

Course 02402 Introduction to Statistics

Forelæsning 8: Simpel lineær regression

DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Regression

Regression handler om sammenhænge mellem (to) kontinuerte variable
Noget har I allerede lært i gymnasiet, men det er utilstrækkeligt på universitetsniveau.

- Stokastiske variable
- Usikkerheder (herunder konfidensintervaller)
- Hypotesetests
- ...

Overview

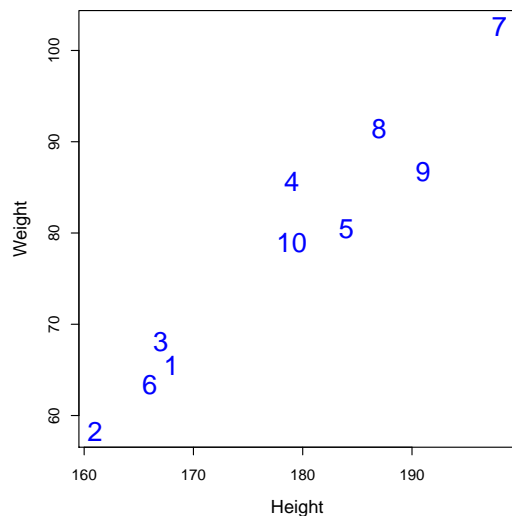
- 1 Eksempel: Højde-vægt
- 2 Lineær regressionsmodel
- 3 Mindste kvadraters metode (least squares)
- 4 Statistik og lineær regression?
- 5 Hypotesetests og konfidensintervaller for β_0 og β_1
- 6 Konfidens- og prædiktionsinterval for regressionslinjen
- 7 Outputtet fra 'summary(lm(y~x))'
- 8 Korrelation
- 9 Residualanalyse: Modelkontrol

Overview

- 1 Eksempel: Højde-vægt
- 2 Lineær regressionsmodel
- 3 Mindste kvadraters metode (least squares)
- 4 Statistik og lineær regression?
- 5 Hypotesetests og konfidensintervaller for β_0 og β_1
- 6 Konfidens- og prædiktionsinterval for regressionslinjen
- 7 Outputtet fra 'summary(lm(y~x))'
- 8 Korrelation
- 9 Residualanalyse: Modelkontrol

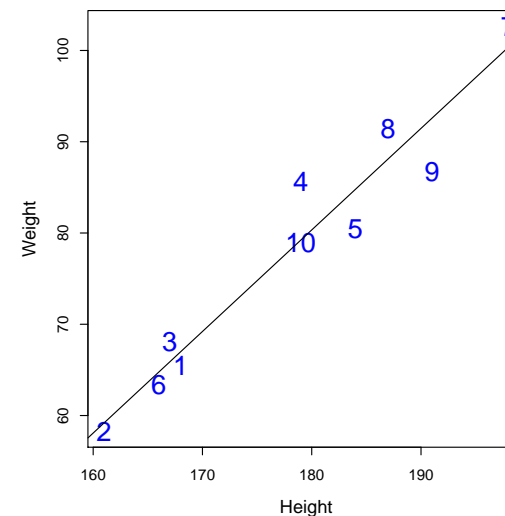
Eksempel: Højde-vægt

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



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

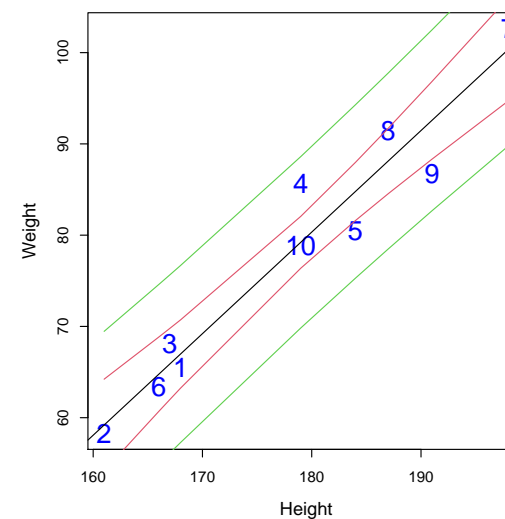


Heights (x_i)	168	161	167	179	184	166	198	187	191	179
Weights (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9

```
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

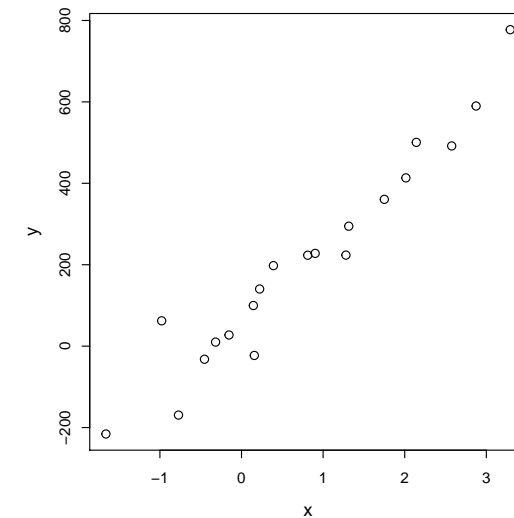


Overview

- 1 Eksempel: Højde-vægt
- 2 **Lineær regressionsmodel**
- 3 Mindste kvadraters metode (least squares)
- 4 Statistik og lineær regression?
- 5 Hypotesetests og konfidensintervaller for β_0 og β_1
- 6 Konfidens- og prædiktionsinterval for regressionslinjen
- 7 Outputtet fra 'summary(lm(y~x))'
- 8 Korrelation
- 9 Residualanalyse: Modelkontrol

Et scatterplot af noget data

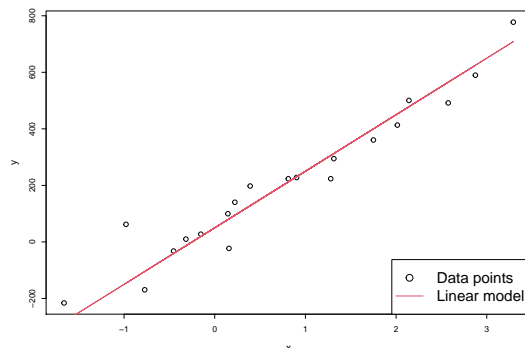
- Vi har n par datapunkter (x_i, y_i) .



Opstil en lineær model

- Opstil en lineær model:

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



- Noget mangler: Beskrivelse af den *tilfældige variation*.

Opstil en lineær regressionsmodel

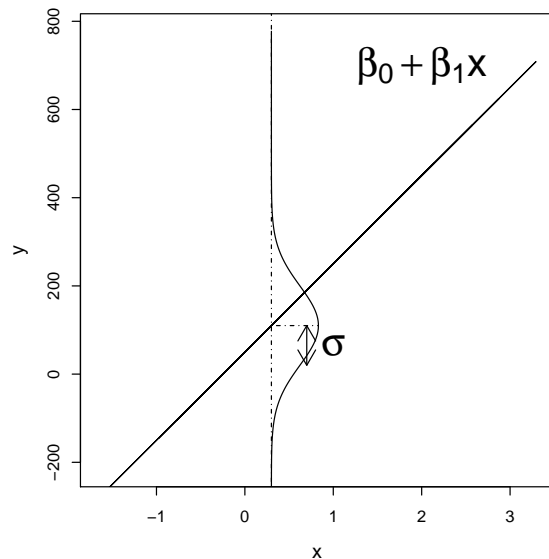
- Opstil den *lineære regressionsmodel*:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n.$$

- Y_i er den *afhængige variabel* (dependent/outcome variable). En stokastisk variabel.
- x_i er en *forklarende variabel*. Deterministiske værdier.
- ε_i er afvigelsen/fejleddet (error). En stokastisk variabel.
- Vi antager at $\varepsilon_i, i = 1, \dots, n$, er *uafhængige og ensfordelte* (i.i.d.), med $\varepsilon_i \sim N(0, \sigma^2)$.

Overvej: Hvilken slags fordeling følger Y_i ? Er Y_i 'erne ensfordelte?

Illustration af statistisk model



Overview

- 1 Eksempel: Højde-vægt
- 2 Lineær regressionsmodel
- 3 Mindste kvadraters metode (least squares)
- 4 Statistik og lineær regression?
- 5 Hypotesetests og konfidensintervaller for β_0 og β_1
- 6 Konfidens- og prædiktionsinterval for regressionslinjen
- 7 Outputtet fra 'summary(lm(y~x))'
- 8 Korrelation
- 9 Residualanalyse: Modelkontrol

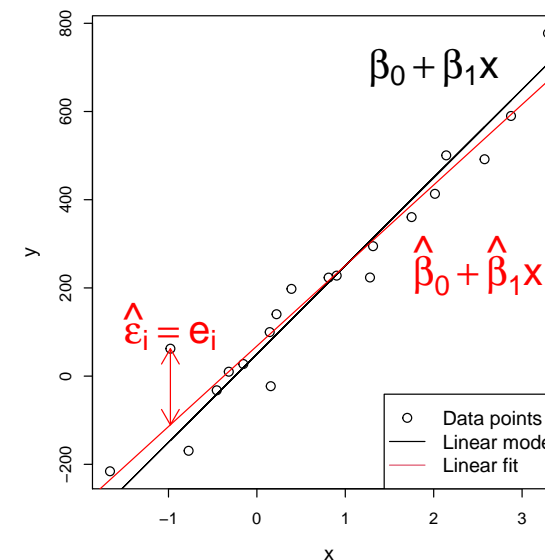
Mindste kvadraters metode

- Hvordan estimerer vi så parametrene β_0 and β_1 ?
- God ide: Minimér variansen σ^2 af residualerne (afvigelsen).
- Men hvordan?
- Minimér 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.

Illustration af model, data og fit



'Least squares'-estimatorer

Theorem 5.4 (her for estimatorer, som i bogen)

'Least squares'-estimatorerne for β_0 og β_1 er givet ved:

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

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

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

'Least squares'-estimerer

Theorem 5.4 (her for *estimerer*)

'Least squares'-estimererne for β_0 og β_1 er givet ved:

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

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

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

Eksempel i R

```
set.seed(100)

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

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

# From here: like for the analysis of 'real data', we have data in x and y:

# Scatter plot of y against x
plot(x, y)

# Find the least squares estimates, use Theorem 5.4
(beta1hat <- sum( (y - mean(y))*(x-mean(x)) ) / sum( (x-mean(x))^2 ))
(beta0hat <- mean(y) - beta1hat*mean(x))

# Use lm() to find the estimates
lm(y ~ x)

# Plot the fitted line
abline(lm(y ~ x), col="red")
```

Overview

- 1 Eksempel: Højde-vægt
- 2 Lineær regressionsmodel
- 3 Mindste kvadraters metode (least squares)
- 4 Statistik og lineær regression?
- 5 Hypotesetests og konfidensintervaller for β_0 og β_1
- 6 Konfidens- og prædiktionsinterval for regressionslinjen
- 7 Outputtet fra 'summary(lm(y~x))'
- 8 Korrelation
- 9 Residualanalyse: Modelkontrol

Parameterestimerne er stokastiske variable

Hvad hvis vi udtrækker en ny stikprøve

Vil estimerne af $\hat{\beta}_0$ og $\hat{\beta}_1$ så blive de samme?

Nej, de er stokastiske variable!

Tager vi en ny stikprøve så får vi en anden realisation.

Hvad er (stikprøve-)fordelingen af parameterestimerne?

...

... i en lineær regressionsmodel med normalfordelte fejldede?

Simulation kan hjælpe os! Lad os se på R...

Fordeling af $\hat{\beta}_0$ og $\hat{\beta}_1$

- $\hat{\beta}_0$ og $\hat{\beta}_1$ er normalfordelte og variansen er givet ved:

Theorem 5.8 (første del)

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

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

Estimer af standardafvigelse på $\hat{\beta}_0$ and $\hat{\beta}_1$

Theorem 5.8 (anden del)

σ^2 erstatter man normalt med dets estimat, $\hat{\sigma}^2$, den *centrale estimator* for σ^2 :

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

Når man bruger estimatet af σ^2 , så bliver varianserne for β_0 og β_1 også estimer. Vi benævner disse $\hat{\sigma}_{\beta_0}^2$ og $\hat{\sigma}_{\beta_1}^2$.

Estimer af standardafvigelsen for $\hat{\beta}_0$ og $\hat{\beta}_1$ (formel 5-43 og 5-44):

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

Overview

- 1 Eksempel: Højde-vægt
- 2 Lineær regressionsmodel
- 3 Mindste kvadraters metode (least squares)
- 4 Statistik og lineær regression?
- 5 Hypotesetests og konfidensintervaller for β_0 og β_1
- 6 Konfidens- og prædiktionsinterval for regressionslinjen
- 7 Outputtet fra 'summary(lm(y~x))'
- 8 Korrelation
- 9 Residualanalyse: Modelkontrol

Hypotesetests for β_0 og β_1

Vi kan udføre hypotesetests for parametrene i en lineær regressionsmodel:

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

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

Theorem 5.12

Under nulhypoteserne ($\beta_0 = \beta_{0,0}$ og $\beta_1 = \beta_{0,1}$) er teststørrelserne

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

t -fordelt med $n - 2$ frihedsgrader.

Hypotesetests for β_0 og β_1

- Se Example 5.13 for et eksempel på et hypotesetest.
- Test om parametrene er signifikant forskellige fra 0:

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

```
# Read data into R
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
y <- c(65.5, 58.3, 68.1, 85.7, 80.5, 63.4, 102.6, 91.4, 86.7, 78.9)

# Fit model to data
fit <- lm(y ~ x)

# Look at model summary to find Tobs-values and p-values
summary(fit)
```

Konfidensintervaller for β_0 og β_1

Method 5.15

$(1 - \alpha)$ konfidensintervaller for β_0 og β_1 er givet ved:

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

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

hvor $t_{1-\alpha/2}$ er $(1 - \alpha/2)$ -fraktilen i t -fordelingen med $n - 2$ frihedsgrader.

- Husk at $\hat{\sigma}_{\beta_0}$ og $\hat{\sigma}_{\beta_1}$ kan findes med ligning 5-43 og 5-44.
- I R kan $\hat{\sigma}_{\beta_0}$ og $\hat{\sigma}_{\beta_1}$ aflæses under "Std. Error" fra `summary(fit)`.

Illustration af konfidensintervaller ved simulation

```
# Number of repetitions (here: CIs)
nRepeat <- 1000

# Empty logical vector of length nRepeat
TrueValInCI <- logical(nRepeat)

# Repeat the simulation and estimation nRepeat times:
for(i in 1:nRepeat){
  # Generate x
  x <- runif(n = 20, min = -2, max = 4)
  # Simulate y
  beta0 = 50; beta1 = 200; sigma = 90
  y <- beta0 + beta1 * x + rnorm(n = length(x), mean = 0, sd = sigma)
  # Use lm() to fit model
  fit <- lm(y ~ x)
  # Use confint() to compute 95% CI for intercept
  ci <- confint(fit, "(Intercept)", level=0.95)
  # Was the 'true' intercept included in the interval? (covered)
  (TrueValInCI[i] <- ci[1] < beta0 & beta0 < ci[2])
}

# How often was the true intercept included in the CI?
sum(TrueValInCI) / nRepeat
```

Overview

- 1 Eksempel: Højde-vægt
- 2 Lineær regressionsmodel
- 3 Mindste kvadraters metode (least squares)
- 4 Statistik og lineær regression?
- 5 Hypotesetests og konfidensintervaller for β_0 og β_1
- 6 Konfidens- og prædiktionsinterval for regressionslinjen
- 7 Outputtet fra 'summary(lm(y~x))'
- 8 Korrelation
- 9 Residualanalyse: Modelkontrol

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

- Konfidensintervallet for $\beta_0 + \beta_1 x_0$ svarer til konfidensintervallet for linjen i punktet x_0 .
- $100(1 - \alpha)\%$ CI kan findes ved:

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

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

- Prædiktionsintervallet for Y_0 beregnes for med et 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 det observerede y_0 .
- For fastholdt α er prædiktionsintervallet større end for konfidensintervallet.

Eksempel med konfidensinterval for linjen

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

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

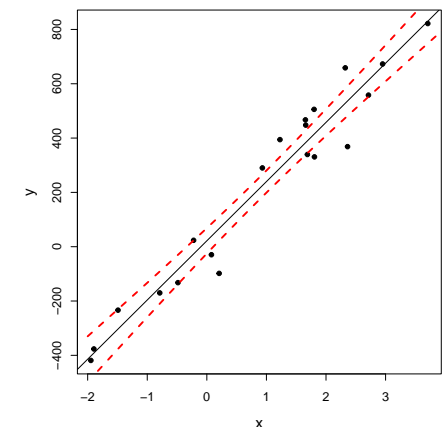
# Use lm() to fit model
fit <- lm(y ~ x)

# Make a sequence of 100 x-values
xval <- seq(from = -2, to = 6, length.out = 100)

# Use the predict function
CI <- predict(fit, newdata = data.frame(x = xval),
             interval = "confidence",
             level = 0.95)

# Check what we got
head(CI)

# Plot the data, model fit and intervals
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 for linjen

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

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

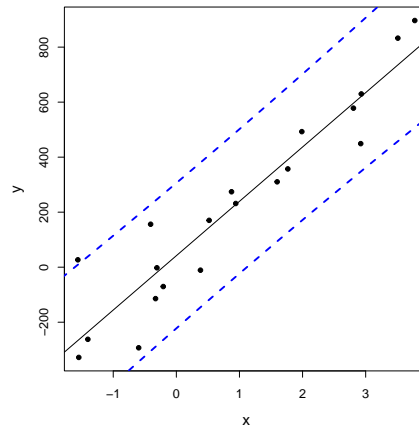
# Use lm() to fit model
fit <- lm(y ~ x)

# Make a sequence of 100 x-values
xval <- seq(from = -2, to = 6, length.out = 100)

# Use the predict function
PI <- predict(fit, newdata = data.frame(x = xval),
             interval = "prediction",
             level = 0.95)

# Check what we got
head(PI)

# Plot the data, model fit and intervals
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)
```



Konfidens- og prædiktionsinterval

- Konfidensintervallet angiver usikkerheden på *regressionslinjen*.
- Prædiktionsintervallet angiver usikkerheden for en *ny observation*.
- Prædiktionsintervallet kan aldrig blive mindre end den tilfældige variation i data (altså den fra fejleddet ε)

Outputtet fra 'summary(lm(y~x))'

Overview

- 1 Eksempel: Højde-vægt
- 2 Lineær regressionsmodel
- 3 Mindste kvadraters metode (least squares)
- 4 Statistik og lineær regression?
- 5 Hypotesetests og konfidensintervaller for β_0 og β_1
- 6 Konfidens- og prædiktionsinterval for regressionslinjen
- 7 **Outputtet fra 'summary(lm(y~x))'**
- 8 Korrelation
- 9 Residualanalyse: Modelkontrol

Outputtet fra 'summary(lm(y~x))'

Hvad (ellers) får vi ud af summary() i R?

```
summary(fit)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -216.86  -66.09   -7.16   58.48  293.37
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    41.8       30.9     1.35   0.19
## x              197.6       16.4    12.05  4.7e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122 on 18 degrees of freedom
## Multiple R-squared:  0.89, Adjusted R-squared:  0.884
## F-statistic: 145 on 1 and 18 DF, p-value: 4.73e-10
```

summary(lm(y~x))

- Residuals: Min 1Q Median 3Q Max
Residualernes: Minimum, 1. kvartil, Median, 3. kvartil, Maximum
- Coefficients:
Estimate Std. Error t value Pr(>|t|) "stars"
Koefficienternes: $\hat{\beta}_i$ $\hat{\sigma}_{\beta_i}$ t_{obs} p -værdi
 - Testet 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
 $\varepsilon_i \sim N(0, \sigma^2)$: outputtet viser $\hat{\sigma}$ og antallet af frihedsgrader ν til hypotesetests, CIs, PIs
- Multiple R-squared: XXX
Forklaret varians r^2 .
- Resten bruger vi ikke i dette kursus.

Overview

- 1 Eksempel: Højde-vægt
- 2 Lineær regressionsmodel
- 3 Mindste kvadraters metode (least squares)
- 4 Statistik og lineær regression?
- 5 Hypotesetests og konfidensintervaller for β_0 og β_1
- 6 Konfidens- og prædiktionsinterval for regressionslinjen
- 7 Outputtet fra 'summary(lm(y~x))'
- 8 **Korrelation**
- 9 Residualanalyse: Modelkontrol

Forklaret varians og korrelation

- Den forklarede varians i en model er r^2 , i summariet "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 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

- Andelen 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.
- Den estimerede (i.e. empiriske) korrelation opfylder

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

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.

Eksempel: Korrelation og R^2 for højde-vægt data

```
# Read data into R
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
y <- c(65.5, 58.3, 68.1, 85.7, 80.5, 63.4, 102.6, 91.4, 86.7, 78.9)

# Fit model to data
fit <- lm(y ~ x)

# Scatter plot of data with fitted line
plot(x,y, xlab = "Height", ylab = "Weight")
abline(fit, col="red")

# See summary
summary(fit)

# Correlation between x and y
cor(x,y)

# Squared correlation is the "Multiple R-squared" from summary(fit)
cor(x,y)^2
```

Overview

- 1 Eksempel: Højde-vægt
- 2 Lineær regressionsmodel
- 3 Mindste kvadraters metode (least squares)
- 4 Statistik og lineær regression?
- 5 Hypotesetests og konfidensintervaller for β_0 og β_1
- 6 Konfidens- og prædiktionsinterval for regressionslinjen
- 7 Outputtet fra 'summary(lm(y~x))'
- 8 Korrelation
- 9 Residualanalyse: Modelkontrol

Residualanalyse

Method 5.28

- Undersøg normalitetsantagelse for residualerne med et qq-plot.
- Undersøg evt. systematiske afvigelser ved at plote residualer, e_i , som en funktion af de fittede værdier \hat{y}_i .

Bemærk!! Det er *residualerne*, IKKE y -værdierne, som indgår i modelkontrol.

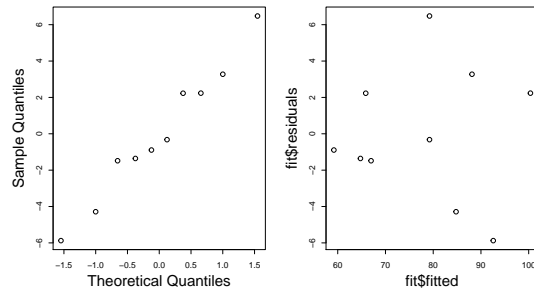
(Method 5.29)

- Er uafhængighedsantagelsen rimelig?

Residualanalyse i R

```
x <- c(168, 161, 167, 179, 184, 166, 198, 187, 191, 179)
y <- c(65.5, 58.3, 68.1, 85.7, 80.5, 63.4, 102.6, 91.4, 86.7, 78.9)
fit <- lm(y ~ x)
```

```
par(mfrow = c(1, 2))
qqnorm(fit$residuals, main = "", cex.lab = 1.5)
plot(fit$fitted, fit$residuals, cex.lab = 1.5)
```



Overview

- 1 Eksempel: Højde-vægt
- 2 Lineær regressionsmodel
- 3 Mindste kvadraters metode (least squares)
- 4 Statistik og lineær regression?
- 5 Hypotesetests og konfidensintervaller for β_0 og β_1
- 6 Konfidens- og prædiktionsinterval for regressionslinjen
- 7 Outputtet fra 'summary(lm(y~x))'
- 8 Korrelation
- 9 Residualanalyse: Modelkontrol