Kursus 02403 Introduktion til Matematisk Statistik

Forelæsning 8: Simpel lineær regression

Jan Kloppenborg Møller

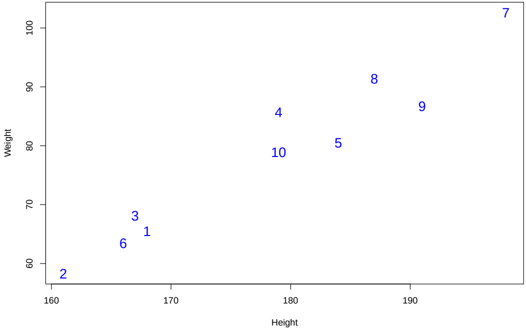
DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 016
Danmarks Tekniske Universitet
2800 Lyngby – Danmark

e-mail: jkmo@dtu.dk

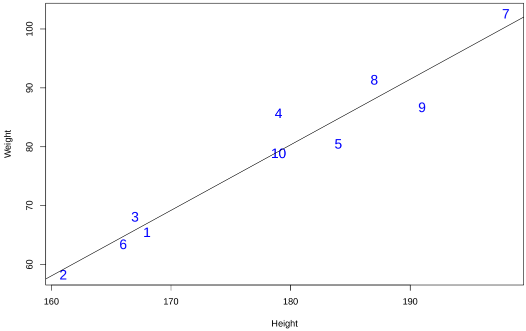
Oversigt

Oversigt

Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9



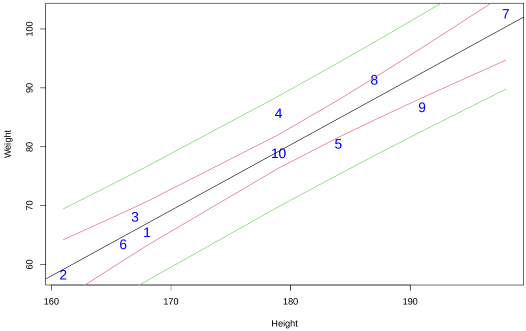
Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9



Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.876 -1.451 -0.608  2.234  6.477
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -119.958     18.897   -6.35  0.00022 ***
## x              1.113      0.106   10.50  5.9e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.88 on 8 degrees of freedom
## Multiple R-squared:  0.932,Adjusted R-squared:  0.924
## F-statistic:  110 on 1 and 8 DF,  p-value: 5.87e-06
```

Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9



Oversigt

Opstil en lineær regressionsmodel

- Opstil den *lineære regressionsmodel*

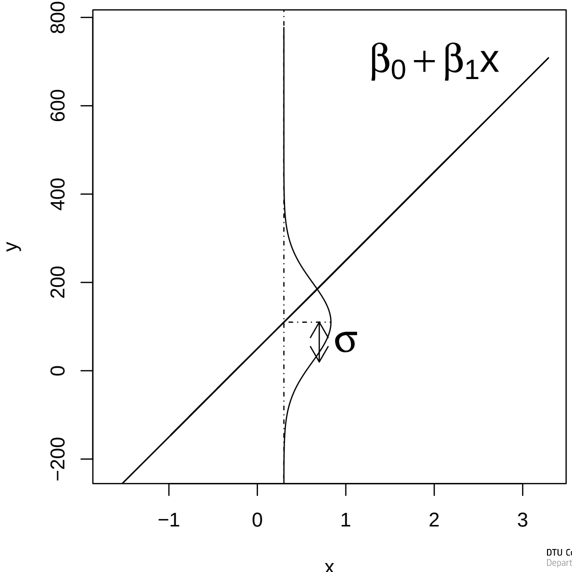
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Y_i er den *afhængige variabel* (dependent variable). En stokastisk variabel.
- x_i er en *forklarende variabel* (explanatory variable)
- ε_i er afvigelsen (error). En stokastisk variabel.

og vi antager

ε_i er independent and identically distributed (i.i.d.) og $N(0, \sigma^2)$

Model-illustration



Oversigt

DTU Compute
Department of Applied Mathematics and Computer Science

Mindste kvadraters metode

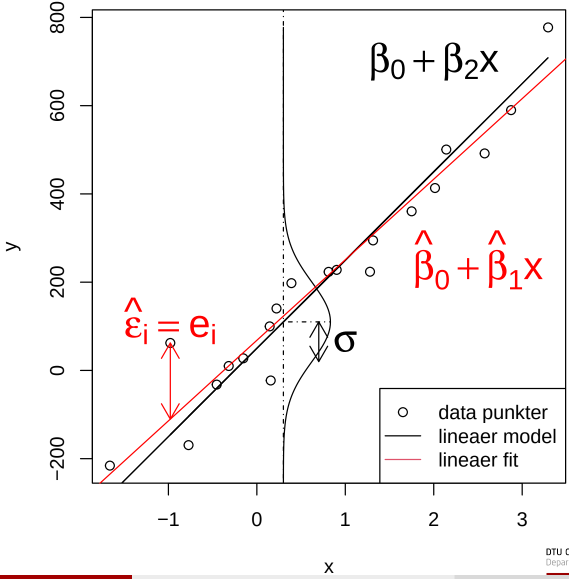
- Minimer variansen σ^2 på afvigelsen. Det er på næsten alle måder det bedste valg i dette setup.
- Formelt: Minimer summen af de kvadrerede afvigelser (Residual Sum of Squares (RSS))

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

$\hat{\beta}_0$ og $\hat{\beta}_1$ minimerer RSS

DTU Compute
Department of Applied Mathematics and Computer Science

Illustration af model, data og fit



DTU Compute
Department of Applied Mathematics and Computer Science

Least squares estimator

Theorem 5.4 (her for estimatorer som i eNoten)

The least squares estimators of β_0 and β_1 are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Least squares estimater

Theorem 5.4 (her for estimater)

The least squares estimatates of β_0 and β_1 are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Oversigt

- Hvordan er parameter estimaterne i en lineær regressionsmodel fordelt (givet normalfordelte afvigelser)?

  De er normalfordelte og deres varians kan estimeres:

Theorem 5.8 (første del)

$$
\begin{aligned}
V[\hat{\beta}_0] &= \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{S_{xx}} \\
V[\hat{\beta}_1] &= \frac{\sigma^2}{S_{xx}} \\
Cov[\hat{\beta}_0, \hat{\beta}_1] &= -\frac{\bar{x}\sigma^2}{S_{xx}}
\end{aligned}
$$

Estimater af standard afvigelserne på $\hat{\beta}_0$ og $\hat{\beta}_1$

Theorem 5.8 (anden del)

Where σ^2 is usually replaced by its estimate ($\hat{\sigma}^2$). The central estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}.$$

When the estimate of σ^2 is used the variances also become estimates and we'll refer to them as $\hat{\sigma}^2_{\beta_0}$ and $\hat{\sigma}^2_{\beta_1}$.

Estimat af standard afvigelserne for $\hat{\beta}_0$ og $\hat{\beta}_1$ (ligningerne (5-73))

$$\hat{\sigma}_{\beta_0} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}; \quad \hat{\sigma}_{\beta_1} = \hat{\sigma}\sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Oversigt

Hypotesetests for parameter estimaterne

- Vi kan altså udføre hypotesetests for parameter estimater i en lineær regressionsmodel:

$$\begin{aligned} H_{0,i}: &\quad \beta_i = \beta_{0,i} \\ H_{1,i}: &\quad \beta_i \neq \beta_{0,i} \end{aligned}$$

- Vi bruger de t-fordelte test størrelser:

Theorem 5.12

Under the null-hypothesis ($\beta_0 = \beta_{0,0}$ and $\beta_1 = \beta_{0,1}$) the statistics

$$T_{\beta_0} = \frac{\hat{\beta}_0 - \beta_{0,0}}{\hat{\sigma}_{\beta_0}}; \quad T_{\beta_1} = \frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}},$$

are t-distributed with $n-2$ degrees of freedom, and inference should be based on this distribution.

Konfidensintervaller for parametrene

Method 5.15

$(1-\alpha)$ confidence intervals for β_0 and β_1 are given by

$$\hat{\beta}_0 \pm t_{1-\alpha/2}\,\hat{\sigma}_{\beta_0}$$
$$\hat{\beta}_1 \pm t_{1-\alpha/2}\,\hat{\sigma}_{\beta_1}$$

where $t_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a t-distribution with $n-2$ degrees of freedom.

- husk at $\hat{\sigma}_{\beta_0}$ og $\hat{\sigma}_{\beta_1}$ findes ved ligningerne (5-74)
- i R kan $\hat{\sigma}_{\beta_0}$ og $\hat{\sigma}_{\beta_1}$ aflæses ved "Std. Error" ved "summary(fit)"

Oversigt

Method 5.18: Konfidensinterval for $\beta_0 + \beta_1 x_0$

- Konfidensinterval for $\beta_0 + \beta_1 x_0$ svarer til et konfidensinterval for linien i punktet x_0
- Beregnes med

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2} \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

- Konfidensintervallet vil i $100(1-\alpha)\%$ af gangene indeholde den rigtige linie, altså $\beta_0 + \beta_1 x_0$

Method 5.18: Prædiktionsinterval for $\beta_0 + \beta_1 x_0 + \varepsilon_0$

- Prædiktionsintervallet (prediction interval) for Y_0 beregnes med en værdi x_0
- Dette gøres *før* Y_0 observeres med

$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\alpha/2} \cdot \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

- Prædiktionsintervallet vil $100(1 - \alpha)$% af gangene indeholde den observerede y_0
- Et prædiktionsinterval bliver altså større end et konfidensinterval for fastholdt α

DTU Compute
Department of Applied Mathematics and Computer Science

Eksempel med konfidensinterval for linien

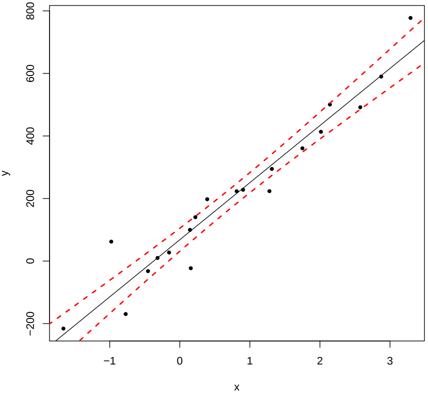
```r
## Eksempel med konfidensinterval for linien

## Lav en sekvens af x værdier
xval <- seq(from=-2, to=6, length.out=100)

## Brug predict funktionen
CI <- predict(fit, newdata=data.frame(x=xval),
interval="confidence",
level=.95)

## Se lige hvad der kom
head(CI)

## Plot data, model og intervaller
plot(x, y, pch=20)
abline(fit)
lines(xval, CI[, "lwr"], lty=2, col="red", lwd=2)
lines(xval, CI[, "upr"], lty=2, col="red", lwd=2)
```



Eksempel med prædiktionsinterval

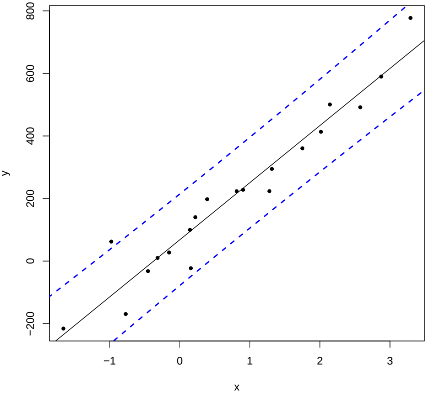
```r
## Eksempel med prædiktionsinterval

## Lav en sekvens a x værdier
xval <- seq(from=-2, to=6, length.out=100)

## Beregn interval for hvert x
PI <- predict(fit, newdata=data.frame(x=xval),
interval="prediction",
level=.95)

## Se lige hvad der kom tilbage
head(PI)

## Plot data, model og intervaller
plot(x, y, pch=20)
abline(fit)
lines(xval, PI[, "lwr"], lty=2, col="blue", lwd=2)
lines(xval, PI[, "upr"], lty=2, col="blue", lwd=2)
```



Oversigt

DTU Compute
Department of Applied Mathematics and Computer Science

Matrix formulering

The simple linear regression problem can be formulated in vector-matrix notation as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \qquad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

or

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}; \qquad \epsilon_i \sim N(0, \sigma^2)$$

RSS in matrix-vector notation

$$RSS = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

Matrix formulering: Parameter estimater

The estimators of the parameters in the simple linear regression model are given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{1}$$

and the covariance matrix of the estimates is

$$V[\hat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{2}$$

and central estimate for the residual variance is

$$\hat{\sigma}^2 = \frac{RSS}{n-2} \tag{3}$$

Oversigt

DTU Compute
Department of Applied Mathematics and Computer Science

Hvad bliver mere skrevet ud af summary?

```
summary(fit)


##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -119.70  -23.74   -4.15   22.44  172.64
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     68.2       17.5     3.9    0.001 **
## x              182.6       11.4    16.0  4.2e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67.4 on 18 degrees of freedom
## Multiple R-squared:  0.935,Adjusted R-squared:  0.931
## F-statistic:  257 on 1 and 18 DF,  p-value: 4.17e-12
```

summary(lm(y∼x)) wrap up

- `Residuals: Min 1Q Median 3Q Max:`
  Residualernes: Minimum, 1. kvartil, Median, 3. kvartil, Maximum

- `Coefficients:`
  `Estimate Std. Error t value Pr(>|t|) "stjerner"`
  Koefficienternes:
  Estimat $\hat{\sigma}_{\beta_i}$ t_{obs} p-værdi
  - Testen er $H_{0,i} : \beta_i = 0$ vs. $H_{1,i} : \beta_i \neq 0$

- `Residual standard error: XXX on XXX degrees of freedom`
  $\varepsilon_i \sim N(0, \sigma^2)$ udskrevet er $\hat{\sigma}$ og ν frihedsgrader (brug til hypotesetesten)

- `Multiple R-squared: XXX`
  Forklaret varians r^2

- Resten bruger vi ikke i det her kursus

Forklaret varians og korrelation

- Forklaret varians af en model er r^2, i summary "Multiple R-squared"
- Beregnes med

$$r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

hvor $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

- Andel af den totale varians der er forklaret med modellen

Forklaret varians og korrelation

- Korrelationen ρ er et mål for *lineær sammenhæng* mellem to stokastiske variable

- Estimeret (i.e. empirisk) korrelation

$$\hat{\rho} = r = \sqrt{r^2}\, sgn(\hat{\beta}_1)$$

hvor $sgn(\hat{\beta}_1)$ er: -1 for $\hat{\beta}_1 \leq 0$ og 1 for $\hat{\beta}_1 > 0$

- Altså:
  - Positiv korrelation ved positiv hældning
  - Negativ korrelation ved negativ hældning

Test for signifikant korrelation

- Test for signifikant korrelation (lineær sammenhæng) mellem to variable

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

er ækvivalent med

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

hvor $\hat{\beta}_1$ er estimatet af hældningen i simpel lineær regressionsmodel

Oversigt

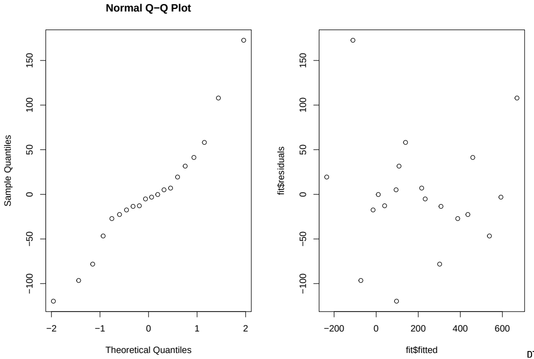
Residual Analysis

Method 5.28

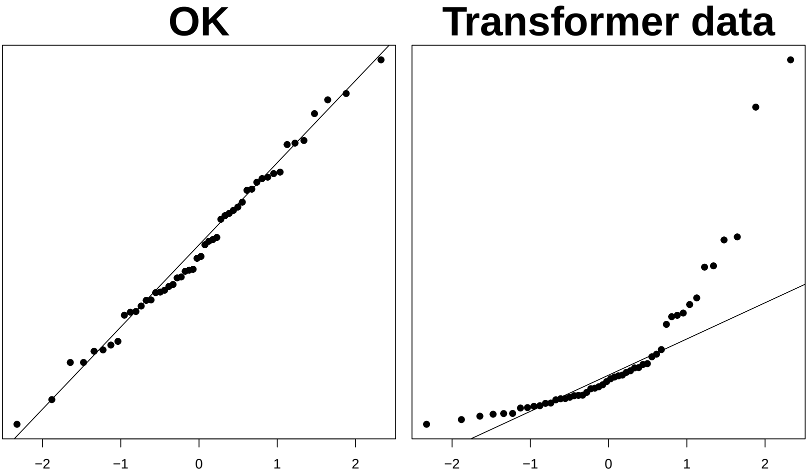
- Check normality assumption with qq-plot.
- Check (non)systematic behavior by plotting the residuals e_i as a function of fitted values $\hat{y}_i$

Residual Analysis in R

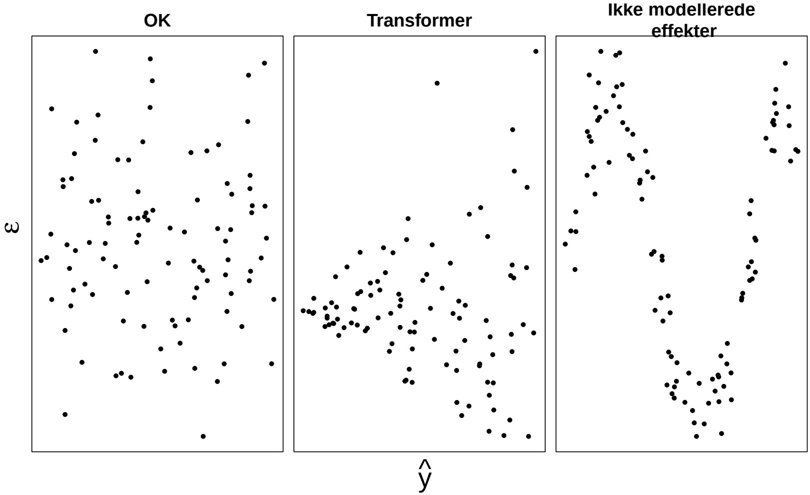
```
fit <- lm(y ~ x)
par(mfrow = c(1, 2))
qqnorm(fit$residuals)
plot(fit$fitted, fit$residuals)
```



Residual Analyse - Normal antagelsen



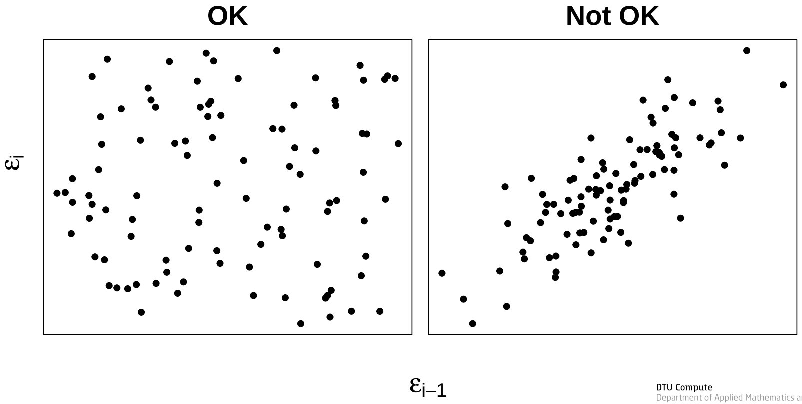
Residual Analyse - Systematiske effekter



Residual Analyse - uafhængighedsantagelsen

For tidsrække data bør uafhængigheds antagelsen også tjekkes, to simple tjek er

- Plot ϵ_i vs. ϵ_{i-1}
- Udregn $cor(\epsilon_i, \epsilon_{i-1})$



Oversigt

Modellering af phytoplankton

Formuler en lineær model for phytoplankton i Skive fjord, estimer modellens parametre og foretag modelkontrol.

Outline