

Kursus 02403

Introduktion til Statistik

Forelæsning 11: Kapitel 11: Regressionsanalyse

Jan Kloppenborg Møller

DTU Informatik

Bygning 305 - rum 219

Danmarks Tekniske Universitet

2800 Lyngby – Danmark

e-mail: jkm@imm.dtu.dk

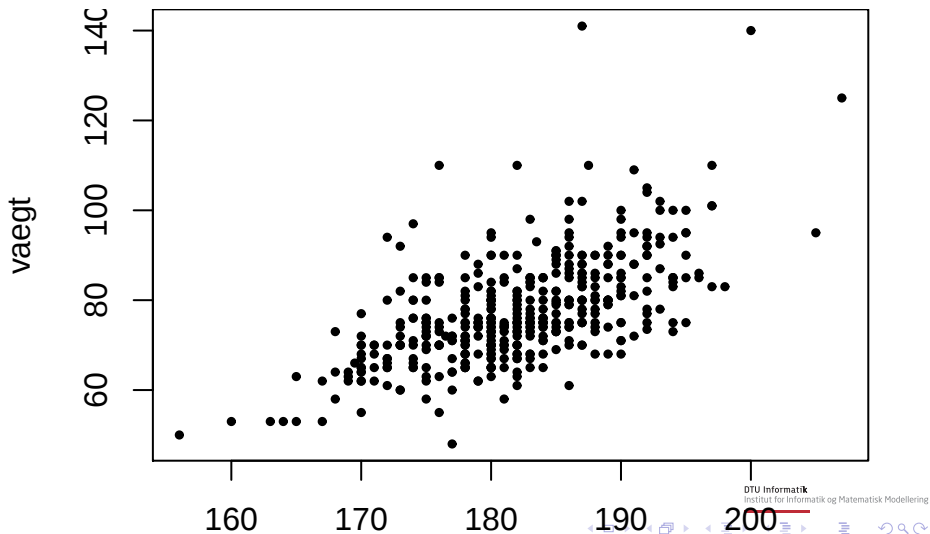
Oversigt

- 1 Eksempel: Højde og vægt
- 2 Korrelation
- 3 Regressionsanalyse (kap 11)
- 4 Mindste kvadraters metode
- 5 Inferens i regressionsmodel
 - Inferens for afskæring og hældning
 - Konfidensinterval for linien
 - Prædiktionsinterval for linien
- 6 Korrelation og regression
- 7 R (R note 10)

Oversigt

- 1 Eksempel: Højde og vægt
- 2 Korrelation
- 3 Regressionsanalyse (kap 11)
- 4 Mindste kvadraters metode
- 5 Inferens i regressionsmodel
 - Inferens for afskæring og hældning
 - Konfidensinterval for linien
 - Prædiktionsinterval for linien
- 6 Korrelation og regression
- 7 R (R note 10)

Højde og vægt af unge danske mænd



Oversigt

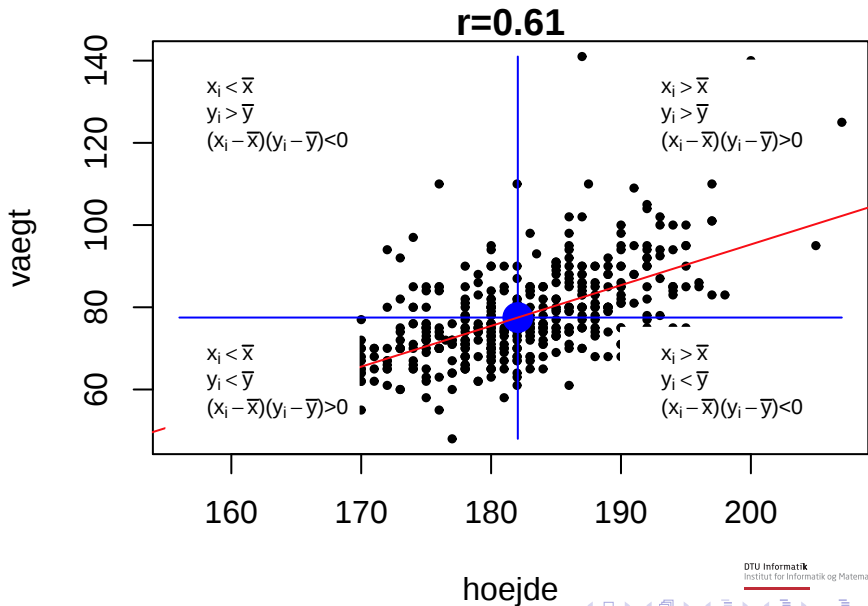
- 1 Eksempel: Højde og vægt
- 2 **Korrelation**
- 3 Regressionsanalyse (kap 11)
- 4 Mindste kvadraters metode
- 5 Inferens i regressionsmodel
 - Inferens for afskæring og hældning
 - Konfidensinterval for linien
 - Prædiktionsinterval for linien
- 6 Korrelation og regression
- 7 R (R note 10)

Korrelation

- Korrelationskoefficienten r angiver den lineære sammenhæng mellem variablene x og y
- Korrelationskoefficienten mellem 2 variable x og y estimeres ved

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- Det antages her, at observationerne (x_i, y_i) er sammenhørende værdier. Der gælder $r \in [-1, 1]$



Oversigt

- 1 Eksempel: Højde og vægt
- 2 Korrelation
- 3 Regressionsanalyse (kap 11)**
- 4 Mindste kvadraters metode
- 5 Inferens i regressionsmodel
 - Inferens for afskæring og hældning
 - Konfidensinterval for linien
 - Prædiktionsinterval for linien
- 6 Korrelation og regression
- 7 R (R note 10)

Regressionsanalyse (kap 11)

- Antag at Y er en stokastisk variabel. Vi er interesseret i at modellere Y 's afhængighed af en *forklarende variabel* x
- Vi undersøger en *lineær* sammenhæng mellem Y og x , dvs. ved en regressionsmodel på formen

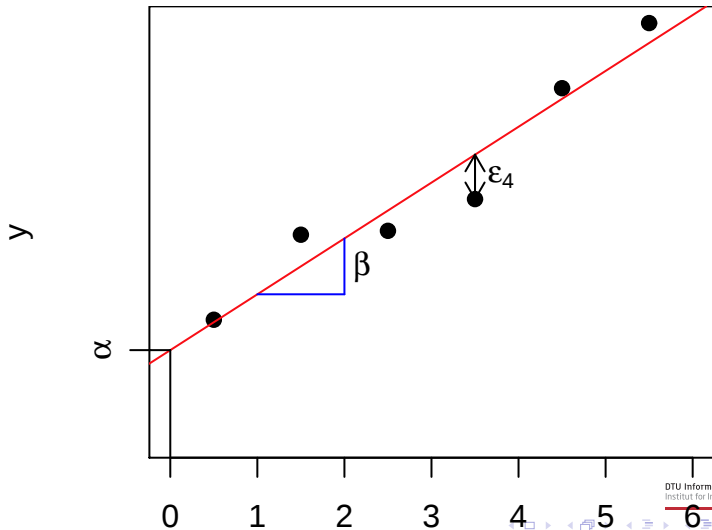
$$Y = \alpha + \beta x + \varepsilon$$

Simpel lineær regressionsmodel

$$Y = \underbrace{\alpha + \beta x}_{\text{model}} + \underbrace{\varepsilon}_{\text{residual}}$$

- Y afhængig variabel
- x uafhængig variabel
- α skæring med Y -akse
- β hældning
- ε residual (tilfældig fejl)

Simpel lineær model



Oversigt

- 1 Eksempel: Højde og vægt
- 2 Korrelation
- 3 Regressionsanalyse (kap 11)
- 4 Mindste kvadraters metode**
- 5 Inferens i regressionsmodel
 - Inferens for afskæring og hældning
 - Konfidensinterval for linien
 - Prædiktionsinterval for linien
- 6 Korrelation og regression
- 7 R (R note 10)

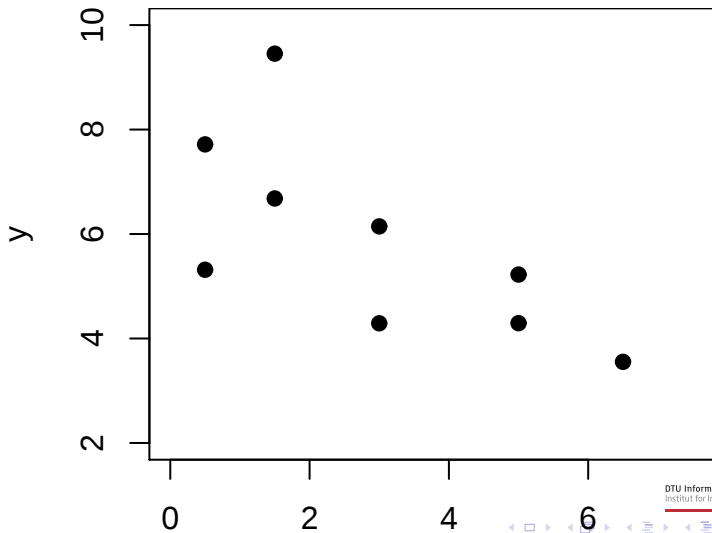
Muslinge biomasser fra Skive fjord (1998)

- Antag at vi har observationerne

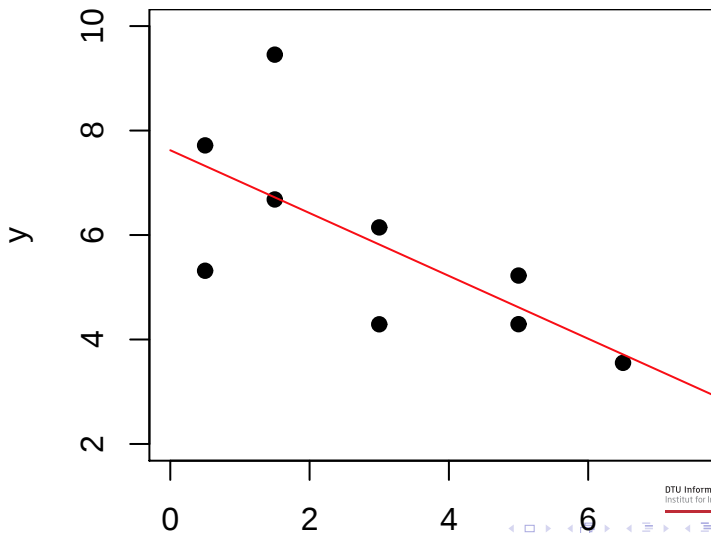
x	0.5	0.5	1.5	1.5	3.0	3.0	5.0	5.0	6.5
y	7.72	5.32	9.45	6.68	6.15	4.29	4.29	5.22	3.55

- x er dybde i meter, y er den fjerde rod af biomasse per kvadratmeter (gr/m^2)
- Er der en sammenhæng mellem x og y ?
- Vi foreslår en model på formen $\hat{y} = a + bx$
- Hvordan beregnes a og b ?

Mindste kvadraters metode



Mindste kvadraters metode



Mindste kvadraters metode

Vi definerer

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

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Mindste kvadraters metode

- a og b bestemmes ved

$$b = \frac{S_{xy}}{S_{xx}}$$

$$a = \bar{y} - b \cdot \bar{x}$$

- a og b er nu de værdier, der giver den regressionslinie, der minimerer den kvadratiske afstand mellem punkter og linie
- a er et *estimat* for α og b er et *estimat* for β

Mindste kvadraters metode

Estimater for α og β :

$$b = \frac{S_{xy}}{S_{xx}} = \frac{-22.42}{37.22} = -0.602$$

$$a = \bar{y} - b \cdot \bar{x} = 5.85 - (-0.602) \cdot 2.94 = 7.626$$

Modellen bliver:

$$\hat{y} = 7.626 - 0.602 \cdot x$$

Oversigt

- 1 Eksempel: Højde og vægt
- 2 Korrelation
- 3 Regressionsanalyse (kap 11)
- 4 Mindste kvadraters metode
- 5 Inferens i regressionsmodel**
 - Inferens for afskæring og hældning
 - Konfidensinterval for linien
 - Prædiktionsinterval for linien
- 6 Korrelation og regression
- 7 R (R note 10)

Inferens i regressionsmodel

Vi antager at de observerede data (Y_i, x_i) kan beskrives ved modellen

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

hvor det antages at ε_i er uafhængige normalfordelte stokastiske variable med middelværdi 0 og konstant varians σ^2

Et estimat af σ^2 bliver

$$s_e^2 = \frac{S_{yy} - (S_{xy})^2 / S_{xx}}{n - 2}$$

Inferens i regressionsmodel

Antag at vi vil teste en hypotese om skæring med y-aksen

$$H_0 : \alpha = a_0$$

$$H_1 : \alpha \neq a_0$$

Teststørrelsen bliver

$$t = \frac{(a - a_0)}{s_e} \sqrt{\frac{nS_{xx}}{S_{xx} + n(\bar{x})^2}}$$

Kritisk værdi findes i t-fordelingen, $t_{\alpha/2}(n - 2)$

Inferens i regressionsmodel

Antag at vi vil teste en hypotese om hældningen β

$$H_0 : \beta = b_0$$

$$H_1 : \beta \neq b_0$$

Teststørrelsen bliver

$$t = \frac{(b - b_0)}{s_e} \sqrt{S_{xx}}$$

Kritisk værdi findes i t-fordelingen, $t_{\alpha/2}(n - 2)$

Konfidensintervaller for α og β

Konfidensinterval for α

$$a \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}$$

Konfidensinterval for β

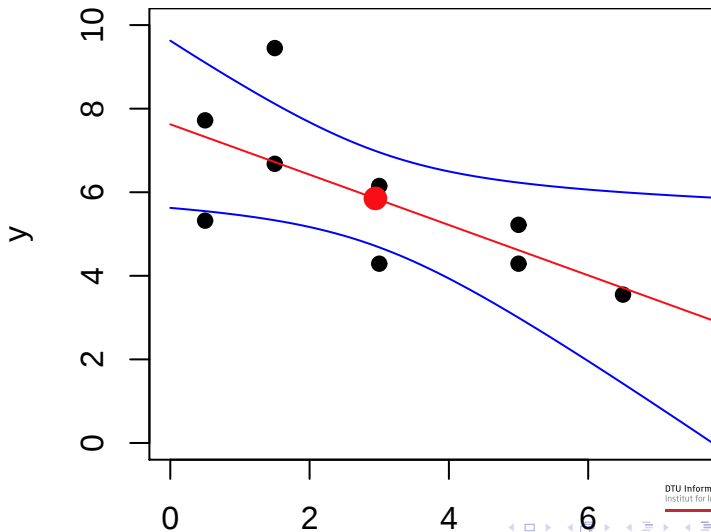
$$b \pm t_{\alpha/2} \cdot s_e \frac{1}{\sqrt{S_{xx}}}$$

Konfidensinterval for $\alpha + \beta x_0$

Konfidensinterval for $\alpha + \beta x_0$ svarer til et konfidensinterval for modellen i punktet x_0

$$(a + bx_0) \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Konfidensinterval



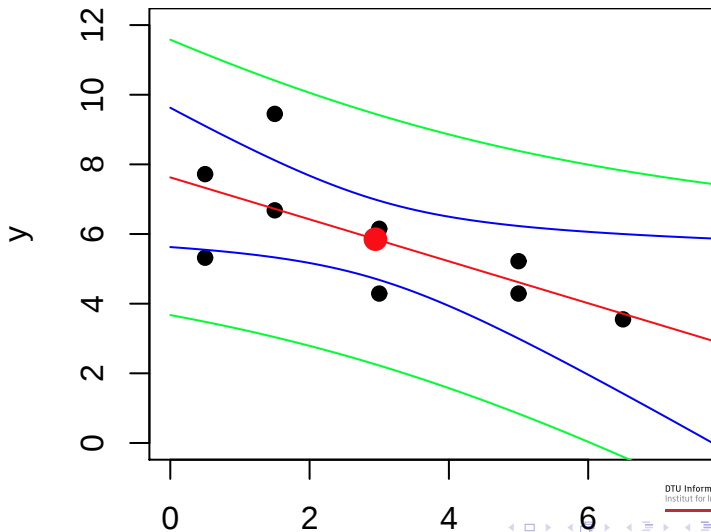
Prædiktionsinterval for $\alpha + \beta x_0$

Prædiktionsinterval for $\alpha + \beta x_0$ svarer til et prædiktionsinterval for modellen i punktet x_0

$$(a + bx_0) \pm t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Et prædiktionsinterval bliver altså større end et konfidensinterval for fastholdt α

Prædiktionsinterval



Oversigt

- 1 Eksempel: Højde og vægt
- 2 Korrelation
- 3 Regressionsanalyse (kap 11)
- 4 Mindste kvadraters metode
- 5 Inferens i regressionsmodel
 - Inferens for afskæring og hældning
 - Konfidensinterval for linien
 - Prædiktionsinterval for linien
- 6 Korrelation og regression
- 7 R (R note 10)

Korrelation og regression

- Korrelation og regression:

$$r = \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}}b, \quad r^2 = \frac{S_{xx}}{S_{yy}}b^2$$

- Korrelationen r udtrykker graden af lineær sammenhæng.
- Korrelationen kvadreret r^2 udtrykker "forklaringsgraden":

S_{yy} = Variation forklaret af linien + Uforklaret variation

$$S_{yy} = \frac{S_{xy}^2}{S_{xx}} + \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right)$$

Inferens for korrelation

- Antager både y og x er tilfældige tal (IKKE kun y)
- r er et estimat for ρ - den underliggende sande lineære sammenhæng mellem y og x .
- Side 340-341 (7ed: 380-381): Formler for hypotesetest og konfidensinterval for korrelationskoefficienten.
- $\rho = 0$ er ensbetydende med $\beta = 0$
- $r = 0$ er ensbetydende med $b = 0$
- Hypotesetest om $\rho = 0$ kan udføres ved at teste $\beta = 0$

Oversigt

- 1 Eksempel: Højde og vægt
- 2 Korrelation
- 3 Regressionsanalyse (kap 11)
- 4 Mindste kvadraters metode
- 5 Inferens i regressionsmodel
 - Inferens for afskæring og hældning
 - Konfidensinterval for linien
 - Prædiktionsinterval for linien
- 6 Korrelation og regression
- 7 R (R note 10)

R (R note 10)

```
> fit.evap <- lm(evap ~ velocity)
> summary(fit.evap)
Call: lm(formula = evap ~ velocity)
Residuals:
    Min       1Q   Median       3Q      Max
-0.201 -0.1467  0.05261  0.1232  0.1747

Coefficients:
            Value Std. Error t value Pr(>|t|)
(Intercept) 0.0692  0.1010     0.6857  0.5123
velocity    0.0038  0.0004     8.7460  0.0000
```

Residual standard error: 0.1591 on 8 degrees of freedom
 Multiple R-Squared: 0.9053
 F-statistic: 76.49 on 1 and 8 degrees of freedom,
 the p-value is 2.286e-05

Oversigt

- 1 Eksempel: Højde og vægt
- 2 Korrelation
- 3 Regressionsanalyse (kap 11)
- 4 Mindste kvadraters metode
- 5 Inferens i regressionsmodel
 - Inferens for afskæring og hældning
 - Konfidensinterval for linien
 - Prædiktionsinterval for linien
- 6 Korrelation og regression
- 7 R (R note 10)