Statistics in Practice: Longitudinal Data Analysis

Geert Verbeke

geert.verbeke@med.kuleuven.be

Geert Molenberghs

geert.molenberghs@uhasselt.be

Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat)

Katholieke Universiteit Leuven & Universiteit Hasselt, Belgium

www.ibiostat/be

universiteit hasselt | I-BioStat | KU LEUVEN

Interuniversity Institute for Biostatistics
and statistical Bioinformatics

Bremen, March 13, 2014

Case Study 1: Lizard Data

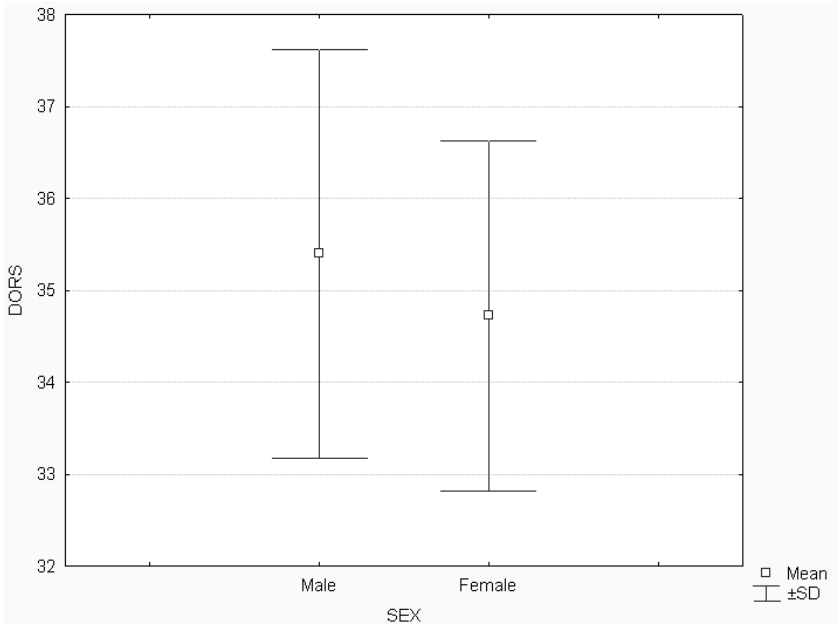
- Example
- Two-way ANOVA
- Mixed models
- Fitting mixed models in SAS
- Remarks

Lizard Data

- Data on 102 lizards
- Response of interest: Number of dorsal shells
- Research question:

> Is number of dorsal shells gender-related ?

- Graphically:

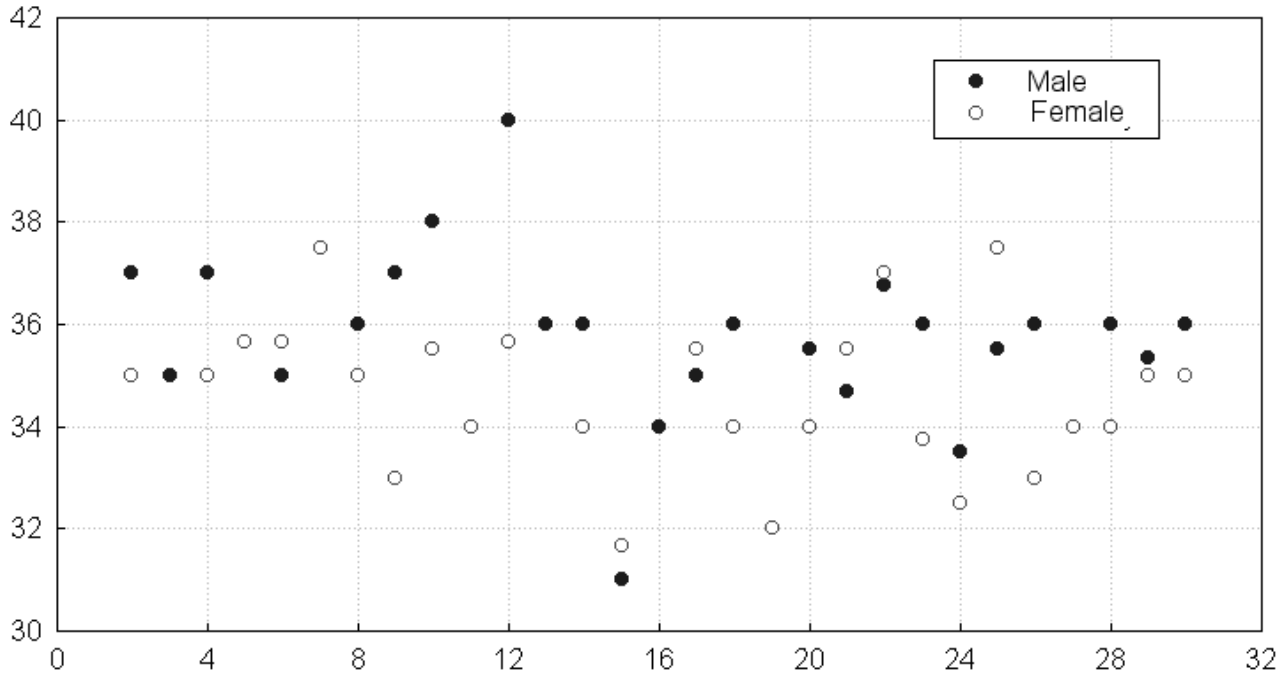


- Two-sample t-test:

Variable	T-tests; Grouping: SEX (schildformengels.STA) Group 1: Male Group 2: Female **Mean Male**	Mean Female	t-value	df	p
DORS	35,39583	34,72222	1,648480	100	0,102393

- Hence, the small observed difference is not significant ($p = 0.1024$).
- A typical aspect of the data is that some animals have the same mother.
- We have 102 lizards from 30 mothers
- Mother effects might be present
- Hence a comparison between male and female animals should be based on within-mother comparisons.

- Graphically:



- Observations:
  - ▷ Much between-mother variability
  - ▷ Often, males (considerably) higher than females
  - ▷ In cases where females higher than males, small differences

- Hence the non-significant t-test result may be due to the between-mother variability
- This is an example of **clustered data**: observations are clustered within mothers
- It is to be expected that measurements within mothers are more alike than measurements from different mothers.
- We expect correlated observations within mothers and independent observations between mothers.
- How to correct for differences between mothers ?

Two-way ANOVA

- An obvious first choice to test for a 'sex' effect, correcting for 'mother' effects, is 2-way ANOVA with factors 'sex' and 'mother'.

- The mother effect then represents the variability between mothers.

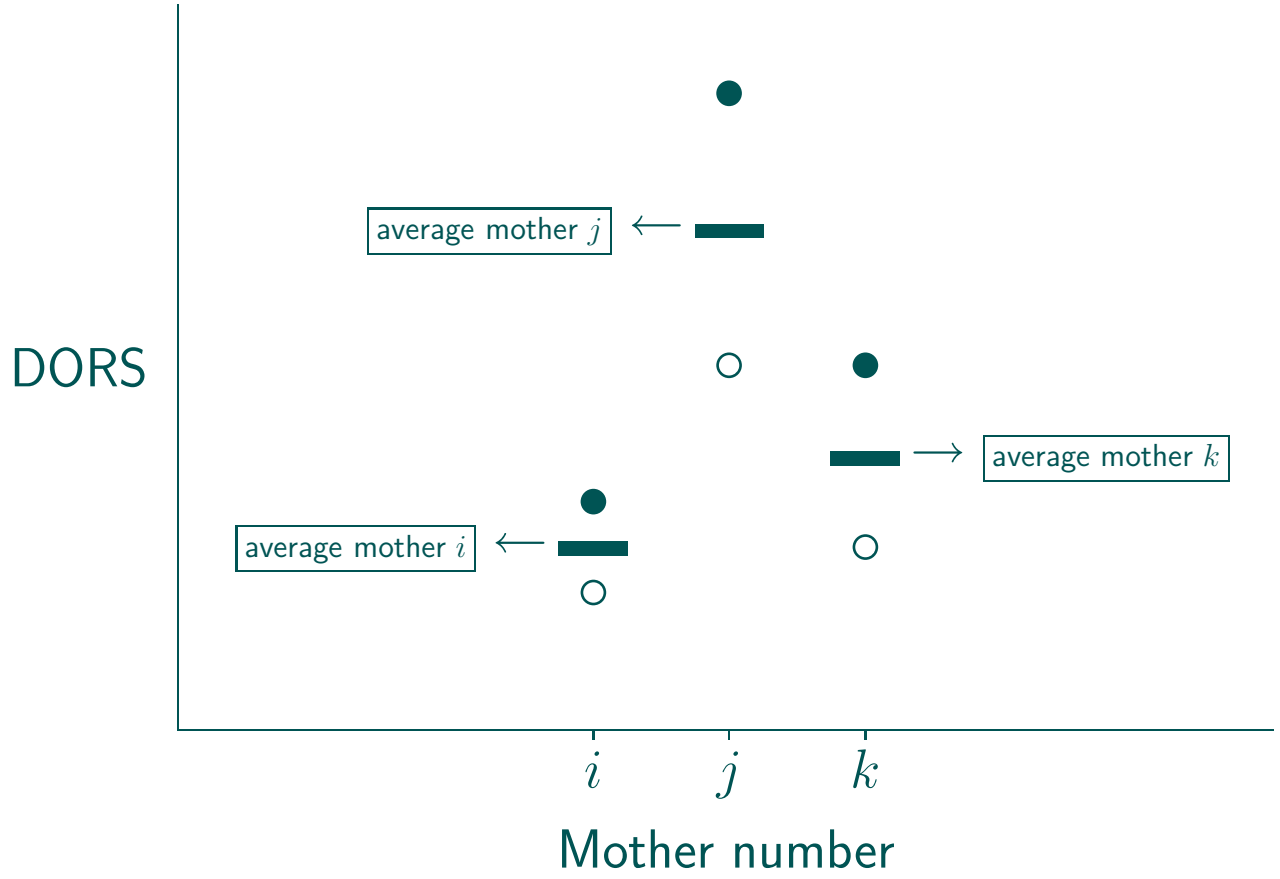
- Let Y_{ij} be the jth measurement on the ith mother, and let t_{ij} be 1 for males and 0 for females.

- The model then equals: $$Y_{ij} = \mu + \alpha_i + \beta t_{ij} + \varepsilon_{ij}$$

- β is the parameter of interest, and we need the usual restrictions on the parameters α_i, e.g., $\Sigma_i \alpha_i = 0$

- Residual distribution: $\varepsilon_{ij} \sim N(0, \sigma^2_{res})$

- Graphically:



- SAS program:

```
proc glm data = lizard;
class sex mothc;
model dors = sex mothc;
run;
```

- Relevant SAS output:

```
                    Class Level Information

Class       Levels  Values

SEX              2  1 2

MOTHC           30  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


Dependent Variable: DORS

                                    Sum of
Source                      DF          Squares     Mean Square     F Value    Pr > F

Model                       29      268.4685062       9.2575347        3.98    <.0001

Error                       72      167.3746310       2.3246477

Corrected Total            101      435.8431373


R-Square     Coeff Var      Root MSE     DORS Mean

0.615975      4.351352      1.524680      35.03922
```

Source	DF	Type III SS	Mean Square
SEX	1	16.7253690	16.7253690
MOTHC	28	256.9378690	9.1763525

Source	F Value	Pr > F
SEX	7.19	0.0091
MOTHC	3.95	<.0001

- Note the highly significant mother effect.
- We now also obtain a significant gender effect.
- Many degrees of freedom are spent to the estimation of the mother effect, which is not even of interest.

Mixed Models

- Note the different nature of the two factors:
  - ▷ SEX: defines 2 groups of interest
  - ▷ MOTHER: defines 30 groups not of real interest. A new sample would imply other mothers.
- In practice, one therefore considers the factor 'mother' as a **random factor**.
- The factor 'sex' is a **fixed effect**.
- Thus the model is a **mixed model**.
- In general, models can contain multiple fixed and/or random factors.

- The model is still of the form:

$$Y_{ij} = \mu + \alpha_i + \beta t_{ij} + \varepsilon_{ij}$$

- But the fact that mothers can be assumed to be randomly selected from a population of mothers is reflected in the additional assumption

$$\alpha_i \sim N(0, \sigma^2_{moth})$$

- Note that we still have that the α_i have mean zero. Before, we had the restriction $\Sigma_i \, \alpha_i = 0$

Fitting Mixed Models in SAS

- Mixed model with 'sex' as fixed and 'mother' as random effect:

```
proc mixed data = lizard;
class sex mothc;
model dors = sex;
random mothc;
run;
```

- Fixed effects are specified in the MODEL statement.

- Random effects are specified in the RANDOM statement.

- Relevant SAS-output:

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	436.17789431	
1	3	407.96849207	0.00072385
2	1	407.88032382	0.00001530
3	1	407.87858406	0.00000001

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Estimate
MOTHC	1.7799
Residual	2.2501

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
SEX	1	72	6.63	0.0121

- Covariance parameter estimates:
  - ▷ Total variability, correcting for gender is decomposed as:

$$\sigma^2 = \sigma^2_{moth} + \sigma^2_{res}$$

$$4.03 = 1.78 + 2.25$$

  - ▷ σ^2_{moth} represents the variability between mothers
  - ▷ σ^2_{res} represents the variability within mothers
  - ▷ The 'mother' factor explains $1.78/4.03 = 44\%$ of the total variability, after correction for gender

- Note the significant difference between male and female animals ($p = 0.0121$)

- With the t-test, ignoring the mother effect, this was $p = 0.1024$.

- The mixed model implies a specific correlation structure:
  - ▷ Observations from different mothers are independent.
  - ▷ Observations within mothers are positively correlated:

$$\text{corr(within mother)} = \frac{\sigma^2_{moth}}{\sigma^2_{moth} + \sigma^2_{res}} = \frac{1.78}{1.78 + 2.25} = 0.44$$

Remarks

- The simplest example of clustered data are paired observations, typically analyzed using a paired t-test.

- In our example, this would mean that we have exactly one male and one female animal per mother.

- The mixed models can be viewed as an extension of the paired t-test to :
  - ▷ more than 2 observations per cluster
  - ▷ unbalanced data: unequal number of measurements per cluster
  - ▷ models with covariates, e.g., 'sex', or others
  - ▷ models with multiple random effects (see later)

Case Study 2: Growth Curves

- ▷ Example
- ▷ The model
- ▷ ESTIMATE and CONTRAST statements
- ▷ Random intercepts model
- ▷ Remarks
- ▷ The linear mixed model

Growth Curves

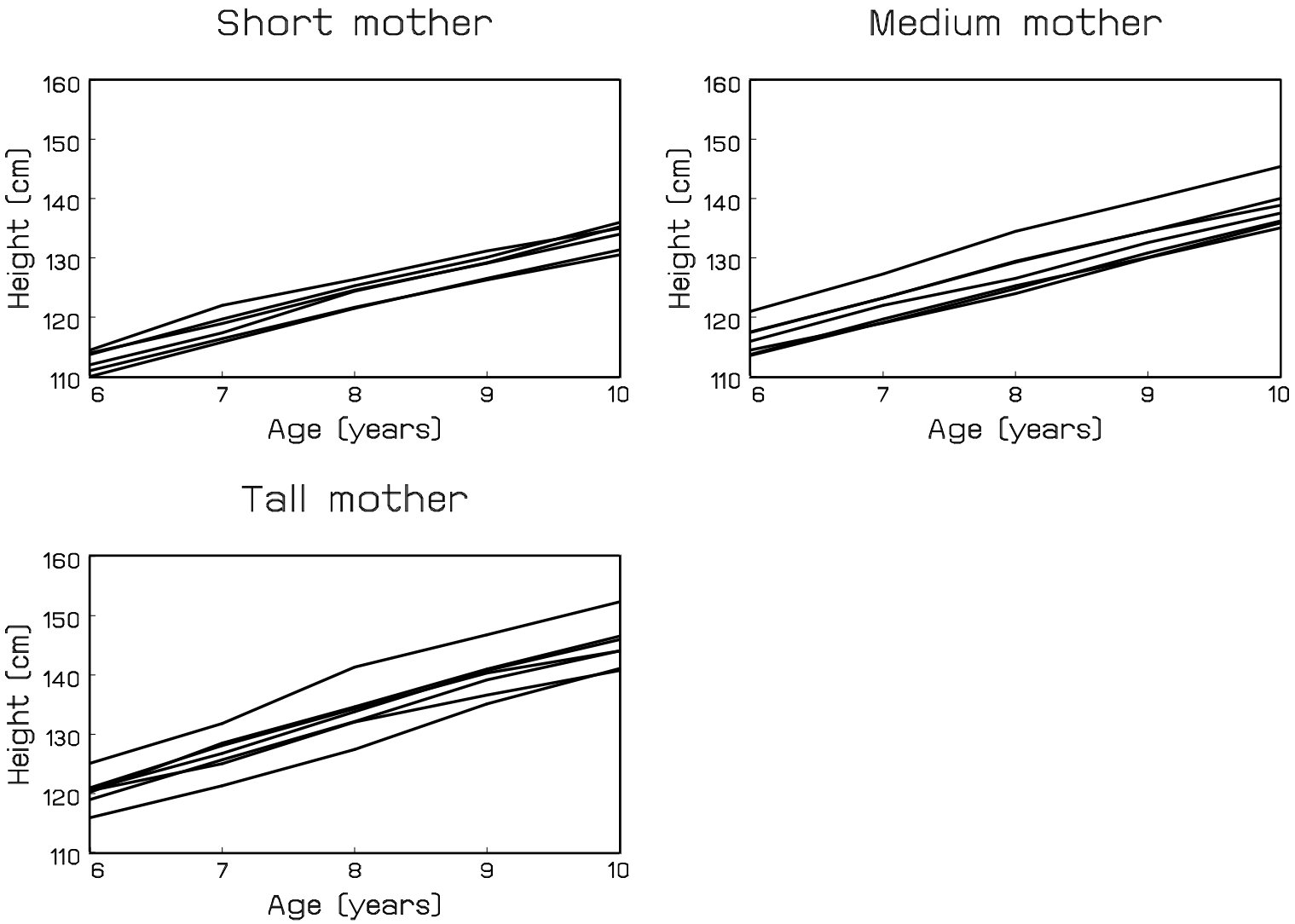
- Taken from Goldstein (1979).
- Research question:

  Is growth related to height of mother ?

- The height of 20 schoolgirls, with small, medium, or tall mothers, was measured over a 4-year period:

	Mothers height	Children numbers
Small mothers	< 155 cm	$1 \rightarrow 6$
Medium mothers	$[155\text{cm};\ 164\text{cm}]$	$7 \rightarrow 13$
Tall mothers	> 164 cm	$14 \rightarrow 20$

- Individual profiles:



- Remarks:
  - ▷ Almost perfect linear relation between Age and Height
  - ▷ Much variability between girls
  - ▷ Little variability within girls
  - ▷ Fixed number of measurements per subject
  - ▷ Measurements taken at fixed time points

The Model

- As for the lizard data, the observations are clustered within children.
- Correction for the variability between children is done through a random child effect.
- Further, we will assume a linear relation between Age and Height, possibly different for the different groups.
- Ignoring the clustered nature of the data, the following ANOCOVA could be used:

```
proc glm data = growth;
class group;
model height = age group age*group;
run;
```

- Inclusion of a random child effect is obtained by:

```
proc mixed data = growth;
class group child;
model height = age group age*group / solution;
random child;
run;
```

- As before, let Y_{ij} be the jth measurement of height for the ith cluster (child), taken at time t_{ij} (age). Our model is then of the form:

$$Y_{ij} = \begin{cases} \beta_1 + b_i + \beta_2 t_{ij} + \varepsilon_{ij}, & \text{if short mother} \\ \beta_3 + b_i + \beta_4 t_{ij} + \varepsilon_{ij}, & \text{if medium mother} \\ \beta_5 + b_i + \beta_6 t_{ij} + \varepsilon_{ij}, & \text{if tall mother} \end{cases}$$

- As before, it is assumed that random effects b_i are normal with mean zero and variance σ^2_{child}.

- The errors ε_{ij} are normal with mean zero and variance σ^2_{res}.

- Relevant SAS output:

Covariance Parameter Estimates

Cov Parm	Estimate
CHILD	8.9603
Residual	0.7696

Solution for Fixed Effects

Effect	GROUP	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		83.1229	1.4162	17	58.69	<.0001
AGE		6.2486	0.1049	77	59.59	<.0001
GROUP	1	-1.8229	2.0846	77	-0.87	0.3846
GROUP	2	-0.1486	2.0028	77	-0.07	0.9411
GROUP	3	0	.	.	.	.
AGE*GROUP	1	-0.9786	0.1543	77	-6.34	<.0001
AGE*GROUP	2	-0.6814	0.1483	77	-4.60	<.0001
AGE*GROUP	3	0	.	.	.	.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
AGE	1	77	8385.15	<.0001
GROUP	2	77	0.46	0.6330
AGE*GROUP	2	77	21.66	<.0001

- Covariance parameter estimates:
  - ▷ Total variability, correcting for age and group is decomposed as:

$$\sigma^2 = \sigma^2_{child} + \sigma^2_{res}$$

$$9.73 = 8.96 + 0.77$$

  - ▷ σ^2_{child} represents the variability between children
  - ▷ σ^2_{res} represents the variability within children
  - ▷ The 'child' factor explains $8.96/9.73 = 92\%$ of the total variability, after correction for group and age

- Note the significant difference in slopes between the groups ($p < 0.0001$)

- The mixed model again implies a specific correlation structure:
  - ▷ Observations from different children are independent.
  - ▷ Observations within children are positively correlated:

$$\text{corr(within child)} = \frac{\sigma^2_{child}}{\sigma^2_{child} + \sigma^2_{res}} = \frac{8.96}{8.96 + 0.77} = 0.92$$

ESTIMATE and CONTRAST Statements

- As in many other SAS procedures, ESTIMATE and CONTRAST statements can be used to obtain inferences about specific contrasts of the fixed effects.

- Slopes for each group separately, as well as pairwise comparisons are obtained using the following program:

```
proc mixed data=growth;
class child group;
model height = group age*group / noint solution;
random child;
contrast 'small-medium' group*age 1 -1 0;
contrast 'small-tall' group*age 1 0 -1;
contrast 'medium-tall' group*age 0 1 -1;
estimate 'small' group*age 1 0 0 / cl;
estimate 'medium' group*age 0 1 0 / cl;
estimate 'tall' group*age  0 0 1 / cl;
run;
```

- Note the different parameterization for the fixed effects, when compared to the

original program:

```
proc mixed data = growth;
class group child;
model height = age group age*group / solution;
random child;
run;
```

- Relevant SAS output:

The Mixed Procedure

Solution for Fixed Effects

Effect	GROUP	Estimate	Standard Error	DF	t Value	Pr > |t|
GROUP	1	81.3000	1.5297	77	53.15	<.0001
GROUP	2	82.9743	1.4162	77	58.59	<.0001
GROUP	3	83.1229	1.4162	77	58.69	<.0001
AGE*GROUP	1	5.2700	0.1133	77	46.53	<.0001
AGE*GROUP	2	5.5671	0.1049	77	53.10	<.0001
AGE*GROUP	3	6.2486	0.1049	77	59.59	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
GROUP	3	77	3234.13	<.0001
AGE*GROUP	3	77	2845.30	<.0001

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
small	5.2700	0.1133	77	46.53	<.0001	0.05	5.0445	5.4955
medium	5.5671	0.1049	77	53.10	<.0001	0.05	5.3584	5.7759
tall	6.2486	0.1049	77	59.59	<.0001	0.05	6.0398	6.4574

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
small-medium	1	77	3.71	0.0579
small-tall	1	77	40.20	<.0001
medium-tall	1	77	21.12	<.0001

- The differences in slopes is mainly explained from the difference between the third group on one hand, and the other two groups on the other hand.

Random Intercepts Model

- Our fitted model was:

$$Y_{ij} = \begin{cases} (\beta_1 + b_i) + \beta_2 t_{ij} + \varepsilon_{ij}, & \text{if short mother} \\ (\beta_3 + b_i) + \beta_4 t_{ij} + \varepsilon_{ij}, & \text{if medium mother} \\ (\beta_5 + b_i) + \beta_6 t_{ij} + \varepsilon_{ij}, & \text{if tall mother} \end{cases}$$

- This can be interpreted as a ANOCOVA model, but with child-specific intercepts b_i

- Such a b_i represents the deviation of the intercept of a specific child from the average intercept in the group to which that child belongs, i.e., deviation from β_1, β_2, or β_3.

- An alternative way to fit a random intercepts model in PROC MIXED is:

```
proc mixed data = growth;
class group child;
model height = age group age*group / solution;
random intercept / subject=child;
run;
```

- The results are identical to those discussed earlier.

Remarks

- The growth-curve dataset is an example of a longitudinal dataset
- In longitudinal data, there is a natural ordering of the measurements within clusters
- The ordering is of primary interest
- Our random-intercepts model implies very strong assumptions:
  - ▷ Parallel profiles within all 3 groups
  - ▷ Constant variance $\sigma^2 = \sigma^2_{child} + \sigma^2_{res}$
  - ▷ Constant correlation within children: $\sigma^2_{child}/(\sigma^2_{child} + \sigma^2_{res})$
- In many longitudinal settings, these assumptions are too restrictive

Linear Mixed Models

- One way to extend the random-intercepts model is to allow also the slopes to be subject-specific:

$$Y_{ij} = \begin{cases} (\beta_1 + b_{1i}) + (\beta_2 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if short mother} \\ (\beta_3 + b_{1i}) + (\beta_4 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if medium mother} \\ (\beta_5 + b_{1i}) + (\beta_6 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if tall mother} \end{cases}$$

- This is an example of the general linear mixed model

- As before, the random effects are assumed to be normally distributed with mean zero:

$$\boldsymbol{b_i} = (b_{1i}, b_{2i})' \sim N(\mathbf{0}, D)$$

- D then equals the 2×2 covariance matrix of the random effects:

$$D = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix}$$

- Interpretation of the parameters:
  - ▷ d_{11} equals the variance of the intercepts b_{1i}
  - ▷ d_{22} equals the variance of the slopes b_{2i}
  - ▷ d_{12} equals the covariance between the intercepts b_{1i} and the slopes b_{2i}.
  - ▷ The correlation between the intercepts and slopes then equals:

$$\text{Corr}(b_{1i}, b_{2i}) = \frac{d_{12}}{\sqrt{d_{11}}\sqrt{d_{22}}}$$

- Random-intercepts models imply constant variance and constant correlation between any two outcomes of the same cluster (see earlier).

- The above model with random intercepts and slopes implies:

  ▷ Variance function:

$$\text{Var}(\boldsymbol{Y_i}(t)) = d_{22}t^2 + 2d_{12}t + d_{11} + \sigma^2$$

  ▷ Correlation function:

$$\text{Corr}(\boldsymbol{Y_i}(t_1), \boldsymbol{Y_i}(t_2)) = \frac{d_{22}t_1t_2 + d_{12}(t_1 + t_2) + d_{11}}{\sqrt{d_{22}t_1^2 + 2d_{12}t_1 + d_{11} + \sigma^2}\,\sqrt{d_{22}t_2^2 + 2d_{12}t_2 + d_{11} + \sigma^2}}$$

- More complicated random-effects structures will yield more complicated variance and correlations functions.

- SAS program:

```
proc mixed data=growth;
class child group;
model height=age group age*group;
random intercept age / type=un subject=child g gcorr;
run;
```

- As before, fixed effects are to be specified in the MODEL statement, while random effects are specified in the RANDOM statement.

- Relevant SAS output:

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
UN(1,1)	CHILD	7.6028
UN(2,1)	CHILD	-0.4437
UN(2,2)	CHILD	0.1331
Residual		0.4758

Estimated G Matrix

Row	Effect	CHILD	Col1	Col2
1	Intercept	1	7.6028	-0.4437
2	AGE	1	-0.4437	0.1331

Estimated G Correlation Matrix

Row	Effect	CHILD	Col1	Col2
1	Intercept	1	1.0000	-0.4412
2	AGE	1	-0.4412	1.0000

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
AGE	1	17	3572.36	<.0001
GROUP	2	60	0.60	0.5514
AGE*GROUP	2	60	9.23	0.0003

- Note the differences in test results for the fixed effects, when compared to the random-intercepts model:

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
AGE	1	77	8385.15	<.0001
GROUP	2	77	0.46	0.6330
AGE*GROUP	2	77	21.66	<.0001

The General Linear Mixed Model

- Let Y_{ij} be response j for cluster i, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$

- Examples:
  - ▷ Y_{ij} is the number of dorsal shells for lizard j within mother i
  - ▷ Y_{ij} is the height of child i at visit j

- The response vector for cluster i equals:

$$\boldsymbol{Y_i} = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$$

- A linear mixed model is a linear regression model for each cluster separately, with fixed as well as random regression coefficients.

- Formally:

$$\begin{cases} \boldsymbol{Y_i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i} \\ \boldsymbol{b_i} \sim N(\mathbf{0}, D), \\ \boldsymbol{\varepsilon_i} \sim N(\mathbf{0}, \sigma^2 I_{n_i}), \\ \boldsymbol{b_1}, \ldots, \boldsymbol{b_N}, \boldsymbol{\varepsilon_1}, \ldots, \boldsymbol{\varepsilon_N} \text{ independent,} \end{cases}$$

- X_i and Z_i are design matrices
- The vector $\boldsymbol{\beta}$ contains all regression parameters which are the same for all clusters
- The vector $\boldsymbol{b_i}$ contains all cluster-specific parameters
- $\boldsymbol{\beta}$ describes average trends in the population

- $\boldsymbol{b_i}$ describes how a specific cluster deviates from the average trend
- As before, the $\boldsymbol{b_i}$ are normally distributed with mean zero and covariance matrix D
- The vector $\boldsymbol{\varepsilon_i}$ contains the measurement error components which are normally distributed with mean zero and variance σ^2
- Terminology:
  - ▷ **Fixed effects:** $\boldsymbol{\beta}$
  - ▷ **Random effects:** $\boldsymbol{b_i}$
  - ▷ **Variance components:** σ^2 and all elements in D

Case Study 1: The Lizard Data

- Our model was given by:

$$Y_{ij} = \mu + \alpha_i + \beta t_{ij} + \varepsilon_{ij}$$

- Fixed effects μ and β, random effects α_i
- The average response is given by μ for females and $\mu + \beta$ for males
- α_i represents how mother i deviates from the overall mean (the mother-effect).

Case Study 2: The Growth Curves

- Our extended model was given by:

$$Y_{ij} = \begin{cases} (\beta_1 + b_{1i}) + (\beta_2 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if short mother} \\ (\beta_3 + b_{1i}) + (\beta_4 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if medium mother} \\ (\beta_5 + b_{1i}) + (\beta_6 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if tall mother} \end{cases}$$

- Fixed effects $\beta_1, \ldots, \beta_6$, random effects b_{1i} and b_{2i}.

- $\beta_1, \beta_3, \beta_5$ represent the average intercepts

- $\beta_2, \beta_4, \beta_6$ represent the average slopes

- b_{1i} expresses how much the intercept of child i deviates from the average intercept in the group to which this child belongs

- b_{2i} expresses how much the slope of child i deviates from the average slope in the group to which this child belongs

Case Study 3: The Rat Data

- The data
- A linear mixed model
- Fitting the model in SAS

The Rat Data

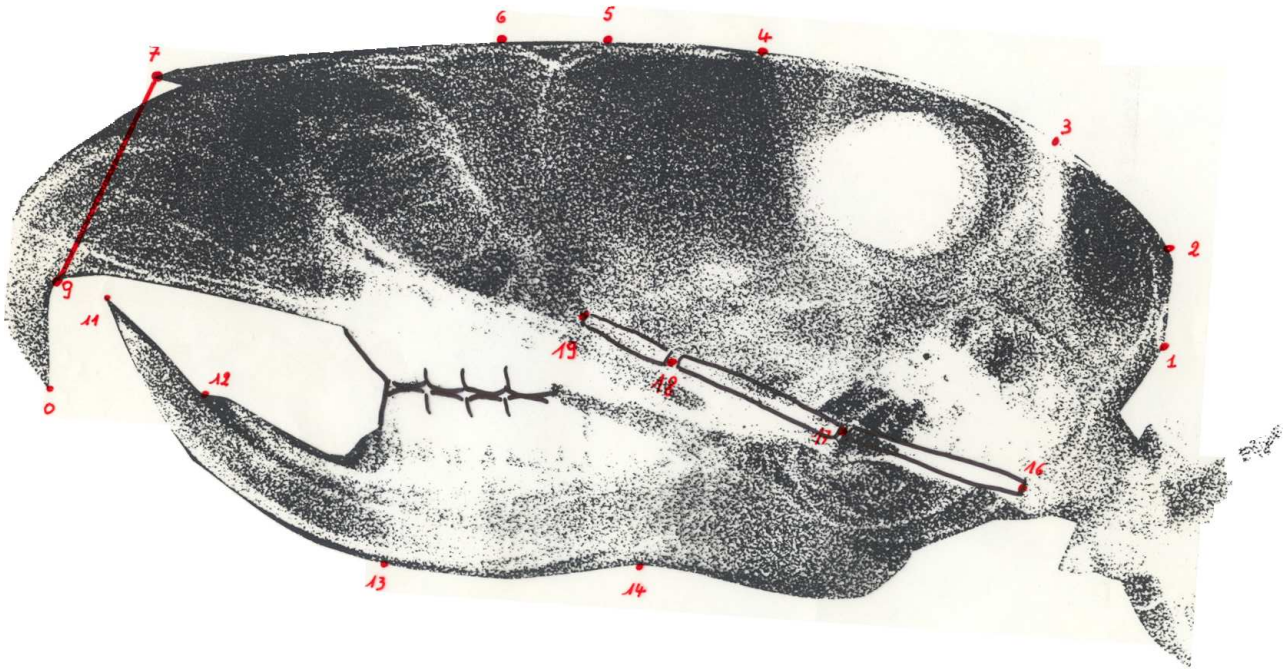
- Research question (Dentistry, K.U.Leuven):

> How does craniofacial growth depend on testosteron production ?

- Randomized experiment in which 50 male Wistar rats are randomized to:
  - ▷ Control (15 rats)
  - ▷ Low dose of Decapeptyl (18 rats)
  - ▷ High dose of Decapeptyl (17 rats)

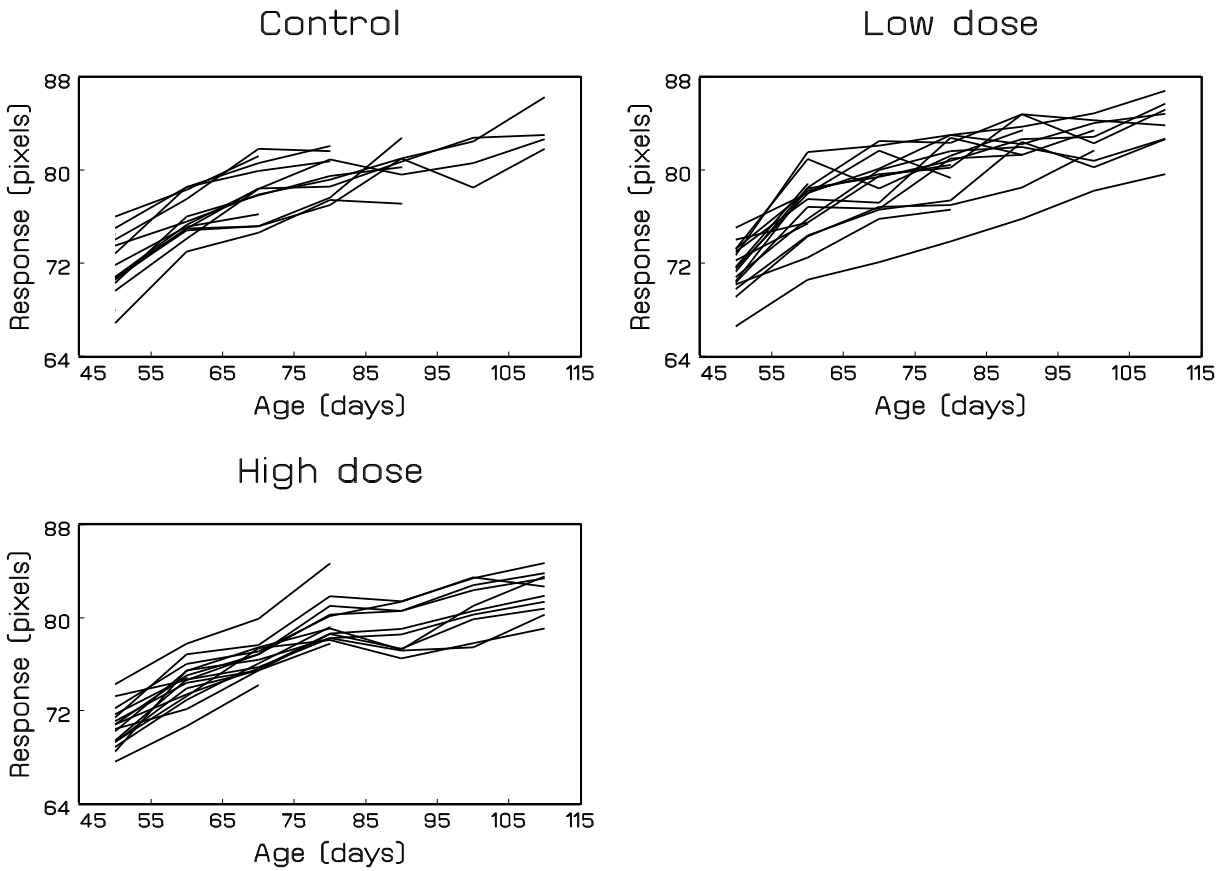
- Treatment starts at the age of 45 days; measurements taken every 10 days, from day 50 on.

- The responses are distances (pixels) between well defined points on x-ray pictures of the skull of each rat:



- Measurements with respect to the roof, base and height of the skull. Here, we consider only one response, reflecting the height of the skull.

- Individual profiles:



- Complication: Dropout due to anaesthesia (56%):

	# Observations			
Age (days)	Control	Low	High	Total
50	15	18	17	50
60	13	17	16	46
70	13	15	15	43
80	10	15	13	38
90	7	12	10	29
100	4	10	10	24
110	4	8	10	22

- Remarks:
  - ▷ Much variability between rats, much less variability within rats
  - ▷ Fixed number of measurements scheduled per subject, but not all measurements available due to dropout, for known reason.
  - ▷ Measurements taken at fixed time points

A Linear Mixed Model

- Since linear mixed models assume a linear regression for each cluster separately, they can also be used for unbalanced data, i.e., data with unequal number of measurements per cluster.

- Note that this was also the case for the lizard data.

- Individual profiles show very similar evolutions for all rats (apart from measurement error)

- This suggests a random-intercepts model

- Non-linearity can be accounted for by using a logarithmic transformation of the time scale:

$$\text{Age}_{ij} \longrightarrow t_{ij} = \ln[1 + (\text{Age}_{ij} - 45)/10)]$$

- We then get the following model:

$$Y_{ij} = (\beta_0 + b_i) + (\beta_1 L_i + \beta_2 H_i + \beta_3 C_i) t_{ij} + \varepsilon_{ij}$$

$$= \begin{cases} \beta_0 + b_i + \beta_1 t_{ij} + \varepsilon_{ij}, & \text{if low dose} \\ \beta_0 + b_i + \beta_2 t_{ij} + \varepsilon_{ij}, & \text{if high dose} \\ \beta_0 + b_i + \beta_3 t_{ij} + \varepsilon_{ij}, & \text{if control.} \end{cases}$$

- L_i, H_i, and C_i are indicator variables:

$$L_i = \begin{cases} 1 & \text{if low dose} \\ 0 & \text{otherwise} \end{cases} \qquad H_i = \begin{cases} 1 & \text{if high dose} \\ 0 & \text{otherwise} \end{cases} \qquad C_i = \begin{cases} 1 & \text{if control} \\ 0 & \text{otherwise} \end{cases}$$

- Parameter interpretation:
  - ▷ β_0: average response at the start of the treatment (independent of treatment)
  - ▷ β_1, β_2, and β_3: average time effect for each treatment group
  - ▷ b_i: subject-specific intercepts

Fitting the Model in SAS

- The following SAS program can be used:

```
data rats;
set rats;
t=log(1+(age-45)/10);
run;
```

```
proc mixed data = rats ;
class treat rat;
model y = treat*t / solution;
random intercept / type=un subject=rat g;
contrast 'treatment effect' treat*t 1 -1 0, treat*t 1 0 -1;
run;
```

- Note the parameterization of the fixed effects

- Relevant SAS output:

```
 Covariance Parameter Estimates

Cov Parm     Subject    Estimate

UN(1,1)      RAT          3.5649
Residual                  1.4448
```

Solution for Fixed Effects

Effect	TREAT	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		68.6074	0.3312	49	207.13	<.0001
t*TREAT	con	7.3138	0.2808	199	26.05	<.0001
t*TREAT	hig	6.8711	0.2276	199	30.19	<.0001
t*TREAT	low	7.5069	0.2252	199	33.34	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
t*TREAT	3	199	734.11	<.0001

Contrasts

Label	Num DF	Den DF	F Value	Pr > F
treatment effect	2	199	2.32	0.1013

- Note the difference between the test for 't*treat' and the test for the treatment effect

- A lot of variability between rats, while little variability within rats:
  - ▷ $\sigma^2_{rat} = 3.565$ represents the variability between rats
  - ▷ $\sigma^2_{res} = 1.445$ represents the variability within rats

- No significant difference between the treatment groups with respect to the average evolution over time ($p = 0.1013$)

Case Study 4: The BLSA Prostate Data

- The data
- A linear mixed model
- Fitting the model in SAS

The Prostate Data

- References:
  - ▷ Carter *et al* (1992, Cancer Research).
  - ▷ Carter *et al* (1992, Journal of the American Medical Association).
  - ▷ Morrell *et al* (1995, Journal of the American Statistical Association).
  - ▷ Pearson *et al* (1994, Statistics in Medicine).

- Prostate disease is one of the most common and most costly medical problems in the United States

- Important to look for markers which can detect the disease at an early stage

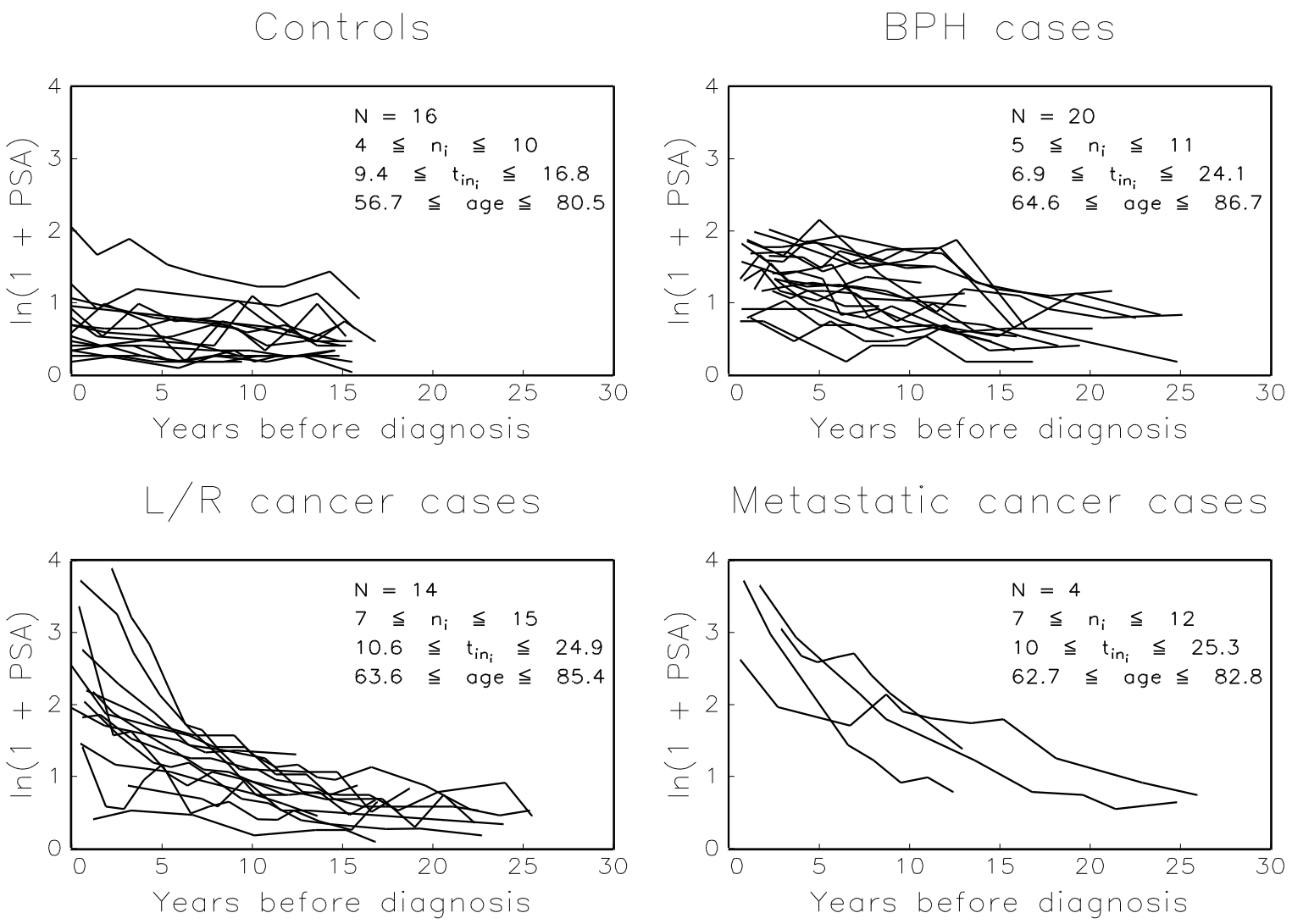
- **P**rostate-**S**pecific **A**ntigen is an enzyme produced by both normal and cancerous prostate cells

- PSA level is related to the volume of prostate tissue.
- Problem: Patients with **B**enign **P**rostatic **H**yperplasia also have an increased PSA level
- Overlap in PSA distribution for cancer and BPH cases seriously complicates the detection of prostate cancer.
- Research question (hypothesis based on clinical practice):

> Can longitudinal PSA profiles be used to detect prostate cancer in an early stage ?

- A retrospective case-control study based on frozen serum samples:
  - ▷ 16 control patients
  - ▷ 20 BPH cases
  - ▷ 14 local cancer cases
  - ▷ 4 metastatic cancer cases
- Complication: No perfect match for age at diagnosis and years of follow-up possible
- Hence, analyses will have to correct for these age differences between the diagnostic groups.

- Individual profiles:



- Remarks:
  - ▷ Much variability between subjects
  - ▷ Little variability within subjects
  - ▷ Highly unbalanced data

A Linear Mixed Model

- A model for the prostate data:

$$\begin{aligned}\ln(\mathsf{PSA}_{ij}+1) \;=\; & \beta_1\mathsf{Age}_i+\beta_2 C_i+\beta_3 B_i+\beta_4 L_i+\beta_5 M_i\\ & +(\beta_6\mathsf{Age}_i+\beta_7 C_i+\beta_8 B_i+\beta_9 L_i+\beta_{10} M_i)\,t_{ij}\\ & +(\beta_{11}\mathsf{Age}_i+\beta_{12} C_i+\beta_{13} B_i+\beta_{14} L_i+\beta_{15} M_i)\,t_{ij}^2\\ & +b_{1i}\;+\;b_{2i}t_{ij}\;+\;b_{3i}t_{ij}^2\;+\;\varepsilon_{ij}.\end{aligned}$$

- C_i, B_i, L_i, M_i are indicators for the 4 diagnostic groups.

- Parameter interpretation:
  - ▷ Average age-corrected quadratic profiles for all groups, modeled through the fixed effects in $\boldsymbol{\beta}$
  - ▷ Random effects b_{1i}, b_{2i}, and b_{3i} allowing subject-specific evolutions to differ from the average in that diagnostic group, even correcting for age differences

Fitting the Model in SAS

- SAS program:

```
proc mixed data=prostate;
class id group;
model lnpsa = group age group*time age*time group*time2 age*time2 / noint solution;
random intercept time time2 / type=un subject=id g gcorr ;
run;
```

- Note again the particular parameterization for the fixed effects

- Relevant SAS output:

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
UN(1,1)	XRAY	0.4518
UN(2,1)	XRAY	-0.5178
UN(2,2)	XRAY	0.9153
UN(3,1)	XRAY	0.1625
UN(3,2)	XRAY	-0.3356
UN(3,3)	XRAY	0.1308
Residual		0.02820

Estimated G Matrix

Effect	XRAY	Col1	Col2	Col3
Intercept	19	0.4518	-0.5178	0.1625
time	19	-0.5178	0.9153	-0.3356
time2	19	0.1625	-0.3356	0.1308

Estimated G Correlation Matrix

Effect	XRAY	Col1	Col2	Col3
Intercept	19	1.0000	-0.8053	0.6686
time	19	-0.8053	1.0000	-0.9700
time2	19	0.6686	-0.9700	1.0000

Solution for Fixed Effects

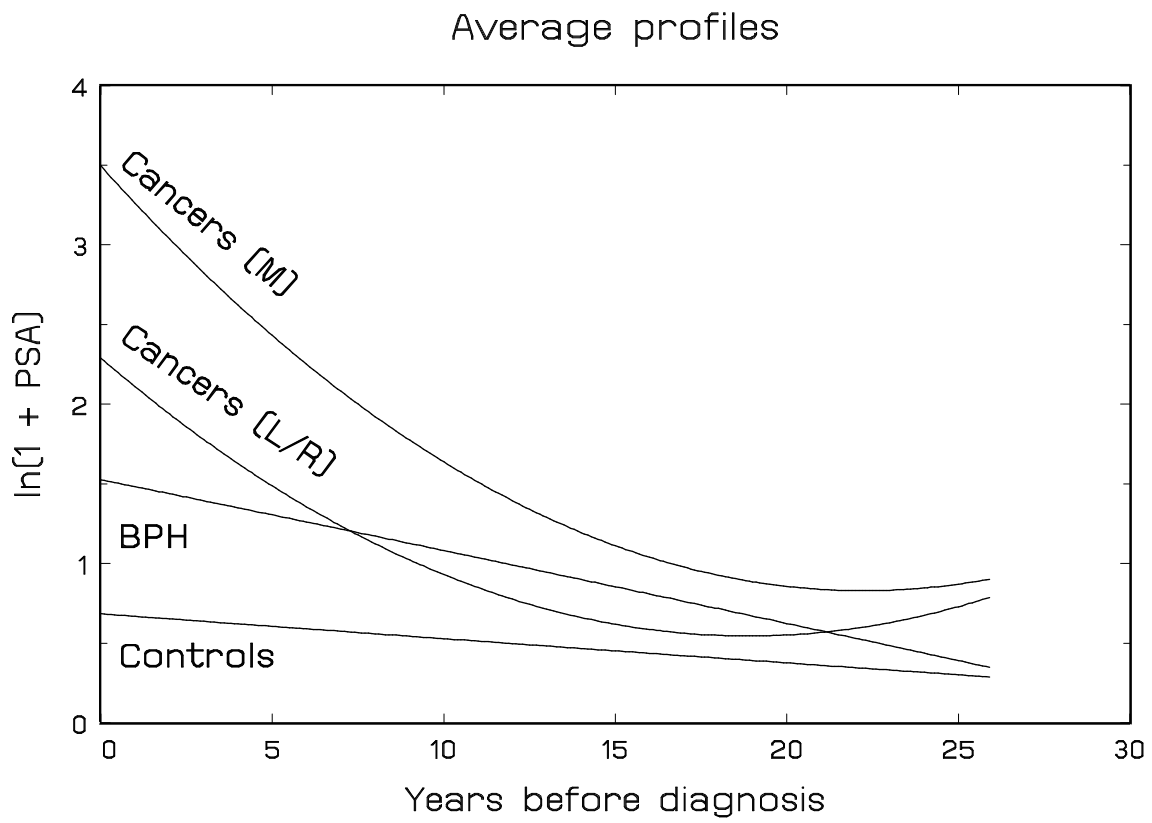
Effect	group	Estimate	Standard Error	DF	t Value	Pr > |t|
group	1	-1.0984	0.9763	299	-1.13	0.2615
group	2	-0.5228	1.0895	299	-0.48	0.6317
group	3	0.2964	1.0587	299	0.28	0.7797
group	4	1.5494	1.0856	299	1.43	0.1546
AGEDIAG		0.02655	0.01423	299	1.87	0.0631
time*group	1	0.5681	1.4725	299	0.39	0.6999
time*group	2	0.3956	1.6377	299	0.24	0.8093
time*group	3	-1.0359	1.5928	299	-0.65	0.5159
time*group	4	-1.6049	1.6258	299	-0.99	0.3244
AGEDIAG*time		-0.01117	0.02142	299	-0.52	0.6026
time2*group	1	-0.1295	0.6100	299	-0.21	0.8320
time2*group	2	-0.1585	0.6723	299	-0.24	0.8138
time2*group	3	0.3419	0.6563	299	0.52	0.6028
time2*group	4	0.3951	0.6660	299	0.59	0.5535
AGEDIAG*time2		0.002259	0.008829	299	0.26	0.7982

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
group	4	299	15.90	<.0001
AGEDIAG	1	299	3.48	0.0631
time*group	4	299	7.85	<.0001
AGEDIAG*time	1	299	0.27	0.6026
time2*group	4	299	4.44	0.0017
AGEDIAG*time2	1	299	0.07	0.7982

- Note the very strong correlations between random effects
- CONTRAST statements can be used to test for group differences

- Based on the fixed effects, fitted average profiles can be plotted (at median age at diagnosis):



Estimation of Random Effects

- ▷ Empirical Bayes estimation
- ▷ Example: Growth curves
- ▷ Example: Prostate data
- ▷ Average versus cluster-specific prediction

Empirical Bayes Estimation

- Random effects $\boldsymbol{b_i}$ reflect how the ith cluster deviates from the population average
- For example, for the growth curves:

$$Y_{ij} = \begin{cases} (\beta_1 + b_{1i}) + (\beta_2 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if short mother} \\ (\beta_3 + b_{1i}) + (\beta_4 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if medium mother} \\ (\beta_5 + b_{1i}) + (\beta_6 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if tall mother} \end{cases}$$

- b_{1i} and b_{2i} express how much the intercept and slope of child i deviate from the average intercept and slope in the group to which this child belongs
- Estimation of the $\boldsymbol{b_i}$ helpful for detecting outlying profiles or clusters
- Since the parameters $\boldsymbol{b_i}$ are assumed to be stochastic, Bayesian methods are applied.

- Posterior means:

$$\widehat{\boldsymbol{b}_i} = E(\boldsymbol{b}_i \mid \boldsymbol{Y}_i = \boldsymbol{y}_i)$$

- The so-obtained estimates $\widehat{\boldsymbol{b}_i}$ are called Empirical Bayes (EB) estimates.

- In practice histograms and/or scatterplots of EB estimates are used to detect outlying clusters

Case Study 2: Growth Curves

- We re-consider the extended model:

$$Y_{ij} = \begin{cases} (\beta_1 + b_{1i}) + (\beta_2 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if short mother} \\ (\beta_3 + b_{1i}) + (\beta_4 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if medium mother} \\ (\beta_5 + b_{1i}) + (\beta_6 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if tall mother} \end{cases}$$

- SAS program for calculation of EB estimates:

```
proc mixed data=growth;
class child group;
model height=age group age*group;
random intercept age / type=un subject=child solution;
ods listing exclude solutionr;
ods output solutionr=out;
run;
```

- The ODS statements are used to write the EB estimates into a SAS output data set, and to prevent SAS from printing them in the output window.

- SAS data management steps:

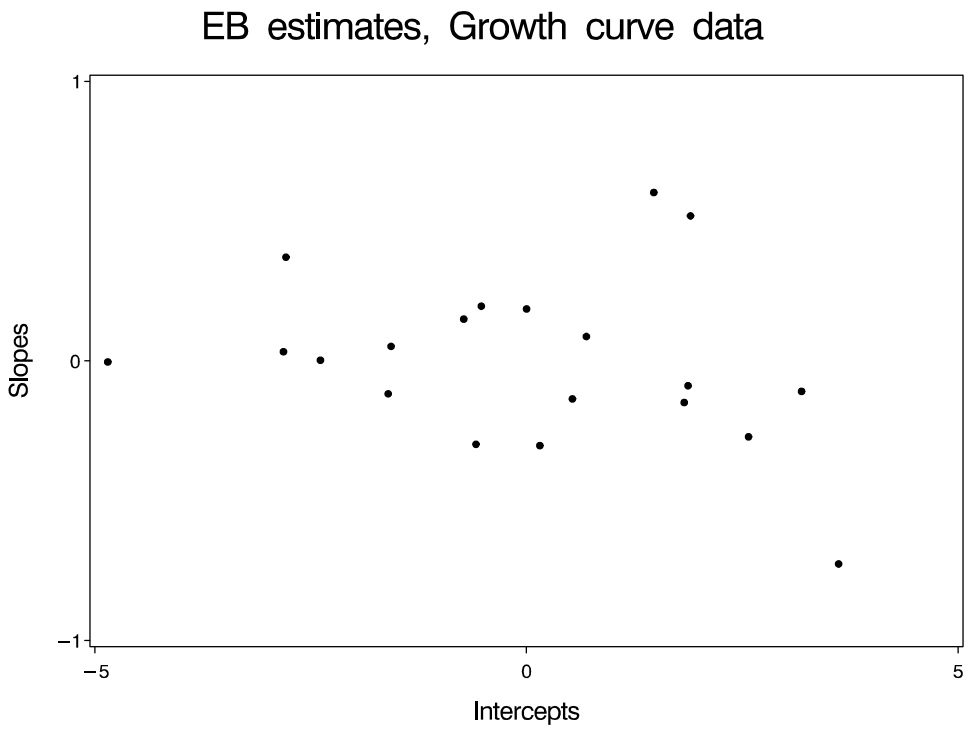
```
data int;set out;if effect='Intercept';
         int=estimate;keep int child;
data slope;set out;if effect='AGE';
         slope=estimate;keep slope child;
data eb;merge int slope; by child;
proc print;run;
```

- SAS program for scatterplot of slopes versus intercepts:

```
proc gplot data=eb;
plot slope*int  / haxis=axis1 vaxis=axis2;
symbol c=black value=dot w=2 i=none;
axis1 label=(h=2  'Intercepts') value=(h=1.5) order=(-5 to 5 by 5) minor=none;
axis2 label=(h=2 A=90 'Slopes') value=(h=1.5) order=(-1 to 1 by 1) minor=none;
title h=3 'EB estimates, Growth curve data';
run;quit;
```

- EB estimates (and scatterplot):

Obs	CHILD	int	slope
1	1	0.15504	-0.30272
2	2	-2.38577	0.00284
3	3	0.00204	0.18604
4	4	1.82784	-0.14857
5	5	3.18613	-0.10875
6	6	-2.78528	0.37117
7	7	0.53293	-0.13570
8	8	1.87116	-0.08839
9	9	1.90073	0.51893
10	10	-0.58431	-0.29783
11	11	0.69442	0.08735
12	12	-1.60125	-0.11740
13	13	-2.81368	0.03303
14	14	3.61566	-0.72571
15	15	-0.52214	0.19608
16	16	-1.56764	0.05223
17	17	-0.72546	0.14969
18	18	2.57185	-0.27115
19	19	-4.84795	-0.00349
20	20	1.47568	0.60234

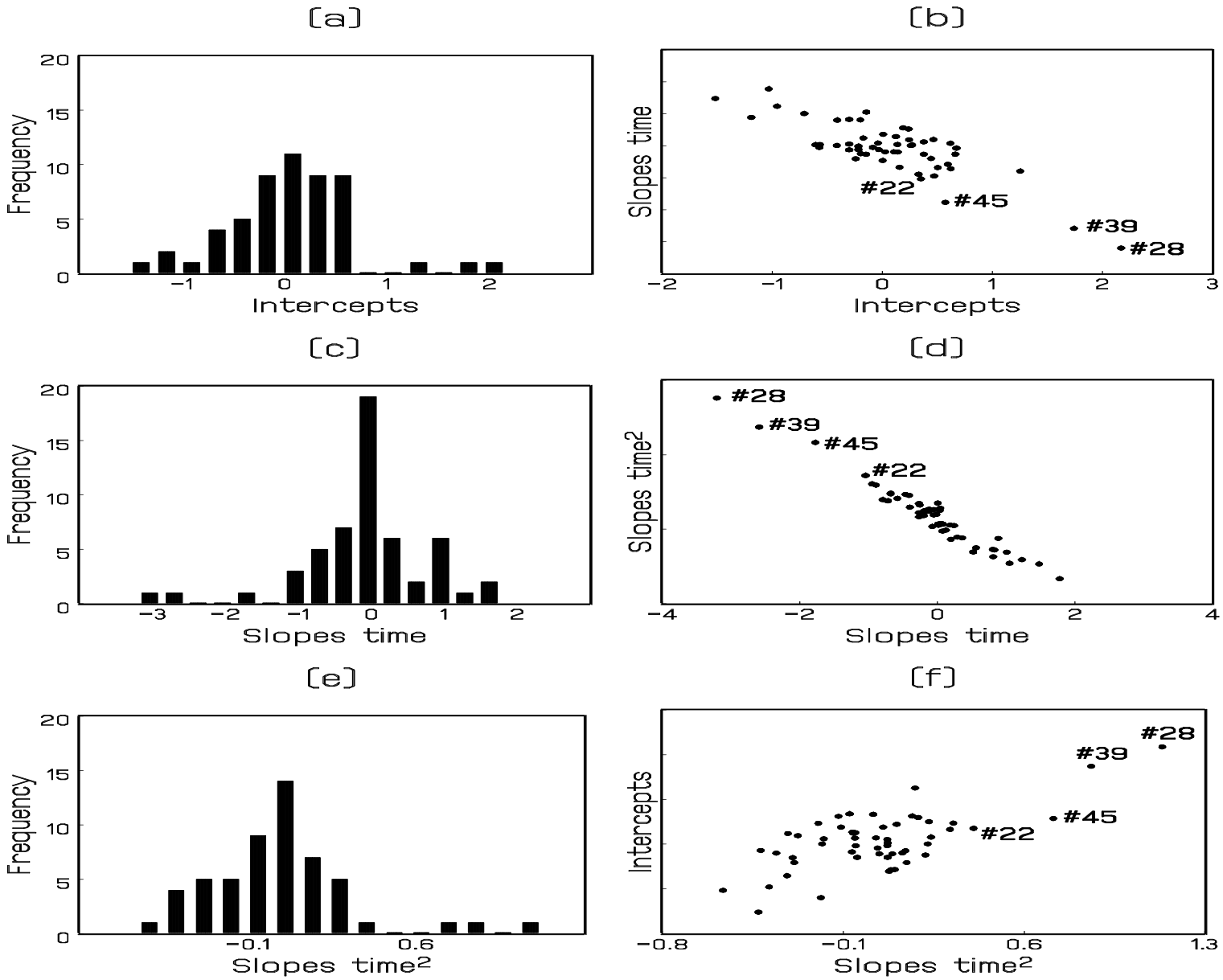


Case Study 4: Prostate Data

- We re-consider the model

$$\begin{aligned}
&\ln(\mathsf{PSA}_{ij}+1)\\
&= \beta_1\mathsf{Age}_i + \beta_2 C_i + \beta_3 B_i + \beta_4 L_i + \beta_5 M_i\\
&\quad + (\beta_6\mathsf{Age}_i + \beta_7 C_i + \beta_8 B_i + \beta_9 L_i + \beta_{10} M_i)\, t_{ij}\\
&\quad + (\beta_{11}\mathsf{Age}_i + \beta_{12} C_i + \beta_{13} B_i + \beta_{14} L_i + \beta_{15} M_i)\, t_{ij}^2\\
&\quad + b_{1i} + b_{2i} t_{ij} + b_{3i} t_{ij}^2 + \varepsilon_{ij}.
\end{aligned}$$

- Again, histograms and scatterplots of components of $\widehat{\boldsymbol{b_i}}$ can be used to detect model deviations or subjects with 'exceptional' evolutions over time


(a)
Frequency
Intercepts
(b)
Slopes time
Intercepts
#22
#45
#39
#28
(c)
Frequency
Slopes time
(d)
Slopes time^2
Slopes time
#28
#39
#45
#22
(e)
Frequency
Slopes time^2
(f)
Intercepts
Slopes time^2
#28
#39
#45
#22

- Strong negative correlations in agreement with correlation matrix corresponding to fitted D:

$$\widehat{D}_{\text{corr}} = \begin{pmatrix} 1.000 & -0.805 & 0.669 \\ -0.805 & 1.000 & -0.970 \\ 0.669 & -0.970 & 1.000 \end{pmatrix}$$

- Histograms and scatterplots show outliers

- Subjects #22, #28, #39, and #45, have highest four slopes for time^2 and smallest four slopes for time, i.e., with the strongest (quadratic) growth.

- Subjects #22, #28 and #39 have been further examined and have been shown to be metastatic cancer cases which were misclassified as local cancer cases.

- Subject #45 is the metastatic cancer case with the strongest growth

Average versus Cluster-specific Prediction

- Once the EB estimates have been calculated, predictions can be obtained both at the cluster level, as well as on the population average level.
- Re-consider the general linear mixed model:

$$\boldsymbol{Y_i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i}$$

- Predictions:
  - ▷ At population average level:

$$E(\boldsymbol{Y_i}) = X_i\widehat{\boldsymbol{\beta}}$$

  - ▷ At cluster level:

$$\widehat{\boldsymbol{Y_i}} = X_i\widehat{\boldsymbol{\beta}} + Z_i\widehat{\boldsymbol{b_i}}$$

Case Study 2: Growth Curves

- We re-consider the extended model:

$$Y_{ij} = \begin{cases} (\beta_1 + b_{1i}) + (\beta_2 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if short mother} \\ (\beta_3 + b_{1i}) + (\beta_4 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if medium mother} \\ (\beta_5 + b_{1i}) + (\beta_6 + b_{2i})t_{ij} + \varepsilon_{ij}, & \text{if tall mother} \end{cases}$$

- SAS program for predictions:

```
proc mixed data=growth;
class child group;
model height= group age*group / noint solution
      outpm=predmean outp=pred;
random intercept age / type=un subject=child solution;
id child age height;
run;
```

```
proc print data=predmean;
proc print data=pred;
run;
```

- Table of predicted means:

Obs	CHILD	AGE	HEIGHT	Pred	StdErr Pred
1	1	6	111.0	112.920	1.10709
2	1	7	116.4	118.190	1.15855
3	1	8	121.7	123.460	1.23249
4	1	9	126.3	128.730	1.32516
5	1	10	130.5	134.000	1.43293
6	2	6	110.0	112.920	1.10709
98	20	8	141.3	133.111	1.14107
99	20	9	146.8	139.360	1.22686
100	20	10	152.3	145.609	1.32664

- Table with predictions on subject level:

Obs	CHILD	AGE	HEIGHT	Pred	StdErr Pred
1	1	6	111.0	111.259	0.47163
2	1	7	116.4	116.226	0.35258
3	1	8	121.7	121.193	0.30673
4	1	9	126.3	126.161	0.36296
5	1	10	130.5	131.128	0.48713
6	2	6	110.0	110.551	0.47163
98	20	8	141.3	139.406	0.30668
99	20	9	146.8	146.257	0.36253
100	20	10	152.3	153.108	0.48571

- Components needed to calculate predictions:

Solution for Fixed Effects

Effect	GROUP	Estimate	Standard Error	DF	t Value	Pr > |t|
GROUP	1	81.3000	1.3381	60	60.76	<.0001
GROUP	2	82.9743	1.2388	60	66.98	<.0001
GROUP	3	83.1229	1.2388	60	67.10	<.0001
AGE*GROUP	1	5.2700	0.1735	60	30.37	<.0001
AGE*GROUP	2	5.5671	0.1606	60	34.66	<.0001
AGE*GROUP	3	6.2486	0.1606	60	38.90	<.0001

Solution for Random Effects

Effect	CHILD	Estimate	Std Err Pred	DF	t Value	Pr > |t|
Intercept	1	0.1550	1.7264	60	0.09	0.9287
AGE	1	-0.3027	0.2195	60	-1.38	0.1730

- Population average prediction for Child #1 (in group 1), at the age of 6 years old:

$$E(Y_{11}) \;=\; 81.30 + 6 \times 5.27 \;=\; 112.92$$

- Subject-specific prediction for Child #1 (in group 1), at the age of 6 years old:

$$\widehat{Y_{11}} \;=\; (81.30 + 0.1550) + 6 \times (5.27 - 0.3027) \;=\; 111.259$$

Case Study 5:
The Belgian Health Interview Survey

- **Conducted in years:** 1997 – 2001 – 2004 — 2008

- **Commissioned by:**
  - ▷ Federal government
  - ▷ Flemish Community
  - ▷ French Community
  - ▷ German Community
  - ▷ Walloon Region
  - ▷ Brussels Region

- **Executing partners:**
  - ▷ Scientific Institute Public Health–Louis Pasteur
  - ▷ National Institute of Statistics
  - ▷ Hasselt University (formerly known as Limburgs Universitair Centrum)
  - ▷ Website: `http://www.iph.fgov.be/epidemio/epien/index4.htm`

- **Goals:**
  - ▷ Subjective health, from the respondent's perspective
  - ▷ Identification of health problems
  - ▷ Information that cannot be obtained from care givers, such as
    - ∗ Estimation of prevalence and distribution of health indicators
    - ∗ Analysis of social inequality in health and access to health care
    - ∗ Study of possible trends in the health status of the population

Design At-a-Glance

- **Regional stratification:** *fixed a priori*
- **Provincial stratification:** *for convenience*
- **Three-stage sampling:**
  - ▷ Primary sampling units (PSU): Municipalities: *proportional to size*
  - ▷ Secondary sampling units (SSU): Households
  - ▷ Tertiary sampling units (TSU): Individuals
- Over-representation of German Community
- Over-representation of 4 (**2**) provinces in 2001 (**2004**):

Limburg	Hainaut
Antwerpen	**Luxembourg**

- Sampling done in 4 quarters: Q1, Q2, Q3, Q4

Regional Stratification

	1997		2001		2004	
Region	Goal	Obt'd	Goal	Obt'd	Goal	Obt'd
Flanders	3500	3536	3500+550=4050	4100	3500+450	
+ elderly					+450=4400	4513
Wallonia	3500	3634	3500+1500=5000	4711	3500+900	
+ elderly					+450=4850	4992
Brussels	3000	3051	3000	3006	3000	
+ elderly					+350=3350	3440
Belgium	10,000	10,221	10,000+2050=12,050	12,111	10,000+1350	
+ elderly					+1250=12,600	12,945

Provincial Stratification in 1997

Province	sample #	sample %	pop. %
Antwerpen	945	26.7	27.7
Oost-Vlaanderen	812	23.0	23.0
West-Vlaanderen	733	20.7	19.1
Vlaams-Brabant	593	16.8	17.0
Limburg	453	12.8	13.2
Hainaut	1325	36.5	38.7
Liège	1210	33.3	30.6
Namur	465	12.8	13.2
Brabant-Wallon	356	9.8	10.3
Luxembourg	278	7.6	7.3
Brussels	3051		

Analysis of Belgian Health Interview Survey

Body Mass Index (BMI):

- ▷ Defined as:

$$\text{BMI} = \frac{\text{weight } (kg)}{\text{height }^2 (m^2)} \quad \left[\frac{\text{kg}}{\text{m}^2}\right]$$

- ▷ A continuous measure
- ▷ Frequently analyzed on the log scale: $\ln(\text{BMI})$

General Health Questionnaire–12 (GHQ-12):

- ▷ Comprises 12 questions, yielding a 13 category outcome
- ▷ The focus is on mental health
- ▷ Can be dichotomized as well

"Vragenlijst voor Onderzoek naar de Ervaren Gezondheid" (VOEG):

- ▷ Dutch instrument, leading to a sum score
- ▷ "Questionnaire for Research Regarding Subjective Health Score"
- ▷ translated into French for Belgium
- ▷ to obtain a more symmetric score, the analysis takes place on the log scale: $\ln(\text{VOEG} + 1)$

Stable General Practioner (SGP):

- ▷ "Do you have a steady general practitioner?" (GP)
- ▷ Obviously a binary indicator

Conventional Design-based Survey Analysis

Logarithm of Body Mass Index				
Analysis	**Belgium**	**Brussels**	**Flanders**	**Wallonia**
Simple Rand. Sampl.	3.187218(0.001845)	3.175877(0.003372)	3.182477(0.002993)	3.201530(0.003216)
Stratification	3.187218(0.001840)	3.175877(0.003373)	3.182477(0.002989)	3.201530(0.003217)
Clustering	3.187218(0.001999)	3.175877(0.003630)	3.182477(0.003309)	3.201530(0.003429)
Weighting	3.185356(0.002651)	3.171174(0.004578)	3.180865(0.003870)	3.198131(0.004238)
All combined	3.185356(0.003994)	3.171174(0.004844)	3.180865(0.004250)	3.198131(0.004403)

Logarithm of VOEG Score				
Analysis	**Belgium**	**Brussels**	**Flanders**	**Wallonia**
SRS	1.702951(0.008954)	1.809748(0.016203)	1.516352(0.015201)	1.801107(0.014550)
Stratification	1.702951(0.008801)	1.809748(0.016206)	1.516352(0.015207)	1.801107(0.014427)
Clustering	1.702951(0.010355)	1.809748(0.018073)	1.516352(0.017246)	1.801107(0.016963)
Weighting	1.634690(0.013233)	1.802773(0.021831)	1.511927(0.019155)	1.803178(0.020426)
All combined	1.634690(0.014855)	1.802773(0.023135)	1.511927(0.021409)	1.803178(0.023214)

General Health Questionnaire – 12				
Analysis	**Belgium**	**Brussels**	**Flanders**	**Wallonia**
SRS	1.661349(0.029584)	1.862745(0.056894)	1.385381(0.046246)	1.772148(0.051023)
Stratification	1.661956(0.029452)	1.864301(0.056939)	1.385857(0.046211)	1.772148(0.050823)
Clustering	1.661349(0.032824)	1.862745(0.062739)	1.385381(0.052202)	1.772148(0.055780)
Weighting	1.626201(0.044556)	1.924647(0.076313)	1.445957(0.061910)	1.858503(0.078566)
All combined	1.626781(0.048875)	1.924647(0.080508)	1.446286(0.068931)	1.858503(0.084047)

Stable General Practitioner (0/1)				
Analysis	**Belgium**	**Brussels**	**Flanders**	**Wallonia**
SRS	0.903540(0.003196)	0.805632(0.007826)	0.952285(0.003908)	0.938646(0.004382)
Stratification	0.903540(0.003116)	0.805632(0.007827)	0.952285(0.003902)	0.938646(0.004366)
Clustering	0.903540(0.003963)	0.805632(0.009766)	0.952285(0.004709)	0.938646(0.005284)
Weighting	0.932702(0.003498)	0.782448(0.011563)	0.954757(0.004722)	0.943191(0.005417)
All combined	0.932702(0.003994)	0.782448(0.013836)	0.954757(0.005379)	0.943191(0.006159)

- Weighting and clustering each increase the standard error,
- the combined analysis does more so.
- The point estimate is identical to the weighted one.

Design Effects

Outcome	Belgium	Brussels	Flanders	Wallonia
Design Effects for Clustering				
LNBMI	1.2	1.2	2.1	1.1
LNVOEG	1.3	1.2	1.3	1.4
GHQ–12	2.3	1.8	1.8	2.4
SGP	1.5	1.6	1.5	1.5
Design Effects for Weighting				
LNBMI	2.1	1.8	2.8	1.7
LNVOEG	2.2	1.8	1.6	2.0
GHQ–12	2.3	1.8	1.8	2.4
SGP	1.2	2.2	1.5	1.5

Modern Model-Based Analysis

$$Y_{ij} = \mu + b_i + \varepsilon_{ij}$$

- Y_{ij} is the observation for subject j in cluster i
- **fixed effect:** μ is the overall, population mean
- **mixed effect:** $\mu + b_i$ is the cluster-specific average
- **random effect:**

$$b_i \sim N(0, \tau^2) \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

- This is an instance of a **linear mixed model**.
- Verbeke and Molenberghs (2000)

One Level Deeper

$$Y_{ijk} = \mu + b_i + c_{ij} + \varepsilon_{ijk}$$

- Y_{ijk} is the observation for subject k in household j in town i
- μ is the overall, population mean
- b_i is the town-level effect: $b_i \sim N(0, \tau^2_{\text{town}})$
- c_{ij} is the household-level effect: $c_{ij} \sim N(0, \tau^2_{\text{HH}})$
- ε_{ij} is the individual-level deviation: $\varepsilon_{ijk} \sim N(0, \tau^2_{\text{ind}})$
- When μ and/or b_i and/or c_{ij} are made functions of covariates, we have a so-called **multilevel approach**.

linear mixed model $\equiv$ **multilevel model**

Logarithm of Body Mass Index					
Analysis	**Procedure**	**Belgium**	**Brussels**	**Flanders**	**Wallonia**
SRS	SURVEYMEANS	3.1872(0.0018)	3.1759(0.0034)	3.1825(0.0030)	3.2015(0.0032)
SRS	MIXED	3.1872(0.0018)	3.1759(0.0034)	3.1825(0.0030)	3.2015(0.0032)
Stratification	SURVEYMEANS	3.1872(0.0018)	3.1759(0.0034)	3.1825(0.0030)	3.2015(0.0032)
Clustering	SURVEYMEANS	3.1872(0.0020)	3.1759(0.0036)	3.1825(0.0033)	3.2015(0.0034)
Clustering	MIXED	3.1880(0.0020)	3.1761(0.0036)	3.1840(0.0033)	3.2022(0.0034)
Weighting	SURVEYMEANS	3.1853(0.0027)	3.1712(0.0046)	3.1809(0.0039)	3.1981(0.0042)
Weighting	MIXED	3.1854(0.0018)	3.1712(0.0034)	3.1809(0.0030)	3.1981(0.0032)
All combined	SURVEYMEANS	3.1853(0.0040)	3.1712(0.0048)	3.1809(0.0043)	3.1981(0.0044)
Clust+Wgt	MIXED	3.1865(0.0023)	3.1706(0.0039)	3.1817(0.0036)	3.1994(0.0038)

Binary Data: Generalized Estimating Equations

- When an outcome is binary, one can calculate a proportion π, which is the probability to belong to a group, to have a certain characteristic, etc.

- Alternatively, the logit can be calculated:

$$\beta = \text{logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right), \qquad \pi = \frac{e^{\beta}}{1+e^{\beta}}$$

  and

$$\text{logit}[P(Y_i = 1)] = \beta$$

- Estimation of β typically proceeds through maximum likelihood estimation, which necessitates numerical optimization, since no closed form exists.

- For SRS, this can be implemented using logistic regression.
- For the clustered case, the correlation can be incorporated into the model:

$$\text{logit}[P(Y_{ij} = 1)] = \beta, \qquad \text{Corr}(Y_{ij}, Y_{ik}) = \alpha$$

- β is the logit of the population proportion.
- α is the within-household correlation.
- Full maximum likelihood estimation is tedious.
- Liang and Zeger (Biometrika 1986) have developed a convenient estimation method: generalized estimating equations (GEE).
- A way to think about it is: correlation-corrected logistic regression.

Stable General Practitioner (0/1) — Marginal Models						
Analysis	**Procedure**	**Par.**	**Belgium**	**Brussels**	**Flanders**	**Wallonia**
SRS	SURVEYMEANS	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	SURVEYLOGISTIC.	$-\beta$	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	SURVEYLOGISTIC.	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	GENMOD	$-\beta$	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	GENMOD	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
Strat.	SURVEYMEANS	π	0.9035(0.0031)	0.8056(0.0078)	0.9522(0.0039)	0.9386(0.0044)
Strat.	SURVEYLOGISTIC	$-\beta$	2.3272(0.0358)	1.4219(0.0050)	2.9936(0.0859)	2.7278(0.0758)
Strat.	SURVEYLOGISTIC	π	0.9035(0.0031)	0.8056(0.0078)	0.9522(0.0039)	0.9386(0.0044)
Clust.	SURVEYMEANS	π	0.9035(0.0040)	0.8056(0.0098)	0.9523(0.0047)	0.9386(0.0053)
Clust.	SURVEYLOGISTIC	$-\beta$	2.2372(0.0455)	1.4219(0.0624)	2.9936(0.1037)	2.7278(0.0918)
Clust.	SURVEYLOGISTIC	π	0.9035(0.0040)	0.8056(0.0098)	0.9523(0.0047)	0.9386(0.0053)
Clust.	GENMOD	$-\beta$	2.1504(0.0435)	1.3784(0.0591)	2.9188(0.1019)	2.6470(0.0890)
Clust.	GENMOD	π	0.8957(0.0040)	0.7987(0.0095)	0.9488(0.0050)	0.9338(0.0055)
Wgt.	SURVEYMEANS	π	0.9327(0.0035)	0.7824(0.0116)	0.9548(0.0047)	0.9432(0.0054)
Wgt.	SURVEYLOGISTIC	$-\beta$	2.6290(0.0557)	1.2800(0.0679)	3.0494(0.1093)	2.8096(0.1011)
Wgt.	SURVEYLOGISTIC	π	0.9327(0.0035)	0.7824(0.0116)	0.9548(0.0047)	0.9432(0.0054)
Wgt.	GENMOD	$-\beta$	2.6290(0.0642)	1.2800(0.0813)	3.0494(0.1245)	2.8096(0.1150)
Wgt.	GENMOD	π	0.9327(0.0040)	0.7824(0.0138)	0.9548(0.0054)	0.9432(0.0062)
All	SURVEYMEANS	π	0.9327(0.0040)	0.7824(0.0138)	0.9548(0.0054)	0.9432(0.0062)
All	SURVEYLOGISTIC	$-\beta$	2.6290(0.0636)	1.2800(0.0813)	3.0494(0.1245)	2.8096(0.1150)
All	SURVEYLOGISTIC	π	0.9327(0.0040)	0.7824(0.0138)	0.9548(0.0054)	0.9432(0.0062)
Cl.+Wt.	GENMOD	$-\beta$	2.5233(0.0659)	1.2014(0.0839)	2.9693(0.1284)	2.7251(0.1186)
Cl.+Wt.	GENMOD	π	0.9258(0.0045)	0.7688(0.0149)	0.9512(0.0060)	0.9385(0.0068)

Binary Data: Generalized Linear Mixed Models

- We already considered two models to account for multi-level structures:

▷ The LMM, through random effects:

▷ $Y_{ij} = \mu + b_i + \varepsilon_{ij}$

▷ $b_i \sim N(0, \tau^2)$

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

▷ GEE, through marginal correlation:

▷ $P(Y_{ij} = 1) = \frac{e^\beta}{1+e^\beta}$

▷ $\mathsf{Corr}(Y_{ij}, Y_{ik}) = \alpha$

- Aspects of both can be combined, to produce the generalized linear mixed model (GLMM):

$$P(Y_{ij} = 1) = \frac{e^{\beta + b_i}}{1 + e^{\beta + b_i}}$$

$$b_i \sim N(0, \tau^2)$$

- There are a few important differences:
  - ▷ Unlike with the LMM and GEE, it is not straightforward to calculate/obtain the intra-cluster correlation.
  - ▷ ML is an obvious candidate for parameter estimation.

▷ But: the likelihood contribution for cluster (household) i is:

$$L_i = \int \prod_{j=1}^{n_i} \frac{y_{ij} \cdot e^{\beta + b_i}}{1 + e^{\beta + b_i}} \cdot \varphi(b_i|\tau^2)\ db_i$$

where $\varphi(b_i|\tau^2)$ is the normal density.

▷ There exists no closed-form solution for this integral.

- The stated problem has led to two main approximation approaches:
  - ▷ Numerical integration: the SAS procedure NLMIXED/GLIMMIX.
    - ∗ Allows for high accuracy.
    - ∗ Time consuming.
    - ∗ A bit harder to program.
  - ▷ Taylor series expansions: implemented in the SAS procedure GLIMMIX.
    - ∗ Bias due to poor approximation.
    - ∗ As easy to use as the MIXED and GENMOD procedures.

Case Study 5:
The Belgian Health Interview Survey

- We will estimate the mean (probability) for SGP:
  - ▷ For Belgium and the regions.
  - ▷ Under SRS and two-stage (cluster) sampling.
  - ▷ Using a Taylor series.
  - ▷ Using numerical integration.

- Results ignoring correlation:

$$\widehat{\beta} = -2.2372(\text{ s.e. } 0.0367) \quad \Rightarrow \quad \widehat{\pi} = 0.9035(\text{ s.e. } 0.0032)$$

- Results taking correlation into account:
  - ▷ We obtain the following probability:

$$\widehat{\beta} = -2.3723(\text{ s.e. } 0.0443) \quad \Rightarrow \quad \widehat{\pi} = 0.9147(\text{ s.e. } 0.0035)$$

  - ▷ The estimate for β is supplemented with an estimate for the random effects variance: $\widehat{\tau}^2 = 1.75$ with s.e. 0.12.
  - ▷ $\widehat{\beta}$ and its standard error is not very different from what was obtained with GEE.

Method	Procedure	Estimate (s.e.) $\widehat{\beta}$	Estimate (s.e.) $\widehat{\pi}$
	Marginal approaches		
logistic	SURVEYMEANS	—	0.9035 (0.0040)
logistic	SURVEYLOGISTIC	2.2372 (0.0455)	0.9035 (0.0040)
GEE	GENMOD	2.1504 (0.0435)	0.8957 (0.0040)
	Random-effects approaches		
GLMM	GLIMMIX	2.3723 (0.0443)	0.9147 (0.0035)
GLMM	NLMIXED	4.3770 (0.1647)	0.9876 (0.0020)

- ▷ This difference is spectacular and requires careful qualification.
- ▷ Note that the 'true' value is the number of people in the dataset with a stable GP divided by the total number of people:

$$\text{pragmatic estimate of } \pi \ = \frac{7709}{7709 + 823} = 0.9035$$

which, of course, is in agreement with all of the SRS analyses.

▷ Further:

* The survey-design based procedures are spot on.
* GEE is a little different, but close.
* GLIMMIX is a little different, but close, with the deviation going the other way.
* NLMIXED is spectacularly different.

▷ The strong differences can be explained as follows:

* Consider our GLMM:

$$Y_{ij}|b_i \sim \text{Bernoulli}(\pi_{ij}), \qquad \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + b_i$$

- The conditional means $E(Y_{ij}|b_i)$, are given by

$$E(Y_{ij}|b_i) = \frac{\exp(\beta_0 + b_i)}{1 + \exp(\beta_0 + b_i)}$$

- The marginal means are now obtained from averaging over the random effects:

$$E(Y_{ij}) = E[E(Y_{ij}|b_i)] = E\left[\frac{\exp(\beta_0 + b_i)}{[1 + \exp(\beta_0 + b_i)]}\right] \neq \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

▷ Hence, the parameter vector $\boldsymbol{\beta}$ in the GEE model needs to be interpreted completely differently from the parameter vector $\boldsymbol{\beta}$ in the GLMM:

- GEE: marginal interpretation
- GLMM: conditional interpretation, conditionally upon level of random effects

▷ In general, the model for the marginal average is not of the same parametric form as the conditional average in the GLMM.

▷ For logistic mixed models, with normally distributed random random intercepts, it can be shown that the marginal model can be well approximated by again a logistic model, but with parameters approximately satisfying

$$\frac{\widehat{\boldsymbol{\beta}}^{\mathsf{RE}}}{\widehat{\boldsymbol{\beta}}^{\mathsf{M}}} \;=\; \sqrt{c^2\tau^2 \,+\, 1} \;>\; 1, \qquad\qquad \tau^2 \;=\; \text{variance random intercepts}$$

$$c \;=\; 16\sqrt{3}/(15\pi)$$

▷ For our case:

$$\frac{\widehat{\boldsymbol{\beta}}^{\mathsf{RE}}}{\widehat{\boldsymbol{\beta}}^{\mathsf{M}}} \;=\; \frac{4.3770}{2.1504} \;=\; 2.0354$$

$$\sqrt{c^2\tau^2 \,+\, 1} \;=\; \sqrt{0.5881^2 \times 7.3232 + 1} \;=\; 1.8795$$

▷ The relationship is not exact, but sufficiently close.

- ▷ The interpretation of the random-effects-based β is:

  The logit of having a stable GP for someone with HH-level effecgt $b_i = 0$.

- ▷ The interpretation of the random-effects-based π is:

  The probability of having a stable GP for someone with HH-level effect $b_i = 0$.

- ▷ Thus, the probability corresponding to the average household is different from the probability averaged over all households.

- ▷ All of these relationships would also hold for the GLIMMIX procedure, if it were not so biased!

• We can further expand the summary table for SGP with our new analyses:

Stable General Practitioner (0/1) — Marginal and Random-effects Models						
Analysis	**Procedure**	**Par.**	**Belgium**	**Brussels**	**Flanders**	**Wallonia**
SRS	SURVEYMEANS	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	SURVEYLOGISTIC	$-\beta$	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	SURVEYLOGISTIC	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	GENMOD	$-\beta$	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	GENMOD	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	GLIMMIX	β	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	GLIMMIX	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	NLMIXED	β	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	NLMIXED	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
Strat.	SURVEYMEANS	π	0.9035(0.0031)	0.8056(0.0078)	0.9522(0.0039)	0.9386(0.0044)
Strat.	SURVEYLOGISTIC	$-\beta$	2.3272(0.0358)	1.4219(0.0050)	2.9936(0.0859)	2.7278(0.0758)
Strat.	SURVEYLOGISTIC	π	0.9035(0.0031)	0.8056(0.0078)	0.9522(0.0039)	0.9386(0.0044)
Clust.	SURVEYMEANS	π	0.9035(0.0040)	0.8056(0.0098)	0.9523(0.0047)	0.9386(0.0053)
Clust.	SURVEYLOGISTIC	$-\beta$	2.2372(0.0455)	1.4219(0.0624)	2.9936(0.1037)	2.7278(0.0918)
Clust.	SURVEYLOGISTIC	π	0.9035(0.0040)	0.8056(0.0098)	0.9523(0.0047)	0.9386(0.0053)
Clust.	GENMOD	$-\beta$	2.1504(0.0435)	1.3784(0.0591)	2.9188(0.1019)	2.6470(0.0890)
Clust.	GENMOD	π	0.8957(0.0040)	0.7987(0.0095)	0.9488(0.0050)	0.9338(0.0055)
Clust.	GLIMMIX	β	2.3723(0.0441)	1.5213(0.0628)	3.1433(0.0988)	—
Clust.	GLIMMIX	π	0.9147(0.0034)	0.8207(0.0092)	0.9586(0.0039)	—
Clust.	NLMIXED	β	4.3770(0.1647)	3.4880(0.3134)	8.4384(1.5434)	6.9047(0.8097)
Clust.	NLMIXED	π	0.9876(0.0020)	0.9703(0.0090)	0.9998(0.0003)	0.9990(0.0008)

Stable General Practitioner (0/1) — Marginal and Random-effects Models						
Analysis	**Procedure**	**Par.**	**Belgium**	**Brussels**	**Flanders**	**Wallonia**
Wgt.	SURVEYMEANS	π	0.9327(0.0035)	0.7824(0.0116)	0.9548(0.0047)	0.9432(0.0054)
Wgt.	SURVEYLOGISTIC	$-\beta$	2.6290(0.0557)	1.2800(0.0679)	3.0494(0.1093)	2.8096(0.1011)
Wgt.	SURVEYLOGISTIC	π	0.9327(0.0035)	0.7824(0.0116)	0.9548(0.0047)	0.9432(0.0054)
Wgt.	GENMOD	$-\beta$	2.6290(0.0642)	1.2800(0.0813)	3.0494(0.1245)	2.8096(0.1150)
Wgt.	GENMOD	π	0.9327(0.0040)	0.7824(0.0138)	0.9548(0.0054)	0.9432(0.0062)
Wgt.	GLIMMIX	β	2.6290(0.0557)	1.2800(0.0679)	3.0494(0.1093)	2.8096(0.1011)
Wgt.	GLIMMIX	π	0.9327(0.0035)	0.7824(0.0116)	0.9548(0.0047)	0.9432(0.0054)
All	SURVEYMEANS	π	0.9327(0.0040)	0.7824(0.0138)	0.9548(0.0054)	0.9432(0.0062)
All	SURVEYLOGISTIC	$-\beta$	2.6290(0.0636)	1.2800(0.0813)	3.0494(0.1245)	2.8096(0.1150)
All	SURVEYLOGISTIC	π	0.9327(0.0040)	0.7824(0.0138)	0.9548(0.0054)	0.9432(0.0062)
Cl.+Wgt.	GENMOD	$-\beta$	2.5233(0.0659)	1.2014(0.0839)	2.9693(0.1284)	2.7251(0.1186)
Cl.+Wgt.	GENMOD	π	0.9258(0.0045)	0.7688(0.0149)	0.9512(0.0060)	0.9385(0.0068)
Cl.+Wgt.	GLIMMIX	β	7.8531(0.1105)	5.1737(0.1906)	9.8501(0.1962)	8.7535(0.1850)
Cl.+Wgt.	GLIMMIX	π	0.9996(0.0000)	0.9944(0.0011)	0.9999(0.0000)	0.9998(0.0000)

- In summary, we note the following:
  - ▷ Compared to the marginal approaches, β and π are not generally interpretable as meaningful population quantities.
  - ▷ It is possible to derive the marginal parameters, but this involves extra numerical integration.
  - ▷ Relative to the integration-based estimates, the Taylor-series estimates are biased downwards.
  - ▷ Important uses for the GLMM method:
    - ∗ When estimates are required at more than one level at the same time, e.g., town and/or HH and/or individual.
    - ∗ As a flexible tool for regression, rather than for simple population-level estimates (means, totals).