Forelesning 14

TMET1190

Plan:

1. Lineær regresjon
2. Statistiske egenskaper

Repetisjon:

Forelesning 12-13: Hypotesetester

$\begin{cases} H_0: & \text{komplementet til } H_1 \\ H_1: & \text{det som skal "vises"} \end{cases}$

Metoder:

- finne forkastningsområde til testobservatør (basert på α)

  $\Rightarrow$ Forkaste H_0 hvis testobservatoren havner i forkast.omr.

  Beholde H_0 ellers

- bruke p-verdi

  p = sannsynlighet for å få en minst like ekstrem verdi på testobs. som den realiserte

Forkaste H_0: $p < \alpha$

Beholde H_0: ellers

(1) Linear regresjon

X: forklaringsvariabel (kontrollert)

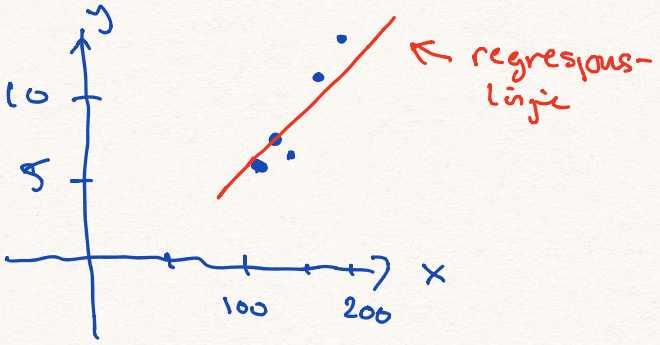
Y: responsvariabel (stokastisk)

Eks: Boligpriser

X: ant. kvm.

y: pris (mill. kr.)

x	y	$\hat{y}$
109	9.5	
181	15.9	
112	9.39	
102	8.75	
158	11.5	



multippel linear regr. → mer enn én forklaringsvar.

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

α, β: parametre med ukjent verdi (α: skj. med y-aksen, β: stign.tall)

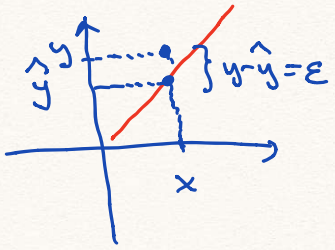
ε: feil-ledd, tilfeldig variasjon

Antagelser:

X: gitt verdi

ε: normalfordelt $N(0, \sigma^2)$ (σ ukjent parameter)

$$Y = \underbrace{\alpha + \beta X}_{\text{regr. linjen } \hat{y}} + \varepsilon$$



$y = \alpha + \beta x$

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

Tilpasning:

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

datasett

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

Husk: α, β ukjente

Minste kvadrekers metode:

Velg α, β slik at $\varepsilon_1^2 + \varepsilon_2^2 + \dots + \varepsilon_n^2$ er minst mulig.

Husk: α, β er de ukjente (vercaldene)

$$\varepsilon_1 = y_1 - (\alpha + \beta x_1)$$
$$\varepsilon_2 = y_2 - (\alpha + \beta x_2)$$
$$\vdots$$
$$\varepsilon_n = y_n - (\alpha + \beta x_n)$$

$$\varepsilon_1^2 + \varepsilon_2^2 + \dots + \varepsilon_n^2 = [y_1 - (\alpha + \beta x_1)]^2 + \dots + [y_n - (\alpha + \beta x_n)]^2$$

$$= y_1^2 - 2y_1 \cdot (\alpha + \beta x_1) + (\alpha + \beta x_1)^2 + \dots$$
$$\dots + y_n^2 - 2y_n (\alpha + \beta x_n) + (\alpha + \beta x_n)^2$$

$$= y_1^2 - 2y_1\alpha - 2x_1y_1\beta + \alpha^2 + 2\alpha\beta x_1 + \beta^2 x_1^2 + \dots$$

$$f(\alpha, \beta) = n\alpha^2 + 2\alpha\beta(x_1 + x_2 + \dots + x_n) + \beta^2(x_1^2 + x_2^2 + \dots + x_n^2) - 2\alpha(y_1 + y_2 + \dots + y_n) - 2\beta(x_1y_1 + x_2y_2 + \dots + x_ny_n) + (y_1^2 + y_2^2 + \dots + y_n^2)$$

Mål:
- Setter inn $(x_1, y_1), \dots, (x_n, y_n)$ fra a datasettet
- løser minimumsproblemet

$$\min f(\alpha, \beta)$$

Sett $f(\alpha,\beta) = \varepsilon_1^2 + \dots + \varepsilon_n^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$

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

Vi løser optimeringsproblemet: $\min f(\alpha,\beta)$

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

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

$$\Sigma y_i = \Sigma(\alpha + \beta x_i) = n \cdot \alpha + (\Sigma x_i)\beta$$

$$n\bar{y} = n\alpha + n\bar{x}\beta \quad | : n$$

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

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

$$= -2 \Sigma (x_i y_i - \alpha x_i - \beta x_i^2) = 0 \quad | : (-2)$$

$$\Sigma x_i y_i = \alpha \Sigma x_i + \beta \Sigma x_i^2$$

fra

$$= (\bar{y} - \beta\bar{x}) \cdot \Sigma x_i + \beta \Sigma x_i^2$$

$$\Sigma x_i y_i - \bar{y} \Sigma x_i = \beta (\Sigma x_i^2 - \bar{x} \Sigma x_i)$$

$$\Sigma x_i y_i - n\bar{x}\bar{y} = \beta (\Sigma x_i^2 - n\bar{x}^2)$$

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

$$\beta = \frac{S_{xy}}{S_x^2} = \frac{S_{xy}}{S_x \cdot S_y} \cdot \frac{S_y}{S_x} = r \frac{S_y}{S_x}$$

$$\beta = r \cdot \frac{S_y}{S_x}$$

Oppsummering:

$f'_\alpha = 0$
$f'_\beta = 0$

gir

$\alpha^* = \bar{y} - \beta \bar{x}$
$\beta^* = r \cdot S_y / S_x$

dvs stasjonære punkt

Klassifisering:

$$H(f): \quad f''_{\alpha\alpha} = \sum_{i=1}^{n} 2 \cdot (-1) \cdot (-1) = \underline{2n}$$

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

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

$$= \Sigma -2x_i(-x_i) = \underline{2 \sum_{i=1}^{n} x_i^2}$$

$$\det H(f) = (2n)(2 \Sigma x_i^2) - (2n\bar{x})^2$$

$$= 4n \Sigma x_i^2 - 4n^2 \bar{x}^2$$

$$= 4n \left(\Sigma x_i^2 - n\bar{x}^2 \right) = 4n \cdot (n-1) s_x^2 \geq 0$$

Siden $f(\alpha, \beta)$ er kvadratisk utbryggbar

med $\det H(f) = 4n(n-1)s_x^2 \geq 0$

og $\operatorname{tr} H(f) = 2n + 2\sum x_i^2 > 0$

så er det stasjonære punkt.

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

$$\beta^* = r \cdot \frac{s_y}{s_x}$$

et minimum for $f(\alpha, \beta)$.

Dette punktet gir derfor den rette linjen $y = \alpha + \beta x$ som passer best til datasettet.

Hovedtrekk:

i) Finn stasjonære pkt.: (α^*, β^*)

$$f'_\alpha = 0$$
$$f'_\beta = 0$$

ii) Klassifisere stasjonære pkt.

$$H(f)\,|\,(\alpha^*, \beta^*)$$

Resultat: et stasjonært pkt, lokalt og globalt min.

Kan vise mellomregningene på notatene etter forelesning: Resultat:

$$\beta^* = \frac{S_{xy}}{S_x^2} \qquad \alpha^* = \bar{y} - \beta^* \bar{x}$$

Konklusjon:

Regresjonslinjen er gitt ved $y = \hat{\alpha} + \hat{\beta} x$

der

$$\hat{\beta} = \frac{S_{xy}}{S_x^2} = r \cdot \frac{S_y}{S_x}$$

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

$\hat{\alpha}, \hat{\beta}$: estimatorer for α, β

Beregning av $\hat{\alpha}$, $\hat{\beta}$:

x	y
109	9.5
(8)	15.9
112	9.39
102	8.75
158	11.5

$$\hat{\beta} = r \cdot \frac{s_y}{s_x}$$

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

[STAT]

109 [INPUT] 9.5 [Σ+]

⋮

158 [INPUT] 11.5 [Σ+]

[$\hat{x}$, r] [SWAP] 0.946 = r

[s_x] 35.0 = s_x

[s_x, s_y] [SWAP] 2.92 = s_y

[$\bar{x}$] 132.4 = $\bar{x}$

[$\bar{x}$, $\bar{y}$] [SWAP] 11 = $\bar{y}$

x (første) : forkl. var.
y (andre) : respons var.

$\hat{\alpha}$, $\hat{\beta}$ direkte:

$\hat{\beta}$: [$\hat{y}$,m] [SWAP]

$\hat{\beta} = 0.079$

En kvadratmeter extra øker prisen med 79.000 kr.

$\hat{\alpha}$: [$\bar{x}$ w 1b] [SWAP]

$\hat{\alpha} = 0.555$

0 kvm gir pris 555.000 kr

Estimat for regresjonslinjen:

$$y = 0.555 + 0.079x$$

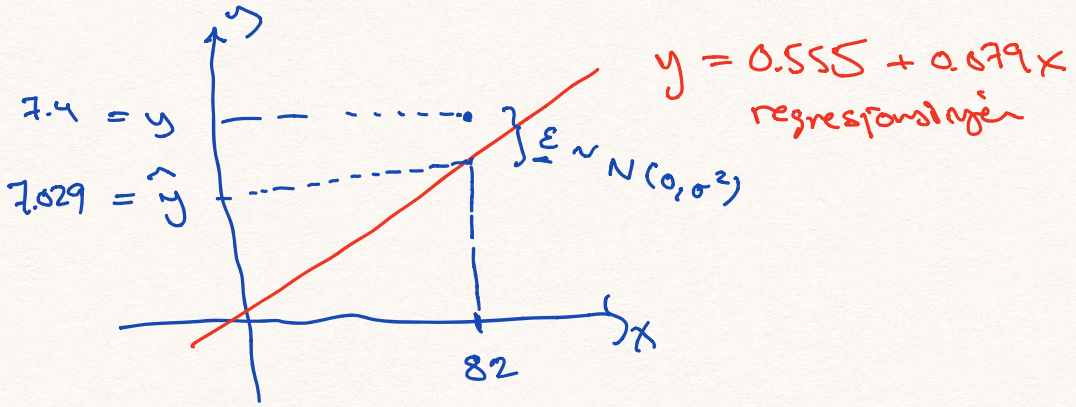
Predikspjner: $y = 0.555 + 0.079x$

$x = 82$: $\hat{y} = 0.555 + 0.079 \cdot 82$

(82 kvm) Kalk: 82 $\boxed{\hat{y}_m}$ $\rightarrow 7.029 = \hat{y}$

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

$x = 82$: $y \sim N(7.029, \sigma^2)$



Husk: Anter at $\varepsilon \sim N(0, \sigma^2)$
Anter at vi har et tilfeldig utvalg
slik at $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ er uavhengige

(2) Statistiske egenskaper

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

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

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

Resultat:

$$Var(\hat{\alpha}) = \frac{\sigma^2 \cdot \sum_{i=1}^{n} x_i^2}{n \cdot \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}$$

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

$$SE(\hat{\alpha}) = \sqrt{Var(\hat{\alpha})}$$
$$SE(\hat{\beta}) = \sqrt{Var(\hat{\beta})}$$
} Std. feil = Std. avvik

Praktisk bruk: Må estimere σ^2 ved å bruke s^2

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{SS_E}{n-2}$$

$$SS_E = \sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + \dots + e_n^2$$

Eks: Hypotesetest for β

(X: forklaringsvariabel
Y: responsvariabel)

$H_0: \beta = 0 = \beta_0 = 0$ ← ingen sammenheng mellom X og Y

$H_1: \beta \neq 0$ ← det finnes en sammenheng mellom X og Y

eller

$H_0: \beta \leq 0 = \beta_0$ ← ingen eller neg. sammenheng mellom X og Y

$H_1: \beta > 0$ ← pos. sammenheng

testobservator: ~~t~~ eller $\hat{\beta}$

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

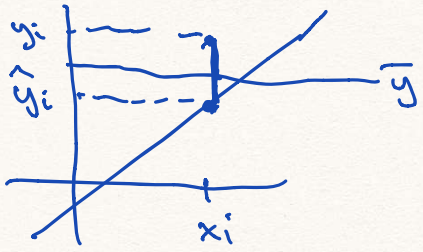
- kan regne ut $\hat{\beta} = r \cdot \frac{S_Y}{S_X}$ fra datasettet
- må estimere $SE(\hat{\beta})$ for å finne T.

Tolkning av R^2:

Oppdeling av avvik:

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

totalt avvik — avvik forklart av regresjonen — feilledd ε_i



Man kan vise:

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

$SS_E = \varepsilon_1^2 + \dots + \varepsilon_n^2$

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

$$= \frac{1}{n-1}\cdot SS_T$$

$$\Updownarrow$$

$$\boxed{SS_T = (n-1)\,s_y^2}$$

↑ kan finne dette på kalkulator

$$SS_T = SS_R + SS_E$$

$$\Updownarrow$$

$$\boxed{SS_E = SS_T - SS_R}$$

↑ enklest å regne ut

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

$\Downarrow$

$$SS_R = r^2 \cdot SS_T$$

Tolkning: R^2 er andelen av total varians forklart av regresjonsmodellen

Vi kombinerer disse formlene:

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

$$SS_E = (n-1) s_y^2 \cdot (1 - r^2) \quad = \varepsilon_1^2 + \varepsilon_2^2 + \dots + \varepsilon_n^2$$

Brukes dette slik:

σ^2 estimeres av $s^2 = \frac{1}{n-2} \cdot SS_E$

$$= \frac{1}{n-2} \cdot (n-1) s_y^2 \cdot (1 - r^2)$$

$$s^2 = \frac{n-1}{n-2} s_y^2 (1 - r^2)$$

$$\frac{1}{n-1} SS_T = \frac{1}{n-1} SS_R + \frac{1}{n-1} SS_E$$

s_y^2

variansen (total varias) i y

den delen av total varias forklart av modellen (regr. linje)

den delen av variansen som ikke er forklart ulka modellen.