

# STK4900/9900 - Lecture 10

## Program

1. Repeated measures and longitudinal data
  2. Simple analysis approaches
  3. Random effects models
  4. Generalized estimating equations (GEE)
  5. GEE for binary data (and GLMs)
  6. Time series data
- Sections 7.1, 7.2, 7.3, 7.4 (except 7.4.5), 7.5
  - Additional material on time series

## Example: Fecal fat

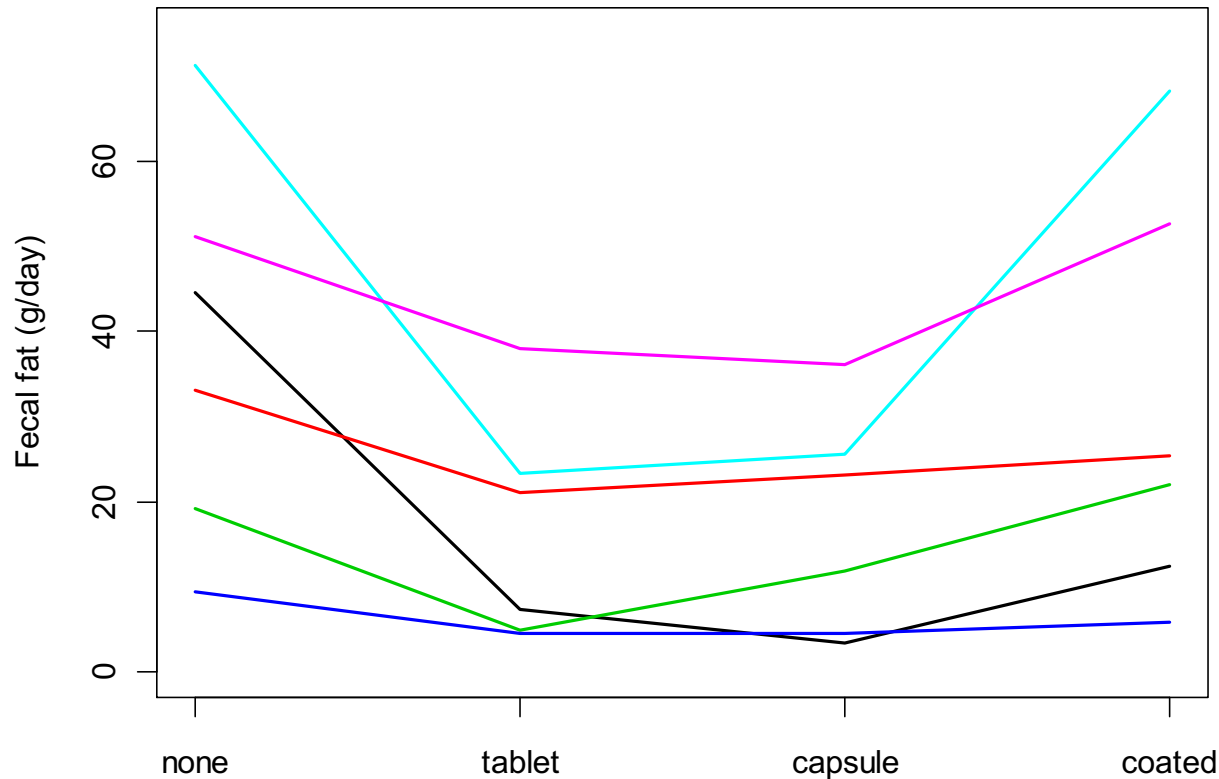
Lack of digestive enzymes in the intestine can cause bowel absorption problems, which will be indicated by excess fat in the feces. Pancreatic enzyme supplements can reduce the problem. The data are from a study to determine if the form of the supplement makes a difference

**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)



In this plot each patients responses are displayed as a line

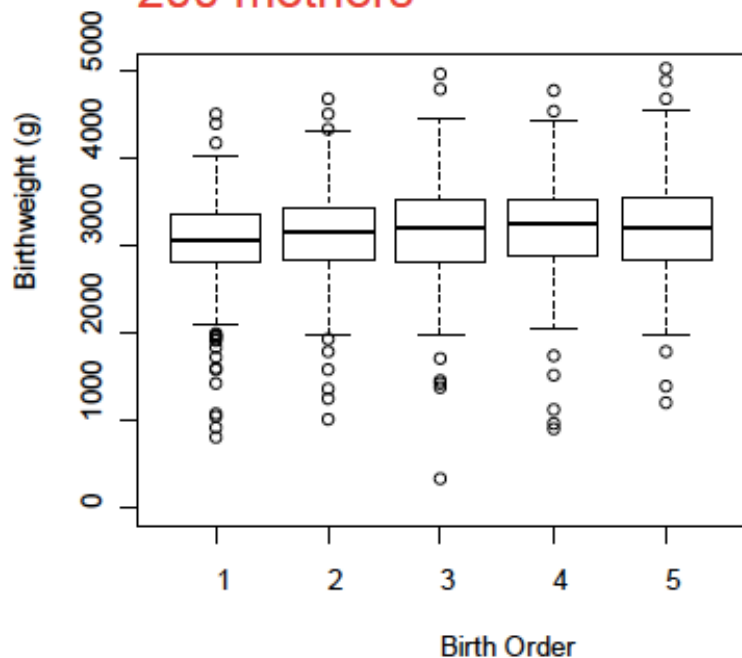
The plot shows that some patients tend to have high values for all pill types, while other patients tend to have low values

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

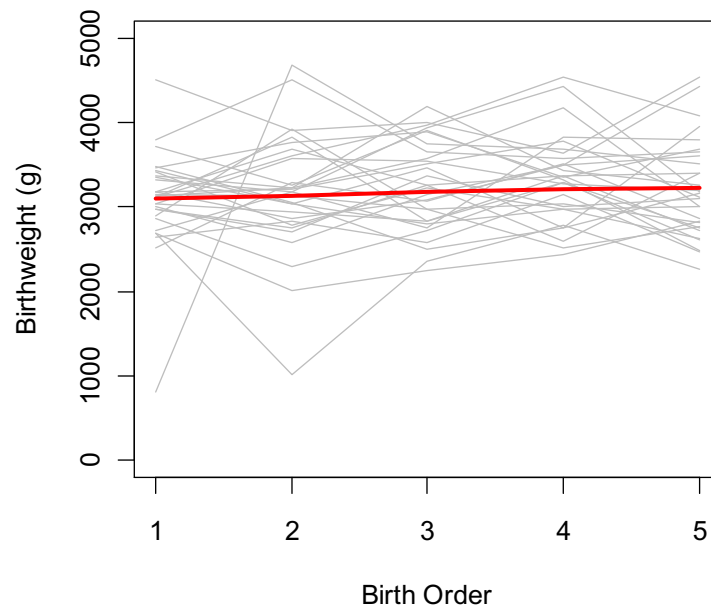
## Example: Birth weight and birth order

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!**  
Not 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.

## Pairwise comparisons: 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/data/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 %

## Pairwise comparisons: Birth weight example

### R commands:

```
babies=read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/data/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			

## 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  $j$

$\alpha_j$  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:  $\sim 1 \mid \text{subject}$   
(Intercept)      Residual  
StdDev:      15.900      10.344

The diagram shows two blue arrows. One arrow points from the label  $S_{\text{subj}}$  to the value 15.900 under the (Intercept) column. The other arrow points from the label  $S_{\epsilon}$  to the value 10.344 under the Residual column.

### 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 ( $j \neq k$ ):

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

Variance for a measurement:

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

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:

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

This analysis is also found on p. 288 in the book.

## 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  $Y_{i1}, Y_{i2}, \dots, Y_{im}$

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

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

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

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):

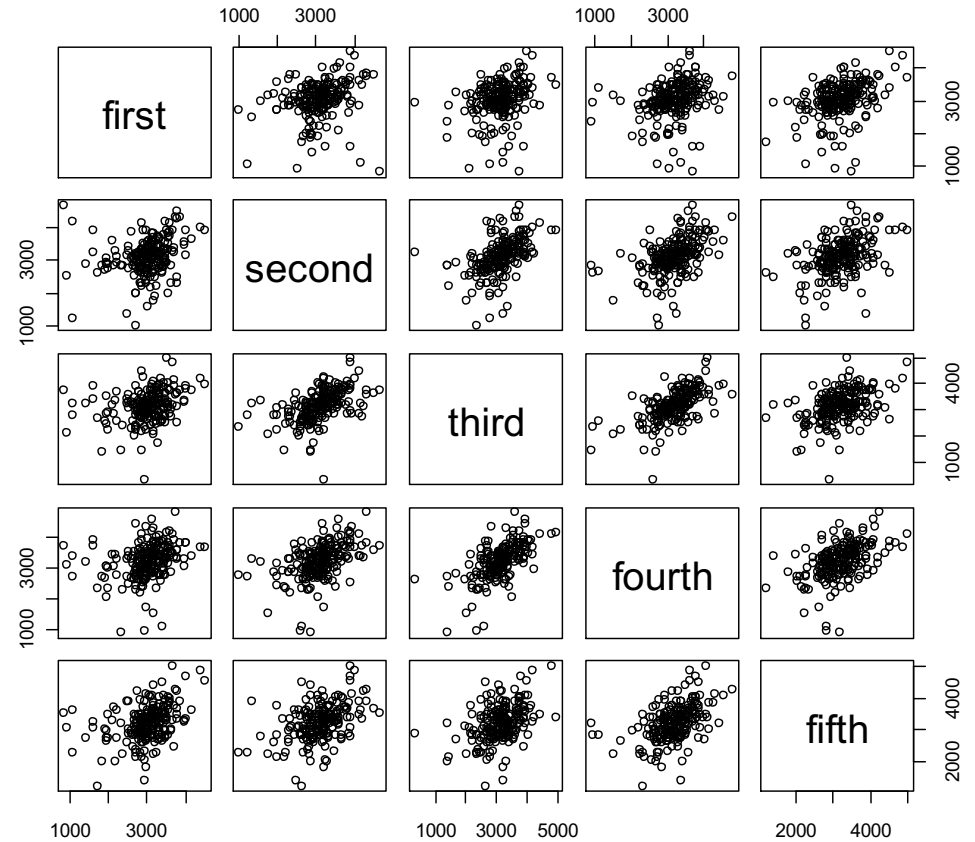
	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
birthord	46.61	9.96	4.68	10.00	4.66
initage	26.73	8.97	2.98	10.09	2.65
birthord	47.31	13.83	3.42	10.49	4.51
initage	27.41	7.83	3.50	9.67	2.83
birthord	44.70	9.95	4.49	9.82	4.55
initage	28.07	8.81	3.19	9.12	3.08
birthord	46.61	12.76	3.65	10.00	4.66
initage	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 17 ) or by using a random effects model (slide 14)

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 weight is less likely with increasing order and age.

## GEE and Birth weight data, contd.

Even the homemade `expcoef` function from Lecture 7 works on `gee`-objects

### More R output (edited):

```
> expcoef(geefit)
```

	<code>expcoef</code>	<code>lower</code>	<code>upper</code>
<code>(Intercept)</code>	3.84	1.18	12.45
<code>birthord</code>	0.92	0.85	0.99
<code>initage</code>	0.91	0.85	0.97

## More on dependent data: Time series data (not in Vittinghoff et al.)

In this lecture we have so far considered data with

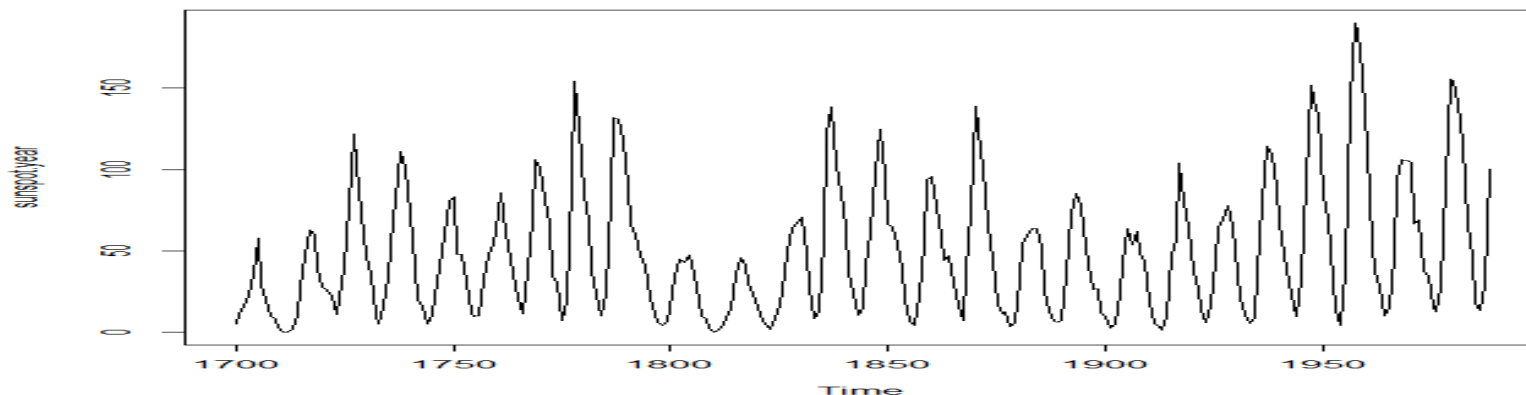
- many **small independent groups** of subjects (measurements)
- dependence within each group

Time series data is a different dependent data structure with

- **long sequences** of correlated data  $Y_1, Y_2, Y_3, \dots$
- **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



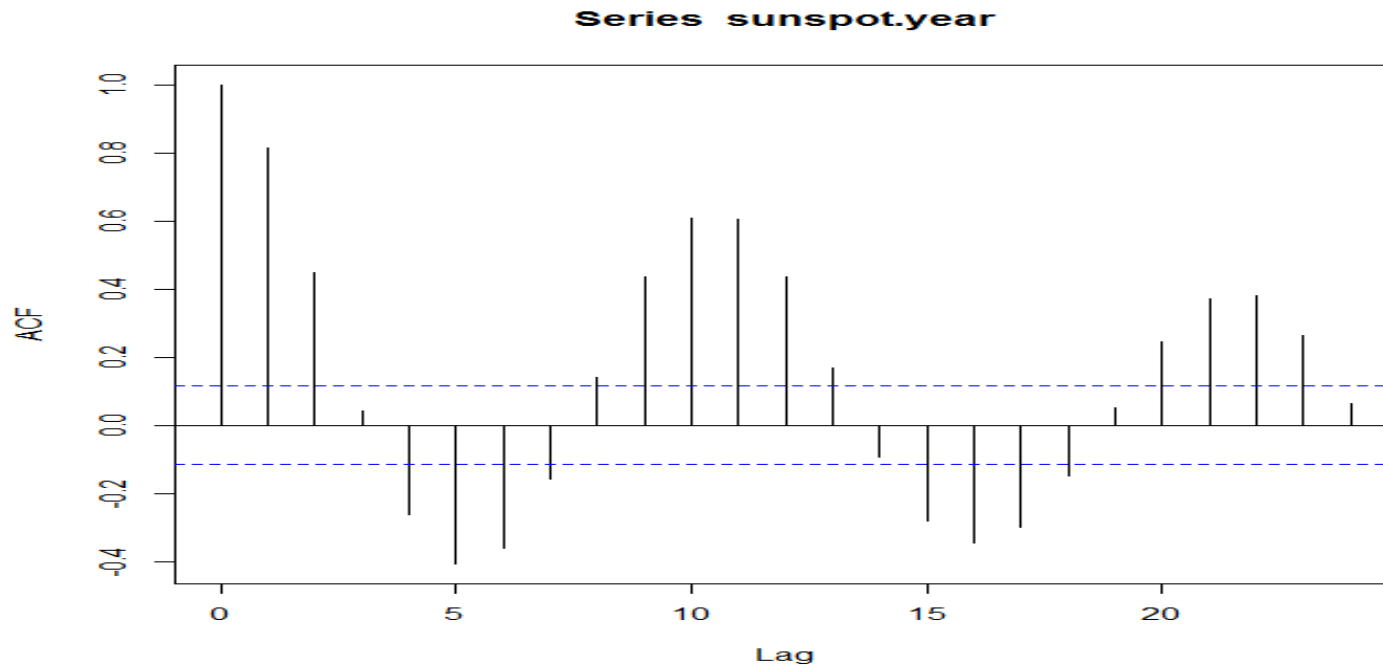
# Autocorrelation function (ACF)

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

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

The  $\hat{\rho}(k), k = 0, 1, 2, \dots$  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 lie at values  $\pm 1.96 / \sqrt{n}$   
In particular for the sunspot data with  $n=289$  this becomes  $\pm 0.118$

These limits correspond to the test in Lecture 2 for  $H_0 : \rho = 0$   
using test statistic

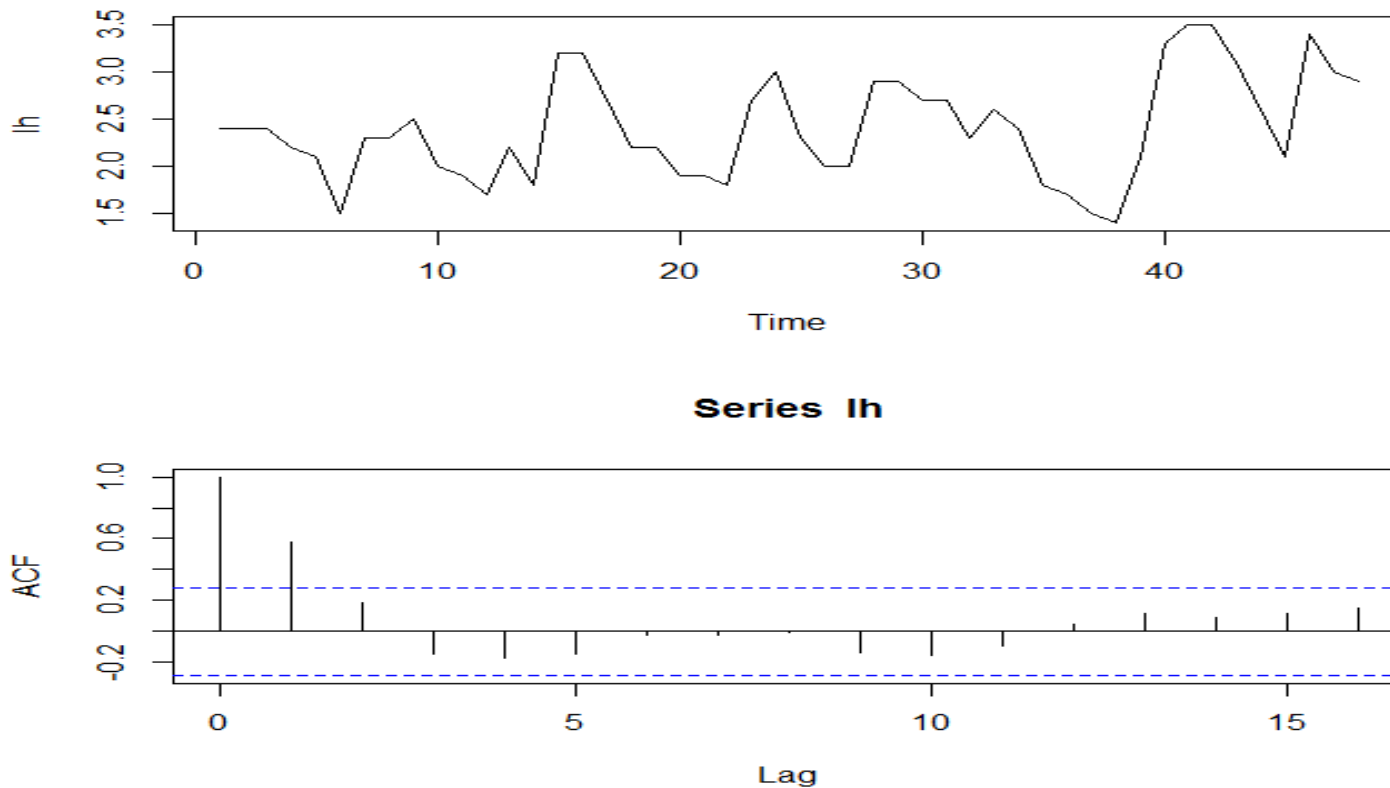
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

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

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

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

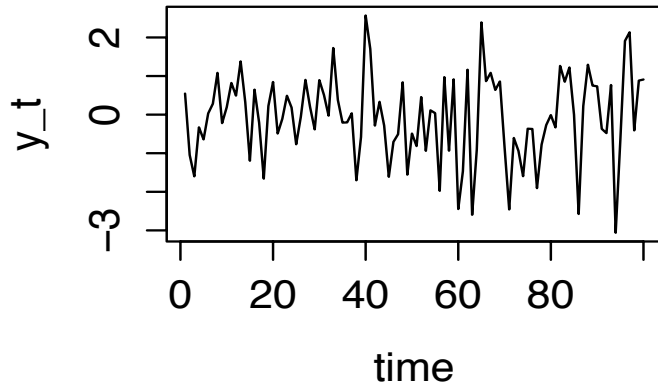
For the special case AR(1) we have

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

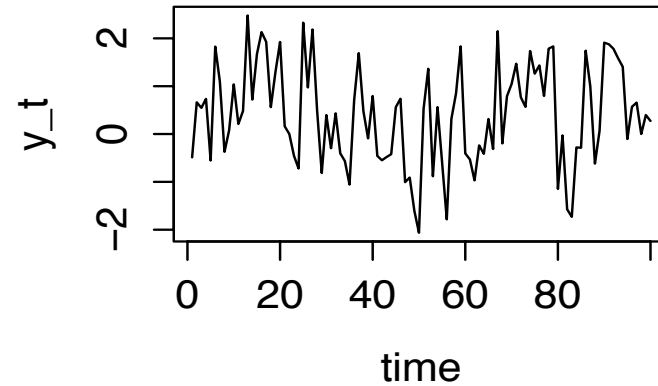
and the current value only depends on the previous value. This is referred to as a Markov property.

Simulated time series from an AR(1) model  $Y_t = a_1 Y_{t-1} + \varepsilon_t$   
with  $a_1 = 0.1, 0.5, 0.9$  and  $-0.9$

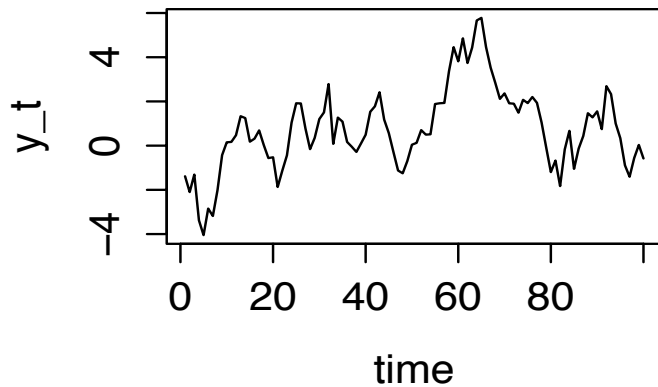
**0.1**



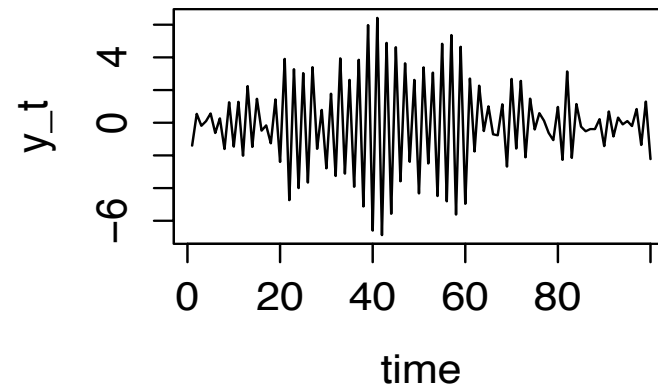
**0.5**



**0.9**



**-0.9**



## R commands:

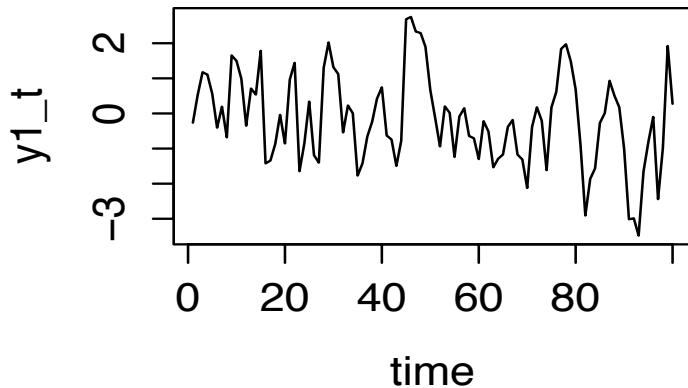
```
par(mfrow=c(2,2))
time =seq(1,100,1)
a = c(0.1, 0.5, 0.9,-0.9)
for(i in 1:4)
{
  y_t<- arima.sim(n,model=list(ar=a[i]),sd=1)
  plot(time, y_t, type="l", main=a[i])
}
```

AR(p) is a special case of the more general **ARIMA(p,d,q)** models for time series. AR-I-MA stands for AutoRegressive Integrated Moving Average, and **p** is the order of the autoregressive process, **d** is the order of a difference operator (helps against nonstationarity) and **q** is the order of the moving average process (where the time series relates to the q past values of the noise).

The parameters in such models can be estimated via maximum likelihood estimation!

**R command:** `arima(y,order=c(p,d,q))` # in its most simple version

**true a1 = 0.5**



```
> arima(y1_t, order=c(1,0,0))
```

Call:

```
arima(x = y1_t, order = c(1, 0, 0))
```

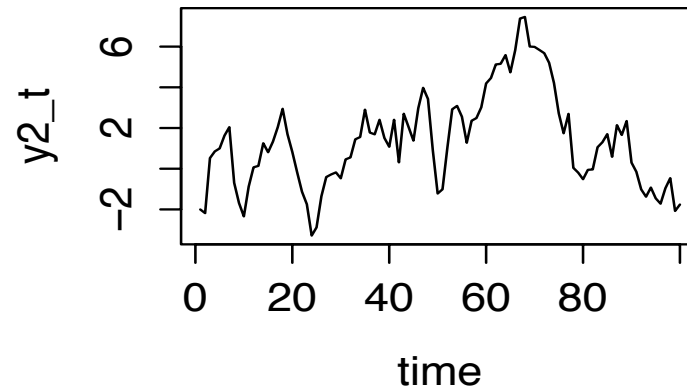
Coefficients:

	ar1	intercept
	<u>0.5846</u>	-0.1587
s.e.	0.0799	0.2491

sigma<sup>2</sup> estimated as 1.101:

log likelihood = -146.89, aic = 299.79

**true a1 = 0.9**



```
> arima(y2_t, order=c(1,0,0))
```

Call:

```
arima(x = y2_t, order = c(1, 0, 0))
```

Coefficients:

	ar1	intercept
	<u>0.9013</u>	0.8586
s.e.	0.0432	1.0011

sigma<sup>2</sup> estimated as 1.105:

log likelihood = -147.73, aic = 301.46

## THE END OF THE CURRICULUM

It has been a strange situation and new teaching experience for the second part. We hope that it has been possible to follow the course in an acceptable way.

Please send feedback on how things have worked out to [osamuels@math.uio.no](mailto:osamuels@math.uio.no)  
Is there anything we can do that would be helpful?

We intend to proceed as normal as possible and so I will make the **second mandatory assignment** available in the middle of next week with the original deadline  
Thursday April 16<sup>th</sup> 2020 before 14:30

The original plan was that the final exam would be a

Written 4 hours exam 14:30:00-18:30 Friday May 29<sup>th</sup> 2020 (Silurveien)

It is absolutely possible that the time and exam form will be different, but such a decision has not been made yet (as of Wednesday 18<sup>th</sup>)

Lets make the best of it, Good luck to everybody!

