Variance component models

Analysis of repeated measurements, NFA 2016

Julie Lyng Forman & Lene Theil Skovgaard

Department of Biostatistics, University of Copenhagen

Topics for today

Linear mixed models for clustered data and repeated measurements in general, i.e. not just for longitudinal data.

New concepts:

- random effects
- variance components
- multi-level models

Suggested reading:

- Fitzmaurice et al. (2011): chapters 8, 21, 22.
- Bland & Altman: *Statistical methods for assessing agreement between two methods of clinical measurement*, Lancet (1986).
- Merlo et al: *Diastolic blood pressure and area of residence: multilevel versus ecological analysis of social inequity*, J. Epedimiol. Community Health, (2001)

Outline

Analysis of repeated measurements

Many applications:

- Longitudinal data (lecture 2)
- Cluster randomized trials/multi-center studies.
- Reproducibility/reliability of measurement methods.
- Treatments applied to multiple limbs, teeth, etc within the same subject.
- Cross-over trials (lecture 4).

ATT: Measurements belonging to the same subject/cluster are correlated. If we fail to take correlation into account our statistical results may be biased.

Sources of variation / correlation

Measurements belonging to the same subject/cluster tend to be correlated (look alike) due to e.g.

- Environmental variation.
  - Between regions, hospitals or work places.
- Biological variation.
  - Between individuals, families or animals.

Today: Use random effects (variance components) to model various sources of variation in a linear mixed model framework.

Outline

One-way analysis of variance – with **random** variation

The simplest possible model for clustered data.

- Comparison of k groups or clusters, satisfying:
- The groups are of no individual interest and it is of no relevance to test whether they have identical means.
- The groups may be thought of as representatives from a population, that we want to describe.

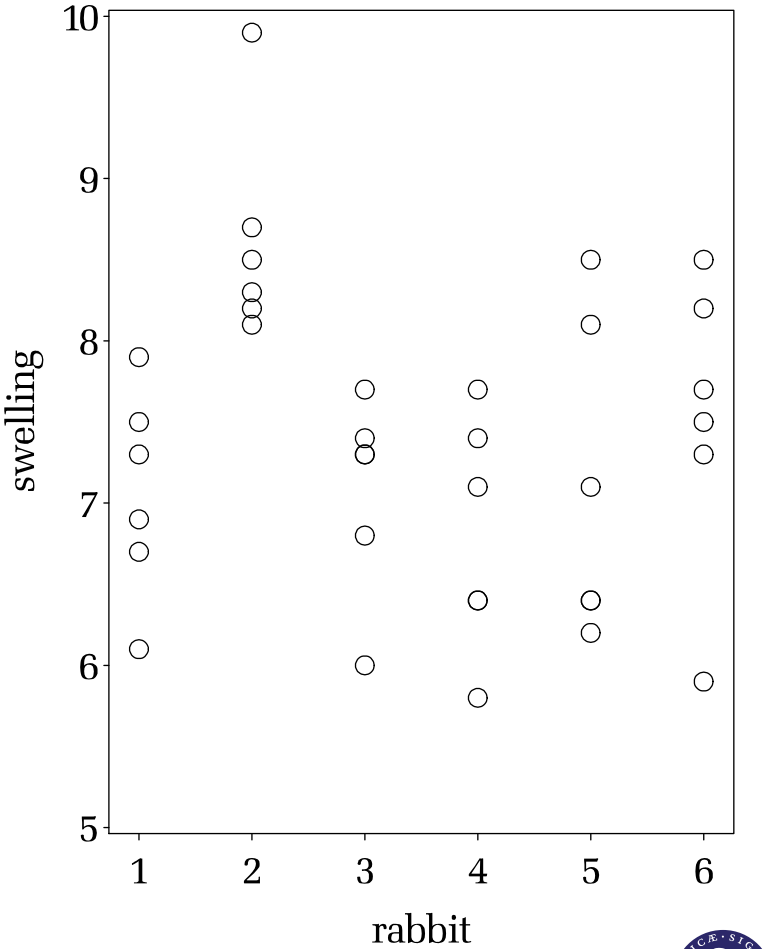
Example: Rabbit data

- $R = 6$ rabbits vaccinated.
- In $S = 6$ spots on the back.

Response: swelling in cm^2

Research question:

How much swelling can be expected in reaction to the vaccine?



Random effects anova (the two-level model)

We let each rabbit have its own level of swelling described as

$$Y_{rs} = A_r + \varepsilon_{rs}$$

- We **assume** that these individual levels are randomly sampled from a normally distributed population,

$$A_r \sim \mathcal{N}(\mu, \omega_B^2)$$

- The error terms are considered to be independent normal,

$$\varepsilon_{rs} \sim \mathcal{N}(0, \sigma_W^2)$$

The rabbit levels are so-called random effects and the variances ω_B^2 and σ_W^2 are so-called variance components describing the variance **between rabbits** and **within rabbits**, respectively.

Implications of random effects anova

All observations are considered as randomly sampled measurements from the **same population**. Thus, the model implies that all measurements follow the same normal distribution:

$$Y_{rs} \sim N(\mu, \omega_B^2 + \sigma_W^2)$$

- Population mean μ, the grand mean.
- Population variance $\omega_B^2 + \sigma_W^2$, the total variation.

But: Measurements made on the same rabbit are correlated with the so-called intra-class correlation

$$\text{Corr}(y_{r1}, y_{r2}) = \rho = \frac{\omega_B^2}{\omega_B^2 + \sigma_W^2}$$

Compound symmetry

The implied covariance of the repeated measurements has a compound symmetry pattern:

$$\begin{pmatrix} \omega_B^2 + \sigma_W^2 & \omega_B^2 & \dots & \omega_B^2 \\ \omega_B^2 & \omega_B^2 + \sigma_W^2 & \dots & \omega_B^2 \\ \vdots & \vdots & & \vdots \\ \omega_B^2 & \omega_B^2 & \dots & \omega_B^2 + \sigma_W^2 \end{pmatrix}$$

In particular all pairs of spots on the same rabbit are assumed to be equally correlated (with the intra-class correlation).

Exchangeability

If any two pairs of measurements are equally correlated we say that the measurements are exchangeable.

- Are the spots randomly selected - ???

If not, an unstructured covariance is more apropriate

- Some spots are expected to respond more similarly than others (physiological/spatial correlation pattern).

In other situations exchangeability is obvious

- E.g. patients sampled randomly from several GPs.

Random effects anova in PROC MIXED

```
PROC MIXED DATA=rabbit;
  CLASS rabbit;
  MODEL swelling = / SOLUTION;
  RANDOM rabbit;
RUN;
```

Covariance Parameter Estimates

Cov Parm	Estimate
rabbit	0.3304
Residual	0.5842

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	7.3667	0.2670	5	27.59	<.0001

Estimation of variance components

Level	Variation	Variance component	Estimate	%of variation
1	Between	ω_B^2	0.3304	36%
2	Within	ω_W^2	0.5842	64%
	Total	$\omega_B^2 + \sigma_W^2$	0.9146	100%

$$\text{ICC} = \frac{\omega_B^2}{\omega_B^2 + \sigma_W^2} = 0.36.$$

Quite a lot of variability **within** rabbits - ?

- Are there systematic differences between the spots?
- Or perhaps measurements just aren't that precise.

Beware not to overinterpret the estimates in a small dataset!

Interpreation of variance components

Typical differences between spots on **different** rabbits:

$$\begin{aligned} y_{r_1 s_1} - y_{r_2 s_2} &= \alpha_{r_1} - \alpha_{r_2} + \varepsilon_{r_1 s_1} - \varepsilon_{r_2 s_2} \\ &\sim N(0, 2 \cdot (\sigma_B^2 + \omega_W^2)) \end{aligned}$$

- 95% normal range: $0 \pm 2\sqrt{2\sigma_B^2 + 2\omega_W^2} = \pm 2.70 \ cm^2$

Typical differences between spots on the **same** rabbit:

$$\begin{aligned} y_{rs_1} - y_{rs_2} &= \varepsilon_{rs_1} - \varepsilon_{rs_2} \\ &\sim N(0, 2\omega_W^2) \end{aligned}$$

- 95% normal range: $0 \pm 2\sqrt{2\omega_W^2} = \pm 2.16 \ cm^2$

Why not use traditional one-way anova?

Focus on rabbit means and test $H_0: \ \mu_1 = \ldots = \mu_6$.

One-way anova table:

	SS	df	MS=SS/df	F
Between rabbits	12.8333	$R-1 = 5$	2.5667	4.39
Within rabbit	17.5266	$R(S-1) = 30$	0.5842	
Total	30.3599	$RS-1 = 35$	0.8674	

Test for identical rabbits means: $F = 4.39 \sim F(5, 30), \ P = 0.004$.

But: We are not interested in these particular 6 rabbits, only in rabbits in general, as a **species**! Presumably these 6 rabbits have been **randomly sampled** from the species.

One-way anova with and without random variation

Classical one-way anova

- The rabbit means μ_r are fixed parameters,
  - supposedly of an interest of their own.
- We say that the rabbit factor is a fixed effect.

Random effects one-way anova

- The rabbit levels A_r are considered random and their population mean μ and variance $\omega_B^2 + \sigma_W^2$ is the major interest.
- We say that the rabbit factor is a random effect.
- (If data is from a pilot study used in the planning of some trial, the intra-class correlation will also be of interest).

Estimation of individual rabbit means

Sometimes estimates of individual random effects are used for e.g. prediction of future disease status.

How do we estimate them?

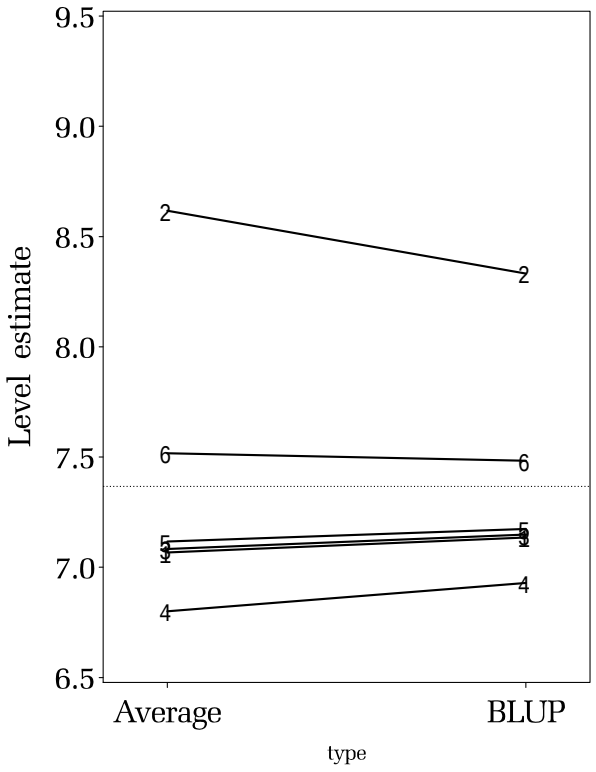
- Simple averages $\bar{y}_{r.}$ of the individual measurements.
- Best unbiased linear predictors (BLUPs) are **weighted averages** of the individual and the population mean:

$$\frac{\tilde{\omega}_B^2}{\tilde{\omega}_B^2 + \frac{\tilde{\sigma}_W^2}{S}} \bar{y}_{r.} + \frac{\frac{\tilde{\sigma}_W^2}{S}}{\tilde{\omega}_B^2 + \frac{\tilde{\sigma}_W^2}{S}} \bar{y}_{..}$$

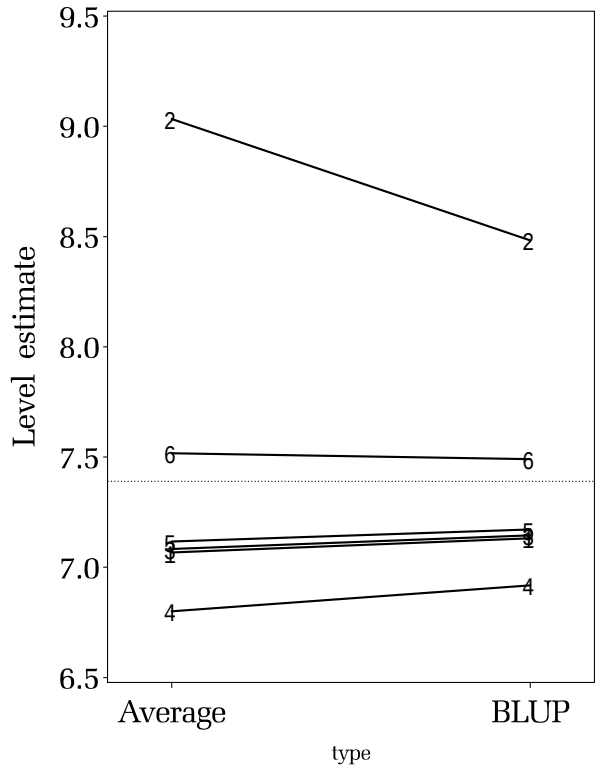
They have been **shrinked** towards the grand mean, $\bar{y}_{..}$.

BLUPs vs averages

Full data



Reduced data



Note: We see larger shrinkage for rabbit no. 2 when the 3 smallest measurements from this rabbit have been removed (i.e. we are *borrowing strenght from the neighbours*).

Outline

Fixed or random effect?

Fixed effects such as treatment, gender, and time.

- Typically a limited number of carefully selected groups.
- Group names are specific and cannot be shuffled.
- Each group must have a decent size in order to reach interesting conclusions (statistical power).

Random effect such as subject, rat or familly.

- Possibly a large number of different groups.
- Group names are non-informative (number of subject, rat or family) and could be shuffled without consequence.
- Allows inference to be extended beyond the subjects in the experiment and to the population they were sampled from.
- The number of groups matters not the size of the groups.

Testing fixed effects

Imagine that rabbits are grouped in two (e.g. treatments):

level	variation	covariates
1	within rabbit	`spot`
2	between rabbits	`group`

- Part of the variation *between rabbits* could be explained by systematic differences between groups.
- Part of the variation *within rabbits* could be explained by systematic differences between spots.

Testing fixed effects with PROC MIXED

```
PROC MIXED DATA=rabbit;
  CLASS group rabbit spot;
  MODEL swelling = group spot / SOLUTION CL DDFM=KR;
  RANDOM rabbit;
RUN;
```

Output:

```
Covariance Parameter Estimates

Cov Parm          Estimate

rabbit              0.3694  <-------------- smaller than before
Residual            0.5477  <-------------- smaller than before
```

Testing fixed effects with PROC MIXED

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
group	1	4	0.64	0.4675
spot	5	25	1.40	0.2584

Solution for Fixed Effects

Effect	spot	group	Estimate	StdError	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept			6.9111	0.4792	4	14.42	0.0001	0.05	5.5807	8.2416
group		1	0.4444	0.5542	4	0.80	0.4675	0.05	-1.0942	1.9831
group		2	0	.	.	.	.	.	.	.
spot	a		0.6500	0.4273	25	1.52	0.1408	0.05	-0.2300	1.5300
spot	b		0.05000	0.4273	25	0.12	0.9078	0.05	-0.8300	0.9300

...

Disregarding repeated measurements

When the **random rabbit variation** is ignored:

```
PROC GLM DATA=rabbit;
  CLASS group spot;
  MODEL swelling=group spot / SOLUTION CLPARM;
RUN;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
group	1	1.77777778	1.77777778	2.08	0.1596
spot	5	3.83333333	0.76666667	0.90	0.4954

Parameter		Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		6.911111111 B	0.40735835	16.97	<.0001	6.077969737	7.744252485
group	1	0.444444444 B	0.30793397	1.44	0.1596	-0.185351236	1.074240125
group	2	0.000000000 B	.	.	.	.	.
spot	a	0.650000000 B	0.53335728	1.22	0.2328	-0.440838117	1.740838117
spot	b	0.050000000 B	0.53335728	0.09	0.9260	-1.040838117	1.140838117

...

Too small standard errors for estimates of difference between groups and too large standard errors for estimates of differences between spots!

Outline

General variance component models

Generalisations of ANOVA and GLM models involving several sources of random variation, so-called variance components.

Examples of sources of random variation:

- Environmental variation.
  - Between regions, hospitals or work places.
- Biological variation.
  - Between individuals, families or animals.
- Within-individual variation.
  - Between arms, teeth, days.
- Variation due to uncontrollable circumstances.
  - E.g. time of day, temperature, observer.
- Measurement error.

Multilevel models

Variance component models are also called multilevel models.

- Levels are most often hierarchical.
- We have variation, i.e. a variance component, on each level.
- And possibly systematic effects (covariates) on each level.

individual	$\rightarrow$	*context/cluster*	$\rightarrow$	*context/cluster*
level 1	$\rightarrow$	**level 2**	$\rightarrow$	**level 3**
students	$\rightarrow$	classes	$\rightarrow$	schools
patient	$\rightarrow$	clinic	$\rightarrow$	regions
visit	$\rightarrow$	girl	$\rightarrow$	
spot	$\rightarrow$	rabbit	$\rightarrow$	

Merits of multilevel models

We get a better understanding of the various sources of variation.

Effects *within* may be estimated more precisely (higher power), since some sources of variation are eliminated, e.g. by making comparisons within a family. This is analogous to the **paired comparison** situation.

When planning investigations, estimates of the variance components are needed in order to compare the power of various designs, and help us decide

- How many replicates do we need at each level?
- Should we randomize entire clusters or randomize *within* the clusters?

Design considerations

(**Note** in analogy with cluster-randomized trials.)

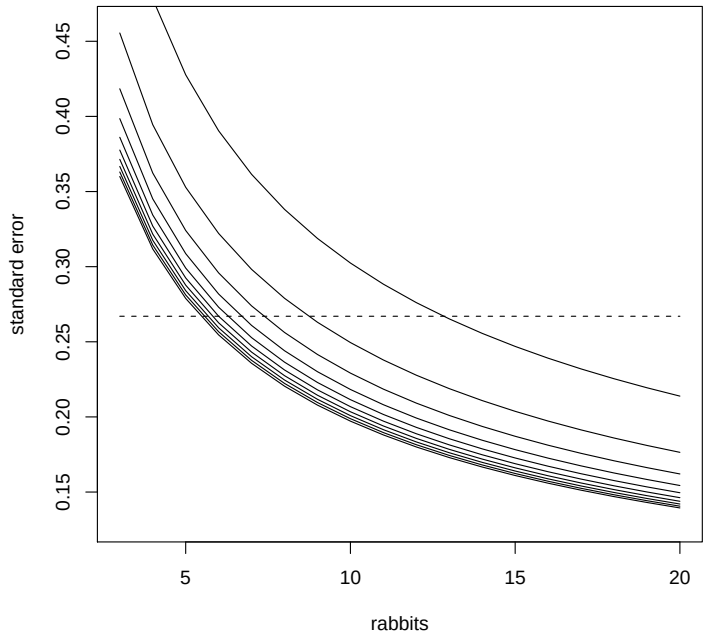
Plan an experiment with:

- R rabbits.
- S spots for each rabbit.
- $R \times S$ measurements.

Std. error of grand mean,

$$\mathrm{var}(\bar{y}) = \frac{\omega_B^2}{R} + \frac{\sigma_W^2}{RS},$$

decreases with R and S.



The different curves correspond to S varying from 1 to 10.

Effective sample size

How many rabbits would we need to obtain the same precision in estimating the grand mean if we had **only one measurement** on each of R_1 rabbits?

Solve an equation to get:

$$R_1 = \frac{R \times S}{1 + \rho(S - 1)}$$

where ρ is the within rabbit correlation.

- Estimate: $\rho = \frac{\omega_B^2}{\omega_B^2 + \sigma_W^2} = \frac{0.3304}{0.3304 + 0.5842} = 0.361 \Rightarrow R_1 = 12.8$

I.e. one measurement on each of thirteen rabbits gives the **same precision** as six measurements on each of six rabbits.

Drawbacks of multilevel models

Their statistical analysis is more difficult.

- When making inference (estimation and testing), it is important to take all sources of variation into account, and effects have to be evaluated against the relevant variation.

If we fail to take the correlation into account, we will experience:

- Possible bias in the mean value estimates.
- Too small standard errors (type 1 error) for estimates of level 2 covariates (between-cluster effects).
- Too large standard errors (type 2 error) for estimates of level 1 covariates (within-cluster effects)

Case: Cortisol and stress-response

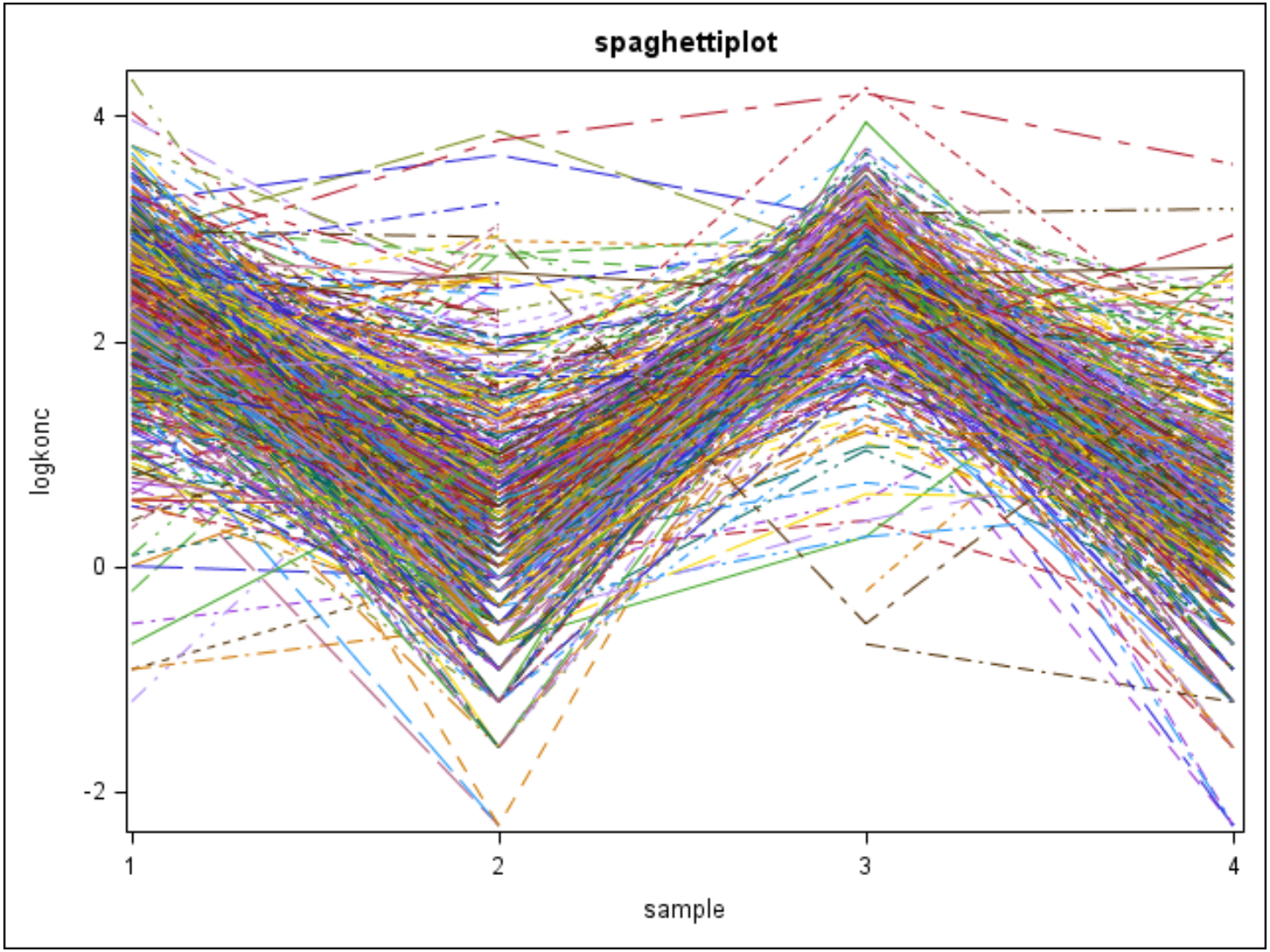
Outcome: Concentration of cortisol in salvia samples taken **mornings and evenings** in workers in *Aarhus amt and kommune* in 2007 (3536 participants, 786 men) with follow-up in 2009 (2408 participants, 520 men).

Interest: effect of stressors: life events, Effort Reward Index.

level	**variation**	**covariates**
3	between persons	`gender, age`
2	within person: between days	`bmi, stressors, year`
1	within person: within days	`timeday (morning/evening)`

Reference: from PRISM study, personal cummunication with Sigurd Mikkelse

Log-transformed concentrations



Three-level model

```
title1 'variance components';
PROC MIXED DATA=prism_men;
  CLASS idnr year (ref='2007') timeday;
  MODEL logkonc = timeday year / SOLUTION CL DDFM=KR;
  RANDOM idnr idnr*year;
RUN;
```

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	6077.88355058	
1	3	6050.14347396	0.00008342
2	1	6050.09026809	0.00000005
3	1	6050.09023526	0.00000000

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Estimate
idnr	0.05943
idnr*year	0
Residual	0.5374

One of the variance component estimates is a **zero**!

Estimated variance components

Level	Variation	Estimate
3	between persons (ω^2)	0.0594 (10.0%)
2	between days (τ^2)	0.0000 (0.0%)
1	within days (σ^2)	0.5374 (90.0%)
	Total	0.5984 (100%)

Level 2 covariates (stressors) can only have **very little impact on individual cortisol koncentrations**!

Negative variance components

In case on of the variance component estimates becomes negative, SAS repports a zero.

What does it mean?

- The zero-estimate may be a chance finding due to statistical uncertainty.
- Or it might be the result of truly negative correlation within clusters - e.g. competition between plants grown in same pot.

What can we do about it?

- Re-fit the model without the problematic random effect.
- Use an unstructured covariance allowing negative correlation
- Include more level 1 covariates, e.g. exact sampling time.

Systematic effects

Solution for Fixed Effects

Effect	year	timeday	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept			2.3916	0.02494	2382	95.88	<.0001	0.05	2.3426	2.4405
timeday		evening	-2.0137	0.02869	1802	-70.19	<.0001	0.05	-2.0699	-1.9574
timeday		morning	0	.	.	.	.	.	.	.
year	2009		0.08465	0.03016	2421	2.81	0.0051	0.05	0.02550	0.1438
year	2007		0	.	.	.	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
timeday	1	1802	4927.33	<.0001
year	1	2421	7.88	0.0051

Cortisol is measured on **log-scale**. Backtransformation $\exp(2.0137) \simeq 7.49$ yields that median levels of kortisol is an estimated 7.5 times higher in the morning than in the evening.

Exact time of measurement should be taken into account!!!

Explained variation (R^2)

We consider only the simplest case, i.e. the two-level model

- we have several variances that can be explained.

Variation within individuals (residual variation):

- decreases when we include an important level 1 covariate (x_1)
- may also decrease when we include an important level 2 covariate (x_2).

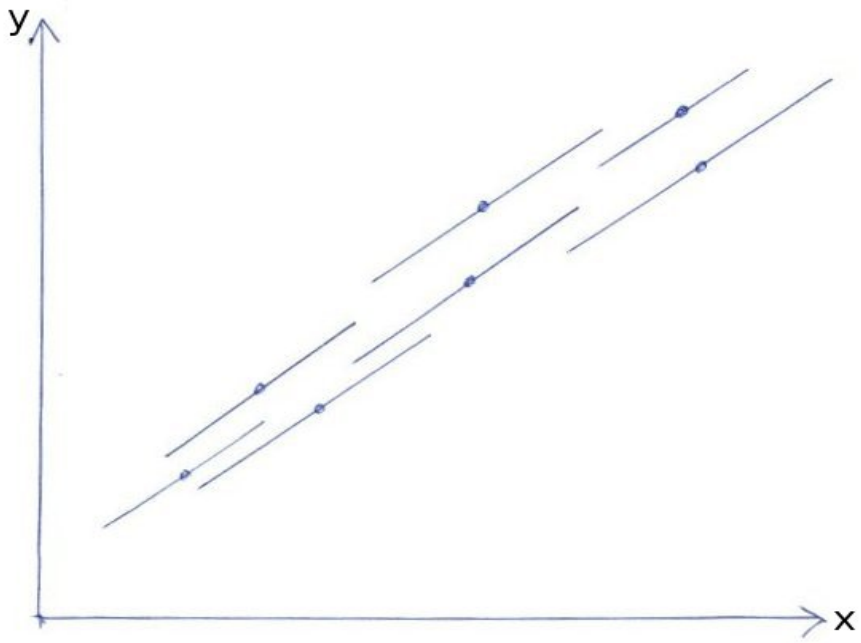
Variation between individuals:

- decreases when we include an important level 2 covariate (x_2)
- *may increase or decrease* when we include an important level 1 covariate (x_1)

Total variance decreases when including an important covariate

Hypothetical example I

Covariate x_1 varies between individuals, and the variation in individual averages ($\bar{y}$) is mostly due to this variation.

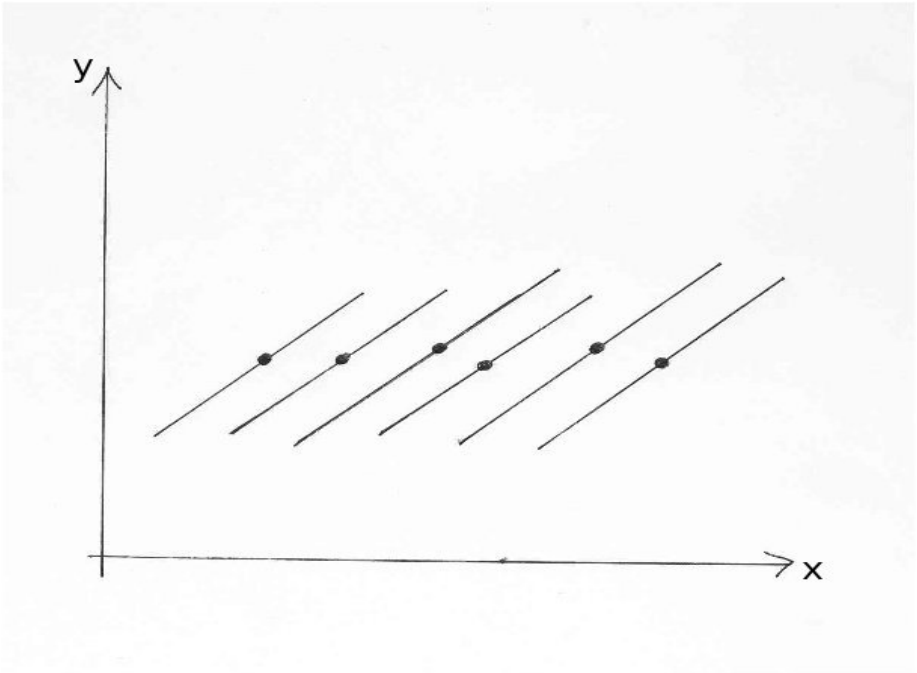


Levels of y, for fixed x are quite alike:

- ω_B^2 **decreases** when x_1 is included.

Hypothetical example II

Covariate x_1 vary between individuals, but the average outcomes ($\bar{y}$) are almost identical:



Levels of y, for fixed x are very different:

- ω_B^2 **increases** when x is included.

Technical explanation*

A balanced design (same number of observations per cluster):

Explicit solution for the two-level model:

$$\tilde{\sigma}^2_W = \mathrm{MS}_W \quad \text{and} \quad \tilde{\omega}^2_B = \mathrm{MS}_B - \frac{\mathrm{MS}_W}{n}$$

- MS_W and MS_B are Mean Squares within and between clusters, defined as in one-way ANOVA.
- n is the number of observations per cluster.

This is deduced from $E(\mathrm{MS}_B) = n\omega^2_B + \sigma^2_W$ and $E(\mathrm{MS}_W) = \sigma^2_W$.

Outline

Ecological analyses

The easy way of dealing with repeated measurements:

- Compute summary statistics for each cluster/individual.
- Perform a traditional analysis on the sample of summary statistics rightfully assuming that these are independent.

Summary statistics could be:

- Sample mean or standard deviation.
- AUC (area under the curve).
- Intercept and slope of regression line.

BUT: Beware of loosing important information.

Ecological vs two-level analysis

Blood pressure and social inequity:
15569 women in 17 regions of Malmø.

Covariates:

- Individual (level 1):
  - low educational achievement (x)
    (less than 9 years of school)
  - age group
- Regional (level 2):
  - rate of people with low educational achievement (z)
    from the 'Skåne Council Statistics Office'
    An aggregated covariate.

Reference: Merlo et al (2001), J. Epidemiol Community Health **55**.

Abstract

***Study objectives*—To study geographical differences in diastolic blood pressure and the influence of the social environment (census percentage of people with low educational achievement) on individual diastolic blood pressure level, after controlling for individual age and educational achievement. To compare the results of multilevel and ecological analyses.**

***Design*—Cross sectional analysis performed by multilevel linear regression modelling, with women at the first level and urban areas at the second level, and by single level ecological regression using areas as the unit of analysis.**

***Setting*—Malmö, Sweden (population 250 000).**

***Participants*—15 569 women aged 45 to 73, residing in 17 urban areas, who took part in the Malmö Diet and Cancer Study (1991–1996).**

Ecological analysis

Average blood pressure in region vs rate of people with low educational achievement.

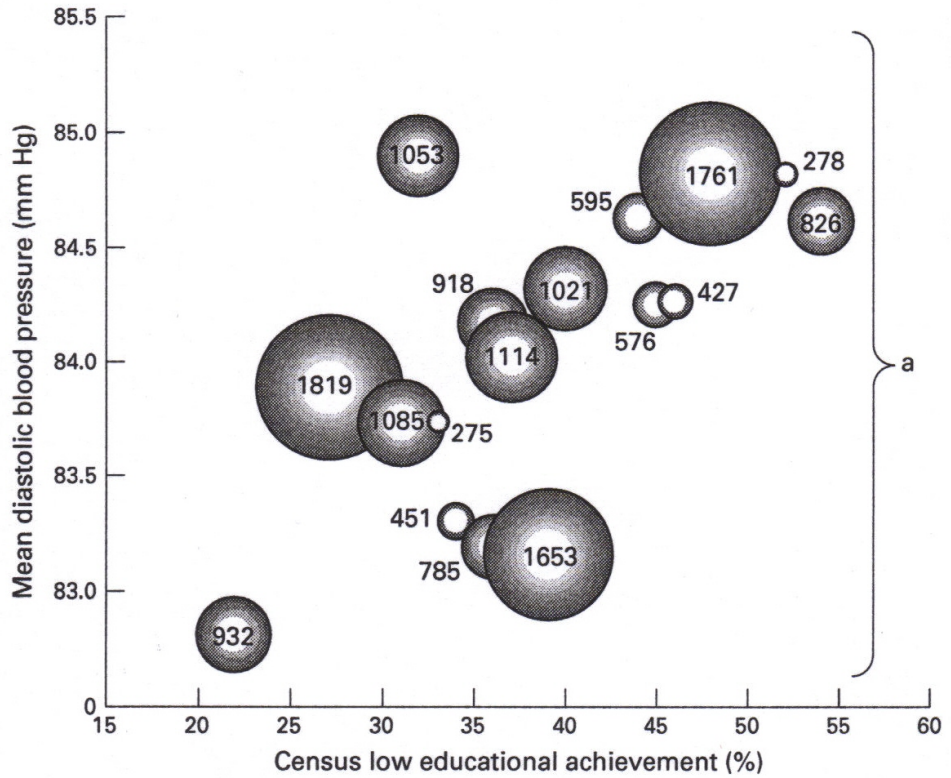


Figure 1 Actual ecological association between mean diastolic blood pressure and census percentage of people with low educational achievement in 17 urban areas of Malmö. The size of the circles corresponds to the number of women. This figure permits observation of the variance between the areas (a).

Size of circle indicates size of investigation.

Estimated slope:
4.66 (SE 1.42).

Seems an important explanatory variable?!?

Estimates from two-level model

What is the effect of individual educational achivement (x_1) vs regional educational achievement (x_2)?

Included covariates	Estimate (SE) x_1 (individual)	Estimate (SE) x_2 (region)	Variation between regions	Variation within regions	R^2 (of total)
none			0.35	96.03	0% (ref)
age			0.26	92.21	26%
x_1, age	1.15 (0.17)		0.14	91.83	59%
x_2, age	-	4.06 (1.35)	0.12	91.48	65%
x_1, x_2, age	1.09 (0.17)	2.97 (1.25)	0.09	91.26	75%

Table 1 in Merlo et al. (2001)

Ecological analysis vs the two-level model

Region as a random effect could only account for 0.36% of the variation in blood pressures (0.35 of 0.35+96.03).

Thus, regional variables such as *rate of low-income* will have very little impact on individual blood presures!

The ecological analysis '*sums up*' the individual and the regional effects, but is *not able to distinguish* between the two.

- It overestimates the level 2 effect.
- It cannot be interpreted as a level 1 effect.

Individual vs regional blood presure

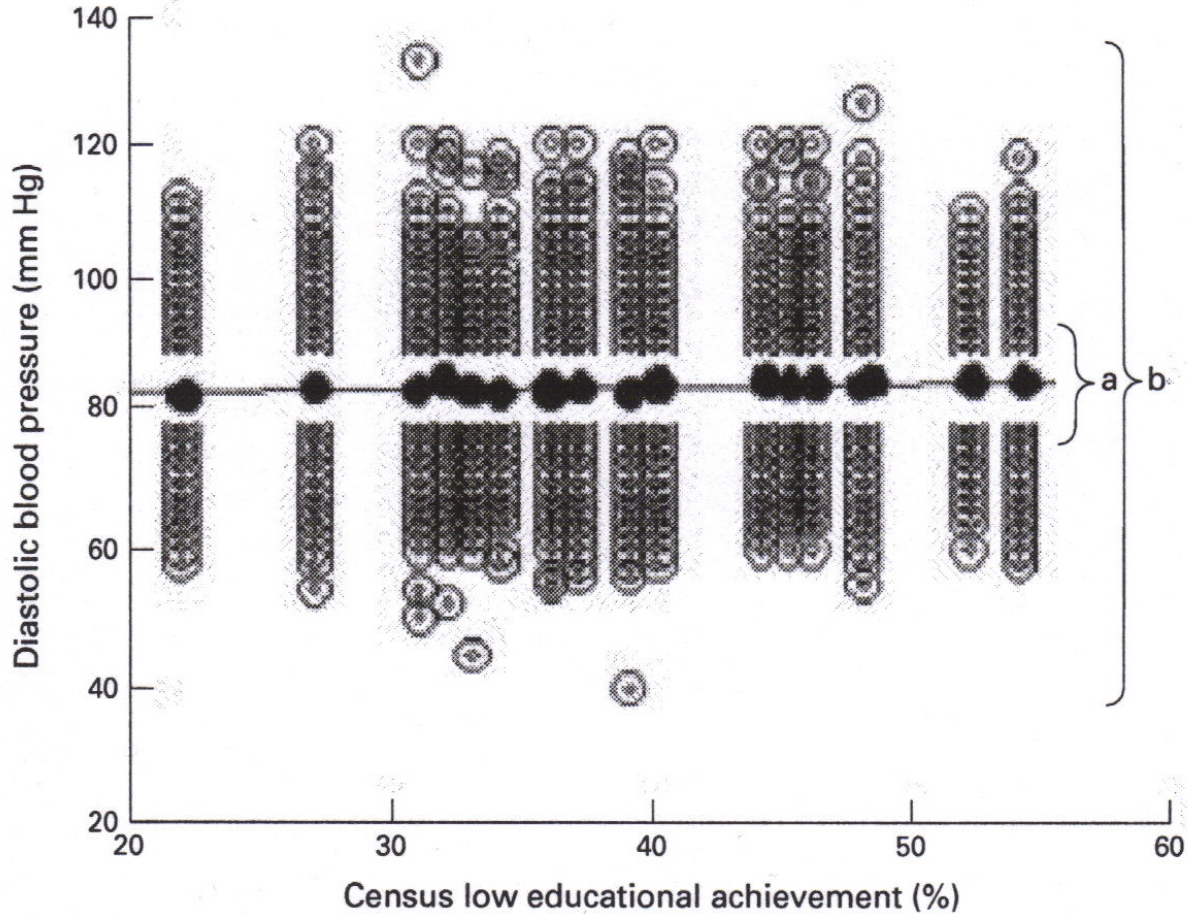


Figure 2 Actual individual association between individual diastolic blood pressure and census percentage of people with low educational achievement in 17 urban areas of Malmö. The same ecological association as in figure 1 is represented as black circles. This figure permits observation of the variance between the areas (a) (see also fig 1) and between the individual women (b).

Example: suicide and religion

Ecological analysis: Percent of suicides increases with percent of protestants in region.

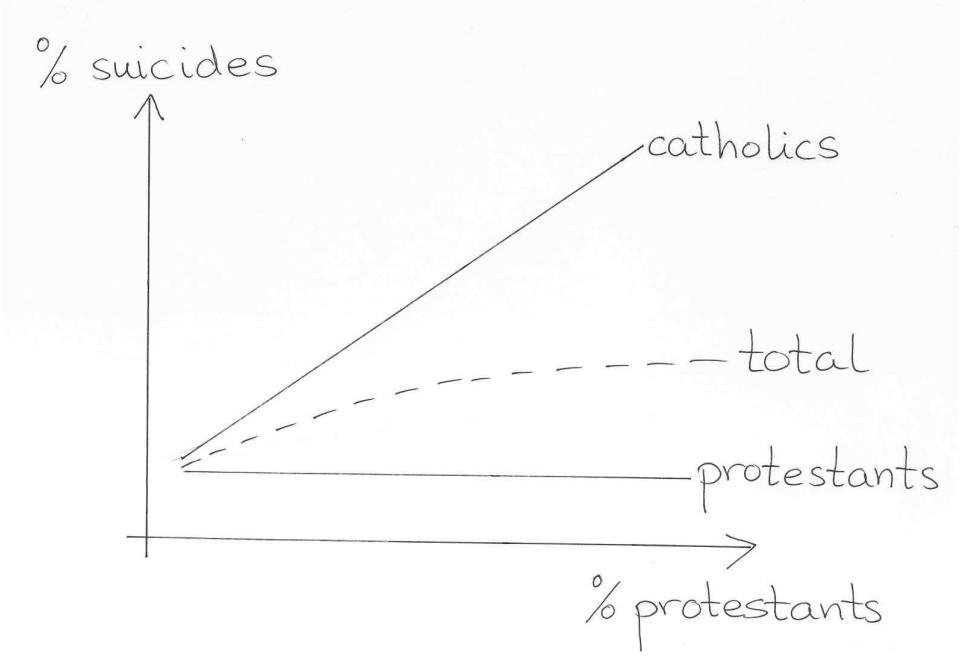
- Are protestants more likely to commit suicide?

Two-level model:

level	unit	variation	covariates
1	individuals	within region, σ^2_W	`religion`, x
2	regions	between regions, ω^2_B	`% protestants`, z

Finding: Interaction between individual effect (x) and region covariate (z) ...

Another example: suicide and religion



More suicides among catholics in regions with many protestants.

Outline

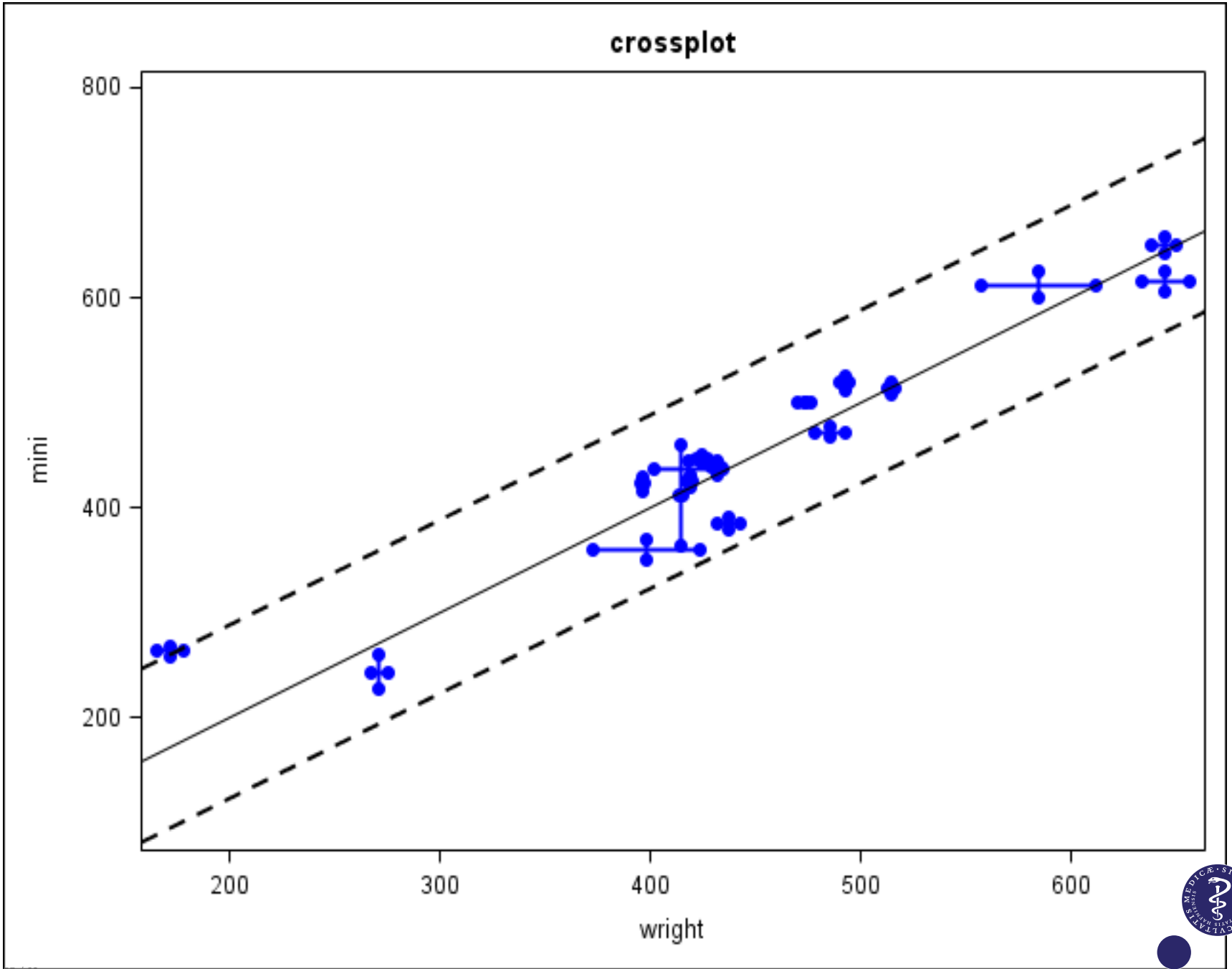
Comparing measurement devices

Example: Peak expiratory flow rate, l/min:

- 17 subjects, 2 measurement devices,
- two replicates with **each method**.

subject	*Wright*		*mini Wright*	
id	Y_{1p1}	Y_{1p2}	Y_{2p1}	Y_{2p2}
1	494	490	512	525
2	395	397	430	415
3	516	512	520	508
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
15	178	165	259	268
16	423	372	350	370
17	427	421	451	443
Average	450.35	445.41	452.47	455.35
SD	116.31	119.61	113.12	111.32

Reference: Bland and Altman, Lancet (1986).


crossplot
800
600
400
200
mini
200
300
400
500
600
wright

Aim of investigation

Quantify the **precision** of each measuring device

- Repeatability (variability=measurement error)

Quantify the **agreement** between the two devices.

- Bias of one method compared to the other.
- Variance of one method compared to the other.

Can the devices be used interchangably?

Simple approaches

For reliability of **each method separately** we could:

- make Bland Altman plots of differences vs averages.
- compute limits of agreement, i.e. the 95% normal range of the differences.

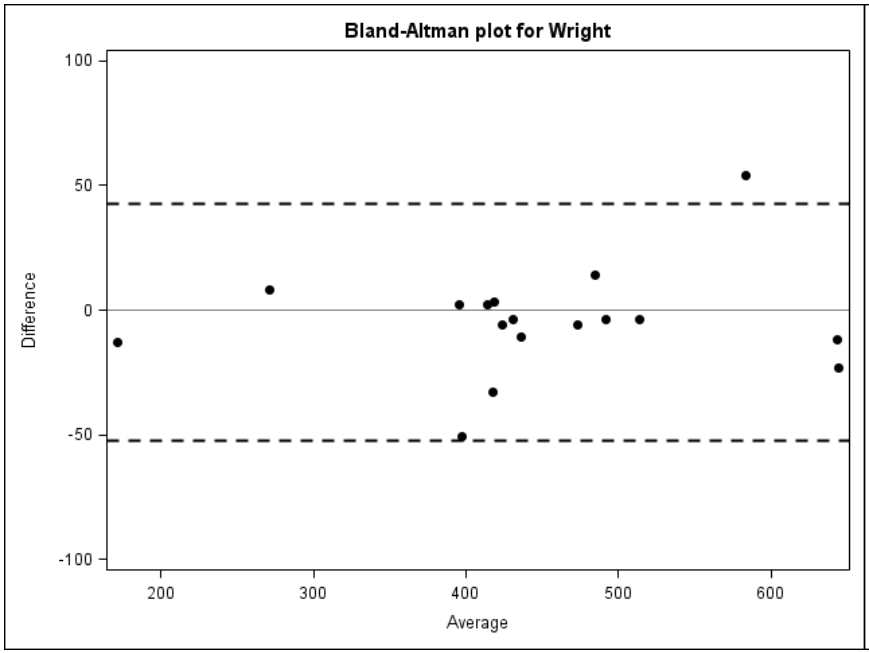
For reproducibility (method comparison) we might:

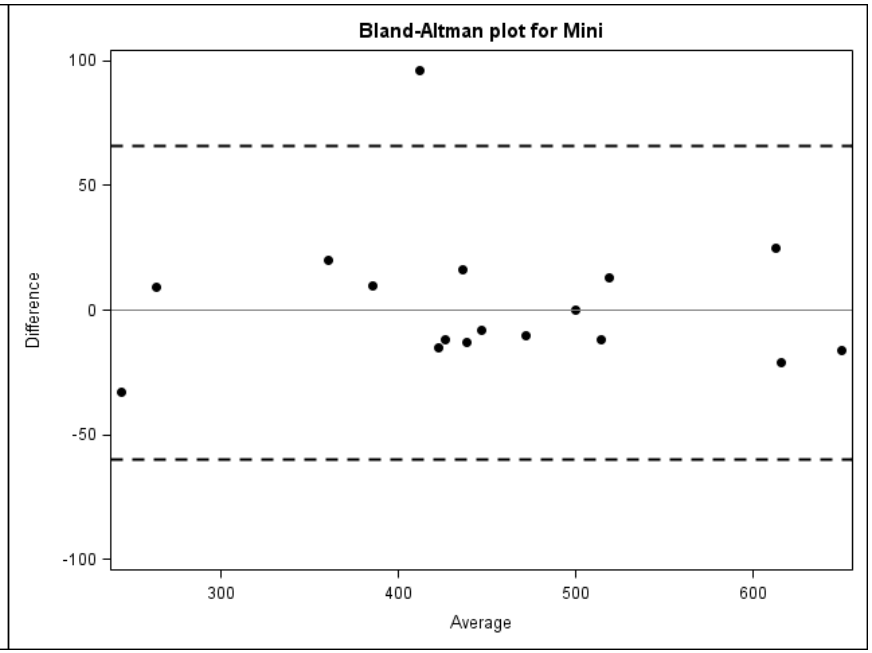
- compare the **averages** in a Bland-Altman plot . . . ?
- Not good - unless you also do averages in clinic!

For both at the same time:

- Mixed model for variance between and within methods.

Repeatability





Method	Estimated bias	95% limits of agreement
Wrigth	-4.94 (-16.11;6.22)	(-52.33;42.45)
Mini Wright	2.88 (-11.96;17.73)	(-60.11;65.86)

Two-level models

For each method $(i = 1, 2)$ we have a two-level model

$$Y_{ijk} = \mu_i + a_{ij} + \varepsilon_{ijk}$$

- μ_i population mean as anticipated by method i.
- a_{ij} deviation of subject j from population mean, assumed normally distributed $N(0, \sigma_i^2)$.
- ε_{ijk} deviation for replicate k (measurement error), assumed normally distributed $N(0, \omega_i^2)$.

PROC MIXED: Stratified analyses

```
PROC MIXED DATA=wright;  BY method;
CLASS id;
MODEL flow = / SOLUTION CL;
RANDOM id;
RUN;
```

method=mini

Cov Parm	Subject	Estimate
Intercept	id	12188
Residual		396.44

Effect	Estimate	Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept	453.91	26.9921	16	16.82	<.0001	0.05	396.69	511.13

method=wright

Cov Parm	Subject	Estimate
Intercept	id	13683
Residual		234.29

Effect	Estimate	Error	DF	t Value	Pr > |t|	Alpha	Lower	U
Intercept	447.88	28.4914	16	15.72	<.0001	0.05	387.48	50

Joint model for both methods

For methods $(i = 1, 2)$:

$$Y_{ijk} = \mu_i + a_{ij} + \varepsilon_{ijk}$$

- ε_{ijk} assumed normally distributed $N(0, \omega_i^2)$ and **independent across methods**.
- a_{ij} assumed normally distributed $N(0, \sigma_i^2)$ and correlated with $\rho = \mathrm{Cor}(a_{i1}, a_{i2})$.

Anticipated means for the same subject ought to look a lot like each other, so the a_{ij}'s are likely to be correlated across methods.

- Note that SAS models the covariance parameter $\sigma_{12} = \mathrm{Cov}(a_{1j}, a_{2j}) = \sigma_1 \cdot \sigma_2 \cdot \rho$.

PROC MIXED: Joint analysis

```
PROC MIXED DATA=wright;
CLASS method id;
MODEL flow=method / SOLUTION CL;
RANDOM method / TYPE=UN SUBJECT=id;
REPEATED / TYPE=simple GROUP=method SUBJECT=id*method;
RUN;
```

Covariance Parameter Estimates

Cov Parm	Subject	Group	Estimate
UN(1,1)	id		12188
UN(2,1)	id		12542
UN(2,2)	id		13683
Residual	method*id	method mini	396.44
Residual	method*id	method wright	234.29

Solution for Fixed Effects

Effect	method	Estimate	StdError	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Intercept		447.88	28.4914	32	15.72	<.0001	0.05	389.85	505.92
method	mini	6.0294	8.0532	32	0.75	0.4595	0.05	-10.3744	22.4332
method	wright	0	.	.	.	.	.	.	.

Repeatability

Typical differences (approximate 95% normal range) between two measurement with the **same method**:

$$\text{Wright: } \hat{\omega}_1^2 = 234.29 \rightarrow \pm 2\sqrt{2\omega_1^2} \simeq \pm 43.3$$

$$\text{Mini: } \hat{\omega}_2^2 = 396.44 \rightarrow \pm 2\sqrt{2\omega_2^2} \simeq \pm 56.3$$

Seemingly Wright is more precise, but is the difference significant?

$$F = \frac{396.44}{234.29} = 1.69 \sim F(17, 17) \rightarrow P = 0.14$$

Don't form too firm a conclusion with too small data.

Reproducibility

No evidence of **systematic** differences between the two methods.

- Estimated bias +6.0 (-10.4;22.4) for mini vs wright. P=0.46.

Typical differnces between the two methods:

$$\begin{aligned}
\mathrm{var}(Y_{1jk} - Y_{2jk}) &= \mathrm{var}(a_{1j} - a_{2j} + \varepsilon_{1jk} - \varepsilon_{2jk}) \\
&= \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} + \omega_1^2 + \omega_2^2 \\
&= 12188 + 13683 - 2 \cdot 12542 + 396.44 + 234.29 \\
&= 1417.73
\end{aligned}$$

Limits-of-agreement: $6.03 \pm 2\sqrt{1417.7} = (-69.3, 81.3)$.

Not a multi-level model!

level	variation	covariates
3	between subjects (ω^2)	
2	between methods (τ^2)	method
1	within methods (σ^2)	

Specified as:

$$Y_{ijk} = \mu_j + a_i + b_{ij} + \varepsilon_{ijk}$$

- $A_i \sim \mathcal{N}(0, \omega^2)$ for subjects $i = 1, \ldots, 17$,
- $B_{ij} \sim \mathcal{N}(0, \tau^2)$ for methods $j = 1, 2$,
- $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$ for replicate $k = 1, 2$.

This is assuming the same variance for both methods.

Estimated variance components

```
PROC MIXED DATA=wright;
  CLASS method id;
  MODEL flow=method / SOLUTION CL;
  RANDOM intercept method / SUBJECT=id;
RUN;
```

```
 Covariance Parameter Estimates
Cov Parm      Subject    Estimate
Intercept     id            12542
method        id           393.57
Residual                   315.37

           Fit Statistics
-2 Res Log Likelihood          676.0
AIC (smaller is better)        681.6
```

What does this tell us about the precision of the measurements?

Typical differences

Between replicate measurements using the same method:

$$\begin{aligned} Y_{ijk_1} - Y_{ijk_2} &= \varepsilon_{ijk_1} - \varepsilon_{ijk_2} \\ &\sim \mathcal{N}(0, 2\sigma^2) \end{aligned}$$

Limits-of-agreement: $\pm 2\sqrt{2\sigma^2} \simeq \pm 50.23.$

Between measurements using the different methods:

$$\begin{aligned} Y_{ij_1k_1} - Y_{ij_2k_1} &= \mu_{j_1} - \mu_{j_2} + b_{ij_1} - b_{ij_2} + \varepsilon_{ij_1k_1} - \varepsilon_{ij_2k_1} \\ &\sim \mathcal{N}(\mu_{j_1} - \mu_{j_2}, 2\tau^2 + 2\sigma^2) \end{aligned}$$

Limits-of-agreement: $\mu_1 - \mu_2 \pm 2\sqrt{2\tau^2 + 2\sigma^2} \simeq 6.03 \pm 75.31.$

(where we include the non-significant systematic difference).

Systematic difference?

Solution for Fixed Effects

Effect	method	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		447.88	27.7519	16	16.14	<.0001
method	mini	6.0294	8.0532	16	0.75	0.4649
method	wright	0	.	.	.	.

Conclusion: No evidence of **systematic** differences between the measurement methods.

BUT: Do we really want to assume that variances are equal when the power for testing this is poor?