Kursus 02402/02323 Introducerende Statistik

Forelæsning 8: Simpel lineær regression

Peder Bacher

DTU Compute, Dynamiske Systemer
Bygning 303B, Rum 009
Danmarks Tekniske Universitet
2800 Lyngby – Danmark
e-mail: pbac@dtu.dk

Forår 2017

Kapitel 5: Simpel lineær regressions analyse

To variable: x og y

- Beregn mindstekvadraters estimat af ret linje

Inferens med simpel lineær regressionsmodel

- Statistisk model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Estimation, konfidensintervaller og tests for β_0 og β_1
- $1-\alpha$ konfidensinterval for linjen *(Stor sikkerhed for den rigtige linje ligger indenfor)*
- $1-\alpha$ prædiktionsinterval for punkter *(Stor sikkerhed for at nye punkter er indenfor)*

ρ, R og R^2

- ρ er korrelationen ($= sign_{\beta_1} R$) er graden af lineær sammenhæng mellem x og y
- R^2 er andelen af den totale variation som er forklaret af modellen
- Afvises $H_0: \beta_1 = 0$ så afvises også $H_0: \rho = 0$

Chapter 5: Simple linear Regression Analysis

Two quantitative variables: x and y

- Calculate the least squares line

Inferences for a simple linear regression model

- Statistical model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Estimation, confidence intervals and tests for β_0 and β_1.
- $1-\alpha$ confidence interval for the line *(high certainty that the real line will be inside)*
- $1-\alpha$ prediction interval for punkter *(high certainty that new points will be inside)*

ρ, R and R^2

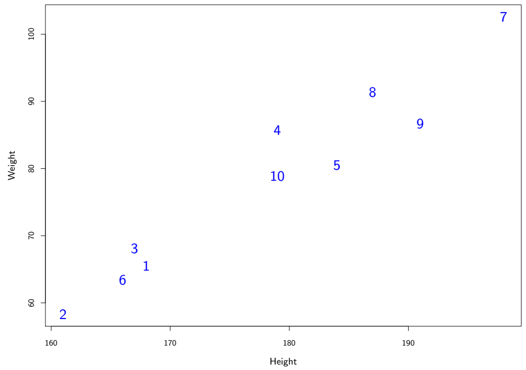
- ρ is the correlation ($= sign_{\beta_1} R$) is the strength of linear relation between x and y
- R^2 is the fraction of the total variation explained by the model
- If $H_0: \beta_1 = 0$ is rejected, then $H_0: \rho = 0$ is also rejected

Oversigt

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. summary(lm()) wrap up
8. Korrelation
9. Model validering: Residual analyse

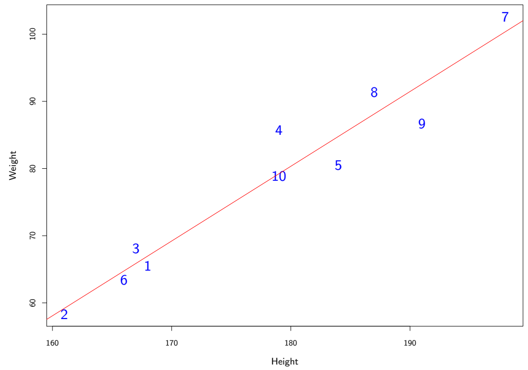
Motiverende eksempel: Højde-vægt

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



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



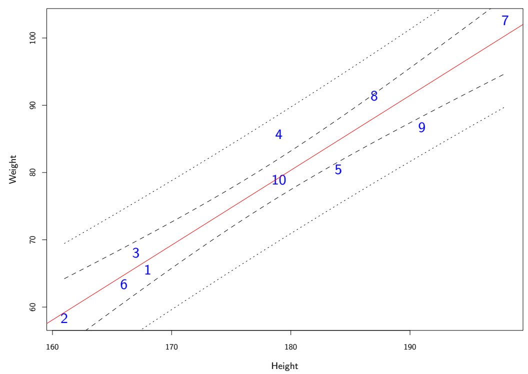
Motiverende eksempel: Højde-vægt

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

```
##
## 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: 3.9 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
```

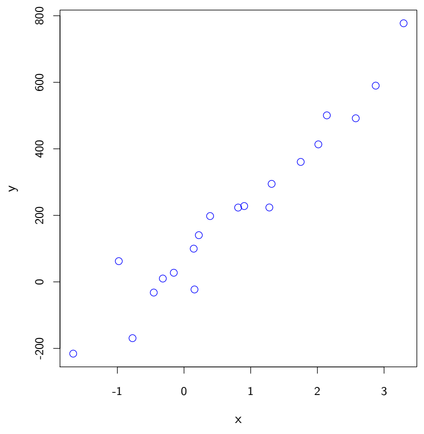

Motiverende eksempel: Højde-vægt

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



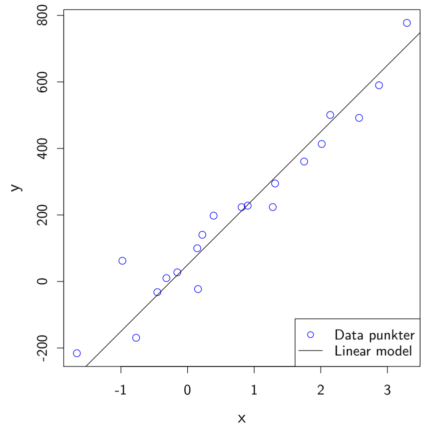
Et scatter plot af nogle punkter. Hvilken model?

- Datapunkter (x_i, y_i)



Kommer de fra en almindelig lineær model?

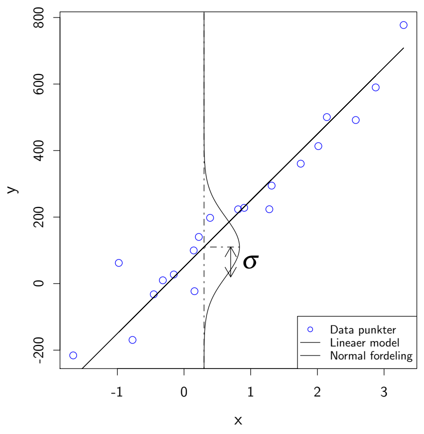
- Opstil en lineær model: $y_i = \beta_0 + \beta_1 x_i$



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

De kommer fra en lineær regressionsmodel

- Opstil en lineær regressionsmodel: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ hvor $\varepsilon_i \sim N(0, \sigma^2)$



Den tilfældige variation er beskrevet med en normalfordeling om linien

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 (epsilon) er afvigelsen (deviation). En stokastisk variabel

og vi antager

ε_i *er independent and identically distributed* (i.i.d.) og $N(0, \sigma^2)$

Mindste kvadraters metode

- Hvis vi kun har datapunkterne, hvordan kan vi 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.

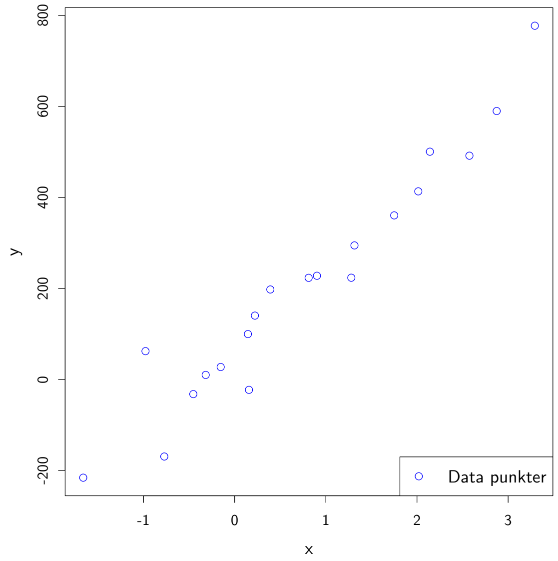
- But how!?

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

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

Dvs. estimaterne $\hat{\beta}_0$ og $\hat{\beta}_1$ er dem som minimerer RSS

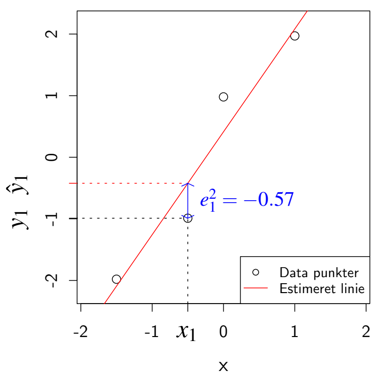
Simuleret eksempel af model, data og fit



Spørgsmål om beregning af residual (socrative.com-ROOM:PBAC)

Udregning af residual for punkt i:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i = \hat{y}_i + e_i \Leftrightarrow$$
$$e_i = y_i - \hat{y}_i$$



Hvad er e_1 her?

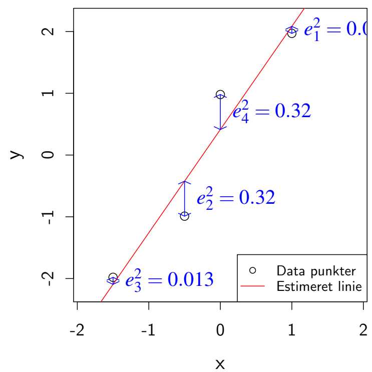
A: ca. -0.57 B: ca. 0.33 C: ca. 1.3 D: Ved ikke

Svar A: ca. -0.57

Spørgsmål om beregning af RSS (socrative.com-ROOM:PBAC)

Beregn:
Residual Sum of Squares (RSS)

Fire punkter, så n=4



Hvad er $RSS = \sum_{i=1}^{n} e_i^2$ her?

A: ca. 0.67 B: ca. 1.65 C: ca. 3.4 D: Ved ikke

Svar A: $RSS = 0.013 + 0.32 + 0.32 + 0.013 = 0.67$

Least squares estimator minimerer RSS

Theorem 5.4 (her for estimatorer som i bogen)

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 estimater minimerer RSS

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

Vi går ikke dybere ind forskellen mellem estimatorer og estimater her i kurset

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 estimaterne, 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 estimaterne
lm(y ~ x)

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


Parameter estimaterne er stokastiske variabler

Hvis vi gentager forsøget vil estimaterne $\hat{\beta}_0$ og $\hat{\beta}_1$ have samme udfald hver gang?

Nej, de er stokastiske variabler. Tager vi en ny stikprøve så vil vi have en anden realisation af dem.

Hvordan er parameter estimaterne fordelt (givet normalfordelte afvigelser)?

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

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

  De er normalfordelte og deres varians kan estimeres:

Theorem 5.7 (første del)

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

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

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

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

Estimater af standardafvigelserne 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

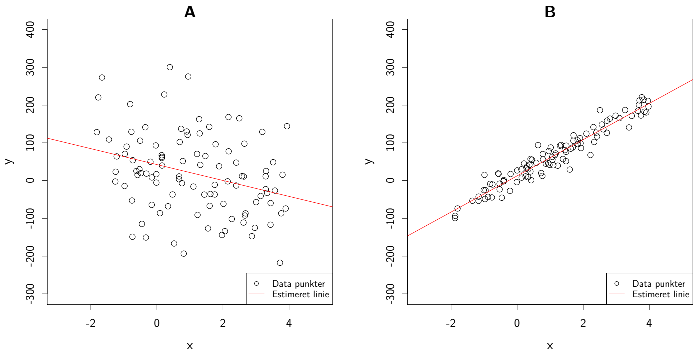
$$\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}.$$

When the estimate of σ^2 is used the variances also become estimates and we"ll refer to them as $\hat{\sigma}^2_{\beta_0}$ and $\hat{\sigma}^2_{\beta_1}$.

- Estimat af standardafvigelserne for $\hat{\beta}_0$ og $\hat{\beta}_1$ (ligningerne (5-41) og (5-42))

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

Spørgsmål: Om fejlenes spredning σ (socrative.com-ROOM:PBAC)

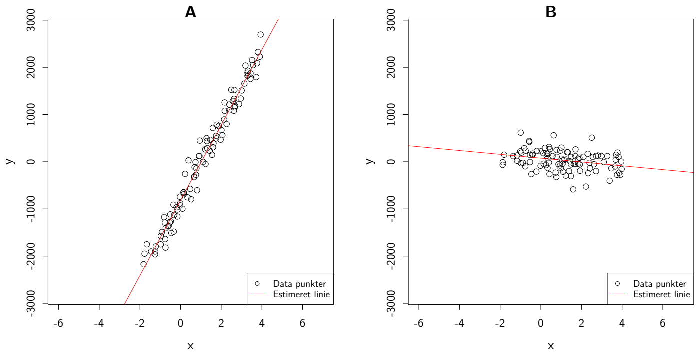


For hvilken er residual variansen $\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$ størst?

A: For fit i plot A B: For fit i plot B C: Lige stor for begge D: Ved ikke

Svar A: For fit i plot A er $\hat{\sigma}$ ca. 100 og for fit i plot B ca. 20

Spørgsmål: Om fejlenes spredning σ (socrative.com-ROOM:PBAC)



For hvilken er residual variansen $\hat{\sigma}^2 = \frac{RSS(\hat{\beta}_0, \hat{\beta}_1)}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}$ størst?

A: For fit i plot A B: For fit i plot B C: Lige stor for begge D: Ved ikke

Svar C: Lige stor for begge, omkring 200

Hypotesetests for parameter parametrene

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

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

- 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.

Eksempel: Hypotesetest for parametrene

- Se Eksempel 5.12 for eksempel på hypotesetest, samt Metode 5.13
- Test om parametrene er signifikant forskellige fra 0

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

- Se resultatet med simulering i R

```
## Hypotesetests for 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 estimaterne
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

$$
\begin{aligned}
&\hat{\beta}_0 \pm t_{1-\alpha/2}\, \hat{\sigma}_{\beta_0} \\
&\hat{\beta}_1 \pm t_{1-\alpha/2}\, \hat{\sigma}_{\beta_1}
\end{aligned}
$$

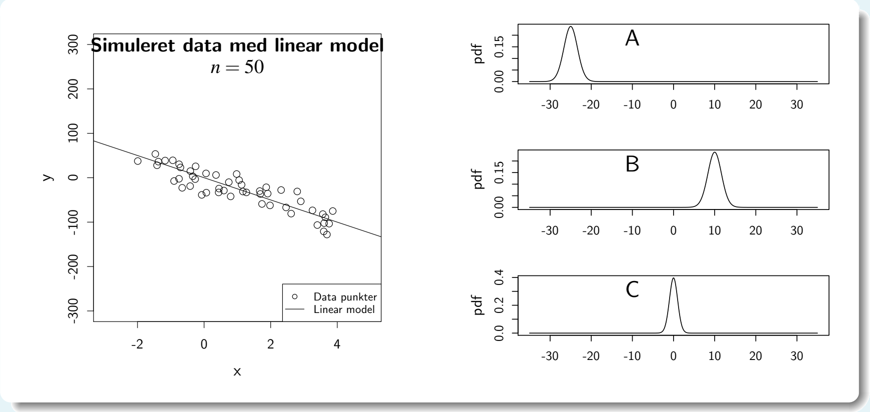
where $t_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of a t-distribution with $n-2$ degrees of freedom.

- husk at $\hat{\sigma}_{\beta_0}$ og $\hat{\sigma}_{\beta_1}$ findes ved ligningerne (5-41) og (5-42)
- i R kan $\hat{\sigma}_{\beta_0}$ og $\hat{\sigma}_{\beta_1}$ aflæses ved "Std. Error" ved "summary(fit)"

Simuleringseksempel: Konfidensintervaller for parametrene

```
## 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 estimaterne
  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])
}

## Hvor ofte blev den rigtige værdi "fanget"?
sum(TrueValInCI) / nRepeat
```

Spørgsmål: Om fordelingen af $\hat{\beta}_1$ (socrative.com-ROOM:PBAC)

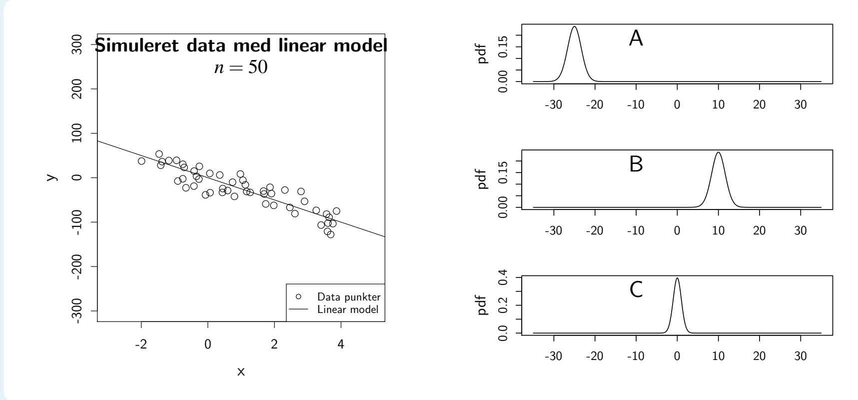


Hvilket plot repræsenterer fordelingen af $\hat{\beta}_1$?

A: Plot A   B: Plot B   C: Plot C   D: Ved ikke

Svar A: β_1 er negativ ($\beta_1 = -25$) og fordelingen af $\hat{\beta}_1$ er centreret i β_1

Spørgsmål: Om fordelingen af $\frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}}$ (socrative.com-ROOM:PBAC)



Hvilket plot repræsenterer fordelingen af $\frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}}$ under H_0: $\beta_{0,1} = -25$?

A: Plot A   B: Plot B   C: Plot C   D: Ved ikke

Svar C: $\frac{\hat{\beta}_1 - \beta_{0,1}}{\hat{\sigma}_{\beta_1}}$ følger under H_0 en t-fordeling, dvs. centreret i 0

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

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

- Der er $100(1-\alpha)\%$ sandsynlighed for at den rigtige linie, altså $\beta_0 + \beta_1 x_0$, er inde i konfidensintervallet

Method 5.17: Prædiktionsinterval for $\beta_0 + \beta_1 x_0 + \varepsilon_0$

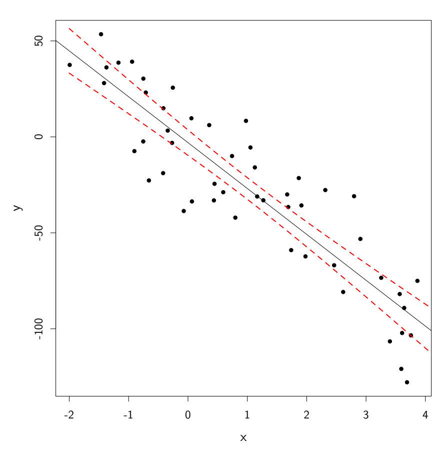
- Prædiktionsintervallet (prediction interval) for Y_0 beregnes for en "ny" værdi af x_i, her kaldt x_0
- Dette gøres *før* Y_0 observeres ved

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

- Der er $100(1-\alpha)\%$ sandsynlighed for at den observerede y_0 vil falde inde i prædiktionsintervallet
- Et prædiktionsinterval bliver altid større end et konfidensinterval for fastholdt α

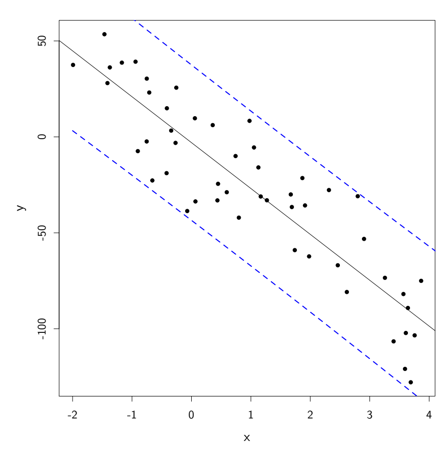
Eksempel med konfidensinterval for linien

```
## Eksempel med konfidensinterval for linien

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

## Brug predict funktionen
CI <- predict(fit, newdata=data.frame(x=xval),
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)
```



Eksempel med prædiktionsinterval

```
## Eksempel med prædiktionsinterval

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

## 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)
```



Hvad bliver mere skrevet ud af summary?

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37.35 -14.08   0.61  14.05  38.96
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.99       3.29   -0.91     0.37
## x             -23.91       1.67  -14.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20 on 48 degrees of freedom
## Multiple R-squared:  0.811,Adjusted R-squared:  0.807
## F-statistic:  206 on 1 and 48 DF,  p-value: <2e-16
```


summary(lm(y∼x)) wrap up

- `Residuals:     Min     1Q   Median    3Q    Max`
  Residualernes: Minimum, 1. kvartil, Median, 3. kvartil, Maximum
- `Coefficients:`
  `Estimate Std. Error t value Pr(>|t|) "stjerner"`
  Koefficienternes:
  Estimat $\hat{\sigma}_{\beta_i}$ t_{obs} 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

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 i data (y_i) der er forklaret med modellen*

Forklaret varians og korrelation

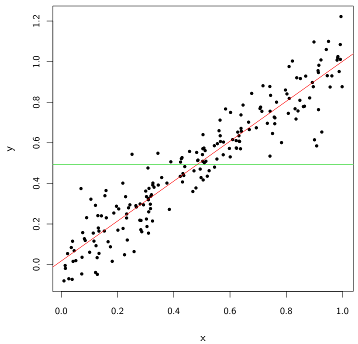
- Korrelationen ρ er et mål for *lineær sammenhæng* mellem to stokastiske variable
- Estimeret (i.e. empirisk) korrelation

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

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

Spørgsmål om korrelation (socrative.com-ROOM:PBAC)



$$r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{1.97}{18.02} = 1 - 0.11 = 0.89 \Leftrightarrow$$

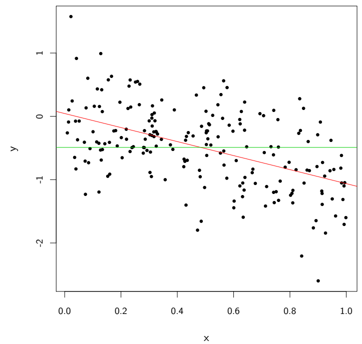
$$r = 0.94$$

Hvad er korrelationen mellem x og y?

A: ca. -0.95 B: ca. 0 C: ca. 0.95

Svar) C: ca. 0.95

Spørgsmål om korrelation (socrative.com-ROOM:PBAC)



$$r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{57.98}{78.32} = 1 - 0.74 = 0.26 \Leftrightarrow$$

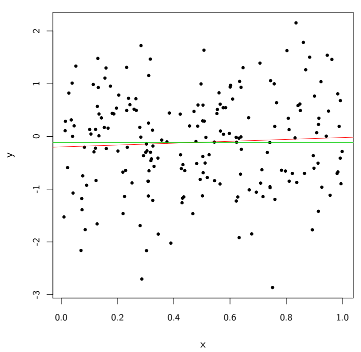
$$r = 0.51$$

Hvad er korrelationen mellem x og y?

A: ca. -0.5 B: ca. 0 C: ca. 0.5

Svar) A: ca. -0.5

Spørgsmål om korrelation (socrative.com-ROOM:PBAC)



$$r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{168.66}{169.18} = 1 - 1 = 0 \Leftrightarrow$$

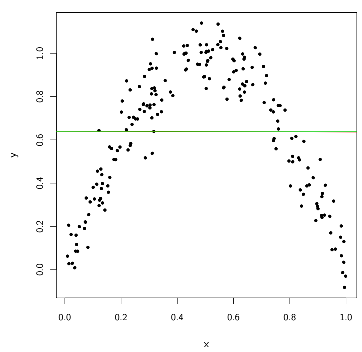
$$r = 0.06$$

Hvad er korrelationen mellem x og y?

A: ca. -0.5 B: ca. 0 C: ca. 0.5

Svar) B: ca. 0

Spørgsmål om korrelation (socrative.com-ROOM:PBAC)



$$r^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{19.81}{19.81} = 1 - 1 = 1.52 \times 10^{-5} \Leftrightarrow$$

$$r = 0$$

Hvad er korrelationen mellem x og y?

A: ca. -0.5 B: ca. 0 C: ca. 0.5

Svar) B: ca. 0

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

Simuleringseksempel om korrelation

```
## 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
```

Model validering: Residual analyse

Method 5.26

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

Residual Analysis in R

```
## Model validering: residual analysis
fit <- lm(y ~ x)
par(mfrow = c(1, 2))
qqnorm(fit$residuals)
qqline(fit$residuals)
plot(fit$fitted, fit$residuals, xlab="Fitted values", ylab="Residuals")
```