

# STK4900/9900 - Lecture 10

## Program

1. Repeated measures and longitudinal data
  2. Simple analysis approaches
  3. Random effects models
  4. Generalized estimating equations
- Sections 8.1, 8.2, 8.3, 8.4 (except 8.4.5), 8.5 (except 8.5.2-4)

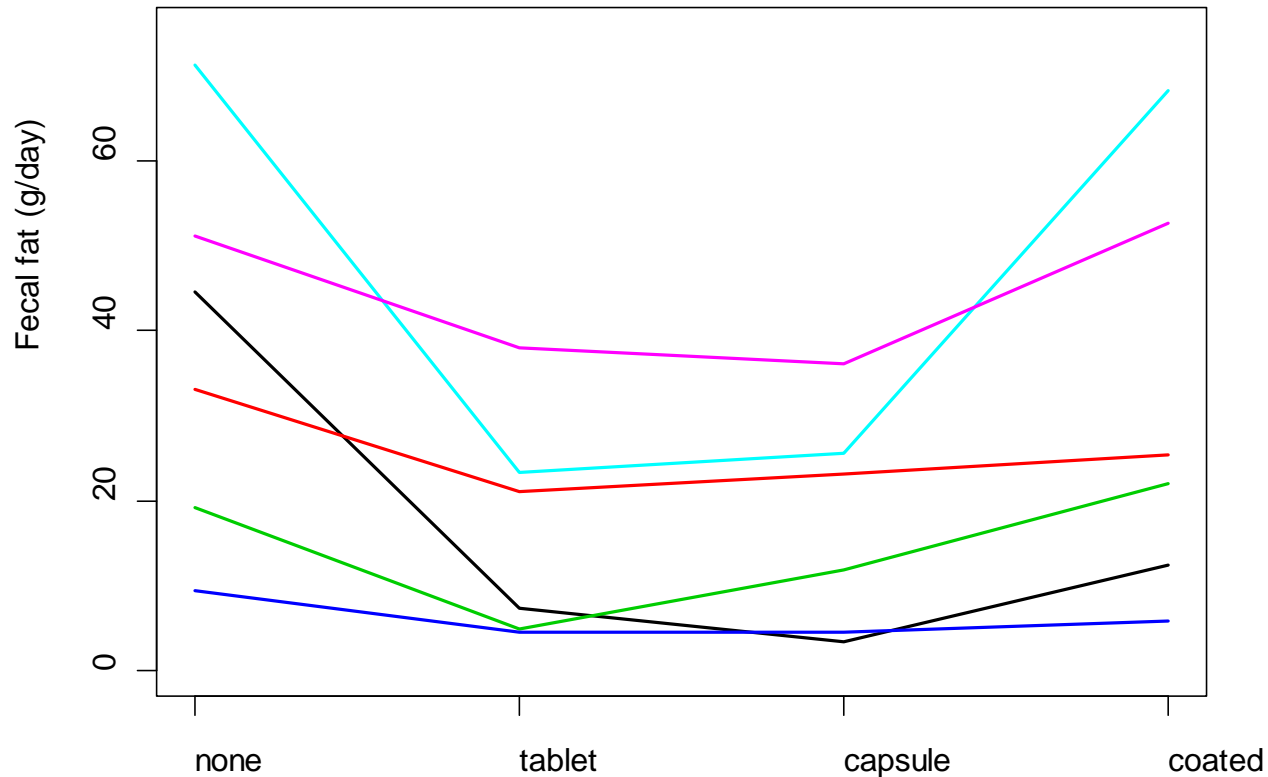
## Example: Fecal fat

Lack of digestive enzymes in the intestine can cause bowel absorption problems, which will be indicated by excess fat in the feces. Pancreatic enzyme supplements can reduce the problem. The data are from a study to determine if the form of the supplement makes a difference

**Table 8.1.** Fecal Fat (g/day) for Six Subjects

Subject number	Pill type				Subject average
	None	Tablet	Capsule	Coated	
1	44.5	7.3	3.4	12.4	16.9
2	33.0	21.0	23.1	25.4	25.6
3	19.1	5.0	11.8	22.0	14.5
4	9.4	4.6	4.6	5.8	6.1
5	71.3	23.3	25.6	68.2	47.1
6	51.2	38.0	36.0	52.6	44.5
Pill type average	38.1	16.5	17.4	31.1	25.8

This is an example with **repeated measurements** (more than one observation per subject)



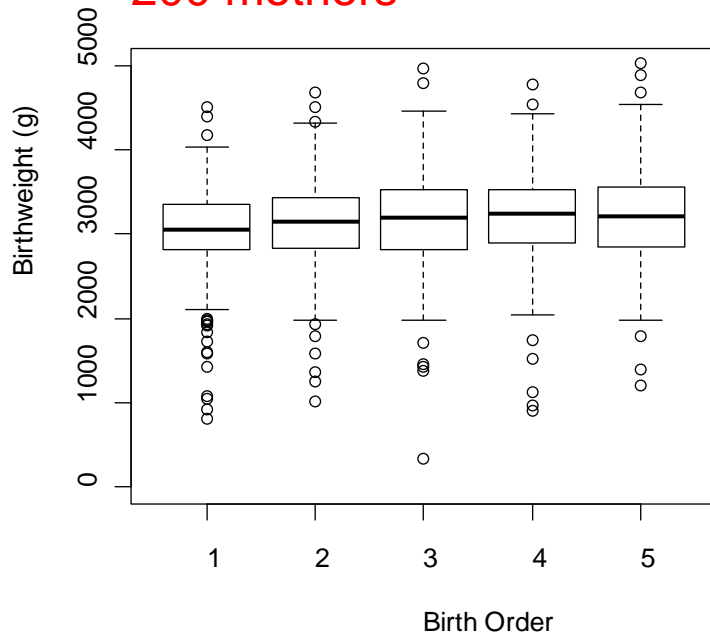
The plot shows that some patients tend to have high values for all pill types, while other patients tend to have low values

The values for a patients are *not* independent, and this has to be taken into account when we analyze the data

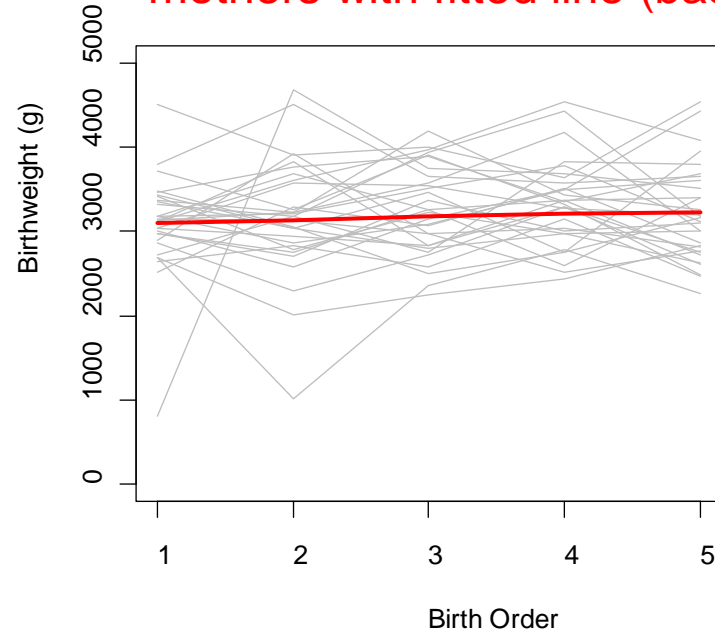
## Example: Birth weight and birth order

We have recorded the weights of the babies of 200 mothers who all have five children. We are interested in studying the effect of birth order and the age of the mother on the birth weight

Box plots of birth weights for all 200 mothers



Birth weights for a sample of 30 mothers with fitted line (based on all)



This is an example of **longitudinal data** (repeated measures taken over time)

The birth weights for a mother are *not* independent, and this has to be taken into account when we analyze the data

# Simple approaches to analyzing repeated measures and longitudinal data

When analyzing repeated measures data or longitudinal data we must be careful *not* to disregard the dependence of the data for an individual

By only looking at parts of the data for an individual one may avoid the dependence problem

For the [fecal fat example](#) one option is to compare two pill types at a time using the paired t-test

For the [birth weight example](#) one option is to look at the difference in weight between the fifth and the first child

## Fecal fat example:

### R commands (comparing pill types 1 and 2):

```
fecfat=read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v11/fecfat.txt",  
                 header=T)  
x=fecfat$fecfat[fecfat$pilltype==1]  
y=fecfat$fecfat[fecfat$pilltype==2]  
t.test(x,y,paired=T)
```

### R output :

Paired t-test

$t = 3.109$ ,  $df = 5$ ,  $p\text{-value} = 0.027$

95 percent confidence interval:

3.731      39.369

mean of the differences: 21.55

Estimated difference / P-value for the six comparisons of two pill types at a time (column minus row)

	None	Tablet	Capsule
Tablet	21.6 / 2.7 %	*	*
Capsule	20.7 / 3.7 %	-0.9 / 58.9 %	*
Coated	7.0 / 23.9 %	-14.5 / 7.8 %	-13.7 / 8.0 %

## Birth weight example:

### R commands:

```
babies=read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v11/gababies.txt",  
                 header=T)  
first=babies$bweight[babies$birthord==1]  
fifth= babies$bweight[babies$birthord==5]  
diff=fifth-first  
t.test(diff)
```

### R output :

One Sample t-test

data: diff

t = 4.211, df = 199, p-value = 3.849e-05

95 percent confidence interval:

101.90 281.38

mean of x

191.64

On average the fifth child weighs 191.6 grams more than the first

A 95 % confidence interval is from 101.9 grams to 281.4 grams

If we divide by four we get the average increase per child

## Random effects model

A useful approach for analysing repeated measures is to consider a random effects model

We will describe the random effects model using the [fecal fat example](#)

Here we consider the model:

$$Y_{ij} = \mu + \beta_j + B_i + \varepsilon_{ij}$$

where

$Y_{ij}$  is the fecal fat for patient  $i$  when using pill type  $j$

$\beta_j$  is the effect of pill type  $j$  (relative to type 1)

the  $B_i$  are the effects of patients, assumed independent  $N(0, \sigma_{subj}^2)$

the  $\varepsilon_{ij}$  are random errors, assumed independent  $N(0, \sigma_{\varepsilon}^2)$

To fit a random effects model, we use the "nlme" library



## R commands:

```
library(nlme)
fit.fecfat=lme(fecfat~factor(pilltype), random=~1|subject, data=fecfat)
summary(fit.fecfat)
anova(fit.fecfat)
```

## R output (edited):

Linear mixed-effects model fit by REML

### Random effects:

Formula: ~1 | subject

	(Intercept)	Residual
StdDev:	15.900	10.344

### Fixed effects: fecfat ~ factor(pilltype)

	Value	Std.Error	DF	t-value	p-value
(Intercept)	38.083	7.742	15	4.919	0.0002
factor(pilltype)2	-21.550	5.972	15	-3.608	0.0026
factor(pilltype)3	-20.667	5.972	15	-3.461	0.0035
factor(pilltype)4	-7.017	5.972	15	-1.175	0.2583

	numDF	denDF	F-value	p-value
(Intercept)	1	15	14.266	0.0018
factor(pilltype)	3	15	6.257	0.0057

## Correlation within subjects

Covariance for two measurements from the same patient ( $j \neq k$ ):

$$\begin{aligned}\text{Cov}(Y_{ij}, Y_{ik}) &= \text{Cov}(\mu + \beta_j + B_i + \varepsilon_{ij}, \mu + \beta_k + B_i + \varepsilon_{ik}) \\ &= \text{Cov}(B_i + \varepsilon_{ij}, B_i + \varepsilon_{ik}) = \text{Cov}(B_i, B_i) = \text{Var}(B_i) = \sigma_{subj}^2\end{aligned}$$

Variance for a measurement:

$$\begin{aligned}\text{Var}(Y_{ij}) &= \text{Var}(\mu + \beta_j + B_i + \varepsilon_{ij}) \\ &= \text{Var}(B_i + \varepsilon_{ij}) = \text{Var}(B_i) + \text{Var}(\varepsilon_{ij}) = \sigma_{subj}^2 + \sigma_{\varepsilon}^2\end{aligned}$$

Correlation for two measurements from the same patient:

$$\text{corr}(Y_{ij}, Y_{ik}) = \frac{\text{Cov}(Y_{ij}, Y_{ik})}{\sqrt{\text{Var}(Y_{ij})}\sqrt{\text{Var}(Y_{ik})}} = \frac{\sigma_{subj}^2}{\sigma_{subj}^2 + \sigma_{\varepsilon}^2}$$

Estimate of correlation:  $\frac{15.900^2}{15.900^2 + 10.344^2} = 0.71$

We will then analyze the **birth weight example** using a random effects model

We here consider the model:

$$Y_{ij} = \mu + \beta_1 x_{1ij} + \beta_2 x_{2ij} + B_i + \varepsilon_{ij}$$

where

$Y_{ij}$  is the birth weight for the  $j$ -th baby of the  $i$ -th mother

$x_{1ij} = j$  is the birth order (parity) of the  $j$ -th baby of the  $i$ -th mother

$\beta_1$  is the effect of increasing the birth order by one

$x_{2ij}$  is the age of the  $i$ -th mother when she had her first baby

$\beta_2$  is the effect of one year's increase in the age of the mother

the  $B_i$  are the effects of mothers, assumed independent  $N(0, \sigma_{subj}^2)$

the  $\varepsilon_{ij}$  are random errors, assumed independent  $N(0, \sigma_{\varepsilon}^2)$

## R commands:

```
fit.babies=lme(bweight~birthord+initage, random=~1|momid, data=babies)
summary(fit.babies)
```

## R output (edited):

Linear mixed-effects model fit by REML

### Random effects:

Formula: ~1 | momid

	(Intercept)	Residual
StdDev:	358.18	445.02

### Fixed effects: bweight ~ birthord + initage

	Value	Std.Error*	DF	t-value*	p-value
(Intercept)	2526.62	163.34	799	15.469	0.0000
birthord	46.61	9.951	799	4.684	0.0000
initage	26.73	9.003	198	2.969	0.0034

Estimate of correlation for two babies by the same mother:

$$\frac{358.18^2}{358.18^2 + 445.02^2} = 0.39$$

\*) Standard errors and t-values differ slightly from those on page 277 in the text book, since we here use REML estimation.

## Longitudinal data and correlation structures

A random effects model for longitudinal data assumes that the correlation between any two observations for the same individual are the same

E.g. for the [birth weight example](#) a random effects model assumes the same correlation between the birth weights of the first and second child as for the first and fifth child

In general for longitudinal data, where observations are taken consecutively over time, it may be the case that observations that are close to each other in time are more correlated than those further apart

In order to fit a model for longitudinal data , we need to take into account the correlation of the observations

Assume that the observations for the  $i$ -th subject are  $Y_{i1}, Y_{i2}, \dots, Y_{im}$

Common assumptions on the correlation structure of the  $Y_{ij}$  are ( $j \neq k$ ):

Exchangeable:  $\text{corr}(Y_{ij}, Y_{ik}) = \rho$

Autoregressive of order 1 (AR1):  $\text{corr}(Y_{ij}, Y_{ik}) = \rho^{|k-j|}$

Independence:  $\text{corr}(Y_{ij}, Y_{ik}) = 0$

Note that a random effects model implies an exchangeable correlation structure

We may use the "gee" library to fit models with these correlation structures (using a method called generalized estimating equations)

## R commands:

```
library(gee)
```

```
summary(gee(bweight~birthord+initage,id=momid,data=babies,corstr="exchangeable"))
```

```
summary(gee(bweight~birthord+initage,id=momid,data=babies,corstr="AR-M"))
```

```
summary(gee(bweight~birthord+initage,id=momid,data=babies,corstr="independence"))
```

## R output (edited):

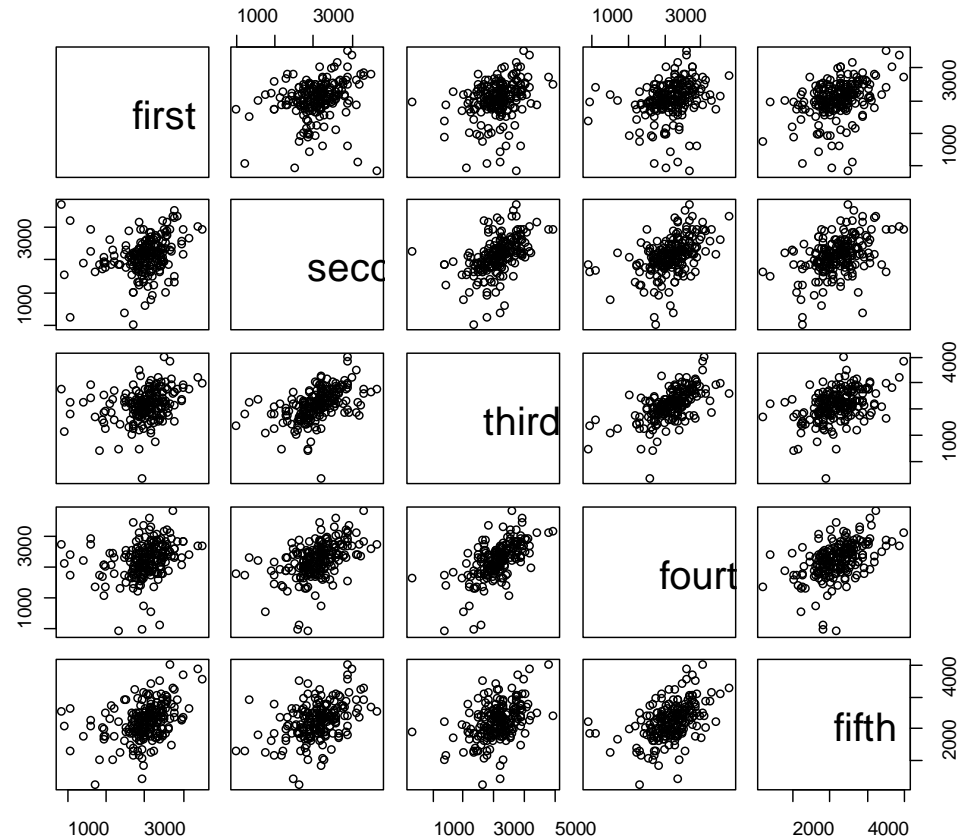
	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
birthord	46.61	9.96	4.68	10.00	4.66
initage	26.73	8.97	2.98	10.09	2.65
birthord	47.31	13.83	3.42	10.49	4.51
initage	27.41	7.83	3.50	9.67	2.83
birthord	46.61	12.76	3.65	10.00	4.66
initage	26.73	5.61	4.77	10.09	2.65

The naive SE and z are valid if the assumed correlation structure is true

Inference should be based on the robust SE and z, since these are valid also if the assumed "working correlation" does not hold

For the example with birth weight and birth order we have the following relation between the birth weights

The weight of the first baby is less correlated with the others. Otherwise the weights have about the same correlation. An exchangeable correlation structure is a reasonable "working assumption"



Correlations:

	first	second	third	fourth	fifth
first	1.000	0.228	0.295	0.258	0.381
second	0.228	1.000	0.483	0.468	0.426
third	0.295	0.483	1.000	0.619	0.423
fourth	0.258	0.468	0.619	1.000	0.464
fifth	0.381	0.426	0.423	0.464	1.000



In conclusion the birth weight data may be analyzed using generalized estimating equations with an exchangeable correlation structure (slide ) or by using a random effects model (slide)

The two models give comparable results in this example

However, if we extend the generalized estimating approach and the random effects model to generalized linear models (like logistic regression), the results no longer agree

This is, however, a topic beyond the scope of these lectures.