02402 Statistik (Polyteknisk grundlag)

Uge 8: Simpel lineær regression

Nicolai Siim Larsen
DTU Compute
Danmarks Tekniske Universitet
2800 Kgs. Lyngby

Dagsorden

1. Opsummering
2. Eksempel: Højde og vægt
3. Lineære regressionsmodeller
4. Mindste kvadraters metode (Least squares)
5. Statistik og lineær regression
6. Hypotesetest og konfidensintervaller for β_0 og β_1
7. Konfidens- og prædiktionsintervaller
8. Outputtet fra summary
9. Korrelation
10. Modelkontrol - Analyse af residualer

Dagsorden

Opsummering af kursets første halvdel

Sandsynlighedsregning

- Fordelinger som statistiske modeller
- Regneregler til udledninger

Hypotesetest med en eller to stikprøver

- Konfidensintervaller
- Kritiske værdier
- p-værdier

Bootstrapping

- Parametrisk bootstrapping (m. antagelser)
- Ikke-parametrisk bootstrapping (u. antagelser)

Reformulering og modeller

Lad $Y \sim N(\mu, \sigma^2)$. Man kan så opskrive en model med en middelværdi og støj:

$$Y = \mu + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

Estimation af populationsmiddelværdien svarer altså til at estimere en modelparameter.

Stikprøvegennemsnittet $\overline{Y}$ er en middelret (unbiased) estimator for μ:

$$\mathbb{E}\left[\overline{Y}\right] = \mu.$$

Hvis variansen er kendt:

$$\frac{\overline{Y} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1^2).$$

Hvis variansen ikke er kendt (og skal estimeres med S^2):

$$\frac{\overline{Y} - \mu}{\sqrt{S^2/n}} \sim t(n-1).$$

Reformulering og modeller

Lad $X = \mu_1 + \varepsilon_1$, hvor $\varepsilon_1 \sim N(0, \sigma_1^2)$, og lad $Y = \mu_2 + \varepsilon_2$, hvor $\varepsilon_2 \sim N(0, \sigma_2^2)$.

Hypotesetest med en stikprøve

$$\begin{aligned} H_0: &\quad \mu_1 = \mu_0, \\ H_1: &\quad \mu_1 \neq \mu_0. \end{aligned}$$

Hypotesetest med to stikprøver - Ikke parret

$$\begin{aligned} H_0: &\quad \mu_1 - \mu_2 = \delta, \\ H_1: &\quad \mu_1 - \mu_2 \neq \delta. \end{aligned}$$

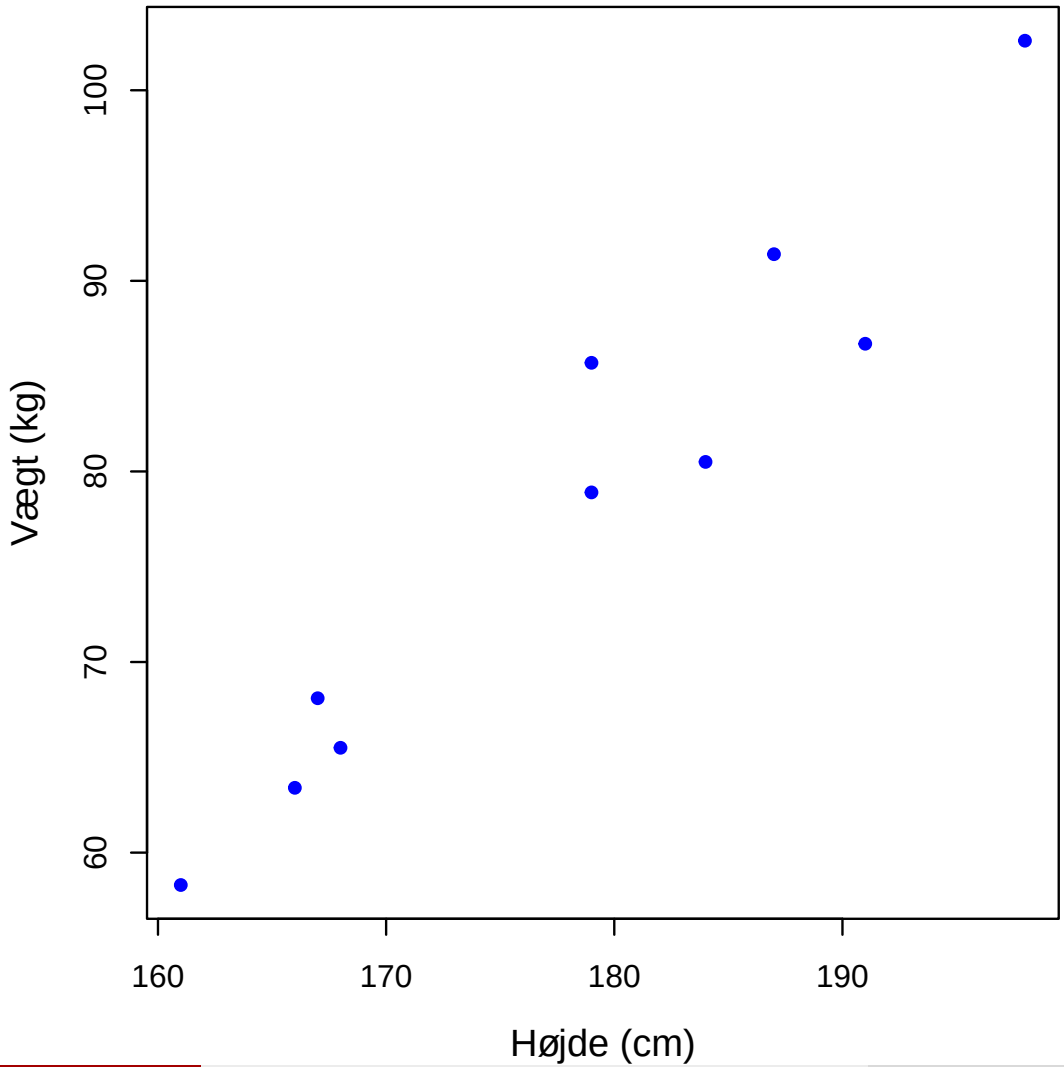
Hypotesetest med to stikprøver - Parret
Ny model: $Z = X - Y = \mu + \varepsilon$, hvor $\varepsilon \sim N(0, \sigma^2)$.

$$\begin{aligned} H_0: &\quad \mu = \mu_1 - \mu_2 = \mu_0, \\ H_1: &\quad \mu = \mu_1 - \mu_2 \neq \mu_0. \end{aligned}$$

Dagsorden

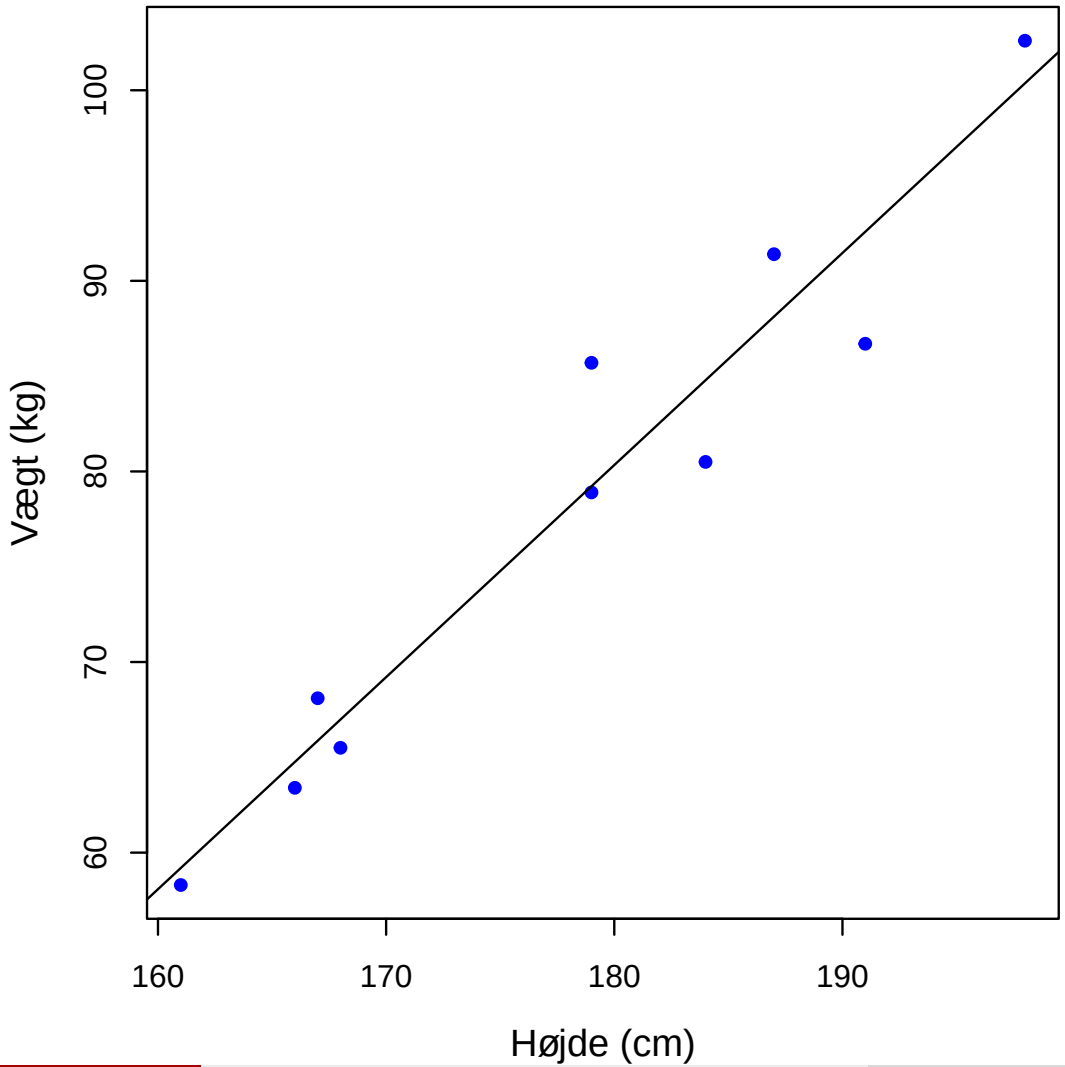
Eksempel: Højde og vægt

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



Eksempel: Højde og vægt

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

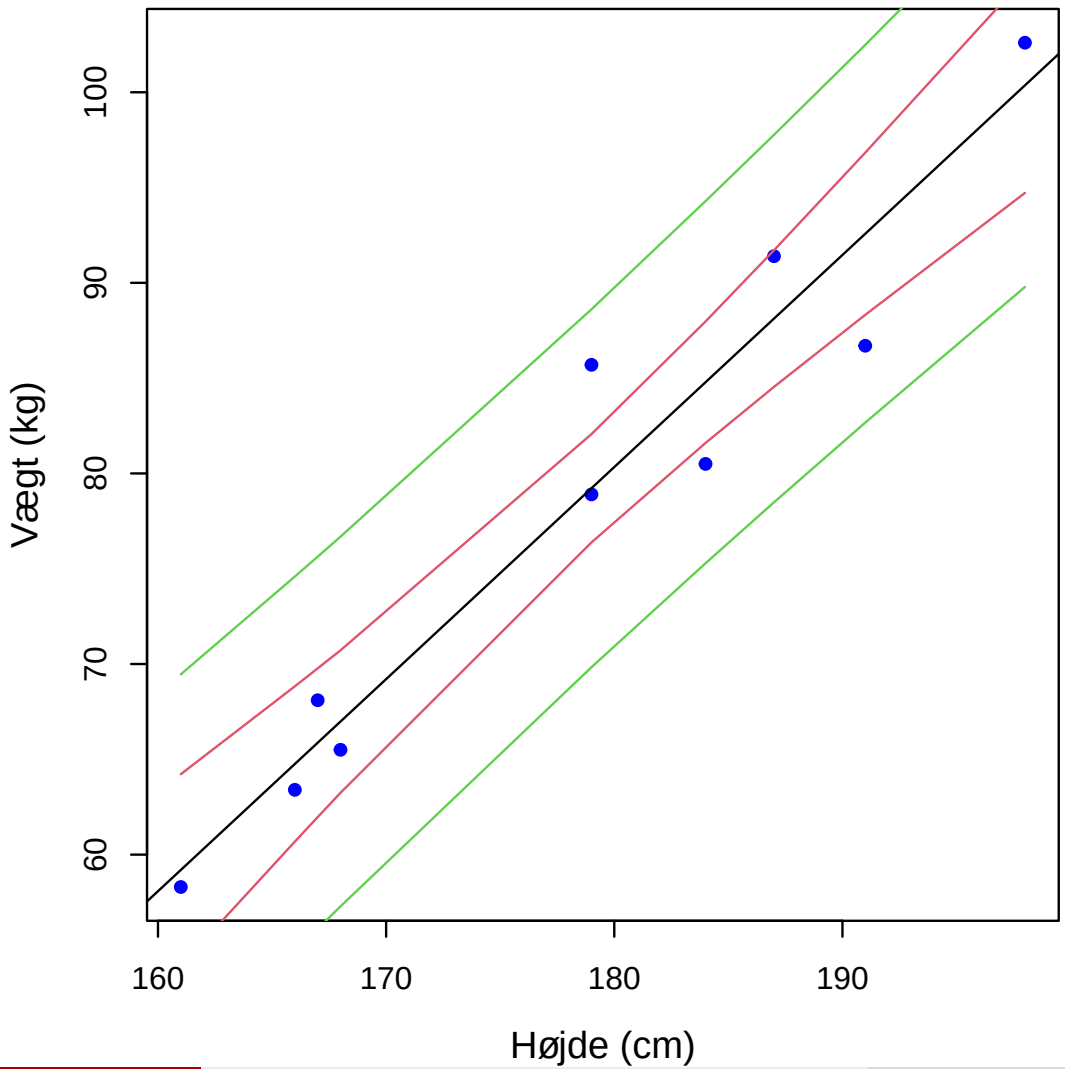


Højde (x_i)	168	161	167	179	184	166	198	187	191	179
Vægt (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9

```
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
```

Eksempel: Højde og vægt

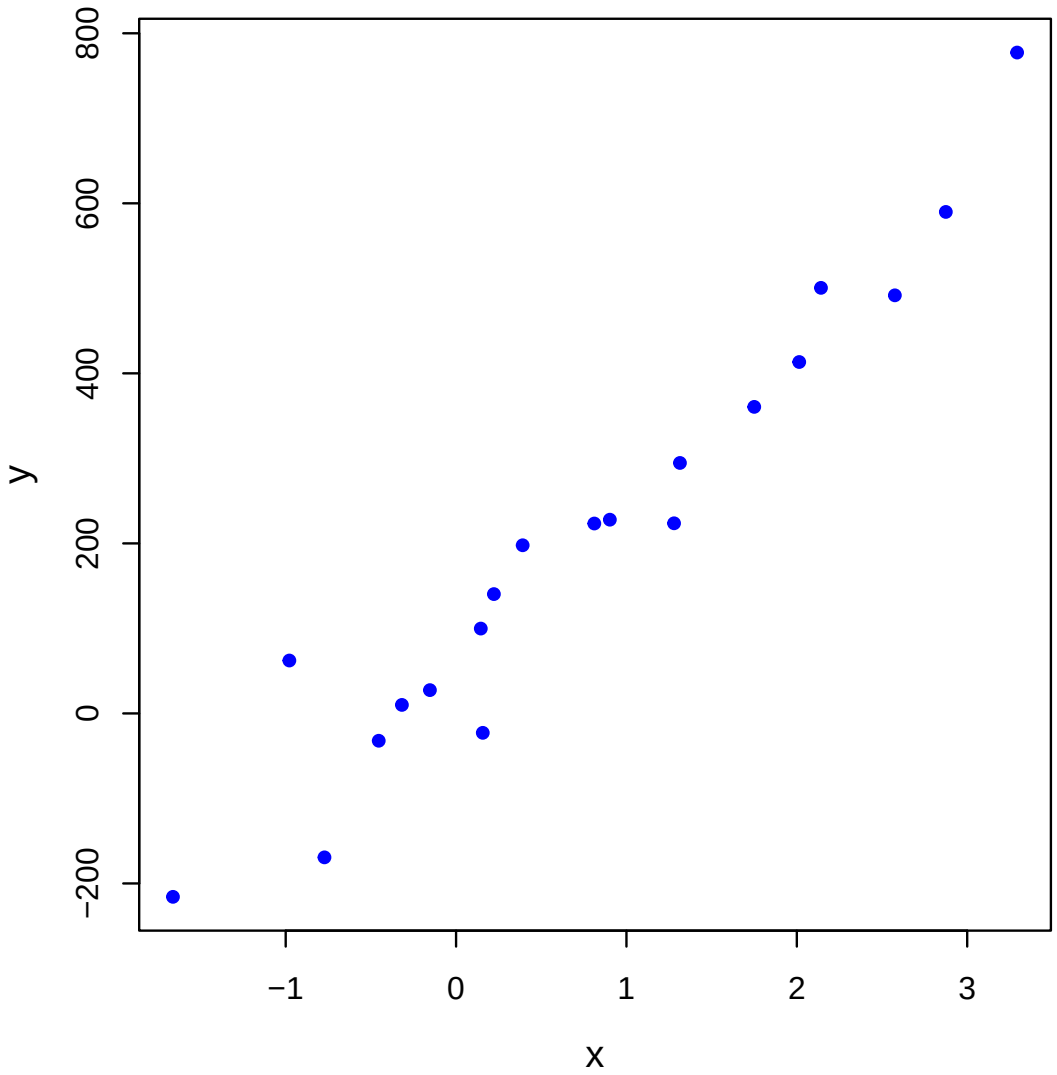
Højde (x_i)	168	161	167	179	184	166	198	187	191	179
Vægt (y_i)	65.5	58.3	68.1	85.7	80.5	63.4	102.6	91.4	86.7	78.9



Dagsorden

Et scatterplot

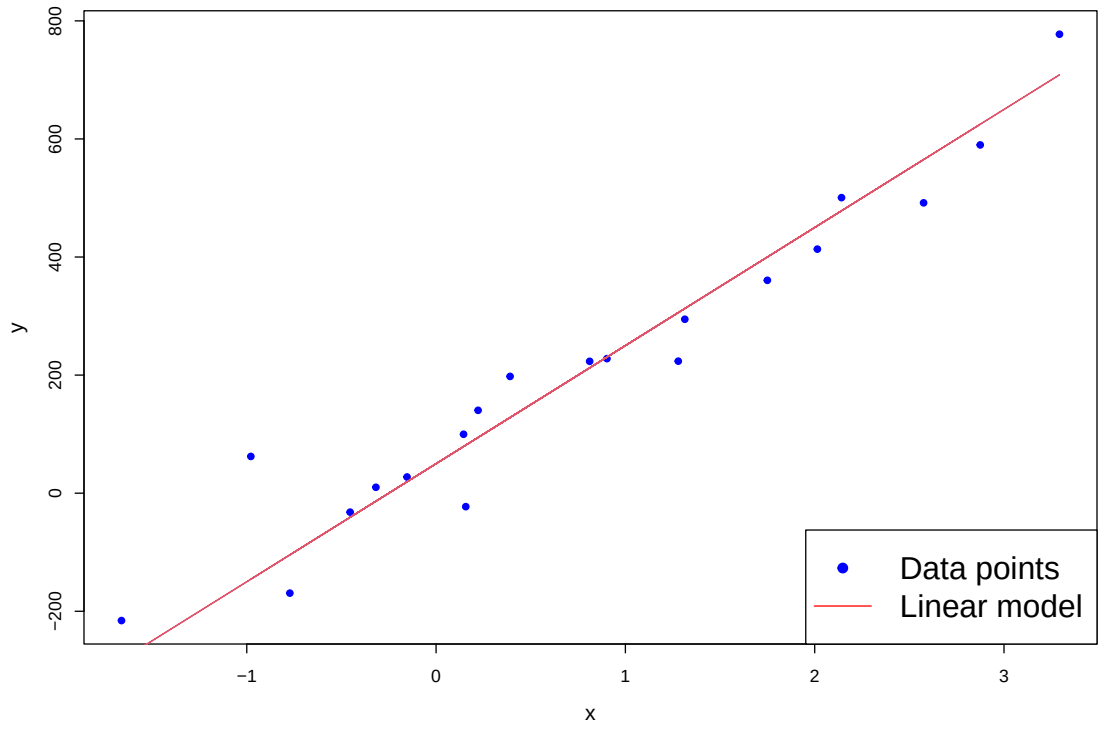
- Vi har n par datapunkter (x_i, y_i).



En lineær model

Hvis datapunkterne ligger på en linje, kan sammenhængen mellem x- og y-værdierne beskrives ved ligningen:

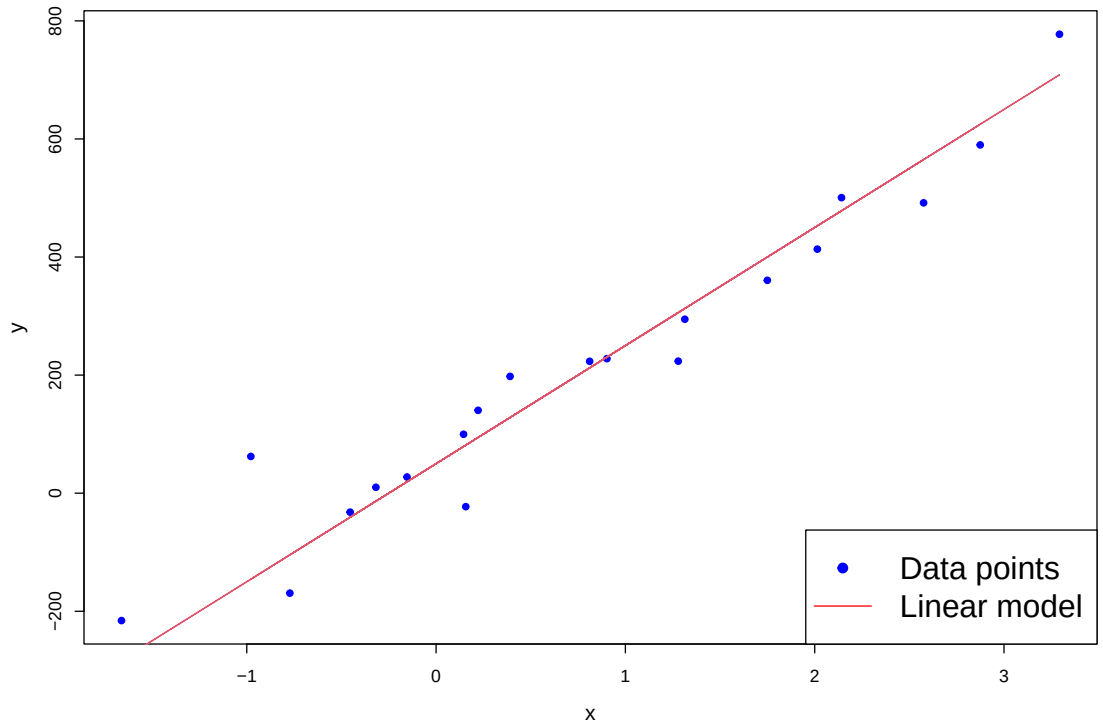
$$y_i = \beta_0 + \beta_1 x_i.$$



En lineær model

Hvis datapunkterne ligger på en linje, kan sammenhængen mellem x- og y-værdierne beskrives ved ligningen:

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



- Vi mangler en beskrivelse af den *tilfældige variation*.

Den lineære regressionsmodel

- Den *lineære regressionsmodel*:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, \ldots, n).$$

- Y_i er en *afhængige variabel* (dependent/outcome variable) - En stokastisk variabel.
- x_i er en *forklarende variabel* (independent/predictor/regressor/explanatory variable, covariate) - En deterministisk værdi.
- ε_i er en støj/afvigelse/fejl (error) - En stokastisk variabel.
- Vi antager, at fejlene ε_i $(i = 1, \ldots, n)$ er *uafhængige og ensfordelte* (i.i.d.) med $\varepsilon_i \sim N(0, \sigma^2)$.

Den lineære regressionsmodel

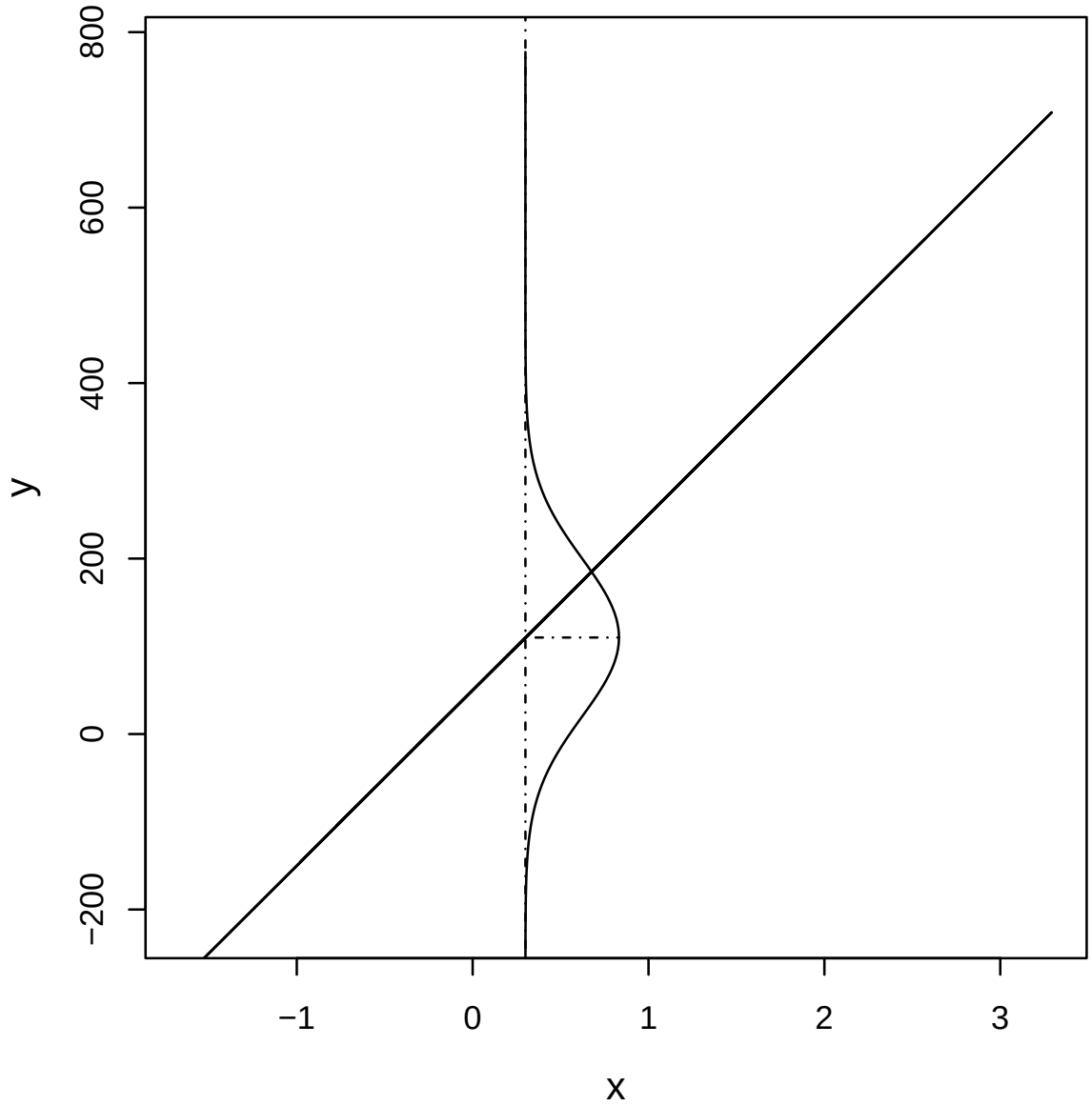
- Den *lineære regressionsmodel*:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, \ldots, n).$$

- Y_i er en *afhængige variabel* (dependent/outcome variable) - En stokastisk variabel.
- x_i er en *forklarende variabel* (independent/predictor/regressor/explanatory variable, covariate) - En deterministisk værdi.
- ε_i er en støj/afvigelse/fejl (error) - En stokastisk variabel.
- Vi antager, at fejlene ε_i $(i = 1, \ldots, n)$ 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 den statistiske model



Dagsorden

Mindste kvadraters metode

- Vi ønsker at estimere parametrene β_0 og β_1.

Mindste kvadraters metode

- Vi ønsker at estimere parametrene β_0 og β_1.
- God ide: Lad os minimere variansen af residualerne/afvigelsen (σ^2).

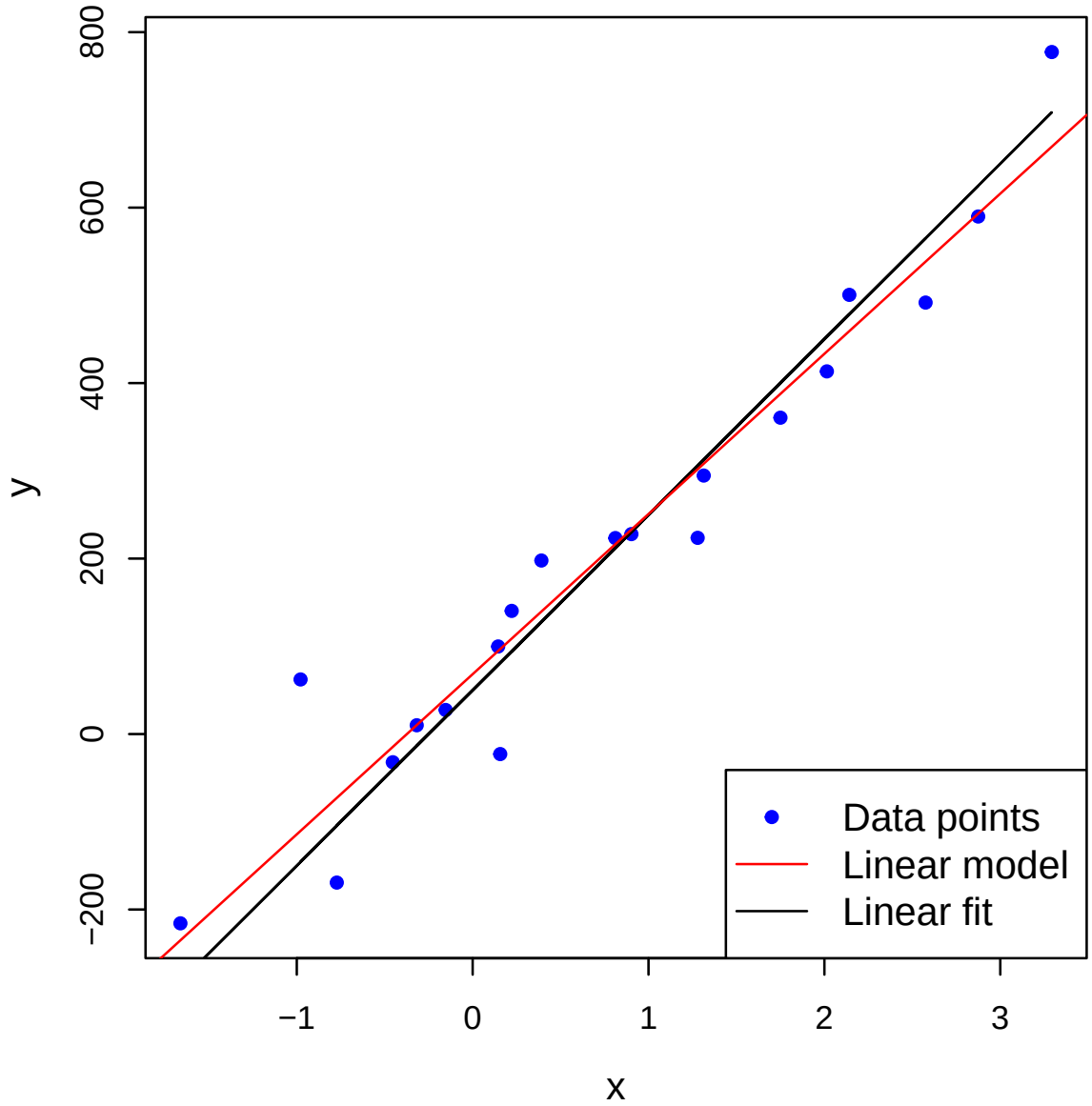
Mindste kvadraters metode

- Vi ønsker at estimere parametrene β_0 og β_1.
- God ide: Lad os minimere variansen af residualerne/afvigelsen (σ^2).
- Vi minimerer summen af de kvadrerede afvigelser (Residual Sum of Squares, RSS):

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

Dvs. at vi vælger $\hat{\beta}_0$ og $\hat{\beta}_1$, sådan at de minimerer RSS.

Illustration af model, data og fit



'Least squares'-estimatorer

Sætning 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'-estimater

Sætning 5.4 (her for *estimater*)

'Least squares'-estimaterne 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

```r
set.seed(100)

# Simuler værdier for x
x <- runif(n = 20, min = -2, max = 4)

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

# Scatter plot af y mod x
plot(x, y)

# Find LS estimater (Sætning 5.4)
(beta1hat <- sum( (y - mean(y))*(x-mean(x)) ) / sum( (x-mean(x))^2 ))
(bet0hat <- mean(y) - beta1hat*mean(x))

# Brug lm() til at finde LS estimater
lm(y ~ x)

# Plot den bedste rette linje
abline(lm(y ~ x), col="red")
```

Dagsorden

Variation i parameterestimaterne

Vi udtager en ny stikprøve

Vil estimaterne af $\hat{\beta}_0$ and $\hat{\beta}_1$ så blive de samme?

Variation i parameterestimaterne

Vi udtager en ny stikprøve

Vil estimaterne af $\hat{\beta}_0$ and $\hat{\beta}_1$ så blive de samme?

Nej - Der er en variation!

En ny stikprøve giver anledning til nye realiseringer af estimatorerne, dvs. nye estimater.

Variation i parameterestimaterne

Vi udtager en ny stikprøve

Vil estimaterne af $\hat{\beta}_0$ and $\hat{\beta}_1$ så blive de samme?

Nej - Der er en variation!

En ny stikprøve giver anledning til nye realiseringer af estimatorerne, dvs. nye estimater.

Hvad er fordelingerne af parameterestimatorerne?

Vi skal kende dem for at lave hypotesetest mv.

Variation i parameterestimaterne

Vi udtager en ny stikprøve

Vil estimaterne af $\hat{\beta}_0$ and $\hat{\beta}_1$ så blive de samme?

Nej - Der er en variation!

En ny stikprøve giver anledning til nye realiseringer af estimatorerne, dvs. nye estimater.

Hvad er fordelingerne af parameterestimatorerne?

Vi skal kende dem for at lave hypotesetest mv.

Simulation kan hjælpe os! R kan give os en intuitiv ide!

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

Estimatorerne $\hat{\beta}_0$ og $\hat{\beta}_1$ er normalfordelte med varianserne:

Sætning 5.8 (første del)

$$V[\hat{\beta}_0] = \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{S_{xx}},$$

$$V[\hat{\beta}_1] = \frac{\sigma^2}{S_{xx}},$$

$$Cov[\hat{\beta}_0, \hat{\beta}_1] = -\frac{\bar{x}\sigma^2}{S_{xx}}.$$

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

Standardafvigelserne for $\hat{\beta}_0$ og $\hat{\beta}_1$

Sætning 5.8 (anden del)

Da σ^2 er ukendt, benyttes det *centrale estimat for* σ^2:

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

Vi estimerer altså variansen (standardafvigelsen) for fejlen og derved også varianserne (standardafvigelserne) for estimatorerne. Vi benævner disse $\hat{\sigma}^2_{\beta_0}$ og $\hat{\sigma}^2_{\beta_1}$.

Standardafvigelserne for $\hat{\beta}_0$ og $\hat{\beta}_1$

Sætning 5.8 (anden del)

Da σ^2 er ukendt, benyttes det *centrale estimat for* σ^2:

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

Vi estimerer altså variansen (standardafvigelsen) for fejlen og derved også varianserne (standardafvigelserne) for estimatorerne. Vi benævner disse $\hat{\sigma}^2_{\beta_0}$ og $\hat{\sigma}^2_{\beta_1}$.

Man får følgende estimater af standardafvigelserne for $\hat{\beta}_0$ og $\hat{\beta}_1$:

$$\hat{\sigma}_{\beta_0} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}, \quad \hat{\sigma}_{\beta_1} = \hat{\sigma}\sqrt{\frac{1}{S_{xx}}}.$$

Dagsorden

Hypotesetest for β_0 og β_1

Vi kan udføre hypotesetest for parametrene i en lineær regressionsmodel:

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

Hypotesetest for β_0 og β_1

Vi kan udføre hypotesetest for parametrene i en lineær regressionsmodel:

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

Sætning 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-fordelte med $n-2$ frihedsgrader.

Hypotetest for β_0 og β_1

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

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

```
# Indlæs data

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 til data
fit <- lm(y ~ x)

# Find teststørrelser og p-værdier
summary(fit)
```

Konfidensintervaller for β_0 og β_1

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

Dagsorden

Metode 5.18: Konfidensinterval for regressionslinjen

En simpel lineær regressionsmodel kan skrives som

$$Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

eller

$$Y \sim N(\mu(x), \sigma^2),$$

hvor $\mu(x) = \beta_0 + \beta_1 x$.
For $x = x_0$ kan vi derfor finde et konfidensinterval for $\mu(x_0) = \beta_0 + \beta_1 x_0$.

$(1-\alpha)$-konfidensintervallet for regressionslinjen i $x = x_0$ (for $\mu(x_0)$) kan findes ved:

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

Metode 5.18: Prædiktionsinterval for en ny observation

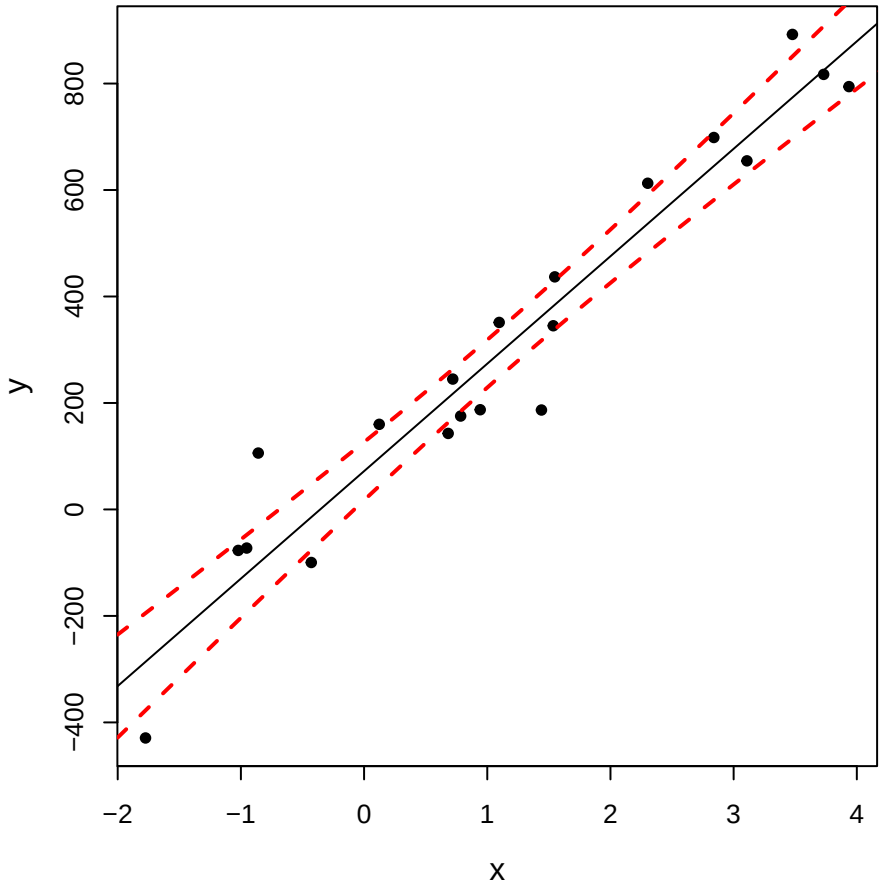
- Vi ønsker et prædiktionsinterval for en ny observation Y_0 for $x = x_0$.
- $(1-\alpha)$-prædiktionsintervallet for en ny observation Y_0 for $x = x_0$ kan findes ved:

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

- Prædiktionsintervallet vil $100(1-\alpha)\%$ af gangene indeholde det observerede y_0.
- For fastholdt α er prædiktionsintervallet større end konfidensintervallet.

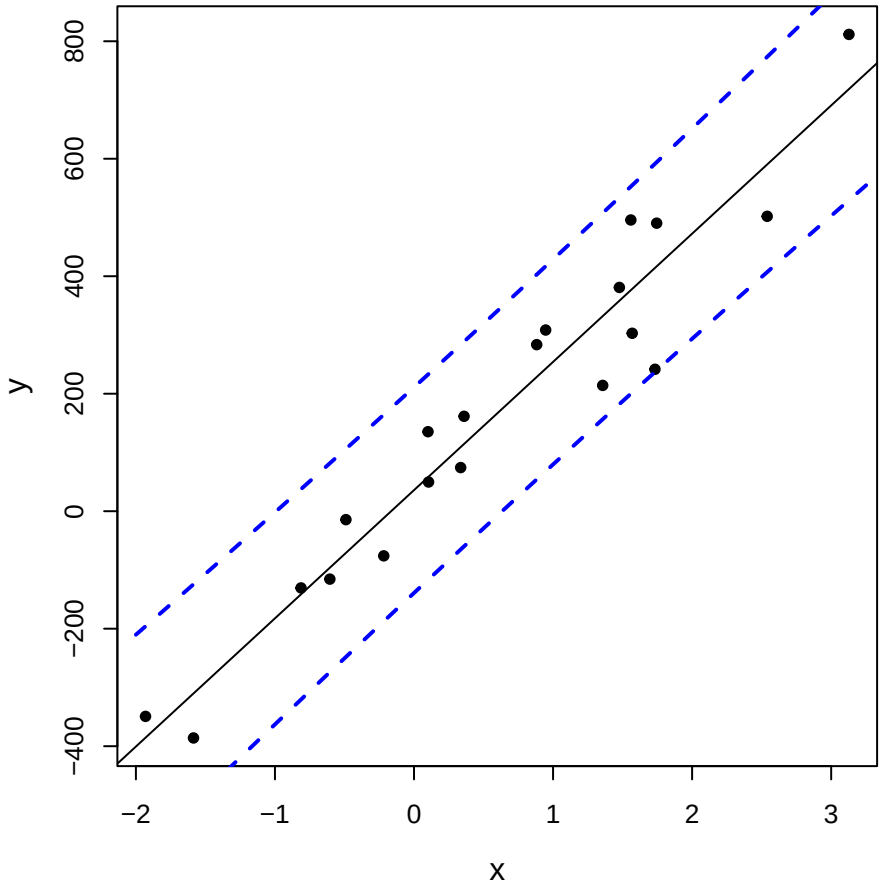
Eksempel: Konfidensinterval for linjen

```r
# Simuler 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), sd = sigma)

# Brug lm() til at fitte modellen
fit <- lm(y ~ x)

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

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

# Tjek værdierne
head(CI)

# Plot  data, regressionslinjen og intervallerne
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: Prædiktionsinterval for observation

```r
# Simuler 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), sd = sigma)

# Brug lm() til at fitte modellen
fit <- lm(y ~ x)

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

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

# Tjek værdierne
head(PI)

# Plot data, regressionslinjen og intervallerne
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 prædiktionsintervaller

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

Dagsorden

Hvad får vi ud af summary() i R?

```
summary(fit)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -172.5  -67.2   16.7   58.9  119.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     35.9       20.0    1.79     0.09 .
## x              218.3       14.0   15.56    7e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 80.9 on 18 degrees of freedom
## Multiple R-squared:  0.931,Adjusted R-squared:  0.927
## F-statistic:  242 on 1 and 18 DF,  p-value: 6.97e-12
```

Summary(lm(y∼x))

- `Residuals:  Min  1Q  Median  3Q  Max`

Summary(lm(y∼x))

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

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"
```

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"`

  Koefficienterne:

  $\hat{\beta}_i$ $\hat{\sigma}_{\beta_i}$ t_{obs} p-værdi
  - Testen er $H_{0,i} : \beta_i = 0$ mod $H_{1,i} : \beta_i \neq 0$.
  - Stjernerne er sat efter p-værdien.

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
  - Testen er $H_{0,i} : \beta_i = 0$ mod $H_{1,i} : \beta_i \neq 0$.
  - Stjernerne er sat efter p-værdien.
- `Residual standard error: XXX on XXX degrees of freedom`

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
  - Testen er $H_{0,i} : \beta_i = 0$ mod $H_{1,i} : \beta_i \neq 0$.
  - 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 brugt i hypotesetest samt konfidens- og prædiktionsintervaller

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
  - Testen er $H_{0,i} : \beta_i = 0$ mod $H_{1,i} : \beta_i \neq 0$.
  - 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 brugt i hypotesetest samt konfidens- og prædiktionsintervaller
- `Multiple R-squared: XXX`

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
  - Testen er $H_{0,i} : \beta_i = 0$ mod $H_{1,i} : \beta_i \neq 0$.
  - 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 brugt i hypotesetest samt konfidens- og prædiktionsintervaller
- `Multiple R-squared: XXX`
  Forklaret varians R^2

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
  - Testen er $H_{0,i} : \beta_i = 0$ mod $H_{1,i} : \beta_i \neq 0$.
  - 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 brugt i hypotesetest samt konfidens- og prædiktionsintervaller
- `Multiple R-squared: XXX`
  Forklaret varians R^2
- Resten bruger vi ikke kurset.

Dagsorden

Forklaret varians og korrelation

- Den forklarede varians i en model er R^2 (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 (dvs. empiriske) korrelation opfylder

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

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

Forklaret varians og korrelation

- Korrelationen ρ er et mål for *lineær sammenhæng* mellem to stokastiske variable.
- Den estimerede (dvs. empiriske) korrelation opfylder

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

  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 β_1 er hældningen i den simple lineære regressionsmodel.

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

```r
# Indlæs data

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 modellen
fit <- lm(y ~ x)

# Scatter plot af data mod regressionslinjen
plot(x,y, xlab = "Height", ylab = "Weight")
abline(fit, col="red")

# Summary
summary(fit)

# Korrelationen mellem X og Y
cor(x,y)

# Den kvadrerede korrelation er "Multiple R-squared"
cor(x,y)^2
```

Dagsorden

Residualanalyse

Metode 5.28

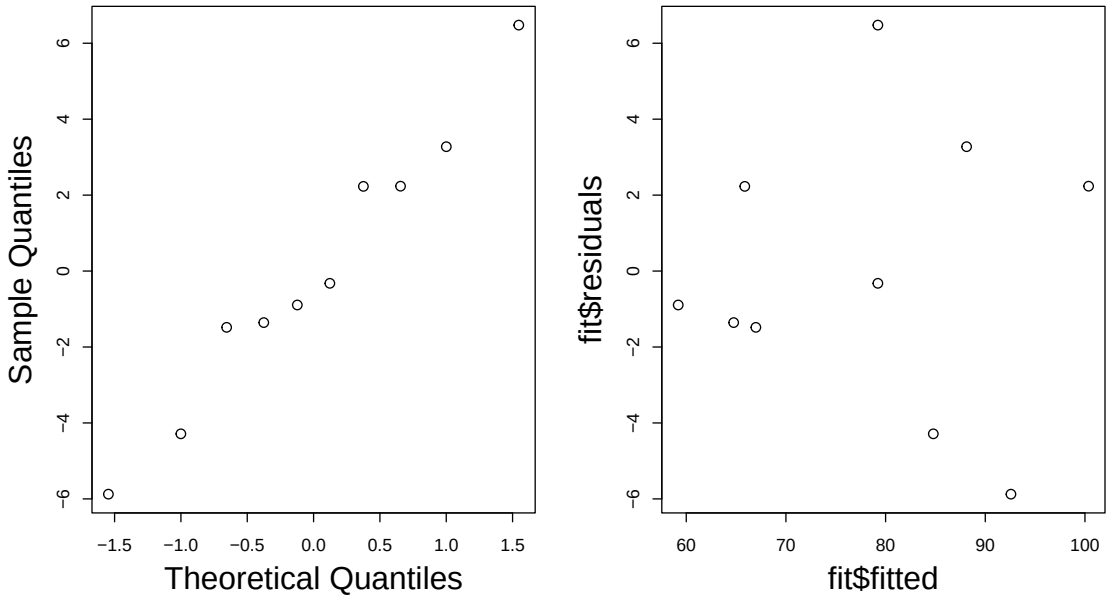
- Undersøg normalitetsantagelse med et qq-plot.
- Undersøg evt. systematiske afvigelser ved at plotte residualerne (e_i) som en funktion af de fittede værdier $(\hat{y}_i)$.

(Metode 5.29)

- Er uafhængighedsantagelsen rimelig?

Modelkontrol 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)
```



Dagsorden

1. Opsummering
2. Eksempel: Højde og vægt
3. Lineære regressionsmodeller
4. Mindste kvadraters metode (Least squares)
5. Statistik og lineær regression
6. Hypotesetest og konfidensintervaller for β_0 og β_1
7. Konfidens- og prædiktionsintervaller
8. Outputtet fra summary
9. Korrelation
10. Modelkontrol - Analyse af residualer