Lecture 5
Three level variance component models

Three levels models

- In three levels models the clusters themselves are nested in superclusters, forming a hierarchical structure.
- For example, we might have repeated measurement occasions (units) for patients (clusters) who are clustered in hospitals (superclusters).

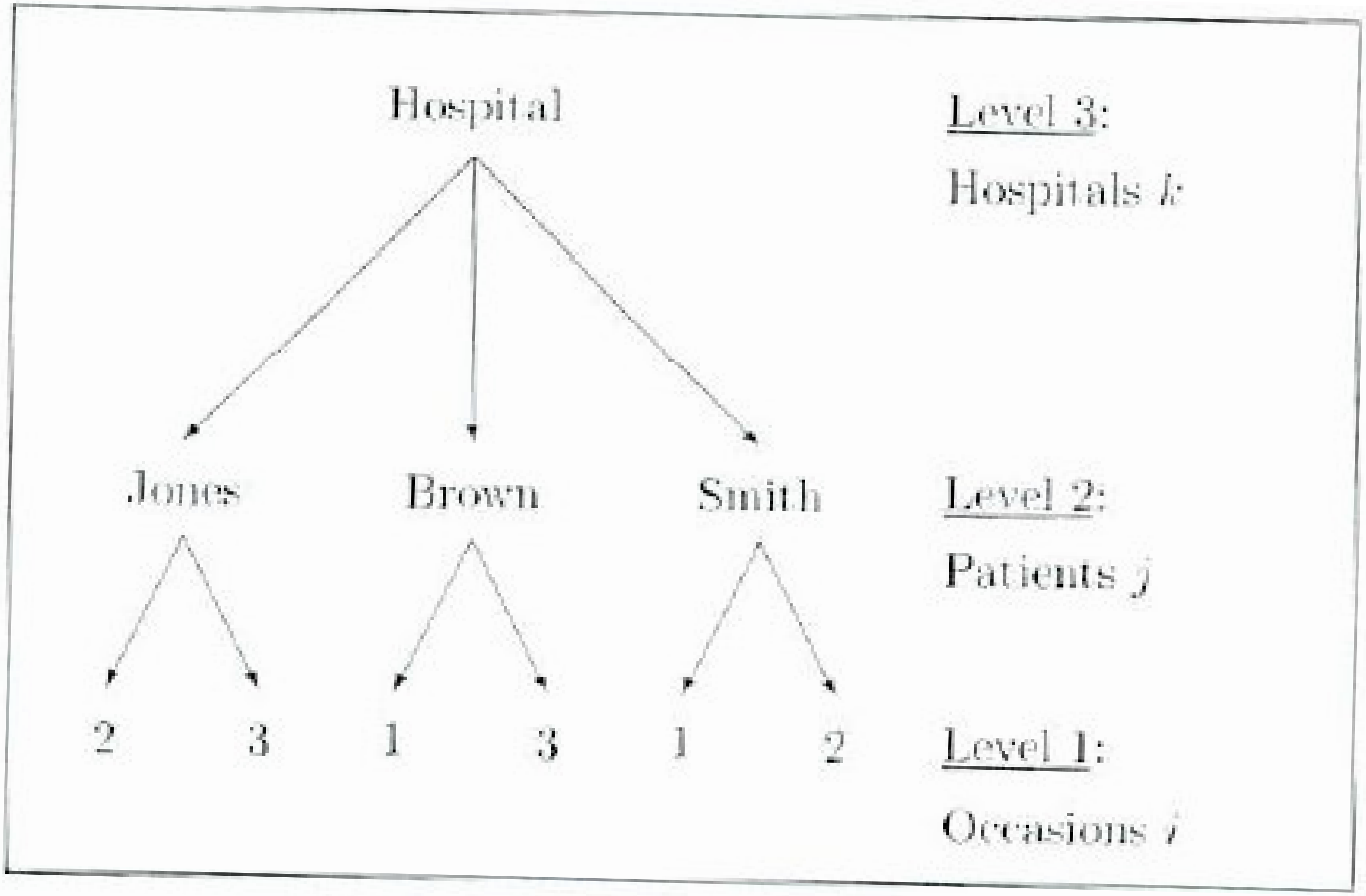


Figure 7.1: Illustration of three-level design

Which method is best for measuring respiratory flow?

- Peak respiratory flow (PEFR) is measured by two methods, the standard Wright peak flow and the Mini Wright meter, each on two occasions on 17 subjects.

Table 1.1: Peak respiratory flow rate measured on two occasions using both the Wright and the Mini Wright meter (Bland and Altma, Lancet 1986)

	Wright peak flow meter		Mini Wright meter	
Subject	First	Second	First	Second
1	494	490	512	525
2	395	397	430	415
3	516	512	520	508
4	434	401	428	444
5	476	470	500	500
6	557	611	600	625
7	413	415	364	460
8	442	431	380	390
9	650	638	658	642
10	433	429	445	432
11	417	420	432	420
12	656	633	626	605
13	267	275	260	227
14	478	492	477	467
15	178	165	259	268
16	423	372	350	370
17	427	421	451	443

Level 1: occasion (i)
Level 2: method (j)
Level 3: individual (k)

Model 1: two-level

We fitted a two-level model to all 4 measurements ignoring the fact the different methods were used

- Occasion *i*, method *j*, subject *k*

$$y_{ijk} = \beta_1 + \zeta_k^{(3)} + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

Variance of the measurements within subjects

$$\zeta_k^{(3)} \sim N(0, \tau^2)$$

Variance of the measurements across subjects

Here we made no distinction between the two methods

Model 2: two-level

- Occasion *i*, method *j*, subject *k*

$$y_{ijk} = \beta_1 + \beta_2 x_j + \zeta_k^{(3)} + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

$$\zeta_k^{(3)} \sim N(0, \tau^2)$$

Here we might add a binary variable for estimating the methods' effect - this variable allows for a systematic difference between the 2 methods

Intraclass correlation coefficient

$$\frac{\tau^2}{\tau^2 + \sigma^2} = \frac{109.2^2}{109.2^2 + 23.8^2} = 0.95$$

Correlation between the 4 repeated measures on
the same individual (the method used for the measurement
Is ignored)

The % of the total variance of the measurements
(within + between) that is explained by the variance of the measu
individuals

Why we need three stage?

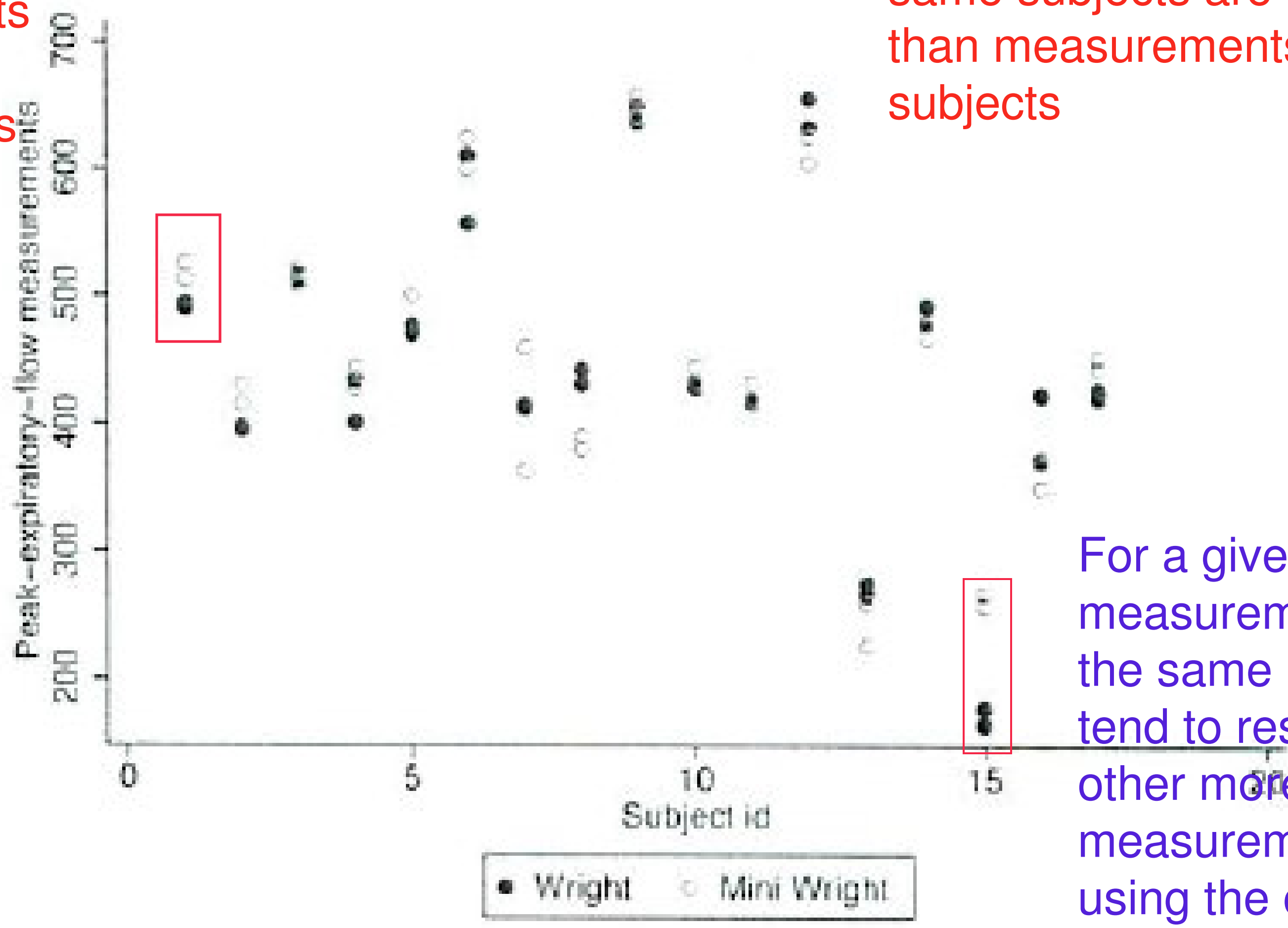
occasion (i), method(j), individual (k)

- Both two-level variance component models assume that the four measurements using the two methods, were all mutually independent, conditional on the random intercept (that is, they ignore the possibility that the measurements obtained with the same method might be more similar to each other than the measurements obtained with two different methods). In other words the measurements are “nested” within the “method”
- To see if this appears reasonable, we can plot all four measurements against subject id

Fig 7.2: Scatterplot of peak-respiratory flow measured by two methods versus subject id

The shift between the measurements taken from the 2 methods varies across subjects

measurements on the same subjects are more similar than measurements on different subjects



For a given subject, the measurements using the same method tend to resamble each other more than measurements using the other method

Why we need three-level models?

- As expected, measurements on the same subjects are more similar than measurements on different subjects. This between subject heterogeneity is modeled by the subject-level intercept $\varsigma_k^{(3)}$

Why we need three-level models

- The figure suggests that for a given subject, the measurements using the same method tend to be more similar to each other, violating the conditional independence assumption of model (1)
- The difference between methods is not due to some constant shift of the measurements using one method relative to the other, but due to shifts that vary between subjects, thus violating the assumption in model (2)

Model 3: three-level variance component models

$$y_{ijk} = \beta_1 + \zeta_{jk}^{(2)} + \zeta_k^{(3)} + \varepsilon_{ijk}$$

account for between-method within-subject heterogeneity

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

$$\zeta_{jk}^{(2)} \sim N(0, \tau_2^{\ 2})$$

Variance of the measurements across the two methods for the same subject

$$\zeta_k^{(3)} \sim N(0, \tau_3^{\ 2})$$

Variance of the measurements across subjects

Parameters interpretations

$$y_{ijk} = \beta_1 + \zeta_{jk}^{(2)} + \zeta_k^{(3)} + \varepsilon_{ijk}$$

β_1 Population average of all measurements (across occasions, methods, and subjects)

$\beta_1 + \zeta_k^{(3)}$ Average of the measurements for subject k (across occasions and methods)

$\beta_1 + \zeta_{jk}^{(2)} + \zeta_k^{(3)}$ Average of the measurements for method j and for subject k (across occasions)

Model 4: three-level variance component models

$$y_{ijk} = \beta_1 + \boxed{\beta_2 x_j} + \zeta_{jk}^{(2)} + \zeta_k^{(3)} + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

$$\zeta_{jk}^{(2)} \sim N(0, \tau_2^{\ 2})$$

$$\zeta_k^{(3)} \sim N(0, \tau_3^{\ 2})$$

Different types of intraclass correlation

$$\rho(subject) = cor(y_{ijk}, y_{i'j'k} \mid x_j, x_{j'}) =$$

$$= \frac{\tau_3^2}{\tau_2^2 + \tau_3^2 + \sigma^2}$$

correlation between the 4 measurements within the subject (same subject, different method, and different occasion)

$$\rho(method, subject) = cor(y_{ijk}, y_{i'jk} \mid x_j) =$$

$$= \frac{\tau_2^2 + \tau_3^2}{\tau_2^2 + \tau_3^2 + \sigma^2}$$

correlation between the measurements obtained with the same method and for the same subject (same subject, same method, different occasions)

Intraclass correlations

- Note that cor(method,subject)> cor(subject). This makes sense since, as we saw in Figure 7.2, measurements using the same method are more similar than measurements using different methods for the same person.

Three-stage formulation

Random effect　　Fixed effect

$$y_{ijk} = \eta_{jk} + \beta_2 x_j + \varepsilon_{ijk}$$ Stage 1

$$\eta_{jk} = \pi_k + \zeta_{jk}^{(2)}$$ Stage 2

$$\pi_k = \beta_1 + \zeta_{jk}^{(3)}$$ Stage 3

$$y_{ijk} = \beta_1 + \beta_2 x_j + \zeta_{jk}^{(2)} + \zeta_{jk}^{(3)} + \varepsilon_{ijk}$$

Table 7.1: Maximum likelihood estimates for two-level and three-level models for expiratory flow data

	Two-level models		Three-level models	
	Model 1	Model 2	Model 3	Model 4
	Est (SE)	Est (SE)	Est (SE)	Est (SE)
Fixed part				
β_1	450.9 (26.6)	447.9 (26.8)	447.9 (26.9)	450.90 (26.6)
β_2		6.0 (5.7)	6.0 (7.8)	
Random part				
`xtmixed`				
$\sqrt{\psi^{(2)}}$			19.0 (4.8)	19.5 (4.8)
$\sqrt{\psi^{(3)}}$	109.2 (18.9)	109.2 (18.9)	108.6 (19.0)	108.6 (19.1)
$\sqrt{\theta}$	23.8 (2.4)	23.6 (2.3)	17.8 (2.2)	17.8 (2.2)
Log likelihood	−349.89	−349.33	−345.00	−345.29

Television school and family smoking cessation project (TVSFP)

- The TVSFP is a study designed to determine the efficacy of a school-based smoking prevention program in conjunction with a television-based prevention program, in terms of preventing smoking onset and increasing smoking cessation (Flay et al 1995)

TVSFP: outcome

- Outcome: a tobacco and health knowledge scale (THKS) assessing the student's knowledge of tobacco and health
- Linear model for THKS post-intervention, with THKS pre-intervention as a covariate

TVSFP: study design

- 2x2 factorial design, with four intervention conditions determined by cross-classification of a school-based social resistant curriculum (CC: coded as 0 or 1) with a television-based program (TV, coded as 0 or 1)
- Randomization to one of the four intervention conditions was at the school level
- Intervention was delivered at the classroom level
- 1600 seventh-grades students from 135 classes in 28 schools in Los Angeles

Three-level model for the TVSFP

i (student), *j* (classroom), *k* (school)

postTHKS

$$Y_{ijk} = \beta_1 + \beta_2 preTHKS + \beta_3 CC + \beta_4 TV + \beta_5 (CC \times TV) +$$

$$+ b_k^{(3)} + b_{jk}^{(2)} + \varepsilon_{ijk}$$

$\varepsilon_{ijk} \sim N(0, \sigma_1^2)$ Within classroom, across students

$b_{jk}^{(2)} \sim N(0, \sigma_2^2)$ Within school, across classrooms

$b_k^{(3)} \sim N(0, \sigma_3^2)$ Across schools

Intraclass correlation coefficients

- Correlation among THKS scores for classmates (or children within the same class and same school) is 0.061

$$\frac{\sigma_3^2 + \sigma_2^2}{\sigma_3^2 + \sigma_2^2 + \sigma_1^2} = \frac{0.039 + 0.065}{0.039 + 0.065 + 1.602}$$

Intraclass correlation coefficients

- Correlation among THKS scores for children for different classrooms within the same school is 0.023

$$\frac{\sigma_3^2}{\sigma_3^2+\sigma_2^2+\sigma_1^2}=\frac{0.039}{0.039+0.065+1.602}$$

Table 17.3 Fixed and random effects estimates for the THKS scores from the Television, School and Family Smoking Prevention and Cessation Project.

Variable	Estimate	SE	Z
Intercept	1.702	0.1254	13.57
Pre-Intervention THKS	0.305	0.0259	11.79
CC	0.641	0.1609	3.99
TV	0.182	0.1572	1.16
CC × TV	−0.331	0.2245	−1.47
Level 3 Variance:			
σ_3^2	0.039	0.0253	1.52
Level 2 Variance:			
σ_2^2	0.065	0.0286	2.26
Level 1 Variance:			
σ_1^2	1.602	0.0591	27.10

Should we ignore the intraclass correlation?

- The intraclass correlation coefficients were relatively small at both the school and at the classroom levels.
- We might be tempted to think that the clustering of the data would not affect the intervention effects
- Such conclusion would be erroneous
- Although the intraclass correlations are small, they have substantial impact on the inferences

Linear model for the TVSFP without random effects

i (student), *j* (classroom), *k* (school)

postTHKS

$$Y_{ijk} = \beta_1 + \beta_2 preTHKS + \beta_3 CC + \beta_4 TV + \beta_5 (CC \times TV) + \varepsilon_{i}$$

$$\varepsilon_{ijk} \sim N(0, \sigma_1^2)$$

This model ignores clustering in the data at a classroom and school levels. This is a standard linear regression model and assumes that the responses are independent

Table 17.4 Fixed effects estimates from analysis that ignores clustering in the THKS scores from the Television, School and Family Smoking Prevention and Cessation Project.

Variable	Estimate	SE	Z
Intercept	1.661	0.0844	19.69
Pre-Intervention THKS	0.325	0.0258	12.58
CC	0.641	0.0921	6.95
TV	0.199	0.0900	2.21
CC × TV	−0.322	0.1302	−2.47

Comparing results

- Model-based standard errors (assuming no clustering) and misleading small for the randomized intervention effects and lead to substantially different conclusions
- Bottom line: even a very modest intra-cluster correlation can have a discernable impact on the inferences