

① Lineær regresjon

I denne delen av kurset ser vi på sammenhengen mellom to variable X og Y ved å bruke lineær regresjon.

Modellens forutsetninger:

For hver verdi av X , antar vi at

$$Y = \alpha + \beta X + \varepsilon$$

der α, β er parametre og feil-leddet $\varepsilon \sim N(0, \sigma)$ for en parameter σ . Vi antar også at når $x_1, x_2, \dots, x_n$ er ulike verdier av X , så er de tilsvarende feil-leddene $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ uavhengige.

Data-analyse: to variables

Utvælgelse : n

Variabler : x, y

||

Datasett:

i	x_i	y_i
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
$\vdots$		
n	x_n	y_n

Utvælgingskovarians:

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

utvælgings-
korrrelasjonskoeffisient:

$$r = \frac{S_{xy}}{S_x \cdot S_y}$$

Merk: Vi har $-1 \leq r \leq 1$
for alle datasett

Tolkning: $r > 0$ betyr positiv sammenheng
mellom x og y , dvs større verdier
av x har en tendens til å gi
større verdier av y

$r < 0$ betyr negativ sammenheng,
dvs større verdier av x korre-
leres til å gi mindre verdier
av y (og omvendt)

Jo nærmere $|r|$ er 1, jo sterkere
sammenheng.

$$\begin{array}{ll} r=1: & y = \alpha + \beta x \quad \text{med } \beta > 0 \\ r=-1: & y = \alpha + \beta x \quad \text{" } \beta < 0 \end{array}$$

perfekt lineær korrelasjon

Beregning:

r (og andre deskriptive mål) kan finnes ved å bruke kalkulatorens statistikk-funksjon

Eks:

x	y
7	23
5	19
8	21

Legge inn på kalkulator:

C STAT

tømmer
Statistikk-
minne

7 INPUT 23 $\Sigma+$

5 INPUT 19 $\Sigma+$

8 INPUT 21 $\Sigma+$

Merk:

For å få fram r ,

bruk:

$\hat{x}, r$ SWAP

Tilsvarende for

$\bar{y}, s_y$

Kan bruke statistikk-
funksjoner:

r $\rightarrow$ korrelasjons-
koeffisient

$\bar{x}$

$\bar{y}$

s_x

s_y

} En-variabel
Statistikk

Minste kvadraters metode

Gitt et datasett, ønsker vi å finne den rette linjen $y = \alpha + \beta x$ som "passer best" med datasettet. Det kalles regresjonslinjen.

Mer presist:

For hver α, β kan vi finne ε_i (feil-ledd) slik at

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Vi velger α, β slik at

$$\varepsilon_1^2 + \varepsilon_2^2 + \dots + \varepsilon_n^2$$

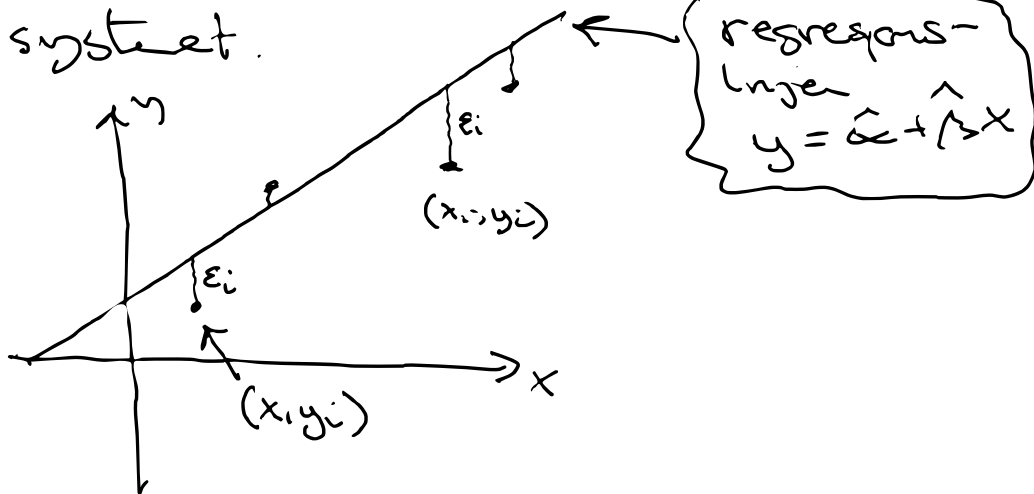
er minst mulig (minste kvadraters metode).

Løsning: Bedste estimat for α om vi velger

$$\hat{\beta} = r_{xy} \cdot \frac{s_y}{s_x} \quad \hat{\alpha} = \bar{y} - \hat{\beta} \bar{x}$$

Spredningsplott:

Et datasett (x_i, y_i) kan fremstilles i et spredningsplott (spredingsdiagram) ved å plotta punktene i xy -koordinatsystemet.



Vi kan tegne inn regresjonslinjen i spredningsplottet. Den har:

$\hat{\beta}$: stigningstall

$\hat{\alpha}$: skjæring med y -aksen

Visuelt kan vi se om linjen passer "godt" til dataene.

Statistiske egenskaper:

$$\left. \begin{aligned} E(\hat{\alpha}) &= \alpha \\ E(\hat{\beta}) &= \beta \end{aligned} \right\} \text{forventningsrette} \\ \text{estimatorer}$$

$$SE(\hat{\alpha})^2 = \text{Var}(\hat{\alpha}) = \frac{\sigma^2 \cdot \sum x_i^2}{n \sum (x_i - \bar{x})^2}$$

$$SE(\hat{\beta})^2 = \text{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

Praktisk bruk:

Estimer σ^2 ved hjelp av

$$S^2 = \frac{1}{n-2} \sum_{i=1}^n \varepsilon_i^2 = \frac{1}{n-2} \cdot \sum_{i=1}^n (y_i - \hat{y})^2 \cdot (1-r^2)$$

⇓

Estimat for $SE(\hat{\beta})^2$:

$$\frac{1}{n-2} \cdot \frac{\sum (y_i - \bar{y})^2}{\sum (x_i - \bar{x})^2} (1-r^2) = \frac{1}{n-2} \cdot \frac{s_y^2}{s_x^2} (1-r^2)$$

Beregne estimat for regresjonslignen:

Med datasettet lagt inn i statistikk-
minne på kalkulator:

$$\boxed{m} \rightarrow \hat{\beta}$$

$$\boxed{b} \rightarrow \hat{\alpha}$$

Anvendelser:

i) Konfidensintervall for β :

$$\hat{\beta} \pm t_{\alpha/2}^{n-2} \cdot SE(\hat{\beta})$$

Mark: $n-2$ dt.

$$SE(\hat{\beta}) = \sqrt{\frac{1}{n-2} \frac{S_y^2}{S_x^2} (1-r^2)}$$

(Tilsvarende for α)

2) Hypotesetest for β

To-sidig: $H_0: \beta = 0$ ($\beta_0 = 0$)
 $H_1: \beta \neq 0$

Testobservator: $T = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})}$

Forlestringsnædd: $|T| > t_{\alpha/2}^{n-2}$

(tilsvarende for højre/venstresidig,
for α)

Mark:

$n-2$ df
 $SE(\hat{\beta}) = \sqrt{\frac{1}{n-2} \frac{s_y^2}{s_x^2} (1-r^2)}$

$\beta \neq 0$ kan tolkes som

"sammenhæng mellem X og Y "

Tolkning av R^2 :

r^2 kan tolkes som

"andelen av total varians i
 y som förklaras av regressions-
linjen", dvs

$$r^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

hvar

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$