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

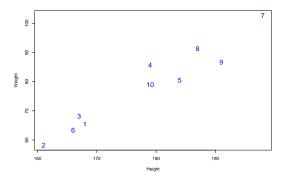
Per Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

Foråret 2015

Motiverende eksempel: Højde-vægt

Per Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

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



Foråret 2015 4 / 43 Oversigt

Motiverende eksempel: Højde-vægt

Lineær regressionsmodel

Mindste kvadraters metode (least squares)

Statistik og lineær regression??

5 Hypotesetests og konfidensintervaller for $\hat{\beta}_0$ og $\hat{\beta}_1$

Konfidensinterval og prædiktionsinterval

Konfidensinterval for linien

Prædiktionsinterval

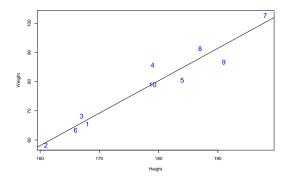
Korrelation

Residual Analysis: Model control

DTU Compute Foråret 2015 er Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8 2 / 43

Motiverende eksempel: Højde-vægt

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



Motiverende eksempel: Højde-vægt

Heights (x_i) | 168 179 198 Weights (y_i) 65.5 80.5 102.6 78.9 58.3 68.1 85.7 63.4 91.4 86.7

```
summary(lm(y ~ x))
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
     Min
             1Q Median
## -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 ***
                1.113
                           0.106 10.50 5.9e-06 ***
## Signif. codes: 0 '***' 0.001 '**' 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
```

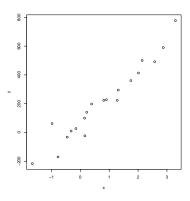
Per Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

Foråret 2015 6 / 43

Lineær regressionsmodel

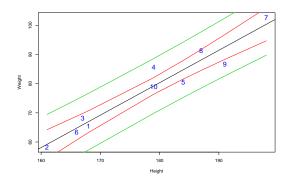
Et scatter plot af noget data

• Vi har n par datapunkter (x_i, y_i)



Motiverende eksempel: Højde-vægt

Heights (x_i) 168 161 167 179 184 166 198 179 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

Per Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

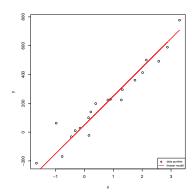
Foråret 2015 7 / 43

Lineær regressionsmodel

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!

Per Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

Foråret 2015

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 $N(0,\sigma^2)$

Foråret 2015 11 / 43

er Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

Mindste kvadraters metode (least squares)

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!?

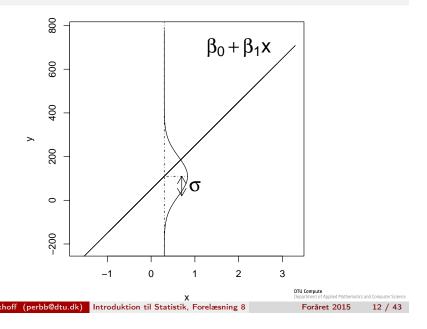
Minimer summen af de kvadrerede afvigelser (Residual Sum of Squares (RSS)

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

 $\hat{\beta_0}$ og $\hat{\beta_1}$ minimerer RSS

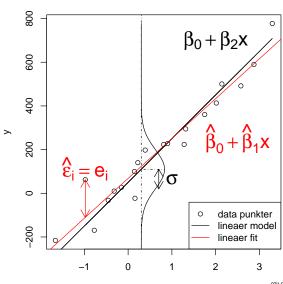
Foråret 2015 14 / 43

Model-illustration



Mindste kvadraters metode (least squares)

Illustration af model, data og fit



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

Foråret 2015 16 / 43

Mindste kvadraters metode (least squares)

Per Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

R eksempel

```
## Simuler en lineær model med normalfordelt afvigelse og estimer parametrene
## Generer n værdier af input x som uniform fordelt
x <- runif(n=20, min=-2, max=4)
## Simuler linear 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 estimaterne, brug Theorem 5.4
(beta1hat \leftarrow sum((y-mean(y))*(x-mean(x))) / sum((x-mean(x))^2))
(beta0hat <- mean(y) - beta1hat*mean(x))
## Brug lm() til at udregne estimaterne
lm(y ~ x)
## Plot den estimerede linie
abline(lm(y ~ x), col="red")
```

Mindste kvadraters metode (least squares)

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

Tænk ikke længere over det for nu!

DTII Compute

Per Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

Foråret 2015

17 / 43

Statistik og lineær regression??

Parameter estimaterne er stokastiske variabler

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

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

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

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

DTII Compute

20 / 43

Foråret 2015 18 / 43

Per Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

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

Theorem 5.7 (første del)

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

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

ruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

Foråret 2015 21 / 43

Hypotesetests og konfidensintervaller for $\hat{\beta}_0$ og $\hat{\beta}_1$

Hypotesetests for parameter estimaterne

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

$$H_{0,i}: \beta_i = \beta_{0,i}$$

$$H_{1,i}: \beta_i \neq \beta_{1,i}$$

• 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.

Estimater 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

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

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))

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

Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

Foråret 2015 22 / 43

Hypotesetests og konfidensintervaller for $\hat{\beta}_0$ og $\hat{\beta}_1$

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

 $H_{0i}: \beta_i = 0$

 $H_{1,i}: \beta_i \neq 0$

Se resultatet i R

Hypotesetests om signifikante parametre x <- runif(n=20, min=-2, max=4) beta0=50; beta1=200; sigma=90 y <- beta0 + beta1 * x + rnorm(n=length(x), mean=0, sd=sigma) ## Brug lm() til at udregne estimaterne fit <- lm(v ~ x) ## Se summary, deri står hvad vi har brug for

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-2degrees 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)"

Foråret 2015 26 / 43

er Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

Konfidensinterval og prædiktionsinterval Konfidensinterval for linien

Method 5.17: Konfidensinterval for $\beta_0 + \beta_1 x_0$

- Konfidensinterval for $\beta_0 + \beta_1 x_0$ swarer 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 $\beta_0 + \beta_1 x_0$

Hypotesetests og konfidensintervaller for $\hat{\beta}_0$ og $\hat{\beta}_1$

```
## Lav konfidensintervaller for parametrene
## Antal gentagelser
nRepeat <- 100
## Fangede vi den rigtige parameter
TrueValInCI <- logical(nRepeat)</pre>
## Gentag simuleringen og estimeringen nRepeat gange
for(i in 1:nRepeat){
  ## Generer x
  x \leftarrow 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 estimaterne
  fit <-lm(y ~x)
  ## Heldiquis kan R bereque konfidensintervallet (level=1-alpha)
  (ci <- confint(fit, "(Intercept)", level=0.95))</pre>
  ## Var den rigtige parameterværdi "fanget" af intervallet?
  (TrueValInCI[i] <- ci[1] < beta0 & beta0 < ci[2])
```

Per Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

Foråret 2015

sum(TrueValInCI) / nRepeat

Konfidensinterval og prædiktionsinterval Prædiktionsinterval

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

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

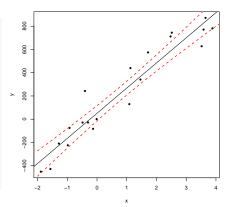
- \bullet Prædiktionsintervallet vil $100(1-\alpha)\%$ af gangene indeholde den observerede y_0
- Et prædiktionsinterval bliver altså større end et konfidensinterval for fastholdt α

Foråret 2015 29 / 43

Per Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

Eksempel med konfidensinterval for linien

```
## Eksempel med konfidensinterval for linien
## Lav en sekvens af x værdier
xval <- seq(from=-2, to=6, length.out=100)</pre>
## Brug predict funktionen
CI <- predict(fit, newdata=data.frame(x=xval),</pre>
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)
```



er Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

Foråret 2015 31 / 43

Korrelation

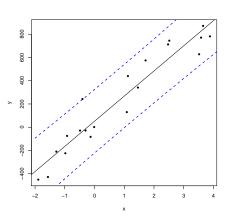
Hvad bliver mere skrevet ud af summary?

```
summary(fit)
##
## Call:
## lm(formula = y ~ x)
## Residuals:
     Min
             1Q Median
                          3Q
## -184.7 -96.4 -20.3 86.6 279.1
## Coefficients:
              Estimate Std. Error t value Pr(>|t|)
                            31.1 1.66 0.12
## (Intercept)
                  51.5
## x
                 216.3
                            15.2 14.22 3.1e-11 ***
## Signif. codes: 0 '***' 0.001 '**' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 126 on 18 degrees of freedom
## Multiple R-squared: 0.918, Adjusted R-squared: 0.914
## F-statistic: 202 on 1 and 18 DF, p-value: 3.14e-11
```

Foråret 2015 34 / 43

Eksempel med prædiktionsinterval

```
## Eksempel med prædiktionsinterval
## Lav en sekvens a x værdier
xval <- seq(from=-2, to=6, length.out=100)</pre>
## 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)
```



er Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

Korrelation

Foråret 2015 32 / 43

DTII Compute

summary($lm(y \sim x)$) wrap up

- Residuals: Min 10 Median 30 Max: Residualernes: Minimum, 1. kvartil, Median, 3. kvartil, Maximum
- Coefficients:

Estimate Std. Error t value Pr(>|t|) "stjerner"

Koefficienternes:

Estimat $\hat{\sigma}_{\beta_i}$ p-værdi

- Testen er $H_{0,i}: \beta_i = 0$ vs. $H_{1,i}: \beta_i \neq 0$
- Stjernerne er sat efter p-værdien
- 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

35 / 43

Foråret 2015

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

er Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

Foråret 2015 36 / 43

Korrelation

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

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 $sqn(\hat{\beta}_1)$ 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

DTII Compute

r Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

Foråret 2015 37 / 43

Korrelation

```
## Korrelation
## Generer x
x <- runif(n=20, min=-2, max=4)
## Simuler 11
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)
```

Foråret 2015 38 / 43

DTII Compute

Residual Analysis: Model control

Residual Analysis

Method 5.26

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

Per Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

Foråret 2015 41 / 43

Outline

Outline

- Statistik og lineær regression??
- Konfidensinterval og prædiktionsinterval
 - Konfidensinterval for linien
 - Prædiktionsinterval
- Korrelation

Per Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

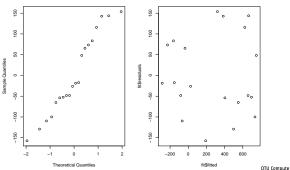
Foråret 2015 43 / 43

Residual Analysis: Model control

Normal Q-Q Plot

Residual Analysis in R

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



Per Bruun Brockhoff (perbb@dtu.dk) Introduktion til Statistik, Forelæsning 8

Foråret 2015 42 / 43