Lecture 10 – Program

1. Time-dependent data in general

2. Repeated Measurements

3. Time series

4. Time series that depend on "covariate" time-series

Time-dependent data:

Outcomes that are measured at several times, for instance:

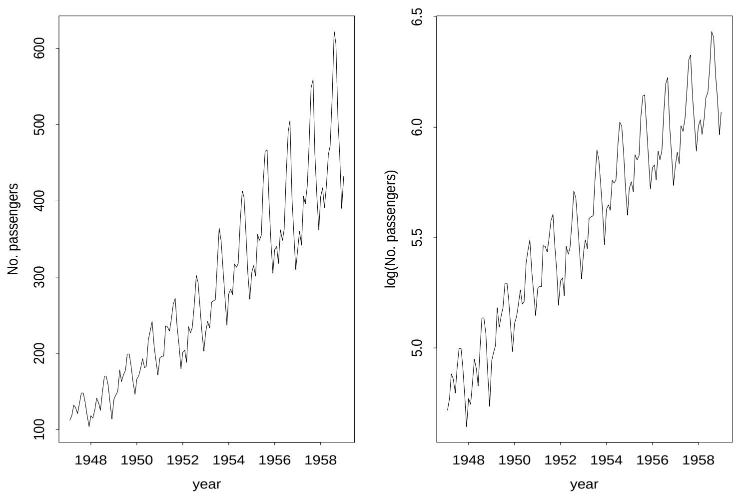
- $y_t =$ temperature day $t = 0, 1, 2, \ldots$
- $y_t =$ precipitation day t
- $y_t =$ price of a stock day t
- $y_{it} =$ weight rat no. i day t.

The outcomes can in general be

- on a continous scale (often assumed normally distributed)
- counts (perhaps Poisson-distributed)
- binary (0/1)

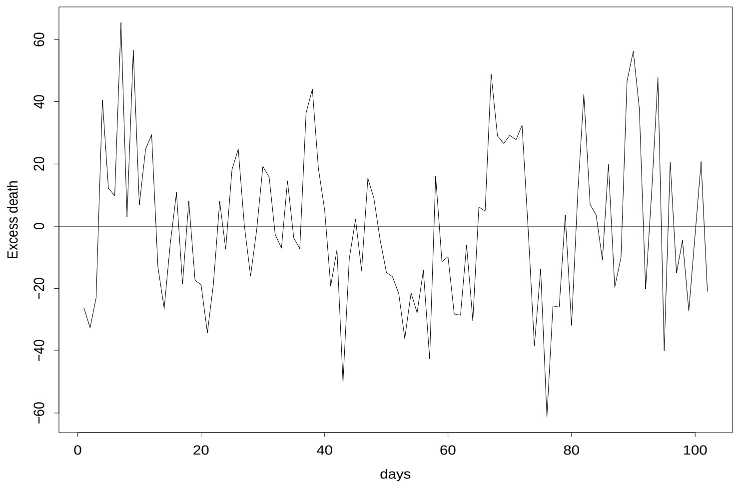
In this lecture: Only continues measurements.

Example: airline passengers (original and log-scale:)



- Time-trend
- Seasonal variation

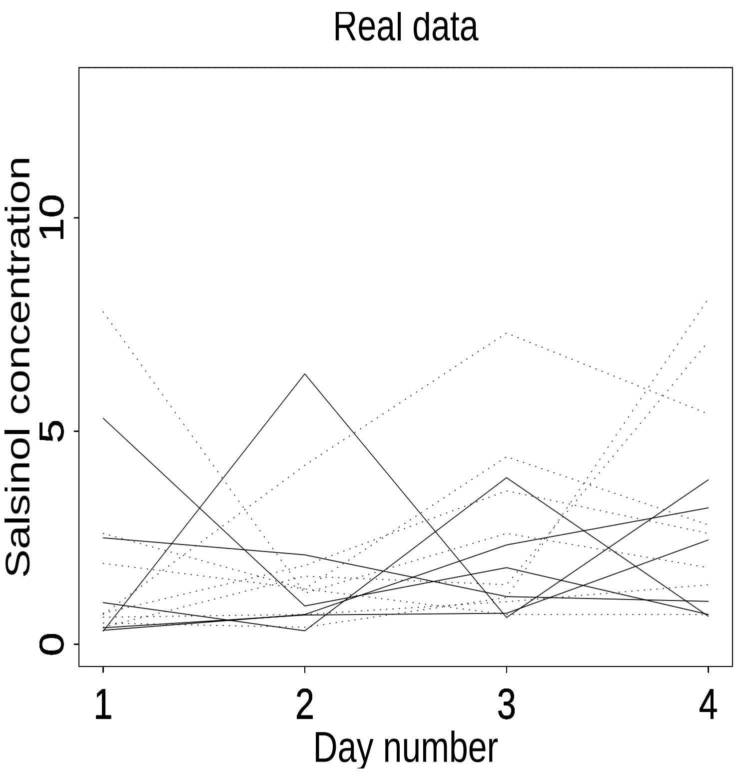
Example: excess deaths in London



(centered and removed effect of flu epidemics)

Example: salsinol-data

$y_{it} =$ salsinol-measurement
at day $t = 1, 2, 3, 4$
for individual $i = 1, 2, \ldots, 14$



Each line represents the measurements on one individual.

Time series vs. repeated measurments

Useful to distinguish between:

- Time series: One (or a few) very long series of measurements
- Repeated measurements: Many short series of measurements

In the examples:

- Airline passengers: Time serie
- Excess deaths: Time serie (with parallel series of temperature and smoke)
- Salsinol data: Repeated measurements

Will typically use different methods to analyze time-series and repeated measurments.

A simple model for repeated measurements:

$$y_{it} = a_i + b_i t + \varepsilon_{it}$$

where the ε_{it} are all independent and the a_i and b_i are specific to individual i

Note that this is just assuming different linear regression models for different individuals.

The data may then be tranformed to least squares estimates $(\widehat{a}_i, \widehat{b}_i)$ for $i = 1, 2, \ldots, n$

Example: salsinol-data

These data consist of measurements on two groups

- Moderately alcohol dependent individuals

- Severly alcohol dependent individuals

A possible question is then:
Are the lines for the two groups different?

A model for making it possible to test this statement could be (with some awkward notation)

$$\widehat{a}_i \sim \mathrm{N}(\alpha_j, \sigma^2)$$

$$\widehat{b}_i \sim \mathrm{N}(\beta_j, \tau^2)$$

where α_j and β_j are the expectations in the groups $j = 1, 2$.

Example, contd. : salsinol-data

We could then test whether intercepts and slopes are the same in the two groups

$$\text{H}_0 : \alpha_1 = \alpha_2 \quad \text{and} \quad \text{H}_0 : \beta_1 = \beta_2$$

by means of standard t-tests.

Let $\bar{a}_j$ and $\bar{b}_j$ be the averages in of $\widehat{a}_i$ and $\widehat{b}_i$ in group j. Then the statistics for the t-tests can be written as:

$$t_\alpha = \frac{\bar{a}_2 - \bar{a}_1}{\text{se}(\bar{a}_2 - \bar{a}_1)} \quad \text{and} \quad t_\beta = \frac{\bar{b}_2 - \bar{b}_1}{\text{se}(\bar{b}_2 - \bar{b}_1)}$$

On the next pages follows R-code for doing these t-tests.

R-code for salsinol-data: reading the data

```
> salsinol0<-matrix(scan("salsinol.dat"),byrow=T,ncol=6)
Read 84 items
> salsinol0
      [,1] [,2] [,3] [,4] [,5] [,6]
 [1,]    2    1 0.33 0.70 2.33 3.20
 [2,]    8    1 5.30 0.90 1.80 0.70
 [3,]    9    1 2.50 2.10 1.12 1.01
 [4,]   11    1 0.98 0.32 3.91 0.66
 [5,]   12    1 0.39 0.69 0.73 2.45
 [6,]   13    1 0.31 6.34 0.63 3.86
 [7,]    1    2 0.64 0.70 1.00 1.40
 [8,]    3    2 0.73 1.85 3.60 2.60
 [9,]    4    2 0.70 4.20 7.30 5.40
[10,]    5    2 0.40 1.60 1.40 7.10
[11,]    6    2 2.60 1.30 0.70 0.70
[12,]    7    2 7.80 1.20 2.60 1.80
[13,]   10    2 1.90 1.30 4.40 2.80
[14,]   14    2 0.50 0.40 1.10 8.10
```

R-code for salsinol-data: fitting individual regressions

```
> I<-seq(1,4)
> coefest<-numeric(0)
> for (i in 1:14) {
+   newlm<-lm(salsinol0[i,3:6]~I)
+   coefest<-rbind(coefest,newlm$coef)
+   }
> coefest
      (Intercept)      I
 [1,]      -0.920  1.024
 [2,]       5.400 -1.290
 [3,]       3.045 -0.545
 [4,]       0.810  0.263
 [5,]      -0.490  0.622
 [6,]       1.550  0.494
 [7,]       0.290  0.258
 [8,]       0.355  0.736
 [9,]       0.100  1.720
[10,]      -2.350  1.990
[11,]       2.900 -0.630
[12,]       7.500 -1.660
[13,]       1.150  0.580
[14,]      -3.350  2.350
```

R-code for salsinol-data: t-tests

```
> t.test(coefest[1:6,2],coefest[7:14,2],var.equal=T)

        Two Sample t-test

data:  coefest[1:6, 2] and coefest[7:14, 2]
t = -0.9019, df = 12, p-value = 0.3848
alternative hypothesis:
true difference in means is not equal to 0
95 percent confidence interval:
 -1.9583198  0.8116531
sample estimates:
 mean of x  mean of y
0.09466667 0.66800000

> t.test(coefest[1:6,2],coefest[7:14,2])

        Welch Two Sample t-test

data:  coefest[1:6, 2] and coefest[7:14, 2]
t = -0.9644, df = 11.75, p-value = 0.3543
alternative hypothesis:
true difference in means is not equal to 0
95 percent confidence interval:
 -1.871725  0.725058
sample estimates:
 mean of x  mean of y
0.09466667 0.66800000
```

Example: Salsinol, cont.

Same results as in B&S (rounding error?).

Remark that the default is **not** to assume equal variances in the two samples.

Other models for repeated measurements

1) Ante-dependence which allows for dependence on previous measurements:

$$y_{it} = a_i + \gamma_t \, y_{i,t-1} + \varepsilon_{it}$$

This model can be extended in various ways, for instance:

$$y_{it} = a_i + b_i \, t + \gamma_t \, y_{i,t-1} + \varepsilon_{it}$$

But the extensions would typically require more than 4 measurements for the individual series.

2) Two-way ANOVA with time and individuals as factors

3) Vector respons. To be treated later

Time series analysis

Common models:

- Autoregressiv models: AR(p)
- Moving average models: MA(q)
- ARMA(p,q)
- ARIMA(p,q,d) where I is for "integrated"

We shall only discuss Autoregressiv models in any detail.

Autoregressiv models: AR(p)

The present observation y_t depends on the previous p observations:

$$y_t = a_1\, y_{t-1} + a_2\, y_{t-2} + \cdots + a_p\, y_{t-p} + \varepsilon_t$$

Note that this is linear regression model

- with response variable y_t
- and covariates $y_{t-1}, \ldots, y_{t-p}$

The model may thus be fitted with standard software for linear regression.

However, specially designed software for such data is widely available and may be more convenient.

Example: excess deaths in London

R contains a function `ar` for fitting autoregressiv models:

```
>ar(london$exc)

Call:
ar(x = london$exc)

Coefficients:
      1        2        3        4        5
 0.2889   0.1893   0.0686   0.0072  -0.0923
      6        7
-0.3302   0.2355
Order selected 7  sigma^2 estimated as  518.5
```

which gives (approx) the fitted relation:

$$
\begin{aligned}
y_t = \; & 0.29y_{t-1} + 0.19y_{t-2} + 0.07y_{t-3} \\
& +0.007y_{t-4} - 0.09y_{t-5} \\
& -0.33y_{t-6} + 0.24y_{t-7}
\end{aligned}
$$

Alternatively by `lm`

We need to set up the data differently:

```
> excess<-cbind(london$exc[1:95],london$exc[2:96],
+ london$exc[3:97],london$exc[4:98],london$exc[5:99],
+ london$exc[6:100],london$exc[7:101],london$exc[8:102])
> excess<-as.data.frame(excess)
> names(excess)<-c("y1","y2","y3","y4","y5","y6","y7","y8")
```

and then the model is fitted as

```
> lm(y8~y7+y6+y5+y4+y3+y2+y1-1,data=excess)$coef
         y7          y6          y5          y4
 0.27435162  0.21850415  0.04409311  0.02323063
          y3          y2          y1
 -0.07236904 -0.33202094   0.23933006
```

which similar, but not identical, to the what the `ar` function did.

Example, cont: excess death

In order to check for significance:

```
> round(summary(lm(y8~y7+y6+y5+y4+y3+y2+y1,
  data=excess))$coef,4)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5547     2.2470 -0.2469   0.8056
y7            0.2745     0.1013  2.7094   0.0081
y6            0.2183     0.1003  2.1764   0.0322
y5            0.0439     0.1020  0.4306   0.6678
y4            0.0240     0.1010  0.2376   0.8128
y3           -0.0719     0.1000 -0.7188   0.4742
y2           -0.3317     0.0977 -3.3952   0.0010
y1            0.2391     0.0997  2.3993   0.0186
```

so that 1st, 2nd, 6th and 7th lag appears to affect todays value.

Two useful function

- Auto-covariance function:

  $\gamma(k) = \text{Cov}(y_t, y_{t-k})$

- Auto-correlation function:

  $\rho(k) = \text{corr}(y_t, y_{t-k})$

with estimates $\widehat{\gamma}(k)$ and $\widehat{\rho}(k)$.

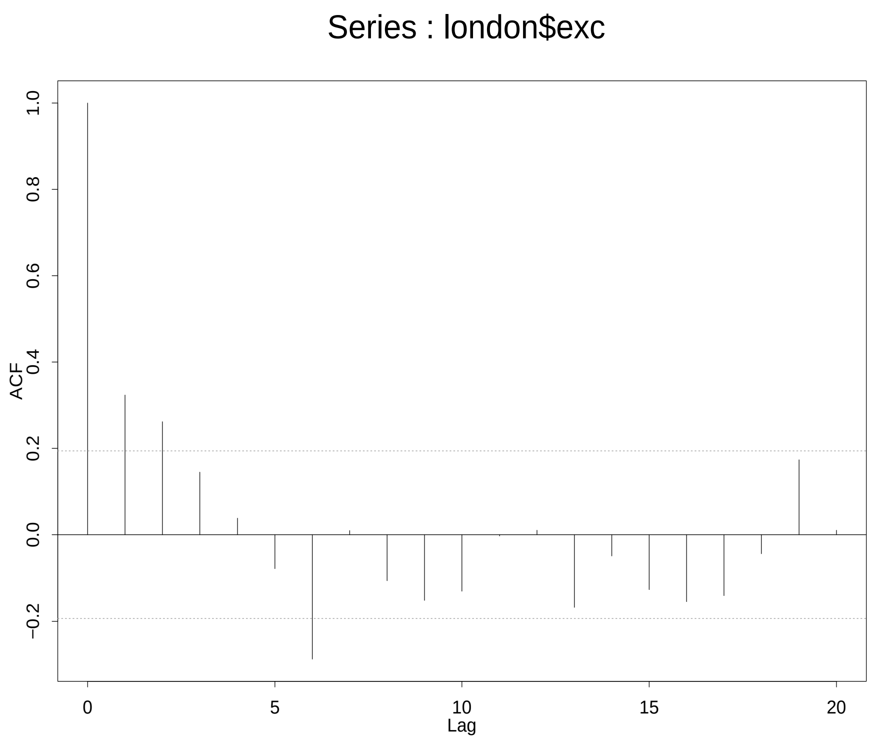
If $\rho(k) = 0$ and we have observed the time series for T days, $\widehat{\rho}(k)$ has standard error approximately equal to $1/\sqrt{T}$.

In this case we would expect $\widehat{\rho}(k)$ to lie within

$$[-2/\sqrt{T}\ ,\ +2/\sqrt{T}]$$

Plot of autocorrelation coefficient (ACF)

Excess death data:



Special case: AR(1)

$$y_t = a y_{t-1} + \varepsilon_t$$

This model have the **Markov** property:

$$P(y_t \,|\, y_{t-1}, y_{t-2}, \ldots) = P(y_t \,|\, y_{t-1})$$

In words this may be expressed as:

- The distribution of y_t given the history $y_{t-1}, y_{t-2}, \ldots$ only depend on the previous observation y_{t-1}

or more loosely

- The previous respons y_{t-1} contains all available information about y_t.

Stationary time series

Stationary time series is by definition a time series for which any subsequence of length $k+1$ starting at t

$$y_t, y_{t+1}, \cdots, y_{t+k}$$

has the same distribution as another subsequence of length $k+1$ starting at any other time s

$$y_s, y_{s+1}, \cdots, y_{s+k}$$

In particular all the y_t have the same distribution and $\text{Var}(y_t) = \text{Var}(y_s) = \sigma^2$. Thus the autocorrelation function (ACF) becomes

$$\rho(k) = \frac{\gamma(k)}{\sigma^2}$$

ACF for AR(1) processes

It is then not very hard to show that

- $\rho(0) = 1$
- $\rho(1) = a$
- $\rho(2) = a^2$
- $\rho(k) = a^k$

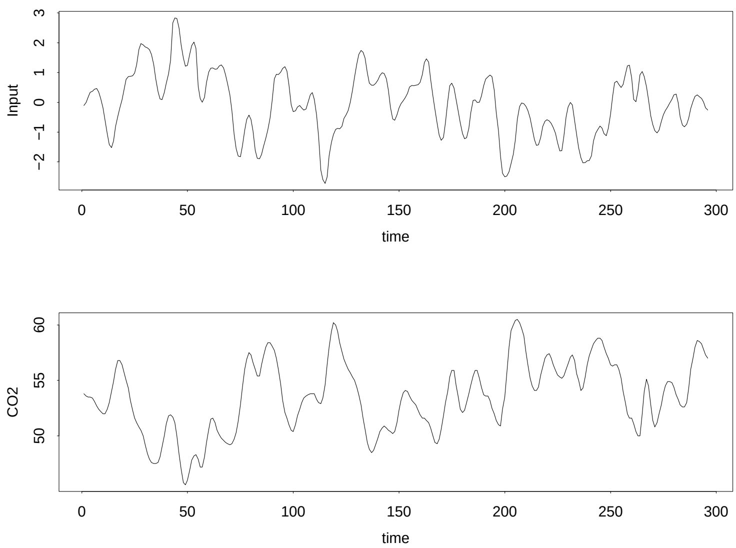
thus the ACF decreases exponentially

Time series may depend on other time series!

Example: Gas furnace data

- $x_t =$ input gas rate at time t

- $y_t =$ output of %CO2 at time t

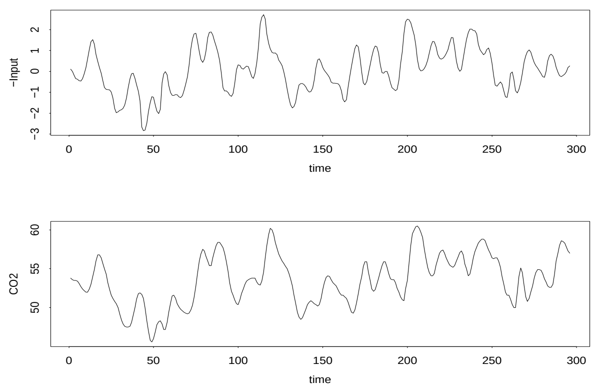


Gas furnace, contd.

The point here is that we have that

$$y_t \approx a + bx_{t-5}$$

with a negative b, that is it depends inversely on the lag-5 value of the x_t series. This can be illustrated by plotting both y_t and $-x_t$:



Now the series have maxima and minima close to each other!

A model for dependent series

Suppose we have one response series y_t and two covariate series $x_{tj}, j = 1, 2$.

The previous example indicates that a useful model may be

$$\begin{aligned} y_t &= a_0 + a_1\, y_{t-1} + a_2\, y_{t-2} + \cdots \\ &\quad + b_0\, x_{1t} + b_1\, x_{1t-1} + b_2\, x_{1t-2} + \cdots \\ &\quad + c_0\, x_{2t} + c_1\, x_{2t-1} + c_2\, x_{2t-2} + \cdots + \varepsilon_t \end{aligned}$$

This model combines

- an AR-process
- dependence on lags of the covariate processes.

This is a regression model, so after reorganizing the data standard software may be used, but special software is likely available.

Cross-correlation function (CCF)

For such data it may be useful to look at the CCF

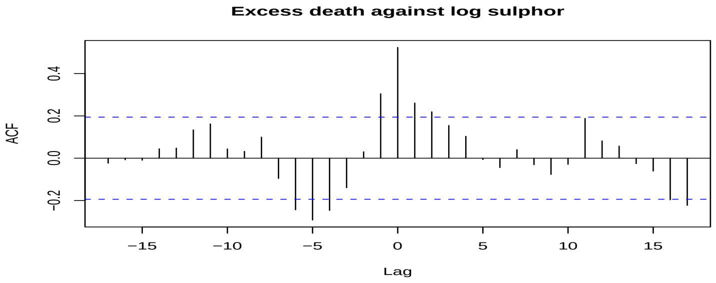
$$\rho_{xy}(u) = \text{corr}(y_{t+u}, x_t)$$

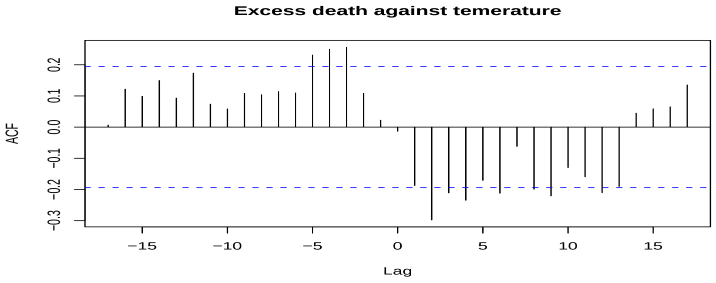
Example I: Excess death data with covariate processes

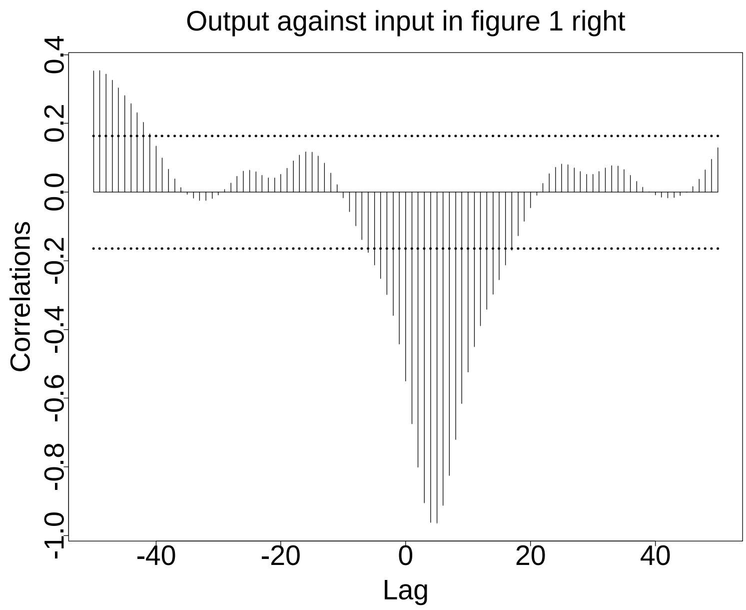
- Temperature
- Smoke

Example II: gasfurnace data

CCF-Examples







Parameter estimates excess death

			Temperature			
	Intercept	log(smoke)	lag 2	lag 0	Unexpl.	
Model	a_0	b_0	c_2	c_0	SS	R^2
1	0.29				63.89	
2	−157.30	25.69			44.78	0.30
3	13.47		−2.62		57.42	0.10
4	−135.53	23.55	−1.72		42.13	0.34
5	−152.48	25.28	−2.60	2.15	38.72	0.39