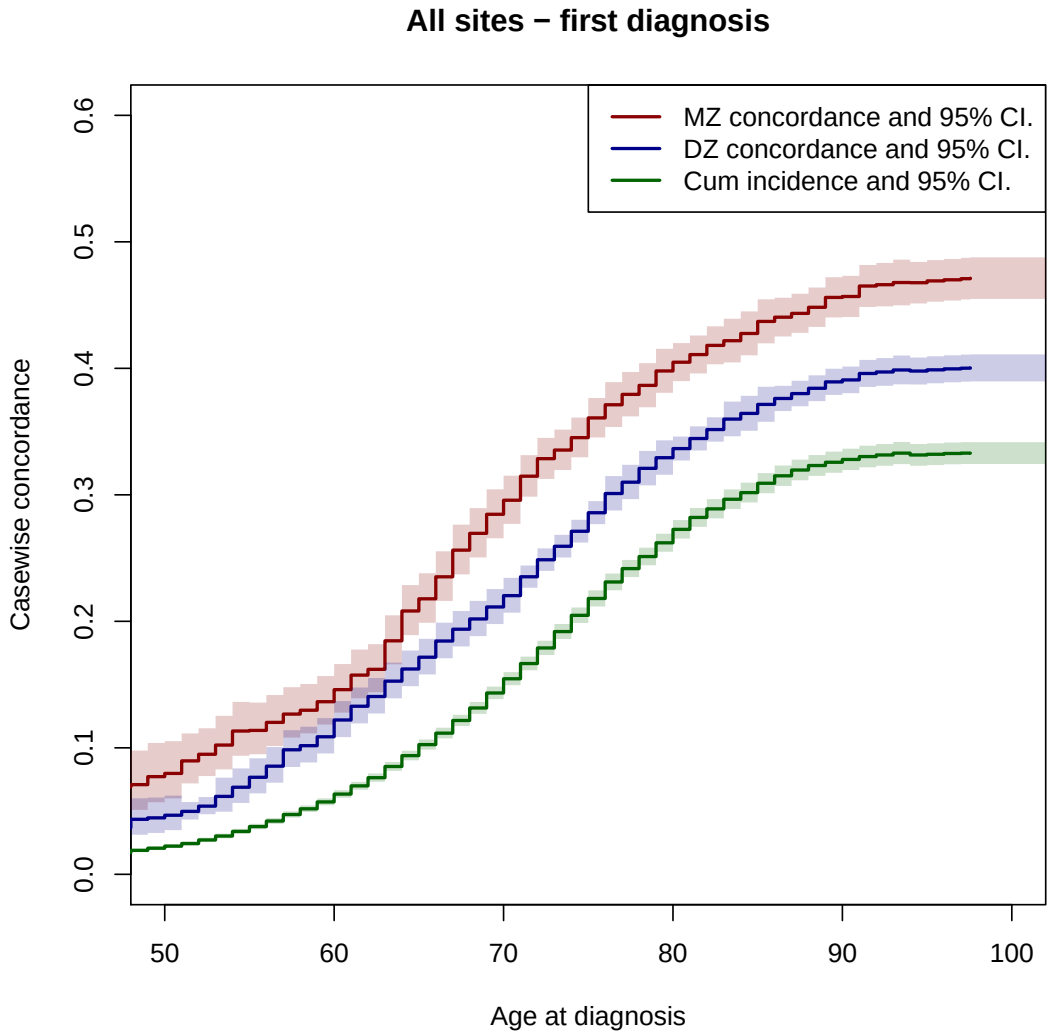
Welcome!

Analysis of Twin Data in Health Science:

- The Course homepage - click here



All Cancer

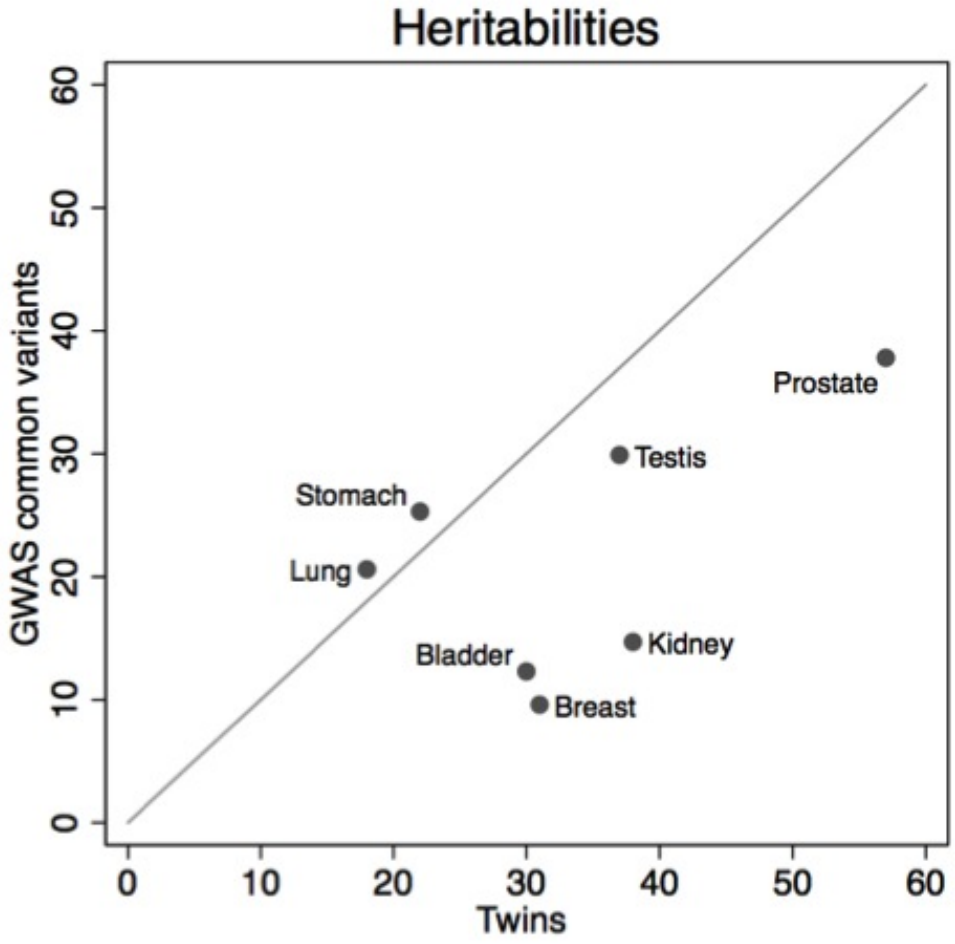


L. Mucci, J. Hjelmborg, T. Scheike, K. Holst, A. Skytthe, H. Adami, N. Holm, K. Christensen, J. Harris, J. Kaprio et al. JAMA (2016)

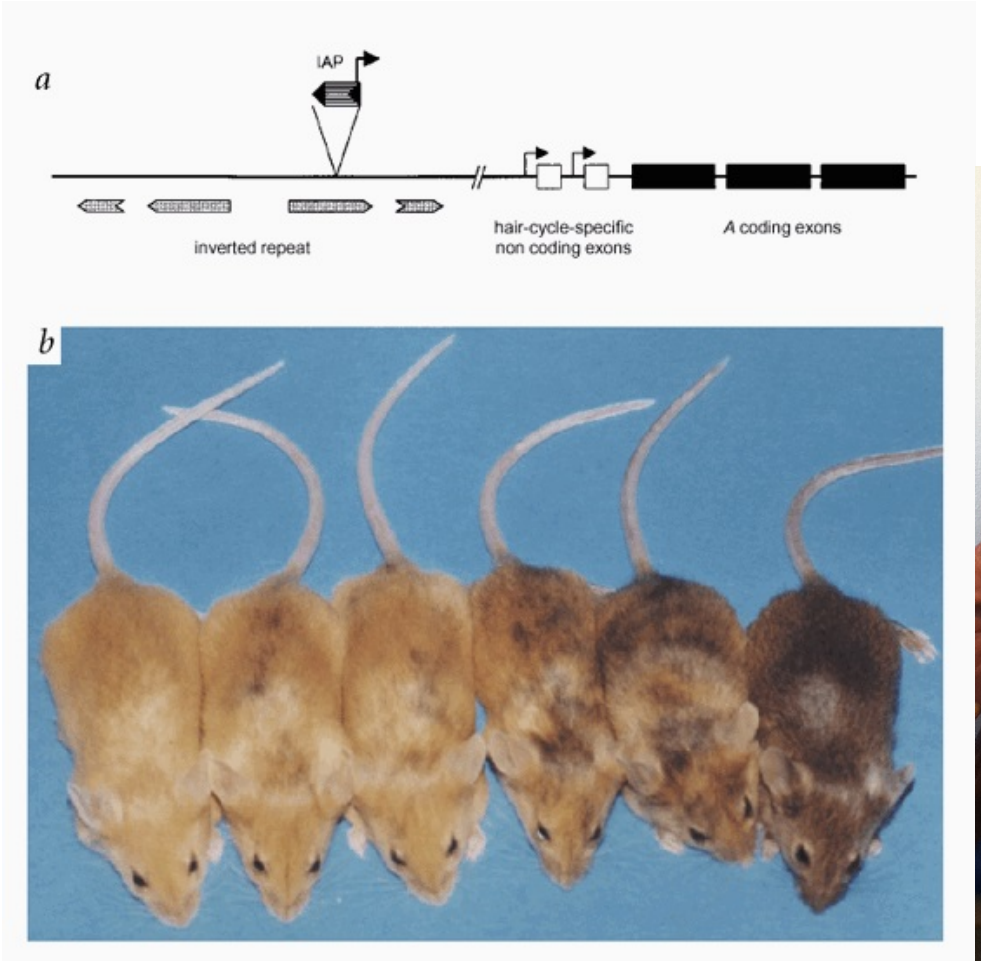
Cancer site	Cumulative risk[1] (%)	N Twin pairs concordant/ discordant		Familial risk[2] (95% CI) – MZ twins	Familial risk (95% CI) – DZ twi
		MZ	DZ		
Overall cancer	32.4%	1383/5887	1933/11461	45.9% (44.1%-47.7%)	37.1% (35.7-38.4
Head and neck[3]	0.8%	5/191	6/361	6.0% (2.4-14.4%)	5.1% (2.2-11.3%
Esophagus	0.4%	0/87	0/183	--	--
Stomach	1.6%	14/338	15/648	6.8% (3.9-11.4%)	4.4% (2.6-7.3%
Small intestine	0.1%	0/32	0/59	--	--
Colon	2.9%	30/577	31/1156	10.9% (7.4-15.8%)	7.9% (5.4-11.4%
Rectum and anus	1.9%	14/440	13/771	6.6% (3.7-11.4%)	5.8% (3.4-9.7%
Liver	0.5%	0/124	2/208	--	--
Gallbladder, extrahepatic bile duct	0.5%	1/110	1/187	0.5% (0-4.7%)	0.3% (0-1.0%)
Pancreas	1.1%	4/234	6/508	4.3% (1.5-11.6%)	3.7% (1.5-8.6%
Nose, sinuses	0.1%	0/21	0/36	--	--
Larynx	0.2%	2/53	1/113	8.4% (2.3-26.4%)	2.7% (1.1-6.1%
Lung, trachea and bronchus	3.2%	50/682	74/1366	17.5% (13.4-22.5%)	13.4% (10.8-16.6
Pleura	0.1%	1/22	0/38	--	--
Bone	0.1%	0/20	0/35	--	--
Melanoma of skin	1.2%	11/342	6/585	19.6% (11.5-31.3%)	6.1% (2.7-13.2%
Skin, non-melanoma	3.0%	16/395	10/618	14.5% (7.5-26.2%)	4.6% (2.4-8.6%
Connective and soft tissues	0.2%	0/57	0/110	--	--
Breast	9.4%	124/1175	141/2223	28.1% (23.9-32.8%)	19.9% (17.0-23.2
Cervix uteri	1.0%	1/210	3/324	--	--
Corpus uteri	2.2%	9/272	6/481	7.0% (3.4-14.0%)	3.6% (1.6-8.0%
Uterus, other	0.1%	0/24	0/36	--	--
Ovary	1.6%	6/234	4/427	8.7% (4.0-17.9%)	2.9% (1.1-7.4%
Other female genital organs	0.4%	0/47	1/84	--	--
Penis and other genital organs	0.1%	0/15	0/34	--	--
Prostate	10.5%	197/807	148/1719	38.0% (33.9-42.2%)	22.0% (18.8-25.7
Testis	0.5%	5/90	3/123	13.8% (5.7-29.6%)	6.0% (1.9-16.9%
Kidney	0.8%	5/196	2/374	6.7% (2.8-15.1%)	1.8% (0.4-6.8%
Bladder, other urinary organs	2.2%	18/471	13/870	9.9% (6.2-15.5%)	5.5% (3.1-9.7%
Eye	0.1%	2/30	0/64	--	--
Brain, central nervous system	0.9%	1/343	3/522	1.7% (0.5-6.2%)	1.8% (0.3-12.0%
Thyroid	0.2%	0/85	1/132	--	--
Hodgkin's disease	0.1%	0/57	0/69	--	--
Multiple myeloma	0.4%	0/114	0/174	--	--
Non-Hodgkin lymphoma	0.7%	1/254	3/466	--	--
Leukemia, acute	0.3%	0/77	0/139	--	--
Leukemia, other	0.6%	5/128	3/259	15.2% (6.1-33.2%)	4.1% (1.3-11.9%

	NorTwinCan 2014		Lichtenstein 2000
	Heritability (%, 95% CI)	**Shared environment (%, 95% CI)**	**Heritability**
Overall cancer	33% (30-37%)	0%	N/A
Head and neck	9% (0-60%)	26% (0-65%)	N/A
Stomach	22% (0-55%)	6% (0-31%)	28%
Colon	15% (0-45%)	16% (0-38%)	35%*
Rectum and anus	14% (0-50%)	10% (0-38%)	35%*
Lung	18% (0-42%)	24% (7-40%)	26%
Skin, melanoma	58% (43-73%)	0%	N/A
Skin, non-melanoma	43% (26-59%)	0%	N/A
Breast	31% (11-51%)	16% (0-31%)	27%
Corpus uteri	27% (11-43%)	0%	0%
Ovary	39% (23-55%)	0%	22%
Prostate	57% (51-63%)	0%	42%
Testis	37% (0-93%)	24% (0-70%)	N/A
Kidney	38% (21-55%)	0%	N/A
Bladder, other urinary organs	30% (0-67%)	0%	31%
Leukemia, other	57% (0-100%)	0%	N/A

* Lichtenstein *et al* presented data for colon and rectum combined


Heritabilities
GWAS common variants
Twins
0
10
20
30
40
50
60
Prostate
Testis
Stomach
Lung
Kidney
Bladder
Breast

The Epigenome



History: Statistical genetics

How is variation at phenotypic level governed by variation at genetic level?

- R.A. Fisher (1918): Two landmark papers.

Biometrical Genetics

XV.—**The Correlation between Relatives on the Supposition** of Mendelian Inheritance. By **R. A. Fisher**, B.A. *Communicated by* Professor J. Arthur Thomson. (With Four Figures in Text.)

(MS. received June 15, 1918. Read July 8, 1918. Issued separately October 1, 1918.)

CONTENTS.

Prologue

Effect?

Exposure$\rightarrow$Outcome

- Outcome: Continuous variable (eg. time to event, BMI, . . .).
- What is the contribution of genetic and environmental factors to the variation in outcome?

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

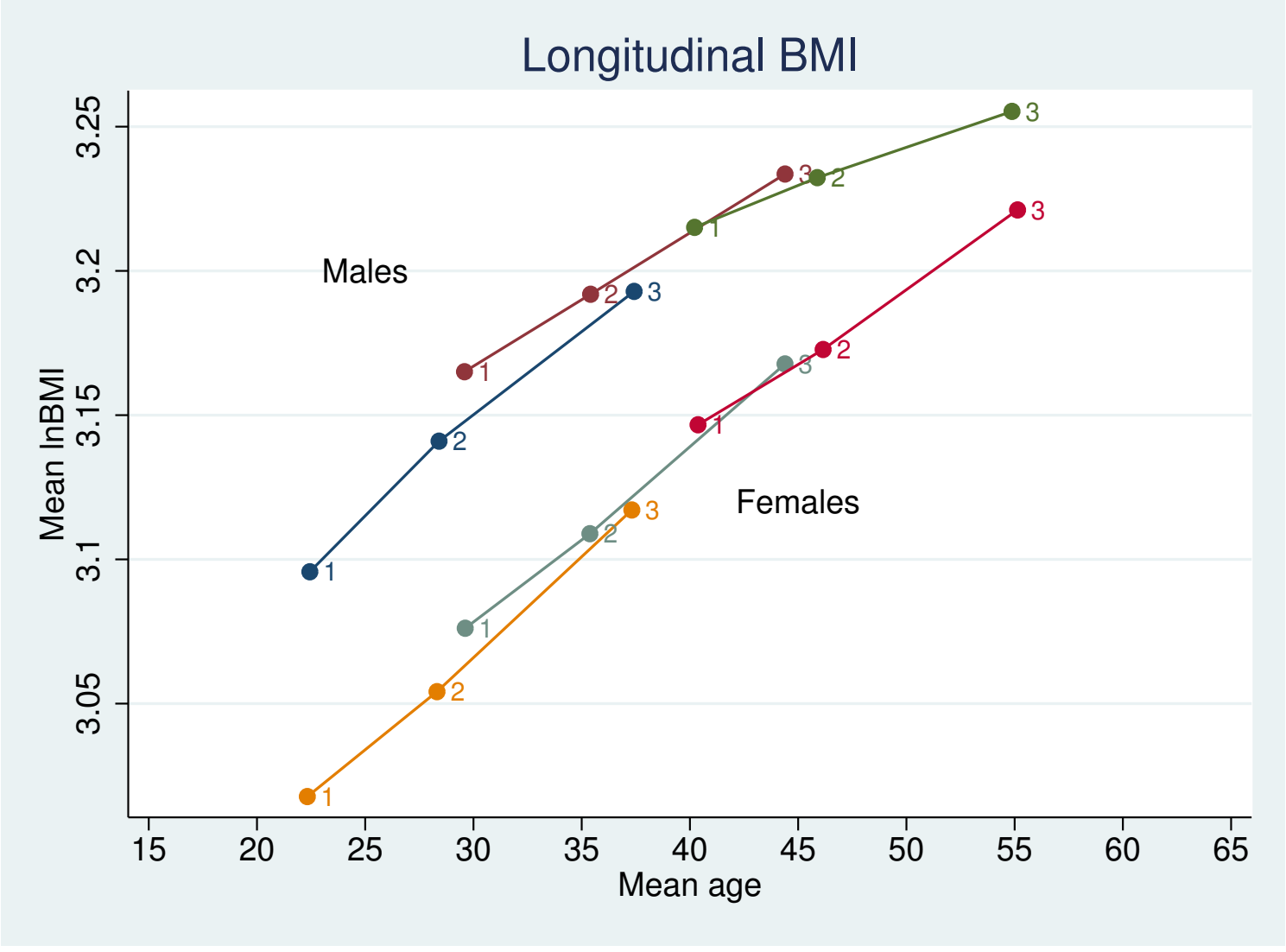
- What kind of genetic and environmental influences to expect?
- Example: SNPedia.com - click here

University of Southern Denmark

Overview

University of Southern Denmark

Case study: Body Mass Index



Case study: Background

- Body Mass Index defined as weight (kg) per squared height (m^2) is a *complex trait* related to several health related factors (eg. obesity, diabetes, aging).
- The index may allow for comparison among individuals with different height, but is not regarded invariant between different sexes.
- Estimated heritability of 0.60-0.70 in BMI has been reported from 'The GenomeEUtwin Study' using 8 cohorts (Schousboe et al. 2003)) with the remaining 30-40% due to a unique environmental variance component.
- The genetic influence is a complex action of several genes. Only few genetic variants identified so far.
- The interplay with environmental factors is under intense investigation.

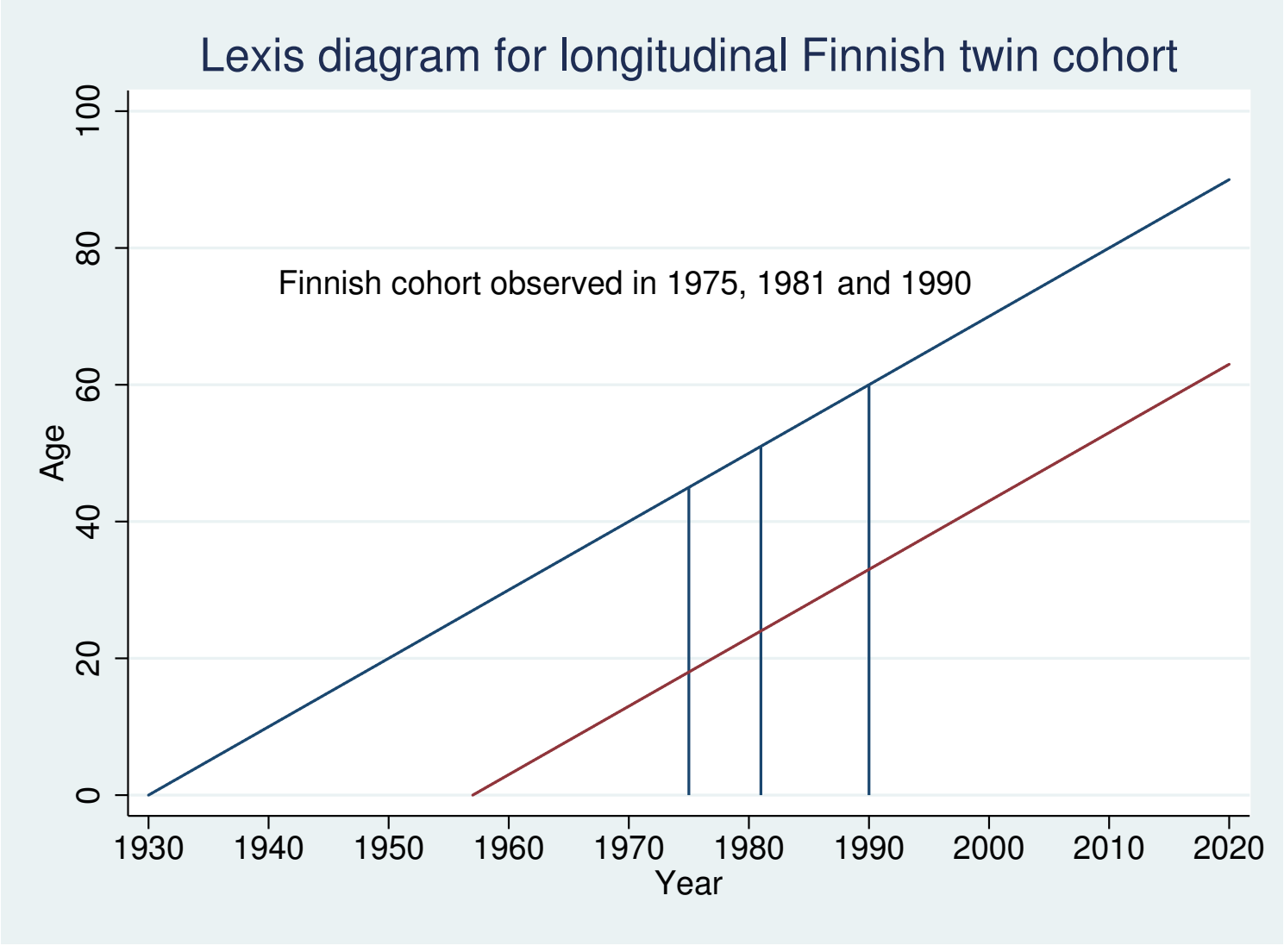
University of Southern Denmark

The Methodology

- Genetic influence on continuous trait
- Correlation: *measure of similarity to be compared for MZ and DZ pairs*
- The polygenic model allows for modelling type and magnitude of genetic influence on BMI by decomposing *the variance in BMI* into genetic and environmental components.

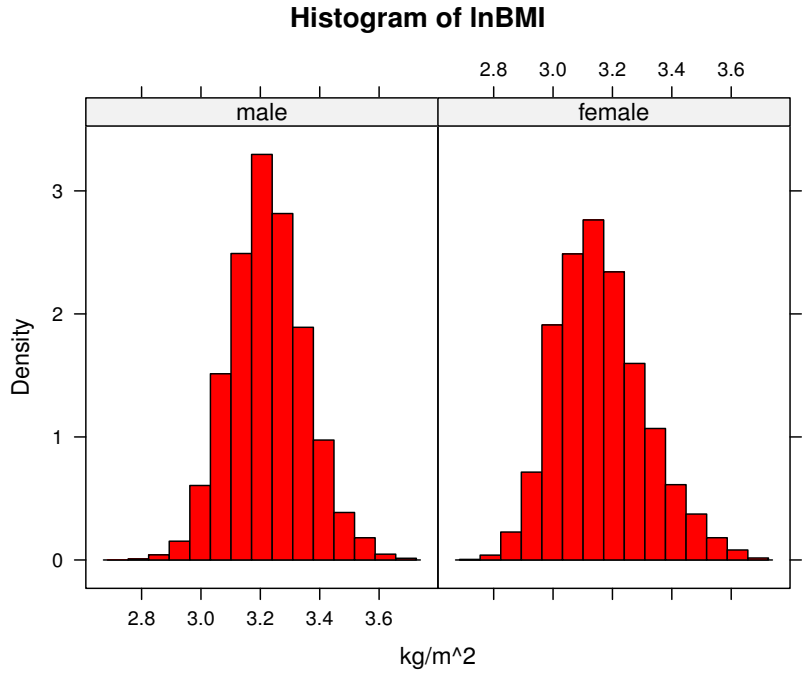
University of Southern Denmark

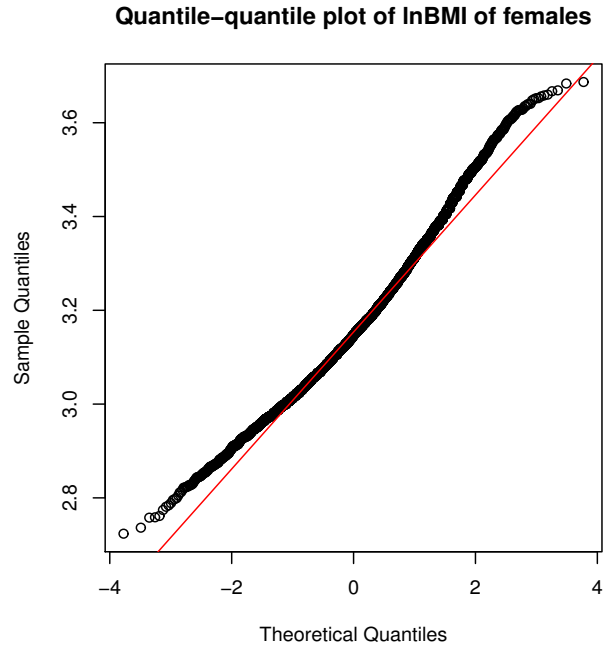
The Material



The Data

	id	tvparnr	nr	bmi	sex	age	zygocity
1.	1000011	100001	1	26.33289	male	57.57974	DZ
2.	1000012	100001	2	25.46939	male	57.57974	DZ
3.	1000021	100002	1	28.65014	male	57.0486	MZ
4.	1000031	100003	1	28.40909	male	57.6783	DZ
5.	1000041	100004	1	27.25089	male	53.51677	DZ
	...	...	.	...	..	...	.





The Data

- How about marginal effects of eg. age and gender?
- Regression model for the response of i'th individual in j'th pair:

$$y_{ij} = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{sex}_{ij} + \beta_3 \text{ageXsex}_{ij} + u_j + \epsilon_{ij},$$

  where u_j is a term that varies in pairs - a random intercept that models the within pair covariance.
- *Why complicate matter?*
- For inference, ie., konfidence and tests, independent observations are needed and u_j models the dependence in twin pairs giving adjusted inference.
- In Stata (see accompanying script for model diagnostics etc.):

```
> xi: xtmixed  lnbmi i.sex*age || tvparnr: , var mle

Log likelihood =  7214.1296                     Prob > chi2       =    0.0000

------------------------------------------------------------------------------
       lnbmi |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_2 |  -.1619652   .0166531    -9.73   0.000    -.1946047   -.1293257
         age |   .0035526   .0002722    13.05   0.000     .0030191    .0040861
 _IsexXage_2 |   .0022916   .0003692     6.21   0.000      .001568    .0030153
       _cons |   3.065863   .0123446   248.36   0.000     3.041668    3.090058
------------------------------------------------------------------------------
```

The Data

- How about marginal effects of eg. age and gender?
- Regression model for the response of i'th individual in j'th pair:

$$y_{ij} = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{sex}_{ij} + \beta_3 \text{ageXsex}_{ij} + u_j + \epsilon_{ij},$$

where u_j is a term that varies in pairs - a random intercept that models the within pair covariance.

- *Why complicate matter?*
- For inference, ie., konfidence and tests, independent observations are needed and u_j models the dependence in twin pairs giving adjusted inference.
- In Stata (see accompanying script for model diagnostics etc.):

```
> xi: xtmixed  lnbmi i.sex*age || tvparnr: , var mle

Log likelihood =  7214.1296                     Prob > chi2       =    0.0000

------------------------------------------------------------------------------
       lnbmi |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_2 |  -.1619652   .0166531    -9.73   0.000    -.1946047   -.1293257
         age |   .0035526   .0002722    13.05   0.000     .0030191    .0040861
 _IsexXage_2 |   .0022916   .0003692     6.21   0.000      .001568    .0030153
       _cons |   3.065863   .0123446   248.36   0.000     3.041668    3.090058
------------------------------------------------------------------------------
```

The Data

- How about marginal effects of eg. age and gender?
- Regression model for the response of i'th individual in j'th pair:

$$y_{ij} = \beta_0 + \beta_1 \text{age}_{ij} + \beta_2 \text{sex}_{ij} + \beta_3 \text{ageXsex}_{ij} + u_j + \epsilon_{ij},$$

  where u_j is a term that varies in pairs - a random intercept that models the within pair covariance.
- *Why complicate matter?*
- For inference, ie., konfidence and tests, independent observations are needed and u_j models the dependence in twin pairs giving adjusted inference.
- In Stata (see accompanying script for model diagnostics etc.):

```
> xi: xtmixed  lnbmi i.sex*age || tvparnr: , var mle

Log likelihood =  7214.1296                     Prob > chi2       =    0.0000

------------------------------------------------------------------------------
       lnbmi |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Isex_2 |  -.1619652   .0166531    -9.73   0.000    -.1946047   -.1293257
         age |   .0035526   .0002722    13.05   0.000     .0030191    .0040861
 _IsexXage_2 |   .0022916   .0003692     6.21   0.000      .001568    .0030153
       _cons |   3.065863   .0123446   248.36   0.000     3.041668    3.090058
------------------------------------------------------------------------------
```

The Data

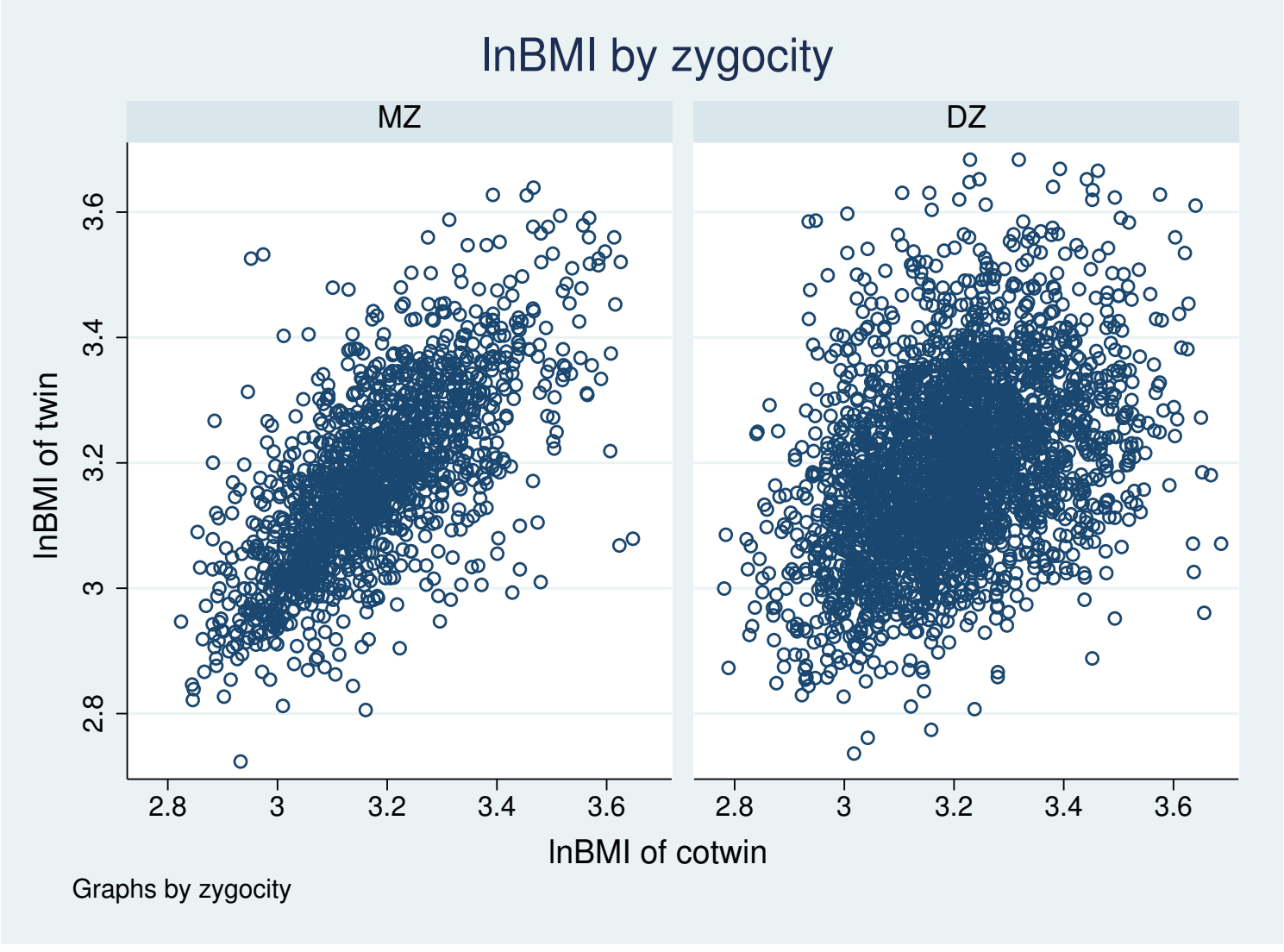
- We consider the logarithm of BMI, in notation ‘lnBMI’.
- The outcome is associated with gender and age.
- Lets load the data into R and head for descriptives:

Use R

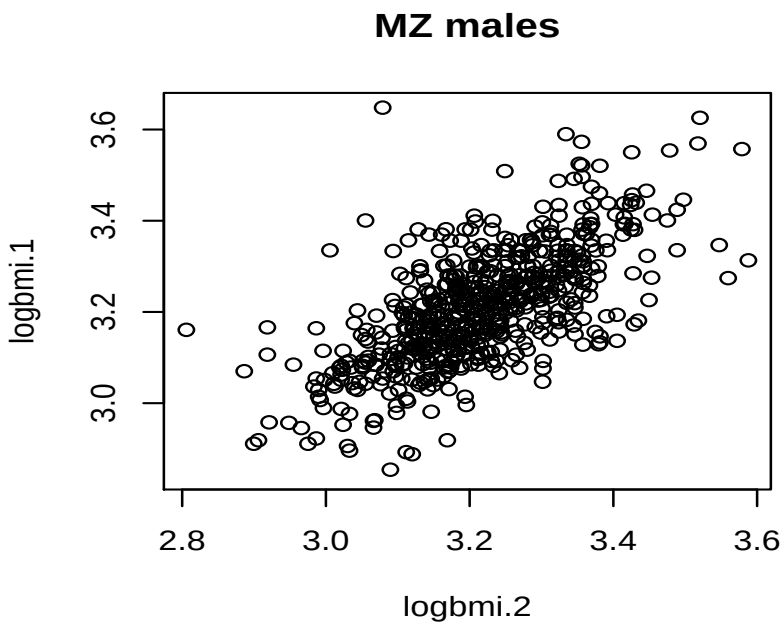
- Start R and open the R-script ’twinbmi.R’.
- Run lines in script till Section *Pairs* begins.

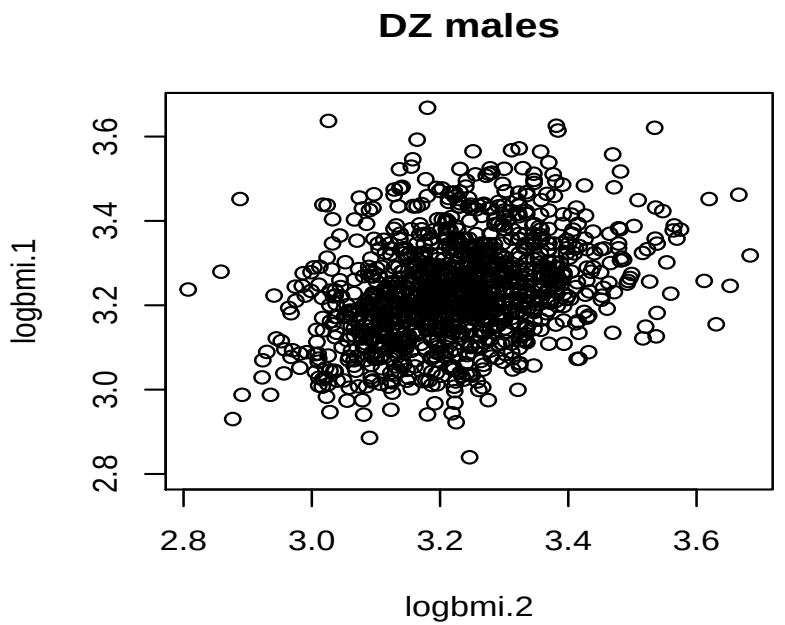
- We then go on considering the paired structure.

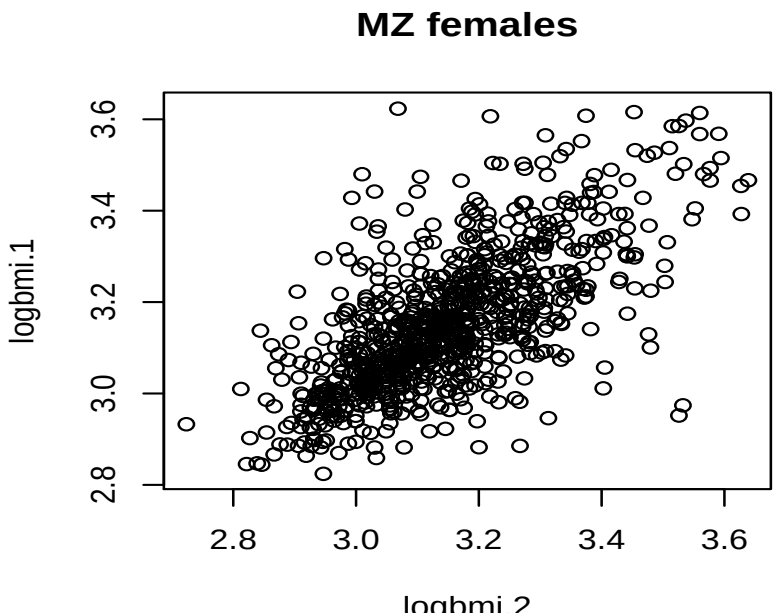
BMI of twin versus BMI of cotwin

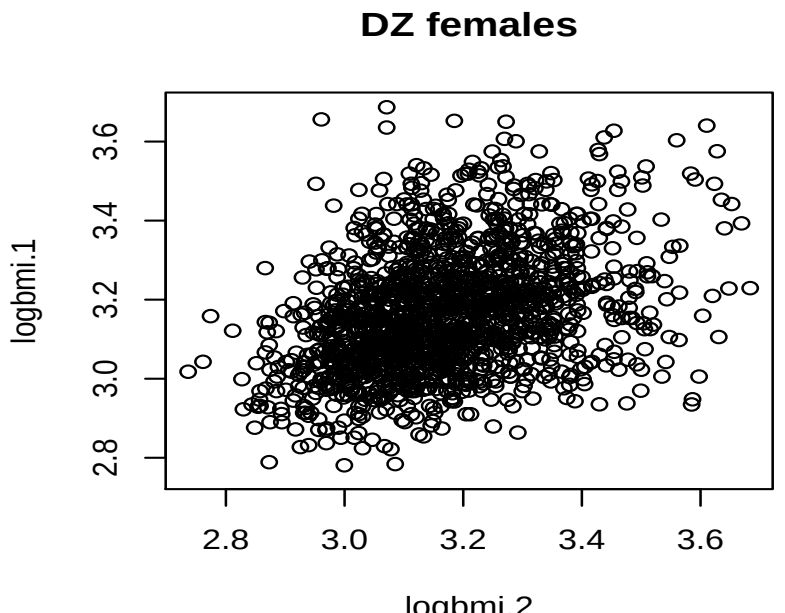


BMI of twin versus BMI of cotwin









Genetic Influence on lnBMI

	Polygenic model		
	Number of pairs	Correlation (95% CI)	Heritability (95% CI)
MZ pairs	1483	? (?,?)	? (?,?)
DZ pairs	2788	? (?,?)	Biometric model

Overview

University of Southern Denmark

Statistical genetics

How is variation at phenotypic level governed by variation at genetic level?

- Two structures for modelling: mean and variance-covariance.
- R.A. Fisher (1918): The variance-covariance matrix varies by type of twin pairs.

$$\Sigma = \begin{pmatrix} \text{variance of first twin} & \text{covariance of twins} \\ \text{covariance of twins} & \text{variance of second twin} \end{pmatrix}$$

- We begin seeking a measure of twin similarity: ρ

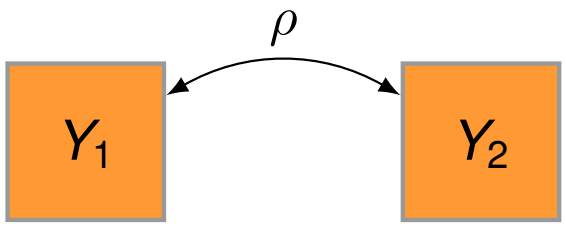
Biometrical Genetics

XV.—**The Correlation between Relatives on the Supposition of Mendelian Inheritance.** By **R. A. Fisher**, B.A. *Communicated by* Professor J. Arthur Thomson. (With Four Figures in Text.)

(MS. received June 15, 1918. Read July 8, 1918. Issued separately October 1, 1918.)

CONTENTS.

SEM - Correlation Path Diagram representation



Within pair intraclass correlation

What's on?

- *How to measure twin similarity?*
- Given pairs (y_{1j}, y_{2j}) of observations of a continuous trait the correlation within pairs is the usual (product-moment) correlation assuming equal mean and variance for twin 1 and twin 2.
- *Why assume equal mean and variance for twin 1 and twin 2?*
- Twin 1 and twin 2 can be interchanged when there is no ordering of twin and co-twin.
- *What is the interpretation of the within pair correlation?*
- This is the amount of variance between pairs of the total variance in the trait.
- *What is the purpose?*
- Higher correlation in MZ than in DZ pairs indicate genetic influence on the trait.
- *But for this comparison you should assume equal mean and variance for MZ and DZ twins!*
- Yes! MZ and DZ twins do not differ (on average) as singletons.

Within pair intraclass correlation

What's on?

- *How to measure twin similarity?*
- Given pairs (y_{1j}, y_{2j}) of observations of a continuous trait the correlation within pairs is the usual (product-moment) correlation assuming equal mean and variance for twin 1 and twin 2.
- *Why assume equal mean and variance for twin 1 and twin 2?*
- Twin 1 and twin 2 can be interchanged when there is no ordering of twin and co-twin.
- *What is the interpretation of the within pair correlation?*
- This is the amount of variance between pairs of the total variance in the trait.
- *What is the purpose?*
- Higher correlation in MZ than in DZ pairs indicate genetic influence on the trait.
- *But for this comparison you should assume equal mean and variance for MZ and DZ twins!*
- Yes! MZ and DZ twins do not differ (on average) as singletons.

Within pair intraclass correlation

What's on?

- *How to measure twin similarity?*
- Given pairs (y_{1j}, y_{2j}) of observations of a continuous trait the correlation within pairs is the usual (product-moment) correlation assuming equal mean and variance for twin 1 and twin 2.
- *Why assume equal mean and variance for twin 1 and twin 2?*
- Twin 1 and twin 2 can be interchanged when there is no ordering of twin and co-twin.
- *What is the interpretation of the within pair correlation?*
- This is the amount of variance between pairs of the total variance in the trait.
- *What is the purpose?*
- Higher correlation in MZ than in DZ pairs indicate genetic influence on the trait.
- *But for this comparison you should assume equal mean and variance for MZ and DZ twins!*
- Yes! MZ and DZ twins do not differ (on average) as singletons.

Within pair intraclass correlation

What's on?

- *How to measure twin similarity?*
- Given pairs (y_{1j}, y_{2j}) of observations of a continuous trait the correlation within pairs is the usual (product-moment) correlation assuming equal mean and variance for twin 1 and twin 2.
- *Why assume equal mean and variance for twin 1 and twin 2?*
- Twin 1 and twin 2 can be interchanged when there is no ordering of twin and co-twin.
- *What is the interpretation of the within pair correlation?*
- This is the amount of variance between pairs of the total variance in the trait.
- *What is the purpose?*
- Higher correlation in MZ than in DZ pairs indicate genetic influence on the trait.
- *But for this comparison you should assume equal mean and variance for MZ and DZ twins!*
- Yes! MZ and DZ twins do not differ (on average) as singletons.

Within pair intraclass correlation

What's on?

- *How to measure twin similarity?*
- Given pairs (y_{1j}, y_{2j}) of observations of a continuous trait the correlation within pairs is the usual (product-moment) correlation assuming equal mean and variance for twin 1 and twin 2.
- *Why assume equal mean and variance for twin 1 and twin 2?*
- Twin 1 and twin 2 can be interchanged when there is no ordering of twin and co-twin.
- *What is the interpretation of the within pair correlation?*
- This is the amount of variance between pairs of the total variance in the trait.
- *What is the purpose?*
- Higher correlation in MZ than in DZ pairs indicate genetic influence on the trait.
- *But for this comparison you should assume equal mean and variance for MZ and DZ twins!*
- Yes! MZ and DZ twins do not differ (on average) as singletons.

Within pair intraclass correlation

What's on?

- *How to measure twin similarity?*
- Given pairs (y_{1j}, y_{2j}) of observations of a continuous trait the correlation within pairs is the usual (product-moment) correlation assuming equal mean and variance for twin 1 and twin 2.
- *Why assume equal mean and variance for twin 1 and twin 2?*
- Twin 1 and twin 2 can be interchanged when there is no ordering of twin and co-twin.
- *What is the interpretation of the within pair correlation?*
- This is the amount of variance between pairs of the total variance in the trait.
- *What is the purpose?*
- Higher correlation in MZ than in DZ pairs indicate genetic influence on the trait.
- *But for this comparison you should assume equal mean and variance for MZ and DZ twins!*
- Yes! MZ and DZ twins do not differ (on average) as singletons.

Within pair intraclass correlation

What's on?

- *How to measure twin similarity?*
- Given pairs (y_{1j}, y_{2j}) of observations of a continuous trait the correlation within pairs is the usual (product-moment) correlation assuming equal mean and variance for twin 1 and twin 2.
- *Why assume equal mean and variance for twin 1 and twin 2?*
- Twin 1 and twin 2 can be interchanged when there is no ordering of twin and co-twin.
- *What is the interpretation of the within pair correlation?*
- This is the amount of variance between pairs of the total variance in the trait.
- *What is the purpose?*
- Higher correlation in MZ than in DZ pairs indicate genetic influence on the trait.
- *But for this comparison you should assume equal mean and variance for MZ and DZ twins!*
- Yes! MZ and DZ twins do not differ (on average) as singletons.

Within pair intraclass correlation

What's on?

- *How to measure twin similarity?*
- Given pairs (y_{1j}, y_{2j}) of observations of a continuous trait the correlation within pairs is the usual (product-moment) correlation assuming equal mean and variance for twin 1 and twin 2.
- *Why assume equal mean and variance for twin 1 and twin 2?*
- Twin 1 and twin 2 can be interchanged when there is no ordering of twin and co-twin.
- *What is the interpretation of the within pair correlation?*
- This is the amount of variance between pairs of the total variance in the trait.
- *What is the purpose?*
- Higher correlation in MZ than in DZ pairs indicate genetic influence on the trait.
- *But for this comparison you should assume equal mean and variance for MZ and DZ twins!*
- Yes! MZ and DZ twins do not differ (on average) as singletons.

Within pair intraclass correlation

What's on?

- *How to measure twin similarity?*
- Given pairs (y_{1j}, y_{2j}) of observations of a continuous trait the correlation within pairs is the usual (product-moment) correlation assuming equal mean and variance for twin 1 and twin 2.
- *Why assume equal mean and variance for twin 1 and twin 2?*
- Twin 1 and twin 2 can be interchanged when there is no ordering of twin and co-twin.
- *What is the interpretation of the within pair correlation?*
- This is the amount of variance between pairs of the total variance in the trait.
- *What is the purpose?*
- Higher correlation in MZ than in DZ pairs indicate genetic influence on the trait.
- *But for this comparison you should assume equal mean and variance for MZ and DZ twins!*
- Yes! MZ and DZ twins do not differ (on average) as singletons.

Correlation in twins

Assumptions

- A measure of twin similarity: ρ
- Given pairs of observations of a continuous trait,

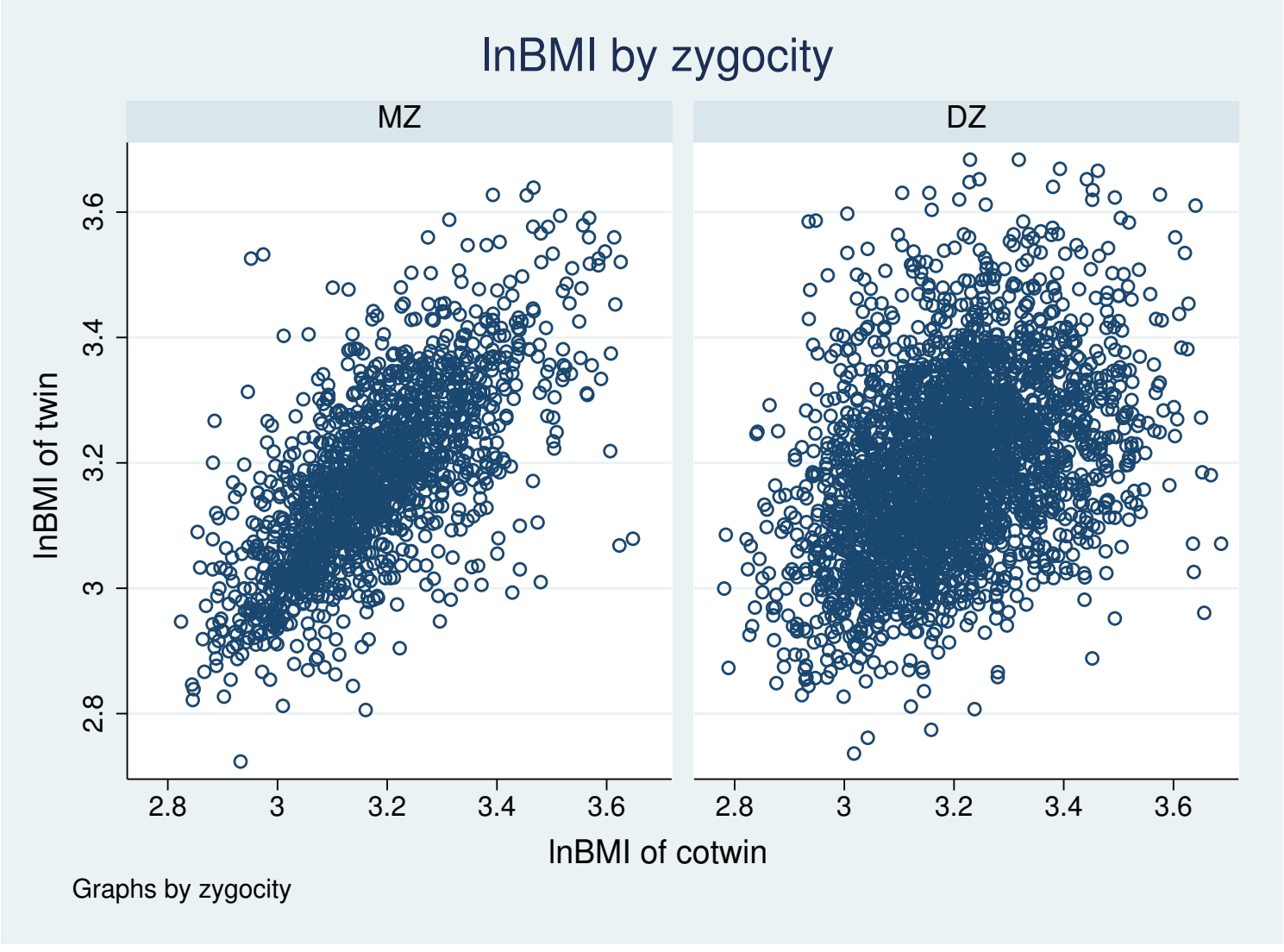
$$\{(y_{1j}, y_{2j})\} \quad j = 1 \ldots n \quad \text{(pairs)}$$

  the correlation is defined by

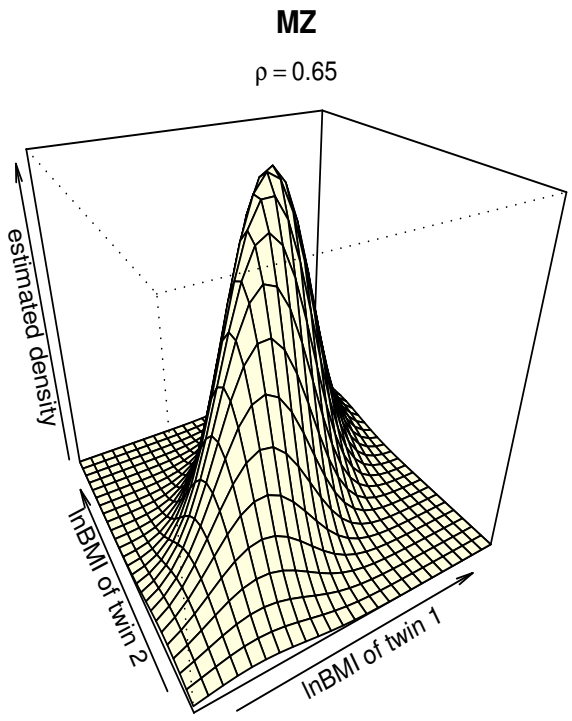
$$\rho(y_1, y_2) = \frac{\text{cov}(y_1, y_2)}{\sqrt{\text{var}(y_1)}\sqrt{\text{var}(y_2)}}$$

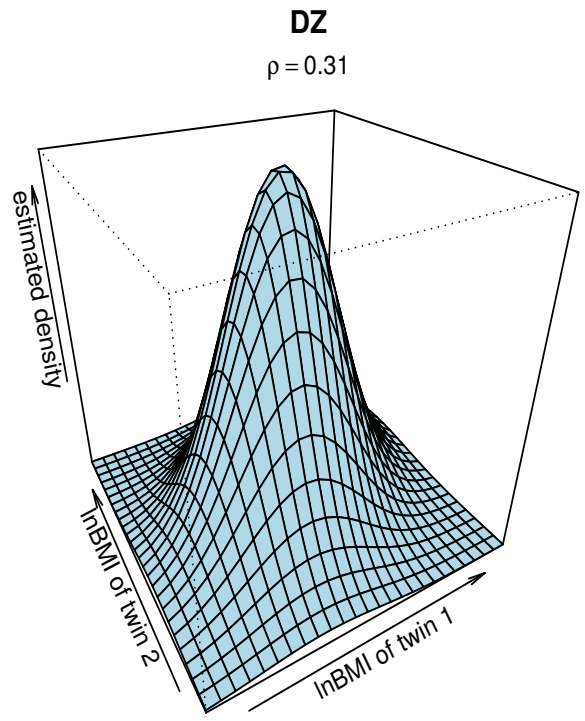
- Assumption: Equal mean and variance for twin 1 and twin 2.
- Assumption: Equal mean and variance for MZ and DZ twins.
- Estimation: By maximum likelihood assuming bivariate normal distribution.

BMI of twin versus BMI of cotwin



Bivarite normal distribution plot





Correlation in twins

- The pair (Y_1, Y_2) is bivariate normal distributed with mean (μ_1, μ_2) and variance-covariance matrix given by

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

- NB! Variance in diagonal and covariance in off-diagonal.
- The variance in Y is denoted by σ^2.
- The covariance between twin 1 and twin 2 known as $\mathrm{cov}(Y_1, Y_2)$ is $\rho\sigma_1\sigma_2$.
- *How likely are observed data given parameters?*
- **Principle:** We choose parameters that make our observations most likely. ▸ Appendix: Maximum Likelihood Estimation

Correlation in twins - practicals

Estimation

- We will estimate twin correlations and test assumptions.
- We start using Mets and then try out OpenMx.

Practicals

- Carry on executing lines from the R-script "twinbmi.R":
- -try to obtain a twin-twin plot.
- -run the saturated model.
- A successful estimation is when parameters maximizes the log-likelihood function of data.
- -a criteria is that the score-command evaluates to very low values (typically below 10^{-5}). See script.

mean : What is the effect of sex and age for twin 1 and twin 2, mz and dz type?
covariance : What is the correlation in pairs?

- The log likelihood function of these parameters is termed 'log Lik.'. It's value together with the degrees of freedom is a measure of goodness of fit to data.

Correlation in twins - practicals

Practicals

- We go on looking for correlations and their inference in model that meets the assumptions $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$ for twin 1 and twin 2 in MZ and DZ pairs.
- The assumptions induce submodels of the saturated model - you may find these specified in the script.
- What is the outcome of the submodel obtained by constraining equal regressions, intercepts and residual variances for twin 1 and twin 2 in the saturated model?
- What is the fit of the submodel and how does it compare to that of the saturated model in terms of a χ^2 test of the difference $-2(\log(L_1) - \log(L_2))$ on the difference in degrees of freedom?
- Carry on constraining marginals for MZ and DZ twins and consider the same issues as above.

WARNING! Busy slides coming up

```
> lnbmi.sat
```

Group 1: MZ (n=1483)

	Estimate	Std. Error	Z value	Pr(>|z|)
Regressions:				
logbmi.1~age.1	0.00596	0.00058	10.22059	<1e-12
logbmi.1~gendermale.1	0.16464	0.04180	3.93862	8.195e-05
logbmi.1~age:gendermale.1	-0.00247	0.00092	-2.67792	0.007408
logbmi.2~age.1	0.00626	0.00057	11.06885	<1e-12
logbmi.2~gendermale.1	0.18606	0.04052	4.59192	4.392e-06
logbmi.2~age:gendermale.1	-0.00274	0.00089	-3.06415	0.002183
Intercepts:				
logbmi.1	2.88926	0.02611	110.66670	<1e-12
logbmi.2	2.87071	0.02531	113.42068	<1e-12
Additional Parameters:				
log(var(MZ)).1	-4.00912	0.03681	-108.91220	<1e-12
log(var(MZ)).2	-4.07144	0.03682	-110.58848	<1e-12
atanh(rhoMZ)	0.76958	0.02609	29.49281	<1e-12

Group 2: DZ (n=2788)

	Estimate	Std. Error	Z value	Pr(>|z|)
Regressions:				
logbmi.1~age.1	0.00565	0.00043	13.11664	<1e-12
logbmi.1~gendermale.1	0.15530	0.03004	5.16951	2.347e-07
logbmi.1~age:gendermale.1	-0.00198	0.00066	-3.00731	0.002636
logbmi.2~age.1	0.00571	0.00043	13.26981	<1e-12
logbmi.2~gendermale.1	0.16833	0.03001	5.60935	2.031e-08
logbmi.2~age:gendermale.1	-0.00254	0.00066	-3.86482	0.0001112
Intercepts:				
logbmi.1	2.91365	0.01943	149.92756	<1e-12
logbmi.2	2.91383	0.01941	150.09632	<1e-12
Additional Parameters:				
log(var(DZ)).1	-4.02373	0.02685	-149.86729	<1e-12
log(var(DZ)).2	-4.02586	0.02685	-149.94604	<1e-12
atanh(rhoDZ)	0.31399	0.01894	16.57373	<1e-12

	Estimate	2.5%	97.5%
Correlation within MZ:	0.64669	0.61594	0.67546
Correlation within DZ:	0.30406	0.26999	0.33737

```
'log Lik.' 5629.137 (df=22)
AIC: -11214.27
BIC: -11074.36
```

```
> lnbmi.flex
```

```
Group 1: MZ (n=1483)
                              Estimate Std. Error    Z value  Pr(>|z|)
Regressions:
   logbmi.1~age.1              0.00611    0.00052   11.79643    <1e-12
   logbmi.1~gendermale.1       0.17535    0.03735    4.69507 2.665e-06
   logbmi.1~age:gendermale.1  -0.00260    0.00082   -3.16076  0.001574
Intercepts:
   logbmi.1                    2.87999    0.02319  124.19495    <1e-12
Additional Parameters:
   log(var(MZ))               -4.03940    0.03100 -130.31863    <1e-12
   atanh(rhoMZ)                0.76793    0.02600   29.53193    <1e-12
```

```
Group 2: DZ (n=2788)
                              Estimate Std. Error    Z value  Pr(>|z|)
Regressions:
   logbmi.1~age.1              0.00568    0.00035   16.33574    <1e-12
   logbmi.1~gendermale.1       0.16182    0.02425    6.67419 2.486e-11
   logbmi.1~age:gendermale.1  -0.00226    0.00053   -4.25499 2.091e-05
Intercepts:
   logbmi.1                    2.91374    0.01569  185.74505    <1e-12
Additional Parameters:
   log(var(DZ))               -4.02407    0.01985 -202.75292    <1e-12
   atanh(rhoDZ)                0.31296    0.01897   16.49464    <1e-12

                       Estimate 2.5%    97.5%
Correlation within MZ: 0.64572  0.61503 0.67447
Correlation within DZ: 0.30312  0.26898 0.33650

'log Lik.' 5623.369 (df=12)
AIC: -11222.74
BIC: -11146.42

> compare(lnbmi.sat,lnbmi.flex) > #comparison with saturated model

- Likelihood ratio test -

data:
chisq = 11.537, df = 10, p-value = 0.3172
sample estimates:
log likelihood (model 1) log likelihood (model 2)
                5629.137                 5623.369
```

```
> lnbmi.u
```

```
Group 1: MZ (n=1483)
                              Estimate Std. Error    Z value  Pr(>|z|)
Regressions:
   logbmi.1~age.1              0.00583    0.00029   20.15100    <1e-12
   logbmi.1~gendermale.1       0.16606    0.02036    8.15634    <1e-12
   logbmi.1~age:gendermale.1  -0.00236    0.00045   -5.27784 1.307e-07
Intercepts:
   logbmi.1                    2.90261    0.01302  222.92574    <1e-12
Additional Parameters:
   log(var)                   -4.02560    0.01673 -240.67809    <1e-12
   atanh(rhoMZ)                0.77529    0.02313   33.52059    <1e-12
```

```
Group 2: DZ (n=2788)
                              Estimate Std. Error    Z value  Pr(>|z|)
Regressions:
   logbmi.1~age.1              0.00583    0.00029   20.15100    <1e-12
   logbmi.1~gendermale.1       0.16606    0.02036    8.15634    <1e-12
   logbmi.1~age:gendermale.1  -0.00236    0.00045   -5.27784 1.307e-07
Intercepts:
   logbmi.1                    2.90261    0.01302  222.92574    <1e-12
Additional Parameters:
   log(var)                   -4.02560    0.01673 -240.67809    <1e-12
   atanh(rhoDZ)                0.31314    0.01870   16.74748    <1e-12

                       Estimate 2.5%    97.5%
Correlation within MZ: 0.65000  0.62304 0.67541
Correlation within DZ: 0.30329  0.26965 0.33618

'log Lik.' 5614.387 (df=7)
AIC: -11214.77
BIC: -11170.26
>
> compare(lnbmi.u,lnbmi.flex)

- Likelihood ratio test -

data:
chisq = 17.962, df = 5, p-value = 0.002994
sample estimates:
log likelihood (model 1) log likelihood (model 2)
                5614.387                 5623.369
```

Saturated model - model selection

Submodels of saturated model						
Submodel	'log Lik.'	df	$-2\Delta X^2$	Δdf	p	AIC
Saturated	5629.137	22				-11214.27
"equal 1 and 2"	5623.369	12	11.537	10	0.3172	-11222.74
"equal MZ and DZ"	5614.387	7	17.962	5	0.002994	-11214.77

- Saturated model: No constraints on mean and variance structures.
- Note that the mean, μ, is actually the mean of residuals $y_{ij} - \beta_1 \text{sex}_{ij} - \beta_2 \text{age}_{ij} - \beta_3 \text{sex}_{ij}\text{age}_{ij}$.
- We insist on natural assumptions although data may not greatly support these.

University of Southern Denmark

Genetic Influence on lnBMI

	Saturated model w. constraints		
	Number of pairs	Correlation (95% CI)	Heritability (95% CI)
MZ pairs	1483	0.65 (0.62,0.68)	? (?,?)
DZ pairs	2788	0.30 (0.27,0.34)	biometric model

- Correlations are adjusted for effects of sex ($\hat{\beta}_{sex} = 0.17$ coded females zero and males one) and age (by an increment of 0.0058 in lnBMI for each year, slightly lower if male (see interaction term)).

University of Southern Denmark

Overview

University of Southern Denmark

Aims

- Difference in correlations between MZ and DZ twins suggests genetic influence on trait.
- *What type and magnitude of genetic and environmental influences to expect?*
- We consider classical twin analysis using the polygenic model, known as the ADCE-model, in which the individual outcome, Y_i decomposes into

$$Y_i = A_i + D_i + C_i + E_i,$$

where

  - A: Additive genetic effects of alleles
  - D: Dominant genetic effects
  - C: Shared environmental effects
  - E: Unique environmental effects

University of Southern Denmark

Biometric analyses - polygenic model

- Contributing factors to the variation in outcome:

$$\Sigma_Y = \begin{pmatrix} \sigma_A^2 & z\sigma_A^2 \\ z\sigma_A^2 & \sigma_A^2 \end{pmatrix} + \begin{pmatrix} \sigma_D^2 & u\sigma_D^2 \\ u\sigma_D^2 & \sigma_D^2 \end{pmatrix} + \begin{pmatrix} \sigma_C^2 & \sigma_C^2 \\ \sigma_C^2 & \sigma_C^2 \end{pmatrix} + \begin{pmatrix} \sigma_E^2 & 0 \\ 0 & \sigma_E^2 \end{pmatrix}$$

where $z = u = 1$ for MZ pairs, $z = \frac{1}{2}$ and $u = \frac{1}{4}$ for DZ pairs.

In particular, we obtain

- Heritability:

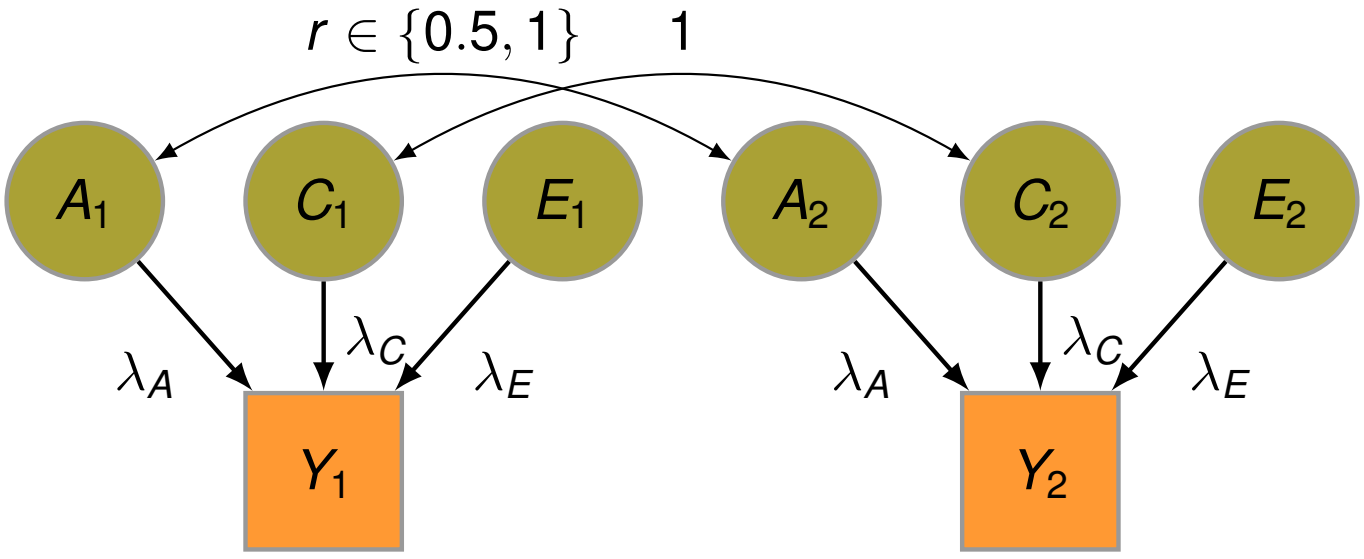
$$h_Y^2 = \frac{\sigma_A^2 + \sigma_D^2}{\sigma_A^2 + \sigma_D^2 + \sigma_C^2 + \sigma_E^2}$$

- Shared environmental effect:

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

University of Southern Denmark

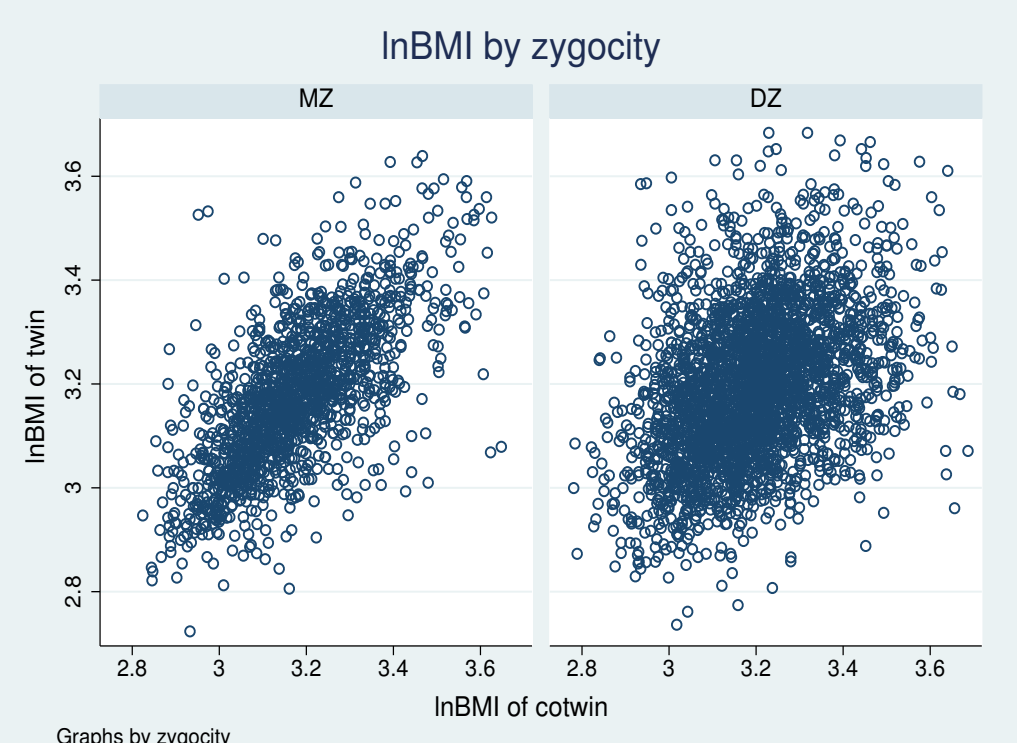
SEM - ACE Path Diagram representation



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).



Biometric analyses - practicals

Estimation

- We will fit appropriate polygenic model to data, estimate heritability and test assumptions.
- NB! Only three of four variance-components are estimable in same model since only three equations in model.

Practicals - Use R

- Carry on estimating the ACE model and its submodels (AE, CE and E).
- Try to estimate the ADE model as well and compare it to ACE wrt. AIC (lowest is most parsimoniuos).
- Which model and hence which results to report?

```
> lnbmi.ace
                          Estimate  Std. Error  Z value  Pr(>|z|)
logbmi                  2.9025e+00  1.3088e-02 221.7584 < 2.2e-16
sd(A)                   1.0743e-01  1.6632e-03  64.5951 < 2.2e-16
sd(C)                   1.1059e-07  1.5885e-02   0.0000         1
sd(E)                   7.9726e-02  1.3531e-03  58.9208 < 2.2e-16
logbmi~age              5.8343e-03  2.9099e-04  20.0498 < 2.2e-16
logbmi~gendermale       1.6611e-01  2.0481e-02   8.1105 5.041e-16
logbmi~age:gendermale -2.3563e-03  4.4889e-04  -5.2492 1.528e-07

MZ-pairs DZ-pairs
    1483     2788

Variance decomposition:
  Estimate 2.5%    97.5%
A 0.64486  0.61920 0.67053
C 0.00000  0.00000 0.00000
E 0.35514  0.32947 0.38080


                         Estimate 2.5%    97.5%
Broad-sense heritability 0.64486  0.61920 0.67053

                       Estimate 2.5%    97.5%
Correlation within MZ: 0.64486  0.61847 0.66981
Correlation within DZ: 0.32243  0.30954 0.33520

'log Lik.' 5613.624 (df=7)
AIC: -11213.25
BIC: -11168.73
```

```
> lnbmi.ade
                          Estimate  Std. Error  Z value  Pr(>|z|)
logbmi                  2.90261390  0.01302012 222.9330 < 2.2e-16
sd(A)                   0.10026867  0.00609022  16.4639 < 2.2e-16
sd(D)                   0.03937491  0.01553051   2.5353   0.01123
sd(E)                   0.07904756  0.00142834  55.3424 < 2.2e-16
logbmi~age              0.00583261  0.00028944  20.1515 < 2.2e-16
logbmi~gendermale       0.16606029  0.02035965   8.1563 3.453e-16
logbmi~age:gendermale -0.00235505  0.00044622  -5.2778 1.307e-07

MZ-pairs DZ-pairs
    1483     2788

Variance decomposition:
  Estimate 2.5%     97.5%
A  0.56315  0.43372  0.69258
D  0.08684 -0.04772  0.22141
E  0.35000  0.32387  0.37614


                         Estimate 2.5%    97.5%
Broad-sense heritability 0.65000  0.62386 0.67613

                       Estimate 2.5%    97.5%
Correlation within MZ: 0.65000  0.62309 0.67537
Correlation within DZ: 0.30329  0.27058 0.33530

'log Lik.' 5614.387 (df=7)
AIC: -11214.77
BIC: -11170.26
>
> #comparison of non-nested models
> AIC(lnbmi.ace,lnbmi.ade)
          df       AIC
lnbmi.ace  7 -11213.25
lnbmi.ade  7 -11214.77
```

```
> lnbmi.ae
                          Estimate  Std. Error  Z value  Pr(>|z|)
logbmi                 2.90248121  0.01308848 221.7584 < 2.2e-16
sd(A)                  0.10743239  0.00166316  64.5953 < 2.2e-16
sd(E)                  0.07972558  0.00135310  58.9208 < 2.2e-16
logbmi~age             0.00583434  0.00029099  20.0498 < 2.2e-16
logbmi~gendermale      0.16611110  0.02048100   8.1105 5.041e-16
logbmi~age:gendermale -0.00235629  0.00044889  -5.2492 1.528e-07

MZ-pairs DZ-pairs
    1483     2788

Variance decomposition:
  Estimate 2.5%    97.5%
A 0.64486  0.61920 0.67053
E 0.35514  0.32947 0.38080


                         Estimate 2.5%    97.5%
Broad-sense heritability 0.64486  0.61920 0.67053

                       Estimate 2.5%    97.5%
Correlation within MZ: 0.64486  0.61847 0.66981
Correlation within DZ: 0.32243  0.30954 0.33520

'log Lik.' 5613.624 (df=6)
AIC: -11215.25
BIC: -11177.09
>
> compare(lnbmi.ade,lnbmi.ae)

- Likelihood ratio test -

data:
chisq = 1.5276, df = 1, p-value = 0.2165
sample estimates:
log likelihood (model 1) log likelihood (model 2)
                5614.387                 5613.624
```

```
> lnbmi.ce
                          Estimate  Std. Error  Z value  Pr(>|z|)
logbmi                 2.90078541  0.01306870 221.9643 < 2.2e-16
sd(C)                  0.08684814  0.00171221  50.7228 < 2.2e-16
sd(E)                  0.10148741  0.00109924  92.3255 < 2.2e-16
logbmi~age             0.00585634  0.00029065  20.1491 < 2.2e-16
logbmi~gendermale      0.16676183  0.02048187   8.1419 3.890e-16
logbmi~age:gendermale -0.00237224  0.00044908  -5.2825 1.275e-07

MZ-pairs DZ-pairs
    1483     2788

Variance decomposition:
  Estimate 2.5%    97.5%
C 0.42274  0.39805 0.44743
E 0.57726  0.55257 0.60195


                         Estimate 2.5% 97.5%
Broad-sense heritability 0        0    0

                       Estimate 2.5%    97.5%
Correlation within MZ: 0.42274  0.39774 0.44711
Correlation within DZ: 0.42274  0.39774 0.44711

'log Lik.' 5495.683 (df=6)
AIC: -10979.37
BIC: -10941.21
>
> AIC(lnbmi.ae,lnbmi.ce)
         df       AIC
lnbmi.ae  6 -11215.25
lnbmi.ce  6 -10979.37
> #not good at all.
> #We report the AE model.
```

Polygenic model - model selection

Biometric analyses - model selection						
Models	'log Lik.'	df	$-2\Delta X^2$	Δdf	p	AIC
Saturated	5629.137	22				-11214.27
ACE	5613.624	7				-11213.25
ADE	5614.387	7				-11214.77
AE (*)	5613.624	6	1.5276	1	0.2165†	-11215.25
CE	5495.683	6	235.88	1	< 0.0001	-10979.37

- The additive genetic effect A is significant in all models (i.e. CE and E models are significantly worse).
- The ADE model has a slightly better fit than the ACE model in terms of Akaike's criterion having lowest AIC value (given by $-2\ln(L) - 2\text{df}$).
- The AE model is chosen by comparison with ADE being the most *parsimonious* model
- †this p-value is too conservative and can be halved (Dominicus et al. 2006).
- The C component in the ACE model vanishes at zero, otherwise we should report it.

University of Southern Denmark

Genetic Influence on lnBMI

	Polygenic model		
	Number of pairs	Correlation (95% CI)	Heritability (95% CI)
MZ pairs	1483	0.65 (0.62,0.68)	0.64 (0.62,0.67)
DZ pairs	2788	0.30 (0.27,0.34)	AE model

- The biometric polygenic model assuming additive genetic and unique environmental components in lnBMI and adjusting for effects of sex and age gave the best fit to observations.

Overview

University of Southern Denmark

OpenMx theoretical background

- Statistical program for Structural Equation Modelling (SEM)
- The focus in SEM is the basically the covariance matrix, $\Sigma = \Sigma(\theta)$.
- Great many statistical methods can be formulated via SEM.
- 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.
- 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.
- -indeed the purpose of programs OpenMx, Mx, LISREL, M-Plus and others.

Practicals using OpenMx

- SEM is implemented by specifying the mean and covariance structures.
- SEM for twin data has the following structure in OpenMx:

```
Model <- mxModel("name",
    mxModel("MZ",
        mxMatrix(),
        mxAlgebra(),
        mxData( observed=mzData, type="raw" ),
        mxFIMLObjective( covariance, means)
    ),
    mxModel("DZ",
        mxMatrix(),
        mxAlgebra(),
        mxData( observed=dzData, type="raw" ),
        mxFIMLObjective( covariance, means)
    ),
    mxAlgebraObjective(MZ.objective + DZ.objective)
)

ModelFit <- mxRun(Model)
-manage output-
```

Biometric analyses - practicals using OpenMx

Practicals

- To get familiar with OpenMx we carry out the analysis of BMI (omitting covariates).
- Using the script 'twinbmiOpenMx.R', let's
  - -fit univariate saturated model.
  - -constrain to same mean for twin 1 and 2, mz and dz.
  - -then constrain to same mean and variance for twin 1 and 2, mz and dz.
- The ACE, ADE and submodels AE, CE and E may then be fitted and compared as above.

Overview

University of Southern Denmark

Remarks

- Within pair similarity is measured by correlations.
- Correlations are further modelled by genetic and environmental variance components via the polygenic ADCE model.
- For instance, the polygenic ACE model relates to correlations via $\rho_{mz} = h^2 + c^2$ and $\rho_{dz} = \frac{1}{2}h^2 + c^2$.

Heuristics of MZ and DZ correlations

	Interpretation		
Relation	Genetics	Environment	Examples
$\rho mz > 4\rho dz$	Epistasis		albinism
$\rho mz > 2\rho dz$	Genetic dominance D		
$\rho mz = 2\rho dz$	Additive effect A (mono- or polygenic) and small D	Small C	BMI
$2\rho dz > \rho mz > \rho dz$	Additive genes A	Shared environment C	longevity
$\rho mz = \rho dz > 0$	No genetic effect	C	
$\rho mz = \rho dz = 0$	No genetic effect	No familial aggregation	

Remarks

- How to do ususal exposure-outcome analysis with twins treated as singletons?
- Regression model for the response of i'th individual in j'th pair:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_j + \epsilon_{ij},$$

- For inference, ie., konfidence and tests, independent observations are needed and u_j models the dependence in twin pairs giving adjusted inference.
- This may also be achieved by robust variance estimation using independence between pairs or similarly by generalised estimating equations (gee). (Implemented in standard software, eg. R and Stata).

Overview

University of Southern Denmark

Aims of multivariate twin analyses

- 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?

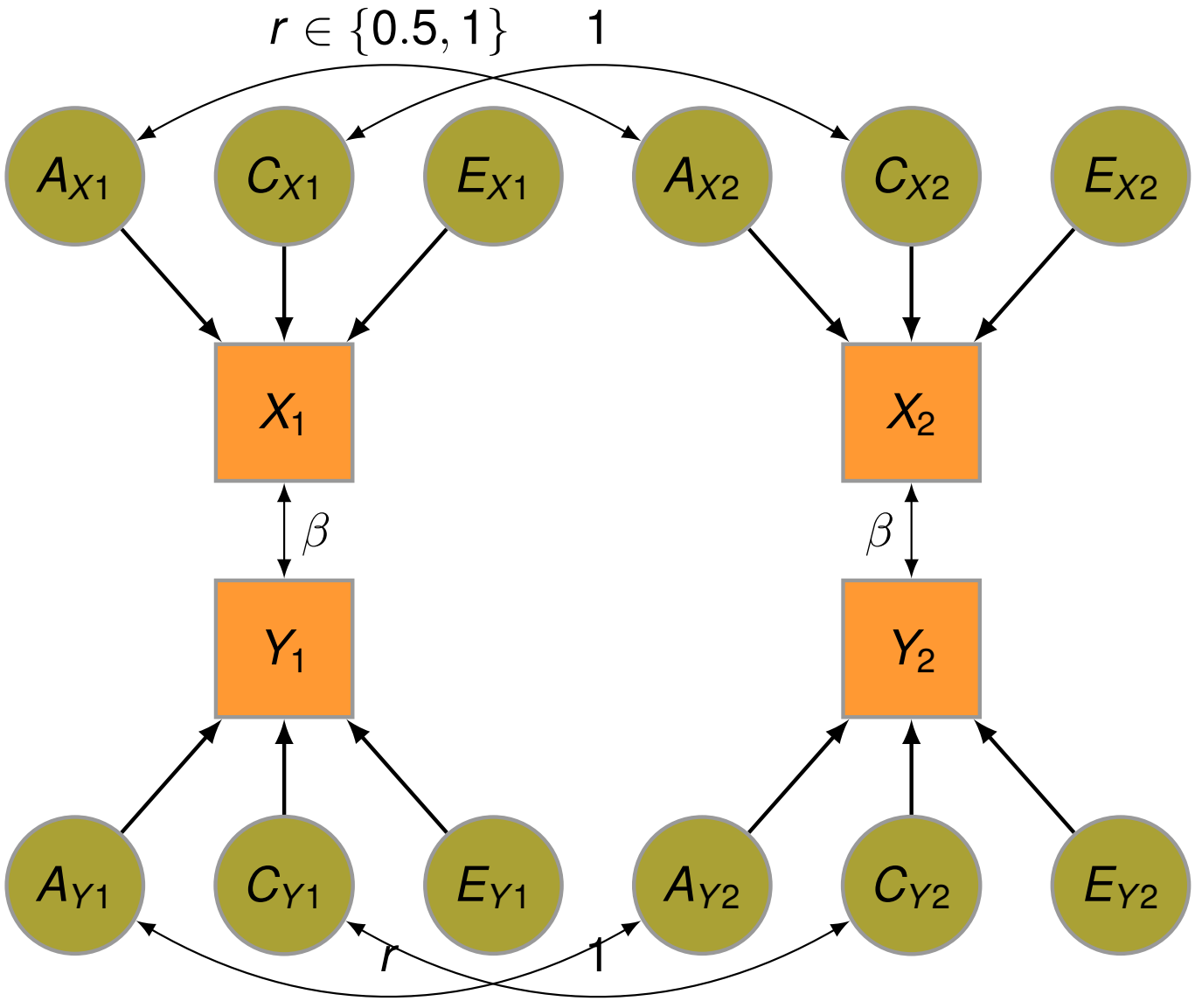
$$\begin{cases} Y = \text{Genes} + \text{Environment} \\ \Sigma_Y = \Sigma_{\text{Genes}} + \Sigma_{\text{Environment}} \end{cases}$$

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

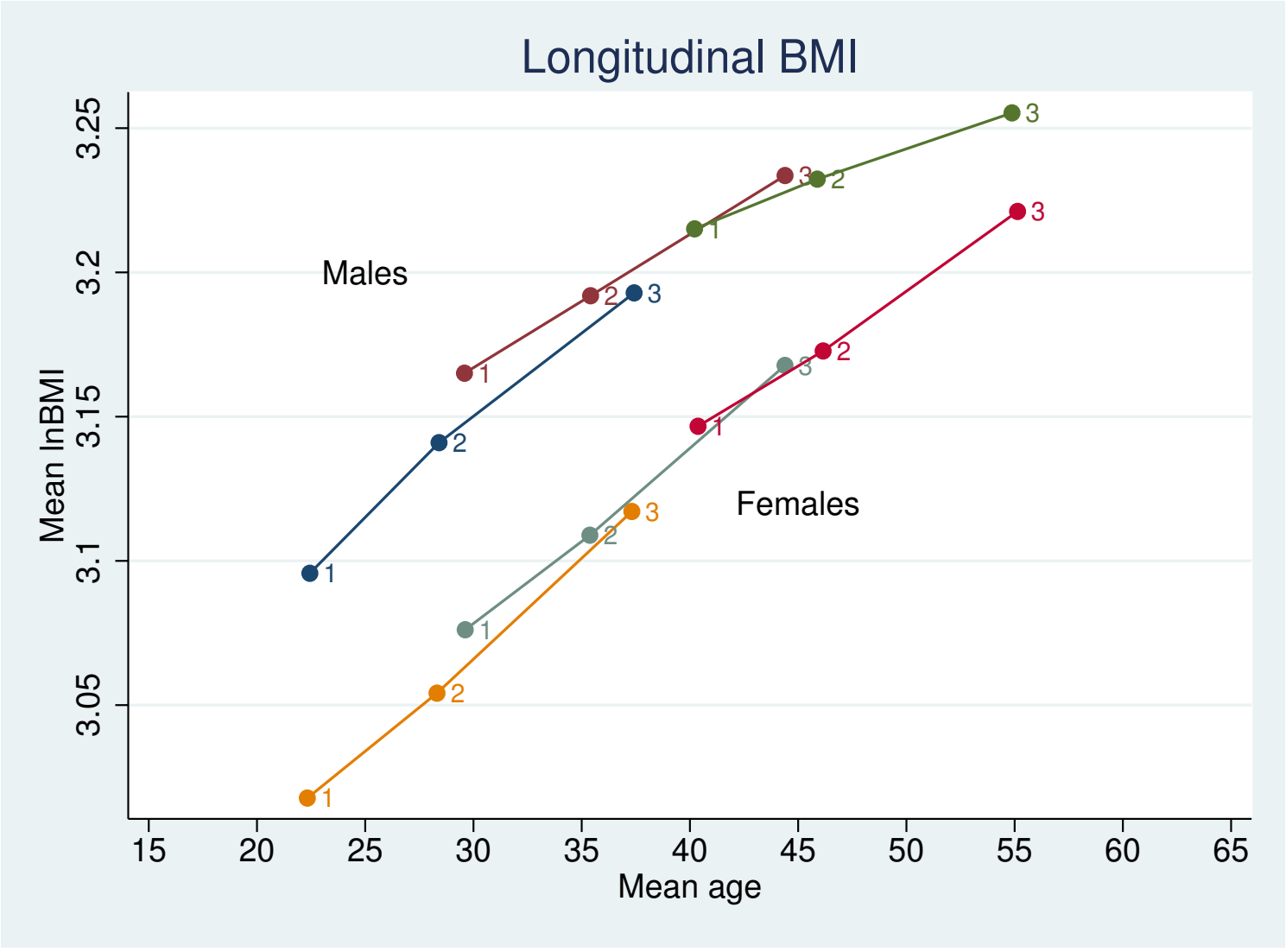
Scope of study

- 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).

SEM - Path Diagram representation



Longitudinal twin data



-genetic influence on change in lnBMI? (to follow later in course)

Aims in Time to Event Twin Studies

Effect?

Exposure$\rightarrow$Outcome

- Outcome: Time to occuence of event. Event may not occur - can be censored at follow-up.
- What is the contribution of genetic and environmental factors to the variation in risk of 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?
- How does this influence vary with time?

University of Southern Denmark

Overview

University of Southern Denmark

Appendix: Correlation in twins

- **Principle:** We choose parameters that makes our observations most likely.
- First, the probability of observing the values (y_{1j}, y_{2j}) in $j'th$ pair given parameters is

$$f((y_{1j}, y_{2j})|\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \exp\{\frac{-1}{2(1-\rho^2)} Q(y_{1j}, y_{2j})\},$$

where

$$Q(y_{1j}, y_{2j}) = (\frac{y_{1j}-\mu_1}{\sigma_1})^2 - 2\rho(\frac{y_{1j}-\mu_1}{\sigma_1})(\frac{y_{2j}-\mu_2}{\sigma_2}) + (\frac{y_{2j}-\mu_2}{\sigma_2})^2$$

- Second, since pairs of observations are independent the likelihood of all data is given by

$$L = \prod_{j=1}^{n} f((y_{1j}, y_{2j})|\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$$

.

- Finally we maximize (the logarithm) of this function, known as the log likelihood function to obtain the parameters, in particular the correlation, for which our observations are most likely.

Back