

Kursus 02402/02323 Introducerende Statistik

Forelæsning 8: Simpel lineær regression

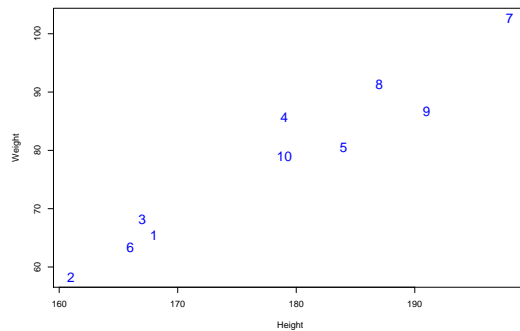
Per Bruun Brockhoff

DTU Compute, Statistik og Dataanalyse
Bygning 324, Rum 220
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: perbb@dtu.dk

DTU Compute
Department of Applied Mathematics and Computer Science

Motiverende eksempel: Højde-vægt

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



DTU Compute
Department of Applied Mathematics and Computer Science

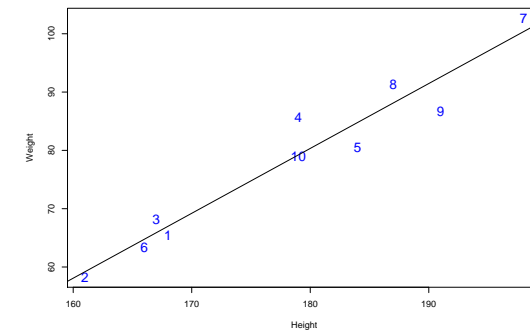
Overzicht

- 1 Motiverende eksempel: Højde-vægt
- 2 Lineær regressionsmodel
- 3 Mindste kvadraters metode (least squares)
- 4 Statistik og lineær regression??
- 5 Hypotesetests og konfidensintervaller for $\hat{\beta}_0$ og $\hat{\beta}_1$
- 6 Konfidensinterval og prædiktionsinterval
 - Konfidensinterval for linien
 - Prædiktionsinterval
- 7 Korrelation
- 8 Residual Analysis: Model control

DTU Compute
Department of Applied Mathematics and Computer Science

Motiverende eksempel: Højde-vægt

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



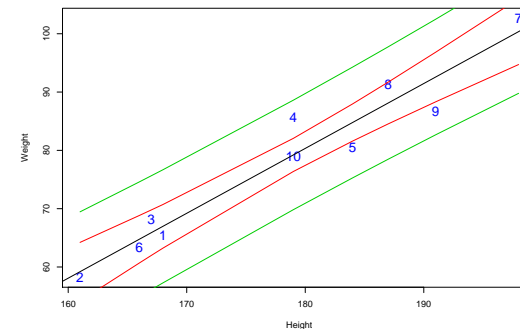
DTU Compute
Department of Applied Mathematics and Computer Science

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

```
summary(lm(y ~ x))

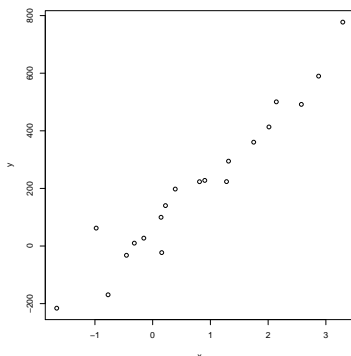
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.876 -1.451 -0.608  2.234  6.477
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -119.958    18.897   -6.35  0.00022 ***
## x              1.113     0.106   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



Et scatter plot af noget data

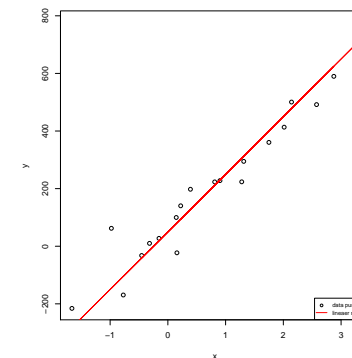
- Vi har n par datapunkter (x_i, y_i)



Opstil en lineær model

- Opstil en lineær model

$$y_i = \beta_0 + \beta_1 x_i$$



men den der mangler noget til at beskrive den *tilfældige variation!*

Opstil en lineær regressionsmodel

- Opstil den *lineære regressionsmodel*

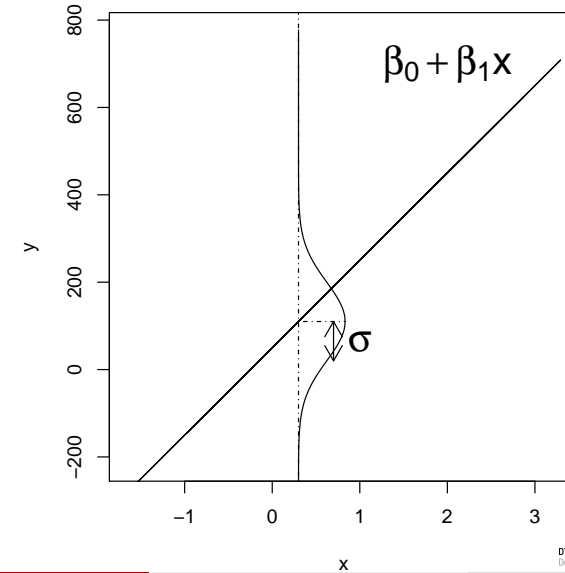
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- 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



Mindste kvadraters metode

- Hvad kan vi gøre for at estimere parametrene β_0 og β_1 ?

God ide: Minimer variansen σ^2 på afvigelsen. Det er på næsten alle måder det bedste valg i dette setup.

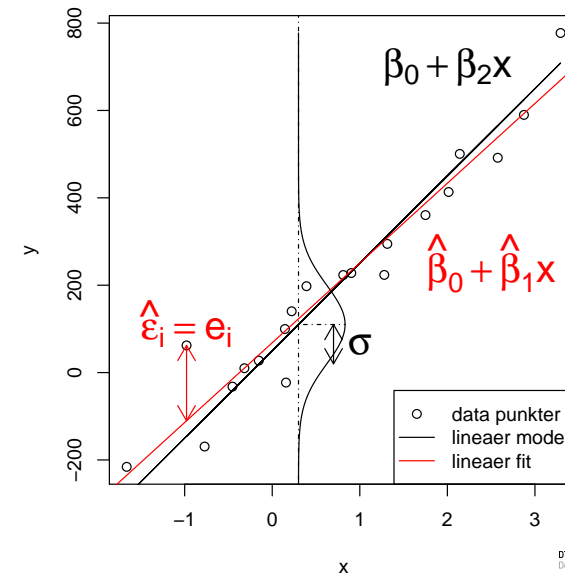
- But how!?

Minimer summen af de kvadrerede afvigelser (Residual Sum of Squares (RSS))

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

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

Illustration af model, data og fit



Least squares estimator

Theorem 5.4 (her for estimatører 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 estimator

Theorem 5.4 (her for estimator)

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

Tænk ikke længere over det for nu!

R eksempel

```
## Simuler en lineær model med normalfordelt afvigelse og estimer parametrene
## FØRST LAV DATA:
## Generer n værdier af input x som uniform fordelt
x <- runif(n=20, min=-2, max=4)

## Simuler lineær regressionsmodel
beta0=50; beta1=200; sigma=90
y <- beta0 + beta1 * x + rnorm(n=length(x), mean=0, sd=sigma)

## HERFRA ligesom virkeligheden, vi har dataen i x og y:
## Et scatter plot af x og y
plot(x, y)

## Udregn least squares estimatorerne, brug Theorem 5.4
(beta1hat <- sum( (y-mean(y))*(x-mean(x)) ) / sum( (x-mean(x))^2 ))
(beta0hat <- mean(y) - beta1hat*mean(x))

## Brug lm() til at udregne estimatorerne
lm(y ~ x)

## Plot den estimerede linie
abline(lm(y ~ x), col="red")
```

Parameter estimatorerne er stokastiske variable

Hvis vi tog en ny stikprøve ville estimatorerne $\hat{\beta}_0$ og $\hat{\beta}_1$ have samme udfald?

Nej, de er stokastiske variable. Tog vi en ny stikprøve så ville vi have en anden realisation.

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

Prøv lige at simulere for at se på det...

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

De er normalfordelte (for $n < 30$ brug t -fordeling) og deres varians kan estimeres:

Theorem 5.7 (første del)

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

- Kovariansen $Cov[\hat{\beta}_0, \hat{\beta}_1]$ (covariance) gør vi ikke mere ud af her.

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

Theorem 5.7 (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 $\hat{\sigma}_{\beta_0}^2$ and $\hat{\sigma}_{\beta_1}^2$.

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

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

Hypotesetests for parameter estimerne

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

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

- Vi bruger de t -fordelte statistikker:

Theorem 5.11

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.

- Se Eksempel 5.12 for eksempel på hypotesetest.
- Test om parametrene er signifikant forskellige fra 0

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

- Se resultatet i R

```
## Hypotesetests om signifikante parametre

## Generer x
x <- runif(n=20, min=-2, max=4)

## Simuler Y
beta0=50; beta1=200; sigma=90
y <- beta0 + beta1 * x + rnorm(n=length(x), mean=0, sd=sigma)

## Brug lm() til at udregne estimerne
fit <- lm(y ~ x)

## Se summary, deri står hvad vi har brug for
summary(fit)
```

Konfidensintervaller for parametrene

Method 5.14

$(1 - \alpha)$ 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)"

```
## Lav konfidensintervaller for parametrene
```

```
## Antal gentagelser
```

```
nRepeat <- 100
```

```
## Fangede vi den rigtige parameter
```

```
TrueValInCI <- logical(nRepeat)
```

```
## Gentag simuleringen og estimeringen nRepeat gange
```

```
for(i in 1:nRepeat){
```

```
  ## Generer x
```

```
  x <- runif(n=20, min=-2, max=4)
```

```
  ## Simuler y
```

```
  beta0=50; beta1=200; sigma=90
```

```
  y <- beta0 + beta1 * x + rnorm(n=length(x), mean=0, sd=sigma)
```

```
  ## Brug lm() til at udregne estimatorne
```

```
  fit <- lm(y ~ x)
```

```
  ## Heldigvis kan R beregne konfidensintervallet (level=1-alpha)
```

```
  (ci <- confint(fit, "(Intercept)", level=0.95))
```

```
  ## Var den rigtige parameterværdi "fanget" af intervallet?
```

```
  (TrueValInCI[i] <- ci[1] < beta0 & beta0 < ci[2])
```

```
}
```

```
sum(TrueValInCI) / nRepeat
```

Method 5.17: 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 5.17: 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

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

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

hvor $\operatorname{sgn}(\hat{\beta}_1)$ 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

```
## Korrelation
## Generer x
x <- runif(n=20, min=-2, max=4)
## Simuler y
beta0=50; beta1=200; sigma=90
y <- beta0 + beta1 * x + rnorm(n=length(x), mean=0, sd=sigma)

## Scatter plot
plot(x,y)

## Brug lm() til at udregne estimaterne
fit <- lm(y ~ x)

## Den rigtige linie
abline(beta0, beta1)
## Plot fittet
abline(fit, col="red")

## Se summary, deri står hvad vi har brug for
summary(fit)

## Korrelation mellem x og y
cor(x,y)

## Kvadreret er den "Multiple R-squared" fra summary(fit)
cor(x,y)^2
```


Residual Analysis

Method 5.26

- 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

Outline

Outline

- 1 Motiverende eksempel: Højde-vægt
- 2 Lineær regressionsmodel
- 3 Mindste kvadraters metode (least squares)
- 4 Statistik og lineær regression??
- 5 Hypotesetests og konfidensintervaller for $\hat{\beta}_0$ og $\hat{\beta}_1$
- 6 Konfidensinterval og prædiktionsinterval
 - Konfidensinterval for linien
 - Prædiktionsinterval
- 7 Korrelation
- 8 Residual Analysis: Model control

Residual Analysis in R

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

