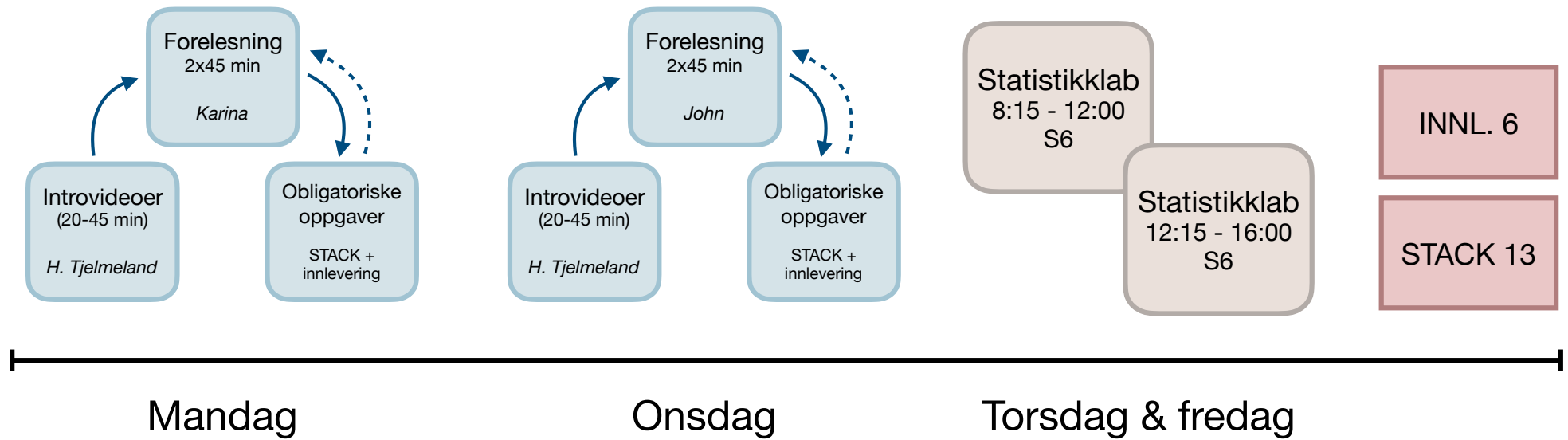


TMA4245 Statistikk

Uke 15 - mandag

TMA4245 Statistikk

Uke 15 - mandag

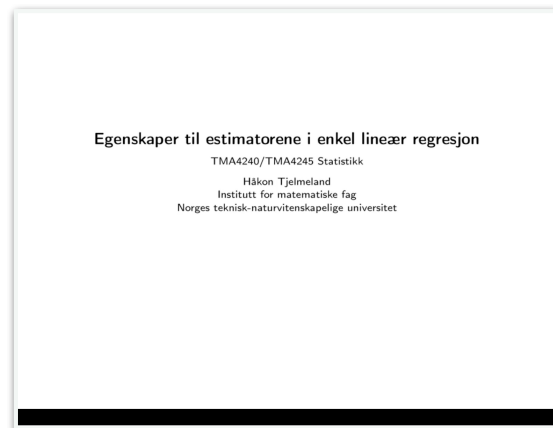


NB: forelesning onsdag
08:15-10:00 i S3

Uke 15 (man 8. april - fredag 12. april)



Introvideoer



Oppgaver

STACK 13: Oppg. 1, 3
INN. 6: Oppg. 1-3

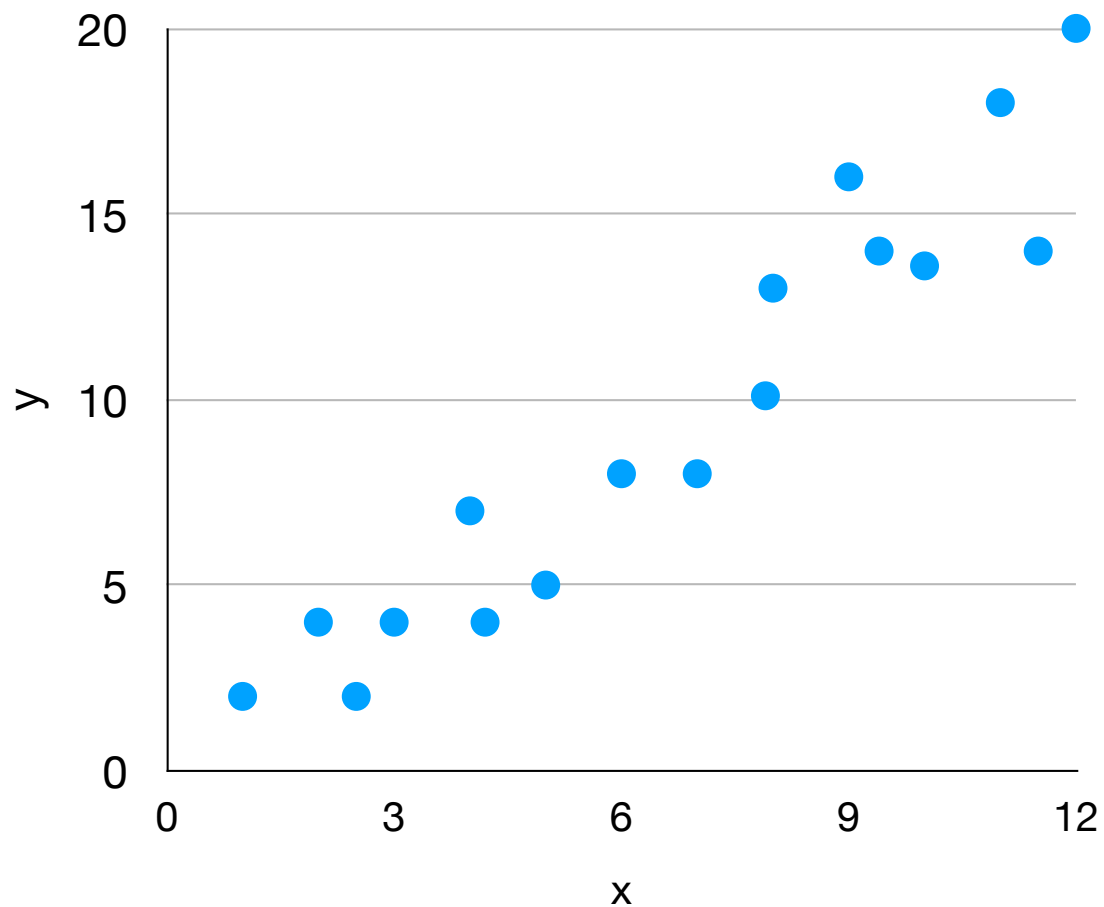
Enkel lineær regresjon

x	y
x_1	y_1
x_2	y_2
...	...
x_n	y_n

- Du får et datasett med to variabler
- Finnes det noen sammenheng mellom disse? Hvordan ser denne sammenhengen ut?
 - Visualiserer dataen

Enkel lineær regresjon

x	y
x ₁	y ₁
x ₂	y ₂
...	...
x _n	y _n



Uavhengighet?

Hvordan type sammenheng har vi?

Modell: Enkel lineær regresjon

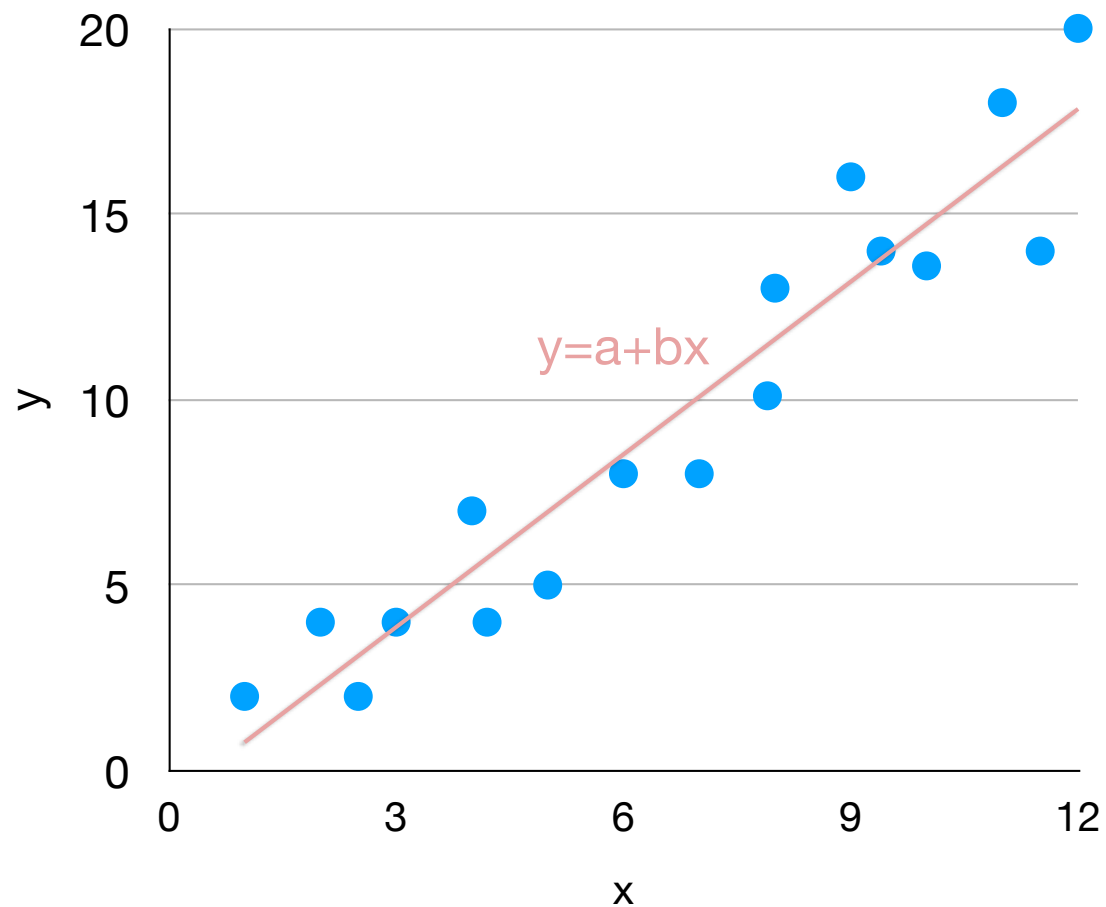
Anta at hver observert y -verdi er en realisasjon av en tilhørende stokastiske variabel Y , og at sammenhengen mellom tilhørende x og Y er gitt av ligningen

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

der β_0 og β_1 er parametre, mens ε er en stokastisk variabel med forventningsverdi $\mathbf{E}[\varepsilon] = 0$ og varians $\mathbf{Var}[\varepsilon] = \sigma^2$. Man har dermed en modell med tre parametre, β_0 , β_1 og σ^2 .

Enkel lineær regresjon

x	y
x ₁	y ₁
x ₂	y ₂
...	...
x _n	y _n



Minste kvadraters metode:

Teorem: Minste kvadraters estimer

Anta at vi har observert parene $(x_i, y_i), i = 1, 2, \dots, n$ og at vi antar en enkel lineær regresjonsmodell for disse dataene. Minste kvadraters estimatene blir da gitt ved

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

der $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ og $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

Sannsynlighetsmaksimeringsprinsippet:

Teorem: SME for enkel lineær regresjon

Anta at $Y_1, Y_2, \dots, Y_n$ er uavhengige stokastiske variabler og at $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ for $i = 1, 2, \dots, n$. Sannsynlighetsmaksimeringsestimatorene for β_0, β_1 og σ^2 basert på $Y_1, Y_2, \dots, Y_n$ er da

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x},$$
$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2,$$

der $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ og $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$.

Egenskaper til estimatorene i enkel lineær regresjon

-Forventningsverdi og varians vises i video, så det antar vi er kjent derfra:

Utgledning av $E[\hat{\beta}]$ og $\text{Var}[\hat{\beta}]$

- * $Y_1, Y_2, \dots, Y_n$ er uavhengige og $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$
- * Estimator for β

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$E[\hat{\beta}] = E \left[\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\text{Var}[\hat{\beta}] = \text{Var} \left[\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\begin{aligned} E[aX] &= aE[X] \\ E[b] &= b \\ E[X + Y] &= E[X] + E[Y] \end{aligned}$$

Utgledning av $E[\hat{\alpha}]$ og $\text{Var}[\hat{\alpha}]$

- * $Y_1, Y_2, \dots, Y_n$ er uavhengige og $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$
- * Estimator for α

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{x}$$

$$E[\hat{\alpha}] = E[\bar{Y} - \hat{\beta} \bar{x}]$$

$$\text{Var}[\hat{\alpha}] = \text{Var}[\bar{Y} - \hat{\beta} \bar{x}]$$

$$\begin{aligned} E[aX] &= aE[X] \\ E[b] &= b \\ E[X + Y] &= E[X] + E[Y] \\ \text{Var}[aX] &= a^2 \text{Var}[X] \\ \text{Var}[b] &= 0 \\ \text{Hvis } X, Y \text{ uavh.:} \\ \text{Var}[X + Y] &= \text{Var}[X] + \text{Var}[Y] \end{aligned}$$

Egenskaper til estimatorene i enkel lineær regresjon

-Forventningsverdi og varians vises i video, så det antar vi er kjent derfra:

Utgledning av $E[\hat{\beta}]$ og $\text{Var}[\hat{\beta}]$

- * $Y_1, Y_2, \dots, Y_n$ er uavhengige og $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$
- * Estimator for β

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$E[\hat{\beta}] = E \left[\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\text{Var}[\hat{\beta}] = \text{Var} \left[\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\begin{aligned} E[aX] &= aE[X] \\ E[b] &= b \\ E[X + Y] &= E[X] + E[Y] \end{aligned}$$

Utgledning av $E[\hat{\alpha}]$ og $\text{Var}[\hat{\alpha}]$

- * $Y_1, Y_2, \dots, Y_n$ er uavhengige og $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$
- * Estimator for α

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{x}$$

$$E[\hat{\alpha}] = E[\bar{Y} - \hat{\beta} \bar{x}]$$

$$\text{Var}[\hat{\alpha}] = \text{Var}[\bar{Y} - \hat{\beta} \bar{x}]$$

$$\begin{aligned} E[aX] &= aE[X] \\ E[b] &= b \\ E[X + Y] &= E[X] + E[Y] \\ \text{Var}[aX] &= a^2 \text{Var}[X] \\ \text{Var}[b] &= 0 \\ \text{Hvis } X, Y \text{ uavh.:} \\ \text{Var}[X + Y] &= \text{Var}[X] + \text{Var}[Y] \end{aligned}$$

-To “triks” som blir utelatt. Viser disse:

(Se notater fra forelesning)

Egenskaper til estimatorene i enkel lineær regresjon

-Forventningsverdi og varians vises i video, så det antar vi er kjent derfra:

Utgledning av $E[\hat{\beta}]$ og $\text{Var}[\hat{\beta}]$

* $Y_1, Y_2, \dots, Y_n$ er uavhengige og $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$

* Estimator for β

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$E[\hat{\beta}] = E \left[\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\text{Var}[\hat{\beta}] = \text{Var} \left[\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\begin{aligned} E[aX] &= aE[X] \\ E[b] &= b \\ E[X + Y] &= E[X] + E[Y] \end{aligned}$$

Utgledning av $E[\hat{\alpha}]$ og $\text{Var}[\hat{\alpha}]$

* $Y_1, Y_2, \dots, Y_n$ er uavhengige og $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$

* Estimator for α

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{x}$$

$$E[\hat{\alpha}] = E[\bar{Y} - \hat{\beta} \bar{x}]$$

$$\text{Var}[\hat{\alpha}] = \text{Var}[\bar{Y} - \hat{\beta} \bar{x}]$$

$$\begin{aligned} E[aX] &= aE[X] \\ E[b] &= b \\ E[X + Y] &= E[X] + E[Y] \\ \text{Var}[aX] &= a^2 \text{Var}[X] \\ \text{Var}[b] &= 0 \\ \text{Hvis } X, Y \text{ uavh.:} \\ \text{Var}[X + Y] &= \text{Var}[X] + \text{Var}[Y] \end{aligned}$$

-To “triks” som blir utelatt.

-Kovarians mellom $\bar{Y}$ og $\hat{\beta}$.

Egenskaper til estimatorene i enkel lineær regresjon

-Forventningsverdi og varians vises i video, så det antar vi er kjent derfra:

Utgledning av $E[\hat{\beta}]$ og $\text{Var}[\hat{\beta}]$

* $Y_1, Y_2, \dots, Y_n$ er uavhengige og $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$

* Estimator for β

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$E[\hat{\beta}] = E \left[\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\text{Var}[\hat{\beta}] = \text{Var} \left[\frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$E[aX] = aE[X]$$

$$E[b] = b$$

$$E[X + Y] = E[X] + E[Y]$$

Utgledning av $E[\hat{\alpha}]$ og $\text{Var}[\hat{\alpha}]$

* $Y_1, Y_2, \dots, Y_n$ er uavhengige og $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$

* Estimator for α

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{x}$$

$$E[\hat{\alpha}] = E[\bar{Y} - \hat{\beta} \bar{x}]$$

$$E[aX] = aE[X]$$

$$E[b] = b$$

$$E[X + Y] = E[X] + E[Y]$$

$$\text{Var}[aX] = a^2 \text{Var}[X]$$

$$\text{Var}[b] = 0$$

Hvis X, Y uavh.:

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$$

$$\text{Var}[\hat{\alpha}] = \text{Var}[\bar{Y} - \hat{\beta} \bar{x}]$$

-To “triks” som blir utelatt.

-Kovarians mellom $\bar{Y}$ og $\hat{\beta}$

<https://tma4245.math.ntnu.no/enkel-lineær-regresjon/egenskaper-til-estimatorene-i-enkel-lineær-regresjon/>

Se under “Bevis: Egenskaper til (hat) Beta_0”

Normalantakelse og fordeling til estimatorene

- SME bestemmes ved å anta $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$
- I minste kvadraters metode antar vi ingen spesifikk fordeling for Y_i
- Legg merke til at regningen av egenskaper til estimatorene ikke bruker normalfordeling, men kun $E[Y_i]$ og $\text{Var}[Y_i]$, og disse er like for begge metoder.
- Vi kan få forventningsrette estimatorer så lenge regningen gjort av Håkon fortsatt er gyldig. Se selv på overgangene - hvilke antakelser gjøres?
- Fordeling til estimatorene bestemmes fra fordelingen til Y_i
- Forventningsverdi til $\hat{\sigma}^2$ bestemmes ved å bruke normalantakelsen av Y_i

6 Noen resultater fra regresjonsanalysen

La $Y_1, \dots, Y_n$ være uavhengige variabler med samme varians σ^2 og forventningsverdier

$$E(Y_i) = \alpha + \beta x_i, \quad i = 1, \dots, n.$$

Minste kvadratsumsestimatorene for α og β er da

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x},$$

og en forventningsrett estimator for σ^2 er gitt ved

$$S^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2.$$

Dersom i tillegg $Y_1, \dots, Y_n$ er normalfordelte vil

$$\frac{(n-2)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

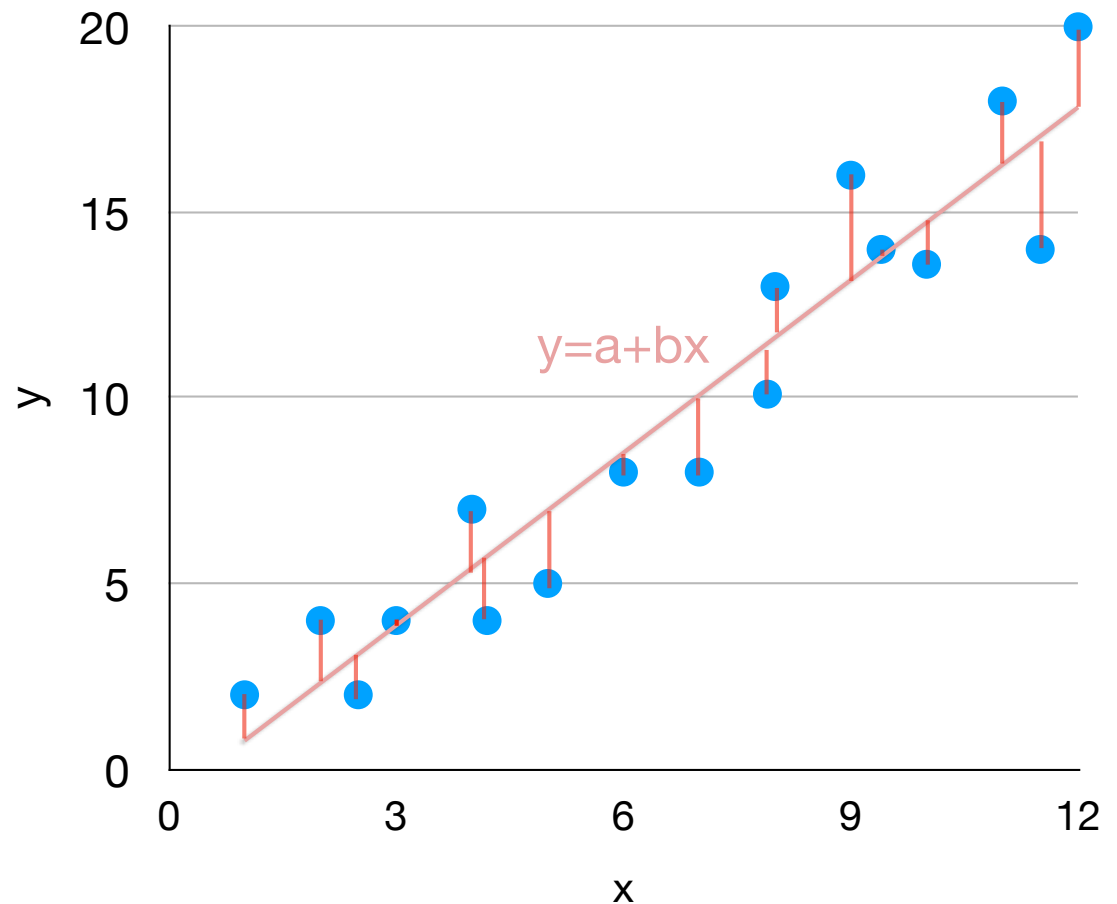
være χ^2 -fordelt med $n-2$ frihetsgrader. Det kan også vises at $(n-2)S^2/\sigma^2$ er uavhengig av $\hat{\alpha}$ og $\hat{\beta}$.

Bli kjent med det som står her!

Residualer - “støyledd”

-Fanger opp hvor “feil” modellen vår er (avvikene i dataen)

x	y
x ₁	y ₁
x ₂	y ₂
...	...
x _n	y _n



Residualer - “støyledd”

- Fanger opp hvor “feil” modellen vår er (avvikene i dataen)
- Antar forventningsverdi lik 0
- Ofte (i TMA4240/TMA4245) normalfordelt, men *må* ikke være det
- Ofte uavhengige og identisk fordelt (iid)
- Hvorfor er det nyttig å studere residualene? Oppgave 2 på innlevering 6 illustrerer dette

Se “Residualer.ipynb” i JupyterHub under “Forelesninger”.

Eksamensoppgave, H2020 oppg. 13 (versjon B)

13B

Innledning: Anta at $Y_1, Y_2, \dots, Y_n$ er uavhengige stokastiske variabler og at

$$Y_i \sim N(\beta \ln(x_i), \sigma^2 x_i^2),$$

der β og σ^2 er parametre, mens $x_1, x_2, \dots, x_n$ er kjente tall. Vi skal i denne oppgaven anta at verdien til β er ukjent, mens verdien til σ^2 er kjent.

Det oppgis at sannsynlighetsmaksimeringsestimatoren for β da er gitt som

$$\hat{\beta} = \frac{\sum_{i=1}^n Y_i \frac{\ln(x_i)}{x_i^2}}{\sum_{i=1}^n \frac{(\ln(x_i))^2}{x_i^2}}.$$

La $y_1, y_2, \dots, y_n$ betegne verdiene vi har observert for de stokastiske variablene $Y_1, Y_2, \dots, Y_n$.

Vis at SME for beta-hatt blir som oppgitt.

Bestem

- forventningsverdi og
- varians til sannsynlighetsmaksimeringsestimatoren, og
- bestem fordeling til estimatoren.

Diverse "regnereregler" som dukker opp (med "bevis"):

$$\sum_{i=1}^n a = \underbrace{a+a+\dots+a}_{n \text{ ganger}} = n \cdot a$$

$$\sum_{i=1}^n ax_i = ax_1 + ax_2 + \dots + ax_n = a(x_1 + \dots + x_n) = a \sum_{i=1}^n x_i$$

$$E\left[\sum_{i=1}^n a_i x_i\right] = E[a_1 x_1 + a_2 x_2 + \dots + a_n x_n] = a_1 E[x_1] + a_2 E[x_2] + \dots + a_n E[x_n] \\ = \sum_{i=1}^n a_i E[x_i]$$

$$\text{Var}\left[\sum_{i=1}^n a_i x_i\right] = \text{Var}[a_1 x_1 + a_2 x_2 + \dots + a_n x_n] \stackrel{\text{uavh.}}{=} \text{Var}[a_1 x_1] + \text{Var}[a_2 x_2] + \dots + \text{Var}[a_n x_n] \\ = a_1^2 \text{Var}[x_1] + \dots + a_n^2 \text{Var}[x_n] \\ = \sum_{i=1}^n a_i^2 \text{Var}[x_i]$$

$$\prod_{i=1}^n a = \underbrace{a \cdot a \cdot a \cdot \dots \cdot a}_{n \text{ ganger}} = a^n$$

$$\prod_{i=1}^n ax_i = ax_1 \cdot ax_2 \cdot \dots \cdot ax_n = \underbrace{a \cdot a \cdot \dots \cdot a}_{n \text{ ganger}} \cdot x_1 \cdot \dots \cdot x_n = a^n \prod_{i=1}^n x_i$$

$$\prod_{i=1}^n a \cdot b^{x_i} = a \cdot b^{x_1} \cdot a \cdot b^{x_2} \cdot \dots \cdot a b^{x_n} = a \cdot a \cdot \dots \cdot a \cdot b^{x_1} \cdot b^{x_2} \cdot \dots \cdot b^{x_n} \\ = a^n b^{x_1 + x_2 + \dots + x_n} \\ = a^n b^{\sum_{i=1}^n x_i}$$

NB (igjen): Merk at forelesning på onsdag 08:15-10:00 er i S3, og ikke R7