STK4900/9900 - Lecture 10

Program

1. Repeated measures and longitudinal data
2. Simple analysis approaches
3. Random effects models
4. Generalized estimating equations

- Sections 7.1, 7.2, 7.3, 7.4 (except 7.4.5), 7.5

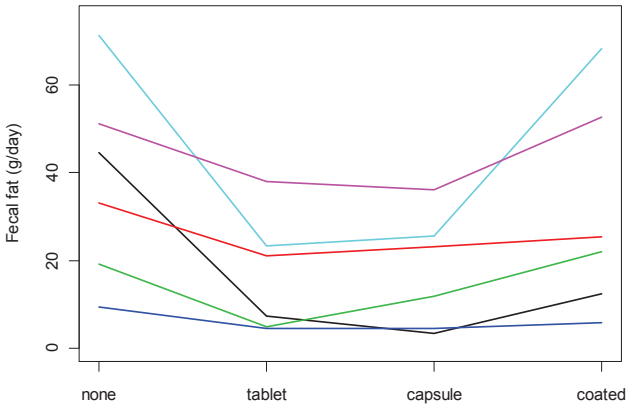
Example: Fecal fat

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)



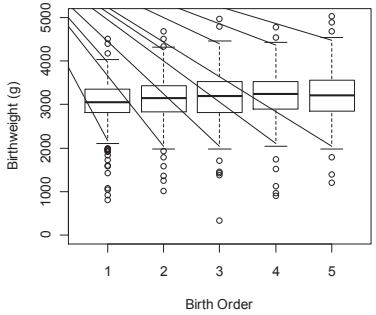
The plot shows that some patients tend to have high values for all pill types, while other patients tend to have low values

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

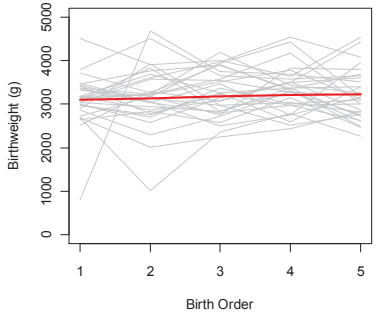
Example: Birth weight and birth order

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

1. Standard analysis ignoring the dependence
   Fecal fat ex: One-way ANOVA on pill type
   Birth weight ex: Linear regression on birth order
   However: Ignoring dependence is WRONG!
   Nor pursued further.

2. Looking at parts of the data for an individual to avoid the dependence problem
   Fecal fat ex: one option is to compare two pill types at a time using the paired t-test
   Birth weight example one option is to look at the difference in weight between the fifth and the first child.
   This is not wrong, but ignores part of the data (birth weight) or gives many comparisons (fecal fat)
   See slides 7-8.

Approaches, contd.

3. Including dependence as fixed factor variable
   Fecal fat ex: Two-way ANOVA on pill type and individual
   Birth weight ex: Linear regression on birth order with mother as factor variable
   However: Generally not interested in factors individual/mother. Also in birth weight ex many factor levels.

4. Including dependence as random factor: Random effects model
   Fecal fat ex: Two-way ANOVA on pill type and individual as random factor
   Birth weight ex: Linear regression on birth order with mother as random factor variable
   Pro: Individual/mother modeled as random variable.

Fecal fat example:

R commands (comparing pill types 1(none) and 2(Tablet)):

```
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-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 weights 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

Approach 3: Individual as fixed factor

Fecal fat example: Two way ANOVA

R commands and output:

```
> anova(lm(fecfat~factor(pilltype)+factor(subject),data=fecfat))

Response: fecfat
                Df Sum Sq Mean Sq F value    Pr(>F)
factor(pilltype)  3 2008.6  669.53  6.2574 0.0057406 **
factor(subject)   5 5588.4 1117.68 10.4457 0.0001821 ***
Residuals        15 1605.0  107.00
```

Birth weight example: Two way ANOVA

R commands and output:

```
> babiesanova=lm(bweight~birthord+initage+factor(momid),data=babies)
> anova(babiesanova)

Response: bweight
               Df    Sum Sq Mean Sq F value    Pr(>F)
birthord        1   4344611 4344611 21.9375 3.310e-06 ***
initage         1   7401953 7401953 37.3750 1.523e-09 ***
factor(momid) 198 166220126  839496  4.2389 < 2.2e-16 ***
Residuals     799 158238219  198045
```


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 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

Approach 4: Random effects model

A drawback of Approach 3 is that an effect is estimated for every individual. The interest, however, lies in how such effects will vary over a population.

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: $Y_{ij} = \mu + \alpha_j + B_i + \varepsilon_{ij}$

where

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

α_j is the effect of pill type j (relative to type 1)

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

the ε_{ij} are random errors, assumed independent $N(0, \sigma^2_{\varepsilon})$

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

Correlation within subjects

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

$$\text{Cov}(Y_{ij}, Y_{ik}) = \text{Cov}(\mu + \alpha_j + B_i + \varepsilon_{ij},\ \mu + \alpha_k + B_i + \varepsilon_{ik})$$

$$= \text{Cov}(B_i + \varepsilon_{ij},\ B_i + \varepsilon_{ik}) = \text{Cov}(B_i,\ B_i) = \text{Var}(B_i) = \sigma^2_{subj}$$

Variance for a measurement:

$$\text{Var}(Y_{ij}) = \text{Var}(\mu + \alpha_j + B_i + \varepsilon_{ij})$$

$$= \text{Var}(B_i + \varepsilon_{ij}) = \text{Var}(B_i) + \text{Var}(\varepsilon_{ij}) = \sigma^2_{subj} + \sigma^2_{\varepsilon}$$

Correlation for two measurements from the same patient:

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

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

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

β_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

β_2 is the effect of one year's incerase in the age of the mother

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

the ε_{ij} are random errors, assumed independent $N(0, \sigma^2_{\varepsilon})$

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 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 form 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 is 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 type of correlation between the observations

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

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

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

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

Unstructured: $\text{corr}(Y_{ij}, Y_{ik}) = \rho_{jk}$

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

R output (edited):

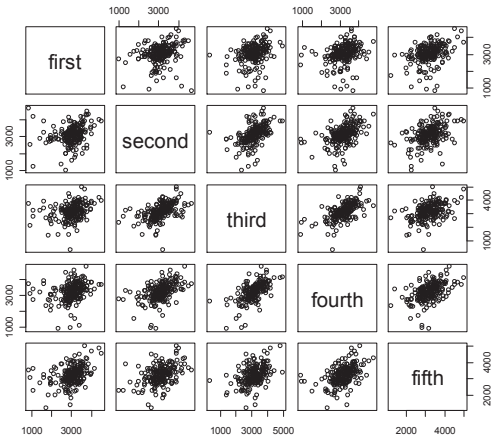
	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
birthord	46.61	9.96	4.68	10.00	4.66
initage	26.73	8.97	2.98	10.09	2.65
birthord	47.31	13.83	3.42	10.49	4.51
initage	27.41	7.83	3.50	9.67	2.83
birthord	44.70	9.95	4.49	9.82	4.55
initage	28.07	8.81	3.19	9.12	3.08
birthord	46.61	12.76	3.65	10.00	4.66
initage	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 14) or by using a random effects model (slide 11)

The two models give comparable results in this example

However, if we extend the generalized estimating (GEE) approach and the random effects model to generalized linear models (like logistic regression), the results need not longer agree.

The GEE approach can be used for all glms (distributional families, link functions). We will only consider extension of logistic regression to dependent data.

GEE and binary data (logistic model)

Example: Birth weight data, but with an indicator of low birth weight (lowbrth), i.e. < 3000g.

```
geefit<-gee(lowbrth~birthord+initage,id=momid,family=binomial,data=babies,
                                                    corstr="unstructured")
```

R output (edited):

```
> summary(geefit)

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

Coefficients:
             Estimate Naive S.E.   Naive z Robust S.E.  Robust z
(Intercept)      1.34     0.60      2.24       0.60       2.24
birthord       -0.082    0.038     -2.16      0.038      -2.15
initage        -0.093    0.034     -2.76      0.034      -2.77
```

Negative coefficients for birth order and initial age indicate that low birth weigh is less likely with increasing order and age.

GEE and Birth weight data, contd.

More R output (edited):

```
Working Correlation
     [,1]   [,2]   [,3]   [,4]   [,5]
[1,] 1.00   0.16   0.24   0.28   0.28
[2,] 0.16   1.00   0.47   0.31   0.30
[3,] 0.24   0.47   1.00   0.40   0.36
[4,] 0.28   0.31   0.40   1.00   0.27
[5,] 0.28   0.30   0.36   0.27   1.00
```

Even the homemade expcoef function works on gee-objects

More R output (edited):

```
> expcoef(geefit)
                expcoef     lower      upper
(Intercept)        3.84      1.18      12.45
birthord           0.92      0.85       0.99
initage            0.91      0.85       0.97
```


Hierarchical models (Vittinghoff et al., Ch. 7.2)

- Example: Lung capacity at patient visits within physician
- Example: Average grades in classes within schools

Here we are generally not interested in neither patients nor individual physicians. Should be modeled as random effects. Similarly for the school example.

This generates more general dependency structures than discussed earlier and should be analyzed accordingly. We will not go into details.

Time series data (not in Vittinghoff et al.)

In this lecture we have so far considered data with

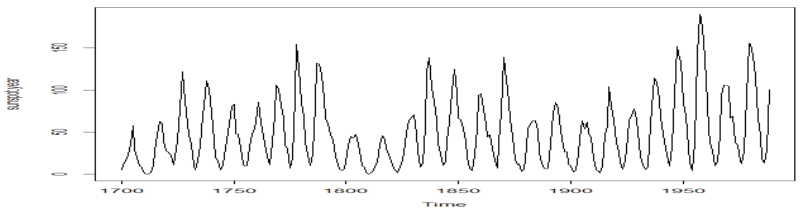
- many small independent groups of subjects (measurements)
- dependence within each group

Time series data is a different dependent data structure with

- long sequences of correlated data $Y_1, Y_2, Y_3, ...$
- only one (or maybe a few) such sequences

Examples

- Temperature on consecutive days (weeks)
- Stock prices on consecutive days (weeks)
- Sunspot activity years 1700-1988



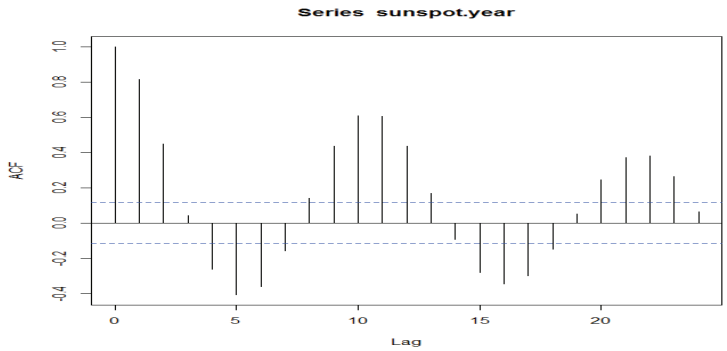
Autocorrelation function (ACF)

The correlation between observation at time t, Y_t, and its lag at time t-k, Y_{t-k}, is given

$$\hat{\rho}(k) = corr(Y_t, Y_{t-k})$$

The $\hat{\rho}(k), k = 0, 1, 2, ...$ is referred to as the autocorrelation function.

For the sunspot data we have ACF with high positive correlations at 11 year cycles



Uncertainty limits for the ACF

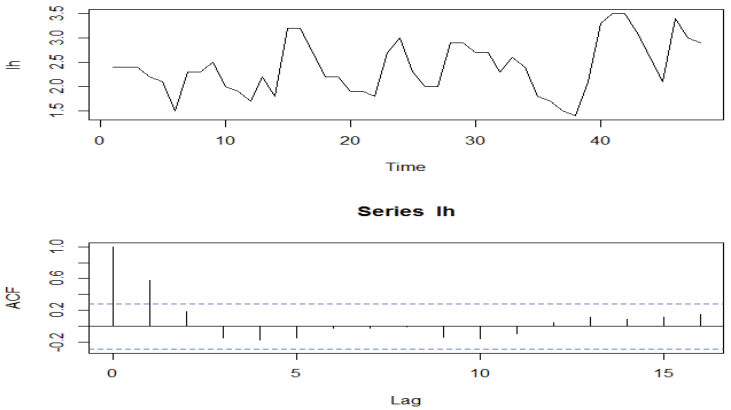
The dashed horizontal lines in the ACF plots lies at values $\pm 1.96/\sqrt{n}$
In particular for the sunspot data with n=289 this becomes ±0.118

These limits correspond to the test in Lecture 2 for $H_0 : \rho = 0$
using test statistic $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

Then correlations within the uncertainty limits are not significantly different from zero.

Example: Luteinizing hormone in blood samples at 10 min. intervals from a human female, 48 samples.

The sunspot example is a bit unusual in the 11-year correlation pattern. The luteinizing hormone example may be more typical of a time-series in the sense only the first (few) correlations are significantly different from 0.



Autoregressive models AR(p)

Another method for analyzing the dependence in a time series is through autoregressive models where the current value Y_t is regressed on previous p values $Y_{t-1}, Y_{t-2}, Y_{t-3}, ..., Y_{t-p}$ of the series, i.e. through a model

$$Y_t = a_1 Y_{t-1} + a_2 Y_{t-2} \cdots + a_p Y_{t-p} + \varepsilon_t$$

where the a_k are regression coefficients and the ε_t independent error terms. This is called an AR(p) model.

For the special case AR(1) we have

$$Y_t = a_1 Y_{t-1} + \varepsilon_t$$

and the current value only depends on the previous value. This is referred to as a Markov property.