Plan:

1. Linear regresjon
2. Statistiske egenskaper
3. Tolkning av R^2

Pensum:

[L] 7.3
(ikke regel 7.6, 7.7)

Husk:
1. Veiledning kl. 11 i dag.
2. Neste forelesning: Repetisjon
   Prøve-eksamen
   Ekstra forelesning: For-eksamen.

Repetisjon:

Korrelasjonskoeffisient: (i et utvalg)

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

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

(utvalgs-kovarians)

$$S_x = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$S_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

utvalgs-std. avvik

$-1 \leq R \leq 1$

$R > 0$: positiv sammenheng

$R < 0$: negativ sammenheng

$|R|$: nær 1 gir sterk sammenheng
nær 0 gir svak — " —

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

$$(n-1) \cdot S_x^2 = \sum_{i=1}^{n} x_i^2 - n \bar{x}^2$$

$$(n-1) S_y^2 = \sum_{i=1}^{n} y_i^2 - n \bar{y}^2$$

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x} \bar{y}) = \sum x_i y_i - \bar{x} \cdot \sum y_i - \bar{y} \sum x_i + \bar{x} \bar{y} \cdot n = \sum x_i y_i - \bar{x} \cdot n \bar{y} - \cancel{\bar{y} n \bar{x}} + \cancel{n \bar{x} \bar{y}}$$

① Lineær regresjon

X: forklaringsvariabel (kontrollert)
Y: responsvariabel (stokastisk)

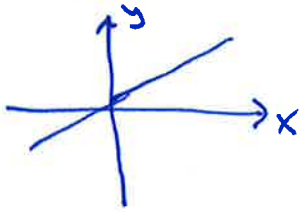
datasett

i	x_i	y_i
1	x_1	y_1
2	x_2	y_2
⋮		
n	x_n	y_n

Antar: Lineær sammenheng

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

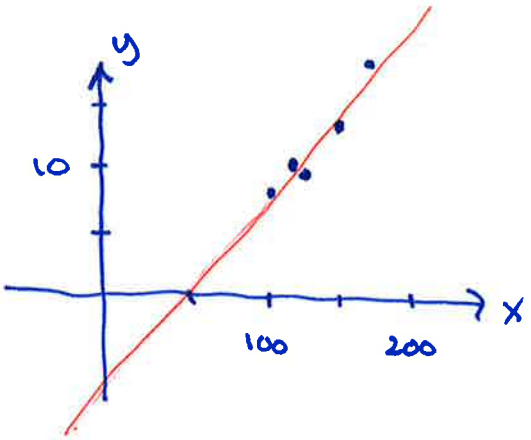
for (ukjente) parametre α, β.



α: skjæring med y-aksen
β: helning / stigningstall

Eks: Boligpriser
X: ant. kvm areal
y: pris (mill. kr)

x	y
109	9.5
181	15.9
112	9.39
102	8.75
158	11.5



Multippel lineær regresjon: Mer enn én forklaringsvariabel

Tilpasse den rette linjen til datasettet

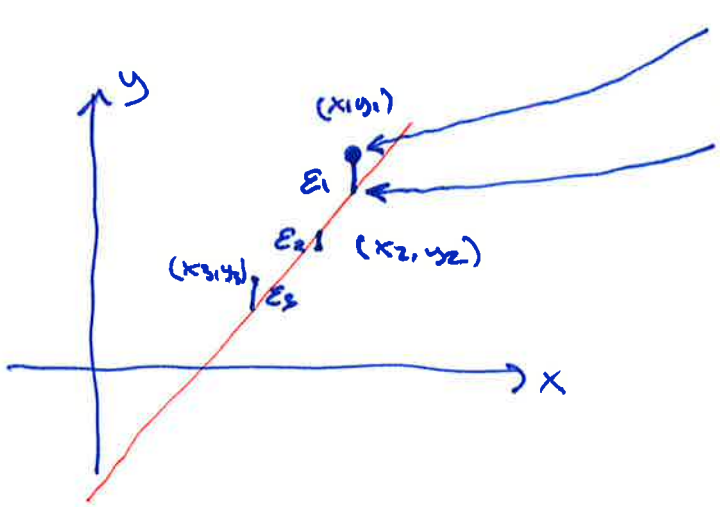
Modell: $Y = \alpha + \beta X + \varepsilon$

ε: feil-ledd

Antagelser:

- x: gitt
- ε_i: $N(0, \sigma^2)$ for en konstant σ
- $\varepsilon_1, \ldots, \varepsilon_n$: uavhengige

$$Y_1 = \alpha + \beta x_1 + \varepsilon_1$$
$$Y_2 = \alpha + \beta x_2 + \varepsilon_2$$
$$\vdots \qquad \vdots$$
$$Y_n = \alpha + \beta x_n + \varepsilon_n$$



$y_i - \hat{y}_i = \varepsilon_i$

Beste tilpasning: (Minste kvadraters metode)

Velg α, β slik at:

$\varepsilon_1^2 + \varepsilon_2^2 + \ldots + \varepsilon_n^2$ ← kvadratiske feilfunksjon

er minst mulig.

Minste kvadraters metode: Finn α, β slik at $\varepsilon_1^2+\varepsilon_2^2+\dots+\varepsilon_n^2$ er minimal

$$\min f(\alpha,\beta) = \varepsilon_1^2+\varepsilon_2^2+\dots+\varepsilon_n^2 = (y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+\dots+(y_n-\hat{y}_n)^2$$

$$= (y_1-\alpha-\beta x_1)^2+(y_2-\alpha-\beta x_2)^2+\dots$$

$$= \sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2$$

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

Løsning:

$$f'_\alpha = \sum_{i=1}^{n} 2(y_i-\alpha-\beta x_i)\cdot(-1)$$

$$= -2\cdot\sum_{i=1}^{n} y_i-\alpha-\beta x_i = 0 \quad |:(-2)$$

$$\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)=0 \qquad \sum_{i=1}^{n}\hat{y}_i-\sum_{i=1}^{n}\alpha-\sum_{i=1}^{n}\beta x_i=0$$

$$n\cdot\bar{y}-n\alpha-\beta\cdot n\bar{x}=0 \qquad \boxed{\bar{y}=\alpha+\beta\bar{x}}$$

$$\alpha = \bar{y}-\beta\bar{x}$$

$$f'_\beta = \sum_{i=1}^{n} 2(y_i-\alpha-\beta x_i)\cdot(-x_i) = 0$$

$$-2\sum_{i=1}^{n}(x_iy_i-\alpha x_i-\beta x_i^2)=0 \quad |:(-2)$$

$$\sum_{i=1}^{n} x_iy_i-\alpha\cdot\sum x_i-\beta\sum x_i^2=0$$

$$\sum x_iy_i-(\bar{y}-\beta\bar{x})\cdot\sum x_i-\beta\cdot\sum x_i^2=0$$

$$\sum x_iy_i-\bar{y}\sum x_i = \beta\left(\sum x_i^2-\bar{x}\sum x_i\right)$$

$$\beta = \frac{\sum x_iy_i-n\bar{x}\bar{y}}{\sum x_i^2-n\bar{x}^2} = \frac{(n-1)\cdot s_{xy}}{(n-1)s_x^2} = \frac{s_{xy}}{s_x^2}$$

Stasjonære punkt: (α, β) gitt ved:

$$\beta = \frac{s_{xy}}{s_x^2} = \frac{s_{xy}}{s_x \cdot s_y} \cdot \frac{s_y}{s_x}$$

r

$$\beta = r \cdot \frac{s_y}{s_x}$$
$$\alpha = \bar{y} - \beta \cdot \bar{x}$$

Merk: ① Dette gir alltid et stasjonært punkt, og vi kan vise at det er et globalt minimum for $f(\alpha, \beta) = \varepsilon_1^2 + \cdots + \varepsilon_n^2$

Se detaljer neste side → $H(f) = \begin{pmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\sum x_i^2 \end{pmatrix}$ $\det > 0$, $A > 0$

② Vi tenker på $\hat{\alpha}, \hat{\beta}$ som estimater for α, β.

Beste estimat for (α, β) = estimatorer for (α, β) →

$$\hat{\beta} = r \cdot \frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2}$$
$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Eks:

x	y
109	9.5
181	15.9
112	9.39
102	8.95
158	11.5

$\hat{\beta} = 0.0790$ ($79.000 kr/m^2$) ← m ← $\bar{y}_m$ SWAP

$\hat{\alpha} = 0.555$ ← b ← $\bar{x}_{w,b}$ SWAP

Beste tilpasning = regresjonslinjen

$$y = 0.555 + 0.079x$$

Forklaring: (α, β) gitt ved $\beta = \frac{S_{xy}}{S_x^2}$, $\alpha = \bar{y} - \beta\bar{x}$ (det stasjonære punkt) gir minimum:

$$f''_{\alpha\alpha} = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)'_\alpha = -2 \cdot \sum_{i=1}^{n}(-1) = -2 \cdot (-1) \cdot n = 2n$$

$$f''_{\alpha\beta} = -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)'_\beta = -2\sum_{i=1}^{n}(-x_i) = 2\sum x_i = 2n\bar{x}$$

$$f''_{\beta\beta} = \sum_{i=1}^{n}\left[2(y_i - \alpha - \beta x_i)(-x_i)\right]'_\beta = 2\sum_{i=1}^{n}(-x_i y_i + \alpha x_i + \beta x_i^2)'_\beta$$

$$= 2\sum_{i=1}^{n} x_i^2$$

$$\Downarrow$$

$$H(f) = \begin{pmatrix} 2n & 2n\bar{x} \\ 2n\bar{x} & 2\sum_{i=1}^{n} x_i^2 \end{pmatrix}$$

konstant matrise siden f har grad 2 i (α, β)

$$\det = \left(2n \cdot 2\sum x_i^2\right) - \left(2n\bar{x}\right)^2$$

$$= 4n \cdot \left(\sum x_i^2\right) - 4n^2\bar{x}$$

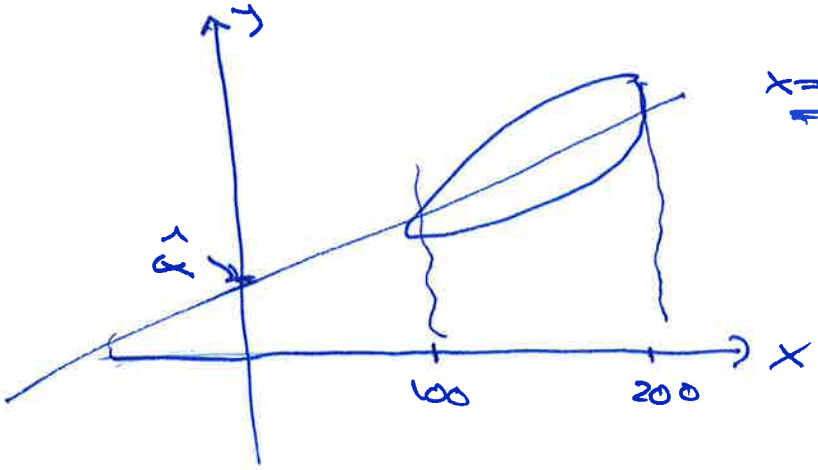
$$= 4n\left(\sum x_i^2 - n\bar{x}\right)$$

$$= 4n\left((n-1)S_x^2\right) = 4n(n-1)S_x^2 > 0$$

$A = 2n > 0$

$$\Downarrow$$

(α, β) er minimumspkt



$x=100$: $\hat{y} = 0.555 + 7.9 = 8.455$

(forventet pris for leil. på 100 m²)

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

$E(Y) = E(\alpha+\beta x+\varepsilon)$

$= \alpha + \beta x + 0$

$= \alpha + \beta x = \hat{y}$

$Var(Y) = Var(\varepsilon) = \sigma^2$

For en gitt x:

$Y \sim N(\hat{y}, \sigma^2)$

$Y = 0.555 + 0.079X$

$\hat{\alpha}$ $\hat{\beta}$

$SS_E = \sum_{i=1}^{n} \varepsilon_i^2$

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

$x=112$: $\hat{y} = 9.397$

112 → $\hat{y}_{|x}$ → 9.397

(2) Statistiske egenskaper

Resultat: $E(\hat{\alpha}) = \alpha$, $E(\hat{\beta}) = \beta$ ← forventningsrette estimatorer

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

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

Praktisk bruk: Må estimere σ^2

$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} \varepsilon_i^2$ ← estimat for σ^2

$= \frac{1}{n-2} \cdot SSE$

Hvordan finne std.feilen: $\hat{\alpha}$, $\hat{\beta}$:

- Finn regresjonslinjen $(\hat{\alpha}, \hat{\beta})$
- Regn ut $SS_E = \sum \varepsilon_i^2 = \sum (y_i - \hat{y}_i)^2$
  og sett $s^2 = \frac{1}{n-2} \cdot SSE$.
- Bruk $\sigma^2 = s^2$ i formlene for $Var(\hat{\alpha})$, $Var(\hat{\beta})$:

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

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum x_i^2 - n\bar{x}^2} = \frac{\sigma^2}{(n-1) \cdot s_x^2} \Rightarrow SE(\hat{\beta}) = \sqrt{Var(\hat{\beta})}$$

Bruk: Hypotesetest for β:

$H_0: \beta = 0 = \beta_0$
$H_1: \beta \neq 0 = \beta_0$

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

Forkastningsområde: $|T| > t_{\alpha/2}$ α = sign.nivå

$t_{\alpha/2}$ ← t-fordelt med $n-2$ d.f.

Konfidensintervall for β:

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

↑ $n-2$ d.f. (frihetsgrader)

③ Tolkning av r^2

Oppdeling av avvik:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

$y_i - \bar{y}$: totalt avvik

$\hat{y}_i - \bar{y}$: avvik forklart av modellen

$y_i - \hat{y}_i$: feilledd ε_i = tilfeldig avvik

Man kan vise:

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

Def:

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = (n-1)s_y^2$$ total sum of squares

$$SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i) = \varepsilon_1^2 + \varepsilon_2^2 + \ldots + \varepsilon_n^2$$ summen av kvadratene av feilledd (residuals)

$$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$ summen av kvadratene av forklarte feil

$$\Downarrow$$

$$SS_T = SS_R + SS_E$$

Merk: Jeg følger her skrivemåten i boken. I litteraturen er det vanlig å bruke TSS for SS_T (total sum of squares), RSS for SS_E (residual sum of squares) og ESS for SS_R (explained sum of squares). Altså:

$$SS_E = RSS$$
$$SS_R = ESS$$

Man kan vise: $r^2 = \frac{SS_R}{SS_T}$

r^2: kvadratet av korrelasjons-koeffisienten r

$\frac{SS_R}{SS_T}$: andelen av total varians som forklares av modellen

Dette betyr:

i) r^2 kan tolkes som andelen av variasjonen i Y som forklares av modellen (regresjonslinjen)

ii) Når vi skal regne ut SS_E for å finne $SE(\hat{\beta})$ kan vi bruke følgende formel:

$$SS_E = SS_T - SS_R$$
$$= SS_T - r^2 \cdot SS_T$$
$$= SS_T \cdot (1 - r^2)$$
$$SS_E = (n-1) s_y^2 \cdot (1 - r^2)$$