

# To Be or Not to Be “Genetic”

## Introduction

---

## Basic concepts

---

- Hereditary versus inborn
    - Early life “environments” may be attributed to inborn
  - Hereditary versus (biological) genetic
    - Environments shared by a family can resemble genetics
  - (overall) genetic versus specific genetic markers (e.g. SNP)
    - Genetic parameters estimated from family relationships are **OVERALL** genetic effects from ANY hereditary mechanisms
- 

3

## Methodology Workshop Plans

---

- Current Workshop
    - Estimating Genetic-Environmental Components and Epidemiologic Approach
  - Next Topics
    - Gene Identification by GWA and Linkage
    - Population Genetics and Advanced Topics
- 

4

# Topics Overview

---

- Heritability estimation by variance component methods
  - Familial aggregation
  - Cotwin & sibling regression method
  - Random effect model
  - Intraclass Correlation
  - Estimating genetic variation among unrelated (with genetic markers)
- 

5

## Pedigree Database

---

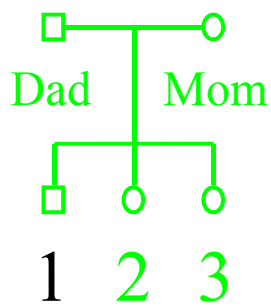
6

# Pedigree data

---

## □ Limitation of conventional DB

- “Relationship Code” is NOT FIXED and changes as the relationship varies.
- Only a “matrix structure” can capture entire relationship codes



	D	M	1	2	3
D	X	sp	o	o	o
M	sp	X	o	o	o
1	D	M	X	sb	sb
2	D	M	sb	X	sb
3	D	M	sb	sb	X

## Thus, Basic Structure of Pedigree Data

---

ID	Father ID	Mother ID	Fam ID	Age	sex
105	0	0	15	67	1
106	0	0	15	65	2
107	105	106	15	39	1
108	105	106	15	38	2
109	105	106	15	35	2
110	107	197	15	11	1

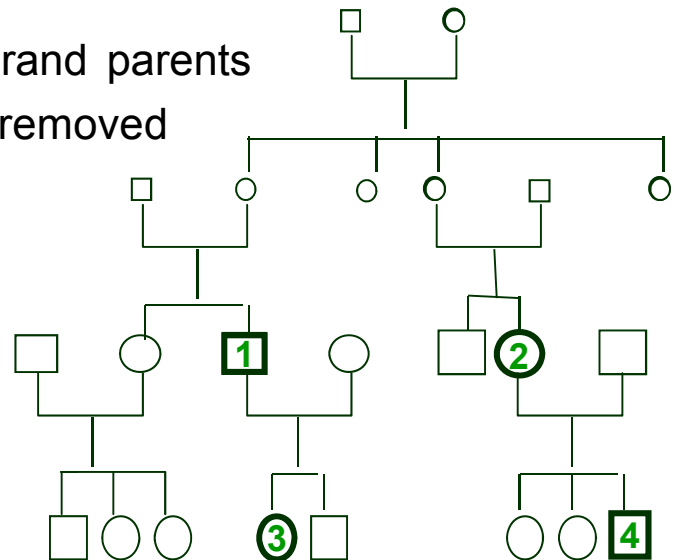
# Family Relationship

- 0 degree – MZ twin
- 1<sup>st</sup> degree – Parent-Offspring / Sib-Pairs
- 2<sup>nd</sup> degree – Grandparents-Offspring / half-sibs / Avuncular (uncle & aunt)
- 3<sup>rd</sup> degree – cousins, great-grand parents
- 4<sup>th</sup> degree – 1<sup>st</sup> cousin once-removed
- 5<sup>th</sup> degree – 2<sup>nd</sup> cousin

Between 1) & 2)?

2) & 3)?

3) & 4)?



9

# Intrapair Correlations

- First things to look at!!
- Analogous to **univariate analysis** in Modeling
  - IC(spouse): control (familial environment)
  - IC(P-O): F-S / F-D / M-S / M-D
    - Difference between F- and M- ?
  - IC(Sib) : B-B / B-S / S-S
    - Difference between B-B and S-S?
    - Difference between P-O and Sibs?
  - IC(MZ) / IC (DZ)
    - Difference between DZ and Sibs?
  - IC (other pairs if available)

# Biometric Dissection of phenotypic variance

---

11

## Mean, variance, covariance

---

### 1. Mean ( $X$ )

$$\mu(X) = \frac{\sum_i x_i}{n}$$

$$\mu = E(X) = \sum_i x_i f(x_i)$$

**X**  
x<sub>1</sub>  
x<sub>2</sub>  
x<sub>3</sub>  
x<sub>4</sub>  
...  
x<sub>n</sub>

# Means, Variances and Covariances

---

## 2. Variance (x) and Covariance (x,y)

$$\begin{aligned} \text{Var}(X) &= E(X - \mu)^2 \\ &= \sum_i (x_i - \mu)^2 f(x_i) \end{aligned}$$

X	X-μ	(X-μ) <sup>2</sup>
x <sub>1</sub>	x <sub>1</sub> -μ	(x <sub>1</sub> -μ) <sup>2</sup>
x <sub>2</sub>	x <sub>2</sub> -μ	(x <sub>2</sub> -μ) <sup>2</sup>
x <sub>3</sub>	x <sub>3</sub> -μ	(x <sub>3</sub> -μ) <sup>2</sup>
x <sub>4</sub>	x <sub>4</sub> -μ	(x <sub>4</sub> -μ) <sup>2</sup>
...	...	...
x <sub>n</sub>	x <sub>n</sub> -μ	(x <sub>n</sub> -μ) <sup>2</sup>

$$\begin{aligned} \text{Cov}(X, Y) &= E(X - \mu_X)(Y - \mu_Y) \\ &= \sum_i (x_i - \mu_X)(y_i - \mu_Y) f(x_i, y_i) \end{aligned}$$

X	Y	X-μ <sub>X</sub>	Y-μ <sub>Y</sub>
x <sub>1</sub>	y <sub>1</sub>	x <sub>1</sub> -μ <sub>X</sub>	y <sub>1</sub> -μ <sub>Y</sub>
x <sub>2</sub>	y <sub>2</sub>	x <sub>2</sub> -μ <sub>X</sub>	y <sub>2</sub> -μ <sub>Y</sub>
x <sub>3</sub>	y <sub>3</sub>	x <sub>3</sub> -μ <sub>X</sub>	y <sub>3</sub> -μ <sub>Y</sub>
x <sub>4</sub>	y <sub>4</sub>	x <sub>4</sub> -μ <sub>X</sub>	y <sub>4</sub> -μ <sub>Y</sub>
...	...	...	...
x <sub>n</sub>	y <sub>n</sub>	x <sub>n</sub> -μ <sub>X</sub>	y <sub>n</sub> -μ <sub>Y</sub>

## Covariance Algebra

---

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

$$\text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$$

Forms Basis for Path Tracing Rules

---

# Covariance and Correlation

---

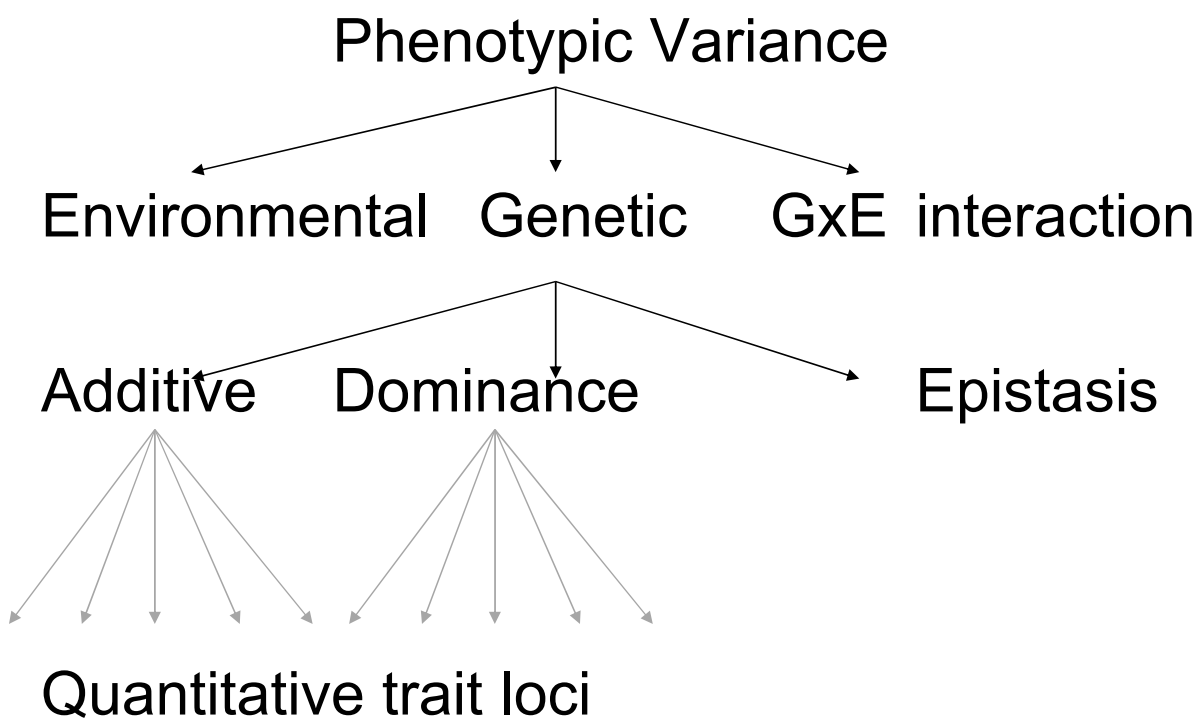
$$r_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Correlation is covariance scaled to range [-1,1].

---

## Components of variance

---



# (Unmeasured) Environmental components

---

## □ Shared (C)

- Correlation = 1

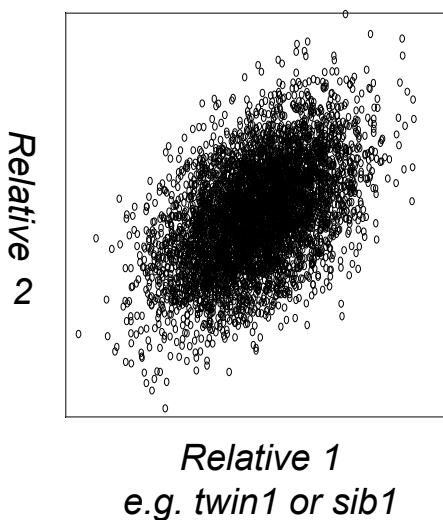
## □ Nonshared (E)

- Correlation = 0

---

## Familial Covariation, Var-Cov matrix

---



*Bivariate normal distribution*

$$X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{matrix} & \text{sib1 or twin1} \\ \text{sib2} & \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_2^2 \end{bmatrix} \\ \text{twin2} & \end{matrix}$$

**Cov** – exactly same  
**Var** – assumption to be the same

# Implied covariance matrices

$$\Sigma_{MZ} = \begin{bmatrix} a^2 + c^2 + e^2 & \\ & a^2 + c^2 + e^2 \end{bmatrix}$$

$$\Sigma_{DZ} = \begin{bmatrix} a^2 + c^2 + e^2 & \\ \frac{1}{2} a^2 + c^2 & a^2 + c^2 + e^2 \end{bmatrix}$$

**Resemblance (=Cov) = A+C + (D)**

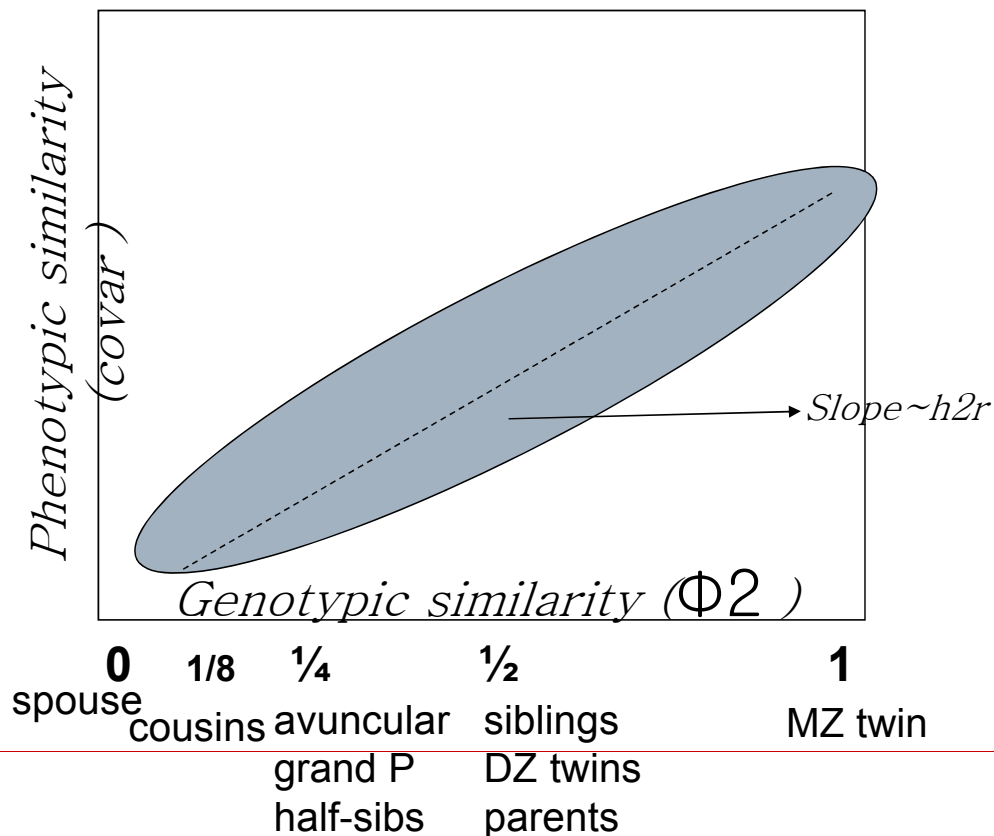
**Difference : (= Var – Cov) = E**

## Covariances between family relationships

Relationship	Covariance (within pair)	Difference (between pair)	ACDE notation	Kinship (Φ <sup>2</sup> )
MZ twin (self)	$a^2+c^2$	$e^2$	A+C	1.0
DZ / Sib	$\frac{1}{2} a^2+ c^2$	$\frac{1}{2} a^2+e^2$	$\frac{1}{2}A+C$	0.5
Parents- offspring	$\frac{1}{2} a^2+c^2$	$\frac{1}{2} a^2+e^2$	$\frac{1}{2} A+C$	0.5
Avuncular(2 <sup>nd</sup> )	$\frac{1}{4} a^2+c^2$	$\frac{3}{4} a^2+e^2$	$\frac{1}{4}A+C$	0.25
Grandparent s (2 <sup>nd</sup> )	$\frac{1}{4} a^2+c^2$	$\frac{3}{4} a^2+e^2$	$\frac{1}{4}A+C$	0.25
Cousins (3 <sup>rd</sup> )	$\frac{1}{8} a^2+c^2$	$\frac{7}{8} a^2+e^2$	$\frac{1}{8} A+C$	0.125
Spouses	$c^2$	$a^2+e^2$	C	0.0

# Graphic understanding of heritability

---



21

## General risk factor studies

---

- Conventional Epi studies
    - e.g. lean body mass and bone mineral density (dissecting “G or E” is not a topic)
    - e.g. hs-CRP and insulin sensitivity
  - Family relationship is a “nuisance” rather than a tool
    - Correlation structure should be adjusted – e.g. random effect model
- 

22

## Must we have “family relationship”?

---

- ❑ With genetic markers (SNP, etc) certain SW tools allow “genetic variation (GV)” attributable to **ALL MEASURED MARKERS** (not the effect of the markers)
- ❑ The gap between heritability and GV from specific markers?
- ❑ GCTA is a pioneer program for this purpose

---

23

## Heritability Analysis using SOLAR



서울대학교 보건대학원 유전체역학교실  
양윤주



## Table of contents

Heritability Analysis  
using SOLAR

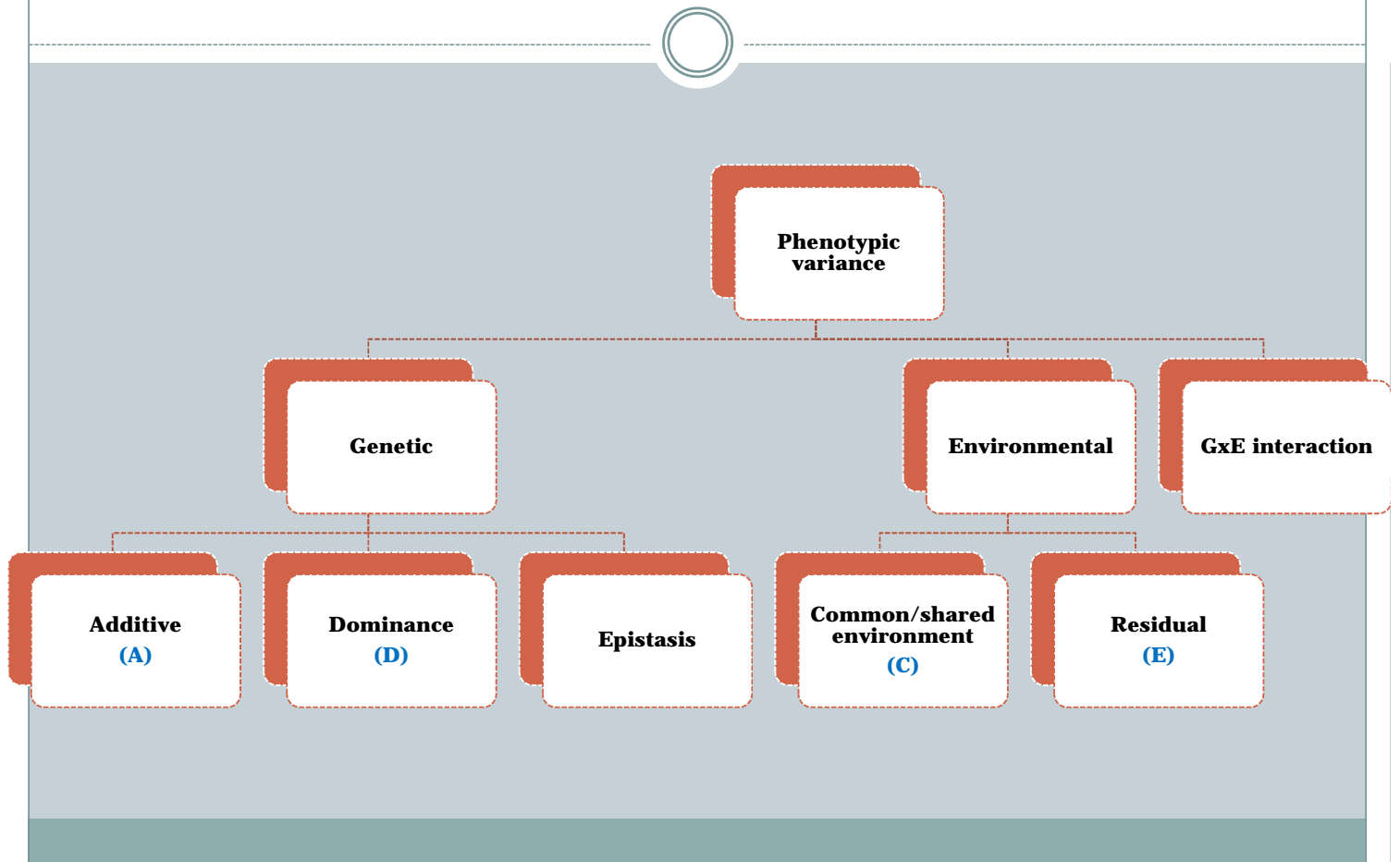
- Basics to Heritability
- Introduction to SOLAR
  - Preparation for SOLAR execution
- SOLAR modeling
  - AE model and ACE model (Ex.1 and 2)
- Tips for effective analysis (Ex.3)
- Genetic correlation analysis (Ex.4)

## Heritability( $h^2$ )



- **Heritability** is the proportion of phenotypic variation in a population that is due to genetic variation between individuals
- Phenotypic variation among individuals may be due to **genetic** and/or **environmental** factors.
- $\text{Var}(P) = \text{Var}(G) + \text{Var}(E) + 2 \text{Cov}(G, E)$

# Components of variance



## Heritability( $h^2$ )

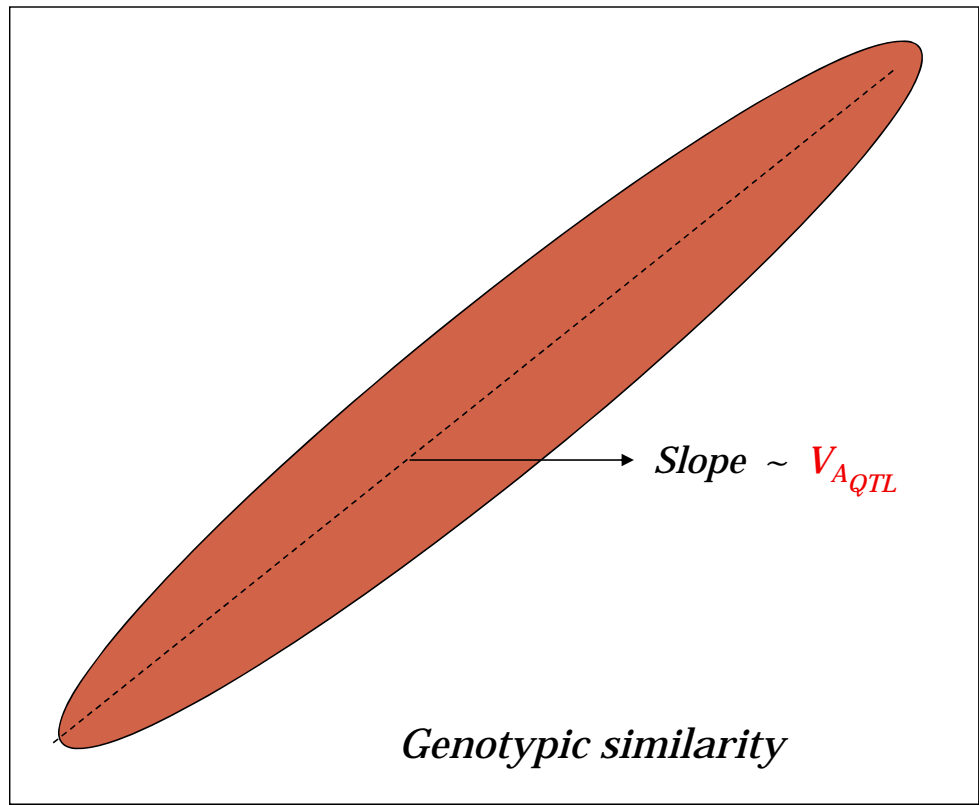
- Heritability analyses estimate the relative contributions of differences in **genetic** and **non-genetic** factors **to the total phenotypic variance** in a population.
- Assuming  $Cov(G,E)$  can be controlled to 0,

**Broad sense heritability**    **Narrow sense heritability**

$$H^2 = \frac{Var(G)}{Var(P)}$$

$$h^2 = \frac{Var(A)}{Var(P)}$$

Phenotypic similarity



0    0.125    0.25    0.5    1

spouse    cousins    Avuncular  
grand P  
half-sibs    Siblings    DZ twins    MZ twins  
parents

## almost 100% heritability

Intraclass-correlation between

spouses: 0.01

cousins: 0.13

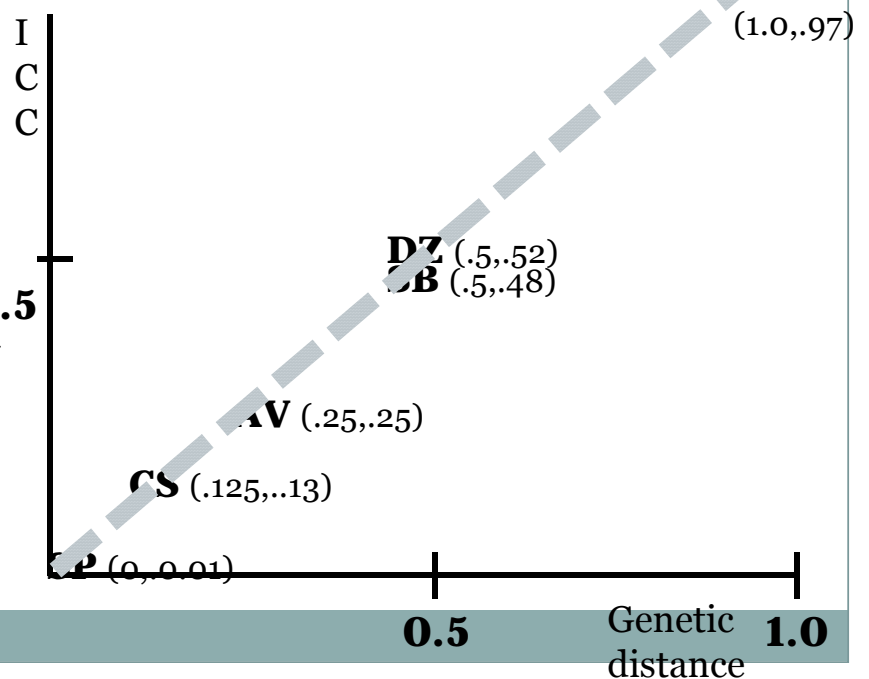
avunculars: 0.25

siblings: 0.48

dizygotic twins: 0.52

monozygotic twins: 0.97

the slope is the heritability



# SOLAR(Sequential Oligogenic Linkage Analysis Routines)



- **SOLAR** is a package of software to perform several kinds of statistical genetic analysis, including linkage analysis, quantitative genetic analysis, and covariate screening.
- <http://www.biostat.wustl.edu/genetics/geneticssoft/manuals/solar210/01.chapter.html>

## SOLAR 실행 \_server-based



- **SSH & Linux**
- SSH secure cell로 서버에 접속하기  
Quick Connect  
>>> Host name: 147.47.67.171
- Basic Linux commands
  - ls** listing of directory contents
  - top** show the latest works being performed
  - more** allow file contents or piped output to be sent to the screen one page at a time
  - cd** change directory
  - chmod** change file access permissions (x: execute or run the file as a program)
  - cp** copy files

# SOLAR modeling\_preparation for HA

- Requires 2 types of file (to calculate  $h^2$ )

- Phenotype file
- e.g. BMI

ID	age	smoke	TW1_f089	TW1_f089_PY	FTND
476588	63	1	1		0
9244659	61	3	1		7
4712412	55	1	1		0
1533638	51	3	1		2
4139052	53	1	1		0
7825997	18	1	1		0

- Pedigree file

FAMID	ID	FA	MO	SEX	MZTWIN	DZTWIN_y	DZTWIN	HHID
476588	476588	0	0	1				1000001
476588	9244659	0	0	2				1000001
9244659	4712412	0	0	1				1000002
7825997	1533638	0	0	1				1000006
583106	4139052	0	0	1				1000007
4519469	7825997	0	0	1				1000008
4519469	583106	0	0	2				1000008
7125561	4519469	0	0	1				1000011

## SOLAR \_preparation

- 파일 실행

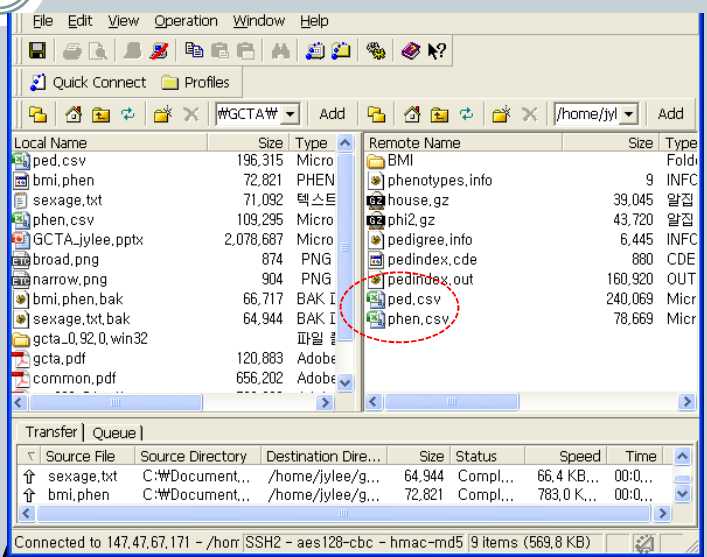
>load ped name.csv

>load phen name.csv

예제)

> load ped obesity\_ped.csv

> load phen obesity\_phen.csv



# SOLAR\_preparation



- Files made after loading a pedigree file
  - house.gz, phi2.gz
  - pedigree.info, pediindex.cde, pedindex.out
- Files made after loading a phenotype file
  - phenotypes.info

# SOLAR modeling



- A (additive genetic)
- C (common / shared environment)
- E (error)

## AE model

```
Solar > model new
Solar > trait BMI
Solar > cov age sex
Solar > polygenic –all (/–screen) +
options
```

## ACE model

```
Solar > model new
Solar > trait BMI
Solar > cov age sex
Solar > house
Solar > polygenic –all (/–screen) +
options
```

# Exercise 1. AE and ACE modeling



- Choose one of obesity-related traits i.e. BMI, LDL, triglyceride etc. and analyze the heritability of it with AE and ACE model respectively.
- Let us interpret the result.

```
*****
*                               Summary of Results                               *
*****
```

```
Pedigree:   obesity_ped.csv
Phenotypes: obesity_phen.csv
Trait:      BMI                               Individuals: 3069
```

```
1)          H2r is 0.6945588  p = 8.8221385e-101 (Significant)
           H2r Std. Error: 0.0260236
```

```
C2 is 0.0000000
```

Since it was zero, the C2 parameter has been deleted.  
To keep C2 parameters even when they are all zero,  
use the -keephouse option.

```
2)          Proportion of Variance Due to All Final Covariates Is
           0.0857784
```

```
Output files and models are in directory BMI/
Summary results are in BMI/polygenic.out
Loglikelihoods and chi's are in BMI/polygenic.logs.out
Best model is named poly and null0 (currently loaded)
Final models are named housepoly, house, poly, spor, nocovar
```

```
Warning: 3) Residual Kurtosis is 3.0311 which is too high.
See note 5 in "help polygenic".
```

# Exercise\_interpretation



- **SOLAR command**

- Covariate? Set up the beta parameters
- Polygenic? Set up the variance parameters

1) Confidence interval =  $\widehat{h^2} \pm z_{\alpha/2} \text{S.E.}$  e.g.  $z_{\alpha/2} = 1.96$  for 95% CI.

2) SOLAR result shows variance percent explained by total covariates counted.

3) if kurtosis is too high, See p.19.

## ACE modeling with different HHID



- Our data has three kinds of HHID, namely HHID, HHID\_sib and HHID\_gen

- HHID : members of same family has a same number.
- HHID\_sib : siblings who have same parent got a same number.
- HHID\_gen : In a family, each generation is numbered differently.

- How to change the current HHID.

```
solar> field HHID(default) HHID_sib/HHID_gen  
solar> load ped ped.csv  
solar> load phen phen.csv
```

Or modify the original file ( change the name of variables)

# ACE modeling with different HHID



FID	ID	FA	MO	SEX	MZTW	HHID	HHID_sib	HHID_gen	
3769850	9638891	3769850	3860627		2	1001	1000001	2	1002001
3769850	5332573	3769850	3860627		2	1001	1000001	2	1002001
3860627	5494079	6680787	5772595		2	1002	1000002	4	1002002
3860627	1343600	6680787	5772595		2		1000002	4	1002002
3860627	2865522	6680787	5772595		1		1000002	4	1002002
3860627	9359710	6680787	5772595		2	1002	1000002	4	1002002
6680787	2307396	1771914	442392		1	1003	1000003	6	1002003
6680787	1530397	1771914	442392		1	1003	1000003	6	1002003
6680787	9123209	4826505	442392		1		1000003	7	1002003
8225394	3252990	9633746	2860961		2		1000004	9	1002004
8225394	3084132	9633746	2860961		2		1000004	9	1002004
8225394	3072408	6584464	3252990		2	1004	1000004	10	1003004
8225394	9797325	6584464	3252990		2	1004	1000004	10	1003004

## Exercise 2. AE and ACE modeling(2)



- Do Exercise 1 again with HHID\_sib.
- Let us interpret the result.

## terms



### **solar> model**

Pvar : phenotypic variance

I : identity matrix

Phi2 : 2\*kinship coefficient

Delta7 : dominance effect

h2r : total additive genetic heritability

e2 : residual genetic variance

c2 : common/shared environment

## Tips for mitigating kurtosis



1) Use tdist command before polygenic statement. e.g.

```
solar > model new
```

```
solar > trait BMI
```

```
solar > cov age sex
```

```
solar > tdist
```

```
solar > polygenic
```

2) Normalizing trait e.g.

```
solar> model new
```

```
solar> define i_BMI = inormal BMI
```

```
solar> trait i_BMI
```

```
solar> cov age sex
```

```
solar> house
```

```
solar> polygenic -screen -all
```

## Tips for saving model



- Save a model

```
>>save model BMI/a
```

```
>>ls BMI
```

```
AE_BMI.mod housepoly.mod nocovar.mod etc.
```

```
>>load model BMI/a
```

## Tips for the Effective execution



### Running SOLAR with batchfiles

(예) solar EOF<<  
model new  
trait BMI  
cov age sex  
polygenic -screen  
model new  
trait SBP  
cov age sex  
polygenic -screen  
model new  
trait DBP  
cov age sex  
polygenic -screen  
quit  
EOF

```
>>>이 script를 하나의 파일로 저장한다 (name.cmd)  
>>> 서버에 업로드한 후  
>>> SOLAR를 실행하고, ped, phen 파일을 로드한 후에  
>>> SOLAR에서 나간 후 chmod +x name.cmd  
>>> ./name.cmd > out.log
```

같은 모델로 여러 trait을 분석하고자 할 때나 옵션만 변경해서 실행하고자 하는 경우, 한 분석이 끝나기를 매번 기다릴 필요 없이 한 번에 실행

## Exercise 3. batchfile usage



- Make a batchfile for the analyses you did in both Ex.1 and 2. Then execute in server and compare the result with previous ones.

## Genetic correlation



- $\rho_p = \rho_g + \rho_e$
- ∴ Phenotypic correlation within a individual can be divided into genetic correlation and environmental correlation.
- Correcting the equation above for related individuals

$$\rho_p = \rho_g \sqrt{h_1^2} \sqrt{h_2^2} + \rho_e \sqrt{(1 - h_1^2)} \sqrt{(1 - h_2^2)}$$

## Genetic correlation\_Definition



- **Genetic correlation** is the proportion of variance that two traits share due to genetic causes. (pleiotropy)
- $\rho_g > 0$  → two traits are influenced by common genes

## Genetic correlation



- Bivariate analysis using solar e.g. BMI, LDL  
→ extend to multivariate analysis

```
solar> model new
```

```
solar> trait BMI LDL (Just specify two traits here!)
```

```
solar> cov age sex
```

```
solar> polygenic
```

## Exercise 4. calculate the genetic correlation



- Let us estimate the genetic correlation between low-density-lipoprotein and body mass index.

- $$\rho_p = \rho_g \sqrt{h_{BMI}^2} \sqrt{h_{LDL}^2} + \rho_e \sqrt{(1 - h_{BMI}^2)} \sqrt{(1 - h_{LDL}^2)}$$

## Familial aggregation의 평가방법과 응용

이동훈

# Familial aggregation

- Familial aggregation of diseases is generally taken as evidence for **the existence of a genetic etiologic mechanism, environmental factors common to family members**, or a combination of both
- An initial goal for many genetic epidemiological studies is to demonstrate familial aggregation of a disease

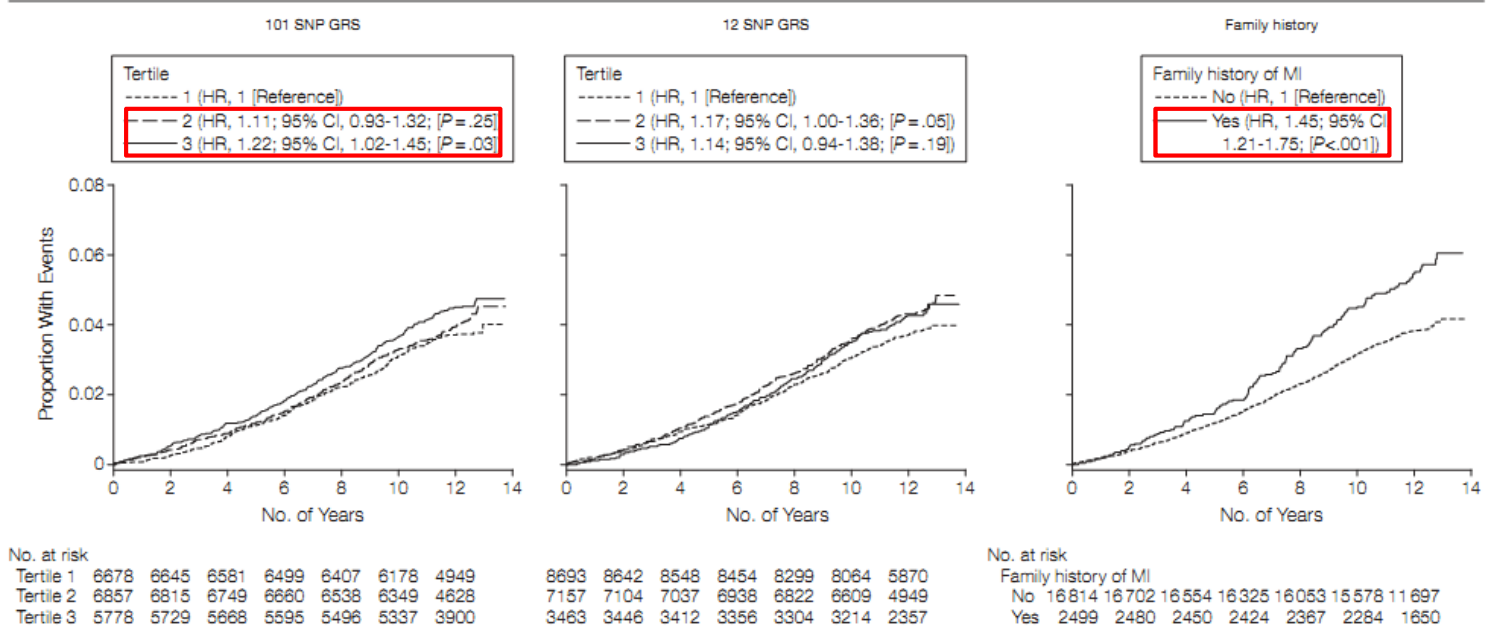
## Risk prediction of disease from Family History

- Genetic risk prediction has raised a lot of attention recently, and many companies are already offering such services
- However, family history is ignored in genetic risk prediction by these companies
- Family history can be obtained easily and at no extra cost

# GWAS (SNP) vs. Family History

- GWAS (SNP)
  - Only genotyped individuals
  - Specific markers
  - Cost ↑
- Family History
  - Overall genetic factors + Shared environmental factors
  - Prediction ↑

**Figure 1.** Cumulative Incidence of Cardiovascular Events by Genetic Risk Score (GRS) Tertile and Family History of Myocardial Infarction (MI)



For the 101 single-nucleotide polymorphisms (SNPs) GRS tertile 1, the mean was 95 (range, 73-99); tertile 2, the mean was 102 (range, 100-105); tertile 3, the mean was 110 (range, 106-125). For the 12 SNP GRS tertile 1, the mean was 9 (range, 4-10); tertile 2, the mean was 11 (range, 11-12); tertile 3, the mean was 14 (range, 13-19).

- The genetic risk score was not associated with cardiovascular disease risk
- In contrast, self-reported family history remained significantly associated with cardiovascular disease in multivariable models

# Limitations of simple Family History

- **Family size**
  - 가족크기가 고려 안됨
    - 5명이 있는 가족에 고혈압 환자 2명
    - 10명이 있는 가족에 고혈압 환자 2명
- **Age**
  - 같은 가족관계라도 **연령**이 다를 때 적용 문제
    - 80세 노인 5명이 있는 가족에 치매 환자 2명
    - 50세 성인 5명이 있는 가족에 치매 환자 2명
- **Relative type**
  - 다양한 **가족관계** 별로 위험도가 다르게 나올 수 있음
    - 내 형제가 암에 걸렸을 때 나의 위험도
    - 내 부모가 암에 걸렸을 때 나의 위험도

## Familial aggregation 평가 방법

Design		Measure
Case-control	Population-based	Odds ratio (OR)
	Family-based	
Cohort	Population-based	Standardized incidence ratio (SIR)
	Family-based	
	Family-based	Family History Score (FHS)
	Family-based	$\lambda_R$

# The conventional case-control design

	Family history		
	Positive	Negative	
Cases	50	55	105
Controls	23	56	79
	73	111	184

- Compare the prevalence of a family history of the disease b/w cases and controls
- 47.6% ( $=50/105$ ) of cases had a positive family history
- 29.1% ( $=23/79$ ) of controls
- OR = 2.21

## Family case-control design

	Disease status		
	Affected	Unaffected	
Case relative	71	173	244
Control relative	29	134	163
	100	307	407

- 29% ( $=71/244$ ) of case relatives had a disease
- 18% ( $=29/163$ ) of control relatives
- OR = 1.90

# OR 평가방법의 장단점

- 장점
  - 쉽고 간단하게 계산 가능
- 단점
  - family size, age, relative type 모두 고려 안됨

## Standardized incidence ratio (SIR)

- The expected number of disease was calculated from the **age- sex- and period-specific incidence rates**
- SIR was calculated from the ratio of observed number (O) to expected number (E) of disease
- $$SIR = \frac{\text{Number of observed cases}}{\text{Number of expected cases}} = \frac{O}{E}$$

(predicted by incidence data)

Table 3  
Relative risk of CRC in relatives according to their age and gender

Relative characteristics	O	E	RR	95% CI	P value	P <sub>trend</sub>
<b>Gender</b>						
Female	53	34.11	1.55	(1.16–2.03)	0.002	
Male	54	35.28	1.53	(1.15–2.00)	0.002	NS
<b>Age (years)</b>						
≤50	10	4.84	2.07	(0.99–3.80)	0.03	
51–60	17	10.16	1.67	(0.97–2.68)	0.03	NS
61–70 years	28	21.87	1.28	(0.85–1.85)	NS	
> 70	52	32.52	1.60	(1.19–2.10)	<0.001	
<b>Gender and age (years)</b>						
<b>Female</b>						
≤50	4	2.32	1.72	(0.46–4.41)	NS	
51–60	5	4.49	1.11	(0.36–2.60)	NS	NS
61–70	14	9.80	1.43	(0.78–2.40)	NS	
> 70	30	17.49	1.72	(1.16–2.45)	0.004	
<b>Male</b>						
≤50	6	2.52	2.38	(0.87–5.18)	NS	
51–60	12	5.66	2.12	(1.09–3.70)	0.01	NS
61–70	14	12.07	1.16	(0.63–1.95)	NS	
> 70	22	15.03	1.46	(0.92–2.22)	NS	

NS, non-significant; O, observed; E, expected; RR, relative risk; 95% CI, 95% Confidence Interval.

Table 2 Familial relative risk for breast cancer by ER status of the index case tumour

	OBS	EXP	FRR	95%CI
<b>Mothers</b>				
All cases	663	398.81	1.66	1.53 - 1.79
ER-negative	84	45.09	1.86	1.46 - 2.27
ER-positive	332	199.36	1.67	1.49 - 1.85
<b>Sisters</b>				
All cases	416	207.25	2.01	1.81 - 2.21
ER-negative	37	22.97	1.61	1.05 - 2.17
ER-positive	226	106.81	2.12	1.83 - 2.41
<b>All relatives</b>				
All cases	1079	606.06	1.78	1.68 - 1.89
ER-negative	121	68.06	1.78	1.44 - 2.11
ER-positive	558	306.17	1.82	1.67 - 1.98

CI: confidence interval; ER: estrogen receptor; EXP: expected, FRR: familial relative risk; OBS: observed.

## SIR 평가방법의 장단점

- 장점
  - 비교적 쉽게 계산 가능
  - age, sex 고려 가능
  - family-based data에서는 family size 고려 가능
- 단점
  - age sex-specific incidence rate가 있어야 됨
  - relative type 고려 안됨

# Family History Score

- 실제 관찰된 환자수(O) - 기대되는 환자수(E)
- 기대되는 환자수(E)
  - 해당 성별 연령별 누적 위험도
- FHS
  - = (실제 관찰된 환자수 - 기대되는 환자수)를 표준화한 점수

# Family History Score

- FHS for i-th family

$$Z_i = \frac{(\sum_j O_{ij} - \sum_j E_{ij})}{\{\sum_j E_{ij}(1 - E_{ij})\}^{1/2}}$$

← 실제환자수의 합 - 기대환자수의 합

← 분산 값을 이용하여 표준화

- Cumulative risk

$$E_{ij} = 1 - \exp\left(-\sum_k \lambda_k I_{ijk}\right)$$

← 해당 연령까지의 누적위험도

**TABLE 4. Summary of estimates from proportional hazard and logistic regression analyses by different family history classifications: Cancer Prevention Study II, United States, 1982–1991**

Family history	Proportional hazard analysis		Logistic regression analysis				
	Relative hazard*	95% CI†	Relative risk*	95% CI	R <sup>2</sup>	R <sup>2</sup> <sub>adj</sub>	C statistic
Observed breast cancer cases‡	1.49	1.30–1.73	1.50	1.30–1.73	0.033	0.085	0.73
FHS‡,§	1.10	1.07–1.13	1.10	1.07–1.12	0.036	0.091	0.75
Observed no. of breast cancer cases§							
0	1.00		1.00				
1	1.59	1.36–1.27	1.59	1.36–1.28			
≥2	1.71	0.94–3.09	1.72	0.95–3.10	0.035	0.087	0.73
FHS§							
0	1.00		1.00				
1	1.06	0.78–1.43	1.06	0.78–1.45			
2	1.64	1.27–2.14	1.66	1.29–2.19			
3	2.31	1.80–2.96	2.30	1.80–2.95	0.039	0.099	0.77

\* Estimates were adjusted for menopausal status, age at menarche, age when first living child was born, history of breast cysts, oral contraceptive use, other estrogen use, body mass index, diethylstilbestrol (DES), education, religion, race, alcohol use, smoking status, and, among postmenopausal women, the age at which periods stopped.

† CI, confidence interval; FHS, family history score.

‡ Observed breast cancer cases and FHS used as continuous variables.

§ Observed breast cancer cases and FHS used as categorical variables. All families with positive FHS were divided into 3 equal groups, i.e., 1 = low FHS, first 33%; 2 = medium FHS, second 33%; and 3 = high FHS, third 33%.

## FHS 평가방법의 장단점

- 장점
  - age, sex, family size를 고려 가능
- 단점
  - age sex-specific incidence rate가 있어야 됨
  - relative type 고려 안됨

# Familial aggregation 평가 방법

Design		Measure
Case-control	Population-based	Odds ratio (OR)
	Family-based	
Cohort	Population-based	Standardized incidence ratio (SIR)
	Family-based	
	Family-based	Family History Score (FHS)
	Family-based	$\lambda_R$

## $\lambda_R$ 정의

- $\lambda_R$   
= Familial Relative Risk  
= Recurrence Risk Ratio  
= Familial Recurrence Risk  
= Relative Recurrence Risk  
= Relative Risk
- Define  $\lambda_R$  as the risk ratio for a type R relative of an affected individual compared with population prevalence (K)
- $\lambda_R = K_R/K$

# Familial Relative Risk (FRR) = $\lambda_R$

- The familial relative risk (ratio) of disease is a standard parameter used in genetic analysis to indicate the increased risk of disease in relatives of affected cases compared with the risk of disease in the general population
- $\lambda_{rel} = \frac{P(\text{family member 2} \mid \text{family member 1})}{\text{Life time prevalence in general population}}$
- An increased risk of disease in relatives indicates a host genetic component to susceptibility

$$\lambda_R$$

- $\lambda_R = \frac{P(\text{family member 2} \mid \text{family member 1})}{\text{Life time prevalence in general population}}$

		Family member 1	
		D+	D-
Family member 2	D+	a	b
	D-	c	d

- $P(\text{family member 2} \mid \text{family member 1}) = a/(a+b) = a/(a+c) \quad \because (b=c)$

# $\lambda_R$ 평가방법의 장단점

- 장점
  - relative type에 대해 고려 가능 ( $\lambda_{MZ}$ ,  $\lambda_{Sib}$ , ...)
- 단점
  - age, family size는 고려 안됨

## Familial aggregation 평가방법 정리

	평가방법			
	OR	SIR	FHS	$\lambda_R$
Age	X	O	O	X
Family size	X	$\Delta$	O	X
Relative type	X	X	X	O

# Exercises

<표 1> 연령별·성별 **비만** 유병률

연령	전체			남자			여자		
	유병수	유병률(%)	95%신뢰구간	유병수	유병률(%)	95%신뢰구간	유병수	유병률(%)	95%신뢰구간
20~29	1,471,814	19.4	15.8-23.1	986,146	25.2	19.4-31.1	484,888	13.2	9.3-17.2
30~39	2,474,979	29.0	26.1-32.0	1,661,220	38.0	33.4-42.6	811,742	19.5	16.2-22.9
40~49	2,889,485	35.2	32.0-38.3	1,718,451	41.1	36.4-45.8	1,168,010	29.0	25.1-33.0
50~59	2,157,719	42.0	38.7-45.4	1,053,992	41.0	35.8-46.2	1,106,253	43.1	38.1-48.1
60~69	1,425,564	39.5	35.3-43.8	518,103	31.0	25.1-36.9	910,729	47.0	41.3-52.8
70+	845,502	31.5	27.2-35.8	268,377	27.5	19.5-35.5	579,086	33.9	28.5-39.3
20세 이상	11,264,533	31.5	29.8-33.2	6,207,078	35.1	32.6-37.6	5,061,403	28.0	25.9-30.1
30세 이상	9,804,466	34.8	33.0-36.6	5,219,097	37.9	35.2-40.5	4,580,170	31.8	29.6-34.0

1) 비만 : 체질량지수(Body mass index, kg/m<sup>2</sup>)가 25kg/m<sup>2</sup>이상인 경우  
 2) 유병수 : 2005년 추계인구로 유병수 추정

연령	<b>고LDL콜레스테롤혈증</b>								
	전체			남자			여자		
	유병수	유병률(%)	95%신뢰구간	유병수	유병률(%)	95%신뢰구간	유병수	유병률(%)	95%신뢰구간
20~29	113,800	1.5	0.1-3.0	74,352	1.9	0.0-4.5	44,081	1.2	0.4-0.9
30~39	349,911	4.1	2.8-5.5	275,413	6.3	3.9-8.6	79,093	1.9	0.8-3.1
40~49	591,031	7.2	5.5-8.9	351,216	8.4	5.8-11.1	237,630	5.9	3.7-8.1
50~59	493,193	9.6	7.6-11.7	151,672	5.9	3.3-8.5	336,239	13.1	9.7-16.6
60~69	339,248	9.4	7.1-11.7	96,935	5.8	2.9-8.8	240,277	12.4	8.7-16.1
70+	230,835	8.6	5.8-11.4	46,844	4.8	1.9-7.7	184,487	10.8	6.6-15.0
20세 이상	2,109,865	5.9	5.1-6.7	990,303	5.6	4.5-6.6	1,102,663	6.1	5.0-7.2
30세 이상	2,000,337	7.1	6.2-8.0	922,637	6.7	5.5-7.9	1,065,826	7.4	6.1-8.8

# Example - SIR

```
data ex_data;
set example_data;

/* 성별 연령별 유병률 입력 */
if sex=1 then do;
  if agegp=20 then do; obe_exp = 0.252 ; lip_exp = 0.019 ; end;
  else if agegp=30 then do; obe_exp = 0.38 ; lip_exp = 0.063 ; end;
  else if agegp=40 then do; obe_exp = 0.411 ; lip_exp = 0.084 ; end;
  else if agegp=50 then do; obe_exp = 0.41 ; lip_exp = 0.059 ; end;
  else if agegp=60 then do; obe_exp = 0.31 ; lip_exp = 0.058 ; end;
  else if agegp=70 then do; obe_exp = 0.275 ; lip_exp = 0.048 ; end;
end;

else if sex=2 then do;
  if agegp=20 then do; obe_exp = 0.132 ; lip_exp = 0.012 ; end;
  else if agegp=30 then do; obe_exp = 0.195 ; lip_exp = 0.019 ; end;
  else if agegp=40 then do; obe_exp = 0.29 ; lip_exp = 0.059 ; end;
  else if agegp=50 then do; obe_exp = 0.431 ; lip_exp = 0.131 ; end;
  else if agegp=60 then do; obe_exp = 0.47 ; lip_exp = 0.124 ; end;
  else if agegp=70 then do; obe_exp = 0.339 ; lip_exp = 0.108 ; end;
end;

/* 표준화를 위한 분산 계산 */
obe_pq=obe_exp*(1-obe_exp);
lip_pq=lip_exp*(1-lip_exp);

/* 보정을 위한 연령그룹 생성 */
if agegp in (20 30 40) then agegp2 = 1;
if agegp = 50 then agegp2 = 2;
if agegp = 60 then agegp2 = 3;
if agegp = 70 then agegp2 = 4;
run;
```

```
/* SIR */
proc means data=ex_data sum maxdec=2;
class sex; /* 성별 */
var obe_obs obe_exp lip_obs lip_exp ;
run;

proc means data=ex_data sum maxdec=2;
class agegp2; /* 연령그룹별 */
var obe_obs obe_exp lip_obs lip_exp ;
run;

proc means data=ex_data sum maxdec=2;
class sex agegp2; /* 성별 연령그룹별 */
var obe_obs obe_exp lip_obs lip_exp ;
run;
```

# Example - FHS

```
/* 비만 */
/* 가족별 실제환자수, 기대환자수, 분산의 합 계산 */
proc means data=ex_data noprint;class fid;var obe_obs;output
out=obe_obs(where=(fid~=.) drop=_TYPE_ _FREQ_) sum=obe_obs_sum;run;
proc means data=ex_data noprint;class fid;var obe_exp;output
out=obe_exp(where=(fid~=.) drop=_TYPE_ _FREQ_) sum=obe_exp_sum;run;
proc means data=ex_data noprint;class fid;var obe_pq;output out=obe_pq(where=(fid~=.)
drop=_TYPE_ _FREQ_) sum=obe_pq_sum;run;
data obe;
merge obe_obs obe_exp obe_pq ;
by fid;
obe_std=sqrt(obe_pq_sum); /* 표준편차 계산 */
obe_t = (obe_obs_sum-obe_exp_sum)/obe_std; /* Family History Score 계산 */

/* FHS ranking */
if obe_t<=0 then obe_rank=0;
else if 0<obe_t<=0.517344614 then obe_rank=1;
else if 0.517344614<obe_t<=1.5105424144 then obe_rank=2;
else if 1.5105424144<obe_t then obe_rank=3;

if obe_obs_sum>3 then obe_obs_sum=3;
if obe_t<0 then obe_t=0;
run;
/* FHS 3등분 계산 */
proc univariate data=obe noprint;where obe_t>0;var obe_t;output out=quantile pctlpre=
percent_ pctlpts= 33.33, 66.66;run;
```

```
/* 고LDL */
/* 가족별 실제환자수, 기대환자수, 분산의 합 계산 */
proc means data=ex_data noprint;class fid;var lip_obs;output
out=lip_obs(where=(fid~=.) drop=_TYPE_ _FREQ_) sum=lip_obs_sum;run;
proc means data=ex_data noprint;class fid;var lip_exp;output
out=lip_exp(where=(fid~=.) drop=_TYPE_ _FREQ_) sum=lip_exp_sum;run;
proc means data=ex_data noprint;class fid;var lip_pq;output
out=lip_pq(where=(fid~=.) drop=_TYPE_ _FREQ_) sum=lip_pq_sum;run;
data lip;
merge lip_obs lip_exp lip_pq ;
by fid;
lip_std=sqrt(lip_pq_sum); /* 표준편차 계산 */
lip_t = (lip_obs_sum-lip_exp_sum)/lip_std ; /* Family History Score 계산 */

/* FHS ranking */
if lip_t<=0 then lip_rank=0;
else if 0<lip_t<=1.279676565 then lip_rank=1;
else if 1.279676565<lip_t<=2.1376882107 then lip_rank=2;
else if 2.1376882107<lip_t then lip_rank=3;

if lip_obs_sum>3 then lip_obs_sum=3;
if lip_t<0 then lip_t=0;
run;
/* FHS 3등분 계산 */
proc univariate data=lip noprint;where lip_t>0;var lip_t;output
out=quantile pctlpre= percent_ pctlpts= 33.33, 66.66;run;
```

# Example - FHS

```
proc sort data=ex_data;by fid;run;
data ex_data2;
merge
ex_data(keep=fid sex agegp obe_obs obe_exp lip_obs lip_exp)
obe(keep=fid obe_obs_sum obe_exp_sum obe_t obe_rank)
lip(keep=fid lip_obs_sum lip_exp_sum lip_t lip_rank)
;
by fid;
run;

/* FHS */
/* 비만 */
proc logistic data=ex_data2 descending;
class agegp2(ref='1') sex(ref='1') ;
model obe_obs = obe_obs_sum agegp2 sex /rsquare rl lackfit;
run;
proc logistic data=ex_data2 descending;
class agegp2(ref='1') sex(ref='1') ;
model obe_obs = obe_t agegp2 sex /rsquare rl lackfit ;
run;

/* 고 LDL */
proc logistic data=ex_data2 descending;
class agegp2(ref='1') sex(ref='1') ;
model lip_obs = lip_obs_sum agegp2 sex /rsquare rl lackfit;
run;
proc logistic data=ex_data2 descending;
class agegp2(ref='1') sex(ref='1') ;
model lip_obs = lip_t agegp2 sex /rsquare rl lackfit ;
run;
```

# Example - $\lambda_{Sib}$

```
/* lambda sibling */
proc sort data=ex_data;by fid fa mo;run;
proc transpose data=ex_data(keep=fid id fa mo obe_obs
where=(fa~=0)) out=ex_data_t(where=(col2~=.));
by fid fa mo;
/* 가족번호 부 모를 기준으로 transpose */
var id ;
run;
```

```
data sib;
set /* pair로 만들기 */
ex_data_t(keep=col1 col2 rename=(col1=sib1 col2 =sib2))
ex_data_t(keep=col1 col3 rename=(col1=sib1 col3 =sib2))
ex_data_t(keep=col1 col4 rename=(col1=sib1 col4 =sib2))
ex_data_t(keep=col1 col5 rename=(col1=sib1 col5 =sib2))
ex_data_t(keep=col1 col6 rename=(col1=sib1 col6 =sib2))
ex_data_t(keep=col1 col7 rename=(col1=sib1 col7 =sib2))
ex_data_t(keep=col1 col8 rename=(col1=sib1 col8 =sib2))
ex_data_t(keep=col1 col9 rename=(col1=sib1 col9 =sib2))
ex_data_t(keep=col1 col10 rename=(col1=sib1 col10 =sib2))
ex_data_t(keep=col2 col3 rename=(col2=sib1 col3 =sib2))
ex_data_t(keep=col2 col4 rename=(col2=sib1 col4 =sib2))
ex_data_t(keep=col2 col5 rename=(col2=sib1 col5 =sib2))
ex_data_t(keep=col2 col6 rename=(col2=sib1 col6 =sib2))
ex_data_t(keep=col2 col7 rename=(col2=sib1 col7 =sib2))
ex_data_t(keep=col2 col8 rename=(col2=sib1 col8 =sib2))
ex_data_t(keep=col2 col9 rename=(col2=sib1 col9 =sib2))
ex_data_t(keep=col2 col10 rename=(col2=sib1 col10 =sib2))

.....

ex_data_t(keep=col8 col9 rename=(col8=sib1 col9 =sib2))
ex_data_t(keep=col8 col10 rename=(col8=sib1 col10 =sib2))
ex_data_t(keep=col9 col10 rename=(col9=sib1 col10 =sib2))
;
if sib1="" or sib2="" then delete;
run;
```

# Example - $\lambda_{Sib}$

```
proc sort data=sib;by sib1;run;  
proc sort data=ex_data;by id;run;  
data sib2; /* 각각의 값들 merge */  
merge sib(in=x) ex_data(keep=id obe_obs lip_obs  
rename=(id=sib1 obe_obs=sib1_obe lip_obs=sib1_lip));  
by sib1; if x;  
run;  
proc sort data=sib2;by sib2;run;  
data obe_sib lip_sib; /* 각각의 값들 merge */  
merge sib2(in=x) ex_data(keep=id obe_obs lip_obs  
rename=(id=sib2 obe_obs=sib2_obe lip_obs=sib2_lip));  
by sib2; if x;  
drop tmp;  
if sib1_obe=0 & sib2_obe=1 then do;  
tmp=sib1; sib1=sib2; sib2=sib1;  
sib1_obe=1; sib2_obe=0;  
tmp=sib1_lip; sib1_lip=sib2_lip; sib2_lip=sib1_lip;  
end; output obe_sib;  
if sib1_lip=0 & sib2_lip=1 then do;  
tmp=sib1; sib1=sib2; sib2=sib1;  
sib1_lip=1; sib2_lip=0;  
tmp=sib1_obe; sib1_obe=sib2_obe; sib2_obe=sib1_obe;  
end; output lip_sib;  
run;
```

```
proc freq data=obe_sib;  
table  
sib1_obe * sib2_obe  
sib1_obe * sib2_lip  
/nocol norow nopercent nocum  
;  
run;  
proc freq data=lip_sib;  
table  
sib1_lip * sib2_lip  
sib1_lip * sib2_obe  
/nocol norow nopercent nocum  
;  
run;
```

## Appendix A

- Define the random variable  $X_1$  to be 1 if individual 1 is affected, and 0 if unaffected
- Define  $X_2$  for a related individual 2 of type R
- Population prevalence =  $E(X_1) = K$
- Define  $K_R = E(X_2|X_1=1)$  to be the recurrence risk for a type R relative of an affected individual

- Then the probability that a proband and type R relative are both affected is  

$$K \cdot K_R = E(X_1 X_2) = \text{Cov}(X_1, X_2) + K^2$$
- $K_R = K + (1/K)\text{Cov}(X_1, X_2)$
- Define  $\lambda_R$  as the risk ratio for a type R relative of an affected individual compared with population prevalence
- $\lambda_R = K_R/K = 1 + (1/K^2)\text{Cov}(X_1, X_2)$

- $V_A$  : additive genetic variance
- $V_D$  : dominance variance
- $\phi$  : kinship coefficient
- $\pi$  : the probability the two relatives share two alleles identical by descent

- $\lambda_R = 1 + (1/K^2)(2\phi V_A + \pi V_D)$

- $\lambda_{\text{Spouse}} = 1$  : non-genetic parameter
- $\lambda_{\text{MZ}} = 1 + (1/K^2)(V_A + V_D)$  : 100% genetic
- $\lambda_{\text{Sib}} = 1 + (1/K^2)(\frac{1}{2}V_A + \frac{1}{4}V_D)$
- $\lambda_{\text{Parents}} = 1 + (1/K^2)(\frac{1}{2}V_A)$
- .....

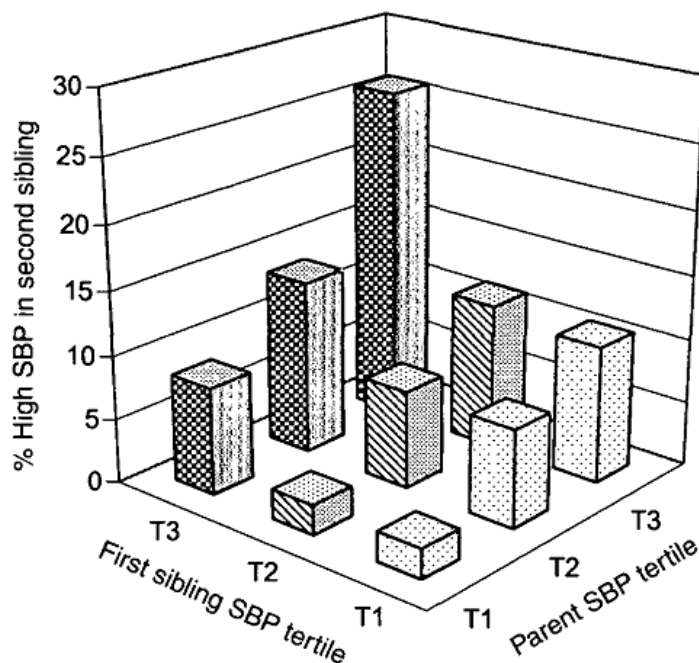
# Appendix B

## Familial correlation

**TABLE 2.** Crude and adjusted\* correlation coefficients of parental and sibling blood pressure, Anqing, China, 1994–1997

	Father	Mother	First sibling
<b>Systolic blood pressure</b>			
Crude value			
Mother	0.277		
First sibling	0.248	0.296	
Second sibling	0.285	0.345	0.416
Adjusted value			
Mother	0.113		
First sibling	0.168	0.206	
Second sibling	0.218	0.216	0.328
<b>Diastolic blood pressure</b>			
Crude value			
Mother	0.150		
First sibling	0.212	0.290	
Second sibling	0.228	0.282	0.373
Adjusted value			
Mother	0.090		
First sibling	0.160	0.192	
Second sibling	0.176	0.174	0.300

\* Adjusted by sex, age, height, weight, smoking status, alcohol consumption, and educational level; all *p* values <0.01.



**FIGURE 1.** Percentages of high systolic blood pressure (SBP) in second siblings in nine groups based on the blood pressure status of the parents and first sibling (defined by the tertile (T) of percentage-predicted SBP), Anqing, China, 1994–1997. High blood pressure was defined as higher than the 90th percentile of the percentage-predicted SBP.

**TABLE 5. Adjusted odds ratios† and 95% confidence intervals of high blood pressure in second siblings, by parental and first-sibling blood pressure tertile,‡ Anqing, China, 1994–1997**

Blood pressure tertile§		Systolic blood pressure		Diastolic blood pressure	
Parents	First sibling	OR¶	95% CI¶	OR	95% CI
<i>Mean of father and mother</i>					
Low	Low	1.0		1.0	
	Middle	1.3	0.3–6.1	3.9*	1.1–14.2
	High	3.9*	1.1–14.4	2.8	0.7–10.7
Middle	Low	3.9*	1.1–14.4	3.2	0.8–11.9
	Middle	3.5	0.9–13.0	3.1	0.8–11.8
	High	5.9**	1.7–20.7	5.5**	1.5–19.4
High	Low	4.3*	1.2–15.6	3.2	0.8–11.9
	Middle	6.7**	1.9–23.5	8.1**	2.4–27.8
	High	14.3**	4.3–48.2	14.4**	4.3–48.2
<i>Mother only</i>					
Low	Low	1.0		1.0	
	Middle	2.6	0.5–13.4	1.8	0.7–5.2
	High	7.2**	1.6–32.6	1.0	0.3–3.6
Middle	Low	5.3*	1.1–25.7	1.5	0.5–4.3
	Middle	5.9*	1.3–27.1	1.5	0.5–4.2
	High	7.9**	1.8–34.9	3.5*	1.3–8.9
High	Low	5.4*	1.2–25.1	0.9	0.3–3.1
	Middle	7.5**	1.7–33.6	3.3*	1.3–8.6
	High	18.3**	4.3–78.6	6.2**	2.5–15.5
<i>Father only</i>					
Low	Low	1.0		1.0	
	Middle	1.2	0.3–3.9	0.8	0.2–2.6
	High	2.0	0.7–6.1	2.8*	1.0–7.3
Middle	Low	1.2	0.4–3.9	2.0	0.7–5.6
	Middle	1.9	0.6–5.7	1.9	0.7–5.4
	High	4.1**	1.5–11.2	2.6	1.0–6.4
High	Low	2.1	0.7–6.4	0.9	0.3–3.1
	Middle	3.3*	1.2–9.3	2.1	0.8–5.8
	High	6.1**	2.3–16.5	5.1**	2.0–12.8

\*  $p < 0.05$ ; \*\*  $p < 0.01$ .

† The odds ratio for the low-low group was used as the reference value (1.0).

‡ Based on the percentage-predicted values.

§ Logistic regression models were used; high blood pressure, >90th percentile of percentage-predicted values.

¶ CI, confidence interval; OR, odds ratio.

**Table 3 – Correlation coefficients ( $r \pm$  standard error) between family members and  $h^2$  estimates for systolic and diastolic arterial pressure**

Degree of family relationship	N of de pairs	SBP*	DBP†
		$r \pm se$	$r \pm se$
Spouses	58	0.24±0.13	0.25±0.13
Father – son	53	0.26±0.14	0.26±0.14
Father – daughter	75	0.32±0.11	0.24±0.12
Mother – son	75	0.01±0.12	0.29±0.11
Mother – daughter	117	0.24±0.09	0.29±0.09
Brother – brother	11	0.35±0.28	0.50±0.24
Sister – sister	33	0.21±0.17	0.40±0.15
Brother – sister	57	0.29±0.13	0.29±0.13
$h^2$		0.43±0.10 ( $p < 0.001$ )	0.49±0.10 ( $p < 0.001$ )

SBP - systolic blood pressure, DBP - diastolic blood pressure ; \* - adjusted to the covariables: age, sex,  $age^2$ ,  $age \times sex$ ,  $age^2 \times sex$  and BMI; † - adjusted to the covariables  $age^2$ ,  $age \times sex$ ,  $age^2 \times sex$  and BMI.

# Appendix C

- 95% Confidence Interval of SIR

$$SIR \times \{ 1 \pm z_{\alpha/2} \sqrt{O/E^2} \}$$

# Appendix D

# Family History Score

- The FHS for i-th family

$$Z_i = \frac{(\sum_j O_{ij} - \sum_j E_{ij})}{\{\sum_j E_{ij}(1 - E_{ij})\}^{1/2}} \quad E_{ij} = 1 - \exp\left(-\sum_k \lambda_k t_{ijk}\right)$$

- $O_{ij}$  is the disease indicator of i-th family's j-th member
- The expected risk  $E_{ij}$  is given by cumulative risk of the disease
- $\lambda_k$  is the external reference rate for the k-th stratum
- If disease is rare,  $E_{ij}$  is approximately equal to  $\sum_k \lambda_k t_{ijk}$  and  $E_{ij}(1 - E_{ij}) \approx E_{ij}$

$$Z_i = \frac{(\sum_j O_{ij} - \sum_j E_{ij})}{(\sum_j E_{ij})^{1/2}}$$

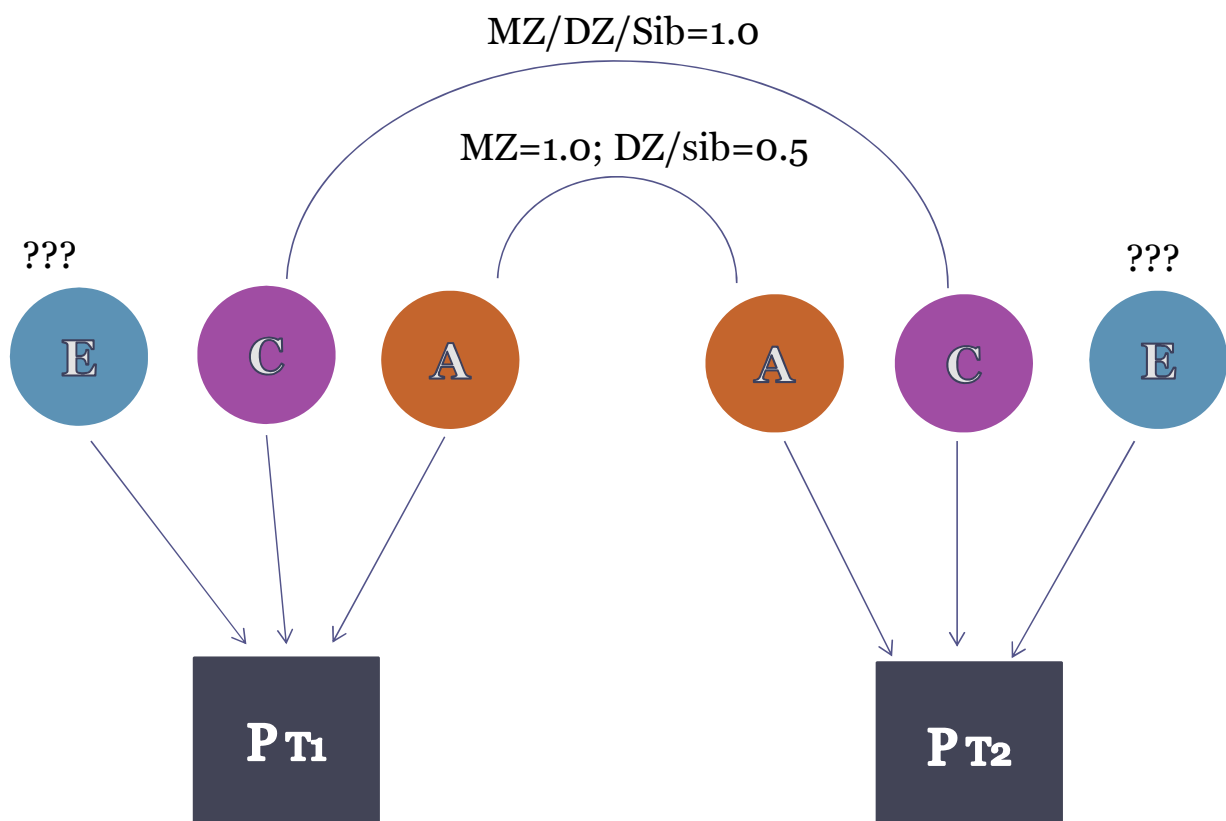
## References

- Risch, N., *Linkage strategies for genetically complex traits. I. Multilocus models*. American journal of human genetics, 1990. **46**(2): p. 222-8.
- van Duijn, C.M., et al., *Familial aggregation of Alzheimer's disease and related disorders: a collaborative re-analysis of case-control studies*. International journal of epidemiology, 1991. **20 Suppl 2**: p. S13-20.
- Kerber, R.A., *Method for calculating risk associated with family history of a disease*. Genetic epidemiology, 1995. **12**(3): p. 291-301.
- Guo, S.W., *Inflation of sibling recurrence-risk ratio, due to ascertainment bias and/or overreporting*. American journal of human genetics, 1998. **63**(1): p. 252-8.
- Yang, Q., et al., *Family history score as a predictor of breast cancer mortality: prospective data from the Cancer Prevention Study II, United States, 1982-1991*. American journal of epidemiology, 1998. **147**(7): p. 652-9.
- Wang, X., et al., *Familial aggregation of blood pressure in a rural Chinese community*. American journal of epidemiology, 1999. **149**(5): p. 412-20.
- Liang, K.Y. and T.H. Beaty, *Statistical designs for familial aggregation*. Statistical methods in medical research, 2000. **9**(6): p. 543-62.
- Sveinbjornsdottir, S., et al., *Familial aggregation of Parkinson's disease in Iceland*. New England Journal of Medicine, 2000. **343**(24): p. 1765-1770.
- Brauer, P.M., et al., *Familial aggregation of diabetes and hypertension in a case-control study of colorectal neoplasia*. American journal of epidemiology, 2002. **156**(8): p. 702-13.
- Andrieu, N., et al., *Familial relative risk of colorectal cancer: a population-based study*. European journal of cancer, 2003. **39**(13): p. 1904-11.
- Haralambous, E., et al., *Sibling familial risk ratio of meningococcal disease in UK Caucasians*. Epidemiology and infection, 2003. **130**(3): p. 413-8.
- Andrieu, N., et al., *Estimation of the familial relative risk of cancer by site from a French population based family study on colorectal cancer (CCREF study)*. Gut, 2004. **53**(9): p. 1322-8.
- Verhage, B.A., et al., *Site-specific familial aggregation of prostate cancer*. International journal of cancer. Journal international du cancer, 2004. **109**(4): p. 611-7.
- Yasui, Y., et al., *Familial relative risk estimates for use in epidemiologic analyses*. American journal of epidemiology, 2006. **164**(7): p. 697-705.
- Fermino, R.C., et al., *Genetic factors in familial aggregation of blood pressure of Portuguese nuclear families*. Arquivos brasileiros de cardiologia, 2009. **92**(3): p. 199-204, 203-9.
- Mavaddat, N., et al., *Familial relative risks for breast cancer by pathological subtype: a population-based cohort study*. Breast cancer research : BCR, 2010. **12**(1): p. R10.
- Paynter, N.P., et al., *Association between a literature-based genetic risk score and cardiovascular events in women*. JAMA : the journal of the American Medical Association, 2010. **303**(7): p. 631-7.
- Scheurer, M.E., et al., *Familial aggregation of glioma: a pooled analysis*. American journal of epidemiology, 2010. **172**(10): p. 1099-107.
- So, H.C., et al., *Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening*. American journal of human genetics, 2011. **88**(5): p. 548-65.

# Co-twin & sibling regression method

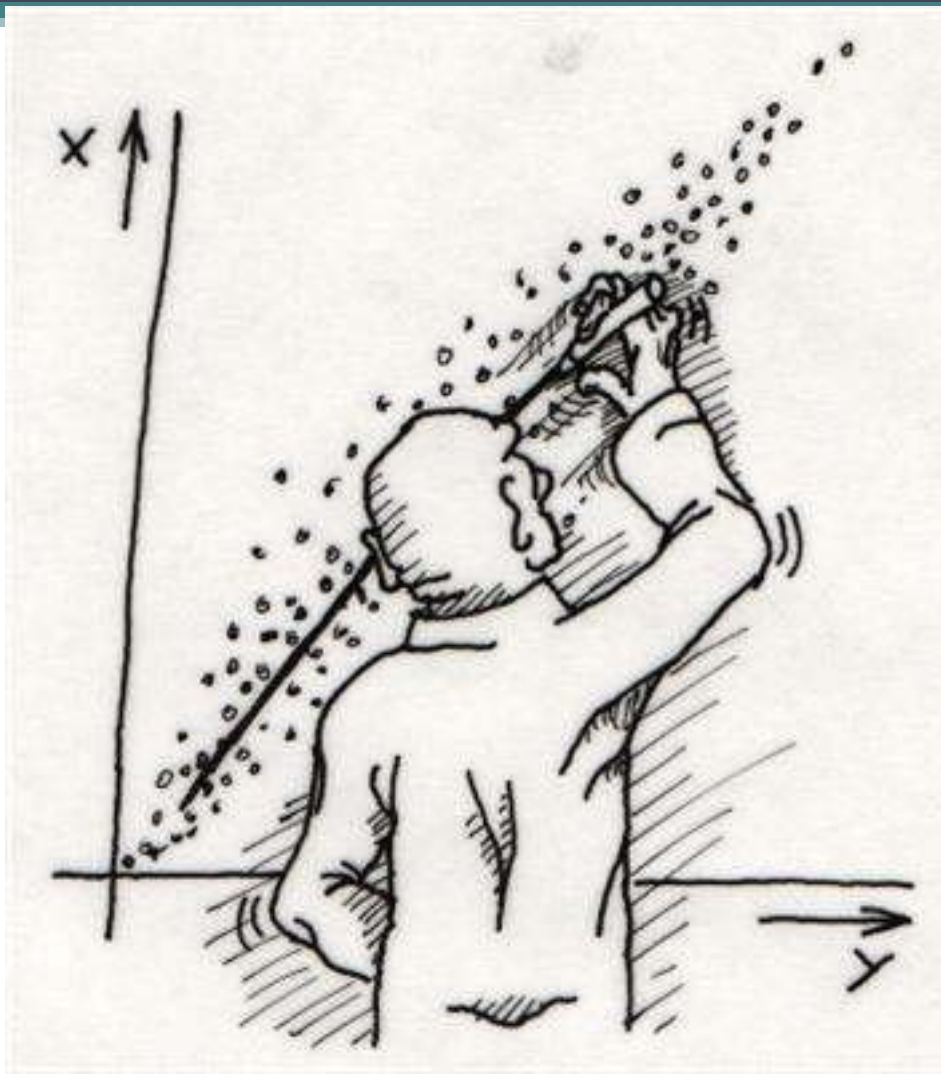
서울대학교 보건대학원 유전체역학연구실  
양사라

## MZ vs. DZ/SIB



# Regression Model

- Finding relationship between a dependent variables and one or more independent variables
- Seek to 'explain' the variation in an outcome of interest in terms of differences in one or more risk factors (covariates)
- $Y \approx f(X, \beta)$  (relates Y to a function of X and  $\beta$ )
- $y = \alpha + \beta x + e$  /  $y = \beta_0 + \beta_1 X + e$



# Topic

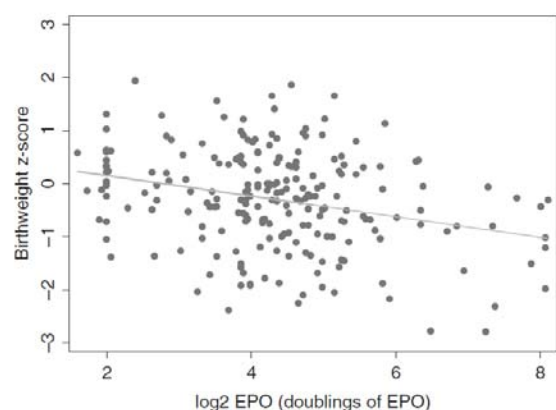
- Co-Twin data : MZ, DZ, SIB
- Relationship between BMI and LDL
- 1) Regression on X alone (treating twin as individuals)
- 2) Multiple regression: including the co-twin x value in the model
- 3) Co-twin control regression (co-twin/sib difference)
- Using SAS

## Regression on X alone (treating twins as individuals)

- The simplest approach to find the best fitting values of  $\beta_0$  and  $\beta_c$

$$E(Y_{ij}) = \beta_0 + \beta_c X_{ij}.$$

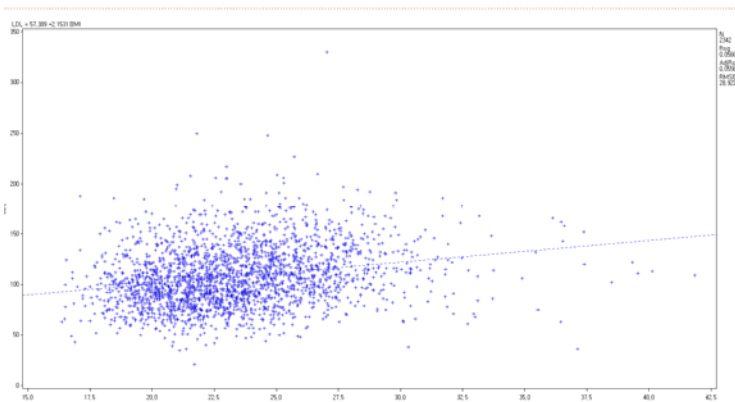
- Presented only as a point of comparison (not statistically independent)



## Regression on X alone (treating twins as individuals)

- $\beta_c$  represents the average rate of change in Y for every unit increase in X
- Y= LDL X= BMI
- Ordinary Least Squares (OLS) – treats all Y values as independent given the corresponding covariate values
- However, the standard method for calculating standard errors when using OLS is not correct in the context of twin data (assumes independency wrongly)

## 1<sup>st</sup> model result



$$\beta_c = 2.15(0.18)$$
$$p < 0.0001$$

## Multiple regression: including the co-twin x value in the model

- Allows the covariate effect to differ within and between twin pairs (본인의 LDL을 Co-Twin의 BMI로 Prediction)
- information about the value of  $Y_{ij}$  in the co-twin's X value as well as in the subject's own X.

$$E(Y_{ij}) = \beta_0 + \beta_W(X_{ij} - \bar{X}_i) + \beta_B \bar{X}_i.$$

- $\beta_W$  gives the expected change in Y for a one-unit change in the difference between the individual X and the twin-pair average X value (within-pair coefficients)
- $\beta_B$  gives the expected change in Y for a one-unit change in the twin-pair average X (between-pair coefficient)

## Interpretation

- $\beta_W = E(MZ) - \frac{1}{2} A + E(DZ/SIB)$
- $\beta_B = A + C + E$  (overall effect)

# Interpretation

- If  $\beta_B \approx 0$  - only association is a dependence of the outcome on within-pair difference in covariate
- If  $\beta_W \approx 0$  - any association of Y with X, which would then only appear in the form of a non-zero  $\beta_B$ , can be explained by shared twin-pair factors.
- If  $\beta_B \approx \beta_W$  - expected change in the outcome for a given change in the covariate is the same irrespective of whether the comparison is made between two twins or between two unrelated individuals in the twin population
- If both coefficients are non-zero but not equal, the interpretation becomes more complex

## 2<sup>nd</sup> model result

- **MZ**
  - $\beta_W = 2.70(0.88)$   $p=0.0021$
  - $\beta_B = 2.85(0.33)$   $p<0.0001$
- **DZ**
  - $\beta_W = 0.81(0.96)$   $p=NS$
  - $\beta_B = 1.08(0.60)$   $p=NS$
- **Sib**
  - $\beta_W = 1.56(0.20)$   $p<0.0001$
  - $\beta_B = 1.60(0.15)$   $p<0.0001$
- **DZ&SIB combined**
  - $\beta_W = 1.64(0.27)$   $p<0.0001$
  - $\beta_B = 4.28(0.07)$   $p<0.0001$
- **Pooled**
  - $\beta_W = 1.59(0.20)$   $p<0.0001$
  - $\beta_B = 4.45(0.04)$   $p<0.0001$

# Regression using twin-pair difference values

- widely used, approach to regression with twin data based on analyzing paired-difference values
- X and Y values within each pair by ordering the twins
- Compare pooled data VS.  
MZ data /DZ data/Sib data/ Dz&Sib Data

$$E(D_i^Y) = \beta_w D_i^X.$$

## Interpretation

- If  $\beta_w$  from pooled data  $<$   $\beta_w$  from MZ data, then we can assume that there are considerably high environmental effect on the phenotype.
- If  $\beta_w$  from pooled data  $<$   $\beta_w$  from Sib/DZ&SIB data, then we can both assume that there are both environmental and genetic effect on the phenotype
- If  $\beta_w$  from MZ  $>$   $\beta_w$  from Sib/DZ&SIB data, then we can assume that there are more environmental effect on the phenotype than genetic effect

## 3<sup>rd</sup> model result

- Pooled  $\beta_w = 1.59(0.15) p < 0.0001$
- MZ  $\beta_w = 2.69(0.50) p < 0.0001$
- DZ  $\beta_w = 0.81(0.72) p = \text{NS}$
- SIB  $\beta_w = 1.56(0.16) p < 0.0001$
- DZ&SIB Combined  $\beta_w = 1.63(0.22) p < 0.0001$

## Conclusion

- **Multiple Regression**- allows simultaneous examination of both within-pair and between pair effects
- **Simple paired-difference analysis**- when interested in within-pair effects and the association of Y with X

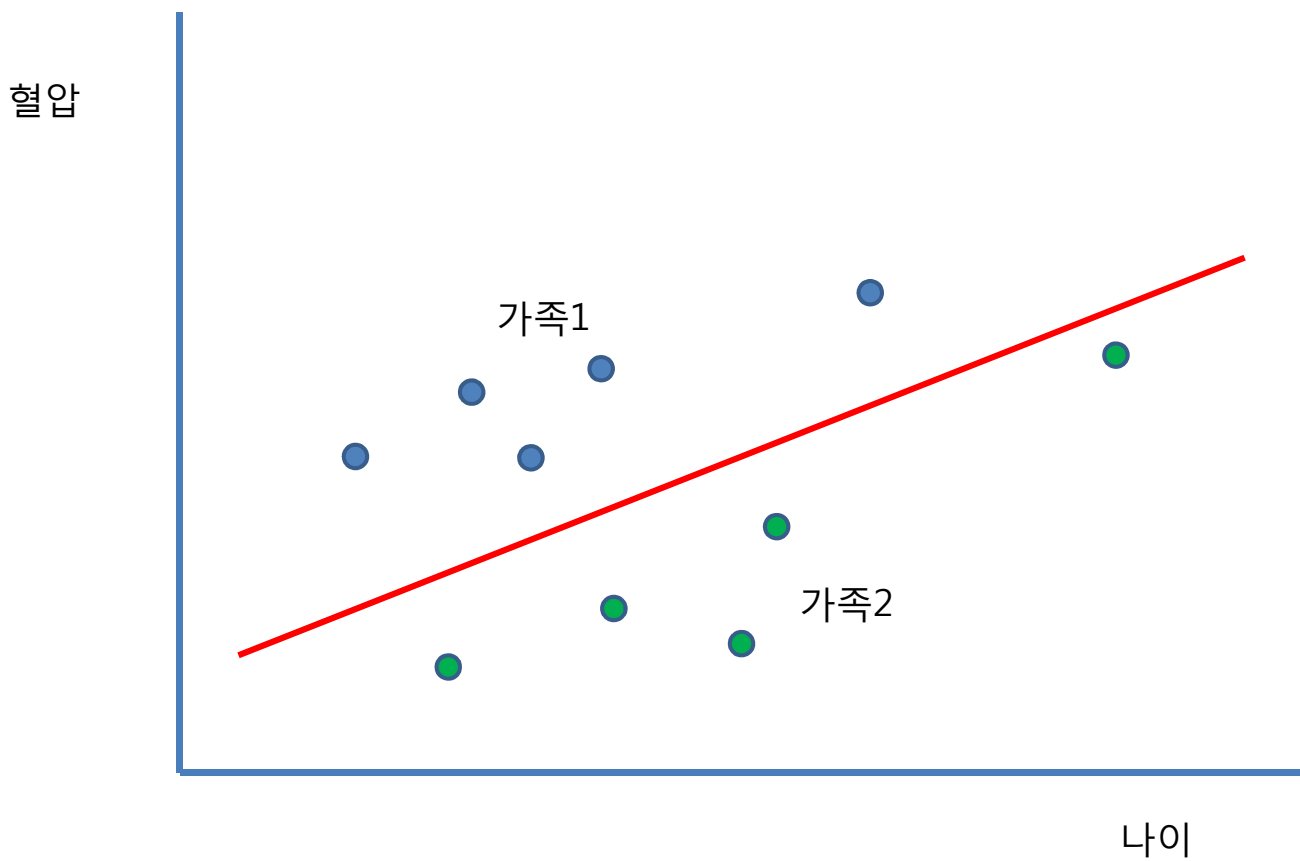
# 가족쌍둥이 분석 1차 연구진 워크샵

서울대학교 보건대학원  
유전체역학교실 석사과정 3학기  
예방의학교실 전공의 2년차  
김진섭

## 차례

1. Random Effect Model의 개념과 일반적인 역학적 관련성 분석
2. Intraclass Correlation(ICC)의 개념과 가족간 correlation의 분석
3. 실습 (SAS & R)

# 1. Random Effect Model

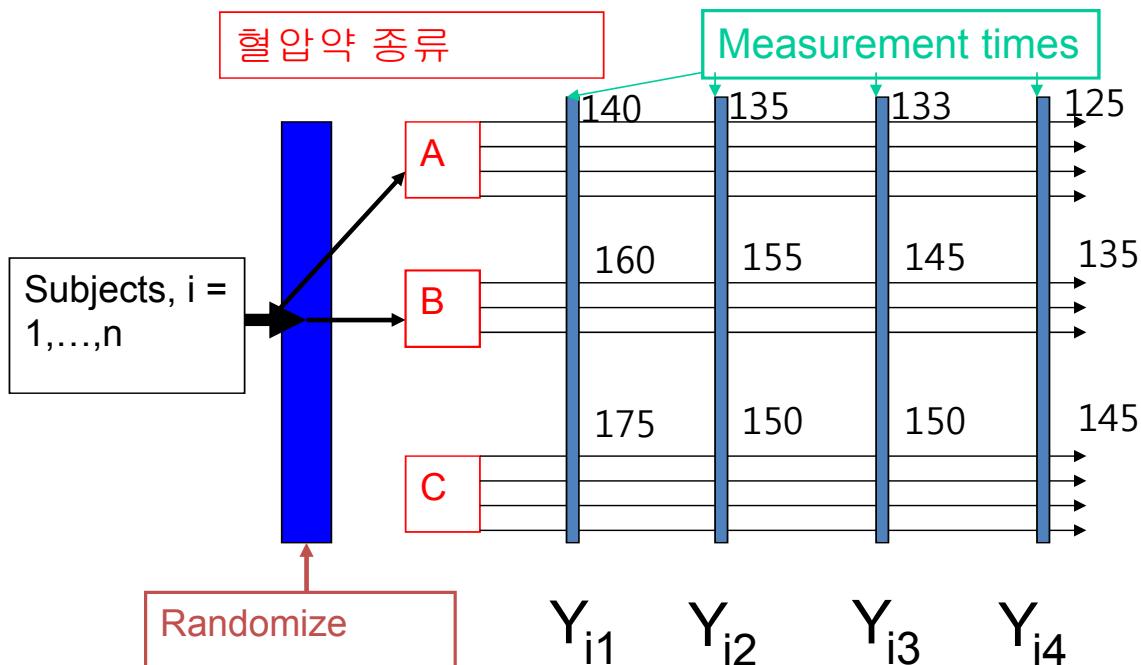


# 일반선형모형의 한계..

- 가족이나 쌍둥이들의 표현형은 상관관계가 있는데, 이를 고려하지 않고 일반적인 분석(독립가정)을 하면 올바른 추정량을 얻을 수 없다.
- 그렇다고 가족이나 쌍둥이를 다 빼고 분석하면 n수가 크게 줄어들어 power가 떨어진다.
- 가족, 쌍둥이 구조를 고려한 분석을 해야 한다. 이를 고려할 수 있는 최소한의 방법이 random effect model!!!

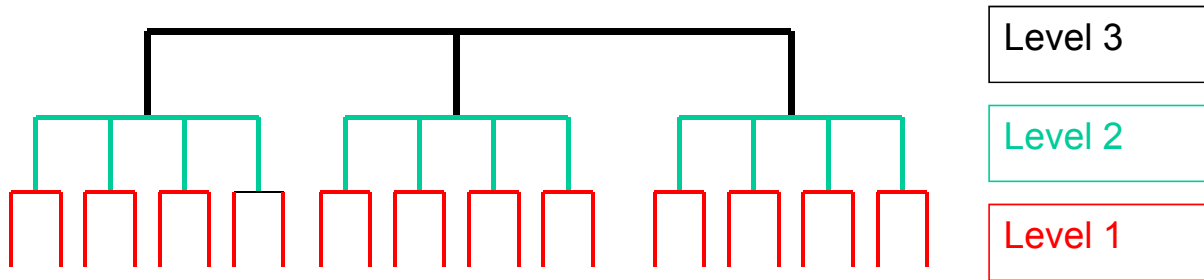
## Correlated data의 예

1. **Repeated measures:** same subjects, same measure, successive times – expect successive measurements to be correlated



# Correlated data의 예(2)

## 2. Clustered/multilevel studies



E.g., Level 3: populations

Level 2: family

Level 1: sibling or twin

We expect correlations within populations and within family or sibling/twin

## Fixed effect model 개념

- 일반적인 회귀분석. **Effect size: beta, OR 등..**
- Effect size가 하나의 고정된 값이라고 생각,  
-> **Beta나 OR 자체에 관심**  
-> 적절한 통계적 방법을 통해 이를 추정한다.
- 연구자가 관심 있는 effect size의 수준이 일정(Fixed)하다는 가정.
- 예: 나이가 혈압에 미치는 effect가 일정할 것이다.

# Fixed effect model 일반수식

- 일반적인 회귀모형
- $Y = X\beta + \epsilon \rightarrow y_i = \beta_0 + x_i\beta_1 + \epsilon_i$

가정: X의 효과는 고정된 값(분산0)  $\rightarrow$  beta1 은 고정된 어떤 수  
 $Var(y_i) = Var(\beta_0 + x_i\beta_1 + \epsilon_i) = Var(\epsilon_i) = \sigma^2$

- 각 관측치는 서로 독립

예: y=혈압, x=나이

- 일반적인 회귀분석으로 beta 값을 추정한다.

## Fixed VS Random??

- 특정 혈압약의 치료효과를 분석.  
**(1) 3개의 병원에서 자료를 뽑았다면?**  
 $\rightarrow$  그냥 병원변수를 더미변수로 주어 일반 회귀분석.
- (2) 5개의 병원**  
 $\rightarrow$  일반 회귀분석??
- (3) 100개의 병원**  
 $\rightarrow$  일반회귀분석 어려워짐....변수의 개수가 너무 많다..

# Random effect model 개념

- 고정된 값이 아니고 분포를 가진 Effect size
- 전체 분포에서 하나의 실현된 경우를 데이터를 통하여 관측한 것으로 생각한다.
- **Effect size의 분산을 추정(평균은 0이라 가정)**  
->> **Effect size 자체에는 관심 없음.**
- Effect size의 수준이 다양하고 자료에서 관측하는 effect size들은 전체 effect size의 일부분이라 생각한다.
- **예) 환자의 effect size:** 환자 수가 많고 연구자가 모든 환자들의 effect size를 수집할 수 없다. 자료에서의 환자의 effect size는 전체 연구집단에서 random sampling 된 것이라고 가정.

## Random effect model 일반수식

- $Y = Za + \epsilon \rightarrow y_{ij} = \beta_0 + a_i + \epsilon_{ij}$

가정: Z의 효과는 random & Z의 효과들은 서로 독립.

즉  $Var(a_i) = \sigma_1^2$ (독립),  $Var(\epsilon_{ij}) = \sigma^2$ (독립)

$$Var(y_i) = Var(\beta_0 + a_i + \epsilon_{ij}) = \sigma_1^2 + \sigma^2$$

예: y=혈압, z= 가족

- 최대가능도추정량(Maximum likelihood estimator,MLE)을 이용하여 구한다.

: 계수와 분산을 예측, 추정한다.

# Mixed model의 접근(1)

- $Y = X\beta + Za + \epsilon$  (위의 두 model의 결합)

가정: X는 fixed effect, Z는 random effect.  $\epsilon$ 는 독립

$$\text{Var}(y_i) = \sigma_1^2 + \sigma^2$$

예: y- 혈압 X-나이 Z-가족

Restrictive maximum likelihood 방법을 통해 근사한다.  
(`proc mixed/glimmix` in SAS, `lme4` in R)

Estimation **beta**, **sigma**, Prediction a

# Mixed model의 접근(2)

- $Y = X\beta + \epsilon$ ,
- X의 effect size는 고정, 단 error가 독립이 아님(가족, 반복측정 등)
- Random effect가 error에 포함되어 있는 model.  
-> Random effect의 분산이 error의 분산에 포함되어 있다.

Proc GENMOD(Gee in R) 이용

GEE(Generalized Estimating Equation) 이용

-최대가능도 추정량(MLE)을 구한다.

-> Score function을 이용하여 분산추측 -> 계수추정 -> 분산추정 -> 계수추정...을 반복한다.

# 접근(2)에서 조심할 점

- 접근(1)에서는 random effect와 error에 대한 분산이 가정되어 있었다.
- 접근(2)에서는 error의 분산이 접근(1)에서의 random effect의 분산도 포함하고 있으므로 이를 고려하여 분산 구조를 지정해주어야 한다. (가족이나 쌍둥이를 고려한)

## 조심할 점(2)

### Types of correlation

1. **Independent:**  $V_i$  is diagonal
2. **Exchangeable:** All measurements on the same unit are **equally correlated**  $\rho_{im} = \rho$

->>> 가족/쌍둥이끼리 상관정도는 같다. **가족의 경우 위험한 가정..**

(부모끼리의 상관정도와 부모자식간의 상관정도가 같다???????? Common environment factor의 보정에 더 가깝다...)

- **Beta값과 V값을 추정한다!!!!**

# 이어서..

## Correlation

For unit i

$$\mathbf{V}_i = \sigma^2 \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & & \\ \vdots & & \ddots & \rho_{..} \\ \rho_{n1} & & \rho_{..} & 1 \end{bmatrix}$$

For repeated measures = correl between times l and m

For clustered data = correl between measures l and m

For all models considered here  $\mathbf{V}_i$  is assumed to be same for all units

$$\mathbf{V}_i = \sigma^2 \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & & \\ \vdots & & \ddots & \rho_{..} \\ \rho_{n1} & & \rho_{..} & 1 \end{bmatrix}$$

$$= \begin{pmatrix} \sigma_1 + \sigma^2 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 + \sigma^2 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 + \sigma^2 \end{pmatrix}$$

# Notation

- Repeated measurements:  $y_{ij}$ ,  $i = 1, \dots, N$ , subjects;
- $j = 1, \dots, n_i$ , times for subject  $i$
- Clustered data:  $y_{ij}$ ,  $i = 1, \dots, N$ , clusters;  $j = 1, \dots, n_i$ , measurements within cluster  $i$
- Use “unit” for subject or cluster

- Vector of measurements for unit  $i$   $\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix}$

Vector of measurements for all units  $\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_N \end{bmatrix}$

- For unit  $i$ :  $E(\mathbf{y}_i) = \boldsymbol{\mu}_i = \mathbf{X}_i \boldsymbol{\beta}$ ;  $\mathbf{y}_i \sim N(\boldsymbol{\mu}_i, \mathbf{V}_i)$
- $\mathbf{X}_i$ :  $n_i \times p$  design matrix
- $\boldsymbol{\beta}$ :  $p \times 1$  parameter vector
- $\mathbf{V}_i$ :  $n_i \times n_i$  variance-covariance matrix,
- e.g.,  $\mathbf{V}_i = \sigma^2 \mathbf{I}$  if measurements are independent
- -> **correlated** 인 경우 다르다.

For all units:  $E(\mathbf{y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ ,  $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{V})$

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_N \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_N \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & 0 & 0 \\ 0 & \ddots & \\ 0 & & \mathbf{V}_N \end{bmatrix}$$

This  $\mathbf{V}$  is suitable if the units are independent

# Choosing the Best Model

## 1. Mixed Model

: -2 Log-likelihood, AIC, AICC, BIC 등.

-> 작을수록 좋은 Model이다!!!!

## 2. GEE

: QIC, QICu

-> 작을수록 좋은 Model이다!!!!

## AIC의 의미??

- Likelihood ratio(가능도비)

일반적으로 우리의 자료가 나올 최대 가능성

-----  
귀무가설(beta=0) 하에서 우리의 자료가 나올 가능성

: 이것이 클수록 일어날 가능성이 큰 Model이다!!!

<-> - **2\*log likelihood** 가 작을수록 일어날 가능성이  
큰 Model이다.

- **AIC = - 2\*log likelihood + 2\*(#parameters)**

## 2. Intraclass correlation

### 정의

- **Intercorrelation** : 변수들 사이의 상관관계  
(예: 혈압과 BMI의 상관관계), 일반적으로 쓰는 상관계수.
- **Intraclass correlation**: 변수에 대한 집단 내 상관관계.  
(예: 가족 내에서 혈압의 상관관계, 쌍둥이 내에서 혈압의 상관관계)

# Review..

$$Y_{ij} = \mu + u_i + e_{ij}$$

$Y_{ij}$ : i번째 가족의 j번째 phenotype

$\mu$ : population의 평균 phenotype

$u_i$ : i번째 가족의 random effect  $\sim N(0, \sigma_u^2)$ ;

$e_{ij}$ : error  $\sim N(0, \sigma_e^2)$

$E(Y_{ij}) = \mu$ ;  $\text{var}(Y_{ij}) = \sigma_u^2 + \sigma_e^2$  ;

$\text{cov}(Y_{ij}, Y_{km}) = \sigma_u^2$ , ( $i=k$ ),  $\text{cov}(Y_{ij}, Y_{km}) = 0$  (otherwise).

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} = \text{ICC}$$

(ICC: intraclass correlation coefficient)

## ICC의 극단적 예

- 즉 ICC는 가족으로 설명되는 분산/전체분산
- **ICC=0** ->  $\sigma_u^2 = 0$ , 즉 random effect의 분산이 0이다. 따라서 가족 내에서의 상관관계가 없다 -> 가족의 영향이 없다.
- **ICC=1** ->  $\sigma_e^2 = 0$ , 즉 random effect로 전체 분산을 설명할 수 있다. -> 가족간의 상관관계가 1

# ICC의 계산

- Random effect를 이용하여 분산을 추정하고 그것으로 ICC를 구할 수 있다.

- Mixed model(1):

-> Random effect의 분산

-----  
(Random effect의 분산+residual의 분산)

- Mixed model(2): GEE

->  $\rho$ 가 ICC다.

$$\mathbf{V}_i = \sigma^2 \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & & \\ \vdots & & \ddots & \rho_{..} \\ \rho_{n1} & & \rho_{..} & 1 \end{bmatrix}$$

$$= \begin{pmatrix} \sigma_1 + \sigma^2 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 + \sigma^2 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 + \sigma^2 \end{pmatrix}$$

$$\rho = \frac{\sigma_1^2}{\sigma_1^2 + \sigma^2}$$

# 계산(2)

- ANOVA table이용하여 구할 수 있다.(SAS에서 macro써 야함....)

Source of Variation	Sums of Squares	Degrees of Freedom	Mean Squares	E(MS)
Between schools	$SS_B = \sum_{i=1}^n \sum_{j=1}^{n_i} w_{ij} (\bar{y}_{i.} - \bar{y}_{..})^2$	$n - 1$	$MS_B = \frac{SS_B}{n - 1}$	$\sigma_W^2 + n' \sigma_B^2$
Within schools	$SS_W = \sum_{i=1}^n \sum_{j=1}^{n_i} w_{ij} (y_{ij} - \bar{y}_{i.})^2$	$\sum_{i=1}^n (n_i - 1)$	$MS_W = \frac{SS_W}{\sum_{i=1}^n (n_i - 1)}$	$\sigma_W^2$
Total	$SS_T = \sum_{i=1}^n \sum_{j=1}^{n_i} w_{ij} (y_{ij} - \bar{y}_{..})^2$	$\sum_{i=1}^n n_i - 1$		

$$IC = \frac{MS_B - MS_W}{MS_B + (n' - 1) MS_W} \quad n' = \frac{1}{n - 1} \left[ \sum_{i=1}^n n_i - \frac{\sum_{i=1}^n n_i^2}{\sum_{i=1}^n n_i} \right]$$

## 3. 실습 (SAS & R)

# 1. Continuous data(SAS)

## Proc mixed

```
proc mixed data=kkk2;
class FID ;
    **범주형변수 입력
model BMI= LDL /solution;
    ** Fixed effect 입력
random FID / solution;
    ** Random effect 입력
run;
```

/solution : Fixed effect의 estimator,  
Random effect 의 predictor를 구한다.

## Proc GENMOD

```
proc genmod data=kkk2;
class FID ;
model BMI=LDL;
repeated subject=FID/ type=exch corrw;
run;
```

## exch: correlation structure 가정.  
-> 가족 내 에서의 상관 정도는 같다.

```
-----
The Mixed Procedure
Class Level Information
Class  Levels  Values
FID      756    1 2 3 4 5 6 7 8 9 10 11 12 13
          14 15 16 17 18 19 20 21 22 23
          24 25 26 27 28 29 30 31 32 33
          34 35 36 37 38 39 40 41 42 43
          44 45 46 47 48 49 50 51 52 53
          54 55 56 57 58 59 60 61 62 63
          64 65 66 67 68 69 70 71 72 73
          74 75 76 77 78 79 80 81 82 83
          84 85 86 87 88 89 90 91 92 93
          94 95 96 97 98 99 100 101 102
          103 104 105 106 107 108 109
          110 111 112 113 114 115 116
          117 118 119 120 121 122 123
          124 125 126 127 128 129 130
          131 132 133 134 135 136 137
          138 139 140 141 142 143 144
          145 146 147 148 149 150 151
          152 153 154 155 156 157 158
          159 160 161 162 163 164 165
          166 167 168 169 170 171 172
          173 174 175 176 177 178 179
```

```
Covariance Parameter
Estimates
Cov Parm  Estimate
FID       2.6593
Residual  7.6857
```

```
Fit Statistics
-2 Res Log Likelihood    15655.5
AIC (smaller is better)  15659.5
AICC (smaller is better) 15659.5
BIC (smaller is better)  15668.7
```

```
The Mixed Procedure
Solution for Fixed Effects
Effect      Estimate      Standard Error      DF      t Value      Pr > |t|
Intercept  20.8010          0.2208              755      94.20        <.0001
LDL         0.02579          0.001854           2322     13.91        <.0001
```

$$ICC = 2.6593 / (2.6593 + 7.6857) = 0.2571$$

```
Solution for Random Effects
Effect  FID  Estimate      Std Err      DF      t Value      Pr > |t|
FID     1    0.2519        1.2541       2322     0.20         0.8408
FID     2   -1.0963        0.9886       2322    -1.11         0.2676
FID     3    0.3950        0.9884       2322     0.34         0.7347
FID     4   -0.2530        0.8834       2322    -0.29         0.7746
FID     5   -0.7676        1.0572       2322    -0.73         0.4678
FID     6   -0.4215        0.6786       2322    -0.62         0.5345
FID     7   -1.2914        1.2541       2322    -1.03         0.3032
FID     8    2.4830        0.9883       2322     2.51         0.0121
FID     9    0.0554        1.0572       2322     0.52         0.6006
```

Exchangeable Working  
Correlation

Correlation 0.2096770207

GEE Fit Criteria

QIC 3083.0963  
QICu 3081.0000

Analysis Of GEE Parameter Estimates  
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr >  Z
Intercept	20.8232	0.2309	20.3707	21.2758	90.18	<.0001
LDL	0.0256	0.0020	0.0217	0.0295	13.01	<.0001

- ICC=0.2097

# 1. Continuous data(R)

## Package(lme4)

- `install.packages("lme4")`
- `library(lme4)`
- `a=read.csv("C:/kkk3.csv")`
- `fm1=lmer(BMI ~ LDL +(1 | FID),a)`
- `print(fm1)`
- `ranef(fm1, drop=TRUE)`
- `fm2=lmer(BMI ~ LDL +(1 | FID)+(1 | MZTWIN),a)`
- `print(fm2)`

## Package(gee)

- `install.packages("gee")`
- `library(gee)`
- `gm1=gee(BMI~LDL, id=MZTWIN, data=a, corstr="exchangeable")`

```

> fml=lmer(BMI ~ LDL +(1 | FID),a)
> print(fml)
Linear mixed model fit by REML
Formula: BMI ~ LDL + (1 | FID)
Data: a
   AIC   BIC logLik deviance REMLdev
15663 15688  -7828   15642   15655
Random effects:
Groups   Name             Variance Std.Dev.
FID      (Intercept)  2.6593   1.6307
Residual                    7.6857   2.7723
Number of obs: 3079, groups: FID, 756

Fixed effects:
              Estimate Std. Error t value
(Intercept) 20.800969   0.220826   94.20
LDL          0.025791   0.001854   13.91

Correlation of Fixed Effects:
(Intr)
LDL -0.931

> print(gml)
GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
Link: Identity
Variance to Mean Relation: Gaussian
Correlation Structure: Exchangeable

Call:
gee(formula = BMI ~ LDL, id = FID, data = a, corstr = "exchangeable")

Number of observations : 3079

Maximum cluster size : 17

Coefficients:
(Intercept)          LDL
20.82321286  0.02560247

Estimated Scale Parameter: 10.20392
Number of Iterations: 2

Working Correlation[1:4,1:4]
      [,1] [,2] [,3] [,4]
[1,] 1.000000 0.209677 0.209677 0.209677
[2,] 0.209677 1.000000 0.209677 0.209677
[3,] 0.209677 0.209677 1.000000 0.209677
[4,] 0.209677 0.209677 0.209677 1.000000

```

# Data 변환

- **data** kkk3;set kkk2;
- if BMI <25 then obe=0;
- if BMI >=25 then obe=1;
- if BMI <=21 then BMI\_C=0;
- else if BMI <=24 then BMI\_C=1;
- else if BMI <=26 then BMI\_C=2;
- else if BMI <=28 then BMI\_C=3;
- else if BMI <=30 then BMI\_C=4;
- else BMI\_C=5;
- **run;**

-> Binomial data로 바뀌서 obe에 저장(비만이 1).

Count data로 바뀌서 BMI\_C에 저장(임의로 했음, 포아송분포 따르는지 꼭 그림 그려 확인해야..)

## 2. Binomial data(SAS)

- **Proc genmod or proc glimmix** 사용
  - glimmix가 계산이 복잡하여 프로그램이 실행되지 않을 경우 많음..
- Logistic regression에 random effect를 추가한 개념.
  
- **proc genmod** data=kkk3 descending;
- class FID ;
- model obe=LDL / dist=binomial;
- repeated subject=FID/ type=exch corrw;
- **run;**

### Exchangeable Working Correlation

Correlation      0.1300790578

### GEE Fit Criteria

QIC              3713.9530  
QICu             3712.4807

### Analysis Of GEE Parameter Estimates Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr >  Z
Intercept	-2.3946	0.1665	-2.7210	-2.0682	-14.38	<.0001
LDL	0.0143	0.0014	0.0115	0.0170	10.21	<.0001

# 2. Binomial data(R)

## Package(lme4)

- `km1=glmer(obe~LDL+(1 | FID), data=a, family=binomial)`
- `Print(km1)`

## Package(gee)

- `gm1=gee(obe~LDL, id=FID,data=a,family=binomial, corstr="exchangeable")`
- `Print(gm1)`

```
> km1=glmer(obe~LDL+(1 | FID), data=a, family=binomial)
> print(km1)
Generalized linear mixed model fit by the Laplace approximation
Formula: obe ~ LDL + (1 | FID)
Data: a
   AIC   BIC logLik deviance
3624 3642  -1809    3618
Random effects:
Groups Name          Variance Std.Dev.
FID      (Intercept) 0.88471  0.94059
Number of obs: 3079, groups: FID, 756

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.830942    0.181652  -15.58  <2e-16 ***
LDL          0.016798    0.001502   11.19  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
      (Intr)
LDL -0.951
```

```

> gm2=gee(obe~LDL, id=FID,data=a,family=binomial, corstr="exchangeable")
Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
running glm to get initial regression estimate
(Intercept)          LDL
-2.35271503  0.01387753
> print(gm2)

GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
Link:                Logit
Variance to Mean Relation: Binomial
Correlation Structure: Exchangeable

Call:
gee(formula = obe ~ LDL, id = FID, data = a, family = binomial,
    corstr = "exchangeable")

Number of observations : 3079

Maximum cluster size   : 17

Coefficients:
(Intercept)          LDL
-2.39459479  0.01425132

Estimated Scale Parameter: 1.001359
Number of Iterations: 2

Working Correlation[1:4,1:4]
      [,1]      [,2]      [,3]      [,4]
[1,] 1.0000000 0.1300793 0.1300793 0.1300793
[2,] 0.1300793 1.0000000 0.1300793 0.1300793
[3,] 0.1300793 0.1300793 1.0000000 0.1300793
[4,] 0.1300793 0.1300793 0.1300793 1.0000000

```

## 3. Count data(SAS)

- 예: 자녀수, 교통사고 사망자 수 등.
- Proc genmod or glimmix 이용
- Poisson regression에 random effect 적용한 개념
- **proc genmod** data=kkk3;
- class FID ;
- model BMI\_C=LDL/ dist=poi;
- repeated subject=FID/ type=exch corrw;
- **run;**

## 2. Count data(R)

### Package(lme4)

- `km1=glmer(obe~LDL+(1 | FID), data=a, family=poisson)`
- `Print(km1)`

### Package(gee)

- `gm1=gee(obe~LDL, id=MZTWIN,data=a,family=poisson, corstr="exchangeable")`

## 참고문헌

- **SAS System for Mixed Models, Ramon et al. ,1996 SAS Institute Inc.**
- **Linear Models in Statistics(2<sup>nd</sup> Edition), Rancher et al. ,WILEY**
- **R-project:** [www.r-project.org/](http://www.r-project.org/)
- **Generlized Estimating Equation:** [hisdu.sph.uq.edu.au/lisu/SSAI%20course/presentations/Annette.ppt](http://hisdu.sph.uq.edu.au/lisu/SSAI%20course/presentations/Annette.ppt)

경청해 주셔서 감사합니다.

**Estimating the variance  
explained by  
all genotyped markers**

**Minji Han**

Graduate School of Public Health  
Seoul National University

# Contents

📖 Introduction

📖 What is GCTA?

📖 Functions of GCTA

📖 Estimate the variance explained by SNPs

📖 Options of GCTA

📖 Examples

## GCTA

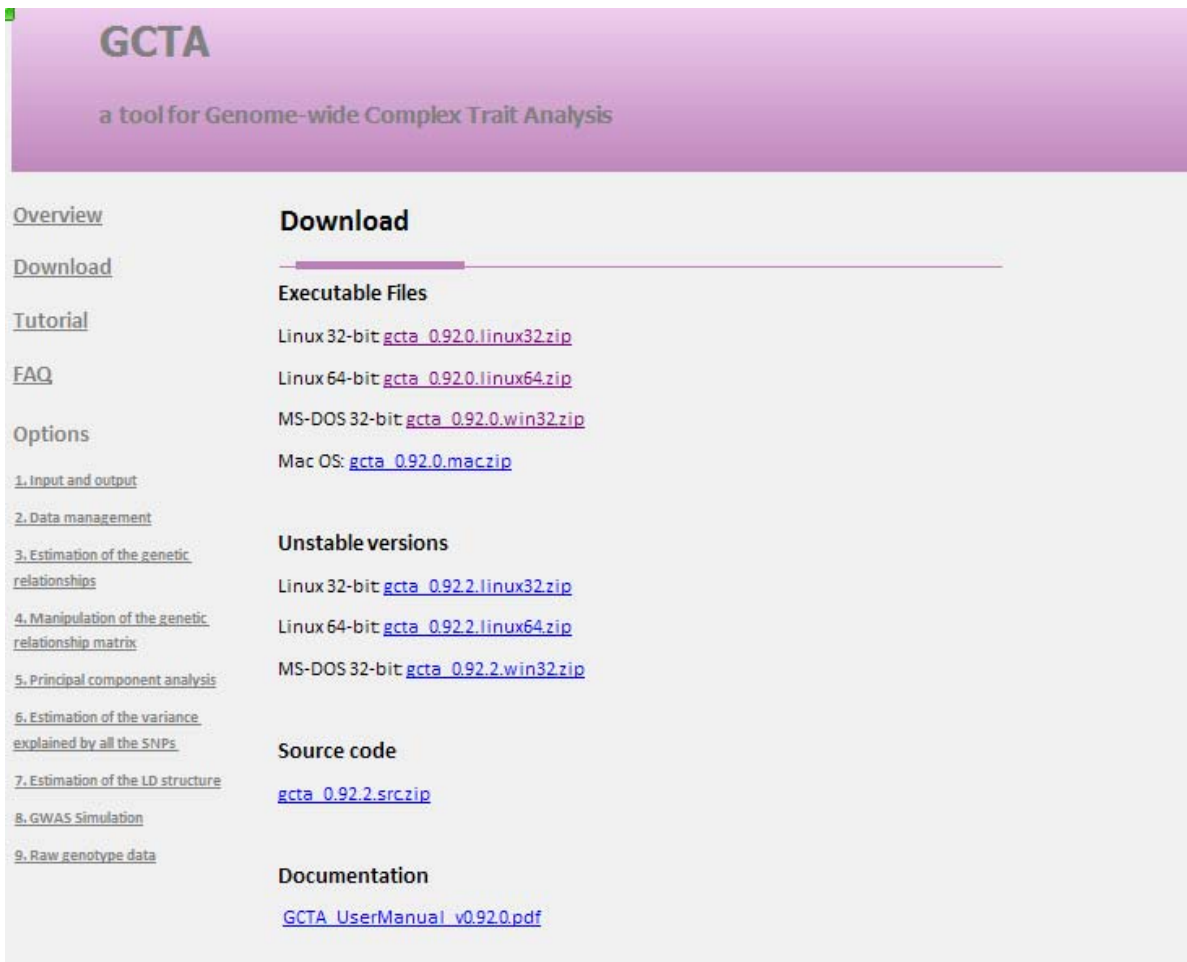
a tool for Genome-wide Complex Trait Analysis

### **GCTA (Genome-wide Complex Trait Analysis)**

is designed to estimate the proportion of phenotypic variance explained by genome- or chromosome-wide SNPs for complex traits.

**GCTA** was developed by [Jian Yang](#), [Hong Lee](#), [Mike Goddard](#) and [Peter Visscher](#) and is maintained in [Peter Visscher's lab](#) at the [Queensland Institute of Medical Research](#) .

# Download GCTA <http://gump.qimr.edu.au/gcta>



**GCTA**  
a tool for Genome-wide Complex Trait Analysis

**Overview**

**Download**

**Executable Files**

- Linux 32-bit: [gcta\\_0.92.0.linux32.zip](#)
- Linux 64-bit: [gcta\\_0.92.0.linux64.zip](#)
- MS-DOS 32-bit: [gcta\\_0.92.0.win32.zip](#)
- Mac OS: [gcta\\_0.92.0.mac.zip](#)

**Unstable versions**

- Linux 32-bit: [gcta\\_0.92.2.linux32.zip](#)
- Linux 64-bit: [gcta\\_0.92.2.linux64.zip](#)
- MS-DOS 32-bit: [gcta\\_0.92.2.win32.zip](#)

**Source code**

- [gcta\\_0.92.2.src.zip](#)

**Documentation**

- [GCTA UserManual v0.92.0.pdf](#)

**Options**

- [1. Input and output](#)
- [2. Data management](#)
- [3. Estimation of the genetic relationships](#)
- [4. Manipulation of the genetic relationship matrix](#)
- [5. Principal component analysis](#)
- [6. Estimation of the variance explained by all the SNPs](#)
- [7. Estimation of the LD structure](#)
- [8. GWAS Simulation](#)
- [9. Raw genotype data](#)

## Citations

### Methodology:

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010 Jul 42(7): 565-9. [[PubMed ID: 20562875](#)].

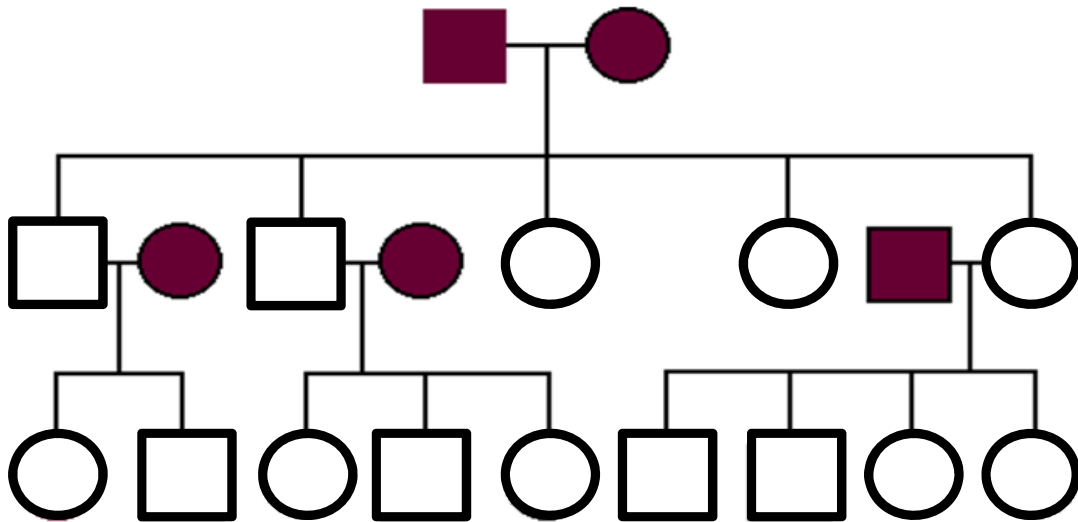
Lee SH, Wray NR, Goddard ME and Visscher PM. Estimating Missing Heritability for Disease from Genome-wide Association Studies. Am J Hum Genet. 2011 Mar 88(3): 294-305. [[PubMed ID: 21376301](#)]

Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, de Andrade M, Feenstra B, Feingold E, Hayes MG, Hill WG, Landi MT, Alonso A, Lettre G, Lin P, Ling H, Lowe W, Mathias RA, Melbye M, Pugh E, Cornelis MC, Weir BS, Goddard ME, Visscher PM: Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet. 2011 Jun 43(6): 519-525. [[PubMed ID: 21552263](#)]

### Software tool:

Yang J, Lee SH, Goddard ME and Visscher PM. GCTA: a tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 2011 Jan 88(1): 76-82. [[PubMed ID: 21167468](#)]

# Founders



Use only **unrelated** data

## Basic Concept

- To fit the effects of all the SNPs as **random effects** by a mixed linear model (MLM)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \text{ with } \text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{W}\mathbf{W}'\sigma_u^2 + \mathbf{I}\sigma_\varepsilon^2, \quad (\text{Equation 1})$$

- $\mathbf{y}$  :  $n \times 1$  vector of phenotypes (n: sample size)
- $\boldsymbol{\beta}$ : vector of fixed effects such as age, sex and/or principal components
- $\mathbf{u}$ : vector of SNP effects with  $\mathbf{u} \sim N(0, \mathbf{I}\sigma_u^2)$
- $\mathbf{I}$ :  $n \times n$  identity matrix
- $\boldsymbol{\varepsilon}$ : vector of residual effects with  $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma_\varepsilon^2)$
- $\mathbf{W}$ : standadized genotype matrix with  $ij^{\text{th}}$  element where  $x_{ij}$  is the number of copies of the reference allele for the  $i^{\text{th}}$  SNP of the  $j^{\text{th}}$  individual and  $p_i$  is the frequency of the reference allele

If we define  $\mathbf{A} = \mathbf{W}\mathbf{W}'/N$   $\sigma_g^2 = N\sigma_u^2$

# Total genetic effects

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\varepsilon} \text{ with } \mathbf{V} = \mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_\varepsilon^2, \quad (\text{Equation 2})$$

- $\mathbf{g}$ :  $n \times 1$  vector of the total genetic effects of the individuals  
with  $\mathbf{g} \sim N(0, \mathbf{A}\sigma_g^2)$
  - $\mathbf{A}$ : genetic relationship matrix (GRM) between individuals
- >>> we can estimate  $\sigma_g^2$  by the restricted maximum likelihood (REML) approach, relying on the GRM estimated from all the SNPs

## GCTA's main functions

- 1) Data management
- 2) Estimation of **the genetic relationships** from SNPs
- 3) Manipulation of **the genetic relationship matrix**
- 4) Estimation of **the variance explained by all the SNPs**
  - Estimate the variance explained by all the autosomal SNPs;
  - Partition the genetic variance onto individual chromosomes;
  - Estimate the genetic variance associated with the X-chromosome;
  - Test the effect of dosage compensation on the X-chromosome;
- 6) Estimation of the linkage disequilibrium(LD) structure
- 7) GWAS Simulation

# Estimation of the Genetic Relationships from Genome-wide SNPs

- **genetic relationships** between individuals
- Including close relatives would result in the estimate of genetic variance being driven by the phenotypic correlations, and this estimate could be a biased estimate of total genetic variance, for example because of common environmental effects.
- For data collected from family or twin studies, it is recommended to exclude one individual of a pair whose relationship is greater than a specified cutoff value, e.g., 0.025.

## Estimation of the Variance explained by Genome-wide SNPs by REML

- The GRM estimated from the SNPs can be fitted subsequently in an MLM to estimate the variance explained by these SNPs via the REML method.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^r \mathbf{g}_i + \boldsymbol{\varepsilon}, \quad \mathbf{V} = \sum_{i=1}^r \mathbf{A}_i \sigma_i^2 + \mathbf{I} \sigma_{\varepsilon}^2$$

- $\mathbf{g}_i$ : vector of random genetic effects, which could be the total genetic effects for the whole genome or for a single chromosome
- $\sigma_i^2$  : variance of the  $i^{\text{th}}$  genetic factor with its corresponding GRM,  $\mathbf{A}_i$

# Options

## Input and output

`--bfile test`

Input **PLINK** binary PED files, e.g. **test.fam**, **test.bim** and **test.bed** (see **PLINK** user manual for details).

`--out test`

Specify output root filename.

## DATA format (PLINK format)

- PED file

Family ID Individual ID Paternal ID Maternal ID Sex Phenotype Genotype

```
1 1 0 0 1 2 A A A C 0 0 0 0
2 1 0 0 2 2 A A A C 0 0 A A 0 0
2 2 0 0 1 1 A A A C 0 0 A A 0 0
9 1 1 2 0 0 0 0 0 0 0 0 0 0
```

- FAM file

```
1 1 0 0 1 2
2 1 0 0 2 2
2 2 0 0 1 1
9 1 1 2 0 0
```

- MAP file

Chromosome SNP identifier Genetic distance (morgans) Base-pair position (bp units)

```
1 snp2 0 2
2 snp4 0 4
1 snp1 0 1
1 snp3 0 3
5 snp5 0 1
```

# Binary format

- .PED, .FAM, .MAP files

>>> binary format (.bed, .fam, .bim)

file명을 같은 이름으로 저장할 것

- 실행하기

```
gcta --bfile #filename --out  
#out_filename
```

## Options

### Estimation of the genetic relationships from the SNPs

#### --make-grm

Estimate the genetic relationship matrix (GRM) between pairs of individuals from a set of SNPs. By default, **GCTA** will save the lower triangle of the genetic relationship matrix in a compressed text file (e.g. **test.grm.gz**) and save the IDs in a plain text file (e.g. **test.grm.id**).

#### --make-grm-xchr

Estimate the GRM from SNPs on the X-chromosome.

The GRM will be saved in the same format as above. Due to the speciality of the GRM for the X-chromosome, it is not recommended to manipulate the matrix by --grm-cutoff or --grm-adj, or merge it with the GRMs for autosomes (see below for the options of manipulating the GRM).

## GCTA 예시 Manipulation of the genetic relationship matrix & Estimation of the phenotypic variance explained by the SNPs using the REML method

```
mjhan@med2:~/gcta> gcta --mgrm grm_chrs.txt --pheno
test2.phen --mphenos 7 --reml --out test_b_all_chrs
```

Computational time: 0:0:6

\*\*\*\*\*

```
* Genome-wide Complex Trait Analysis (GCTA)
* version 0.92.9
* (C) 2010 Jian Yang, Hong Lee, Michael Goddard and Peter Visscher
* GNU General Public License, v2
* Queensland Institute of Medical Research
```

\*\*\*\*\*

Analysis started: Thu Jul 28 10:42:47 2011

Options:

```
--mgrm grm_chrs.txt
--pheno test2.phen
--mphenos 7
--reml
--out test_b_all_chrs
```

## Output Manipulation of the genetic relationship matrix & Estimation of the phenotypic variance explained by the SNPs using the REML method

Summary result of REML analysis:

Source	Variance	SE	V(1)/Vp	0.019357	0.139129
V(1)	0.255906	1.839386	V(2)/Vp	0.087940	0.140971
V(2)	1.162626	1.865685	V(3)/Vp	0.000001	0.124318
V(3)	0.000008	1.643560	V(4)/Vp	0.058272	0.130698
V(4)	0.770395	1.728121	V(5)/Vp	0.000001	0.128368
V(5)	0.000008	1.697104	V(6)/Vp	0.000001	0.111012
V(6)	0.000008	1.467657	V(7)/Vp	0.000001	0.127114
V(7)	0.000008	1.680529	V(8)/Vp	0.021692	0.106195
V(8)	0.286785	1.403722	V(9)/Vp	0.000001	0.111700
V(9)	0.000008	1.476749	V(10)/Vp	0.010131	0.114726
V(10)	0.133944	1.516662	V(11)/Vp	0.098720	0.113298
V(11)	1.305139	1.501200	V(12)/Vp	0.000001	0.122558
V(12)	0.000008	1.620294	V(13)/Vp	0.007631	0.101151
V(13)	0.100893	1.337145	V(14)/Vp	0.000001	0.099306
V(14)	0.000008	1.312890	V(15)/Vp	0.031845	0.091080
V(15)	0.421014	1.204768	V(16)/Vp	0.000001	0.097689
V(16)	0.000008	1.291507	V(17)/Vp	0.000001	0.090777
V(17)	0.000008	1.200127	V(18)/Vp	0.010522	0.094413
V(18)	0.139110	1.247934	V(19)/Vp	0.029107	0.071358
V(19)	0.384818	0.944184	V(20)/Vp	0.006020	0.082144
V(20)	0.079584	1.085799	V(21)/Vp	0.052781	0.073648
V(21)	0.697802	0.976383	V(22)/Vp	0.020504	0.067922
V(22)	0.271083	0.898294			
V(e)	7.211488	4.073288			
Vp	13.220661	0.817974			

**Output** Manipulation of the genetic relationship matrix  
& Estimation of the phenotypic variance explained by the SNPs  
using the REML method

	BMI		LDL		(BMI)(LDL)	
	Variance	SE	Variance	SE	Variance	SE
V(1)/Vp	0.02	0.14	0.15	0.15	0.00	0.18
V(2)/Vp	0.08	0.14	0.00	0.15	0.00	0.18
V(3)/Vp	0.00	0.13	0.00	0.14	0.00	0.17
V(4)/Vp	0.04	0.13	0.00	0.14	0.06	0.17
V(5)/Vp	0.00	0.13	0.03	0.13	0.05	0.16
V(6)/Vp	0.00	0.11	0.02	0.13	0.00	0.16
V(7)/Vp	0.00	0.13	0.03	0.12	0.19	0.15
V(8)/Vp	0.04	0.11	0.07	0.12	0.00	0.15
V(9)/Vp	0.00	0.11	0.00	0.12	0.00	0.14
V(10)/Vp	0.01	0.12	0.06	0.13	0.00	0.16
V(11)/Vp	0.09	0.11	0.00	0.11	0.05	0.12
V(12)/Vp	0.00	0.12	0.01	0.13	0.00	0.14
V(13)/Vp	0.03	0.10	0.00	0.12	0.00	0.13
V(14)/Vp	0.00	0.10	0.02	0.10	0.00	0.11
V(15)/Vp	0.03	0.09	0.07	0.10	0.13	0.12
V(16)/Vp	0.00	0.10	0.07	0.11	0.00	0.13
V(17)/Vp	0.00	0.09	0.00	0.10	0.00	0.12
V(18)/Vp	0.02	0.10	0.00	0.10	0.11	0.13
V(19)/Vp	0.01	0.07	0.01	0.08	0.00	0.10
V(20)/Vp	0.01	0.08	0.02	0.09	0.03	0.12
V(21)/Vp	0.01	0.07	0.07	0.07	0.00	0.09
V(22)/Vp	0.01	0.07	0.00	0.07	0.00	0.09

**Output** Manipulation of the genetic relationship matrix  
& Estimation of the phenotypic variance explained by the SNPs  
using the REML method

	BMI		LDL		(BMI)(LDL)	
	Variance	SE	Variance	SE	Variance	SE
V(g)	2.33	3.04	574.86	241.89	0.00	3320.57
V(e)	7.69	3.04	380.37	235.70	8908.60	3328.82
Vp	10.02	0.53	955.22	50.57	8908.60	469.53
V(g)/Vp		0.3		0.25	0.00	0.37

## GCTA 실습 예제))

1. Estimate the GRM from SNPs on chromosome 21 with  $MAF > 0.05$
2. Estimate the GRM from SNPs on all autosomal chromosome with  $MAF > 0.01$  ( make batch file )
3. Estimation of the phenotypic variance explained by the SNPs using the REML method (all autosomal chromosome, 2nd phenotype)

참고)

## Missing Heritability

- **MISSING HERITABILITY**: SNPs discovered by GWASs account for only a small fraction of the genetic variation of complex traits in human populations.
- The heritability of height has been estimated to be  $\sim 0.8$ , but  $\sim 50$  variants that are associated with height account for only  $\sim 5\%$  of phenotypic variation.
- In this study, 45% of variance can be explained by considering all SNPs simultaneously. Thus, most of the heritability is hiding rather than missing because of many SNPs with small effects.